HW1

Mengying Lin. SID: 3038737132

1. Bias-Variance Tradeoff Review

(a) **Show that we can decompose the expected mean squared error into three parts: bias, variance, and irreducible error σ^2:**

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \sigma^2$$

Formally, suppose we have a randomly sampled training set $\mathcal{D}$ (drawn independently of our test data), and we select an estimator denoted $\theta = \hat{\theta}(\mathcal{D})$ (for example, via empirical risk minimization). The expected mean squared error for a test input x can be decomposed as below:

$$\mathbb{E}_{Y\sim p(y|x),\mathcal{D}}[(Y - f_{\hat{\theta}(\mathcal{D})}(x))^2] = \text{Bias}(f_{\hat{\theta}(\mathcal{D})}(x))^2 + \text{Var}(f_{\hat{\theta}(\mathcal{D})}(x)) + \sigma^2$$

You may find it helpful to recall the formulaic definitions of Variance and Bias, reproduced for you below:

$$\text{Bias}(f_{\hat{\theta}(\mathcal{D})}(x)) = \mathbb{E}_{Y\sim p(Y|x),\mathcal{D}}[f_{\hat{\theta}(\mathcal{D})}(x) - Y]$$
$$\text{Var}(f_{\hat{\theta}(\mathcal{D})}(x)) = \mathbb{E}_{\mathcal{D}}\left[(f_{\hat{\theta}(\mathcal{D})}(x) - \mathbb{E}[f_{\hat{\theta}(\mathcal{D})}(x)])^2\right]$$

Pf: For any variable x, we have $\text{Var}(x) = E(x^2) - (E(x))^2$

Let $D[x]$ denote the variance of x.

$R.H.S = \left(E[f_{\hat{\theta}(\mathcal{D})}(x) - Y]\right)^2 + D[f_{\hat{\theta}(\mathcal{D})}(x)] + \sigma^2$

$= E[(f_{\hat{\theta}(\mathcal{D})}(x) - Y)^2] - D[f_{\hat{\theta}(\mathcal{D})}(x) - Y] + D[f_{\hat{\theta}(\mathcal{D})}(x)] + \sigma^2$

$= E[(f_{\hat{\theta}(\mathcal{D})}(x) - Y)^2] - D[f_{\hat{\theta}(\mathcal{D})}(x)] - D[Y] + D[f_{\hat{\theta}(\mathcal{D})}(x)] + \sigma^2$

$= E[(f_{\hat{\theta}(\mathcal{D})}(x) - Y)^2] - D[Y] + \sigma^2.$

Let $\sigma^2 = D[Y]$, we have R.H.S. = L.H.S. Q.E.D.

(b) Suppose our training dataset consists of $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where the only randomness is coming from the label vector $Y = \mathbf{X}\theta^* + \varepsilon$ where θ^* is the true underlying linear model and each noise variable ε_i is i.i.d. with zero mean and variance 1. We use ordinary least squares to estimate a $\hat{\theta}$ from this data. **Calculate the bias and covariance of the $\hat{\theta}$ estimate and use that to compute the bias and variance of the prediction at particular test inputs x.** Recall that the OLS solution is given by

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y,$$

where $X \in \mathbb{R}^{n\times d}$ is our (nonrandom) data matrix, $Y \in \mathbb{R}^n$ is the (random) vector of training targets. For simplicity, assume that $X^\top X$ is diagonal.

For a particular input x:

① bias $= f_{\hat{\theta}}(x) - f_{\theta^*}(x) = x^T(X^TX)^{-1}X^TY - x^T\theta^* + \varepsilon$

② variance $= (f_{\hat{\theta}}(x) - E[f_{\theta^*}(x)])^2 = (x^T(X^TX)^{-1}X^TY - x^T\theta^*)^2$

Consider the corresponding expectations:

$$E[\text{bias}] = E[x^T(X^TX)^{-1}X^TY - x^T\theta^* + \varepsilon]$$

$$= E[x^T((X^TX)^{-1}X^TY - \theta^*)]$$

$$= E[x^T(\hat{\theta} - \theta^*)] = 0.$$

$$E[\text{variance}] = E[(x^T\hat{\theta} - x^T\theta^*)^2] = E[x^T(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T x]$$

$$= x^T E[(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T] x$$

$$= x$$

2. Least Squares and the Min-norm problem from the Perspective of SVD

Consider the equation $X\mathbf{w} = \mathbf{y}$, where $X \in \mathbb{R}^{m \times n}$ is a non-square data matrix, w is a weight vector, and y is vector of labels corresponding to the datapoints in each row of X.

Let's say that $X = U\Sigma V^T$ is the (full) SVD of X. Here, U and V are orthonormal square matrices, and Σ is an $m \times n$ matrix with non-zero singular values (σ_i) on the "diagonal".

For this problem, we define $\Sigma^\dagger$ an $n \times m$ matrix with the reciprocals of the singular values ($\frac{1}{\sigma_i}$) along the "diagonal".

(a) First, consider the case where $m > n$, i.e. our data matrix X has more rows than columns (tall matrix) and the system is overdetermined. **How do we find the weights w that minimizes the error between $X\mathbf{w}$ and y?** In other words, we want to solve $\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$.

a) Let $t(w) = \|Xw - y\| = (Xw-y)^T(Xw-y)$.

$$= (w^TX^T - y^T)(Xw-y)$$

$$= w^TX^TXw - y^TXw - w^TX^Ty + y^Ty.$$

$$\frac{dt}{dw} = 2X^TXw - 2X^Ty.$$

Let $\frac{dt}{dw} = 0$, we have $X^TXw = X^Ty$.

Assume X^TX is nonsingular, we have $\hat{w} = (X^TX)^{-1}X^Ty$.

