SWCON253 Machine Learning

Lecture 5. Least Squares and Optimization in Machine Learning

Given

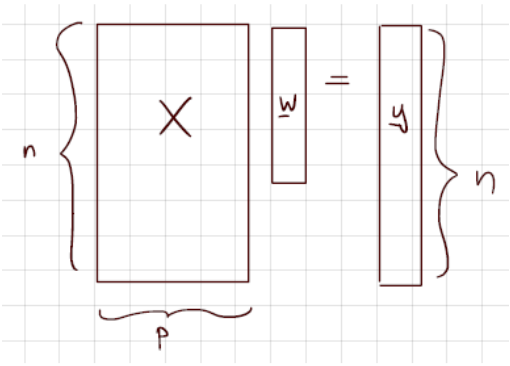
Labels $\underline{y} \in \mathbb{R}^n$ for n training samples

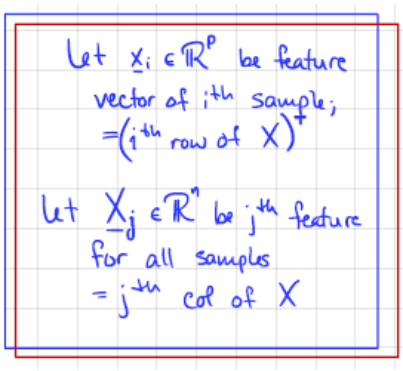
Features $X \in \mathbb{R}^{nxp}$ (p features)

We assume that n ≥ p, rank (X) = p (X has p linearly independent columns)

We want to find $\underline{w}$ so that $\hat{\underline{y}} = X\underline{w} \approx \underline{y}$

Pictorially,





n equations, p unknowns

In classification,

$y_i \in \{-1, +1\}$

We will compute the weight vector $\hat{\underline{w}}$ that minimizes the sum of squared errors. So, we want to minimize the distance between $\underline{y}$ and $X\underline{w}$ that we predict using the linear model.

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

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

Let $\hat{\underline{y}} = X\hat{\underline{w}}$. The elements of $\hat{\underline{y}}$ are not $\in \{-1, +1\}$

That is, we are not getting a bunch of plus and minus ones here. We are actually getting real numbers. For example, for one sample, y might be 0.5, for another sample it might be -0.7.

We have to come up with a classification rule.

Classification rule

We predict +1 if $\hat{y}_i > 0$ and -1 if $\hat{y}_i < 0$

In other words, $\tilde{y}_i = sign(\hat{y}_i) \in \{-1, +1\}$

We can actually see not only if $\hat{y}_i$ is close to y, but also whether the labels $\tilde{y}_i$ that we're predicting equal the labels that we actually observed.

For a new sample, $x_{new} \in \mathbb{R}^p$ we want to predict its label (new unknown label)

$$\hat{y}_{new} = \langle \underline{\hat{w}}, \underline{x}_{new} \rangle$$

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

Overall this is pipeline of the classification system.
It takes all the training data.
We try to learn a good weight factor in the context of this linear model and then
we take a new sample and
we plugged the learned the weight vector $\underline{w}$ that we computed with our training data into the linear model and
use that to come up with a predictive label.

$$\underline{\hat{w}} = \underset{\underline{w}}{\operatorname{argmin}} \left\| \underline{y} - X\underline{w} \right\|_2^2$$
$$\underline{\hat{w}} = (X^T X)^{-1} X^T \underline{y}$$

What we said is $\underline{\hat{w}}$ is the value w that minimizes the length $\underline{y}$ minus $X\underline{w}$ (2-norm squared)
We have arrived this equation for $\underline{\hat{w}}$ using this geometric argument.
What we would like to do today is to derive that same expression from different way by thinking about it as an optimization problem, instead of thinking about it geometrically.

