Nonparametric Bayesian Inference and Image Segmentation

October 7-21, 2011

- Dirichlet Process
- Dependent Dirichlet Process
- Pitman-Yor Process
- Gaussian Process
Image Segmentation

We will start with nonparametric Bayesian approach to image segmentations.

- What is a Bayesian approach?
- What is a nonparametric Bayesian approach?
- What is image segmentation?

References

Segmentation is the partitioning of an image into multiple segments with the aim of simplifying the image representation. In particular, segments are supposed to be coherent regions in the image.

Two important and often difficult design decisions:

- model for individual segments (histogram clustering? Mixture of Gaussians? k-means clustering?)
- number of segments (how do we determine the correct number?)

With nonparametric Bayesian approach, the number of segments can be determined directly from the data given the specified model.
Bayesian and Nonparametric Bayesian

Inference is usually a backward process, given a known forward process.
For example, the segmentation algorithm usually assume the following generative model:

\[ x_i \sim F(\cdot | \theta_i), \quad \theta_i \sim G; i = 1, \ldots, n \]
\[ G \sim \text{SP}(\alpha G_0). \]

where \( x_i \) are the observed data and SP the given stochastic process with Bayesian prior specified by a constant \( \alpha \) and distribution \( G_0 \).
In this setup, only \( x_i \), and the Bayeisan prior \((\alpha, G_0)\) are given. \( \theta_i \) are to be inferred from the data and the given model.
Image Segmentation Example

\[ x_i \sim F(\cdot | \theta_i), \quad \theta_i \sim G; \quad i = 1, \ldots, n \]

\[ G \sim \text{SP}(\alpha G_0). \]

How do we relate the abstract model to images?

The data \( x_i \) are to be thought of as features derived/computed from the image. For example, color, intensity, and others. Each \( \theta_i \) is assumed to determine one segment in the image (\( \theta_i \) can have the same value).
Bayesian and Nonparametric Bayesian

The Bayesian approach is parametric if we assume the form of the distribution responsible for each class of data.

It is nonparametric if no such form is assumed. The Bayesian prior is then formulated as distribution on the space of probability distributions. The distribution on the space of probability distributions (measures) is a stochastic process.

\[ M(X), \text{ the space of probability distributions (measures) on } X. \]
The MDP/MRF model is applied to synthetic aperture radar (SAR) images and magnetic resonance imaging (MRI) data.

**Fig. 1.** Segmentation results on real-world radar data. Original image (left), unconstrained MDP segmentation (middle), MDP segmentation with smoothness constraint (right).

**Fig. 2.** A SAR image with a high noise level and ambiguous segments (left). Solutions without (middle) and with smoothing (right).
Fig. 3. MR frontal view image of a monkey’s head. Original image (left), smoothed MDP segmentation (middle), original image overlaid with segment boundaries (right).

Fig. 4. Segmentation result for multichannel data: A SAR image with three channels (left), segmentation result obtained with the MDP/MRF model, and the original image overlaid with segment boundaries (right).
Segmentation Results from [2]
Segmentation Results from [2]
Segmentation Results from [2]
Segmentation Results from [4]
Dirichlet Process Mixture

In many recent publications on nonparametric Bayesian approach to machine learning, a commonly used generated model is

\[ x_i \sim F(\cdot|\theta_i), \quad \theta_i \sim G; \quad i = 1, \ldots, n \]

\[ G \sim \text{DP}(\alpha G_0). \]

where \( \text{DP}(\alpha G_0) \) is the {Dirichlet Process}.  

1. \( \alpha \) the concentration parameter,  
2. \( G_0 \) the base measure.  
3. \( F(\cdot|\theta_i) \) is a Gaussian distribution given \( \theta_i \).

The entire model is called Dirichlet Process Mixture or DPM for short. To understand it better, we will start with  

1. Dirichlet Distribution  
2. Dirichlet Process
Dirichlet Distribution is a family $\text{Dir}_k(\alpha)$ of continuous multivariate probability distributions defined on the standard $k - 1$ simplex $S_k$ in $\mathbb{R}^k$ parameterized by $\alpha = (\alpha_1, \cdots, \alpha_K)$, $\alpha_1, \cdots, \alpha_K > 0$.

