# * Bayesian Approches :

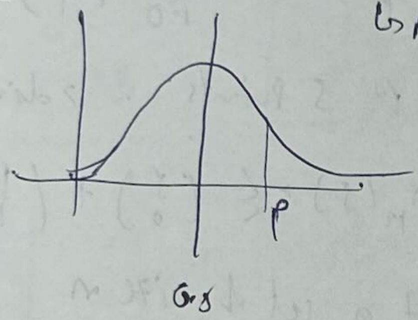
→ Frequentist Approches :- does not use prior prob.

but the Bayesian approches are use prior prob.

eg $D = \{H, T, T, H, \ldots\}$ find ~~of~~ bias of the coins.

$$P(H) = [0, 1]$$

* Bayesian : distribution of values of bias
  ↳ Means some perameter of that distⁿ.



→ eth every ~~ka~~ itteration prior will be updated.

$\theta$ → bias, Parameters.

$$\boxed{\underbrace{P(\theta \mid D, \alpha)}_{\text{Posterior}} \propto \underbrace{P(D \mid \theta, \alpha)}_{\text{likelyhood}} \cdot \underbrace{P(\theta \mid \alpha)}_{\text{prior}}}$$

* so Main thing is How get $\theta$'s.
  → two Approches → ① MLE - use the frequentist Approches
  ② MAP → Maxim A Posteries

Ex. Head with prob. $p$, tails w.p. $(1-p)$ $n$ coins tosses.

$y_i = 0$ tails

$y_i = 1$ Heads

$D = \{H\ T\ H\ H\ T\}$

$$L(\theta) = \boxed{\prod_{i=1}^{n} p^{y_i} \cdot (1-p)^{(1-y_i)}}$$

→ $\theta$ - parameters → likelyhood equestion
- this is not prob we need to mormalize them we get the prob.

$$\theta_{MLE} = \underset{\theta}{\arg\max}\ L(\theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(y_i / x_i, \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log P(y_i / x_i, \theta)$$

$$= \arg\min_{\theta} \sum_{i=1}^{n} - \cancel{\sum} \log P(y_i / x_i, \theta)$$

$\downarrow$

(*) $\theta_{MAP} = \arg\max \; - \cancel{\Sigma} \; L(\theta) \cdot P(\theta)$

$D: \{(x_i, y_i)\}_{i=1 \text{ to } n}$ $\quad x_i \in R^d$

$y_i \in R$

$y_i = w^T x_i + \epsilon_i$ $\quad X = R^d$

$\hookrightarrow N(0, 1/\beta)$ $\quad Y = R$

(gaussian noise)

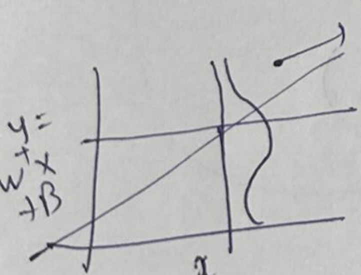
~~Likelihood~~ Likelihood $= \prod_{i=1}^{n} P(y_i / x_i, w, \beta)$

→ here we assume that $y_i$ is also distrⁿ on gaussian distⁿ.

mean is $w^T x + \epsilon$

$$= \cancel{\prod_{i=1}^{n} P(\epsilon_i /}$$

$$= \prod_{i=1}^{n} P(y_i / x_i, w, \beta) = \prod_{i=1}^{n} N(y_i \mid w^T x_i, \underbrace{1/\beta}_{\text{variance}})$$




$$\boxed{= \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \beta (y_i - w^T x_i)^2\right)}$$

$\vdots$

$$= \arg\min_{(w, \beta)} \sum_{i=1}^{n} - \log \sqrt{\frac{\beta}{2\pi}} + \sum_{i=1}^{n} \frac{1}{2} \beta (y_i - w^T x_i)^2$$

(if we know $\beta$ then we are to ERM problem)

→ but $\beta$ is unknown than find the $\beta$ & get the gradient 1 that $\frac{\partial L(w,\beta)}{\partial w} = 0$, $\frac{\partial L(w,\beta)}{\partial \beta} = 0$

Assumption

No auto correlation

① $\epsilon$ is independent $w^T x_i$

② $\epsilon_1, \epsilon_2 \ldots \epsilon_n$ are independent sample dist

③ Identical variant across datapoints.

