HOMEWORK 1 - V0

STA 414/2104 Winter 2021

University of Toronto

Version history: V0 → V1:

- **Deadline:** Mon, Feb. 8, at 13:59.
- **Submission:** You need to submit your solutions through Crowdmark, including all your derivations, plots, and your code. You can produce the file however you like (e.g. LaTeX, Microsoft Word, etc), as long as it is readable. Points will be deducted if we have a hard time reading your solutions or understanding the structure of your code. For each question, you should append your code right after your answers to that question.

1. Probability and Calculus - 20 pts.

1.1. *Variance and covariance - 10 pts.* Let X, Y be two independent random vectors in $\mathbb{R}^m$.

(a) Find their covariance.
(b) For a constant matrix $A \in \mathbb{R}^{m \times m}$, show the following two properties:

$$\mathbb{E}(X + AY) = \mathbb{E}(X) + A\mathbb{E}(Y)$$
$$\text{Var}(X + AY) = \text{Var}(X) + A\text{Var}(Y)A^T$$

(c) Using part (b), show that if $X \sim \mathcal{N}(\mu, \Sigma)$, then $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$.

What to submit?

a) Here, simply stating the answer is sufficient.
b) Show that the equalities hold for each entry.
c) Here, you may use the fact that linear transformation of a Gaussian random vector is again Gaussian. Therefore you only need to compute the expected value and variance of AX.

1.2. *Calculus - 10 pts.* Let $x, y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times m}$. In vector notation, what is

(a) the gradient with respect to x of $x^T y$?
(b) the gradient with respect to x of $x^T x$?
(c) the gradient with respect to x of $\frac{1}{2} x^T A x$?
(d) the gradient with respect to x of $\exp(x^T A x)$?

What to submit?

a-d) Use the definition of gradient and find its each entry. Results should be in vector form.

2. Regression - 40 pts. In this question, you will derive certain properties of linear regression.

2.1. *Linear regression - 20 pts.* Suppose that $\mathbf{\Phi} \in \mathbb{R}^{n \times m}$ with $n \geq m$ and $\mathbf{t} \in \mathbb{R}^n$, and that $\mathbf{t}|(\mathbf{\Phi}, \mathbf{w}) \sim \mathcal{N}(\mathbf{\Phi w}, \sigma^2 \mathbf{I})$. We know that the maximum likelihood estimate $\hat{\mathbf{w}}$ of $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{t}.$$

(a) Write the log-likelihood implied by the model above, and compute its gradient w.r.t. $\mathbf{w}$. By setting it equal to 0, derive the above estimator $\hat{\mathbf{w}}$.
(b) Find the distribution of $\hat{\mathbf{w}}$, its expectation and covariance matrix.

What to submit?

a) Log-likelihood, its gradient, and your entire derivation.
b) Use the property in 1.1 c) of multivariate Gaussian random vectors, and find the distribution, and calculate its expectation and variance.

2.2. *Ridge regression and MAP - 20 pts.* Suppose that we have $\mathbf{t}|(\mathbf{\Phi}, \mathbf{w}) \sim \mathcal{N}(\mathbf{\Phi w}, \sigma^2 \mathbf{I})$ and we place a normal prior on $\mathbf{w}|\mathbf{\Phi}$, i.e., $\mathbf{w} \sim \mathcal{N}(0, \tau^2 \mathbf{I})$. Recall from the first lecture (also in preliminaries.pdf) that MAP estimate of $\mathbf{w}$ is given as the maximum of the posterior density

$$\hat{\mathbf{w}}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} \{p(\mathbf{w}|\mathbf{\Phi}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{\Phi})\}.$$

Here, $\propto$ notation means *proportional to*, and is used since we dropped the term $p(\mathbf{t}|\mathbf{\Phi})$ in the denominator as it doesn't have $\mathbf{w}$ in it, thus it doesn't contribute to the maximization problem.

