Least Squares Problems

Recommended reading:

- Lecture 11 in [4]
- Section II.2 in [3]
- This handout by Homer F. Walker:
  https://users.wpi.edu/∼walker/MA3257/HANDOUTS/least-squares_handout.pdf

[1] R. Rannacher.
Numerik 0 - Einführung in die Numerische Mathematik.
Heidelberg University Publishing, 2017.

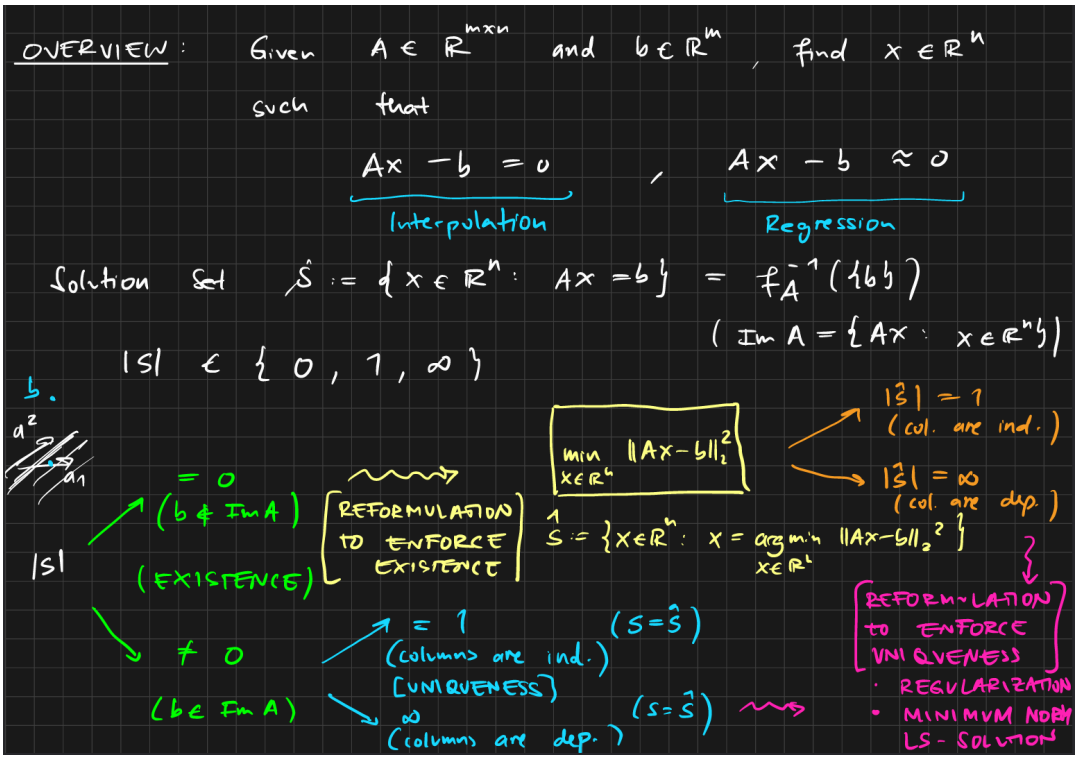
[2] G. Strang.
Introduction to Linear Algebra.
Wellesley-Cambridge Press, 2003.

[3] G. Strang.
Linear Algebra and Learning from Data.
Wellesley-Cambridge Press, 2019.

[4] L.N. Trefethen and D. Bau.
Numerical linear algebra.
SIAM, Soc. for Industrial and Applied Math., Philadelphia, 1997.