} homoskedasticity

$$\frac{\partial l(w,\beta)}{\partial \beta} = \sum_{i=1}^{n}\left(-\frac{1}{2\beta} + \frac{1}{2}(y_i - w^T x_i)^2\right) = 0$$

$$\therefore \quad \frac{1}{\beta^*} = \frac{1}{n}\sum (y_i - w^T x_i)^2$$

$$\boxed{\therefore \sigma^2_{MLE} = \frac{1}{n}\sum (y_i - w^T x_i)^2}$$

↳ irreducible variance

$$\Rightarrow E_D[w_{MLE}] = E_D\left((X^T X)^{-1} X^T Y\right) \qquad Y = X^T w + \epsilon$$

expectation of 'w'.

this is Randomness because of Noise & which random Noise $N(0, 1/\beta)$

$\epsilon_1, \ldots, \epsilon_n$ $N(0, 1/\beta)$

$$= (X^T X)^{-1} X^T E_D(Y)$$

$$= (X^T X)^{-1} X^T \left(E(Xw) + E(\epsilon)\right)$$

$$= \underbrace{(X^T X)^{-1} X^T X}_{I} w$$

$$= w$$

$$\boxed{E_D(w_{MLE}) = w}$$

↳ $w_{MLE}$ is unbias estimator of $w$.

$$* \; E_D\left(\frac{1}{\beta_{MLE}}\right) = \qquad \left(\frac{1}{\beta_{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2\right.$$

$$\cancel{((x y)^T = y^T x^T}$$

$$= E_D\left[\frac{1}{n}(Xw_{MLE} - y)^T (Xw_{MLE} - y)\right]$$

$$= \frac{1}{n} E_D\left( \underbrace{w_{MLE}^T}_{= ((X^TX)^{-1}X^Ty)^T = (y^T X (X^TX)^{T-1})} X^T X w_{MLE} - 2 w_{MLE}^T X^T y + y^T y\right]$$

$$= \cancel{\frac{1}{n} E_D(}$$

$$= \frac{1}{n} E_D\left[ y^T X (X^TX)^{-1} \underbrace{X^T X (X^TX)^{-1}}_{I} X^T y - 2 y^T X (X^TX)^{-1} X^T y + y^T y\right)$$

$$= \frac{1}{n} E_D\left( y^T X (X^TX)^{-1} X^T y - 2 y^T X (X^TX)^{-1} X^T y + y^T y\right)$$

$$= \frac{1}{n} E_D\left( -y^T X (X^TX)^{-1} X^T y + y^T y\right)$$

$$= \frac{1}{n} E_D\left( y^T \underbrace{(I - X (X^TX)^{-1} X^T)}_{Z} y\right)$$

$$= \frac{1}{n} E_D\left[(Xw + \epsilon)^T Z (Xw + \epsilon)\right]$$

this is original '$w$' not '$w_{MLE}$'

$$= \frac{1}{n} E_D\left[ w^T X^T Z X w + \cancel{\epsilon^T Z X w}^{\,0} + \cancel{w^T X^T Z \epsilon}^{\,0} + \epsilon^T Z \epsilon\right]$$

$$= \frac{1}{n}\left[ w^T X^T Z X w + 0 + 0 + \epsilon^T Z \epsilon \right] \rightarrow E(\epsilon^T \epsilon) \neq 0)$$

$$\hookrightarrow \sum_{i=1}^{n} Z_{ij}\, E(\epsilon_i \epsilon_j)$$

$$(\because E(\epsilon_i \epsilon_j) = E(\epsilon_i)\,E(\epsilon_j)$$

$0 \cdot 0$ $\hookrightarrow$ this is for $i \neq j \rightarrow$ but for ($i = j = 0$) )

# scaller values

$tr(ABC) = tr(CBA)$

$= \frac{1}{n}$

$= \frac{1}{n}\left(w^T X^T Z X w + 0 + 0 + \frac{1}{\beta}\sum_{i=1}^{n} diag(Z)\right)$

$= \frac{1}{n}\left[\cancel{w^T X^T X w - w^T \underbrace{X^T X (X^T X)^{-1}}_{I} X^T X w} + \frac{1}{\beta}\sum_{i=1}^{n} diag(Z)\right]$

$= \frac{1}{n}\left(\frac{1}{\beta}\sum_{i=1}^{n} diag(Z)\right) = \boxed{\frac{1}{n}\left(\frac{1}{\beta} Tr(Z)\right)}$