Show that the MAP estimate of $\mathbf{w}$ given $(\mathbf{t}, \mathbf{\Phi})$ in this context is

$$\hat{\mathbf{w}}_{MAP} = (\mathbf{\Phi}^\top \mathbf{\Phi} + \lambda \mathbf{I})^{-1} \mathbf{\Phi}^\top \mathbf{t} \tag{2.1}$$

where $\lambda = \sigma^2 / \tau^2$.

What to submit?

a) Submit all your derivations.

3. Cross validation - 40 pts. In this problem, you will write a function that performs K-fold cross validation (CV) procedure to tune the penalty parameter λ in Ridge regression. CV procedure is one of the most commonly used methods for tuning hyperparameters. *In this question, you shouldn't use the package* `scikit-learn` to perform CV. *You should implement all of the below functions yourself. You may use* `numpy` *and* `scipy` *for basic math operations such as linear algebra, sampling etc.*

In class we learned *training*, *test*, and *validation* procedures which assumes that you have enough data and you can set aside a validation set and a test set to use it for assessing the performance of your machine learning algorithm. However in practice, this may be problematic since we may not have enough data. A remedy to this issue is K-fold cross- validation which uses a part of the available data to fit the model, and a different part to test it. K-fold CV procedure splits the data into K equal-sized parts; for example, when K = 5, the scenario looks like this:

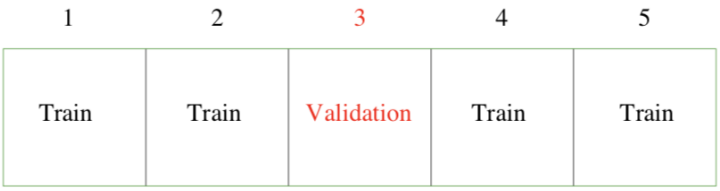

Fig 1: credit: Elements of Statistical Learning

1. We first set aside a test dataset and never use it until the training and parameter tuning procedures are complete. We will use this data for final evaluation. In this question, test data is provided to you as a separate dataset.
2. CV error estimates the test error of a particular hyperparameter choice. For a particular hyperparameter value, we split the training data into K blocks (See the figure), and for $k = 1, 2, ..., K$ we use the k-th block for validation and the remaining $K - 1$ blocks are for training. Therefore, we train and validate our algorithm K times. Our CV estimate for the test error for that particular hyperparameter choice is given by the average validation error across these K blocks.
3. We repeat the above procedure for several hyperparameter choices and choose the one that provides us with the smalles CV error (which is an estimate for the test error).

Below, we will code the above procedure for tuning the regularization parameter in linear regression which is a hyperparameter. Your `cross_validation` function will rely on 6 short functions which are defined below along with their variables.

- `data` is a variable and refers to a $(\mathbf{t}, \mathbf{\Phi})$ pair (can be test, training, or validation) where $\mathbf{t}$ is the target (response) vector, and $\mathbf{\Phi}$ is the feature matrix.
- `model` is a variable and refers to the coefficients of the trained model, i.e. $\hat{\mathbf{w}}_\lambda$.
- `data_shf = shuffle_data(data)` is a function and takes `data` as an argument and returns its randomly permuted version along the samples. Here, we are considering a uniformly random permutation of the training data. Note that $\mathbf{t}$ and $\mathbf{\Phi}$ need to be permuted the same way preserving the target-feature pairs.
- `data_fold, data_rest = split_data(data, num_folds, fold)` is a function that takes `data`, number of partitions as `num_folds` and the selected partition `fold` as its arguments and returns the selected partition (block) `fold` as `data_fold`, and the remaining data as `data_rest`. If we consider 5-fold cross validation, `num_folds`=5, and your function splits the data into 5 blocks and returns the block `fold` ($\in \{1, 2, 3, 4, 5\}$) as the validation fold and the remaining 4 blocks as `data_rest`. Note that `data_rest` $\cup$ `data_fold` = `data`, and `data_rest` $\cap$ `data_fold` $= \emptyset$.
- `model = train_model(data, lambd)` is a function that takes `data` and `lambd` as its arguments, and returns the coefficients of ridge regression with penalty level λ. For simplicity, you may ignore the intercept and use the expression in equation (2.1).
- `predictions = predict(data, model)` is a function that takes `data` and `model` as its arguments, and returns the `predictions` based on `data` and `model`.

