Robotics: Science and Systems

Optimization I

Zhibin Li
School of Informatics
University of Edinburgh

Content

- Concept of optimization
- Unconstrained optimization
  - Least Square (LS) optimization
  - Tikhonov regularisation
- Gradient-based optimization
- Constrained optimization
  - Lagrange multipliers

Mathematical optimization

Minimization or maximization the value of a function ***f*** (also written as ***J***).

The function ***f*** is called an **objective function**. Usually, the **objective function** refers to the norm of errors, so we also use the name **cost function**, followed by an exchangeable notion of ***J*** as well.

Minimization or maximization depends on the formulation of ***f*** :

❑ Maximize rewards;
❑ Minimize penalties and errors.

Mathematical optimization

Here, we only introduce **continuous optimization**, where the variables are allowed to take on any value within a range of values, usually **real numbers**.

Out of scope:

1. Discrete optimization, where some or all of the variables may be binary, integer, or more abstract objects drawn from sets with finitely many elements, eg mixed-integer programming (MIP).
2. Nonlinear optimization.

Unconstrained optimization

Unconstrained optimization

Unconstrained optimization considers the problem of minimizing an objective function $\boldsymbol{f}$ that depends on real variables with no restrictions on their values.

Mathematically, let $\mathbf{x} \in \mathbf{R}^n$ be a n-dimensional vector ($n \geq 1$); and let $\boldsymbol{f}(\mathbf{R}^n) \rightarrow \mathbf{R}$ be a smooth function.

Unconstrained optimization problem is simply to find $\mathbf{x}$ such that

$$\arg\min \boldsymbol{f}(\mathbf{x}).$$

Often it is also practical to replace the constraints of an optimization problem with penalized terms in the objective function and to solve the problem faster as an unconstrained problem (see “Tikhonov regularization” later).

Least Square (LS) optimization

Suppose we want to find the value of **x** that minimizes

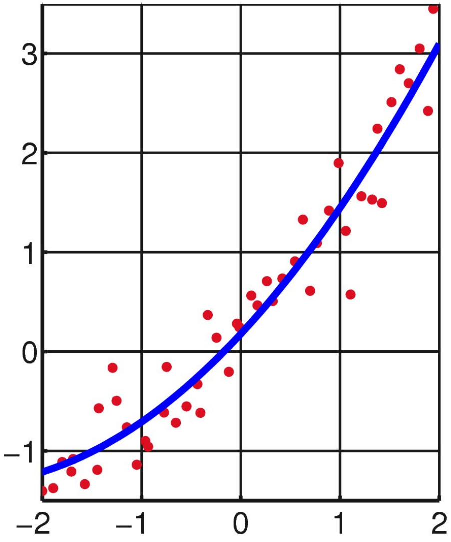
$$\boldsymbol{f}(\mathbf{x}) = ||\ \mathbf{Ax} - \mathbf{b}\ ||^2$$

Let **Ax**=y, a more intuitive form of this minimization is

$$\boldsymbol{f}(\mathbf{x}) = \sum (y_i - b_i)^2$$

How to interpret this?

Recall the discussion in system identification



Least Square (LS) optimization

So, given

$$\boldsymbol{f}(\mathbf{x}) = ||\ \mathbf{Ax} - \mathbf{b}\ ||^2$$

For minimizing $\boldsymbol{f}(\mathbf{x})$, ideally we would like to have

$$\mathbf{Ax} = \mathbf{b},\ (\boldsymbol{f}(\mathbf{x}) = 0)$$

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{b}$$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Least Square (LS) optimization

In order to have $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ (eg $\mathbf{A}^{-1}$) to be valid, there are several assumptions (recap of linear algebra):

1. **A** is a *n* x *n* square matrix
2. **A** has full rank, ie rank(**A**) = n
3. **A** is nonsingular (its determinant is *Not* 0, det(**A**)≠0)

Why pseudo inverse? $\mathbf{A}^{+}$ or $\mathbf{A}^{\#}$?

However, in most cases, the assumption of **A** is **not** valid. Usually, **A** is a m x n matrix, ie $\mathbf{A} \in M(m,n)$, and m > n. **A** is a tall rectangular matrix.

Least Square (LS) optimization

So, given

$$\boldsymbol{f}(\mathbf{x}) = ||\ \mathbf{Ax} - \mathbf{b}\ ||^2$$

we have specialized linear algebra algorithms that can solve $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ problem efficiently by pseudo inverse $\mathbf{x} = \mathbf{A}^{+}\mathbf{b}$**.**

Moore-Penrose pseudoinverse

Pseudoinverse is to compute a 'best fit' (least squares) solution to a system of linear equations that have multiple solutions.

A pseudoinverse $\mathbf{A}^+$ of a matrix **A** is a generalization of its inverse matrix $\mathbf{A}^{-1}$. The Moore-Penrose pseudoinverse is the most widely known pseudoinverse.

Left pseudoinverse (to be multiplied on the left side):

$$\mathbf{A}^+ = (\mathbf{A}^\mathrm{T} \mathbf{A})^{-1} \mathbf{A}^\mathrm{T}$$

Right pseudoinverse (to be multiplied on the right side):

$$\mathbf{A}^+ = \mathbf{A} (\mathbf{A} \mathbf{A}^\mathrm{T})^{-1}$$

Least Square (LS) optimization

So, the smallest-norm solution to the unconstrained least squares problem

$$\boldsymbol{f}(\mathbf{x}) = || \mathbf{A}\mathbf{x} - \mathbf{b} ||^2$$

is:

$$\mathbf{x} = \mathbf{A}^{+}\mathbf{b}$$

where $\mathbf{A}^{+}$ is the left pseudoinverse

$$\mathbf{x} = [\ (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}] \cdot \mathbf{b}$$

Weighted least square optimization

Previous LS solution assumes that each element in data **b** is equally treated. However, some data may not be equally reliable, eg recent data is more relevant than the past data/history.

To improve the fitting, weighted least-squares can be used, and the weight is used in the fitting process.

Similarly, Let **Ax**=y, weighted least-squares is formulated as

$$\boldsymbol{f}(\mathbf{x}) = \sum w_i(y_i - b_i)^2$$

to minimize the weighted sum of squares. w_i is the weight for each element in vector **b**.

