Hidden Markov Models & their Applications to Statistical Genetics

Sofia Barragan, Spring 2021, Macalester College

Contents

Welcome

This bookdown on discrete Hidden Markov Models & their applications to statistical genetics is my capstone! I made this for my Mathematical Statistics course taught by Kelsey Grinde. Big thanks to her for her guidance and help.

Content was written and gathered by Sofia Barragan with appropriate citations for print materials. Wherever possible, I try to provide direct link citations for any digital materials or resources.

Embedded Youtube videos are under the sole ownership of their linked creator.

Note: This is a very brief primer & I assume minimal mathematical background.

Chapter 1

Mathematical Intuition

Markov Chains are cool! Hidden Markov Models are also cool, but require more preparation! In this section, we'll go through conditional probabilities & set up the basis to study Hidden Markov Models by getting comfortable with chains first.

1.1 Conditional Probability & Bayes Rule

Basic Concepts

Probability, as a field, formalizes how we predict events with some equally beautiful & ugly notation, intuitive concepts, and complex mathematical principles. But the premise is simple: by ascribing a numeric value to the outcomes of an event, we can abstract the real world and study it with math.

The process of ascribing numeric values to the outcome of an event is called mapping & by mapping all possible probabilities of an event's outcomes, we create a **random variable**.

> **NOTE:** This can be confusing! The "random" part of the word doesn't mean all outcomes have an equal chance of happening; really, it means that within an event, there are multiple possible outcomes.

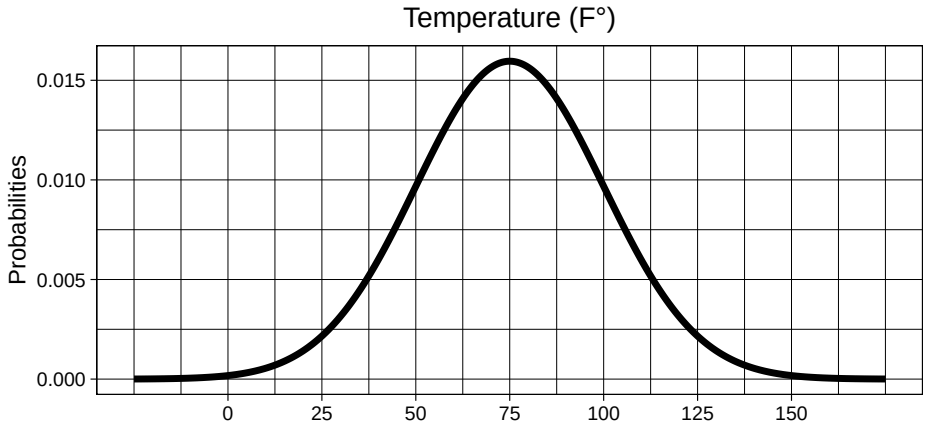
Example: Weather & Temperature

Let's say that the weather is the event, W, whose only **outcomes** are sunny (w_s), rainy (w_r) or cloudy(w_c). By mapping numeric values (e.g., probabilities) to the outcomes, we can turn W into a random variable. Below, we list all mappings in the **probability mass function**, $p_{(W)}(w)$.

$$p_{(W)}(w) = \begin{cases} 0.5, & \text{for } w_s \\ 0.2, & \text{for } w_r \\ 0.3, & \text{for } w_c \\ 0.0, & \text{otherwise} \end{cases}$$

> **NOTE:** The sub-probability of *all* probability mass functions must sum to 1.