- `error = loss(data, model)` is a function which takes `data` and `model` as its arguments and returns the average squared `error` loss based on `model`. This means if `data` is composed of $\mathbf{t} \in \mathbb{R}^n$ and $\mathbf{\Phi} \in \mathbb{R}^{n \times p}$, and model is $\hat{\mathbf{w}}$, then the return value is $\|\mathbf{t} - \mathbf{\Phi}\hat{\mathbf{w}}\|^2/n$.
- `cv_error = cross_validation(data, num_folds, lambd_seq)` is a function that takes the training `data`, number of folds `num_folds`, and a sequence of λ's as `lambd_seq` as its arguments and returns the cross validation error across all λ's. Take `lambd_seq` as evenly spaced 50 numbers over the interval (0.02, 1.5). This means `cv_error` will be a vector of 50 errors corresponding to the values of `lambd_seq`. Your function will look like:

```
data = shuffle_data(data)
for i = 1,2,...,length(lambd_seq)
       lambd = lambd_seq(i)
       cv_loss_lmd = 0.
       for fold = 1,2, ...,num_folds
            val_cv, train_cv = split_data(data, num_folds, fold)
            model = train_model(train_cv, lambd)
            cv_loss_lmd += loss(val_cv, model)
       cv_error(i) = cv_loss_lmd / num_folds
return cv_error
```

Download the dataset from the course webpage `hw1_data.zip` and place and extract in your working directory, or note its location `file_path`. For example, file path could be `/Users/yourname/Desktop/`

- In `Python`:

```
import numpy as np
data_train = {'X': np.genfromtxt('data_train_X.csv', delimiter=','),
              't': np.genfromtxt('data_train_y.csv', delimiter=',')}
data_test = {'X': np.genfromtxt('data_test_X.csv', delimiter=','),
             't': np.genfromtxt('data_test_y.csv', delimiter=',')}
```

Here, the design matrix $\mathbf{\Phi}$ is loaded as `data_??['X']`, and target vector $\mathbf{t}$ is loaded as `data_??['t']`, where `??` is either `train` or `test`.

(a) Write the above 6 functions, and identify the correct order and arguments to do cross validation.

(b) Find the training and test errors corresponding to each λ in `lambd_seq`. This part does not use the `cross_validation` function but you may find the other functions helpful.

(c) Plot training error, test error, and 5-fold and 10-fold cross validation errors on the same plot for each value in `lambd_seq`. What is the value of λ proposed by your cross validation procedure? Comment on the shapes of the error curves.

What to submit?

a) The functions your wrote.
b) Report the errors you find.
c) The plot containing 4 curves: i) training ii) test, iii) 5-fold CV iv) 10-fold CV errors, where x axis is lambda.
d) Your entire code should be attached to the end of your answers.

1.1. *Variance and covariance - 10 pts.* Let X, Y be two independent random vectors in $\mathbb{R}^m$.

(a) Find their covariance.

(b) For a constant matrix $A \in \mathbb{R}^{m \times m}$, show the following two properties:

$$\mathbb{E}(X + AY) = \mathbb{E}(X) + A\mathbb{E}(Y)$$
$$\mathrm{Var}(X + AY) = \mathrm{Var}(X) + A\mathrm{Var}(Y)A^T$$

(c) Using part (b), show that if $X \sim \mathcal{N}(\mu, \Sigma)$, then $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$.

