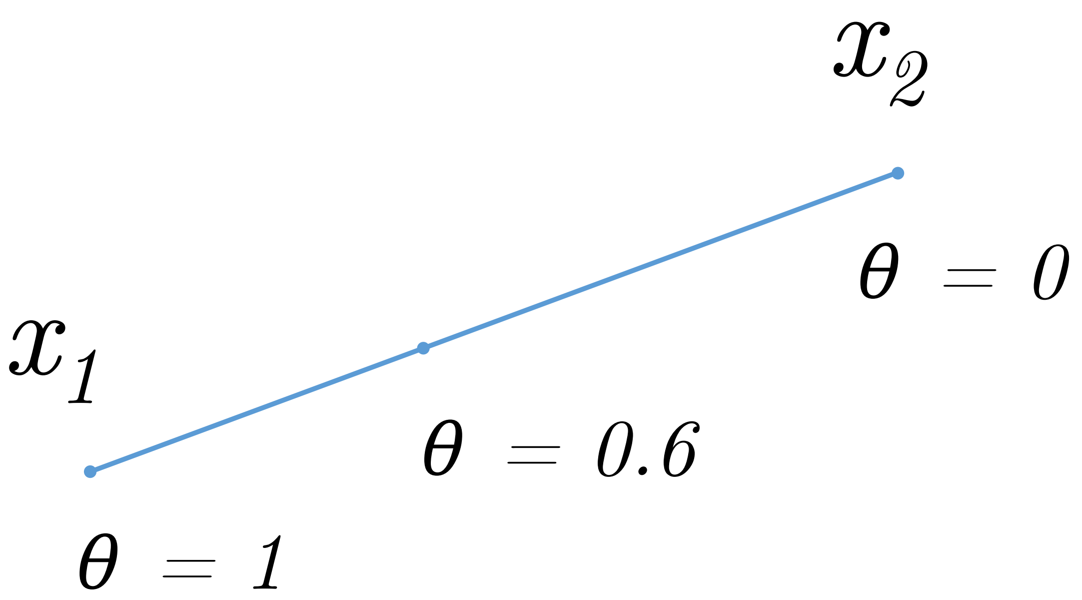
Convex set

Line segment

Suppose x_1, x_2 are two points in $\mathbb{R}^n$. Then the line segment between them is defined as follows:

$$x = \theta x_1 + (1 - \theta)x_2, \ \theta \in [0, 1]$$



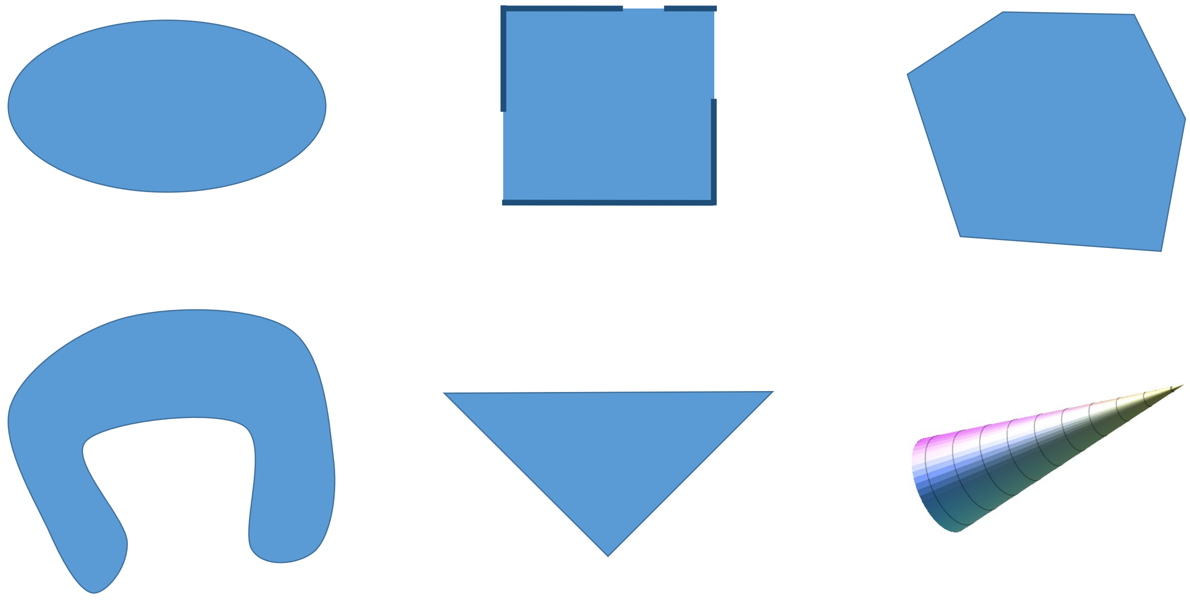
Convex set

The set S is called **convex** if for any x_1, x_2 from S the line segment between them also lies in S, i.e.

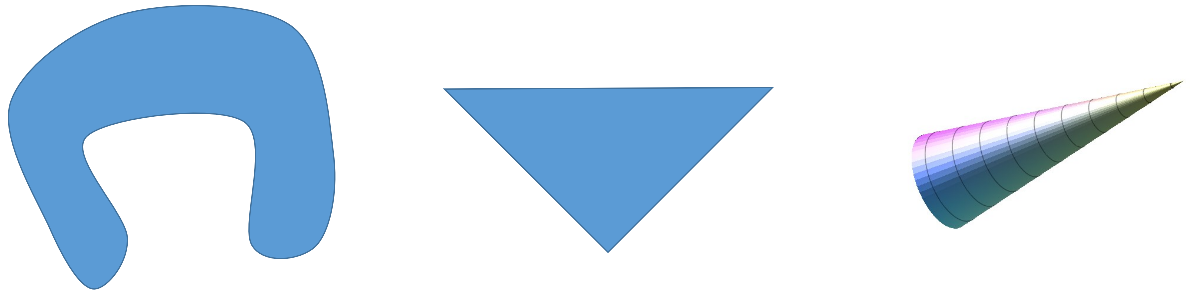
$$\forall \theta \in [0, 1], \ \forall x_1, x_2 \in S :$$
$$\theta x_1 + (1 - \theta)x_2 \in S$$

Examples:

- Any affine set
- Ray
- Line segment



BRO NOT BRO BRO



NOT BRO BRO BRO

Related definitions

Convex combination

Let $x_1, x_2, \ldots, x_k \in S$, then the point $\theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k$ is called the convex combination of points $x_1, x_2, \ldots, x_k$ if $\sum\limits_{i=1}^{k} \theta_i = 1, \; \theta_i \geq 0$

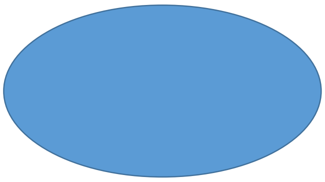
Convex hull

The set of all convex combinations of points from S is called the convex hull of the set S.

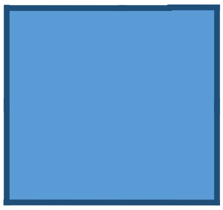
$$\mathbf{conv}(S) = \left\{ \sum_{i=1}^{k} \theta_i x_i \mid x_i \in S, \sum_{i=1}^{k} \theta_i = 1, \; \theta_i \geq 0 \right\}$$

- The set $\mathbf{conv}(S)$ is the smallest convex set containing S.
- The set S is convex if and only if $S = \mathbf{conv}(S)$.

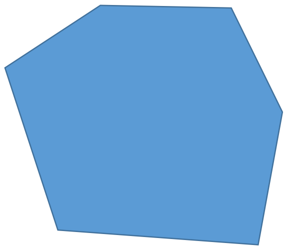
Examples:



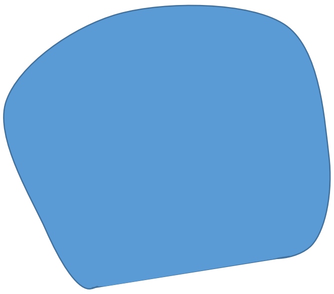
BRO



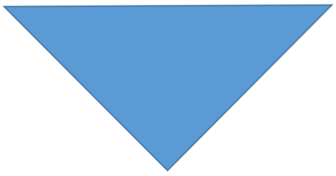
BRO



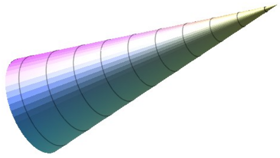
BRO



BRO



BRO



BRO

Finding convexity

In practice it is very important to understand whether a specific set is convex or not. Two approaches are used for this depending on the context.

- By definition.
- Show that S is derived from simple convex sets using operations that preserve convexity.

By definition

$$x_1, x_2 \in S, \ 0 \le \theta \le 1 \quad \rightarrow \quad \theta x_1 + (1-\theta)x_2 \in S$$

Preserving convexity

The linear combination of convex sets is convex

Let there be 2 convex sets S_x, S_y, let the set $S = \{s \mid s = c_1 x + c_2 y, \ x \in S_x, \ y \in S_y, \ c_1, c_2 \in \mathbb{R}\}$

Take two points from S: $s_1 = c_1 x_1 + c_2 y_1, s_2 = c_1 x_2 + c_2 y_2$ and prove that the segment between them $${theta s_1 + (1 - \theta)s_2, \theta \in [0,1]$$ also belongs to $$S$$

$$\theta s_1 + (1-\theta)s_2$$

$$\theta(c_1 x_1 + c_2 y_1) + (1-\theta)(c_1 x_2 + c_2 y_2)$$

$$c_1(\theta x_1 + (1-\theta)x_2) + c_2(\theta y_1 + (1-\theta)y_2)$$

$$c_1 x + c_2 y \in S$$

The intersection of any (!) number of convex sets is convex

If the desired intersection is empty or contains one point, the property is proved by definition. Otherwise, take 2 points and a segment between them. These points must lie in all intersecting sets, and since they are all convex, the segment between them lies in all sets and, therefore, in their intersection.

The image of the convex set under affine mapping is convex

$$
S \subseteq \mathbb{R}^n \text{ convex} \quad \rightarrow \quad f(S) = \{f(x) \mid x \in S\} \text{ convex} \quad (f(x) = \mathbf{A}x + \mathbf{b})
$$

Examples of affine functions: extension, projection, transposition, set of solutions of linear matrix inequality $\{x \mid x_1 A_1 + \ldots + x_m A_m \preceq B\}$ Here $A_i, B \in \mathbf{S}^p$ are symmetric matrices $p \times p$.

Note also that the prototype of the convex set under affine mapping is also convex.

$$
S \subseteq \mathbb{R}^m \text{ convex} \;\rightarrow\; f^{-1}(S) = \{x \in \mathbb{R}^n \mid f(x) \in S\} \text{ convex} \;\; (f(x) = \mathbf{A}x + \mathbf{b})
$$

Example 1

Prove, that ball in $\mathbb{R}^n$ (i.e. the following set $\{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_c\| \leq r\}$) - is convex.

