Stochastic Processes, Kernel Regression, Infinite Mixture Models

Gabriel Huang

(TA for Simon Lacoste-Julien)

IFT 6269 : Probabilistic Graphical Models - Fall 2018
Stochastic Process
= 
Random Function
Today

• Motivate **Gaussian** and **Dirichlet** distribution in Bayesian Framework.
• Kolmogorov’s extension theorem.
• Define **Gaussian Process** and **Dirichlet Process** from finite-dimensional marginals.
• Gaussian Process:
  • Motivating Applications: Kriging, Hyperparameter optimization.
  • Properties: Conditioning/Posterior distribution.
  • Demo.
• Dirichlet Process:
  • Motivating Application: Clustering with unknown number of clusters.
  • Construction: stick-breaking, Polya urn, Chinese Restaurant Process.
  • De Finetti theorem.
  • How to use.
  • Demo.
Disclaimer

I will be skipping the more theoretical building blocks of stochastic processes (e.g. measure theory) in order to be able to cover more material.
Recall some distributions

Gaussian Distribution
Samples $X$ in $\mathbb{R}^d$.

Dirichlet Distribution
Samples $\pi$ in Simplex $\Delta_{d-1}$ verifies $\pi_1 + \cdots + \pi_d = 1$
Why Gaussian and Dirichlet?  

*They are often used as priors*
Bayesians like to use those distributions as priors over model parameters $p(x)$

Why?
Because they are very convenient to represent/update.

Conjugate Priors
Conjugate Prior means:
Posterior in same family as prior

\[ p(\theta|x) \propto p(x|\theta)p(\theta) \]
\[ p(\theta|x) \propto p(x|\theta)p(\theta) \]

\( \theta|x \sim \text{Gaussian}(\mu', \Sigma') \)

\( x|\theta \sim \text{Gaussian}(\theta, \Sigma_L) \)

Gaussian is conjugate prior for Gaussian likelihood model.
\[ p(\theta|x) \propto p(x|\theta)p(\theta) \]

\[ \theta|x \sim \text{Dirichlet}(\alpha') \]

\[ x|\theta \sim \text{Categorical}(\theta) \]

\textbf{Dirichlet} is conjugate prior for \textbf{Categorical/Multinoulli} likelihood model.
So taking the posterior is simply a matter of updating the parameters of the prior.
Back to Gaussian and Dirichlet

Gaussian Distribution

Samples $X$ in $R^d$.

Dirichlet Distribution

Samples $\pi$ in Simplex $\Delta_{d-1}$ verifies $\pi_1 + \cdots + \pi_d = 1$
Gaussian and Dirichlet are indexed with a finite set of integers \( \{1, \ldots, d\} \).

They are random vectors.

\[
(X_1, X_2, \ldots, X_d) \\
(\pi_1, \pi_2, \ldots, \pi_d)
\]
Can we index random variables with infinite sets as well?

In other words, define random functions.
Defining stochastic processes from their marginals.
Suppose we want to define a random function (stochastic process)

$$X: t \in T \rightarrow \mathbb{R},$$

where $T$ is an infinite set of indices.

Imagine a joint distribution over all the $(X_t)$. 
Kolmogorov Extension Theorem

informal statement

Assume that for any $k \geq 1$, and every finite subset of indices $(t_1, t_2, \ldots, t_k)$, we can define a marginal probability (finite-dimensional distribution)

$$p_{t_1,t_2,\ldots,t_k}(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$$

Then, if all marginal probabilities agree, there exists a unique stochastic process $X: t \in T \rightarrow \mathbb{R}$ which satisfies the given marginals.
So Kolmogorov’s extension theorem gives us a way to implicitly define stochastic processes. (However it does not tell us how to construct them.)
Defining Gaussian Process from finite-dimensional marginals.
Characterizing Gaussian Process

Samples $X \sim GP(\mu, \Sigma)$ of a Gaussian Process are **random functions**

$$X : T \rightarrow \mathbb{R}$$

defined on the domain $T$ (*such as time* $T = \mathbb{R}$, or *vectors* $T = \mathbb{R}^d$).

We can also see them as an infinite collection $(X_t)_{t \in T}$ indexed by $T$.

