Bayesian Nonparametric Priors and their Clusterings

Florian Schimmel

Institut für Mathematik Universität Heidelberg

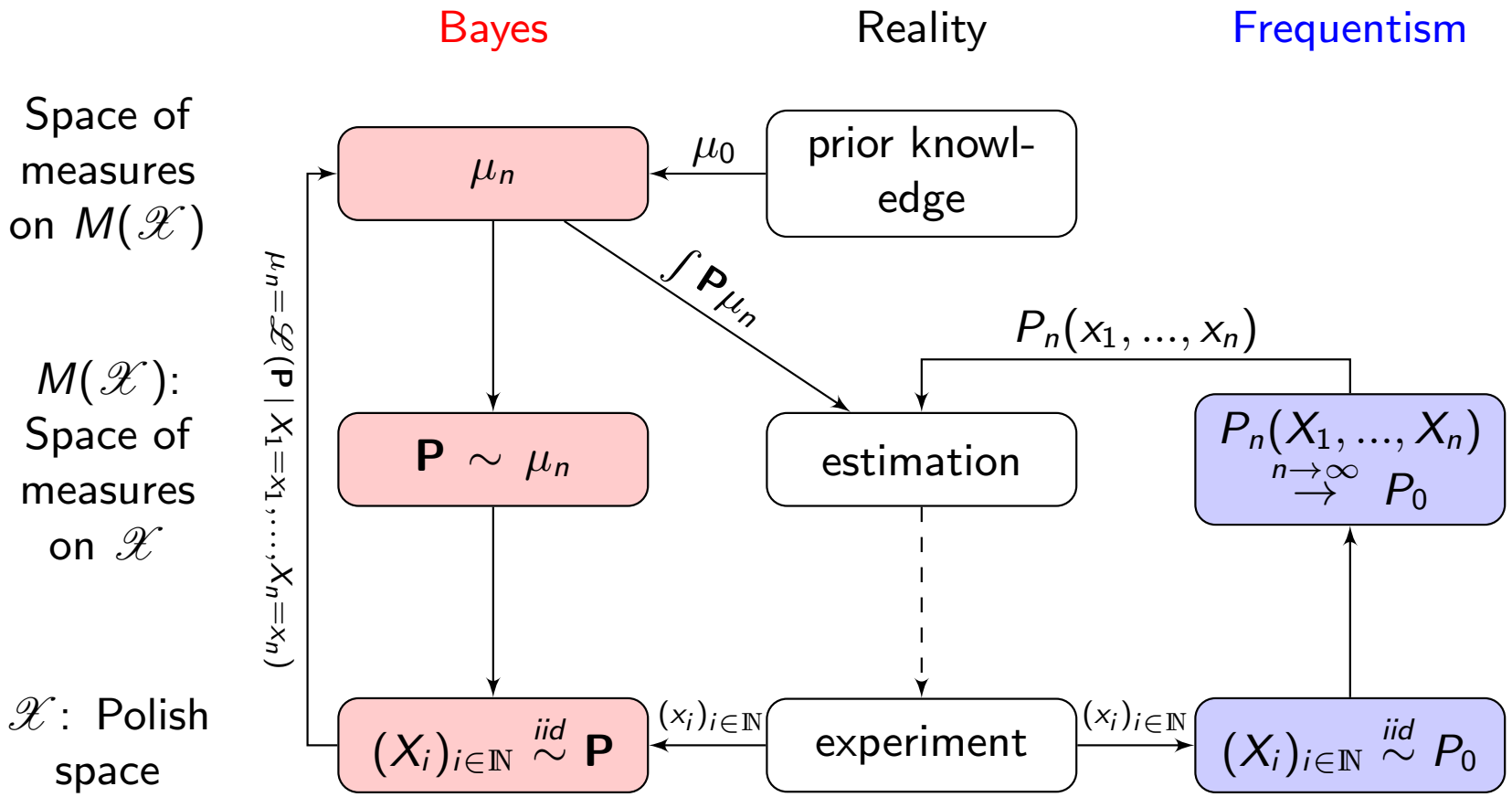
April 5, 2017

Overview

1. Bayesian Statistics
   - basic concepts
   - definitions and theorems
2. Dirichlet Process
   - construction
   - properties
   - clustering
3. Pitman-Yor process
   - construction
   - properties
   - clustering
   - relationship to Dirichlet process
4. Application
   - language modeling

1. Bayes Statistics

Bayes Statistics



Basics and Conventions

- $(\mathscr{X}, \mathscr{B})$ measurable Polish space equipped with Borel σ-algebra
- $(M(\mathscr{X}), \mathscr{B}_M)$ measurable Polish space of probability measures on $(\mathscr{X}, \mathscr{B})$
- weak convergence $\mu_n \xrightarrow{w} \mu / \underset{n\to\infty}{w\text{-lim}}\ \mu_n = \mu$ if for all bounded continuous functions f

$$\lim_{n\to\infty} \int f \, d\mu_n = \int f \, d\mu$$

- conditional independent and identically distributed sequences

$$(X_n)_{n\in\mathbb{N}} \mid \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$$

for $P \in M(X)$ and a random variable $\mathbf{P}$ with values in $M(\mathscr{X})$

Exchangeability and De Finetti

Definition 1

A sequence $(X_i)_{i\in I}$ is exchangeable, if for every finite $J \subseteq I$, $(B_j)_{j\in J}$ in $\mathscr{B}$ and permutation σ of J, it holds

$$\mathbb{P}[X_j \in B_j, j \in J] = \mathbb{P}[X_{\sigma(j)} \in B_j, j \in J].$$

- random variables are exchangeable if distribution is independent of order

Theorem 2 (De Finetti)

A sequence $(X_n)_{n\in\mathbb{N}}$ is exchangeable if and only if there is a measure ρ, such that for $\mathbf{P} \sim \rho$ it holds

$$(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$$

for $P \in M(\mathscr{X})$. Furthermore this ρ is almost surly unique and

$$\underset{n\to\infty}{w\text{-}lim}\ \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} = \rho \qquad \mathbb{P}\text{-a.s.}$$

- independence of distribution and order is equivalent to conditional i.i.d.

2. Dirichlet Process

Dirichlet Distribution

Definition 3

For $\alpha_1, ..., \alpha_n > 0$ a random variable $\mathbf{p}$ with values in the simplex

$$\Delta_n := \left\{ (p_1, ..., p_n) \in [0,1]^n \,\middle|\, \sum_{i=1}^{n} p_i = 1 \right\}$$

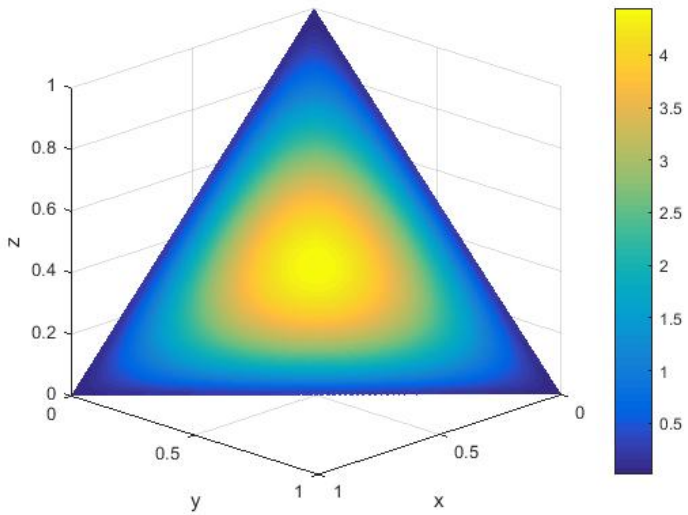
is Dirichlet distributed $\mathbf{p} \sim Dir_{(\alpha_1,...,\alpha_n)}$ with parameter $(\alpha_1, ..., \alpha_n)$, if its density is given by

$$\begin{aligned} f(p_1, ..., p_n) &= \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\Gamma(\alpha_1) ... \Gamma(\alpha_n)} p_1^{\alpha_1 - 1} ... p_n^{\alpha_n - 1} \\ &= \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\Gamma(\alpha_1) ... \Gamma(\alpha_n)} p_1^{\alpha_1 - 1} ... p_{n-1}^{\alpha_{n-1} - 1} \left(1 - \sum_{i=1}^{n-1} p_i\right)^{\alpha_n - 1}. \end{aligned}$$

- Δ_n is set of probability measures on a finite set
- In case $\alpha_i = 0$ for some parameters, the respective values will be set $p_i = 0$ almost surly

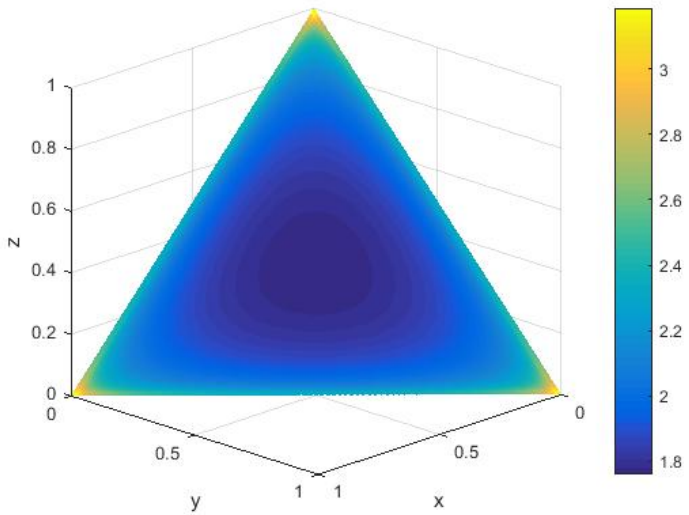
Values of the Dirichlet Distribution

$\alpha = (2, 2, 2)$



- symmetric
- high values in center

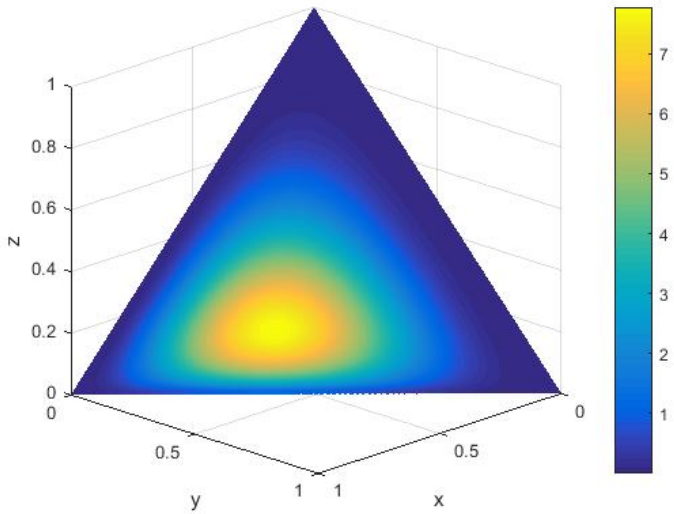
$\alpha = (0.9, 0.9, 0.9)$



- symmetric
- high values at the margin

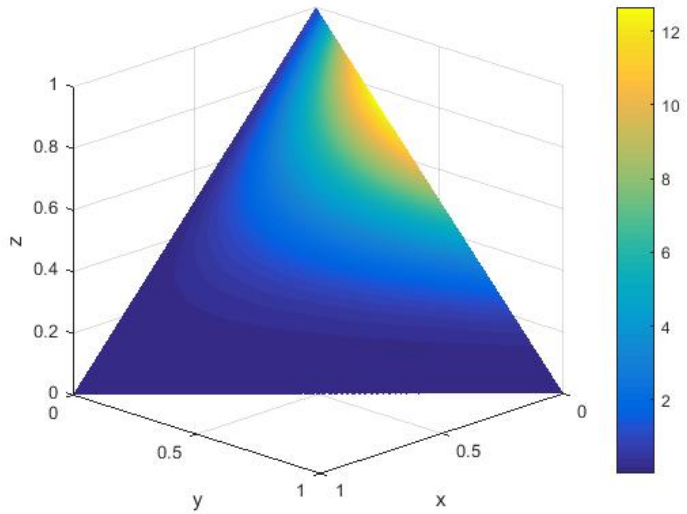
Values of the Dirichlet Distribution

$\boldsymbol{\alpha} = (4, 3, 2)$



- higher values at larger x and y
- shift to x

$\boldsymbol{\alpha} = (1, 2, 4)$



- high values at large z
- shift to y

Properties of the Dirichlet Distribution

Theorem 4

For $(\mathbf{p}_1, ..., \mathbf{p}_n) \sim Dir_{(\alpha_1,...,\alpha_n)}$ *it holds*

$$(\mathbf{p}_1, ..., \mathbf{p}_i + \mathbf{p}_j, ..., \mathbf{p}_{j-1}, \mathbf{p}_{j+1}, ..., \mathbf{p}_n) \sim Dir_{(\alpha_1,...,\alpha_i+\alpha_j,...,\alpha_{j-1},\alpha_{j+1},...,\alpha_n)}$$

- let $(\mathbf{p}_1, ..., \mathbf{p}_n) \sim Dir_{(\alpha_1,...,\alpha_n)}$
- finite set $D_n = \{d_1, ..., d_n\}$
- define random measure $\mathbf{p}$ by $\mathbf{p}(d_i) := \mathbf{p}_i$
- define measure α by $\alpha(d_i) := \alpha_i$
- for a partition $B_1, ..., B_m$ of D_n it holds

$$(\mathbf{p}(B_1), ..., \mathbf{p}(B_m)) \sim Dir_{(\alpha(B_1),...,\alpha(B_m))}$$

Dirichlet Process

Definition 5

Let $\mathbb{H}$ be an probability measure on a measurable Polish space $(\mathscr{X}, \mathscr{B})$ and $\theta > 0$. Then the Dirichlet process $D_{(\theta,\mathbb{H})}$ with parameter $(\theta, \mathbb{H})$ is uniquely defined by

$$(\mathbf{P}(B_1), ..., \mathbf{P}(B_n)) \sim Dir_{(\theta\mathbb{H}(B_1),..,\theta\mathbb{H}(B_n))},$$

for every measurable Partition $B_1, ..., B_n$ and $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$.

