LECTURE 18

Best approximation

What to do when a system has no solution, and how Gauss found ceres

Dr Ilia Pirashvili
Department of Mathematics, University of Galway
4 November 2026 · Winter 2026

Where we got to

LAST LECTURE

- Orthogonal sets are independent; with an orthogonal basis $\mathcal{C} := \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ the coordinates any vector v in $\mathcal{C}$ are $c_i = \mathbf{v} \cdot \mathbf{u}_i / \mathbf{u}_i \cdot \mathbf{u}_i$.
- $\mathrm{Proj}_W(\mathbf{v})$ is the sum of the projections onto the basis lines, and every $\mathbf{v} = \mathrm{Proj}_W(\mathbf{v}) + \mathbf{z}$ with $\mathbf{z} \in W^\perp$, uniquely.
- Gram–Schmidt turns any basis into an orthogonal one.
- An **orthogonal matrix** has $A^{-1} = A^\mathsf{T}$ and preserves all lengths and angles.

One fact remains to be drawn out of the decomposition theorem, and it might just be **the** most useful thing in the module.

Warm-up

WARM-UP

Is there an $\mathbf{x}$ with $\begin{bmatrix}1\\1\end{bmatrix} x = \begin{bmatrix}2\\3\end{bmatrix}$? ▸ If not, which multiple of $\begin{bmatrix}1\\1\end{bmatrix}$ comes closest to $\begin{bmatrix}2\\3\end{bmatrix}$ ▸?

Warm-up

WARM-UP

Is there an $\mathbf{x}$ with $\begin{bmatrix}1\\1\end{bmatrix} x = \begin{bmatrix}2\\3\end{bmatrix}$? ▸ If not, which multiple of $\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ comes closest to $\left[\begin{smallmatrix}2\\3\end{smallmatrix}\right]$ ▸?

SOLUTION

Clearly no, the rows force $x = 2$ and $x = 3$ simultaniously; a contradiction. The system is inconsistent, as $\left[\begin{smallmatrix}2\\3\end{smallmatrix}\right] \notin \operatorname{Im} A$. For the closest multiple, project: $\frac{5}{2}\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right] = \left[\begin{smallmatrix}2.5\\2.5\end{smallmatrix}\right]$.

Here, $\frac{5}{2}$ comes from $\frac{\left[\begin{smallmatrix}2\\3\end{smallmatrix}\right] \cdot \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]}{\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right] \cdot \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]}$. That second step is the foundation of today's lecture.

SECTION 1

The best approximation theorem

Projection minimises distance

THEOREM 18.1 ▸ · Best Approximation

Let $W \subseteq \mathbb{R}^n$ be a subspace and $\mathbf{v} \in \mathbb{R}^n$. Then $\hat{\mathbf{v}} = \text{Proj}_W(\mathbf{v})$ is the unique closest point of W to $\mathbf{v}$. That is, for every $\mathbf{w} \in W$ with $\mathbf{w} \neq \hat{\mathbf{v}}$,

$$\|\mathbf{v} - \hat{\mathbf{v}}\| \; < \; \|\mathbf{v} - \mathbf{w}\|.$$

Projection minimises distance

THEOREM 18.1 ▸ · Best Approximation

Let $W \subseteq \mathbb{R}^n$ be a subspace and $\mathbf{v} \in \mathbb{R}^n$. Then $\hat{\mathbf{v}} = \mathrm{Proj}_W(\mathbf{v})$ is the unique closest point of W to $\mathbf{v}$. That is, for every $\mathbf{w} \in W$ with $\mathbf{w} \neq \hat{\mathbf{v}}$,

$$\|\mathbf{v} - \hat{\mathbf{v}}\| \; < \; \|\mathbf{v} - \mathbf{w}\|.$$

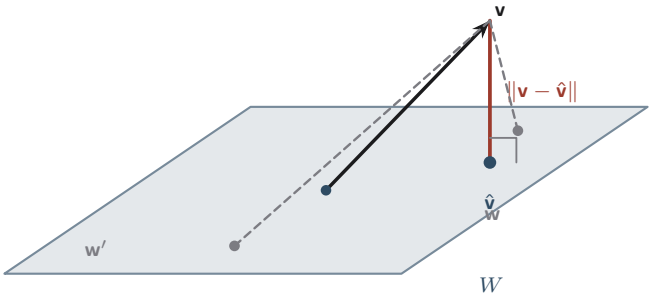
PROOF.

