

# CS 395T: Homework 4

Brady Zhou  
brady.zhou@utexas.edu

November 10, 2017

## 1 Proximal Gradient Theory

Consider the following optimization problem

$$\sum_{i=1}^n \|A_i x_i - b_i\|^2 + \lambda \sum_{1 \leq i < j \leq n} \|x_i - x_j\|_1 \quad (1)$$

Here, we seek to optimize a function that is a sum of a convex, smooth function  $g$  and a convex function  $h$ . We can do this by making an initial guess  $x^{(0)}$ , and iterating with the recurrence

$$x^{(\tau+1)} \leftarrow \text{prox}_h(x^{(\tau)} - \gamma \nabla_x g(x^{(\tau)})) \quad (2)$$

where  $\gamma$  is some step size and

$$\text{prox}_h(x) = \underset{u}{\operatorname{argmin}} \lambda \sum_{1 \leq i < j \leq n} \|u_i - u_j\|_1 + \frac{1}{2} \sum_{i=1}^n \|u_i - x_i\|^2 \quad (3)$$

This  $\text{prox}_h(\cdot)$  is still not so easy to evaluate, but a key observation is that we can solve this one dimension at a time, that is

$$\text{prox}_h(x^{(k)}) = \underset{u}{\operatorname{argmin}} \lambda \sum_{1 \leq i < j \leq n} |u_i - u_j| + \frac{1}{2} \sum_{i=1}^n (u_i - x_i^{(k)})^2 \quad (4)$$

Now,  $u$  is a list of scalars and from here on the superscript  $x_i^{(k)}$  will be omitted.

In this form, we can make two observations.

$$\text{if } x_i < x_j, \text{ then } u_i \leq u_j \quad (5)$$

$$\text{if } x_i = x_j, \text{ then } u_i = u_j \quad (6)$$

To show (5), consider the contrary. That  $x_i < x_j$  and  $u_i > u_j$ . Now, we claim there is an  $\hat{u}_i, \hat{u}_j$  such that  $\hat{u}_i = u_j, \hat{u}_j = u_i$  that will decrease the objective function. For brevity, the left term is omitted since the sum of absolute values does not change with a swap. Similarly, the right term only changes in two terms.

$$\begin{aligned} \frac{1}{2}(u_i - x_i)^2 + \frac{1}{2}(u_j - x_i)^2 &\leq \frac{1}{2}(\hat{u}_i - x_i)^2 + \frac{1}{2}(\hat{u}_j - x_j)^2 && \text{By claim } u_i, u_j \text{ is optimal} \\ (u_i - x_i)^2 + (u_j - x_j)^2 &\leq (\hat{u}_i - x_i)^2 + (\hat{u}_j - x_j)^2 \\ (u_i^2 - 2u_i x_i + x_i^2) + (u_j^2 - 2u_j x_j^2 + x_j^2) &\leq (\hat{u}_i^2 - 2\hat{u}_i x_i + x_i^2) + (\hat{u}_j^2 - 2\hat{u}_j x_j + x_j^2) \\ (u_i^2 - 2u_i x_i + x_i^2) + (u_j^2 - 2u_j x_j^2 + x_j^2) &\leq (u_j^2 - 2u_j x_i + x_i^2) + (u_i^2 - 2u_i x_j + x_j^2) \\ -2u_i x_i - 2u_j x_j &\leq -2u_j x_i - 2u_i x_j \\ 2u_i(x_j - x_i) &\leq 2u_j(x_j - x_i) \\ u_i &\leq u_j && \text{Contradiction} \end{aligned}$$

A similar argument follows to show (6).

Now, we will assume  $x$  is sorted and add constraints to  $prox_h(\cdot)$  to make it into a form more manageable (getting rid of absolute values). (5), (6) will now be encoded into the Lagrangian of the form

$$L(u, \alpha) = -\lambda \sum_{1 \leq i < j \leq n} (u_j - u_i) + \frac{1}{2} \sum_{i=1}^n (u_i - x_k)^2 + \sum_{i=1}^{n-1} \alpha_i (u_{i+1} - u_i) \quad (7)$$

However, this form is not entirely correct as it is lacking the equality constraints. To fix this, we will “bin” the  $x_i$  into unique tuples of the form  $(x_i, l_i)$ , where  $l_i$  denotes the count. The new Lagrangian is of the form

$$L(u, \alpha) = -\lambda \sum_{1 \leq i < j \leq n} l_i l_j (u_j - u_i) + \frac{1}{2} \sum_{i=1}^n l_i (u_i - x_i)^2 + \sum_{i=1}^{n-1} \alpha_i (u_{i+1} - u_i) \quad (8)$$

Taking the gradient,

$$\nabla_{u_k} L = -\lambda \left( -\sum_{i=1}^{k-1} l_i l_k + \sum_{i=k+1}^n l_i l_k \right) + l_k (u_k - x_k) + \alpha_{k-1} - \alpha_k \quad (9)$$

