1.(a) I did this homework with Luning Zhao and Lei Teng.

I did all the problems on my own

(b) I rertify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

2. (a) If Charlie has m types of cards, the probability he gets another type of card is $P = \frac{40-m}{40}$, so the expected number of days it takes is $\epsilon = \frac{1}{P} = \frac{40}{40-m}$.

To get all the 40 types card, it takes: $\sum_{i=0}^{39} \frac{40}{40-i} = 40 \sum_{i=0}^{39} \frac{1}{40-i} = 171$ days.

(b) Assume Charlie has x distinct card types at the end of the game, the probability of having the type that Willy will draw among d card types is: $\frac{x}{d}$.

If Charlie wants his winning probability to be at least $1-\delta$:

$\frac{x}{d} \geq 1-\delta$, so $x \geq d(1-\delta)$

So he should have $d(1-\delta)$ distinct card types.

(c) Assume Willy randomly draw on the type A card, then the probability that Charlie wins a prize = the probability that Charlie has type A card.

= 1- the probability that Charlie doesn't have type A card.

For each day, the probability that Charlie doesn't have type A card is $\frac{d-1}{d}$

so the probability that Charlie doesn't have type A card $= (\frac{d-1}{d})^n$.

so the probability that Charlie wins a prize $= 1-(\frac{d-1}{d})^n$

(d) When $n = \alpha d$, $P(\text{winning}) = 1-(1-\frac{1}{d})^{\alpha d}$.

Assume $y = (1-\frac{1}{d})^{\alpha d}$ $x = \frac{1}{d}$.

When $d \to \infty$, $x \to 0$.

$y|_{x\to 0} = (1-x)^{\frac{\alpha}{x}}$

$$\ln y = \frac{d}{x}\ln(1-x) = d\,\frac{\ln(1-x)}{x}.$$

using l'hopital's rule, $\ln y|_{x\to 0} = d\,\frac{\ln(1-x)}{x}\Big|_{x\to 0}$

$$= d\,\frac{\frac{-1}{1-x}}{1}\Big|_{x\to 0}$$

$$= -d.$$

so $y|_{x\to 0} = e^{-d}.$

so $P(\text{winning})_{d\to\infty} = 1-e^{-d}.$

(e) Assume our dataset has a size of x,

then P(successfully estimate the function at a random point)

$= 1- (1-\frac{1}{d})^{x}$ [similar as question (c)]

$$1-(1-\frac{1}{d})^{x} \geq 1-\delta$$

$$(1-\frac{1}{d})^{x} \leq \delta$$

$$x\ln(1-\frac{1}{d}) \leq \ln\delta$$

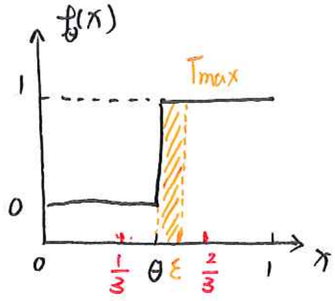
$$x \geq \frac{\ln\delta}{\ln(1-\frac{1}{d})}$$

So we need a training set at least at a size of $\frac{\ln\delta}{\ln(1-\frac{1}{d})}$.

When $d\to\infty$, $\frac{1}{d}\to 0$, $1-\frac{1}{d}\to 1$, $\ln(1-\frac{1}{d})\to 0$

So $\frac{\ln\delta}{\ln(1-\frac{1}{d})} \to \infty.$

3. (a)

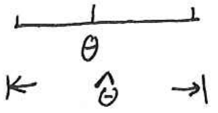


The probability that $T_{max} - \theta > \varepsilon$ is that no point is in $[\theta, \theta+\varepsilon]$ as showed in the yellow region.

so $$P(T_{max} - \theta > \varepsilon) = \begin{cases} (1-\varepsilon)^n & (\varepsilon < \frac{2}{3} - \theta) \\ 0 & (\varepsilon \geq \frac{2}{3} - \theta) \end{cases}$$

Similarly, $$P(\theta - T_{min} > \varepsilon) = \begin{cases} (1-\varepsilon)^n & (\varepsilon < \theta - \frac{1}{3}) \\ 0 & (\varepsilon \geq \theta - \frac{1}{3}) \end{cases}$$

(b) $P(|\hat{\theta} - \theta_0| < \varepsilon)$ represents the probability that the estimation $\hat{\theta}$ is within ε to the true value θ_0.



