Important Announcements and Notes (Oct 10)

- Midterm marks probably released this weekend

- Fill out mid-term evalution survey please

The Nobel Prize in Chemistry 2024

David Baker

"for computational protein design"

David Baker. Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

"for protein structure prediction"

Demis Hassabis. Ill. Niklas Elmehed © Nobel Prize Outreach

John Jumper

"for protein structure prediction"

John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach

The Nobel Prize in Physics 2024

John Hopfield

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

John Hopfield. Ill. Niklas Elmehed © Nobel Prize Outreach

Geoffrey Hinton

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

Geoffrey Hinton. Ill. Niklas Elmehed © Nobel Prize Outreach

Important Announcements and Notes (Oct 1)

Midterm details:

- 25 multiple choice (select all that apply) questions using scantron (Bring a pencil!!)
- Covers up to (and including) today lecture up to (and including) sec 6.2 in the course notes
- Formula sheet will be provided (no cheat sheet) will post tonight
- You can bring a calculator
- Similar to assignment questions and examples in course notes
- Will make an announcement tonight with all these same details

- Will use some of Oct 3 Lecture to do a midterm review

Optimization

finding the best solution from a set of possible solutions

Usually this means finding the minimum or maximum value of some function

we will care about:

$$\min_{w \in \mathcal{W}} g(w)$$

minimum value of $g(w)$ over all $w \in \mathcal{W}$

or $$w^* = \operatorname*{argmin}_{w \in \mathcal{W}} g(w)$$

the $w \in \mathcal{W}$ that achieves the minimum value of $g(w)$

w^* is a "minimizer"

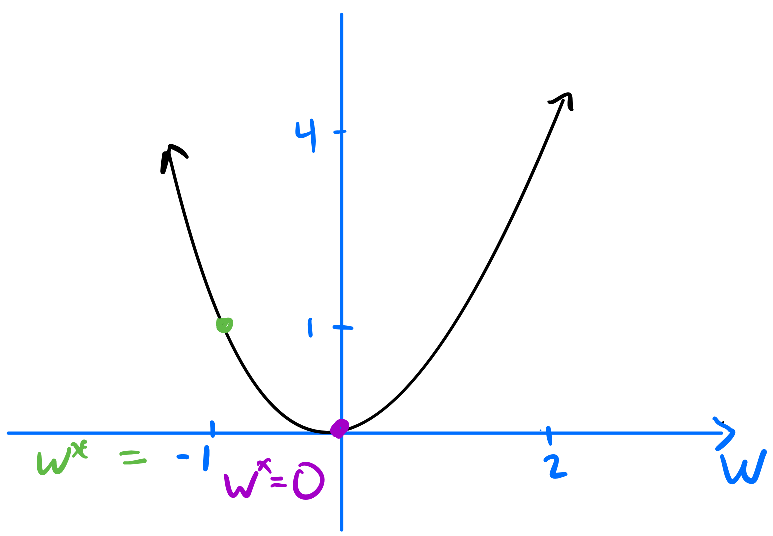
$$\min_{w \in \mathcal{W}} g(w) = g(w^*)$$

Ex: $g(w) = w^2$

$\mathcal{W} = \mathbb{R}$ $\quad w^* = \operatorname*{argmin}_{w \in \mathcal{W}} g(w)$