Weighted least square optimization

Define each variance as σ^2_i, as the weight as the inverse of the variance $w_i=1/\sigma^2_i$, then the diagonal weight matrix **W** is:

$$\mathbf{W}=[\,\begin{array}{llll} w_1 & 0 & \ldots & 0; \\ 0 & w_2 & \ldots 0; & \\ \vdots & 0 & \ddots & \\ 0 & \ldots 0 & w_n & \end{array}].$$

So, the solution of weighted least square estimate is:

$$\mathbf{x} = [(\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}]\cdot\mathbf{b}$$

Weighted least square optimization

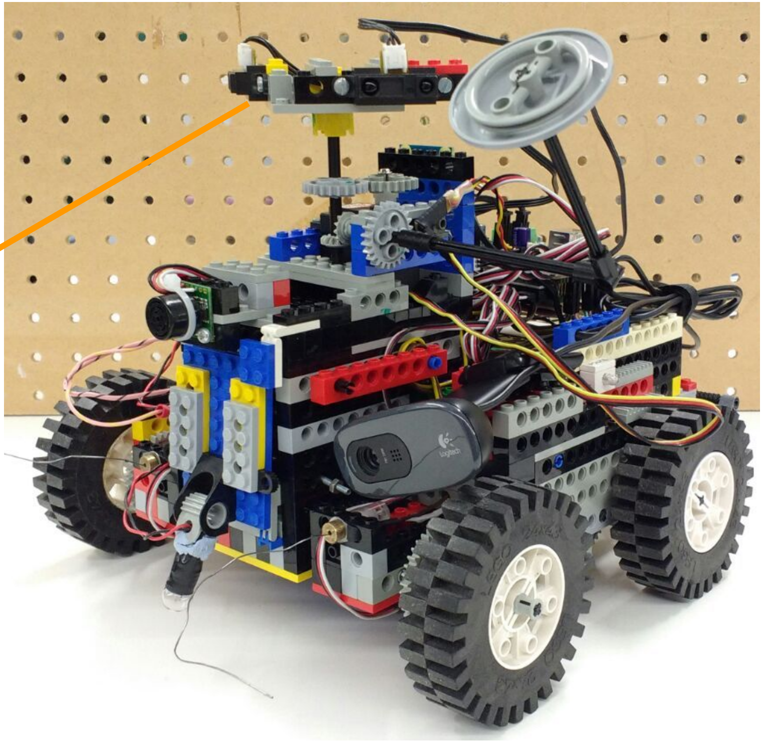
From the solution

$$\mathbf{x} = [(\mathbf{A}^{\mathsf{T}}\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{W}] \cdot \mathbf{b}$$

we can see that ordinary LS is a simplified version where **W**=**I**.

Application in RSS practical

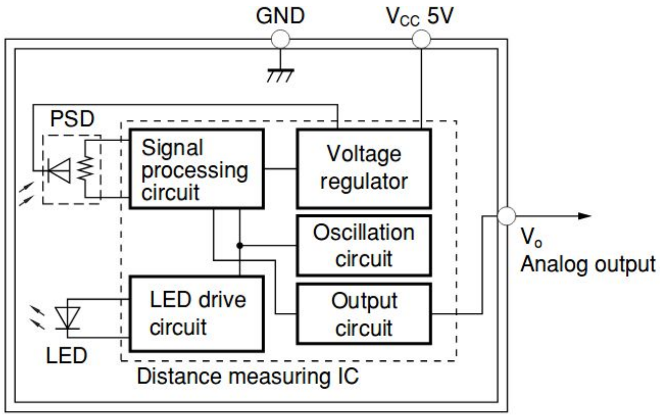
More accurate estimation of distance from sensor readings.

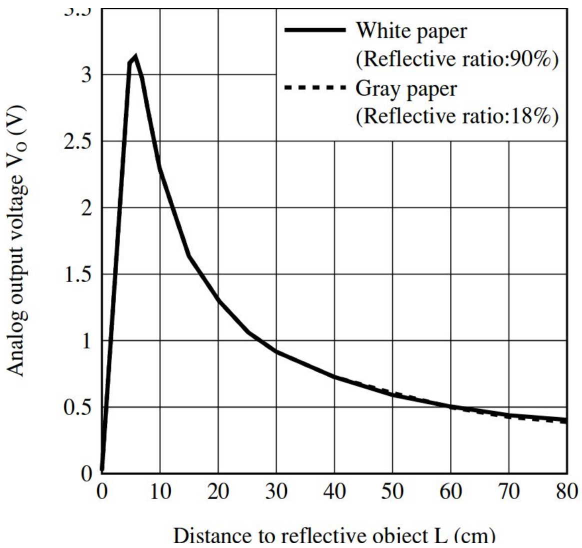


Application in RSS practical

Strength of reflection is inverse proportional to the distance.

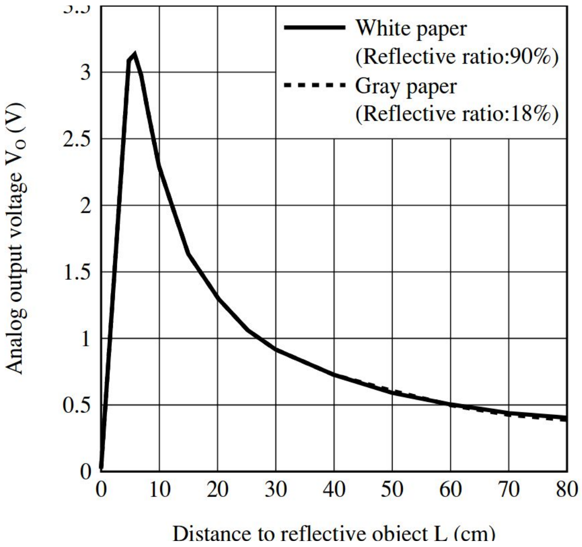
Voltage is *inverse proportional* to the distance.

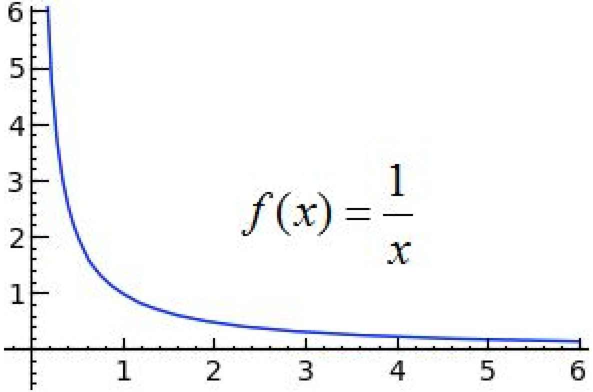


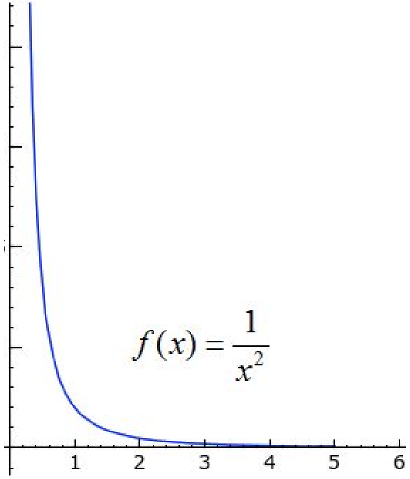


Correlating distance and the voltage

Voltage is *inverse proportional* to the distance. Some basic functions that reflect the inverse proportional relation.







