L-Lip: Convergence Guaranteed

2024年10月17日 星期四 09:41

L-Lip is a property of function

How many iter should we do GD : $f(x^t) - f(x^*) = 0$

Def: A function $f: \Omega \to \mathbb{R}$ where $\Omega \in \mathbb{R}^n$ is L-Lipschitz if $\forall x, y \in \Omega$

$$\underbrace{|f(x) - f(y)|}_{\text{dist between } f \text{ value}} \le L \underbrace{\|x - y\|}_{\text{dist between points}}$$

L that is not ∞, minimal one is the true description ($\uparrow L$, the less tight the theorem)

Bound on the derivatives, this holds in differential geometry, if you step one step in x, the function only step by 1 or less

Ex: $f(x) = |x|$ $\quad ||x| - |y|| \le |x - y|$ $\quad f$ is 1-Lip $\quad$ ⊖ ⊖ ⊖ ⊕ / ⊕ ⊕ ⊕ ⊖

When we say L-Lipschitz, we say the minimum Lip

Ex: $f(x) = x^2$ $\quad |x^2 - y^2| \le |x+y|\,|x-y| \quad \forall x, y \in \mathbb{R}$

Unbounded : x^2 is not Lip $\quad$ (constant depends on x, y)

L-Lipschitz Bound $\nabla f(x)$

Lemma: If f is L-Lip, Convex, and diff, then the $\|\nabla f(x)\| \le L \quad \forall x$

(techniqually doesn't need to be here)

Pf: $\forall x, y$, $f(x) - f(y) \ge \nabla f(x)^T (y - x)$ $\quad$ Convexity definition (function always $\ge$ Tangent Plain)

$$L\|x - y\| \underset{\text{L-Lip}}{\ge} |f(x) - f(y)| \underset{\text{convex}}{\ge} |\nabla f(x)^T (y - x)|$$

Pick $y = x + \nabla f(x)$

$$L\|\nabla f(x)\| \ge \dots \ge \underbrace{|\nabla f(x)^T \nabla f(x)|}_{\|\nabla f(x)\|^2} \Rightarrow \|\nabla f(x)\| \le L$$

$L = \|\nabla f(x)\|$ is the min L

L-Lip is the least complex optimization

Reverse is True

Theorem: $\|\nabla f(x)\| \le L \quad \forall x \in \Omega$, then f is L-Lip

Proof : Taylor Theorem

$$f(x) - f(y) = \nabla f(tx + (1-t)y)^T (x - y) \quad \text{for some } t \in (0,1)$$

$$|f(x) - f(y)| = |\nabla f(tx + (1-t)y)^T (x - y)|$$

Cauchy-Schwartz : $|W^T z| = |\,\|W\|\,\|z\| \cos\theta\,| \le \|W\|\,\|z\|$

L-Lip : $|f(x) - f(y)| \le \|\nabla f(\dots)\|\,\|x - y\|$
$\le L\|x - y\|$

Convergence Garantee

Thm: Let f be Convex, diff, and L-Lip

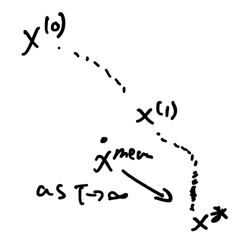
1. assume $\|x^{(0)} - x^*\| < R \longrightarrow$ quality of first guess (R is abstract, can be whole domain)
2. fix num of iteration of GD at T
3. Choose learning rate that is $\mu = \frac{R}{L\sqrt{T}} \longrightarrow$ Fixed learning rate of all

Does it garantee to go to zero? Do we know if zero exist at all?

