Nonconvex Optimization

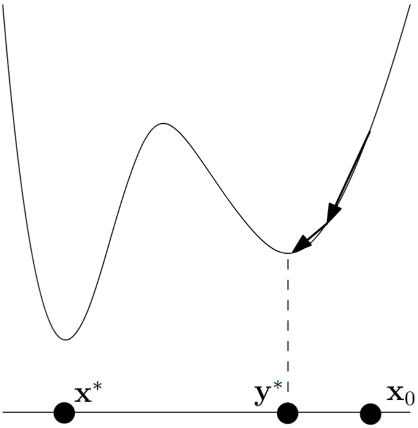
Axel Böhm

November 17, 2021

1. Introduction
2. GD for linear networks

Gradient Descent in the nonconvex world

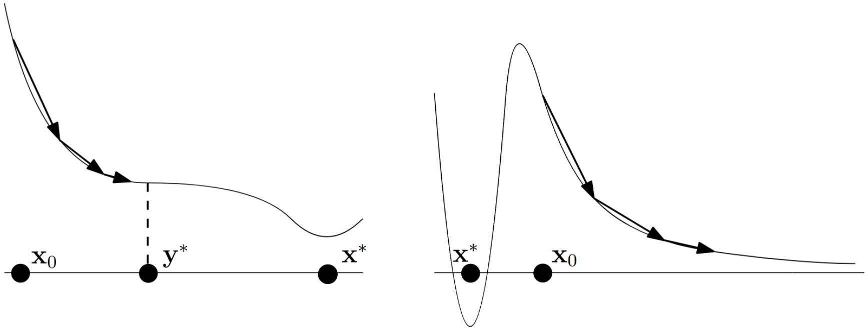
may get stuck in a local minimum and miss the global minimum



Gradient Descent in the nonconvex world II

Even if there is a unique local minimum (equal to the global minimum), we

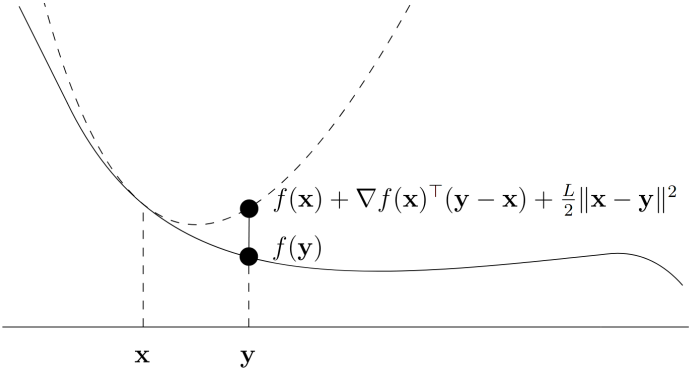
- may get stuck in a saddle point;
- run off to infinity;
- possibly encounter other bad behaviors.



Smooth (but not necessarily convex) functions

Recall: A differentiable $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth over a convex set X if

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



Bounded Hessians $\Rightarrow$ smooth

Lemma

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable and

$$\|\nabla^2 f(x)\| \leq L$$

where $\|\cdot\|$ is spectral norm. Then f is L-smooth

Examples:

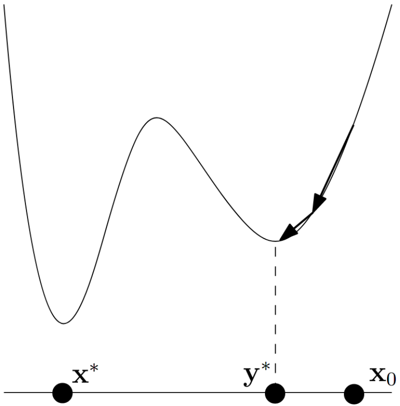
- all quadratic functions $f(x) = x^T A x + b^T x + c$
- $f(x) = \sin(x)$ (many global minima)

Gradient descent on smooth functions

Will prove: $\|\nabla f(x_k)\|^2 \to 0$ for $k \to \infty$ ...

