Lecture 4: Least Squares and Optimization

Example

Least Squares & Classification

Setup: n training samples, $(\underline{x}_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}$ for $i = 1, \dots, n$

let $X = \begin{bmatrix} - & \underline{x}_1^T & - \\ - & \underline{x}_2^T & - \\ & \vdots & \\ - & \underline{x}_n^T & - \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

Assume $n \geq p$, X is rank-p. Then

compute $\underline{\hat{w}} = \underset{\underline{w}}{\arg\min} \|y - X\underline{w}\|_2^2$,

$\tilde{y} = X\underline{\hat{w}} \in \mathbb{R}^n$ (i.e. not +1 or -1 labels)

alternatively, we might be tempted to consider

$\underline{\hat{w}} = \underset{\underline{w}}{\arg\min} \|y - \text{sign}(X\underline{w})\|_2^2$

but this is very hard to solve computationally

Classification rule: predict +1 if $\tilde{y}_i > 0$ and -1 if $\tilde{y}_i < 0$

$\Rightarrow \hat{y}_i = \text{sign}(\tilde{y}_i)$

$\hat{y} = \text{sign}(\tilde{y})$

for a new sample $\underline{x}_{new} \in \mathbb{R}^p$, want to predict new, unknown label y_{new}.

$\tilde{y}_{new} = \langle \underline{x}_{new}, \underline{\hat{w}} \rangle \in \mathbb{R}$

$\hat{y}_{new} = \text{sign}(\tilde{y}_{new})$

Optimization approach

$\hat{\underline{w}}$ = "argument $\underline{w}$ that minimizes" $\sum_{i=1}^{n} r_i^2(\underline{w})$

$$= \underset{\underline{w}}{\text{argmin}} \sum_{i=1}^{n} r_i^2(\underline{w}) \qquad = \underset{\underline{w}}{\text{argmin}} \; \|\underline{r}\|_2^2$$

2-norm or Euclidean norm: $\|\underline{a}\|_2 := \left(\sum_{i=1}^{n} a_i^2\right)^{1/2}$

$$= \underset{\underline{w}}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} w_j x_{ij}\right)^2 \qquad = \underset{\underline{w}}{\text{argmin}} \; \underbrace{\|\underline{y} - X\underline{w}\|_2^2}_{f(\underline{w})}$$

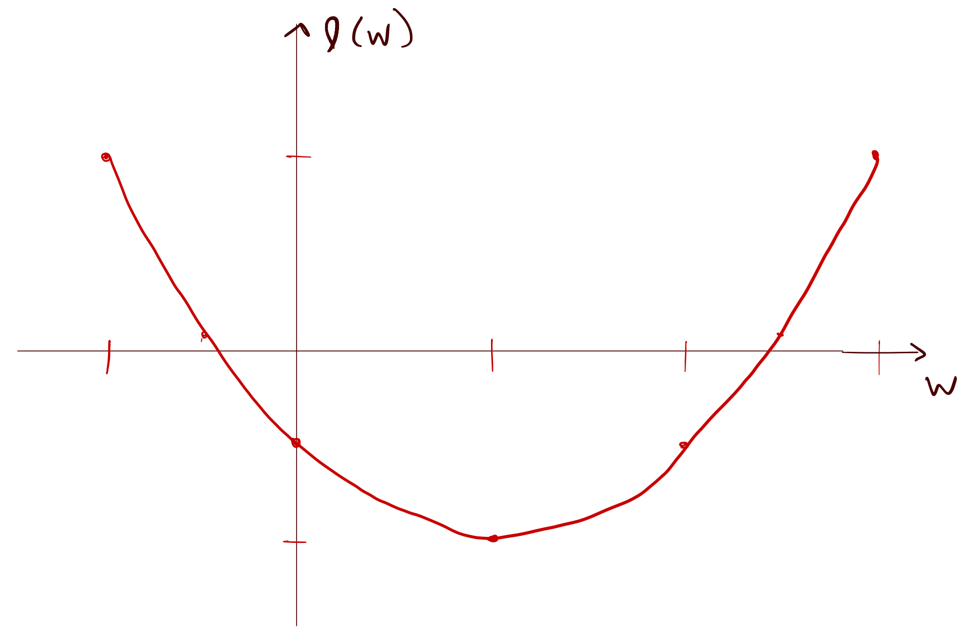
$$= \underset{\underline{w}}{\text{argmin}} \; \underline{y}^T\underline{y} - \underline{y}^T X \underline{w} - \underline{w}^T X^T \underline{y} + \underline{w}^T X^T X \underline{w}$$

Warmup.

$$\ell(w) = \frac{1}{2} w^2 - w - \frac{1}{2}$$

$$\hat{w} = \underset{w}{\text{argmin}} \; \ell(w)$$

$$\frac{d\ell}{dw} = 2 \cdot \frac{1}{2} w - 1 = 0 \implies \hat{w} = 1$$



Positive definite matrices

From the geometric perspective, we saw that it was important for finding a unique least squares solution $\hat{\underline{w}}$ that $X^T X$ be invertible.