(b) **Plug in the SVD $X = U\Sigma V^T$ and simplify.** Be careful with dimensions!

b) Plug in $X = U\Sigma V^T$, we have

$$\hat{w} = [(U\Sigma V^T)^T(U\Sigma V^T)]^{-1}(U\Sigma V^T)^Ty$$

$$= (V\Sigma^TU^TU\Sigma V^T)^{-1}(V\Sigma^TU^T)y.$$

Because U and V are orthonormal, $U^TU = I_m$, $V^TV = I_n$,

$$\hat{w} = (V\Sigma^T\Sigma V^T)^{-1}(V\Sigma^TU^T)y$$

$$= \left[V\begin{pmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2\end{pmatrix}V^T\right]^{-1}(V\Sigma^TU^T)y$$

$$= V\begin{pmatrix}\frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n^2}\end{pmatrix}V^TV\Sigma^TU^Ty$$

$$= V\begin{pmatrix}\frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n^2}\end{pmatrix}\Sigma^TU^Ty$$

$$= V\Sigma^\dagger(\Sigma^\dagger)^T\Sigma^TU^Ty$$

$$= V\Sigma^\dagger\begin{pmatrix}I_n & 0\end{pmatrix}U^Ty$$

Denote $\begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_n\end{pmatrix}$ as Σ_n, $\Sigma^\dagger = (\Sigma_n^{-1}\ 0)$

$$\hat{w} = V\begin{pmatrix}\Sigma_n^{-1} & 0\end{pmatrix}\begin{pmatrix}I_n \\ 0\end{pmatrix}U^Ty$$

$= V(\Sigma_n^{-1}\ 0)U^T y$

$= V\Sigma^+ U^T y.$

(c) You'll notice that the least-squares solution is in the form $\mathbf{w}^* = A\mathbf{y}$. **What happens if we left-multiply X by our matrix A?** This is why the matrix A of the least-squares solution is called the left-inverse.

c) $A = V\Sigma^+ U^T$. $AX = V\Sigma^+ U^T U\Sigma V^T = V\Sigma^+\Sigma V^T$

$= V(\Sigma_n^{-1}\ 0)\begin{pmatrix}\Sigma_n\\ 0\end{pmatrix}V^T = V I_n V^T = I_n.$

We get a n-D identity matrix.

(d) Now, let's consider the case where $m < n$, i.e. the data matrix X has more columns than rows and the system is underdetermined. There exist infinitely many solutions for w, but we seek the minimum-norm solution, ie. we want to solve $\min \|\mathbf{w}\|^2 \text{s.t.} X\mathbf{w} = \mathbf{y}$. **What is the minimum norm solution?**

d) Using Lagrange multipliers: $L(w, \lambda) = \frac{1}{2}w^T w - \lambda^T(Xw - y)$.

let $\frac{\partial L}{\partial w} = w - X^T\lambda = 0$

$\begin{cases} X^T\lambda = w. \\ Xw = y \end{cases} \Rightarrow \begin{cases} \lambda = (XX^T)^{-1}y \\ w = X^T(XX^T)^{-1}y. \end{cases}$

(e) **Plug in the SVD $X = U\Sigma V^T$ and simplify.** Be careful with dimensions!

e) $w = V\Sigma^T U^T (U\Sigma V^T V\Sigma^T U^T)^{-1} y$

$= V\Sigma^T [(U\Sigma\Sigma^T U^T)U]^{-1} y$

$= V\Sigma^T [U\Sigma\Sigma^T]^{-1} y$

$= V\Sigma^T (\Sigma\Sigma^T)^{-1} \cdot U^{-1} y$.

$= V\Sigma^T (\Sigma\Sigma^T)^{-1} U^T y$

$= V\Sigma^+ U^T y.$

(f) You'll notice that the min-norm solution is in the form $\mathbf{w}^* = B\mathbf{y}$. **What happens if we right-multiply X by our matrix B?** This is why the matrix B of the min-norm solution is called the right-inverse.

f) $B = V\Sigma^+ U^T$. $XB = U\Sigma V^T V\Sigma^+ U^T = U\Sigma(V^T V)\Sigma^+ U^T = U\Sigma\Sigma^+ U^T$

Denote $\begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_m\end{pmatrix}$ as Σ_m.

$XB = U(\Sigma_m\ 0)\begin{pmatrix}\Sigma_m^{-1}\\ 0\end{pmatrix}U^T$

$= UU^T = I_m$

We get a m-D identity matrix.

3. The 5 Interpretations of Ridge Regression

(a) *Perspective 1: Optimization Problem.* Ridge regression can be understood as the unconstrained optimization problem

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ is a data matrix, and $\mathbf{y} \in \mathbb{R}^n$ is the target vector of measurement values. What's new compared to the simple OLS problem is the addition of the $\lambda\|\mathbf{w}\|^2$ term, which can be interpreted as a "penalty" on the weights being too big.

Use vector calculus to expand the objective and solve this optimization problem for w.

a) Sol: Denote $f(w) = \|y - Xw\|_2^2 + \lambda\|w\|_2^2$

$$= (y-Xw)^T(y-Xw) + \lambda w^T w.$$

$$= (y^T - w^T X^T)(y - Xw) + \lambda w^T w$$

$$= y^T y - w^T X^T y - y^T X w + w^T X^T X w + \lambda w^T w.$$

$$\frac{df(w)}{dw} = 2X^T X w - 2X^T y + 2\lambda w.$$

let $\frac{df}{dw} = 0$. $\quad (X^T X + \lambda I_d) w = X^T y.$

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

(b) *Perspective 2: "Hack" of shifting the Singular Values.* In the previous part, you should have found the optimal w is given by

$n \geq d$.

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

(If you didn't get this, you should check your work for the previous part).

Let $X = U\Sigma V^T$ be the (full) SVD of the X. Recall that U and V are square orthonormal (norm-preserving) matrices, and Σ is a $n \times d$ matrix with singular values σ_i along the "diagonal". **Plug this into the Ridge Regression solution and simplify. What happens when the singular values when $\sigma_i << \lambda$? What about when $\sigma_i >> \lambda$?**

b) Sol: Plug in $X = U\Sigma V^T$. We have:

$$w = ((U\Sigma V^T)^T U\Sigma V^T + 2\lambda I_d)^{-1} (U\Sigma V^T)^T y$$

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

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

When $\sigma_i \ll \lambda$: $w \approx [2\lambda I_d]^{-1} V\Sigma^T U^T y = \frac{1}{2\lambda} I_d V\Sigma^T U^T y = \frac{1}{2\lambda} V\Sigma^T U^T y.$

Compared with the soln $\hat{w} = V\Sigma^{\dagger} U^T y$ in 1(b),

its components are likely to be smaller due to $\frac{1}{s_i}$.

when $\sigma_i >> \lambda$: $w \approx V\Sigma^{+}U^T y$, which is consistent with the soln in 1(b).

