IE 598: Big Data Optimization **Fall 2016**

Lecture 10: Lower bounds & Projected Gradient Descent– September 22

Lecturer: Niao He *Scriber: Meghana Bande*

Overview: In this lecture, we discuss the lower bounds on the complexity of first order optimization algorithms. We introduce the Projected Gradient Descent for constrained optimization problems and discuss their convergence rates. We illustrate how to find the projections for a few convex sets using examples.

10.1 Recap

In the previous lecture, we have seen the convergence rates of gradient descent and accelerated gradient descent algorithms which can be summarized below. Note that $\kappa = \frac{L}{\mu}$.

	Gradient Descent	Accelerated Gradient Descent
f is convex, L smooth	$O(\frac{LD^2}{t})$	$O(\frac{LD^2}{t^2})$
f is μ strongly convex, L smooth	$O((\frac{\kappa-1}{\kappa+1})^{2t})$	$O((\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2t})$

In order to show the optimality of these algorithms in terms of complexity, lower bounds will be discussed in this lecture.

10.2 Lower Bounds

We look at lower complexity bounds for convex optimization problems which use first order methods for objective functions belonging to certain classes .

10.2.1 Lower Bound for Smooth Convex Problems

Theorem 10.1 *For any t, $1 \leq t \leq \frac{1}{2}(n-1)$, and any $x_0 \in \mathbb{R}^n$, there exists an L-smooth convex function f such that, for any first order method $\mathcal{M}$ that generates a sequence of points x_t such that $x_t \in x_0 + \text{span}\{\nabla f(x_0), \ldots, \nabla f(x_{t-1})\}$, we have*

$$f(x_t) - f(x^*) \geq \frac{3L\|x_0 - x^*\|^2}{32(t+1)^2}.$$

Proof: The sequence of iterates for $f(x)$ starting from x_0 is just a shift of the sequence generated for $f(x+x_0)$ from the origin. Without loss of generality, we assume that $x_0 = 0$.

Consider function f where

$$f(x) = \frac{L}{4} \cdot (\frac{1}{2}x^T A x - e_1^T x),$$

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & \ddots \\ 0 & 0 & \ddots & \ddots \end{pmatrix}_{(n\times n)} \quad \text{and } e_1 = \begin{pmatrix} 1 \\ 0 \\ . \\ 0 \end{pmatrix}_{n\times 1}.$$

1. We show that the function f is convex and L smooth. We have $\nabla^2 f(x) = \frac{L}{4}A$ and we need to show that $0 \preceq A \preceq 4I$. For any $x \in \mathbb{R}^n$, we have

$$\begin{aligned} x^T Ax &= 2\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n-1} x_i x_{i+1} \\ &= x_1^2 + x_n^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2. \end{aligned}$$

   From this, we have $0 \preceq A$. In order to show $A \preceq 4I$ consider the following,

$$\begin{aligned} x^T Ax &= x_1^2 + x_n^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 \\ &\leq x_1^2 + x_n^2 + 2\sum_{i=1}^{n-1}(x_i^2 + x_{i+1}^2) \\ &\leq 4\sum_{i=1}^{n} x_i^2. \end{aligned}$$

   where we have used $(a-b)^2 \leq 2(a^2+b^2)$ since $(a-b)^2 + (a+b)^2 = 2(a^2+b^2)$.

2. We now find the optimal value and optimal solution of the function f. To this end, we set $\nabla f(x) = 0$ which gives us $Ax - e_1 = 0$ or the set of equations

$$2x_1 - x_2 = 1, \quad -x_{i-1} + 2x_i - x_{i+1} = 0, \forall i \geq 2.$$

   We verify that $x_i^* = 1 - \frac{i}{n+1}$ satisfies the above equations.

$$\begin{aligned} 2(1 - \frac{1}{n+1}) - (1 - \frac{2}{n+1}) &= 1 \\ -(1 - \frac{i-1}{n+1}) + 2(1 - \frac{i}{n+1}) - (1 - \frac{i+1}{n+1}) &= 0, \forall i \geq 2. \end{aligned}$$