Is this important in the optimization setting as well? YES!

The following two things are equivalent for $X \in \mathbb{R}^{n \times p}$ with $n \geq p$, rank$(X) = p$ (X has p linearly independent columns)

1. $X^T X \in \mathbb{R}^{p \times p}$ is invertible ($(X^T X)^{-1}$ exists)
2. $X^T X$ is positive definite

A matrix Q is positive definite (p.d.) if

$$\underline{w}^T Q \underline{w} > 0 \quad \text{for all } \underline{w} \neq 0$$

Shorthand: $Q \succ 0$

A matrix Q is positive semi-definite (p.s.d.) if

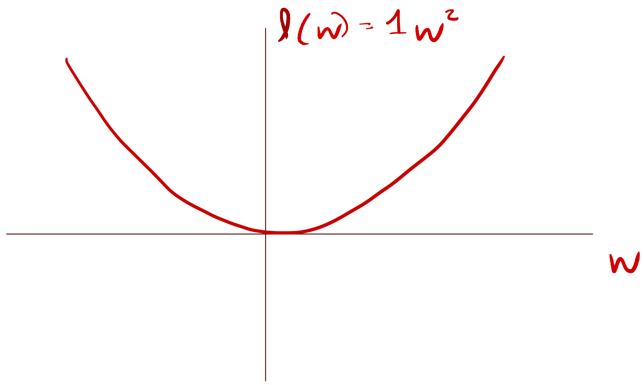
$$\underline{w}^T Q \underline{w} \geq 0 \quad \text{for all } \underline{w} \neq 0$$

Shorthand: $Q \succeq 0$

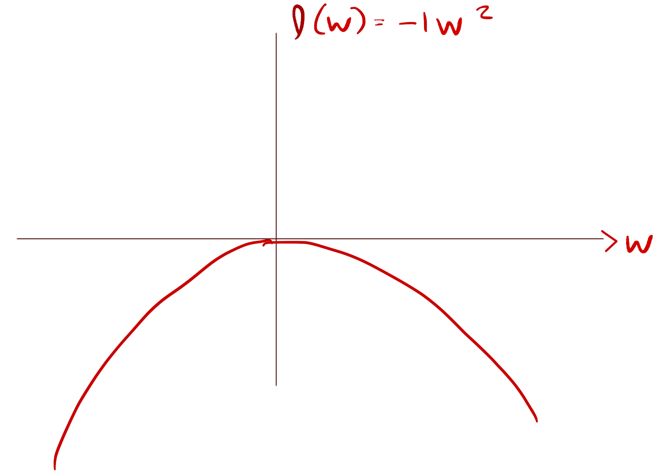
this does not mean all elements of Q are positive or non-negative!

Ex: $w \in \mathbb{R},\ Q \in \mathbb{R} \implies w^T Q w = Q w^2 > 0$ if $Q > 0$

imagine trying to minimize $l(w) = Q w^2$



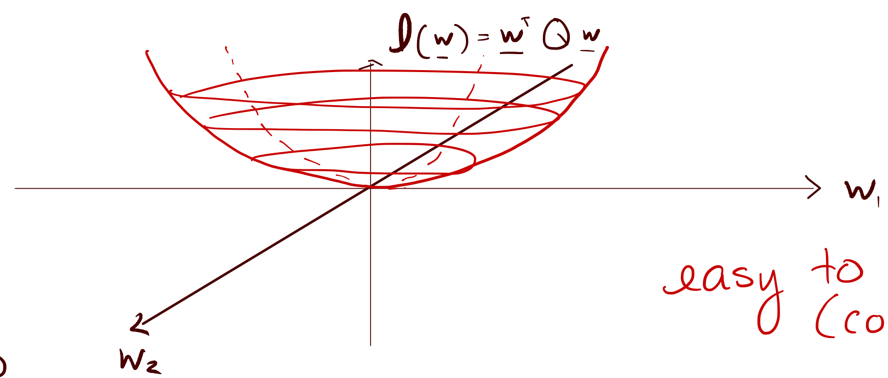
easy to minimize
(convex)



