Bayesian Nonparametrics

Lorenzo Rosasco

9.520 Class 21

April 26, 2010

About this class

Goal To give an overview of some of the basic concepts in Bayesian Nonparametrics. In particular, to discuss Dirichlet processes and their several characterizations and properties.

Plan

- Parametrics, nonparametrics and priors
- A reminder on distributions
- Dirichlet processes
  - Definition
  - Stick Breaking
  - Pólya Urn Scheme and Chinese processes

References and Acknowledgments

This lecture heavily draws (sometimes literally) from the list of references below, which we suggest as further readings.
Figures are taken either from Sudderth PhD thesis or Teh Tutorial.

Main references/sources:

- Yee Whye Teh, *Tutorial in the Machine Learning Summer School*, and his notes *Dirichlet Processes*.
- Erik Sudderth, PhD Thesis.
- Gosh and Ramamoorthi, *Bayesian Nonparametrics*, (book).

See also:

- Zoubin Ghahramani, Tutorial ICML.
- Michael Jordan, Nips Tutorial.
- Rasmussen, Williams, *Gaussian Processes for Machine Learning*, (book).
- Ferguson, paper in Annals of Statistics.
- Sethuraman, paper in *Statistica Sinica*.
- Berlinet, Thomas-Agnan, *RKHS in Probability and Statistics*, (book).

Parametrics vs Nonparametrics

We can illustrate the difference between the two approaches considering the following prototype problems.

1. function estimation
2. density estimation

(Parametric) Function Estimation

- Data, $S = (X, Y) = (x_i, y_i)_{i=1}^n$
- Model, $y_i = f_\theta(x_i) + \epsilon_i$,
  e.g. $f_\theta(x) = \langle \theta, x \rangle$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma > 0$.
- prior $\theta \sim P(\theta)$
- posterior

$$P(\theta|X, Y) = \frac{P(\theta)P(Y|X, \theta)}{P(Y|X)}$$

- prediction

$$P(y^*|x^*, X, Y) = \int P(y^*|x^*, \theta)P(\theta|X, Y)d\theta$$

(Parametric) Density Estimation

- Data, $S = (x_i)_{i=1}^n$
- Model, $x_i \sim F_\theta$
- prior $\theta \sim P(\theta)$
- posterior

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)}$$

- prediction

$$P(x^*|X) = \int P(x^*|\theta)P(\theta|X)d\theta$$

Nonparametrics: a Working Definition

- In the above models the number of parameters available for learning is fixed a priori.
- Ideally the more data we have, the more parameters we would like to explore.

This is in essence the idea underlying nonparametric models.

The Right to a Prior

- A finite sequence is *exchangeable* if its distribution does not change under permutation of the indeces.
- A sequence is *infinitely exchangeable* if any finite subsequence is exchangeable.

De Finetti's Theorem

If the random variables $(x_i)_{i=1}^{\infty}$ are infinitely exchangeable, then there exists some space Θ and a corresponding distribution $p(\theta)$, such that the joint distribution of n observations is given by:

$$P(x_1, \ldots, x_n) = \int_{\Theta} P(\theta) \prod_{i=1}^{n} P(x_i|\theta) d\theta.$$

Question

The previous classical result is often advocated as a justification for considering (possibly infinite dimensional) priors.

Can we find **computationally efficient** nonparametric models?

We already met one when we considered the Bayesian interpretation of regularization...

Reminder: Stochastic Processes

Stochastic Process

A family $(X_t) : (\Omega, P) \to \mathbb{R}$, $t \in T$, of random variables over some index set T.

Note that:
$X_t(\omega)$, $\omega \in \Omega$, is a number,
$X_t(\cdot)$ is a *random* variable,
$X_{(\cdot)}(\omega) : T \to \mathbb{R}$ is a function and is called *sample path*.

