An introduction to Bayesian nonparametrics
Lecture 1: The Dirichlet process

Sinead Williamson

Department of Statistics and Data Sciences
McCombs School of Business

THE UNIVERSITY OF TEXAS AT AUSTIN

Lecture Overview

THE UNIVERSITY OF TEXAS AT AUSTIN

- Lecture 1: The Dirichlet process
  - What is Bayesian nonparametrics?
  - From Dirichlet distribution to Dirichlet Process
  - Representations
  - Inference
- Lecture 2: The Indian buffet process
- Lecture 3: Hierarchical nonparametric models

What is Bayesian nonparametrics?

THE UNIVERSITY OF TEXAS AT AUSTIN

- In this summer school, you've seen various examples of Bayesian modeling.
- General framework:
  - Come up with a class of models, parametrized by some set of parameters Θ.
  - Place a prior distribution over the parameters.
  - Update our *posterior* distribution as we see observations.

What is Bayesian nonparametrics?

THE UNIVERSITY OF TEXAS AT AUSTIN

- In this summer school, you've seen various examples of Bayesian modeling.
- General framework:
  - Come up with a class of models, parametrized by some set of parameters Θ.
  - Place a prior distribution over the parameters.
  - Update our *posterior* distribution as we see observations.

- *All models are wrong, but some are useful* – George Box

- We want a model that captures our intuitions about the data, while minimizing erroneous assumptions... this can be hard!
  - How to choose the number of topics to model the New York Times?
  - What if our test set contains features not present in our training set?

What is Bayesian nonparametrics?

THE UNIVERSITY OF TEXAS AT AUSTIN

- A *parametric* Bayesian model is one with a fixed, pre-specified number of global parameters:
  - Bayesian linear regression: $y_i \sim \text{Normal}(\mathbf{x}_t^T \boldsymbol{\beta}, \sigma^2)$, $\boldsymbol{\beta}$ is of fixed size.
  - Mixture of K Gaussians: K means, K covariances, one probability vector.

What is Bayesian nonparametrics?

THE UNIVERSITY OF TEXAS AT AUSTIN

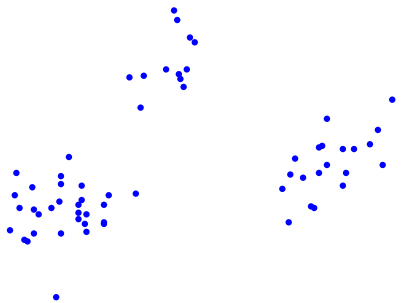
- A *parametric* Bayesian model is one with a fixed, pre-specified number of global parameters:
  - Bayesian linear regression: $y_i \sim \text{Normal}(\mathbf{x}_t^T \boldsymbol{\beta}, \sigma^2)$, $\boldsymbol{\beta}$ is of fixed size.
  - Mixture of K Gaussians: K means, K covariances, one probability vector.
- A *nonparametric* Bayesian model is *not* a model with no parameters...
- It is a model where the number of parameters can grow with parameter size.

What is Bayesian nonparametrics?

THE UNIVERSITY OF TEXAS AT AUSTIN

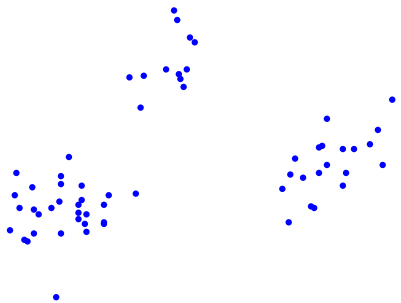
- A *parametric* Bayesian model is one with a fixed, pre-specified number of global parameters:
  - Bayesian linear regression: $y_i \sim \text{Normal}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$, $\boldsymbol{\beta}$ is of fixed size.
  - Mixture of K Gaussians: K means, K covariances, one probability vector.
- A *nonparametric* Bayesian model is *not* a model with no parameters...
- It is a model where the number of parameters can grow with parameter size.
- We achieve this by allowing an infinite number of parameters *a priori*.
- However, a finite data set will only ever use a finite number of data points.
  - ~~Bayesian linear regression~~ Gaussian processes – we need infinitely many values to pin down the function.
  - ~~Mixture of K Gaussians~~ Dirichlet process mixture model – infinitely many mixture components *a priori*.

Clustering data: Bayesian mixture models



- Obvious model: Mixture of three Gaussians, parametrized by a probability vector $\pi = (\pi_1, \pi_2, \pi_3)$, three means μ_1, μ_2, μ_3, three covariances $\Sigma_1, \Sigma_2, \Sigma_3$.
- For each data point,
  - Sample cluster indicator $z_i \sim \boldsymbol{\pi}$
  - Sample $x_i \sim Normal(\mu_{z_i}, \Sigma_{z_i})$

Clustering data: Bayesian mixture models



- Obvious model: Mixture of three Gaussians, parametrized by a probability vector $\pi = (\pi_1, \pi_2, \pi_3)$, three means μ_1, μ_2, μ_3, three covariances $\Sigma_1, \Sigma_2, \Sigma_3$.
- For each data point,
  - Sample cluster indicator $z_i \sim \boldsymbol{\pi}$
  - Sample $x_i \sim Normal(\mu_{z_i}, \Sigma_{z_i})$
- This gives us a likelihood

$$p(x_1, \ldots, x_N | \boldsymbol{\pi}, \{\mu_k\}, \{\Sigma_k\}) = \prod_{i=1}^{N} \sum_{k=1}^{3} \pi_k \text{Normal}(x_n | \mu_k, \Sigma_k)$$

Clustering data: Bayesian mixture models

THE UNIVERSITY OF TEXAS AT AUSTIN

- How to choose the mixing weights $\boldsymbol{\pi}$ and the mixture parameters $\{\mu_k, \Sigma_k\}$?
- Bayesian choice: Put a prior on them and integrate out:

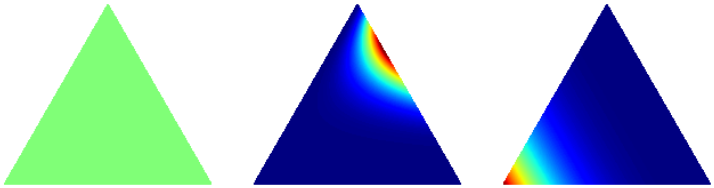
$$p(x_1, \ldots, x_N) = \int\int\int p(x_1, \ldots, x_N | \boldsymbol{\pi}, \{\mu_k\}, \{\Sigma_k\}) p(\boldsymbol{\pi}) d\boldsymbol{\pi} \prod_{k=1}^{3} p(\mu_k, \Sigma_k) d\mu_k d\Sigma_k$$

- Where possible, use conjugate priors:
  - Gaussian-inverse Wishart for mixture parameters
  - Dirichlet distribution for mixing weights
- Let's think about the Dirichlet distribution for a bit...

The Dirichlet distribution: A distribution over probability vectors

- The Dirichlet distribution is a distribution over the $(K-1)$-dimensional simplex – in other words, the space of all K-dimensional probability vectors.
- It is parametrized by a K-dimensional vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ such that each $\alpha_k \geq 0$ and $\sum_k \alpha_k > 0$.
- The expected value of a Dirichlet random variable π is given by $\mathbb{E}[\pi] = \frac{(\alpha_1, \ldots, \alpha_K)}{\sum_k \alpha_k}$

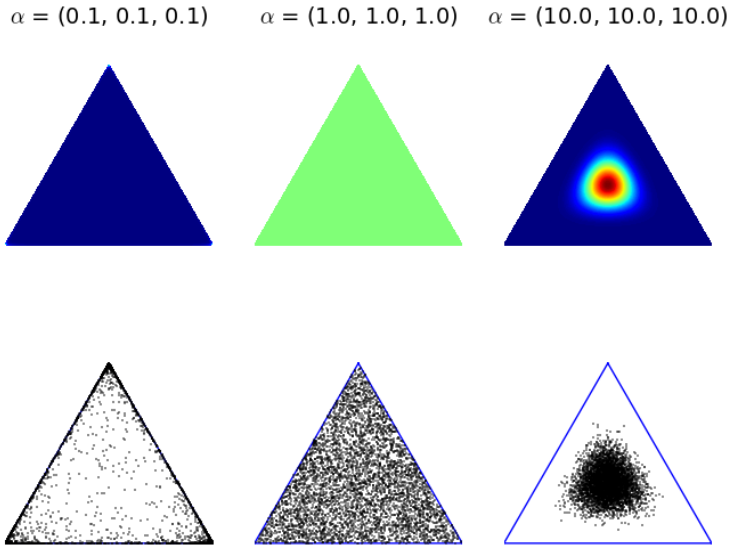
α = (1.0, 1.0, 1.0) α = (1.0, 3.0, 7.0) α = (5.0, 1.0, 1.0)



- The Dirichlet(1,1,1) distribution is the uniform distribution on the 2-simplex.

The Dirichlet distribution: A distribution over probability vectors

- The magnitude $\sum_k \alpha_k$ of the parameters acts as an inverse variance.
- Larger magnitude $\rightarrow$ more similar samples.
- Smaller magnitude $\rightarrow$ sparser samples.



The Dirichlet distribution: Conjugacy to the multinomial

THE UNIVERSITY OF TEXAS AT AUSTIN

- There are a number of distributions over probability vectors... but we like the Dirichlet because it is conjugate to the multinomial.
- If $\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$, then

$$p(\pi) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

The Dirichlet distribution: Conjugacy to the multinomial

THE UNIVERSITY OF TEXAS AT AUSTIN

- There are a number of distributions over probability vectors... but we like the Dirichlet because it is conjugate to the multinomial.
- If $\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$, then

$$p(\pi) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

- If $x_i \overset{iid}{\sim} \pi$ for $i = 1, \ldots, N$, then

$$p(x_1, \ldots, x_N | \pi) = \frac{N!}{m_1! \cdots m_K!} \prod_{k=1}^K \pi_k^{m_k}$$

where $m_k = \sum_i \mathbb{I}(x_i = k)$

The Dirichlet distribution: Conjugacy to the multinomial

- There are a number of distributions over probability vectors... but we like the Dirichlet because it is conjugate to the multinomial.
- If $\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$, then

$$p(\pi) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

- If $x_i \overset{iid}{\sim} \pi$ for $i = 1, \ldots, N$, then

$$p(x_1, \ldots, x_N | \pi) = \frac{N!}{m_1! \cdots m_K!} \prod_{k=1}^K \pi_k^{m_k}$$

  where $m_k = \sum_i \mathbb{I}(x_i = k)$
- So, the posterior takes the form

$$\begin{aligned} p(\pi | x_1, \ldots, x_N) &\propto p(x_1 \ldots x_N | \pi) p(\pi) \\ &\propto \frac{\Gamma(\sum_{k=1}^K \alpha_k + m_k)}{\prod_{k=1}^K \Gamma(\alpha_k + m_k)} \prod_{k=1}^K \pi_k^{\alpha_k + m_k - 1} \\ &= \text{Dirichlet}(\pi | \alpha_1 + m_1, \ldots, \alpha_K + m_K) \end{aligned}$$

Properties of the Dirichlet distribution: Combining and splitting elements

The Dirichlet distribution has a number of nice properties...

