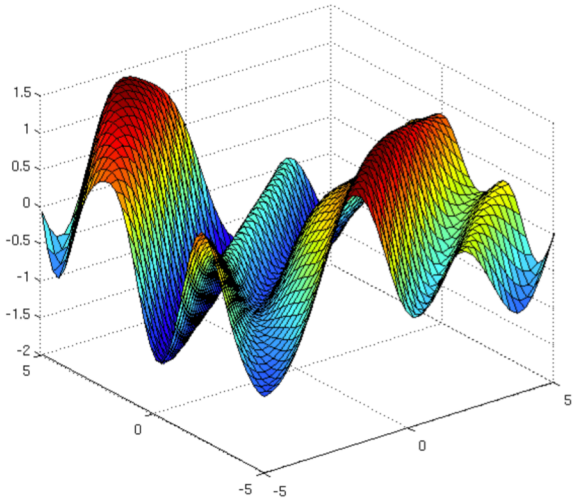
Stochastic Processes, Kernel Regression, Infinite Mixture Models



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.

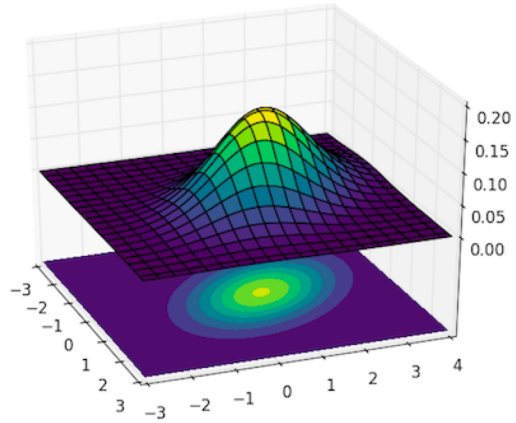
Disclaimer

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

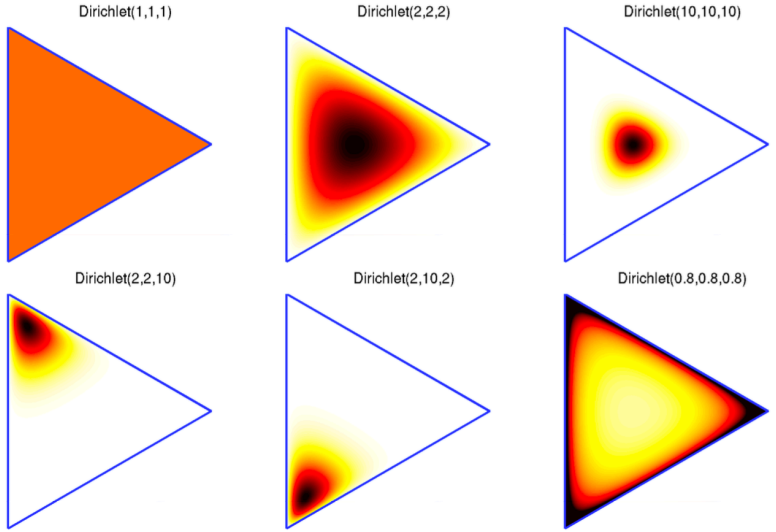
Samples X in R^d.



Dirichlet Distribution

Samples π in Simplex Δ_{d-1}

verifies $\pi_1 + \cdots + \pi_d = 1$



Why Gaussian and Dirichlet?

They are often used as priors

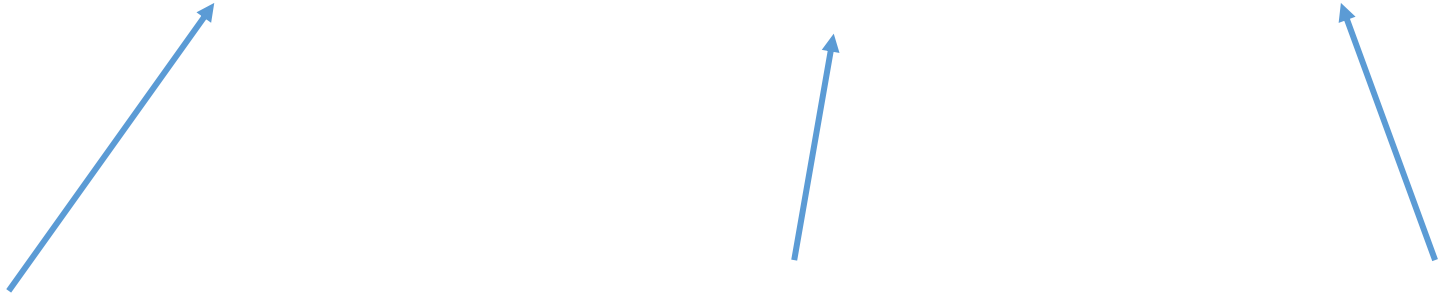
Bayesians like to use those distributions as priors over model parameters $p(x)$

Why?

Because they are very convenient to represent/update.

Conjugate Priors

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$



Posterior Likelihood model Prior

Conjugate Prior means:

Posterior in same family as prior

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

Posterior

$\theta|x \sim \mathbf{Gaussian}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$

Prior

$\theta|x \sim \mathbf{Gaussian}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Likelihood

$x|\theta \sim \mathbf{Gaussian}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_L)$

Gaussian is conjugate prior for **Gaussian** likelihood model.

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

Posterior

$\theta|x \sim \textbf{Dirichlet}(\boldsymbol{\alpha'})$

Likelihood

$x|\theta \sim \textbf{Categorical}(\boldsymbol{\theta})$

Prior

$\theta|x \sim \textbf{Dirichlet}(\boldsymbol{\alpha})$

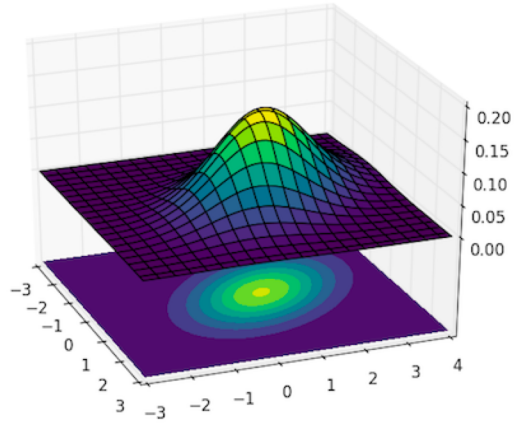
Dirichlet is conjugate prior for **Categorical/Multinoulli** likelihood model.

So taking the posterior is simply
a matter of updating
the parameters
of the prior.

Back to Gaussian and Dirichlet

Gaussian Distribution

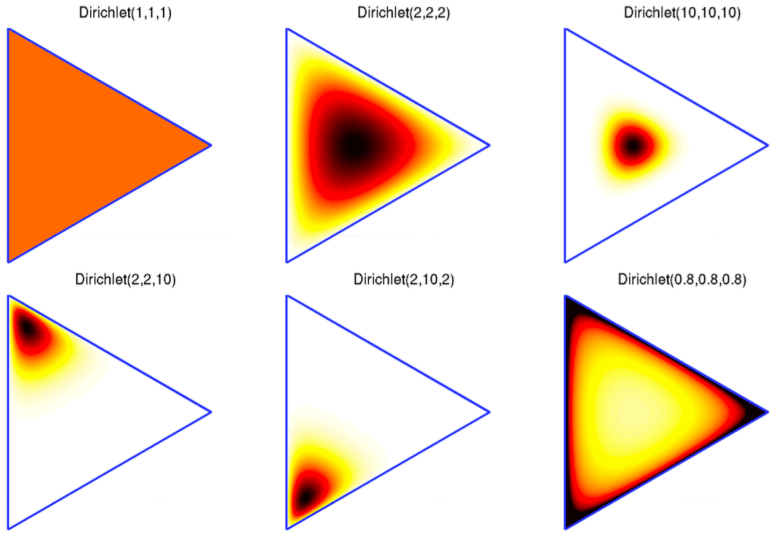
Samples X in R^d.



Dirichlet Distribution

Samples π in Simplex Δ_{d-1}

verifies $\pi_1 + \cdots + \pi_d = 1$



Gaussian and Dirichlet
are indexed with a
finite set of integers $\{1, \dots, d\}$.

They are **random vectors**.

$$(X_1, X_2, \dots, X_d)$$
$$(\pi_1, \pi_2, \dots, \pi_d)$$

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)

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

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 $\boldsymbol{X}: t \in T \rightarrow \mathbf{R}$ 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 $\boldsymbol{X \sim GP(\mu, \Sigma)}$ of a Gaussian Process are **random functions**

$$X: T \to \mathbf{R}$$

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

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

Parameters are the **Mean** function $\boldsymbol{\mu(t)}$ and **Covariance** function $\boldsymbol{\Sigma(t, t')}$**.**

For any $t_1, t_2, \ldots, t_k \in T$ we define the following finite-dimensional distributions $\mathrm{p}(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$.

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

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

$$\boldsymbol{X \sim GP(\mu, \Sigma)}$$

Characterizing Gaussian Process

Some properties are immediate consequences of definition:

- $\mathbf{E}[X_t] = \mu(t)$
- $\mathbf{Cov}(X_t, X_{t'}) = E[\left(X_t - \mu(t)\right)\left(X_{t'} - \mu(t')\right)^T] = \Sigma(t, t')$
- Any linear combination of distinct dimensions is still a Gaussian:

$$\sum_{i=1}^{k} \alpha_i * X_{t_i} \sim \mathcal{N}(\cdot, \cdot)$$

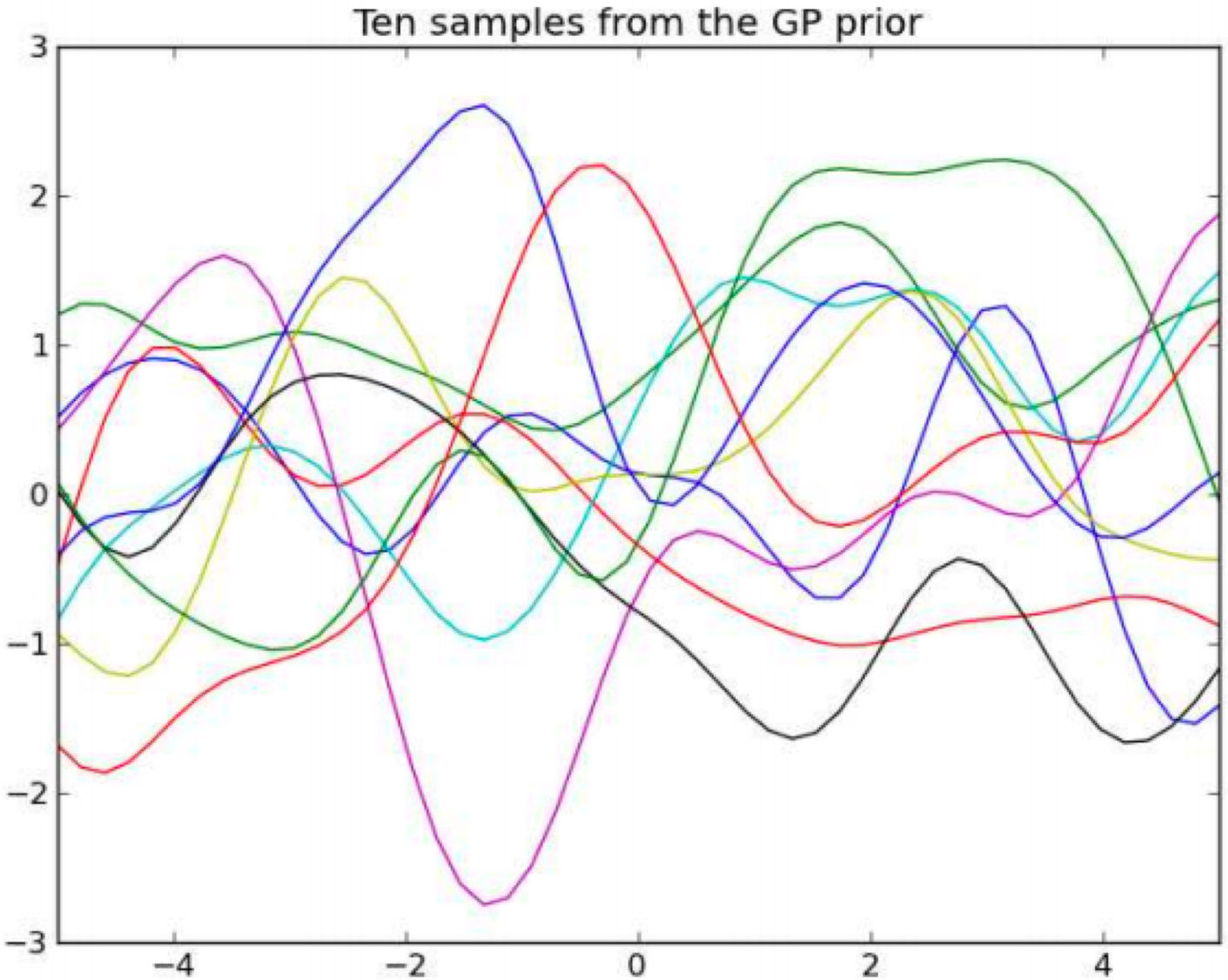
Characterizing Gaussian Process

Some properties are immediate consequences of definition:

- Stationarity: $\Sigma(t, t') = \Sigma(t - t')$
  does not depend on the positions
- Continuity: $\lim_{\boldsymbol{t' \to t}} \boldsymbol{\Sigma}(t, t') = \boldsymbol{\Sigma}(t, t)$
- 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:

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

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

Controls uncertainty

Controls smoothness (bandwidth/length-scale)

WARNING: Change of notation!
x is now the index and
$f(x)$ is the random function

Now, assume we observe a training set

$$X_n = (x_1, x_2, \ldots, x_n), y_n = (y_1, y_2, \ldots, y_n)$$

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

$$\boldsymbol{p(f|X_n, y_n) \propto\ p(y_n|f, X_n)\, p(f)}$$

- **Gaussian Process Prior**:

$$p(f) = GP\big(\mu(x), \Sigma(\mathrm{x}, \mathrm{x}')\big)$$

- **Gaussian Likelihood**:

$$p(y_n|f, X_n) = \mathcal{N}(f(X), \epsilon^2 I_n)$$

- **-> Gaussian Process Posterior**:

$$p(f|X_n, y_n) = GP\big(\mu'(x), \Sigma'(\mathrm{x}, \mathrm{x}')\big)$$

for some $\mu'(x), \Sigma'(x, x')$.

Remember: Gaussian Process is conjugate prior for Gaussian likelihood model.

Bayes' Rule

$$\boldsymbol{p(f|X_n, y_n) \propto p(y_n|f, X_n)\, p(f)}$$

- **Gaussian Process Prior**:

$$p(f) = GP\big(\mu(x), \Sigma(\mathrm{x}, \mathrm{x}')\big)$$

- **Dirac Likelihood**: $(\epsilon \to 0)$

$$p(y_n|f, X_n) = \delta\big(y_n - f(X_n)\big)$$

that is, y_n is now deterministic after observing f, X_n.

$$y_n = f(X_n)$$

- **-> Gaussian Process Posterior**:

$$p(f|X_n, y_n) = GP\big(\mu'(x), \Sigma'(\mathrm{x}, \mathrm{x}')\big)$$

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: $\boldsymbol{p(y^* | X_n, y_n, x^*)}$

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

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

$$\begin{bmatrix} y_n \\ y^* \end{bmatrix} \Big| \begin{bmatrix} X_n \\ x^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(X_n) \\ \mu(x^*) \end{bmatrix}, \begin{bmatrix} \Sigma(X_n, X_n) & \Sigma(X_n, x^*) \\ \Sigma(x^*, X_n) & \Sigma(x^*, x^*) \end{bmatrix} \right)$$

Theorem: For a Gaussian vector with distribution

$$\begin{matrix} X_1 \\ X_2 \end{matrix} \sim \mathcal{N}\left(\begin{matrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{matrix}, \begin{matrix} \boldsymbol{\Sigma_{1,1}} & \boldsymbol{\Sigma_{1,2}} \\ \boldsymbol{\Sigma_{2,1}} & \boldsymbol{\Sigma_{2,2}} \end{matrix} \right)$$

the conditional distribution $X_2 | X_1$ is given by

$$X_2 \,\middle|\, X_1 \sim \mathcal{N}\left(\boldsymbol{\mu_2 + \Sigma_{2,1}\Sigma_{1,1}^{-1}(x_1 - \mu_1)}, \boldsymbol{\Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}} \right)$$

This Theorem will be useful for the Kalman filter, later on …

[Schur's complement]

Applying the previous theorem gives us the posterior y^*.

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

$$\begin{bmatrix} y^* \\ \hline y_n \\ \hline X_n \\ \hline x^* \end{bmatrix} \sim \mathcal{N}\left(\mu' \;,\; \Sigma \right)$$

$$\boldsymbol{\Sigma' = \Sigma(x^*, x^*) - \Sigma(x^*, X_n)\Sigma(X_n, X_n)^{-1}\Sigma(X_n, x^*)}$$

Active Learning with Gaussian Process.

Active Learning

Active Learning is 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

$$\pi = (\pi_1, \pi_2, \dots) \sim GEM(\alpha)$$

scalar weights
sum up to 1

$$\theta_1, \theta_2, \dots \sim_{iid} G_0$$

parameters, sampled
from base distribution

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

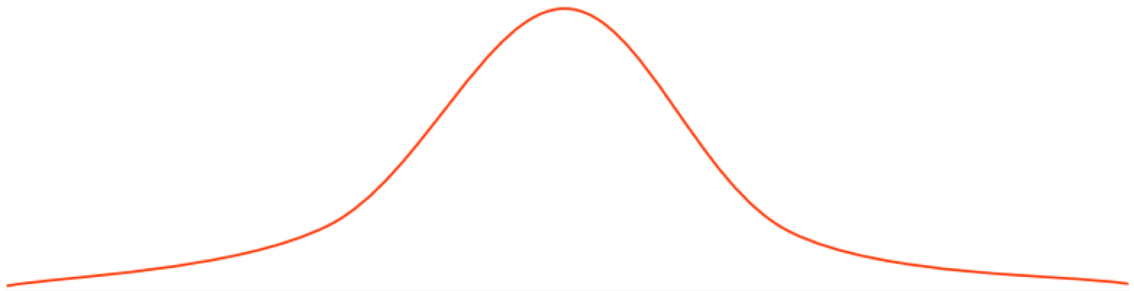
Diracs concentrate
probability mass π_k
at θ_k

G is a **random probability measure**:

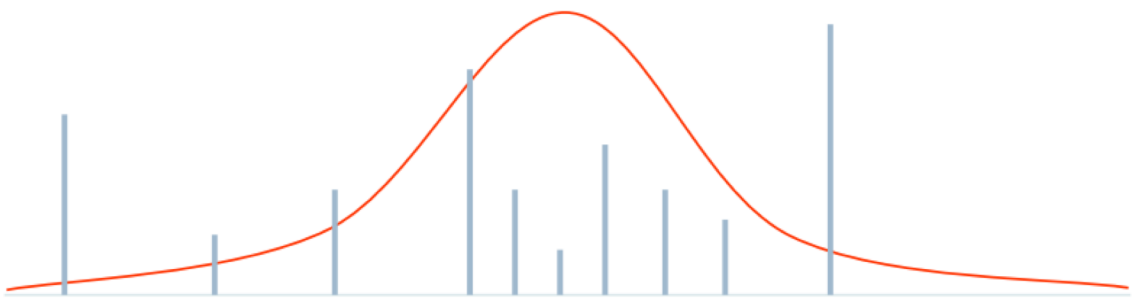
- **random**: both π and θ are random
- **probability measure**: it is a convex combination of Diracs, which are probability measures