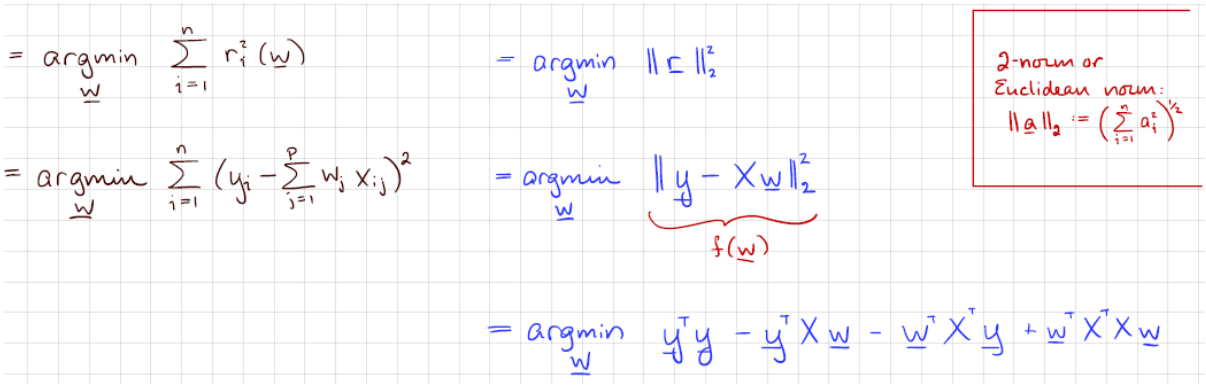
Optimization approach

Recall

$$\|\underline{x}\|_2^2 = \sum_{i=1}^{n} x_i{}^2 = \underline{x}^T \underline{x} = \langle \underline{x}, \underline{x} \rangle$$

$$\underline{\hat{w}} = \underset{\underline{w}}{\operatorname{argmin}} \left\| \underline{y} - X\underline{w} \right\|_2^2$$
$$= \underset{\underline{w}}{\operatorname{argmin}} (\underline{y} - X\underline{w})^T (\underline{y} - X\underline{w})$$
$$= \underset{\underline{w}}{\operatorname{argmin}} \underline{y}^T \underline{y} - (X\underline{w})^T y - \underline{y}^T Xw + (X\underline{w})^T (X\underline{w})$$
$$= \underset{\underline{w}}{\operatorname{argmin}} \underline{y}^T \underline{y} - \underline{w}^T X^T y - \underline{y}^T Xw + \underline{w}^T X^T X\underline{w}$$
$$= \underset{\underline{w}}{\operatorname{argmin}} \underline{y}^T \underline{y} - 2\underline{w}^T X^T y + \underline{w}^T X^T X\underline{w}$$

We started off by saying we wanted to minimize the sum of squared errors or sum of squared residuals and now
we have taken that objective function we want to minimize and we've rewritten it as this long matrix vector equation and now what we want to do is
we want figure out how to actually solve that optimization problem.



Warmup

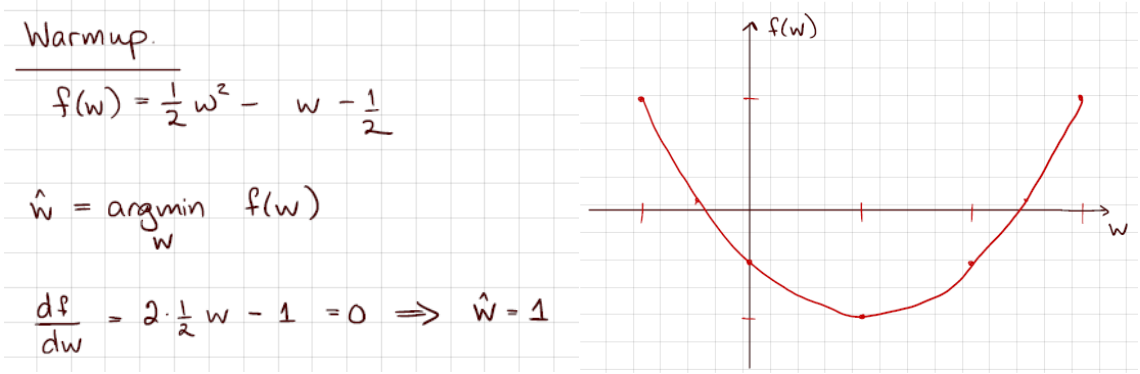
$$f(w) = \frac{1}{2}w^2 - w - \frac{1}{2}$$

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

$$\frac{df}{dw} = w - 1 = 0$$

Thus, $\hat{w} = 1$

We take the derivative and set it equal to zero. That is going to give us the minimum.



