Notes on Stochastic Processes

Danny Chen

Contents

Preface

April 8, 2021

This is meant to be an informal summary of stochastic processes learned over the course of, well, I don't know. At the very least, I will put all that I got out of the reading course done with Professor Peter J. Thomas down. I'll try to keep up with updating this throughout my later education.

With that being said, I would first and foremost thank Professor Thomas for assisting me in the reading course. I would also want to thank my dad for getting me into this messy but very pleasant world of stochastic processes in the first place.

April 20, 2021

I had a bit too much fun writing... I'm elaborating on the intuition too much but without giving concrete examples, and that would just make it unpleasant to read. With that being said, I'll cut down on unnecessary development of intuition because this is meant to be notes rather than an introductory text. I'll keep the humorous (at least *I* feel like they are funny) parts. In addition, I'll try to keep the proofs sparse.

Chapter 1

Some Probability Theory

1.1 A Different Integral

Discrete random variables are just lies. Well, that's not *exactly* true. But after second thought, the formulation of "probability mass functions" seems quite a clumsy one. So, let's take a step at generalizing it. Our first step is to make everything a density. And to do so, we need the notion of a dirac delta function.

Definition (Dirac Delta)**.** *The Dirac Delta function, $\delta(x)$, is a probability distribution concentrated at $x = 0$, that is, the function satisfies the two conditions:*

1. *for all $x \neq 0$, $\delta(x) = 0$, and*
2. *$\int_{-\infty}^{\infty} \delta(x)\ dx = 1$*

Now, we can see that discrete random variables can also be written as a density, comprised of delta functions at discrete values. But, let's think about it a bit more... Is this Riemann integrable?[1] Before diving into measure theory, let's think: what do we have right now? Even for discrete distribution, we always have a (right) continuous cumulative distribution!

Definition (Riemann-Stieljes Integral)**.** *Let $f : [a, b] \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be real-valued bounded functions. Then, the Riemann-Stieljes Integral is defined to be the following limit.*

$$\int_a^b f(x)\ dg(x) = \lim_{n\to\infty} \sum_{k=0}^{n-1} f(\xi_k)(g(x_{k+1}) - g(x_k))$$

with respect to some partition $a = x_0 < x_1 < \cdots < x_n = b$ and $\xi_k \in [x_k, x_{k+1}]$.

This is kind of like the Riemann integral, but integrated with respect to a function! So now, we can write expectations like this:

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\ dF(x)$$

for some arbitrary function g and the cumulative distribution F.

some measure theory

1.2 The Variety of Convergence

Similar to how we have pointwise convergence and uniform convergence in sequences of functions, we have different types of convergence, too, for sequences of random variables.[2]

[1] It turns out, this is the wrong question to ask. It's a distribution, so at the end of the day, it has to be a measure. What I was trying to probe at is that more care is needed than the regular Reimann integral here.

[2] This title is inspired by the chapter "The Variety of Ion Channels" in Ermentrout and Terman's *Mathematical Foundations of Neuroscience* [2].

1.3 Limit Theorems

Chapter 2

Renewal Theory

Renewal Processes is a general framewoork for describing arrivals with stochastic interarrival times. So, it might answer questions like "when the heck is the bus coming?" on a snowy Cleveland day and you're too lazy to walk to the classroom at 9 am in the morning. Not me though... I like snow!

2.1 The General Set Up

Let's suppose we're tired of having to wait an unknown amount of time for the bus. So, we take the initiative and wanted to model the arrival process of buses. What do we need? Generally, there are three quantities of interest.

1. X_n, the waiting time from the n-th arrival to the $(n+1)$-th arrival
2. S_n, the time until the n-th arrival
3. $N(t)$, the number of arrivals in the interval $(0, t)$

And since we're dealing with Renewal Processes, we want the process to *renew* after each arrival. So, X_i's must be iid. Also, we can see that $S_n = \sum_i X_i$, so while this quantity will be of interest, it's not too useful most of the time.

2.2 Stopping Time

We will introduce the concept of stopping time here just for the sake of proving an upcoming theorem. This would come up again, and perhaps more intuitively, in the Martingale section.

Stopping time is, well, the time you stop. So, if you're in a casino, this might be the time when you loss your 20 dollars (which should be done in no time) and decided that the money will be better spent on buying two double cheese burgers with fries and milkshake. If you're doing clinical trials, this might be the time when the number of patients who got sick after the drug surpassed a certain number. The formal definition is as follow.

Definition (Stopping Time)**.** *A stopping time τ for a sequence of random variables $X_1, X_2, \ldots,$ is an indicator random variable $\mathbb{I}_{\tau=n}$ that is a function of $X_1, X_2, \ldots, X_n$.*

The notion of stopping times would be revisited in the Martingale section, and one can gain a stronger intuition there. For now, we'll use it as a tool to get the expected count via Wald's Equality.

Theroem 1 (Wald's Equality)**.** *Let τ be a stopping time for a sequence of iid random variables $X_1, X_2, \ldots$. Then, $\mathbb{E}(S_\tau) = \mathbb{E}(X)\mathbb{E}(\tau)$.*

Proof. The proof here uses a few clever tricks. First, we spot that we can rewrite S_τ as the following.

$$S_\tau = \sum_{n=1}^{\infty} X_n \mathbb{I}_{\{n \leq \tau\}}$$

Then, looking at the complement of what we're indicating over, $n > \tau$, we see that it is independent of everything happening in $X_1, X_2, \ldots, X_n$. The rest would be an exercise. □

expected count to time t

2.3 All Kinds of Asymptotic Behaviors

The asymptotic behavior of renewal processes can be summed up in one sentence: just write down whatever feels the most intuitive and that will be true *almost surely*. Most of the statements will be proved using squeeze theorem, sandwiching the desired quantity by random variables with probability 1.

2.3.1 Strong Law of Renewal Processes

The Strong Law of Renewal Processes kind of gives the asymptotic behavior (as $t \to \infty$) for the average count of the process. We can try to think about what quantities might affect the average count over time, and the only plausible answer seems to be the average interarrival times. Intuitively, if there are n arrivals within the interval $(0, t)$, then we would guess that the average waiting time is t/n. We can see the relation between this intuition with the Strong Law of Renewal Processes quite clearly once we see the theorem.

Theroem (Strong Law of Renewal Processes)**.** *Consider a renewal process with mean interarrival time as* $\mathbb{E}(X) < \infty$*, then*

$$\lim_{t \to \infty} \frac{N(t)}{t} \to \frac{1}{\mathbb{E}(X)}$$

Proof. The proof here is done by cleverly spotting the following inequality. Let S_n be the time of the n-th epoch.

$$\frac{N(t)}{S_{N(t)+1}} \leq \frac{N(t)}{t} \leq \frac{N(t)}{S_{N(t)}}$$

Both the upper and lower bound converges of $1/\mathbb{E}(X)$ almost surely. So, the quantity being sandwiched also converges to the same value with probability 1.[1,2] □

2.3.2 The Elementary Renewal Theorem

asymptotic time average of count

2.3.3 Renewal-Reward Processes

waiting time

Here, we will consider

2.3.4 Delay Renewal Processes

asymptotics hold true even when the first is delayed

[1] There are a bit more subtlety here. First, by the mean interarrival time being finite, we now that the count as $t \to \infty$ must be infinite almost surely. If this is not true, then there are sample paths for which the convergence is not true and we wouldn't have the result almost surely.

