Exercises – Lecture 1
Stochastic Processes and Markov Chains, Part1

Question 1 *(without* `R`*)*
The transition matrix of a 1st order Markov chain with state space $\mathcal{S} = \{0, 1\}$ is:

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

If the initial probability distribution (at time $t = 0$) is $\boldsymbol{\pi} = (0.8, 0.2)^T$, what is the probability that at time $t = 3$ the state occupied is 0?

Question 2
In a clinical trial patients with a 100% terminal cancer are being treated with a new medicine that aims to increase life expectancy. At the start of each day the medicine is administered. At the moment of administration the state of the patient may change (instantaneous effectiveness!). The state changes of a patient over time can be described by a first order Markov process. At the onset of the clinical trial each patient starts in the same state 'treatment', in which medication is received. With probability α the medicine causes the patient to go into remission (the cancer diminishes). In remission, medication stops and the patient does not risk death. However, while in remission a patient may get a relapse (the cancer returns) with probability β. In relapse the treatment is not re-initiated. Once in relapse, the patient may stay in relapse, or may die (with probability δ). If the treatment is unsuccesful, the state of the patient is unchanged, but the patients has a (daily) risk of death (equal to δ).

a) Give the state space and the transition matrix of this process. In this specify the restrictions on α, β and δ. Also draw the *state diagram* and include the probabilities of each transition.

b) Does this process have an absorbing state? Give the n-step transition matrix for this process when n approaches ∞.

c) Calculate the probability of death of a patient given that s/he has never been in remission. Also, give the probability of death of a patient given that s/he has been in remission.

d) It has been decided that patients who relapse will instantaneously restart with the treatment. Effectively, the state space reduces to three states as 'treatment' and 'relapse' now coincide. It is now possible, with probability β, to return to 'treatment' from 'remission'. Repeat question 2c for this modified Markov process.

Question 3 *(without* `R`*)*
There are two models described below for a signal of length five: i.i.d. (identically and independently distributed), and first order Markov. For each of the sequences `CCGAT` and `CATAT` find the probability of the sequence given the model, for each of the two models (so your answer should consist of four probabilities).

i) i.i.d.: The probabilities of the four nucleotides are $\{p_A = 0.2, p_C = 0.1, p_G = 0.1, p_T = 0.6\}$.

ii) First order Markov: The initial distribution is $\{p_A = 0.2, p_C = 0.1, p_G = 0.1, p_T = 0.6\}$, and the transition matrix (for the nucleotide ordering `A`, `C`, `G`, `T`) is:

$$\begin{pmatrix} 0.10 & 0.80 & 0.05 & 0.05 \\ 0.35 & 0.10 & 0.10 & 0.45 \\ 0.30 & 0.20 & 0.20 & 0.30 \\ 0.60 & 0.10 & 0.25 & 0.05 \end{pmatrix}.$$

Question 4

The `sample` function can be used to generate random sequences in R. For example, the syntax:

```
> # specify state space
> nucleotides <- c("A", "C", "G", "T")
> # probability vector of each nucleotide (in same order)
> p0 <- c(0.2, 0.3, 0.3, 0.2)
> # sample 10 nucleotides from the state space with probabilities above
> sample(nucleotides, 10, replace=TRUE, prob=p0)
[1] "A" "C" "A" "T" "A" "G" "C" "G" "C" "A"
```

generates a DNA sequence from state space $\mathcal{S}$ = `nucleotides` (as defined in the syntax above) with length=10 and multinomial probabilities equal to `p0` (as defined above). This mimicks the i.i.d. model (a 0^{th} order Markov chain) of Exercise 3.

This can also be done using a `for`-loop:

```
> DNAseq <- sample(nucleotides, 1, replace=TRUE, prob=p0)
> for (i in 2:10){
>    DNAseq <- c(DNAseq, sample(nucleotides, 1, replace=TRUE, prob=p0))
> }
```

Although this is less efficient.

a) *(without* `R`*)* Write down the transition matrix of Markov chain from which the DNA sequence above has been generated.

b) *(with* `R`*)* The `GC` content of a DNA sequence is defined as the percentage of `C`'s and `G`'s on the total number of bases of the sequence: (#`C`+#`G`)/(#`A`+#`C`+#`G`+#`T`)∗100%. Calculate the `GC` content for an infinitely long DNA sequence generated in accordance with the sampling model above. Confirm this by simulation: generate a DNA sequence of (say) length 100000 and calculate its `GC` content.

