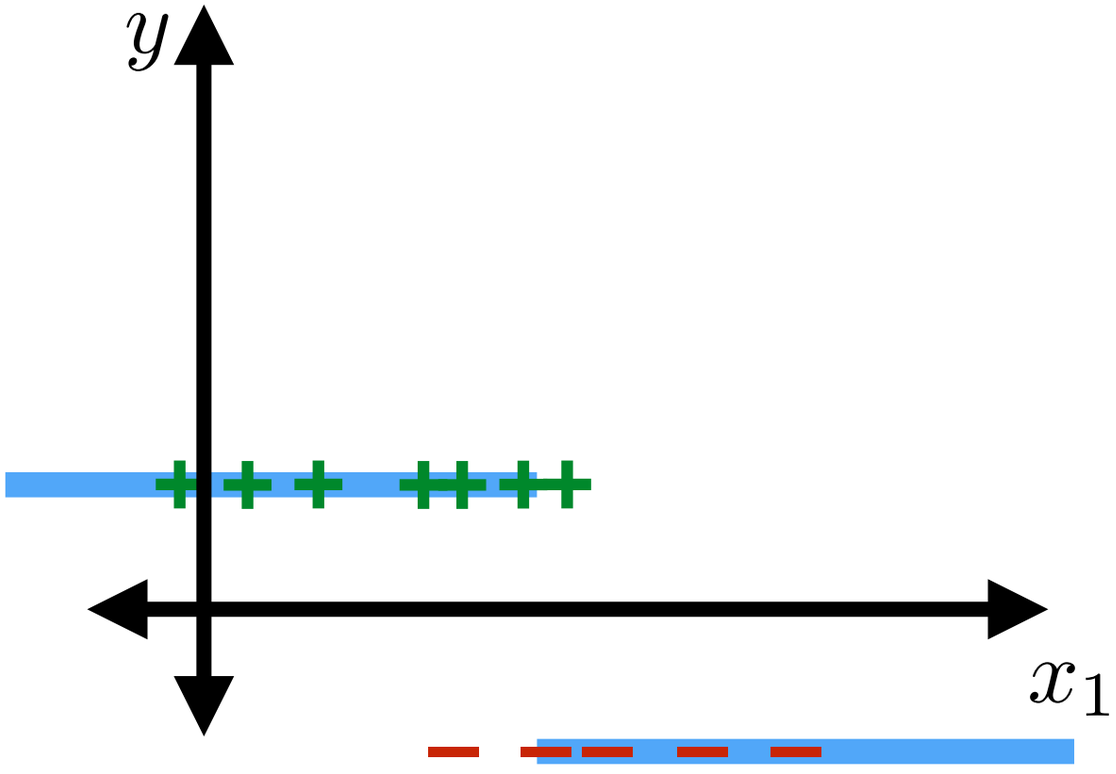
# Recall

## Classification

- Datum $i$: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \{-1, +1\}$
- Hypothesis $h : \mathbb{R}^d \to \{-1, +1\}$
- Loss: 0-1, asymmetric, NLL
- Example: linear classification



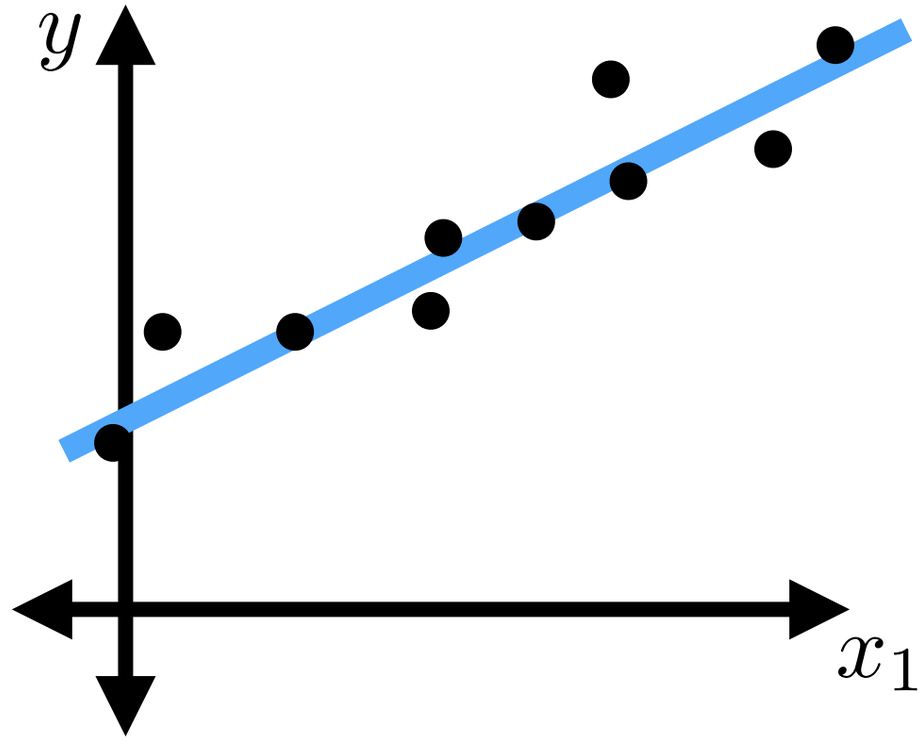

# Compare

## Regression

- Datum $i$ : feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$
- Loss: $L(g, a) = (g - a)^2$
- Example: linear regression



