Solutions week 4

4.1

Programming exercise...

4.2

(a)

For a vector v we have that $v^T v = \|v\|_2^2$, so if $y = Ux$ and $U^T U = I$ we get

$$\|y\|_2^2 = y^T y = x^T U^T U x = x^T x = \|x\|_2^2$$

This means that unitary transformations does not change lengths.

(b)

Let's first inspect what $A^T A$ and AA^T are when expressed using singular value decomposition.

$$A^T A = V S^T U^T U S V^T = V S^T S V^T = V S^T S V^{-1}$$
$$AA^T = U S V^T V S^T U^T = U S S^T U^T = U S S^T U^{-1}$$

We notice that this is exactly the diagonalization of a matrix using eigenvalues and eigenvectors ($X = V D V^{-1}$) and we can identify $S^T S$ and SS^T as diagonal matrices containing eigenvalues of $A^T A$ and AA^T respectively, while V and U have columns with eigenvectors for $A^T A$ and AA^T respectively.

(c)

For a general X matrix, the SVD will be

$$X = USV^T = \begin{matrix} m \end{matrix} \overset{\begin{matrix} r & m-r \end{matrix}}{\begin{bmatrix} U_1 & U_2 \end{bmatrix}} \begin{matrix} r \\ m-r \end{matrix} \overset{\begin{matrix} r & n-r \end{matrix}}{\begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix}} \begin{matrix} r \\ n-r \end{matrix} \overset{\begin{matrix} n \end{matrix}}{\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}} = U_1 S_1 V_1^T.$$

With $U \in \mathcal{R}^{m \times m}$ and $V \in \mathcal{R}^{n \times n}$ unitary and where $S_1 \in \mathcal{R}^{r \times r}$ is diagonal, with r positive diagonal elements.

Note that X^TX will be invertible precisely when S^TS is invertible (since $X^TX = US^TSU^T$ and U is invertible). Since

$$S^TS = \begin{array}{c} \\ r \\ n-r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{bmatrix} S_1^2 & 0 \\ 0 & 0 \end{bmatrix} \end{array}$$

this happens precisely when $r = n$, which means that X has full column rank. (The interpretation is that there are no "useless feature-combinations" in our linear regression model.). Note also that when $r = n$ the matrix V_2 will disappear (become an empty matrix), and we will have $V = V_1$.

Let's now look at how the least squares formula is transformed, using the economy version SVD $X = U_1S_1V_1^T$

$$\begin{aligned} X^TX\theta &= X^Ty \\ V_1S_1^2V_1^T\theta &= V_1S_1U_1^Ty \\ S_1^2V_1^T\theta &= S_1U_1^Ty \\ V_1^T\theta &= S_1^{-1}U_1^Ty \\ V_1V_1^T\theta &= V_1S_1^{-1}U_1^Ty \\ \theta &= V_1S_1^{-1}U_1^Ty \end{aligned}$$

The interpretation of this is then that we transform y by projecting it onto each of the vectors in U_1. This projection will take us to the room where basis vectors express the data the best, and here we scale with the inverse of the singular values.

The last step is actually somewhat dangerous (inverting the singular values). It means that we find θ by allowing θ to get large values to be able to fit the parts of the data with small singular values, which corresponds to directions with little information in the data. This is a good motivation for using regularization, as in the next part of this exercise.

(d)

Updating our earlier calculations we get

$$\begin{aligned} (X^TX + \gamma I)\theta &= X^Ty \\ (V_1S_1^2V_1^T + \gamma I)\theta &= V_1S_1U_1^Ty \\ V_1(S_1^2 + \gamma I)V_1^T\theta &= V_1S_1U_1^Ty \\ (S_1^2 + \gamma I)V_1^T\theta &= S_1U_1^Ty \\ \theta &= V_1(S_1^2 + \gamma I)^{-1}S_1U_1^Ty \end{aligned}$$

