Analytical Functions

Example: 1-D

$$f_1(x) = a(x - x_0)^2 + b$$

where $x, x_0, b \in \mathbb{R}, a \in \mathbb{R}$

Generalization:

convex quadratic function

$$f_2(x) = \frac{1}{2}(x - x_0)^T A\,(x - x_0) + b$$

where $x, x_0 \in \mathbb{R}^n, b \in \mathbb{R}\quad, A \in \mathbb{R}^{\{n \times n\}}$

and A symmetric positive definite (SPD)

Exercise:
What is the minimum of $f_2(x)$?

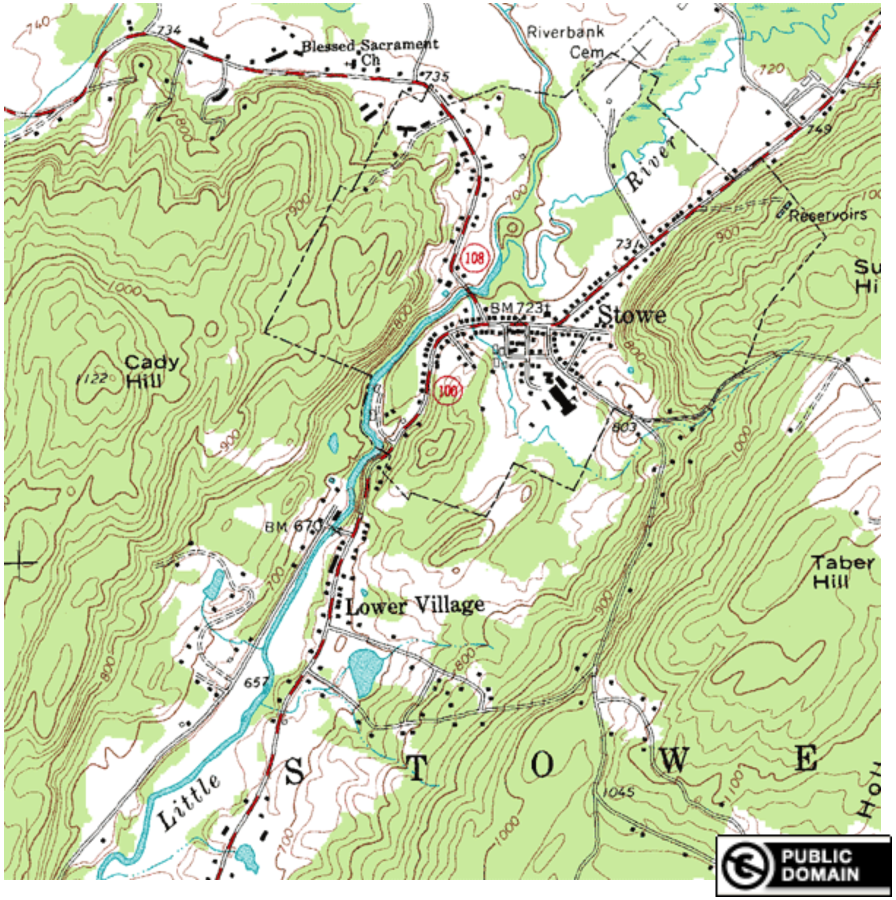
Levels Sets of Convex Quadratic Functions

Continuation of exercise:
What are the level sets of f_2?

Reminder: level sets of a function

$$L_c = \{x \in \mathbb{R}^n \mid f(x) = c\}$$

(similar to topography lines / level sets on a map)

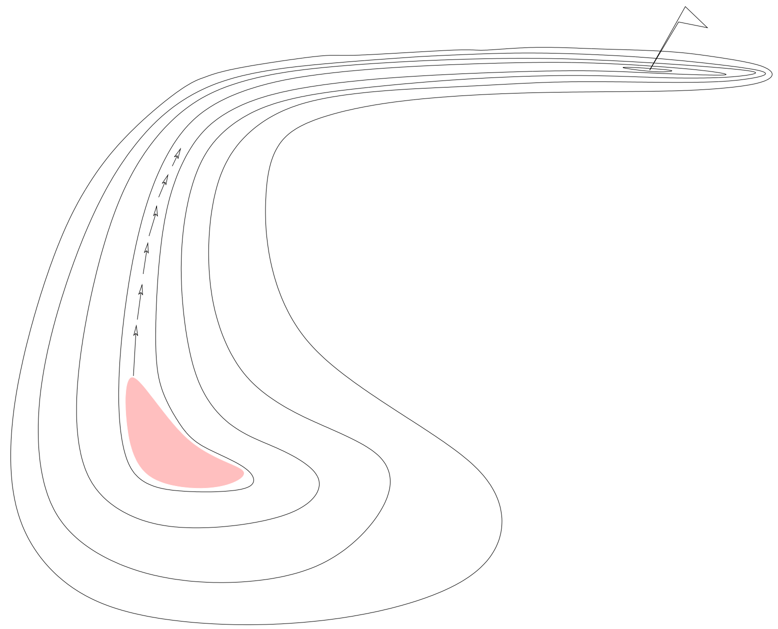


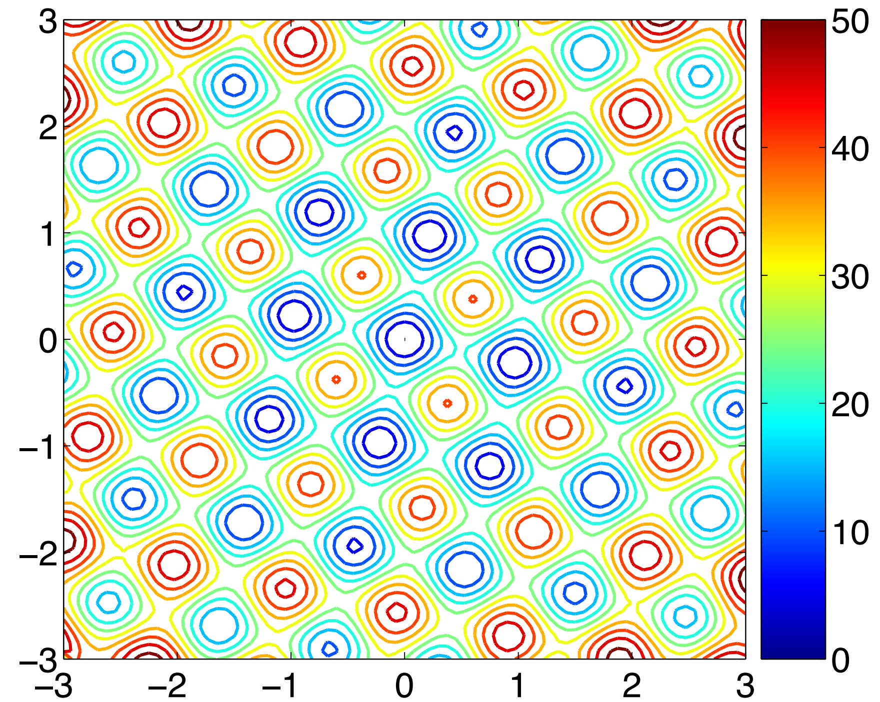
Level Sets: Visualization of a Function

One-dimensional (1-D) representations are often misleading (as 1-D optimization is "trivial", see slides related to curse of dimensionality), we therefore often represent level-sets of functions

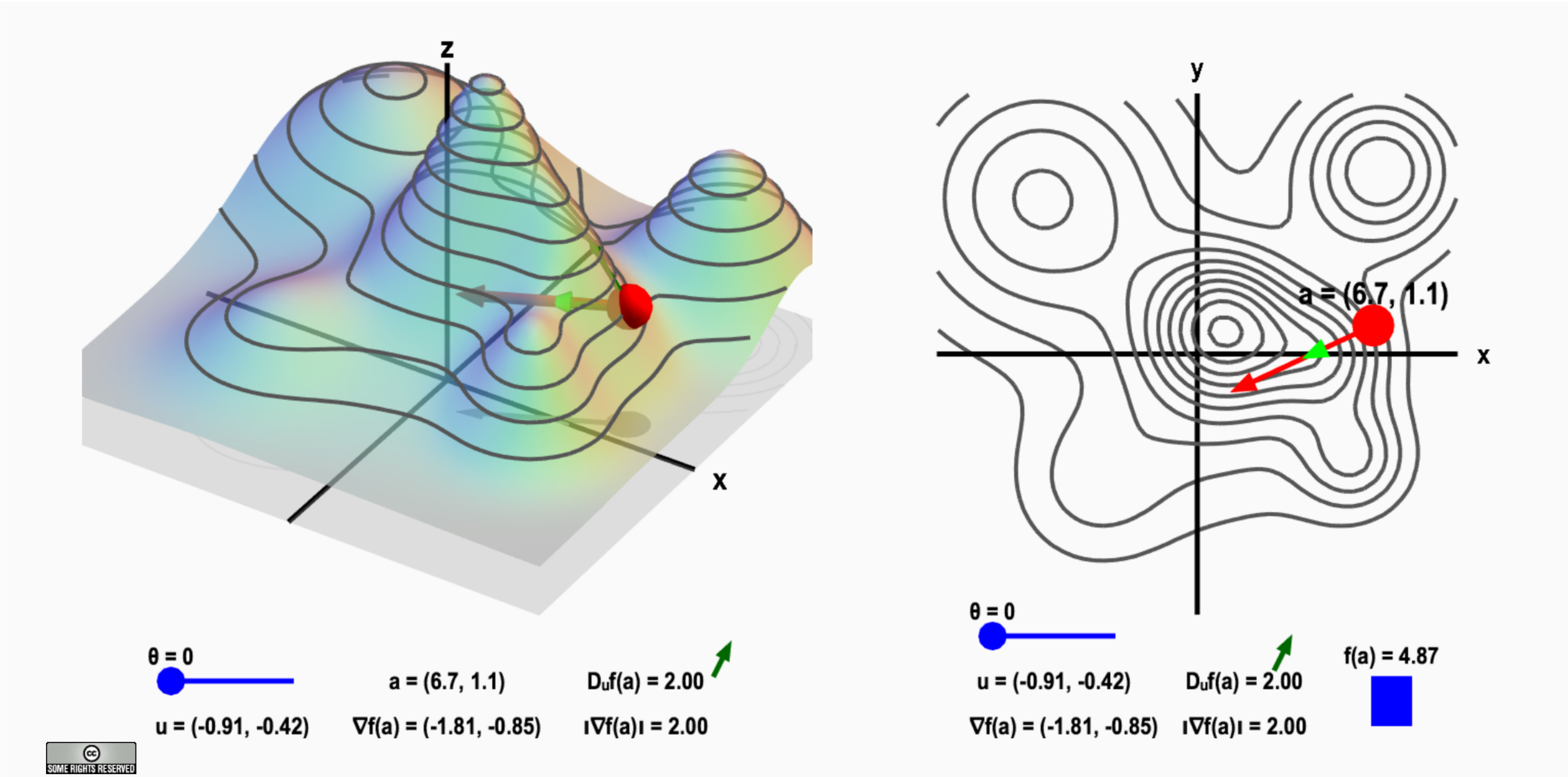
$$\mathscr{L}_c = \{x \in \mathbb{R}^n \,|\, f(x) = c,\}, c \in \mathbb{R}$$