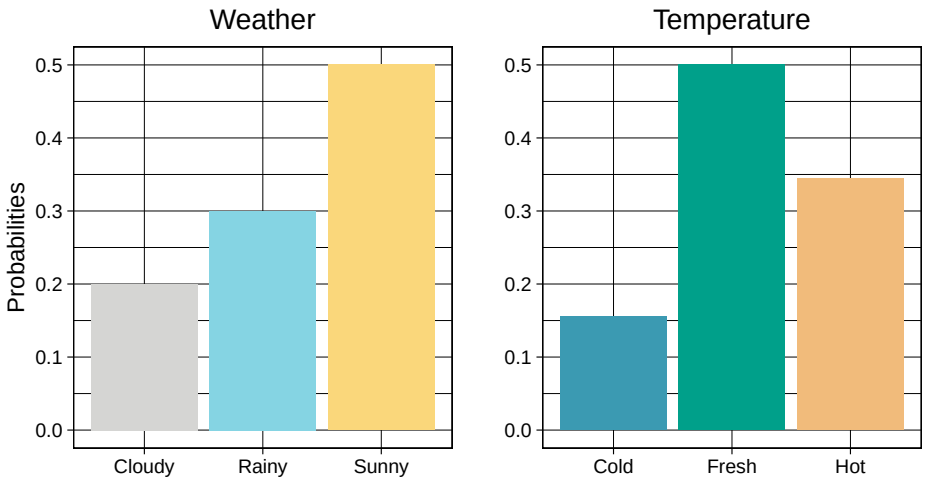
Now, let's see that the daily temperature (F°) is the event T. Normally, we could say that T follows a normal distribution with $\mu = 75$ **F°** and $\sigma^2 = 625$ **F°**, since its a continuous variable. So, $T \sim N(75, 625)$ and would look something like this



Instead, let's say that T is a discrete random variable whose only potential outcomes are cold (t_c), fresh (t_n), or hot (t_h). Then, T has the probability mass function

$$p_{(T)}(t) = \begin{cases} 0.155, & \text{for } t_c \\ 0.5, & \text{for } t_f \\ 0.345, & \text{for } t_h \\ 0.0, & \text{otherwise} \end{cases}$$

Below are two graphs summarizing what we know so far about W and T



But what happens if the weather depends on the temperature?

Conditional Probabilities

Let's study the two arbitrary events A, B & learn some definitions about probability.

Conditional Probability: The conditional probabilities of A's, given B and B, given A are written below.

$$P(A \mid B) \qquad\qquad P(B \mid A)$$

Independence: We say that the two events, A & B are independent if the conditional probabilities provide us no new-information about either event. So,

$$P(A \mid B) = P(A) \qquad\qquad P(B \mid A) = P(B)$$

Joint Probability: The probability of two events, A & B happening at the same time is called a **joint probability** and is typically denoted by $P(A \cap B)$. It is calculated below.

$$\begin{aligned} P(A \cap B) &= P(A \mid B) \cdot P(B) \\ &= P(A) \cdot P(B) \qquad\qquad \text{if independent} \end{aligned}$$

It's generally true that the weather on a particular day, depends on the temperature. This implies that W and T are conditional events with conditional probabilities. So,

$$P(W \mid T) \neq P(W)$$

Bayes' Rule & LOTP

Conceived by Reverend Thomas Bayes in the 18th Century, posthumously published by his friend Richard Price, and then formalized into an equation by Pierre-Simon Laplace, Bayes' Rule is a cornerstone equation in modern statistics & probability[1]. I've written Bayes Rule below[2]

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

> **NOTE**: Above you'll notice that we're looking at the probability of A & B, divided by the probability of B. We are dividing by $P(B)$ to normalize & isolate the probability of A, under the conditions we observe B.
>
> **Quick Check:** If A & B are independent, how would you further simplify the numerator of Bayes' Rule?

The denominator of Bayes' Rule, $P(B)$ is called the **marginal probability** of B.

The marginal probability can either be

[1] https://www.bayesrulesbook.com/chapter-1.html#a-quick-history-lesson
[2] https://www.bayesrulesbook.com/chapter-2.html

- given
- computed with the **law of total probability** or (**LOTP**).