Dirichlet Process

- Consider Gaussian G_0

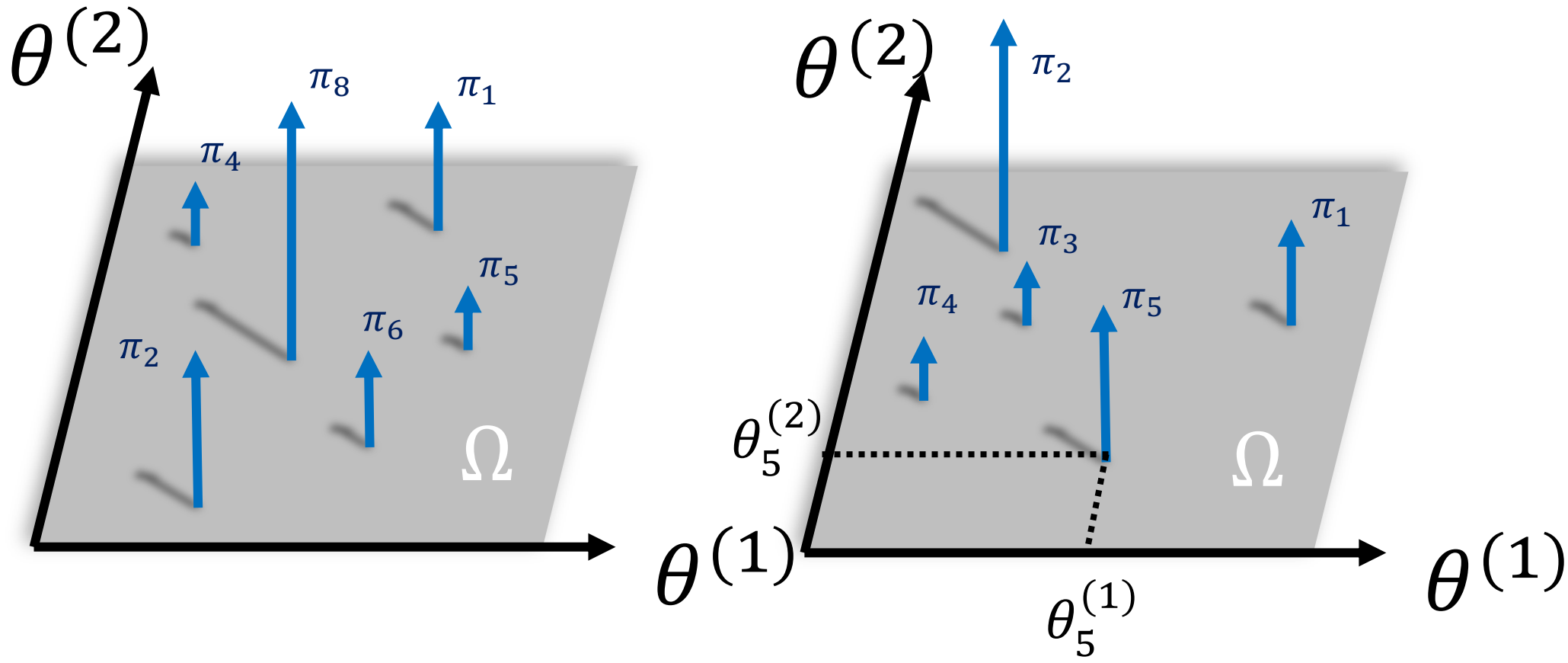


- $G \sim DP(\alpha, G_0)$



Courtesy of Khalid El-Arini

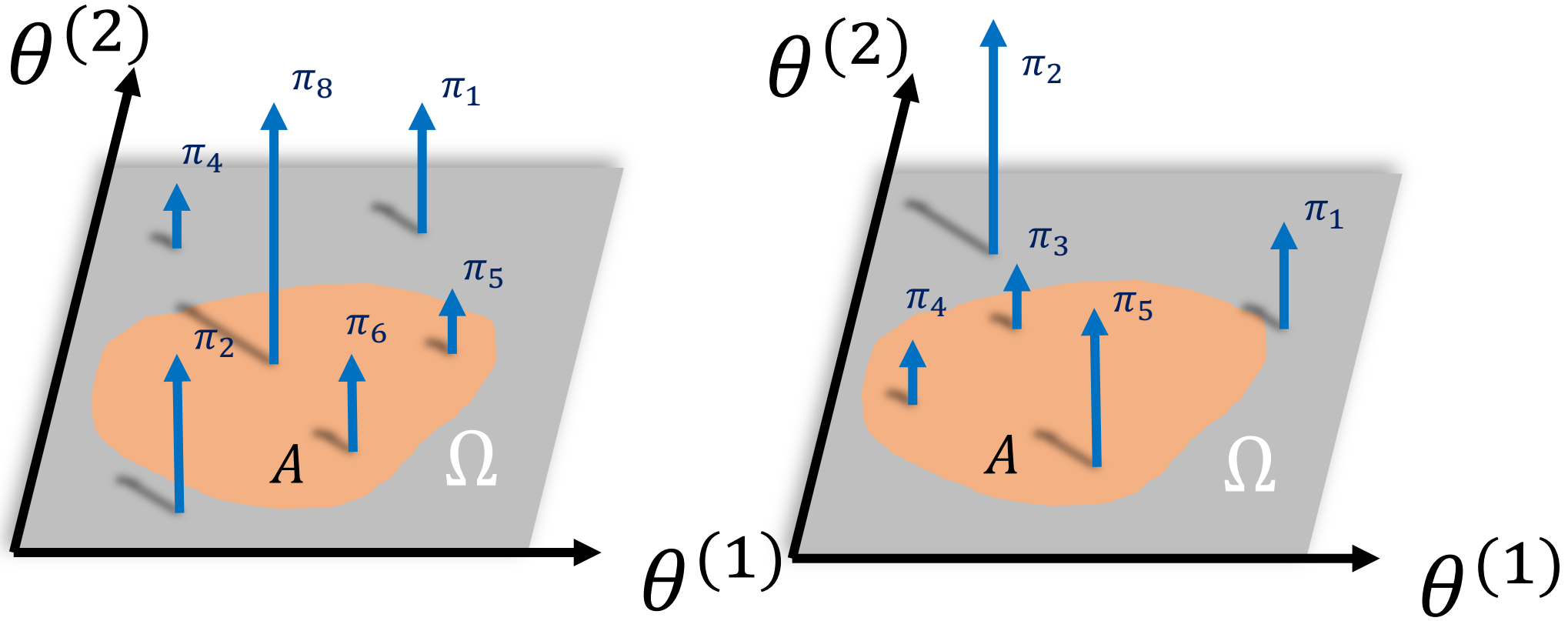
Two independent samples G from $DP(\alpha, G_0)$



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 $G(A) = \sum_{k=1}^{+\infty} \pi_k * 1\{\theta_k \in A\}$



$G_1(A) = \pi_5 + \pi_6 + \pi_8$

$G_1(A) = 0.05+0.1+0.3=0.45$

$G_2(A) = \pi_3 + \pi_4 + \pi_5$

$G_2(A) = 0.05+0.05+0.2=0.3$

For a fixed subset A, notice how G(A) is random.
In fact even the π_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

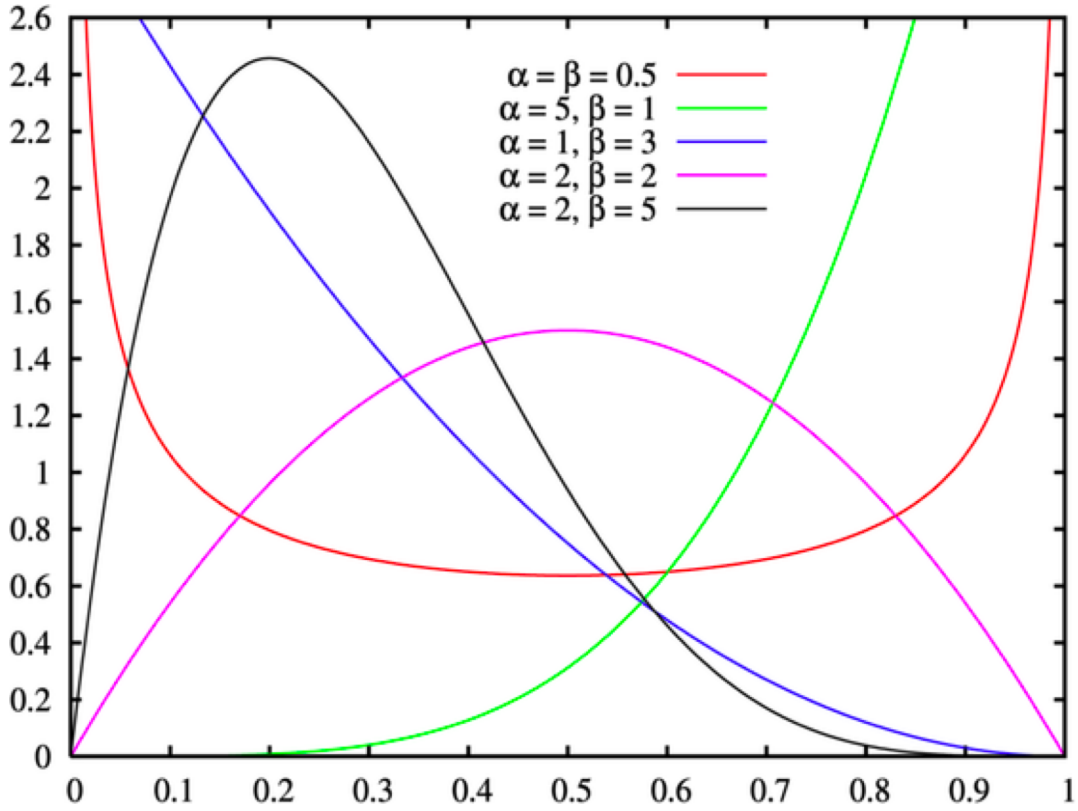
$$\pi \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$$

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**

$$\pi \sim GEM(1, \alpha)$$

Beta Distribution

$$\pi_1 \sim \mathbf{Beta}(\alpha, \beta)$$
$$\pi_2 = 1 - \pi_1$$
$$p(\pi_1|\alpha, \beta) \propto \pi_1^{\alpha-1}(1-\pi_1)^{\beta-1}$$

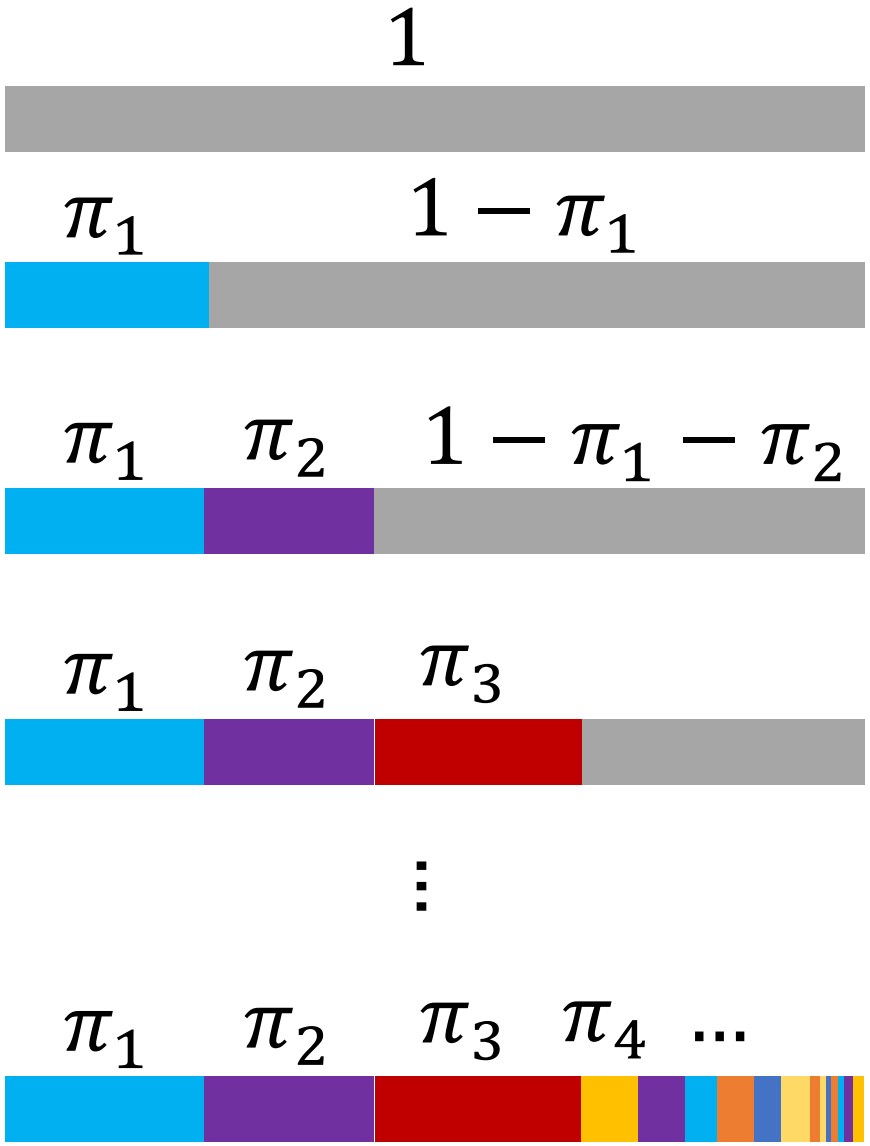


Equivalent to:

$$\pi_1, \pi_2 \sim \mathbf{Dirichlet}(\alpha, \beta)$$
$$p(\pi_1, \pi_2|\alpha, \beta) \propto \pi_1^{\alpha-1}\pi_2^{\beta-1}$$

