Session 12: Gradient descent

Optimization and Computational Linear Algebra for Data Science

Léo Miolane

Contents

Gradient descent

Gradient descent algorithm

Goal: minimize a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$.

Starting from a point $x_0 \in \mathbb{R}^n$, perform the updates:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t).$$

"step size"
"learning rate"

Cauchy (~1850)

$$x_t \xrightarrow{-\alpha_t \nabla f(x_t)} x_{t+1} \longrightarrow$$

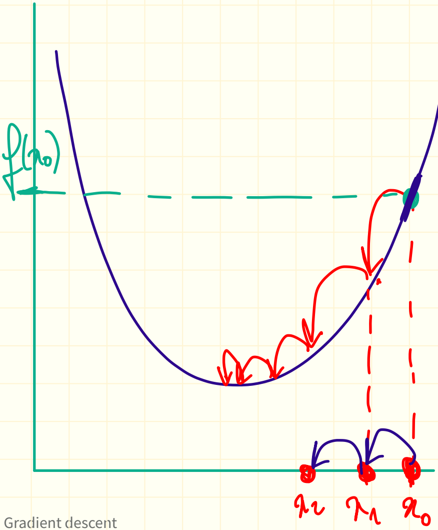
IDEA: $f(x_t+h) \simeq f(x_t) + \underbrace{h \cdot \nabla f(x_t)}$

$$f(x_{t+1}) \simeq f(x_t) \underbrace{- \alpha_t \|\nabla f(x_t)\|^2}_{\leq 0}$$

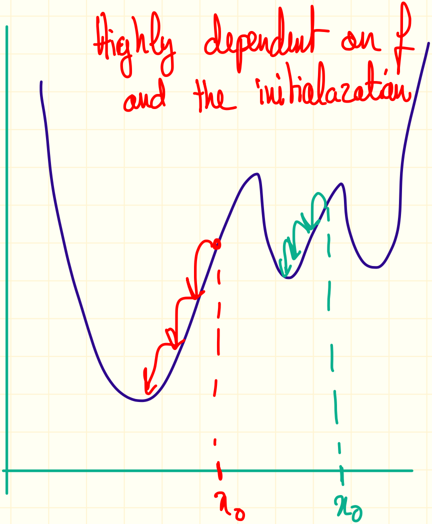
for α_t small enough.

Convex vs non-convex

Convex



Non-convex



Numerical observations

- If the step size α is small enough, gradient descent converges to $x^\star$ **but** this may take a while.
- If the step size α is large, gradient descent moves faster **but** it may oscilate or even diverge.
- The convergence is faster when the eigenvalues of the Hessian H_f are of close to each other.

Convergence analysis for convex functions

Smoothness and strong convexity

Definition

Given $L, \mu > 0$, we say that a twice-differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$ is

- L-smooth if for all $x \in \mathbb{R}^n$, $\lambda_{\max}(H_f(x)) \leq L$.
- μ-strongly convex if for all $x \in \mathbb{R}^n$, $\lambda_{\min}(H_f(x)) \geq \mu$.

Remark: if f is $\left| \begin{array}{l} L\text{-smooth} \\ \mu\text{-strongly convex} \end{array} \right.$ then:

$$\underbrace{f(x) + \nabla f(x) \cdot h}_{} + \frac{\mu}{2}\|h\|^2 \leq f(x+h) \leq \underbrace{f(x) + \nabla f(x) \cdot h}_{} + \frac{L}{2}\|h\|^2$$

Speed for L-smooth functions

Proposition

Assume that f is convex, L-smooth and admits a global minimizer $x^\star \in \mathbb{R}^n$. Then, gradient descent with constant step size $\alpha_t = 1/L$ verifies:

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|^2}{t+4}. \leq \frac{\text{Constant}}{t}$$

Why step size $\alpha_t = \frac{1}{L}$?

$$f(x_t + h) \leq \underbrace{f(x_t) + \nabla f(x_t) \cdot h + \frac{L}{2}\|h\|^2}_{\text{this is minimal for } \boxed{h = -\frac{1}{L}\nabla f(x_t)}}$$

$$\rightarrow x_{t+1} = x_t + h = x_t - \frac{1}{L}\nabla f(x_t)$$

L-smooth + μ-strongly cvx functions

Theorem

Assume that f is convex, L-smooth and μ-strongly convex. Then, gradient descent with constant step size $\alpha_t = 1/L$ verifies:

$$f(x_t) - f(x^\star) \leq \underbrace{\left(1 - \frac{\mu}{L}\right)^t}_{\leq e^{-\frac{\mu}{L}t}} \underbrace{(f(x_0) - f(x^\star))}_{\text{Constant}}.$$

Remark: • GD with step size $\alpha_t = \frac{1}{L}$ is "adaptive" to strong convexity"

• The quantity $\kappa = \frac{L}{\mu} \geq 1$ is called the "condition number"