hard to minimize

Ex $\underline{w} \in \mathbb{R}^2,\ Q \in \mathbb{R}^{2\times 2}$

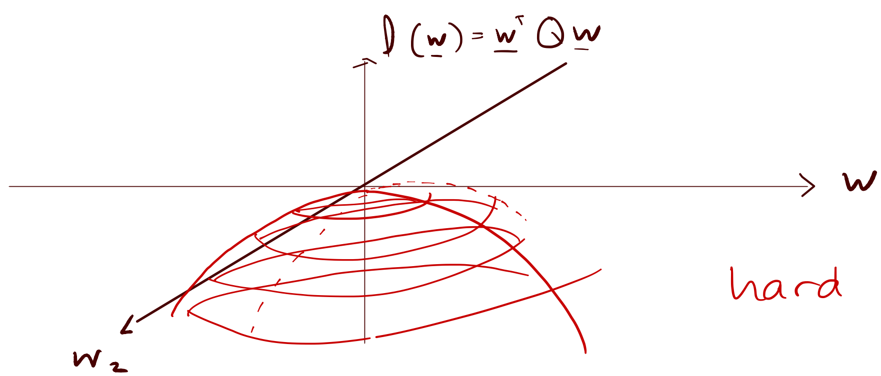
$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \implies$$



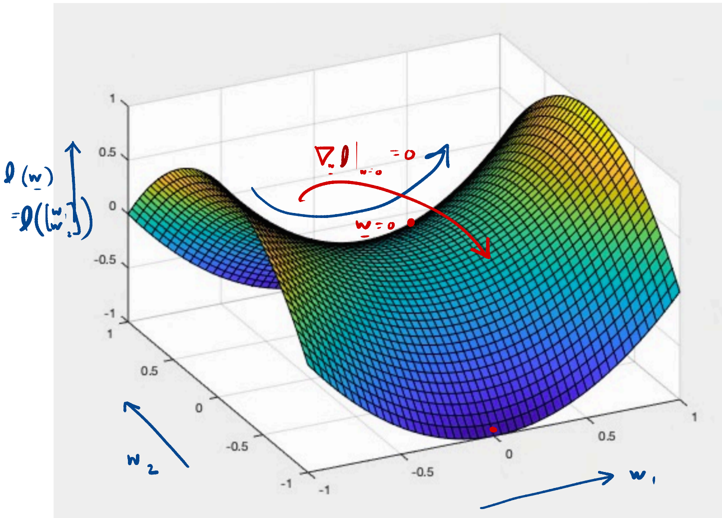
easy to minimize
(convex)

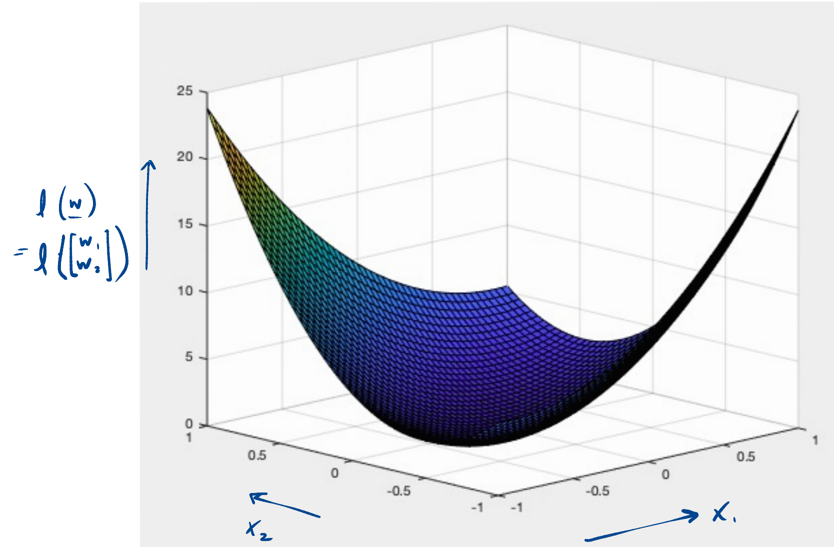
$Q \succ 0$ b/c $\underline{w}^T Q \underline{w} = w_1^2 + w_2^2 > 0$

$$Q = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \implies$$



hard to minimize





$$Q = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

$$l(\underline{w}) = \underline{w}^T Q \underline{w}$$

this an example of a "saddle point"