Examples of level sets in 2D





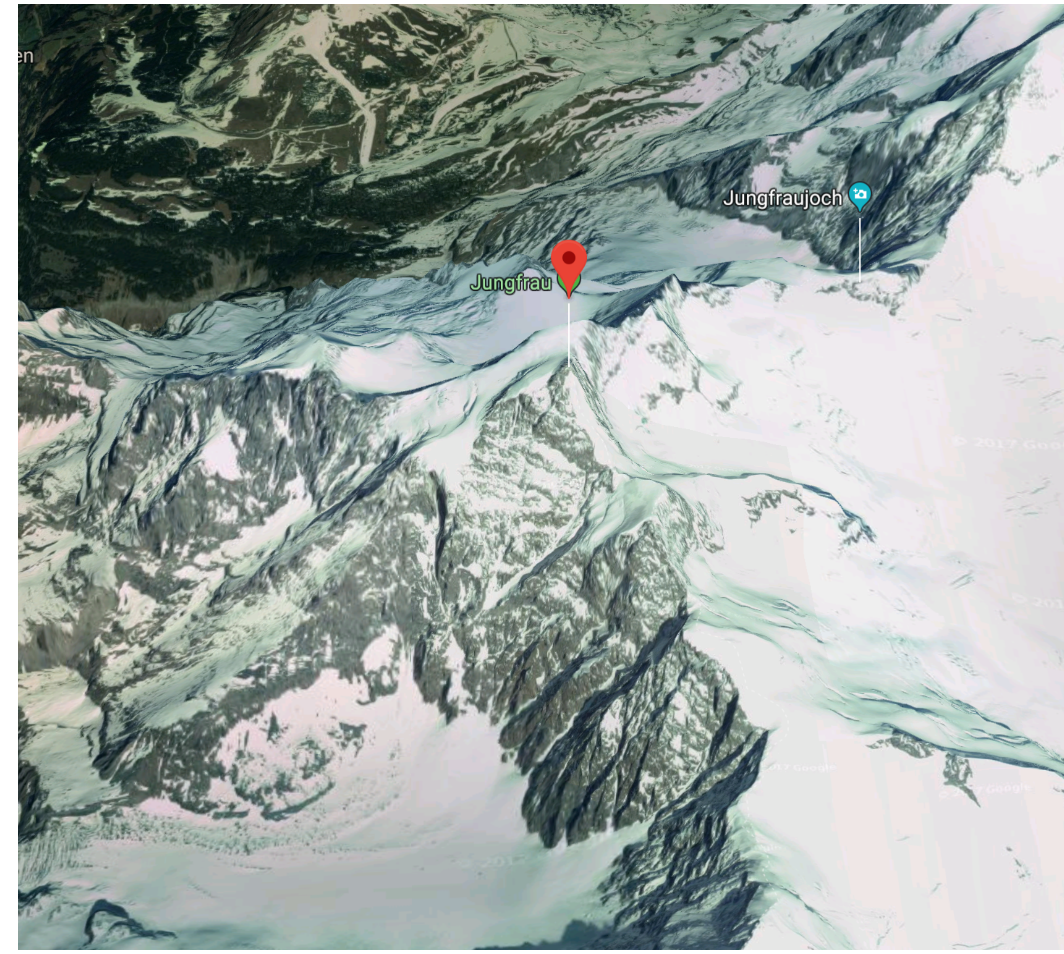
Level Sets: Visualization of a Function



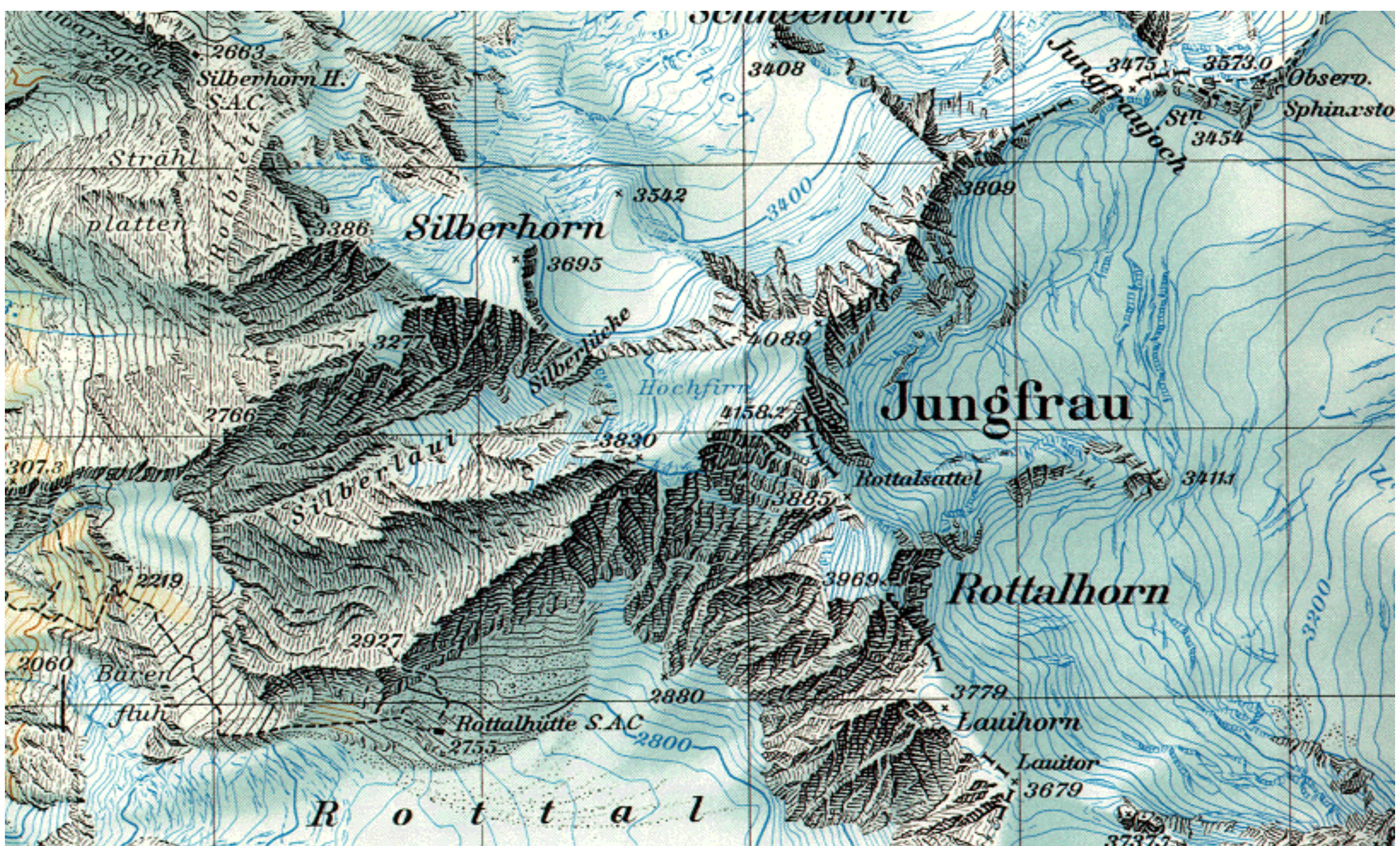
Source: Nykamp DQ, "Directional derivative on a mountain." From *Math Insight*. http://mathinsight.org/applet/directional_derivative_mountain

Level Sets: Topographic Map

The function is the altitude



3-D picture



Topographic map

Levels Sets of Convex Quadratic Functions

Continuation of exercise:
What are the level sets of f_2?