Example 2

Which of the sets are convex: 1. Stripe, $\{x \in \mathbb{R}^n \mid \alpha \leq a^\top x \leq \beta\}$ 1. Rectangle, $\{x \in \mathbb{R}^n \mid \alpha_i \leq x_i \leq \beta_i, i = \overline{1, n}\}$ 1. Kleen, $\{x \in \mathbb{R}^n \mid a_1^\top x \leq b_1, a_2^\top x \leq b_2\}$ 1. A set of points closer to a given point than a given set that does not contain a point, $\{x \in \mathbb{R}^n \mid \|x - x_0\|_2 \leq \|x - y\|_2, \forall y \in S \subseteq \mathbb{R}^n\}$ 1. A set of points, which are closer to one set than another, $\{x \in \mathbb{R}^n \mid \mathbf{dist}(x, S) \leq \mathbf{dist}(x, T), S, T \subseteq \mathbb{R}^n\}$ 1. A set of points, $\{x \in \mathbb{R}^n \mid x + X \subseteq S\}$, where $S \subseteq \mathbb{R}^n$ is convex and $X \subseteq \mathbb{R}^n$ is arbitrary. 1. A set of points whose distance to a given point does not exceed a certain part of the distance to another given point is $\{x \in \mathbb{R}^n \mid \|x - a\|_2 \leq \theta \|xb\|_2, a, b \in \mathbb{R}^n, 0 \leq 1\}$

Example 3

Let $x \in \mathbb{R}$ is a random variable with a given probability distribution of $\mathbb{P}(x = a_i) = p_i$, where $i = 1, \dots, n$, and $a_1 < \dots < a_n$. It is said that the probability vector of outcomes of $p \in \mathbb{R}^n$ belongs to the probabilistic simplex, i.e.
$P = \{p \mid \mathbf{1}^T p = 1, p \succeq 0\} = \{p \mid p_1 + \dots + p_n = 1, p_i \geq 0\}$. Determine if the following sets of p are convex: 1. $\alpha < \mathbb{E}f(x) < \beta$, where $\mathbb{E}f(x)$ stands for expected value of $f(x) : \mathbb{R} \to \mathbb{R}$, i.e.
$\mathbb{E}f(x) = \sum\limits_{i=1}^{n} p_i f(a_i)$ 1. $\mathbb{E}x^2 \leq \alpha$ 1. $\mathbb{V}x \leq \alpha$

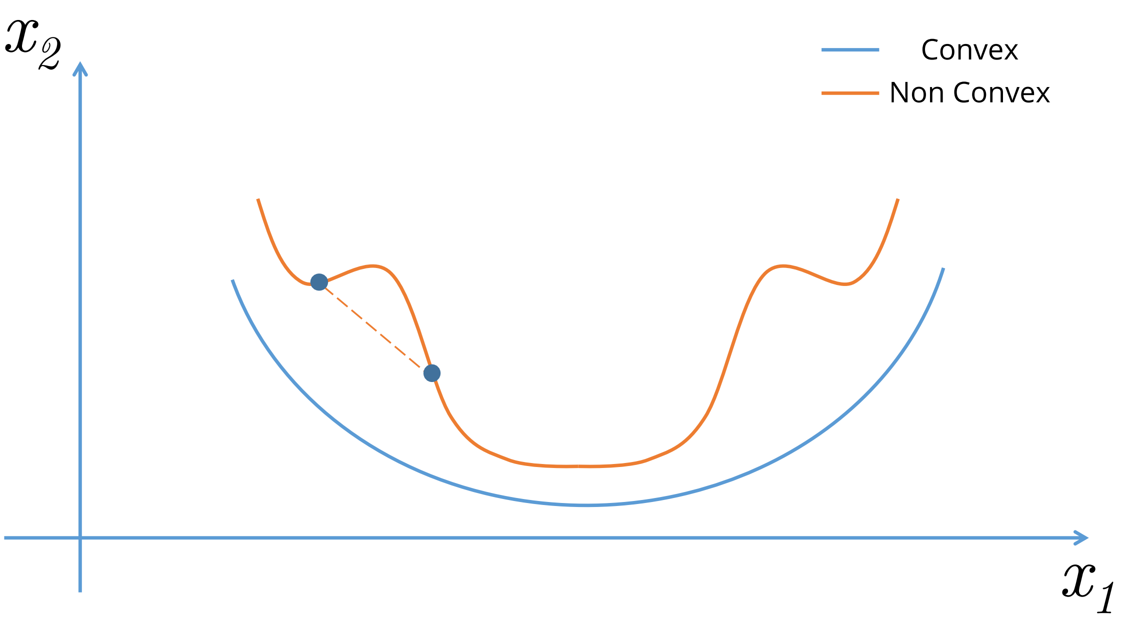
Convex function

Convex function

The function $f(x)$, **which is defined on the convex set** $S \subseteq \mathbb{R}^n$, is called **convex** S, if:

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

for any $x_1, x_2 \in S$ and $0 \leq \lambda \leq 1$.
If above inequality holds as strict inequality $x_1 \neq x_2$ and $0 < \lambda < 1$, then function is called strictly convex S



Examples

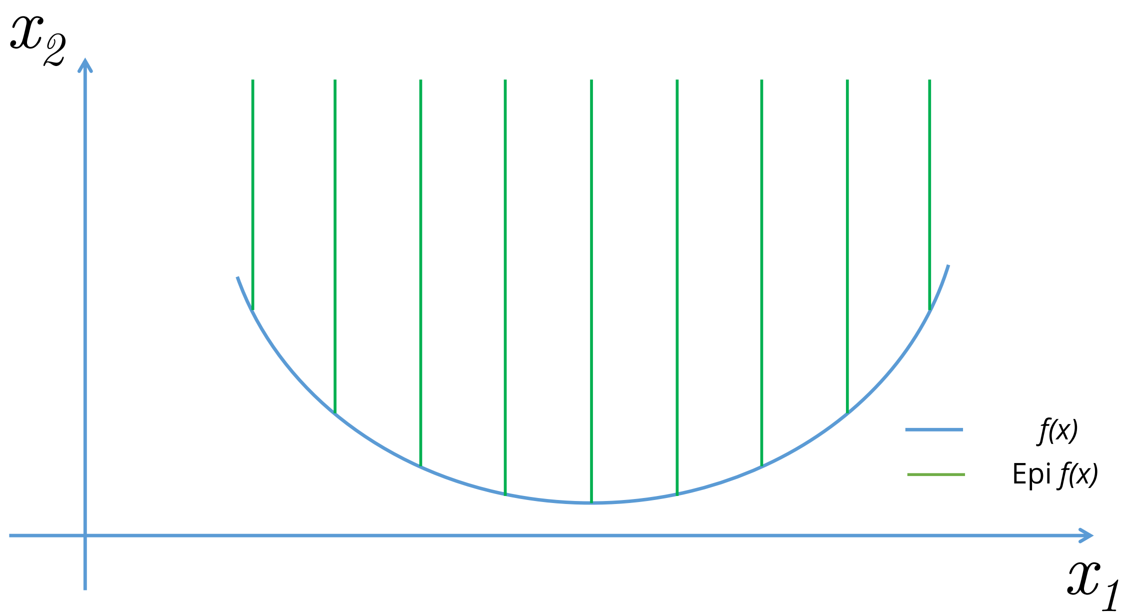
- $f(x) = x^p, p > 1, \quad S = \mathbb{R}_+$
- $f(x) = \|x\|^p, \quad p > 1, S = \mathbb{R}$
- $f(x) = e^{cx}, \quad c \in \mathbb{R}, S = \mathbb{R}$
- $f(x) = -\ln x, \quad S = \mathbb{R}_{++}$
- $f(x) = x \ln x, \quad S = \mathbb{R}_{++}$
- The sum of the largest k coordinates $f(x) = x_{(1)} + \ldots + x_{(k)}, \quad S = \mathbb{R}^n$
- $f(X) = \lambda_{max}(X), \quad X = X^T$
- $f(X) = -\log \det X, \quad S = S^n_{++}$

Epigraph

For the function $f(x)$, defined on $S \subseteq \mathbb{R}^n$, the following set:

$$\text{epi } f = \{[x, \mu] \in S \times \mathbb{R} : f(x) \le \mu\}$$

is called **epigraph** of the function $f(x)$

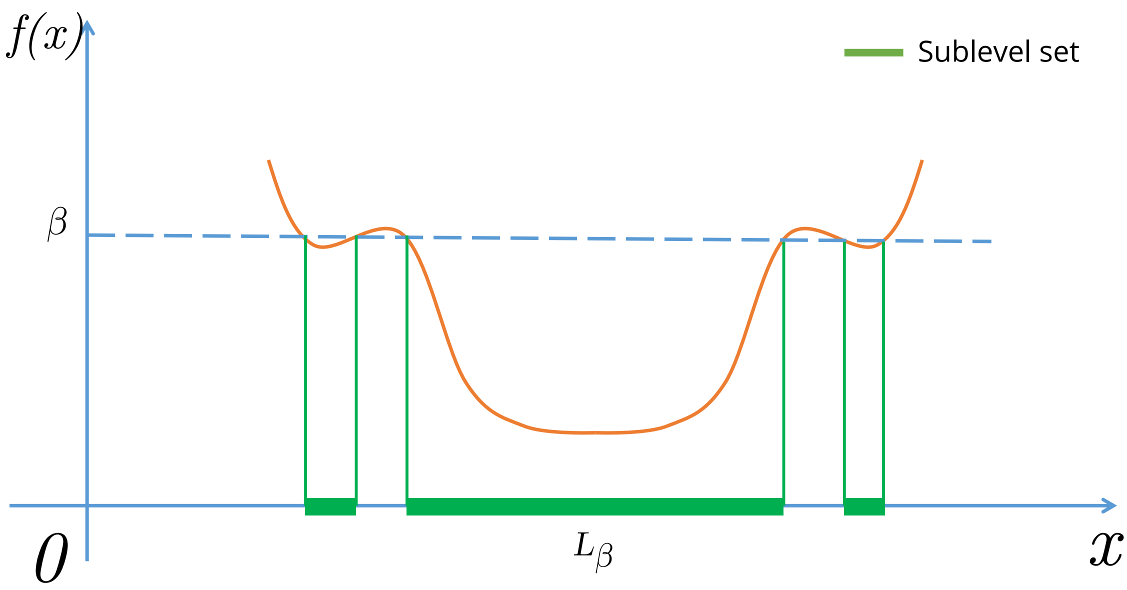


Sublevel set

For the function $f(x)$, defined on $S \subseteq \mathbb{R}^n$, the following set:

