Bayesian Nonparametric Modeling and Inference for Multiple Object Tracking

Bahman Moraffah

Arizona State University
Signal Processing and Adaptive Sensing Laboratory

Probability Does Not Exist.

Bruno de Finetti (1906-1985)

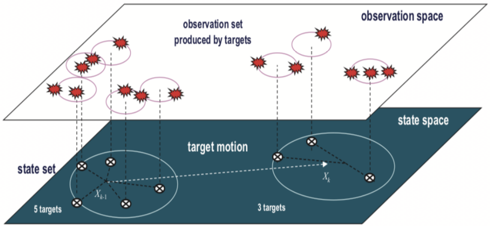
Problem Formulation and Challenges

Problem Statement: Multi-object tracking (MOT) aims to jointly estimate the number of objects and path, location, characteristics of objects from sensor data

- Each object may leave or stay in the field of view with a time dependent probability and transition to next time according to a transition probability kernel
- Each survived object transitions to the next time according to a transition kernel
- New objects can join the scene at random
- Object cardinality is unknown

Challenges:

- Unknown time-varying number of objects
- Robustly associate objects at each time step
- Uncertainty on parameters such as measurement noise or clutter



Multiple Object Tracking Model

Motion Model

- Each existing object $\mathbf{x}_{\ell,k-1}$ may
  - leave the scene w.p. $1 - \mathrm{P}_{\ell,k|k-1}$
  - stay in the scene w.p. $\mathrm{P}_{\ell,k|k-1}$ and transition to the next time using transition probability kernel $\mathbb{Q}_{\boldsymbol{\theta}}(\mathbf{x}_{\ell,k-1}, \cdot)$
- A random number of new objects can appear from random locations in the state space

Measurement Model

- Each object $\mathbf{x}_{\ell,k}$ generates an observation $\mathbf{z}_{l,k}$ with likelihood $p(\mathbf{z}_{l,k}|\mathbf{x}_{\ell,k})$

Tasks:

A. Construct a prior to capture the dependency among the objects $\rightarrow$ Prior

B. Estimate trajectory $\rightarrow$ Inference

Related Work

Object tracking has been studied in various ways:

1. Bayesian methods for a single object tracking
   - Kalman filter, Particle filter, The interacting multiple model for maneuvering, The nearest neighbor for tracking, The probabilistic data association filter
2. Random finite set theory for multi object tracking
   - Multiple hypothesis tracking filter, Joint probability data association filter, Probability hypothesis density filter, Labeled multi-Bernoulli filter, Generalized labeled multi-Bernoulli filter
3. Deep Learning models for multi object tracking
   - Deep affinity network for multiple object tracking, Deep network flow for multi object tracking , Data association for multi object tracking via deep neural networks
4. Bayesian nonparametrics for tracking
   - Evolutionary Clustering, Dynamic Clustering, Bayesian inference for linear dynamic model through Dirichlet processes, Hierarchical Dirichlet process for maneuvering

Contributions

Bayesian Nonparametric Modeling for Multiple Object Tracking:

- Dependent Bayesian nonparametric models as a prior on object states
  - (a) Survival (b) Birth, and (c) Death
  - Adjust the probabilities among new and transitioned objects
  - Dependent models to update object cardinality and posterior distribution
  - Simple inferential methods such as MCMC and VB
  - Captures full dependency with a well known nonparametric marginal distribution
- Achieve higher estimation accuracy and lower computational cost at lower SNR values
- Consistent dependent process
- Achieve optimal frequentist minimax rate of convergence

Dependent Dirichlet Process for Multiple Object Tracking

DDP Prior Construction at time k

- Construct a dependent Dirichlet process prior as follows:

(C1) The ℓth object is assigned to one of the survived and transitioned clusters from time $(k-1)$ which is occupied by at least one of the previous $\ell - 1$ previous objects. The object selects one of these clusters with probability:

Number of objects at jth cluster at time k

Number of objects after transitioningat ith cluster

$$\Pi^{1}_{j,k}(\text{Choosing jth cluster}|\boldsymbol{\theta}^{\ell-1}_{1,k}) \propto [V_k]_j + \sum_{i=1}^{D_{k-1}} \left[V_{k|k-1}\right]_i \; \lambda_{i,k|k-1} \; \delta_i(c_{j,k})$$

Cluster indicator $\lambda_{i,k|k-1} \in \{0,1\}$

Cluster assignment

where the normalizing constant is $(\ell - 1) + \sum_{j}^{\ell-1} \sum_{i=1}^{D_{k-1}} \left[V_{k|k-1}\right]_i \lambda_{i,k|k-1} \delta_i(c_{j,k}) + \alpha$

Cont'd

(C2) The ℓth object is assigned to one of the survived and transitioned clusters from time $(k-1)$. However, this cluster has not yet been assigned to any of the first $\ell - 1$ objects. The object selects such a cluster with probability:

$$\Pi^{2}_{j,k}\left(\textit{Choosing jth cluster that has not been selected yet}\middle|\boldsymbol{\theta}_{1,k}^{\ell-1}\right) \propto \sum_{i=1}^{D_{k-1}} \left[V_{k|k-1}\right]_i \lambda_{i,k|k-1}\delta_i(c_{j,k})$$

- The cluster parameter $\boldsymbol{\theta}^{\star}_{\ell,k-1}$ transitions with transition kernel $\nu(\boldsymbol{\theta}^{\star}_{\ell,k-1},\cdot)$

(C3) The object does not belong to any of the existing clusters; a new cluster parameter is drawn with probability:

$$\Pi_k^3(\textit{Creating new cluster}|\boldsymbol{\theta}_{1,k}^{\ell-1}) \propto \alpha$$

Hyper-parameter

- A new cluster parameter is drawn from the base distribution

(Moraffah & Papandreou 2018)

Is the marginal distribution a Dirichlet Process?