- generalization of Dirichlet distribution
- often only a measure as parameter with θ encoded in the whole mass

Stick-Breaking Construction

Theorem 6

Let $(V_n)_{n\in\mathbb{N}} \sim \mathbb{H}^{\otimes\mathbb{N}}$ *and* $(\beta_i)_{i\in\mathbb{N}} \sim Beta(1,\theta)^{\otimes\mathbb{N}}$ *all independent. Define*

$$W_k := \beta_k \prod_{i=1}^{k-1}(1-\beta_i)$$

then

$$\sum_{k=1}^{\infty} W_k \delta_{V_k} =: \mathbf{P} \sim D_{(\theta,\mathbb{H})}.$$

- developed by Sethuraman
- discrete measure
- common analogy
  - start with stick of length 1
  - in each step break of β_i from the remaining stick to get W_i as length of broken off part

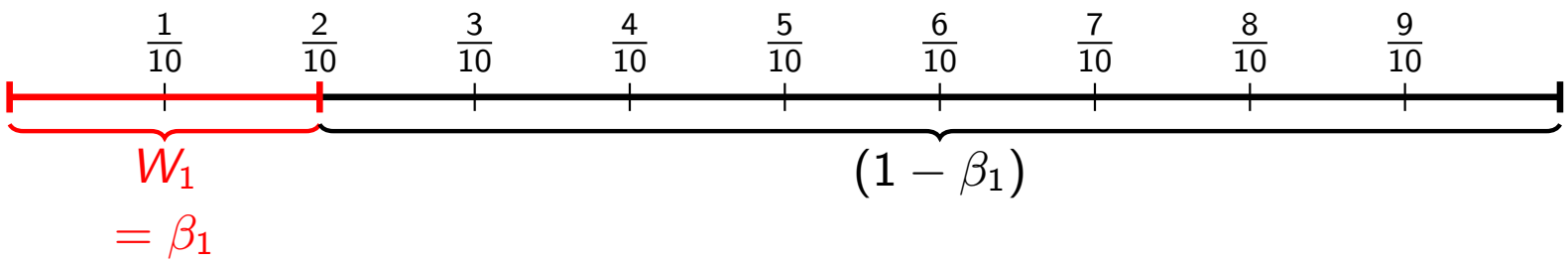
Stick-breaking construction

- start with stick of length 1



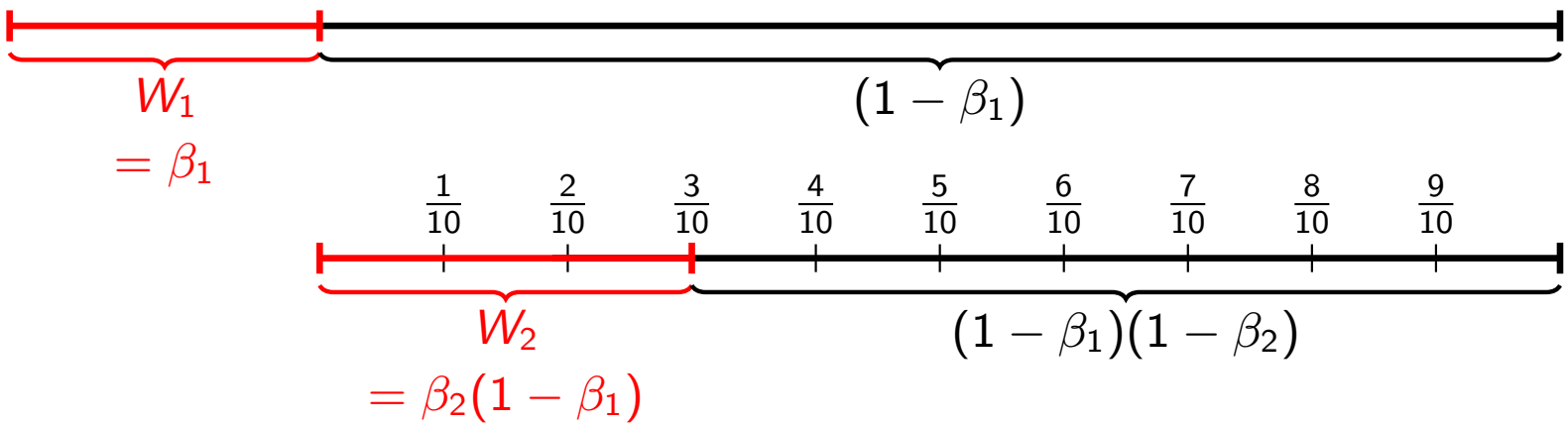
Stick-breaking construction

- break of $\beta_1 = 0.2$ from the stick to get W_1



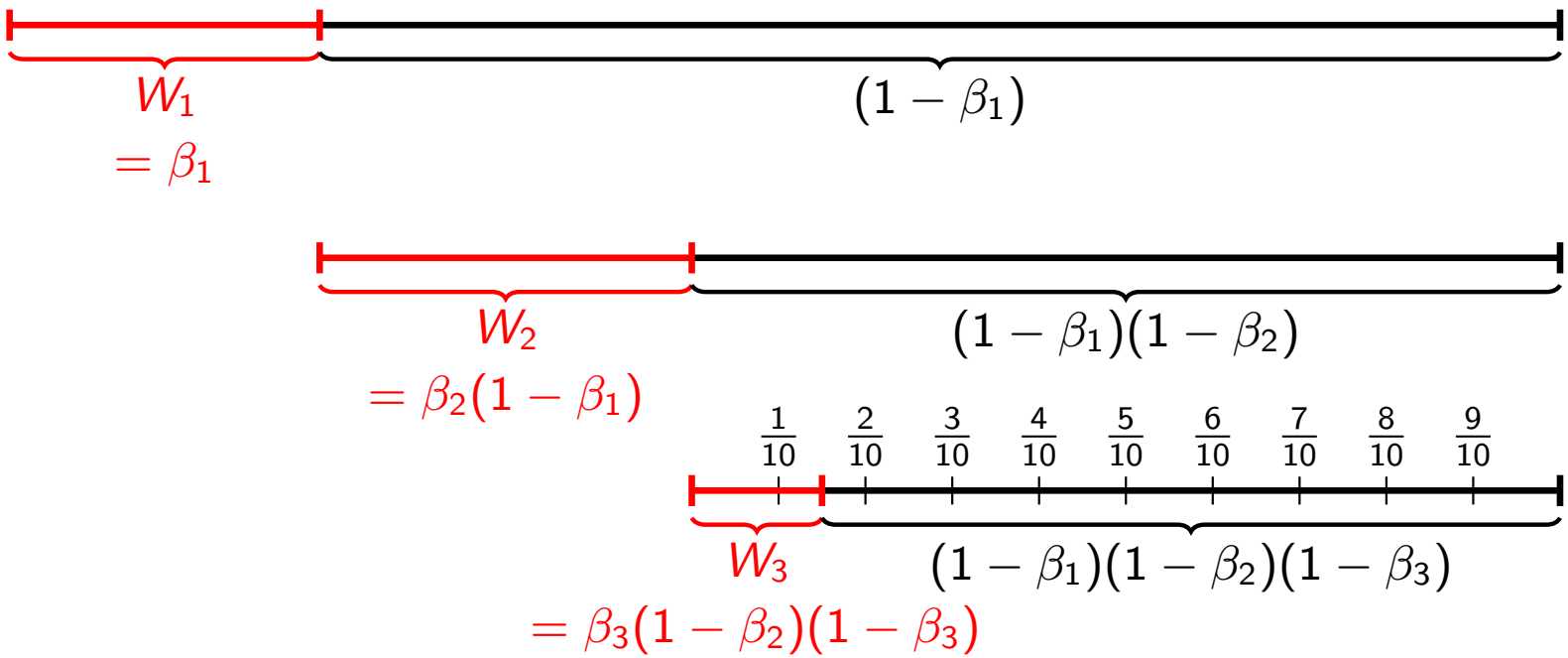
Stick-breaking construction

- break of $\beta_2 = 0.3$ from the remaining stick to get W_2



Stick-breaking construction

- break of $\beta_3 = 0.15$ from the remaining stick to get W_3



Marginal Distribution and Conjugacy

Theorem 7

Let $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ *and* $X \mid \mathbf{P} = P \sim P$ *for* $P \in (M(\mathscr{X}), \mathscr{B}_M)$, *then* $X \sim \mathbb{H}$.

- $\mathbb{H}$ should reflect prior knowledge

Theorem 8

Let $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ *and* $(X_n)_{n\in\mathbb{N}} \mid \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$ *with* $P \in M(\mathscr{X})$, *then*

$$D_{\left(\theta+n,\ \frac{1}{\theta+k}\left(\theta\mathbb{H}+\sum_{i=1}^{k}\delta_{x_i}\right)\right)}$$

is the posterior distribution for given $X_1 = x_1, \ldots, X_k = x_k$.

- Dirichlet process is conjugate
- each observation gives weight to its value
- influence of base measure $\mathbb{H}$ can be adjusted by θ

Posterior Predictive

Theorem 9

Let $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ *and* $(X_i)_{i\in\mathbb{N}} \mid \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$, *then*

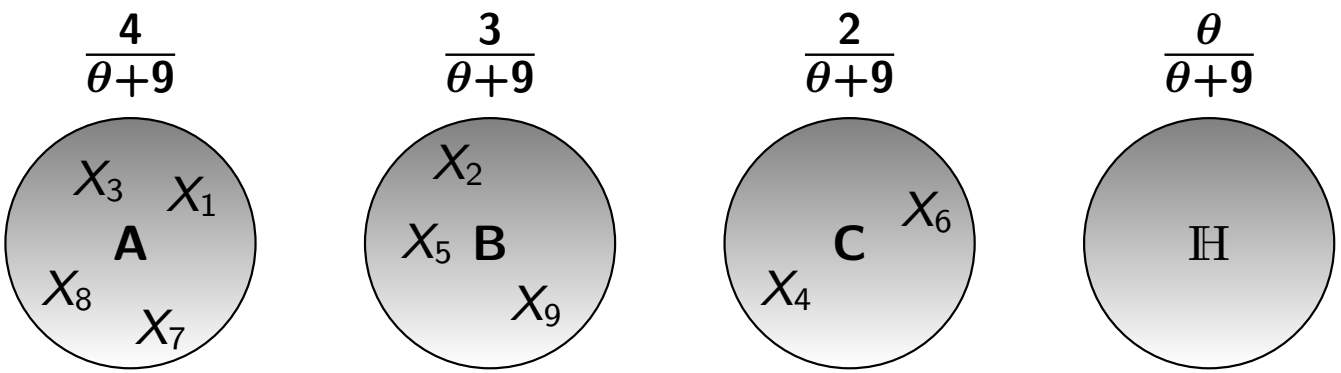
$$X_{n+1} \mid X_1, ..., X_n \sim \frac{1}{\theta + n}\left(\theta\mathbb{H} + \sum_{j=1}^{M_n} K_j^n \delta_{Y_j}\right).$$

- M_n number of values up to the n-th observation
- Y_j unique values
- K_j^n number of appearances of Y_j up to the n-th observation
- proof follows directly by conjugacy and marginal distribution

Clustering Property

- $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

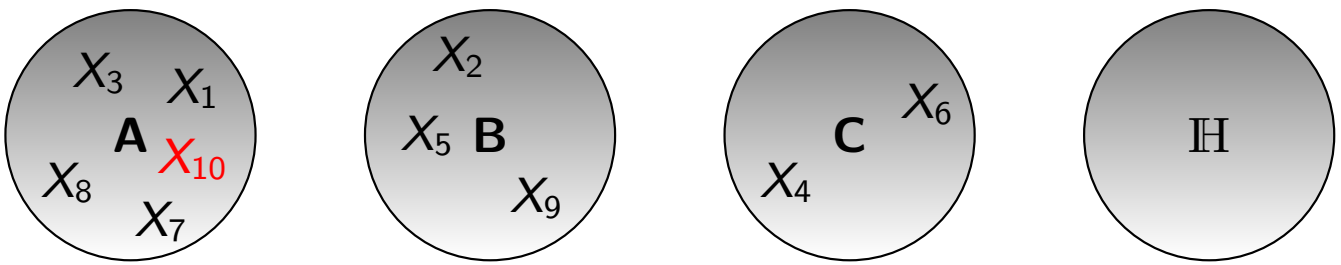
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	A	B	A	C	B	C	A	A	B	



