**Indian Institute of Technology Kanpur**
**CS777 Topics in Learning Theory**

*Scribe:* Harshvardhan
*Instructor:* Purushottam Kar
*Date:* March 14, 2018

**LECTURE**
# 17

## Sparse Recovery-I

## 1 Introduction

In the previous three lectures we covered some standard techniques used in convex optimization like Projected Gradient Descent and Frank-Wolfe and their convergence analysis when the objective function and the constraint set satisfied certain 'nice' properties. The previous lecture was devoted to analyzing convergence of Projected Gradient Descent in convex settings and it ended with an introduction of Non-Convex Optimization and Sparse Recovery.

Non-Convex optimization in itself is NP- Hard but if the function and constraint set satisfy certain properties it can be performed and even shown to have a good convergence rate. The problem of Sparse recovery is one such example. It is a popular technique used in Statistics and Compressive Sensing for obtaining a sparse representation of the data.

In this lecture, we will study the structure of the Sparse Recovery problem and analyse its solution using the PGD algorithm on non-convex sets.

## 2 Recap

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

The optimization problem in Non-Convex Optimization has the above form where optimization objective $f(\mathbf{x})$ and the constraint set $\mathcal{C}$ may be non-convex.

The problem of Sparse Recovery can be expressed in terms of finding a sparse linear model $x \in \mathcal{B}_0(k)$ for a positive semi-definite design or sketch matrix $A \in \mathbb{R}^{n\times d}$ (which may be hand-crafted or the set of observations) and a vector $b \in \mathbb{R}^d$. Here $\mathcal{C} = \mathcal{B}_0(k) = \left\{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_0 \leq k\right\}$, which is the set of $k$-sparse vectors. The objective functio takes the following form for sparse recovery -

$$f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2$$

### 2.1 $\alpha - RSC$ and $\beta - RSS$

We repeat the definitions of $\alpha$-Restricted Set Convexity and $\beta$-Restricted Set Smoothness which are essential for analysis of functions over possibly non-convex sets. Note that these restricted properties are defined with respect to an objective function $f$ and a set $\mathcal{C}$.

A function $f$ is said to be $\alpha$-RSC and $\beta$-RSS over a set $\mathcal{C}$ if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$,

$$f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \in \left(\frac{\alpha}{2}, \frac{\beta}{2}\right) \|\mathbf{x} - \mathbf{y}\|_2^2$$

.

As shown in previous lecture , for the sparse recovery settings, the objective function can be represented as a sum of its linear approximation and a quadratic term.

$$f(\mathbf{x}) = f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{y})^\top A^\top A (\mathbf{x} - \mathbf{y})$$

From the above expression it is evident that for $f$, $\alpha = \lambda_{min}\left(A^\top A\right)$ and $\beta = \lambda_{max}\left(A^\top A\right)$. The convexity condition can be easily violated if the smallest eigenvalue is 0. For an ill-posed problem ($n < d$), $A$ can never be full rank and thus $A^\top A$ has minimum eigenvalue 0.

**Remark 17.1.** The values of $\alpha$ and $\beta$ can be different from the minimum and maximum eigenvalues of $A^\top A$ if the eigenvectors corresponding to these vectors don't lie in the constraint set $\mathcal{C}$.

# 3 Ensuring $\alpha - RSC$ and $\beta - RSS$

We first state the exact optimization problem for sparse recovery.

$$\min_{\mathbf{x} \in \mathcal{B}_0(k)} \frac{1}{n} \|A\mathbf{x} - \mathbf{b}\|_2^2$$

Here, $k << d$ and the factor of $\frac{1}{n}$ ensures that eigenvalues don't scale with $n$.

Sparse Sets offer very rich structure with 'nice' properties in a lot of different scenarios. As seen in earlier lectures , k-sparse sets had small Rademacher complexity and regression over these sets required only $\mathcal{O}(k \log d)$ samples . Similarly, $\alpha$-RSC and $\beta$-RSS can be guaranteed with high probability for the sparse recovery objective over the set of k-sparse vectors using sampling techniques with low sample complexity.

**Theorem 17.1.** If $A \in \mathbb{R}^{n \times d}$, such that $n = \Omega\left(\frac{1}{(1-\alpha)^2} k \log d\right)$ and each element $A_{i,j}$ is sampled i.i.d from a $\mathcal{N}(0,1)$ (or $\{-1,1\}$ Rademacher distribution ), then $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2$ is $\alpha$-RSC over the set $\mathcal{B}_0(k)$ with high probability $(1 - \exp(-\Omega(d)))$

A similar result exists for $\beta$-RSS.

