Problem 1 (Regression):

Let \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \) where \( x_i \in \mathbb{R}^d \) and \( y_i \in \mathbb{R} \) be the training data that you are given. To predict \( y \) as \( w^T x \) for some parameter vector \( w \in \mathbb{R}^d \) we can use

The ordinary least square optimization (OLS) problem:

\[
\text{argmin}_w R(w) = \text{argmin}_w \sum_{i=1}^{n} (y_i - w^T x_i)^2. \tag{1}
\]

The ridge regression optimization problem with parameter \( \lambda > 0 \):

\[
\text{argmin}_w R_{\text{ridge}}(w) = \text{argmin}_w \left[ \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda w^T w \right]. \tag{2}
\]

We define the OLS and ridge estimator as, \( \hat{w} = (X^T X)^{-1} X^T y \) and \( \hat{w}_{\text{ridge}}(\lambda) = (X^T X + \lambda I_d)^{-1} X^T y \), respectively.

Regression and Shrinkage

1. Let \( U \Sigma V^T \) be the Singular Value Decomposition (SVD) of \( X \). What is \( \hat{w} \)?

   Here we use the compact SVD. \( X_{n \times d} = U_{n \times r} \Sigma_{r \times r} V_{d \times r}^T \), where \( r \leq \min\{m, n\} \). Assume \( X^T X \) is invertible.

   (a) \( V \Sigma U^T y \)
   (b) \( V \Sigma^{-1} U^T y \)
   (c) \( V \Sigma^{-1} \Sigma U^T y \)
   (d) \( V \Sigma^{-2} U^T y \)

   Solution:

   (b) and (d) are both correct solutions.

   Both the OLS and the ridge estimators can be rewritten in term of the SVD matrices.

   \[
   \hat{w} = (X^T X)^{-1} X^T y = (V \Sigma U^T U \Sigma V^T)^{-1} V \Sigma U^T y = (V \Sigma^2 V^T)^{-1} V \Sigma U^T y = V \Sigma^{-2} V^T V \Sigma U^T y = V \Sigma^{-2} U^T y
   \]
2. What is \( \hat{w}_{\text{ridge}} \)?
   (a) \( V(\Sigma + \lambda I)^{-1}\Sigma U^T y \)
   (b) \( V(\Sigma^2 + \lambda I)^{-1}\Sigma U^T y \)
   (c) \( V(\lambda I)^{-1}\Sigma U^T y \)
   (d) \( V(\Sigma^2 + \lambda I)\Sigma U^T y \)

**Solution:**
The correct answer is (b).

\[
\hat{w}_{\text{ridge}}(\lambda) = (X^TX + \lambda I)^{-1}X^Ty = (V\Sigma^2 V^T + \lambda I)^{-1}V\Sigma U^T y = V(\Sigma^2 + \lambda I)^{-1}V^T V\Sigma U^T y = V(\Sigma^2 + \lambda I)^{-1}\Sigma U^T y
\]

3. The ridge penalty term, \( \lambda w^T w \), :
   (a) shrinks the low variance components
   (b) shrinks the high variance components
   (c) amplifies the low variance components
   (d) does not change the components

**Solution:**
The correct answer is (a).

Writing \( \Sigma_{jj} = d_{jj} \), we have: \( d_{jj}^{-1} \geq \frac{d_{jj}}{d_{jj} + \lambda} \) for all \( \lambda > 0 \)

Thus, the ridge penalty will shrink the singular values and the low variance components will be shrunk to a greater extent.

**Regression and Bias**

4. Compute \( E_{e|X}[\hat{w}] \).
   (a) \( w \)
   (b) \( (X^TX)w \)
   (c) \( (X^TX)^{-1}w \)
   (d) \( 2w \)

**Solution:**
The correct answer is (a).

\[
E_{e|X}[\hat{w}] = E_{e|X}[(X^TX)^{-1}(X^Ty)] = E_{e|X}[(X^TX)^{-1}(X^T(Xw + e))] = E_{e|X}[(w + (X^TX)^{-1}(X^T e))] = w
\]

5. Compute \( E_{e|X}[\hat{w}_{\text{ridge}}] \).
   (a) \( (X^TX + \lambda I)^{-1}(X^TX)w \)
   (b) \( w \)
   (c) \( (X^TX)w \)
\[ \text{(d) } (X^TX - \lambda I)^{-1}(X^TX)w \]

**Solution:**
The correct answer is (a).

\[
\mathbb{E} [\hat{w}_{\text{ridge}}(\lambda)] = \mathbb{E} \left[ (X^TX + \lambda I)^{-1} X^T y \right] \\
= \mathbb{E} \left[ (X^TX + \lambda I)^{-1} (X^TX) (X^TX)^{-1} X^Ty \right] \\
= \mathbb{E} \left[ (X^TX + \lambda I)^{-1} (X^TX) \hat{w} \right] \\
= (X^TX + \lambda I)^{-1} (X^TX) \mathbb{E} (\hat{w}) \\
= (X^TX + \lambda I)^{-1} (X^TX) w
\]

We can see that \( \mathbb{E} [\hat{w}_{\text{ridge}}(\lambda)] \neq w \) for any \( \lambda > 0 \). Hence, the ridge estimator is biased.

