Plan

Recap

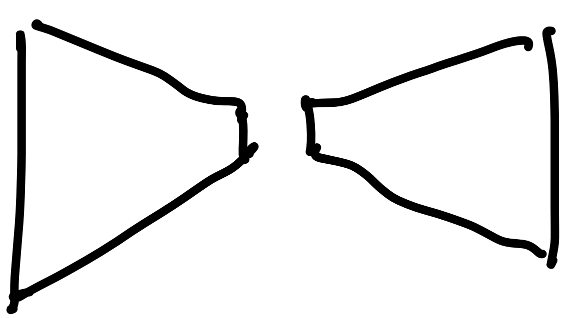
Logistics

Regularization

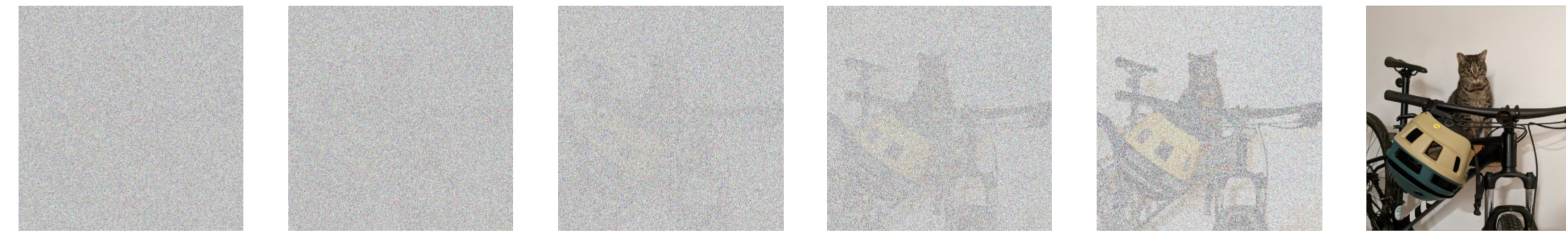
Implicit Regularization

Parameterization

CRF



Diffusion



x_t $\quad$ x_{t-1}

$x_t \in \mathbb{R}^{m \times m}$

$\epsilon_t \in \mathbb{R}^{m \times m}$

$$x_t = x_{t-1} + \epsilon_t$$

$$\epsilon_t \approx f_\theta(x_t, t)$$

$$x_{t-1} = x_t - \epsilon_t$$

Stable Diffusion

↳ do diffusion in latent space

↳ text conditioning

Logistics

Check Canvas Assignment

- Last day of new content
  - last form tonight
  - CRF
- Homework tomorrow
- Tomorrow
  - no lecture/demo
  - I'll be around
  - games at 4
- Presentation