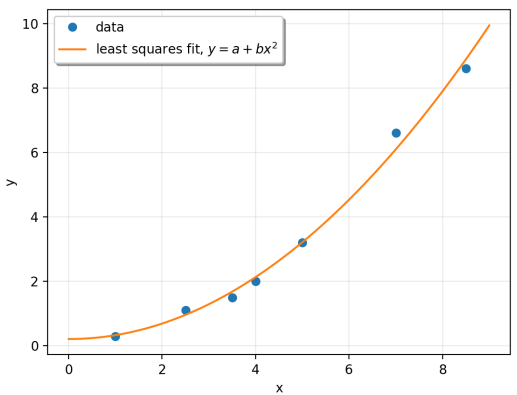
6 Least Squares Problems

6.1 Overview



Situation: We allow for $b \notin \mathrm{Im}(A)$
$\Rightarrow$ The system $Ax = b$ is not solvable, i.e., there is **no** $x^* \in \mathbb{R}^n$ so that $Ax^* = b$

Example: Curve fitting
The situation above typically occurs when trying to explain a set of data by just a few parameters leading to over-determined systems: more equations than unknowns ($m \gg n$).



Corresponding system:

$$\begin{pmatrix} 1 & z_1^2 \\ \vdots & \vdots \\ 1 & z_n^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Approach: Minimize the error (/residual/defect) $\|Ax - b\|$

We obtain existence by reformulating the problem:

Definition 6.1 (*Least Squares Solution*) *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$. Then $\widehat{x} \in \mathbb{R}^n$ is called a* ***least squares solution of*** *$Ax = b$, if $\widehat{x}$ is a minimizer of the problem*

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2,$$

i.e., $\|A\widehat{x} - b\|_2^2 \leq \|Ax - b\|_2^2$ for all $x \in \mathbb{R}^n$.

Remark

- We recall that $\|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2}$ which explains the naming:

$$\|Ax - b\|_2^2 = \sum_{i=1}^m \underbrace{(\dots)^2}_{\text{"squares"}} \rightarrow \underbrace{\min}_{\text{"least squares"}}.$$

- The norm is always nonnegative, i.e., $\|x\|_2 \geq 0$, so that the minimal possible value of the objective function is zero (which would imply $Ax = b$ due to the definiteness of the norm).
- Also, since squaring $x \mapsto x^2$ is a monotonically increasing function on nonnegative numbers we can minimize the squared residual without changing the set of minimizers, i.e.,

$$\widehat{S} := \{x \in \mathbb{R}^n : \ x := \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2\} = \{x \in \mathbb{R}^n : \ x := \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2\}.$$

  As a result we get rid of the square root which is advantageous in terms of derivatives (see optimality conditions later). Note that the optimal value of the objective function may differ, but this is not important here.

6.2 The Normal Equation

The minimization problem is equivalent to a linear system:

Theorem 6.2 (*Normal Equation*) *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then $\widehat{x} \in \mathbb{R}^n$ is a least squares solution of $Ax = b$ if and only if $\widehat{x}$ solves the* ***normal equation***

$$A^T A x = A^T b.$$

Proof sketch:

(1) Optimization perspective:

Let us define the objective function

$$f\colon \mathbb{R}^n \to \mathbb{R}, \quad f(x) := \|Ax - b\|_2^2.$$

Then one can show that f is convex, i.e.,

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall\, \lambda \in [0,1].$$

In fact, this is an immediate consequence of the triangle inequality and the absolute homogeneity of the norm as well as the monotonicity of the square.

The convexity of the objective function then implies the existence of a minimizer as well as the necessary *and* sufficient first–order optimality condition:

$$\widehat{x} \text{ minimizer} \iff 0 = f'(\widehat{x}) = 2A^T(A\widehat{x} - b) \quad \text{(normal equation)}.$$

(2) Geometric Perspective:

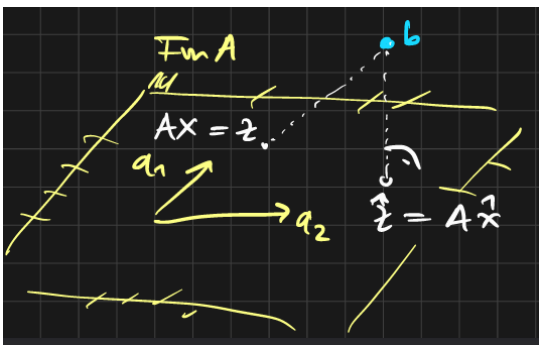
We recall that

$$\mathrm{Im}(A) = \{Ax \colon x \in \mathbb{R}^n\} = \mathrm{span}(a_1, \ldots, a_n) = \{x_1 a_1 + \ldots + x_n a_n \colon x \in \mathbb{R}^n\} \subset \mathbb{R}^m.$$

Therefore the least squares problem also reads as

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = \min_{z \in \mathrm{Im}(A)} \|z - b\|_2^2.$$

Example: Let $A \in \mathbb{R}^{3\times 2}$ and $b \in \mathbb{R}^3$.



