Fitting a Line $y = mx + b$ via Least Squares (Vector/Matrix and Geometric Interpretation)

0) One-line goal

Choose a line $y = mx + b$ such that

$$\sum_{i=1}^{4} \left(y_i - (mx_i + b)\right)^2$$

is minimized.

1) Why write this in vector/matrix form?

Collect the data into a vector:

$$\boldsymbol{B} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 3 \end{bmatrix}$$

Since the line produces four predicted values at once,

$$A\boldsymbol{x} = \begin{bmatrix} mx_1 + b \\ mx_2 + b \\ mx_3 + b \\ mx_4 + b \end{bmatrix}$$

The unknowns are m and b, so define

$$\boldsymbol{x} = \begin{bmatrix} m \\ b \end{bmatrix}$$

Then A is the matrix that stacks the "ingredients (features)" needed to form $mx_i + b$ at each data point:

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix}$$

Why are the columns $(1, 2, 3, 4)^T$ and $(1, 1, 1, 1)^T$?

- First column: the x_i values multiplied by m
- Second column: the constant 1 multiplied by b

In other words, the line $mx + b$ is built from two basis components: the x term and the constant term.

2) Least squares in one line

Define the residual (error) vector

$$\boldsymbol{r} = \boldsymbol{B} - A\boldsymbol{x}$$

Then the quantity we minimize is simply

$$\|\boldsymbol{r}\|^2 = \|\boldsymbol{B} - A\boldsymbol{x}\|^2$$

3) Geometric interpretation (the key idea)

- $A\boldsymbol{x}$ always lies in the column space of A.
- The column space is

$$\mathrm{Col}(A) = \mathrm{span}\left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

  that is, the 2D subspace consisting of all vectors formed from the x-column and the constant column.

Thus, in words, the least squares problem says:
Project $\boldsymbol{B}$ orthogonally onto $\mathrm{Col}(A)$. The projected point is $A\boldsymbol{x}^*$, and the distance is minimized.

Question: "If we project $\boldsymbol{B}$ onto the space spanned by 1 and x, is that the closest point?" Yes. That projection is exactly the least squares solution.

4) "Why perpendicular?" What is perpendicular to what?

A fundamental property of orthogonal projection:

$$\boldsymbol{r} = \boldsymbol{B} - A\boldsymbol{x}^\star \quad \perp \quad \mathrm{Col}(A)$$

That is, the residual vector $\boldsymbol{r}$ is orthogonal to the entire column space.
In particular, for every vector $\boldsymbol{v}$ in the column space,

$$\boldsymbol{v}^T\boldsymbol{r} = 0$$

Important caution

It is NOT that $\boldsymbol{B}$ is perpendicular to $A\boldsymbol{x}$.

What is perpendicular is the **residual** $\boldsymbol{B} - A\boldsymbol{x}^*$ to the *plane* (the column space).
Since $A\boldsymbol{x}^* \in \mathrm{Col}(A)$, it also follows that

$$(A\boldsymbol{x}^*)^T(\boldsymbol{B} - A\boldsymbol{x}^*) = 0$$

5) Turning orthogonality into equations: why A^T appears

"Residual orthogonal to the column space" can be written using the column space generators (the columns of A):

- Every vector in $\mathrm{Col}(A)$ has the form $A\boldsymbol{z}$.
- $\boldsymbol{r} \perp \mathrm{Col}(A)$ means $\boldsymbol{r} \perp A\boldsymbol{z}$ for all $\boldsymbol{z}$.

Therefore,

$$(A\boldsymbol{z})^T\boldsymbol{r} = 0 \quad \forall \boldsymbol{z}$$
$$\boldsymbol{z}^T(A^T\boldsymbol{r}) = 0 \quad \forall \boldsymbol{z}$$

For this to hold for all $\boldsymbol{z}$, we must have

$$A^T\boldsymbol{r} = 0$$

that is,

$$A^T(\boldsymbol{B} - A\boldsymbol{x}) = 0$$

This is the *normal equation.*

Question: "Why do we multiply by A^T?" Because expressing "orthogonal to the column space" as "inner product with every column of A equals zero" naturally bundles all those conditions into the single equation $A^T(\boldsymbol{B} - A\boldsymbol{x}) = 0$.

