Optimization: Algorithms, Complexity & Approximations

Anastasios Kyrillidis *

*Instructor, Computer Science at Rice University

Contributors: Nick Sapoval, Carlos Quintero Pena, Delaram Pirhayatifard, McKell Stauffer, Mohammad Taha Toghani, Senthil Rajasekaran, Gaurav Gupta, Pranay Mittal

Chapter 8

In this chapter we discuss the problem of sparse model selection, i.e., how to perform optimization when the desired/unknown model is constrained by sparsity. While the problem has natural convex translations, we study the iterative hard thresholding (IHT) algorithm and prove results about its performance.

Sparse model selection | Iterative hard thresholding

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

$$\min_{x\in\mathcal{C}\subseteq\mathbb{R}^p} f(x)$$

where $f(\cdot)$ represents a convex objective, and $x \in \mathcal{C}$ represents some constraint. In this lecture, we will be discussing a case in which $\mathcal{C} \subseteq \mathbb{R}^p$ is *not* a convex constraint, namely, when it is the requirement that x be k-sparse. In order to think about this problem, let us introduce its simplest non-trivial version of this problem: the *sparse linear regression* problem, defined as follows:

$$\begin{aligned} \underset{x\in\mathbb{R}^p}{\text{minimize}} \quad & \tfrac{1}{2}\|y - Ax\|_2^2 \\ \text{subject to} \quad & \|x\|_0 \leq k. \end{aligned}$$

Here, $A \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$. Linear regression problems can have a "teacher" generative model assumption where $y = Ax^\star$ + noise, where $x^\star$ is the unknown k-sparse signal we look for. What makes this problem interesting is when we impose the restriction that $n \ll p$; i.e., the problem is ill-posed and classical linear algebra solvers on sets of linear equations do not necessarily recover $x^\star$.

Let us first discuss some procedures that deal with this problem.

- The above problem is non-convex: the inclusion of the ℓ_0-pseudonorm makes the problem non-convex. Classical approaches include convexification: the tightest convex relaxation of the ℓ_0-pseudonorm is that of the ℓ_1-norm (assuming bounded-energy on the initial non-convex set).This leads to the re-definition of the problem as:

  $$\begin{aligned} \underset{x\in\mathbb{R}^p}{\text{minimize}} \quad & \tfrac{1}{2}\|y - Ax\|_2^2 \\ \text{subject to} \quad & \|x\|_1 \leq \lambda. \end{aligned}$$

  There is a long-listed literature on this subject (to be updated). One caveat of this approach is that λ hyperparameter/regularization parameter is not intuitive to be set up correctly (while sparsity k is easier to set up).
- An alternative formulation uses the notion of proximal operators and proximal gradient descent:

  $$\underset{x\in\mathbb{R}^p}{\text{minimize}} \quad \tfrac{1}{2}\|y - Ax\|_2^2 + \rho\|x\|_1$$

  This formulation "moves" the convex ℓ_1-norm constraint into the objective, and uses the following update rule

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

  In words, the ℓ_1-norm in the objective 'biases' the solution towards sparsity (could be seen as an approximation to the exact ℓ_1-norm projection). Similar to the case above, it is not clear how we can select ρ value to achieve good performance.
- Finally, one could keep the non-convexity, and use non-convex projected gradient descent. This leads to

  $$x_{t+1} = \mathcal{H}_k(x_t - \eta\nabla f(x_t)).$$

  This is perhaps somewhat like 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 $H_k(z) = \text{argmin}_{\|x_0\|\leq k}\|x - z\|_2^2$

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).

Note that verifying if a matrix satisfies the RIP is NP-hard. Therefore, let us take for granted we have such a 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,

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

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

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

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

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

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

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:

- $$\langle x_{t+1} - x^*, A_U^\top w\rangle = \langle A_U(x_{t+1} - x^*), w\rangle$$
$$\leq \|A_U(x_{t+1} - x^*)\|_2 \cdot \|w\|_2$$
$$\leq \sqrt{1-\delta} \cdot \|x_{t+1} - x^*\|_2 \cdot \|w\|_2$$
- $$\langle x_{t+1} - x^*, (I - A_U^\top A_U)(x_t - x^*)\rangle$$
$$\leq \|x_{t+1} - x^*\|_2 \cdot \|(I - A_U^\top A_U)(x_t - x^*)\|_2$$
$$\leq \|x_t + 1 - x^*\|_2 \cdot \|I - A_U^\top A_U\|_2 \cdot \|x_t - x^*\|_2$$

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

Combining both the above derivations yields the following

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

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

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

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

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

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

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

Thus, to obtain $\|x_{t+1} - x^*\|_2 \leq \epsilon$, 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

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

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

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.

