Indian Institute of Technology Kanpur
CS774 Optimization Techniques

Scribe: Ninad Khargonkar
Instructor: Purushottam Kar
Date: October 27, 2016

LECTURE 21

Non-convex optimization

Introduction

In the previous lecture we looked at the problem of sparse recovery and the restricted isometric property for a matrix.
In this lecture we will look at some convex relaxations like the Morozov, Tikhonov and Ivanov formulations for the problem. In particular we will look at the Iterative Hard Thresholding (IHT) algorithm and its analysis under the assumption of RIP. We will also briefly look at some of its applications.

Recap

Definition 21.1. A matrix A is said to satisfy the (ϵ, s)-RIP condition if $\forall x \in B_0(0,s)$;

$$(1-\epsilon)||x||_2{}^2 \leq ||Ax||_2{}^2 \leq (1+\epsilon)||x||_2{}^2$$

It is assumed that we have $b = Ax^*$ such that the sparse vector $x^* \in B_0(0,s)$. We wish to find $\hat{x}$ where:

$$\hat{x} = \text{arg min } ||x||_0$$
$$\text{s.t: } Ax = b$$

Claim 21.1. If A satisfies the (ϵ, $2s$)-RIP condition for any $\epsilon > 0$, then $\hat{x} = x^*$.

Proof. $A\hat{x} = b = Ax^*$
$\implies A(\hat{x} - x^*) = 0$
$\implies (\hat{x} - x^*) \in B_0(0, 2s)$
Now use the (ϵ, $2s$)-RIP condition. □

Convex Relaxations

As the l_0 norm is unbounded, we will attempt to 'convexify' it. The tightest 'convexification' will be given by l_1 norm. All other l_p norms with $p > 1$ are still valid 'convexifiers'. See Fig. 1 for reference.

Note: Assume all norms mentioned in this lecture are l_2 norms unless specified otherwise.

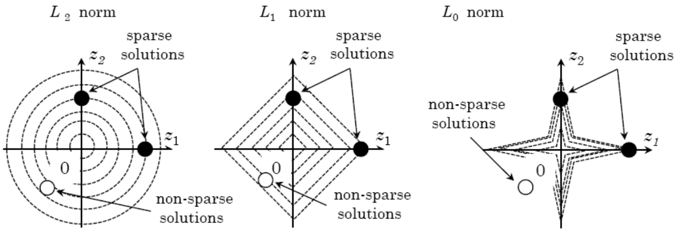

Figure 1: Norms

Morozov formulation

$$\hat{x} = \text{arg min } ||x||_1$$
$$\text{s.t: } Ax = b$$

Claim 21.2. If A satisfies the (ϵ, $2s$)-RIP condition for $\epsilon \leq \frac{1}{\sqrt{2}+1}$, then $\hat{x} = x^*$.

Lasso formulation

$$\hat{x} = \text{arg min}_x \; ||Ax - b||_2{}^2 \; +\lambda ||x||_1$$

Lasso stands for 'least absolute shrinkage and selection operator'. There is an equivalent formulation which looks like:

$$\hat{x} = \text{arg min}_x \; ||Ax - b||_2{}^2$$
$$\text{s.t: } ||x||_1 = \leq R$$

Ivanov formulation

$$\hat{x} = \text{arg min}_x \; ||Ax - b||_2{}^2$$
$$\text{s.t: } ||x||_0 = \leq s$$

This is an NP-hard problem and we will look at the IHT (Iterative Hard Thresholding) algorithm which attempts to find a solution to this problem. Specifically we attempt to find $x^* \in \mathbb{R}^n$ such that $||x^*||_0 \leq s << n$ and $b = Ax^*$ where $b \in \mathbb{R}^m$.

Algorithm 1: Iterative Hard Thresholding

```
Input: $b \in \mathbb{R}^m, A \in \mathbb{R}^{m*n}$
$\mathbf{x}^0 \in B_0(0, s) := B_0(s)$
for $t = 0, 1, \ldots, T$ do
    $\mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - A^T(A\mathbf{x}^t - b)$
    $\mathbf{x}^{t+1} \leftarrow \Pi_{B_0(s)}(\mathbf{z}^{t+1})$
end for
```

This algorithm ensures that we get sparse vectors in every iteration which may not be possible for other methods as they give dense vectors. Here, the projection step essentially involves solving the non-trivial problem.

$$\hat{z} := \Pi_{B_0(s)}(z) = \text{arg min}_{y \in B_0(s)} \; ||z - y||_2{}^2$$

$\hat{z}$ can found out by the following steps:

1. Sort elements of z according to their absolute values i.e $|z_i|$'s
2. Let σ be the permutation such that $|z_{\sigma_{(1)}}| \geq |z_{\sigma_{(2)}}| \geq \dots$
3. $\hat{z_i} = z_i$, if $\sigma^{-1}(i) \leq s$ or equal to 0 otherwise.

