




Machine Learning for Physical Sensing & Perception

CMSC 818V : Fall 2026

Lecture 3: Inverse Problems -- From measurements to Physical States

Nirupam Roy

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Tu-Th 12:30 – 1:45pm

IRB 2107



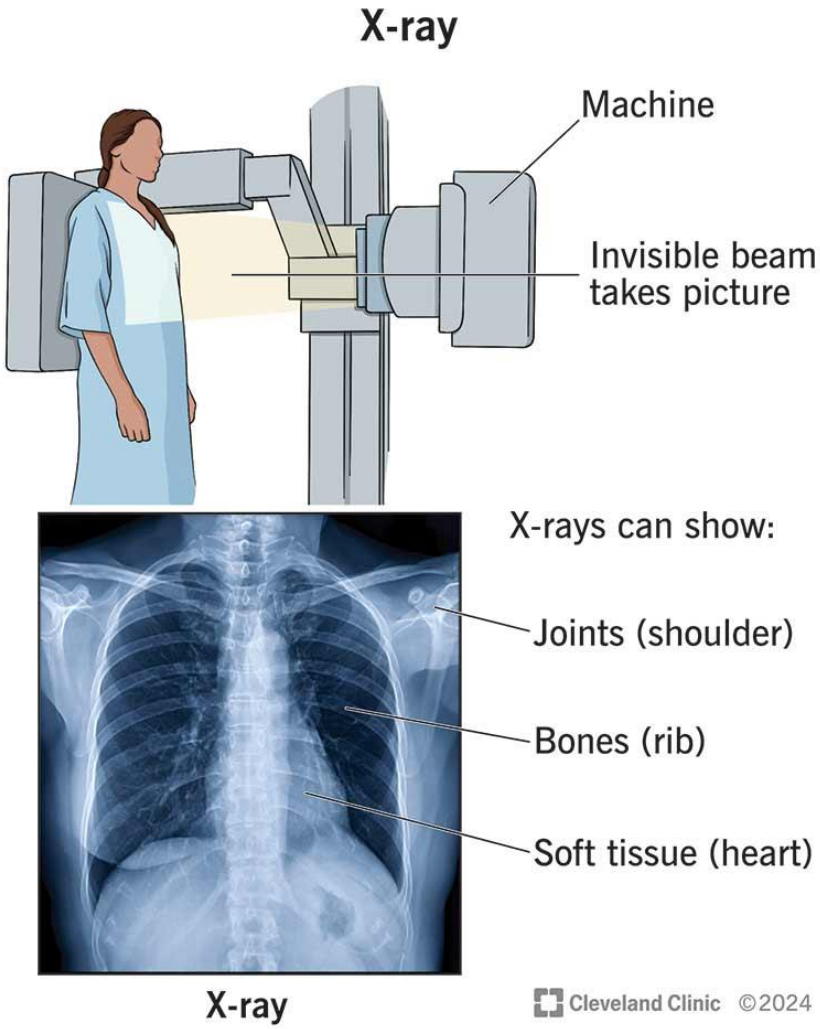




Computed Tomography



Computed Tomography: From 2D projection to 3D cross-sections

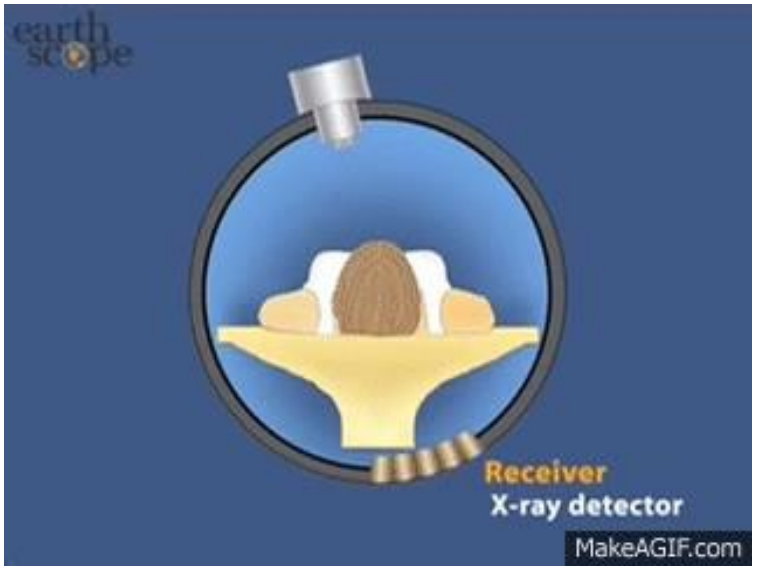


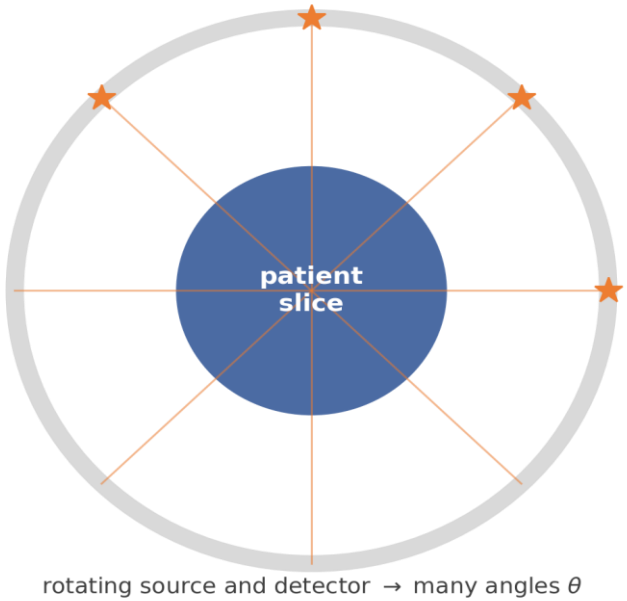
The **Hounsfield scale** is the standardized numerical scale used in CT images to represent the X-ray attenuation of different materials/tissues relative to water.

Hounsfield Unit (HU)

$$HU = 1000 \frac{\mu_{\text{tissue}} - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}}$$

The central question





A CT detector never measures this image directly.

So how can measurements taken outside the body tell us what is inside it?

Today we build a CT scanner from scratch — and mathematical ideas in this lecture solves a problem that scanner runs into.

How the lecture builds, one problem at a time

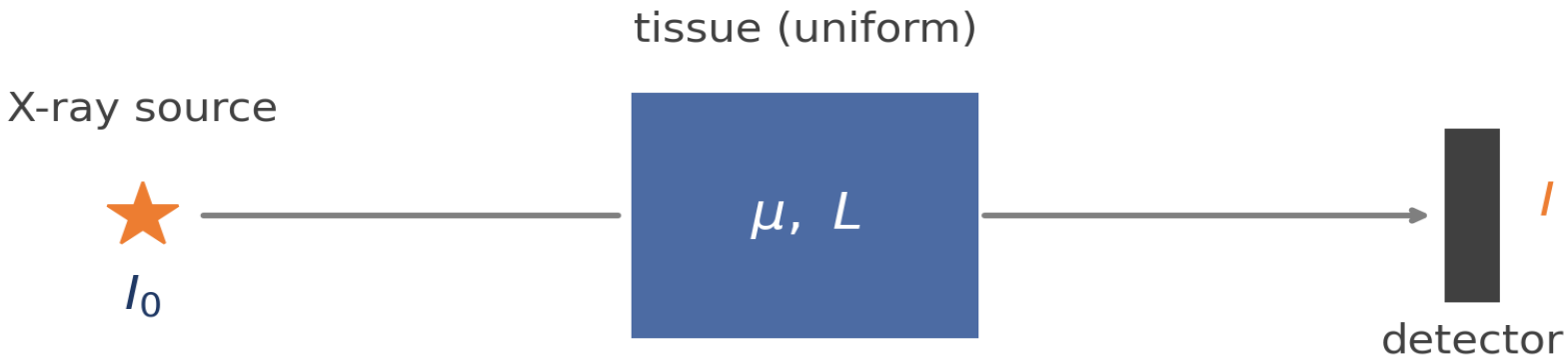
X-ray physics	*what does a detector measure?*
Forward model y = Ax	*discretize the body into voxels*
Identifiability	*can the scanner even see the object?*
Noise y = Ax + n	*measurements are never exact*
ML → least squares	*which image best explains the data?*
Conditioning → SVD	*which patterns are seen strongly, weakly, not at all?*
Prior → MAP → regularization	*what to do when data alone isn't enough*
Learned reconstruction	→ *bridge to the rest of the course*

Every concept below is introduced because the CT scanner we are building needs it.

1. What does an X-ray detector (the sensor) measure?

The simplest possible patient: one uniform slab

Forward model: Beer-Lambert Law



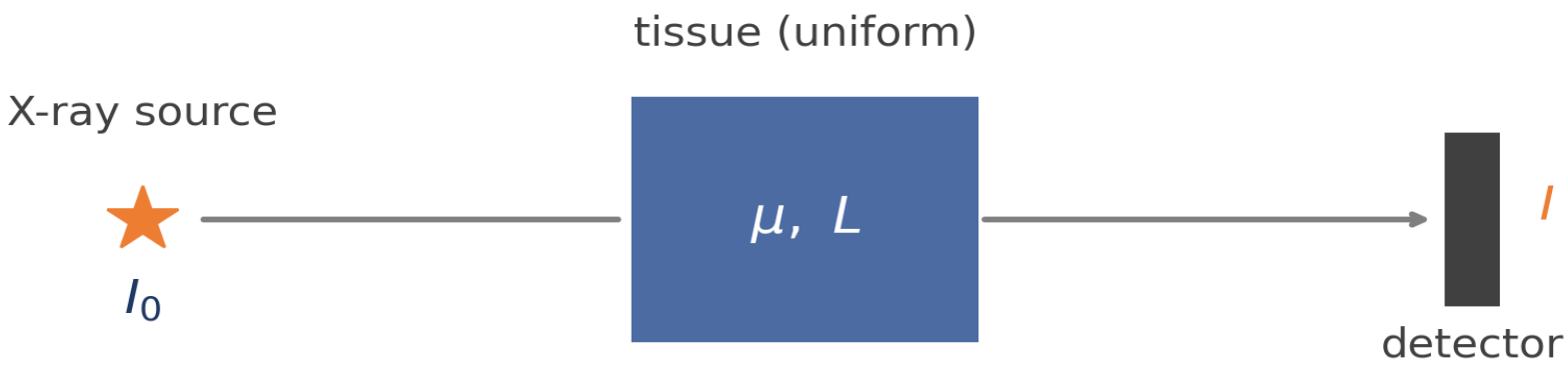
The **Beer–Lambert law** states that the intensity of radiation passing through an absorbing material decreases exponentially with the material's attenuation coefficient and the distance traveled through the material.

$$I = I_0 e^{-\mu L}$$

I_0: incident X-ray intensity
I: measured X-ray intensity after passing through the material
μ: linear attenuation coefficient of the material $[\mathrm{length}^{-1}]$
L: path length through the material

The physical interpretation of the **exponential**: every additional unit of distance removes approximately the same *fraction* of the photons that remain, rather than subtracting the same absolute amount.

Forward model: Beer-Lambert Law



The **Beer–Lambert law** states that the intensity of radiation passing through an absorbing material decreases exponentially with the material's attenuation coefficient and the distance traveled through the material.

$$I = I_0 e^{-\mu L}$$

I_0: incident X-ray intensity
I: measured X-ray intensity after passing through the material
μ: linear attenuation coefficient of the material $[\mathrm{length}^{-1}]$
L: path length through the material

Log transformed measurements: $$y \equiv -\ln\left(\frac{I}{I_0}\right) = \mu L$$

Forward vs. Inverse

Forward problem

$\mu \rightarrow y$

"If I knew the material, what would the detector read?"

Inverse problem

$y \rightarrow \mu$

"I measured the attenuation. What material caused it?"

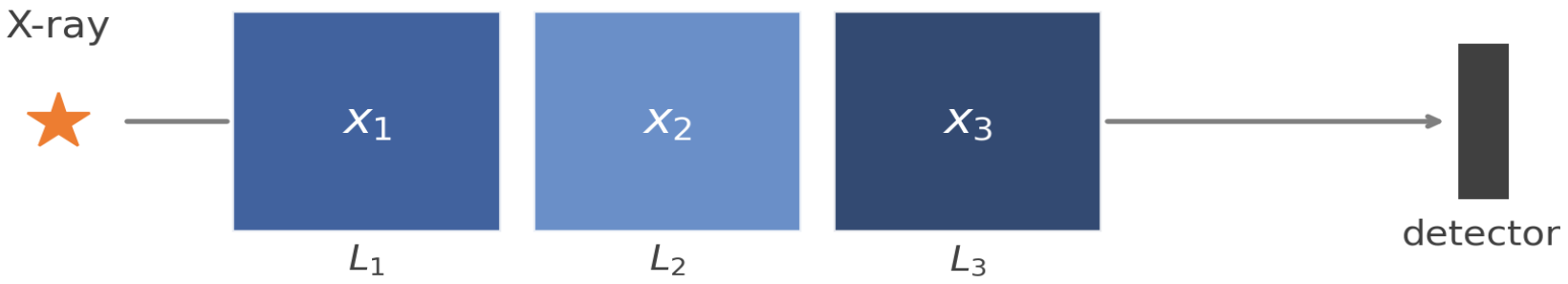
latent physical state → **physics** → **sensor measurement**

This chain — physical state → physics → measurement — is the pattern every sensing problem in this course follows.

