Session 12: Gradient Descent

Optimization and Computational Linear Algebra for Data Science

BYOD & VANDENBHERGE - CHAP 9

FRANCIS BACH → CHAP 5.

Marylou Gabrié (based on material by Léo Miolane)

Contents

Clarification about saddle points

- A critical point is always either a local minimum or a local maximum, a saddle point.

- **Definitions:**
  - A critical point x^* is a local extrema for a small $\delta > 0$ for any $x \in B(x^*, \delta)$, $f(x)$ is bigger/smaller than $f(x^*)$.
  - If a critical point is not a local extrema, then it is a saddle point.

h

- **Caracterizations (sufficient but not necessary conditions):** Examine Hessian $H_f(x^*)$:
  - is positive definite $\Rightarrow$ local minimum.
  - has strictly positive and strictly negative eigenvalues $\Rightarrow$ saddle

1. 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:

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

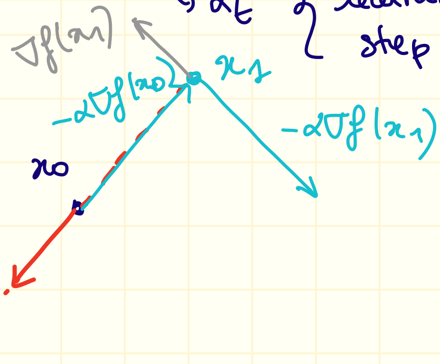
α_t $\left\{ \begin{array}{l} \text{learning rate} \\ \text{step size} \end{array} \right.$

$\alpha_t \in \mathbb{R}$

$\alpha_t > 0$

CAUCHY (~1850)

$n=2$



IDEA: $f(x_t + h) = f(x_t) + h \cdot \nabla f(x_t)$

$h = -\alpha \nabla f(x_t)$

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

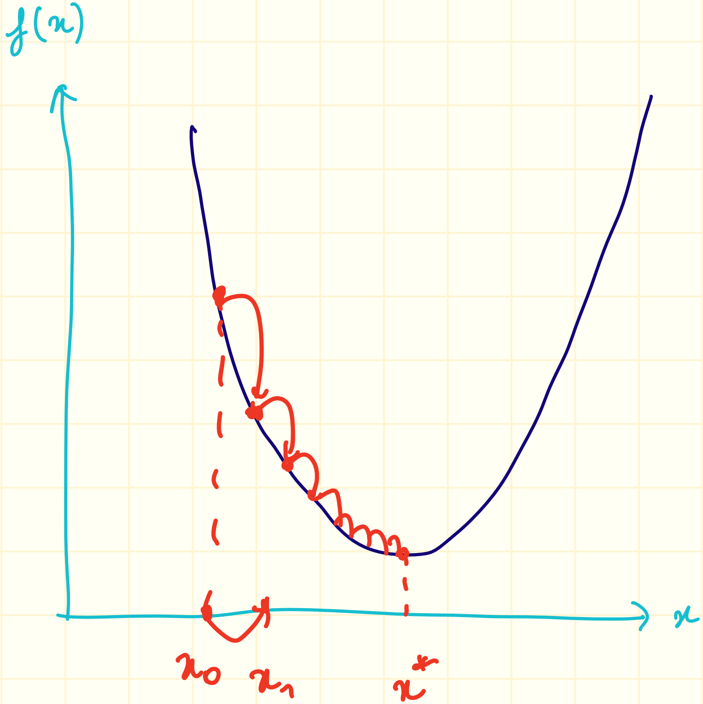
for α small.

