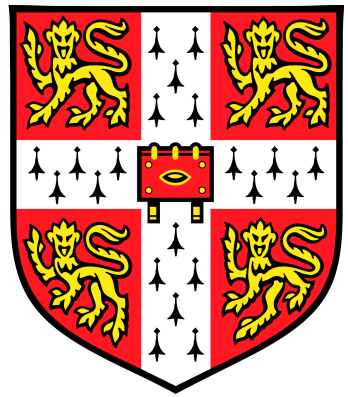
Diffusion trees as priors

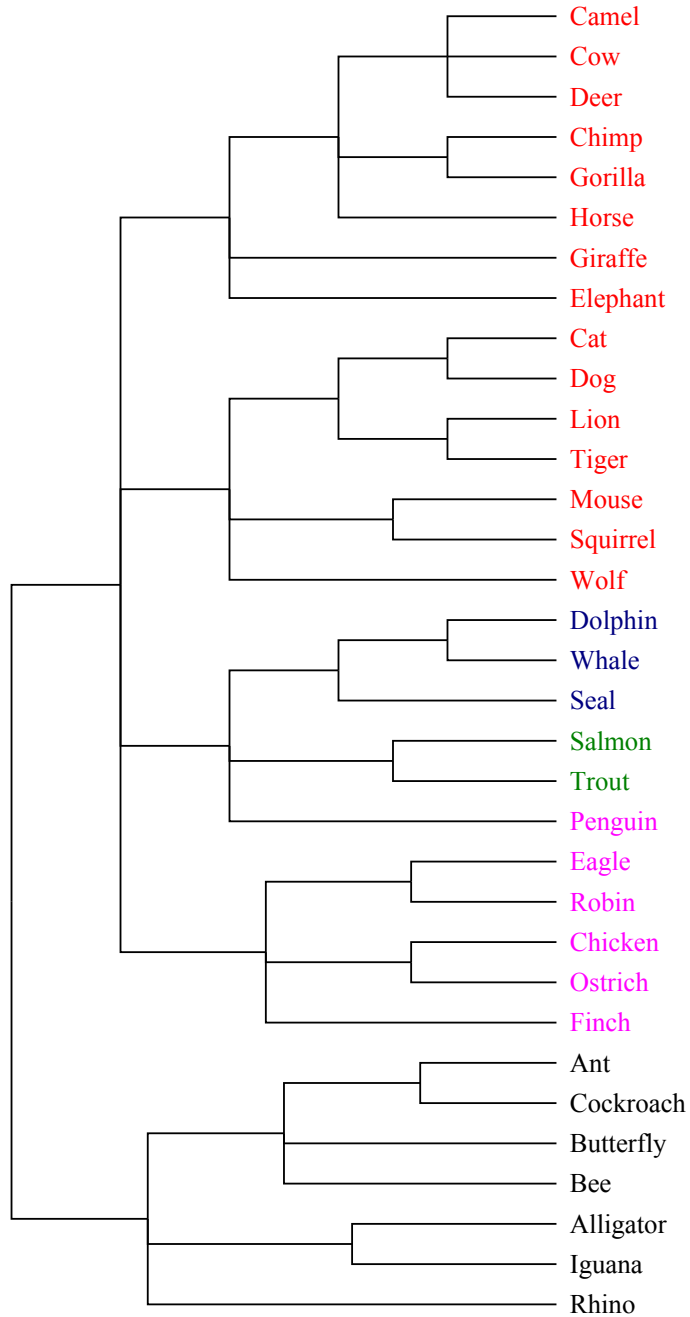
David A. Knowles
University of Cambridge

April 20, 2012



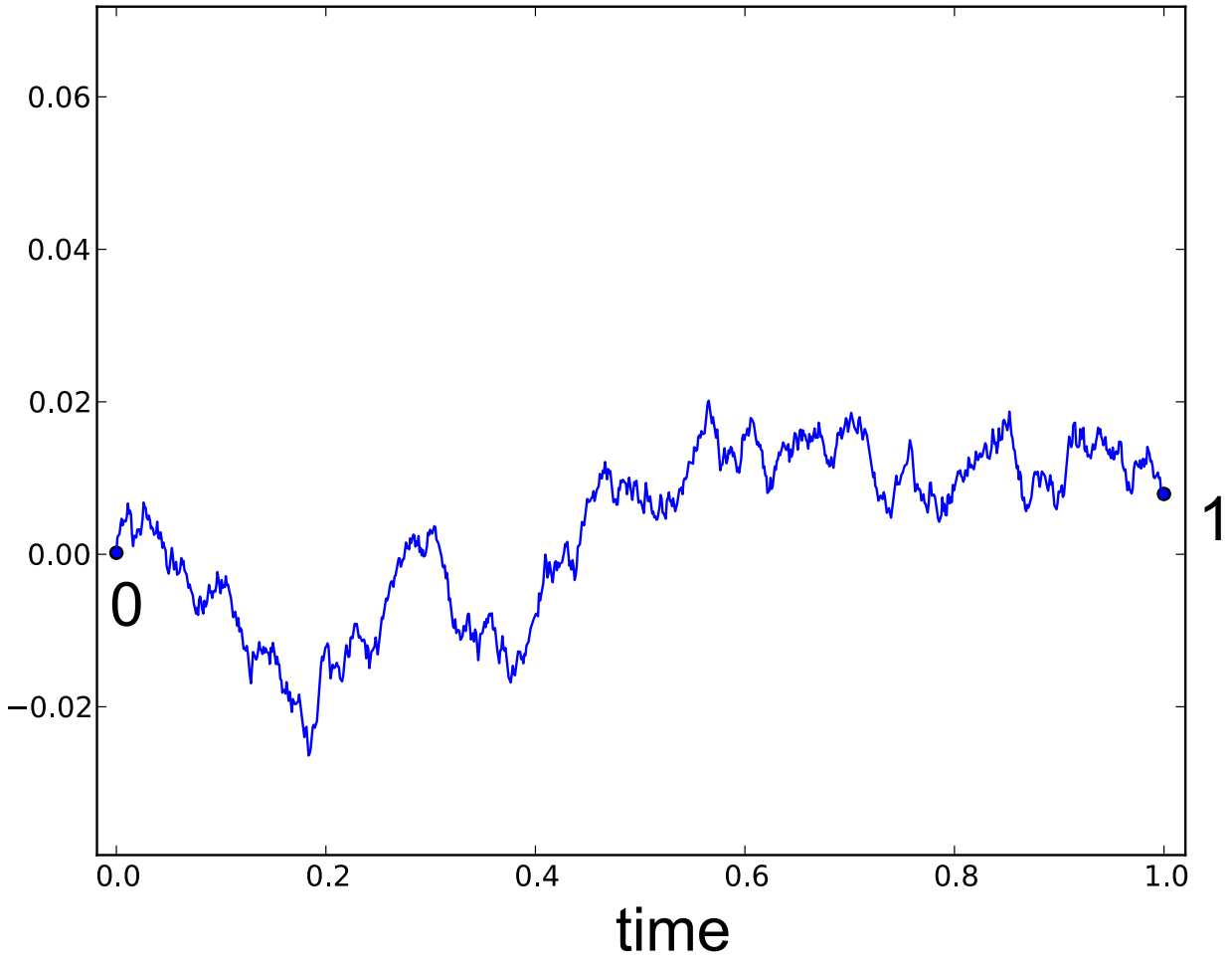
Motivation

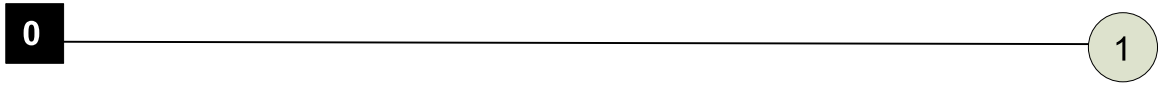
- True hierarchies
- Parameter tying
- Visualisation and interpretability

