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

$P(y;n) = b(y)\exp(ny - a(n))$

$\frac{\partial}{\partial n} \int P(y;n)\,dy = \int \frac{\partial}{\partial n} P(y;n)\,dy$

$\int \frac{\partial}{\partial n} P(y;n)\,dy = \int \frac{\partial}{\partial n}\left[b(y)\exp(ny - a(n))\,dy\right] = 0$

$= \frac{\partial}{\partial n}\left[b(y)\exp(ny - a(n))\right] = b(y)\exp(ny - a(n)) \cdot (y - a'(n))$

$\int b(y)\exp(ny - a(n)) \cdot (y - a'(n))\,dy = 0$

$\int (y - a'(n))\,P(y;n)\,dy = 0$

$\int y P(y;n)\,dy - a'(n)\int P(y;n)\,dy = 0$

$E[Y;n] - a'(n) = 0 \rightarrow E[Y;n] = \frac{\partial}{\partial n} a(n)$

$\boxed{\therefore\ E[Y|X;\theta] = E[Y;n] = \frac{\partial}{\partial n} a(n)}$

1.b Show $Var(Y;n) = \frac{\partial^2}{\partial n^2} a(n)$

From 1.A) $E[Y;n] = a'(n) = \mu(n)$

$Var(Y;n) = E[(Y-\mu)^2] = E[Y^2] - \mu^2$, Find Y

$\partial/\partial n \, E[Y;n] = \partial/\partial n \, a'(n) = a''(n)$, $\partial/\partial n \, E[Y;n] = \partial/\partial n \int y \cdot P(y;n) \, dy$

$= \int y \cdot \partial/\partial n \, P(y;n) \, dy$

$\therefore \int y \cdot P(y;n) \cdot (y - a'(n)) \, dy = \int y(y-\mu) \cdot P(y;n) \, dy$

$= E[Y(Y-\mu)] = E[Y^2] - \mu^2 = Var(Y)$

$$\boxed{\therefore \partial/\partial n \, E[Y] = Var(Y) = a''(n)}$$

1.c Show Hessian is always PSD

let $\eta = \theta^T x$, $\frac{\partial L(\theta)}{\partial \theta} = (a'(\eta) - y) \cdot x$

$\nabla_\theta^2 L(\theta) = \partial / \partial \theta \left[(a'(\eta) - y) \cdot x \right] = a''(\eta) \cdot x x^T$

Since in 1.B) $a''(\eta) = Var(Y; \eta) \geq 0$ and xx^T is one-rank PSD,

$$\boxed{\begin{array}{l} \therefore \ \nabla_\theta^2 L(\theta) = Var(Y; \eta) \cdot x x^T \geq 0 \\ \text{Hessian is always PSD} \end{array}}$$

$loss(\theta) = a(\theta^T x) - y \cdot \theta^T x$

Hessian $= \nabla_\theta^2 L(\theta) = Var(Y; \eta) \cdot x x^T$

Since $Var(Y; \eta) \geq 0$, Hessian is <u>PSD</u>

2.a

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

Differentiating this objective, we get:

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

The gradient descent update rule is

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

which reduces here to:

$$\theta := \theta - \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\left(\theta^T \hat{x}^{(i)} - y^{(i)}\right)\hat{x}^{(i)}\right)$$

$$= \boxed{\theta - \frac{\lambda}{n}\sum_{i=1}^{n}\left(\theta^T \hat{x}^{(i)} - y^{(i)}\right)\hat{x}^{(i)}}$$

2.c

As $k\uparrow$, we can fit the data much closer, allows us to better represent complex data with a lot of curves.

2.e

The Sine waves generally need less parameters in the feature map than the polynomial. $k=3$ fits all the small points in the sine plots, where in the poly plot it is very underfitted.

2.g The data overfits the data at very low k counts. As k increases, the overfitting gets worse.

3.a
3.b
3.c

3.a) Dataset B never Converges

3.B) The data in CSV_B is linearly seperable, and thus we have no finite solution for maximum likelihood estimation. See plot "plot-a.Prg" and "plot-b.Prg".

3.c) i) no, because there is no closed form solution to converge to.

ii) no, same as i

iii) no, same as ii and i

iv) yes, gives the data a finite minimum

v) yes, will make data non-linearly seperable