02610
Optimization and Data Fitting
Week 1: Fundamentals of Unconstrained Optimization

Yiqiu Dong

DTU Compute
Technical University of Denmark

What is a solution?

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

- $\boldsymbol{x} \in \mathbb{R}^n$ is a real vector with $n \geq 1$ components.
- $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function.

We would most like to have a global minimizer, i.e.,

A points $\boldsymbol{x}^*$ is a global minimizer if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

But the global minimizer can be difficult to find due to limited knowledge of f, most algorithms are able to find only a local minimizer, i.e.,

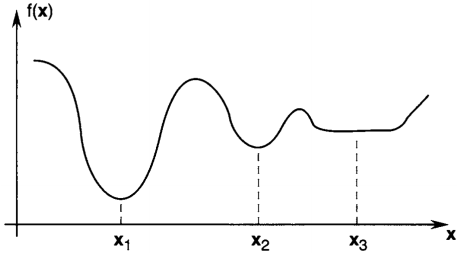
A points $\boldsymbol{x}^*$ is a local minimizer if there is a neighborhood $\mathcal{N}$ of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}$.

What is a solution?

A point $\boldsymbol{x}^*$ is a

- global minimizer, if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.
- local minimizer, if there is a neighborhood $\mathcal{N}$ of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}$.
  - **strict** local minimizer, if there is a neighborhood $\mathcal{N}$ of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}$ and $\boldsymbol{x} \neq \boldsymbol{x}^*$.
  - **isolated** local minimizer, if there is a neighborhood $\mathcal{N}$ of $\boldsymbol{x}^*$ such that $\boldsymbol{x}^*$ is the only local minimizer in $\mathcal{N}$.
  - Some strict local minimizers are not isolated. For example:

$$f(x) = x^4 \cos(1/x) + 2x^4, \qquad f(0) = 0.$$

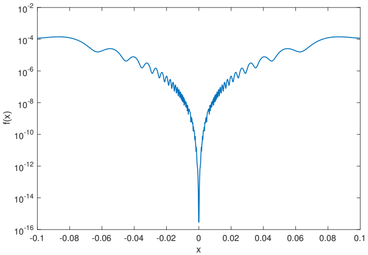


What is a solution?

A point $\boldsymbol{x}^*$ is a

- global minimizer, if $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.
- local minimizer, if there is a neighborhood $\mathcal{N}$ of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) \le f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}$.
  - **strict** local minimizer, if there is a neighborhood $\mathcal{N}$ of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}$ and $\boldsymbol{x} \neq \boldsymbol{x}^*$.
  - **isolated** local minimizer, if there is a neighborhood $\mathcal{N}$ of $\boldsymbol{x}^*$ such that $\boldsymbol{x}^*$ is the only local minimizer in $\mathcal{N}$.
  - Some strict local minimizers are not isolated. For example:

$$f(x) = x^4\cos(1/x) + 2x^4, \qquad f(0) = 0.$$



Univariate example 1

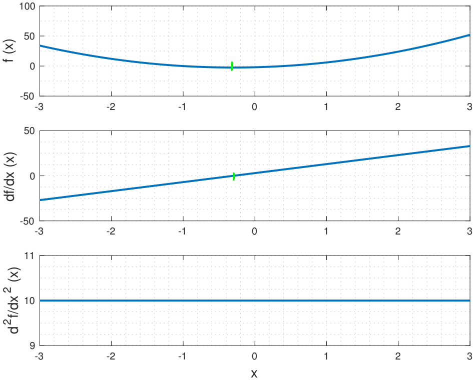
$$\min_{x \in \mathbb{R}} f(x) = 5x^2 + 3x - 2$$

Optimality conditions:

$$\frac{df}{dx}(x^*) = 10x^* + 3 = 0$$
$$\frac{d^2 f}{dx^2}(x^*) = 10 > 0$$

Solution:

$$\text{Minimizer:} \quad x^* = -0.3$$
$$\text{Minimum value:} \quad f(x^*) = -2.45$$

