Answers – Lecture 1
Stochastic Processes and Markov Chains, Part1

Question 1
The exercise does not specify anything about the states visited prior to X_3. Hence, in the calculation of the probability $P(X_3 = \mathtt{0})$ we need to account for the uncertainty at X_0, X_1 and X_2. Note that the third order transition matrix $\mathbf{P}^{(3)}$ is equal to $\mathbf{P}^3$ and

$$\mathbf{P}^3 = \begin{pmatrix} 0.583 & 0.417 \\ 0.556 & 0.444 \end{pmatrix}.$$

This gives the transition probability over three time steps, e.g. $(\mathbf{P})_{2,1} = P(X_3 = \mathtt{0} | X_0 = \mathtt{1})$. This probability accommodates the stochasticity of X_1 and X_2. It is left to take into account the start position:

$$\begin{aligned} P(X_3 = \mathtt{0}) &= P(X_0 = \mathtt{0})P(X_3 = \mathtt{0}|X_0 = \mathtt{0}) + P(X_0 = \mathtt{1})P(X_3 = \mathtt{0}|X_0 = \mathtt{1}) \\ &= 0.8 \times 0.583 + 0.2 \times 0.556. \end{aligned}$$

Question 2
Question 2a)
The state space is given by ‘treatment’, ‘remission’, ‘relapse’, ‘death’; the initial distribution by $\boldsymbol{\pi} = (1, 0, 0, 0)^T$, and the transition matrix by:

$$\mathbf{P} = \begin{pmatrix} 1-\alpha-\delta & \alpha & 0 & \delta \\ 0 & 1-\beta & \beta & 0 \\ 0 & 0 & 1-\delta & \delta \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where $\alpha \geq 0$, $\beta \geq 0$, $\delta \geq 0$ and $\alpha + \delta \leq 1$. Also give the state diagram.

Question 2b)

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Question 2c)
Note that: $P(X_t = \text{treatment}) = \sum_{s=1}^{t}(1-\alpha-\delta)^{s-1}$. Then, the probability of death of a patient who has never been in remission equals: $\delta \sum_{s=1}^{\infty}(1-\alpha-\delta)^{s-1}$. The only other route to death is

via remission. The probability of death via this route is simply: $1 - \delta \sum_{s=1}^{\infty}(1-\alpha-\delta)^{s-1}$.

Question 2d)
The answer is exactly as in 2c (as the first answer of 2c does not change).

Question 3
IID: $P(\texttt{CCGAT}) = P(\texttt{C})P(\texttt{C})P(\texttt{G})P(\texttt{A})P(\texttt{T}) = 0.1 \times 0.1 \times 0.1 \times 0.2 \times 0.6 = 0.00012$. Simarly, $P(\texttt{CCGAT}) = 0.00144$.

Markov: $P(\texttt{CCGAT}) = P(\texttt{C})P(\texttt{C}|\texttt{C})P(\texttt{G}|\texttt{C})P(\texttt{A}|\texttt{G})P(\texttt{T}|\texttt{A}) = 0.1 \times 0.1 \times 0.1 \times 0.3 \times 0.05 = 1.5 * 10^{-5}$. Simarly, $P(\texttt{CCGAT}) = 5.25 * 10^{-5}$.

Question 4
Question 4a

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.2 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.2 & 0.3 & 0.3 & 0.2 \end{pmatrix}$$

Question 4b
$[P(C)+P(G)]/[P(A)+P(C)+P(G)+P(T)] = (0.3+0.3)/1 = 0.6$.

Question 5
Question 5a
Multiply by $1 = P(X_t = \texttt{T})/P(X_t = \texttt{T})$ and apply the definition of conditional probability

$$\begin{aligned} P(X_{t+1} = \texttt{A}, X_t = \texttt{T}) &= P(X_{t+1} = \texttt{A}, X_t = \texttt{T})\frac{P(X_t = \texttt{T})}{P(X_t = \texttt{T})} \\ &= \frac{P(X_{t+1} = \texttt{A}, X_t = \texttt{T})}{P(X_t = \texttt{T})}P(X_t = \texttt{T}) \\ &= P(X_{t+1} = \texttt{A}|X_t = \texttt{T})P(X_t = \texttt{T}) \\ &= p_{\texttt{TA}} \times \varphi_{\texttt{T}} \end{aligned}$$

Question 5b

Multiply by '1′ and apply the definition of conditional probability, repetitively:

$$\begin{aligned}
P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-1} = \mathtt{T}) &= \frac{P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-1} = \mathtt{T})}{P(X_t = \mathtt{T}, X_{t-1} = \mathtt{T})} P(X_t = \mathtt{T}, X_{t-1} = \mathtt{T}) \\
&= P(X_{t+1} = \mathtt{A} \,|\, X_t = \mathtt{T}, X_{t-1} = \mathtt{T}) P(X_t = \mathtt{T}, X_{t-1} = \mathtt{T}) \\
&= P(X_{t+1} = \mathtt{A} \,|\, X_t = \mathtt{T}) P(X_t = \mathtt{T}, X_{t-1} = \mathtt{T}) \\
&= P(X_{t+1} = \mathtt{A} \,|\, X_t = \mathtt{T}) \frac{P(X_t = \mathtt{T})}{P(X_{t-1} = \mathtt{T})} P(X_{t-1} = \mathtt{T}) \\
&= P(X_{t+1} = \mathtt{A} | X_t = \mathtt{T}) P(X_t = \mathtt{T} | X_{t-1} = \mathtt{T}) P(X_{t-1} = \mathtt{T})
\end{aligned}$$

