# Introduction to Algorithms for Data Mining and Machine Learning

## Chapter I: Introduction to Optimization

Xin-She Yang

For details, please read the book:

Xin-She Yang, **Introduction to Algorithms for Data Mining and Machine Learning**, Academic Press/Elsevier, (2019).

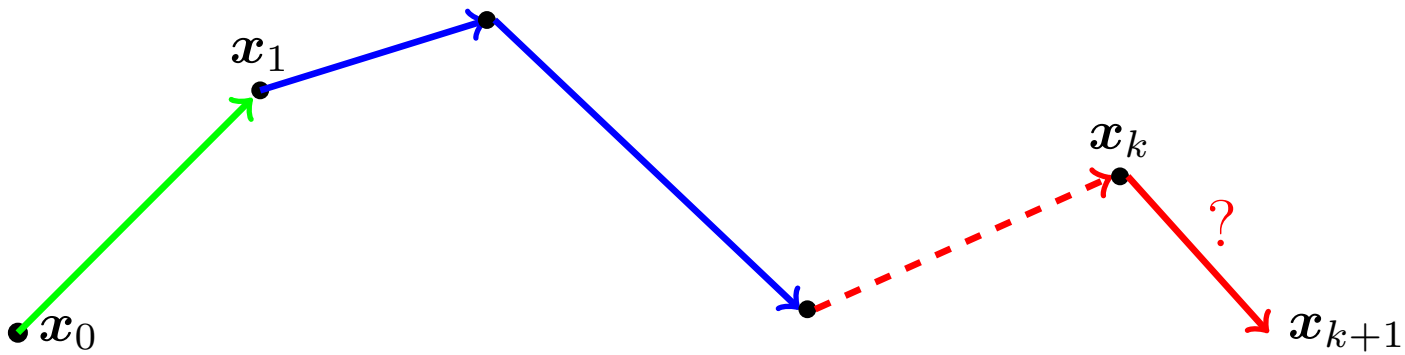
# Essence of an Algorithm

## Algorithm = Iterative Procedure

For a given problem (e.g., optimization), a solution is represented as a vector $\boldsymbol{x}$. To obtain the correct solution, we usually start with an initial guess $\boldsymbol{x}_0$ (an educated guess) and try to modify or improve the solution iteratively.

To generate a better solution point $\boldsymbol{x}_{k+1}$ from an existing solution $\boldsymbol{x}_k$ at iteration $k$, we have

$$\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k \ \ (\text{modification}).$$




## Different algorithms

Different ways for generating new solutions!

# Example: Finding the square root

To find the square of any positive number $a > 0$, we can use

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right), \quad (k = 0, 1, 2, ...), \quad x_0 = 1.$$

## Example: $a = 4$

$$\begin{aligned}
x_0 &= 1, \quad [\text{an educated guess}] \\
x_1 &= \frac{1}{2}(1 + \frac{4}{1}) = 2.5, \\
x_2 &= \frac{1}{2}(2.5 + \frac{4}{2.5}) = 2.05, \\
x_3 &= \frac{1}{2}(2.05 + \frac{4}{2.05}) \approx 2.0061, \\
x_4 &= \frac{1}{2}(2.0061 + \frac{4}{2.0061}) \approx 2.00000927,
\end{aligned}$$

This is very close to the true value of $\sqrt{4} = 2$.

As we know that $\sqrt{4} = \pm 2$, so how do we get the $-2$ root?

## Example: $a = 4$

Starting with $x_0 = -1$, we have

$$x_1 = \frac{1}{2}(-1 + \frac{4}{-1}) = -2.5,$$
$$x_2 = \frac{1}{2}(-2.5 + \frac{4}{-2.5}) = -2.05,$$
$$x_3 = \frac{1}{2}(-2.05 + \frac{4}{-2.05}) \approx -2.0061,$$
$$x_4 = \frac{1}{2}(-2.0061 + \frac{4}{-2.0061}) \approx -2.00000927.$$

## Where to start?

If we choose $x_0 = 0$, the algorithm will simply fail (due to division by zero).

This highlights an issue that the final solution can depend on the initial starting point. In fact, such dependence exists in almost all local search algorithms.

Exercise: `Write a short R code to implement this algorithm.`

# Optimization: An Example

A simple optimization problem is to find the maximum or minimum of a (real) function $f(x)$. For example, we know that $f(x) = x^2$ has the minimum value $f_{\min} = 0$ at $x_* = 0$ because the square of any real number is non-negative. Here we use $x_*$ with a star $*$ to highlight the fact that it is a special point, not any point.

$$\text{Minimize } f(x) = x^2, \quad x \in \mathbb{R},$$

which has a global minimum $f_{\min}(x_*) = 0$ at $x_* = 0$. We can even guess it.