2. The body is spatially varying

One ray through several tissues

Several tissues along one ray



$$I = I_0\, e^{-(L1\, x1 + L2\, x2 + L3\, x3)}$$

$$I = I_0 \exp\left(-\int_{\text{ray}} \mu(\mathbf{r})\, d\ell\right)$$

Taking logs of both sides collapses the physics into one linear equation:

$$y = L_1 x_1 + L_2 x_2 + L_3 x_3 = a^T \mathbf{x}$$

$$y = -\ln \frac{I}{I_0} = \int_{\text{ray}} \mu(\mathbf{r})\, d\ell.$$

One X-ray = one linear equation. To reconstruct an image, we will need many such equations.

Discretization: from a continuous field to a vector x

A continuous attenuation field μ(x, y) is approximated by a grid of voxels:

$$x = [x_1, x_2, \ldots, x_N]^T \qquad \mu(\mathbf{r}) \to [x_1, x_2, \ldots, x_N]^T,$$

For one X-ray i, the measurement sums attenuation along its path:

$$y_i \approx \sum_j A_{ij} x_j.$$

A_{ij} has a physical meaning: the length ray i travels through voxel j.

Every entry of A is geometry — nothing about A is arbitrary or abstract.

3. One ray is obviously not enough

The first appearance of non-uniqueness

Same measurement, many possible objects

$$y = x_1 + x_2 + x_3 = 6$$

Was the object [1, 2, 3]? Or [2, 2, 2]? Or [0, 1, 5]?

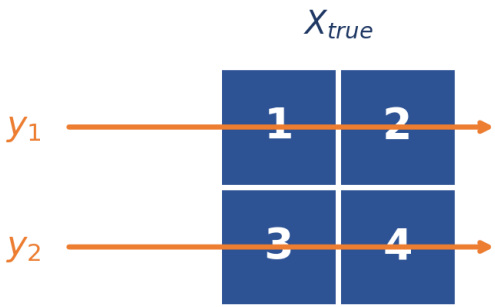
Candidate object	Predicted y
[1, 2, 3]	6
[2, 2, 2]	6
[0, 1, 5]	6

One measurement is not enough — but the fix is not simply "more numbers."

More measurements only help if they reveal new information about the object. That is exactly what CT rotation is for.

4. Build a toy CT scanner

Two horizontal projections through a 2×2 object



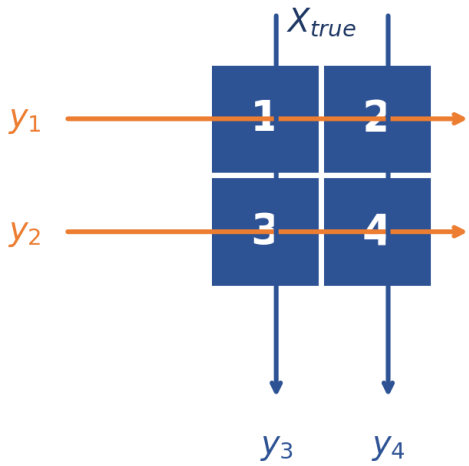
$$y_1 = x_1 + x_2 = 3, \quad y_2 = x_3 + x_4 = 7$$

Two equations cannot determine four unknowns. So we rotate the scanner and add another angle.

5. Add another projection angle

Vertical measurements — now 4 equations, 4 unknowns

Four measurements, four unknowns — solved?



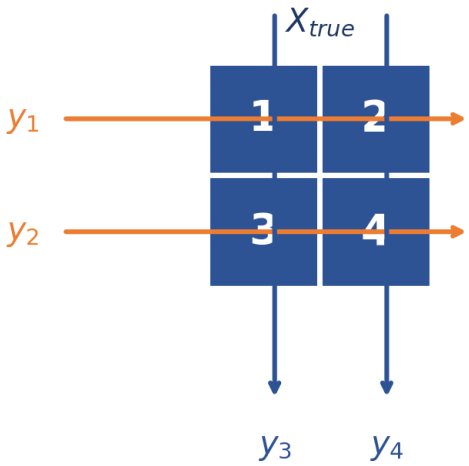
$$y_1 = x_1 + x_2 = 3, \quad y_2 = x_3 + x_4 = 7$$
$$y_3 = x_1 + x_3 = 4, \quad y_4 = x_2 + x_4 = 6$$

Can we reconstruct the image now?

6. Four numbers ≠ four pieces of information

A hidden redundancy in the measurements

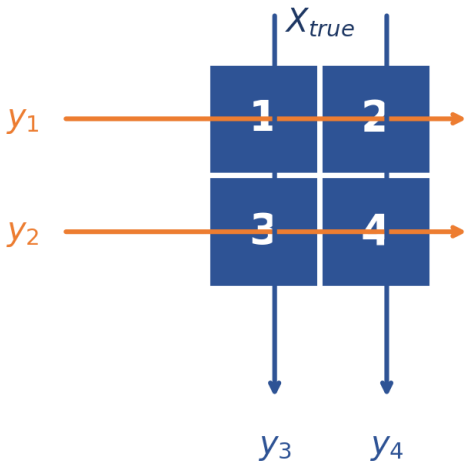
Four measurements, four unknowns — solved?



$$y_1 = x_1+x_2 = 3, \quad y_2 = x_3+x_4 = 7$$
$$y_3 = x_1+x_3 = 4, \quad y_4 = x_2+x_4 = 6$$

$$y_1+y_2 = 3+7 = 10 \qquad y_3+y_4 = 4+6 = 10$$

Four measurements, four unknowns — solved?



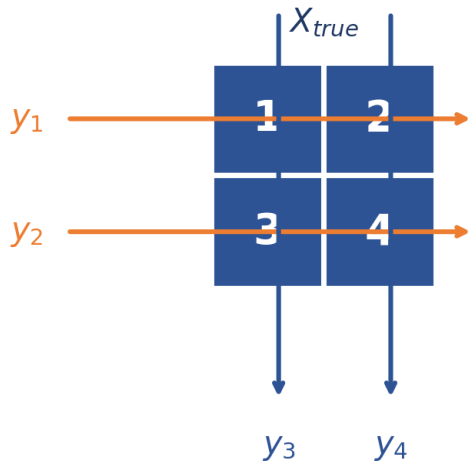
$$y_1 = x_1+x_2 = 3, \quad y_2 = x_3+x_4 = 7$$
$$y_3 = x_1+x_3 = 4, \quad y_4 = x_2+x_4 = 6$$

$$y_1+y_2 = 3+7 = 10 \qquad y_3+y_4 = 4+6 = 10$$

Whose fault is this?

Four measurements, four unknowns — solved?

Physical sensing geometry



$y_1 = x_1+x_2 = 3, \quad y_2 = x_3+x_4 = 7$

$y_3 = x_1+x_3 = 4, \quad y_4 = x_2+x_4 = 6$

Mathematical sensing model

$$y = Ax,$$

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}.$$

The redundancy is sitting in plain sight

$$y_1+y_2 = 3+7 = 10 \qquad y_3+y_4 = 4+6 = 10$$

Both sums equal the same thing — the total attenuation of all four pixels:

$$x_1+x_2+x_3+x_4$$

So one equation carries no new information — it is redundant with the other three.

rank(A) = 3

even though $A \in \mathbb{R}^{4x4}$

det(A) = 0

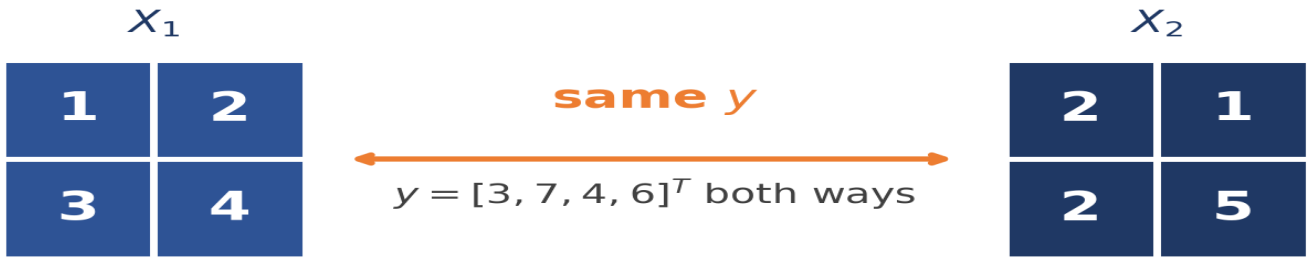
so A^{-1} does not exist

What does a "singular CT scanner" actually mean for the image?

7. What does “singular” mean, physically?

Two different patients, identical scan

The scanner cannot tell these two patients apart



Check X_2's projections: horizontal 2+1=3, 2+5=7. Vertical: 2+2=4, 1+5=6 — exactly the same as X_1.

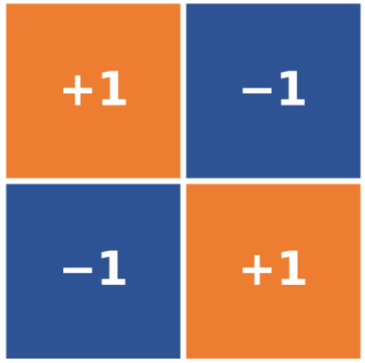
This is what singularity means physically: the acquisition geometry has failed to encode some spatial pattern of the object.

8. Find exactly what the scanner cannot see

The central explanation of singularity

Subtract the two objects — a checkerboard appears

v = checkerboard



$Av = 0$ **(invisible to A)**

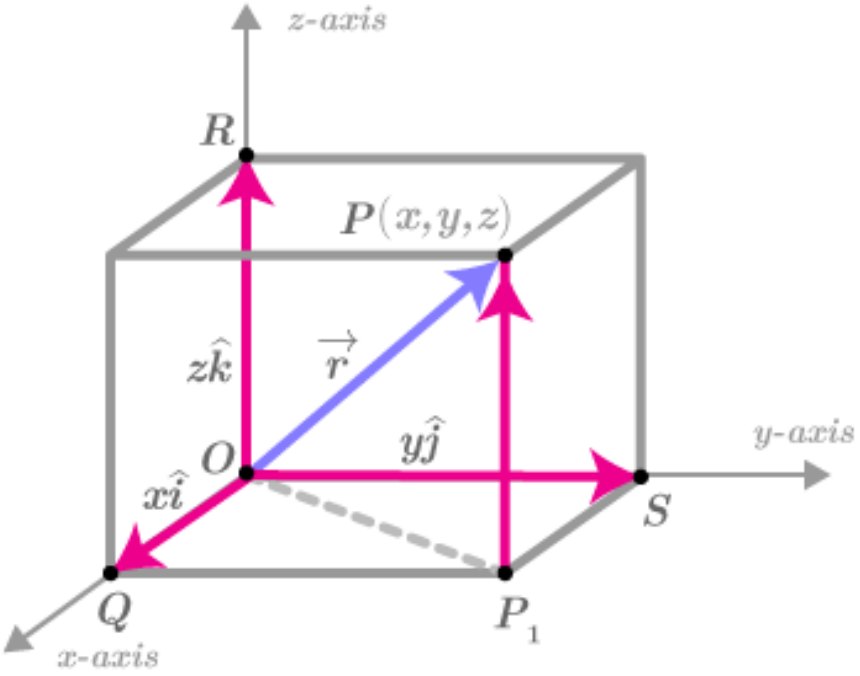
$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

$$X_2 - X_1 \rightarrow v = [1, -1, -1, 1]^T$$

$$\boldsymbol{Av = 0}$$

$$\boldsymbol{Ax = A\,(x + \alpha v)}$$

Image is a point/vector, measurement matrix “A” transforms it



The Scanner as Linear Map and Its Null Space

1. The image is a vector

An image x is a vector in a high-dimensional space $\mathbb{R}^n$.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

Each entry x_i is a pixel (voxel) value.

Example ($n = 4$):

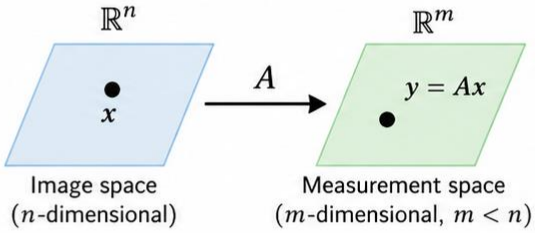
$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \in \mathbb{R}^4$$

2. The scanner is a linear map

The scanner is a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$ (with $m < n$):

$$y = Ax$$

It maps an n-dimensional image to m measurements.



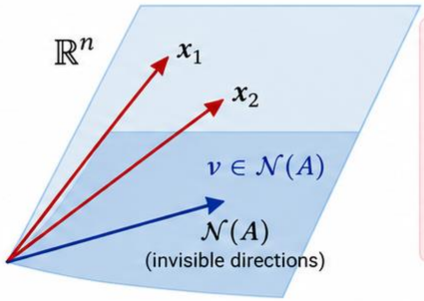
Many different x can map to the same y because $m < n$.