- We can get a $K-1$-dimensional Dirichlet distribution from a K-dimensional distribution.
- If

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

then

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

Properties of the Dirichlet distribution: Combining and splitting elements

The Dirichlet distribution has a number of nice properties...

- We can get a $K-1$-dimensional Dirichlet distribution from a K-dimensional distribution.
- If

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

then

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

- We can get a $K+1$-dimensional Dirichlet$(\alpha_1 b, \alpha_1(1-b), \alpha_2, \ldots, \alpha_K)$ distribution from a K-dimensional Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_K)$ distribution.
- If

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

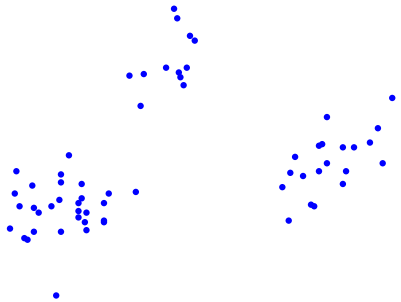
and

$$\theta \sim \text{Beta}(\alpha_1 b, \alpha_1(1-b)),\ 0 < b < 1$$

then

$$(\pi_1\theta_1, \pi_1(1-\theta_1), \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b, \alpha_1(1-b) \ldots, \alpha_K)$$

Returning to the Bayesian mixture of Gaussians



- Now we know about the Dirichlet distribution... we will return to our mixture of Gaussians.
  - Sample $\pi \sim$ Dirichlet$(\alpha_1, \alpha_2, \alpha_3)$
  - For each cluster, sample $\mu_k, \Sigma_k \sim$ Normal-inverse Wishart$(\mu_0, \lambda, \Psi, \nu)$
  - For the ith data point...
    - Sample a cluster indicator $z_i \sim \pi$.
    - Sample a location $x_i \sim$ Normal$(\mu_{z_i}, \Sigma_{z_i})$
- Instead of explicitly sampling π, we can integrate it out.

An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

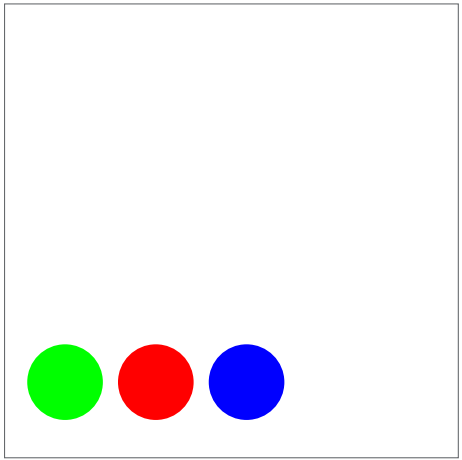
$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.

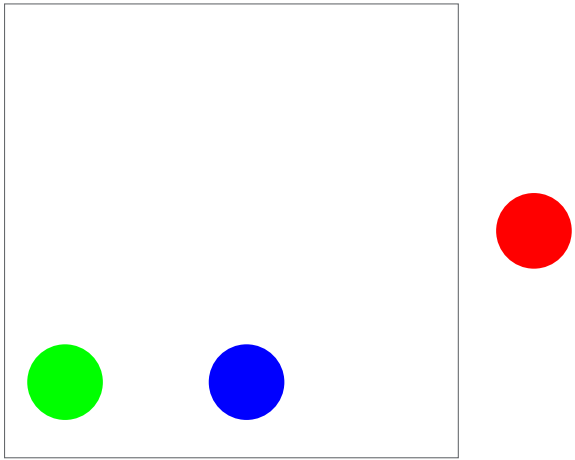


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.

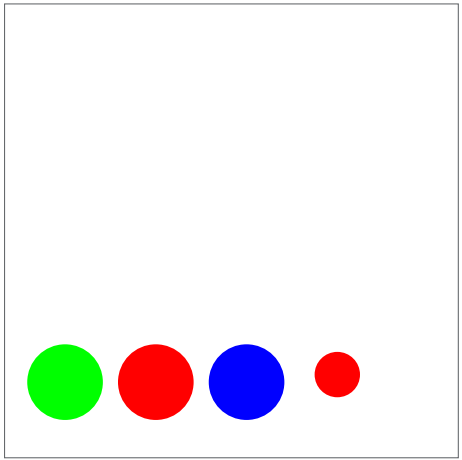


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.

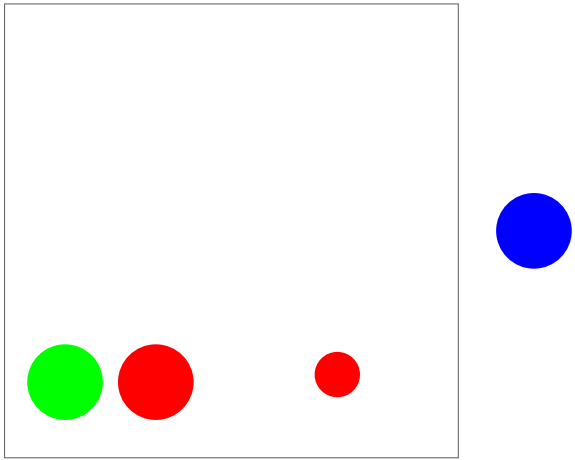


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.

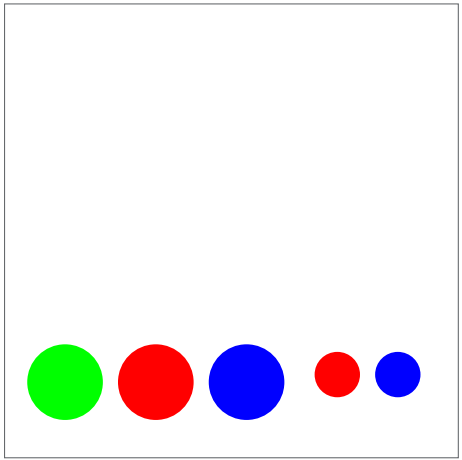


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.

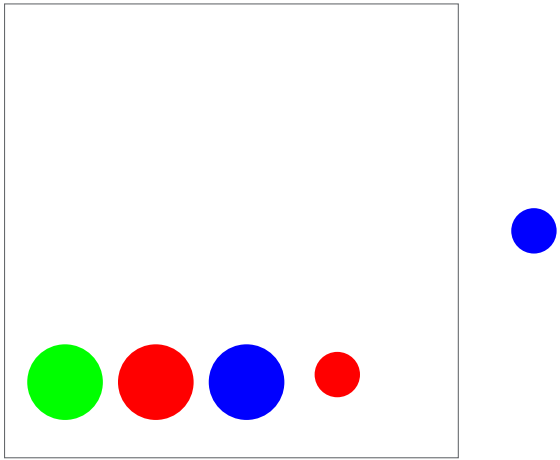


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.

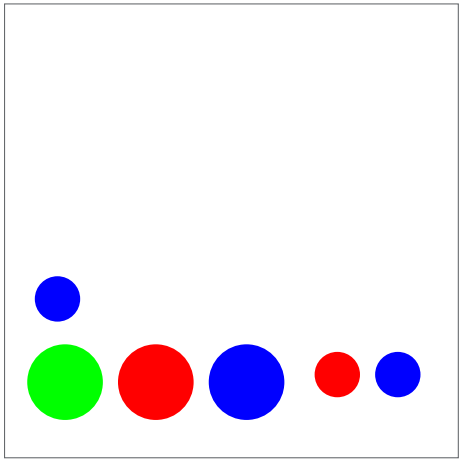


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.

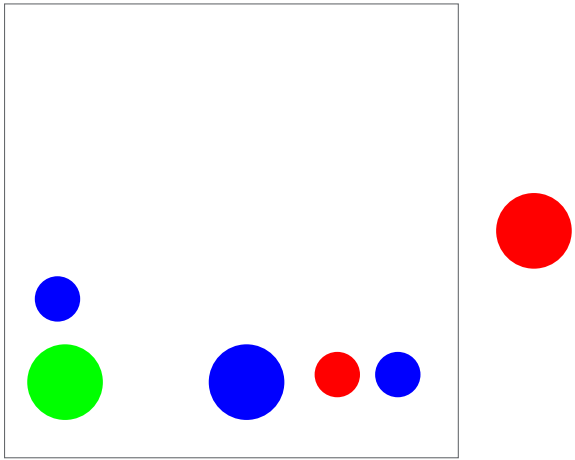


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.

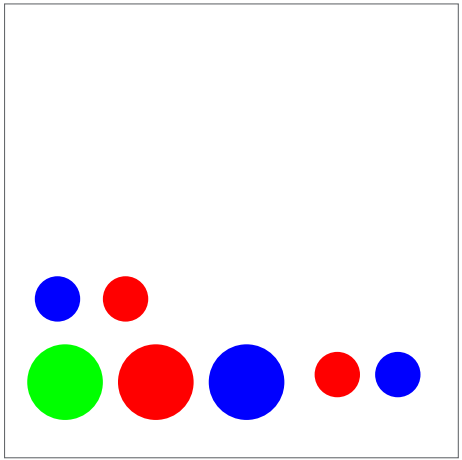


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.

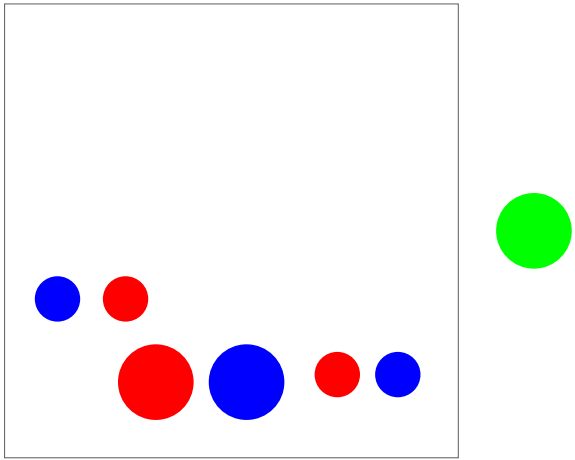


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.