$$\begin{aligned} & P(|\hat{\theta} - \theta_0| < \varepsilon) \\ = & P(T_{max} - \theta_0 < \varepsilon) \cdot P(\theta_0 - T_{min} < \varepsilon) \\ = & [1 - P(T_{max} - \theta_0 > \varepsilon)] \cdot [1 - P(\theta_0 - T_{min} > \varepsilon)] \\ = & [1 - (1-\varepsilon)^n] \cdot [1 - (1-\varepsilon)^n] \\ = & [1 - (1-\varepsilon)^n]^2 \end{aligned}$$

When $\varepsilon \ll 1$, $(1-\varepsilon)^n = 1 - n\varepsilon$ (only keep the 1st order term in Taylor expansion).

$$P(|\hat{\theta} - \theta_0| < \varepsilon) = [1 - (1-n\varepsilon)]^2 = n^2 \varepsilon^2.$$

If we want $P(|\hat{\theta} - \theta_0| < \varepsilon) \geq 1 - \delta$, then $n^2 \varepsilon^2 \geq 1 - \delta$.

$$n \geq \frac{\sqrt{1-\delta}}{\varepsilon}$$

(c) Because we know $\theta \in [\frac{1}{3}, \frac{2}{3}]$, we could place n Xs evenly between $[\frac{1}{3}, \frac{2}{3}]$. Then the distance between 2 adjascent points in $\frac{1}{3(n+1)}$

$$P(|\hat{\theta} - \theta_0| < \varepsilon) = \frac{2\varepsilon}{\frac{1}{3(n+1)}} = 6\varepsilon(n+1) \quad (\varepsilon n < \frac{1}{6})$$

$$= 1 \quad (\varepsilon n \geq \frac{1}{6})$$

$P(|\hat{\theta}-\theta_0|<\varepsilon) \geq 1-\delta \Rightarrow \quad 6\varepsilon(n+1) \geq 1-\delta$

$$n \geq \frac{1-\delta}{6\varepsilon} - 1$$

(d) If we could sample adaptively, we could first put a point in the middle of $[\frac{1}{3}, \frac{2}{3}]$. Then, with this point, we can calculate Tmin and Tmax. Next point will be put in the middle of [Tmin, Tmax]. Repeat this process for n times.

The length between Tmin and Tmax after n times is: $\frac{1}{3 \cdot 2^n}$

$$P(|\hat{\theta}-\theta_0|<\varepsilon) = \frac{2\varepsilon}{\frac{1}{3\cdot 2^n}} = 3\varepsilon \cdot 2^{n+1}$$

$P(|\hat{\theta}-\theta_0|<\varepsilon) \geq 1-\delta \Rightarrow \quad 3\varepsilon \cdot 2^{n+1} \geq 1-\delta$

$$n \geq \log_2\left(\frac{1-\delta}{3\varepsilon}\right) - 1$$

(e) For random case, $n \propto \frac{1}{\varepsilon}$

For deterministic case, $n \propto \frac{1}{\varepsilon}$

For adaptive case, $n \propto \log_2 \frac{1}{\varepsilon}$

So when $\varepsilon\downarrow$, $n\uparrow$ propotionally for random and deterministic case. but $n\uparrow$ much slower as logarithmly for adaptive case.

For random case, $n \propto \sqrt{1-\delta}$, when $\delta\downarrow$, $n\uparrow$ as $\sqrt{1-\delta}$

(f) Adaptive method is more officient at constraining models.
So in Machine Learning, we should adjust our model based on new data.
This will help us constrain our model faster with fewer data.

4. (a) $X^T A = \lambda X^T$

$(X^T A)^T = (\lambda X^T)^T$

$A^T X = \lambda X$

so the left eigenvalues and eigenvectors for A is the same as right eigenvalues and eigenvectors for A^T.