Parameters are the **Mean** function $\mu(t)$ and **Covariance** function $\Sigma(t, t')$. 
For any $t_1, t_2, \ldots, t_k \in T$ we define the following finite-dimensional distributions $p(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$.

$$X_{t_1}, X_{t_2}, \ldots, X_{t_k} \sim \mathcal{N}((\mu(t_i)_i, (\Sigma(t_i, t_j))_{i,j})$$

Since they are consistent with each other, Kolmogorov’s extension theorem states that they define a unique stochastic process, we will call **Gaussian Process:**

$$X \sim GP(\mu, \Sigma)$$
Characterizing Gaussian Process

Some properties are immediate consequences of definition:

• \( E[X_t] = \mu(t) \)

• \( \text{Cov}(X_t, X_{t'}) = E[(X_t - \mu(t))(X_{t'} - \mu(t'))^T] = \Sigma(t, t') \)

• Any linear combination of distinct dimensions is still a Gaussian:

\[
\sum_{i=1}^{k} \alpha_i \ast X_{t_i} \sim \mathcal{N}(\cdot, \cdot)
\]
Characterizing Gaussian Process

Some properties are immediate consequences of definition:

• **Stationarity**: $\Sigma(t, t') = \Sigma(t - t')$
  
  *does not depend on the positions*

• **Continuity**: $\lim_{t' \to t} \Sigma(t, t') = \Sigma(t, t)$

• Any linear combination is still a Gaussian:

\[
\sum_{i=1}^{k} \alpha_i * X_{t_i} \sim \mathcal{N} (\cdot; \cdot)
\]
Example Samples
Posteriors of Gaussian Process.

*How to use them for regression?*
Interactive Demo

need a volunteer

http://chifeng.scripts.mit.edu/stuff/gp-demo/
Gaussian processes are very useful for doing regression on an unknown function $f$:

$$y = f(x).$$

Say we don’t know anything about that function, except the fact that it is smooth.
Before observing any data, we represent our belief on the unknown function $f$ with the following prior:

$$ f \sim GP(\mu(x), \Sigma(x, x')) $$

For instance $\mu(x) = 0$ and $\Sigma(x, x') = C \cdot \exp\left(-\frac{||x-x'||^2}{\sigma^2}\right)$

**WARNING:** Change of notation!

$x$ is now the index and $f(x)$ is the random function
Now, assume we observe a training set

\[ X_n = (x_1, x_2, ..., x_n), y_n = (y_1, y_2, ..., y_n) \]

and we want to predict the value \( y^* = f(x^*) \)
associated with a new test point \( x^* \).

One way to do that is to compute the posterior \( f|X_n, y_n \) after observing the evidence (training set).
Bayes’ Rule

\[ p(f | X_n, y_n) \propto p(y_n | f, X_n) \ p(f) \]

- **Gaussian Process Prior:**
  \[ p(f) = GP(\mu(x), \Sigma(x, x')) \]

- **Gaussian Likelihood:**
  \[ p(y_n | f, X_n) = \mathcal{N}(f(X), \epsilon^2 I_n) \]

- **\( \rightarrow \) Gaussian Process Posterior:**
  \[ p(f | X_n, y_n) = GP(\mu'(x), \Sigma'(x, x')) \]
  for some \( \mu'(x), \Sigma'(x, x') \).

**Remember:** Gaussian Process is conjugate prior for Gaussian likelihood model.
Bayes’ Rule
\[ p(f|X_n, y_n) \propto p(y_n|f, X_n) \ p(f) \]

- **Gaussian Process Prior:**
  \[ p(f) = \text{GP}(\mu(x), \Sigma(x, x')) \]

- **Dirac Likelihood:** \((\epsilon \to 0)\)
  \[ p(y_n|f, X_n) = \delta(y_n - f(X_n)) \]
  that is, \(y_n\) is now deterministic after observing \(f, X_n\).
  \[ y_n = f(X_n) \]

- **→ Gaussian Process Posterior:**
  \[ p(f|X_n, y_n) = \text{GP}(\mu'(x), \Sigma'(x, x')) \]
  for some \(\mu'(x), \Sigma'(x, x')\).
The problem is there is no easy way to represent the parameters of the posterior $\mu(x), \Sigma(x, x')$ efficiently.

Instead of computing the full posterior $f$, we will just evaluate the posterior at one point $y^* = f(x^*)$.