$\alpha, \beta \to +\infty$ gives peaked distribution around $\alpha/(\alpha+\beta)$

Stick Breaking $\pi \sim GEM(\alpha)$



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

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

$$V_3 \sim Beta(1, \alpha)$$

$$\pi_3 = V_2(1 - \pi_1 - \pi_2)$$

Griffiths, Engen, McCloskey

Defining Dirichlet Process from finite-dimensional marginals.

Dirichlet Process

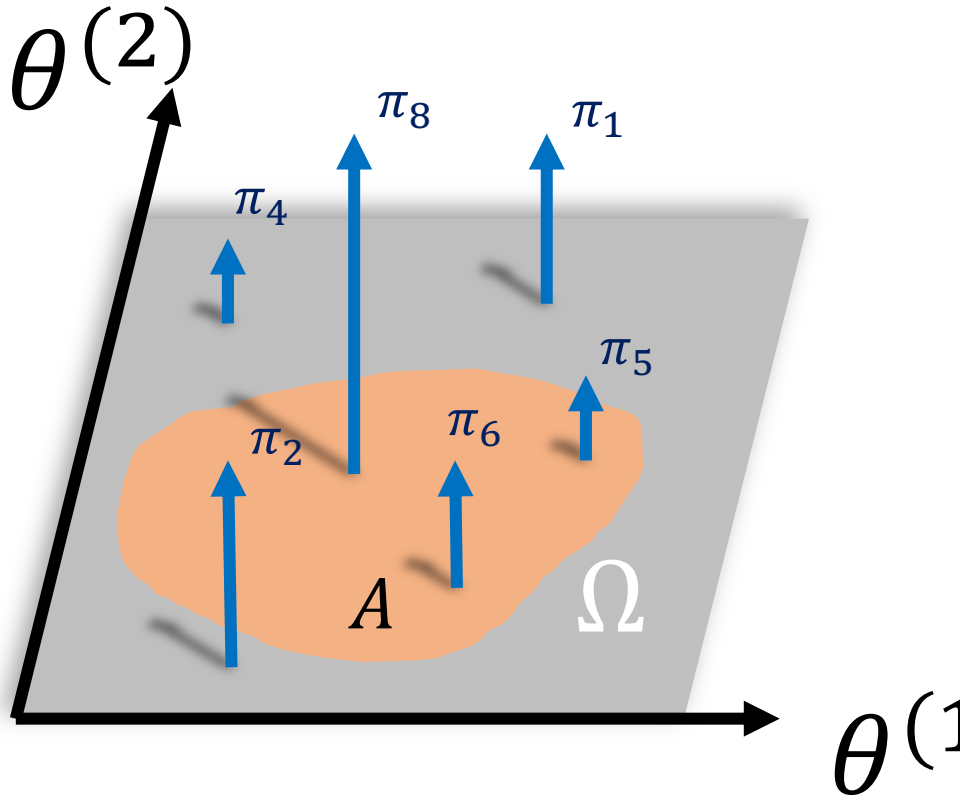
Samples $\boldsymbol{G \sim DP(\alpha, G_0)}$ of a Dirichlet Process are themselves probability measures (i.e. distributions) over a measurable space $(\Omega, \mathcal{F})$.

$$G: \mathcal{F} \rightarrow \mathbf{R}_+$$

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** $\boldsymbol{G_0}$ (over Ω) and the parameter $\boldsymbol{\alpha} > 0$.



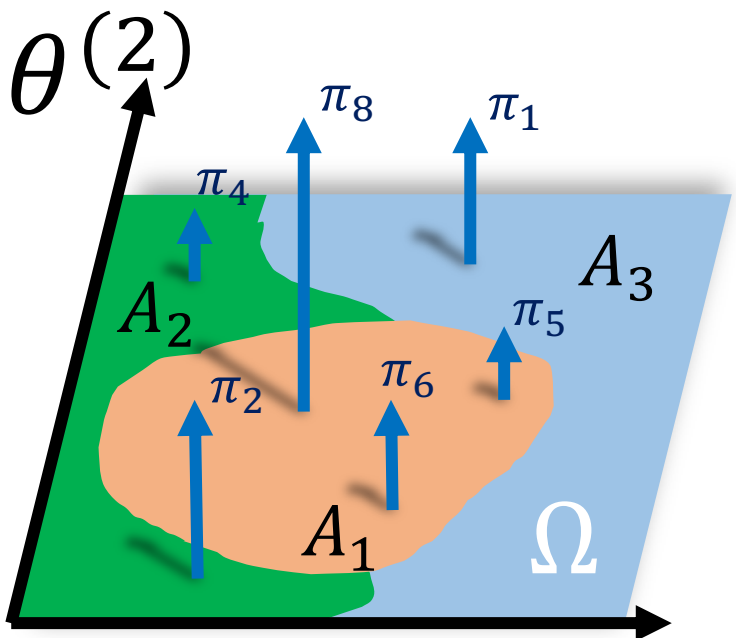
Kolmogorov Consistency Construction

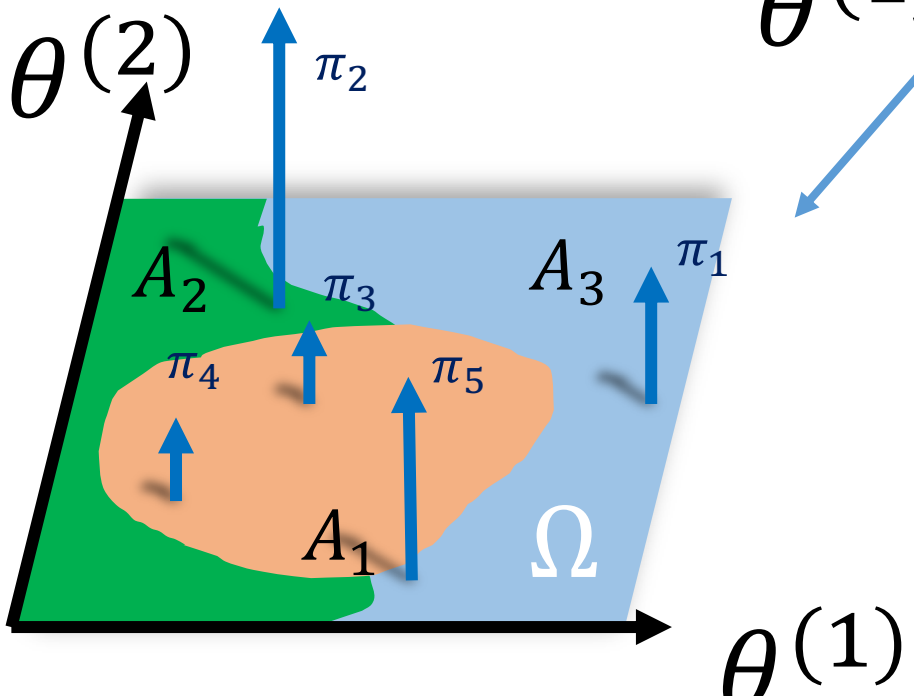
For any $k \geq 0$, consider any partition $A_1, A_2, \dots, A_k$ of the space Ω. We define the following finite-dimensional distributions

$$G(A_1), \dots, G(A_k) \sim Dirichlet(\boldsymbol{\alpha * G_0(A_1), \dots, \alpha * G_0(A_k)})$$

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:**

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





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

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

$$G(A_1), G(A_2), G(A_3) \sim \textbf{Dirichlet}(8,2,2)$$

All constructions match.

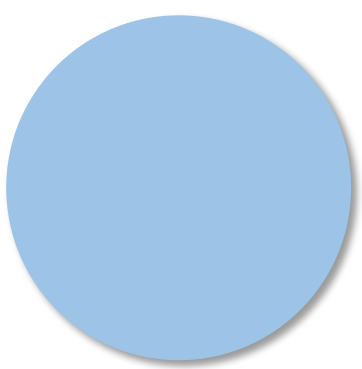
It can be shown that Stick-Breaking and Kolmogorov consistency definitions match.

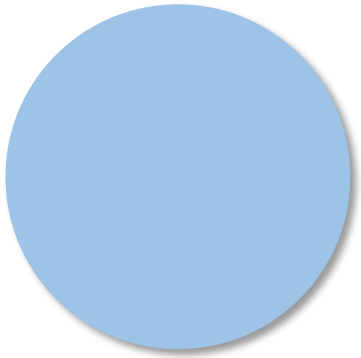
https://www.stat.ubc.ca/~bouchard/courses/stat547-sp2011/notes-part2.pdf

Defining Dirichlet Process from Chinese Restaurant Process /BlackWell-McQueen Urn.

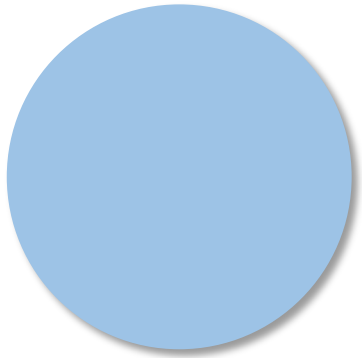
Chinese Restaurant Process (CRP)

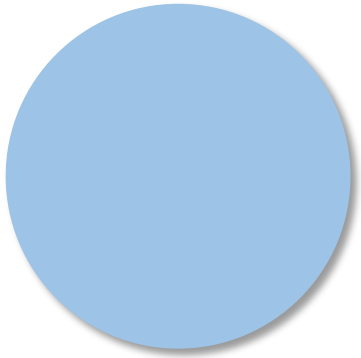
Infinity of Tables





...