- The point $\hat{z} \in \mathbb{R}^m$ ist the point in the plane $\mathrm{Im}(A)$, which is as close as possible to b in terms of the Euclidean norm $\|\cdot\|_2$.

  $\rightarrow$ The vector $(\hat{z} - b)$ is orthogonal to this plane!
- By definition of the image $\mathrm{Im}(A)$, each z in this plane can be written as $z = Ax$ for some $x \in \mathbb{R}^n$.
- The parameter vector $\hat{x}$ with $\hat{z} = A\hat{x}$ is the desired least squares solution.

One can show that the orthogonal projection yields shortest distance:

Lemma 6.3 (*Orthogonal Projection*) *Let $V \subset \mathbb{R}^m$ be a linear subspace and $b \in \mathbb{R}^m$. Then*

$$\widehat{z} = \arg\min_{z \in V} \|z - b\|_2^2 \quad \Leftrightarrow \quad \widehat{z} - b \in V^\perp := \{w \in \mathbb{R}^n : w^\top z = 0 \ \ \forall z \in V\}, \ \ \widehat{z} \in V.$$

Proof. The following equation is crucial for the proof (compare Pythagorean identity): Let $\widehat{z} \in V$ be fixed, then for all $z \in V$ we have

$$\begin{aligned} \|z - b\|_2^2 &= \|b - \widehat{z} + \widehat{z} - z\|_2^2 \\ &= \|b - \widehat{z}\|_2^2 + \|\widehat{z} - z\|_2^2 + 2(b - \widehat{z})^\top (z - \widehat{z}) \end{aligned} \tag{9}$$

Exemplary, we only proof direction "$\Leftarrow$" here.
Now let $\widehat{z} - b \in V^\perp$, so that $\forall z \in V$,

$$(\widehat{z} - b)^\top (z - \widehat{z}) = 0.$$

Inserting this into equation (9) and exploiting the positivity of the norm we thus obtain

$$\begin{aligned} \|z - b\|_2^2 &= \|b - \widehat{z}\|_2^2 + \|\widehat{z} - z\|_2^2 + 2(b - \widehat{z})^\top (z - \widehat{z}) \\ &= \|b - \widehat{z}\|_2^2 + \|\widehat{z} - z\|_2^2 \\ &\geq \|b - \widehat{z}\|_2^2, \end{aligned}$$

for all $z \in V$.

The reverse direction also relies on (9). See, e.g., the third proof in this Wikipedia section. □

Now let us apply this lemma to our setting $V = \mathrm{Im}(A)$ and exploit the orthogonality of the fundamental subspaces ($\mathrm{Im}(A)^\perp = \ker(A^\top)$). We obtain

$$\begin{aligned}
\widehat{z} = \arg\min_{z \in \mathrm{Im}(A)} \|z - b\|_2^2 \iff & \quad \widehat{z} - b \in \mathrm{Im}(A)^\perp = \ker(A^\top), \ \ \widehat{z} \in \mathrm{Im}(A) && \text{(Lemma 6.3)} \\
\iff & \quad A^\top(\widehat{z} - b) = 0, \ \ \widehat{z} \in \mathrm{Im}(A) && \text{(definition of } \ker(A^\top)) \\
\iff & \quad A^\top(A\widehat{x} - b) = 0 \quad \text{for some } \widehat{x} \in \mathbb{R}^n && \text{(definition of } \widehat{z} \in \mathrm{Im}(A))
\end{aligned}$$

- The normal equation then reads as:

  The least squares solution $\widehat{x}$ is so that
  the residual vector $A\widehat{x} - b$ is orthogonal to all columns of A.

