Machine Learning for Big Data: Unconstrained optimization

Lionel Fillatre

fillatre@unice.fr

Topics

- Introduction
- Convexity
- Conditions of optimality
- Gradient descent
- Conclusion

1 Introduction

Examples of machine learning optimization problems

Linear Classification

$$\arg\min_{w} \sum_{i=1}^{n} ||w||^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } 1 - y_i x_i^T w \leq \xi_i$$

$$\xi_i \geq 0$$

K-Means

$$\arg \min_{\mu_1, \mu_2, \ldots, \mu_k} J(\mu) = \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2$$

Optimization problems

- Generic unconstrained minimization problem

$$\min_{x \in \mathbb{X}} f(x)$$

where

- Vector space $\mathbb{X}$ is the search space
- $f: \mathbb{X} \to \mathbb{R}$ is a cost (or objective) function
- A solution $x^* = \underset{x \in \mathbb{X}}{\operatorname{argmin}} f(x)$ is the minimizer of $f(x)$
- The value $f(x^*)$ is the minimum

- Maximization can be converted to minimization

$$\max f(x) = \min(-f(x))$$

Existence of global minimum

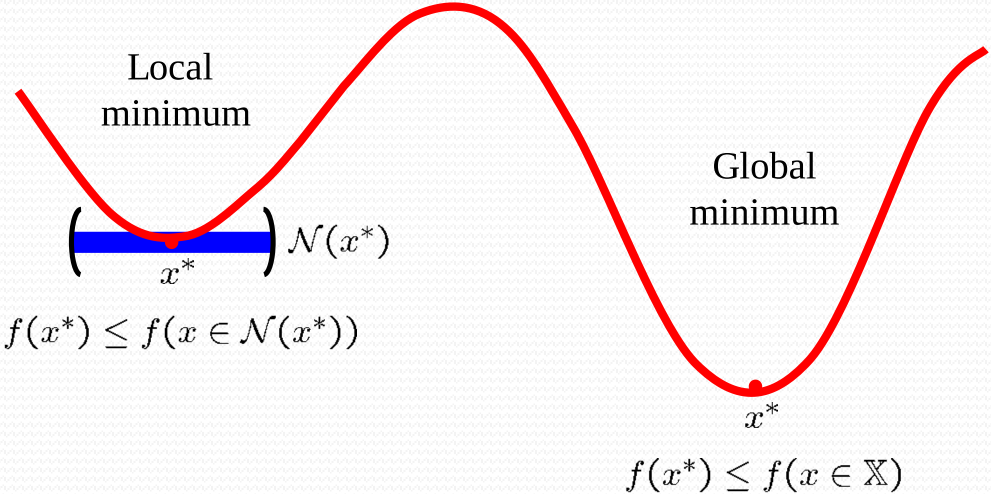
- If $f(x)$ is continuous on the set S which is closed and bounded, then $f(x)$ has at least one global minimum in S.
  - A set S is closed if it contains all its boundary points.
  - A set S is bounded if it is contained in the interior of some circle:

$\|x^T x\|^2 \leq c, \forall x \in S$ where c is a finite value.

S compact = S closed and bounded

Local vs. Global minimum

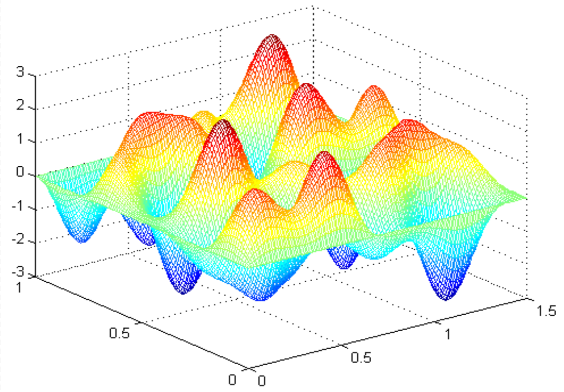
Find minimum by analyzing the local behavior of the cost function