$\kappa \nearrow$ the convergence speed $\searrow$

Proof

Recall: $f(x+h) \leq f(x) + \nabla f(x) \cdot h + \frac{L}{2}\|h\|^2$

Apply it for $x = x_t$, $h = -\frac{1}{L}\nabla f(x_t)$

- $\rightarrow$ $\boxed{f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2}$
- By strong convexity: $f(x_t) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x_t)\|^2$ (exercise)

Combining the two inequalities:

$$f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) - \frac{\mu}{L}\left(f(x_t) - f(x^*)\right)$$

$$= \left(f(x_t) - f(x^*)\right)\left(1 - \frac{\mu}{L}\right)$$

∎

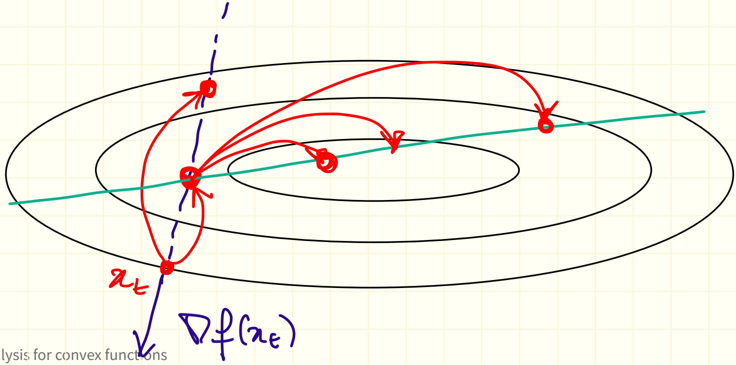
Choosing the step size

Backtracking line search

Start with $\alpha = 1$ and while

$$f\left(x_t - \alpha \nabla f(x_t)\right) \geq f(x_t) - \frac{\alpha}{2}\|\nabla f(x_t)\|^2,$$

update let's say $\alpha = 0.8\alpha$.

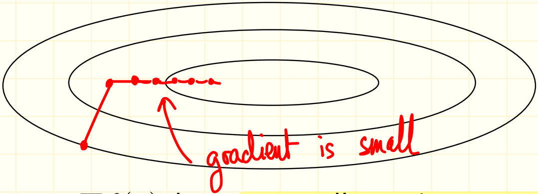


Improvements

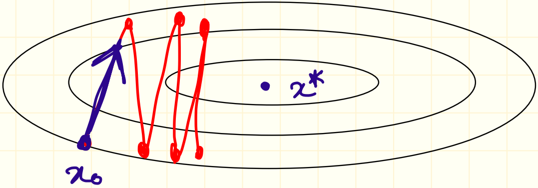
Issues with gradient descent

When the condition number $\kappa = L/\mu$ is large:

1. the norm $\|\nabla f(x)\|$ is sometimes too small.
   $\rightarrow$ gradient descent steps are too small.



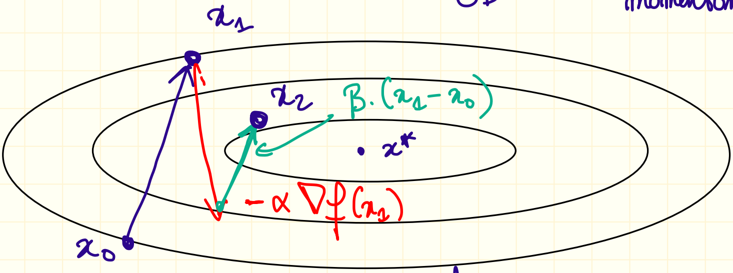
2. The vector $-\nabla f(x)$ does « not really » points towards the minimizer $x^\star$.
   $\rightarrow$ gradient descent oscilates.



Gradient descent + momentum

Idea: mimic the trajectory of an « heavy ball » that goes down the slope:

$$x_{t+1} = x_t + v_t \quad \text{where} \quad v_t = \underbrace{-\alpha_t \nabla f(x_t)}_{\text{GD}} + \underbrace{\beta_t \overbrace{v_{t-1}}^{x_t - x_{t-1}}}_{\text{momentum}}.$$



Momentum damps the oscillations

+ accumulate momentum in the horizontal direction.

Newton's method

Assume that f is μ-strongly convex and L-smooth.

eigenvalues $\geq \mu > 0$
hence invertible.