Gaussian Processes

$GP(f_0, K)$, Gaussian Process (GP) with mean f_0 and covariance function K

A family $(G_x)_{x \in X}$ of random variables over X such that: for any $x_1, \ldots, x_n$ in X, $G_{x_1}, \ldots, G_{x_n}$ is a multivariate Gaussian.

We can define the mean $f_0 : X \to \mathbb{R}$ of the GP from the mean $f_0(x_1), \ldots, f_0(x_n)$ and the covariance function $K : X \times X \to \mathbb{R}$ settting $K(x_i, x_j)$ equal to the corresponding entries of covariance matrix. Then K is symm., pos. def. function.

A sample path of the GP can be thought of as a random function

$$f \sim GP(f_0, K).$$

(Nonparametric) Function Estimation

- Data, $S = (X, Y) = (x_i, y_i)_{i=1}^n$
- Model, $y_i = f(x_i) + \epsilon_i$
- prior $f \sim GP(f_0, K)$
- posterior

$$P(f|X, Y) = \frac{P(f)P(Y|X, f)}{P(Y|X)}$$

- prediction

$$P(y^*|x^*, X, Y) = \int P(y^*|x^*, f)P(f|X, Y)df$$

We have seen that the last equation can be computed in closed form.

(Nonparametric) Density Estimation

Dirichlet Processes (DP) will give us a way to build nonparametric priors for density estimation.

- Data, $S = (x_i)_{i=1}^n$
- Model, $x_i \sim F$
- prior $F \sim DP(\alpha, H)$
- posterior

$$P(F|X) = \frac{P(F)P(X|F)}{P(X)}$$

- prediction

$$P(x^*|X) = \int P(x^*|F)P(F|X)dF$$

Plan

- Parametrics, nonparametrics and priors
- A reminder on distributions
- Dirichlet processes
  - Definition
  - Stick Breaking
  - Pólya Urn Scheme and Chinese processes

