UNIVERSITY OF CAMBRIDGE

Pitfalls in the use of Parallel Inference for the Dirichlet Process

Yarin Gal, Zoubin Ghahramani

mlg.eng.cam.ac.uk/yarin

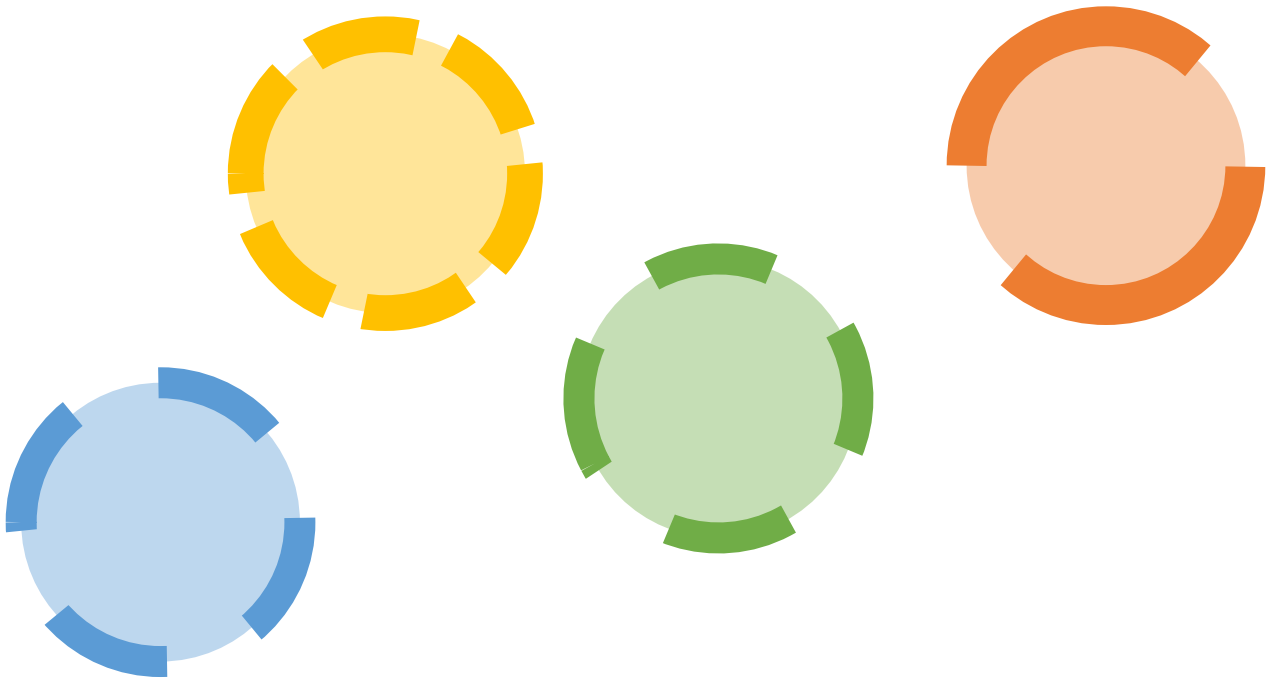
June 24, 2014

Outline

UNIVERSITY OF CAMBRIDGE

- The Dirichlet process
- Parallel inference
- Non-approximate parallel inference in the Dirichlet process
- What can go wrong
- How can we try to fix it

Sampling from the Dirichlet process – the Chinese restaurant process



- A restaurant with 4 tables and 2, 4, 4, and 6 customers sitting around each one

Real world applications – Natural Language Processing

- Language modelling
  - A derivative model (the Hierarchical Pitman–Yor process) was shown to correspond to the state-of-the-art in language modelling
- Machine Translation
  - Used to obtain state-of-the-art results in Bayesian word alignment
- Working with huge datasets (tens of GBs)
- Development cycle taking weeks at a time
- Usually using small values for the concentration parameter ($\alpha = 0.1$ is common)