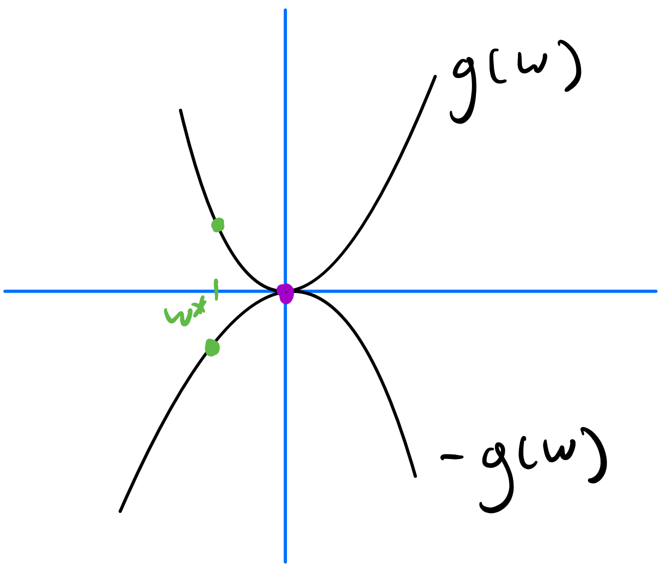
$= 0$

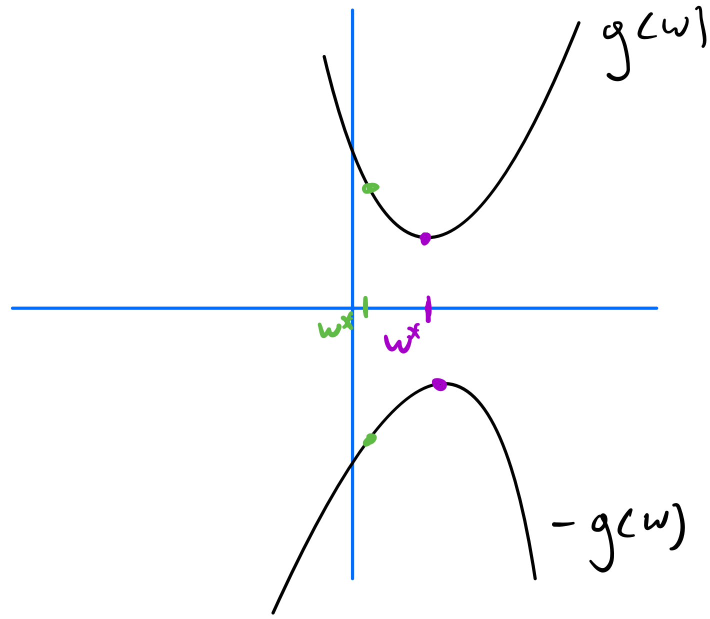
$$\min_{w \in \mathcal{W}} g(w) = 0 = g(w^*)$$



$\mathcal{W} = \{-1, 2\}$ $\quad \min\limits_{w \in \mathcal{W}} g(w) = 1 = g(w^*)$ $\quad \operatorname*{argmin}\limits_{w \in \mathcal{W}} g(w) = -1 = w^*$

Note: There is a relationship between minimizing and maximizing





$$w^* = \operatorname*{argmin}_{w \in \mathcal{W}} g(w) = \operatorname*{argmax}_{w \in \mathcal{W}} -g(w)$$

$$g(w^*) = \min_{w \in \mathcal{W}} g(w) = -\left(\max_{w \in \mathcal{W}} -g(w)\right) = -\left(-g(w^*)\right)$$

How do we solve minimization problems?

Cases:

1. If $\mathcal{W}$ is discrete (not continuous) we compare $g(w)$ for all $w \in \mathcal{W}$

2. If $\mathcal{W}$ is continuous we can sometimes use derivatives

$$\boxed{\text{We will focus on } \mathcal{W} \text{ continuous}}$$

If $g(w)$ is convex and twice differentiable then:

Cases: 1. If $\mathcal{W} = \mathbb{R}$ then w^* is the solution to $g'(w) = 0$

2. If $\mathcal{W} = [a, b]$ then w^* is the solution to $g'(w) = 0$ if this solution is in $[a, b]$. Otherwise, w^* is a or b

Twice differentiable: The second derivative of $g(w)$ written $g''(w)$ exists for all $w \in \mathcal{W}$

Convex: $g(w)$ is convex if $g''(w) \geq 0$ for all $w \in \mathcal{W}$

"Usually $g(w)$ is bowl shaped"