$$\mathcal{L}_\beta = \{x \in S : f(x) \le \beta\}$$

is called **sublevel set** or Lebesgue set of the function $f(x)$



Criteria of convexity

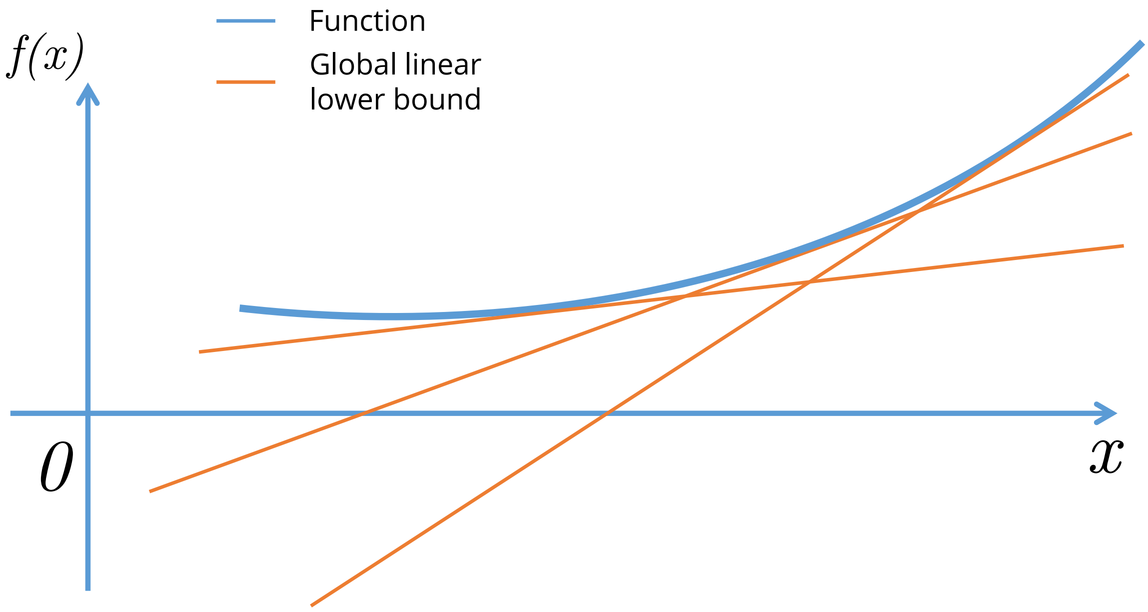
First order differential criterion of convexity

The differentiable function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is convex if and only if $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y - x)$$

Let $y = x + \Delta x$, then the criterion will become more tractable:

$$f(x + \Delta x) \geq f(x) + \nabla f^T(x)\Delta x$$



Second order differential criterion of convexity

Twice differentiable function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is convex if and only if $\forall x \in \mathbf{int}(S) \neq \emptyset$:

$$\nabla^2 f(x) \succeq 0$$

In other words, $\forall y \in \mathbb{R}^n$:

$$\langle y, \nabla^2 f(x) y \rangle \geq 0$$

Connection with epigraph

The function is convex if and only if its epigraph is convex set.

Connection with sublevel set

If $f(x)$ - is a convex function defined on the convex set $S \subseteq \mathbb{R}^n$, then for any β sublevel set $\mathcal{L}_\beta$ is convex.

The function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is closed if and only if for any β sublevel set $\mathcal{L}_\beta$ is closed.

Reduction to a line

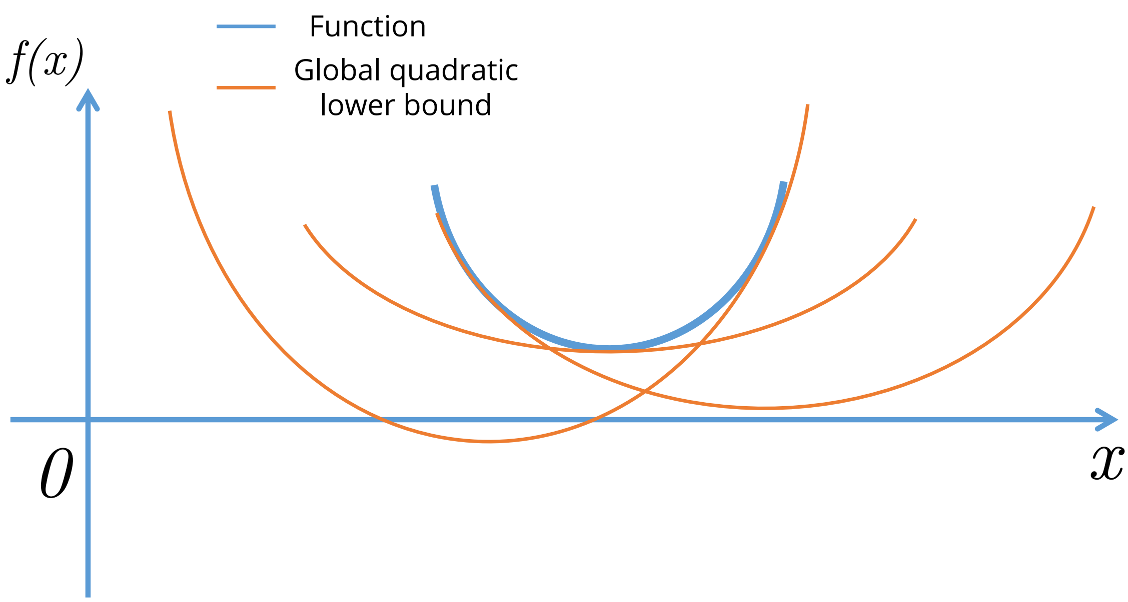
$f : S \to \mathbb{R}$ is convex if and only if S is convex set and the function $g(t) = f(x + tv)$ defined on $\{t \mid x + tv \in S\}$ is convex for any $x \in S, v \in \mathbb{R}^n$, which allows to check convexity of the scalar function in order to establish covexity of the vector function.

Strong convexity

$f(x)$, **defined on the convex set** $S \subseteq \mathbb{R}^n$, is called μ-strongly convex (strogly convex) on S, if:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) - \mu\lambda(1-\lambda)\|x_1 - x_2\|^2$$

for any $x_1, x_2 \in S$ and $0 \leq \lambda \leq 1$ for some $\mu > 0$.



Criteria of strong convexity

First order differential criterion of strong convexity

Differentiable $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ μ-strongly convex if and only if $\forall x, y \in S$:

$$f(y) \geq f(x) + \nabla f^T(x)(y - x) + \frac{\mu}{2}\|y - x\|^2$$

Let $y = x + \Delta x$, then the criterion will become more tractable:

$$
f(x + \Delta x) \geq f(x) + \nabla f^T(x)\Delta x + \frac{\mu}{2}\|\Delta x\|^2
$$

Second order differential criterion of strong convexity

Twice differentiable function $f(x)$ defined on the convex set $S \subseteq \mathbb{R}^n$ is called μ-strongly convex if and only if $\forall x \in \mathbf{int}(S) \neq \emptyset$:

$$
\nabla^2 f(x) \succeq \mu I
$$

In other words:

$$
\langle y, \nabla^2 f(x) y\rangle \geq \mu \|y\|^2
$$

Facts

- $f(x)$ is called (strictly) concave, if the function $-f(x)$ - (strictly) convex.
- Jensen's inequality for the convex functions:
$$
f\left(\sum_{i=1}^n \alpha_i x_i\right) \leq \sum_{i=1}^n \alpha_i f(x_i)
$$
for $\alpha_i \geq 0; \quad \sum\limits_{i=1}^n \alpha_i = 1$ (probability simplex)
For the infinite dimension case:
$$
f\left(\int\limits_S x p(x)dx\right) \leq \int\limits_S f(x)p(x)dx
$$
If the integrals exist and $p(x) \geq 0, \quad \int\limits_S p(x)dx = 1$
- If the function $f(x)$ and the set S are convex, then any local minimum $x^* = \arg\min\limits_{x \in S} f(x)$ will be the global one. Strong convexity guarantees the uniqueness of the solution.

Operations that preserve convexity

- Non-negative sum of the convex functions: $\alpha f(x) + \beta g(x), (\alpha \geq 0, \beta \geq 0)$
- Composition with affine function $f(Ax + b)$ is convex, if $f(x)$ is convex
- Pointwise maximum (supremum): If $f_1(x), \ldots, f_m(x)$ are convex, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex
- If $f(x, y)$ is convex on x for any $y \in Y$: $g(x) = \sup\limits_{y \in Y} f(x, y)$ is convex
- If $f(x)$ is convex on S, then $g(x, t) = tf(x/t)$ - is convex with $x/t \in S, t > 0$
- Let $f_1 : S_1 \to \mathbb{R}$ and $f_2 : S_2 \to \mathbb{R}$, where $\operatorname{range}(f_1) \subseteq S_2$. If f_1 and f_2 are convex, and f_2 is increasing, then $f_2 \circ f_1$ is convex on S_1

Other forms of convexity

- Log-convex: $\log f$ is convex; Log convexity implies convexity.
- Log-concavity: $\log f$ concave; **not** closed under addition!
- Exponentially convex: $[f(x_i + x_j)] \succeq 0$, for $x_1, \ldots, x_n$
- Operator convex: $f(\lambda X + (1 - \lambda)Y) \preceq \lambda f(X) + (1 - \lambda) f(Y)$
- Quasiconvex: $f(\lambda x + (1 - \lambda)y) \leq \max\{f(x), f(y)\}$
- Pseudoconvex: $\langle \nabla f(y), x - y\rangle \geq 0 \longrightarrow f(x) \geq f(y)$

* Discrete convexity: $f : \mathbb{Z}^n \to \mathbb{Z}$; “convexity + matroid theory.”

References

* Steven Boyd lectures
* Suvrit Sra lectures
* Martin Jaggi lectures

Example 4

Show, that $f(x) = c^\top x + b$ is convex and concave.

Example 5

Show, that $f(x) = x^\top Ax$, where $A \succeq 0$ - is convex on $\mathbb{R}^n$.

