Flexible Martingale Priors for Deep Hierarchies

Jacob Steinhardt and Zoubin Ghahramani

Summary

•We present a new family of Bayesian hierarchical models based on the nested Chinese restaurant process, and show that every completely exchangeable hierarchical model can be represented as a member of this family

•We do this by giving a criterion (the *martingale criterion*) that allows substantial generalization of the nested Chinese restaurant process beyond topic models

•Using this criterion, we construct infinitely deep hierarchical Dirichlet and beta processes

•Our construction circumvents issues present in the tree-structured stick-breaking model

Motivation

•Priors over tree structures are crucial for performing Bayesian hierarchical modeling

•To date, all proposals for priors over discrete trees have undesirable properties
- •Tree-structured stick-breaking has a constant depth under the prior
- •Nested Chinese restaurant processes are hard to extend beyond topic models
- •Dirichlet diffusion trees are designed for continuous, not discrete, data

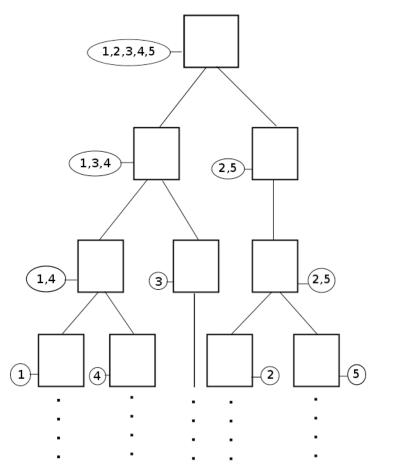
•To flexibly learn the structure of models such as hierarchical Dirichlet and beta processes, we need something better

•Our solution: build machinery to extend the nCRP to these models

Review: The nCRP

•The nested Chinese restaurant process, or nCRP, is a prior for Bayesian hierarchical models

•Each datum is associated with a path down the tree, as shown below (each of the numbers indicates a datum)

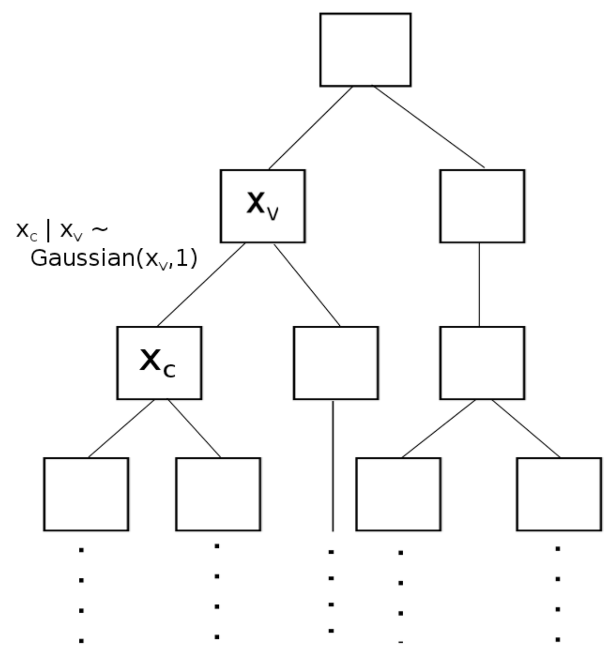


•If X is a datum and its path has reached v, the probability that it continues to a child c of v is given by a Chinese restaurant process

•The distribution over X given its path depends only on the latent parameters along the path

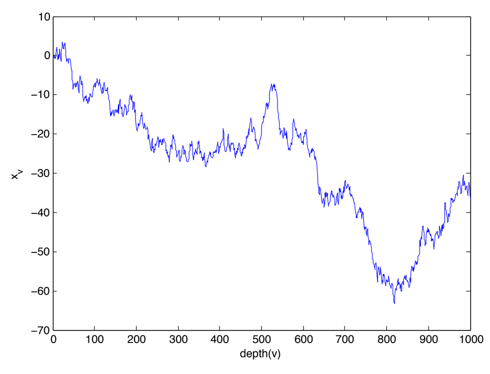
Example: An Infinite Random Walk

•Suppose that each node v contains a real number x_v and that for a child c of v, the distribution for x_c given x_v is $N(x_v,1)$