$f_2(x) = \frac{1}{2}(x-x_0)^T A (x-x_0) + b$

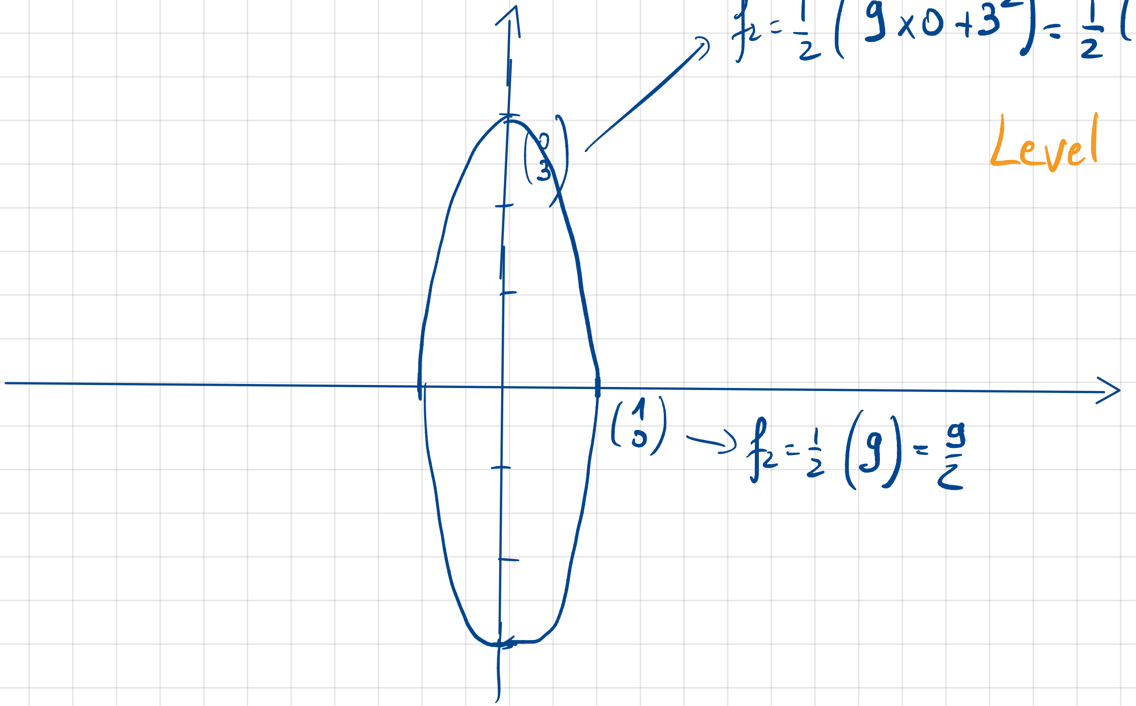
A SPD

- Probably too complicated in general, thus an example here
- Consider $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}, b = 0, n = 2$
  a) Compute $f_2(x)$.
  b) Plot the level sets of $f_2(x)$.
  c) More generally, for $n = 2$, if A is SPD with eigenvalues $\lambda_1 = 9$ and $\lambda_2 = 1$, what are the level sets of $f_2(x)$?

Not necessarily diagonal

$$A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \qquad f_2(x) = \frac{1}{2}\left(9x_1^2 + x_2^2\right) \qquad x = (x_1, x_2)$$

$$f_2 = \frac{1}{2}\left(9 \times 0 + 3^2\right) = \frac{1}{2}(9) = \frac{9}{2}$$

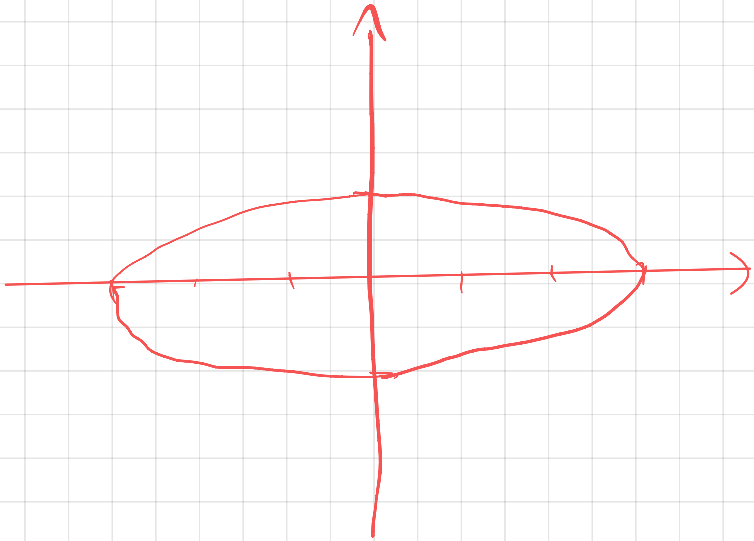


$\begin{pmatrix} 0 \\ 3 \end{pmatrix}$

$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \rightarrow f_2 = \frac{1}{2}(9) = \frac{9}{2}$

Level sets are ellipsoid, long-axis : y-axis
small axis : x-axis

If $A = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix}$, $f_2(x) = \frac{1}{2}\left(x_1^2 + 9x_2^2\right) \longrightarrow$



A is symmetric, positive, definite:

$A = PDP^T$ from the spectral theorem.

P is orthogonal

P contains the eigenvectors of A

$$f_2(x) = \frac{1}{2} x^T A x = \frac{1}{2} x^T P D P^T x$$

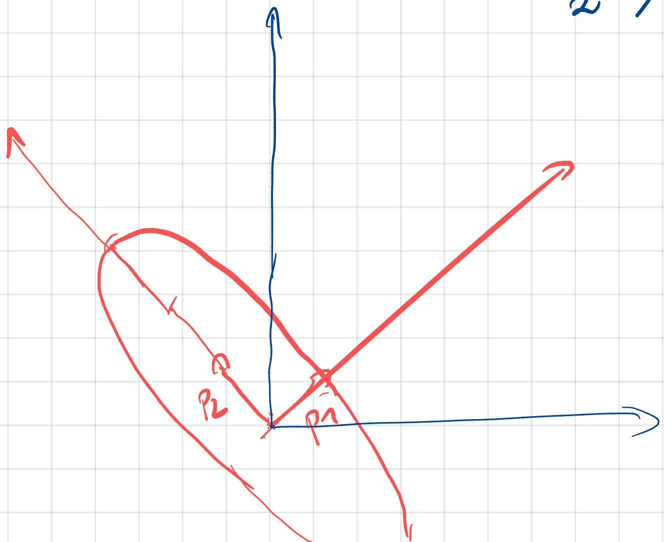
$$= \frac{1}{2} \underbrace{(P^T x)^T}_{y^T} D \underbrace{P^T x}_{y}$$

$$= \frac{1}{2} y^T D y$$

$$= \frac{1}{2} y^T \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} y = \frac{1}{2}\left(9 y_1^2 + y_2^2\right)$$

$y = P^T x$

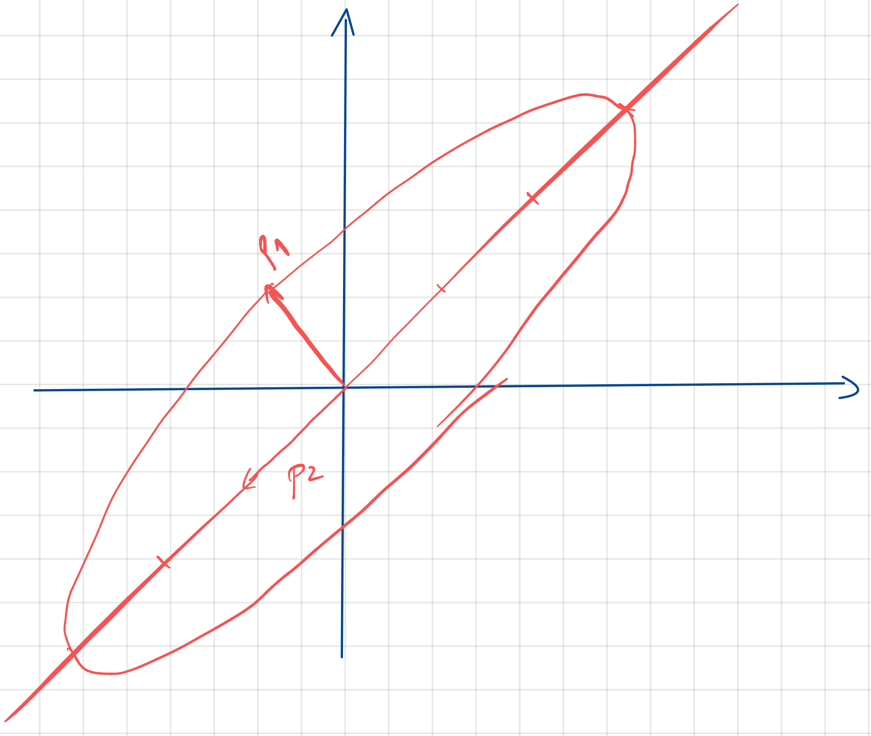
$$D = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$$



