EE5137 : Stochastic Processes (Spring 2021)
Some Additional Notes on Markov Chains and Classification of States

Vincent Y. F. Tan

February 27, 2021

In this document, we provide some supplementary material to Lecture 7. You need to know Sections 1, 2 and 3 here. Section 4 is interesting but completely optional reading.

1 Classification of States

Recall that for a finite-state Markov chain, two distinct states i and j communicate (abbreviated $i \leftrightarrow j$) if i is accessible from j and j is accessible from i.

Proposition 1. *Communication is an equivalence relation. That is*

1. $i \leftrightarrow i$;
2. *if* $i \leftrightarrow j$, *then* $j \leftrightarrow i$;
3. *if* $i \leftrightarrow j$ *and* $j \leftrightarrow k$, *then* $i \leftrightarrow k$;

Proof. The first two parts follow directly from the definition. For part 3, suppose that $i \leftrightarrow j$ and $j \leftrightarrow k$; then there exists m and n such that $P_{ij}^m > 0$ and $P_{jk}^n > 0$. Hence,

$$P_{ik}^{m+n} = \sum_{r=0}^{\infty} P_{ir}^m P_{rk}^n \geq P_{ij}^m P_{jk}^n > 0. \tag{1}$$

Similarly, we can show that there exists an s such that $P_{ki}^s > 0$. □

For any states i and j define f_{ij}^n to be the probability that, starting in i, the first transition into j occurs at time n. Formally,

$$f_{ij}^0 = 0, \quad f_{ij}^n = \Pr(X_n = j, X_k \neq j, k = 1, \ldots, n-1 | X_0 = i). \tag{2}$$

Let

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^n. \tag{3}$$

Then f_{ij} denotes the probability of ever making a transition into state j give that the process starts in i. Note that for $i \neq j$, $f_{ij} > 0$ if and only if j is accessible from i. State j is said to be *recurrent* if $f_{jj} = 1$, and *transient* otherwise. These definitions are consistent with that in the book.

Proposition 2. *State j is recurrent if, and only if,*

$$\sum_{n=1}^{\infty} P_{jj}^n = \infty. \tag{4}$$

Proof. State j is recurrent if, with probability 1, a process starting at state j with eventually return. However, by the Markovian property, it follows that the process probabilistically restarts itself upon returning to state j. Hence, with probability 1, it will return again to j. Repeating this argument, we see that, with probability 1, the number of visits to state j will be infinite and tus will have infinite expectation. On the other hand, suppose j is transient. Then each time the process returns to j there is a positive probability of $1 - f_{jj}$ that it will never again return; hence, the number of visits is geometric with finite mean $1/(1 - f_{jj})$.

By the above argument, we see that state j is recurrent if and only if

$$\mathbb{E}[\text{number of visits to } j | X_0 = j] = \infty \tag{5}$$

But letting

$$I_n = \begin{cases} 1 & \text{if } X_n = j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

it follows that $\sum_{n=0}^{\infty} I_n$ denotes the number of visits to j. Since

$$\mathbb{E}\left[\sum_{n=0}^{\infty} I_n \,\middle|\, X_0 = j\right] = \sum_{n=0}^{\infty} \mathbb{E}\left[I_n \,|\, X_0 = j\right] = \sum_{n=0}^{\infty} P_{jj}^n \tag{7}$$

the result follows. □

Corollary 3. *If i is recurrent and $i \leftrightarrow j$, then j is recurrent.*

Proof. Let m and n be such that $P_{ij}^n > 0$ and $P_{ji}^m > 0$. Now for any $s \geq 0$,

$$P_{jj}^{m+n+s} \geq P_{ji}^m P_{ii}^s P_{ij}^n \tag{8}$$

and thus,

$$\sum_s P_{jj}^{m+n+s} \geq P_{ji}^m P_{ij}^n \sum_s P_{ii}^s = \infty \tag{9}$$

and the result follows from Proposition 2. □

2 The Simple Random Walk

The Markov chain whose state space is the set of all integers and has transition probabilities

$$P_{i,i+1} = p = 1 - P_{i,i-1}, \quad i \in \mathbb{Z} \tag{10}$$

where $p \in (0, 1)$, is called the simple random walk. One interpretation of this process is that it represents the wanderings of a drunken man as he walks along a straight line. Another is that it represents the winnings of a gambler who on each play of the game either wins or loses one dollar.