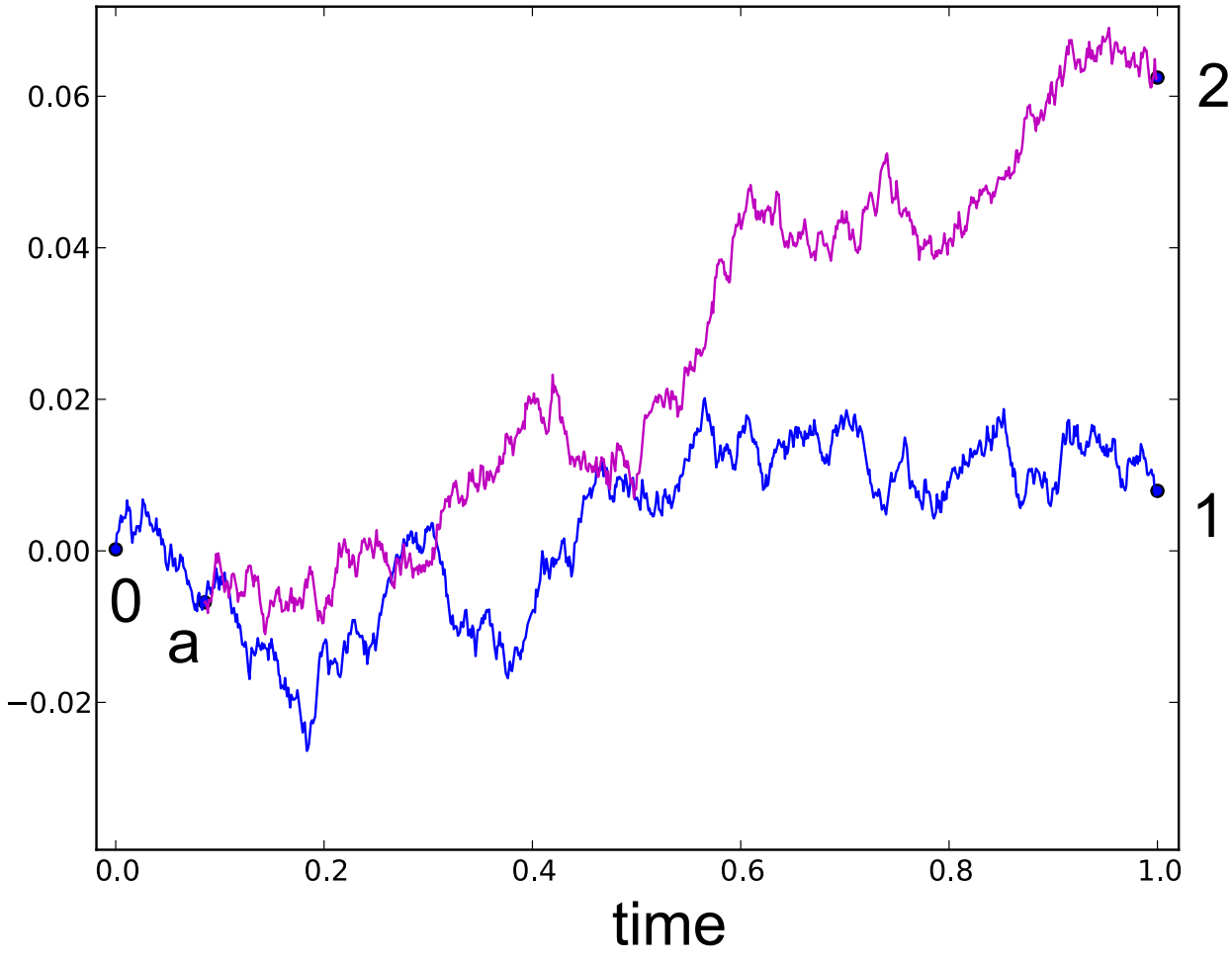


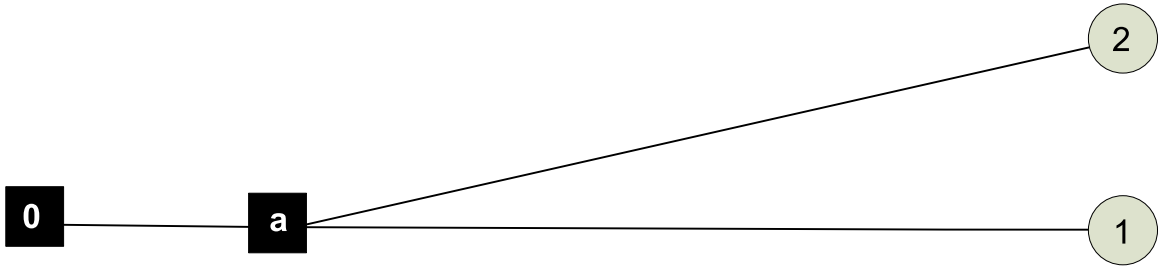
Setup: Pitman-Yor diffusion tree

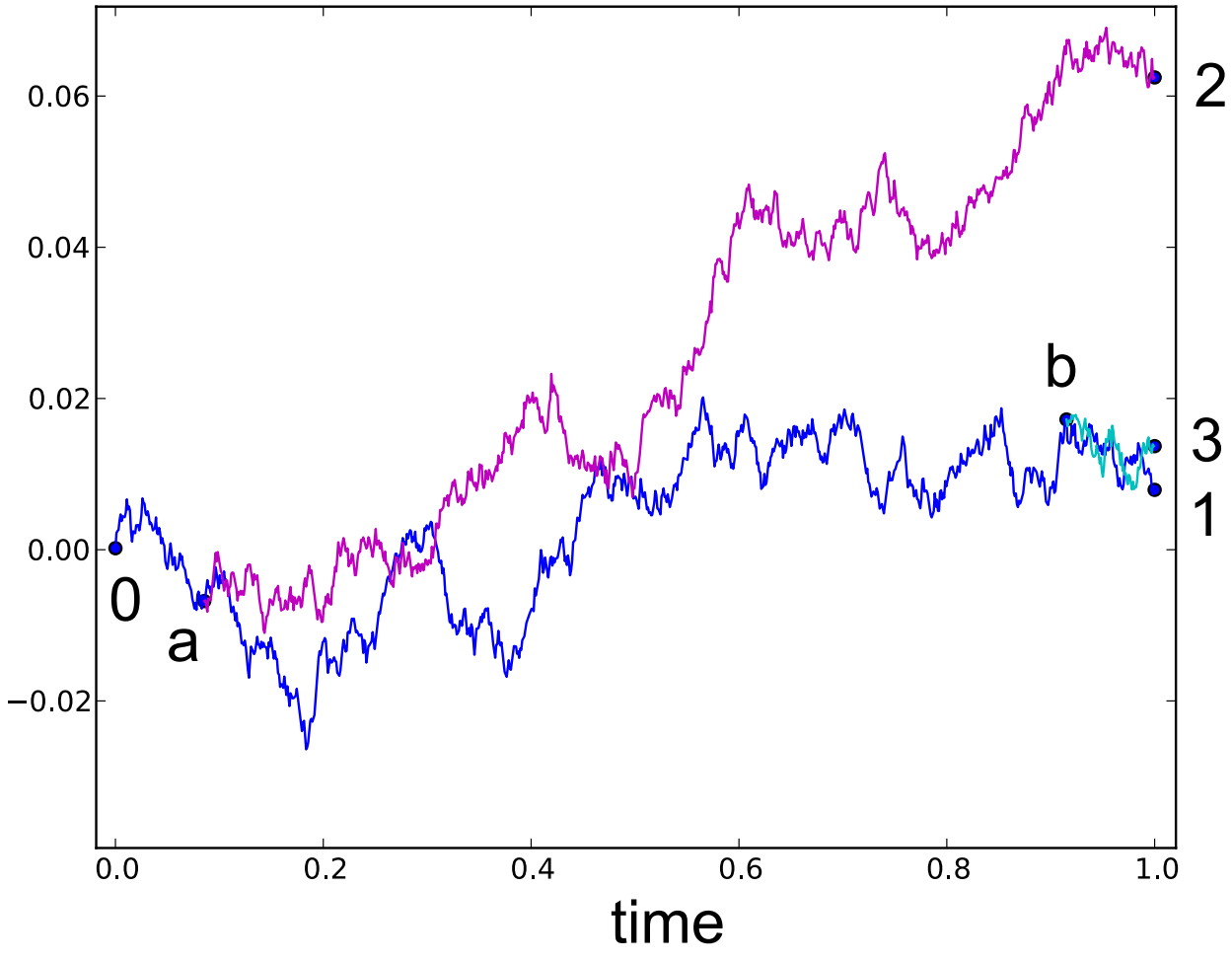
- Generalisation of the Dirichlet Diffusion Tree (Neal, 2001)
- A top-down generative model for trees over N datapoints $x_1, x_2, \cdots, x_N \in \mathbb{R}^D$
- Points start at "time" $t = 0$ and follow Brownian diffusion in a D-dimensional Euclidean space until $t = 1$, where they are observed
- Model based approach allows uncertainty over trees to be quantified, and integration into larger models

