

Convergence of GD for convex Functions

Theorem Consider the GD algorithm

$$x_{k+1} = x_k - \alpha \nabla f(x_k), \quad k=0, 1, \dots$$

Assume that f has Lipschitz gradient with Lipschitz constant L . Further assume that :

(a) f is a convex function

(b) $\exists x^* \text{ s.t. } f(x^*) = \min f(x)$

Then for sufficiently small α :

$$(i) \lim_{k \rightarrow \infty} f(x_k) = \min f(x) = f(x^*)$$

$$(ii) f(x_k) \rightarrow f(x^*) \text{ at rate } \frac{1}{k}$$

Proof Let x^* be a min.

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

$$= \|x_k - x^*\|^2 + \alpha^2 \|\nabla f(x_k)\|^2 - 2\alpha \nabla f(x_k)^T (x_k - x^*)$$

By convexity,

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

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

Thus,

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

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

Sum over $\sum_{k=0}^N$ on both sides

$$2\alpha \sum_{k=0}^N (f(x_k) - f(x^*)) \leq \|x_0 - x^*\|^2 - \|x_{N+1} - x^*\|^2 + \alpha^2 \sum_{k=0}^N \|\nabla f(x_k)\|^2$$

$$\leq \|x_0 - x^*\|^2 + \alpha^2 \sum_{k=0}^N \|\nabla f(x_k)\|^2$$

But from proof of previous result on convergence of steepest descent with constant step size, if $0 < \alpha < \frac{2}{L}$,

$$\sum_{k=0}^N \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - f(x^*)}{\alpha (1 - \frac{1}{2} L \alpha)}$$

Also, $f(x_{k+1}) - f(x_k) \leq -\alpha (1 - \frac{1}{2} L \alpha) \|\nabla f(x_k)\|^2 \leq 0$

$$\Rightarrow f(x_N) \leq f(x_k), \text{ for } k=0, 1, \dots, N.$$

$$\Rightarrow \sum_{k=0}^N (f(x_k) - f(x^*)) \geq (N+1) (f(x_N) - f(x^*))$$

i.e. $f(x_N) - f(x^*) \leq \frac{1}{N+1} \sum_{k=0}^N (f(x_k) - f(x^*))$

$$\leq \frac{1}{2\alpha(N+1)} \left(\|x_0 - x^*\|^2 + \frac{\alpha(f(x_0) - f(x^*))}{1 - \frac{1}{2} L \alpha} \right)$$

$$\rightarrow 0 \text{ as } N \rightarrow \infty$$

Rate of convergence is $\frac{1}{N}$.

To make $f(x_N) - f(x^*) < \varepsilon$, we need $N \sim O(\frac{1}{\varepsilon})$

- Can show that GD with Armijo's Rule also converges at rate $\frac{1}{N}$ if ∇f is Lipschitz, without prior knowledge of L . But need $\Gamma \in [\frac{1}{2}, 1)$.

Definition (Strong Convexity) A twice continuously differentiable function is strongly convex if

$$\exists m > 0 \text{ s.t. } \nabla^2 f(x) \succcurlyeq mI \quad \forall x.$$

Strong convexity \Rightarrow strict convexity

$$\begin{aligned} \text{Proof } \nabla^2 f(x) \succcurlyeq mI &\Rightarrow \nabla^2 f(x) - mI \succeq 0 \\ &\Rightarrow \forall z \neq 0 \quad z^T (\nabla^2 f(x) - mI) z \geq 0 \\ &\Rightarrow z^T \nabla^2 f(x) z \geq m z^T z > 0 \quad \blacksquare \end{aligned}$$

Converse is not true: e.g. $f(x) = x^4$ is strictly convex but $\nabla^2 f(0) = 0$

Lemma $\nabla^2 f(x) \succcurlyeq mI \quad \forall x$

$$\Rightarrow f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} m \|y-x\|^2$$

(i.e. f is lower bounded by a quadratic)

Proof By Taylor's Theorem,

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} (y-x)^T \nabla^2 f((1-\beta)x + \beta y) (y-x)$$

for some $\beta \in [0,1]$.

$$\begin{aligned} &\geq f(x) + \nabla f(x)^T (y-x) + \underbrace{\frac{1}{2} (y-x)^T m (y-x)}_{\frac{1}{2} m \|y-x\|^2} \end{aligned}$$

Strong convexity with parameter m , along with
M-Lipschitz gradient assumption (with $M \geq m$)

$$\Rightarrow \frac{1}{2}m\|y-x\|^2 \leq f(y) - f(x) - \nabla f(x)^T(y-x) \leq \frac{1}{2}M\|y-x\|^2$$

Rate of convergence of GD with above assumption

Consider steepest descent with fixed stepsize

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

Since f is (strictly) convex, and has Lipschitz gradient, we know from previous result that $\{x_k\}$ converges to (unique) minimum x^* at rate $\frac{1}{k}$ as $k \rightarrow \infty$ if $0 < \alpha < \frac{2}{M}$.

Now, what happens with strong convexity added?