3. The null space: what the scanner cannot see

The **null space** of A is the set of all image vectors that produce zero measurements:

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}$$

$\mathcal{N}(A)$ is a vector subspace of $\mathbb{R}^n$ with dimension $n - \text{rank}(A)$.



For any $v \in \mathcal{N}(A)$:

$$Av = 0$$

Adding this pattern to any image does not change the measurements.

The scanner cannot see any component of the image in the null space.
All images that differ by a vector in $\mathcal{N}(A)$ produce exactly the same measurements.

$$A \text{ invertible} \iff \mathcal{N}(A) = \{0\} \iff \text{rank}(A) = n \iff \det(A) \neq 0$$

Any amount of this pattern is invisible to A

$$x' = x + \alpha v \qquad Ax' = Ax + \alpha Av = Ax$$

The null space contains anatomical patterns that this acquisition geometry cannot see — no matter how large α is.

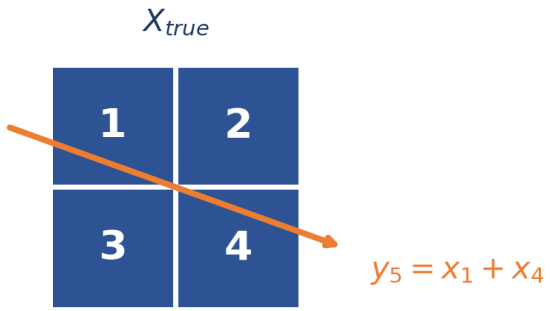
- "Invertible" and "identifiable" are properties of A and the acquisition geometry, not of the patient.
- A null-space vector v is a direction you can add to the true image with zero effect on any measurement.
- This is the single picture to keep in mind for the rest of the lecture: singularity, rank deficiency, and (soon) small singular values are all the same idea, just in degree.

Null space: The set of input directions that a linear forward model maps to zero; in sensing, these represent changes in the physical state that are invisible to the measurements.

9. How should the scanner fix this?

A measurement sensitive to the missing checkerboard mode

Change measurement geometry to “sees” the checkerboard



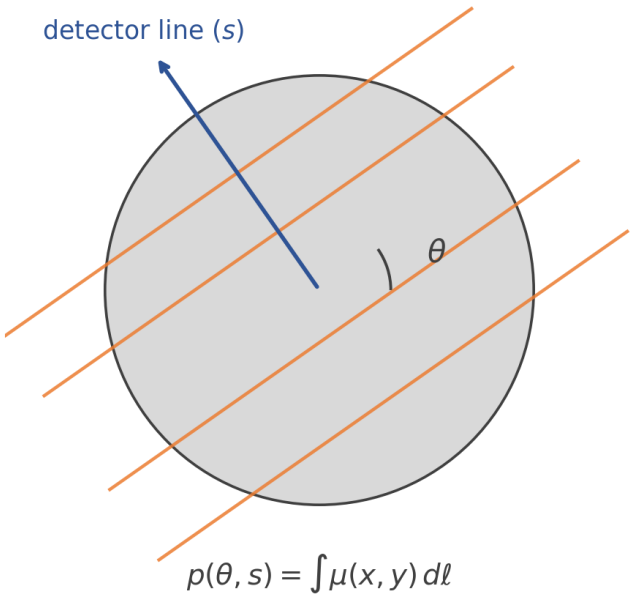
$$y_5 = x_1 + x_4 \rightarrow \text{true: } 1+4=5, \quad \text{alt: } 2+5=7$$

Now A_new v = $[0,0,0,0,2]^T \neq 0$ — the previously invisible mode is observable. CT rotation isn't just “more images”; different angles expose different spatial modes.

10. From toy CT to actual tomography

The Radon transform and the sinogram

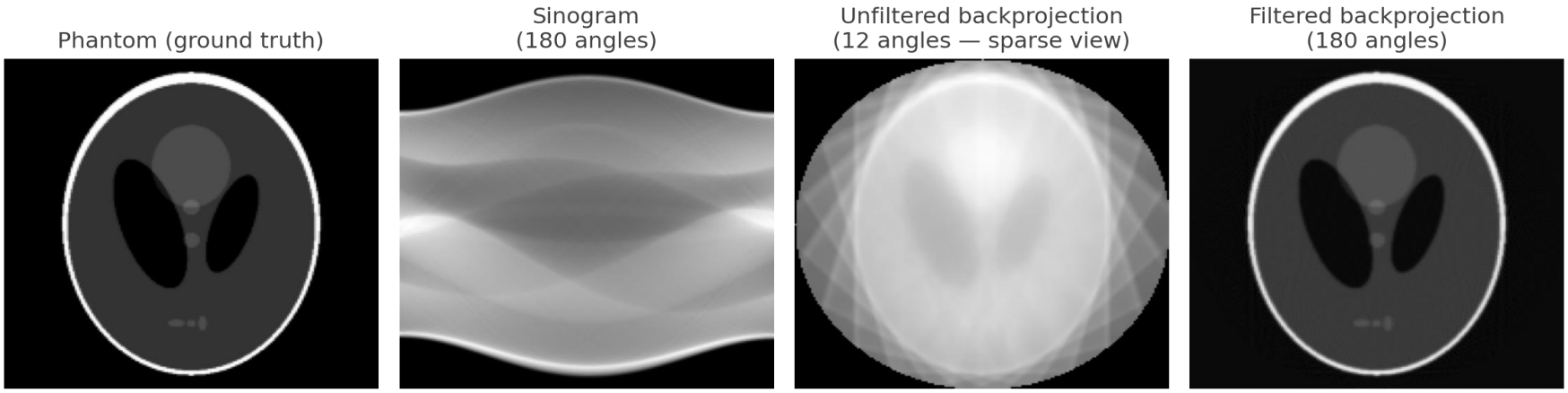
Line integrals at every angle and offset



$$p(\theta, s) = \sum_{ray(\theta,s)} \mu(x, y)\, d\ell$$

- One angle θ, all detector offsets s → one projection
- All angles together → a sinogram
- After discretizing: y = Ax — the same equation as our 2×2 toy scanner, just enormous

A genuine reconstruction: what angle count buys you



Too few projection angles (sparse view) leaves the image under-determined, exactly like our 2-equation toy scanner — filtered backprojection with enough angles recovers it cleanly.

The giant matrix A represents the discretized geometry of every ray in the scan.

11. Real measurements are not exact

From y = Ax to y = Ax + n

The residual: what a candidate image gets wrong

$$y = Ax + n$$

For a candidate reconstruction $\hat{x}$, it predicts $\hat{y} = A\hat{x}$. Define the residual:

$$r = y - A\hat{x}$$

- r is the disagreement between what the real scanner measured and what it would measure if $\hat{x}$ were the true anatomy
- if $\hat{x} = x_{true}$, then $r = n$ — pure noise

A correct reconstruction leaves discrepancies consistent with the noise model — not necessarily zero.

12. Maximum likelihood

Which CT image best explains the measurements?

Choosing among competing reconstructions

Candidates x_A, x_B, x_C each predict different projections: Ax_A, Ax_B, Ax_C. Which one should we choose?

$$\hat{x}_{ML} = arg\ max_x\ \boldsymbol{p(y \mid x)}$$

Among all possible internal attenuation maps, find the one for which the actual detector readings would be most probable.

This is where statistics enters — because the CT problem itself demands it, not because it's the next topic on a syllabus.

13. Gaussian noise → least squares

The sequence becomes logically clean

From a noise model to a concrete objective

$$n \sim N(0, \sigma^2 I) \quad ; \quad p(y|x) \propto exp(-||y-Ax||^2 / 2\sigma^2)$$

$$\hat{x}_{ML} = arg\ min_x\ ||y - Ax||^2$$

CT physics → y = Ax+n → noise stats → p(y|x) → ML → least squares

Least squares in CT means: find the attenuation map whose simulated X-ray projections are closest, in squared error, to what was measured.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

14. Solving least squares

What A^T actually does: backprojection

Setting the gradient to zero

$$J(x) = ||y - Ax||^2 \qquad \nabla_x J = -2A^T(y - Ax)$$

$$A^T A x = A^T y$$

$$\hat{x} = (A^T A)^{-1} A^T y$$

But the formula is not the point — the physical meaning of the two operators is:

A projects forward; A^T backprojects



- A (forward projection): take a proposed image and predict the X-ray projections it would produce
- A^T (backprojection): take projection-domain discrepancies and smear them back across the pixels the rays traversed

$A^T(y - Ax)$ means: backproject the projection errors to determine how the image should change.

15. Turning least squares into an algorithm

Simulate — compare — backproject — update



$$x^{k+1} = x^k + \eta\, A^T(y - Ax^k)$$

1. Simulate the scanner: $\hat{y}_k = Ax_k$
2. Compare with measurements: $r_k = y - \hat{y}_k$
3. Backproject the discrepancy: $g_k = A^T r_k$
4. Update the image: $x_{k+1} = x_k + \eta g_k$

This connects optimization directly to iterative reconstruction — no longer an abstract "Gradient-Descent" formula.

Lecture 3: Day 2

Revisiting Least Square and more

Data Science → Physical Science → Computational Physical Science

Data Science:

"What patterns are in the data?"

Physical Modeling:

"What physical process generated the data?"

Computational Physical Science:

"How do we computationally recover, simulate, and predict the physical world from measurements?"

A question we quietly skipped

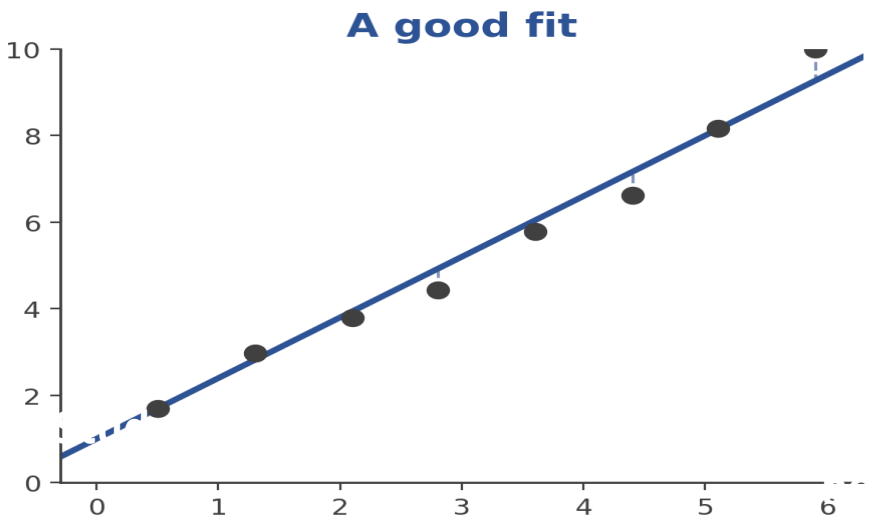
Every time we solved an inverse problem, we minimized squared error:

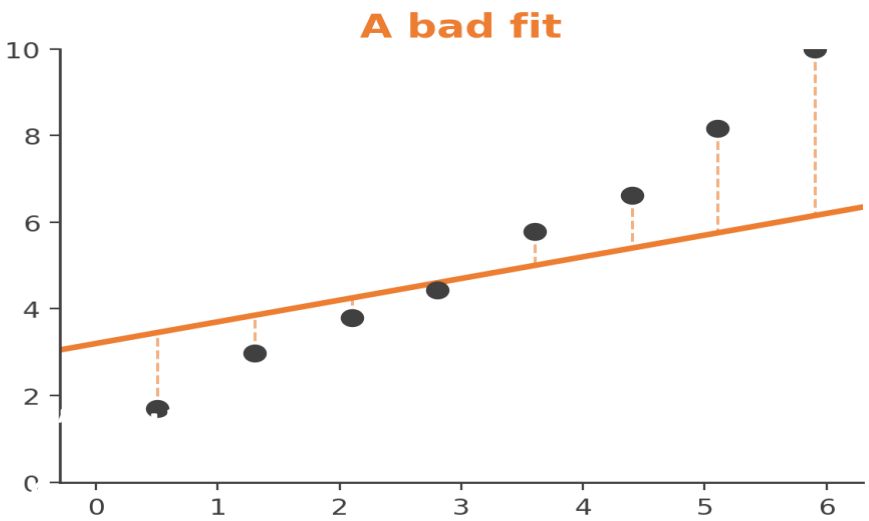
$$\hat{x} = arg\ min_x\ ||y - Ax||^2$$

Why square the error? Why not absolute error, fourth power, or the single worst error?

What does it mean for one line to fit better?

