Convergence of GD with constant stepsize

Background: Lipschitz continuity

Definition A function $g: \mathbb{R}^n \to \mathbb{R}^m$ is called Lipschitz (continuous) if $\exists\ L > 0$ s.t.

$$\|g(y) - g(x)\| \leq L \|y - x\|, \quad \forall\ x, y \in \mathbb{R}^n$$

(L: Lipschitz constant)

Examples (see HW2)

1. Any $g: \mathbb{R} \to \mathbb{R}$ with bounded derivative is Lipschitz, e.g. $g(x) = x$, $g(x) = \sin x$
2. $g: \mathbb{R} \to \mathbb{R}$ does not have to be differentiable, e.g. $g(x) = |x|$ is Lipschitz

Proof: By Triangle inequality,

$|x| = |x - y + y| \leq |x - y| + |y|$, $|y| \leq |y - x| + |x|$

$\Rightarrow |x| - |y| \leq |x - y|$, $|y| - |x| \leq |x - y| \Rightarrow ||x| - |y|| \leq |x - y|$

3. $g: \mathbb{R}^n \to \mathbb{R}$, $g(x) = \|x\|$ is Lipschitz

How about $g(x) = \|x\|^2$?

Consider $y = 2x$. Then if g is Lipschitz we must have $|\|x\|^2 - 4\|x\|^2| \leq L \|x - 2x\|$, $\forall\ x \in \mathbb{R}^n$

$\Rightarrow \quad 3\|x\|^2 \leq L\|x\| \quad \forall\ x \in \mathbb{R}^n$

Does not hold for $\|x\| > \frac{L}{3}$

Lipschitz Gradient

Special case of $g : \mathbb{R}^n \to \mathbb{R}^n$ of interest in GD convergence analysis is $g(x) = \nabla f(x)$

$\nabla f(x)$ is Lipschitz if $\exists\ L > 0$ s.t.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall\ x, y$$

Examples

1. $f(x) = \|x\|^2$, $\nabla f(x) = 2x$ is Lipschitz with $L = 2$

2. $f(x) = \|x\|^4$

$$\nabla f(x) = 4\|x\|^2 x$$

Test Lipschitz condition with $y = -x$.
For ∇f to be Lipschitz, we must have

$$\|4\|x\|^2 x + 4\|x\|^2 x\| \leq L\|2x\| \quad \forall\ x \in \mathbb{R}^n$$

i.e. $8\|x\|^2\|x\| \leq 2L\|x\|$

Not possible if $\|x\|^2 > \frac{L}{4}$.

3. If f is twice continuously differentiable with $\nabla^2 f(x) \succcurlyeq -MI$ and $\nabla^2 f(x) \preccurlyeq MI$ then

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall\ x, y \in \mathbb{R}^n$$

Notation: $A \succcurlyeq B$ means $A - B \succcurlyeq 0$
$A \preccurlyeq B$ means $B - A \succcurlyeq 0$

$-MI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI \;\; \forall x \Rightarrow \|\nabla f(x) - \nabla f(y)\| \le M \|y-x\|$

Proof for symmetric A,

1. $x^T A x \le \lambda_{max}(A) \|x\|^2$
2. $\lambda_i(A^2) = \lambda_i^2(A)$
3. $-MI \preccurlyeq A \preccurlyeq MI \Rightarrow \lambda_{min}(A) \ge -M, \; \lambda_{max}(A) \le M$

Define $g(t) = \frac{\partial f}{\partial x_i}(x + t(y-x))$. Then

$$g(1) = g(0) + \int_0^1 g'(s)\, ds.$$

$$\Rightarrow \frac{\partial f}{\partial x_i}(y) = \frac{\partial f}{\partial x_i}(x) + \int_0^1 \sum_{j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(x + s(y-x))(y_j - x_j)\, ds$$

$$\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f(x + s(y-x))(y-x)\, ds$$

$$\|\nabla f(y) - \nabla f(x)\| = \left\| \int_0^1 \nabla^2 f(x + s(y-x))(y-x)\, ds \right\|$$

Triangle ineq. $\le \int_0^1 \|\nabla^2 f(x + s(y-x))(y-x)\|\, ds$

$$= \int_0^1 \sqrt{(y-x)^T [\nabla^2 f(x + s(y-x))]^2 (y-x)}\, ds$$

(H = $\nabla^2 f(x + s(y-x))$)

$$\le \int_0^1 \sqrt{\lambda_{max}(H^2)\, \|y-x\|^2}\, ds$$

$$\le M \|y-x\| \int_0^1 ds = M\|y-x\|$$

$\lambda_{max}(H^2) = \max\{\lambda_{max}^2(H), \lambda_{min}^2(H)\} \le M^2$

Descent Lemma Let $f: \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with a Lipschitz gradient with Lipschitz constant L. Then

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

Proof Let $g(t) = f(x + t(y-x))$.