Ex: $g(w) = w^2$, $\mathcal{W} = \mathbb{R}$

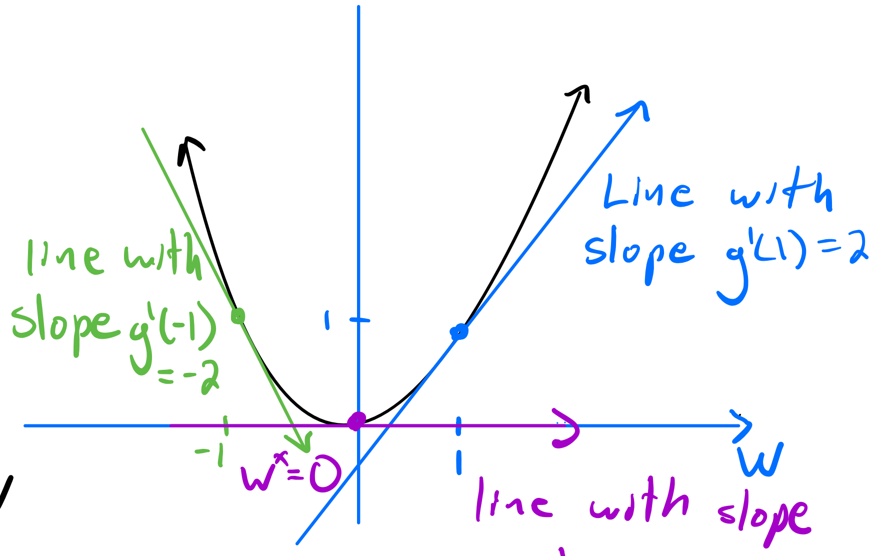
$w^* = \underset{w}{\operatorname{argmin}}\, g(w)$