Clustering Property

- $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

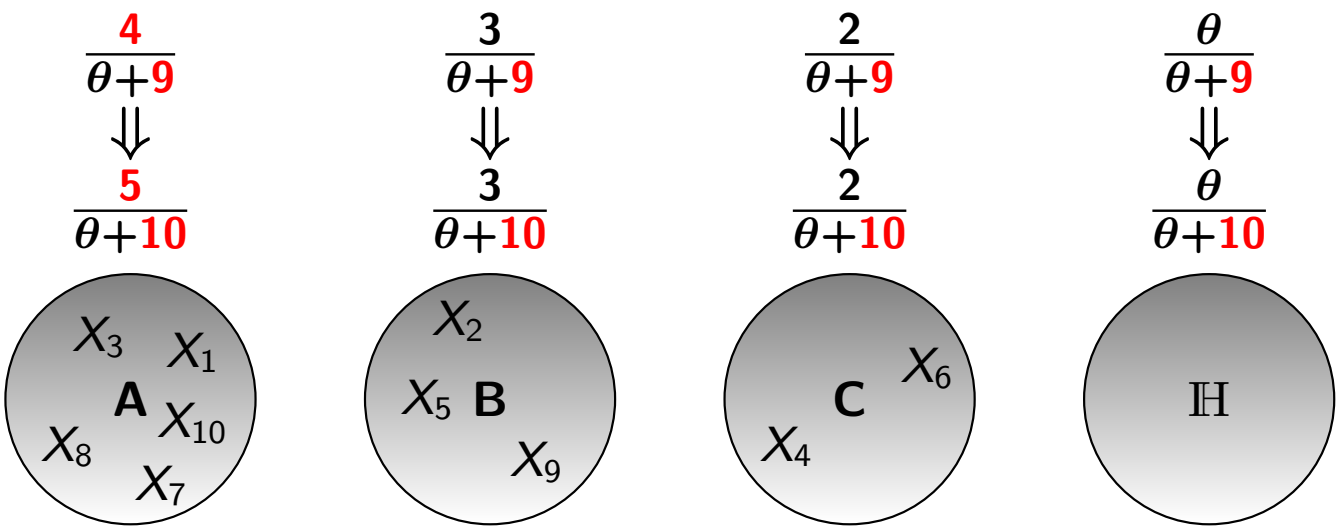
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	A	B	A	C	B	C	A	A	B	A



Clustering Property

- $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

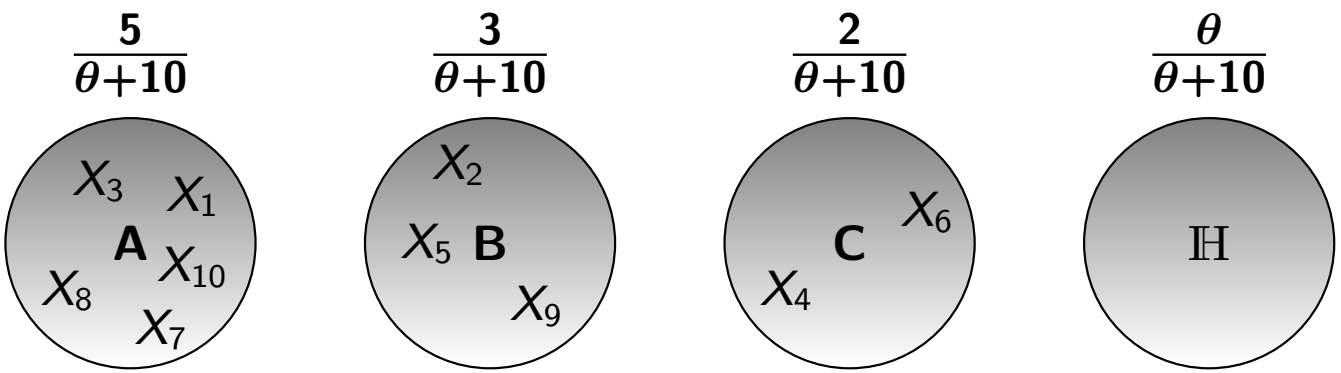
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	A	B	A	C	B	C	A	A	B	A



Clustering Property

- $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

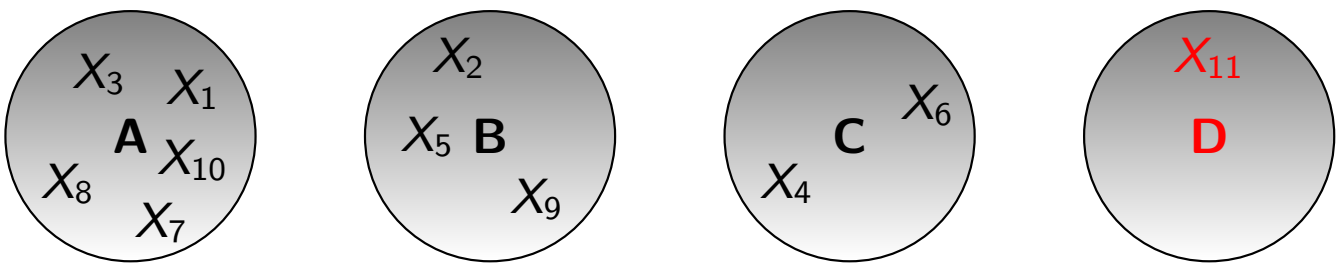
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
Value	A	B	A	C	B	C	A	A	B	A	



Clustering Property

- $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

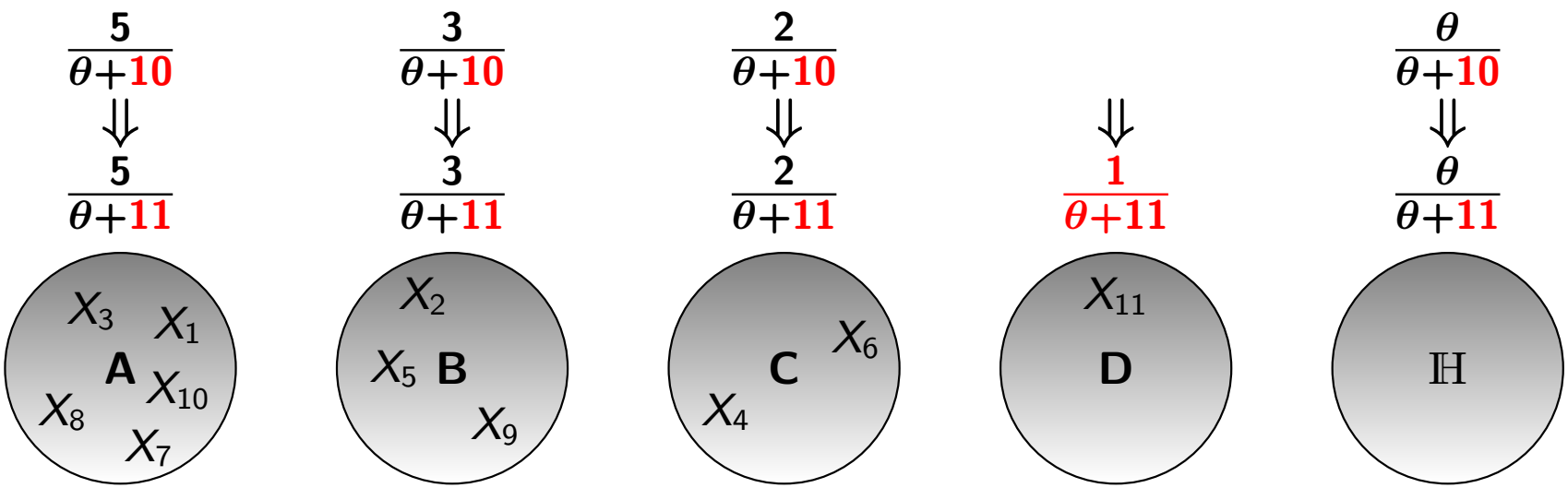
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
Value	A	B	A	C	B	C	A	A	B	A	D



Clustering Property

- $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

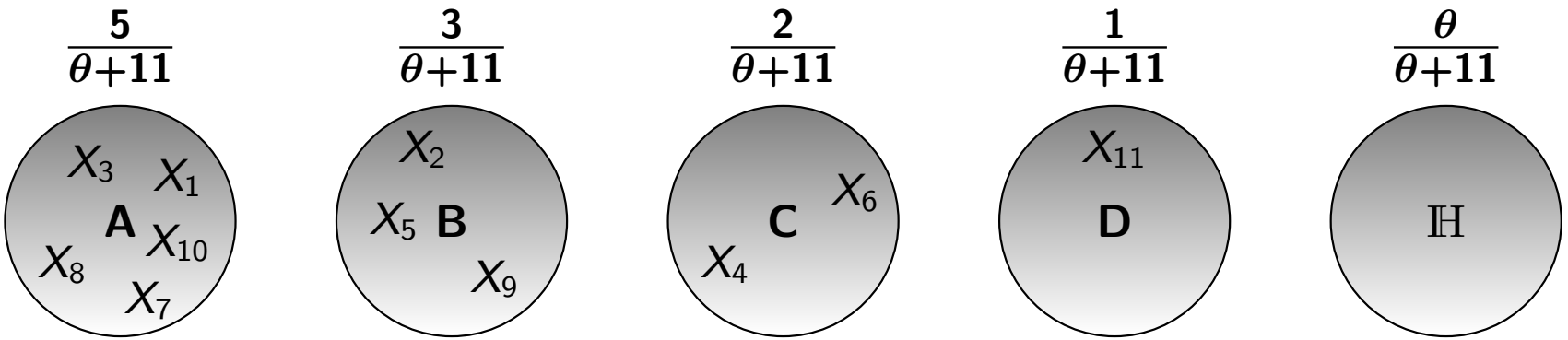
Obs.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Value	A	B	A	C	B	C	A	A	B	A	D	



Chinese Restaurant Process

- $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

Obs.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Value	A	B	A	C	B	C	A	A	B	A	D	



- Chinese Restaurant with infinite number of tables with infinite number of seats
- new Customer gets seated at an empty table with probability $\frac{\theta}{\theta+n}$ or an occupied table with probability proportional to number of sitting people
- rich get richer property

Asymptotic

Theorem 10

Let $\mathbf{P} \sim D_{(\theta,\mathbb{H})}$ *with continuous* $\mathbb{H}$, $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$ *and*

$$M_n := \#\{m \in \{1,...,n\} \,|\, X_m \notin \{X_1,...,X_{m-1}\}\}.$$

Then

$$\lim_{n\to\infty} \frac{M_n}{\log n} = \theta \qquad \mathbb{P}\text{-a.s.}$$

- M_n number of distinct values up to the n-th observation
- logarithmic growth of distinct values
- main reason for generalization to Pitman-Yor process

3. Pitman-Yor Process

Pitman-Yor process

Definition 11

Let $0 \leq \alpha < 1$, $\theta > -\alpha$, $\beta_n \overset{iid}{\sim} Beta(1-\alpha, \theta+n\alpha)$ and $(V_n)_{n\in\mathbb{N}} \sim \mathbb{H}^{\otimes\mathbb{N}}$ all independent. Define

$$W_k := \beta_k \prod_{i=1}^{k-1}(1-\beta_i),$$

then

$$\sum_{i=1}^{\infty} W_i \delta_{V_i} =: \mathbf{P} \sim D_{(\theta,\mathbb{H})}.$$

- generalization of Dirichlet process
- sorted weighting sequence is called two-parameter Poisson-Dirichlet distribution
- also known as Poisson-Dirichlet process
- α discount parameter
- θ strength parameter

Posterior Predictive

Theorem 12

Let $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ *with continuous base measure and* $(X_i)_{i\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$. *Then*

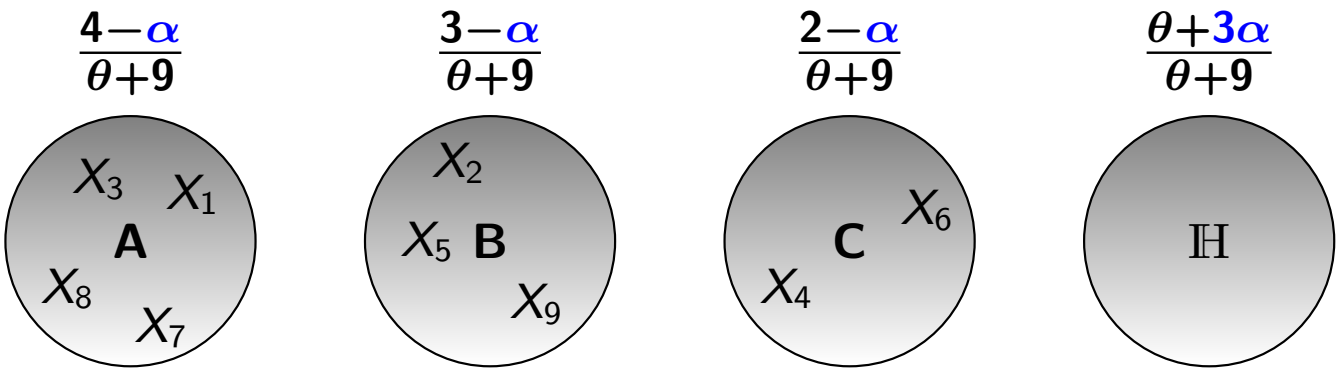
$$X_{n+1} \,|\, X_1, ..., X_n \sim \frac{1}{\theta + n}\left((\theta + M_n\alpha)\mathbb{H} + \sum_{j=1}^{M_n}(K_j^n - \alpha)\delta_{Y_j}\right).$$

- M_n number of different values up to the n-th observation
- Y_j unique values
- K_j^n number of appearances of Y_j up to the n-th observation
- α reduces initial probability of clusters
- continuity of $\mathbb{H}$ ensures that recurrent values are from the respective cluster

Clustering Property

- $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

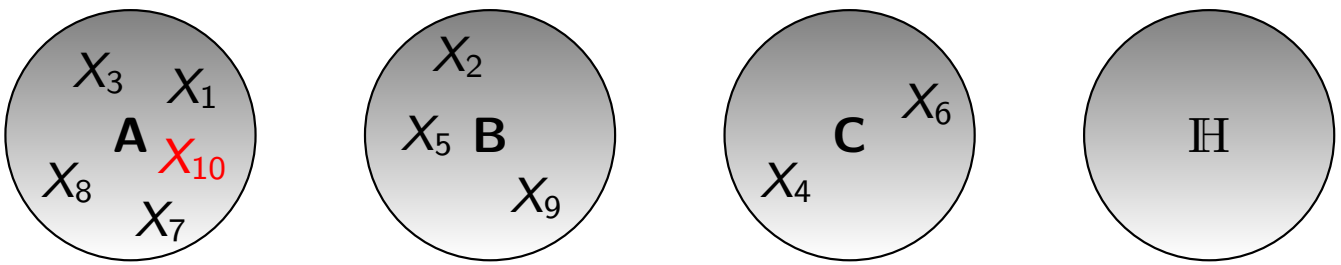
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	A	B	A	C	B	C	A	A	B	