Newton's method perform the updates:

$$x_{t+1} = x_t - H_f(x_t)^{-1}\nabla f(x_t).$$

IDEA: $f(x_t + h) \simeq \underbrace{f(x_t) + h \cdot \nabla f(x_t) + \frac{1}{2} h^T H_f(x_t) h}_{}.$

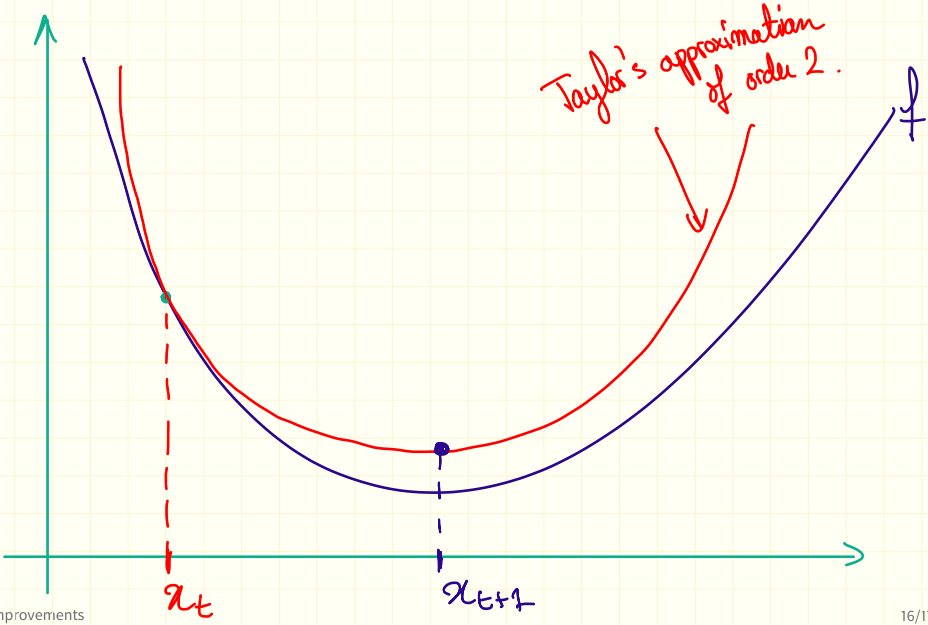
$\stackrel{\text{def}}{=} Q(h)$, minimal for $h = -H_f(x_t)^{-1}\nabla f(x_t)$

Proof

- Q is convex. $\left(H_Q(h) = H_f(x_t) \leftarrow \text{PSD}\right)$
- Let's solve $\nabla Q(h) = 0$: $\nabla f(x_t) + H_f(x_t) h = 0$

$$\rightarrow h = -H_f(x_t)^{-1}\nabla f(x_t)$$

Graphical interpretation



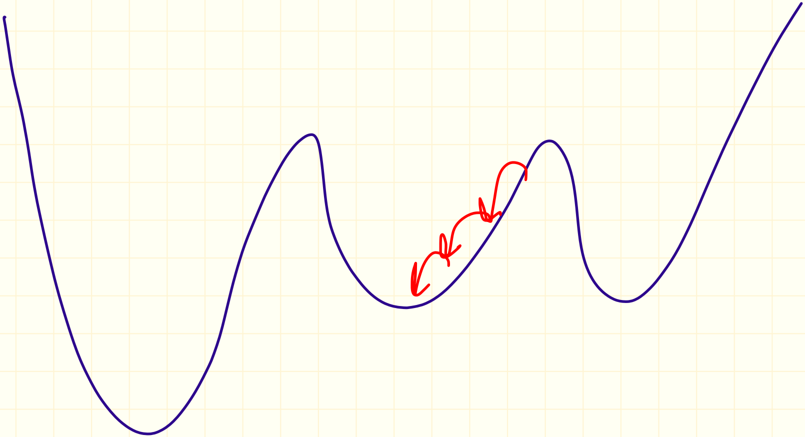
Advantages and drawbacks

- Extremly fast there exists $C, \rho > 0$ such that
$$\boxed{\|x_t - x^\star\|^2 \leq \underline{C} e^{-\rho 2^t}.}$$
$e^{-\rho 2^t}$
- Computationally expensive: requires $\sim n^3$ operations to compute the inverse of the $n \times n$ matrix $H_f(x_t)$.
- In non-convex setting, Newton's method gets attracted by any critical points (which could be saddle points/maximas...).

Quasi-Newton methods: try to approximate $H_f(x_t)^{-1}$ by matrices B_t that are easier to compute.

$x_{t+1} = x_t - B_t \nabla f(x_t)$

Questions?



Questions?

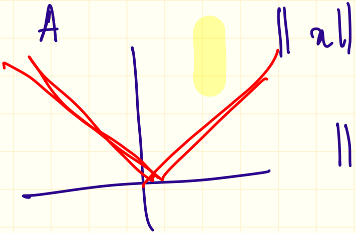
$$A = U \Sigma V^T$$

$$\min_{x \in \mathbb{R}^n} \|Ax - y\|^2 = \min_{u \in \operatorname{Im}(A)} \|u - y\|^2$$

$$\hookrightarrow \quad x^* = A^+ y$$

$$u^* = A A^+ y$$

$$u^* = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{pmatrix} \begin{pmatrix} - & u_1 & - \\ & \vdots & \\ - & u_r & - \end{pmatrix} y.$$



$$\|x\| = |x|$$