$g'(w) = 2w$, $g''(w) = 2$

$g''(w) = 2 \geq 0$ for all $w \in \mathcal{W}$

$g(w)$ is convex

$g'(w) = 2w = 0 \Rightarrow w^* = 0$



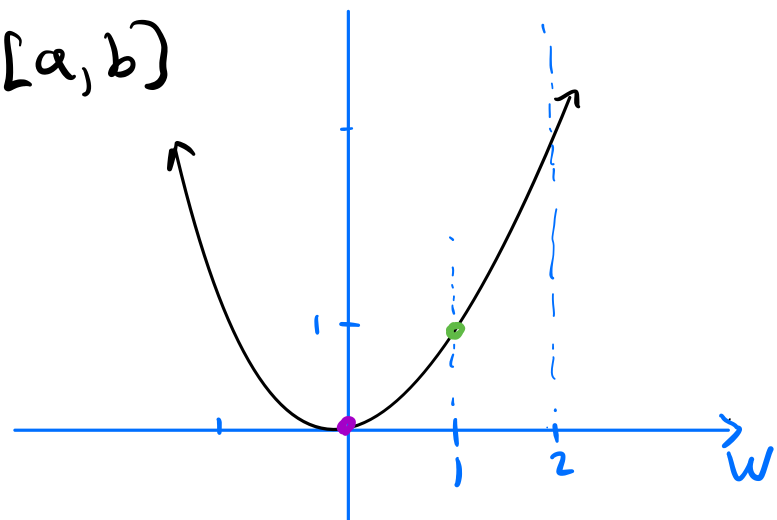
$g'(1) = 2 \cdot 1$

$g'(-1) = 2 \cdot (-1) = -2$

Ex: $g(w) = w^2$, $\mathcal{W} = [1, 2] = [a, b]$

$2w = 0 \Rightarrow w = 0 \notin [1, 2]$

$g(1) = 1$, $g(2) = 4$



$w^* = 1 \Rightarrow g(w^*) = \min_w g(w) = 1$

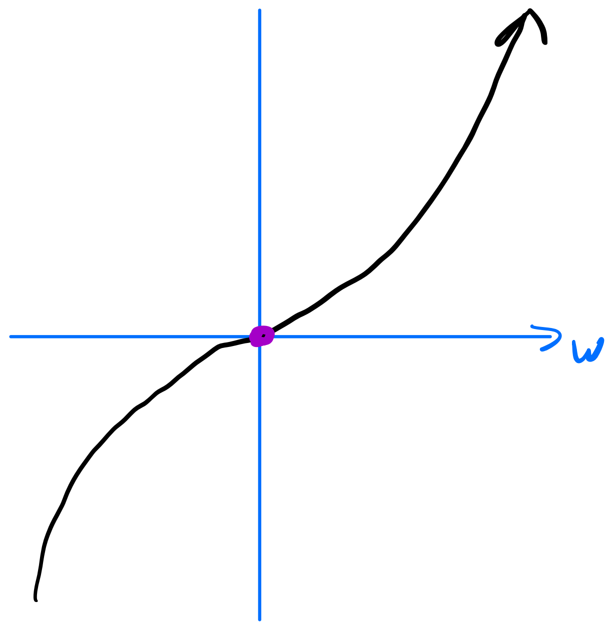
Ex: $g(w) = w^3$, $\mathcal{W} = \mathbb{R}$

$g'(w) = 3w^2$, $g''(w) = 6w$

$g'(w) = 3w^2 = 0$

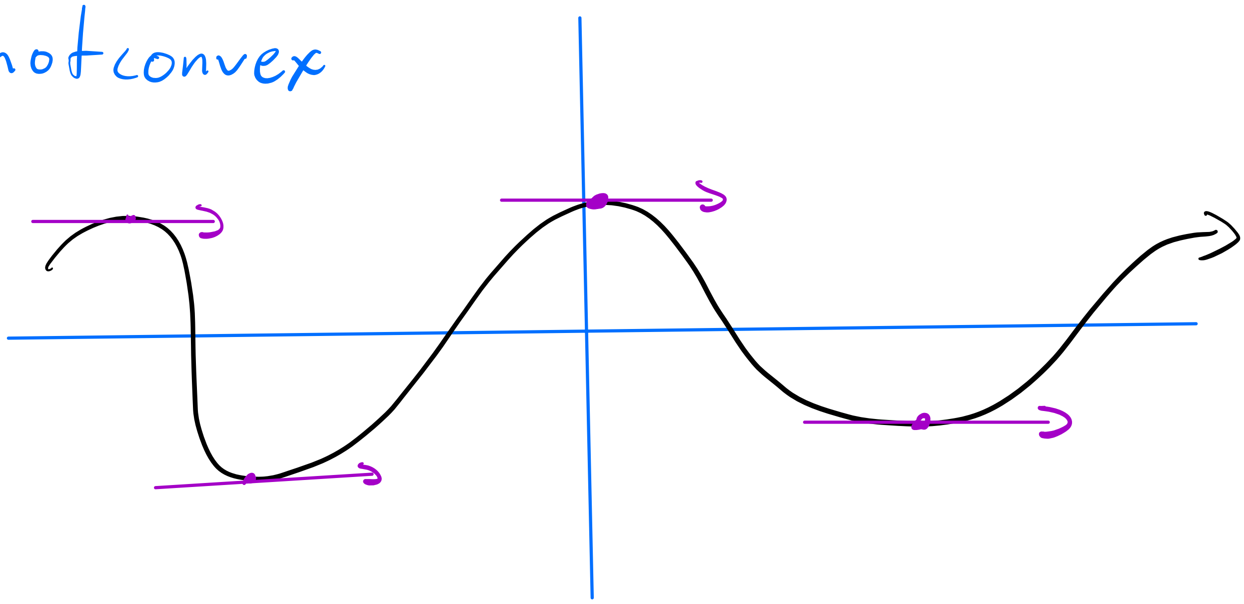
$\Rightarrow w = 0$

$g''(w) = 6w < 0$ when $w < 0$



$g(w)$ is not convex

Ex: not convex



Multidimensional Minimization

If $\mathcal{W} = \mathbb{R}^d$ for $d > 1$, and $g(\vec{w})$ is convex

Note: it is more complicated to check if $g(\vec{w})$ is convex if $d > 1$. So, I will just tell you

