CS189: Intro to Machine Learning
Summer 2018

Lecture 1: Linear Regression

Josh Tobin
UC Berkeley EECS

Outline for today

- Reminders
- Review of OLS
- Intro to features

Outline for today

- **Reminders**
- Review of OLS
- Intro to features

Reminders

- Sign up for piazza
- Sign up for gradescope
- HW0 due Thursday
- Lecture slides will be posted tomorrow morning

Outline for today

- Reminders
- **Review of OLS**
- Intro to features

Ordinary least squares

$$y \approx Xw$$

Ordinary least squares

$$y \approx Xw$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

Goal: find the best w

Four levels for ML problems

1. **Data & application**
2. Model
3. Optimization problem
4. Optimization algorithm

Data & application layer

$$y \approx Xw$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

Goal: find the best w

Four levels for ML problems

1. Data & application
2. **Model**
3. Optimization problem
4. Optimization algorithm

2. Model

$$y \approx Xw$$

$$\hat{y} = Xw$$

Four levels for ML problems

1. Data & application
2. Model
3. **Optimization problem**
4. Optimization algorithm

3. Optimization problem

$$y \approx Xw$$

$$\min_{w} ||Xw - y||_2^2$$

Goal: find the best w

Four levels for ML problems

1. Data & application
2. Model
3. Optimization problem
4. **Optimization algorithm**

4. Optimization algorithms

$$\min_{w} ||Xw - y||_2^2$$

$$\hat{w} = \left(X^T X\right)^{-1} X^T y$$

4. Optimization algorithms

$$\min_w ||Xw - y||_2^2$$

$$X^T X \hat{w} = X^T y$$

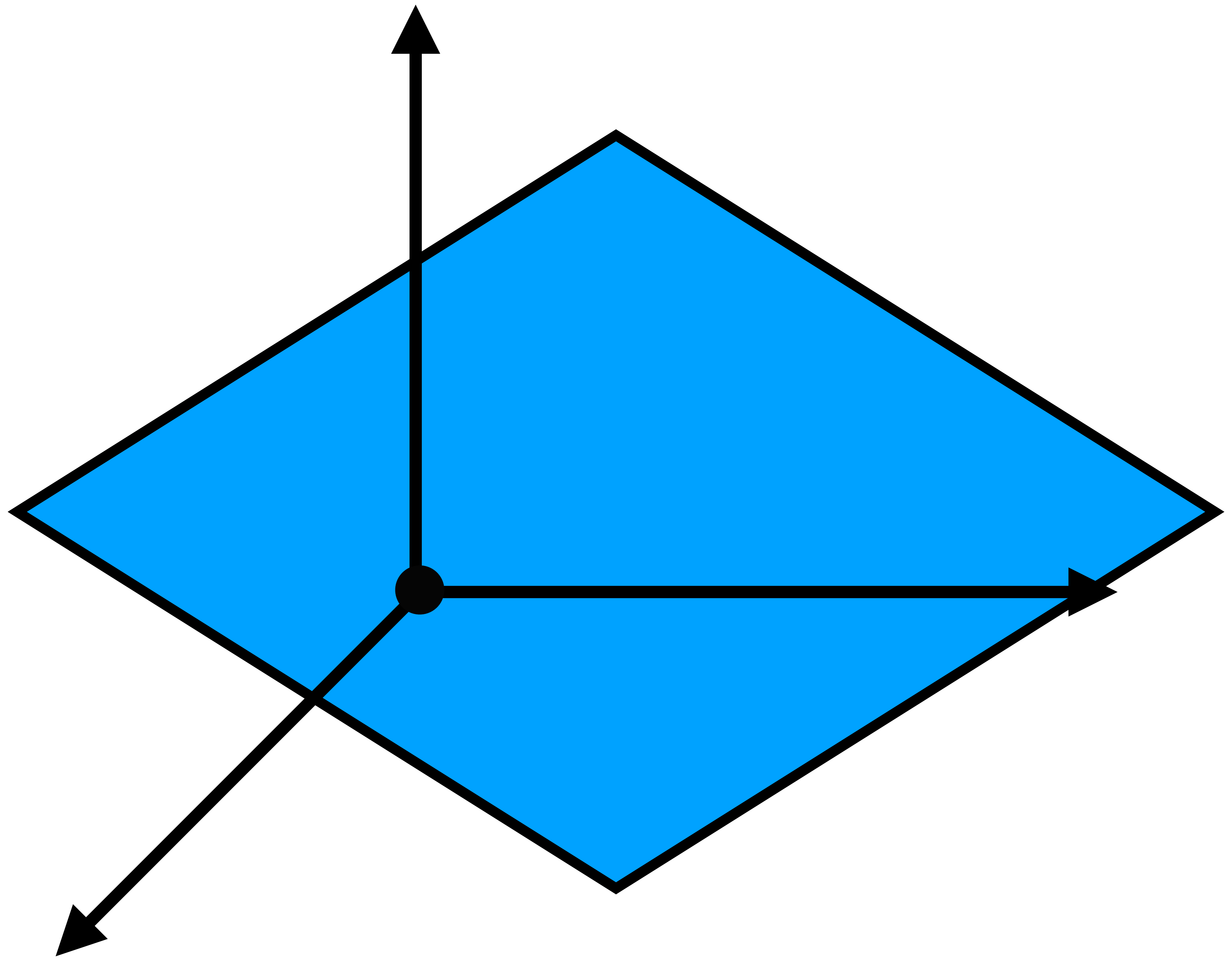
Why are normal equations the solution?

