CS229 Problem Set 1

Tianyu Du

Monday 8th July, 2019

1 Question 1: Linear Classifiers

1.1 Question 1(a)

Lemma 1.1.

$$g'(z) = g(z)\,(1 - g(z)) \tag{1.1}$$

Lemma 1.2. For every $x, z \in \mathbb{R}^n$,

$$\sum_i \sum_j z_i x_i z_j x_j = (x^T z)^2 \geq 0 \tag{1.2}$$

$$\begin{aligned}
\nabla_\theta J(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}\frac{g'(\theta^T x^{(i)})}{g(\theta^T x^{(i)})} - (1-y^{(i)})\frac{g'(\theta^T x^{(i)})}{1-g(\theta^T x^{(i)})}\right)x^{(i)} && (1.3)\\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}(1-g(\theta^T x^{(i)})) - (1-y^{(i)})g(\theta^T x^{(i)})\right)x^{(i)} && (1.4)\\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - g(\theta^T x^{(i)})\right)x^{(i)} && (1.5)\\
\implies \frac{\partial J(\theta)}{\partial \theta_j} &= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - g(\theta^T x^{(i)})\right)x_j^{(i)} \;\forall j \in [d] && (1.6)\\
\implies \forall k \in [d],\; \frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k} &= -\frac{1}{n}\sum_{i=1}^{n}\left(-g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_k^{(i)}\right)x_j^{(i)} && (1.7)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_j^{(i)}x_k^{(i)}\right) && (1.8)
\end{aligned}$$

Therefore, $H_J(\theta)$ can be constructed from the array of second order derivatives of $J(\theta)$ as

$$H_J(\theta)_{j,k} := \frac{1}{n}\sum_{i=1}^{n}\left(g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_j^{(i)}x_k^{(i)}\right) \tag{1.9}$$

Notice that since $g(\theta^T x^{(i)}) \in (0,1)$, therefore $g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)})) > 0$ for every θ and $x^{(i)}$.

Proof. Show that $H_J(\theta) \succeq 0$: let $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, then

$$z^T H_J(\theta) \in \mathbb{R}^{1\times d} \tag{1.10}$$

Then the β^{th} column of $z^T H_J(\theta)$ is

$$z^T H_J(\theta)_\beta = \frac{1}{n}\sum_{\alpha=1}^{d}\sum_{i=1}^{n} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))z_\alpha x_\alpha^{(i)} x_\beta^{(i)} \tag{1.11}$$

Therefore

$$z^T H_J(\theta) z = \frac{1}{n} \sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) z_\alpha x_\alpha^{(i)} x_\beta^{(i)} z_\beta \tag{1.12}$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) \sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} z_\alpha x_\alpha^{(i)} x_\beta^{(i)} z_\beta \tag{1.13}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)}))}_{>0 \ \because g(\cdot) \in (0,1)} \underbrace{(z^T x)^2}_{\geq 0} \geq 0 \tag{1.14}$$

Hence, $H_J(\theta) \succeq 0$ is shown by showing $z^T H_J(\theta) z$ for every $z \in \mathbb{R}^d$. ■

1.2 Question 1(c)

Proof. By Bayes' theorem,

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} \tag{1.15}$$

Define

$$z := \frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} \tag{1.16}$$

$$\Theta := \{\phi,\mu_0,\mu_1,\Sigma\} \tag{1.17}$$

Conditioned on particular x, y is either 0 or 1, therefore,

$$p(y=0|x;\Theta) = 1-z \tag{1.18}$$

$$\implies \frac{z}{1-z} = \frac{p(y=1|x;\Theta)}{p(y=0|x;\Theta)} \tag{1.19}$$

$$= \frac{p(x|y=1;\Theta)p(y=1;\Theta)}{p(x|y=0;\Theta)p(y=0;\Theta)} \tag{1.20}$$

$$= \frac{\phi}{1-\phi}\frac{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)} \tag{1.21}$$

