Lecture 2 : Linear Regression.

Goal: $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ $\quad x^{(i)} \mapsto y^{(i)}$

$x^{(i)} \in \mathbb{R}^d$ (features, samples) $\quad y^{(i)} \in \mathbb{R}$ (targets, labels)

$(x, y) \sim \mathbb{P}(x, y)$

Ex. 1] [Housing Price Prediction] (Regression)

x : [# bedrooms, # bathrooms, Sq. footage] $\rightarrow$ y : [Price of House]

Ex. 2] Image Classification (Classification).

x : [pixel location + color $[0, 255] \subset \mathbb{Z}$]

y : [dogs, cats, birds] one-hot-vector

Q1: We want $\hat{f}(x) \approx y$ for all points $(x, y) \sim P_{(x,y)}$

($\hat{f}$: predictor)

What does $\approx$ mean?

Q2: What class of functions $\hat{f}$ are we considering?

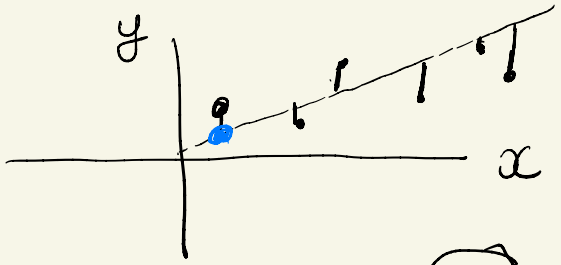
Q3: How do we find a good $\hat{f}$ from this class?

A1: Performance metric: Mean squared error (MSE) ←

$\{y_i, \hat{y}_i\}_{i=1}^n$

$y_i \in \mathbb{R}$, $\hat{y}_i \in \mathbb{R}$

$$MSE := \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$



A2: $\{ f : \mathbb{R}^d \to \mathbb{R} \; ; \; f(x) = \underbrace{w}_{\mathbb{R}^{1 \times d}} \underbrace{x}_{\mathbb{R}^d} \}$

A3: $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$

$x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$

Empirical Risk Minimization

$$\text{minimize} \quad \frac{1}{2} \sum_{i=1}^{n} (w x^{(i)} - y^{(i)})^2$$

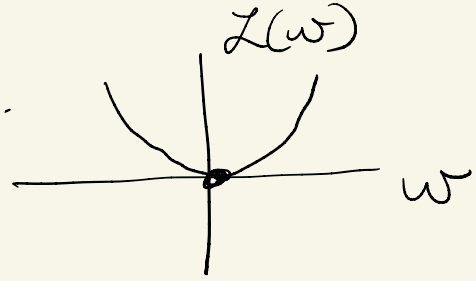
Objective: minimize $\frac{1}{2}\sum_{i=1}^{n}(wx^{(i)}-y^{(i)})^2 \iff \frac{1}{2}\|wX-y\|_2^2$

$w \in \mathbb{R}^{1\times d}$

$$X = \begin{bmatrix} 1 & 1 & & 1 \\ x^{(1)} & x^{(2)} & \cdots & x^{(n)} \\ 1 & 1 & & 1 \end{bmatrix} \quad ; \quad y = [\, y^{(1)} \;\; \cdots \;\; y^{(n)} \,]$$

Ex. 1 $\mathcal{L}(w) = \frac{1}{2}w^2$; Datapoint $x=1, y=0$.

$w \in \mathbb{R}^1$ $\quad \frac{1}{2} \in \mathbb{R}^1$



$\mathcal{L}'(w) = \boxed{w = 0.}$

convex.

Ex. 2 $\mathcal{L}(w) = \frac{1}{2}\|wX-y\|_2^2$

$\nabla\mathcal{L}(w) = (wX - y)X^T = 0.$

$wXX^T = yX^T$

$w \in \mathbb{R}^{1\times d}$, $X \in \mathbb{R}^{d\times n}$, $X^T \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^{1\times n}$, $yX^T \in \mathbb{R}^{1\times d}$

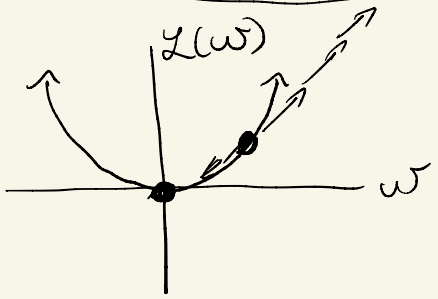
(1) XX^T invertible. $n \gg d$.