(i) ■ right, eigenvalues and eigenvectors :

$A x = \lambda x$ $\quad \begin{vmatrix} 2-\lambda & -4 \\ -1 & -1-\lambda \end{vmatrix} = -(2-\lambda)(1+\lambda) - 4 = 0$

$(A - \lambda I) x = 0$

- $\lambda_1 = 3$ $\quad \begin{bmatrix} -1 & -4 \\ -1 & -4 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 4 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{4}{\sqrt{17}} \\ -\frac{1}{\sqrt{17}} \end{bmatrix}$

- $\lambda_2 = -2$ $\quad \begin{bmatrix} 4 & -4 \\ -1 & 1 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

■ left eigenvalues and eigenvectors :

$A^T = \begin{bmatrix} 2 & -1 \\ -4 & -1 \end{bmatrix} \quad \begin{vmatrix} 2-\lambda & -1 \\ -4 & -1-\lambda \end{vmatrix} = -(2-\lambda)(\lambda+1) - 4 = 0$

- $\lambda_1 = 3$ $\quad \begin{bmatrix} -1 & -1 \\ -4 & -4 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

- $\lambda_2 = -2$ $\quad \begin{bmatrix} 4 & -1 \\ -4 & 1 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{17}} \\ \frac{4}{\sqrt{17}} \end{bmatrix}$

(ii) ■ right eigenvalues and eigenvectors :

$B = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \quad \begin{vmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1 = 0$

- $\lambda_1 = 2$ $\quad \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

- $\lambda_2 = 4$ $\quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

▪ left eigenvalues and eigenvectors :

because B is a symmetric matrix, $B^T = B$.

so the left eigenvalues and eigenvectors for B is the same as right ones.

(iii) ▪ left eigenvalues and eigenvectors:

$Ax = \lambda x$, here, λ and x are eigenvalues and eigenvectors for A.

$AAx = A\lambda x = \lambda Ax = \lambda^2 x$.

$A^2 x = \lambda^2 x$.

so evals for A^2 is $\lambda_1 = 3^2 = 9$, $x_1 = \begin{bmatrix} \frac{4}{\sqrt{17}} \\ -\frac{1}{\sqrt{17}} \end{bmatrix}$

$\lambda_2 = (-2)^2 = 4$, $x_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

▪ right evals and evecs : $\lambda_1 = 3^2 = 9$, $x_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

$\lambda_2 = (-2)^2 = 4$, $x_2 = \begin{bmatrix} \frac{1}{\sqrt{17}} \\ \frac{4}{\sqrt{17}} \end{bmatrix}$

(iv) ▪ left eigenpairs : $\lambda_1 = 2^2 = 4$, $x_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

$\lambda_2 = 4^2 = 16$, $x_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

▪ right eigenpairs : same as left ones.

(v) • left pairs:

$AB = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$

$= \begin{bmatrix} 2 & -10 \\ -4 & -4 \end{bmatrix}$

$\begin{vmatrix} 2-\lambda & -10 \\ -4 & -4-\lambda \end{vmatrix} = (2-\lambda)(-4-\lambda) - 40 = 0$

• $\lambda_1 = 6$ $\begin{bmatrix} -4 & -10 \\ -4 & -10 \end{bmatrix} x = 0$ $x = \begin{bmatrix} 5 \\ -2 \end{bmatrix} = \begin{bmatrix} \frac{5}{\sqrt{29}} \\ -\frac{2}{\sqrt{29}} \end{bmatrix}$

• $\lambda_2 = -8$ $\begin{bmatrix} 10 & -10 \\ -4 & 4 \end{bmatrix} x = 0$ $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

▪ right pairs:

$(AB)^T = \begin{bmatrix} 2 & -4 \\ -10 & -4 \end{bmatrix}$ $\qquad \begin{vmatrix} 2-\lambda & -4 \\ -10 & -4-\lambda \end{vmatrix} = -(2-\lambda)(4+\lambda) - 40 = 0$

- $\lambda_1 = 6 \quad \begin{bmatrix} -4 & -4 \\ -10 & -10 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

- $\lambda_2 = -8 \quad \begin{bmatrix} 10 & -4 \\ -10 & 4 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{29}} \\ \frac{5}{\sqrt{29}} \end{bmatrix}$

(vi) ▪ right pairs:

$BA = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$ $\qquad \begin{vmatrix} 5-\lambda & -13 \\ -1 & -7-\lambda \end{vmatrix} = -(5-\lambda)(7+\lambda) - 13 = 0$

$= \begin{bmatrix} 5 & -13 \\ -1 & -7 \end{bmatrix}$

- $\lambda_1 = 6 \quad \begin{bmatrix} -1 & -13 \\ -1 & -13 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 13 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{13}{\sqrt{170}} \\ -\frac{1}{\sqrt{170}} \end{bmatrix}$

- $\lambda_2 = -8 \quad \begin{bmatrix} 13 & -13 \\ -1 & 1 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