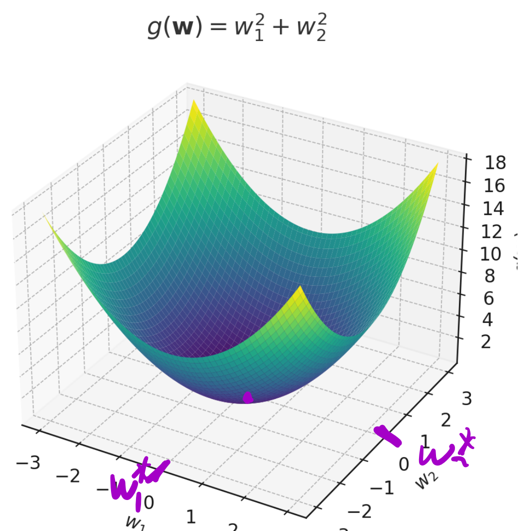
Then we calculate

$$\vec{w}^* = (w_1^*, \ldots, w_d^*)^T = \underset{\vec{w} \in \mathcal{W}}{\operatorname{argmin}}\, g(\vec{w})$$

by setting w_j^* as the solution to

$$\frac{\partial g}{\partial w_j}(\vec{w}) = 0 \qquad \text{for all } j \in \{1, \ldots, d\}$$

Ex: $g(\vec{w}) = g(w_1, w_2) = w_1^2 + w_2^2$, $\qquad \mathcal{W} = \mathbb{R}^2$



$$\frac{\partial g}{\partial w_1}(\vec{w}) = 2w_1 = 0 \Rightarrow w_1^* = 0$$

$$\frac{\partial g}{\partial w_2}(\vec{w}) = 2w_2 = 0 \Rightarrow w_2^* = 0$$

$$\vec{w}^* = (w_1^*, w_2^*) = (0,0)$$

$$g(\vec{w}^*) = \min_{\vec{w}} g(\vec{w})$$

Finding a good predictor (Linear Regression)

Optimization step of ERM

$\mathcal{X} = \mathbb{R}^{d+1}, \mathcal{Y} = \mathbb{R}$ regression

$\mathcal{F} \subset \{f \mid f: \mathbb{R}^{d+1} \to \mathbb{R}\}$

$\mathcal{D} = ((\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n))$

$\hat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\vec{x}_i), y_i)$ an estimate of $L(f)$ for all $f \in \mathcal{F}$

we want $\boxed{\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{L}(f)}$

pick $\mathcal{F} = \{f \mid f: \underset{\mathbb{R}^{d+1}}{\overset{\shortparallel}{\mathcal{X}}} \to \underset{\mathbb{R}}{\overset{\shortparallel}{\mathcal{Y}}}$ and $f(\vec{x}) = \vec{x}^T \vec{w}$ where $\vec{w} \in \mathbb{R}^{d+1}\}$

"Linear functions"

$f(\vec{x}) = \vec{x}^T \vec{w} + b = (x_1, \ldots, x_d)(w_1, \ldots, w_d)^T + b$

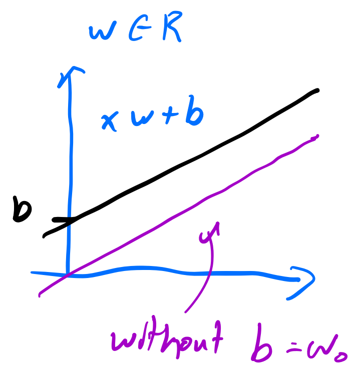
$= b + x_1 w_1 + \cdots + x_d w_d$

$b = w_0$

$x_0 = 1$

$= x_0 w_0 + x_1 w_1 + \cdots + x_d w_d$

$= \vec{x}_0^T \vec{w}_0$



Assume $\vec{x} = (x_0 = 1, x_1, \ldots, x_d) \in \mathbb{R}^{d+1}$ $\quad \vec{w} = (w_0, w_1, \ldots, w_d)^T$

