Introduction to Machine Learning
Lecture 7 - Introduction to Optimization

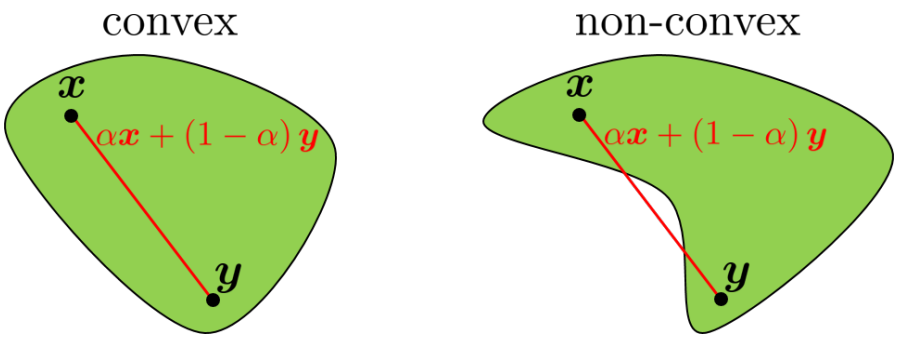
1 Convexity

1.1 Convex set

A set $\mathcal{C} \subseteq \mathbb{R}^d$ is said to be **convex** if, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and for all $\alpha \in [0,1]$:

$$\boxed{\alpha \boldsymbol{x} + (1-\alpha)\,\boldsymbol{y} \in \mathcal{C}}$$

For example:



Exercise 1 Let $\mathcal{C}$ be a convex set and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3 \in \mathcal{C}$.
Prove that any **convex combination** is also in $\mathcal{C}$. namely:

$$\sum_{i=1}^{3} \alpha_i \boldsymbol{x}_i \in \mathcal{C}$$

where $\alpha_i \geq 0$ and $\sum_{i=1}^{3} \alpha_i = 1$.
Solution:

$$\begin{aligned}\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \alpha_3\boldsymbol{x}_3 &= \underbrace{(\alpha_1+\alpha_2)}_{1-\alpha_3}\underbrace{\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\boldsymbol{x}_1 + \frac{\alpha_2}{\alpha_1+\alpha_2}\boldsymbol{x}_2\right)}_{\triangleq \boldsymbol{y} \in \mathcal{C}} + \alpha_3\boldsymbol{x}_3 \\ &= (1-\alpha_3)\,\boldsymbol{y} + \alpha_3\boldsymbol{x}_3 \in \mathcal{C}\end{aligned}$$

■

Exercise 2 Let $\mathcal{C}$ and $\mathcal{D}$ be convex sets.
Prove that $\mathcal{C} \cap \mathcal{D}$ is also convex set.

Solution:
Consider $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \cap \mathcal{D}$.
Since both $\mathcal{C}$ and $\mathcal{D}$ are convex, for any $\alpha \in [0,1]$ we have:

$$\alpha \boldsymbol{x} + (1-\alpha)\,\boldsymbol{y} \in \mathcal{C}, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$$

And also:

$$\alpha \boldsymbol{x} + (1-\alpha)\,\boldsymbol{y} \in \mathcal{D}, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$$

Thus:

$$\alpha \boldsymbol{x} + (1-\alpha)\,\boldsymbol{y} \in \mathcal{C} \cap \mathcal{D}, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \cap \mathcal{D}$$

■

1.2 Convex function

Let $\mathcal{C}$ be a convex set and let $f : \mathcal{C} \longrightarrow \mathbb{R}$ be a function.

- f is called **convex** if, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{C}$ and for all $\alpha \in [0,1]$:

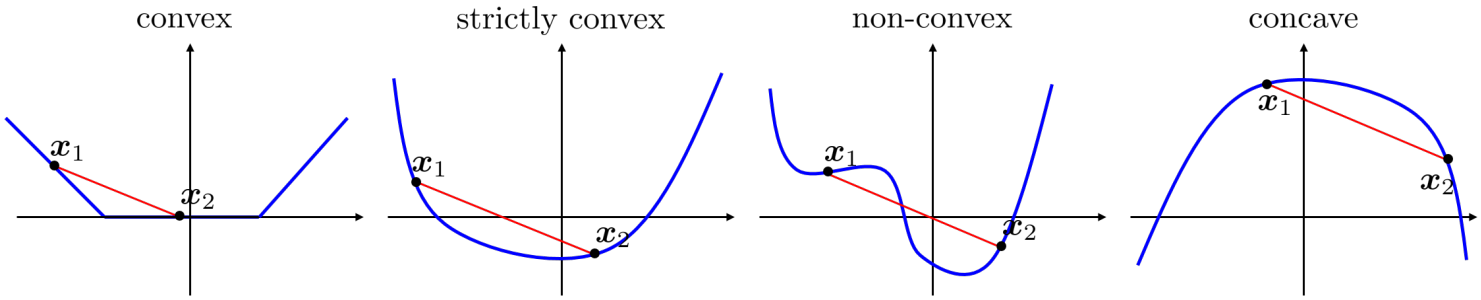
$$\boxed{f\left(\alpha \boldsymbol{x}_1 + (1-\alpha)\,\boldsymbol{x}_2\right) \leq \alpha f\left(\boldsymbol{x}_1\right) + (1-\alpha)\, f\left(\boldsymbol{x}_2\right)}$$

