

S&DS 365 / 665  
Intermediate Machine Learning

# **Nonparametric Bayes: Gaussian and Dirichlet Processes**

(continued)

February 28

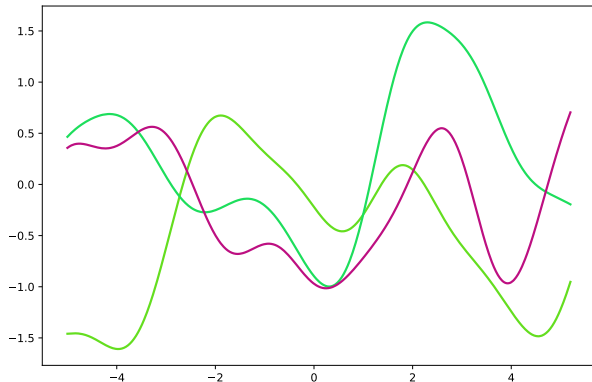
# Reminders

- Assignment 2 is out
- Quiz 2 on Wednesday (CNN, GP, DP)
- Midterm on March 16 in class
  - ▶ practice exam next week
  - ▶ review week of March 14

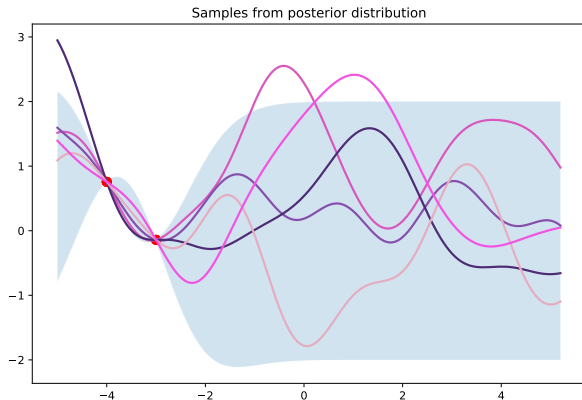
# For Today

- Dirichlet process demos and definitions
- Next topic: Approximate inference

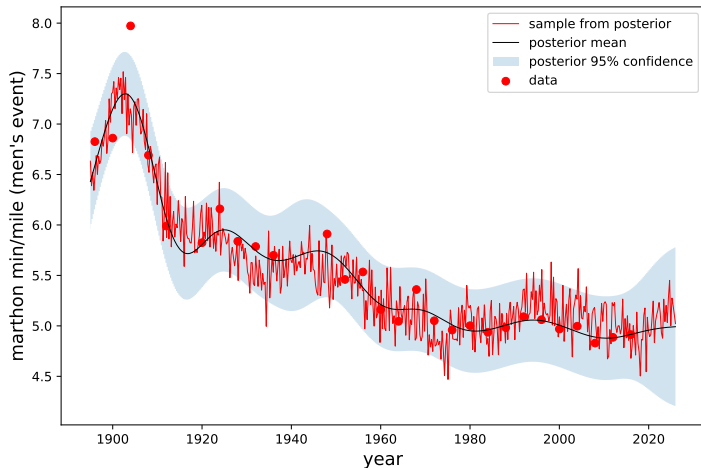
# Last week's demo: GP samples



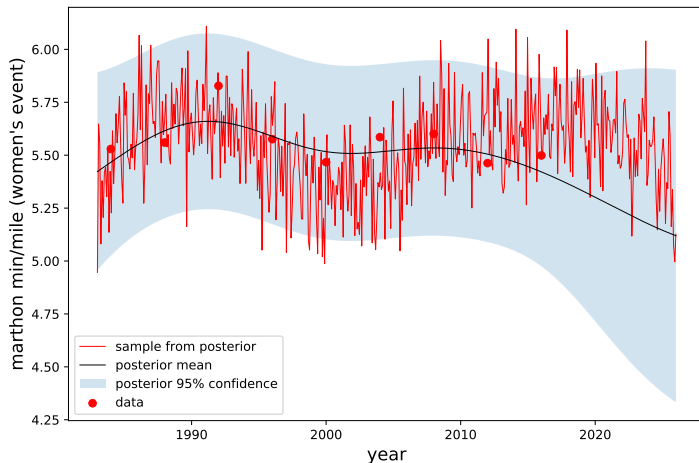
# Last week's demo: GP samples



# Olympic marathon times (men's race)



# Olympic marathon times (women's race)



# The Dirichlet Process

- The Dirichlet process is analogous to the Gaussian process
- Every partition of sample space has a Dirichlet distribution (more precise shortly)
- GPs are tools for regression functions; DPs are tools for distributions and densities
- DPs finesse the problem of choosing the number of components in a mixture model
