Linear Regression, Regularization & the Bias-Variance Trade off.

Dataset $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^{n} \subset \mathbb{R}^d \times \mathbb{R}$

y_t: real number ($\mathbb{R}$)

Assume y_t is approximately an affine line of x_t, i.e.

$y_t \sim \theta^T x_t + \theta_0$ for some (θ, θ_0)

y_t: random; $\theta^T x_t + \theta_0$: not random

→ can use kernel trick on inputs x_t

More precisely, $\mathbb{E}[y_t | x_t, \theta, \theta_0] = \theta^T x_t + \theta_0$

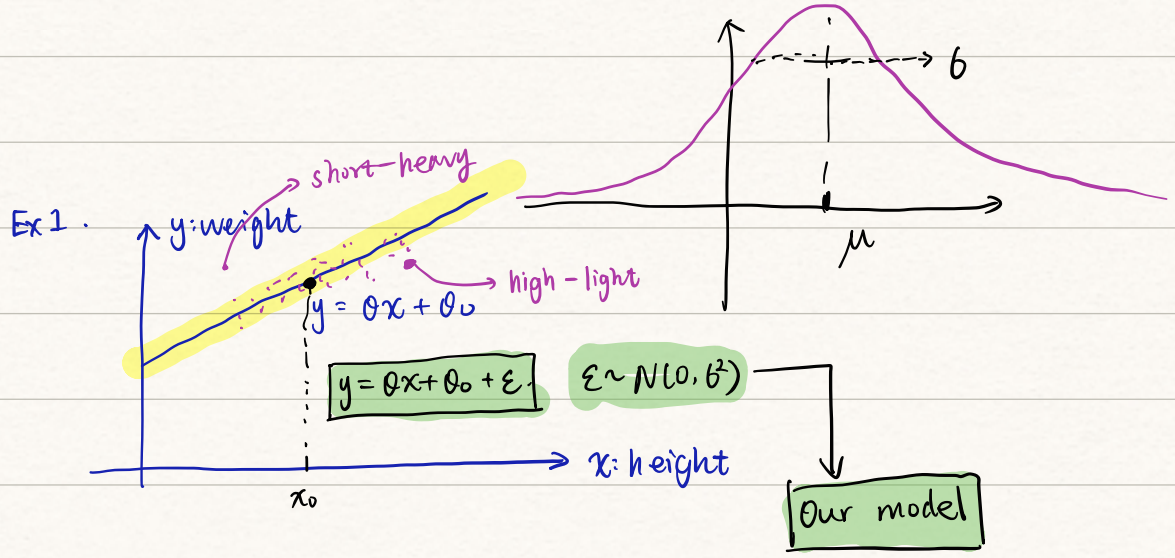
Associate a Prob. distribution of y given parameters (θ, θ_0) & input x

$\rightarrow P(y | x, \theta, \theta_0)$

$\Downarrow$ tractable

Gaussian: $P(y | x, \theta, \theta_0) = \mathcal{N}(y; \theta^T x + \theta_0, \sigma^2)$

$$\mathcal{N}(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \quad y \in \mathbb{R}$$



Clearly, $\mathbb{E}[y | x, \theta, \theta_0] = \mathbb{E}[\theta^T x + \theta_0 + \varepsilon | x, \theta, \theta_0]$

$= \theta^T x + \theta_0$

Dataset $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^n$

Generative Model:

$y_t = \underline{\theta}^T x_t + \theta_0 + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma^2)$

$\Downarrow$

$\varepsilon_1, \dots, \varepsilon_n \overset{i.i.d}{\sim} N(0, \sigma^2)$

Goal: Given $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^n$, estimate the parameters $(\underline{\theta}, \theta_0) \in \mathbb{R}^d \times \mathbb{R}$.