Clustering Property

- $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

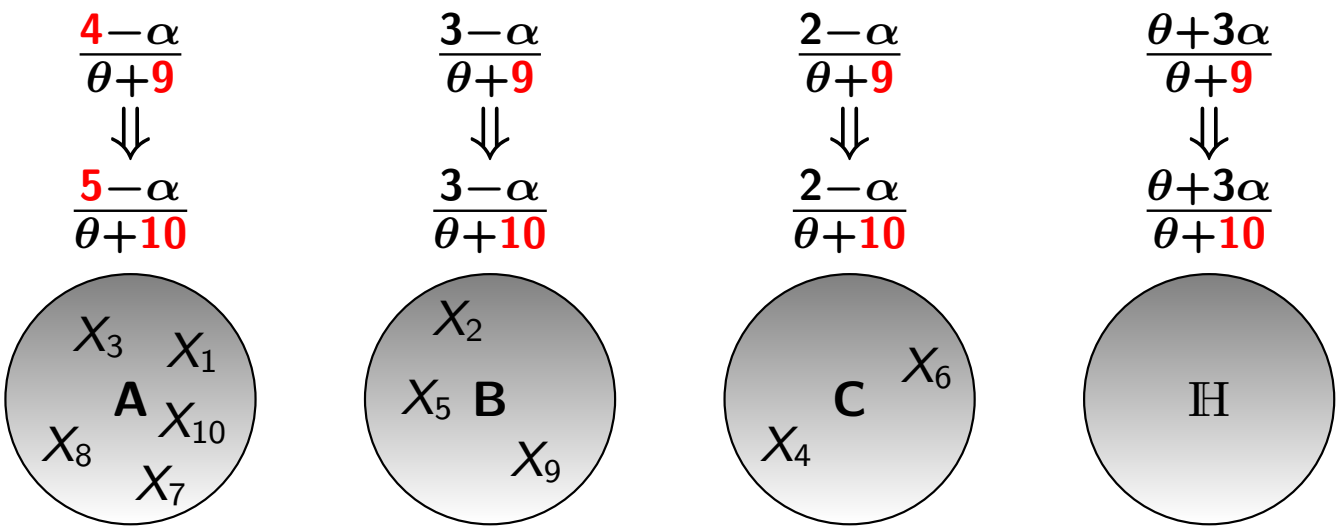
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	A	B	A	C	B	C	A	A	B	A



Clustering Property

- $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

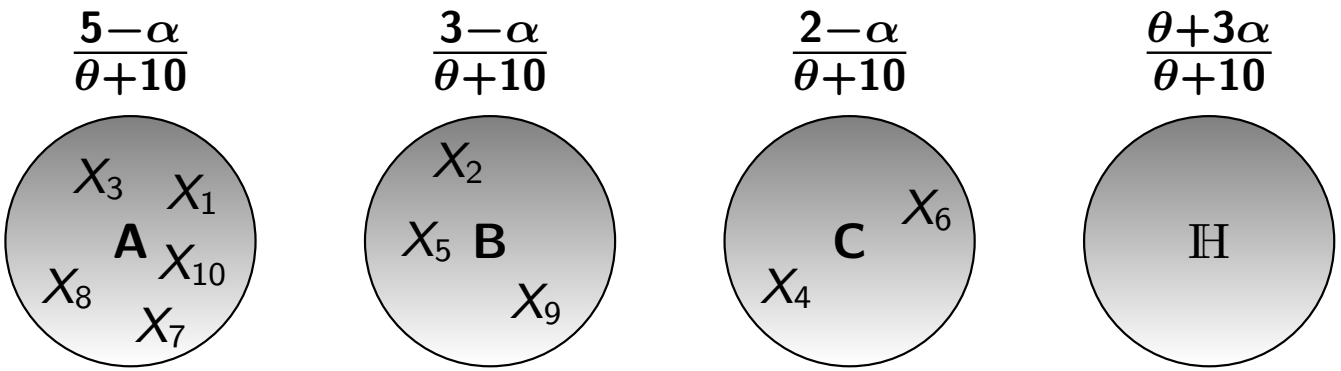
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	A	B	A	C	B	C	A	A	B	A



Clustering Property

- $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

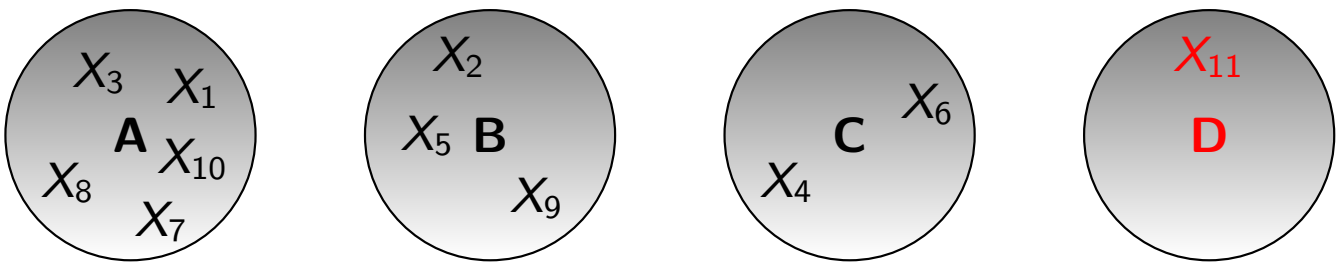
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
Value	A	B	A	C	B	C	A	A	B	A	



Clustering Property

- $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

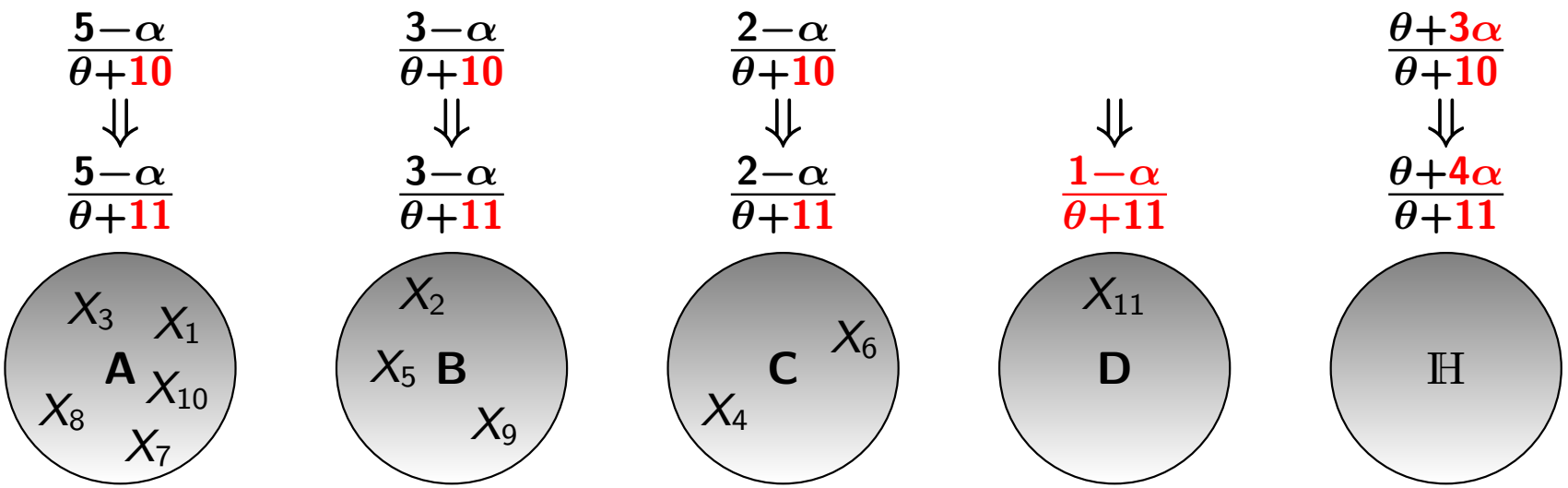
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
Value	A	B	A	C	B	C	A	A	B	A	D



Clustering Property

- $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

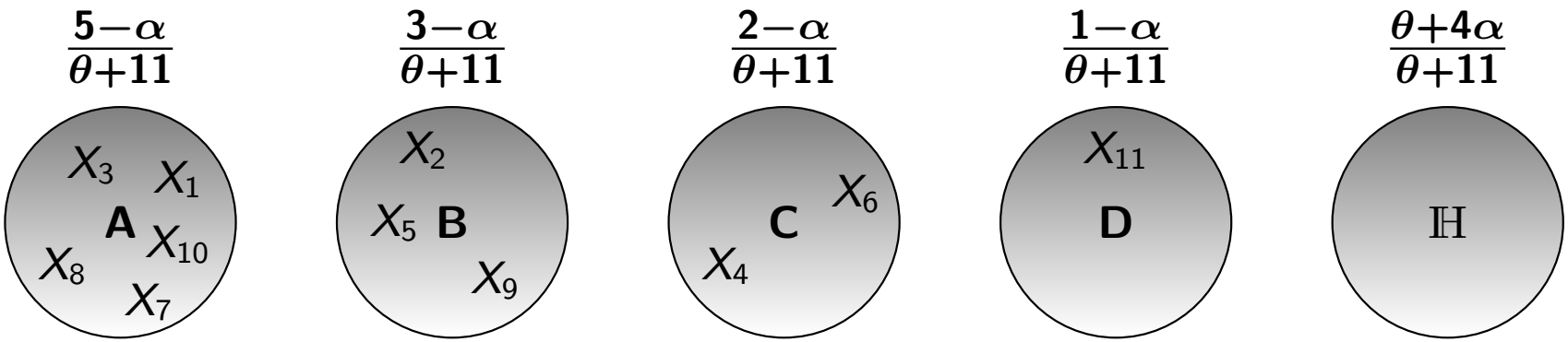
Obs.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Value	A	B	A	C	B	C	A	A	B	A	D	



Clustering Property

- $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ and $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$

Obs.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Value	A	B	A	C	B	C	A	A	B	A	D	



Proof

Theorem 13

Let $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ *with continuous base measure and* $(X_i)_{i\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$. *Then*

$$X_{n+1} \,|\, X_1, ..., X_n \sim \frac{1}{\theta + n}\left((\theta + M_n\alpha)\mathbb{H} + \sum_{j=1}^{M_n}(K_j^n - \alpha)\delta_{Y_j}\right).$$

- M_n number of different values up to the n-th observation
- Y_j unique values
- K_j^n number of appearances of Y_j up to the n-th observation

Proof (de Finetti)

- show exchangeability of $(X_n)_{n\in\mathbb{N}}$
- by de Finetti there is unique $\mathbf{P}'$ with

$$(X_i)_{i\in\mathbb{N}} \mid \mathbf{P}' = P' \sim P'^{\otimes\mathbb{N}}$$

  for $P' \in M(\mathscr{X})$
- further by de Finetti

$$\mathscr{L}(\mathbf{P}') = w\text{-}\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} = w\text{-}\lim_{n\to\infty}\sum_{i=1}^{M_n}\frac{K_i^n}{n}\delta_{Y_i} = \sum_{i=1}^{\infty}\delta_{Y_i}\lim_{n\to\infty}\frac{K_i^n}{n} \quad \mathbb{P}\text{-a.s.}$$

- show $lim_{n\to\infty}\frac{K_i^n}{n}$ has same distribution as respective element of weighting sequence

Asymptotic

Theorem 14

Let $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$ *with continuous* $\mathbb{H}$ *and*

$$(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$$

for $P \in M(\mathscr{X})$. *Define the number of clusters*

$$M_n := \#\{m \in \{1,...,n\} \,|\, X_m \notin \{X_1,...,X_{m-1}\}\}.$$

Then the almost sure limit of $\frac{M_n}{n^\alpha}$ *exists*

$$0 < \lim_{n\to\infty} \frac{M_n}{n^\alpha} < \infty \qquad \mathbb{P}\text{-a.s.}$$

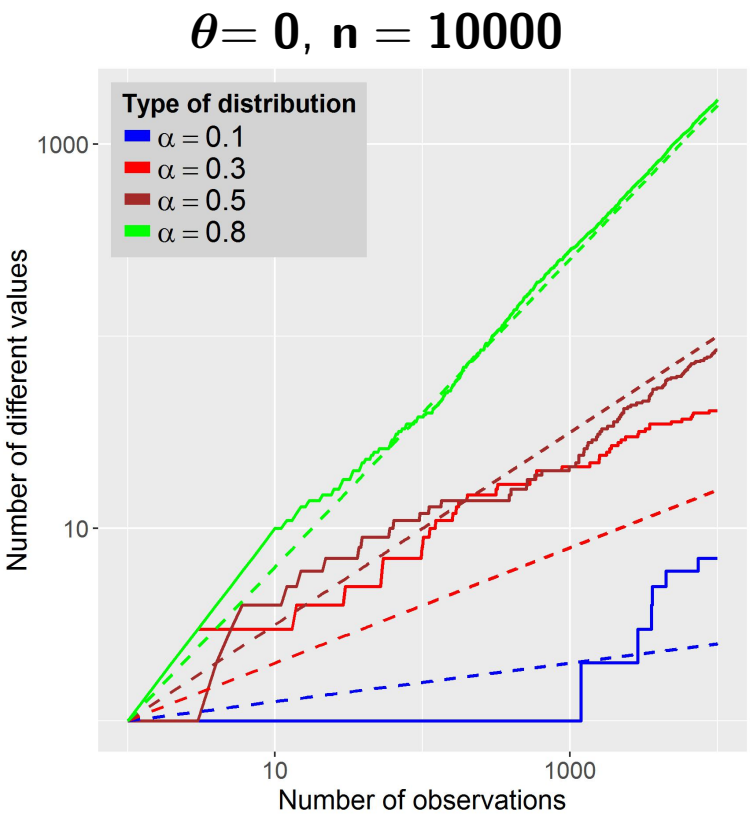
- main reason for generalization of Dirichlet process
- now $M_n \overset{n\to\infty}{\cong} n^\alpha$ instead of $M_n \overset{n\to\infty}{\cong} \log n$

