Least Squares

Practical Linear Algebra | Lecture 7

Problem Setup

$$y = Ax$$

$$A \in \mathbb{R}^{m \times n}$$

- For a given y, we want to solve for x
- If A is square and invertible, we can solve for x exactly, where $x = A^{-1}y$
- Interpretation: same number of equations and unknowns, and all equations are unique

Problem Setup

$$y = Ax$$

$$A \in \mathbb{R}^{m \times n}$$

- In real life, we rarely have the same number of equations and unknowns
- Instead, we often have more equations than unknowns ($m > n$)
- In general, we can't satisfy all equations simultaneously
- *The next best thing we can do is to try to satisfy all equations as closely as possible*

Residuals

- How do we measure closeness?
- Each element of Ax should be close to each corresponding element of y
- For a specific value of x, the residual vector $y - Ax$ represents how closely each equation is satisfied
- If we *square* each element of the residual vector and *sum* them, we get a single number that represents how closely all equations are satisfied
- We call this number the residual sum of squares (RSS)

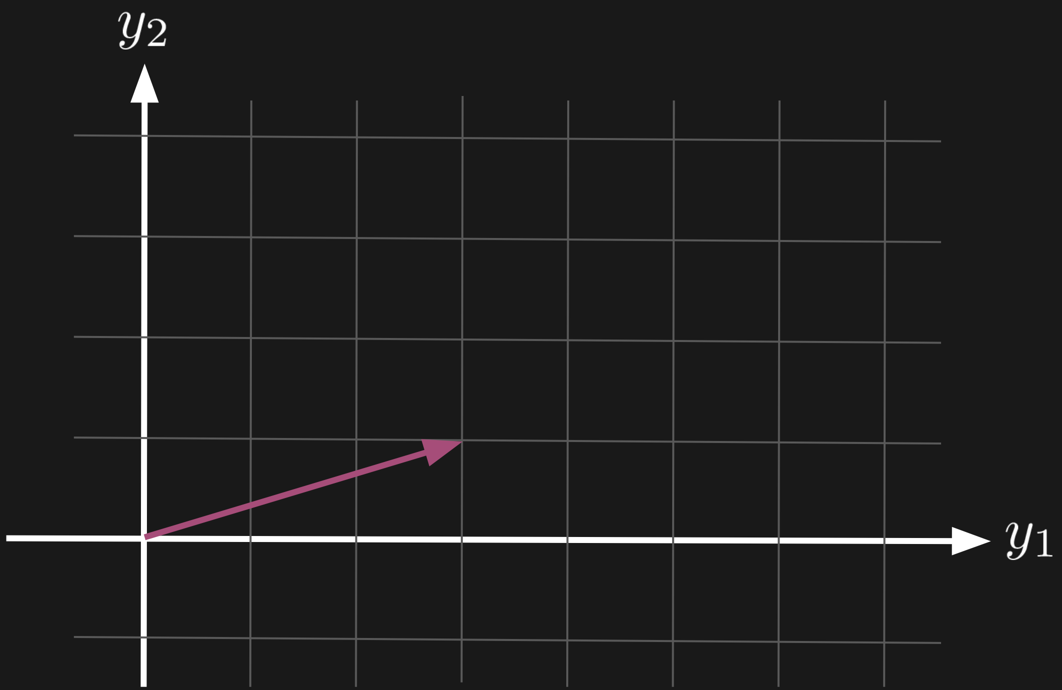
Least Squares

- Idea: Let's choose x to minimize the RSS
- This is the method of least squares

Geometric Interpretation

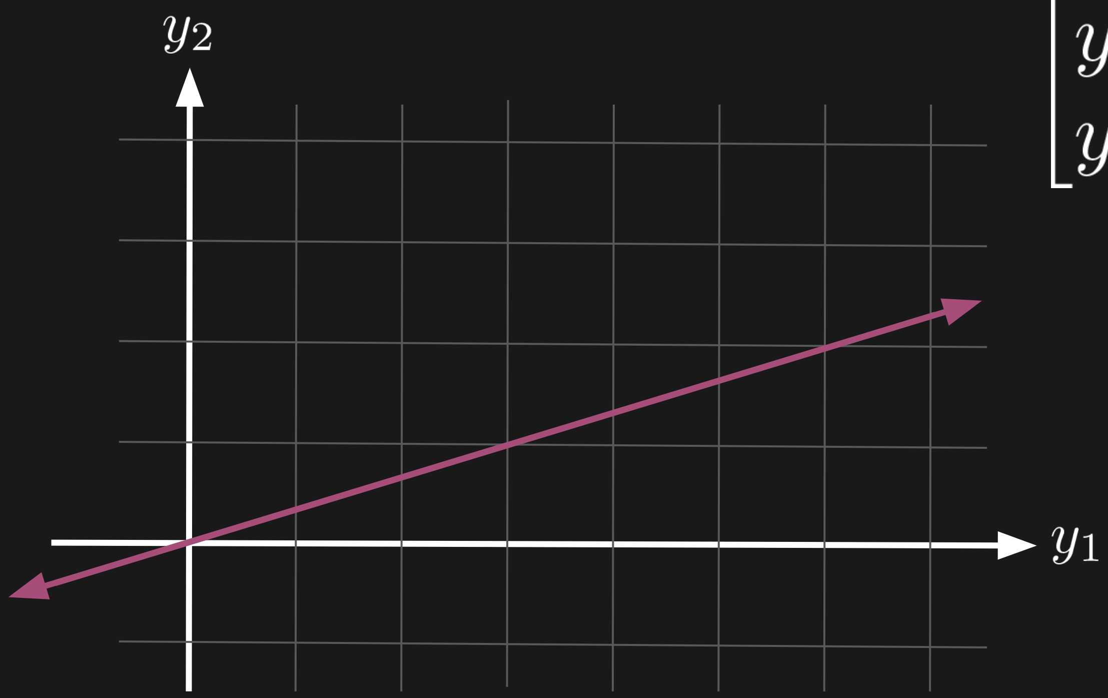
- Recall the definition of $\mathbf{range}(A)$
- Ax is the set of all possible output vectors – it forms a linear subspace
- In general, $y \notin \mathbf{range}(A)$
- The next best thing we can do is the find the vector Ax that is closest to y in terms of distance
- The distance between the vector y and Ax is just the norm of $y - Ax$
- Geometrically, we can just project y onto the subspace Ax