We want: $p(y^*|X_n, y_n, x^*)$
We want: \( p(y^*|X_n, y_n, x^*) \)

The finite-dimensional marginals of the Gaussian process give that:

\[
\begin{bmatrix}
y_n \\
y^* \\
x_n \\
x^*
\end{bmatrix} \sim \mathcal{N}
\begin{pmatrix}
\mu(X_n) \\
\mu(x^*) \\
\Sigma(X_n, X_n) \\
\Sigma(x^*, x^*)
\end{pmatrix}
\]
Theorem: For a Gaussian vector with distribution

\[
\begin{bmatrix}
X_1 \\
X_2
\end{bmatrix}
\sim \mathcal{N}
\left(\begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix},
\begin{bmatrix}
\Sigma_{1,1} & \Sigma_{1,2} \\
\Sigma_{2,1} & \Sigma_{2,2}
\end{bmatrix}\right)
\]

the conditional distribution \( X_2 | X_1 \) is given by

\[
\begin{bmatrix}
X_2 \\
X_1
\end{bmatrix}
\sim \mathcal{N}
\left(\begin{bmatrix}
\mu_2 + \Sigma_{2,1} \Sigma_{1,1}^{-1} (x_1 - \mu_1) \\
\mu_2 + \Sigma_{2,1} \Sigma_{1,1}^{-1} (x_1 - \mu_1)
\end{bmatrix},
\begin{bmatrix}
\Sigma_{2,2} - \Sigma_{2,1} \Sigma_{1,1}^{-1} \Sigma_{1,2} & \Sigma_{2,1} \Sigma_{1,1}^{-1} \\
\Sigma_{2,1} \Sigma_{1,1}^{-1} & \Sigma_{1,1}^{-1}
\end{bmatrix}\right)
\]

This Theorem will be useful for the
Kalman filter, later on ... [Schur’s complement]
Applying the previous theorem gives us the posterior $y^*$.

\[
\mu' = \mu(x^*) + \Sigma(x^*, X_n)\Sigma(X_n, X_n)^{-1}(X_n - \mu(X_n))
\]

\[
\begin{pmatrix}
\begin{array}{c}
y^* \\
x^*
\end{array}
\end{pmatrix}
\sim 
\mathcal{N}
\left(
\begin{pmatrix}
\mu' \\
\Sigma
\end{pmatrix}
\right)
\]

\[
\Sigma' = \Sigma(x^*, x^*) - \Sigma(x^*, X_n)\Sigma(X_n, X_n)^{-1}\Sigma(X_n, x^*)
\]
Active Learning with Gaussian Process.
Active Learning

Active Learning is an iterative process:
• Generate a question $x^*$.
• Query the world with the question (by acting, can be costly)
• Obtain an answer $y^* = f(x^*)$.
• Improve model by learning from the answer.
• Repeat.
Active Learning

Gaussian process is good for cases where it is expensive to evaluate $y^* = f(x^*)$.

- **Kriging.** $y^*$ is the amount of natural resource, $x^*$ is new 2D/3D location to dig. Every evaluation is mining and can cost millions.

- **Hyperparameter** optimization (Bayesian optimization). $y^*$ is the validation loss, $x^*$ is set of hyperparameters to test. Every evaluation is running an experiment and can take hours.
Back to the demo

(Talk about utility function)

http://chifeng.scripts.mit.edu/stuff/gp-demo/
Formal equivalence with Kernelized Linear Regression.

[blackboard if time]

Rasmussen & Williams (2006)
http://www.gaussianprocess.org/gpml/chapters/RW2.pdf
Dirichlet Processes.

*Stick Breaking Construction*
\[ \pi = (\pi_1, \pi_2, \ldots) \sim GEM(\alpha) \]

scalar weights sum up to 1

\[ G = \sum_{k=1}^{+\infty} \pi_k \delta_{\theta_k} \]

\[ \theta_1, \theta_2, \ldots \sim iid \ G_0 \]

parameters, sampled from base distribution

G is a **random probability measure**:
- **random**: both \( \pi \) and \( \theta \) are random
- **probability measure**: it is a convex combination of Diracs, which are probability measures

Diracs concentrate probability mass \( \pi_k \) at \( \theta_k \)
Dirichlet Process

- Consider Gaussian $G_0$

- $G \sim \text{DP}(\alpha, G_0)$

Courtesy of Khalid El-Arini
Two independent samples $G$ from $DP(\alpha, G_0)$

Each sample $G$ is a probability distribution (e.g. over parameters) and can be written as a mixture of diracs.

$$G = \sum_{k=1}^{+\infty} \pi_k \delta_{\theta_k}$$
Measuring is counting $G(A) = \sum_{k=1}^{+\infty} \pi_k \ast 1\{\theta_k \in A\}$

