NMK40403 ARTIFICIAL INTELLIGENCE

SUPPORT VECTOR MACHINES 03: Unconstrained minimization

Mohamed Elshaikh

Unconstrained minimization

- In previous slide, we discovered that to maximize the margin we need to minimize the norm of w.
- It means we need to solve the following optimization problem:

  Minimize in (w,b) $\|w\|$

  subject to $y_i(w \cdot x_i + b) \geq 1$

  (for any i=1,…,n)

- The first thing to notice about this optimization problem is that it has constraints. They are defined by the line which begins with "subject to". You may think that there is only one constraint, but there is, in fact, n constraints. (this is because of the last line "for any"...)

Unconstrained minimization

- Unconstrained minimization
- Before tackling such a complicated problem, let us start with a simpler one. We will first look at how to solve an unconstrained optimization problem, more specifically, we will study unconstrained minimization.
- That is the problem of finding which input makes a function return its minimum. (Note: in the SVM case, we wish to minimize the function computing the norm of w, we could call it f and write it $f(w)=\|w\|$).
- Let us consider a point x^* (you should read it "x star", we just add the star so that you know we are talking about a specific variable, and not about any x).

Unconstrained minimization

- Theorem:
  - Let $f:\Omega \rightarrow R$ be a continuously twice differentiable function at x^*.
  - If x^* satisfies $\nabla f(x^*)=0$ and $\nabla^2 f(x^*)$ is positive definite then x^* is a local minimum.
  - The hard truth with such a theorem is that although being extremely concise, it is totally impossible to understand without some background information. What is $\nabla f(x^*)=0$? What is $\nabla^2 f(x^*)$? What do we mean by positive definite?

Unconstrained minimization

- **Theorem (with more details):**
- If x^* satisfies:

  f has a zero gradient at x^*:

  $$\nabla f(x^*)=0$$

  and

  the Hessian of f at x^* is positive definite:

  $$z^T((\nabla^2 f(x^*))z>0, \forall z \in R^n$$

  - where
    - $$\nabla^2 f(x)=\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$
- then x^* is a local minimum.

Unconstrained minimization

- What does this all mean?
- Let us examine this definition step by step.

- Step 1:

- Let $f:\Omega \rightarrow R$ be a continuously twice differentiable function at x^*.

- First, we introduce a function which we call f, this function takes its values from a set Ω (omega) and returns a real number. There is the first difficulty here because we do not state what Ω is, but we will be able to guess it in the next line. This function f should be continuous and twice differentiable, or the rest of the definition will not be true.

Unconstrained minimization

- Step 2:

- x^* is a local minimum of f(x) if and only if:

- We want to find a value to give to f for it to produce its minimum. We simply name this value x^*.

- From the notation we can tell two things:
- x^* is written in bold, so it is a vector. It means that f is a multivariate function.
- As a result, the set Ω we saw earlier is the set from which we pick values to give to f. It means that the set Ω is a set of vectors and $x^* \in \Omega$ ("x stars belongs to Omega")

Unconstrained minimization

- Step 3:

- f has a zero gradient at x^*
- This one is the first condition which must hold if we want x^* to be a local minimum of f(x). We must check that the gradient of the function f at x^* is equal to zero.

- What is the gradient? Just think of it as a derivative on steroids.

- Definition: "the gradient is a generalization of the usual concept of derivative of a function in one dimension to a function in several dimensions" (Wikipedia)

Unconstrained minimization

- A gradient is, in fact, the same thing as a derivative, but for functions like f which take vectors as input.
- That is why we wanted f to be a differentiable function in the first place, if it is not the case we cannot compute the gradient, and we are stuck.
- In calculus, when we want to study a function, we often study the sign of its derivative.
- It allows you to determine if the function is increasing or decreasing and to identify minimum and maximum.
- By setting the derivative to zero, we can find the "critical points" of the function at which it reaches a maximum or a minimum.
- When we work with functions having more variable, we need to set each partial derivative to zero.

Unconstrained minimization

- It turns out, the gradient of a function is a vector containing each of its partial derivatives.
- By studying the sign of the gradient, we can gather important pieces of information about the function.
- In this case, checking if the gradient equals zero for $x*$ allow us to determine if $x*$ is a critical point (and that the function f possibly has a minimum at this point).
- (Note: Checking if the gradient equals zero at a point means checking that each partial derivative equals zero for this point)

Unconstrained minimization

- The gradient of a function is denoted by the symbol ∇ (nabla).
- The line
- $\nabla f(x^*)=0$

- is just a repetition of "f has a zero gradient at x∗" in mathematical notation.
- For a vector $x^*(x_1,x_2,x_3)$, $\nabla f(x^*)=0$ means:

$$\frac{\partial f}{\partial x_1}(\mathbf{x}^*) = 0$$

$$\frac{\partial f}{\partial x_2}(\mathbf{x}^*) = 0$$

$$\frac{\partial f}{\partial x_3}(\mathbf{x}^*) = 0$$

Unconstrained minimization

- Step 4:

- the Hessian of f at x^* is positive definite

- That is where most people get lost. This single sentence requires a lot of backgrounds.
- You need to know:
- that the Hessian is a matrix of second-order partial derivatives
- how we can tell if a matrix is positive definite

The Hessian matrix

- The Hessian is a matrix, and we give it a name.
- We could call it H but instead we call it $\nabla^2 f(x)$ which is more explicit.
- We keep the symbol ∇ used for the gradient, and add a 2 to denote we the fact that this time we are talking about second-order partial derivative.
- Then we specify the name of the function (f) from which we will compute these derivates.
- By writing f(x) we know that f takes a vector x as input and that the Hessian is computed for a given x.

Unconstrained minimization

- To sum up, we need to compute a matrix called the Hessian matrix for x^*.
- So we take the function f, we take the value of x^* and we compute the value for each cell of the matrix using the following formula:
- Eventually we get the Hessian matrix and it contains all the numbers we have computed.