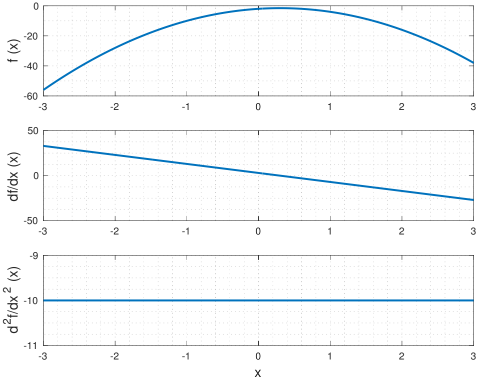


Univariate example 2

$$\min_{x\in\mathbb{R}} f(x) = -5x^2 + 3x - 2$$

Optimality conditions:

$$\frac{df}{dx}(x^*) = -10x^* + 3 = 0$$
$$\frac{d^2 f}{dx^2}(x^*) = -10 < 0$$



Solution:

Minimizer:	*no*
Minimum value:	*no*

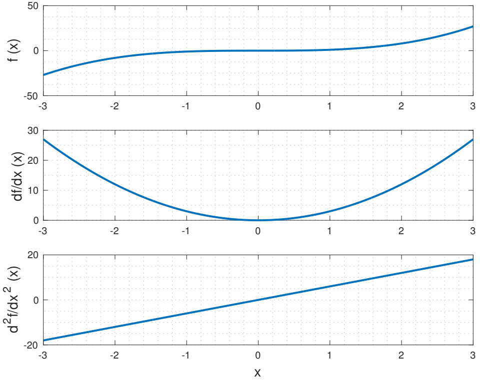
Note that there is a maximizer.

Univariate example 3

$$\min_{x \in \mathbb{R}} f(x) = x^3$$

Optimality conditions:

$$\frac{df}{dx}(x^*) = 3x^{*2} = 0$$
$$\frac{d^2 f}{dx^2}(x^*) = 6x^*$$



Solution:

Minimizer: *no*

Minimum value: *no*

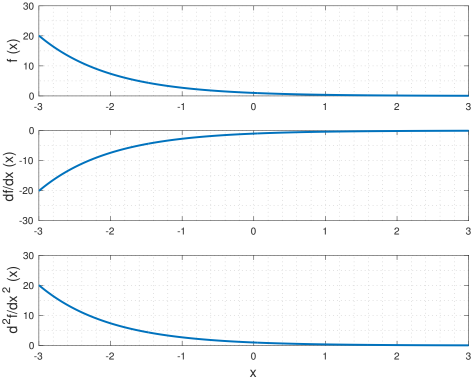
Note that there is a *stationary point*, i.e., the point satisfying $\frac{df}{dx}(x^*) = 0$.

Univariate example 4

$$\min_{x \in \mathbb{R}} f(x) = e^{-x}$$

Optimality conditions:

$$\frac{df}{dx}(x^*) = -e^{-x}$$
$$\frac{d^2 f}{dx^2}(x^*) = e^{-x} > 0$$



Solution:

Minimizer: *no*

Minimum value: *no*

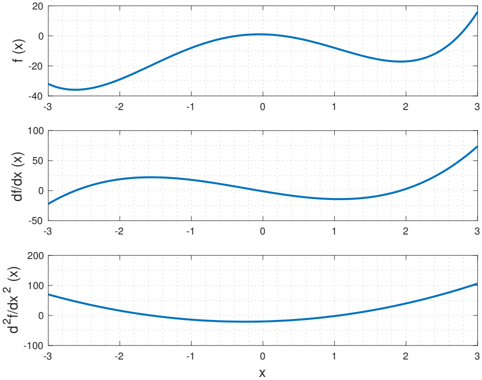
Note that the infimum is 0.

Univariate example 5

$$\min_{x \in \mathbb{R}} f(x) = x^4 + x^3 - 10x^2 - x + 1$$

Optimality conditions:

$$\frac{df}{dx}(x^*) = 4x^{*3} + 3x^{*2} - 20x^* - 1$$
$$\frac{d^2 f}{dx^2}(x^*) = 12x^{*2} + 6x^* - 20$$



Solution:

Minimizer:	*2 local*
Minimum value:	*2 local*

Note that there is also a maximizer.

Optimality condition with univariate problems

$$\min_{x\in\mathbb{R}} f(x) \qquad f:\mathbb{R}\to\mathbb{R}, f\in\mathcal{C}^2(\mathbb{R}) \tag{1}$$

Theorem (Necessary Optimality Conditions)

Let $x^* \in \mathbb{R}$ be a local minimizer of (1). Then

1. $\frac{df}{dx}(x^*) = 0$
2. $\frac{d^2f}{dx^2}(x^*) \geq 0$

Theorem (Sufficient Optimality Conditions)

Let $x^* \in \mathbb{R}$ satisfy

1. $\frac{df}{dx}(x^*) = 0$
2. $\frac{d^2f}{dx^2}(x^*) > 0$

Then, x^* is a local minimizer of (1).