Dirichlet Distribution

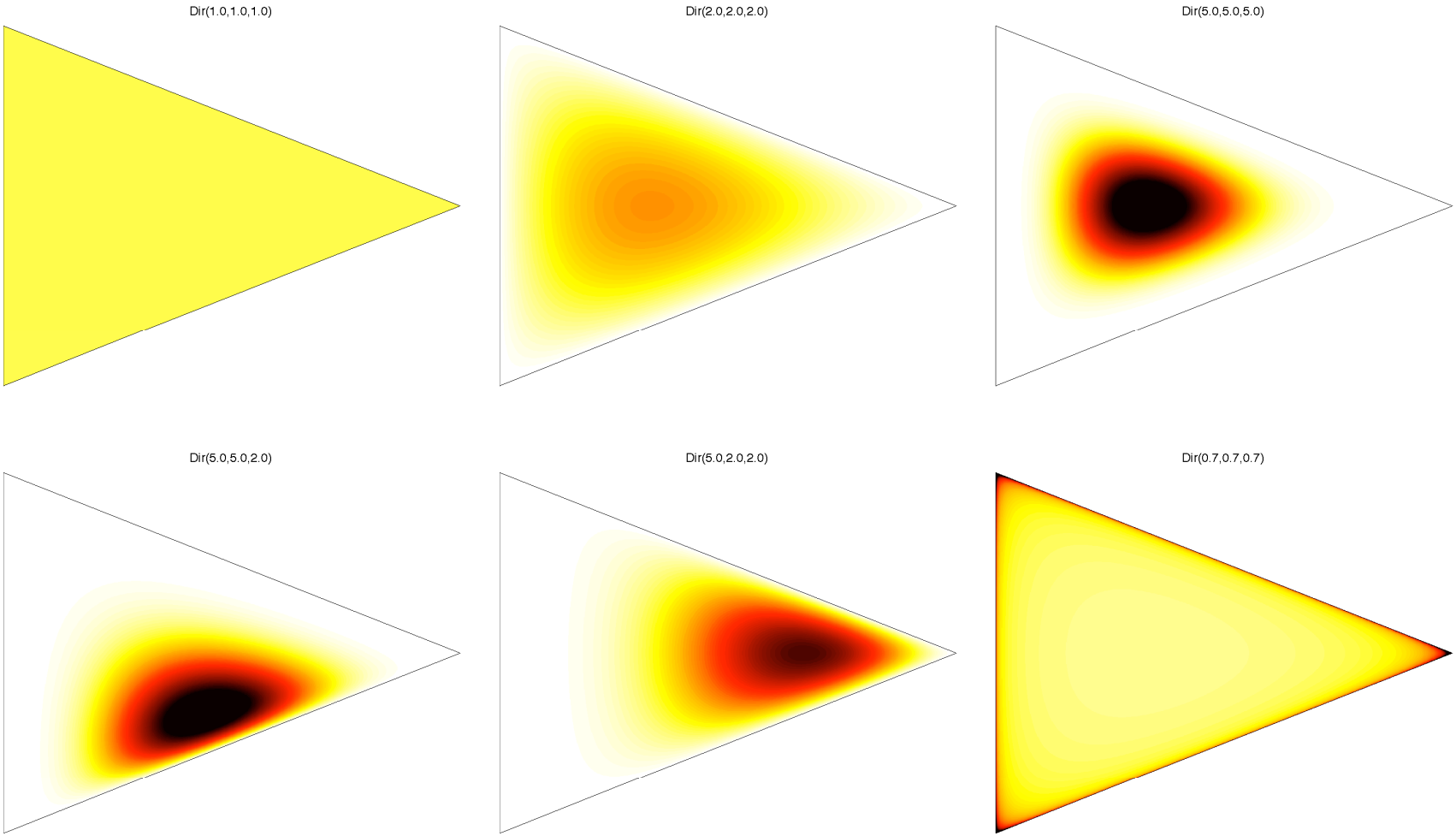
It is a distribution over the K-dimensional simplex $\mathbb{S}^K$, i.e. $x \in \mathbb{R}^K$ such that $\sum_{i=1}^K x^i = 1$ and $x^i \geq 0$ for all i.

The Dirichlet distribution is given by

$$P(x) = P(x^1, \ldots, x^K) = \frac{\Gamma(\sum_{i=1}^K \alpha^i)}{\sum_{i=1}^K \Gamma(\alpha^i)} \prod_{i=1}^K (x^i)^{\alpha^i - 1}$$

where $\alpha = (\alpha^1, \ldots, \alpha^K)$ is a parameter vector and Γ is the Gamma function.
We write $x \sim \text{Dir}(\alpha)$, i.e. $x^1, \ldots, x^K \sim \text{Dir}(\alpha^1, \ldots, \alpha^K)$.

Dirichlet Distribution



Reminder: Gamma Function and Beta Distribution

The Gamma function

$$\gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

It is possible to prove that $\Gamma(z+1) = z\Gamma(z)$.

