

An Introduction to Bayesian Nonparametric Modelling

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

March 6, 2009 / MLII



Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Indian Buffet and Beta Processes

Completely Random Measure

Summary

Bibliography

Modelling Data

All models are wrong, but some are useful.

—George E. P. Box, Norman R. Draper (1987).

- ▶ Models are never correct for real world data.
- ▶ How do we deal with model misfit?
 1. Quantify closeness to true model, and optimality of fitted model;
 2. Model selection or averaging;
 3. Increase the flexibility of your model class.

Nonparametric Modelling

- ▶ What is a nonparametric model?
 - ▶ A parametric model where the number of parameters increases with data;
 - ▶ A really large parametric model;
 - ▶ A model over infinite dimensional function or measure spaces.
 - ▶ A family of distributions that is dense in some large space.
- ▶ Why nonparametric models in Bayesian theory of learning?
 - ▶ broad class of priors that allows data to “speak for itself”;
 - ▶ side-step model selection and averaging.
- ▶ How do we deal with the infinite parameter space?
 - ▶ Marginalize out all but a finite number of parameters;
 - ▶ Define infinite space implicitly (akin to the kernel trick) using either Kolmogorov Consistency Theorem or de Finetti’s theorem.

Classification and Regression

- ▶ Learn a mapping $f : \mathbb{X} \rightarrow \mathbb{Y}$.

Data: Pairs of data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Model: $y_i | x_i, w \sim F(\cdot | x_i, w)$

Classification: $\mathbb{Y} = \{+1, -1\}$ or $\{1, \dots, C\}$.

Regression: $\mathbb{Y} = \mathbb{R}$

- ▶ Prior over parameters

$$p(w)$$

- ▶ Posterior over parameters

$$p(w | \mathbf{x}, \mathbf{y}) = \frac{p(w)p(\mathbf{y} | \mathbf{x}, w)}{p(\mathbf{y} | \mathbf{x})}$$

- ▶ Prediction with posterior:

$$p(y_* | x_*, \mathbf{x}, \mathbf{y}) = \int p(y_* | x_*, w) p(w | \mathbf{x}, \mathbf{y}) dw$$

Nonparametric Classification and Regression

- ▶ Learn a mapping $f : \mathbb{X} \rightarrow \mathbb{Y}$.

Data: Pairs of data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Model: $y_i | x_i, w \sim F(\cdot | x_i, f)$

Classification: $\mathbb{Y} = \{+1, -1\}$ or $\{1, \dots, C\}$.

Regression: $\mathbb{Y} = \mathbb{R}$

- ▶ Prior over parameters

$$p(f)$$

- ▶ Posterior over parameters

$$p(f | \mathbf{x}, \mathbf{y}) = \frac{p(f)p(\mathbf{y} | \mathbf{x}, f)}{p(\mathbf{y} | \mathbf{x})}$$

- ▶ Prediction with posterior:

$$p(y_* | x_*, \mathbf{x}, \mathbf{y}) = \int p(y_* | x_*, f) p(f | \mathbf{x}, \mathbf{y}) df$$

Density Estimation

- ▶ Parametric density estimation (e.g. Gaussian, mixture models)

Data: $\mathbf{x} = \{x_1, x_2, \dots\}$

Model: $p(x_i|w) = g(x_i|w)$

- ▶ Prior over parameters

$$p(w)$$

- ▶ Posterior over parameters

$$p(w|\mathbf{x}) = \frac{p(w)p(\mathbf{x}|w)}{p(\mathbf{x})}$$

- ▶ Prediction with posterior

$$p(x_*|\mathbf{x}) = \int p(x_*|w)p(w|\mathbf{x}) dw$$

Nonparametric Density Estimation

- ▶ Nonparametric density estimation

Data: $\mathbf{x} = \{x_1, x_2, \dots\}$

Model: $p(x_i|f) = f(x_i)$

- ▶ Prior over densities

$$p(f)$$

- ▶ Posterior over densities

$$p(f|\mathbf{x}) = \frac{p(f)p(\mathbf{x}|f)}{p(\mathbf{x})}$$

- ▶ Prediction with posterior

$$p(x_*|\mathbf{x}) = \int f(x_*)p(f|\mathbf{x}) df$$

Semiparametric Modelling

- ▶ Linear regression model for inferring effectiveness of new medical treatments.

$$y_{ij} = \beta^\top x_{ij} + \mathbf{b}_i^\top \mathbf{z}_{ij} + \epsilon_{ij}$$

y_{ij} is outcome of j th trial on i th subject.

x_{ij}, \mathbf{z}_{ij} are predictors (treatment, dosage, age, health...).

β are fixed-effects coefficients.

\mathbf{b}_i are random-effects subject-specific coefficients.

ϵ_{ij} are noise terms.

- ▶ Care about inferring β . If x_{ij} is treatment, we want to determine $p(\beta > 0 | \mathbf{x}, \mathbf{y}, \mathbf{z})$.
- ▶ Usually we assume Gaussian noise $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Is this a sensible prior? Over-dispersion, skewness,...
- ▶ May be better to model noise nonparametrically: $\epsilon_{ij} \sim F$.
- ▶ Also possible to model subject-specific random effects nonparametrically: $\mathbf{b}_i \sim G$.

Model Selection/Averaging

- ▶ Data: $\mathbf{x} = \{x_1, x_2, \dots\}$
Models: $p(\theta_k|M_k)$, $p(\mathbf{x}|\theta_k, M_k)$
- ▶ Marginal likelihood

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k) d\theta_k$$

- ▶ Model selection

$$M = \operatorname{argmax}_{M_k} p(\mathbf{x}|M_k)$$

- ▶ Model averaging

$$p(x_*|\mathbf{x}) = \sum_{M_k} p(x_*|M_k)p(M_k|\mathbf{x}) = \sum_{M_k} p(x_*|M_k) \frac{p(\mathbf{x}|M_k)p(M_k)}{p(\mathbf{x})}$$

- ▶ But: is this computationally feasible?

Model Selection/Averaging

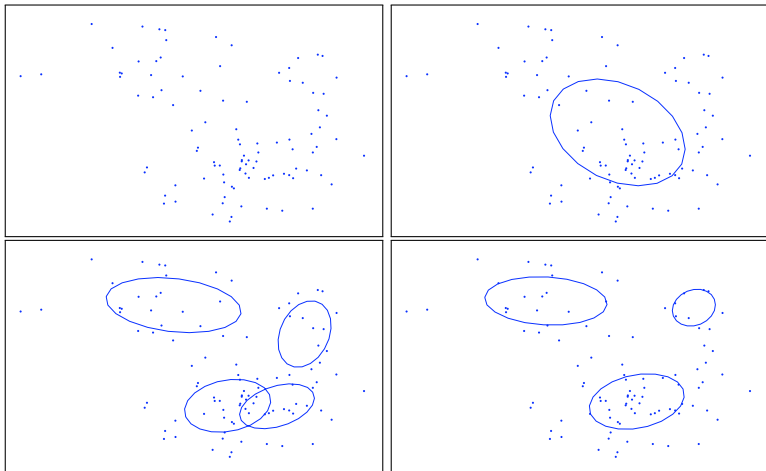
- ▶ Marginal likelihood is usually extremely hard to compute.

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k) d\theta_k$$

- ▶ Model selection/averaging is to prevent underfitting and overfitting.
- ▶ But reasonable and proper Bayesian methods should not overfit [Rasmussen and Ghahramani 2001].
- ▶ Use a really large model M_∞ instead, and *let the data speak for themselves*.

Model Selection/Averaging

How many clusters are there?



Other Tutorials on Bayesian Nonparametrics

- ▶ Zoubin Ghahramani, UAI 2005.
- ▶ Michael Jordan, NIPS 2005.
- ▶ Volker Tresp, ICML nonparametric Bayes workshop 2006.
- ▶ Workshop on Bayesian Nonparametric Regression, Cambridge, July 2007.
- ▶ My Machine Learning Summer School 2007 tutorial and practical course.
- ▶ I have an introduction to Dirichlet processes [Teh 2007], and another to hierarchical Bayesian nonparametric models [Teh and Jordan 2009].

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Indian Buffet and Beta Processes

Completely Random Measure

Summary

Bibliography

A Tiny Bit of Measure Theoretic Probability Theory

- ▶ A σ -*algebra* Σ is a family of subsets of a set Θ such that
 - ▶ Σ is not empty;
 - ▶ If $A \in \Sigma$ then $\Theta \setminus A \in \Sigma$;
 - ▶ If $A_1, A_2, \dots \in \Sigma$ then $\bigcup_{i=1}^{\infty} A_i \in \Sigma$.
- ▶ (Θ, Σ) is a *measure space* and $A \in \Sigma$ are the *measurable sets*.
- ▶ A *measure* μ over (Θ, Σ) is a function $\mu : \Sigma \rightarrow [0, \infty]$ such that
 - ▶ $\mu(\emptyset) = 0$;
 - ▶ If $A_1, A_2, \dots \in \Sigma$ are disjoint then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.
 - ▶ Everything we consider here will be measurable.
 - ▶ A probability measure is one where $\mu(\Theta) = 1$.
- ▶ Given two measure spaces (Θ, Σ) and (Δ, Φ) , a function $f : \Theta \rightarrow \Delta$ is *measurable* if $f^{-1}(A) \in \Sigma$ for every $A \in \Phi$.

