EE C227C Convex Optimization and Approximation	Lecture 11 - 2/22/2022

Lecture 11: Level Method

Lecturer: Jiantao Jiao *Scribe: Yulong Dong, Baturalp Yalcin*

1 Level Method

In this lecture, we will give the proof for the level method. Given search points $x_1, x_2, \cdots, x_i$, we can construct a sequence of models where the i-th model is given by a piecewise linear function

$$f_i(x) = \max_{1 \le j \le i} \left(f(x_j) + \langle \nabla f(x_j), x - x_j \rangle \right). \tag{1}$$

Here, f is a convex function and the terms inside the max are the global lower bound of the $f(x)$. Hence, $f_i(x)$ is a global lower bound of the function f. Let G be a convex compact set. We consider a sequence of upper bounds and lower bounds.

$$f_i^- = \min_{x \in G} f_i(x), \quad f_i^+ = \min_{1 \le j \le i} f(x_j), \quad \text{and } \Delta_i = f_i^+ - f_i^-. \tag{2}$$

Following the construction, we have the monotonicity

$$f_1^- \le f_2^- \le \cdots \le f^*, \quad f_1^+ \ge f_2^+ \ge \cdots \ge f^*, \quad \text{and } \Delta_1 \ge \Delta_2 \ge \cdots \ge 0. \tag{3}$$

We have the nondecreasing sequence of f_i^- because we are adding another piecewise linear function to $f_i(x)$. As a result, the global lower bound does not decrease as more lower bound is included. On the other hand, f_i^+ is a nonincreasing sequence because we look for the best point we have searched at each f_i^+ and we have more points as the iterate number increases. Consequently, Δ_i is a nonincreasing sequence. The algorithm for the level method is given as follows.

ALGORITHM 1: Level Method

for $i = 2, 3, \cdots$ **do**
  Solve $\min_{x \in G} f_i(x)$ and get f_i^-.
  Form the level $l_i = (1-\lambda) f_i^- + \lambda f_i^+$.
  Project $x_{i+1} = \Pi_{Q_i}(x_i)$ onto the convex set $Q_i = \{x \in G | f_i(x) \le l_i\}$.
end for

Remarks.

(i) The level l_i lies between the lower bound and the upper bound. The upper bound is the best performance one can ever achieve until the given step. Therefore, the search points in the previous steps will not be in the set Q_i since $l_i < f_i^+$.

(ii) Note that Q_i is a convex set because level set of any convex function is a convex set and intersection of convex sets is also a convex set.

(iii) In addition, the projection onto a convex set is not necessarily easy. However, Q_i has a desirable structure because we add piecewise linear constraints only. The number of such constraints increases significantly as the iteration number increases. Thus, it might be desirable to drop some of the constraints to allow for an easier projection operation. This idea leads to the truncated level method.

(iv) Furthermore, if Q_i is an empty set, it implies that we reached the desired level of accuracy.

(v) Last but not at least, choosing λ close to 0 or 1 leads to quite slow convergence. Therefore, λ should be chosen carefully in $(0, 1)$.

Interpretations. We define parametrized points as

$$x(d) = \arg\min_{x \in G} \left(f_i(x) + \frac{d}{2} \|x - x_i\|^2 \right). \tag{4}$$

Then, we will show that the x_i generated by the level method is the parametrized point for some d.

When $d = 0$, we have $f_i(x(d)) = \min_{x \in G} f_i(x) = f_i^-$. When $d \to \infty$, we claim $x(d) \to x_i$ and $f_i(x(d)) \to f_i(x_i) = f(x_i) \geq f_i^+$. Therefore, by continuity, there exists d so that $l(d) = l_i$. Note that we have $f_i(x_i) = f(x_i)$ because $f_i(x) \geq f(x_i)$ by the definition of $f_i(x)$ and $f_i(x) \leq f(x)$ by the convexity of the function f.

We also claim that $x(d)$ is the closest point to x_i in the set $B = \{x \in G | f_i(x) \leq f_i(x(d))\}$. We denote $l(d) = f_i(x(d))$. Note that $x(d) \in B$ trivially. We want to show that whether there are any point in the set that is closest to x_i but not $x(d)$. It follows the argument below.