$$X = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_n^T \end{bmatrix} = \begin{bmatrix} \vec{x}^1 & \cdots & \vec{x}^d \end{bmatrix} = \begin{bmatrix} x_1^1 & \cdots & x_1^d \\ \vdots & & \\ \vdots & & \\ x_n^1 & \cdots & x_n^d \end{bmatrix} \qquad n >> d$$

$$\hat{w} = \left(X^T X\right)^{-1} X^T y$$

Why are normal equations the solution?

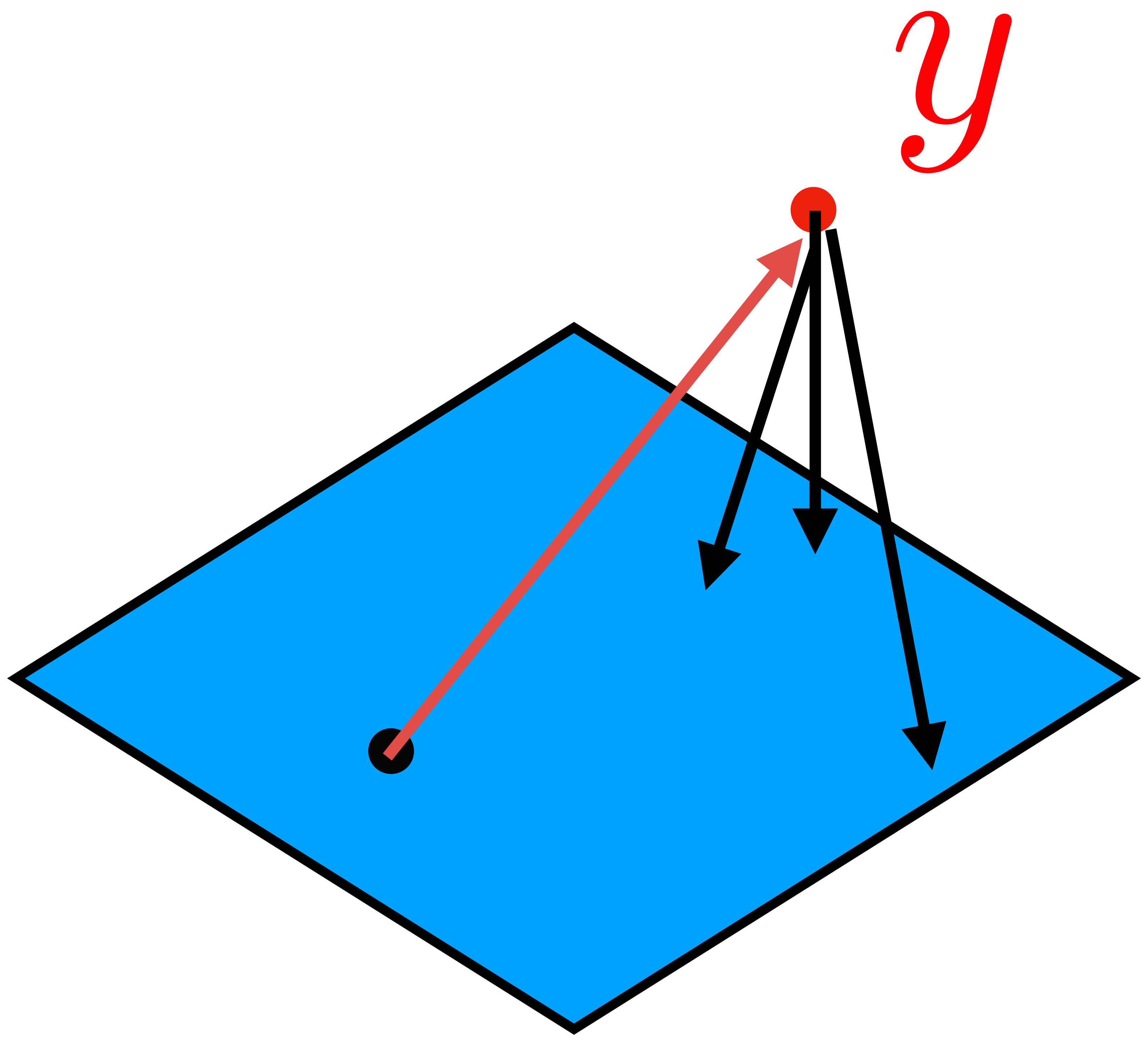
$$Xw \in \mathrm{col}(X) \subset \mathbb{R}^n$$



$$\hat{w} = \left(X^T X\right)^{-1} X^T y$$

Why are normal equations the solution?