The density function is

$$f(p_1, \cdots, p_k; \alpha_1, \cdots, \alpha_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i-1}$$

where $\Gamma(n)$ is the Gamma function and $(p_1, \cdots, p_k)$ are points in $S_k$:

$$(p_1, \cdots, p_k) \in S_k \rightarrow p_1, \cdots, p_k \geq 0, p_1 + \cdots + p_k = 1.$$
First Properties

When $k = 2$, $\text{Dir}_2(\alpha_1, \alpha_2)$ is the Beta distribution $\text{Beta}(\alpha_1, \alpha_2)$, distribution defined on the unit interval in $\mathbb{R}^2$.

Let $\bar{\alpha} = \alpha_1 + \cdots + \alpha_K$.

$$
\mathbb{E}[p_i] = \frac{\alpha_i}{\bar{\alpha}}, \quad \text{Var}[p_i] = \frac{\mathbb{E}[p_i](1 - \mathbb{E}[p_i])}{\bar{\alpha} + 1}
$$
Relation with Gamma Distribution

If $Z_1, \cdots, Z_k$ are independent Gamma random variable with parameter $\alpha_i \geq 0$, and $Z = Z_1 + \cdots Z_k$, then

$$\left( \frac{Z_1}{Z}, \cdots \frac{Z_k}{Z} \right)$$

is distributed as $\text{DD}_k(\alpha_1, \cdots, \alpha_k)$.

Recall that Gamma distribution has the density function

$$\gamma(x, \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

for $x \geq 0$. $\alpha$ is the shape parameter and $\beta$ the inverse scale parameter.
Interpreting $\alpha$

Let $X$ denote the finite set with $k$ elements, $X = \{1, 2, \cdots, k\}$. Each point $p \in S_k$ defines a probability distribution (measure) on $X$. The simplex $S_k$ is the space $m(X)$ of probability measures on $X$, and $\text{Dir}_k(\alpha)$ is a probability distribution on $m(X)$. $\alpha$ will be viewed as a measure on $X$ with total measure $\bar{\alpha} = \alpha_1 + \cdots + \alpha_k$. Dirichlet distribution $\text{Dir}_k(\alpha)$ is then parameterized by measures $\alpha$ on $X$. 
Measure on Finite Sets

For finite set $\mathbb{X}$, a measure $\mu$ on $\mathbb{X}$ is specified by the value of $\mu$ at each point in $\mathbb{X}$.

For example, $\mathbb{X} = \{1, 2, ..., k\}$. A subset $U = \{2, 3, 4\}$.

$$\mu(U) = \mu(2) + \mu(3) + \mu(4).$$

$\mu(i)$ is the measure assigned to $\{i\}$, and $U$ is the disjoint union of $\{2\}, \{3\}, \{4\}$. By countable additivity, $\mu(U)$ must be $\mu(2) + \mu(3) + \mu(4)$.

Clearly, $\delta_1(U) = 0$ and $\delta_2(U) = 1$, where $\delta_i$ is the delta measure at $i$.

The $\alpha$ exponent $(\alpha_1, ..., \alpha_k)$ in the density formula of Dirichlet distribution can be treated as a measure on $\mathbb{X}$. 
Using the relation with Gamma distribution, it is easy to show

**Proposition**

*If* $p = (p_1, \cdots, p_k)$ *is distributed as* $\text{Dir}_k(\alpha)$, *then for any partition* $U_1, \cdots, U_m$ *of* $X = \{1, 2, \cdots, k\}$, *the random vector*

$$(p(U_1), \cdots, p(U_m)) = \left( \sum_{i \in U_1} p_i, \cdots, \sum_{i \in U_m} p_i \right)$$

*is distributed as*

$$\text{Dir}_m(\alpha(U_1), \cdots, \alpha(U_m))$$
For example, take $k = 10$ and $\alpha = \alpha_1, \alpha_2, ..., \alpha_{10}$. $U_1 = \{1, 2, 3, 4\}, U_2 = \{5, 6, 7\}, U_3 = \{8, 9, 10\}$ is a partition of $X$.

