



S&DS 365 / 665  
Intermediate Machine Learning

# **Approximate Inference: Gibbs Sampling for DP Mixtures**

March 2

Yale

# Reminders

- Assignment 2 due next Wednesday
- Quiz 2 available starting at 1pm today (CNN, GP, DP)
  - ▶ available for 48 hours
  - ▶ 30 minutes once started
- Midterm on March 16 in class
  - ▶ practice exam next week
  - ▶ review week of March 14

# For Today

- Recap: Dirichlet process algos and definitions
- DP process demo, comparison to parametric Bayes
- Dirichlet process mixtures
- Approximate inference with Gibbs sampling

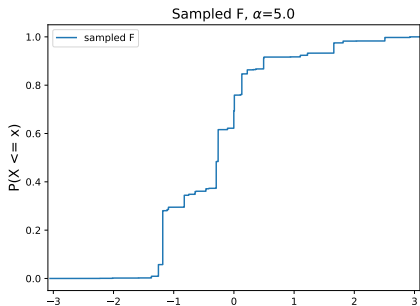
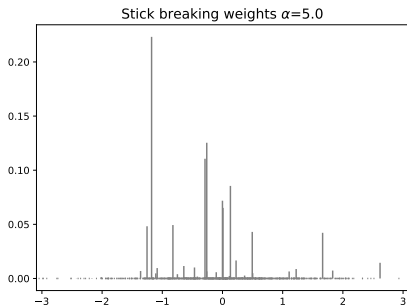
# The Dirichlet Process

- The Dirichlet process is analogous to the Gaussian process
- Every partition of sample space has a Dirichlet distribution
- GPs are tools for regression functions; DPs are tools for distributions and densities

# Dirichlet process

Each sample from a Dirichlet process prior has a *random collection of weights*, assigned to a *random selection of data*

# Sample from DP prior



# The posterior distribution

Let  $X_1, \dots, X_n \sim F$  and let  $F$  have prior  $\pi = DP(\alpha, F_0)$

Then the posterior  $\pi$  for  $F$  given  $X_1, \dots, X_n$  is

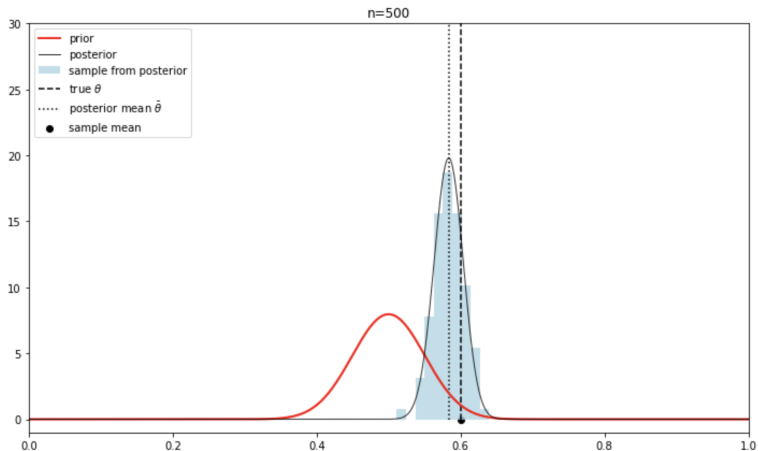
$$DP(\alpha + n, \bar{F}_n)$$

where

$$\bar{F}_n = \frac{n}{n + \alpha} F_n + \frac{\alpha}{n + \alpha} F_0.$$

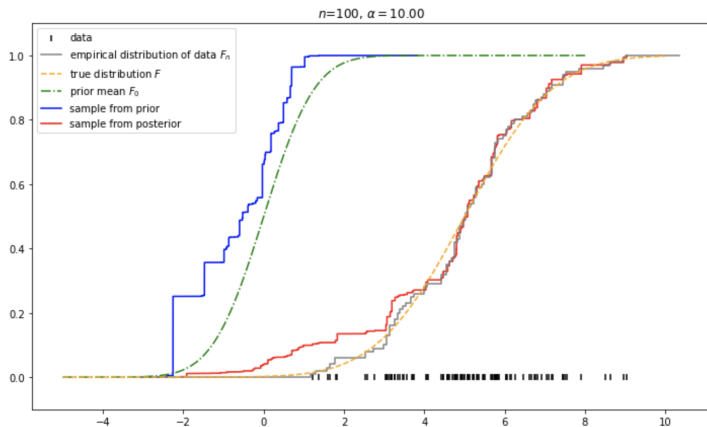
Here  $F_n$  is the empirical distribution of  $X_1, \dots, X_n$