Supervised

(x, y), n samples

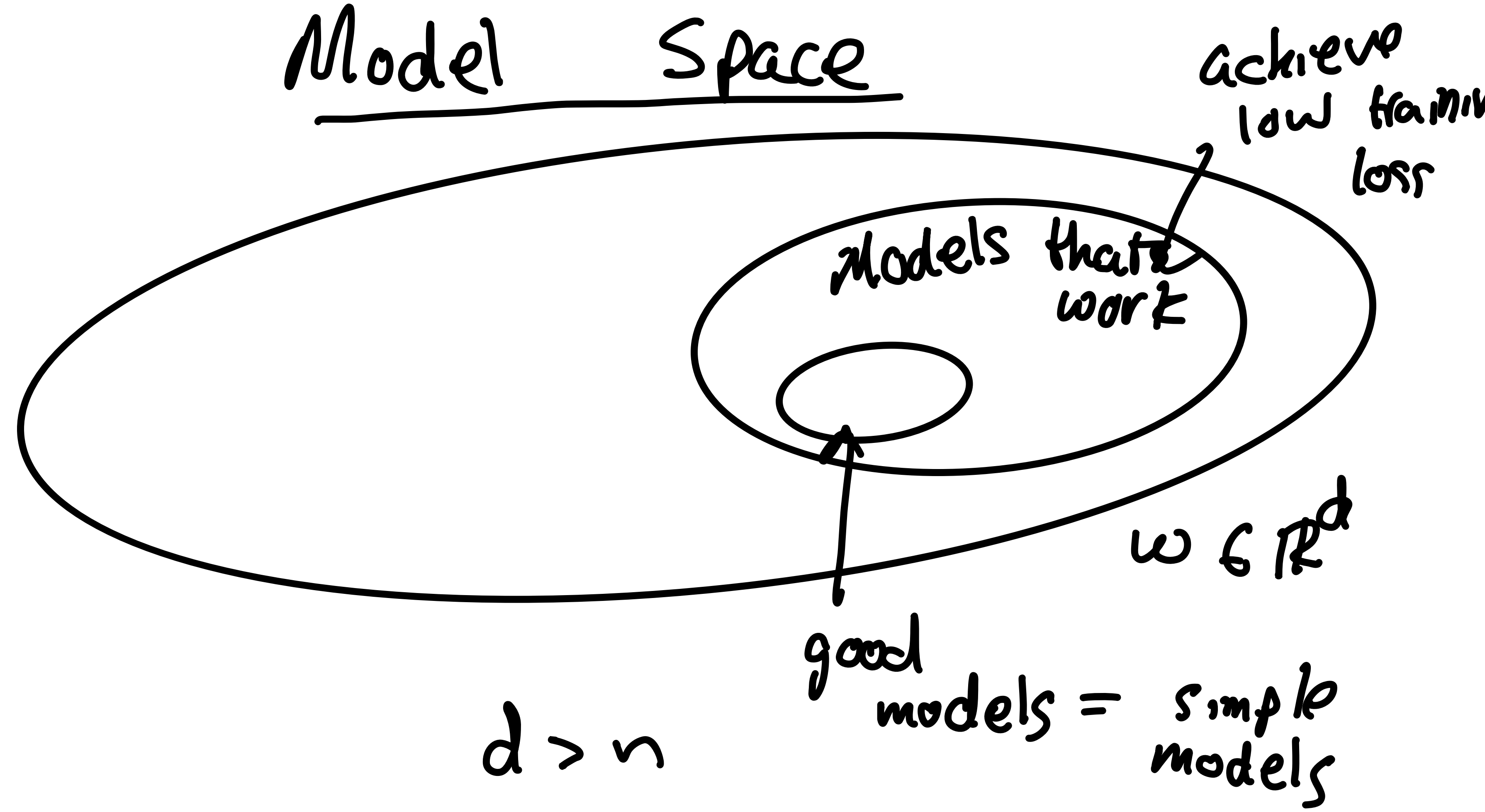
$f_w(x) \approx y$

$\mathcal{L}(w) = \|f_w(x) - y\|_2^2$

training (x,y) test (x,y)

Conceivably f_w could memorize the training data

Model Space



$d > n$

Frumpf	Clon
Apple	Pear
Moose	Kiwi
Computer	Tea
Board	Pen
Phone	Ball
	Horse

Strategies

- last letter
- first letter
- length
- vowels

What is a simple model?

↳ few parameters

What is a simple model in the space of fixed number of parameters?

↳ $||w||_2^2 \sim$ complexity

Large $||w||_2^2$ → big changes when we perturb

Regularization

$$\mathcal{L}(w) + \lambda ||w||_2^2$$

"lambda" tradeoff

how well we solve task

complexity of model

we do this implicitly

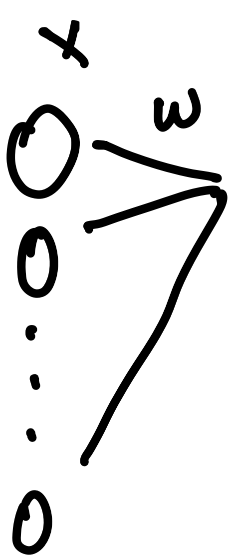
Implicit Regularization

Linear Regression

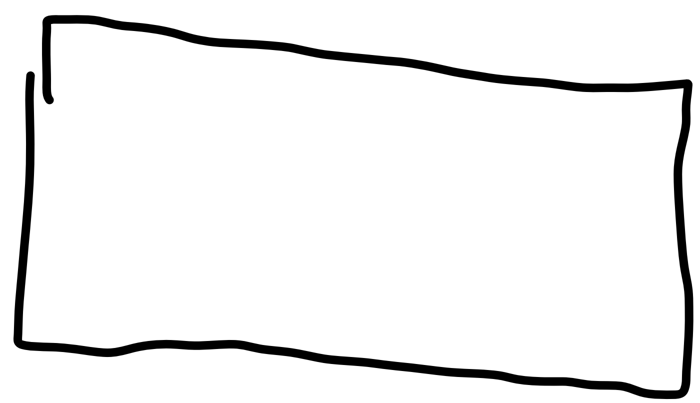
(y, x)