- f is called **strictly convex** if, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{C}$ such that $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ and for all $\alpha \in (0,1)$:

$$\boxed{f\left(\alpha \boldsymbol{x}_1 + (1-\alpha)\,\boldsymbol{x}_2\right) < \alpha f\left(\boldsymbol{x}_1\right) + (1-\alpha)\, f\left(\boldsymbol{x}_2\right)}$$

- f is called (strictly) concave if $-f$ is (strictly) convex.

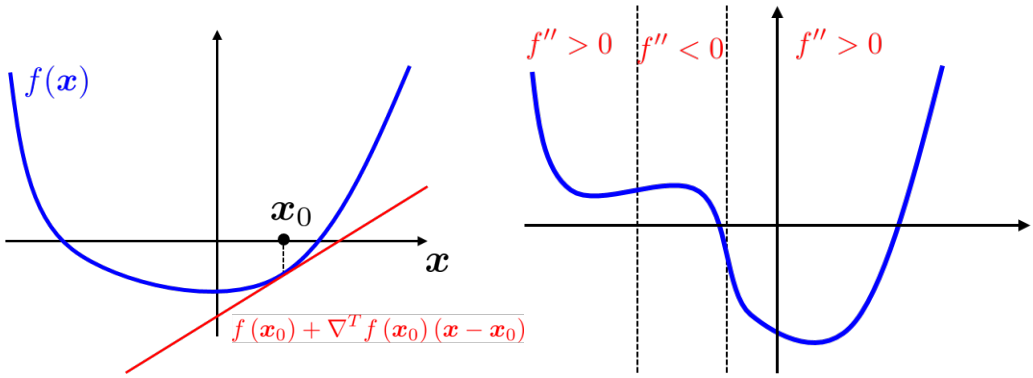
For example:



Theorem 1. *Suppose $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is twice differentiable.*
Then the following are equivalent:

1. f is convex.
2. $f\left(\boldsymbol{y}\right) \geq f\left(\boldsymbol{x}\right) + \nabla^T f\left(\boldsymbol{x}\right)\left(\boldsymbol{y} - \boldsymbol{x}\right)$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$.
3. $\nabla^2 f(\boldsymbol{x}) \succeq 0$ for all $\boldsymbol{x}$.

Condition 2 states that the Taylor expansion at any point is a global under estimator of f.
Condition 3 states that the function f has non-negative curvature everywhere.



f is strictly convex if, for all $\boldsymbol{x}$: $\nabla^2 f\left(\boldsymbol{x}\right) \succ 0$ (it is enough but it is not a necessary condition).

We prove the theorem for dimension one: $n = 1$.

1 ⇒ 2 Since f is convex, by definition for $\alpha \in (0,1)$:

$$\begin{aligned} f(\alpha \boldsymbol{y} + (1-\alpha)\boldsymbol{x}) &\leq \alpha f(\boldsymbol{y}) + (1-\alpha) f(\boldsymbol{x}) \\ f(\boldsymbol{x} + \alpha(\boldsymbol{y}-\boldsymbol{x})) &\leq f(\boldsymbol{x}) + \alpha(f(\boldsymbol{y}) - f(\boldsymbol{x})) \\ \frac{f(\boldsymbol{x} + \alpha(\boldsymbol{y}-\boldsymbol{x})) - f(\boldsymbol{x})}{\alpha} &\leq f(\boldsymbol{y}) - f(\boldsymbol{x}) \end{aligned}$$

For $\alpha \to 0^+$ we have:

$$f'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) \leq f(\boldsymbol{y}) - f(\boldsymbol{x})$$

$$\Rightarrow \boxed{f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + f'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})}$$

2 ⇒ 1 We have that:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + f'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}) \qquad \forall \boldsymbol{x}, \boldsymbol{y}$$