$$\implies \log\frac{z}{1-z} = \log\frac{\phi}{1-\phi} \tag{1.22}$$

$$+\left(-\frac{1}{2}x^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}x - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right) \tag{1.23}$$

$$-\left(-\frac{1}{2}x^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}x - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0\right) \tag{1.24}$$

$$= \log\frac{\phi}{1-\phi} + \left((\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right) \tag{1.25}$$

$$\implies \frac{z}{1-z} = \exp\underbrace{\left(\log\frac{\phi}{1-\phi} + \left((\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right)\right)}_{=:\Delta} \tag{1.26}$$

$$\implies z = \frac{\exp(\Delta)}{1+\exp(\Delta)} = \frac{1}{1+\exp(-\Delta)} \tag{1.27}$$

Therefore

$$\frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} = \frac{1}{1+\exp(-(\theta^Tx+\theta_0))} \tag{1.28}$$

where

$$\theta = (\Sigma^{-1})^T(\mu_1-\mu_0) \tag{1.29}$$

$$\theta_0 = \log\frac{\phi}{1-\phi} + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 \tag{1.30}$$

■

1.3 Question 1(d)

1.3.1 ϕ

Proof.

$$\frac{\partial}{\partial \phi}\ell(\phi,\cdot) = \frac{\partial}{\partial \phi}\sum_{i=1}^{n}\underbrace{\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)}_{\perp\phi} + \log p(y^{(i)};\phi) \tag{1.31}$$

$$= \frac{\partial}{\partial \phi}\sum_{i=1}^{n}\log p(y^{(i)};\phi) \tag{1.32}$$

$$= \frac{\partial}{\partial \phi}\sum_{i=1}^{n}\log \phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}} \tag{1.33}$$

$$= \frac{\partial}{\partial \phi}\sum_{i=1}^{n}y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi) \tag{1.34}$$

$$= \sum_{i=1}^{n}y^{(i)}\frac{1}{\phi} - (1-y^{(i)})\frac{1}{1-\phi} \tag{1.35}$$

The first order condition of maximizing likelihood becomes

$$\sum_{i=1}^{n}y^{(i)}\frac{1}{\phi} - (1-y^{(i)})\frac{1}{1-\phi} = 0 \tag{1.36}$$

$$\implies \sum_{i=1}\frac{y^{(i)}}{\phi} + \frac{y^{(i)}}{1-\phi} - \frac{1}{1-\phi} = 0 \tag{1.37}$$

$$\implies \sum_{i=1}^{n}y^{(i)}\frac{1-\phi+\phi}{\phi(1-\phi)} - \frac{1}{1-\phi} = 0 \tag{1.38}$$

$$\implies \sum_{i=1}^{n}y^{(i)}\frac{1}{\phi(1-\phi)} = n\frac{1}{1-\phi} \tag{1.39}$$

$$\implies \phi = \frac{1}{n}\sum_{i=1}^{n}y^{(i)} \tag{1.40}$$

∎

1.3.2 μ_0

Proof.

$$\frac{\partial}{\partial \mu_0}\ell(\mu_0,\cdot) = \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n} \log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \underbrace{\log p(y^{(i)};\phi)}_{\perp \mu_0} \tag{1.41}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n} \log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) \tag{1.42}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n} \Bigg\{ \overbrace{y^{(i)} \log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right]}^{\perp \mu_0} \tag{1.43}$$

$$+ (1-y^{(i)}) \log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right] \Bigg\} \tag{1.44}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}(1-y^{(i)}) \left(\underbrace{\log \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}_{\perp \mu_0} - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0) \right) \tag{1.45}$$

$$= \frac{\partial}{\partial \mu_0}(-1)\sum_{i=1}^{n}(1-y^{(i)})\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0) = 0 \tag{1.46}$$

$$\implies \sum_{i=1}^{n}(1-y^{(i)})\Sigma^{-1}(x^{(i)}-\mu_0) = 0 \tag{1.47}$$

$$\tag{1.48}$$

$$\implies \sum_{i=1}^{n}\Sigma^{-1}(1-y^{(i)})x^{(i)} = \sum_{i=1}^{n}\Sigma^{-1}(1-y^{(i)})\mu_0 \tag{1.49}$$