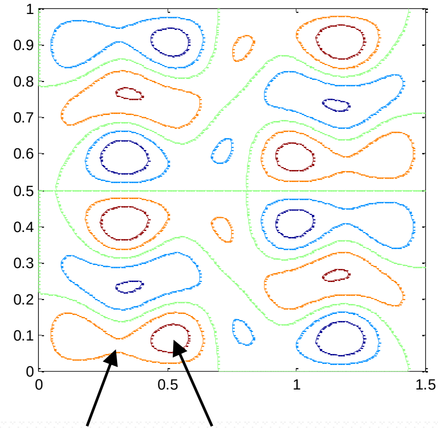


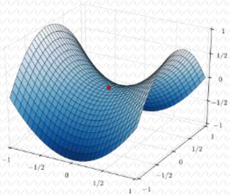
Local vs. Global in real life

Broad Peak (K3), 12th highest mountain on Earth

Some numerical difficulties



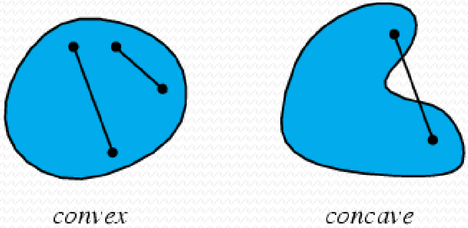


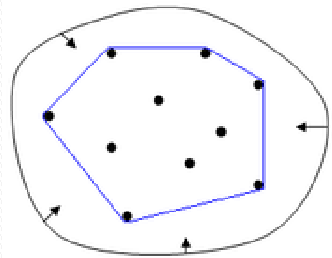


2 Convexity

Convex Hull

- A set C is **convex** if every point on the line segment connecting *x* and *y* is in *C*.
- The **convex hull** for a set of points X is the minimal convex set containing X.



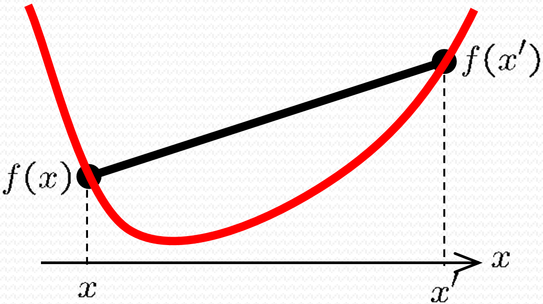


Convex functions

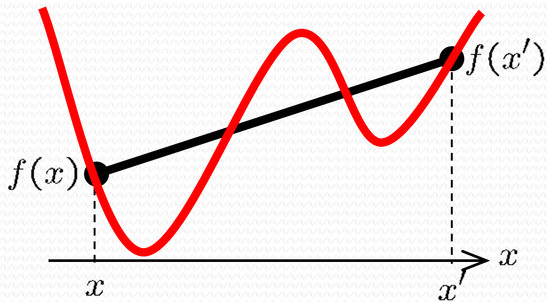
A function $f : A \subseteq \mathbb{X} \to \mathbb{R}$ defined on a convex set A is called convex if $f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$