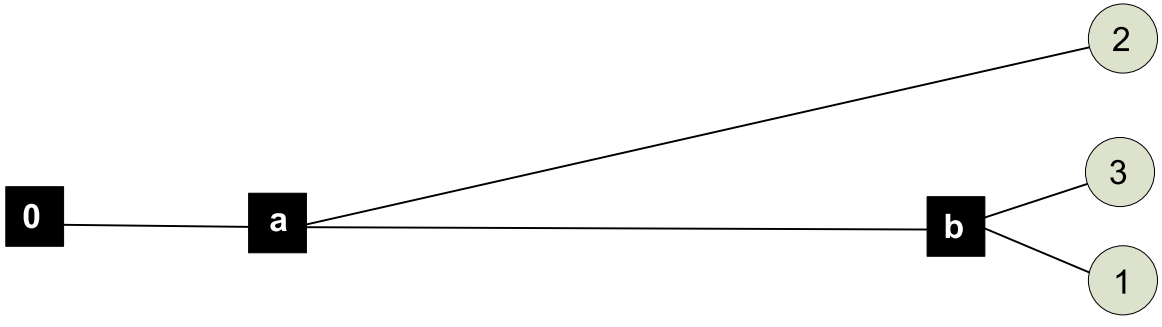

0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0
time
0
1

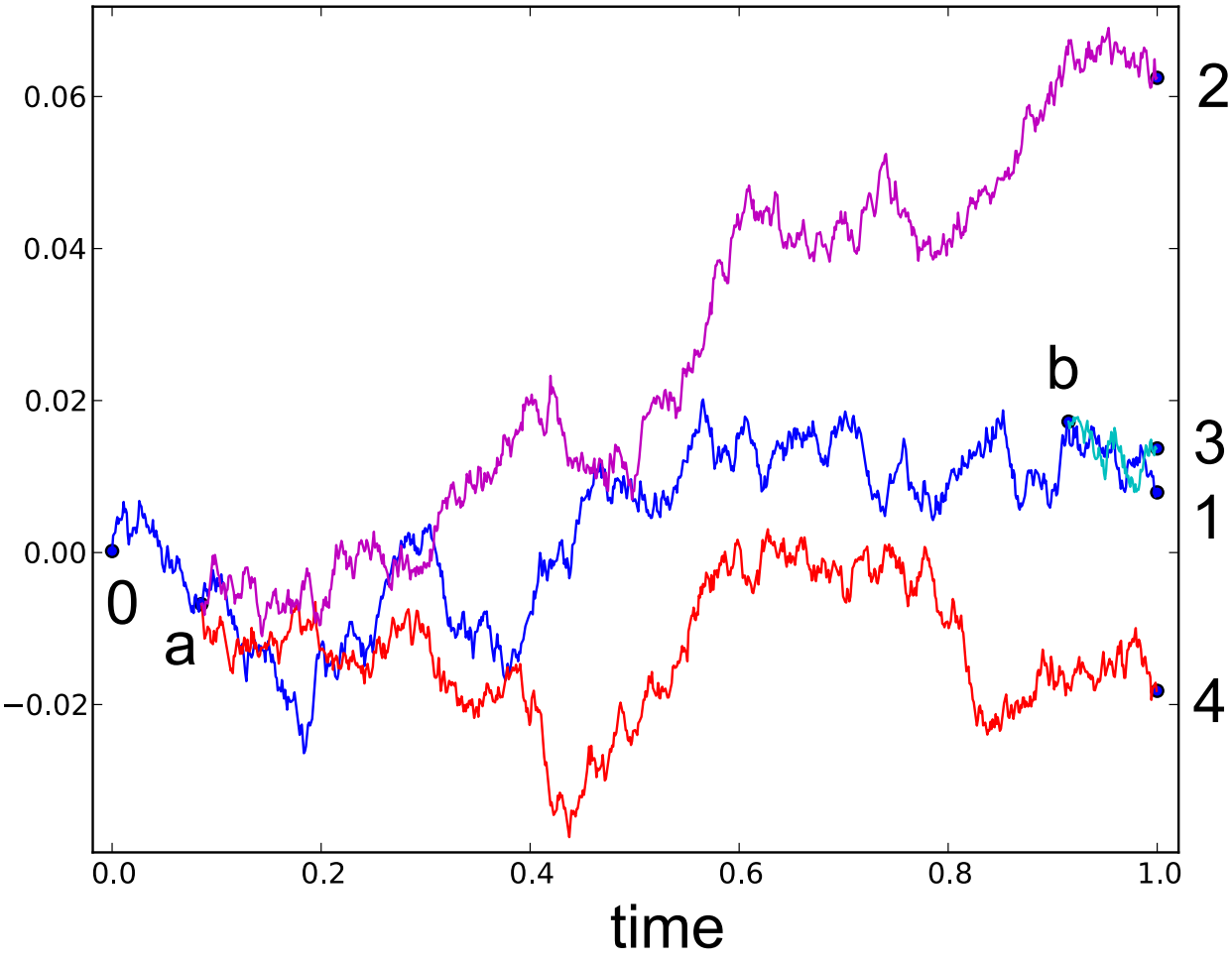

0
1

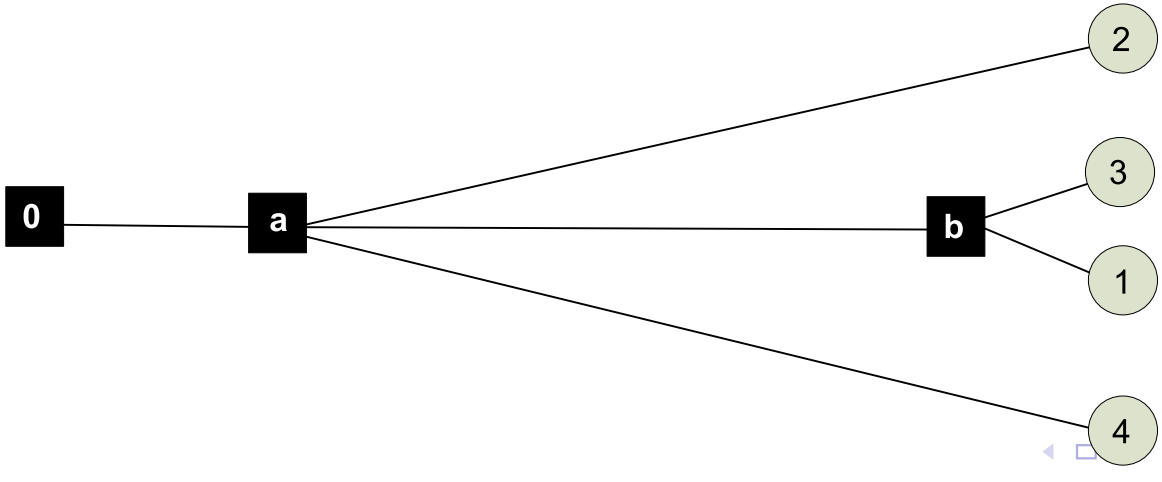

0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0
time
0
a
2
1


0
a
2
1


0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0
time
0
a
b
2
3
1


0
a
b
2
3
1


0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0
time
0
a
b
2
3
1
4


0
a
b
2
3
1
4

Branching probability

At a branch point,

$$P(\text{following branch } k) = \frac{n_k - \alpha}{m + \theta},$$
$$P(\text{diverging}) = \frac{\theta + \alpha K}{m + \theta},$$

where

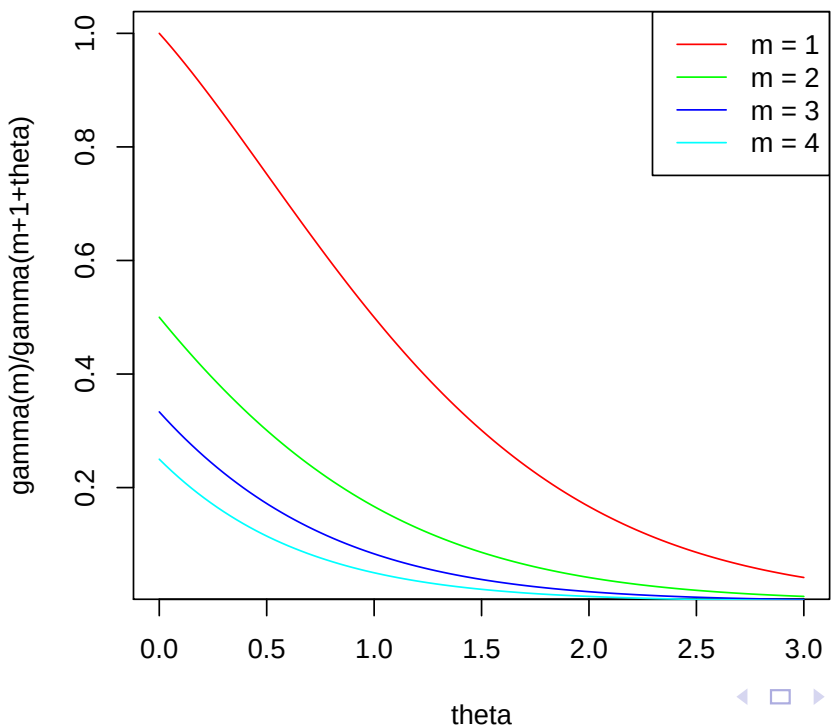
- n_k: number of samples which previously took branch k
- K: current number of branches from this branch point
- $m = \sum_{k=1}^{K} n_k$: number of samples which previously took the current path
- θ, α are hyperparameters

Probability of diverging

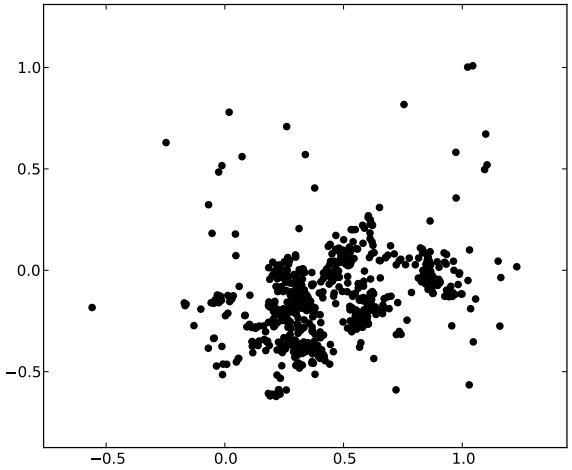
To maintain exchangeability, probability of diverging becomes

$$P\left(\begin{array}{c}\text{diverging}\\ \text{in } [t, t+dt]\end{array}\right) = \frac{a(t)\Gamma(m-\alpha)dt}{\Gamma(m+1+\theta)}$$

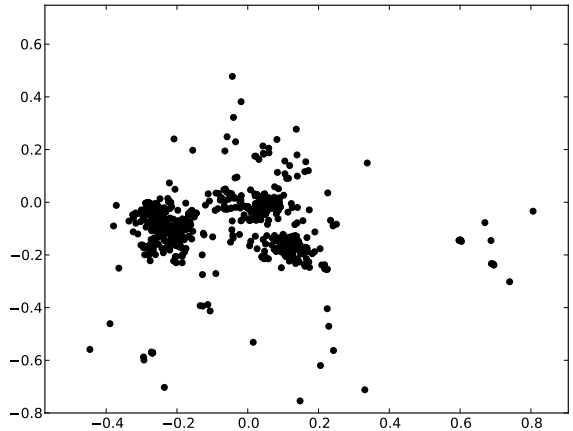
where we use $a(t) = c/(1-t)$. Note that $\int_{[0,1]} a(t)dt = \infty$ gives divergence before $t = 1$ a.s., and therefore a continuous distribution on x.