Consider:

$$\boldsymbol{z} \triangleq \alpha \boldsymbol{y} + (1-\alpha)\boldsymbol{x}$$

$$\Rightarrow \begin{cases} f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + f'(\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{z}) & (1) \\ f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + f'(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{z}) & (2) \end{cases}$$

$$\begin{aligned} \alpha(1) + (1-\alpha)(2) \Longrightarrow \alpha f(\boldsymbol{y}) + (1-\alpha) f(\boldsymbol{x}) &\geq f(\boldsymbol{z}) + f'(\boldsymbol{z})(\alpha(\boldsymbol{y}-\boldsymbol{z}) + (1-\alpha)(\boldsymbol{x}-\boldsymbol{z})) \\ \alpha f(\boldsymbol{y}) + (1-\alpha) f(\boldsymbol{x}) &\geq f(\boldsymbol{z}) + f'(\boldsymbol{z}) \left(\underbrace{\alpha \boldsymbol{y} + (1-\alpha)\boldsymbol{x}}_{=z} - \boldsymbol{z} \right) \\ \alpha f(\boldsymbol{y}) + (1-\alpha) f(\boldsymbol{x}) &\geq f(\boldsymbol{z}) \end{aligned}$$

$$\Rightarrow \boxed{\alpha f(\boldsymbol{y}) + (1-\alpha) f(\boldsymbol{x}) \geq f(\alpha \boldsymbol{y} + (1-\alpha)\boldsymbol{x})}$$

2 ⇒ 3 Consider $y > x$. From 2 we have:

$$\begin{cases} f(y) \geq f(x) + f'(x)(y-x) \\ f(x) \geq f(y) + f'(y)(x-y) \end{cases} \Rightarrow \begin{cases} f(y) - f(x) \geq f'(x)(y-x) \\ f(y) - f(x) \leq f'(y)(y-x) \end{cases}$$

$$\begin{aligned} \Rightarrow f'(y)(y-x) &\geq f'(x)(y-x) \\ f'(y)(y-x) - f'(x)(y-x) &\geq 0 \\ \frac{f'(y) - f'(x)}{y-x} &\geq 0 \end{aligned}$$

For $y \to x^+$, we have:

$$\Rightarrow \boxed{f''(x) \geq 0}$$

3 ⇒ 2 Using Taylor series (with the mean-value form of the remainder):

$$f(y) = f(x) + f'(x)(y-x) + \frac{1}{2} f''(z)(y-x), \qquad z \in [x, y]$$

$$\Rightarrow \boxed{f(y) \geq f(x) + f'(x)(y-x)}$$

1.2.1 Results

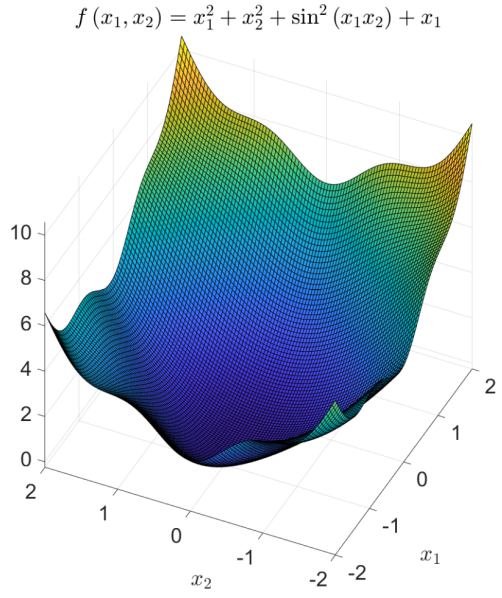
Consider the convex function $f : \mathcal{C} \longrightarrow \mathbb{R}$.:

1. Any local minimum point is also a global minimum point.
2. The level set $\mathcal{D}_c = \{\boldsymbol{x} : f(\boldsymbol{x}) \leq c\}$ is a convex set.
3. If f is strictly convex, then the (global) minimum point (if exist) is unique.
4. If f is differentiable and $\boldsymbol{x}^* \in \mathcal{C}$ is not on the boundary of $\mathcal{C}$ ($\boldsymbol{x}^* \notin \partial\mathcal{C}$) then:

$$\boxed{\boldsymbol{x}^* \text{ is a minimum point} \Leftrightarrow \nabla f(\boldsymbol{x}^*) = \mathbf{0}}$$

2 Unconstrained Optimization (gradient descent)

Consider the following function:



Find the minimum point of f:

$$\{x_1^*, x_2^*\} = \arg\min_{x_1, x_2} f(x_1, x_2) = ?$$

We can try to compare the gradient to zero:

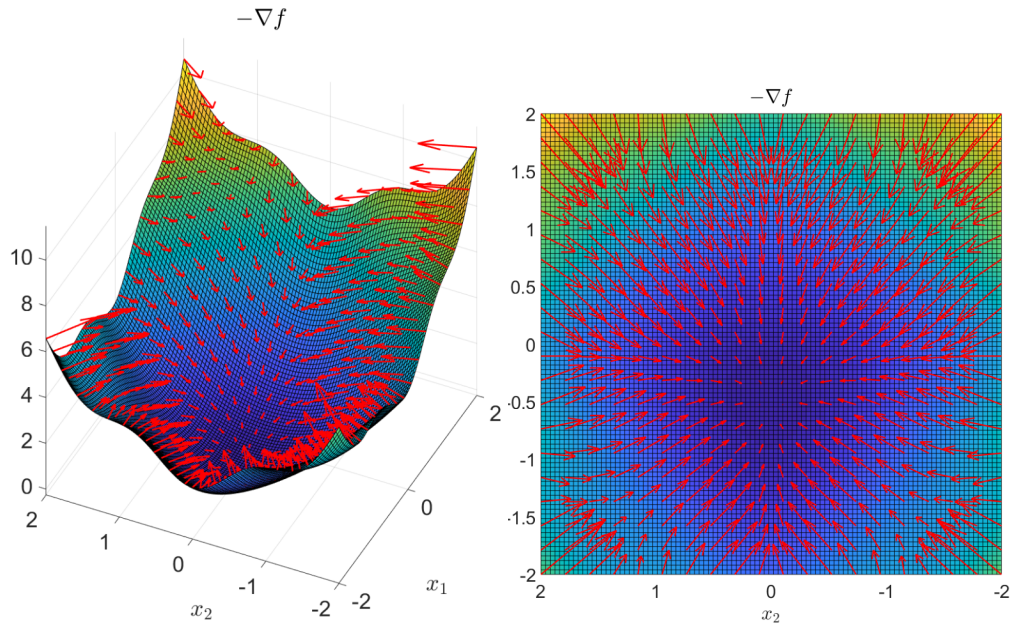
$$\begin{aligned}
\nabla f &= \mathbf{0} \\
\begin{bmatrix} \frac{\partial}{\partial x_1} f(x_1, x_2) \\ \frac{\partial}{\partial x_2} f(x_1, x_2) \end{bmatrix} &= \mathbf{0} \\
\begin{bmatrix} 2x_1 + x_2 \sin(2x_1 x_2) + 1 \\ 2x_2 + x_1 \sin(2x_1 x_2) \end{bmatrix} &= \mathbf{0}
\end{aligned}$$

$$\Rightarrow \begin{cases} 2x_1 + x_2 \sin(2x_1 x_2) + 1 = 0 \\ 2x_2 + x_1 \sin(2x_1 x_2) = 0 \end{cases}$$

In general, this equations has no analytical solution.

2.1 Gradient descent

Consider the (minus) gradient of f at each point:



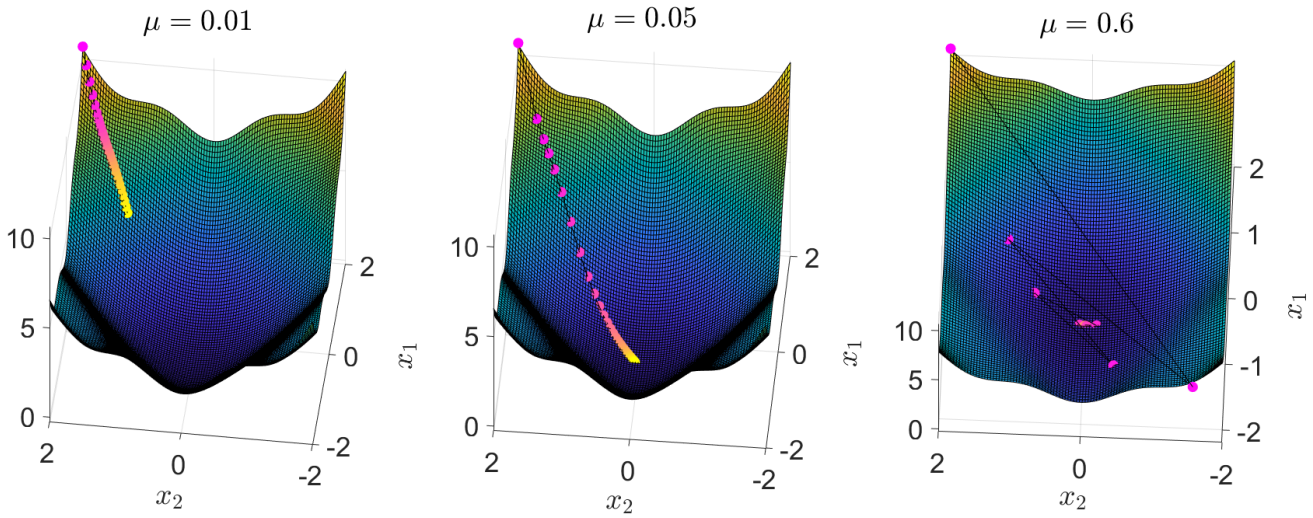
At each point, the gradient ∇f points in the direction of (locally) maximum increase in f.
Thus, to find a minimum point of f one can "go" in the opposite direction: $-\nabla f$.