Then $g(0) = f(x)$ and $g(1) = f(y)$

$$g(1) = g(0) + \int_0^1 g'(t)\,dt$$

and $g'(t) = \nabla f(x+t(y-x))^T (y-x)$

$$\Rightarrow f(y) = f(x) + \int_0^1 \nabla f(x+t(y-x))^T (y-x)\,dt$$

$$= f(x) + \int_0^1 (\nabla f(x+t(y-x)) - \nabla f(x))^T (y-x)\,dt + \int_0^1 \nabla f(x)^T (y-x)\,dt$$

$$\leq f(x) + \nabla f(x)^T (y-x) + \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\| \, \|y-x\|\,dt$$

(Cauchy-Schwarz)

$$\leq f(x) + \nabla f(x)^T (y-x) + L \int_0^1 \|t(y-x)\| \, \|y-x\|\,dt$$

$$= f(x) + \nabla f(x)^T (y-x) + L \|y-x\|^2 \underbrace{\int_0^1 t\,dt}_{=\frac{1}{2}}$$

Convergence of Steepest Descent with Fixed Stepsize

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. Then if α is sufficiently small and $f(x) \geq f_{min}$ for all $x \in \mathbb{R}^n$, every limit point of $\{x_k\}$ is a stationary point of f

Proof Applying The Descent Lemma,

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

$$= f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k) + \frac{1}{2} L \alpha^2 \|\nabla f(x_k)\|^2$$

$$= f(x_k) + \alpha \left(\frac{1}{2} L \alpha - 1\right) \|\nabla f(x_k)\|^2$$

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

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

$$\leq f(x_0) - f_{min}$$

If $0 < \alpha < \frac{2}{L}$, i.e. $\alpha\left(1 - \frac{1}{2} L \alpha\right) > 0$,

$$\sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - f_{min}}{\alpha\left(1 - \frac{1}{2} L \alpha\right)} < \infty, \quad \forall N$$

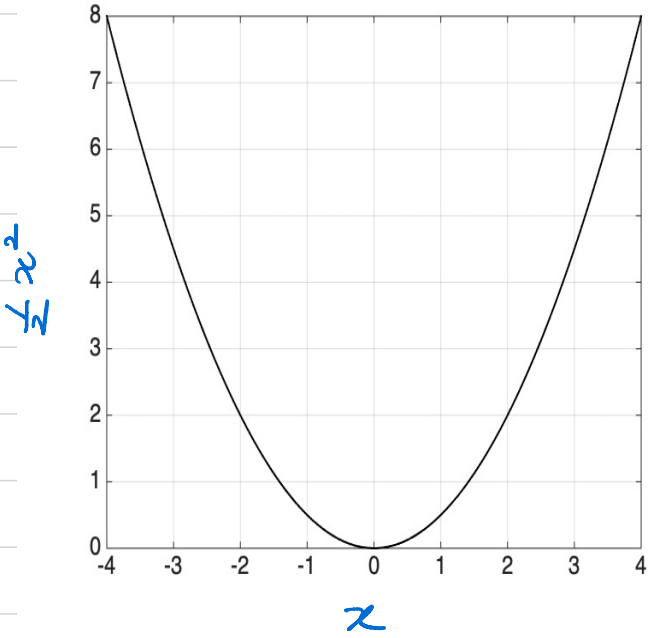
$$\Rightarrow \lim_{k \to \infty} \nabla f(x_k) = 0$$

If $\bar{x}$ is a limit point of $\{x_k\}$, $\lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} x_k = \bar{x}$

By continuity of ∇f, $\nabla f(\bar{x}) = 0$

Example

$f(x) = \frac{1}{2} x^2, \quad x \in \mathbb{R}$



$\nabla f(x) = x$

Lipschitz with $L = 1$

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

$= x_k (1 - \alpha)$

$0 < \alpha < \frac{2}{L} = 2$ needed for convergence

Test case 1: $\alpha = 1.5$. Then $x_{k+1} = x_k (-0.5)$

$\Rightarrow \quad x_k = x_0 (-0.5)^k \to 0$ as $k \to \infty$

Test case 2: $\alpha = 2.5$. Then $x_{k+1} = x_k (-1.5)$

$\Rightarrow \quad x_k = x_0 (-1.5)^k \Rightarrow |x_k| \to \infty$

Test case 3: $\alpha = 2$. Then $x_{k+1} = x_k (-1)$

$\Rightarrow \quad x_k = x_0 (-1)^k \Rightarrow$ oscillation between $-x_0, x_0$.

Example

What if gradient is not Lipschitz?

e.g. $f(x) = x^4, \quad x \in \mathbb{R}$

$\nabla f(x) = 4x^3$

$x = 0$ only stationary point (global min)

$x_{k+1} = x_k - 4\alpha x_k^3 = x_k (1 - 4\alpha x_k^2)$

$x_{k+1} = x_k(1 - 4\alpha x_k^2)$

- If $|x_1| = |x_0|$, then $|x_k| = |x_0|$ for all k, and $\{x_k\}$ stays bounded away from 0, except if $x_0 = 0$
- $|x_1| < |x_0| \iff |x_0|\,|1 - 4\alpha x_0^2| < |x_0|$

$\iff -1 < 1 - 4\alpha x_0^2 < 1$

$\iff 0 < x_0^2 < \frac{1}{2\alpha} \iff 0 < |x_0| < \frac{1}{\sqrt{2\alpha}}$

- Therefore, if $|x_1| < |x_0|$, then $|x_1| < |x_0| < \frac{1}{\sqrt{2\alpha}}$

$\Rightarrow |x_2| < |x_1|, \ldots |x_{k+1}| < |x_k|, \forall k \Rightarrow \{|x_k|\}$ converges.

- And if $|x_1| > |x_0|$, then $|x_{k+1}| > |x_k|$ for all k and $\{x_k\}$ stays bounded away from 0

Claim $0 < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_k| \to 0$

Proof Suppose $|x_k| \to c > 0$. Then

$$\frac{|x_{k+1}|}{|x_k|} \to 1$$

But $\frac{|x_{k+1}|}{|x_k|} = |1 - 4\alpha x_k^2| \to |1 - 4\alpha c^2|$

Thus $|1 - 4\alpha c^2| = 1 \Rightarrow c = \frac{1}{\sqrt{2\alpha}}$ — (1)

But $c < |x_0| < \frac{1}{\sqrt{2\alpha}}$ — (2)

(1) and (2) are contradictory

$\Rightarrow c = 0$.