

Convergence of GD with constant stepsizeBackground: Lipschitz Continuity

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

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

$\underbrace{}_{\text{Lipschitz constant}}$

Examples

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

Proof : By triangle inequality,

$$\begin{aligned} |x| &= |x-y+y| \leq |y| + |x-y| \Rightarrow |x|-|y| \leq |x-y| \\ |y| &= |y-x+x| \leq |x| + |x-y| \Rightarrow |y|-|x| \leq |x-y| \end{aligned} \quad \Rightarrow |x|-|y| \leq |x-y|$$

$\uparrow L=1$

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

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

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

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

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

Lipschitz Gradient

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

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

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

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

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

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

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

3. If f is twice continuously differentiable

with

$$\nabla^2 f(x) \succcurlyeq -M\mathbf{I} \quad \text{and} \quad \nabla^2 f(x) \leq M\mathbf{I}$$

then

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|$$

$$\forall x, y \in \mathbb{R}^n$$

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

$A \prec B$ means $B - A \succcurlyeq 0$

$$-MI \leq \nabla^2 f(x) \leq MI \quad \forall x \Rightarrow \|\nabla f(x) - \nabla f(y)\| \leq M \|y-x\|$$

Proof For symmetric A ,

$$1. \quad x^T A x \leq \lambda_{\max}(A) \|x\|^2$$

$$2. \quad \lambda_i(A^2) = \lambda_i^2(A)$$

$$3. \quad -MI \leq A \leq MI \Rightarrow \lambda_{\min}(A) \geq -M, \quad \lambda_{\max}(A) \leq M$$

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

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

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

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

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

$$\begin{aligned} \text{Triangle ineq.} \quad & \leq \int_0^1 \|\nabla^2 f(x + s(y-x))(y-x)\| ds \\ & = \int_0^1 \sqrt{(y-x)^T [\underbrace{\nabla^2 f(x + s(y-x))}_H]^2 (y-x)} ds \end{aligned}$$

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

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

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

Descent Lemma Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $g(1) = f(y)$

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

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

$$\begin{aligned} \Rightarrow f(y) &= f(x) + \int_0^1 \nabla f(x + t(y-x))^T (y-x) dt \\ &= f(x) + \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x))^T (y-x) dt \\ &\quad + \int_0^1 \nabla f(x)^T (y-x) dt \\ &\stackrel{\text{Cauchy-Schwarz}}{\leq} f(x) + \nabla f(x)^T (y-x) \\ &\quad + \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\| \|y-x\| dt \\ &\leq f(x) + \nabla f(x)^T (y-x) + L \int_0^1 t \|y-x\| dt \\ &= f(x) + \nabla f(x)^T (y-x) + L \|y-x\|^2 \underbrace{\int_0^1 t dt}_{=\frac{1}{2}} \end{aligned}$$

Convergence of steepest Descent with Fixed Stepsize

Theorem consider the GD algorithm

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

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,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2} L \|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \alpha \nabla f(x_k)^T \nabla f(x_k) + \frac{1}{2} L \alpha^2 \|\nabla f(x_k)\|^2 \\ &= f(x_k) + \alpha \left(\frac{1}{2} L \alpha - 1 \right) \|\nabla f(x_k)\|^2 \\ \Rightarrow \|\nabla f(x_k)\|^2 \alpha \left(1 - \frac{1}{2} L \alpha \right) &\leq f(x_k) - f(x_{k+1}) \\ \Rightarrow \alpha \left(1 - \frac{1}{2} L \alpha \right) \sum_{k=0}^N \|\nabla f(x_k)\|^2 &\leq f(x_0) - f(x_{N+1}) \\ &\leq f(x_0) - f_{\min} \end{aligned}$$

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

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

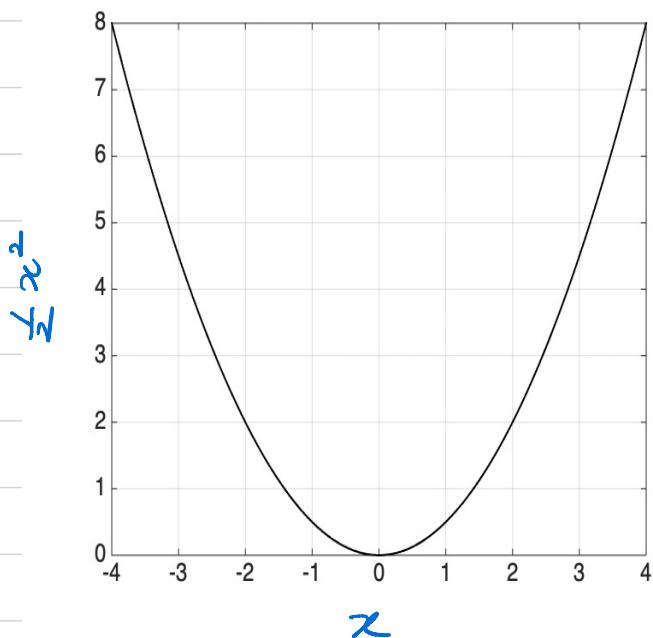
$$\Rightarrow \lim_{k \rightarrow \infty} \nabla f(x_k) = 0$$

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

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

Example

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



$$\nabla f(x) = x$$

Lipschitz with $L = 1$

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

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

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

$$\Rightarrow x_k = x_0 (-0.5)^k \rightarrow 0 \text{ as } k \rightarrow \infty$$

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

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

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

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

Example What if gradient is not Lipschitz?

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

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

$x = 0$ 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 $x_0 = 0$
- $|x_1| < |x_0| \Leftrightarrow |x_0| |1 - 4\alpha x_0^2| < |x_0|$
 $\Leftrightarrow -1 < 1 - 4\alpha x_0^2 < 1$
 $\Leftrightarrow 0 < x_0^2 < \frac{1}{2\alpha} \Leftrightarrow 0 < |x_0| < \frac{1}{\sqrt{2\alpha}}$
- Therefore, if $|x_1| < |x_0|$, then $|x_1| < |x_0| < \frac{1}{\sqrt{2\alpha}}$
 $\Rightarrow |x_2| < |x_1|, \dots |x_{k+1}| < |x_k|, \forall k \Rightarrow \{|x_k|\}$ 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 $0 < |x_0| < \frac{1}{\sqrt{2\alpha}} \Rightarrow |x_k| \rightarrow 0$, i.e. $x_k \rightarrow 0$ as $k \rightarrow \infty$

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

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

$$\text{Bnt } \frac{|x_{k+1}|}{|x_k|} = |1 - 4\alpha x_k^2| \rightarrow |1 - 4\alpha c^2|$$

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

$$\text{Bnt } c < |x_0| < \frac{1}{\sqrt{2\alpha}} \quad \text{--- (2)}$$

(1) and (2) are contradictory $\Rightarrow c = 0$.

$$\text{If } |x_0| = \frac{1}{\sqrt{2\alpha}}, |x_1| = |x_0| |1 - 2| = |x_0| = |x_k| \quad \forall k$$