For a fixed subset $A$, notice how $G(A)$ is random. In fact even the $\pi_k$ change value for each sample.
To generate a **finite** sequence of (mixture) weights \( \pi = \pi_1, \pi_2, \ldots, \pi_k \) that sum up to 1, we can use the **Dirichlet** distribution

\[
\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)
\]

How to generate an **infinite** sequence of (mixture) weights \( \pi = \pi_1, \pi_2, \ldots \) which sum up to 1? We can use **stick-breaking**

\[
\pi \sim \text{GEM}(1, \alpha)
\]
Beta Distribution

\[ \pi_1 \sim \text{Beta}(\alpha, \beta) \]
\[ \pi_2 = 1 - \pi_1 \]
\[ p(\pi_1 | \alpha, \beta) \propto \pi_1^{\alpha-1}(1 - \pi_1)^{\beta-1} \]

Equivalent to:

\[ \pi_1, \pi_2 \sim \text{Dirichlet}(\alpha, \beta) \]
\[ p(\pi_1, \pi_2 | \alpha, \beta) \propto \pi_1^{\alpha-1}\pi_2^{\beta-1} \]

\( \alpha, \beta \to +\infty \) gives peaked distribution around \( \alpha/(\alpha + \beta) \)
Stick Breaking $\pi \sim GEM(\alpha)$

\[ \pi_1 \quad 1 - \pi_1 \]
\[ \pi_1 \quad \pi_2 \quad 1 - \pi_1 - \pi_2 \]
\[ \pi_1 \quad \pi_2 \quad \pi_3 \]
\[ \vdots \]
\[ \pi_1 \quad \pi_2 \quad \pi_3 \quad \pi_4 \quad \ldots \]

$V_1 \sim Beta(1, \alpha)$ \quad $\pi_1 = V_1$

$V_2 \sim Beta(1, \alpha)$ \quad $\pi_2 = V_2(1 - \pi_1)$

$V_3 \sim Beta(1, \alpha)$ \quad $\pi_3 = V_2(1 - \pi_1 - \pi_2)$

Griffiths, Engen, McCloskey
Defining Dirichlet Process from finite-dimensional marginals.
Dirichlet Process

Samples $G \sim DP(\alpha, G_0)$ of a Dirichlet Process are themselves probability measures (i.e. distributions) over a measurable space $(\Omega, \mathcal{F})$.

$$G : \mathcal{F} \to \mathbb{R}_+$$

which associate a probability to every measurable subset $A \in \mathcal{F}$.

Note: $\mathcal{F}$ is the set of all measurable subsets $A \subseteq \Omega$.

Parameters are the **base probability distribution** $G_0$ (over $\Omega$) and the parameter $\alpha > 0$. 
Kolmogorov Consistency Construction

For any $k \geq 0$, consider any partition $A_1, A_2, \ldots, A_k$ of the space $\Omega$. We define the following finite-dimensional distributions

$$G(A_1), \ldots, G(A_k) \sim \text{Dirichlet}(\alpha \ast G_0(A_1), \ldots, \alpha \ast G_0(A_k))$$

Since they can be proved* to be consistent with each other, Kolmogorov’s extension theorem states that they define a unique stochastic process, we will call Dirichlet Process:

$$G \sim DP(\alpha, G_0)$$
Here $A_1, A_2, A_3$ is a partition of the parameter space $\Omega$. Assume $\alpha = 10, G_0 = \mathcal{N}(0, I_2)$. Draw two distributions $G_1, G_2 \sim \text{iid } DP(\alpha, G_0)$.

First sample

$G_1(A_1) = \pi_5 + \pi_6 + \pi_8$
$G_1(A_2) = \pi_2 + \pi_4$
$G_1(A_3) = \pi_1$

Second sample

$G_2(A_1) = \pi_3 + \pi_4 + \pi_5$
$G_2(A_2) = \pi_2$
$G_2(A_3) = \pi_1$

Probability masses for base distribution (deterministic)

$G_0(A_1) = \mathcal{N}(0, I_2)(A_1) = 0.8$
$G_0(A_2) = \mathcal{N}(0, I_2)(A_2) = 0.2$
$G_0(A_3) = \mathcal{N}(0, I_2)(A_3) = 0.2$

Then we have that

$G(A_1), G(A_2), G(A_3) \sim \text{Dirichlet}(8, 2, 2)$
All constructions match.

It can be shown that Stick-Breaking and Kolmogorov consistency definitions match.

