Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

April 24, 2018

22 Lecture 22: Non-convex constraints part 2: projected gradient descent

Last time we discussed optimization of a convex function f over non-convex sets. A typical example for such a problem is min $\|Ax - y\|_2^2$ subject to $\|x\|_0 \leqslant s$. Another type of non-convex optimization is to minimize a non-convex function over a convex set. A typical example for this kind of non-convex optimization is known as l_0-minimization where min $\|x\|_0$ subject to an affine convex constraint like $Ax = y$. In sum, the l_0-minimization is:

$$\min_{Ax=y} \|x\|_0$$

One option for solving the above non-convex optimization problem is to relax the l_0-objective to the convex l_1-objective as:

$$\min_{Ax=y} \|x\|_1$$

Note that under appropriate assumptions on A, x and y, we observed that ℓ_1-minimization still gives the right answer (assumptions such as RIP and the restricted nullspace property).

In our setting, we have

nonconvex constraint

$\downarrow$

cvx relaxation

$\downarrow$

PGD (Projected Gradient Descent),
or more directly, nonconvex constraint $\rightarrow$ PGD.

We discuss some takeaways from the jupyter notebook. While convex relaxations can be solved efficiently (e.g., using interior point methods), scaling to large problem instances is an issue. Hence it makes sense to consider first order methods such as projected gradient descent (PGD) to speed up computation. We find that it works directly with projections onto the non-convex set. And when we're running PGD, it's natural to ask whether we need the convex relaxation in the first place or can just directly run PGD for the non-convex set.

Consider the test problem of form $y = Ax$ with s-sparse x where A has dimension $n \times d$ and we sample the entries of A as i.i.d. Gaussians so the matrix satisfies the restricted isometry property if we take enough samples n. We can check how many samples are needed for ℓ_1-relaxation to work. We have that an i.i.d. Gaussian matrix requires $O(s \log d/s)$ rows to satisfy RIP, so this corresponds to $n = O(s \log d/s)$ samples. The running time of using just the interior point methods is somewhat slow.

- $d = 100, n = 50$: 10 ms,
- $d = 1000, n = 500$: 5-6 seconds,
- $d = 4000, n = 2000$: 112 seconds.

Therefore, to solve very large instances, we should also consider first-order methods. We may directly run projected gradient descent with the non-convex set of sparse vectors, also known as Iterative Hard Thresholding since the projection step (to find the closest s-sparse vector) corresponds to hard thresholding the vector (keep only the s largest entries and set the rest to 0). It is 1000 times faster as it took 0.0357 seconds.

Now, we discuss the Iterative Hard Thresholding (IHT) which is known as a projected gradient descent (PGD) for sparse vectors.

Iterative Hard Thresholding (IHT) is used with Gradient Descent for the following problem.

Given the setup:

$$y = Ax + e \tag{1}$$

where $y \in \mathbb{R}^M, x \in \mathbb{R}^N, A \in \mathbb{R}^{M \times N}$, e is observation noise, the goal is to estimate x given y and A when $M << N$ and x is approximately K-sparse.
The IHT algorithm uses the iteration

$$x^{n+1} = P_K(x^n + A^\top(y - Ax^n)) \tag{2}$$

where P_K is a hard thresholding operator that keeps the largest K elements of a vector.
Consider the following objective function.

$$f(x) = \frac{1}{2}\|Ax - y\|_2^2 \tag{3}$$

$$\nabla f(x) = A^\top(Ax - y) \tag{4}$$

If it is possible to directly optimize over nonconvex sets, then we don't need convex relaxation.
We want to show that the algorithm IHT outputs a solution. The algorithm is defined as:

$$\begin{aligned}
&\text{IHT}(y, A, t, s) \\
&\qquad x^1 \leftarrow 0 \\
&\qquad \text{for } i = 1, \cdots, t, \\
&\qquad \tilde{x}^{i+1} \leftarrow x^i - A^\top(Ax - y) \\
&\qquad x^{i+1} \leftarrow P_S(\tilde{x}^{i+1}) \\
&\qquad \text{return } \widehat{x} \leftarrow x^{t+1}
\end{aligned}$$

Note that P_s is a projection onto a set of s-sparse vectors. In the rest of the lecture we will see how this projection works and how fast the method is.
We study a nice property of matrices, the Restricted Isometry Property. It is useful since it implies that the difference between two s-sparse vectors cannot be mapped to 0, and the RIP also implies the restricted nullspace property. Here, the Restricted Isometry Property allows optimization over nonconvex sets.