$\rightarrow w = (yX^T)(XX^T)^{-1}$

(2) XX^T is not invertible. $n < d$.

Gradient Descent: $\mathcal{L}(w)$; $w^{(0)}$ initialization, η step size.

$$\boxed{w^{(t+1)} = w^{(t)} - \eta \nabla \mathcal{L}(w^{(t)})}$$



$\mathcal{L}(w) = \frac{1}{2} w^2$; $w^{(0)}$

Step 1: $w^{(1)} = w^{(0)} - \eta w^{(0)} = \underbrace{(1-\eta)}_{c} w^{(0)}$

Step 2: $w^{(2)} = (1-\eta) w^{(1)} = \underbrace{(1-\eta)^2}_{c} w^{(0)}$

$\boxed{c < 1}$: $w^{(t)} = c^t \cdot w^{(0)}$

$t \to \infty$ $\quad w^{(t)} \to 0$.

$c < 1 \Rightarrow 1 - \eta < 1$

$\boxed{0 < \eta < 2}$

$\mathcal{L}(w) = \frac{1}{2}\|wX - y\|^2$; Run g.d. step size η, $w^{(0)} = 0$.

$\nabla\mathcal{L}(w) = (wX - y)X^T$

Step 1: $w^{(1)} = w^{(0)} - \eta(w^{(0)}X - y)X^T = w^{(0)}\underbrace{[I - \eta XX^T]}_{A} + \underbrace{\eta yX^T}_{B}$

$w^{(1)} = w^{(0)}A + B$

Step 2: $w^{(2)} = w^{(1)}A + B = (w^{(0)}A + B)A + B$

$= w^{(0)}A^2 + BA + B$

$= w^{(0)}A^2 + B[A + I]$

$w^{(1)} = B$

$w^{(2)} = B[A + I]$

$w^{(3)} = B[A^2 + A + I]$ $\rightarrow$ $w^{(t)} = B[A^{t-1} + A^{t-2} + \dots + A + I]$

$A := Q\Delta Q^T$; $A^K = Q\Delta^K Q^T$

$w^{(t)} = B[Q\Delta^{t-1}Q^T + Q\Delta^{t-2}Q^T + \dots + Q\Delta Q^T + QIQ^T]$

$= BQ[\Delta^{t-1} + \Delta^{t-2} + \dots + \Delta + I]Q^T$

$$w^{(t)} = BQ\left[\Delta^{t-1} + \cdots + \Delta + I\right]Q^T$$

$A = I - \eta X X^T = Q \Delta Q^T$

$B = \eta y X^T$

$$= BQ\left[\begin{bmatrix}\lambda_1^{t-1} & & \\ & \ddots & \\ & & \lambda_n^{t-1}\end{bmatrix} + \cdots + \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix} + \begin{bmatrix}1 & & \\ & \ddots & \\ & & 1\end{bmatrix}\right]Q^T$$

$$= BQ\begin{bmatrix}\frac{1-\lambda_1^t}{1-\lambda_1} & & \\ & \ddots & \\ & & \frac{1-\lambda_n^t}{1-\lambda_n}\end{bmatrix}Q^T$$

$$1 + \lambda_i + \lambda_i^2 + \cdots + \lambda_i^{t-1} = \frac{1-\lambda_i^t}{1-\lambda_i}$$

$A = I - \eta X X^T$; $X = U \Sigma V^T$

$= I - \eta U \Sigma^2 U^T$; $\Sigma = \begin{bmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2\end{bmatrix}$

$= UU^T - \eta U \Sigma^2 U^T$ ($UU^T = I$)