Correlating distance and the voltage

Define y as the distance and x as the voltage, recall that voltage is *inverse proportional* to the distance, use polynomial to correlate the relation as:

$$y = a_0 + a_1 x^{-1} + a_2 x^{-2} + a_3 x^{-3} + a_4 x_1^{-4}$$

Stack data from multiple measurements into matrix form as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} 1, x_1^{-1}, x_1^{-2}, x_1^{-3}, x_1^{-4} \\ 1, x_2^{-1}, x_2^{-2}, x_2^{-3}, x_2^{-4} \\ \vdots \\ 1, x_k^{-1}, x_k^{-2}, x_k^{-3}, x_k^{-4} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}$$

Correlating distance and the voltage

$$\underbrace{\begin{bmatrix} 1, x_1^{-1}, x_1^{-2}, x_1^{-3}, x_1^{-4} \\ 1, x_2^{-1}, x_2^{-2}, x_2^{-3}, x_2^{-4} \\ \vdots \\ 1, x_k^{-1}, x_k^{-2}, x_k^{-3}, x_k^{-4} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}}_{b}$$

Solution of pseudo inverse

$$x = A^{\dagger} b$$

Correlating distance and the voltage

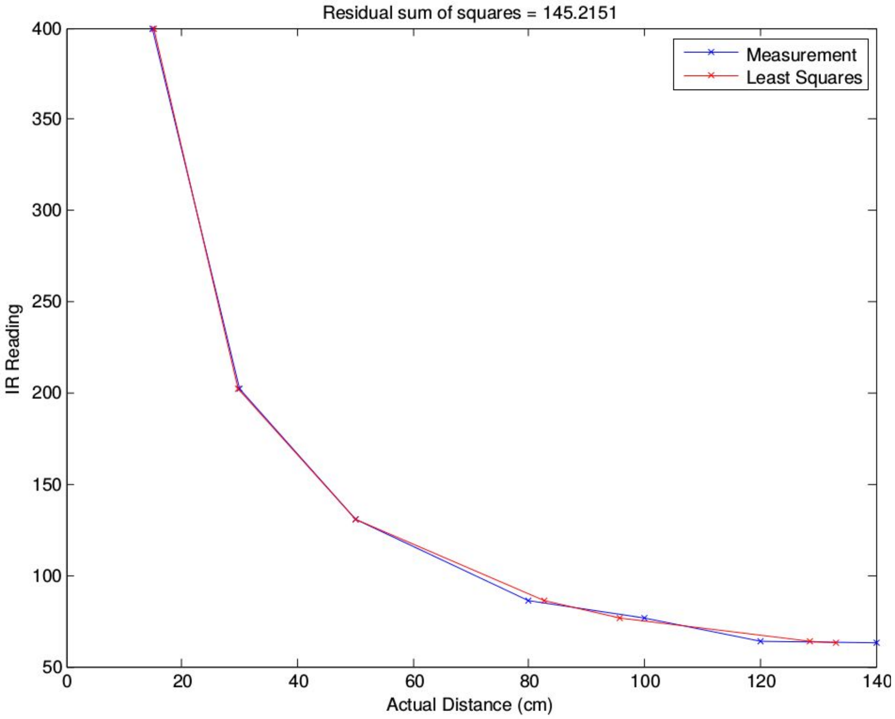
Solution of pseudo inverse:

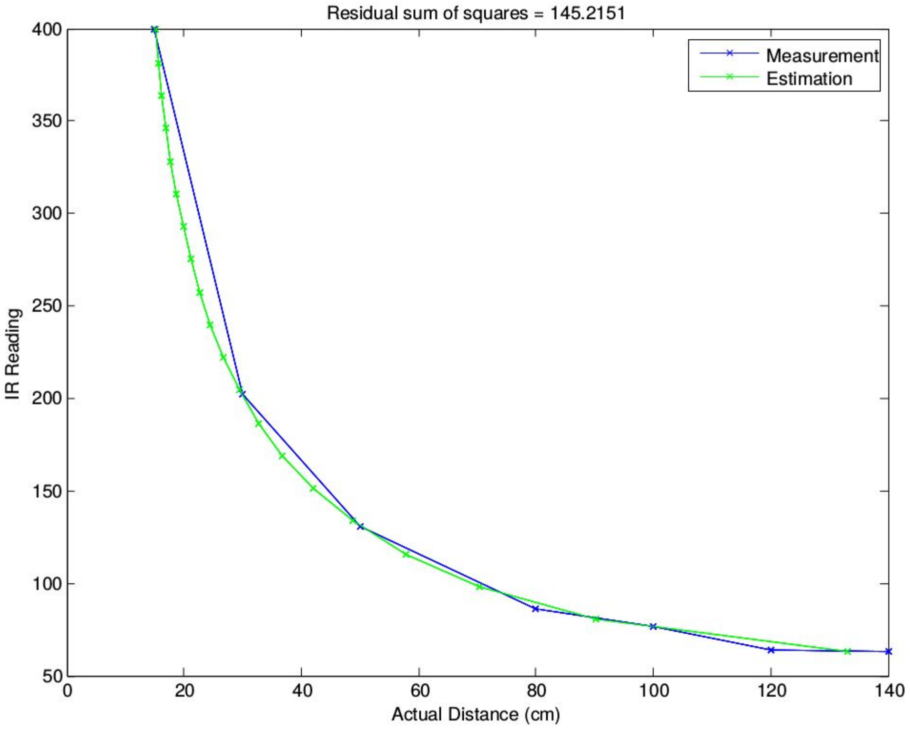
$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} 1, x_1^{-1}, x_1^{-2}, x_1^{-3}, x_1^{-4} \\ 1, x_2^{-1}, x_2^{-2}, x_2^{-3}, x_2^{-4} \\ \vdots \\ 1, x_k^{-1}, x_k^{-2}, x_k^{-3}, x_k^{-4} \end{bmatrix}^{\dagger} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}$$

After obtaining coefficients, we can apply equation below for any new measured voltage to calculate the distance

$$y = a_0 + a_1 x^{-1} + a_2 x^{-2} + a_3 x^{-3} + a_4 x_1^{-4}$$

Application in RSS practical: results





Regularized least squares

Regularized least squares (RLS) is a family of methods for solving the least-squares problem while using regularization to further constrain the resulting solution.

