Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Rushab Munot
Instructor: Purushottam Kar
Date: March 5, 2018

LECTURE
14

Optimization Techniques-I

Introduction

So far, we have looked at the statistical part of Learning Theory. From this lecture onwards, we will be focusing on the algorithmic part of learning theory. For the next few lectures, we will deal with optimizing convex functions over convex domains. We will start off with a set of algorithms collectively known as the First Order Methods, specifically, we will introduce and analyze the Projected Gradient Descent and the Frank Wolfe algorithms. Before proceeding to these methods we need to formalize and develop the required notation and Tools.

Required Tools

We will now formalize the notions of convexity, convex combinations, projections onto sets, etc.

Definition 14.1 (Convex Combination)**.** Given two points x and y in a vector space V, for any $\lambda \in [0,1]$, $z_\lambda = \lambda x + (1-\lambda)y$ is called a linear cmobination of x and y.

For n points $x_1, \ldots, x_n \in V$, for $\lambda_1, \ldots, \lambda_n \in \mathbb{R}^+ \cup \{0\}$ such that $\sum_{i=1}^n \lambda_i = 1$, the corresponding linear combination is given as

$$z_{\boldsymbol{\lambda}} = \sum_{i=1}^{N} \lambda_i x_i$$

Definition 14.2 (Convex Set)**.** A set $\mathcal{C}$ is said to be convex iff $\forall x, y \in \mathcal{C}$, every convex combination of x and y is also in $\mathcal{C}$.

In simple terms, a set is convex iff for any two points in the set, any point on the line segment between these two points is also in the set.

Definition 14.3 (Convex Functions - definition-I)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex iff its epigraph is convex.

The epigraph of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$epi(f) = \Big\{(x, y) : f(x) \leq y, \quad x \in \mathbb{R}^d, \quad y \in \mathbb{R}\Big\}$$

This particular definition for the convexity of a function is not of much practical use because of the difficulty in verifying it. Given below are several alternative definitions.

Definition 14.4 (Convex (differentiable) Functions - definition-II)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex iff $\forall x, y \in \mathbb{R}^d$, we have

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

This definition captures the intuition the a convex function lies *above* the tangent at any point. We can extend this definition to non-differentiable functions by introducing an alternative to gradients called *subgradients*.

Definition 14.5 (Subgradients)**.** The subgradient of a function f is defined as

$$\partial f = \left\{ g : f(y) \geq f(x) + \langle g, y - x \rangle, \quad x, y \in \mathbb{R}^d \right\}$$

Note that, the subgradient of a function at any point is a set of vectors as opposed to the gradient which is a vector. Also, if f is differentiable at some point x, its subgradient at x is a singleton set consisting of the normal to the tangential hyperplane. Also, observe that the size of the subgradient at any point can be either 0, 1 or infinite.

For example, consider the function $f(x) = [x]_+ = \begin{cases} 0 \text{ if } x \leq 0 \\ x \text{ if } x > 0 \end{cases}$
f is differentiable at all points other than x, thus the subgradient at $x \neq 0$ is $\{0\}$ if $x < 0$ and $\{1\}$ if $x \geq 0$. However, $\partial f(0) = [0, 1]$.

Definition 14.6 (Convex Functions - definition-III)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex iff $\forall x \in \mathbb{R}^d$, $\partial f(x) \neq \emptyset$.

Definition 14.7 (Strong Convexity)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to $\alpha-$strongly convex for $\alpha > 0$ iff $\forall x, y \in \mathbb{R}^d$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\alpha}{2} \|x - y\|_2^2$$

For example, $\|x\|_2^2$ is strongly convex with $\alpha = 2$.

Every strongly convex function is convex. Further, every convex function is trivially strongly convex, i.e. with $\alpha = 0$.

Strong convexity forces the function curve to move away from the tangent as one moves farther from the point of contact. The complementary idea is *strong smoothness*, where the function is forced to move towards the tangent.