Since all states communicate with one, it follows from Corollary 3 that the states are either all recurrent or all transient. Let's just focus on state 0 and attempt to determine whether $\sum_n P_{00}^n$ is finite or infinite.

Since it is impossible to be even (win 0 dollars) after an odd number of steps,

$$P_{00}^{2n+1} = 0, \quad n \in \mathbb{N}. \tag{11}$$

On the other hand, the gambler would be even after $2n$ trials if and only if he won n of those trials and lost n. This probability is

$$P_{00}^{2n} = \binom{2n}{n} p^n (1-p)^n = \frac{(2n)!}{(n!)^2} (p(1-p))^n, \quad n \in \mathbb{N}. \tag{12}$$

By using the Stirling approximation,[1]

$$n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi}, \tag{13}$$

[1] We write $a_n \sim b_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$

we obtain

$$P_{00}^{2n} \sim \frac{(4p(1-p))^n}{\sqrt{\pi n}}. \tag{14}$$

Hence, $\sum_n P_{00}^n < \infty$ if and only if $p \neq 1/2$ (note that $\sum_n \frac{1}{\sqrt{n}} = \infty$). Thus the chain is recurrent if $p = 1/2$ and transient if $p \neq 1/2$.

When $p = 1/2$, the above process is called a *one-dimensional symmetric random walk*. We could also look at symmetry random walks in more than one dimension. For instance, in the two-dimensional symmetric random walk, the process would, at each transition, either take a step to the left, right, up or down, each with probability 1/4. Similarly in three dimensions, the process would, with probability 1/6, make a transition to any of the six adjacent points. By using the same method as the one-dimensional random walk, it can be shown that the two-dimensional symmetric random walk is recurrent, but all higher-dimensional symmetric random walks are transient.

3 Clarification about the Proof of Theorem 4.2.8

We know that

$$d(i) \mid t$$

where t is any number in the set $T := \{t : P_{jj}^t > 0\}$. Thus $d(i)$ is a common divisor of the elements of T. By definition, $d(j)$ is the *greatest* common divisor of T. It is known that every common divisor of a set of numbers T divides the greatest common divisor, i.e., it holds that

$$d(i) \mid d(j).$$

For a proof of this non-trivial fact, see `https://proofwiki.org/wiki/Common_Divisor_Divides_GCD` or `https://www.cut-the-knot.org/Generalization/gcd.shtml`. Also see Lemma 4 below for a self-contained proof. Note that $d(j)$ need not be in set T.

Consider the example in which $T = \{t : P_{jj}^t > 0\} = \{4, 8, 10, \ldots\}$ for any $j \in \{1, \ldots, 9\}$ in Fig. 4.2(b) in Gallager's book. Note that $d(j) = \gcd(T) = 2$. Note that $d(j) = 2$ need not be in T, i.e., $P_{jj}^2 = P_{jj}^{d(j)}$ could be (and in fact is) 0. However, $d(i) = 2$ divides every element in the set T. It also divides $d(j)$, i.e.,

$$d(i) \mid d(j).$$

In fact, both $d(i)$ and $d(j)$ are 2 in this case.