# Optimization: An Example

A simple optimization problem is to find the maximum or minimum of a (real) function $f(x)$. For example, we know that $f(x) = x^2$ has the minimum value $f_{\min} = 0$ at $x_* = 0$ because the square of any real number is non-negative. Here we use $x_*$ with a star $*$ to highlight the fact that it is a special point, not any point.

$$\text{Minimize } f(x) = x^2, \quad x \in \mathbb{R},$$

which has a global minimum $f_{\min}(x_*) = 0$ at $x_* = 0$. We can even guess it.

**Any function $f(x)$?**

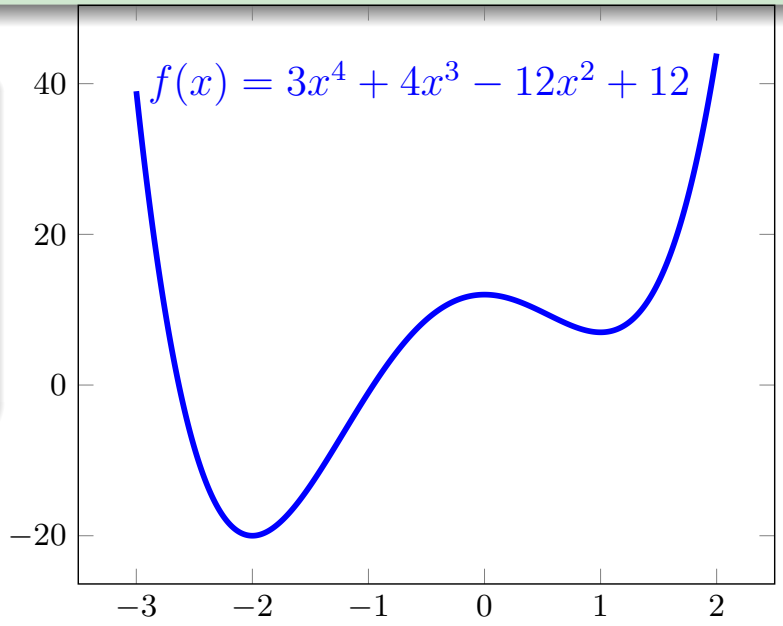
Find the maxima and minima of

$$f(x) = 3x^4 + 4x^3 - 12x^2 + 12$$

where $x \in \mathbb{R}$.

The global minimum at $x_* = -2$?

Its global maximum is $+\infty$ at $\pm\infty$.
How do we know?

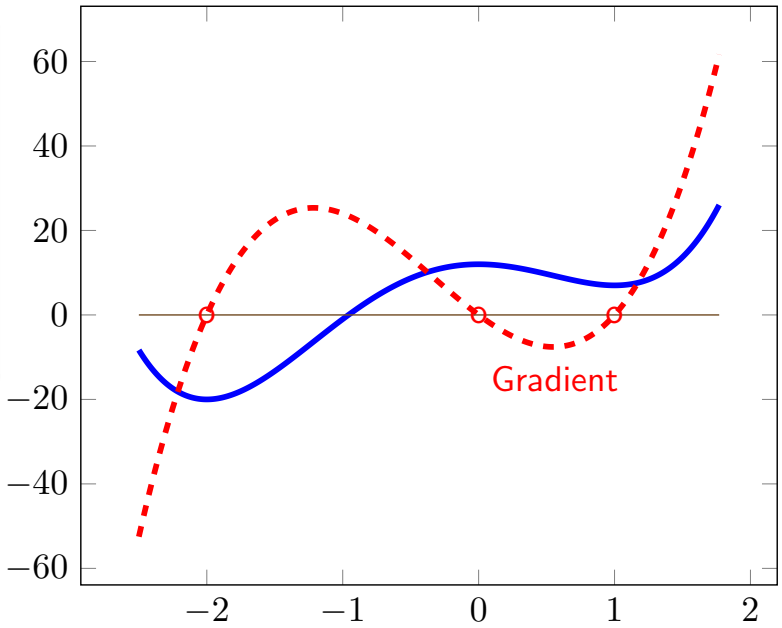


## Any function $f(x)$?

Find the maxima and minima of

$$f(x) = 3x^4 + 4x^3 - 12x^2 + 12$$

where $x \in \mathbb{R}$.