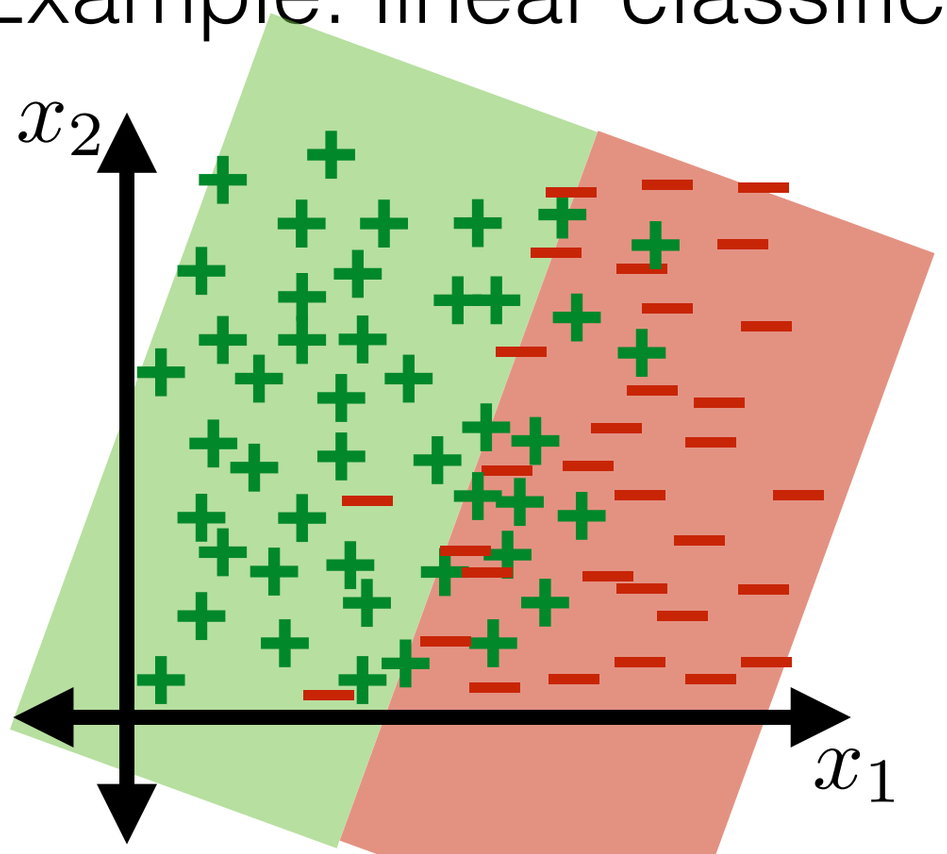
# Recall

## Classification

- Datum $i$: feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \{-1, +1\}$
- Hypothesis $h : \mathbb{R}^d \to \{-1, +1\}$
- Loss: 0-1, asymmetric, NLL
- Example: linear classification