6. Pick the true statements.

(a) The Ordinary Least Squares estimator is biased.

(b) The ridge regression estimator is biased.

**Solution:**
Only (b) is True.
We can see that \( \mathbb{E} [\hat{w}_{\text{ridge}}(\lambda)] \neq w \) for any \( \lambda > 0 \). Hence, the ridge estimator is biased.

7. When \( \lambda \to \infty \), all the regression weights converge to:

(a) 1

(b) 0

(c) \( \infty \)

(d) \( \pi \)

**Solution:**
The correct answer is (b).
When \( \lambda \to \infty \) :

\[
\lim_{\lambda \to \infty} \mathbb{E} [\hat{w}_{\text{ridge}}(\lambda)] = \lim_{\lambda \to \infty} (X^TX + \lambda I)^{-1} (X^TX) w = 0_d
\]

All the regression coefficients are shrunken towards zero as the penalty parameter increases.

**Variance of Regression Estimates**

8. Compute the variance of \( \hat{w} \):

\[ Var(AY) = AVar(Y)A^T \]

(a) \( (X^TX)\sigma^2 \)
(b) \((X^T X)^{-1}\sigma^2\)
(c) \(\sigma^2/2\)
(d) \(2\sigma^2\)

Solution:
The correct answer is (b).

\[
\begin{align*}
V ar(\hat{w}) &= V ar((X^T X)^{-1} X^T y) \\
&= V ar((X^T X)^{-1} X^T (Xw + \epsilon)) \\
&= V ar((X^T X)^{-1} X^T (\epsilon)) \\
&= (X^T X)^{-1} X^T V ar(\epsilon) X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{align*}
\]

9. Compute the variance of \(\hat{w}_{\text{ridge}}\).

(a) \(\sigma^2 (X^T X + \lambda I)^{-1} (X^T X) \left[(X^T X + \lambda I)^{-1}\right]^T\)
(b) \(\sigma^2 (X^T X - \lambda I)^{-1} (X^T X) \left[(X^T X - \lambda I)^{-1}\right]^T\)
(c) \(\sigma^2 (X^T X + 2\lambda I)^{-1} (X^T X) \left[(X^T X + 2\lambda I)^{-1}\right]^T\)
(d) \(\sigma^2 (X^T X + \frac{\lambda}{2} I)^{-1} (X^T X) \left[(X^T X + \frac{\lambda}{2} I)^{-1}\right]^T\)

Solution:
The correct answer is (a).

We have: \(\hat{w}_{\text{ridge}}(\lambda) = (X^T X + \lambda I)^{-1} (X^T X) \hat{w}\)
We define: \(\Omega_\lambda = (X^T X + \lambda I)^{-1} (X^T X)\)
It can be seen that,

\[
V ar[\hat{w}_{\text{ridge}}(\lambda)] = V ar[\Omega_\lambda \hat{w}] = \Omega_\lambda V ar[\hat{w}] \Omega_\lambda^T = \sigma^2 \Omega_\lambda (X^T X)^{-1} \Omega_\lambda^T = \sigma^2 (X^T X + \lambda I)^{-1} (X^T X) \left[(X^T X + \lambda I)^{-1}\right]^T
\]

Note that we have used the fact that \(V ar(AX) = AV ar(X)A^T\) for a non random matrix \(A\), and the fact that \(V ar(\hat{w}) = \sigma^2 (X^T X)^{-1}\)

10. \(V ar(\bar{w}) \preceq \text{Var}\bar{w}_{\text{ridge}}\). This statement is:
(Try to prove your statement)

(a) True
(b) False
Solution:
The given statement is False.
Comparing it to the variance of the OLS estimator,
\[
\text{Var}[\hat{w}] - \text{Var}[\hat{w}_{\text{ridge}}(\lambda)] = \sigma^2 \left[ (X^T X)^{-1} - \Omega \left( X^T X \right)^{-1} \Omega^T \right]
\]
\[
= \sigma^2 \Omega \left[ \left( I + \lambda \left( X^T X \right)^{-1} \right) \left( X^T X \right)^{-1} \left( I + \lambda \left( X^T X \right)^{-1} \right)^T - \left( X^T X \right)^{-1} \right] \Omega^T
\]
\[
= \sigma^2 \Omega \left[ 2\lambda \left( X^T X \right)^{-2} + \lambda^2 \left( X^T X \right)^{-3} \right] \Omega^T
\]
\[
= \sigma^2 \left( X^T X + \lambda I \right)^{-1} \left[ 2\lambda I + \lambda^2 \left( X^T X \right)^{-1} \right] \left( X^T X + \lambda I \right)^{-1} \Omega^T
\]
The difference is non-negative definite. Hence, the variance of the OLS estimator exceeds that of the ridge estimator.

\[
\text{Var}[\hat{w}] \geq \text{Var}[\hat{w}_{\text{ridge}}(\lambda)]
\]

11. When \( \lambda \to \infty \), the variance of the ridge estimator,
(a) reduces to zero
(b) converges to 1
(c) increases to \( \infty \)

Solution:
The correct answer is (a).