The gradient of $f(x)$ is

$$f'(x) = 12x^3 + 12x^2 - 24x.$$

We know $f'(x) = 0$ or

$$f'(x) = 12x^3 + 12x^2 - 24x = 12x(x-1)(x+2) = 0$$

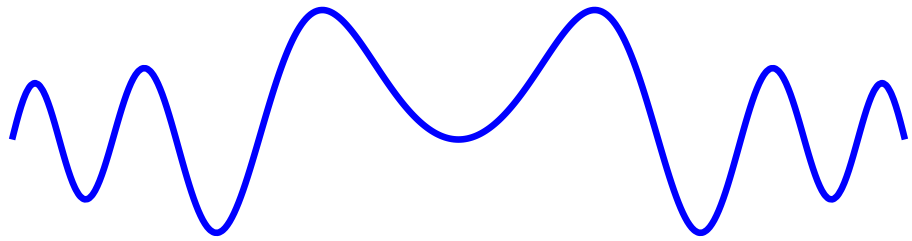
has three solutions $x = 0$, $+1$ and $-2$.

By going through all the three points, we know $f(-2) = -20$ is the minimum since it curves up at that point, as seen on the graph. More specifically, $f''(x) = 36x^2 + 24x - 24$, which gives $f''(-2) = 72 > 0$ (so a minimum).

Both $f'(x)$ and $f''(x)$ are needed to determine whether a given point is a maximum or minimum. However, this does not apply to the case when $x_* = \pm\infty$.

# Optimality

In general, for a simple univariate function $f(x)$ where $x \in \mathbb{R}$, its maximum or minima occurs at $f'(x) = 0$.



For example, there are 6 maxima and 5 minima, shown in the graph.

## Optimality Condition (for univariate functions)

$$\text{Minimize or Maximize} \quad f(x), \quad x \in \mathbb{R}.$$

The critical condition $f'(x_*) = 0$ determines the locations $(x_*)$ of optimality.

$$\begin{cases} f''(x_*) > 0 & \text{local minimum} \\ f''(x_*) < 0 & \text{local maximum} \end{cases}$$

Exercise: `Using R to visualize the univariate function` $f'(x) = 12x^3 + 12x^2 - 24x$ `in the domain [-10,+10] and find all its optima.`

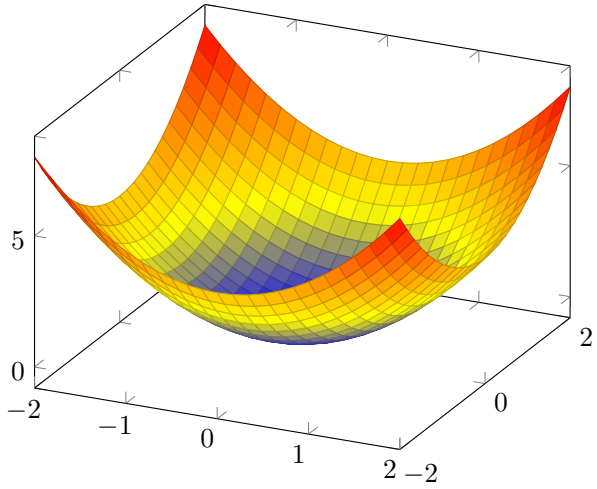
# Functions with two variables

The following function of two variables

$$f(x, y) = x^2 + y^2, \quad x, y \in \mathbb{R}$$

has a global minimum $f_{\min} = 0$ at (0,0) because both the squares of any real numbers ($x$ and $y$) are non-negative.

How do we solve it formally?

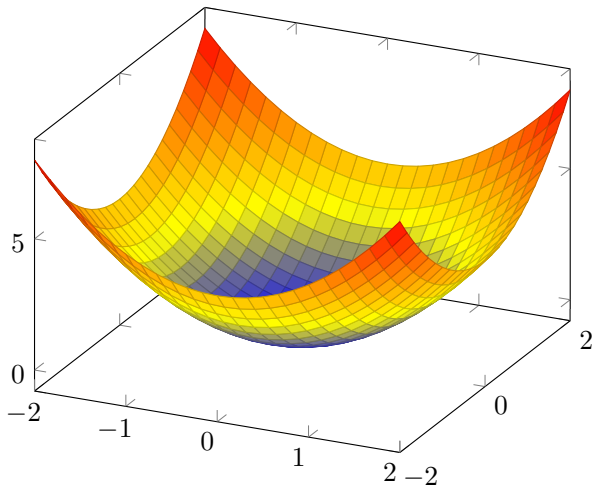


# Functions with two variables

The following function of two variables

$$f(x, y) = x^2 + y^2, \quad x, y \in \mathbb{R}$$

has a global minimum $f_{\min} = 0$ at (0,0) because both the squares of any real numbers ($x$ and $y$) are non-negative.