Q is <u>NOT</u> positive definite

$$Q = \begin{bmatrix} 5 & -5 \\ -5 & 1 \end{bmatrix}$$

Recall Q is pos. def. if $\underline{w}^T Q \underline{w} > 0$ for all $\underline{w} \neq 0$

$$l(\underline{w}) = \underline{w}^T Q \underline{w}$$

$l(\underline{w}) > 0$ for $\underline{w} \neq 0$

$\Rightarrow Q$ is positive definite

Properties of Positive Definite Matrices

1) if $P \succ 0$ and $Q \succ 0$, then $P+Q \succ 0$

$$\underline{w}^T P \underline{w} > 0 \;\&\; \underline{w}^T Q \underline{w} > 0 \Rightarrow \underline{w}^T (P+Q) \underline{w} = \underline{x}^T P \underline{x} + \underline{x}^T Q \underline{x} > 0$$

2) if $Q \succ 0$ and $a > 0$, then $aQ \succ 0$

$$\underline{w}^T Q \underline{w} > 0 \Longrightarrow \underline{w}^T (aQ) \underline{w} = a\, \underline{w}^T Q \underline{w} > 0$$

3) for <u>any</u> A, $A^T A \succeq 0$ and $AA^T \succeq 0$

if the columns of A are linearly independent, then $A^T A \succ 0$

Note $\underline{w}^T \underline{w} \geq 0$ always, and $\underline{w}^T \underline{w} = 0$ only if $\underline{w} = 0$.

Let $\underline{\tilde{w}} := A \underline{w}$

$\underline{w}^T A^T A \underline{w} = \underline{\tilde{w}}^T \underline{\tilde{w}} \geq 0$. now $\underline{\tilde{w}}^T \underline{\tilde{w}} = 0$ only if $\underline{\tilde{w}} = A \underline{w} = 0$. $A \underline{w} = 0$ if either ⓐ $\underline{w} = 0$

or ⓑ columns of A are linearly dependent.

4) if $Q \succ 0$, then Q^{-1} exists (consider special case where Q is diagonal $Q = \begin{bmatrix} q_1 & & 0 \\ & q_2 & \\ 0 & & q_n \end{bmatrix}$)

$\Rightarrow Q \succ 0 \Longleftrightarrow$ all $q_i > 0$

5) $Q \succ P$ means $Q - P \succ 0$

If $Q = X^T X$, is Q positive definite?

For Q to be P.D., need $\underline{w}^T Q \underline{w} > 0$ for all $\underline{w} \neq 0$

now $\underline{w}^T Q \underline{w} = \underline{w}^T X^T X \underline{w} = \underline{\tilde{w}}^T \underline{\tilde{w}} \geq 0$ where we define $\underline{\tilde{w}} := X\underline{w}$

For Q to be positive definite (not positive semi definite, where $\underline{w}^T Q \underline{w}$ might $= 0$)
we need to ensure $\underline{\tilde{w}}^T \underline{\tilde{w}} > 0$

Now $\underline{\tilde{w}}^T \underline{\tilde{w}} = 0$ only if $\underline{\tilde{w}} = 0$ (because $\underline{\tilde{w}}^T \underline{\tilde{w}} = \tilde{w}_1^2 + \tilde{w}_2^2 + \cdots + \tilde{w}_p^2$)

$\underline{\tilde{w}} = 0$ if $X\underline{w} = 0$. Can $X\underline{w} = 0$ for some $\underline{w} \neq 0$?

Recall $X\underline{w}$ = weighted sum of columns of X. I.e. $X\underline{w} = w_1 \underline{X}_1 + w_2 \underline{X}_2 + \cdots + w_p \underline{X}_p$.

Recall defn of Linear independence: the columns of X are L.I. if no weighted sum $= 0$ (unless all weights $= 0$)

If cols. of X are LI, then $X\underline{w} \neq 0$ unless $\underline{w} = 0$

$\implies w^T X^T X w > 0 \implies X^T X$ is pos definite

Least squares optimization problem

$$\hat{\underline{w}} = \underset{\underline{w}}{\operatorname{argmin}}\ \ell(\underline{w}) \qquad \text{where } \ell(\underline{w}) = \underline{y}^T\underline{y} - \underline{y}^T X \underline{w} - \underline{w}^T X^T \underline{y} + \underline{w}^T X^T X \underline{w}$$

$\ell(\underline{w})$ maps $\underline{w} \in \mathbb{R}^p$ to $\mathbb{R}$

Assume $\ell(\underline{w})$ is convex (more on this later):

When w is a scalar, we set derivative $\frac{d\ell}{dw}$ to zero and solve for w.