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Unconstrained minimization

- Let us look at the definition to see if we understand it well:

- Definition: In mathematics, the Hessian matrix or Hessian is a square matrix of second-order partial derivatives of a scalar-valued function. It describes the local curvature of a function of many variables. (Wikipedia)

- (Note: A scalar valued function is a function that takes one or more values but returns a single value. In our case f is a scalar valued function.)

Positive definite

- Now that we have the Hessian matrix, we want to know if it is positive definite at x^*.
- Definition:
  - A symmetric matrix A is called positive definite if $x^TAx>0$, for all $x \in Rn$.
- This time, we note that once again we were given the definition in the first place. It was just a little bit harder to read because of our notational choice. If we replace A by $\nabla 2f(x^*)$ and x by z we get exactly the formula:
- $z^T((\nabla 2f(x^*))z>0, \forall z \in Rn$
- The problem with this definition is that it is talking about a symmetric matrix. A symmetric matrix is a square matrix this is equal to its transpose.

Unconstrained minimization

- The Hessian matrix is square, but is it symmetric?
- Luckily for us yes!
- "if the second derivatives of f are all continuous in a neighborhood D, then the Hessian of f is a symmetric matrix throughout D" (Wikipedia)
- But even with the definition, we still don't know how to check that the Hessian is positive definite. That is because the formula $z^T((\nabla^2 f(x^*))z>0$, is for all z in Rn
- We can't try this formula for all z in Rn!

Unconstrained minimization

- That is why we will use the following theorem:
- Theorem:
- The following statements are equivalent:
  - The symmetric matrix A is positive definite.
  - All eigenvalues of A are positive.
  - All the leading principal minors of A are positive.
  - There exists nonsingular square matrix B such that $A=B^TB$
- So we have three ways of checking that a matrix is positive definite:
  - By computing its eigenvalues and checking they are positive.
  - By computing its leading principal minors and checking they are positive.
  - By finding a nonsingular square matrix B such that $A=B^TB$.

Computing the leading principal minors

Minors: To compute the minor M_{ij} of a matrix we remove the i^{th} line and the j^{th} column, and compute the determinant of the remaining matrix.

- Example: : Let us consider the following 3 by 3 matrix:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

  - To compute the minor M_{12} of this matrix we remove the line number 1 and the column number 2. We get:

$$\begin{pmatrix} \square & \square & \square \\ d & \square & f \\ g & \square & i \end{pmatrix}$$

  - so we compute the determinant of:

$$\begin{pmatrix} d & f \\ g & i \end{pmatrix}$$

  - which is : ***di−fg***

Computing the leading principal minors

- Principal minors
- A minor M_{ij} is called a ***principal minor*** when ***i=j***.

- For our 3x3 matrix, the principal minors are :

  - $M_{11}=ei-fh$
  - $M_{22}=ai-cg$
  - $M_{33}=ae-bd$

- But that is not all ! Indeed, minors also have what we call an order.

Computing the leading principal minors

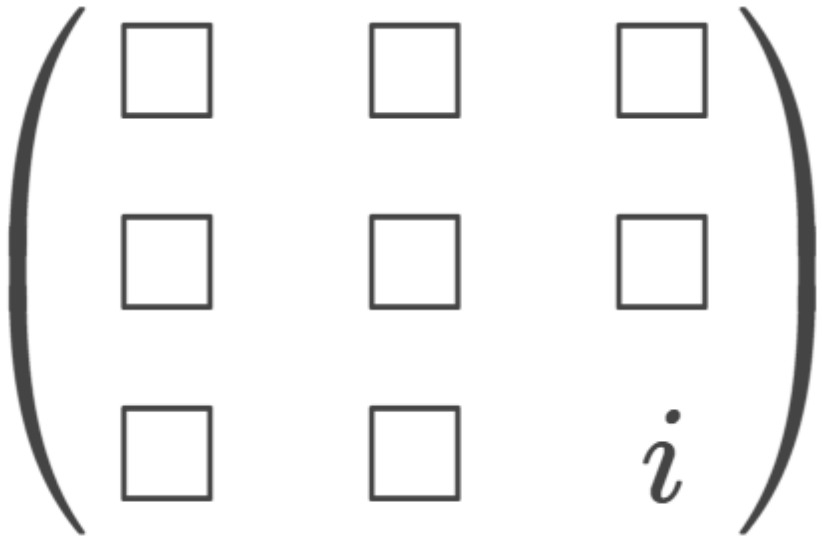
- Definition:
- A minor of A of order k is principal if it is obtained by deleting $n-k$ rows and the $n-k$ columns with the same numbers.

- In our previous example, the matrix is 3×3 so $n=3$ and we deleted 1 line, so we computed principal minors of order 2.

- There are $\binom{n}{k}$□ principal minors of order k, and we write Δk for any of the principal minors of order k.

Computing the leading principal minors

- **To sum-up**:
- Δ_0: does not exist because if we remove three lines and three columns we have deleted our matrix!
- Δ_1: We delete (3-1) = 2 lines and 2 columns with the same number.
- So we remove lines 1 and 2 and column 1 and 2.