Lemma 4. *Say we have two natural numbers m and n, whose greatest common divisor is $d = \gcd(m, n)$. Let a be any common divisor of m and n. It holds that $a \mid d$.*

Proof. Assume, to the contrary, that a cannot exactly divide d, then by definition of exact division, there exist x, y and z such that $a = xy$, and $d = xz$, but $y > 1$ and $z > 1$ are relatively prime. Since a and d divide m, $xy \mid m$ and $xz \mid m$ where y, z are relatively prime. Thus, there exists $a_1, d_1 \in \mathbb{N}$ such that $m = xya_1 = xzd_1$. This implies that $ya_1 = zd_1$. This implies that $z \mid ya_1$ and since y and z are relatively prime, $z \mid a_1$. This implies that $xyz \mid xya_1$ and so $xyz \mid m$.

Similarly, there exists $a_2, d_2 \in \mathbb{N}$ such that $n = xya_2 = xzd_2$. This implies that $ya_2 = zd_2$. This implies that $z \mid ya_2$ and since y and z are relatively prime, $z \mid a_2$. This implies that $xyz \mid xya_2$ and so $xyz \mid n$.

Hence, xyz is a common divisor of m and n. But $xyz > xz = d$. This contradicts the fact that d is the *greatest* common divisor of m and n. □

4 Motivation for Markov Chains

This is an addendum to the lecture on Markov chains. This is completely optional but hopefully interesting.

4.1 Background

Deoxyribonucleic acid (DNA) is the genetic material of a cell which is a code containing instructions for the make-up of human beings and other organisms. The DNA code is made up of four chemical bases: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The sequence of these bases determines information necessary to build and maintain an organism, similar to the way in which the arrangement of letters determine a word. DNA bases are paired together as (A-T) and (C-G) to form base pairs which are attached to a sugar-phosphate backbone (deoxyribose). The combination of a base, sugar and phosphate is called a nucleotide, which is arranged in two long strands that form a twisted spiral famously known as the double helix. See Figure 1.

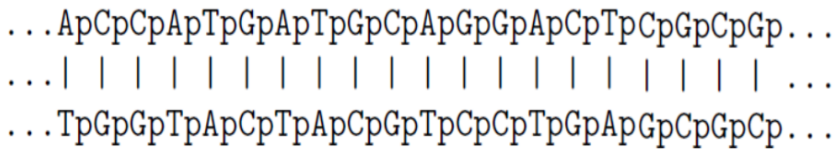

Figure 1: Strand of DNA in the form of a double helix where the base pairs are separated by a phosphate backbone (p)

Since the 1960s, it has been known that the pattern in which the four bases occur in a DNA sequence is not random. Early research into the composition of DNA relied on indirect methods such as base composition determination or the analysis of nearest neighbour frequencies. It was only when Elton [Elt74] noticed that models which assumed a homogeneous DNA structure where inappropriate when modelling the compositional heterogeneity of DNA and thus it was proposed that DNA should be viewed as a sequence of segments, where each segment follows its own distribution of bases. The seminal paper by Churchill [Chu89] was the first to apply HMMs to DNA sequence analysis where a heterogeneous strand of DNA was assumed to comprise of homogeneous segments. Using the hidden states of the hidden Markov model it was possible to detect the underlying process of the individual segments and categorise the entire sequence in terms of shorter segments.

4.2 CpG Islands

To illustrate the use of hidden Markov models in DNA sequence analysis we will consider an example given by Durbin et al. [DEKM99].

In the human genome wherever the dinucleotide CG (sequence of two base pairs) occurs where a cytosine nucleotide is found next to a guanine nucleotide in a linear sequence of bases along its length. We use the notation CpG (-C-phosphate-G-) to separate the dinucleotide CG from the base pair C-G. Typically wherever the CG dinucleotide occurs the C nucleotide is modified by the process of methylation where the cytosine nucleotide is converted into methyl-C before mutating into T, thus creating the dinucleotide TG. The consequence of this is that the CpG dinucleotides are rarer in the genome than would be expected. For biological reasons the methylation process is suppressed in short stretches of the genome, such as around the start regions of genes. In these regions we see more CpG dinucleotides than elsewhere in the gene sequence. These regions are referred to as *CpG Islands* [Bir87] and are usually anywhere from a few hundred to a few thousand bases long.

