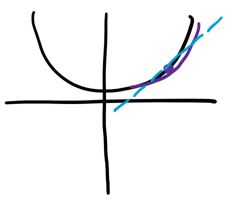
Strongly Convex

2024年11月19日 星期二 09:44

Strong Convexity

A function $f\ \mathbb{R}^n \to \mathbb{R}$ is strongly convex if $\exists\ c > 0$ s.t. $x, y \in \mathbb{R}^n$

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{c}{2}\|y-x\|^2$$



Instead of tangent line, we actually have a tangent curve

This derive into this

Size of the gradient tells you how close you are to the x^* when your condition is under strongly convex, and the relation is squared

$$\Rightarrow \quad f(x) - f(x^*) \le \frac{\|\nabla f(x)\|^2}{2c} \qquad \text{(proved in hw)}$$

← This is a property of the function, it is given to you

Exponential Convergence Rate

We will show that Strong Convexity converges super quick

1. GD: $x^{(t+1)} = x^{(t)} - \mu_t \nabla f(x^{(t)})$
2. Assuming L-smooth, bounds gradient ($\|\nabla^2 f(x)\| \le L$)

$$f(x^{(t+1)}) \le f(x^{(t)} - \underbrace{\mu_t f(x^{(t)})}_{h})$$

Taylor theory

$$f(x^{(t)}) - h^T \nabla f(x^{(t)}) + \underbrace{\frac{1}{2} h^T \nabla^2 f(\xi) h}$$

$\le L\|h\|^2$

Cauchy-Sharwz : $\langle h, \nabla^2 f(\xi) h \rangle \le \|h\|\,\|\nabla^2 f(\xi) h\|$

$\le \|h\|\,\|\nabla^2 f(\xi)\|\,\|h\|$

Biggest the hessian can be : L

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

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

Now nice property show up : $f(x^{(t+1)}) \le f(x^{(t)}) - \mu_t\left(1 - \frac{L\mu_t}{2}\right)\|\nabla f(x^{(t)})\|^2$

Pick $\mu_t \le \frac{1}{L}$ (L-smooth)

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

2 Directions now

A: No Assumption of Strong Convexity

Telescoping Series : $\frac{\mu_t}{2}\sum_{t=0}^{T-1}\|\nabla f(x^{(t)})\|^2 \le \sum_{t=0}^{T-1}\left[f(x^{(t)}) - f(x^{(t+1)})\right]$

$= f(x^{(0)}) - f(x^T)$

$\le f(x^{(0)}) - f(x^*)$

As $T \to \infty$, $\|\nabla f(x^{(t)})\|^2$ must $\to 0$

Done this before

B: Assuming Strong Convexity

$$\|\nabla f(z)\|^2 \ge 2c\left[f(z) - f(x^*)\right]$$

Then $f(x^{(t+1)}) \le f(x^{(t)}) - \frac{\mu_t}{2}\left[2c\left[f(x^{(t)}) - f(x^*)\right]\right]$ Do $-f(x^*)$ on both side

$$f(x^{(t+1)}) - f(x^*) \le (1 - \mu_t c)\left(f(x^{(t)}) - f(x^*)\right)$$

Assume $\mu_t \le \frac{1}{L}$ $\quad \le (1 - c/L)\left(f(x^{(t)}) - f(x^*)\right)$

$$f(x^{(t+1)}) - f(x^*) \le \left(1 - \frac{c}{L}\right)\left(f(x^{(t)}) - f(x^*)\right)$$

$$\le \left(1 - \frac{c}{L}\right)^t \left(f(x^{(0)}) - f(x^*)\right)$$

Chase all the way back to $x^{(0)}$

This is true with GD + ① Strong Convexity

② L-Smooth $\|\nabla^2 f(x)\| \le L$

③ $\mu_t = 1/L$

This is a really really fast exponential convergence rate

Generalization by Squeeze

When our function is quadratic $f(x) = \frac{1}{2}x^T A x$

$\frac{c}{L} = \frac{1}{K}$ ← Condition #

Smallest Eigenvalue of A, direction of least convexity

Longest Eigenvalue of $\nabla^2 f(x)$ (A) , Can't stretch it more than L direction of max convexity

If $\frac{c}{L} = 1$, It's saying GD of the quadratic with identity matrix

Closer c to L, easier

Condition # for general strongly convex function :

1. L-Smooth

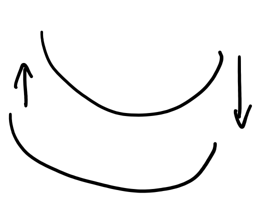
$\|\nabla^2 f(x)\| \le L$

Can't be "too convex"

Bounded above, Can't be too bended

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

$$\le \underbrace{f(x) + \nabla f(x)^T h + \frac{L}{2}\|h\|^2}_{\to Q_L(h)}$$



Constant Linear Quadratic in h

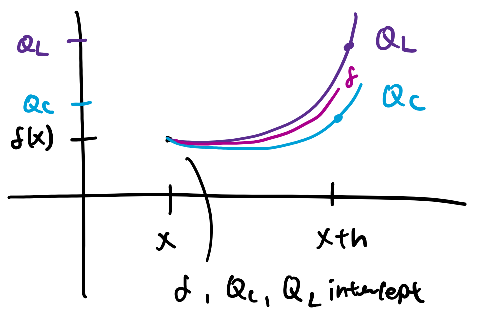
(This is similar notion to Condition # K)

2. Strongly Convex Part

$$f(x+h) \ge \underbrace{f(x) + \nabla f(x)^T h + \frac{c}{2}\|h\|^2}_{Q_c(h)}$$

Must be "Somewhat Convex"

Can't be too flat



function is stuck between Q_c and Q_L

→ Squeeze theorem

Note: $f(x)$ doesn't need to quadratic any more,

as long as it satisfy ① and ②, it doesn't matter what f is,

we have auto quadratic bounding

At this sweet spot, GD converges exponentially fast

Under strongly convex, previous method all generalizes well

When $\frac{c}{L} \equiv K$ $\quad$ GD w/ momentum have conv rate of $\frac{\sqrt{K}-1}{\sqrt{K}+1}$ for strong convexity and $\|\nabla^2 f(x)\| \le L$

GD w/ N.A. have convergence rate of $\sqrt{\frac{\sqrt{K}-1}{\sqrt{K}}}$

Turns out the things we talked about quadratic function is true for function with strongly convexity

When removed strongly convexity, momentum would fail, but N.A. would still work

What ridge regression is doing is to artificially create a c and make a function strongly convex