**Theorem 17.2.** If $A \in \mathbb{R}^{n \times d}$, such that $n = \Omega\left(\frac{1}{(1-\beta)^2} k \log d\right)$ and each element $A_{i,j}$ is sampled i.i.d from a $\mathcal{N}(0,1)$ (or $\{-1,1\}$ Rademacher distribution ), then $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2$ is $\beta$-RSS over the set $\mathcal{B}_0(k)$ with high probability $(1 - \exp(-\Omega(d)))$

Here, $k \log d << n$. Thus, with sufficient number of samples we can make the restricted minimum eigenvalue and in turn $\alpha$ as large as we want.

## 3.1 Implications of $\alpha$-RSC

$\alpha$-RSC in the sparse settings leads to uniqueness in the algorithm output for sparse recovery,i.e, when $f$ is $\alpha$-RSC over $\mathcal{B}_0(k)$ then $\forall \mathbf{x}, \mathbf{y} \in \mathcal{B}_0(k)$

$$(\mathbf{x} - \mathbf{y})^\top A^\top A (\mathbf{x} - \mathbf{y}) \geq \alpha \|\mathbf{x} - \mathbf{y}\|_2^2$$
$$\implies \|A(x - y)\|_2^2 \geq \alpha \|\mathbf{x} - \mathbf{y}\|_2^2$$

. The contrapositive of the above statement says that for some distinct $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathcal{B}_0(k)$

$$\|A(\hat{\mathbf{x}} - \hat{\mathbf{y}})\|_2^2 = 0$$
$$\implies A\hat{\mathbf{x}} = A\hat{\mathbf{y}}$$

This leads to an ill-posed problem, where the algorithm cannot distinguish between two distinct vectors and thus cannot give a unique solution . $\alpha$-RSC ensures that this never happens.

**Exercise 17.1.** Prove that the optima $\mathbf{x}^* \in \mathcal{B}_0(k)$ for an $\alpha$-RSC $f$ is unique.

### 3.2 Deterministic Construction of $A$

The design/sketch matrix can be obtained by 2 very different methods. In the settings of Sparse Regression, each row of $A$ corresponds to the input feature vector for a datapoint. However, for compressive sensing applications , each row of $A$ corresponds to a 'carefully' handcrafted functional. The degree of control over $A$ in the later case is very high which motivates us for constructing these design matrices deterministically. It turns out that the best algorithms for deterministic construction of $A$ are quadratic or nearly quadratic in the sparsity level $k$ as opposed to sampling based techniques which are linear in $k$. The interested reader can refer to Jain and Kar (2017) for a discussion on this.

## 4 PGD for Sparse Recovery

### 4.1 Algorithm

Algorithm 1: Projected Gradient Descent

**Input:** Data $A, \mathbf{b}$, step lengths $\eta_t$, constraint set $\mathcal{C}$
**Output:** $\mathbf{x} \in \mathcal{C}$
1: $\mathbf{x}^1 \leftarrow \mathbf{0}$
2: **for** $t = 1, 2, \ldots, T-1$ **do**
3: $\quad \mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - \eta_t \cdot A^\top (A\mathbf{x}^t - \mathbf{b})$
4: $\quad \mathbf{x}^{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{z}^{t+1})$ `//Projection onto Non-convex set`
5: **end for**
6: **return** $\mathbf{x}^\top$

The algorithm we will be using to solve this optimization problem is exactly similar to PGD. The only difference here is that projection step projects onto a non-convex set $\mathcal{C}$. Projection onto a non-convex set is NP-Hard but simplifies for the cases of structured non-convex sets like sparse vectors ($\mathcal{B}_0(k)$) or low rank matrices($\mathcal{B}_{rank}(k)$).

**Exercise 17.2.** Show that projection onto the set of sparse vectors $\mathcal{B}_0(k)$ is Hard Thresholding.

**Exercise 17.3.** Show that projection onto the set of low rank matrices $\mathcal{B}_{rank}(k)$ is SVD.

### 4.2 Analysis

**The analysis scheme involves the following steps**

1. Apply $\alpha$-RSC between $\mathbf{x}^t$ and $\mathbf{x}^*$
2. Apply $\beta$-RSS between $\mathbf{x}^{t+1}$ and $\mathbf{x}^t$
3. Apply Projection Property 0

**Remark 17.2.** Projection Property 0 ($\|\hat{\mathbf{x}} - \mathbf{z}\|_2 \leq \|\mathbf{x} - \mathbf{z}\|_2 \, \forall \mathbf{x} \in \mathcal{C}$) is always satisfied by the definition of projection.