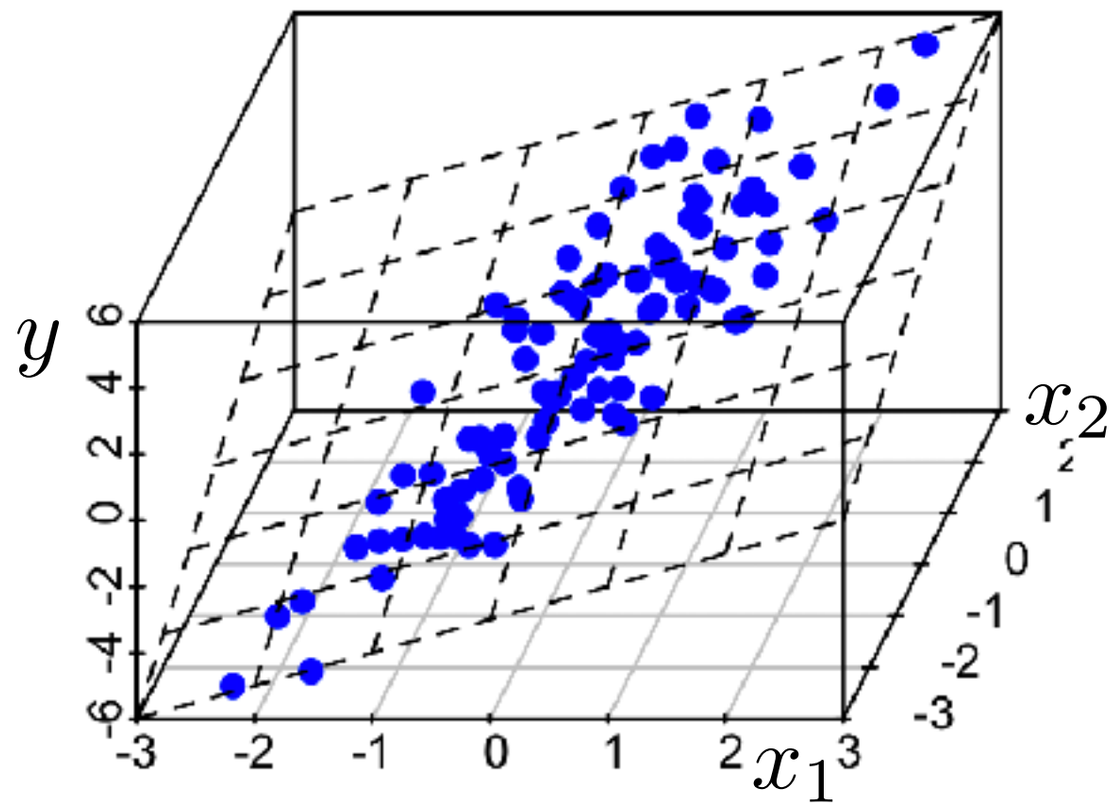

# Compare

## Regression

- Datum $i$ : feature vector

$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top \in \mathbb{R}^d$$

  - Label $y^{(i)} \in \mathbb{R}$
- Hypothesis $h : \mathbb{R}^d \to \mathbb{R}$
- Loss: $L(g, a) = (g - a)^2$
- Example: linear regression



# Linear regression

- Hypotheses for linear regression: $h(x;\theta,\theta_0) = \theta^\top x + \theta_0$
- Training error (using squared error loss)

$$J(\theta,\theta_0) = \frac{1}{n}\sum_{i=1}^{n} L(h(x^{(i)};\theta,\theta_0), y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2$$

- Training error if we augment features with feature "1"

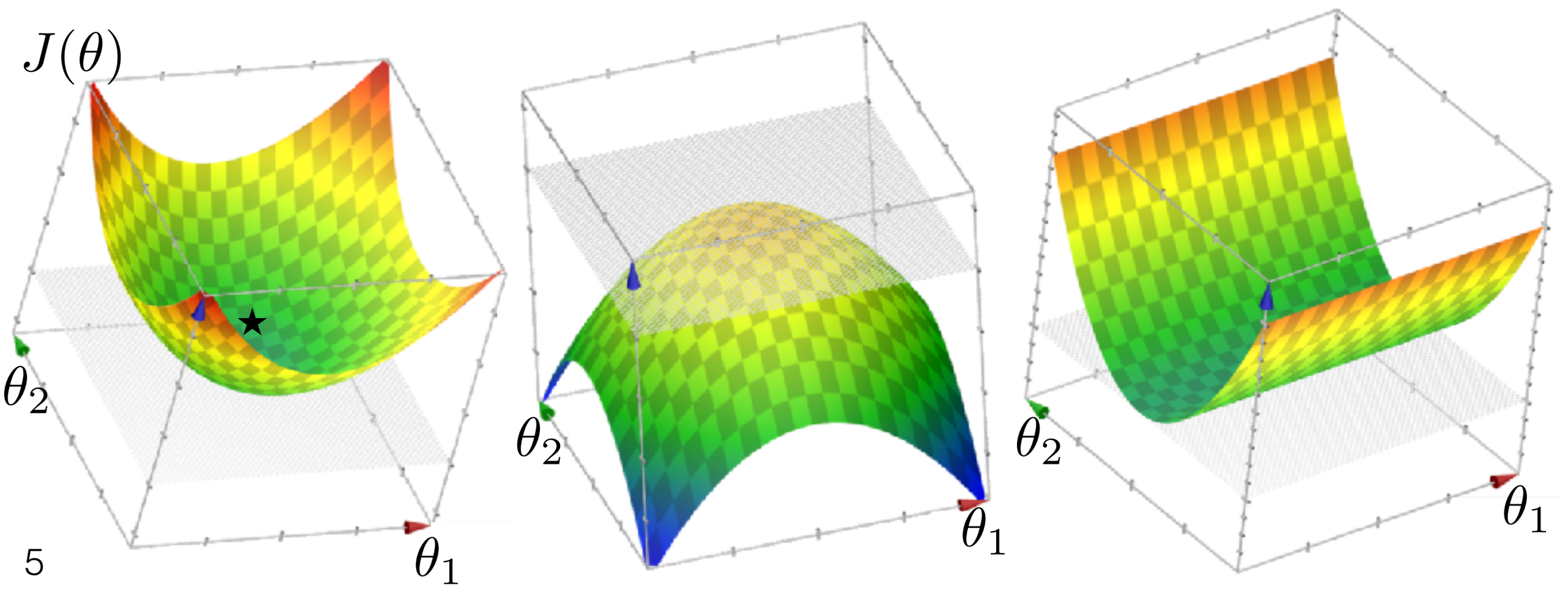
$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} - y^{(i)})^2 = \frac{1}{n}\sum_{i=1}^{n}((\underbrace{x^{(i)}}_{1\text{xd}})^\top \underbrace{\theta}_{\text{dx1}} - \underbrace{y^{(i)}}_{1\text{x1}})^2$$