▪ left pairs:

$(BA)^T = \begin{bmatrix} 5 & -1 \\ -13 & -7 \end{bmatrix}$

- $\lambda_1 = 6 \quad \begin{bmatrix} -1 & -1 \\ -13 & -13 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

- $\lambda_2 = -8 \quad \begin{bmatrix} 13 & -1 \\ -13 & 1 \end{bmatrix} x = 0 \quad x = \begin{bmatrix} 1 \\ 13 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{170}} \\ \frac{13}{\sqrt{170}} \end{bmatrix}$

(b) SVD is defined as: $A = U \Sigma V^T$

- $A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$ $\qquad A = \begin{bmatrix} 0.99 & 0.11 \\ 0.11 & -0.99 \end{bmatrix} \begin{bmatrix} 4.50 & 0 \\ 0 & 1.33 \end{bmatrix} \begin{bmatrix} 0.42 & -0.91 \\ 0.91 & 0.42 \end{bmatrix}$

$U = \begin{bmatrix} 0.99 & 0.11 \\ 0.11 & -0.99 \end{bmatrix} \quad V = \begin{bmatrix} 0.42 & 0.91 \\ -0.91 & 0.42 \end{bmatrix}$ singular values: 4.50, 1.33

- $B = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ $\quad B\vec{x}_1 = \lambda_1 \vec{x}_1$, $\quad B\vec{x}_2 = \lambda_2 \vec{x}_2$ $\quad B[\vec{x}_1\ \vec{x}_2] = [\vec{x}_1\ \vec{x}_2]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$

$$B = [\vec{x}_1\ \vec{x}_2]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\left([\vec{x}_1\ \vec{x}_2]\right)^T$$

$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ $\quad V = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ singular values: 2, 4.

- $AA = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 8 & -4 \\ -1 & 5 \end{bmatrix}$ $\quad U = \begin{bmatrix} -0.92 & 0.39 \\ 0.39 & 0.92 \end{bmatrix}$ $\quad V = \begin{bmatrix} -0.81 & 0.59 \\ 0.59 & 0.81 \end{bmatrix}$ singular values: 9.59, 3.76

- $BB = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$ $\quad U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ $\quad V = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ $\quad \Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 16 \end{bmatrix}$

- $AB = \begin{bmatrix} 2 & -10 \\ -4 & -4 \end{bmatrix}$ $\quad \Sigma = \begin{bmatrix} 10.78 & 0 \\ 0 & 4.45 \end{bmatrix}$ $\quad U = \begin{bmatrix} 0.93 & 0.36 \\ 0.36 & -0.93 \end{bmatrix}$ $\quad V = \begin{bmatrix} 0.04 & 0.10 \\ -0.10 & 0.04 \end{bmatrix}$

- $BA = \begin{bmatrix} 5 & -1 \\ -13 & -7 \end{bmatrix}$ $\quad \Sigma = \begin{bmatrix} 15.3 & 0 \\ 0 & 3.14 \end{bmatrix}$ $\quad U = \begin{bmatrix} -0.27 & 0.96 \\ 0.96 & 0.27 \end{bmatrix}$ $\quad V = \begin{bmatrix} -0.91 & 0.42 \\ -0.42 & -0.91 \end{bmatrix}$

- $C = \begin{bmatrix} 3 & 1 \\ 1 & 3 \\ 2 & -4 \\ -1 & -1 \end{bmatrix}$ $\quad \Sigma = \begin{bmatrix} 5.20 & 0 \\ 0 & 3.86 \end{bmatrix}$ $\quad U = \begin{bmatrix} -0.14 & 0.80 \\ -0.56 & 0.32 \\ 0.08 & 0.43 \\ 0.18 & -0.28 \end{bmatrix}$ $\quad V = \begin{bmatrix} -0.08 & 1.00 \\ -1.00 & 0.08 \end{bmatrix}$

(c) If $\lambda, \vec{x}$ is the eval and evec for A, then $A\vec{x} = \lambda\vec{x}$.

