Introduction to Dirichlet Processes and Their Use

Wray Buntine

Monash University
http://topicmodels.org

Thanks to: **Lan Du** (Monash) and **Kar Wai Lim** (NUS) for some content,
Lancelot James of HKUST for teaching me some of the material

2017-09-12

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs

4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion

What are Non-parametric Methods?

By classical statistics: modelling without parameters,
e.g., nearest neighbour methods

What are Non-parametric Methods?

By classical statistics: modelling without parameters,
e.g., nearest neighbour methods

By Bayesian statistics: modelling with an infinite number of parameters.

What are Non-parametric Methods?

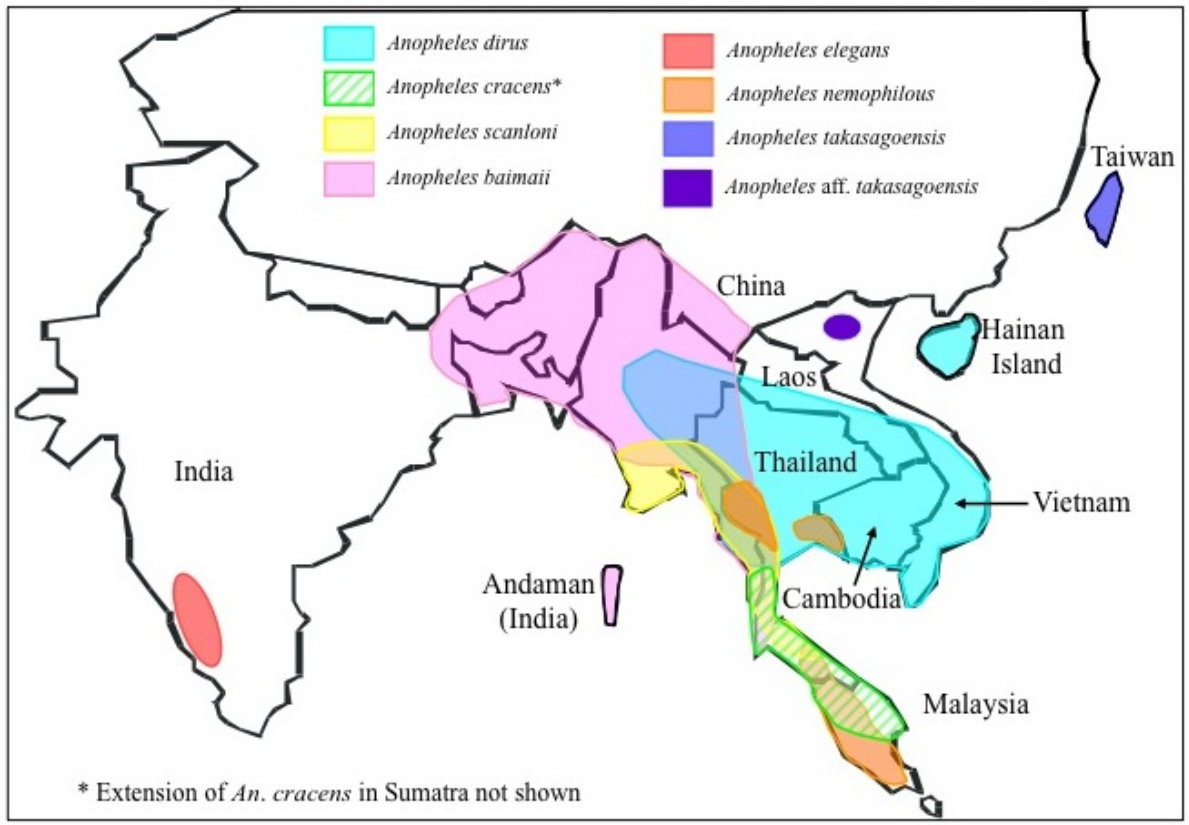
By classical statistics: modelling without parameters,
e.g., nearest neighbour methods

By Bayesian statistics: modelling with an infinite number of parameters.

More accurately: modelling:

- with a finite but variable number of parameters;
- more parameters are "unfurled" as data requires it.

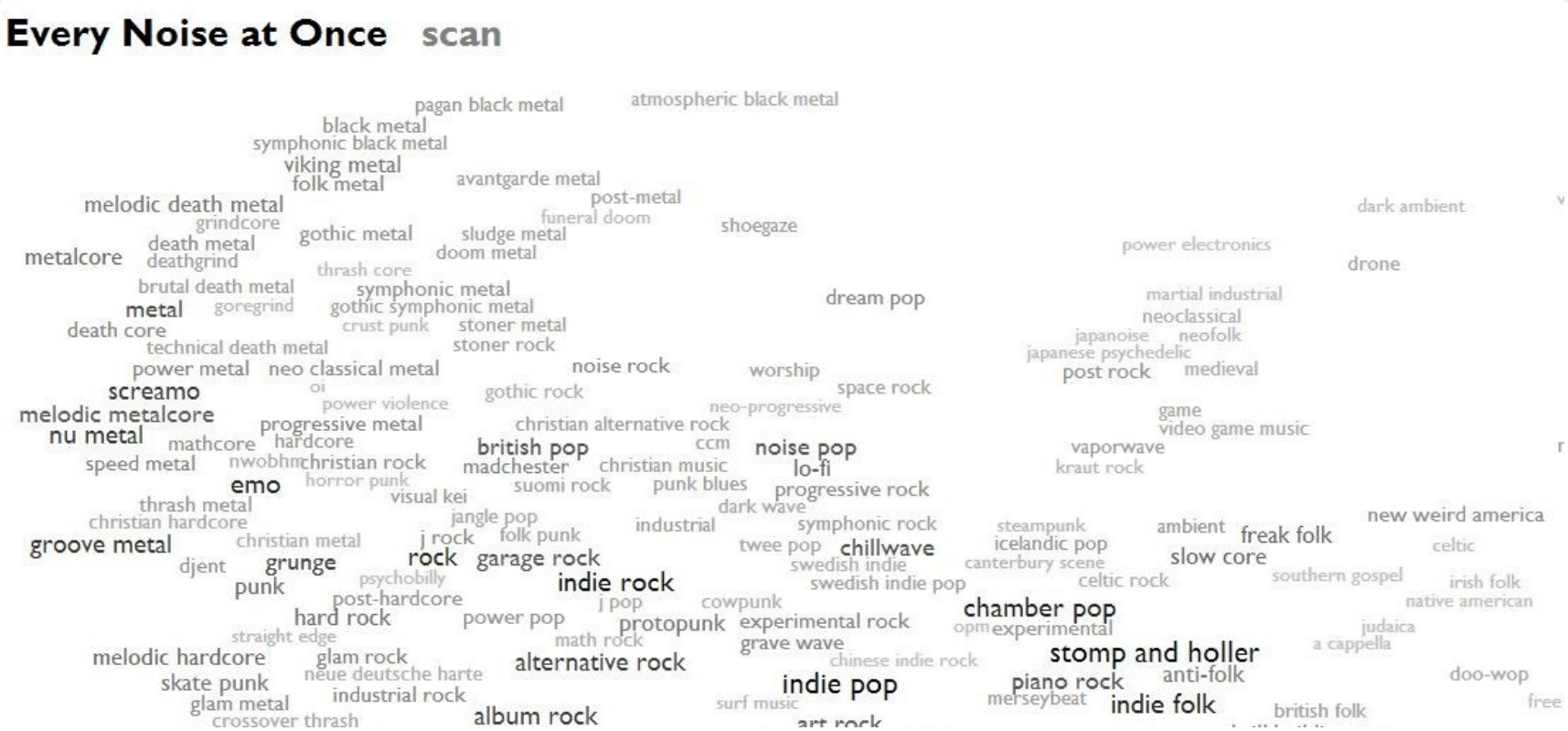
How Many Species of Mosquitoes are There?



e.g. Given some measurement points about mosquitoes in Asia, how many species are there?

K=4? K=5? K=6 K=8?

Which Music Genre's do You Listen to?



(see http://everynoise.com)

Music genre's are constantly developing.

Which ones do you listen to?

What is the chance that a new genre is seen?

What is an Unknown Dimension?

- Some dimensions are fixed but unknown:
  - this uses parametric statistics
  - this is not Bayesian non-parametrics
- Some dimensions keep on growing as we get more data:
  - this is Bayesian non-parametrics

Probability and Parameter Vectors

Problems in modern natural language processing and intelligent systems often have probability vectors for:

- the next word given $(n-1)$ previous,
- an author/conference/corporation to be linked to/from a webpage/patent/citation,
- part-of-speech of a word in context,

Probability and Parameter Vectors

Problems in modern natural language processing and intelligent systems often have probability vectors for:

- the next word given $(n-1)$ previous,
- an author/conference/corporation to be linked to/from a webpage/patent/citation,
- part-of-speech of a word in context,

We need to work with distributions over probability vectors to model these sorts of phenomena well.

Probability and Parameter Vectors

Problems in modern natural language processing and intelligent systems often have probability vectors for:

- the next word given $(n-1)$ previous,
- an author/conference/corporation to be linked to/from a webpage/patent/citation,
- part-of-speech of a word in context,

We need to work with distributions over probability vectors to model these sorts of phenomena well.

- the layers of a deep neural network
- user preferences

Probability and Parameter Vectors

Problems in modern natural language processing and intelligent systems often have probability vectors for:

- the next word given $(n-1)$ previous,
- an author/conference/corporation to be linked to/from a webpage/patent/citation,
- part-of-speech of a word in context,

We need to work with distributions over probability vectors to model these sorts of phenomena well.

- the layers of a deep neural network
- user preferences

Likewise for distributions over parameter vectors.

Example Priors on Probability Vectors

- The advantage of working with probability vectors as a first class object to place distributions over is that we can use them hierarchically.
- This allows us to have much better quality priors.

N-grams: An Example Hierarchy

To model a sequence of words $p(w_1, w_2, ..., w_N)$ we can use:

Unigram (1-gram) model: $p(w_n|\vec{\theta}_1)$

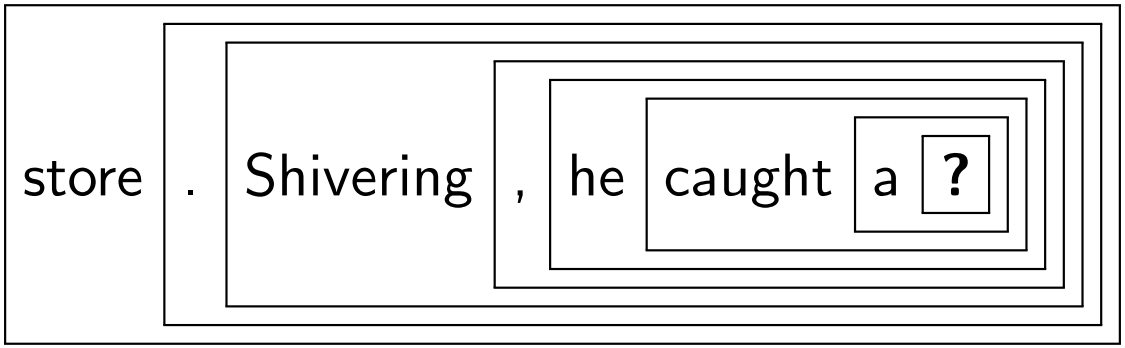
Bigram (2-gram) model: $p(w_n|w_{n-1}, \vec{\theta}_2)$

Trigram (3-gram) model: $p(w_n|w_{n-1}, w_{n-2}, \vec{\theta}_3)$

Or we can interpolate them somehow:

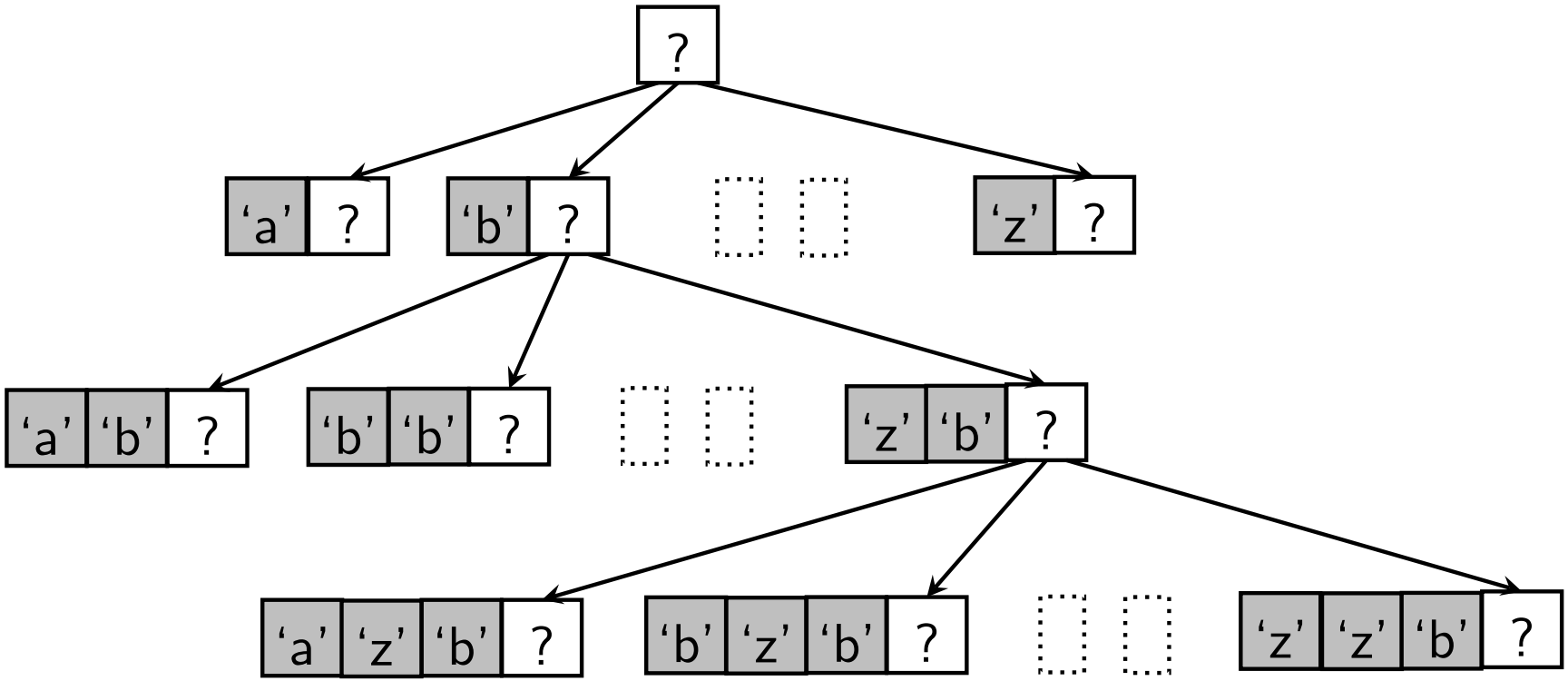
$$\hat{p}(w_n|w_{n-1}, w_{n-2}, \vec{\theta}_3) + \lambda_2 \hat{p}(w_n|w_{n-1}, \vec{\theta}_2)$$

Bayesian Idea: Similar Context Means Similar Word



- Words in [a [?]] should be like words in [?]
  - though no plural nouns
- Words in [caught [a [?]]] should be like words in [a [?]]
  - though a suitable object for "caught"
- Words in [he [caught [a [?]]]] be *very like* words in [caught [a [?]]]
  - "he" shouldn't change things much

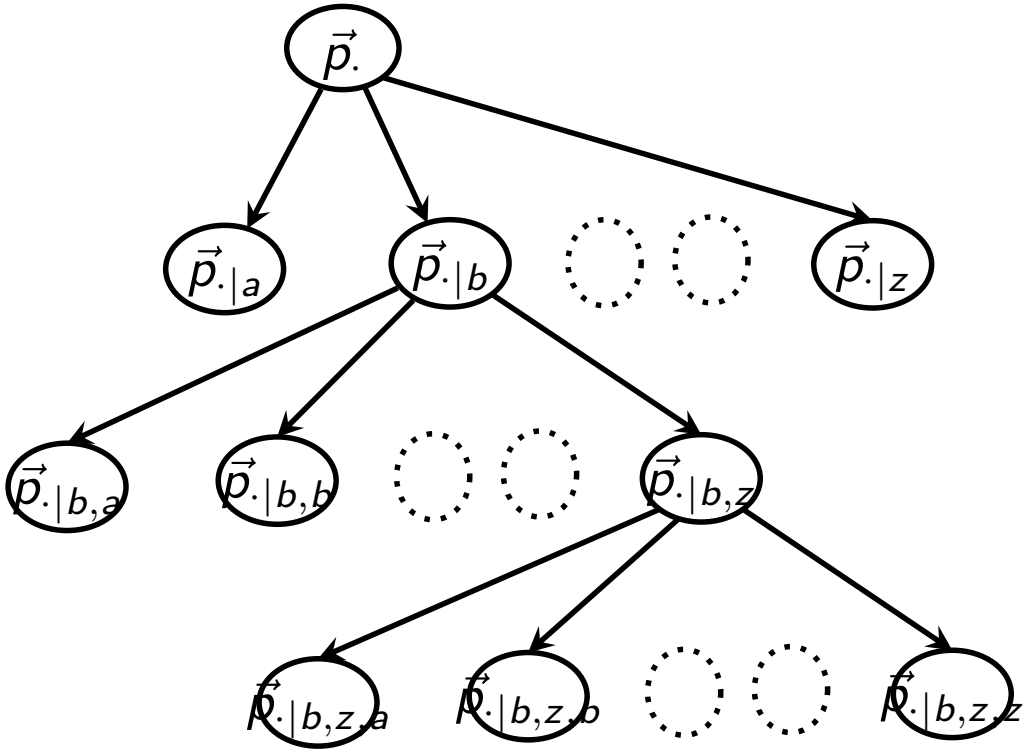
Bayesian N-grams



Build all three together, making the 1-gram a prior for the 2-grams, and the 2-grams a prior for the 3-grams, *etc.*

For this, we need to say each probability vector in the hierarchy is a variant of its parent.

Bayesian N-grams, cont.

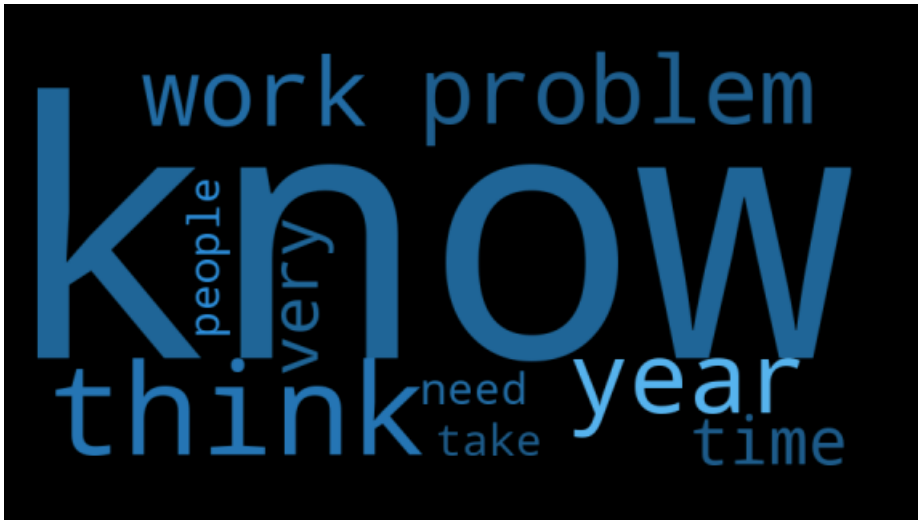


$$\begin{aligned}
\mathcal{S} &= \text{symbol set, fixed or possibly countably infinite} \\
\vec{p}_{\cdot} &\sim \text{prior on prob. vectors (initial vocabulary)} \\
\vec{p}_{\cdot|x_1} &\sim \text{dist. on prob. vectors with mean } \vec{p}_{\cdot} \qquad \forall_{x_1 \in \mathcal{S}} \\
\vec{p}_{\cdot|x_1,x_2} &\sim \text{dist. on prob. vectors with mean } \vec{p}_{\cdot|x_1} \qquad \forall_{x_1,x_2 \in \mathcal{S}}
\end{aligned}$$

Other Hierarchical Priors: Sentiment

If you wanted good sentiment words about cameras or mobile phones, it might be better to use a prior from general sentiment (shown left) rather than from general word frequency.

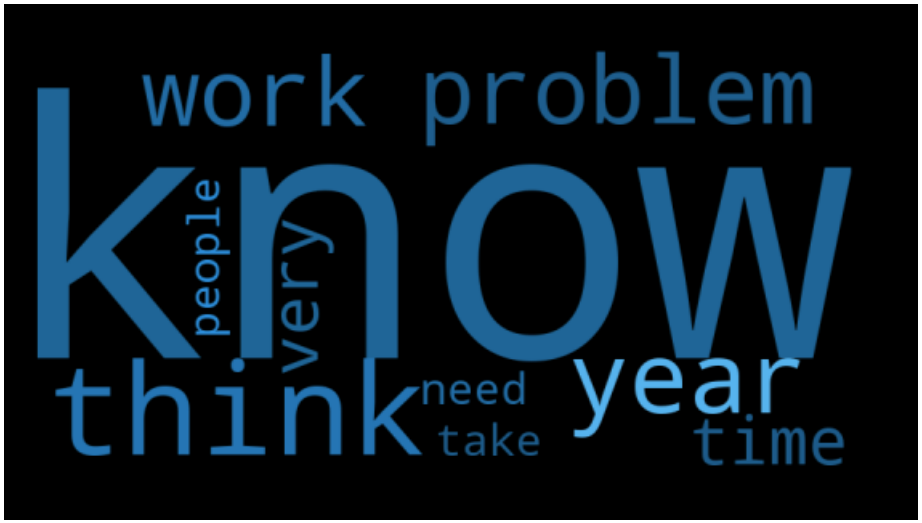
Other Hierarchical Priors: Background Words



Background words in the domain of "news about obesity."

- Have a "background words" prior for topics in a topic model:
  - should model typical (and non-topical) words in any document".

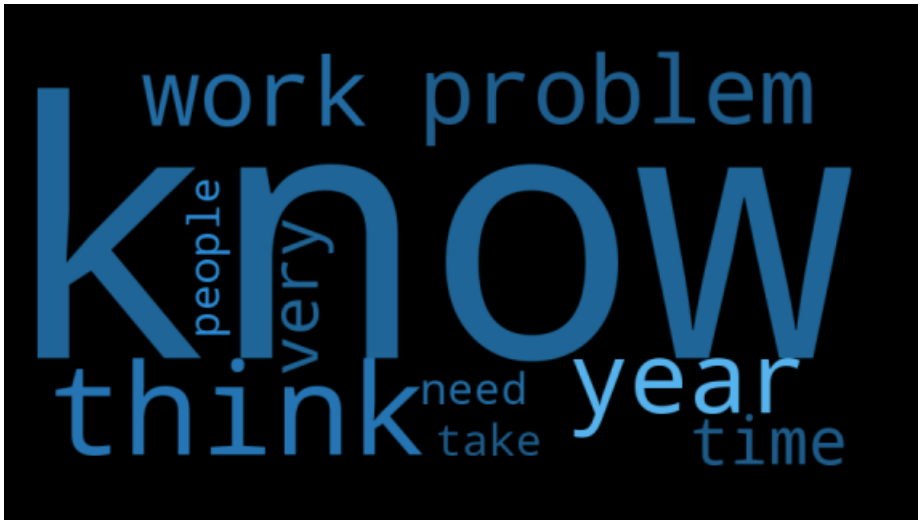
Other Hierarchical Priors: Background Words



Background words in the domain of "news about obesity."

- Have a "background words" prior for topics in a topic model:
  - should model typical (and non-topical) words in any document".
- A topic word probability vector then has this as a "mean."
  - To make it work, we use Zipfian priors more suited to word probability vectors.