2D Example



$$\begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

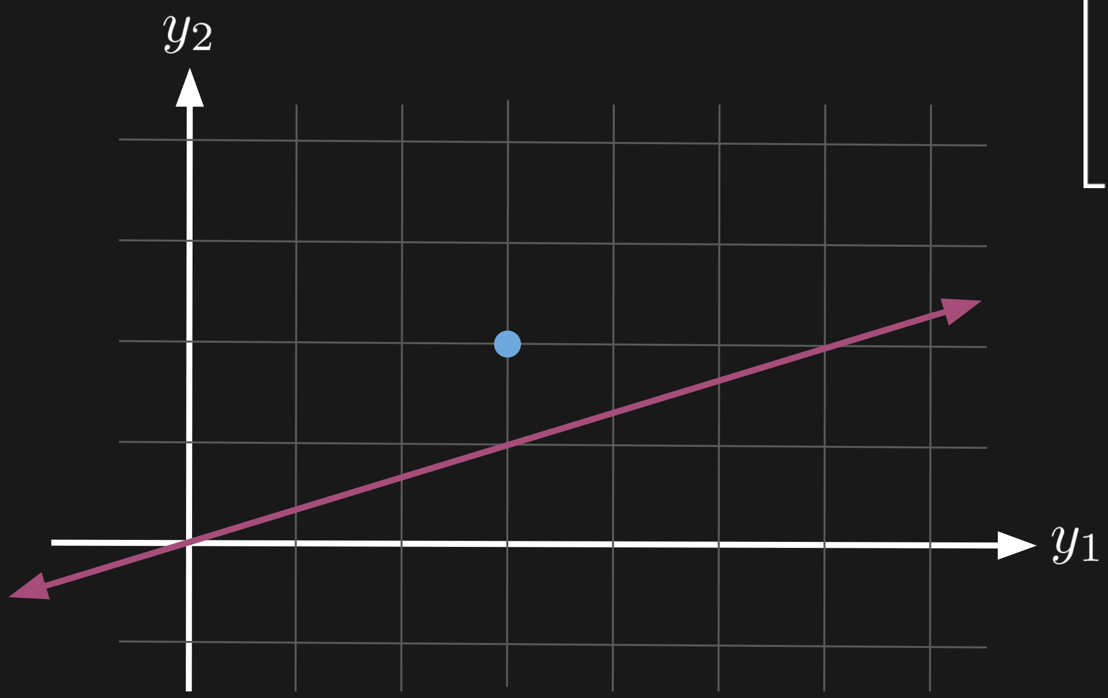
2D Example



$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \end{bmatrix}$$

