Regression from finite data:

$\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathcal{X} \times \mathbb{R}$

drawn i.i.d from $D$

$f: \mathcal{X} \Rightarrow \mathbb{R}$

$R[f] = \mathbb{E}_{x,y}[(f(x) - y)^2]$

The problem is also called "least squares regression" to emphasise the risk function used.

Example: If $\mathcal{X} = \mathbb{R}^d$, $f$ is a function from $\mathbb{R}^d$ to $\mathbb{R}$

---

ERM - Framework and motivation.

Step 1

$R[f] \approx \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 = \hat{R}[f]$

$R[f]$: True risk

$\hat{R}[f]$: Empirical risk

Step 2

Choose a set of function $\mathcal{F} \subseteq \{f: \mathcal{X} \to \mathbb{R}\}$

Step 3

Return $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{R}[f]$

(Break ties arbitrarily)

## Main example: Linear regression

$$\mathcal{F} = \{ f_w : \mathbb{R}^d \to \mathbb{R} \;,\; f_w(x) = w^T x \text{ for } w \in \mathbb{R}^d \}$$

(i.e) $\mathcal{F}$ is the set of "linear" functions from $\mathbb{R}^d$ to $\mathbb{R}$.

e.g. if $d=2$ $\mathcal{F}$ contains functions like $x_1 + 3x_2$, $-x_1 + x_2$ etc.

$$\hat{R}[f_w] = \hat{R}(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (w^T x_i - y_i)^2 \qquad w^T x_i = y_i \;\; \forall i \in [n]$$

$$\nabla_w \hat{R}(w) = \frac{1}{n} \sum_{i=1}^{n} 2 (w^T x_i - y_i) x_i$$

---

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & \ddots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

Data matrix

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

Target / label vector.

$w \in \mathbb{R}^d$ — weight vector

$$\hat{R}(w) = \frac{1}{n} \sum_i (x_i^T w - y_i)^2$$

($X$: $n \times d$, $w$: $d \times 1$, $y$: $n \times 1$)

$$= \frac{1}{n} \| Xw - y \|_2^2$$

$$\nabla_w \hat{R}(w) = \frac{1}{n} \sum_{i=1}^{n} 2 (w^T x_i - y_i) x_i$$

$$= \frac{1}{n} \cdot 2 \cdot X^T (Xw - y)$$

($X^T$: $d \times n$, $(Xw - y)$: $n \times 1$)

$$\begin{bmatrix} x_1^T w - y_1 \\ x_2^T w - y_2 \\ \vdots \\ x_n^T w - y_n \end{bmatrix}$$

Setting $\nabla \hat{R}(\hat{w}) = 0$ we get.

$$X^T(X\hat{w} - y) = 0$$

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

$Aw = b$

$d$ equations $d$ unknowns.

Ex. P.T this system of eqns. always has a solution.

If $X^T X$ is invertible then

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

Pseudo Inverse of $X \in \mathbb{R}^{d \times n}$

Exercise: What if $X^T X$ is not invertible?

$\therefore$ The function returned is $\hat{f}(x) = \hat{w}^T x$

---

## Polynomial Regression: $k^{th}$ degree polynomial

Informally: Find the best $k^{th}$-degree polynomial that fits the data.

Spl Case: $d=1$, $k=5$

$$\mathcal{F}_{1,5} = \{ f : \mathbb{R} \to \mathbb{R} : \quad f(z) = w_0 + w_1 z + w_2 z^2 + \cdots + w_5 z^5 \text{ for some } w \in \mathbb{R}^6 \}$$

Spl case $d=2$, $k=3$

$$\mathcal{F}_{2,3} = \{ f : \mathbb{R}^2 \to \mathbb{R} : f(z) = w_0 + w_1 z_1 + w_2 z_2 + w_3 z_1^2 + w_4 z_2^2 + w_5 z_1 z_2 + w_6 z_1^3 + w_7 z_2^3 + w_8 z_1^2 z_2 + w_9 z_1 z_2^2, \text{ for } w \in \mathbb{R}^{10} \}$$

Exercise: Give $\mathcal{F}_{3,2}$, $\mathcal{F}_{3,3}$, and $\mathcal{F}_{4,2}$

Feature Mapping

$\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ converts a low dimensional $x \in \mathbb{R}^d$ to a higher dimensiond $\phi(x) \in \mathbb{R}^{d'}$.

Key observation: We can represent a k-degree polynomial regression on $x \in \mathbb{R}^d$, by a linear regression on a higher dimensional $\phi(x) \in \mathbb{R}^{d'}$

Exercise: Give $d'$ as a function of $d$ and $k$.

$d' =$

$= {}^{d+k}C_k$

---

For example let $d=2$

$$\min_{f \in F_{2,3}} \hat{R}[f] = \min_{f \in F_{2,3}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

$$= \min_{w \in \mathbb{R}^{10}} \frac{1}{n} \sum_{i=1}^{n} (w^T \phi(x_i) - y_i)^2$$

where $\phi(x_i) = [1, x_{i1}, x_{i2}, x_{i1}^2, x_{i2}^2, x_{i1}x_{i2}, x_{i1}^3, x_{i2}^3, x_{i1}^2 x_{i2}, x_{i1} x_{i2}^2]$

$\phi : \mathbb{R}^2 \to \mathbb{R}^{10}$

$\therefore$ A 3-degree polynomial regression problem in $\mathbb{R}^2$, can be reduced to a linear regression problem in $\mathbb{R}^{10}$.

Algorithmically:

The data matrix $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$ is replaced by a feature matrix $\Phi = \begin{bmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix} \in \mathbb{R}^{n \times d'}$.

$\therefore$

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

($d' \times 1$, $d' \times d'$, $d' \times n$, $n \times 1$)

is the solution if $\Phi^T \Phi$ is inv.

o.w: Solve for $w$.

