Assignment 1

1

1.1 Proximal Gradient Descent

$$arg \min_{w \in R^d} ||w||_1 + \sum_{i=1}^{n} \left(y^i - < w^i, x^i > \right)^2$$

$$arg \min_{w \in R^d} ||w||_1 + || Xw - y ||_2^2$$

$Regularizer: r(w) = || w ||_1$
$Loss\ function: l(w) = || Xw - y ||_2^2$

We know that l(w) is a differentiable function hence

$$\nabla l(w) = 2X^T (Xw - y)$$

$$prox_r(u) = arg \min_z \left(r(z) + \frac{1}{2} ||z-u||_2^2 \right)$$

$$prox_r(u) = arg \min_z \left(||z||_1 + \frac{1}{2} ||z-u||_2^2 \right)$$

$Let\ f(z) = ||z||_1 + \frac{1}{2}||z-u||_2^2$
We know that $||z||_1$, $\frac{1}{2}||z-u||_2^2$ are convex functions, hence $prox_r(u)$ is a convex function. So the local minima, we obtain by first order optimality will give us the global minima.
But f(z) is not differentiable, so we calculate the sub-differential for the i^{th} co-ordinate.

$$\partial_i f(z) = \begin{cases} -1 + z_i - u_i , & z_i < 0 \\ 1 + z_i - u_i , & z_i > 0 \\ k \mid k \in [-1 + z_i - u_i , 1 + z_i - u_i] , & z_i = 0 \end{cases}$$

By $1st$ order optimality we have $\partial f(z_0) = 0$.

$$arg \min_z f(z) = z_0$$

$$prox_r (u_i) = \begin{cases} u_i + 1 , & u_i < -1 \\ u_i - 1 , & u_i > 1 \\ 0 , & -1 \le u_i \le 1 \end{cases}$$

It can be mathematically proven that for $prox_r(u)$ to converge coordinate need to be shrunk by value step length η
Hence prox functionused by us is

$$prox_r (u_i) = \begin{cases} u_i + \eta , & u_i < -1 \\ u_i - \eta , & u_i > 1 \\ 0 , & -1 \le u_i \le 1 \end{cases}$$

Preprint. Under review.

In t^{th} iteration
$g^t \in \partial l(w)$

$$g^t = \nabla l(w^t) = 2X^T (Xw - y)$$
$$u^{t+1} = w^t - \eta g^t$$
$$w^{t+1} = prox_r (u^{t+1})$$

For proximal gradient descent we found step length of **0.04** to be best suited.

We also used a simple acceleration function given below to speed up convergence

$$w^{t+1} = prox_r(u^{t+1}) + \frac{t-1}{t+2}(prox_r(u^{t+1}) - w^t)$$

We initialized w vector with random values st. $w_i \in [0, 1]$ for all $0 \leq i \leq d$ as it yields the best result.

1.2 Stochastic Coordinate Descent

$$arg \min_{w \in R^d} ||w||_1 + \sum_{i=0}^{n} \left(y^i - < w^i, x^i > \right)^2$$

$$f(w) = arg \min_{w \in R^d} ||w||_1 + \sum_{i=0}^{n} \left(y^i - \sum_{j=0}^{d} w_j x_j^i \right)^2$$

We observe that the given function is convex and non-differential.
Therefore to calculate f(x) we need to deal with the sub-differential of the given function.

$$\partial_j ||w||_1 = \begin{cases} -1 , & w_j < 0 \\ [-1 , 1] , & w_j = 0 \\ 1 , & w_j > 0 \end{cases}$$

$$\partial_j f(w) = \begin{cases} -1 , & w_j < 0 \\ [-1 , 1] , & w_j = 0 \\ 1 , & w_j > 0 \end{cases} - 2x_j^i \sum_{i=0}^{n} \left(y^i - \sum_{k \neq j,\ k=0}^{d} w_k x_k^i - w_j x_j^i \right)$$

$$\partial_j f(w) = \begin{cases} -1 , & w_j < 0 \\ [-1 , 1] , & w_j = 0 \\ 1 , & w_j > 0 \end{cases} - 2 \sum_{i=0}^{n} \left(x_j^i y^i \sum_{k \neq j,\ k=0}^{d} w_k x_k^i \right) + 2w_j \sum_{i=0}^{n} x_j^2$$

Let $\sum_{i=0}^{n} \left(x_j^i y^i \sum_{k \neq j,\ k=0}^{d} w_k x_k^i \right) = p_j$ and $\sum_{i=0}^{n} x_j^2 ; = z_j$

$$\partial_j f(w) = \begin{cases} 2w_j z_j - 2p_j - 1 , & w_j < 0 \\ [\, 2w_j z_j - 2p_j - 1 , 2w_j z_j - 2p_j + 1 \,] , & w_j = 0 \\ 2w_j z_j - 2p_j + 1 , & w_j > 0 \end{cases}$$

Using first order optimality condition

$$w_j = \begin{cases} \frac{p_j + \frac{1}{2}}{z_j} , & p_j < \frac{-1}{2} \\ 0 , & \frac{-1}{2} \leq p_j \leq \frac{1}{2} \\ \frac{p_j - \frac{1}{2}}{z_j} , & p_j > \frac{1}{2} \end{cases}$$

We used cyclic coordinate selection for selecting next coordinate, i.e.

$$next\_i = \begin{cases} 0 , & i = n \\ i + 1 , & otherwise \end{cases}$$

We initialized w vector with all zeroes as it yields the best result.