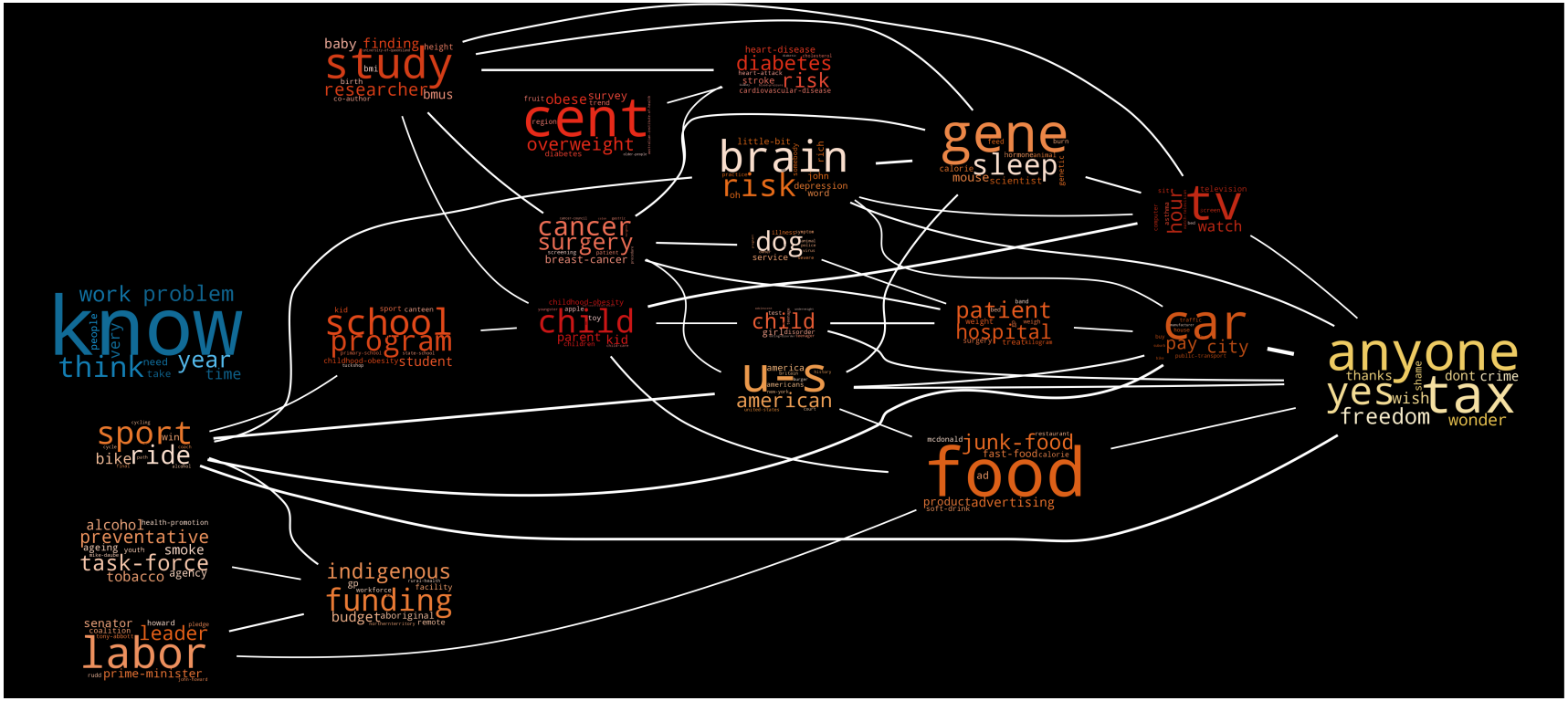
Other Hierarchical Priors: Background Words



Background words in the domain of "news about obesity."

- Have a "background words" prior for topics in a topic model:
  - should model typical (and non-topical) words in any document".
- A topic word probability vector then has this as a "mean."
  - To make it work, we use Zipfian priors more suited to word probability vectors.
- The topic is characterised by departure from the mean.
  - i.e., what words does it emphasise over those in the mean.

Other Hierarchical Priors: Background Words



Topics for "news about obesity" showing the background topic.

The background topic (in blue) is a prior for all other topics!

Bayesian Inference – General Methodology

ASIDE: Good Wikipedia content is marked so$^{\mathcal{W}}$.

Bayesian Inference – General Methodology

Bayesian inference$^{\mathcal{W}}$ is particularly suited for intelligent systems in the context of the previous requirments:

- Bayesian model combination$^{\mathcal{W}}$ and Bayes factors$^{\mathcal{W}}$ for model selection$^{\mathcal{W}}$ can be used;
- marginal likelihood$^{\mathcal{W}}$, a.k.a. the evidence for efficient estimation;
- collapsed Gibbs samplers$^{\mathcal{W}}$, a.k.a. Rao-Blackwellised samplers, for Monte-Carlo Markov chain$^{\mathcal{W}}$ (MCMC) estimation;
- also blocked Gibbs samplers$^{\mathcal{W}}$.

Bayesian Inference – General Methodology

Bayesian inference$^{\mathcal{W}}$ is particularly suited for intelligent systems in the context of the previous requirments:

- Bayesian model combination$^{\mathcal{W}}$ and Bayes factors$^{\mathcal{W}}$ for model selection$^{\mathcal{W}}$ can be used;
- marginal likelihood$^{\mathcal{W}}$, a.k.a. the evidence for efficient estimation;
- collapsed Gibbs samplers$^{\mathcal{W}}$, a.k.a. Rao-Blackwellised samplers, for Monte-Carlo Markov chain$^{\mathcal{W}}$ (MCMC) estimation;
- also blocked Gibbs samplers$^{\mathcal{W}}$.

Wikipedia coverage of Bayesian non-parametrics is patchy.

I'm not bad, I was just drawn that way

Jessica Rabbit in "Who Shot Roger Rabbit?"

I'm not complex, I was just written that way

For any topological space $\mathcal{T}$, $\mathcal{B}(\mathcal{T})$ will denote the Borel σ-field of subsets of $\mathcal{T}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space and $\mathbb{X}$ be complete, separable and endowed with a metric $d_{\mathbb{X}}$. Define on $(\Omega, \mathcal{F}, \mathbb{P})$ a Poisson random measure $\tilde{N}$ on $\mathbb{S} = \mathbb{R}^+ \times \mathbb{X}$ with intensity measure ν. This means that

(i) for any C in $\mathcal{B}(\mathbb{S})$ such that $\nu(C) = \mathbb{E}[\tilde{N}(C)] < \infty$, the probability distribution of the random variable $\tilde{N}(C)$ is Poisson$(\nu(C))$;

(ii) for any finite collection of pairwise disjoint sets, $A_1, \ldots, A_k$, in $\mathcal{B}(\mathbb{S})$, the random variables $\tilde{N}(A_1), \ldots, \tilde{N}(A_k)$ are mutually independent.

Moreover, the measure ν must satisfy the following conditions,

$$\int_{(0,1)} s\nu(\mathrm{d}s, \mathbb{X}) < \infty, \quad \nu([1,\infty) \times \mathbb{X}) < \infty.$$

We refer to Daley & Vere-Jones (1988) for an exhaustive account on Poisson random measures.

From James, Lijoi & Prünster, 2009,

"Posterior analysis for normalized random measures with independent increm

Disclaimer: excellent paper, the complexity is a necessary part of the precision required for mathematical analysis ... but is not so important in machine learning!

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms
- Bayesian inference gives guidance for engineering machine learning.
  - regularisation
  - using side information

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms
- Bayesian inference gives guidance for engineering machine learning.
  - regularisation
  - using side information
- Non-parametrics lets us deal with variable length parameters vectors and other structures.
  - sometimes is little more complex than parametric methods

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms
- Bayesian inference gives guidance for engineering machine learning.
  - regularisation
  - using side information
- Non-parametrics lets us deal with variable length parameters vectors and other structures.
  - sometimes is little more complex than parametric methods
- Apparent complexity of non-parametrics is largely an artifact of the theoretical literature.
  - oftentimes is little more complex than parametric methods

Applying Bayesian Computation

- Approximate Bayesian inference can be made practical.
  - variational and online algorithms
  - collapsed and newer MCMC algorithms
- Bayesian inference gives guidance for engineering machine learning.
  - regularisation
  - using side information
- Non-parametrics lets us deal with variable length parameters vectors and other structures.
  - sometimes is little more complex than parametric methods
- Apparent complexity of non-parametrics is largely an artifact of the theoretical literature.
  - oftentimes is little more complex than parametric methods
- Deep neural networks is a computational methodology, not a statistical one.
  - it is therefore complementary to Bayesian methods

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs
- Introduction to DPs and PYPs
- Behaviour of the DP and PYP

3 Definitions of DPs and PYPs

4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion

What is it All About?

- All our processes are mathematical functions on sets with the form

$$\mu(A) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(A)$$

i.e. sum up those w_k whose θ_k are in A

What is it All About?

- All our processes are mathematical functions on sets with the form

$$\mu(A) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(A)$$

i.e. sum up those w_k whose θ_k are in A

- Understand these, and how to generate the $\vec{p}$, and its behaviour, and you have it mastered.

Introduction to Dirichlet Processes

- Early 2000's Prof. Michael Jordan introduced the Dirichlet Process[W] to the Machine Learning community as a non-parametric method.
- Good tutorials and reviews can be found at:
  - *"Tutorial on Dirichlet Processes and Hierarchical Dirichlet Processes"* by Yeh Whye Teh, 2007
  - *"Completely Random Measures, Hierarchies and Nesting"* by Michael Jordan, 2010
- The now classic paper is *"Hierarchical Dirichlet Processes,"* Teh, Jordan, Beal & Blei, JASA, 2006.

Measures

A measure corresponds to the notion of the size or area.

Definition of Additivity for Functions

A function $m(\cdot)$ on domain $\mathcal{X}$ is additive if whenever $A, B \subset \mathcal{X}$ and $A \cap B = \emptyset$ then $m(A \cup B) = m(A) + m(B)$.

Definition (roughly) of a Measure

A measure$^{\mathcal{W}}$ on domain $\mathcal{X}$ is a non-negative additive function $m(\cdot)$ on (certain well behaved) subsets of X so that $m(\emptyset) = 0$.

Measures

A measure corresponds to the notion of the size or area.

Definition of Additivity for Functions

A function $m(\cdot)$ on domain $\mathcal{X}$ is additive if whenever $A, B \subset \mathcal{X}$ and $A \cap B = \emptyset$ then $m(A \cup B) = m(A) + m(B)$.

Definition (roughly) of a Measure

A measure$^{\mathcal{W}}$ on domain $\mathcal{X}$ is a non-negative additive function $m(\cdot)$ on (certain well behaved) subsets of X so that $m(\emptyset) = 0$.

- If $m(\mathcal{X}) = 1$ then the measure is a probability measure$^{\mathcal{W}}$
  - generalising a PDF (density) and PMF (mass).
- Integrable non-negative functions on continuous spaces are measures.
- In formal theory, the space $\mathcal{X}$ is made to be closed under finite union and complement, *i.e.* Borel σ-field.

Dirichlet Process Introduction

- A Dirichlet Process[W] (DP) is a distribution over probability measures (densities, masses, distributions).
- A DP denoted DP $(\alpha, H(\cdot))$ has two parameters (inputs):

  concentration: which is like an *inverse variance*, denoted α

  mean: called the base distribution, denoted $H(\cdot)$
- Or alternatively one parameter (input)

  measure: denoted $m(\cdot)$

  so $m(A) = \alpha H(A)$ and $\alpha = m(\mathcal{X})$.

Dirichlet Process Introduction

- A Dirichlet Process[W] (DP) is a distribution over probability measures (densities, masses, distributions).
- A DP denoted DP $(\alpha, H(\cdot))$ has two parameters (inputs):

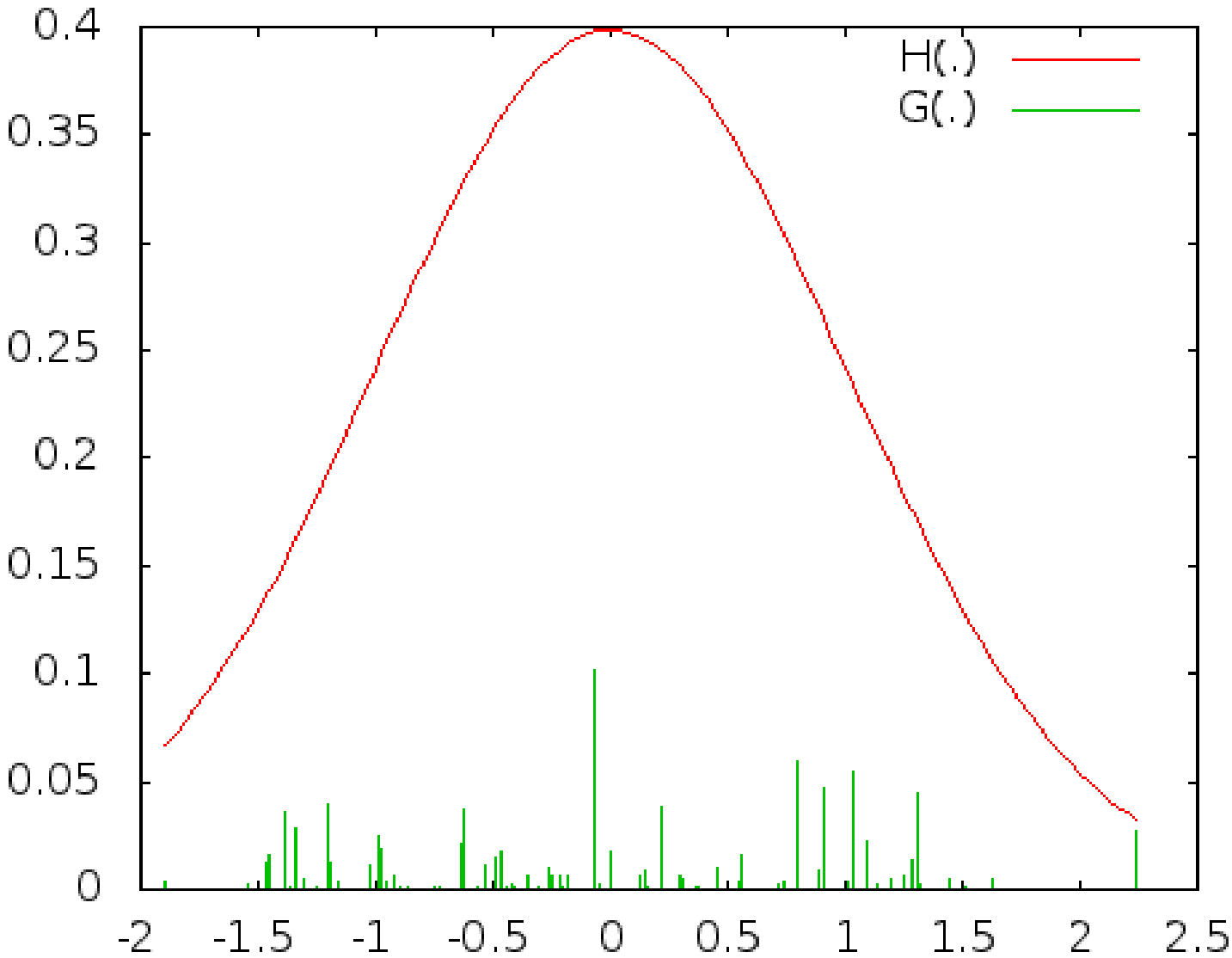
  concentration: which is like an *inverse variance*, denoted α

  mean: called the base distribution, denoted $H(\cdot)$
- Or alternatively one parameter (input)

  measure: denoted $m(\cdot)$

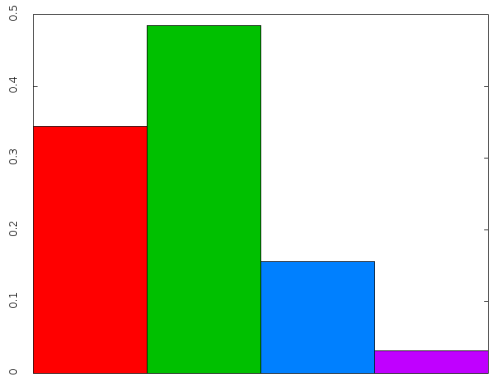
  so $m(A) = \alpha H(A)$ and $\alpha = m(\mathcal{X})$.
- DP defined as the natural extension of a Dirichlet to arbitrary probability measures.
  - See Dirichlet process (Formal definition)[W].
- More definitions later.

Example: $G(\cdot) \sim \mathrm{DP}(1, \mathrm{Gaussian}(0, 1))$

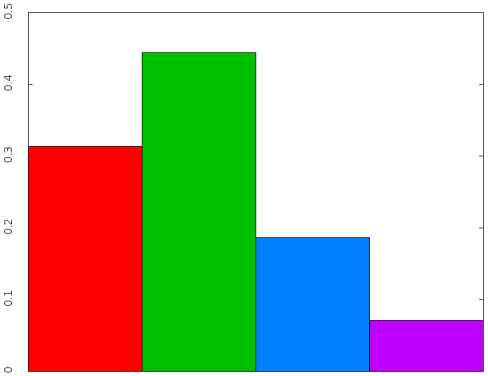


Example: DP on a 4-D vector

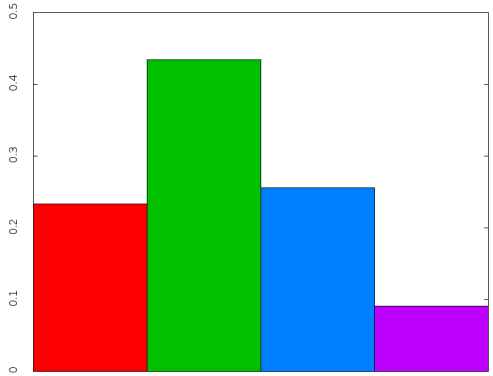
$\vec{p}_0$



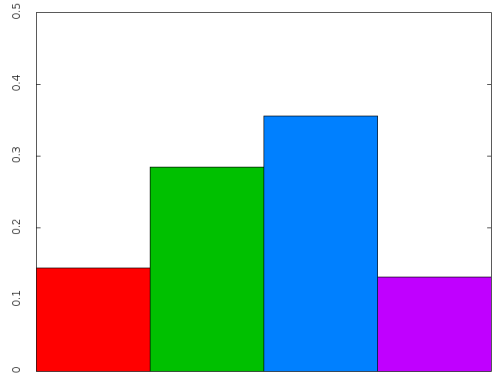
$\vec{p}_1 \sim \text{DP}(500, \vec{p}_0)$



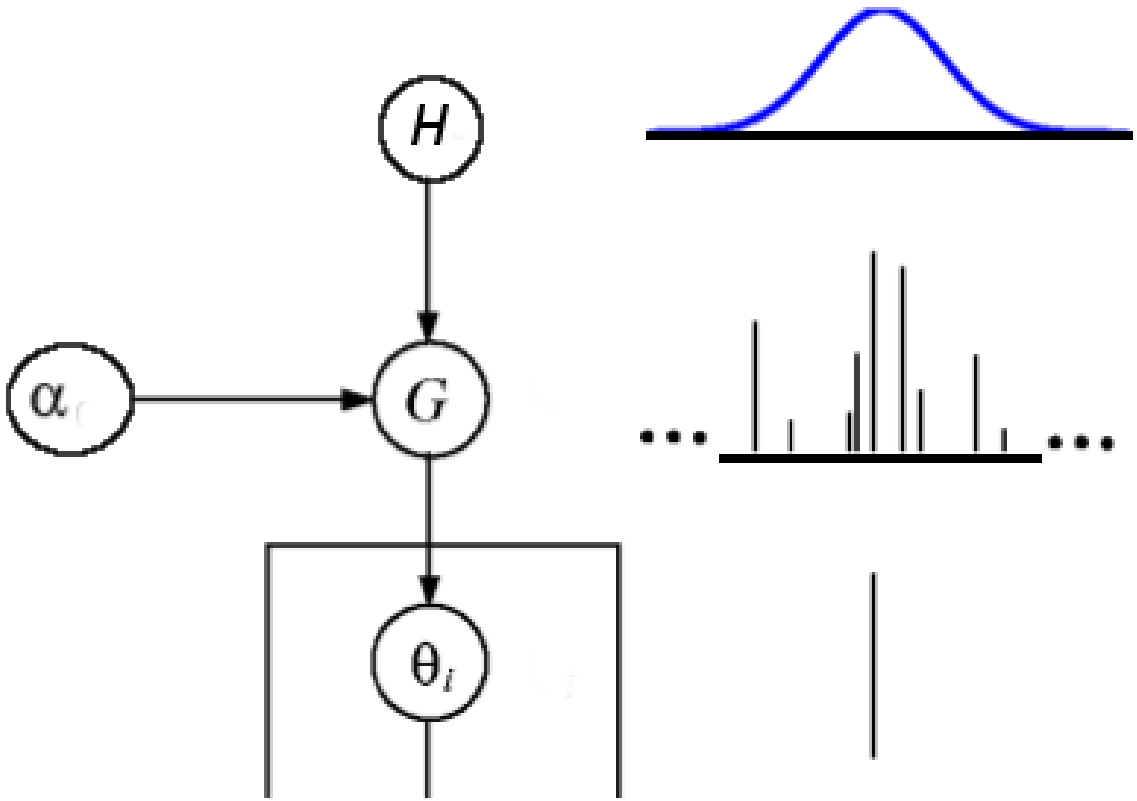
$\vec{p}_2 \sim \text{DP}(5, \vec{p}_0)$



$\vec{p}_3 \sim \text{DP}(0.5, \vec{p}_0)$

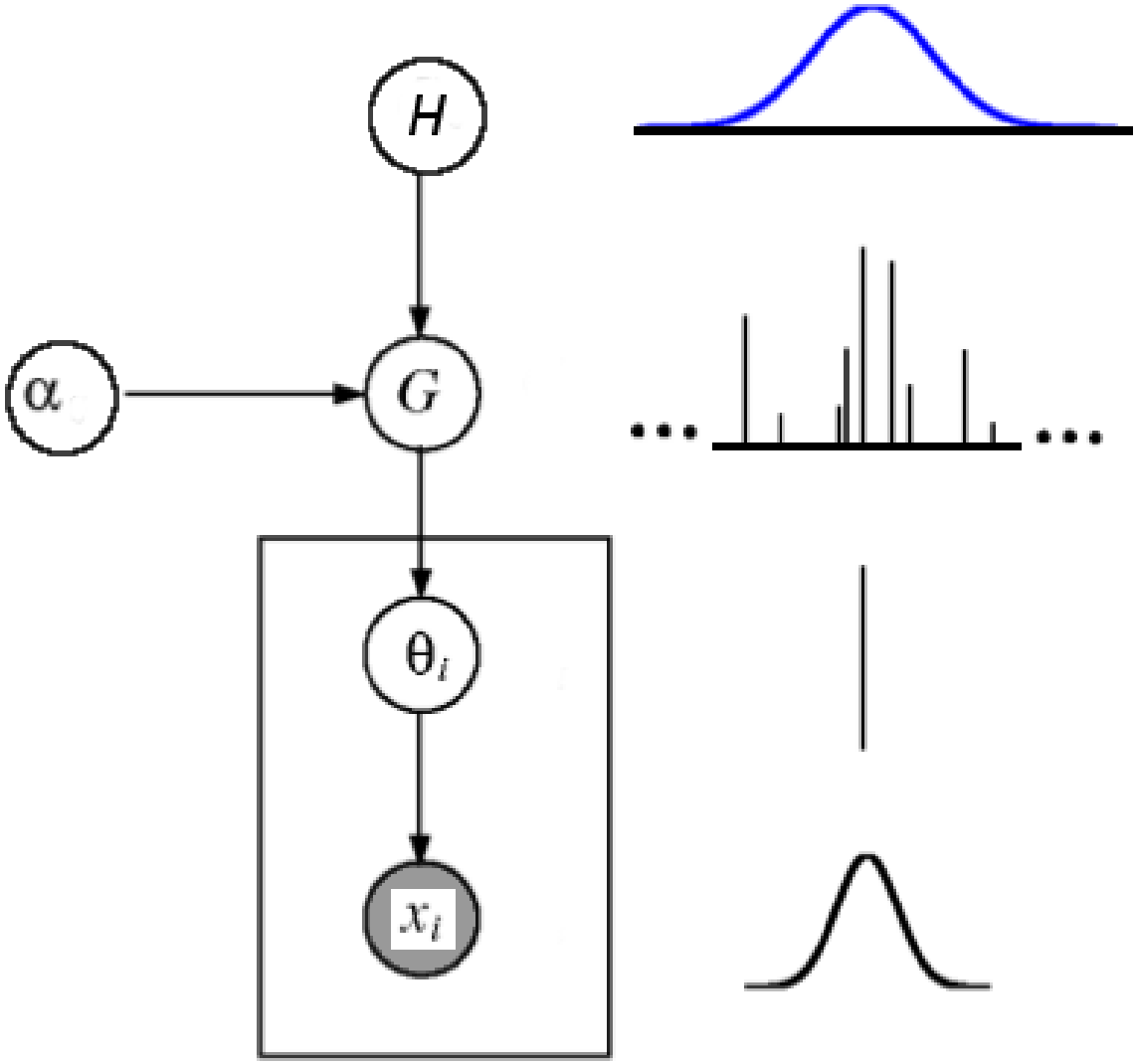


Dirichlet Process on Gaussian



- $H(\cdot)$ is a Gaussian
- G is DP sample of points from a Gaussian weighted according to a DP probability vector
- θ_i is a sample from G, originally from Gaussian $H(\cdot)$

Dirichlet Process for Gaussian Mixture



- θ_i is a sample from G, so have repeats
- $x_i \sim \text{Gaussian}(\theta_i, \sigma)$ (so x_i are clustered)

The Pitman-Yor Process (PYP)

- The Pitman-Yor Process (PYP) has three arguments PYP $(d, \alpha, H(\cdot))$ (or 2 for the 1 parameter DP), extending the DP:
  - discount d is the Zipfian slope for the PYP.
  - Concentration α is inversely proportional to variance.
  - Base distribution $H(\cdot)$ that seeds the distribution and is the mean.
- The DP is also a PYP: PYP $(0, \alpha, H(\cdot)) \equiv$ DP $(\alpha, H(\cdot))$.