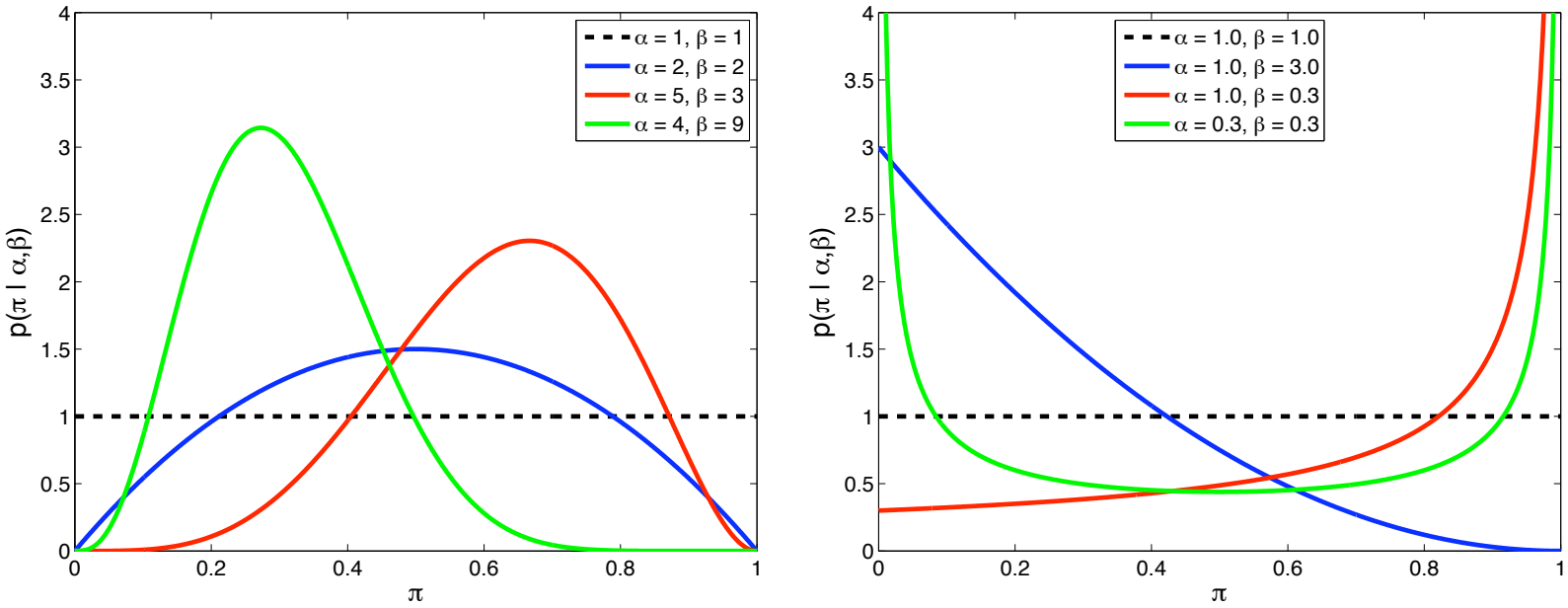
Beta Distribution

Special case of the Dirichlet distribution given by $K = 2$.

$$P(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)}$$

Note that here $x \in [0, 1]$ whereas for the Dirichlet distribution we would have $x = (x^1, x^2)$ with $x^1, x^2 > 0$ and $x^1 + x^2 = 1$.

Beta Distribution



For large parameters the distribution is unimodal. For small parameters it favors biased binomial distributions.

Properties of the Dirichlet Distribution

Note that the K-simplex $\mathbb{S}^K$ can be seen as the space of probabilities of a discrete (categorical) random variable with K possible values.
Let $\alpha_0 = \sum_{i=1}^{K} \alpha^i$.

- Expectation

$$\mathbb{E}[x^i] = \frac{\alpha^i}{\alpha_0}.$$

- Variance

$$\mathbb{V}[x^i] = \frac{\alpha^i(\alpha_0 - \alpha^i)}{\alpha_0^2(\alpha_0 + 1)}.$$

- Covariance

$$Cov(x^i, x^j) = \frac{\alpha^i \alpha^j}{\alpha_0^2(\alpha_0 + 1)}.$$

Properties of the Dirichlet Distribution

- Aggregation: let $(x^1, \ldots, x^K) \sim Dir(\alpha^1, \ldots, \alpha^K)$ then

$$(x^1 + x^2, \ldots, x^K) \sim Dir(\alpha^1 + \alpha^2, \ldots, \alpha^K).$$

  More generally, aggregation of any subset of the categories produces a Dirichlet distribution with parameters summed as above.
- The marginal distribution of any single component of a Dirichlet distribution follows a beta distribution.