When $\underline{w}$ is a vector, we set gradient $\nabla_{\underline{w}}\ell$ to zero and solve for $\underline{w}$

$$\nabla_{\underline{w}}\ell := \begin{bmatrix} df/dw_1 \\ df/dw_2 \\ \vdots \\ df/dw_p \end{bmatrix}$$

Ex. $\ell(\underline{w}) = \underline{w}^T\underline{c} = c_1w_1 + c_2w_2 + \cdots + c_pw_p$

$$\nabla_w \ell = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} = \underline{c}$$

"Rule 1"

Ex. $\ell(\underline{w}) = \|\underline{w}\|^2 = \underline{w}^T\underline{w} = w_1^2 + w_2^2 + \cdots + w_p^2$

$$\nabla_w \ell = \begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_p \end{bmatrix} = 2\underline{w}$$

Ex. $l(\underline{w}) = \underline{w}^T Q \underline{w} = \sum_{i=1}^{p} \sum_{j=1}^{p} w_i Q_{ij} w_j$

$$\frac{d(w_i Q_{ij} w_j)}{dw_k} = \begin{cases} 2Q_{kk} w_k & \text{if } k = i = j \\ Q_{kj} w_j & \text{if } i = k \neq j \\ Q_{ik} w_i & \text{if } j = k \neq i \end{cases}$$

$$\frac{dl}{dw_k} = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{d(w_i Q_{ij} w_j)}{dw_k}$$

$\Rightarrow \nabla_{\underline{w}} l = Q\underline{w} + Q^T \underline{w}$.

if Q is symmetric (ie. $Q = Q^T$), then

$$\nabla_w \underline{w}^T Q \underline{w} = 2Q\underline{w}$$

"Rule 2"

Ex. Least squares: $l(\underline{w}) = y^T y - 2\underline{w}^T X^T y + \underline{w}^T X^T X \underline{w}$

$$l(\underline{w}) = \sum_{i=1}^{n} (y_i - \underline{x}_i^T \underline{w})^2 = \Bigg\| \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_T \end{bmatrix}}_{\begin{bmatrix} y_1 - x_1^T \underline{w} \\ y_2 - x_2^T \underline{w} \\ \vdots \\ y_n - x_n^T \underline{w} \end{bmatrix}} \Bigg\|_2^2 = \|y - X\underline{w}\|_2^2$$

$$= (y - X\underline{w})^T (y - X\underline{w})$$

$$= y^T y - y^T X \underline{w} - \underline{w}^T X^T y + \underline{w}^T X^T X \underline{w}$$

$$= y^T y - \underbrace{2\underline{w}^T X^T y}_{\text{use Rule 1}} + \underbrace{\underline{w}^T X^T X \underline{w}}_{\text{use Rule 2}} \longrightarrow \underbrace{X^T X}_{\text{symmetric}} \text{ here is like } Q \text{ in Rule 2}$$

$$\nabla_w l = 0 - 2X^T y + 2X^T X \underline{w}$$

$\nabla_{\underline{w}} \ell = 0 - 2X^T y + 2X^T X \underline{w}$

If $Q = X^T X \succ 0$, then we can compute gradient and set it to zero because f is convex

$$\nabla_{\underline{w}} \ell = 0$$

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

$$\Rightarrow \hat{\underline{w}} = (X^T X)^{-1} X^T y$$ (what we got from geometric perspective!)

So far, we've assumed $X \in \mathbb{R}^{n \times p}$ has $n \geq p$ and p cols of X are LI.

- Colums of X are LI $\Rightarrow$ $Q = X^T X \succ 0$ (i.e. $X^T X$ is positive definite)
- $X^T X$ is pos. definite $\Rightarrow$ least square loss $y^T y - 2\underline{w}^T X^T y + \underline{w}^T X^T X \underline{w}$ is convex
- if $X^T X \succ 0$, then $X^T X$ has inverse
- $X^T X$ has inverse $\Rightarrow$ $\hat{\underline{w}} = (X^T X)^{-1} X^T y$ exists and is unique