Algorithm 1 Gradient Descent Algorithm

1. Set some initial point $\boldsymbol{x}_0$
2. **for** $k = 0, 1, 2, \ldots$ Update:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \mu \nabla f\left(\boldsymbol{x}_k\right)$$

The value of μ control the size of the step (learning rate).
We set $\boldsymbol{x}_0 = \begin{bmatrix} 2 & 2 \end{bmatrix}$ and apply the algorithm for 30 iterations:

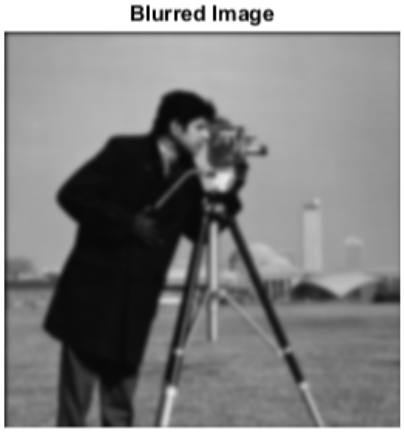


2.2 Matrix inversion example

Consider the following 256×256 image:

We denote the column stack of the image with $\boldsymbol{x} \in \mathbb{R}^{65,536}$
Let $\boldsymbol{H}$ be a blurring matrix and let $\boldsymbol{y}$ be the blurred version (in column stack) of the image:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}, \qquad \boldsymbol{H} \in \mathbb{R}^{65,536 \times 65,536}$$



To restore the original image $\boldsymbol{x}$ from $\boldsymbol{y}$ we need to compute:

$$\hat{\boldsymbol{x}} = \boldsymbol{H}^{-1}\boldsymbol{y}, \qquad \text{Assuming } \boldsymbol{H} \text{ is invertible}$$

But, most computers cannot invert a $65,536 \times 65,536$ matrix.
Instead, we can defined the following optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2$$

The optimal solution is indeed $\hat{\boldsymbol{x}} = \boldsymbol{H}^{-1}\boldsymbol{y}$.
The gradient is given by:

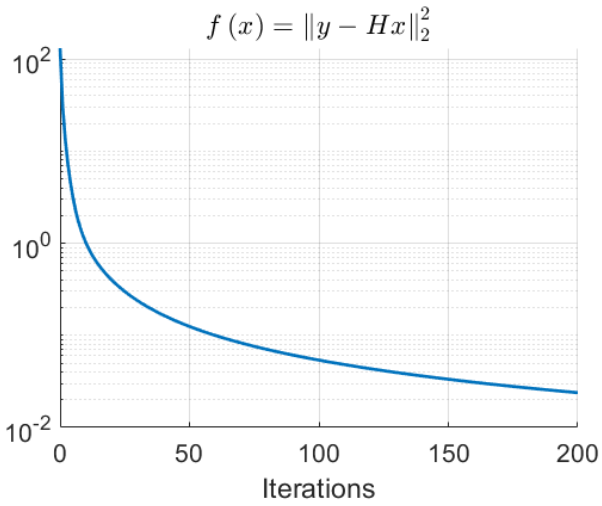
$$\nabla f(\boldsymbol{x}) = -2\boldsymbol{H}^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})$$

We set the initial solution:

$$\boldsymbol{x}_0 = \boldsymbol{y}$$

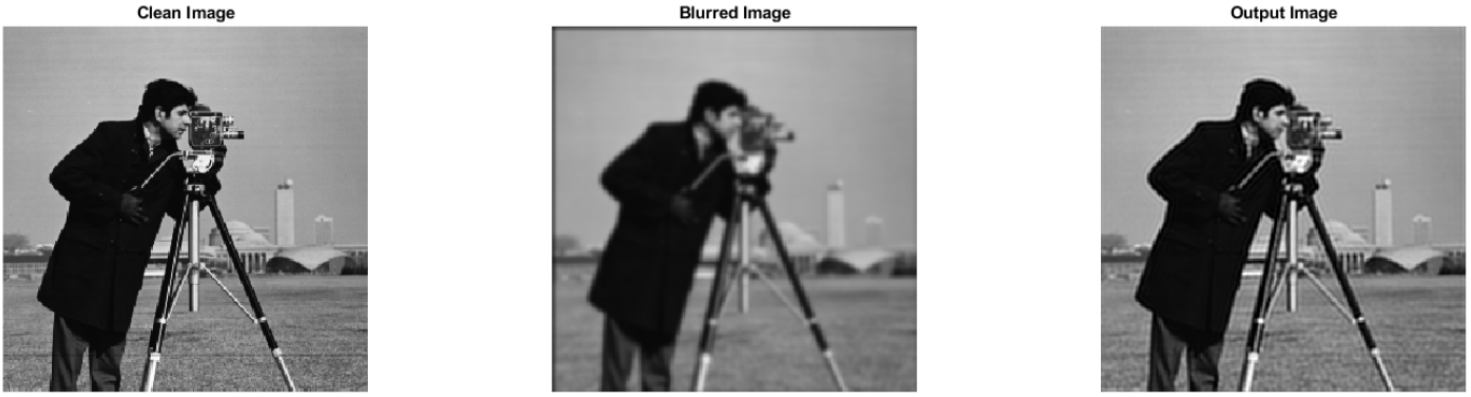
and $\mu = 0.9$.
We apply 200 iterations of the gradient descent algorithm.