LOTP states that

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n)$$
$$= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_n)P(A_n)$$

> **Note:** Looks scary! Really it's like calculating the probability of B in each subcategory of A, multiplying by the probability of that sub-A, then adding it all together.

1.2 Markov Chains

Stochastic processes are events that have some element of randomness in their outcomes. The amount of 'randomness' & the type of events can vary depending on the context. In turn, studying the properties of stochastic processes often requires many different techniques which go well beyond the boundaries of statistics. And with a litany of applications across so many domains of knowledge, studying stochastic processes also involves many techniques from Physics, Linguistics, Sociology, Public Health, Geography & more.

> **Note:** There is *some* nuance in the language we use to describe randomness, probabilistic, and stochastic, but it is murky. So, for now, let's stick with the above definition & enjoy some interchangeability between random, stochastic, and probabilistic.

So, this section will only be a tiny snippet of the wide topics covered in studying stochastic processes.

Intuition

Markov chains are a subclass of stochastic processes that describe partially random events occurring in succession, typically in succession. We will formalize this definition soon, but for now, let's talk about the weather.

Example: Weather

Let's ignore the fact that temperature determines weather. Instead, let's assume that we can compute the probability of tomorrow's weather by only looking at today's. This results in three important ideas:

- General weather patterns before today are irrelevant when predicting tomorrow's weather. This is an example of a Markov process, more specifically a Markov Chain.
- The outcomes of the weather are rainy, sunny, or cloudy. These are examples of states.
- The probability of tomorrow's weather depends on whether it was rainy, sunny, or cloudy to day. Meaning there are probabilities associated with tomorrow's events, strictly defined by today's. These are examples of transition probabilities.

Let's be rigorous now!

Mathematical Definitions

Let $X_i = X_1, \dots, X_{n-1}, X_n$ be a collection of n successively indexed events.

The discrete X_i are a **Markov Chain** if

- They exhibit a **Markov Property**.
  - Where the probability of some new or predicted event $X_{n+1} = x_{n+1}$ is

$$P(X_{n+1} = x_{n+1} \mid X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = x_n) = P(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

    NOTE: This is similar to the conditional probability of independent events! But instead we specify that prior events $X_1, \dots X_{n-1}$ are independent of the outcome of X_{n+1}, but X_n is not.

- There is some countable set of outcomes called the **State Space** which contains every possible outcome or **State**
  - The outcomes $x_1, \dots, x_{n-1}, x_n, x_{n+1}$ are all common elements of the state space, $\mathbb{S}$.

    NOTE: This may seem complicated, but this means we can define what are possible and impossible outcomes.

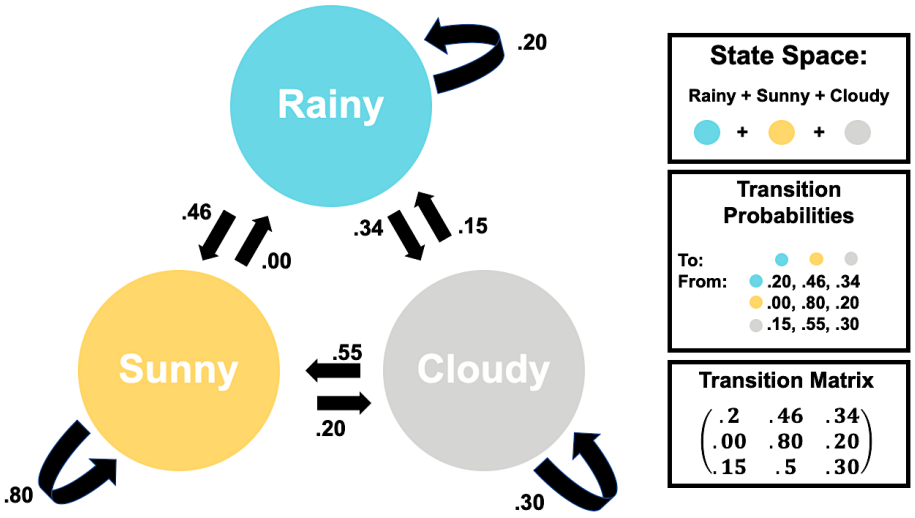
- The probability of changing states is a **Transition Probability** and if we were to write them out for each *state*, they would sum to 1.
  - Transition Probabilities sum to 1 because they cumulatively define the *probability mass function* of the transition from each state to another.
  - These probabilities can be placed in a **Transition Matrix** where the columns indicate a next state & the rows indicate the current state.

