Chapter 8

Abstract

In this chapter we discuss the problem of sparse model selection, i.e. how to perform optimization when the desired result is constrained by rank. As the problem does not have a natural convex translation, we introduce the iterative hard thresholding (IHT) algorithm and prove results about its performance.

Over the run of this course, we have mostly been discussing problems of the form

$$\min_{x \in C} f(x)$$

where $f(x)$ represents a convex objective and $x \in C$ represents some convex constraint. In this lecture, we will be discussing a case in which $x \in C$ is *not* a convex constraint, namely, when it is the requirement that x be k - sparse.

In order to think about this problem, let's introduce the simplest non trivial version of this problem. The *sparse linear regression* problem is defined as follows.

$$\min \|y - Ax\|_2^2$$

such that

$$\|x\|_0 \leq k$$

And furthermore, we impose the restriction that $x \in \mathbb{R}^p$ then $k << p$. How should we deal with this constraint? We can try to think in terms of previously introduced concepts and ideas.

- We can try convexification. Thus, we redefine the problem to read

  $$\min \|y - Ax\|_2^2$$

  such that

  $$\|x\|_1 \leq \lambda$$

  But it's really not clear how to do this, or how to pick λ.

- We can try to convexify with an eye towards proximal gradient descent.

  $$\min \|y - Ax\|_2^2 + p \|x\|_1$$

  and use the following update rule

  $$x_{t+1} = \text{Prox}_{p\|\cdot\|_1}(x_t - \eta \nabla f(x_t))$$

  And try to 'bias' the solution towards sparsity using the L_1 norm. But again, it's not clear how we can do this safely.

- We keep the non convexity and use non convex projected gradient descent. So we keep the formulation of

  $$\min \|y - Ax\|_2^2$$

  such that

  $$\|x\|_0 \leq k$$

  But we introduce a new update rule

  $$x_{t+1} = H_k(x_t - \eta \nabla f(x_t))$$

  Perhaps something along the lines of sorting the input with respect to magnitude and selecting the k largest ones.

In general, the L_0 pseudonorm introduces exponentially hard solutions. If we try to select the k size subsets from p, we experience combinatorial explosion. However, the key to focus on here is the words "in general". This lecture will focus on problems of this type that will prove an exception to this rule and admit polynomial complexity.

To this end we introduce the iterative hard thresholding algorithm. We have

$$x_{t+1} = H_k(x_t - \eta \nabla f(x_t))$$

where $H_k(z) = \arg\min_{\|x_0\| \leq k} \|x - z\|_2^2$

Before continuing discussion on this algorithm, lets first get a sense of what the hyperparameters we are dealing with are.

- Starting point
- Step size (η)
- Sparsity level

For a second, let's imagine that we were dealing with the simple case of $A = I$. Then, we would end up with a new problem formulation - to minimize

$$f(x) = \|y - x\|_2^2$$

subject to the k - sparsity restriction. We have seen this problem before in an earlier homework assignment, this is not something that is difficult to solve. If possible, we should try to pin down the property of I that makes such a solution possible.

This property is called *restricted isometry*. We have that there exists some δ for all x_1, x_2 such that

$$(1 - \delta) \|(\| x_1 - x_2)_2^2 \leq \|I(x_1 - x_2)\|_2^2 \leq (1 + \delta) \|x_1 - x_2\|_2^2$$

Intuitively, this means that the matrix I does not perturb the distance between x_1 and x_2 too much. This can be easily seen for the case of I, but that does not mean it is the only matrix with such a property. In the literature, other properties like nullspace properties and eigenvalues properties are considered, but for this lecture we will be honing in on the restricted isometry property (henceforth RIP).

As an aside, not that verifying if a matrix satisfies the RIP is NP-hard. Therefore, let us take for granted we have sucha a matrix and now attempt to prove convergence given this restriction.