1. If $x_i \in B$, then $f_i(x_i) \leq f_i(x(d))$ by the construction of B. At the same time, the definition of $x(d)$ implies that $f_i(x(d)) \leq f_i(x(d)) + \frac{d}{2} \|x(d) - x_i\|^2 \leq f_i(x_i)$. Therefore, $f_i(x_i) = f_i(x(d))$ and $x(d) = x_i$.

2. If $x_i \notin B$, then it equivalently says that we can not find $y \in B$ so that $\|y - x_i\| < \|x(d) - x_i\|$ and $f_i(y) \leq f_i(x(d))$. If so, these two inequalities yield $f_i(y) + \frac{d}{2} \|y - x_i\|^2 < f_i(x(d)) + \frac{d}{2} \|x(d) - x_i\|^2$. Then, $x(d)$ is not the minimizer of the objective function in the defining equation, which is a contradiction.

2 Theorem Statement and Proof

Theorem 1 (Theorem 8.2.1 [1])**.** *Assume $\sup_{x,y \in G} \|x - y\|_2 \leq D$. The function f is convex and L-Lipschitz. Then,*

$$T > \frac{1}{(1-\lambda)^2 \lambda (2-\lambda)} \frac{L^2 D^2}{\epsilon^2} \Rightarrow \Delta_T \leq \epsilon \tag{5}$$

Before the proof, we can interpret the results of the theorem.

(i) Whenever the parameter λ is close to 0 or 1, the number of required iterations for ϵ optimality gap explodes.

(ii) In addition, the analytical complexity $T := \Omega\left(\frac{L^2 D^2}{\epsilon^2}\right)$ is at the same order as subgradient method and the mirror descent discussed in the earlier lectures.

(iii) In fact, this bound is tight for non-smooth minimization in the black box theory. Despite achieving the same worst-case result in the subgradient method, the level method converges much faster than subgradient method in the average case. For instance, the level method converges much faster than subgradient method for the MAXQUAD objective function.

(iv) Since the analysis of average complexity is challenging, the gap between the theory and practice is closed by trying out all the possible appropriate algorithms to solve the problems.

Proof

Let's define a set of intervals $I = \{1, 2, \ldots, T\}$. The definition of intervals are shown in Figure 1. I_1 includes every iterates such that $\Delta_i \leq \frac{\Delta_T}{1-\lambda}$. Note that we might not need to define T intervals necessarily as shown in the Figure 1 and $k \geq 1$ intervals might be sufficient. In addition, $i(s)$ denotes the right-most point in the interval I_s. Formally,

$$i(s) = \arg\min \left\{ i | \Delta_i \geq \frac{\Delta_{i(s-1)}}{1-\lambda} \right\} \tag{6}$$

Note that $i(1) = T$ and $\Delta_{i(1)} = \Delta_T$.

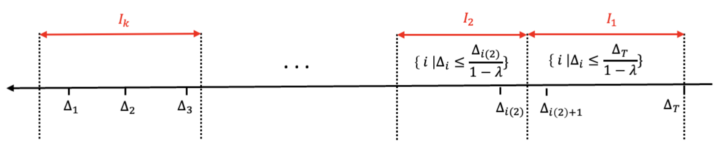

Figure 1: Illustration of Intervals

Key Observation:

$$\cap_{i \in I_s} Q_i \ni U_s \text{ where } U_s \triangleq \arg\min_{x \in G} f_{i(s)}(x) \tag{7}$$

In words, the above equation implies that there exists a point U_s in every Q_i where i is in the index set that belongs to I_s. In addition, U_s is the global minimizer of the $f_{i(s)}(x)$.

First, we can prove the correctness of the observation. For $i \in I_s$, we have $i \leq i(s)$ and

$$f_i(U_s) \leq f_{i(s)}(U_s) = f^-_{i(s)} = f^+_{i(s)} - \Delta_{i(s)} \tag{8}$$

The above inequality holds because $i \leq i(s)$ and when we increase the index, we increase the global lower bound of the function f. The equalities follows from the definitions because U_s is the global minimizer of the $f_{i(s)}$ Therefore, we obtain

$$f^+_{i(s)} - \Delta_{i(s)} \leq f^+_i - \Delta_{i(s)} \leq f^+_i - (1-\lambda)\Delta_i = l_i \tag{9}$$

The first inequality is due to the fact that $f^+_i \geq f^+_{i(s)}$. The second inequality follows from the definition of the points within the interval, i.e. $\Delta_{i(s)} \geq (1-\lambda)\Delta_i, \forall i \in I_s$ and the last equality follows from the definition of Δ_i and l_i. We obtained the inequality $f_i(U_s) \leq l_i, \forall i \in I_s$. As a result, $U_s \in \cap_{i \in I_s} Q_i$.

