ROB 101 - Fall 2021

Optimization: First-Order Unconstrained (Gradient Descent)

November 15, 2021

UNIVERSITY OF MICHIGAN

Learning Objectives

- Mathematics is used to describe physical phenomena, pose engineering problems, and solve engineering problems.
- We show how linear algebra and computation allow you to use a notion of "optimality" as a criterion for selecting among a set of solutions to an engineering problem.

- Arg min should be thought of as another function in your toolbox,
$$x^* = \arg\min_{x \in \mathbb{R}^m} f(x).$$
- Extrema of a function occur at places where the function's first derivative vanishes.
- The gradient of a function points in the direction of maximum rate of growth.

We define a norm ball in $\mathbb{R}^m$ as

$$\mathcal{B}(x_c,r) := \{x \in \mathbb{R}^m : \|x - x_c\| \leq r\}.$$

- Objective function $f : \mathbb{R}^m \to \mathbb{R}$ and decision variable $x \in \mathbb{R}^m$

$$\underset{x\in\mathbb{R}^m}{\text{minimize}}\ f(x), \qquad x^* = \underset{x\in\mathbb{R}^m}{\arg\min}\, f(x)$$

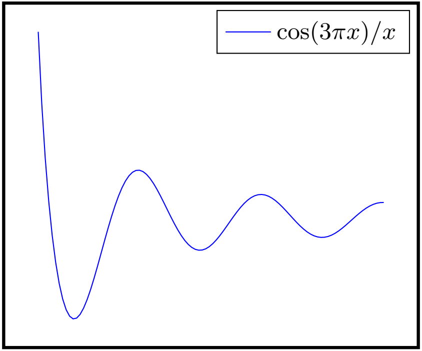
- Global minimum

$$f(x^\star) \le f(x) \qquad \underbrace{\text{for all } x \in \mathbb{R}^m}_{\text{global}}$$

- Local minimum

$$f(x^*) \le f(x) \qquad \underbrace{\text{for all } x \in \mathcal{B}_{r>0}(x^*)}_{\text{local}}$$

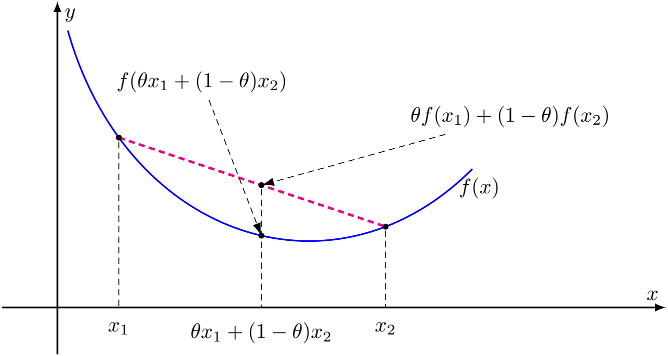
Example



$f : \mathbb{R}^m \to \mathbb{R}$ $(\mathrm{dom} f = \mathbb{R}^m)$ is convex iff:

1. For all $x_1, x_2 \in \mathbb{R}^m$ and all $\theta \in [0,1]$:

$$f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta) f(x_2)$$



$f : \mathbb{R}^m \to \mathbb{R}$ $(\mathrm{dom} f = \mathbb{R}^m)$ is convex iff:

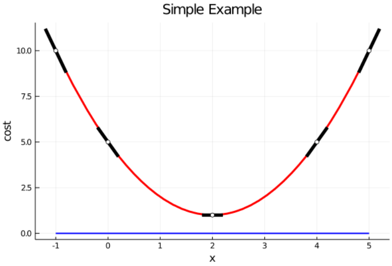
1. For all $x_1, x_2 \in \mathbb{R}^m$ and all $\theta \in [0,1]$:

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2)$$

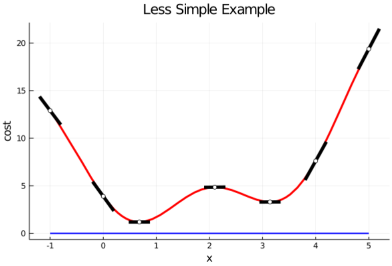
2. First-order condition: For all $x, x_0 \in \mathbb{R}^m$:

$$f(x) \geq f(x_0) + \nabla f(x_0)(x - x_0)$$

The Derivatives of the Objective Functions



(a)



(b)

Fact

First-order necessary condition for x to be a local extremum of f is

$$\nabla f(x) = 0.$$

Objective function: $f(x) = \frac{1}{2}\|Ax - b\|^2$

- Gradient: $\nabla f(x) = A^\mathsf{T} Ax - A^\mathsf{T} b$,
- $\nabla f(x^\star) = 0 \Rightarrow A^\mathsf{T} A x^\star = A^\mathsf{T} b$ (Normal Equations).