So now we want to know is how can we actually go about solving optimization problems like this where instead of optimizing over scaler numbers like a scale W. we instead want to optimize over weight vectors

When we look at this geometric perspective. we said it was really important that x^Tx is invertible, and that's how we came up with our solution which is the function of x^Tx, so something similar shows up to when we think about the optimization perspective as well.

Positive Definite Matrices

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

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

The following two things are equivalent for $X \in \mathbb{R}^{nxp}$ with n ≥ p, rank(X) = p

(X has p linearly independent columns)

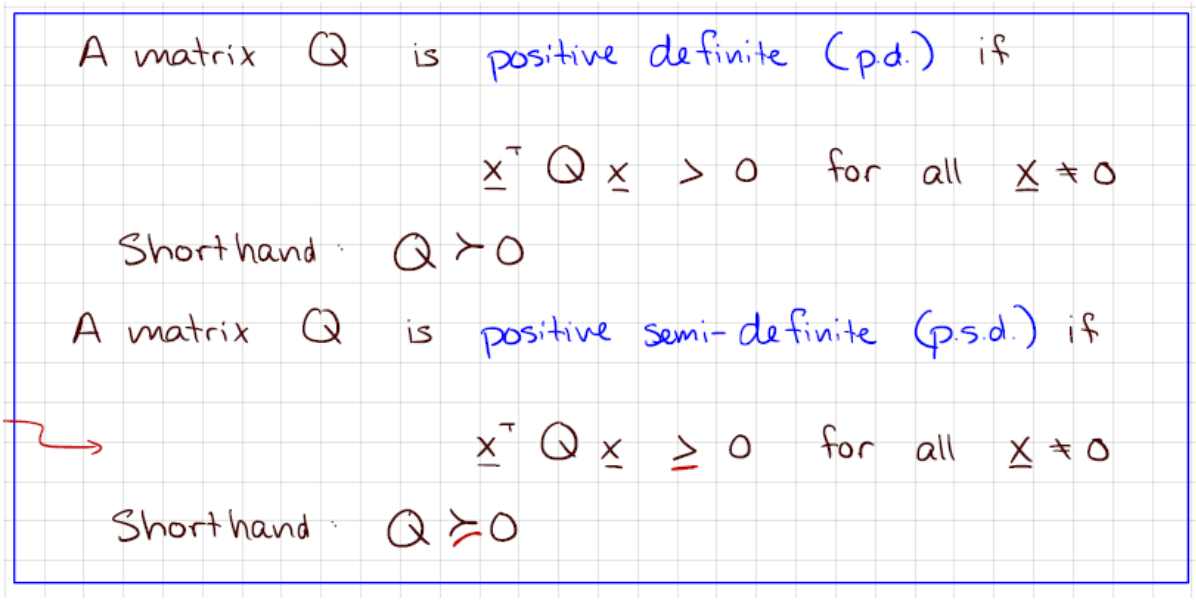
1. $X^TX \in \mathbb{R}^{nxp}$ is invertible (in other words, $(X^TX)^{-1}$ exists)
2. X^TX is positive definite

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

$\underline{x}^TQ\underline{x} > 0$ for all $\underline{x} \neq 0$ (shorthand: $Q \succ 0$) Q curly greater than zero

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

$\underline{x}^TQ\underline{x} \geq 0$ for all $\underline{x} \neq 0$ (shorthand: $Q \succcurlyeq 0$) Q curly greater than or equal to zero



Ex.