p_1, p_2 eigenvectors of A associated to $\lambda_1 = 9$, $\lambda_2 = 1$

"Same" ellipsoid than before but rotated

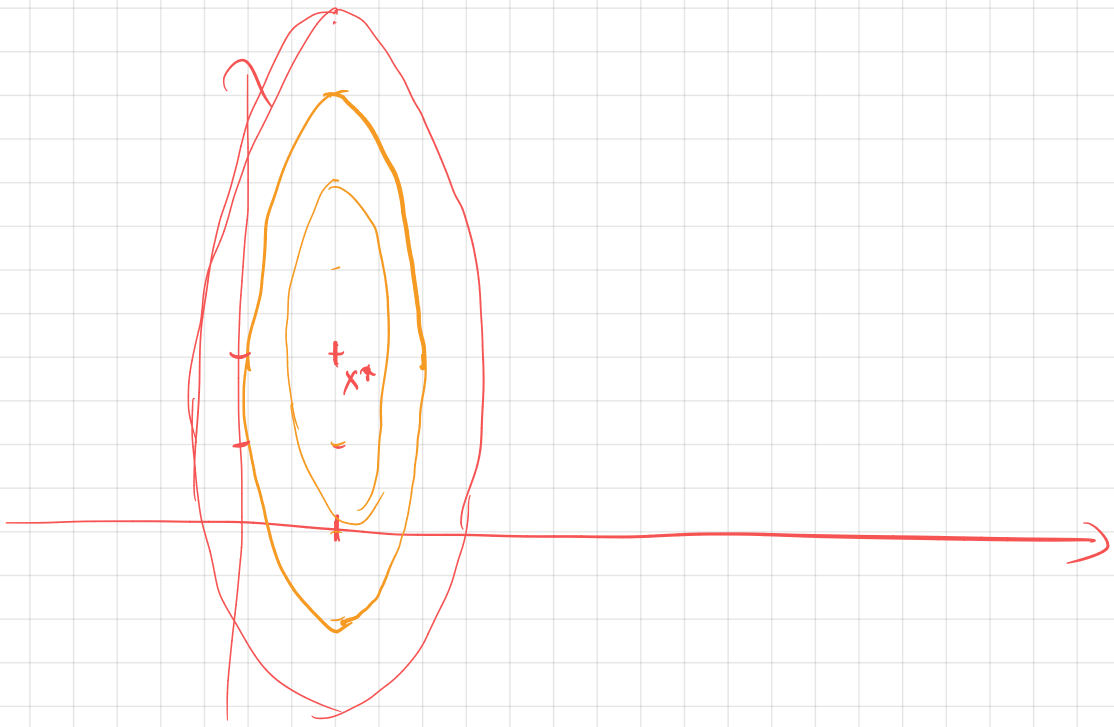
The main axis of ellipsoid are the eigenvectors of A.


P1
P2

We have assumed before $x^* = 0$, if $x^* = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and we consider

$$f(x) = \frac{1}{2} (x - x^*)^T \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} (x - x^*)$$

Then the optimum of f is in x^* and the ellipsoid are centered around x^*, i.e.

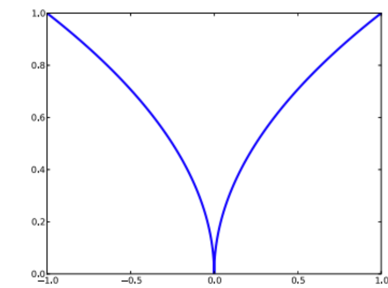


What Makes a Function Difficult to Solve?

~~Why stochastic search?~~

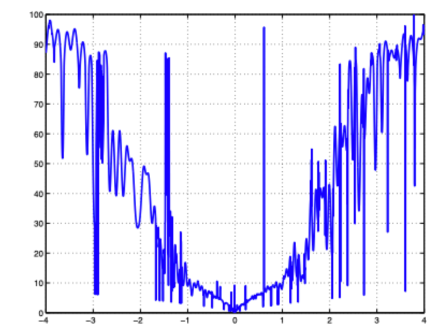
- non-linear, non-quadratic, non-convex

  on linear and quadratic functions much better search policies are available



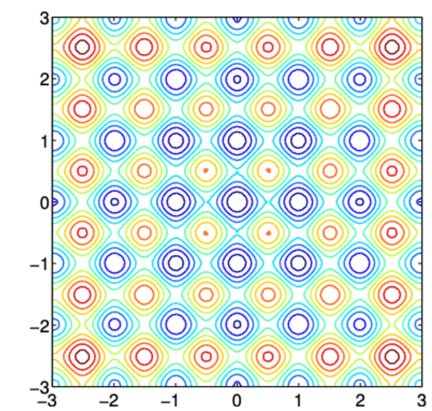
- ruggedness

  non-smooth, discontinuous, multimodal, and/or noisy function



- dimensionality (size of search space)

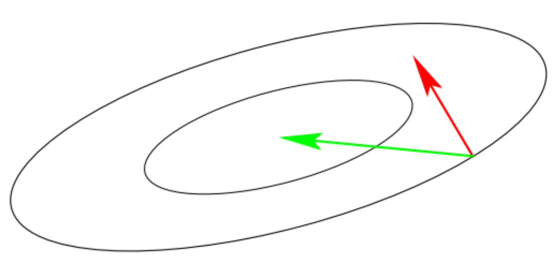
  (considerably) larger than three



- non-separability

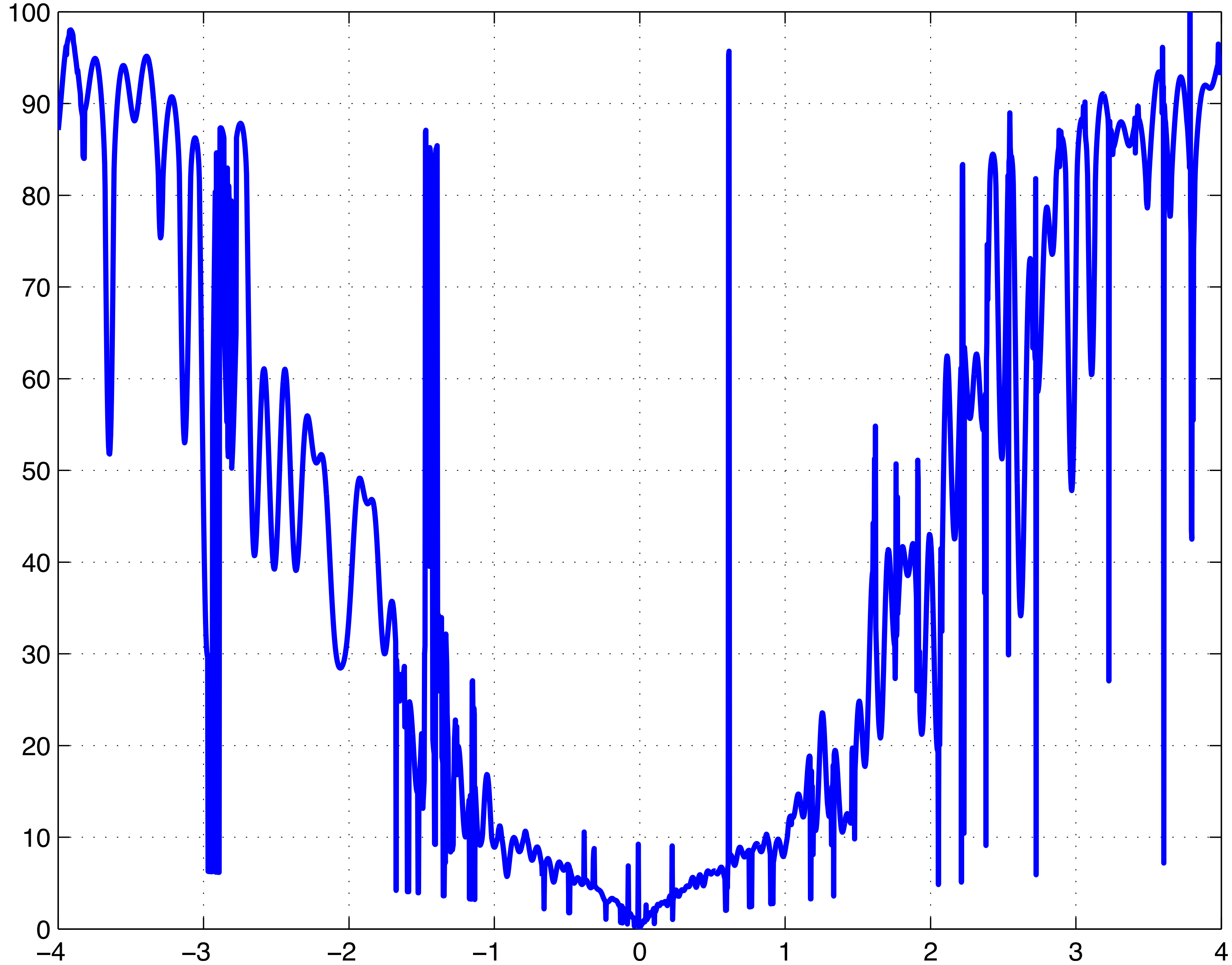
  dependencies between the objective variables

- ill-conditioning



gradient direction Newton directio

Ruggedness



A cut of a 4-D function that can easily be solved with the CMA-ES algorithm

Why is Optimization a non-trivial Problem?

Curse of dimensionality

if n=1, which simple approach could you use to minimize:

$$f : [0,1] \to \mathbb{R} \quad ?$$

Why is Optimization a non-trivial Problem?