Convex vs non-convex

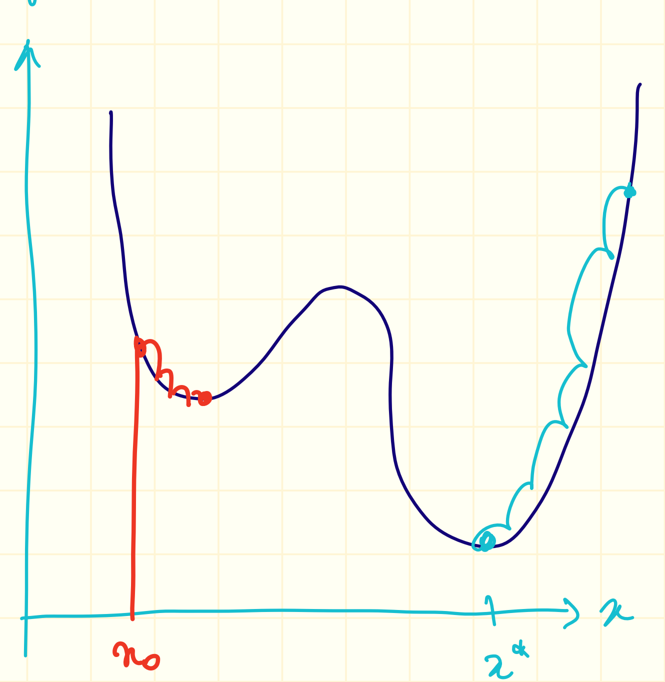
$m=1$

highly dependent of the initialization —

Convex

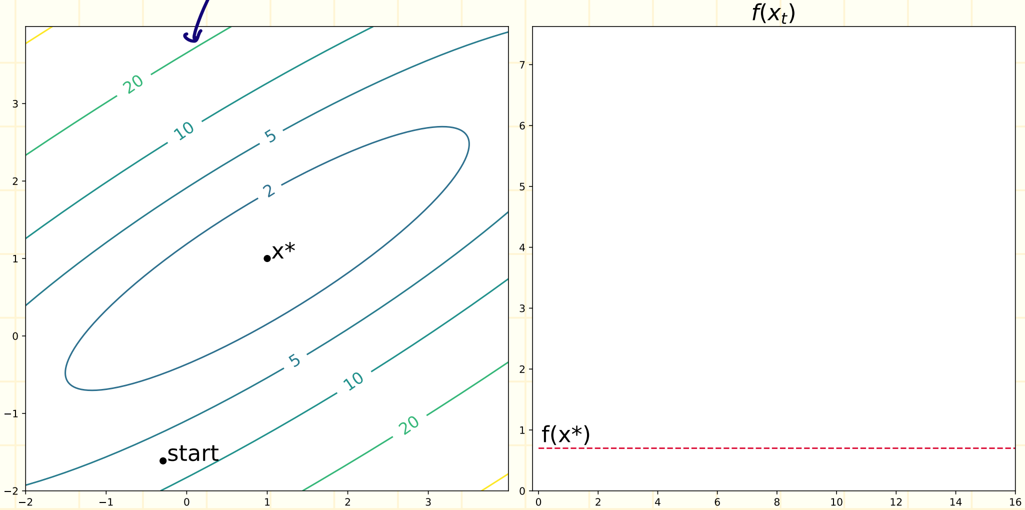


Non-convex

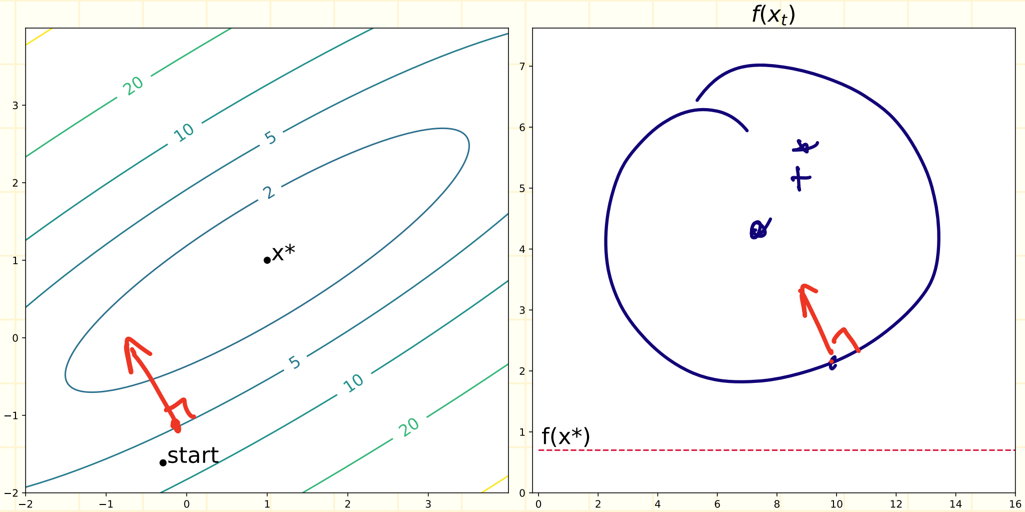


Numerical observations

$n = 2$. contour lines .



