Gradient descent on nonconvex smooth functions

Mert Pilanci

EE364b, Stanford University

Goal

In convex optimization, gradient descent is often analyzed through distance to an optimizer.

In nonconvex optimization, global optimality is usually out of reach. The natural target becomes **stationarity**:

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

We will study

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

for differentiable nonconvex objectives.

Main questions:

- what does smoothness mean
- why does it make gradient descent work at all
- what can still be proved in the nonconvex setting

What smoothness says

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is L-**smooth** if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \qquad \forall x, y \in \mathbb{R}^n.$$

Equivalent interpretation: the gradient cannot change arbitrarily fast.

For twice differentiable f, this is the same as

$$\|\nabla^2 f(x)\|_2 \leq L \qquad \forall x.$$

So L controls the local curvature scale. Large L means sharp bending. Small L means the graph bends gently.

The quadratic upper model

The most useful consequence of L-smoothness is

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

Interpretation:

- the first-order Taylor model is not exact
- but its error is at most quadratic
- smoothness gives a global upper bound on how wrong the linearization can be

This is exactly the statement needed to analyze one gradient step.

without smoothness, a local linear model may be too optimistic and a step can overshoot badly

Why the step size $1/L$ appears

Take the upper model at x and minimize it over y:

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

The minimizer is

$$y = x - \frac{1}{L}\nabla f(x).$$

So one gradient step with step size $1/L$ is exactly **minimizing the quadratic upper bound supplied by smoothness**.

This is a strong motivation for the algorithm:

- the linear term wants to move along $-\nabla f(x)$
- the quadratic term penalizes moving too far
- L determines the correct trust in the local linear model

Descent lemma

Let $g^{(k)} = \nabla f(x^{(k)})$ and

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}.$$

Apply the smoothness upper bound with $y = x^{(k)} - \alpha_k g^{(k)}$:

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \alpha_k \|g^{(k)}\|_2^2 + \frac{L\alpha_k^2}{2}\|g^{(k)}\|_2^2.$$

Hence if $0 < \alpha_k \leq 1/L$,

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{\alpha_k}{2}\|g^{(k)}\|_2^2.$$

Every nonstationary step decreases the objective.

Other implications of smoothness

Smoothness does more than give descent.

If f is L-smooth, then for all x, y,

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

So first-order approximation error is second order.

Also,

$$f\left(x - \frac{1}{L} \nabla f(x)\right) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|_2^2.$$

This gives a direct link between decrease in function value and gradient norm.

Consequences:

- large gradient norm forces significant decrease
- if function values stop decreasing much, gradients must be getting small

Global stationarity guarantee

Assume

$$f_{\inf} = \inf_x f(x) > -\infty, \qquad 0 < \alpha_k \leq 1/L.$$

Summing the descent lemma,

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

Hence

$$\sum_{i=1}^{\infty} \alpha_i \|\nabla f(x^{(i)})\|_2^2 \leq 2\Big(f(x^{(1)}) - f_{\inf}\Big) < \infty.$$

Therefore:

- if $\inf_i \alpha_i > 0$, then $\|\nabla f(x^{(k)})\|_2 \to 0$
- more generally, if $\sum_i \alpha_i = \infty$, then infinitely many iterates have very small gradient

Best-so-far bound

Define

$$g_{\text{best}}^{(k)} = \min_{i=1,\ldots,k} \|\nabla f(x^{(i)})\|_2^2.$$

Then

$$g_{\text{best}}^{(k)} \sum_{i=1}^{k} \alpha_i \leq \sum_{i=1}^{k} \alpha_i \|\nabla f(x^{(i)})\|_2^2 \leq 2\Big(f(x^{(1)}) - f_{\text{inf}}\Big).$$

So

$$g_{\text{best}}^{(k)} \leq \frac{2\Big(f(x^{(1)}) - f_{\text{inf}}\Big)}{\sum_{i=1}^{k} \alpha_i}.$$

For the constant step size $\alpha_i = 1/L$,

$$\min_{i=1,\ldots,k} \|\nabla f(x^{(i)})\|_2^2 \leq \frac{2L\Big(f(x^{(1)}) - f_{\text{inf}}\Big)}{k}.$$

guarantee: $O(1/k)$ in squared gradient norm.

What this does not say

The guarantee is only about stationarity.

It does **not** distinguish among:

- local minima
- local maxima
- saddle points

A point with $\nabla f(x) = 0$ may still be undesirable.

So after proving stationarity, the next question is dynamical:

which stationary points does gradient descent actually approach?

To build intuition, it helps to inspect concrete nonconvex smooth examples.