Strategy: Maximum Likelihood Estimation

① Likelihood f^n: $L(\underline{\theta}, \theta_0, \sigma^2 | \mathcal{D}) = P(\mathcal{D} | \underline{\theta}, \theta_0, \sigma^2)$

$$\overset{(i.i.d)}{=} \prod_{t=1}^n P(y_t | x_t, \underline{\theta}, \theta_0, \sigma^2)$$

$$= \prod_{t=1}^n N(y_t; \underline{\theta}^T x_t + \theta_0, \sigma^2)$$

Note: How likely we see the dataset under the current parameters $(\underline{\theta}, \theta_0, \sigma^2)$

Notation:

σ^2: models error that are not captured by Linear Model.

② log-likelihood:

$$\ell(\underline{\theta}, \theta, \sigma^2 | \mathcal{D}) = \sum_{t=1}^n \log\{N(y_t; \underline{\theta}^T x_t + \theta_0, \sigma^2)\}$$

$$= \sum_{t=1}^n \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[y_t - (\theta^T x_t + \theta_0)]^2\right\}\right].$$

$$= \sum_{t=1}^n \left[\text{const} - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_t - (\theta^T x_t + \theta_0))^2\right]$$

$$= \text{const} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^n (y_t - \theta^T x_t - \theta_0)^2$$

(Simplify)

Note: the estimation of $(\underline{\theta}, \theta_0)$ is decoupled from that of σ^2.

$\Rightarrow$ we can estimate $(\underline{\theta}, \theta_0)$ first

$$\text{Minimize}_{\underline{\theta}, \theta_0} \boxed{\sum_{t=1}^{n} (y_t - \underline{\theta}^T x_t - \theta_0)^2} \longrightarrow \text{massage into matrix form!}$$

① Define the target vector $\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$ y_t : t-th label/target.

② Define the design matrix $X = \begin{bmatrix} \underline{x}_1^T & 1 \\ & \vdots \\ \underline{x}_n^T & 1 \end{bmatrix}_{n \times d+1}$

$$\sum_{t=1}^{n} (y_t - \underline{\theta}^T x_t - \theta_0)^2 = \sum_{t=1}^{n} \left(y_t - [\underline{x}_t^T \; 1] \begin{bmatrix} \underline{\theta} \\ \theta_0 \end{bmatrix} \right)^2$$

$$= \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \underline{x}_1^T & 1 \\ & \vdots \\ \underline{x}_n^T & 1 \end{bmatrix} \begin{bmatrix} \underline{\theta} \\ \theta_0 \end{bmatrix} \right\|^2$$

$\underline{y}$: $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ X : $\begin{bmatrix} \underline{x}_1^T & 1 \\ & \vdots \\ \underline{x}_n^T & 1 \end{bmatrix}$ parameters : $\begin{bmatrix} \underline{\theta} \\ \theta_0 \end{bmatrix}$

$$= \| y - X \begin{bmatrix} \underline{\theta} \\ \theta_0 \end{bmatrix} \|^2 = \| y - X\tilde{\theta} \|^2 \quad \left(\tilde{\theta} = \begin{bmatrix} \underline{\theta} \\ \theta_0 \end{bmatrix}_{(d+1) \times 1} \right)$$

(Understanding of Decouple)

Firstly, we maximize log-likelihood:

$$\ell(\underline{\theta}, \theta_0, \sigma^2 | \mathcal{D}) = \cancel{\text{const}} - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (y_t - \underline{\theta}^T x_t - \theta_0)^2$$

$$= \underbrace{f_1(\sigma^2) - \overbrace{f_2(\sigma^2)}^{>0} \cdot \overbrace{f_3(\underline{\theta}, \theta_0)}^{\min}}_{\text{maximize}} \qquad \tilde{\theta} = \begin{pmatrix} \underline{\theta} \\ \theta_0 \end{pmatrix}$$

$\frac{\partial}{\partial \underline{\theta}} \ell = 0 \Leftrightarrow \frac{\partial}{\partial \underline{\theta}} f_3 = 0$ ⟶ decouple!