Theorem

Suppose that the space of state parameters is Polish. The dependent Dirichlet process in C1-C3 define a Dirichlet process at each time step given the previous time configurations, i.e.,

$$E_k | E_{k-1} \sim DP\left(\alpha, \sum_{\Theta_k} \Pi^1_{\ell,k} \delta_{\boldsymbol{\theta}_{\ell,k}} + \sum_{\Theta^\star_{k|k-1} \setminus \Theta_k} \Pi^2_{\ell,k} \nu(\boldsymbol{\theta}^\star_{\ell,k-1}, \boldsymbol{\theta}_{\ell,k}) \delta_{\boldsymbol{\theta}_{\ell,k}} + \Pi^3_k H\right)$$

(Moraffah & Papandreou 2018, 2019)

Density Estimation

Theorem

Assume that the space of object state parameters is separable and complete. Given the past configurations, the state distribution is

$$p(\mathbf{x}_{\ell,k}|\mathbf{x}_{1,k}^{\ell-1}, \mathbf{X}_{k|k-1}, \Theta^{\star}_{k|k-1}, \Theta_k) = \begin{cases} \mathbb{Q}_{\underline{\boldsymbol{\theta}}}(\mathbf{x}_{\ell,k-1}, \mathbf{x}_{\ell,k}) f(\mathbf{x}_{\ell,k}|\boldsymbol{\theta}^{\star}_{\ell,k}) & \text{w.p. } C1 \\ \mathbb{Q}_{\underline{\boldsymbol{\theta}}}(\mathbf{x}_{\ell,k-1}, \mathbf{x}_{\ell,k}) \nu(\boldsymbol{\theta}^{\star}_{\ell,k-1}, \boldsymbol{\theta}_{\ell,k}) f(\mathbf{x}_{\ell,k}|\boldsymbol{\theta}^{\star}_{\ell,k}) & \text{w.p. } C2 \\ \int_{\boldsymbol{\theta}} f(\mathbf{x}_{\ell,k}|\boldsymbol{\theta}) dH(\boldsymbol{\theta}) & \text{w.p. } C3 \end{cases}$$

for some density f.

This method is called Dependent Dirichlet Process-Evolutionary Markov Modeling (DDP-EMM)

(Moraffah & Papandreou 2018, 2019)

Learning Model

Upon receiving $\mathcal{Z}_k = \{\mathbf{z}_{l,k},\ l = 1, \ldots, M_k\}$ at time k

- Update our belief and learn using $p(\mathbf{z}_{l,k} | \boldsymbol{\theta}_{\ell,k}, \mathbf{x}_{\ell,k})$ is drawn from to the following hierarchy:

$$
\begin{aligned}
&\boldsymbol{\theta}_{\ell,k} \sim \text{DDP-EMM}(\alpha, H) \\
&\mathbf{x}_{\ell,k} \mid \boldsymbol{\theta}^{\star}_{\ell,k} \sim F(\boldsymbol{\theta}^{\star}_{\ell,k}) \\
&\mathbf{z}_{l,k} | \boldsymbol{\theta}^{\star}_{\ell,k}, \mathbf{x}_{\ell,k} \sim R(\mathbf{z}_{l,k} | \boldsymbol{\theta}^{\star}_{\ell,k}, \mathbf{x}_{\ell,k})
\end{aligned}
$$

  for some distribution R

How to find the posterior distribution? How to do inference?

Bayesian Inference: Gibbs Sampler

- Predictive Distribution: The Bayesian posterior can be solved through the following:

$$P(\mathbf{x}_{\ell,k}|\mathcal{Z}_k) = \int_{\boldsymbol{\theta}} P(\mathbf{x}_{\ell,k}|\mathcal{Z}_k, \boldsymbol{\theta}) dG(\boldsymbol{\theta}|\mathcal{Z}_k)$$

where

$$P(\mathbf{x}_{\ell,k}|\mathcal{Z}_k, \boldsymbol{\theta}) = P(\mathbf{x}_{\ell,k}|\boldsymbol{\theta})$$
$$P(\mathbf{x}_{\ell,k}|\Theta) = \int P(\mathbf{x}_{\ell,k}|\boldsymbol{\theta}_{\ell,k}) d\pi(\boldsymbol{\theta}_{\ell,k}|\Theta)$$

and the posterior distribution $\pi(\boldsymbol{\theta}_{\ell,k}|\Theta)$ is given by

$$\pi(\boldsymbol{\theta}_{\ell,k}|\Theta) = \sum_{\boldsymbol{\theta}\in\Theta_k - \{\boldsymbol{\theta}_{\ell,k}\}} \Pi^1_{j,k}\delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{\ell,k}) + \sum_{\substack{\boldsymbol{\theta}\in\Theta^{\star}_{k|k-1}\backslash\Theta \\ \boldsymbol{\theta}\neq\boldsymbol{\theta}_{\ell,k}}} \Pi^2_{j,k}\nu(\boldsymbol{\theta}^{\star}_{\ell,k-1}, \boldsymbol{\theta}_{\ell,k})\delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{\ell,k}) + \Pi^3_k H(\boldsymbol{\theta}_{\ell,k})$$

- How to compute $G(\boldsymbol{\theta}|\mathcal{Z}_k)$? $\rightarrow$ Direct computation is expensive $\rightarrow$ Gibbs sampling

Cont'd

Theorem

(Gibbs Sampler) In the hierarchical model the conditional posterior distribution is given by

$$\boldsymbol{\theta}_{\ell,k} \mid \boldsymbol{\theta}_{-\ell,k}, \mathcal{Z}_k \sim \sum_{j=1}^{|\mathcal{C}_k|} \zeta_{j,k}\, \delta_{\boldsymbol{\theta}_{j,k}}(\boldsymbol{\theta}_{\ell,k}) + \sum_{\substack{j=1 \\ j\notin \mathcal{C}_k}}^{D_{k|k-1}} \beta_{j,k}\, K_{j,k}(\boldsymbol{\theta}_{\ell,k}) + \gamma_{\ell,k}\, H_\ell(\boldsymbol{\theta}_{\ell,k}),$$

where $\boldsymbol{\theta}_{-\ell,k}$ by convention is the set $\{\boldsymbol{\theta}_{j,k},\, j \neq \ell\}$, where

$$\zeta_{j,k} = \frac{[V_k]_j + \sum_{i=1}^{D_{k|k-1}} \left[V_{k|k-1}\right]_i \lambda_{i,k|k-1}\delta_i(c_{j,k})}{(\ell-1) + \sum_{i=1}^{D_{k|k-1}} \left[V_{k|k-1}\right]_i \lambda_{i,k|k-1} + \alpha} R(\mathbf{z}_{\ell,k}|\mathbf{x}_{j,k}, \boldsymbol{\theta}_{j,k}), \qquad \beta_{j,k} = \frac{\sum_{\substack{i=1 \\ i\notin \mathcal{C}_k}}^{D_{k|k-1}} \left[V_{k|k-1}\right]_j \lambda_{j,k|k-1}}{(\ell-1) + \sum_{i=1}^{D_{k|k-1}} \left[V_{k|k-1}\right]_i \lambda_{i,k|k-1} + \alpha}$$

$$\sum_{j=1}^{|\mathcal{C}_k|} \zeta_{j,k} + \sum_{\substack{j=1 \\ j\notin \mathcal{C}_k}}^{D_{k|k-1}} \beta_{j,k} + \gamma_{\ell,k} = 1$$

$K_{j,k} = R(\mathbf{z}_{\ell,k}|\mathbf{x}_{j,k}, \boldsymbol{\theta}_{j,k})$ and $dH_\ell(\boldsymbol{\theta}) \propto R(\mathbf{z}_{\ell,k}|\mathbf{x}_{j,k}, \boldsymbol{\theta})dH(\boldsymbol{\theta})$, H: base distribution on $\boldsymbol{\theta}$.