A Tiny Bit of Measure Theoretic Probability Theory

- ▶ If p is a probability measure on (Θ, Σ) , a *random variable* X taking values in Δ is simply a measurable function $X : \Theta \rightarrow \Delta$.
 - ▶ Think of the probability space (Θ, Σ, p) as a black-box random number generator, and X as a function taking random samples in Θ and producing random samples in Δ .
 - ▶ The probability of an event $A \in \Phi$ is $p(X \in A) = p(X^{-1}(A))$.
- ▶ A *stochastic process* is simply a collection of random variables $\{X_i\}_{i \in \mathbb{I}}$ over the same measure space (Θ, Σ) , where \mathbb{I} is an index set.
 - ▶ What distinguishes a stochastic process from, say, a graphical model is that \mathbb{I} can be infinite, even uncountably so.
 - ▶ This raises issues of how do you even define them and how do you ensure that they can even exist (mathematically speaking).
- ▶ Stochastic processes form the core of many Bayesian nonparametric models.
 - ▶ Gaussian processes, Poisson processes, gamma processes, Dirichlet processes, beta processes...

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Consistency

Poisson Processes

Gamma Processes

Dirichlet Processes

Indian Buffet and Beta Processes

Completely Random Measure

Summary

Bibliography

Gaussian Processes

- ▶ A *Gaussian process* (GP) is a random function $f : \mathbb{X} \rightarrow \mathbb{R}$ such that for any finite set of input points x_1, \dots, x_n ,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} c(x_1, x_1) & \dots & c(x_1, x_n) \\ \vdots & \ddots & \vdots \\ c(x_n, x_1) & \dots & c(x_n, x_n) \end{bmatrix} \right)$$

where the parameters are the mean function $m(x)$ and covariance kernel $c(x, y)$.

- ▶ GPs can be visualized by iterative sampling $f(x_n) | f(x_1), \dots, f(x_{n-1})$ on a sequence of input points x_1, x_2, \dots
 - ▶ Demonstration.
- ▶ Note: a random function f is a stochastic process. It is a collection of random variables $\{f(x)\}_{x \in \mathbb{X}}$ one for each possible input value x .

[Rasmussen and Williams 2006]

Posterior and Predictive Distributions

- ▶ How do we compute the posterior and predictive distributions?
- ▶ Training set $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and test input x_{n+1} .
- ▶ Out of the (uncountably infinitely) many random variables $\{f(x)\}_{x \in \mathbb{X}}$ making up the GP only $n + 1$ has to do with the data:

$$f(x_1), f(x_2), \dots, f(x_{n+1})$$

- ▶ Training data gives observations $f(x_1) = y_1, \dots, f(x_n) = y_n$. The predictive distribution of $f(x_{n+1})$ is simply

$$p(f(x_{n+1}) | f(x_1) = y_1, \dots, f(x_n) = y_n)$$

which is easy to compute since $f(x_1), \dots, f(x_{n+1})$ is Gaussian.

- ▶ This can be generalized to noisy observations $y_i = f(x_i) + \epsilon_i$ or non-linear effects $y_i \sim D(f(x_i))$ where $D(\theta)$ is a distribution parametrized by θ .

Consistency and Existence

- ▶ The definition of Gaussian processes only give finite dimensional marginal distributions of the stochastic process.
- ▶ Fortunately these marginal distributions are *consistent*.
 - ▶ For every finite set $\mathbf{x} \subset \mathbb{X}$ we have a distinct distribution $\rho_{\mathbf{x}}([f(x)]_{x \in \mathbf{x}})$. These distributions are said to be consistent if

$$\rho_{\mathbf{x}}([f(x)]_{x \in \mathbf{x}}) = \int \rho_{\mathbf{x} \cup \mathbf{y}}([f(x)]_{x \in \mathbf{x} \cup \mathbf{y}}) d[f(x)]_{x \in \mathbf{y}}$$

for disjoint and finite $\mathbf{x}, \mathbf{y} \subset \mathbb{X}$.

- ▶ The marginal distributions for the GP are consistent because *Gaussians are closed under marginalization*.
- ▶ The *Kolmogorov Consistency Theorem* guarantees existence of GPs, i.e. the whole stochastic process $\{f(x)\}_{x \in \mathbb{X}}$.

Poisson Processes

- ▶ A *Poisson process* (PP) is a random function $f : \Sigma \rightarrow \mathbb{R}$ such that:
 - ▶ Σ is the σ -algebra over \mathbb{X} .
 - ▶ For any measurable set $A \subset \mathbb{X}$,

$$f(A) \sim \text{Poisson}(\lambda(A)),$$

where the parameter is the rate measure λ (a function from the measurable sets of \mathbb{X} to \mathbb{R}_+).

- ▶ And if $A, B \subset \mathbb{X}$ are disjoint then $f(A)$ and $f(B)$ are independent.
- ▶ The above family of distributions is consistent, since the sum of two independent Poisson variables is still Poisson with the rate parameter being the sum of the individual rates.
- ▶ Note that f is also a measure, a *random measure*. It always consists of point masses:

$$f = \sum_{i=1}^n \delta_{x_i}$$

where $x_1, x_2, \dots \in \mathbb{X}$ and $n \sim \text{Poisson}(\lambda(\mathbb{X}))$, i.e. f is a *point process*.

Gamma Processes

- ▶ A *Gamma process* (GP) is a random function $f : \Sigma \rightarrow \mathbb{R}$ such that:
 - ▶ For any measurable set $A \subset \mathbb{X}$,

$$f(A) \sim \text{Gamma}(\lambda(A), 1),$$

where the parameter is the shape measure λ .

- ▶ And if $A, B \subset \mathbb{X}$ are disjoint then $f(A)$ and $f(B)$ are independent.
- ▶ The above family of distributions is also consistent, since the sum of two independent gamma variables (with same scale parameter 1) is still gamma with the shape parameter being the sum of the individual shape parameters.
- ▶ f is also a random measure. It always consists of weighted point masses:

$$f = \sum_{i=1}^{\infty} w_i \delta_{x_i}$$

with total weight $\sum_{i=1}^{\infty} w_i \sim \text{Gamma}(\lambda(\mathbb{X}))$.

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

- Representations of Dirichlet Processes

- Applications of Dirichlet Processes

- Exchangeability

- Pitman-Yor Processes

Indian Buffet and Beta Processes

Completely Random Measure

Summary

Dirichlet Distributions

- ▶ A *Dirichlet distribution* is a distribution over the K -dimensional probability simplex:

$$\Delta_K = \{(\pi_1, \dots, \pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\}$$

- ▶ We say (π_1, \dots, π_K) is Dirichlet distributed,

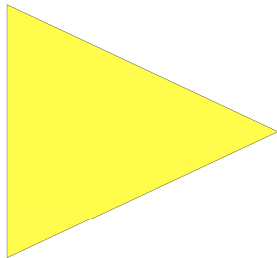
$$(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\lambda_1, \dots, \lambda_K)$$

with parameters $(\lambda_1, \dots, \lambda_K)$, if

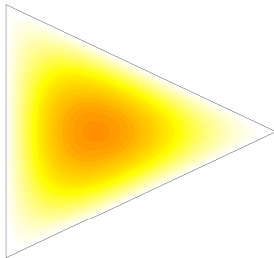
$$p(\pi_1, \dots, \pi_K) = \frac{\Gamma(\sum_k \lambda_k)}{\prod_k \Gamma(\lambda_k)} \prod_{k=1}^n \pi_k^{\lambda_k - 1}$$

Dirichlet Distributions

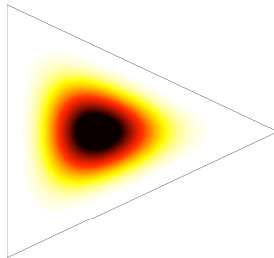
$\text{Dir}(1,0,1,0,1,0)$



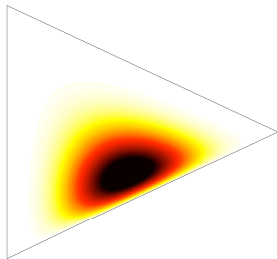
$\text{Dir}(2,0,2,0,2,0)$



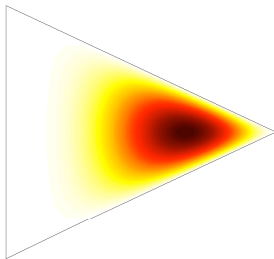
$\text{Dir}(5,0,5,0,5,0)$



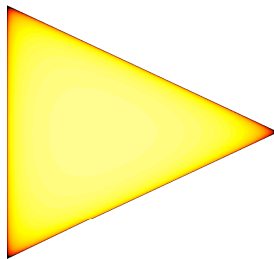
$\text{Dir}(5,0,5,0,2,0)$



$\text{Dir}(5,0,2,0,2,0)$



$\text{Dir}(0,7,0,7)$



Normalizing a Gamma Process

- ▶ We can obtain a sample of (π_1, \dots, π_K) by drawing K independent Gamma samples and normalizing:

$$\gamma_k \sim \text{Gamma}(\lambda_k, 1) \quad \text{for } k = 1, \dots, K$$

$$\pi_k = \gamma_k / \sum_{\ell} \gamma_{\ell}.$$

- ▶ Similarly a Dirichlet process is obtained by normalizing a gamma process:

$$\gamma \sim \text{GP}(\lambda)$$

$$G = \gamma / \gamma(\Theta)$$

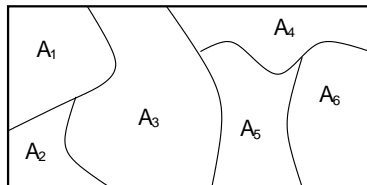
where λ is a base measure.

