Lecture 05, 24/09/2019

Summary of previous lecture

We assumed that

1) f is differentiable
2) $\nabla f(x)$ is Lipschitz continuous

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L \|y - x\|_2$$

for some positive constant L.

Lipschitz continuity of the gradient implies that the gradient cannot change arbitrarily fast.

Lipschitz continuity is a common assumption for Machine Learning models.

For example many ML objective functions are Lipschitz continuous: least squares, logistic regression, deep neural networks.

We defined gradient descend

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k).$$

and we showed that at each iteration gradient descend decreases the objective function

$$f(x_{k+1}) < f(x_k)$$

(assuming that x_k is not a stationary point $\nabla f(x_k) \neq 0$)

We also showed that if a function is differentiable and $\nabla f(x)$ is Lipschitz continuous, then we can upper bound the function

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|y-x\|_2^2$$

$$\forall x, y \in \mathbb{R}^n$$

Outline

1) Amount of decrease of the objective function

2) Convergence rate of gradient descent.

Decrease of the objective function

We will use

1) $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

2) $f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|y-x\|_2^2$

to prove

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2$$

proof

We have $f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2 \quad \forall\, x, y \in \mathbb{R}^n$

Set 1) $y = x_k - \frac{1}{L} \nabla f(x_k)$ and 2) $x = x_k$

to get

$$f\left(x_k - \frac{1}{L}\nabla f(x_k)\right) \le f(x_k) + \nabla f(x_k)^T \left(x_k - \frac{1}{L}\nabla f(x_k) - x_k\right) + \frac{L}{2} \left\| x_k - \frac{1}{L} \nabla f(x_k) - x_k \right\|_2^2$$

$\Leftrightarrow$

$$f\left(x_k - \frac{1}{L}\nabla f(x_k)\right) \le f(x_k) + \nabla f(x_k)^T \left(-\frac{1}{L}\nabla f(x_k)\right) + \frac{L}{2} \left\| -\frac{1}{L} \nabla f(x_k) \right\|_2^2$$

$$f\left(x_k - \frac{1}{L}\nabla f(x_k)\right) \leq f(x_k) - \frac{1}{L}\|\nabla f(x_k)\|_2^2 + \frac{1}{2L}\|\nabla f(x_k)\|_2^2$$

$$= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$$

Comments

1) This shows that gradient descent with step-size $a_k = \frac{1}{L}$ is guaranteed to decrease the objective function

2) Amount of decrease depends on the length of the gradient $\|\nabla f(x)\|_2^2$

Convergence rate

Discussion

Using the inequality

$$f(\underbrace{x_k - \tfrac{1}{L}\nabla f(x_k)}_{x_{k+1}}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$$

we got that

$$\|\nabla f(x_k)\|_2^2 \le 2L\left(f(x_k) - f(x_{k+1})\right)$$

If we assume that f is bounded from below

$$\tilde{f} \le f(x) \quad \forall x \in \mathbb{R}^n$$

Then because $f(x_{k+1}) < f(x_k)$ then as $k \to +\infty$

we must have $f(x_k) - f(x_{k+1}) \to 0$, which

in combination with the above inequality gives

us $\|\nabla f(x_k)\|_2^2 \to 0$.

<u>However, we would like to know how fast</u>
<u>the gradient goes to zero.</u>

In particular, the termination criterion of gradient descent is $\|\nabla f(x_*)\|_2 \le \varepsilon$

for some positive constant $\varepsilon > 0$.

We would like to know : how many iterations of gradient descend are required to guarantee that $\|\nabla f(x_k)\|_2 \le \varepsilon$.

In other words, given a tolerance parameter ε, we would like to know how many iterations does it take to get $\|\nabla f(x_k)\|_2 \le \varepsilon$.

Assumptions

1) $\nabla f(x)$ is Lipschitz continuous
2) Step-size $a_k = \frac{1}{L}$ (simplifies the analysis)
3) Function f is bounded below.

$\exists \tilde{f}$ such that $\tilde{f} \le f(x) \ \forall x \in \mathbb{R}^n$.

example least-squares is at least 0.

proof: convergence rate

We proved that the guaranteed progress is

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2$$

Since we want to bound $\|\nabla f(x_k)\|_2^2$

let's rearrange as

$$\|\nabla f(x_k)\|_2^2 \leq 2L\left(f(x_k) - f(x_{k+1})\right) \quad \forall k$$

Let's sum-up the squared norms of all gradients up to iteration t.

$$\sum_{k=0}^{t} \|\nabla f(x_k)\|_2^2 \leq 2L \sum_{k=0}^{t} [f(x_k) - f(x_{k+1})]$$

