$$\hat{w}_{ML} = \underbrace{(XX^T)^{-1}X}\,y$$

$y \longrightarrow$ Random part

$\longrightarrow$ Linear transformations.

$$\mathbb{E}[\hat{w}_{ML}] = \mathbb{E}\left[(XX^T)^{-1}Xy\right]$$

$$= (XX^T)^{-1}X\,\mathbb{E}[y].$$

$$= \underline{(XX^T)^{-1}}\,\underline{X(X^T w)}$$

$$= w$$

$\Rightarrow \hat{w}_{ML}$ is unbiased!

$y_i := w^T x_i + \epsilon$

$y_i = \mathcal{N}(w^T x_i, \sigma^2)$

$\mathbb{E}[y_i] = w^T x_i$

$\mathbb{E}[y] = X^T w$

What is variance of $\hat{w}_{ML}$?

$$\hat{w}_{ML} = (XX^T)^{-1} X y = A y \qquad \left[A = (XX^T)^{-1} X \right]$$

$$\text{covar}(\hat{w}_{ML}) = \mathbb{E}\left[(\hat{w}_{ML} - \underbrace{\mathbb{E}[\hat{w}_{ML}]}) (\hat{w}_{ML} - \mathbb{E}[\hat{w}_{ML}])^T \right]$$

$$= \mathbb{E}\left[(Ay - w)(Ay - w)^T \right]$$

$$= \mathbb{E}\left[Ayy^TA^T - Ayw^T - wy^TA^T + ww^T \right]$$

on simplification [Please do This]

$$\text{Covar}(\hat{w}_{ML}) = \sigma^2 (XX^T)^{-1}$$

$$\underbrace{\mathbb{E}\left(\|\hat{w}_{ML} - w\|^2\right)} = \mathbb{E}\Big(\underbrace{(\overbrace{\hat{w}_{ML} - w}^{a})^T (\hat{w}_{ML} - w)}\Big)$$

$$= \mathbb{E}\left(\text{trace}\left((\hat{w}_{ML} - w)(\hat{w}_{ML} - w)^T\right)\right)$$

$$= \text{trace}\left(\mathbb{E}\Big[\underbrace{(\hat{w}_{ML} - w)(\hat{w}_{ML} - w)^T}\Big]\right)$$

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

M.S.E.

$$aa^T = \begin{bmatrix} a_1^2 & a_1a_2 & a_1a_1 \\ & a_2^2 & \\ & & \ddots \\ & & & a_k^2 \end{bmatrix}$$

a^Ta

$\text{trace}(aa^T) = a^Ta$. $\text{trace}(aa^T)$

$$= \operatorname{trace}\left(\operatorname{cov}(\hat{w}_{ML}) \right)$$

$$= \operatorname{trace}\left(\sigma^2 (XX^T)^{-1} \right)$$

$$= \sigma^2 \operatorname{trace}(XX^T)^{-1}$$

Let the eigenvalues of XX^T be $\{\lambda_1, \ldots, \lambda_d\}$

Eigenvalues of $(XX^T)^{-1}$ $= \left\{ \frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_d} \right\}$

3)

Mean squared error

$$\mathbb{E}\left(\|\hat{w}_{ML} - w\|^2\right) = \sigma^2 \underbrace{\left(\sum_{i=1}^{d} \frac{1}{\lambda_i}\right)}$$

$$\hat{w}_{ML} = (XX^T)^{-1} Xy$$

Consider the following modified estimator

$$\hat{w}_{\text{modified}} = (XX^T + \lambda I)^{-1} Xy$$

- For some matrix A, say e.val are $\{\lambda_1, \ldots, \lambda_d\}$

$$Au_k = \lambda_k u_k \quad \text{for some } u_k \in \mathbb{R}^d$$

- What are e.values of $A + \lambda I$

$$(A + \lambda I) u_k = Au_k + \lambda u_k$$

$$= \lambda_k u_k + \lambda u_k$$

$$= (\lambda_k + \lambda) u_k$$

EXISTENCE THEOREM (INFORMAL)

$\exists \lambda \in \mathbb{R}_+$ s.t

$\hat{w}_{modified} = (XX^T + \lambda I)^{-1} Xy$ has lesser

m.s.e than $\hat{w}_{ML}$

Ridge Regression: Biased Estimation for Nonorthogonal Problems

Arthur E. Hoerl & Robert W. Kennard

SO FAR

LINEAR REGRESSION

$y = X^T w + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$

$w^* = (XX^T)^{-1} Xy$;

$\hat{w}_{ML} = (XX^T)^{-1} Xy$

$\mathbb{E}[\hat{w}_{ML}] = w$

$\mathrm{Cov}(\hat{w}_{ML}) = \sigma^2 (XX^T)^{-1}$

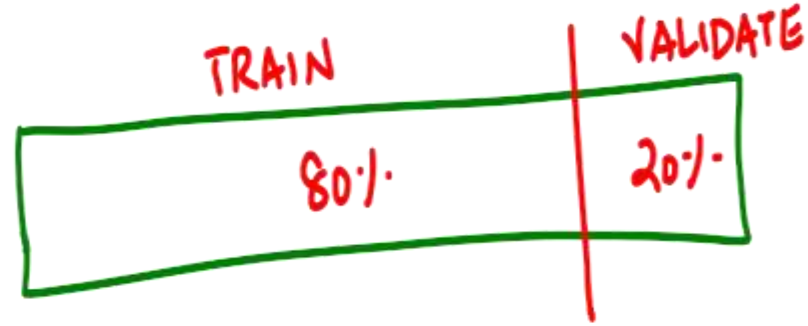
$MSE = \sigma^2 \, \underbrace{\mathrm{trace}\left((XX^T)^{-1}\right)}$