$y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times d}$

$y \approx f_\theta(X) = Xw \quad w \in \mathbb{R}^d$



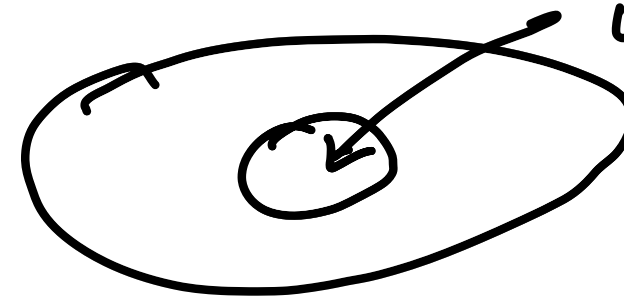
$x^T w$



X

$d > n \Rightarrow$ lots of optimal w

$w: \; y = Xw$



w

$$\mathcal{L}(w) = \frac{1}{2}\|y - Xw\|_2^2$$

$$-\frac{\partial \mathcal{L}}{\partial w_i} = x_i^T(y - Xw)$$

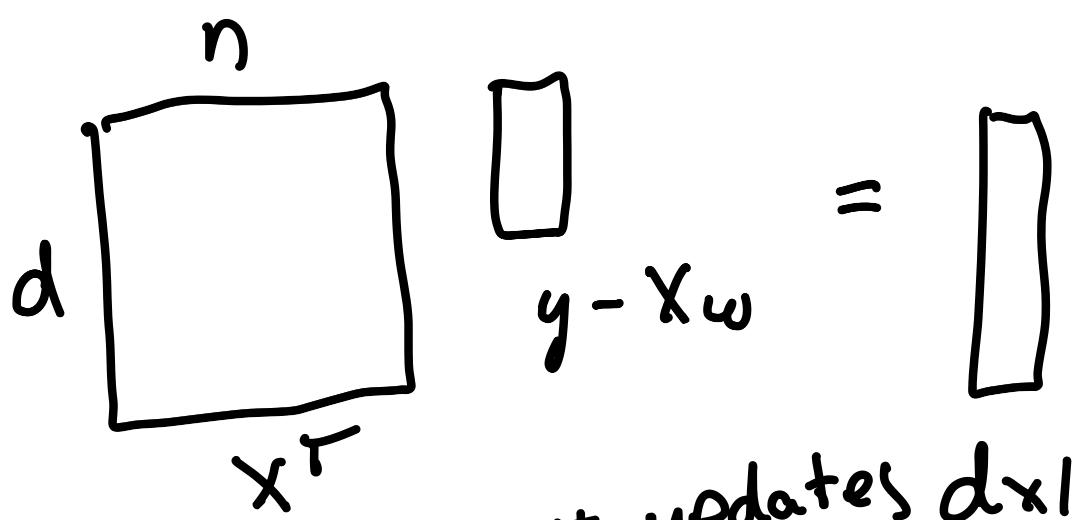
$$-\nabla_w \mathcal{L} = X^T(y - Xw)$$

$d \times n \quad n \times 1 \quad n \times d \; d \times 1$

direction of update

$$-\nabla_w \mathcal{L} = X^T(y - Xw) = \sum_{i=1}^{n} x_i \cdot \text{something}$$

span of $\{x_i\}$



T ← # updates

$$\underbrace{w_0}_{0} - \underbrace{\sum_{t=1}^{T} \nabla_w \ell(w_t)}_{\text{span of } \{x_i\}} = \underbrace{\hat{w}}_{\text{span of } \{x_i\}}$$