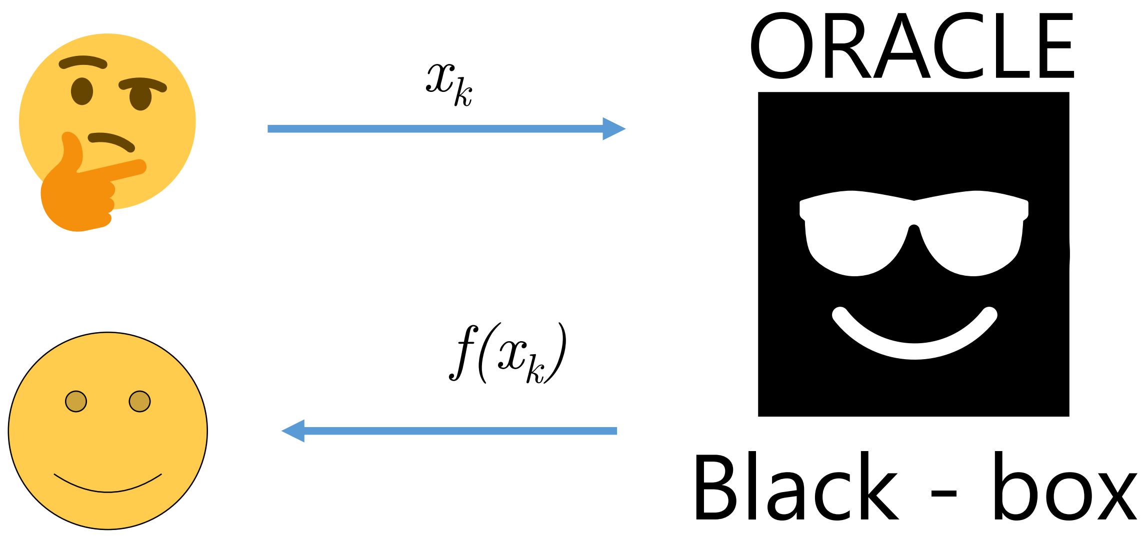
Example 6

Show, that $f(x)$ is convex, using first and second order criteria, if $f(x) = \sum\limits_{i=1}^n x_i^4$.

Example 7

Find the set of $x \in \mathbb{R}^n$, where the function $f(x) = \dfrac{-1}{2(1 + x^\top x)}$ is convex, strictly convex, strongly convex?

Zero order methods



Now we have only zero order information from the oracle. Typical speed of convegence of these methods is sublinear. A lot of methods are referred both to zero order methods and global optimization.

Code

- Global optimization illustration - Open in Colab
- Nevergrad library - Open in Colab
- Optuna quickstart Open in Colab

Simulated annealing

Problem

We need to optimize the global optimum of a given function on some space using only the values of the function in some points on the space.

$$\min_{x \in X} F(x) = F(x^*)$$

Simulated Annealing is a probabilistic technique for approximating the global optimum of a given function.

Algorithm

The name and inspiration come from annealing in metallurgy, a technique involving heating and controlled cooling of a material to increase the size of its crystals and reduce their defects. Both are attributes of the material that depend on its thermodynamic free energy. Heating and cooling the material affects both the temperature and the thermodynamic free energy. The simulation of annealing can be used to find an approximation of a global minimum for a function with many variables.

Steps of the Algorithm

Step 1 Let $k = 0$ - current iteration, $T = T_k$ - initial temperature.

Step 2 Let $x_k \in X$ - some random point from our space

Step 3 Let decrease the temperature by following rule $T_{k+1} = \alpha T_k$ where $0 < \alpha < 1$ - some constant that often is closer to 1

Step 4 Let $x_{k+1} = g(x_k)$ - the next point which was obtained from previous one by some random rule. It is usually assumed that this rule works so that each subsequent approximation should not differ very much.

Step 5 Calculate $\Delta E = E(x_{k+1}) - E(x_k)$, where $E(x)$ - the function that determines the energy of the system at this point. It is supposed that energy has the minimum in desired value x^*.

Step 6 If $\Delta E < 0$ then the approximation found is better than it was. So accept x_{k+1} as new started point at the next step and go to the step **Step 3**

Step 7 If $\Delta E >= 0$, then we accept x_{k+1} with the probability of $P(\Delta E) = \exp^{-\Delta E/T_k}$. If we don't accept x_{k+1}, then we let $k = k + 1$. Go to the step **Step 3**

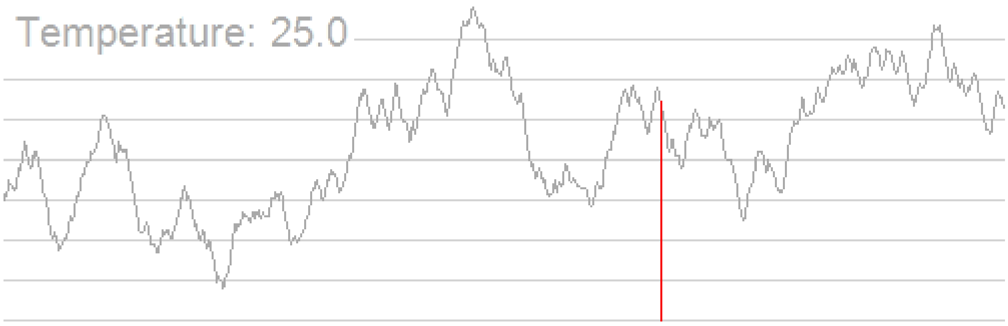
The algorithm can stop working according to various criteria, for example, achieving an optimal state or lowering the temperature below a predetermined level T_{min}.

Convergence

As it mentioned in Simulated annealing: a proof of convergence the algorithm converges almost surely to a global maximum.

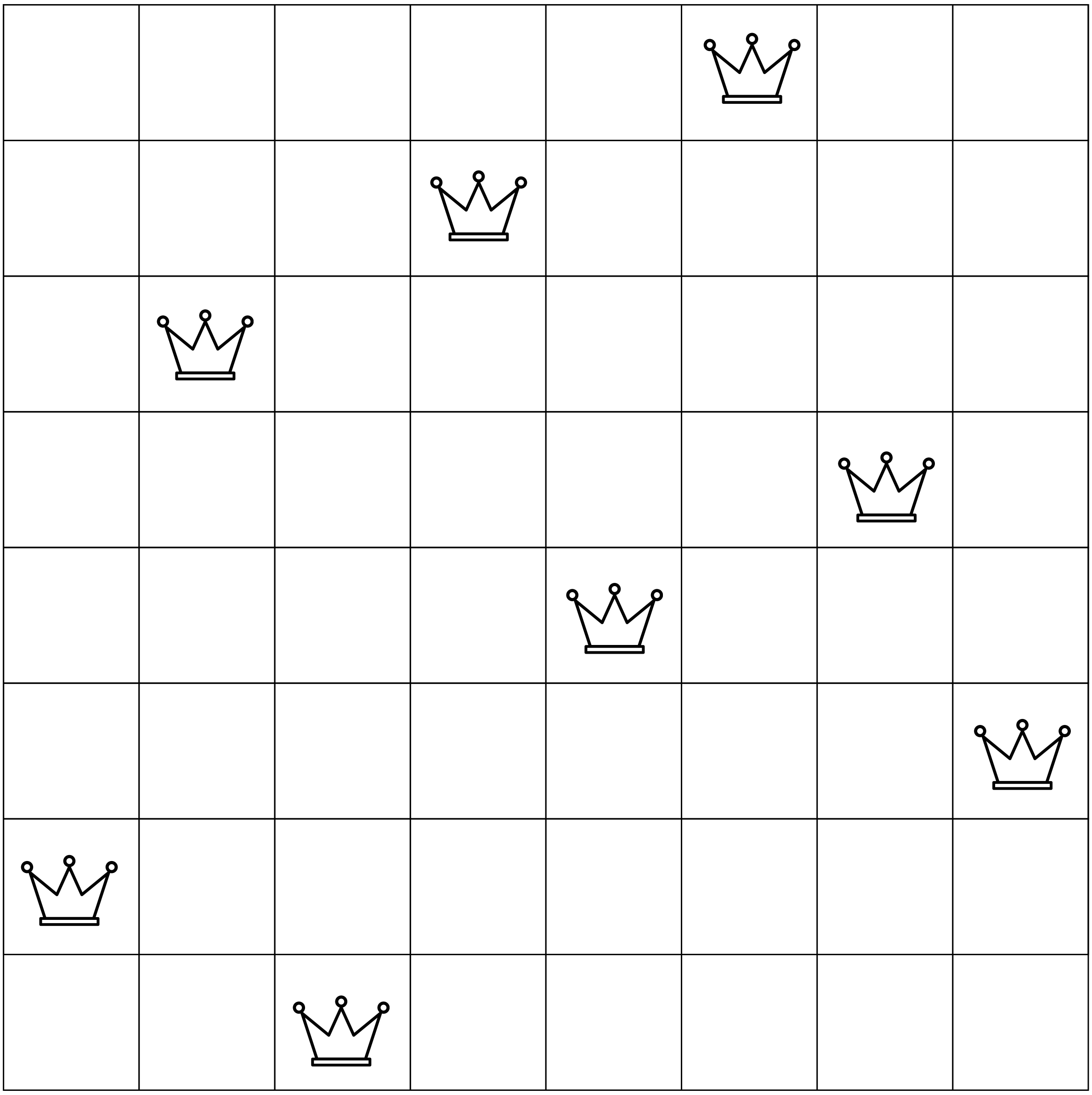
Illustration

A gif from Wikipedia:



Example

In our example we solve the N queens puzzle - the problem of placing N chess queens on an N×N chessboard so that no two queens threaten each other.



The Problem

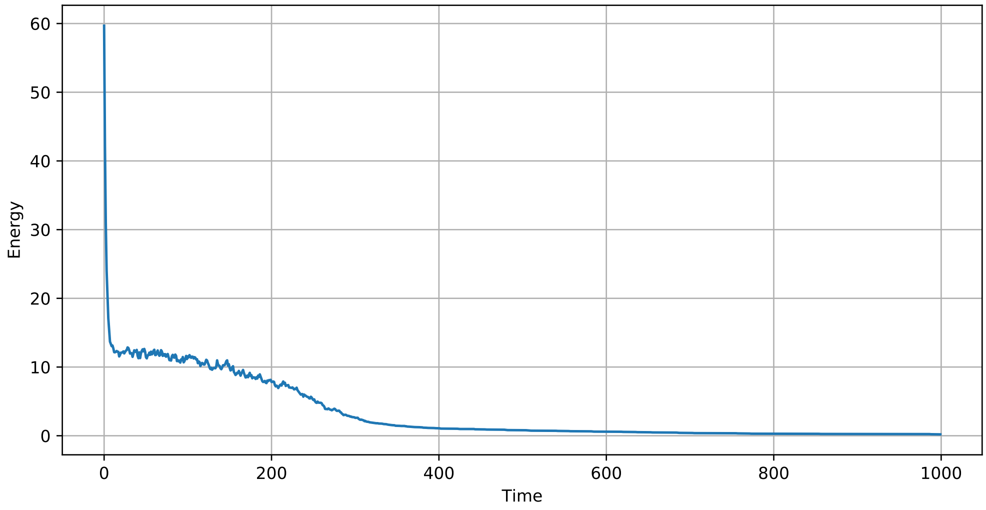
Let $E(x)$ - the number of intersections, where x - the array of placement queens at the field (the number in array means the column, the index of the number means the row).