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - \alpha \nabla f(x_k) - x^*\|^2 \\ &= \|(x_k - x^*) - \alpha (\nabla f(x_k) - \nabla f(x^*))\|^2 \\ &\stackrel{\nabla f(x^*) = 0}{=} \|x_k - x^*\|^2 + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2 \\ &\quad - 2\alpha (x_k - x^*)^T (\nabla f(x_k) - 0) \end{aligned}$$

$$\underbrace{\nabla f}_{M\text{-Lipschitz}} \leq \|x_k - x^*\|^2 + \alpha^2 M^2 \|x_k - x^*\|^2 + 2\alpha (x^* - x_k)^T \nabla f(x_k)$$

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \alpha^2 M^2 \|x_k - x^*\|^2 + 2\alpha (x^* - x_k)^T \nabla f(x_k)$$

By strong convexity of f

$$f(x^*) \geq f(x_k) + (x^* - x_k)^T \nabla f(x_k) + \frac{m}{2} \|x_k - x^*\|^2$$

$$\Rightarrow \|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 (1 + \alpha^2 M^2 - \alpha m) + 2\alpha (f(x^*) - f(x_k))$$

By strong convexity of f

$$f(x_k) \geq f(x^*) + \nabla f(x^*)^T (x_k - x^*) + \frac{m}{2} \|x_k - x^*\|^2$$

$$\Rightarrow f(x^*) - f(x_k) \leq -\frac{1}{2} m \|x_k - x^*\|^2$$

$$\text{Thus, } \|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 (1 + \alpha^2 M^2 - 2m\alpha)$$

$$\Rightarrow \|x_N - x^*\|^2 \leq \|x_0 - x^*\|^2 (1 + \alpha^2 M^2 - 2m\alpha)^N$$

If $\alpha > 0$ is s.t. $|1 + \alpha^2 M^2 - 2m\alpha| < 1$, then

$x_N \rightarrow x^*$ geometrically as $N \rightarrow \infty$.

Note that just having $0 < \alpha < \frac{2}{M}$ does not guarantee geometric convergence to x^* .

$$\text{e.g. } \alpha = \frac{1}{M} \Rightarrow 1 + \alpha^2 M^2 - 2m\alpha = \underbrace{2(1 - \frac{m}{M})}_{\geq 1 \text{ if } \frac{m}{M} \leq 0.5}$$

$$1 + \alpha^2 M^2 - 2m\alpha = (\alpha M)^2 - 2\alpha M \cdot \frac{m}{M} + 1$$

$$= \left(\alpha M - \frac{m}{M}\right)^2 + 1 - \frac{m^2}{M^2}$$

minimized by setting $\alpha = \alpha^* = \frac{m}{M^2}$

$$\min_{\alpha > 0} 1 + \alpha^2 M^2 - 2m\alpha = 1 - \frac{m^2}{M^2} \in [0, 1)$$

Note that $\alpha^* \leq \frac{1}{M} < \frac{2}{M}$

with $\alpha = \alpha^*$, $\|x_N - x^*\|^2 \leq \left(1 - \frac{m^2}{M^2}\right)^N \|x_0 - x^*\|^2$

$\frac{M}{m}$ is called the condition number

If $\frac{M}{m} \gg 1$, then $1 - \frac{m^2}{M^2}$ is close to 1

and convergence is slow.

If $\frac{M}{m} = 1$, $\alpha^* = \frac{1}{M}$, and $x_N = x^*, \forall N \geq 1$.

(convergence in one step)

Note that since $\nabla f(x^*) = 0$,

$$f(x_N) - f(x^*) \leq \frac{M}{2} \|x_N - x^*\|^2$$

$$\leq \left(1 - \frac{m^2}{M^2}\right)^N \frac{M}{2} \|x_0 - x^*\|^2$$

To make $f(x_N) - f(x^*) < \varepsilon$, we only need $N \sim O(\log \frac{1}{\varepsilon})$ - called "linear" convergence

Example $f(x) = \frac{1}{2} x^T Q x + b^T x + c, Q > 0$

$$\nabla^2 f(x) = Q$$

Let λ_{\min} and λ_{\max} be the min. and max. eigenvalues of Q . Then we know (from lec 2), for all $\vec{z} \in \mathbb{R}^n$,

$$\lambda_{\min} \|\vec{z}\|^2 \leq \vec{z}^T Q \vec{z} \leq \lambda_{\max} \|\vec{z}\|^2$$

Thus for all $\vec{z} \in \mathbb{R}^n$

$$\vec{z}^T (Q - \lambda_{\min} I) \vec{z} \geq 0$$

$$\Rightarrow Q - \lambda_{\min} I \succcurlyeq 0 \equiv Q \succcurlyeq \lambda_{\min} I$$

Similarly, \downarrow $Q \preccurlyeq \lambda_{\max} I$

Thus $\lambda_{\min} I \preccurlyeq \nabla^2 f(x) \preccurlyeq \lambda_{\max} I$

Condition number = $\frac{\lambda_{\max}}{\lambda_{\min}}$

Special Case $\mathbf{Q} = \mu \mathbf{I}$, $\mu > 0$.

$$\lambda_{\min} = \lambda_{\max} = \mu = m = M$$

$$f(\mathbf{x}) = \frac{\mu}{2} \|\mathbf{x}\|^2 + \mathbf{b}^T \mathbf{x} + c$$

$$\nabla f(\mathbf{x}) = \mu \mathbf{x} + \mathbf{b}$$

$$\mathbf{x}^* = -\frac{\mathbf{b}}{\mu}$$

$$\alpha^* = \frac{m}{M^2} = \frac{1}{\mu}$$

$$\begin{aligned}\mathbf{x}_1 &= \mathbf{x}_0 - \alpha^* \nabla f(\mathbf{x}_0) \\ &= \mathbf{x}_0 - \frac{1}{\mu} (\mu \mathbf{x}_0 + \mathbf{b}) \\ &= -\frac{\mathbf{b}}{\mu} = \mathbf{x}^*\end{aligned}$$

Convergence in one step!