Curse of dimensionality

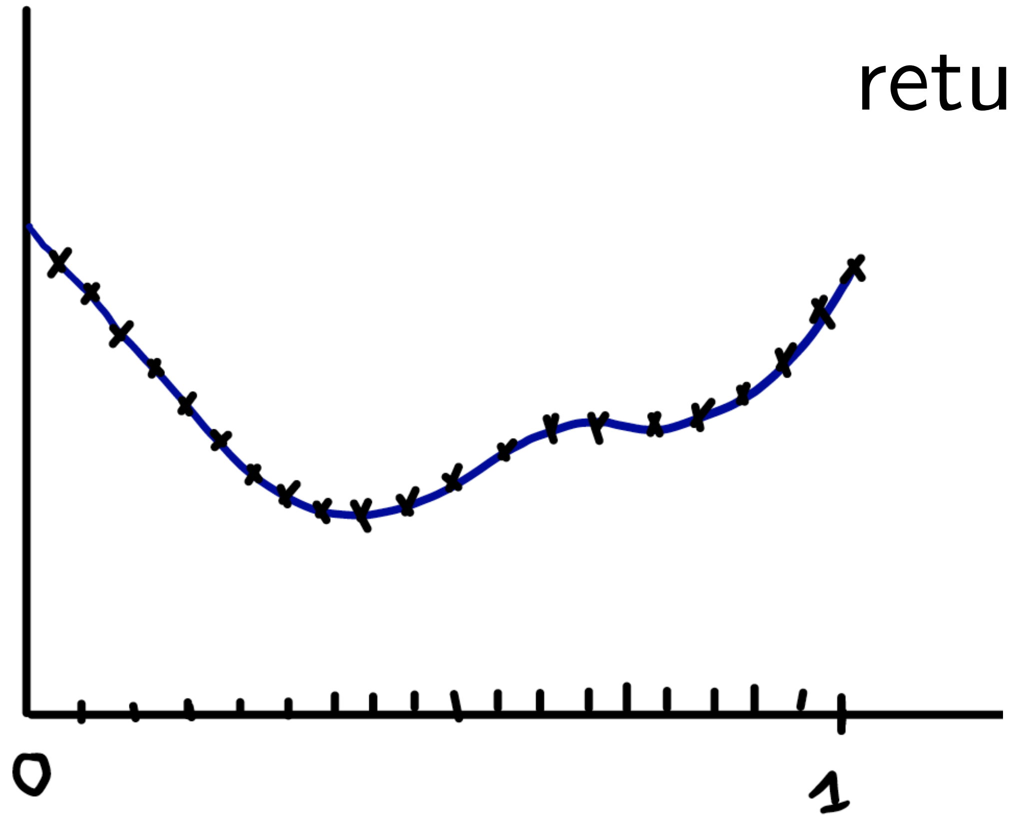
if n=1, which simple approach could you use to minimize:

$$f : [0,1] \to \mathbb{R} \quad ?$$

set a regular grid on [0,1]

evaluate on f all the points of the grid

return the lowest function value



Why is Optimization a non-trivial Problem?

Curse of dimensionality

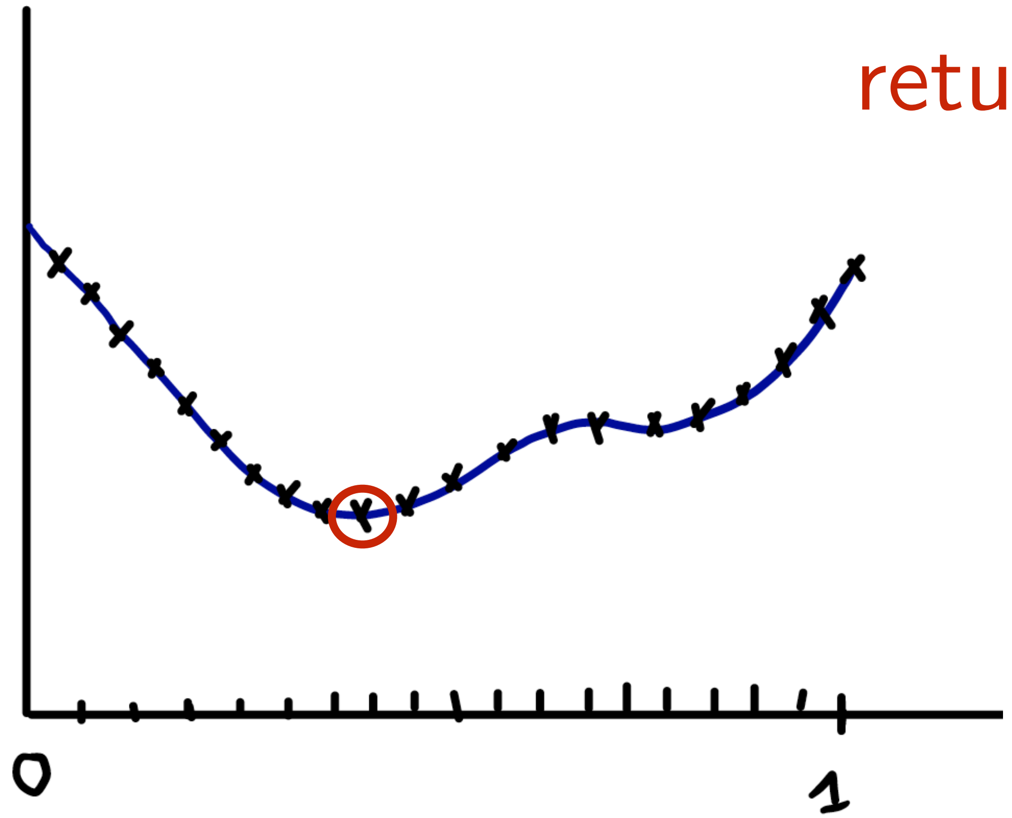
if n=1, which simple approach could you use to minimize:

$$f : [0,1] \to \mathbb{R} \quad ?$$

set a regular grid on [0,1]

evaluate on f all the points of the grid

return the lowest function value



Why is Optimization a non-trivial Problem?

Curse of dimensionality

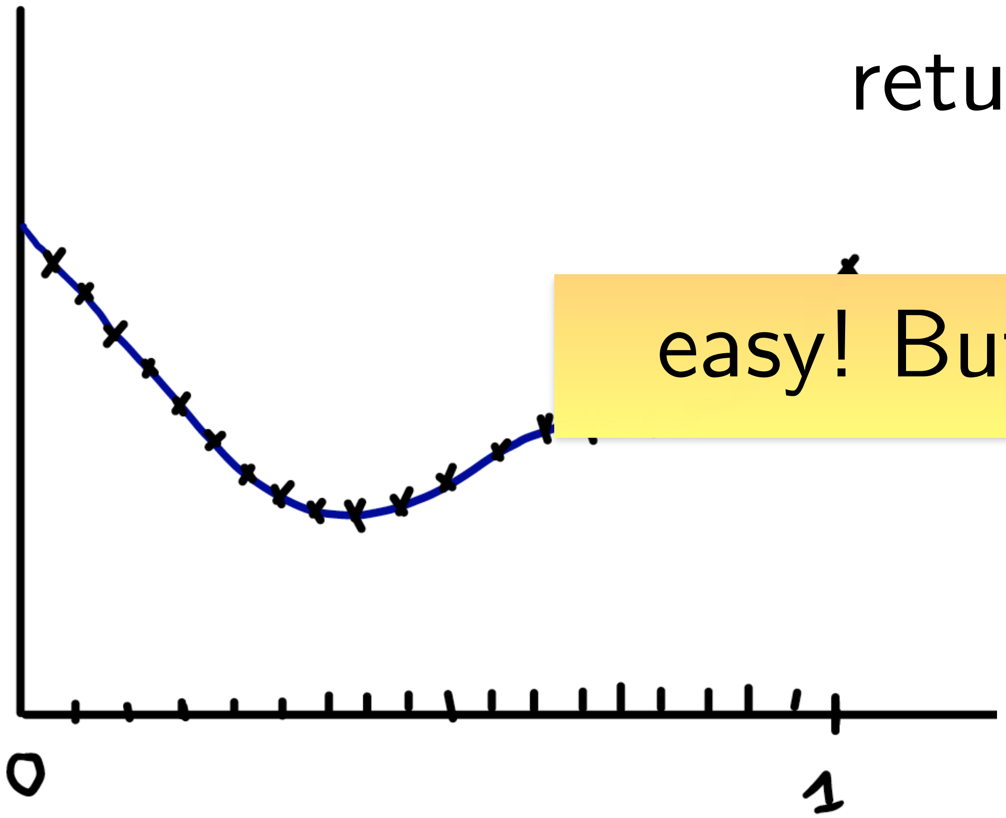
if n=1, which simple approach could you use to minimize:

$$f : [0,1] \to \mathbb{R} \quad ?$$

set a regular grid on [0,1]

evaluate on f all the points of the grid

return the lowest function value



easy! But how does it scale when n increases?

1-D optimization is trivial

Curse of Dimensionality

The term curse of dimensionality (Richard Bellman) refers to problems caused by the rapid increase in volume associated with adding extra dimensions to a (mathematical) space.

Example: Consider placing 100 points onto a real interval, say [0,1].

How many points would you need to get a similar coverage (in terms of distance between adjacent points) in dimension 10?

Curse of Dimensionality

The term curse of dimensionality (Richard Bellman) refers to problems caused by the rapid increase in volume associated with adding extra dimensions to a (mathematical) space.

Example: Consider placing 100 points onto a real interval, say $[0,1]$. To get similar coverage, in terms of distance between adjacent points, of the 10-dimensional space $[0,1]^{10}$ would require $100^{10} = 10^{20}$ points. A 100 points appear now as isolated points in a vast empty space.

Consequence: a search policy (e.g. exhaustive search) that is valuable in small dimensions might be useless in moderate or large dimensional search spaces.

Curse of Dimensionality

How long would it take to evaluate 10^{20} points?