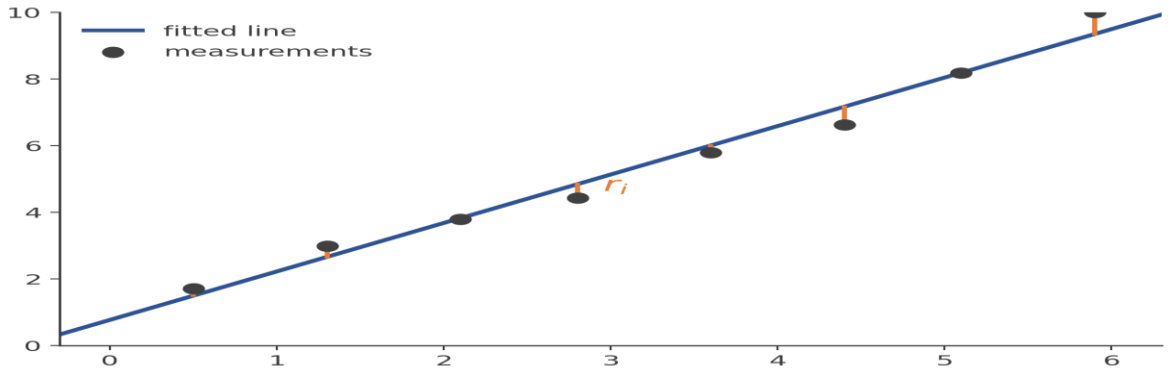
$$(t_i, y_i),\ i = 1, \ldots, N \qquad y_i = x_1 t_i + x_2 + \epsilon_i$$





Define the residual — then ask how to combine them

$$r_i = y_i - (x_1 t_i + x_2)$$



There is nothing mathematically inevitable about squaring.

Combine residuals as…	Named
$\Sigma_i\ r_i$	signed sum — cancels out, useless
$\Sigma_i\ \lvert r_i \rvert$	sum of absolute errors
$\Sigma_i\ r_i^2$	sum of squared errors — least squares
$\max_i\ \lvert r_i \rvert$	worst-case error

The general L_p norm formula is:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \qquad p \geq 1$$

For $p = 1$, this gives L_1; for $p = 2$, it gives the Euclidean L_2 norm.

Norm	Notation	What it measures / intuition	Common use
L_0 "norm"*	$\|x\|_0$	Number of nonzero elements	Sparsity, compressed sensing
L_1 norm	$\|x\|_1$	Total absolute magnitude	Robust fitting; sparse regularization
L_2 norm	$\|x\|_2$	Ordinary Euclidean length	Least squares, Gaussian noise
Squared L_2*	$\|x\|_2^2$	Squared Euclidean length	Least-squares objective
L_p norm	$\|x\|_p$	Generalized vector magnitude	General optimization, $p \geq 1$
L_∞ norm	$\|x\|_\infty$	Largest absolute component	Worst-case error

Line-fitting is y = Ax + n in disguise

Lecture 3, general form

$$y = Ax + n$$

Today, line fitting

$$y_i = x_1 t_i + x_2 + n_i$$

The unknown parameters are θ = [t; 1], and the forward model predicts:

$$\hat{y}_i(\theta) = x_1 t_i + x_2 \qquad r_i(\theta) = y_i - \hat{y}_i(\theta)$$

forward model → prediction → measurement → residual → inference

Least squares isn't just a geometric rule for a good-looking line — it can be derived from a probabilistic model of how the measurements were generated.

"A" tells us **how the unknown parameters x generate measurements y.**

1. In line fitting, where does A come from?

Suppose the model is

$$y(t) = x_1 t + x_2.$$

Here x_1, x_2 are unknown. If we observe at $t = 1, 2, 3$,

$$\begin{aligned} y_1 &= 1x_1 + x_2 \\ y_2 &= 2x_1 + x_2 \\ y_3 &= 3x_1 + x_2. \end{aligned}$$

Therefore,

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x}.$$

So the **straight-line rule determines the structure of A.**

The forward model says:

given line parameters x, $\quad$ A predicts what we measure.

2. CT has a different physical rule

In CT, the physical rule is not

$$y = x_1 t + x_2.$$

It is approximately the **line integral of attenuation along an X-ray path**:

$$y_i = \int_{\text{ray } i} \mu(\mathbf{r})\, d\mathbf{r}.$$

After discretizing the body into pixels:

$$\boxed{y_i = \sum_j A_{ij} x_j}$$

where x_j is the attenuation coefficient of pixel j, and

$$A_{ij} \approx \text{length of X-ray } i \text{ passing through pixel } j.$$

For example:

$$y_1 = 0.8x_1 + 1.0x_2 + 0x_3 + 0.4x_4.$$

That equation is the CT counterpart of

$$y_i = t_i x_1 + x_2.$$

3. So the more precise analogy is

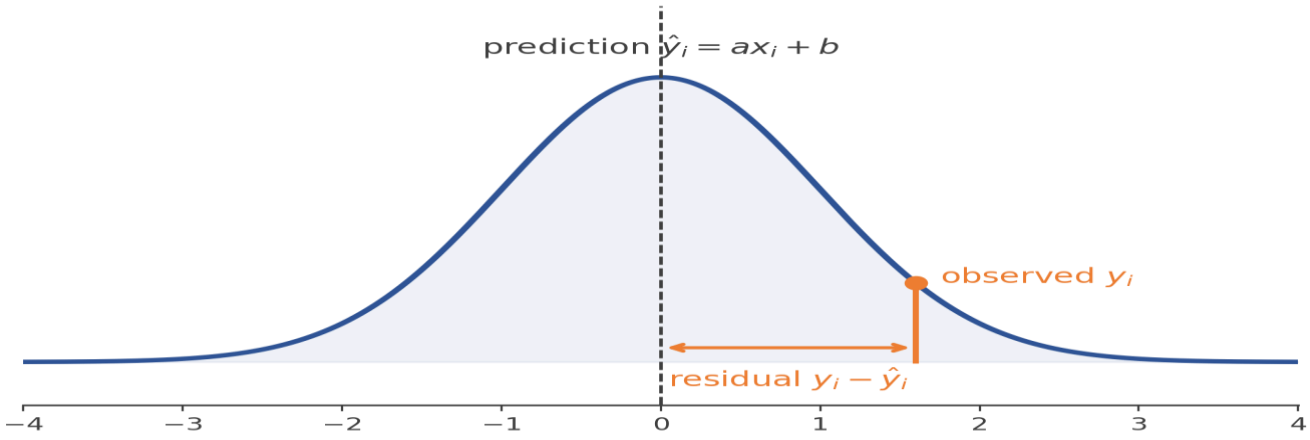
Line fitting	CT
Physical/model assumption: $y(t) = x_1 t + x_2$	Physical model: $y_i = \int_{\text{ray } i} \mu(\mathbf{r}) d\mathbf{r}$
Unknown x_1, x_2	Unknown pixel attenuations $x_1, \ldots, x_N$
$A_{i,:} = [t_i, 1]$	$A_{i,:}$ = ray i's intersection lengths with pixels
$y_i = t_i x_1 + x_2$	$y_i = \sum_j A_{ij} x_j$
All measurements: $y = Ax$	All measurements: $y = Ax$

So I would slightly correct my previous statement: **the straight line itself isn't the best direct analogue of the CT image.** The deeper analogy is:

$$\begin{array}{ccc} \text{line parameters } x & \xrightarrow[\text{straight-line rule}]{A} & \text{point measurements } y \\ \text{CT image } x & \xrightarrow[\text{X-ray physics}]{A} & \text{projection measurements } y \end{array}$$

The probability model tells us how to penalize errors

$$y_i = x_1 t_i + x_2 + n_i \qquad n_i \sim N(0, \sigma^2)$$



$$p(y_i \mid x_1, x_2) = 1/\sqrt{(2\pi\sigma^2)}\ \ exp{-(y_i-(x_1 t_i+ x_2))^2/2\sigma^2}$$

The term inside the exponent is precisely the squared residual: a line producing a large residual makes the observed measurement exponentially less probable.

Independent noise ⇒ multiply the probabilities

$$p(y \mid x_1, x_2) = \prod_{Ni=1} p(y_i \mid x_1, x_2)$$

$$L(x_1, x_2) = \prod_i \; 1/\sqrt{(2\pi\sigma^2)} \;\; exp{-(y_i - x_1\, t_i - x_2)^2/2\sigma^2}$$

What are we trying to find? Not the most probable y — we already measured y.

We want the parameters that make the measurements we actually observed most plausible.

$$(\hat{x}1, \hat{x}2)_{ML} = arg\ max_{x1,\, x2}\ \mathbf{p(y \mid x_1, x_2)}$$

The key payoff

$$\log L(x_1, x_2) = -N/2 \log(2\pi\sigma^2) - 1/2\sigma^2 \sum_i (y_i - x_1 t_i - x_2)^2$$

- The first term doesn't depend on x1, x2 — it drops out of the arg max
- The multiplier $1/2\sigma^2$ is positive and constant — it doesn't change where the max is

$$\arg\max_{x1,\,x2} \log L(x_1, x_2) = \arg\min_{x1,x2} \sum_i (y_i - x_1 t_i - x_2)^2$$

Gaussian noise + Maximum Likelihood ⇒ Least Squares

So: why did we square the residual? Because under Gaussian measurement noise, the negative log-probability of a residual is proportional to its square — a much deeper reason than "squares stop positive and negative errors from canceling."

Exactly the same derivation, in vector form

$$y = Ax + n, \quad n \sim N(0, \sigma^2 I) \qquad \hat{x} = \arg\min_x ||y - Ax||^2$$

$$J(x) = (y - Ax)^T (y - Ax) \qquad \nabla_x J = -2A^T (y - Ax)$$

$$A^T (y - A\hat{x}) = 0 \qquad A^T A\hat{x} = A^T y$$

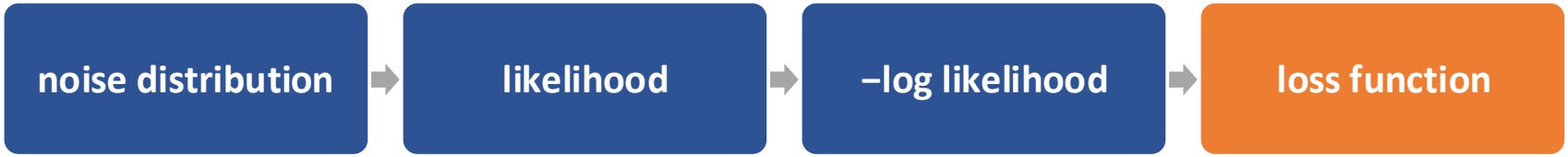
$$\hat{x} = (A^T A)^{-1} A^T \mathbf{y}$$

This connects directly back to CT:

$$y_{\text{measured projections}} = A_{\text{CT forward operator}}\, x_{\text{unknown image}} + n_{\text{measurement noise}}$$

Least squares no longer looks like an arbitrary optimization objective — it's what Gaussian noise plus maximum likelihood forces on us.

One general principle governs every case



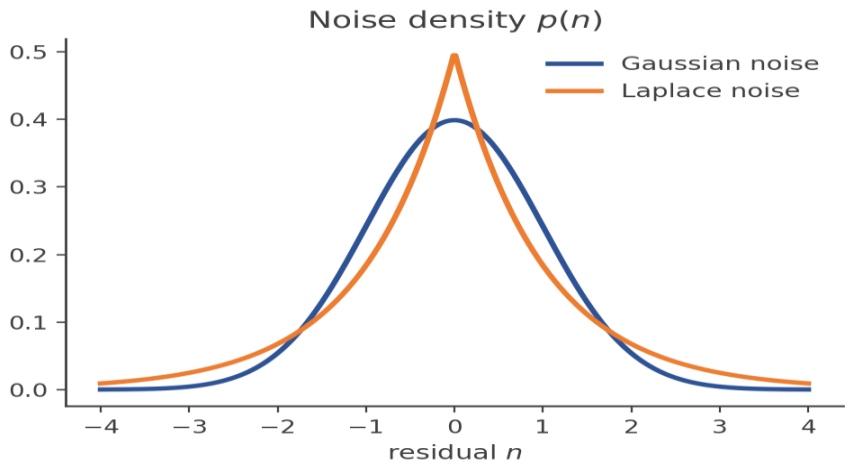
Change the noise assumption, and the "right" loss function changes with it — least squares was never the only option, just the Gaussian one.

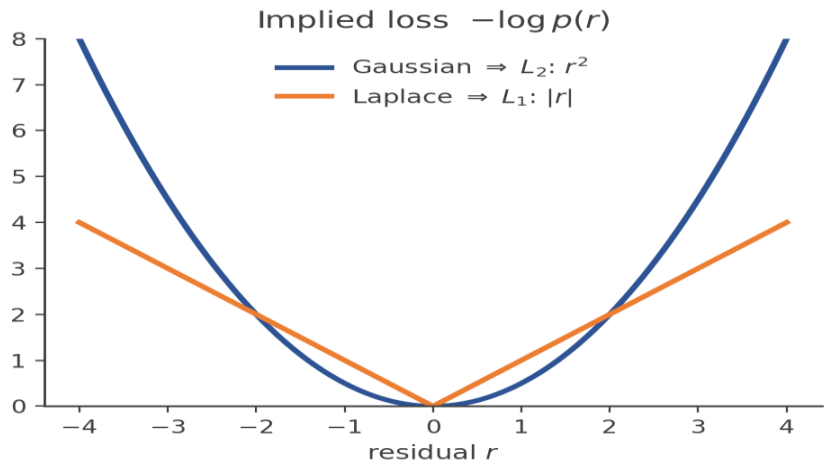
Let's derive a few concrete cases.

