

On the Convergence Properties of Iterative Optimization Methods

Demo Author

March 16, 2026

Abstract

This document demonstrates a complex multi-file $\text{\LaTeX}$ project structure with nested references, cross-citations, and custom packages. We study the convergence of gradient descent (Section 1.1), present theoretical guarantees (Theorem 3.2), and validate with numerical results (Table 1).

Contents

1	Introduction	2
1.1	Problem Statement	2
1.2	Research Objectives	2
2	Background	3
2.1	Convex Analysis	3
2.2	Lipschitz Continuity	3
2.3	Gradient Descent	3
3	Methodology	4
3.1	Main Theorems	4
3.2	Proofs	4
3.3	Numerical Results	5
4	Conclusion	6
4.1	Summary of Contributions	6
4.2	Future Work	6
A	Supplementary Derivations	6
A.1	Descent Lemma	6
A.2	Connection to Proximal Methods	7

1 Introduction

Optimization lies at the heart of modern machine learning and scientific computing. Given an objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the fundamental problem is to find

$$x^* = \arg \min_{x \in \mathbb{R}^n} f(x). \quad (1)$$

This work addresses two key questions. First, we formalize the problem statement and identify the assumptions under which convergence can be guaranteed (Section 1.1). Second, we outline the research objectives that guide our theoretical analysis (Section 1.2).

1.1 Problem Statement

Consider the unconstrained optimization problem from Eq. (1). We make the following standing assumptions:

1. $f \in C^2(\mathbb{R}^n)$ is twice continuously differentiable.
2. f is μ -strongly convex (Definition 3.1) with $\mu > 0$.
3. ∇f is L -Lipschitz continuous (Eq. (4)) with $L \geq \mu$.

Under these assumptions, the optimization problem (1) has a unique minimizer $x^* \in \mathbb{R}^n$ satisfying $\nabla f(x^*) = 0$. The condition number $\kappa = L/\mu \geq 1$ characterizes the difficulty of the problem [3].

1.2 Research Objectives

The objectives of this work are threefold:

1. **Convergence rate.** Establish the exact convergence rate of gradient descent (5) under the assumptions of Section 1.1. We aim to show that the iterates satisfy

$$f(x_k) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f(x^*)). \quad (2)$$

2. **Optimal step size.** Determine the step size η^* that minimizes the contraction factor in (2).
3. **Numerical validation.** Compare theoretical predictions with empirical convergence curves across varying condition numbers (see Table 1).

2 Background

We review the necessary mathematical background for our analysis.

2.1 Convex Analysis

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \quad (3)$$

Convexity ensures that every local minimum is a global minimum [1].

2.2 Lipschitz Continuity

A differentiable function f has L -Lipschitz continuous gradients if

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^n. \quad (4)$$

This condition, combined with strong convexity (Definition 3.1 in Section 3), yields linear convergence rates as shown by Nesterov [2].

2.3 Gradient Descent

The classical gradient descent iteration is

$$x_{k+1} = x_k - \eta \nabla f(x_k), \quad (5)$$

where $\eta > 0$ is the step size. The convergence rate depends on the condition number $\kappa = L/\mu$, as we formalize in Theorem 3.2.

3 Methodology

We now present the theoretical framework. We begin with core definitions and convergence theorems (Section 3.1), followed by detailed proofs (Section 3.2), and conclude with numerical validation (Section 3.3).

Definition 3.1 (Strong Convexity). *A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ -strongly convex ($\mu > 0$) if for all $x, y \in \mathbb{R}^n$:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2. \quad (6)$$

This extends the convexity condition from Eq. (3) in Section 2.1.

3.1 Main Theorems

We now state the main convergence result.

Theorem 3.2 (Linear Convergence of Gradient Descent). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly convex with L -Lipschitz continuous gradients. Consider the gradient descent iteration (5) with step size $\eta = 1/L$. Then for all $k \geq 0$:*

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|^2. \quad (7)$$

Corollary 3.3 (Optimal Step Size). *Under the conditions of Theorem 3.2, the step size $\eta^* = 1/L$ minimizes the contraction factor $\rho(\eta)$ defined in Lemma 3.4.*

Lemma 3.4 (Contraction Lemma). *For step size $0 < \eta < 2/L$, define the contraction factor*

$$\rho(\eta) = \max\{|1 - \eta\mu|, |1 - \eta L|\}. \quad (8)$$

Then $\|x_{k+1} - x^*\| \leq \rho(\eta) \|x_k - x^*\|$ for all $k \geq 0$.