The obtain values of $f(\boldsymbol{x}_k)$ (in logarithmic scale):



Some images during the iterations:

Final result:

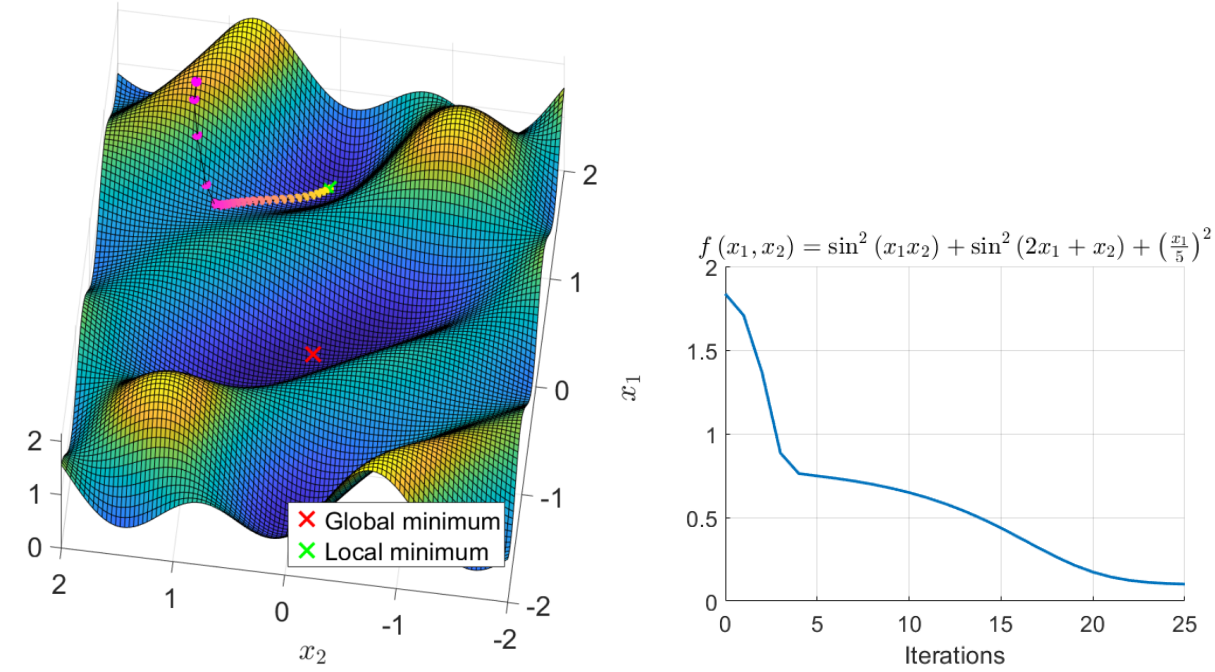


2.3 Non-convex function

- If the function f is convex, then, under some conditions on the step size μ, the gradient descent algorithm converges to the global minimum.
- If the function f is non-convex, then, the gradient might only converge to a local minimum.

Example:

$$f(x_1, x_2) = \sin^2(x_1 x_2) + \sin^2(2x_1 + x_2) + \left(\tfrac{x_1}{5}\right)^2$$



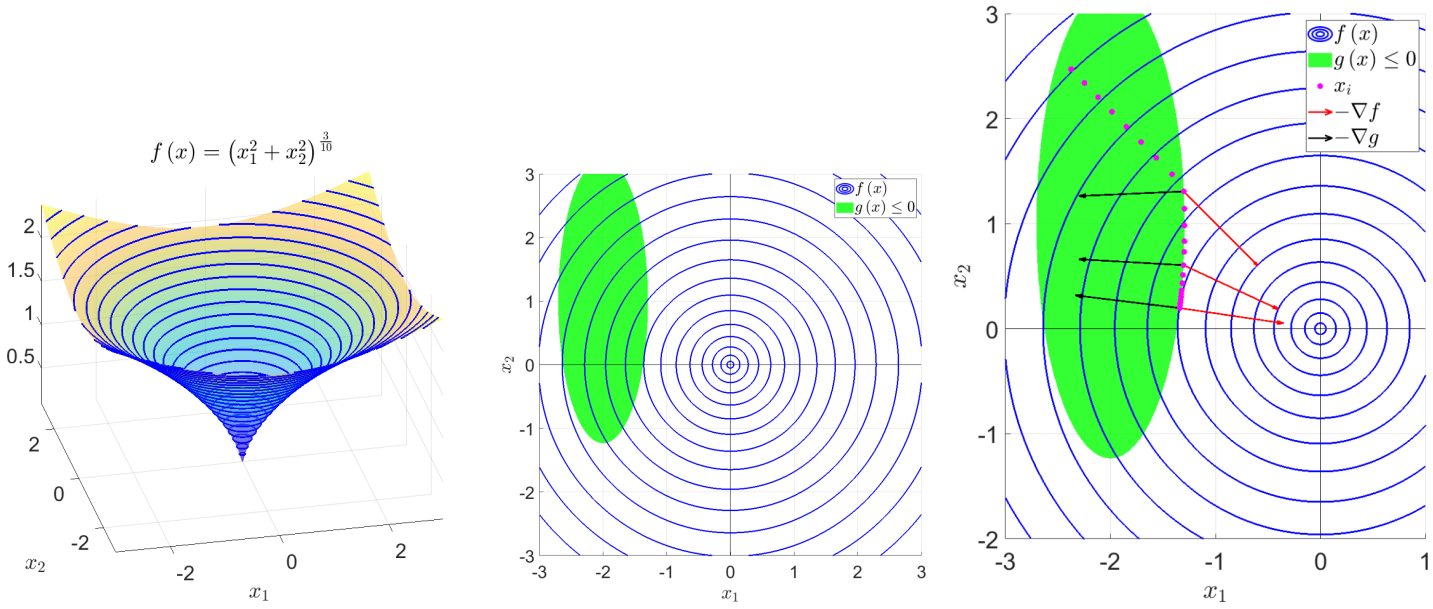
The gradient descent algorithm converge to a local minimum (note that $f(0,0) = 0$).