Curse of Dimensionality

How long would it take to evaluate 10^{20} points?

```
import timeit
timeit.timeit('import numpy as np ;
np.sum(np.ones(10)*np.ones(10))', number=1000000)
> 7.0521080493927
```

7 seconds for 10^6 evaluations of $f(x) = \sum_{i=1}^{10} x_i^2$

We would need more than 10^8 days for evaluating 10^{20} points

[As a reference: origin of human species: roughly 6 x 10^8 days]

Separability

Given $f : x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$, let us define the 1-D functions that are cuts of f along the different coordinates:

$$f^i_{(x^i_1,\ldots,x^i_n)}(y) = f(x^i_1, \ldots, x^i_{i-1}, y, x^i_{i+1}, \ldots, x^i_n)$$

for $(x^i_1, \ldots, x^i_n) \in \mathbb{R}^{n-1}$, with $(x^i_1, \ldots, x^i_n) = (x^i_1, \ldots, x^i_{i-1}, x^i_{i+1}, \ldots, x^i_n)$

Definition: A function f is **separable** if for all i, for all $(x^i_1, \ldots, x^i_n) \in \mathbb{R}^{n-1}$, for all $(\hat{x}^i_1, \ldots, \hat{x}^i_n) \in \mathbb{R}^{n-1}$

$$\mathrm{argmin}_y f^i_{(x^i_1,\ldots,x^i_n)}(y) = \mathrm{argmin}_y f^i_{(\hat{x}^i_1,\ldots,\hat{x}^i_n)}(y)$$

a weak definition of separability

Separability (cont)

Proposition: Let f be a **separable** then for all x_i^j

$$\operatorname{argmin} f(x_1, \ldots, x_n) = \left(\operatorname{argmin} f^1_{(x_2^1, \ldots, x_n^1)}(x_1), \ldots, \operatorname{argmin} f^n_{(x_1^n, \ldots, x_{n-1}^n)}(x_n) \right)$$

and f can be optimized using n minimization along the coordinates.

Exercice: prove the previous proposition

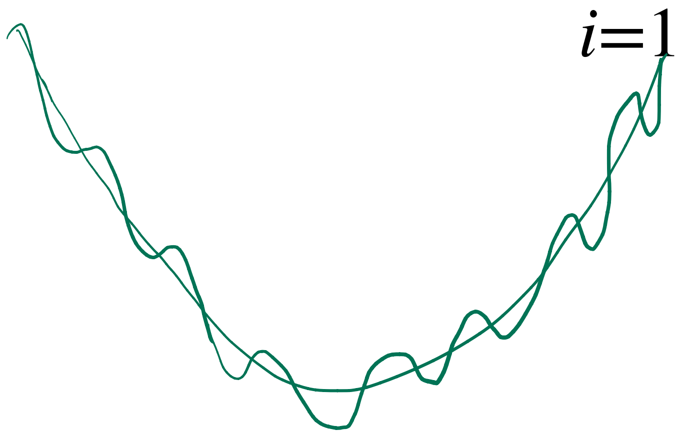
HOME EXERCICE

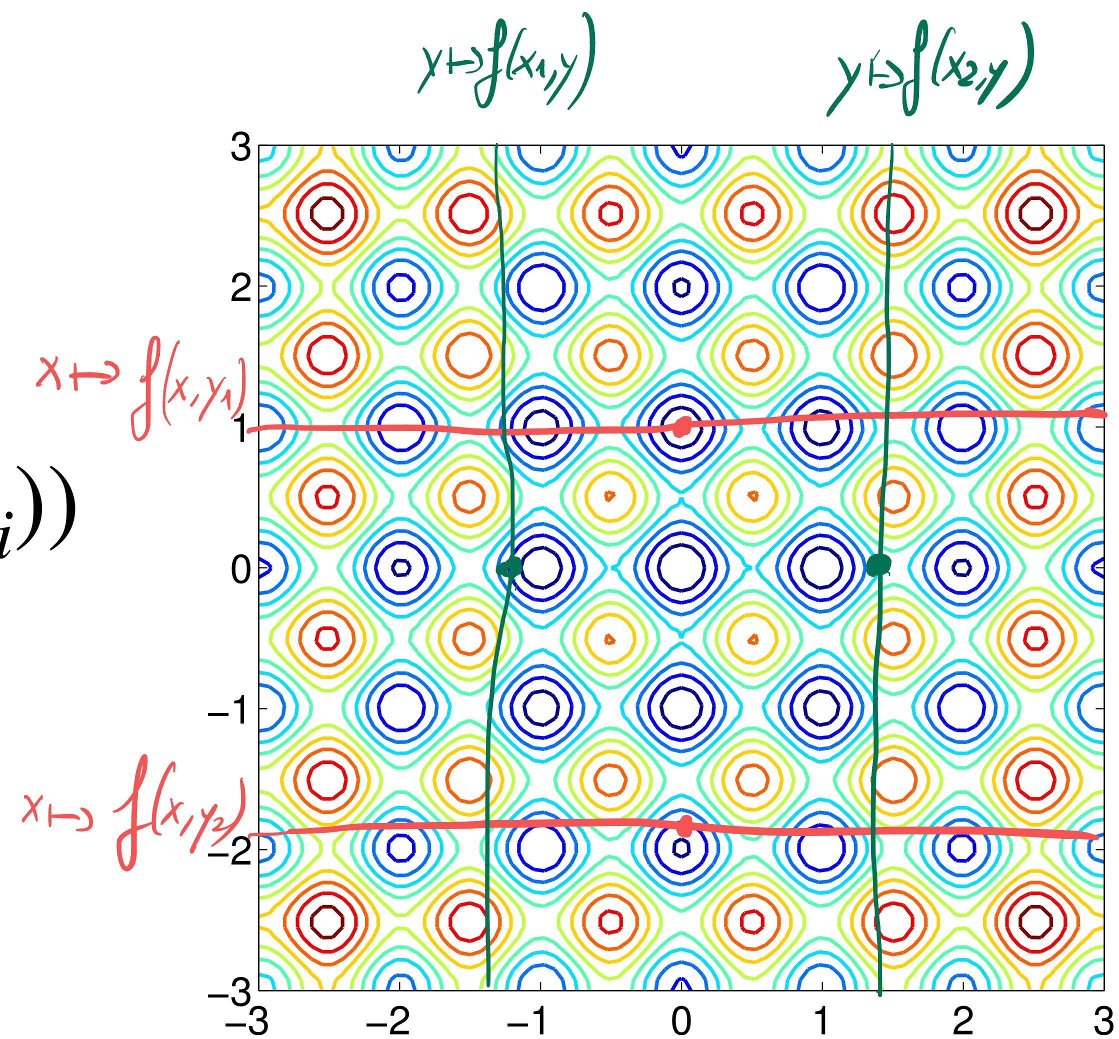
Example: Additively Decomposable Functions

Exercice: Let $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} h_i(x_i)$ for h_i having a unique argmin. Prove that f is separable. We say in this case that f is additively decomposable.

Example: Rastrigin function

$$f(x) = 10n + \sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i))$$





Non-separable Problems

Separable problems are typically easy to optimize. Yet **difficult real-word problems are non-separable**.

One needs to be careful when evaluating optimization algorithms that not too many test functions are separable and if so that the *algorithms do not exploit separability*.

> ***Otherwise:*** *good performance on test problems will not reflect good performance of the algorithm to solve difficult problems*

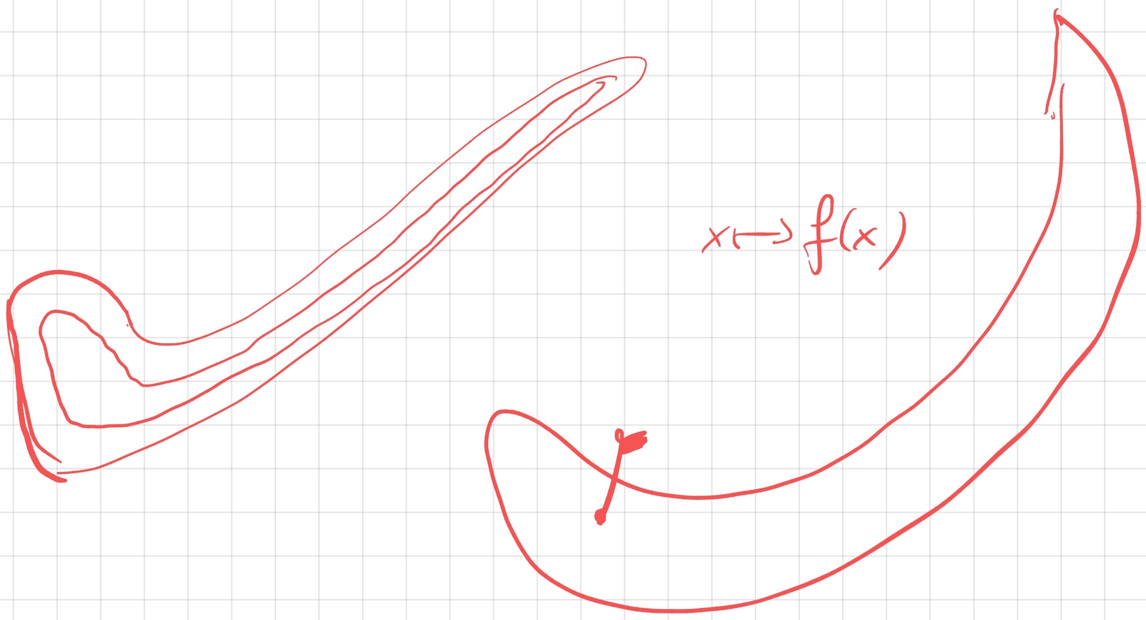
Algorithms known to exploit separability:

> Many Genetic Algorithms (GA), Most Particle Swarm Optimization (PSO)

If I give you $f : \begin{cases} \mathbb{R}^n \longrightarrow \mathbb{R} \\ x \longmapsto f(x) \end{cases}$ which is separable

How can you build a non-separable function?