$= U\left[I - \eta \Sigma^2\right]U^T$

$\lambda_i = 1 - \eta \sigma_i^2$ ←

$\sigma_1 > \sigma_2 > \cdots > \sigma_n$

$|1 - \eta \sigma_i^2| < 1$

$$w^{(t)} = \eta y X^T Q \begin{bmatrix}\frac{1-(1-\eta\sigma_1^2)^t}{\eta\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1-(1-\eta\sigma_n^2)^t}{\eta\sigma_n^2}\end{bmatrix}Q^T$$

$\eta < \frac{2}{\sigma_1^2}$

$t \to \infty \quad w(t) = \underline{y X^T U \begin{bmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_n^2 \end{bmatrix} U^T}$

$A = I - \eta X X^T \; ; \; X = U \Sigma V^T$

$= U \left[I - \eta \Sigma^2 \right] U^T$

$= Q \Lambda Q^T$

$w(\infty) = y V \Sigma^T U^T U \begin{bmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_n^2 \end{bmatrix} U^T$

$= y V \underbrace{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}}_{\Sigma^T} \underbrace{\begin{bmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_n^2 \end{bmatrix}}_{\Sigma^{+2}} U^T$

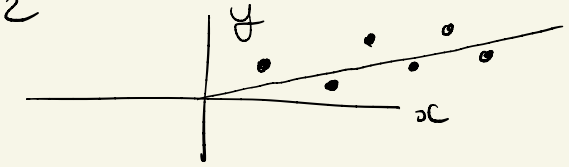
$= \boxed{y V \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_n \end{bmatrix} U^T} \quad \Sigma^+$

$= \boxed{y V \Sigma^+ U^T}$

$= \boxed{y X^+}$

(1) Underparameterized Regime

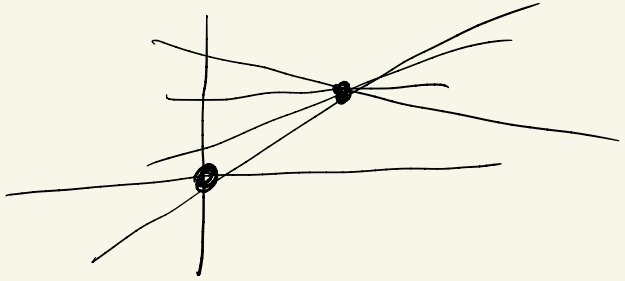
$\frac{1}{2} \| wX - y \|_2^2$ $\quad n >> d.$ $\quad (XX^T)$



$$w = (yX^T)(XX^T)^{-1}$$

$$= yX^+$$

(2) Overparameterized Regime



Last time: **Linear Regression**

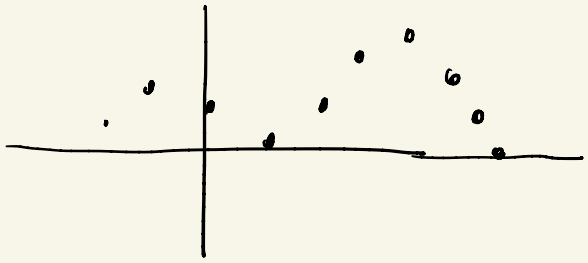
$\{ f : \mathbb{R}^d \to \mathbb{R} ; \quad f(x) = w x \}$; Solution $\hat{f}(x) = (y X^+) x$,

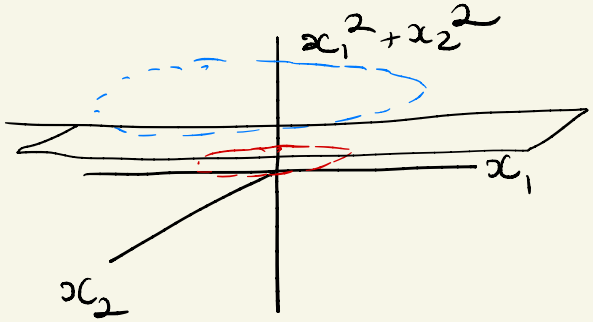
($w \in \mathbb{R}^{1\times d}$, $x \in \mathbb{R}^d$)

