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 $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$.

a) $\mathrm{cov}(X, Y) = E(XY) - \mu_X \mu_Y$

But, since X, Y are independent R.V then $\mathrm{cov}(X, Y) = 0$

b) Let $A \in \mathbb{R}^{m \times m}$ be a constant matrix

$$E(X + AY) = E(X) + E(AY)$$
$$= E(X) + AE(Y) \quad \text{since } A \text{ is constant matrix}$$

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

Note: $\mathrm{Var}(AY) = E[(A(Y-\mu_Y))(A(Y-\mu_Y))^T]$

$$= E(A(Y-\mu_Y)(Y-\mu_Y)^T A^T)$$
$$= A\,\mathrm{var}(Y)A^T$$

c) Assume $X \sim N(\mu, \Sigma)$

Then $E(AX) = AE(X) = A\mu$ and $\mathrm{Var}(AX) = A\mathrm{Var}(X)A^T = A\Sigma A^T$

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

∎

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 $\frac{1}{2}x^T Ax$?
(d) the gradient with respect to x of $\exp(x^T Ax)$?

Let $x^T = [x_1, \cdots, x_m]$ $\quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$

a) $x^T y = x_1 y_1 + x_2 y_2 + \cdots + x_m y_m$

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

b) $x^T x = x_1^2 + x_2^2 + \cdots + x_m^2$

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

c) $$\frac{\partial}{\partial x} \frac{1}{2} x^T A x = \frac{1}{2} \times \frac{d \sum_{k=1}^m \sum_{j=1}^m a_{k,j} x_j x_k}{d x_i}$$

$$= \frac{1}{2} \sum_{j=1}^m a_{ij} x_j + \sum_{k=1}^m a_{ki} x_k$$

$$= \frac{1}{2}(Ax + A^T x)$$

$$= \frac{1}{2}(A + A^T)x$$

if A is symmetric $\Rightarrow A = A^T$

so, $\frac{\partial}{\partial x} \frac{1}{2} x^T A x = Ax$

d) $$\frac{\partial}{\partial x} \exp\{x^T A x\} = \frac{\partial}{\partial x} x^T A x \times \exp\{x^T A x\}$$

$$= 2Ax \exp\{x^T A x\}$$

■

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 $\mathbf{t}|(\mathbf{\Phi}, \mathbf{w}) \sim \mathcal{N}(\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I})$. 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.

a) given Φ, w we can write the log likelihood:

$$p(t \mid \Phi w, \sigma^2 I) = \prod_{i=1}^{N} \mathcal{N}(t_i \mid \Phi w, \sigma^2 I)$$

$$= \prod_{i=1}^{N} \frac{1}{(2\pi)^{D/2}} \frac{1}{|\sigma^2 I|^{1/2}} \exp\left\{-\frac{1}{2}(t_i - \Phi w)^T (\sigma^2 I)^{-1} (t_i - \Phi w)\right\}$$

$$\ln(p(t \mid \Phi w, \sigma^2 I)) = \sum_{i=1}^{N} \ln\left(\frac{1}{(2\pi)^{D/2}} \frac{1}{|\sigma^2 I|^{1/2}}\right) + \left(-\frac{1}{2}(t_i - \Phi w)^T (\sigma^2 I)^{-1} (t_i - \Phi w)\right)$$

$$= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\sigma^2 I| - \frac{1}{2}\sum_{i=1}^{N}(t_i - \Phi w)^T(\sigma^2 I)^{-1}(t_i - \Phi w)$$

$$\propto \frac{1}{2\sigma^2 I}\sum_{i=1}^{N}(t_i - \Phi w)^T (t_i - \Phi w)$$ * constant terms removed

<u>Computing the gradient wrt w</u>

$$\frac{\partial}{\partial w} \ln p(t \mid \Phi w, \sigma^2 I) = \frac{1}{2\sigma^2 I}\sum_{i=1}^{N} \frac{\partial}{\partial w}(t_i - \Phi w)^T (t_i - \Phi w)$$

$$= \frac{1}{2\sigma^2 I} \times \sum_{i=1}^{N} \frac{\partial}{\partial w}\left(t_i^T t_i - t_i^T \Phi w - w^T \Phi^T t_i + w^T \Phi^T \Phi w\right)$$

$$= \frac{1}{2\sigma^2 I} \times 2(\Phi^T \Phi w - \Phi^T t)$$

$$= \frac{\Phi^T(t - \Phi w)}{\sigma^2}$$

<u>Setting it to 0 then finding $\hat{w}$:</u>

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

$$\Rightarrow \quad \Phi^T t - \Phi^T \Phi \hat{w} = 0$$

$$\Phi^T \Phi \hat{w} = \Phi^T t$$

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

■

b) $E(\hat{w}) = E[(\Phi^T\Phi)^{-1}\Phi^T t]$

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

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

$= w$

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

$= [(\Phi^T\Phi)^{-1}\Phi^T t] \, Var(t) \, [(\Phi^T\Phi)^{-1}\Phi^T t]^T$

$= [(\Phi^T\Phi)^{-1}\Phi^T t] \, \sigma^2 I \, [(\Phi^T\Phi)^{-1}\Phi^T t]^T$

$= \sigma^2 [(\Phi^T\Phi)^{-1}\Phi^T t] I [(\Phi^T\Phi)^{-1}\Phi^T t]^T$

$= \sigma^2$

Then we have $\hat{w} \sim N(w, \sigma^2)$

because linear transformation of a Gaussian R.V is Gaussian again. ∎

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

$$\hat{w}_{MAP} = \underset{w}{\arg\max}\ \{P(w|\Phi,t) \propto P(t|\Phi,w)\, P(w|\Phi)$$

We know the distribution of $P(t|\Phi,w)$ and $P(w|\Phi)$ then,

$$\log\big(P(t|\Phi,w)\,P(w|\Phi)\big) = \log\left[\frac{1}{(2\pi)^{D/2}\cdot|\tau^2 I|^{1/2}} e^{\left(-\frac{w^Tw}{2\tau^2}\right)} \times \frac{1}{(2\pi)^{N/2}\cdot|\sigma^2 I|^{1/2}} e^{-\frac{(t-\Phi w)^T(t-\Phi w)}{2\sigma^2}}\right]$$

$$= -\frac{w^Tw}{2\tau^2} - \frac{(t-\Phi w)^T(t-\Phi w)}{2\sigma^2} + C$$

* all constant terms bundled to C

Then taking gradient with respect to w :

$$\frac{d \log\big(P(t|\Phi,w)\,P(w|\Phi)\big)}{dw} = \frac{d}{dw}\left(-\frac{w^Tw}{2\tau^2} - \frac{t^Tt - t^T\Phi w - w^T\Phi^T t + w^T\Phi^T\Phi w}{2\sigma^2}\right)$$

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

Setting it to equal to 0 :

$$\frac{w}{\tau^2} - \frac{\Phi^T(t-\Phi w)}{\sigma^2} = 0$$

$$\frac{\sigma^2}{\tau^2}w - \Phi^T t + \Phi^T\Phi w = 0$$

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

$$\hat{w}_{map} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T t \quad \text{where } \lambda = \frac{\sigma^2}{\tau^2}$$

$\blacksquare$