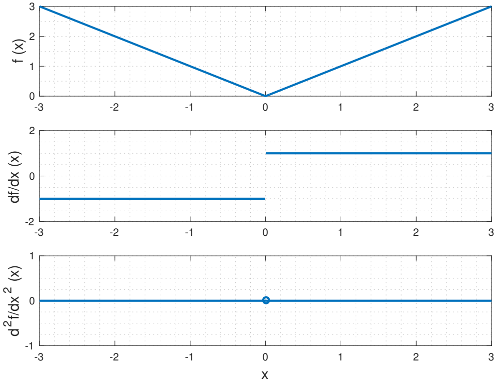
Non-smooth function

$$\min_{x \in \mathbb{R}} f(x) = |x|$$

Optimality conditions:

$$\frac{df}{dx}(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \end{cases}$$

$$\frac{d^2 f}{dx^2}(x) = 0$$



Solution:

$$\text{Minimizer:} \quad x^* = 0$$

$$\text{Minimum value:} \quad f(x^*) = 0$$

It is impossible in general to identify a minimizer of a general non-smooth function. But in some special cases, we can. $\rightarrow$ 02611 “Optimization for Data

Multivariate problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad f:\mathbb{R}^n\to\mathbb{R},\ f\in\mathcal{C}^2(\mathbb{R}^n)$$

- **Gradient:**

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \frac{\partial f}{\partial x_2}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^n$$

- **Hessian:**

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} \nabla\frac{\partial f}{\partial x_1}(\boldsymbol{x}) & \nabla\frac{\partial f}{\partial x_2}(\boldsymbol{x}) & \cdots & \nabla\frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_1\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_n}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_2\partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_2\partial x_n}(\boldsymbol{x}) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_n\partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_n\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^{n\times n}$$

Necessary optimality conditions with multivariate problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad f:\mathbb{R}^n \to \mathbb{R},\ f \in \mathcal{C}^2(\mathbb{R}^n) \tag{2}$$

Theorem (Necessary Optimality Conditions)

Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be a local minimizer of (2). Then

1. $\nabla f(\boldsymbol{x}^*) = 0$
2. $\nabla^2 f(\boldsymbol{x}^*) \succeq 0$ (positive semi-definite)

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is *positive semi-definite*, if it satisfies $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \neq 0$.
- A is positive semi-definite if and only if all its eigenvalues are non-negative.

Sufficient optimality conditions with multivariate problems

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \qquad f : \mathbb{R}^n \to \mathbb{R},\ f \in \mathcal{C}^2(\mathbb{R}^n) \tag{2}$$

Theorem (Sufficient Optimality Conditions)

Let $\boldsymbol{x}^* \in \mathbb{R}^n$ satisfy

1. $\nabla f(\boldsymbol{x}^*) = 0$
2. $\nabla^2 f(\boldsymbol{x}^*) \succ 0$ (positive definite)

Then, $\boldsymbol{x}^*$ is a local minimizer of (2).

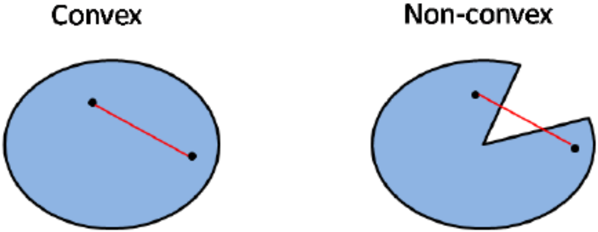
- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is *positive definite*, if it satisfies $\boldsymbol{x}^T A \boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq 0$.
- A is positive definite if and only if all its eigenvalues are positive.

Convexity

- **Convex set:** A set $S \in \mathbb{R}^n$ is a *convex set*, if

$$\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 \in S$$

for any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ and all $\alpha \in [0,1]$.



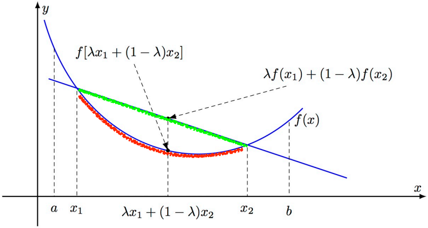
- **Convex functions:** A function f is a *convex function*, if its domain S is a convex set and it satisfies

$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \leq \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2)$$

for any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ and all $\lambda \in [0,1]$.