  - ▶ Example: Don't need to specify the number of topics in a topic model



# The Dirichlet Process

Dirichlet processes have some fun mnemonic metaphors, which help understand the concepts:

- Stick breaking
- Chinese restaurants

But it's easy to get confused—we're working with probability distributions over probability distributions

## Starting point: CDF

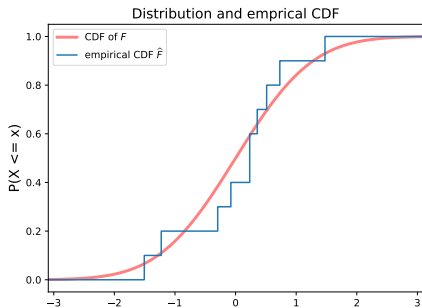
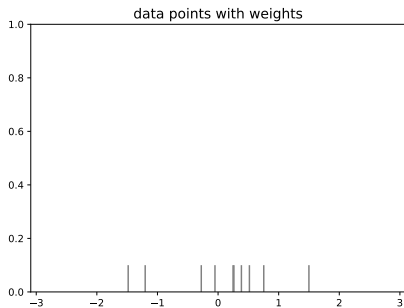
The *empirical distribution* of a set of data is the probability distribution that places probability mass  $\frac{1}{n}$  on each data point  $x_1, x_2, \dots, x_n$ .

The *empirical CDF* is the function

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_i \leq x)$$

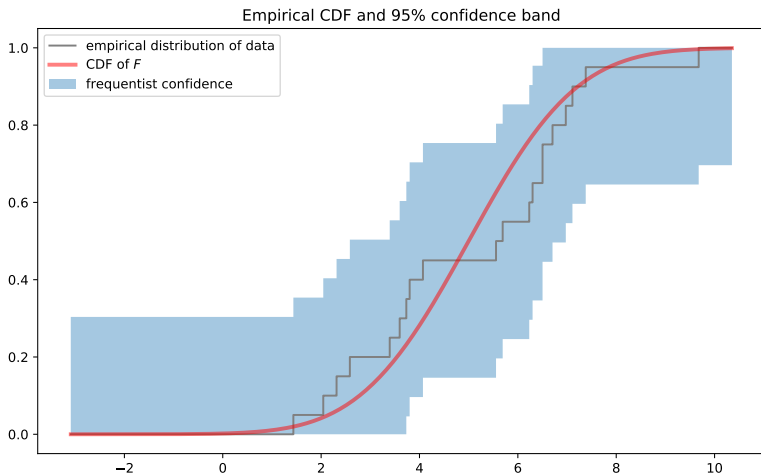
This is a step function with steps of size  $\frac{1}{n}$  on each data point.

# Empirical CDF



# Empirical CDF

A frequentist 95% confidence band is given by  $\hat{F}(x) \pm \sqrt{\frac{1}{2n} \log\left(\frac{2}{.05}\right)}$

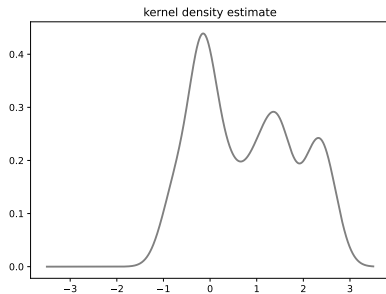
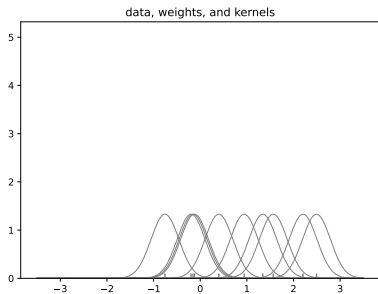