If $p \in S_{10}$ is distributed as $\text{Dir}_{10}(\alpha)$, then $(p_1 + p_2 + p_3 + p_4, p_5 + p_6 + p_7, p_8 + p_9 + p_{10})$ is distributed as $
\begin{align*}
\text{Dir}_3(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4, \alpha_5 + \alpha_6 + \alpha_7, \alpha_8 + \alpha_9 + \alpha_{10}).
\end{align*}$

Or

$(p(U_1), p(U_2), p(U_3)) \sim \text{Dir}_3(\alpha(U_1), \alpha(U_2), \alpha(U_3))$.

Recall that $\bar{\alpha} = \alpha(U_1) + \alpha(U_2) + \alpha(U_3)$.

In particular,

$p_i \sim \text{Beta}(\alpha_i, \bar{\alpha} - \alpha_i)$. 

View the proposition from a probabilistic/geometric angle (instead of algebraic):

Taking any partition of $\mathbb{X}$. For example, $\mathbb{X} = U_1 \cup U_2 \cup U_3$. This gives a (measurable) map

$$e^k_3 : m(\mathbb{X}) \rightarrow m(\mathbb{X}_3), \quad p \in m(\mathbb{X}_3) \rightarrow (p(U_1), p(U_2), p(U_3)).$$

$m(\mathbb{X})$ is the space of probability measures (distributions) on $\mathbb{X}$. Our original Dirichlet distribution is a distribution on $m(\mathbb{X})$, and it can be pushed down to (lower-dimensional simplices) $m(\mathbb{X}_d)$ for $d < k$ using the map $e^k_d$. 
In particular, we have the obvious compatibility (i.e., summing over $\alpha_i$):

\[ m(X) \xrightarrow{e_k} m(X_m) \xrightarrow{e_n} e_n^m \Rightarrow m(X_n) \]

$X_m$ comes from the partition $U_1, \ldots, U_m$ and $X_n$ comes from the partition $V_1, \ldots, V_n$ such that $V_i$ are the disjoint unions of $U_i$. In particular, a (finite-dimensional) Dirichlet distribution can be characterized by a collection of compatible lower-dimensional Dirichlet distributions. Any connection with what we have learnt so far?
Properties of Dirichlet Distribution

For each \( p \in \mathbf{M}(\mathbb{X}) \), let \( X_1, X_2, ..., X_n \) be i.i.d. \( \sim p \) and \( p \sim \text{Dir}_k(\alpha) \). Then the posterior is proportional to

\[
\prod_{i=1}^{k} \frac{\alpha_i - 1 + n_i}{\alpha_i - 1 + n_i},
\]

where \( n_i = \#\{j : X_j = i\} \). Since Posterior \( \propto \) likelihood \( \times \) prior, we have

\[
\text{posterior} = P(p|X_1, ..., X_n) \propto \prod_{i=1}^{n} P(X_i|p)P(p).
\]

\( P(X_i|p) = p_w, \) if \( X_i = w \) and \( P(p) \) is Dirichlet. The result follows.
Properties of Dirichlet Distribution

The posterior distribution of $p$ given $X_1, X_2, ..., X_n$ is $\text{Dir}_k(\alpha + \sum \delta_{X_i})$.

Notice that $\text{Dir}_k(\alpha + \sum \delta_{X_i}) = \text{Dir}_k(\alpha_1 + n_1, ..., \alpha_k + n_k)$.

Need to get use to the notation (the payback is substantial)!

For example, let $\alpha' = \alpha + \sum \delta_{X_i}$. Then for $w \in \mathbb{X}$,

$$\alpha'(w) = \alpha(w) + \sum_{i=1}^{n} \delta_{X_i}(w) = \alpha_i + n_w,$$

which give the exponents in the posterior distribution.

The Posterior distribution depends of course on the data, $X_1, ..., X_n$. 
The predictive distribution (from the posterior) is a distribution on $X$, i.e., an element in $m(X)$. In nonparametric Bayesian, this distribution is given as

$$P(X_{n+1} | X_1, \ldots, X_n) = \int_{m(X)} p \, dP(p | X_1, \ldots, X_n).$$

This gives the following formula (from the formula on Slide 15) for the predictive distribution of $X_{n+1}$ given $X_1, \ldots, X_n$ is

$$P(X_{n+1} | X_1, \ldots, X_n) = \frac{\alpha + \sum_{i=1}^{n} \delta x_i}{\alpha(X) + n}$$
What Does the Formula Say?