2

2.1 Proximal Gradient Descent

For proximal gradient descent, we found acceleration function

$$w^{t+1} \;=\; prox_r(u^{t+1}) \;+\; \frac{t-1}{t+2}\,(prox_r(u^{t+1}) - w^t)$$

and step length as best solutions.
We used held out validation for this, keeping 200 data points as validation set.
Firstly we tried various step lengths $\{0.1,0.5,1,0.05,0.08,0.07,0.03,0.02,0.01\}$ and saw which ones reduced the objective value well. We realized going above 0.08 was causing overstepping and below 0.01 results in no significant change in the objective value for long.
Then we tried some acceleration function with step lengths of $\{0.03,0.04,0.05,0.06,0.07,0.08\}$ on our validation set.
We tried

$$w^{t+1} \;= prox_r(u^{t+1}) \;+\; f(t)(prox_r(u^{t+1}) \;-\; w^t)$$

$$f(t) \;=\; \left\{\frac{t}{t+1}\,,\; \frac{t-1}{t+2}\,,\; \frac{t-2}{t+1}\,,\; \frac{t-1}{t+1}\right\}$$

We found $f(t) \;=\; \frac{t-1}{t+2}$ worked best for us with **step length = 0.04**
After this we tried step lengths in $\{0.035,0.038,0.039,0.04,0.041,0.042,0.0401,0.0402,0.045\}$
We found that 0.04 was the most suited step length.
We also tried to batch update the gradient, trying various batch sizes $\{10,50,100,150,200,500,800\}$, But we found that 800 was the best suited batch size. Hence we dropped the idea of batch updating the gradient.
We tried 3 initialization for w as vector with all zeroes, all ones and to a random value vector st. $w_i \;\in$ [0,1] for all $0 \;\leq\; i \leq\; d$. We got best results when we used the last case. So we initialised w in that fashion.

2.2 Stochastic Coordinate Descent

For coordinate descent we tried cyclic and random coordinate selection methods and found cyclic worked best for us.
We did this using 3 fold validation dividing the given data in 3 sets of 250, 250, 300.
Also we tried 3 initialization for w as vector with all zeroes, all ones and to a random value vector st. $w_i \;\in$ [0,1] for all $0 \;\leq\; i \leq\; d$. But we did not observe much improvements in the latter two cases, so we chose to keep w as vector with all zeroes.

3

On implementing both the models (Accelerated Proximal Gradient Descent and Stochastic Coordinate Descent), we found that both the models converged on same values, but Accelerated Proximal Gradient Descent converged faster.
So according to us **Accelerated Proximal Gradient Descent** is the best.

1 CS771 Graph.jpeg

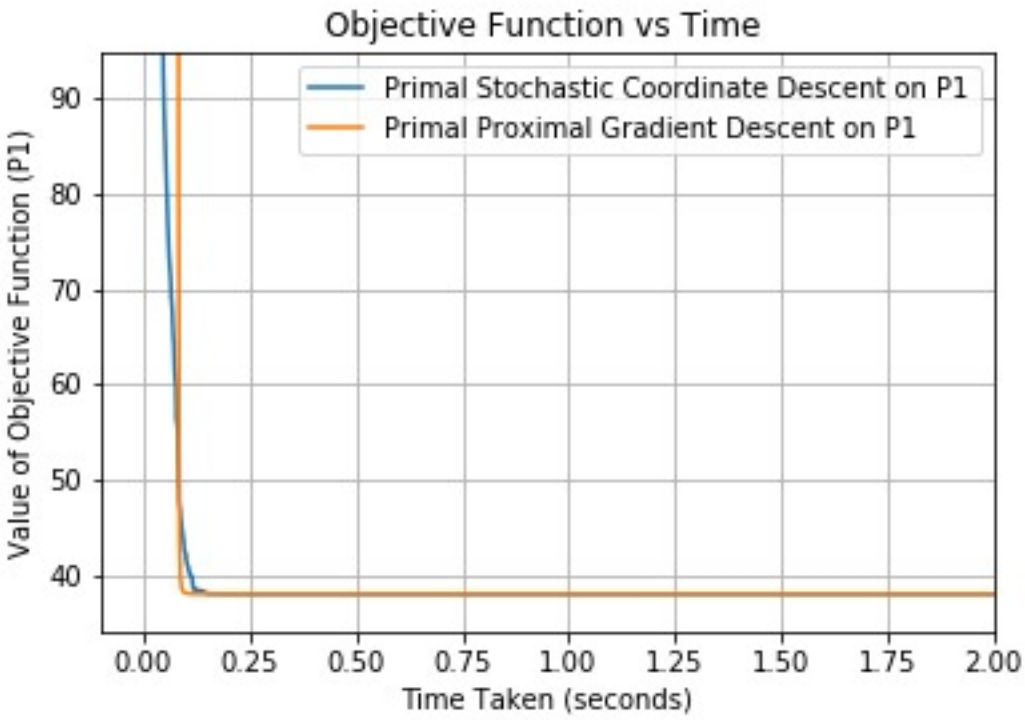

Figure 1: Graph between Objective Function(P1) and time