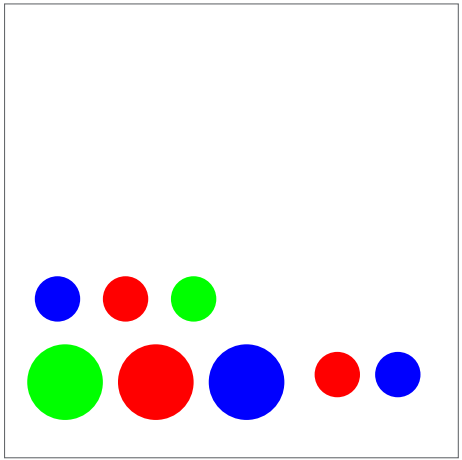


An urn representation

- Conditioned on π, the cluster indicators are independent: $p(z_i = k|\pi) = \pi_k$.
- When we integrate out π, they are no longer independent, and we have

$$p(z_i = k|z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- We can describe this using an urn analogy.
- Start with K different colored balls, each of size α_k.
- Pick a ball with probability proportional to its size.
- Return that ball, plus a unit-size ball of the same color.
- Repeat to build up dataset.



Exchangeability

- Does the probability of the ith ball being red depend on how many of the first $i-1$ balls are red?

Exchangeability

- Does the probability of the ith ball being red depend on how many of the first $i-1$ balls are red?
- Of course! More red balls $\rightarrow$ more likely to pick a red ball.
- The balls are *not* i.i.d.

Exchangeability

- Does the probability of the ith ball being red depend on how many of the first $i-1$ balls are red?
- Of course! More red balls $\rightarrow$ more likely to pick a red ball.
- The balls are *not* i.i.d.
- Does changing the order of the sequence matter? Does $p(r,r,r,b,g) = p(g,r,b,r,r)$?

Exchangeability

- Does the probability of the ith ball being red depend on how many of the first $i-1$ balls are red?
- Of course! More red balls $\rightarrow$ more likely to pick a red ball.
- The balls are *not* i.i.d.
- Does changing the order of the sequence matter? Does $p(r,r,r,b,g) = p(g,r,b,r,r)$?
- No! But this might not be as obvious... so we can double check

Exchangeability

- Does the probability of the ith ball being red depend on how many of the first $i-1$ balls are red?
- Of course! More red balls $\rightarrow$ more likely to pick a red ball.
- The balls are *not* i.i.d.
- Does changing the order of the sequence matter? Does $p(r,r,r,b,g) = p(g,r,b,r,r)$?
- No! But this might not be as obvious... so we can double check

$$p(r,r,r,b,g) = \frac{\alpha_r}{\sum_k \alpha_k} \frac{\alpha_r+1}{\sum_k \alpha_k + 1} \frac{\alpha_r+2}{\sum_k \alpha_k + 2} \frac{\alpha_b}{\sum_k \alpha_k + 3} \frac{\alpha_g}{\sum_k \alpha_k + 3}$$

Exchangeability

- Does the probability of the ith ball being red depend on how many of the first $i-1$ balls are red?
- Of course! More red balls $\rightarrow$ more likely to pick a red ball.
- The balls are *not* i.i.d.
- Does changing the order of the sequence matter? Does $p(r,r,r,b,g) = p(g,r,b,r,r)$?
- No! But this might not be as obvious... so we can double check

$$p(r,r,r,b,g) = \frac{\alpha_r}{\sum_k \alpha_k} \frac{\alpha_r+1}{\sum_k \alpha_k + 1} \frac{\alpha_r+2}{\sum_k \alpha_k + 2} \frac{\alpha_b}{\sum_k \alpha_k + 3} \frac{\alpha_g}{\sum_k \alpha_k + 3}$$

$$p(g,r,b,r,r) = \frac{\alpha_g}{\sum_k \alpha_k} \frac{\alpha_r}{\sum_k \alpha_k + 1} \frac{\alpha_b}{\sum_k \alpha_k + 2} \frac{\alpha_r+1}{\sum_k \alpha_k + 3} \frac{\alpha_r+2}{\sum_k \alpha_k + 3}$$

Exchangeability

- Does the probability of the ith ball being red depend on how many of the first $i-1$ balls are red?
- Of course! More red balls $\rightarrow$ more likely to pick a red ball.
- The balls are *not* i.i.d.
- Does changing the order of the sequence matter? Does $p(r,r,r,b,g) = p(g,r,b,r,r)$?
- No! But this might not be as obvious... so we can double check

$$p(r,r,r,b,g) = \frac{\alpha_r}{\sum_k \alpha_k} \frac{\alpha_r+1}{\sum_k \alpha_k + 1} \frac{\alpha_r+2}{\sum_k \alpha_k + 2} \frac{\alpha_b}{\sum_k \alpha_k + 3} \frac{\alpha_g}{\sum_k \alpha_k + 3}$$

$$p(g,r,b,r,r) = \frac{\alpha_g}{\sum_k \alpha_k} \frac{\alpha_r}{\sum_k \alpha_k + 1} \frac{\alpha_b}{\sum_k \alpha_k + 2} \frac{\alpha_r+1}{\sum_k \alpha_k + 3} \frac{\alpha_r+2}{\sum_k \alpha_k + 3}$$

- This property is known as exchangeability – the probability of a sequence is invariant to permutations

Why does exchangeability matter?

THE UNIVERSITY OF TEXAS AT AUSTIN

- Exchangeability allows us treat every data point as if it were the last one that we've seen.
- We know that

$$p(z_i = k | z_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathbb{I}(z_j = k) + \alpha_k}{i - 1 + \sum_k \alpha_k}$$

- Instead of just conditioning on the first $i-1$ data points, we can pretend the ith data point is actually the last one we saw, so that

$$p(z_i = k | z_{-i}) = \frac{\sum_{j \neq i} \mathbb{I}(z_j = k) + \alpha_k}{N - 1 + \sum_k \alpha_k}$$

- We can combine this with the cluster likelihood to get the posterior distribution

$$p(z_i = k | x_i z_{-i}, \{\mu_k\}, \{\Sigma_k\}) \propto \frac{\sum_{j \neq i} \mathbb{I}(z_j = k) + \alpha_k}{N - 1 + \sum_k \alpha_k} \text{Normal}(x_i | \mu_k, \Sigma_k)$$

- This makes it easy to construct a Gibbs sampler!

Collapsed inference using an urn scheme

This suggests a Gibbs sampler of the form:

- For $i = 1, \ldots, N$:
  - Sample the cluster allocation of the ith data point, given the conditional distribution

$$P(z_i = k | x_i, z_{-i}) \propto (m_k^{-i} + \alpha_k) \text{Normal}(x_i | \mu_k, \Sigma_k)$$

- For $k = 1 : K_+$:
  - Sample the cluster parameters from their conditional distribution

Choosing the number of clusters

THE UNIVERSITY OF TEXAS AT AUSTIN

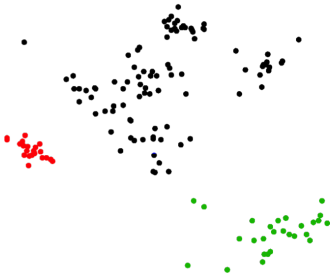
- The Dirichlet distribution is a great choice when there is a clear, fixed number of clusters... but sometimes that's not the case.
- Sometimes it's hard to tell what the right number of clusters is...

- Even if we have a good idea of how many clusters we have today... what if we see new clusters tomorrow?
- We should make sure we have more clusters than we actually need.

Choosing the number of clusters

THE UNIVERSITY OF TEXAS AT AUSTIN

- The Dirichlet distribution is a great choice when there is a clear, fixed number of clusters... but sometimes that's not the case.
- Sometimes it's hard to tell what the right number of clusters is...

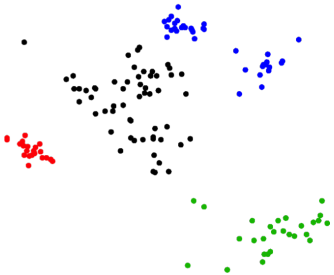


- Even if we have a good idea of how many clusters we have today... what if we see new clusters tomorrow?
- We should make sure we have more clusters than we actually need.

Choosing the number of clusters

THE UNIVERSITY OF TEXAS AT AUSTIN

- The Dirichlet distribution is a great choice when there is a clear, fixed number of clusters... but sometimes that's not the case.
- Sometimes it's hard to tell what the right number of clusters is...

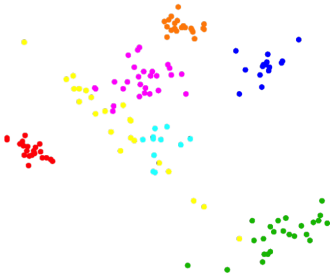


- Even if we have a good idea of how many clusters we have today... what if we see new clusters tomorrow?
- We should make sure we have more clusters than we actually need.

Choosing the number of clusters

THE UNIVERSITY OF TEXAS AT AUSTIN

- The Dirichlet distribution is a great choice when there is a clear, fixed number of clusters... but sometimes that's not the case.
- Sometimes it's hard to tell what the right number of clusters is...



- Even if we have a good idea of how many clusters we have today... what if we see new clusters tomorrow?
- We should make sure we have more clusters than we actually need.

Bayesian nonparametric mixture models

- The finite mixture model had K mixture components:

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{K} \pi_k \text{Normal}(x_n|\mu_k, \Sigma_k)$$

Bayesian nonparametric mixture models

- The finite mixture model had K mixture components:

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{K} \pi_k \text{Normal}(x_n|\mu_k, \Sigma_k)$$

- To make sure we never run out of clusters, no matter how many data points we see, we need (countably) infinite clusters!

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{\infty} \pi_k \text{Normal}(x_n|\mu_k, \Sigma_k)$$

- N data points will use at most N clusters.
- However, if some of the π_k's are bigger than others, there will probably be fewer than N.
- So, a finite data set will always use a finite—but random—number of clusters.