3 Constrained Optimization

Consider the following constrained optimization problem:

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \text{subject to} \\ g(\boldsymbol{x}) \leq 0 \end{cases} = \begin{cases} \arg\min\limits_{\boldsymbol{x}} \left(x_1^2 + x_2^2\right)^{\frac{3}{10}} \\ \text{subject to} \\ 2(x_1+2)^2 + \frac{1}{5}(x_2-1)^2 - 1 \leq 0 \end{cases}$$

Consider a starting point $\boldsymbol{x}_0$ inside the feasible area: $g(\boldsymbol{x}_0) \leq 0$.
Applying gradient descend to f while satisfying the condition $g \leq 0$ yields the following iterations (right figure):



Note that the algorithm converged to a point such that:

$$\nabla f(\boldsymbol{x}^*) = -\lambda \nabla g(\boldsymbol{x}^*), \qquad \text{for some } \lambda > 0$$

$$\Rightarrow \boxed{\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0}$$

Thus, we write the Lagrangian (λ is called Lagrangian multiplier):

$$\boxed{\mathcal{L}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})} \Longrightarrow \boxed{\nabla_x \mathcal{L} = \nabla f(\boldsymbol{x}) + \lambda \nabla g(\boldsymbol{x})}$$

The optimal point $\boldsymbol{x}^*$ satisfying the KKT conditions:

1. Optimality under the constraint: $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \lambda^*) = 0$
2. Lagrange multiplier is non-negative: $\lambda^* \geq 0$
3. Solution is in the feasible area: $g(\boldsymbol{x}^*) \leq 0$
4. Complimentary slackness: $\lambda^* g(\boldsymbol{x}^*) = 0$

Condition 4 states that at least $g(\boldsymbol{x}^*) = 0$ (active constraint) or $\lambda^* = 0$ (non-active constraint).

3.1 Examples

3.1.1 Active constraint

Solve the following constraint problem:

$$\begin{cases} \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_2^2 & \boldsymbol{x} \in \mathbb{R}^2 \\ \text{subject to} & \\ (x_1+2)^2 + (x_2)^2 - 1 \leq 0 & \end{cases}$$

Note that $\boldsymbol{x} = \mathbf{0}$ is the minimizer of $\|\boldsymbol{x}\|^2$ but it is not satisfy the constraint.

Solution:

The Lagrangian is given by $\left(\|\boldsymbol{x}\|^2 = x_1^2 + x_2^2\right)$:

$$\mathcal{L} = x_1^2 + x_2^2 + \lambda\left((x_1+2)^2 + (x_2)^2 - 1\right)$$

$$\Rightarrow \nabla_{\boldsymbol{x}}\mathcal{L} = \mathbf{0}$$

$$\begin{bmatrix} 2x_1 + \lambda(2x_1+4) \\ 2x_2 + 2\lambda x_2 \end{bmatrix} = \mathbf{0}$$

$$\Rightarrow \boldsymbol{x}^* \triangleq \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} -\frac{2\lambda}{1+\lambda} \\ 0 \end{bmatrix}$$