Computing the leading principal minors

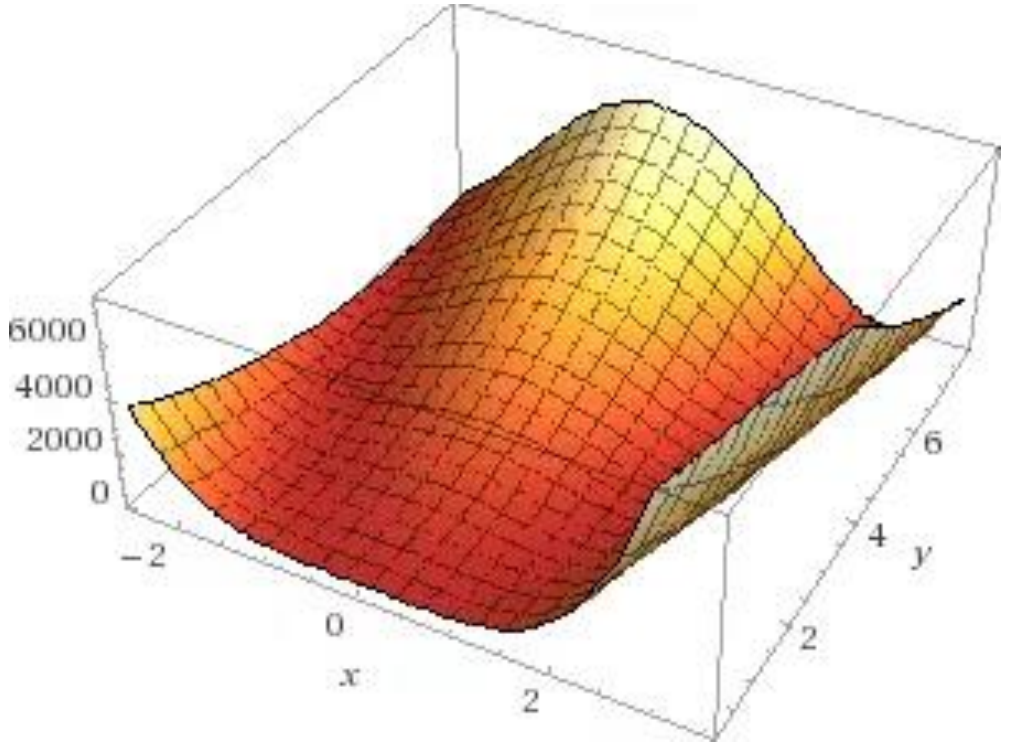
- It means that one of the principal minors of order 1 is i. Let us find the others:
    - We delete lines 2 and 3 and column 2 and 3 and we get a.
    - We delete lines 1 and 3 and column 1 and 3 and we get e
- $\boldsymbol{\Delta_2}$: is what we have seen before:
    - $\mathrm{M}_{11}=\mathrm{ei}-\mathrm{fh}$
    - $\mathrm{M}_{22}=\mathrm{ai}-\mathrm{cg}$
    - $\mathrm{M}_{33}=\mathrm{ae}-\mathrm{bd}$
- $\boldsymbol{\Delta_3}$: We delete nothing. So it is the determinant of the matrix :
    - $\boldsymbol{aei+bfg+cdh-ceg-bdi-afh}$.

Leading principal minor

- Definition:
- The leading principal minor of A of order k is the minor of order k obtained by deleting the last $n-k$ rows and columns.
- So it turns out leading principal minors are simpler to get. If we write D_k for the leading principal minor of order k we find that:
    - $D_1=a$ (we deleted the last two lines and column)
    - $D_2=ae-bd$ (we removed the last line and the last column)
    - $D_3=aei+bfg+cdh-ceg-bdi-afh$
- Now that we can compute all the leading principal minors of a matrix, we can compute them for the Hessian matrix at x^* and if they are all positive, we will know that the matrix is positive definite.
- We now have fully examined what we have to know, and you should be able to understand how to solve an unconstrained minimization problem.

Unconstrained minimization

- **Computing the leading principal minors**
- **Example**
- In this example we will try to find the minimum of the function:
  $f(x,y)=(2-x)^2+100(y-x^2)^2$
  which is known as the Rosenbrock's banana function.



The Rosenbrock function for a = 2 and b = 100

Computing the leading principal minors

Solution

1- search for which point it gradient equals zero:

$$\nabla f(x,y) = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}$$

2- compute the partial derivatives, and we find:

$$\frac{\partial f}{\partial x} = 2(200x^3 - 200xy + x - 2)$$

$$\frac{\partial f}{\partial y} = 200(y - x^2)$$

Computing the leading principal minors

Our goal is to find when they are both at zero, so we need to solve the following system of equations:

$$2(200x^3 - 200xy + x - 2) = 0 \tag{1}$$
$$200(y - x^2) = 0 \tag{2}$$

We distribute to get:

$$400x^3 - 400xy + 2x - 4 = 0 \tag{3}$$
$$200y - 200x^2 = 0 \tag{4}$$

We multiply (2) by $2x$ to obtain:

$$400xy - 400x^3 = 0 \tag{5}$$

We now add (3) and (5) to get:

$$400x^3 - 400xy + 2x - 4 + 400xy - 400x^3 = 0 \tag{6}$$

Computing the leading principal minors

which simplifies into:

$$2x - 4 = 0$$

$$x = 2$$

We substitute x in (4)

$$200y - 200 \times 2^2 = 0$$

$$200y - 800 = 0$$

$$y = \frac{800}{200}$$

$$y = 4$$

It looks like we have found the point $(x, y) = (2, 4)$ for which $\nabla f(x, y) = 0$. But is this a minimum?

The hessian Matrix

The Hessian matrix is :

$$\nabla^2 f(x,y) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial xy} \\ \frac{\partial^2 f}{\partial yx} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

$$\frac{\partial^2 f}{\partial x^2} = 1200x^2 - 400y + 2$$

$$\frac{\partial^2 f}{\partial xy} = -400x$$

$$\frac{\partial^2 f}{\partial yx} = -400x$$

$$\frac{\partial^2 f}{\partial y^2} = 200$$

Let us now compute the Hessian for $(x, y) = (2, 4)$

$$\nabla^2 f(x,y) = \begin{pmatrix} 3202 & -800 \\ -800 & 200 \end{pmatrix}$$

The hessian Matrix

The matrix is symetric, we can check its leading principal minors:

Minors of rang 1:

If we remove the last line and last column the minor M_{11} is 3202.