[2] The other subtlety here is that regular Strong Law of Large Number states that the average converges to the expectation almost surely, but reciprocal might not be true. We have to use the Continuity Theorem here which states that continuous transformation of an almost sure convergence also converges almost surely. Then, since waiting times are positive, we can transform both sides by $f(x) = 1/x$ to yield the final result.

2.4 A Special Case: Poisson Processes

Poisson Processes are great. They are quite simple to analyze and consequently, widely used for modelling many things like how long the line is for the small ramen place (queues, that is) and spike trains.

2.4.1 Different Interpretations

There are many different ways a Poisson Process can be thought of. Let's begin with the most generic, but my personal favorite interpretation, which views the Poisson Process as sequences of interarrival times.

Definition (Poisson Process)**.** *A Poisson Process with arrival rate λ is a renewal process defined by the interarrival times $\{X_i; i \in \mathbb{N}\}$, where X_i's are iid exponential random variables with mean $1/\lambda$.*

Many good properties can be derived out of this definition easily using the memoryless property of exponential distributions. More specifically, given any time Z after the n-th arrival, the distribution of the first arrival after Z is always an exponential. Of course, while the memoryless property is great mathematically, it's quite mind-boggling intuitively: *is the bus ever coming?*

Deriving S_n, the distribution of the n-th arrival time, and $N(t)$, the number of arrivals before t, also becomes quite straightforward by envoking a few results from probability theory.

Theroem. *Suppose there is a Poisson Process with interarrival time exponentially distributed with rate λ. Then, the time of the n-th epoch, denoted S_n, and the number of arrivals from $(0, t)$, denoted $N(t)$, has the following distributions.*

$$S_n \sim \text{Gamma}(n, \lambda), \quad N(t) \sim \text{Poisson}(\lambda t)$$

Proof. To show that $S_n \sim \text{Gamma}(n.\lambda)$, just recall that the sum of exponential distributions follow a Gamma distribution. There are many ways of deriving the Poisson count. My favorite is to write the probability of n arrivals before time t as the marginal of the joint distribution with S_n [3].

$$\text{Prob}(N(t) = n) = \int_0^t \text{Prob}(N(t) = n | S_n = \tau) f_{S_n}(\tau)\ d\tau$$

This integral should evaluate to the Poisson pmf. □

Alternatively, we can define Poisson Processes by zooming in at the probability of arrival at a small window δ. For those Calculus wiz out there, this might be more appealing since it might offer a intuition than taking the distributions for granted.

Definition (Poisson Process)**.** *The Poisson Process with rate λ is a counting process for which the arrivals are stationary and independent in non-overlapping intervals while satisfying the following relation: let $\tilde{N}(t, \tau) = N(t) - N(\tau)$, for $\tau \leq t$, then*

$$\begin{aligned}
\text{Prob}(\tilde{N}(t, t+\delta) = 0) &= 1 - \lambda\delta + o(\delta) \\
\text{Prob}(\tilde{N}(t, t+\delta) = 1) &= \lambda\delta + o(\delta) \\
\text{Prob}(\tilde{N}(t, t+\delta) > 1) &= o(\delta)
\end{aligned}$$

Proof. Just a reminder, $o(\delta)$ is any function that is asymptotically (strictly) smaller than δ as $\delta \to 0$.[4] Let's define $p_n(t)$ be the probability of having n spikes at time t. Then, the above condition is sufficient to show that the following differential equations are holds.

$$\begin{aligned}
\frac{d}{dt} p_0(t) &= -\lambda p_0(t), \quad p_0(0) = 1 \\
\frac{d}{dt} p_n(t) &= \lambda p_n(t) + \lambda p_{n-1}(t), \quad p_n(0) = 1
\end{aligned}$$

[3] This is from Exercise 2.3 in [3].

[4] In mathematical terms, if a function $f(\delta) = o(\delta)$, then $\lim_{\delta \to 0} f(\delta)/\delta = 0$. Intuitively, one can think of it as having δ as an upperbound approaching 0. For example, $\delta^2 = o(\delta)$. One can see that, at when δ is small and positive, δ^2 is upperbounded by δ. So, in the limit, δ squeezes δ^2 down to 0.

The initial condition makes sense because, well, the probability of having any number of spikes at time 0 is 1. Solving this will tell you that the interarrival times are exponentially distributed (by stationary) and the total count to time t is Poisson. □

2.4.2 Poisson Processes under Different Circumstances

Now, let's consider what happens when combining two Poisson Processes. For example, you have two neurons with different spiking rates and you want to consider them as one spike train (I don't know why you would want to do that, but it's doable). Intuitively, this just corresponds to adding up the rate, and this intuition is very much correct.

Let X and Y be the interarrival times of two independent Poisson processes with rate λ and μ respectively. If we think about the interarrival times conditioned on one of the process just arrived, the waiting time of each process still follows an exponential. Then, we can find distribution the minimum of the arrival times (or the first of the two to arrive) as follow.

$$\text{Prob}(\min\{X, Y\} < z) = 1 - \text{Prob}(\{X \geq z\} \cap \{Y \geq z\})$$

Then, by independence, we can see that this follows an exponential distribution with rate $\lambda + \mu$. This can be easily extended to finitely many (I'm suspecting countably many, too, as long as the sum converges) Poisson Processes, in which the rates are just added. This will be useful for interpreting Markov Processes in the future.

Of course, you might wonder, "what if the rate is not constant?" The spiking rate of neurons are definitely not constant at all times, neither is the line in front of the small ramen shop at different time of the day. So, let's model the arrival rate as a function of time, $\lambda(t)$. Intuitively, one might suggest that the count of arrivals within any interval (τ, t) is just the integral over the interval of the rate. This turns out to be also true, which is very convenient.

Just as an aside, doing inference on this continuous rate function might be difficult. One way that Kass, Ventura, and Brown [4] suggested is to plot the histogram of spikes and then smooth the histogram from, I guess, the usual kernel density estimation tricks. But, I have not yet study those carefully, so I will not elaborate any further than this.

Chapter 3

Markov Process

Markov Processes are everywhere. It is used for a lot in robotics and artificial intelligence. For example, one might argue that the probability of winning a chess game is only dependent on the current board, which is exactly the Markov property. Also, the reason why Bayesian statistics gets such a boost in attention was due to the development of Markov Chain Monte Carlo, a method that enables sampling from analytically intractable posterior distributions.

3.1 Markov Chain

The fundamental concept is actually quite simple. We're given a bunch of states (this can be the location of your robot, the amount of money you have in the casino, or the number of activated gating proteins in a nerve cell), and we have our stochastic process be indexed by time-like increments, $\{X_n; n \in \mathbb{N}\}$, where X_n represents the state taken at the n-th time step. Then, the stochastic process is said to have the Markov Property if the distribution of states for the next time step is only dependent on the current time step, i.e.

$$\text{Prob}(X_{n+1} = i \mid X_n = j, X_{n-1} = k, \ldots, X_0 = \ell) = \text{Prob}(X_{n+1} = i \mid X_n = j)$$

And this property gives some great properties, as we will see later on.

3.1.1 Finite State Markov Chain

We will begin the discussion when we have a finite number of states. One common way to visualize this is to think about a graph where the nodes are states and edges are weighted according to the probability that one state will go to another. And the most common way to mathematically represent a graph is through a matrix. Let's call this matrix P, and P_{ij} is the weight of the edge going from state i to state j. In the context of a Markov Chain, this is the probability of ending up in state j starting from state i in one step.

It's noteworthy that P is a stochastic matrix, meaning that the entries are non-negative and the rows sum up to 1. This has some desirable linear algebraic properties.