Let $\mathbb{X} = \{1, 2, 3, \ldots, 10\}$, and take $\alpha = \{2, 2, \ldots, 2\}$. Given $X_1 = 2, X_2 = 3, X_3 = 2$. What is the distribution $P(X_4|X_1, X_2, X_3)$ according to the formula?

$$P(X_{n+1}|X_1, \ldots, X_n) = \frac{\alpha + \sum_{i=1}^n \delta_{X_i}}{\alpha(\mathbb{X}) + n}$$

$$P(X_4 = 3|X_1 = 2, X_2 = 3, X_3 = 2) = \frac{2 + 1}{20 + 3} = \frac{3}{23}.$$  

$$P(X_4 = 2|X_1 = 2, X_2 = 3, X_3 = 2) = \frac{2 + 2}{20 + 3} = \frac{4}{23}.$$  

$$P(X_4 \in \{2, 3\}|X_i) = \frac{\alpha(\{2, 3\}) + \sum_{i=1}^3 \delta_{X_i}(\{2, 3\})}{20 + 3} = \frac{7}{23}.$$
What Does the Formula Say?

It is straightforward (for finite-dimensional case) to see that as $n \to \infty$

$$\frac{\alpha + \sum_{i=1}^{n} \delta X_i}{\alpha(X) + n} \to \text{Empirical Distribution} \propto \{n_1/n, \ldots, n_k/n\}.$$ 

The total measure $\alpha(X)$ control the rate of the convergence, i.e., the strength of ones’ belief.
An urn with $\alpha(X)$ balls of which $\alpha_j$ are of color $i$, $i = 1, \ldots, k$. Draw balls at random from the urn, replacing each ball drawn by two balls of the same color.

$$P(X_1 = j) = \frac{\alpha_j}{\bar{\alpha}}, \quad P(X_2 = j|X_1) = \frac{\alpha_j + \delta_{X_1}(j)}{\bar{\alpha} + 1}, \ldots$$

In general,

$$P(X_{n+1} = j|X_1, \ldots, X_n) = \frac{\alpha_j + \sum_{i=1}^{n} \delta_{X_i}(j)}{\bar{\alpha} + n}.$$
Exchangeable Sequence

For example, take $k = 2$ so there are two colors, red (r) and black (b). Take $n = 3$. The following two joint distributions are the same

$$P(X_1 = r, X_2 = b, X_3 = r) = \frac{\alpha_1}{\bar{\alpha}} \frac{\alpha_2}{\bar{\alpha} + 1} \frac{\alpha_1 + 1}{\bar{\alpha} + 2}$$

and

$$P(X_1 = r, X_2 = r, X_3 = b) = \frac{\alpha_1}{\bar{\alpha}} \frac{\alpha_1 + 1}{\bar{\alpha} + 1} \frac{\alpha_2}{\bar{\alpha} + 2}$$

The general formula

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{\alpha_1^{[n_1]} \ldots \alpha_k^{[n_k]}}{\bar{\alpha}^{[n]}}$$

(where $a^{[n]}$ is the ascending factorial) showing the sequence $X_1, X_2, \ldots, X_n$ is exchangeable.
De Finett’s theorem on exchangeable sequences show that there is a probability distribution $\Pi$ on $m(X)$ such that

$$P(X_1 = x_1, X_2 = x_2, .., X_n = x_n) = \int_{m(X)} p_1^{n_1} ... p_k^{n_k} d\Pi(p).$$

The joints are given by the previous formula, which are also the formula for the moments of $\text{Dir}_k(\alpha)$. This shows that $\Pi = \text{Dir}_k(\alpha)$. 

Exchangeable Sequence
Dirichlet Process (Ferguson, 1973)

Let $X$ denote an abstract topological space (typically required to be separable and a complete metric space (e.g., $\mathbb{R}^1$)) and $\mathcal{B}(X)$ a $\sigma$-algebra of subsets of $X$. $m(X)$ denote the space of probability measures on $X$.