The RHS is called the "telescoping sum"

$$\sum_{k=0}^{t} f(x_k) - f(x_{k-1}) = f(x_0) - f(x_1) + f(x_1) - f(x_2) + f(x_2) - \dots f(x_{t+1})$$

$$= f(x_0) - f(x_{t+1})$$

This gives us

$$\sum_{k=0}^{t} \|\nabla f(x_k)\|_2^2 \leq 2L\left(f(x_0) - f(x_{t+1})\right)$$

We can also simplify the LHS

$$(t+1)\min_{0\leq k\leq t} \|\nabla f(x_k)\|_2^2 \leq \sum_{k=0}^{t} \|\nabla f(x_k)\|_2^2$$

Thus we get

$$(t+1)\min_{0\leq k\leq t} \|\nabla f(x_k)\|_2^2 \leq 2L(f(x_0) - f(x_{t+1}))$$

Let's use the fact that f is bounded below

$$\tilde{f} \leq f(x) \quad \forall x \in \mathbb{R}^n$$

to get

$$(t+1) \min_{0\leq k\leq t} \|\nabla f(x_k)\|_2^2 \leq 2L(f(x_0) - \tilde{f})$$

Now divide by $t+1$ to get

$$\min_{0 \le k \le t} \|\nabla f(x_k)\|_2^2 \le \frac{2L(f(x_0) - \tilde{f})}{t+1}$$

$$= O\left(\frac{1}{t}\right) \quad \text{(convergence rate)}$$

Thus after t iterations we have that $\exists$ at least one x_k such that

$$\|\nabla f(x_k)\|_2^2 = O\left(\frac{1}{t}\right)$$

How many iterations will it take to

$$\min_{0 \le k \le t} \|\nabla f(x_k)\|_2^2 \le \varepsilon$$

We need

$$\frac{2L(f(x_0) - \tilde{f})}{t+1} \le \varepsilon$$

$$\Leftrightarrow \quad \boxed{\frac{2L(f(x_0) - \tilde{f})}{\varepsilon} - 1 \le t}$$

(11)

This means that after $\frac{2L(f(x_0)-\tilde{f})}{\varepsilon}-1$ iterations gradient descent is guaranteed to produce at least one x_k such that

$$\|\nabla f(x_k)\|_2^2 \le \varepsilon .$$

Comments

1) Similar result can be shown when using the line-search techniques to compute the step-size α_k. Only some constants change.

2) The rate $O\left(\frac{1}{t}\right)$ is dimension independent. (assuming that L is a constant)

3) We showed that after iterations t, $\exists\, 0\le k\le t$ such that

$$\min_{0\le k\le t} \|\nabla f(x_k)\|_2^2 \le \frac{2L(f(x_0)-\tilde{f})}{t+1}$$

It is not necessary that the last iteration t achieves this bound. Since this is a worst-case result earlier iterations might satisfy this bound too.

4) For ML problems bounds like

$$\min_{0 \le k \le t} \|\nabla f(x_k)\|^2 \le \frac{2L(f(x_0) - \hat{f})}{t+1}$$

are often very loose.

In practise gradient descent might be much faster.

There is a practical and theoretical component to understanding how gradient descent works.

5) Since our function is not necessarily convex, gradient descent is only guaranteed to converge to a stationary point.

Convergence rate for convex functions

Assumptions

1) ~~$\nabla f(x)$~~ f is differentiable

2) $\nabla f(x)$ is Lipschitz continuous

3) f is convex

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \quad \forall x, y \in \mathbb{R}^n$$

We will first need the following Lemma

Lemma: If f is convex & ∇f is Lipschitz continuous then

$$f(y) - f(x) \leq \nabla f(y)^T (y-x) - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$$

$$(\nabla f(y) - \nabla f(x))^T (y-x) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2$$

proof:

Let's start by proving the first inequality.

$$f(y) - f(x) = f(y) - f(z) + f(z) - f(x) \quad (i)$$

Using convexity of f:

$$f(z) - f(y) \geq \nabla f(y)^T (z - y)$$

$$\Rightarrow$$

$$f(y) - f(z) \leq \nabla f(y)^T (y - z) \quad (ii)$$

we get that (i) + (ii) gives

$$f(y) - f(x) \leq \nabla f(y)^T (y - z) + f(z) - f(x) \quad (iii)$$

using $f(z) - f(x) \leq \nabla f(x)^T (z - x) + \frac{L}{2} \| z - x \|_2^2$

(Corollary of FT.C) (iv) ~~(4)~~

we get that combining (iii) + (iv) we have

$$f(y) - f(x) \leq \nabla f(y)^T (y - z) + \nabla f(x)^T (z - x) + \frac{L}{2} \| z - x \|_2^2$$