The problem is to find x^* where $E(x^*) = \min_{x \in X} E(x)$ - the global minimum, that is predefined and equals to 0 (no two queens threaten each other).

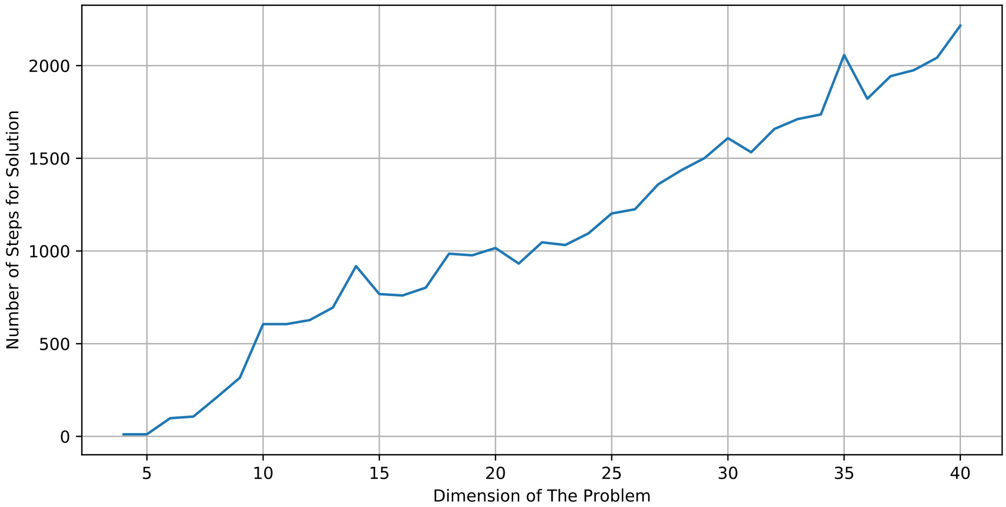
In this code $x_0 = [0, 1, 2, \dots, N]$ that means all queens are placed at the board's diagonal . So at the beginning $E = N(N-1)$, because every queen intersects others.

Results

Results of applying this algorithm with $\alpha = 0.95$ to the N queens puzzle for $N = 10$ averaged by 100 runs are below:



Results of running the code for N from 4 to 40 and measuring the time it takes to find the solution averaged by 100 runs are below:



[Open in Colab]

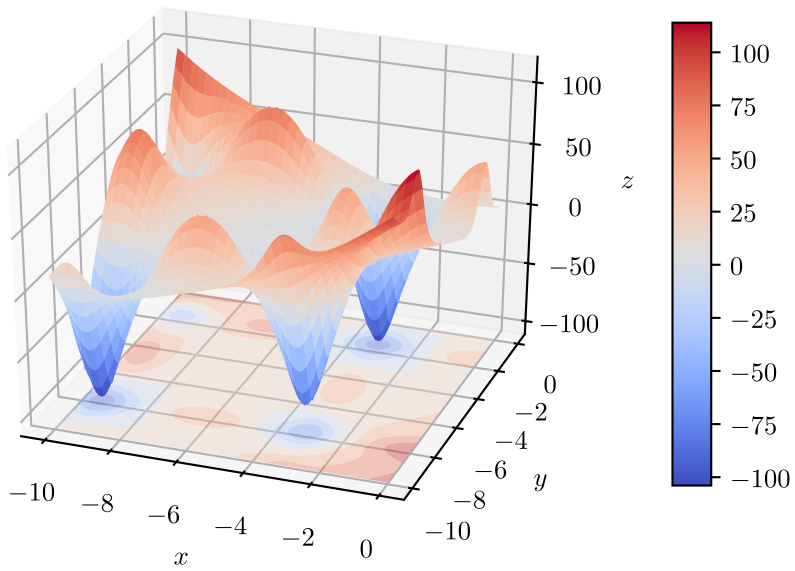
Nelder–Mead

Problem

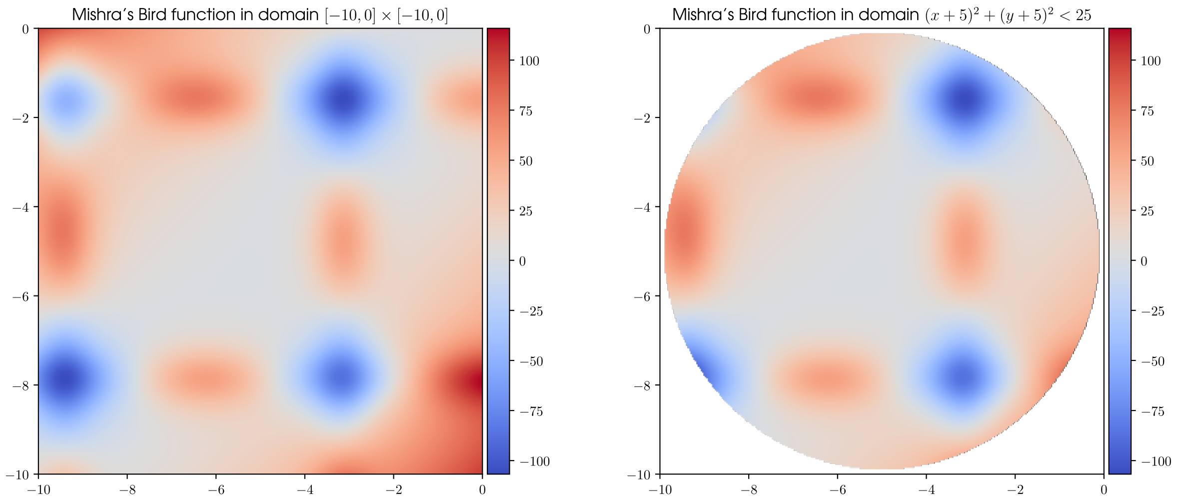
Sometimes the multidimensional function is so difficult to evaluate that even expressing the 1^{st} derivative for gradient-based methods of finding optimum becomes an impossible task. In this case, we can only rely on the values of the function at each point. Or, in other words, on the 0 order oracle calls.

Let's take, for instance, Mishra's Bird function:

$$f(x, y) = \sin y \cdot e^{(1-\cos x)^2} + \cos x \cdot e^{(1-\sin y)^2} + (x-y)^2$$



This function is usually subjected to the domain $(x+5)^2 + (y+5)^2 < 25$, but for the sake of picture beauty we will mainly use domain $[-10; 0] \times [-10; 0]$.



Algorithm

Related definitions:

- **Simplex** -- polytope with the least possible number of vertices in n-dimensional space. (So, it's $(n+1)$-polytope.) In our $2D$ case it will be triangle.
- **Best point** x_1 -- vertex of the simplex, function value in which is the smallest among all vertices.
- **Worst point** x_{n+1} -- vertex of the simplex, function value in which is the largest among all vertices.
- **Other points** $x_2, \dots, x_n$ -- vertices of the simplex, ordered in such way that $f(x_1) \leqslant f(x_2) \leqslant \dots \leqslant f(x_n) \leqslant f(x_{n+1})$. This implies that $\{x_1, x_2, \dots, x_n\}$ are best points in relation to x_{n+1} and $\{x_2, \dots, x_n, x_{n+1}\}$ are worst points in relation to x_1.
- **Centroid** x_o -- center of mass in the polytope. In Nelder-Mead the centroid is calculated for the polytope, constituted by best vertices. In our $2D$ case it will be the center of the triangle side, which contains 2 best points $x_o = \dfrac{x_1 + x_2}{2}$.

Main idea

The algorithm maintains the set of test points in the form of simplex. For each point the function value is calculated and points are ordered accordingly. Depending on those values, the simplex exchanges the worst point of the set for the new one, which is closer to the local minimum. In some sense, the simplex is crawling to the minimal value in the domain.

The simplex movements finish when its sides become too small (termination condition by sides) or its area becomes too small (termination condition by area). I prefer the second condition, because it takes into account cases when simplex becomes degenerate (three or more vertices on one axis).

Steps of the algorithm

1. Ordering

Order vertices according to values in them:

$$f(x_1) \leqslant f(x_2) \leqslant \ldots \leqslant f(x_n) \leqslant f(x_{n+1})$$

Check the termination condition. Possible exit with solution $x_{\min} = x_1$.

2. Centroid calculation

$$x_o = \frac{\sum\limits_{k=1}^{n} x_k}{n}$$

3. Reflection

Calculate the reflected point x_r:

$$x_r = x_o + \alpha\left(x_o - x_{n+1}\right)$$

where α -- reflection coefficient, $\alpha > 0$. (If $\alpha \leqslant 0$, reflected point x_r will not overlap the centroid)

The next step is figured out according to the value of $f(x_r)$ in dependency to values in points x_1 (best) and x_n (second worst):

- $f(x_r) < f(x_1)$: Go to step 4.
- $f(x_1) \leqslant f(x_r) < f(x_n)$: new simplex with $x_{n+1} \rightarrow x_r$. Go to step 1.
- $f(x_r) \geqslant f(x_n)$: Go to step 5.

4. Expansion

Calculate the expanded point x_e:

$$x_e = x_o + \gamma\left(x_r - x_o\right)$$

where γ -- expansion coefficient, $\gamma > 1$. (If $\gamma < 1$, expanded point x_e will be contracted towards centroid, if $\gamma = 1$: $x_e = x_r$)

The next step is figured out according to the ratio between $f(x_e)$ and $f(x_r)$:

- $f(x_e) < f(x_r)$: new simplex with $x_{n+1} \rightarrow x_e$. Go to step 1.
- $f(x_e) > f(x_r)$: new simplex with $x_{n+1} \rightarrow x_r$. Go to step 1.

5. Contraction

Calculate the contracted point x_c:

$$x_c = x_o + \beta\left(x_{n+1} - x_o\right)$$

where β -- contraction coefficient, $0 < \beta \leqslant 0.5$. (If $\beta > 0.5$, contraction is insufficient, if $\beta \leqslant 0$, contracted point x_c overlaps the centroid)

The next step is figured out according to the ratio between $f(x_c)$ and $f(x_{n+1})$:

- $f(x_c) < f(x_{n+1})$: new simplex with $x_{n+1} \to x_c$. Go to step 1.
- $f(x_c) \geqslant f(x_{n+1})$: Go to step 6.