The goal is to define a probability measure $\mathcal{P}$ on $m(X)$.

An Important Point  Given the probability measure $\mathcal{P}$ and a Borel set $B \subset X$, $\mathcal{P}(B)$ is a random variable.

The idea is to define $\mathcal{P}$ indirectly by specifying the joint distribution of

$$(P(B_1), \ldots, P(B_m)), \quad P \sim \mathcal{P}$$

for every finite measurable partition $B_1, \ldots, B_m$ of $X$. 
Dirichlet Process

The data on joint distributions of finite measurable partitions is sufficient to specify the joint distribution of $P(A_1, \ldots, A_m)$ for arbitrary measure sets $A_1, \ldots, A_m$. The trick is to use the $2^m$ disjoint sets obtained by taking intersections of $A_i$ and their complements

$$B_{v_1,\ldots,v_m} = \bigcup_{j=1}^m A_{v_j}^j.$$

We define the Dirichlet Process $\mathcal{DP}(\alpha)$ as an (abstract) probability measure on $\mathcal{m}(\mathbb{X})$ (with $\alpha$ a measure on $\mathbb{X}$) such that for an arbitrary measurable partition $B_1, \ldots, B_m$ of $\mathbb{X}$, the joint

$$P(P(B_1), \ldots, P(B_m)) \sim \text{Dir}_m(\alpha(B_1), \ldots, \alpha(B_m)).$$
Consistency

To ensure the data indeed define an abstract probability measure on $m(X)$, one needs to verify the Kolmogorov consistency conditions (follow directly from the properties of Dirichlet Distributions).

Condition

Let $\mathcal{B} = (B_1, \cdots, B_k)$ and $\mathcal{C} = (C_1, \cdots, C_l)$ be two measurable partitions of $X$ such that $\mathcal{C}$ is a refinement of $\mathcal{B}$. That is, every $B_i$ is a union $\bigcup_{i \in b_j} C_i$ of sets in $\mathcal{C}$. Then the distribution of

$$\left(\sum_{i \in b_1} P(C_i), \cdots, \sum_{i \in b_k} P(C_i)\right)$$

as determined from the joint distribution of $(P(C_1), \cdots, P(C_l))$, is identical to the distribution of $(P(B_1), \cdots, P(B_k))$. 
Recall that a stochastic process is a collection of random variables \( \{X_i\}_{i \in I} \) where \( I \) is the index set. If \( I = \mathbb{R} \), we have the continuous-time stochastic process.

Take \( X = \mathbb{R}^1 \), and identify each element in \( m(X) \) with its cumulative distribution function \( F \). The probability measure \( \mathcal{P} \) on \( m(X) \) now define a family of random variables parameterized by \( \mathbb{R} \) through the evaluation map \( \text{ev}_t \):

\[
\text{ev}_t : m(X) \to \mathbb{R},
\]

\[
\text{ev}_t(F) = F(t), \quad t \in \mathbb{R}^1.
\]

The probability measure \( \mathcal{DP} \) gives a stochastic process whose sample paths are the graphs of cumulative distribution functions.
It is not immediately clear that random probability measures of a Dirichlet Process are discrete. Blackwell and MacQueen showed that with probability 1, random probability measures from a DP are discrete (i.e., counting measures).

\[
P = \sum_{i=1}^{\infty} P_i \delta_{X_i}, \quad \sum_{i=1}^{\infty} P_i = 1.
\]

**Generalized Pólya Urn** Let \( \{X_n, n \geq 1\} \) denote a sequence of random variables with values in \( X \). It is a *Pólya Sequence with parameter* \( \mu \) (\( \mu \) a measure on \( X \)) if for every \( B \subset X \) we have

\[
P(X_1 \in B) = \frac{\mu(B)}{\mu(X)}
\]

and

\[
P\{X_{n+1} \in B | X_1, \ldots, X_n\} = \frac{\mu_n(B)}{\mu_n(X)}
\]

where \( \mu_n = \mu + \sum_{1}^{n} \delta_{X_i} \). Note that \( \mu_n(X) = \mu(X) + n \).
This is a direct generalization of finite-dimensional Dirichlet distribution with predictive distribution (with $\mu$ in place of $\alpha$).