6) The normal equation, summarized

$$A^T(\boldsymbol{B} - A\boldsymbol{x}) = 0 \quad \Rightarrow \quad A^TA\boldsymbol{x} = A^T\boldsymbol{B}$$
$$\boldsymbol{x} = (A^TA)^{-1}A^T\boldsymbol{B} \quad (\text{when } A^TA \text{ is invertible})$$

7) Least Squares from an ML Perspective (High-Dimensional Generalization)

1. Model setup

Consider the linear model

$$h_w(x) = \langle w, x \rangle = w^Tx$$

- $x_i \in \mathbb{R}^d$: input vector
- $w \in \mathbb{R}^d$: weight vector
- If a bias term is present, it can be absorbed by augmenting the input as $x_i = (x_i, 1)$.

This is the high-dimensional generalization of $y = mx + b$.

2. Objective function (Loss function)

Given a dataset $\{(x_i, y_i)\}_{i=1}^m$, we consider

$$\min_w \frac{1}{m}\sum_{i=1}^m (w^Tx_i - y_i)^2$$

For convenience, the constant factor $\frac{1}{m}$ can be omitted without changing the minimizer:

$$L(w) = \sum_{i=1}^m (w^Tx_i - y_i)^2$$

This is the **least squares** objective.

3. Gradient computation (key step)

Differentiating each term yields

$$\nabla_w (w^T x_i - y_i)^2 = 2x_i(w^T x_i - y_i)$$

Hence, the full gradient is

$$\nabla_w L(w) = 2\sum_{i=1}^{m} x_i(w^T x_i - y_i)$$

At the minimum, $\nabla_w L(w) = 0$, so

$$\sum_{i=1}^{m} x_i(w^T x_i - y_i) = 0$$

Up to this point, the derivation follows the **ML (calculus-based) perspective**.

4. Separating terms

Expanding the equation above gives

$$\sum_{i=1}^{m} x_i \langle w, x_i \rangle = \sum_{i=1}^{m} x_i y_i$$

We now reinterpret this expression in matrix form.

5. Matrix reinterpretation (key transformation)

Define the data matrix

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times d}, \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

Left-hand side

$$\sum_{i=1}^{m} x_i(w^T x_i) = \sum_{i=1}^{m} x_i x_i^T w = \left(\sum_{i=1}^{m} x_i x_i^T \right) w$$

But

$$X^T X = \sum_{i=1}^{m} x_i x_i^T$$

Right-hand side

$$\sum_{i=1}^{m} x_i y_i = X^T y$$

6. Normal equation

Therefore, we obtain

$$X^T X w = X^T y$$

This is exactly the same equation as

$$A^T A x = A^T B$$

with the notation change

$$A = X, \quad x = w, \quad B = y$$

7. Closed-form solution

If $X^T X$ is invertible, the solution is

$$\boxed{w = (X^T X)^{-1} X^T y}$$

Remarks

- X is typically not a square matrix, so it does not have an inverse.
- However, $X^T X$ is square and invertible when X has full column rank.

8. Geometric interpretation (connection to the previous figure)

- Xw : a vector in the column space $\mathrm{Col}(X)$
- y : the true target vector
- $y - Xw$: the residual

The least squares condition is

$$y - Xw \perp \mathrm{Col}(X)$$

which can be written as

$$X^T(y - Xw) = 0 \quad \Longleftrightarrow \quad X^T X w = X^T y$$

Thus,

the ML gradient condition = the linear algebra orthogonal projection condition.

9. One-page summary (for exams and interviews)

- Linear model:
$$h_w(x) = w^T x$$
- Objective:
$$\min_w \|Xw - y\|^2$$
- Optimality condition:
$$X^T(Xw - y) = 0$$
- Solution:
$$w = (X^T X)^{-1} X^T y$$
- Interpretation:
$$\text{Orthogonal projection of } y \text{ onto } \mathrm{Col}(X)$$

8) Inner Product Spaces of Functions: Why, What, and How

1. Bottom line first

The phrase **"inner product space of functions"** means:

"We will treat functions not as points, but as *vectors*."