Generally λ controls how much constraint we want to put on the magnificance of w.

(c) *Perspective 3: Maximum A Posteriori (MAP) estimation.* Ridge Regression can be viewed as finding the MAP estimate when we apply a prior on the (now viewed as random parameters) $\mathbf{W}$. In particular, we can think of the prior for $\mathbf{W}$ as being $\mathcal{N}(\mathbf{0}, I)$ and view the random Y as being generated using $Y = \mathbf{x}^T\mathbf{W} + \sqrt{\lambda}N$ where the noise N is distributed iid (across training samples) as $\mathcal{N}(0, 1)$. At the vector level, we have $\mathbf{Y} = X\mathbf{W} + \sqrt{\lambda}\mathbf{N}$. Note that the X matrix whose rows are the n different training points are not random.

Show that (1) is the MAP estimate for $\mathbf{W}$ given an observation $\mathbf{Y} = \mathbf{y}$.

c) Pf: $P(W=w \mid Y=y) = \dfrac{P(Y=y \mid W=w)\,P(W=w)}{P(Y=y)}$.

Since $P(Y=y)$ is a constant, we only need to consider maximize

$P(Y=y \mid W=w)\,P(W=w)$ (denoted as $f(w)$).

$P(XW+\sqrt{\lambda}N = y \mid W=w) = P\left(N = \dfrac{y-Xw}{\sqrt{\lambda}} \,\middle|\, W=w\right)$

Note that N and W are independent.

$P\left(N = \dfrac{y-Xw}{\sqrt{\lambda}} \,\middle|\, W=w\right) = P\left(N = \dfrac{y-Xw}{\sqrt{\lambda}}\right)$

$f(w) = P\left(N = \dfrac{y-Xw}{\sqrt{\lambda}}\right) P(W=w)$.

Because $N \sim \mathcal{N}(0, I)$, $W \sim \mathcal{N}(0, I)$.

$$f(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(y-Xw)^T(y-Xw)}{\lambda}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{w^Tw}{2}}.$$

$$\frac{df}{dw} = \frac{1}{2\pi} e^{-\left(\frac{1}{2}\frac{(y-Xw)^T(y-Xw)}{\lambda} + \frac{w^Tw}{2}\right)} \cdot \left[-\frac{X^Ty + X^TXw}{\lambda} + w\right]$$

Let $\frac{df}{dw} = 0$, $\quad -\frac{x^Ty + x^Tx\,w}{\lambda} + w = 0$

$$(x^Tx + \lambda I_n)w = x^Ty.$$

Therefore $w = (x^Tx + \lambda I_n)^{-1} x^Ty$ is the MAP estimate. Q.E.D.

(d) *Perspective 4: Fake Data*. Another way to interpret "ridge regression" is as the ordinary least squares for an augmented data set — i.e. adding a bunch of fake data points to our data. Consider the following augmented measurement vector $\hat{\mathbf{y}}$ and data matrix $\hat{\mathbf{X}}$:

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_d \end{bmatrix} \quad \hat{\mathbf{X}} = \begin{bmatrix} X \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix},$$

where $\mathbf{0}_d$ is the zero vector in $\mathbb{R}^d$ and $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ is the identity matrix. **Show that the classical OLS optimization problem** $\operatorname{argmin}_{\mathbf{w}} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w}\|_2^2$ **has the same minimizer as** (1)**.**

d). Note $\|\hat{y} - \hat{X}w\|_2^2 = \left\| \begin{bmatrix} y \\ 0_d \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda} I_d \end{bmatrix} w \right\|_2^2$

$$= \left\| \begin{bmatrix} y - Xw \\ \sqrt{\lambda} I_d w \end{bmatrix} \right\|_2^2$$

$$= \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

which is of the same form with (1).

(e) *Perspective 5: Fake Features.* For this last interpretation, let's instead construct an augmented design matrix in the following way:

$$\check{\mathbf{X}} = [X \ \sqrt{\lambda}\mathbf{I_n}]$$

i.e. we stack X with $\sqrt{\lambda}\mathbf{I}_n$ horizontally. Now our problem is underdetermined: the new dimension $d+n$ is larger than the number of points n. Therefore, there are infinitely many values $\boldsymbol{\eta} \in \mathbb{R}^{d+n}$ for which $\check{\mathbf{X}}\boldsymbol{\eta} = \mathbf{y}$. We are interested in the **min-norm** solution, ie. the solution to

$$\underset{\boldsymbol{\eta}}{\operatorname{argmin}} \|\boldsymbol{\eta}\|_2^2 \text{ s.t. } \check{\mathbf{X}}\boldsymbol{\eta} = \mathbf{y}. \tag{2}$$

Show that this is yet another form of ridge regression and that the first d coordinates of η^* form the minimizer of (1).

e). Using Lagrange multipliers:

$$L(\eta, \lambda) = \frac{1}{2}\eta^T\eta + \lambda^T(y - \hat{X}\eta)$$

let $\frac{\partial L}{\partial \eta} = \eta - \hat{X}^T\lambda = 0 \quad \cdots ①$

$$y - \hat{X}\eta = 0 \quad \cdots ②$$

From ①.②. we can deduce $\begin{cases} \lambda = (\hat{X}\hat{X}^T)^{-1}y. \\ \eta = \hat{X}^T(\hat{X}\hat{X}^T)^{-1}y \end{cases}$

$$\eta = \begin{bmatrix} X^T \\ \sqrt{\lambda}I_n \end{bmatrix}\left([X \ \sqrt{\lambda}I_n]\begin{bmatrix} X^T \\ \sqrt{\lambda}I_n \end{bmatrix}\right)^{-1}y$$

$$= \begin{bmatrix} X^T \\ \sqrt{\lambda}I_n \end{bmatrix}(XX^T + \lambda I_n)^{-1}y$$

The first d dimention of η is $X^T(XX^T + \lambda I_n)^{-1}y$,

The solu in (1) is $(X^TX + \lambda I_d)^{-1}X^Ty$.