Conjugate Priors

Let $X \sim F$ and $F \sim P(\cdot|\alpha) = P_\alpha$.

$$P(F|X,\alpha) = \frac{P(F|\alpha)P(X|F,\alpha)}{P(X,\alpha)}$$

We say that $P(F|\alpha)$ is a **conjugate prior** for the likelihood $P(X|F)$ if, for any X and α, the posterior distribution $P(F|X,\alpha)$ is in the same family of the prior. Moreover in this case the prior and the posterior distributions are then called conjugate distributions.

The Dirichlet distribution is conjugate to the multinomial distribution

Multinomial Distribution

Let X have values in $\{1, \ldots, K\}$. Let $\pi_1, \ldots, \pi_K$ define the probability mass function,

$$P(X|\pi_1, \ldots, \pi_K) = \prod_{i=1}^{K} \pi_i^{\delta_i(X)},$$

with $X \in \{1, \ldots, K\}$.

multinomial distribution

Given n observations the total probability of all possible sequences of length n taking those values is

$$P(x^1, \ldots, x^n|\pi_1, \ldots, \pi_K) = \frac{n!}{\prod_{i=1}^{K} C_i!} \prod_{i=1}^{K} \pi_i^{C_i},$$

where $C_i = \sum_{i=1}^{n} \delta_i(X^i)$.

For $K = 2$ this is just the binomial distribution.

Conjugate Posteriors and Predictions

Given n observations $S = x^1, \ldots, x^n$ from a multinomial distribution $P(\cdot|\theta)$ with a Dirichlet prior $P(\theta|\alpha)$ we have

$$P(\theta|S, \alpha) \propto P(\theta|\alpha)P(S|\theta) \propto$$

$$\prod_{i=1}^{K} (\theta^i)^{(\alpha_i + C_i - 1)} \propto \text{Dir}(\alpha_1 + C_1, \ldots, \alpha_K + C_K)$$

where C_i is the number of observations with value i.

Plan

- Parametrics, nonparametrics and priors
- A reminder on distributions
- Dirichlet processes
  - Definition
  - Stick Breaking
  - Pólya Urn Scheme and Chinese processes

Dirichlet Processes

Given a space X we denote with F a distribution on X and with $\mathcal{F}(X)$ the set of all possible distributions on X.

Informal Description

A Dirichlet process (DP) will be a distribution over $\mathcal{F}(X)$.
A sample from a DP can be seen as a (random) probability distribution on X.

Dirichlet Processes (cont.)

A partition of X is a collection of subsets $B_1, \ldots, B_N$ is such that, if $B_i \cap B_j = \emptyset$, $\forall i \neq j$ and $\cup_{i=1}^{N} B_i = X$.

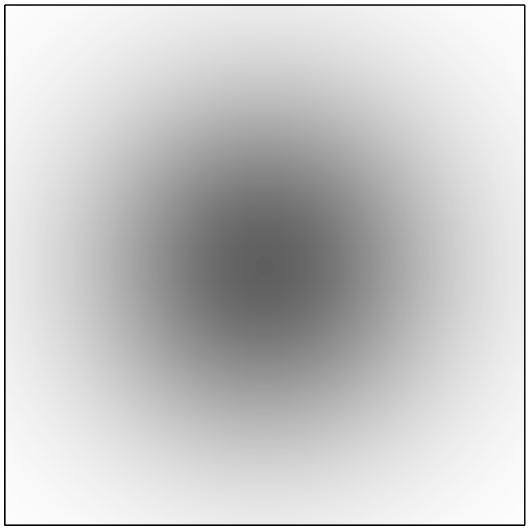
Definition (Existence Theorem)