Write $\mathbf{v} - \mathbf{w} = (\mathbf{v} - \hat{\mathbf{v}}) + (\hat{\mathbf{v}} - \mathbf{w})$. The first bracket lies in $W^\perp$ (definition of orthogonality), the second in W ($\hat{\mathbf{v}} \in W$ and $\mathbf{w} \in W$), so they are orthogonal and Pythagoras applies:

$$\|\mathbf{v} - \mathbf{w}\|^2 \; = \; \|\mathbf{v} - \hat{\mathbf{v}}\|^2 + \|\hat{\mathbf{v}} - \mathbf{w}\|^2.$$

The last term is > 0 unless $\mathbf{w} = \hat{\mathbf{v}}$. □

Why the projection minimises distance: illustrated



Any other point of W reaches $\mathbf{v}$ along the hypotenuse of a right triangle whose vertical leg is already $\|\mathbf{v} - \hat{\mathbf{v}}\|$. The hypotenuse is longer, always.

SECTION 2

When a system has no solution

The situation ▸

In practice one often meets $A\mathbf{x} = \mathbf{b}$ with *more equations than unknowns*, and they do not just linear combinations that can be discarded. Real life measurements are never precise, and we do not just want to randomly throw out data, especialy since we do not know which is "right" and which is "wrong".

The situation ▸

In practice one often meets $A\mathbf{x} = \mathbf{b}$ with *more equations than unknowns*, and they do not just linear combinations that can be discarded. Real life measurements are never precise, and we do not just want to randomly throw out data, especialy since we do not know which is "right" and which is "wrong".

LEAST-SQUARE METHOD

If $A\mathbf{x} = \mathbf{b}$ is not solvable (contradictory), ask for the next best thing: The $\hat{\mathbf{x}}$ making $A\hat{\mathbf{x}}$ as close to $\mathbf{b}$ as possible, meaning

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\| \qquad \text{for all } \mathbf{x}.$$

Such an $\hat{\mathbf{x}}$ is called the **least-squares solution**.

The situation ▸

In practice one often meets $A\mathbf{x} = \mathbf{b}$ with *more equations than unknowns*, and they do not just linear combinations that can be discarded. Real life measurements are never precise, and we do not just want to randomly throw out data, especialy since we do not know which is "right" and which is "wrong".

LEAST-SQUARE METHOD

If $A\mathbf{x} = \mathbf{b}$ is not solvable (contradictory), ask for the next best thing: The $\hat{\mathbf{x}}$ making $A\hat{\mathbf{x}}$ as close to $\mathbf{b}$ as possible, meaning

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\| \qquad \text{for all } \mathbf{x}.$$

Such an $\hat{\mathbf{x}}$ is called the **least-squares solution**.

As the name implies, it is minimising $\|\mathbf{b} - A\mathbf{x}\|$ is the *sum of the squares* of its entries.

Reducing it to a projection ▸

As $\mathbf{x}$ runs over $\mathbb{R}^n$, the vector $A\mathbf{x}$ runs over exactly $\operatorname{Im} A$. So we are asking:

> *Which point of* $\operatorname{Im} A$ *is closest to* $\mathbf{b}$*?*

Reducing it to a projection ▸

As $\mathbf{x}$ runs over $\mathbb{R}^n$, the vector $A\mathbf{x}$ runs over exactly $\operatorname{Im} A$. So we are asking:

> *Which point of* $\operatorname{Im} A$ *is closest to* $\mathbf{b}$*?*

The Best Approximation Theorem answers it as the point $\operatorname{Proj}_{\operatorname{Im} A}(\mathbf{b})$.

KEY IDEA

So $\hat{\mathbf{x}}$ is a least-squares solution exactly when

$$A\hat{\mathbf{x}} = \operatorname{Proj}_{\operatorname{Im} A}(\mathbf{b}),$$

and this system *is* consistent, by construction.

The normal equations

We do not have to compute the projection, which needs an orthogonal basis of $\operatorname{Im} A$.