So for ith row, $\sum_j A_{ij} x_j = \lambda x_i$

$$\sum_{j\neq i} A_{ij} x_j = \lambda x_i - A_{ii} x_i = (\lambda - A_{ii}) x_i .$$

(d) $(\lambda - A_{ii}) x_i = \sum_{j\neq i} A_{ij} x_j \Rightarrow |\lambda - A_{ii}|\,|x_i| = \sum_{j\neq i} |A_{ij}|\,|x_j|$

because $|x_i| > |x_j|$ for all $j \neq i$, $\sum_{j\neq i} |A_{ij}|\,|x_j| \leq \left(\sum_{j\neq i} |A_{ij}|\right)|x_i|$ $\Bigg\} \Rightarrow |\lambda - A_{ii}| \leq \sum_{j\neq i} |A_{ij}|$

5. (a) This problem is a supervised learning, because the measurements $\vec{x}_i$ is in a sequence, so it is labeled. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function for mapping new examples.

(b) If our training set only $\vec{x} \in R^t$, we want to predict the unobserved target by $\hat{y}^t = \vec{x}^T \hat{w}$ where $\hat{w} \in R^t$. minimizes $\sum_{i=1}^{n} (x_i^T \hat{w} - y_i)^2 = \|Xw - y\|^2$

$$X = \overset{n}{\Bigg\downarrow} \underbrace{\begin{bmatrix} (X_1)_1 & (X_1)_2 \cdots & (X_1)_t \\ \vdots & & \\ (X_n)_1 & (X_n)_2 \cdots & (X_n)_t \end{bmatrix}}_{\xrightarrow{\quad t \quad}} \qquad w = \begin{bmatrix} w_1 \\ \vdots \\ w_t \end{bmatrix} \Bigg\downarrow t \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \Bigg\downarrow n$$

(c) · The way we predict $\hat{x}^t$ from $x^1, x^2, \cdots x^{t-1}$ is similar as (b).

Our training data sets contain $\vec{x} \in R^{t-1}$.

$\hat{x}^t$ is predicted by $\hat{x}^t = [x^1\ x^2 \cdots x^{t-1}]\,\hat{w}$, $\hat{w} \in R^{t-1}$.

$\hat{w}$ minimizes the least-squares training cost $\sum_{i=1}^{n} \|Xw - X^t\|$.

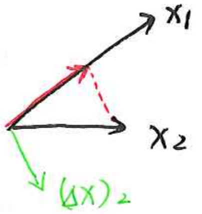
$$X = \begin{bmatrix} (X_1)_1 & (X_1)_2 \cdots & (X_1)_{t-1} \\ \vdots & & \\ (X_n)_1 & (X_n)_2 \cdots & (X_n)_{t-1} \end{bmatrix} \qquad w = \begin{bmatrix} w_1 \\ \vdots \\ w_{t-1} \end{bmatrix} \qquad X^t = \begin{bmatrix} (X_1)_t \\ \vdots \\ (X_n)_t \end{bmatrix}.$$

- $\tilde{y}^t$ then can be calculated by $\tilde{y}^t = \tilde{x}^t A$, where $\tilde{x}^t = (\Delta x^1, \cdots, \Delta x^t)$.
- The predictions of $\hat{y}^t$ is same in one-stage training (b) and two-stage training (c).

When $t=2$, $X'w = X^2$

$$w = (X'^T X')^{-1} X'^T X^2$$

$$\hat{X^2} = X'W = \frac{X'X'^T X^2}{X'^T X'} = \frac{X'^T X^2}{X'^T X'} X'$$



$\hat{x^2}$ is the projected x_2 in x_1.

$(\Delta X)_2 = X^2 - \hat{X^2}$, so $(\Delta X)_2$ is the vector that is perpendicular to x_1.

Similarly, $(\Delta x)_3$ is perpendicular to x_1, x_2.

Thus, linear regression to $[x^1, \cdots x^t]$ is the same as to $[(\Delta x)^1, \cdots, (\Delta x)^t]$.

(d) From $[x^1, x^2]$ to $[(\Delta x)^1, (\Delta x)^2]$ is the Gram-Schmidt process.