Question 5c

$$\begin{aligned}
P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T} \,|\, X_{t-1} = \mathtt{T}) &= P(X_{t+1} = \mathtt{A} \,|\, X_t = \mathtt{T}, X_{t-1} = \mathtt{T}) P(X_t = \mathtt{T} \,|\, X_{t-1} = \mathtt{T}) \\
&= P(X_{t+1} = \mathtt{A} \,|\, X_t = \mathtt{T}) P(X_t = \mathtt{T} \,|\, X_{t-1} = \mathtt{T})
\end{aligned}$$

Or, use the definition of conditional probability to note that:

$$P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T} \,|\, X_{t-1} = \mathtt{T}) = \frac{P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-1} = \mathtt{T})}{P(X_{t-1} = \mathtt{T})}$$

and substitute the answer to Exercise 5b.

Question 5d
Using the definition of conditional probability to note that:

$$P(X_{t+1} = \mathtt{A}, X_{t-1} = \mathtt{T} \,|\, X_t = \mathtt{T}) = P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-1} = \mathtt{T}) / P(X_t = \mathtt{T})$$

Now as before.

Question 5e
Using the total probability law we rewrite:

$$P(X_{t+1} = \mathtt{A}, X_{t-1} = \mathtt{T}) = \sum_{x_t} P(X_{t+1} = \mathtt{A}, X_t = x_t, X_{t-1} = \mathtt{T})$$

The term inside the sum is similar to the probability calculated in Question 5b). Thus:

$$P(X_{t+1} = \mathtt{A}, X_{t-1} = \mathtt{T}) = \sum_{x_t} P(X_{t+1} = \mathtt{A} | X_t = x_t) P(X_t = x_t | X_{t-1} = \mathtt{T}) P(X_{t-1} = \mathtt{T})$$

Question 5f
Subsequently use the definition of conditional probability followed by the total probability law to write:

$$\begin{aligned}
P(X_{t+1} = \mathtt{A}, X_{t-2} = \mathtt{C} \,|\, X_t = \mathtt{T}) &= P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-2} = \mathtt{T}) / P(X_t = \mathtt{T}) \\
&= \sum_{x_{t-1}} P(X_{t+1} = \mathtt{A}, X_t = \mathtt{T}, X_{t-1} = x_{t-1}, X_{t-2} = \mathtt{T}) / P(X_t = \mathtt{T})
\end{aligned}$$

Now as before.

Question 6

```
> nucleotides <- c("A", "C", "G", "T")
> p0 <- c(0.2, 0.3, 0.3, 0.2)
> P <- matrix(c(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), ncol=4, byrow=TRUE)
> DNAseq <- sample(nucleotides, 1, replace=TRUE, prob=p0)
> for (j in 1:1000000){
+ DNAseq <- c(DNAseq, sample(nucleotides, 1, replace=TRUE,
+ P[nucleotides==DNAseq[1], ]))
+ }
> table(DNAseq) / length(DNAseq)
```

Question 7

Question 7a

```
> slh <- read.table("sequence.txt")
> slh <- levels(slh[1,1])[slh[1,1]]
> DNAseq <- character()
> for (j in 1:nchar(slh)){
>    DNAseq[j] <- substr(slh, j, j)
> }
> table(DNAseq)
> P <- table(DNAseq[-1000], DNAseq[-1])
> P <- P / rowSums(P)
```

Question 7b

```
> P[1,1] * P[1,1] * P[1,1] * P[1,1] * P[1,1]
> P[2,4] * P[4,3] * P[3,2] * P[2,1] * P[1,3]
> P[1,2] * P[2,2] * P[2,3] * P[3,3] * P[3,4]
```

Question 7c

```
> nuclFreq <- matrix(table(DNAseq), ncol=1)
> dimerFreq <- table(DNAseq[1:(length(DNAseq)-1)], DNAseq[2:length(DNAseq)])
> dimerExp <- nuclFreq %*% t(nuclFreq)/(length(DNAseq)-1)
> teststat <- sum((dimerFreq - dimerExp) ^ 2 / dimerExp)
> pval <- exp(pchisq(teststat, 9, lower.tail=FALSE, log.p=TRUE))
```

Question 7d

To be inserted.

Question 8

$$\begin{aligned}
P(\text{stretch } 1|\mathbf{P}_1) = \exp(-31.9551) &\quad\leq\quad \exp(-29.4885) = P(\text{stretch } 1|\mathbf{P}_2) \\
P(\text{stretch } 2|\mathbf{P}_1) = \exp(-36.5958) &\quad\leq\quad \exp(-31.1944) = P(\text{stretch } 2|\mathbf{P}_2) \\
P(\text{stretch } 3|\mathbf{P}_1) = \exp(-25.2954) &\quad\geq\quad \exp(-29.8534) = P(\text{stretch } 3|\mathbf{P}_2)
\end{aligned}$$