Definition 22.1. RIP: matrix A satisfies the (s, δ)-RIP if for all s-sparse vectors we have:

$$(1 - \delta)\|x\|_2^2 \leqslant \|Ax\|_2^2 \leqslant (1 + \delta)\|x\|_2^2$$

Consequences: For all supports S of size s, we have:

$$\|I - X_S^\top X_S\|_2 \leqslant \delta$$

where X_S is X restricted to S (for detailed description please see the notes for lecture 21).
Consider the following setup:

- $y = Ax^* + e$
- x^* is s-sparse with support S^*
- A has $(3s, \frac{1}{4})$-RIP
- e is arbitrary noise.

Then the following holds,

Theorem 22.2.

$$\|x^{i+1} - x^*\|_2 \leqslant \frac{1}{2}\|x^i - x^*\|_2 + \max_{|S|\leqslant 3s} \|A_S^\top e\|_2$$

Proof. Let S_i be the support of (x^i) and let $S\prime = S^{i+1} \cup S^i \cup S^*$ (so $|S\prime| \leqslant 3s$) . Then:

$$\begin{aligned}
\|x^{i+1} - x^*\|_2 &\leqslant \|x^{i+1} - \tilde{x}^{i+1}_{S\prime}\|_2 + \|\tilde{x}^{i+1}_{S\prime} - x^*\|_2 \qquad \text{(by triangle inequality)} \\
&\leqslant 2\|\tilde{x}^{i+1}_{S\prime} - x^*\|_2 \\
&= 2\|x^i_{S\prime} - A^\top_{S\prime}(Ax^i - y) - x^*\| \\
&= 2\|x^i - A^\top_{S\prime}(A_{S\prime}x^i - A_{S\prime}x^* - e) - x^*\| \\
&= 2\|x^i - x^* - A^\top_{S\prime}A_{S\prime}(x^i - x^*) + A^\top_{S\prime}e\| \\
&\leqslant 2\|I - A^\top_{S\prime}A_{S\prime}(x^i - x^*)\|_2 + 2\|A^\top_{S\prime}e\|_2 \\
&\leqslant 2\delta\|(x^i - x^*)\|_2 + 2\max_{|S|\leqslant 3s} \|A_S^\top e\|_2
\end{aligned}$$

■

The second inequality follows from the first inequality because x^{i+1} is the s-sparse projection of $\tilde{x}^{i+1}$. In particular, this also means that x^{i+1} is the best s-sparse approximation of $\tilde{x}^{i+1}_{S\prime}$ and hence we have:

$$\|x^{i+1} - x^{i+1}_{S\prime}\| \leqslant \|\tilde{x}^{i+1}_{S\prime} - x^*\|_2$$

In the theorem above, we have shown that the error goes down by a factor $1/2$ in every iteration (up to the noise threshold). From there, it is fairly straightforward to get a linear convergence rate as in the following corollary:

Corollary 22.3. *$\|\widehat{x} - x^*\|_2 \leqslant (\frac{1}{2})^t\|x^*\|_2 + 5\|e\|_2$ so $t = \log\frac{\|x^*\|_2}{\epsilon}$ iterations suffice for $\|\widehat{x} - x^*\|_2 \leqslant \epsilon + 5\|e\|_2$.*

We have a linear rate and PSG, but the analysis looks somewhat different (no convexity/smoothness) and we did not need a step size.
Now consider function $f = \frac{1}{2}\|Ax - y\|_2^2$ and hence $\nabla f(x) = A^\top(Ax - y)$.

Definition 22.4. Smoothness:

$$f(x+\Delta) \leqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{L}{2}\|\Delta\|_2^2$$

which is valid for all $\Delta \in \mathbb{R}^d$.