$\Phi^T \Phi w = \Phi^T y$

---

Given a new test instance $x \in \mathbb{R}^d$, it is converted to $\phi(x) \in \mathbb{R}^{d'}$ and the prediction is given by

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

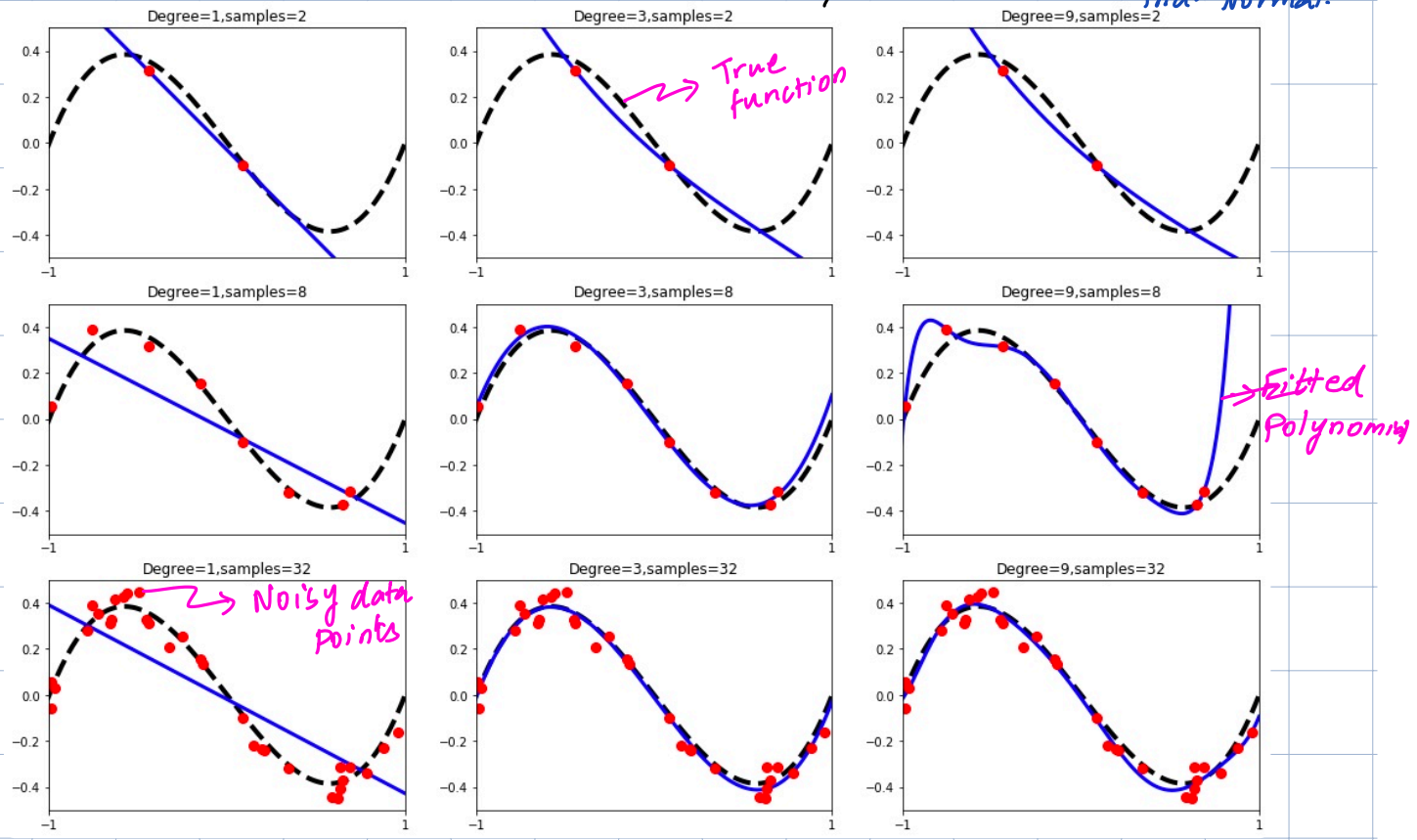
Exercise: Computational complexity of $k$-degree polynomial regression over $\mathbb{R}^d$. (Assume constant time for add, mul, & exponentiation. Assume $n^3$ time for inverting $n \times n$ matrix.)

d = 1

Example Polynomial regression:

$X \sim \text{Unif}[-1, 1]$

$Y = X^3 - X + \epsilon$ → zero mean ind. normal.



# Bias Variance Analysis

$A : S \to f$

Assumption: $y = f^*(x) + e$, $e$ ind. of $x$, zero mean

$S \sim D^m$, $(x, y) \sim D$

An algorithm maps a sample $S \in (X \times \mathbb{R})^m$ to a function $f : X \to \mathbb{R}$ which we will denote by $f_S$.

$$MSE[f] = E_{x,y}(f(x) - y)^2$$

$$\text{or } R[f]$$

$$MSE(\text{Algo}) = E_{S,x,y}\left[(f_S(x) - y)^2\right]$$

$$= E_S\left[E_{x,e}\left[(f_S(x) - f^*(x) - e)^2\right]\right]$$

$$= E_S\, E_{x,e}\left[(f_S(x) - f^*(x))^2 + e^2 - 2e\,(f_S(x) - f^*(x))\right]$$

---

$$= E_S\, E_x\left[(f_S(x) - f^*(x))^2\right] + E_e[e^2] - \underbrace{2E\left[e\,(f_S(x) - f^*(x))\right]}_{=\,0}$$

