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.

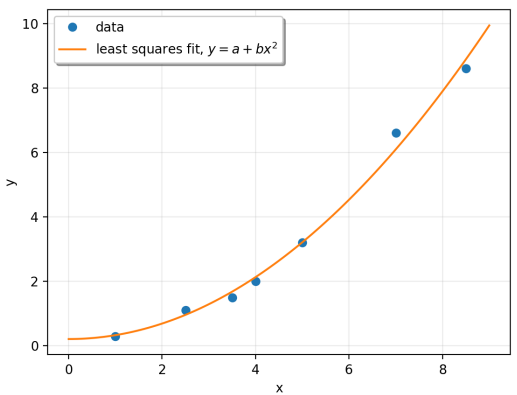
[4] L.N. Trefethen and D. Bau.
Numerical linear algebra.
SIAM, Soc. for Industrial and Applied Math., Philadelphia, 1997.

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 $x \in \mathbb{R}^n$.

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:

(2) Geometric Perspective:

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*

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

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

Analysis of the Normal Equation

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

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

(3) Consistent reformulation:

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

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

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: $A^T A$ often *ill-conditioned* and numerical elimination may fail due to rounding errors!

Can we work without $A^T A$*?* Yes!
(2) Using reduced QR **Decomposition** $A = \widehat{Q}\widehat{R}$

(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 $Ax = b$)

$\rightarrow$ 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*):

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

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:

Analysis of the “regularized” normal equation

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

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 $m \neq n$.

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

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

6.5 Small Tour: Inverse Problems in Imaging

$\rightarrow$ presented in an ipython notebook.