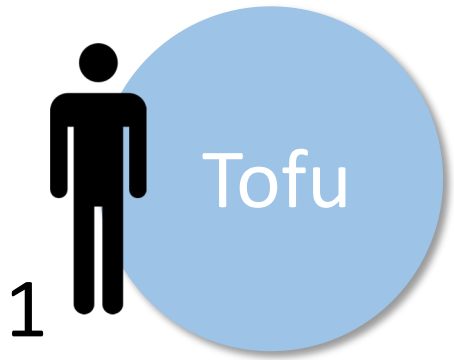


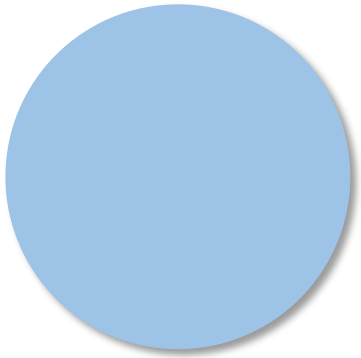


$$G_0 = Uniform(\{Fish, Pork, Tofu\})$$

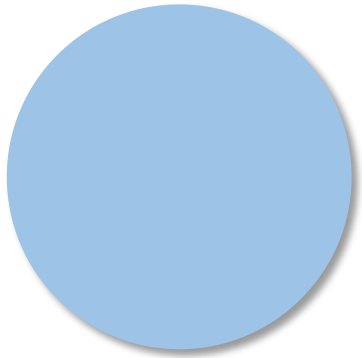
Chinese Restaurant Process (CRP)

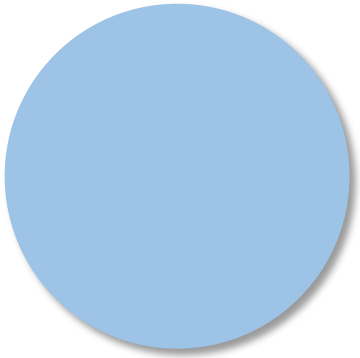
Infinity of Tables





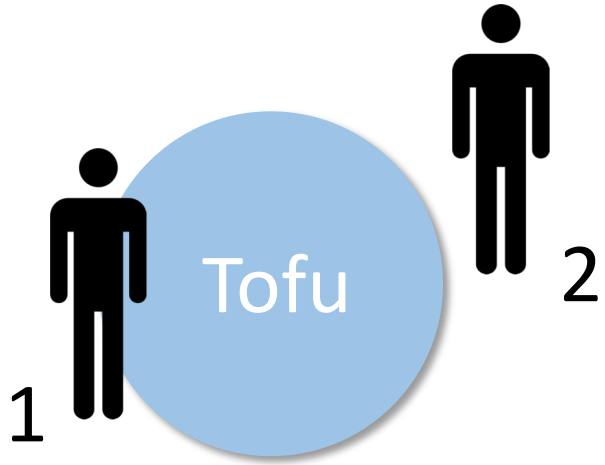
...



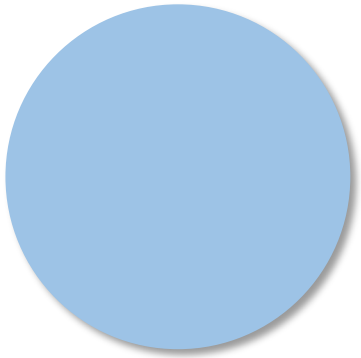


- Customer 1 arrives.
- Takes any free table.
- Sample a dish $\theta_1 \sim G_0$-> Tofu
- **state={{1}}**, n=1 customers

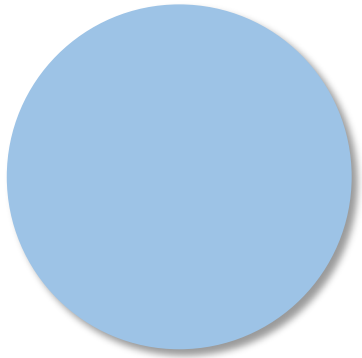
Chinese Restaurant Process (CRP)

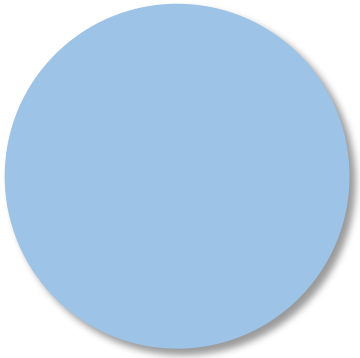


Infinity of Tables



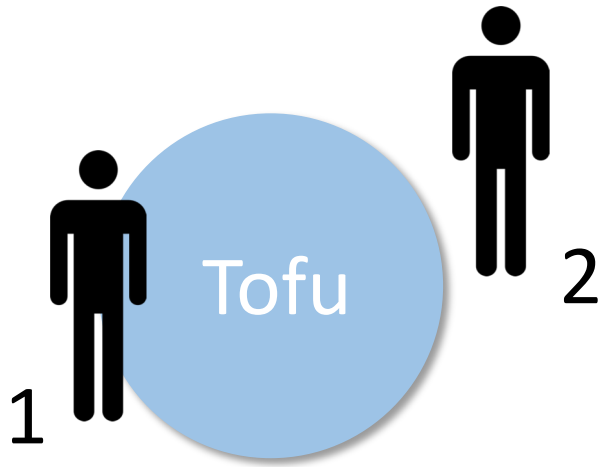
…



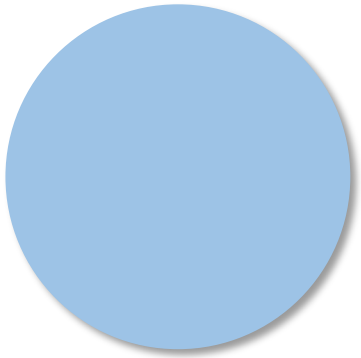


- Customer 2 arrives.
- P(new table)$\propto \alpha$
- P(table $\{1\}$)$\propto |\{1\}| = 1$
- Decides to sit at $\{1\}$
- Share dish: $\theta_2 = \theta_1$ =Tofu
- **{{1,2}},** n=2 customers

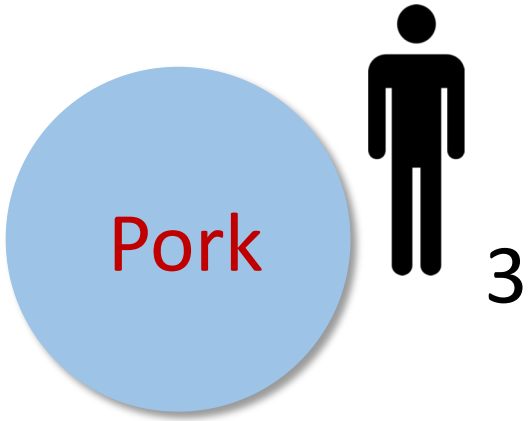
Chinese Restaurant Process (CRP)

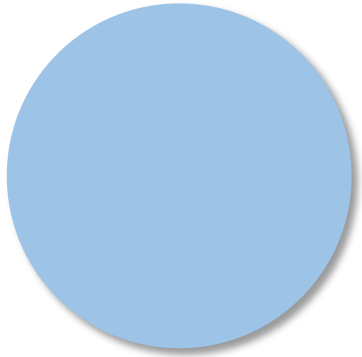


Infinity of Tables



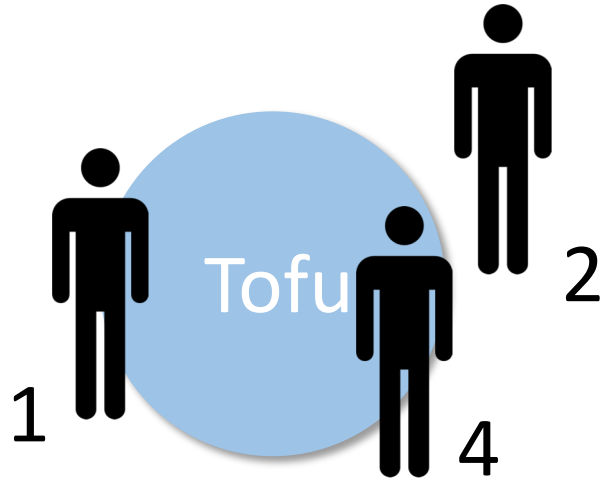
…



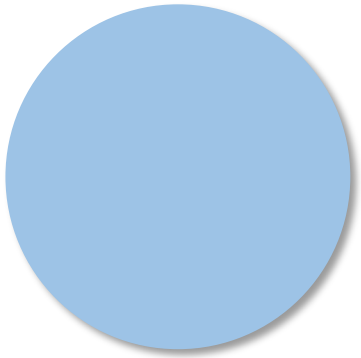


- Customer 3 arrives.
- P(new table)$\propto \alpha$
- P(table $\{1,2\}$)$\propto |\{1,2\}| = 2$
- Decides to sit at new table
- Sample a dish $\theta_3 \sim G_0$-> Pork
- **{{1,2},{3}},** n=3 customers

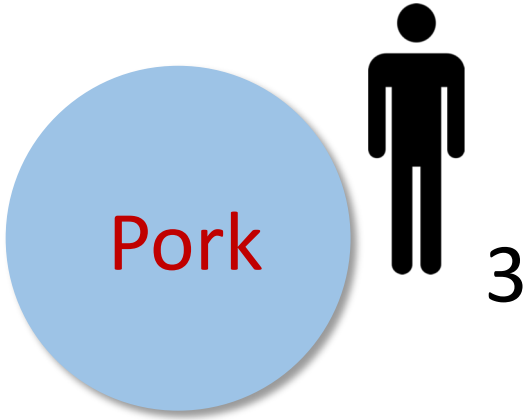
Chinese Restaurant Process (CRP)