Theroem. *(Some Properties of the transition Matrix) Let P be a stochastic matrix describing the transition probabilities of a Markov Chain, then*

1. *Th entry P_{ij}^n represents the probability of going from state i to state j in n steps.*
2. *$P_{ij}^{m+n} = \sum_k P_{ik}^m P_{kj}^n$; this is also called the Chapman-Kolmogorov equation.*
3. *The largest eigenvalue of P is 1, and this corresponds to $\mathbf{1}$, the all-one vector, as the right eigenvector.*

Keep these properties in mind. They will be useful in the next section when we discuss properties of Markov Chains.

3.1.2 Countable-State Markov Chain

Of course, the notion of Markov Chains can be expanded to countable number of state spaces. One special case of such Markov Chains is something like this:

First, consider a finite state Markov Chain with organize the states into a line. Then, at every point, you go right with probability p and left with probability $1-p$ (of course, there's only one direction to go if you're at the endpoints). Solving for the equilibrium here is quite simple (well, tedious, but conceptually simple), and one will see that the steady-state probabilities looks like a geometric sequence. Then, if you take $n \to \infty$, you still have a geometric sequence, but depending on the odds of going left as oppose to right, you will either have an actual steady-state distribution or things out get concentrated on the right end, when really there's no end.[1]

This example shows that, like like what elementary real analysis taught us, everything is wonky in the limit. But, we will not focus on specific examples like this in this section. Instead, we'll use it as a generalization to talk about important properties of Markov Chains.

3.2 Properties of Markov Chains

Now, we consider a few properties or problems that one would often consider when looking at Markov Chains. First passage time and recurrence looks at when you would first arrive at a place and when would you return. Steady state probability concerns the distribution after walking on the a graph for a long time. Reversibility peeks at what happens when you walk backwards in time.

3.2.1 First Passage and Recurrence

First passage time, as its name suggests, is the time when you visit a particular state for the first time. It's not hard to imagine the numerous applications of this.

Let's begin with the finite case. We're going to cheat a bit and modify our chain so that once we enter our target, say (without loss of generality), the first state, it enters a self-loop with probability 1. The path leading up to it is unchanged, but this *trapping state* construction makes our lives slightly easier (for reason you'll see soon).

First consider the expected first passage time given that we start in state i, and let's call this value ν_i. Obviously, if we start at state 1, then we're done! So, let $\nu_1 = 0$. For the cases of starting at the other states, the expected first passage time has to satisfy the following:

$$\nu_i = 1 + \sum_j P_{ij}\nu_j$$

The 1 comes from taking one step into the future, and the expected first passage time from the our current position is just a weighted sum of the expected first passage time starting from the next state. So, we can write it into a matrix form.

$$\nu = r + P\nu$$

where r is the vector that is 0 at the first entry and 1 everywhere else. Now, we see the purpose of the trapping state construction! This system should have a solution as long as the states are transient.

For the expected recurrence, the time until I return to my starting state, we can do the same thing. We will again solve for the vector of first passage times. Then, we can weight that according to the probability of transitioning out of the starting state. And since the expected first passage starting from the starting state is defined to be 0, we have to take this into account and add the probability of self-loop to the final weighted sum.

Recurrence on infinite state-spaces is like the dog who went on a walk on his own or your ex: it's not just a matter of *when*, it's a matter of *will it ever* come back? Let's first set up a few notations.

[1] Exercise 4.9 of [3].

Definition (First-passage time)**.** *Given a Markov Chain with countable state-spaces, we define the first-passage time from state i to state j, denote $f_{ij}(n)$, as the following.*

$$f_{ij}(n) = \text{Prob}(X_n = j, X_{n-1} \neq j, \ldots, X_1 \neq j | X_0 = i)$$

As this is a probability distribution over $\mathbb{N}$, we can define the cumulative distribution of the first passage time, $F_{ij}(n) = \sum_k f_{ij}(k)$. Furthermore, define $F_{ij}(\infty) = \lim_{n\to\infty} F_{ij}(n)$.

We can find the first passage time recursively via the following.

$$f_{ij}(n) = \sum_{k\neq j} P_{kj} f_{kj}(n-1)$$

This is true for $F_{ij}(n)$ and $F_{ij}(\infty)$, too. This looks pretty handy, but not really because we can't really numerically take the limit, and we're interested in the limiting behavior more often than not.

Then, let's define an important property that will matter much more in the next section when we look at steady state probability.

Definition (Recurrent State)**.** *A state j is called recurrent if $F_{jj}(\infty) = 1$. On the other hand, it's transient if $F_{jj}(\infty) < 1$.*

So, a state is recurrent if I will return to it with probability 1. And sadly, the adventurous dog and your ex are more likely to be one of the transient states. Anyways, doesn't this look awfully familiar? Let's try to apply some renewal theory ! First, each recurrence is iid, so renewal theory applies. Then, we can track the number of recurrence and the time it takes to the next recurrence. A few of the results are summarized in the theorem below.

Theroem (Renewal Theory on Recurrent States)**.** *Let $N_{jj}(t)$ be the number of recurrences of state j, and T_{jj} be the time for a recurrence to happen. Then, the following conditions are equivalent:*

1. *State j is recurrent*
2. $\lim_{t\to\infty} N_{jj}(t) = \infty$
3. $\lim_{t\to\infty} \mathbb{E}(N_{jj}(t)) = \infty$
4. $\lim_{n\to\infty} \sum_n P_{jj}^n = \infty$

Furthermore, by the strong law of renewal process, if state j is recurrent, then

$$\lim_{t\to\infty} \frac{N_{jj}(t)}{t} \xrightarrow{a.s.} \frac{1}{\mathbb{E}(T_{jj})}$$

To conclude this section, we will introduce the different notions of recurrent states. This would be useful also in the upcoming section.

Definition (Positive/Null Recurrent)**.** *A state j is positive recurrent if $\mathbb{E}(T_{jj}) < \infty$. On the other hand, it is null recurrent if $\mathbb{E}(T_{jj}) = \infty$.*

3.2.2 Steady State Probability

One of the most desirable properties of Markov Chains is that, under the right conditions, the probability of observing a state becomes asymptotically constant in time, hence the name "steady state".[2] This will be the center of discussion for many topics and applications of Markov Chains. The full-on analysis takes some

[2]Identifying when the steady state exists involves the classification of states in a Markov Chain that I don't particularly enjoy. Overall, for steady states to exist, we want to guarantee that all states in the chain can be visited (recurrent clas) and in an un-periodic fashion. We call this type of chains an *ergodic* Markov Chain.

time, so we'll go over the intuition. If a system with finitely many states is in steady state and we list the probability of observing each state in a row vector, π, then, we would want the following to hold true.

$$\pi = P\pi$$

This just means that upon walking one step, the distribution of states does not change. If we look at this algebraically, we see that π is nothing but the left eigenvector corresponding to eigenvalue 1, which we know exists. And that's it! One other thing to note is that, by looking at the eigenvalue decomposition, as the number of steps we take increases, all other eigenvalues of the n-th order transition matrix goes to 0 since they are less than 1. Thus, the rate at which the chain mixes (or approach steady state) is roughly geometric with the second largest eigenvalue.[3]

Of course, when we move to countable state spaces, we don't have the luxury of doing linear algebra like this. What we do have, however, is renewal theory! First, the definition of steady state doesn't change: we still want

$$\pi_i = \sum_k \pi_k P_{ki}, \quad \sum_i \pi_i = 1$$

Keeping this in mind, the description of the steady state distribution is presented in the theorem below.