The Pitman-Yor Process (PYP)

- The Pitman-Yor Process (PYP) has three arguments $\text{PYP}(d, \alpha, H(\cdot))$ (or 2 for the 1 parameter DP), extending the DP:
  - discount d is the Zipfian slope for the PYP.
  - Concentration α is inversely proportional to variance.
  - Base distribution $H(\cdot)$ that seeds the distribution and is the mean.
- The DP is also a PYP: $\text{PYP}(0, \alpha, H(\cdot)) \equiv \text{DP}(\alpha, H(\cdot))$.
- PYP originally called "two-parameters Poisson-Dirichlet process" (Ishwaran and James, 2003).
- Proper definitions later.

DP and PYP Behaviour

- The output/sample of a DP/PYP always looks like

$$\mu(A) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(A)$$

  where each $\theta_k \sim H(\cdot)$ and $\mu(\mathcal{X}) = \sum_{k=1}^{\infty} p_k = 1$.
- The general functional form is called a species sampling model (SSM):
  - the different θ_k denote species, also called atoms
  - assume for now they are distinct
- Think of $\mu(\cdot)$ as an an infinite probability vector $\vec{p}$ indexed by points $\theta_k \in \mathcal{X}$.

DP and PYP Behaviour

- The output/sample of a DP/PYP always looks like

$$\mu(A) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(A)$$

  where each $\theta_k \sim H(\cdot)$ and $\mu(\mathcal{X}) = \sum_{k=1}^{\infty} p_k = 1$.
- The general functional form is called a species sampling model (SSM):
  - the different θ_k denote species, also called atoms
  - assume for now they are distinct
- Think of $\mu(\cdot)$ as an an infinite probability vector $\vec{p}$ indexed by points $\theta_k \in \mathcal{X}$.
- The $\mu(\cdot)$ are samples from the DP/PYP, and are not continuous distributions, even if the input $H(\cdot)$ is.

Infinite Probability Vectors $\vec{p}$

- The value of p_{58153} is almost surely infinitessimal.
  - and for $p_{9356483202}$, *etc.*

Infinite Probability Vectors $\vec{p}$

- The value of p_{58153} is almost surely infinitessimal.
  - and for $p_{9356483202}$, *etc.*
- At most a smaller few of the p_k can be larger and significant.

  i.e. The number of $p_k > \delta$ must be less than $(1/\delta)$.

  e.g. there can be no more than 1000 p_k greater than 0.001.

Infinite Probability Vectors $\vec{p}$

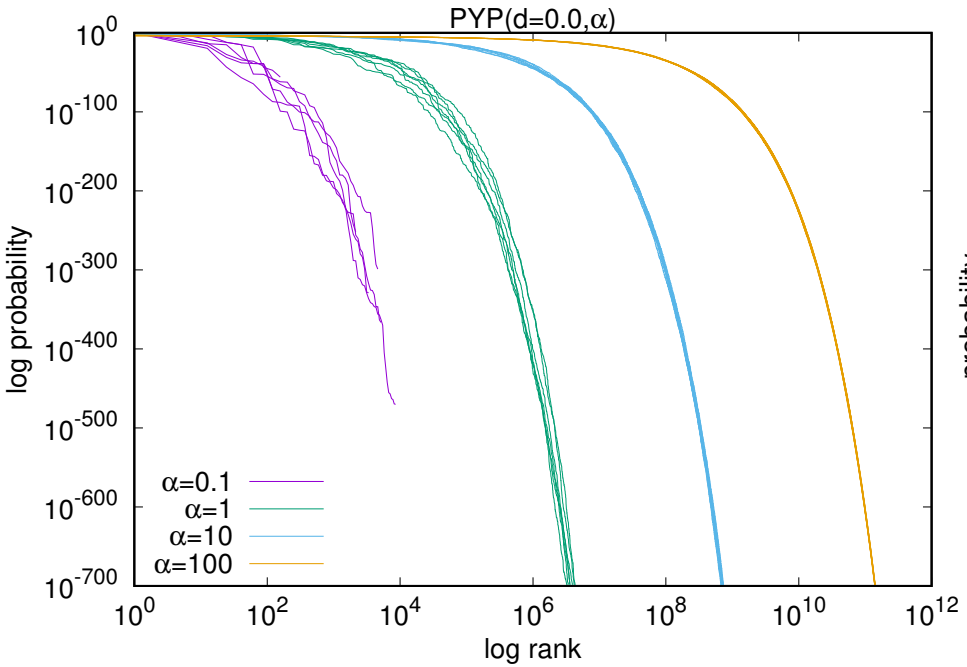
- The value of p_{58153} is almost surely infinitessimal.
  - and for $p_{9356483202}$, *etc.*
- At most a smaller few of the p_k can be larger and significant.

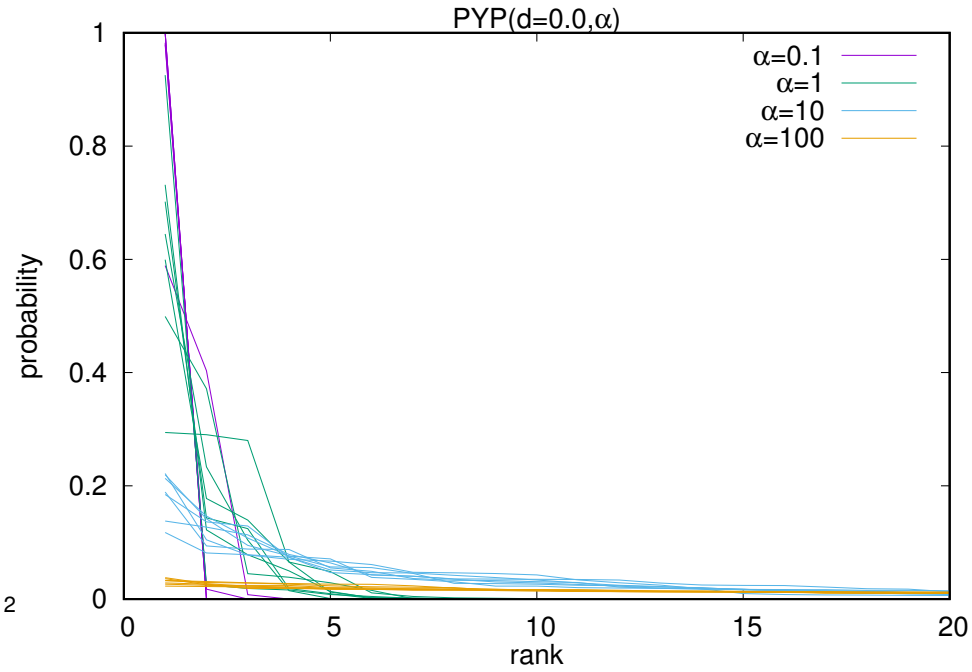
  i.e. The number of $p_k > \delta$ must be less than $(1/\delta)$.

  e.g. there can be no more than 1000 p_k greater than 0.001.
- It is meaningless to consider a p_k without:
  - defining some kind of ordering on indices,
  - only considering those greater than some δ, **or**
  - ignoring the indices and only considering the partitions of data induced by the indices.

Ranked plots of sampled DP probability vectors

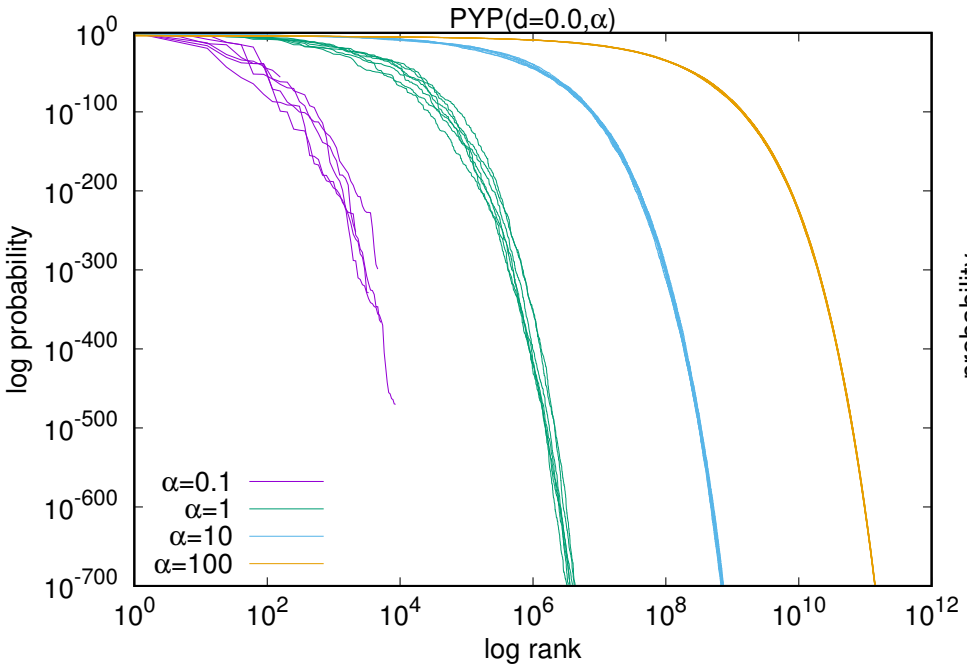
The probabilities in $\vec{p}$ are ranked in decreasing order and then plotted.

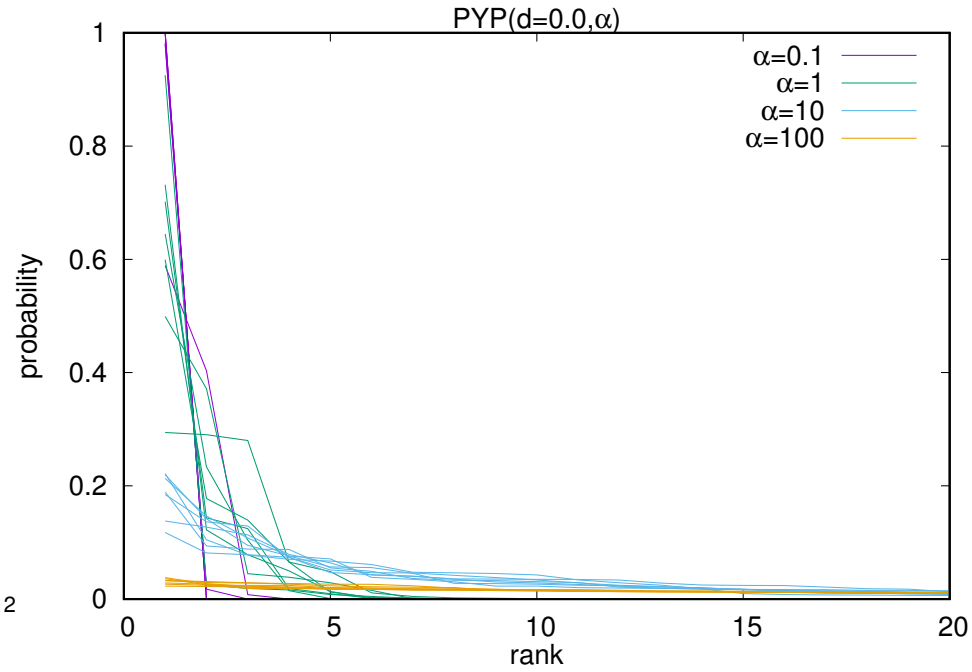




Ranked plots of sampled DP probability vectors

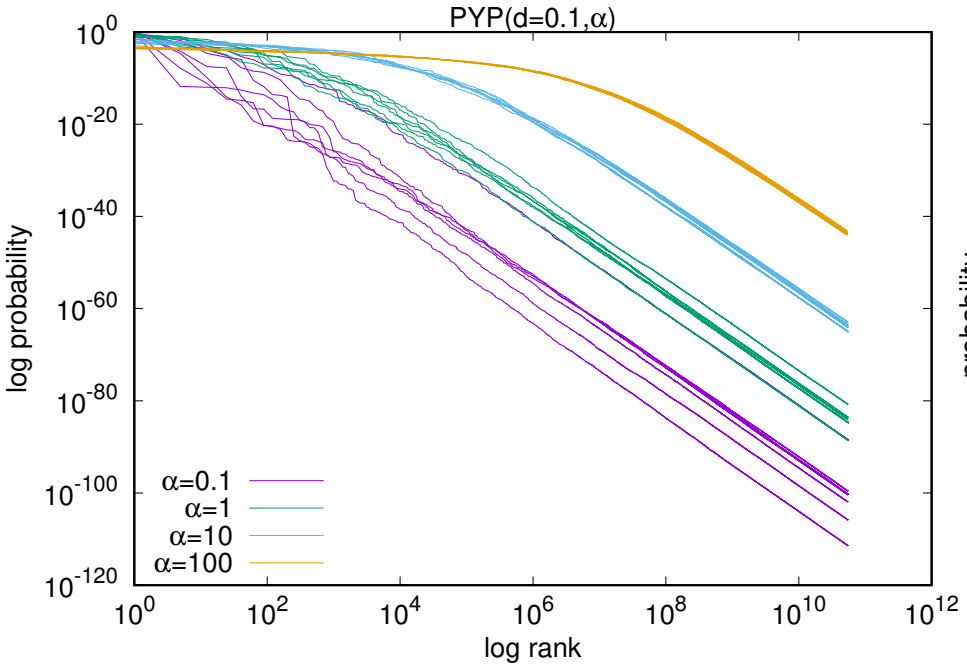
The probabilities in $\vec{p}$ are ranked in decreasing order and then plotted.

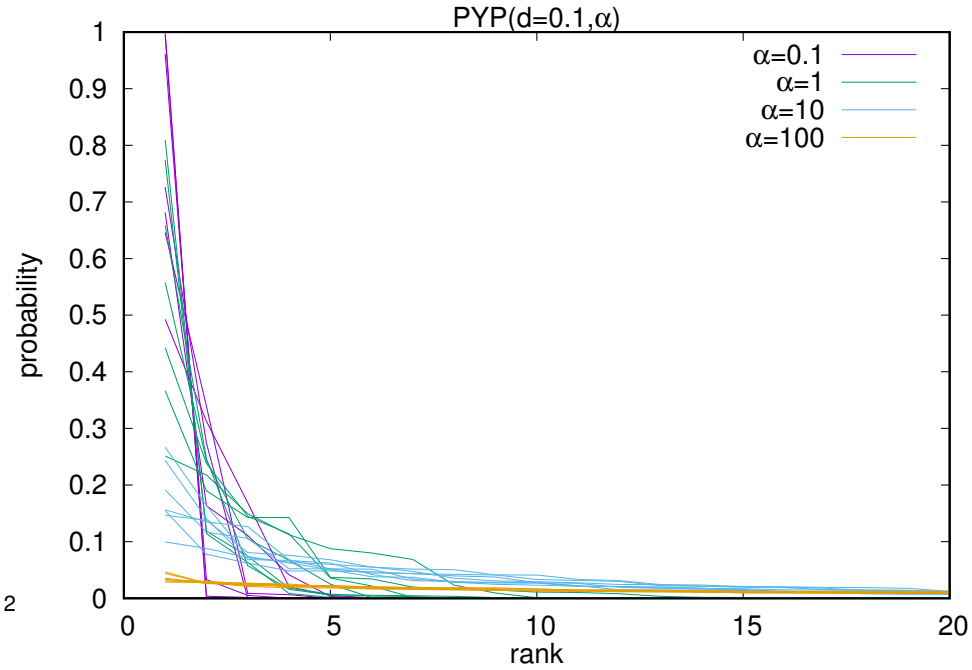




Note: how separated the different plots are for different concentration

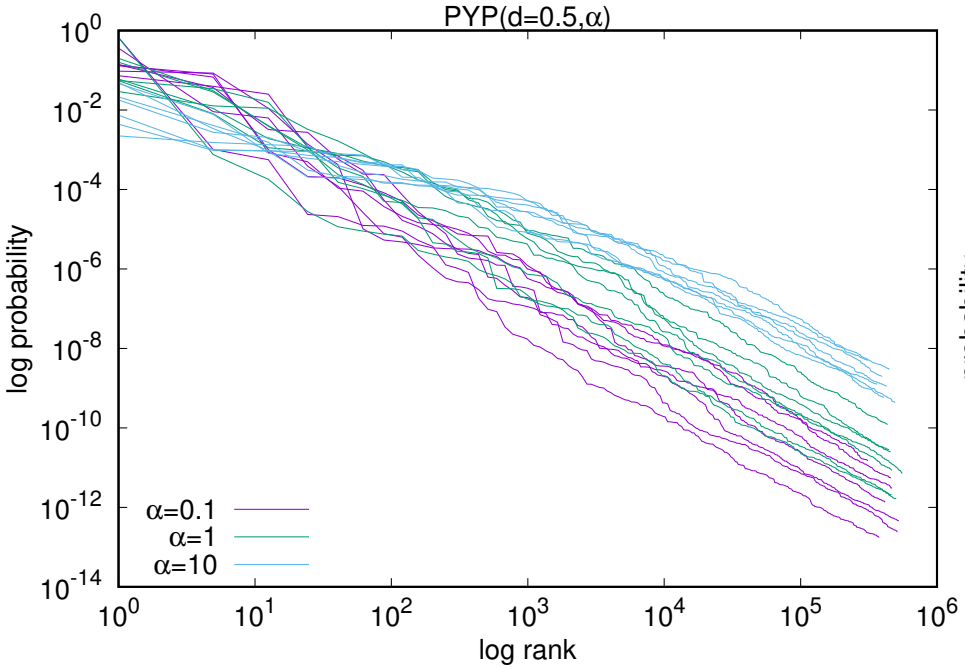
Ranked plots of sampled PYP prob. vectors with $d = 0.1$

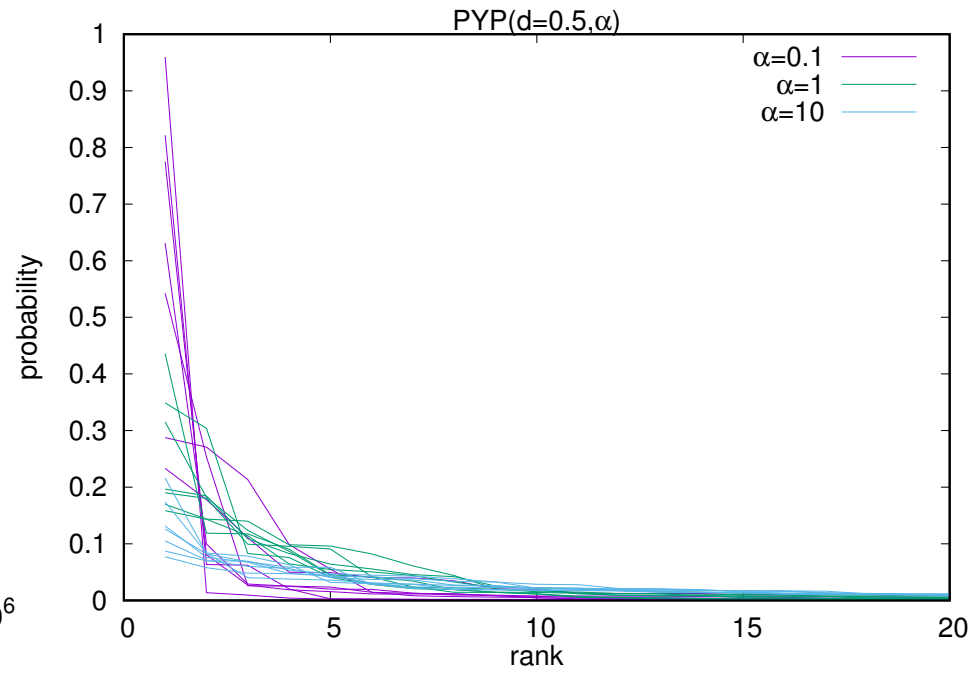




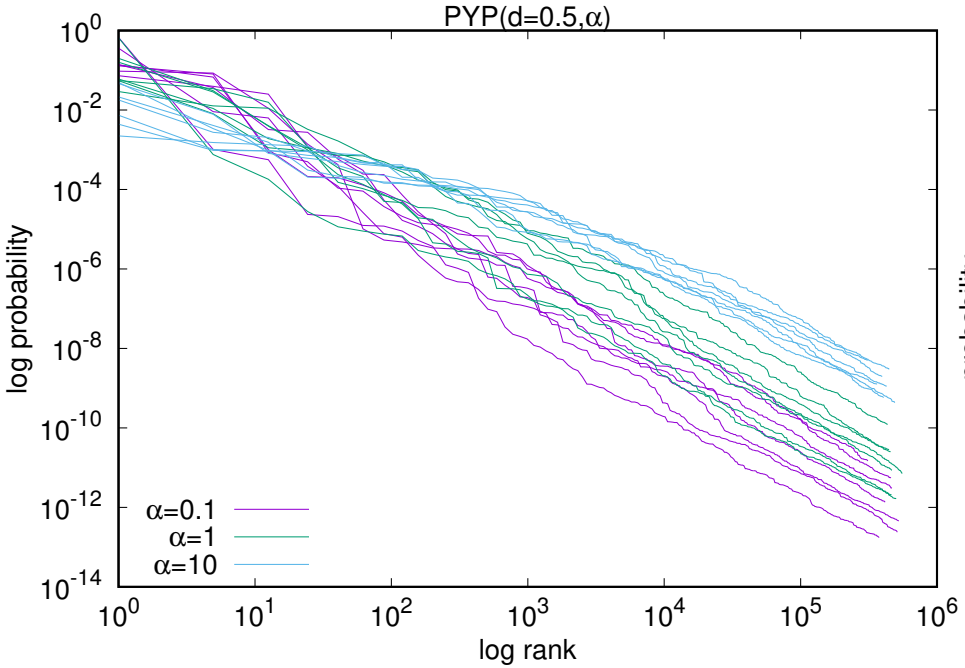
Note: how separated the different plots are for different concentration

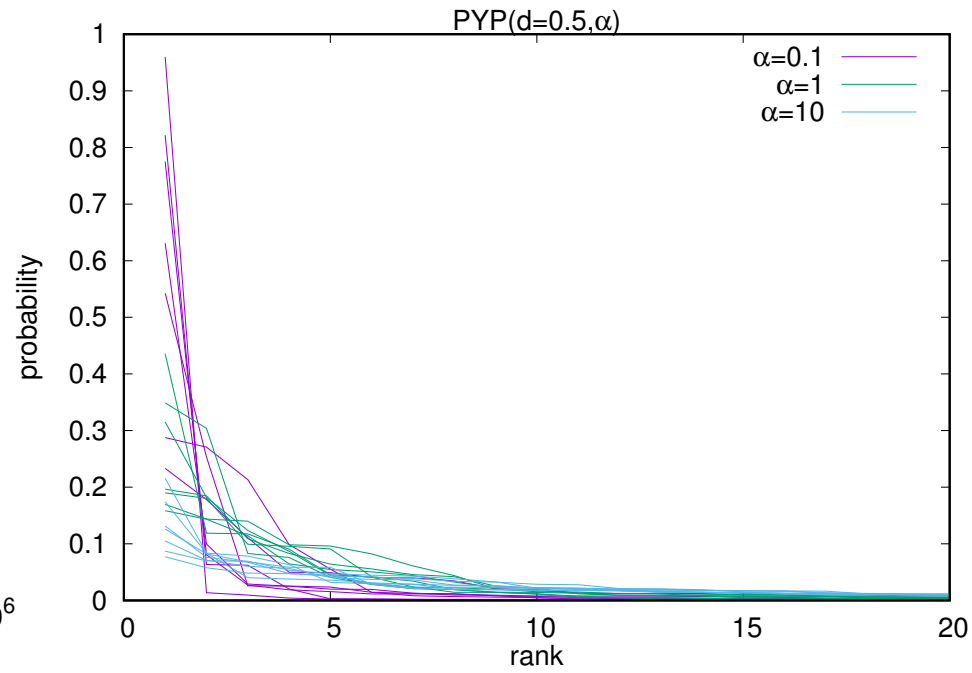
Ranked plots of sampled PYP prob. vectors with $d = 0.5$





Ranked plots of sampled PYP prob. vectors with $d = 0.5$





Note: separation not as clear for different plots for different concentration

Sneak Peak: DP and PYP behaviour

The DP and PYP samples look like $\sum_{k=1}^{\infty} p_k \delta_{\theta_k}(\cdot)$.

- with DP$(\alpha, H(\cdot))$, the $\vec{p}$ probabilities, when rank ordered, behave like a geometric series $p_k \tilde{\propto} \frac{1}{e^{k/\alpha}}$.
- with PYP$(d, \alpha, H(\cdot))$, the $\vec{p}$ probabilities, when rank ordered, behave like a Dirichlet series $p_k \tilde{\propto} \frac{1}{k^{1/d}}$.
- probabilities for the PYP are thus Zipfian (somewhat like Zipf's law$^{\mathcal{W}}$) and converge slower.

Evidence for the PYP

- Evidence or marginal likelihood$^{\mathcal{W}}$ is probability of seeing the data after marginalising out the parameter vector $\vec{p}$ from the SSM.
- Given by $p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot))$.

Evidence for the PYP

- Evidence or marginal likelihood$^{\mathcal{W}}$ is probability of seeing the data after marginalising out the parameter vector $\vec{p}$ from the SSM.
- Given by $p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot))$.
- Used in clustering with a DP/PYP.
  - don't need to know the probabilities $\vec{p}$ when clustering with a DP/PYP!