Question 5 *(without* `R`*)*

Assume a first order Markov model with state space $\mathcal{S} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, initial distribution $\boldsymbol{\pi}$ and transition matrix $\mathbf{P}$. Let $\ldots, X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, \ldots$ be the sequence generated by this model. Furthermore, write $P(X_t = \mathtt{A}) = \varphi_{\mathtt{A}}$, $P(X_t = \mathtt{C}) = \varphi_{\mathtt{C}}$, $P(X_t = \mathtt{G}) = \varphi_{\mathtt{G}}$, and $P(X_t = \mathtt{T}) = \varphi_{\mathtt{T}}$ for all t.

a) Write $P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T})$ in terms of the elements of $\mathbf{P}$ and the φ's.

b) Write $P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-1} = \mathtt{T})$ in terms of the elements of $\mathbf{P}$ and the φ's.

c) Write $P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T} \,|\, X_{t-1} = \mathtt{T})$ in terms of the elements of $\mathbf{P}$ and the φ's.

d) Write $P(X_{t+1} = \mathtt{A}, X_{t-1} = \mathtt{T} \,|\, X_t = \mathtt{T})$ in terms of the elements of $\mathbf{P}$ and the φ's.

e) Write $P(X_{t+1} = \mathtt{A}, X_{t-1} = \mathtt{T})$ in terms of the elements of $\mathbf{P}$ and the φ's.

f) Write $P(X_{t+1} = \mathtt{A}, X_{t-2} = \mathtt{C} \,|\, X_t = \mathtt{T})$ in terms of the elements of $\mathbf{P}$ and the φ's.

Question 6 *(with* `R`*)*
Consider the transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.1500 & 0.3500 & 0.3500 & 0.1500 \\ 0.1660 & 0.3340 & 0.3340 & 0.1660 \\ 0.1875 & 0.3125 & 0.3125 & 0.1875 \\ 0.2000 & 0.3000 & 0.3000 & 0.2000 \end{pmatrix}.$$

Now use `sample()` to generate DNA sequences according to a first order Markov chain. Hereto use the transition matrix given above. We can only simulate one symbol at a time, because we need to keep track of the current state.

```
> DNAseq <- sample(nucleotides, 1, replace=TRUE, prob=p0)
> DNAseq <- c(DNAseq, sample(nucleotides, 1, replace=TRUE,
                  P[nucleotides==DNAseq[1], ]))
> DNAseq
[1] "C" "A"
```

The `P[nucleotides==DNAseq[1], ]` statement returns the row in the transition matrix that corresponds to the matching character in the alphabet `nucleotides`. Try `P[nucleotides=="A", ]`, `P[nucleotides=="C", ]`, `P[nucleotides=="G",]` and `P[nucleotides=="T",]` to see what happens.

Use the transition matrix given in Question 4 and simulate a DNA sequence of (say) length 100000. Hint: use a `for`-loop. Calculate the `GC` content for this sequence and compare it to the answer of Question 4b.