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.

2. 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

function not to be infinitely steep.

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

thu 9 or 10

Remark: if f convex if $\begin{cases} L \text{ smooth} \\ \mu \text{ strongly convex} \end{cases}$, then for h small:

$$\frac{1}{2} h^T H_f(x) h$$
$$\wedge|$$

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

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:

$$\underbrace{f(x_t) - f(x^\star)}_{\text{how close in terms of function value we are after } t \text{ step of GD}} \leq \frac{2L\overbrace{\|x_0 - x^\star\|^2}^{\text{initial distance to the solution}}}{t+4}. = O\left(\frac{1}{t}\right)$$

Why step $\alpha_t = \frac{1}{L}$: $f(x_t + h) \leq \underbrace{f(x_t) + \nabla f(x_t) \cdot h + \frac{L}{2}\|h\|^2}_{\text{min w.r.t } h}$

$$h^* = -\underbrace{\frac{1}{L}}_{=\alpha}\nabla f(x_t)$$

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:

$$f(x_t) - f(x^\star) \leq \Big(1 - \underbrace{\frac{\mu}{L}}_{1/K}\Big)^t \underbrace{(f(x_0) - f(x^\star))}_{\text{distance to solution at initialization}}. = O\big(e^{-\mu/L\, t}\big)$$

Remark: $K = \dfrac{L}{\mu} \geqslant \dfrac{\max_x \lambda_{\max} H_f(x)}{\min_x \lambda_{\min} H_f(x)} \geqslant 1.$ CONDITION NUMBER

$\searrow$ speed of convergence if $K \nearrow$

Proof

Recall that $f(x+h) \leqslant f(x) + \nabla f(x) \cdot h + L/2 \, \|h\|^2$

Apply this for: $x = x_t$ and $h = -\frac{1}{L} \nabla f(x_t)$

$$\Rightarrow f(x_{t+1}) \leqslant f(x_t) - \frac{1}{2L} \|\nabla f(x_t)\|^2 \quad (1)$$

exercize

By strong convexity: $f(x_t) - f(x^*) \leqslant \frac{1}{2\mu} \|\nabla f(x_t)\|^2 \quad (2)$

global minimum

$(1) - f(x^*)$

$(1) + (2)$: $f(x_{t+1}) - f(x^*) \leqslant f(x_t) - f(x^*) - \frac{1}{2L} \|\nabla f(x_t)\|^2$