As a result, for functions we can perform exactly the same geometric operations as for vectors:

- addition
- scalar multiplication
- measuring length
- measuring angles
- orthogonal projection

2. Why are functions vectors?

The essence of a vector is not its *coordinates*, but its *algebraic structure*.

To be a vector space, only three conditions are required:

1. addition is defined
2. scalar multiplication is defined
3. the result stays in the same space

For functions,

$$(f+g)(x) = f(x) + g(x), \qquad (cf)(x) = cf(x)$$

are always well-defined.

Therefore,

$$\{f : [0,1] \to \mathbb{R}\}$$

is an **infinite-dimensional vector space**.

3. Then what is an "inner product"?

In Euclidean space, the inner product is

$$\langle v, w \rangle = v^T w$$

and it encodes

- alignment of directions
- orthogonality
- the building block of length

We want the same roles to hold for functions. This motivates the definition

$$\boxed{\langle f, g \rangle = \int_0^1 f(x)g(x)\,dx}$$

This definition is not arbitrary.

4. Why this integral is a genuine inner product

To qualify as an inner product, three properties must hold.

(1) Linearity

$$\langle af + bg, h\rangle = a\langle f, h\rangle + b\langle g, h\rangle$$

This follows directly from the linearity of integration.

(2) Symmetry

$$\langle f, g\rangle = \langle g, f\rangle$$

This follows from the symmetry of multiplication.

(3) Positive definiteness

$$\langle f, f\rangle = \int_0^1 f(x)^2 dx \geq 0, \quad = 0 \iff f = 0 \text{ (almost everywhere)}$$

Hence, this is a **valid inner product**.

5. Geometric meaning of this inner product

Length

$$\|f\| = \sqrt{\langle f, f\rangle} = \left(\int_0^1 f(x)^2 dx\right)^{1/2}$$

This represents the "energy" or total magnitude of the function.

Angle

$$\cos\theta = \frac{\langle f, g\rangle}{\|f\|\,\|g\|}$$

The more similar the shapes of two functions, the smaller the angle.

Orthogonality

$$\langle f, g\rangle = 0$$

means the functions are uncorrelated on average.

For example,

$$1 \perp x - \tfrac{1}{2}$$

because the constant component and the left–right oscillating component are independent.

6. Approximation = orthogonal projection

The core problem is to find, within the function space

$$\text{span}\{1, x\},$$

the point that is closest to a given function f.

This is exactly an **orthogonal projection**, whose defining property is always

$$f - \hat{f} \;\perp\; \text{span}\{1, x\}.$$

7. Why a system of equations appears

Since the basis $\{1, x\}$ is not orthogonal, the orthogonality condition must be imposed separately:

$$\langle f - a - bx, 1\rangle = 0$$

$$\langle f - a - bx, x\rangle = 0$$

This leads directly to a system of equations.

8. The true role of Gram–Schmidt

Gram–Schmidt does not invent new formulas. It performs exactly one task:

"It redefines mutually orthogonal axes within the same plane."

That is,

$$\{1, x\} \longrightarrow \left\{1,\ x - \tfrac{1}{2}\right\}$$

With an orthogonal basis, the projection formula simplifies to

$$\text{coefficient} = \frac{\text{inner product}}{\text{self inner product}}$$

one clean line at a time.

9. Core intuition summary

- Functions are vectors
- Integrals are dot products
- Least squares minimizes distance
- Gram–Schmidt straightens the coordinate system

Therefore,

"An inner product space of functions"

is a language for treating functions as *geometric objects*, not pictures.