Now we want to: $\min_{\widetilde{\theta}} \|y - X\widetilde{\theta}\|^2$

$$\Leftrightarrow \min_{\widetilde{\theta}} (y - X\widetilde{\theta})^T (y - X\widetilde{\theta})$$

$$\Leftrightarrow \min_{\widetilde{\theta}} \underbrace{\widetilde{\theta}^T X^T X \widetilde{\theta} - 2\widetilde{\theta}^T X^T y}_{g(\widetilde{\theta})}.$$

$$\frac{\partial g}{\partial \underline{\theta}} = 2X^T X\widetilde{\theta} - 2X^T y = 0 \Rightarrow \widetilde{\theta} = (X^T X)^{-1} X^T y$$

Actually is $\hat{\widetilde{\theta}}$!

That is to say, $\begin{pmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{pmatrix} = (X^T X)^{-1} X^T y$

↳ X has full column rank

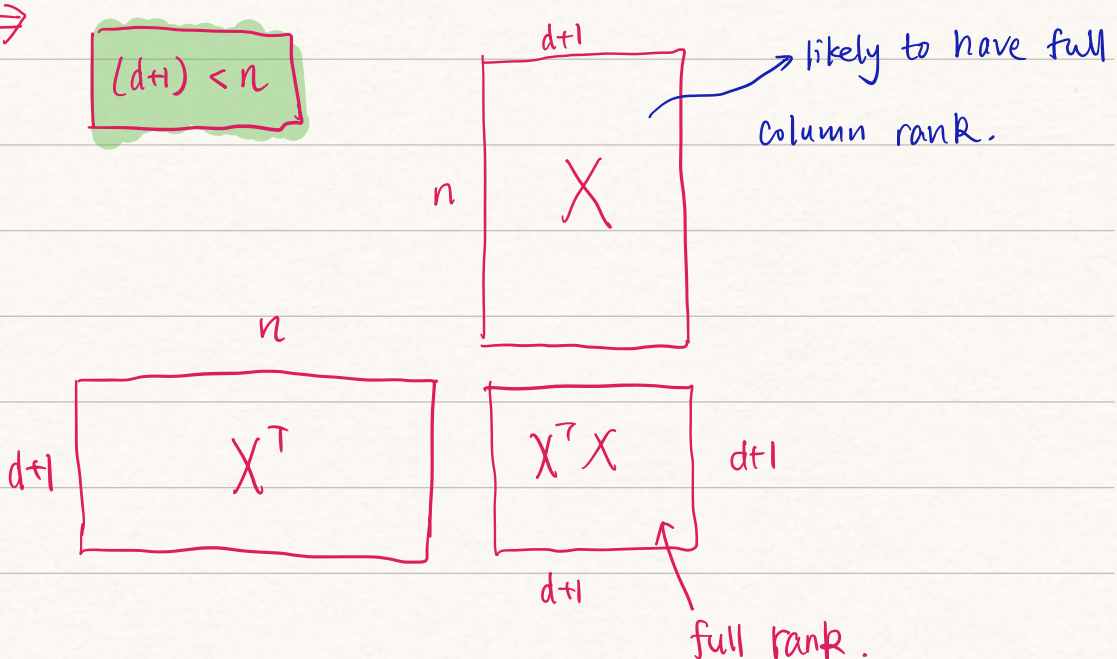
Question 1: Why is $\hat{\widetilde{\theta}}$ the globally minimum?