Laplace noise → L1 loss — robust to outliers

$$p(r) \propto \exp(-|r|/b) \qquad -\log p(r) = |r|/b + \text{const.}$$

$$\hat{x} = \arg\min_x \Sigma_i |y_i - [Ax]_i|$$





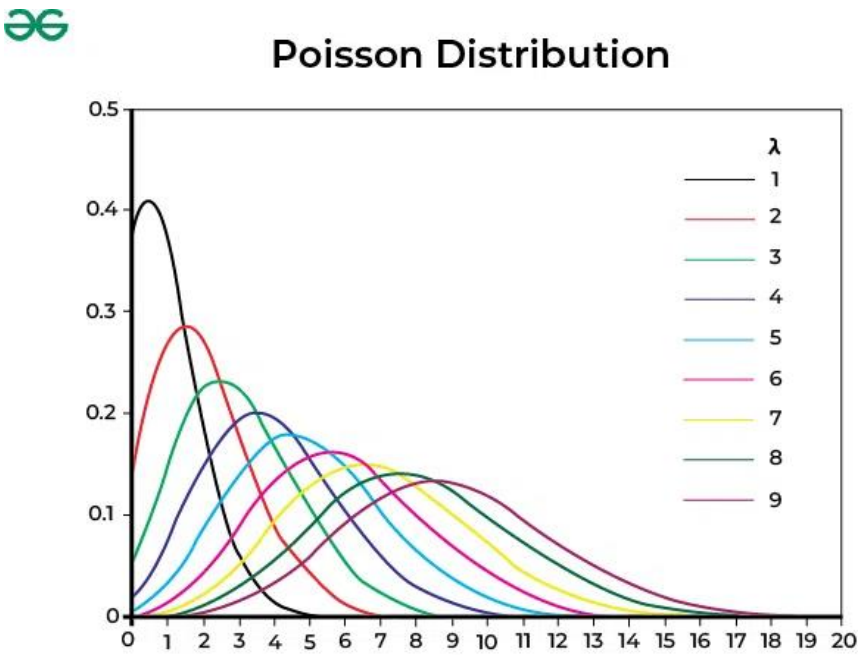
Laplace noise → absolute-error loss (L1). L1 grows only linearly with a large residual, so a single bad outlier can't dominate the fit the way it does under L2.

Poisson measurements → the CT callback

Particularly relevant for CT: X-ray detectors fundamentally count photons.

$$y_i \sim Poisson(\lambda_i(x)) \qquad p(y_i|\lambda_i) = \lambda_i y_i e^{-\lambda}{}_i / y_i!$$

$$min_x \; \Sigma_i \, [\lambda_i(x) - y_i \log \lambda_i(x)]$$



Poisson measurements → the CT callback

Particularly relevant for CT: X-ray detectors fundamentally count photons.

$$y_i \sim Poisson(\lambda_i(x)) \qquad p(y_i|\lambda_i) = \lambda_i y_i e^{-\lambda}{}_i / y_i !$$

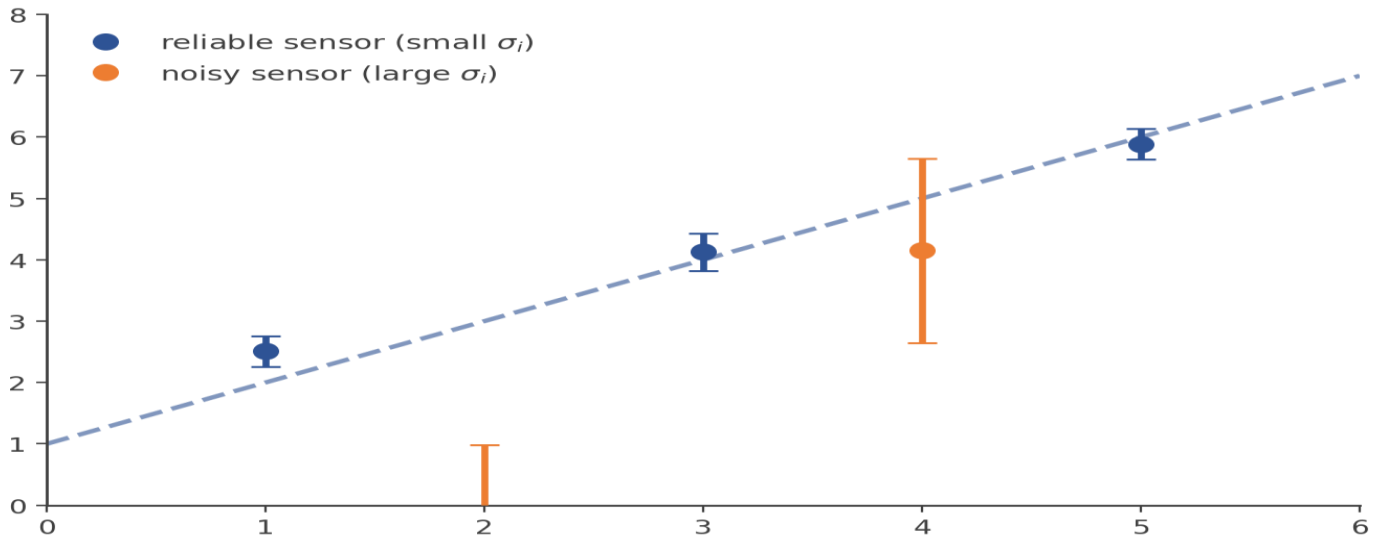
$$min_x \; \Sigma_i \, [\lambda_i(x) - y_i \, log \, \lambda_i(x)]$$

If CT measurements fundamentally involve photon counting, should ordinary least squares always be the statistically correct objective? Not necessarily — the physics of sensing shapes the statistics, and the statistics shape the loss function.

Noise model	Implied loss
Gaussian, equal variance	L2 — squared error (“least squares”)
Laplace	L1 — absolute error
Poisson (photon counts)	λ(x) − y logλ(x)

Not every sensor is equally trustworthy

$n_i \sim N(0, \sigma_i^2)$ — *each measurement has its own noise level*



$$min_x \ \Sigma_i \ r_i^2 / \sigma_i^2$$

Weighted least squares: a reliable sensor (small σ_i) gets high weight $1/\sigma_i^2$ in the objective; a noisy sensor (large σ_i) gets down-weighted automatically.

The general case — and a misconception to retire

$$n \sim N(0, \Sigma) \qquad \hat{x} = \arg\min_x \ (y - Ax)^T \Sigma^{-1} (y - Ax)$$

Σ is the full noise covariance — it reduces to $\sigma^2 I$ only in the special i.i.d., equal-variance case.

Gaussian noise does not automatically mean ordinary $\|y - Ax\|_2^2$. Plain least squares follows specifically from i.i.d., equal-variance Gaussian noise — the moment sensors disagree in reliability, the correct objective is weighted.

We have a formula for x̂. Why isn't that the end of the story?

$$\hat{x} = (A^T A)^{-1} A^T y = A^{\dagger} y$$

The pseudoinverse tells us what the least-squares solution is. Iterative methods tell us how to compute it practically — and let us solve far more than plain least squares.

Two separate reasons push us toward iterative inversion:

Computational scale

The formula exists, but forming and inverting A^TA may be computationally infeasible at real sensing scale.

More realistic problems

Noise amplification, priors, constraints, and nonlinear forward models — none of these have a simple closed-form inverse.

“We have a formula for x. Why don't we just use it?” — that question is today's motivation.

$$\boxed{y = Ax + n}, \qquad A \in \mathbb{R}^{m \times n}$$

where $m =$ number of measurements and $n =$ number of unknowns.

System	Dimensions	Measurements vs. unknowns	Typical situation	Exact inverse A^{-1}?	Typical solution
Square, full rank	$m = n,\ \mathrm{rank}(A) = n$	Same	Unique solution	**Yes**	$x = A^{-1}y = A^{\dagger}y$
Square, rank deficient	$m = n,\ \mathrm{rank}(A) < n$	Same, but redundant information	Singular system	**No**	Pseudoinverse / regularization
Overdetermined	$m > n$	More measurements than unknowns	Usually no exact x with noise	**No**	Least squares: $x = A^{\dagger}y$
Underdetermined	$m < n$	Fewer measurements than unknowns	Infinitely many possible x's	**No**	Minimum-norm pseudoinverse or regularization/prior

Four reasons the closed form isn't the end of the story

1. A can be enormous — *storing / inverting A^TA is infeasible at scale*

↓

2. Pseudoinverse amplifies noise — *small singular values → huge errors*

↓

3. No room for prior knowledge — *smoothness, sparsity, nonnegativity, …*

↓

4. The forward model may be nonlinear — *$y = f(x) + n$ has no matrix to invert*

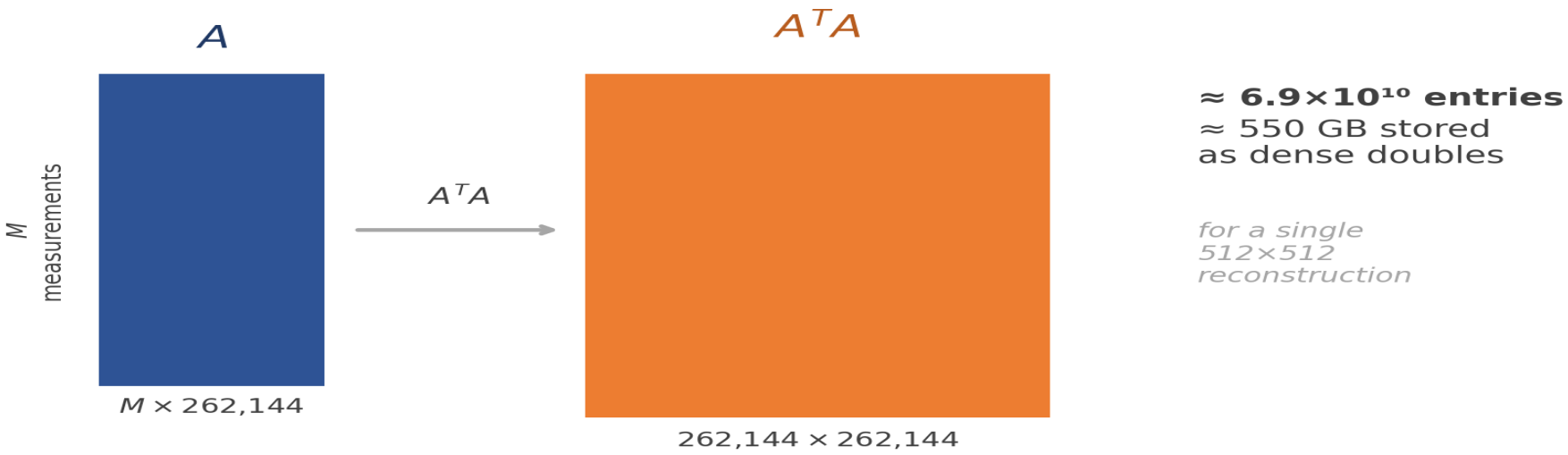
Each limitation motivates a piece of the same answer: treat reconstruction as a computational procedure, not a formula.

Limitation 1: Measurement matrix is huge

512×512 image $\quad x \in \mathbb{R}_{262,144}$

With hundreds of thousands to millions of X-ray measurements, A can conceptually have dimensions:

$$A \in \mathbb{R}_{M \times 262,144}$$



Closed form ≠ practical algorithm

Mathematical closed form ≠ computationally practical algorithm.

- Explicitly forming A^TA already produces a 262,144 × 262,144 matrix — roughly 6.9×10^{10} entries
- At 8 bytes per double, that's about 550 GB just to store A^TA densely
- And then you'd still have to solve that huge linear system

So in real imaging systems, we often don't even store A explicitly.

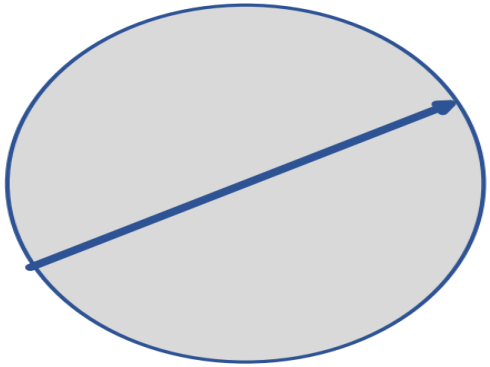
implement Ax → **implement A^Tr** → **never form A or A^TA**

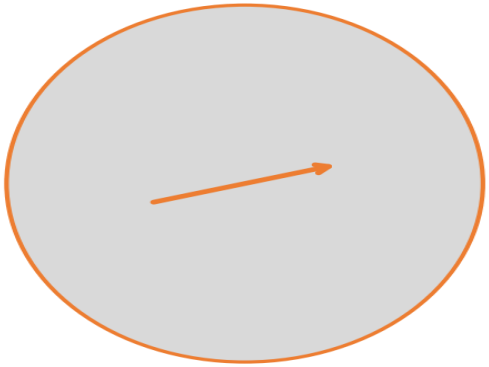
We implement the operations that compute Ax and A^Tr efficiently — as algorithms, not as stored matrices.