NOTE: The above definitions were cumulatively drawn from the following sources [3] [4]

Visualizations

Weather Graphs

Below we've redefined our weather example as a Markov Chain, using our new definitions, and place it into a Weighted Directed Graph. Isn't she pretty! Note that the *weights* on our arrows correspond to the transition probability associated with that change-arrow.



Also, check out that Transition Matrix! It's the first one you've looked at, but they can quickly get very complex.

Weather Animations

To simulate how a Markov Chain of weather behaves, let's animate it!

In this visualization, there are 10 multicolored dots representing different arbitrary days colored by that particular day's weather. To see how Markov Chains evolve over 15 days, we shuffle them 15 times in a row according to our transition matrix

[3] Lay, David C. 2012. "Applications to Markov Chains." In Linear Algebra and Its Applications, 4th ed., 253–62. Boston: Pearson College Division.

[4] https://setosa.io/ev/markov-chains/

Rainy	Sunny	Cloudy
.20	.46	.34
.00	.80	.20
.15	.50	.30

Written text describes the proportion of weather outcomes following simulation. That is, how many rainy, sunny, or cloudy days happen after a given shuffle.

Look at that! We converged on our initial probability distribution. That isn't necessarily true for all simulations, but it's true for this one.

> **NOTE:** This Markov Chain visualization was designed by Will Hipson, a graduate student in Psychology at Carleton University. You can find links for reproduction at his page[5] or check out my GitHub to see my edits.

1.3 Conclusion

We've covered the necessary probability concepts. In the next section, we'll find out out how we can leverage Markov Chains to predict another set of variables, even when we can't see the outcomes of our chain. Prepare for bivariate distributions, many subscripts, and a lot of summation notation!

If you have any lingering questions, I've linked some great YouTube videos that may be helpful below.

Video Resources

Conditional Probability

Bayes Rule

Markov Chains

References

[5] https://willhipson.netlify.app/post/markov-sim/markov_chain/

Chapter 2

Frog Families & HMMs

Previously, we wanted to study the weather. Let's up the stakes and now look at how the weather can determine whether or not a cute, infinitely reproducing, frog-family survives the month.

If you were a frog, you wouldn't be able to check the weather before leaving your tree hide– you wouldn't even be able to read!– but you do know that when a family member leaves for their daily hop, their chance of living depends on the weather. If the weather is

- **Sunny:** they would have a 10% chance of living (and eating a nice bug) with a 90% chance of dying.
- **Cloudy:** they would have a 75% chance of living (and eating a nice bug) with a 25% chance of dying.
- **Rainy:** they would have a 98% chance of living (and eating a nice bug) with a 2% chance of dying.

Since we, as non-frog statisticians, know that the weather is a Markov Chain, we can say the survival of this frog-family is actually a Hidden Markov Model. In the following sections, we will establish what that implies theoretically for us & our frog family.

2.1 Hidden Markov Models

Hidden Markov Models are another class of probabilistic models that model Markov processes whose outcomes cannot be directly observed, but their dependent observed events can be. Often, we're really interested in these unobserved

events and predicting them, so we can work backward and use the observed ones[1].

Intuitively, it's like using the shadow of an animal to guess what it is!

Figure 2.1: Fig 1. Frog Shadow Made by Author

Properties

Let

- $X_i = X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = x_n$ be a collection of n successively indexed events.
- $Y_i = Y_1 = y_1, \dots, Y_{n-1} = y_{n-1}, Y_n = y_n$ be a collection of n successively indexed observations.
- $\lambda = (X, Y)$.