Asymptotic Plots

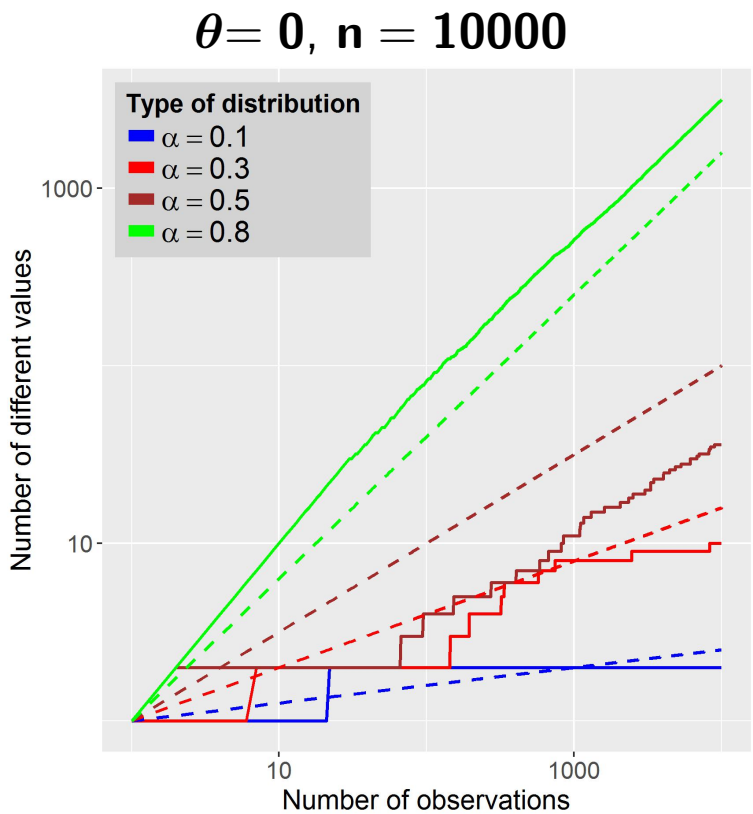
- various plots, showing development of new values in samples
- number of observations n and θ varies
- in each plot several α
- if $\theta > 0$ Dirichlet process is plotted
- log-log plots show power-law by a linear function

$$\frac{M_n}{n^\alpha} \approx \phi \Leftrightarrow \log(M_n) \approx \log(\phi) + \alpha \log(n)$$

Asymptotic Plots

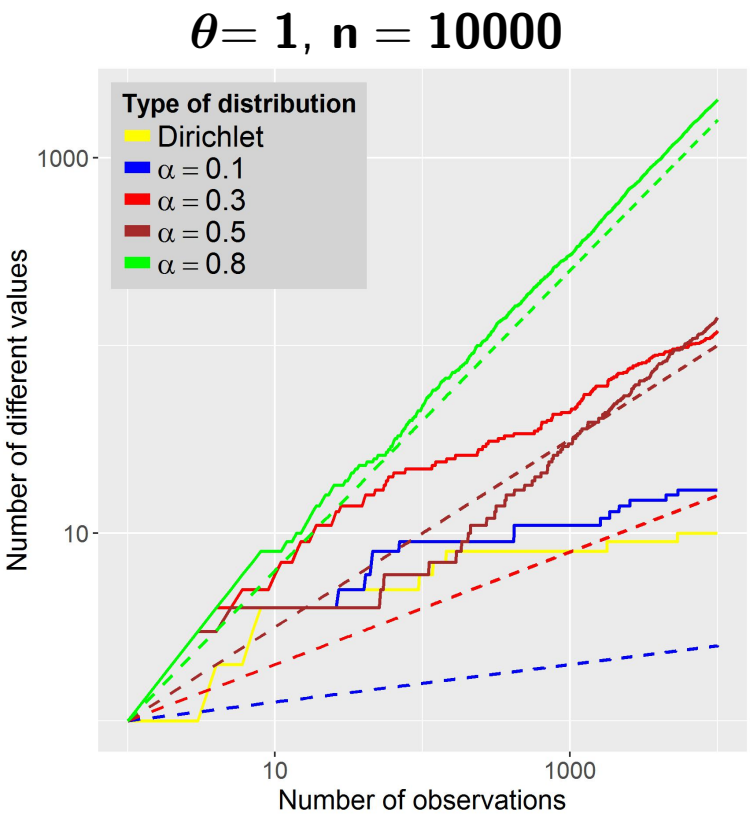


- roughly fit prediction
- larger values fit better

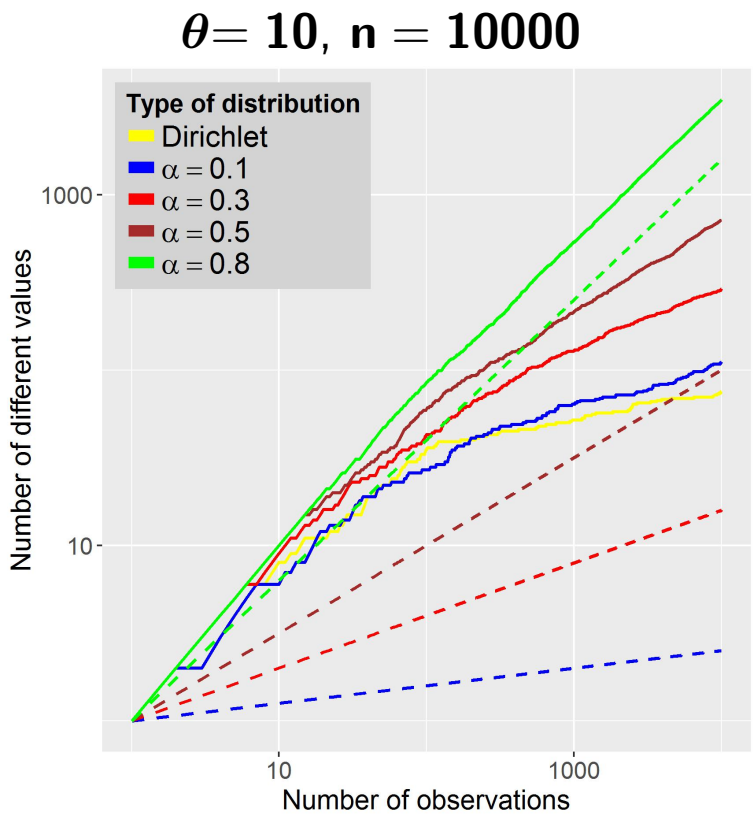


- same observations as left
- developments differ noticeable at smaller values

Asymptotic Plots



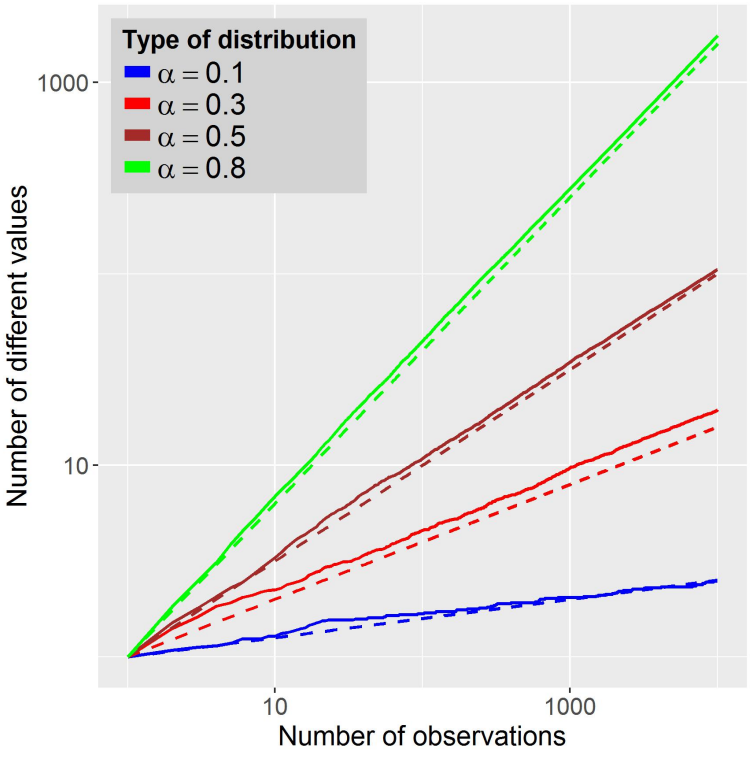
- stronger slope for smaller values of α
- complete balance is missing



- stronger slope clearly visible for all values of α
- slight compensation in the end

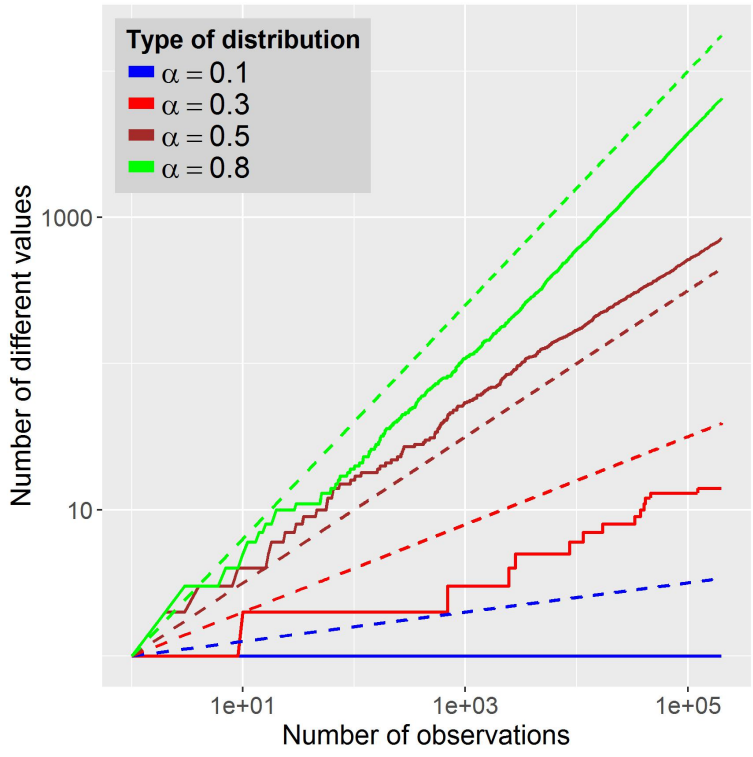
Asymptotic Plots

$\theta = 0$, n $= 10000$, averaged



- no statistical deviation
- power law clearly visible

$\theta = 0$, n $= 200000$



- smoother result
- still deviations at small α

Consistency

Theorem 15

Let $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$, $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$ *for* $P \in M(\mathscr{X})$ *and* $(Y_n)_{n\in\mathbb{N}} \sim \mathbb{Q}^{\otimes\mathbb{N}}$ *with a discrete measure* $\mathbb{Q}$. *If it holds*

$$\frac{M_n}{n} := \frac{\#\{m \in \{1,...,n\} \,|\, X_m \notin \{X_1,...,X_{m-1}\}\}}{n} \overset{n\to\infty}{\to} 0 \qquad \mathbb{P}\text{-a.s.}$$

then

$$\mathscr{L}\left(\mathbf{P} \,|\, X_1 = Y_1, ..., X_n = Y_n\right) \overset{w}{\to} \delta_{\mathbb{Q}} \qquad \mathbb{P}\text{-a.s.}$$

- consistent if proportion of distinct observations asymptotically diminishes
- no consistency for continuous $\mathbb{Q}$
- stronger result possible

Consistency

Corollary 16

Let $\mathbf{P} \sim PY_{(\alpha,\theta,\mathbb{H})}$, $(X_n)_{n\in\mathbb{N}} \,|\, \mathbf{P} = P \sim P^{\otimes\mathbb{N}}$ *for* $P \in M(\mathscr{X})$ *and* $(Y_n)_{n\in\mathbb{N}} \sim \mathbb{Q}^{\otimes\mathbb{N}}$ *with a discrete measure* $\mathbb{Q}$*, then*

$$\mathscr{L}\left(\mathbf{P} \,|\, X_1 = Y_1, ..., X_n = Y_n\right) \xrightarrow{w} \delta_{\mathbb{Q}} \qquad \mathbb{P}\text{-a.s.}$$

- show for all discrete measures

$$\frac{M_n}{n} := \frac{\#\{m \in \{1, ..., n\} \,|\, X_m \notin \{X_1, ..., X_{m-1}\}\}}{n} \xrightarrow{n\to\infty} 0 \qquad \mathbb{P}\text{-a.s.}$$

4. Application

Simulation

- language modeling
- only cluster development without base measure
- create sample, conditioned on initial clustered text
- compared to whole clustered text
- plots comparing sorted cluster masses
- absolute error