$\boxed{\hat{w}_{\text{modified}}} = \underbrace{(XX^T + \lambda I)^{-1} Xy}$

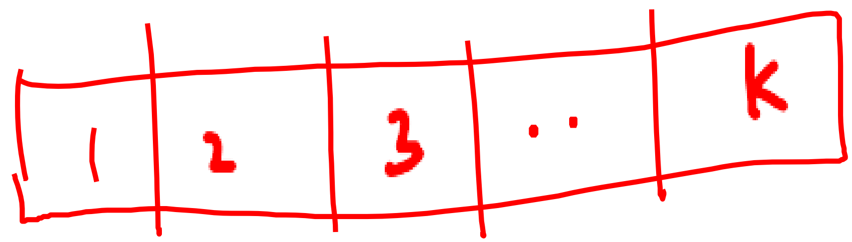
$\exists \lambda$ s.t $\hat{w}_{mod}$ has smaller m.s.e than $\hat{w}_{ML}$

$\underbrace{\mathbb{E}\left[\|\hat{w}_{ML} - w\|^2\right]}$

CROSS-VALIDATION



K-fold Cross Validation



Leave one out cross validation – K = n

BAYESIAN MODELING

need

- PRIOR on parameter
- i.e., $P(\underline{w})$ $w \in \mathbb{R}^d$

LIKELIHOOD

$$y|x \sim N(w^T x, 1)$$

For simplicity. Can use σ^2 as well

PRIOR

$$w \sim \mathcal{N}(\underbrace{0}_{\in \mathbb{R}^d},\ \underbrace{\nu^2 I}_{\in \mathbb{R}^{d \times d}}) \longrightarrow \begin{bmatrix} \nu^2 & & \\ & \nu^2 & \\ & & \ddots \ \nu^2 \end{bmatrix}$$

As usual

$$\underbrace{P\left(w \,\middle|\, \{(x_1,y_1)\ldots,(x_n,y_n)\}\right)}_{\text{Posterior}} \propto \underbrace{P\left(\{(x_1,y_1)\ldots(x_n,y_n)\} \,\middle|\, w\right)}_{\text{Likelihood}} \cdot \underbrace{P(w)}_{\text{Prior}}$$

$$\propto \left(\prod_{i=1}^{n} e^{-\frac{(y_i - w^T x_i)^2}{2}}\right) \cdot \left(\prod_{i=1}^{d} e^{-\frac{w_i^2}{2\gamma^2}}\right)$$

$$\propto \left(e^{-\sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2}}\right) \; e^{-\underbrace{\left(\sum_{i=1}^{d} w_i^2\right)}_{\|w\|^2} \cdot \frac{1}{2\gamma^2}}$$

$$\log\left(P(w \mid \text{Data})\right) \equiv - \sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2} - \frac{1}{2\gamma^2} \|w\|^2$$

MAP ESTIMATE

$$\hat{w}_{MAP} = \arg\min_{w} \underbrace{\overbrace{\sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2}}^{\frac{1}{2}\|X^T w - y\|^2} + \overbrace{\frac{1}{2\gamma^2}\|w\|^2}^{\frac{1}{2\gamma^2} w^T w}}_{\longrightarrow\, f(w)}$$

$$\boxed{\hat{\omega}_{MAP} = \left(XX^T + \frac{1}{\gamma^2} I\right)^{-1} Xy}$$

CROSS VALIDATE

CONCLUSION

- MAP estimation for lin. reg with a Gaussian prior $\mathcal{N}(0, \gamma^2 I)$ for ω is "equivalent" to $\hat{\omega}_{modified}$ estimator that we used earlier.

$$\hat{w}_R \rightarrow \arg\min_{w} \underbrace{\sum_{i=1}^{n} (w^T x_i - y_i)^2}_{\text{LOSS}} + \overset{\text{HYPER PARAMETER}}{\lambda} \underbrace{\|w\|^2}_{\text{REGULARIZATION}}$$