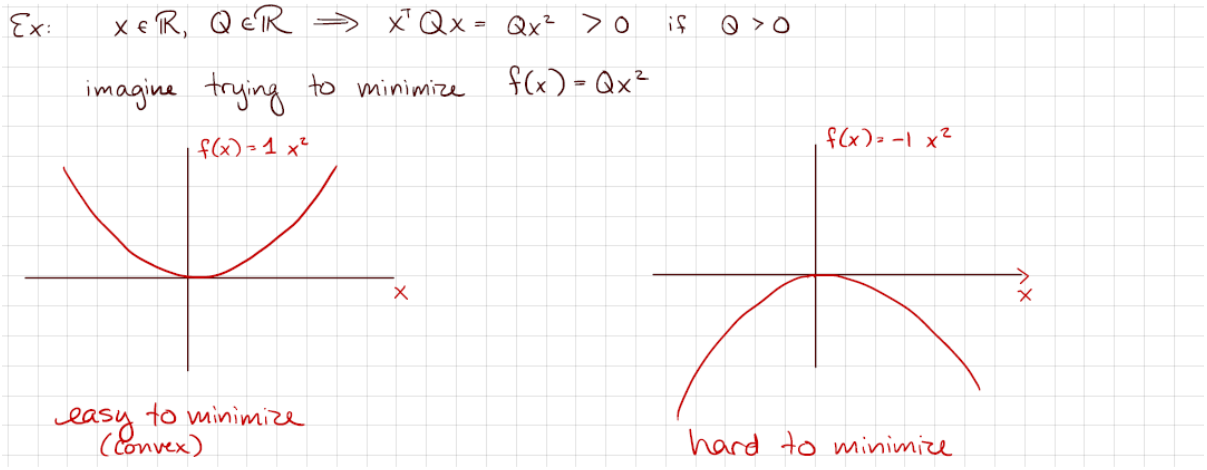
$x \in \mathbb{R}^n,\ Q \in \mathbb{R}^n$

$x^TQx = Qx^2$

$Qx^2 > 0$ if $Q > 0$

Imagine minimizing $f(x)=Qx^2$

convex (easy to minimize) vs. non-convex (hard to minimize)



Now what we wanna do is think about this in 2D instead of 1D, what happens when Q is a matrix instead of a scalar. okay. so for example x now is a length 2 vector and Q is a two by two matrix.

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

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

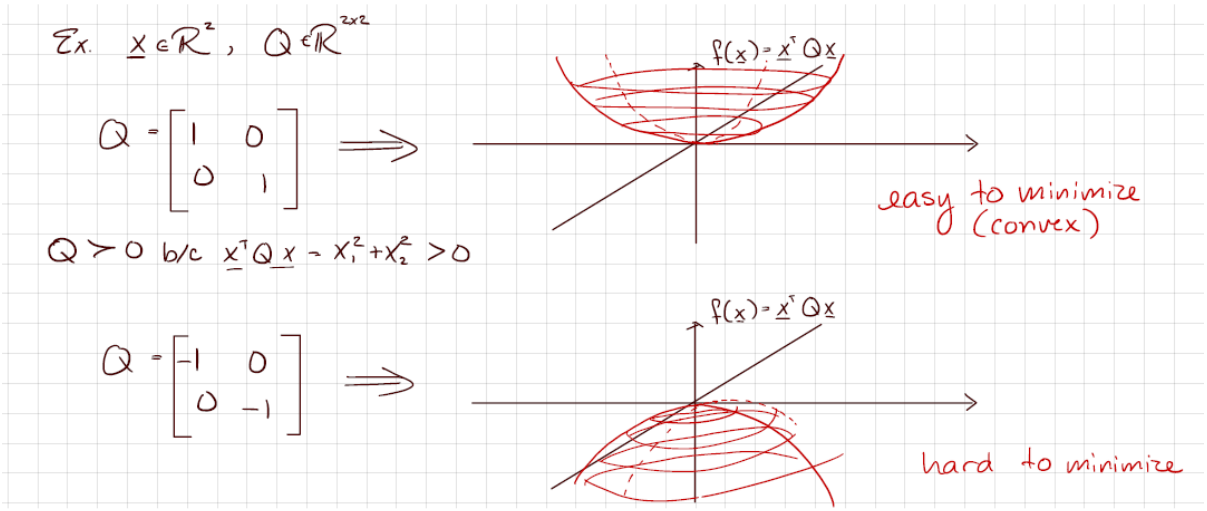
f(x) = $\underline{x}^T Q \underline{x}$