Quadratic Optimization

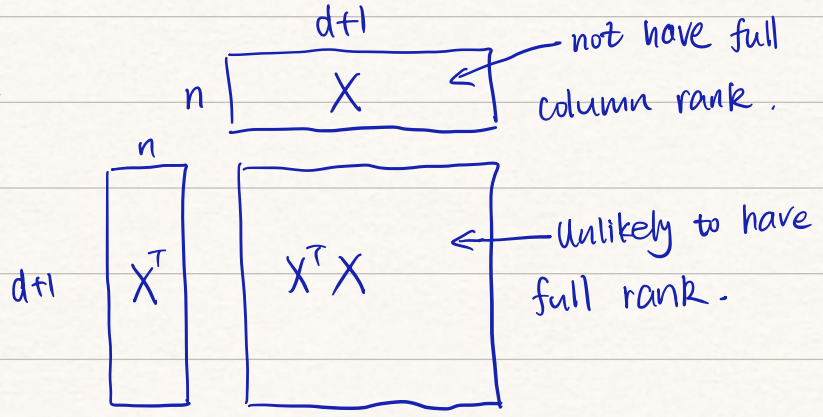
Question 2: What does X need to satisfy?

Q2. $X \rightarrow$ has full column rank. $\Rightarrow (X^T X)^{-1}$ exists

$X \in \mathbb{R}^{n \times (d+1)} \Rightarrow$ $(d+1) < n$



Problem: if X looks like



Rmk: i) $\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix}$ is a linear function of target vector $\underline{y}$.

ii) $\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix}$ depends in a highly non-linear way on $\boxed{\underline{x}_1, \ldots, \underline{x}_n}$

And it is coupled through the pseudo-inverse $(X^TX)^{-1}X^T$

iii) To be able to invert X^TX, we need X to have full column rank. (i.e. all columns in X are linearly independent).

↓

Not likely to happen if $d+1 \gg n$

↓

High-dimension curse.

- $d \approx 19K \sim 20K$ protein coding genes
- $n \approx 1{,}000$ patients (subjects)

Now we have $\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} = (X^TX)^{-1}X^T\underline{y}$

$$\ell(\underline{\theta}, \theta_0, \sigma^2 | \mathcal{D}) = \text{const} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \underline{\theta}^T\underline{x}_t - \theta_0)^2$$

↓

$$\frac{d}{d\sigma^2}\ell = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{\underline{\theta}}^T\underline{x}_t - \hat{\theta}_0)^2$$

(

Meaning of this (Does it makes sense?)

This is the average squared prediction error & caputures the "loss" non-ideal nature of the linear model!

↳ Step 1: $\hat{\underline{\theta}}, \hat{\theta}_0 \longrightarrow$ Step 2: $\hat{\sigma}^2$.

Bias - Variance TradeOff.

Note that: $\hat{\underline{\theta}}, \hat{\theta}_0, \hat{\sigma}^2$ are RANDOM.

$$\begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\theta}_0 \end{bmatrix} = \underbrace{(X^T X)^{-1} X^T}_{\text{not random}} \underbrace{y}_{\text{Random!}} \Rightarrow y_t = \underline{\theta}^T \underline{x}_t + \theta_0 + \varepsilon_t$$

Suppose $\mathcal{D}$ is indeed generated from some linear model.

i.e. $y_t = \underline{\theta}^* \underline{x}_t + \theta_0^* + \varepsilon_t \quad t = 1, 2, \dots, n$

for some $\underline{\theta}^* \in \mathbb{R}^d$, $\theta_0^* \in \mathbb{R}$

Here $\underline{\theta}^*$ & θ_0^* are the ground truth parameters

$$\underline{y} = X \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix} + \underline{\varepsilon} \qquad \begin{cases} \mathbb{E}(\underline{\varepsilon}) = \underline{0} \\ \text{cov}(\underline{\varepsilon}) = \mathbb{E}(\underline{\varepsilon}^T \underline{\varepsilon}) = \sigma^2 I_n \end{cases}$$

$$\Updownarrow$$

$$\begin{bmatrix} \mathbb{E}(\varepsilon_1^2) & \cdots & \mathbb{E}(\varepsilon_1 \varepsilon_t) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(\varepsilon_t \varepsilon_1) & \cdots & \mathbb{E}(\varepsilon_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$$

By the Previous result: $\begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\theta}_0 \end{bmatrix} = (X^T X)^{-1} X^T \underline{y} \rightsquigarrow \underline{y} = X \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix} + \underline{\varepsilon}$

$$\begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\theta}_0 \end{bmatrix} = (X^T X)^{-1} X^T \left(X \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix} + \underline{\varepsilon} \right)$$

$$= \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} + (X^T X)^{-1} X^T \underline{\varepsilon}$$

Q1: What's the **BIAS**?