Properties of the Model

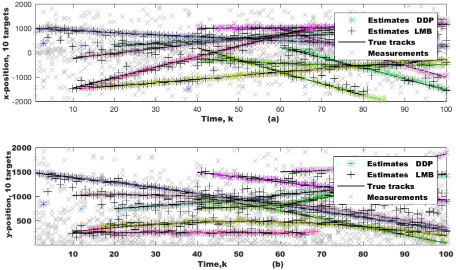
- At each time step, convergence to the posterior distribution $P_\theta(\cdot|\mathcal{Z}_k)$ does not depend on the starting value and almost surely converges in total variation norm
- The posterior is weakly/strongly consistent
- Due to Markov property of the process the EPPF on the partition on N_k and $(N_k - 1)$ objects given the configuration at time $(k-1)$ satisfies

$$p_{N_k-1}([V_k]_1^\star, \ldots, [V_k]_{D_k}^\star) = \sum_{j=1}^{D_k} p_{N_k}([V_k]_1^\star, \ldots, [V_k]_j^\star + 1, \ldots [V_k]_{D_k}^\star) + p_{N_k}([V_k]_1^\star, \ldots, [V_k]_{D_k}^\star, 1)$$

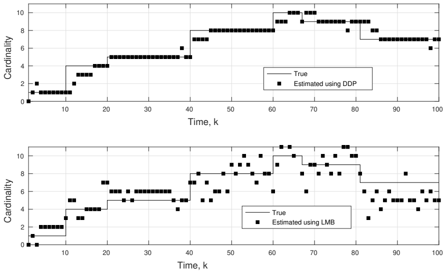
- Assuming the Holder space $\mathcal{H}_\kappa([0,1]^{n_z})$, under some mild condition contraction rate is $n^{-\frac{\kappa}{2\kappa+n_z}}$, which matches the optimal frequentist rate

(Moraffah & Papandreou 2019)

Example 1: Comparison to LMB Tracker

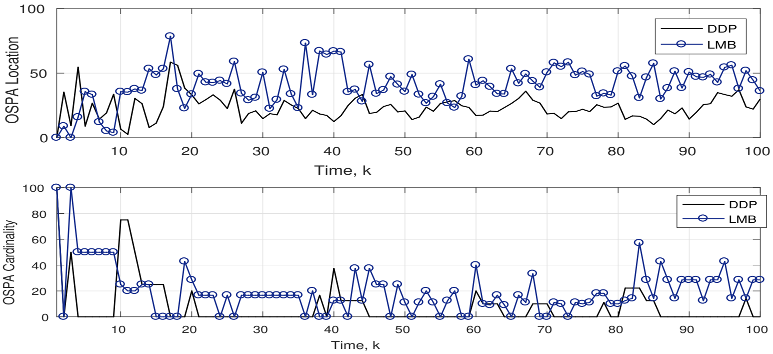


Actual and estimated x (top) and y(bottom) position versus time k using DDP-EMM and LMB methods



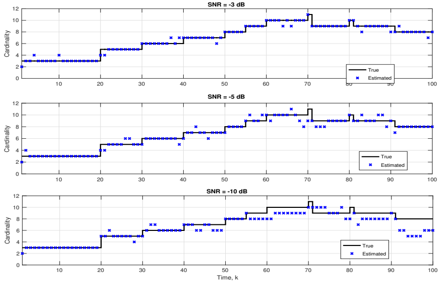
Comparison between cardinality estimation for DDP (top) and LMB (bottom) when tracking 10 objects

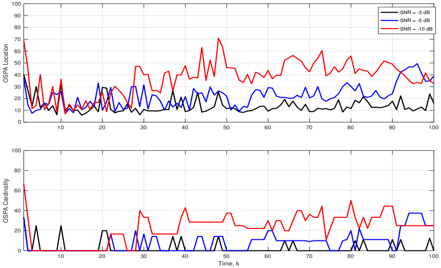
OSPA Comparison to LMB Tracker



OSPA location (top) and cardinality (bottom) of order $p = 1$ and cut-off $c = 100$

Example 2: DDP-EMM under Different SNRs

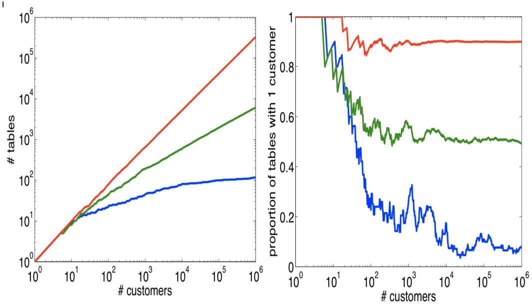




DDP-EMM performance for SNR $= -3$ dB, SNR $= -5$ dB, and SNR $= -10$ dB

Dependent Pitman-Yor Process for Multiple Object Tracking

Why Dependent Pitman-Yor processes?



$\alpha = 10$ and for $d = 0.9$, $d = 0.5$, and $d = 0$

- Object state benefits from a larger number of available clusters to capture full dependency $\rightarrow$ More plausible to have less popular clusters $\rightarrow$ Pitman-Yor process follows the power law
- Rich gets richer
- DPY-STP to model a collection of random distributions that are related but not identical

	DP	PY
Number of unique cluster in N points	$\mathcal{O}(\alpha \log N)$	$\mathcal{O}(\alpha N^b)$

Dependent Pitman-Yor Modeling for MOT: DPY-STP Construction

- Construct a dependent Pitman-Yor process prior as follows:

(C1) The ℓth object belongs to one of the survived and transitioned clusters from time $(k-1)$ and occupied at least by one of the previous $\ell - 1$ objects. The object selects one of these clusters with probability:

$$\Gamma^{1}_{j,k}(\text{Choosing jth cluster}|\boldsymbol{\theta}^{\ell-1}_{1,k}) \propto \sum_{i=1}^{D_{k-1}} \left[\mathbf{V}^{\star}_{k|k-1}\right]_{i} \eta_{i,k|k-1}\delta_i(c_{j,k}) + [\mathbf{V}_k]_j - d$$

$$\text{Normalizing constant}: \sum_{j=1}^{\ell-1}\sum_{i=1}^{D_{k-1}} \left[\mathbf{V}^{\star}_{k|k-1}\right]_{i} \eta_{i,k|k-1}\delta_i(c_{j,k}) + \sum_{j=1}^{\ell-1}[\mathbf{V}_k]_j + \alpha$$