Dirichlet Processes

- ▶ A *Dirichlet Process* (DP) is a random probability measure G over (Θ, Σ) such that for any finite set of measurable partitions $A_1 \dot{\cup} \dots \dot{\cup} A_K = \Theta$,

$$(G(A_1), \dots, G(A_K)) \sim \text{Dirichlet}(\lambda(A_1), \dots, \lambda(A_K))$$

where λ is a base measure.



- ▶ The above family of distributions is consistent (next slide), and Kolmogorov Consistency Theorem can be applied to show existence (but there are technical conditions restricting the generality of the definition).

[Ferguson 1973, Blackwell and MacQueen 1973]

Consistency of Dirichlet Marginals

- ▶ Because Dirichlet variables are normalized gamma variables and sums of gammas are gammas, if (I_1, \dots, I_J) is a partition of $(1, \dots, K)$,

$$\left(\sum_{i \in I_1} \pi_i, \dots, \sum_{i \in I_J} \pi_i \right) \sim \text{Dirichlet} \left(\sum_{i \in I_1} \lambda_i, \dots, \sum_{i \in I_J} \lambda_i \right)$$

- ▶ If we have two partitions (A_1, \dots, A_K) and (B_1, \dots, B_J) of Θ , form the common refinement (C_1, \dots, C_L) where each C_ℓ is the intersection of some A_k with some B_j . Then:

By definition, $(G(C_1), \dots, G(C_L)) \sim \text{Dirichlet}(\lambda(C_1), \dots, \lambda(C_L))$

$$\begin{aligned} (G(A_1), \dots, G(A_K)) &= \left(\sum_{C_\ell \subset A_1} G(C_\ell), \dots, \sum_{C_\ell \subset A_K} G(C_\ell) \right) \\ &\sim \text{Dirichlet}(\lambda(A_1), \dots, \lambda(A_K)) \end{aligned}$$

Similarly, $(G(B_1), \dots, G(B_J)) \sim \text{Dirichlet}(\lambda(B_1), \dots, \lambda(B_J))$

so the distributions of $(G(A_1), \dots, G(A_K))$ and $(G(B_1), \dots, G(B_J))$ are consistent.

- ▶ Demonstration.

Parameters of Dirichlet Processes

- ▶ Usually we split the λ base measure into two parameters $\lambda = \alpha H$:
 - ▶ *Base distribution* H , which is like the *mean* of the DP.
 - ▶ *Strength parameter* α , which is like an *inverse-variance* of the DP.
- ▶ We write:

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

if for any partition (A_1, \dots, A_K) of Θ :

$$(G(A_1), \dots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \dots, \alpha H(A_K))$$

- ▶ The first and second moments of the DP:

$$\text{Expectation:} \quad \mathbb{E}[G(A)] = H(A)$$

$$\text{Variance:} \quad \mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

where A is any measurable subset of Θ .

Representations of Dirichlet Processes

- ▶ Since draws of gamma processes consist of weighted point masses, so will draws from Dirichlet processes:

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

where $\sum_k \pi_k = 1$ and $\theta_k^* \in \Theta$.

- ▶ What is the joint distribution over π_1, π_2, \dots and $\theta_1^*, \theta_2^*, \dots$?
- ▶ Since G is a (random) probability measure over Θ , we can treat it as a distribution and draw samples from it. Let

$$\theta_1, \theta_2, \dots \sim G$$

be random variables with distribution G .

- ▶ What is the marginal distribution of $\theta_1, \theta_2, \dots$ with G integrated out?
- ▶ There is positive probability that sets of θ_i 's can take on the same value θ_k^* for some k , i.e. the θ_i 's cluster together. How do these clusters look like?
- ▶ For practical modelling purposes this is sufficient. But is this sufficient to tell us all about G ?

Stick-breaking Construction

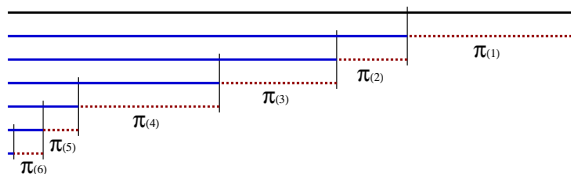
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

- ▶ There is a simple construction giving the joint distribution of π_1, π_2, \dots and $\theta_1^*, \theta_2^*, \dots$ called the *stick-breaking construction*.

$$\theta_k^* \sim H$$

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

$$v_k \sim \text{Beta}(1, \alpha)$$



- ▶ Also known as the *GEM* distribution, write $\pi \sim \text{GEM}(\alpha)$.
- ▶ If we order π_1, π_2, \dots in decreasing order, the resulting distribution is called the *Poisson-Dirichlet* distribution.

[Sethuraman 1994]

Pólya Urn Scheme

$$\theta_1, \theta_2, \dots \sim G$$

- ▶ The marginal distribution of $\theta_1, \theta_2, \dots$ has a simple generative process called the *Pólya urn scheme*.

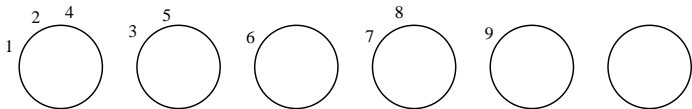
$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- ▶ Picking balls of different colors from an urn:
 - ▶ Start with no balls in the urn.
 - ▶ with probability $\propto \alpha$, draw $\theta_n \sim H$, and add a ball of color θ_n into urn.
 - ▶ With probability $\propto n - 1$, pick a ball at random from the urn, record θ_n to be its color and return two balls of color θ_n into urn.
- ▶ Pólya urn scheme is like a “representer” for the DP—a finite projection of an infinite object G .
- ▶ Also known as the *Blackwell-MacQueen urn scheme*.

[Blackwell and MacQueen 1973]

Chinese Restaurant Process

- ▶ According to the Pólya urn scheme, and because \mathbf{G} consists of weighted point masses, $\theta_1, \dots, \theta_n$ take on $K < n$ distinct values, say $\theta_1^*, \dots, \theta_K^*$.
- ▶ This defines a partition of $(1, \dots, n)$ into K clusters, such that if i is in cluster k , then $\theta_i = \theta_k^*$.
- ▶ The distribution over partitions is a *Chinese restaurant process* (CRP).
- ▶ Generating from the CRP:
 - ▶ First customer sits at the first table.
 - ▶ Customer n sits at:
 - ▶ Table k with probability $\frac{n_k}{\alpha + n - 1}$ where n_k is the number of customers at table k .
 - ▶ A new table $K + 1$ with probability $\frac{\alpha}{\alpha + n - 1}$.
 - ▶ Customers \Leftrightarrow integers, tables \Leftrightarrow clusters.
- ▶ The CRP exhibits the *clustering property* of the DP.
 - ▶ *Rich-gets-richer* effect implies small number of large clusters.
 - ▶ Expected number of clusters is $K = O(\alpha \log n)$.