- We also find an equivalent characterization for the solution set, namely,

$$\widehat{S} = \{x \in \mathbb{R}^n \colon A^T A x = A^T b\}.$$

Analysis of the Normal Equation

(1) Properties of the system matrix A^TA (*Gramian matrix*)

- A^TA is of size $n \times n$ (for typically $n \ll m$)
- A^TA is symmetric and positive semi-definite ($\Rightarrow$ nonnegative eigenvalues)
- $\ker(A) = \ker(A^TA)$, which implies:

  A independent columns $\iff \ker(A) = \{0\} \iff \ker(A^TA) = \{0\} \iff A^TA$ is invertible.

  With other words, the least squares solution is unique if and only if A has independent columns (also compare to the geometric interpretation above).

(2) For any A, b there exists a least squares solution (existence enforced ✓)

Due to $\ker(A) = \ker(A^TA)$ and $\mathrm{Im}(A) = \ker(A^T)^\perp$, we have

$$\begin{aligned}\text{existence: } \exists x \in \mathbb{R}^n\colon A^TAx = A^Tb &\iff A^Tb \in \mathrm{Im}(A^TA) = \ker(A^TA)^\perp = \ker(A)^\perp \\ &\iff 0 = (A^Tb)^Tv = b^T(Av) \quad \forall\, v \in \ker(A)\end{aligned}$$

The latter statement on the right-hand side is true for any matrix A and any vector b.

(3) Consistent reformulation:

If $b \in \mathrm{Im}(A)$, i.e., if the original system $Ax = b$ is solvable, then

$$\{x \in \mathbb{R}^n \colon Ax = b\} =: S = \widehat{S} := \{x \in \mathbb{R}^n \colon A^T A x = A^T b\}.$$

Proof:

- "$S \subset \widehat{S}$":
  Let $x \in S = \{x \in \mathbb{R}^n : \ Ax = b\}$ (such an element x exists because we assume $b \in \mathrm{Im}(A)$), then

$$Ax - b = 0 \quad \overset{A^T \cdot |}{\Rightarrow} \quad A^T(Ax - b) = 0 \quad \Rightarrow \quad x \in \widehat{S} = \{x \in \mathbb{R}^n : \ A^T A x = A^T b\}.$$

- "$\widehat{S} \subset S$":
  Let $\widehat{x} \in \widehat{S}$, i.e., $\widehat{x} := \underset{x \in \mathbb{R}^n}{\mathrm{argmin}} \|Ax - b\|_2^2$, then because $b \in \mathrm{Im}(A)$,

$$\|A\widehat{x} - b\|_2^2 = 0.$$

  Thus

$$A\widehat{x} - b = 0 \quad \Rightarrow \quad \widehat{x} \in S.$$

6.3 Solving the Normal Equation

Assumptions:

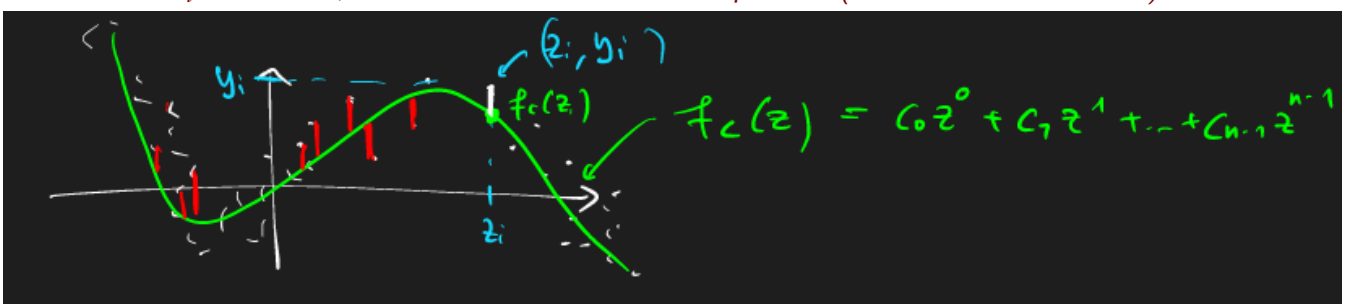
- A has independent columns (existence and uniqueness ✓)
- n is of moderate size (direct methods applicable ✓)