$$\implies \sum_{i=1}^{n}(1-y^{(i)})x^{(i)} = \sum_{i=1}^{n}(1-y^{(i)})\mu_0 \tag{1.50}$$

$$\implies \mu_0 = \frac{\sum_{i=1}^{n}(1-y^{(i)})x^{(i)}}{\sum_{i=1}^{n}(1-y^{(i)})} = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=0\}} \tag{1.51}$$

■

1.3.3 μ_1

Proof.

$$\frac{\partial}{\partial \mu_1}\ell(\mu_1, \cdot) = \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \underbrace{\log p(y^{(i)};\phi)}_{\perp \mu_1} \tag{1.52}$$

$$= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) \tag{1.53}$$

$$= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\Big\{y^{(i)}\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right] \tag{1.54}$$

$$+ \underbrace{(1-y^{(i)})\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right]\Big\}}_{\perp \mu_1} \tag{1.55}$$

$$= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}y^{(i)}\left(\underbrace{\log\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}_{\perp \mu_1} - \frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right) \tag{1.56}$$

$$= \frac{\partial}{\partial \mu_1}(-1)\sum_{i=1}^{n}y^{(i)}\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1) = 0 \tag{1.57}$$

$$\implies \sum_{i=1}^{n}y^{(i)}\Sigma^{-1}(x^{(i)}-\mu_1) = 0 \tag{1.58}$$

$$\implies \sum_{i=1}^{n}\Sigma^{-1}y^{(i)}x^{(i)} = \sum_{i=1}^{n}\Sigma^{-1}y^{(i)}\mu_1 \tag{1.59}$$

$$\implies \sum_{i=1}^{n}y^{(i)}x^{(i)} = \sum_{i=1}^{n}y^{(i)}\mu_1 \tag{1.60}$$

$$\implies \mu_1 = \frac{\sum_{i=1}^{n}y^{(i)}x^{(i)}}{\sum_{i=1}^{n}y^{(i)}} = \frac{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=1\}} \tag{1.61}$$

■

1.3.4 Σ^{-1}

Proof. TODO ■

2 Question 2: Incomplete, Positive-Only Labels

2.1 Question 2(c)

Proof.

$$
\begin{aligned}
p(t^{(i)} = 1 | y^{(i)} = 1, x^{(i)}) &= \frac{p(y^{(i)} = 1 | t^{(i)} = 1, x^{(i)}) p(t^{(i)} = 1 | x^{(i)})}{p(y^{(i)} = 1 | x^{(i)})} && (2.1) \\
&= \frac{p(y^{(i)} = 1 | t^{(i)} = 1, x^{(i)}) p(t^{(i)} = 1 | x^{(i)})}{p(y^{(i)} = 1 | t^{(i)} = 1, x^{(i)}) p(t^{(i)} = 1 | x^{(i)}) + p(y^{(i)} = 1 | t^{(i)} = 0, x^{(i)}) p(t^{(i)} = 0 | x^{(i)})} && (2.2) \\
&= \frac{\alpha p(t^{(i)} = 1 | x^{(i)})}{\alpha p(t^{(i)} = 1 | x^{(i)}) + 0 p(t^{(i)} = 0 | x^{(i)})} && (2.3) \\
&= \frac{\alpha p(t^{(i)} = 1 | x^{(i)})}{\alpha p(t^{(i)} = 1 | x^{(i)})} = 1 && (2.4)
\end{aligned}
$$

■

2.2 Question 2(d)

Proof.