The normal equations

We do not have to compute the projection, which needs an orthogonal basis of $\operatorname{Im} A$. The residual $\mathbf{r} := \mathbf{b} - A\hat{\mathbf{x}}$ is orthogonal to $\operatorname{Im} A$, hence to the columns of A as they span it, i.e.,

$$\mathbf{a}_1 \cdot \mathbf{r} = 0, \qquad \mathbf{a}_2 \cdot \mathbf{r} = 0, \quad \dots \quad \mathbf{a}_n \cdot \mathbf{r} = 0$$

As the rows of A^{T} are exactly the columns of A, we have

$$A^{\mathsf{T}}\mathbf{r} = \begin{bmatrix} \mathbf{a}_1^{\mathsf{T}}\mathbf{r} \\ \mathbf{a}_2^{\mathsf{T}}\mathbf{r} \\ \vdots \\ \mathbf{a}_n^{\mathsf{T}}\mathbf{r} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 \cdot \mathbf{r} \\ \mathbf{a}_2 \cdot \mathbf{r} \\ \vdots \\ \mathbf{a}_n \cdot \mathbf{r} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

That is to say,

$$A^{\mathsf{T}}\mathbf{r} = 0 \iff A^{\mathsf{T}}(\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0} \iff \boxed{A^{\mathsf{T}}\mathbf{b} = A^{\mathsf{T}}A\hat{\mathbf{x}}}.$$

The main formula

THEOREM 18.2 ▸ · Normal equations

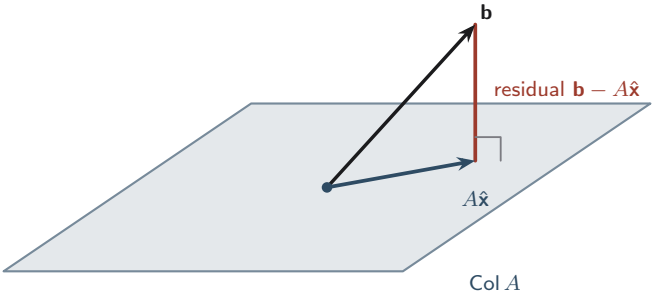
$\hat{\mathbf{x}}$ *is a least-squares solution of* $A\mathbf{x} = \mathbf{b}$ *if and only if*

$$A^{\mathsf{T}} A\,\hat{\mathbf{x}} \;=\; A^{\mathsf{T}}\mathbf{b}.$$

If the columns of A *are independent,* $A^{\mathsf{T}}A$ *is invertible and* $\hat{\mathbf{x}} = (A^{\mathsf{T}}A)^{-1}A^{\mathsf{T}}\mathbf{b}$ *is unique.*

A system of n equations in n unknowns, however many data points were given.

In pictures



b is out of reach; $A\hat{\mathbf{x}}$ is the nearest thing to it inside Im A. The residual is perpendicular to everything we could have reached. That perpendicularity *is* the transpose equation.

SECTION 3

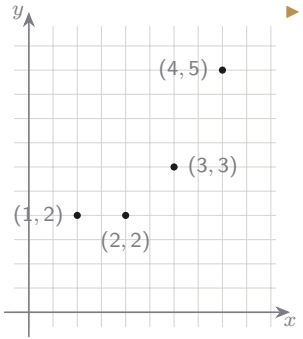
Lines of best fit

The problem, in its most common form

Four measurements: $(1,2)$, $(2,2)$, $(3,3)$, $(4,5)$. We want the line $y = c + dx$ fitting them best.

An exact fit would need $c + dx_i = y_i$ for each i, that is

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{A} \begin{bmatrix} c \\ d \end{bmatrix} = \underbrace{\begin{bmatrix} 2 \\ 2 \\ 3 \\ 5 \end{bmatrix}}_{\mathbf{b}}.$$



Visually, we can see that these four points do not lie on a common line.

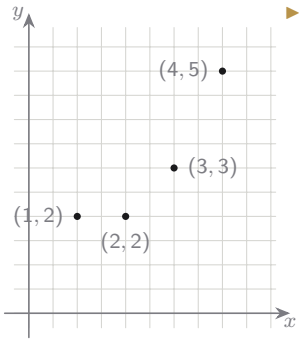
The problem, in its most common form

Four measurements: $(1,2)$, $(2,2)$, $(3,3)$, $(4,5)$. We want the line $y = c + dx$ fitting them best.