The key idea of the proof is to upper bound $N_s \triangleq |I_s|$. Mainly, we want to show that N_s is not large and the number of elements decreases exponentially fast. Let $i(s)$: last point in I_s and $j(s)$: first point in I_s. Since we know $U_s \in Q_i$ and x_{i+1} is projection of x_i onto Q_i, we can utilize inverse triangle inequality. The idea is we have an obtuse angle between the vectors as depicted in the Figure 2.

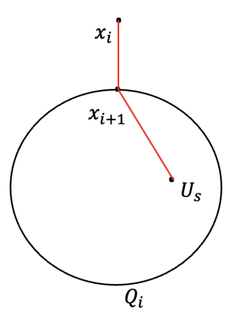

Figure 2: Geometry of the Projection onto Compact Convex Set

Specifically, we have the following since $U_s \in Q_i$ and x_{i+1} is the projection point of x_i.

$$\|x_{i+1} - U_s\|_2^2 \leq \|x_i - U_s\|_2^2 - \|x_i - x_{i+1}\|_2^2 \tag{10}$$

We can use telescoping sum argument to obtain the following.

$$\sum_{i \in I_s} \|x_i - x_{i+1}\|_2^2 \leq \|x_{j(s)} - U_s\|_2^2 - \|x_{i(s)} - U_s\|_2^2 \leq \|x_{j(s)} - U_s\|_2^2 \leq D^2 \tag{11}$$

where the second inequality follows from the fact that $\|x_{i(s)} - U_s\|_2^2 \geq 0$.

Next, we can use Lipschitz property of the function f. If the improvement in the function value is large, we know that distance between the successive iterates cannot be small. We will utilize this property to conclude the proof. We know that the objective value f_i improves at least $(1-\lambda)\Delta_i$ from x_i to x_{i+1} by the definition l_i. Since the function f is L-Lipschitz and f_i is piecewise linear functions of f, f_i is L-Lipschitz. Specifically, the gradient of the $f_i(x)$ is upper bounded by the maximum of $\nabla f(x_j)$ which are upper bounded by L. Hence, we have

$$\|x_i - x_{i+1}\| \geq \frac{(1-\lambda)\Delta_i}{L} \geq \frac{(1-\lambda)\Delta_{i(s)}}{L} \tag{12}$$

The inequality (11) and the Lipschitz condition (12) implies the movement between the successive iterates cannot be too large, but cannot be too small either. Combining the last two inequalities, we obtain

$$N_s \leq \frac{1}{(1-\lambda)^2} \frac{L^2 D^2}{\Delta_{i(s)}^2} \tag{13}$$

By the definition of the intervals, we observe

$$\Delta_{i(s)} > \frac{\Delta_{i(1)}}{(1-\lambda)^{s-1}} \tag{14}$$

Suppose for contradiction $\Delta_T > \epsilon$. Then, by using the upper bound on the N_s, we obtain

$$N_s \leq \frac{1}{(1-\lambda)^2} \frac{L^2 D^2 (1-\lambda)^{2(s-1)}}{\epsilon^2} \tag{15}$$

Consequently,

$$T = \sum_{s \geq 1} N_s \leq \sum_{s \geq 1} \frac{1}{(1-\lambda)^2} \frac{L^2 D^2 (1-\lambda)^{2(s-1)}}{\epsilon^2} \leq \frac{1}{(1-\lambda)^2 \lambda (2-\lambda)} \frac{L^2 D^2}{\epsilon^2} \tag{16}$$

In the last inequality, we use the fact that $\sum_{s \geq 1} ((1-\lambda)^2)^{s-1}$ is a geometric series with $|(1-\lambda)^2| < 1$ and we use the infinite geometric series sum formula. As a result, whenever $T > \frac{1}{(1-\lambda)^2 \lambda (2-\lambda)} \frac{L^2 D^2}{\epsilon^2}$, we must have $\Delta_T \leq \epsilon$.

□

In the next lecture, we will talk about the acceleration of gradient descent.

References

[1] A. Nemirovski, "Lectures on modern convex optimization," in *Society for Industrial and Applied Mathematics (SIAM.* Citeseer, 2001.