Bayesian nonparametric mixture models

- The finite mixture model had K mixture components:

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{K} \pi_k \text{Normal}(x_n|\mu_k, \Sigma_k)$$

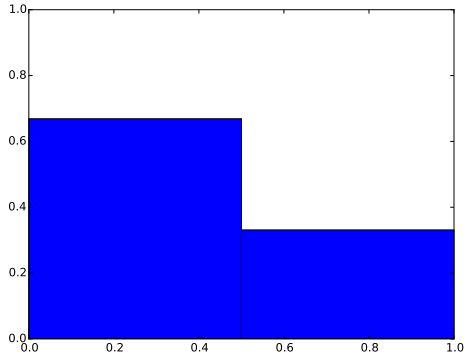
- To make sure we never run out of clusters, no matter how many data points we see, we need (countably) infinite clusters!

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{\infty} \pi_k \text{Normal}(x_n|\mu_k, \Sigma_k)$$

- N data points will use at most N clusters.
- However, if some of the π_k's are bigger than others, there will probably be fewer than N.
- So, a finite data set will always use a finite—but random—number of clusters.
- How to choose an appropriate prior?
- We want something *like* a Dirichlet prior... but with an infinite number of components.

Constructing an appropriate prior

- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$



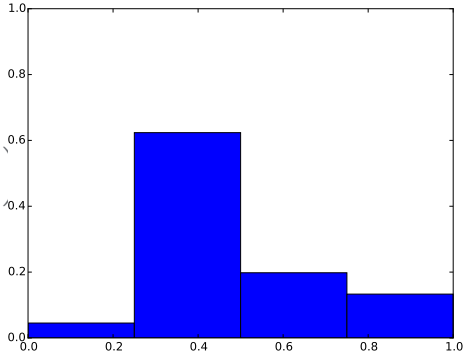
Constructing an appropriate prior

- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$



Constructing an appropriate prior

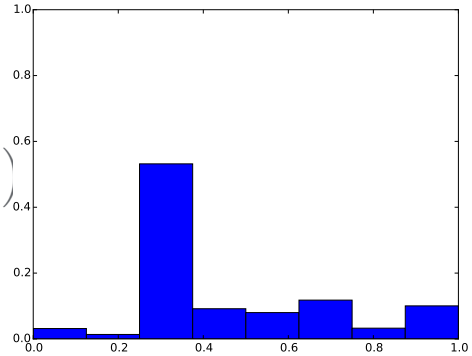
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.



Constructing an appropriate prior

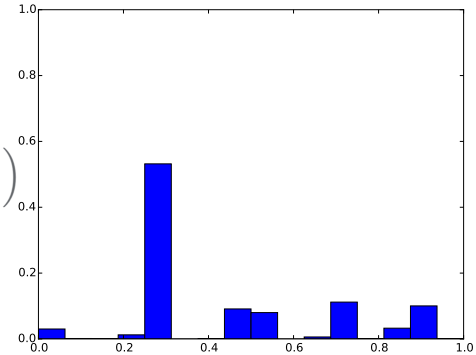
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.



Constructing an appropriate prior

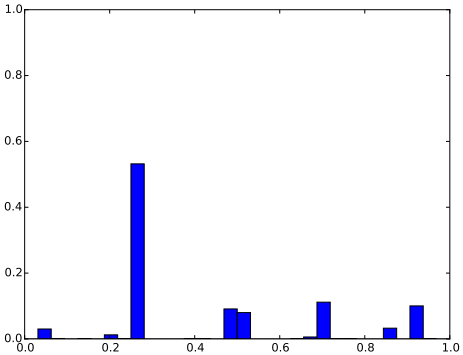
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.



Constructing an appropriate prior

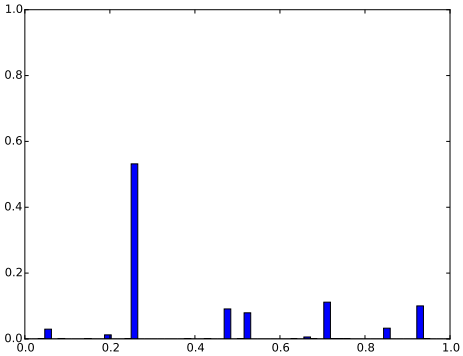
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.



Constructing an appropriate prior

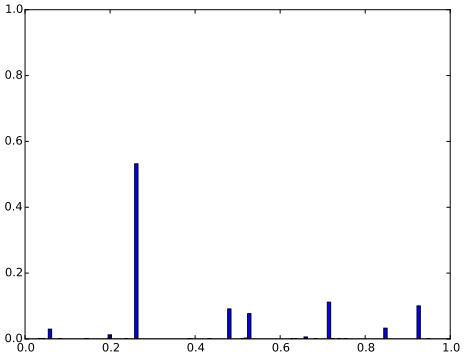
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.



Constructing an appropriate prior

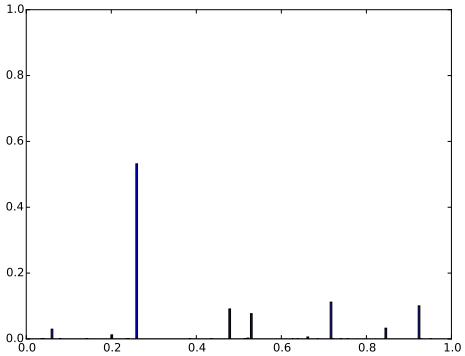
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.



Constructing an appropriate prior

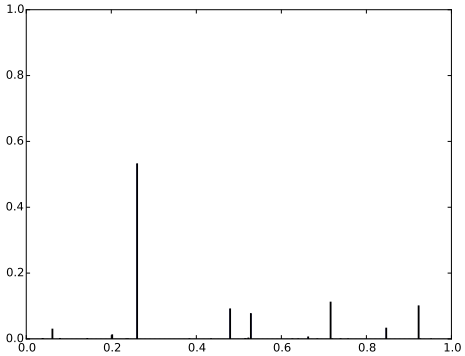
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.



Constructing an appropriate prior

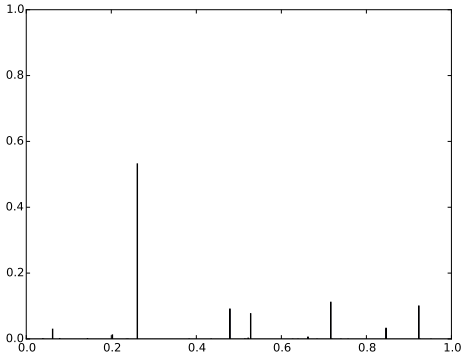
- Start off with
  $\boldsymbol{\pi}^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$
- Split each component according to our beta splitting rule:

$$\theta_1, \theta_2 \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

$$\boldsymbol{\pi}^{(4)} = \left(\theta_1\pi_1^{(2)}, (1-\theta_1)\pi_1^{(2)}, \theta_2\pi_2^{(2)}, (1-\theta_2)\pi_2^{(2)}\right)$$

$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

- Repeat to get
  $\boldsymbol{\pi}^{(K)} \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- As $K \to \infty$, we get a vector with infinitely many components.

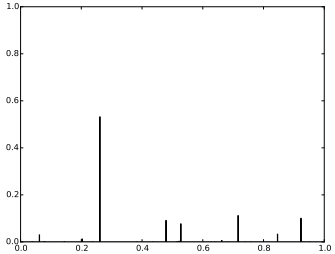


The Dirichlet process [Ferguson, 1973]

- We can combine this with a mechanism for generating parameter values.

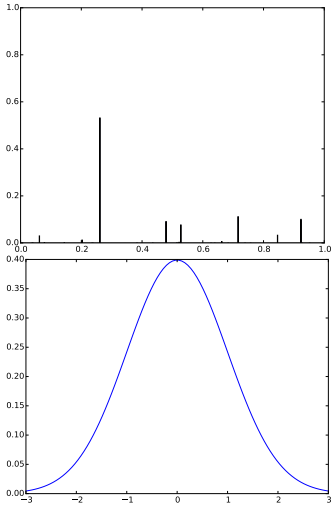
The Dirichlet process [Ferguson, 1973]

- We can combine this with a mechanism for generating parameter values.
- Let $\pi \sim \lim_{K\to\infty} \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$



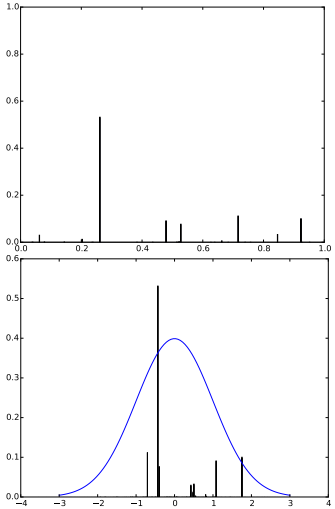
The Dirichlet process [Ferguson, 1973]

- We can combine this with a mechanism for generating parameter values.
- Let $\pi \sim \lim_{K\to\infty} \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$
- Let H be a distribution on some space Ω... e.g. a Gaussian distribution on the real line.



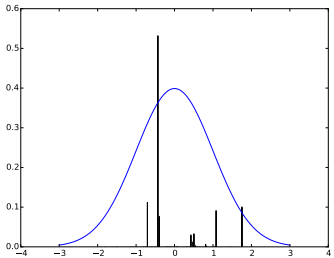
The Dirichlet process [Ferguson, 1973]

- We can combine this with a mechanism for generating parameter values.
- Let $\pi \sim \lim_{K\to\infty} \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$
- Let H be a distribution on some space Ω... e.g. a Gaussian distribution on the real line.
- For $k = 1, 2, \ldots,$ sample $\theta_k \sim H$
- Then $G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ is a probability distribution over Ω.
- Samples from the Dirichlet process are *discrete*. We call the point masses, *atoms*



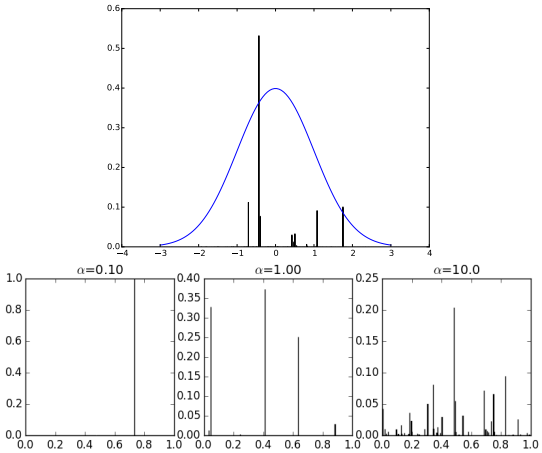
Samples from the Dirichlet process

- We write $G \sim DP(\alpha, H)$
- The **base measure** H determines the *locations* of the atoms.