   Thus, the optimal solution x^* is given by

$$x_i^* = 1 - \frac{i}{n+1}, \forall 1 \leq i \leq n.$$

   and the optimal value of the function is given by

$$f(x^*) = \frac{L}{4}(-\frac{1}{2}e_1^T x^*) = \frac{L}{8}(-1 + \frac{1}{(n+1)}).$$

3. We now obtain bounds in order to prove the theorem.

   We have $x_0 = 0$, $\nabla f(x) = Ax - e_1$ and the iterate $x_1 \in \text{span}\{\nabla f(x_0)\} = \text{span}\{e_1\}$. It is easy to see from the structure of A that $\text{span}\{Ae_1, \ldots, Ae_i\} = \text{span}\{e_1, \ldots, e_{i+1}\}$. Thus, we have $x_t \in \text{span}\{e_1, \ldots, e_t\}$.

Using this, we bound $f(x_t)$.

$$f(x_t) \geq \min_{x \in \text{span}\{e_1,\ldots,e_t\}} f(x) = \frac{L}{8}(-1 + \frac{1}{(t+1)}).$$

Now we try to obtain bounds on $\|x_0 - x^*\|^2$. Note that $x_0 = 0$.

$$\begin{aligned} \|x^*\|^2 &= \sum_{i=1}^{n}(1 - \frac{i}{n+1})^2 \\ &= \sum_{i=1}^{n}(\frac{i}{n+1})^2 = \frac{\sum_{i=1}^{n} i^2}{(n+1)^2} \\ &= \frac{n(n+1)(2n+1)}{6(n+1)^2} \\ &\leq \frac{n+1}{3}. \end{aligned}$$

We have used $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$.

Using the above bounds, we obtain the following,

$$\begin{aligned} \frac{f(x_t) - f(x^*)}{L\|x_0 - x^*\|^2} &\geq \frac{\frac{1}{8}(-1 + \frac{1}{(t+1)}) + \frac{1}{8}(1 - \frac{1}{(n+1)})}{\frac{n+1}{3}} \\ &\geq \frac{\frac{3}{8}(\frac{1}{t+1} - \frac{1}{2t+2})}{2t+2} = \frac{3}{32(t+1)^2}. \end{aligned}$$

We have used $t \leq \frac{1}{2}(n-1)$ which gives $n \geq 2t+1$ in the above to obtain the required bounds.

■

We define l_2 space as the space containing all sequences $x = \{x_i\}_{i=1}^{\infty}$ with finite norm. For convex L smooth functions which are μ strongly convex, we present the following lower bound.

10.2.2 Lower Bound for Smooth and Strongly Convex Problems

Theorem 10.2 *For any $x_0 \in l_2$, and any constants $\mu > 0$ and $L > \mu$, there exists a μ-strongly convex and L-smooth function f such that for any first order method $\mathcal{M}$ that generates a sequence of points x_t such that $x_t \in x_0 + \text{span}\{\nabla f(x_0), \ldots, \nabla f(x_{t-1})\}$, we have*

$$f(x_t) - f(x^*) \geq \frac{\mu}{2}(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2t}\|x_0 - x^*\|^2.$$

where $\kappa = \frac{L}{\mu}$.

Proof:

Consider the function f as follows

$$f(x) = \frac{L-\mu}{4}(\frac{1}{2}x^T A x - e_1^T x) + \frac{\mu}{2}\|x\|^2$$

where A and e_1 are as defined in the proof of Theorem 10.1.