Definition 14.8 (Strong Smoothness)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to $\beta-$strongly smooth for $\alpha > 0$ iff $\forall x, y \in \mathbb{R}^d$

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\beta}{2} \|x - y\|_2^2$$

Remark 14.1. Strong convexity implies convexity. Convexity does not imply strong convexity (non-trivially i.e. with $\alpha > 0$). However, strong smoothness does not imply convexity. Neither does convexity imply strong smoothness.

Exercise 14.1. Give an example of a strongly smooth function which is not convex.

Theorem 14.1. Consider a doubly differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. then

1. f is $\alpha-$strongly convex $\implies \nabla^2 f(x) \succeq \alpha I$ which is equivalent to saying that $\lambda_{min}(\nabla^2 f(x)) \geq \alpha$ where $\lambda_{min}(M)$ denotes the smallest eigenvalue of matrix M.

2. f is $\beta-$strongly smooth $\implies \nabla^2 f(x) \preceq \beta I$ which is equivalent to saying that $\lambda_{max}(\nabla^2 f(x)) \leq \beta$ where $\lambda_{max}(M)$ denotes the largest eigenvalue of matrix M.

3. **Lipschitz Gradients:** f is $\beta-$strongly smooth iff $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$.

Remark 14.2. In all the above definitions we have used the second order norm. However, these definitions are valid for any norm.

Projections

We will now define projections from a point onto a set and prove some basic results.

Definition 14.9 (Projection of a point onto a set)**.** Let set $\mathcal{C} \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$. Then the projection of x on the set $\mathcal{C}$ is the set given by

$$\pi_{\mathcal{C}}(x) \triangleq \arg\min_{y \in \mathcal{C}} \|y - x\|_2$$

The projection of a point on a set need not be unique. A point can have multiple projections on a set. Projections onto convex sets, called convex projections, however, must be unique. Such projections also have other interesting properties. Before getting to those we state a trivial property of all projections.

Theorem 14.2 (Projection Property-0)**.** For a set $\mathcal{C}$ and a point x, both in $\mathcal{R}^d$,

$$\|\pi_{\mathcal{C}}(x) - x\|_2 \leq \|y - x\|_2 \, \forall y \in \mathcal{C}$$

Proof. Follows trivially from the definition. □

Projection Property-0 holds for *all* sets, convex or not. However, the next two properties hold only for convex sets. We now look at some properties of projections onto convex sets.

Definition 14.10 (Convex Projections)**.** A convex projection is a projection from a point on a convex set.

Theorem 14.3 (Uniqueness of projections)**.** The projection from a point on a convex set is unique.

Proof. Use the fact that square of L_2 norm is strongly conex. The proof is left as an exercise. □

Remark 14.3. This theorem is not restricted only to L_2 norms. A convex projection is unique for all norms that are strongly convex.

Theorem 14.4 (Projection Property - I)**.** For a convex set $\mathcal{C} \in \mathbb{R}^d$ and a point $x \in \mathbb{R}^d$, let $\hat{x} = \pi_{\mathcal{C}}(x)$. Then

$$\langle x - \hat{x}, y - \hat{x} \rangle \leq 0$$

Proof. The idea of the proof is given below. The proof is left as an exercise.

- Suppose that there exists a point y_0 such that $\langle y - \hat{x}, y - x \rangle > 0$. Argue that $\hat{x}$ cannot be the projection of x, since all points on the line segment between $\hat{x}$ and y must belong to $\mathcal{C}$ or be boundary points (or both). Show that in this case $\hat{x}$ cannot be the least distance from x to $\mathcal{C}$.

□

Projection Property-I says that for a point x outside $\mathcal{C}$ and a point $y \in \mathcal{C}$, the angle between $x - \hat{x}$ and $y - \hat{x}$ is obtuse or right (non-acute).

