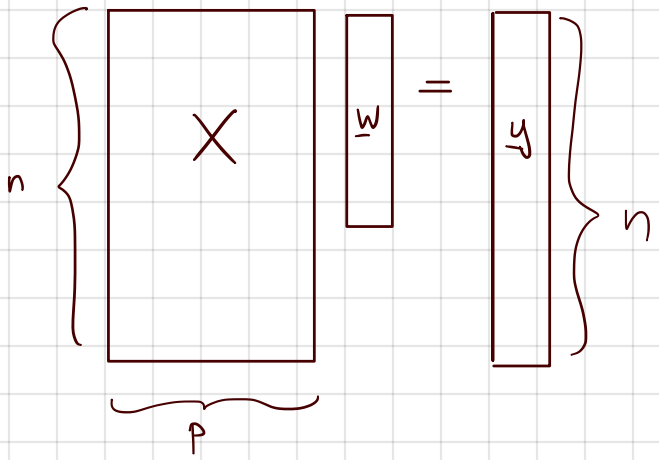
Lecture 4: Least squares and optimization

Mathematical Foundations of Machine Learning
University of Chicago

Given: vector of labels $\underline{y} \in \mathbb{R}^n$

matrix of features $X \in \mathbb{R}^{n \times p}$

Want: vector of weights $\underline{w} \in \mathbb{R}^p$



$\Longleftarrow$ n equations
p unknowns

Let $\underline{x}_i \in \mathbb{R}^p$ be feature vector of i^{th} sample;
$= (i^{th}$ row of $X)^T$

Let $\underline{X}_j \in \mathbb{R}^n$ be j^{th} feature for all samples
$= j^{th}$ col of X

Assume: $n \geq p$

$\text{Rank}(X) = p$ (X has p linearly independent columns)

If $\underline{y} = X\underline{w}$, then we have a system of n linear equations;

i^{th} equation: $y_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip}$

$$= \sum_{j=1}^{p} w_j x_{ij} = \langle \underline{w}, \underline{x}_i \rangle$$

↑ i^{th} row of X (transposed)

In general, $\underline{y} \neq X\underline{w}$ for any $\underline{w}$ (because of modeling errors, noise)

Define residual $r_i = r_i(\underline{w}) = y_i - \langle \underline{w}, \underline{x}_i \rangle$

LEAST SQUARES ESTIMATION: if columns of X are linearly independent

find $\underline{w}$ to minimize $\underbrace{\sum_{i=1}^{n} |r_i(\underline{w})|^2}_{= \|\underline{r}\|^2} \Rightarrow \hat{w} = (X^T X)^{-1} X^T y = X^{\#} y$ $(X^{\#} = (X^T X)^{-1} X^T =$ "pseudoinverse")

$y = X\hat{w} = X(X^T X)^{-1} X^T y = P_X y =$ projection of y onto range(X)

Least Squares & Classification

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

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

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

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

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

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

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

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

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

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

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

Optimization approach

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

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

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

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

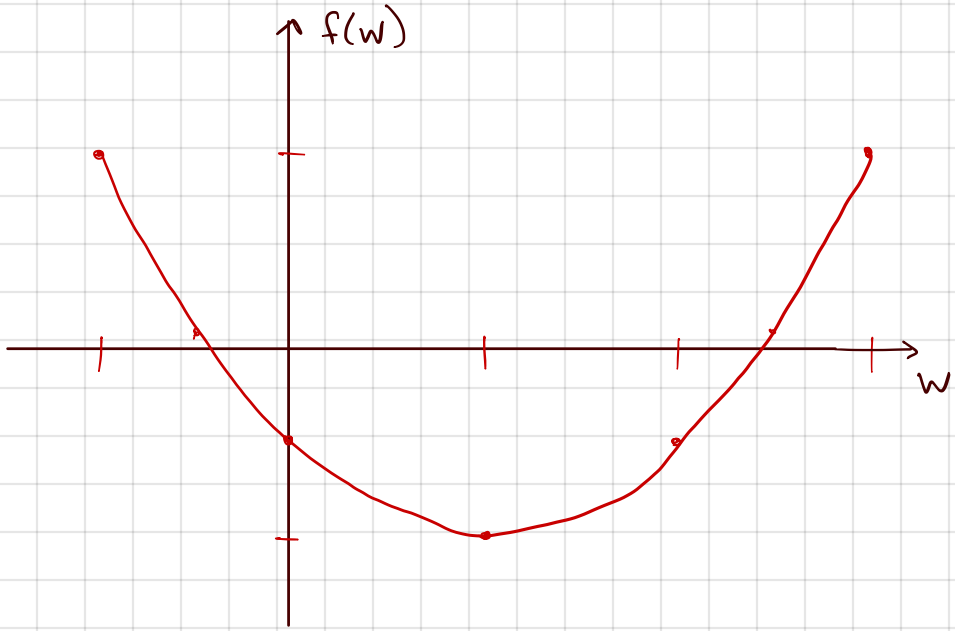
2-norm or Euclidean norm: $\|\underline{a}\|_2 := \left(\sum_{i=1}^{n} a_i^2\right)^{1/2}$

Warmup.

