On the Convergence of Gradient Descent in Non-Convex Optimization

J. Smith , A. Doe
Department of Mathematics

March 27, 2026

Abstract

We study convergence properties of gradient descent in non-convex optimization landscapes under standard L-smoothness assumptions. Our main result establishes an $\mathcal{O}(1/\sqrt{T})$ rate.

1. Introduction

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function, i.e., its gradient satisfies $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$. The *gradient descent* iteration is defined by

$$x_{t+1} = x_t - \eta \nabla f(x_t), \qquad \eta \in \left(0,\, \tfrac{1}{L}\right],$$

where η is the step size. Understanding the convergence of this simple procedure in the non-convex setting is fundamental to modern machine learning theory.

Theorem 1.1: Main Convergence Result

Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and bounded below. Running gradient descent with step size $\eta = 1/\sqrt{T}$ for T iterations yields

$$\min_{0 \leq t < T} \|\nabla f(x_t)\|^2 \;\leq\; \frac{2L\big(f(x_0) - f^*\big)}{\sqrt{T}} \;=\; \mathcal{O}\!\left(\frac{1}{\sqrt{T}}\right).$$

Proof. Applying the L-smoothness descent lemma repeatedly and telescoping yields the bound. □

Lemma 1.2: Descent Lemma

For any L-smooth f and step size $\eta \leq 1/L$,

$$f(x_{t+1}) \;\leq\; f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2.$$

2. Methodology

2.1 Convergence Analysis

Summing the descent lemma over $t = 0, \ldots, T-1$ and using the bounded-below assumption $f(x) \geq f^*$ gives a total decrease bound. Dividing by T and minimizing over all iterates establishes the convergence rate.

2.2 Practical Considerations

- **Step size selection:** Setting $\eta = 1/L$ (if L is known) yields the sharpest constant.
- **Stochastic variant:** Mini-batch SGD achieves the same $\mathcal{O}(1/\sqrt{T})$ rate in expectation under standard noise assumptions.
- **Adaptive methods:** Adam and AdaGrad adapt the step size per coordinate but lack matching worst-case guarantees in general.

3. Related Work

Classical convergence analysis of first-order methods dates to Nesterov (1983, 2004). The non-convex analysis presented here follows the framework of Ghadimi and Lan (2013). Stochastic extensions and variance-reduction techniques have been extensively studied; see the survey by Bottou et al. (2018) for an overview.

4. Conclusion

We have established an $\mathcal{O}(1/\sqrt{T})$ convergence rate for gradient descent on L-smooth non-convex functions — a fundamental guarantee underpinning the empirical success of gradient-based optimization in deep learning.