Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Kanishk Gandhi
Instructor: Purushottam Kar
Date: March 12, 2018

LECTURE 16

Optimization Techniques - III

1 Introduction

In the previous two lectures, we were introduced with the basic concepts, algorithmic tools, and analysis techniques used in the design and analysis of optimization algorithms. We looked at properties of projections onto convex sets and then introduced the two broad categories of optimizations methods, namely Projected First Order Methods and Interior Point First Order Methods.

In the last lecture we analyzed the interior point methods, particularly the Frank-Wolfe method and then introduced analysis for Projected Gradient Descent for various constraints on functions and the set that they were defined over. With two cases already analyzed, we now move towards the analysis for other constraints over the objective and the set $\mathcal{C}$.

This lecture while trying to analyze and provide concrete bounds, also seeks to help us understand the intuitions behind the effectiveness of gradient descent style methods in optimizing non-convex objectives. In the analysis we see the common trade-off between generalization of the bound and the tightness that it provides. Towards the end of this lecture, we introduce non-convex optimization which is an NP-Hard problem. We realize that although being an NP-Hard problem, we get promising results.

2 Recap

2.1 Projected Gradient Descent Algorithm

We first look at the analysis for a few convex optimization cases for PGD and then introduce the non-convex cases, the analysis for which is also somewhat similar. Let X be defined over a set $\mathcal{C}$ and $f(X)$ be the objective function.

$$min_{X \in \mathcal{C}} f(X) \triangleq f^*$$

$$x^* = \arg\min_{x \in \mathcal{C}} f(X)$$

The PGD algorithm is specified in 1

2.2 Analysis for two cases

CASE 1: f is convex with bounded gradients

We found in the previous class that the potential function $\sum \Phi_t$ is bounded by $\mathcal{O}\left(\sqrt{T}\right)$

Algorithm 1: Projected Gradient Descent

```
Output: A model $\mathbf{x} \in \mathcal{C}$
$\mathbf{x} \leftarrow \mathcal{C}$
for $t = 1, 2, \ldots, T-1$ do
    $\mathbf{z}^{t+1} \leftarrow \mathbf{x}^t - \eta_t \nabla F(\mathbf{x}^t)$        //Or $g^t \in \delta f(\mathbf{x}^t)$(subgradient)
    $\mathbf{x}^{t+1} \leftarrow \Pi_{\mathcal{C}}(\mathbf{z}^{t+1})$        //This is a projection
    $\Phi_t = f(\mathbf{x}^t) - f(\mathbf{x}^*)$        //Potential Function
    $\mathcal{D}_t = \left\| \mathbf{x}^t - \mathbf{x}^* \right\|$
end for
return $\mathbf{x}^T$
```

CASE 2: f is α-sc with bounded gradients
We found that in this case,

$$\sum \Phi_t \leq \mathcal{O}\left(logT\right)$$

$$\sum t\Phi_t \leq \mathcal{O}\left(T\right)$$

Here, a higher weight is assigned to samples of the later iterations, which intuitively makes sense as the later iterations are closer to the optima. It should be noted that this makes sense only in the case of strong convexity. In this lecture, we see how the analysis fares for strongly smooth objectives.

2.3 Strong Convexity & Strong Smoothness

Strong Convexity and Strong Smoothness in some sense give us an idea of how much the first order approximation of the function deviates from the actual function. It is upper bounded by the smoothness term (function should not vary much) while lower bounded by the strongly convex (function should vary by at least some value). If a function f is $\alpha - sc$ and $\beta - ss$,

$$\frac{\alpha}{2} \left\| X - Y \right\|_2^2 \leq f(X) - f(Y) - \langle \nabla f(Y), X - Y \rangle \leq \frac{\beta}{2} \left\| X - Y \right\|_2^2$$

It is important to note that convexity properties are usually global and bounds involve $\mathbf{x}^t, \mathbf{x}^*$, while smoothness is based on local properties, the bounds in this case involve $\mathbf{x}^t, \mathbf{x}^{t+1}$.

3 Analysis of PGD

3.1 f is β Strongly Smooth

From the definition of smoothness, we can say that,

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

It should be noted here that analysis of Φ_t here with $\mathbf{x}^t$ and $\mathbf{x}^*$ does not give us much. In general, non-convex optimization is np-hard. We get around this by ensuring in our analysis that $f(\mathbf{x}^t)$ keeps decreasing (Maybe over average).