Minor of rang 2:

This is the determinant of the Hessian:

$$3202 \times 200 - (-800) \times (-800) = 400$$

$(2, 4)$

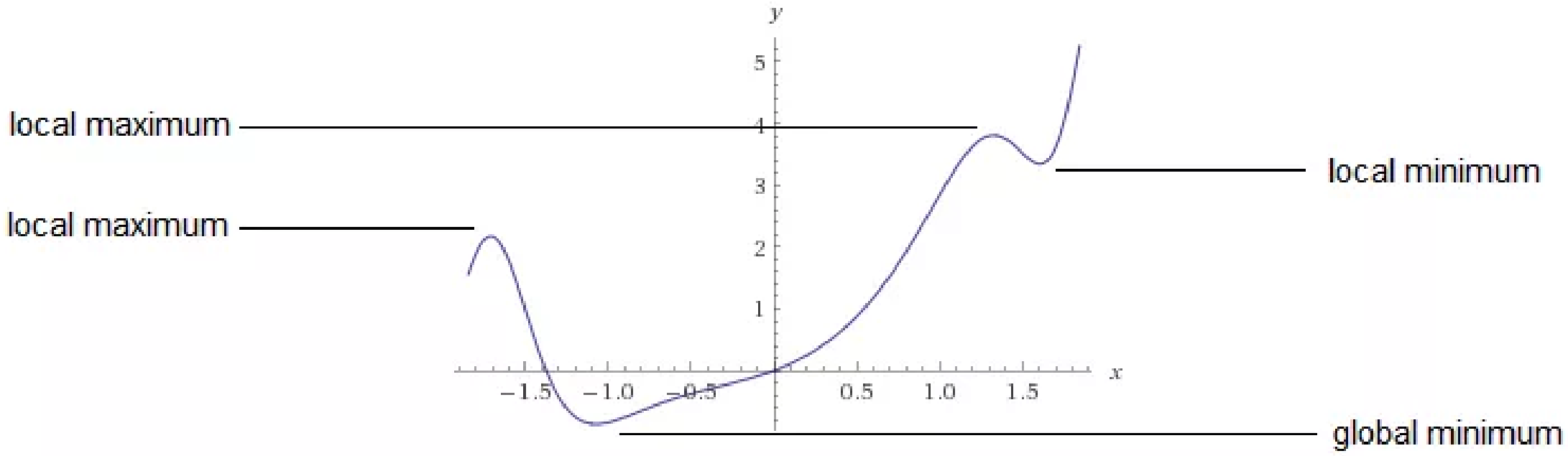
$$\begin{pmatrix} 3202 & -800 \\ -800 & 200 \end{pmatrix}$$

All the leading principal minors of the Hessian are positives. It means that the Hessian is positive definite.

The two conditions we needed are met, and we can say that the point $(2, 4)$ is a local minimum.

LOCAL minimum?

- A point is called a local minimum when it is the smallest value within a range. More formally:
  - Given a function f defined on a domain X, a point x^* is said to be a local minimum if there exists some $\epsilon > 0$ such that $f(x^*) \leq f(x)$ for all x in X within distance ϵ of x^*.
- This is illustrated in the figure below:



GLOBAL minimum

- A global minimum, however, is true for the whole domain of the function:
    - Given a function f defined on a domain X, a point x^* is said to be a global minimum if $f(x^*) \leq f(x)$ for all x in X
- So all our hard work was just to find a local minimum, but in real life, we often want to find the global minimum.

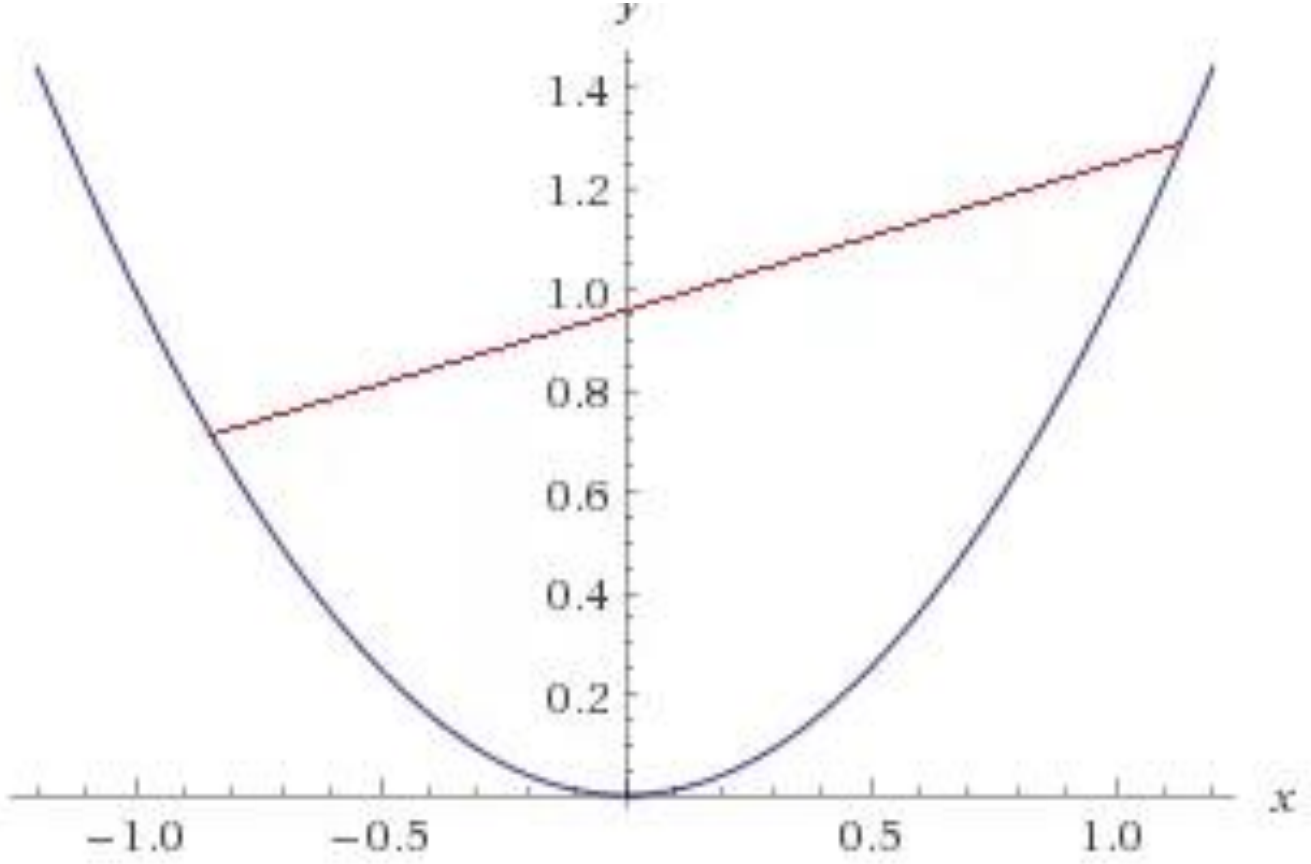
How can we find a GLOBAL minimum

- There is one simple way to find the global minimum:
    1. Find all the local minima
    2. Take the smallest one; it is the global minimum.
- Another approach is to study the function we are trying to minimize. If this function is ***convex***, then we are sure its local minimum is a global minimum.