- Inference is slow!
- A common problem with non-parametric techniques
- Possible solutions:
  - Variational inference - an approximate approach
  - Parallel MCMC inference

Parallel inference

UNIVERSITY OF CAMBRIDGE

Given a network with many nodes (computers in a network or cores in a cluster), we would like to have an inference that:

- distributes the computational load evenly across the nodes,
- scales favourably with the number of nodes,
- has low overhead in the global steps,
- and converges to the true posterior distribution

Parallel inference in the DP

UNIVERSITY OF CAMBRIDGE

- Approximate parallel inference (Asuncion, Smyth, and Welling [2008])
  - Gives slower convergence (Williamson et al., [2013])
- Non-approximate parallel inference using a re-parametrisation of the Dirichlet process
  - Recently suggested, independently, by Lovell, Adams, and Mansingka [2012] and Williamson, Dubey, and Xing [2013]

UNIVERSITY OF CAMBRIDGE

Two-staged Chinese restaurant process [Lovell et al., 2012]:

- Each data point (customer) chooses one of the K nodes (tables) according to its popularity:

$$P(\text{data point } n \text{ chooses node } k \mid \alpha) = \frac{\alpha\mu_k + \sum_{i=1}^{n-1} \mathbb{I}(s_{z_i} = k)}{\alpha + n - 1},$$

  with some weights μ_k where s_{z_i} is the node allocation of point i – this is equivalent to the Dirichlet-Categorical distribution.

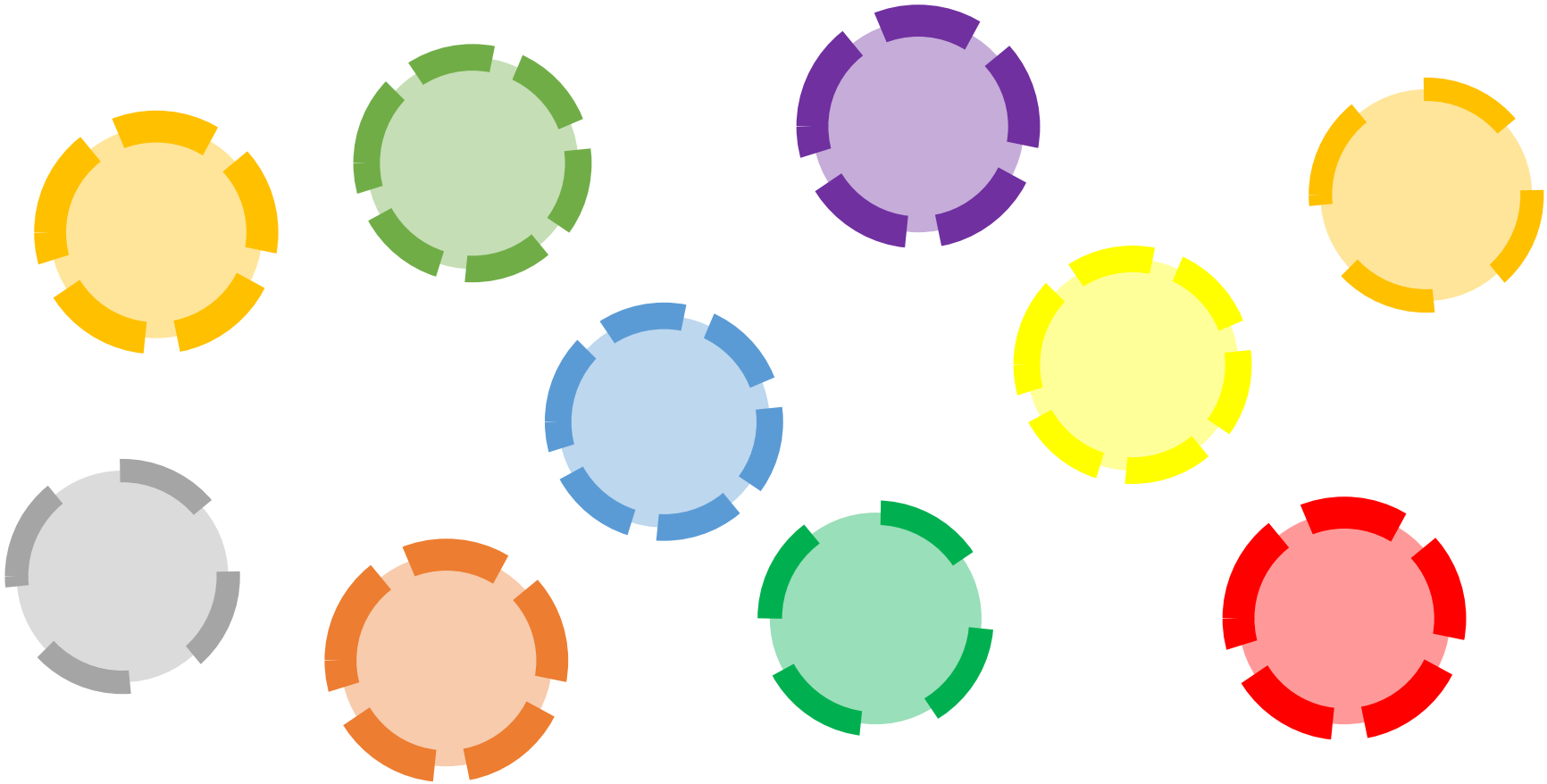
- In each node k the data points follow the usual Chinese restaurant process (CRP) with parameter $\alpha\mu_k$.