What to submit?

a) Here, simply stating the answer is sufficient.
b) Show that the equalities hold for each entry.
c) Here, you may use the fact that linear transformation of a Gaussian random vector is again Gaussian. Therefore you only need to compute the expected value and variance of AX.

(a). $\mathrm{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)^T]$

$= E[XY^T - X\mu_Y^T - \mu_X Y^T + \mu_X \mu_Y^T]$

$= E[XY^T] - E[X\mu_Y^T] - E[\mu_X Y^T] + E[\mu_X \mu_Y^T]$

$\because$ X, Y are two independent random vectors in $\mathbb{R}^m$

$\therefore E[XY^T] = E[X]E[Y^T]$

$= E[X]E[Y^T] - E[X]\mu_Y^T - \mu_X E[Y^T] + \mu_X \mu_Y^T$

$= \mu_X \mu_Y^T - \mu_X \mu_Y^T - \mu_X \mu_Y^T + \mu_X \mu_Y^T$

$= 0$

$\therefore \mathrm{cov}(X, Y) = 0$

(b).

Show: $E[X + AY] = E[X] + AE[Y]$

$E(X+AY)$

$$= E\left[\begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} + \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}\right]$$

$$= E\begin{bmatrix} x_1 + \sum_{i=1}^{m} a_{1i} y_i \\ \vdots \\ x_m + \sum_{i=1}^{m} a_{mi} y_i \end{bmatrix}$$

$$= \begin{bmatrix} E\left[x_1 + \sum_{i=1}^{m} a_{1i} y_i\right] \\ \vdots \\ E\left[x_m + \sum_{i=1}^{m} a_{mi} y_i\right] \end{bmatrix}$$

$$= \begin{bmatrix} E(x_1) + \begin{bmatrix} a_{11} & \cdots & a_{1m} \end{bmatrix} \begin{bmatrix} E(y_1) \\ \vdots \\ E(y_m) \end{bmatrix} \\ \vdots \\ E(x_m) + \begin{bmatrix} a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} E(y_1) \\ \vdots \\ E(y_m) \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_m) \end{bmatrix} + \begin{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1m} \end{bmatrix} \begin{bmatrix} E(y_1) \\ \vdots \\ E(y_m) \end{bmatrix} \\ \vdots \\ \begin{bmatrix} a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} E(y_1) \\ \vdots \\ E(y_m) \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_m) \end{bmatrix} + \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} E(y_1) \\ \vdots \\ E(y_m) \end{bmatrix}$$

$$= E(X) + A\,E(Y)$$

$$\therefore E[X + AY] = E[X] + A\,E[Y]$$

show: $Var(X + AY) = Var(X) + A\,Var(Y)A^T$

$$\because Var(X) = E[(X - \mu_X)(X - \mu_X)^T]$$

$$\therefore \mathrm{Var}(X+AY)$$

$$= E[(X+AY-\mu_X-A\mu_Y)(X+AY-\mu_X-A\mu_Y)^T]$$

$$= E[((X-\mu_X)+(AY-A\mu_Y))((X-\mu_X)+(AY-A\mu_Y))^T]$$

$$= E[(X-\mu_X)(X-\mu_X)^T + (X-\mu_X)(AY-A\mu_Y)^T + (AY-A\mu_Y)(X-\mu_X)^T + (AY-A\mu_Y)(AY-A\mu_Y)^T]$$

$$= E[(X-\mu_X)(X-\mu_X)^T] + E[(X-\mu_X)(AY-A\mu_Y)^T] + E[(AY-A\mu_Y)(X-\mu_X)^T] + E[(AY-A\mu_Y)(AY-A\mu_Y)^T]$$