$= \frac{1}{n}\left(\frac{1}{\beta} Tr(I - X(X^T X)^{-1} X^T)\right)$

$= \frac{1}{n}\left(\frac{1}{\beta} \underbrace{Tr(I)}_{n} - \underbrace{Tr(X(X^T X)^{-1} X^T)}\right)$

$\hookrightarrow X^T X \Rightarrow d \times d$

$\downarrow$

d-dimention

$Tr(X^T X) = d$

$= \frac{1}{n}\left(\frac{1}{\beta}(n - D)\right)$

$= \left(\frac{n-D}{n}\right)\left(\frac{1}{\beta}\right)$

dimention

$$\boxed{E_D\left(\frac{1}{\beta_{MLE}}\right) = \left(1 - \frac{D}{n}\right)\left(\frac{1}{\beta}\right)}$$

- Always under estimate the true variance
- If it is small rank ~~A~~ square matrix $E_D\left(\frac{1}{\beta_{MLE}}\right) = 0$
- if increase the $n\uparrow$ then bias will decrease.

* $Var(w_{MLE}) = \cancel{E_D(w_{MLE}^2) - (E(w_{MLE}))^2}$ (Var. of vector is valid. cover matrix)

$= E_D\left((w_{MLE} - E_D(w_{MLE}))(w_{MLE} - E_D(w_{MLE}))^T\right)$

$= E_D\left((w_{MLE} - w)(w_{MLE} - w)^T\right)$

$= E_D\left[\cancel{w_{MLE}^T w_{MLE} - 2 w_{MLE} w^T + w^T w}\right]$

$$= E_D\Big( \cancel{\left((X^TX)^{-1}X^Ty\right)^T (X^TX)^{-1}X^Ty}.$$

$$= E_D\Big( \left((X^TX)^{-1}X^Ty - w\right)\left((X^TX)^{-1}X^Ty - w\right)^T \Big)$$

$$= E_D\Big( \left((X^TX)^{-1}X^Ty\right)^T (X^TX)^{-1}\cancel{X^Ty} -$$

$$= E_D\Big( \left((X^TX)^{-1}X^T(w^TX + \epsilon) - w\right) \cdot$$

$$\left((X^TX)^{-1}X^T(w^TX + \epsilon) - w\right)^T \Big)$$

$$= E_D\Big( \left((X^TX)^{-1}X^Tw^TX + (X^TX)^{-1}X^T\epsilon - w\right) \cdot$$

$$\left((X^TX)^{-1}X^Tw^TX + (X^TX)^{-1}X^T\epsilon - w\right)^T \Big)$$

$$= E_D\Big( \left[(X^TX)^{-1}X^Tw^TX + (X^TX)^{-1}X^T\epsilon - w\right] \cdot$$

$$\left(\left((X^TX)^{-1}X^Tw^TX\right)^T + \epsilon^T(X^TX)^{-1}X - w^T\right)_0$$

$$= E_D\Big( \left((X^TX)^{-1}X^Tw^TX\right)^T \left((X^TX)^{-1}X^Tw^TX\right) + (X^TX)^{-1}X^Tw^T\cancel{X}\,\epsilon^T(X^TX)^{-1}X$$

$$- (X^TX)^{-1}X^Tw^TX\cancel{w}^T + 0 + 0 + 0 - \cancel{w}\left((X^TX)^{-1}X^Tw^TX\right)\Big)^T$$

$$=$$

$$= \frac{1}{\beta}(X^TX)^{-1}$$

* Posterior $\propto$ prior $\cdot$ likelyhood
  ↓ maximize this

* Gauss markov them.

$W_{MLE}$ is BLUE of $w$ (we want best variance (small))

best linear unbiased estimator

$\tilde{\omega}$ is another linear unbiased estimator

$$\underbrace{V(\tilde{\omega}) \geq Var(W_{MLE})}_{\text{need to prove this}}$$

$\to Var(\tilde{\omega}) = E[(\tilde{\omega} - E(\tilde{\omega}))(\tilde{\omega} - E(\tilde{\omega}))^T]$

$= E((\tilde{\omega} - AXw)(\tilde{\omega} - AXw)^T)$ ($\because E(\tilde{\omega}) = w$ because it is unbiased)

$= E((A\epsilon)(A\epsilon)^T)$

$= E(A\epsilon\epsilon^T A^T) = A E(\epsilon\epsilon^T) A^T$

$$\boxed{Var(\tilde{\omega}) = \frac{1}{\beta} AA^T}$$

$\tilde{\omega} = A \cdot Y$ some linear estimator

$= A(Xw + \epsilon)$

$\to \tilde{\omega} = AXw + A\epsilon$

$AXw = \tilde{\omega}$   $\boxed{AX = I}$   $E(A\epsilon) = 0$

$\therefore$ Now: $Var(\tilde{\omega}) - Var(W_{MLE})$

$= \frac{1}{\beta} AA^T - \frac{1}{\beta}(XX^T)^{-1}$

$= \frac{1}{\beta}(AA^T - (XX^T)^{-1})$

↳ some ~~settle~~ settle up is here.