Let $g(x) = E_{S'}[f_{S'}(x)]$

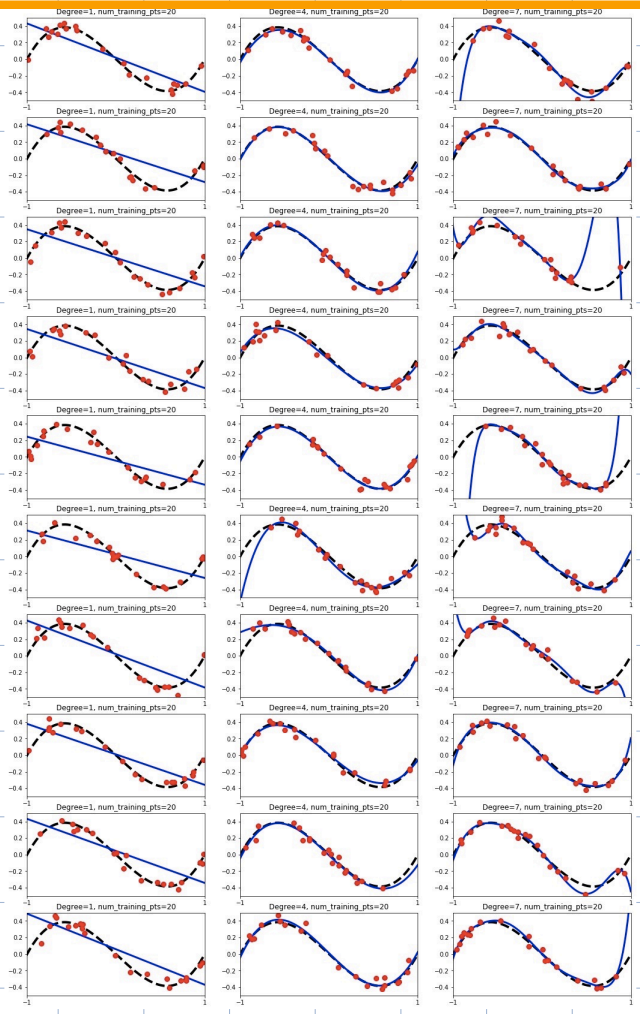
$$MSE(\text{Algo}) = E_S\, E_x\left[(f_S(x) - g(x) + g(x) - f^*(x))^2\right] + E[e^2]$$

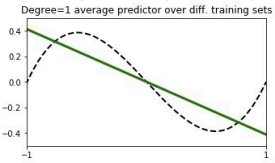
$$= \underbrace{E_S E_x\left[(f_S(x) - g(x))^2\right]}_{\text{Variance}} + \underbrace{E_x\left[(g(x) - f^*(x))^2\right]}_{\text{Bias squared}} + \underbrace{E[e^2]}_{\text{Noise}}$$

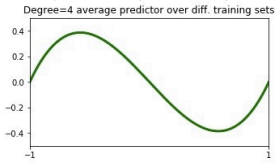
$$+ 2E_S E_x\left[(f_S(x) - g(x))(g(x) - f^*(x))\right]$$