We need to further prove $X^T(XX^T + \lambda I_n)^{-1} = (X^TX + \lambda I_d)^{-1}X^T$

which is $(X^TX + \lambda I_d)X^T(XX^T + \lambda I_n)^{-1} = X^T$

$$\text{L.H.S.} = (X^TXX^T + \lambda X^T)(XX^T + \lambda I_n)^{-1} = X^T(XX^T + \lambda I_n)(XX^T + \lambda I_n)^{-1}$$

$$= X^T.$$

Q.E.D

(g) *We know that the solution to ridge regression (1) is given by* $\hat{\mathbf{w}}_\mathbf{r} = (X^\top X + \lambda\mathbf{I})^{-1}X^\top\mathbf{y}$. **What happens when** $\lambda \to \infty$**?** *It is for this reason that sometimes ridge regularization is referred to as "shrinkage."*

when $\lambda \to \infty$. $\hat{w_r} \to (\lambda I)^{-1}X^Ty = \lambda^{-1}X^Ty$.

Thus $\|\hat{w_r}\| = \frac{1}{\lambda}\|X^Ty\| \to 0$.

(h) **What happens to the solution of ridge regression when you take the limit $\lambda \to 0$?** *Consider both the cases when X is wide (underdetermined system) and X is tall (overdetermined system).*

when $\lambda \to 0$, $\hat{w}_r \to (X^TX)^{-1}X^Ty$.

The solution becomes the OLS solution no matter X is tall or wide.

(f) We know that the Moore-Penrose pseudo-inverse for an underdetermined system (wide matrix) is given by $A^\dagger = A^T(AA^T)^{-1}$, which corresponds to the min-norm solution for $A\eta = \mathbf{z}$. That is, the optimization problem

$$\text{argmin } \|\eta\|^2 \text{s.t.} A\eta = \mathbf{z}$$

is solved by $\eta = A^\dagger \mathbf{z}$. Let $\hat{\mathbf{w}}$ be the minimizer of (1).

Use the pseudo-inverse to show that solving to the optimization problem in (2) yields

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

Then, show that this is equivalent to the standard formula for Ridge Regression

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

pf: ① Use pseudo-inverse:

$$\eta = \hat{X}^T(\hat{X}\hat{X}^T)^{-1}y$$
$$= \begin{bmatrix} X^T \\ \sqrt{\lambda} I_n \end{bmatrix} \left(\begin{bmatrix} X & \sqrt{\lambda} I_n \end{bmatrix} \begin{bmatrix} X^T \\ \sqrt{\lambda} I_n \end{bmatrix} \right)^{-1} y$$
$$= \begin{bmatrix} X^T \\ \sqrt{\lambda} I_n \end{bmatrix} (XX^T + \lambda I_n)^{-1} y.$$
$$= \begin{bmatrix} X^T(XX^T + \lambda I_n)^{-1}y \\ \sqrt{\lambda}(XX^T + \lambda I_n)^{-1}y \end{bmatrix}$$

Therefore $\hat{w} = X^T(XX^T + \lambda I_n)^{-1}y$.

② To prove $(X^TX + \lambda I_n)^{-1}X^Ty = X^T(XX^T + \lambda I_n)^{-1}y$,

we can first prove $(X^TX + \lambda I_n)^{-1}X^T = X^T(XX^T + \lambda I_n)^{-1}$

Since $(X^T)^{-1}(X^TX + \lambda I_n)^{-1}X^T = (XX^TX^T + \lambda X^T)^{-1}X^T$

$$= [X^T(XX^T + \lambda I_n)]^{-1}X^T$$

$$= (XX^T + \lambda I_n)^{-1}(X^T)^{-1}X^T = (XX^T + \lambda I_n)^{-1}.$$

Hence $(X^TX + \lambda I_n)^{-1}X^T = X^T(XX^T + \lambda I_n)^{-1}$

4. General Case Tikhonov Regularization

Consider the optimization problem:

$$\min_{\mathbf{x}} ||W_1(A\mathbf{x} - \mathbf{b})||_2^2 + ||W_2(\mathbf{x} - \mathbf{c})||_2^2$$

Where W_1, A, and W_2 are matrices and $\mathbf{x}$, $\mathbf{b}$ and $\mathbf{c}$ are vectors. W_1 can be viewed as a generic weighting of the residuals and W_2 along with c can be viewed as a generic weighting of the parameters.

(a) **Solve this optimization problem manually** by expanding it out as matrix-vector products, setting the gradient to $\mathbf{0}$, and solving for $\mathbf{x}$.

$f(x) = (Ax-b)^T W_1^T W_1 (Ax-b) + (x-c)^T W_2^T W_2 (x-c)$.

$= (x^T A^T - b^T) W_1^T W_1 (Ax-b) + (x^T - c^T) W_2^T W_2 (x-c)$.

$\frac{df}{dx} = 2A^T W_1^T W_1 Ax - 2A^T W_1^T W_1 b + 2W_2^T W_2 x - 2W_2^T W_2 c$.

Let $\frac{df}{dx} = 0$. $(A^T W_1^T W_1 A + W_2^T W_2) x = A^T W_1^T W_1 b + W_2^T W_2 c$

Assume $A^T W_1^T W_1 A + W_2^T W_2$ is invertible,

$x = (A^T W_1^T W_1 A + W_2^T W_2)^{-1} (A^T W_1^T W_1 b + W_2^T W_2 c)$

(b) **Construct an appropriate matrix $\mathbf{C}$ and vector $\mathbf{d}$ that allows you to rewrite this problem as**

$$\min_x ||C\mathbf{x} - \mathbf{d}||^2$$

and use the OLS solution ($\mathbf{x}^* = (C^T C)^{-1} C^T \mathbf{d}$) **to solve.** Confirm your answer is in agreement with the previous part.