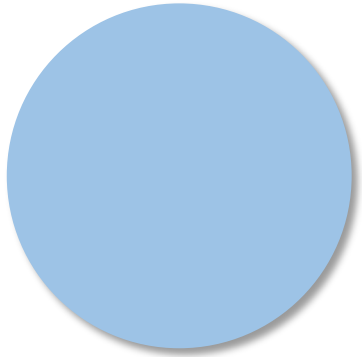


Infinity of Tables



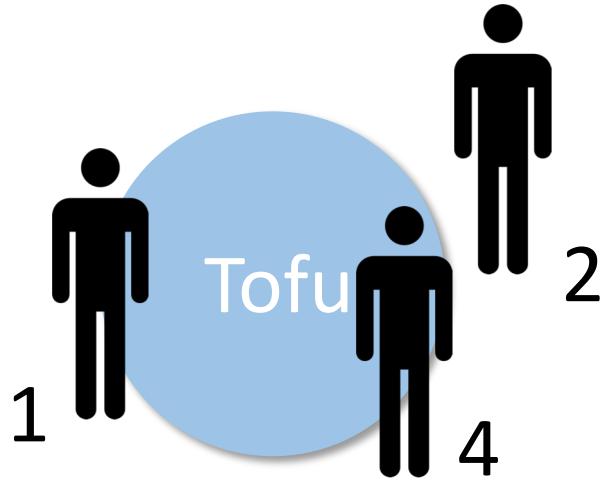
…

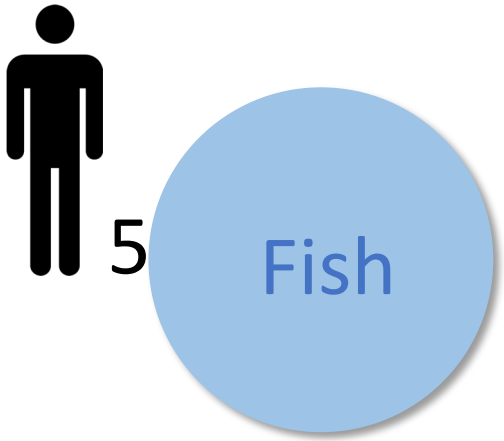




- Customer 4 arrives.
- P(new table)$\propto \alpha$
- P(table $\{1,2\}$)$\propto |\{1,2\}| = 2$
- P(table $\{3\}$)$\propto |\{3\}| = 1$
- Share dish, $\theta_4 = \theta_1$ =Tofu
- **{{1,2,4},{3}},** n=4 customers

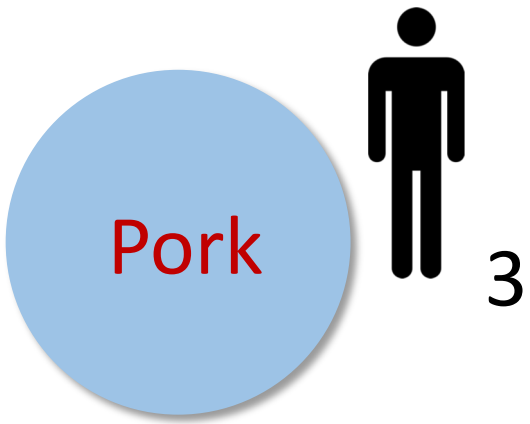
Chinese Restaurant Process (CRP)

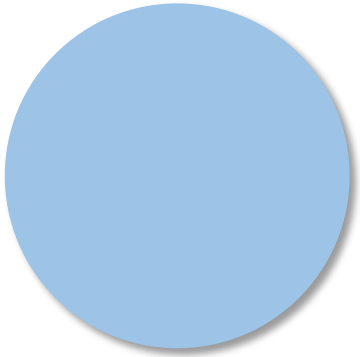




Infinity of Tables

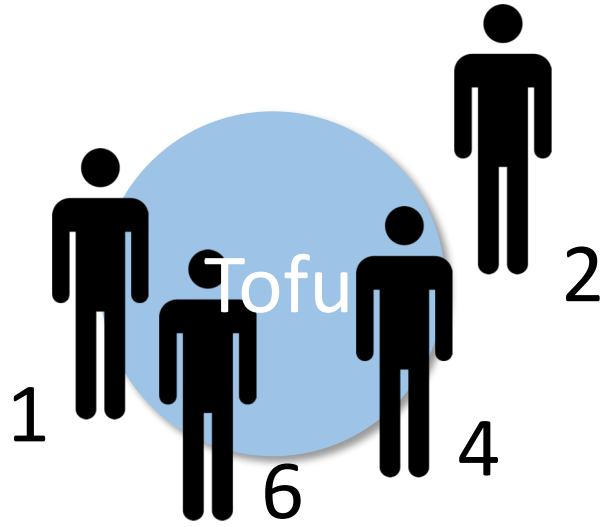
…

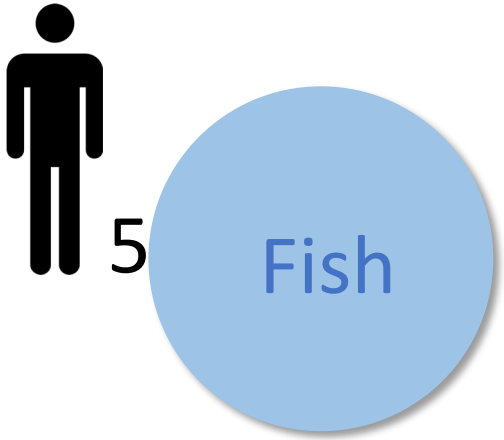




- Customer 5 arrives.
- P(new table)$\propto \alpha$
- P(table $\{1,2,4\}$)$\propto |\{1,2,4\}| = 3$
- P(table $\{3\}$)$\propto |\{3\}| = 1$
- Pick new table
- Sample new dish $\theta_5 = Fish$
- **{{1,2,4},{3},{5}},** n=5 customers

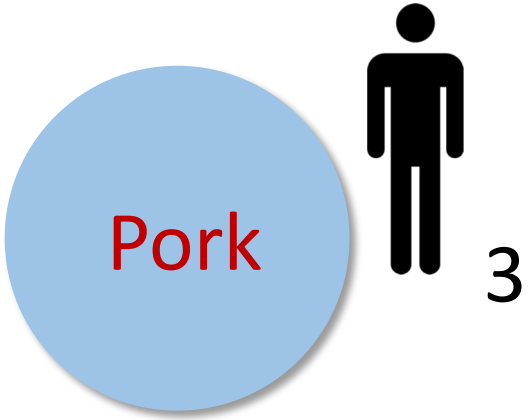
Chinese Restaurant Process (CRP)

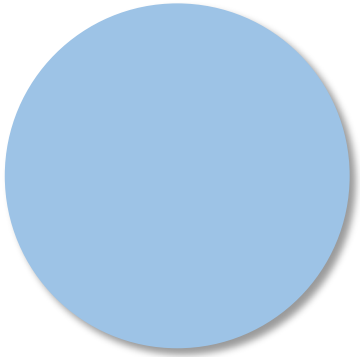




Infinity of Tables

…





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

Chinese Restaurant Process (CRP)

We can look at the sequence of dishes

$$\theta_1 = Tofu$$
$$\theta_2 = Tofu$$
$$\theta_3 = Pork$$
$$\theta_4 = Tofu$$
$$\theta_5 = Fish$$
$$\theta_6 = Tofu$$

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

$$p(\theta_1 = u_1, \theta_2 = u_2, \dots) =$$
$$p\left(\theta_1 = u_{\sigma(1)}, \theta_2 = u_{\sigma(2)}, \dots\right)$$

The **order** in which the customers arrive is actually **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*:

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

  for some random variable θ.

- 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 http://faculty.dbmi.pitt.edu/day/Bioinf2132-advanced-Bayes-and-R/Bioinf2132-documents-2017/2017-11-30/nips-tutorial05.pdf

De Finetti's Theorem

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

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

Here: $\boldsymbol{G_0} = \frac{1}{3}\delta_{fish} + \frac{1}{3}\delta_{pork} + \frac{1}{3}\delta_{tofu}$, $\boldsymbol{\alpha}$ is the same ($\propto new\ table$)
Let $\pi = (\pi_1, \pi_2, \dots) \sim \mathbf{GEM}(\alpha)$ stick-breaking.
Sample $\theta = (\theta_{k=1}, \theta_{k=2}, \dots) \sim_{iid} \mathbf{G_0}$
Now we can form our random measure $\mathbf{G} = \sum_{k=1}^{+\infty} \pi_k * \delta_{\theta_k}$.
And we sample $\theta_{i=1}, \theta_{i=2}, \dots \sim_{iid} \mathbf{G}$

*Unique in distribution.

θ_i is the parameter for data point i (customer i)
θ_k is the parameter for component k (table k)

Blackwell-McQueen Urn
Polya Urn

Same process, different story.

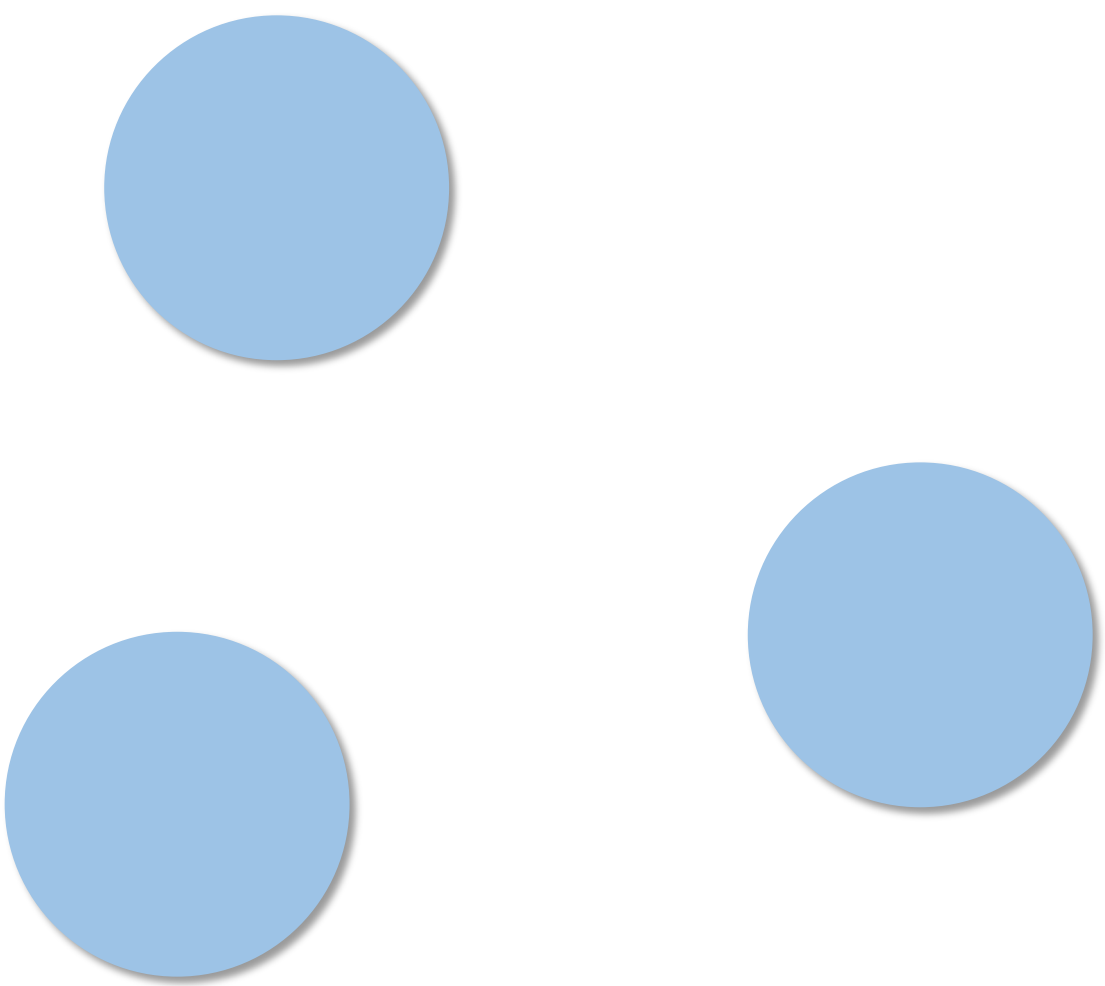
Each dish is a set of unique ball colors.