Cases where RLS is needed:

1. Number of measurements is less than the number of variables in LS (not enough data). Then, the least-squares problem is ill-posed, matrix **A** is a flat rectangular matrix, and it becomes impossible to fit. RLS introduces further constraints that uniquely determine the solution.
2. The learned model represented by the vector of model parameters **x** suffers from poor generalization. RLS can add some “*prior knowledge*” about the problem, eg initial guess of the solution.

Tikhonov regularisation

Tikhonov regularization solves problems in the form of

$$|| \mathbf{A}\ \mathbf{x} - \mathbf{b}\ ||_{\mathbf{P}}^{2} + ||\ \mathbf{x} - \mathbf{x}_0||_{\mathbf{Q}}^{2},$$

where $//\ \mathbf{x}\ //_{Q}^{2}$ is the weighted norm $\mathbf{x}^{T}\mathbf{Q}\mathbf{x}$. Similarly, $||\ \mathbf{A}\ \mathbf{x} - \mathbf{b}\ ||_{\mathbf{P}}^{2}$ stands for $(\mathbf{A}\ \mathbf{x} - \mathbf{b})^{T}\mathbf{P}(\mathbf{A}\ \mathbf{x} - \mathbf{b})$.

This formulation particularly has a closed form solution:

$$\mathbf{x} = (\mathbf{A}^{T}\mathbf{P}\mathbf{A} + \mathbf{Q})^{-1}(\ \mathbf{A}^{T}\mathbf{P}\mathbf{b} + \mathbf{Q}\mathbf{x}_0).$$

or equivalently

$$\mathbf{x} = \mathbf{x}_0 + (\mathbf{A}^{T}\mathbf{P}\mathbf{A} + \mathbf{Q})^{-1}\ [\mathbf{A}^{T}\mathbf{P}(\mathbf{b} - \mathbf{A}\mathbf{x}_0)].$$

Tikhonov regularisation

Solution of Tikhonov regularization

$$\mathbf{x} = \boxed{\mathbf{x}_0} + \boxed{(\mathbf{A}^T\mathbf{P}\mathbf{A} + \mathbf{Q})^{-1} [\mathbf{A}^T\mathbf{P}(\mathbf{b} - \mathbf{A}\mathbf{x}_0)]}.$$

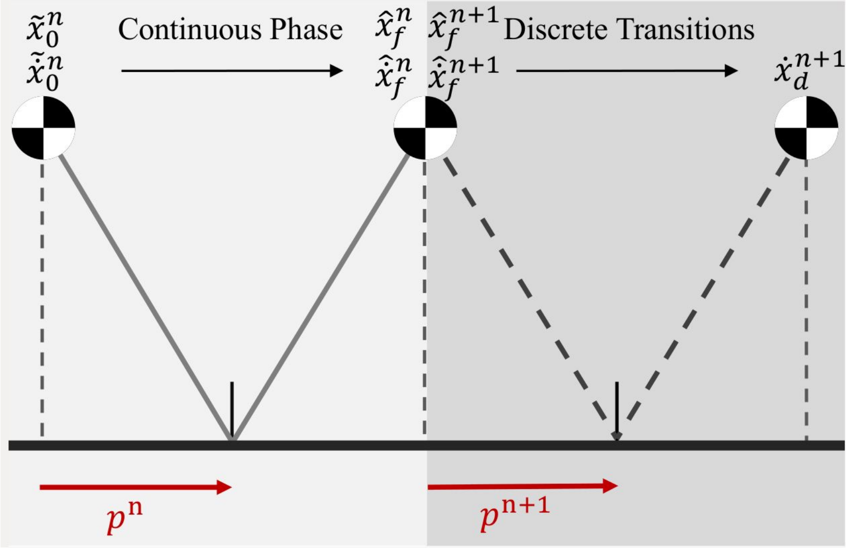
It is a penalized LS problem with weighting matrices (**P**, **Q**) and initial values $\mathbf{x}_0$.

Sometimes, due to noises and stochastic measurements, a true minimum of $||\mathbf{A}\mathbf{x}-\mathbf{b}\ ||^2$ is given by a particular **x** that can be very far away from the nominal $\mathbf{x}_0$. This is dangerous if **x** contains control actions, eg the angle of flap that is far away from operational range.

Suitable case, a fine tuning with weighted norm of deviation around the initial guess $\mathbf{x}_0$, find an optimal solution **x** around a nominal $\mathbf{x}_0$.

Application of Tikhonov regularisation

Example: dynamic walking control



Collect data (k=6)

$$\mathbf{p} = \begin{bmatrix} \tilde{p}^{n-k} \\ \vdots \\ \tilde{p}^{n-1} \end{bmatrix}_{k\times 1}, \mathbf{X}_2 = \begin{bmatrix} \tilde{\dot{x}}_f^{n-k} & \tilde{\dot{x}}_f^{n-k+1} & 1 \\ \vdots & \vdots & \vdots \\ \tilde{\dot{x}}_f^{n-1} & \tilde{\dot{x}}_f^{n} & 1 \end{bmatrix}_{k\times 3}$$

$$\min_{\boldsymbol{\beta}} \|\mathbf{X}_2\boldsymbol{\beta} - \mathbf{p}\|_{\mathbf{P}_2}^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_{\mathbf{Q}_2}^2$$

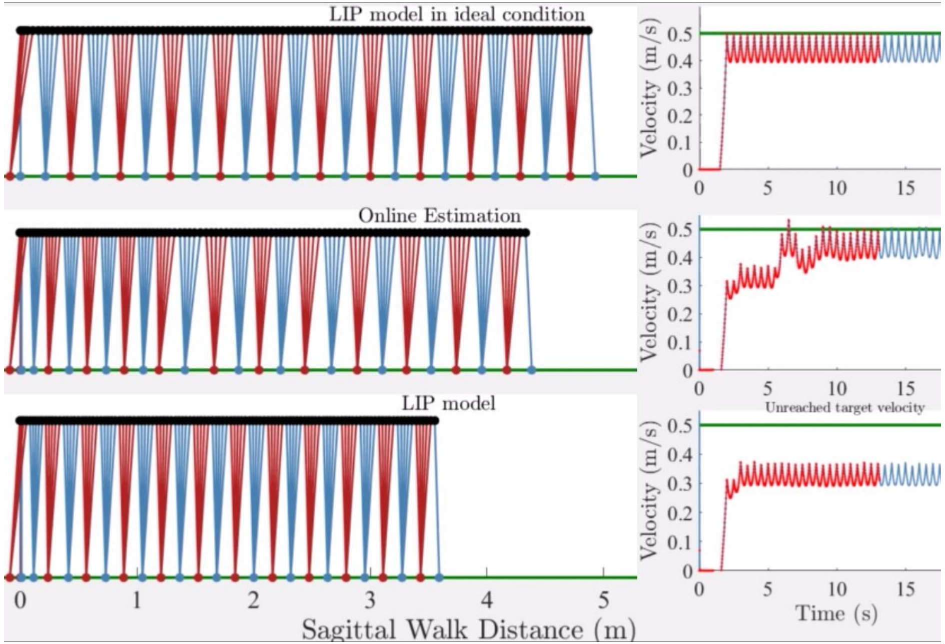
$\boldsymbol{P}_2$ = diag(0.1, 0.2, 0.3, 0.4, 0.5, 0.6),
$\boldsymbol{Q}_2$ =diag(0.1, 0.1, 0.001)

$$\Downarrow$$

$$p^{n+1} = \mathbf{x}^{n+1}\boldsymbol{\beta}$$

Application in dynamic walking

Video demo will be shown during lecture.