An exact fit would need $c + dx_i = y_i$ for each i, that is

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{A} \begin{bmatrix} c \\ d \end{bmatrix} = \underbrace{\begin{bmatrix} 2 \\ 2 \\ 3 \\ 5 \end{bmatrix}}_{\mathbf{b}}.$$

Four equations, two unknowns ▸. It is immediately that no line passes through all four points, so this is inconsistent.



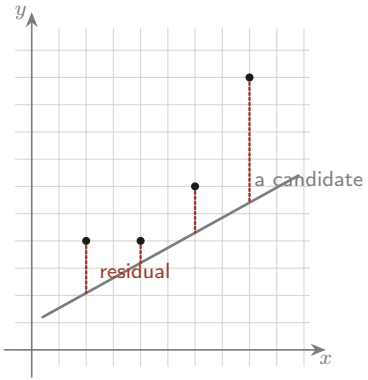
Visually, we can see that these four points do not lie on a common line.

What is being minimised

For a candidate line, the i-th **residual** is the vertical gap between the data point and the line. Least squares minimises ▸

$$\sum_i (\text{residual}_i)^2,$$

which is exactly $\|\mathbf{b} - A\mathbf{x}\|^2$.



The last remark about projection would require a few more words to fully explain.

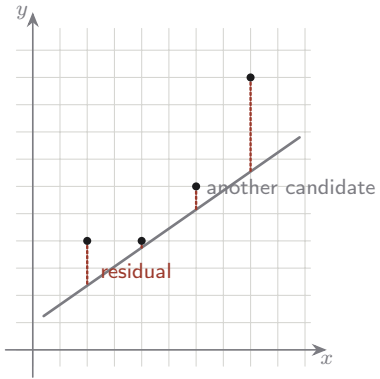
What is being minimised

For a candidate line, the i-th **residual** is the vertical gap between the data point and the line. Least squares minimises ▸

$$\sum_i (\text{residual}_i)^2,$$

which is exactly $\|\mathbf{b} - A\mathbf{x}\|^2$.

Squaring stops signs cancelling and punishes large gaps disproportionately. It is also what makes the answer a *projection*, which is why this measure of error is the one we can solve.



The last remark about projection would require a few more words to fully explain.

Solving it (1/2) ▸

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 3 \\ 5 \end{bmatrix}.$$

Solving it (1/2) ▸

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 3 \\ 5 \end{bmatrix}.$$

$$A^{\mathsf{T}}A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}, \qquad A^{\mathsf{T}}\mathbf{b} = \begin{bmatrix} 12 \\ 35 \end{bmatrix}.$$

Solving it (2/2) ▸

The normal equations $A^{\mathsf{T}}A\hat{\mathbf{x}} = A^{\mathsf{T}}\mathbf{b}$ read

$$4c + 10d = 12, \qquad 10c + 30d = 35.$$

Solving it (2/2) ▸

The normal equations $A^{\mathsf{T}}A\hat{\mathbf{x}} = A^{\mathsf{T}}\mathbf{b}$ read

$$4c + 10d = 12, \qquad 10c + 30d = 35.$$

From the first, $c = 3 - 2.5d$; substituting into the second,

$$10(3 - 2.5d) + 30d = 35 \implies 30 + 5d = 35 \implies d = 1.$$

Solving it (2/2) ▸

The normal equations $A^{\mathsf{T}}A\hat{\mathbf{x}} = A^{\mathsf{T}}\mathbf{b}$ read

$$4c + 10d \;=\; 12, \qquad 10c + 30d \;=\; 35.$$

From the first, $c = 3 - 2.5d$; substituting into the second,

$$10(3 - 2.5d) + 30d \;=\; 35 \quad\Longrightarrow\quad 30 + 5d \;=\; 35 \quad\Longrightarrow\quad d \;=\; 1.$$

Then $c = 3 - 2.5(1) = 0.5$.
Check in the first equation: $4(0.5) + 10(1) = 12$. **Yes.**

Solving it (2/2) ▸

The normal equations $A^{\mathsf{T}}A\hat{\mathbf{x}} = A^{\mathsf{T}}\mathbf{b}$ read

$$4c + 10d \;=\; 12, \qquad 10c + 30d \;=\; 35.$$

From the first, $c = 3 - 2.5d$; substituting into the second,

$$10(3 - 2.5d) + 30d \;=\; 35 \quad\Longrightarrow\quad 30 + 5d \;=\; 35 \quad\Longrightarrow\quad d \;=\; 1.$$

Then $c = 3 - 2.5(1) = 0.5$.