Limitation 2: Pseudoinverse Can Amplify Noise

$$\hat{x} = (A^T A)^{-1} A^T y = A^{\dagger} y$$

Even with enough measurements, not all directions are equal





v_2 isn't completely invisible — $Av_2 \neq 0$ — but it is almost invisible. This is the noisy counterpart of the earlier null-space problem.

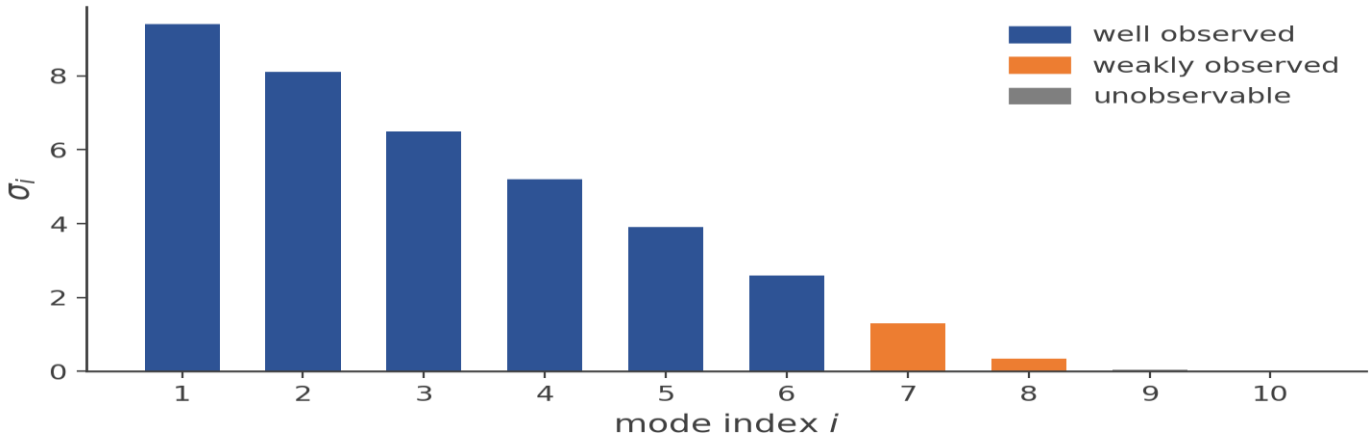
Regime	Condition	Meaning
Exact singularity	Av = 0	scanner cannot see the pattern at all

This is precisely why inversion can be unstable even when a unique mathematical solution exists.

SVD: A = UΣV^T: object modes and how strongly they're seen

$$A = U\Sigma V^T \qquad Av_i = \sigma_i u_i$$

The right singular vectors v_i are object/image modes; σ_i measures how strongly mode i appears in the measurements.

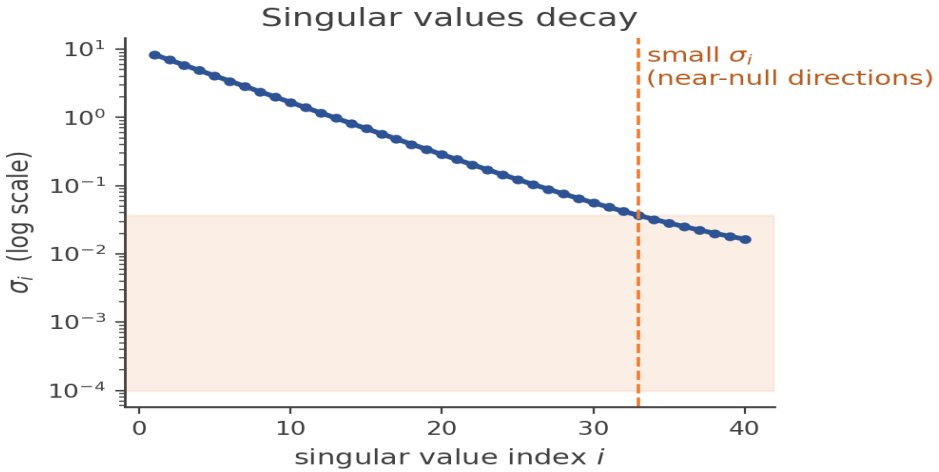


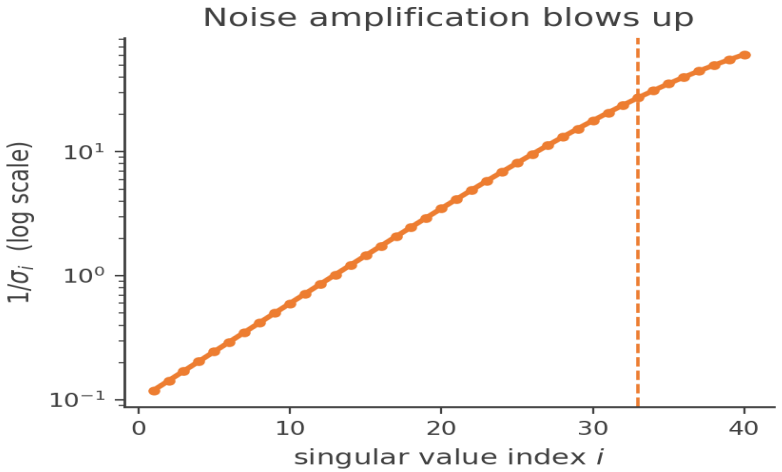
σ_i	What it means
large	strong detector change — well observed
small	very little detector change — poorly observed
zero	no detector change — invisible (null space)

Limitation 2: Pseudoinverse Can Amplify Noise

$$A = U\Sigma V^T \qquad A^\dagger = V\Sigma^\dagger U^T$$

$$\hat{x} = \Sigma_i \; u_{iT} y \,/\, \sigma_i \; v_i$$





If $\sigma_i \approx 0$, then $1/\sigma_i \gg 1$ — even a tiny amount of measurement noise along that direction becomes a huge error in $\hat{x}$.

Well-defined mathematically, unstable in practice

$$\textit{small } \sigma_i \quad \rightarrow \quad \textit{noise amplification}$$

- An ill-conditioned sensing system has some directions that are only weakly observed
- The pseudoinverse is mathematically well-defined in many such cases
- But the resulting reconstruction can still be practically unstable

This is an inverse-problem limitation, not purely a computational one — no amount of extra compute fixes a direction the sensor barely measures.

This is exactly the near-singularity / conditioning story from Lecture 3 — reappearing here as a property of the pseudoinverse itself.

Limitation 3: No room for “prior knowledge”

$$\min_x \; ||y - Ax||_2^2$$

But you may already know things about x — say, a CT image should be spatially smooth, nonnegative, piecewise smooth, or sparse in some transform domain.

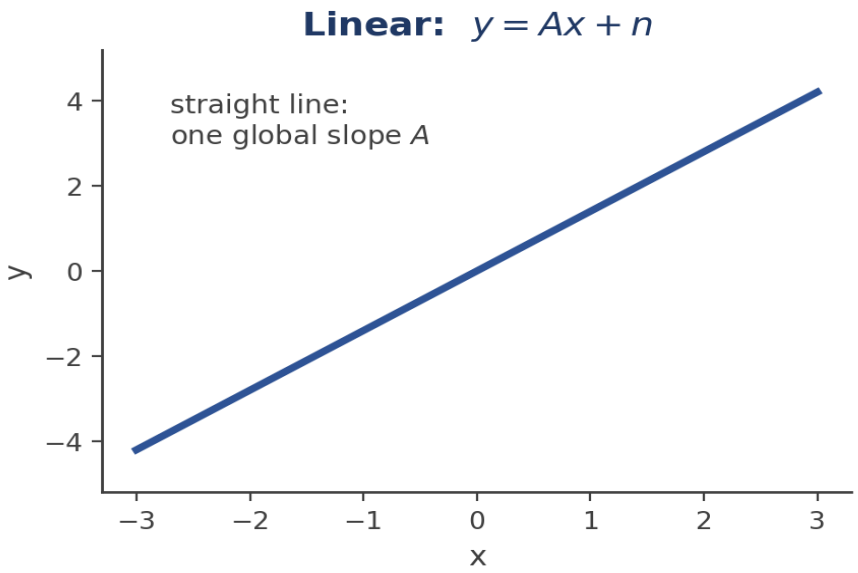
$$\min_x \; \|y - Ax\|_2^2 \; + \; \lambda \boldsymbol{R(x)}$$

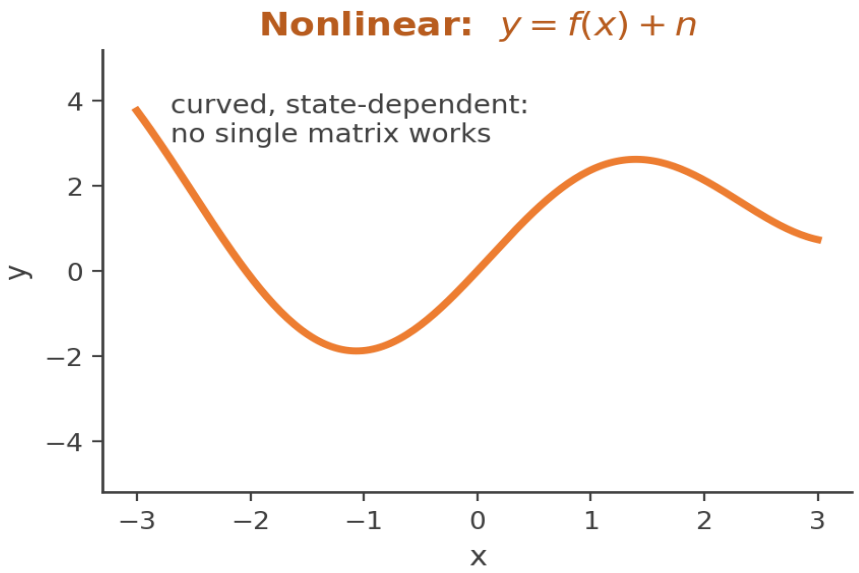
measurement consistency *prior / regularization*

You might additionally require $x_i \geq 0$ — a hard constraint plain least squares has no way to express.

Once R(x) or a constraint enters the picture, there's usually no simple pseudoinverse formula — iterative optimization becomes the natural approach.

Limitation 4: Forward model may not be linear





$$\hat{x} = arg\ min_x\ ||y - f(x)||^2$$

There is no matrix $A^\dagger$ to calculate. You generally need an iterative nonlinear optimization procedure.

A connection to ML: a neural network or a differentiable physical simulator can become part of f(x) — the same optimization principles still apply.

What Does Iterative Inversion Actually Do?

Four steps, repeated: $J(x) = \|y - Ax\|^2_2$, start from $x^{(0)}$

1 **Forward projection** $\hat{y}^{(k)} = Ax^{(k)}$

If this were the true image, what would the scanner have measured?

2 **Compare with reality** $r^{(k)} = y - Ax^{(k)}$

How inconsistent is the current image with the actual measurements?

3 **Map error back to image space** $\nabla J(x^{(k)}) = -2A^T(y - Ax^{(k)})$

A^T maps information from measurement space back into image space.

4 **Correct the estimate** $x^{(k+1)} = x^{(k)} + 2\eta\, A^T(y - Ax^{(k)})$

Update, then repeat from Step 1.

The intuitive physical loop, not just the formula



1. guess physical state
2. simulate measurement
3. compare to real data
4. backproject error
5. correct, repeat

This is the reading I'd emphasize more than the gradient-descent equation itself: guess → simulate → compare → backproject → correct — the same loop underlies every iterative inverse method.

Least Squares via Gradient Descent

$y = Ax + n$

1. Problem formulation

Given measurements y and sensing matrix A, solve the least squares problem

$$\hat{x} = \arg\min_x \|Ax - y\|_2^2$$

Define the cost function

$$f(x) = \|Ax - y\|_2^2 = (Ax - y)^T (Ax - y)$$

2. Gradient

The gradient of the least squares cost is

$$\nabla f(x) = 2A^T(Ax - y)$$

3. Gradient descent update

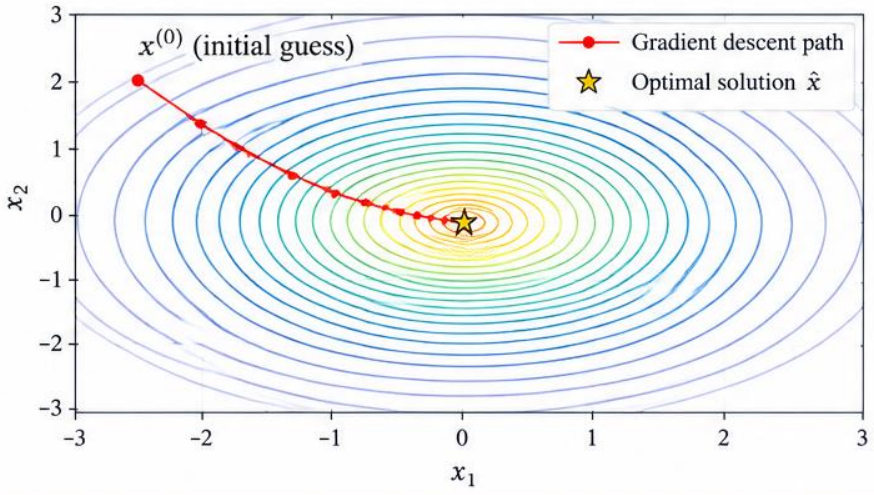
Starting from an initial guess $x^{(0)}$, iterate

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha \nabla f(x^{(k)}) \\ &= x^{(k)} - 2\alpha A^T(Ax^{(k)} - y) \end{aligned}$$

where $\alpha > 0$ is the step size (learning rate). For convergence (when A has full column rank), choose $0 < \alpha < 1/\lambda_{\max}(A^T A)$.