The set $\mathcal{C} \in \mathbf{R}^d$ has no constraints in the following analysis.

$$\begin{aligned} f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) + \frac{1}{\eta_t}\left\langle \mathbf{x}^t - \mathbf{x}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\beta}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|_2^2 \\ &\leq f(\mathbf{x}^t) - (\frac{1}{\eta_t} - \frac{\beta}{2})\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|_2^2 \end{aligned} \tag{1}$$

If $(\frac{1}{\eta_t} - \frac{\beta}{2})$ is positive, then surely $f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t)$. This implies,

$$\eta_t \leq \frac{2}{\beta}$$

Which means that the value of step length is less than the curvature of the function. For convenience of analysis, we assume that we set

$$\eta_t \leq \frac{1}{\beta}$$

This gives us,

$$\begin{aligned} f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) - \frac{\beta}{2}\left\|\mathbf{x}^{t+1} - \mathbf{x}^t\right\|_2^2 \\ &\leq f(\mathbf{x}^t) - \frac{\beta}{2}\eta^2\left\|\nabla f(x^t)\right\|_2^2 \end{aligned} \tag{2}$$

Thus, as long as gradient is large, $f(\mathbf{x}^t)$ makes huge progress to reach towards the optimum. The gradient cannot remain to be large, else $f(\mathbf{x})$ will got to negative infinity. The gradient after while has to go to zero, reaching a stationary point. So, based on the above analysis we can say that PGD with a sufficiently small η_t reaches stationary points even if the function is non-convex.

For any ϵ, so long as

$$\left\|\nabla f(x^t)\right\|^2 \geq \epsilon$$

We can say that,

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) - \frac{\beta}{2}\eta^2\epsilon$$

If the function is lower bounded,

$$-B \leq f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) - \frac{\beta}{2}\eta^2\epsilon$$

Indicating that,

$$T \leq 2\frac{f(\mathbf{x}_o) + B}{\eta^2\epsilon}$$

Remark 16.1. $\exists t \leq t_o = \mathcal{O}\left(\frac{1}{\epsilon}\right)$ s.t. $\left\|\nabla f(x^t)\right\|^2 \leq \epsilon$ for at least one point.

Exercise 16.1. Try to prove the above result for a constrained case.

Recent papers have shown that the rate can be accelerated to $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$. Further, there is a lot of recent work that seeks to avoid saddle points. The methods for these cases are quite involved and require structure. Also, there are certain non-convex problems where achieving event the local minima is NP-Hard.

3.2 f is β Strongly Smooth and Convex

We write the following from the definition of strong smoothness and convexity,

$$\begin{aligned} f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\ f(\mathbf{x}^{t+1}) &\geq f(\mathbf{x}^t) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \right\rangle \end{aligned} \tag{3}$$

Remark 16.2. Here the bound is on t+1 compared to case 1. We are jumping steps due to the constraint of strong smoothness. We cannot apply this when β is very large like the order of number of samples.

$$\begin{aligned} \Phi_{t+1} &\leq \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^* \right\|_2^2 \\ &= \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{z}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\ &\leq \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{x}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \quad (Projection\ Property - I) \\ &= \frac{1}{2\eta_t} (\mathbf{D}_t^2 - \mathbf{D}_{t+1}^2) - \frac{1}{2} (\frac{1}{\eta_t} - \beta) \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \qquad (2ab = a^2 + b^2 - (a-b)^2) \end{aligned} \tag{4}$$

Usually for analysis, we look to simplify the inner product terms. Depending upon the simplicity of subsequent expressions, we use one of the following equations: $2ab = a^2 + b^2 - (a-b)^2$ or $2ab = (a+b)^2 - a^2 - b^2$.

In the expression above, as long as $(\frac{1}{\eta} - \beta) \geq 0$, we can show that the function is decreasing. Setting the value of η to be small enough, helps us to ensure this. Thus, we set $\eta = \frac{1}{\beta}$ giving,

$$\begin{aligned} \sum_{t=1}^{t=T} \Phi_t &\leq \frac{D_\circ^2}{2\eta} \\ \frac{1}{T} \sum_{t=1}^{t=T} \Phi_t &\leq \frac{D_\circ^2}{2T\eta} \\ \Phi_T &\leq \frac{D_\circ^2}{2T\eta} \end{aligned} \tag{5}$$

So, we can say that after T values, there is a value of ϵ for which

$$f(\mathbf{x}^T) \leq f(\mathbf{x}^*) + \epsilon$$

Exercise 16.2. Show that $f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t)$

What we have shown here is that if the step length is small enough, we should get an appreciable decrease in the function. If the decrease is not seen, we make η smaller. To find the appropriate value of the step length is important while optimizing. Popular optimizers like ADAM and Adagrad are essentially searching for this.