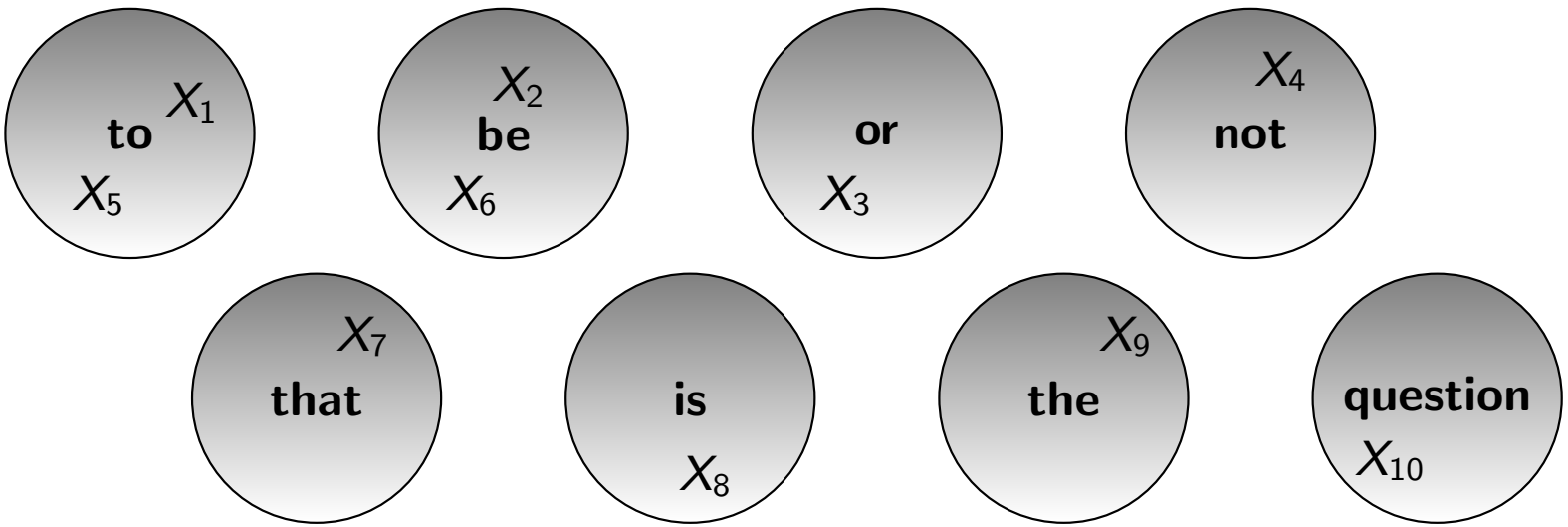
$$\Delta_{\theta,\alpha,n} = \sum_{i=1}^{\max(M_n,c)} |K_i^n - W_i^n|$$

- K_i^n and W_i^n denote mass of i-th cluster and appearance of i-th word
- M_n and c denote number of clusters and unique words

Approach

- sort words in clusters

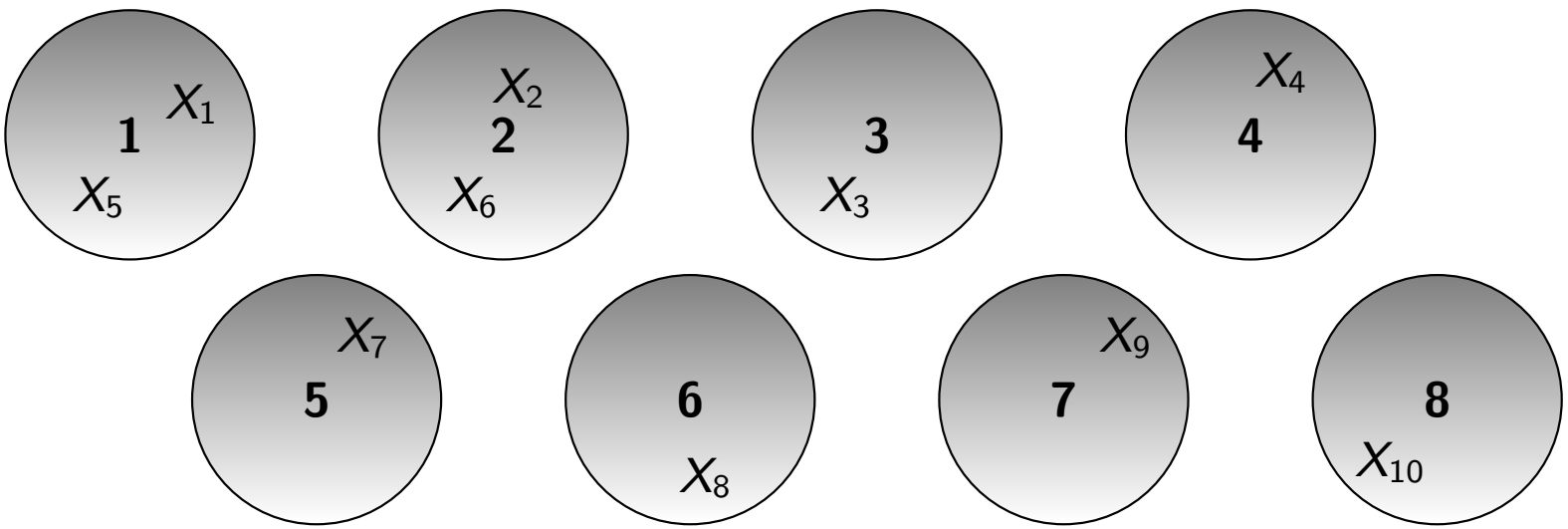
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Word	to	be	or	not	to	be	that	is	the	question



Approach

- exchange cluster labels in numbers

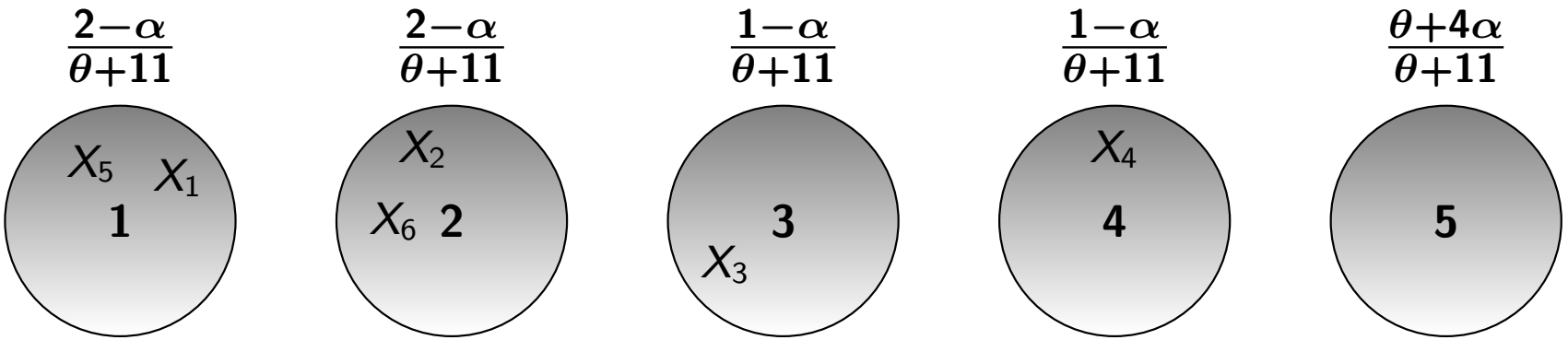
Word	to	be	or	not	to	be	that	is	the	question
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	1	2	3	4	1	2	5	6	7	8



Approach

- get initial values and start sampling process
- use sampling formula

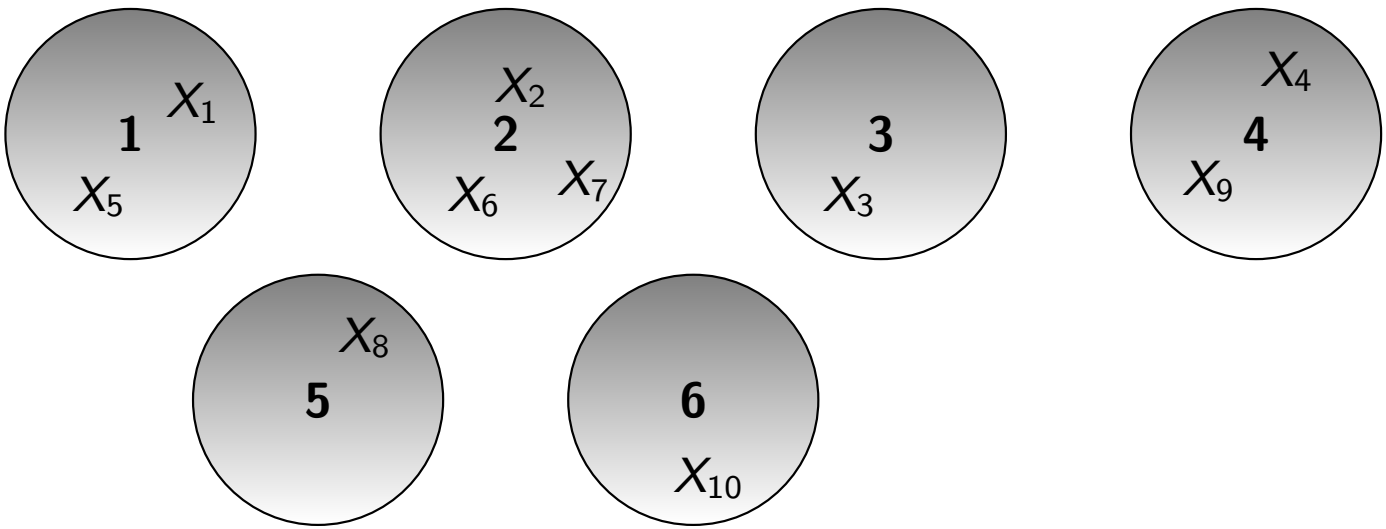
Words	to	be	or	not	to	be	
Obs.	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Value	1	2	3	4	1	2	



Approach

- get sample value for all observations

Word	to	be	or	not	to	be				
Observation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Value	1	2	3	4	1	2	2	5	4	6



Approach

- compare sample clusters and text clusters

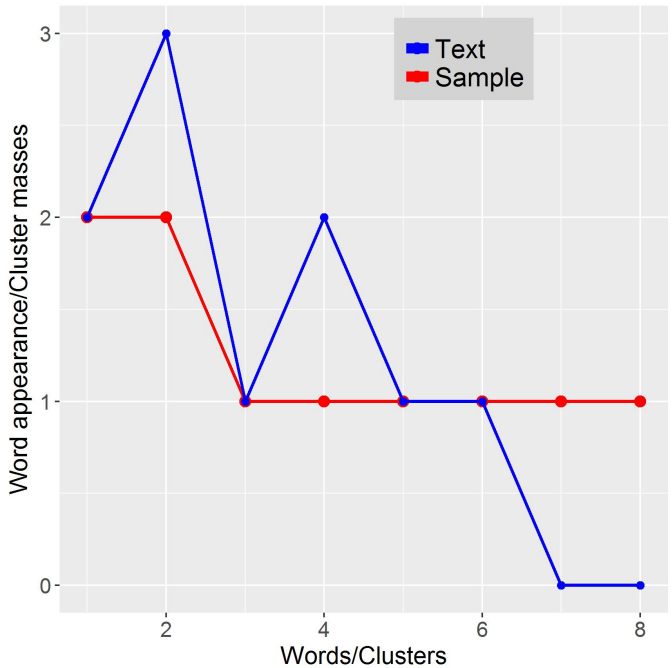
Contained Words	2	2	1	1	1	1	1	1
Contained Samples	2	3	1	2	1	1	0	0
Difference	0	1	0	1	0	0	1	1

- $\Rightarrow$ absolute error $\Delta_{\theta,\alpha,n} = 4$

Approach

- plot data for easier comparison

cluster	1	2	3	4	5	6	7	8
contained words	2	2	1	1	1	1	1	1
contained samples	2	3	1	2	1	1	0	0

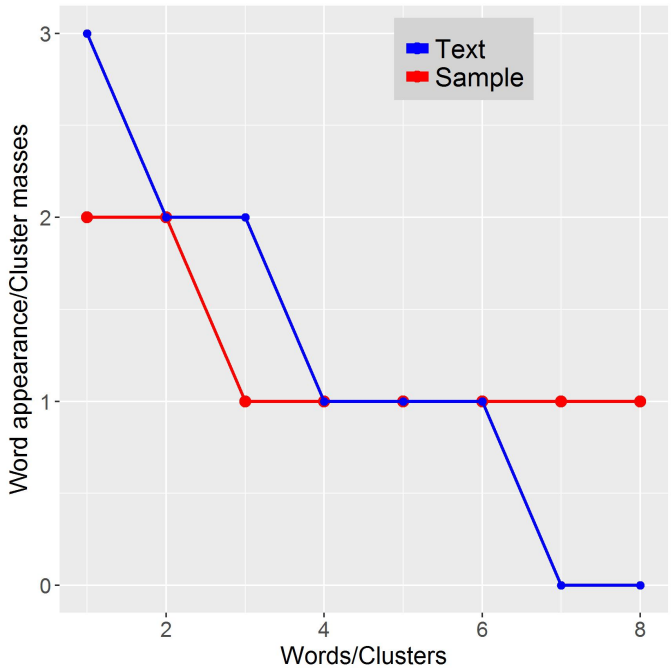


Cluster masses and word appearances, unsorted

Approach

- plot sorted data for clearer plot

cluster	1	2	3	4	5	6	7	8
contained words	2	2	1	1	1	1	1	1
contained samples	2	3	1	2	1	1	0	0