↳ $\mathbb{E}\left[\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} - \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} \middle| X \right] = \underline{0} \Rightarrow$ BIAS is $\underline{0}$.

(for the Least square Sol^n)

↳ data.

Rmk: In Other Words, the estimate $\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} = (X^T X)^{-1} X^T \underline{y}$ is **UNBIASED**!

Q2: What is the **uncertainty** of the Parameter Estimate?

↳ $\mathrm{cov}\left(\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} \middle| X \right) = \mathbb{E}\left[\left(\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} - \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} \right) \left(\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} - \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} \right)^T \middle| X \right]$

Name of matrix:

① $X^T X \rightsquigarrow$ Gram Matrix

② $X \rightsquigarrow$ Design Matrix

$$= \mathbb{E}\left[(X^T X)^{-1} X^T \underline{\varepsilon}\, \underline{\varepsilon}^T X \left((X^T X)^{-1} \right)^T \middle| X \right]$$

Symmetry

$$= (X^T X)^{-1} X^T \, \mathbb{E}\left[\underline{\varepsilon}\, \underline{\varepsilon}^T \middle| X \right] X (X^T X)^{-1}$$

[Prop] → if $A = A^T$, then $(A^{-1})^T = A^{-1}$

$$= \sigma^2 (X^T X)^{-1}$$

Rmk: Uncertainty depends on the inputs X

$$X^T X = \begin{bmatrix} x_1 & \cdots & x_n \\ 1 & & 1 \end{bmatrix} \begin{bmatrix} x_1^T & 1 \\ \vdots & \\ x_n^T & 1 \end{bmatrix} = \sum_{t=1}^{n} \begin{bmatrix} x_t \\ 1 \end{bmatrix} \begin{bmatrix} x_t^T & 1 \end{bmatrix}$$

$$= n \left(\frac{1}{n} \sum_{t=1}^{n} \begin{bmatrix} x_t \\ 1 \end{bmatrix} \begin{bmatrix} x_t^T & 1 \end{bmatrix} \right)$$

Average of some matrices

① (Large Number Thm)

Note: If $x_t \overset{i.i.d}{\sim} P(x)$, then we have:

$$\frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix}x_t\\1\end{bmatrix}\begin{bmatrix}x_t^T & 1\end{bmatrix} \xrightarrow{P} C \rightsquigarrow C = \mathop{\mathbb{E}}_{x\sim P}\left[\begin{bmatrix}x\\1\end{bmatrix}\begin{bmatrix}x^T & 1\end{bmatrix}\right]$$

Roughly, $(X^TX) \xrightarrow{P} n\cdot C \Rightarrow (X^TX)^{-1} \xrightarrow{P} \frac{1}{n}\cdot\widetilde{C}$, $\widetilde{C} = C^{-1}$

Therefore $\mathrm{cov}[\underline{\varepsilon}] \xrightarrow{P} \frac{\sigma^2}{n}C^{-1}$

(2)

Fact 1: $\mathbb{E}[\|z - z^*\|^2] = \|\mathbb{E}[z] - z^*\|^2 + \mathbb{E}\|z - \mathbb{E}[z]\|^2$

MSE (Mean Squared Error); z: estimator (r.v.); z^*: ground truth parameter; $\|\mathbb{E}[z] - z^*\|^2$: Bias^2; $\mathbb{E}\|z - \mathbb{E}[z]\|^2$: Variance of z.