6. Shrinkage

Replace all points of simplex x_i with new ones, except for the best point x_1:

$$x_i = x_1 + \sigma\left(x_i - x_1\right)$$

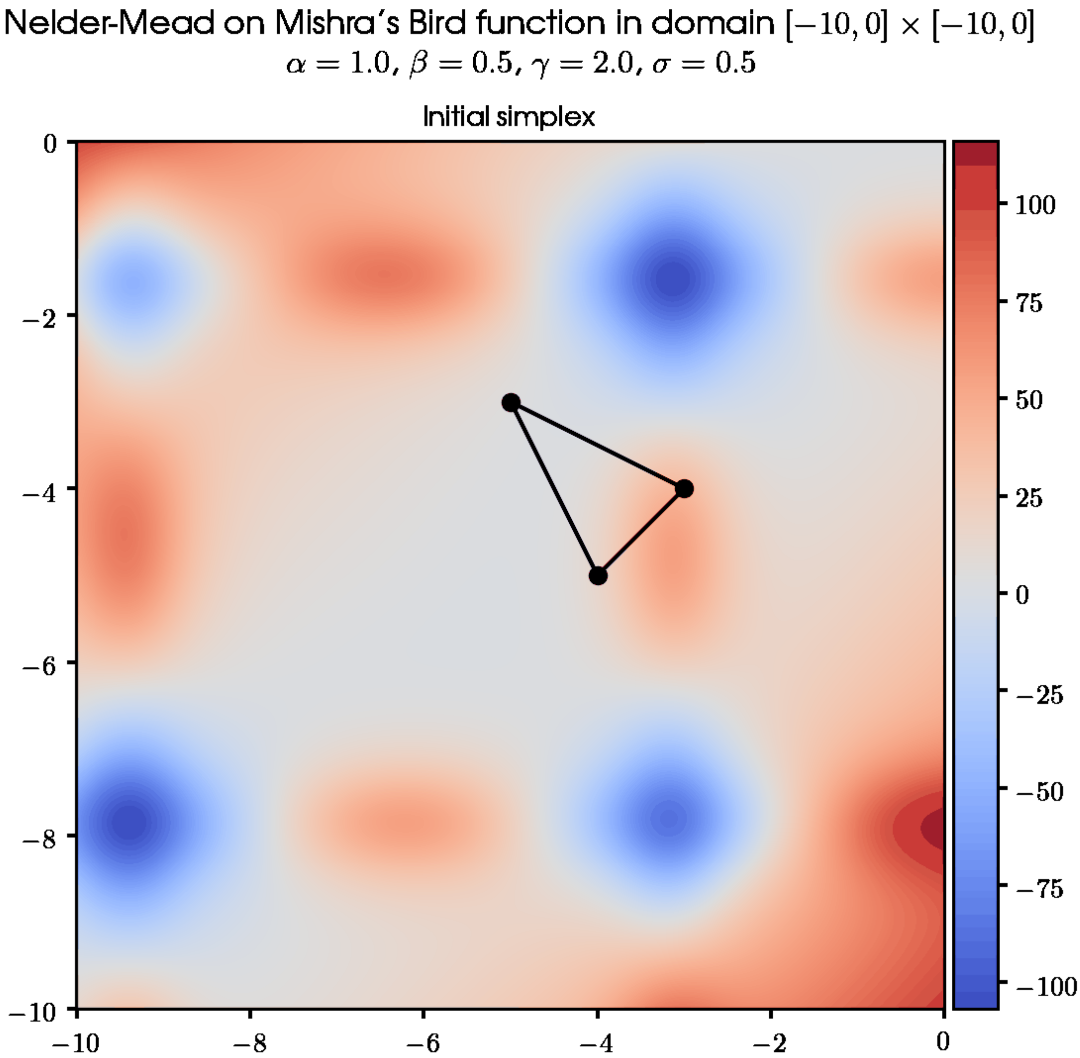
where σ -- shrinkage coefficient, $0 < \sigma < 1$. (If $\sigma \geqslant 1$, shrinked point x_i overlaps the best point x_1, if $\sigma \leqslant 0$, shrinked point x_i becomes extended)

Go to step 1.

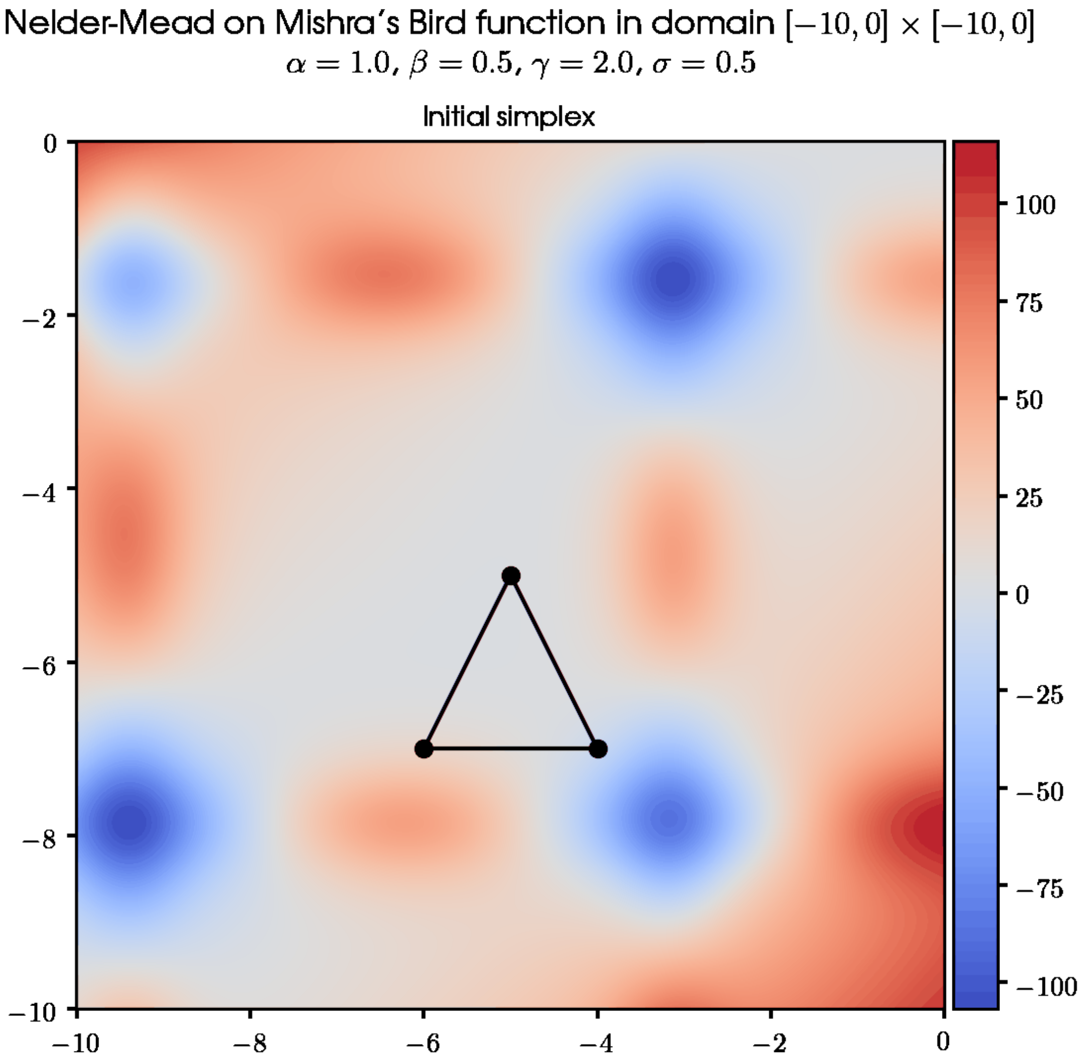
Examples

This algorithm, as any method in global optimization, is highly dependable on the initial coonditions. For instance, if we use different initial simplex or different set of parameters $\{\alpha, \beta, \gamma, \sigma\}$ the resulting optimal point will differ.

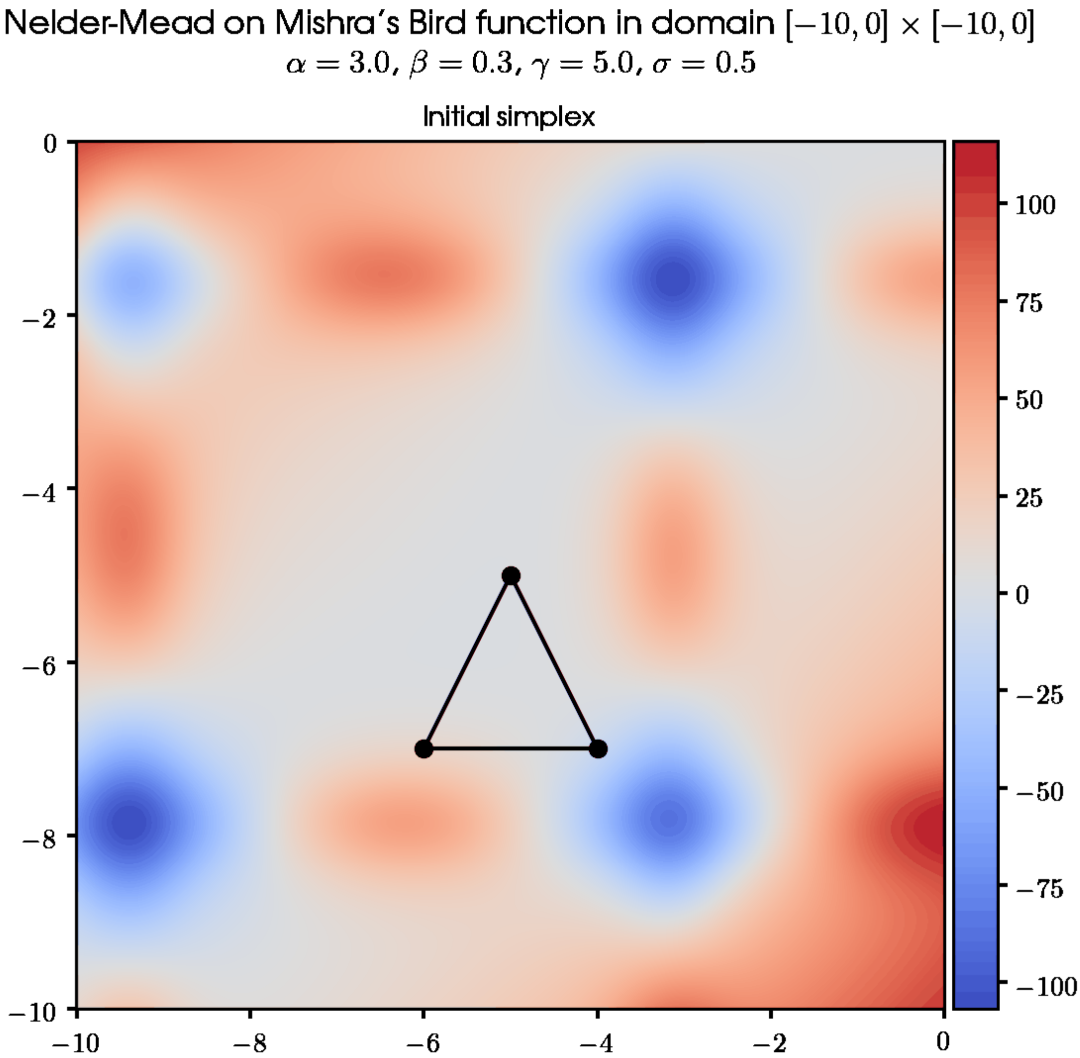
Some random initial simplex and default set of parameters