Evidence for the PYP

- Evidence or marginal likelihood[W] is probability of seeing the data after marginalising out the parameter vector $\vec{p}$ from the SSM.
- Given by $p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot))$.
- Used in clustering with a DP/PYP.
  - don't need to know the probabilities $\vec{p}$ when clustering with a DP/PYP!
- Various ways of deriving this. Depends on the definition used.
  - Cleanest theory uses stochastic processes.

  $\longrightarrow$ We will briefly touch on.
  - Best known uses the Chinese restaurant sampler.

  $\longrightarrow$ We will briefly touch on.

Evidence for the PYP, cont.

- We have a list of K distinct atoms $x_1^*, ..., x_K^*$ and their occurrence counts in the data $n_1, ..., n_K$.
- Then $p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot))$ given by

$$\frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(x_k^*)$$

where:

- rising factorial
$$(\alpha)_N = \alpha(\alpha+1)\cdots(\alpha+N-1) = \frac{\Gamma(\alpha+N)}{\Gamma(\alpha)},$$
- discounted rising factorial
$$(\alpha|d)_K = \alpha(\alpha+d)\cdots(\alpha+d(K-1)) = d^K \frac{\Gamma(\alpha/d+K)}{\Gamma(\alpha/d)},$$
- so $(1-d)_{n_k-1} = (1-d)(2-d)\cdots(n_k-1-d)$, which is 1 when $n_k = 1$.

Evidence for the PYP, Example

$$p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot)) = \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(x_k^*)$$

Example: data is (a, b, a, a, c), then $N = 5, K = 3, \vec{n} = (3, 1, 1)$.

$$\begin{aligned}
& p(\vec{x}|PYP, d, \alpha, H(\cdot)) \\
&= \frac{(\alpha|d)_3}{(\alpha)_5}(1-d)_{3-1}(1-d)_{1-1}(1-d)_{1-1}H(a)H(b)H(c) \\
&= \frac{\alpha(\alpha+d)(\alpha+2)}{\alpha(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)}(1-d)(2-d)H(a)H(b)H(c) \\
&= \underbrace{\left(\frac{\alpha}{\alpha}H(a)\right)}_{\text{1st 'a'}} \underbrace{\left(\frac{\alpha+d}{\alpha+1}H(b)\right)}_{\text{1st 'b'}} \underbrace{\left(\frac{1-d}{\alpha+2}\right)}_{\text{2nd 'a'}} \underbrace{\left(\frac{2-d}{\alpha+3}\right)}_{\text{3rd 'a'}} \underbrace{\left(\frac{\alpha+2d}{\alpha+4}H(c)\right)}_{\text{1st 'c'}}
\end{aligned}$$

Evidence for the PYP, Example

$$p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot)) = \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(x_k^*)$$

Example: data is (a, b, a, a, c), then $N = 5, K = 3, \vec{n} = (3, 1, 1)$.

$$\begin{aligned}
& p(\vec{x}|PYP, d, \alpha, H(\cdot)) \\
&= \frac{(\alpha|d)_3}{(\alpha)_5}(1-d)_{3-1}(1-d)_{1-1}(1-d)_{1-1}H(a)H(b)H(c) \\
&= \frac{\alpha(\alpha+d)(\alpha+2)}{\alpha(\alpha+1)(\alpha+2)(\alpha+3)(\alpha+4)}(1-d)(2-d)H(a)H(b)H(c) \\
&= \underbrace{\left(\frac{\alpha}{\alpha}H(a)\right)}_{\text{1st 'a'}} \underbrace{\left(\frac{\alpha+d}{\alpha+1}H(b)\right)}_{\text{1st 'b'}} \underbrace{\left(\frac{1-d}{\alpha+2}\right)}_{\text{2nd 'a'}} \underbrace{\left(\frac{2-d}{\alpha+3}\right)}_{\text{3rd 'a'}} \underbrace{\left(\frac{\alpha+2d}{\alpha+4}H(c)\right)}_{\text{1st 'c'}}
\end{aligned}$$

Last line is the sequential sampling, CRP, interpretation.

The Chinese Restaurant Process, briefly

- Given $p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot))$

$$\frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(x_k^*) \ .$$

- Then the marginal sequential sampler, the CRP, is

$$p(x_{N+1}|x_1, ..., x_N, PYP, d, \alpha, H(\cdot))$$
$$= \begin{cases} \frac{n_k - d}{N+\alpha} & \text{if } x_{N+1} = x_k^* \text{ for } 1 \le k \le K \\ \frac{dK+\alpha}{N+\alpha} H(x_{K+1}^*) & \text{otherwise, letting } x_{K+1}^* = x_{N+1} \end{cases}$$

The Chinese Restaurant Process, briefly

- Given $p(x_1, ..., x_N | PYP, d, \alpha, H(\cdot))$

$$\frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(x_k^*) \ .$$

- Then the marginal sequential sampler, the CRP, is

$$p(x_{N+1}|x_1, ..., x_N, PYP, d, \alpha, H(\cdot))$$
$$= \begin{cases} \frac{n_k - d}{N+\alpha} & \text{if } x_{N+1} = x_k^* \text{ for } 1 \leq k \leq K \\ \frac{dK+\alpha}{N+\alpha} H(x_{K+1}^*) & \text{otherwise, letting } x_{K+1}^* = x_{N+1} \end{cases}$$

- Few stochastics processes have such simple marginal sequential samplers.

Sampling points from the DP and PYP

- Another way to understand the DP/PYP is to draw samples from $\mu(A) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(A)$,
  e.g. using a Chinese restaurant process.
- By discreteness of $\mu(\cdot)$ samples must repeat, for instance:

$$\theta_{394563},\ \theta_{37},\ \theta_{72345},\ \theta_{37}, \ldots$$

- With a sequence of size N, how many distinct entries K are there?

Sampling points from the DP and PYP

- Another way to understand the DP/PYP is to draw samples from $\mu(A) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(A)$,
  e.g. using a Chinese restaurant process.
- By discreteness of $\mu(\cdot)$ samples must repeat, for instance:

$$\theta_{394563},\ \theta_{37},\ \theta_{72345},\ \theta_{37}, \ldots$$

- With a sequence of size N, how many distinct entries K are there?

e.g. suppose we observe sample $c, f, e, d, c, e, a, c, e, b, c, f, d$ then the sufficient statistics of the sample are:

N:	sample size 13
K:	distinct atoms 6
atoms:	in "first observed" ordering, c, f, e, d, a, b
counts:	for atoms, $\vec{n} = (3, 2, 3, 2, 1, 1)$

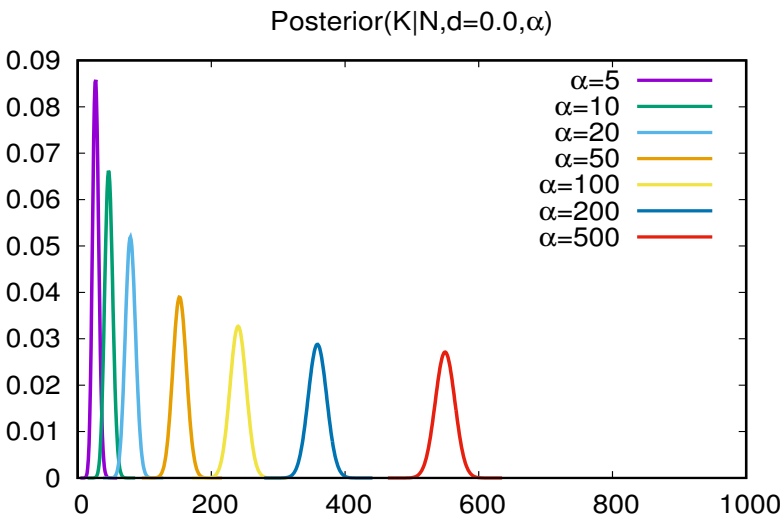
Sampling points from the DP and PYP

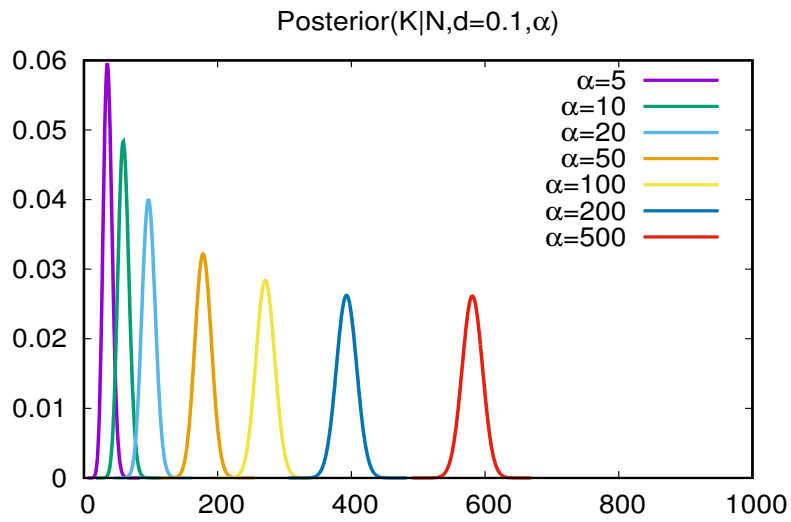
e.g. given 5Gb of text, how many distinct words are there?

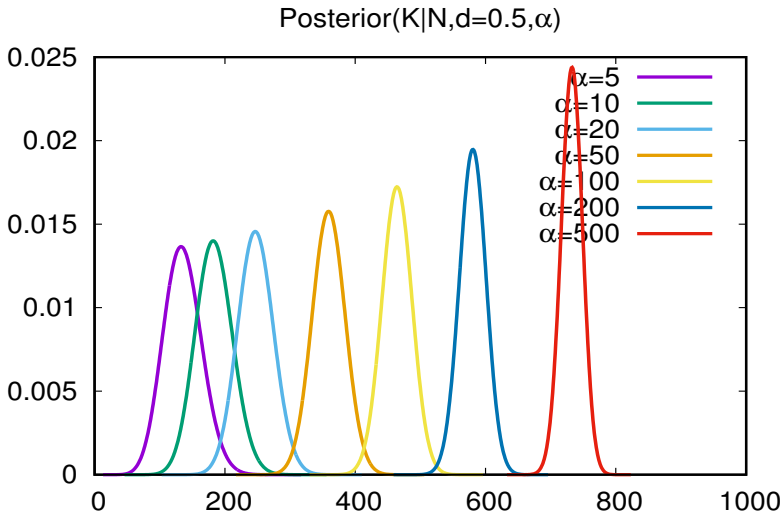
e.g. given a year's worth of music, how many distinct genres are there?

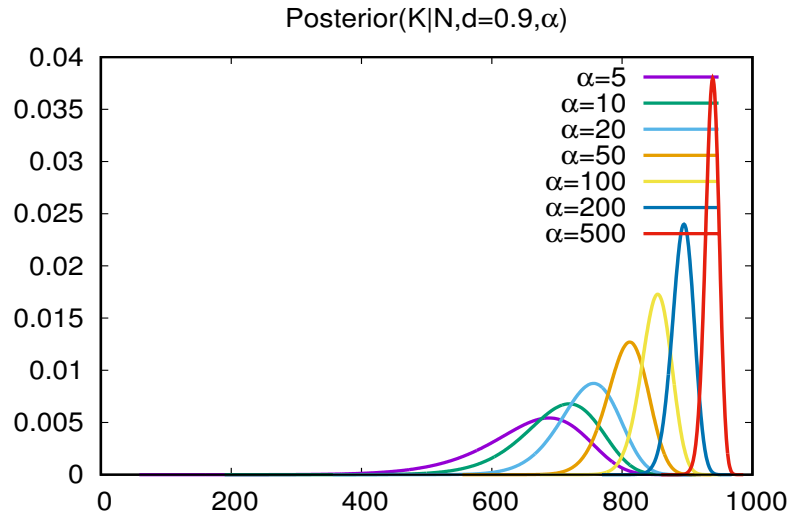
- formally, $\mu(\cdot)$ is called a species sampler, so
  - how many distinct species are there?
- a marginalised version of a DP/PYP is called the Chinese restaurant process, so
  - how many tables in your Chinese restaurant?
  - the distribution on tables is called the Chinese restaurant distribution (CRD)

How Many Species/Tables are There?









Posterior probability on K given $N = 1000$ and different d, α.

Number of Species/Tables

- Theory is well understood with good formula (see Buntine and Hutter, 2012).
- The number of species/tables varies dramatically depending on the discount and the concentration.

Number of Species/Tables

- Theory is well understood with good formula (see Buntine and Hutter, 2012).
- The number of species/tables varies dramatically depending on the discount and the concentration.
- Is approximately Gaussian with a smallish standard-deviation.
- In applications, we should probably sample discount and/or the concentration.
- Note concentration has fast effective samplers, but sampling discount is slow.

DP and PYP Review

- When the base (input) distribution is continous on domain $\mathcal{X}$, the DP/PYP produces a probability vector over a countable number of items in the domain $\mathcal{X}$.
- DP/PYP can be used to build "infinite" mixture models.
- DP/PYP is ideal for non-parametrics!

DP and PYP Review

- When the base (input) distribution is continous on domain $\mathcal{X}$, the DP/PYP produces a probability vector over a countable number of items in the domain $\mathcal{X}$.
- DP/PYP can be used to build "infinite" mixture models.
- DP/PYP is ideal for non-parametrics!

- Some misunderstandings and poor algorithms in the community.

e.g. "A simple example of Dirichlet process mixture inconsistency for the number of components" oral paper at NIPS 2013

e.g. variational stick breaking methods for some hierarchical models

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs
- Alternatives
- Stick Breaking

4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion

Definitions?

- There are several ways of defining DPs and PYPs.

Definitions?

- There are several ways of defining DPs and PYPs.
- Will look at the simplest one, "stick breaking", first.

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs
- Alternatives
- Stick Breaking

4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion

Definitions for DP and PYP

There are several ways of defining a DP/PYP of the form PYP($d, \alpha, H(\cdot)$), or DP($\alpha, H(\cdot)$).

- Generate a $\vec{p}$ with a GEM(d, α), then form an SSM by independently sampling $\theta_k \sim H(\cdot)$ (Pitman and Yor, 1997).

  $\longrightarrow$ We will cover next.
- Proport its existence via the de Finetti-Hewitt-Savage Theorem
  - the Boolean version is de Finetti's Theorem$^{\mathcal{W}}$!

  by saying it has marginal posterior sampler given by a Chinese Restaurant Process, CRP($d, \alpha, H(\cdot)$).

  $\longrightarrow$ We covered briefly earlier.

More Definitions for DP and PYP

There are other ways of defining a DP.

For the DP($\alpha, H(\cdot)$):

- As a natural extension to the Dirichlet in non-discrete or countably infinite domains, see Dirichlet process (Formal definition)$^{\mathcal{W}}$.

  $\longrightarrow$ We will cover later.

- Generate an infinite vector $\vec{w}$ with a Gamma process with rate α and scale 1. Then set $\vec{p} = \frac{1}{\sum_{k=1}^{\infty} w_k} \vec{w}$.

  $\longrightarrow$ Extremely important! We will cover later.

More Definitions for DP and PYP

There are other ways of defining a PYP.

For the PYP($d, \alpha, H(\cdot)$):

- Sample a rate $M \sim \text{gamma}(\alpha/d, 1)$. Generate an infinite vector $\vec{w}$ with a generalised gamma process with discount d, rate M and scale 1. Then set $\vec{p} = \frac{1}{\sum_{k=1}^{\infty} w_k} \vec{w}$.

  $\longrightarrow$ Important! We will cover later.

More Definitions for DP and PYP

There are other ways of defining a PYP.

For the PYP($d, \alpha, H(\cdot)$):

- Sample a rate $M \sim \text{gamma}(\alpha/d, 1)$. Generate an infinite vector $\vec{w}$ with a generalised gamma process with discount d, rate M and scale 1. Then set $\vec{p} = \frac{1}{\sum_{k=1}^{\infty} w_k} \vec{w}$.
  $\longrightarrow$ Important! We will cover later.

NB. You think this is complicated.
You should have seen the original by Pitman!

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

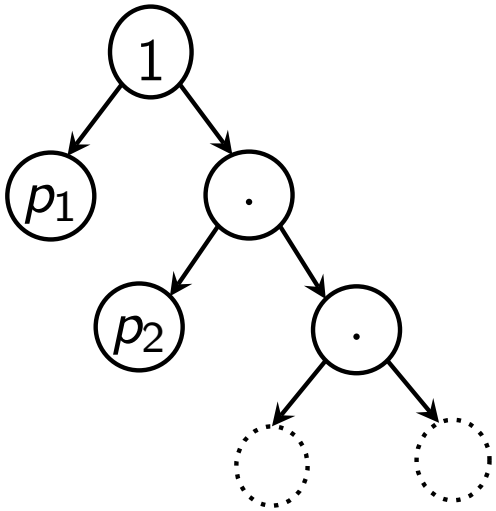
3 Definitions of DPs and PYPs
- Alternatives
- Stick Breaking

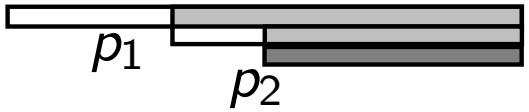
4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion

Stick Breaking

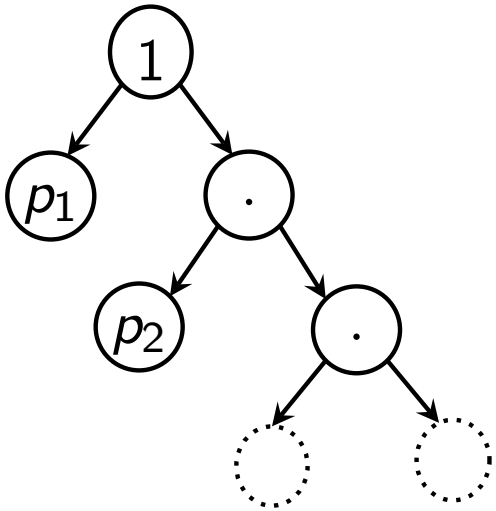


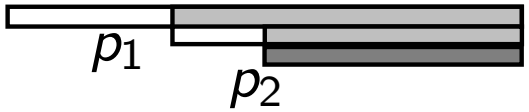


$1 \quad (1 - V_1) \quad (1 - V_1)(1 - V_2)$

- break piece, take remainder, break piece, take remainder, repeat!
- repeatedly sample a proportion using a Beta distribution to break a piece off the remainder stick of length $\left(1 - \sum_{j=1}^{k-1} p_j\right)$
- due to the construction, remainder sticks vanish quickly so p_k usually get vanishingly small as k grows
- see Ishwaran and James (2001)

Stick Breaking, cont





$1 \quad (1 - V_1) \quad (1 - V_1)(1 - V_2)$

The infinite probability vector $\vec{p}$ can be generated incrementally:

$$
\begin{aligned}
p_1 &= V_1 && \text{where } V_1 \sim \text{Beta}(1-d, \alpha+d) \\
p_2 &= V_2\,(1-p_1) && \text{where } V_2 \sim \text{Beta}(1-d, \alpha+2d) \\
p_3 &= V_3\,(1-p_1-p_2) && \text{where } V_3 \sim \text{Beta}(1-d, \alpha+3d) \\
p_k &= V_k \left(1 - \sum_{j=1}^{k-1} p_j\right) && \text{where } V_k \sim \text{Beta}(1-d, \alpha+k\,d)
\end{aligned}
$$

Definition of the GEM

Definition of the GEM distribution

The infinite probability vector $\vec{p}$ generated by the stick breaking construction with discount $d = 0$ and concentration α is said to be from the GEM(α) distribution.

Definition of the GEM

Definition of the GEM distribution

The infinite probability vector $\vec{p}$ generated by the stick breaking construction with discount $d = 0$ and concentration α is said to be from the GEM(α) distribution.

- GEM stands for Griffiths, Engen and McCloskey.
- Use GEM(d, α) for the PYP-like extension with a discount, and it is the "stick breaking" distribution.

Definition of the GEM

Definition of the GEM distribution

The infinite probability vector $\vec{p}$ generated by the stick breaking construction with discount $d = 0$ and concentration α is said to be from the GEM(α) distribution.

- GEM stands for Griffiths, Engen and McCloskey.
- Use GEM(d, α) for the PYP-like extension with a discount, and it is the "stick breaking" distribution.
- Note the p_k are implicitly ordered by construction, earlier values tend to be larger,
  - this is called size-biased ordering, and has a very specific mathematical meaning we will consider later.

Evidence for the GEM

Data for the GEM are in the form of indices for a size-biased ordering.
e.g. 1,1,2,3,2,3,1,4,3,5,2,...

The distribution has a clear form as a set of independent betas, so one can, with difficulty, obtain the evidence.

$$p(x_1, ..., x_N | GEM, d, \alpha)$$
$$= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} \prod_{k=1}^{K} \frac{n_k - d}{\alpha + (k-1)d + \sum_{l=k}^{K} n_l}$$

where n_k is number of occurrences of index k, $n_k > 0$.

Evidence for the GEM

Data for the GEM are in the form of indices for a size-biased ordering.
e.g. 1,1,2,3,2,3,1,4,3,5,2,...

The distribution has a clear form as a set of independent betas, so one can, with difficulty, obtain the evidence.

$$p(x_1, ..., x_N | GEM, d, \alpha)$$
$$= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} \prod_{k=1}^{K} \frac{n_k - d}{\alpha + (k-1)d + \sum_{l=k}^{K} n_l}$$

where n_k is number of occurrences of index k, $n_k > 0$.

Same as the evidence for PYP, but with a final penalty term.

Evidence for the GEM is Different

$$p(x_1, ..., x_N | GEM, d, \alpha)$$
$$= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} \prod_{k=1}^{K} \frac{n_k - d}{\alpha + (k-1)d + \sum_{l=k}^{K} n_l}$$

- Is approximated by:

$$\frac{n_1}{N} \frac{n_2}{N - n_1} \frac{n_3}{N - n_1 - n_2} \frac{n_4}{N - n_1 - n_2 - n_3} \cdots$$

i.e. (probability type 1 will be seen 1st) *times*
(probability type 2 will be seen 2nd given type 1 is seen 1st) *times* ...

Evidence for the GEM is Different

$$p(x_1, ..., x_N | GEM, d, \alpha)$$
$$= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} \prod_{k=1}^{K} \frac{n_k - d}{\alpha + (k-1)d + \sum_{l=k}^{K} n_l}$$

- Is approximated by:

$$\frac{n_1}{N} \frac{n_2}{N - n_1} \frac{n_3}{N - n_1 - n_2} \frac{n_4}{N - n_1 - n_2 - n_3} \cdots$$

i.e. (probability type 1 will be seen 1st) *times*
(probability type 2 will be seen 2nd given type 1 is seen 1st) *times* ...

- Final penalty term corresponds to the choice of ordering of indices.

Historical Context for Stick-breaking + GEM

2001: Ishwaran and James publish the general stick breaking scheme and sampling methods

2003: Pitman generalises theory for other processes in "Poisson-Kingman partitions"

2008: Teh, Furikawa and Welling develop a variational stick breaking scheme for a HDP version of topic models

2008-2014: used widely in variational approaches with HDPs

Historical Context for Stick-breaking + GEM

2001: Ishwaran and James publish the general stick breaking scheme and sampling methods

2003: Pitman generalises theory for other processes in "Poisson-Kingman partitions"

2008: Teh, Furikawa and Welling develop a variational stick breaking scheme for a HDP version of topic models

2008-2014: used widely in variational approaches with HDPs

- Implications of the difference between the GEM and the earlier marginal likelihood are unexplored, though some authors have reported improvements by periodically sorting the GEM.
- Stick breaking gets way too much attention in ML!

Outline

1. Motivation and Context
2. Basic Properties of (non-hierarchical) DPs and PYPs
3. Definitions of DPs and PYPs
4. Stochastic Processes
   - Poisson Point Processes
   - The Gamma Process
   - Constructing a Gamma Process from its Rate
5. Inference on Hierarchies

Definitions?

- Hierarchical DPs and PYPs behave very differently to non-hierarchical.

Definitions?

- Hierarchical DPs and PYPs behave very differently to non-hierarchical.
- To understand them we need to understand Poisson Point Processes.

Outline

1. Motivation and Context
2. Basic Properties of (non-hierarchical) DPs and PYPs
3. Definitions of DPs and PYPs
4. Stochastic Processes
   - Poisson Point Processes
   - The Gamma Process
   - Constructing a Gamma Process from its Rate
5. Inference on Hierarchies

Poisson Point Processes – Motivation

Poisson Point Processes[W] (PPPs) are a special class of spatial processes that behave like Poisson distributions on finite subsets.