$$\|w\|^2 \leftarrow \left(\sum_{i=1}^{d} w_i^2\right)$$

$$\lambda \leftarrow \approx \frac{1}{2\gamma^2}$$

RIDGE REGRESSION

LINEAR REGRESSION

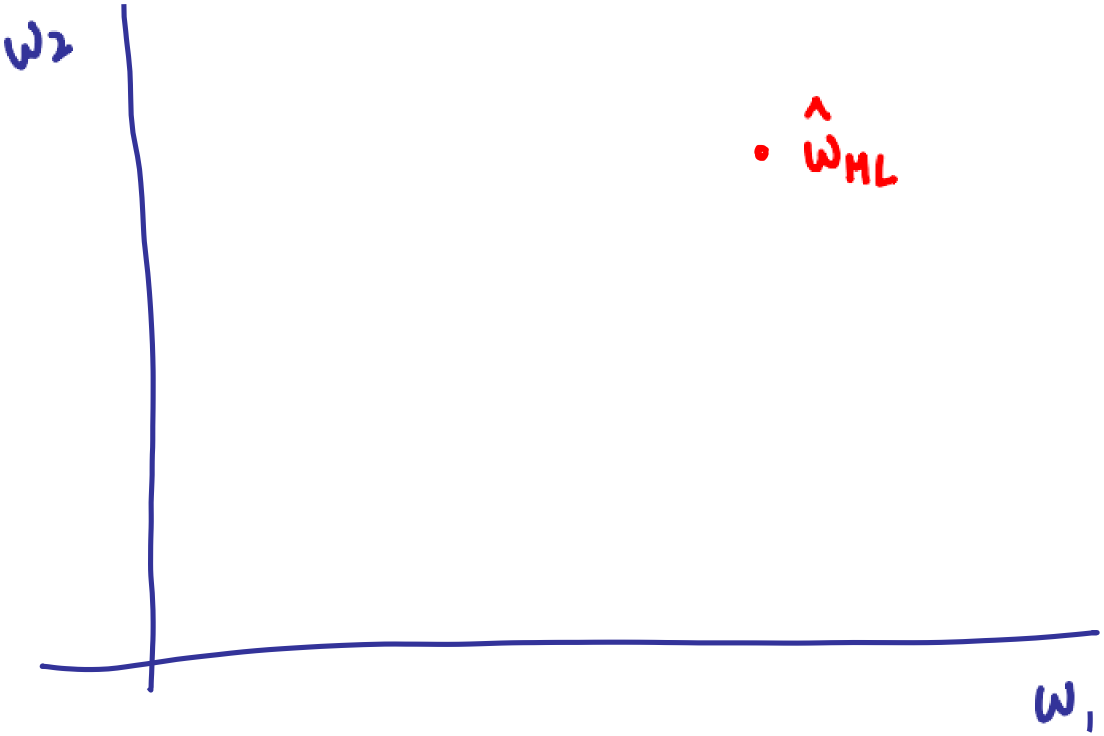
$$\hat{w}_{ML} = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

RIDGE REGRESSION

HYPER PARAMETER

$$\hat{w}_R = \arg\min_{w \in \mathbb{R}^d} \underbrace{\sum_{i=1}^{n} (w^T x_i - y_i)^2}_{\text{LOSS}} + \underbrace{\lambda \|w\|^2}_{\text{REGULARIZATION}}$$

LOSS + REGULARIZATION



$$\min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2 \rightarrow A$$

$$\equiv$$

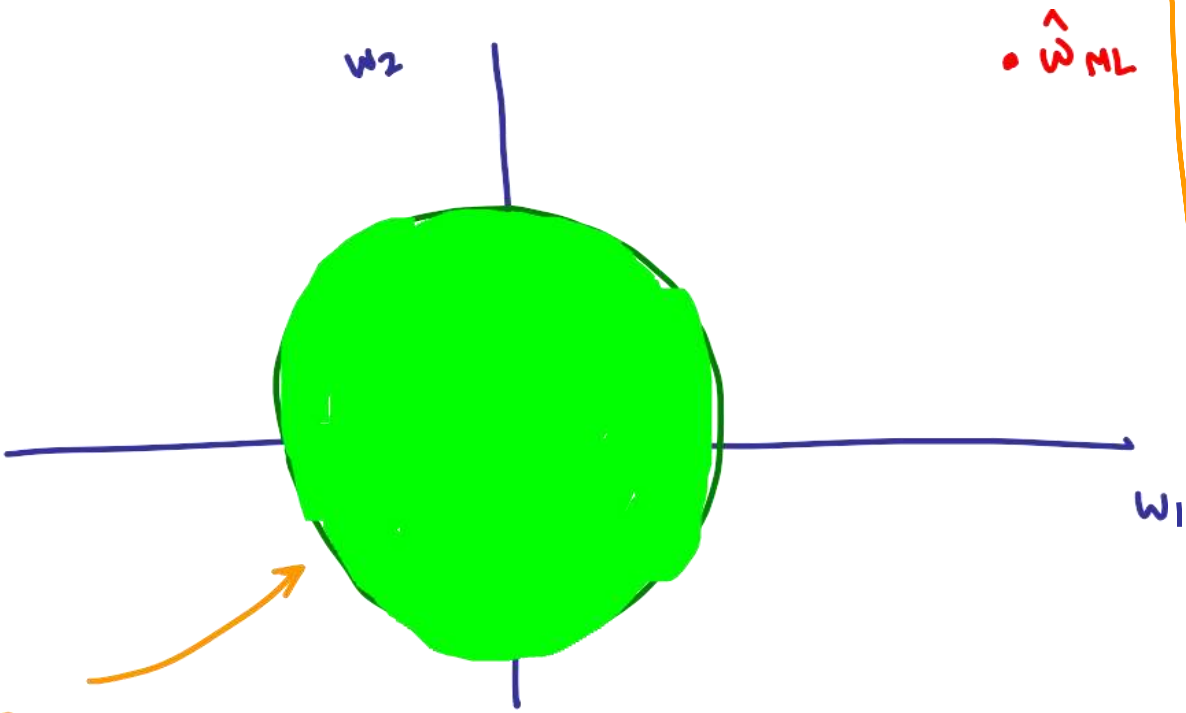
$$\min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \rightarrow B$$

$$\text{s.t } \|w\|^2 \leq \theta$$

depends on λ

For every choice of λ, $\exists\, \theta$ (depending on λ) s.t

A and B give the same solution.