$\Updownarrow$

$\mathrm{Tr}(\mathrm{Cov}(z))$.

Pf: $\mathbb{E}[\|z - z^*\|^2] = \mathbb{E}[\|z - \mathbb{E}[z] + \mathbb{E}[z] - z^*\|^2]$

$= \mathbb{E}[\|z - \mathbb{E}[z]\|^2 + \|\mathbb{E}[z] - z^*\|^2 + 2\langle z - \mathbb{E}[z], \mathbb{E}[z] - z^*\rangle]$

($z - \mathbb{E}[z]$: r.v.; $\mathbb{E}[z] - z^*$: const; $= 0$)

$= \mathbb{E}[\|z - \mathbb{E}[z]\|^2 + \|\mathbb{E}[z] - z^*\|^2$

(3)

Observation $\|a\|^2 = \mathrm{Tr}(a\cdot a^T) = a^Ta$.

Fact 2. Variance $= \mathbb{E}[\|z - \mathbb{E}[z]\|^2] = \mathrm{Tr}(\mathrm{Cov}(z))$

Pf: $\mathbb{E}[\|z - \mathbb{E}[z]\|^2]$

$= \mathbb{E}[(z - \mathbb{E}[z])^T(z - \mathbb{E}[z])]$

$= \mathrm{Tr}(\mathbb{E}[(z - \mathbb{E}[z])^T(z - \mathbb{E}[z])])$

$= \mathrm{Tr}(\mathbb{E}[(z - \mathbb{E}(z))(z - \mathbb{E}(z))^T])$

$= \mathrm{Tr}(\mathrm{Cov}[z])$

For this Problem, we already have:

$$\begin{cases}\mathbb{E}[\hat{\theta} - \theta^*] = 0\\ \mathrm{Cov}[\hat{\theta}] = \sigma^2(X^TX)^{-1}\end{cases}$$

$$\Rightarrow \mathrm{MSE} = \mathbb{E}[\|\hat{\theta} - \theta^*\|^2] = \mathrm{Tr}(\sigma^2(X^TX)^{-1}) = \sigma^2\,\mathrm{Tr}((X^TX)^{-1})$$

Remind: $C = \mathop{\mathbb{E}}_{x\sim P}\left[\begin{bmatrix}x\\1\end{bmatrix}\begin{bmatrix}x^T & 1\end{bmatrix}^T\right]$

$(X^TX)^{-1} \xrightarrow{P} \frac{1}{n}\cdot C^{-1}$

$C \in \mathbb{R}^{(d+1)\times(d+1)}$

$$\approx \frac{\sigma^2}{n} \mathrm{Tr}(C^{-1})$$

Rmk: i) If $n \to +\infty$, MSE $\to 0$

ii) MSE $\propto \sigma^2$

iii) $\mathrm{Tr}(C^{-1})$ is of order $d+1$ (is # of parameters in $(\underline{\theta}, \theta_0)$)

Penalized log-likelihood & Bayesian Inference.

Problem: $X^T X$ may not be invertible !

↓

Idea: Put a prior probability distribution on $\begin{bmatrix} \underline{\theta} \\ \theta_0 \end{bmatrix}$.