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

$y = X\hat{w}$ ← b/c GD gives an optimal solution

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

$XX^T\alpha = y$ assuming XX^T is full rank $(d > n)$

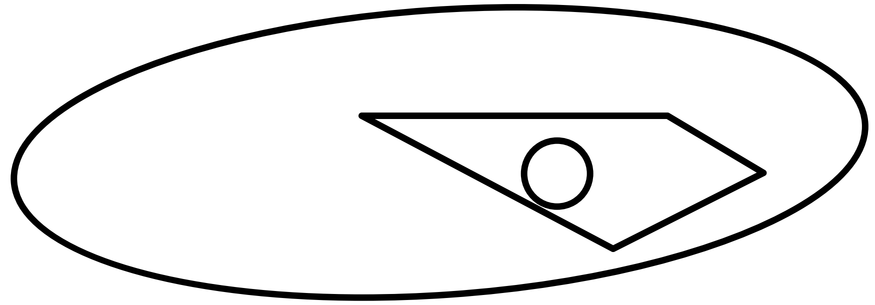
$n \times d \; d \times n$

$$(XX^T)^{-1} X X^T \alpha = (XX^T)^{-1} y$$

$$\alpha = (XX^T)^{-1} y$$

$$\hat{w} = X^T\alpha = X^T(XX^T)^{-1} y$$

$$\min \|w\|_2^2$$
$$w : Xw = y$$



$$\Updownarrow$$

$$\min_{w \in \mathbb{R}^d} \max_{\beta \in \mathbb{R}^n} \overbrace{\|w\|_2^2 + \beta^T(y - Xw)}^{\mathcal{L}}$$

Solution is saddle point

Optimal w, b...