$\|w\|^2 \leq \theta$

$w_1^2 + w_2^2 \leq \theta$

What is the objective value of linear reg at $\hat{w}_{ML}$?

$$\sum_{i=1}^{n} \left(\hat{w}_{ML}^T x_i - y_i\right)^2$$

$$= f\left(\hat{w}_{ML}\right)$$

Consider the following set

$$S_c = \left\{ w \in \mathbb{R}^2 : \quad f(w) = f(\hat{w}_{ML}) + c \right.$$

$c \in \mathbb{R}_+$

$\forall\, w \in S_c$

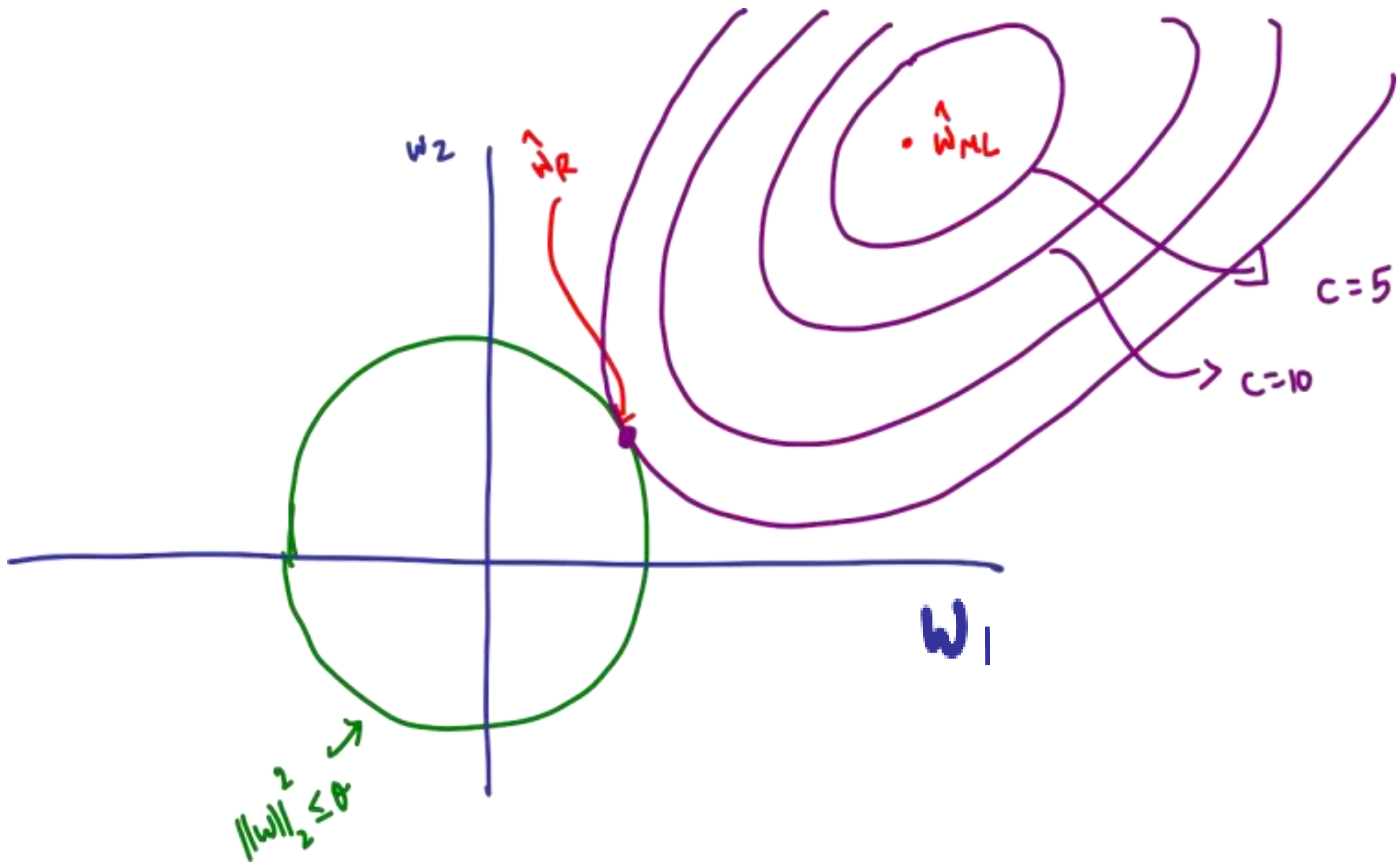
$$\sum_{i=1}^{n} (w^T x_i - y_i)^2 = \sum_{i=1}^{n} (\hat{w}_{ML}^T x_i - y_i)^2 + c$$

$$\equiv \quad \underbrace{\| X^T w - y \|^2} = \| X^T \hat{w}_{ML} - y \|^2 + c$$