$E[(X-\mu_X)(AY-A\mu_Y)^T]$

$\because$ X, Y are independent

$= E[(X-\mu_X)]\,E[(AY-A\mu_Y)^T]$

$= (E(X)-\mu_X)(E(AY)-A\mu_Y)^T$

$= (\mu_X-\mu_X)(AE(Y)-A\mu_Y)^T$

$= 0$

$$E[(AY - A\mu_Y)(X - \mu_X)^T]$$

$\because$ independent X & Y

$$\therefore = E(AY - A\mu_Y)\,E[(X - \mu_X)]^T$$

$$= (E(AY) - A\mu_Y)\,(E(X) - \mu_X)^T$$

$$= (A\mu_Y - A\mu_Y)\,(E(X) - \mu_X)^T$$

$$= 0$$

$$E[(AY - A\mu_Y)(AY - A\mu_Y)^T]$$

$$= E[A(Y - \mu_Y)(Y - \mu_Y)^T A^T]$$

$$= A\,E[(Y - \mu_Y)(Y - \mu_Y)^T]A^T$$

$$= A\,Var(Y)A^T$$

$\therefore$ $Var(X + AY)$

$$= Var(X) + 0 + 0 + A\,Var(Y)A^T$$

$$= Var(X) + A\,Var(Y)A^T$$

$\therefore$ $$Var(X + AY) = Var(X) + A\,Var(Y)A^T$$

(c). if $X \sim N(\mu, \Sigma)$

from part (b),

$\Rightarrow E(AX) = AE(X)$

$\because E(X) = \mu$

$\therefore \boxed{E(AX) = A\mu}$

$\Rightarrow Var(AX) = A\,Var(X)\,A^T$

$\because Var(X) = \Sigma$

$\therefore \boxed{Var(AX) = A\Sigma A^T}$

$\because$ linear transformation of a Gaussian random vector is again Gaussian

$\therefore AX \sim N(A\mu, A\Sigma A^T)$

1.2. *Calculus - 10 pts.* Let $x, y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times m}$. In vector notation, what is

(a) the gradient with respect to x of $x^T y$?
(b) the gradient with respect to x of $x^T x$?
(c) the gradient with respect to x of $\frac{1}{2} x^T A x$?
(d) the gradient with respect to x of $\exp(x^T A x)$?

What to submit?

a-d) Use the definition of gradient and find its each entry. Results should be in vector form.

(a). $$\frac{\partial}{\partial x}(x^T y) = \begin{bmatrix} \frac{\partial}{\partial x_1}(x^T y) \\ \vdots \\ \frac{\partial}{\partial x_m}(x^T y) \end{bmatrix}$$

$$x^T y = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

$$= x_1 y_1 + \ldots + x_m y_m$$

$$\frac{\partial}{\partial x}(x^T y) = \begin{bmatrix} \frac{\partial}{\partial x_1}(x_1 y_1 + \ldots + x_m y_m) \\ \vdots \\ \frac{\partial}{\partial x_m}(x_1 y_1 + \ldots + x_m y_m) \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = y$$

$$\therefore \frac{\partial}{\partial x}(x^T y) = y.$$

(b). $X^TX = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$

$= x_1^2 + \cdots + x_m^2$

$$\frac{\partial}{\partial x}(X^TX) = \begin{bmatrix} \frac{\partial}{\partial x_1}(x_1^2 + \cdots + x_m^2) \\ \vdots \\ \frac{\partial}{\partial x_m}(x_1^2 + \cdots + x_m^2) \end{bmatrix} = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_m \end{bmatrix} = 2X.$$

$$\therefore \frac{\partial}{\partial x}(X^TX) = 2X.$$

(c). $\frac{\partial}{\partial x}\left(\frac{1}{2}X^TAX\right)$

$= \frac{1}{2}\frac{\partial}{\partial x}(X^TAX)$

$$X^TAX = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} x_i x_j$$

$$= \underset{1\times m}{\begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}} \underset{m\times 1}{\begin{bmatrix} a_{11}x_1 + \cdots + a_{1m}x_m \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mm}x_m \end{bmatrix}}$$

$$= a_{11}x_1^2 + \cdots + a_{1m}x_1x_m + \cdots + a_{m1}x_1x_m + \cdots + a_{mm}x_m^2$$