How do we solve it formally?

## Optimality

There are two first partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, thus the stationary conditions mean that

$$\frac{\partial f}{\partial x} = 2x = 0, \quad \frac{\partial f}{\partial y} = 2y = 0,$$

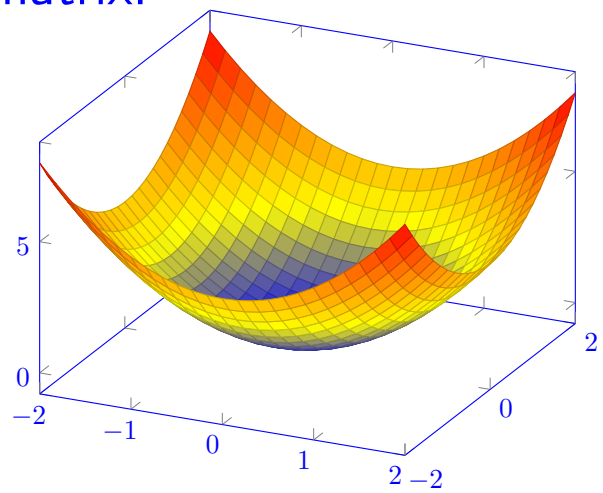
which can be written compactly as a gradient vector $\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}) = 0$.

There are four second-order derivatives $f_{xx} = \frac{\partial^2 f}{\partial x^2}$, $f_{yy} = \frac{\partial^2 f}{\partial y^2}$, $f_{xy} = \frac{\partial^2 f}{\partial x \partial y}$, and $f_{yx} = \frac{\partial^2 f}{\partial y \partial x}$. We have to use them to form a so-called Hessian matrix.

# Multivariate functions

In general, for an $n$-dimensional multivariate function $f(\boldsymbol{x})$ with $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, the optima occur at $\nabla f(\boldsymbol{x}) = 0$. The second-derivative criterion becomes the definiteness of the Hessian matrix:

$$\boldsymbol{H} = \nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{pmatrix}.$$




## Previous Example Revisited

For $f(\boldsymbol{x}) = x^2 + y^2$, its gradient $\nabla f = (2x,\ 2y) = 0$ gives $\boldsymbol{x}_* = (0, 0)$. Its Hessian matrix is

$$\boldsymbol{H} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

which is positive definite. This means $f$ is convex and the point (0,0) corresponds to a minimum (in fact, the global minimum).

Let us try a more complicated function

$$f(x,y) = (x-1)^2 + x^2 y^2, \quad x, y \in \mathbb{R}.$$

Its stationary conditions are

$$\frac{\partial f}{\partial x} = 2(x-1) + 2xy^2 = 0, \quad \frac{\partial f}{\partial y} = 0 + 2x^2 y = 0.$$

From $2x^2 y = 0$, we have either $y = 0$ or $x = 0$.

- Substituting $y = 0$ into the first condition, we have $x = 1$, which is a solution or a stationary point $x_* = 1$ and $y_* = 0$.
- The other condition $x = 0$ does not satisfy the first condition.