Exercise: Prove that the above steps get us the desired solution. (Hint: show $\forall i \in \text{supp}(\hat{z})$, show that $\hat{z_i} = z_i$)

Exercise: Property of 'Contractivity' of the projection step. If C is a convex set then prove that:
$||\Pi_C(x) - \Pi_C(y)||_2 \leq ||x - y||_2$ (This property is violated for $B_0(s)$)

Claim 21.3. If A satisfies the (ϵ, $3s$)-RIP condition for $\epsilon \leq \frac{1}{2}$, then $||x^t - x^*||_2 \leq e^{-\Omega(t)}$ and $||x||_0 \leq s$ (it converges in linear time)

Proof: By the zeroth-order property we have by construction:

$$||x^{t+1} - z^{t+1}||_2{}^2 \leq ||x^* - z^{t+1}||_2{}^2$$

Define $I_t := \text{supp}(x^t) \cup \text{supp}(x^{t+1}) \cup \text{supp}(x^*)$
For simplicity, in the following steps I_t is denoted by I.
Define V_I by $(V_I)_i = V_i$ if $i \in I$ and is zero otherwise.

$$\begin{gathered}
||x_I^{t+1} - z_I^{t+1}||_2 \leq ||x_I^* - z_I^{t+1}||_2 \\
z_I^{t+1} = A_I^T(Ax^t - b) \\
||x_I^{t+1} - z_I^{t+1}||_2 = ||x_I^{t+1} - x_I^t + A_I^T(Ax^t - b)|| \\
\geq ||x_I^{t+1} - x_I^*|| - ||x_I^* - x_I^t + A_I^T(Ax^t - b)|| \\
= ||x_I^{t+1} - x_I^*|| - ||x_I^* - z_I^{t+1}|| \\
||x_I^{t+1} - x_I^*|| \leq 2 * ||x_I^* - z_I^{t+1}|| \\
||x^{t+1} - x^*|| \leq 2\epsilon||x^t - x^*||
\end{gathered}$$

The justification for the last step is shown a little bit later. Now consider $||x^* - z^{t+1}||_2{}^2$. It will be equal to $||x^* - x^t - A_I^T(Ax^t - b)||^2$. Using structure of x^t and x^* w.r.t I, we get the expression to be: $||x^* - x^t - A_I^T A_I(x^t - x^*)||^2$ which further reduces to: $||(I - A_I^T A_I)(x^t - x^*)||^2$

Remark 21.1. Verify that $||Ax||_2^2 \geq (1 - \epsilon)$ $||x||_2^2$ is equivalent to $\lambda_{min}(A_s^T A_s) \geq$ (1-ϵ) where s = supp(x).
Also prove that (ϵ,$2s$)-RIP $\iff$ $\lambda_{min}(A_s^T A_s) \geq$ (1-ϵ) $\forall$ $|s| < 2\epsilon$.

Using the above statements, we get: (assume that $m \geq 2s$)

$$||(I - A_I^T A_I)(x^t - x^*)||^2 \leq \epsilon||(x^t - x^*)||^2$$

$$\implies ||(x^t - x^*)|| \leq Ce^{-2\epsilon t}$$

The constant C is proportional to $||(x^0 - x^*)||$

Some applications

1. Multi-label learning / Tagging

There are label vectors $y \in \mathbb{R}^L$ where L is of the order of 10^6. But the label vectors are sparse, i.e $||y||_0 \leq s$ where $s \sim 10$.
The training set consists of the documents $x_n \in \mathbb{R}^d$ along with their labels y_n.

For the purpose of learning, we compress y_i's to z_i's $\in \mathbb{R}^k$ where $k << L$. We perform the training with $(x_n,\ z_n)$ and then for testing perform sparse recovery on the output from the learned model to obtain the entire vector of labels.

2. Gene-expression Analysis

The gene expressions are given by $x_i \in \mathbb{R}^n$ and the extent of the disease by $y_i \in \mathbb{R}$. Prediction y_t is approximately given by $< x_t,\ v^* > +\eta$; $t = 1, 2, \ldots m$ where $m << n$.
Here, $v^* \in \mathbb{R}^n$ and $||v^*||_0 \leq s << n$

3. Robust Learning

$b_i = < a_i, x^* > +e$ (Here e represents an adversary giving improper feedback)
This is equivalent to $b = Ax+e$ with $e, b \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m*n}$ where $||e||_0 \leq k << m$

We take the problem of $\min_{x,e} ||b - e - Ax||$ with the constraint $||e||_0 \leq k$
Note that for a given e, it simply reduces to solving a linear regression problem for x.

References

Wattanit H. lp Norms. *https://rorasa.wordpress.com/2012/05/13/l0-norm-l1-norm-l2-norm-l-infinity-norm/*.

Jeff M Phillips. Lasso Regularized Regression.

Wikipedia. Lasso. *https://en.wikipedia.org/wiki/Lasso_(statistics)*.