$$\frac{\partial}{\partial x}(x^TAx) = \begin{bmatrix} \frac{\partial}{\partial x_1}\left(\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}x_ix_j\right) \\ \vdots \\ \frac{\partial}{\partial x_m}\left(\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}x_ix_j\right) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial x_1}\left(a_{11}x_1^2 + \cdots + a_{1m}x_1x_m + \cdots + a_{m1}x_1x_m + \cdots + a_{mm}x_m^2\right) \\ \vdots \\ \frac{\partial}{\partial x_m}\left(a_{11}x_1^2 + \cdots + a_{1m}x_1x_m + \cdots + a_{m1}x_1x_m + \cdots + a_{mm}x_m^2\right) \end{bmatrix}$$

$$= \begin{bmatrix} 2a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m + a_{21}x_2 + \cdots + a_{m1}x_m \\ \vdots \\ a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{(m-1)m}x_{m-1} + a_{m1}x_1 + a_{m2}x_2 + \cdots + 2a_{mm}x_m \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1m}x_m \\ \vdots \\ a_{1m}x_1 + a_{2m}x_2 + \dots + a_{mm}x_m \end{bmatrix} + \begin{bmatrix} a_{11}x_1 + a_{21}x_2 + \dots + a_{m1}x_m \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mm}x_m. \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{m} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{m} a_{mj}x_j \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{m} a_{i1}x_i \\ \vdots \\ \sum_{i=1}^{m} a_{im}x_i \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{m} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{m} a_{mj}x_j \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{m} a_{i1}x_i \\ \vdots \\ \sum_{i=1}^{m} a_{im}x_i \end{bmatrix}$$

row vectors

$$= \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{m1} & \cdots & a_{mm} \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} + \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \\ a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

$$= A^T x + Ax$$

$$= (A^T + A)x.$$

$$\therefore \frac{\partial}{\partial x}\left(\frac{1}{2} x^T A x\right) = \frac{1}{2}(A^T + A)x.$$

(d). $\frac{\partial}{\partial x}\left(e^{x^T A x}\right)$

$$= e^{x^T A x} \cdot \frac{\partial}{\partial x}(x^T A x)$$

from part (c). $\frac{\partial}{\partial x}(x^T A x) = (A^T + A)x$

$$\therefore \frac{\partial}{\partial x}\left(e^{x^T A x}\right) = (A^T + A)x\, e^{x^T A x}$$

2.1. *Linear regression - 20 pts.* Suppose that $\mathbf{\Phi} \in \mathbb{R}^{n \times m}$ with $n \geq m$ and $\mathbf{t} \in \mathbb{R}^n$, and that $\mathbf{t}|(\mathbf{\Phi}, \mathbf{w}) \sim \mathcal{N}(\mathbf{\Phi w}, \sigma^2 \mathbf{I})$. We know that the maximum likelihood estimate $\hat{\mathbf{w}}$ of $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{t}.$$

(a) Write the log-likelihood implied by the model above, and compute its gradient w.r.t. $\mathbf{w}$. By setting it equal to 0, derive the above estimator $\hat{\mathbf{w}}$.
(b) Find the distribution of $\hat{\mathbf{w}}$, its expectation and covariance matrix.

What to submit?

a) Log-likelihood, its gradient, and your entire derivation.
b) Use the property in 1.1 c) of multivariate Gaussian random vectors, and find the distribution, and calculate its expectation and variance.

