Answers – Lecture 4
Hidden Markov models

Question 1
Question 1a)
The underlying sequence comprises three positions. As the state space of the underlying Markov process has three states, the possible number of underlying sequences amounts to $3^3 = 27$. However, the initial distribution of the Markov process rules out the possibility of starting in S_2 and S_3. This leaves 3^2 sequences. Starting from S_1 the transition matrix specifies the possible states of the second position (e.g. exclusing S_1). Et cetera. Eventually, two sequences are feasible (S_1, S_2, S_1) and (S_1, S_3, S_2). Now using the emission matrix verify whether both can produce the observed sequence. Only the latter sequence can. Hence, the state sequence that leads to observed series: (S_1, S_3, S_2). Further $P((S_1, S_3, S_2)) = \frac{1}{2}$ and $P((a, b, c)|(S_1, S_3, S_2)) = (\frac{1}{2})^3 = \frac{1}{8}$. The probability $P((a, b, c))$ is now: $P((a, b, c)|(S_1, S_3, S_2))P((S_1, S_3, S_2)) = \frac{1}{8}\frac{1}{2} = \frac{1}{16}$.

Question 1b)
Possible underlying sequences: (S_1, S_2, S_1) en (S_1, S_3, S_2), both have probability of $\frac{1}{2}$. Further: $P((a, c, a)|(S_1, S_2, S_1)) = \frac{1}{8} = P((a, c, a)|(S_1, S_3, S_2))$. Hence, $P((a, c, a) = \frac{1}{8}$.

Question 2
Question 2a)

$$\begin{aligned} & P(Y_{t-1} = 0, Y_{t+1} = 1 \,|\, X_t = \text{exon}) \\ & \quad = \quad P(Y_{t-1} = 0 \,|\, X_t = \text{exon})\, P(Y_{t+1} = 1 \,|\, X_t = \text{exon}) \end{aligned}$$

using the total probability law:

$$\begin{aligned} = \quad & \left\{ \sum_{x_{t-1} \in \{\mathtt{I},\mathtt{E}\}} P(Y_{t-1} = 0, X_{t-1} = x_{t-1} \,|\, X_t = \text{exon}) \right\} \\ & \quad \times \left\{ \sum_{x_{t-1} \in \{\mathtt{I},\mathtt{E}\}} P(Y_{t+1} = 1, X_{t+1} = x_{t+1} \,|\, X_t = \text{exon}) \right\} \end{aligned}$$

using $P(A, B \,|\, C) = P(A, B, C)/P(C) = (P(A, B, C)/P(B, C)) \times (P(B, C)/P(C)) = P(A \,|\, B, C) \times$

$P(B \,|\, C)$ (that is, using the definition of conditional probability repetitively):

$$= \left\{ \sum_{x_{t+1} \in \{\mathtt{I},\mathtt{E}\}} P(Y_{t-1} = 0 \,|\, X_{t-1} = x_{y-1})\, P(X_{t-1} = x_{t-1} \,|\, X_t = \text{exon}) \right\}$$
$$\times \left\{ \sum_{x_{t-1} \in \{\mathtt{I},\mathtt{E}\}} P(Y_{t+1} = 1 \,|\, X_{t+1} = x_{t+1})\, P(X_{t+1} = x_{t+1} \,|\, X_t = \text{exon}) \right\}$$

using the fact that an exon cannot emit a 1 and the reversibility of the Markov chain

$$\begin{aligned}
&= \sum_{x_{t-1} \in \{\mathtt{I},\mathtt{E}\}} P(Y_{t-1} = 0 \,|\, X_{t-1} = x_{y-1})\, P(X_{t-1} = x_{t-1} \,|\, X_t = \text{exon}) \\
&\qquad \times P(Y_{t+1} = 1 \,|\, X_{t+1} = \text{intron})\, P(X_{t+1} = \text{intron} \,|\, X_t = \text{exon}) \\
&= \sum_{x_{t-1} \in \{\mathtt{I},\mathtt{E}\}} P(Y_{t-1} = 0 \,|\, X_{t-1} = x_{t-1})\, P(X_t = \text{exon} \,|\, X_{t-1} = x_{t-1}) \\
&\qquad \times P(Y_{t+1} = 1 \,|\, X_{t+1} = \text{intron})\, P(X_{t+1} = \text{intron} \,|\, X_t = \text{exon}) \\
&= \sum_{x_{t-1} \in \{\mathtt{I},\mathtt{E}\}} P(Y_{t-1} = 0 \,|\, X_{t-1} = x_{t-1})\ P(X_t = \text{exon} \,|\, X_{t-1} = x_{t-1})\, \frac{1}{2}\,\alpha \\
&= \frac{1}{4}\alpha^2 + \frac{1}{2}\alpha(1-\alpha).
\end{aligned}$$