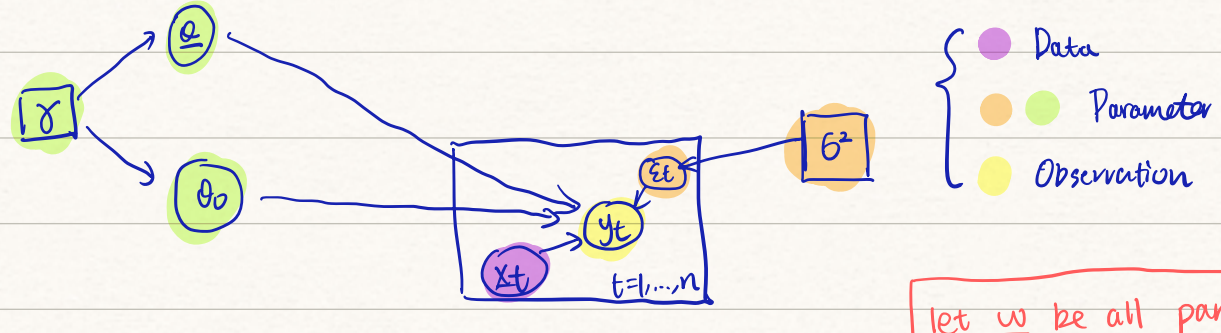
$$P(\underline{\theta}, \theta_0, \gamma) = N\left(\begin{bmatrix} \underline{\theta} \\ \theta_0 \end{bmatrix}; \begin{bmatrix} \underline{0} \\ 0 \end{bmatrix}, \gamma I\right)$$

$$\equiv N(\theta_0; 0, \gamma) \cdot \prod_{i=1}^{d} N(\theta_i; 0, \gamma)$$

Motivation:

Encouraging the parameters $(\underline{\theta}, \theta_0)$ shrunk to be $(\underline{0}, 0)$

Graphical Model (Bayesian Network)



let $\underline{w}$ be all parameters !

$$y_t = \underline{\theta}^T x_t + \theta_0 + \varepsilon_t \quad t = 1, 2, \ldots, n$$

Before, we use MLE $\Rightarrow$ maxmize $L(\underline{\theta}, \theta_0, \sigma^2 | \mathcal{D}) = P(\underline{y} | X, \underline{w})$

Now, we make $\underline{w} \to$ r.v. $\Rightarrow$ MAP $\Rightarrow$ max $P(\underline{w} | X, \underline{y})$

$$\Leftrightarrow \max \frac{P(\underline{w}, \underline{y} | X)}{P(\underline{y} | X)} \Leftrightarrow P(\underline{y} | X, \underline{w}) \cdot P(\underline{w})$$

$$\Leftrightarrow \max \underbrace{P(\underline{y} | X, \underline{w})}_{\text{posterior}} \cdot \underbrace{P(\underline{w})}_{\text{prior}}$$

Penalized log-likelihood (Regularization)

$$\hookrightarrow \ell'(\underline{\theta}, \theta_0, \sigma^2, \gamma) = \text{const} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_t (y_t - \underline{\theta}^T x_t - \theta_0)^2 - \frac{1}{2\gamma}(\theta_0^2 + \sum_{j=1}^d \theta_j^2) - \frac{d+1}{2}\log(\gamma)$$

Prior of $(\underline{\theta}, \theta_0) \rightsquigarrow \mathcal{N}([\frac{\underline{\theta}}{\theta_0}]; [\frac{0}{0}], \gamma I)$

Suppose σ^2 & γ are coupled as $\lambda = \frac{\sigma^2}{\gamma}$ (FIXED)

$$= \text{const} - \frac{n+d+1}{2}\log(\sigma^2) + \frac{d+1}{2}\log\left(\frac{\sigma^2}{\gamma}\right) - \frac{1}{2\sigma^2}\left[\sum_t (y_t - \underline{\theta}^T x_t - \theta_0)^2 + \lambda(\theta_0^2 + \sum_{j=1}^d \theta_j^2)\right]$$

$\frac{\sigma^2}{\gamma} \searrow \lambda$

Note that we have NO γ!

a function $f^* = f^*(\underline{\theta}, \theta_0, \sigma^2, \lambda)$

Also have the decoupled structure of $(\underline{\theta}, \theta_0)$

$\hookrightarrow$ We have the solution: $\begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\theta}_0 \end{bmatrix} = \underbrace{(X^T X + \lambda I)^{-1} X^T \underline{y}}_{\text{guarantee } (X^T X + \lambda I) \text{ invertible}}$

