Homework assignments Week 14 (Students should submit their homework before 21 p.m. on January 1, 2022.)

6. The proximal operator of the l_0-norm (2%):

Let $\mathbb{I}\{\mathcal{A}\}$ be an indicator function such that $\mathbb{I}\{\mathcal{A}\} = 1$ if $\mathcal{A}$ is true, and $\mathbb{I}\{\mathcal{A}\} = 0$ otherwise. The l_0-norm of a p-dimensional vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ is defined by

$$||\boldsymbol{\beta}||_0 = \sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\},$$

i.e. the number of non-zero valued elements in $\boldsymbol{\beta}$. The proximal operator of l_0-norm on vector $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ is defined by

$$\mathrm{HT}_\alpha(\mathbf{x}) = \arg\min_{\boldsymbol{\beta}} \left\{ \alpha||\boldsymbol{\beta}||_0 + \frac{1}{2}||\boldsymbol{\beta} - \mathbf{x}||_2^2 \right\}.$$

Here the proximal operator $\mathrm{HT}_\alpha(\mathbf{x})$ is a p-dimensional vector, and its jth element is denoted by $[\mathrm{HT}_\alpha(\mathbf{x})]_j$.

Which of the following statements are *true*?

a. The proximal operator of l_0-norm is also called the **hard-thresholding** operator since it can be expressed as

$$[\mathrm{HT}_\alpha(\mathbf{x})]_j = x_j \mathbb{I}\{|x_j| \geq \sqrt{2\alpha}\}.$$

b. If $\boldsymbol{\beta}^* = \mathrm{HT}_\alpha(\mathbf{x})$, then we have

$$\alpha||\boldsymbol{\beta}^*||_0 + \frac{1}{2}||\boldsymbol{\beta}^* - \mathbf{x}||_2^2 = \sum_{j=1}^{p} \min\{(1/2)|x_j|, \alpha\}.$$

c. Since $||\boldsymbol{\beta}||_0 \in \{0, 1, 2, \cdots, p\}$ is discrete-valued, the optimization problem associated with $\mathrm{HT}_\alpha(\mathbf{x})$ is a combinatorial optimization problem, and the most efficient way for computing $\mathrm{HT}_\alpha(\mathbf{x})$ will still cost $O(2^p)$ flops

Programming work

7. l_0-norm regularized estimation via the fast proximal gradient algorithm (6%)

In this programming work we will build an iterative scheme based on the **fast proximal gradient algorithm** to find an l_0-norm regularized estimate of regression coefficients in the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{\text{true}} + \boldsymbol{\epsilon}.$$

The l_0-norm regularized estimate for $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}^{\text{HT}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_0 \right\},$$

where $\alpha \geq 0$ is a user-specified tuning parameter.

To build a fast proximal gradient algorithm to find $\widehat{\boldsymbol{\beta}}^{\text{HT}}$, you need to:

1. Derive the gradient of the loss function $(1/2)||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$ with respect to $\boldsymbol{\beta}$.

2. Specify the stepsize c_r. Here you can use whatever way you like to specify the stepsize c_r. But I would like to suggest the following value:

   $$c_r = \frac{1}{2\lambda_1(\mathbf{X}^T\mathbf{X})},$$

   where $\lambda_1(\mathbf{X}^T\mathbf{X})$ is the maximum eigenvalue of the Gram matrix $\mathbf{X}^T\mathbf{X}$.

3. Specify the initial values for running the iterative scheme. You may follow what has been suggested in the course slides to set the initial values.

4. Specify (a) a **stopping criterion**, (b) a **tolerance** for the error and (c) the **maximum number of iterations** for stopping the iterative scheme. Here you are allowed to use whatever way you like to specify the **stopping criterion** like the following one:

   Some measure on error $\leq$ tol OR The number of iterations $>$ max_iter.

However the tolerance for the error and the maximum number of iterations should be

$$\begin{aligned} \text{tol} &= 5 \times 10^{-6}, \\ \text{max\_iter} &= 10000. \end{aligned}$$

5. Like the lasso estimation, the l_0-norm regularized estimation also has a tuning parameter α. You need to decide what value α should have in order to yield a good estimate for the regression coefficients. Be aware that in practice, the tuning parameter may be scaled by the stepsize c_r and becomes

$$\tilde{\alpha} = c_r \alpha. \tag{1}$$

Therefore when you choose a tuning parameter value, what you should consider is the value of $\tilde{\alpha}$. You may choose a value of $\tilde{\alpha}$ from the following interval:

$$\tilde{\alpha} \in [0.001\tau, \tau],$$

where $\tau \geq 0$ is the upper bound of the interval. You should think about how to decide the value of τ. One principle is that you should have $\widehat{\boldsymbol{\beta}}^{\text{HT}} = \mathbf{0}$ when you carry out your own l_0-norm regularized estimation at τ.

Data generation: We use a setting the same as the one used in **Question 5 in Week 13 homework assignment** to generate the data. However this time you can specify whatever size of **training data** n you like to carry out your own l_0-norm regularized estimation. In addition, this time I want you to generate **test data** to evaluate performance of **prediction** from your own model. The size of the test data $(\mathbf{y}^{\text{test}}, \mathbf{X}^{\text{test}})$ you use may be $n^{\text{test}} = 10 \times n$, where n is the size of the training data.

Tasks: You need to report two things: one is the plots of iteration error against iteration number, and the other one is prediction performance (in table format. See below). For the former:

1. Fix tuning parameter $\tilde{\alpha}$ at 4 different values you like and run two iterative schemes based on the (a) **fast proximal gradient algorithm** and (b) **proximal gradient algorithm** under the 4 different values of $\tilde{\alpha}$ separately. Produce 4 plots according to the 4 different values of $\tilde{\alpha}$ with the following format: The x-axis is the number of iterations r and the y-axis is the **Euclidean norm** of $\boldsymbol{\beta}^{r+1} - \boldsymbol{\beta}^{r}$, difference between update at $(r+1)$th iteration and rth iteration, of the iterative scheme. Please report the iteration error for (a) the **fast proximal gradient algorithm** in **red** color and (b) the **proximal gradient algorithm** is **black** color.

For the prediction performance, report the following values from your prediction:

1. The size of training data n;
2. The size of test data n^{pred}.
3. The chosen tuning parameter value $\tilde{\alpha}^*$ that leads to the best performance in prediction.
4. The mean squared error of $\widehat{\boldsymbol{\beta}}^{\text{HT}}$, which is defined by
$$\text{MSE} = ||\widehat{\boldsymbol{\beta}}^{\text{HT}} - \boldsymbol{\beta}^{\text{true}}||_2^2,$$
5. The training error, which is defined by
$$\text{Err}^{\text{train}} = \frac{||\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{HT}} - \mathbf{y}||_2^2}{n},$$
where $\mathbf{y}$ is the response vector in the training data.
6. The test error, which is defined by
$$\text{Err}^{\text{test}} = \frac{||\mathbf{X}^{\text{test}}\widehat{\boldsymbol{\beta}}^{\text{HT}} - \mathbf{y}^{\text{test}}||_2^2}{n^{\text{test}}}.$$

You may report the results in the Table 1 format.

n	n^{test}	$\tilde{\alpha}^*$	MSE	$\text{Err}^{\text{train}}$	Err^{test}
-	-	-	-	-	-

Table 1: Table format.