Question 2b)

The only possible intron-exon sequences that may yield 010 are: `IIE` en `III`. Then:

$$\begin{aligned}
P(010 \,|\, \mathtt{IIE}) &= \frac{1}{4} \\
P(010 \,|\, \mathtt{III}) &= \frac{1}{8} \\
P(\mathtt{III}) &= (1-\alpha)^2 \\
P(\mathtt{IIE}) &= \alpha(1-\alpha) \\
P(010 \,|\, \mathtt{IIE})\, P(\mathtt{IIE}) &= \frac{1}{4}\alpha(1-\alpha) \\
P(010 \,|\, \mathtt{III})\, P(\mathtt{III}) &= \frac{1}{8}(1-\alpha)^2 \\
P(010) &= \frac{1}{4}\alpha(1-\alpha) + \frac{1}{8}(1-\alpha)^2
\end{aligned}$$

And, thus ($\alpha = 1/4$):

$$\begin{aligned}
P(\mathtt{IIE} \,|\, 010) &= \frac{2}{5} \\
P(\mathtt{III} \,|\, 010) &= \frac{3}{5}
\end{aligned}$$

The required sequence is thus: `III`.

Question 3

Question 3a

Define the state space for the latent Markov chain $S = \{I, II\}$ and the emission alphabet $\{\texttt{A}, \texttt{A}^2, \texttt{A}^3, \texttt{C}, \texttt{C}^2, \texttt{C}^3\}$. It remains to specify the transition matrix and the emission matrix:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

Question 3b

Many parametrizations are possible (in fact, a HMM is not even necessary). Hence, here only a possible one is given. Define the state space for the latent Markov chain $S = \{I, II, III\}$ and the emission alphabet $\{\texttt{AC}^3, \texttt{A}^2\texttt{C}^2, \texttt{A}^3\texttt{C}\}$. It remains to specify the transition matrix and the emission matrix:

$$\mathbf{P} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Question 4

Question 4a

Define the state space for the latent Markov chain $S = \{\neg\texttt{CpG}, \texttt{CpG}\}$ and the emission alphabet $\{hypo, normal, hyper\}$. It remains to specify the transition matrix and the emission matrix. Given is the stationary distribution of the hidden Markov chain: $\varphi_{CpG} = 0.10$ and $\varphi_{\neg CpG} = 0.90$. Furthermore, we know that the first row in the transition matrix of this Markov chain is given by $(0.95, 0.05)$. Also we now that the stationary distribution satisfies $\boldsymbol{\varphi}^T\mathbf{P} = \boldsymbol{\varphi}^T$. Hence, $\varphi_1\, p_{11} + \varphi_2\, p_{21} = \varphi_1$. Or, $0.90 \times 0.95 + 0.10 \times p_{21} = 0.90$. Ergo, $p_{21} = 0.45$. The transition and emission matrix are thus:

$$\mathbf{P} = \begin{pmatrix} 0.95 & 0.05 \\ 0.45 & 0.55 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 0 & 2/3 & 1/3 \end{pmatrix}.$$

Question 4b

$$\begin{aligned} P(X_j = \texttt{CpG} \,|\, Y_j = \text{normal}) &= P(X_j = \texttt{CpG}, Y_j = \text{normal}) \,/\, P(Y_j = \text{normal}) \\ &= P(Y_j = \text{normal} \,|\, X_j = \texttt{CpG}) \frac{P(X_j = \texttt{CpG})}{P(Y_j = \text{normal})} \\ &= \frac{2}{3}\frac{1}{10} \,/\, P(Y_j = \text{normal}) \end{aligned}$$

It remains to determine $P(Y_j = \text{normal})$. Hereto observe:

$$\begin{aligned} P(Y_j = \text{normal}) &= P(Y_j = \text{normal} \,|\, X_j = \texttt{CpG})\, P(X_j = \texttt{CpG}) \\ &\quad + P(Y_j = \text{normal} \,|\, X_j = \neg\texttt{CpG})\, P(X_j = \neg\texttt{CpG}) \\ &= \frac{2}{3} \times 0.10 + \frac{2}{3} \times 0.90 \;=\; \frac{2}{3}. \end{aligned}$$

Combining the above yields the desired probability: 1/10.