$f(x) = x^T (A^T W_1^T W_1 A + W_2^T W_2) x$

$- (b^T W_1^T W_1 A + c^T W_2^T W_2) x - x^T (A^T W_1^T W_1 b + W_2^T W_2 c) + b^T b$

$(Cx-d)^T (Cx-d) = x^T C^T C x - d^T C x - x^T C^T d + d^T d$

let $\begin{cases} A^T W_1^T W_1 A + W_2^T W_2 = C^T C \\ b^T W_1^T W_1 A + c^T W_2^T W_2 = d^T C. \end{cases}$

[I have no idea how to solve them].

Therefore $x^* = (C^T C)^{-1} C^T d = (A^T W_1^T W_1 A + W_2^T W_2)^{-1} (b^T W_1^T W_1 A + c^T W_2^T W_2)^T$

$= (A^T W_1^T W_1 A + W_2^T W_2)^{-1} (A^T W_1^T W_1 b + W_2^T W_2 c)$

(c) **Choose a W_1, W_2, and $\mathbf{c}$** such that this reduces to the simple case of ridge regression that you've seen in the previous problem, $\mathbf{x}^* = (A^T A + \lambda I)^{-1} A^T b$.

Let $\begin{cases} W_1 = I \\ W_2 = \lambda I \\ c = 0 \end{cases}$, we have $f(x) = \|Ax - b\|_2^2 + \lambda \|x\|_2^2$,

which is the same as 3(a).

5. Coding Fully Connected Networks

In this coding assignment, you will be building a fully-connected neural network from scratch using NumPy. You will have the choice between two options:

Use Google Colab (Recommended). Open this url and follow the instructions in the notebook.

Use a local Conda environment. Clone `https://github.com/gonglinyuan/cs182hw1` and refer to `README.md` for further instructions.

For this question, please submit a .zip file your completed work to the Gradescope assignment titled "Homework 1 (Code)". Please answer the following question in your submission of the written assignment:

(a) **Did you notice anything about the comparative difficulty of training the three-layer net vs training the five layer net?**

① It seems that five layer net is more sensitive to weight-scale, the loss of which may explode even if there is only slight change to the parameters

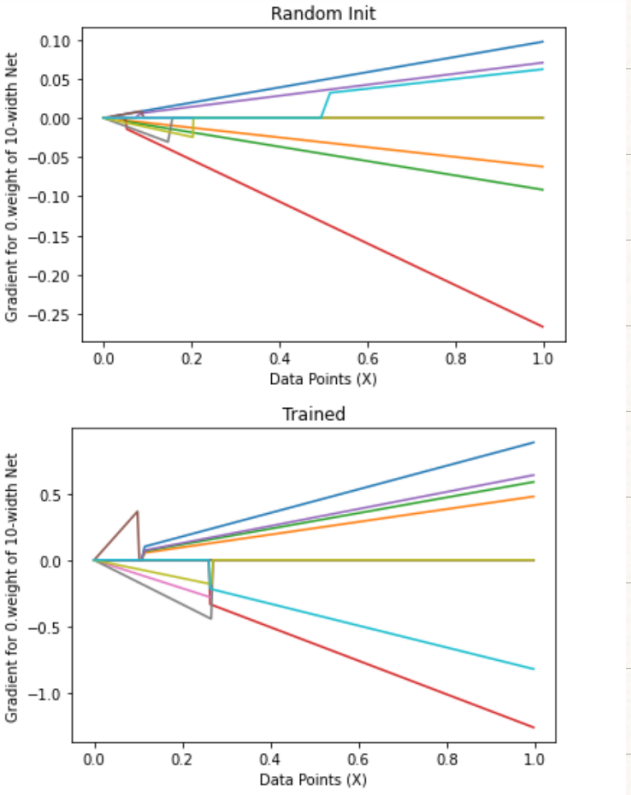
② It takes longer time to train a five layer net.

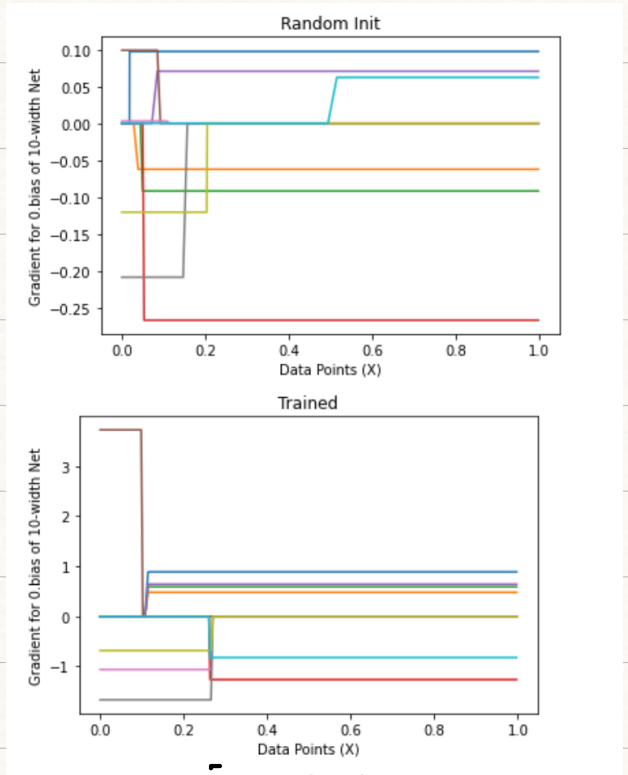
6. Visualizing features from local linearization of neural nets