Summary of unconstrained optimization

All the above problems (unconstrained optimization) have analytic solutions.

Pros: analytic solution, fast computation, suitable for real-time implementation.

Cons: limited complexity.

Note: In real systems, physical quantities are bounded, ie systems are subject to physical constraints, because materials, energy and power are not infinite.

Quiz: what are the constraints in a physical system?

Gradient-Based Optimization

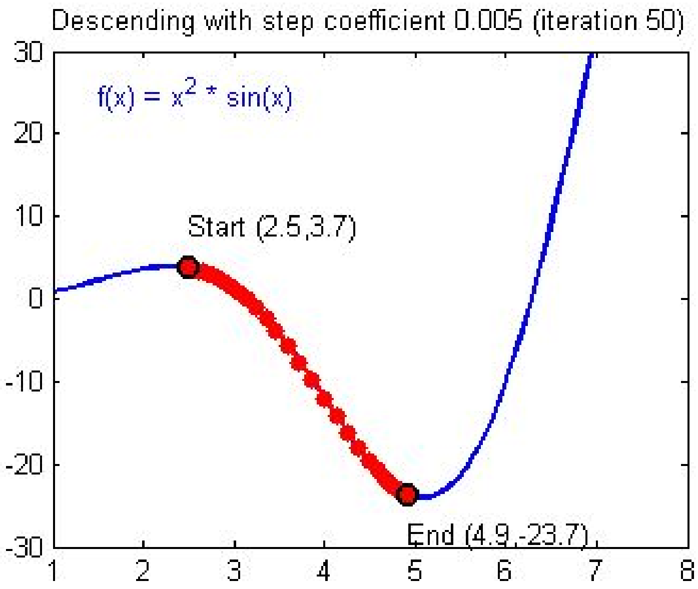
Optimization: gradient descent

Gradient descent optimization: use the first-order iterative to find the *minimum* of a function. The derivatives of a function can be used to follow the function downhill to a minimum. This technique is called *gradient descent*.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \cdot \nabla \boldsymbol{f}(\mathbf{x}_k)$$

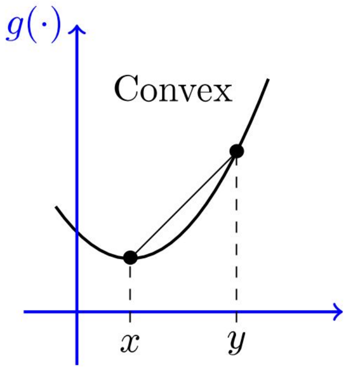
The gradient descent algorithm takes incremental steps proportional to the negative of the gradient of the function at the current point.

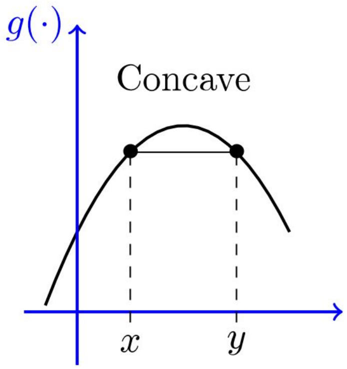
Similarly, *gradient ascent* is to find the local maxima, it takes steps proportional to the positive of the gradient.

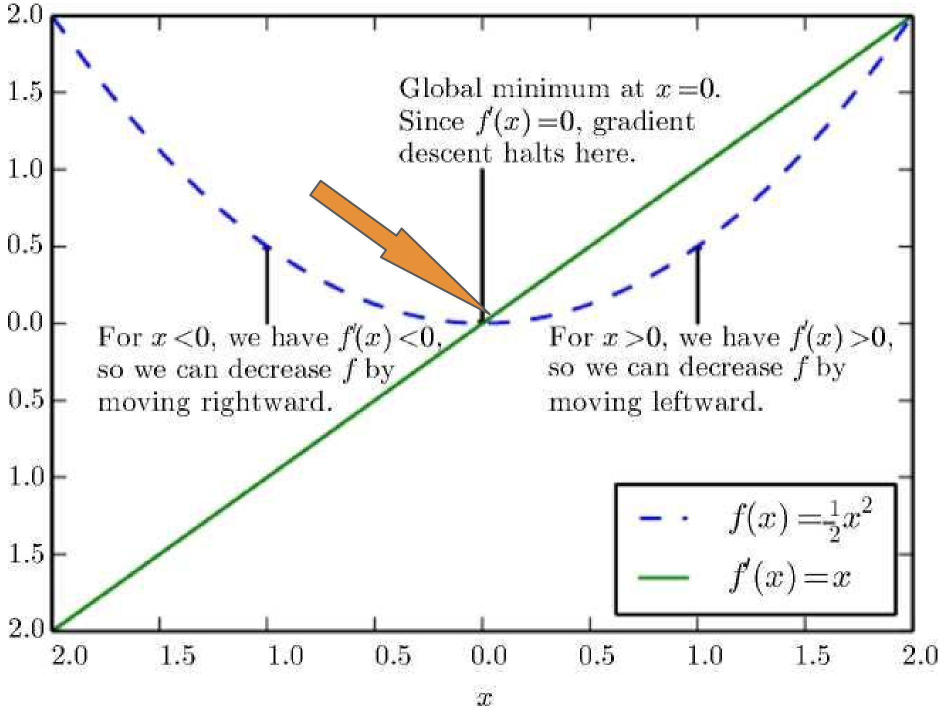


Gradient descent

∇*f*(x) = 0 is at x=0, where convex or concave function ***f***(x) reaches extrema.