$= \frac{1}{\beta}(((X^TX)^{-1}X^T + B)((X^TX)^{-1}X^T + B)^T)$

$= \frac{1}{\beta}(((X^TX)^{-1}X^TX(X^TX)^{-1} + \underbrace{BX(X^TX)^{-1}}_{T_1} + BB^T + \underbrace{(X^TX)^{-1}X^TB}_{T_2}$

$$Var(\tilde{\omega}) = \underbrace{\frac{1}{\beta}(X^TX)^{-1}}_{Var(W_{MLE})} + \underbrace{\frac{1}{\beta}BB^T}_{\text{this always posi. sym. matrix}} + \frac{T_1}{\beta} + \frac{T_2}{\beta}$$

$$(T_1 = BX(X^TX)^{-1} \;\&\; T_2 = (X^TX)^{-1}X^TB)$$

$$\text{so } A \Rightarrow (X^TX)^{-1}X^T + B$$

$$AX = (X^TX)^{-1}X^TX + BX$$

$$AX = I + BX$$

$$(AX - I) = BX$$

$$\hookrightarrow \cancel{AX - I} = (\because AX = I)$$

$$\boxed{BX = 0} \rightarrow \text{so } T_1 = 0, \; T_2 = 0$$

$$\therefore Var(\tilde{w}) - Var(w_{MLE}) = \frac{1}{\beta} BB^T$$

$$\therefore Var(\tilde{w}) \succeq Var(w_{MLE})$$

$$\boxed{Var(\tilde{w}) - Var(w_{MLE}) \succeq 0}$$

→ If we project $w_{MLE} \rightarrow \tilde{w}$ then point of $w_{MLE}$ is tighter

* <u>MAP</u>:-

$$y_i = w^T x_i + \epsilon_i \hookrightarrow N(0, 1/\beta)$$

$$\uparrow N(\mu_0, \Sigma_0) \hookrightarrow \text{Prior}$$

$$P(w \mid \underset{\text{data}}{D}, 1/\beta, \text{otherParam}) \propto P(D \mid w, 1/\beta) \cdot P(w)$$

$$: P(w \mid (x_1, y_1) \ldots (x_n, y_n), 1/\beta) \propto P(y_1, y_2 \ldots y_n \mid w, 1/\beta, x_1, x_2 \ldots x_n) \cdot P(w \mid \mu_0, \Sigma_0)$$

$$P(w \mid y_1 \ldots y_n)(x_1 \ldots x_n)) = \underbrace{\prod_{i=1}^{n} P(y_i \mid x_i, w, 1/\beta)}_{\text{likelyhood}} P(w \mid \mu_0, \Sigma_0)$$

$$= \prod_{i=1}^{n} N(y_i \mid w^T x_i, 1/\beta) \cdot P(w \mid \mu_0, \Sigma_0)$$

$$= \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\,(y_i - w^T x_i)^2 \cdot \beta\right)$$

$$\cdot \frac{1}{2\pi^{d/2}\,|\Sigma_0|} \exp\left(-\tfrac{1}{2}\,(w-\mu_0)^T \Sigma_0^{-1} (w-\mu_0)\right)$$

$$-\log p(w \mid D, \Sigma_0, \mu_0, 1/\beta) = -n \log\sqrt{\beta} + n\log\sqrt{2\pi} + \frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + d\log\sqrt{2\pi} + \frac{1}{2}\log|\Sigma_0| + \frac{1}{2}(w-\mu_0)^T \Sigma^{-1}(w-\mu_0)$$

$$\cancel{0 = 0 + 0 - \beta\sum_{i=1}^{n}(y_i - w^T x_i)\,x_i + 0 + 0 + (w-\mu_0)\,\Sigma_0^{-1}}$$