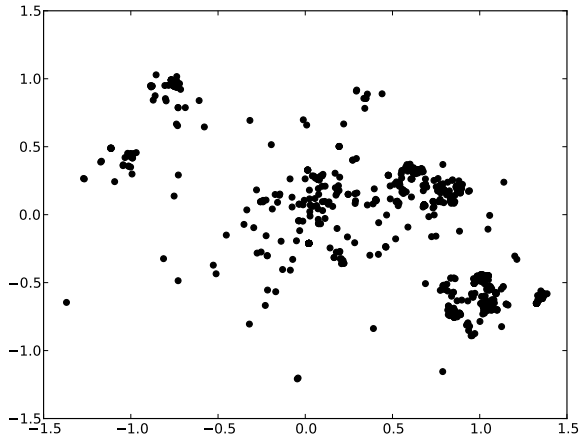
Example draws in $\mathbb{R}^2$



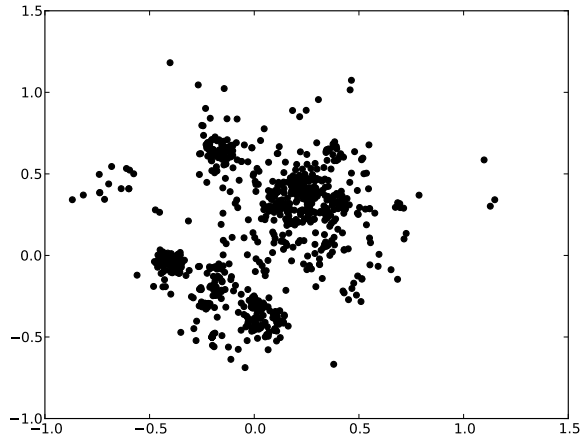
(a) $c = 1, \theta = 0, \alpha = 0$ (DDT)



(b) $c = 1, \theta = 0.5, \alpha = 0$



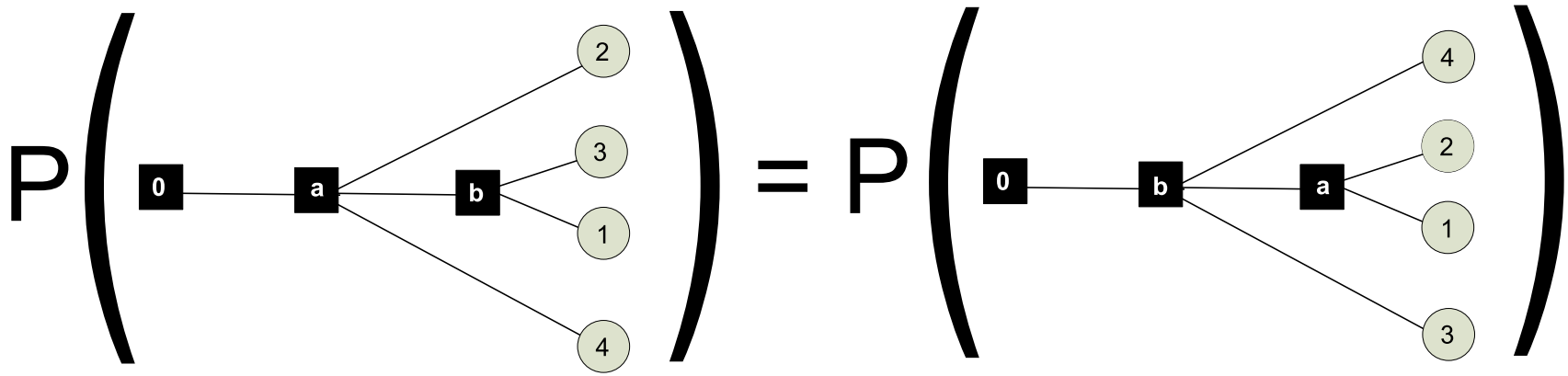
(c) $c = 1, \theta = 1, \alpha = 0$



(d) $c = 3, \theta = 1.5, \alpha = 0$

Lemma

The probability of generating a specific tree structure, divergence times, divergence locations and corresponding data set is invariant to the ordering of data points.



Proof.

Probability of tree structure:

$$\prod_{[ab]\in\text{internal edges}} \frac{\prod_{k=3}^{K_b}[\theta+(k-1)\alpha]\prod_{l=1}^{K_b}\Gamma(n_l^b-\alpha)}{\Gamma(m(b)+\theta)\Gamma(1-\alpha)^{K_b-1}} \tag{1}$$

Probability of divergence times:

$$\prod_{[ab]\in\text{internal edges}} a(t_b)\exp\left[(A(t_a)-A(t_b))H_{m(b)-1}^{\theta,\alpha}\right]$$

where we define $H_n^{\theta,\alpha}=\sum_{i=1}^{n}\frac{\Gamma(i-\alpha)}{\Gamma(i+1+\theta)}$.

Probability of node locations:

$$\prod_{[ab]\in\text{edges}} \mathrm{N}(x_b;x_a,\sigma^2(t_b-t_a)I)$$

None of these depend on the order of data points! □

Proposition

The Pitman-Yor Diffusion Tree defines an infinitely exchangeable distribution over data points.

Proof.

Summing over all possible tree structures, and integrating over all branch point times and locations, by Lemma 1 we have infinite exchangeability. □

Corollary

There exists a prior ν on probability measures on $\mathbb{R}^D$ such that the samples $x_1, x_2, \ldots$ generated by a PYDT are conditionally independent and identically distributed (iid) according to $\mathcal{F} \sim \nu$, that is, we can represent the PYDT as

$$PYDT(x_1, x_2, \ldots) = \int \left(\prod_i \mathcal{F}(x_i) \right) d\nu(\mathcal{F})$$

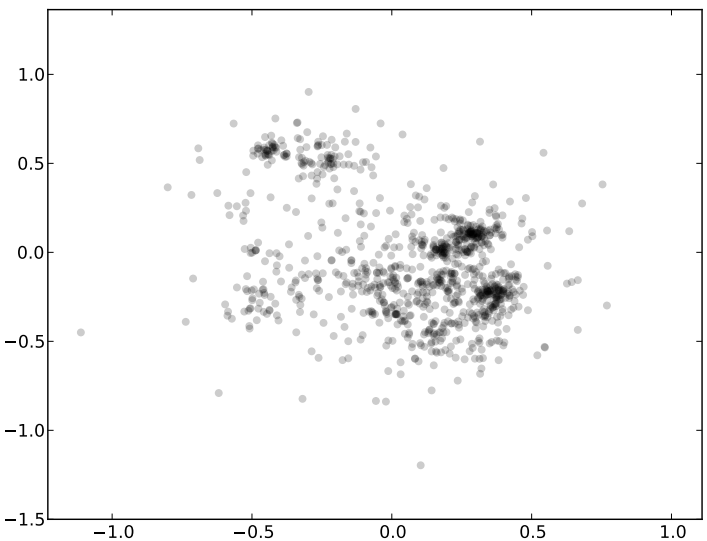
.

Proof.

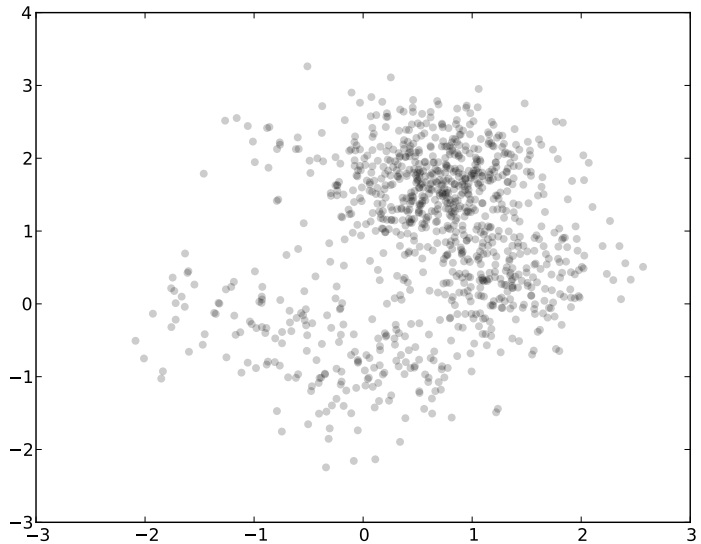
Since the PYDT defines an infinitely exchangeable process on data points, the result follows directly by de Finetti's Theorem. □

Comparing to the DPM

It is difficult for the DPM to model fine structure: it has to choose between using many small clusters whose parameters will be difficult to fit, or large clusters that would oversmooth the data.