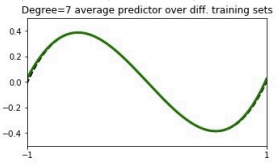
Show that this $= 0$

---



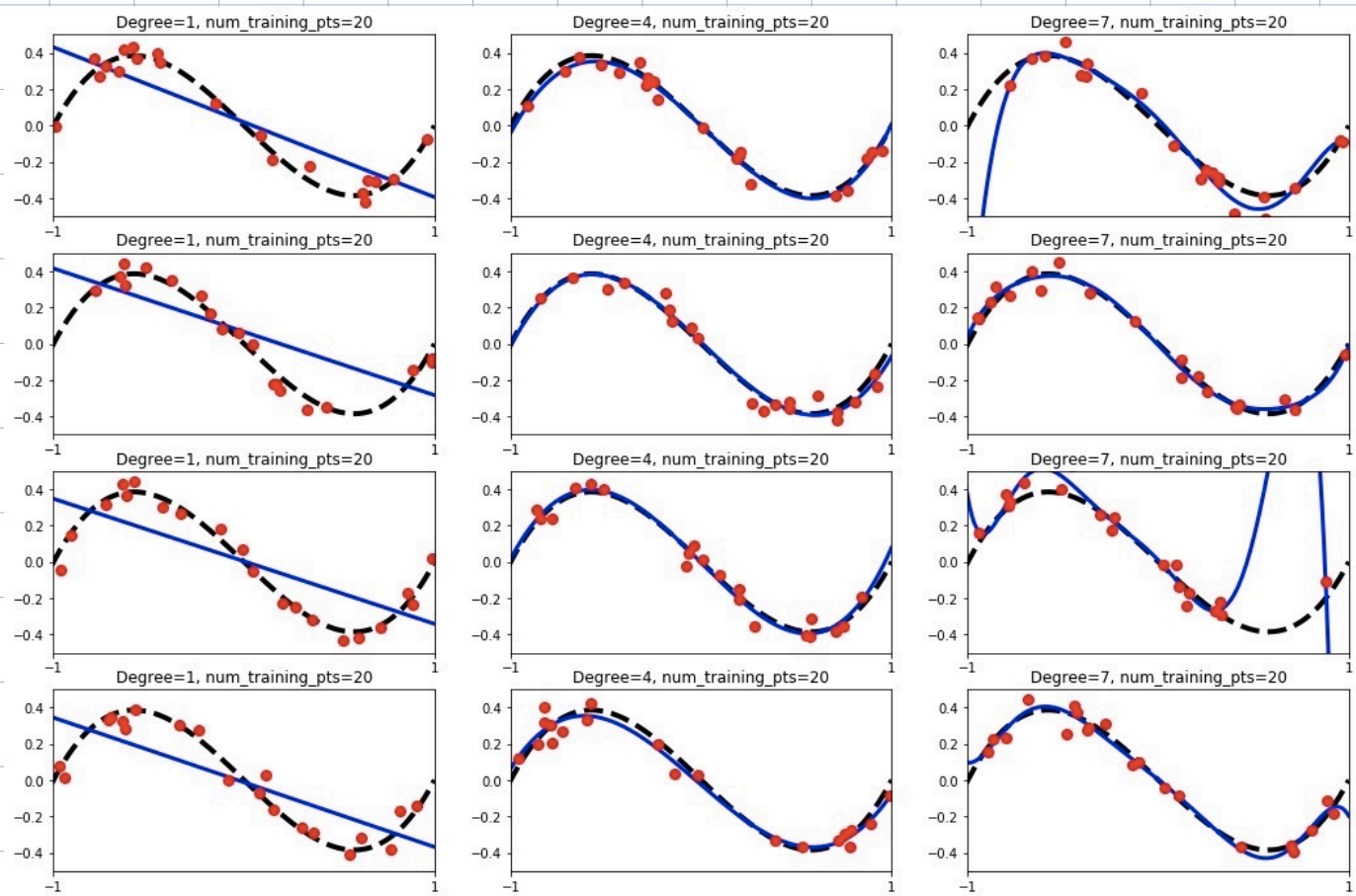



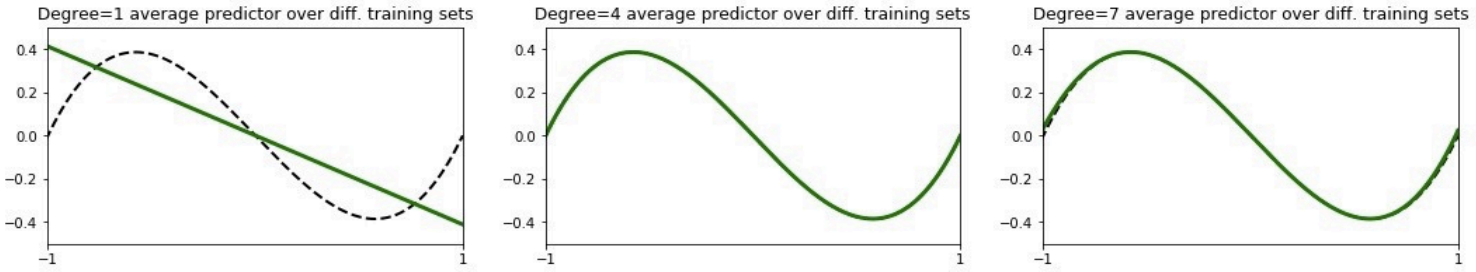




Overfit model → High variance alg

Underfit → High bias



# $g(x)$ Plot for various $f_s$




$f_s^1$ → Returns the best fitting 1st degree polynomial with bias $b_1$, variance $v_1$ and noise $\sigma_1$

$f_s^4$ → Returns the best fitting 4th degree polynomial with bias $b_4$, variance $v_4$ and noise $\sigma_4$

$f_s^7$ → Returns the best fitting 7th degree polynomial with bias $b_7$, variance $v_7$ and noise $\sigma_7$

Arrange $b_1, b_4, b_7$ in ascending order.

Arrange $v_1, v_4, v_7$ in ascending order.

Disc. model for reg.

Let $(X, Y)$ ~ $D$ over $\mathcal{X} \times \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^d$

Let $Y_i = w^T X_i + \epsilon_i$ — $w$ is a fixed but unknown constant here.

where $w$ is some vector in $\mathbb{R}^d$.

$\epsilon_i \sim N(0, \sigma_i^2)$ is independent of $X$

Question: Given $(x_1, y_1), \ldots (x_m, y_m)$ drawn i.i.d from $D$. Can $w$ be identified?

Assumption $Y | X = x \sim N(w^T x, \sigma^2)$

$$P(x_1, y_1, x_2, y_2 \cdots x_m, y_m | w)$$

---

$$P(y_1 \cdots y_m | x_1 \cdots x_m, w) = P(y_1 | x_1, w) \cdot P(y_2 | x_2, w) \cdots P(y_m | x_m, w)$$

$$P(Y_i = y_i | X_i = x_i, w) = P(\epsilon_i = y_i - w^T x_i | X = x_i, w)$$

$$= P(\epsilon_i = y_i - w^T x_i)$$

$$= \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - w^T x_i)^2\right)$$

$$Y | X = x \sim N(w^T x, \sigma^2)$$

$$P(y_1 \cdots y_m | x_1 \cdots x_m, w) = \prod_{i=1}^{m} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2} \epsilon_i^2\right)$$

$$\mathcal{L}(w) = \prod_{i=1}^{m} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma_i^2} (y_i - w^T x_i)^2\right)$$

$$\mathcal{LL}(w) = \sum_{i=1}^{m} \frac{-1}{2\sigma_i^2} (y_i - w^T x_i)^2 + \text{const.}$$

$$-\mathcal{LL}(w) = C \sum_{i=1}^{m} \frac{(w^T x_i - y_i)^2}{\sigma_i^2} = C \hat{R}(w)$$

$\therefore$ minimisation of $\hat{R}$ can also be motivated via Max. Likelihood

$$= \left(w^T \frac{x_i}{\sigma_i} - \frac{y_i}{\sigma_i}\right)^2$$

# Maximum Aposteriori Estimation (MAP)

captures prior knowledge about $w$

Assumption: $w \sim N([0], \tau^2 I_d)$ } Independent.

same as before:
$x \sim D_x$,
$\epsilon \sim N(0, \sigma^2)$
$y = w^T x + \epsilon$

$N\left(\begin{bmatrix} 100 \\ 200 \end{bmatrix}, \begin{bmatrix} 50 & 0 \\ 0 & 70 \end{bmatrix}\right)$

Question: Given $(x_1, y_1), \ldots (x_m, y_m)$ drawn i.i.d from $D$. Can $w$ be identified?

$$P(y_{1:m} | x_{1:m}, w) = \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} \epsilon_i^2\right)$$

$\epsilon_i^2 \rightarrow y_i - w^T x_i$

$$P(w | x_{1:m}, y_{1:m}) = \frac{P(y_{1:m} | x_{1:m}, w)\ P(w | x_{1:m})}{P(y_{1:m} | x_{1:m})}$$