Each customer is a successive draw.

Using Dirichlet Process for infinite mixture models.

Chinese Restaurant Process (CRP)

Infinity of **Components**

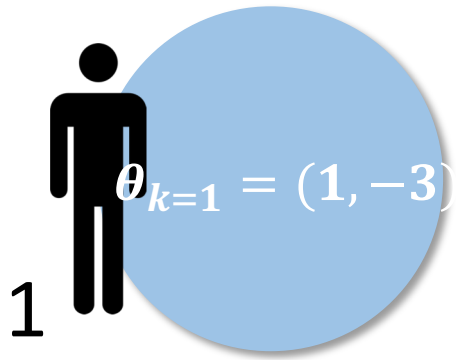


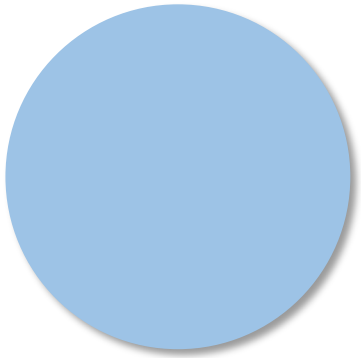
...

$$G_0 = \mathcal{N}(\mu_0, \Sigma_0)$$

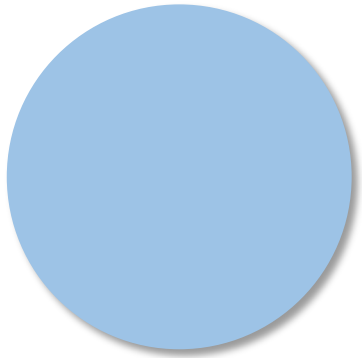
Chinese Restaurant Process (CRP)

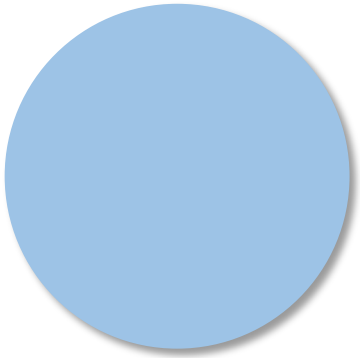
Infinity of Tables





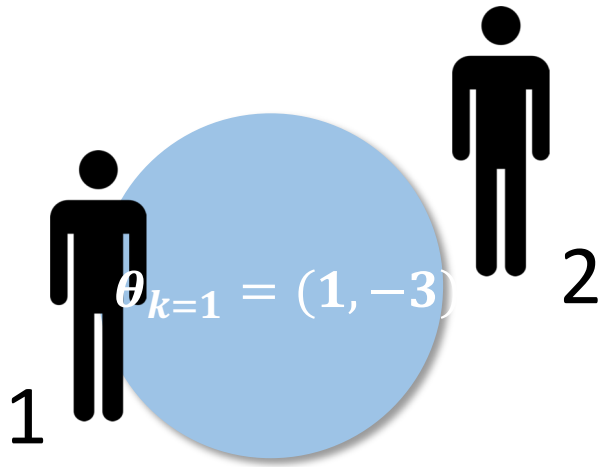
...



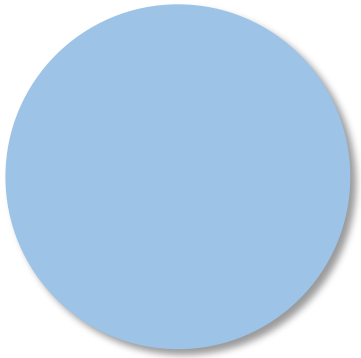


- Sample parameter for data point 1.
- Takes any free table.
- Sample a parameter $\theta_1 \sim G_0$
- **state={{1}}**, n=1 customers

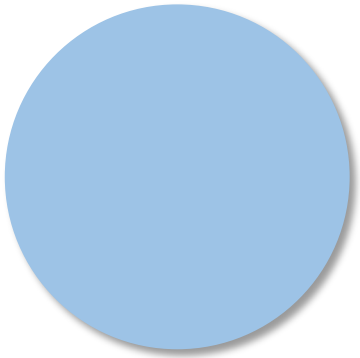
Chinese Restaurant Process (CRP)

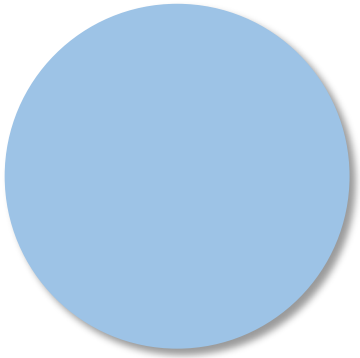


Infinity of Tables



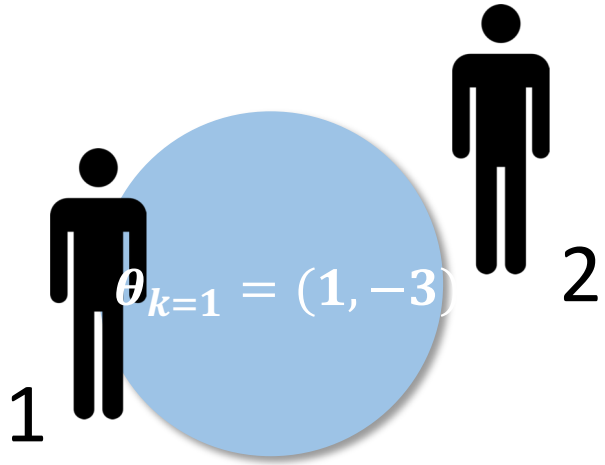
…



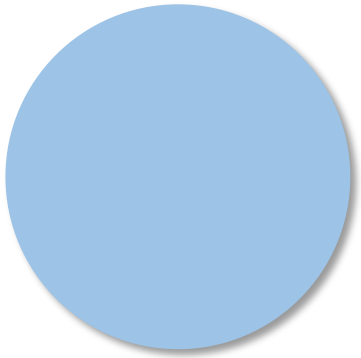


- Sample parameter for data point 2.
- P(new table)$\propto \alpha$
- P(table $\{1\}$)$\propto |\{1\}| = 1$
- Decides to sit at $\{1\}$
- Share dish: $\theta_2 = \theta_1$ =Tofu
- **{{1,2}},** n=2 customers

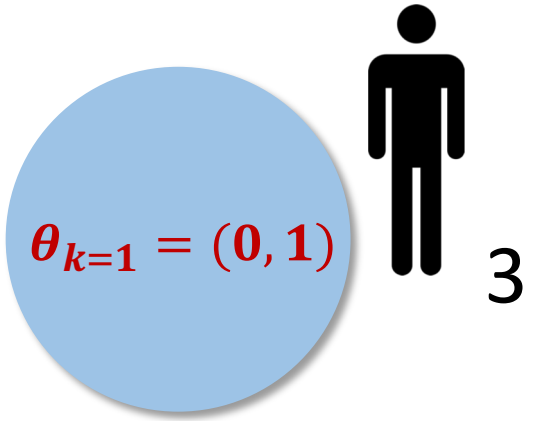
Chinese Restaurant Process (CRP)

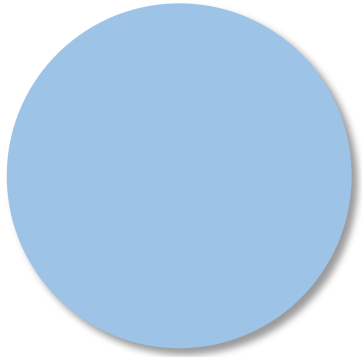


Infinity of Tables



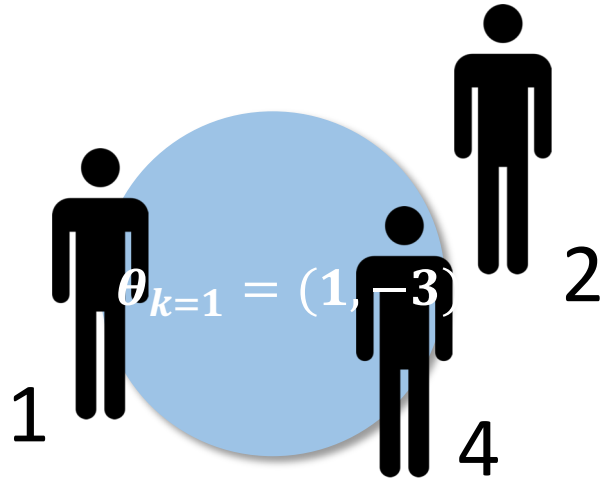
…



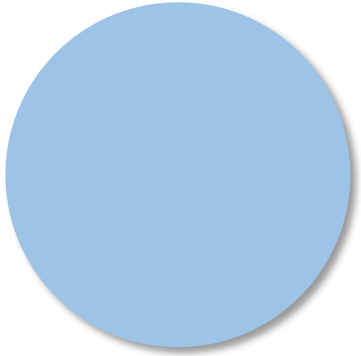


- Sample parameter for data point 3.
- P(new table)$\propto \alpha$
- P(table $\{1,2\}$)$\propto |\{1,2\}| = 2$
- Decides to sit at new table
- Sample a dish $\theta_3 \sim G_0$-> Pork
- **{{1,2},{3}},** n=3 customers

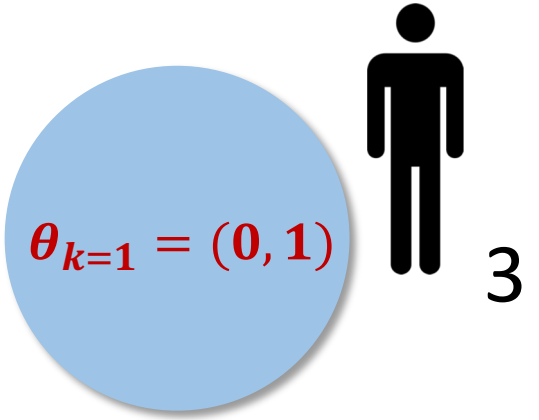
Chinese Restaurant Process (CRP)