Yes, $\tilde{y}^t = \tilde{X}^t w^t \Rightarrow w^t = \underbrace{[(\tilde{X}^t)^T \tilde{X}^t]}_{\text{diagonal matrix}}^{-1} (\tilde{X}^t)^T \tilde{y}^t$

$$= \begin{bmatrix} (\Delta X_1)^2 & 0 & 0 \\ 0 & (\Delta X_2)^2 & 0 \\ & & \ddots \\ 0 & 0 & (\Delta X_t)^2 \end{bmatrix}$$

$$= \begin{bmatrix} A_1 (\Delta X^1)^T \tilde{y}^1 \\ \vdots \\ A_n (\Delta X^t)^T \tilde{y}^t \end{bmatrix}$$

So
$$w^{t-1} = \begin{bmatrix} A_1 (\Delta X^1)^T \tilde{y}^1 \\ \vdots \\ A_n (\Delta X^{t-1})^T \tilde{y}_n^{t-1} \end{bmatrix}$$

If we know w^{t-1}, for row in $[1, t-1]$, $(w^t)_{row} = (w^{t-1})_{row}$ is the same.

and row for t, $(w^t)_{t\,row} = [(\tilde{X}^t)^T \tilde{X}^t]_{t\,row} (\Delta X^t)^T \tilde{y}^t$.

$$= |\Delta x^t|^2 (\Delta x^t)^T \tilde{y}^t$$

(e) To minimize $\|Y - WX\|_F^2$ over w,

we need to calculate $\frac{\partial \|Y - WX\|_F^2}{\partial W} = \frac{\partial tr(Y-WX)^T (Y-WX)}{\partial W}$

Let $f(w) = tr((Y-WX)^T (Y-WX))$

$$= tr(Y^T Y - 2Y^T WX + X^T W^T WX)$$

$$f(W+\Delta) = tr(Y^TY - 2Y^T(W+\Delta)X + X^T(W+\Delta)^T(W+\Delta)X)$$

$$= tr(Y^TY - 2Y^TWX + X^TW^TWX - 2Y^T\Delta X$$

$$+ X^T\Delta^TWX + X^TW^T\Delta X)$$

$$= f(W) + tr[(-2XY^T + 2XX^TW^T)\Delta]$$

So $$\frac{\partial f(W)}{\partial W} = 2(XX^TW^T - XY^T) = 0$$

$$XX^TW^T = XY^T$$

$$WXX^T = YX^T \Rightarrow W = (YX^T)(XX^T)^{-1}$$

(f) because y_n are independent of each other.

6. (a) $x_{t+1} \approx A x_t + B u_t$ can be written as: $\begin{bmatrix} x_1 & u_1 \\ x_2 & u_2 \\ \vdots & \vdots \\ x_{n-1} & u_{n-1} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$.

To $\min_{x} \|Ax - b\|^2$, here $A = \begin{bmatrix} x_1 & u_1 \\ x_2 & u_2 \\ \vdots & \vdots \\ x_{n-1} & u_{n-1} \end{bmatrix}$ $b = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$,

$$\begin{bmatrix} A \\ B \end{bmatrix} = (A^T A)^{-1} A^T b.$$

see code $\Rightarrow$ $A = 0.98$

$B = -0.09$

(b) $\vec{x}_{t+1} \approx A \vec{x}_t + B \vec{u}_t$ can be written as: $[A \quad B] \begin{bmatrix} x_1 \cdots & x_{n-1} \\ u_1 \cdots & u_{n-1} \end{bmatrix} = [x_2 \cdots \quad x_n]$

Let $w = [A \quad B]$ $x = \begin{bmatrix} x_1 \cdots & x_{n-1} \\ u_1 \cdots & u_{n-1} \end{bmatrix}$ $y = [x_2 \cdots \quad x_n]$

To $\min_{w} \|wx - y\|_F^2 = tr((wx-y)^T(wx-y))$

$$\frac{\partial tr(x^T w^T w x - x^T w^T y - y^T w x + y^T y)}{\partial w}$$

$$= \frac{\partial tr(x x^T w^T w)}{\partial w} - 2 \frac{\partial tr(x y^T w)}{\partial w}$$

Let $f(w) = tr(x x^T w^T w)$

$$f(w+\Delta) = tr(x x^T (w+\Delta)^T (w+\Delta))$$

$$= tr(x x^T w^T w) + tr(x x^T (w^T + w^T)\Delta) + \|\Delta\|^2$$

$$\approx f(w) + 2 tr(x x^T w^T \Delta)$$

$$= f(w) + tr(\frac{\partial f}{\partial w} \Delta) \Rightarrow \frac{\partial f}{\partial w} = \frac{\partial tr(x x^T w^T w)}{\partial w} = 2 x x^T w^T$$

Similarly, let $g(w) = tr(xy^T w)$

$$g(w+\Delta) = tr(xy^T(w+\Delta))$$

$$= g(w) + tr(xy^T\Delta) \Rightarrow \frac{\partial tr(xy^T w)}{\partial w} = xy^T$$

