1.a

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

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

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

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

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

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

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

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

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

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

$$\frac{\partial a(q)}{\partial q} = E[y; q]$$

1.6

$$\mathrm{Var}[y;\eta] = E[y^2;\eta] - (E[y;\eta])^2$$

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

we can get next:

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

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

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

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

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

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

$$= \int_{-\infty}^{+\infty} y^2 p(y)$$

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

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

$$= \mathrm{Var}[y;\eta]$$

1c.

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

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

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

we get

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

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

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

so we can conclude that $\frac{\partial^2 l(\theta)}{\partial \eta^2}$ is always PSD.

2.a

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

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

thus we get:

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

the updated rule of the batch gradient descent will be:

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

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

2.c

As we increase k from 1 to 20 the polynomial regression model transitions from underfitting the sample dataset to overfitting it. This trend can be observed @ k=1, k=2, k=3 (underfitting) and at k=20 (overfitting). It becomes too complex and starts to fit the noise in the data, resulting in poor generalization performance on new, unseen data.

2.c

Adding $\sin(x)$ as additional feature decreases the underfitting when $k$ is small.

2.g

Increasing the value of $k$ in polynomial regression leads to a transition from underfitting to overfitting of the training dataset. Although using sinusoidal features in polynomial regression improves its performance it still suffers from overfitting on the training dataset.

3.a

When training on a dataset A, gradient descent converges in 30372 iterations. On the other hand, training on a dataset B fails to converge.

3.b.

Dataset B appears to have perfect linear separability, meaning that it is possible to draw a hyperplane to completely separate the data points, but it is not the case for dataset A.

Logistic regression attempts to maximize the log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{m} y^i h_\theta(x^i) + (1-y^i)(1-h_\theta(x^i))$$

Where $h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$ is the sigmoid function.

The issue with dataset B is a scaling problem. As the data is linearly separable, $\|\theta\|$ can be arbitrarily large and this will increase the log-likelihood indefinitely, as $\forall i, h_\theta(x^i) \to 1$ when $\|\theta\| \to \infty$. This happens so long as $\theta$ describes a hyperplane that separates the data perfectly.

It is also can be said that $\nabla_\theta = X^T(y - h_\theta(X))$ can be made arbitrarily close to 0 by increasing $\theta$, so long as $\theta$ describes a hyperplane that separates the data perfectly.

It doesn't happen when the data is not linearly separable. The log-likelihood would start decreasing due to the misclassified data points.

$\exists! \theta$ such that $X^T(y - h_\theta(X)) = 0$.

3.c

i. It is scaling problem. Changing the learning rate would not solve the problem.

ii. Changing the learning rate wouldn't solve the scaling problem.

iii. This would help solving our problem. We would be actively penalizing large values of $\theta$ by including a term of proportional to $\|\theta\|^2$ in the objective function.

iv. The problem cannot be solved by applying a linear transformation since it will not change the linear separability of the data.

v. This could help. The data could be made to be linearly inseparable. However, this approach is not entirely reliable because the convergence of the algorithm depends on the set of noise samples. Therefore, it cannot be guaranteed that the algorithm will converge for all possible sets of noise samples, making it not very robust.