Representations of Dirichlet Processes

- ▶ Posterior Dirichlet process:

$$\begin{array}{l} G \sim \text{DP}(\alpha, H) \\ \theta | G \sim G \end{array} \iff \begin{array}{l} \theta \sim H \\ G | \theta \sim \text{DP} \left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1} \right) \end{array}$$

- ▶ Pólya urn scheme:

$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- ▶ Chinese restaurant process:

$$p(\text{customer } n \text{ sat at table } k | \text{past}) = \begin{cases} \frac{n_k}{n-1+\alpha} & \text{if occupied table} \\ \frac{\alpha}{n-1+\alpha} & \text{if new table} \end{cases}$$

- ▶ Stick-breaking construction:

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \quad \beta_k \sim \text{Beta}(1, \alpha) \quad \theta_k^* \sim H \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

Density Estimation

- ▶ Parametric density estimation (e.g. Gaussian, mixture models)

Data: $\mathbf{x} = \{x_1, x_2, \dots\}$

Model: $x_j | \mathbf{w} \sim F(\cdot | \mathbf{w})$

- ▶ Prior over parameters

$$p(\mathbf{w})$$

- ▶ Posterior over parameters

$$p(\mathbf{w} | \mathbf{x}) = \frac{p(\mathbf{w})p(\mathbf{x} | \mathbf{w})}{p(\mathbf{x})}$$

- ▶ Prediction with posteriors

$$p(x_* | \mathbf{x}) = \int p(x_* | \mathbf{w})p(\mathbf{w} | \mathbf{x}) d\mathbf{w}$$

Density Estimation

- ▶ Bayesian nonparametric density estimation with Dirichlet processes

Data: $\mathbf{x} = \{x_1, x_2, \dots\}$

Model: $x_j \sim G$

- ▶ Prior over distributions

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

- ▶ Posterior over distributions

$$p(G|\mathbf{x}) = \frac{p(G)p(\mathbf{x}|G)}{p(\mathbf{x})}$$

- ▶ Prediction with posteriors

$$p(x_*|\mathbf{x}) = \int p(x_*|G)p(G|\mathbf{x}) dF = \int G(x_*)p(G|\mathbf{x}) dG$$

- ▶ *Not quite feasible, since G is a discrete distribution, in particular it has no density.*

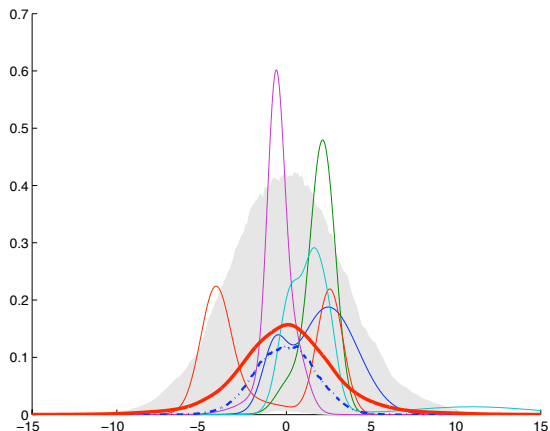
Density Estimation

- ▶ Solution: Convolve the DP with a smooth distribution:

$$\begin{aligned} G &\sim \text{DP}(\alpha, H) \\ F(\cdot) &= \int F(\cdot|\theta)dG(\theta) \\ x_i &\sim F_x \end{aligned} \quad \Rightarrow \quad \begin{aligned} G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \\ F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \\ x_i &\sim F_x \end{aligned}$$

- ▶ Demonstration.

Density Estimation

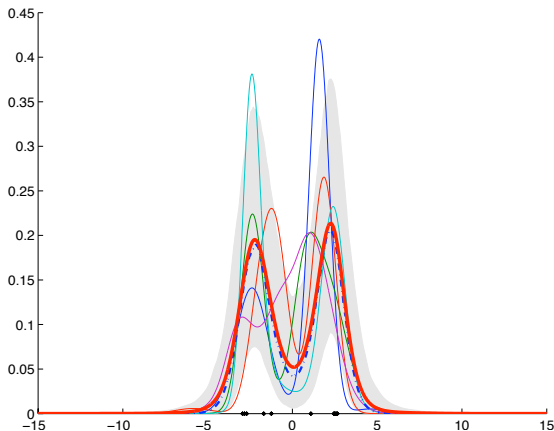


$F(\cdot|\mu, \Sigma)$ is Gaussian with mean μ , covariance Σ .

$H(\mu, \Sigma)$ is Gaussian-inverse-Wishart conjugate prior.

Red: mean density. Blue: median density. Grey: 5-95 quantile. Others: draws. Black: data points.

Density Estimation



$F(\cdot|\mu, \Sigma)$ is Gaussian with mean μ , covariance Σ .

$H(\mu, \Sigma)$ is Gaussian-inverse-Wishart conjugate prior.

Red: mean density. Blue: median density. Grey: 5-95 quantile. Others: draws. Black: data points.

Clustering

- ▶ Recall our approach to density estimation:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \sim \text{DP}(\alpha, H)$$
$$F_x(\cdot) = \sum_{k=1}^{\infty} \pi_k F(\cdot | \theta_k^*)$$
$$x_i \sim F_x$$

- ▶ Above model equivalent to:

$$z_i \sim \text{Discrete}(\pi)$$
$$\theta_i = \theta_{z_i}^*$$
$$x_i | z_i \sim F(\cdot | \theta_i) = F(\cdot | \theta_{z_i}^*)$$

- ▶ This is simply a mixture model with an *infinite* number of components. This is called a *DP mixture model*.

Clustering

- ▶ DP mixture models are used in a variety of clustering applications, where the number of clusters is not known a priori.
- ▶ They are also used in applications in which we believe the number of clusters grows without bound as the amount of data grows.
- ▶ DPs have also found uses in applications beyond clustering, where the number of latent objects is not known or unbounded.
 - ▶ Nonparametric probabilistic context free grammars.
 - ▶ Visual scene analysis.
 - ▶ Infinite hidden Markov models/trees.
 - ▶ Haplotype inference.
 - ▶ ...
- ▶ In many such applications it is important to be able to model the same set of objects in different contexts.
- ▶ This corresponds to the problem of *grouped clustering* and can be tackled using *hierarchical Dirichlet processes*.

Exchangeability

- ▶ Instead of deriving the Pólya urn scheme by marginalizing out a DP, consider starting directly from the conditional distributions:

$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- ▶ For any n , the joint distribution of $\theta_1, \dots, \theta_n$ is:

$$p(\theta_1, \dots, \theta_n) = \frac{\alpha^K \prod_{k=1}^K h(\theta_k^*) (m_{nk} - 1)!}{\prod_{i=1}^n i - 1 + \alpha}$$

where $h(\theta)$ is density of θ under H , $\theta_1^*, \dots, \theta_K^*$ are the unique values, and θ_k^* occurred m_{nk} times among $\theta_1, \dots, \theta_n$.

- ▶ The joint distribution is *exchangeable* wrt permutations of $\theta_1, \dots, \theta_n$.
- ▶ *De Finetti's Theorem* says that there must be a random probability measure G making $\theta_1, \theta_2, \dots$ iid. This is the DP.

De Finetti's Theorem

Let $\theta_1, \theta_2, \dots$ be an infinite sequence of random variables with joint distribution p . If for all $n \geq 1$, and all permutations $\sigma \in \Sigma_n$ on n objects,

$$p(\theta_1, \dots, \theta_n) = p(\theta_{\sigma(1)}, \dots, \theta_{\sigma(n)})$$

That is, the sequence is *infinitely exchangeable*. Then there exists a latent random parameter G such that:

$$p(\theta_1, \dots, \theta_n) = \int p(G) \prod_{i=1}^n p(\theta_i | G) dG$$

where p is a joint distribution over G and θ_i 's.

- ▶ θ_i 's are *independent* given G .
- ▶ Sufficient to define G through the conditionals $p(\theta_n | \theta_1, \dots, \theta_{n-1})$.
- ▶ G can be *infinite dimensional* (indeed it is often a *random measure*).
- ▶ The set of infinitely exchangeable sequences is *convex* and it is an important theoretical topic to study the set of *extremal points*.
- ▶ Partial exchangeability: Markov, group, arrays,...

Pitman-Yor Processes

- ▶ Two-parameter generalization of the Chinese restaurant process:

$$p(\text{customer } n \text{ sat at table } k | \text{past}) = \begin{cases} \frac{n_k - \beta}{n - 1 + \alpha} & \text{if occupied table} \\ \frac{\alpha + \beta K}{n - 1 + \alpha} & \text{if new table} \end{cases}$$

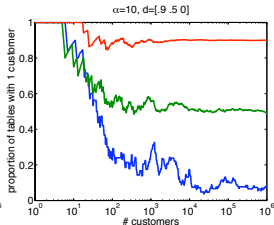
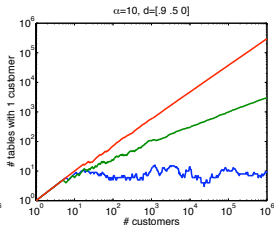
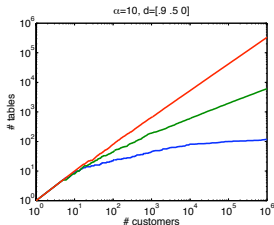
- ▶ Associating each cluster k with a unique draw $\theta_k^* \sim H$, the corresponding Pólya urn scheme is also exchangeable.
- ▶ De Finetti's Theorem states that there is a random measure underlying this two-parameter generalization.
 - ▶ This is the *Pitman-Yor process*.
- ▶ The Pitman-Yor process also has a stick-breaking construction:

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i) \quad \beta_k \sim \text{Beta}(1 - \beta, \alpha + \beta k) \quad \theta_k^* \sim H \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

[Pitman and Yor 1997, Perman et al. 1992]

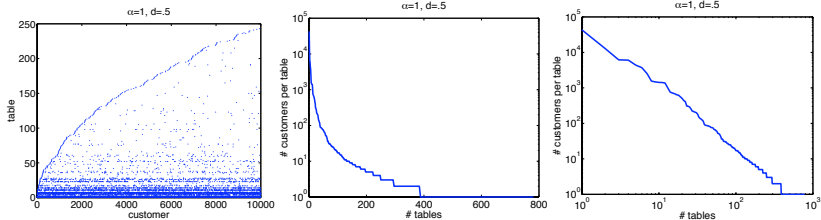
Pitman-Yor Processes

- ▶ Two salient features of the Pitman-Yor process:
 - ▶ With more occupied tables, the chance of even more tables becomes higher.
 - ▶ Tables with smaller occupancy numbers tend to have lower chance of getting new customers.
- ▶ The above means that Pitman-Yor processes produce Zipf's Law type behaviour, with $K = O(\alpha n^\beta)$.