The Hessian matrix is

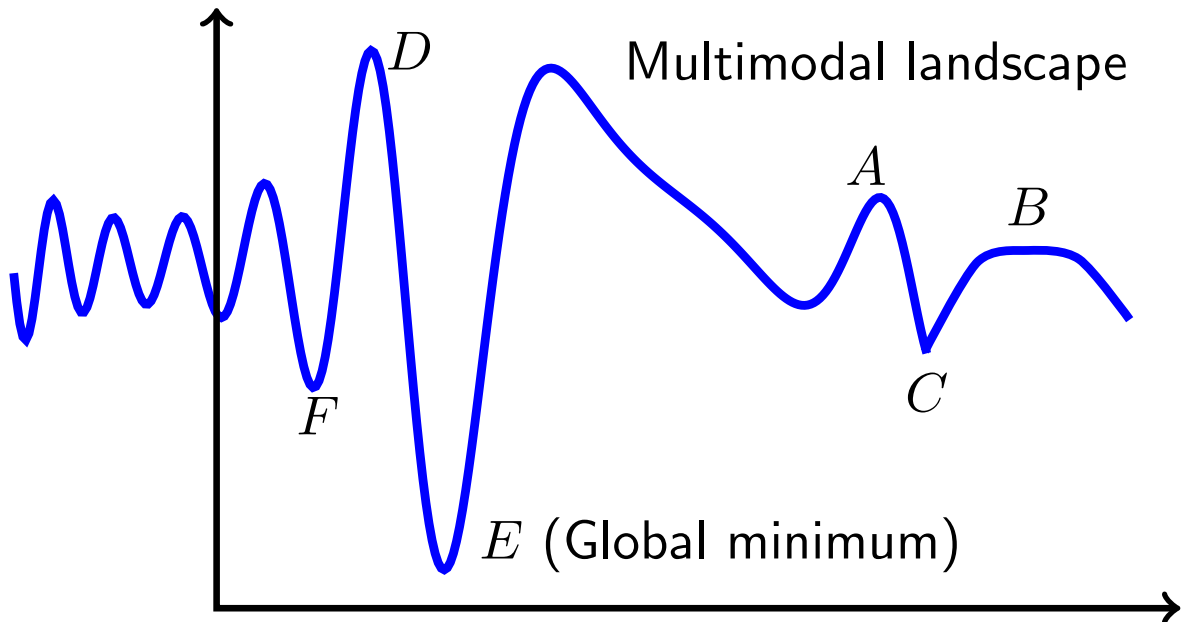
$$\boldsymbol{H} = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial x^2} \end{pmatrix} = \begin{pmatrix} 2y^2 + 2 & 4xy \\ 4xy & 2x^2 \end{pmatrix}.$$

At the stationary point, we have $x_* = 1$ and $y_* = 0$ and the Hessian matrix becomes

$$\boldsymbol{H} = \begin{pmatrix} 2 \times 0^2 + 2 & 4 \times 1 \times 0 \\ 4 \times 1 \times 0 & 2 \times 1^2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

which is positive definite. Therefore, (1,0) corresponds to a minimum with $f_{\min} = 0$.

# Optimality




## Strong and Weak Optimality

- Point $A$ is a strong local maximum, whereas point $B$ within a small flat region (potentially many $x_*$) is a weak local maximum.
- Point $D$ is the global maximum, and point $E$ is the global minimum.
- Point $F$ is a strong local minimum.

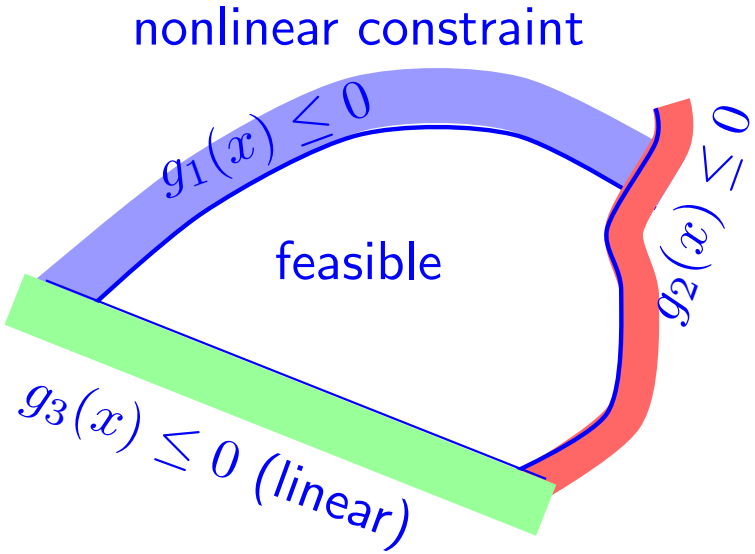
# Constrained Optimization

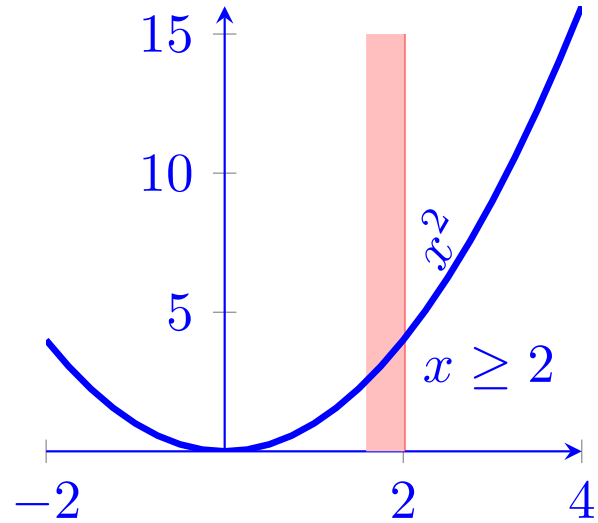
The objective is to minimize (or maximize)

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

subject to

$$h_i(\boldsymbol{x}) = 0, \quad (i = 1, 2, ..., M), \qquad \text{[equality constraints]}$$
$$g_j(\boldsymbol{x}) \le 0, \quad (j = 1, 2, ..., N). \qquad \text{[inequality constraints]}$$




Feasible domain with nonlinear inequality constraints $g_1(x)$ and $g_2(x)$ as well as a linear inequality $g_3(x)$ (left). Minimization of $f(x) = x^2$ subject to $x \ge 2$ (right).