Used to:

- analyse variable length parameter vectors and other structures with varying dimensions and elements
  - e.g. stochastic graphs and trees
- analyse time/space varying affects
  - e.g. event bursts in Twitter data
  - e.g. crime incidents in a city

Poisson Point Processes – Motivation

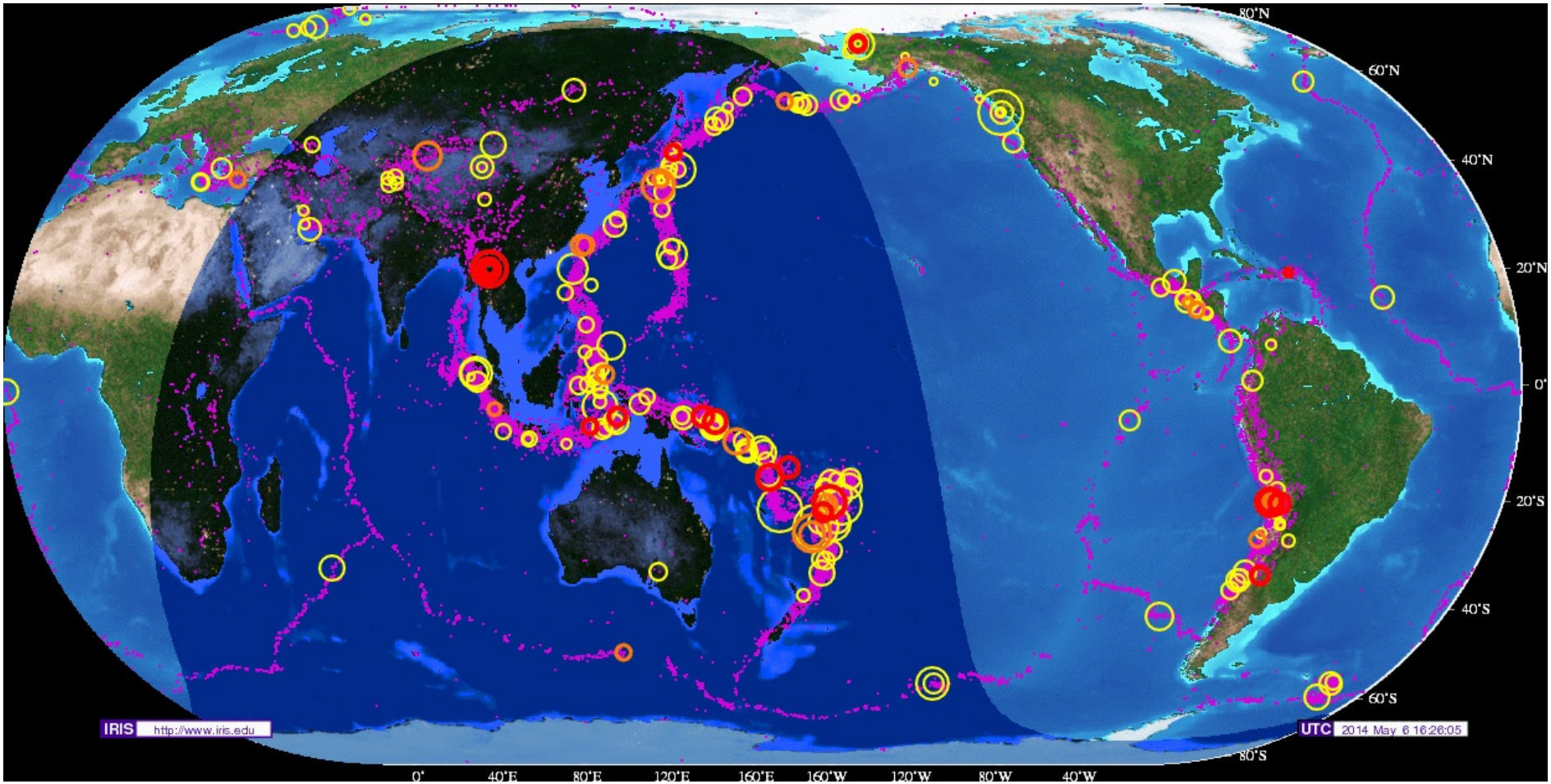
Poisson Point Processes$^{\mathcal{W}}$ (PPPs) are a special class of spatial processes that behave like Poisson distributions on finite subsets.

Used to:

- analyse variable length parameter vectors and other structures with varying dimensions and elements
  - e.g. stochastic graphs and trees
- analyse time/space varying affects
  - e.g. event bursts in Twitter data
  - e.g. crime incidents in a city

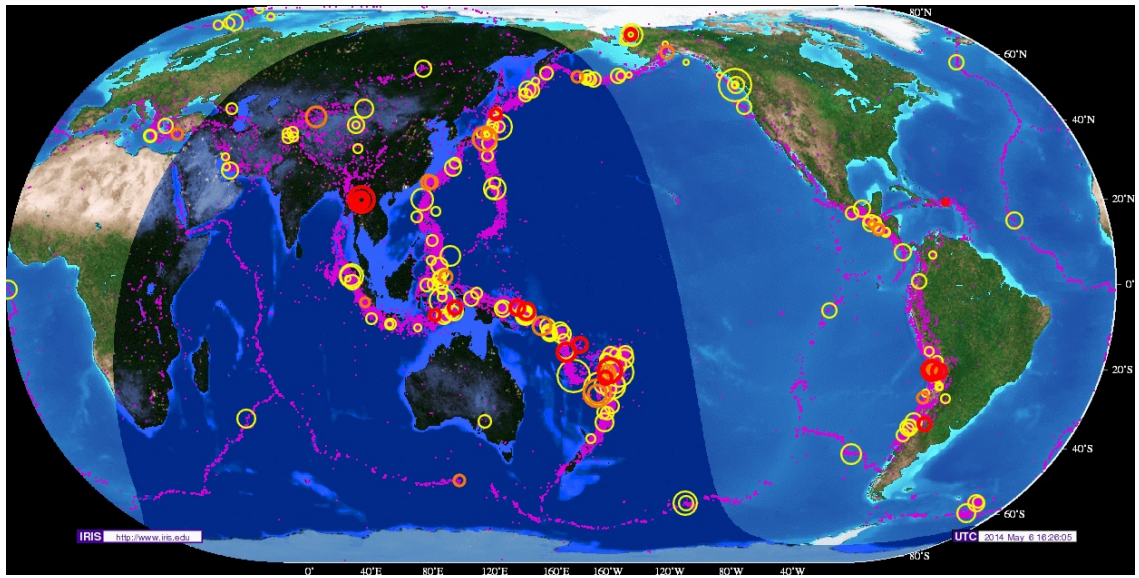
Before giving a proper definition, we will work through the background.

Earthquakes in 2014: A Spatial Process



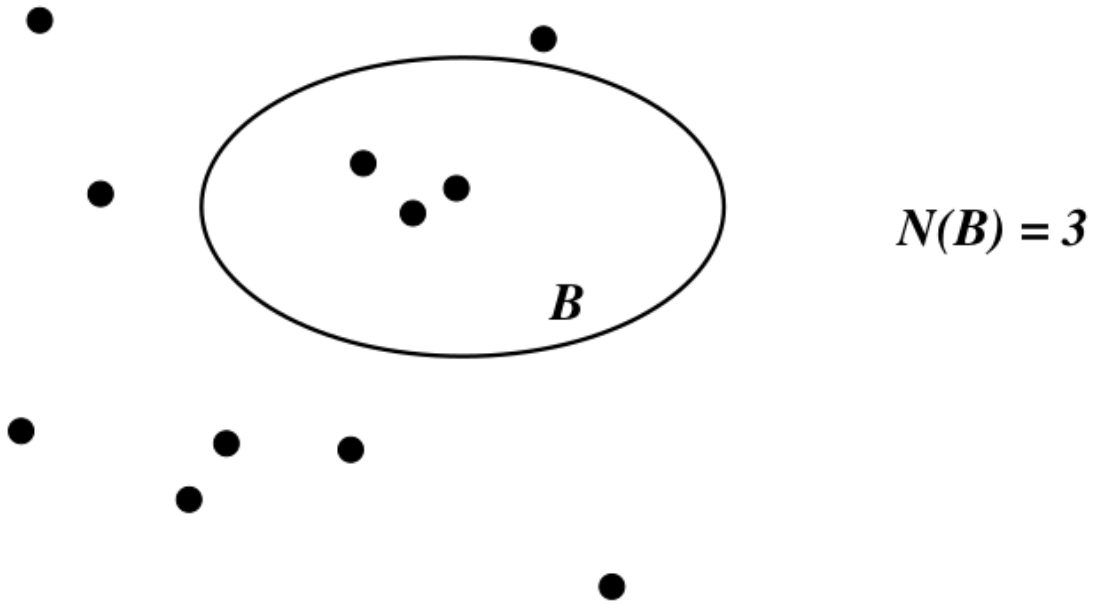
- We have points from a process given by (*latitude*, *longitude*),
- each point marked with the *magnitude*.

Earthquakes in 2014, cont.



- The process part, points generated by a rate, denoted $\rho(\mathit{latitude}, \mathit{longitude})$:
  - number of earthquakes in 2014 per unit area, *i.e.* this rate is spatial,
  - rate is not the same as a probability,
  - can generate a countably infinite number of points too.
- The marked part attaches auxiliary variables to the points in the process generated by a probability $p(\mathit{magnitude} \mid \mathit{latitude}, \mathit{longitude})$.

What is a Spatial Process?



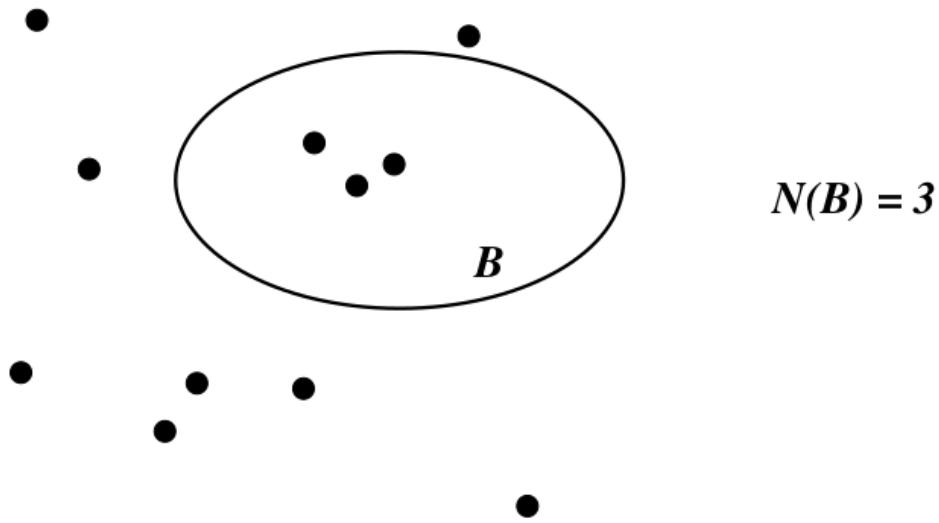
- Spatial process gives a sample of points on a space $\mathcal{X}$.
- Useful statistic: $N(B)$ is the number of points inside subset $B \subseteq \mathcal{X}$.
- We can characterise spatial processes by properties of $N(B)$ for different subsets B.

few following examples from Adrian Badderly, "Spatial Point Processes and their Applications," 2007

What is a Spatial Process?

Definition of Spatial Process

A spatial process[W] on $\mathcal{X}$ is a collection of random variables, representing a countable set of random values in the space $\mathcal{X}$.

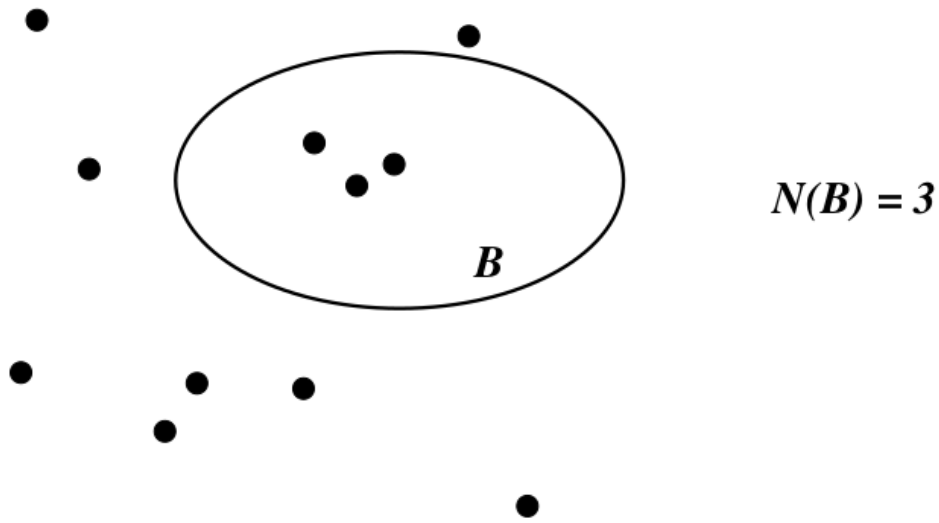


- a single sample from a spatial process depicted in the figure
- $N(B)$ is the number of points inside subset $B \subseteq \mathcal{X}$

What is a Spatial Process?

Definition of Spatial Process

A spatial process[W] on $\mathcal{X}$ is a collection of random variables, representing a countable set of random values in the space $\mathcal{X}$.



- a single sample from a spatial process depicted in the figure
- $N(B)$ is the number of points inside subset $B \subseteq \mathcal{X}$

Note: points in the process can also be marked, by attaching auxiliary variables generated by another PDF.

Additivity for Poissons

For Poissons, data and parameters are jointly additive:

- **if** $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$
  then $(x_1 + x_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$;

Additivity for Poissons

For Poissons, data and parameters are jointly additive:

- **if** $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$
  then $(x_1 + x_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$;
- related property referred to as infinite divisibility (probability)$^{\mathcal{W}}$.

Additivity for Poissons

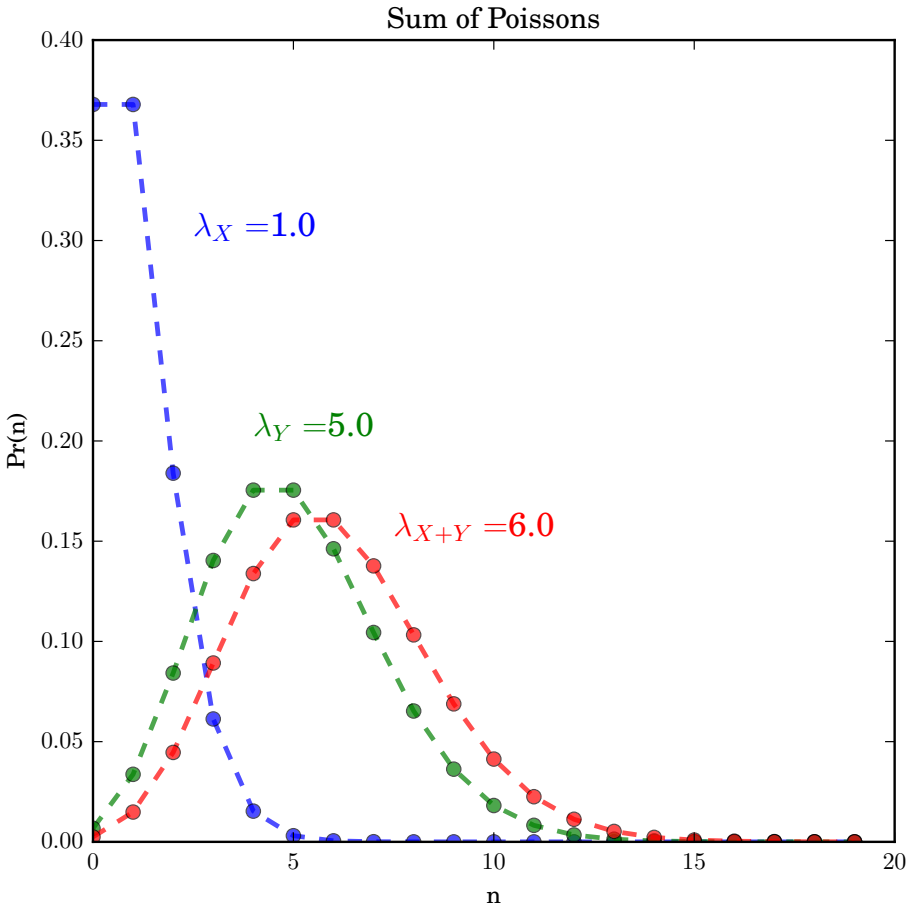
For Poissons, data and parameters are jointly additive:

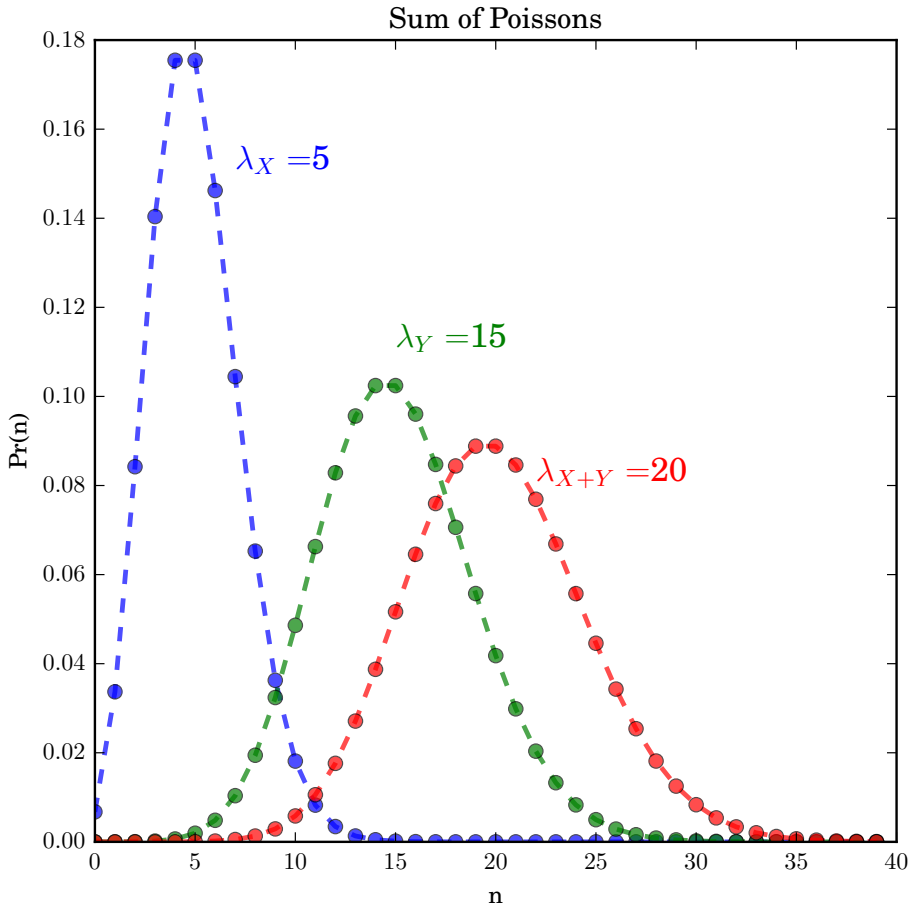
- **if** $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$
  then $(x_1 + x_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$;
- related property referred to as infinite divisibility (probability)$^{\mathcal{W}}$.

Same property holds for the gamma, negative binomial, Gaussian, stable distributions (and their key parameter).

More on this later ....