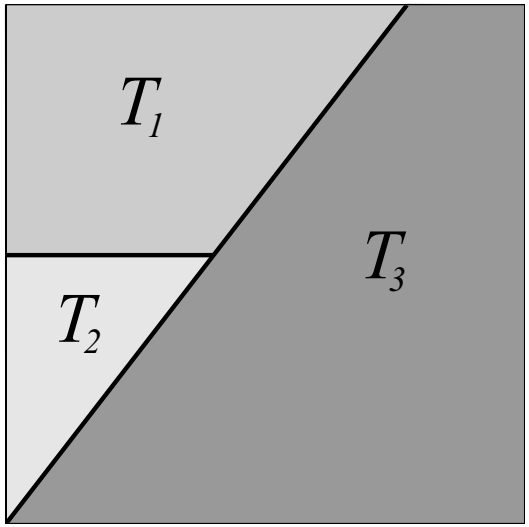
Let $\alpha > 0$ and H a probability distribution on X.
One can prove that there exists a unique distribution $DP(\alpha, H)$ on $\mathcal{F}(X)$ such that, if $F \sim DP(\alpha, H)$ and $B_1, \ldots, B_N$ is a partition of X then

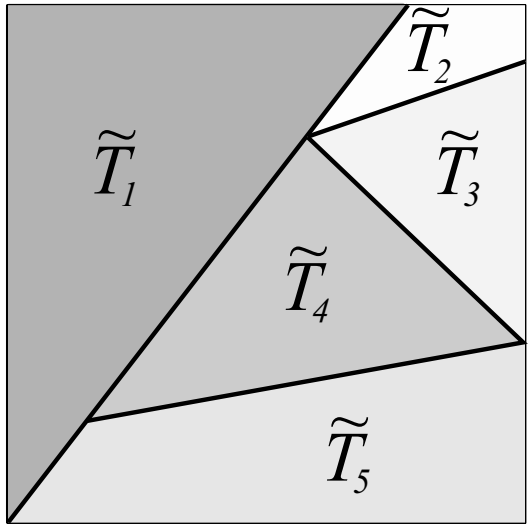
$$(F(B_1), \ldots, F(B_N)) \sim \text{Dir}(\alpha H(B_1), \ldots, \alpha H(B_N)).$$

The above result is proved (Ferguson '73) using Kolmogorov's Consistency theorem (Kolmogorov '33).

Dirichlet Processes Illustrated







Dirichlet Processes (cont.)

The previous definition is the one giving the name to the process.
It is in fact also possible to show that a Dirichlet process corresponds to a stochastic process where the sample paths are probability distributions on X.

Properties of Dirichlet Processes

Hereafter $F \sim DP(\alpha, H)$ and A is a measurable set in X.

- Expectation: $\mathbb{E}[F(A)] = \alpha H(A)$.
- Variance: $\mathbb{V}[F(A)] = \frac{H(A)(1-H(A))}{\alpha+1}$

Properties of Dirichlet Processes (cont.)

- Posterior and Conjugacy: let $x \sim F$ and consider a fixed partition $B_1, \ldots, B_N$, then

$$P(F(B_1), \ldots, F(B_N)|x \in B_k) =$$
$$\text{Dir}(\alpha H(B_1), \ldots, \alpha H(B_k) + 1, \ldots, \alpha H(B_N)).$$

It is possible to prove that if $S = (x_1, \ldots, x_n) \sim F$, and $F \sim DP(\alpha, H)$, then

$$P(F|S, \alpha, H) = DP\left(\alpha + n, \frac{1}{n+\alpha}\left(\alpha H + \sum_{i=1}^{n} \delta_{x_i}\right)\right)$$

Plan

- Parametrics, nonparametrics and priors
- A reminder on distributions
- Dirichlet processes
  - Definition
  - Stick Breaking
  - Pólya Urn Scheme and Chinese processes

A Qualitative Reasoning

From the form of the posterior we have that

$$\mathbb{E}(F(A)|S,\alpha,H) = \frac{1}{n+\alpha}\left(\alpha H(A) + \sum_{i=1}^{n}\delta_{x_i}(A)\right).$$

If $\alpha < \infty$ and $n \to \infty$ ona can argue that

$$\mathbb{E}(F(A)|S,\alpha,H) = \sum_{i=1}^{\infty}\pi_i\delta_{x_i}(A)$$

where $(\pi_i)_{i=1}^{\infty}$ is the sequence corresponding to the limit the empirical frequencies of the observations $(x_i)_{i=1}^{\infty}$.

