Hassan OMRAN

Lecture 4: Least-Squares Optimization

Télécom Physique Strasbourg
Université de Strasbourg

Outline of the talk

1. Introduction
2. Solving $Ax = b$
   - Least-Squares solution to an overdetermined $Ax = b$
   - Solution to Ax=b minimizing $||x||$
   - The general solution to $Ax = b$
3. Nonlinear Least-Squares

1. Introduction

2. Solving $Ax = b$
   - Least-Squares solution to an overdetermined $Ax = b$
   - Solution to Ax=b minimizing $||x||$
   - The general solution to $Ax = b$

3. Nonlinear Least-Squares

Introduction

- *Least squares problem is an old important problem*
- *Early contributions by Gauss and Legendre in the beginning of the* 19th *century*
- *Fitting a model to measurements and observations subject to errors is a basic problem in science*
- *This lecture is a brief review of linear and nonlinear least-squares optimization*

Carl Friedrich Gauss

Adrien-Marie Legendre

1. Introduction

2. Solving $Ax = b$
   - Least-Squares solution to an overdetermined $Ax = b$
   - Solution to Ax=b minimizing $||x||$
   - The general solution to $Ax = b$

3. Nonlinear Least-Squares

Least-Squares solution to an overdetermined $Ax = b$

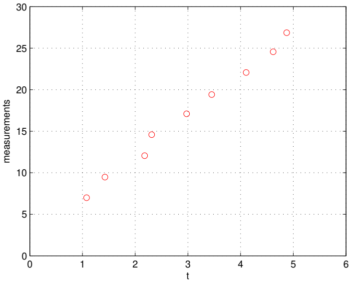
Motivational example:

*Consider the problem of **data fitting**, where we want to find the best line fit to the measurement of the output of a process*

That is, using a set of measurements $\{y_1, \ldots, y_m\}$ *at* $\{t_1, \ldots, t_m\}$*, we want to find the best linear model*

$$y(t) = at + b$$

In other words, we want to find the "optimal" a *and* b*, such that the model "fits" the data*



Measured data of a process

Least-Squares solution to an overdetermined $Ax = b$

Motivational example:

We want to find a and b such that

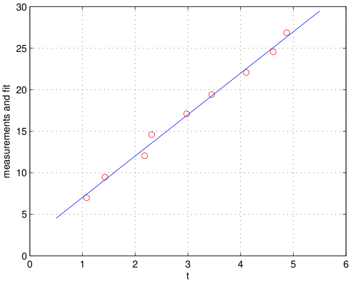
$$\begin{aligned} y_1 &= a t_1 + b \\ &\vdots \\ y_m &= a t_m + b \end{aligned}$$

Which can be written as

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

However, the above system of equations is (often) inconsistent (no solution can be found). Instead, we can find the line which minimizes

$$\sum_{i=1}^{m} (y_i - a t_i - b)^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$$



The best line fit

Least-Squares solution to an overdetermined $Ax = b$

Consider a system of linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^{m}$. *We consider here the case where* $m \geq n$ *rank*$(\mathbf{A}) = n$. *When* $\mathbf{b}$ *does not belong to the range of* $\mathbf{A}$ ($\mathbf{b} \notin \mathscr{R}(\mathbf{A})$), *in this case there is no solution to* (1). *In this case we say that* (1) *is overdetermined (or inconsistent)*

We aim to find the vector $\mathbf{x}^\star$ *such that*

$$\begin{aligned} \mathbf{x}^\star = \arg\min_{\mathbf{x}} \quad & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \\ s.t. \quad & \mathbf{x} \in \mathbb{R}^n \end{aligned} \tag{2}$$

Note that $\mathbf{x}^\star$ *is equal to the solution of* (1) *when it does have a solution, otherwise it minimizes the difference between its left and right sides. Before presenting the solution to* (2), *we need the following interesting Theorem.*

Theorem 1

Consider $\mathbf{A} \in \mathbb{R}^{m\times n}$, *such that* $m \geq n$. *In this case, rank*$(\mathbf{A}) = n$ *if and only if rank*$(\mathbf{A}^T\mathbf{A}) = n$.

The above Theorem states that $\mathbf{A}$ *is a full column rank if and only if* $\mathbf{A}^T\mathbf{A}$ *is nonsingular.*

Least-Squares solution to an overdetermined $Ax = b$

Theorem 2

The vector $\mathbf{x}^\star$ that minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is given by the solution to the equation $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. That is

$$\mathbf{x}^\star = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{3}$$

Proof.