Check in the first equation: $4(0.5) + 10(1) = 12$. **Yes.**

So the line of best fit is

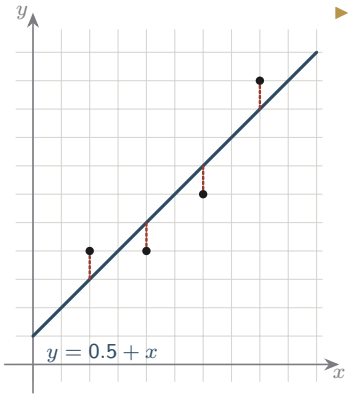
$$y \;=\; 0.5 + x.$$

The answer, drawn

$$y = 0.5 + x$$

Predicted values 1.5, 2.5, 3.5, 4.5 against observed $2, 2, 3, 5$, so the residuals are

$$0.5, \quad -0.5, \quad -0.5, \quad 0.5.$$



The answer, drawn

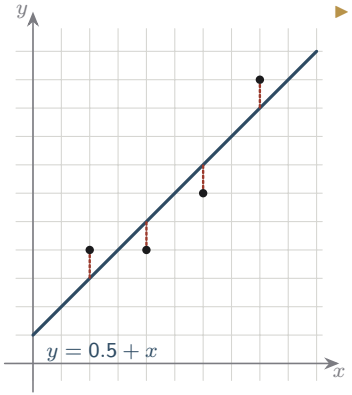
$$y = 0.5 + x$$

Predicted values 1.5, 2.5, 3.5, 4.5 against observed $2, 2, 3, 5$, so the residuals are

$$0.5, \quad -0.5, \quad -0.5, \quad 0.5.$$

Their sum of squares is $4 \times 0.25 = 1$. It is not possible to find a line closer to all of the points.

Notice the residuals sum to 0: that is the first normal equation, $\sum(y_i - c - dx_i) = 0$.



Where this is used

- **Linear regression**: the tried and true method of statistics and econometrics. The "line of best fit" in a scientific paper is what we did above.
- **Data science**: fitting a model to many more observations than parameters is exactly the over-determined case. See also curve fitting in the next chapter.
- **Principal component analysis**: finding the subspace that best approximates a cloud of data. This is least squares with the subspace itself as the unknown.
- **Calibration and signal processing**: recovering a few parameters from a great many noisy readings.

Where this is used

- **Linear regression**: the tried and true method of statistics and econometrics. The "line of best fit" in a scientific paper is what we did above.
- **Data science**: fitting a model to many more observations than parameters is exactly the over-determined case. See also curve fitting in the next chapter.
- **Principal component analysis**: finding the subspace that best approximates a cloud of data. This is least squares with the subspace itself as the unknown.
- **Calibration and signal processing**: recovering a few parameters from a great many noisy readings.

In every case the arithmetic is what we just did; only the interpretation changes.

SECTION 4

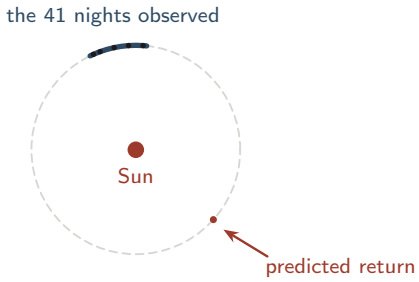
Where it came from

Ceres, 1801

On 1 January 1801 Giuseppe Piazzi spotted a new object between Mars and Jupiter. He tracked it for forty-one nights — a few degrees of its orbit — and then it passed behind the Sun and was lost.

Finding it again meant predicting where it would re-emerge, from a short arc of imprecise measurements.

It became a prestigious question that the foremost astronomers and mathematicians of that time tried to solve.



Carl Friedrich Gauss, then twenty-four, computed an orbit and in December 1801, Ceres was found again, close to his prediction. Noone else came close enough to find it.

NOT EXAMINABLE

What Gauss actually did ▸

His method, published in 1809, though he already used it since 1795, was the least squares.

- An orbit is fixed by a few parameters. Each observation gives one equation in them.
- There were far more observations than parameters, each carrying error, so no exact solution existed.
- Rather than discard data, use *all* of it, minimising the sum of squared residuals.