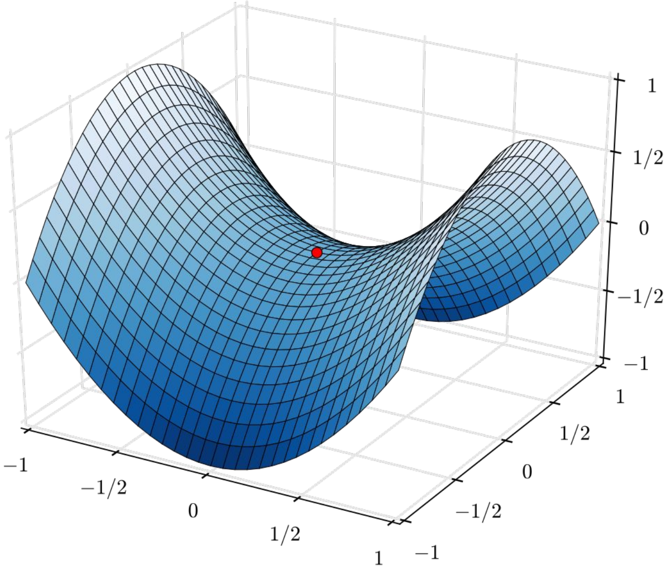


Picture source: Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

Saddle point

$\nabla \boldsymbol{f}(x) = 0 \Rightarrow \boldsymbol{f}(x)$ at extrema is not necessarily true when $\boldsymbol{f}(x)$ becomes more complicated.

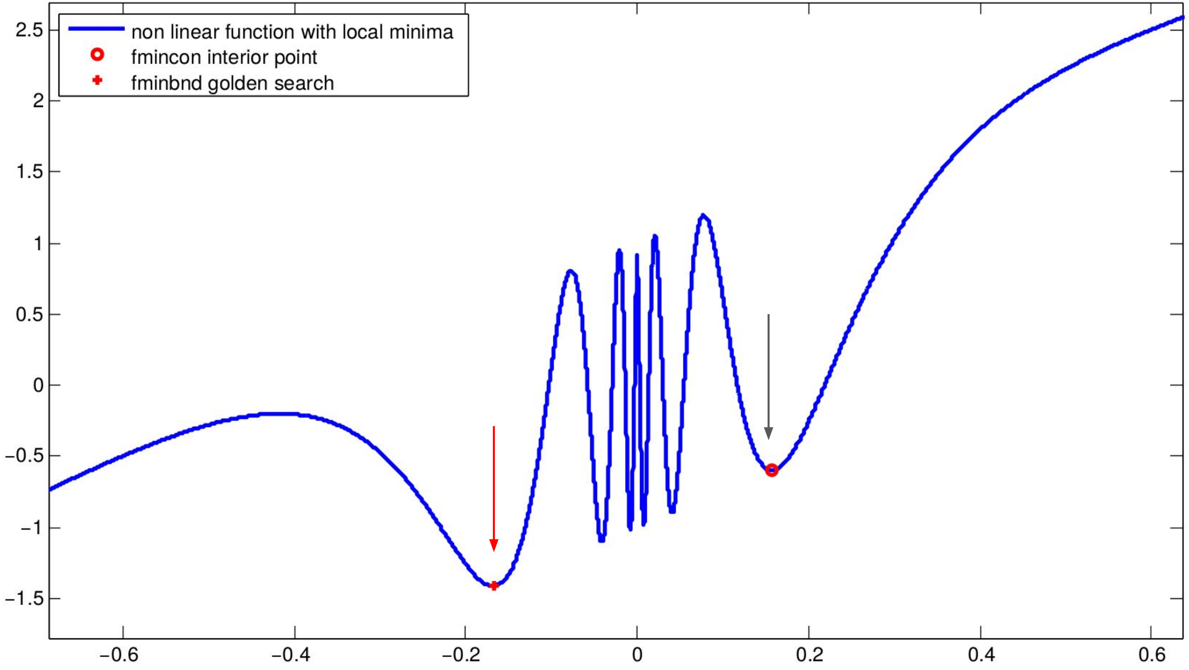
Saddle point
A point on the surface where the derivatives) of orthogonal function are zero, ie *stationary point*.



A saddle point (in red) on the function $z=x^2-y^2$
(Picture source: wikipedia)

Local minima problem

Search by gradient or continuous search can fail in complicated scenarios.



Constrained optimization

Constrained optimization

Practically, apart from maximizing or minimizing a function f(x) over all possible values of **x**, the range of feasible **x** are usually limited to some set **S**.

Thus, we have *constrained optimization* problems. Formally, if we define **S**,

$$\mathbf{S} = \{\ \mathbf{x} \mid \forall\ i,\ g_i(x) = 0 \text{ and } \forall j,\ h_j(x) \leq 0\},$$

where $g_i(\mathbf{x}) = c_i$ (i = 1 , … , n) are the equality constraints to be satisfied;

and $h_j(\mathbf{x}) \leq d_j$ (j=1 , … , m) are the inequality constraints to be satisfied.

Constrained optimization

A general form of constrained optimization problem

$$\underset{\mathbf{x}\in\mathbf{S}}{\arg\min}\ \boldsymbol{f}(\mathbf{x})$$

Or more specifically with the constraints as,

$$\begin{gathered}\arg\min\ \boldsymbol{f}(\mathbf{x})\\ \mathbf{g}(\mathbf{x}) = \mathbf{c}\\ \mathbf{h}(\mathbf{x}) \leq \mathbf{d}\end{gathered}$$

by using vectors (**c**, **d**) and matrices (**g**, **h**).

The objective function ***f*** is actually the sum of different cost functions.

Lagrange Multiplier

Lagrange multiplier

The method of Lagrange multipliers is to maximize or minimize functions with subject to equality constraints. More generally, using vector $\mathbf{x}=[x_1,...x_n]^T$, $\mathbf{x} \in \mathbf{R}^n$,

$$\text{maximize } \boldsymbol{f}(\mathbf{x})$$

$$\text{subject to } \boldsymbol{g}(\mathbf{x}) = 0$$

Often, it is easier to express the constraint as a multivariable function that equals a constant, $\boldsymbol{g}(\mathbf{x})=c$. Mathematically, it yields no difference in principle,

$$\text{maximize } \boldsymbol{f}(\mathbf{x})$$

$$\text{subject to } \boldsymbol{g}(\mathbf{x}) - c = 0$$

* same principle for minimization, ie arg min case.

Lagrange multiplier

Optimization problem:

$$\text{maximize } \boldsymbol{f}(\mathbf{x})$$

$$\text{subject to } \boldsymbol{g}(\mathbf{x}) = 0$$

Assume both ***f*** and ***g*** have continuous first partial derivatives, ie ***f*** and ***g*** are differentiable.

We introduce a dummy variable λ called a Lagrange multiplier. Define the Lagrange function ***L*** by

$$\boldsymbol{L}(\mathbf{x}, \lambda) = \boldsymbol{f}(\mathbf{x}) - \lambda \cdot \boldsymbol{g}(\mathbf{x})$$

Lagrange multiplier

Assume ***f***($\mathbf{x}_0$) is a maximum of ***f***($\mathbf{x}$) for the original constrained problem, then there exists λ_0 such that vector ($\mathbf{x}_0$, λ_0) is a point for the Lagrange function ***L*** where the partial derivatives of ***L*** is zero.