$$\begin{aligned}
p(t^{(i)} = 1|x^{(i)}) &= p(t^{(i)}, y^{(i)} = 1|x^{(i)}) + p(t^{(i)} = 1, y^{(i)} = 0|x^{(i)}) && (2.5)\\
&= p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)})p(y^{(i)} = 1|x^{(i)}) && (2.6)\\
&+ p(y^{(i)} = 0|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)}) && (2.7)\\
&= 1p(y^{(i)} = 1|x^{(i)}) + (1 - \alpha)p(t^{(i)} = 1|x^{(i)}) && (2.8)\\
\implies p(t^{(i)} = 1|x^{(i)}) &= \frac{1}{\alpha}p(y^{(i)} = 1|x^{(i)}) && (2.9)
\end{aligned}$$

■

2.3 Question 2(e)

Proof.

$$
\begin{aligned}
h(x^{(i)}) &= p(y^{(i)} = 1|x^{(i)}) && (2.10)\\
\implies \mathbb{E}[h(x^{(i)})|y^{(i)} = 1] &= \mathbb{E}[p(y^{(i)} = 1|x^{(i)})|y^{(i)} = 1] && (2.11)\\
&= \mathbb{E}\{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)}) && (2.12)\\
&+ p(y^{(i)} = 1|t^{(i)} = 0, x^{(i)})p(t^{(i)} = 0|x^{(i)})|y^{(i)} = 1\} && (2.13)\\
&= \mathbb{E}[\alpha p(t^{(i)} = 1|x^{(i)}) + 0|y^{(i)} = 1] && (2.14)\\
&= \alpha\mathbb{E}[p(t^{(i)} = 1|x^{(i)})|y^{(i)} = 1] && (2.15)
\end{aligned}
$$

From part (c), we proved that given $y^{(i)} = 1$, $t^{(i)} = 1$ with probability 1, conditioned on $x^{(i)}$. Hence,

$$
\begin{aligned}
\mathbb{E}[p(t^{(i)} = 1|x^{(i)})|y^{(i)} = 1] = 1 && (2.16)\\
\implies \mathbb{E}[h(x^{(i)})|y^{(i)} = 1] = \alpha && (2.17)
\end{aligned}
$$

■

3 Question 3: Poisson Regression

3.1 Question 3(a)

Proof.

$$
\begin{aligned}
p(y;\lambda) &= \frac{\exp(-\lambda)\lambda^y}{y!} && (3.1)\\
&= \exp\log(\frac{\exp(-\lambda)\lambda^y}{y!}) && (3.2)\\
&= \exp\left(-\lambda + y\log(\lambda) - \log(y!)\right) && (3.3)\\
&= \frac{1}{y!}\exp(\log(\lambda)y - \lambda) && (3.4)
\end{aligned}
$$

therefore, Poisson distribution belongs to the exponential family with

$$
\begin{aligned}
b(y) &:= \frac{1}{y!} && (3.5)\\
\eta(\lambda) &:= \log(\lambda) && (3.6)\\
T(y) &:= y && (3.7)\\
a(\eta) &:= \exp(\eta) = \lambda && (3.8)
\end{aligned}
$$

∎

3.2 Question 3(b)

Answer. By definition, the canonical response function maps η to the expectation $\mathbb{E}[T(y);\eta]$, which equals $\mathbb{E}[y;\eta] = \lambda$ here. Based on the fact that $\eta(\lambda) = \log(\lambda)$, the exponential function maps $\eta(\lambda)$ to $\mathbb{E}[y;\eta]$. Hence, the canonical response function here is the exponential function. ■

3.3 Question 3(c)

Derive.

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}\log(p(y^{(i)}|x^{(i)};\theta)) &= \frac{\partial}{\partial\theta_j}\left(\log(b(y)) + \eta^T y^{(i)} - a(\eta)\right) && (3.9)\\
&= \frac{\partial}{\partial\theta_j}\left(\theta^T x^{(i)} y^{(i)} - \exp(\theta^T x^{(i)})\right) && (3.10)\\
&= x_j^{(i)} y^{(i)} - \exp(\theta^T x^{(i)}) x_j^{(i)} && (3.11)\\
&= \left(y^{(i)} - \exp(\theta^T x^{(i)})\right) x_j^{(i)} && (3.12)
\end{aligned}$$