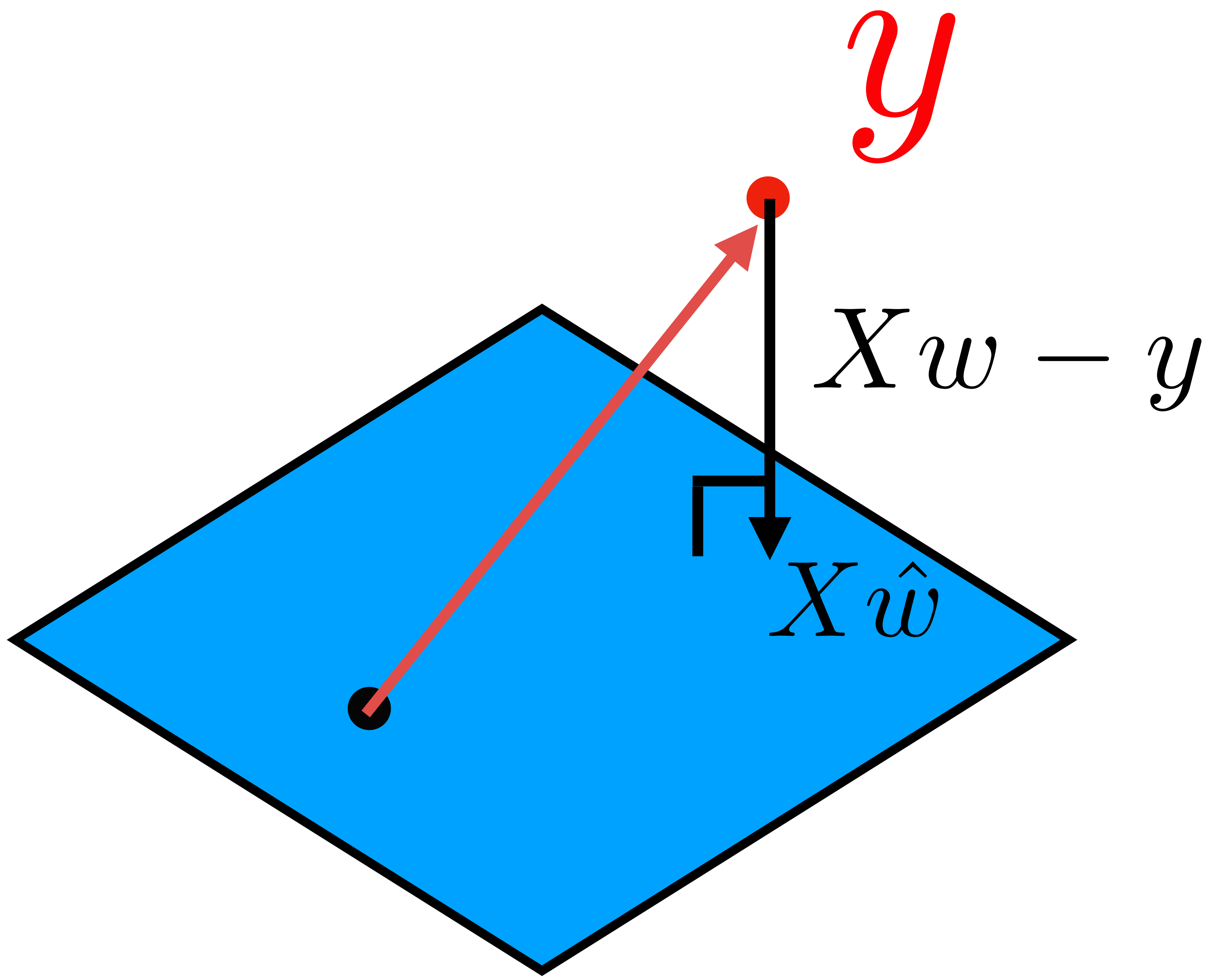
$$Xw - y$$



$$\hat{w} = \left(X^T X\right)^{-1} X^T y$$

Why are normal equations the solution?

