Please provide complete and well-written solutions to the following exercises.

Due February 7, at the beginning of class.

Homework 2

Exercise 1. Recall that a random variable T is exponential with parameter λ if T has the density function given by $f_T(t) = \lambda e^{-\lambda t}$ for all $t \geq 0$, and $f_T(t) = 0$ otherwise.

Let $\lambda > 0$. Let $\tau_1, \tau_2, \ldots$ be independent exponential random variables with parameter λ. Let $T_0 = 0$, and for any $n \geq 1$, let $T_n := \tau_1 + \cdots + \tau_n$. A **Poisson Process** with parameter $\lambda > 0$ is a set of integer-valued random variables $\{N(s)\}_{s \geq 0}$ defined by $N(s) := \max\{n \geq 0 \colon T_n \leq s\}$.

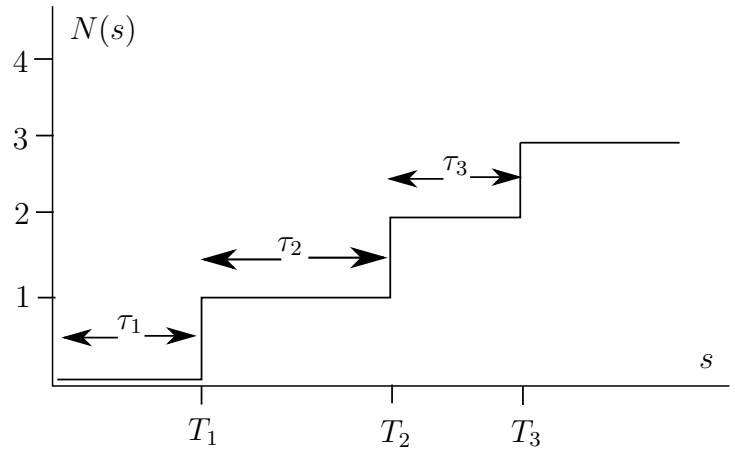

FIGURE 1. One Sample Path of a Poisson Process.

We can think of $N(s)$ as the total number of customers that have arrived in a store, or the total number of people that have visited a website, where the time between each visit is an exponential random variable with an unknown parameter $\lambda > 0$. The use of the exponential distribution in this model is reasonable due to the memoryless property of the exponential random variable. (If τ is an exponential and $t, s > 0$, then $\mathbf{P}(\tau > t + s \,|\, \tau > t) = \mathbf{P}(\tau > s)$.)

Suppose we have observed that $\tau_1 = t_1, \ldots, \tau_m = t_m$ for some $t_1, \ldots, t_m > 0$. The goal of this exercise is to estimate the unknown parameter $\lambda > 0$. One way to estimate λ is via the maximum likelihood. The maximum likelihood estimator is $Y = r(\tau_1, \ldots, \tau_m)$, where $r \colon \mathbf{R}^m \to \mathbf{R}$, and $r(t_1, \ldots, t_m)$ is defined to be any value of $\lambda > 0$ that maximizes

$$\prod_{i=1}^{m} \lambda e^{-\lambda t_i}$$

Show that the maximum likelihood estimator of λ has the form

$$Y = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} \tau_i}.$$

For example, suppose $\tau_1, \ldots, \tau_{17}$ are observed to be

$$138\ 257\ 418\ 263\ 134\ 336\ 145\ 36\ 94\ 213\ 171\ 407\ 69\ 19\ 16\ 215\ 96.$$

Estimate λ.

Exercise 2. This exercise deals with sunspot data from the following files (the same data appears in different formats)

txt file csv (excel) file

These files are taken from http://www.sidc.be/silso/datafiles#total

To work with this data, e.g. in Matlab you can use the command

```
x=importdata('SN_d_tot_V2.0.txt')
```

to import the .txt file.

The format of the data is as follows.

- Columns 1-3: Gregorian calendar date (Year, Month, then Day)
- Column 4: Date in fraction of year
- Column 5: Daily total number of sunspots observed on the sun. A value of -1 indicates that no number is available for that day (missing value).
- Column 6: Daily standard deviation of the input sunspot numbers from individual stations.
- Column 7: Number of observations used to compute the daily value.
- Column 8: Definitive/provisional indicator. A blank indicates that the value is definitive. A '*' symbol indicates that the value is still provisional and is subject to a possible revision (Usually the last 3 to 6 months)

It is known that the number of sunspots on the sun follows an approximately 11-year sinusoidal pattern. So, if we plot the number of sunspots over several years, the distance between the highest observed numbers of sunspots should be around 11 years.

Let U_t be the number of sunspots at time t, where t is measured in years. We model U_t as

$$U_t = m_t + a\cos(2\pi\theta t) + b\sin(2\pi\omega t)) + Y_t, \qquad \forall\, t \in \mathbf{R},$$

where $a, b, \theta, \omega \in \mathbf{R}$ are unknown (deterministic) parameters, m_t is an unknown deterministic function of t that is assumed to be a "slowly varying" function of t, and $\{Y_t\}_{t\in\mathbf{R}}$ are i.i.d. mean zero random variables. The quantity m_t is called the **trend** and the quantity $s_t := a\cos(2\pi\theta t) + b\sin(2\pi\omega t))$ is called the **seasonal component** of the time series $\{U_t\}_{t\in\mathbf{R}}$.