for any $x, x' \in \mathbb{X}$ and $\lambda \in [0, 1]$

For convex function, local minimum = global minimum

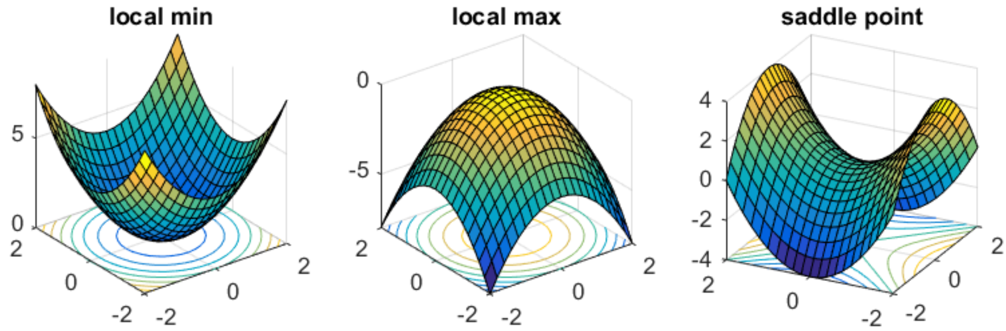


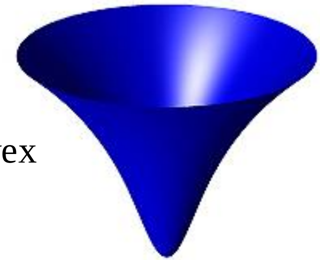
Convex



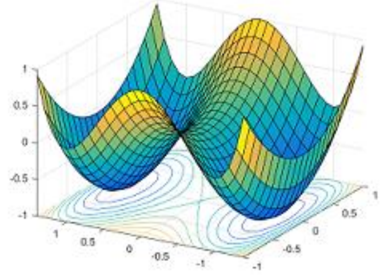
Non-convex

Convex and non-convex functions





Quasi-convex



Non-convex

3 Conditions of optimality

One-dimensional optimality conditions

Point x^* is the local minimizer of a $\mathcal{C}^2$-function $f : \mathbb{R} \to \mathbb{R}$ if

- $f'(x^*) = 0$
- $f''(x^*) > 0$

Approximate a function around x^* as a parabola using Taylor expansion

$$f(x^* + dx) \approx f(x^*) + f'(x^*)dx + \frac{1}{2}f''(x^*)dx^2$$

$f'(x^*) = 0$ guarantees the minimum at x^*

$f''(x^*) > 0$ guarantees the parabola is convex

Gradient

- In multidimensional case, linearization of the function according to Taylor
$$f(x+h) \approx f(x) + h^T g(x)$$
for a small vector h.

- The function $g(x)$, denoted as $\nabla f(x)$, is called the gradient of $f(x)$ at point x.

- We can show that
$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)$$

Example

- Calculate the gradient to determine the direction of the steepest slope at point (2,1) for the function

$$f(x,y) = x^2y^2$$

- **Solution:** To calculate the gradient we would need to calculate

$$\frac{\partial f}{\partial x}(x,y) = 2xy^2 \quad \frac{\partial f}{\partial y}(x,y) = 2x^2y$$