Cluster masses and word appearances, sorted

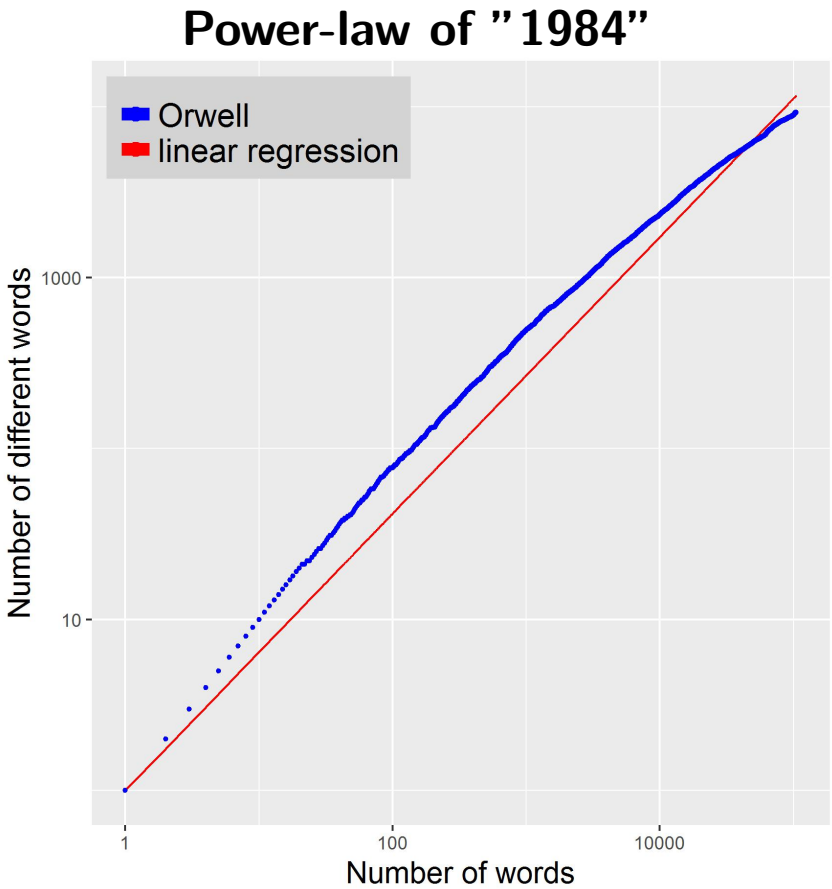
Application

- "1984" by George Orwell
- $n = 104671$ words and $c = 9264$ unique words in "1984"
- initial observations $m \in \{0, 100, 1000, 10000\}$
- parameter $\theta \in \{1, 100, 500\}$
- parameter $\alpha \in \{0, 0.3, 0.6, 0.801\}$
- simulation by use of posteriori predictive

$$X_{n+1} \mid X_1, ..., X_n \sim \frac{1}{\theta + n} \left((\theta + M_n \alpha) \mathbb{H} + \sum_{j=1}^{M_n} (K_j^n - \alpha) \delta_{Y_j} \right)$$

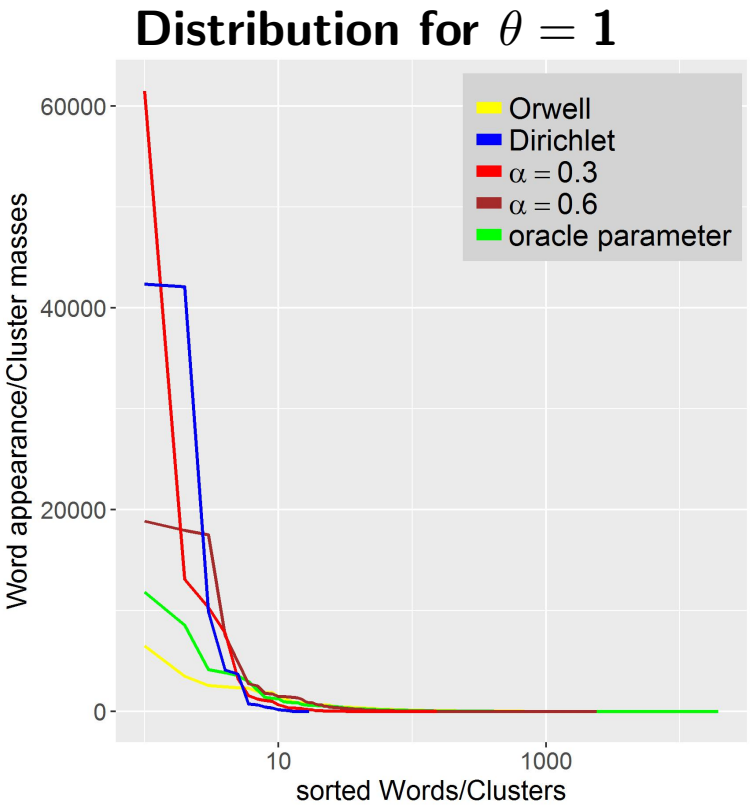
- larger θ implies faster growth of cluster count at the beginning
- larger α implies persistent faster growth of cluster count

Power-Law

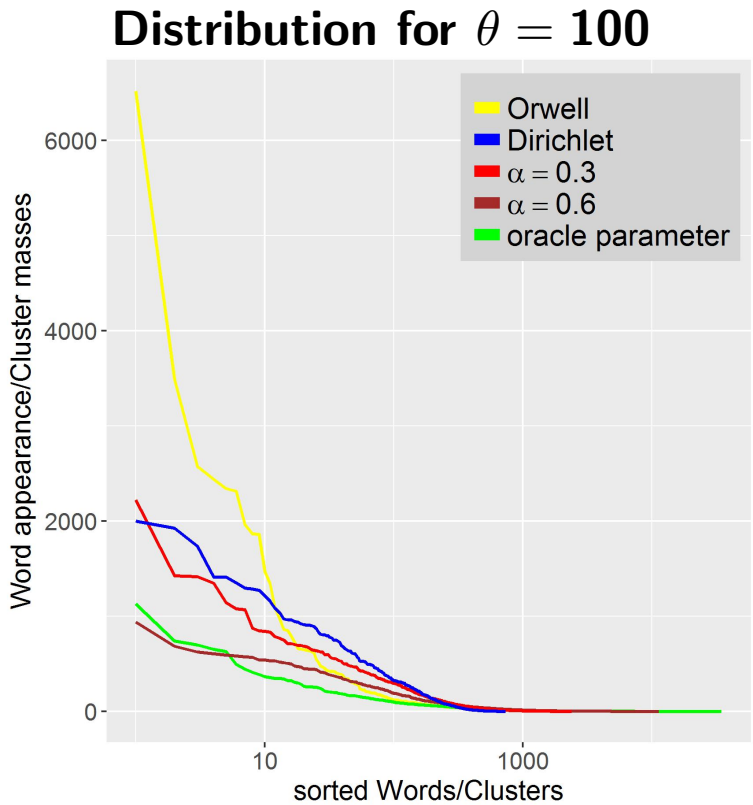


- approximately a power-law
- stronger growth at the beginning
- slower growth in the end
- linear regression acquires slope of $\alpha = 0.801$
- oracle parameter

Simulation Plots (m=0,averaged)

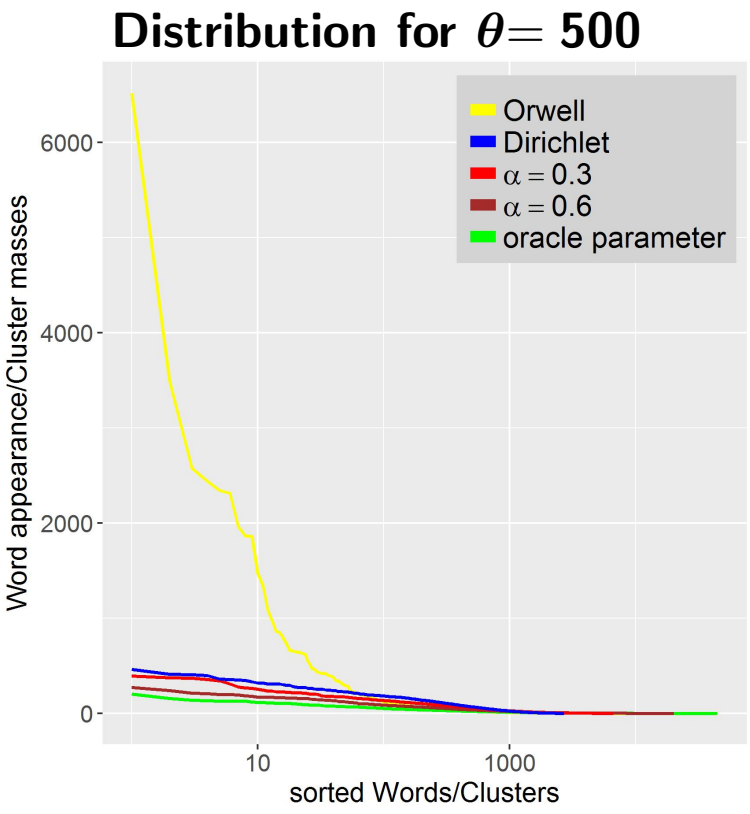


- oracle parameter fits best

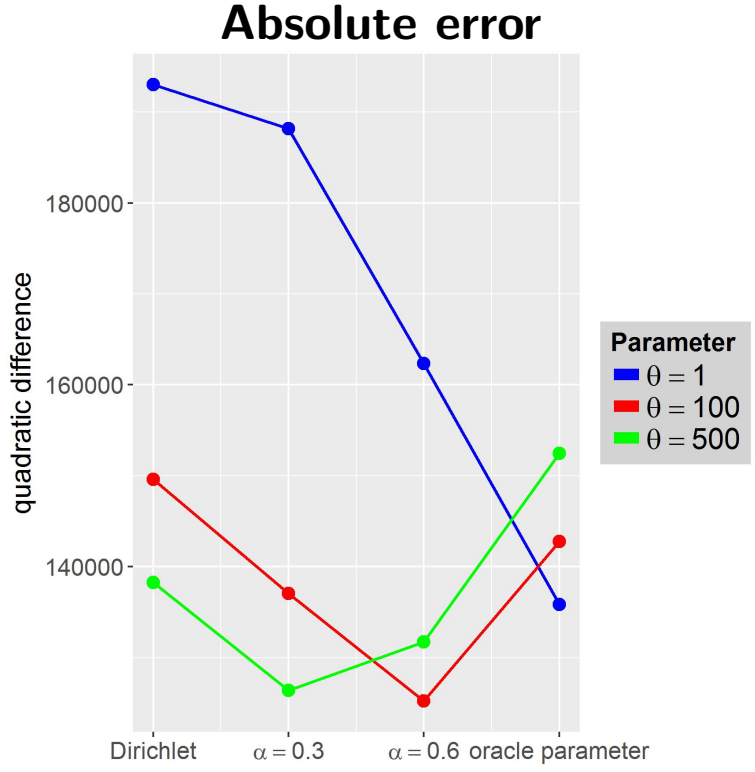


- underestimation for all samples

Simulation Plots (m=0,averaged)

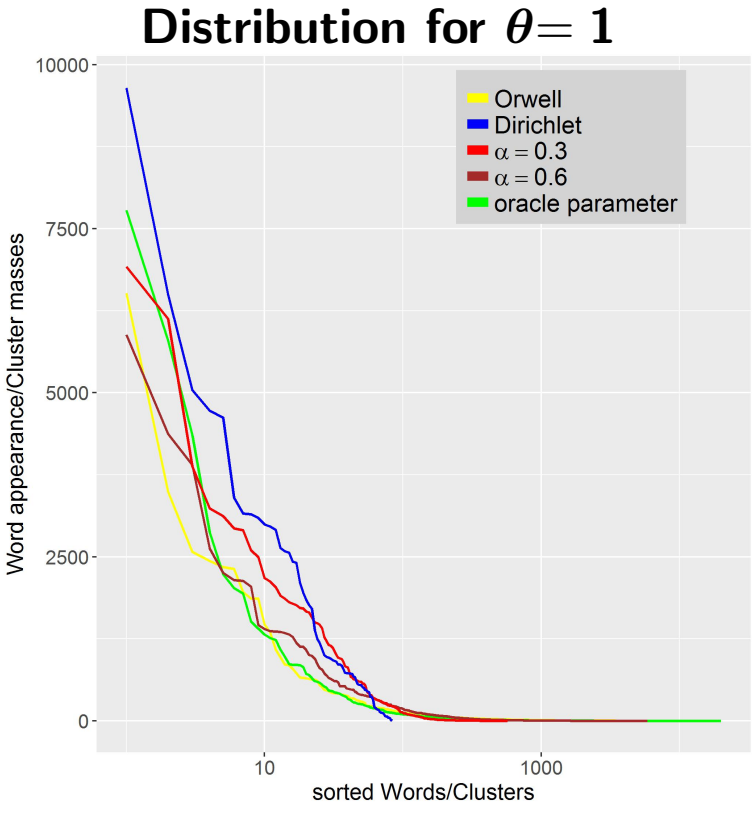


- massive underestimation for all samples

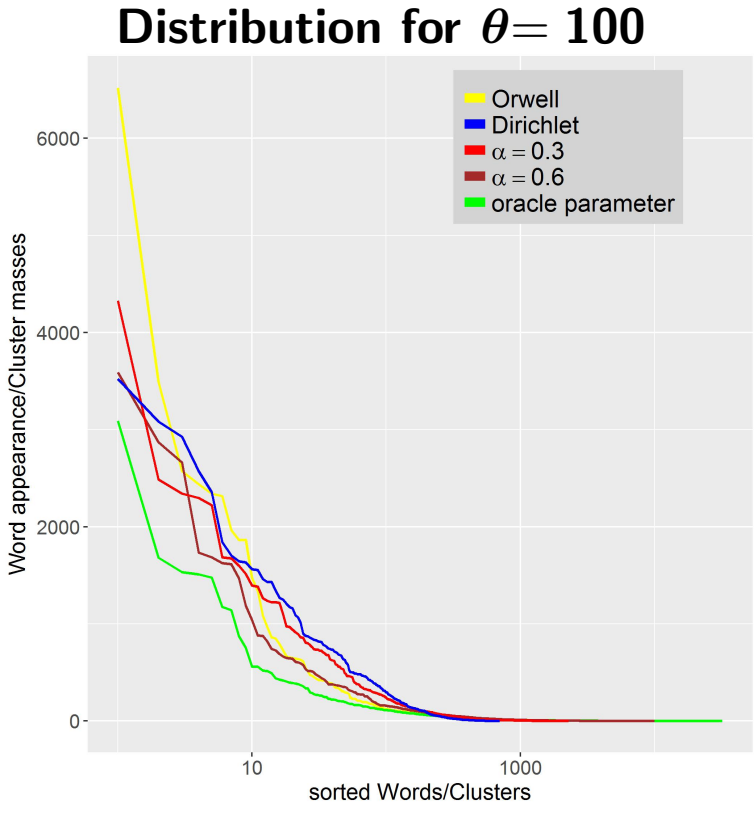


- samples fit better for Pitman-Yor
- smallest error for large θ because of imitating behavior of text