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

$$\hat{w} = \underset{w}{\operatorname{argmin}} \; f(w)$$

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



Positive definite matrices

From the geometric perspective, we saw that it was important for finding a unique least squares solution $\hat{w}$ that X^TX 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 $n \geq p$, $\text{rank}(X) = p$

(X has p linearly independent columns)

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

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

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

Shorthand: $Q \succ 0$

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

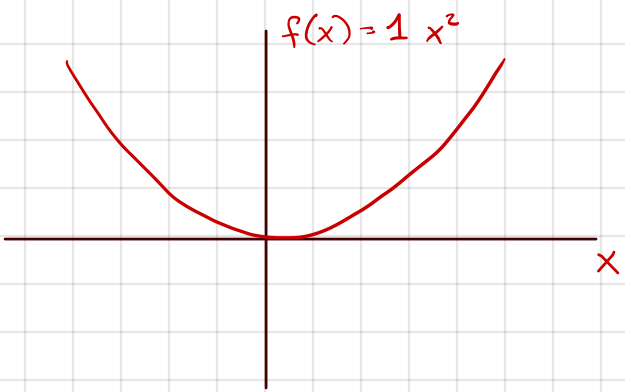
$$\underline{x}^T Q \underline{x} \geq 0 \quad \text{for all} \quad \underline{x} \neq 0$$

Shorthand: $Q \succeq 0$

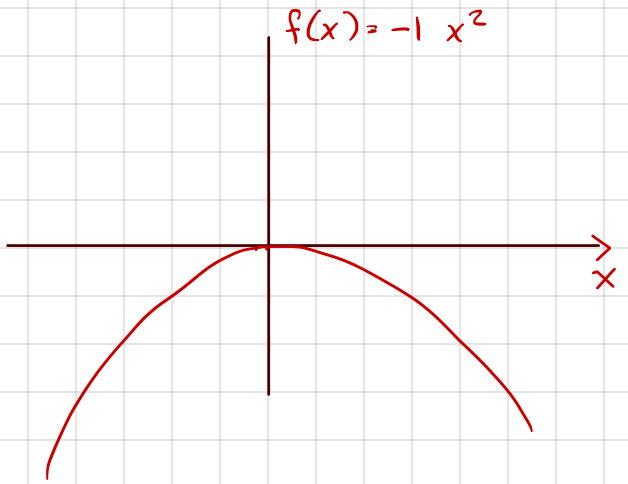
this does not mean all elements of Q are positive or non-negative!

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

imagine trying to minimize $f(x) = Q x^2$



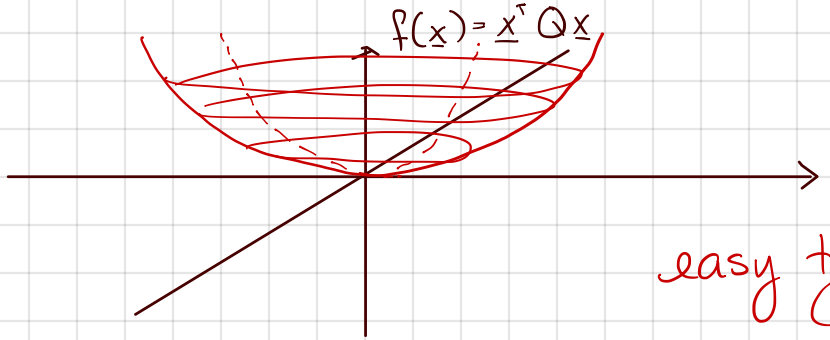
easy to minimize (convex)



hard to minimize

Ex. $\underline{x} \in \mathbb{R}^2$, $Q \in \mathbb{R}^{2x2}$

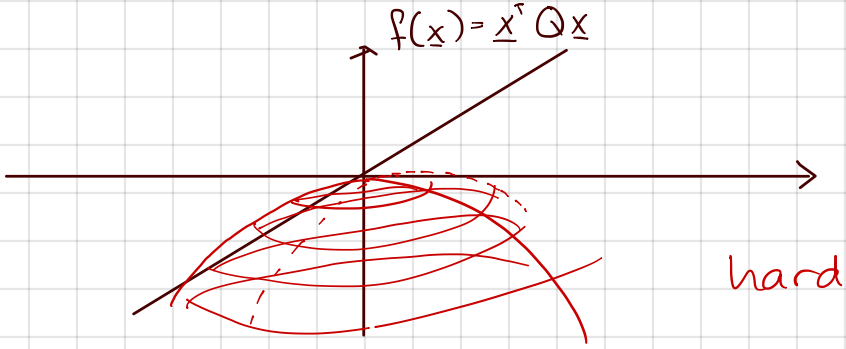
$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow$$



easy to minimize (convex)

$Q > 0$ b/c $\underline{x}^T Q \underline{x} = x_1^2 + x_2^2 > 0$

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



hard to minimize

Properties of Positive Definite Matrices

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

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

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

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

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

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

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

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

$\underline{x}^T A^T A \underline{x} = \tilde{\underline{x}}^T \tilde{\underline{x}} \geq 0$. now $\tilde{\underline{x}}^T \tilde{\underline{x}} = 0$ only if $\tilde{\underline{x}} = Ax = 0$. $Ax = 0$ if either (a) $x = 0$ or (b) columns of A are linearly dependent.

4) if $A \succ 0$, then A^{-1} exists

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

Least squares optimization problem

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

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

Assume $f(\underline{w})$ is convex (more on this later):
When w is a scalar, we set derivative $\frac{df}{dw}$ to zero and solve for w.

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

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

Ex. $f(\underline{w}) = \underline{w}^T \underline{c} = c_1 w_1 + c_2 w_2 + \cdots + c_p w_p$

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

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

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

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

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

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

$$\Rightarrow \nabla_{\underline{w}} f = Qx + Q^T x.$$

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

$$\nabla_{\underline{w}} \; w^T Q w = 2Qw$$

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

if $Q \succ 0$, then we can compute gradient and set it to zero because f is convex

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

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

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