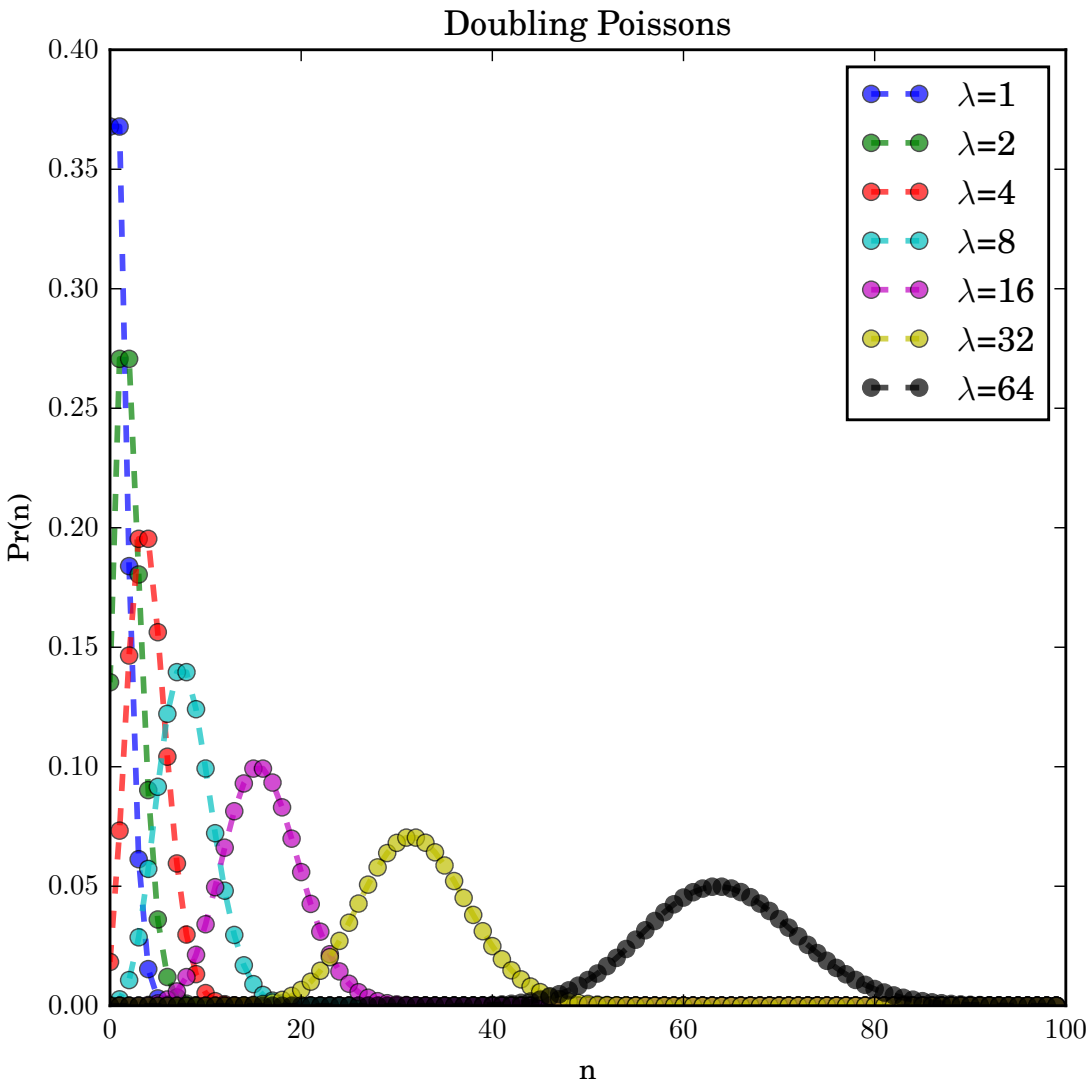
Adding Poissons



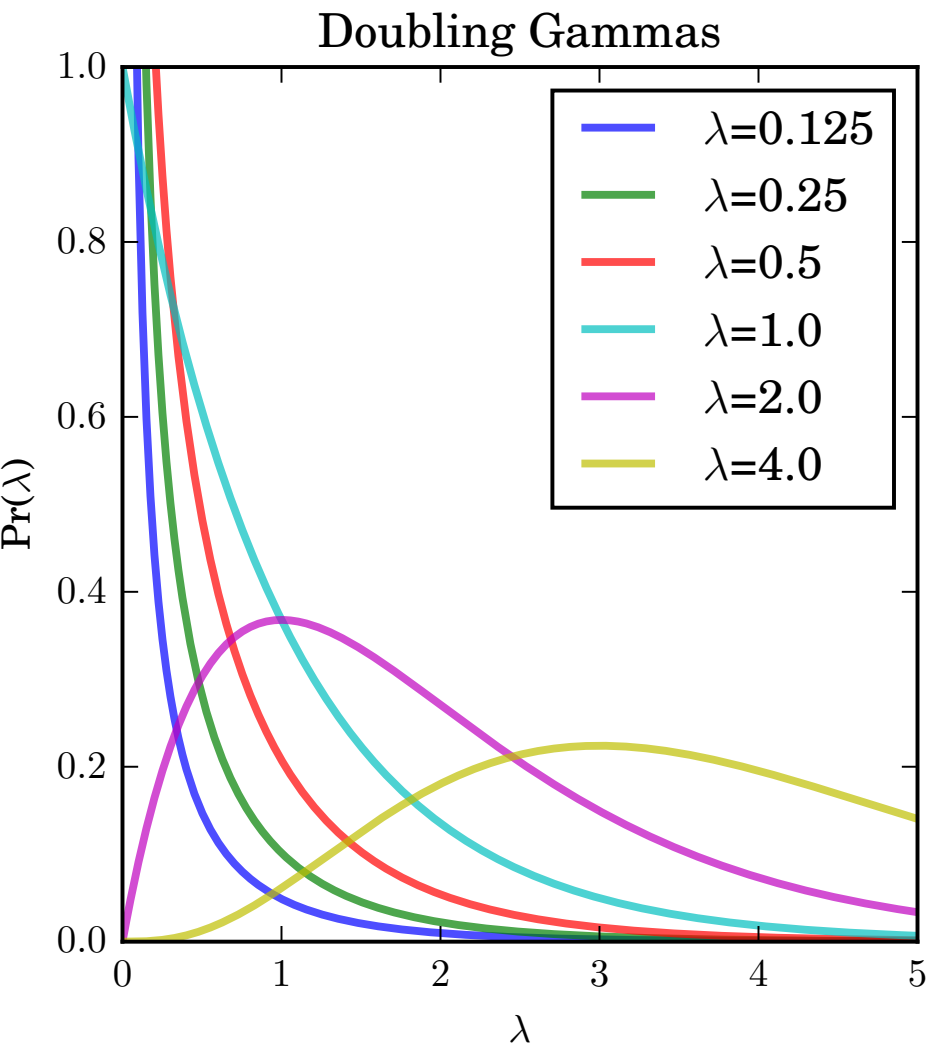


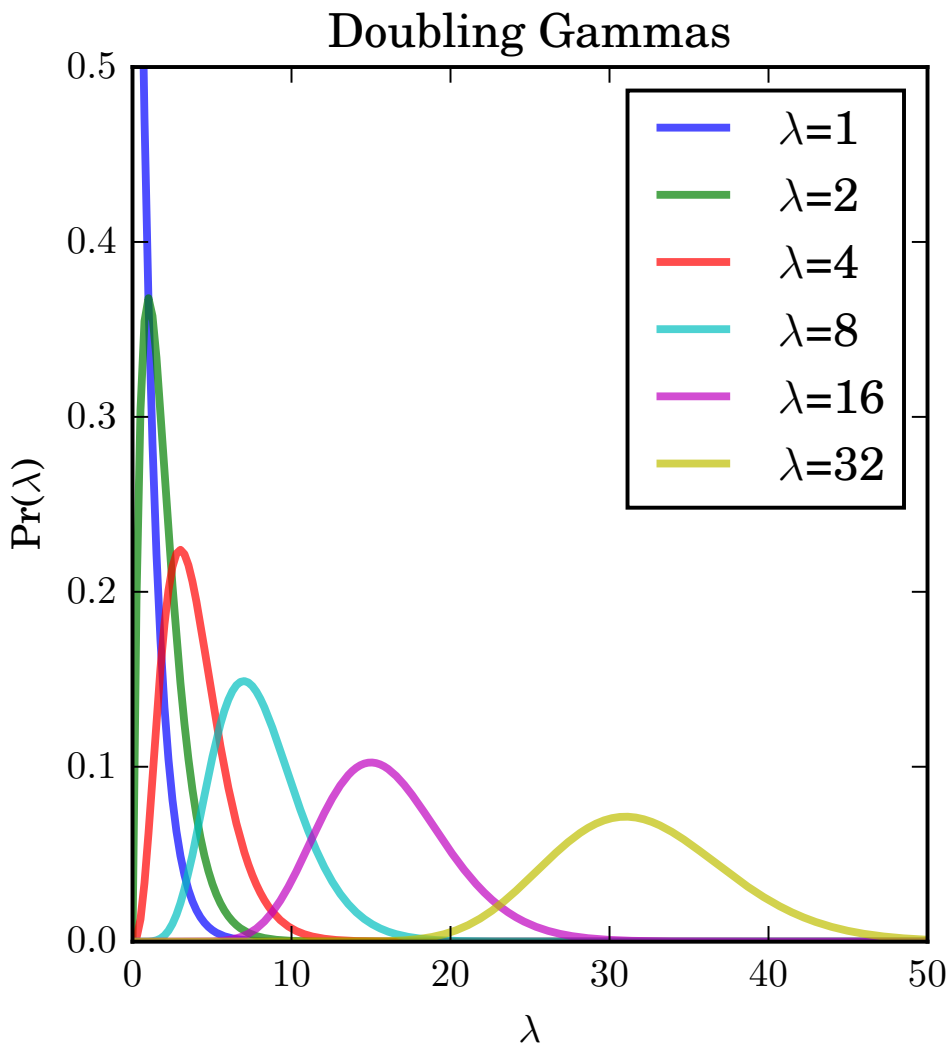
if $X \sim \text{Poisson}(\lambda_X)$ and $Y \sim \text{Poisson}(\lambda_Y)$
then $(X + Y) \sim \text{Poisson}(\lambda_X + \lambda_Y)$

Doubling Poissons



Doubling Gammas





Measures and Poissons

- A measure corresponds to the notion of size or area.
- Note:
  - measures are additive functions, and
  - Poisson distribution is additive.

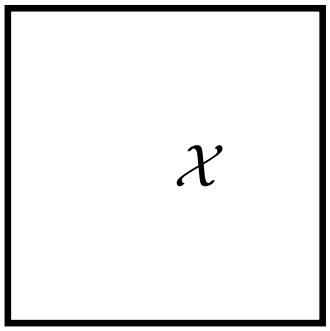
Measures and Poissons

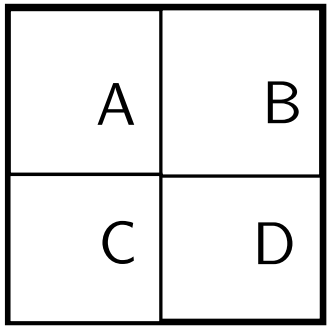
- A measure corresponds to the notion of size or area.
- Note:
  - measures are additive functions, and
  - Poisson distribution is additive.
- **Idea:** use a measure to define a Poisson-based spatial process,
  - so the measure gives the Poisson rate for count of points in any subset of the space,
  - and make non-overlapping subspaces independent.
- Called a **Poisson Point Process**.

Measures and Poissons

- A measure corresponds to the notion of size or area.
- Note:
  - measures are additive functions, and
  - Poisson distribution is additive.
- **Idea:** use a measure to define a Poisson-based spatial process,
  - so the measure gives the Poisson rate for count of points in any subset of the space,
  - and make non-overlapping subspaces independent.
- Called a **Poisson Point Process**.
- Same can be done for the gamma, negative binomial, Gaussian, stable distributions.

A Spatial Interpretation of PPPs

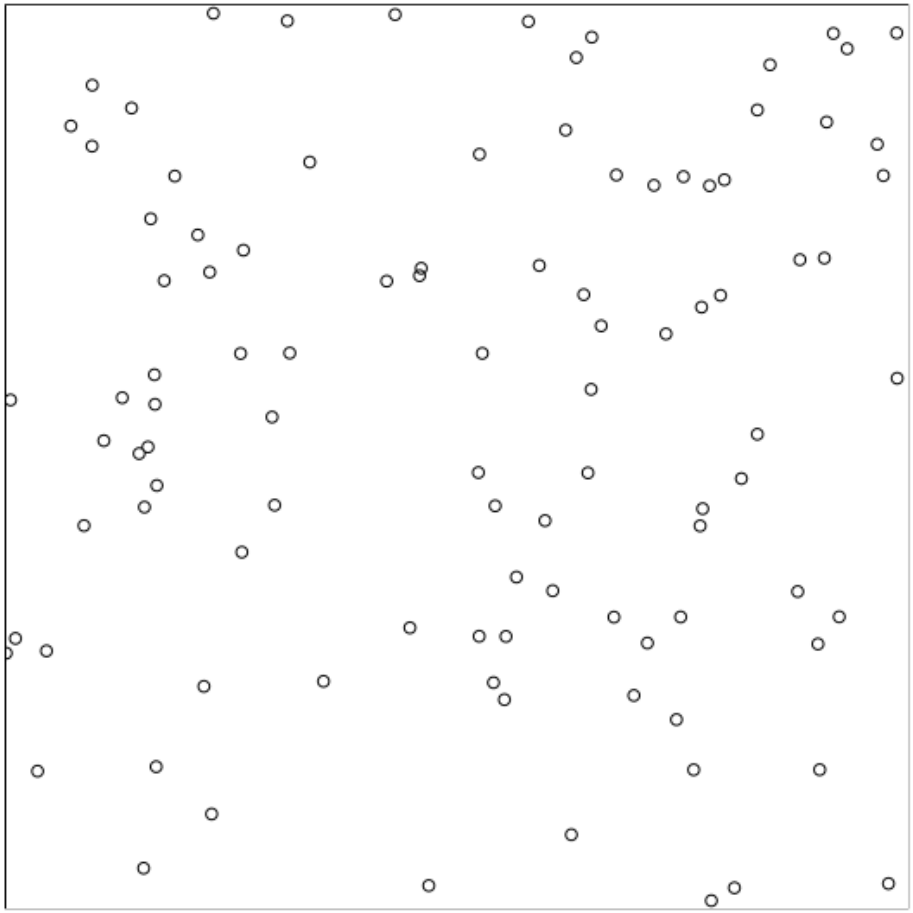




partition $\mathcal{X}$ into A, B, C, D

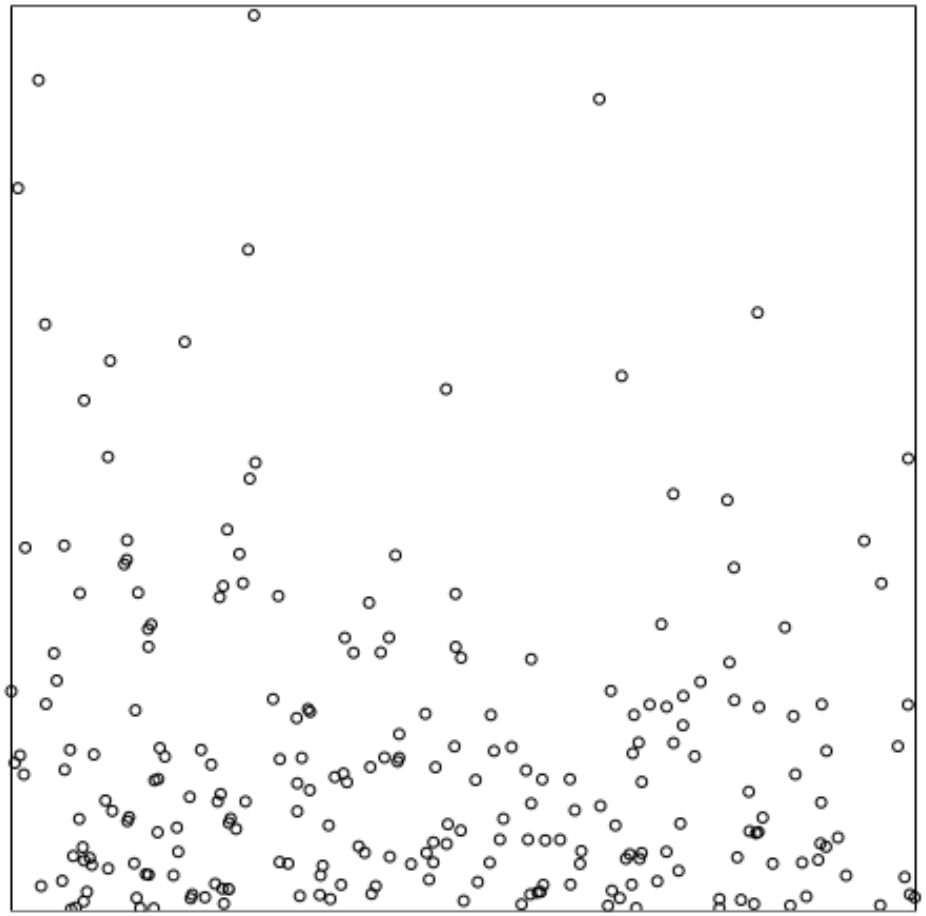
- **Rough Definition:** for any $X \subseteq \mathcal{X}$, $N(X) \sim \text{Poisson}(m(X))$.
- So we have
  - $N(\mathcal{X}) \sim \text{Poisson}(m(\mathcal{X}))$,
  - $N(A) \sim \text{Poisson}(m(A))$,
  - $N(B) \sim \text{Poisson}(m(B))$,
  - $N(C) \sim \text{Poisson}(m(C))$, etc.
- Definition is consistent because Poisson is additive:
  - $N(\mathcal{X})$ is distributed the same as $N(A) + N(B) + N(C) + N(D)$.

Example of a PPP



- PPP is a spatial process
- domain $\mathcal{X} = [0, 1]^2$ with constant rate $\rho(x, y) = 100$
- total rate is $\int_0^1 \int_0^1 \rho(x, y) \mathrm{d}\, x, y = 100$
- generate $N \sim \text{Poisson}(100)$ points uniformly on unit square
- both the number of points N and their locations are sampled

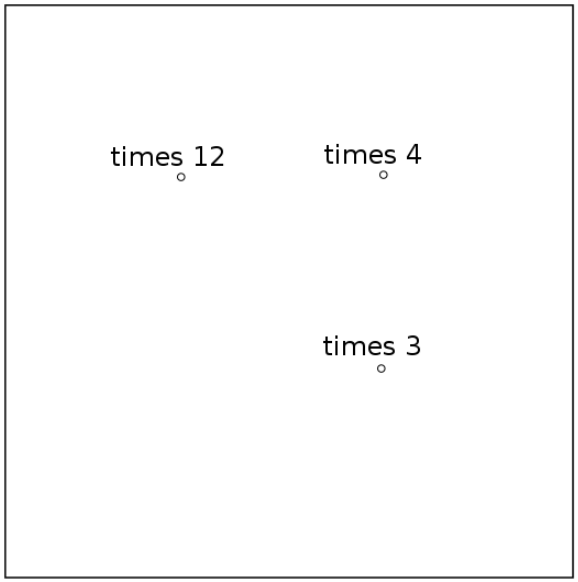
Example of a PPP: Continuous Rate



- PPP on unit square with rate function $\rho(x, y) = e^{7-5y}$
- total rate is $\int_0^1 \int_0^1 \rho(x, y) \mathrm{d}\, x, y = 217.85$
- generate $N \sim \text{Poisson}(217.85)$ points according to PDF

$$p(x, y) = \frac{1}{217.85} e^{7-5y} \ .$$

Example of a PPP: Discrete Rate



- PPP is a unit square with rate function

$$\begin{aligned}\rho(x,y) &= 10\,\delta_{x=1/3,y=2/3} + 5\,\delta_{x=2/3,y=2/3} \\ &\quad +2\,\delta_{x=2/3,y=1/3} + 1\,\delta_{x=1/3,y=1/3}\end{aligned}$$

- total rate is $\int_0^1 \int_0^1 \rho(x,y)\mathrm{d}\,x,y = 18$
- but rate is discrete, so to sample:
  1. sample Poisson(10) points at $(1/3, 2/3)$,
  2. sample Poisson(5) points at $(2/3, 2/3)$,
  3. sample Poisson(2) points at $(2/3, 1/3)$,
  4. sample Poisson(1) points at $(1/3, 1/3)$,
- so behaves like a Poisson at the individual points

The Formal Definition

For any topological space $\mathcal{T}$, $\mathscr{B}(\mathcal{T})$ will denote the Borel σ-field of subsets of $\mathcal{T}$. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be some probability space and $\mathbb{X}$ be complete, separable and endowed with a metric $d_{\mathbb{X}}$. Define on $(\Omega, \mathscr{F}, \mathbb{P})$ a Poisson random measure $\tilde{N}$ on $\mathbb{S} = \mathbb{R}^+ \times \mathbb{X}$ with intensity measure ν. This means that

(i) for any C in $\mathscr{B}(\mathbb{S})$ such that $\nu(C) = \mathbb{E}[\tilde{N}(C)] < \infty$, the probability distribution of the random variable $\tilde{N}(C)$ is Poisson$(\nu(C))$;

(ii) for any finite collection of pairwise disjoint sets, $A_1, \ldots, A_k$, in $\mathscr{B}(\mathbb{S})$, the random variables $\tilde{N}(A_1), \ldots, \tilde{N}(A_k)$ are mutually independent.

Moreover, the measure ν must satisfy the following conditions,

$$\int_{(0,1)} s\nu(\mathrm{d}s, \mathbb{X}) < \infty, \quad \nu([1,\infty) \times \mathbb{X}) < \infty.$$

We refer to Daley & Vere-Jones (1988) for an exhaustive account on Poisson random measures.

What is a Poisson Point Process?

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process$^{\mathcal{W}}$. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $\rho(A), \rho(B) < \infty$ then $A \cap B =$ implies $N(A) \perp\!\!\!\perp N(B)$.

- All works due to additivity of the Poisson.
- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).

What is a Poisson Point Process?

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process$^{\mathcal{W}}$. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $\rho(A), \rho(B) < \infty$ then $A \cap B =$ implies $N(A) \perp\!\!\!\perp N(B)$.

- All works due to additivity of the Poisson.
- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).
- Rate $\rho(x)$ also called intensity, or Lévy measure (in some contexts).

What is a Poisson Point Process?

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process$^{\mathcal{W}}$. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x) \mathrm{d}\, x$, and denote the number of points in A, $N(A) = |X \cap A|$:

- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}\,(\rho(A))$, and
- if $\rho(A), \rho(B) < \infty$ then $A \cap B =$ implies $N(A) \perp\!\!\!\perp N(B)$.

- All works due to additivity of the Poisson.
- Able to work with a rate whose total integral is infinite:
  - as long as it can be partitioned into finite pieces (the measure is σ-finite).
- Rate $\rho(x)$ also called intensity, or Lévy measure (in some contexts).
- Points in sample sometimes called "events".

Poisson Point Process Definition Details

Definition of Poisson point process

A domain $\mathcal{X}$ and a non-negative rate function $\rho(x)$ for $x \in \mathcal{X}$, define a Poisson point process[W]. It generates a countable set of points $X \subset \mathcal{X}$. For any subset $A \subset \mathcal{X}$, define the measure $\rho(A) = \int_A \rho(x)\mathrm{d}x$, and denote the number of points in A, $N(A) = |X \cap A|$:

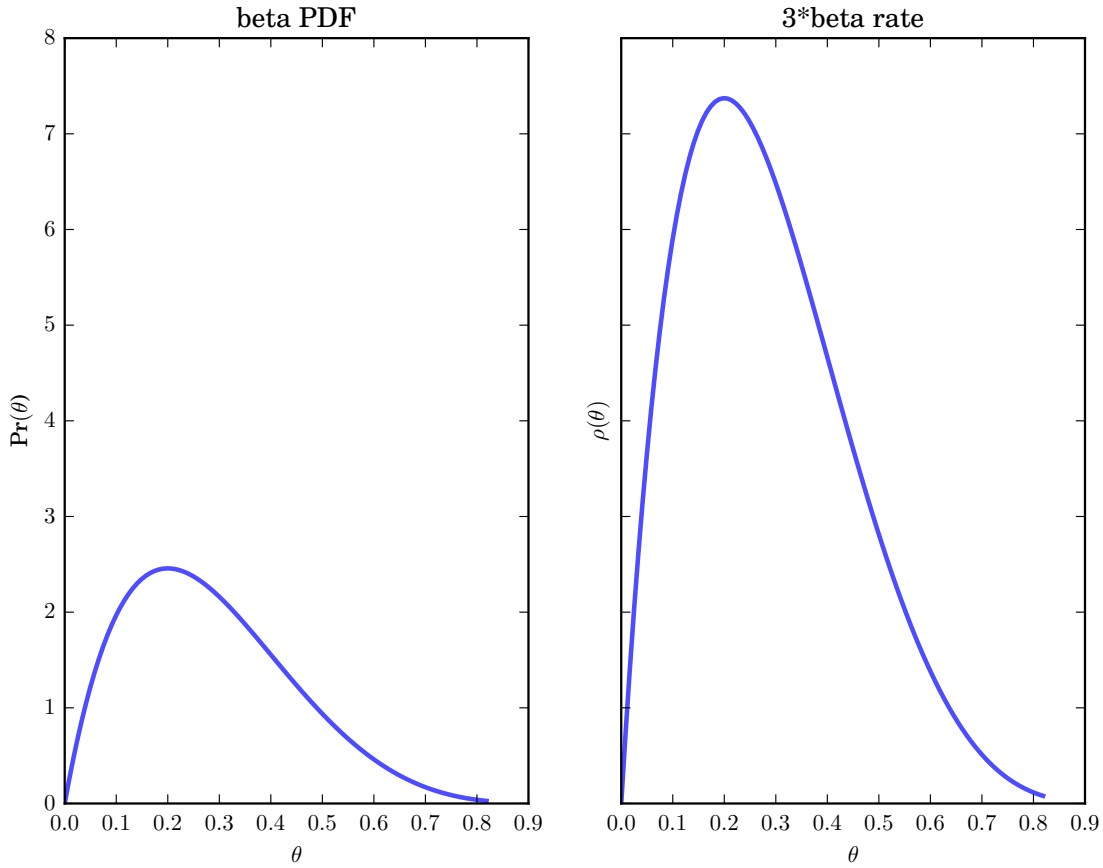
- if $\rho(A) < \infty$ then $N(A) \sim \text{Poisson}(\rho(A))$, and
- if $\rho(A), \rho(B) < \infty$ then $A \cap B = \text{implies } N(A) \perp\!\!\!\perp N(B)$.

constructive

axiomatic

- There are two parts to the definition that define the behaviour of the PPP.

Poisson Point Process versus Probability Density Function



domain $\mathcal{X} = (0, 1)$

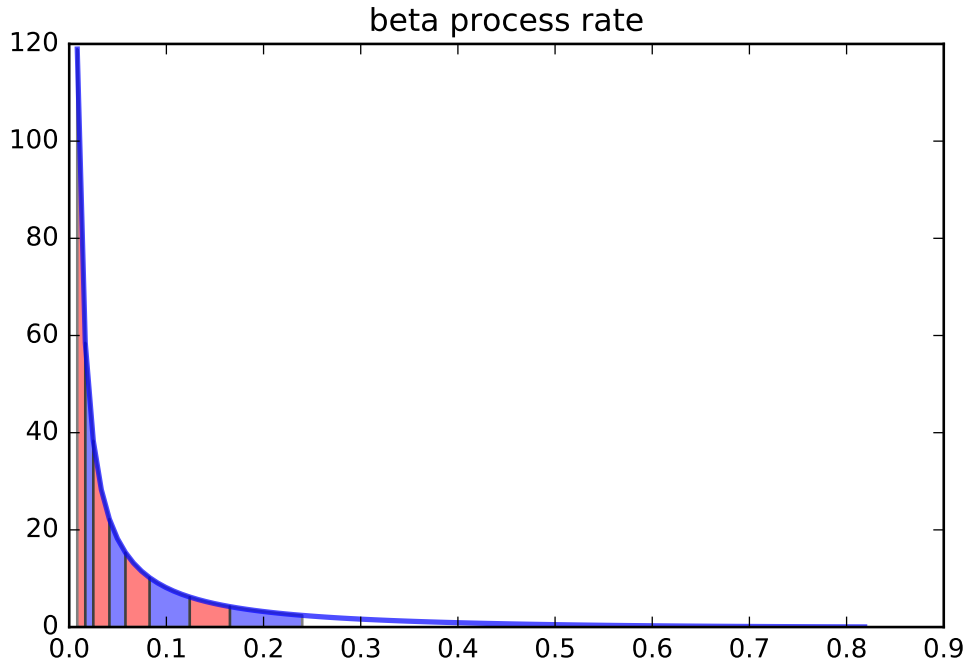
- left is PDF for beta$(5, 2)$, $p(x) = x(1-x)^4/5$
- right is PPP with rate $\rho(x) = 3 * p(x)$
  - note the rate does not sum/integrate to 1

PPP versus PDF, cont.

- A PPP may generate a countably infinite number of points in the full space:
  - is OK as long as you can partition the space into parts with finite measure (so they have finite number of points).
- This is exactly what we need for non-parametrics:
  - to generate an infinite number of possibilities to work with.

PPP versus PDF, cont.

- A PPP may generate a countably infinite number of points in the full space:
  - is OK as long as you can partition the space into parts with finite measure (so they have finite number of points).
- This is exactly what we need for non-parametrics:
  - to generate an infinite number of possibilities to work with.



- rate for "improper" beta$(0, 3)$ has $\rho(x) = x^{-1}(1-x)^2$,
- integral diverges as left boundary $x \to 0$
- so PPP get infinite number of points near zero

Poisson Point Process in the Discrete Case

Interpretation of the *discrete case*:

- $\rho(x) = \sum_k \lambda_k \delta_{x_k}(x)$,
- generate $N_k \sim \text{Poisson}(\lambda_k)$ occurrences of x_k
- superimpose the sets of points

The axiomatic part of the definition says a PPP acts like a Poisson at discrete points.

Poisson Point Process in the Finite Intensity Case

Interpretation of the non-discrete *finite intensity case*:

- assume $\rho(\mathcal{X})$ is finite,
- generate count of points $N \sim \text{Poisson}\left(\rho(\mathcal{X})\right)$
- generate points X_n for $n = 1, ..., N$ by sampling according to the PDF $\rho(x)/\rho(\mathcal{X})$.

The constructive part of the definition says the PPP is like a PDF with probability proportional to $\rho(x)$.

