

Optimization for Deep Learning

Gradient Descent

Kun Yuan

Peking University

Main contents in this lecture

- Convex sets, functions, and problems
- Strong convexity and smoothness
- Gradient descent
- Convergence analysis

Convex sets

Definition 1 (Convex set)

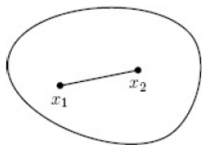
A set $\mathcal{X} \subseteq \mathbb{R}^d$ is called convex, if for $\forall x, y \in \mathcal{X}$, it holds that

$$\theta x + (1 - \theta)y \in \mathcal{X}, \quad \forall \theta \in [0, 1].$$

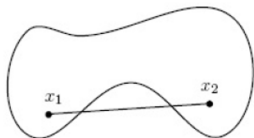
Examples:

- Hyperplane $\{x | a^T x = b\}$ and hyperspace $\{x | a^T x \leq b\}$
- Euclidian ball $\{x | \|x - x_c\| \leq r\}$
- Polyhedron $\{x | a_j^T x \leq b_j, j = 1, \dots, m, c_j^T x = d_j, j = 1, \dots, p\}$

Convex sets: illustration



Convex set



Non-convex set

Convex function

Definition 2 (Convex function)

Function $f : \mathcal{X} \rightarrow \mathbb{R}$ is said to be convex if $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set and $\forall x, y \in \mathcal{X}$, it holds that

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y), \quad \forall \theta \in [0, 1].$$

Examples:

- Exponential e^{ax} is convex on \mathbb{R} for any $a \in \mathbb{R}$
- Norms $\|x\|_1$ and $\|x\|_2$ are convex on \mathbb{R}^d
- Linear regression loss function $\|Ax - b\|^2$ is convex on \mathbb{R}^d
- Logistic regression $\frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-b_i a_i^T x))$ is convex on \mathbb{R}^d

Convex function: illustration

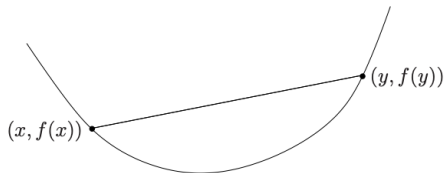


Figure: An illustration of a convex function ¹

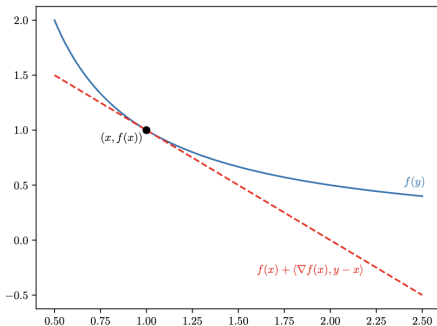
¹This figure is from (?)

Convex function: property

Lemma 1 (Convex property)

Suppose $f : \mathcal{X} \rightarrow \mathbb{R}$ is differentiable, then f is convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathcal{X}.$$

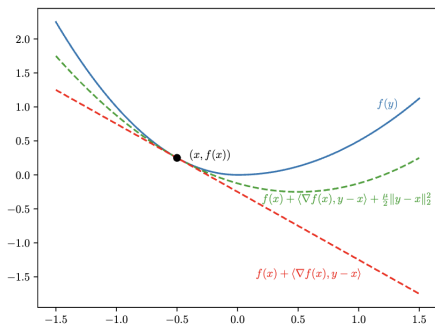


Strongly-convex function

Definition 3 (μ -strongly convex function)

Function $f : \mathcal{X} \rightarrow \mathbb{R}$ is μ -strongly convex if there exists a constant $\mu > 0$ such that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathcal{X}.$$



Definition 4 (L -smoothness)

A differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be L -smooth if $\forall x, y \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2,$$

where $L > 0$ is the Lipschitz constant of ∇f .

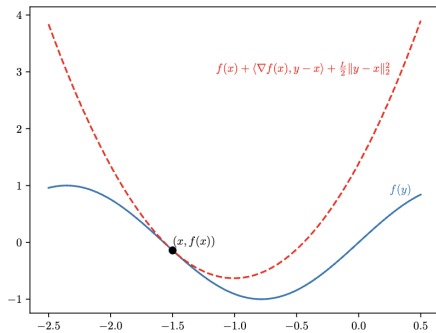
In other words, the gradient cannot vary too quickly

It is easy to show that the above inequality is equivalent to (see notes)

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$

which implies that $f(y)$ can be upper bounded by a quadratic function

L -smoothness: illustration



Gradient descent

- Consider the following **smooth** and **unconstrained** optimization

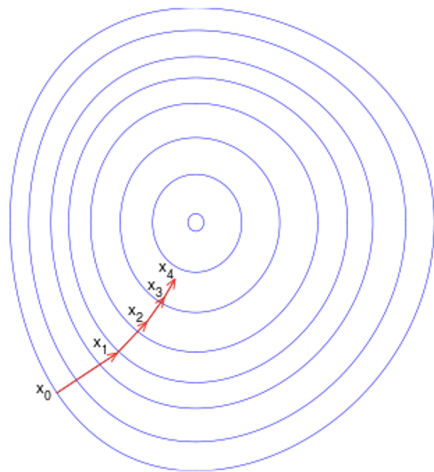
$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x)$$

- Gradient descent (GD)** is very effective to solve the above problem

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

where γ is the learning rate (or step size), and x_0 initializes arbitrarily.