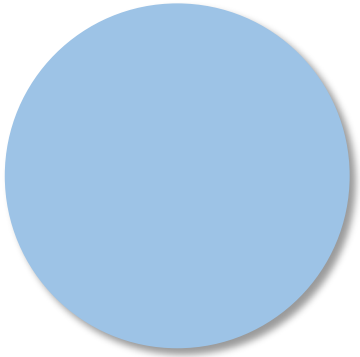


Infinity of Tables



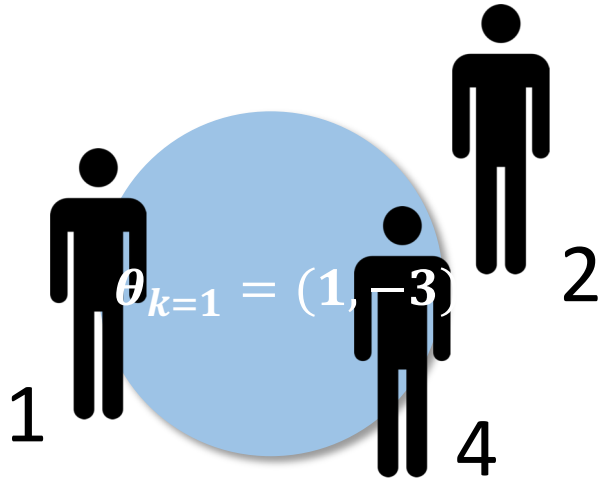
…

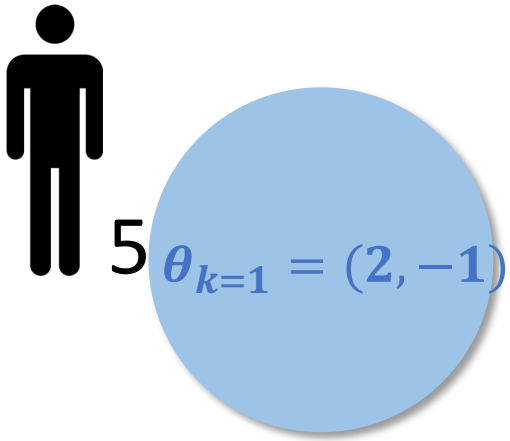




- Customer 4 arrives.
- P(new table)$\propto \alpha$
- P(table $\{1,2\}$)$\propto |\{1,2\}| = 2$
- P(table $\{3\}$)$\propto |\{3\}| = 1$
- Share dish, $\theta_4 = \theta_1$ =Tofu
- **{{1,2,4},{3}},** n=4 customers

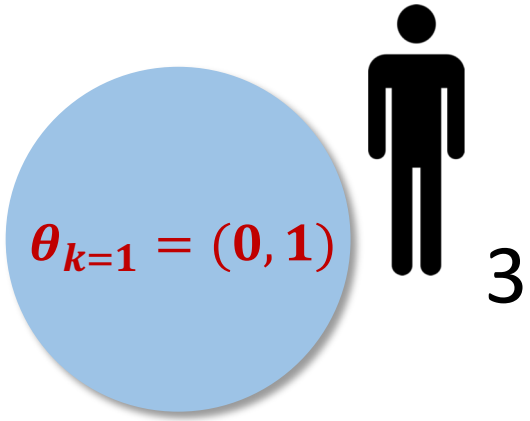
Chinese Restaurant Process (CRP)

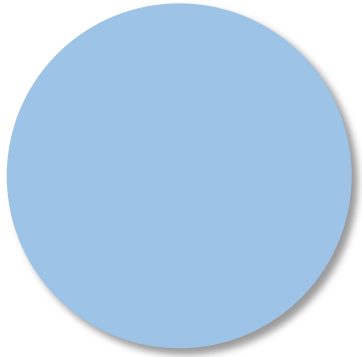




Infinity of Tables

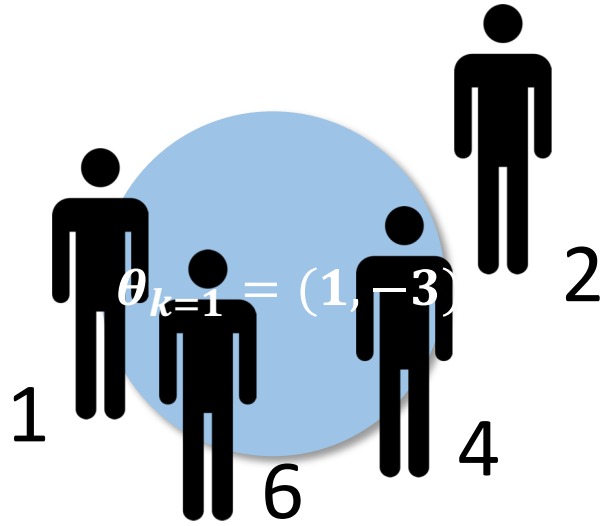
...

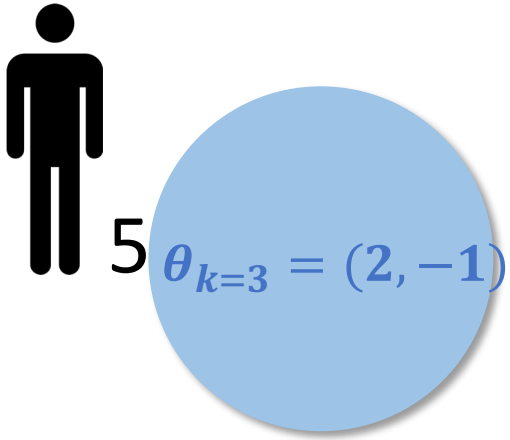




- Customer 5 arrives.
- P(new table)$\propto \alpha$
- P(table $\{1,2,4\}$)$\propto |\{1,2,4\}| = 3$
- P(table $\{3\}$)$\propto |\{3\}| = 1$
- Pick new table
- Sample new dish $\theta_5 = Fish$
- **{{1,2,4},{3},{5}},** n=5 customers

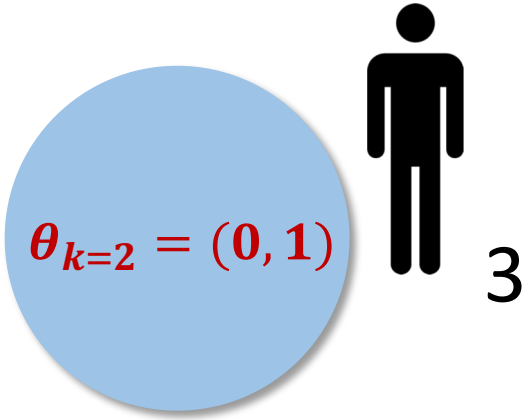
Chinese Restaurant Process (CRP)

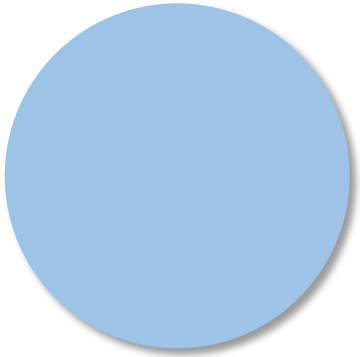




Infinity of Tables

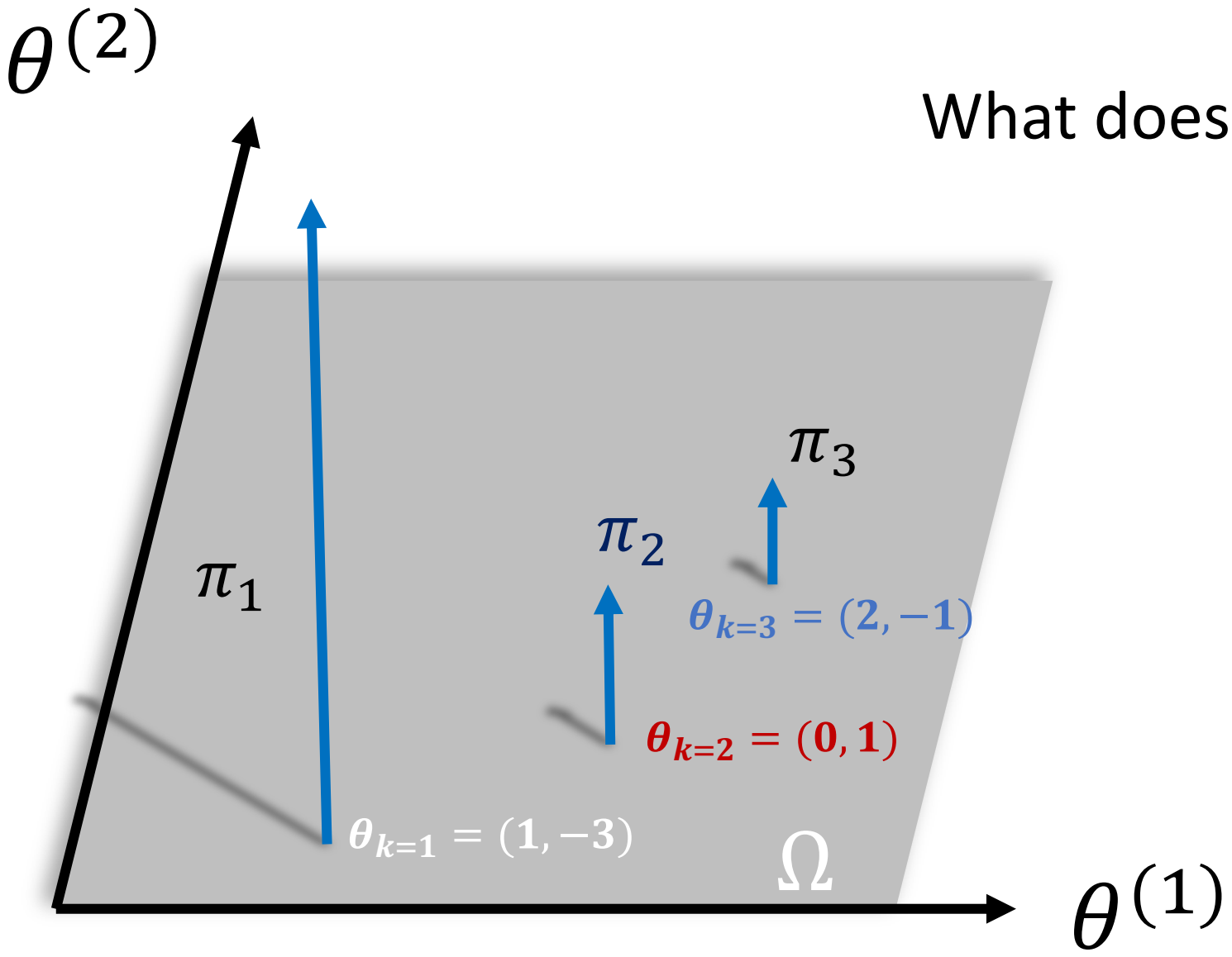
...





- Customer 6 arrives.
- P(new table)$\propto \alpha$
- P(table $\{1,2,4\}$)$\propto |\{1,2,4\}| = 3$
- P(table $\{3\}$)$\propto |\{3\}| = 1$
- P(table $\{5\}$)$\propto |\{5\}| = 1$
- 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?



We can describe a generative process of data points.

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