$$y = Ax$$

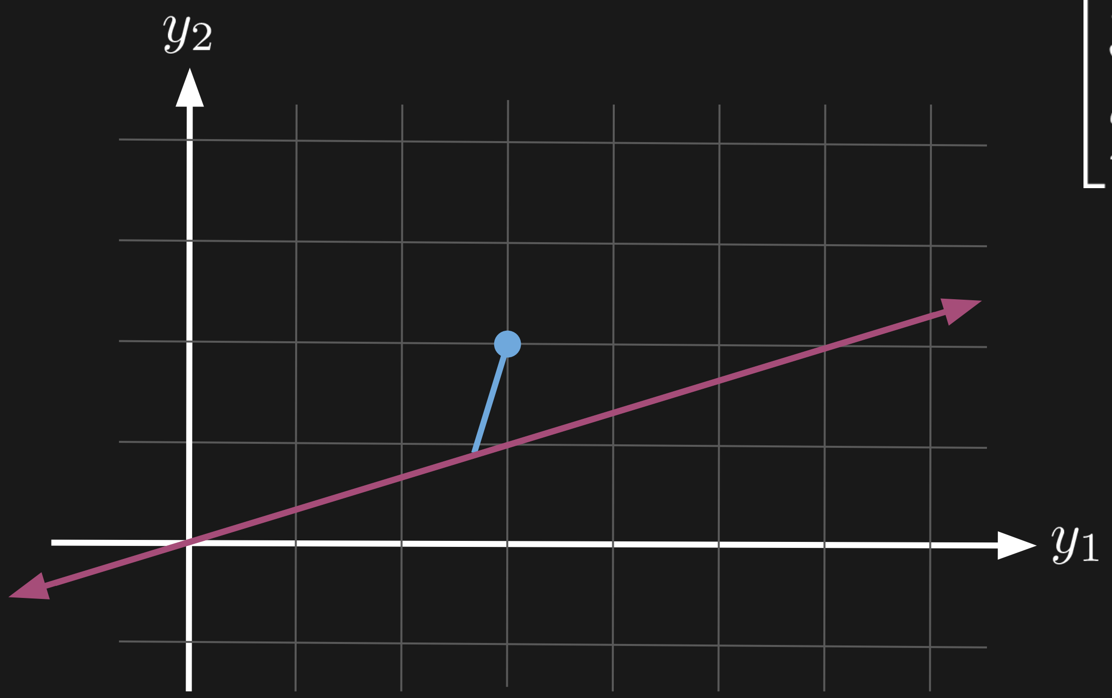
2D Example



$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \end{bmatrix}$$

$$y = Ax$$

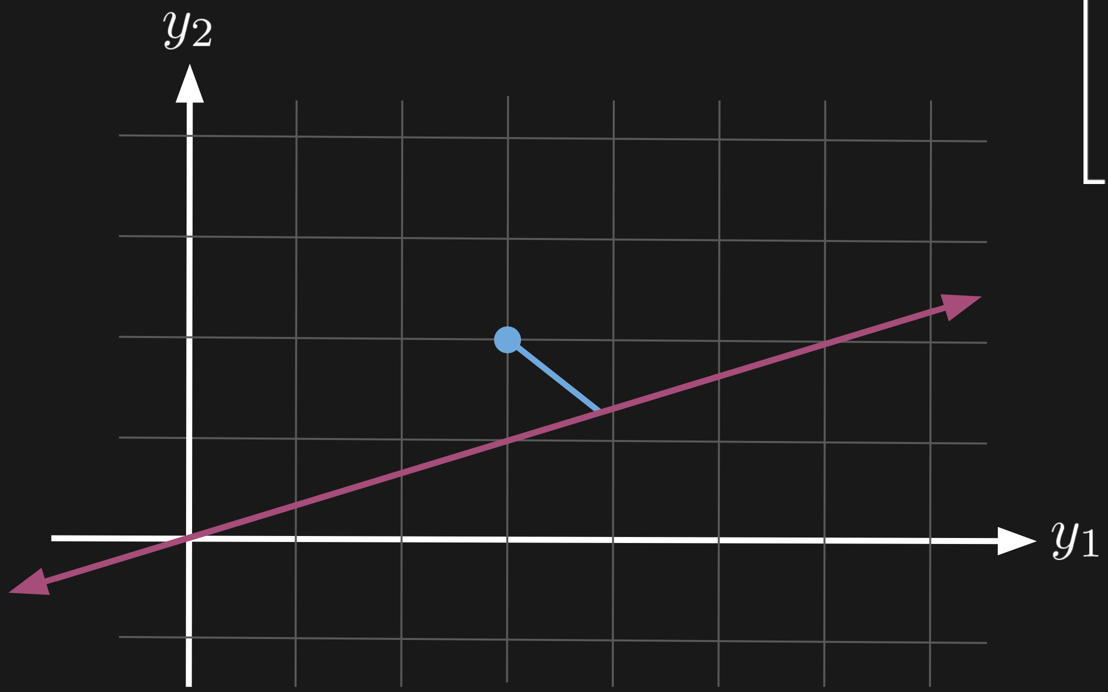