Convergence Proof

We will start be recalling some relevant details to the proof.

$$x_{t+1} = H_k(x_t - \eta \nabla f(x_t))$$

For the linear regression problem $\nabla f(x_t) = A^\top (y - Ax_t)$. Therefore, we can rewrite

$$x_{t+1} = H_k(x_t - \eta A^\top (y - Ax_t))$$

And now, we may ask : can't we just use the result from convex projected gradient descent?

But the answer is unfortunately no. Before, we use the fact that

$$\|H_k(x) - H_k(y)\|_2 \leq \|x - y\|_2$$

But that's not true anymore. Consider the following two matrices. $\begin{bmatrix} 1 \\ 10 \end{bmatrix} \begin{bmatrix} 10 \\ 1 \end{bmatrix}$ We would have that

$$\|H_1(x) - H_1(y)\|_2 \leq \|x - y\|_2$$

or that $10\sqrt{2} \leq 9\sqrt{2}$

So we cannot use this property.

But what can we say about our projection?

Denote $\tilde{x_t} = x_t + A^\top (y - Ax_t)$. With this notation,

$$\|x_{t+1} - \tilde{x_t}\|_2^2 \leq \|x^* - \tilde{x_t}\|_2^2 \implies$$

$$\left\| (x_{t+1} - x^*) + (x^* - \tilde{x_t})_2^2 \right\| \leq \|x^* - \tilde{x_t}\|_2^2 \implies$$

$$\|x_{t+1}\|_2^2 + \|x^* - \tilde{x_t}\|_2^2 + 2\langle x_{t+1} - x^*, x^* - \tilde{x_t} \rangle \, \|x^* - \tilde{x_t}\|_2^2 \implies$$

$$\|x_{t+1} - x^*\|_2^2 \leq 2\langle x_{t+1} - x^*, x^* - \tilde{x_t} \rangle$$

We will define $U := \text{supp}(x_t) \cup \text{supp}(x_{t+1}) \cup \text{supp}(x^*)$ where supp is the support function.

Since

$$\tilde{x_t} = x_t + A^\top (y - Ax_t)$$
$$= x_t + A^\top (Ax^* + w - Ax_t)$$

$$= x_t + A^\top A(x - x^* - x_t) + A^\top w$$

Then combining the two derivations we have

$$\|x_{t+1} - x^*\|_2^2 \le 2\langle x_{t+1} - x^*, x_t + A^\top A(x - x^* - x_t) + A^\top w - x^*\rangle$$

where the RHS equals

$$2\langle x_{t+1} - x^*, I - A_U^\top A_U(x_t - x^*)\rangle + 2\langle x_{t+1} - x^*, A_U^\top w\rangle$$

A key product of the inner product operator to note (This is forbidden knowledge, you may not mention this to anyone else)

$$\langle x, A^\top y\rangle = x^\top A^\top y = (Ax)^\top y = \langle Ax, y\rangle$$

Thus:

- $$\begin{aligned}\langle x_{t+1} - x^*, A_U^\top w\rangle &= \langle A_U(x_{t+1} - x^*), w\rangle \\ &\le \|A_U(x_{t+1} - x^*)\|_2 \cdot \|w\|_2 \\ &\le \sqrt{1-\delta} \cdot \|x_{t+1} - x^*\|_2 \cdot \|w\|_2\end{aligned}$$

- $$\begin{aligned}\langle x_{t+1} - x^*, (I - A_U^\top A_U)(x_t - x^*)\rangle &\le \|x_{t+1} - x^*\|_2 \cdot \left\|(I - A_U^\top A_U)(x_t - x^*)\right\|_2 \\ &\le \|x_t + 1 - x^*\|_2 \cdot \left\|I - A_U^\top A_U\right\|_2 \cdot \|x_t - x^*\|_2\end{aligned}$$

  Where one can show that : $\left\|I - A_U^\top A_U\right\| \le \max\{(1 + \delta - 1, 1 - (1 - \delta)\}$