## Recall: KDE

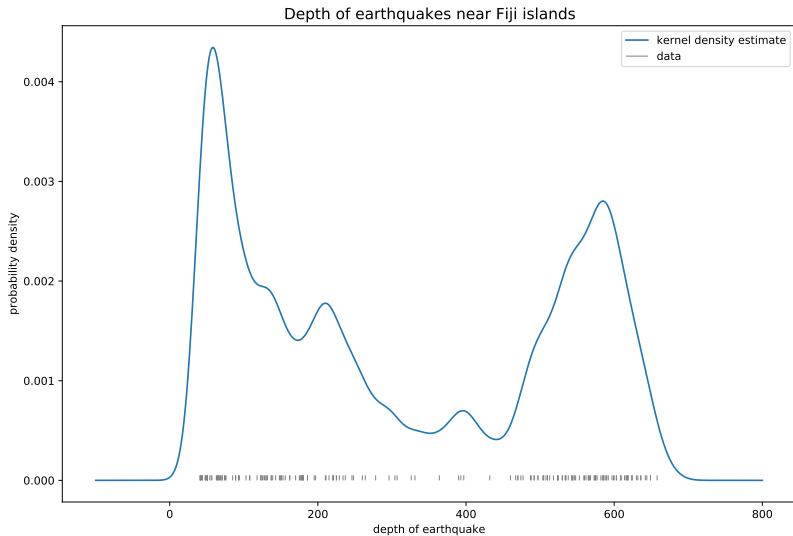
The *kernel density estimate* is the mixture model that places weight  $\frac{1}{n}$  on the kernel bump function centered on each data point:

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

# Recall: KDE



# Recall: KDE



# Getting rid of the data

Both the empirical CDF and kernel density estimate involve the data

We want to construct a *prior* distribution over these objects, before we see any data

Solution: Use synthetic or “imaginary” data!



# Dirichlet process

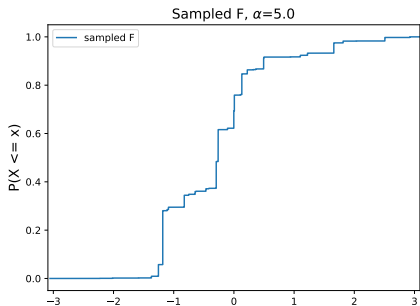
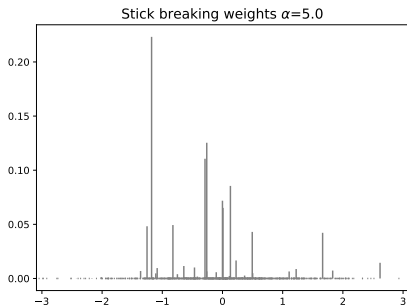
Each sample from a Dirichlet process prior has a *random collection of weights*, assigned to a *random selection of data*

Each sample from Dirichlet process mixture has a random collection of weights assigned to a random selection of *model parameters*

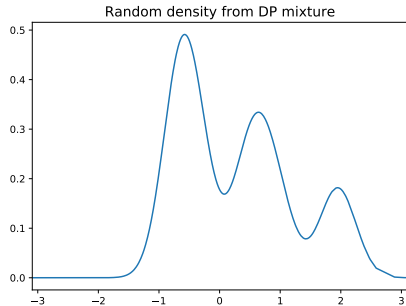
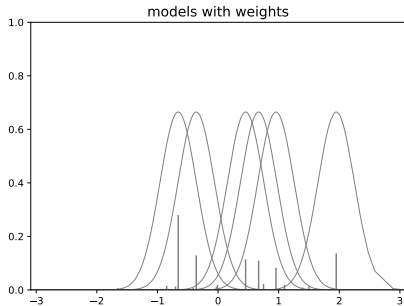
## Recall our sticking breaking demo