2D Example



$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \end{bmatrix}$$

$$y = Ax$$

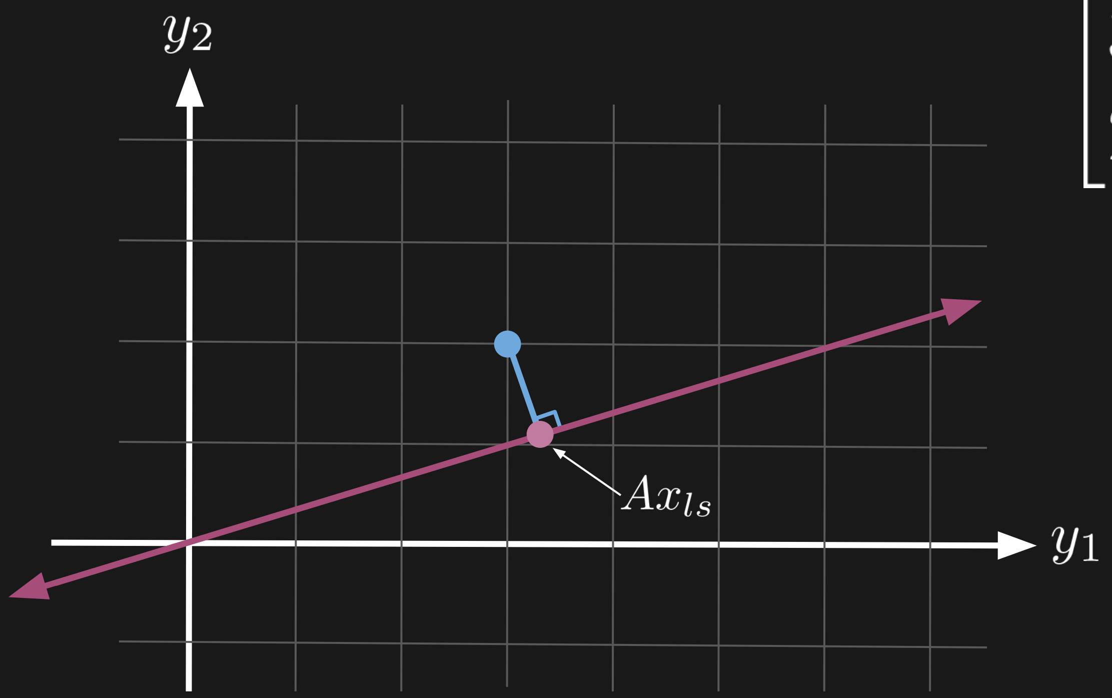
2D Example



$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \end{bmatrix}$$