Thus there is a unique least squares solution given by (also revisit the section on projections)

$$\widehat{x} = (A^T A)^{-1} A^T b.$$

Example: Polynomial regression

- Here we typically have many measurements $(z_1, y_1), \ldots, (z_m, y_m) \in \mathbb{R}^2$ (i.e., m large).
- Polynomial model is given by $f_c(z) := \sum_{j=0}^{n-1} c_j z^j$ (for n rather small because we want to smoothen the data).
- The corresponding design matrix is then given by $A = (z_i^{j-1})_{ij}$ (revisit section on curve fitting).
- One can show:
  If all the z_i are distinct, then the columns of A are independent (see Vandermonde matrix)!



Approaches:

(1) Using Cholesky decomposition $A^\top A = LL^\top$

$\rightarrow$ Problem: A^TA often *ill-conditioned* and numerical elimination may fail due to rounding errors!

Can we work without A^TA*?* Yes!

(2) Using reduced QR **Decomposition** $A = \widehat{Q}\widehat{R}$

Let us recall that

$$\forall A \in \mathbb{R}^{m\times n} \; \exists \; \widehat{Q} \in \mathbb{R}^{m\times n} \text{ orthonormal columns, } \; \widehat{R} \in \mathbb{R}^{n\times n} \text{ triangular : } \; A = \widehat{Q}\widehat{R}.$$

Now we insert $A = \widehat{Q}\widehat{R}$ into the normal equation to obtain

$$A^TAx = A^Tb \overset{A=\widehat{Q}\widehat{R}}{\Leftrightarrow} (\widehat{Q}\widehat{R})^T(\widehat{Q}\widehat{R})x = (\widehat{Q}\widehat{R})^Tb \;\Leftrightarrow\; \widehat{R}^T\widehat{Q}^T\widehat{Q}\widehat{R}x = \widehat{R}^T\widehat{Q}^Tb \;\Leftrightarrow\; \widehat{R}^T\widehat{R}x = \widehat{R}^T\widehat{Q}^Tb.$$

If A has independent columns, we know that $\widehat{R}$ is invertible and so is $\widehat{R}^\top$. Thus we end up with the system

$$\widehat{R}x = \widehat{Q}^\top b$$

which can be solved via *backward substitution*.

Remarks:

- Recall: Using the reduced QR decomposition to solve $Ax = b$ results in the same system. Now we see that for the case $b \notin \mathrm{Im}(A)$, we solve the normal equation.
- If the columns of A are independent, then $\widehat{R}$ is invertible and without loss of generality one can require $r_{ii} > 0$, otherwise one multiplies the i-th column in $\widehat{Q}$ and row in $\widehat{R}$ by "-1". Thus, the factorization $\widehat{R}^\top\widehat{R}$ can be considered a Cholesky decomposition of $A^\top A$. However, the difference here is that we obtain it by the QR decomposition of A and not by applying the Cholesky algorithm to $A^\top A$. The former is roughly twice as expensive. One can show that this additional effort pays off in terms of improved stability against rounding errors.

(3) Using the Pseudoinverse A^+ (see below)
This is typically not done in practice since the computation of the singular value decomposition (which has to be iterative in higher dimensions since we need to solve eigenvalue problems) is more expensive than a direct method. However it offers interesting theoretical insights as we will see below.

(4) Randomized algorithms:
If A^TA is large (n large), then in particular A^TA cannot (and should not) be computed.
In such cases one can use *randomized* algorithms which only work with subsamples of the columns of A (not addressed in this course; see for example [3, II.4])

6.4 Regularization and Minimum Norm Least Squares Solution (enforce uniqueness)

6.4.1 Motivation and Overview

Situation: Columns of A are possibly linearly dependent $(\ker(A) \supsetneq \{0\})$
$\Rightarrow A^TA$ *not* invertible
$\Rightarrow$ there are infinitely many solutions of $A^TAx = A^Tx$ (or if $b \in \mathrm{Im}(A)$, of $Ax = b$)

- Geometric perspective:
  Draw a picture with dependent columns of A and consider an example $b \in \mathrm{Im}(A)$ and an example $b \notin \mathrm{Im}(A)$.
- Algebraic perspective:
  Let $\widehat{x}$ be a solution of the normal equation, then for all $x_0 \in \ker(A)$ we have that $\widehat{x} + x_0$ is also a solution. In fact, we find
  $$A^\top A(\widehat{x} + x_0) = A^\top(A\widehat{x} + Ax_0) = A^\top A\widehat{x}.$$

$\rightarrow$ We say the minimization problem is *ill-posed* ($\neq$ well-posed $=$ existence+uniqueness)

Question: Which solution to pick?