- Theorem: ***A local minimum of a convex function is a global minimum***

Convex functions

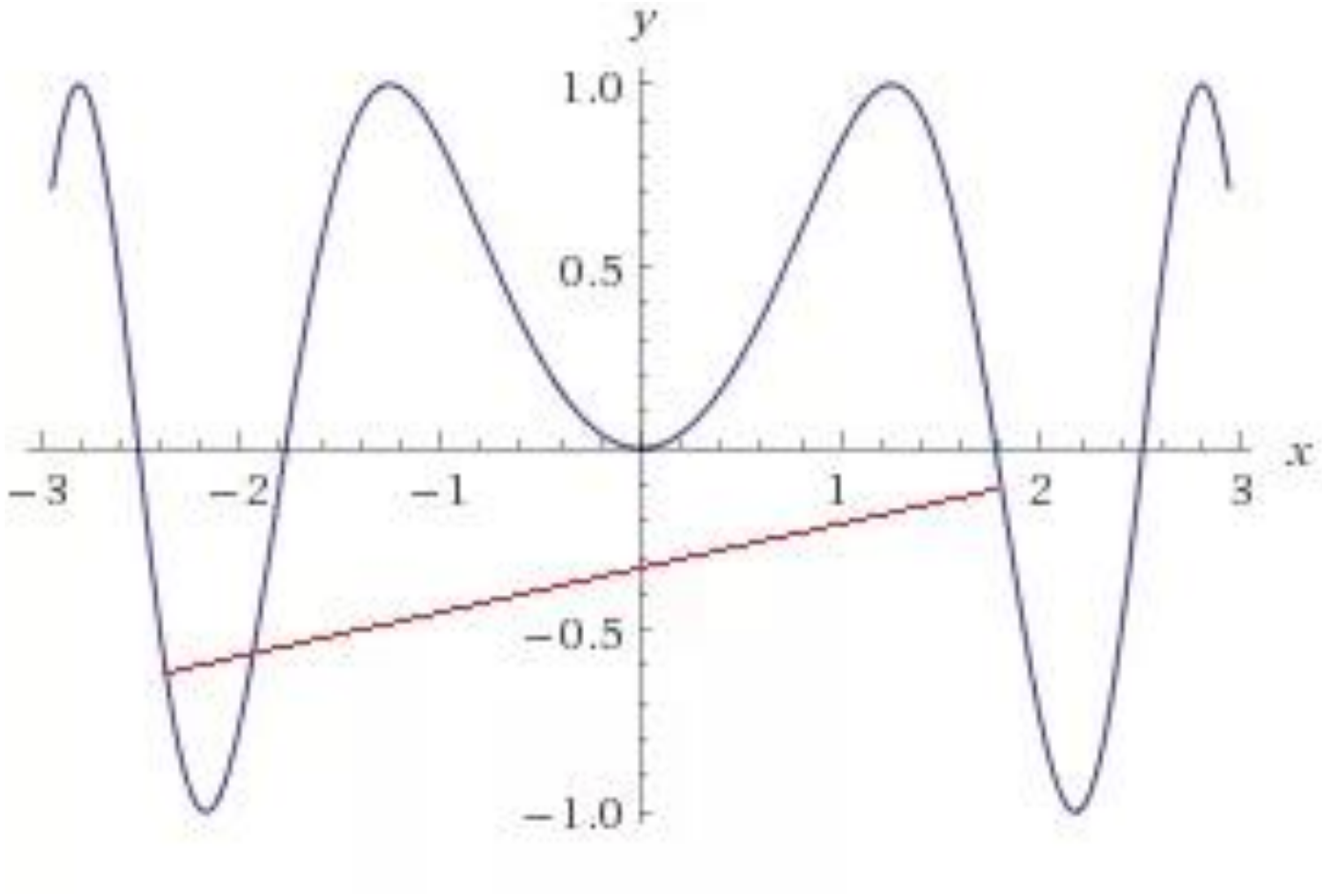
- What is a convex function?
  - A function is convex if you can trace a line between two of its points without crossing the function line.



A convex function

Non convex functions

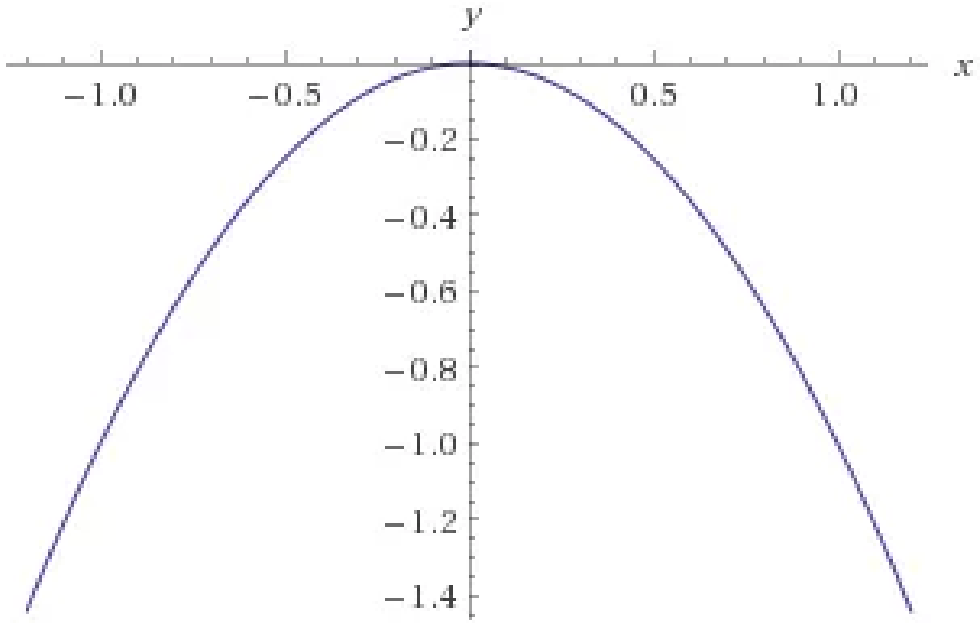
- **Non convex functions**
- However, if you cross the function line, then the function is non-convex.
- As you can see in the figure below, the red line crosses the function, which means it is non-convex. Note, however, that the function is convex on some intervals, for instance on [-1,+1].