Then $\underbrace{f\left(\frac{1}{T}\sum_{s=0}^{T-1} x^{(s)}\right)}_{\text{function value at the average of iteration}} - f(x^*) \le \frac{RL}{\sqrt{T}}$ $\quad$ Error

Convergence Rate: How fast does function $\to$ zero as a function of T

Average error on the order of $\frac{1}{\sqrt{T}}$



Ex: f is 2-Lip, $x^{(0)}$ is $\|x^{(0)} - x^*\| \le 10$, $T = 10{,}000$

Then, $\mu = 10/200 = 1/20$

The avg of all iteration gives that $f\left(\frac{1}{T}\sum_{s=0}^{T-1} x^{(s)}\right) - f(x^*) \le \frac{20}{\sqrt{10{,}000}} = \frac{2}{10} = \frac{1}{5}$

(L can be higher, but it would just not be tight bond)

Proof

B/c f is Convex : $\boxed{f(x^*) - f(x^{(s)}) \ge \nabla f(x^{(s)})^T (x^* - x^{(s)})}$ ① $\quad$ Any other point beneath Tangent Plain

B/c $\boxed{\begin{aligned} x^{(s+1)} &= x^{(s)} - \mu \nabla f(x^{(s)}) \\ \nabla f(x^{(s)}) &= \frac{x^{(s)} - x^{(s+1)}}{\mu} \end{aligned}}$ ②

From ① and ② : $f(x^*) - f(x^{(s)}) \ge \frac{1}{\mu} \langle x^{(s)} - x^{(s+1)}, x^* - x^{(s)} \rangle$

multiply by -1

$$\boxed{\underbrace{f(x^{(s)}) - f(x^*)}_{\text{where you are}} \le \frac{1}{\mu} \langle x^{(s)} - x^{(s+1)}, \underbrace{x^{(s)} - x^*}_{\text{How far from } x^*} \rangle} \quad ③$$

These 2 are pointing at opposite direction, equation flipped sustained

$$\boxed{\text{from LA: } \langle a, b \rangle = \frac{\|a\|^2 + \|b\|^2 - \|a - b\|^2}{2}}$$

$$f(x^{(s)}) - f(x^*) \le \frac{1}{2\mu}\Big(\underbrace{\|x^{(s)} - x^{(s+1)}\|^2}_{\|a\|^2} + \underbrace{\|x^{(s)} - x^*\|^2}_{\|b\|^2} - \underbrace{\|x^{(s+1)} - x^*\|}_{\|b - a\|^2}\Big)$$

the same

$$\boxed{\text{From GD: } x^{(s)} - x^{(s+1)} = \mu \nabla f(x^{(s)})} \quad ④$$

$x^{(s)} - x^{(s+1)}$ cancle out $\quad \left(\frac{\mu^2}{2\mu} = \frac{\mu}{2}\right)$

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

s term $\quad$ $s+1$ term

⑤

Telescoping Theory : Sum from $0 \to T-1$, there are T equation here, all middle ones cancle, only 1st and last term exist

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

R^2 since it's quality of guess $\quad$ ≥ 0 value (don't care) $\quad$ L^2 since $\|\nabla f(x^{(s)})\| \le L$ for L-Lip f

$$\dots \le \frac{R^2}{2\mu} + \frac{\mu}{2} T L^2$$

$$\frac{1}{T}\sum_{s=0}^{T-1}\left(f(x^{(s)}) - \underbrace{f(x^*)}_{\text{no } (s)}\right) \le \frac{1}{2\mu T} R^2 + \frac{\mu}{2} L^2$$

$$\boxed{\text{Convex } f \Rightarrow f\left(\frac{1}{T}\sum_{s}^{T-1} x^{(s)}\right) \le \frac{1}{T}\sum_{s=0}^{T-1} f(x^{(s)})} \quad ⑥$$

Convexity for a bunch of point, Proof by induction: 1✓, T-1✓, T✓

$$f\left(\frac{1}{T}\sum_{s=0}^{T-1} x^{(s)}\right) - f(x^*) \le \frac{1}{2\mu T} R^2 + \frac{\mu}{2} L^2$$

Choose a min μ, $\mu = \frac{R}{L\sqrt{T}}$ Tightest (Choosing a μ so 2 terms are equal to each other)

Then $\quad f\left(\frac{1}{T}\sum_{s=0}^{T-1} x^{(s)}\right) - f(x^*) \le \frac{RL}{\sqrt{T}}$

GD Converges and we know it for sure it works in T iteration when initial guess is in R and the function is Convex, differentiable, and L-Lip