**Remark 17.3.** Projection Properties I and II may be violated for a non-convex set $\mathcal{C}$.

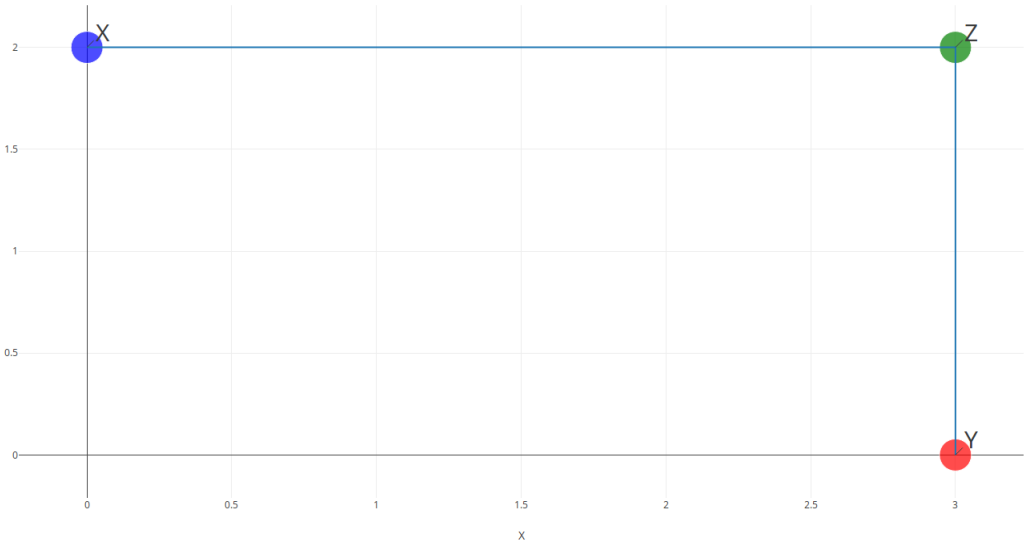


Figure 1: Violation of Projection Property II for 1-sparse vectors

**Example 17.1.** In $\mathbb{R}^2$, $\mathcal{B}_0(1)$ is the union of the 2 axes. Let $\mathbf{z} \in \mathbb{R}^2$ and $\mathbf{y}$ be its projection on $\mathcal{B}_0(1)$. From the figure, $\|\mathbf{x}-\mathbf{y}\| > \|\mathbf{z}-\mathbf{x}\|$ but $\mathbf{x} \in \mathcal{B}_0(1)$

**Assumption 17.1.** $\nabla f(\mathbf{x}^*) = 0$

This assumption is essential for easy analysis for non-convex sets and the result can be proved without this assumption. Moreover this can be ensured for the noiseless setting ($\mathbf{b} = A\mathbf{x}^*$) when global minima lies in the set $\mathcal{C}$. This assumption is required to obtain $\langle \nabla f(\mathbf{x}^*), \hat{\mathbf{x}} - \mathbf{x}^* \rangle = 0$ $\forall \hat{x} \in \mathcal{C}$

**Claim 17.3.** If $\mathcal{C}$ is a convex set, then $\forall \hat{x} \in \mathcal{C}$ $\langle \nabla f(\mathbf{x}^*), \hat{\mathbf{x}} - \mathbf{x}^* \rangle \geq 0$

While analysing PGD, we did not combine the $\alpha$-RSS and $\beta$-RSC conditions together as done in previous lectures(shown below) since the other Projection Properties are invalid. Inability to apply Projection Property I gets us stuck when following this approach.

$$\Phi_{t+1} \leq \left\langle \nabla f\left(\mathbf{x}^t\right), \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{\beta}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|_2^2 - \frac{\alpha}{2} D_t^2$$

However, such a convergence analysis might also be successful but it might not be optimal due to very lossy intermediate steps. We separately deal with the $\alpha$-RSC and the $\beta$-RSS inequalities and combine them only when they have suitable forms.

By $\beta$-RSS between $\mathbf{x}^{t+1}$ and $\mathbf{x}^t$ we get

$$f\left(\mathbf{x}^{t+1}\right) - f\left(\mathbf{x}^t\right) \leq \frac{1}{\eta}\left\langle \mathbf{x}^t - \mathbf{z}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\beta}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|_2^2$$

Using the identity $2ab = (a+b)^2 - a^2 - b^2$

$$\leq \frac{1}{2\eta}\left(\left\|\mathbf{x}^{t+1} - \mathbf{z}^{t+1}\right\|_2^2 - \left\|\mathbf{x}^t - \mathbf{z}^{t+1}\right\|_2^2 - \left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|_2^2\right) + \frac{\beta}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|_2^2$$