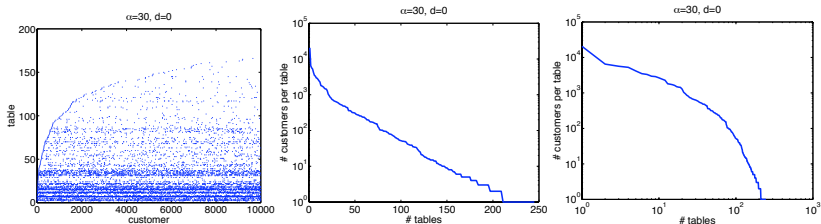


Pitman-Yor Processes

Draw from a Pitman-Yor process



Draw from a Dirichlet process



Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Indian Buffet and Beta Processes

Completely Random Measure

Summary

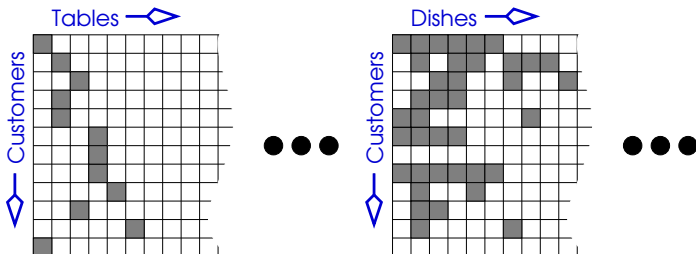
Bibliography

Beyond Clustering

- ▶ Dirichlet and Pitman-Yor processes are nonparametric models of clustering.
- ▶ Can nonparametric models go beyond clustering to describe data in more expressive ways?
 - ▶ Hierarchical (e.g. taxonomies)?
 - ▶ Distributed (e.g. multiple causes)?

Indian Buffet Processes

- ▶ The *Indian Buffet Process* (IBP) is akin to the Chinese restaurant process but describes each customer with a binary vector instead of cluster.
- ▶ Generating from an IBP:
 - ▶ Parameter α .
 - ▶ First customer picks $\text{Poisson}(\alpha)$ dishes to eat.
 - ▶ Subsequent customer i picks dish k with probability $\frac{n_k}{i}$; and picks $\text{Poisson}(\frac{\alpha}{i})$ new dishes.



Infinite Independent Components Analysis

- ▶ Each image X_i is a linear combination of sparse features:

$$X_i = \sum_k \Lambda_k y_{ik}$$

where y_{ik} is activity of feature k with sparse prior. One possibility is a mixture of a Gaussian and a point mass at 0:

$$y_{ik} = z_{ik} a_{ik} \quad a_{ik} \sim \mathcal{N}(0, 1) \quad Z \sim \text{IBP}(\alpha)$$

- ▶ An ICA model with infinite number of features.

[Knowles and Ghahramani 2007]

Indian Buffet Processes and Exchangeability

- ▶ The IBP is infinitely exchangeable, though this is much harder to see.
- ▶ De Finetti's Theorem again states that there is some random measure underlying the IBP.
- ▶ This random measure is the Beta process.

[Griffiths and Ghahramani 2006, Thibaux and Jordan 2007]

Beta Processes

- ▶ A *beta process* $B \sim \text{BP}(c, \alpha H)$ is a random discrete measure with form:

$$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

where the points $P = \{(\theta_1^*, \mu_1), (\theta_2^*, \mu_2), \dots\}$ are spikes in a 2D Poisson process with rate measure:

$$c\mu^{-1}(1 - \mu)^{c-1} d\mu \alpha H(d\theta)$$

- ▶ The beta process with $c = 1$ is the de Finetti measure for the IBP. When $c \neq 1$ we have a two parameter generalization of the IBP.
- ▶ This is an example of a *completely random measure*.
- ▶ A beta process *does not* have Beta distributed marginals.

[Hjort 1990]

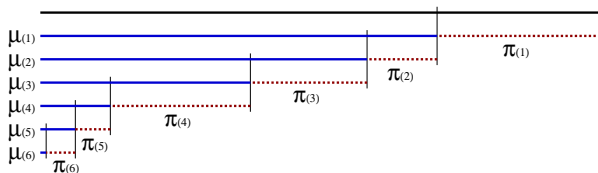
Stick-breaking Construction for Beta Processes

- ▶ When $c = 1$ it was shown that the following generates a draw of B :

$$v_k \sim \text{Beta}(1, \alpha) \quad \mu_k = (1 - v_k) \prod_{i=1}^{k-1} (1 - v_i) \quad \theta_k^* \sim H$$

$$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

- ▶ The above is the complement of the stick-breaking construction for DPs!



Survival Analysis

- ▶ The Beta process was first proposed as a Bayesian nonparametric model for survival analysis with right-censored data.
- ▶ The hazard rate B is given a $\text{BP}(c, \alpha H)$ prior. $B(\theta)d\theta$ is the chance of death in an infinitesimal interval $[\theta, \theta + d\theta)$ given that the individual has survived up to time θ .
- ▶ Data consists of a set of death times τ_1, τ_2, \dots and censored times $\gamma_1, \gamma_2, \dots$, and can be summarized as:

Death measure:
$$D = \sum_i \delta_{\tau_i}$$

Number-at-risk function:
$$R(\theta) = D([\theta, \infty)) + \sum_i \mathbb{I}(\gamma_i \geq \theta)$$

- ▶ The posterior of B is:

$$B|D, R \sim \text{BP}(c + R, \alpha H + D)$$

Note: the above is a generalization to c being a function of θ .

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Indian Buffet and Beta Processes

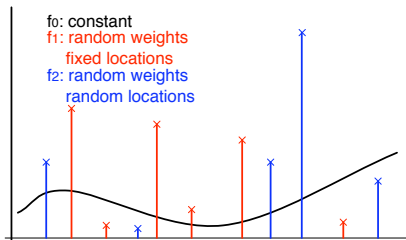
Completely Random Measure

Summary

Bibliography

Completely Random Measures

- ▶ A *completely random measure* (CRM) is a random measure f such that $f(A)$ and $f(B)$ are independent whenever A and B are disjoint.
- ▶ A CRM f can always be decomposed into three distinct and independent components: $f = f_0 + f_1 + f_2$ where:
 - ▶ f_0 is a constant measure;
 - ▶ $f_1 = \sum_{i=1}^n u_i \delta_{y_i}$ where w_i are independent, and y_i are *fixed*;
 - ▶ $f_2 = \sum_{i=1}^n v_i \delta_{x_i}$ where $(x_1, v_1), (x_2, v_2), \dots$ are the atoms of a Poisson process with non-atomic rate measure λ over the space $\mathbb{X} \times (0, \infty]$.
- ▶ Examples of CRMs: Poisson, gamma, beta processes. DPs are *normalized random measures*.



