Gradient Descent in Nonconvex Optimization

Mikhailova Olena

27.03.2025

1 Introduction

In convex optimization, Gradient Descent (GD) guarantees convergence to a global minimum under certain conditions. However, many real-world problems involve *nonconvex* objectives. While GD cannot guarantee global optimality in such cases, it remains widely used to find *ϵ-stationary points* where the gradient norm is small ($\|\nabla f(x)\|_2 \leq \epsilon$). This document analyzes GD's behavior for nonconvex functions with Lipschitz-continuous gradients.

2 Problem Setup

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and satisfies the Lipschitz gradient condition:

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

This implies the quadratic upper bound:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2.$$

3 Convergence Analysis

Theorem 1 (6.4 in Notes)**.** *Let f be nonconvex with L-Lipschitz gradient. GD with fixed step size $t \leq 1/L$ satisfies:*

$$\min_{i=0,\dots,k} \|\nabla f(x^{(i)})\|_2 \leq \sqrt{\frac{2(f(x^{(0)}) - f^*)}{t(k+1)}},$$

where f^* is a lower bound of f.

Proof. Using the descent lemma for $x^{(i+1)} = x^{(i)} - t\nabla f(x^{(i)})$:

$$f(x^{(i+1)}) \leq f(x^{(i)}) - \frac{t}{2}\|\nabla f(x^{(i)})\|_2^2.$$

Summing over $i = 0$ to k:

$$f(x^{(k+1)}) - f(x^{(0)}) \leq -\frac{t}{2} \sum_{i=0}^{k} \|\nabla f(x^{(i)})\|_2^2.$$

Rearranging and using $f(x^{(k+1)}) \geq f^*$:

$$\sum_{i=0}^{k} \|\nabla f(x^{(i)})\|_2^2 \leq \frac{2(f(x^{(0)}) - f^*)}{t}.$$

The minimum gradient norm satisfies:

$$(k+1) \min_i \|\nabla f(x^{(i)})\|_2^2 \leq \sum_{i=0}^{k} \|\nabla f(x^{(i)})\|_2^2.$$

Combining these gives the result. □

4 Key Observations

- **Rate:** GD achieves $O(1/\sqrt{k})$ convergence rate, requiring $O(1/\epsilon^2)$ iterations to find ϵ-stationary points.
- **Optimality:** This rate is tight; no deterministic first-order method can improve it for this problem class.
- **Step Size:** The step size $t \leq 1/L$ ensures monotonic decrease of the objective.

5 Example: Nonconvex Function

Consider $f(x) = x^2 + 3\sin(x)$ with $L = 5$. GD updates:

$$x^{(k+1)} = x^{(k)} - t(2x^{(k)} + 3\cos(x^{(k)})).$$

Figure 1 (simulated) shows convergence to stationary points.

6 Comparison with Convex Case

	Nonconvex	**Strongly Convex**
Convergence	$O(1/\sqrt{k})$	$O(\gamma^k)$ (linear)
Goal	$\|\nabla f(x)\| \leq \epsilon$	$f(x) - f^* \leq \epsilon$
Step Size	$t \leq 1/L$	$t \leq 2/(m+L)$

7 Conclusion

For nonconvex optimization, GD efficiently finds stationary points but cannot guarantee global minima. The $O(1/\epsilon^2)$ complexity is fundamental, and acceleration methods (e.g., Nesterov) do not improve this rate in the nonconvex setting.