Simulation Plots (m=100)

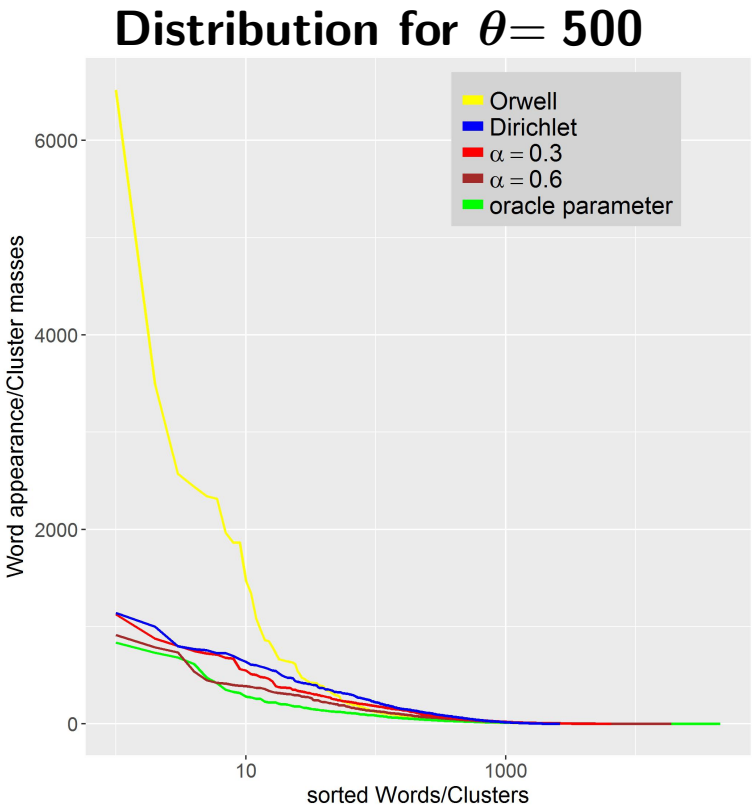


- oracle parameter fits mostly best

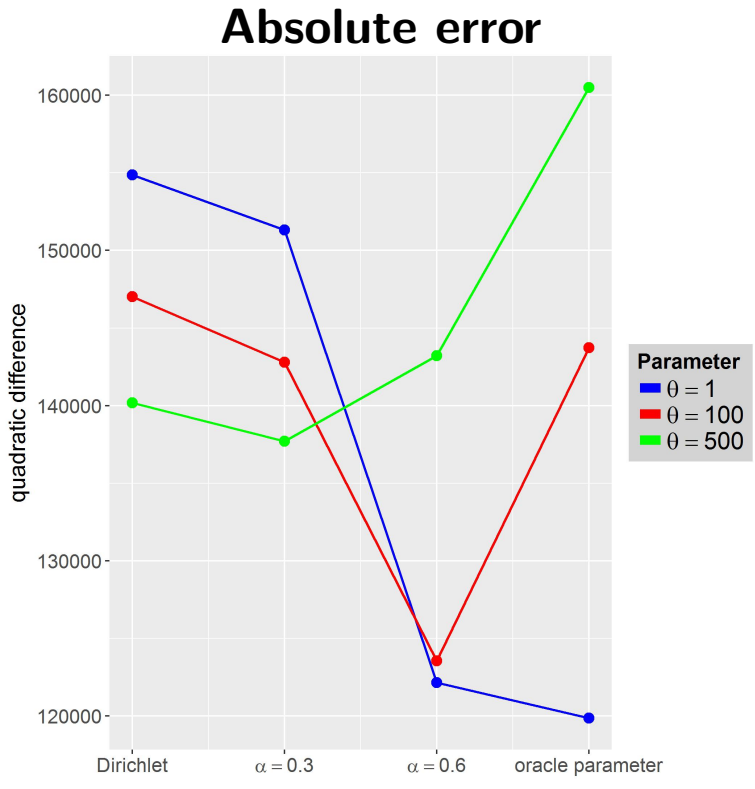


- in general slight underestimation

Simulation Plots (m=100)

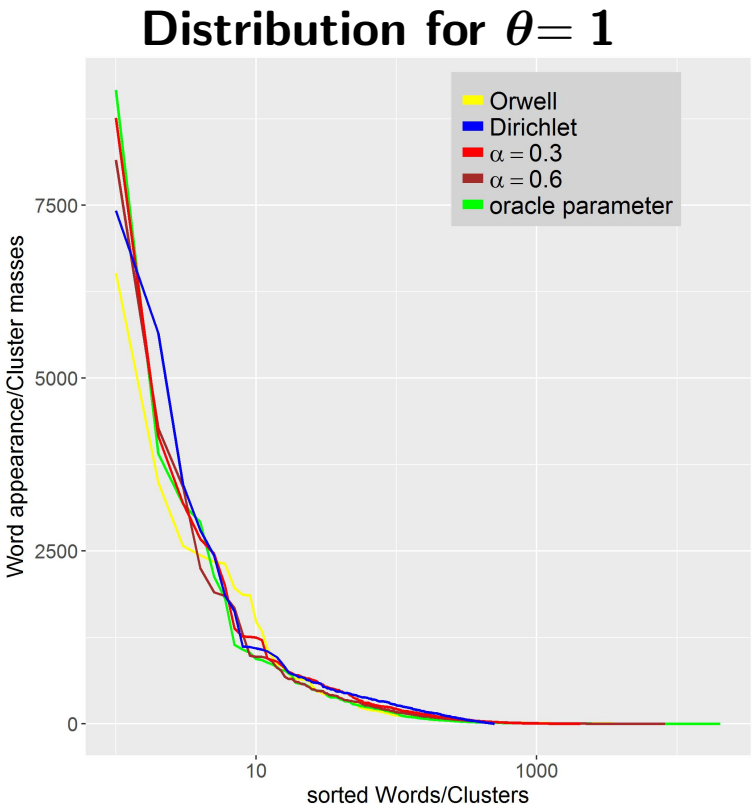


- massive underestimation for all samples

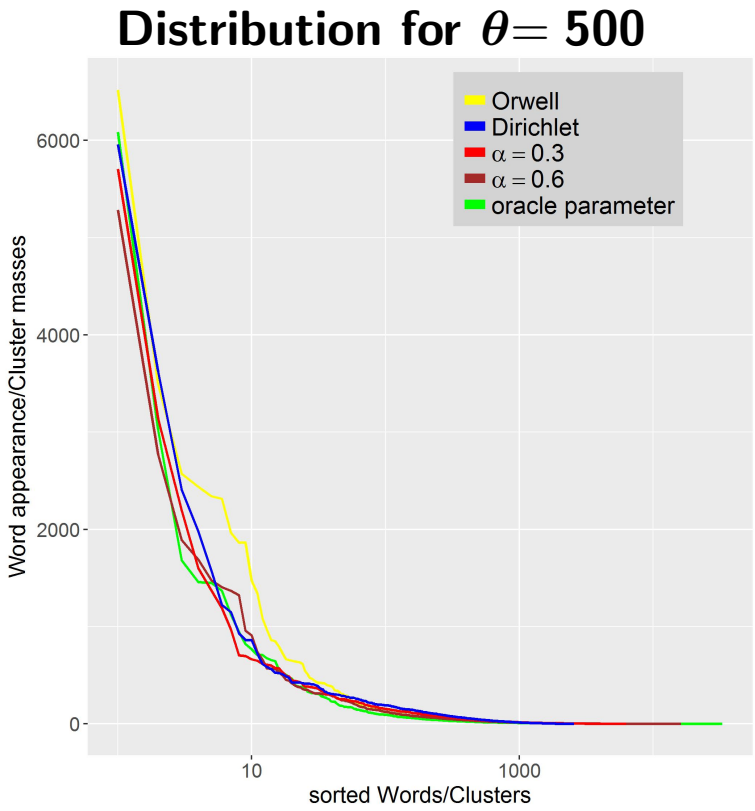


- samples fit better for Pitman-Yor
- now small θ gives best error

Simulation Plots (m=1000)

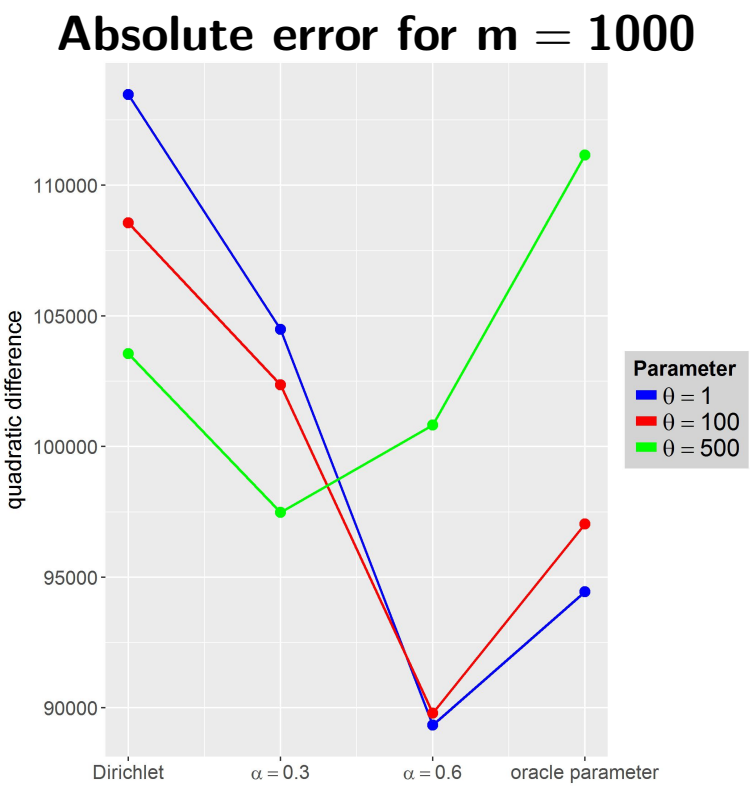


- no distinction possible

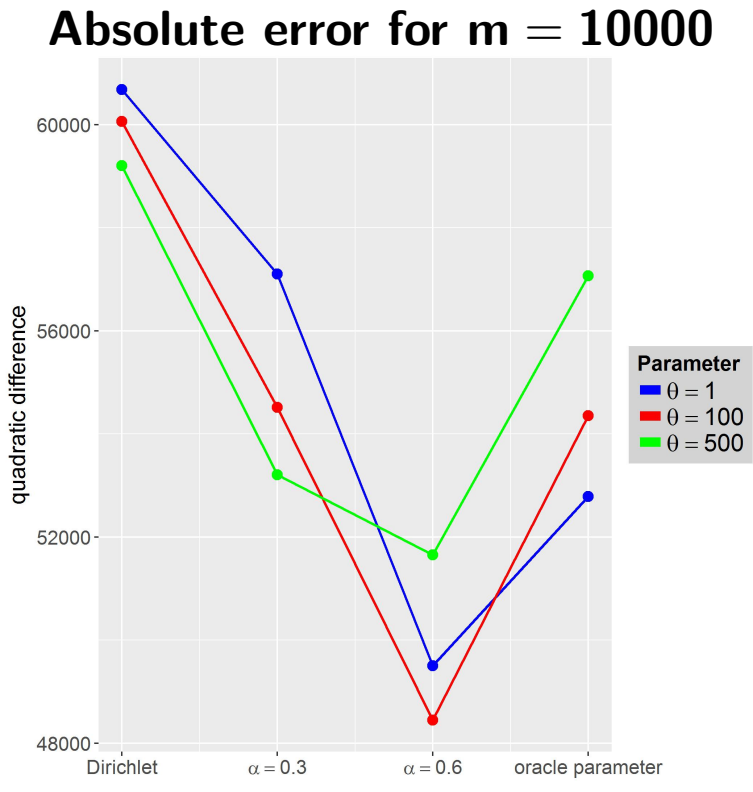


- only slight underestimation

Absolute Errors



- influence of small θ diminishes
- best error for small θ
- best error not at oracle parameter



- influence even of larger θ decreases
- best error for small θ
- best error not at oracle parameter

Thank you for your attention!

Bayesian Nonparametric Priors and their Clusterings

Florian Schimmel

Institut für Mathematik Universität Heidelberg

April 5, 2017