$$\min_{w} ||Xw - y||_2^2$$

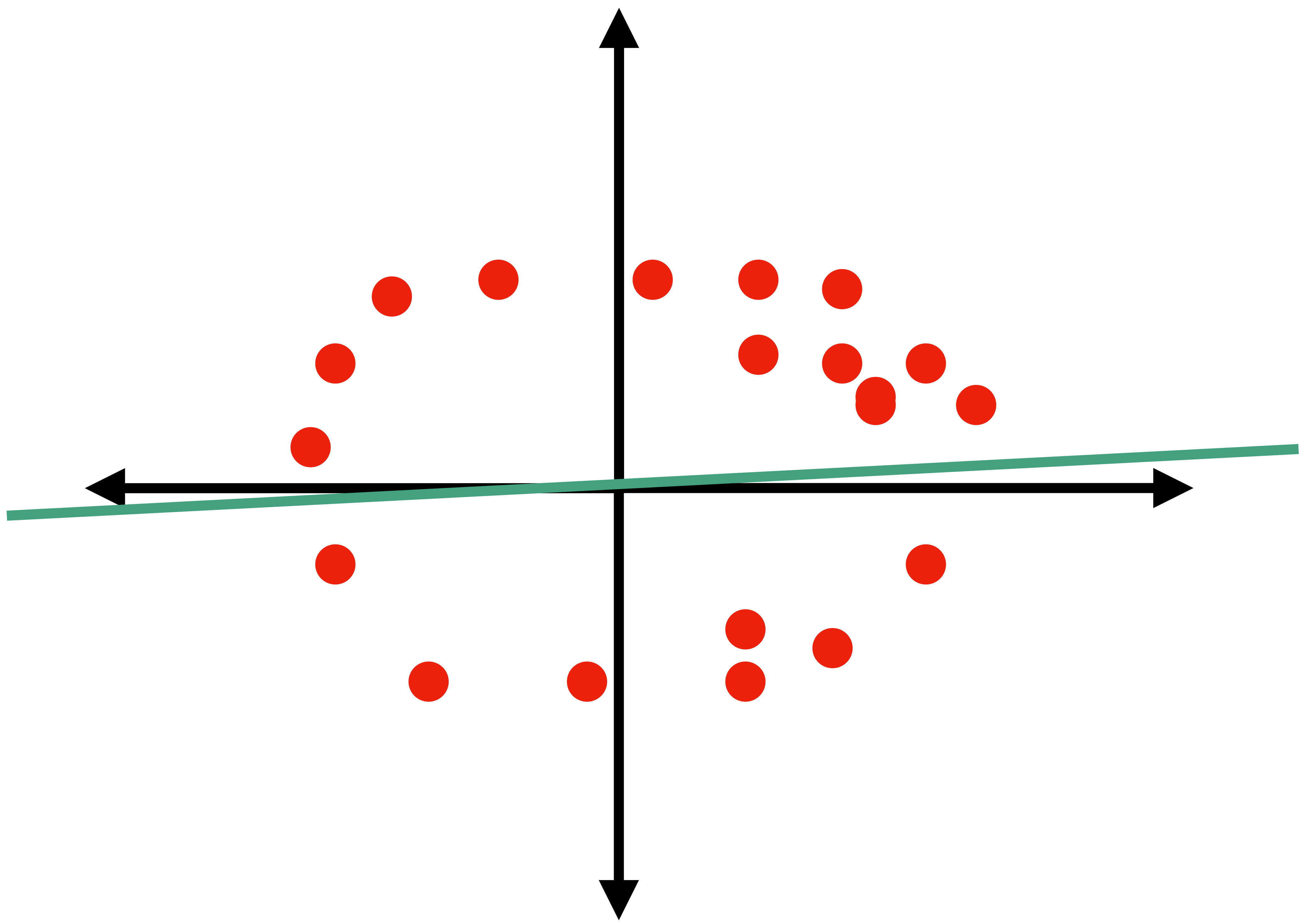


$$\hat{w} = \left(X^T X\right)^{-1} X^T y$$

Outline for today

- Reminders
- Cool ML application of the day
- Quick review of OLS
- **Intro to features**