Theroem (Steady State of Countable Markov Chains)**.** *Suppose we have a countable state Markov Chain for which all states commute with transition probabilities $\{P_{ij}\}$. Then, the chain is positive recurrent if and only if a unique steady state exists. Furthermore, $\pi_j = 1/\mathbb{E}(T_{jj})$.*

3.2.3 Reversibility

3.3 Markov Chain Monte Carlo

I would feel bad if I'm writing about Markov Chains and not talk about Markov Chain Monte Carlo (MCMC). It probably single-handedly popularized Bayesian statistics, the thing that every college students interested in machine learning dreamed of doing.[4] However, I'm not an expert at it. So, I'll list the few things that I do know and add more once I learn more about it.

3.3.1 Metropolis-Hastings Algorithm

3.3.2 Gibbs Sampling

3.4 Markov Process

[3] This gets more complicated as we get specific into the structure of the chain. This is true for ergodic chains (chains where you can go to any other states in infinite amount of time with a non-zero probability). However, if we have transient states, we would have to take into account the time spent outside until it wanders into the recurrent state.

[4] I have no evidence for either statements, but for the latter one, I was one of those a year ago. But I guess I still am one of them now. As you can see, there's a chapter on Bayesian Nonparametrics.

Chapter 4

Martingale

4.1 Threshold Crossing

4.2 Martingales

4.2.1 Stopping Process

Chapter 5

Brownian Motion

Appendix A

Dirichlet Process

This was my project for *STAT 448: Bayesian Theory with Applications* class. It involves some stochastic process... so, here it is.

A.1 Introduction

Imagine you're stranded on an uninhabited island. For survival, you decided to record the animals you see to better understand what you're working with. And for better organization, you want to group those animals by their features. Suppose you're very diligent in recording the species. Eventually, you discovered so many new animals that you have to store it into a computer (that conveniently appeared in your backpack), and you decide to have the computer do the species grouping for you. However, you immediately realize that, as more animals are logged into the system, new groups might be formed and you have to continuously update your model. Is there a model flexible enough so that you don't have to do that?

This is the motivation behind nonparametric bayesian methods. A lot of the time, you're unsure about what's going to happen. In the example above, you have no idea how many species will be present; in fact, new species are being discovered every day. If you're doing topic modeling for, say, Wikipedia, you'll expect yourself to find more and more exotic topics going from one click to the next. Instead of placing a prior on the number of species or the number of topics as you would normally do for parametric methods, Bayesian nonparametric places a prior on the space of distributions. This greatly improves the flexibility of the model since one goes from thinking about a finite-dimensional parameter space to an infinite-dimensional one.

A.2 Dirichlet Process

This section introduces some of the properties of the Dirichlet process, including the definition, methods for sampling it, and useful variants of the process. We begin by showing the famous Stick Breaking construction of Dirichlet processes and the Chinese Restaurant process – a related stochastic process. Then, we will introduce the mathematical formalism that connects the two processes.

A.2.1 Stick Breaking and the GEM distribution

We introduce the Dirichlet process first not by giving the definition, but by showing a popular way of sampling from it. It is often called the *stick breaking construction*, and it turns out to be a good name for it as we will later see.

We begin by reviewing some properties of the Beta distribution that will give valuable intuition for the upcoming materials. Recall that, if a random variable $X \sim \text{Beta}(\alpha, \beta)$, then X has the following density.

$$f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

Let's look at a special case of the beta distribution where we fix the first parameter to be 1 and the second parameter we'll call α. Figure A.1 shows the density function of $\text{Beta}(1, \alpha)$ with a few different values of α.

Observe that when α is small, the draws from the distribution tends to be closer to 1. On the other hand, when α is big, the draws would be closer to 0. This is an important observation, and we'll soon see the significance.

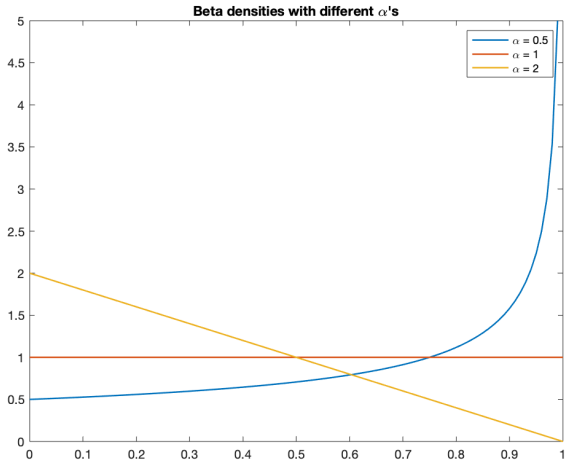

Figure A.1: The density of Beta distributions with different parameters.

Now, we'll begin the stick-breaking. Imagine a stick of unit length, and hold a value $\alpha > 0$ fixed. Then, define $\rho_1 \sim \text{Beta}(1, \alpha)$ to be a draw from the beta distribution. Since this number is between 0 and 1, we'll break the stick off at ρ_1 and put that piece aside; now, we have a stick of length $1 - \rho_1$. Then, we repeat! Draw $\pi_2 \sim \text{Beta}(1, \alpha)$ and break the remaining stick at proportion ρ_2 and set it aside. Now, we have a stick of length $(1 - \rho_1)(1 - \rho_2)$ left. Let's denote the length of each broken stick we set aside as π_k and we can write these lengths down recursively.

$$\rho_i \sim \text{Beta}(1, \alpha), \quad \pi_1 = \rho_1, \quad \pi_k = \prod_{i=1}^{k-1} (1 - \rho_i)\rho_k,$$

This infinite collection of lengths follows the *GEM distribution*, $(\pi_1, \pi_2, \dots) \sim \text{GEM}(\alpha)$.[1] It is not hard to convince yourself that this sequence converges to 1 almost surely and that overall, π_k's would be of a decreasing trend.[2] Figure A.2 shows two examples of the stick-breaking process with different α's.

We're half way done with the construction. Now, suppose we have a distribution, G_0, on some preferred space. Then, we sample $\phi_k \sim G_0$, $k = 1, 2, \dots$, and we match each ϕ_k to the corresponding π_k such that for each k, we have a delta function[3] at each ϕ_k weighted by π_k. This is the Dirichlet process. It has two parameters: $\alpha \in \mathbb{R}$ is called the *concentration parameter* and G_0 is called the *base measure*. And for each draw out of the Dirichlet process, i.e. $G \sim \text{DP}(\alpha, G_0)$, it is the infinite mixture of the weights and delta functions.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \quad (\pi_k)_{k=1}^{\infty} \sim \text{GEM}(\alpha), \quad \phi_k \sim G_0$$

Figure A.3 gives an example of such construction. There are a few observations. First, note that G is a distribution (we'll deal with this more in the next section), and it is almost surely discrete. The atoms

[1] This distribution was originally used by Griffiths, Engen, and McCloskey (hence the name GEM) for modeling genetic processes. Accidentally, they found that it can be used to construct Dirichlet processes and subsequently furthered the field of Bayesian nonparametrics. cite

[2] The concept of almost sure convergence is seldomly mentioned in this report, so we'll settle on intuition here. If a sequence of random variable X_n converges to X almost surely, then the probability of all events in X_n converges to X with the exception of the set of events with zero probability.

[3] The delta here is the dirac-delta function, $\delta_a(x)$ where $\delta_a(x) = 0$ for all $x \neq a$ and $\int_{\mathbb{R}} \delta_a(x)\, dx = 1$. This can be extended to spaces other than the reals, and we will provide a more detailed description of that later.