- The resulting random partition has the same distribution as the CRP with parameter α.

Parallel inference in the DP

UNIVERSITY OF CAMBRIDGE

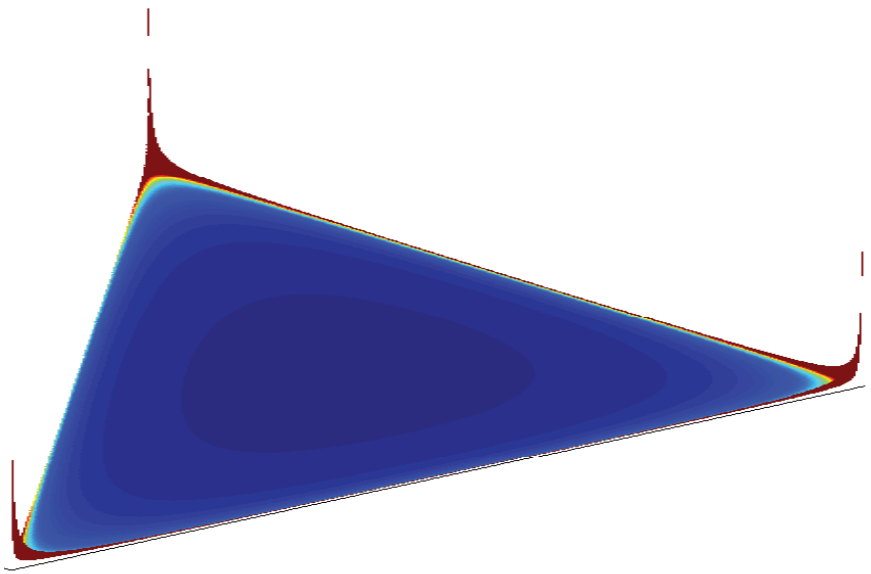
For a network with 10 nodes we split the data using a sample from a Dirichlet distribution with 10 components:



- Each table corresponding to a single node and each customer to a data point sent to that node

However...

- Samples from the Dirichlet distribution with parameter smaller than 1 have most of the mass concentrated around the corners of the simplex



However...