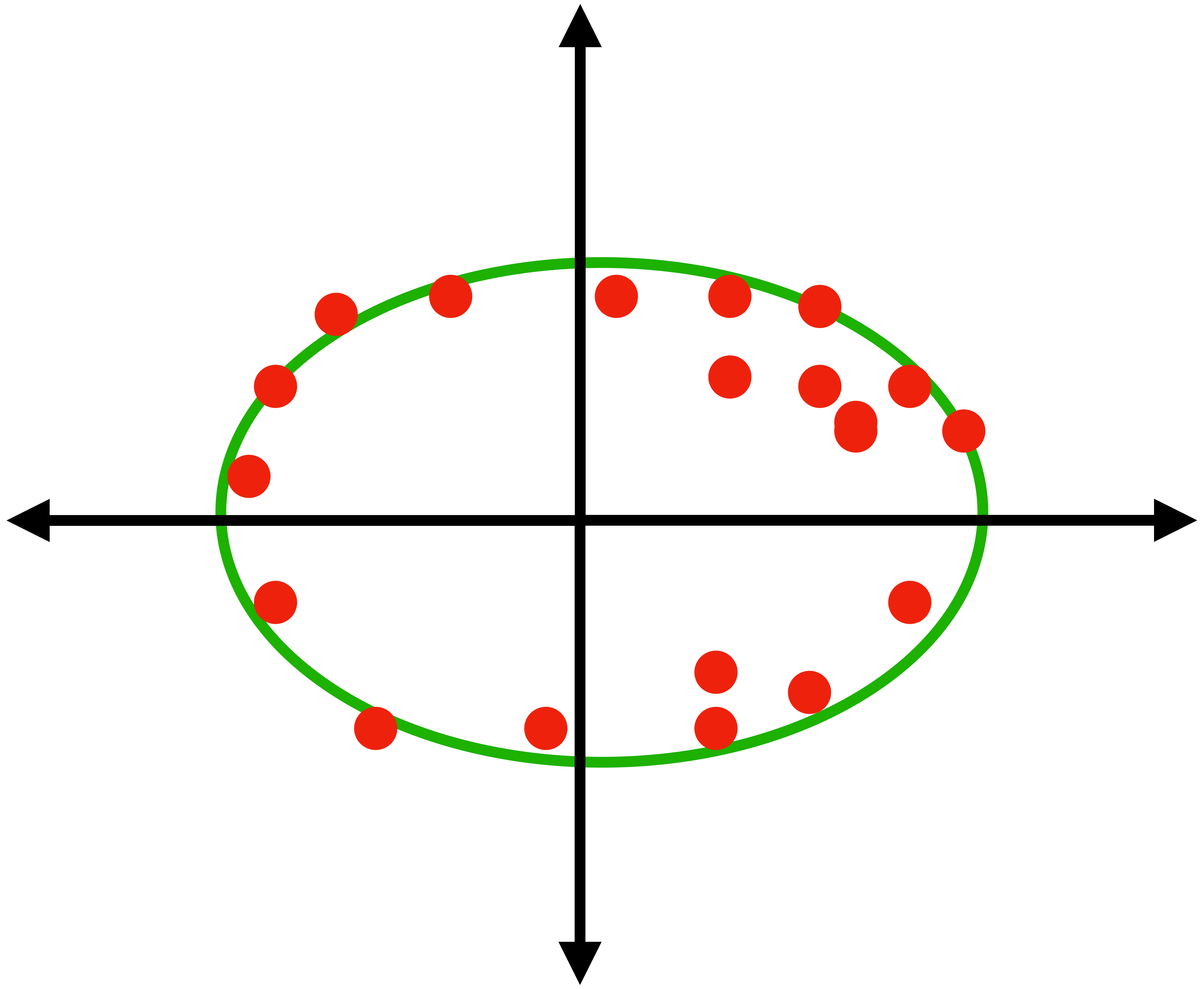
Can we model this linearly?



$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\vdots$$
$$(x_n, y_n)$$

Can we model this linearly?