If the posterior concentrates about its mean the above reasoning suggests that the obtained distribution is discrete.

Stick Breaking Construction

Explicit construction of a DP.

Let $\alpha > 0$, $(\pi_i)_{i=1}^{\infty}$ such that

$$\pi_i = \beta_i \prod_{j=1}^{i-1}(1-\beta_j) = \beta_i(1-\sum_{j=1}^{i-1}\pi_j)$$

where $\beta_i \sim Beta(1,\alpha)$, for all i.
Let H be a distribution on X and define

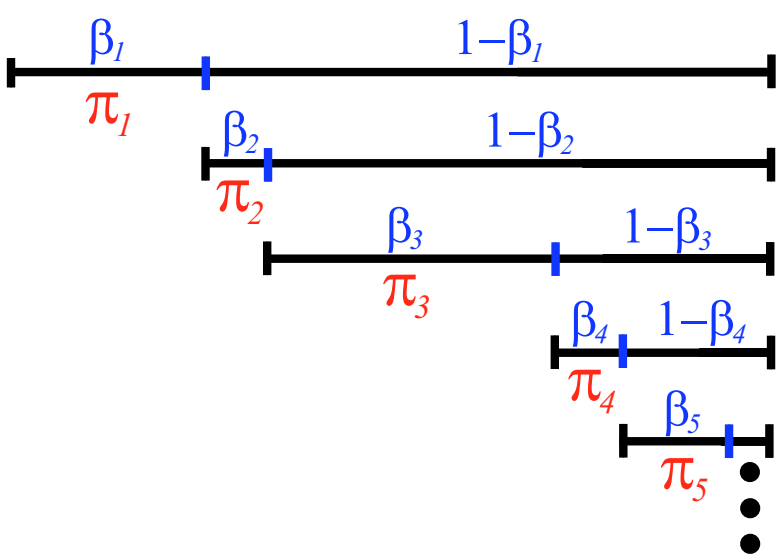
$$F = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$$

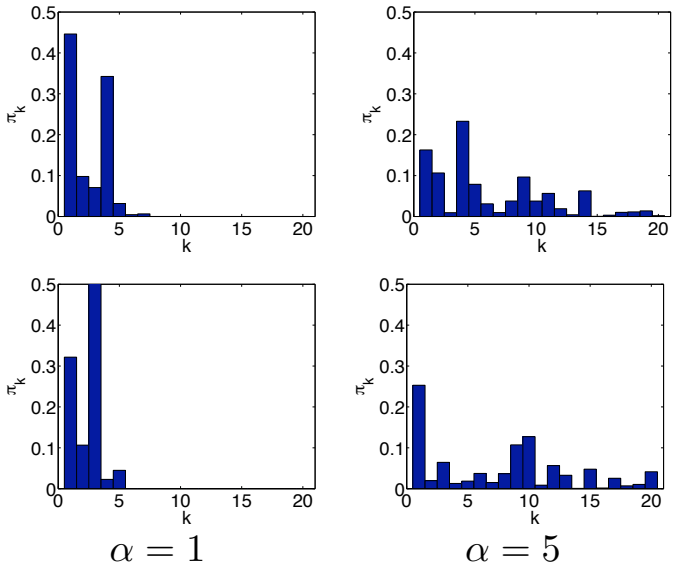
where $\theta_i \sim H$, for all i.

Stick Breaking Construction (cont.)

it is possible to prove (Sethuraman '94) that the previous construction returns a DP and conversely a Dirichlet process is discrete almost surely.