(e) PYDT



(f) DPM

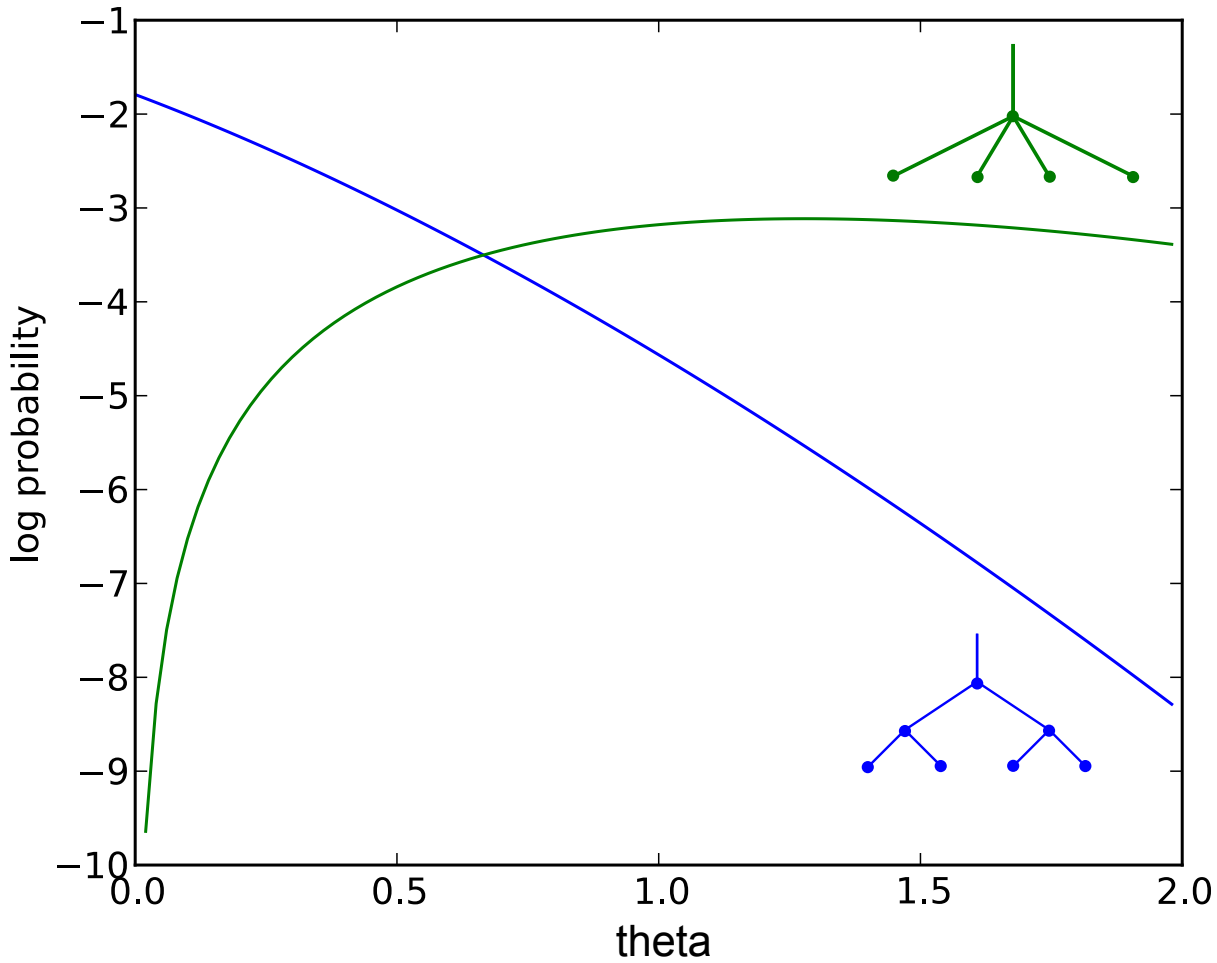
Parameter ranges

There are several valid ranges of the parameters (θ, α):

- $0 \leq \alpha < 1$ and $\theta > -2\alpha$. General multifurcating case with arbitrary branching degree.
- $\alpha < 0$ and $\theta = -\kappa\alpha$ where $\kappa \in \mathbb{Z} \geq 3$ is the maximum outdegree of a node.
- $\alpha < 1$ and $\theta = -2\alpha$. Binary branching, and specifically the DDT for $\alpha = \theta = 0$. A parameterised family of priors proposed by MacKay and Broderick (2007).
- $\alpha = 1$ gives instantaneous divergence so data points are independent.

Effect of varying θ

Fix $\alpha = 0$. Large θ: flat clusterings. Small θ: hierarchical clusterings.



Tree balance

Binary branching parameter range: $\alpha < 1$ and $\theta = -2\alpha$.
Probability of going left is

$$\frac{n_l - \alpha}{n_l + n_r - 2\alpha} \tag{2}$$

This reinforcement is equivalent to hypothesising a per node "probability of going left", with prior

$$p \sim \text{Beta}(-\alpha, -\alpha) \tag{3}$$

Conditioning on the previous data points

$$p|n_r, b_l \sim \text{Beta}(n_l - \alpha, n_r - \alpha) \tag{4}$$

Thus marginalising out p gives (2). For α close to 1, p will be close to 0 or 1, so the tree will be very unbalanced. For $\alpha \to -\infty$, p will be close to $\frac{1}{2}$ giving balanced trees.

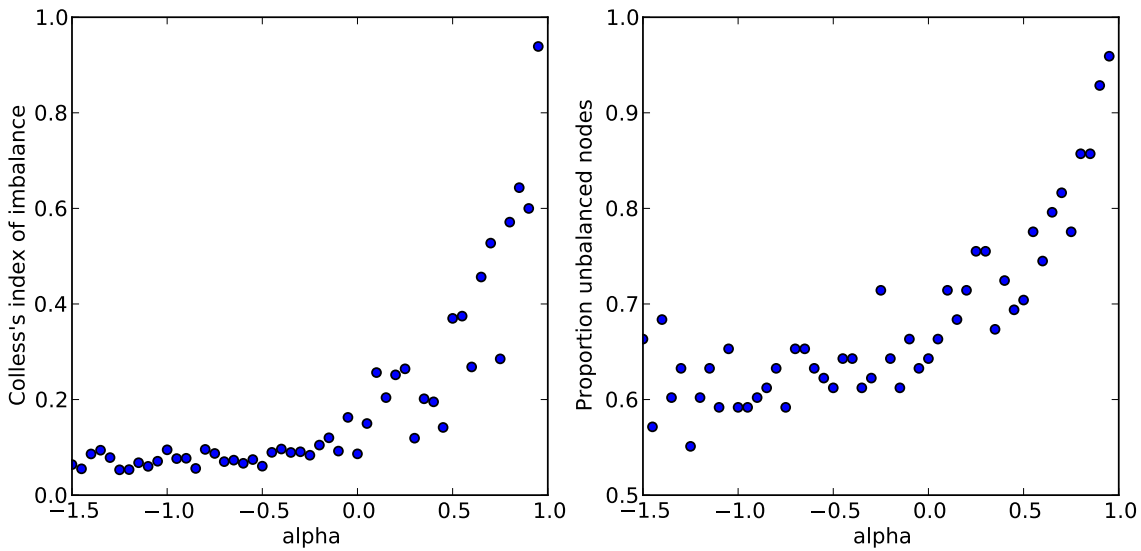
Tree balance

A measure of tree imbalance is Colless's I (Colless, 1982)

$$I = \frac{2}{(n-1)(n-2)} \sum_{a \in \mathcal{T}} |l(a) - r(a)| \tag{5}$$

The normalised no. of unbalanced nodes in a tree, J (Rogers, 1996), i.e.

$$J = \frac{1}{(n-2)} \sum_{a \in \mathcal{T}} (1 - \mathbb{I}[l(a) = r(a)]) \tag{6}$$



Generalises the Dirichlet diffusion tree

$\theta = \alpha = 0$ recovers the DDT of Neal (2001).
Probability of diverging off a branch

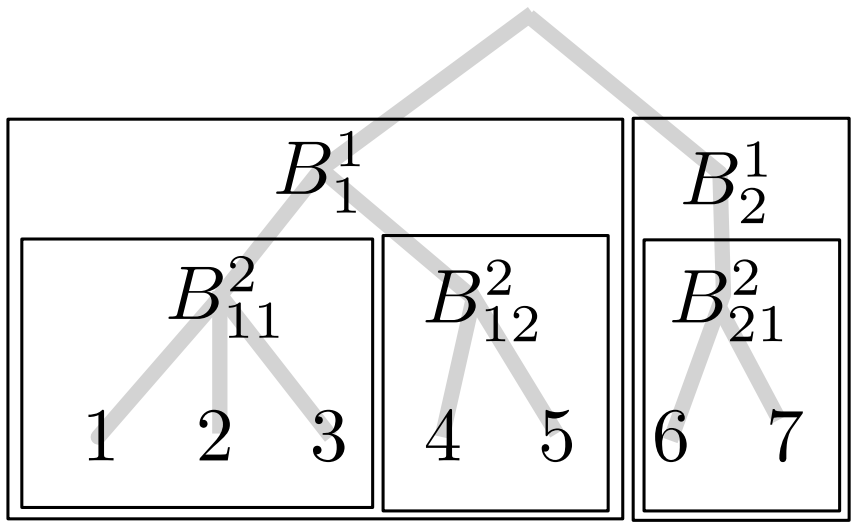
$$\frac{a(t)\Gamma(m-0)dt}{\Gamma(m+1+0)} = \frac{a(t)(m-1)!dt}{m!} = \frac{a(t)dt}{m}, \tag{7}$$

Probability of following a branch at an existing branch point is proportional to the number of previous datapoints having followed that branch

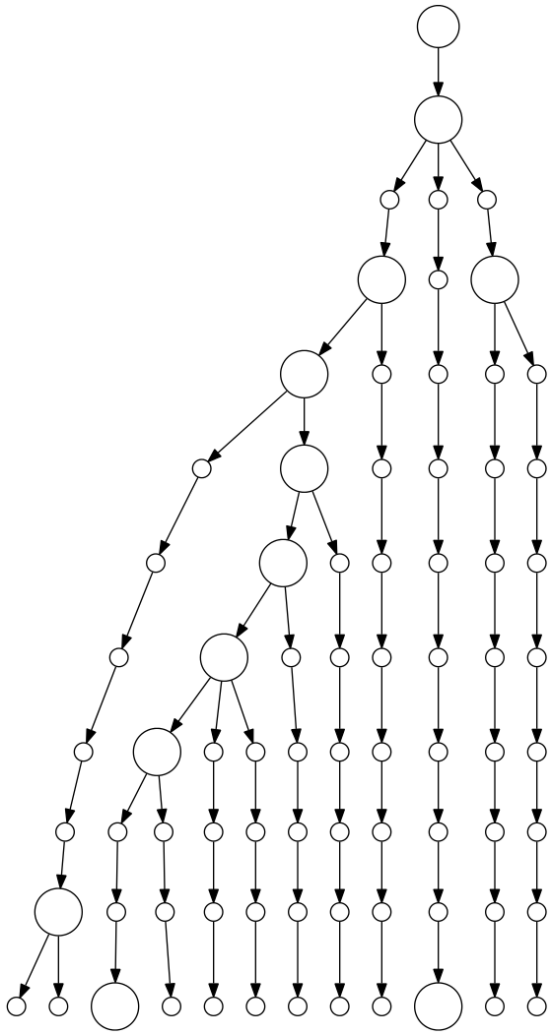
$$\frac{\prod_{l=1}^{K_b=2}\Gamma(n_l^b-0)}{\Gamma(m(b)+0)} = \frac{(n_1^b-1)!(n_2^b-1)!}{(m(b)-1)!}, \tag{8}$$