<u>Rosenbrock function</u>:



$\exists\,!$ A linear $x \longmapsto f(Ax)$ is separable. <u>No.</u>

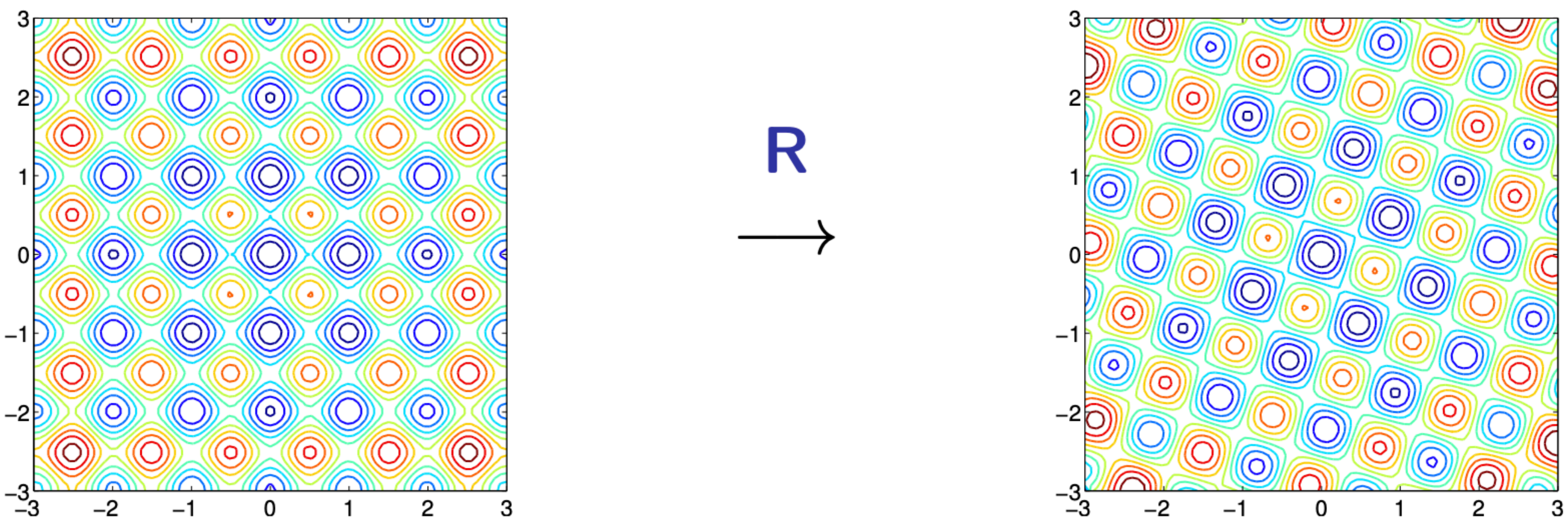
Non-separable Problems

Building a non-separable problem from a separable one

Rotating the coordinate system

- $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$ separable
- $f : \boldsymbol{x} \mapsto f(\mathbf{R}\boldsymbol{x})$ non-separable

R rotation matrix



[1] Hansen, Ostermeier, Gawelczyk (1995). On the adaptation of arbitrary normal mutation distributions in evolution strategies: The generating set adaptation. Sixth ICGA, pp. 57-64, Morgan Kaufmann

[2] Salomon (1996). "Reevaluating Genetic Algorithm Performance under Coordinate Rotation of Benchmark Functions; A survey of some theoretical and practical aspects of genetic algorithms." BioSystems, 39(3):263-278

Let $f(x) = \frac{1}{2} x^T A x$ where A is symmetric positive definite.

Is f separable?

If $A = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$, is f separable?

$$f(x) = \underbrace{\frac{9}{2} x_1^2}_{h_1(x_1)} + \underbrace{\frac{1}{2} x_2^2}_{h_2(x_2)} = h_1(x_1) + h_2(x_2)$$

f is then additively decomposable, so it is separable.

If A is diagonal, then f is separable.

If A si not diagonal, then f not separable.

$$f(x) = \frac{1}{2} x^T A x$$ where A is not diagonal

I can write f as the rotation of a separable function:

From the spectral theorem $A = P \underset{\text{Diagonal}}{D} P^T$

$$\underbrace{g(x) = \frac{1}{2} x^T D x}_{\text{Is separable because } D \text{ is diagonal}}$$

$$\begin{aligned} f(x) &= \frac{1}{2} x^T P D P^T x \\ &= \frac{1}{2} (P^T x)^T D \, P^T x \\ &= \underset{\text{separable}}{g} (\underset{\text{orthogonal}}{P^T} x) \end{aligned}$$

Rotation

Let f be convex quadratic, ie $f = \frac{1}{2}(x - x_0)^T A (x - x_0) + b$ where A is SPD.

$$\left(f \text{ is separable} \right) \iff \left(A \text{ is diagonal} \right)$$

In addition, any convex quadratic function can be written as

$$f(x) = g(Px) \quad \text{where } g \text{ is separable}$$
$$P \text{ is orthogonal}$$

Ill-conditioned Problems - Case of Convex-quadratic functions

Exercice: Consider a convex-quadratic function $f(x) = \frac{1}{2}(x - x^\star)^{\top} H(x - x^\star)$ with H a symmetric, positive, definite (SPD) matrix.

1. ~~why is it called a convex-quadratic function? What is the H~~ ~~matrix of~~ ~~f~~ ~~?~~

The condition number of the matrix H (with respect to the Euclidean norm) is defined as

$$\mathrm{cond}(H) = \frac{\lambda_{\max}(H)}{\lambda_{\min}(H)}$$

with $\lambda_{\max}()$ and $\lambda_{\min}()$ being respectively the largest and smallest eigenvalues.

Ill-conditioned Problems

Ill-conditioned means a high condition number of the [redacted] matrix H.

Consider now the specific case of the function $f(x) = \dfrac{1}{2}(x_1^2 + 9x_2^2)$

1. Compute its H[redacted] matrix, its condition number

2. Plots the level sets of f, relate the condition number to the axis ratio of the level sets of f

3. Generalize to a general convex-quadratic function

Real-world problems are often ill-conditioned.

4. Why to you think it is the case?

5. why are ill-conditioned problems difficult?

~~(see also **Exercice 2.5**)~~

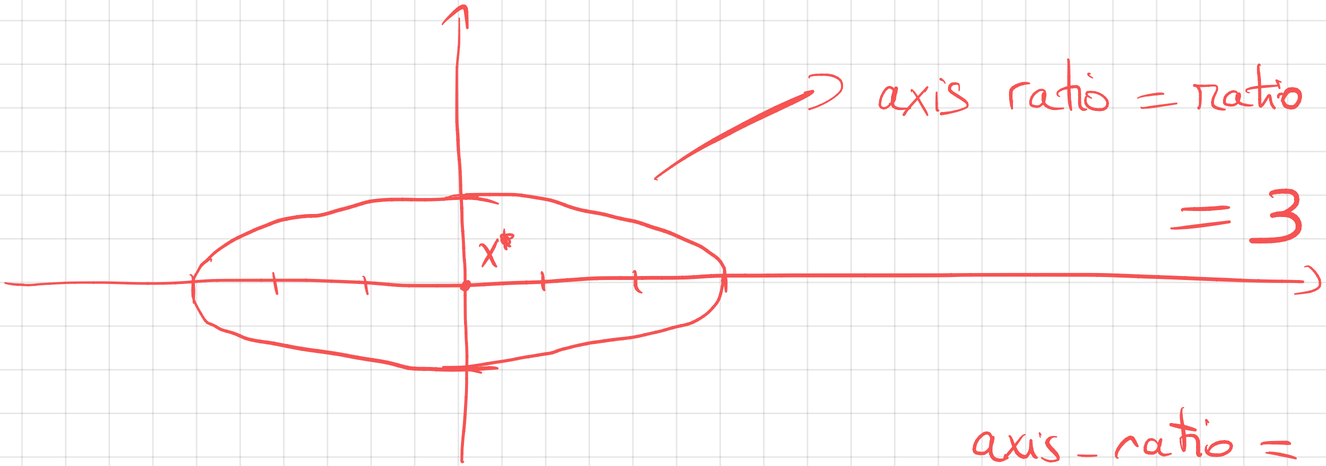
$$f(x) = \frac{1}{2}\left(x_1^2 + 9x_2^2\right) = \frac{1}{2}(x - x^*)^T H (x - x^*)$$

$$= \frac{1}{2} x^T \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} x \qquad \left(\underline{\text{i.e}} \quad x^* = 0, \quad H = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \right)$$

$$x = (x_1, x_2)$$

Then $H = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix}$ (and $x^* = 0$)

$$\text{cond}(H) = \frac{\lambda_{max}(H)}{\lambda_{min}(H)} = \frac{9}{1} = 9$$



axis ratio = ratio between largest and smallest axis of ellipsoid

$= 3$

$$\text{axis\_ratio} = \sqrt{\text{cond}(H)}$$

An ill-conditioned convex-quadratic problem will have a large ratio between the largest axis and smallest axis of the ellipsoid level set.