The derivatives of ***L*** at ($\mathbf{x}_0$, λ_0) is zero at all dimensions, ie gradient at ($\mathbf{x}_0$, λ_0) is zero,

$$\begin{gathered} \partial \boldsymbol{L}/\partial x_1=0 \\ \vdots \\ \partial \boldsymbol{L}/\partial x_n=0 \\ \partial \boldsymbol{L}/\partial \lambda=0 \end{gathered}$$

← Solving these equations yields the solution of ($\mathbf{x}_0$, λ_0)

thus ($\mathbf{x}_0$, λ_0) is called the **stationary point** (**critical point**).

Lagrange multiplier

$$\boldsymbol{L}(\mathbf{x}, \lambda) = \boldsymbol{f}(\mathbf{x}) - \lambda \cdot \boldsymbol{g}(\mathbf{x})$$

The constrained optimum of the original problem, ***f***, coincides with a stationary point of ***L***.

Evaluating the Lagrangian itself at a solution ($\mathbf{x}_0$, λ_0) will give the maximum value ***f***.

This is because particularly at stationary point, ***g***(**x**) or (***g***(**x**)-c) in the Lagrangian goes to zero.

Lagrange multiplier: how to understand it?

To understand better, we use the alternative expression that includes constraint c, $\boldsymbol{g}(\mathbf{x})$=c. The optimization problem:

$$\text{maximize } \boldsymbol{f}(\mathbf{x})$$

$$\text{subject to } \boldsymbol{g}(\mathbf{x}) - c = 0$$

The Lagrange function

$$\boldsymbol{L}(\mathbf{x}, \lambda) = \boldsymbol{f}(\mathbf{x}) - \lambda \cdot [\, \boldsymbol{g}(\mathbf{x}) - c \,]$$

Quiz: $\boldsymbol{L}$ is a multivariable function, what is $\partial \boldsymbol{L}/\partial c$? Answer: λ

Lagrange multiplier: how to understand it?

The Lagrange function

$$\boldsymbol{L}(\mathbf{x}, \lambda) = \boldsymbol{f}(\mathbf{x}) - \lambda \cdot [\ \boldsymbol{g}(\mathbf{x}) - c\ ]$$

This partial derivative $\partial \boldsymbol{L}/\partial c=\lambda$ shows that λ is the rate of change of the ***L*** as a function of the constrained parameter c.

In other words, Lagrange multiplier λ tells how much the value of ***L*** changes due to the change of a given constraint c.

Lagrange multiplier: practice

Let **x**=[x,y], the optimization problem is

$$\arg\max \boldsymbol{f}(x, y) = x + y$$

$$\text{subject to } \boldsymbol{g}(x,y) = x^2 + y^2 = 1$$

Recall

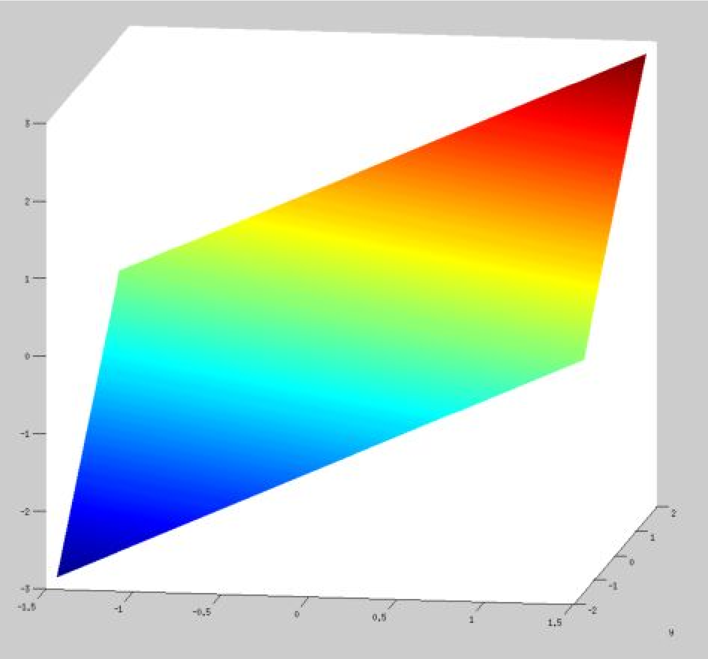
$$\boldsymbol{L}(\mathbf{x}, \lambda) = \boldsymbol{f}(\mathbf{x}) - \lambda \cdot [\, \boldsymbol{g}(\mathbf{x}) - c \,]$$

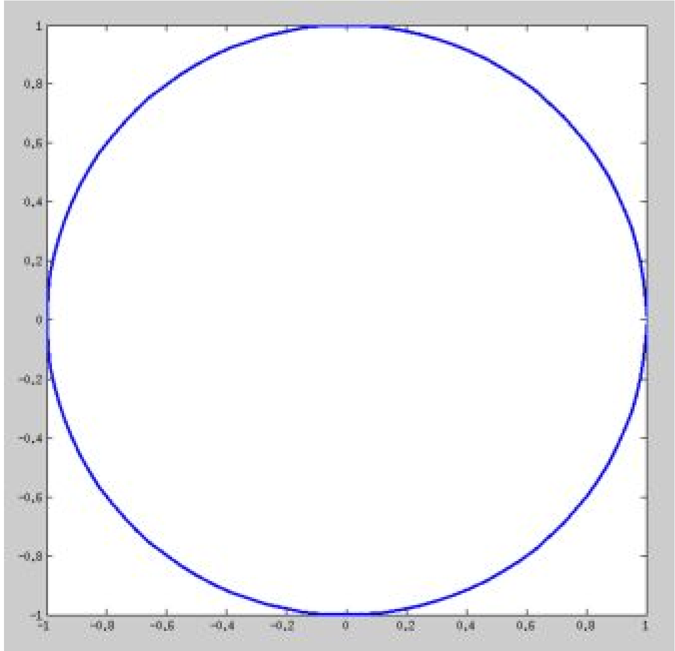
Hence

$$\boldsymbol{L}(\mathbf{x}, \lambda) = x + y - \lambda \cdot (x^2 + y^2 - 1)$$

Lagrange multiplier: practice

Intuitively, ***f***(x, y) = x + y is an inclined plane, and ***g***(x,y) = x^2 + y^2 = 1 is a circle.

