This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

$$\frac{\partial}{\partial \eta}\int p(y;\eta)\,dy = 0$$

$$\frac{\partial}{\partial \eta}\int p(y;\eta)\,dy = \int \frac{\partial}{\partial \eta} p(y;\eta)\,dy \underset{\uparrow}{=} \int \frac{\partial}{\partial \eta} b(y)\exp(\eta y - a(\eta))\,dy$$

$$p(y;\eta) = b(y)\exp(\eta y - a(\eta))$$

$$= \int b(y)\underbrace{\frac{\partial}{\partial \eta}\exp(\eta y - a(\eta))}_{\text{use chain rule}}dy$$

$$= \int \underbrace{b(y)\exp(\eta y - a(\eta))}\cdot\left(y - \frac{\partial a(\eta)}{\partial \eta}\right)dy$$

$$= \int p(y;\eta)\cdot\left(y - \frac{\partial a(\eta)}{\partial \eta}\right)dy$$

$$= \int p(y;\eta)\cdot y - p(y;\eta)\cdot\frac{\partial a(\eta)}{\partial \eta}\,dy$$

$$= \int p(y;\eta)\,y\,dy - \int p(y;\eta)\cdot\frac{\partial a(\eta)}{\partial \eta}\,dy$$

$$= \underbrace{\int p(y;\eta)\,y\,dy} - \frac{\partial a(\eta)}{\partial \eta}\underbrace{\int p(y;\eta)\,dy}_{1}$$

$$= E(Y;\eta) - \frac{\partial a(\eta)}{\partial \eta}$$

$$\because \frac{\partial}{\partial \eta}\int p(y;\eta)\,dy = 0$$

$$\therefore E(Y;\eta) = \frac{\partial a(\eta)}{\partial \eta} \underset{\uparrow\ \eta = \theta^T x}{\Rightarrow} E(Y \mid \theta^T x) = \frac{\partial a(\eta)}{\partial \eta}$$

1.b In 1(a), we know $\int y\, p(y;\eta)\,dy = E(Y;\eta) = \frac{\partial}{\partial \eta} a(\eta)$

$$\therefore \frac{\partial^2}{\partial \eta^2} a(\eta) = \frac{\partial}{\partial \eta} \int y\, p(y;\eta)\,dy$$

$\therefore$ It suffice to show $\frac{\partial}{\partial \eta} \int y\, p(y;\eta)\,dy = Var(Y;\eta)$

$$\frac{\partial}{\partial \eta} \int y\, p(y;\eta)\,dy = \int y \frac{\partial}{\partial \eta} p(y;\eta)\,dy$$

$p(y;\eta) = b(y)\exp(\eta y - a(\eta))$

$$= \int y \frac{\partial}{\partial \eta} b(y)\exp(\eta y - a(\eta))\,dy$$

$$= \int y\, \underbrace{b(y)\exp(\eta y - a(\eta))} \cdot \left(y - \frac{\partial a(\eta)}{\partial \eta}\right) dy$$

$$= \int y\, p(y;\eta)\left(y - \frac{\partial a(\eta)}{\partial \eta}\right) dy$$

$$= \int y^2 p(y;\eta) - y\, p(y;\eta) \frac{\partial a(\eta)}{\partial \eta}\, dy$$

$$= \int y^2 p(y;\eta)\,dy - \frac{\partial a(\eta)}{\partial \eta} \int y\, p(y;\eta)\,dy$$

$$= E[Y^2;\eta) - \frac{\partial a(\eta)}{\partial \eta} E[Y;\eta]$$

$\because$ by 1(a) $\frac{\partial a(\eta)}{\partial \eta} = E[Y;\eta]$

$$= E[Y^2;\eta) - E^2[Y;\eta]$$

$$= Var[Y;\eta]$$

$$\therefore Var[Y;\eta] = \frac{\partial^2}{\partial \eta^2} a(\eta)$$

1.c Max $\log p(y^{(i)} | x^{(i)}; \theta)$ is equivalent as $\min \underbrace{-\log p(y^{(i)}|x^{(i)};\theta)}_{\text{lost function}}$

$$\therefore \ell(\theta) = -\sum_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta),\ m \text{ is \# of samples}$$

$$p(y^{(i)};\eta) = b(y^{(i)})\cdot \exp(\eta y^{(i)} - a(\eta)) \Rightarrow = -\sum_{i=1}^{m} \log\left[b(y^{(i)})\cdot \exp(\eta y^{(i)} - a(\eta))\right]$$

$$= -\sum_{i=1}^{m} \log b(y^{(i)}) + (\eta y^{(i)} - a(\eta))$$

$$= \sum_{i=1}^{m} -\log b(y^{(i)}) - \theta^T x^{(i)} y^{(i)} + a(\theta^T x^{(i)})$$