$$= \frac{1}{n}\|\underbrace{\tilde{X}}_{\text{nxd}}\underbrace{\theta}_{\text{dx1}} - \underbrace{\tilde{Y}}_{\text{nx1}}\|^2 = \frac{1}{n}(\underbrace{\tilde{X}\theta - \tilde{Y}}_{\text{1xn}})^\top(\underbrace{\tilde{X}\theta - \tilde{Y}}_{\text{nx1}})$$

Define $\underset{\text{nxd}}{\tilde{X}} = \begin{bmatrix} x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \cdots & x_d^{(n)} \end{bmatrix}$ $\underset{\text{nx1}}{\tilde{Y}} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$

# Linear regression: A Direct Solution

- Goal: minimize $J(\theta) = \frac{1}{n}(\tilde{X}\theta - \tilde{Y})^\top(\tilde{X}\theta - \tilde{Y})$
- Uniquely minimized at a point if gradient at that point is zero and function "curves up" [see linear algebra]



# Linear regression: A Direct Solution

- Goal: minimize $J(\theta) = \frac{1}{n}(\tilde{X}\theta - \tilde{Y})^\top(\tilde{X}\theta - \tilde{Y})$
- Uniquely minimized at a point if gradient at that point is zero and function "curves up" [see linear algebra]
- Gradient $\nabla_\theta J(\theta) = \frac{2}{n}\tilde{X}^\top(\tilde{X}\theta - \tilde{Y}) \overset{\text{set}}{=} 0$

dx1 dxn nxd,dx1 nx1

Exercise: check the vector elements

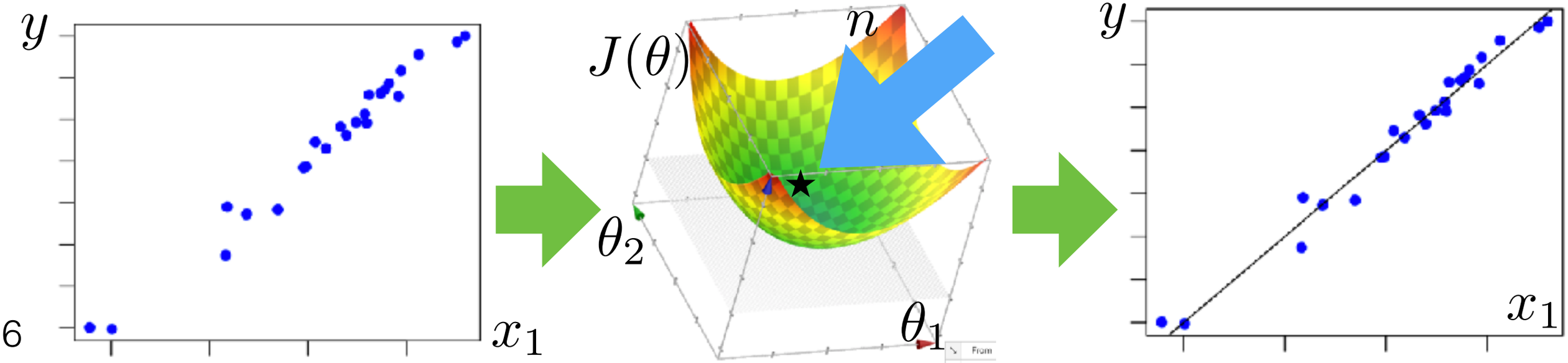
Exercise: check n,d=1

$$\tilde{X}^\top\tilde{X}\theta - \tilde{X}^\top\tilde{Y} = 0$$

$$(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{X}\theta = (\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{Y}$$

$$\theta = (\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{Y}$$

- Matrix of second derivatives $\frac{2}{n}\tilde{X}^\top\tilde{X}$



# Linear regression: A Direct Solution

- Goal: minimize $J(\theta) = \frac{1}{n}(\tilde{X}\theta - \tilde{Y})^\top(\tilde{X}\theta - \tilde{Y})$
- Uniquely minimized at a point if gradient at that point is zero and function "curves up" [see linear algebra]
- Gradient $\nabla_\theta J(\theta) = \frac{2}{n}\tilde{X}^\top(\tilde{X}\theta - \tilde{Y}) \overset{\text{set}}{=} 0$