We briefly discuss two approaches: **Tikhonov Regularization** and **Minimum norm solution**

6.4.2 Tikhonov Regularization

Tikhonov Regularization of the least squares problem (*ridge regression*):

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 + \frac{\delta}{2}\|x\|_2^2, \quad \text{for } \delta > 0 \text{ small.} \tag{10}$$

Remarks

- We enforce uniqueness by reformulating the problem. In fact, the idea here is to add a strictly convex regularization term $\frac{\delta}{2}\|x\|_2^2$ to the original objective function (*convexification*).
- The parameter $\delta > 0$ is sometimes called *regularization parameter*. The smaller it is, the closer do we get to the original problem, the more is the minimization of the residual emphasized.
- One can generalize the regularization term to a rather generic strictly convex function. Thereby one can control properties of the solution. For example, choosing the $\|\cdot\|_1$– instead of the $\|\cdot\|_2$–norm enforces sparsity on the solution, which is a desirable feature in many applications.

Characterization of the "regularized" solution

$$x_\delta := \arg\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 + \tfrac{\delta}{2}\|x\|_2^2$$

Theorem 6.4 (*"Regularized" Normal Equation*) *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then $x_\delta \in \mathbb{R}^n$ solves the regularized problem* (10) *if and only if x_δ solves the* ***"regularized" normal equation***

$$(A^TA + \delta I)x = A^Tb. \tag{11}$$

Proof:
Defining the (strictly convex) function $f(x) := \|Ax-b\|_2^2 + \frac{\delta}{2}\|x\|_2^2$, one can show for the sufficient and necessary first-order optimality conditions:

$$f'(x) = 0 \iff (A^TA + \delta I)x = A^Tb.$$

Analysis of the "regularized" normal equation

- The matrix $A^TA + \delta I$ is symmetric and positive definite *for all* $\delta > 0$ and thus **invertible**. More precisely:

  - Symmetry: *(Recall: A matrix B is called symmetric, if $B^T = B$ holds.)*
    $(A^TA + \delta I)^T = (A^TA)^T + \delta I^T = A^TA + \delta I$

  - Positivity *(Recall: A matrix B is called positive definite, if $x^TBx > 0$ holds $\forall x \in \mathbb{R}^n$.)*
    $x^T(A^TA + \delta I)x = \underbrace{x^T(A^TA)x}_{\geq 0} + \underbrace{\delta}_{>0}\underbrace{x^Tx}_{>0} > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$

  Therefore, the equation (11) has the unique solution

  $$x_\delta = (A^TA + \delta I)^{-1}A^T.$$

- The smaller δ, the more is the error minimization emphasized, the more do we approach the normal equation.

- We note that the vector x_δ (for $\delta > 0$) does neither solve $Ax = b$ nor $A^TAx = A^Tb$!

6.4.3 Minimum Norm Solution and the Moore–Penrose Pseudoinverse

Idea: Among the infinitely many solutions we pick the one with *smallest* norm, i.e.,

$$\min_{x \in \hat{S}} \|x\|_2^2 \qquad \left(\hat{S} := \{x \in \mathbb{R}^n : A^T A x = A^T b\}\right). \tag{12}$$

$\rightarrow$ We enforce uniqueness by determining a specific selection criterion.

Characterization of the minimum-norm solution

$$x^+ := \arg\min_{x \in \hat{S}} \|x\|_2^2$$

Theorem 6.5 (*Minimum-Norm Least Squares*) *Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then*

$$x^+ = \lim_{\delta \to 0} x_\delta$$

solves the minimum-norm least squares problem (12). *Here, $x_\delta = (A^T A + \delta I)^{-1} A^T b$ is the solution of the regularized least squares problem* (10).