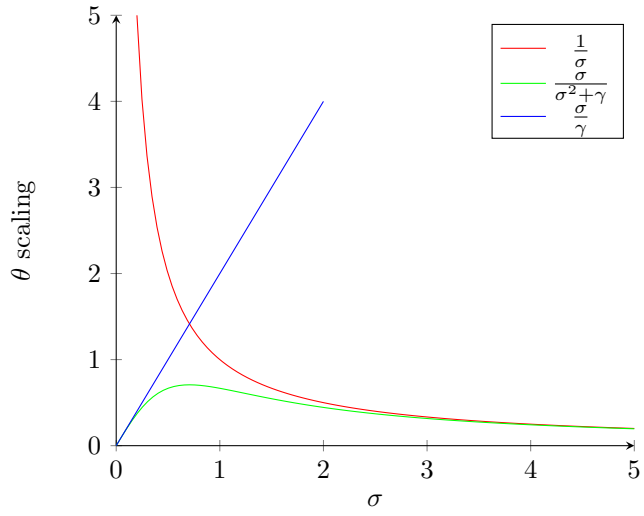
Looking at the difference from the previous problem we have

$$S_1^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & & \\ & \frac{1}{\sigma_2} & \\ & & \ddots \end{bmatrix}$$

$$(S_1^2 + \gamma I)^{-1} S_1 = \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\gamma} & & \\ & \frac{\sigma_2}{\sigma_2^2+\gamma} & \\ & & \ddots \end{bmatrix}$$

where if we plot one of the diagonal elements we see that for no regularization we move towards infinite scaling for theta as the singular values become small (i.e. we can allow for arbitrary sized values in θ for weak signals to fit as good as possible) while for regularization with a factor γ we have the same asymptotic behaviour, but we have a linear behaviour for small σ which mean that we will allow θ to change less for signals in directions that have too small σ.

We can also note that the peak of this scaling curve for Tikhonov regularization is at $\sqrt{\gamma}$, so with this you set at what point lower singular value will vanish quickly.



(e)

With $P_1 = U_1 U_1^T$ and using $U_1^T U_1 = I$ we get

$$\begin{aligned} P_1^T &= (U_1 U_1^T)^T = U_1 U_1^T = P_1 \\ P_1^2 &= U_1 U_1^T U_1 U_1^T = U_1 I U_1^T = P_1 \\ P_1(I_m - P_1) &= P_1 - P_1^2 = 0 \end{aligned}$$

The cases with $P_2 = U_2 U_2^T$, $Q_1 = V_1 V_1^T$, and $Q_2 = V_2 V_2^T$ are similar.

(f)

Programming exercise...

4.3

Programming exercise...

4.4

Programming exercise...

4.5

For linear discriminant analysis we model $p(x \mid y = i)$ as a Gaussian where all classes have the same covariance matrix.

$$p(x \mid y = i) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i)\right)$$

We then make predictions based on the optimal bayes classifier,

$$p(y = i \mid x) = \frac{p(x \mid y = i)p(y = i)}{\sum_j p(x \mid y = j)p(y = j)},$$

and as we know we should pick which ever y gives the highest probability which (since the denominator is constant for any x) is the same as the y giving the largest $p(x \mid y)p(y)$. We look at the log of the probability which makes it nicer (log is a strictly increasing function for $x > 0$ so the y giving the highest probability will also give the highest log probability).

$$\begin{aligned}
\log\left(p(x \mid y = i)p(y = i)\right) &= \log p(x \mid y = i) + \log p(y = i) \\
&= \log\left(\frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i)\right)\right) + C_1 \\
&= -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + C_2
\end{aligned}$$

What we would now like to do is to find the boundary where two classes are equally likely to be picked, and show that this is linear for any two classes.

$$\begin{aligned}
-\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + C_i &= -\frac{1}{2}(x - \mu_j)^T \Sigma^{-1}(x - \mu_j) + C_j \\
x^T \Sigma^{-1} x - 2x^T \Sigma^{-1}\mu_i + \mu_i^T \Sigma^{-1}\mu_i - x^T \Sigma^{-1} x + 2x^T \Sigma^{-1}\mu_j - \mu_j^T \Sigma^{-1}\mu_j &= 2(C_i - C_j) \\
x^T \Sigma^{-1}(\mu_j - \mu_i) &= \mu_j^T \Sigma^{-1}\mu_j/2 - \mu_i^T \Sigma^{-1}\mu_i/2 + C_i - C_j
\end{aligned}$$

where we have used that $x^T\Sigma^{-1}\mu = \mu^T\Sigma^{-1}x$ since Σ^{-1} is a symmetric matrix.

Therefore, we get a decision boundary of the form $x^T w = b$ which defines a hyperplane (with normal vector w). Hence all class-boundaries are described by linear functions.

4.6

Programming exercise...