Since the 11-year sinusoidal pattern is known, we assume for now that $\theta = \omega = 1/11$. Note that

$$\sum_{s=t,t+1/365,t+2/365,\ldots,t+11} \cos(2\pi\theta s) \approx 0, \qquad \sum_{s=t,t+1/365,t+2/365,\ldots,t+11} \cos(2\pi\theta s) \approx 0, \qquad \forall\, t \in \mathbf{R}.$$

So, if m_t is slowly varying in the sense that $m_t \approx \sum_{s=t-5.5,t-5.5+1/365,t+2/365,\ldots,t+5.5} m_s$, an unbiased estimator for m_t is

$$M_t := \frac{1}{11 \cdot 365.25} \sum_{s=t-5.5,t-5.5+1/365,t+2/365,\ldots,t+5.5} U_s.$$

M_t defined in this way is called a **moving average**.

• Plot M_t versus t. Do you observe any fluctuations in M_t or does it seem to be roughly constant? If so, what is this constant?

Once we have the estimate M_t, we can then use the approximation

$$U_t - M_t \approx a\cos(2\pi\theta t) + b\sin(2\pi\omega t)) + Y_t, \qquad \forall\, t \in \mathbf{R},$$

and then try to estimate a, b. A general way to estimate $s_t := a\cos(2\pi\theta t) + b\sin(2\pi\omega t))$ is to use a (smaller) moving average such as

$$S_t := \frac{1}{11} \sum_{s=t-5/365,t-4/365,\ldots,t+5/365} [U_s - M_s].$$

Note that S_t is unbiased.

• Plot S_t versus t. Does it look like a sinusoidal curve? Note that S_t removed the trend from the time series.

Another way to estimate s_t is to estimate the constants a and b directly. By the double angle formula, note that

$$\sum_{s=t,t+1/365,t+2/365,\ldots,t+11} \cos(2\pi\theta s)\sin(2\pi\theta s) = \sum_{s=t,t+1/365,t+2/365,\ldots,t+11} \frac{1}{2}\sin(4\pi\theta s) \approx 0.$$

Also,

$$\frac{1}{365.25} \sum_{s=t,t+1/365,t+2/365,\ldots,t+11} \cos^2(2\pi s/11) \approx \int_0^{11} \cos^2(2\pi x/11)dx \approx 11/2.$$

So, an unbiased estimator for a is

$$A_t := \frac{2}{11 \cdot 365.25} \sum_{s=t,t+1/365,t+2/365,\ldots,t+11} (U_s - M_s)\cos(2\pi\theta s), \qquad \forall\, t \in \mathbf{R}.$$

Similarly, an unbiased estimator for b is

$$B_t := \frac{2}{11 \cdot 365.25} \sum_{s=t,t+1/365,t+2/365,\ldots,t+11} (U_s - M_s)\sin(2\pi\theta s), \qquad \forall\, t \in \mathbf{R}.$$

• Plot A_t versus t. Plot B_t versus t. Are they close to being constant in t?

• Plot $U_t - [M_t + A_t \cos(2\pi t/11) + B_t \sin(2\pi t/11))]$ versus t. This is the time series with the trend and seasonal components removed. Does this plot "resemble" a stationary process?

• Our modeling assumptions used a period of 11 for the seasonal component of the time series. Does the data reflect this assumption? For example, would it be more accurate to have $\theta = \omega = 1/(10.9)$ in our modeling assumption?

Exercise 3.

- Let X, Y be real-valued random variables with $\mathbf{E}\,|X| < \infty$. Let $f\colon \mathbf{R} \to \mathbf{R}$ be a bounded (measurable) function. Using the definition of conditional expectation, show that
$$\mathbf{E}(Xf(Y)\,|\,Y) = f(Y)\mathbf{E}(X|Y).$$
- Let X, Y be independent, real-valued random variables with $\mathbf{E}\,|X| < \infty$. Using the definition of conditional expectation, show that
$$\mathbf{E}(X|Y) = \mathbf{E}X.$$

Exercise 4. Let H be a Hilbert space. Let $g, h \in H$. Prove the Cauchy-Schwarz inequality
$$|\langle g, h\rangle| \le ||g||\,||h||\,.$$
Show also the triangle inequality $||g + h|| \le ||g|| + ||h||$, and the parallelogram law $||g + h||^2 + ||g - h||^2 = 2\,||g||^2 + 2\,||h||^2$.

Exercise 5. Let H be a Hilbert space. Let $h \in H$. Let $h_1, h_2, \ldots \in H$ be a sequence in H such that $\lim_{n\to\infty} ||h_n - h|| = 0$ (i.e. the sequence converges to h). Conclude that the sequence $h_1, h_2, \ldots$ converges weakly to h, in the sense that
$$\lim_{n\to\infty} \langle h_n, g\rangle = \langle h, g\rangle, \qquad \forall\, g \in G.$$
Put another way, the inner product function is continuous. in one argument.

More generally, if $g_1, g_2, \ldots \in H$ satisfies $\lim_{n\to\infty} ||g_n - g|| = 0$, then
$$\lim_{n\to\infty} \langle h_n, g_n\rangle = \langle h, g\rangle.$$
That is, the inner product function is continuous in both arguments.