Gradient descent: illustration

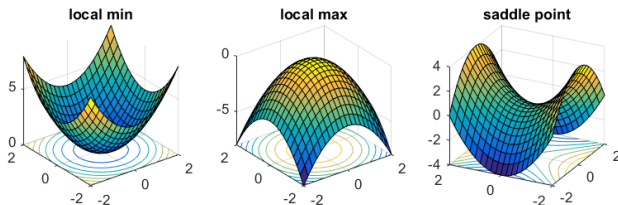


Stationary solution

Given a differentiable function $f(x)$, x^* is the **stationary solution** if and only if

$$\nabla f(x^*) = 0$$

Illustration of the stationary solution²



For convex functions, stationary solutions are global solutions.

²Figure is from Prof. Rong Ge's online post

Convergence analysis: non-convex scenario

With small γ , $\{f(x_k)\}$ is a strictly decreasing sequence

Lemma 1 (Decay in function value)

Assume $f(x)$ to be L -smooth. If $\gamma \leq 1/L$, it holds that

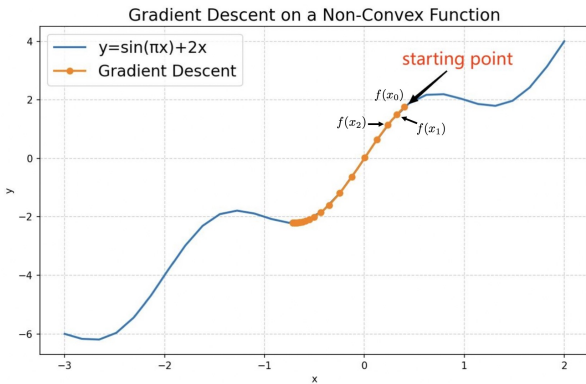
$$f(x_{k+1}) \leq f(x_k) - \frac{\gamma}{2} \|\nabla f(x_k)\|^2$$

From the above theorem, we conclude that

$$f(x_0) > f(x_1) > \cdots f(x_k) > f(x_{k+1}) > \cdots$$

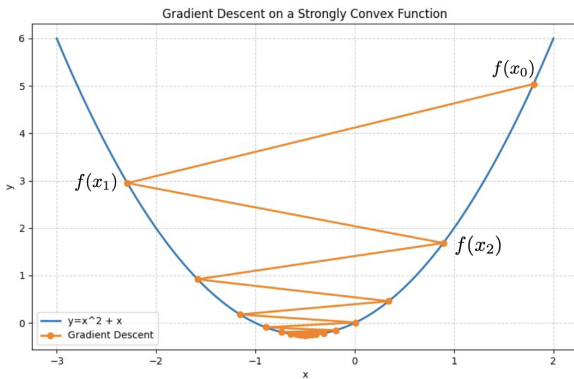
Convergence analysis: non-convex scenario

Illustration:



Convergence analysis: non-convex scenario

Illustration:



Convergence analysis: non-convex scenario

Theorem 1

Assume $f(x)$ to be L -smooth. If $\gamma \leq 1/L$, gradient descent converges as

$$\frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f^*)}{\gamma(K+1)}.$$

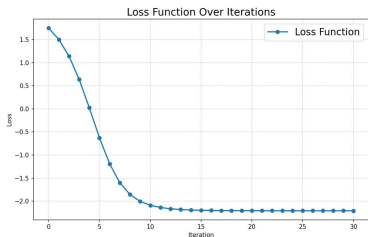
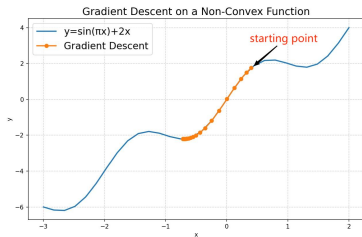
If we further set $\gamma = 1/L$, it holds that

$$\frac{1}{K+1} \sum_{k=0}^K \|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0) - f^*)}{K+1}.$$

- If the ergodic average converges to 0, then it holds that $\|\nabla f(x_k)\| \rightarrow 0$
- Smaller γ leads to slower convergence
- Convergence rate is $O(L/K)$, which implies $O(L/\epsilon)$ iterations to achieve an ϵ -accurate solution, i.e., the iteration complexity is $O(L/\epsilon)$.