A non-convex function

Unconstrained minimization

- As often, there is also an "opposite" concept: a function f is concave if −f is convex.
- The problem here is that the original definition of a convex function is also true, we can trace a line between two points of the function without crossing the line...



A concave function

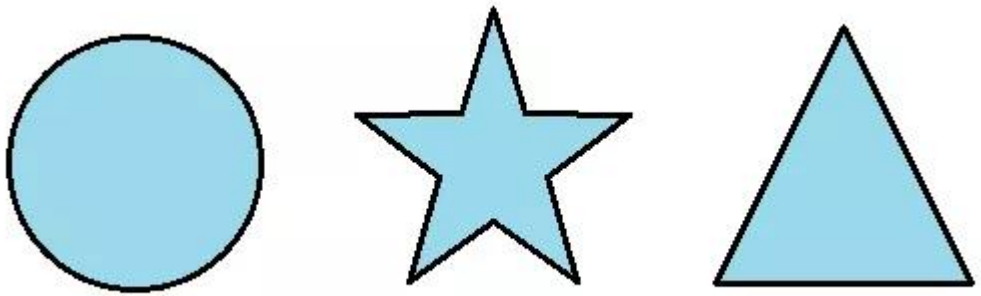
Convex Functions

- So the mathematicians have been a little bit more specific, and they say that:
  - ***A function is convex if its epigraph (the set of points on or above the graph of the function) is a convex set.***

- But what is a convex set?
  - ***In Euclidean space, a convex set is the region such that, for every pair of points within the region, every point on the straight line segment that joins the pair of points is also within the region.***

Convex Functions

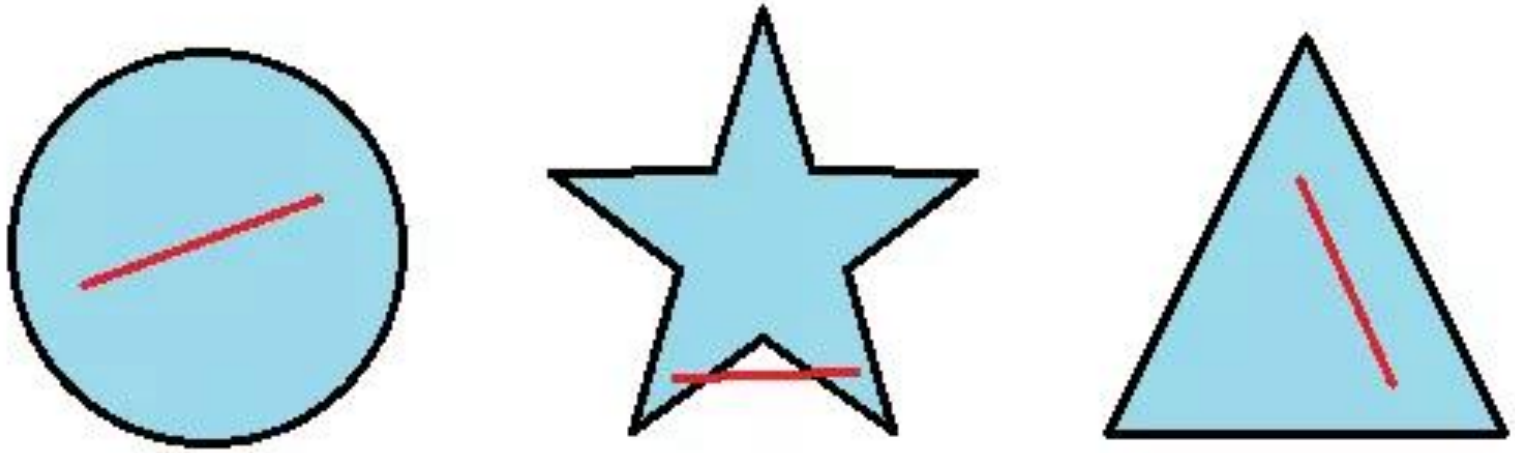
- By using the same logic as before. A set of points is convex if when we pick two points belonging to the set and we trace a line between them then the line is inside the set..



Which set is convex and which set is not convex?

Convex Functions

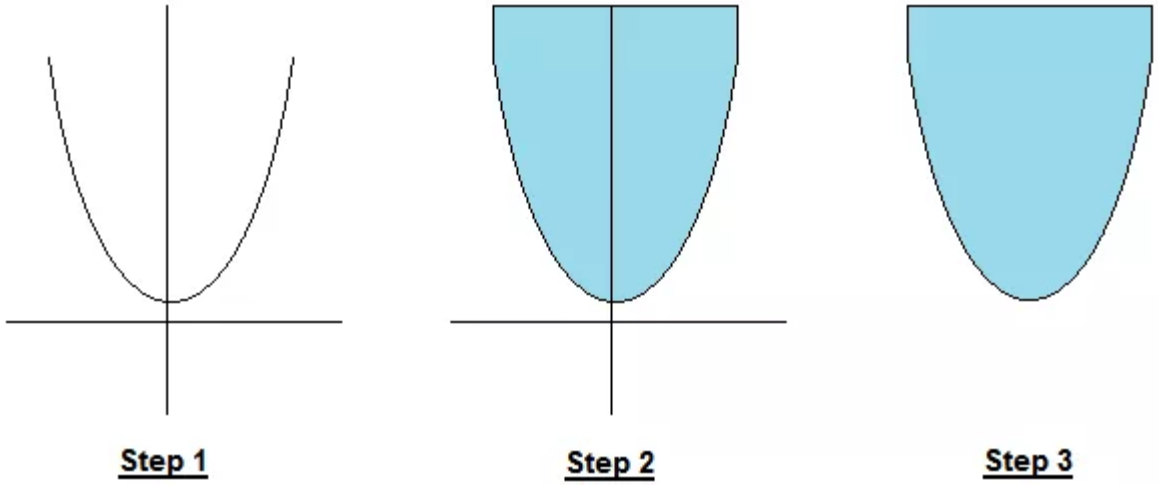
- If you guessed right, the circle and the triangles are convex sets. In the figure below I traced a red line between two points. As you can see, the line joining two points of the star leave the figure indicating that it is not a convex set.