($0 \leq d < 1$ and $\alpha > -d$ are the discount and strength parameters in the Pitman-Yor process, respectively)

Cont'd

(C2) The ℓth object belongs to one of the survived and transitioned clusters from time $(k-1)$ but this cluster has not yet been occupied by any one the first $\ell - 1$objects. The object selects such a cluster with probability:

$$\Gamma^2_{j,k}(j\text{th cluster not been selected yet}|\boldsymbol{\theta}_{1,k},\ldots,\boldsymbol{\theta}_{\ell-1,k}) \propto \sum_{i=1}^{D_{k-1}} \left[\mathbf{V}^\star_{k|k-1}\right]_i \eta_{i,k|k-1}\delta_i(c_{j,k}) - d$$

ℓth survived object parameter at time $(k-1)$ evolves according to a transition kernel:

$$\boldsymbol{\theta}_{\ell,k} \sim \zeta(\boldsymbol{\theta}^\star_{\ell,k-1},\cdot)$$

(C3) The object does not belong to any of the existing clusters, thus a new cluster parameter is drawn from some base distribution H, corresponding to the base distribution in Pitman-Yor process, with probability:

$$\Gamma_k^3(\text{Creating new cluster}|\boldsymbol{\theta}_1(k), \ldots, \boldsymbol{\theta}_{\ell-1}(k)) \propto |D_k|_{\ell-1} d + \alpha$$

$|D_k|_{\ell-1}$: total number of unique clusters at time k created by the first $(\ell - 1)$ objects

Is the marginal distribution a Pitman-Yor?

Theorem

Suppose that the space of state parameters is separable and complete metrizable space. This process defines a Pitman-Yor process at each time step given the previous time configurations, i.e.,

$$\textit{DPY-STP}_k | \textit{DPY-STP}_{k-1} \sim \mathcal{PY}\Big(d, \alpha, \sum_{\Theta_k} \Gamma^1_{j,k} \delta_{\boldsymbol{\theta}_{\ell,k}} + \sum_{\Theta^\star_{k|k-1} \setminus \Theta_k} \Gamma^2_{j,k} \zeta(\boldsymbol{\theta}^\star_{\ell_k - 1}, \boldsymbol{\theta}_{\ell,k}) \delta_{\boldsymbol{\theta}_{\ell,k}} + \Gamma^3_k H\Big)$$

(Moraffah & Papandreou 2019)

Given previous Theorem, we provide an object density estimator.

Theorem

Assume the space of states, $\mathcal{X}$, is separable and complete metrizable topological space. Given Equations in cases C1-C3, distribution over states follows:

$$p(\mathbf{x}_{\ell,k}|\mathbf{x}_{1,k},\ldots,\mathbf{x}_{\ell-1,k},\mathbf{X}_{k|k-1},\Theta^{\star}_{k|k-1},\Theta_k) = \begin{cases} \mathbb{Q}_{\underline{\boldsymbol{\theta}}}(\mathbf{x}_{\ell,k-1},\mathbf{x}_{\ell,k})f(\mathbf{x}_{\ell,k}|\boldsymbol{\theta}^{\star}_{\ell,k}) & \textit{If C1} \\ \mathbb{Q}_{\underline{\boldsymbol{\theta}}}(\mathbf{x}_{\ell,k-1},\mathbf{x}_{\ell,k})\zeta(\boldsymbol{\theta}^{\star}_{\ell,k-1},\boldsymbol{\theta}^{\star}_{\ell,k})f(\mathbf{x}_{\ell,k}|\boldsymbol{\theta}^{\star}_{\ell}(k)) & \textit{If C2} \\ \int_{\boldsymbol{\theta}} f(\mathbf{x}_{\ell,k}|\boldsymbol{\theta})dH(\boldsymbol{\theta}) & \textit{If C3} \end{cases}$$

for some density $f(\cdot|\boldsymbol{\theta})$, distribution H on parameters, and $\mathbf{X}_{k|k-1}$ the set of survived state objects.

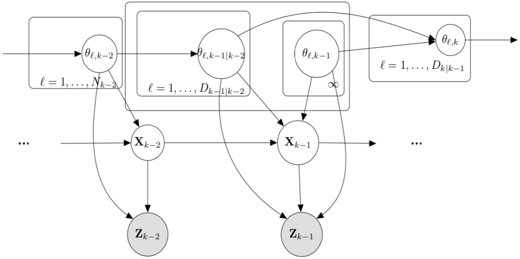
Upon receiving $\mathcal{Z}_k = \{\mathbf{z}_{l,k},\ l = 1, \ldots, M_k\}$ at time k

- Update our belief and learn using $p(\mathbf{z}_{l,k}|\boldsymbol{\theta}_{\ell,k}, \mathbf{x}_{\ell,k})$ is drawn from to the following hierarchy:

$$\mathbf{x}_{\ell,k}|\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{\ell-1,k}, \mathbf{X}_{k|k-1}, \Theta_k \sim \text{DPY-STP}$$
$$\mathbf{z}_{l,k}|\mathbf{x}_{\ell,k}, \boldsymbol{\theta}^{\star}_{\ell,k} \sim R(\mathbf{z}_{l,k}|\mathbf{x}_{\ell,k}, \boldsymbol{\theta}^{\star}_{\ell,k})$$

  for some distribution R

How to find the posterior distribution?
How to do inference?



Graphical model representing this procedure from time $(k-1)$ to k

Inference: Gibbs Sampler

- $\mathcal{C}_k = \{c_{1,k}, \ldots, c_{N_k,k}\}$: Cluster indicator at time k
  - $c_{i,k} = c_{j,k}$ if and only if $\boldsymbol{\theta}_{i,k} = \boldsymbol{\theta}_{j,k}$
  - $c_{i,k} = \ell$ if and only if $\boldsymbol{\theta}_{i,k} = \boldsymbol{\theta}^{\star}_{\ell,k}$
  - $\mathcal{C}_k$ provides a partition on $\{1, \ldots, N_k\}$
- Successive conditional Blackwell-MacQueen distribution:

$$\boldsymbol{\theta}_{\ell,k} | \Theta \sim \sum_{\Theta_k - \{\boldsymbol{\theta}_{\ell,k}\}} \Gamma^1_{j,k} \delta_{\boldsymbol{\theta}}(\theta_{\ell,k}) + \sum_{\substack{\boldsymbol{\theta} \in \Theta^{\star}_{k|k-1} \setminus \Theta \\ \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\ell,k}}} \Gamma^2_{j,k} \nu(\boldsymbol{\theta}^{\star}_{\ell,k-1}, \boldsymbol{\theta}_{\ell,k}) \delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{\ell,k}) + \Gamma^3_k H(\boldsymbol{\theta}_{\ell,k})$$

