Homework 6 Solutions

1. (10 points) *Uniform distribution.*

   a. u has to be larger than $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$.

   b. The likelihood function is

$$\mathcal{L}_{\mathbf{x}}(u) = \prod_{i=1}^{n} f_{X_i}(x_i, u) \tag{1}$$
$$= \prod_{i=1}^{n} 1/u = 1/u^n \tag{2}$$

   for $u \geq x_{(n)}$, and zero else. The likelihood monotonically decreases as u moves away from $x_{(n)}$, therefore the maximum of the likelihood is achieved exactly at $X_{(n)} = \widehat{U}_{\mathrm{ML}}$.

   c. We need to find the density of $X_{(n)}$, to find densities of extreme values its useful to appeal to the cdf:

$$F_{X_{(n)}}(x) = \mathrm{P}(X_{(n)} \leq x) = \mathrm{P}(X_1 \leq x, \ldots, X_n \leq x) \tag{3}$$
$$= \mathrm{P}(X_1 \leq x) \ldots \mathrm{P}(X_n \leq x) \tag{4}$$
$$= \frac{x}{u} \cdots \frac{x}{u} \tag{5}$$

   taking derivative wrt x gives

$$f_{\widehat{U}_{\mathrm{ML}}}(x) = \frac{nx^{n-1}}{u^n} \text{ for } x \in [0, u] \tag{6}$$

   d. It is expected to be biased since $\widehat{U}_{\mathrm{ML}} \leq u$, indeed,

$$\mathrm{E}(\widehat{U}_{\mathrm{ML}}) = \int_1^u x \frac{nx^{n-1}}{u^n} dx \tag{7}$$
$$= \frac{nu}{(n+1)} \tag{8}$$

   e. For any positive epsilon,

$$\mathrm{P}(|\widehat{U}_{\mathrm{ML}} - u| > \epsilon) = \mathrm{P}(u - \widehat{U}_{\mathrm{ML}} > \epsilon) \tag{9}$$
$$\leq \frac{u - \frac{nu}{(n+1)}}{\epsilon} = \frac{u}{(n+1)\epsilon} \text{ by Markov's inequality} \tag{10}$$

   which converges to zero as n tends to infinity.

2. (10 points) *Half life*

   a. Let us denote the half life by h and the parameter of the exponential by λ. We have

$$P(T \leq h) = \int_0^h \lambda e^{-\lambda x} \, \mathrm{d}x \tag{11}$$
$$= 1 - e^{-\lambda h} \tag{12}$$
$$= \frac{1}{2}. \tag{13}$$

This implies that

$$\lambda = \frac{\log 2}{h}. \tag{14}$$

The parameters are $\lambda_1 := 0.283$ for carbon-15, $\lambda_2 := 3.59\,10^{-2}$ for carbon-10 and $\lambda_3 := 2.31\,10^{-2}$ for seaborgium-266.

b. The likelihood is equal to

$$\mathcal{L}_{\mathbf{x}}(p) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{15}$$

$$= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \tag{16}$$

and the log likelihood to

$$\log \mathcal{L}_{\mathbf{x}}(p) = \sum_{i=1}^{n} (\log \lambda - \lambda x_i). \tag{17}$$

We compute the derivative and second derivative of the log-likelihood function,

$$\frac{\mathrm{d} \log \mathcal{L}_{\mathbf{x}}(\lambda)}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i, \tag{18}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{\mathbf{x}}(p)}{\mathrm{d}\lambda^2} = -\frac{n}{\lambda^2} < 0. \tag{19}$$

The function is concave, as the second derivative is negative. The maximum is consequently at the point where the first derivative equals zero, namely

$$\hat{\lambda}_{\mathrm{ML}} = \frac{n}{\sum_{i=1}^{n} x_i}. \tag{20}$$

This maximum will almost surely not be one of the predefined values for the parameter of the exponential. It would make more sense to restrict the maximization to those three points, i.e. compute their likelihoods and choose the one with the highest value.

c. Let Λ be a discrete random variable such that if p_i denotes the probability that the sample comes from isotope i,

$$p_\Lambda(\lambda_i) = p_i \quad \text{for } i \in \{1,2,3\}. \tag{21}$$

By Bayes rule, the posterior is of the form

$$p_{\Lambda|\mathbf{X}}(\lambda_i|\mathbf{x}) = \frac{p_\Lambda(\lambda_i)\, f_{\mathbf{X}|\Lambda}(\mathbf{x}|\lambda_i)}{\sum_{j=1}^{3} p_\Lambda(\lambda_j)\, f_{\mathbf{X}|\Lambda}(\mathbf{x}|\lambda_j)} \tag{22}$$

$$= \frac{p_i \lambda_i^n e^{-\lambda_i \sum_{k=1}^{n} x_k}}{\sum_{j=1}^{3} p_j \lambda_j^n e^{-\lambda_j \sum_{k=1}^{n} x_k}}. \tag{23}$$

d. The posterior mean is equal to

$$\mathrm{E}\left(\Lambda|\mathbf{X}\right) = \sum_{i=1}^{3} \lambda_i p_{\Lambda|\mathbf{X}}\left(\lambda_i|\mathbf{x}\right). \tag{24}$$

It will not be equal to λ_1, λ_2 and λ_3, so it does not make a lot of sense to choose it. A better idea is to choose the value of λ_i $i \in \{1, 2, 3\}$ that has higher probability. As we will see later in the course, this is the Maximum A Posteriori (MAP) estimate of the parameter.

e. The code is

```
def post_dist(p, x, lambda_vec):
    dist = np.zeros(3)
    n = len(x)
    for i in range(3):
        dist[i] = p[i] * (lambda_vec[i] ** n) * np.exp(-lambda_vec[i] * np.sum(x))
    dist = dist / np.sum(dist)
    return dist
```

For just two samples, carbon-10 has higher probability. This makes sense since the prior is strongly biased towards it and its corresponding λ is very close to the λ of seaborgium-266. For 100 samples there is enough data for likelihood term to override the prior and seaborgium-266 has higher probability.