given by minimizing $\frac{1}{2} \| y - wX \|_2^2$

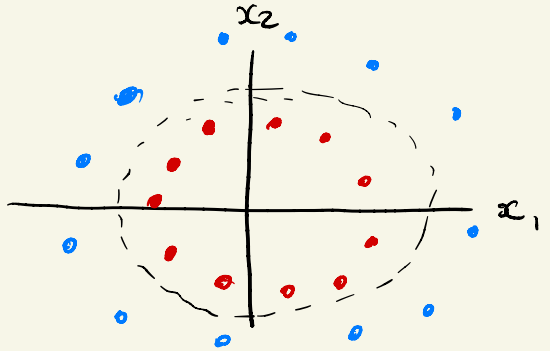
square loss or mean squared error.

Today: Kernel Regression





Ex. 1



$\{x^{(i)}\}_{i=1}^n \longrightarrow \{\psi(x^{(i)})\}_{i=1}^n \to$ L.R.

$\psi : \mathbb{R}^2 \to \mathbb{R}^3$

$(x_1, x_2) \longmapsto (x_1, x_2, x_1^2 + x_2^2)$

$$\mathcal{F} = \{ f : \mathbb{R}^d \to \mathbb{R} \; ; \; f(x) = \underbrace{\langle w, \psi(x) \rangle}_{w^T \psi(x)} \; ; \; \psi : \mathbb{R}^d \to \mathbb{R}^p \}$$

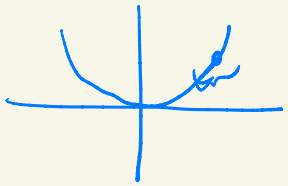
Q1: How do we deal with $p >> d$?

$$\mathcal{L}(w) = \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - w x^{(i)})^2 = \frac{1}{2} \| y - wX \|_2^2$$

$$\begin{bmatrix} - x^{(1)} - \\ - x^{(2)} - \\ \vdots \\ - x^{(n)} - \end{bmatrix}$$

$$w^{(t+1)} = w^{(t)} + \eta \underbrace{(y - w^{(t)} X)}_{1 \times n} X^T \leftarrow$$

$$\beta^{(t)} \quad [\beta_1^{(t)}, \beta_2^{(t)}, \dots, \beta_n^{(t)}]$$



$$= w^{(t)} + \eta \underbrace{\sum_{j=1}^{n} \beta_j^{(t)} x^{(j)T}}$$

$$w^{(0)} = 0$$

$$w^{(1)} = \eta \sum_{j=1}^{n} \beta_j^{(0)} x^{(j)T}$$

Guess: $w = \underbrace{\sum_{i=1}^{n} \alpha_i x^{(i)T}}$

Guess: $w = \sum_{i=1}^{n} \alpha_i x^{(i)^T}$; $\mathcal{L}(w) = \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - w x^{(i)})^2$

$$\mathcal{L}(w) = \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \sum_{j=1}^{n} \alpha_j x^{(j)^T} x^{(i)} \right)^2$$

$\langle x^{(j)}, x^{(i)} \rangle$

$x^{(i)} \in \mathbb{R}^d$

$\psi(x^{(i)}) \in \mathbb{R}^p$

$\langle \psi(x^{(i)}), \psi(x^{(j)}) \rangle$

$[\alpha_1, \dots, \alpha_n]$ $\quad \alpha \in \mathbb{R}^{1 \times n}$

$$\begin{bmatrix} \langle x^{(1)}, x^{(i)} \rangle \\ \vdots \\ \langle x^{(n)}, x^{(i)} \rangle \end{bmatrix}$$

$K(X, x^{(i)}) \in \mathbb{R}^n$

$$= \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \alpha \underbrace{K(X, x^{(i)})}_{\mathbb{R}^n})^2$$