dx1 dxn nxd,dx1 nx1

Exercise: check the vector elements

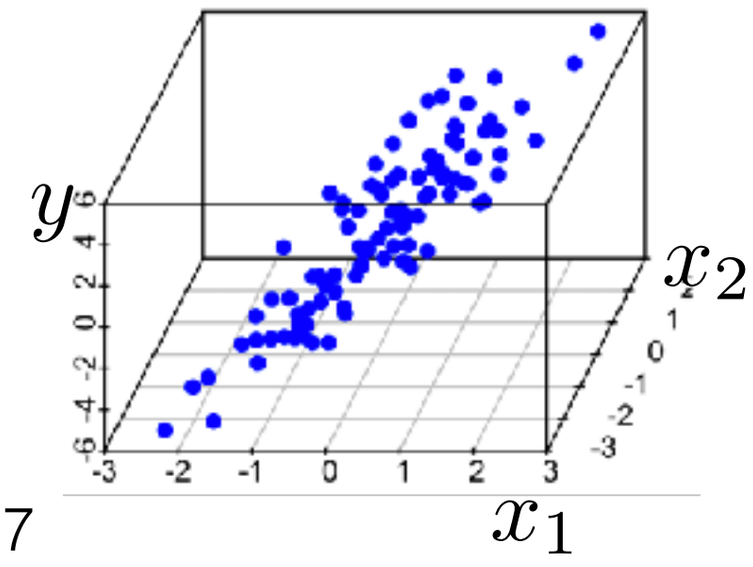
Exercise: check n,d=1

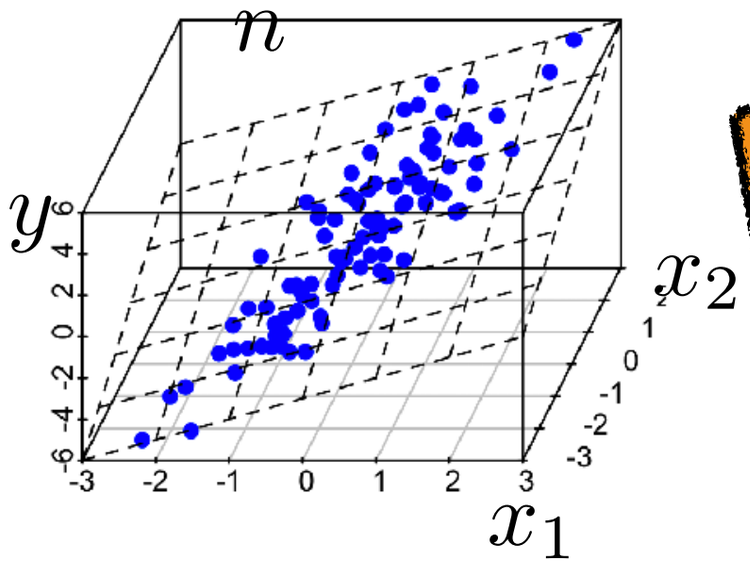
$$\tilde{X}^\top\tilde{X}\theta - \tilde{X}^\top\tilde{Y} = 0$$

$$(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{X}\theta = (\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{Y}$$