3. (10 points) *Empirical probability mass function.*

a. We can calculate the empirical pmf at a point k: $\hat{p}_n(k) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{k=X_i}$. This is just the histogram from the data we have. It approximates the real histogram if n is large.

b. Note that the sum above corresponds to the number of people in the sample at age k.

$$\begin{aligned}
\mathrm{E}(\hat{p}_n(k)) &= \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{k=X_i}) && (25)\\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\mathbf{1}_{k=X_i}) && (26)\\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}(k = X_i) && (27)\\
&= p(k) && (28)
\end{aligned}$$

c. In the mean square norm,

$$\begin{aligned}
\mathrm{E}(|p(k) - \hat{p}_n(k)|^2) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{k=X_i}) && (29)\\
&= \frac{1}{n^2}\sum \mathrm{Var}(\mathbf{1}_{k=X_i}) && (30)\\
&= \frac{\mathrm{P}(k = X_i) - \mathrm{P}^2(k = X_i)}{n} && (31)
\end{aligned}$$

which tends to zero as n tends to infinity.

d. If the age were measured in minutes then the estimated histograms would be very spiky around the individual samples and zero else, giving rise to an irregular shape. You would need to use kernels because the number of data is small with respect to the number of possible values.

4. (10 points) *Call center*

a. The likelihood is equal to

$$\mathcal{L}_{\mathbf{x}}(p) = \prod_{i=1}^{n} p_{X_i}(x_i) \tag{32}$$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \tag{33}$$

and the log likelihood to

$$\log \mathcal{L}_{\mathbf{x}}(p) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log(x_i!) \right). \tag{34}$$

We compute the derivative and second derivative of the log-likelihood function,

$$\frac{\mathrm{d} \log \mathcal{L}_{\mathbf{x}}(\lambda)}{\mathrm{d}\lambda} = \sum_{i=1}^{n} \frac{x_i}{\lambda} - 1, \tag{35}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{\mathbf{x}}(p)}{\mathrm{d}\lambda^2} = -\sum_{i=1}^{n} \frac{x_i}{\lambda^2} < 0. \tag{36}$$

The function is concave, as the second derivative is negative. The maximum is consequently at the point where the first derivative equals zero, namely

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{37}$$

which is just the sample mean.

b. The code is

```
# Function to estimate the parameters of a Poisson r.v. from iid data
def poisson_ml_estimator(data):
    return np.mean(data)


# Function to estimate the empirical pmf from iid data
def empirical_pmf(data):
    max_val = np.amax(data)
    emp_pmf_aux = np.zeros(max_val+1)
    for i in range(len(data)):
        emp_pmf_aux[np.int(data[i])] += 1.
    emp_pmf = emp_pmf_aux / len(data)
    return emp_pmf
```

```
# Function to compute pmf of Poisson with parameter param at x
def poisson_pmf(param, x):
    return np.exp(-param) * (param ** x) / math.factorial(x)
```

The errors are equal to 0.433 (ML estimation) and 0.519 (nonparametric estimation) for the 2-day data and 0.507 (ML estimation) and 0.403 (nonparametric estimation) for the September data.

c. The parametric method performs better for the 2-day data, whereas the nonparametric method performs better for the September data. When we do not have a lot of data, assuming a parametric method helps to find a stable estimate as we only have to estimate a small number of parameters (in this case one). When we have a lot of data, a nonparametric method is often able to adapt more effectively to the data than a parametric model, especially taking into account that the data usually follows a parametric model only approximately (it is obvious from the October data that the true pmf is not Poisson).

5. (10 points) *Method of moments*

a. For a Bernoulli random variable $\mu = p$, so $\hat{p}$ equals the sample mean which is equal to the ML estimator (from the notes).

b. For a geometric random variable $\mu = 1/p$, so $\hat{p} = 1/\bar{x}_n$. The likelihood is equal to

$$\mathcal{L}_{\mathbf{x}}(p) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{38}$$

$$= \prod_{i=1}^{n} (1-p)^{x_i - 1} p \tag{39}$$

and the log likelihood to

$$\log \mathcal{L}_{\mathbf{x}}(p) = \sum_{i=1}^{n} \left((x_i - 1) \log (1-p) + \log p \right) = \log (1-p) \left(\sum_{i=1}^{n} x_i - n \right) + n \log p. \tag{40}$$

We compute the derivative and second derivative of the log-likelihood function,

$$\frac{\mathrm{d} \log \mathcal{L}_{\mathbf{x}}(\lambda)}{\mathrm{d}p} = -\frac{1}{1-p} \left(\sum_{i=1}^{n} x_i - n \right) + \frac{n}{p}, \tag{41}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{\mathbf{x}}(p)}{\mathrm{d}p^2} = -\frac{1}{(1-p)^2} \left(\sum_{i=1}^{n} x_i - n \right) - \frac{n}{p^2} < 0. \tag{42}$$

The function is concave, as the second derivative is negative. The maximum is consequently at the point where the first derivative equals zero, namely

$$\hat{p}_{\mathrm{ML}} = \frac{n}{\sum_{i=1}^{n} x_i}, \tag{43}$$

which is equal to method-of-moments estimator.

c. For a Poisson random variable $\mu = \lambda$ so $\hat{\lambda}$ equals the sample mean which is equal to the ML estimator (from Problem 4).

d. For an exponential random variable $\mu = 1/\lambda$ so $\hat{\lambda} = 1/\bar{x}_n$ which is equal to the ML estimator (from Problem 2).