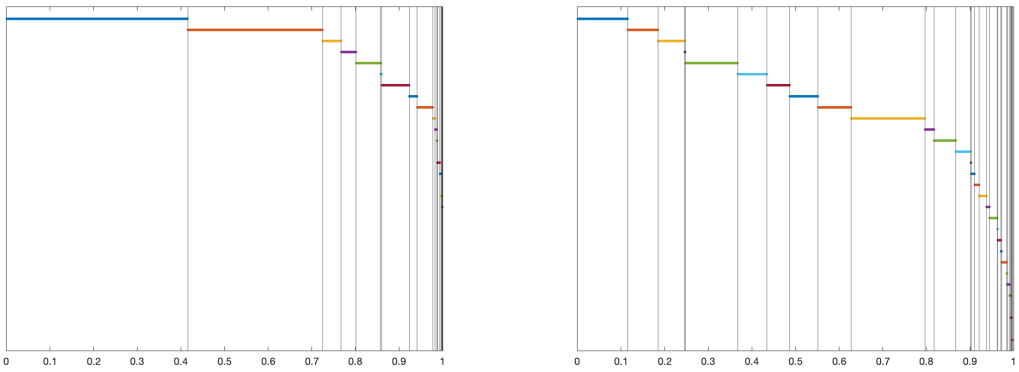

Figure A.2: A visual representation of the stick breaking process with (left) $\alpha = 3$, (right) $\alpha = 6$. One can see that increasing the concentration parameters tends to produce shorter sticks.

will concentrate around the regions where G_0 has the most mass. However, the Dirichlet process is not a discretized version of G_0 because the weight of each atom is not related to G_0.

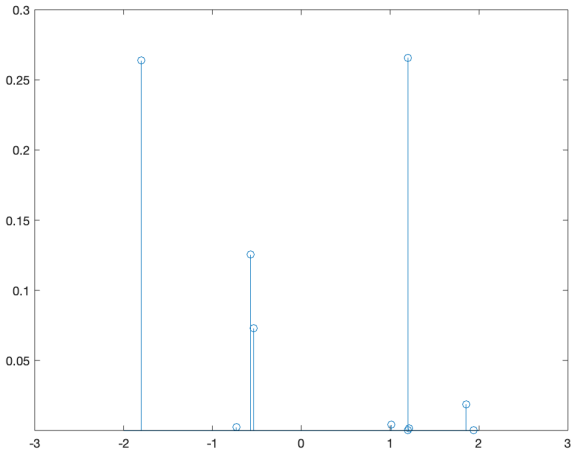

Figure A.3: A draw from the Dirichlet process. Each spike is a dirac delta (point mass) with the base distribution as a normal distribution. Theoretically, there are infinitely many spikes but due to numerical issues, only a portion can be shown.

Recall the observations on Beta distributions earlier this subsection, it is clear why α is the *concentration* parameter. If α is small, then we would have a few large weights that dominate G. If α is small, we have a much more uniform distribution of weights. This can turn out to be important in problems like clustering, where the number of clusters is unknown. By varying the concentration parameter, the Dirichlet process can express a varying number of "visible" clusters while still modeling infinite "invisible" clusters.

A.2.2 Chinese Restaurant Process and Polya Urn

We will digress onto a different, but related stochastic process: the Chinese Restaurant process (CRP). We will come back to the relationship between the CRP and the Stick Breaking construction in the next section. The CRP interpretation would prove to be invaluable in applications such as clustering, where some information from the Stick Breaking construction is not of importance. In essence, the CRP is the marginal distribution of the Stick Breaking process.

Imagine that it is five o'clock on a Saturday evening in the Boston China town and customers start flooding in one by one. The first customer goes in and takes a seat at table 1. The second customer then enters and now he has an option: he can sit with customer 1 or he can start a new table. Suppose he starts a new table with probability 1/2 and joins customer 1 otherwise. Suppose that customer 2 joins customer 1. Then, the third customer arrives. And since people who go to Chinese restaurants are talkative, he is more likely to join a pre-occupied and lively table than starting a new table. So, the third customer joins the first table with probability proportional to the number of people already sitting there, 2/3, and starts a new table otherwise. This pattern repeats as an infinite number of customers flood into the infinitely large restaurant. This is the *Chinese Restaurant process*.

Let's describe the CRP more formally. Let's choose a fixed $\alpha \in \mathbb{N}$. Let π_n denote a partition of the set $\{1, 2, \ldots, n\}$. Each element of π_n will denote a table, and the elements inside each table denotes the customer sitting at that table. For example, one instance of π_6 can be $\{\{1, 3, 5\}, \{2\}, \{4, 6\}\}$. Given π_n, the $(n+1)$-th customer then obeys the following rule:

$$P(\text{customer n+1 joins table c}) = \frac{|c|}{\alpha + n}$$
$$P(\text{customer n+1 starts a new table}) = \frac{\alpha}{\alpha + n}$$

where $|c|$ denotes the cardinality of the set, or the number of people in the table. The example in the previous paragraph was done with $\alpha = 1$. One can see that the greater the α, the more inclined a customer is to start a new table. Then, we can ask "what's the probability of observing a particular partition?" We can do this by filling in the customers one by one, and we will get the following expression:

$$P(\pi_n) = \frac{\alpha^n}{\alpha^{(n)}} \prod_{c \in \pi_n} (|c| - 1)!$$

where $\alpha^{(n)} = \alpha \cdot (\alpha + 1) \cdots (n)$ is the ascending factorial. We see that order does not matter! This would turn out to be important when relating the CRP to the Stick breaking distribution.

One other popular analogy for the CRP is the Polya urn, a classic example in probability theory. There are many different Polya urns, but we will focus on one particular model. Consider an urn containing balls of possibly different colors, and there are always α many magic balls. At each iteration, we draw a ball out of the urn. If it is not one of the magic balls, we put an extra ball of the same color in. If we got lucky and draw the magic ball, then we will put in a brand-new colored ball while putting the magic ball back. If we start off with just the magic balls in the urn, one can see that this is identical to the CRP. The only difference here is that each table is now distinguishable from one another via color. However, the probability of seeing any configuration of balls in the urn is the same and is still invariant under reordering. It turns out that this is the preferable interpretation in the upcoming subsection.

A.2.3 Formal Definition

The stick-breaking gives a set of instructions for sampling from a Dirichlet process, and in a sense, if we can sample from it, we know the distribution. However, we still would like to build the formal definition of a Dirichlet Process formally. The following remark gives a high-level introduction to various aspects of probability from a measure-theoretic point of view.

Remark (Probability, Measure, and Integration)**.** *We start by giving a formal definition of a probability space. This will serve as the foundation of the upcoming formalism.*

Definition (Probability Space)**.** *A Probability space is defined by a tuple $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is a set, $\mathcal{A}$ is the σ-algebra on the set Ω, and $\mathbb{P}$ is a probability measure $\mathbb{P} : \mathcal{A} \to [0, 1]$.*

The set Ω is often called the sample space. One can think of the σ-Algebra as the set of subsets of Ω. And $\mathbb{P}$ is the function that maps elements in $\mathcal{A}$ to a probability, some value in $[0, 1]$. For example, the dirac measure, $\delta_x : \mathcal{A} \to \mathbb{R}$, where $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. This is the same δ that we saw in the Stick Breaking construction. To gain more intuition, let's formally define a random variable, too.