$$y = Ax$$

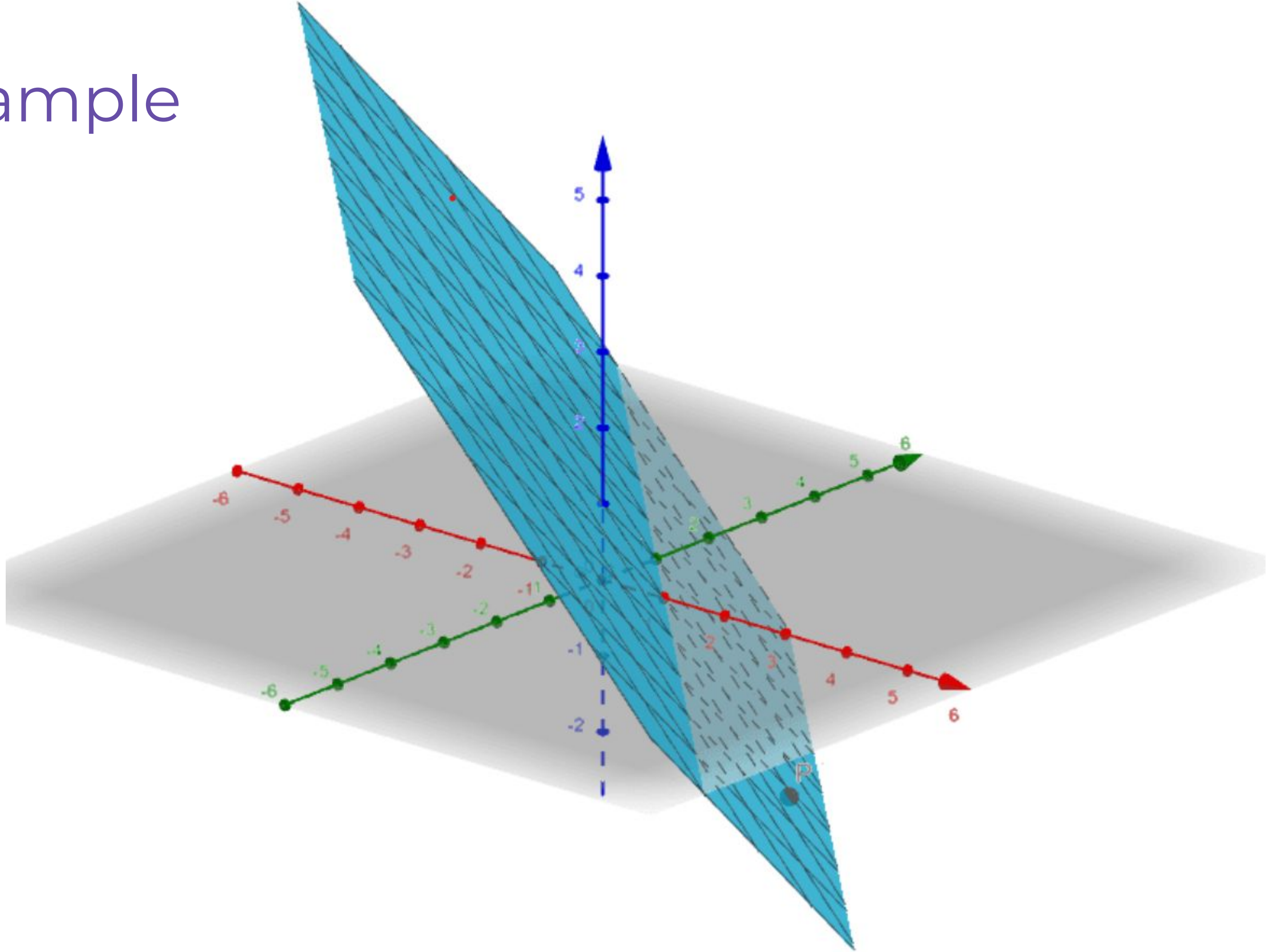
2D Example



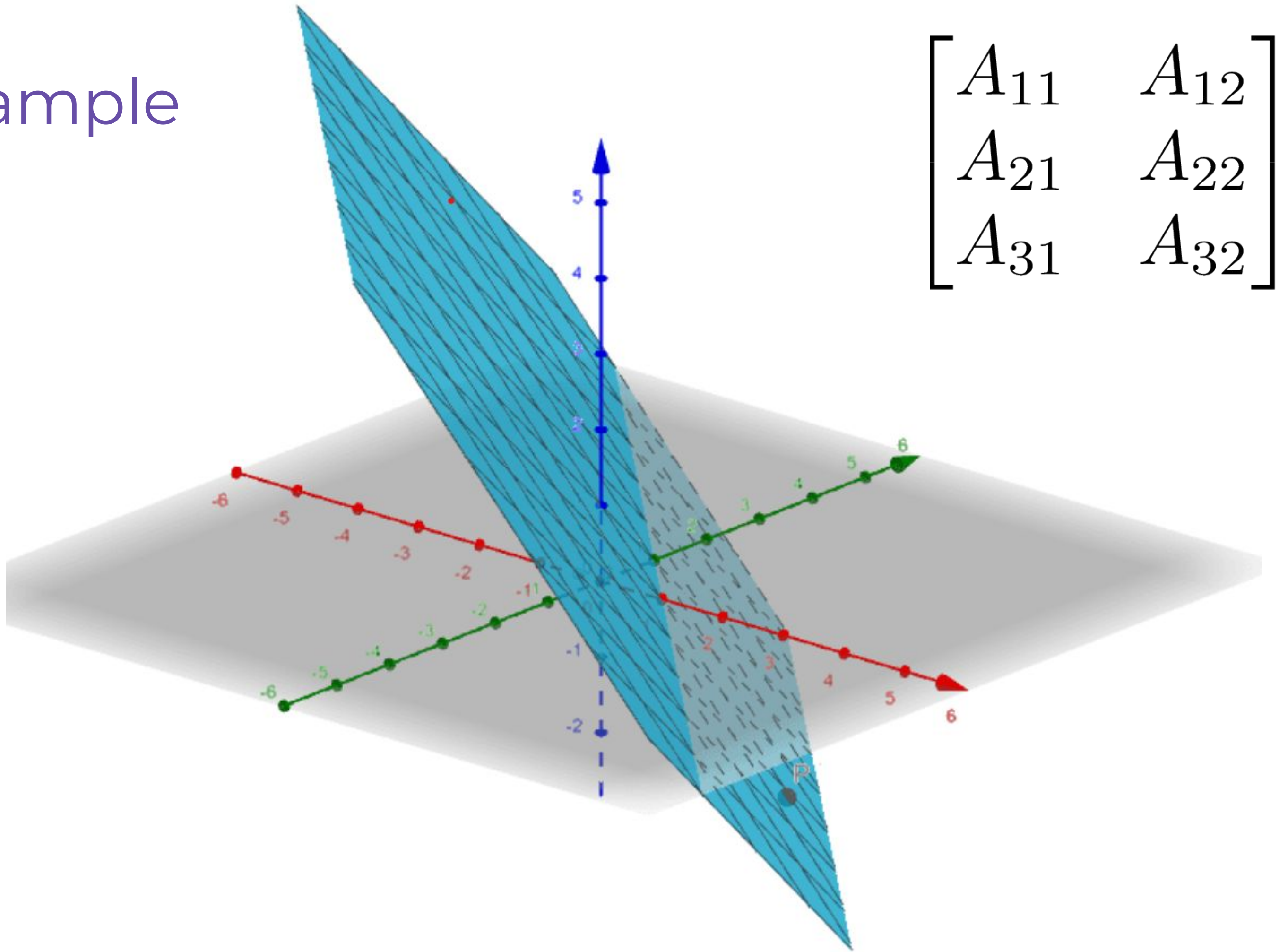
$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \end{bmatrix}$$