Poisson Point Process in the Infinite Intensity Case

Interpretation of the non-discrete *infinite intensity case*:

- partition up $\mathcal{X}$ into sets A_k for $k = 1, ..., \infty$ so that each $\rho(A_k)$ is finite,
- apply the finite case to the Poisson process on A_k to get samples X_k,
- merge the resulting samples to get sample $X = \bigcup_k X_k$,

Poisson Point Process in the Infinite Intensity Case

Interpretation of the non-discrete *infinite intensity case*:

- partition up $\mathcal{X}$ into sets A_k for $k = 1, ..., \infty$ so that each $\rho(A_k)$ is finite,
- apply the finite case to the Poisson process on A_k to get samples X_k,
- merge the resulting samples to get sample $X = \bigcup_k X_k$,

Method of combining independent PPPs called superposition.

Processes other than Poisson

- PPPs works because the Poisson is additive in its parameter.
- But, so are negative binomials, gammas, stable distributions, *etc.*.
  - look up infinite divisibility (probability)$^{\mathcal{W}}$ for more!

Processes other than Poisson

- PPPs works because the Poisson is additive in its parameter.
- But, so are negative binomials, gammas, stable distributions, *etc.*.
  - look up infinite divisibility (probability)$^{\mathcal{W}}$ for more!
- We can similarly define:
  - negative binomial process,
  - gamma process,
    - no longer integer counts but now a discrete measure (weights on points)
  - stable process,
  - Dirichlet process
    - now normalised weights, a probability vector
    - by normalising a gamma process.

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs

4 Stochastic Processes
- Poisson Point Processes
- The Gamma Process
- Constructing a Gamma Process from its Rate

5 Inference on Hierarchies

A Gamma Process, Axiomatically

Axiomatic definition of gamma process

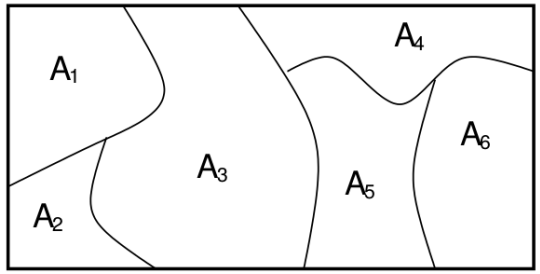
A domain $\mathcal{X}$ and a measure $m(x)$ for $x \in \mathcal{X}$ with finite integral ($\int_{\mathcal{X}} m(x)\mathrm{d}\,x < \infty$) define a gamma process$^{\mathcal{W}}$ with scale β, denoted $\mathrm{GP}(m(\cdot), \beta)$. The gamma process generates a random measure $H(\cdot)$ also on $\mathcal{X}$ that must satisfy for $A \subseteq \mathcal{X}$

- $H(A) \sim \mathrm{Gamma}\,(m(A), \beta)$, and
- if $A \cap B =$ then $H(A) \perp\!\!\!\perp H(B)$.

- Called an axiomatic definition because it only says what properties a $H(\cdot)$ must have.
- Main modification needed from the PPP:
  - no longer returning number of points $N(\cdot)$ but a measure $H(\cdot)$ (the $N(\cdot)$ is called a counting measure)

Gamma Process versus Dirichlet Process

Consider a partition $(A_1, A_2, ..., A_K)$ of $\mathcal{X}$, and measure $m(\cdot)$ on $\mathcal{X}$.



- Equivalently, $H(\cdot) \sim \text{GP}(m(\cdot), \beta)$ when for any such partition, for all k

$$H(A_k) \sim \text{Gamma}(m(A_k))$$

and moreover for all $j \neq k$, $H(A_j) \perp\!\!\!\perp H(A_k)$.

- Axiomatic definition of the DP is $G(\cdot) \sim \text{DP}(m(\cdot))$ when for any such partition

$$(G(A_1), G(A_2), ..., G(A_K)) \sim \text{Dirichlet}(m(A_1), m(A_2), ..., m(A_K))$$

- Note the first implies the second by the normalisation property of gamma distributions.

Hierarchical Processes

- If the measure is discrete, the axiomatic definition is constructive.
- The gamma process does an element-wise gamma sample for the input vector.
  - i.e. draw a Gamma distribution at each atom
    - To sample from GP $(\sum_k w_k \delta_{x_k}(x), \beta)$, sample $\mu_k \sim \text{gamma}(w_k, \beta)$ for all k and return $H(x) = \sum_k \mu_k \delta_{x_k}(x)$.
- Likewise, the DP is just a Dirichlet distribution on the input measure.
  - i.e. To sample from DP $(\sum_k w_k \delta_{x_k}(x))$, sample $\vec{\mu} \sim \text{Dirichlet}(\vec{w})$ then return $H(x) = \sum_k \mu_k \delta_{x_k}(x)$.

Hierarchical Processes, cont.

- Likewise, the PYP (with discount d) is got by normalising an element-wise Tweedie distribution$^{\mathcal{W}}$ (with exponent d) on the input measure.
  - i.e. To sample from PYP $(d, \sum_k w_k \delta_{x_k}(x))$,
    sample a rate $M \sim \text{gamma}(\alpha/d, 1)$,
    sample element-wise $\mu_k \sim \text{Tweedie}(d, (Mw_k)^{1/d}, 1)$
    then return $H(x) = \frac{1}{\sum_k \mu_k} \sum_k \mu_k \delta_{x_k}(x)$.

Hierarchical Processes, cont.

- Likewise, the PYP (with discount d) is got by normalising an element-wise Tweedie distribution$^{\mathcal{W}}$ (with exponent d) on the input measure.
  - i.e. To sample from PYP $(d, \sum_k w_k \delta_{x_k}(x))$,
    sample a rate $M \sim \text{gamma}(\alpha/d, 1)$,
    sample element-wise $\mu_k \sim \text{Tweedie}(d, (Mw_k)^{1/d}, 1)$
    then return $H(x) = \frac{1}{\sum_k \mu_k} \sum_k \mu_k \delta_{x_k}(x)$.
- Why a Tweedie distribution? It is another name for the distribution of the sum (see slide few pages on) for the generalised gamma process.

Outline

1 Motivation and Context

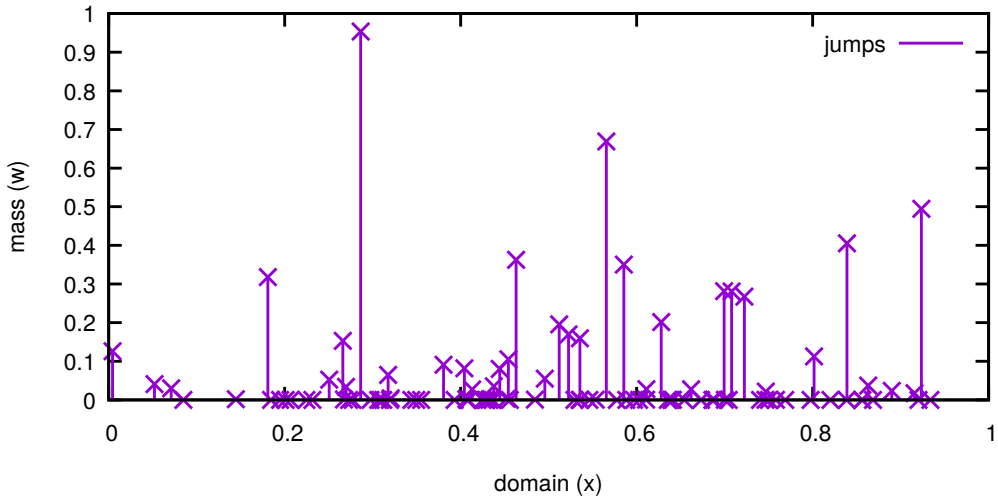
2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs

4 Stochastic Processes
- Poisson Point Processes
- The Gamma Process
- Constructing a Gamma Process from its Rate

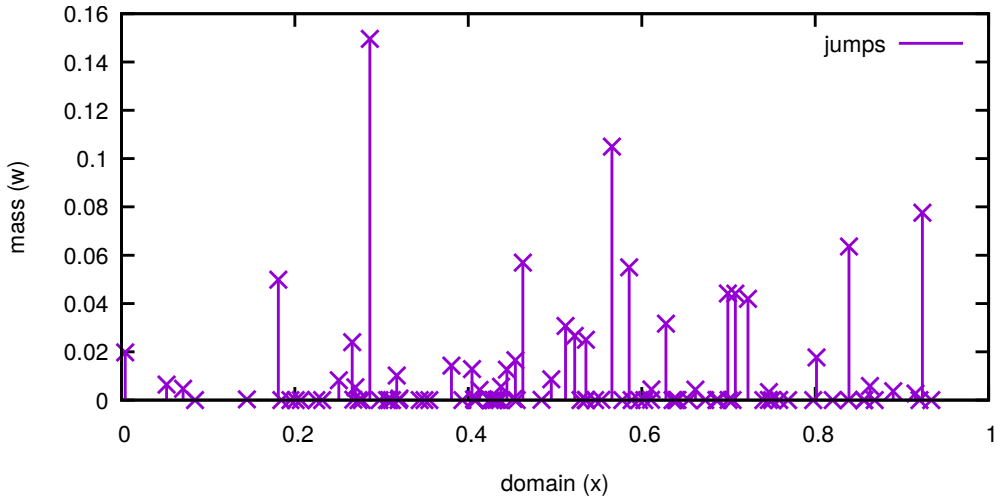
5 Inference on Hierarchies

Discrete Functions from a PPP



- Have a PPP on $\mathcal{R}^+ \times \mathcal{X}$.
- Interpret each point (w, x) as (mass,domain) pair.
- Sample set of points $N \sim$ PPP, as shown.
- Usually the rate factors: $\rho(w, x) = \rho_1(w)\rho_2(x)$, said to be homogeneous.

Discrete Functions from a PPP, cont.



- Sample $N \sim$ PPP as before.
- Define a function on $\mathcal{X}$ by $\mu(A) = \sum_{(w,x)\in N} w\delta_x(A)$.
- The function, now shown on the figure, must be discrete.
- Want total $\mu(\mathcal{X}) = \sum_{(w,x)\in N} w < \infty$.
- If there are an infinite number of points, they must be close to the 0-axis (mass $w = 0$) to get $\mu(\mathcal{X}) < \infty$.

A Gamma Process, Constructively

Constructive definition of gamma process

A domain $\mathcal{X}$ and a measure $m(x)$ for $x \in \mathcal{X}$ with finite integral ($\int_{\mathcal{X}} m(x)\mathrm{d}\,x < \infty$) define a gamma process$^{\mathcal{W}}$ with scale β, denoted $\mathrm{GP}(m(\cdot), \beta)$. The gamma process $H(\cdot)$ is a measure on $\mathcal{X}$ given, for $A \subseteq \mathcal{X}$,

$$H(A) = \sum_{k=1}^{\infty} w_k \delta_{x_k}(A)$$

where $\{(w_1, x_1), (w_2, x_2), \ldots\}$ is a sample from a PPP with rate $\rho(w, x) = w^{-1} e^{-\beta w} m(x)$.

- To prove this is equivalent to the axiomatic definition, we need to show $H(A) \sim \mathrm{gamma}(m(A), \beta)$.
- The "how" is given on the next slide.

Distribution for the Sum

In Lévy process$^{\mathcal{W}}$ theory, the distribution of the total

$$H(A) = \sum_{k=1}^{\infty} w_k \delta_{x_k}(A)$$

can be obtained from the rate $M\rho(w)m(x)$ by the Lévy-Khintchine formula (see Lévy process (Lévy-Khintchine representation)$^{\mathcal{W}}$).

- These distributions are always additive in the parameter M.
- This is intimately related to the property of infinite divisibility (Lévy process)$^{\mathcal{W}}$
- Used to map from constructive to axiomatic definitions and back.
  - The forward version computes the characteristic function (probability theory)$^{\mathcal{W}}$ of the distribution
  - The reverse is done similarly to inverting a Fourier transform from the characteristic function.

Historical Context for Stochastic Processes

See Lijoi and Prünster, "Models beyond the Dirichlet process" 2010.

1967-1993: Kingman establishes foundations of completely random measures and partition structures and the book (1993) on Poisson processes.

1995-2006: Pitman and colleagues develops the Pitman-Yor process based on subordinator theory, and develops various extensions and relationships.

2009-2016: James and colleagues establish general Bayesian results for analysis of CRMs and NRMIs, including variants of the the Indian buffet process.

Historical Context for Stochastic Processes

See Lijoi and Prünster, "Models beyond the Dirichlet process" 2010.

1967-1993: Kingman establishes foundations of completely random measures and partition structures and the book (1993) on Poisson processes.

1995-2006: Pitman and colleagues develops the Pitman-Yor process based on subordinator theory, and develops various extensions and relationships.

2009-2016: James and colleagues establish general Bayesian results for analysis of CRMs and NRMIs, including variants of the the Indian buffet process.

Note widely enough known in ML that hierarchical processes are usually equivalent to regular distributions, though its a standard result from subordinator theory.

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs

4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion

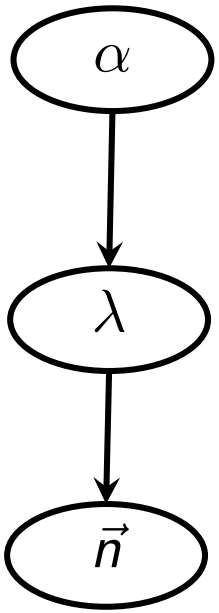
Hierarchical Processes Are Just Distributions

- Hierarchical GPs are just gamma distributions.
- Hierarchical DPs are just Dirichlet distributions.

Hierarchical Processes Are Just Distributions

- Hierarchical GPs are just gamma distributions.
- Hierarchical DPs are just Dirichlet distributions.
- How do we work with them then?

The Hierarchical Gamma



Context:

- α is a parent variable, $\lambda \sim \text{gamma}(\alpha, \beta)$, and data vector $\vec{n}$ has each entry $n_n \sim \text{Poisson}(\lambda)$ for $n = 1, ..., N$.
- How can our data about λ be used to influence α?

More generally, we have a whole vector of α's and we want to do this for a whole vector of λ's.

The Gamma Distribution

- The gamma distribution$^{\mathcal{W}}$ is $p(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}$ for $\lambda,\alpha,\beta \in \mathcal{R}^+$.
- Given data in the form of $(n_1,...,n_N)$, each distributed as $n_n \sim \text{Poisson}(\lambda)$.
- The marginal likelihood $p(n_1,...,n_N|\text{Poisson},\alpha,\beta)$ is

$$\frac{1}{\prod_{n=1}^N n_n!}\frac{\beta^\alpha}{\Gamma(\alpha)}\int \lambda^{\alpha+n_\cdot-1}e^{-(N+\beta)\lambda}\mathrm{d}\,\lambda = \frac{1}{\prod_{n=1}^N n_n!}\frac{\beta^\alpha\Gamma(\alpha+n_\cdot-1)}{(N+\beta)^{\alpha+n_\cdot}\Gamma(\alpha)}$$

where $n_\cdot = \sum_{n=1}^N n_n$.

The Gamma Distribution

- The gamma distribution[W] is $p(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}$ for $\lambda,\alpha,\beta \in \mathcal{R}^+$.
- Given data in the form of $(n_1,...,n_N)$, each distributed as $n_n \sim \text{Poisson}(\lambda)$.
- The marginal likelihood $p(n_1,...,n_N|\text{Poisson},\alpha,\beta)$ is

$$\frac{1}{\prod_{n=1}^N n_n!}\frac{\beta^\alpha}{\Gamma(\alpha)}\int \lambda^{\alpha+n_\cdot-1}e^{-(N+\beta)\lambda}\mathrm{d}\,\lambda = \frac{1}{\prod_{n=1}^N n_n!}\frac{\beta^\alpha\Gamma(\alpha+n_\cdot-1)}{(N+\beta)^{\alpha+n_\cdot}\Gamma(\alpha)}$$

  where $n_\cdot = \sum_{n=1}^N n_n$.
- When used hierarchically, this becomes the likelihood for α.

The Gamma Distribution

- The gamma distribution[W] is $p(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda\beta}$ for $\lambda,\alpha,\beta \in \mathcal{R}^+$.
- Given data in the form of $(n_1,...,n_N)$, each distributed as $n_n \sim \text{Poisson}(\lambda)$.
- The marginal likelihood $p(n_1,...,n_N|\text{Poisson},\alpha,\beta)$ is

$$\frac{1}{\prod_{n=1}^N n_n!}\frac{\beta^\alpha}{\Gamma(\alpha)}\int \lambda^{\alpha+n_\cdot-1}e^{-(N+\beta)\lambda}\mathrm{d}\,\lambda = \frac{1}{\prod_{n=1}^N n_n!}\frac{\beta^\alpha\Gamma(\alpha+n_\cdot-1)}{(N+\beta)^{\alpha+n_\cdot}\Gamma(\alpha)}$$

  where $n_\cdot = \sum_{n=1}^N n_n$.
- When used hierarchically, this becomes the likelihood for α.
- Too complex to easily use in a Gibbs sampler!

Augmenting the Marginal

- The likelihood terms in α are $\frac{\beta^\alpha}{(N+\beta)^{\alpha+n_\cdot}} \frac{\Gamma(\alpha+n_\cdot)}{\Gamma(\alpha)}$, too complex to use!
- We use the identity, for $n > 0$,

$$\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} = \sum_{t=1}^{n} S_{t,0}^{n} \alpha^t$$

(I challenge you to find this in the Wikipedia!)

- $S_{t,0}^{n}$ is an unsigned Stirling number of the first kind
  - a combinatoric quantity easily tabulated,
  - see Buntine and Hutter, 2012.

Augmenting the Marginal

- The likelihood terms in α are $\frac{\beta^\alpha}{(N+\beta)^{\alpha+n.}} \frac{\Gamma(\alpha+n.)}{\Gamma(\alpha)}$, too complex to use!
- We use the identity, for $n > 0$,

$$\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} = \sum_{t=1}^{n} S_{t,0}^{n} \alpha^t$$

(I challenge you to find this in the Wikipedia!)

- $S_{t,0}^{n}$ is an unsigned Stirling number of the first kind
  - a combinatoric quantity easily tabulated,
  - see Buntine and Hutter, 2012.
- The corresponding distribution is the distribution on the number of tables for a CRP,

$$p(t|CRP, n., d=0, \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n.)} S_{t,0}^{n.} \alpha^t$$

- Call this a Chinese restaurant distribution with parameters discount $d = 0$, concentration α and count $n.$

Augmenting the Marginal, cont.

- The original likelihood terms in α are $\frac{\beta^\alpha}{(N+\beta)^{\alpha+n.}} \frac{\Gamma(\alpha+n.)}{\Gamma(\alpha)}$.
- With augmentation, we get

$$\overbrace{\frac{\beta^\alpha}{(N+\beta)^{\alpha+n.}} \frac{\Gamma(\alpha+n.)}{\Gamma(\alpha)}}^{\text{likelihood}} \overbrace{\frac{\Gamma(\alpha)}{\Gamma(\alpha+n.)} S_{t,0}^{n.} \alpha^t}^{\text{augmentation}} \propto \left(\frac{\beta}{N+\beta}\right)^\alpha \alpha^t$$

- Thus the **marginal likelihood above** simplifies to a **Poisson likelihood**.

Augmented Reasoning for the Hierarchical Gamma

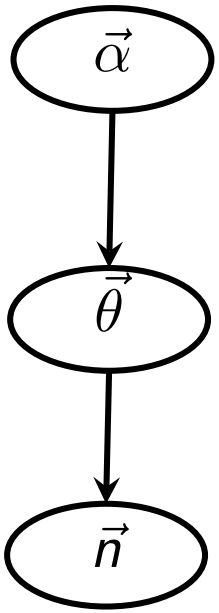
Given a gamma likelihood about λ, with N Poisson counts of total $n_\cdot$ sampled from it:

- Sample $t \sim \text{CRD}(0, \alpha, n_\cdot)$.
- Then a marginal likelihood (with λ marginalised out) for α has the form

$$p(\vec{n}|\alpha) \propto \left(\frac{\beta}{N+\beta}\right)^{\alpha} \alpha^{t}$$

- This is in turn a Poisson likelihood to be used in reasoning about α.

The Hierarchical Dirichlet



Context:

- $\vec{\alpha}$ is a K-dimensional parent probability vector, $\vec{\theta} \sim \text{Dirichlet}_K(\beta\vec{\alpha})$, and data vector $\vec{n} \sim \text{multinomial}(\vec{\theta}, N)$ for total count N.
- How can our data about $\vec{\theta}$ be used to influence $\vec{\alpha}$?
- Note in the DP case, $\vec{\alpha}$ is infinite, but we will get to that later.

NB. similar setup to the hierarchical gamma.

The Dirichlet Distribution

- The Dirichlet distribution[W] is $p(\vec{\theta}|K, \beta\vec{\alpha}) = \frac{\Gamma(\beta)}{\prod_k \Gamma(\beta\alpha_k)} \prod_k \theta_k^{\beta\alpha_k - 1}$ for $\beta \in \mathcal{R}^+$ and K-dim. probability vectors $\vec{\theta}$ and $\vec{\alpha}$.
- Given data in the form of K-dim counts $\vec{n}$, of total N distributed multinomial$(\vec{\theta}, N)$.
- The marginal likelihood $p(\vec{n}|\text{multinomial}, \vec{\alpha}, \beta, N)$ is

$$\binom{N}{\vec{n}} \frac{\Gamma(\beta)}{\prod_k \Gamma(\beta\alpha_k)} \int \prod_k \theta_k^{\beta\alpha_k + n_k - 1} \mathrm{d}\,\vec{\theta} = \binom{N}{\vec{n}} \frac{\Gamma(\beta) \prod_k \Gamma(n_k + \beta\alpha_k)}{\Gamma(N + \beta) \prod_k \Gamma(\beta\alpha_k)}$$

The Dirichlet Distribution

- The Dirichlet distribution[W] is $p(\vec{\theta}|K, \beta\vec{\alpha}) = \frac{\Gamma(\beta)}{\prod_k \Gamma(\beta\alpha_k)} \prod_k \theta_k^{\beta\alpha_k - 1}$ for $\beta \in \mathcal{R}^+$ and K-dim. probability vectors $\vec{\theta}$ and $\vec{\alpha}$.
- Given data in the form of K-dim counts $\vec{n}$, of total N distributed multinomial$(\vec{\theta}, N)$.
- The marginal likelihood $p(\vec{n}|\text{multinomial}, \vec{\alpha}, \beta, N)$ is

$$\binom{N}{\vec{n}} \frac{\Gamma(\beta)}{\prod_k \Gamma(\beta\alpha_k)} \int \prod_k \theta_k^{\beta\alpha_k + n_k - 1} \mathrm{d}\,\vec{\theta} = \binom{N}{\vec{n}} \frac{\Gamma(\beta) \prod_k \Gamma(n_k + \beta\alpha_k)}{\Gamma(N + \beta) \prod_k \Gamma(\beta\alpha_k)}$$

- When used hierarchically, this becomes the marginal likelihood for $\vec{\alpha}$.

The Dirichlet Distribution

- The Dirichlet distribution[W] is $p(\vec{\theta}|K, \beta\vec{\alpha}) = \frac{\Gamma(\beta)}{\prod_k \Gamma(\beta\alpha_k)} \prod_k \theta_k^{\beta\alpha_k - 1}$ for $\beta \in \mathcal{R}^+$ and K-dim. probability vectors $\vec{\theta}$ and $\vec{\alpha}$.
- Given data in the form of K-dim counts $\vec{n}$, of total N distributed multinomial$(\vec{\theta}, N)$.
- The marginal likelihood $p(\vec{n}|\text{multinomial}, \vec{\alpha}, \beta, N)$ is

$$\binom{N}{\vec{n}} \frac{\Gamma(\beta)}{\prod_k \Gamma(\beta\alpha_k)} \int \prod_k \theta_k^{\beta\alpha_k + n_k - 1} \mathrm{d}\,\vec{\theta} = \binom{N}{\vec{n}} \frac{\Gamma(\beta) \prod_k \Gamma(n_k + \beta\alpha_k)}{\Gamma(N + \beta) \prod_k \Gamma(\beta\alpha_k)}$$

- When used hierarchically, this becomes the marginal likelihood for $\vec{\alpha}$.
- As before too complex to use!

Augmenting the Marginal

- The likelihood terms in $\vec{\alpha}$ are $\frac{\prod_k \Gamma(n_k+\beta\alpha_k)}{\prod_k \Gamma(\beta\alpha_k)}$, as before too complex to use!
- Augment similarly to before with $t_k \sim \text{CRD}(0, \beta\alpha_k, n_k)$.
- Thus, the likelihood becomes $\beta^{\sum_k t_k} \prod_k \alpha_k^{t_k}$ which is a multinomial likelihood for $\vec{\alpha}$.
- Thus the **marginal likelihood above** for $\vec{\alpha}$ simplifies to a **multinomial likelihood**.

Augmenting the Marginal, cont.

- For β, there is an additional term $\frac{\Gamma(\beta)}{\Gamma(N+\beta)}$.
- We augment this using $q \sim \text{beta}(N, \beta)$.
- With augmentation, we get

$$\overbrace{\frac{\Gamma(\beta)}{\Gamma(N+\beta)}}^{\text{likelihood}} \overbrace{\frac{\Gamma(N+\beta)}{\Gamma(N)\Gamma(\beta)} q^{\beta-1}(1-q)^{N-1}}^{\text{augmentation}} \propto q^{\beta}$$

- Combining the two augmentations, the **marginal likelihood** for β is $q^{\beta}\beta^{\sum_k t_k}$ which is a **Poisson likelihood**.