Convexity

- **Convex set:** A set $S \in \mathbb{R}^n$ is a *convex set*, if
$$\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 \in S$$
for any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ and all $\alpha \in [0,1]$.
- **Convex functions:** A function f is a *convex function*, if its domain S is a convex set and it satisfies
$$f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) \le \lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2)$$
for any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ and all $\lambda \in [0,1]$.



Convex functions

- **Definition 1:** A function f is a *convex function*, if its domain S is a convex set and it satisfies

$$f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \le \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2)$$

  for any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S$ and all $\alpha \in [0,1]$.
- **Definition 2:** Let f be twice differentiable. Then f is a *convex function* if and only if its domain S is a convex set and its Hessian is positive semidefinite, i.e. for all $\boldsymbol{x} \in S$

$$\nabla^2 f(\boldsymbol{x}) \succeq 0.$$

- f is **strict convex**, if $f(\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) < \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2)$ whenever $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$.
- f is *concave*, if $-f$ is convex.

Examples of univariate convex functions

- Affine: $f(x) = ax + b$ with $x \in \mathbb{R}$ for all $a, b \in \mathbb{R}$
- Exponential: $f(x) = e^{ax}$ with $x \in \mathbb{R}$ for all $a \in \mathbb{R}$
- Powers: $f(x) = x^a$ with $x \in \mathbb{R}_{++}$ when $a \geq 1$ or $a \leq 0$
- Negative roots: $f(x) = -x^a$ for $x \in \mathbb{R}_{++}$ when $0 \leq a \leq 1$
- Powers of absolute value: $f(x) = |x|^p$ with $x \in \mathbb{R}$ for $p \geq 1$
- Negative logarithm: $f(x) = -\log(x)$ for $x \in \mathbb{R}_{++}$
- Negative entropy: $f(x) = x \log(x)$ for $x \in \mathbb{R}_{++}$

Examples of multivariate convex functions

- Norms: Every norm on $\mathbb{R}^n$ is convex. For example,

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,\dots,n} |x_i|$$

- Max function: $f(\boldsymbol{x}) = \max\{x_1, \dots, x_n\}$
- Log-sum-exp: $f(\boldsymbol{x}) = \log(e^{x_1} + \cdots + e^{x_n})$
- Negative geometric mean: $f(\boldsymbol{x}) = -(\prod_{i=1}^{n} x_i)^{1/n}$ with $\boldsymbol{x} \in \mathbb{R}^n_{++}$

Operations that preserve convexity

- Non-negative weighted sums: If f_i with $i = 1, \ldots, m$ are convex and $w_i \geq 0$ for all $i = 1, \ldots, m$, then $f(\boldsymbol{x}) = w_1 f_1(\boldsymbol{x}) + \cdots + w_m f_m(\boldsymbol{x})$ is convex.
- Non-negative weighted integrals: If $f(\boldsymbol{x}, \boldsymbol{y})$ is convex in $\boldsymbol{x}$ for each $\boldsymbol{y} \in S$ and $w(\boldsymbol{y}) \geq 0$ for each $\boldsymbol{y} \in S$, then $g(\boldsymbol{x}) = \int_S w(\boldsymbol{y}) f(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y}$ is convex in $\boldsymbol{x}$.
- Composition with an affine mapping: Suppose $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n \times m}$, and $\boldsymbol{b} \in \mathbb{R}^n$. If f is convex, then $g(\boldsymbol{x}) = f(A\boldsymbol{x} + \boldsymbol{b})$ is convex.
- Pointwise maximum: If f_i with $i = 1, \ldots, m$ are convex, then $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\}$ is convex.