$\vec{x}_i = (x_{i0} = 1, x_{i1}, \ldots, x_{id}) \in \mathbb{R}^{d+1}$

Notice $\hat{f} \in \mathcal{F}$ so $\hat{f}(\vec{x}) = \vec{x}^T \hat{\vec{w}}$ for some $\hat{\vec{w}} \in \mathbb{R}^d$

$$\hat{\vec{w}} = \underset{\underset{\mathcal{V}}{\overset{\shortparallel}{\vec{w} \in \mathbb{R}^{d+1}}}}{\operatorname{argmin}} \underset{g}{\underset{\downarrow}{\hat{L}(\vec{w})}} \quad \text{where } \hat{L}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\vec{x}_i^T \vec{w}, y_i)$$

Pick ℓ to be the squared loss. Then $\hat{L}(\vec{w})$ is convex.

Solve for $\hat{w}$:

$$\hat{L}(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i^T \vec{w} - y_i)^2$$

$$= \frac{1}{n}\left[(\vec{x}_1^T \vec{w} - y_1)^2 + \cdots + (\vec{x}_n^T \vec{w} - y_n)^2\right]$$

$\vec{x}_i = (x_{i0} = 1, x_{i1}, \ldots, x_{id})^T$, $\vec{w} = (w_0, w_1, \ldots, w_d)^T$

$$= \frac{1}{n}\Big[(x_{10} w_0 + x_{11} w_1 + \cdots + x_{1d} w_d - y_1)^2 + \ldots$$

$$+ (x_{n0} w_0 + x_{n1} w_1 + \cdots + x_{nd} w_d - y_n)^2\Big]$$

$$\hat{L}_i(\vec{w}) = (x_i^T \vec{w} - y_i)^2 = (x_{i0}w_0 + x_{i1}w_1 + \dots + x_{id}w_d - y_i)^2$$

$$= \frac{1}{n}\left[\hat{L}_1(\vec{w}) + \dots + \hat{L}_n(\vec{w})\right]$$

$$\frac{\partial \hat{L}}{\partial w_j}(\vec{w}) = \frac{1}{n}\left[\frac{\partial \hat{L}_1}{\partial w_j}(\vec{w}) + \dots + \frac{\partial \hat{L}_n}{\partial w_j}(\vec{w})\right]$$

$$\hat{L}_i(\vec{w}) = g(u_i) = u_i^2, \quad u_i = \vec{x}_i^T \vec{w} - y_i$$

$$\frac{\partial \hat{L}_i}{\partial w_j}(\vec{w}) = \frac{\partial g}{\partial u_i} \frac{\partial u_i}{\partial w_j}(\vec{w})$$

$$= 2 u_i x_{ij}$$

$$= 2 x_{ij}(\vec{x}_i^T \vec{w} - y_i)$$

$$= \frac{2}{n}\left[x_{1j}(\vec{x}_1^T \vec{w} - y_1) + \dots + x_{nj}(\vec{x}_n^T \vec{w} - y_n)\right]$$

$$= \frac{2}{n} \sum_{i=1}^{n} x_{ij}(\vec{x}_i^T \vec{w} - y_i)$$

$$\frac{\partial \hat{L}}{\partial w_j}(\vec{w}) = \frac{2}{n}\sum_{i=1}^{n} x_{ij}(\vec{x}_i^T\vec{w} - y_i) = 0$$

$$\Rightarrow \frac{2}{n}\sum_{i=1}^{n} x_{ij}\vec{x}_i^T\vec{w} - \frac{2}{n}\sum_{i=1}^{n} x_{ij}y_i = 0$$

$$\sum_{i=1}^{n} x_{ij}\vec{x}_i^T\vec{w} = \sum_{i=1}^{n} x_{ij}y_i \quad \text{for all } j \in \{0, 1, \ldots, d\}$$

$$\sum_{i=1}^{n} x_{i0}\vec{x}_i^T\vec{w} = \sum_{i=1}^{n} x_{i0}y_i$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{id}\vec{x}_i^T\vec{w} = \sum_{i=1}^{n} x_{id}y_i$$