Experiments: non-convex scenario

We minimize $f(x) = \sin(\pi x) + 2x$



Convergence analysis: convex scenario

Theorem 2

Suppose $f(x)$ is convex and L -smooth, if $\gamma = 1/(2L)$, gradient descent converges as

$$f(x_K) - f^* \leq \frac{2L\|x_0 - x^*\|^2}{K+1}.$$

For convex functions, it holds that $f(x_k) \rightarrow f^*$ at rate $O(L/K)$

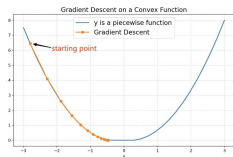
The iteration complexity $O(L/\epsilon)$

Experiments: convex scenario

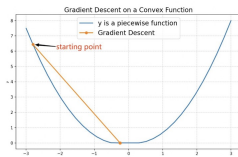
We minimize

$$f(x) = \begin{cases} \frac{3}{2}x^2, & \text{if } x \leq 0 \\ \frac{9}{2}x^2, & \text{if } x > 0 \end{cases}$$

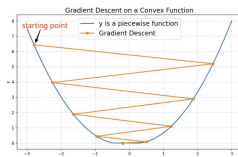
Gradient descent is sensitive to the choice of learning rate



$$\gamma = \frac{1}{5L}$$



$$\gamma = \frac{1}{L}$$



$$\gamma = \frac{2}{L}$$

Convergence analysis: strongly-convex scenario

Theorem 3

Assume $f(x)$ is L -smooth and μ -strongly convex, if $\gamma = 2/(L + \mu)$, gradient descent converges as

$$\|x_K - x^*\| \leq \left(\frac{L - \mu}{L + \mu} \right)^{K+1} \|x_0 - x^*\|$$

For strongly-convex functions, it holds that $x_k \rightarrow x^*$ at rate $O\left((1 - \frac{1}{\kappa})^k\right)$ where $\kappa = L/\mu$ is regarded as the condition number of the strongly-convex function

GD converges exponentially (or linearly) fast for strongly-convex problems

The iteration complexity is $O(\kappa \log(1/\epsilon))$

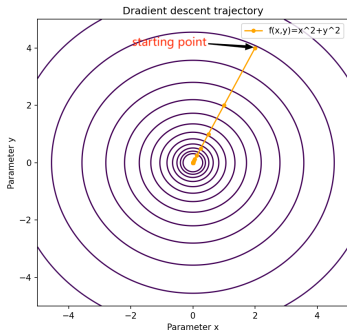
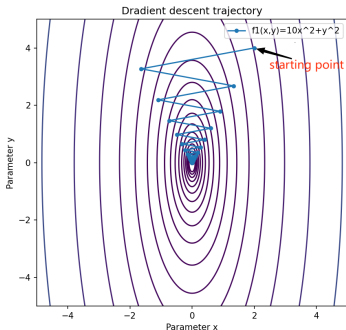
Experiments: strongly-convex scenario

We minimize

$$f_1(x, y) = x^2 + y^2$$

$$f_2(x, y) = 10x^2 + y^2$$

The condition number $\kappa = L/\mu$ has a significant influence on convergence



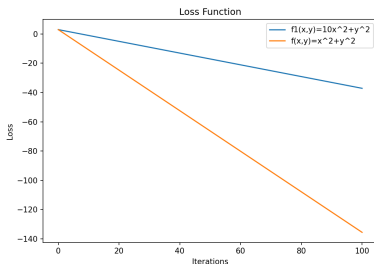
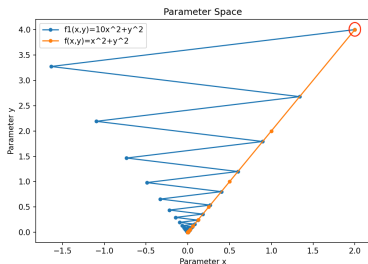
Experiments: strongly-convex scenario

We minimize

$$f_1(x, y) = x^2 + y^2$$

$$f_2(x, y) = 10x^2 + y^2$$

The condition number $\kappa = L/\mu$ has a significant influence on convergence



Convergence analysis: proof details

All proof details are in the notes.

Convergence rate summary

	Convergence rate	Iteration complexity
Non-convex	$O(L/K)$	$O(L/\epsilon)$
Generally-convex	$O(L/K)$	$O(L/\epsilon)$
Strongly-convex	$O((1 - \frac{\mu}{L})^K)$	$O(\frac{L}{\mu} \log(1/\epsilon))$

Summary

- Gradient descent is very popular for unconstrained and smooth optimization
- For non-convex problems, GD converges at rate $O(L/K)$
- For generally-convex problems, GD converges at rate $O(L/K)$
- For strongly-convex problems, GD converges at rate $O((\frac{L-\mu}{L+\mu})^K)$

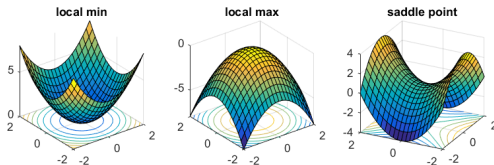
Advanced topic

- In this lecture, we show that

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

which implies that GD will converge to the stationary solution

- Will GD converge to the local maximum of saddle point?



- No! GD will converge to the local minimum with probability 1

References I