Nested CRP

- Distribution over hierarchical partitions
- Denote the K blocks in the first level as $\{B_k^1 : k = 1, ..., K\}$
- Partition these blocks with independent CRPs
- Denote the partitioning of B_k^1 as $\{B_{kl}^2 : l = 1, ..., K_k\}$
- Recurse for S iterations, forming a S deep hierarchy



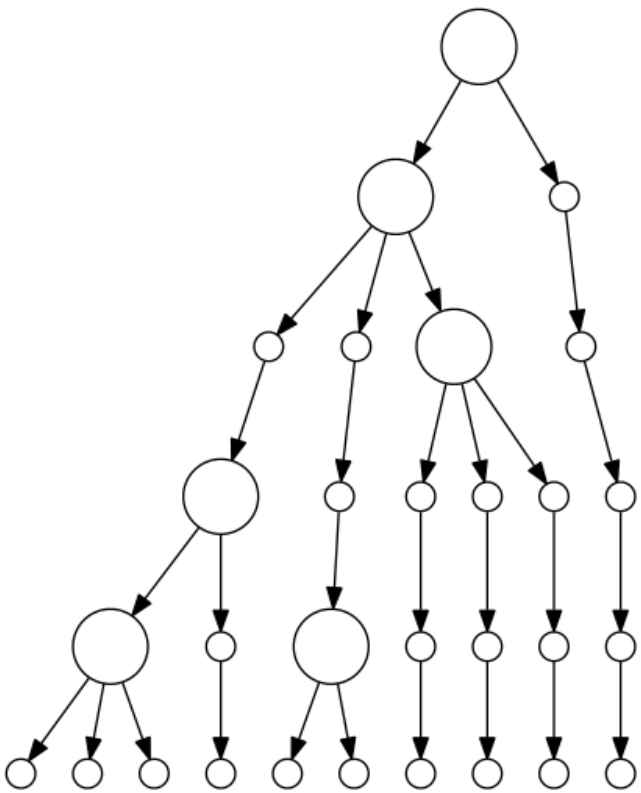
Nested CRP



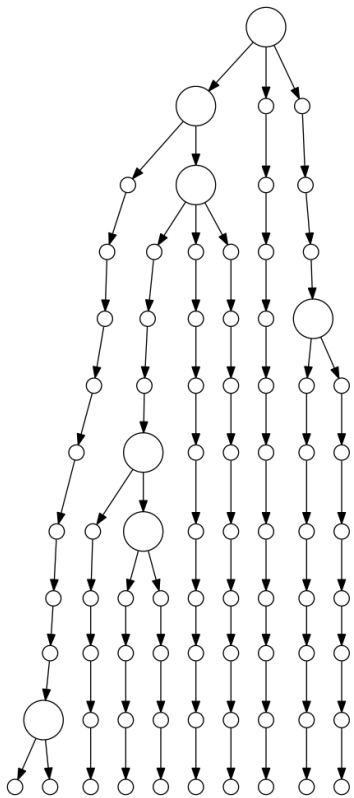
A draw from a $S = 10$-level nested Chinese restaurant process with 15 leaves.

Continuum limit of a nested CRP

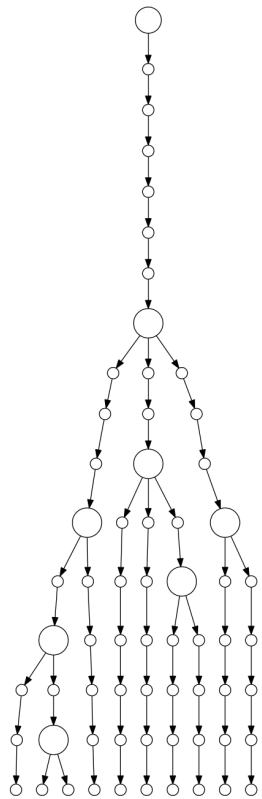
Associate each level s in an S-level nCRP with "time" $t_s = \frac{s-1}{S} \in [0, 1)$, and let the concentration parameter at level s be $a(t_s)/S$, where $a : [0, 1] \mapsto \mathbb{R}^+$. Taking the limit $S \to \infty$ recovers the Dirichlet Diffusion Tree with divergence function $a(t)$.



$S = 5$



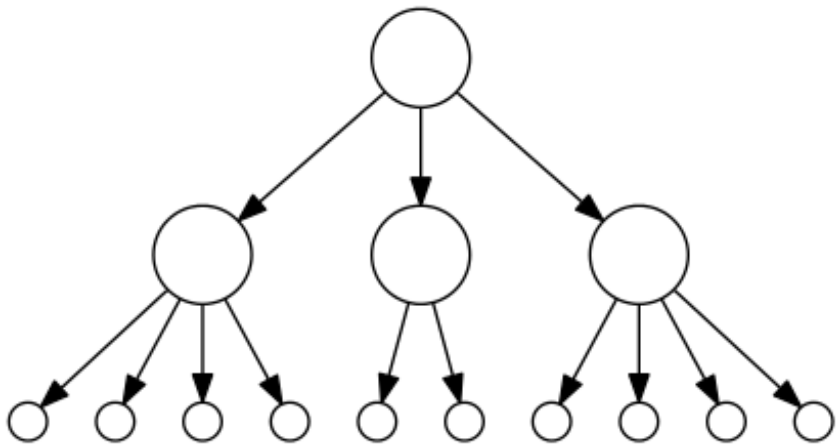
$S = 10$



$S = 15$

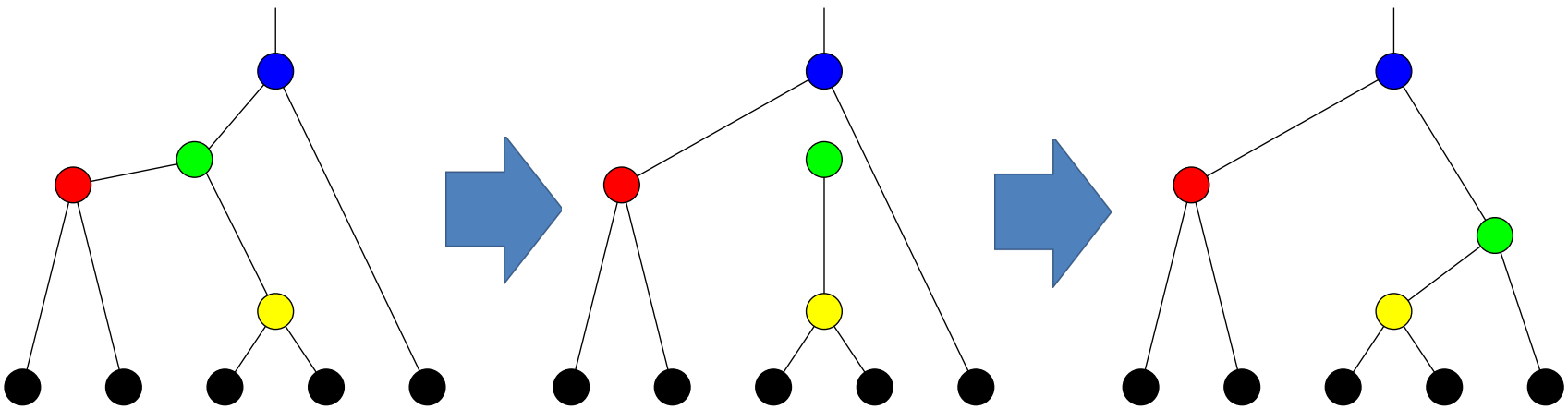
Other properties of the PYDT

- Generalisation of DP mixture of Gaussians (with specific variance structure)
- Prior over tree structures is a multifurcating Gibbs fragmentation tree (McCullagh et al., 2008), the most general Gibbs type, Markovian, exchangeable, consistent distribution over trees