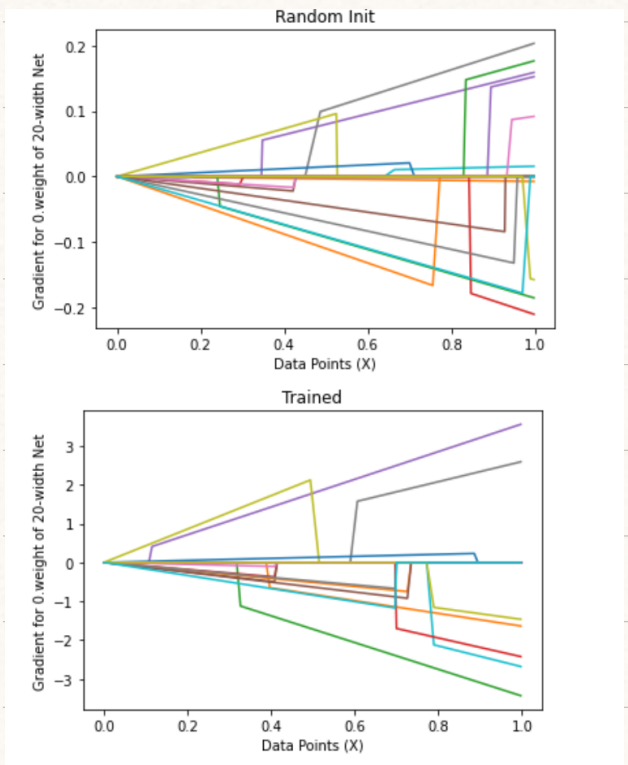
This problem expects you to modify the Jupyter Notebook you were given in the first discussion section for the course to allow the visualization of the effective "features" that correspond to the local linearization of the network in the neighborhood of the parameters.

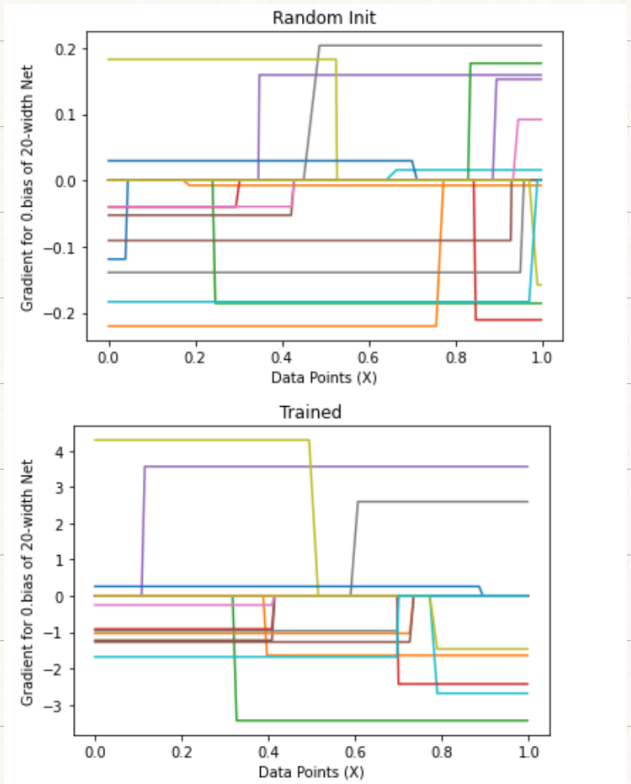
We provide you with some starter code on Google Colab. For this question, **please do *not* submit your code to Gradescope**. Instead, just include your plots and comments regarding the questions in the subparts.

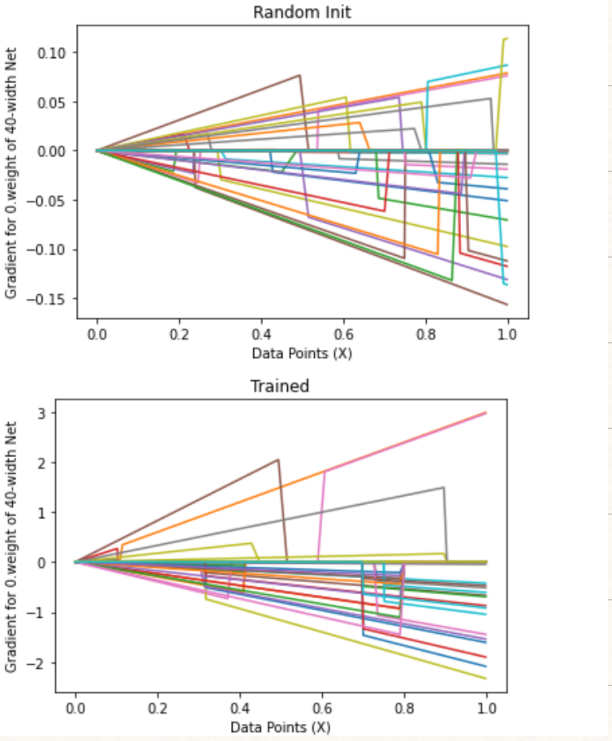
(a) **Visualize the features corresponding to $\frac{\partial}{\partial w_i^{(1)}} y(x)$ and $\frac{\partial}{\partial b_i^{(1)}} y(x)$ where $w_i^{(1)}$ are the first hidden layer's weights and the $b_i^{(1)}$ are the first hidden layer's biases.** These derivatives should be evaluated at at least both the random initialization and the final trained network. When visualizing these features, plot them as a function of the scalar input x, the same way that the notebook plots the constituent "elbow" features that are the outputs of the penultimate layer.

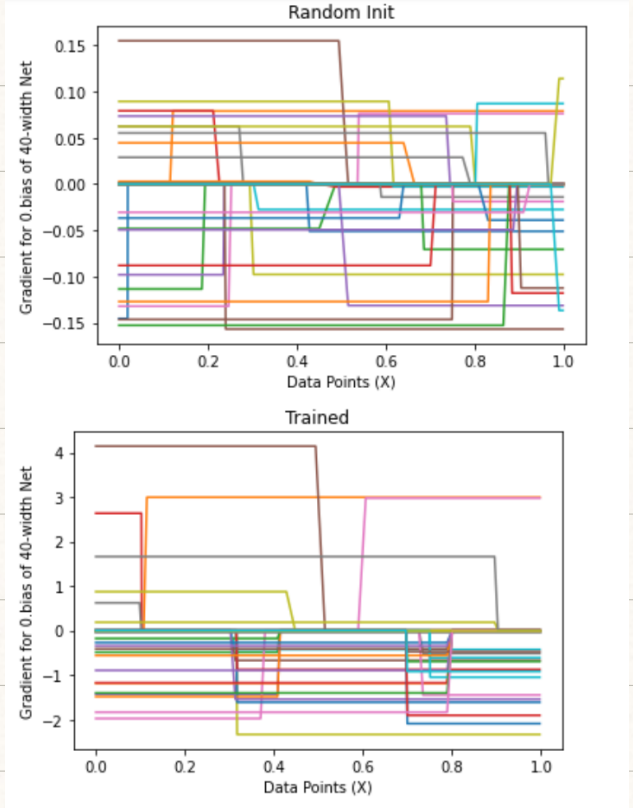












As illustrated, the gradients of $W^{(1)}$ and $b^{(1)}$ are ramdomly distributed, and after trained, they tend to cluster around certain values.

