Gradient Method: Part 3

Lecture 4-3 - CMSE 382

Prof. Firas Khasawneh

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Fri, Feb 6, 2026

This Lecture

Topics:

- Review: Lipschitz continuity
- Convergence of the gradient method

Announcements:

- None

Section 1

Lipschitz Continuity

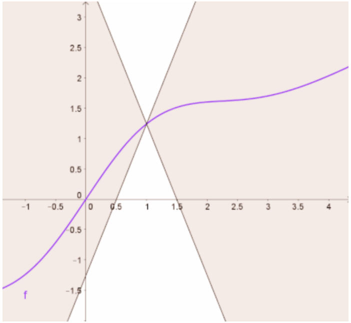
Definition

A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is **Lipschitz continuous** if there exists an $L > 0$ such that

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

- If f is Lipschitz with constant L, then it is Lipschitz with any constant $L' \geq L$.
- We are usually interested in the smallest Lipschitz constant.



▸ View Lipschitz Continuity GIF

We assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that ∇f is **Lipschitz continuous**:

- There exists an $L > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

- If ∇f is Lipschitz with constant L, then it is Lipschitz with any constant $L' \geq L$.
- The class of functions over a set D whose gradient satisfies the Lipschitz condition with constant L is denoted $C_L^{1,1}(D)$.
  - $C_L^{1,1}(D)$ is the class of functions "C-one-comma-one L-Lipschitz".
  - If we don't care about the value of L, we write $C^{1,1}(D)$ to denote functions that are Lipschitz for some L.

Theorem: $f \in C_L^{1,1}$ if and only if $\|\nabla^2 f(\mathbf{x})\| \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$.

- **Linear functions**: Given $\mathbf{a} \in \mathbb{R}^n$, $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ is in $C^{1,1}$.
- **Quadratic functions**: If $\mathbf{A}$ is an $n \times n$ symmetric matrix, $\mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$, then $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top + c$ is a $C^{1,1}$ function. ($\|\nabla^2 f\| = 2\|A\|$).

$$
\begin{aligned}
\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| &= 2\|(\mathbf{A}\mathbf{x} + \mathbf{b}) - (\mathbf{A}\mathbf{y} + \mathbf{b})\| \\
&= 2\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\| \\
&= 2\|\mathbf{A}(\mathbf{x} - \mathbf{y})\| \\
&\leq 2\|\mathbf{A}\| \, \|\mathbf{x} - \mathbf{y}\| \\
&\leq L\|\mathbf{x} - \mathbf{y}\|
\end{aligned}
$$

Choice of norm:

- The value of the Lipschitz constant L depends on the norm used
- The choice of the norm does not impact Lipschitz continuity, only the the value of L .

Section 2

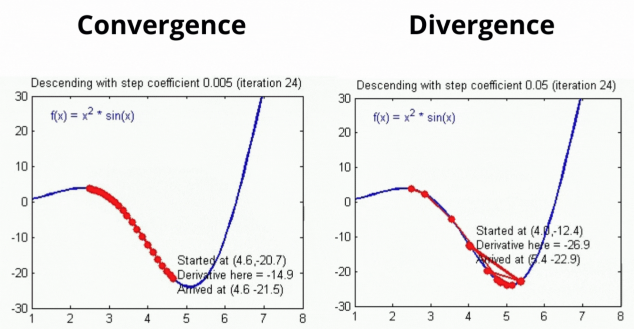
Convergence of Gradient Method

Convergence of the Gradient Method

Given an unconstrained optimization problem

$$\min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}$$

- When will gradient descent converge?
- How many iterations to reach the stopping criteria $\|\nabla f(\mathbf{x}_k)\|^2 < \varepsilon$?



View Convergence GIF

Convergence of the Gradient Method

Theorem

Let $f \in C_L^{1,1}(\mathbb{R}^n)$, and let $\{\mathbf{x}_k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$ with one of the following stepsize strategies:

- *constant stepsize $\bar{t} \in (0, \frac{2}{L})$*
- *exact line search*
- *backtracking procedure with $s \in \mathbb{R}_{++}, \alpha \in (0,1)$, and $\beta \in (0,1)$.*

Assume there exists $m \in \mathbb{R}$ such that $f(\mathbf{x}) > m$ for all $\mathbf{x} \in \mathbb{R}^n$. Then:

(a) *The sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is nonincreasing. In addition, for any $k \geq 0$, $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ unless $\nabla f(\mathbf{x}_k) = \mathbf{0}$.*

(b) *$\nabla f(\mathbf{x}_k) \to \mathbf{0}$ as $k \to \infty$.*
  - *Cannot guarantee convergence to a global optima, but can show convergence to a stationary point.*

Convergence of the Gradient Method

Rate of Convergence of Gradient Norms

Theorem

Under the setting of the previous theorem, let f^* be the limit of the convergent sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$. Then for $n+1$ iterations $\exists k$ such that

$$\min_{k=0,1,\dots,n} \|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{f(\mathbf{x}_0) - f^*}{M(n+1)}$$

where

$$M = \begin{cases} \bar{t}\left(1 - \frac{\bar{t}L}{2}\right) & \text{constant stepsize} \\ \frac{1}{2L} & \text{exact line search} \\ \alpha \min\left\{s, \frac{2\beta(1-\alpha)}{L}\right\} & \text{backtracking} \end{cases}$$

- Independent of the data vector size d (but L can grow with d).

How many iterations to reach the stopping criteria $\|\nabla f(\mathbf{x}_k)\|^2 < \varepsilon$?

- How many iterations to reach the stopping criteria $\|\nabla f(\mathbf{x}_k)\|^2 < \varepsilon$?
- $\|\nabla f(\mathbf{x}_k)\|^2 \leq \frac{f(\mathbf{x}_0)-f^*}{M(n+1)}$, so for $\|\nabla f(\mathbf{x}_k)\|^2 \leq \varepsilon$,

$$(n+1) \geq \frac{f(\mathbf{x}_0) - f^*}{M\varepsilon}$$

  - We need $(n+1)$ of order $1/\varepsilon$ for $\|\nabla f(\mathbf{x}_k)\|^2 < \varepsilon$
  - In practice, gradient descent converges much faster

Advantages and limitations of gradient descent method:

+ Simple and easy to implement
+ Very fast for well-conditioned objective functions (can find global optima for convex functions)
- Often slow for non-convex problems
- Inapplicable to non-differentiable functions

Groups - Round 2

Group 1
Abigail, Shanze, Jack, Quang Minh,

Group 2
Igor, Atticus, K M Tausif, Long,

Group 3
Yousif, Zheng, Jake, Purvi,

Group 4
Maye, Alice, Arjun, Kyle,

Group 5
Monirul Amin, Jay, Brandon, Luis,

Group 6
Scott, Ha, Lora, Tianjian,

Group 7
Braedon, Sai, Joseph, Noah,

Group 8
Michal, Aidan, Jonid, Dev,

Group 9
Vinod, Saitej, Anthony, Breena,

Group 10
Karen, Dori, Lowell, Aaron,

Group 11
Jamie, Sanskaar, Dominic, Lauryn,

Group 12
Andrew, Arya, Daniel, Morgan,