# Penalty Method – Dealing with constraints

For a constrained optimization problem

$$\min f(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, x_2, ..., x_n)^T \in \mathbb{R}^n,$$

subject to

$$\phi_i(\boldsymbol{x}) = 0, \ (i = 1, 2, ..., M), \quad \psi_j(\boldsymbol{x}) \le 0, \ \ (j = 1, 2, ..., N),$$

the main idea of the penalty method is to use a penalty function to transform the unconstrained optimization problem into an unconstrained one with a modified objective

$$\Pi(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{M} \mu_i \phi_i^2(\boldsymbol{x}) + \sum_{j=1}^{N} \nu_j \max\left\{0, \psi_j(\boldsymbol{x})\right\}^2,$$

where $\mu_i \ge 0$ and $\nu_j \ge 0$ are penalty parameters to be given.

Then, we can solve this new unconstrained problem using optimality conditions such as $\nabla \Pi = 0$, though the algebraic calculations may be more tedious.

Let us look at a simple example

$$\min f(x) = 40(x-1)^2, \quad x \in \mathbb{R},$$

subject to a single inequality (with a given value of $a$)

$$g(x) = x - a \geq 0.$$

**Observations**

- Without any constraint, the minimum $f_{\min} = 0$ occurs $x = 1$.
- Even with this constraint, if $a \leq 1$, the result is not affected.
- If $a > 1$, the minimum should occur at the boundary $x_* = a$ with $f_{\min} = 40(a-1)^2$.

Using a penalty parameter $\mu$, we have

$$\Pi = f(x) + \mu[g(x)]^2 = 40(x-1)^2 + \mu(x-a)^2,$$

From $\Pi'(x) = 0$, we have

$$\Pi'(x) = 80(x-1) - 2\mu(x-a) = 0, \quad \text{or} \quad x_* = \frac{40 - \mu a}{40 - \mu},$$

which depends on $\mu$.

It seems that the optimal solution $x_*$ depends on the penalty parameter $\mu$, which highlights a potential problem for the penalty method (the choice of $\mu$).

**$a = 5$**

- For $\mu = 1$, we have $x_*(\mu) = \frac{(40-\mu a)}{(40-\mu)} = \frac{40-1\times 5}{40-1} \approx 0.897$.
- Similarly, for $\mu = 100$, we have $x_* \approx 7.66667$.
- For $\mu = 10000$, we have $x_* \approx 5.01606$.
- For $\mu = 1000000$, we have $x_* \approx 5.00016$.

The true optimal solution $x_* = 5$ can be approximated if $\mu$ is sufficiently large.

In fact, if $\mu \gg 1$ (very, very large), we have

$$x_* = \frac{40 - \mu a}{40 - \mu} \approx \frac{-\mu a}{-\mu} \approx a,$$

which leads to the correct optimal solution.

Now the question is "How large is enough enough?"
The answer seems to be "It depends" (the desired quality of the solution).

However, a very large $\mu$ may cause numerical problems.

# Method of Lagrange Multipliers

The method of Lagrange multipliers can deal with equality constraints effectively (without any undetermined parameters).

## Equality Constraints

For an objective $f(\boldsymbol{x})$ subject to an equality constraint $h(\boldsymbol{x}) = 0$, the idea is to use a Lagrange multiplier $\lambda$ to form a Lagrangian function

$$L = f(\boldsymbol{x}) + \lambda h(\boldsymbol{x}),$$

which converts the original constrained problem into an unconstrained one. Then, we can treat $\lambda$ as the extra parameter/variable and solve this new problem as usual (in terms of stationary conditions).

In case of multiple equalities $h_j(j = 1, 2, ..., M)$, we can use $M$ multipliers $\lambda_j$ and we have

$$L = f(\boldsymbol{x}) + \sum_{j=1}^{M} \lambda_j h_j(\boldsymbol{x}).$$

For an $n$-dimensional problem with $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, we have

$$\frac{\partial L}{\partial x_i} = \frac{\partial f}{\partial x_i} + \sum_{j=1}^{M} \lambda_j \frac{\partial h_j}{\partial x_i} = 0, \quad (i = 1, ..., n), \tag{1}$$

$$\frac{\partial L}{\partial \lambda_j} = h_j = 0, \quad (j = 1, 2, ..., M), \text{ (The original equalities)} \tag{2}$$

which form $M + n$ equations.