$$\nabla_w \mathcal{L} = 0 = 2w - X^T\beta$$

d×n n×1

$$\Rightarrow w = X^T\beta \cdot \frac{1}{2}$$

$$\nabla_\beta \mathcal{L} = 0 = y - Xw$$

$$0 = y - XX^T\beta \cdot \frac{1}{2}$$

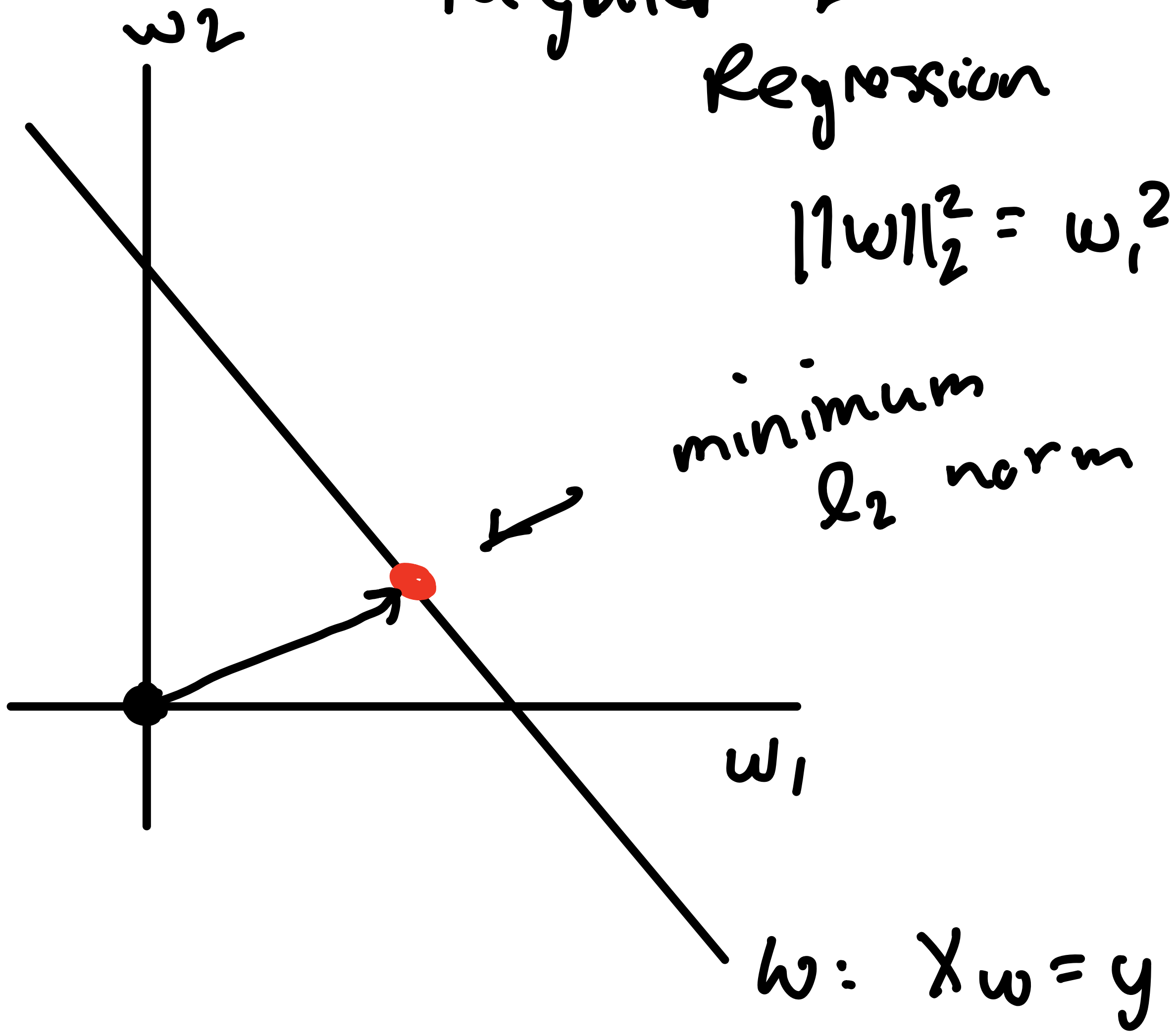
$$2y = XX^T\beta$$

$$\beta = 2 \cdot (XX^T)^{-1} y$$

$$w = X^T\beta \cdot 1/2 = X^T (XX^T)^{-1} y$$

Regular Linear Regression

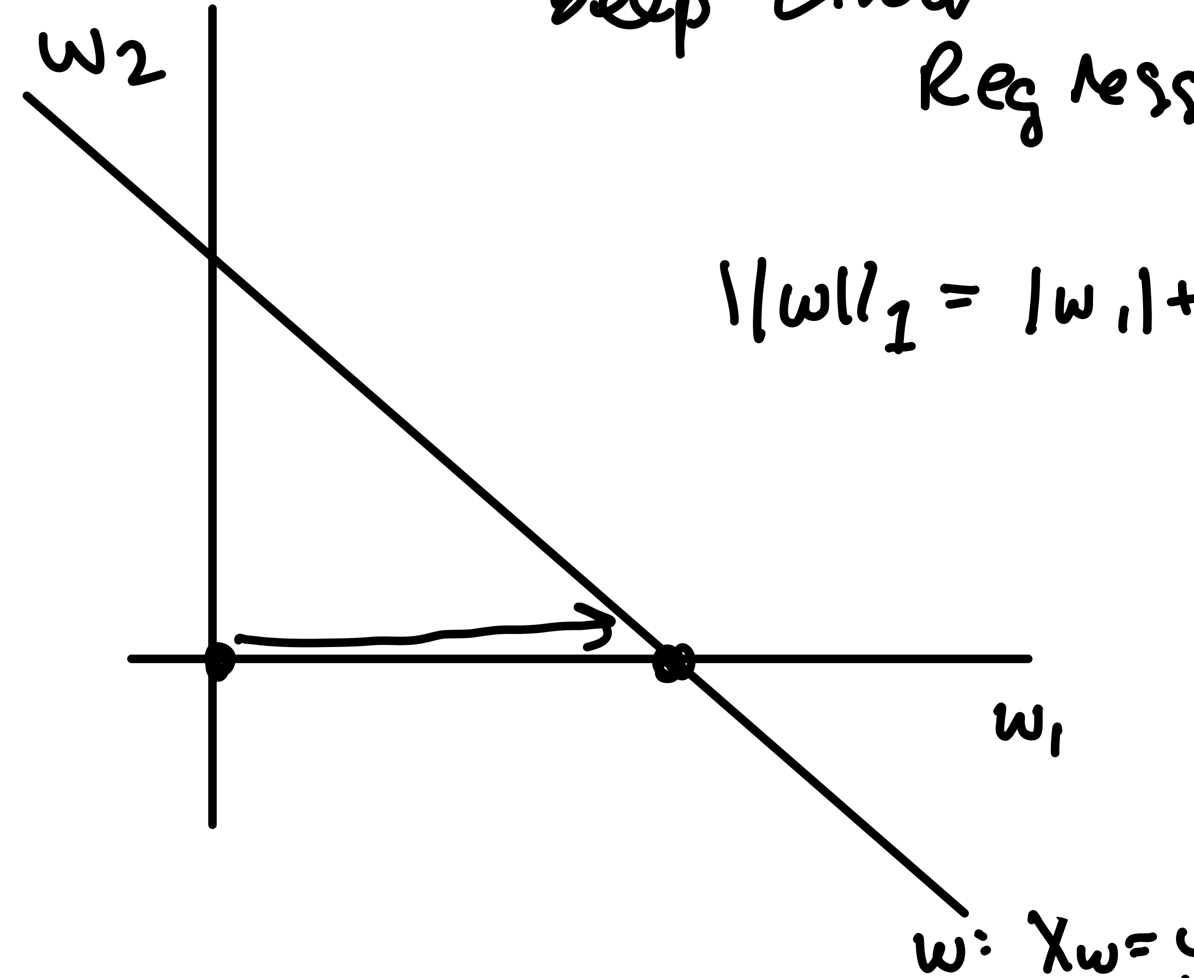
$\|w\|_2^2 = w_1^2 + w_2^2$



GD from $w_0 = 0$ bring us to optimal with minimum ℓ_2 norm

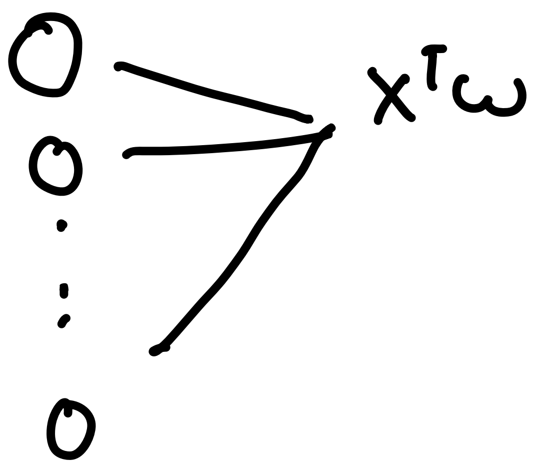