λ is a discrete **Hidden Markov Model** if it fulfills the following definitions:

Discreteness: Both X_i and Y_i are sequential, discrete, random variables with the n observations $x_{1,\dots,n}$ and $y_{1,\dots,n}$.

> **Note:** Neither X_i nor Y_i *must* be discrete events, but this project is explicitly dedicated to discrete Hidden Markov Models, so we will assume they are.

Markov Assumptions:

- The X_i *must* be a Markov Chain.
  - So,

$$P(X_n = x_n \mid X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = x_n) = P(X_n = x_n \mid X_n = x_n)$$

[1] https://www.nature.com/articles/nbt1004-1315#citeas

- The Y_i exhibit a Markov property where the Y_k^{th} observation can be solely predicted using the k^{th} state of the X_i, for some arbitrary k.
  - So, Y_i observations are solely dependent on the state of the hidden X_i and nothing else. Implying that

$$P(Y_n \mid Y_1 = y_1, ..., Y_n = y_n, \quad X_1 = x_1, ..., X_n = x_n) = P(Y_n \mid X_n = x_n)$$

Hidden States:

- **Hidden States:** The unobservable possible outcomes of X_i. Denoted as $x_1, ..., x_{n-1}, x_n$.
- Together, the outcomes form the **Hidden State Space**, which we will denote as $\mathbb{S}$.

Transition, Emission, and Initial Probabilities:

- **Transition Probability:** The probability of changing X_i-states. These are typically placed into a transition matrix.
  - We typically denote this transition matrix as $\mathbb{o}_{ij}$, where we change from state i to state j.
- **Emission Probability:** The probability of Y_i, given the X_i's state.
  - So, $P(Y_n \mid X_n = x_n)$ is the **Emission Probability** which can be placed into the **Emission Matrix**
    - We typically denote these emission probabilities as b_n, where n is the state of the X_i.
- **Initial Probability:** An estimation of X_i's state at some arbitrary point a. This is often denoted by π and is typically as a start guess.

  NOTE: The above definitions were cumulatively drawn from the following sources [2] [3] [4]

Example: Frogs

Let's formalize the frog-family's situation, $\mathbb{F}$, with the above definitions!

Let $W_i = W_1, ..., W_n$ be the weather for n days. We know from previous work that the weather is a Markov Chain. So, for the frogs, the **hidden states** of W_i are sunny, rainy, and cloudy, while the hidden state space will be $\mathbb{W}$.

From the previous sections, we also know that the transition matrix of W_i is

	Rainy	Sunny	Cloudy
Rainy	.20	.46	.34

[2] https://medium.com/@kangeugine/hidden-markov-model-7681c22f5b9
[3] https://jwmi.github.io/ASM/5-HMMs.pdf
[4] https://web.stanford.edu/~jurafsky/slp3/A.pdf

	Rainy	Sunny	Cloudy
Sunny	.00	.80	.20
Cloudy	.15	.50	.30

And let our **initial probabilities** (π) of W be

Rainy	Sunny	Cloudy
.2	.5	.3

Finally, let $D_i = D_1, ..., D_n$ be whether or not we see a frog die after leaving the tree. We know that the **emission probabilities** of the D_is are therefore

	$\mathbf{P}(\mathbf{D_i} \mid \text{Rainy})$	$\mathbf{P}(\mathbf{D_i} \mid \text{Sunny})$	$\mathbf{P}(\mathbf{D_i} \mid \text{Cloudy})$
Survives	.98	.10	.75
Dead	.02	.90	.25

> **NOTE:** These are actually 3 separate emission matrices squashed together for simplicity. The columns contain the desired matrices.

In the next section, we'll show some visual examples of this frog family's HMM, $\mathbb{F}$.