Nothing in the method required a line, as the example of ceres shows. We just need finitely many parameters appearing linearly. Note it does not mean that the curve has to be linear. For example $y = c_0 + c_1 x + c_2 x^2$ has the coefficients c_0, c_1, c_2 as parameters, which are linear. On the other hand $y = c_0 \cdot c_1 + c_1 x + c_2 x^2$ would not work.

NOT EXAMINABLE

Fitting a curve instead of a line ▸

To fit $y = c_0 + c_1 x + c_2 x^2$, treat the c_i as the unknowns. Its a linear system of 3 unknowns. Write one row per data point:

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

NOT EXAMINABLE

Fitting a curve instead of a line ▸

To fit $y = c_0 + c_1 x + c_2 x^2$, treat the c_i as the unknowns. Its a linear system of 3 unknowns. Write one row per data point:

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Then solve $A^\mathsf{T} A\hat{\mathbf{x}} = A^\mathsf{T}\mathbf{b}$, exactly as before.

DECREASE THE DEGREE OF A CURVE BY INCREASING THE VARIABLES

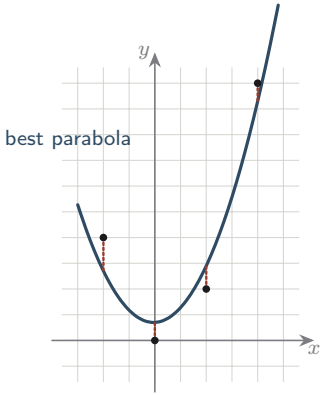
The curve may be non-linear, what must be linear is the dependence on the *unknowns*. They indeed appear only linearly and that is why this is still linear algebra.

NOT EXAMINABLE

A quadratic fit, drawn

The points $(-1, 2)$, $(0, 0)$, $(1, 1)$, $(2, 5)$ lie on no parabola. The same machinery gives the best one.

Here $A^{\mathsf{T}}A$ is 3×3 and the arithmetic is longer, but no idea is new: build A, form $A^{\mathsf{T}}A$ and $A^{\mathsf{T}}\mathbf{b}$, solve.



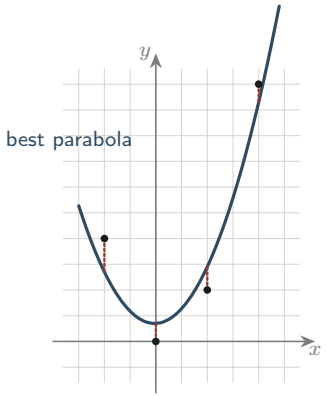
NOT EXAMINABLE

A quadratic fit, drawn

The points $(-1, 2)$, $(0, 0)$, $(1, 1)$, $(2, 5)$ lie on no parabola. The same machinery gives the best one.

Here $A^{\mathsf{T}}A$ is 3×3 and the arithmetic is longer, but no idea is new: build A, form $A^{\mathsf{T}}A$ and $A^{\mathsf{T}}\mathbf{b}$, solve.

Gauss's case was of this kind: an orbit under gravity is a conic section, fitted from many more observations than unknowns.



NOT EXAMINABLE

Back on the syllabus

KEY IDEA

Examinable: the Best Approximation Theorem, the normal equations $A^{\mathsf{T}}A\hat{\mathbf{x}} = A^{\mathsf{T}}\mathbf{b}$, and fitting a *line* to data.

The historical material and the quadratic fit are context, not requirements.

Back on the syllabus

KEY IDEA

Examinable: the Best Approximation Theorem, the normal equations $A^{\mathsf{T}}A\hat{\mathbf{x}} = A^{\mathsf{T}}\mathbf{b}$, and fitting a *line* to data.

The historical material and the quadratic fit are context, not requirements.

One practical caution before we finish: $A^{\mathsf{T}}A$ can be badly behaved numerically, and forming it explicitly is not how good software proceeds. There are better routes — through orthogonal bases, and later through the singular value decomposition — but they compute the same $\hat{\mathbf{x}}$.

We will meet some of that machinery in the coming weeks.