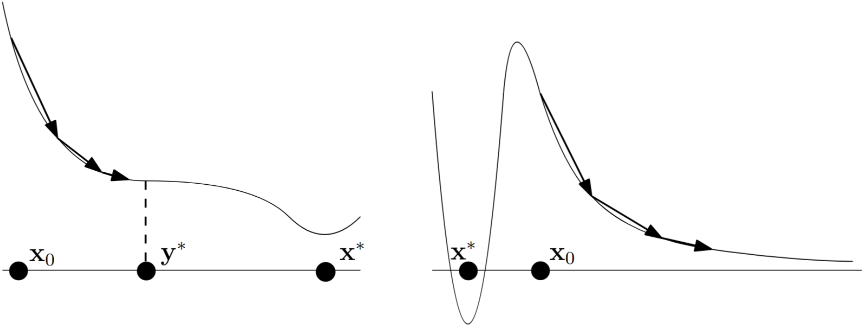
...at the same rate as $f(x_k) - f(x^*) \to 0$ in the convex case.

- $f(x_k) - f(x^*)$ itself may not converge to 0 in the nonconvex case:



What does $\|\nabla f(x_k)\|^2 \to 0$ mean?

- May or may not mean convergence to a critical point $\nabla f(y^*) = 0$
- critical point might not be even local minimum



Figure

Gradient descent on smooth (not necessarily convex) functions

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth with a global minimum x^*. Choosing stepsize $\alpha := \frac{1}{L}$ gradient descent yields

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 \leq \frac{2L}{K}\left(f(x_0) - f(x^*)\right).$$

In particular, same bound hold for "best" iterate

$$\min_{0\leq k\leq K-1} \|\nabla f(x_k)\|^2 \leq \frac{2L}{K}\left(f(x_0) - f(x^*)\right)$$

and

$$\lim_{k\to\infty} \|\nabla f(x_k)\|^2 = 0.$$

Gradient descent on smooth functions II: Proof

Smoothness gives:

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

Use $y = x_{k+1}$ and $x = x_k$

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), -\nabla \alpha f(x_k) \rangle + \frac{L\alpha^2}{2} \|\nabla f(x_k)\|^2$$

to obtain sufficient decrease:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

Proof II

sufficient decrease:

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

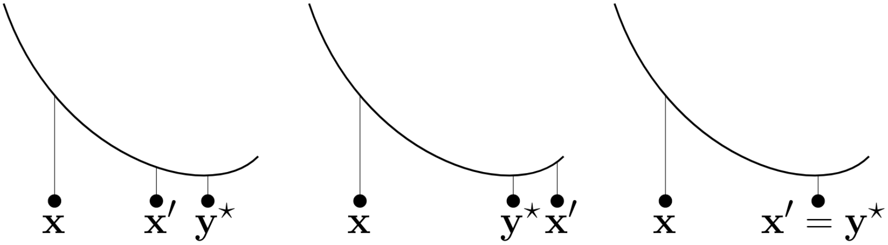
Sum up from $k = 0, 1, \ldots, K-1$ to get

$$\frac{1}{2L}\sum_{k=0}^{K-1}\|\nabla f(x_k)\|^2 \le f(x_0) - f(x_k) \le f(x_0) - f(x^*).$$

Multiply by $2L/K$ to get the statement of the theorem.

No overshooting

Under the **smoothness** assumption and appropriate stepsize $\alpha \leq 1/L$, GD cannot pass a critical point:



$\mathbf{x}' = \mathbf{x} - \gamma \nabla f(\mathbf{x}), \gamma < 1/L$ overshooting may happen with $\gamma = 1/L$

Trajectory Analysis

Even if the "landscape" (graph) of a nonconvex function has local minima, saddle points, and flat parts, gradient descent may avoid them and still converge to a global minimum. For this, one needs a good starting point and some theoretical understanding of what happens when we start there—this is trajectory analysis.

Linear models with several outputs

Recall: Learning linear models

- n inputs $x_1, \ldots, x_n$, where each input $x_i \in \mathbb{R}^d$
- n outputs $y_1, \ldots, y_n \in \mathbb{R}$
- Hypothesis (after centering):

$$y_i \approx w^T x_i,$$

for a weight vector $\boldsymbol{w} = (w_1, ..., w_d) \in \mathbb{R}^d$ to be learned.

Now more than one output value:

- n outputs $y_1, ..., y_n$, where each output $y_i \in \mathbb{R}^m$
- Hypothesis:

$$y_i \approx W x_i,$$

for a weight matrix $W \in \mathbb{R} m \times d$ to be learned