Minimizing w.r.t z in the RHS

(Note RHS is convex) we get that the

minimizer is $\quad z = x - \frac{1}{L}\left(\nabla f(x) - \nabla f(y)\right)$

(we minimize because we want to get the smallest possible RHS)

Replacing $\quad z = x - \frac{1}{L}\left(\nabla f(x) - \nabla f(y)\right)$ in RHS

we get

$$
\begin{aligned}
f(y) - f(x) &= \nabla f(y)^T\left(y - x + \frac{1}{L}\left(\nabla f(x) - \nabla f(y)\right)\right) \\
&\quad + \nabla f(x)^T\left(x - \frac{1}{L}\left(\nabla f(x) - \nabla f(y)\right) - x\right) \\
&\quad + \frac{L}{2}\left\| x - \frac{1}{L}\left(\nabla f(x) - \nabla f(y)\right) - x\right\|_2^2 \\
&= \nabla f(y)^T (y - x) + \frac{1}{L}\nabla f(y)^T\left(\nabla f(x) - \nabla f(y)\right) \\
&\quad - \frac{1}{L}\nabla f(x)^T\left(\nabla f(x) - \nabla f(y)\right) \\
&\quad + \frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|_2^2
\end{aligned}
$$

$$= \nabla f(y)^T (y-x) - \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

$$+ \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

$$= \nabla f(y)^T (y-x) - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y \in \mathbb{R}^n$$

which proves the first inequality.

The second inequality follows from applying the first inequality and interchanging the roles of x & y to get

$$f(x) - f(y) \leq \nabla f(x)^T (x-y) - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$$

Adding together

$$f(x) - f(y) \leq \nabla f(x)^T (x-y) - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$$

and

$$f(y) - f(x) \leq \nabla f(y)^T (y-x) - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

we get

$$0 \leq (\nabla f(y) - \nabla f(x))^T (y-x) - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$$

which proves the second inequality.

Theorem: let f be convex and differentiable and $\nabla f(x)$ is Lipschitz continuous.

Let x_k for $k = 0 \ldots t$ be the sequence of iterates generated by gradient descent.

It follows that

$$f(x_k) - f(x^*) \le \frac{2L\|x_0 - x^*\|_2^2}{t+1}$$

proof:

$$\|x_{k+1} - x^*\|_2^2 = \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|_2^2$$

$$= \left(x_k - x^* - \frac{1}{L}\nabla f(x_k)\right)^T \left(x_k - x^* - \frac{1}{L}\nabla f(x_k)\right)$$

$$= \|x_k - x^*\|_2^2 - \frac{2}{L}(x_k - x^*)^T \nabla f(x_k) + \frac{1}{L^2}\|\nabla f(x_k)\|_2^2 \quad (i)$$

~~Note that using convexity we have~~

$$\cancel{f(x^*) \ge f(x_k) + \nabla f(x_k)^T (x_k - x^*)}$$

$$\Rightarrow \cancel{f(x^*) - f(x_k) \ge \nabla f(x_k)^T (x_k - x^*)} \Rightarrow$$

using the second inequality of our Lemma with $y = x$ & $x = x^*$ we get

$$(\nabla f(x)^T - \nabla f(x^*))^T (x - x^*) \geq \frac{1}{L} \|\nabla f(x^*) - \nabla f(x)\|_2^2$$

But $\nabla f(x^*) = 0$ thus the above simplifies to

$$\nabla f(x)^T (x - x^*) \geq \frac{1}{L} \|\nabla f(x)\|_2^2$$

Setting $x = x_k$ we get

$$\nabla f(x_k)^T (x_k - x^*) \geq \frac{1}{L} \|\nabla f(x_k)\|_2^2 \quad \text{(ii)}$$

Combining (i) & (ii) we get

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - \frac{2}{L^2} \|\nabla f(x_k)\|_2^2 + \frac{1}{L^2} \|\nabla f(x_k)\|_2^2$$

$$= \|x_k - x^*\|_2^2 - \frac{1}{L^2} \|\nabla f(x_k)\|_2^2$$

calling upon the amount of decrease of f from gradient descent we have that

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2$$

Adding and subtracting $f(x^*)$ we get

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$$

Using convexity we get

$$f(x_k) - f(x^*) \leq \nabla f(x_k)^T (x_k - x^*)$$

$$\leq \|\nabla f(x_k)\|_2 \|x_k - x^*\|_2 \quad \text{(iii)}$$

Note that

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - \frac{1}{L^2}\|\nabla f(x_k)\|_2^2$$

implies that distance to x^* is decreasing at each iteration. Thus

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_0 - x^*\|_2^2$$

Using this in (iii) we get

$$f(x_k) - f(x^*) \leq \|\nabla f(x_k)\| \cdot \|x_0 - x^*\|_2$$

$$\Leftrightarrow$$

$$\frac{f(x_k) - f(x^*)}{\|x_0 - x^*\|_2} \leq \|\nabla f(x_k)\|_2 \quad \text{(iv)}$$