$$\theta = (\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{Y}$$

- Matrix of second derivatives $\frac{2}{n}\tilde{X}^\top\tilde{X}$

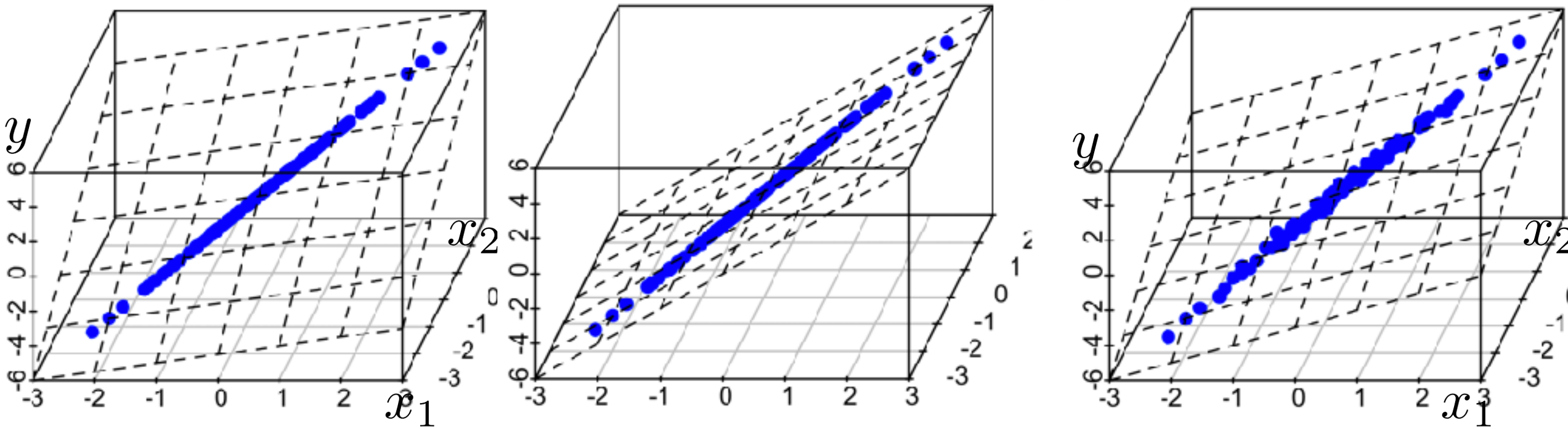


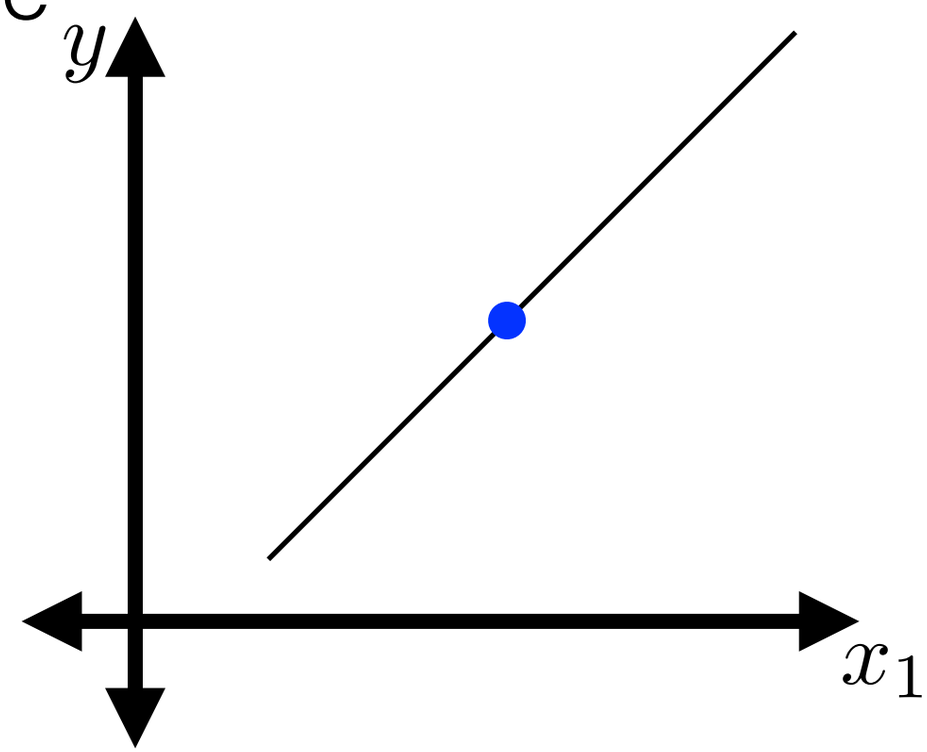

Note: hypothesis is a hyperplane!

# What can go wrong in practice?

- Sometimes there isn't a unique best hyperplane




- Sometimes there's technically a unique best hyperplane, but just because of noise
- Feature encodings (cf. one-hot); real-life features often correlated; many feature dimensions
- How to choose among planes? Preference for $\theta$ components being near zero



# Regularizing linear regression

- Linear regression with square penalty: ridge regression

$$J_{\text{ridge}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2 + \lambda\|\theta\|^2 \quad (\lambda > 0)$$

- Special case: ridge regression with no offset

$$J_{\text{ridge}}(\theta) = \frac{1}{n}(\tilde{X}\theta - \tilde{Y})^\top(\tilde{X}\theta - \tilde{Y}) + \lambda\|\theta\|^2$$

$$= \frac{1}{n}\|\tilde{X}\theta - \tilde{Y}\|^2 + \lambda\|\theta\|^2$$

What happens if $\lambda < 0$ ?

  - Min at: $\nabla_\theta J_{\text{ridge}}(\theta) = 0 \;\Rightarrow \theta = (\tilde{X}^\top\tilde{X} + n\lambda I)^{-1}\tilde{X}^\top\tilde{Y}$

dxn,nxd dxd

    - Matrix of second derivatives: $\tilde{X}^\top\tilde{X} + n\lambda I$ (always "curves up" & invertible when $\lambda > 0$)

- Can also solve for minimizing parameters in case with offset; just a bit more math