$P(w | x_{1:m}, y_{1:m}) \propto P(y_{1:m} | x_{1:m}, w)\ P(w | x_{1:m})$

$\propto P(w)$

Posterior probability of $w$ given

$\begin{bmatrix} w \\ y \end{bmatrix} \in \mathbb{R}^{d+m}$

$w \sim N(\ )$

$y | w \sim N(wx, \square)$ $\Rightarrow$ $w | y$ is also Normal

$$\mathcal{L}^P(w) = \frac{1}{C} \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(y_i - w^T x_i)^2\right) \prod_{i=1}^{d} \frac{1}{\tau\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{2\tau^2}\right)$$

$$\mathcal{LL}^P(w) = \sum_{i=1}^{m} \frac{-1}{2\sigma^2}(y_i - w^T x_i)^2 + \sum_{i=1}^{d} \frac{-1}{2\tau^2}(w_i^2) + \text{const.}$$

$$-\mathcal{LL}^P(w) = \frac{1}{2\sigma^2} \sum_{i=1}^{m} (w^T x_i - y_i)^2 + \frac{1}{2\tau^2} \|w\|^2 + \text{const.}$$

$$= C\left(\sum_{i=1}^{m} (w^T x_i - y_i)^2 + \frac{\sigma^2}{\tau^2} \|w\|^2\right) + \text{const.}$$

$\sigma^2 \rightarrow$ Data label variance, $\tau^2 \rightarrow$ Prior variance

## L2 Regularised Regression: (Ridge)

$$\hat{R}(w) = \frac{1}{2} \sum_{i=1}^{m} (w^T x_i - y_i)^2 + \frac{\lambda}{2} \|w\|^2$$ (Regularised empirical risk)

$$= \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$$

$$\nabla \hat{R}(w) = X^T(Xw - y) + \lambda w$$

$$\nabla \hat{R}(\hat{w}) = X^T(X\hat{w} - y) + \lambda \hat{w} = 0$$

$$\Rightarrow \quad (X^T X \hat{w} - X^T y) = -\lambda \hat{w}$$

$$X^T X \hat{w} + \lambda \hat{w} = X^T y$$

$$(X^T X + \lambda I)\hat{w} = X^T y$$

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

S.T if $\lambda > 0$, this is always invertible.

---

Regularised empirical risk with feature vectors.

$$\hat{R}_\lambda(w) = \frac{1}{2} \sum_{i=1}^{m} (w^T \phi(x_i) - y_i)^2 + \frac{\lambda}{2} \|w\|^2$$

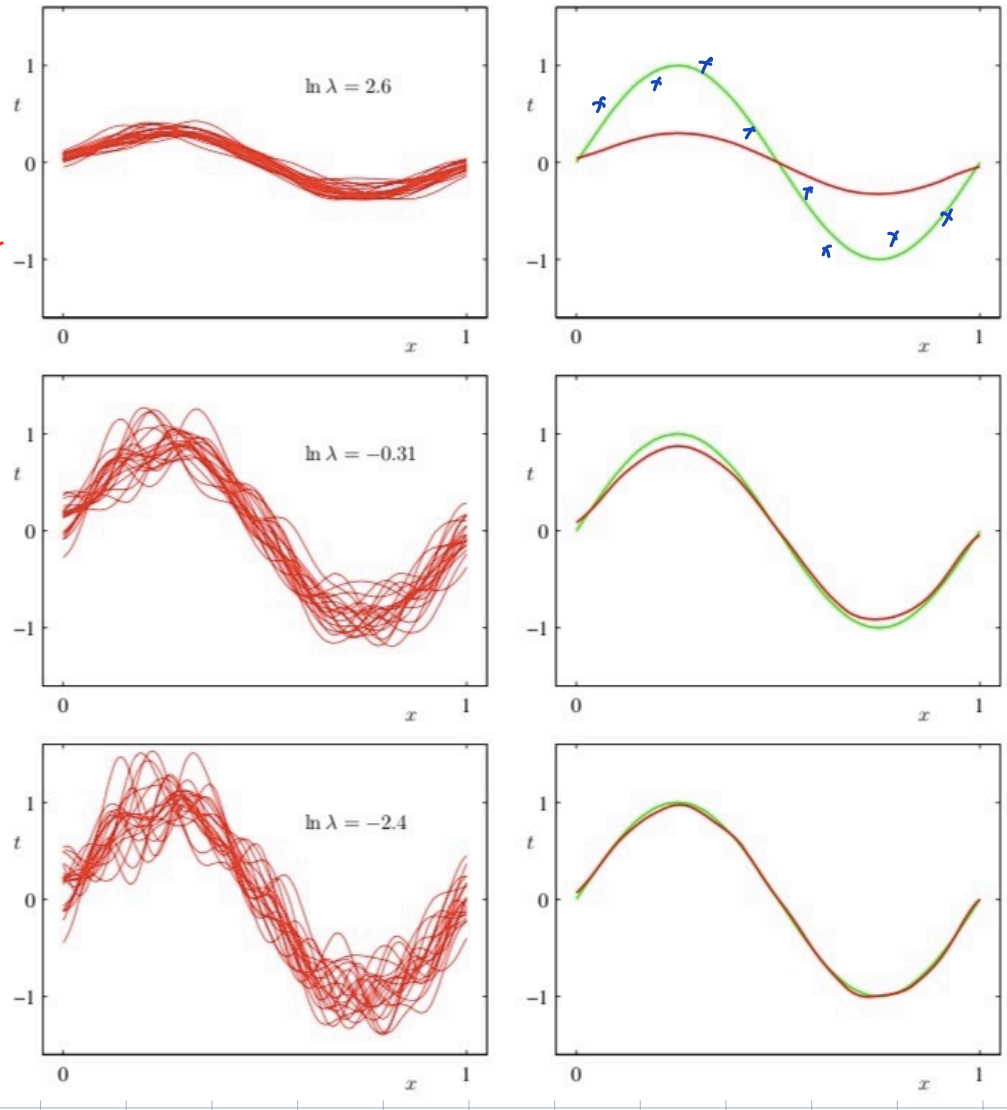
Solution:

$$\hat{w} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y$$

⇒ Two handles on the algorithm

i) Choosing $\phi$

ii) Choosin $\lambda$




Low Variance
High bias

high Variance
Low bias

Bishop

$\phi$ :

$x \to [1, x, x^2 \ldots x^{20}]$

Question: What role does $\lambda$ play ?