9) A Complete Generalization of Least Squares via Gram–Schmidt

0. What we have really been doing

In fact, we have already been working with the following setup:

- Space: a function space
- Inner product: $\langle f, g\rangle = \displaystyle\int_0^1 f(x)g(x)\,dx$
- Approximation space: $\text{span}\{1, x\}$
- Goal: **orthogonal projection** of f onto this space

The only reason a system of equations appeared is that

- we kept the original basis $\{1, x\}$, and
- that basis was not orthogonal.

Gram–Schmidt is the tool that **automates** this process.

1. General problem formulation (the real generalization)

Instead of approximating by a line, consider the general approximation

$$f(x) \approx a_0\phi_0(x) + a_1\phi_1(x) + \cdots + a_n\phi_n(x).$$

Here,

$$\{\phi_0, \phi_1, \ldots, \phi_n\}$$

is an arbitrary set of functions (e.g. $1, x, x^2, \ldots$).

2. Without orthogonalization, a system always appears

The least squares condition is

$$\left\langle f - \sum_{k=0}^{n} a_k\phi_k,\ \phi_j \right\rangle = 0 \qquad (j = 0, \ldots, n).$$

Expanding this gives

$$\sum_{k=0}^{n} a_k\langle\phi_k, \phi_j\rangle = \langle f, \phi_j\rangle.$$

That is,

$$Ga = b,$$

where

$$G_{jk} = \langle\phi_k, \phi_j\rangle.$$

The matrix G is the **Gram matrix**, and a system of equations is unavoidable in this form.

3. Why Gram–Schmidt enters

Now we change the basis:

$$\{\phi_0, \phi_1, \ldots, \phi_n\} \quad \longrightarrow \quad \{u_0, u_1, \ldots, u_n\}.$$

There is only one requirement:

$$\langle u_i, u_j\rangle = 0 \quad (i \neq j).$$

This transformation is precisely the **Gram–Schmidt orthogonalization**.

4. The magic of an orthogonal basis

Using the same approximation in an orthogonal basis,

$$f \approx c_0u_0 + c_1u_1 + \cdots + c_nu_n,$$

the least squares condition becomes

$$\left\langle f - \sum_{k=0}^{n} c_ku_k,\ u_j \right\rangle = 0.$$

Expanding,

$$\langle f, u_j\rangle - c_j\langle u_j, u_j\rangle = 0.$$

Therefore,

$$\boxed{c_j = \frac{\langle f, u_j\rangle}{\langle u_j, u_j\rangle}}$$

The system of equations **disappears completely**.

5. Plugging your problem into this framework

(1) Original basis

$$\phi_0 = 1, \qquad \phi_1 = x.$$

(2) Gram–Schmidt result

$$u_0 = 1,$$
$$u_1 = x - \frac{\langle x, 1\rangle}{\langle 1, 1\rangle} \cdot 1 = x - \frac{1}{2}.$$

(3) Coefficients

$$c_0 = \frac{\langle f, 1\rangle}{\langle 1, 1\rangle}, \qquad c_1 = \frac{\langle f, x - \frac{1}{2}\rangle}{\langle x - \frac{1}{2}, x - \frac{1}{2}\rangle}.$$

These are exactly the same expressions you already computed.

6. Why this is a true generalization

Now the procedure is independent of the choice of basis.

For example,

$$\text{span}\{1, x, x^2\}$$

1. Apply Gram–Schmidt to obtain an orthogonal basis
2. Project onto each direction
3. Coefficients are always given by

$$c_k = \frac{\langle f, u_k\rangle}{\langle u_k, u_k\rangle}$$

- As the dimension increases,
- as the target function changes,
- or as the interval changes,

the structure never changes.

7. One key sentence

To generalize via Gram–Schmidt means

"to replace solving a least squares problem via systems of equations with a sequence of one-dimensional orthogonal projections."