- Hence,

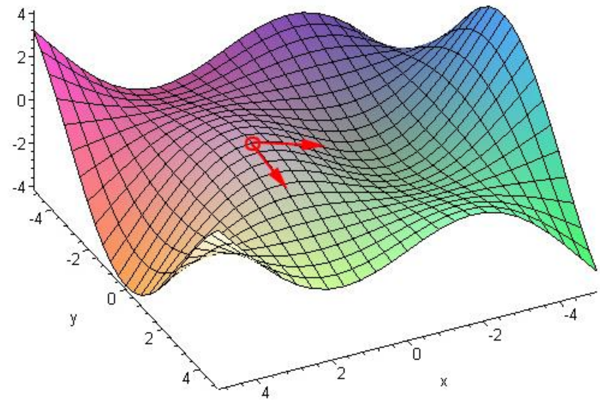
$$\nabla f(2,1) = \binom{4}{8}$$

Interpreting the gradient

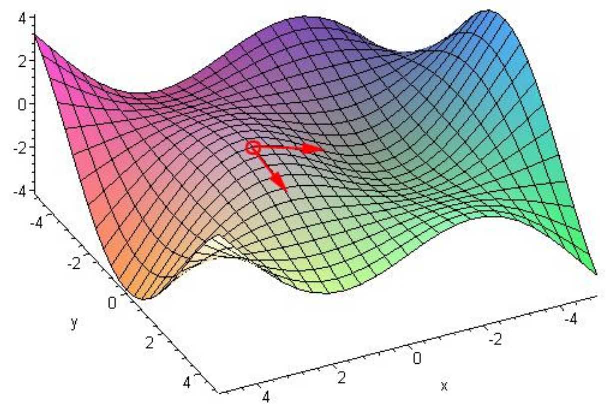
- Along the axes…

$$\frac{\partial f}{\partial y}(x, y)$$

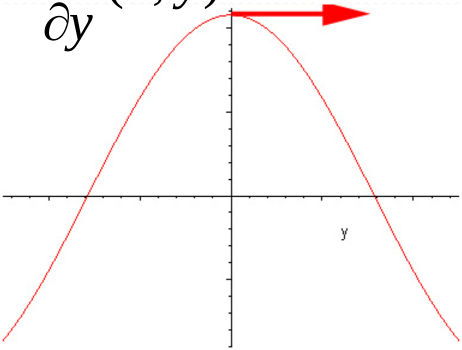
$$\frac{\partial f}{\partial x}(x, y)$$



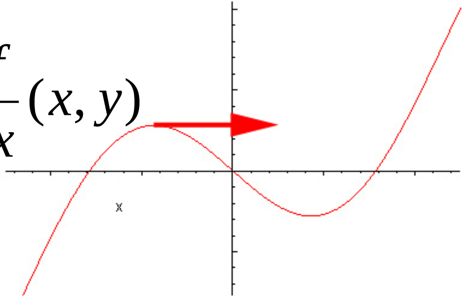
Interpreting the gradient



$$\frac{\partial f}{\partial y}(x, y)$$



$$\frac{\partial f}{\partial x}(x, y)$$



Continously differentiable function

- Definition: A real-valued function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be continuously differentiable if the partial derivatives

$$\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$$

exist for each x in $\mathbb{R}^n$ and are continuous functions of x.

- In this case, we say $f \in \mathcal{C}^1$: it is a $\mathcal{C}^1$smooth function

- If the function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ has second order derivatives continuous with respect to x, then we say $f \in \mathcal{C}^2$.

Taylor expansion

- A function $f \in \mathcal{C}^2$ may be approximated locally by its Taylor series expansion about a point x^*

$$f(x^* + h) \approx f(x^*) + \nabla f(x^*)^T h + \frac{1}{2} h^T \nabla^2 f(x^*) h$$

where $\nabla f(x^*)$ is the gradient and $H(x^*) = \nabla^2 f(x^*)$ is the Hessian.

- The Hessian is the symmetric matrix of second order derivatives at point x:

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}$$

Optimality conditions

- Point x^* is the local minimizer of a C^2 function $f: \mathbb{X} \to \mathbb{R}$ if
  - $\nabla f(x^*) = 0$
  - $x^T \nabla^2 f(x^*) x > 0$ for all $x \neq 0$, i.e., the Hessian is a positive definite, which is denoted $\nabla^2 f(x^*) > 0$
- Approximate a function around x^* as a parabola using Taylor expansion

$$f(x^* + h) = f(x^*) + \nabla f(x^*)^T h + \frac{1}{2} h^T \nabla^2 f(x^*) h$$

$\nabla f(x^*) = 0$ guarantees the minimum at x^*

$\nabla^2 f(x^*) > 0$ guarantees the parabola is convex

Quadratic functions

$$f(\mathbf{x}) = a + \mathbf{g}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$$

- The vector **g** and the Hessian **H** are constant.
- Second order approximation of any function by the Taylor expansion is a quadratic function.

Necessary conditions for a minimum

$$f(\mathbf{x}) = a + \mathbf{g}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$$

- Expand $f(x)$ about a stationary point $\mathbf{x}^*$ in a vector direction $\mathbf{p}$

$$\begin{aligned} f(\mathbf{x}^* + \alpha\mathbf{p}) &= f(\mathbf{x}^*) + \mathbf{g}(\mathbf{x}^*)^T\alpha\mathbf{p} + \frac{1}{2}\alpha^2\mathbf{p}^T\mathbf{H}\mathbf{p} \\ &= f(\mathbf{x}^*) + \frac{1}{2}\alpha^2\mathbf{p}^T\mathbf{H}\mathbf{p} \end{aligned}$$

since at a stationary point $\mathbf{g}(\mathbf{x}^*) = 0$

- At a stationary point the behavior is determined by $\mathbf{H}$

Behavior related to eigenvalue

- **H** is a symmetric matrix, and so has orthogonal eigenvectors u_i and eigenvalues λ_i:

$$\mathbf{H}\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad \|\mathbf{u}_i\| = 1$$

$$\begin{aligned} f(\mathbf{x}^* + \alpha \mathbf{u}_i) &= f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \mathbf{u}_i^T \mathbf{H} \mathbf{u}_i \\ &= f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \lambda_i \end{aligned}$$

- As $|\alpha|$ increases, $f(\mathbf{x}^* + \alpha \mathbf{u}_i)$ increases, decreases or is unchanging according to whether λ_i is positive, negative or zero.