(a). $p(t|\phi, w) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(t-\phi w)^T \Sigma^{-1} (t-\phi w)\right\}$

$\ln p(t|\phi, w) = \ln (2\pi)^{-n/2} + \ln |\Sigma|^{1/2} - \frac{1}{2}(t-\phi w)^T \Sigma^{-1} (t-\phi w)$

$= -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(t-\phi w)^T \Sigma^{-1} (t-\phi w)$

∴ log-likelihood is

$$\ln p(t|\phi, w) = -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(t-\phi w)^T \Sigma^{-1} (t-\phi w)$$

$\frac{\partial}{\partial w}(\ln p(t|\phi, w) = 0 + 0 - \frac{1}{2}\frac{\partial}{\partial w}\left((t-\phi w)^T \Sigma^{-1} (t-\phi w)\right)$

$\text{Var}(t) = \sigma^2 I = \Sigma.$

$\Rightarrow \Sigma^{-1} = \frac{1}{\sigma^2} I$

suppose $A = t - \phi w$

$$\Rightarrow A^T \frac{1}{\sigma^2} I A$$

$$= [a_1 \ \dots \ a_n] \begin{bmatrix} \frac{1}{\sigma^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma^2} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

$$= \left[\frac{1}{\sigma^2} a_1 \ \dots \ \frac{1}{\sigma^2} a_n \right] \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

$$= \frac{1}{\sigma^2} [a_1 \ \dots \ a_n] \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

$\Rightarrow$

$$\frac{\partial}{\partial w} (\ln p(t|\phi, w) = -\frac{1}{2\sigma^2} \frac{\partial}{\partial w} ((t-\phi w)^T (t-\phi w))$$

$$= -\frac{1}{2\sigma^2} \frac{\partial}{\partial w} (t^T t - t^T \phi w - w^T \phi^T t + w^T \phi^T \phi w)$$

($t^T \phi w$ → number, $w^T \phi^T t$ ← number)

$$= -\frac{1}{2\sigma^2} \frac{\partial}{\partial w} (t^T t - 2 w^T \phi^T t + w^T \phi^T \phi w)$$

($w^T \phi^T t$: x; $\phi^T \phi$: A)

$\frac{\partial}{\partial w}(w^T x) = x$ $\quad \frac{\partial}{\partial w}(w^T A w) = (A + A^T) w$

$$= -\frac{1}{2\sigma^2} (0 - 2\phi^T t + (\phi^T \phi + \phi^T \phi) w)$$

$$= -\frac{1}{2\sigma^2} (-2\phi^T t + 2\phi^T \phi w)$$

$\therefore$ gradient is $\frac{\partial}{\partial w} (\ln p(t|\phi, w) = -\frac{1}{2\sigma^2} (-2\phi^T t + 2\phi^T \phi w)$

$$-\frac{1}{2\sigma^2}\left(-2\Phi^T t + 2\Phi^T\Phi w\right) = 0$$

$$2\Phi^T t = 2\Phi^T\Phi w$$

$$\Phi^T t = \Phi^T\Phi w$$

$$\hat{w} = (\Phi^T\Phi)^{-1}\Phi^T t$$

$\therefore$ the estimator $\hat{w} = (\Phi^T\Phi)^{-1}\Phi^T t$

(b). $E(\hat{w}) = E\left((\Phi^T\Phi)^{-1}\Phi^T t\right)$

$= (\Phi^T\Phi)^{-1}\Phi^T \underbrace{E(t)}_{\Phi w}$

$= (\Phi^T\Phi)^{-1}\Phi^T\Phi w$

$= \boxed{w}$

$$\text{Var}(\hat{w}) = \text{Var}\left((\phi^T\phi)^{-1}\phi^T t\right)$$

$$= (\phi^T\phi)^{-1}\phi^T \underbrace{\text{Var}(t)}_{\sigma^2 I} \underbrace{\left((\phi^T\phi)^{-1}\phi^T\right)^T}_{\phi(\phi^T\phi)^{-1}}$$