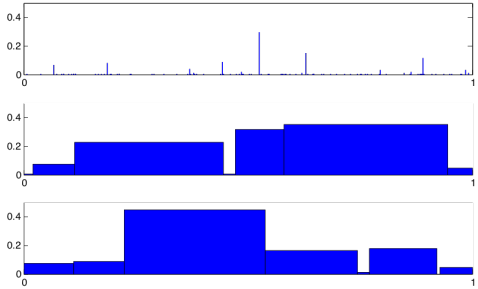
Samples from the Dirichlet process

- We write $G \sim DP(\alpha, H)$
- The **base measure** H determines the *locations* of the atoms.
- The **concentration parameter** α determines the distribution over atom sizes.
- Small values of α give sparser distributions



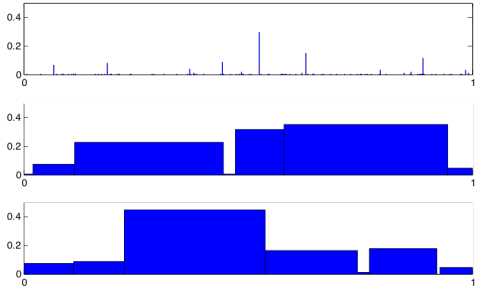
Dirichlet process and Dirichlet marginals

- Consider a Dirichlet process on $(0, 1)$ with uniform base measure H.
- Pick any partition $A_1, \ldots, A_K$ of $(0, 1)$, and sum up the atoms in each partition.



Dirichlet process and Dirichlet marginals

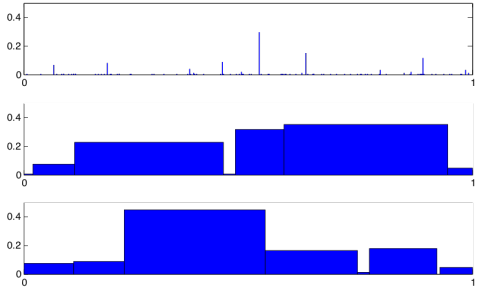
- Consider a Dirichlet process on $(0, 1)$ with uniform base measure H.
- Pick any partition $A_1, \ldots, A_K$ of $(0, 1)$, and sum up the atoms in each partition.



- Remember: If $(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$, then $(\pi_1 + \pi_2, \pi_3 \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$
- So, the weights assigned to the partition are $\text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$

Dirichlet process and Dirichlet marginals

- Consider a Dirichlet process on $(0, 1)$ with uniform base measure H.
- Pick any partition $A_1, \ldots, A_K$ of $(0, 1)$, and sum up the atoms in each partition.



- Remember: If $(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$, then $(\pi_1 + \pi_2, \pi_3 \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$
- So, the weights assigned to the partition are $\text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$
- This gives an alternative definition of the Dirichlet process: The (unique) distribution over Ω such that, for a partition $A_1, \ldots, A_K$ of Θ,

$$(P(A_1), \ldots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$

The Dirichlet process mixture model [Antoniak, 1974]

THE UNIVERSITY OF TEXAS AT AUSTIN

- We can use the Dirichlet process in place of the Dirichlet distribution to construct a mixture model with infinitely many clusters.
  - Sample a probability distribution $G \sim \text{DP}(\alpha, H)$ where H is a normal-inverse Wishart distribution.
  - This gives us $G = \sum_k \pi_k \delta_{\theta_k}$.

The Dirichlet process mixture model [Antoniak, 1974]

THE UNIVERSITY OF TEXAS AT AUSTIN

- We can use the Dirichlet process in place of the Dirichlet distribution to construct a mixture model with infinitely many clusters.
  - Sample a probability distribution $G \sim \text{DP}(\alpha, H)$ where H is a normal-inverse Wishart distribution.
  - This gives us $G = \sum_k \pi_k \delta_{\theta_k}$.
  - For each observation, sample a parameter $\phi_i := (\mu_i, \Sigma_i) \sim G$

The Dirichlet process mixture model [Antoniak, 1974]

THE UNIVERSITY OF TEXAS AT AUSTIN

- We can use the Dirichlet process in place of the Dirichlet distribution to construct a mixture model with infinitely many clusters.
  - Sample a probability distribution $G \sim \text{DP}(\alpha, H)$ where H is a normal-inverse Wishart distribution.
  - This gives us $G = \sum_k \pi_k \delta_{\theta_k}$.
  - For each observation, sample a parameter $\phi_i := (\mu_i, \Sigma_i) \sim G$
  - Equivalently, sample a cluster indicator $z_i \sim \pi$, and set $\phi_i = \theta_{z_i}$.

The Dirichlet process mixture model [Antoniak, 1974]

THE UNIVERSITY OF TEXAS AT AUSTIN

- We can use the Dirichlet process in place of the Dirichlet distribution to construct a mixture model with infinitely many clusters.
  - Sample a probability distribution $G \sim \text{DP}(\alpha, H)$ where H is a normal-inverse Wishart distribution.
  - This gives us $G = \sum_k \pi_k \delta_{\theta_k}$.
  - For each observation, sample a parameter $\phi_i := (\mu_i, \Sigma_i) \sim G$
  - Equivalently, sample a cluster indicator $z_i \sim \pi$, and set $\phi_i = \theta_{z_i}$.
  - Then, sample the observation $x_i \sim \text{Normal}(\mu_i, \Sigma_i)$

The Dirichlet process mixture model [Antoniak, 1974]

THE UNIVERSITY OF TEXAS AT AUSTIN

- We can use the Dirichlet process in place of the Dirichlet distribution to construct a mixture model with infinitely many clusters.
  - Sample a probability distribution $G \sim \text{DP}(\alpha, H)$ where H is a normal-inverse Wishart distribution.
  - This gives us $G = \sum_k \pi_k \delta_{\theta_k}$.
  - For each observation, sample a parameter $\phi_i := (\mu_i, \Sigma_i) \sim G$
  - Equivalently, sample a cluster indicator $z_i \sim \pi$, and set $\phi_i = \theta_{z_i}$.
  - Then, sample the observation $x_i \sim \text{Normal}(\mu_i, \Sigma_i)$
- The Dirichlet process has some similar properties to the Dirichlet distribution, that make inference feasible.

Conjugacy to the multinomial

THE UNIVERSITY OF TEXAS AT AUSTIN

- We saw that the Dirichlet distribution was conjugate to the multinomial.
- This is also true of the Dirichlet process!

- Pick a partition $A_1, \ldots, A_K$ of Ω, and let $P(A_k)$ be the mass assigned to A_k by $G \sim \text{Dirichlet}(\alpha, H)$.
- Then $(P(A_1), \ldots, P(A_k)) \sim \text{Dirichlet}\left(\alpha H(A_1), \ldots, \alpha H(A_K)\right)$.

Conjugacy to the multinomial

THE UNIVERSITY OF TEXAS AT AUSTIN

- We saw that the Dirichlet distribution was conjugate to the multinomial.
- This is also true of the Dirichlet process!

- Pick a partition $A_1, \ldots, A_K$ of Ω, and let $P(A_k)$ be the mass assigned to A_k by $G \sim \text{Dirichlet}(\alpha, H)$.
- Then $(P(A_1), \ldots, P(A_k)) \sim \text{Dirichlet}\left(\alpha H(A_1), \ldots, \alpha H(A_K)\right)$.
- If we see an observation in the jth segment, then we must have

$$(P(A_1), \ldots, P(A_j), \ldots, P(A_k)) \sim \text{Dirichlet}\left(\alpha H(A_1), \ldots, \alpha H(A_j) + 1, \ldots, \alpha H(A_K)\right)$$

Conjugacy to the multinomial

THE UNIVERSITY OF TEXAS AT AUSTIN

- We saw that the Dirichlet distribution was conjugate to the multinomial.
- This is also true of the Dirichlet process!

- Pick a partition $A_1, \ldots, A_K$ of Ω, and let $P(A_k)$ be the mass assigned to A_k by $G \sim \text{Dirichlet}(\alpha, H)$.
- Then $(P(A_1), \ldots, P(A_k)) \sim \text{Dirichlet}\left(\alpha H(A_1), \ldots, \alpha H(A_K)\right)$.
- If we see an observation in the jth segment, then we must have

$$(P(A_1), \ldots, P(A_j), \ldots, P(A_k)) \sim \text{Dirichlet}\left(\alpha H(A_1), \ldots, \alpha H(A_j) + 1, \ldots, \alpha H(A_K)\right)$$

- This must be true for *all possible partitions* of Ω.
- This is only possible if the posterior of G is given by

$$G|X_1 = x \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_x}{\alpha + 1}\right)$$

Predictive distribution

- Remember, for the Dirichlet distribution we could integrate out π to get $P(z_k = k|z_{-n}) \propto \sum_{i \neq j} \mathbb{I}(z_j = k) + \alpha_k$
- We can do something similar for the Dirichlet process!

- Let m_k be the number of times we have seen $X_i = \theta_k$ in the first n observations – or equivalently the number of times that $Z_i = k$ – and let K_+ be the number of values we've seen so far.
- The posterior distribution over G given n observations is

$$\text{DP}\left(\alpha + n, \frac{\alpha H + \sum_{k=1}^{K_+} m_k \delta_{\theta_k}}{\alpha + n}\right)$$

- So, we have

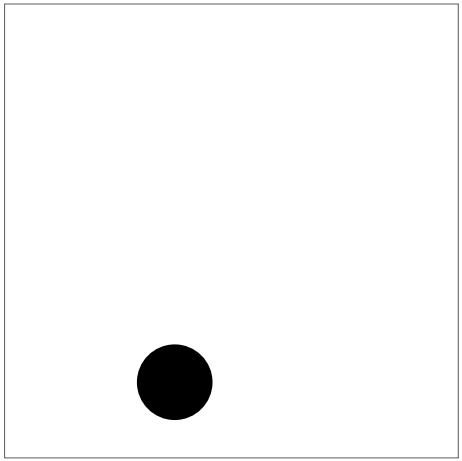
$$P(Z_{n+1} = k|Z_{1:n}) = \begin{cases} \frac{m_k}{n+\alpha} & \text{if } k \leq K_+ \\ \frac{\alpha}{n+\alpha} & \text{for new cluster} \end{cases}$$

- If we pick a new cluster, we sample it's parameter from H.

Polya urn scheme [Blackwell and MacQueen, 1973]

THE UNIVERSITY OF TEXAS AT AUSTIN

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.