We know at a stationary point, the gradient is 0, so

$$u_k = x_k + \lambda \left( -\sum_{i=1}^{k-1} l_i + \sum_{i=k+1}^n l_i \right) + \frac{1}{l_k} (\alpha_k - \alpha_{k-1}) \quad (10)$$

Finally, to solve  $prox_h(\cdot)$ , we will evaluate

$$b_k = x_k + \lambda \left( -\sum_{i=1}^{k-1} l_i + \sum_{i=k+1}^n l_i \right) \quad (11)$$

If for all  $k$ , we have  $b_k < b_{k+1}$ , then we set  $u_k = b_k$ , and we are done. This is true because all the constraints are satisfied, so the problem is equivalent to the unconstrained optimization (that is,  $\alpha_i = 0$ ). Using the unconstrained gradient, we can see these conditions at  $b_k$  are sufficient to be a minimizer.

On the other hand, if there exists a  $b_k$  such that  $b_k \geq b_{k+1}$ , we must do something a bit counter intuitive and change the objective function. We will now update  $\hat{x}_k = \hat{x}_{k+1} = \frac{l_i x_i + l_j x_j}{l_i + l_j}$ , and merge the two counts, that is, update  $\hat{l}_i = \hat{l}_j = l_i + l_j$ . We will end up removing the duplicate tuple. We can in fact perform this change to the optimization problem because this updated problem is equivalent to the previous, up to an additive constant that is not dependent on  $u$ , the variable we are optimizing over.

## 2 Proximal Gradient Coding

The runtime of this algorithm was  $O(dn \log n)$  per iteration, where  $n$  is the number of vectors, and  $d$  is the number of dimensions each vector has.

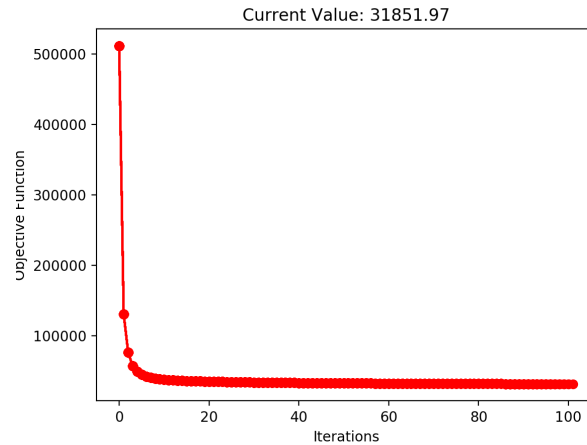


Figure 1: Proximal Gradient

**Note:** This implementation is not fully correct. If at each iteration, we simply update  $x^{\tau+1} \leftarrow x^{\tau} - \gamma \nabla f(x^{\tau})$ , we end up with an objective value score of 40000, but with the true proximal gradient update, we end up with a score of 32000, as shown above. However, this proximal gradient uses a hack, which is shown in the code.

### 3 Convex Set Hyperplane Separation

Let  $\mathcal{A}, \mathcal{B}$  be two disjoint, nonempty, **compact**, convex sets. We wish to show that there exists a hyperplane with normal  $w$  and scalar  $b$  such that

$$\langle x, w \rangle \geq b \quad \forall x \in \mathcal{A} \quad (12)$$

$$\langle y, w \rangle \leq b \quad \forall y \in \mathcal{B} \quad (13)$$

To find  $w, b$ , consider the solution  $u^*, v^*$  to the minimization problem

$$\underset{\substack{u \in \mathcal{A} \\ v \in \mathcal{B}}}{\operatorname{argmin}} ||u - v||^2 \quad (14)$$

Note that  $u^*, v^*$  are not necessarily unique. Then, we have

$$w = (u^* - v^*) \quad (15)$$

$$b = w^\top \left( \frac{u^* + v^*}{2} \right) \quad (16)$$

To show this hyperplane does in fact separate  $\mathcal{A}$  and  $\mathcal{B}$ , consider the contrary - that  $\exists x \in \mathcal{A}$  such that  $\langle x, w \rangle < b$ .

We can reformulate the previous minimization problem into a function  $d(u) = ||u - v^*||^2$ , where  $u \in \mathcal{A}$ . Taking the gradient, we can see  $\nabla d(u) = 2(u - v^*)$ .

Now we consider an infinitesimal change in the  $x - u^*$  direction at a point  $u^*$ .

$$\begin{aligned} \left( \nabla d(u^*) \right)^\top (x - u^*) &= 2(u^* - v^*)^\top (x - u^*) \\ &= 2w^\top (x - u^*) \\ &= 2w^\top x - 2w^\top u^* \\ &< b - 2w^\top u^* \\ &= w^\top (u^* - v^*) \\ &= -||w||^2 \\ &< 0 \end{aligned}$$

Since this dot product is negative, this direction is in fact a descent direction to the function  $d$ . This direction is still in  $\mathcal{A}$  by convexity, so this is a clear contradiction to the assumption that  $u^*, v^*$  are optimal. We can conclude that this hyperplane separates  $\mathcal{A}$  and  $\mathcal{B}$ . □