Cont'd

Theorem

Suppose the base measure is nonatomic, the required conditional distribution to do local inference is derived by marginalizing over the mixing measures:

$$p(c_{i,k} = \ell | \mathcal{C}_k \setminus \{c_{i,k}\}, \mathbf{Z}_k, rest) \propto \begin{cases} \Gamma_{\ell,k}^{1,-i} R(\mathbf{z}_{l,k} | \mathbf{x}_{\ell,k}, \boldsymbol{\theta}_{\ell,k}^{\star}) & \text{for cluster } \ell \text{ that has been selected} \\ \Gamma_{\ell,k}^{2,-i} R(\mathbf{z}_{l,k} | \mathbf{x}_{\ell,k}, \boldsymbol{\theta}_{\ell,k}^{\star}) & \text{for cluster } \ell \text{ that has not yet been selected} \\ \Gamma_{k}^{3,-i} \int R(\mathbf{z}_{l,k} | \mathbf{x}_{\ell,k}, \boldsymbol{\theta}) dH(\boldsymbol{\theta}) & \text{new cluster is created} \end{cases}$$

Cont'd

How to find $\Gamma^{1,-i}_{\ell,k}$, $\Gamma^{2,-i}_{\ell,k}$, and $\Gamma^{3,-i}_{\ell,k}$?

$$\Gamma^{1,-i}_{\ell,k} = \frac{\left[\sum_{j=1}^{D_{k-1}} \left[\mathbf{V}^\star_{k|k-1}\right]_j \eta_{j,k|k-1}\delta_j(c_{\ell,k}) + [\mathbf{V}_k]_\ell\right]_{-i} - d}{\left[\sum_{t=1}^{\ell-1}\sum_{j=1}^{D_{k-1}} \left[\mathbf{V}^\star_{k|k-1}\right]_j \eta_{j,k|k-1}\delta_j(c_{t,k}) + \sum_{t=1}^{\ell-1}[\mathbf{V}_k]_t\right]_{-i} + \alpha}$$

$$\Gamma^{2,-i}_{\ell,k} = \frac{\left[\sum_{j=1}^{D_{k-1}} \left[\mathbf{V}^\star_{k|k-1}\right]_j \eta_{j,k|k-1}\delta_j(c_{\ell,k})\right]_{-i} - d}{\left[\sum_{t=1}^{\ell-1}\sum_{j=1}^{D_{k-1}} \left[\mathbf{V}^\star_{k|k-1}\right]_j \eta_{j,k|k-1}\delta_j(c_{t,k}) + \sum_{t=1}^{\ell-1}[\mathbf{V}_k]_t\right]_{-i} + \alpha}$$

$$\Gamma^{3,-i}_{k} = \frac{|D_k|_{-i} d + \alpha}{\left[\sum_{t=1}^{\ell-1}\sum_{j=1}^{D_{k-1}} \left[\mathbf{V}^\star_{k|k-1}\right]_j \eta_{j,k|k-1}\delta_j(c_{t,k}) + \sum_{t=1}^{\ell-1}[\mathbf{V}_k]_t\right]_{-i} + \alpha}$$

(Moraffah & Papandreou 2019)

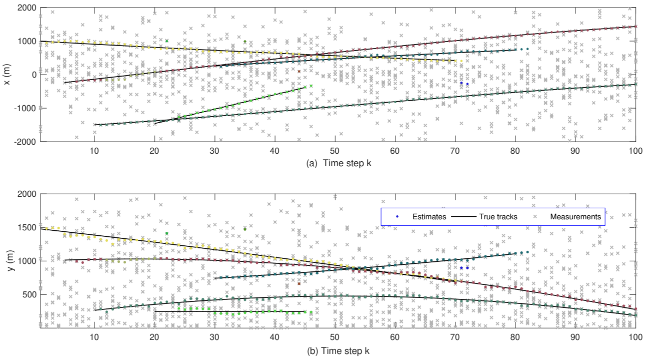
Posterior Consistency of DPY-STP

- Under strict conditions the posterior distribution is consistent
- Most of discrete nonparametric prior (with the exception of the Dirichlet process) are inconsistent
- When discrete nonparametric priors are used in hierarchical mixture models, the generally lead to a consistent density estimator

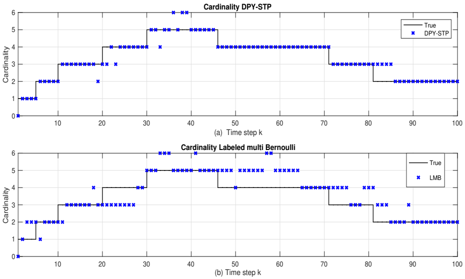
(Lijoi, et.al 2008, 2010, Moraffah & Papandreou 2019)

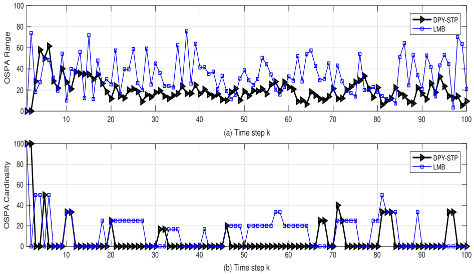
Example 1: Comparison to LMB Tracker

Multi Object Tracking via DDP-STP for Five Moving Objects



True and estimated (a) x-coordinate and (b) y-coordinate as a function of the time step k for five objects

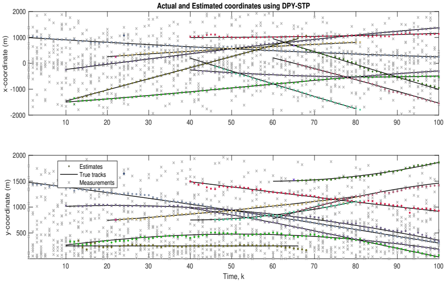




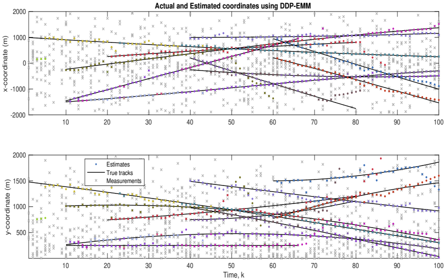
True and learned object cardinality as a function of time step k for 5 objects

OSPA (order $p = 1$ and cut-off $c = 100$ for range (top) Cardinality (bottom) for the DPY-STP and the labeled multi-Bernoulli (LMB) tracker