Q1: But what's the BIAS? Suppose dataset is generated from the noisy linear model: $y_t = \langle \underline{\theta}^*, x_t \rangle + \theta_0^* + \varepsilon_t$

$\hookrightarrow$ Answer is NO

$$\mathbb{E}\left[\begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\theta}_0 \end{bmatrix} \middle| X\right] = \mathbb{E}\left[(X^T X + \lambda I)^{-1} X^T \left(X \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix} + \underline{\varepsilon}\right) \middle| X\right]$$

$$= (X^T X + \lambda I)^{-1} X^T X \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix}$$

$$= (X^T X + \lambda I)^{-1} (X^T X + \lambda I - \lambda I) \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix}$$

$$= \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix} - \lambda (X^T X + \lambda I)^{-1} \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix}$$

BIAS.

Conclusion: If $\lambda \neq 0$, then BIAS $\neq 0$.

Q2: Recall MSE = bias^2 + variance

$$= \begin{cases} 0 + \sigma^2 \operatorname{tr}[(X^TX)^{-1}], & \lambda = 0 \\ \text{some bias} + \text{small variance}, & \lambda \neq 0 \end{cases}$$

Review

Linear regression → Unregularized Case → High dimension curse!

- Model.
- Solution $(X^TX)^{-1}X\underline{y}$
- Property of Solution
  - Unbiased.
  - $\operatorname{cov}(\hat{\theta}) = \sigma^2(X^TX)^{-1}$

Bias-Variance Trade-off

↳ MSE = bias^2 + variance

↳ In no regularization case:

- bias = 0
- variance = $\operatorname{tr}(\sigma^2(X^TX)^{-1})$ ↓ This may be quite big

To deal with the high dimension curse

Regularization Case.

⇐ By introducing a prior distribution over $(\underline{\theta}, \theta_0)$

- Model
- Solution → $(X^TX + \lambda I)^{-1}X^T\underline{y}$
- Property →
  - Biased.
  - Smaller variance

⇒ Bias-Variance Tradeoff

① The calculation of Penalized - LR.

$$\text{const} - \frac{n+d+1}{2}\log(\sigma^2) + \frac{d+1}{2}\log\left(\frac{\sigma^2}{\gamma}\right) - \frac{1}{2\sigma^2}\left[\sum_t (y_t - \underline{\theta}^T \underline{x}_t - \theta_0)^2 + \lambda(\theta_0^2 + \sum_{j=1}^{d}\theta_j^2)\right].$$

$\gamma \searrow \lambda$

$$\min \sum_t (y_t - \underline{\theta}^T \underline{x}_t - \theta_0)^2 + \lambda\left[\sum \theta_j^2 + \theta_0^2\right]$$

$$\Leftrightarrow \min \|y - X\theta\|_2^2 + \lambda \|\theta\|^2$$

$$\Leftrightarrow \min (y - X\theta)^T (y - X\theta) + \lambda \theta^T \theta$$

$$\Leftrightarrow \min y^T y - 2y^T X\theta + \theta^T X^T X\theta + \lambda \theta^T \theta.$$

$$\Leftrightarrow \min y^T y - 2\theta^T X^T y + \theta^T (X^T X + \lambda I)\theta$$

$$\Rightarrow \frac{\partial I}{\partial \theta} = 0 \Rightarrow \theta = (X^T X + \lambda I)^{-1} X^T y$$

② $\hat{\theta} = (X^T X + \lambda I)^{-1} X^T y$

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

$= \underbrace{(X^T X + \lambda I)^{-1} X^T X \theta^*}_{\text{Expectation}} + \underbrace{(X^T X + \lambda I)^{-1} X^T \varepsilon}_{\text{Random}}$

Expectation $\Rightarrow$ $\theta^* - \underbrace{\lambda (X^T X + \lambda I)^{-1} \theta^*}_{\text{Bias}}$

Random $\Rightarrow$ Cov

Bias