Definition (Random Variable)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ be another measurable space. Then, a function $X : \Omega \to E$ is a random variable if for every $A \in \mathcal{E}$, $X^{-1} = \{\omega \in \Omega : X(\omega) \in A\}$ is an event in $\mathcal{A}$.*

Viewing random variables as a map from one measurable space to another gives flexibility in defining a random variable. Take a roll of a die, for example, we can have the number of the side facing up as the random variable, or the parity of the side facing up. However, a lot of the time, we just have these two sets as the same. Combining the previous concepts together, let X be a random variable following the Poisson distribution, E is the set of natural numbers (including 0); $\mathcal{A}$ would then be the set of all subsets of the natural numbers, so the set $\{1, 3, 5, \dots\}$ would be one of such sets; the probability measure $\mathbb{P}$ on a set $A \in \mathcal{A}$ would be the following sum

$$\mathbb{P}(A) = \mathbb{P}\{\omega : X(\omega) \in A\} = \sum_{k \in A} \frac{e^{-\lambda}\lambda^k}{k!}$$

Let's take the Gaussian distribution as an example for the continuous case. Our sample space is now the real line, $\mathbb{R}$. The σ-algebra now the set of all intervals in the reals. And the probability measure is then the integral over the corresponding set.

$$\mathbb{P}(A) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(\frac{-1}{2\sigma^2}(x-\mu)^2\big)\ dx$$

We will go over integration with respect to the probability measure by considering examples and settle on building an intuition rather than a rigorous foundation of Lebesgue integration. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be some probability space, and define a random variable X on that space. Then, we would write the expectation of that random variable as the following

$$\mathbb{E}(X) = \int_\Omega X(\omega)\ \mathbb{P}(d\omega)$$

This looks a lot like the usual expectation: it is a weighted sum of events, $X(\omega)$, by the probability (the measure) of that event occurring, $\mathbb{P}(d\omega)$. Reconsidering the Gaussian example. Since our sample space is the real line, a little piece of the sample space, $d\omega$, is now a little piece of the real line, which we'll denote by the familiar dx. Then, measuring this dx is then the length of dx times the density, i.e. $\mathbb{P}(dx) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(\frac{-1}{2\sigma^2}(x-\mu)^2\big)\ dx$. So, we recover the familiar integral for the expectation of a Gaussian.

$$\mathbb{E}(X) = \int_{\mathbb{R}} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\big(\frac{-1}{2\sigma^2}(x-\mu)^2\big)\ dx$$

Now, we're ready to define the Dirichlet Process formally.

Definition (Dirichlet Process)**.** *Let $(\Omega, \mathcal{A})$ be a measurable space, and G_0 be a probability measure on that space. Then, the Dirichlet Process has is a stochastic process that has parameters α and G_0, and for any finite partition $A_1, A_2, \dots, A_k$ of Ω,*

$$\big(\mathrm{DP}(A_1), \mathrm{DP}(A_2), \dots, \mathrm{DP}(A_k)\big) \sim \mathrm{Dirichlet}\big(\alpha G_0(A_1), \alpha G_0(A_2), \dots, \alpha G_0(A_k)\big)$$

In plain words, a draw from a Dirichlet process is a distribution! More specifically, a distribution when you applied onto a finite partition, it looks like a Dirichlet distribution. This is what it's meant by the Dirichlet process being a *random measure*: it is a distribution on the space of measures.

However, to formally define the notion of a random measure, we need to put it into the context of *stochastic processes.*

Definition (Stochastic Process)**.** *A stochastic process is a collection of random variables, indexed by some set T, on the same probability space.*

For example, a Poisson process is a counting process, $N(t)$, characterized by the number of arrivals in time $[0, t)$ where interarrival times are distributed exponentially with mean $1/\lambda$. Here, the random variable is $N(t)$, the count of arrivals, and it is indexed by $t \in \mathbb{R}$, which is often interpreted as time. Each $N(t)$ is a random variable following the Poisson distribution with parameter λt.

Lemma. *Let G be a Dirichlet process defined as follow*

$$\pi \sim \text{GEM}(\alpha), \quad \phi_k \sim G_0, \quad G = \sum_{k=0}^{\infty} \pi_k \delta_{\phi_k} \sim \text{DP}(\alpha, G_0)$$

on some measurable space E. Then, G is a stochastic process with the index set as the element of the σ-algebra, $\mathcal{E}$, i.e. $\{G(A); A \in \mathcal{E}\}$ is a stochastic process.

So, we should think of random measures as a special stochastic process whose index set is the subsets of the underlying space. We can check that for any fixed A, the Dirichlet process is indeed a measure.

$$G(A) = \sum_k \pi_k \delta_{\phi_k}(A) = \sum_{k:\phi_k \in A} \pi_k$$

And since π_k's sum up to 1 almost surely and is non-negative, G is indeed a probability measure. And since for each draw from the Dirichlet process, π's and ϕ 's would be different, $G(A)$ is a random variable. So, intuitively, G is random and it's a measure, hence, a random measure.

One important consequence of the stochastic process view is the *De Finetti's Theorem*. It is sometimes called the "Fundamental Theorem of Bayesian Analysis"; this illustrates the importance of the theorem. A lot of times in doing Bayesian statistics, we assume exchangeability – that the order of the random variables doesn't affect the distribution – rather than iid to provide a much richer model. For example, if we throw a bag of coins and observe the sequence of 0s and 1s, assuming that all the coins are iid is oversimplifying the problem by a bit too much: we don't know if all coins are identical! However, it is perfectly fine to assume that the coins are exchangeable – they might not be the same, I just don't know the order! So, De Finetti's theorem states the following.

Theroem (De Finetti's Theorem)**.** *Let $X_1, X_2, \ldots$ be an infinite sequence of random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. Then, X_i's are exchangeable, i.e.*

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \mathbb{P}(X_{\sigma(1)} \in A_1, X_{\sigma(2)} \in A_2, \ldots, X_{\sigma(n)} \in A_n)$$

for any permutation σ if and only if

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \int \prod_{i=1}^{n} G(A_i)\ \mu(dG)$$

holds for some probability measure μ.

To put it in terms of parametric inference, De Finetti's theorem would be saying that: if we assume exchangeablity, then

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta)\ \mu(d\theta)$$

So, we have 1) a distribution over the parameters, 2) the data must be conditionally independent given a specific parameter. The theorem stated gives a more general version of this: if we assume exchangeability, then 1) there is a probability over distributions, a stochastic process, that we can integrate over (one can think of this as a weighted sum), and 2) for any fixed draw G from the stochastic process, the X_i's are conditionally independent of each other. In this more general notion, we got rid of the parametric assumption for μ, and hence the term *Bayesian Nonparametric*.

We can use De Finetti's theorem to relate the CRP with the Stick Breaking construction, which is the Dirichlet process. We recall that the partitions, or more accurately the distribution of colored balls in the urn, is exchangeable; the label of the color doesn't really matter. By De Finetti's theorem, we then know that there is a stochastic process for which we can draw a distribution G from (this would be a distribution on the color space), and by sampling repeated from G, we recover the Polya urn distribution of colors. David Blackwell showed that the stochastic process that accomplished this is the Dirichlet process.

A.3 Application: Density Estimation

In this section, we'll go through an example of estimating the density of a mixture model. This problem has been thoroughly studied and has great practical significance. We will be focusing on particularly its use for clustering and draw a connection between the assignment of data to clusters with the Chinese Restaurant process.

A.3.1 Finite Mixture Model

A mixture model, as the name suggests, is a probability distribution comprising of a mix of different distributions. More formally, let $p_k(x)$ be K different distributions and π_k be the *weight vector* satisfying $\sum_k \pi_k = 1$. Then, the mixture distribution is then a linear combination of the p_k's and the π_k's, i.e.

