COMS 4995-3: Advanced Algorithms April 10, 2017

Lecture 20 – Optimization: Natural Greedy and Gradient Descent

Instructor: *Alex Andoni* Scribes: *Runqing Yang, Shu Fang*

1 Introduction

In the previous lectures we have talked about linear programming, and tried to optimize functions which subject to some constrains. We will step back a lit bit in terms of the variables, talking about some general optimization problems. In particular, we will minimize a function $f(x)$, which is not necessarily linear, where x is unconstrained. Considering that we are still talking about linear programming, here we provide a way to convert constrained optimization to unconstrained optimization.

So, suppose we have a linear programming problem to minimize $f(x)$, where x belongs to a constrained set K, we can transfer $f(x)$ into a function $g(x)$, where x is unconstrained, i.e., $x \in \mathbb{R}^n$. Mathematically, if we have a probelm to find:

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

we can convert it to a new problem that:

$$\min_{x \in \mathbb{R}^n} g(x)$$

where $g(x)$ is defined as:

$$g(x) = \begin{cases} f(x), & x \in K \\ +\infty, & x \notin K \end{cases} \tag{1}$$

Although this $g(x)$ has some "pulses", the solution can be equivalent. We will discuss more details in this lecture.

2 Natural Greedy Algorithm

Natural greedy algorithm is a iterative algorithm to obtain optimal value. It follows two simple operations:

- start at $x_0 \in \mathbb{R}^n$
- jump to a near point, which improves $f(x)$: $x^t \to x^{t+1}, f(x^t) \to f(x^{t+1})$

Here, we make our first assumption to $f(x)$. We assume that $f(x)$ is "nice", which means that the value of $f(x)$ doesn't jump too much with continuous derivatives that all exist, to apply Taylor approximation.

$$f(x+\delta) = f(x) + \nabla f(x)^T \cdot \delta + \frac{1}{2}\delta^T \cdot \nabla^2 f(x) \cdot \delta$$

Here, $\nabla f(x)$ is a n-dimension vector, where $[\nabla f(x)]_i = \frac{\partial f}{\partial x_i}(x)$; $\nabla^2 f(x)$ is a $n \times n$ dimension matrix, where $[\nabla^2 f(x)]_{i,j} = \frac{\partial^2 f}{\partial x_i \partial y_i}(x)$, $\forall i, j \in [n]$; plus, we define δ is the step size and δ is reasonably small.

And then we can compute the new point via $y = \alpha x + (1-\alpha)(x+\delta), \alpha \in [0,1]$

$$\Rightarrow f(x) \approx f(x) + \nabla f(x)^T \delta + \underbrace{O(||\delta||^2)}_{small}$$

At a small scale, we can assume $f(x)$ is linear, so we have:

$$f(x+\delta) \approx f(x) + \nabla f(x)^T \delta$$

Now, keep x, we are trying to optimize $f(x+\delta)$ w.r.t.δ, with the assumption that $||\delta||$ is reasonably small.

$$\begin{aligned} \underset{\delta:||\delta|| \leq \epsilon}{arg.min}\, f(x) + \nabla f(x)^T \delta &= \underset{\delta:||\delta|| \leq \epsilon}{arg.min}\, \nabla f(x)^T \delta \\ &= -\eta \nabla f(x)^T \\ s.t.\ \eta ||f(x)||^2 &= \epsilon \end{aligned} \tag{2}$$

3 Gradient Descent Algorithm

We now introduce gradient descent method to solve a optimal value problem. Actually, it is not a single method, but a general framework with many possible realizations. We describe some concrete variants and analyze their performance. The performance guarantees that we are going to obtain will depend on assumptions that we make about f. Basically, this algorithm can be described as following:

- start at $x_0 \in \mathbb{R}^n$
- at step t: $x^{t+1} = x^t - \eta \nabla f(x^t)$
- do it for T steps

In the above algorithm description, η is fixed in the entire algorithm or depends on t.