Example 2: Comparison to DDP-EMM for 10 Moving Objects

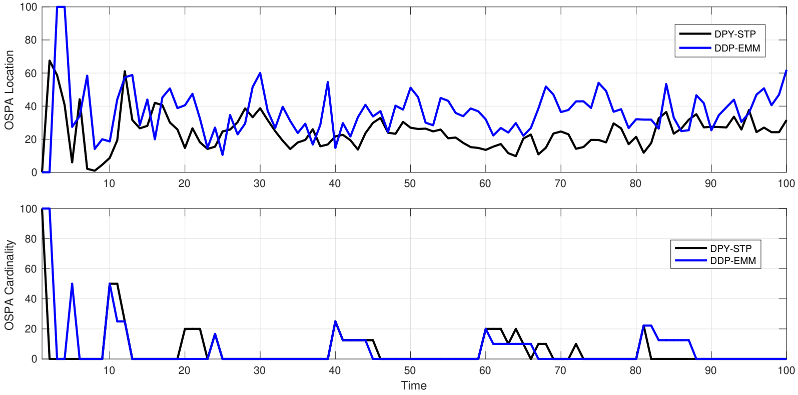


Actual and estimated x and y coordinates through DPY-STP



Actual and estimated x and y coordinates through DDP-EMM

Cont'd: OSPA Comparison



OSPA comparison between DPY-STP and DDP-EMM for cut-off $c = 100$ and order $p = 1$

What Have We Done and What Will Be Done?

What Have We Done?

- Introduce two class of dependent nonparametric models for multi object tracking problem to model object evolution
  - Determine the object identity
  - Estimate the object trajectory as well as object cardinality
- Proposed models that perform well under uncertainties
- Proposed simple MCMC model for Bayesian inference
- Showed consistency of the posterior under introduced prior
- Showed contraction rate of posterior matches the minimax rate
- Compared the performance of proposed methods to one another and also other existing methods

Work to Be Done before Defense

A. **Infinite Random Tree for Multiple Object Tracking**

- Modeling uncertainty over trees; path/branch generated by diffusion process (generate samples using Brownian motion at $t = 0$)
- Branching probability: probability of selecting a branch vs diverging, depends on number of samples previously followed same branch
- Dependent as prior can incorporate time-dependent learned information
  - Place a dependent Diffusion process on parameters
  - Tree leaf/node: object state, branch: cluster of states in a hierarchy
  - Find trajectory of each object by tracing path on tree
  - Predict and update number of objects at each time

Goal: Introduce a dependent nonparametric model over infinite random trees that can robustly estimate the object trajectory as well as object cardinality

(Moraffah& Papandreou 2019)

Work to Be Done before Defense

B. Single Object Tracking with Dependent Measurements

Big picture: Challenges and Solutions:

- Problem Statement: Single object tracking problem when multiple measurements are collected from multiple sensor
- Challenge:
  - Association
  - How to use the dependency among measurements to track accurately
- Solution: Group measurements

How to group measurements so that

(1) Dependency among measurement is held?
(2) Sensor information is preserved?

Solution: Hierarchical Dirichlet process mixture model

Work to Be Done before Defense

C. **Multiple Object Tracking with Dependent Measurements**

- Single object tracking models may not be applied
- Generalize the DDP-EMM multi object tracking model to multiple dependent sensors
- Algorithms should be cable of the following via dependent measurements
  - Dealing with unknown time-dependent object and measurement cardinality
  - Robustly identifying the object identities
- Model Description:
  a. Prior construction over object states
  b. Bayesian Inference
  c. Posterior through an MCMC approach

Future Work

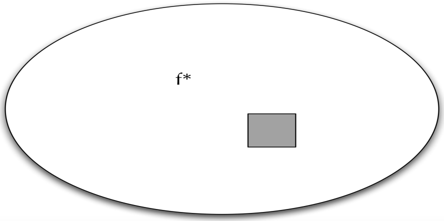
- Use nonparametric models to address other problems in tracking such as high clutter
- Exploit nonparametric models for spawning
- Nonparametric models and causation
- Employ introduced models in other problem such as pattern recognition and image segmentations
- Utilize nonparametric models in health problems such as finding pattern in DNA structure

Appendix

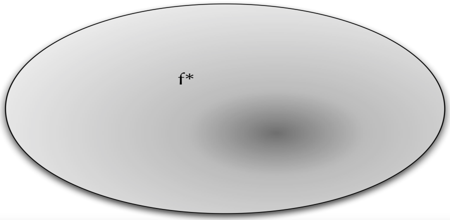
Bayesian Nonparametrics

Bayesian parametric vs. Bayesian nonparametric Models

Parametric model



Nonparametric model



Nonparametric does not mean no parameter; means cannot be described by a finite set of parameters.

No free lunch: Cannot learn from data unless some assumptions are made (less constraints than parametric models).

Motivation for Bayesian Nonparametrics

- A theoretical motivation: de Finetti's Theorem $\longrightarrow$ Nonparametric prior

Definition

A sequence of random variables is **infinitely exchangeable** if the distribution is invariant for any finite sequence, i.e., for any n and permutation σ

$$P(\mathbf{X}_1 \in A_1, \ldots, \mathbf{X}_n \in A_n) = P(\mathbf{X}_{\sigma(1)} \in A_1, \ldots, \mathbf{X}_{\sigma(n)} \in A_n)$$

Theorem

(de Finetti's Theorem) A sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is infinitely exchangeable iff for all n and some distribution G

$$P(\mathbf{X}_1 \in A_1, \ldots, \mathbf{X}_n \in A_n) = \int_{\boldsymbol{\theta}} \prod_{j=1}^{n} P(\mathbf{X}_j \in A_j | \boldsymbol{\theta}) G(d\boldsymbol{\theta})$$

Dirichlet Process

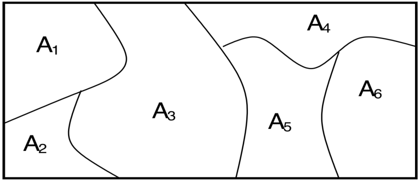
- The most popular nonparametric prior distribution over measures
- Can be viewed as the extension of finite mixture models for density estimation
- Can be derived from different ways:
    1. Ferguson definition of Dirichlet process (Ferguson 1973)
    2. Stick-breaking process (Sethuraman 1994)
    3. Chinese restaurant process (Aldous 1985)
    4. Blackwell-MacQueen process (Pólya urn scheme) (Blackwell 1975)
    5. Dirichlet process and Lévy processes (Gamma Processes)

Ferguson's Definition of Dirichlet Process

Definition

Dirichlet process is a random probability measure over the space Θ satisfying:

- Let $A_1, \ldots, A_n$ be a partition of the Polish space Θ, and $G \sim DP(\alpha, H)$ be a realization of a Dirichlet process with concentration parameter α, and base distribution H, then
  - (i) G is a random measure
  - (ii) G is discrete with probability one
  - (iii) The vector $(G(A_1), \ldots, G(A_n))$ is a probability vector
  - (iv) The marginal distribution of $(G(A_1), \ldots, G(A_n))$ is Dirichlet$(\alpha H(A_1), \ldots, \alpha H(A_n))$