3.3 f is β Strongly Smooth and α Strongly Convex

The analysis for this section is similar to the previous case with an additional term for strong convexity.

$$\begin{aligned} f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 \\ f(\mathbf{x}^{t+1}) &\geq f(\mathbf{x}^t) + \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^* - \mathbf{x}^t \right\rangle - \frac{\alpha D_t^2}{2} \end{aligned} \tag{6}$$

Moving forward similarly,

$$\begin{aligned} \Phi_{t+1} &\leq \left\langle \nabla f(\mathbf{x}^t), \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^* \right\|_2^2 - \frac{\alpha D_t^2}{2} \\ &= \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{z}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha D_t^2}{2} \\ &\leq \frac{1}{\eta_t} \left\langle \mathbf{x}^t - \mathbf{x}^{t+1}, \mathbf{x}^{t+1} - \mathbf{x}^* \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha D_t^2}{2} \quad (Projection\ Property - I) \\ &= \frac{1}{2\eta_t} (\mathbf{D}_t^2 - \mathbf{D}_{t+1}^2) - \frac{1}{2} (\frac{1}{\eta_t} - \beta) \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|_2^2 - \frac{\alpha D_t^2}{2} \qquad (2ab = a^2 + b^2 - (a-b)^2) \end{aligned} \tag{7}$$

We know that by definition $\Phi_{t+1} \geq 0$ and setting $\eta = \frac{1}{\beta}$

$$\begin{aligned} D_{t+1}^2 &\leq (1 - \alpha\eta) D_t^2 \\ D_{t+1}^2 &\leq (1 - \frac{\alpha}{\beta}) D_t^2 \\ D_{t+1}^2 &\leq (1 - \frac{\alpha}{\beta}) T D_\circ^2 \\ D_{t+1}^2 &\leq \exp(\frac{-T\alpha}{\beta}) D_\circ^2 \end{aligned} \tag{8}$$

We know by definition that $\alpha < \beta$, thus, $0 < 1 - \frac{\alpha}{\beta} < 1$. We see that the value of D_t decays with an exponential factor. In general, statisticians know this as an exponential or geometric rate of convergence. In the field of optimization, this is known as a linear rate of convergence. This is because, if for example $\frac{\alpha}{\beta} = \frac{1}{10}$ and $D_t^2 = 0.001$, then $D_{t+1}^2 \leq 0.0001$ which shows a linear decrease in order. Similarly, there are methods which achieve a quadratic rate (like Newton's Method) which show ridiculously fast convergence.

Definition 16.1. (Condition Number - κ) Jain et al. (2017)
The condition number for an objective f is a scalar $f : \mathcal{X} \leftarrow \mathbf{R}$ that bounds how much the function value can change relative to a perturbation of the input. It is a measure of the fragility or sensitivity of f.

A well conditioned problem ensures that the output does not change significantly with a slight change in input. In our analysis, the term $\frac{\alpha}{\beta}$ is known as the condition number of the objective f. Usually convergence rates are of the form $\exp(\frac{-T}{\kappa})$. So, the smaller the condition number, the larger is the rate of convergence.

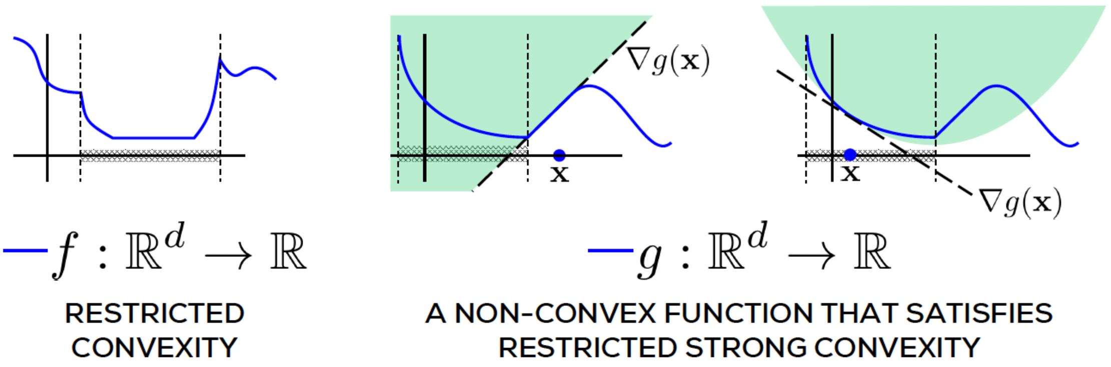
$\nabla g(\mathbf{x})$
$\mathbf{x}$
$\mathbf{x}$
$\nabla g(\mathbf{x})$
$f : \mathbb{R}^d \to \mathbb{R}$
$g : \mathbb{R}^d \to \mathbb{R}$
RESTRICTED CONVEXITY
A NON-CONVEX FUNCTION THAT SATISFIES RESTRICTED STRONG CONVEXITY