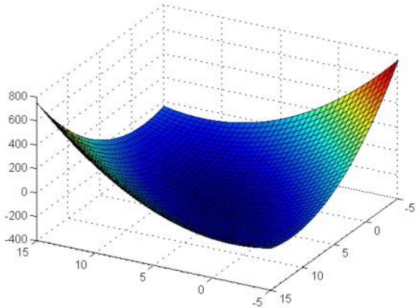
Examples of quadratic functions

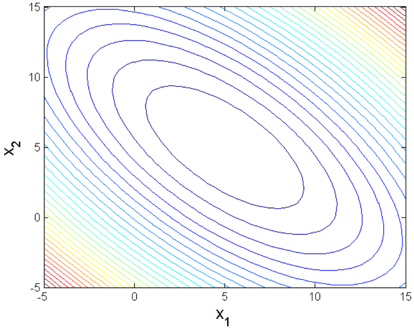
Case 1: both eigenvalues positive

$$f(\mathbf{x}) = a + \mathbf{g}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$$

with

$$a = 0\,, \qquad \mathbf{g} = \begin{bmatrix} -50 \\ -50 \end{bmatrix}, \qquad \mathbf{H} = \begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix}$$ positive definite

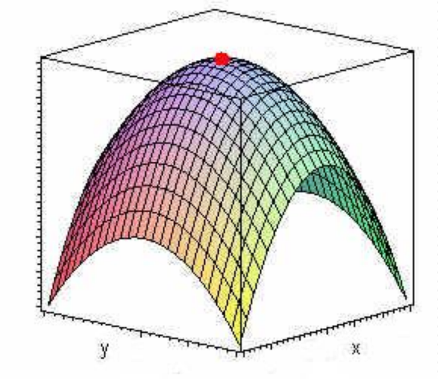




The stationary point is a minimum!

Examples of quadratic functions

Case 2: both eigenvalues negative



The stationary point is a maximum!

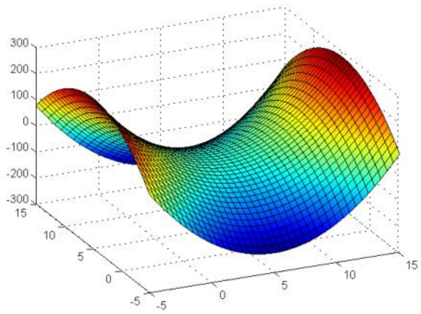
Examples of quadratic functions

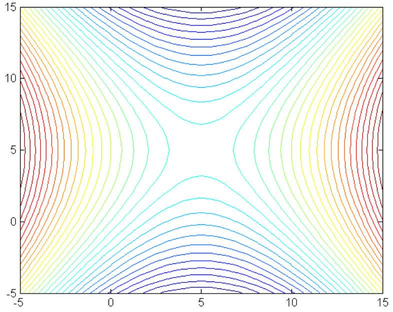
Case 3: eigenvalues have different sign

$$f(\mathbf{x}) = a + \mathbf{g}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$$

with

$$a = 0\,, \qquad \mathbf{g} = \begin{bmatrix} -30 \\ 20 \end{bmatrix}, \qquad \mathbf{H} = \begin{bmatrix} 6 & 0 \\ 0 & -4 \end{bmatrix}$$ indefinite





The stationary point is a saddle point!

Optimization for quadratic functions

- Assume that **H** is positive definite

$$f(\mathbf{x}) = a + \mathbf{g}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$$

$$\nabla f(\mathbf{x}) = \mathbf{g} + \mathbf{H}\mathbf{x}$$

- There is a unique minimum at

$$\mathbf{x}^* = -\mathbf{H}^{-1}\mathbf{g}$$

- If the size of *x* is large, it is not feasible to perform this inversion directly.