Convex optimization problems

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & c_i(\boldsymbol{x}) = 0 \text{ for } i \in \mathcal{E} \\ & c_i(\boldsymbol{x}) \geq 0 \text{ for } i \in \mathcal{I} \end{aligned}$$

- The objective function f must be convex
- The equality constraint functions c_i with $i \in \mathcal{E}$ must be linear
- The inequality constraint functions c_i with $i \in \mathcal{I}$ must be concave

Convex optimization problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x})$$

- The objective function f must be convex
- The equality constraint functions c_i with $i \in \mathcal{E}$ must be linear
- The inequality constraint functions c_i with $i \in \mathcal{I}$ must be concave

Optimality conditions with convex problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad f:\mathbb{R}^n \to \mathbb{R},\ f \in \mathcal{C}(\mathbb{R}^n) \text{ and } f \text{ is convex} \tag{3}$$

Theorem

- Any local minimizer $\boldsymbol{x}^* \in \mathbb{R}^n$ of (3) is a global minimizer.
- If in addition f is differentiable, $\boldsymbol{x}^*$ is a global minimizer of (3) if and only if

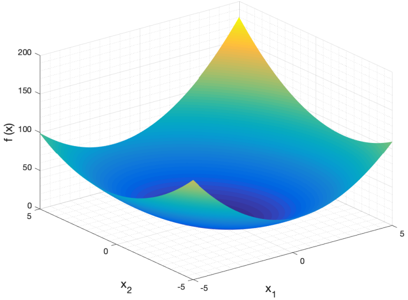
$$\nabla f(\boldsymbol{x}^*) = 0.$$

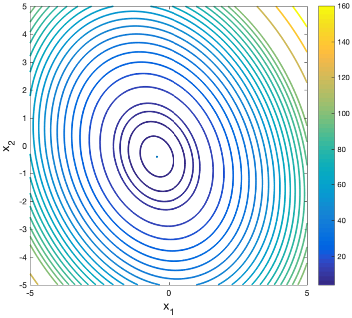
Role of optimality conditions

- **Recognize the solution.** Given a candidate solution, check optimality conditions to verify it is a solution.
- **Measure the quality** of an approximate solution: Measure how "close" a point is to being a solution.
- **Develop algorithms.** Reduce an optimization problem to solving a system of (nonlinear) equations (finding a root of the gradient).

Multivariate example 1

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = 3x_1^2 + 2x_2^2 + x_1x_2 + 3x_1 + 2x_2 + 4$$





Contour plots using Python

```
x1 = numpy.arange(-5, 5, 0.005)
x2 = x1
X, Y = np.meshgrid(x1, x2)
F = 3*X**2+2*Y**2+X*Y+3*X+2*Y+4

fig, ax=matplotlib.pyplot.subplots(subplot_kw={"projection": "3d"})
surf = ax.plot_surface(X, Y, F, cmap=matplotlib.cm.coolwarm,
                       linewidth=0, antialiased=False)
plt.show()

fig, ax = plt.subplots()
v = np.concatenate((np.arange(0, 10, 2), np.arange(10, 100, 5),
                    np.arange(100, 200, 20)), axis=None)
c = ax.contour(X, Y, F, v, linewidths=2)
norm = matplotlib.colors.Normalize(vmin=c.cvalues.min(),
                  vmax=c.cvalues.max())
sm = matplotlib.cm.ScalarMappable(norm=norm, cmap=c.cmap)
fig.colorbar(sm, ax=ax)
plt.show()
```

Multivariate example 1

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = 3x_1^2 + 2x_2^2 + x_1x_2 + 3x_1 + 2x_2 + 4$$

- **Gradient:**

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 6x_1 + x_2 + 3 \\ 4x_2 + x_1 + 2 \end{bmatrix}$$

- **Hessian:**

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) \end{bmatrix} = \begin{bmatrix} 6 & 1 \\ 1 & 4 \end{bmatrix}$$

- **Optimality conditions:**

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} 6x_1 + x_2 + 3 \\ 4x_2 + x_1 + 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 6 & 1 \\ 1 & 4 \end{bmatrix} \succ 0$$

Solve linear systems in Python

$$A\boldsymbol{x} = \mathrm{b}$$

- **LU factorization:** A is a general matrix (unsymmetric, indefinite).

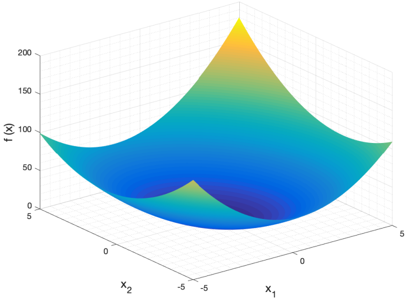
```
>>> import numpy as np
>>> from scipy.linalg import lu

>>> P, L, U = lu(A)

>>> tmp = scipy.linalg.solve_triangular(L,P@b,lower=True)
>>> x = scipy.linalg.solve_triangular(U,tmp,lower=False)
```