Chinese Restaurant Process (CRP)

Infinity of Tables

\[ G_0 = \text{Uniform}\left(\{\text{Fish, Pork, Tofu}\}\right) \]
Chinese Restaurant Process (CRP)

• Customer 1 arrives.
• Takes any free table.
• Sample a dish $\theta_1 \sim G_0$-> Tofu
• $\text{state}=${\{1\}}, n=1 customers
Chinese Restaurant Process (CRP)

1. Customer 2 arrives.
2. \( P(\text{new table}) \propto \alpha \)
3. \( P(\text{table \{1\}}) \propto |\{1\}| = 1 \)
4. Decides to sit at \{1\}
5. Share dish: \( \theta_2 = \theta_1 = \text{Tofu} \)
6. \{\{1,2\}\}, n=2 customers
Chinese Restaurant Process (CRP)

Infinity of Tables

1. Tofu
2. 2
3. Pork

- Customer 3 arrives.
- \( P(\text{new table}) \propto \alpha \)
- \( P(\text{table \{1,2\}}) \propto |\{1,2\}| = 2 \)
- Decides to sit at new table
- Sample a dish \( \theta_3 \sim G_0 \rightarrow \text{Pork} \)
- \{\{1,2\},\{3\}\}, n=3 \text{ customers}
Chinese Restaurant Process (CRP)

Infinity of Tables

- Customer 4 arrives.
- $P(\text{new table}) \propto \alpha$
- $P(\text{table } \{1,2\}) \propto |\{1,2\}| = 2$
- $P(\text{table } \{3\}) \propto |\{3\}| = 1$
- Share dish, $\theta_4 = \theta_1 = \text{Tofu}$
- $\{\{1,2,4\},\{3\}\}$, $n=4$ customers
Chinese Restaurant Process (CRP)

- Customer 5 arrives.
- \( P(\text{new table}) \propto \alpha \)
- \( P(\text{table \{1,2,4\}}) \propto |\{1,2,4\}| = 3 \)
- \( P(\text{table \{3\}}) \propto |\{3\}| = 1 \)
- Pick new table
- Sample new dish \( \theta_5 = Fish \)
- \( \{\{1,2,4\},\{3\},\{5\}\}, \text{n=5 customers} \)
Chinese Restaurant Process (CRP)

- Customer 6 arrives.
- \( P(\text{new table}) \propto \alpha \)
- \( P(\text{table \{1,2,4\}}) \propto |\{1,2,4\}| = 3 \)
- \( P(\text{table \{3\}}) \propto |\{3\}| = 1 \)
- \( P(\text{table \{5\}}) \propto |\{5\}| = 1 \)
- Pick table \{1,2,4\}
- Share dish \( \theta_6 = \theta_1 = \text{Tofu} \)
- \( \{\{1,2,4,6\},\{3\},\{5\}\} \), n=6 customers
Chinese Restaurant Process (CRP)

We can look at the sequence of dishes

\[ \begin{align*}
\theta_1 &= \text{Tofu} \\
\theta_2 &= \text{Tofu} \\
\theta_3 &= \text{Pork} \\
\theta_4 &= \text{Tofu} \\
\theta_5 &= \text{Fish} \\
\theta_6 &= \text{Tofu}
\end{align*} \]

It can be shown that the distribution of \((\theta_t)_t\) is exchangeable. That is:

\[ p(\theta_1 = u_1, \theta_2 = u_2, \ldots) = p(\theta_1 = u_{\sigma(1)}, \theta_2 = u_{\sigma(2)}, \ldots) \]

The order in which the customers arrive is actually \textbf{not important}. 
De Finetti’s Theorem

• Suppose that we agree that if our data are reordered, it doesn’t matter
  – this is generally not an assertion of “independent and identically distributed”; rather, it is an assertion of “exchangeability”

• *Exchangeability*: the joint probability distribution underlying the data is invariant to permutation

**Theorem (De Finetti, 1935).** *If* \( (x_1, x_2, \ldots) \) *are infinitely exchangeable, then the joint probability* \( p(x_1, x_2, \ldots, x_N) \) *has a representation as a mixture:*

\[
p(x_1, x_2, \ldots, x_N) = \int \left( \prod_{i=1}^{N} p(x_i \mid \theta) \right) dP(\theta)
\]

*for some random variable* \( \theta \).

• I.e., if you assert exchangeability, it is reasonable to act as if there is an underlying parameter, there is a prior on that parameter, and the data are conditionally IID given that parameter.