$$w_1 x_i^2 + w_2 y_i^2 + w_3 x_i y_i + w_4 x + w_5 y + w_6 = 0$$

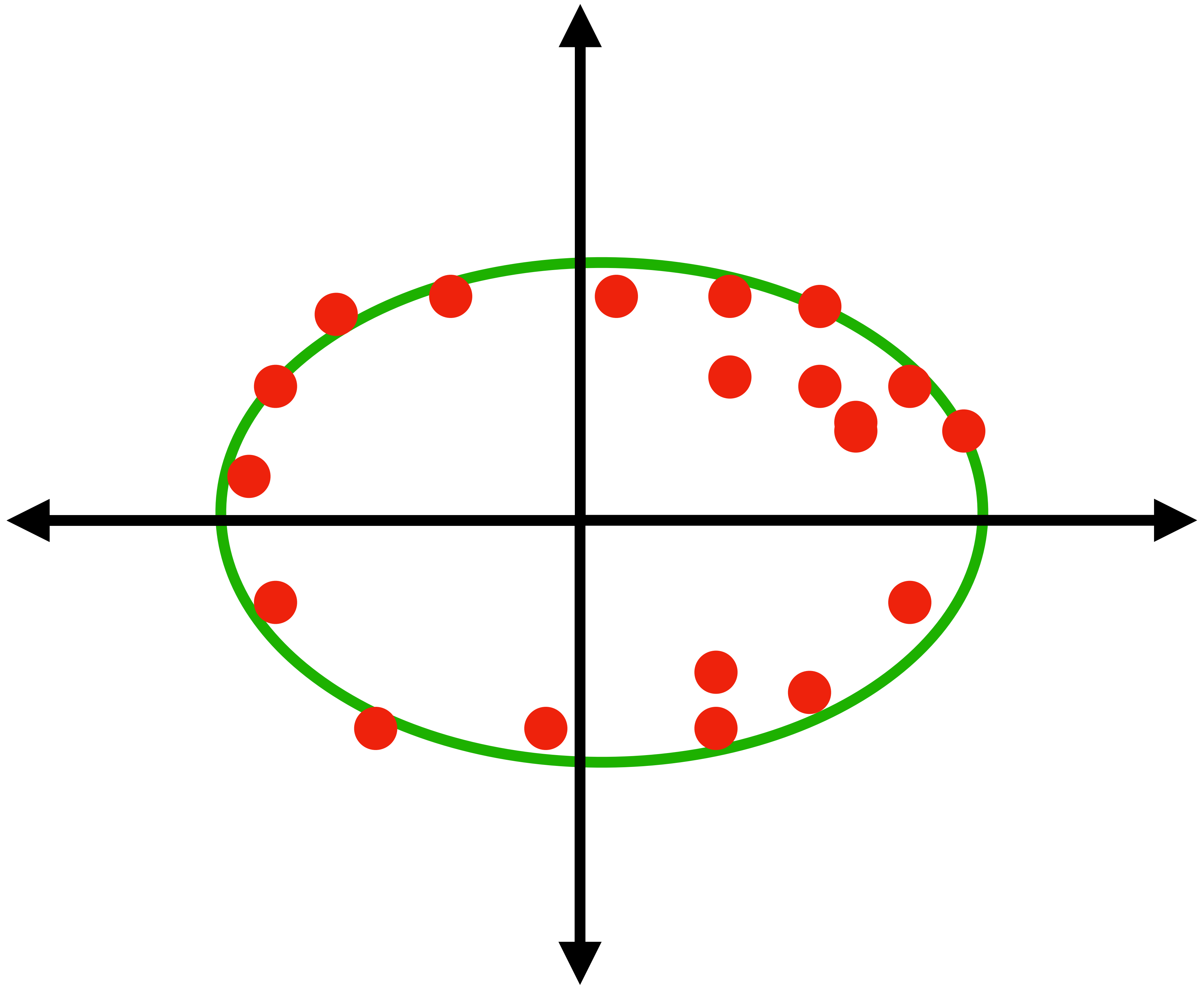


$$\{x^2, y^2, xy, x, y, 1\}$$

$$\bar{X} = \begin{bmatrix} x_1^2 & y_1^2 & x_1 y_1 & x_1 & y_1 & 1 \\ \vdots & & & & & \vdots \\ x_n^2 & y_n^2 & x_n y_n & x_n & y_n & 1 \end{bmatrix}$$

$$\min_w ||Xw - 0||^2$$

Can we model this linearly?



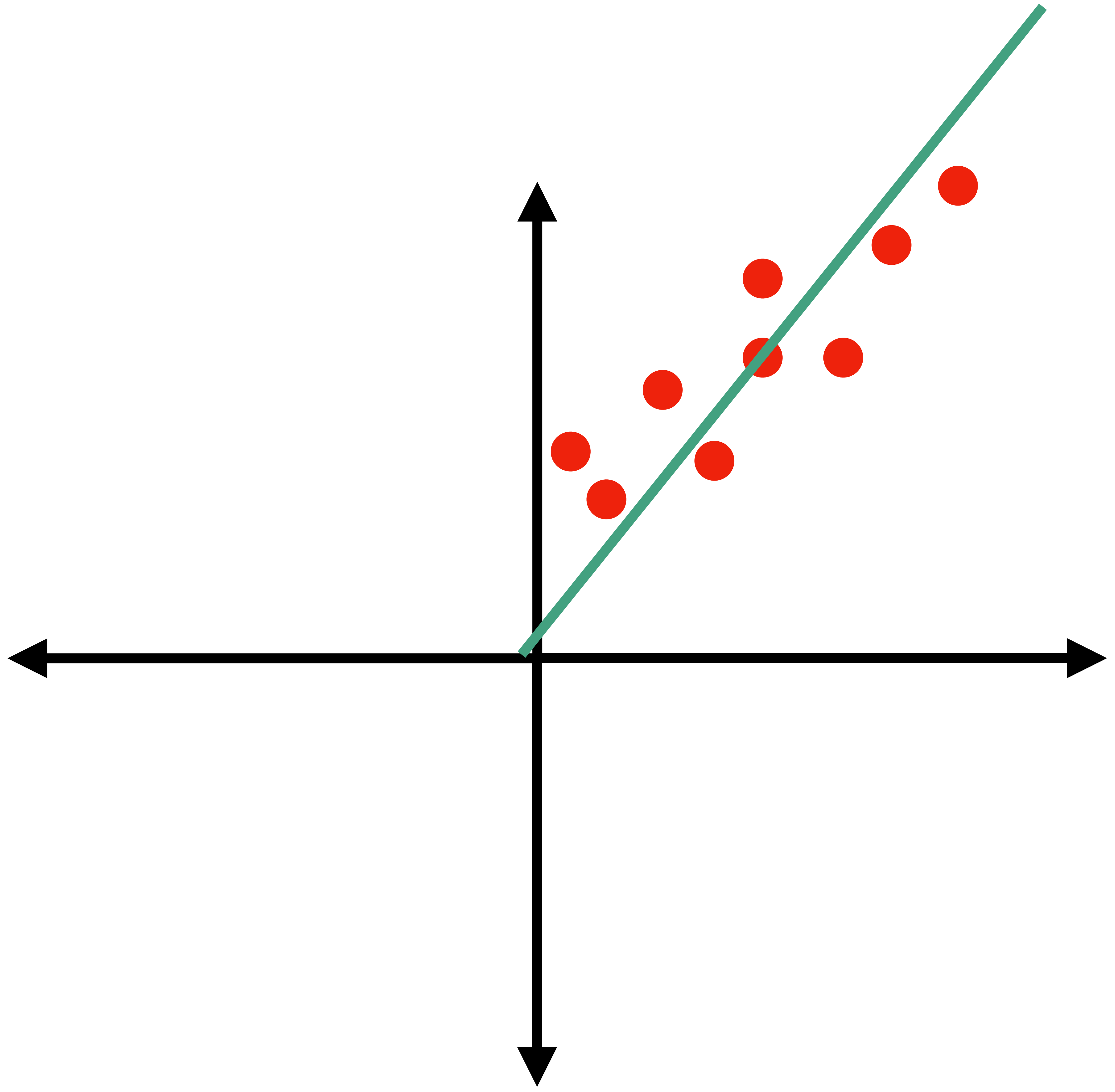
$$w_1 x_i^2 + w_2 y_i^2 + w_3 x_i y_i + w_4 x_i + w_5 y_i = 1$$