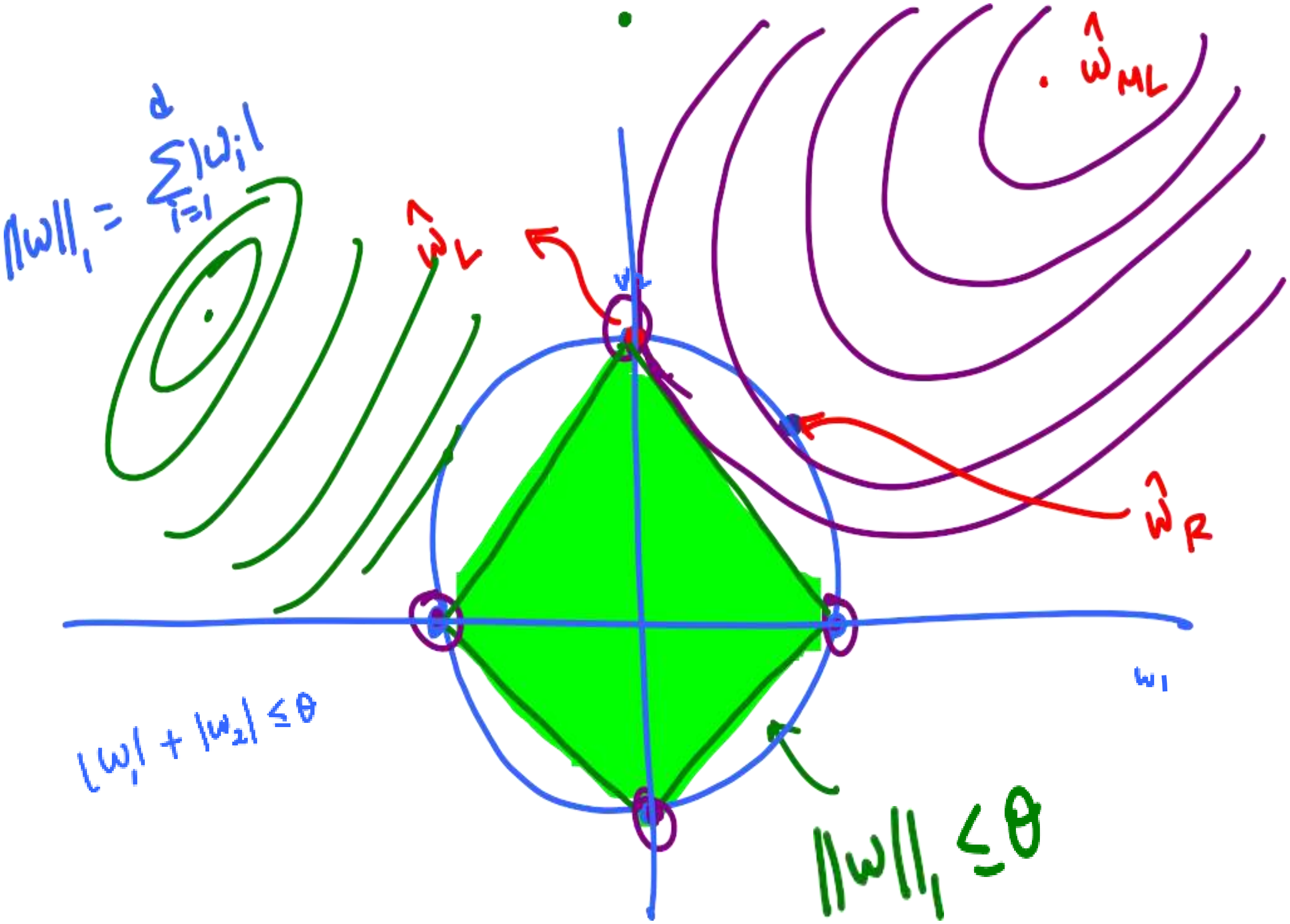
on simplification [do this]

$$\Rightarrow \quad \underline{(w - \hat{w}_{ML})^T \underline{(XX^T)} (w - \hat{w}_{ML})} = C'$$

depends on C, (XX^T), $\hat{w}_{ML}$ but does not depend on w



- Ridge regression pushes feature weights towards 0 but does not necessarily make it 0.



$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

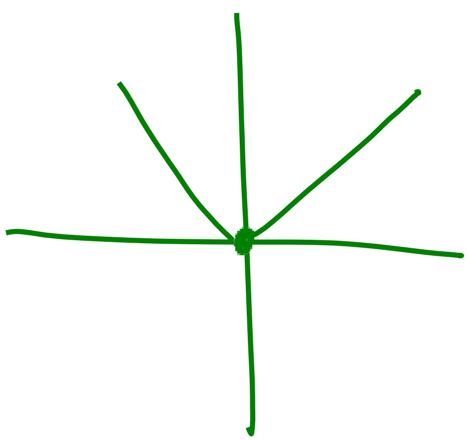
s.t.