3.1 Assumption I

Definition 1. *For $\beta > 0$, f is **β-smooth** if: $\forall$ x and $y = x + \delta$, we have $||\nabla f(y) - \nabla f(x)|| \leq \beta \cdot ||y - x||$*

The β-smooth definition is equivalent to the following: $\forall y \in \mathbb{R}^n$, we have $y^t \cdot \nabla^2 f(x) \cdot y \leq \beta \cdot ||y||^2$. So we have:

$$f(x+\delta) \leq f(x) + \underbrace{\nabla f(x)^T \delta + \frac{\beta}{2}||\delta||^2}_{\delta = -\eta \cdot \nabla f(x),\ we\ choose\ \eta\ to\ minimize}$$

$$
\begin{aligned}
\eta &= \underset{\delta=-\eta\cdot\nabla f(x)}{arg.min}\ \nabla f(x)^T\delta + \frac{\beta}{2}||\delta||^2 \\
&= \underset{\delta=-\eta\cdot\nabla f(x)}{arg.min}\ -\eta\cdot\nabla f(x)^T\cdot\nabla f(x) + \frac{\beta}{2}\cdot\eta^2||\nabla f(x)||^2 \\
&= \underset{\delta=-\eta\cdot\nabla f(x)}{arg.min}\ -\eta\cdot||\nabla f(x)||^2 + \frac{\beta}{2}\cdot\eta^2||\nabla f(x)||^2 \\
&= \underset{\delta=-\eta\cdot\nabla f(x)}{arg.min}\ -\eta + \frac{\beta}{2}\cdot\eta^2 \\
&= \frac{1}{\beta} \qquad\qquad \Rightarrow set\ \eta = \frac{1}{\beta}
\end{aligned} \tag{3}
$$

With the progress in (3), we can now continue to compute $f(x+\delta)$.

$$
\begin{aligned}
f(x+\delta) &\leq f(x) - \frac{1}{\beta}||\nabla f(x)||^2 + \frac{1}{2\beta}||\nabla f(x)||^2 \\
&= f(x) - \frac{1}{2\beta}||\nabla f(x)||^2
\end{aligned} \tag{4}
$$

- if $\nabla f(x) \neq 0$, we need some progress
- if $\nabla f(x) = 0$, then x can be (shown in Fig. 1):
  - global min
  - local min
  - saddle point
    - in some direction, increase
    - in some direction, decrease
    - or stay constant
  - local maximum

When we encounter with *saddle point* and *local maximum*, we'll need some random processing such as giving small perturbation.

3.2 Assumption II

Definition 2. *f is **convex** iff.* $f(\lambda x + (1-\lambda)\eta) \leq \lambda f(x) + (1-\lambda)f(y)$, $\forall x, y \in \mathbb{R}^n, \lambda \in [0,1]$

This definition is equivalent to that, $\forall \delta, x \in \mathbb{R}^n, \delta^T\cdot\nabla^2 f(x)\cdot\delta \geq 0$, which means that $\nabla^2 f(x)$ is a positive-definite matrix with all eigenvalue positive.

Claim 3. *critical points $\Leftrightarrow$ local global minimum*

Proof.

$$
\begin{aligned}
f(x+\delta) &= f(x) + \nabla f(x)^T\delta + \underbrace{\frac{1}{2}\delta^T\nabla^2 f(y)\delta}_{\geq 0} \\
&\geq f(x) + \nabla f(x)^T\delta
\end{aligned} \tag{5}
$$