$$P(X_{n+1} | X_1, \ldots, X_n) = \frac{\mu + \sum_{i=1}^{n} \delta X_i}{\mu(X) + n}.$$ 

Except now $X$ is an arbitrary complete and separable metric space, and the base measure $\mu$ can be a non-discrete (or continuous) probability measure. More concretely,

$$P(X_{n+1} | X_1, \ldots, X_n) = \frac{\mu(X)}{\mu(X) + n} \frac{\mu}{\mu(X)} + \frac{n}{\mu(X) + n} \frac{1}{n} \sum_{i=1}^{n} \delta X_i.$$ 

In particular, to sample $X_{n+1}$ according to the predictive distribution, $X_{n+1}$ has probability $\frac{\mu(X)}{\mu(X) + n}$ as a sample from $\frac{\mu}{\mu(X)}$, and probability $\frac{n}{\mu(X) + n}$ as a sample from $\frac{1}{n} \sum_{i=1}^{n} \delta X_i$. 
Chinese Restaurant Process (CRP)

This gives the Chinese Restaurant Process

\[
\frac{\mu(X)}{\mu(X)+n} \quad - \text{probability of sitting in an empty table}
\]
\[
\frac{n}{\mu(X)+n} \quad - \text{probability of sitting in a non-empty table.}
\]

The main difference between the finite Dirichlet distributions and infinite Dirichlet process is the number of tables in the Dirichlet process is infinite.
In machine learning, an important modeling problem is to understand (assume, study) the unknown process that gives the (potentially infinite) data. In the CRP analogy, the customers are the data and the tables are the clusters.

- number of cluster is infinite,
- there are statistical laws that govern the sizes of clusters.
Theorem
Let \{X_n\} be a Polya sequence with parameter \(\mu\). Then

1. \(m_n = \frac{\mu_n}{\mu_n(X)}\) converges with probability 1 as \(n \to \infty\) to a limiting discrete probability measure \(\mu^*\).

2. \(\mu^*\) is a random probability measure from a Dirichlet Process with base measure \(\mu\),

3. given \(\mu^*\), the variables \(X_1, X_2, \ldots\) are independent with distribution \(\mu^*\).

where \(\mu_n = \mu + \sum_{i=1}^{n} \delta_{X_i}\).
Dirichlet Process

Knowing the random measure $P \sim \mathcal{DP}(\alpha)$ is purely atomic (i.e., discrete)

$$P = \sum_{i=1}^{\infty} P_i \delta_{X_i}.$$

What can we say about the masses $P_i$?

Experiment: Take $\text{Dir}_n(\alpha, \cdots, \alpha)$ and let $n \to \infty$ and $\alpha \to 0$ such that $\bar{\alpha} = n\alpha \to \lambda$

If $p \sim \text{Dir}_n(\alpha, \cdots, \alpha)$, we know that $p_1 \sim \text{Dir}_2(\alpha, (n-1)\alpha)$. Its limit as $\alpha \to 0$ is a degenerated Beta distribution (not interesting). The size-biased sampling gives more information:

Let $\nu$ be a random variable having values $1, 2, 3, \cdots, n$ such that

$$\mathbb{P}(\nu = t|p) = p_t, \quad t \in \{1, \cdots, n\}.$$
Dirichlet Process

It then easy to show that the vector

\[(p_\nu, p_1, \cdots, p_{\nu-1}, p_{\nu+1}, \cdots, p_n) \sim \text{Dir}_n(\alpha + 1, \alpha, \cdots, \alpha).\]

Marginalize to the first component as before, we see that

\[p_\nu \sim \text{Dir}_2(\alpha + 1, (n-1)\alpha) \sim \frac{\Gamma(n\alpha + 1)}{\Gamma(\alpha + 1)\Gamma(n\alpha - \alpha)} p^\alpha (1-p)^{(n-1)\alpha - 1}.\]

Take the limit \(\alpha \to 0, n\alpha \to \lambda\), the above distribution converges to \(\lambda(1-p)^{\lambda-1}\). Furthermore, The remaining \(p_i\), after normalizing by \((1-p_\nu)\) distributed as \(\text{Dir}_{n-1}(\alpha, \cdots, \alpha)\). Repeat the process, we get

\[p_1 = \nu_1, p_2 = (1-\nu_1)\nu_2, p_3 = (1-\nu_1)(1-\nu_2)\nu_3, \cdots,\]

where \(\nu_i \sim \lambda(1-p)^{\lambda-1}\).
This gives an example of Stick Breaking.
Stick-breaking Process

The previously mentioned construction can be visualized as a stick-breaking process.