The presence of a CpG island can be an indication to the start of a gene. Therefore, identifying CpG islands helps to determine the location of genes across the DNA. We would like to answer the following two questions:

1. Question 1: given a short sequence, is it from a CpG island or not?
2. Question 2: given a long sequence, does it contain a CpG island or not?

In terms of our hidden Markov model we can define the genomic sequence as being a sequence of bases which are either within the CpG island or are not. This then gives us our two hidden states $\{\text{CpG island}, \text{Non-CpG island}\}$ which we wish to uncover by observing the sequence of bases. As all four bases can occur in both the CpG island and non-CpG island regions, we first must define a sensible notation to differentiate between C in a CpG island region and C in a non-CpG island region. For A, C, G, T in a CpG island we have $\{A_+, C_+, G_+, T_+\}$ and for those bases that are not in a CpG island we have $\{A_-, C_-, G_-, T_-\}$.

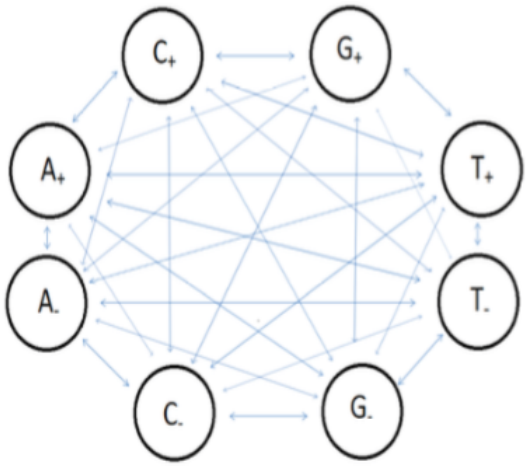

Figure 2: Possible transitions between the 8 states

Figure 2 illustrates the possible transitions between bases, where it is possible to transition between all bases in both CpG island states and non-CpG island states. The transitions which occur do so according to two sets of probabilities which specify firstly the state and then the given observed chemical base from the state.

Once the observations $Y_t = \{A, C, G, T\}$ and the hidden states $S_i = \{\text{CpG island}, \text{Non-CpG island}\}$ have been established, we are then able to construct a direct acyclic graph (DAG) which we shall use to illustrate the dependent structure of the model. The DAG given in Figure 3 shows that observations Y_t are dependent on the hidden states S_i and that both the states and observations are dependent on probability matrices A and B, respectively. The matrix $\text{A} = \{a_{ij}\}$ represents the transition between the two hidden states $\Pr(q_t = S_j | q_{t-1} = S_i) = a_{ij}$ and B denotes the observable state probabilities for the 2-hidden states $\text{B} = (p_+, p_-)$. See Figure 3 for the DAG and Figure 4 for possible values of p_+ and p_-. Note that the transition from C to G in p_+ (0.274) is much higher than the counterpart in p_- (0.078).

4.3 Estimation: Answering Question 1

We will estimate the transition probabilities from statistical data about CpG islands and non-CpG islands. We will therefore build two Markov chains, one for each. Then given a sequence, we compute the probability p of obtaining the sequence in the CpG island Markov chain, and the probability q of obtaining the sequence in the non-CpG island Markov chain. The odds ratio or log-odds ratio of these two probabilities can be used to determine whether the sequence is coming from a CpG island or not. Here are the steps:

1. Bring a set of short DNA sequences labeled + for CpG islands and − for non-CpG islands (therefore these sequences are known to be either coming from CpG islands or not)