$+ (2)$

(2) $\|\nabla f(x_t)\|^2 \geqslant 2\mu (f(x_t) - f(x^*))$

$-\frac{1}{2L} \|\nabla f(x_t)\|^2 \leqslant -\frac{\mu}{L} (f(x_t) - f(x^*))$

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

⊕ FORMAL WRITING OF THE INDUCTION

Choosing the step size

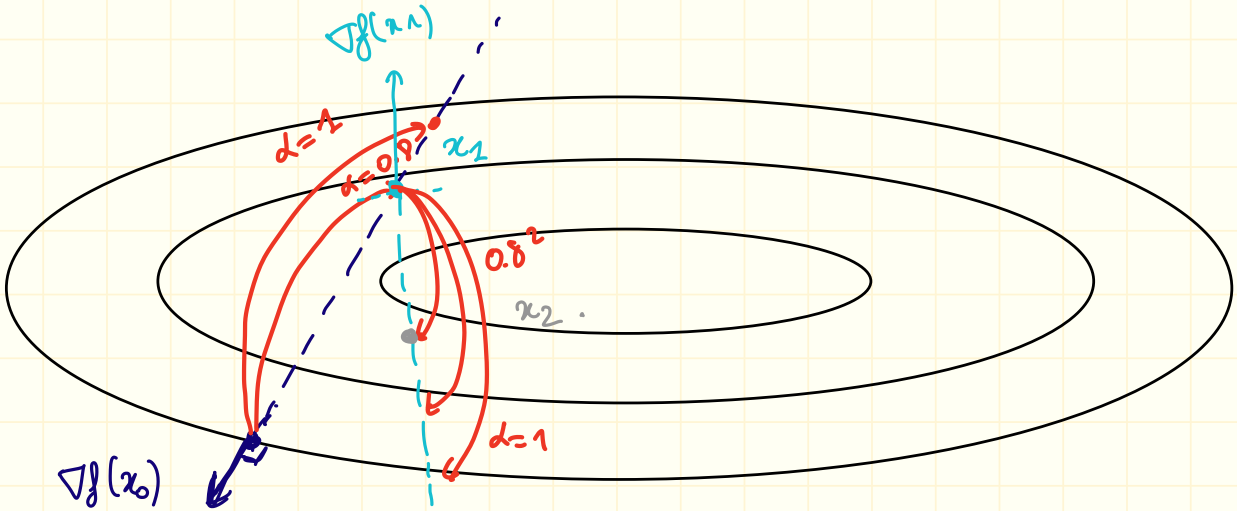
Backtracking line search

decrease by at least $\|\frac{\alpha}{2}\nabla f(x_t)\|^2$

Start with $\alpha = 1$ and while

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

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



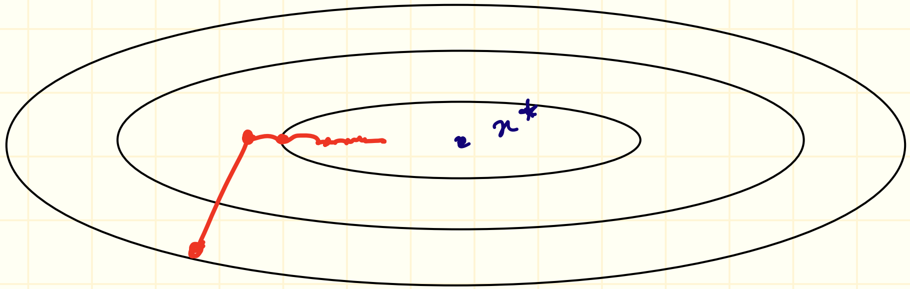
3. Improvements

Issues with gradient descent

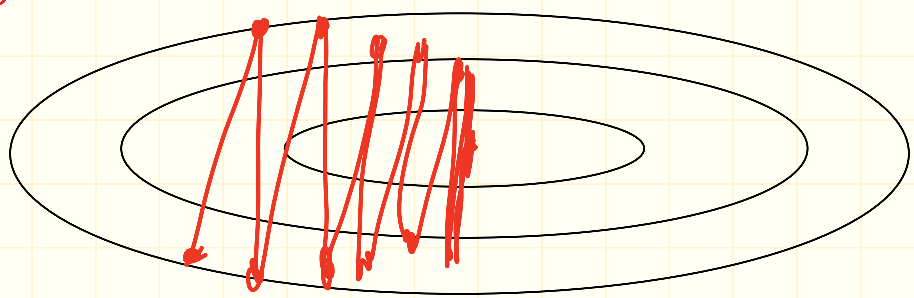
When the condition number $\kappa = L/\mu$ is large:

→ or learning rate

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.