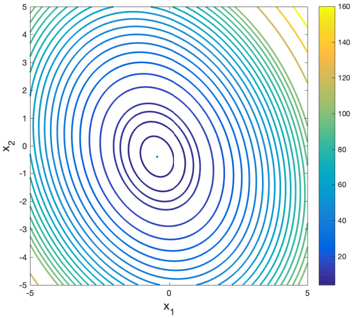
- **Cholesky factorization:** A is a symmetric positive definite matrix.

```
>>> c, low = scipy.linalg.cho_factor(A)
>>> x = scipy.linalg.cho_solve((c, low), b)
```

Multivariate example 1

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = 3x_1^2 + 2x_2^2 + x_1x_2 + 3x_1 + 2x_2 + 4$$

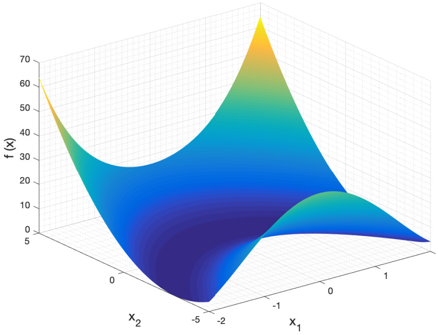


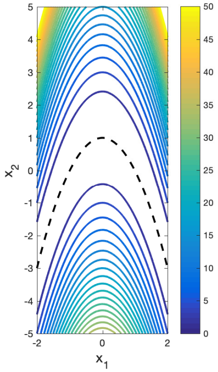


$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -0.4348 \\ -0.3913 \end{bmatrix}$$

Multivariate example 2

$$\min_{\boldsymbol{x} \in \mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 1)^2$$





Multivariate example 2

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 1)^2$$

- Obviously, the minimizers are located on the curve $\mathcal{S} = \{(x_1, x_2) : x_1^2 + x_2 - 1 = 0\}$.
- **Gradient:**

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 4x_1(x_1^2 + x_2 - 1) \\ 2(x_1^2 + x_2 - 1) \end{bmatrix}$$

- **Hessian:**

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) \end{bmatrix}$$

$$= \begin{bmatrix} 12x_1^2 + 4x_2 - 4 & 4x_1 \\ 4x_1 & 2 \end{bmatrix}$$

Multivariate example 2

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 1)^2$$

The points on the curve $\mathcal{S} = \{(x_1, x_2) : x_1^2 + x_2 - 1 = 0\}$ satsify

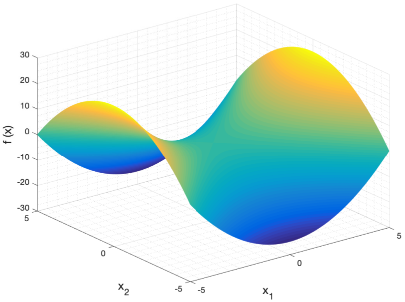
$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} 4x_1(x_1^2 + x_2 - 1) \\ 2(x_1^2 + x_2 - 1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

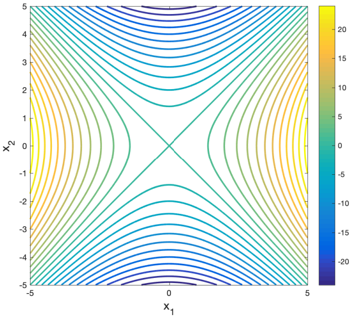
$$\nabla^2 f(x) = \begin{bmatrix} 8x_1^2 & 4x_1 \\ 4x_1 & 2 \end{bmatrix} \succeq 0$$

At the minimizer $\boldsymbol{x}^* \in \mathcal{S}$, the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite with eigenvalues 0 and 2.

Multivariate example 3

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = x_1^2 - x_2^2$$





Multivariate example 2

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = x_1^2 - x_2^2$$

- **Gradient:**

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 2x_1 \\ -2x_2 \end{bmatrix}$$

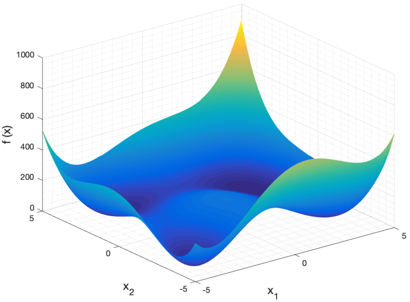
- **Hessian:**

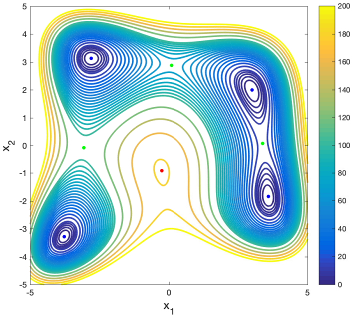
$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