# Sample from DP prior



# Sample from DP mixture



# Stick breaking process

Stick breaking:

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

“Imaginary data”:

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

# Stick breaking process

To draw a single random distribution  $F$  from  $DP(\alpha, F_0)$ :

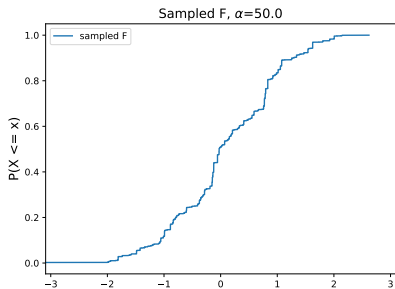
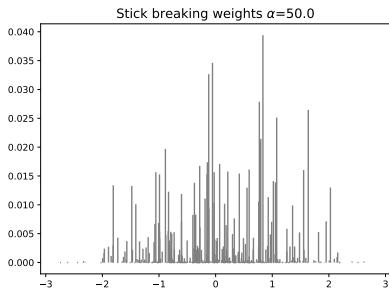
- 1 Draw  $s_1, s_2, \dots$  independently from  $F_0$ .
- 2 Draw  $V_1, V_2, \dots \sim \text{Beta}(1, \alpha)$  and set  $w_j = V_j \prod_{i=1}^{j-1} (1 - V_i)$
- 3 Let  $F$  be the discrete distribution that puts mass  $w_j$  at  $s_j$

# Stick breaking process

The mean of  $\text{Beta}(1, \alpha)$  is  $\frac{1}{1+\alpha}$ .

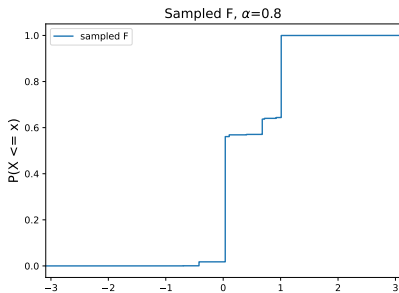
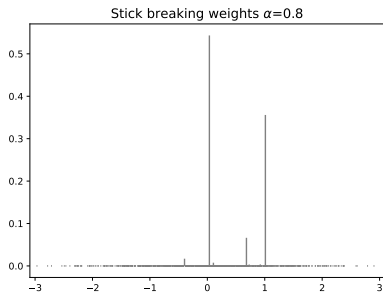
- As  $\alpha$  gets larger, the weights get smaller
- Weights always sum to one

# Different $\alpha$





# Different $\alpha$



# Clustering/repeats

Suppose we draw data  $F$ , drawn 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, the samples  $X_i$  are clustered according to which mixture component they are sampled from.

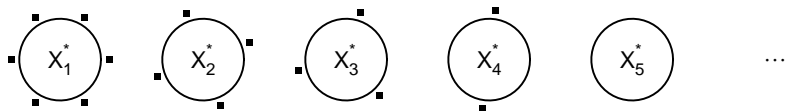
The “Chinese restaurant process” captures this

# Chinese restaurant mnemonic



Inspired by the large Chinese restaurants in San Francisco

# 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  $X_1, \dots, X_n$ , draw

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

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 given  $X_1, \dots, X_n$ ,

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

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

# 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  $X_1, \dots, X_n$

# DP Demo

RAM

Disk

⌂

🔍

<>

{x}

📁

+ Code

+ Text

Cannot save changes

⬆

⬇

🔗

⚙

⌵

⬆

⬇

🔗

✎

📄

🗑

⋮

## Dirichlet process demo

In this notebook we demonstrate the Dirichlet process, using the stick breaking construction. First we sample from the prior distribution  $\pi = DP(\alpha, F_0)$ , then we sample from the posterior distribution  $p(F | x)$  given a data set  $x_1, \dots, x_n$  that is drawn from a distribution that is different from the base distribution  $F_0$ .

```
[1] import numpy as np
import pandas as pd
from scipy.stats import norm
from IPython.display import clear_output
from time import sleep
import matplotlib.pyplot as plt
%matplotlib inline
```

Below we define three "helper" functions.

