CS 165 | Winter 2023
Instructor: Prof. A. Anandkumar | Caltech

Solution to Homework 1

1 **Statistics and Sufficient Statistics** Consider experiments that produce n i.i.d. observation $Y_i \overset{i.i.d.}{\sim} p(y;\theta)$. For each of the following model, find the log-likelihood function $L(\mathbf{y};\theta)$ and a sufficient statistic of as "low dimension" as possible.

(a) Normal distribution with unknown mean $\mathcal{N}(\theta, 1)$.

(b) Exponential distribution with unknown mean $\mathcal{E}(\theta)$.

(c) Poisson distribution with unknown mean $\mathcal{P}(\theta)$.

(d) Bernoulli with unknown mean $\mathcal{B}(\theta)$.

(e) Uniform distribution $\mathcal{U}(0, \theta)$.

Solution: First, let's recall the Neyman-Fisher factorization theorem: A statistic $t(\mathbf{y})$ is sufficient if and only if the pdf $p(\mathbf{y};\theta)$ has factorization $p(\mathbf{y};\theta) = g(t(\mathbf{y}),\theta)h(\mathbf{y})$ for some non-negative functions g, h. Equivalently, $t(\mathbf{y})$ is sufficient if and only if the *log-likelihood* function $L(\mathbf{y};\theta) = \log p(\mathbf{y};\theta)$ has factorization $L(\mathbf{y};\theta) = \tilde{g}(t(\mathbf{y}),\theta) + \tilde{h}(\mathbf{y})$ for some functions $\tilde{g}, \tilde{h}$. In the following $\log(\cdot)$ denotes the natural logarithm.

(a) $t(\mathbf{y}) = \sum_{i=1}^n y_i$.

$$p(\mathbf{y};\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(y_i-\theta)^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n (y_i-\theta)^2/2}$$

$$L(\mathbf{y};\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (y_i - \theta)^2 = \underbrace{-\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n y_i^2}_{\tilde{h}(\mathbf{y})} + \underbrace{\theta\sum_{i=1}^n y_i + n\theta^2}_{\tilde{g}(t(\mathbf{y}),\theta)}.$$

Neyman-Fisher Theorem shows that $\sum_{i=1}^n y_i$ is sufficient.

(b) $t(\mathbf{y}) = \sum_{i=1}^n y_i$.

$$p(\mathbf{y};\theta) = \prod_{i=1}^n \frac{1}{\theta} e^{-y_i/\theta} = \frac{1}{\theta^n}\left(-\frac{\sum_{i=1}^n y_i}{\theta}\right)$$

$$L(\mathbf{y};\theta) = \underbrace{-n\log\theta - \frac{1}{\theta}\sum_{i=1}^n y_i}_{\tilde{g}(t(\mathbf{y}),\theta)}.$$

We can apply Neyman-Fisher by taking $\tilde{h}(\mathbf{y}) = 0$.
Note: Some students in the class start by $p(\mathbf{y};\theta) = \theta^n(\theta \sum y_i)$. This is also OK, but beware that this pdf has mean $1/\theta$. The one above has mean θ.

(c) $t(\mathbf{y}) = \sum_{i=1}^n y_i$.

$$p(\mathbf{y};\theta) = \prod_{i=1}^n e^{-\theta}\frac{\theta^{y_i}}{y_i!} = e^{-n\theta}\frac{\theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}$$

$$L(\mathbf{y};\theta) = \underbrace{-n\theta + (\log\theta)\sum_{i=1}^n y_i}_{\tilde{g}(t(\mathbf{y}),\theta)} \underbrace{- \log\prod_{i=1}^n y_i!}_{\tilde{h}(\mathbf{y})}.$$

(d) $t(\mathbf{y}) = \sum_{i=1}^n y_i$.

$$p(\mathbf{y};\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} = \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n y_i}(1-\theta)^n$$

$$L(\mathbf{y};\theta) = \underbrace{\left(\sum_{i=1}^n y_i\right)\log\frac{\theta}{1-\theta} + n\log(1-\theta)}_{\tilde{g}(t(\mathbf{y}),\theta)}.$$

Take $\tilde{h}(\mathbf{y}) = 0$.

(e) $t(\mathbf{y}) = \max_{i=1,\cdots,n} y_i$.

$$p(\mathbf{y};\theta) = \prod_{i=1}^n \frac{1}{\theta}1(0 \le y_i \le \theta) = \frac{1}{\theta^n}1(0 \le \min_{i=1,\cdots,n} y_i)1(\max_{i=1,\cdots,n} y_i \le \theta),$$

where $1(\cdot)$ is the indicator function (which is equal to 1 if the statement in $(\cdot)$ is true, and equal to 0 otherwise).

$$L(\mathbf{y};\theta) = \underbrace{-n\log\theta + \log 1(\max_{i=1,\cdots,n} y_i \le \theta)}_{\tilde{g}(t(\mathbf{y}),\theta)} + \underbrace{1(0 \le \min_{i=1,\cdots,n} y_i)}_{\tilde{h}(\mathbf{y})}.$$

2 **Gaussian Mixture** Suppose that Y_i is an i.i.d. sequence drawn from $\mathcal{N}(\theta,1)$, and $\mathbf{Y} = (Y_1,\cdots,Y_n)$. We know that $t(\mathbf{Y}) = \sum_i Y_i$ is a sufficient statistic. Consider next the model involving a Bernoulli random variable $X \sim \mathcal{B}(\frac{1}{4})$ in which

$$Y \sim \begin{cases} \mathcal{N}(\theta,1) & X=0 \\ \mathcal{N}(\theta,2) & X=1 \end{cases}$$

(a) Show that $(\sum_i Y_i, X)$ is a sufficient statistic.

(b) Is $\sum_i Y_i$ a sufficient statistic?

Solution: When X is observable, the likelihood of Y is conditioned on X.

$$p_{\mathbf{Y}}(\mathbf{y};\theta|X=0) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}}e^{-(y_i-\theta)^2/2}$$

$$p_{\mathbf{Y}}(\mathbf{y};\theta|X=1) = \prod_{i=1}^n \frac{1}{\sqrt{4\pi}}e^{-(y_i-\theta)^2/4}$$

Since $\sum_i Y_i$ is sufficient statistic for both the conditional distributions, $(X, \sum_i Y_i)$ is sufficient.

When X is not observable,the likelihood of Y is a mixture gaussian,

$$p_{\mathbf{Y}}(\mathbf{y};\theta) \quad = \quad P(X=0)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-(y_i-\theta)^2/2} + P(X=1)\prod_{i=1}^{n}\frac{1}{\sqrt{4\pi}}e^{-(y_i-\theta)^2/4}$$

The likelihood function cannot be factorized in terms of $\sum_i Y_i$. Therefore, by factorization theorem, $\sum_i Y_i$ is not sufficient.