Question 7

a) (with `R`*)* Given the following sequence (available at the lecture website):

```
CGGAACAACCACTTGACCTTGGTTGTGTATGTCCCGGACTCGGGGGTCCATAATCCTGTTGATCGTGTTTCCCTA
GTTCAACGGTTTACGTGGGTGTGTTCTAATCTCACCCCGTGAACTCTCGTAAAATATGTTGTCCCACTCTTGTCT
CCTACGATCGGTCGTGTTCCCGTCCTCGTTCGTTCATACGTGATCACGGTATGTGGGATTCGTCGTTACCCCCGG
TGACGGATGGATGTCAATGTGTCGATTTCGTGGTCAATGTGGGGGTGTCGTGGGGTGGTATCTTCGGTGGTTCCG
ACGGGGGGTCCCTCTCGTGGGGTCTATATGGTTCTACGTCTACCCTCAATGAACGTCGTTGTTCACCCATGATTT
CGGATCAAATGGTGGTTATGTTTCGTGAATCGACTGAACCCCGGGGTGGTGGTCATCAAATCTTGTCGAAACCGG
TGGGTGTTGAACAATAATGACACTCCGTGATCTTGTTCATTAACGTGACCGAACGTCCGGTCGAATGTCGGTCCT
CAAATCGGAATTGTTAAACTGGTCAATCGACACCGTTCGTTCCTGTGTGGTATTATGTACCTCCCTTCTCCGTGT
CGTACTCGTCGTTTCGTGTCATATGGATACTCAACGGGGGTTCCACGTCCCCTAATGAATGTCGGTGATGTCCCC
CCGTATTCCTCACCGGTGTAACCCACGAAATTGTGGGGACGGTACGTCAACCTTGTCGATACGAAACGTTCGTGT
ATGTGATGTGTGGGTGTTTCCAATAACGAATACACGTCCTGTGGGATTCCCTATAACGTGGTCCCGGTCGGGGGT
GAAACCACGGTCCTACCGGTTCATACGTGTGTCGTAACCGTGGGTTAAATGTCCGGATCTCGGAACAACCACTTG
ACCTTGGTTGTGTATGTCCCGGACTCGGGGGTCCATAATCCTGTTGATCGTGTTTCCCTAATCTGTCCACCCGTA
TCAAATTTGTCCAATTTCGGGGATG
```

Maximize the likelihood with respect to the parameters, give the estimated transition matrix, and draw the state diagram.

To read in a string in `R`:
`> DNAseq <- c("AAGTCAGT")`
To select a letter, say the 5th, from this string:
`> substr(DNAseq, 5, 5)`
To obtain the number of characters of a string:

`> nchar(DNAseq)`

b) *(in principle without* `R`*, although a calculator is practical)* Using the estimated transition matrix in part *a)*, calculate the likelihood of the following motifs: `AAAAAA`, `CTGCAG` and `ACCGGT`. You may wish to assume that the first nucleotide is given, hence $P(Y_0 = y_0) = 1$.

c) *(in principle without* `R`*, although a calculator is practical)* For the sequence given in part *a)*, test using the χ^2-test whether the independence model could have been assumed.

d)* *(with and without* `R`*)* Repeat part *a)*, but now assume that the transition matrix is symmetric. Hint: use the derivation of the likelihood and the corresponding estimators presented in Lecture 1, and simplify as much as possible using the symmetry assumption.

Question 8 *(in principle without* `R`*, although a calculator is practical)*
Regions on the genome that have many more `CG` dimers (and in fact more C and G nucleotides) than elsewhere on the DNA are called CpG islands. From a set of human DNA sequences we have learned that the nucleotide sequence of a CpG island and of remainder of the DNA are modeled by first order Markov models with the transition matrices:

$$\mathbf{P}_1 = \begin{pmatrix} 0.180 & 0.274 & 0.426 & 0.120 \\ 0.171 & 0.368 & 0.274 & 0.188 \\ 0.161 & 0.339 & 0.375 & 0.125 \\ 0.079 & 0.355 & 0.384 & 0.182 \end{pmatrix} \quad \text{and} \quad \mathbf{P}_2 = \begin{pmatrix} 0.300 & 0.205 & 0.285 & 0.210 \\ 0.322 & 0.298 & 0.078 & 0.302 \\ 0.248 & 0.246 & 0.298 & 0.208 \\ 0.177 & 0.239 & 0.292 & 0.292 \end{pmatrix},$$

respectively.

Given the following stretches of genomic sequence, decide whether they come from a CpG island or not. Hint: calculate the likelihood of a stretch under both first order Markov models. You may wish to assume that the first nucleotide is given, hence $P(Y_0 = y_0) = 1$.

- Stretch 1: `GGTGGTCATCAAATCTTGTCGA`
- Stretch 2: `ACGTTTAATATGGTACCAATGT`
- Stretch 3: `ACGGGGGGTCCCTCTCGTGGGG`