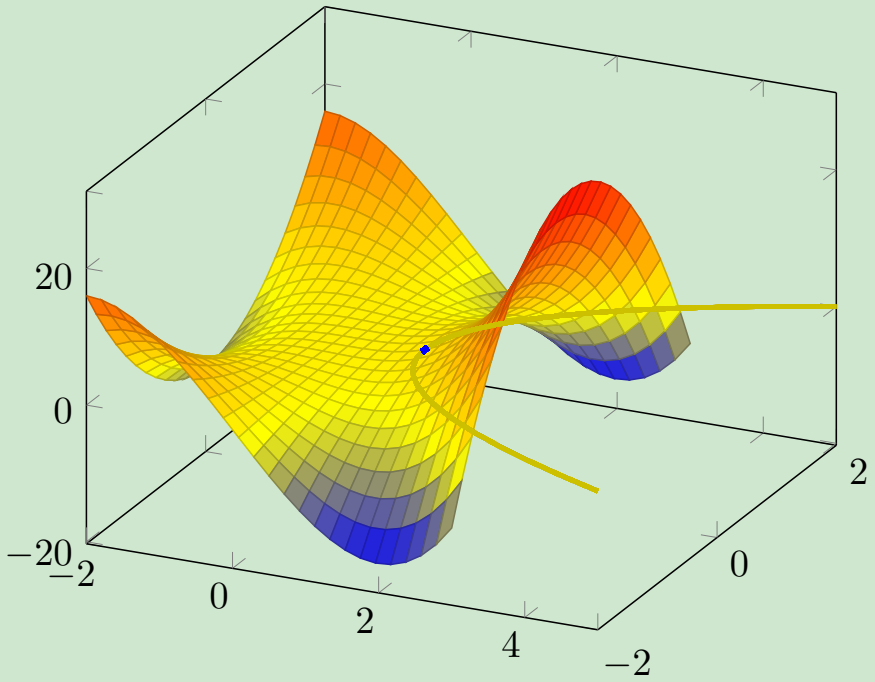
## Example

The Monkey surface defined by

$$f(x, y) = x^3 - 3xy^2$$

has no unique maximum or minimum. The point $x = y = 0$ is a saddle point.

With an additional equality $x - y^2 = 1$, it becomes a constrained optimization problem with a unique minimum (the blue dot).



From $f = x^3 - 3xy^2$ and $h = x - y^2 - 1$, we have

$$\Phi = f + \lambda h = x^3 - 3xy^2 + \lambda(x - y^2 - 1).$$

So we have

$$\frac{\partial \Phi}{\partial x} = 3x^2 - 3y^2 + \lambda = 0, \quad \frac{\partial \Phi}{\partial y} = 0 - 6xy + (-2\lambda y) = 0, \quad \frac{\partial \Phi}{\partial \lambda} = x - y^2 - 1 = 0.$$

The second condition $-6xy - 2\lambda y = -2y(3x + \lambda) = 0$ means either $y = 0$ or $\lambda = -3x$.

From $f = x^3 - 3xy^2$ and $h = x - y^2 - 1$, we have

$$\Phi = f + \lambda h = x^3 - 3xy^2 + \lambda(x - y^2 - 1).$$

So we have

$$\frac{\partial \Phi}{\partial x} = 3x^2 - 3y^2 + \lambda = 0, \quad \frac{\partial \Phi}{\partial y} = 0 - 6xy + (-2\lambda y) = 0, \quad \frac{\partial \Phi}{\partial \lambda} = x - y^2 - 1 = 0.$$

The second condition $-6xy - 2\lambda y = -2y(3x + \lambda) = 0$ means either $y = 0$ or $\lambda = -3x$.

- If $y = 0$, $x - y^2 - 1 = 0$ gives $x = 1$. Substituting it into the first condition, we have $\lambda = -3$. Thus, $f(1, 0) = 1$ is a minimum (not a maximum) (the blue dot).
- If $\lambda = -3x$, the first condition becomes $3x^2 - 3y^2 - 3x = 0$. Using the third condition $x = y^2 + 1$, we have

  $$3(y^2 + 1)^2 - 3y^2 - 3(y^2 + 1) = 0,$$

  or simply $3(y^4 + 2) = 0$, which is impossible to satisfy for any real numbers.