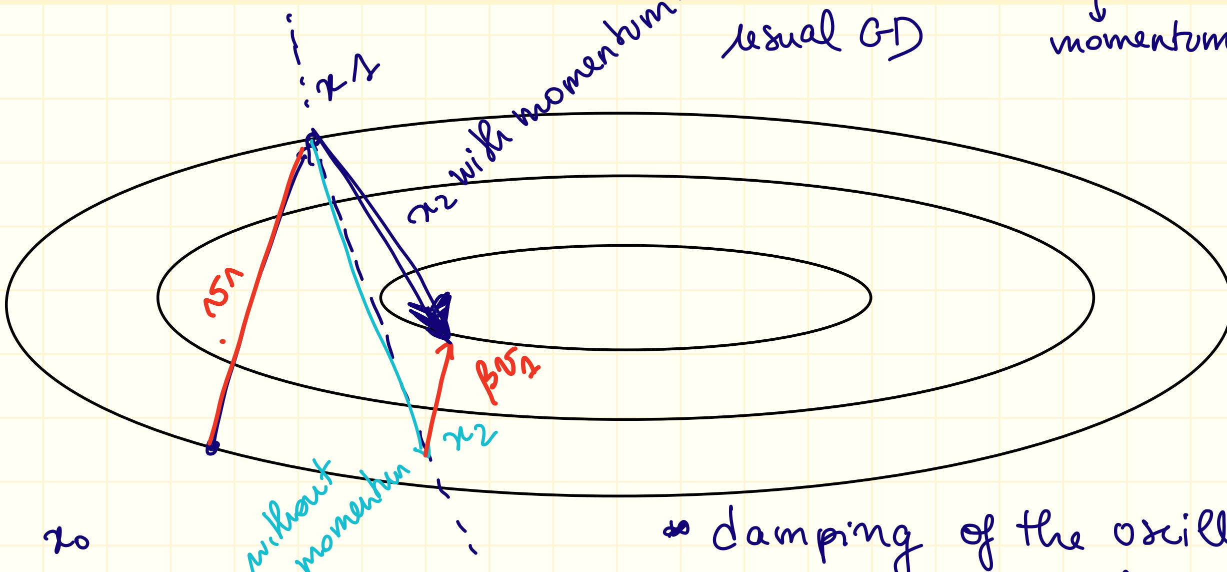


Gradient descent + momentum

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

parameter β.

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



* damping of the oscillations
* promote direction towards minimum.

Newton's method

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

Newton's method perform the updates:

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

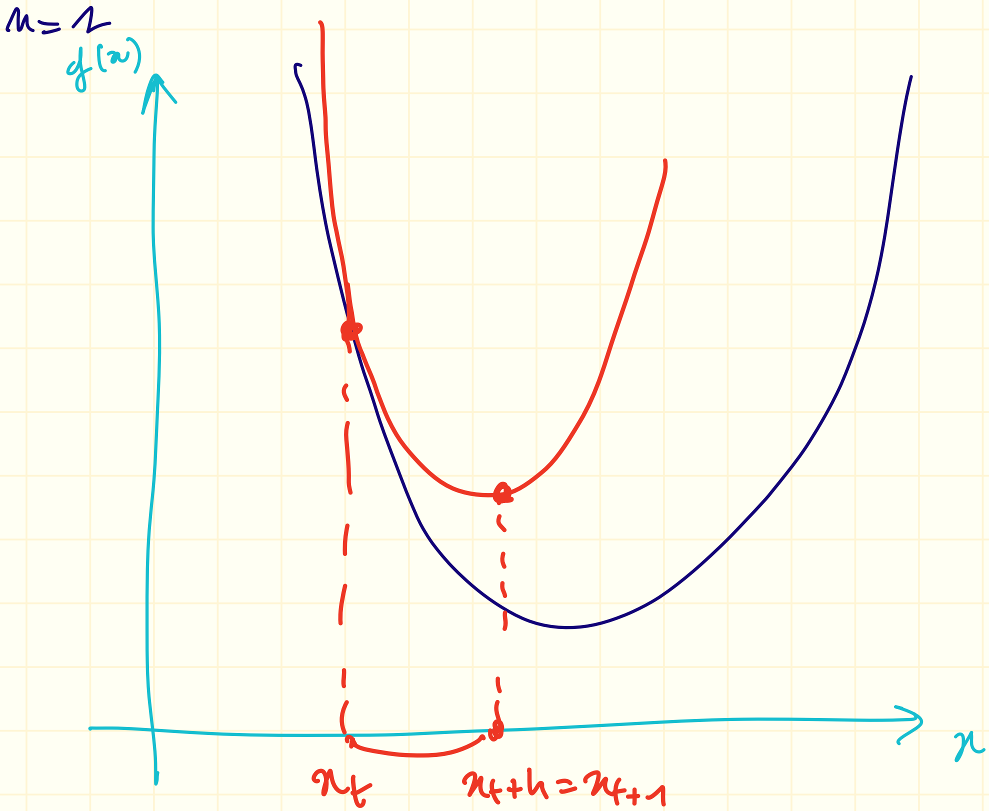
IDEA: Optimizing the learning rate by considering the second order Taylor expansion.

$$f(x_{t+1}) = f(x_t + h) = \underbrace{f(x_t) + h \cdot \nabla f(x_t) + \frac{1}{2} h^T H_f(x_t) h}_{= Q(h)}.$$

- Q is convex $\quad H_Q(h) = H_f(x_t) \quad H_Q(h)$ is PSD.
- $\nabla Q(h) = 0 \quad \Rightarrow \quad \nabla f(x_t) + H_f(x_t) h = 0$

$$\Rightarrow \quad h = - H_f^{-1}(x_t) \nabla f(x_t).$$

Graphical interpretation



Advantages and drawbacks

- Extremly fast there exists $C, \rho > 0$ such that

$$\|x_t - x^\star\|^2 \leq Ce^{-\rho 2^t}.$$

- 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.

Questions?

Questions?