3.2 Proofs

Proof of Lemma 3.4. By the gradient descent update (5) and optimality $\nabla f(x^*) = 0$:

$$\begin{aligned} x_{k+1} - x^* &= x_k - \eta \nabla f(x_k) - x^* \\ &= (x_k - x^*) - \eta (\nabla f(x_k) - \nabla f(x^*)). \end{aligned} \quad (9)$$

For a quadratic $f(x) = \frac{1}{2}x^\top Ax$ with $\mu I \preceq A \preceq LI$, we have $\nabla f(x_k) - \nabla f(x^*) = A(x_k - x^*)$, giving

$$x_{k+1} - x^* = (I - \eta A)(x_k - x^*). \quad (10)$$

The spectral radius of $I - \eta A$ is $\rho(\eta) = \max\{|1 - \eta\mu|, |1 - \eta L|\}$, which proves the claim. $\square$

Proof of Theorem 3.2. Applying Lemma 3.4 recursively:

$$\|x_k - x^*\| \leq \rho(\eta)^k \|x_0 - x^*\|. \quad (11)$$

With $\eta = 1/L$, we get $\rho(1/L) = \max\{1 - \mu/L, 0\} = 1 - \mu/L$. Squaring both sides of (11) yields Eq. (7). $\square$

Proof of Corollary 3.3. The optimal step size minimizes $\rho(\eta)$. Setting $|1 - \eta\mu| = |1 - \eta L|$ gives $\eta^* = 2/(\mu + L)$. However, since we restrict to f being general strongly convex (not necessarily quadratic), the tighter analysis of Nesterov [2] shows that $\eta^* = 1/L$ achieves the rate in Eq. (7). $\square$

3.3 Numerical Results

We validate Theorem 3.2 on quadratic problems $f(x) = \frac{1}{2}x^\top Ax$ with controlled condition numbers.

Table 1: Convergence of gradient descent with $\eta = 1/L$. Iterations to reach $\|x_k - x^*\| < 10^{-8}$ for $n = 100$.

$\kappa = L/\mu$	Theory: k^*	Observed: k	Ratio
10	212	198	0.93
100	2120	2089	0.99
1000	21203	21157	1.00

The theoretical bound $k^* = \frac{\kappa}{\ln(1/\epsilon)}$ from Eq. (7) closely matches the observed iteration counts. For $\kappa = 1000$, the prediction is within 0.2% of the empirical value, confirming the tightness of the bound established in Lemma 3.4.

4 Conclusion

We have presented a rigorous analysis of gradient descent for strongly convex functions with Lipschitz continuous gradients.

4.1 Summary of Contributions

Our main contributions are:

1. We proved the linear convergence rate of gradient descent (Theorem 3.2), showing that the error contracts at rate $1 - \mu/L$ per iteration.
2. We derived tight bounds on the optimal step size (Corollary 3.3), confirming $\eta^* = 1/L$.
3. We validated the theoretical predictions against numerical experiments (Table 1), demonstrating excellent agreement across condition numbers $\kappa \in \{10, 100, 1000\}$.

4.2 Future Work

Several directions remain open:

- Extending the analysis to the non-convex setting (cf. the problem formulation in Section 1.1).
- Incorporating momentum terms via Nesterov acceleration [2].
- Studying stochastic variants as suggested by Robbins and Monro [5].

See Appendix A for supplementary derivations.

A Supplementary Derivations

A.1 Descent Lemma

We provide the standard descent lemma used implicitly in the proof of Theorem 3.2.

Lemma A.1 (Descent Lemma). *If ∇f is L -Lipschitz continuous (Eq. (4)), then for all $x, y \in \mathbb{R}^n$:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2. \quad (12)$$

Proof. By the fundamental theorem of calculus and Cauchy–Schwarz:

$$\begin{aligned}
f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt \\
&= \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\
&\leq \langle \nabla f(x), y - x \rangle + \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \|y - x\| dt \\
&\leq \langle \nabla f(x), y - x \rangle + \int_0^1 Lt \|y - x\|^2 dt \\
&= \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.
\end{aligned} \tag{13}$$

□

A.2 Connection to Proximal Methods

When f is not smooth, gradient descent (5) generalizes to the proximal gradient method. The convergence analysis of Section 3.1 extends naturally; see Parikh and Boyd [4] for a comprehensive treatment.

Recall that the strong convexity condition (Definition 3.1) remains essential for linear convergence. Without it, the best achievable rate is $O(1/k)$ [1].

References

- [1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [2] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.
- [3] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.
- [4] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.
- [5] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.