At the point $\boldsymbol{x}^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\nabla f(\boldsymbol{x}^*) = 0$ is satisfied, but $\boldsymbol{x}^*$ is neither a local minimizer nor a local maximizer. In this case, we call $\boldsymbol{x}^*$ a *saddle point*. Note that the Hessian matrix at $\boldsymbol{x}^*$ is indefinite.

Multivariate example 4

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$





Multivariate example 4

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

- **Gradient:**

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 4x_1(x_1^2 + x_2 - 11) + 2(x_1 + x_2^2 - 7) \\ 2(x_1^2 + x_2 - 11) + 4x_2(x_1 + x_2^2 - 7) \end{bmatrix}$$

- **Hessian:**

$$\begin{aligned} \nabla^2 f(x) &= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) \end{bmatrix} \\ &= \begin{bmatrix} 12x_1^2 + 4x_2 - 42 & 4x_1 + 4x_2 \\ 4x_1 + 4x_2 & 4x_1 + 12x_2^2 - 26 \end{bmatrix} \end{aligned}$$

Multivariate example 4

$$\min_{\boldsymbol{x} \in \mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

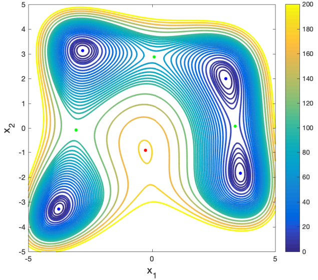
The stationary points can be found by using a nonlinear system solver such as `scipy.optimize.fsolve` in Python to solve

$$\nabla f(\boldsymbol{x}) = 0.$$

In the following slides, we list the stationary points together with its Hessian, $\nabla^2 f(\boldsymbol{x})$, and the corresponding eigenvalues, λ, of the Hessian matrix.

Multivariate example 4

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$



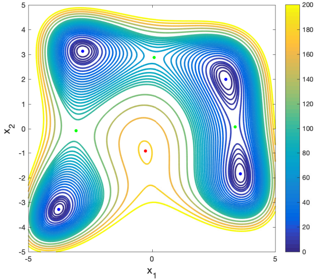
Local minimizers (blue)

$\boldsymbol{x}$	$\nabla^2 f(\boldsymbol{x})$	λ
$\begin{bmatrix} -2.8051 \\ 3.1313 \end{bmatrix}$	$\begin{bmatrix} 64.9495 & 1.3048 \\ 1.3048 & 80.4409 \end{bmatrix}$	$\begin{bmatrix} 64.8404 \\ 80.5501 \end{bmatrix}$
$\begin{bmatrix} 3 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 74 & 20 \\ 20 & 34 \end{bmatrix}$	$\begin{bmatrix} 25.7157 \\ 82.2843 \end{bmatrix}$
$\begin{bmatrix} -3.7793 \\ -3.2832 \end{bmatrix}$	$\begin{bmatrix} 116.2655 & -28.2500 \\ -28.2500 & 88.2345 \end{bmatrix}$	$\begin{bmatrix} 70.7144 \\ 133.7856 \end{bmatrix}$
$\begin{bmatrix} 3.5844 \\ -1.8481 \end{bmatrix}$	$\begin{bmatrix} 104.7850 & 6.9452 \\ 6.9452 & 29.3246 \end{bmatrix}$	$\begin{bmatrix} 28.6907 \\ 105.4189 \end{bmatrix}$

The Hessian matrices are positive definite, i.e., all eigenvalues are positive.