Given a parameter  $\alpha$ , the `stick_break` function returns a set of weights  $w_1, w_2, \dots, w_N$  given by

$$w_i = \begin{cases} V_0 & \text{if } i = 0 \\ V_i(1 - V_{i-1}) \cdots (1 - V_0) & \text{if } i > 0 \end{cases}$$

where the random variables  $V_0, V_1, \dots, V_N$  are independent draws from a  $\text{Beta}(1, \alpha)$  distribution. This gives a set of weights that sums to one (if  $N$  is large).

```
[2] def stick_break(alpha, N):
    v = np.random.beta(1, alpha, size=N)
```



# But what actually is a DP?

Recall:

A random function  $m$  is distributed according to a Gaussian process if for every  $x_1, x_2, \dots, x_n$  the random vector  $m(x_1), \dots, m(x_n)$  has a multivariate Gaussian distribution

$$N(\mu(x), K(x))$$

## But what actually is a DP?

A random distribution  $F$  is distributed according to a Dirichlet process  $DP(\alpha, F_0)$  if for every partition  $A_1, \dots, A_n$  of the sample space the random vector  $F(A_1), \dots, F(A_n)$  has a Dirichlet distribution

$$\text{Dir}(\alpha F_0(A_1), \alpha F_0(A_2), \dots, \alpha F_0(A_n))$$

## But what actually is a DP?

As a special case, if the sample space is the real line we can take the partition to be

$$A_1 = \{z : z \leq x\}$$

$$A_2 = \{z : z > x\}$$

and then

$$F(x) \sim \text{Beta}(\alpha F_0(x), \alpha(1 - F_0(x)))$$

# Big picture

The definition tells us the precise sense in which a DP is an infinite Dirichlet distribution

But this is not concrete

The sticking breaking and Chinese restaurant processes give us *algorithms* for working with a DP

# Big picture

Historically:

DP definition    $\longrightarrow$    CRP    $\longrightarrow$    Stick breaking

# Big picture

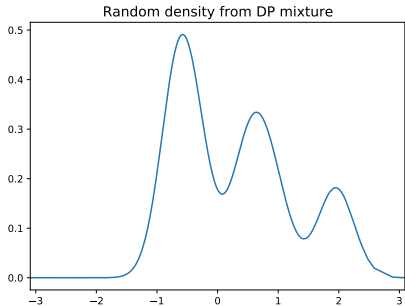
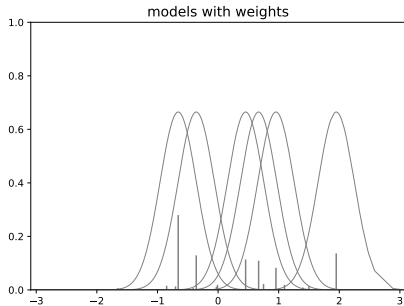
Conceptually, algorithmically:

DP definition  $\longleftarrow$  CRP  $\longleftarrow$  Stick breaking

# 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
- Subject to the curse of dimensionality!
- In stick breaking we replace  $X_i$  by  $\theta_i$
- In Chinese restaurant process we replace  $X_i^*$  by  $\theta_i^*$

# 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  $\text{DPM}(\alpha, F_0)$ :

- 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  $\theta_1, \dots, \theta_n$ , draw

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

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 given  $\theta_1, \dots, \theta_n$

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

# 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

# Summary

- A Dirichlet process is a prior over distribution functions
- The stick breaking process tells us how to sample  $F$
- The Chinese restaurant process tells us how to sample  $X$
- A Dirichlet process is a Bayesian version of the empirical CDF
- A Dirichlet process mixture is a Bayesian version of kernel density estimation
- Bayesian nonparametric methods require a lot of conceptual machinery and computation