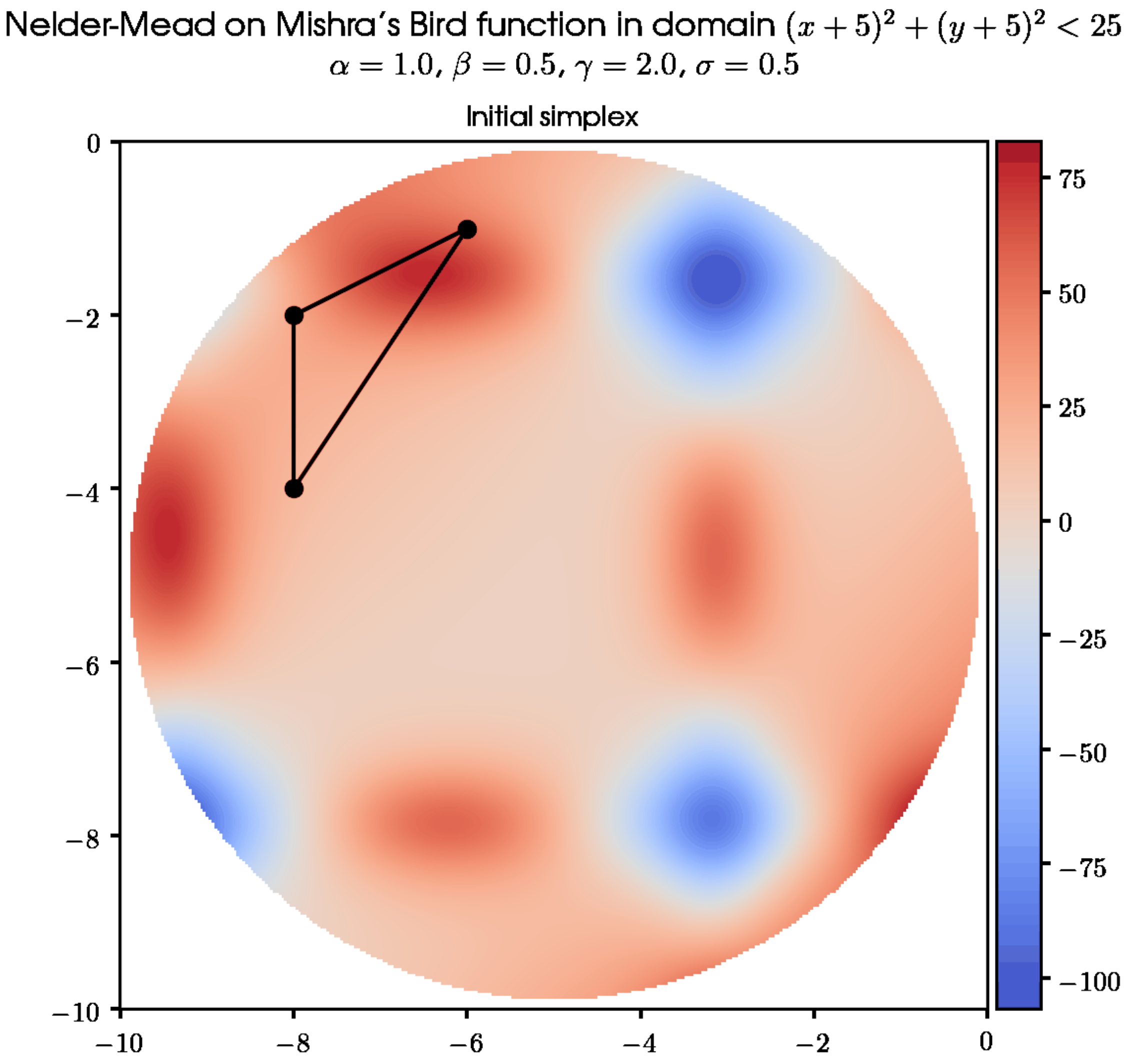
Different initial simplex and same set of parameters



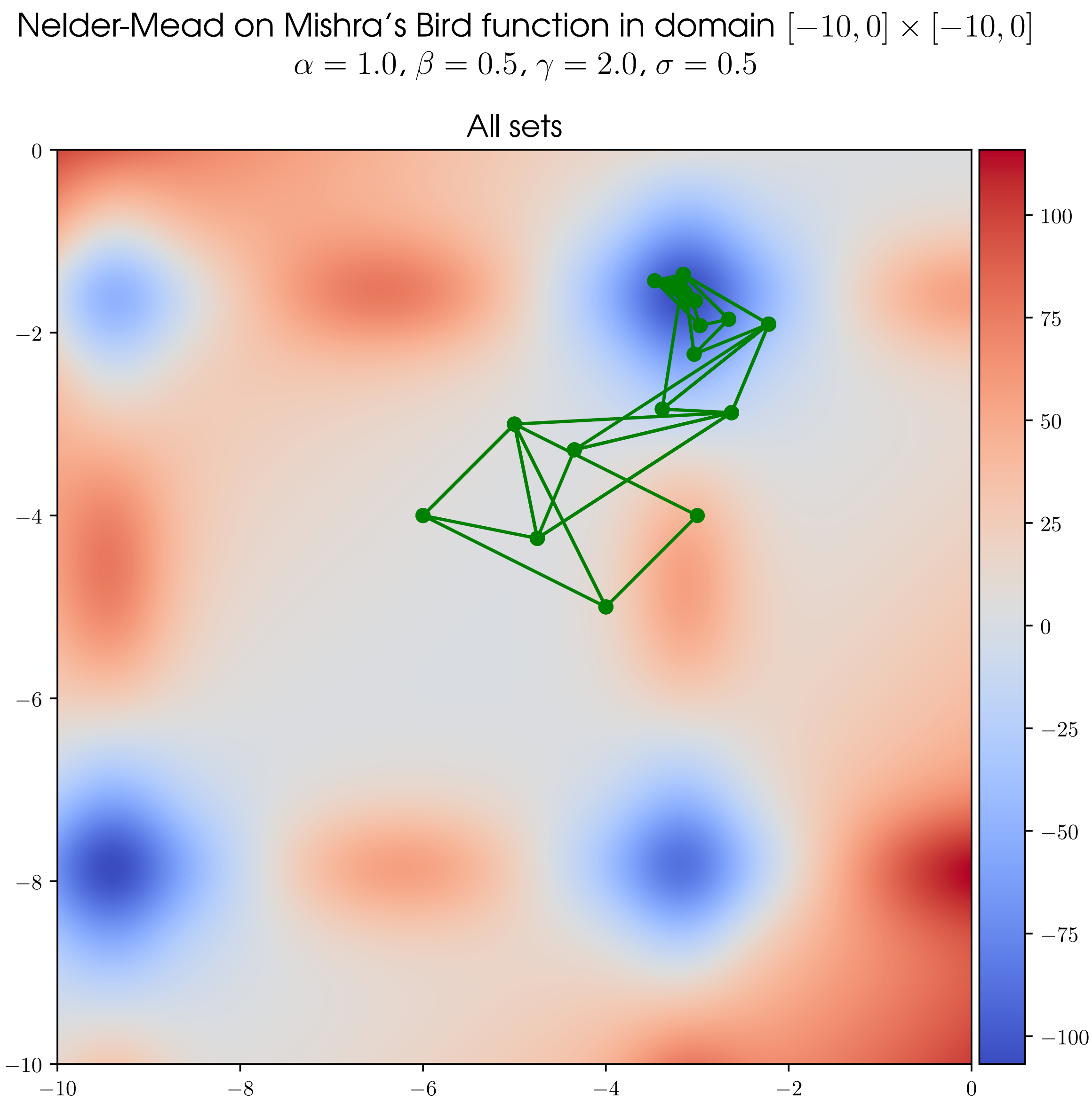
Same initial simplex and different set of parameters