UNIVERSITY OF CAMBRIDGE

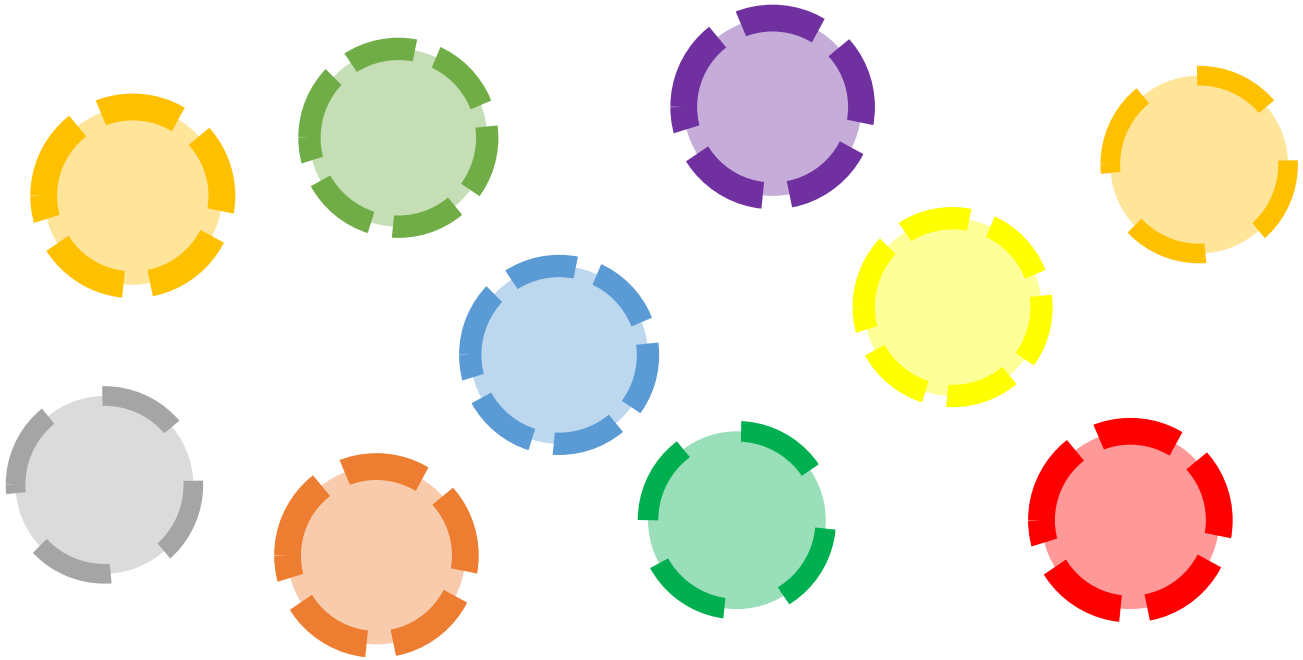
- and in the limit of K we obtain samples from the Dirichlet process with parameter α:

$$\mathrm{DP}(\alpha)$$

- This means that the expected number of nodes used is the same as the expected number of tables in a restaurant with parameter α
  - (we can augment the number of nodes by sending multiple jobs to the same machine)

However...

Actual samples from a Dirichlet process with 50 data points don't look like this:



- The expected number of tables in a restaurant with n customers is given by

$$\alpha \log(n)$$

However...

UNIVERSITY OF CAMBRIDGE

So a sample from a Dirichlet process with 50 data points would look more like this:

Which means that only a constant number of nodes, dependent on the number of data points, would be used.

However...