$$\|w\|_1 \leq \theta$$

LASSO

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_1$$

LEAST ABSOLUTE SHRINKAGE & SELECTION OPERATOR

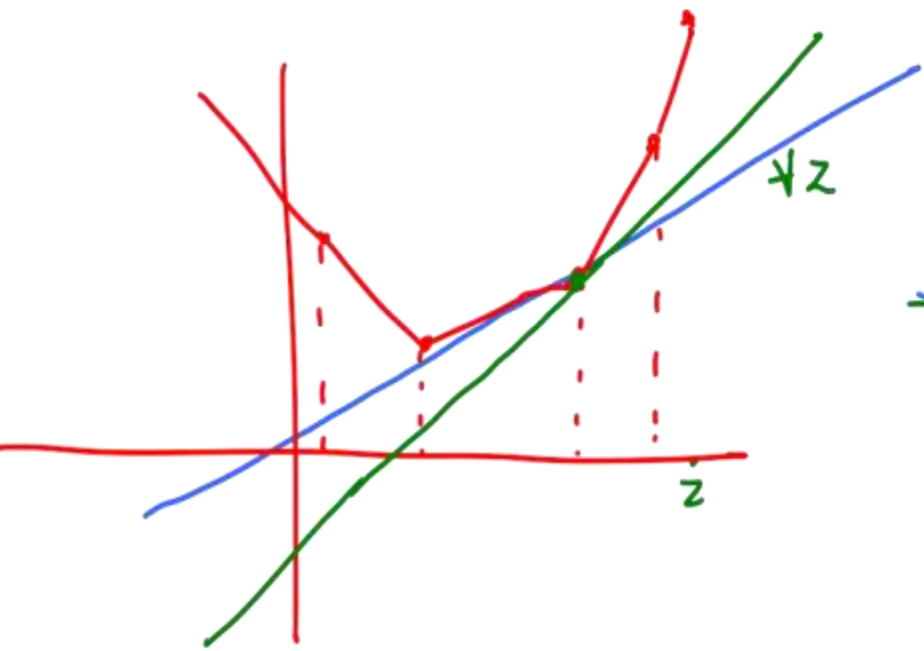


- How to solve this?

↳ Sub-gradient descent

Sub-gradient

A vector $g \in \mathbb{R}^d$ is a sub-grad of $f : \mathbb{R}^d \to \mathbb{R}$ at x if



$$f(z) \geq f(x) + g^T (z - x) \quad \forall z$$

If f is convex, sub-grad descent converges!

MODELING NON LINEAR RELATIONSHIPS

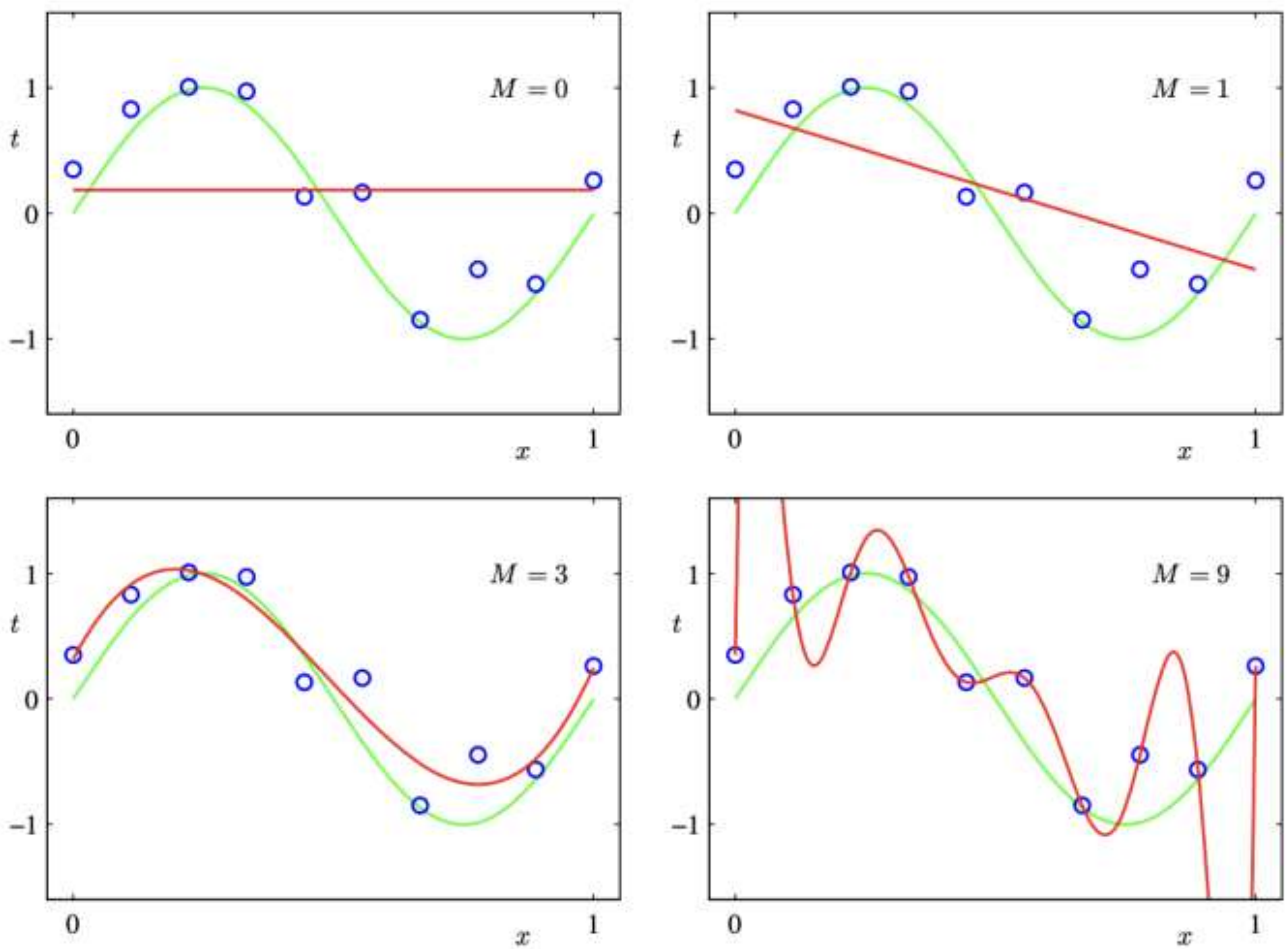
Key idea: Map features to higher dimension.

$$x = [x_1, x_2] \rightarrow . \phi(x) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]$$

In the above example, learning a linear function in the higher dimension is equivalent to learning a quadratic function in the lower dimension.