SECTION 5

Two more examples

An inconsistent system, solved ▸

Find the least-squares solution of

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

An inconsistent system, solved ▸

Find the least-squares solution of

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

$$A^{\mathsf{T}}A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad A^{\mathsf{T}}\mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

An inconsistent system, solved ▸

Find the least-squares solution of

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

$$A^{\mathsf{T}}A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad A^{\mathsf{T}}\mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

$$2\hat{x}_1 + \hat{x}_2 = 1, \quad \hat{x}_1 + 2\hat{x}_2 = 1 \qquad \Longrightarrow \qquad \hat{\mathbf{x}} = \begin{bmatrix} 1/3 \\ 1/3 \end{bmatrix}.$$

An inconsistent system, solved ▸

Find the least-squares solution of

$$A\mathbf{x} = \mathbf{b}, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

$$A^{\mathsf{T}}A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad A^{\mathsf{T}}\mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

$$2\hat{x}_1 + \hat{x}_2 = 1, \quad \hat{x}_1 + 2\hat{x}_2 = 1 \qquad \Longrightarrow \qquad \hat{\mathbf{x}} = \begin{bmatrix} 1/3 \\ 1/3 \end{bmatrix}.$$

Then $A\hat{\mathbf{x}} = \begin{bmatrix} 1/3 \\ 1/3 \\ 2/3 \end{bmatrix}$ and the residual is $\begin{bmatrix} 2/3 \\ 2/3 \\ -2/3 \end{bmatrix}$, orthogonal to both columns. **Yes.**

When the answer is not unique

EXAMPLE 18.3

Take $A = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$ and $\mathbf{b} = [\frac{1}{3}]$. The columns are dependent, so $A^{\mathsf{T}}A = \begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix}$ is singular.

When the answer is not unique

EXAMPLE 18.3

Take $A = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$. The columns are dependent, so $A^{\mathsf{T}}A = \begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix}$ is singular.

The normal equations $2\hat{x}_1 + 4\hat{x}_2 = 4$ (twice over) have infinitely many solutions ▸ — for instance $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

When the answer is not unique

EXAMPLE 18.3

Take $A = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$. The columns are dependent, so $A^{\mathsf{T}}A = \begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix}$ is singular.

The normal equations $2\hat{x}_1 + 4\hat{x}_2 = 4$ (twice over) have infinitely many solutions ▸ — for instance $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

KEY IDEA

The *projection* $A\hat{\mathbf{x}}$ is always unique, by the Best Approximation Theorem. What can fail to be unique is the $\hat{\mathbf{x}}$ producing it — and that happens exactly when $\operatorname{Ker} A \neq \{\mathbf{0}\}$.

Both choices above give $A\hat{\mathbf{x}} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$, as they must.

Summary of Lecture 18

BEST APPROXIMATION	$\text{Proj}_W(\mathbf{v})$ is the unique closest point of W to $\mathbf{v}$. Proof: Pythagoras on $\mathbf{v}-\mathbf{w}=(\mathbf{v}-\hat{\mathbf{v}})+(\hat{\mathbf{v}}-\mathbf{w})$.
LEAST SQUARES	When $A\mathbf{x}=\mathbf{b}$ is inconsistent, seek $\hat{\mathbf{x}}$ minimising $\|\mathbf{b}-A\mathbf{x}\|$; equivalently $A\hat{\mathbf{x}}=\text{Proj}_{\text{Col}\,A}(\mathbf{b})$.
NORMAL EQUATIONS	$A^{\mathsf{T}}A\hat{\mathbf{x}}=A^{\mathsf{T}}\mathbf{b}$, because the residual is orthogonal to every column. Independent columns $\Rightarrow$ $A^{\mathsf{T}}A$ invertible and $\hat{\mathbf{x}}$ unique.
LINE OF BEST FIT	Rows $(1,x_i)$, right-hand side y_i; then $A^{\mathsf{T}}A$ collects m, $\sum x_i$, $\sum x_i^2$ and $A^{\mathsf{T}}\mathbf{b}$ collects $\sum y_i$, $\sum x_iy_i$.
UNIQUENESS	$A\hat{\mathbf{x}}$ is always unique; $\hat{\mathbf{x}}$ is unique exactly when $\text{Ker}\,A=\{\mathbf{0}\}$.