1. We show that it is L smooth and μ strongly convex. We have

$$\nabla^2 f(x) = \frac{L-\mu}{4}A + \mu I.$$

In the proof of Theorem 10.1, we have seen that $0 \preceq A \preceq 4I$. Hence we have,

$$\mu I \preceq \nabla^2 f(x) \preceq LI.$$

2. We now compute the optimal solution of f. From the first order optimality conditions, we have

$$\nabla f(x) = (\frac{L-\mu}{4}A + \mu I)x - (\frac{L-\mu}{4})e_1 = 0$$

which gives $(A + \frac{4}{\kappa-1})x = e_1$. This gives us a set of equations

$$\begin{aligned} 2\frac{\kappa+1}{\kappa-1}x_1 - x_2 &= 1, \\ x_{i+1} + x_{i-1} - 2\frac{\kappa+1}{\kappa-1}x_i &= 0, \forall i \geq 2. \end{aligned}$$

Let $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. We verify that $x_i^* = q^i = (\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^i$ gives us the optimal solution. Substituting q in the optimality conditions gives us

$$q^2 - 2\frac{\kappa+1}{\kappa-1}q + 1 = q(-\frac{(\sqrt{\kappa}+1)^2}{\kappa-1}) + 1 = 0.$$

This solution is also unique due to strong convexity of f.

3. We now find bounds to prove the theorem. From the sum of infinite geometric series, we have,

$$\|x^*\|^2 = \sum_{i=1}^{\infty} q^{2i} = \frac{q^2}{1-q^2}.$$

Similar to the proof of Theorem 10.1, it can be seen that $x_t \in \text{span}\{e_1, \ldots, e_t\}$. Therefore,

$$\|x_t - x^*\|^2 \geq \sum_{i=t+1}^{\infty} q^{2i} = \frac{q^{2(t+1)}}{1-q^2}.$$

From strong convexity of f, we have $f(x_t) - f(x^*) \geq \langle \nabla f(x^*), x_t - x^* \rangle + \frac{\mu}{2}\|x_t - x^*\|^2$. Since x^* is optimal, we have $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for any x. Hence,

$$f(x_t) - f(x^*) \geq \frac{\mu}{2}\|x_t - x^*\|^2.$$

We obtain the final result as follows.

$$\frac{f(x_t) - f(x^*)}{\|x_0 - x^*\|^2} \geq \frac{\mu\|x_t - x^*\|^2}{2\|x_0 - x^*\|^2} \geq \frac{\mu}{2}\frac{\frac{q^{2(t+1)}}{1-q^2}}{\frac{q^2}{1-q^2}} = \frac{\mu}{2}q^{2t}.$$

■

Thus, we see that the accelerated gradient descent algorithms discussed in the previous lecture are optimal in the complexity sense.

10.3 Projected Gradient Descent

So far, we were concerned with finding the optimal solution of an unconstrained optimization problem. In real life, optimization problems we are likely to come across constrained optimization problems. In this section, we discuss how to solve constrained optimization problem:

$$\min_{x \in X} f(x)$$

where f is a convex function and X is a convex set.

If we wish to use gradient descent update to a point $x_t \in X$, it is possible that the iterate $x_{t+1} = x_t - \frac{\nabla f(x_t)}{L}$ may not belong to the constraint set X. In the projected gradient descent, we simply choose the point nearest to $x_t - \frac{\nabla f(x_t)}{L}$ in the set X as x_{t+1} i.e., the projection of $x_t - \frac{\nabla f(x_t)}{L}$ onto the set X.

Definition 10.3 *The projection of a point y, onto a set X is defined as*

$$\Pi_X(y) = \operatorname*{argmin}_{x \in X} \frac{1}{2}\|x - y\|_2^2.$$

Projected Gradient Descent (PGD): Given a starting point $x_0 \in X$ and step-size $\gamma > 0$, PGD works as follows until a certain stopping criterion is satisfied,

$$x_{t+1} = x_t - \gamma \Pi_X(x_t - \nabla f(x_t)), \forall t \geq 1.$$

In this lecture, for an L smooth convex function, we fix the step-size to be $\gamma = \frac{1}{L}$.