Figure 1: A depiction of restricted convexity properties. f is clearly non-convex over the entire real line but is convex within the cross-hatched region bounded by the dotted vertical lines. g is a non-convex function that satisfies restricted strong convexity. Outside the cross-hatched region (again bounded by the dotted vertical lines), g fails to even be convex as its curve falls below its tangent, but within the region, it actually exhibits strong convexity. Jain et al. (2017)

4 Non-Convex Optimization

Here the set $\mathcal{C}$ may be non-convex and and the objective f may be non-convex too. The results that we saw previously for PGD on convex sets and objectives fail to apply in this scenario directly. We need to find a workaround for this problem. We do this by recognizing that the previous analysis for PGD did not require the objective to be convex for the entire space. The properties need to be satisfied only for a constrained set over which optimization is being performed. So, taking this into account, we introduce restricted SC/SS properties.

Definition 16.2. (Restricted SC/SS) A function f is said to be α-RSC and β-RSS over a (possibly non-convex) set $\mathcal{C}$ if $\forall X, Y \in \mathcal{C}$,

$$\frac{\alpha}{2}\|X-Y\|_2^2 \leq f(X) - f(Y) - \langle \nabla f(Y), X-Y \rangle \leq \frac{\beta}{2}\|X-Y\|_2^2$$

Figure 1 (4) shows a depiction of the restricted sc property.

Remark 16.3. The strong smoothness constraint is usually not found to be violated while the strong convexity constraint is violated sometimes.

We now look at the example of a common non-convex optimization problem of sparse recovery.

4.1 Sparse Recovery

Sparse recovery is a popular problem in the fields of statistics, optimization and signal recovery. Our aim is to recover data form a sparse approximation or obtain a sparse approximation of data $\mathcal{X} \in \mathbf{R}^d$ using a p.s.d. design or sketch matrix $A \in \mathbf{R}^{n \times d}$ (that could be learned or handcrafted). The objective is defined as,

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

We analyze the objective.

$$\begin{aligned}
f(\mathbf{x}) &= \|A\mathbf{x} - \mathbf{b}\|_2^2 \\
&= \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{b}^T A\mathbf{x} + \|\mathbf{b}\|_2^2 \\
f(\mathbf{x}) - f(\mathbf{y}) &= \mathbf{x}^T A^T A\mathbf{x} - \mathbf{y}^T A^T A\mathbf{y} - 2\mathbf{b}^T A(\mathbf{x} - \mathbf{y}) \\
&= \mathbf{x}^T A^T A\mathbf{x} + \mathbf{y}^T A^T A\mathbf{y} - 2\mathbf{b}^T A(\mathbf{x} - \mathbf{y}) - 2\mathbf{x}^T A^T A\mathbf{y} + 2(\mathbf{x} - \mathbf{y})^T A^T \mathbf{b} - \langle \nabla f(\mathbf{y} \\
&\text{as } \nabla f(\mathbf{y}) = 2A^T(Ay - b) = 2A^T A\mathbf{y} - 2A^T\mathbf{b} \\
&\text{Completing Squares,} \\
f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle &= (\mathbf{x} - \mathbf{y})^T A^T A(\mathbf{x} - \mathbf{y})
\end{aligned} \tag{9}$$

We find that the objective is m-sc and M-ss where,

$$\begin{aligned}
m &= \lambda_{min}(A^T A) \\
M &= \lambda_{max}(A^T A)
\end{aligned} \tag{10}$$

Remark 16.4. If the dimension n is less than d, we have an under-defined problem and A is not full rank. When $n > d$, we find that A is full rank.

We analyze optimization techniques and algorithms for sparse recovery further in the following two lectures.

References

Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.