Combining both the above derivations yields the following

$$\|x_{t+1} - x^*\|_2^2 \le 2\delta \|x_{t+1} - x^*\|_2 \cdot \|x_t - x^*\|_2 + 2\sqrt{1+\delta}\, \|x_{t+1} - x^*\|_2 \cdot \|w\|_2 \implies$$

$$\|x_{t+1} - x^*\|_2 \le 2\delta \cdot \|x_t - x^*\| + 2\sqrt{1+\delta} \cdot \|w\|_2$$

We will assume that $\delta < \frac{1}{2}$, $p = 2\delta < 1$, and furthermore that $\|w\|_2 \le \theta$

Then, combining these assumptions with the previous derivations we have the following

$$\|x_{t+1} - x^*\|_2 \le p \cdot \|x_t - x^*\|_2 + 2\sqrt{1+\delta} \cdot \theta$$

$$\le p^t \cdot \|x_0 - x^*\|_2 + 2\sqrt{1+\delta} \cdot \theta \sum_{i=0}^{t} p^i$$

$$= p^t \cdot \|x_0 - x^*\|_2 + 2\sqrt{1+\delta} \cdot \theta \cdot \frac{1 - p^{t+1}}{1-p} \le p^t \|x_0 - x^*\|_2 + \frac{\sqrt{1+\delta} \cdot \theta}{1-p}$$

Thus, to obtain $\|x_{t+1} - x^*\|_2 \le \epsilon$, we need $O(\log \frac{\|x_0 -^* \|_2}{\epsilon})$ iterations.

Can we pick a new step size based on the RIP property? We want to use the facts that

- in Convex Optimization, $\eta = \frac{1}{L}$ works well (where L is the Lipschitz constant of the objective function)
- We will compute L in our scenario

So by definition of $f(\cdot)$, for x_1, x_2 that are k - sparse

$$\|\nabla f(x_1) - \nabla f(x_2)\|_2 = \left\|-A^\top (y - Ax_1) + A^\top (y - Ax_2)\right\|_2$$

$$= \left\|A^\top A(x_1 - x_2\right\|_2 \leq \max_{S:|S|\leq 2k} \left\|(A^\top A)_s\right\|_2 \cdot \|x_1 - x_2\|_2 \leq (1+\delta)\, \|x_1 - x_2\|_2$$

By definition of RIP. Furthermore, it means we could take $1+\delta$ as L

Thus, *potentially*, $\eta = \frac{1}{L}$ could work, and for $\delta > 0$, we have $\eta < 1$

Matrices that satisfy

$$\mathbb{P}_{A\sim D^{n\times p}}[\|A_x\|_2^2 - \|x\|_2^2] > \epsilon \cdot \|x\|_2^2 \, (\leq 2e^{-\Omega(n})$$

will also satisfy the RIP property with probability $1 - 2e^{-\Omega(n)}$, whenever $n \geq \Omega(\frac{k}{\delta^2} \log \frac{p}{k})$

So this hints at a way to get RIP matrices (which, as we mentioned before, were computationally expensive to verify). Both Gaussian and Bernoulli matrices $A \in \mathbb{R}^{n\times p}$ will satisfy the above property, and therefore make good candidates.

To close the lecture, we will briefly mention adaptive step sizes. In turns out using

$$\eta := \arg\min_{\eta} \|y - A(x_t - \eta \nabla_\delta f(x_t)\|_2$$

generally performs better in practice than the the $\frac{1}{1+\delta}$ that theory might suggest (recall, $1+\delta$ can be taken as a Lipschitz constant for RIP matrices)

Can we relate η to the RIP?

We have the following

$$1 - \delta \leq \frac{\|Ax\|_2^2}{\|x\|_2^2} \leq 1 + \delta$$

and then furthermore,

$$1 + \delta \leq \eta \leq \frac{1}{1-\delta}$$

The theory for the adaptive step size is not well developed beyond this.