$$\cancel{0 = (w-\mu_0)\,\Sigma_0^{-1} - \beta(y - w\cancel{x})\,x}$$

$$\frac{d}{dw}\left(\frac{\beta}{2}(Xw - y)^T(Xw-y) + \frac{1}{2}(w-\mu_0)^T \Sigma_0^{-1}(w-\mu_0)\right)$$

$$= \cancel{\frac{\beta}{2}\left(2(X^T X)w\right)}$$

$$0 = \frac{\beta}{2}\left[\left(2(X^T X)w - 2(X^T y)\right)\right] + \frac{1}{2}\left(2\Sigma_0^{-1} w - 2\Sigma_0^{-1}\mu_0\right)$$

$$\beta(X^T X + \Sigma_0^{-1})w - \beta X^T y - \Sigma_0^{-1}\mu_0 = 0$$

$$\beta(X^T X + \Sigma_0^{-1})w = \beta X^T y + \Sigma_0^{-1}\mu_0$$

$$w_{MLE} = \left(\beta(X^T X + \Sigma_0^{-1})\right)^{-1}\left(\beta X^T y + \Sigma_0^{-1}\mu_0\right)$$

* 1 lecture is missing

$$\boxed{w_{MAP} = (\underbrace{\beta(X^TX) + \Sigma_0^{-1}}_{\text{Always is invertible}})^{-1} \cdot (\beta X^T y + \Sigma_0^{-1}\mu_0)}$$

$X^TX$ = PSD $\quad x^T A x \geq \forall x \neq \vec{0}$

$\Sigma_0^{-1}$ : PD $\quad \vec{x}^T A x > 0 \ \forall x \neq \vec{0}$

Cases:

① $n = 0$, $\boxed{w_{MAP} = \mu_0}$

② Prior is spherical $\rightarrow \Sigma_0^{-1} = \underbrace{\frac{1}{c}[I]}$ is some scalled version of identity.

$$w_{MAP} = (\beta(X^TX) + \Sigma_0^{-1})^{-1}(\beta X^T y + \Sigma_0^{-1}\mu_0)$$

$$= (\beta(X^TX) + \frac{1}{c}I)^{-1} \beta X^T y + \frac{1}{c}\mu_0$$

③ if spherical prior which is infinitely broad. $c \to \infty$

$$W_{MAP} = \beta(X^TX)^{-1}\beta X^T y$$

$\hookrightarrow w_{MAP} = w_{MLE} = w_{ERM}$

④ $\mu_0 = 0$ with spherical prior

$$(\beta(X^TX)^{-1} + \frac{1}{c}I)^{-1} \beta X^T y$$

$$= (X^TX)^{-1} + \underbrace{\frac{1}{\beta c}}_{\lambda} I)^{-1} \beta X^T y$$

⑤ Laplacian prior

$$P(w|\sigma_0, \mu_0) = \frac{1}{2\sigma_0} \exp\left(-\frac{\|w-\mu_0\|_1}{\sigma_0}\right)$$

$$\Rightarrow -\log P(w|\sigma_0, \mu_0) = -\log^1/_2\sigma_0 + \frac{\|w - \mu_0\|_1}{\sigma_0}$$

$$\partial(\tilde{J}(w)) = 0 + \|w_0 - \mu_0\|$$

$\hookrightarrow \|w - \mu_0\|_1$ $\hookrightarrow$ not define

$\Rightarrow$ $w_{AP}$ ( $\mu_0 = 0$, prior spherical )

$$w_{MAP} = ((X^T X)^{*} + \lambda I)^{-1} X^T y$$

$$E(w_{MAP}) = E((X^T X)^{*} + \lambda I)^{-1} X^T y)$$

$$(y = w^{*T} X + \epsilon)$$

$$= E((X^T X)^{*} + \lambda I)^{-1} X^T (w^{*T} X + \epsilon))$$

$$= E((X^T X)^{*} + \lambda I)^{-1} w^{*T} X^T X + X^T \epsilon)$$

$$= \cancel{E(X^T X)}$$

$$(X^T X)^{*} + \lambda I)^{-1} E(w^{*T} X^T X) + 0$$

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

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

$$E(w_{MAP}) = w^{*} - \lambda (X^T X + \lambda I)^{-1} w^{*}$$

this is bias estimator

* $V(w_{MAP}) = \_$

(do by your self)

$$w_{MAP} = \underbrace{(X^T X + \lambda I)^{-1} X^T}\, Y \qquad \left( Z = (X^T X + \lambda I)^{-1} X^T \right)$$