Why do we often encounter ill-conditioned problems (in the real world)?

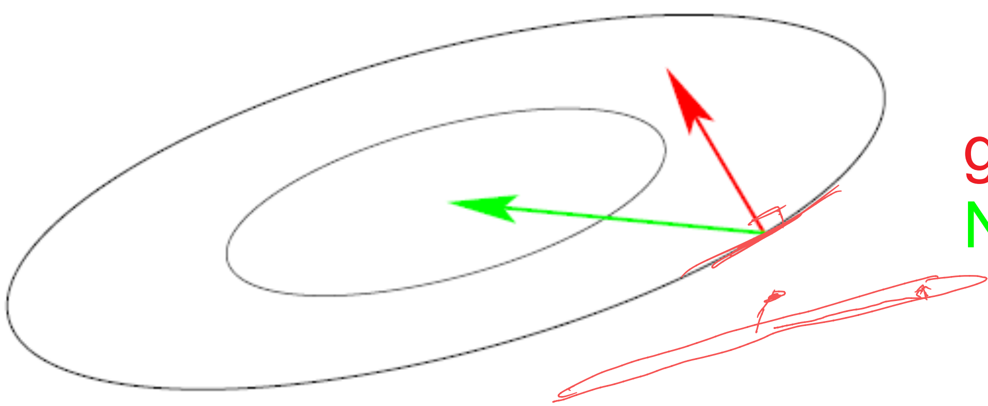
→ Because we optimize often variables that have different units/scales with different order of magnitude.

Ill-Conditioned Problems: Curvature of Level Sets

Consider the convex-quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T H(\boldsymbol{x} - \boldsymbol{x}^*) = \frac{1}{2}\sum_i h_{i,i} x_i^2 + \frac{1}{2}\sum_{i,j} h_{i,j} x_i x_j$$

H is Hessian matrix of f and symmetric positive definite



gradient direction $-f'(x)^T$

Newton direction $-H^{-1}f'(x)^T$

Ill-conditioning means squeezed level sets (high curvature). Condition number equals nine here. Condition numbers up to 10^{10} are not unusual in real-world problems.

If $H \approx I$ (small condition number of H) first order information (e.g. the gradient) is sufficient. Otherwise second order information (estimation of H^{-1}) information necessary.

Reminder: Different Notions of Optimum

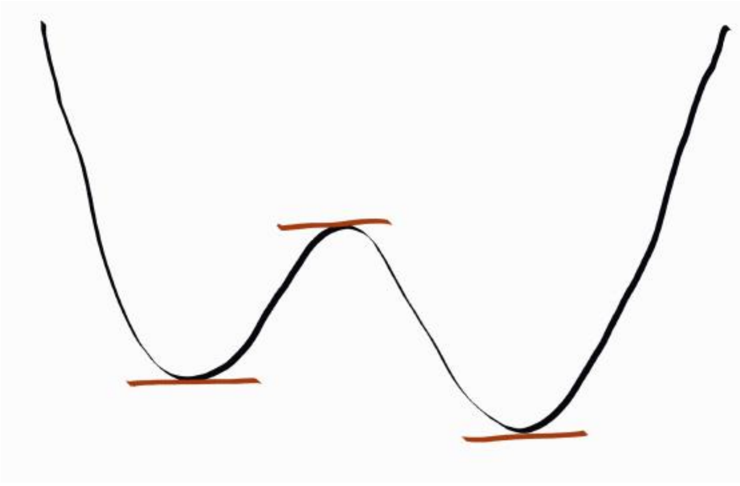
Unconstrained case

- local vs. global
  - local minimum $\boldsymbol{x}^*$: $\exists$ a neighborhood V of $\boldsymbol{x}^*$ such that $\forall \boldsymbol{x} \in \mathrm{V}: f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$
  - global minimum: $\forall \boldsymbol{x} \in \Omega: f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$
- strict local minimum if the inequality is strict

Mathematical Characterization of Optima

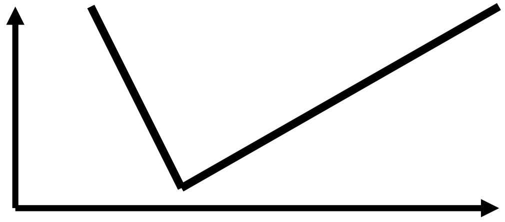
Objective: Derive general characterization of optima

Example: if $f: \mathbb{R} \to \mathbb{R}$ differentiable,
$f'(x) = 0$ at optimal points



- generalization to $f: \mathbb{R}^n \to \mathbb{R}$?
- generalization to constrained problems?

Remark: notion of optimum independent of notion of derivability

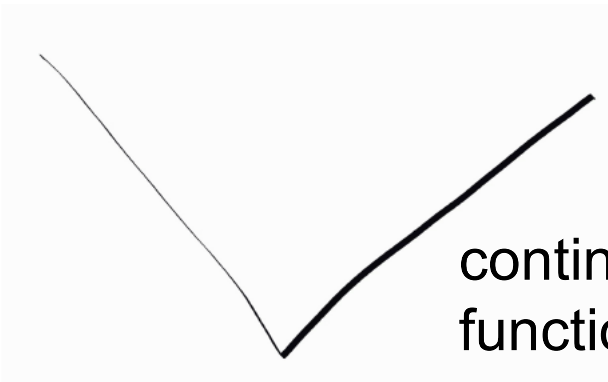


optima of such function can be easily approached by certain type of methods

Reminder: Continuity of a Function

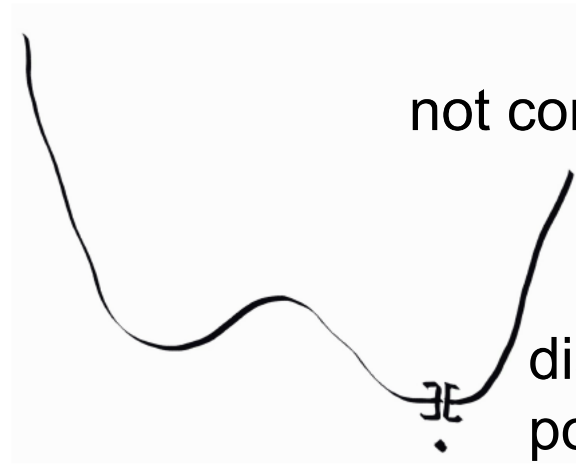
$f : (V, || \, ||_V) \longrightarrow (W, || \, ||_W)$ is continuous in $x \in V$ if

$\forall \epsilon > 0, \exists \eta > 0$ such that $\forall y \in V$: $||x - y||_V \leq \eta$; $||f(x) - f(y)||_W \leq \epsilon$

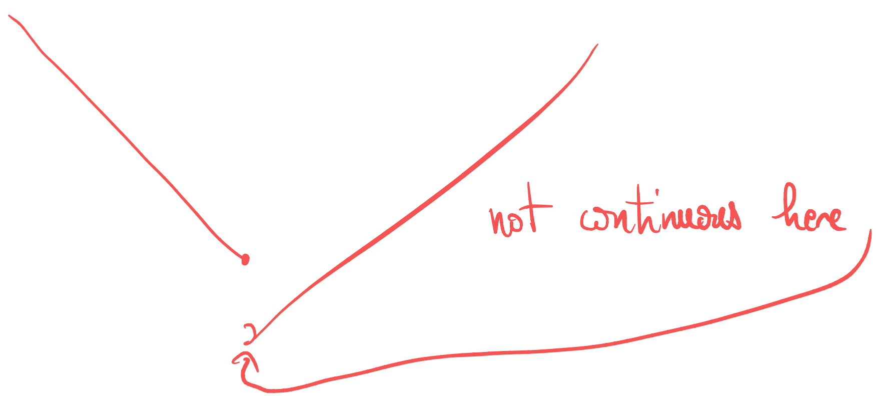


continuous function

not continuous



discontinuity point



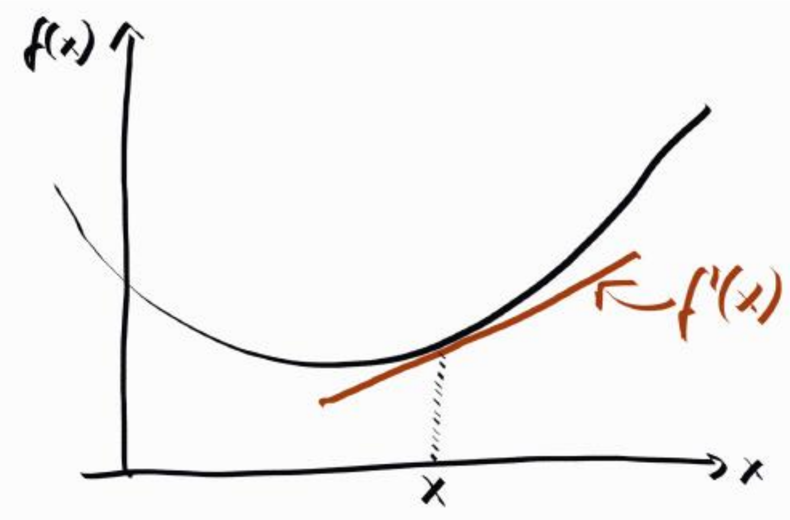
Reminder: Differentiability in 1D (n=1)

$f: \mathbb{R} \to \mathbb{R}$ is differentiable in $x \in \mathbb{R}$ if

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \text{ exists, } h \in \mathbb{R}$$

Notation:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$



The derivative corresponds to the slope of the tangent in x.