Examples of Completely Random Measures

Gamma process: $\alpha > 0$

$$\lambda(d\theta, d\mu) = \mu^{-1} e^{-\mu} d\mu \alpha H(d\theta)$$

Inverse Gaussian process: $\tau > 0, \alpha > 0$

$$\lambda(d\theta, d\mu) = \mu^{-3/2} e^{-\tau\mu} / \sqrt{2\pi} d\mu \alpha H(d\theta)$$

Stable process: $0 < \beta < 1, \alpha > 0$

$$\lambda(d\theta, d\mu) = \beta \mu^{-1-\beta} / \Gamma(1 - \beta) d\mu \alpha H(d\theta)$$

Generalized gamma process: $0 < \beta < 1, \tau \geq 0, \alpha > 0$

$$\lambda(d\theta, d\mu) = \beta \mu^{-1-\beta} e^{-\tau\mu} / \Gamma(1 - \beta) d\mu \alpha H(d\theta)$$

Beta process: $0 < c < 1, \alpha > 0$

$$\lambda(d\theta, d\mu) = c \mu^{-1} (1 - \mu)^{c-1} d\mu \alpha H(d\theta)$$

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Indian Buffet and Beta Processes

Completely Random Measure

Summary

Bibliography

Summary

- ▶ Introduced the major Bayesian nonparametric models and stochastic processes developed in statistics and machine learning:
 - ▶ Gaussian processes, Poisson processes, gamma processes, Dirichlet processes, beta processes, completely random measures.
 - ▶ Missing: hierarchical Dirichlet processes and other hierarchical Bayesian nonparametric models, infinite hidden Markov and other time series models, Dirichlet diffusion trees and other hierarchical clustering models...
- ▶ Described two important theoretical tools used to build such models:
 - ▶ Consistency and Kolmogorov's Consistency Theorem
 - ▶ Exchangeability and de Finetti's Theorem
- ▶ Touched upon a number of prototypical applications of Bayesian nonparametric models.
- ▶ Missing: Inference methods based on MCMC, variational, and on different representations.

Outline

Bayesian Nonparametric Modelling

Measure Theoretic Probability Theory

Gaussian Processes

Dirichlet Processes

Indian Buffet and Beta Processes

Completely Random Measure

Summary

Bibliography

Bibliography I

Dirichlet Processes and Beyond in Machine Learning

Dirichlet Processes were first introduced by [Ferguson 1973], while [Antoniak 1974] further developed DPs as well as introduce the mixture of DPs. [Blackwell and MacQueen 1973] showed that the Pólya urn scheme is exchangeable with the DP being its de Finetti measure. Further information on the Chinese restaurant process can be obtained at [Aldous 1985, Pitman 2002]. The DP is also related to Ewens' Sampling Formula [Ewens 1972]. [Sethuraman 1994] gave a constructive definition of the DP via a stick-breaking construction. DPs were rediscovered in the machine learning community by [Neal 1992, Rasmussen 2000].

Hierarchical Dirichlet Processes (HDPs) were first developed by [Teh et al. 2006], although an aspect of the model was first discussed in the context of infinite hidden Markov models [Beal et al. 2002]. HDPs and generalizations have been applied across a wide variety of fields.

Dependent Dirichlet Processes are sets of coupled distributions over probability measures, each of which is marginally DP [MacEachern et al. 2001]. A variety of dependent DPs have been proposed in the literature since then [Srebro and Roweis 2005, Griffin 2007, Caron et al. 2007]. The infinite mixture of Gaussian processes of [Rasmussen and Ghahramani 2002] can also be interpreted as a dependent DP.

Indian Buffet Processes (IBPs) were first proposed in [Griffiths and Ghahramani 2006], and extended to a two-parameter family in [Ghahramani et al. 2007]. [Thibaux and Jordan 2007] showed that the de Finetti measure for the IBP is the beta process of [Hjort 1990], while [Teh et al. 2007] gave a stick-breaking construction and developed efficient slice sampling inference algorithms for the IBP.

Nonparametric Tree Models are models that use distributions over trees that are consistent and exchangeable. [Blei et al. 2004] used a nested CRP to define distributions over trees with a finite number of levels. [Neal 2001, Neal 2003] defined Dirichlet diffusion trees, which are binary trees produced by a fragmentation process. [Teh et al. 2008] used Kingman's coalescent [Kingman 1982b, Kingman 1982a] to produce random binary trees using a coalescent process. [Roy et al. 2007] proposed annotated hierarchies, using tree-consistent partitions first defined in [Heller and Ghahramani 2005] to model both relational and featural data.

Markov chain Monte Carlo Inference algorithms are the dominant approaches to inference in DP mixtures. [Neal 2000] is a good review of algorithms based on Gibbs sampling in the CRP representation. Algorithm 8 in [Neal 2000] is still one of the best algorithms based on simple local moves. [Ishwaran and James 2001] proposed blocked Gibbs sampling in the stick-breaking representation instead due to the simplicity in implementation. This has been further explored in [Porteous et al. 2006]. Since then there has been proposals for better MCMC samplers based on proposing larger moves in a Metropolis-Hastings framework [Jain and Neal 2004, Liang et al. 2007a], as well as sequential Monte Carlo [Fearhead 2004, Mansinghka et al. 2007].

Other Approximate Inference Methods have also been proposed for DP mixture models. [Blei and Jordan 2006] is the first variational Bayesian approximation, and is based on a truncated stick-breaking representation. [Kurihara et al. 2007] proposed an

Bibliography II

Dirichlet Processes and Beyond in Machine Learning

improved VB approximation based on a better truncation technique, and using KD-trees for extremely efficient inference in large scale applications. [Kurihara et al. 2007] studied improved VB approximations based on integrating out the stick-breaking weights. [Minka and Ghahramani 2003] derived an expectation propagation based algorithm. [Heller and Ghahramani 2005] derived tree-based approximation which can be seen as a Bayesian hierarchical clustering algorithm. [Daume III 2007] developed admissible search heuristics to find MAP clusterings in a DP mixture model.

Computer Vision and Image Processing. HDPs have been used in object tracking

[Fox et al. 2006, Fox et al. 2007b, Fox et al. 2007a]. An extension called the transformed Dirichlet process has been used in scene analysis [Sudderth et al. 2006b, Sudderth et al. 2006a, Sudderth et al. 2008], a related extension has been used in fMRI image analysis [Kim and Smyth 2007, Kim 2007]. An extension of the infinite hidden Markov model called the nonparametric hidden Markov tree has been introduced and applied to image denoising [Kivinen et al. 2007a, Kivinen et al. 2007b].

Natural Language Processing. HDPs are essential ingredients in defining nonparametric context free grammars

[Liang et al. 2007b, Finkel et al. 2007]. [Johnson et al. 2007] defined adaptor grammars, which is a framework generalizing both probabilistic context free grammars as well as a variety of nonparametric models including DPs and HDPs. DPs and HDPs have been used in information retrieval [Cowans 2004], word segmentation [Goldwater et al. 2006b], word morphology modelling [Goldwater et al. 2006a], coreference resolution [Haghighi and Klein 2007], topic modelling [Blei et al. 2004, Teh et al. 2006, Li et al. 2007]. An extension of the HDP called the hierarchical Pitman-Yor process has been applied to language modelling [Teh 2006a, Teh 2006b, Goldwater et al. 2006a]. [Savova et al. 2007] used annotated hierarchies to construct syntactic hierarchies. These on nonparametric methods in NLP include [Cowans 2006, Goldwater 2006].

Other Applications. Applications of DPs, HDPs and infinite HMMs in bioinformatics include

[Xing et al. 2004, Xing et al. 2007, Xing et al. 2006, Xing and Sohn 2007a, Xing and Sohn 2007b]. DPs have been applied in relational learning [Shafto et al. 2006, Kemp et al. 2006, Xu et al. 2006], spike sorting [Wood et al. 2006a, Görür 2007]. The HDP has been used in a cognitive model of categorization [Griffiths et al. 2007]. IBPs have been applied to infer hidden causes [Wood et al. 2006b], in a choice model [Görür et al. 2006], to modelling dyadic data [Meeds et al. 2007], to overlapping clustering [Heller and Ghahramani 2007], and to matrix factorization [Wood and Griffiths 2006].

References I



Aldous, D. (1985).

Exchangeability and related topics.

In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin.



Antoniak, C. E. (1974).

Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.

Annals of Statistics, 2(6):1152–1174.



Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002).

The infinite hidden Markov model.

In *Advances in Neural Information Processing Systems*, volume 14.



Blackwell, D. and MacQueen, J. B. (1973).

Ferguson distributions via Pólya urn schemes.

Annals of Statistics, 1:353–355.



Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004).

Hierarchical topic models and the nested Chinese restaurant process.

In *Advances in Neural Information Processing Systems*, volume 16.



Blei, D. M. and Jordan, M. I. (2006).

Variational inference for Dirichlet process mixtures.

Bayesian Analysis, 1(1):121–144.



Caron, F., Davy, M., and Doucet, A. (2007).

Generalized Polya urn for time-varying Dirichlet process mixtures.

In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 23.

References II



Cowans, P. (2004).

Information retrieval using hierarchical Dirichlet processes.

In *Proceedings of the Annual International Conference on Research and Development in Information Retrieval*, volume 27, pages 564–565.



Cowans, P. (2006).

Probabilistic Document Modelling.

PhD thesis, University of Cambridge.



Daume III, H. (2007).

Fast search for Dirichlet process mixture models.

In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.



Ewens, W. J. (1972).

The sampling theory of selectively neutral alleles.

Theoretical Population Biology, 3:87–112.



Fearnhead, P. (2004).

Particle filters for mixture models with an unknown number of components.

Statistics and Computing, 14:11–21.



Ferguson, T. S. (1973).

A Bayesian analysis of some nonparametric problems.

Annals of Statistics, 1(2):209–230.