What does the smoothness mean for the above function f?

$$\frac{1}{2}\|A(x+\Delta) - y\|_2^2 \leqslant \|Ax - y\|_2^2 + \Delta^\top A^\top (Ax - y) + \frac{L}{2}\|\delta\|_2^2$$

then:

$$\frac{1}{2}(x+\Delta)^\top A^\top A(x+\Delta) + \frac{1}{2}y^\top y \leqslant \frac{1}{2}x^\top A^\top Ax - y^\top Ax + \frac{1}{2}y^\top y + \Delta^\top A^\top Ax - \Delta^\top A^\top y + \frac{L}{2}\|\Delta\|_2^2$$

hence:

$$\frac{1}{2}\Delta^\top A^\top A\Delta \leqslant \frac{L}{2}\|\Delta\|_2^2,$$

and as a result:

$$\|A\Delta\|_2^2 \leqslant L\|\Delta\|_2^2$$

Taking $L = 1 + \delta$ gives that the above equals

$$L/\ell \approx \frac{1+\delta}{1-\delta} \approx 1.$$

Note that the above inequality relates being smooth with following the RIP, just replace the condition "for any $\triangle$" (for smoothness) by "for any s-sparse $\triangle$" (for RIP). We have similar results for strong convexity.

Definition 22.5. Strong Convexity:

$$f(x+\Delta) \geqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{l}{2}\|\Delta\|_2^2$$

which is valid for all $\Delta \in \mathbb{R}^d$.

We can easily conclude that strong convexity is equivalent to $\|A\Delta\|_2^2 \geqslant l\|\Delta\|_2^2$.

Definition 22.6. Restricted Strong Convexity (RSC):

$$f(x+\Delta) \geqslant f(x) + \langle \nabla f(x), \Delta \rangle + \frac{l}{2}\|\Delta\|_2^2$$

which is valid for all s-sparse Δ.

We can say that:

$$\text{RIP} = \text{RSC} + \text{RSM},$$

with very good ($L \approx 1 + \delta$) condition number L/ℓ (hence constant step size $\approx 1/L$). There is a lot of work on weakening this assumption and further constraints sets. Convex relaxation involves mostly optimal dependence on condition number. PGD can be made to work for arbitrary condition number but with a worse statistical rate. Can we match convex relaxation with non-convex PGD? Yes!
Given the sparsity condition, it is possible to do hard thresholding in $O(d)$ time. Given the low-rank condition, one can compute the SVD and find the largest singular values and for a $d_1 \times d_2$ matrix, in $O(d_1 d_2 \min(d_1, d_2))$ time.

Definition 22.7. *Group Sparsity:* We are given a family of groups $G_i \subseteq [d]$, support $\text{supp}(x^*) = \cup_{j \in J} G_j$ for some $|J| \leqslant g$. This is NP-hard via set cover.

Definition 22.8. *Graph Sparsity:* Given graph G defined on $[d]$ (nodes are indices in $\{1, \cdots, d\}$), $\text{supp}(x^*)$ is a connected subgraph in G. Here, projection on set is NP-hard (Steiner trees).

Definition 22.9. *Approximate Projection*: (Tail approximation) Given input $\tilde{x} \in \mathbb{R}^d$, find $x \in C$ (C is the constraint set) such that

$$\|x - \tilde{x}\|_2 \leqslant \alpha \min_{x' \in C} \|x' - \tilde{x}\|_2.$$

For x^* that is 1-sparse and the problem

$$y = Ax^*,$$

$A = \pm\frac{1}{\sqrt{n}}$ gives $n = O(\log d)$ and then we have RIP, and PGD should work.
In the first iteration of PGD,

$$x' \leftarrow 0$$
$$x^2 \leftarrow \widehat{P}_S(A^\top y)$$

where $A^\top y = A^\top A x^* = b$.
Consider the following setup.

$$b_1 = 1$$
$$\mathbb{E}[b_i^2] = \frac{1}{n}$$
$$\mathbb{E}[\|b\|_2^2] = 1 + \frac{d-1}{n}.$$

The approximate error of the best projection is on the order $\frac{d-1}{n}$.

Definition 22.10. Head Projection:
Given $\tilde{x} \in \mathbb{R}^d$, find a support S such that

$$\|\tilde{x}_S\|_2 \geqslant \beta \max_{S' \in \text{Supp}(C)} \|\tilde{x}_{S'}\|_2. \tag{5}$$

When this is combined with approximate projection, PGD still works.

References