Augmented Reasoning for the Hierarchical Dirichlet

Given a Dirichlet likelihood about $\vec{\theta}$, and K-dim multinomial data $\vec{n}$ of total N sampled from it:

- Sample $t_k \sim \text{CRD}(0, \beta\alpha_k, n_k)$ for $k = 1, ..., K$.
- Sample $q \sim \text{beta}(N, \beta)$.
- Then a marginal likelihood (with θ marginalised out) for $\vec{\alpha}$ and β has the form

$$p(\vec{n}|\vec{\alpha}, \beta, N) \propto q^{\beta} \beta^{\sum_k t_k} \prod_k \alpha_k^{t_k}$$

- This is in turn a Poisson likelihood to be used in reasoning about β and a multinomial likelihood for $\vec{\alpha}$.
- In the DP context, $\vec{\alpha}$ is infinite, but only K-dimensions have data observed, so same tricks apply.

And So On ...

- Thus
  - **hierarchies** of gamma processes and Dirichlet processes,
  - **which are really** hierarchies of gamma distributions and Dirichlet distributions,
  - **can be augmented** so one gets regular Poisson likelihoods and multinomial likelihoods from child to parent.

And So On ...

- Thus
  - **hierarchies** of gamma processes and Dirichlet processes,
  - **which are really** hierarchies of gamma distributions and Dirichlet distributions,
  - **can be augmented** so one gets regular Poisson likelihoods and multinomial likelihoods from child to parent.
- There is a rich set of literature and methods for doing this:
  - Mingyuan Zhou (UT Austin) et al.
  - Buntine, Du et al.

And So On ...

- Thus
  - **hierarchies** of gamma processes and Dirichlet processes,
  - **which are really** hierarchies of gamma distributions and Dirichlet distributions,
  - **can be augmented** so one gets regular Poisson likelihoods and multinomial likelihoods from child to parent.
- There is a rich set of literature and methods for doing this:
  - Mingyuan Zhou (UT Austin) et al.
  - Buntine, Du et al.
- The PYP has a slightly more complex augmentation,
  - see Buntine and Hutter, 2012,
  - corresponds to the generalised gamma process,
  - allows Zipfian probabilities to be modelled.

And So On ...

- Thus
  - **hierarchies** of gamma processes and Dirichlet processes,
  - **which are really** hierarchies of gamma distributions and Dirichlet distributions,
  - **can be augmented** so one gets regular Poisson likelihoods and multinomial likelihoods from child to parent.
- There is a rich set of literature and methods for doing this:
  - Mingyuan Zhou (UT Austin) et al.
  - Buntine, Du et al.
- The PYP has a slightly more complex augmentation,
  - see Buntine and Hutter, 2012,
  - corresponds to the generalised gamma process,
  - allows Zipfian probabilities to be modelled.
- The DP and PYP case corresponds to a collapsed, hierarchical CRP (Buntine and Hutter, 2012).

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs

4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion
- Other Issues
- Chinese Restaurant Processes

Questions?

Questions?

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

3 Definitions of DPs and PYPs

4 Stochastic Processes

5 Inference on Hierarchies

6 Conclusion
- Other Issues
- Chinese Restaurant Processes

Distribution for the Sum: Details

In Lévy process$^{\mathcal{W}}$ theory, the distribution of the total $W = \sum_{k=1}^{\infty} w_k$ is got via the Lévy-Khintchine formula.

Lemma on distribution for the sum

For the PPP with rate $M\rho(w)m(x)$ for $(w, x) \in \mathcal{R}^+ \times \mathcal{X}$ where $m(x)$ is a probability measure, let $W = \sum_{k=1}^{\infty} w_k$ be the total mass for a sample from the PPP. Then the moment generating function of W, if it exists, is given by $e^{-M\phi(-t)}$, where

$$\phi(t) = \int_{\mathcal{R}^+} \left(1 - e^{-tw}\right) \rho(w) \mathrm{d}\, w \ .$$

- These distributions are always additive in the parameter M.
- Requires some tricks (find $\mathrm{d}\,\phi(t)/\mathrm{d}\,t$ first) to evaluate the integral.
- Used to map from constructive to axiomatic definitions.
- The reverse is done similarly to inverting a Laplace transform.

Theory — Size-Biased Sampling (Pitman)

Definition of Size-biased Sampling

Assume we have an infinite vector from a PPP with rate $\rho(\lambda)$, and the total is $\Lambda = \sum_{k=1}^{\infty} \lambda_k$. Given a sample $\vec{\lambda}$ generated by the PPP, sampling λ_k proportionally to $\frac{\lambda_k}{\sum_{k=1}^{\infty} \lambda_k}$ is called size-biased sampling. i.e.,

1. Sample j_1 according to probability vector $\frac{1}{\sum_{k=1}^{\infty} \lambda_k} \vec{\lambda}$, and use $\lambda_1^* = \lambda_{j_1}$.
2. Sample j_2 according to probability vector $\frac{1}{\sum_{k=1}^{\infty} \lambda_k} \vec{\lambda}$, but don't allow j_1, and use $\lambda_2^* = \lambda_{j_2}$.
3. Sample j_3 ... but don't allow j_1, j_2, and use $\lambda_3^* = \lambda_{j_3}$.
4. *etc.*

- size-biased sampling builds a size-biased order
- a GEM generates $\vec{p}$ so it follows a size-biased order
- alternatively generate a sample from an SSM and order the λ_k in terms of first occurence of k-th atom in the sample

Outline

1 Motivation and Context

2 Basic Properties of (non-hierarchical) DPs and PYPs

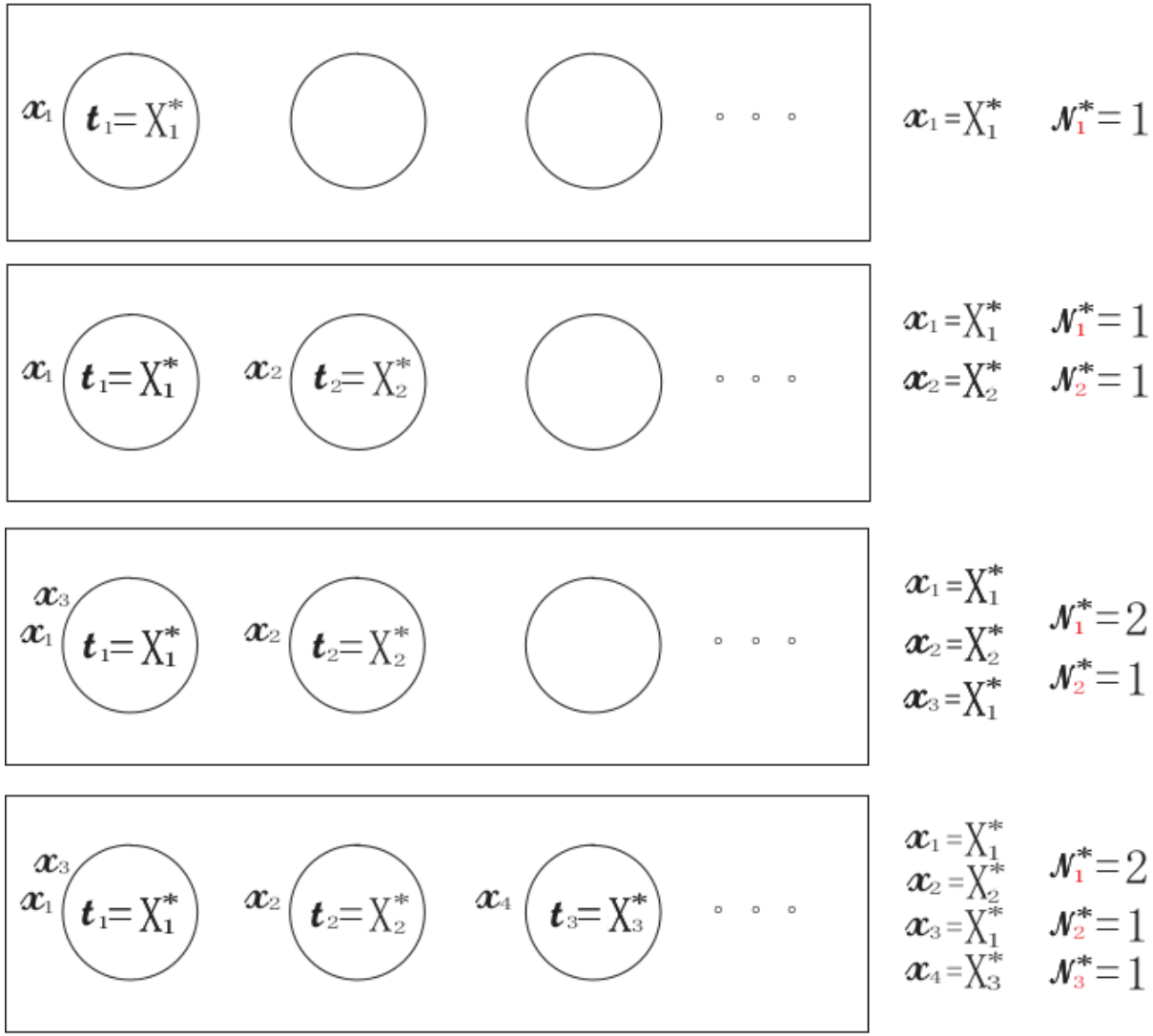
3 Definitions of DPs and PYPs

4 Stochastic Processes

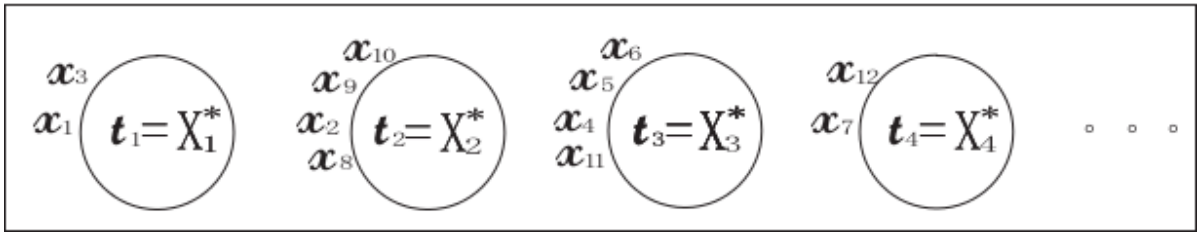
5 Inference on Hierarchies

6 Conclusion
- Other Issues
- Chinese Restaurant Processes

"Chinese Restaurant" Sampling



Chinese Restaurant Process (CRP)

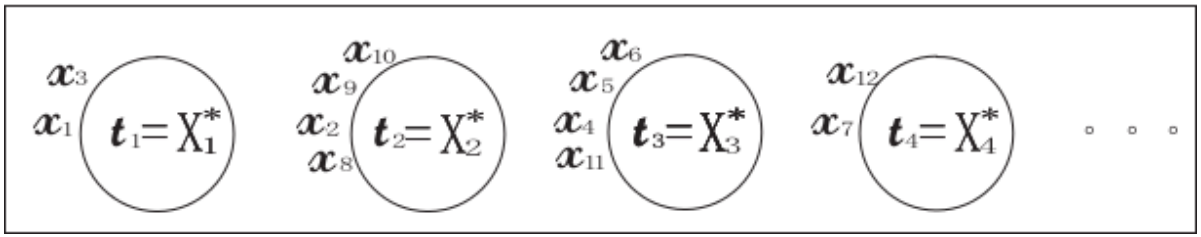


Definition of a Chinese restaurant process

A sample from a Chinese restaurant process$^{\mathcal{W}}$ (CRP) with parameters $(d, \alpha, H(\cdot))$ is generated as follows (we use the two-parameter version):

1. First customer enters the restaurant, sits at the first empty table and picks a dish according to $H(\cdot)$.
2. Subsequent customer numbered $N+1$:
   1. with probability $\frac{\alpha + K\,d}{N+\alpha}$ (K is count of existing tables) sits at an empty table and picks a dish according to $H(\cdot)$;
   2. otherwise, joins an existing table k with probabiliity proportional to $\frac{n_k - d}{N+\alpha}$ (n_k is the count of existing customers at the table) and has the dish served at that table.

CRP Terminology



Restaurant: single instance of a CRP, roughly like a Dirichlet-multinomial distribution.

Customer: one data point.

Table: cluster of data points sharing the one sample from $H(.)$.

Dish: the data value corresponding to a particular table; all customers at the table have this "dish".

Table count: number of customers at the table.

Seating plan: full configuration of tables, dishes and customers.

Historical Context

1990s: Pitman and colleagues in mathematical statistics develop statistical theory of partitions, Pitman-Yor process, *etc.*

2006: Teh develops hierarchical n-gram models using PYs.

2006: Teh, Jordan, Beal and Blei develop hierarchical Dirichlet processes, *e.g.* applied to LDA.

2006-2011: Chinese restaurant processes (CRPs) go wild!

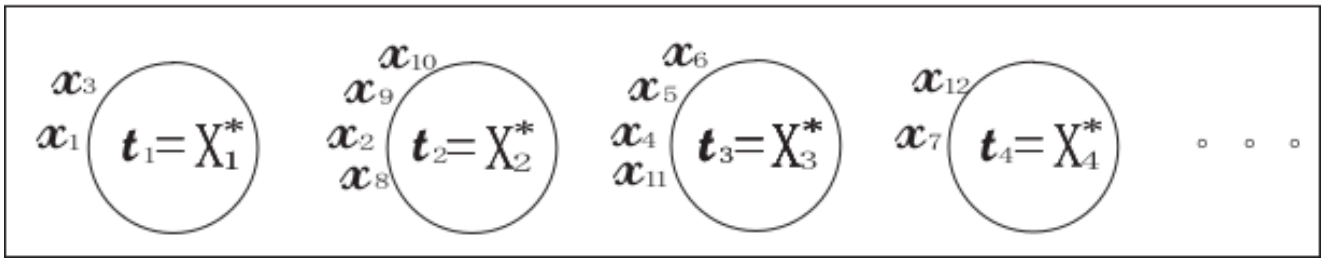
- require dynamic memory in implementation,
- Chinese restaurant franchise,
- multi-floor Chinese restaurant process (Wood and Teh, 2009),
- *etc.*

Opened up whole field of non-parametrics, but generally regarded as slow and unrealistic for scaling up

CRP Example

CRP with base distribution $H(\cdot)$:

$$p(x_{N+1} \mid x_{1:N}, d, \alpha, H(\cdot)) = \frac{\alpha + Kd}{N+\alpha} H(x_{N+1}) + \sum_{k=1}^{K} \frac{n_k - d}{N+\alpha} \delta_{X_k^*}(x_{N+1}) ,$$



$$p(x_{13} = X_1^* \mid x_{1:12}, \ldots) = \frac{2-d}{12+\alpha} \qquad p(x_{13} = X_2^* \mid x_{1:12}, \ldots) = \frac{4-d}{12+\alpha}$$

$$p(x_{13} = X_3^* \mid x_{1:12}, \ldots) = \frac{4-d}{12+\alpha} \qquad p(x_{13} = X_4^* \mid x_{1:12}, \ldots) = \frac{2-d}{12+\alpha}$$

$$p(x_{13} = X_5^* \mid x_{1:12}, \ldots) = \frac{\alpha + 4d}{12+\alpha} H(X_5^*)$$

CRP, cont.

$$p(x_{N+1}|x_1,...,x_N,d,\alpha,H(\cdot)) = \frac{K\,d+\alpha}{N+\alpha}H(x_{N+1}) + \sum_{k=1}^{K}\frac{n_k-d}{N+\alpha}\delta_{X_k^*}(x_{N+1})$$

- is like Laplace sampling, but with a discount $(-d)$ instead of $1/2$ offset
- doesn't define a vector $\vec{p}$; the CRP is exchangable, de Finetti's theorem on exchangable observations proves a vector $\vec{p}$ must exist
- exactly the same process as we saw in posterior sampling with the improper Dirichlet

Evidence for the CRP

It does show how to compute evidence only when $H(\cdot)$ is non-discrete.

$$
\begin{aligned}
& p(x_1, ..., x_N | N, CRP, d, \alpha, H(\cdot)) \\
&= \frac{\alpha(d+\alpha)...((K-1)d+\alpha)}{\alpha(1+\alpha)...(N-1+\alpha)} \prod_{k=1}^{K} ((1-d)...(n_k-1-d)H(X_k^*)) \\
&= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(X_k^*) \ ,
\end{aligned}
$$

where there are K distinct data values X_1^*, ...X_K^*.

Evidence for the CRP

It does show how to compute evidence only when $H(\cdot)$ is non-discrete.

$$
\begin{aligned}
& p(x_1, ..., x_N | N, CRP, d, \alpha, H(\cdot)) \\
&= \frac{\alpha(d+\alpha)...((K-1)d+\alpha)}{\alpha(1+\alpha)...(N-1+\alpha)} \prod_{k=1}^{K} ((1-d)...(n_k-1-d)H(X_k^*)) \\
&= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(X_k^*) \ ,
\end{aligned}
$$

where there are K distinct data values X_1^*, ...X_K^*.

Same as the evidence for the DP.

Evidence for the CRP

It does show how to compute evidence only when $H(\cdot)$ is non-discrete.

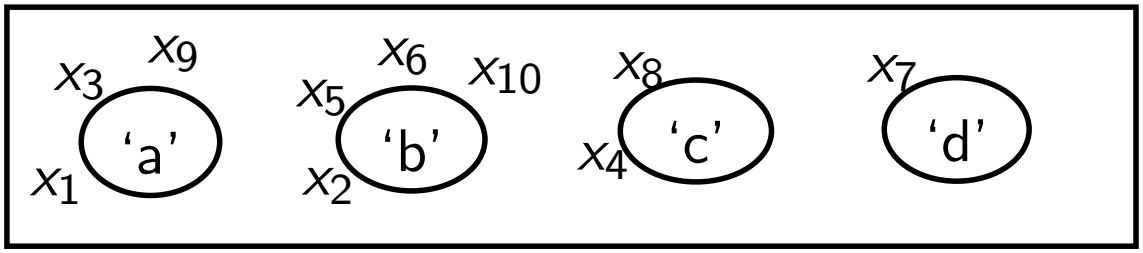
$$
\begin{aligned}
& p(x_1, ..., x_N | N, CRP, d, \alpha, H(\cdot)) \\
&= \frac{\alpha(d+\alpha)...((K-1)d+\alpha)}{\alpha(1+\alpha)...(N-1+\alpha)} \prod_{k=1}^{K} ((1-d)...(n_k-1-d)H(X_k^*)) \\
&= \frac{(\alpha|d)_K}{(\alpha)_N} \prod_{k=1}^{K} (1-d)_{n_k-1} H(X_k^*) \ ,
\end{aligned}
$$

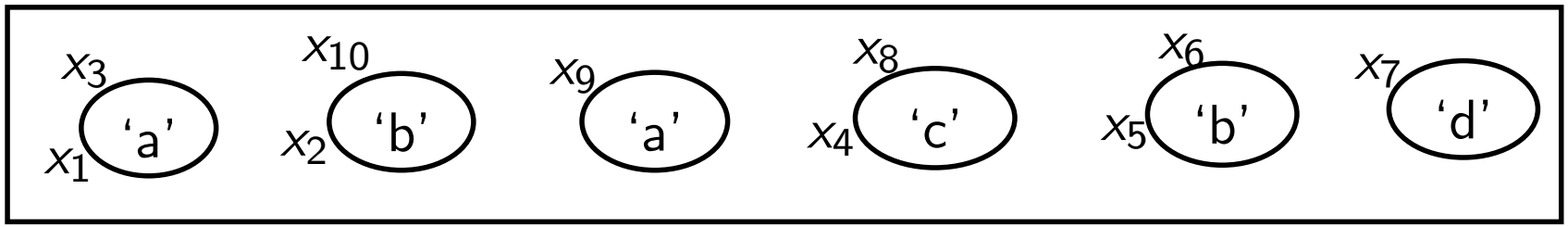
where there are K distinct data values X_1^*, ...X_K^*.

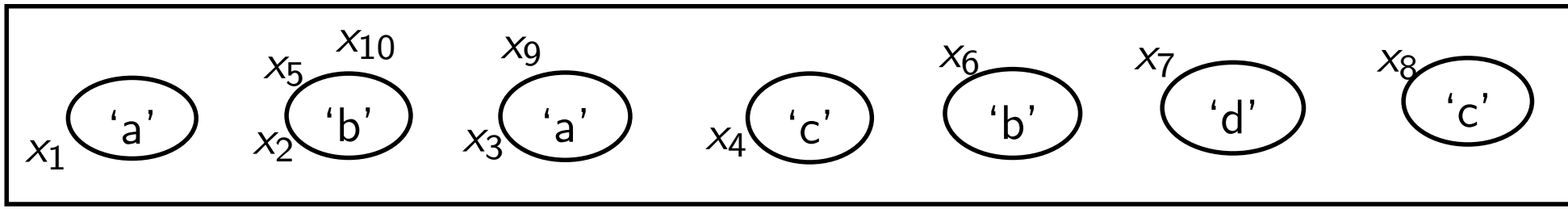
Same as the evidence for the DP.

Hence by the de Finetti-Hewitt-Savage Theorem, from the definition of the CRP, we can say there must exist an underlying generative distribution/measure, i.e., the DP itself.

CRPs on Discrete Data







The above three table configurations all match the data stream:

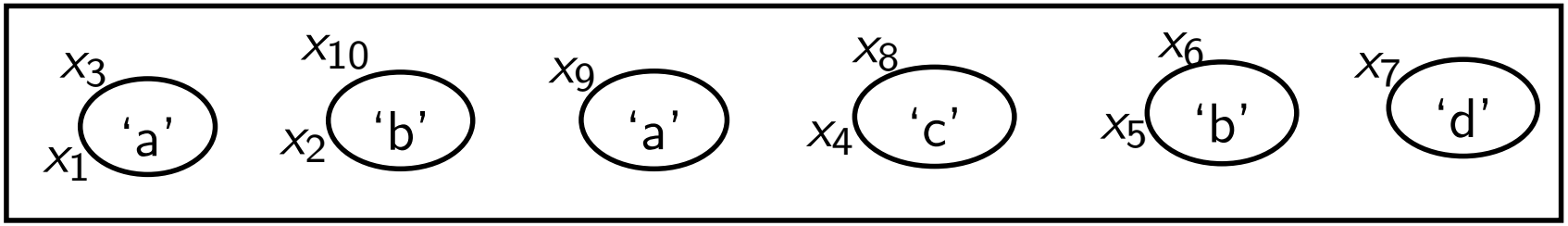
$a, b, a, c, b, b, d, c, a, b$

CRPs on Discrete Data, cont.

Different configurations for the data stream:

a, b, a, c, b, b, d, c, a, b

- $n_a = 3$, $n_b = 4$, $n_c = 2$, $n_d = 1$
- So the 3 data points with 'a' could be spread over 1,2 or 3 tables!
- Thus, inference will need to know the particular configuration/assignment of data points to tables.



The configuration here has: number of tables $t_a = 2$ with table counts $(2, 1)$, $t_b = 2$ with counts $(2, 2)$, $t_c = 1$ with counts (2) and $t_d = 1$ with counts (1).

Evidence of PYP for Probability Vectors

Notation: Tables numbered $t = 1, ..., T$. Data types numbered $k = 1, ..., K$. Full table configuration denoted $\mathcal{T}$. Count of data type k is n_k, and number of tables t_k. Table t has data value (dish) X_t^* with m_t customers.

$$n_k = \sum_{t\,:\,X_t^*=k} m_t \ , \qquad t_k = \sum_{t\,:\,X_t^*=k} 1 \ , \qquad N = \sum_{k=1}^{K} n_k \ , \qquad T = \sum_{k=1}^{K} t_k$$

with constraints $t_k \leq n_k$ and $n_k > 0 \rightarrow t_k > 0$.

Evidence:

$$\begin{aligned} p(\vec{x}, \mathcal{T} | N, \text{PYP}, d, \alpha) &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{t=1}^{T} ((1-d)_{m_t-1} H(X_t^*)) \\ &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \left(\prod_{t\,:\,X_t^*=k} (1-d)_{m_t-1} \right) H(k)^{t_k} \end{aligned}$$

Evidence of PYP for Probability Vectors, cont.

Consider configurations of tables with data type k, and denote them by $\mathcal{T}_k$.

$$\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2 + \ldots + \mathcal{T}_K$$

Let the function tables$(\cdot)$ return the number of tables in a configuration.

$$\begin{aligned} & p\left(\vec{x}, \vec{t} \,\middle|\, N, \text{PYP}, d, \alpha\right) \\ &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \sum_{\text{tables}(\mathcal{T}_k)=t_k} \left(\prod_{t\,:\,X_t^*=k} (1-d)_{m_t-1} \right) H(k)^{t_k} \\ &= \frac{(\alpha|d)_T}{(\alpha)_N} \prod_{k=1}^{K} \mathcal{S}^{n_k}_{t_k,d} H(k)^{t_k} \end{aligned}$$

The simplification uses the definition of a particular second order Stirling number $\mathcal{S}^n_{t,d}$ (see Buntine and Hutter, 2012).