Visualizations

Frog Graphs

Below, we've drawn the Hidden Markov Model, $\mathbb{F}$ & placed it into a Weighted Directed Graph.

Frog Animations

We can reuse our previous animation to simulate state-space transitions and what that implies for our frog family's chance of living.

Now, the dots change colors with the true but unobserved weather of each day. Written text now describes,

- The emission probability of dying on a given day.

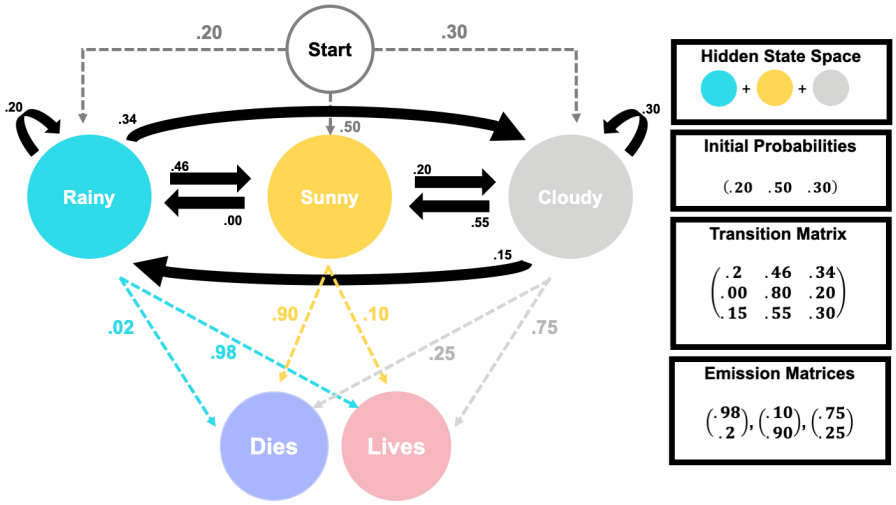


Figure 2.2: Fig 2. HMM Directed Graph. Made by Author

- The number of unobserved, but true, weather states.

What a nice way of looking at the relationship between emission probabilities and the hidden states! In this next section, we'll see how we model the hidden states using observations.

2.2 Algorithms

Let's say the frog family has been really lucky, and all the frogs who left the tree hide in the past week came back! One very curious frog named Betsy wonders what the probability of all her family surviving is, given the weather.

We can answer Betsy's questions using likelihood calculations or the Forward-Backward Algorithm.

Likelihood

Likelihood or **Posterior Probability** are commonly defined as the probability of the data, given a true underlying parameter[5].

[5]https://www.bayesrulesbook.com

Let W_i be the hidden states of the weather in the past week, and let D_i be our observations on whether or not the frogs came back alive. This implies

$$W_i = \{W_1 = w_1, W_2 = w_2, W_3 = w_3, W_4 = w_4, W_5 = w_5, W_6 = w_6, W_7 = w_7\}$$
$$W_i = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$$

and

$$D_i = \{D_1 = d_1, D_2 = d_2, D_3 = d_3, D_4 = d_4, D_5 = d_5, D_6 = d_6, D_7 = d_7$$
$$D_i = \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$$

In this case, the likelihood would be the probability of our observed D_i, given some true underlying set of W_i. This is computed below.

$$\begin{aligned} P(D_i \mid W_i) &= P(d_1 \mid w_1) \times P(d_2 \mid w_2) \\ &\quad \times P(d_3 \mid w_3) \times P(d_4 \mid w_4) \\ &\quad \times P(d_5 \mid w_5) \times P(d_6 \mid w_6) \\ &\quad \times P(d_7 \mid w_7) \\ P(D_i \mid W_i) &= \prod_{i=1}^{7} P(d_i \mid w_i) \end{aligned}$$

NOTE: The capital pi ($\prod$) means to multiply a series indexed by i.