Recall $\lambda \propto \frac{\sigma^2}{\tau^2}$

$\therefore$ High $\lambda \Rightarrow$ Low $\tau$ (Prior variance is low)

Low $\lambda \Rightarrow$ High $\tau$ (Prior variance is high)

Equivalent Problems:

$$\min_w \frac{1}{m} \sum_{i=1}^{m} (w^T \phi(x_i) - y_i)^2$$

$$s.t \quad \|w\|^2 \leq B$$

# Linear Algebra Cheat sheet:

i) $f(x) = x^T A x$ for some $A \in \mathbb{R}^{d \times d}$

$\nabla f(x) = 2Ax$

(ii) $f(x) = w^T x$ for some $w \in \mathbb{R}^d$

$\nabla f(x) = w$

(iii) $(AB)^T = (B^T A^T)$

(iv) $(AB)^{-1} = B^{-1} A^{-1}$ if $AB$ is invertible.

(v) If $A$ is not invertible $\Leftrightarrow \exists x$ s.t $Ax = 0$

$\exists x \quad x^T A x = 0$

(vi) $f(x) = g(Ax)$ $\quad f: \mathbb{R}^d \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}$

$\nabla f(x) = A^T \nabla g(Ax)$ $\quad A \in \mathbb{R}^{n \times d}$

(vii) $\text{RowSpace}(A) \oplus \text{NullSpace}(A) = \mathbb{R}^d$ $\quad A \in \mathbb{R}^{n \times d}$

$\text{ColSpace}(A) \oplus \text{LeftNull}(A) = \mathbb{R}^n$

$\therefore \dim(RS(A)) + \dim(NS(A)) = d$

$\dim(CS(A)) + \dim(LNS(A)) = n$

$\dim(RS(A)) = \dim(CS(A))$

# Geometric Interpretation of Least Squares Regression.

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

$\Updownarrow$ ($n \times d$ matrix)

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|^2 \iff \min_{u \in CS(X)} \|u - y\|^2$$

Rf Columns of $X$ are LI then $X^TX$ is inv.

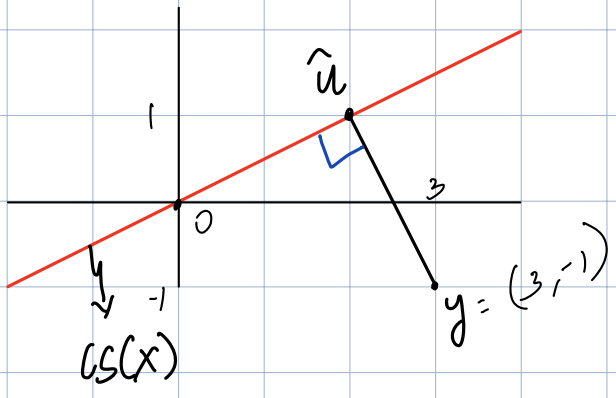
(Substituting $u = Xw$)

$$CS(X) = \{v \in \mathbb{R}^n : v = Xw \text{ for some } w \in \mathbb{R}^d\}$$

Recall: $\hat{w} = (X^TX)^{-1}X^Ty$ is the soln

$\therefore \hat{u} = X(X^TX)^{-1}X^Ty$ is the Proj of $y$ on $CS(X)$

---




$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} \quad y = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix}; \quad X^Ty = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

Solve

$$X^TX\hat{w} = X^Ty$$

Many candidate solns: $\begin{bmatrix} -1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ y_3 \end{bmatrix}$

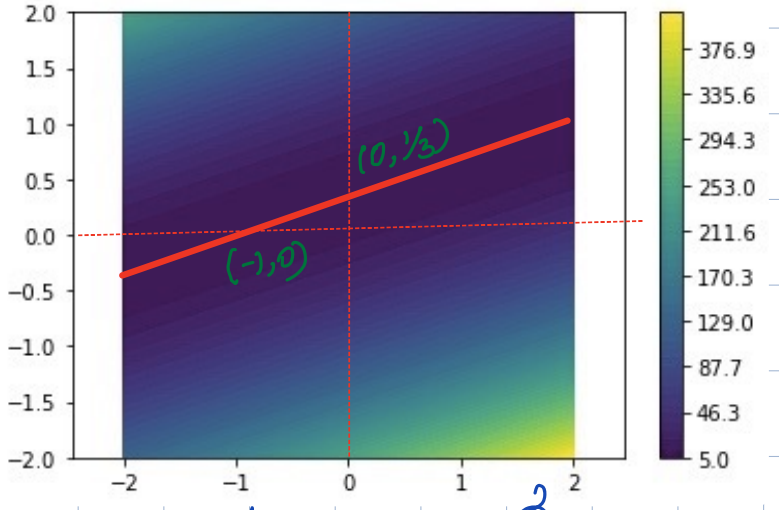
In all cases

$$\hat{u} = X\hat{w} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Default Pseudoinv soln: $\begin{bmatrix} 0 \\ 1/3 \end{bmatrix}$.