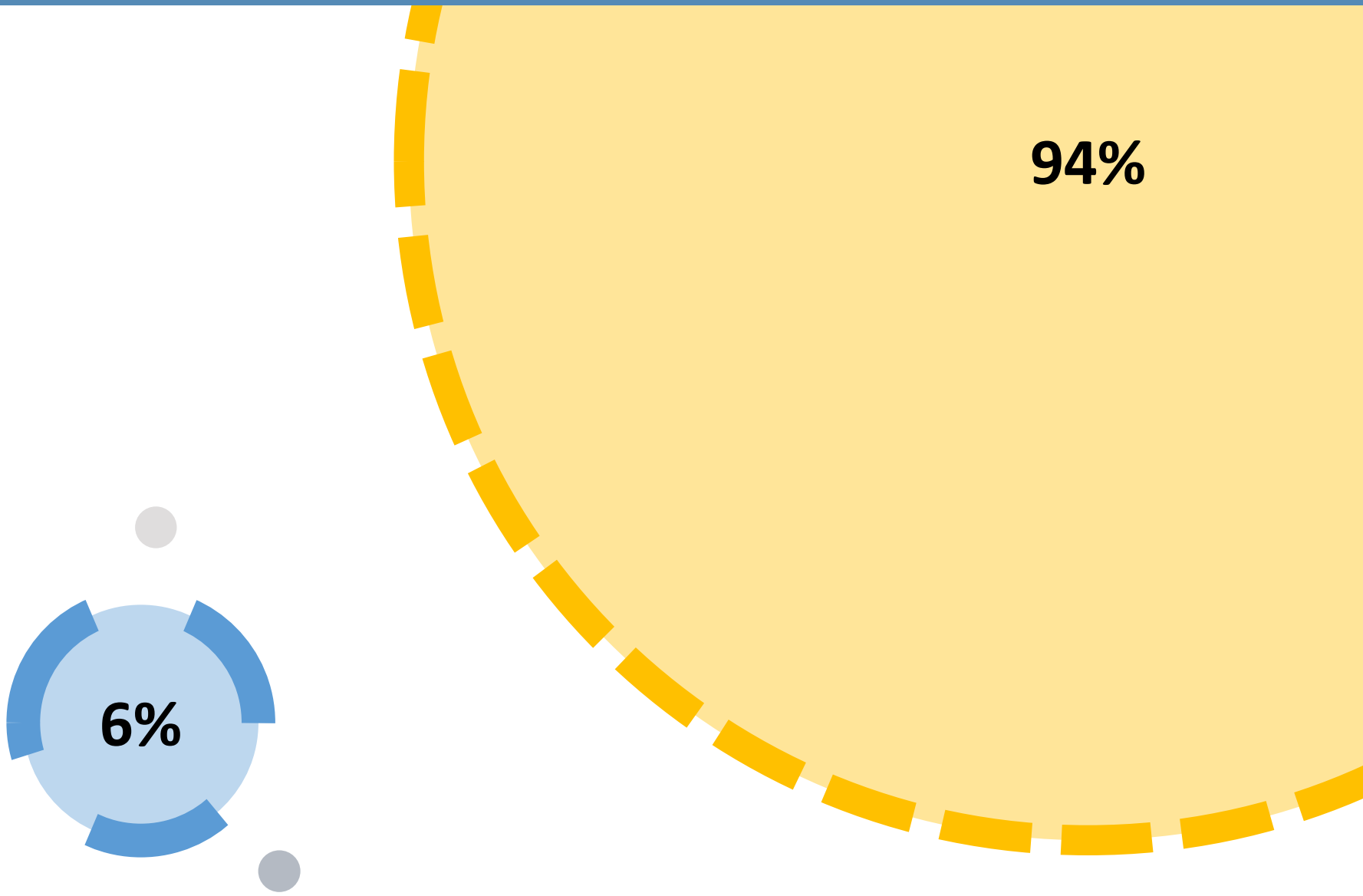
UNIVERSITY OF CAMBRIDGE

Even worse, the sizes of the different tables follows an exponential decay, so the the number of customers sitting next to each table would actually be

$$C, \quad Cq, \quad Cq^2, \quad Cq^3, \quad \cdots$$

for $q = \dfrac{\alpha}{1+\alpha}$ and $C = \dfrac{1}{1+\alpha}$, so an actual sample would be...