f(x) = $x_1^2 + x_2^2 > 0$ for all $\underline{x} \neq 0$

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

f(x) = $\underline{x}^T Q \underline{x}$

f(x) = $-x_1^2 - x_2^2 < 0$ for all $\underline{x} \neq 0$



Properties of Positive Definite Matrices

1. If $P \succ 0$ and $Q \succ 0$, then $P+Q \succ 0$

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

2. If $Q \succ 0$ and $a>0$, then $aQ \succ 0$

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

3. For any A, $A^TA \succcurlyeq 0$ and $A^TA \succcurlyeq 0$

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

   Let $\underline{\tilde{x}} = A\underline{x}$
   $\underline{x}^T A^T A \underline{x} = \underline{\tilde{x}}^T \underline{\tilde{x}} \geq 0$
   now $\underline{\tilde{x}}^T \underline{\tilde{x}} = 0$ only if $\underline{\tilde{x}} = \mathrm{Ax} = 0$. Ax=0 if either x=0 or columns of A are linearly independent

4. If $A \succcurlyeq 0$, then A^{-1} exists
5. $Q \succcurlyeq P$ means $Q - P \succcurlyeq 0$

Recall Optimization Formulation

$$\underline{\hat{w}} = \underset{\underline{w}}{\operatorname{argmin}}\, \underline{y}^T \underline{y} - 2\underline{w}^T X^T y + \underline{w}^T X^T X \underline{w}$$

We try to solve this optimization problem.

Hint: property 3 of positive definite matrices.

We assume

$X^TX \succ 0 \rightarrow f(w) = \underline{y}^T \underline{y} - 2\underline{w}^T X^T y + \underline{w}^T X^T X w$

→ is convex

→ compute "derivative" & set to zero to find minimizer

Only we're not really dealing with scalars anymore.

We have vectors. So instead of derivative, use gradients

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 $\nabla_w f$ to zero and solve for $\underline{w}$.

$$\nabla_w f = \begin{bmatrix} \frac{df}{dw_1} \\ \frac{df}{dw_2} \\ \vdots \\ \frac{df}{dw_p} \end{bmatrix}$$

Ex 1.

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

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

Ex 2.

$$\mathrm{f}(\underline{\mathrm{w}}) = \underline{w}^T \underline{w} = \|\underline{w}\|_2^2 = w_1^2 + w_2^2 \ \ldots + \ w_p^2$$

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

Ex 3.

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

$$\frac{d(w_i Q_{ij} w_j)}{dw_k} = \begin{Bmatrix} 2Q_{ii} w_i & if \ \ k = i = j \\ Q_{ij} w_j & if \ \ i = k \neq j \\ Q_{ii} w_i & if \ \ i \neq k = j \\ 0 & if \ \ k \neq i, k \neq j \end{Bmatrix}$$

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

$$\nabla_w f = Qw + Q^T w$$

If Q is symmetric (i.e., $Q=Q^T$), then

$$\nabla_w f = 2Qw$$

Ex. Least Squares

$f(\underline{w}) = \underline{y}^T\underline{y} - 2\underline{w}^T X^T y + \underline{w}^T X^T X\underline{w}$

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

➔ $\underline{\widehat{w}} = (X^T X)^{-1} X^T\underline{y}$

What we got from geometric perspective!

(based on Ex 2. c = $-2X^T y$ & Ex 3. $Q = X^T X = Q^T$)