So $2xx^Tw^T - 2xy^T = 0$

$$\Rightarrow \quad xx^Tw^T = xy^T$$

$$w = [(xx^T)^{-1}xy^T]^T = yx^T(xx^T)^{-1}$$

see code $\Rightarrow$ $A = \begin{bmatrix} 0.15 & 0.93 & -0.00 \\ 0.04 & 0.31 & 0.87 \\ -0.53 & 0.05 & -0.47 \end{bmatrix}$ $\quad B = \begin{bmatrix} 0.05 & 0.21 & -0.37 \\ -0.05 & -0.93 & 0.13 \\ 0.91 & -0.47 & -0.84 \end{bmatrix}$

(c) $\ddot{x}_i \approx ax_i + b\dot{x}_i + cx_{i-1} + d\dot{x}_{i-1} + e$

can be written as : $\begin{bmatrix} x_1 & \dot{x}_1 & x_{1\text{-prev}} & \dot{x}_{1\text{-prev}} & 1 \\ & & \vdots & & \\ x_n & \dot{x}_n & x_{n\text{-prev}} & \dot{x}_{n\text{-prev}} & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix} = \begin{bmatrix} \ddot{x}_1 \\ \vdots \\ \ddot{x}_n \end{bmatrix}$

Let $X = \begin{bmatrix} x_1 & \dot{x}_1 & x_{1\text{-prev}} & \dot{x}_{1\text{-prev}} & 1 \\ \vdots & & & & \\ x_n & \cdots & & & \end{bmatrix}$ $\quad w = \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix}$ $\quad y = \begin{bmatrix} \ddot{x}_1 \\ \vdots \\ \ddot{x}_n \end{bmatrix}$

$$w = (X^TX)^{-1}X^T y$$

code $\Rightarrow$ $a = -0.012 \quad b = -0.318 \quad c = 0.011 \quad d = 0.275 \quad e = -0.88.$

(d) $\ddot{x}_i = a x_i + b\dot{x}_i + c x_{i-1} + d\dot{x}_{i-1} + e$

$= c(x_{i-1} - x_i) + d(\dot{x}_{i-1} - \dot{x}_i) - (-b-d)(\dot{x}_i - L) + e + (b+d)L$

$(c = -a)$

so $h = c = 0.011$, when $x_{i-1} - x_i \downarrow$, $\ddot{x}_i \downarrow$

meaning when distance to previous car decrease. driver should slow down.

$f = d = 0.275$, when $V_{i-1} - V_i \downarrow$, $a_i \downarrow$

meaning when previous car slow down. driver should slow down too.

$g = -b-d = 0.043$, when $V_i - L > 0$, $a_i \downarrow$

meaning when car's velocity is larger than speed limit. driver should slow down.

$W = e + (b+d)L$, uncertainties that could be controlled by driver.

7. (a) see code

(b) see code

(c) see code.

(d) · We use a seperate test set to evaluate our model performance because this model is trained from training set, but we want to make sure this model is also applicable to more generalized data.
Also training errors always decrease with more model parameters for training set, but it's not necessarily correct for test set.

· The reason that the performance is similar in training set and test set is because we have 11623 sets of data in traing set, and it's much more than our model parameters (40). So we can get a robust estimation of our model.

(e) see code.

Performance:

		[-1, 1]	[0, 1]
without bias	training set:	99.7%	98.9%
	test set:	99.8%	99.1%
with bias	training set:	99.4%	99.4%
	test set:	99.6%	99.6%.

You can see that for the model without bias, regression to [0, 1] is worse than regression to [-1, 1], but for model with bias, they perform comparably.
Adding a bias is helpful when your model has a bias on training set.

8. question: Prove why $(X^TX + \lambda I)$ $(\lambda > 0)$ is invertible.

Answer: First, Let's prove X^TX is positive-semidefinite.

for any vector y, $y^T(X^TX)y = \|Xy\|^2 \geq 0$,

so X^TX is positive-semidefinite.

Then, Let's prove $(X^TX + \lambda I)$ $(\lambda > 0)$ is positive-definite.

for any vector y, $y^T(X^TX + \lambda I)y = \|Xy\|^2 + \lambda\|y\|^2 > 0$

So $X^TX + \lambda I$ is positive definite, therefore it's invertible.