Proposition 10.4 *The following inequalities hold:*

1. *If $y \in X$, then $\Pi_X(y) = y$.*
2. *The projection onto a convex set X is non-expansive.*
   $\|\Pi_X(x) - \Pi_X(y)\|_2 \leq \|x - y\|_2$,
   $\|\Pi_X(x) - \Pi_X(y)\|_2^2 \leq \langle \Pi_X(x) - \Pi_X(y), x - y \rangle \leq \|x - y\|_2^2$.

Proof:

1. We first note that the $\frac{1}{2}\|x - y\|_2^2$ is strictly convex since $\nabla^2 f(x) = 1$. Hence the solution to the optimization problem is unique. If $y \in X$, then we have $\frac{1}{2}\|y - y\|_2^2 = 0$. Since $\frac{1}{2}\|x - y\|_2^2 \geq 0$, zero is the optimal value of the function and y is it's unique minimizer, thus giving $\Pi_X(y) = y$.
2. For any feasible x^*, the optimality conditions are given by

   $$\langle \nabla f(x^*), z - x^* \rangle \geq 0, \forall z \in X.$$

   Hence for x, y, we have

   $$\langle \Pi_X(y) - y, \Pi_X(x) - \Pi_X(y) \rangle \geq 0,$$
   $$\langle \Pi_X(x) - x, \Pi_X(y) - \Pi_X(x) \rangle \geq 0,$$

   since $\Pi_X(y), \Pi_X(x) \in X$.

Combining the above equations, we have

$$\langle \Pi_X(x) - \Pi_X(y) - (x-y), 2(\Pi_X(y) - \Pi_X(x)) \rangle \geq 0,$$

$$\langle y - x, \Pi_X(y) - \Pi_X(x) \rangle \geq \langle \Pi_X(y) - \Pi_X(x), \Pi_X(y) - \Pi_X(x) \rangle,$$

From Cauchy-Schwartz we further have

$$\langle y - x, \Pi_X(y) - \Pi_X(x) \rangle \leq \|\Pi_X(x) - \Pi_X(y)\|_2 \|x - y\|_2$$

giving us $\|\Pi_X(x) - \Pi_X(y)\|_2 \leq \|x - y\|_2$.

■

10.3.1 Interpretation

We present a few useful observations.

1. Each iterate in the PGD can be viewed as the minimizer of a quadratic approximation of the objective function, similar to the case of Gradient Descent (GD).

   $$x_{t+1} = \underset{x \in X}{\operatorname{argmin}} \{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\mu}{2} \|x - x_t\|^2 \}.$$

   We also observe that the objective function is non-increasing with each iteration, $f(x_{t+1}) \leq f(x_t)$.

2. The optimal solution x^* is a fixed point of $x = \Pi_X(x - \frac{\nabla f(x)}{L})$ i.e., $x^* = \Pi_X(x^* - \frac{\nabla f(x^*)}{L})$.

   Proof:

   $$\begin{aligned}
   x^* \text{ is optimal} &\iff \langle \nabla f(x^*), z - x^* \rangle \geq 0, \forall z \in X \\
   &\iff \langle -\frac{1}{L} \nabla f(x^*), z - x^* \rangle \leq 0, \forall z \in X \\
   &\iff \langle (x^* - \frac{1}{L} \nabla f(x^*)) - x^*, z - x^* \rangle \leq 0, \forall z \in X \\
   &\iff \langle x^* - (x^* - \frac{1}{L} \nabla f(x^*)), z - x^* \rangle \geq 0, \forall z \in X \\
   &\iff x^* \text{ is the projection of } x^* - \frac{1}{L} \nabla f(x^*)
   \end{aligned}$$

   where the last line follows because the projection of y is the minimizer of the function $\frac{1}{2}\|x - y\|_2^2$ over the set X. ■

3. We can rewrite the iterates of PGD as follows

   $$x_{t+1} = x_t - \frac{1}{L} g_X(x_t).$$

   by defining $g_X(x) = L(x - x^\dagger)$, where $x^\dagger = \Pi_X(x - \frac{1}{L} \nabla f(x))$. The function $g_X(x)$ is often called the *Gradient Mapping*. Note that if the problem is unconstrained, $x^\dagger = x - \frac{1}{L} \nabla f(x)$, $g_X(x) = \nabla f(x)$, thus reducing to the usual gradient descent case.