A Constructive Method: Stick-Breaking Construction

- The definition of Dirichlet process is not handy!
- To summarize a method to draw from Dirichlet process, stick breaking process is introduced(Sethuraman 1994)

Definition

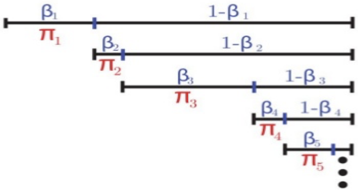
Stick-breaking Construction: To draw a single distribution G from a $DP(\alpha, H)$,

$$\theta_j \overset{i.i.d.}{\sim} H, \qquad \pi_j \sim \text{GEM}(\alpha), \qquad G = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$$

Griffiths-Engen-McCloskey (GEM) distribution:

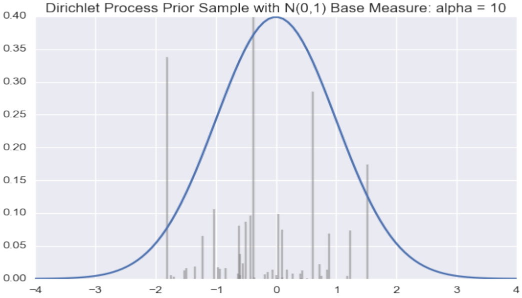
$$\pi'_j \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha)$$

$$\pi_j = \pi'_j \prod_{i=1}^{j-1} (1 - \pi'_i)$$



Cont'd

- A draw from a Dirichlet process is always atomic and $\sum_j \pi_j = 1$
- The weights π_j are decreasing on average but not strictly
- Poisson-Dirichlet process(Kingman 1975) gives an ordering (not computationally tractable)
- A draw from Dirichlet process is discrete with probability one



Posterior Dirichlet Process

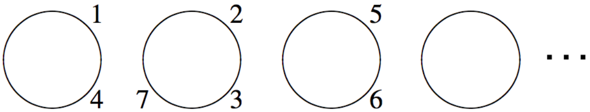
- Let $G \sim DP(\alpha, H)$
- Assume $\boldsymbol{\theta}_j \overset{i.i.d.}{\sim} G$ for $j = 1, \ldots, n$
- Then the posterior distribution given $\boldsymbol{\theta}_j$'s is a Dirichlet process

$$G|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim DP(\alpha + n, \frac{\sum_{j=1}^{n} \delta_{\boldsymbol{\theta}_j} + \alpha H}{\alpha + n})$$

Dirichlet Process as a Clustering Method

- Consider set $[n] = \{1, 2, \dots, n\}$
- Let $G \sim DP(\alpha, H)$
- Assume $\boldsymbol{\theta}_j \overset{i.i.d.}{\sim} G$ for $j = 1, 2, \dots$ (may be repeated)
- Assume $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ takes K distinct values $(K < n) \implies \boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$
- K distinct values defines a partition on $[n]$ such that $j \in \mathcal{C}_k$ iff $\boldsymbol{\theta}_j = \boldsymbol{\theta}_k^\star$
- The induced distribution over all partitions is called Chinese restaurant process(CRP) (Aldous 1985. Pitman 2006).

Chinese Restaurant Process



- CRP(α) is a distribution over partitions, i.e., $\rho \sim \text{CRP}(\alpha)$
- Each customer comes into the restaurant and picks a table at random:

$$\mathbb{P}(\text{Choose table } \mathcal{C}) = \frac{n_{\mathcal{C}}}{\alpha + \sum_{\rho} n_{\mathcal{C}}} \qquad \mathbb{P}(\text{Choose a new table}) = \frac{\alpha}{\alpha + \sum_{\rho} n_{\mathcal{C}}}$$

- Preferential attachment: Rich gets richer
- CRP is exchangeable (not de Finetti exchangeable) and the induced distribution over partitions(no labeling) is called exchangeable partition probability function (EPPF) :

$$\mathbb{P}(n_1, \ldots, n_K | \alpha) = \frac{\alpha^K}{\alpha^{[n]}} \prod_j (n_j - 1)!$$

$$\alpha^{[n]} = \alpha(1+\alpha)\ldots(\alpha + n - 1)$$

How to Relate a Exchangeable Partition to a Exchangeable Sequence

- CRP is an exchangeable random partition not an exchangeable sequence.
- Construct a random partition as follows:
  - For each $\mathcal{C} \in \rho$ define $\boldsymbol{\theta}^\star_{\mathcal{C}} \sim H$
  - For each $j \in [n]$ define $\boldsymbol{\theta}_j = \boldsymbol{\theta}^\star_{\mathcal{C}}$

  where $\mathcal{C} \in \rho$ and $j \in \mathcal{C}$. $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ are de Finetti exchangeable
- What is the underlying distribution G that makes them i.i.d.?
  Answer: Dirichlet Process!

Properties of Dirichlet Process

- DP is discrete with probability one
- DP has atomic distribution $G = \sum_{j=1}^{\infty} \pi_j \boldsymbol{\theta}_j^\star$
- A random sequence can be constructed in the following way

$$\begin{aligned} \rho &\sim \text{CRP}(\alpha) \\ \boldsymbol{\theta}_{\mathcal{C}}^\star &\sim H \quad \text{for each } \mathcal{C} \in \rho \\ \boldsymbol{\theta}_j^\star &= \boldsymbol{\theta}_{\mathcal{C}}^\star \quad \text{for each } j \in [n],\ \mathcal{C} \in \rho, j \in \mathcal{C} \end{aligned}$$

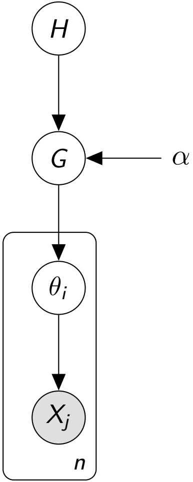
- $\mathbb{E}[G(A)] = H(A)$
- $\mathbb{V}ar[G(A)] = \dfrac{H(A)(1 - H(A))}{\alpha + 1}$
- $\mathbb{E}[K|\alpha, n] = \alpha \log n$

Dirichlet Process Mixture Model

- Dirichlet process is not an appropriate prior for density estimation
- Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim F$,

$$G|\alpha, H \sim \mathrm{DP}(\alpha, H)$$
$$\boldsymbol{\theta}_j|G \overset{i.i.d.}{\sim} G$$
$$\mathbf{X}_j|\boldsymbol{\theta}_j \sim f(\cdot|\boldsymbol{\theta}_j)$$

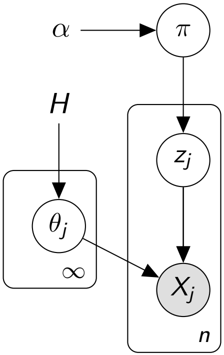
  for some probability density function f.
- Use MCMC methods to find posterior
- The beauty of this model is that due to the discreteness of G, a clustering method is induced. In other words, we have implicitly created a prior on K, the number of distinct $\boldsymbol{\theta}_j$.