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} (w^T \phi(x_i) - y_i)^2$$

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



Coefficient values increases as M increases!

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Regularization helps control the co-efficient values

	$\ln\lambda=-\infty$	$\ln\lambda=-18$	$\ln\lambda=0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

$$x = [f_1 \quad f_2 \quad f_3 \quad f_4]$$

$\Downarrow$ Cubic relation

$$\phi(x) = [1 \quad \underbrace{f_1 \quad f_2 \quad f_3 \quad f_4}$$

$$1 + 4 + 4C^2 + 4C^3$$

$x \in d$

$\phi(x) \in O(d^p)$

ISSUE.

$\phi(x) \in \mathbb{R}^D \rightarrow$ might be too large

EXAMPLE

$$x = \begin{bmatrix} f_1 & f_2 \end{bmatrix} \qquad x' = \begin{bmatrix} g_1 & g_2 \end{bmatrix}$$

$$\boxed{(x^T x' + 1)^2} = \left(\begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + 1 \right)^2$$

$$= (f_1 g_1 + f_2 g_2 + 1)^2$$

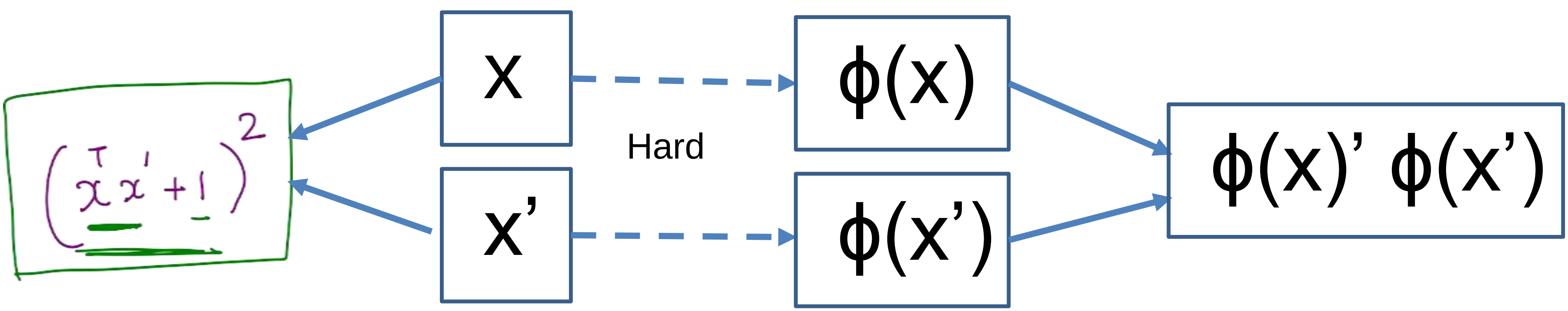
$$= f_1^2 g_1^2 + f_2^2 g_2^2 + 1 + 2 f_1 g_1 f_2 g_2$$

$$+ 2 f_1 g_1 + 2 f_2 g_2$$

$$\boxed{\left(x^T x' + 1\right)^2}$$

$$= \begin{bmatrix} f_1^2 & f_2^2 & 1 & \sqrt{2} f_1 f_2 & \sqrt{2} f_1 & \sqrt{2} f_2 \end{bmatrix} \begin{bmatrix} g_1^2 \\ g_2^2 \\ 1 \\ \sqrt{2} g_1 g_2 \\ \sqrt{2} g_1 \\ \sqrt{2} g_2 \end{bmatrix}$$

$$\phi(x)^T \phi(x')$$



Summary/Questions

-> To capture non-linear relationships, one can "create" non-linear functions of features.

-> But the number of features to create grows exponentially with the degree of non-linearity p that we wish to capture (d^p)

-> For d = 2 and p = 2, it appears there is a trick to get around this.

-> Is this trick general enough to be useful the general case as well? (i.e., for any d and any p?)

MORE EXAMPLES

Polynomial map

$$K(x, x') = (x^T x' + 1)^p \quad \text{for some } p \geq 1$$

→ Can be shown to be a "valid" function.

i.e., $\exists \, \phi \quad \mathbb{R}^d \to \mathbb{R}^D$ such that

$$K(x, x') = \phi(x)^T \phi(x')$$

EXERCISE

Compute ϕ for $p = 3$, $p = 4$.

MORE EXAMPLES.

$$K(\underset{\in \mathbb{R}^d}{x}, \underset{\in \mathbb{R}^d}{x'}) = \exp\left(\frac{-\|x - x'\|^2}{2\sigma^2} \right)$$

for some $\sigma > 0$

RADIAL BASIS FUNCTION

→ Can also shown to be a "valid" map

[Technicalities aside think of $\phi(x)$ as a function and dot product as integrals]

→ Interestingly, ϕ in this case maps x to an "infinite" dimensional space

KERNEL FUNCTION

Any function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ which is a "valid" map is a kernel function

$$K(x, x') = (x^T x' + 1)^p \rightarrow \text{Polynomial kernel}$$

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \rightarrow \text{Gaussian kernel / Radial basis kernel}$$

Question: Given a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, how can we say it's a valid kernel?

METHOD 1: Explicitly construct a ϕ map

[might be hard sometimes]

METHOD 2 : MERCER's THEOREM

A function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a valid kernel

if and only if

(a) k is symmetric.