$$y = Ax$$

3D Example

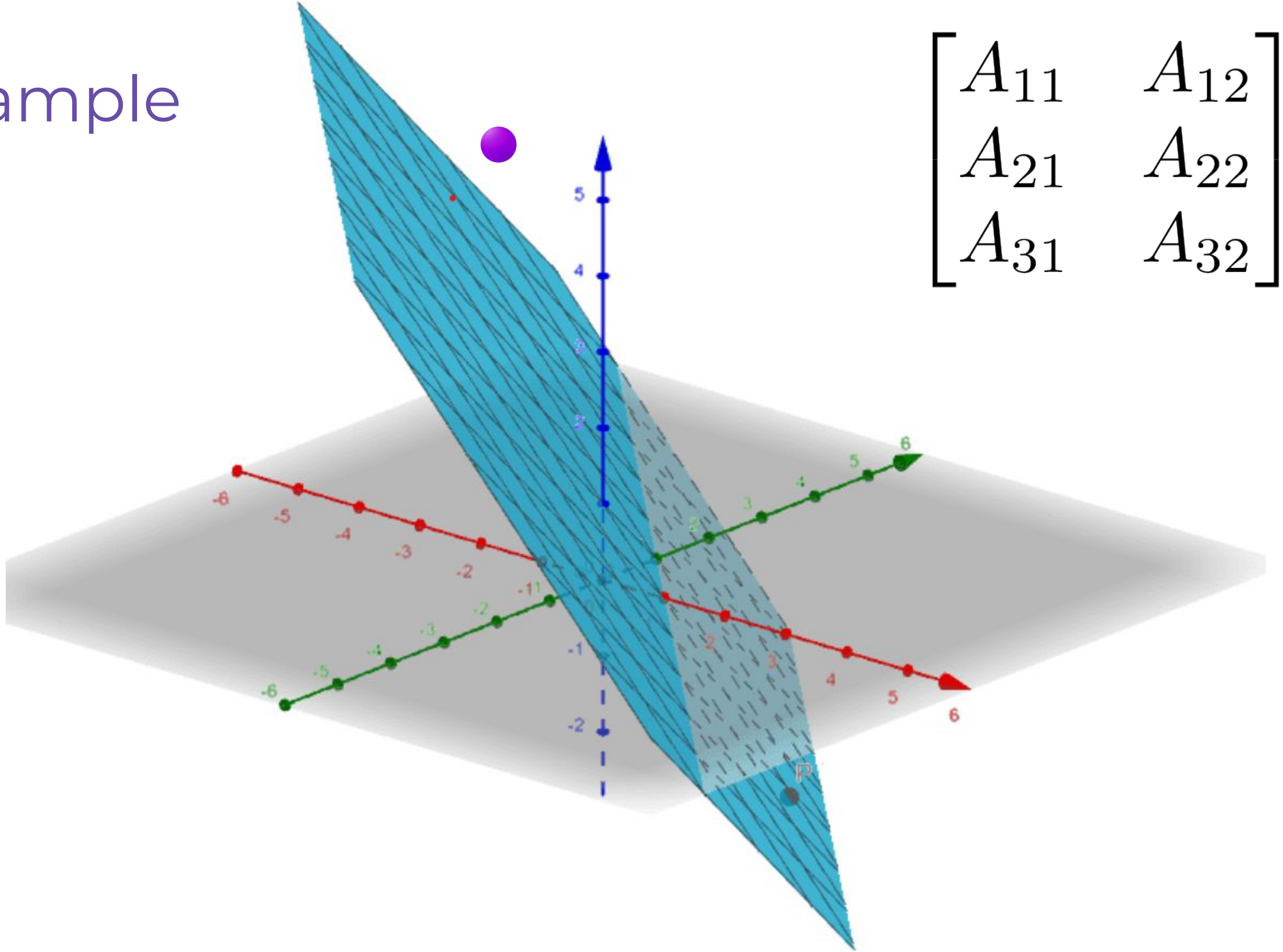


3D Example



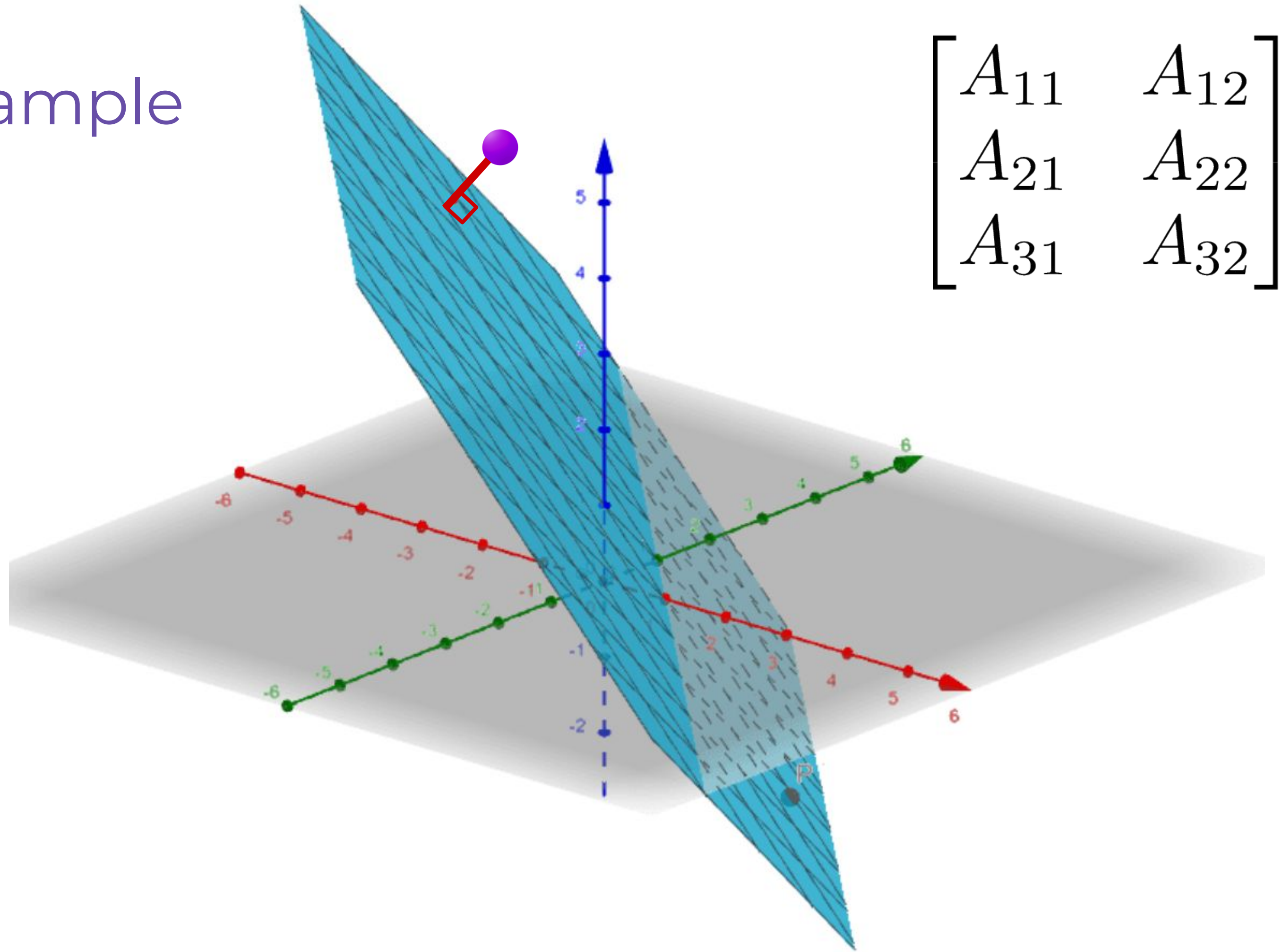
$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

3D Example



$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

3D Example



$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Closed-Form Solution

- Let's actually calculate x_{ls}

$$
\begin{aligned}
\|y - Ax\|^2 &= (y - Ax)^T (y - Ax) \\
&= (y^T - x^T A^T)(y - Ax) \\
&= y^T y - 2y^T Ax + x^T A^T Ax
\end{aligned}
$$

Closed-Form Solution

- Take the gradient with respect to x and set it to zero

$$
\begin{aligned}
\nabla_x \|y - Ax\|^2 &= \nabla_x (y^T y - 2y^T Ax + x^T A^T Ax) \\
&= \nabla_x (y^T y) - \nabla_x (2y^T Ax) + \nabla_x (x^T A^T Ax) \\
&= 0 - 2A^T y + 2A^T Ax = 0
\end{aligned}
$$

$$\Rightarrow A^T Ax = A^T y$$ the normal equation

Pseudoinverse

- Fact: If A is skinny and full rank, then $A^T A$ is invertible (proof left to you)

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

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

$$x_{ls} = A^\dagger y$$

AKA Moore-Penrose inverse

where $A^\dagger \equiv (A^T A)^{-1} A^T$ is the pseudoinverse of A

Projection Matrix

- Note that $\hat{y} \equiv Ax_{ls}$ is the vector in $\mathbf{range}(A)$ that is closest to y

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

$$\hat{y} = AA^\dagger y$$

- We call $AA^\dagger$ the projection matrix because it projects y onto $\mathbf{range}(A)$
- Also known as the hat matrix because it puts a hat on y

Next Time

- Statistical interpretation of least squares
- Applications of least squares