Now, let us look at the case where \( \lambda \to \infty \):
\[
\lim_{\lambda \to \infty} \text{Var}[\hat{w}_{\text{ridge}}(\lambda)] = \lim_{\lambda \to \infty} \sigma^2 \Omega \left( X^T X \right)^{-1} = 0_d
\]
The variance of the ridge estimator vanishes. Hence, the variance of the ridge regression coefficient estimates decreases towards zero as the penalty parameter becomes large.

**Regularized loss for regression**
In this problem you will help Ada solve a linear regression problem. From the domain experts she has learned that it makes sense to use the following regularizer\(^1\),
\[
R(w) = \sum_{i=1}^{d-1} |w_i - w_{i+1}|
\]
for the weight vector \( w \in \mathbb{R}^d \). She is given \( n \) data points \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), where each \( x_i \in \mathbb{R}^d \) and each \( y_i \in \mathbb{R} \). Hence, she has to minimize the following objective
\[
f(w) = \frac{1}{n} \sum_{i=1}^{n} (w_i^T x_i - y_i)^2 + \lambda R(w).
\]

12. Ada wrote a program and then solved the above problem for the same data points and four different positive penalizers \( \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 \). Unfortunately, she has misnamed the files holding the results and does not know which file corresponds to which \( \lambda_i \). Your task is to help Ada by assigning to each file the corresponding \( \lambda_i \) that was used. Try to justify your answer.

Match the following computed weight vectors, \( w^* \), to the corresponding \( \lambda_i \) used.

---
\(^1\)This regularizer makes sense if we would like to prefer solutions whose entries do not change much between adjacent coordinates.
Take any $w$ and $w'$ satisfying $R(w) < R(w')$ that are optimal for some $\lambda \neq \lambda'$. Then, because they are optimal for the corresponding losses

$$L(w) + \lambda R(w) \leq L(w') + \lambda R(w'), \quad \text{and}$$

$$-L(w) - \lambda' R(w) \leq -L(w') - \lambda' R(w').$$

Adding both equations we have $(\lambda - \lambda') R(w) \leq (\lambda - \lambda') R(w')$. Because $R(w) \leq R(w')$, the above is satisfied if $\lambda \geq \lambda'$, and this inequality has to be strict as $\lambda \neq \lambda'$ by assumption.

Because the regularizer for the four parameter vectors evaluates to 2, 9, 3 and 4 respectively, this means that the order is $\lambda_4, \lambda_1, \lambda_3, \lambda_2$.

13. Ada’s colleague Alan wrote another program to solve the same optimization problem, but arrived at a different optimum for the same penalizer $\lambda > 0$.

Does this mean that one of them has an implementation bug? Justify your answer (for yourself).

(a) Yes
(b) No

Solution:
The correct answer is (b). No it does not, consider the case where all $x_i$ and all $y_i$ are equal to zero. Then any constant vector is a solution.

14. To ensure that her algorithm is correctly implemented, Ada wants to implement the following test procedure. First, come up with some synthetic distribution $P(x, y)$ where the data comes from. Then, compute the optimal vector $w^*$ on a finite sample from $P(x, y)$, and finally compute the generalization error of $w^*$. If she defined the distribution generating the data as

$$P(x, y) = \begin{cases} \frac{1}{3} & \text{if } x \in \{0, 1\}^3 \text{ and } y = x_1 + 2x_2 + 2x_3, \text{ or} \\ 0 & \text{otherwise} \end{cases},$$

and she computed the vector $w_* = (2, 2, 2)$ on the finite sample, what is the generalization error?

(a) $\frac{1}{2}$
(b) $\frac{1}{4}$
(c) $\frac{1}{8}$
The correct answer is (a).

Note that there will be no loss if \( x_1 = 0 \), since in this case \( w_\ast^\top x = y \). On the other hand if \( x_1 = 1 \) then the loss is always 1 irrespective of the values of \( x_2 \) and \( x_3 \), since in this case \( w_\ast^\top x = 2x_1 + 2x_2 + 2x_3 = x_1 + y = 1 + y \). Hence, the expected loss is equal to \( 1 \cdot P(x_1 = 1) = \frac{1}{2} \).