8. Essential takeaways

- The problem you solved corresponds to the special case $n = 1$
- Gram–Schmidt extends this to arbitrary degree n
- The goal is not cosmetic simplification, but

fixing the structure of the problem

10) A Zoomed-In View: Least Squares, Projection, and Ellipsoidal Maximization

1. Starting point: Why split $\langle \theta, a \rangle$ this way?

We begin with the simple identity

$$\langle \theta, a \rangle = \langle \hat{\theta}, a \rangle + \langle \theta - \hat{\theta}, a \rangle.$$

Here:

- $\hat{\theta}$: the parameter estimate we currently have
- θ : the true (unknown) parameter
- $u := \theta - \hat{\theta}$: the estimation error (uncertainty vector)

Thus,

$$\langle \theta, a \rangle = \underbrace{\langle \hat{\theta}, a \rangle}_{\text{current prediction}} + \underbrace{\langle u, a \rangle}_{\text{potential increase due to uncertainty}}.$$

Therefore, the question

"How large can this be in the worst (or best) case?"

reduces to the optimization problem

$$\max_{u} \langle u, a \rangle.$$

2. Can u grow arbitrarily? (No $\rightarrow$ constraints)

The uncertainty vector u is not free to grow arbitrarily. In our setting, it satisfies the constraint

$$u^T V u \leq \beta^2.$$

This means:

- length is measured using the metric induced by V
- u must lie within a radius β under this metric

Geometrically, this defines an ellipsoid:

$$\{u : u^T V u \leq \beta^2\}.$$

Why does V appear here? (Connection to least squares) In linear regression, the matrix

$$X^T X$$

represents the amount of information provided by the data:

- rich, well-spread data $\Rightarrow$ stable estimates
- limited or one-directional data $\Rightarrow$ weak directions of $X^T X$

Statistically, one typically has

$$\mathrm{Cov}(\hat{\theta}) \approx \sigma^2 (X^T X)^{-1}.$$

Hence, uncertainty regions naturally take the form of ellipsoids. In this context, V plays the role of an *information matrix*, typically

$$V \approx X^T X \quad (\text{or } X^T X + \lambda I).$$

3. The core geometric problem

We now arrive at the central optimization problem:

$$\max_u \langle u, a \rangle \quad \text{subject to} \quad u^T V u \leq \beta^2.$$

Geometric interpretation:

- $\langle u, a \rangle$ measures how far u extends in the direction a
- $u^T V u \leq \beta^2$ restricts u to lie inside an ellipsoid

In words:

> Find the point inside the ellipsoid that extends farthest in direction a.

This is directly analogous to orthogonal projection:

- least squares: project B onto a subspace
- here: find the *support point* of the ellipsoid in direction a

4. Why the optimizer points in the direction $V^{-1}a$

A key conclusion is

$$u \parallel V^{-1}a.$$

Intuition (the most important part)

- directions where V is large are *tight*: small movements quickly violate the constraint
- directions where V is small are *loose*: larger movements are allowed

To maximize the projection onto a, the optimizer exploits these loose directions in exactly the way encoded by V^{-1}. Thus, the optimal direction becomes $V^{-1}a$.

5. Exact derivation via Lagrange multipliers

Since the maximum occurs on the boundary, impose $u^T V u = \beta^2$ and define

$$\mathcal{L}(u, \lambda) = a^T u - \lambda(u^T V u - \beta^2).$$

Setting the gradient to zero:

$$\nabla_u \mathcal{L} = a - 2\lambda V u = 0$$

which gives

$$Vu = \frac{1}{2\lambda} a, \qquad u = \frac{1}{2\lambda} V^{-1} a.$$

Thus,

$$u = c\, V^{-1} a.$$

6. Determining the scaling constant

Imposing the constraint:

$$u^T V u = \beta^2$$

yields

$$(cV^{-1}a)^T V (cV^{-1}a) = c^2 a^T V^{-1} a = \beta^2.$$

Hence,

$$c = \frac{\beta}{\sqrt{a^T V^{-1} a}}.$$

7. Final result (boxed conclusion)

The maximum value is

$$\max_u a^T u = a^T (cV^{-1}a) = \beta\sqrt{a^T V^{-1} a}.$$

Therefore,

$$\boxed{\max_\theta \langle \theta, a \rangle = \langle \hat{\theta}, a \rangle + \beta\sqrt{a^T V^{-1} a}}$$