Gradient Descent

Cong Ma

University of Chicago, Winter 2026

Outline

- Gradient descent algorithm
- Smooth problems
- Convex and smooth problems
- Strongly convex and smooth problems
- Backtracking line search
- Preconditioned GD

Problem Setup

We consider unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} f(x),$$

where:

- $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable
- Gradient $\nabla f(x)$ is available

Goal: find a point $x^\star$ such that $\nabla f(x^\star) = 0$, which we assume exists.

Descent Directions

Definition 1 (Descent Direction)

A vector d is a *descent direction* for f at x if

$$f(x + td) < f(x)$$

for all sufficiently small $t > 0$.

A simple sufficient characterization is given by the following result.

Lemma 2

If f is continuously differentiable in a neighborhood of x, then any direction d such that

$$d^\top \nabla f(x) < 0$$

is a descent direction.

Gradient Descent Algorithm

Basic idea: move in the direction of negative gradient

Given an initial point x^0, iterate:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k),$$

where:

- $\alpha_k > 0$ is the step size (learning rate)
- $k = 0, 1, 2, \ldots$

Gradient descent is a first-order method: it uses only gradient information.

Steepest Descent Direction

Among all unit directions:

$$\min_{\|d\|=1} \ \nabla f(x)^\top d$$

Solution:

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$

Therefore the steepest descent direction is:

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

(up to scaling)

Geometric Interpretation

- $\nabla f(x)$ points in the direction of steepest ascent
- $-\nabla f(x)$ points in the direction of steepest descent
- Each iteration moves to reduce the objective value locally

Proximal View of Gradient Descent

Gradient descent can be viewed as:

$$x^{k+1} = \arg\min_y \left\{ \underbrace{f(x^k) + \nabla f(x^k)^\top (y - x^k)}_{\text{first-order approx.}} + \underbrace{\frac{1}{2\alpha_k} \|y - x^k\|_2^2}_{\text{proximal term}} \right\}.$$

Step-Size Selection

The step size α_k critically affects performance.

Common choices:

- **Constant step size:** $\alpha_k = \alpha$
- **Diminishing step size:** $\alpha_k \downarrow 0$
- **Line search:** choose α_k to sufficiently decrease f

Too small α_k: slow convergence Too large α_k: divergence or oscillations

Smooth Functions

Definition 3

f is L-smooth if

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

for all x, y.

Equivalent inequality:

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

Second-order characterization

$$\|\nabla^2 f(x)\|_2 \ \leq \ L, \qquad \forall x \quad \text{(for twice differentiable functions)}$$

Descent Lemma

Lemma 4 (Smoothness Upper Bound)

Assume f is L-smooth. Then for any x, direction d, and stepsize α,

$$f(x + \alpha d) \leq f(x) + \alpha \nabla f(x)^\top d + \frac{L\alpha^2}{2}\|d\|^2.$$

This follows from Taylor expansion and Lipschitz continuity of the gradient.

Applying the Descent Lemma

Choose the gradient descent direction:

$$d = -\nabla f(x).$$

Substitute into the lemma:

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha \|\nabla f(x)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x)\|^2.$$

The right-hand side is minimized at

$$\alpha = \frac{1}{L}.$$

Constant Stepsize Guarantee

Set

$$\alpha = \frac{1}{L}.$$

Then gradient descent satisfies:

$$f(x^{k+1}) = f\left(x^k - \frac{1}{L}\nabla f(x^k)\right) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2.$$

Conclusion

Each iteration produces a guaranteed decrease proportional to the squared gradient norm.

General Case: Descent and Gradient Summability

From the one-step descent inequality,

$$f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2,$$

assume f is lower bounded:

$$f(x) \ge \bar{f}.$$

Summing over $k = 0, \ldots, T-1$ and telescoping gives

$$f(x^T) \le f(x^0) - \frac{1}{2L}\sum_{k=0}^{T-1}\|\nabla f(x^k)\|^2.$$

Using $f(x^T) \ge \bar{f}$, we obtain

$$\sum_{k=0}^{T-1}\|\nabla f(x^k)\|^2 \le 2L\big(f(x^0) - \bar{f}\big).$$

Asymptotic Stationarity

From bounded sum:

$$\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 < \infty$$

it follows that

$$\lim_{k\to\infty} \|\nabla f(x^k)\| = 0$$

Interpretation

Gradient descent converges to a stationary point (not necessarily global minimum).