Stick Breaking Construction: Interpretation





$\alpha = 1$ $\alpha = 5$

The weights π partition a unit-length *stick* in an infinite set: the *i*-th weight is a random proportion β_i of the stick remaining after sampling the first $i - 1$ weights.

The Role of the Strength Parameter

Note that $\mathbb{E}[\beta_i] = 1/(1+\alpha)$.

- for small α, the first few components will have all the mass.
- for large α, F approaches the distribution H assigning uniform weights to the samples θ_i.

Plan

- Parametrics, nonparametrics and priors
- A reminder on distributions
- Dirichlet processes
  - Definition
  - Stick Breaking
  - Pólya Urn Scheme and Chinese processes

Pólya Urn Scheme

The observation that a sample from a DP is discrete allows to simplify the form of the prediction distribution,

$$\mathbb{E}(F(A)|S,\alpha,H) = \frac{1}{n+\alpha}\left(\alpha H(A) + \sum_{i=1}^{K} N_i \delta_{x_i}(A)\right).$$

where N_i are the number of observations with value i. In fact,

It is possible to prove (Blackwell and MacQueen '94) that if the base measure admits a density h, then

$$P(x^*|S,\alpha,H) = \frac{1}{n+\alpha}\left(\alpha h(x^*) + \sum_{i=1}^{K} N_i \delta_{x_i}(x^*)\right).$$

Chinese Restaurant Processes

The previous prediction distribution gives a distribution over partitions.
Pitman and Dubins called it Chinese Restaurant Processes (CRP) inspired by the seemingly infinite seating capacity of restaurants in San Francisco's Chinatown.

Chinese Restaurant Processes (cont.)

There is an infinite (countable) set of tables.

- First customer sits in 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.
- Customer n sits at table $k+1$ with probability

$$\frac{\alpha}{\alpha + n + 1}.$$

Number of Clusters and Strength Parameter

It is possible to prove (Antoniak '77??) that the number of clusters K grows as

$$\alpha \log n$$

as we increase the number of observations n.

Dirichlet Process Mixture

The clustering effect in DP arises from assuming that there are multiple observations having the same values.
This is hardly the case in practice.

Dirichlet Process Mixture (DPM)

The above observation suggests to consider the following model, $F \sim DP(\alpha, H)$,

$$\theta_i \sim F$$

and

$$x_i \sim G(\cdot|\theta_i).$$

Usually G is a distribution in the exponential family and $H = H(\lambda)$ a corresponding conjugate prior.

Dirichlet Process Mixture

CRP give another representation of the DPM.
Let z_i denote the unique cluster associated to x_i, then

$$z_i \sim \pi$$

and

$$x_i \sim G(\theta_{z_i}).$$

If we marginalize the indicator variables z_i's we obtain an infinite mixture model

$$P(x|\pi, \theta_1, \theta_2, \dots) = \sum_{i=1}^{\infty} \pi_i f(x|\theta_i)$$

Dirichlet Process and Model Selection

Rather than choosing a finite number of components K, the DP use the stick breaking construction to adapt the number of clusters to the data. The complexity of the model is controlled by the strength parameter α.

Conclusions

- DP provide a framework for nonparametric inference.
- Different characterizations shed light on different properties.

- DP mixtures allow to adapt the number of components to the number of samples...
- ...BUT the complexity of the model is controlled by the strength parameter α.
- Neither the posterior distribution nor the prediction distribution can be found analytically approximate inference is needed- see next class.

What about Generalizatoin Bounds?

Note that ideally $X \sim F$ and $F \sim DP(\alpha^*, H^*)$ for some α^*, H^* and we can compute the posterior

$$P^* = P(F|X, \alpha^*, H^*).$$

In practice we have only samples $S = (x_1, \ldots, x_n) \sim F$ and have to choose α, H to compute

$$P_n = P(F|S, \alpha, H).$$

- **[Consistency]** Does P_n approximate P^* (in some suitable sense)?
- **[Model Selection]** How should we choose α (and H)?