1. a.

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

we can get: $E[Y;\eta] = \int_{-\infty}^{+\infty} y\, p(y;\eta)\, dy$

$$= \int_{-\infty}^{+\infty} y\, b(y)\exp(\eta y - a(\eta))\, dy \quad (1)$$

since $\int_{-\infty}^{+\infty} p(y;\eta)\, dy = 1$

we can get: $\int_{-\infty}^{+\infty} b(y)\exp(\eta y - a(\eta))\, dy = 1$

which yields: $a(\eta) = \log\left[\int_{-\infty}^{+\infty} b(y)\exp(\eta y)\, dy\right] \quad (2)$

replace $a(\eta)$ in (1) with (2), we can get

$$E[Y;\eta] = \frac{\int_{-\infty}^{+\infty} y\, b(y)\exp(\eta y)\, dy}{\exp(a(\eta))}$$

$$= \frac{\int_{-\infty}^{+\infty} y\, b(y)\exp(\eta y)\, dy}{\int_{-\infty}^{+\infty} b(y)\exp(\eta y)\, dy} \quad (3)$$

from (2), we can further get

$$\frac{\partial a(\eta)}{\partial \eta} = \frac{\frac{\partial}{\partial \eta}\int_{-\infty}^{+\infty} b(y)\exp(\eta y)\, dy}{\int_{-\infty}^{+\infty} b(y)\exp(\eta y)\, dy}$$

$$= \frac{\int_{-\infty}^{+\infty} y\, b(y)\exp(\eta y)\, dy}{\int_{-\infty}^{+\infty} b(y)\exp(\eta y)\, dy} \quad (4)$$

And from (3) and (4), we can finally get

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

1.b

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

$$= \int_{-\infty}^{+\infty} y^2 p(y;\eta)\,dy - \left(\int_{-\infty}^{+\infty} y p(y;\eta)\,dy\right)^2$$

we can get

$$= \frac{\partial^2 a(\eta)}{\partial \eta^2} = \frac{\partial}{\partial \eta}\left[\frac{\partial}{\partial \eta} a(\eta)\right]$$

$$= \frac{\partial}{\partial \eta} E[Y;\eta]$$

$$= \frac{\partial}{\partial \eta} \int_{-\infty}^{+\infty} y\, b(y) \exp(\eta y - a(\eta))\,dy$$

$$= \int_{-\infty}^{+\infty} y^2 b(y) \exp(\eta y - a(\eta))\left(y - \frac{\partial}{\partial \eta} a(\eta)\right) dy$$

$$= \int_{-\infty}^{+\infty} y^2 b(y) \exp(\eta y - a(\eta))\,dy$$

$$- \frac{\partial}{\partial \eta} a(\eta) \int_{-\infty}^{+\infty} y\, b(y) \exp(\eta y - a(\eta))\,dy$$

$$= \int_{-\infty}^{+\infty} y^2 p(y;\eta)\,dy - (E[Y;\eta])^2$$

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

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

1.c.

With $\ell(\theta) = -\log\left(\prod_{i=1}^{n} p(y_i;\eta)\right)$

$$= -\log\left(\prod_{i=1}^{n} b(y_i)\exp(\eta y_i - a(\eta))\right)$$

$$= -\sum_{i=1}^{n}\eta y_i + \sum_{i=1}^{n} a(\eta) - \log\prod_{i=1}^{n} b(y_i)$$

We can get

$$\frac{\partial^2 \ell(\theta)}{\partial \eta^2} = \frac{\partial^2}{\partial \eta^2}\sum_{i=1}^{n} a(\eta)$$

$$= n \cdot \frac{\partial^2}{\partial \eta^2} a(\eta)$$

$$= n \cdot \mathrm{Var}[Y;\eta]$$

and we can conclude that $\frac{\partial^2 \ell(\theta)}{\partial \eta^2}$ is always positive semi-definite.

2.a.

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left[\theta^T\hat{x}^{(i)} - y^{(i)}\right]^2$$

$$= \frac{1}{2}(\hat{x}\theta - y)^T(\hat{x}\theta - y)$$

Differentiate this objective, we get

$$\nabla_\theta J(\theta) = \hat{x}^T\hat{x}\theta - \hat{x}^T y$$

The gradient descent update rule is

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

which reduces here to:

$$\theta := \theta - \alpha(\hat{x}^T\hat{x}\theta - \hat{x}^T y)$$

2.d As k increases from 1, 2, ... to 20, the polynomial regression starts from "under-fitting" the sample dataset to "over-fitting" the sample dataset (e.g., k=1,2,3 under fit and k=20 is "overfit")

2.f. Adding sin(x) as additional feature reduces the "under-fitting" problem when k is small.

2.h. As k increases, polynomial regression performs from "underfitting" to "overfitting"; while polynomial regression with sinusoidal features has a better performance, but still "overfitting" the training dataset.