# Some notes on features

- Linear regression with square penalty: ridge regression

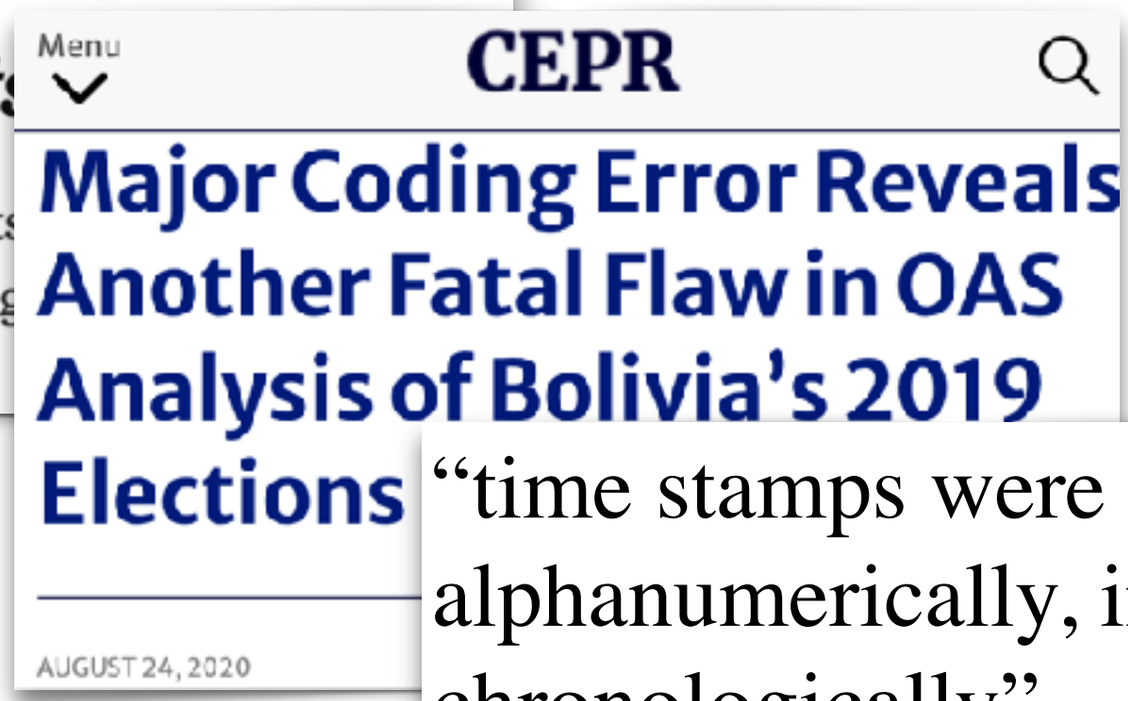
$$J_{\text{ridge}}(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} (\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2 + \lambda \|\theta\|^2 \quad (\lambda > 0)$$

- Assumption: features on same scale (cf. standardization)
- Featurization still matters! (ridge or “ordinary” regression)


The New York Times

***A Bitter Election. Accusations of Fraud. And Now Second Thoughts***

A close look at Bolivian election data suggests
by the O.A.S. that raised questions of vote-rig
force out a president — was flawed.


Menu

CEPR

**Major Coding Error Reveals Another Fatal Flaw in OAS Analysis of Bolivia’s 2019 Elections**

AUGUST 24, 2020

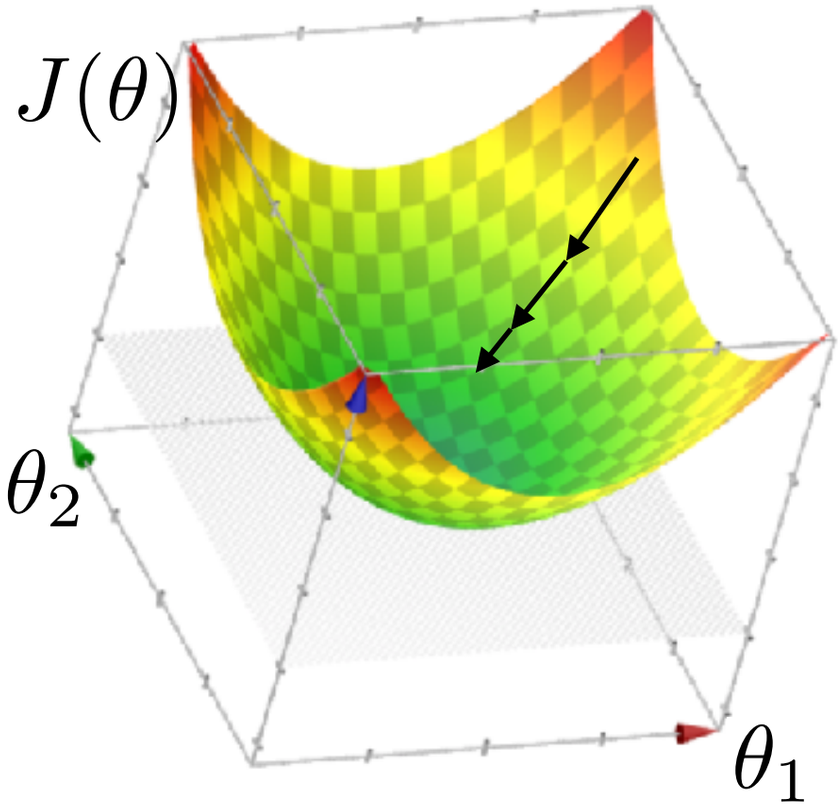
- Can never take a data set blindly
- Share code/data