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$$

For example, we can have $p_k(x)$ to be Gaussians with different mean and variances. This is called the *Gaussian Mixture model*. But of course, these can be done with different distributions altogether as well.

We can see that if choose $p_k(x)$ to be some parameterized distribution with parameter θ_k's, the problem is simple. One can specify a prior on each quantity. For example, π might follow Dirichlet, θ can be chosen to be the conjugate distribution, and K can be some distribution on the natural numbers. Then, all that's left is to "turn on the Bayesian crank".[4] In particular, Gibbs sampling seems like a good choice for finding the parameters. If we just want the *Maximum a Posteriori(MAP)* estimate – the point that maximizes the posterior likelihood – of θ_k's, the *Expectation Maximization (EM) algorithm* is also an adequate choice.

A.3.2 Dirichlet Process Mixture Model

It turns out that specifying a prior on K is not a particularly pleasant task. Michael Escobar, who was a student at the time, though: why don't we take it to be infinite mixtures? So now, the density is a series of functions.

$$p(x) = \sum_{k=1}^{\infty} \pi_k p(x; \theta_k)$$

This looks quite daunting at first sight because there are a lot of parameters (in fact, infinitely many of them). However, we can construct a prior using the Dirichlet process. Let $x_i, i = 1, 2, \ldots, n$ be the data points. Then, we can write our probability model as the following.

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_1, \ldots, \theta_n | G &\sim G \\ x_i | \theta_i &\sim p(x; \theta_i) \end{aligned}$$

Let's strengthen our intuition a bit. Remember that the Stick Breaking construction of Dirichlet processes says that a draw from the Dirichlet process is a discrete distribution, or more specifically, an infinite sum of weighted delta distributions. Then, we say that with probability proportional to the weights, I assign my data point to group k and take on distribution $p(x; \theta_k)$. We do that for every data point.

Generating data, though usually the focus of the problem, can be done. The difficultly comes at we cannot just sample an infinite mixture with a large, but finite one because it fundamentally changes the object. The method that should be used is sampling on demand. To draw from the Dirichlet process, an infinite object, we do the regular inverse sampling procedure but do the stick-breaking on demand. That is, draw a number uniformly within $[0, 1]$, then keep breaking the stick only so much so that we get the weight we want. Then, we draw the corresponding θ if it has not been drawn before. This way, we respect the infinite object and we will succeed with probability 1.[5]

[4] As said by professor Larry Wasserman at Carnegie Mellon University.

[5] The failure case would be if the uniform draw turns out to be 1. Thankfully, that almost surely doesn't happen.

A.3.3 Just the Clusters

We are going to focus on a particular problem of *clustering*. Intuitively, we try to put the data points into groups that "make sense."[6] In this case of a Dirichlet mixture, we have a natural cluster assignment: data points with the same θ's are in the same cluster! Let's set up a random variable z_i representing the cluster that x_i is in.

Now, we will try to devise an algorithm that does clustering using the Dirichlet process prior. By De Finetti's theorem, we know that drawing the θ's is exactly the Polya Urn/Chinese Restaurant scheme. So, let's interpret clustering in terms of the CRP. Imagine the data points as customers and each cluster as a table. So, we can devise a Gibbs sampler for finding the cluster assignments. Let $x_{i-,k} = \{x_j : j \neq i, z_j = k\}$, the set of data points in cluster k that is not the i-th data point. Also, since all data in the same cluster share a θ, let's denote θ_{z_k} denote the shared parameter for each cluster. The pseudo-code is described in Algorithm 1.

Algorithm 1: Gibbs Sampling for Cluster Assignment

```
Assign all customer to table 1
while not converged do
    for each customer i ∈ {1, 2, ..., n} do
        for each non-empty table k do
            compute p_k = (|table k| / (α+n-1)) ∫ p(x_i|θ_{z_k}, x_{i-,k}) p(θ_{z_k}|x_{i-,k}) dθ
        Compute p* = (α/(α+n-1)) ∫ p(x_i|θ) G(dθ)
        Assign customer i to table k/new table with probability proportional to p_k/p*
```

We know that the CRP is exchangeable. So, at each iteration, we take out customer i and pretend that he's the last customer to come into the restaurant. Then, we go about asking: given the current configuration, what is the probability that customer i joins table k or starts a new table? Of course, this is weighted by the likelihood of observing customer i at table k. Since the CRP update rule tends to put more people in already crowded tables, the probability of joining a table becomes a trade-off between the preferential attachment from CRP and the likelihood of observing data i in cluster k.

One trick that was used to speed up the convergence was by marginalizing out θ. If we were to use a regular Gibbs sampler, we need the full conditional $p(z_i = k|x, \theta)$. However, we conveniently marginalize out θ and find

$$\begin{aligned} p(z_i = k|x) &= p(z_i = k|x_i, x_{i-}) \\ &\propto p(x|z_i = k, x_{i-}) p(z_i = k|x_{i-}) \\ &= \int p(x_i|\theta_{z_k}, x_{i-,k}) p(\theta_{z_k}|x_{i-,k}) d\theta \cdot p(z_i = k) \end{aligned}$$

This is sometimes called *collapsed Gibbs*. Of course, we can include inference on θ into the model. Also, we've been assuming that α is fixed. We can also place a prior on that (typically a Gamma distribution) and inference on α.

We coded a toy example to demonstrate the capability and potential downfall of this model. We first generated the clusters using a finite mixture of normal distributions ($\mathcal{N}(\theta_k, 0.5)$ where $\theta_k \sim \mathcal{N}(0, 4)$. Then, we did inference on the cluster assignment using a Dirichlet process mixture model with the "correct" base and mixture distributions. Figure A.4 shows one outcome of this. The left image is the data with the clusters shown, and the right shows the clusters obtained. We can see that, when the clusters are far away enough, the CRP Gibbs sampler can learn the clusters without specifying the number of clusters in the first place. However, when there are clusters close by (like in Figure A.5, the CRP will have a hard time distinguishing the clusters. Changing the number of iterations, α value and prior guesses all had small effects on the outcome.

[6] There is no strict definition for how to cluster. Different metrics can yield different results, and which metric to use really depends on the situation. For an example, see spectral clustering.

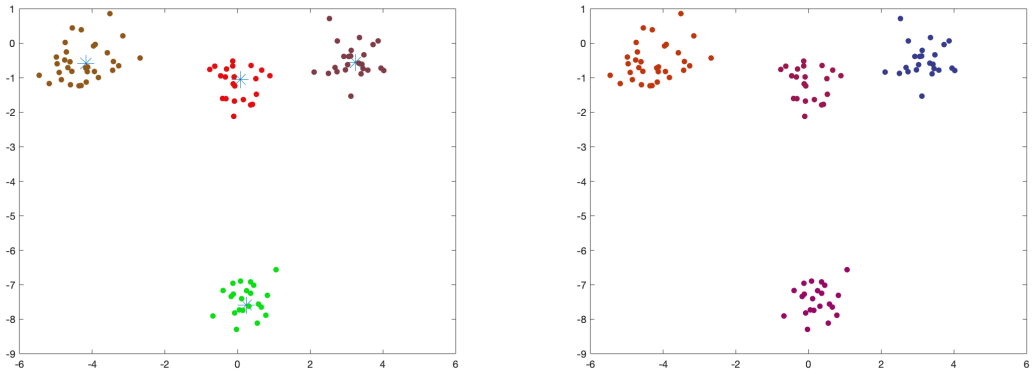

Figure A.4: Example showing Chinese Restaurant process for clustering. (Left) 100 data points were generated with a finite mixture model. (Right) The cluster assignment obtained by looking at just the data. We see that the clusters are correctly recovered if no significant overlapping occurred.