It can be shown that the key inequalities used to show convergence in the gradient descent method follow in a similar way for the projected gradient descent as well by replacing the gradient with the gradient mapping.

10.3.2 Convergence Analysis

Recall that when showing the convergence rates for the gradient descent algorithm, we used the following properties:

(a) For a L-smooth function f, the iterates given by the gradient descent with step size $\gamma = \frac{1}{L}$ satisfy

$$f(x_{t+1}) - f(x_t) \leq -\frac{\|\nabla f(x_t)\|^2}{2L}.$$

(b) If f is aslo μ-strongly convex, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Similar results hold for the projected gradient descent which are presented below. Details can be found in Section 2.2 from [NES'04].

Proposition 10.5 *For a convex and L-smooth function f, we have*

$$f(x^\dagger) - f(x) \leq -\frac{\|g_X(x)\|^2}{2L}.$$

If f is also μ-strongly convex, we have

$$\langle g_X(x), x - x^* \rangle \geq \frac{\mu}{2}\|x - x^*\|^2 + \frac{1}{2L}\|g_X(x)\|^2.$$

Remark. With these facts, we can immediately adapt previous convergence analysis for GD method to analyze PGD and obtain similar results, namely, a sublinear rate $O(1/t)$ for general smooth convex case and a linear rate $O((1-\kappa^{-1})^t)$ for the smooth strongly convex case. Moreover, if we combine the projected gradient descent with Nesterov's acceleration, we will also obtain the optimal convergence results for constained convex optimization, similar to what we have in the unconstrained case.

Theorem 10.6 *For a convex function f that is L-smooth, the iterates given by the projected gradient descent with step size $\gamma = \frac{1}{L}$ satisfy*

$$f(x_t) - f(x^*) \leq \frac{2L}{t}\|x_0 - x^*\|^2.$$

If f is further μ-strongly convex, we have

$$\|x_t - x^*\|^2 \leq (1 - \frac{\mu}{L})^t\|x_0 - x^*\|^2.$$

Proof: The proofs are similar to the gradient descent case. We present the proof for μ-strongly convex case.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \frac{1}{L}g_X(x_t) - x^*\|^2 \\
&= \|x_t - x^*\|^2 + \frac{1}{L^2}\|g_X(x_t)\|^2 - \frac{2}{L}\langle g_X(x_t), x_t - x^* \rangle \\
&\leq \|x_t - x^*\|^2 + \frac{1}{L^2}\|g_X(x_t)\|^2 - \frac{2}{L}(\frac{\mu}{2}\|x_t - x^*\|^2 + \frac{1}{2L}\|\nabla g_X(x_t)\|^2) \\
&= (1 - \frac{\mu}{L})\|x_t - x^*\|^2 \\
&\leq (1 - \frac{\mu}{L})^t\|x_0 - x^*\|^2.
\end{aligned}$$

where we have used Proposition 10.5 to bound the inner product. ■

10.3.3 Examples

We present a few examples to illustrate how to find projection onto different sets.

1. l_2 **ball**: $X = \{x : \|x\|_2 \leq 1\}$. The projection is given by

$$\Pi_X(y) = \begin{cases} y & \text{for } \|x\|_2 \leq 1, \\ \frac{y}{\|y\|_2} & \text{for } \|x\|_2 > 1. \end{cases}$$

2. **Box**: $X = \{x \in \mathbb{R}^n : \ell \leq x \leq u\}$. The i^{th} coordinate of the projection, $\forall 1 \leq i \leq n$, is given by

$$[\Pi_X(y)]_i = \begin{cases} \ell_i & \text{for } x_i < \ell_i, \\ x_i & \text{for } \ell_i \leq x_i \leq u_i, \\ u_i & \text{for } x_i > u_i. \end{cases}$$