Note that in this case $\mathbf{A}^T\mathbf{A}$ is nonsingular and definite positive.

$$\begin{aligned} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 &= (\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= \frac{1}{2}\mathbf{x}^T(2\mathbf{A}^T\mathbf{A})\mathbf{x} + (-2\mathbf{b}^T\mathbf{A})\mathbf{x} + \mathbf{b}^T\mathbf{b} \end{aligned}$$

This show that the problem in (2) is actually the problem of finding the minimum of a convex quadratic function:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \underbrace{(2\mathbf{A}^T\mathbf{A})}_{Q} \mathbf{x} + \underbrace{(-2\mathbf{b}^T\mathbf{A})}_{\mathbf{q}^T} \mathbf{x}$$

Least-Squares solution to an overdetermined $Ax = b$

Proof.

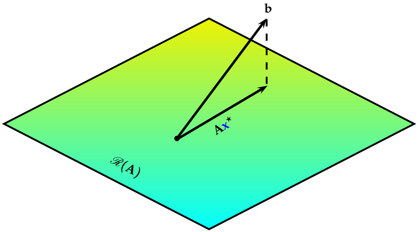
using SOSF we see that the unique minimizer is given by

$$\begin{aligned} \boldsymbol{0} &= \nabla f(\boldsymbol{x}^\star) \\ &= (2\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{x}^\star + (-2\boldsymbol{A}^T\boldsymbol{b}) \end{aligned}$$

That is

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}^\star = \boldsymbol{A}^T\boldsymbol{b} \qquad \Rightarrow \qquad \boldsymbol{x}^\star = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$$

$\square$



The geometric interpretation of the problem (2)

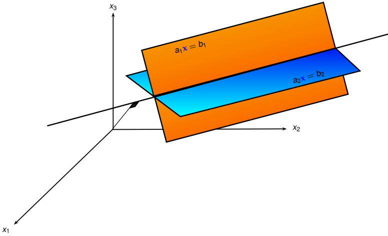
See Gauss–Markov Theorem for the statistical interpretation of the ordinary least-squares problem

Solution to Ax=b minimizing ||x||

Motivational example:

Consider the intersection between two hyperplanes $a_1 \mathbf{x} = b_1$ and $a_2 \mathbf{x} = b_2$ (with $\mathbf{x} \in \mathbb{R}^3$). The points belonging to the intersection verify the system of linear equations

$$\underbrace{\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}}_{\boldsymbol{A}} \mathbf{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}_{\boldsymbol{b}}$$



Here there exists an infinite number of solutions to $\boldsymbol{A}\mathbf{x} = \boldsymbol{b}$.

We are interested in the solution which is the closest to the origin.

Solution to Ax=b minimizing ||x||

Consider a system of linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^{m}$. We consider here the case where $m \leq n$ $rank(\mathbf{A}) = m$.

We aim to find the vector $\mathbf{x}^\star$ such that

$$\begin{aligned} \mathbf{x}^\star &= \arg\min_{\mathbf{x}} \quad \|\mathbf{x}\|^2 \\ s.t. & \quad\quad\quad \mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned} \tag{5}$$

Theorem 3

The unique solution to (4) that minimizes $\|\mathbf{x}\|^2$ is

$$\mathbf{x}^\star = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b} \tag{6}$$

Solution to Ax=b minimizing ||x||

Proof.

Consider $x^\star$ *which any solution to* $Ax = b$ *different from* $x^\star = A^T(AA^T)^{-1}b$. *Note that*

$$\begin{aligned} \|x\|^2 &= \|(x - x^\star) + x^\star\|^2 \\ &= \|x - x^\star\|^2 + \|x^\star\|^2 + 2x^{\star T}(x - x^\star) \end{aligned} \tag{7}$$

The last term is

$$\begin{aligned} x^{\star T}(x - x^\star) &= \left(A^T(AA^T)^{-1}b\right)^T\left(x - A^T(AA^T)^{-1}b\right) \\ &= \left(b^T(AA^T)^{-1}\right)A\left(x - A^T(AA^T)^{-1}b\right) \\ &= \left(b^T(AA^T)^{-1}\right)\left(\underbrace{Ax - b}_{=0}\right) = 0 \end{aligned}$$

Thus, from (7) *we have that*

$$\begin{aligned} \|x\|^2 &= \|x - x^\star\|^2 + \|x^\star\|^2 > \|x^\star\|^2 \qquad (\textit{since } x \neq x^\star \Rightarrow \|x - x^\star\|^2 > 0) \\ \|x\|^2 &> \|x^\star\|^2 \end{aligned}$$

The general solution to $Ax = b$

Consider a system of linear equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{8}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^{m}$, $\mathrm{rank}(\boldsymbol{A}) = r$. *Note that* $r \leq \min(m, n)$. *We are interested in a generalized approach to solve* (8) *using the notion of* ***pseudoinverse***.