Note that it gives a way to partition the unity with infinitely many segments. Depending on the parameter $\lambda$, we can expect that most of the segments have small lengths.
Sethuraman’s result show that sampling from a Dirichlet process
\[ P = \sum_{i=1}^{\infty} P_i \delta_{x_i}, \quad P \sim DP(\alpha), \]
can be done in two-independent steps:

1. Sample a collection of points \( \{x_1, x_2, \cdots \} \subset X \) according to the probability distribution \( \alpha/\alpha(X) \).

2. Sample the weights \( P_i \) using stick-breaking process
\[ p_1 = v_1, \quad p_2 = (1 - v_1)v_2, \quad p_3 = (1 - v_1)(1 - v_2)v_3, \cdots, \]
with \( v_i \sim Beta(1, \alpha(X)) \).
Dirichlet Mixtures

Recall the mixture models we mentioned earlier,

\[ x_i \sim F(\cdot | \theta_i), \quad \theta_i \sim P; \quad i = 1, \ldots, n \]
\[ P \sim \mathcal{DP}(\alpha). \]

We have

\[ P = \sum_{i=1}^{\infty} P_i \delta_{\theta_i}. \]

With this model, we can think of each \( \theta_i \) as an \textit{(a priori unknown)} cluster of the data \( x_i \), and data in each cluster is generated according to \( F(\cdot | \theta_i) \). The stick-breaking construction provides a prior on the sizes of these clusters as \( p_i \) can be interpreted as the proportion of the given cluster \( i \).
Pitman-Yor Process

Pitman-Yor process is a two-parameter stick-breaking process that generalizes the Sethuraman’s stick-breaking construction (which requires one parameter, $\alpha$). Specifically,

$$p_1 = v_1, p_2 = (1 - v_1)v_2, p_3 = (1 - v_1)(1 - v_2)v_3, \ldots,$$

with $v_i \sim \text{Beta}(1 - \beta, \alpha + i\beta)$.

- Unlike the case for Dirichlet Process, the distribution for $v_i$ is different for different $i$ (unless $\beta = 0$).
- $\beta = 0$ gives Sethuraman’s construction.
Initially, these stick-breaking priors are important in ecology and population statistics. In these applications, one needs to know (as prior) some kind of distribution of frequencies of different types of observable objects.

**Missing Species Problem**  
Alexander Corbet, a naturalist, spent two years in Malaysia studying butterflies.

\[ n_j = \#\{\text{Species trapped exactly } j \text{ times in two years}\} . \]

His data: \( n_1 = 118, n_2 = 74, n_3 = 44, n_4 = 24 \). . . .

The question he wanted to answer is  
If he spends an additional one year in Malaysia, how many new species could he expect to capture?
Power Law Distributions

Pitman-Yor process provides Power Law distributions for the frequencies of species that is heavier-tailed. A striking example in image applications is given in

"Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes" by Sudderth and Jordan, NIPS08.

![Graphs and images showing Power Law Distributions](image)

Figure 1: Validation of stick-breaking priors for the statistics of human segmentations of the forest (top) and insidescity (bottom) scene categories. We compare observed frequencies (black) to those predicted by Pitman-Yor process (PY, red circles) and Dirichlet process (DP, green squares) models. For each model, we also display 95% confidence intervals (dashed). (a) Example human segmentations, where each segment has a text label such as sky, tree trunk, car, or person walking. The full segmented database is available from LabelMe [14]. (b) Frequency with which different semantic text labels, sorted from most to least frequent on a log-log scale, are associated with segments. (c) Number of segments occupying varying proportions of the image area, on a log-log scale. (d) Counts of segments of size at least 5,000 pixels in 256 × 256 images of natural scenes.