Complimentary slackness:

$$\lambda g(\boldsymbol{x}^*) = 0$$

If $\lambda = 0$ then $\boldsymbol{x}^* = \mathbf{0}$ but $g(\mathbf{0}) > 0$, namely, $\boldsymbol{x} = \mathbf{0}$ is not in the feasible area.
Thus, $\lambda \neq 0$ (g is an active constraint):

$$\Rightarrow g(\boldsymbol{x}^*) = 0$$

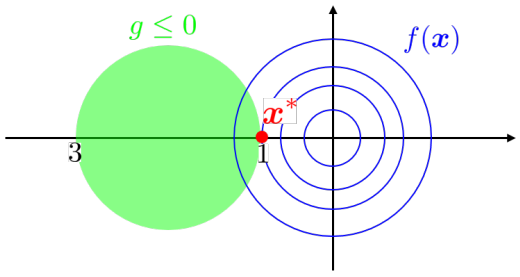
$$(x_1^*+2)^2 + (x_2^*)^2 - 1 = 0$$

$$(x_1^*+2)^2 = 1$$

$$-\frac{2\lambda}{1+\lambda} + 2 = \pm 1$$

$-\frac{2\lambda}{1+\lambda} + 2 = -1$ $2\lambda = 3(1+\lambda)$ $\lambda = -3$	$-\frac{2\lambda}{1+\lambda} + 2 = 1$ $2\lambda = 1+\lambda$ $\lambda = 1$
The Lagrange multiplier need to be non-negative: $\lambda \geq 0$	This is result is valid.
~~$\lambda = -3$~~	$\boxed{\lambda = 1}$

$$\Rightarrow \boxed{\boldsymbol{x}^* = \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} -\frac{2\lambda}{1+\lambda} \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}}$$



3.1.2 Non-active constraint

Solve the following constraint problem:

$$\begin{cases} \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_2^2 & \boldsymbol{x} \in \mathbb{R}^2 \\ \text{subject to} \\ (x_1+2)^2 + (x_2)^2 - 10 \leq 0 \end{cases}$$

Solution:

As before, the Lagrangian is given by $\left(\|\boldsymbol{x}\|^2 = x_1^2 + x_2^2\right)$:

$$\mathcal{L} = x_1^2 + x_2^2 + \lambda\left((x_1+2)^2 + (x_2)^2 - 1\right)$$

$$\Rightarrow \nabla_{\boldsymbol{x}}\mathcal{L} = \mathbf{0}$$

$$\begin{bmatrix} 2x_1 + \lambda(2x_1+4) \\ 2x_2 + 2\lambda x_2 \end{bmatrix} = \mathbf{0}$$

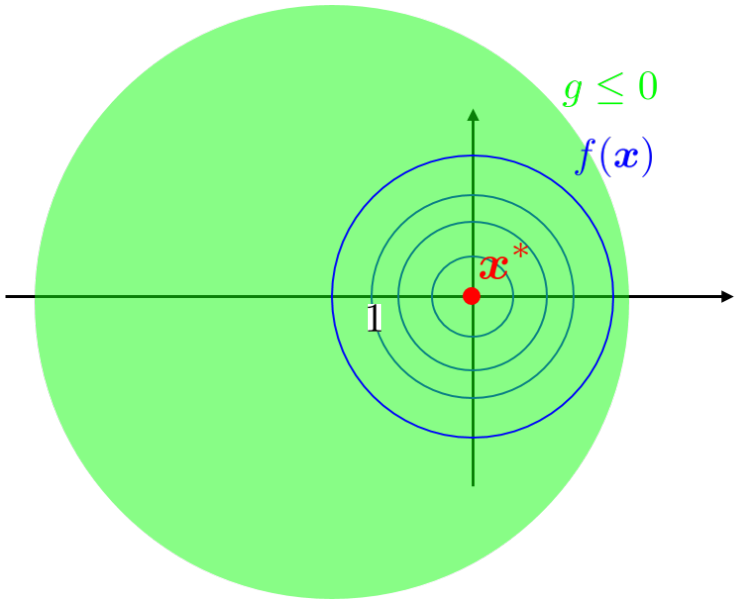
$$\Rightarrow \boldsymbol{x}^* \triangleq \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} -\frac{2\lambda}{1+\lambda} \\ 0 \end{bmatrix}$$

Complimentary slackness:

$$\lambda g(\boldsymbol{x}^*) = 0$$

If $\lambda = 0$ then $\boldsymbol{x}^* = \mathbf{0}$ and also $g(\mathbf{0}) < 0$, namely, $\boldsymbol{x} = \mathbf{0}$ is within the feasible area.
Thus, $\lambda = 0$ and the constraint is not active:

$$\Rightarrow \boxed{\boldsymbol{x}^* = 0}$$



3.2 General Case

A general constrained optimization problem is given by:

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}} f(\boldsymbol{x}) & \\ \text{subject to} & \\ g_i(\boldsymbol{x}) \leq 0 & i = 1, 2 \ldots, N \\ h_j(\boldsymbol{x}) = 0 & j = 1, 2 \ldots, M \end{cases}$$

g_i are the inequality constraints and h_j are the equality constraints.
The Lagrangian is given by:

$$\mathcal{L} = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x}) + \sum_j \mu_j h_j(\boldsymbol{x})$$

If the functions $\{f, \{g_i\}, \{h_j\}\}$ are convex,
the point $\boldsymbol{x}$ is an optimal solution if the generalized KKT (Karus-Khun-Tucker) conditions are satisfied.