Solve normal equations

- Assume we want to minimize $\|Y - Ax\|_2^2$ with respect to x
- Let us consider $f(x) = \frac{1}{2}x^T A^T Ax - Y^T Ax$
- We have $\nabla f(x) = A^T Ax - A^T Y$
- There is a unique minimum at $x^* = (A^T A)^{-1} A^T Y$
- Then the minimization of a function is equivalent to solve the normal equations!
- Easier to code and to use with a distributed computing system.

4 Gradient descent

Optimization algorithms

Generic optimization algorithm

- Start with some $x^{(0)}, k = 0$
- Determine descent direction $d^{(k)}$
- Choose step size $\alpha^{(k)}$ such that

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) < f(x^{(k)})$$

- Update iterate

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$$

- Increment iteration counter $k \leftarrow k + 1$
- Solution $x^* \approx x^{(k)}$

Until convergence

Stopping criteria

- Near local minimum, $\nabla f(x) \approx 0$ or equivalently $\|\nabla f(x)\| \approx 0$

  Stop when gradient norm becomes small

$$\|\nabla f(x^{(k)})\| \leq \epsilon$$

- Stop when step size becomes small

$$\|x^{(k+1)} - x^{(k)}\| \leq \epsilon$$

- Stop when relative objective change becomes small

$$\frac{f(x^{(k)}) - f(x^{(k+1)})}{f(x^{(k)})} \leq \epsilon$$

Example

- Determine the minimum of the function

$$f(x, y) = x^2 + y^2 + 2x + 4$$

- Use the point (2,1) as the initial estimate of the optimal solution.

Solution

Iteration 1: To calculate the gradient; the partial derivatives must be evaluated as

$$\frac{\partial f}{\partial x}(x=2, y=1) = 2x + 2 = 2(2) + 2 = 4 \qquad \frac{\partial f}{\partial y}(x=2, y=1) = 2y = 2(1) = 2$$

Now the function $f(x, y)$ can be expressed along the direction of gradient as

$$f\left(2 + \frac{\partial f}{\partial x}(2,1)h, 1 + \frac{\partial f}{\partial y}(2,1)h\right) = f(2+4h, 1+2h) = (2+4h)^2 + (1+2h)^2 + 2(2+4h) + 4$$

$$g(h) = 20h^2 + 28h + 13$$

Solution Cont.

Iteration 1 continued:
This is a simple function and it is easy to determine $h^* = -0.7$ by taking the first derivative and solving for its roots.

This means that traveling a step size of $h = -0.7$ along the gradient reaches a minimum value for the function in this direction. These values are substituted back to calculate a new value for x and y as follows:

$$x = 2 + 4(-0.7) = -0.8$$
$$y = 1 + 2(-0.7) = -0.4$$

Note that $f(2,1) = 13$ $\quad f(-0.8,-0.4) = 3.2$

Solution Cont.

Iteration 2: The new initial point is $(-0.8,-0.4)$ with $f(-0.8,-0.4)=3.2$
We calculate the gradient at this point as

$$\frac{\partial f}{\partial x}(x=-0.8, y=-0.4) = 2x+2 = 2(-0.8)+2 = 0.4$$

$$\frac{\partial f}{\partial y}(x)-0.8, y=-0.4) = 2y = 2(-0.4) = -0.8$$

$$f\left(-0.8+\frac{\partial f}{\partial x}(-0.8,-0.4)h,-0.4+\frac{\partial f}{\partial y}(-0.8,-0.4)h\right) = f(-0.8+0.4h,-0.4-0.8h)$$

$$=(-0.8+0.4h)^2+(0.4-0.8h)^2+2(-0.8+0.4h)+4$$

$$g(h) = 0.8h^2+0.8h+3.2 \quad \Rightarrow h^* = -0.5$$

$$x = -0.8+0.4(-0.5) = -1$$

$$y = -0.4-0.8(-0.5) = 0 \qquad f(-1,0)=3$$

Solution Cont.