Proof: Uses the singular value decomposition.

Remarks

- By construction x^+ has two properties:

  1) It is a least squares solution, i.e.,

  $$A^T A x^+ = A^T b \quad \text{(or if } b \in \mathrm{Im}(A), \text{ also } Ax^+ = b).$$

  2) It is the one with smallest norm, i.e.,

  $$\|x^+\|_2 \leq \|\widehat{x}\|_2 \quad \forall\, \widehat{x} \in \hat{S}.$$

- By applying Theorem 6.4 we find

  $$x^+ = \lim_{\delta \to 0} x_\delta = \lim_{\delta \to 0} (A^T A + \delta I)^{-1} A^T b = \left(\lim_{\delta \to 0} (A^T A + \delta I)^{-1} A^T \right) b.$$

- One can show that the limiting matrix

  $$A^+ := \lim_{\delta \to 0} (A^T A + \delta I)^{-1} A^T$$

  is precisely the so-called *Moore–Penrose Pseudoinverse* of A (proof below). With the help of the SVD $U\Sigma V^\top = A$, it can be computed by

  $$A^+ = V \Sigma^+ U^\top,$$

  where the pseudoinverse of a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$ is given by

  $$\Sigma^+ = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right).$$

The Moore–Penrose Pseudoinverse

Let us explain why the term *pseudoinverse* is used here.

Let $A \in \mathbb{R}^{m \times n}$ with $m \neq n$, then A and the corresponding function $f_A \colon \mathbb{R}^n \to \mathbb{R}^m$ can't be invertible. In fact, there can't be a one-to-one relation between the spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ in this case.

$\to$ The two spaces have different dimensions. For instance, a single nonzero vector can explain a line ($\mathbb{R}$), but two independent vectors are needed to explain a plane ($\mathbb{R}^2$).

$\to$ One could say that $\mathbb{R}^n$ and $\mathbb{R}^m$ are "differently large" if $m \neq n$.

However: We still aim at solving systems $Ax = b$ for $A \in \mathbb{R}^{m \times n}$ with possibly $m \neq n$.

Recall: If A is invertible (then in particular $m = n$), then $x = A^{-1}b$ is the unique solution. The inverse is a function which maps the right-hand side to the unique solution.

$\to$ As seen above, the concept of an inverse matrix fails if $m \neq n$.

The minimum-norm least squares solution is a generally applicable concept which maps a right-hand side to "some sort of *unique* solution":

$$x^+ := x^+(b) := \arg \min_{\text{s.t. } A^TAx = A^Tb} \|x\|_2^2 = \left(\lim_{\delta \to 0} (A^TA + \delta I)^{-1} A^T \right) b \qquad (x^+ \text{ uniquely exists!})$$

We finally show

i) The limiting matrix is the Moore–Penrose Pseudoinverse:

$$A^+ := \lim_{\delta \to 0} (A^T A + \delta I)^{-1} A^T = V \Sigma^+ U^\top$$

ii) Applying the Moore–Penrose Pseudoinverse to b gives the minimum-norm least squares solution:

$$x^+ = V \Sigma^+ U^\top b.$$

To i)

Let us consider the SVD $A = U \Sigma V^\top$, then

$$A^T A + \delta I = V(\Sigma^\top \Sigma + \delta^2 I) V^\top,$$

where

$$\Sigma^\top \Sigma + \delta^2 I = \mathrm{diag}(\sigma_1^2 + \delta^2, \ldots, \sigma_r^2 + \delta^2, \delta^2, \ldots, \delta^2)$$

with inverse

$$(\Sigma^\top \Sigma + \delta^2 I)^{-1} = \mathrm{diag}(1/(\sigma_1^2 + \delta^2), \ldots, 1/(\sigma_r^2 + \delta^2), 1/\delta^2, \ldots, 1/\delta^2)$$

Thus

$$(A^T A + \delta I)^{-1} A^T = V[(\Sigma^\top \Sigma + \delta^2 I)^{-1} \Sigma^\top] V^\top$$

where

$$(\Sigma^\top \Sigma + \delta^2 I)^{-1} \Sigma^\top = \mathrm{diag}(\frac{\sigma_1}{\sigma_1^2 + \delta^2}, \ldots, \frac{\sigma_r}{\sigma_r^2 + \delta^2}, 0, \ldots, 0) \quad \to \quad \mathrm{diag}(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0) = \Sigma^+$$

for $\delta \to 0$.