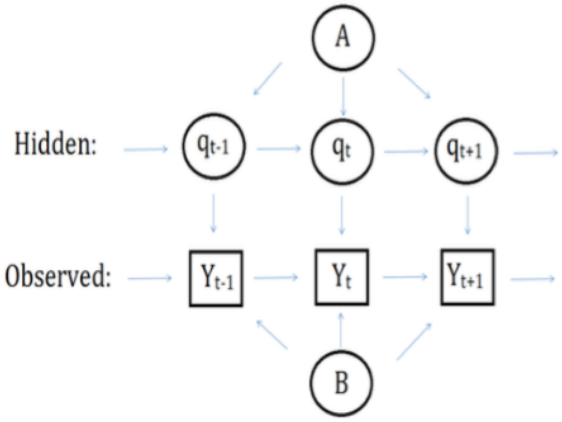

Figure 3: DAG of the hidden Markov model with A representing the state transition probabilities and B representing the observation probabilities for a given state

2. For the CpG island Markov chain, estimate

$$a_{ij}^{+} = \frac{c_{ij}^{+}}{\sum_k c_{ik}^{+}} \tag{15}$$

   where c_{ij}^{+} is the number of times nucleotide j follows nucleotide i in sequences labelled $+$.

3. For the non-CpG island Markov chain, estimate a_{ij}^{-} in a similar way.

Now given a sequence $\mathbf{y} = (y_1, \ldots, y_n)$, compute $p(\mathbf{y})$ for each Markov chain, denote these by $p(\mathbf{y}|+)$ and $p(\mathbf{y}|-)$. Then we use the log-odds ratio $\log \frac{p(\mathbf{y}|+)}{p(\mathbf{y}|-)}$ to determine if x is coming from a CpG island or not: If $\log \frac{p(\mathbf{y}|+)}{p(\mathbf{y}|-)} > 0$, the $\mathbf{y}$ is coming from a CpG island. Assuming that the transitions from the start state and to the end state are the same in both cases, the log-odds ratio can be expressed as

$$\log \frac{p(\mathbf{y}|+)}{p(\mathbf{y}|-)} = \sum_{i=1}^{n-1} \log \frac{p_+(y_i, y_{i+1})}{p_-(y_i, y_{i+1})} \tag{16}$$

Now consider the sequence $CGCG$. The log-odds ratio for this sequence based on Figure 4 is

$$\log \frac{0.274}{0.078} + \log \frac{0.339}{0.246} + \log \frac{0.274}{0.078} > 0 \tag{17}$$

so this sequence is likely to be from from a CpG island.

4.4 Estimation: Answering Question 2

We have developed a startegy to answer Question 1. What about Question 2? We can use the dual Markov chain model that we developed above to find CpG islands in a long sequence of nucleotides. Here's how: consider windows of small size, say 100, in the long sequence. For each window (a short sequence), compute the log-odds ratio as above. Therefore, we can identify windows with positive log-odds ratio and then merge intersecting windows to determine which parts of the long sequence are CpG islands.

The disadvantage of the above approach to Question 2 is that CpG islands tend to have variable length, and a window of 100 might not be appropriate to judge: If the window is too small, then we tend to have