Theorem 14.5 (First Order Convex Optimization Property)**.** Let $\mathcal{C}$ be convex and f be a convex, differentiable function defined on $\mathcal{C}$. Let $x^* = \arg\min_{z \in \mathcal{C}} f(x)$. Then,

$$\langle \nabla f(x^*), y - x^* \rangle \leq 0$$

Proof. The proof is left as an exercise. Making two cases based on whether $x^* \in \mathit{interior}(\mathcal{C})$ (which implies $\nabla f(x) = 0$ since f is convex) or whether $x^* \in \mathit{boundary}(\mathcal{C})$ may help. □

Theorem 14.6 (Projection Property - II)**.** For a convex set $\mathcal{C} \in \mathbb{R}^d$ and a point $x \in \mathbb{R}^d$, $\forall y \in \mathcal{C}$,

$$\|y - \pi_{\mathcal{C}}(x)\|_2 \leq \|y - x\|_2$$

Proof. **Geometric Proof:** Let $u_\parallel$ and $u_\perp$ be the unit vectors along $x - \hat{x}$ and and the direction perpendicular to it in the plane of z, $\hat{x}$ and x. Let $v_\parallel, v_\perp$ be the decomposition of $z - x$ along $u_\parallel$ and $u_\perp$ respectively. Let $\hat{v}_\parallel, \hat{v}_\perp$ be the decomposition of $z - \hat{x}$ along $u_\parallel$ and $u_\perp$ respectively.

$$z - x = v_\perp + v_\parallel$$
$$z - \hat{x} = \hat{v}_\perp + \hat{v}_\parallel$$

Note that, $v_\parallel = \hat{v}_\parallel$, $\hat{v}_\perp \leq v_\perp$. Pythagoras Theorem completes the proof.

Algebraic Proof:

$$\begin{aligned} \|z - x\|_2^2 &= \|z - \hat{x}\|_2^2 + \|\hat{x} - x\|_2^2 + 2\langle z - \hat{x}, \hat{x} - x \rangle \\ &\geq \|z - \hat{x}\|_2^2 + \|\hat{x} - x\|_2^2 \quad \ldots \text{ Projection Property I} \\ &\geq \|z - \hat{x}\|_2^2 \end{aligned}$$

□

First Order Methods

Projected First Order Methods

The objective that we wish to focus upon, is to minimize a convex function over a convex set. In formal terms, given a convex set $\mathcal{C} \in \mathbb{R}^d$, we wish to minimize a convex function f over $\mathcal{C}$. That is to find x^*

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

Example 14.1. Consider the classic SVM objective

$$\arg\min_w f(w) \qquad \text{where} \qquad f(w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}[1 - y_i w_i x_i]_+$$

The fuctions $\left[1 - y_i w_i x_i\right]_+$ are convex (as $C \geq 0$). So is $\|w\|_2^2$. Thus the objective function is a sum of convex functions and hence is convex itself.

The k^{th} order approximation of a function is the sum of the first k terms of its Taylor Expansion. The zeroth, first and second order approximations, which are most commonly used, are given below.

- Zeroth Order Approximation: $f(w) = f(w^t)$
- First Order Approximation: $f(w) = f(w^t) + \langle \nabla f(w^t), w - w^t \rangle$
- Second Order Approximation: $f(w) = f(w^t) + \langle \nabla f(w^t), w - w^t \rangle + \frac{1}{2}\langle w - w^t, \nabla^2 f(w^t)(w - w^t) \rangle$

Projected Gradient Descent

Projected first order methods are generally very convenient and compuationally inexpensive compared to other methods. The algorithm for Projected Gradient Descent (PGD) is given below. PGD is an iterative method that uses convex projections. However, when projection onto $\mathcal{C}$ is difficult to compute, PGD may become expensive to use.

Notice the use of the $SELECT()$ function, which selects one of x^t or a combination of them.