M. I. Jordan NIPS 2017 Tutorial
De Finetti’s Theorem

Applied to the CRP, it means there exist a unique* random variable $G$, such that all $\theta$ become independent conditionally to $G$.

We can show that $G \sim DP(\alpha, G_0)$!

Here: $G_0 = \frac{1}{3} \delta_{fish} + \frac{1}{3} \delta_{pork} + \frac{1}{3} \delta_{tofu}$, $\alpha$ is the same ($\propto$ new table)

Let $\pi = (\pi_1, \pi_2, \ldots) \sim GEM(\alpha)$ stick-breaking.

Sample $\theta = (\theta_{k=1}, \theta_{k=2}, \ldots) \sim iid G_0$

Now we can form our random measure $G = \sum_{k=1}^{+\infty} \pi_k * \delta_{\theta_k}$.

And we sample $\theta_{i=1}, \theta_{i=2}, \ldots \sim iid G$

$\theta_i$ is the parameter for data point $i$ (customer $i$)

$\theta_k$ is the parameter for component $k$ (table $k$)

*Unique in distribution.
Blackwell-McQueen Urn
Polya Urn
Same process, different story.

Each dish is a set of unique ball colors.

Each customer is a successive draw.
Using Dirichlet Process for infinite mixture models.
Chinese Restaurant Process (CRP)

Infinity of Components

\[ G_0 = \mathcal{N}(\mu_0, \Sigma_0) \]
Chinese Restaurant Process (CRP)

\[ \theta_{k=1} = (1, -3) \]

Infinity of Tables

- Sample parameter for data point 1.
- Takes any free table.
- Sample a parameter \( \theta_1 \sim G_0 \)
- \text{state} = \{1\}, n=1 customers
Chinese Restaurant Process (CRP)

• Sample parameter for data point 2.
• \( P(\text{new table}) \propto \alpha \)
• \( P(\text{table } \{1\}) \propto |\{1\}| = 1 \)
• Decides to sit at \( \{1\} \)
• Share dish: \( \theta_2 = \theta_1 = \text{Tofu} \)
• \( \{\{1,2\}\}, \text{n=2 customers} \)

Infinity of Tables
Chinese Restaurant Process (CRP)

Infinity of Tables

- Sample parameter for data point 3.
- $P(\text{new table}) \propto \alpha$
- $P(\text{table } \{1,2\}) \propto |\{1,2\}| = 2$
- Decides to sit at new table
- Sample a dish $\theta_3 \sim G_0 \Rightarrow \text{Pork}$
- $\{\{1,2\},\{3\}\}$, n=3 customers
Chinese Restaurant Process (CRP)

Customer 4 arrives.

- $P(\text{new table}) \propto \alpha$
- $P(\text{table } \{1,2\}) \propto |\{1,2\}| = 2$
- $P(\text{table } \{3\}) \propto |\{3\}| = 1$
- Share dish, $\theta_4 = \theta_1 = \text{Tofu}$
- $\{\{1,2,4\},\{3\}\}, n=4 \text{ customers}$
Chinese Restaurant Process (CRP)

1. Customer 5 arrives.
2. \( P(\text{new table}) \propto \alpha \)
3. \( P(\text{table } \{1,2,4\}) \propto |\{1,2,4\}| = 3 \)
4. \( P(\text{table } \{3\}) \propto |\{3\}| = 1 \)
5. Pick new table
6. Sample new dish \( \theta_5 = Fish \)
7. \( \{\{1,2,4\},\{3\},\{5\}\}, n=5 \text{ customers} \)
Chinese Restaurant Process (CRP)

Customer 6 arrives.

P(new table) ∝ α

P(table \{1,2,4\}) ∝ |\{1,2,4\}| = 3

P(table \{3\}) ∝ |\{3\}| = 1

P(table \{5\}) ∝ |\{5\}| = 1

Pick table \{1,2,4\}

Share dish \(\theta_6 = \theta_1 = Tofu\)

\{\{1,2,4,6\},\{3\},\{5\}\}, n=6 customers
What does $G$ look like?

$\theta^{(2)}$

$\pi_1$

$\pi_2$

$\pi_3$

$\theta_{k=1} = (1, -3)$

$\theta_{k=2} = (0, 1)$

$\theta_{k=3} = (2, -1)$

$\Omega$
We can describe a generative process of data points.

(but first let’s recall the generative process for GMM)