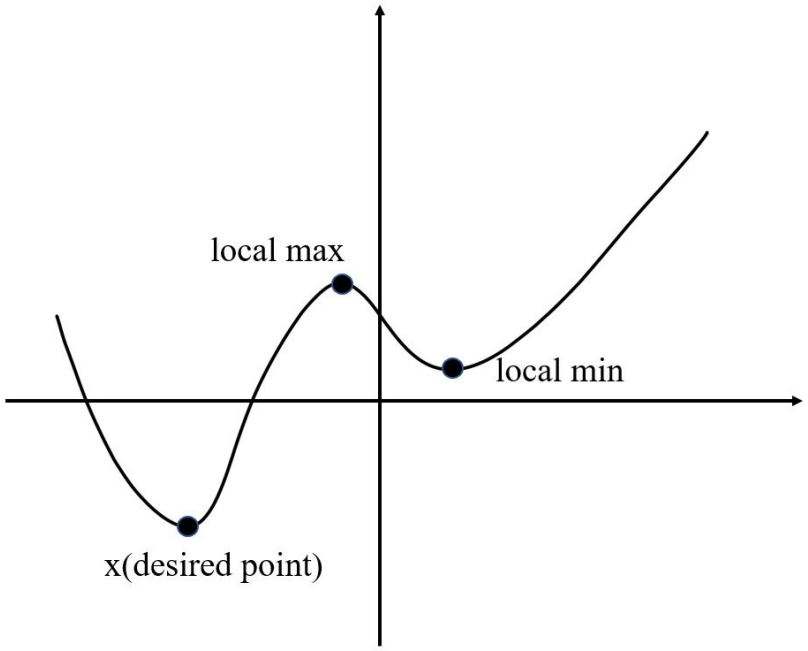

Figure 1: Critical points in a function

If $\nabla f(x) = 0 \Rightarrow \forall \delta, f(x+\delta) \geq f(x)$

- when f is β-smooth $\Rightarrow$ progress in each step as long as $\nabla f(x) \neq 0$, then we have:

$$f(x+\delta) \leq f(x) + \frac{1}{2\beta}||\nabla f(x)||^2$$

- when f is convex, if $\nabla f(x) = 0$, then we conclude that x is a global minimum.

□

So, how fast will we find the global maximum? or convergence.
Let $x^* = opt = arg.min f(x)$, we are looking for some moment T, that:

$$f(x^T) - f(x) \leq \epsilon$$

Here, we track how far from *opt* we are:

$$f(x^*) \geq f(x) + \nabla f(x)^T \cdot (x^* - x) + \underbrace{(x^* - x) \cdot \nabla^2 \ldots (x^2 - x^*)}_{\geq 0}$$

$$\begin{aligned} \Rightarrow f(x) - f(x^*) &\leq -\nabla f(x)^T \cdot (x^* - x) \\ &\leq ||\nabla f(x)|| \cdot ||x^* - x|| \end{aligned}$$

If $||\nabla f(x)|| \leq \frac{\epsilon}{||x^* - x||}$, we can conclude that $f(x) - f(x^*) \leq \epsilon$. Notice that the annoying part here is $||x^* - x||$, which we want to get rid of.

Claim 4. *The time complexity to find* $f(x^T) - f(x^*) \leq \epsilon$ *is subject to* $T = O(\frac{\beta||x^0 - x^*|^2}{\epsilon})$

- steps is exponential in target
- bits of precision

3.3 Assumption III

Definition 5. *f is* ***α-strong convex*** *if,* $\forall \delta, x$*, and for some* $\alpha \geq 0$*:*

$$\delta^T \nabla^2 f(x) \delta \leq \alpha \cdot ||\delta||^2$$

This definition is equivalent to say that $\nabla^2 f(x)$ has a minimum eignvalue λ, and $\lambda \geq \alpha > 0$. This will allow us to lower bound how much to progress. For:

$$f(x + \delta) \geq f(x) + \nabla f(x)^T \cdot \delta + \frac{\alpha}{2}||\delta||^2 \tag{*}$$

First, we can remove the dependence of $||x^0 - x^*||$. We can now connect $||x^0 - x^*||$ to $f(x) - f(x^*)$:

$$(*) \Rightarrow f(x) \geq f(x^*) + \nabla f(x^*)^T \delta + \frac{\alpha}{2}||\delta||^2$$

$$\delta = x - x^*$$

$$f(x) - f(x^*) \geq \frac{\alpha}{2} \cdot ||x - x^*||^2 \tag{**}$$

The progress made per step is $-\frac{1}{2\beta}||\nabla f(x)||^2$, then

$$\begin{aligned} f(x^{t+1}) - f(x^*) &\leq f(x^t) - f(x^*) - \underbrace{\frac{1}{2\beta}||\nabla f(x^t)||^2}_{\geq 0} \\ &\leq f(x^t) - f(x^*) - \frac{1}{2\beta}\frac{[f(x^t) - f(x^*)]^2}{||x^t - x^*||^2} \\ &\leq f(x^t) - f(x^*) - \frac{1}{2\beta}\frac{\alpha}{2}[f(x^t) - f(x^*)] \;\Rightarrow \textit{apply } (**) \\ &\leq [f(x^t) - f(x^*)] \cdot (1 - \frac{\alpha}{4\beta}) \end{aligned}$$

So, within $T = O(\frac{\alpha}{\beta} \cdot \lg \frac{f(x^0) - f(x^*)}{\epsilon})$, we ought to reach the *opt* solution up to ϵ.