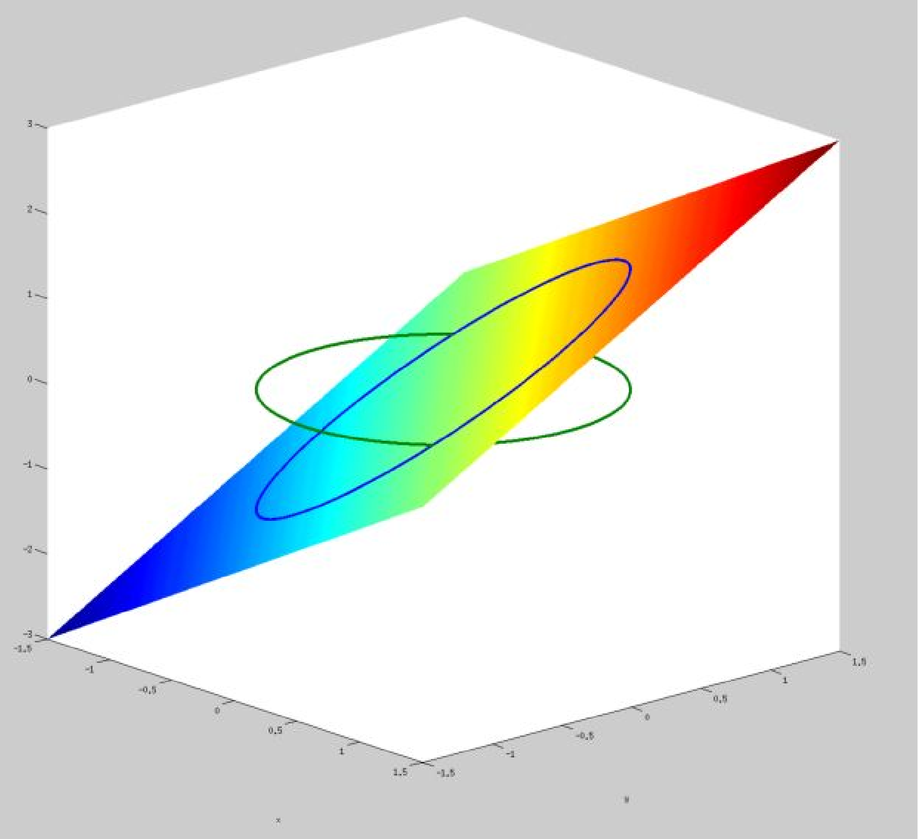




Lagrange multiplier: practice

So intuitively speaking:

1. How our optimization problem looks like?
2. What is the optimal solution?



Lagrange multiplier: whiteboard exercise

Recall that solving equations below yields the solution of $(\mathbf{x}_0, \lambda_0)$

$$\partial \boldsymbol{L}/\partial x_1=0$$

$$\vdots$$

$$\partial \boldsymbol{L}/\partial x_n=0$$

$$\partial \boldsymbol{L}/\partial \lambda=0$$

Quiz: so how these partial derivatives look like in our problem?

$$\boldsymbol{L}(\mathbf{x}, \lambda) = x + y - \lambda \cdot (x^2 + y^2 - 1)$$

$$\partial \boldsymbol{L}/\partial x=?$$

$$\partial \boldsymbol{L}/\partial y=?$$

$$\partial \boldsymbol{L}/\partial \lambda=?$$

Lagrange multiplier: practice

Quiz: so how these partial derivatives look like in our problem?

$$\boldsymbol{L}(\mathbf{x}, \lambda) = x + y - \lambda \cdot (x^2 + y^2 - 1)$$

$$\partial \boldsymbol{L}/\partial x = 2\lambda x+1 = 0$$

$$\partial \boldsymbol{L}/\partial y = 2\lambda y+1 = 0$$

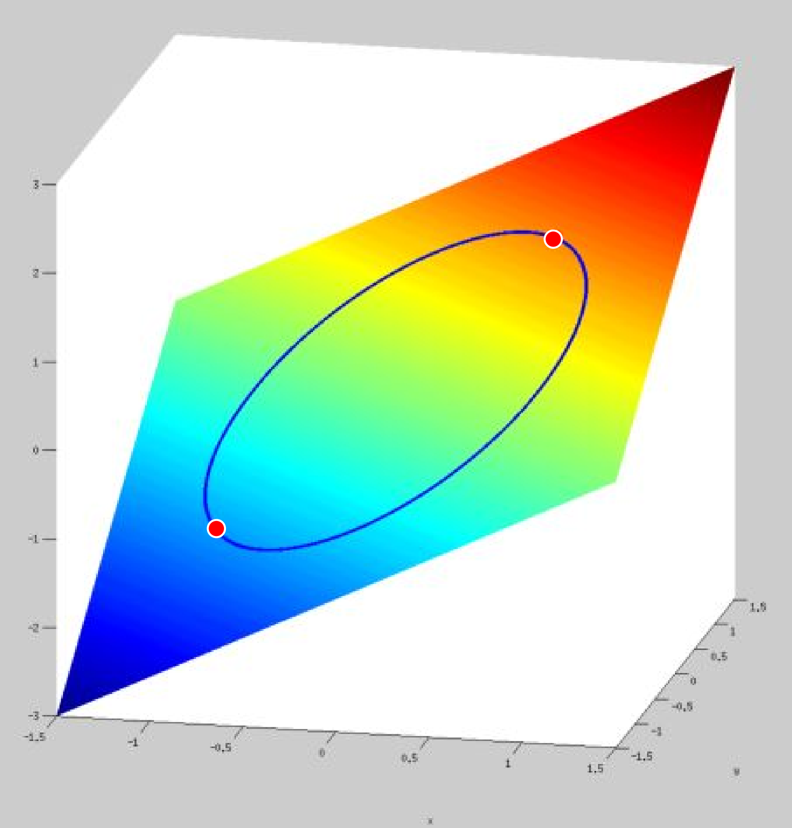
$$\partial \boldsymbol{L}/\partial \lambda = x^2+y^2-1 = 0$$

$\lambda=\pm\sqrt{2}/2$ ($2^{-1/2}$), note $x=y=-1/(2\lambda)$, so, stationary points:

$(-\sqrt{2}/2,-\sqrt{2}/2,\sqrt{2}/2)$, $(\sqrt{2}/2,\sqrt{2}/2,-\sqrt{2}/2)$

$f_{max}=\sqrt{2}$, $f_{min}=-\sqrt{2}$.

Lagrange multiplier: practice



Summary

Unconstrained optimization

- ❏ Least Square (LS) optimization: case study of IR sensor, good estimates of polynomial coefficients
- ❏ Tikhonov regularisation: success application for precious walking speed control

Constrained optimization

- ❏ Lagrange multipliers: maximization and minimization problem subject to equality constraints