In particular we are interested in ***Moore-Penrose inverse*** *which is denoted by* $\boldsymbol{A}^{\dagger}$.

The pseudoinverse

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, *a matrix* $\boldsymbol{A}^{\dagger} \in \mathbb{R}^{n \times m}$ *is called a* ***pseudoinverse*** *of the matrix* $\boldsymbol{A}$ *if*

$$\boldsymbol{A}\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{A}$$

and there exist $U \in \mathbb{R}^{n \times n}$ *and* $V \in \mathbb{R}^{m \times m}$ *such that*

$$\boldsymbol{A}^{\dagger} = U\boldsymbol{A}^{T} \qquad \text{and} \qquad \boldsymbol{A}^{\dagger} = \boldsymbol{A}^{T}V$$

The general solution to $Ax = b$

The left pseudoinverse

For the case $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *with* $m \geq n$ *and* $\operatorname{rank}(\boldsymbol{A}) = n$*, verify that*

$$\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T$$

is a pseudoinverse that satisfies $\boldsymbol{A}^{\dagger} \boldsymbol{A} = I_n$*. Note that this pseudoinverse appears in* (3)*.*

The right pseudoinverse

For the case $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *with* $m \leq n$ *and* $\operatorname{rank}(\boldsymbol{A}) = m$*, verify that*

$$\boldsymbol{A}^{\dagger} = \boldsymbol{A}^T (\boldsymbol{A} \boldsymbol{A}^T)^{-1}$$

is a pseudoinverse that satisfies $\boldsymbol{A} \boldsymbol{A}^{\dagger} = I_m$*. Note that this pseudoinverse appears in* (6)*.*

The general solution to $Ax = b$

The pseudoinverse has the following nice existence and uniqueness property [1]

Theorem 4

*Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the pseudoinverse always **exists** and it is **unique**.*

Finally, the pseudoinverse can be interpreted in the context of solving a system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$

Theorem 5

Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and a vector $\boldsymbol{b} \in \mathbb{R}^{m}$, where $\text{rank}(\boldsymbol{A}) = r$, consider the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. The vector $\boldsymbol{x}^\star = \boldsymbol{A}^\dagger \boldsymbol{b}$ is a solution to the problem of minimizing $||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2$ on $\mathbb{R}^n$.
Moreover, $\boldsymbol{x}^\star = \boldsymbol{A}^\dagger \boldsymbol{b}$ is the unique vector with minimal norm among all solutions that minimize $||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2$.

[1] For more information about generalized inverses see **Generalized Inverses, Theory and Applications** by Adi Ben-Israel and Thomas N. E. Greville

1. Introduction

2. Solving $Ax = b$
   - Least-Squares solution to an overdetermined $Ax = b$
   - Solution to Ax=b minimizing $||x||$
   - The general solution to $Ax = b$

3. Nonlinear Least-Squares

Nonlinear Least-Squares

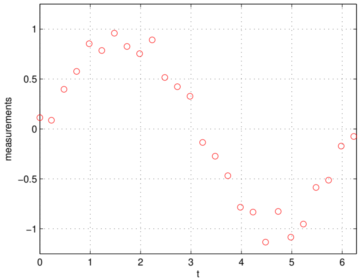
Motivational example:

*Consider the problem of **data fitting**, where we want to find the best fit to the measurement of the output of a process*

That is, using a set of measurements $\{y_1, \ldots, y_m\}$ *at* $\{t_1, \ldots, t_m\}$*, we want to find the best nonlinear model*

$$y(t) = a\sin(\omega t + \phi)$$

In other words, we want to find the "optimal" model that "fits" the data.



Measured data

Nonlinear Least-Squares

Motivational example:

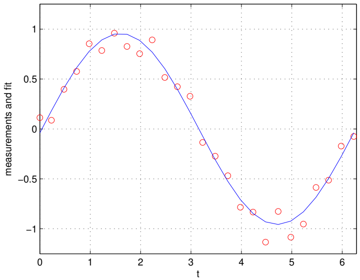
We want to find a, ω and ϕ such that

$$\begin{aligned} y_1 &= a\sin(\omega t_1 + \phi) \\ &\vdots \\ y_m &= a\sin(\omega t_m + \phi) \end{aligned}$$

However, (often) no solution can be found.