# Demo: Parametric vs. Nonparametric Bayes





# Demo: Parametric vs. Nonparametric Bayes



# Clustering/repeats

Suppose we draw data  $F$  from a Dirichlet process, and then sample data from  $F$ :

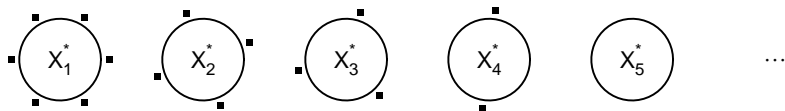
$$F \sim DP(\alpha, F_0)$$

$$X_1, X_2, \dots, X_n \mid F \sim F$$

Since  $F$  is a mixture model (of point masses), the samples  $X_i$  are clustered according to which mixture component they are sampled from.

The “Chinese restaurant process” captures this

# Chinese restaurant mnemonic



A customer (data point) comes into the restaurant and either

- 1 sits at an empty table, with probability proportional to  $\alpha$ , or
- 2 sits at an occupied table with probability proportional to number of customers already seated at that table

# Chinese restaurant process

- 1 Draw  $X_1 \sim F_0$ .
- 2 Given  $X_1, X_2, \dots, X_n$ , sample next point as

$$X_{n+1} \mid X_1, \dots, X_n = \begin{cases} X \sim F_n & \text{with probability } \frac{n}{n+\alpha} \\ X \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha} \end{cases}$$

where  $F_n$  is the empirical distribution of  $X_1, \dots, X_n$

This allows us to sample from the marginal distribution over  $X$ , without explicitly drawing a distribution  $F$  from the DP

# Chinese restaurant process

Let  $X_1^*, X_2^*, \dots$  denote unique values of  $X_1, \dots, X_n$

Define cluster assignment variables  $c_1, \dots, c_n$  where  $c_i = j$  means that  $X_i$  takes the value  $X_j^*$

Let  $n_j = |\{i : c_i = j\}|$ . Then

$$X_{n+1} = \begin{cases} X_j^* & \text{with probability } \frac{n_j}{n+\alpha} \\ X \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha} \end{cases}$$

This allows us to sample from the marginal distribution over  $X$ , without explicitly drawing a distribution  $F$  from the DP

# From DP to DPM

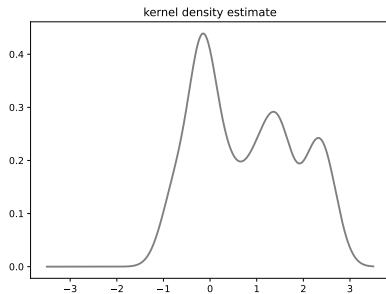
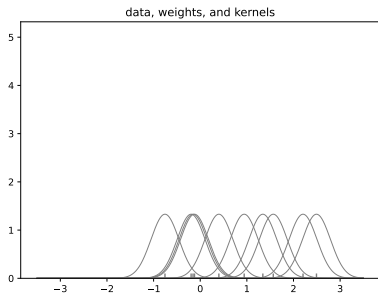
- A DP is a distribution over distributions
- A Dirichlet process mixture is a distribution over mixture models
- DPMs are Bayesian versions of kernel density estimation
- In stick breaking we replace  $X_i$  by  $\theta_i$
- In Chinese restaurant process we replace  $X_i^*$  by  $\theta_i^*$

## Recall: KDE

The *kernel density estimate* is the mixture model that places weight  $\frac{1}{n}$  on the kernel bump function centered on each data point:

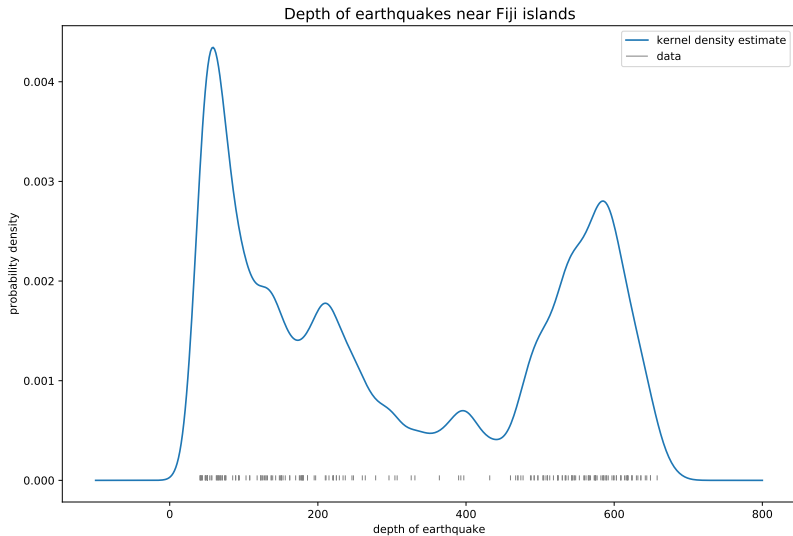
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$$

# Recall: KDE

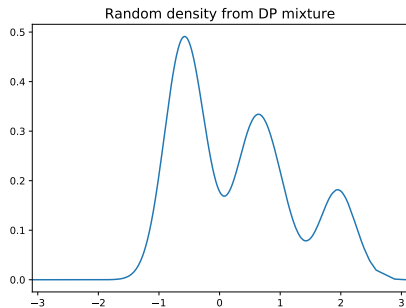
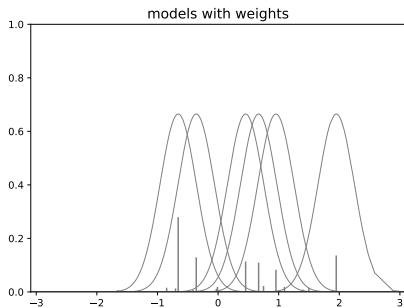




# Recall: KDE



# Sample from DP mixture



# Nonparametric Bayesian mixture model

$$\begin{aligned} F &\sim DP(\alpha, F_0) \\ \theta_1, \dots, \theta_n | F &\sim F \\ X_i | \theta_i &\sim f(x | \theta_i), \quad i = 1, \dots, n. \end{aligned}$$

# Stick breaking process for DPM

Stick breaking:

- At each step, break off a fraction  $V \sim \text{Beta}(1, \alpha)$

Sample model parameters:

- At each step, sample  $\theta \sim F_0$

# Stick breaking process for DPM

To draw a single random mixture from  $DPM(\alpha, F_0)$ :

- 1 Draw  $\theta_1, \theta_2, \dots$  independently from  $F_0$ .
- 2 Draw  $V_1, V_2, \dots \sim \text{Beta}(1, \alpha)$  and set  $w_j = V_j \prod_{i=1}^{j-1} (1 - V_i)$
- 3 Let  $f$  be the (infinite) mixture model

$$f(x) = \sum_{j=1}^{\infty} w_j f(x | \theta_j)$$

# Chinese restaurant process for a DPM

- 1 Draw  $\theta_1 \sim F_0$ .
- 2 Given  $\theta_1, \theta_2, \dots, \theta_n$  sample new model as

$$\theta_{n+1} \mid \theta_1, \dots, \theta_n = \begin{cases} \theta \sim F_n & \text{with probability } \frac{n}{n+\alpha} \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha} \end{cases}$$

where  $F_n$  is the empirical distribution of  $\theta_1, \dots, \theta_n$

# Chinese restaurant process for a DPM

Let  $\theta_1^*, \theta_2^*, \dots$  denote unique values of  $\theta_1, \dots, \theta_n$

Define cluster assignment variables  $c_1, \dots, c_n$  where  $c_i = j$  means that  $\theta_i$  takes the value  $\theta_j^*$

Let  $n_j = |\{i : c_i = j\}|$ . Then

$$\theta_{n+1} = \begin{cases} \theta_j^* & \text{with probability } \frac{n_j}{n+\alpha} \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha} \end{cases}$$

# The posterior for a DPM

- The posterior distribution does not have a closed form — need to approximate it algorithmically
- Two forms of approximations: Gibbs sampling and variational methods — next topic



# Gibbs sampling

We'll use the CRP to approximate the DPM posterior

Let's go to the chalk board!

# Summary

- A Dirichlet process mixture is a Bayesian version of kernel density estimation
- The posterior distribution cannot be computed explicitly—must be approximated
- Gibbs sampling approximates posterior by iteratively re-clustering the data
- Bayesian nonparametric methods require a lot of conceptual machinery and computation