If we knew the states of the past week were `Sunny`, `Cloudy`, `Rainy`, `Rainy`, `Rainy`, `Rainy`, `Rainy` and everyone came back alive. We'd compute the likelihood as

$$\begin{aligned} P(D_i \mid W_i) &= P(Lives \mid Sunny) \times P(Lives \mid Cloudy) \\ &\quad \times P(Lives \mid Rainy) \times P(Lives \mid Rainy) \\ &\quad \times P(Lives \mid Rainy) \times P(Lives \mid Rainy) \\ &\quad \times P(Lives \mid Rainy) \end{aligned}$$

$$\begin{aligned} P(D_i \mid W_i) &= .10 \times .75 \times .98 \times .98 \times .98 \times .98 \times .98 \\ P(D_i \mid W_i) &= 0.06779406 \\ P(D_i \mid W_i) &\approx 0.068 \end{aligned}$$

Bad News: We can't truly know the states of the W_i! Bummer.

Because of this, we would need to add together all likelihoods for all possible states for each event, weighted by their probability[6] . More succinctly, we intend to find the likelihood of all observations $P(D_i)$, given the probability of all states $P(D_i \cap W_i)$. We can do this with the rules of joint probability we covered in the previous section.

$$\begin{aligned} P(D_i \cap W_i) &= P(D_i \mid W_i) \times P(W_i) \\ &= \prod_{i=1}^{7} P(d_i \mid w_i) \times P(W_i) && \text{by above} \\ &= \prod_{i=1}^{7} P(d_i \mid w_i) \times \prod_{i=1}^{7} P(w_i \mid w_{i-1}) && \text{by Markov} \end{aligned}$$

Now, we sum up all the observations given.

$$P(D_i) = \sum_{i=1}^{n} P(D_i \cap W_i) = \sum_{i=1}^{n} P(D_i \mid W_i) \times P(W_i)$$

So, for our 7 consecutive live frogs, this would be

[6] https://web.stanford.edu/~jurafsky/slp3/A.pdf

$$
\begin{aligned}
P(D_i) = {} & P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \\
& \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \\
& + P(Lives \mid Cloudy) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \\
& \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \\
& + P(Lives \mid Sunny) \times P(Lives \mid Cloudy) \times P(Lives \mid Sunny) \\
& \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \\
& + P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Cloudy) \\
& \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \times P(Lives \mid Sunny) \\
& + ... \\
& = P(Lives \mid Rainy) \times P(Lives \mid Rainy) \times P(Lives \mid Rainy) \\
& \times P(Lives \mid Rainy) \times P(Lives \mid Rainy) \times P(Lives \mid Rainy) \times P(Lives \mid Rainy)
\end{aligned}
$$

That is *extremely difficult* to do when we the number of our states (N) are numerous or we have some arbitrarily large number observations, T, since we'd be analyzing N^T different possible sequences. However, we can predict the hidden states of HMMs like this one, using the Forward-Backward Algorithm.

The Forward-Backward Algorithm

The Forward-Backward Algorithm is a dynamic programming algorithm used to infer the probability of seeing observations in a Hidden Markov Model. It contains two sub-algorithms, the Forward Algorithm & the Backward Algorithm, which respectively compute the probability of the data from the beginning and the end of observations until they collide and result in a total probability of the data[7]. The process may seem extra, but as you've seen in the last section's toy computation, we have to be extra out of necessity.

> **NOTE:** The following sections heavily drawn from the following sources [8] [9] [10] [11]

[7] https://web.stanford.edu/~jurafsky/slp3/A.pdf
[8] https://web.stanford.edu/~jurafsky/slp3/A.pdf
[9] https://scholar.harvard.edu/files/adegirmenci/files/hmm_adegirmenci_2014.pdf
[10] https://www.cs.tut.fi/kurssit/SGN-24006/PDF/L08-HMMs.pdf
[11] http://www.columbia.edu/~mh2078/MachineLearningORFE/HMMs_MasterSlides.pdf