$\alpha K(X, x^{(i)}) = y^{(i)}$

$1 \times n$ $\quad n$

$w \in \mathbb{R}^{1 \times d}$ $\quad wX = y$

$1 \times d$ $\quad d \times n$ $\quad 1 \times n$

$$\underbrace{\alpha}_{1 \times n} \; n \left[\begin{matrix} K(X, x^{(1)}) & K(X, x^{(2)}) & \cdots & K(X, x^{(n)}) \end{matrix} \right] = y$$

n $\quad n \times n$

$\frac{1}{2}\|wx - y\|_2^2 \Rightarrow \frac{1}{2}\|y - \alpha K(x,x)\|_2^2$ ← $y = \underbrace{\alpha}_{n\times n} K(x,x)$

$w = yX^T$ $\quad K(x,x) = \begin{bmatrix} 1 & 1 & & 1 \\ K(x,x^{(1)}) & K(x,x^{(2)}) & \cdots & K(x,x^{(n)}) \\ 1 & 1 & & 1 \end{bmatrix}$ $\quad \underline{yK(x,x)^{-1}}$

Step 1: Compute $K(x,x)$; $K(x,x)_{ij} = \langle x^{(i)}, x^{(j)} \rangle$

Step 2: $\alpha = yK(x,x)^{-1}$ $\quad K(x,x)_{ij} = \langle \psi(x^{(i)}), \psi(x^{(j)}) \rangle$

Define: Kernel as a symmetric, positive definite function

$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. $\quad K(x,z) \in \mathbb{R}$.

(1) $K(x,z) = K(z,x)$

(2) $\langle \psi(x), \psi(x) \rangle = \sum_{i=1}^{p} \psi(x)_i^2 \geq 0$. ←

$\forall \{x^{(i)}\}_{i=1}^n$, $\forall \{c_i\}_{i=1}^n \subset \mathbb{R}$ $\quad \sum_{i=1}^n \sum_{j=1}^n c_i c_j K(x^{(i)}, x^{(j)}) \geq 0$.

$$\forall \{x^{(i)}\}_{i=1}^{n}, \quad \forall \{c_i\}_{i=1}^{n} \subset \mathbb{R} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x^{(i)}, x^{(j)}) \geq 0.$$

Ex: $K(x, z) = \langle \psi(x), \psi(z) \rangle$

$n = 2$; $\{c_1, c_2\}$

$$\sum_{i=1}^{2}\sum_{j=1}^{2} c_i c_j K(x^{(i)}, x^{(j)}) = c_1^2 K(x^{(1)}, x^{(1)}) + c_1 c_2 K(x^{(1)}, x^{(2)}) + c_2 c_1 K(x^{(2)}, x^{(1)}) + c_2^2 K(x^{(2)}, x^{(2)})$$

$$= c_1^2 K(x^{(1)}, x^{(1)}) + 2 c_1 c_2 K(x^{(1)}, x^{(2)}) + c_2^2 K(x^{(2)}, x^{(2)})$$

$$= c_1^2 \underbrace{\langle \psi(x^{(1)}), \psi(x^{(1)}) \rangle}_{a^2} + 2 c_1 c_2 \underbrace{\langle \psi(x^{(1)}), \psi(x^{(2)}) \rangle}_{a \cdot b} + c_2^2 \underbrace{\|\psi(x^{(2)})\|_2^2}_{b^2}$$

$$\sim c_1^2 + 2c_1c_2 + c_2^2 \approx (c_1 + c_2)^2$$

$$= \boxed{\| c_1 \psi(x^{(1)}) + c_2 \psi(x^{(2)}) \|_2^2} \geq 0. \quad \leftarrow (c_1 a + c_2 b)^2$$

Examples: (1) $K(x, z) = \langle x, z \rangle$ (Linear)

$n \ll d$.

(2) Laplace kernel

$$K(x, z) = \exp\left\{ - \frac{\|x - z\|_2}{L} \right\}$$

L: bandwidth. $L > 0$.