Question 4c Using the definition of conditional probability:

$$\begin{aligned} & P(X_j = \texttt{CpG} \,|\, Y_j = \text{normal}, Y_{j-1} = \text{hyper}) \\ = \; & P(X_j = \texttt{CpG}, Y_j = \text{normal}, Y_{j-1} = \text{hyper}) \,/\, P(Y_j = \text{normal}, Y_{j-1} = \text{hyper}) \end{aligned}$$

noting that only a CpG island can emit a 'hyper':

$$= \; P(X_j = \texttt{CpG}, X_{j-1} = \texttt{CpG}, Y_j = \text{normal}, Y_{j-1} = \text{hyper}) \,/\, P(Y_j = \text{normal}, Y_{j-1} = \text{hyper})$$

using the definition of conditional probability again:

$$\begin{aligned} = \; & P(Y_j = \text{normal}, Y_{j-1} = \text{hyper} \,|\, X_j = \texttt{CpG}, X_{j-1} = \texttt{CpG}) \\ & \times P(X_j = \texttt{CpG}, X_{j-1} = \texttt{CpG}) / P(Y_j = \text{normal}, Y_{j-1} = \text{hyper}) \end{aligned}$$

using the independence of elements of the observed sequence given the underlying sequence:

$$\begin{aligned} = \; & P(Y_j = \text{normal} \,|\, X_j = \texttt{CpG}) P(Y_{j-1} = \text{hyper} \,|\, X_{j-1} = \texttt{CpG}) \\ & \times P(X_j = \texttt{CpG} \,|\, X_{j-1} = \texttt{CpG}) P(X_{j-1} = \texttt{CpG})) / P(Y_j = \text{normal}, Y_{j-1} = \text{hyper}) \\ = \; & \frac{2}{3}\frac{1}{3} \times 0.55 \times \frac{1}{10} / P(Y_j = \text{normal}, Y_{j-1} = \text{hyper}). \end{aligned}$$

It now remains to calculate $P(Y_j = \text{normal}, Y_{j-1} = \text{hyper})$.

$$\begin{aligned} & P(Y_j = \text{normal}, Y_{j-1} = \text{hyper}) \\ = \; & P(Y_j = \text{normal}, Y_{j-1} = \text{hyper} \,|\, X_j = \texttt{CpG}, X_{j-1} = \texttt{CpG}) P(X_j = \texttt{CpG}, X_{j-1} = \texttt{CpG}) \\ & + P(Y_j = \text{normal}, Y_{j-1} = \text{hyper} \,|\, X_j = \neg\texttt{CpG}, X_{j-1} = \texttt{CpG}) P(X_j = \neg\texttt{CpG}, X_{j-1} = \texttt{CpG}). \end{aligned}$$

Remaining probabilities have been calculated above.

Question 5

Question 5a

Generate a DNA sequence of 1000 nucleotides. Save the sequences of states and nucleotides. Report the R-code and the nucleotide sequence. Hint: use the `sample` function and `for`-loop construction.

```
> iNeXtrons <- c("I", "E")
> nucleotides <- c("A", "C", "G", "T")
> p0 <- c(0.5, 0.5)
> a <- matrix(c(0.9, 0.1, 0.1, 0.9), ncol=2)
> b <- matrix(c(0.49, 0.01, 0.49, 0.01, 0.01, 0.49, 0.01, 0.49), ncol=4, byrow=TRUE)
> iNeXtronSeq <- sample(iNeXtrons, 1, replace=TRUE, prob=p0)
> if (iNeXtronSeq[1] == "I"){
+ nuclSeq <- sample(nucleotides, 1, prob=b[1,])
+ } else {
+ nuclSeq <- sample(nucleotides, 1, prob=b[2,])
+ }
> for (i in 2:10000){
+ iNeXtronSeq <- c(iNeXtronSeq, sample(iNeXtrons, 1,
+ prob=a[iNeXtrons==iNeXtronSeq[i-1], ]))
```

```
+ if (iNeXtronSeq[i] == "I"){
+ nuclSeq <- c(nuclSeq, sample(nucleotides, 1, prob=b[1,]))
+ } else {
+ nuclSeq <- c(nuclSeq, sample(nucleotides, 1, prob=b[2,]))
+ }
+ }
```

Question 5b

```
> table(nuclSeq, iNeXtronSeq)
```

Yes, nucleotide distributions differ considerably between introns and exons.

Question 5c

No, nucleotide distributions (the observed information) are identical for introns and exons.

Question 5d

$$\mathbf{P} \quad = \quad \begin{pmatrix} 0.35 & 0.15 & 0.15 & 0.35 \\ 0.35 & 0.15 & 0.15 & 0.35 \\ 0.35 & 0.15 & 0.15 & 0.35 \\ 0.35 & 0.15 & 0.15 & 0.35 \end{pmatrix}.$$

Question 6

Question 6a

Completely analogous to the example detailed in the lecture notes.

Question 6b

Completely analogous to the example detailed in the lecture notes.