Iteration 3: The new initial point is $(-1,0)$.
We calculate the gradient at this point as

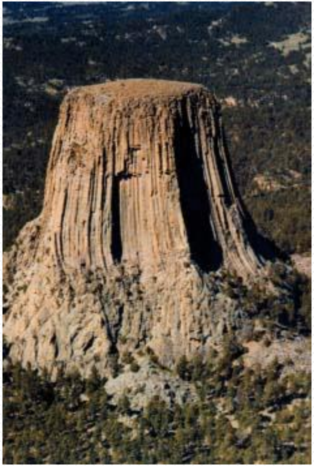
$$\frac{\partial f}{\partial x}(x=-1, y=0) = 2x+2 = 2(-1)+2 = 0$$

$$\frac{\partial f}{\partial y}(x=-1, y=0) = 2y = 2(0) = 0$$

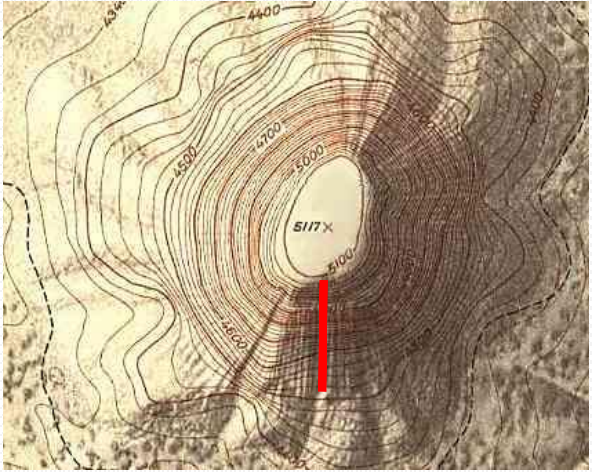
This indicates that the current location is a local optimum along this gradient and no improvement can be gained by moving in any direction. The minimum of the function is at point (-1,0).

How to descend in the fatest way?

Go in the direction in which the height lines are the densest



Devil’s Tower



Topographic map

Steepest descent

$$f(x+d) \approx f(x) + \nabla f(x)^\top d$$

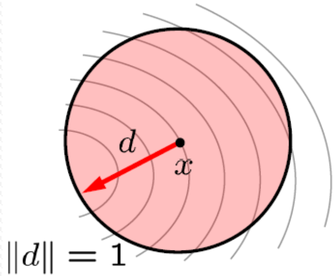
Directional derivative: how much $f(x)$ changes in the direction d (negative for a descent direction)

- Find a unit-length direction minimizing directional derivative:

$$d = \underset{d:\|d\|=1}{\operatorname{argmin}} \nabla f(x)^\top d$$

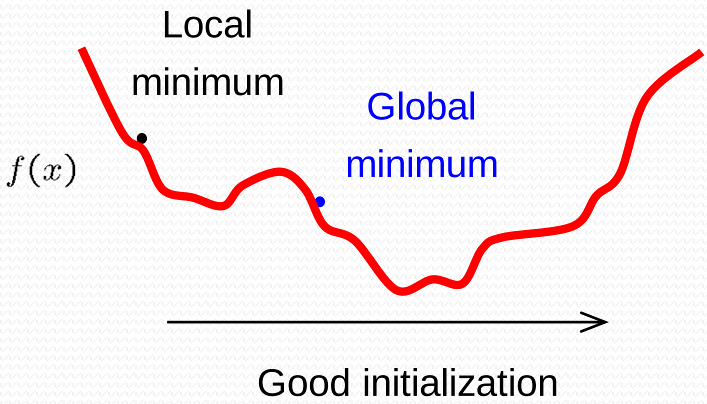
- The opposite of the gradient is the steepest descent:

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$



Non-convex optimization

- Using convex optimization methods with non-convex functions does not guarantee global convergence!
- There is no theoretical guaranteed global optimization, just heuristics



8 Conclusion

Conclusion

- Optimization is one of the founding principles of machine learning
- It is important to understand the limits of the optimization when analysing data
- Optimization usually contains constraints (posivity, etc.)