Instead, we can find the line which minimizes

$$\sum_{i=1}^{m} \Big(y_i - a\sin(\omega t_i + \phi)\Big)^2$$



The best sinusoidal fit

Nonlinear Least-Squares

Consider the following ***nonlinear least-squares*** *problem*

$$\begin{aligned} \mathbf{x}^\star = \arg\min_{\mathbf{x}} \quad & F(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x})^2 \\ s.t. \quad & \mathbf{x} \in \mathbb{R}^n \end{aligned} \tag{9}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$, *for* $i \in \{1, \ldots, m\}$. *By defining*

$$f(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix}$$

the cost function can be written as $F(\mathbf{x}) = f(\mathbf{x})^T f(\mathbf{x})$. *Several methods can be used to solve this optimization problem. Here we consider applying Newton's method. We need to calculate the gradient and the Hessian of* $F(\cdot)$. *First, using the product rules for derivation we have*

$$DF(\mathbf{x}) = 2f(\mathbf{x})^T J(\mathbf{x}) \quad \Leftrightarrow \quad \nabla F(\mathbf{x}) = 2J(\mathbf{x})^T f(\mathbf{x}) \tag{10}$$

Nonlinear Least-Squares

where $J(\cdot)$ is a the Jacobian of $f(\cdot)$

$$J(\boldsymbol{x}) = Df(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} := [J_1(\boldsymbol{x}) \cdots J_n(\boldsymbol{x})] \tag{11}$$

Where $J_i(\boldsymbol{x})$ are the columns of the Jacobian $J(\boldsymbol{x})$. Lets also calculate the Hessian :

$$\begin{aligned} D^2F(\boldsymbol{x}) &= D(\nabla F(\boldsymbol{x})) \\ &= 2D(J(\boldsymbol{x})^T f(\boldsymbol{x})) \\ &= 2D\begin{bmatrix} J_1(\boldsymbol{x})^T f(\boldsymbol{x}) \\ \vdots \\ J_n(\boldsymbol{x})^T f(\boldsymbol{x}) \end{bmatrix} = 2\begin{bmatrix} D(J_1(\boldsymbol{x})^T f(\boldsymbol{x})) \\ \vdots \\ D(J_n(\boldsymbol{x})^T f(\boldsymbol{x})) \end{bmatrix} \\ &= 2\begin{bmatrix} J_1(\boldsymbol{x})^T J(\boldsymbol{x}) + f(\boldsymbol{x})^T D(J_1(\boldsymbol{x})) \\ \vdots \\ J_n(\boldsymbol{x})^T J(\boldsymbol{x}) + f(\boldsymbol{x})^T D(J_n(\boldsymbol{x})) \end{bmatrix} = 2\left(J(\boldsymbol{x})^T J(\boldsymbol{x}) + \begin{bmatrix} f(\boldsymbol{x})^T D(J_1(\boldsymbol{x})) \\ \vdots \\ f(\boldsymbol{x})^T D(J_n(\boldsymbol{x})) \end{bmatrix}\right) \end{aligned}$$

Nonlinear Least-Squares

$$D^2F(\boldsymbol{x}) = 2\Big(J(\boldsymbol{x})^T J(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})\Big) \quad \text{with} \quad \boldsymbol{S}(\boldsymbol{x}) = \begin{bmatrix} f(\boldsymbol{x})^T D(J_1(\boldsymbol{x})) \\ \vdots \\ f(\boldsymbol{x})^T D(J_n(\boldsymbol{x})) \end{bmatrix} \tag{12}$$

From (12) *and* (10) *and by applying Newton's method we get*

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \Big(J(\boldsymbol{x}_k)^T J(\boldsymbol{x}_k) + \boldsymbol{S}(\boldsymbol{x}_k)\Big)^{-1} J(\boldsymbol{x}_k)^T f(\boldsymbol{x}_k) \tag{13}$$

In several applications, the term $\boldsymbol{S}(\cdot)$ can be ignored, which yields

Gauss-Newton method:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \Big(J(\boldsymbol{x}_k)^T J(\boldsymbol{x}_k)\Big)^{-1} J(\boldsymbol{x}_k)^T f(\boldsymbol{x}_k)$$

Nonlinear Least-Squares

In order to overcome the potential problem of not having a positive definite $J(\mathbf{x}_k)^T J(\mathbf{x}_k)$, *the method can be modified using Levenberg-Marquardt modification*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left(J(\mathbf{x}_k)^T J(\mathbf{x}_k) + \mu_k I\right)^{-1} J(\mathbf{x}_k)^T f(\mathbf{x}_k) \tag{14}$$

The term $\mu_k I$ *can be used to approximate* $\boldsymbol{S}(\mathbf{x}_k)$ *in* (13).

Note that Levenberg-Marquardt method was originally modified for nonlinear least-squares problem.

Finally, note that other methods such as quasi-Newton can also be used to solve (9).