Step length is constant and equal to $\frac{1}{\beta}$ for sake of simplicity. Analysis also holds for $\eta \leq \frac{1}{\beta}$

$$= \frac{\beta}{2}\left(\left\|\mathbf{x}^{t+1} - \mathbf{z}^{t+1}\right\|_2^2 - \left\|\mathbf{x}^{t} - \mathbf{z}^{t+1}\right\|_2^2\right)$$

Applying Projection Property 0 with $\hat{x} = x^{t+1}$ and $\mathbf{x} = \mathbf{x}^*$

$$\leq \frac{\beta}{2}\left(\left\|\mathbf{x}^{*} - \mathbf{z}^{t+1}\right\|_2^2 - \left\|\mathbf{x}^{t} - \mathbf{z}^{t+1}\right\|_2^2\right)$$
$$= \frac{\beta}{2}\left(\left\|\mathbf{x}^{*} - \mathbf{x}^t + \mathbf{x}^t - \mathbf{z}^{t+1}\right\|_2^2 - \left\|\mathbf{x}^{t} - \mathbf{z}^{t+1}\right\|_2^2\right)$$

Using $(a+b)^2 = a^2 + b^2 + 2ab$

$$\leq \frac{\beta}{2}\left(D_t^2 + 2\left\langle \mathbf{x}^* - \mathbf{x}^t, \mathbf{x}^t - \mathbf{z}^{t+1}\right\rangle\right)$$
$$= \frac{\beta}{2}D_t^2 + \left\langle \mathbf{x}^* - \mathbf{x}^t, \nabla f\left(\mathbf{x}^t\right)\right\rangle$$

Applying $\alpha$-RSC between $\mathbf{x}^*$ and $\mathbf{x}^t$

$$f\left(\mathbf{x}^*\right) - f\left(\mathbf{x}^t\right) \geq \left\langle \nabla f\left(\mathbf{x}^t\right), \mathbf{x}^* - \mathbf{x}^t\right\rangle + \frac{\alpha}{2}D_t^2$$
$$-\frac{\alpha}{2}D_t^2 - \Phi_t \geq \left\langle \nabla f\left(\mathbf{x}^t\right), \mathbf{x}^* - \mathbf{x}^t\right\rangle$$
$$\implies f\left(x^{t+1}\right) - f\left(x^t\right) \leq \frac{(\beta-\alpha)}{2}D_t^2 - \Phi_t$$
$$\implies \Phi_{t+1} - \Phi_t \leq \frac{(\beta-\alpha)}{2}D_t^2 - \Phi_t$$

To express $D_t$ in terms of $\Phi_t$, we again use $\alpha$-RSC between $\mathbf{x}^t$ and $\mathbf{x}^*$

$$f\left(\mathbf{x}^t\right) - f\left(\mathbf{x}^*\right) \geq \left\langle \nabla f\left(\mathbf{x}^*\right), \mathbf{x}^* - \mathbf{x}^t\right\rangle + \frac{\alpha}{2}D_t^2$$

Here, we use the assumption that $\nabla f\left(\mathbf{x}^*\right) = 0$

$$\implies \frac{2\Phi_t}{\alpha} \leq D_t^2$$
$$\implies \Phi_{t+1} - \Phi_t \leq \left(\frac{\beta}{\alpha} - 1\right)\Phi_t - \Phi_t$$
$$\implies \Phi_{t+1} \leq \Phi_t\left(\frac{\beta}{\alpha} - 1\right)$$
$$\implies \Phi_{t+1} \leq \Phi_t\left(\kappa - 1\right)$$

Here $\kappa = \frac{\beta}{\alpha}$ is the condition number. By definition, $\kappa \geq 1$. We can guarantee exponential decay if $\kappa - 1 \in [0, 1)$. If $\kappa < 2$,

$$\Phi_T \leq \Phi_0 \exp\left(-T\left(2 - \kappa\right)\right)$$

Thus, the value of condition number determines our rate of convergence. Note that here the condition number is defined using the parameters of Restricted Convexity and Restricted Smoothness. Moreover, based on the value of $\kappa$, we can guarantee decrease in the potential function by a constant fraction. For exponential decay we need a small condition number.Thus we need a large value of $\alpha$ which translates to more number of samples. For ill-posed problems with high condition numbers under more assumptions convergence can be guaranteed although the analysis becomes much more complicated.

The subsequent lectures will deal with other methods of solving the sparse recovery problem.

## References

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Found. Trends Mach. Learn.*, 10(3-4):142–336, December 2017. ISSN 1935-8237. doi: 10.1561/2200000058. URL https://doi.org/10.1561/2200000058.