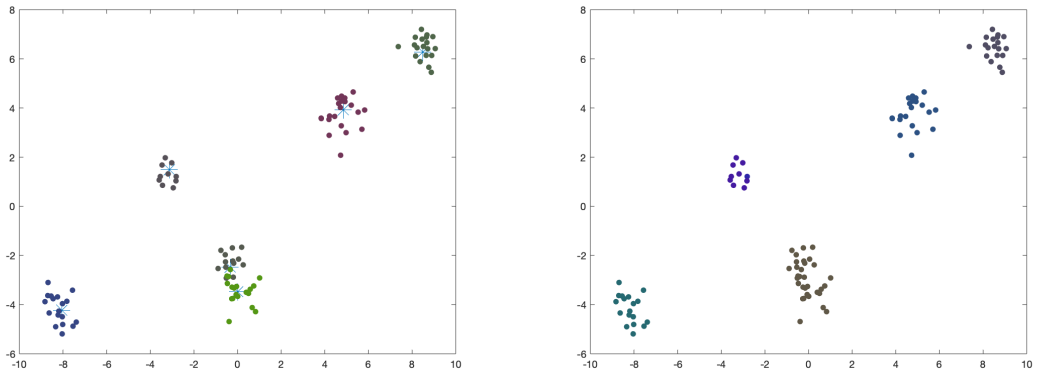

Figure A.5: Example showing Chinese Restaurant process for clustering. (Left) 100 data points were generated with a finite mixture model. (Right) The cluster assignment obtained by looking at just the data. We see that when some clusters are close together, the algorithm is unable to distinguish the two clusters.

A.4 Discussion

Even though Bayesian nonparametrics is a relatively young discipline (in comparison with the rest of statistics), the methods described are in no sense a comprehensive guide. One immediate variant is the *Pitman-Yor process*. It is basically the CRP with a "discount" parameter, σ.

$$P(\text{customer n+1 joins table c}) = \frac{|c| - \sigma}{\alpha + n}$$
$$P(\text{customer n+1 starts a new table}) = \frac{\alpha + \sigma|\pi_n|}{\alpha + n}$$

where $|\pi_n|$ is the number of occupied tables (cardinality of the set of subset partitions). The reason why this is useful is that the number of new tables discovered grows logarithmically to the number of customers. So, as the number of tables goes to infinity, the probability of a customer joining table a new table decreases exponentially. However, we find that this is often not realistic as many real-world distributions obey the power-law. By discounting the preferential attachment, the Pitman-Yor process gives us power-laws, which can be more desirable.

There are many other variants of models involving Dirichlet processes like Hierarchical Dirichlet process, mixtures of Dirichlet process (as opposed to a Dirichlet process for mixture models here), etc. One can also go another route. Notice how De Finetti's theorem related the CRP to the Dirichlet process. We can also go one level above the Dirichlet process into the world of *completely random measures (CRM)*. The CRM framework allows for developing many other similar types of stochastic processes like the Dirichlet process and can be adapted to different problems. Nonparametric bayesian regression also exists, but the common approach is to use a Gaussian process prior rather than a Dirichlet process for continuity (Dirichlet processes are almost surely discrete).

A.4.1 Frequentist vs. Bayesian Nonparametric

To fuel the hatred between Frequentists and Bayesians, we will discuss briefly the difference between nonparametric methods from either side. Just intuitively speaking, Bayesian nonparametric is less studied since it is relatively newer and is typically more work due to the whole stochastic process jargon. However, just like most Bayesian methods, the power of Bayesian nonparametric comes at the ability to specify a prior. When strong prior knowledge is present, the flexibility of nonparametric inference can make the bayesian approach worthwhile. Consider the problem of estimating a cumulative distribution function. The Frequentists typically use the *empircal distribution*.

Definition (Empirical Distribution)**.** *Given n data points X_i, the empirical distribution function $\hat{F}_n$ is the CDF described as follow.*

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)}{n}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The empirical distribution is great: it is unbiased, consistent estimator of the true CDF. Furthermore, the *DWK Inequality* gives a nice confidence set for the estimator.

Theroem (Dvoretzky-Kiefer-Wolfowitz Inequality)**.** *Let $X_1, \ldots, X_n \sim F$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\big(\sup_x \left|F(x) - \hat{F}(x)\right| > \epsilon\big) \leq 2e^{-2n\epsilon^2}$$

On the other hand, for the Bayesians, we can build the following model using the Dirichlet process.

$$F \sim \mathrm{DP}(\alpha, F_0), \quad X_1, \ldots, X_n | F \sim F$$

The posterior distribution of F turns out to be quite intuitive.

$$F|X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + n, \bar{F}), \quad \bar{F} = \frac{n}{\alpha + n}\hat{F}_n + \frac{\alpha}{\alpha + n}F_0$$

So, we see that the posterior distribution is just a convex combination of the empirical distribution and the original guess (base distribution). Then, since we can draw from the Dirichlet process quite easily, we can numerical simulate a confidence set by sampling from the posterior distribution many times.

So, if you're not a hardcore frequentist and you happen to be stuck in a situation with a small amount of data and you have reasons to believe that F_0 is indeed a good guess for the true distribution, nonparametric Bayes seems like a nice option. However, if one or more of the conditions above doesn't hold, the frequentist approach is perhaps safer. This is especially true when the notion of a Jeffery's prior doesn't really exist for F_0.

A.5 Sources and Code

Most of the information was learned from talks/lectures available on Youtube by Tamara Broderick at MIT, Michael I. Jordan at UC Berkeley, and Larry Wasserman at Carnegie Mellon University.[7,8,9] Information on measure-theoretic probability is from *Probability and Stochastics* by Erhan Cinlar [1]. The Gibbs sampling algorithm was partially from a lecture note from Stanford University, partially from Ryan Murphy's *Machine Learning: A Probabilistic Perspective* [5].[10] Frequentists methods are from the book *All of Statistics*, again, by professor Wasserman [6].

All code is made available on **Github** as `stick_breaking_demo.m` and `CRP_demo.m`.[11]

[7] Professor Broderick's talk: `https://www.youtube.com/watch?v=I7bgrZjoRhM`
[8] Professor Jordan's talk: `https://www.youtube.com/watch?v=yfLoxwjCGNY`
[9] Professor Wasserman's lecture: `https://www.youtube.com/watch?v=l0HQXtqnBZY`
[10] Lecture Notes: `http://web.stanford.edu/class/stats362/lec1.pdf`
[11] Link to repository: https://github.com/dannychen0830/STAT-448

Bibliography

[1] Erhan Çınlar. *Probability and stochastics*. Vol. 261. Springer Science & Business Media, 2011.

[2] G Bard Ermentrout and David H Terman. *Mathematical foundations of neuroscience*. Vol. 35. Springer Science & Business Media, 2010.

[3] Robert G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013. DOI: 10.1017/CBO9781139626514.

[4] Robert Kass, Valérie Ventura, and Emery Brown. "Statistical issues in the analysis of neuronal data. Journal of Neurophysiology, 94, 8-25". In: *Journal of neurophysiology* 94 (Aug. 2005), pp. 8–25. DOI: 10.1152/jn.00648.2004.

[5] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[6] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.