Algorithm 1: Projected Gradient Descent

1: $\mathbf{x}^0 \leftarrow$ initilize
2: **for** t = 0, 2, . . . , T-1 **do**
3: $\quad z^{t+1} \leftarrow x^t - \eta_t \nabla f(x^t)$
4: $\quad x^{t+1} \leftarrow \pi_{\mathcal{C}}(z^{t+1})$
5: **end for**
6: $\hat{x} \leftarrow SELECT(x^0, x^1, \ldots, x^t)$

We will see in the next lecture that average of $x^0, \ldots, x^T$ converges instead of x^T, which may or may not converge. Choices for $SELECT(x_0, \ldots, x_T)$, include their average $\frac{1}{T}\sum t \in [T]x^t$, or $\arg\min_{t\in[T]} x^t$.

Interior First Order Methods

Interior methods are aggresive as well as careful, and are based on the idea that - why choose a small update when you can be sure that a larger update is possible.

Let $\hat{f}$ denote the first order approximation of f with basepoint x^t. The method minimizes this approximation to select z^{t+1}.

$$z^{t+1} \leftarrow \arg\min_{z\in\mathcal{C}} \hat{f}(z) = \arg\min_{z\in\mathcal{C}} f(x^t) + \langle z - x^t, \nabla f(x^t) \rangle = \arg\min_{z\in\mathcal{C}} \langle z, \nabla f(x^t) \rangle$$

The first order approzimation $\hat{f}(z)$ is a linear function of z. This implies that if we set $x^{t+1} = z^{t+1}$, the optimal value $\hat{x}$ given by the algorithm can be attained only at the boundary of $\mathcal{C}$. This approach will fail if the actual optimal x^* is in the interior of $\mathcal{C}$. There are several approaches to solve this problem which are mentioned below.

Algorithm 2: Frank-Wolfe Algorithm

$\mathbf{x}^0 \leftarrow$ initilize
for t = 0, 2, . . . , T-1 **do**
  $z^{t+1} \leftarrow \arg\min_{z \in \mathcal{C}} \langle z, \nabla f(x^t) \rangle$
  $x^{t+1} \leftarrow (1 - \eta_t)x^t + \eta_t z^{t+1}$
end for
$\hat{x} \leftarrow x^T$

Frank-Wolfe Method

Frank and Wolfe proposed to use a convex combination of z^{t+1} and x^t to get back in the interior of the set. Thus,

$$x^{t+1} \leftarrow (1 - \eta_t)x^t + \eta_t z^{t+1}$$

We will provide the analysis for Frank-Wolfe in the next lecture and show that unlike PGD, we can use the value of x^T (lastest iteration) as a good estimator of x^*. The complete algorithm is given in Algorithm 2.

Line Search

Instead of taking a linear combination of x^t and z^{t+1}, update x^{t+1} to the point that gives the minimum value of f on the line segment between x^t and z^{t+1}.

$$x^{t+1} \leftarrow \arg\min_{\alpha \in [0,1]} (1 - \alpha)x^t + \alpha z^{t+1}$$

Fully Corrective

This variant sets x^{t+1} to be the linear combination of x^0 and $z_1, \ldots, z^{t+1}$ which minimizes f. Though convergence is very fast the method is computationally very expensive.

$$x^{t+1} \leftarrow \lambda_0 x^0 + \sum_{i=1}^{t+1} \lambda_i z_i \quad \text{subject to} \quad \sum_{i=0}^{t+1} \lambda_i = 1 \ , \ \lambda_i \geq 0$$

Interior First Order Methods are useful when computing projections is expensive. A classic example is maximizing the SVM dual. We will see such examples in detail in the next lecture. We will also look at other useful properties like sparsity that come along with Frank Wolfe but require extra effort with PGD. We will finally give a formal analysis for both these methods.

References

Martin Jaggi. *Revisiting Frank-Wolfe: Projection-Free Convex Optimization.* 2013.

Prateek Jain and Purushottam Kar. Non Convex Optimization for Machine Learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.