(b) During training, we can imagine that we have a generalized linear model with a feature matrix corresponding to the linearized features corresponding to each learnable parameter. We know from our analysis of gradient descent, that the singular values and singular vectors corresponding to this feature matrix are important.

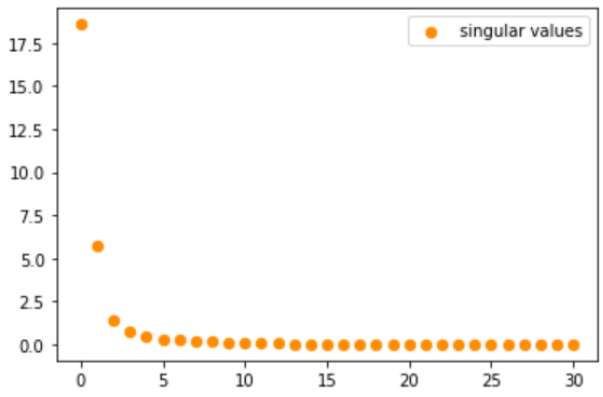
Use the SVD of this feature matrix to plot both the singular values and visualize the "principle features" that correspond to the d-dimensional singular vectors multiplied by all the features corresponding to the parameters.

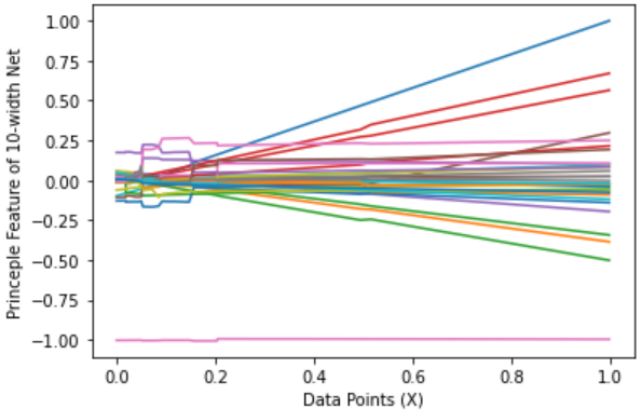
(HINT: Remember that the feature matrix whose SVD you are taking has n rows where each row corresponds to one training point and d columns where each column corresponds to each of the learnable features. Meanwhile, you are going to be plotting/visualizing the "principle features" as functions of x even at places where you don't have training points.)

Consider a $m \times n$ feature matrix X (row: training pts, column: feature vecs).
Perform SVD on X: $X = U\Sigma V^T$. ($U \in \mathbb{R}^{m \times d}$, $\Sigma \in \mathbb{R}^{d \times d}$, $V \in \mathbb{R}^{d \times n}$, $d = \min(m,n)$)
Right multiply X by V, we get a $m \times d$ matrix, which is the matrix containing principle features.

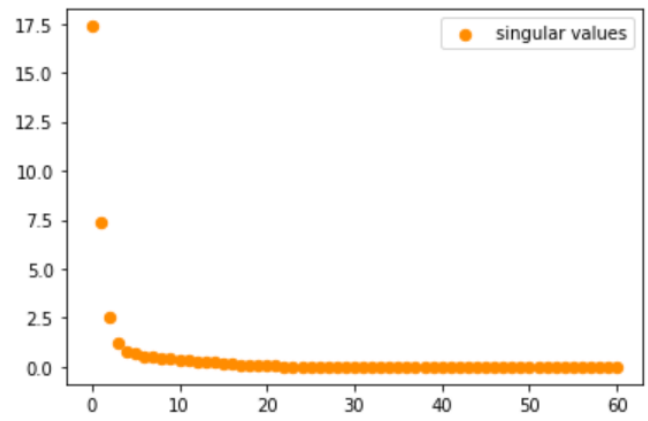
Note: At places where we don't have training pts, we need to corresponding rows with 0.