To ii) 1. Let us start with the simple case: $A \in \mathbb{R}^{m\times n}$ diagonal

$$A = \begin{pmatrix} a_{11} & & & & 0 \\ & \ddots & & & \\ & & a_{rr} & & \\ & & & 0 & \\ 0 & & & & 0 \end{pmatrix} \in \mathbb{R}^{m\times n},\ a_{ii} \neq 0, \quad A^TA = \begin{pmatrix} a_{11}^2 & & & & 0 \\ & \ddots & & & \\ & & a_{rr}^2 & & \\ & & & 0 & \\ 0 & & & & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}$$

Normal equation:

$$A^TAx = A^Tb = \begin{pmatrix} a_{11}b_1 \\ \vdots \\ a_{rr}b_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad \begin{array}{c} a_{11}^2x_1 = a_{11}b_1 \\ \vdots \\ a_{rr}^2x_r = a_{rr}b_r \\ 0\cdot x_i = 0 \\ (i > r) \end{array} \quad \Leftarrow \quad \begin{array}{c} x_1 = \frac{b_1}{a_{11}} \\ \vdots \\ x_r = \frac{b_r}{a_{rr}} \\ x_{r+1} = 0 \\ \vdots \\ x_n = 0 \end{array}$$

$$\Rightarrow \quad x^+ = \begin{pmatrix} \frac{1}{a_{11}}b_1 \\ \vdots \\ \frac{1}{a_{rr}}b_r \\ 0 \\ \vdots \end{pmatrix} \quad 0 = A^+b, \quad A^+ = \begin{pmatrix} \frac{1}{a_{11}} & & & & 0 \\ & \ddots & & & \\ & & \frac{1}{a_{rr}} & & \\ & & & 0 & \\ 0 & & & & 0 \end{pmatrix} \in \mathbb{R}^{n\times m}$$

Note: The x_i for $i > r$ can be chosen arbitrarily, but setting them to zero gives the smallest vector.

2. Now we use these ideas for the general case: $A \in \mathbb{R}^{m \times n}$

By using the SVD $A = U\Sigma V^\top$ we find

$$A^T A = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^T \Sigma V^T,$$

so that the normal equation reads as

$$\begin{aligned} (*) \quad A^T Ax = A^T b \quad &\Leftrightarrow \quad V\Sigma^T \Sigma (V^T x) = V\Sigma^T (U^T b) \\ &\overset{V^T \cdot |}{\Leftrightarrow} \quad \underbrace{\Sigma^T \Sigma (V^T x) = \Sigma^T (U^T b)}_{\text{(normal equation for } (\Sigma,\ U^T b))} \quad (\sharp) \end{aligned}$$

Consequently, x solves $(*)$ if and only if $y := V^T x$ solves $(\sharp)$. Since V is orthogonal, both solutions have the same norm, more precisely,

$$\|x\|_2^2 = x^\top x = x^\top (VV^\top) x = \|V^\top x\|_2^2 = \|y\|_2^2.$$

From 1. above on diagonal matrices we know that $y^+ = \Sigma^+ U^T b$ is the smallest solution of $(\sharp)$. Thus, $x^+ := Vy^+ = V\Sigma^+ U^T b = A^+ b$ is the smallest solution of $(*)$, i.e., the minimum norm least squares solution.

All in all: Since orthogonal matrices (here U and V) are not only invertible but also isometric and the SVD $A = U\Sigma V^\top$ always exists, we could rely on the result for diagonal matrices (here Σ).

6.5 Small Tour: Inverse Problems in Imaging

$\rightarrow$ presented in an ipython notebook.