Inference: MCMC

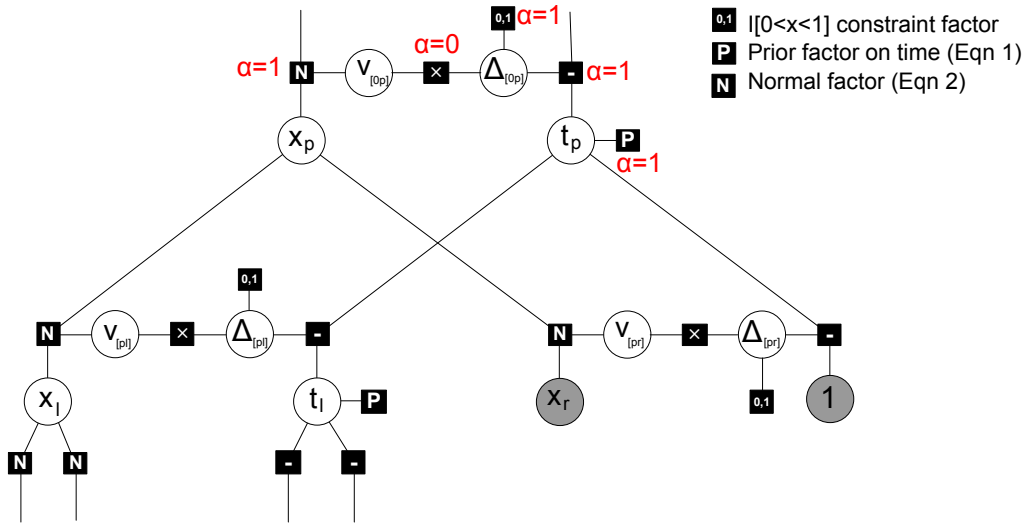
- Not straightforward to extend Neal's slice sampling moves because of atoms in the prior at existing branches
- Propose new subtree locations from the prior: slow!
- Working on Gibbs sampling algorithm using uniformisation ideas from (Rao and Teh, 2011) (with Vinayak Rao)



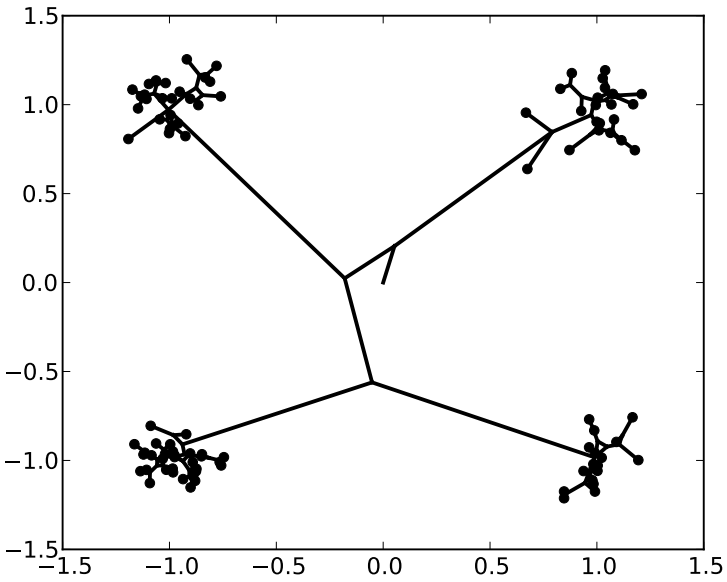
Inference: EM with greedy search

- Power EP or EM to calculate marginal likelihood for a tree
- Use this to drive sequential tree building and search over tree structures
- Warm start inference
- Use local evidence contribution to propose $L = 3$ "good" re-attachment locations

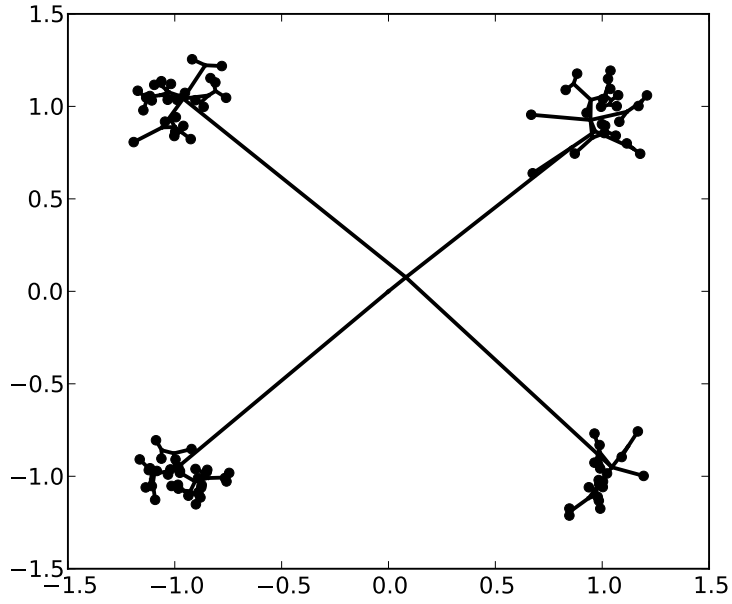
A straightforward extension of the algorithm we presented at ICML 2011 for the DDT (K., 2011)



Results: toy data



(g) DDT



(h) PYDT

Figure: Optimal trees learnt by the greedy EM algorithm for the DDT and PYDT on a synthetic dataset with $D = 2, N = 100$.

Results: Macaques skull measurements

$N_{\text{train}} = 200, N_{\text{test}} = 28, D = 10$ Adams et al. (2008)

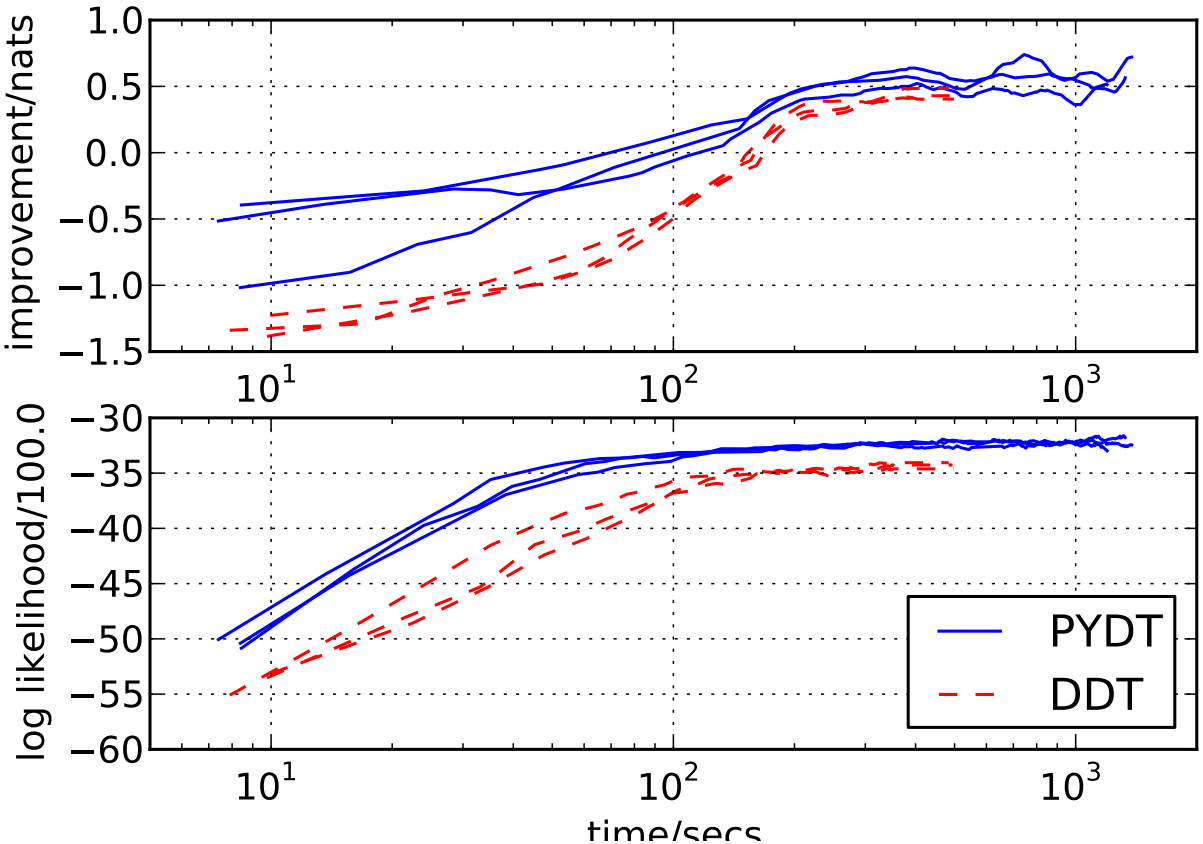


Figure: Density modeling of the $D = 10, N = 200$ macaque skull measurement dataset of Adams et al. (2008). *Top*: Improvement in test predictive likelihood compared to a kernel density estimate. *Bottom*: Marginal likelihood of current tree. The shared x-axis is computation time in seconds.