Adaptive Step Sizes

To close the lecture, we will describe adaptive step sizes. We want to compute η in $x_{t+1} = H_k(x_t - \eta \cdot \nabla f(x_t))$. Some observations:

- x_t is k-sparse
- x_{t+1} is k-sparse
- x_{t+1} has support from x_t, $H_k(-\nabla f(x_t))$ (outside of $\text{supp}(x_t)$), or a combination of both.

- Schematically, it looks like this:

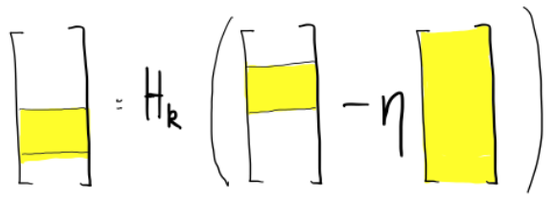

Fig. 1.

In turns out using

$$\eta := \underset{\eta}{\operatorname{argmin}} \|y - A(x_t - \eta \nabla_\delta f(x_t)\|_2$$

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)

Define:

$$s_t = \operatorname{supp}(x_t)$$
$$s_{t+1} = \operatorname{supp}(x_{t+1})$$
$$Q_t = s_t \cup \operatorname{supp}(H_k(\nabla_{s_t c} f(x_t)))$$

Then,

$$H_k(x_t - \eta \nabla f(x_t)) = H_k(x_t - \eta \cdot \nabla_{Q_t} f(x_t))$$

As stated above, we perform line search for η as:

$$\eta = \underset{\eta}{\operatorname{argmin}} \|y - A(x_t - \eta \cdot \nabla_{Q_t} f(x_t))\|_2^2$$

where $\nabla_{Q_t} f(x_t)$ finds the step size that minimizes $f(\cdot)$.

Taking the derivative and setting it equal to zero:

$$\begin{aligned} 0 &= \nabla g(y) \\ &= 2\langle A\nabla_{Q_t} f(x_t), y - Ax_t\rangle + 2\eta \|A\nabla_{Q_t} f(x_t)\|_2^2 \\ \implies \eta &= \frac{-\langle A\nabla_{Q_t} f(x_t), y - Ax_t\rangle}{\|A\nabla_{Q_t} f(x_t)\|_2^2} \\ &= \frac{\|\nabla_{Q_t} f(x_t)\|_2^2}{\|A\nabla_{Q_t} f(x_t)\|_2^2} \end{aligned}$$

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,

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

Proof of Adaptive Step Sizes in IHT

Following the same procedure as in $\eta = 1$, we have

$$\|x_{t+1} - x^\star\|_2 \leq 2\|I - \eta A_u^T A_u\|_2 \cdot \|x_t - x^\star\|_2 + 2\sqrt{1+\delta} \cdot \eta \cdot \|w\|_2$$

By RIP:

$$\begin{aligned} \|I - \eta A_u^T A_u\|_2 &\leq \max\{\eta(1+\delta) - 1, 1 - \eta(1-\delta)\} \\ &\leq \max\{\frac{1+\delta}{1-\delta} - 1, 1 - \frac{1-\delta}{1+\delta}\} \end{aligned}$$

By the property $1 + \delta \leq \eta \leq \frac{1}{1-\delta}$, then

$$\begin{aligned} \|x_{t+1} - x^\star\|_2 &\leq 2\frac{2\delta}{1-\delta} \cdot \|x_t - x^\star\|_2 + 2\frac{2\sqrt{1+\delta}}{1-\delta}\|w\|_2 \\ &= \frac{4\delta}{1-\delta}\|x_t - x^\star\|_2 + 2\frac{2\sqrt{1+\delta}}{1-\delta}\|w\|_2 \end{aligned}$$

Assuming

$$\delta < \frac{1}{5}$$

We get

$$\frac{4\delta}{1-\delta} =: p < 1$$

As shown in the proof of convergence of regular IHT (in which $p < 1$), we get convergence.

Appendix

1. **L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. Siam Review, 60(2):223–311, 2018.**