To show $\ell(\theta)$ is convex, it suffice to show Hessian of $\ell(\theta)$ w.r.t. θ is PSD

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} -\log b(y^{(i)}) - \theta^T x^{(i)} y^{(i)} + a(\theta^T x^{(i)})$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j}\left(-\log b(y^{(i)}) - \theta^T x^{(i)} y^{(i)} + a(\theta^T x^{(i)})\right)$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j} a(\theta^T x^{(i)}) \cdot x_j^{(i)} - y^{(i)} x_j^{(i)} = \sum_{i=1}^{m}\left[\frac{\partial}{\partial \theta_j} a(\theta^T x^{(i)}) - y^{(i)}\right] x_j^{(i)}$$

$$H_{jk} = \frac{\partial^2 \ell(\theta)}{\partial \theta_j \partial \theta_k} = \frac{\partial}{\partial \theta_k}\left(\sum_{i=1}^{m}\left[\frac{\partial}{\partial \theta_j} a(\theta^T x^{(i)}) - y^{(i)}\right] x_j^{(i)}\right)$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \theta_k}\left[\frac{\partial}{\partial \theta_j} a(\theta^T x^{(i)}) - y^{(i)}\right] x_j^{(i)}$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \theta_k}\frac{\partial}{\partial \theta_j} a(\theta^T x^{(i)}) x_j^{(i)} x_k^{(i)}$$

for $z \in \mathbb{R}^n$

$$z^T H z = \sum_{j=1}^{n}\sum_{k=1}^{n} H_{jk} z_j z_k = \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{i=1}^{m} \frac{\partial}{\partial \theta_k}\frac{\partial}{\partial \theta_j} a(\theta^T x^{(i)}) x_j^{(i)} x_k^{(i)} z_j z_k$$

$$= \sum_{i=1}^{m} \frac{\partial^2}{\partial \eta^2} a(\eta)\left[x^{(i)T} z\right]^2,\ (\eta = \theta^T x)$$

$\because$ By 1(b), $\frac{\partial^2}{\partial \eta^2} a(\eta) = \mathrm{var}[Y;\eta] \geq 0$ and $x^{(i)T} z \geq 0$

$\therefore z^T H z \geq 0 \Rightarrow \ell(\theta)$ is convex

2.a

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T\phi(x^{(i)}) - y^{(i)})^2 = \frac{1}{2}\sum_{i=1}^{m}(\theta^T\hat{x}^{(i)} - y^{(i)})^2$$

Differentiating this objective, we get:

$$\nabla_\theta J(\theta) = (y^{(i)} - h_\theta(\hat{x}^{(i)}))\,\hat{x}^{(i)}$$

The gradient descent update rule is

$$\theta := \theta - \lambda \nabla_\theta J(\theta)$$

which reduces here to:

Repeat until convergence {

$$\theta_j := \theta_j + \lambda \sum_{i=1}^{m}(y^{(i)} - h_\theta(\hat{x}^{(i)}))\,\hat{x}_j^{(i)} \text{ for every } j$$

}

OR in a more succinct way

$$\theta := \theta + \lambda \sum_{i=1}^{m}(y^{(i)} - h_\theta(\hat{x}^{(i)}))\,\hat{x}^{(i)} \quad (m \text{ is \# of samples})$$

2.d

when $k=1$, $h_\theta(x) = \theta_0 + \theta_1 x$, the blue line doesn't fit the data

when $k=2$, $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$, the orange line does not fit the data

when $k=3$, $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$, the green line starts to converge to the sample dots. But still underfit

when $k=5$, $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_5 x^5$, the fit of the red curve improves a lot.

when $k=10$, $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$, the fit of the purple curve is kind of similar to the red curve ($k=5$)

when $k=20$, $h_\theta(x) = \theta_0 + \theta_1 x + \cdots \theta_{20} x^{20}$, the brown curve shows overfit

2.f By adding a sin(x) term, the model in 2(e) has better fit than the model in 2(c) especially when k is small.

2.h As k increases, the fit of the training set (small.csv) goes crazier.

The reason is with higher k, we would create more features and start to get bigger than the number of samples in the training set.

Thus as k increases, the model becomes more "flexible" or less robust on a smaller training set.