(b) For any dataset $\{x_1, \ldots, x_n\}$, the

matrix $K \in \mathbb{R}^{n \times n}$

$K_{ij} = k(x_i, x_j)$ is POSITIVE SEMI DEFINITE

eigenvalues of K are all non-negative.

How does kernel help in solving the non-linear regression problem?

$x_1, \dots x_n$
$y_1 \dots y_n$

$\longrightarrow \hat{w}_{ML}$

test point

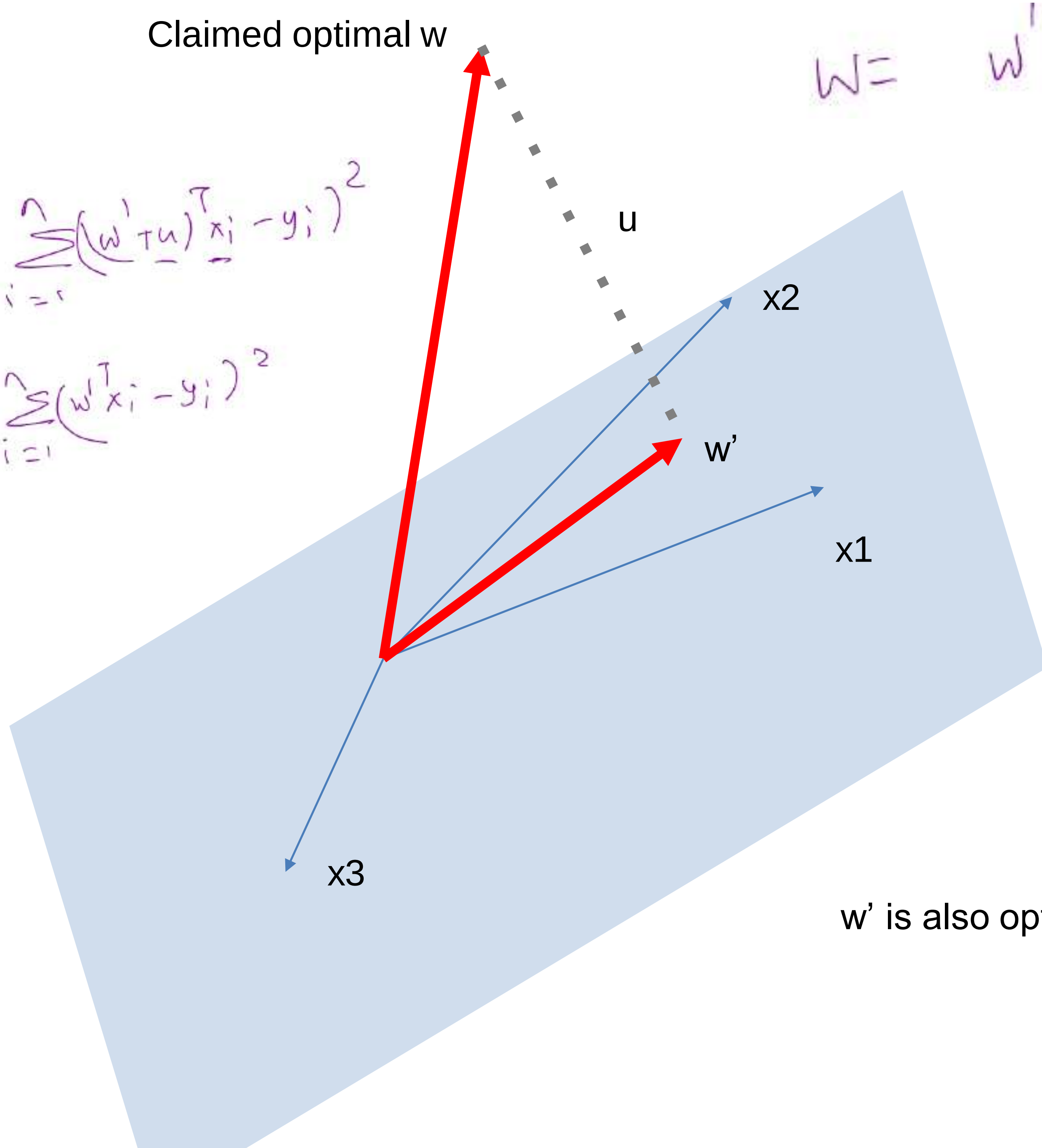
$t \in \mathbb{R}^d$.

$$\boxed{\hat{y} = \hat{w}_{ML}^T t}$$

$$w = w' + u'$$

$$\sum_{i=1}^{n}(w^T x_i - y_i)^2 = \sum_{i=1}^{n}((w'+u)^T x_i - y_i)^2$$

$$= \sum_{i=1}^{n}(w'^T x_i - y_i)^2$$



w’ is also optimal

Note Title 10 Sep 22

$$\hat{w}_{ML} = (XX^T)^{-1}Xy$$

$$\hat{w}_{ML} = X\alpha$$

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

$$X^TXX^TX\alpha = X^TXX^T(XX^T)^{-1}Xy$$

$$(X^TX)^2\alpha = X^TXy$$

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

KERNEL

Kernel- regression ⇐

given a kernel $K \in \mathbb{R}^{n \times n}$,

obtain $\alpha = K^{-1}y \Leftarrow \in \mathbb{R}^n$

For a new point t

$$\hat{y} = \sum_{i=1}^{n} \alpha_i \underbrace{K(t, x_i)}$$