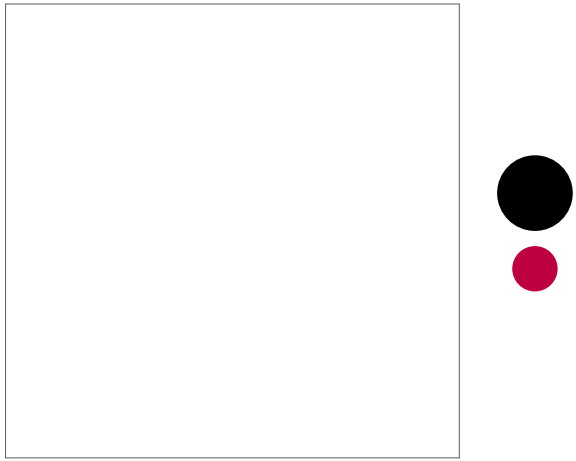
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.

Polya urn scheme [Blackwell and MacQueen, 1973]

THE UNIVERSITY OF TEXAS AT AUSTIN

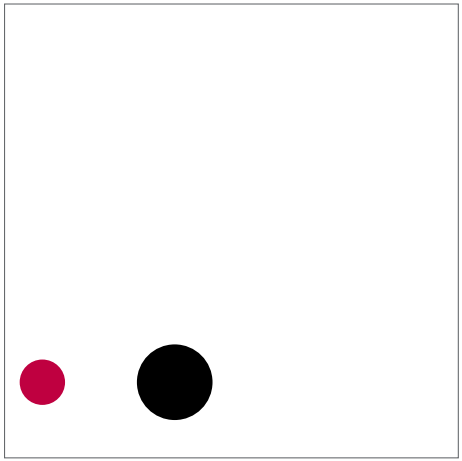
- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.



Polya urn scheme [Blackwell and MacQueen, 1973]

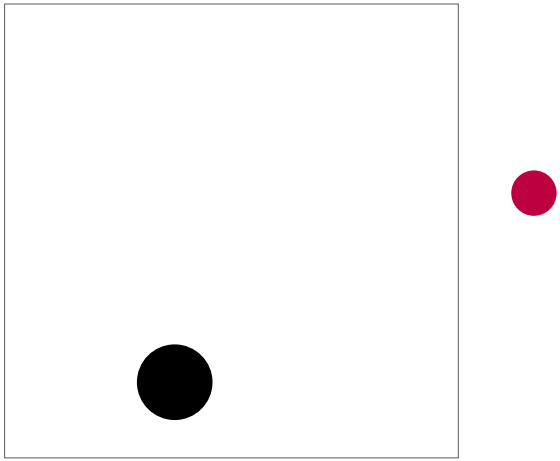
THE UNIVERSITY OF TEXAS AT AUSTIN

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.



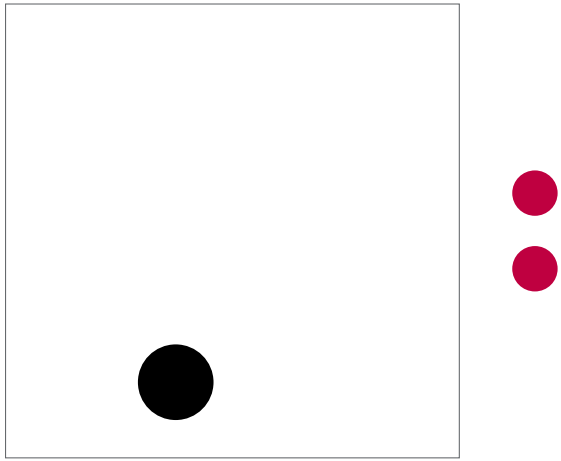
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



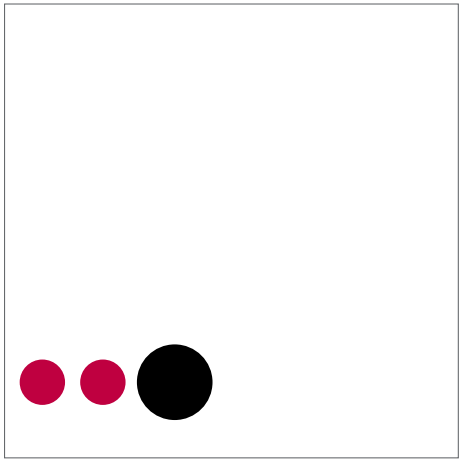
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



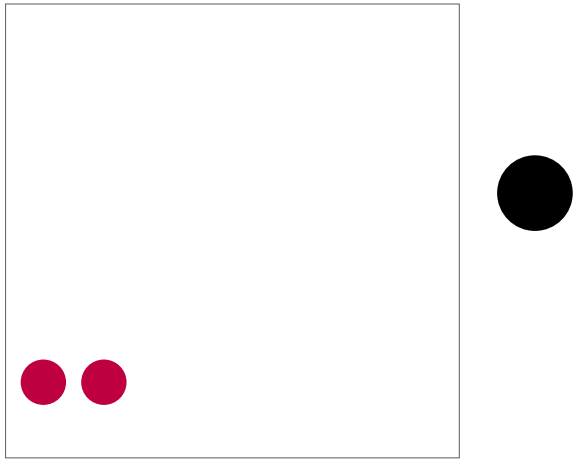
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



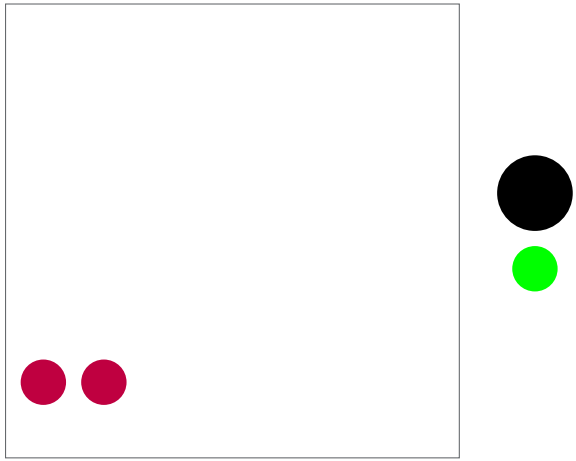
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



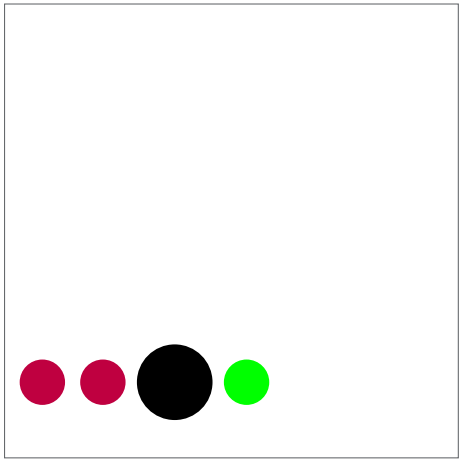
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



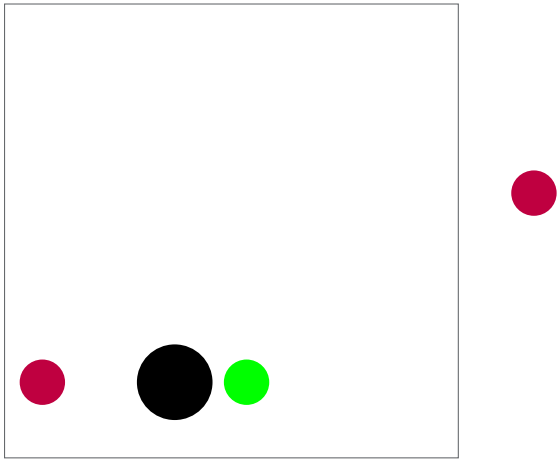
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



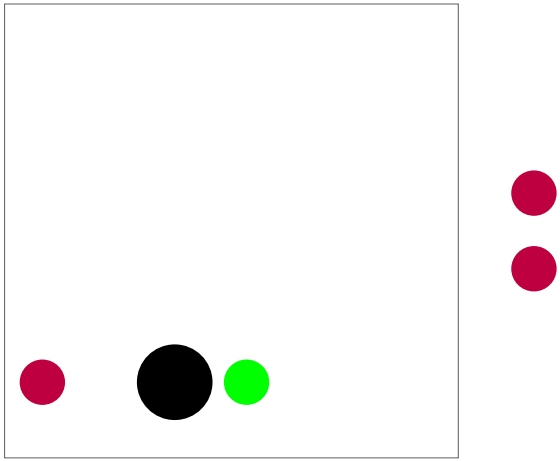
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



Polya urn scheme [Blackwell and MacQueen, 1973]

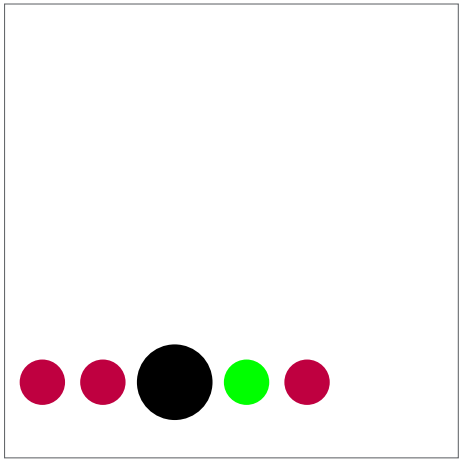
- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



Polya urn scheme [Blackwell and MacQueen, 1973]

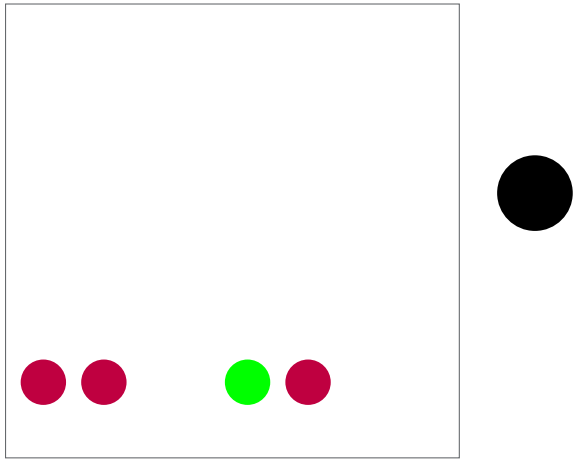
THE UNIVERSITY OF TEXAS AT AUSTIN

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



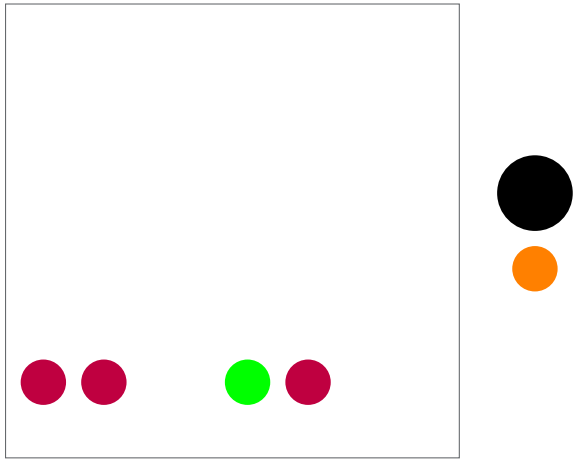
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