3. **Simplex**: $X = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i \leq 1\}$. Computing the projection requires solving the quadratic problem

$$\Pi_X(y) = \underset{x_i \geq 0; \sum x_i \leq 1}{\operatorname{argmin}} \frac{1}{2}\|x - y\|_2^2.$$

   If $y \in X$, then $\Pi_X(y) = y$. Let's consider the case when $y \notin X$. We form the Langrangian dual function for $\lambda \geq 0$ and $\mu \geq 0$,

$$L(x, \lambda, \mu) = \frac{1}{2}\|x - y\|_2^2 + \lambda^T(-x) + \mu(\sum x_i - 1).$$

   We use KKT conditions in order to solve the optimization problem. The optimal solution satisfies

$$\begin{aligned} x_i - y_i - \lambda_i + \mu = 0, \forall i & \quad \text{(optimality)} \\ \lambda^T x = 0 & \quad \text{(complementary slackness)} \\ \mu(\sum x_i - 1) = 0 & \quad \text{(complementary slackness)} \\ \lambda \geq 0, \mu \geq 0 & \quad \text{(feasibility)} \\ x \geq 0, \sum_i x_i \leq 1 & \quad \text{(feasibility)} \end{aligned}$$

   Let μ^* be such that $\sum(y_i - \mu^*)_+ = 1$, where $(x)_+ = \max\{x, 0\}$. Let $x_i^* = (y_i - \mu^*)_+$ and $\lambda_i = x_i^* - (y_i - \mu^*)$ for any i. We can easily verify that the solution $(x^*; \lambda^*, \mu^*)$ satisfies the KKT conditions listed above. Therefore, the projection of y onto the simplex set X is given by,

$$\Pi_X(y) = \begin{cases} y & \text{for } y \geq 0, \sum y_i \geq 0, \\ (y - \mu^*)_+ & \text{otherwise.} \end{cases}$$

   where μ^* is the solution to the constraint $\sum(y_i - \mu)_+ = 1$. Note that μ^* can be found using tools such as the bisection method.

4. **Nuclear ball**: $X = \{x \in \mathbb{R}^{m \times n} : \|x\|_{\text{nuc}} = \sum \sigma_i(x) = \|\sigma(x)\|_1 \leq 1\}$.

   From the singular value decomposition of y, there exist appropriate matrices U, V, Σ such that $y = U\Sigma V^T = \sum \sigma_i u_i v_i^T$.

Let $x = Ux'V^T$. Since we are looking for the projection of a diagonal matrix Σ, the closest x' will be a diagonal matrix and well thus x' denotes the singular values of x. Let the singular values of y be denoted by a vector $\sigma \in \mathbb{R}^{\text{rank}(y)}$ and that of x' by $\sigma' \in \mathbb{R}^{\text{rank}(y)}$.

$$\Pi_X(y) = \underset{\|x\|_{nuc} \leq 1}{\text{argmin}} \frac{1}{2}\|x - y\|_2^2 = \underset{\sigma' \geq 0; \sum \sigma'_i \leq 1}{\text{argmin}} \frac{1}{2}\|\sigma' - \sigma\|_2^2.$$

This is equivalent to projecting onto a simplex. Therefore, the projection of y onto the nuclear ball is given by,

$$\Pi_X(y) = \begin{cases} y & \text{for } \|\sigma(y)\|_1 \leq 1, \\ \sum(\sigma_i - \mu^*)_+ u_i v_i^T & \text{otherwise.} \end{cases}$$

where μ^* is the solution to the constraint $\sum(\sigma_i - \mu)_+ = 1$. Note that in order to compute the projection, we need to find the full singular value decomposition of an $m \times n$ matrix, the complexity of which is known as $O(mn^2)$. Hence using projected gradient descent may be computationally prohibitive for high-dimensional problems.

References

[NES '04] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Springer, 2004.