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

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

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

$\because t \mid (\phi, w) \sim N(\phi w, \sigma^2 I)$

& linear transformation of a Gaussian random vector is again Gaussian

$\therefore$ $\boxed{\hat{w} \sim N\left(w, \sigma^2 (\phi^T\phi)^{-1}\right)}$

2.2. *Ridge regression and MAP - 20 pts.* Suppose that we have $\mathbf{t}|(\mathbf{\Phi}, \mathbf{w}) \sim \mathcal{N}(\mathbf{\Phi w}, \sigma^2\mathbf{I})$ and we place a normal prior on $\mathbf{w}|\mathbf{\Phi}$, i.e., $\mathbf{w} \sim \mathcal{N}(0, \tau^2\mathbf{I})$. Recall from the first lecture (also in preliminaries.pdf) that MAP estimate of $\mathbf{w}$ is given as the maximum of the posterior density

$$\hat{\mathbf{w}}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}}\{p(\mathbf{w}|\mathbf{\Phi}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w})p(\mathbf{w}|\mathbf{\Phi})\}.$$

Here, $\propto$ notation means *proportional to*, and is used since we dropped the term $p(\mathbf{t}|\mathbf{\Phi})$ in the denominator as it doesn't have $\mathbf{w}$ in it, thus it doesn't contribute to the maximization problem.

Show that the MAP estimate of $\mathbf{w}$ given $(\mathbf{t}, \mathbf{\Phi})$ in this context is

$$\hat{\mathbf{w}}_{MAP} = (\mathbf{\Phi}^\top\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^\top\mathbf{t} \tag{2.1}$$

where $\lambda = \sigma^2/\tau^2$.

What to submit?

a) Submit all your derivations.

(a). $p(w|\phi, t) \propto p(t|\phi, w)\, p(w|\phi)$

$\ln p(w|\phi, t) \propto \ln p(t|\phi, w) + \ln p(w|\phi)$

$$\ln p(w|\phi, t) \propto -\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(t-\phi w)^T \Sigma^{-1}(t-\phi w) + \left(-\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| - \frac{1}{2}w^T\Sigma^{-1}w\right)$$

$$\frac{\partial}{\partial w}\left(-\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(t-\phi w)^T \Sigma^{-1}(t-\phi w) + \left(-\frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| - \frac{1}{2}w^T\Sigma^{-1}w\right)\right)$$

$$= 0 + 0 + \frac{\partial}{\partial w}\left(-\frac{1}{2\sigma^2}(t-\phi w)^T(t-\phi w)\right) + 0 + 0 + \frac{\partial}{\partial w}\left(-\frac{1}{2\tau^2}w^T w\right)$$

$$= -\frac{1}{2\sigma^2}\frac{\partial}{\partial w}\left(t^T t - 2w^T\phi^T t + w^T\phi^T\phi w\right)$$

$$-\frac{1}{2\tau^2}\frac{\partial}{\partial w}(w^T w)$$

$$= -\frac{1}{2\sigma^2}\left(-2\phi^T t + 2\phi^T\phi w\right) - \frac{1}{2\tau^2}(2w)$$

$$-\frac{1}{2\sigma^2}\left(-2\phi^T t + 2\phi^T\phi w\right) - \frac{1}{2\tau^2}(2w) = 0$$

$$\frac{1}{\sigma^2}\left(\phi^T t - \phi^T\phi w\right) = \frac{1}{\tau^2}w$$

$$\phi^T t - \phi^T\phi w = \frac{\sigma^2}{\tau^2}w$$

$$\phi^T t = \frac{\sigma^2}{\tau^2}w + \phi^T\phi w$$

$$\left(\phi^T\phi + \frac{\sigma^2}{\tau^2}I\right)w = \phi^T t$$

$$\hat{w}_{MAP} = \left(\phi^T\phi + \underbrace{\frac{\sigma^2}{\tau^2}}_{\lambda}I\right)^{-1}\phi^T t$$

$$= (\phi^T\phi + \lambda I)^{-1}\phi^T t$$

∴ MAP estimate of w given ϕ, t is

$$\hat{w}_{MAP} = = (\phi^T\phi + \lambda I)^{-1}\phi^T t$$

where $\lambda = \frac{\sigma^2}{\tau^2}$