- Marginalize G out:
- Let $\mathbf{X}_1, \dots, \mathbf{X}_n \sim F$,

$$\pi|\alpha \sim \mathrm{GEM}(\alpha)$$
$$\boldsymbol{\theta}_j|H \overset{i.i.d.}{\sim} H$$
$$z_j|\pi \sim \mathrm{Cat}(\pi)$$
$$\mathbf{X}_j|\Theta, z_j \sim f(\cdot|\boldsymbol{\theta}_{z_j})$$

  for some probability density function f.
- Use MCMC methods to find posterior



Hierarchical Dirichlet Process

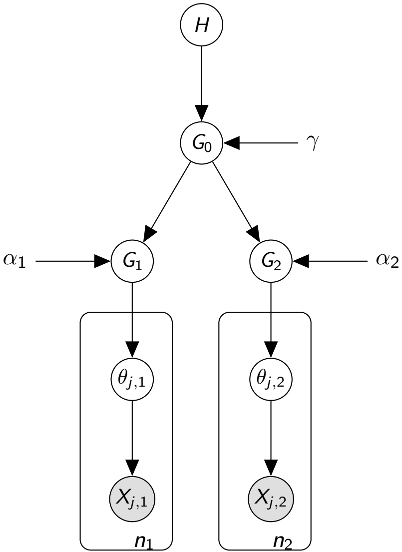
- Hierarchical modeling models shared statistical dependence
- Used for topic modeling
- A hierarchy of Dirichlet processes

$$G_0 \sim DP(\gamma, H)$$
$$G_m | G_0 \sim DP(\alpha_m, G_0)$$
$$\boldsymbol{\theta}_{j,m} | G_m \sim G_m$$

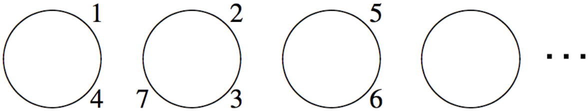
- A hierarchical Dirichlet process mixture can be obtained as follows

$$\mathbf{X}_{j,m} | \boldsymbol{\theta}_{j,m} \sim f(\cdot | \boldsymbol{\theta}_{j,m})$$

- CRP $\longrightarrow$ Chinese restaurant franchise (CRF) (Teh et.al. 2006)



Pitman-Yor Process



- Two parameter CRP: $\text{CRP}([n], d, \alpha)$ with concentration parameter α and discount parameter d over all partitions($\alpha > -d, 0 \leq d < 1$)

$$\mathbb{P}(\text{Choose table } \mathcal{C}) = \frac{n_{\mathcal{C}} - d}{\alpha + \sum_{\rho} n_{\mathcal{C}}} \qquad \mathbb{P}(\text{Choose a new table}) = \frac{\alpha + d|\rho|}{\alpha + \sum_{\rho} n_{\mathcal{C}}}$$

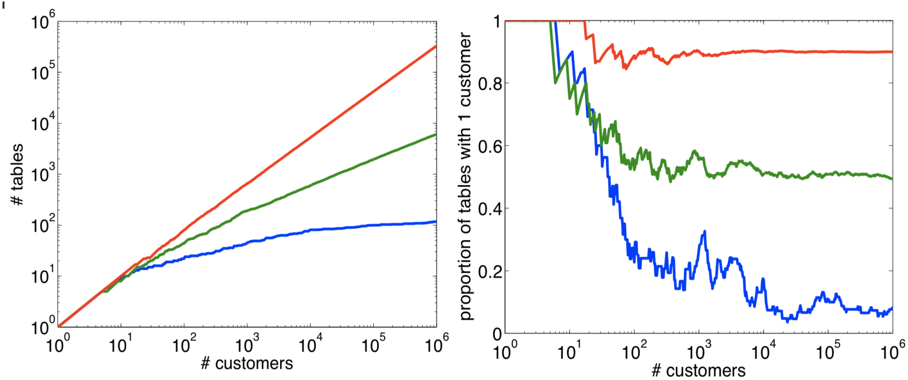
- Two parameter CRP is exchangeable therefore there is an underlying de Finetti distribution such that the data are independent. The de Finetti measure is Pitman-Yor process (Pitman & Yor 1997, Perman et.al. 1992)
- Pitman-Yor process is a generalization of Dirichlet process

Properties of Pitman-Yor Processes

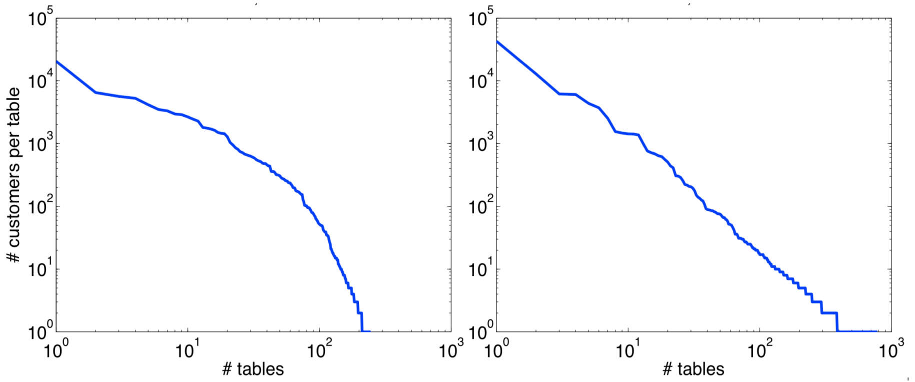
- Bigger d tends to have more tables with few customers
- Rich gets richer: with more occupied tables, chance of even more tables becomes higher
- Tables with small occupancy numbers tend to have lower chance of getting new customers
- $\mathbb{E}[K|\alpha, n] = \alpha n^b \implies$ Power law property of Pitman-Yor process (Goldwater 2005 &Teh 2006)

	DP	PY
Number of unique cluster in n observation	$\mathcal{O}(\alpha \log n)$	$\mathcal{O}(\alpha n^b)$

Dirichlet Process vs. Pitman-Yor Process



$\alpha = 10$ and for $d = 0.9$, $d = 0.5$, and $d = 0$



Pitman-Yor process follows Zipf’s law![1]

(Wood et.al, 2011)

[1]Picture is taken from Teh 2013