However...



for $n = 50$ data points and $\alpha = 0.1$.

However...

UNIVERSITY OF CAMBRIDGE

So for $n = 50$ data points and $\alpha = 0.1$ the parallel inference would send 94% of the data to a single machine.

- Sampling from the *finite* **Dirichlet distribution** with K components (nodes in a network) and different parameter values we get a load balance:

# of nodes	$\alpha = 0.1$	$\alpha = 2$
$K = 10^1$	94%, 6%, 0%, 0%, ...	54%, 23%, 12%, 6%, ...
$K = 10^2$	94%, 6%, 0%, 0%, ...	48%, 22%, 12%, 7%, ...
$K = 10^3$	94%, 6%, 0%, 0%, ...	48%, 21%, 12%, 7%, ...
$K = 10^4$	94%, 6%, 0%, 0%, ...	48%, 21%, 12%, 7%, ...
$K = 10^5$	94%, 6%, 0%, 0%, ...	48%, 21%, 12%, 7%, ...

Figure: Average load on each node in decreasing order

However...

UNIVERSITY OF CAMBRIDGE

And in general, for a Dirichlet process with parameter α, 95% of the data for would be sent to

$$\approx \frac{1.3}{\log(\alpha + 1) - \log(\alpha)}$$

nodes,

- independently of the size of the dataset,
- independently of the number of nodes in the network,
- and dependent only on the **parameter used to model the data**

What can we do?

- We can try to initialise the sampler near the posterior
- We could use approximate inference with Metropolis–Hastings corrections
- We can develop better approximate inference approaches
- Don’t use the Dirichlet process?

What can we do?

We can try to initialise the sampler near the posterior

- When we know the data has **many clusters** which are **evenly balanced**
- Initialise the sampler randomly with many evenly sized clusters
- ... however still doesn't answer many real-world cases
- ... and the distribution of the clusters between the nodes has the same skewed balance

What can we do?

UNIVERSITY OF CAMBRIDGE

We could also use approximate inference with Metropolis–Hastings corrections, splitting the cluster representation among the nodes

- A recent attempt is presented in [Chang and Fisher III, 2013]:
  - Data is decoupled for each finite K (number of components of the DP) conditioned on the probability of each component,
  - This gives approximate inference with a finite mixture model,
  - The sampler transitions between subspaces of possible distributions (finite mixture models with different K) via split-merge Metropolis–Hastings proposals,
  - The split proposals depend linearly on α, while the merge proposals depend linearly on α^{-1}.
- Suitable for the case *when the posterior is known in advance and the initialisation can reflect that*...

What can we do?

UNIVERSITY OF CAMBRIDGE

- This is because for different values of α we might accept no split/merge proposals:

	100 initial clusters		1 initial cluster	
	splits	merges	splits	merges
$\alpha = 0.2$	**0.00**	1.48	**0.03**	0.00
$\alpha = 1$	0.01	1.29	0.03	0.00
$\alpha = 5$	0.32	**0.16**	0.15	**0.00**

- However we suspect that by introducing additional random moves that depend on α in an inverse way this limitation might be overcome.

What can we do?

UNIVERSITY OF CAMBRIDGE

Develop better approximate inference

- Current approach uses Gibbs sampling after distributing the data evenly across the different nodes and in the global step we sync. the state of the nodes (Asuncion, Smyth, and Welling [2008])
- Was reported by Williamson et al., [2013] to have slow convergence

What can we do?

UNIVERSITY OF CAMBRIDGE

And finally, don't use the Dirichlet process

- Recently shown that the Dirichlet process is inconsistent in the number of cluster
- An alternative distribution for clustering has been suggested: using a Poisson distribution mixture of Dirichlet distributions
- Might open the door for more efficient parallel inference

Conclusions

- Scaling up inference for the Dirichlet process is still an open problem...
- ... which has to be solved if we want it to be used in industry and real-world applications!