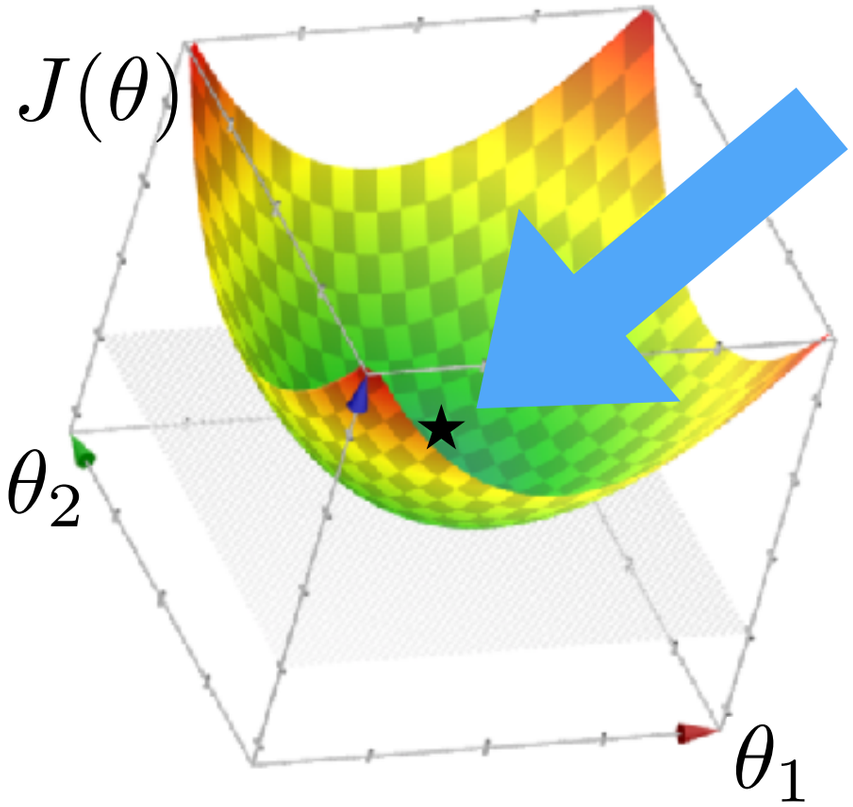
“time stamps were sorted alphanumerically, instead of chronologically”

[https://cepr.net/press-release/major-coding-error-reveals-another-fatal-flaw-in-oas-analysis-of-bolivias-2019-elections/]

# Optimizing linear regression

- Gradient descent vs. analytical/closed-form/direct solution




- Accuracy doesn't mean anything without running time
- Running time doesn't mean anything without accuracy
- Need to measure accuracy for the running time we have

$$\theta = \underbrace{(\tilde{X}^\top \tilde{X} + n\lambda I)}^{-1} \tilde{X}^\top \tilde{Y}$$

dxd

Matrix inversion: $O(d^3)$

# Gradient descent for linear regression

LinearRegression-Gradient-Descent( $\theta_{\text{init}}, \theta_{0,\text{init}}, \eta, T$ )

Initialize $\theta^{(0)} = \theta_{\text{init}}$

Initialize $\theta_0^{(0)} = \theta_{0,\text{init}}$

**for** t = 1 **to** T

Exactly gradient descent with *f* given by linear regression objective

$$\theta^{(t)} = \theta^{(t-1)} - \eta\left\{ \frac{2}{n} \sum_{i=1}^{n} \left[\theta^{(t-1)\top} x^{(i)} + \theta_0^{(t-1)} - y^{(i)}\right] x^{(i)} + 2\lambda\theta^{(t-1)} \right\}$$

$$\theta_0^{(t)} = \theta_0^{(t-1)} - \eta\left\{ \frac{2}{n} \sum_{i=1}^{n} \left[\theta^{(t-1)\top} x^{(i)} + \theta_0^{(t-1)} - y^{(i)}\right] \right\}$$

**Return** $\theta^{(t)}, \theta_0^{(t)}$

# Stochastic gradient descent

- Linear regression objective with $\lambda = 0$ :

$$J_{\text{linreg}}(\Theta) = J_{\text{linreg}}(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n}(\theta^\top x^{(i)} + \theta_0 - y^{(i)})^2$$

- A common machine learning objective:

$$f(\Theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\Theta)$$

- Also recall logistic regression with $\lambda = 0$

Stochastic-Gradient-Descent($\Theta_{\text{init}}, \eta, T$)

  Initialize $\Theta^{(0)} = \Theta_{\text{init}}$

  **for** t = 1 **to** T

    randomly select i from {1,…,n} (with equal probability)

    $\Theta^{(t)} = \Theta^{(t-1)} - \eta(t)\nabla_\Theta f_i(\Theta^{(t-1)})$

  **Return** $\Theta^{(t)}$