Problem 7

a) To show that the Poisson distribution belongs to the single parameter exponential family, we need to express it's probability mass function (PMF) in the following form:

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

where:

- η is the natural parameter.
- $T(y)$ is the sufficient statistic.
- $a(\eta)$ is the log-partition function.
- $b(y)$ is the base measure.

PMF of the Poisson distribution:

$$P(Y=y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots$$

Rewriting the PMF:

$$\lambda^y = e^n$$

where ,

$$n = \log(\lambda^y) \implies n = y\log(\lambda)$$

We can rewrite the PMF as:

$$P(Y=y) = \frac{e^{-\lambda}e^{y\log(\lambda)}}{y!}, \quad y = 0, 1, 2, \ldots$$

$$\implies P(Y=y) = \frac{1}{y!}\exp(-\lambda + y\log(\lambda))$$

Comparing with exponential family form

- $\eta = \log(\lambda)$ (natural parameter)
- $T(y) = y$ (sufficient statistic)
- $a(\eta) = \lambda = \exp(\eta)$ (log-partition function)
- $b(y) = \frac{1}{y!}$ (base measure)

Expressing the PMF in exponential family form:

Substituting these components into the general form, we get:

$$\begin{aligned} P(Y=y) &= \frac{1}{y!}\exp(\log(\lambda)\cdot y - \exp(\log(\lambda))) \\ &= b(y)\exp(\eta^T T(y) - a(\eta)) \end{aligned}$$

b) (a) **The log-likelihood function of θ, given S.**
Given the dataset $S = \{(x^{(i)}, y^{(i)}), x^{(i)} \in R^k, y^{(i)} \in Z^+, i = 1, 2, ..., n\}$, and the assumption that given x, y follows a Poisson distribution with rate $e^{\theta^T x}$, the likelihood function is:

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} P(y^{(i)}|x^{(i)};\theta) \\
&= \prod_{i=1}^{n} \frac{e^{-e^{\theta^T x^{(i)}}}(e^{\theta^T x^{(i)}})^{y^{(i)}}}{y^{(i)}!}
\end{aligned}
$$

The log-likelihood function is then:

$$
\begin{aligned}
l(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{n} \log\left(\frac{e^{-e^{\theta^T x^{(i)}}}(e^{\theta^T x^{(i)}})^{y^{(i)}}}{y^{(i)}!}\right) \\
&= \sum_{i=1}^{n} \left(-e^{\theta^T x^{(i)}} + y^{(i)}\theta^T x^{(i)} - \log(y^{(i)}!)\right)
\end{aligned}
$$

(b) The gradient of the log-likelihood with respect to θ is:

$$
\begin{aligned}
\nabla l(\theta) &= \frac{\partial l(\theta)}{\partial \theta} \\
&= \sum_{i=1}^{n} \left(-e^{\theta^T x^{(i)}} x^{(i)} + y^{(i)} x^{(i)}\right) \\
&= \sum_{i=1}^{n} \left(y^{(i)} - e^{\theta^T x^{(i)}}\right) x^{(i)}
\end{aligned}
$$

(c) Since we want to maximize the log-likelihood, we will use gradient ascent, which is the same algorithm but with a positive step size.
1.Initialization: Choose an initial value for θ, say $\theta^{(0)}$.

2.Iteration: Update θ using the following rule:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \nabla l(\theta^{(t)})$$

where $\alpha > 0$ is the learning rate (step size) and $\nabla l(\theta^{(t)})$ is the gradient of the log-likelihood evaluated at $\theta^{(t)}$.

3.Convergence: Repeat step 2 until convergence, which can be determined by monitoring the change in $l(\theta)$ or in θ itself. For example, stop if $||\theta^{(t+1)} - \theta^{(t)}|| < \epsilon$ for some small tolerance ϵ.

Substituting the gradient we derived earlier, the update rule for gradient ascent becomes:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \sum_{i=1}^{n} \left(y^{(i)} - e^{\theta^T x^{(i)}} \right) x^{(i)}$$