The star is not a convex set

Convex Functions

- We can now use this knowledge to determine if a function is convex.



- Step 1: We have a function and we wish to know if it is convex
- Step 2: We take its epigraph (think of it as filling it with water but the water cannot overflow so it adds up vertically when it reaches the limits of the function)
- Step 3: If the shape of the epigraph is convex, then it is a convex function!

Convex Functions

- How do we know if a function is convex?
- The definition with the epigraph is simple to understand, but with functions with several variables it is kind of hard to visualize. So we need to study the function:
  - ***More generally, a continuous, twice differentiable function of several variables is convex on a convex set if and only if its Hessian matrix is positive semidefinite on the interior of the convex set.***
- If we want to check if a function is convex, one easy way is to use our old friend the Hessian matrix. However, instead of checking if it is positive ***definite***, this time, we need to check if it is positive ***semidefinite***.

Convex Functions

- What is the difference?
- Theorem:
- The following statements are equivalent:
    - The symmetric matrix ***A*** is positive ***semidefinite***.
    - All eigenvalues of ***A*** are ***non-negative***.
    - All the ***principal minors*** of ***A*** are ***nonnegative***.
    - There exists ***B*** such that $\boldsymbol{A=B^TB}$

- As before we will use the minors. The difference here is that we need to check all the principal minors, not only the leading principal minors. Moreover, they need to be nonnegative. (A number is ***positive*** if it is greater than zero. A number is ***non-negative*** if it is greater than or equal to zero).

Convex Functions

- **Example**: is the banana function convex?
- We saw that the Hessian of our banana function was:

$$\nabla^2 f(x,y) = \begin{pmatrix} 1200x^2 - 400y + 2 & -400x \\ -400x & 200 \end{pmatrix}$$

- Its principal minors of rang 1 are:

M_{11} is 200 (we removed line 1 and column 1).

M_{22} is $1200x^2 - 400y + 2$ (we removed line 2 and column 2).

- If the function is convex, these minors should be nonnegative ***on the interior of the convex set***. Which convex set? By definition, the domain of a convex function is a convex set. In our case when we say that a function is convex ***on a convex set***, we are talking about its domain.
- The restriction "on the interior" tells us that we should not pick points which are on the border of the set.

Convex Functions

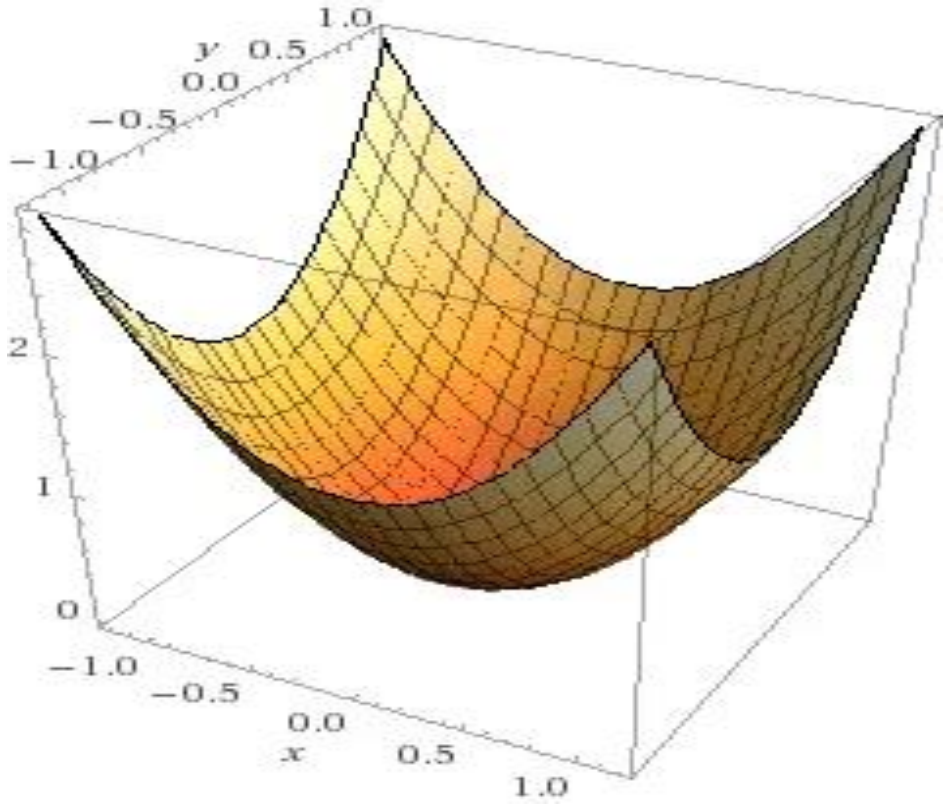
- In our example, the function is defined in R^2 which is a convex set. So we would need to prove that for any point we pick the principal minors are nonnegative.

- We see that that minor $\boldsymbol{M_{11}}$ is always positive. However, we can easily find a point for which $\boldsymbol{M_{22}}$ is negative. For instance for the point ***(1,4)*** $\boldsymbol{M_{22}=-399}$.

- As a result, we can tell the banana function is not convex.

Why are convex functions so cool?

- First, we saw that the local minimum of a convex function is a global minimum. It is a pretty good result to help us find a solution more quickly.
- Moreover, in general, convex optimization problems are easier to solve. Why? To get a better idea let us look at some figures.