Finkel, J. R., Grenager, T., and Manning, C. D. (2007).

The infinite tree.

In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

References III



Fox, E. B., Choi, D. S., and Willsky, A. S. (2006).

Nonparametric Bayesian methods for large scale multi-target tracking.

In *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, volume 40.



Fox, E. B., Sudderth, E. B., Choi, D. S., and Willsky, A. S. (2007a).

Tracking a non-cooperative maneuvering target using hierarchical Dirichlet processes.

In *Proceedings of the Adaptive Sensor Array Processing Conference*.



Fox, E. B., Sudderth, E. B., and Willsky, A. S. (2007b).

Hierarchical Dirichlet processes for tracking maneuvering targets.

In *Proceedings of the International Conference on Information Fusion*.



Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007).

Bayesian nonparametric latent feature models (with discussion and rejoinder).

In *Bayesian Statistics*, volume 8.



Goldwater, S. (2006).

Nonparametric Bayesian Models of Lexical Acquisition.

PhD thesis, Brown University.



Goldwater, S., Griffiths, T., and Johnson, M. (2006a).

Interpolating between types and tokens by estimating power-law generators.

In *Advances in Neural Information Processing Systems*, volume 18.



Goldwater, S., Griffiths, T. L., and Johnson, M. (2006b).

Contextual dependencies in unsupervised word segmentation.

In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

References IV



Görür, D. (2007).

Nonparametric Bayesian Discrete Latent Variable Models for Unsupervised Learning.
PhD thesis, Technische Universität Berlin.



Görür, D., Jäkel, F., and Rasmussen, C. E. (2006).

A choice model with infinitely many latent features.
In Proceedings of the International Conference on Machine Learning, volume 23.



Griffin, J. E. (2007).

The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference.
Technical report, Department of Statistics, University of Warwick.



Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007).

Unifying rational models of categorization via the hierarchical Dirichlet process.
In Proceedings of the Annual Conference of the Cognitive Science Society, volume 29.



Griffiths, T. L. and Ghahramani, Z. (2006).

Infinite latent feature models and the Indian buffet process.
In Advances in Neural Information Processing Systems, volume 18.



Haghighi, A. and Klein, D. (2007).

Unsupervised coreference resolution in a nonparametric Bayesian model.
In Proceedings of the Annual Meeting of the Association for Computational Linguistics.



Heller, K. A. and Ghahramani, Z. (2005).

Bayesian hierarchical clustering.
In Proceedings of the International Conference on Machine Learning, volume 22.

References V



Heller, K. A. and Ghahramani, Z. (2007).

A nonparametric Bayesian approach to modeling overlapping clusters.

In Proceedings of the International Workshop on Artificial Intelligence and Statistics, volume 11.



Hjort, N. L. (1990).

Nonparametric Bayes estimators based on beta processes in models for life history data.

Annals of Statistics, 18(3):1259–1294.



Ishwaran, H. and James, L. F. (2001).

Gibbs sampling methods for stick-breaking priors.

Journal of the American Statistical Association, 96(453):161–173.



Jain, S. and Neal, R. M. (2004).

A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.

Technical report, Department of Statistics, University of Toronto.



Johnson, M., Griffiths, T. L., and Goldwater, S. (2007).

Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models.

In Advances in Neural Information Processing Systems, volume 19.



Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006).

Learning systems of concepts with an infinite relational model.

In Proceedings of the AAAI Conference on Artificial Intelligence, volume 21.



Kim, S. (2007).

Learning Hierarchical Probabilistic Models with Random Effects with Applications to Time-series and Image Data.

PhD thesis, Information and Computer Science, University of California at Irvine.

References VI



Kim, S. and Smyth, P. (2007).
Hierarchical dirichlet processes with random effects.
In Advances in Neural Information Processing Systems, volume 19.



Kingman, J. F. C. (1982a).
The coalescent.
Stochastic Processes and their Applications, 13:235–248.



Kingman, J. F. C. (1982b).
On the genealogy of large populations.
Journal of Applied Probability, 19:27–43.
Essays in Statistical Science.



Kivinen, J., Sudderth, E., and Jordan, M. I. (2007a).
Image denoising with nonparametric hidden Markov trees.
In IEEE International Conference on Image Processing (ICIP), San Antonio, TX.



Kivinen, J., Sudderth, E., and Jordan, M. I. (2007b).
Learning multiscale representations of natural scenes using Dirichlet processes.
In IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil.




Knowles, D. and Ghahramani, Z. (2007).
Infinite sparse factor analysis and infinite independent components analysis.
In International Conference on Independent Component Analysis and Signal Separation, volume 7 of *Lecture Notes in Computer Science*. Springer.




Kurihara, K., Welling, M., and Vlassis, N. (2007).
Accelerated variational DP mixture models.
In Advances in Neural Information Processing Systems, volume 19.


References VII

 Li, W., Blei, D. M., and McCallum, A. (2007).
Nonparametric Bayes pachinko allocation.
In Proceedings of the Conference on Uncertainty in Artificial Intelligence.


 Liang, P., Jordan, M. I., and Taskar, B. (2007a).
A permutation-augmented sampler for Dirichlet process mixture models.
In Proceedings of the International Conference on Machine Learning.

 Liang, P., Petrov, S., Jordan, M. I., and Klein, D. (2007b).
The infinite PCFG using hierarchical Dirichlet processes.
In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

 MacEachern, S., Kottas, A., and Gelfand, A. (2001).
Spatial nonparametric Bayesian models.
Technical Report 01-10, Institute of Statistics and Decision Sciences, Duke University.
<http://ftp.isds.duke.edu/WorkingPapers/01-10.html>.

 Mansinghka, V. K., Roy, D. M., Rifkin, R., and Tenenbaum, J. B. (2007).
AClass: An online algorithm for generative classification.
In Proceedings of the International Workshop on Artificial Intelligence and Statistics, volume 11.

 Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007).
Modeling dyadic data with binary latent factors.
In Advances in Neural Information Processing Systems, volume 19.

 Minka, T. P. and Ghahramani, Z. (2003).
Expectation propagation for infinite mixtures.
Presented at NIPS2003 Workshop on Nonparametric Bayesian Methods and Infinite Models.

References VIII



Neal, R. M. (1992).

Bayesian mixture modeling.

In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211.



Neal, R. M. (2000).

Markov chain sampling methods for Dirichlet process mixture models.

Journal of Computational and Graphical Statistics, 9:249–265.



Neal, R. M. (2001).

Defining priors for distributions using Dirichlet diffusion trees.

Technical Report 0104, Department of Statistics, University of Toronto.



Neal, R. M. (2003).

Density modeling and clustering using Dirichlet diffusion trees.

In *Bayesian Statistics*, volume 7, pages 619–629.



Perman, M., Pitman, J., and Yor, M. (1992).

Size-biased sampling of Poisson point processes and excursions.

Probability Theory and Related Fields, 92(1):21–39.



Pitman, J. (2002).

Combinatorial stochastic processes.

Technical Report 621, Department of Statistics, University of California at Berkeley.

Lecture notes for St. Flour Summer School.



Pitman, J. and Yor, M. (1997).

The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator.

Annals of Probability, 25:855–900.

References IX



Porteous, I., Ihler, A., Smyth, P., and Welling, M. (2006).

Gibbs sampling for (Coupled) infinite mixture models in the stick-breaking representation.
In Proceedings of the Conference on Uncertainty in Artificial Intelligence, volume 22.



Rasmussen, C. E. (2000).

The infinite Gaussian mixture model.
In Advances in Neural Information Processing Systems, volume 12.



Rasmussen, C. E. and Ghahramani, Z. (2001).

Occam's razor.
In Advances in Neural Information Processing Systems, volume 13.



Rasmussen, C. E. and Ghahramani, Z. (2002).

Infinite mixtures of Gaussian process experts.
In Advances in Neural Information Processing Systems, volume 14.



Rasmussen, C. E. and Williams, C. K. I. (2006).

Gaussian Processes for Machine Learning.
MIT Press.



Roy, D. M., Kemp, C., Mansinghka, V., and Tenenbaum, J. B. (2007).

Learning annotated hierarchies from relational data.
In Advances in Neural Information Processing Systems, volume 19.



Savova, V., Roy, D., Schmidt, L., and Tenenbaum, J. B. (2007).

Discovering syntactic hierarchies.
In Proceedings of the Annual Conference of the Cognitive Science Society, volume 29.

References X



Sethuraman, J. (1994).

A constructive definition of Dirichlet priors.

Statistica Sinica, 4:639–650.



Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., and Tenenbaum, J. B. (2006).

Learning cross-cutting systems of categories.

In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 28.



Srebro, N. and Roweis, S. (2005).

Time-varying topic models using dependent Dirichlet processes.

Technical Report UTML-TR-2005-003, Department of Computer Science, University of Toronto.



Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2006a).

Depth from familiar objects: A hierarchical model for 3D scenes.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.



Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2006b).

Describing visual scenes using transformed Dirichlet processes.

In *Advances in Neural Information Processing Systems*, volume 18.



Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2008).

Describing visual scenes using transformed objects and parts.

International Journal of Computer Vision, 77.



Teh, Y. W. (2006a).

A Bayesian interpretation of interpolated Kneser-Ney.

Technical Report TRA2/06, School of Computing, National University of Singapore.

References XI



Teh, Y. W. (2006b).

A hierarchical Bayesian language model based on Pitman-Yor processes.

In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.



Teh, Y. W. (2007).

Dirichlet processes.

Submitted to *Encyclopedia of Machine Learning*.



Teh, Y. W., Daume III, H., and Roy, D. M. (2008).

Bayesian agglomerative clustering with coalescents.

In *Advances in Neural Information Processing Systems*, volume 20.



Teh, Y. W., Görür, D., and Ghahramani, Z. (2007).

Stick-breaking construction for the Indian buffet process.

In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11.



Teh, Y. W. and Jordan, M. I. (2009).

Hierarchical Bayesian nonparametric models with applications.

In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *To appear in Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press.



Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).

Hierarchical Dirichlet processes.

Journal of the American Statistical Association, 101(476):1566–1581.



Thibaux, R. and Jordan, M. I. (2007).

Hierarchical beta processes and the Indian buffet process.

In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

References XII



Wood, F., Goldwater, S., and Black, M. J. (2006a).

A non-parametric Bayesian approach to spike sorting.

In Proceedings of the IEEE Conference on Engineering in Medicine and Biological Systems, volume 28.



Wood, F. and Griffiths, T. L. (2006).

Particle filtering for nonparametric Bayesian matrix factorization.

In Advances in Neural Information Processing Systems, volume 18.



Wood, F., Griffiths, T. L., and Ghahramani, Z. (2006b).

A non-parametric Bayesian method for inferring hidden causes.

In Proceedings of the Conference on Uncertainty in Artificial Intelligence, volume 22.



Xing, E., Sharan, R., and Jordan, M. (2004).

Bayesian haplotype inference via the dirichlet process.

In Proceedings of the International Conference on Machine Learning, volume 21.



Xing, E. P., Jordan, M. I., and Sharan, R. (2007).

Bayesian haplotype inference via the Dirichlet process.

Journal of Computational Biology, 14:267–284.



Xing, E. P. and Sohn, K. (2007a).

Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space.

Bayesian Analysis, 2(2).



Xing, E. P. and Sohn, K. (2007b).

A nonparametric Bayesian approach for haplotype reconstruction from single and multi-population data.

Technical Report CMU-MLD 07-107, Carnegie Mellon University.

References XIII



Xing, E. P., Sohn, K., Jordan, M. I., and Teh, Y. W. (2006).

Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture.

In Proceedings of the International Conference on Machine Learning, volume 23.



Xu, Z., Tresp, V., Yu, K., and Kriegel, H.-P. (2006).

Infinite hidden relational models.

In Proceedings of the Conference on Uncertainty in Artificial Intelligence, volume 22.

Posterior Dirichlet Processes

- ▶ Suppose G is DP distributed, and θ is G distributed:

$$G \sim \text{DP}(\alpha, H)$$
$$\theta|G \sim G$$

- ▶ This gives $p(G)$ and $p(\theta|G)$.
- ▶ We are interested in:

$$p(\theta) = \int p(\theta|G)p(G) dG$$
$$p(G|\theta) = \frac{p(\theta|G)p(G)}{p(\theta)}$$

Posterior Dirichlet Processes

Conjugacy between Dirichlet Distribution and Multinomial.

- ▶ Consider:

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

$$z | (\pi_1, \dots, \pi_K) \sim \text{Discrete}(\pi_1, \dots, \pi_K)$$

z is a multinomial variate, taking on value $i \in \{1, \dots, n\}$ with probability π_i .

- ▶ Then:

$$z \sim \text{Discrete} \left(\frac{\alpha_1}{\sum_i \alpha_i}, \dots, \frac{\alpha_K}{\sum_i \alpha_i} \right)$$

$$(\pi_1, \dots, \pi_K) | z \sim \text{Dirichlet}(\alpha_1 + \delta_1(z), \dots, \alpha_K + \delta_K(z))$$

where $\delta_j(z) = 1$ if z takes on value i , 0 otherwise.

- ▶ Converse also true.

Posterior Dirichlet Processes

- ▶ Fix a partition (A_1, \dots, A_K) of Θ . Then

$$(G(A_1), \dots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \dots, \alpha H(A_K))$$

$$P(\theta \in A_i | G) = G(A_i)$$

- ▶ Using Dirichlet-multinomial conjugacy,

$$P(\theta \in A_i) = H(A_i)$$

$$(G(A_1), \dots, G(A_K)) | \theta \sim \text{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \dots, \alpha H(A_K) + \delta_\theta(A_K))$$

- ▶ The above is true for every finite partition of Θ . In particular, taking a really fine partition,

$$p(d\theta) = H(d\theta)$$

- ▶ Also, the posterior $G | \theta$ is also a Dirichlet process:

$$G | \theta \sim \text{DP} \left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1} \right)$$

Posterior Dirichlet Processes

$$\begin{array}{l} G \sim \text{DP}(\alpha, H) \\ \theta | G \sim G \end{array} \iff \begin{array}{l} \theta \sim H \\ G | \theta \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right) \end{array}$$

Pólya Urn Scheme

- ▶ First sample:

$$\begin{aligned} \theta_1 | G &\sim G & G &\sim \text{DP}(\alpha, H) \\ \iff \theta_1 &\sim H & G | \theta_1 &\sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}) \end{aligned}$$

- ▶ Second sample:

$$\begin{aligned} \theta_2 | \theta_1, G &\sim G & G | \theta_1 &\sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}) \\ \iff \theta_2 | \theta_1 &\sim \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1} & G | \theta_1, \theta_2 &\sim \text{DP}(\alpha + 2, \frac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha + 2}) \end{aligned}$$

- ▶ n^{th} sample

$$\begin{aligned} \theta_n | \theta_{1:n-1}, G &\sim G & G | \theta_{1:n-1} &\sim \text{DP}(\alpha + n - 1, \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}) \\ \iff \theta_n | \theta_{1:n-1} &\sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1} & G | \theta_{1:n} &\sim \text{DP}(\alpha + n, \frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}) \end{aligned}$$

Stick-breaking Construction

- ▶ Returning to the posterior process:

$$\begin{aligned} G &\sim \text{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \theta &\sim H \\ G | \theta &\sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}) \end{aligned}$$

- ▶ Consider a partition $(\theta, \Theta \setminus \theta)$ of Θ . We have:

$$\begin{aligned} (G(\theta), G(\Theta \setminus \theta)) | \theta &\sim \text{Dirichlet}((\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1}(\theta), (\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1}(\Theta \setminus \theta)) \\ &= \text{Dirichlet}(1, \alpha) \end{aligned}$$

- ▶ G has a point mass located at θ :

$$G = \beta \delta_\theta + (1 - \beta) G' \quad \text{with} \quad \beta \sim \text{Beta}(1, \alpha)$$

and G' is the (renormalized) probability measure with the point mass removed.

- ▶ What is G' ?

Stick-breaking Construction

- ▶ Currently, we have:

$$\begin{array}{l} G \sim \text{DP}(\alpha, H) \\ \theta \sim G \end{array} \quad \Rightarrow \quad \begin{array}{l} \theta \sim H \\ G|\theta \sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}) \\ G = \beta \delta_\theta + (1 - \beta)G' \\ \beta \sim \text{Beta}(1, \alpha) \end{array}$$

- ▶ Consider a further partition $(\theta, A_1, \dots, A_K)$ of Θ :

$$\begin{aligned} & (G(\theta), G(A_1), \dots, G(A_K)) \\ &= (\beta, (1 - \beta)G'(A_1), \dots, (1 - \beta)G'(A_K)) \\ &\sim \text{Dirichlet}(1, \alpha H(A_1), \dots, \alpha H(A_K)) \end{aligned}$$

- ▶ The agglomerative/decimative property of Dirichlet implies:

$$\begin{aligned} (G'(A_1), \dots, G'(A_K))|\theta &\sim \text{Dirichlet}(\alpha H(A_1), \dots, \alpha H(A_K)) \\ G' &\sim \text{DP}(\alpha, H) \end{aligned}$$

Stick-breaking Construction

- ▶ We have:

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

$$G = \beta_1 \delta_{\theta_1^*} + (1 - \beta_1) G_1$$

$$G = \beta_1 \delta_{\theta_1^*} + (1 - \beta_1)(\beta_2 \delta_{\theta_2^*} + (1 - \beta_2) G_2)$$

⋮

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

where

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i)$$

$$\beta_k \sim \text{Beta}(1, \alpha)$$

$$\theta_k^* \sim H$$