$L = 10$

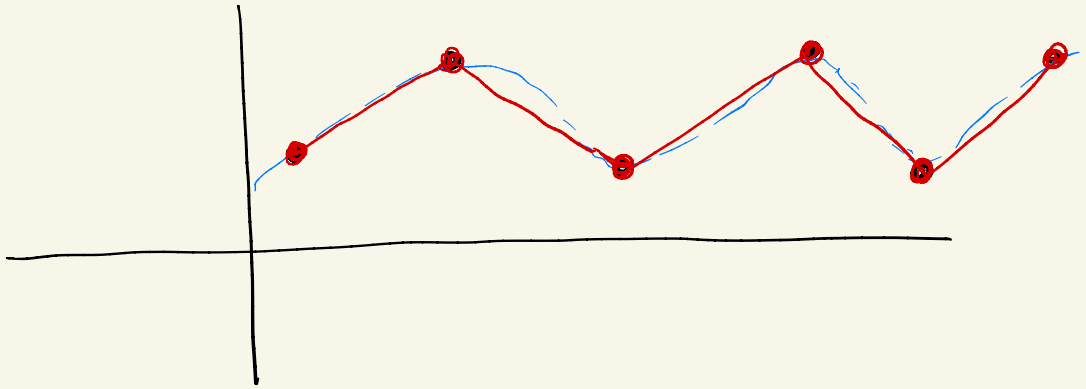
Algorithm: Given: (X_{train}, y_{train}), $K(x, z) \to \mathbb{R}$.

Step 1: $K_{train} = K(X_{train}, X_{train})$

Step 2: $\alpha = y K_{train}^{-1}$

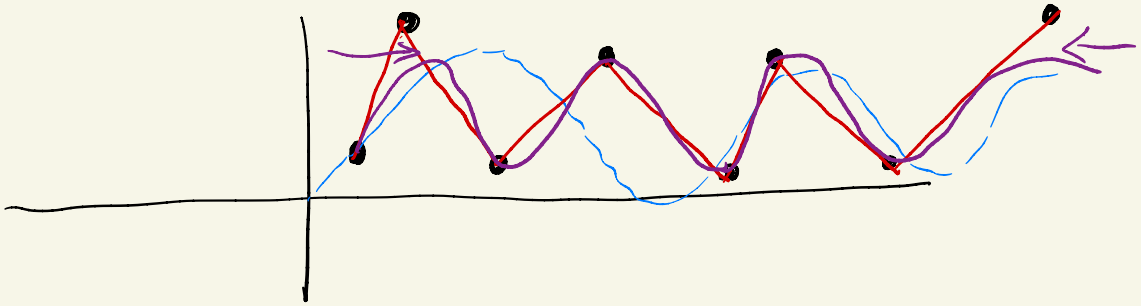
Step 3: $K_{test} = K(X_{train}, X_{test})$

test-pred $= \alpha K_{test}$



$$\alpha = y K^{-1}$$

$L = 10$



Step 1: $K_{train} = K(X_{train}, X_{train})$

Step 2: $\alpha = y(K + \lambda I)^{-1}$

$\lambda \in 10^{-2}, 10^{-3}$

$$K(x,z) = \exp\left\{-\frac{\|x-z\|_2}{L}\right\} \qquad L \to \infty \;;\; \lambda I$$

$$K(X,X) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = J$$

$$\alpha = y\,(\underbrace{K(X,X)}_{J} + \underbrace{\lambda I}_{A})^{-1}$$

Sherman – Morrison

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

$u \to \frac{1}{\lambda}\mathbf{1}$ $\quad \frac{1}{\lambda}\mathbf{1}^T \leftarrow v^T$

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$

$$A^{-1} = \frac{1}{\lambda} I$$

$$\alpha = y\left[\frac{1}{\lambda} I - \frac{\frac{1}{\lambda^2} J}{1 + \frac{1}{\lambda} n}\right]$$

$$\hat{f}(x) = \alpha \underbrace{K(x,x)}_{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}} = y\left[\frac{1}{\lambda}\mathbf{1} - \frac{\frac{n}{\lambda^2}\mathbf{1}}{1 + \frac{1}{\lambda} n}\right] = \underbrace{\sum_{i=1}^{n} y^{(i)}}_{} \underbrace{\left(\frac{1}{\lambda} - \frac{n}{\lambda^2 + \lambda n}\right)}_{C}$$