Rate of Convergence

From averaging:

$$\min_{0\le k\le T-1} \|\nabla f(x^k)\|^2 \le \frac{1}{T}\sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2$$

Using previous bound:

$$\min_{0\le k\le T-1} \|\nabla f(x^k)\| \le \sqrt{\frac{2L\big(f(x^0)-\bar{f}\big)}{T}}$$

This gives an $O(T^{-1/2})$ stationarity rate.

Convex Case: Gradient Descent

We now assume:

- f is **convex**
- f is L-**smooth**
- Global minimizer x^* exists

Define optimal value:

$$f^* = f(x^*)$$

We analyze gradient descent with constant stepsize

$$\alpha = \frac{1}{L}.$$

Convergence Rate (Convex Case)

Theorem 5

Suppose f is convex and L-smooth, and let x^* be a minimizer. Then gradient descent with stepsize $\alpha = 1/L$ satisfies:

$$f(x^T) - f^* \le \frac{L}{2T}\|x^0 - x^*\|^2, \qquad T = 1, 2, \dots$$

Rate

Function value convergence is $O(1/T)$.

Proof Sketch (Convex Case): Key Inequalities

By convexity,

$$f(x^\star) \geq f(x^k)+\nabla f(x^k)^\top(x^\star-x^k) \quad \Rightarrow \quad \nabla f(x^k)^\top(x^k-x^\star) \geq f(x^k)-f^\star.$$

Combining with descent:

$$f(x^{k+1}) \leq f(x^\star) + \nabla f(x^k)^\top(x^k - x^\star) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

which implies

$$f(x^{k+1}) \leq f(x^\star) + \frac{L}{2}\Big(\|x^k - x^\star\|^2 - \|x^{k+1} - x^\star\|^2\Big).$$

Proof Sketch (Convex Case): Telescoping and Rate

Summing over $k = 0, \ldots, T-1$ gives

$$\sum_{k=0}^{T-1} (f(x^{k+1}) - f^\star) \le \frac{L}{2} \|x^0 - x^\star\|^2.$$

Since $f(x^k)$ is nonincreasing,

$$f(x^T) - f^\star \le \frac{1}{T} \sum_{k=0}^{T-1} (f(x^{k+1}) - f^\star).$$

Therefore,

$$f(x^T) - f^\star \le \frac{L}{2T} \|x^0 - x^\star\|^2.$$

Result

Gradient descent achieves $\mathcal{O}(1/T)$ convergence in function value.

Strongly Convex Functions

f is μ-strongly convex if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2,$$

Equivalent second-order characterization

1. $$\nabla^2 f(x) \succeq \mu I, \qquad \forall x \quad \text{(for twice differentiable functions)}$$

Gradient-based error bounds

Lemma 6

Let f be continuously differentiable and m-strongly convex. Then the following inequalities hold:

$$f(x) - f(x^\star) \;\leq\; \frac{\|\nabla f(x)\|^2}{2m},$$

and

$$\|x - x^\star\| \;\leq\; \frac{2}{m}\,\|\nabla f(x)\|.$$

Proof

Suppose f is m-strongly convex. Minimizing both sides of the strong convexity inequality with respect to z, we obtain

$$f(x^\star) \geq f(x) - \nabla f(x)^\top \left(\frac{1}{m}\nabla f(x)\right) + \frac{m}{2}\left\|\frac{1}{m}\nabla f(x)\right\|^2.$$

Simplifying,

$$f(x^\star) = f(x) - \frac{1}{2m}\|\nabla f(x)\|^2.$$

Strong convexity: error bound

Rearranging the previous inequality yields

$$\|\nabla f(x)\|^2 \;\geq\; 2m\big(f(x) - f(x^\star)\big).$$

In particular, if $\|\nabla f(x)\| \leq \delta$, then

$$f(x) - f(x^\star) \;\leq\; \frac{\|\nabla f(x)\|^2}{2m} \;\leq\; \frac{\delta^2}{2m}.$$

Interpretation

For strongly convex functions, a small gradient norm guarantees that we are close to the optimal function value.

Distance to Optimum via Gradient

We estimate the distance to the optimizer $x^\star$ using strong convexity and the Cauchy–Schwarz inequality.