$$\{x_1^2, y_1^2, x_1 y_1, x_1, y_1\}$$

$$\bar{X} = \begin{bmatrix} x_1^2 & y_1^2 & x_1 y_1 & x_1 & y_1 \\ \vdots & & & & \vdots \\ x_n^2 & y_n^2 & x_n y_n & x_n & y_n \end{bmatrix}$$

$$\min_w ||\bar{X} w - \vec{1}||^2$$

Another example



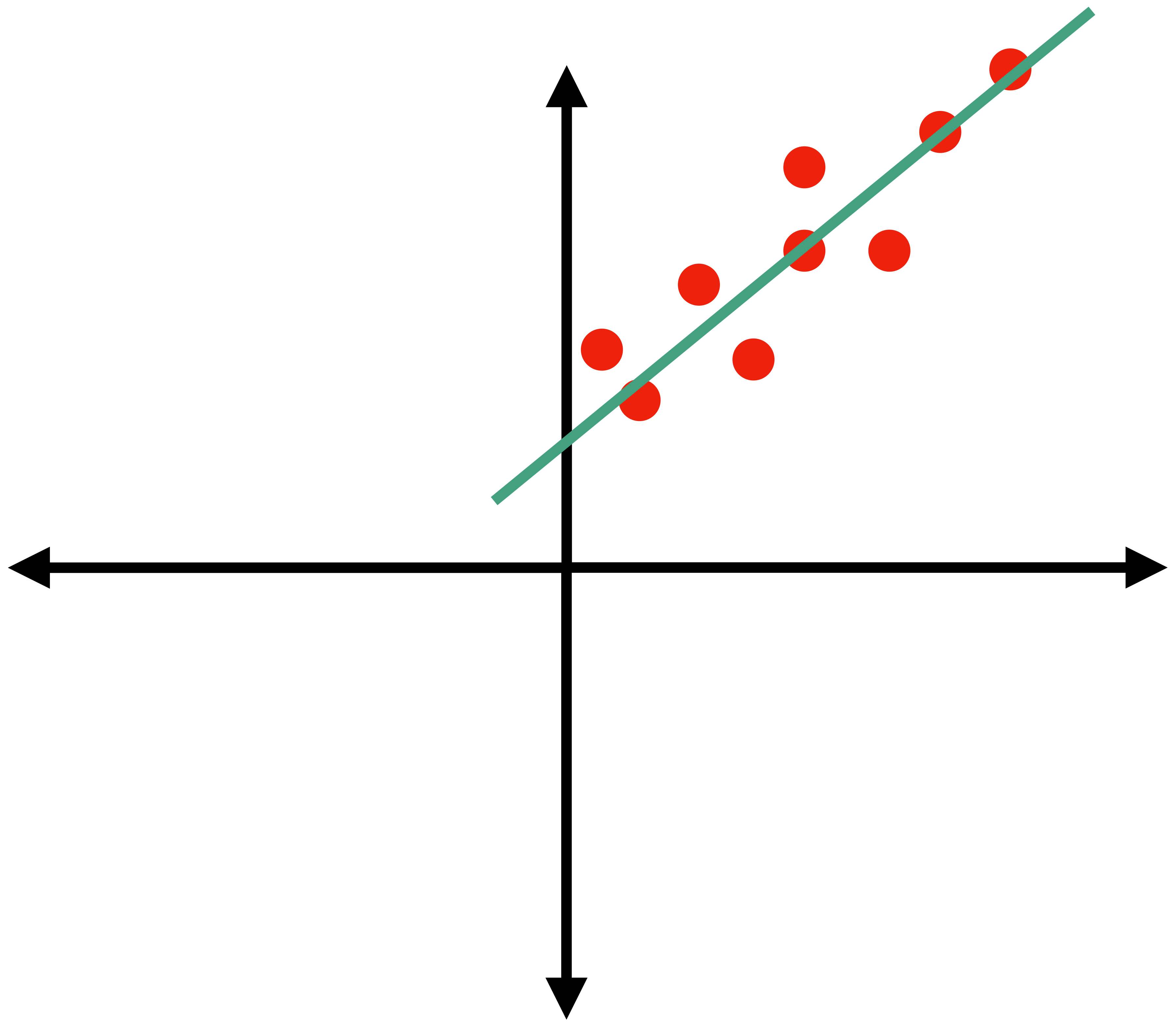
Model

$$\hat{y}_i = w_1 x_i$$

Features

$$\{x\}$$

Linear (affine) model



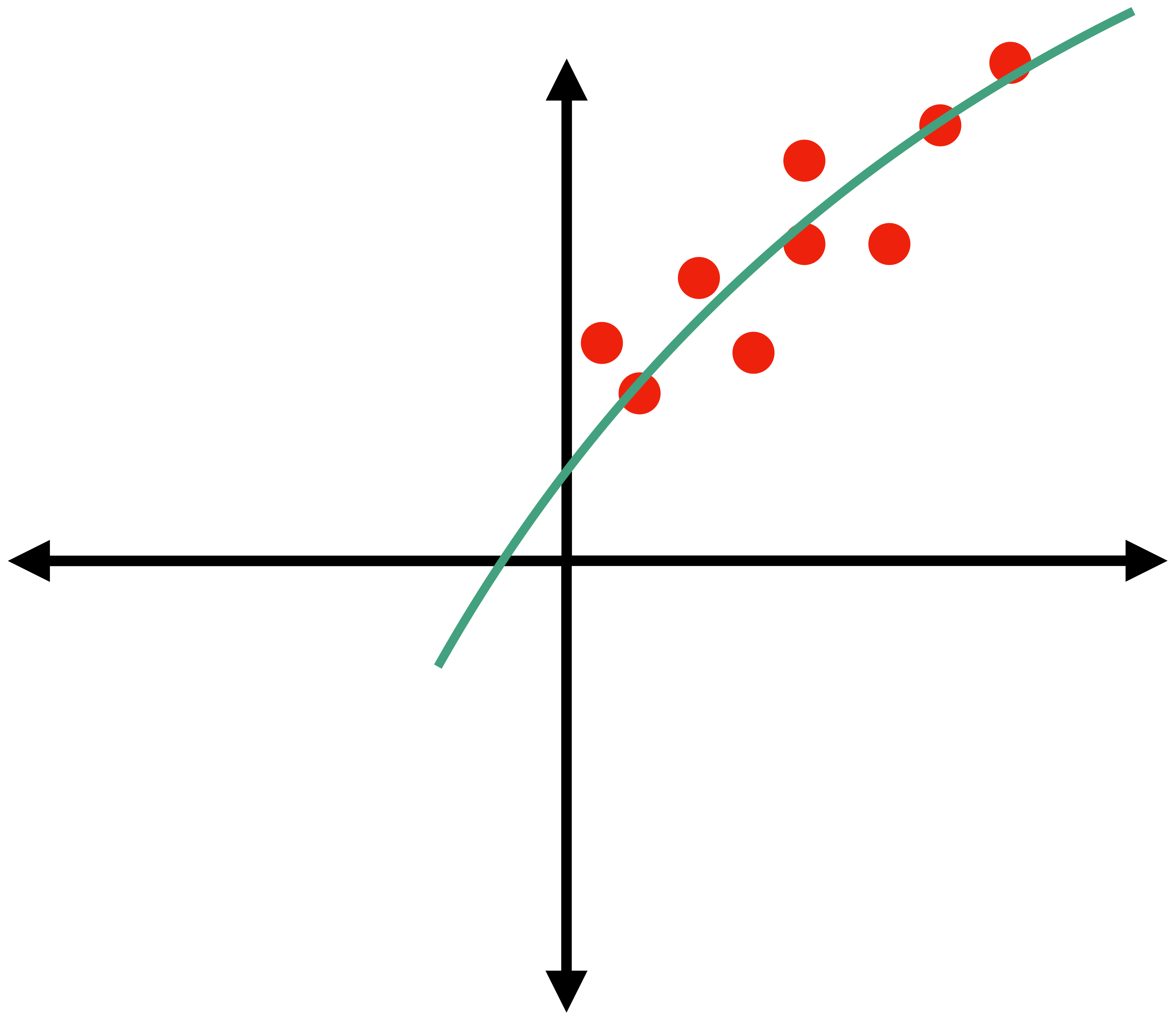
Model

$$\hat{y}_i = w_1 x_i + w_0$$

Features

$$\{1, x\}$$

Quadratic model



Model

$$\hat{y}_i = w_2 x_i^2 + w_1 x_i + w_0$$

Features

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

Polynomial features

Original data

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Model

$$\hat{y}_i = \sum_{j=0}^{p} w_j x_i^j$$

Features

$$\{1, x, x^2, \cdots x^p\}$$

Polynomial features

Original data

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

Model

$$\hat{y}_i = \sum_{j=0}^{p} w_{j0} x_{i1}^p + \cdots + w_{jd} x_{id}^p$$

$$+ w_{j,d+1} x_{i1}^{p-1} x_{i2} + \cdots$$

Features

$$\begin{array}{lllll} \{ & 0, & x_1, & \cdots, & x_d, \\ & x_1^2, & x_2^2, & \cdots, & x_d^2, \\ & x_1 x_2, & x_1 x_3, & \cdots, & x_1 x_d, \\ & & & \cdots, & \\ & x_1^p, & x_2^p, & \cdots, & x_d^p \quad \} \end{array}$$

Polynomial features

- Universal function approximations
- Number of params grows superlinearly in polynomial degree
- Later we will talk about how to avoid a lot of this computation

Polynomial features