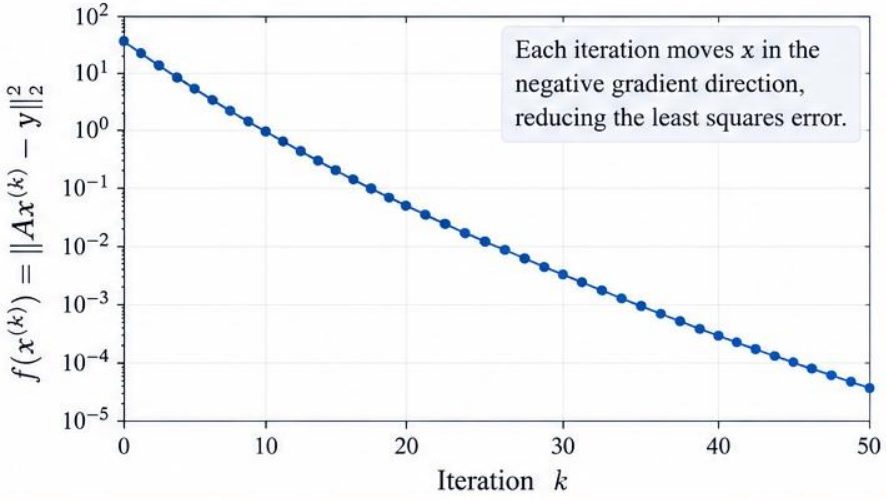
Geometric view (2D example)

Contours of $f(x) = \|Ax - y\|_2^2$ and gradient descent path



Convergence of the cost

The least squares objective decreases monotonically.



Key takeaway: Gradient descent iteratively updates x using $x^{(k+1)} = x^{(k)} - 2\alpha A^T(Ax^{(k)} - y)$ to minimize $\|Ax - y\|_2^2$, converging to the least squares solution $\hat{x} = A^{\dagger}y$ (when A has full column rank).

What Exactly Is “Computational Inverse”?

Broader than “iterative” — a shift in what we're solving

Analytical / direct inversion

$x = A^{\dagger}y$

obtain x through a formula or direct factorization

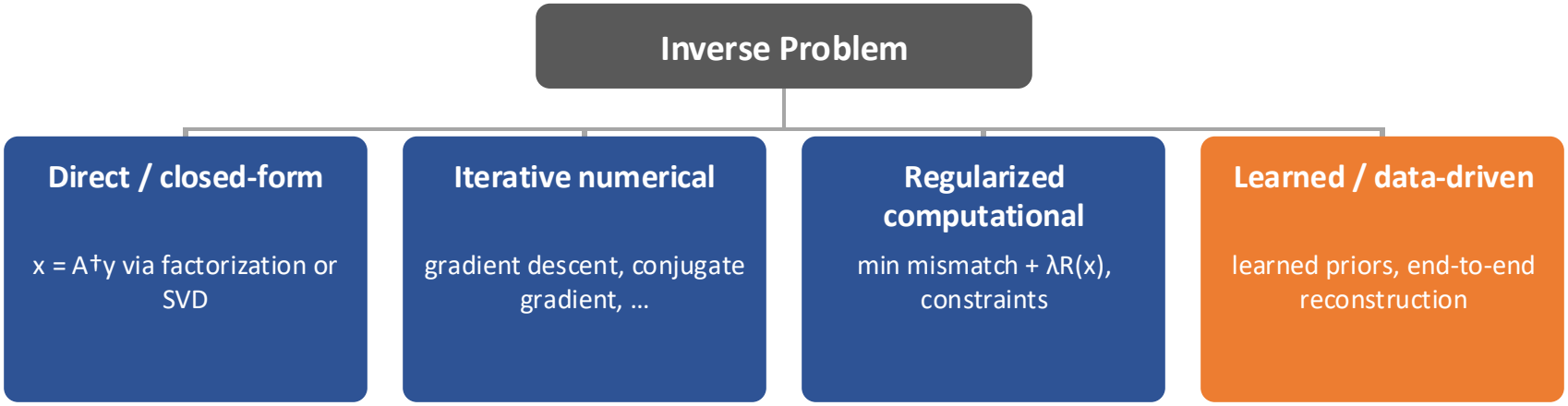
Computational inverse problem

$\hat{x} = arg\ min_x$ ***[data mismatch + prior]***

treat reconstruction as a computational procedure, solved numerically

Iterative optimization is one major way to solve a computational inverse problem — but “pseudoinverse vs. computational inverse” are not two competing definitions of inverse.

I would not teach these as a binary choice. Instead:



Why not just collect data from every angle/all possible combinations of the target variables?

Real acquisition has constraints

An excellent question — and why the answer is no

"Then why not just collect more projections from every possible angle?"

- ☐ Radiation dose to the patient
- ☐ Acquisition time
- ☐ Patient motion during the scan
- ☐ Scanner geometry and detector limitations

Deliberately reducing projections for dose gives sparse-view CT: some spatial modes become weakly constrained or unconstrained — the inverse problem is unstable again.

Now — and only now — prior information becomes necessary.

Maximum Likelihood (MLE) uses one source of information. What about the other?

Continuing directly from the CT example: the likelihood p(y|x) tells us how well an image explains the measurements. But before we ever measured anything, were all images equally plausible?

No. We already know things about what a CT image should look like — and Bayes' rule is how we combine that knowledge with the measurements.

$$p(x|y) \propto p(y|x) \cdot \mathbf{p(x)}$$

measurement evidence *prior knowledge*

- Least squares / maximum likelihood used only p(y|x) — the measurement term
- Brings in p(x) — the prior — and see exactly where regularization comes from

Five steps from Bayes' rule to a regularized inverse problem

1. Bayes' rule	*two sources of information about x*
2. What do we already know?	*physical knowledge about CT images*
3. Encode belief as a probability	*piecewise smooth* $\Rightarrow p(x)$
4. MAP estimation	*balance measurement fit and prior*
5. MAP becomes regularization	$\|\|y-Ax\|\|^2 + \lambda R(x)$

By the end: regularization stops looking like an arbitrary extra term and becomes an explicit statement of prior knowledge about x.

(i) Bayes: Two Sources of Information

We measured y. What should we believe about x?

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

For estimating x, p(y) is just a constant:

$$p(x \mid y) \propto p(y \mid x)p(x)$$

p(y | x)

What does the sensor tell us?

If this were the true attenuation image, how likely are the measured X-ray projections?

p(x)

What do we already know about plausible CT images?

Before looking at these measurements, how plausible is this image as a CT image?

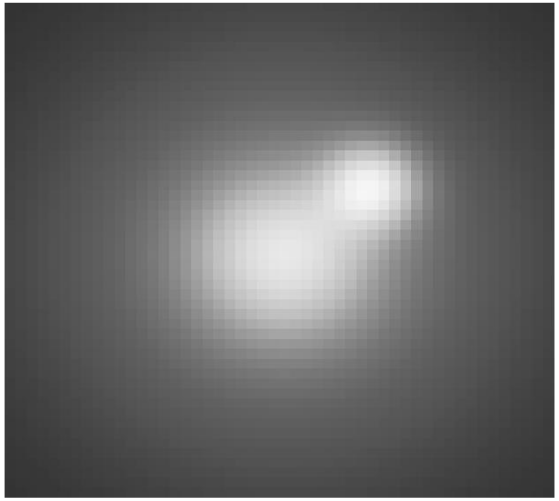
Posterior knowledge = measurement evidence 丐prior knowledge

Least squares / maximum likelihood used only the left box. Today we bring in the right one.

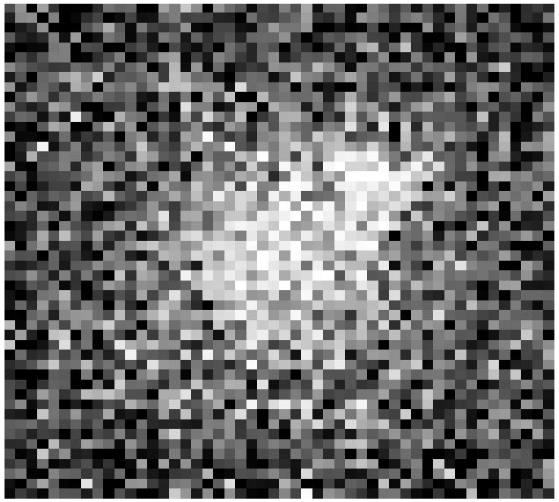
(ii) What Do We Already Know About a CT Image?

Before measuring anything, are all x's equally plausible?

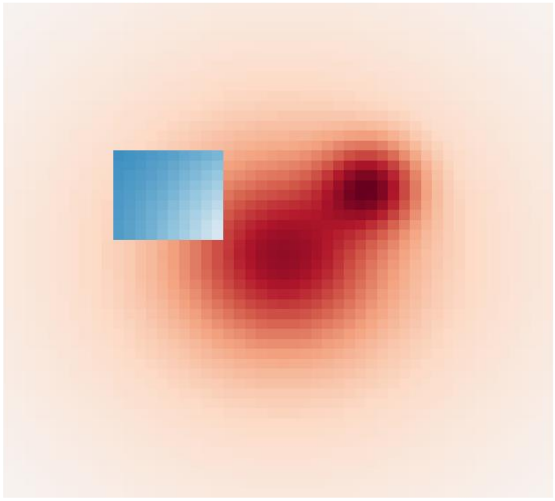
Plausible anatomy



Noisy image

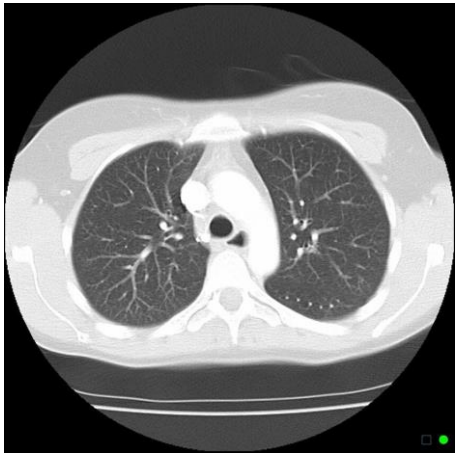


Negative attenuation



blue = $x_i < 0$ (physically impossible)

No. Some images are physically plausible, some are noisy, and some are outright impossible (negative attenuation) — before we've measured a single X-ray.