Assumption

- $A \in \mathbb{R}^{n\times m}$ *and* $b \in \mathbb{R}^n$
- $n \geq m \Leftrightarrow A$ *is a tall matrix*
- $\mathrm{rank}(A) = m$ *(i.e., columns of* A *are linearly independent)*

Claim

The vector $\Delta x_k \in \mathbb{R}^m$ *is a descent direction, then*

$$\langle \nabla f(x_k), \Delta x_k \rangle < 0 \quad \Longleftrightarrow \quad \Delta x_k \text{ is a descent direction}$$

Claim

The vector $\Delta x_k \in \mathbb{R}^m$ is a descent direction, then

$$\langle \nabla f(x_k), \Delta x_k \rangle < 0 \quad \iff \quad \Delta x_k \text{ is a descent direction}$$

Proof.

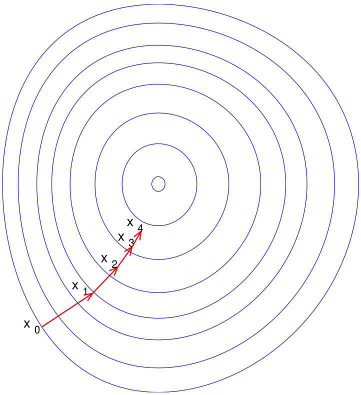
Using the linear approximation of f, we have

$$\begin{aligned} f(x_{k+1}) &\approx f(x_k) + \frac{df(x_k)}{dx}(x_{k+1} - x_k) \\ f(x_{k+1}) - f(x_k) &\approx \nabla f(x_k) \Delta x_k \\ \Delta f(x_k) := f(x_{k+1}) - f(x_k) &\approx \langle \nabla f(x_k), \Delta x_k \rangle \\ \Delta f(x_k) < 0 &\iff \langle \nabla f(x_k), \Delta x_k \rangle < 0. \end{aligned}$$

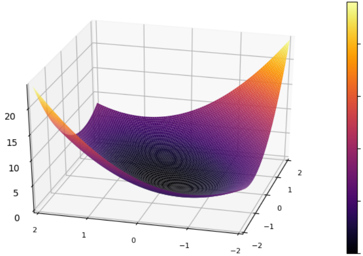
$\square$

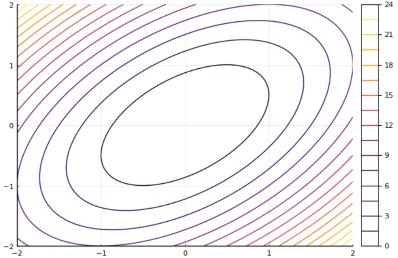
Q. What is a "good" or natural direction (Δx_k) to follow at any point?

It turns out the gradient shows the fastest ascent direction; hence, the negative of the gradient is the fastest descent direction.



To see why we need to take a look at the contour plot of the objective function. The curves in the right figure show the function's level sets (the function is constant along each curve).

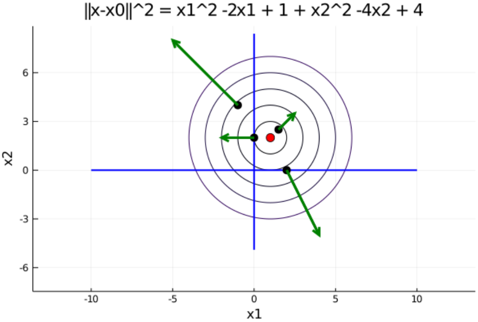




But if the function is constant along a curve, then $\Delta f(x_k) = 0$. Hence, $\boxed{\Delta f(x_k) = \langle \nabla f(x_k), \Delta x_k \rangle = 0.}$

Remark

The only explanation, if $\nabla f(x_k) \neq 0$, is that the gradient is orthogonal to any Δx_k along the curves.



Corollary

We can verify the followings, by setting $\Delta x_k = \pm\nabla f(x_k)$.

1. $\langle \nabla f(x_k), \Delta x_k \rangle = \langle \nabla f(x_k), \nabla f(x_k) \rangle = \|\nabla f(x_k)\|^2 > 0$.
   Hence, $\Delta x_k = \nabla f(x_k)$ *is the fastest ascent direction.*

2. $\langle \nabla f(x_k), \Delta x_k \rangle = \langle \nabla f(x_k), -\nabla f(x_k) \rangle = -\|\nabla f(x_k)\|^2 < 0$.
   Hence, $\Delta x_k = -\nabla f(x_k)$ *is the fastest descent direction.*

We use a step size $\alpha > 0$ to control each update size.

1. Start with an initial guess x_0 ($k = 0$).
2. Evaluate $\nabla f(x_k)$. If $\|\nabla f(x_k)\| = 0$, then the algorithm is converged.
3. Update the decision variable via $x_{k+1} = x_k - \alpha \nabla f(x_k)$.
4. Repeat (go back to 2) until convergence.

Let's switch to the Julia notebook.

Next Time

- Optimization: Second-Order Unconstrained
- Read Chapter 12 of ROB 101 Book