Before next time

AT HOME

1. Find the least-squares solution of $A\mathbf{x} = \mathbf{b}$ ▸ for $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$, and verify the residual is orthogonal to both columns.
2. Find the line of best fit through $(0, 1)$, $(1, 1)$, $(2, 4)$, $(3, 4)$, and compute the sum of squared residuals.
3. Let $W = \text{Span}\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right\}$ and $\mathbf{v} = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}$. Find the point of W closest to $\mathbf{v}$ ▸, and the distance from $\mathbf{v}$ to W.

END OF LECTURE 18

ilia.pirashvili@universityofgalway.ac.ie

Solution: question 1 ▸

SOLUTION

$$A^{\mathsf{T}}A = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \qquad A^{\mathsf{T}}\mathbf{b} = \begin{bmatrix} 6 \\ 13 \end{bmatrix}.$$

So $3c + 6d = 6$ and $6c + 14d = 13$. Doubling the first and subtracting gives $2d = 1$, so $d = 0.5$ and then $c = 1$.

Solution: question 1 ▸

SOLUTION

$$A^{\mathsf{T}}A = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \qquad A^{\mathsf{T}}\mathbf{b} = \begin{bmatrix} 6 \\ 13 \end{bmatrix}.$$

So $3c + 6d = 6$ and $6c + 14d = 13$. Doubling the first and subtracting gives $2d = 1$, so $d = 0.5$ and then $c = 1$.

Then $A\hat{\mathbf{x}} = \begin{bmatrix} 1.5 \\ 2 \\ 2.5 \end{bmatrix}$, and the residual is $\begin{bmatrix} 0.5 \\ -1 \\ 0.5 \end{bmatrix}$. Its inner product with $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ is 0, and with $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ is $0.5 - 2 + 1.5 = 0$. **Yes.**

Solution: question 2 ▸

SOLUTION

With $x_i = 0, 1, 2, 3$ and $y_i = 1, 1, 4, 4$: $m = 4$, $\sum x_i = 6$, $\sum x_i^2 = 14$, $\sum y_i = 10$, $\sum x_i y_i = 21$. So

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 10 \\ 21 \end{bmatrix}.$$

Solution: question 2 ▸

SOLUTION

With $x_i = 0, 1, 2, 3$ and $y_i = 1, 1, 4, 4$: $m = 4$, $\sum x_i = 6$, $\sum x_i^2 = 14$, $\sum y_i = 10$, $\sum x_i y_i = 21$. So

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 10 \\ 21 \end{bmatrix}.$$

From the first, $c = (10 - 6d)/4$; substituting, $15 - 9d + 14d = 21$, so $d = 1.2$ and $c = 0.7$.

Solution: question 2 ▸

SOLUTION

With $x_i = 0, 1, 2, 3$ and $y_i = 1, 1, 4, 4$: $m = 4$, $\sum x_i = 6$, $\sum x_i^2 = 14$, $\sum y_i = 10$, $\sum x_i y_i = 21$. So

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 10 \\ 21 \end{bmatrix}.$$

From the first, $c = (10 - 6d)/4$; substituting, $15 - 9d + 14d = 21$, so $d = 1.2$ and $c = 0.7$. The line is $y = 0.7 + 1.2x$, with predictions $0.7, 1.9, 3.1, 4.3$ and residuals $0.3, -0.9, 0.9, -0.3$; the sum of squares is 1.8.

SOLUTION

W is a line, so we project onto it:

$$\text{Proj}_W(\mathbf{v}) = \frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} = \frac{3+1+0}{2}\begin{bmatrix}1\\1\\0\end{bmatrix} = \begin{bmatrix}2\\2\\0\end{bmatrix}.$$

That is the closest point of W, by the Best Approximation Theorem.

Solution: question 3 ▸

SOLUTION

W is a line, so we project onto it:

$$\text{Proj}_W(\mathbf{v}) = \frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} = \frac{3+1+0}{2}\begin{bmatrix}1\\1\\0\end{bmatrix} = \begin{bmatrix}2\\2\\0\end{bmatrix}.$$

That is the closest point of W, by the Best Approximation Theorem.

The distance is $\|\mathbf{v} - \text{Proj}_W(\mathbf{v})\| = \|\begin{bmatrix}1\\-1\\4\end{bmatrix}\| = \sqrt{18} = 3\sqrt{2}$.

The residual $\begin{bmatrix}1\\-1\\4\end{bmatrix}$ is orthogonal to $\begin{bmatrix}1\\1\\0\end{bmatrix}$, as it must be.