### Problem 2 (Perceptron):

15. Construct a perceptron which correctly classifies the following data. Choose appropriate values for the weights \( w_0, w_1 \) and \( w_2 \)

<table>
<thead>
<tr>
<th>Training Example</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>c</td>
<td>1</td>
<td>1</td>
<td>+1</td>
</tr>
</tbody>
</table>

(a) \( w_0 = -5, w_1 = 2, w_2 = 4 \)
(b) \( w_0 = 5, w_1 = 2, w_2 = -4 \)
(c) \( w_0 = -5, w_1 = 0, w_2 = -4 \)
(d) \( w_0 = 5, w_1 = 2, w_2 = 4 \)

**Solution:**
We can plot the data and trace a separation line. This line has slope \(-1/2\) and \( x_2 \)-intercept \( 5/4 \). \( x_2 = 5/4 - x_1/2 \) i.e. \( 2x_1 + 4x_2 - 5 = 0 \) Thus we can choose \( w_0 = -5, w_1 = 2, w_2 = 4 \)

16. Use the perceptron learning algorithm on the data above, using a learning rate \( \nu \) of 1.0 and initial weight values of \( w_0 = -0.5, w_1 = 0 \) and \( w_2 = 1 \).

Choose the correctly filled table from the options below. In practice, we would apply stochastic gradient descent. But to facilitate this exercise, we do not pick the data-points at random. Instead, we take a, b and c sequentially.

<table>
<thead>
<tr>
<th>Iteration i</th>
<th>( w_0 )</th>
<th>( w_1 )</th>
<th>( w_2 )</th>
<th>Training Example (a, b or c)</th>
<th>Class</th>
<th>( s = w_0 + w_1 x_1 + w_2 x_2 )</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.5</td>
<td>0</td>
<td>1</td>
<td>a.</td>
<td>-</td>
<td>0.5</td>
<td>Update</td>
</tr>
<tr>
<td>2</td>
<td>-1.5</td>
<td>0</td>
<td>0</td>
<td>b.</td>
<td>-</td>
<td>-1.5</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>-1.5</td>
<td>0</td>
<td>0</td>
<td>c.</td>
<td>+</td>
<td>-1.5</td>
<td>Update</td>
</tr>
<tr>
<td>4</td>
<td>-0.5</td>
<td>1</td>
<td>1</td>
<td>a.</td>
<td>-</td>
<td>0.5</td>
<td>Update</td>
</tr>
<tr>
<td>5</td>
<td>-1.5</td>
<td>1</td>
<td>0</td>
<td>b.</td>
<td>-</td>
<td>0.5</td>
<td>Update</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Iteration i</th>
<th>( w_0 )</th>
<th>( w_1 )</th>
<th>( w_2 )</th>
<th>Training Example (a, b or c)</th>
<th>Class</th>
<th>( s = w_0 + w_1 x_1 + w_2 x_2 )</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.5</td>
<td>0</td>
<td>1</td>
<td>a.</td>
<td>+</td>
<td>0.5</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>-0.5</td>
<td>0</td>
<td>1</td>
<td>b.</td>
<td>+</td>
<td>-1.5</td>
<td>Update</td>
</tr>
<tr>
<td>3</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>c.</td>
<td>-</td>
<td>-1.5</td>
<td>None</td>
</tr>
<tr>
<td>4</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>a.</td>
<td>+</td>
<td>0.5</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>b.</td>
<td>+</td>
<td>0.5</td>
<td>None</td>
</tr>
</tbody>
</table>
(b)

<table>
<thead>
<tr>
<th>Iteration i</th>
<th>w0</th>
<th>w1</th>
<th>w2</th>
<th>Training Example (a, b or c )</th>
<th>Class</th>
<th>s=w0+w1x1+w2x2</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.5</td>
<td>0</td>
<td>1</td>
<td>a.</td>
<td>-</td>
<td>0.5</td>
<td>Update</td>
</tr>
<tr>
<td>2</td>
<td>-1.5</td>
<td>1</td>
<td>1</td>
<td>b.</td>
<td>-</td>
<td>1.5</td>
<td>Update</td>
</tr>
<tr>
<td>3</td>
<td>-1.5</td>
<td>0</td>
<td>0</td>
<td>c.</td>
<td>+</td>
<td>-1.5</td>
<td>None</td>
</tr>
<tr>
<td>4</td>
<td>-0.5</td>
<td>1</td>
<td>1</td>
<td>a.</td>
<td>-</td>
<td>0.5</td>
<td>Update</td>
</tr>
<tr>
<td>5</td>
<td>-1.5</td>
<td>1</td>
<td>0</td>
<td>b.</td>
<td>-</td>
<td>0.5</td>
<td>Update</td>
</tr>
</tbody>
</table>

(c)

**Solution:**
The correct answer is (a).