Multivariate example 4

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$



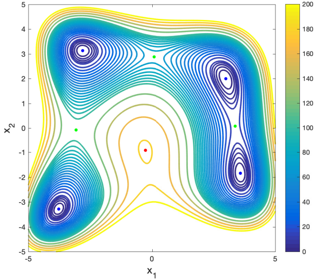
Saddle points (green)

$\boldsymbol{x}$	$\nabla^2 f(\boldsymbol{x})$	λ
$\begin{bmatrix} -3.073 \\ -0.0814 \end{bmatrix}$	$\begin{bmatrix} -49.6184 & -8.3267 \\ -8.3267 & 19.2922 \end{bmatrix}$	$\begin{bmatrix} -50.6102 \\ 20.2840 \end{bmatrix}$
$\begin{bmatrix} 0.0867 \\ 2.8843 \end{bmatrix}$	$\begin{bmatrix} -30.3728 & 11.8837 \\ 11.8837 & 74.1738 \end{bmatrix}$	$\begin{bmatrix} -31.7066 \\ 75.5076 \end{bmatrix}$
$\begin{bmatrix} 3.3852 \\ 0.0739 \end{bmatrix}$	$\begin{bmatrix} 95.8066 & 13.8360 \\ 13.8360 & -12.3939 \end{bmatrix}$	$\begin{bmatrix} -14.1352 \\ 97.5479 \end{bmatrix}$

The Hessian matrices are indefinite, i.e., at least one positive and one negative eigenvalue.

Multivariate example 4

$$\min_{\boldsymbol{x}\in\mathbb{R}^2} f(\boldsymbol{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$



Local maximizer (red)

$\boldsymbol{x}$	$\nabla^2 f(\boldsymbol{x})$	λ
$\begin{bmatrix} -0.2708 \\ -0.9230 \end{bmatrix}$	$\begin{bmatrix} -44.8119 & -4.7755 \\ -4.7755 & -16.8594 \end{bmatrix}$	$\begin{bmatrix} -45.6052 \\ -16.0660 \end{bmatrix}$

The Hessian matrix is negative definite, i.e., all eigenvalues are negative.

Numerical optimization algorithms

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad f:\mathbb{R}^n\to\mathbb{R},\ f\in\mathcal{C}^2(\mathbb{R}^n)$$

Most optimization problems can NOT be solved in a closed form (a single step).

For them, we develop iterative algorithms:

1. Start from an initial candidate solution: $\boldsymbol{x}_0$
2. Generate a sequence of candidate solutions (iterates): $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\cdots$
3. Stop when a certain condition is met; return the candidate solution

In a large number of algorithms, $\boldsymbol{x}_{k+1}$ is generated from $\boldsymbol{x}_k$, that is, using the information of f at $\boldsymbol{x}_k$.

In some algorithms, $\boldsymbol{x}_{k+1}$ is generated from $\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \ldots$ But, for time and memory consideration, most previous iterates are not kept in memory.

Numerical optimization algorithms

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \qquad f:\mathbb{R}^n \to \mathbb{R},\ f\in\mathcal{C}^2(\mathbb{R}^n)$$

The iteration step in most algorithms that we will introduce in this course is essentially in the form of

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k.$$

- $\boldsymbol{x}_k$ is the iterate. We expect as $k \to +\infty$ we have $\boldsymbol{x}_k \to \boldsymbol{x}^*$.
- $\boldsymbol{p}_k$ is the search direction.
- α_k is the step size or step length.
- In addition, we need the stop criteria to define when the algorithm stops.

Two strategies: line search and trust region

- **Line search:** The algorithm chooses a direction $\boldsymbol{p}_k$ and searches along this direction from the current iterate $\boldsymbol{x}_k$ for a new iterate with a lower function value. For example, we can find the next step size by solving the 1D problem:

$$\min_{\alpha>0} f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k).$$

- **Trust region:** The algorithm constructs an approximation of f, m_k, in neighborhood of $\boldsymbol{x}_k$, then searches for a minimizer of m_k in this region. For example, we can find the next moving step by solving

$$\min_{\boldsymbol{p}} m_k(\boldsymbol{x}_k + \boldsymbol{p}), \quad \text{s. t. } \|\boldsymbol{p}\| \leq \Delta.$$

Stopping criteria

The first-order necessary condition $\|\nabla f(\boldsymbol{x}_k)\| = 0$ is not practical.

Some practical stopping criteria:

- Gradient condition: $\|\nabla f(\boldsymbol{x}_k)\| < \epsilon$.
- Successive objective condition: $|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)| < \epsilon$ or the relative one
$$\frac{|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)|}{|f(\boldsymbol{x}_k)|} < \epsilon.$$
- Successive point difference: $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < \epsilon$ or the relative one
$$\frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\|\boldsymbol{x}_k\|} < \epsilon.$$
- Maximum iteration number: $k > k_{\text{max}}$.