$$w_{MAP} = Z Y$$

$$= Z(X w^* + \epsilon)$$

$$\boxed{E_D(w_{MAP}) = Z X w^*}$$

$$Var(w_{MAP}) = E\left( (ZY - ZXw^*)(ZY - ZXw)^T \right)$$

$$= \cancel{E((ZY -} \; E$$

$$= E\left( (ZY - ZXw^*)(ZY - ZXw^*)^T \right)$$

$$\cancel{= E\left( (ZY - ZXw^*)(Y^T Z^T - w^{*T} X^T Z^T) \right)}$$

$$= Z\, E(Y - Xw^*)(Y - Xw^*)^T)$$

$$= Z\, E[((\epsilon)(\epsilon)^T)]\, Z^T$$

$$= Z \frac{1}{\beta} I Z^T$$

$$= (X^T X + \lambda I)^{-1} X^T \cdot \frac{1}{\beta} I \cdot \left( (X^T X + \lambda I)^{-1} X^T \right)^T$$

$$= \frac{1}{\beta} \left( (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \right)$$

$$= \frac{1}{\beta} V \operatorname{diag}\left( \frac{\sigma_i^2}{(\lambda + \sigma_i^2)} \right) V^T$$

$X^T X \qquad U \Sigma V^T \qquad X \in \mathbb{R}^{n \cdots}$

$\hookrightarrow n \times d \cdot n \times d \qquad U \; \Sigma \; V^T$

$n \times n$

$$= V \Sigma \underbrace{U^T U}_{I} \Sigma V$$

$$\boxed{X^T X = V \Sigma^2 V} = V \operatorname{diag}(\sigma_i^2) V^T$$

$$(X^T X)^{-1} = (V \Sigma^2 V^T)^{-1} \qquad (AB)^{-1} = B^{-1} A^{-1}$$

$$= V \Sigma^{-2} V^T$$

$$\boxed{(X^T X)^{-1} = V \, diag\left(\frac{1}{\sigma_i^2}\right) V^2}$$

↳ ~~to~~ this always
full rank matrix

$$\cancel{(X^T X + \lambda I)^{-1}} = \cancel{(X^T X)^{-1} + (\lambda I)^{-1}}$$

$$= \cancel{V \, diag\left(\frac{1}{\sigma_i^2}\right) V^2} + V \, \cancel{diag\left(\frac{1}{\sigma_i^2}\right)} W$$

$$(X^T X + \lambda I)^{-1} = \cancel{X^T X} \, (V \Sigma^2 V^T + \lambda V I V^T)^{-1}$$

$$= (V (\Sigma^2 + \lambda I) V^T)^{-1}$$

$$\boxed{(X^T X + \lambda I)^{-1} = V \, diag\left(\frac{1}{\sigma_i^2 + \lambda I}\right) V^T}$$

$$(X^T X + \lambda I)^{-1} X^T = V \, diag\left(\frac{1}{\sigma_i^2 + \lambda I}\right) V^T \cdot \cancel{diag(\sigma_i^2)} \, (V \Sigma V^T)$$

↓ get

$$= V \, diag\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda I}\right) V^T$$

$$= V \, diag\left(\frac{1}{\sigma_i^2 + \lambda I}\right) \underbrace{V^T V}_{I} \Sigma V^T$$

$$= V \, diag\left(\frac{1}{\sigma_i^2 + \lambda I}\right) diag(\sigma_i) V^T$$

$$\boxed{Z = (X^T X + \lambda I)^{-1} X^T = V \, diag\left(\frac{\sigma_i}{\sigma_i^2 + \lambda I}\right) V^T}$$

$$ZZ^T = V \operatorname{diag}\left(\frac{\sigma_i}{\sigma_i^2 + \lambda I}\right) \underbrace{V^T V}_{I} \operatorname{diag}\left(\frac{\sigma_i}{\sigma_i^2 + \lambda I}\right) V^T$$

$$\boxed{ZZ^T = V \operatorname{diag}\left(\frac{\sigma_i^2}{(\sigma_i^2 + \lambda I)^2}\right) V^T}$$

* Ensemble Methods: Bagging & Boosting

$$\text{Bagging} \downarrow \text{Parallel} \qquad \text{Boosting} \downarrow \text{Sequential}$$

* → Bagging: - Randomforest
  - Use multiple classifier

* Bootstrapping: sampling technique where sample are derived from whole population using replacement procedure.

* Aggregation: [illegible] incorporate all possible