Results: Animal species

- 33 animal species from Kemp and Tenenbaum (2008)
- 102-dimensional binary feature vectors relating to attributes (e.g. being warm-blooded, having two legs)
- Probit regression

Results: Animal species

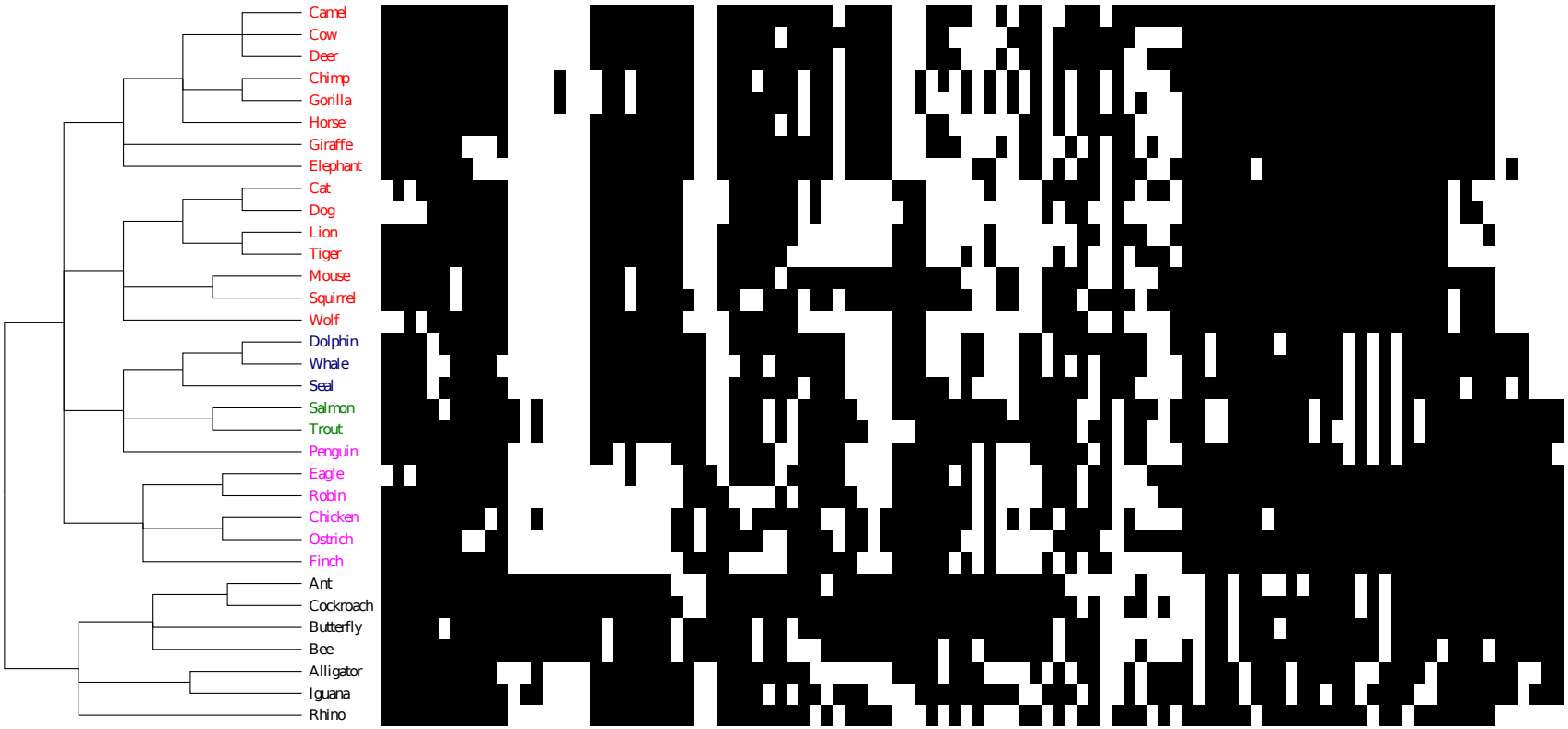


Figure: Tree structure learnt for the animals dataset of Kemp and Tenenbaum (2008).

Other priors over tree structures used in ML

- Kingman's coalscent (KC) (Kingman, 1982; Teh et al., 2008). Points coalesce together rather than fragmenting as in the DDT/PYDT. KC is in a sense the dual process to the DDT, a fact used in Teh et al. (2011).
- Fixed number of generations and individuals per generation where each child chooses its parent (Williams, 2000), a discretisation of KC.
- Nested CRP itself (Blei et al., 2010; Steinhardt and Ghahramani, 2012). How to choose when to stop?
- Tree structured stick breaking (Adams et al., 2010). Extends the stick breaking construction of the CRP to the nested CRP, and adds a per node stopping probability.

Infinite Latent Attributes model for network data (with Konstantina Palla)

- Existing network models explain a "flat" clustering structure
- ILA has features that are partitioned into disjoint groups (subclusters)
- Generalises the IRM (Kemp and Tenenbaum, 2006), LFIRM (Miller et al., 2009), and MAG (Kim and Leskovec, 2011)
- Excellent empirical performance in link prediction

Generative model:

$$\mathbf{Z}|\alpha \sim \text{IBP}(\alpha)$$

$$\mathbf{c}^{(m)}|\gamma \sim \text{CRP}(\gamma)$$

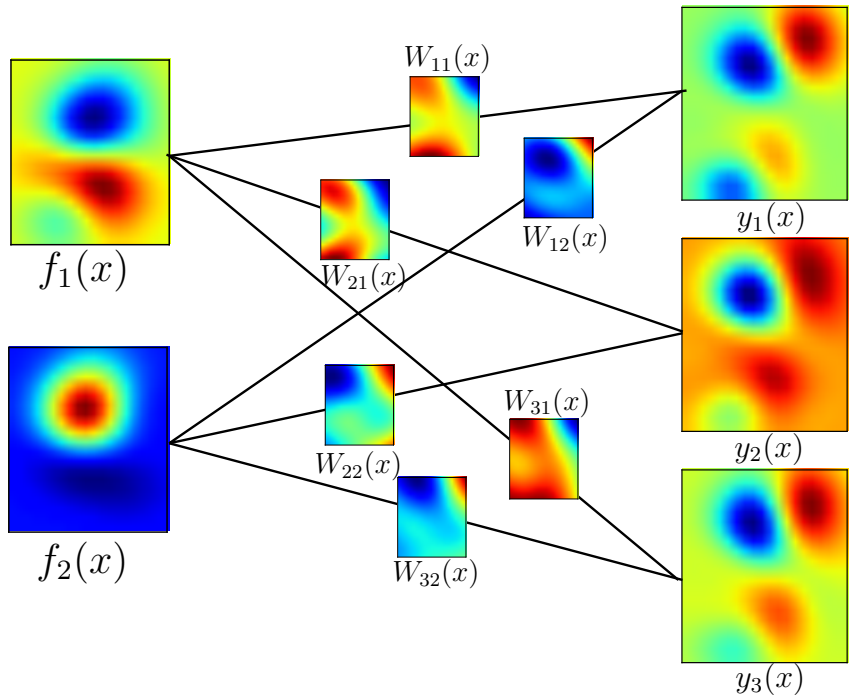
$$w_{kk'}^{(m)}|\sigma_w \sim N(0, \sigma_w^2)$$

$$\Pr(r_{ij} = 1|\mathbf{Z}, \mathbf{C}, \mathbf{W}) = \sigma\left(\sum_m z_{im} z_{jm} w^{(m)}_{c_i^m c_j^m} + s\right).$$

Gaussian Process Regression Networks (with Andrew Wilson)

- Multivariate heteroskadistic regression with covariate dependent signal and noise correlations
- Tractibility of Gaussian processes and multitask advantages of neural networks

$$W(x)_{ij} \sim \mathcal{GP}(0, k_w)$$
$$f_i(x) \sim \mathcal{GP}(0, k_f + \sigma_f^2 \delta)$$
$$\mathbf{y}(x) \sim N(W(x)\mathbf{f}(x), \sigma_y^2 I)$$



Future/ongoing work

- Improved MCMC: uniformisation, slice sampling subtree locations
- Hierarchical structured states in an infinite HMM (e.g. for unsupervised part of speech tagging, modelling genetic variation)
- Topic modelling: hierarchy over topic specific distributions over words
- How to summarise posterior samples?
- Time varying tree structures?

Bibliography I

Adams, R., Murray, I., and MacKay, D. (2008). The Gaussian process density sampler. In *Advances in Neural Information Processing Systems*, volume 21. MIT Press.

Adams, R. P., Ghahramani, Z., and Jordan, M. I. (2010). Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing (NIPS) 23*.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57.

Colless, D. (1982). Phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, 31(1):100–104.

Kemp, C. and Tenenbaum, J. B. (2006). Learning systems of concepts with an infinite relational model. In *21st National Conference on Artificial Intelligence*.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. In *Proceedings of the National Academy of Sciences*, volume 105(31), pages 10687–10692.

Bibliography II

Kim, M. and Leskovec, J. (2011). Modeling social networks with node attributes using the multiplicative attribute graph model. In *UAI*.

Kingman, J. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235 – 248.

Knowles, D. A., Gael, J. V., and Ghahramani, Z. (2011). Message passing algorithms for Dirichlet diffusion trees. In *Proceedings of the 28th Annual International Conference on Machine Learning*.

Knowles, D. A. and Ghahramani, Z. (2011). Pitman-Yor diffusion trees. In *The 28th Conference on Uncertainty in Artificial Intelligence (to appear)*.

MacKay, D. and Broderick, T. (2007). Probabilities over trees: generalizations of the Dirichlet diffusion tree and Kingman's coalescent. Website.

McCullagh, P., Pitman, J., and Winkel, M. (2008). Gibbs fragmentation trees. *Bernoulli*, 14(4):988–1002.

Miller, K., Griffiths, T., and Jordan, M. (2009). Nonparametric latent feature models for link prediction. In *NIPS*.

Bibliography III

Neal, R. M. (2001). Defining priors for distributions using Dirichlet diffusion trees. Technical Report 0104, Dept. of Statistics, University of Toronto.

Rao, V. and Teh, Y. W. (2011). Fast MCMC sampling for Markov jump processes and continuous time Bayesian networks. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*.

Rogers, J. S. (1996). Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic biology*, 45(1):99–110.

Steinhardt, J. and Ghahramani, Z. (2012). Flexible martingale priors for deep hierarchies. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Teh, Y. W., Blundell, C., and Elliott, L. T. (2011). Modelling genetic variations with fragmentation-coagulation processes. In *Advances in Neural Information Processing Systems (NIPS)*.

Teh, Y. W., Daumé III, H., and Roy, D. M. (2008). Bayesian agglomerative clustering with coalescents. *Advances in Neural Information Processing Systems*, 20.

Bibliography IV

Williams, C. (2000). A MCMC approach to hierarchical mixture modelling. *Advances in Neural Information Processing Systems*, 13.