Using (iv) in $f(x_{k+1}) - f(x^*) \le f(x_k) - f(x^*) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$ (20)

we get

$$f(x_{k+1}) - f(x^*) \le f(x_k) - f(x^*) - \frac{(f(x_k)-f(x^*))^2}{\|x_0 - x^*\|^2} \cdot \frac{1}{2L}$$

Set $\beta = \frac{1}{2L} \frac{1}{\|x_0 - x^*\|^2}$

$$\delta_k = f(x_k) - f(x^*)$$

Then the last inequality becomes

$$\delta_{k+1} \le \delta_k - \beta \delta_k^2$$

multiply by $\frac{1}{\delta_k \delta_{k+1}}$ to get

$$\frac{1}{\delta_k} \le \frac{1}{\delta_{k+1}} - \beta \frac{\delta_k}{\delta_{k+1}}$$

$\Rightarrow$

$$\beta \frac{\delta_k}{\delta_{k+1}} \le \frac{1}{\delta_{k+1}} - \frac{1}{\delta_k}$$

since $\delta_{k+1} \le \delta_k$ we get $\beta \le \frac{1}{\delta_{k+1}} - \frac{1}{\delta_k}$

$$\sum_{k=0}^{t} B \le \sum_{k=0}^{t} \frac{1}{\delta_{k+1}} - \frac{1}{\delta_k}$$

$$= \frac{1}{\delta_{t+1}} - \frac{1}{\delta_0} \le \frac{1}{\delta_{t+1}}$$

$$\Rightarrow$$

$$(t+1)\, B \le \frac{1}{\delta_{t+1}}$$

$$\Rightarrow$$

$$f(x_{t+1}) - f(x^*) \le \frac{2L\, \|x_0 - x^*\|_2^2}{t+1}$$

which proves the final result.

Strong Convexity

We can "strengthen" the notion of convexity by defining μ-strong convexity:

That is any function f that satisfies

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|_2^2$$

$$x, y \in \mathbb{R}^n$$

Note that this definition of strong convexity requires f to be differentiable. There are definitions of strong convexity that do not require differentiability. However, we will focus on the above definition for this lecture.

Lemma: If f is μ-strongly convex, then it also satisfies the Polyak-Lojasievicz condition, that is

$$\|\nabla f(x)\|_2^2 \geq 2\mu \left(f(x) - f(x^*) \right)$$

where x^* is the minimizer of f.

proof: Multiply the definition of strong convexity by -1 to get

$$-f(y) \le -f(x) - \nabla f(x)^T (y-x) - \frac{\mu}{2}\|y-x\|_2^2$$

set $y = x^*$ to get

$$f(x) - f(x^*) \le \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x - x^*\|_2^2$$

Complete the square in RHS

$$\nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x - x^*\|_2^2 =$$

$$\frac{1}{\mu}\|\nabla f(x)\|_2^2 - \frac{1}{\mu}\|\nabla f(x)\|_2^2 + \nabla f(x)^T (x - x^*) - \frac{\mu}{2}\|x - x^*\|_2^2$$

$$= -\frac{1}{2}\left\| \sqrt{\mu}\,(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\right\|_2^2 + \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

$$\le \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

which gives that

$$f(x) - f(x^*) \le \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

Theorem: If a function f is strongly convex then it has a unique minimizer.

proof: Let's assume that there exist two unique minimizers x_1^* & x_2^* such that

$$x_1^* \neq x_2^* \quad \& \quad f(x_1^*) = f(x_2^*)$$

From the definition of strong convexity we have that

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|_2^2$$

$$\forall x, y \in \mathbb{R}^n$$

Set $x = x_1^*$ to get

$$f(y) \geq f(x_1^*) + \nabla f(x_1^*)^T (y - x_1^*) + \frac{\mu}{2} \|y - x_1^*\|_2^2$$

Because x_1^* is a minimizer we have that

$$\nabla f(x_1^*) = 0$$

Thus we get

$$f(y) \geq f(x_1^*) + \frac{\mu}{2} \|y - x_1^*\|_2^2 \quad \forall y \in \mathbb{R}^n$$

Set $y = x_2^*$ to get

$$f(x_2^*) \geq f(x_1^*) + \frac{\mu}{2}\|x_2^* - x_1^*\|_2^2$$

Since $x_1^* \neq x_2^*$ then $\|x_2^* - x_1^*\|_2^2 > 0$

Thus

$$f(x_2^*) > f(x_1^*) \implies \text{contradiction}$$

because $f(x_2^*) = f(x_1^*)$ since both x_1^* & x_2^* are minimizers of a convex function.