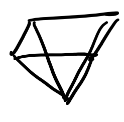
L-Smooth: No Conexity Needed

2024年10月22日 星期二 15:43

Can we do something better

We made assumptions that

1. f is L-Lip
2. $\|x_0 - x^*\| \le R$
3. f convex, diff



No real Hessian,
The ball doesn't pickup momentum

$$\Rightarrow f\left(\frac{1}{T}\sum_{s=0}^{T-1} x_s\right) - f(x^*) \le \frac{RL}{\sqrt{T}} \quad \text{where} \quad \mu = \frac{R}{L\sqrt{T}}$$

L-Smooth Definition

Def: a function $f: \mathbb{R}^n \to \mathbb{R}$ is L-Smooth (stronger than L-Lip)
if gradient is L-Lip, bounding the gradient instead of bounding function

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$$

Much nicer statement is given now

1. Gradient L-Lip
2. Hessian can not be crazy

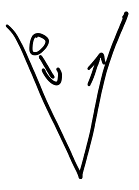
L-Smooth Bounds on Gradient

Thm: If f is L-smooth and twice differentiable (Hessian exist)
then $V^T \nabla^2 f(x) V \le L$, $\forall x \in \Omega$, $V \in \mathbb{R}^n$, $\|V\| = 1$
(upperbounded by L)

No this graph

$$0 \le V^T \nabla^2 f(x) V \le L$$

Min and max both bounded



Can't be this

$\|\nabla^2 f(x)\|_2$
matrix norm: How much it can be stretched

L-Smooth gives Stronger Convergence

Thm: If $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth and convex, and $0 \le \mu \le \frac{1}{L}$ (bound μ)
This binding is saying that when hessian big, L big, take smaller upperbound steps

Then GD satisfies $f(x^{(t)}) - f(x^*) \le \frac{1}{2t\mu}\|x^{(0)} - x^*\|^2$

Remark: $\|x^{(0)} - x^*\| \le R$, $\mu = \frac{1}{L}$ $\Rightarrow$ $f(x^{(t)}) - f(x^*) \le \frac{RL}{2t}$ ← much stronger change at each step

Don't need pre-know how many step in total

Example
$\|x^{(0)} - x^*\| \le 10$
$L = 2 \quad \mu = 1/2 \Rightarrow t = 10{,}000$
$f(x^{(t)}) - f(x^*) \le \frac{1}{1000} = 0.001$ way better bound

Remark: Didn't assume twice differentiable

Pf strategy: $f(y) \le f(x) + \nabla f^T(x)(y-x) + \frac{1}{2}(y-x)^T \overset{\text{Bounded by } L}{(LI)} (y-x)$

$\uparrow$ $\frac{L}{2}\|y - x\|^2$

This shows that even once differentiable, we can prove this theorem

How do we Converge

Does gradient blows up, very small, blows up, ... How smooth does the convergence come

Thm: Let $f: \mathbb{R}^n \to \mathbb{R}$, L-smooth, For any $0 \le \mu \le \frac{1}{L}$,
each step of GD gives

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

It definately goes down, at least at the previous gradient or smaller (converges)

Proof

$$f(x^{(t+1)}) \le f(x^{(t)}) + \nabla f(x^{(t)})^T (x^{(t+1)} - x^{(t)}) + \frac{L}{2}\|x^{(t+1)} - x^{(t)}\|^2$$

GD tells: $x^{(t+1)} = x^{(t)} - \mu \nabla f(x^{(t)}) \longrightarrow \boxed{x^{(t+1)} - x^{(t)} = -\mu \nabla f(x^{(t)})}$

$$\Rightarrow f(x^{(t+1)}) \le f(x^{(t)}) - \nabla f(x^{(t)})^T \mu \nabla f(x^{(t)}) + \frac{L}{2}\mu^2\|\nabla f(x^{(t)})\|^2$$

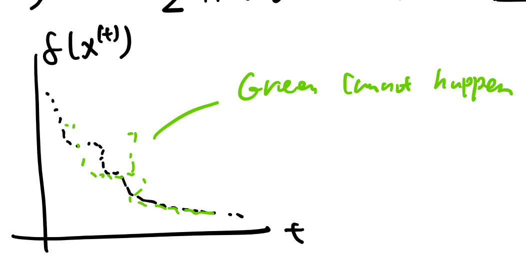
$\mu\|\nabla f(x^{(t)})\|^2$

$$\Rightarrow f(x^{(t+1)}) \le f(x^{(t)}) - \mu\left(1 - \frac{L}{2}\mu\right)\|\nabla f(x^{(t)})\|^2$$

$\boxed{\text{when } \mu \le 1/L \qquad \ge 1/2}$

$$\Rightarrow f(x^{(t+1)}) \le f(x^{(t)}) - \frac{\mu}{2}\|\nabla f(x^{(t)})\|^2 \quad \square$$

If we think about



How Big is the number gonna be

Thm: $f: \mathbb{R}^n \to \mathbb{R}$ L-smooth, $0 \le \mu \le \frac{1}{L}$
Then GD for T iteration means at least at one x_t must satisfy

$$\|\nabla f(x^{(t)})\| \le \sqrt{\frac{2[f(x^{(0)}) - f(x^*)]}{\mu T}}$$

At some point, we must reach a small gradient, gradient always get smaller

Proof

Assume f is lower bounded (not $-\infty$ $f(x^*)$)

$$\boxed{f(x^{(T)}) - f(x^{(0)})} = \sum_{t=0}^{T-1}\left[f(x^{(t+1)}) - f(x^{(t)})\right] \quad \text{Telescoping Theorem}$$

$-\frac{\mu}{2}\|\nabla f(x^{(t)})\|^2$ proved earlier

$$f(x^{(T)}) - f(x^{(0)}) \le -\sum_{t=0}^{T-1}\frac{\mu}{2}\|\nabla f(x^{(t)})\|^2$$

$$\sum_{t=0}^{T-1}\|\nabla f(x^{(t)})\|^2 \le \frac{2}{\mu}\left(f(x^{(0)}) - f(x^{(T)})\right)$$

Since $f(x^{(T)}) \ge f(x^*)$, so can change to $f(x^*)$ maintains the relationship

$$\sum_{t=0}^{T-1}\|\nabla f(x^{(t)})\|^2 \le \frac{2}{\mu}\left(f(x^{(0)}) - f(x^*)\right)$$

We say that there exists at least one $x^{(t)}$ such that
$\|\nabla f(x^{(t)})\|^2 \le \frac{2}{\mu T}\left(f(x^{(0)}) - f(x^*)\right)$

$\sum_{t=0}^{T-1}\|\nabla f(x^{(t)})\|^2$ is a converging series and for a converging series, averaging T terms, one of them must be smaller than or equal to T

$$\Rightarrow \|\nabla f(x^{(t)})\| \le \sqrt{\frac{2(f(x^{(0)}) - f(x^*))}{\mu T}} \quad \square$$

No Convexity Required Remarks

Last 2 theorems of how we converge and how big we converge
Never assumed Convexity!!! Things we care about convex function will transfer to non-convex
Just can't say global minimum for sure, but it will work, function value will go down, and it will stop!!!