•Then the marginal distribution for x_v if v is at depth d is N(0,d)

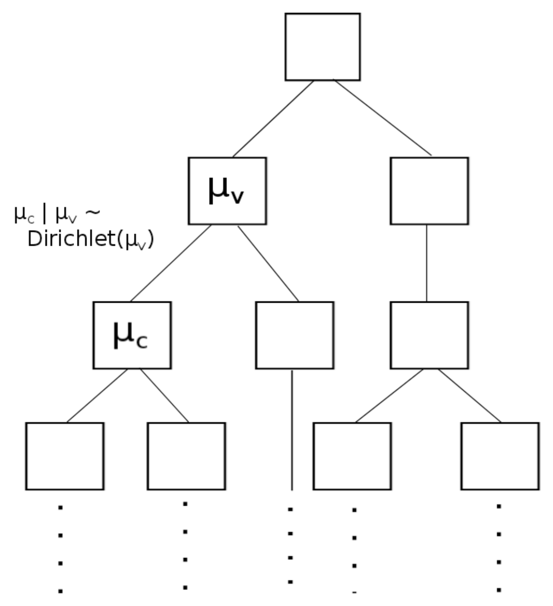
•This diverges as $d\to\infty$:



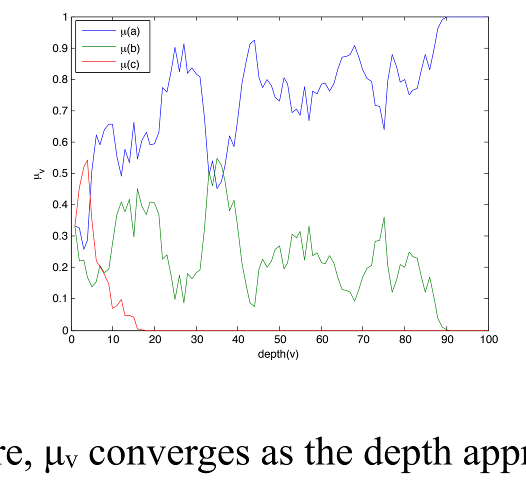
•Therefore, this model is not well-defined

Example: An Infinite Hierarchical Dirichlet Process

•Suppose that each node v contains a probability vector μ_v over 3 outcomes {a,b,c}, and that for a child c of v, the distribution for μ_c given μ_v is Dirichlet(μ_v(a),μ_v(b),μ_v(c))



•Then we can show that $\mu_v(x)$ converges to either 0 or 1 for each x



•Therefore, μ_v converges as the depth approaches ∞

•So, this defines a valid infinitely deep hierarchical Dirichlet process

The Martingale Criterion

•Both for the random walk and the hierarchical Dirichlet process, we have $E[\theta_c \mid \theta_v] = \theta_v$, where θ_v is the collection of parameters at node v

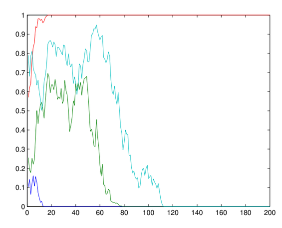
•This condition is called the *martingale criterion*
- •In general, ask that $E[f(\theta_c) \mid \theta_v] = f(\theta_v)$ for some f

•**Theorem (Doob):** All non-negative martingale sequences have a limit with probability 1.

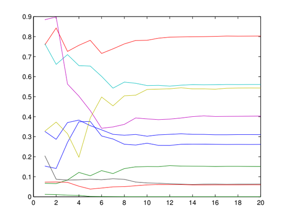
•**Corollary:** The infinite HDP converges. Furthermore, since the limiting variance for μ_c given μ_v must be 0, all the mass of μ_v concentrates on a single atom as the depth approaches ∞.