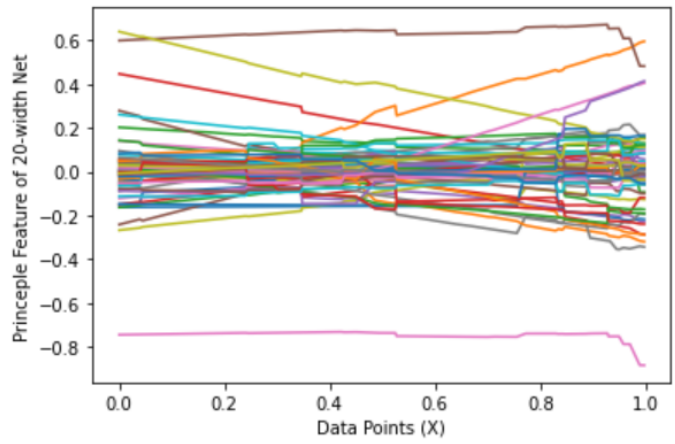
Width 10



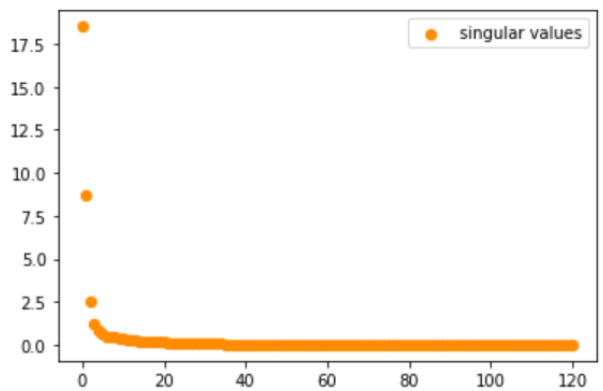


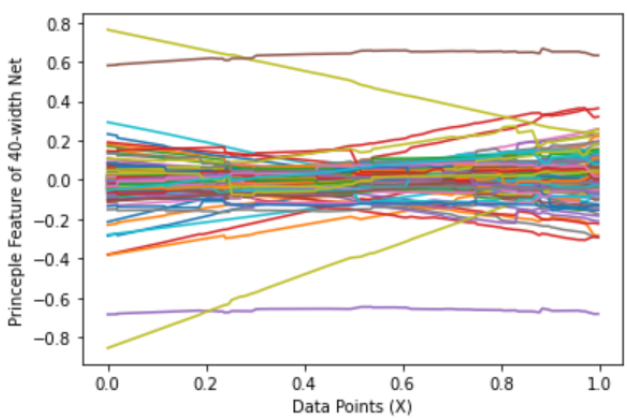
Width 20





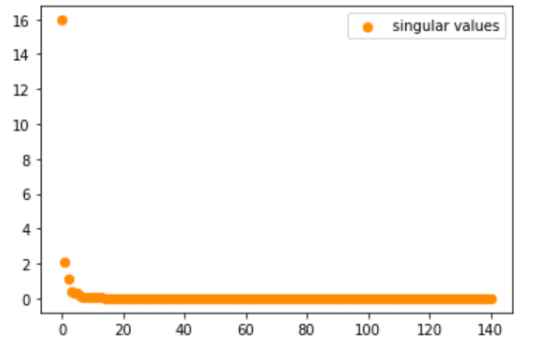
Width 40

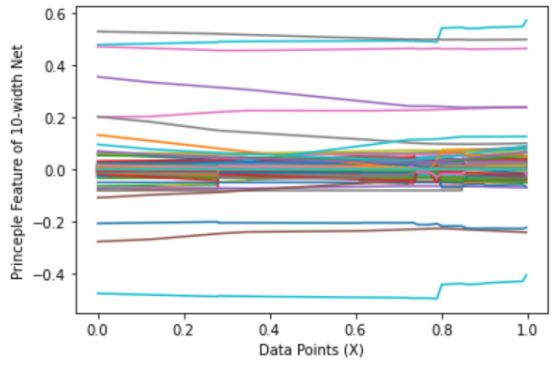




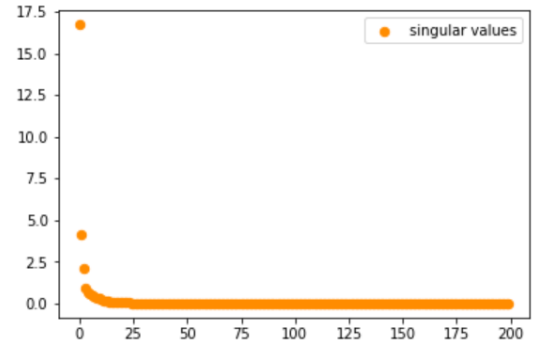
(c) Augment the jupyter notebook to add a second hidden layer of the same size as the first hidden layer, fully connected to the first hidden layer. **Allow the visualization of the features corresponding to the parameters in both hidden layers, as well as the "principle features" and the singular values.**

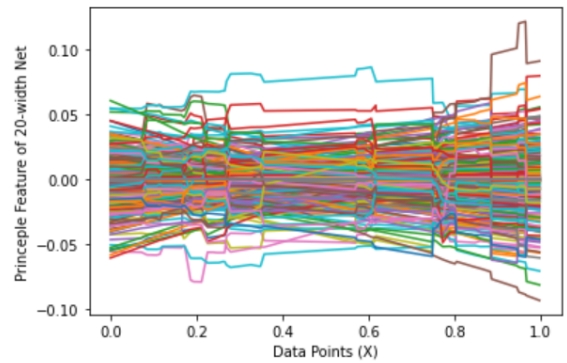
Width 10



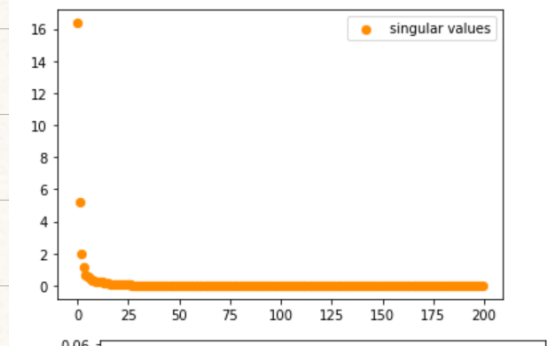


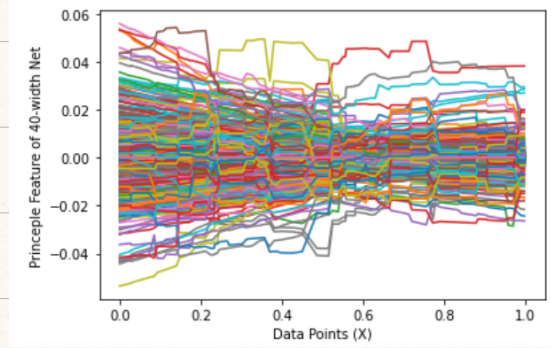
Width 20





Width 40





7. Homework Process and Study Group

Citing sources and collaborators are an important part of life, including being a student!
We also want to understand what resources you find helpful and how much time homework is taking, so we can change things in the future if possible.

(a) **What sources (if any) did you use as you worked through the homework?**

(b) **If you worked with someone on this homework, who did you work with?**
List names and student ID's. (In case of homework party, you can also just describe the group.)

(c) **Roughly how many total hours did you work on this homework? Write it down here where you'll need to remember it for the self-grade form.**

a) https://zhuanlan.zhihu.com/p/29846048

https://numpy.org/doc/stable/reference/routines.linalg.html

b)

Names	Xiang Fei
SID	3038733024.

c) 20 h.