Therefore, the optimality occurs at (1,0) with $f_{\min} = 1$.

From $f = x^3 - 3xy^2$ and $h = x - y^2 - 1$, we have

$$\Phi = f + \lambda h = x^3 - 3xy^2 + \lambda(x - y^2 - 1).$$

So we have

$$\frac{\partial \Phi}{\partial x} = 3x^2 - 3y^2 + \lambda = 0, \quad \frac{\partial \Phi}{\partial y} = 0 - 6xy + (-2\lambda y) = 0, \quad \frac{\partial \Phi}{\partial \lambda} = x - y^2 - 1 = 0.$$

The second condition $-6xy - 2\lambda y = -2y(3x + \lambda) = 0$ means either $y = 0$ or $\lambda = -3x$.

- If $y = 0$, $x - y^2 - 1 = 0$ gives $x = 1$. Substituting it into the first condition, we have $\lambda = -3$. Thus, $f(1, 0) = 1$ is a minimum (not a maximum) (the blue dot).
- If $\lambda = -3x$, the first condition becomes $3x^2 - 3y^2 - 3x = 0$. Using the third condition $x = y^2 + 1$, we have

$$3(y^2 + 1)^2 - 3y^2 - 3(y^2 + 1) = 0,$$

or simply $3(y^4 + 2) = 0$, which is impossible to satisfy for any real numbers.

Therefore, the optimality occurs at (1,0) with $f_{\min} = 1$.

**Limitation**

The method of Lagrange multipliers work well for equality constraints. For inequality constraints, some extra slack variables are needed, which can lead to the so-called Karush-Kuhn-Tucker (KKT) conditions.

# KKT Conditions

## Unconstrained optimization problem

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

subject to $h_i(\boldsymbol{x}) = 0, \ \ (i = 1, 2, ..., M), \quad g_j(\boldsymbol{x}) \leq 0, \ \ (j = 1, 2, ..., N).$

## The KKT Conditions of this Problem

- Stationarity conditions: $\nabla f(\boldsymbol{x}) + \sum_{i=1}^{M} \lambda_i \nabla h_i(\boldsymbol{x}) + \sum_{j=1}^{N} \mu_j \nabla g_j(\boldsymbol{x}) = 0.$
- Primal feasibility: $h_i(\boldsymbol{x}) = 0, \quad g_j(\boldsymbol{x}) \leq 0, \quad (i = 1, 2, ..., M; j = 1, 2, ..., N).$
- Complementary slackness: $\mu_j g_j(\boldsymbol{x}) = 0, \quad (j = 1, 2, ..., N).$
- Dual feasibility: $\mu_j \geq 0, \quad (j = 1, 2, ..., N).$

# KKT Conditions

## Unconstrained optimization problem

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

subject to $h_i(\boldsymbol{x}) = 0, \ \ (i = 1, 2, ..., M), \quad g_j(\boldsymbol{x}) \leq 0, \ \ (j = 1, 2, ..., N).$

## The KKT Conditions of this Problem

- Stationarity conditions: $\nabla f(\boldsymbol{x}) + \sum_{i=1}^{M} \lambda_i \nabla h_i(\boldsymbol{x}) + \sum_{j=1}^{N} \mu_j \nabla g_j(\boldsymbol{x}) = 0.$
- Primal feasibility: $h_i(\boldsymbol{x}) = 0, \quad g_j(\boldsymbol{x}) \leq 0, \quad (i = 1, 2, ..., M; j = 1, 2, ..., N).$
- Complementary slackness: $\mu_j g_j(\boldsymbol{x}) = 0, \quad (j = 1, 2, ..., N).$
- Dual feasibility: $\mu_j \geq 0, \quad (j = 1, 2, ..., N).$

## Limitations

The KKT conditions are neat and useful to prove certain theorems concerning optimization. The proof of the KKT conditions is beyond this part of syllabus.

However, these neat conditions are not very useful to solve the optimization problem in practice.

# References

## References


- Xin-She Yang, **Introduction to Algorithms for Data Mining and Machine Learning**, Academic Press/Elsevier, (2019).
- Xin-She Yang, **Optimization Techniques and Applications with Examples**, John Wiley & Sons, (2018).


## Notes on Software

There are many different software packages for optimization, data mining and machine learning, including R, Python, C++, Java, Julia, and Matlab/Octave implementations.

For example, the following wikipedia links may be useful.

- List of Optimization Software
- Data Mining and Machine Learning Software
- Deep learning software

**Any questions?**

**Thank you :)**