Convergence of GD for convex Functions

Theorem Consider the GD algorithm

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

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

(a) f is a convex function

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

Then for sufficiently small α:

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

(ii) $f(x_N) \longrightarrow f(x^*)$ at rate $\frac{1}{N}$

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 stepsize, if $0 < \alpha < \frac{2}{L}$,

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

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

$\Rightarrow f(x_N) \leq f(x_k)$, for $k = 0, 1, \ldots, N$.

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

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

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

$$\to 0 \quad \text{as } N \to \infty$$

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

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

- 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 $\tau \in \left[\frac{1}{2}, 1\right)$.

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

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

Strong convexity $\Rightarrow$ strict convexity

Proof $\nabla^2 f(x) \succcurlyeq mI \Rightarrow \nabla^2 f(x) - mI \succcurlyeq 0$

$\Rightarrow \forall z \neq 0, \; z^T(\nabla^2 f(x) - mI) z \geq 0$

$\Rightarrow \forall z \neq 0, z^T \nabla^2 f(x) z \geq m z^T z > 0$ (strict convexity)

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

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

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

(second line is sometimes used to define strong convexity)

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

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

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

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

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 $\{x_N\}$ converges to (unique) minimum x^* at rate $\frac{1}{N}$ as $N \to \infty$ if $0 < \alpha < \frac{2}{L}$.

Now, what else does strong convexity give us?

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

$$= \|(x_k - x^*) - \alpha (\nabla f(x_k) - \nabla f(x^*))\|^2 \qquad \nabla f(x^*) = 0$$

$$= \|x_k - x^*\|^2 + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|^2 - 2\alpha (x_k - x^*)^T (\nabla f(x_k) - 0)$$

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

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

By strong convexity of f

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

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

By strong convexity of f

$$f(x_k) \geq f(x^*) + \nabla f(x^*)^T (x_k - x^*) + \frac{1}{2} m \|x_k - x^*\|^2 \quad (\nabla f(x^*) = 0)$$

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

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

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

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

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

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

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

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

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

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

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

Note that $\alpha^* = \frac{m}{L} \cdot \frac{1}{L} \leq \frac{1}{L} < \frac{1}{2L}$

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

$\frac{L}{m}$ is called the <u>condition number</u>

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

and convergence is slow.

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

(convergence in one step)

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

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

descent lemma

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

$r = \alpha^*$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\nabla f(x) = \mu x + b$$

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

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

$$x_1 = x_0 - \alpha^* \nabla f(x_0)$$

$$= x_0 - \frac{1}{\mu}(\mu x_0 + b)$$

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

Convergence in one step!