$$\sum_{i=1}^{n} \overbrace{\vec{x}_i\vec{x}_i^T}^{\text{outer product}}\vec{w} = \sum_{i=1}^{n}\vec{x}_i y_i \qquad A = \sum_{i=1}^{n}\vec{x}_i\vec{x}_i^T$$

$$A\vec{w} = \vec{b} \qquad b = \sum_{i=1}^{n}\vec{x}_i y_i$$

$$\hat{\vec{w}} \stackrel{\text{def}}{=} \vec{w} = A^{-1}\vec{b} \quad \text{if } A^{-1} \text{ exists}$$

inverse of A

What $\mathcal{A}(D)$?

$$D = ((\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$$

$$\mathcal{A}(D) = \hat{f} \in \mathcal{F}$$

Depends on D

where $\hat{f}(\vec{x}) = \vec{x}^T \hat{\vec{w}}$ and $\hat{\vec{w}} = \underbrace{A^{-1} \vec{b}}$

$\vec{x} \in \mathcal{X}$

Algorithm: Closed form linear regression Learner

---

input: $D = ((\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n))$

$A \leftarrow \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T$

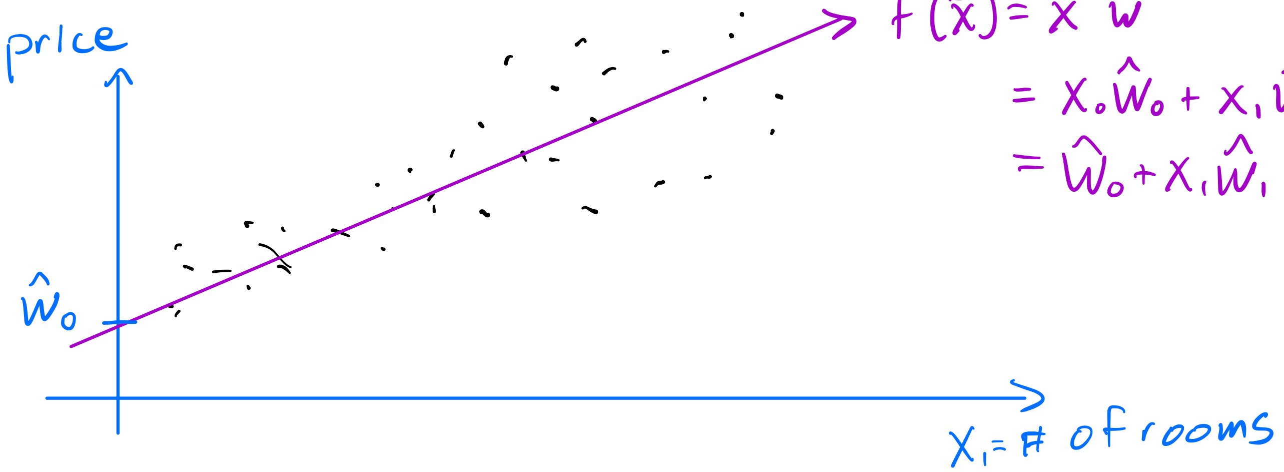
$b \leftarrow \sum_{i=1}^{n} \vec{x}_i y_i$

$\hat{\vec{w}} \leftarrow A^{-1} b$

return $\hat{f}(\vec{x}) = \vec{x}^T \hat{\vec{w}}$

$\underline{\underline{Ex}}$: $d=1$, $\mathcal{X}=\mathbb{R}^{1+1}=\mathbb{R}^2$, $\mathcal{Y}=\mathbb{R}$, $\hat{\vec{w}}=(\hat{w}_0, \hat{w}_1)^T$

$$\mathcal{D}=((\vec{x}_1, y_1), \dots, (\vec{x}_n, y_1))$$



$\hat{f}(\vec{x}) = \vec{x}^T \hat{\vec{w}}$

$= x_0 \hat{w}_0 + x_1 \hat{w}_1$

$= \hat{w}_0 + x_1 \hat{w}_1$