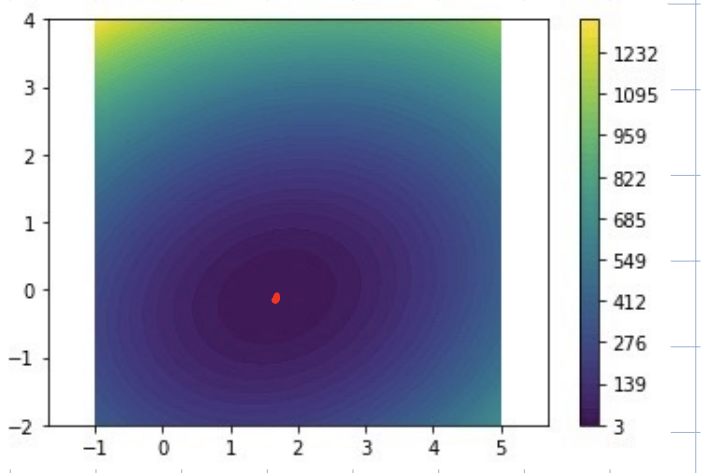
# Visualising contours of $\hat{R}(w) = \|Xw - y\|^2$

$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix}, \quad y = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$



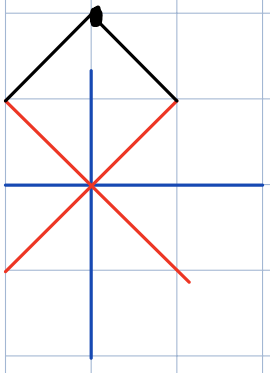

$\|Xw - y\|^2$

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 5 & -5 \end{bmatrix} \quad y = \begin{bmatrix} 3 \\ 4 \\ 9 \end{bmatrix}$$




$\hat{w} = [1.7, -0.1]$

## Representer Theorem:



$$\min_{w \in \mathbb{R}^{d'}} \frac{1}{2}\|\underbrace{\Phi}_{n \times d'} w - y\|^2 + \frac{\lambda}{2}\|w\|^2$$

$$\Phi^T \alpha = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

$$\Updownarrow$$

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2}\|\Phi\Phi^T \alpha - y\|^2 + \frac{\lambda}{2}\|\Phi^T \alpha\|^2$$

Why? Any $w \in \mathbb{R}^{d'} = v + u$ for some $u \in \text{Null}(\Phi)$, $v \in \text{Row}(\Phi)$

and $\|w\|^2 = \|v\|^2 + \|u\|^2$

Let $v = \Phi^T \alpha$

$$R(\alpha) = \frac{1}{2} \| \Phi\Phi^T\alpha - y \|^2 + \frac{\lambda}{2} \| \Phi^T\alpha \|^2$$

$$= \frac{1}{2} (y - \Phi\Phi^T\alpha)^T (y - \Phi\Phi^T\alpha) + \frac{\lambda}{2} \alpha^T \Phi\Phi^T\alpha$$

$$= \frac{1}{2} y^T y + \frac{1}{2} \alpha^T \Phi\Phi^T\Phi\Phi^T\alpha - \alpha^T\Phi\Phi^T y + \frac{\lambda}{2} \alpha^T \Phi\Phi^T\alpha$$

$$= \frac{1}{2} y^T y + \frac{1}{2} \alpha^T K K \alpha - \alpha^T K y + \frac{\lambda}{2} \alpha^T K \alpha$$

(where $K = \Phi\Phi^T$)

$$\nabla R(\hat{\alpha}) = K K\hat{\alpha} - K y + \lambda K\hat{\alpha} = 0$$

$n \times d$

$$K( K\hat{\alpha} - y + \lambda\hat{\alpha}) = 0$$

$K_{i,j} = \phi(x_i)^T \phi(x_j)$

$$K\hat{\alpha} + \lambda\hat{\alpha} = y$$

$$(K + \lambda I)\hat{\alpha} = y$$

s.t This is invertible if $\lambda > 0$

$$\hat{\alpha} = (K + \lambda I)^{-1} y$$

$$\therefore \text{ Soln } \hat{w} = \Phi^T (K + \lambda I)^{-1} y$$

$\lambda = 1$

Computation complexity comparison ($\Phi$ is $n \times d'$, $X$ is $n \times d$)

$$\hat{w} = (\Phi^T \Phi + I)^{-1} \Phi^T y$$

| | |
|---|---|
| $\Phi$ computation | $: nd'$ |
| $\Phi^T \Phi$ computation | $: (d')^2 n$ |
| Inversion | $: (d')^3$ |
| $\hat{w}$ computation | $: d'n + (d')^2$ |
| Total | $: (d')^3 + (d')^2 n$ |

$$\hat{w} = \Phi^T (\Phi \Phi^T + I)^{-1} y$$

| | |
|---|---|
| $\Phi$ computation | $: nd'$ |
| $\Phi \Phi^T$ | $: n^2 d'$ |
| Inversion | $: n^3$ |
| $\hat{w}$ computation | $: n^2 + nd'$ |
| Total | $: n^3 + n^2 d'$ |

---

$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called kernel function, if $K(u, v)$ represents $\phi(u)^T \phi(v)$ for some $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$

Example kernel: RBF/Gaussian

$$K(u, v) = \exp\left(-\frac{\|u - v\|^2}{\sigma^2}\right)$$

$u \in \mathbb{R}^d$, $v \in \mathbb{R}^d$

$\exists \ \phi : \mathbb{R}^d \to \mathcal{H}$

$$K(u, v) = \langle \phi(u), \phi(v) \rangle$$

Another Example Kernel:

$$K(u, v) = (u^T v)^{73}$$

Qn: What functions $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are valid kernel functions?