From strong convexity,

$$f(x^\star) \geq f(x) + \nabla f(x)^\top (x^\star - x) + \frac{m}{2}\|x - x^\star\|^2.$$

Applying Cauchy–Schwarz,

$$\nabla f(x)^\top (x^\star - x) \geq -\|\nabla f(x)\|\,\|x^\star - x\|.$$

Therefore,

$$f(x^\star) \geq f(x) - \|\nabla f(x)\|\,\|x^\star - x\| + \frac{m}{2}\|x - x^\star\|^2.$$

Gradient and Distance Bound

Rearranging the previous inequality yields

$$\|x - x^\star\| \;\leq\; \frac{2}{m}\,\|\nabla f(x)\|.$$

Interpretation

For strongly convex functions, a small gradient norm implies that the current iterate is close to the optimizer in distance.

Linear convergence of gradient descent

Theorem 7 (Linear convergence for m-strongly convex and L-smooth f)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, m-strongly convex, and L-smooth. Consider gradient descent with constant stepsize $\eta = 1/L$:

$$x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k).$$

Let $x^\star \in \arg\min_x f(x)$ and $f^\star := f(x^\star)$. Then for all $k \geq 0$,

$$f(x^{k+1}) - f^\star \leq \left(1 - \frac{m}{L}\right) (f(x^k) - f^\star).$$

Consequently, after T iterations,

$$f(x^T) - f^\star \leq \left(1 - \frac{m}{L}\right)^T (f(x^0) - f^\star).$$

Proof

We analyze the convergence of gradient descent for m-strongly convex and L-smooth functions.

Using the update rule

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k),$$

and substituting the gradient bound, we obtain

$$f(x^{k+1}) = f\left(x^k - \frac{1}{L}\nabla f(x^k)\right) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2.$$

By strong convexity,

$$f(x^{k+1}) \le f(x^k) - \frac{m}{L}\big(f(x^k) - f^\star\big),$$

where $f^\star = f(x^\star)$.

Linear convergence rate

Subtracting $f^\star$ from both sides yields the recursion

$$f(x^{k+1}) - f^\star \leq \left(1 - \frac{m}{L}\right)(f(x^k) - f^\star).$$

Thus, the function values converge linearly to the optimum.

After T iterations,

$$f(x^T) - f^\star \leq \left(1 - \frac{m}{L}\right)^T (f(x^0) - f^\star).$$

Comparison Between Convergence Rates

It is straightforward to convert convergence bounds into iteration complexities.

From the smooth nonconvex guarantee, there exists some $k \le T$ such that

$$\|\nabla f(x^k)\| \le \varepsilon$$

provided that

$$T \ \ge \ \frac{2L\big(f(x^0) - f^\star\big)}{\varepsilon^2}.$$

Interpretation

To find an ε-stationary point, gradient descent requires

$$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

iterations.

Convex vs Strongly Convex Rates

General convex case:
From the convex convergence bound,

$$f(x^k) - f^\star \le \varepsilon$$

whenever

$$k \ \ge \ \frac{L\|x^0 - x^\star\|^2}{2\varepsilon}.$$

Strongly convex case:
From linear convergence,

$$f(x^k) - f^\star \le \varepsilon$$

whenever

$$k \ \ge \ \frac{L}{m} \log\left(\frac{f(x^0) - f^\star}{\varepsilon}\right).$$

To be updated

Motivation for Preconditioning

When level sets are elongated:

- Gradient descent zig-zags
- Convergence becomes slow

This happens when the problem is ill-conditioned.

Preconditioned Gradient Descent

Instead of:

$$x^{k+1} = x^k - \alpha \nabla f(x^k),$$

use:

$$x^{k+1} = x^k - \alpha P \nabla f(x^k),$$

where:

- $P \succ 0$ is a preconditioning matrix

Interpretation:

- Gradient descent in a different metric
- Rescales directions to improve conditioning

Connections and Remarks

- Newton's method is GD with $P = (\nabla^2 f)^{-1}$
- Diagonal preconditioning leads to adaptive methods
- Choice of P can dramatically improve convergence

Preconditioning bridges first-order and second-order methods.

Summary

- Gradient descent is simple and widely applicable
- Step size selection is crucial
- Convergence depends on convexity and smoothness
- Preconditioning improves performance on ill-conditioned problems

Gradient descent is the backbone of modern optimization and machine learning.