Example 1: a smooth double well

Consider

$$f(x) = \frac{1}{4}(x^2 - 1)^2.$$

Then

$$f'(x) = x(x^2 - 1), \qquad f''(x) = 3x^2 - 1.$$

Stationary points:

$$x \in \{-1, 0, 1\}.$$

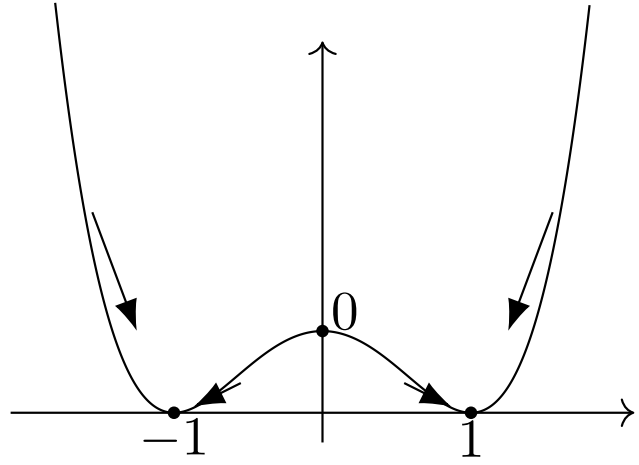
Classification:

- $x = \pm 1$ are local minima
- $x = 0$ is a strict saddle / local maximum in one dimension

Gradient descent becomes

$$x^{(k+1)} = x^{(k)} - \alpha x^{(k)}\Big((x^{(k)})^2 - 1\Big).$$

Double well dynamics



Behavior:

- if $x^{(0)} > 0$, the iterates move toward $+1$
- if $x^{(0)} < 0$, the iterates move toward -1
- $x = 0$ is unstable: exactly starting there keeps you there, any perturbation escapes

This is the simplest picture of attraction basins in a smooth nonconvex problem.

Example 2: monkey saddle

Consider the two-dimensional smooth nonconvex function

$$f(x, y) = x^3 - 3xy^2.$$

Its gradient is

$$\nabla f(x, y) = \begin{bmatrix} 3x^2 - 3y^2 \\ -6xy \end{bmatrix}.$$

At $(0, 0)$ we have $\nabla f(0, 0) = 0$, but this point is not a minimum. It is a highly degenerate saddle.

Near the origin:

- some directions go downhill
- some directions go uphill
- some directions are almost flat

So small gradients do not imply a good point.

Monkey saddle dynamics

Gradient descent is

$$x^{(k+1)} = x^{(k)} - 3\alpha\Big((x^{(k)})^2 - (y^{(k)})^2\Big),$$
$$y^{(k+1)} = y^{(k)} + 6\alpha x^{(k)} y^{(k)}.$$

Qualitative picture:

- if initialized exactly at $(0, 0)$, the iterates stay there
- generic perturbations move away from the saddle
- the trajectory depends strongly on direction because curvature is anisotropic

This shows why the theorem only guarantees stationarity. Stationary points can be unstable saddles.

Example 3: Many saddles

Consider

$$f(x) = \frac{1}{4}\sum_{i=1}^{n}(x_i^2 - 1)^2.$$

Then

$$\nabla f(x)_i = x_i(x_i^2 - 1).$$

This is separable, so each coordinate evolves independently:

$$x_i^{(k+1)} = x_i^{(k)} - \alpha x_i^{(k)}\Big((x_i^{(k)})^2 - 1\Big).$$

Stationary points are all vectors with coordinates in $\{-1, 0, 1\}$.

Most of these are saddles; the only minima are the 2^n sign vectors in $\{-1, 1\}^n$.

This already shows exponential multiplicity of stationary points in a smooth nonconvex problem.

What have we shown?

The analysis of nonconvex smooth gradient descent uses only three ingredients:

- lower boundedness: $f(x) \geq f_{\inf}$
- smoothness: a quadratic upper bound on local model error
- a safe step size: $\alpha_k \leq 1/L$

From these, we get:

- monotonic decrease of function values
- summability of $\alpha_k \|\nabla f(x^{(k)})\|_2^2$
- an $O(1/k)$ best-so-far stationarity bound

What we do **not** get:

- global optimality
- avoidance of saddles
- rates to a specific local minimum

Takeaways

Smoothness gives:

- a global quadratic upper model
- a principled step size scale $1/L$
- descent at every nonstationary iterate
- a complexity guarantee for reaching an approximately stationary point

And the examples show the real geometry behind the theorem:

- attraction basins of local minima
- unstable saddles
- many stationary points in high dimensions