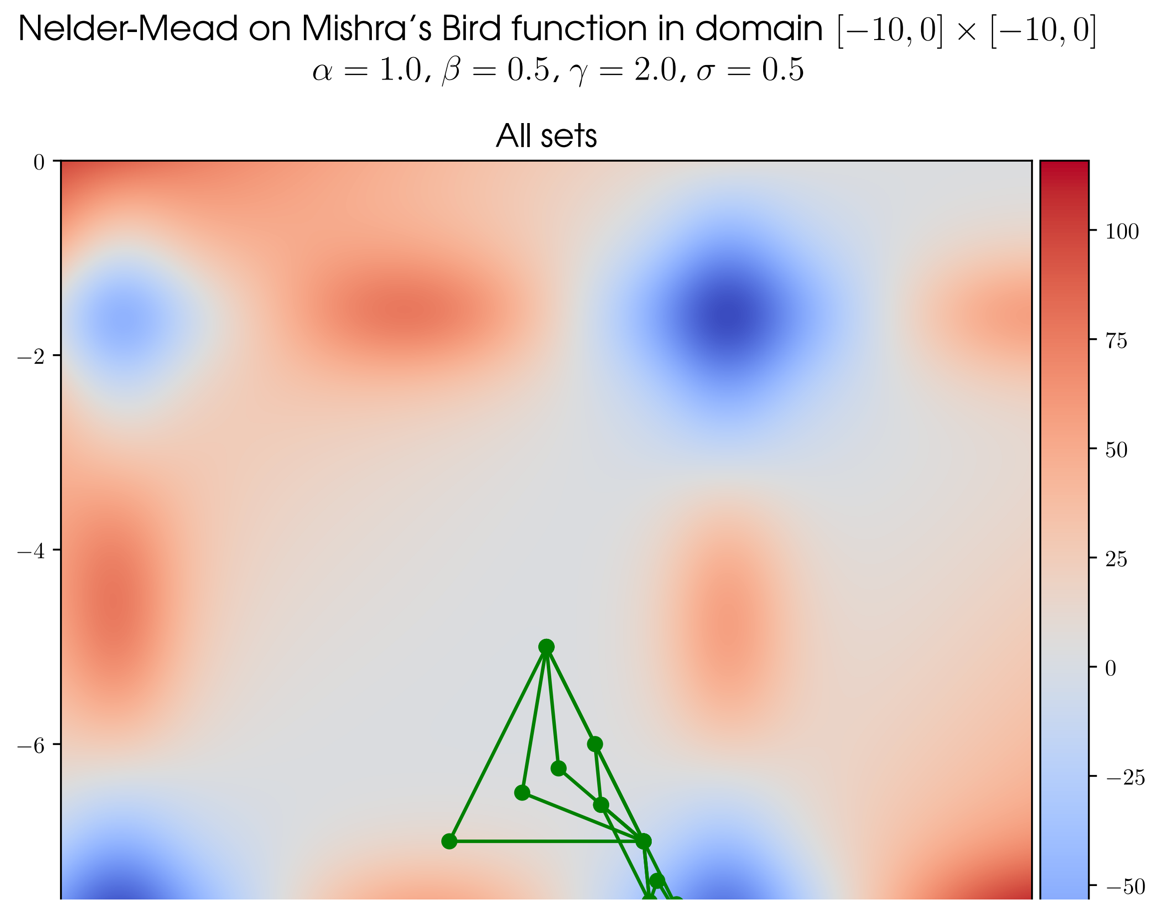


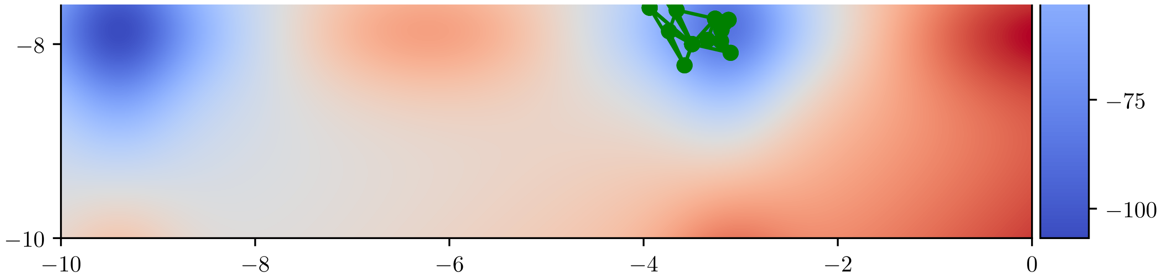
Round domain



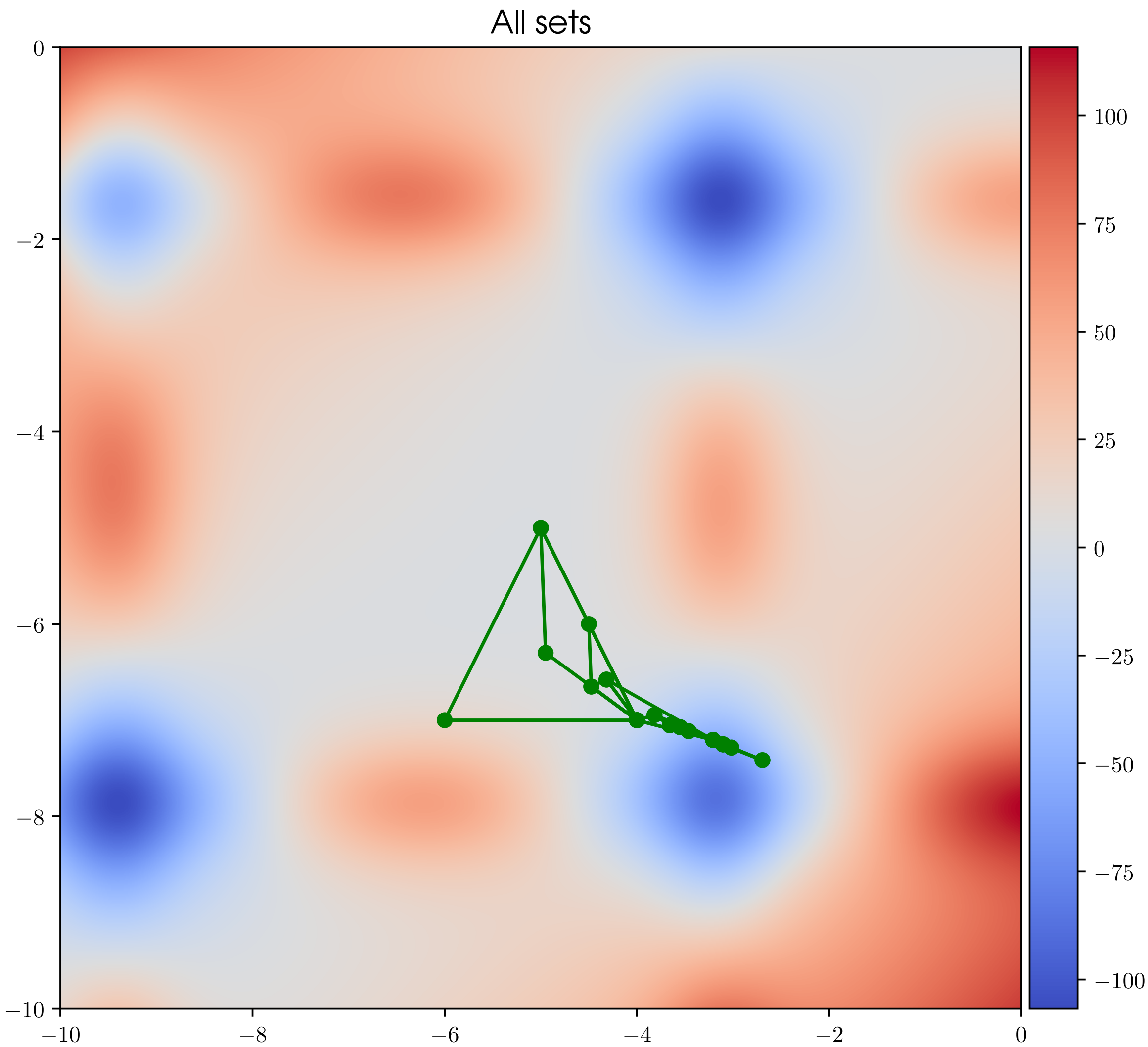
Examples with all sets of simplexes


Nelder-Mead on Mishra's Bird function in domain $[-10,0] \times [-10,0]$
$\alpha = 1.0$, $\beta = 0.5$, $\gamma = 2.0$, $\sigma = 0.5$
All sets

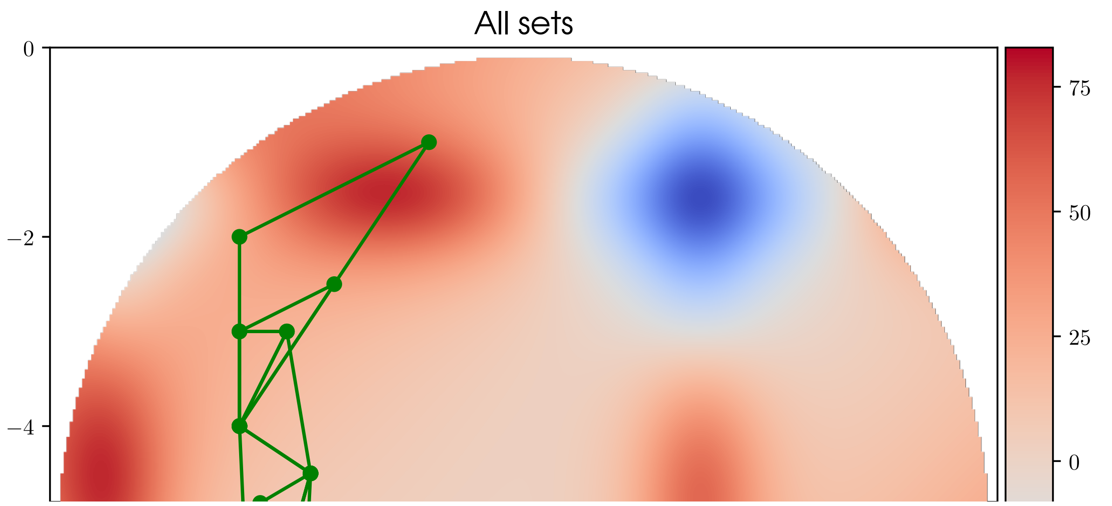

Nelder-Mead on Mishra's Bird function in domain $[-10,0] \times [-10,0]$
$\alpha = 1.0$, $\beta = 0.5$, $\gamma = 2.0$, $\sigma = 0.5$
All sets

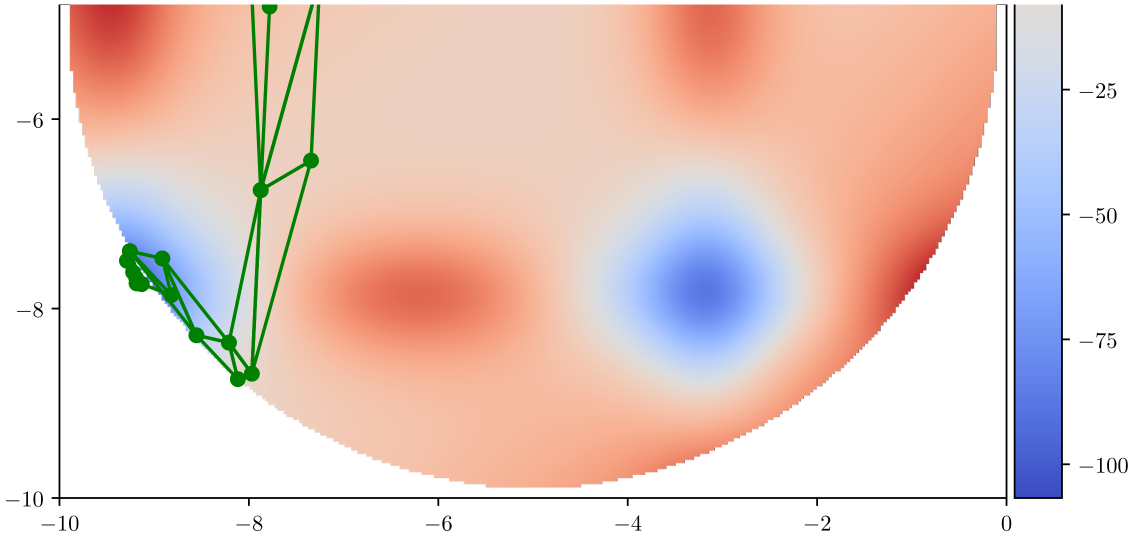


Nelder-Mead on Mishra's Bird function in domain $[-10, 0] \times [-10, 0]$
$\alpha = 3.0$, $\beta = 0.3$, $\gamma = 5.0$, $\sigma = 0.5$



Nelder-Mead on Mishra's Bird function in domain $(x + 5)^2 + (y + 5)^2 < 25$
$\alpha = 1.0$, $\beta = 0.5$, $\gamma = 2.0$, $\sigma = 0.5$





Code

[Open in Colab]()