Deep Linear Regression

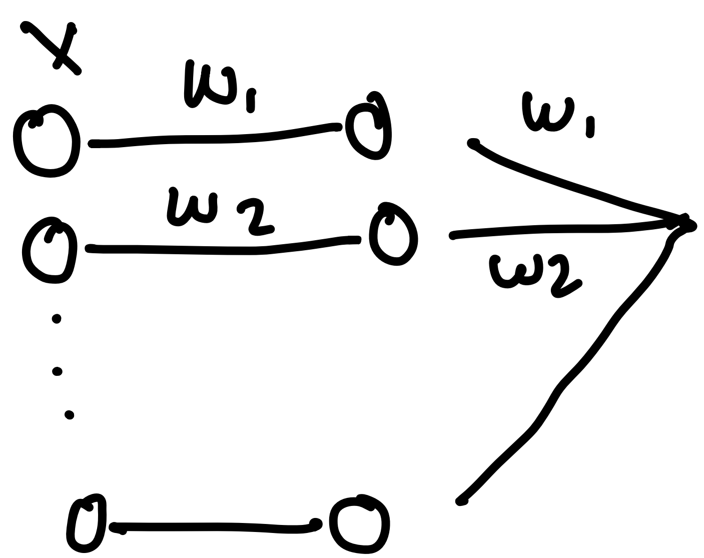
$\|w\|_1 = |w_1| + |w_2|$



GD from $w_0 = \mathbb{1} \cdot$ small constant bring us to optimal with minimum ℓ_1 norm

Deeper Linear Regression





$$\sum_{i=1}^{d} x_i \cdot w_i^2 = x^T (w \circ w)$$

entrywise multiplication

$$\mathcal{L}(w) = \frac{1}{2} \| y - X(w \circ w) \|_2^2$$

$$\frac{\partial \mathcal{L}}{\partial w_i} = x_i^T (y - X(w \circ w)) \cdot \frac{\partial w_i^2}{\partial w_i}$$

$2w_i$

$$\nabla_w \mathcal{L} = X^T (y - X(w \circ w)) 2w$$