Exercise: Identify if following fns are kernels:

1) These kernels are for $d=1$

(a) $K(x,y) = x+y = \phi(x)^T \phi(y)$ for some $\phi: \mathbb{R} \to \mathbb{R}^{d'}$,

(b) $K(x,y) = x-y$

(c) $K(x,y) = xy$

(d) $K(x,y) = x^2$ $\quad \phi(x) = [x, 1]$

(e) $K(x,y) = x^2 + y^2$

2) These kernels are for $d > 1$

(a) $K(x,y) = x^T y$

(b) $K(x,y) = \sum_{i=1}^{d} x_i^2 y_i^2$ $\quad [x_1, x_2 \cdots x_d] \mapsto [x_1^2 \; x_2^2 \cdots x_d^2]$

(c) $K(x,y) = \exp(x^T y)$ $\quad$ (e) $K(x,y) = \exp\left(-\frac{\|x-y\|}{2\sigma^2}\right)$

(d) $K(x,y) = (1 + x^T y)^p$

---

Solution:

i) Find $\phi: \mathbb{R}^d \to \mathbb{R}^{d'}$ s.t $K(x,y) = \phi(x)^T \phi(y)$ to show $K$ is a valid kernel.

ii) For all sets of data points $x_1, \ldots x_n$ with $x_i \in \mathbb{R}^d$, the $n \times n$ Kernel matrix must be symmetric and PSD for $K$ to be a valid kernel.

$$K = \Phi \Phi^T$$

# Illustration of Kernel Function: $\phi : \mathbb{R}^2 \to \mathbb{R}^4$

$$X = \begin{bmatrix} t_1 & t_2 \\ u_1 & u_2 \\ v_1 & v_2 \end{bmatrix} \; ; \; \Phi = \begin{bmatrix} t_1^3 & \sqrt{3}t_1^2 t_2 & \sqrt{3}t_1 t_2^2 & t_2^3 \\ u_1^3 & \sqrt{3}u_1^2 u_2 & \sqrt{3}u_1 u_2^2 & u_2^3 \\ v_1^3 & \sqrt{3}v_1^2 v_2 & \sqrt{3}v_1 v_2^2 & v_2^3 \end{bmatrix} = \begin{bmatrix} \phi(t)^T \\ \phi(u)^T \\ \phi(v)^T \end{bmatrix}$$

2d data & homogenous 3rd degree polynomial regression

3 data points. $\phi(u) = [u_1^3, \sqrt{3}u_1^2 u_2, \sqrt{3}u_1 u_2^2, u_2^3]$

$$\Phi\Phi^T = \begin{bmatrix} \phi(t)^T\phi(t) & \phi(t)^T\phi(u) & \phi(t)^T\phi(v) \\ \phi(u)^T\phi(t) & \phi(u)^T\phi(u) & \phi(u)^T\phi(v) \\ \phi(v)^T\phi(t) & \phi(v)^T\phi(u) & \phi(v)^T\phi(v) \end{bmatrix}$$

---

$$\phi(u)^T\phi(v) = u_1^3 v_1^3 + 3u_1^2 u_2 v_1^2 v_2 + 3u_1 u_2^2 v_1 v_2^2 + u_2^3 v_2^3$$

$$= (u_1 v_1 + u_2 v_2)^3$$

$$= (u^T v)^3 = k(u, v)$$

$\downarrow$

Homogenous 3rd degree polynomial kernel.

(i.e) you can compute $\phi(u)^T\phi(v)$ without computing $\phi(u)$ & $\phi(v)$

Evaluating the learned function on a test point: $x$

$\hat{w}^T \phi(x) = \hat{\alpha}^T \Phi \phi(x)$ (Recall $\hat{w} = \Phi^T \hat{\alpha}$)

($\hat{\alpha}^T$: $1 \times n$, $\Phi$: $n \times d$, $\phi(x)$: $d \times 1$)

$$= [\hat{\alpha}_1 \ \hat{\alpha}_2 \cdots \hat{\alpha}_n] \begin{bmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix} \phi(x)$$

$$= \sum_{j=1}^{n} \hat{\alpha}_j \phi(x_j)^T \phi(x) = \sum_{j=1}^{n} \hat{\alpha}_j K(x_j, x)$$

$\therefore \hat{w}^T \phi(x)$ can also be evaluated without computing $\phi$ or $\hat{w}$.

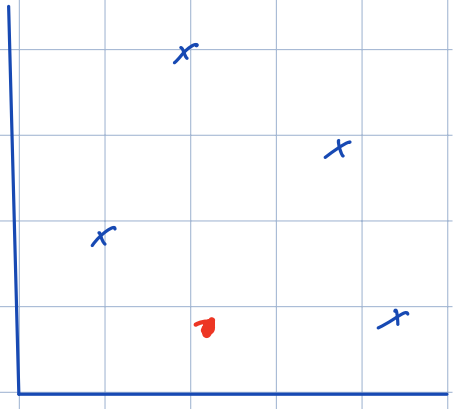
---

To check if $k: \mathbb{R}^d \times \mathbb{R}^d \Rightarrow \mathbb{R}$ is a valid kernel computationally.

Given $n$ datapoints $x_1, x_2 \cdots x_n$ $\quad x_i \in \mathbb{R}^d$

$$K \in \mathbb{R}^{n \times n} \qquad K_{ij} = k(x_i, x_j)$$

$$\hat{y}(x) = \sum_{i=1}^{n} \alpha_i^* k(x_i, x)$$



## Exercise:

Consider the Kernel Ridge regression problem below.

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \| \Phi \Phi^T \alpha - y \|^2 + \frac{\lambda}{2} \| \Phi^T \alpha \|^2$$

Let the data dimension be $d = 1$. $m = 3$.

Let $X = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$


Let Kernel $k$ be $K(u, v) = \exp(-(u-v)^2)$

(a) Let the $\alpha$ solution be $\alpha = \begin{bmatrix} 1 \\ 0.5 \\ -1 \end{bmatrix}$

Give the predicted $y$ value at points $x = 3$ and $x = 0$.

$\hat{y}(3) = \alpha_1 K(1,3) + \alpha_2 K(2,3) + \alpha_3 K(4,3)$

Also plot $\hat{y}(x)$ vs $x$.

(b) Find Optimal $\alpha^*$ for the Optimisation problem above with $\lambda = 1$, and $y = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}$.

(c) Repeat Part (a) with solution from Part (b).