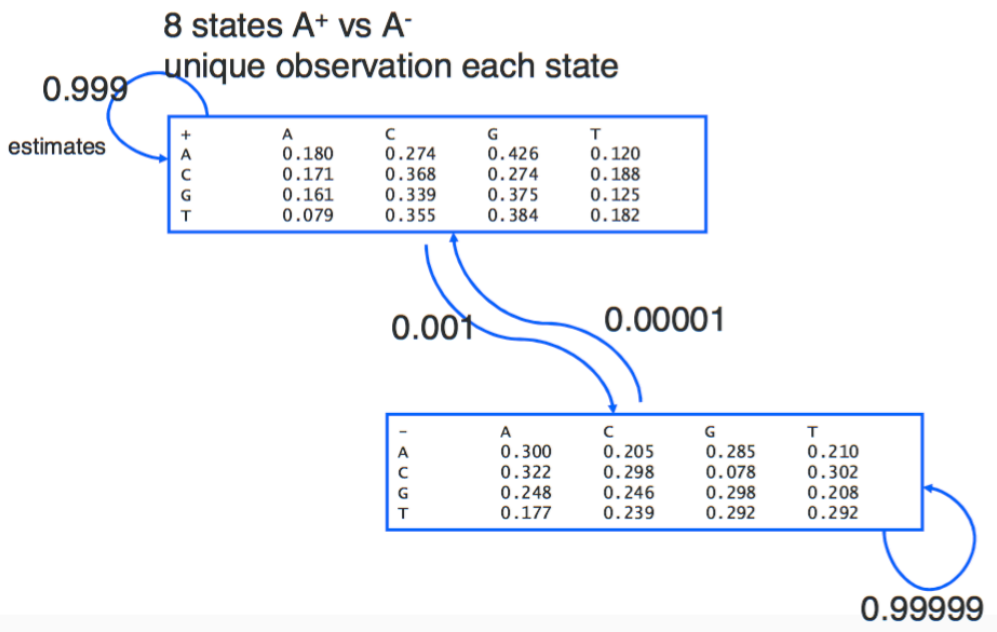

Figure 4: Diagram showing p_+ and p_- and the transitions between CpG island and Non-CpG island

every occurrence of CG as an island by itself. If the window is too large, then we do not achieve enough discrimination (the extreme case being whether the whole sequence is a CpG island or not corresponding to a window size equal to the length of the sequence).

A better way is to incorporate both models (CpG islands and non-CpG islands) into one model. Therefore, we will build a single Markov model consisting of both chains (+) and (−) described above as sub-chains, and with small transition probabilities between the two sub-chains. (See Figure 4 for a global picture.) We rename the states by adding '+' and '−' labels to distinguish them. This relabeling is critical; otherwise, we cannot distinguish states of the new model. See Figure 5.

The advantage of this model is its ability to be trained to reflect reality. As we have done before, we can estimate the transition probabilities between the two sub-chains by relying on known annotated sequences with all their transitions between CpG and non-Cpg islands. This way we remove the dependence on a particular window size.

The problem that we face with this new model, however, is that there it not a one-to-one correspondence between the states and the symbols of the sequence. For instance, the symbol C can be generated by both states C_+ and C_-. Hence, a sequence does not correspond to a path in the model anymore, but to multiple paths. In other terms, a sequence $y_1, \ldots, y_n$ does not uniquely determine the path in the model. The states are hidden in the sense that the sequence itself does not reveal how it was generated. Therefore, we need to use the theory of hidden Markov models and a DAG like in Figure 3 is more appropriate to analyze.

References

[Bir87] A. Bird. CPG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, 3:342–347, 1987.

[Chu89] G. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51:79–94, 1989.

[DEKM99] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, 1999.

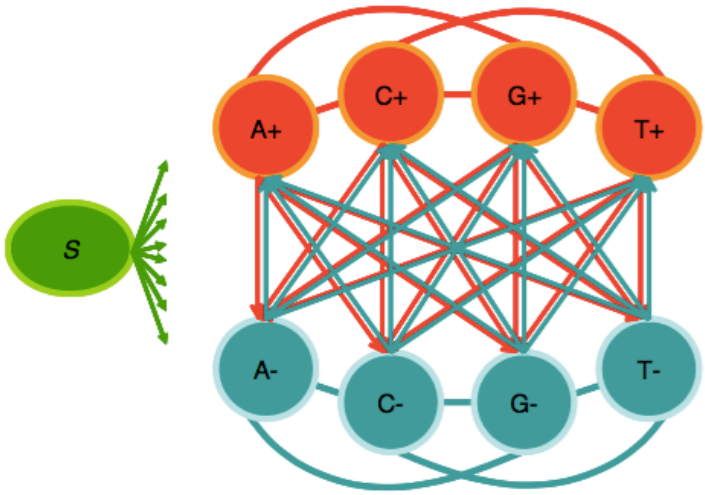

Figure 5: Combining both models

[Elt74] R. A. Elton. Theoretical models for heterogeneity of base composition in DNA. *Journal of Theoretical Biology*, 45(2):533–553, 1974.