The stochastic gradient ascent update rule for parameter θ_j is

$$\theta_j \leftarrow \theta_j + \alpha\left(y^{(i)} - \exp(\theta^T x^{(i)})\right) x_j^{(i)} \tag{3.13}$$

where $(x^{(i)}, y^{(i)})$ is the randomly selected sample, and $\alpha > 0$ denotes the learning rate. ■

4 Question 4: Convexity of Generalized Linear Models

4.1 Question 4(a)

Proof. The mean of y is simply

$$\begin{aligned}
\mathbb{E}[Y;\eta] &= \int_{\mathbb{R}} y p(y;\eta)\ dy && (4.1)\\
&= \int_{\mathbb{R}} y b(y)\exp(\eta y - a(\eta))\ dy && (4.2)
\end{aligned}$$

By definition of probability measure, it must be the case that

$$\int_{\mathbb{R}} p(y;\eta)\ dy = 1 \tag{4.3}$$

for every valid η. Therefore,

$$\begin{aligned}
&\int_{\mathbb{R}} b(y)\exp(\eta y - a(\eta))\ dy = 1 && (4.4)\\
\implies &\int_{\mathbb{R}} b(y)\exp(\eta y)\frac{1}{\exp(a(\eta))}\ dy = 1 && (4.5)\\
\implies &\int_{\mathbb{R}} b(y)\exp(\eta y)\ dy = \exp(a(\eta)) && (4.6)\\
\implies &\frac{\partial \exp(a(\eta))}{\partial \eta} = \frac{\partial}{\partial \eta}\int_{\mathbb{R}} b(y)\exp(\eta y)\ dy && (4.7)\\
\implies &a'(\eta)\exp(a(\eta)) = \int_{\mathbb{R}} b(y)\frac{\partial \exp(\eta y)}{\partial \eta}\ dy && (4.8)\\
\implies &a'(\eta) = \int_{\mathbb{R}} y b(y)\exp(\eta y)\frac{1}{\exp(a(\eta))}\ dy && (4.9)\\
\implies &a'(\eta) = \int_{\mathbb{R}} y b(y)\exp(\eta y - a(\eta))\ dy && (4.10)\\
&= \mathbb{E}[Y;\eta] && (4.11)
\end{aligned}$$

■

4.2 Question 4(b)

Proof. From part (a),

$$
\begin{aligned}
a'(\eta) &= \int_{\mathbb{R}} yb(y) \exp(\eta y - a(\eta))\ dy && (4.12)\\
\implies \frac{\partial^2 a(\eta)}{\partial \eta^2} &= \frac{\partial}{\partial \eta} \int_{\mathbb{R}} yb(y) \exp(\eta y - a(\eta))\ dy && (4.13)\\
&= \int_{\mathbb{R}} yb(y) \exp(ny - a(\eta)) \big(y - a'(\eta)\big)\ dy && (4.14)\\
&= \int_{\mathbb{R}} y^2 b(y) \exp(ny - a(\eta))\ dy - a'(\eta) \int_{\mathbb{R}} yb(y) \exp(ny - a(\eta))\ dy && (4.15)\\
&= \mathbb{E}[Y^2; \eta] - a'(\eta)\mathbb{E}[Y; \eta] && (4.16)\\
&= \mathbb{E}[Y^2; \eta] - \mathbb{E}[Y; \eta]^2 && (4.17)\\
&= \mathbb{V}[Y; \eta] && (4.18)
\end{aligned}
$$

■

4.3 Question 4(c)

Proof. ■

5 Question 5: Linear Regression

5.1 Question 5(a)

$$J(\theta) := \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \phi(x^{(i)}) \right)^2 \tag{5.1}$$

$$\theta \leftarrow \theta + \alpha \sum_{i=1}^{n} \left(y^{(i)} - \theta^T \hat{x}^{(i)} \right) \hat{x}^{(i)} \tag{5.2}$$

where α denotes the learning rate.