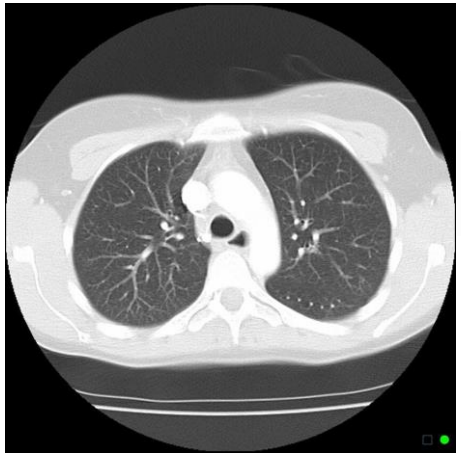


Chest CT: piecewise smoothness / sparse gradients

$$p(x) \propto e^{-\lambda \|\nabla x\|_1}$$

$$\boxed{p(x) \propto e^{-\lambda TV(x)}}$$

$$\boxed{TV(X) = \sum_{i,j} |X[i+1,j] - X[i,j]| + |X[i,j+1] - X[i,j]|}$$



Chest CT: piecewise smoothness / sparse gradients

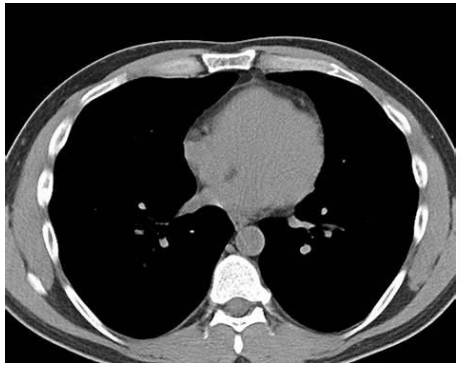
Lung window

$$WL \approx -600\ \text{HU}, \qquad WW \approx 1500\ \text{HU}$$

$$p(x) \propto e^{-\lambda \|\nabla x\|_1}$$

$$\boxed{p(x) \propto e^{-\lambda TV(x)}}$$

$$\boxed{TV(X) = \sum_{i,j} |X[i+1,j] - X[i,j]| + |X[i,j+1] - X[i,j]|}$$

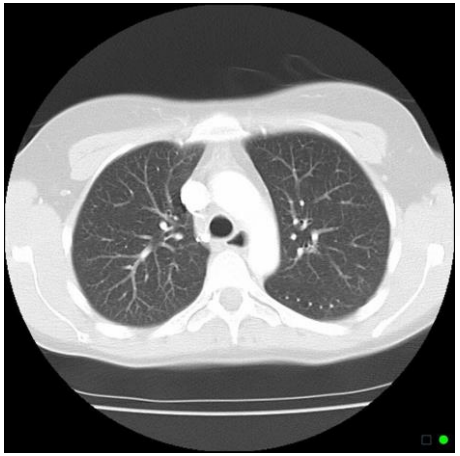


Chest/mediastinal CT: excellent for bounded attenuation and regions

Mediastinal/soft-tissue window

$$WL \approx 40\ \text{HU}, \qquad WW \approx 400\ \text{HU}.$$

$$\boxed{\mu_{\min} \leq X[i,j] \leq \mu_{\max}}$$



Chest CT: piecewise smoothness / sparse gradients

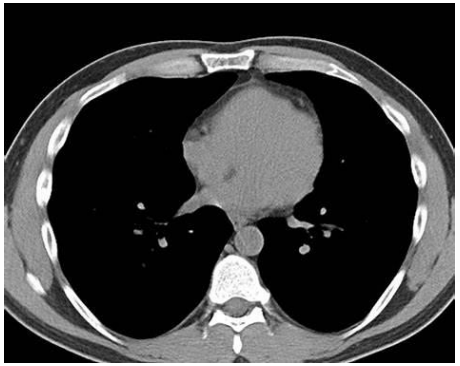
Lung window

$$WL \approx -600\ \text{HU}, \qquad WW \approx 1500\ \text{HU}$$

$$p(x) \propto e^{-\lambda \|\nabla x\|_1}$$

$$\boxed{p(x) \propto e^{-\lambda TV(x)}}$$

$$\boxed{TV(X) = \sum_{i,j} |X[i+1,j] - X[i,j]| + |X[i,j+1] - X[i,j]|}$$

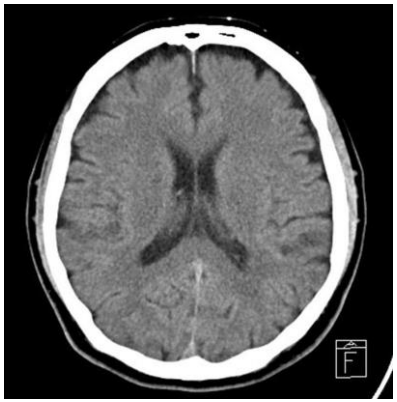


Chest/mediastinal CT: excellent for bounded attenuation and regions

Mediastinal/soft-tissue window

$$WL \approx 40\ \text{HU}, \qquad WW \approx 400\ \text{HU}.$$

$$\boxed{\mu_{\min} \leq X[i,j] \leq \mu_{\max}}$$



Brain CT: anatomy/structure priors

$$\boxed{\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \xrightarrow{\text{learning}} p_\theta(x)}$$

Concrete prior knowledge we could write down

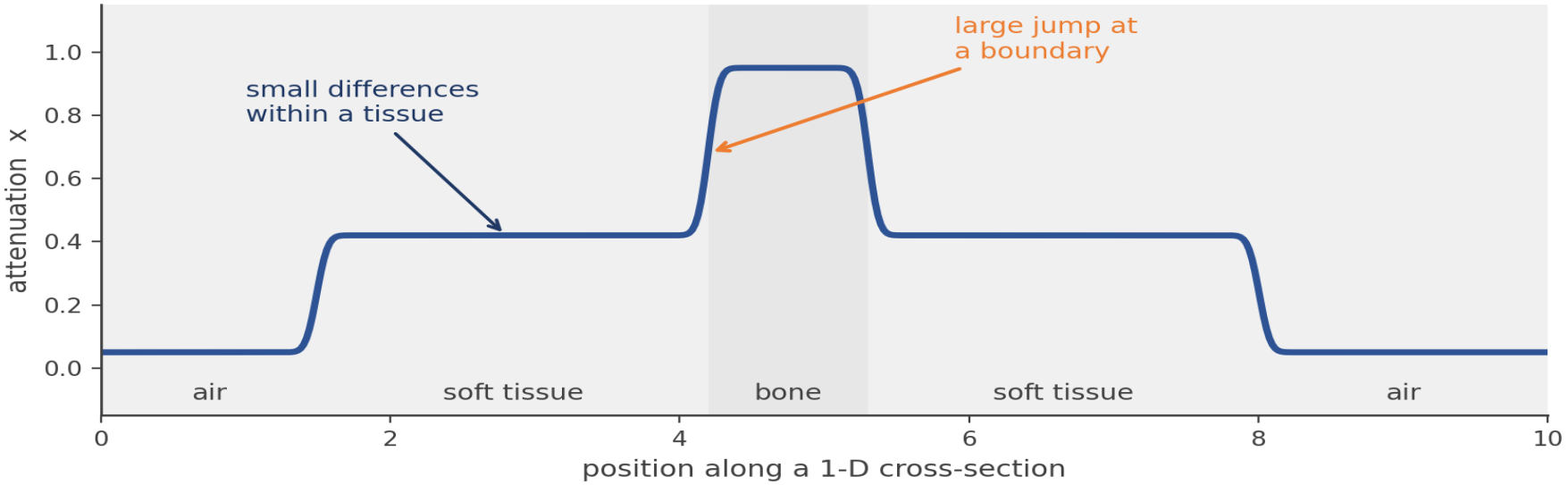
Physical knowledge	Mathematical statement
Attenuation cannot be negative	$x_i \geq 0$
Neighboring pixels within one tissue are similar	$x_i \approx x_j$
Sharp changes mainly occur at tissue boundaries	∇x is sparse
CT images are approximately piecewise smooth	TV(x) tends to be small
Outside the body is approximately air	known / estimated support

Bayesian inference requires us to translate physical knowledge like this into p(x).

We can build p(x) around these: E.g., CT images are approximately piecewise smooth.

(iii) Turning “Piecewise Smooth” Into a Probability

A 1-D cross-section: air | soft tissue | bone | soft tissue

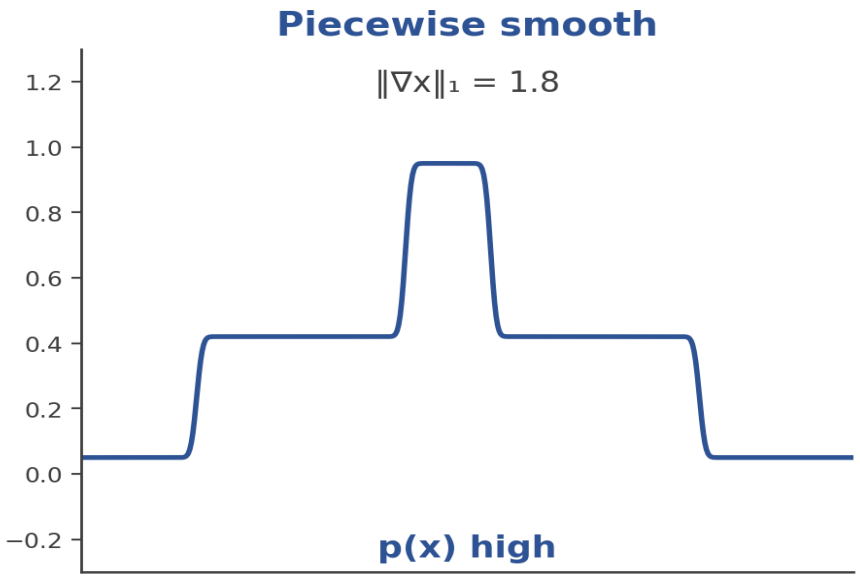


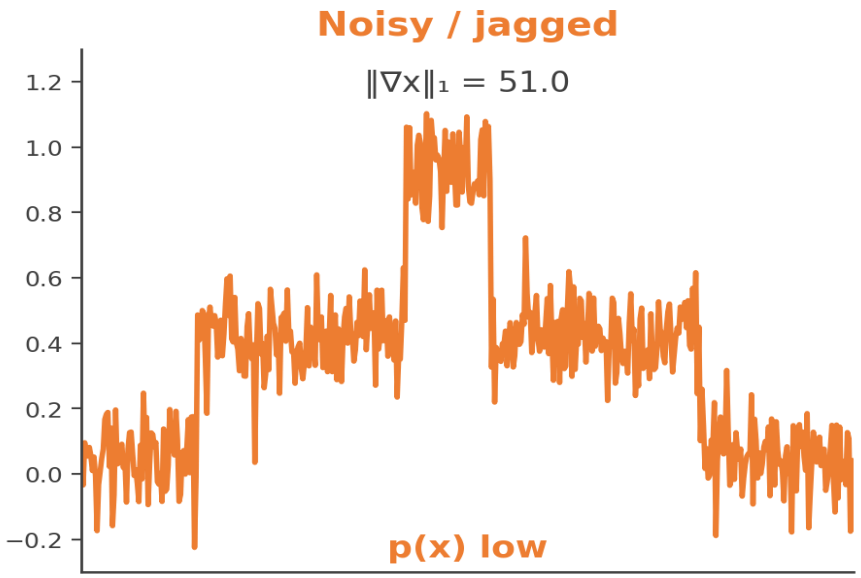
within a tissue: $x_{i+1} - x_i \approx 0$ *only near boundaries are differences large*

So ∇x should contain relatively few large values — measure this with R(x) = $\|\nabla x\|_1$.

Turn the statement into a probability distribution

$$p(x) \propto \exp[-\lambda \parallel \nabla x \parallel_1]$$





Small $\|\nabla x\|_1$ → high p(x). Large $\|\nabla x\|_1$ → low p(x). Important: Bayes did not discover that CT images are piecewise smooth — we encoded that knowledge in p(x).

(iv) MAP: Balance Measurement Fit and Prior

Combine the CT forward model with our piecewise-smooth prior

$$y = Ax + n, \qquad n \sim N(0, \sigma^2 I)$$

$$\boxed{p(y \mid x) \propto \exp\left[-\frac{1}{2\sigma^2} \| y - Ax \|_2^2\right]}$$

Combine it with our CT prior:

$$p(x) \propto \exp[-\lambda \| \nabla x \|_1].$$

Bayes combines them:

$$p(x \mid y) \propto \underbrace{e^{-\frac{1}{2\sigma^2}\|y-Ax\|_2^2}}_{\text{measurement consistency}} \quad \underbrace{e^{-\lambda\|\nabla x\|_1}}_{\text{prior plausibility}}$$

$$\boxed{\hat{x}_{\text{MAP}} = \arg\max_x p(x \mid y)}$$

Find an image that both explains the X-ray measurements AND looks physically plausible.

(v) Bayesian MAP Becomes Regularization

Take the negative logarithm

$$\hat{x}_{MAP} = arg\ min_x\ -log\ p(x|y)$$

Substitute the previous expressions:

$$\hat{x}_{MAP} = arg\ min_x\ [\ 1/2\sigma^2\ \|y-Ax\|_2^2\ +\ \lambda\|\nabla x\|_1\ ]$$

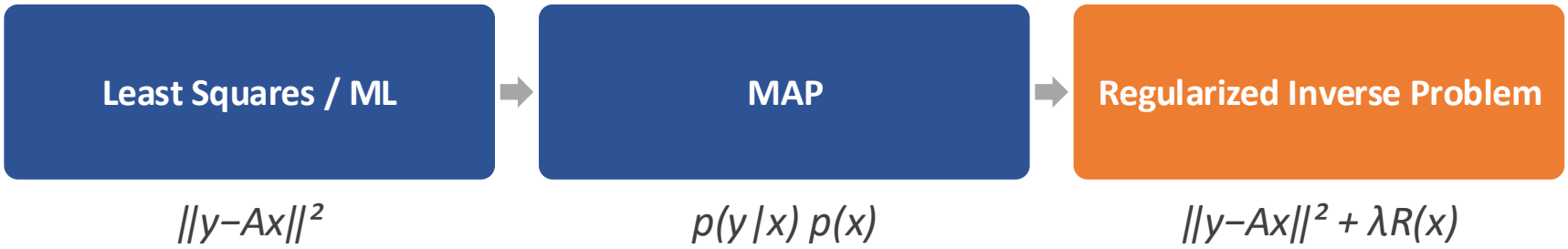
$$\|y-Ax\|_2^2 \quad + \quad \lambda'\|\nabla x\|_1$$

DATA FIDELITY **REGULARIZATION**

Prior probability, passed through −log, becomes a regularization penalty.

$$\hat{x}_{\text{MAP}} = \arg\min_{x} [\text{data fidelity} + \lambda R(x)].$$

The three-stage progression, now complete



Regularization is not an arbitrary extra term bolted onto least squares — it is the explicit mathematical representation of prior knowledge about the physical state x.

This is exactly the gap left open in the computational-inverse deck's "no room for prior knowledge" limitation — now we know precisely where $\lambda R(x)$ comes from.

This sets up iterative CT reconstruction, and eventually deep / unrolled tomography.

Questions?

Extra slides