Polya urn scheme [Blackwell and MacQueen, 1973]

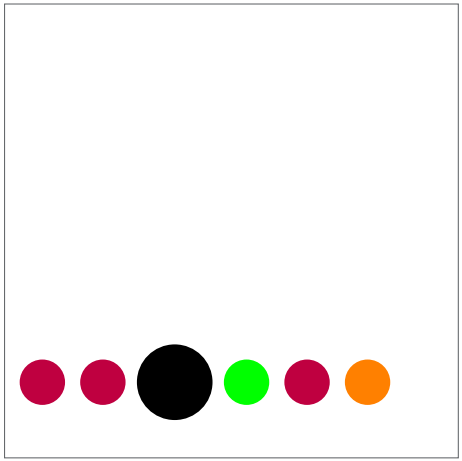
- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



Polya urn scheme [Blackwell and MacQueen, 1973]

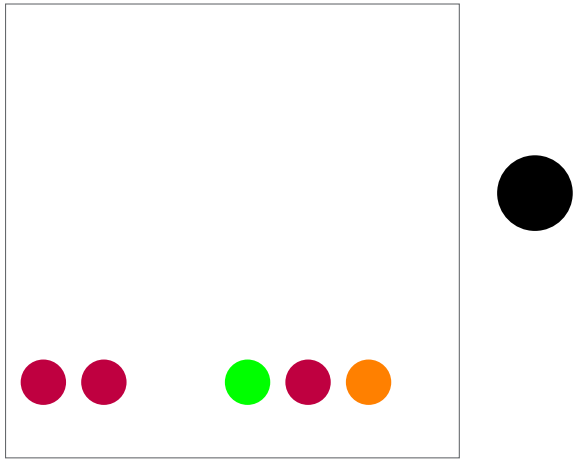
- Again, we can describe this using an urn analogy.

- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.

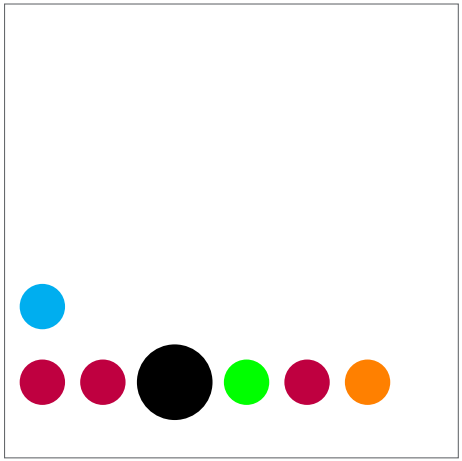


Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.

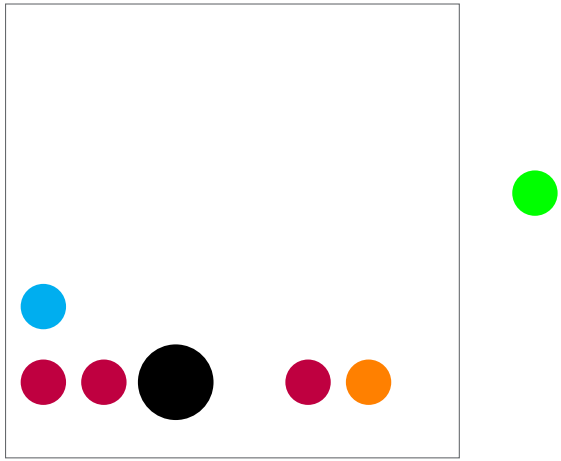
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



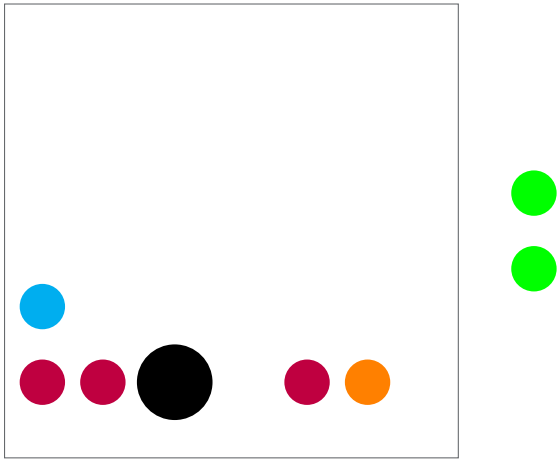
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



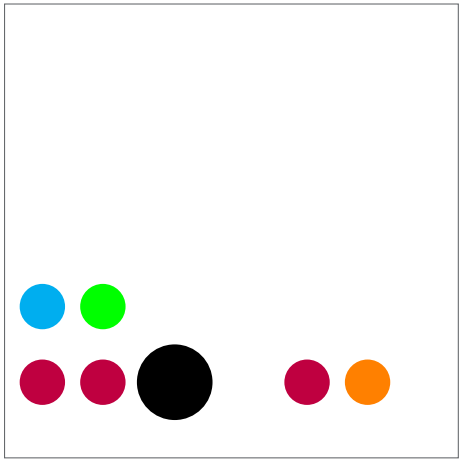
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



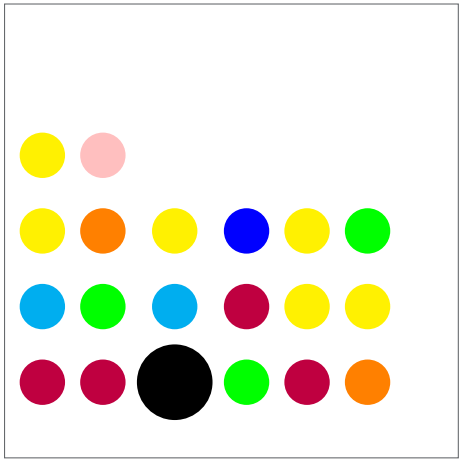
Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.



Polya urn scheme [Blackwell and MacQueen, 1973]

- Again, we can describe this using an urn analogy.
- Start with one black ball, of size α.
- Pick a ball with probability proportional to its size.
- If it's the black ball, sample a new color from H. Return the black ball plus a unit-size ball of the new color.
- If it's a colored ball, return that ball, plus a unit-size ball of the same color.
- Note, we can always sample a new color (the black ball is always there), but it gets less likely as N grows.



The Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.

- Imagine a restaurant with infinitely many tables, each serving a different dish.

The Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

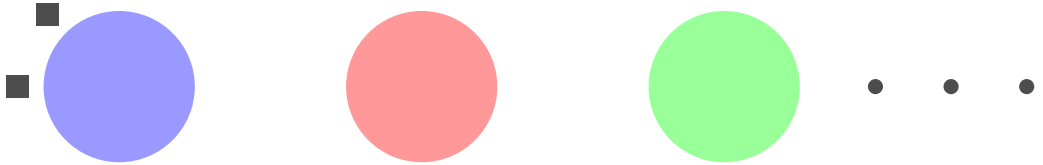
We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.

- Imagine a restaurant with infinitely many tables, each serving a different dish.
- The first customer comes to the restaurant, and sits at the first table.

The Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.

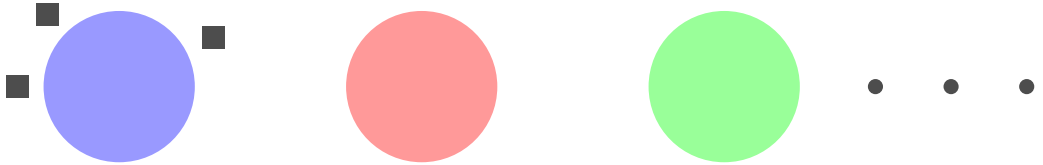


- Imagine a restaurant with infinitely many tables, each serving a different dish.
- The first customer comes to the restaurant, and sits at the first table.
- The second customer enters the restaurant. He sits at the first table with probability $\frac{1}{1+\alpha}$, or sits at a new table with probability $\frac{\alpha}{1+\alpha}$.

The Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.

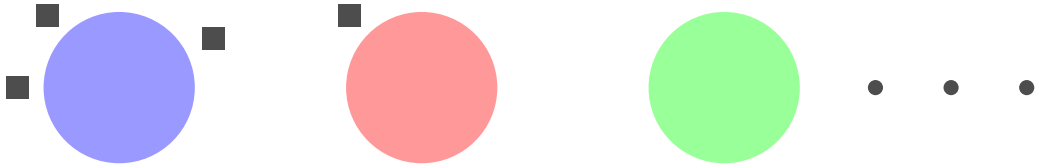


- Imagine a restaurant with infinitely many tables, each serving a different dish.
- The first customer comes to the restaurant, and sits at the first table.
- The second customer enters the restaurant. He sits at the first table with probability $\frac{1}{1+\alpha}$, or sits at a new table with probability $\frac{\alpha}{1+\alpha}$.
- Let m_k be the number of people sat at the kth table. The nth customer sits at the kth table with probability $\frac{m_k}{n-1+\alpha}$, or at a new table with probability $\frac{1}{n-1+\alpha}$.

The Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.

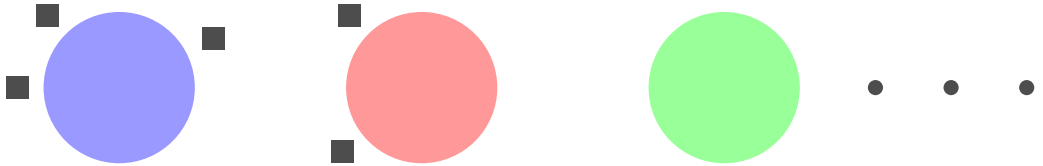


- Imagine a restaurant with infinitely many tables, each serving a different dish.
- The first customer comes to the restaurant, and sits at the first table.
- The second customer enters the restaurant. He sits at the first table with probability $\frac{1}{1+\alpha}$, or sits at a new table with probability $\frac{\alpha}{1+\alpha}$.
- Let m_k be the number of people sat at the kth table. The nth customer sits at the kth table with probability $\frac{m_k}{n-1+\alpha}$, or at a new table with probability $\frac{1}{n-1+\alpha}$.

The Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

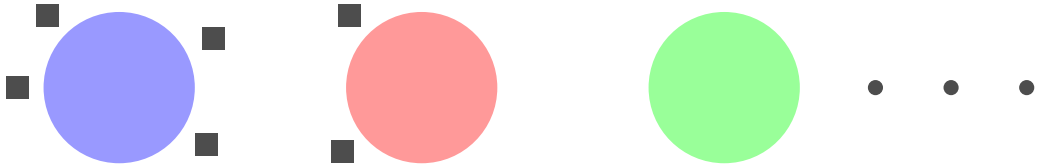
We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.



- Imagine a restaurant with infinitely many tables, each serving a different dish.
- The first customer comes to the restaurant, and sits at the first table.
- The second customer enters the restaurant. He sits at the first table with probability $\frac{1}{1+\alpha}$, or sits at a new table with probability $\frac{\alpha}{1+\alpha}$.
- Let m_k be the number of people sat at the kth table. The nth customer sits at the kth table with probability $\frac{m_k}{n-1+\alpha}$, or at a new table with probability $\frac{1}{n-1+\alpha}$.

The Chinese restaurant process

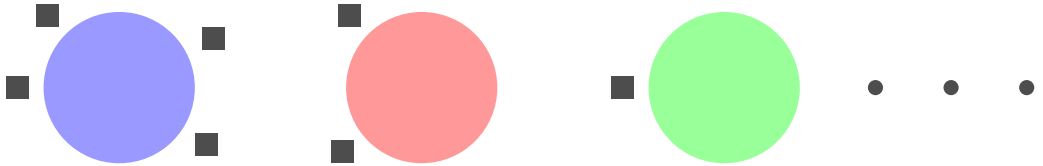
We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.



- Imagine a restaurant with infinitely many tables, each serving a different dish.
- The first customer comes to the restaurant, and sits at the first table.
- The second customer enters the restaurant. He sits at the first table with probability $\frac{1}{1+\alpha}$, or sits at a new table with probability $\frac{\alpha}{1+\alpha}$.
- Let m_k be the number of people sat at the kth table. The nth customer sits at the kth table with probability $\frac{m_k}{n-1+\alpha}$, or at a new table with probability $\frac{1}{n-1+\alpha}$.

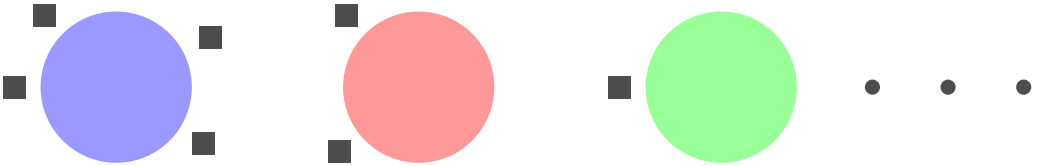
The Chinese restaurant process

We can also describe a sample from a DP-distributed probability distribution in terms of the following restaurant metaphor.



- Imagine a restaurant with infinitely many tables, each serving a different dish.
- The first customer comes to the restaurant, and sits at the first table.
- The second customer enters the restaurant. He sits at the first table with probability $\frac{1}{1+\alpha}$, or sits at a new table with probability $\frac{\alpha}{1+\alpha}$.
- Let m_k be the number of people sat at the kth table. The nth customer sits at the kth table with probability $\frac{m_k}{n-1+\alpha}$, or at a new table with probability $\frac{1}{n-1+\alpha}$.

The Chinese restaurant process



- We tend to sit at popular tables! This is known as the "rich-get-richer" property.
- We can always add new tables – *nonparametric*.
- For a given number of customers, the number of clusters is random.
- We can show that the probability of an assignment of people to tables is exchangeable (if we ignore the table ordering).

Inference using the Chinese restaurant process

- Since the cluster assignments are exchangeable, we can treat each customer as if he is the last.
- So, conditioned on the other $N-1$ observations, we know that the prior predictive probability of the ith data point (customer) being in cluster (table) k is

$$P(z_i = k | z_{-i}) \propto \begin{cases} m_k^{-i} & k \leq K_+ \\ \alpha & \text{new cluster} \end{cases}$$

Inference using the Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

- Since the cluster assignments are exchangeable, we can treat each customer as if he is the last.
- So, conditioned on the other $N-1$ observations, we know that the prior predictive probability of the ith data point (customer) being in cluster (table) k is

$$P(z_i = k|z_{-i}) \propto \begin{cases} m_k^{-i} & k \leq K_+ \\ \alpha & \text{new cluster} \end{cases}$$

- Each cluster (table) k is associated with a parameter (dish) θ_k – e.g. the mean and covariance of a Gaussian.
- So, the conditional probability of the ith data point x_i being in cluster k is:

$$P(z_i = k|x_i, z_{-i}) \propto \begin{cases} m_k^{-i} f(x_i; \theta_k) & k \leq K_+ \\ \alpha \int f(x_i; \theta) dH(\theta) & \text{new cluster} \end{cases}$$

where $f(x; \theta)$ is the appropriate likelihood model.

Collapsed inference using the Chinese restaurant process

THE UNIVERSITY OF TEXAS AT AUSTIN

This suggests a Gibbs sampler of the form:

- For $i = 1, \ldots, N$:
  - Sample the cluster allocation of the ith data point, given the conditional distribution

$$P(z_i = k | x_i, z_{-i}) \propto \begin{cases} m_k^{-i} f(x_i; \theta_k) & k \leq K_+ \\ \alpha \int f(x_i; \theta) dH(\theta) & \text{new cluster} \end{cases}$$

  - If the number of clusters grow or shrink, adjust our representation accordingly (add/delete clusters)
- For $k = 1 : K_+$:
  - Sample the cluster parameters from their conditional distribution (unless they are integrated out)

Problems with the collapsed sampler

THE UNIVERSITY OF TEXAS AT AUSTIN

The collapsed sampler is easy to implement – but can have some problems

- We are only updating one data point at a time.
- Imagine two "true" clusters are merged into a single cluster – a single data point is unlikely to "break away".
- Getting to the true distribution involves going through low probability states, so mixing can be slow.
- If the likelihood is not conjugate, integrating out parameter values for new features can be difficult.

- An alternative approach is to instantiate G, so we can update multiple data points at once.
- Problem: G is infinite-dimensional!
- Luckily, there is a nice representation that can help us...

Stick breaking construction [Sethuraman, 1994]

THE UNIVERSITY OF TEXAS AT AUSTIN

- Imagine a stick of unit length, representing the total probability.
- For $k = 1, 2, \ldots$
  - Sample a Beta$(1, \alpha)$ random variable b_k
  - Break off a fraction b_k of the stick. This is the first atom.
  - Sample a random location for this atom.
  - Recurse on the remaining stick to get:

$$b_k \sim \text{Beta}(1, \alpha) \qquad \pi_k = b_k \prod_{j=1}^{k-1}(1 - b_k) \qquad \theta_k \sim H \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

Stick breaking construction [Sethuraman, 1994]

THE UNIVERSITY OF TEXAS AT AUSTIN

- Imagine a stick of unit length, representing the total probability.
- For $k = 1, 2, \ldots$
  - Sample a Beta$(1, \alpha)$ random variable b_k
  - Break off a fraction b_k of the stick. This is the first atom.
  - Sample a random location for this atom.
  - Recurse on the remaining stick to get:

$$b_k \sim \text{Beta}(1, \alpha) \qquad \pi_k = b_k \prod_{j=1}^{k-1}(1 - b_k) \qquad \theta_k \sim H \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

- We can use the b_k directly to obtain the cluster assignment.
- Starting at the first cluster, choose cluster k with probability b_k, else move on to the next cluster.

Blocked Gibbs sampling using the stick breaking construction

THE UNIVERSITY OF TEXAS AT AUSTIN

This gives us an alternative inference approach.

- Pick a truncation K, so we are working with an approximation to the DP,

$$b_k \sim \text{Beta}(1, \alpha) \qquad \pi_k = b_k \prod_{j=1}^{k-1}(1 - b_k) \qquad \theta_k \sim H \qquad G^K = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$$

- Conditioned on G_k, we can sample a cluster assignment using

$$P(z_i = k | G^K, x_i) \propto \pi_k f(x_i | \theta_k)$$

- Contitioned on the cluster allocations, we can update each b_k.
- Remember, b_k is the probability of belonging to the kth cluster, conditioned on not belonging to any previous clusters.
- So,

$$b_k | \mathbf{z} \sim \text{Beta}\left(1 + m_k, \alpha + \sum_{j=k+1}^{K} m_j\right)$$

Alternative inference approaches

THE UNIVERSITY OF TEXAS AT AUSTIN

These are just a few of the inference methods we can use. Here are a taste of some other methods.

- Slice sampling: Adapt the batch sampler to have a random truncation $\rightarrow$ increases computational cost, but removes truncation error.
- Split/merge sampler: Adapt the collapsed sampler so that we can split and combine clusters $\rightarrow$ increases computational cost, but much faster mixing.
- Weak limit sampling: Approximate the DP with a very big Dirichlet distribution $\rightarrow$ easy, but not really nonparametric.
- Variational Bayes $\rightarrow$ less accurate, but much faster.
- MadBayes (basically, a nonparametric version of K-means) $\rightarrow$ less accurate, but much faster.

Summary

THE UNIVERSITY OF TEXAS AT AUSTIN

- We've introduced the Dirichlet process, which we can think of as an infinitely large Dirichlet distribution.
- We've explored different ways of representing the DP:
  - Dirichlet marginals
  - Urns
  - Chinese restaurant process
  - Stick-breaking construction
- We've explored the main ways of doing inference... if you can code up a collapsed Gibbs sampler for a Dirichlet mixture of Gaussians, you should be able to code up a sampler for a DP mixture of Gaussians.
- This afternoon we'll look at a different nonparametric model: The Indian buffet process.

References

Antoniak, C. (1974).
Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.
The Annals of Statistics, pages 1152–1174.

Blackwell, D. and MacQueen, J. B. (1973).
Ferguson distributions via Pólya urn schemes.
The annals of statistics, pages 353–355.

Ferguson, T. S. (1973).
A Bayesian analysis of some nonparametric problems.
The Annals of Statistics, pages 209–230.

Sethuraman, J. (1994).
A constructive definition of Dirichlet priors.
Statistica sinica, pages 639–650.