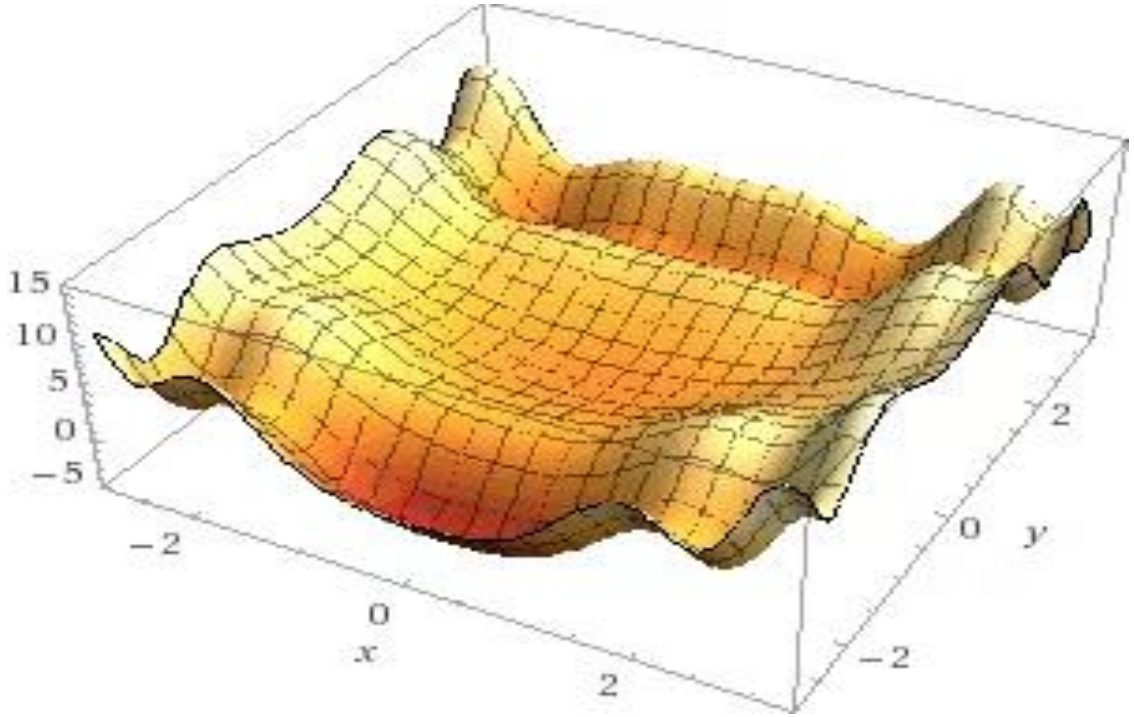


A convex surface

- Imagine that solving the optimization problem is like throwing a marble onto a surface. In the case of the convex surface, like the one in the figure above, no matter where you put the marble, it will go directly to the center of the bowl which is the minimum of the function.

Why are convex functions so cool?

- What if the surface is non-convex? Well as you can see throwing a marble randomly onto the surface has very few chances of hitting the global minimum. Instead, it is likely that the marble will fall into one of the many local minima. And when this is the case, what do you do? Do you try to push the marble to get somewhere else? As you can see, the problem is much more complicated.



A nonconvex surface

- The marble analogy is interesting because it is basically what does an optimization algorithm called gradient descent. Another way to solve an optimization problem is to use the well-known Newton's method.

Convex Functions

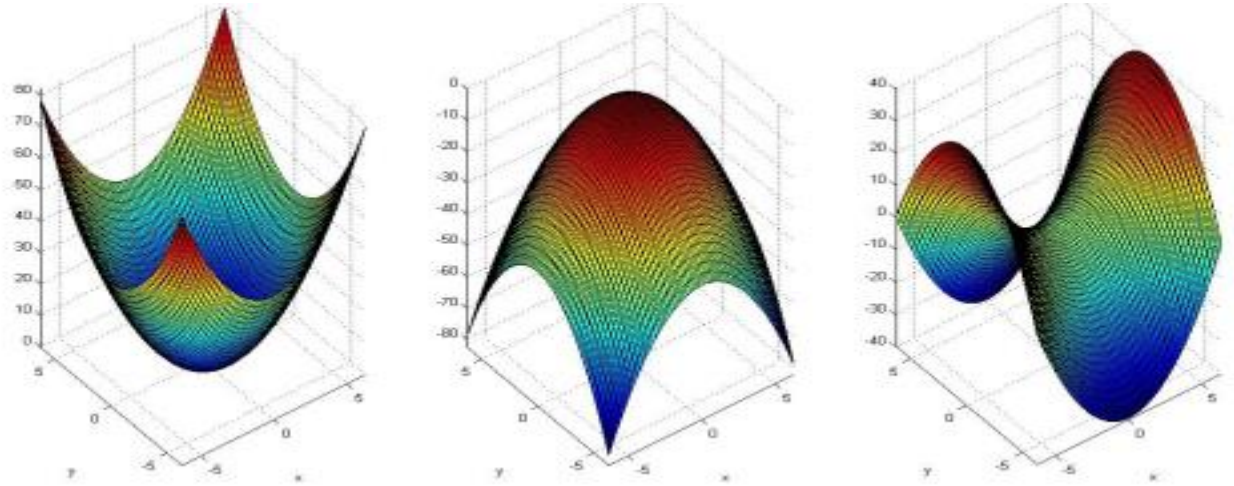
- **Why are convex functions so cool?**
- In this part, we learned what a convex set is and how to tell if a function is convex. Moreover, we saw a visual representation showing us why convex optimization is usually much simpler than non-convex optimization: because there are no local minima.

- Convexity is an important concept to understand when studying optimization.

Ex.

- Convex or Not:

1



2

$$f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2 + 2x_1 - 3x_2 + e^{x_1}.$$

3

$$f(x_1, x_2, x_3) = e^{x_1 - x_2 + x_3} + e^{2x_2} + x_1$$

4

$$f(x_1, x_2) = -\log(x_1x_2)$$

Q&A