•**Remark:** The infinite random walk is not non-negative, which is why Doob's theorem does not apply.

•Examples of martingales:



Ex. 1: Parameters of a hierarchical Beta process. $\theta_{d+1} \mid \theta_d \sim \mathrm{Beta}(50\theta_d, 50(1-\theta_d))$



Ex. 2: A martingale given by $\theta_d=\alpha_d/(\alpha_d+\beta_d)$, where $\alpha_{d+1} \mid \alpha_d \sim \alpha_d+\mathrm{Gamma}(\alpha_d,1)$, $\beta_{d+1} \mid \beta_d \sim \beta_d+\mathrm{Gamma}(\beta_d,1)$.

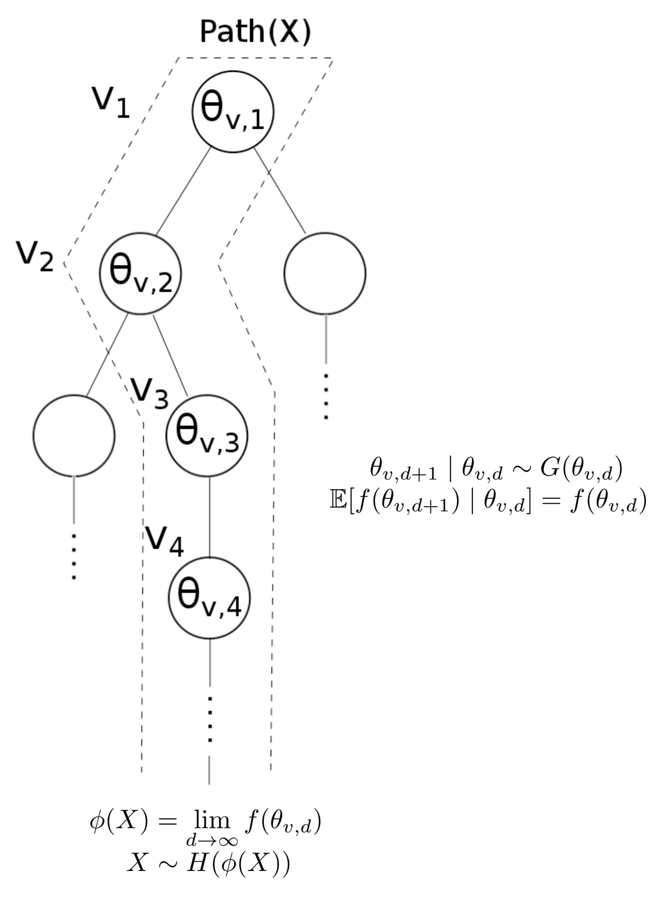
General Construction

•Take any desired prior over infinite trees (such as the nCRP), and let θ_v denote the latent parameter at node v

•Let $\theta_c \mid \theta_v \sim G(\theta_v)$ such that $E[f(\theta_c) \mid \theta_v] = f(\theta_v)$ for some non-negative function f

•For a datum X associated with a path $v_1,v_2,\ldots$, define $\varphi(X)$ as $\phi(X) = \lim_{d\to\infty} f(\theta_{v_d})$
- •By Doob's theorem, $\varphi(X)$ exists

•Sample X from some distribution $H(\varphi(X))$



•Example: infinite HDP
- •θ_v is the probability distribution at node v
- •$G(\theta) = \mathrm{Dirichlet}(\theta)$
- •$f(\theta) = \theta$
- •$H(\varphi) = \mathrm{Multinomial}(\varphi)$

Universality

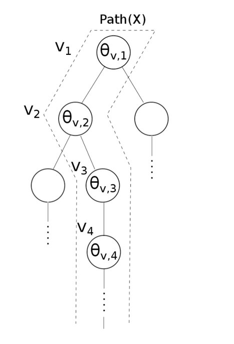
•A hierarchical model is *completely exchangeable* if, for a node c with parent v, the distribution for θ_c depends only on θ_v and the depth of c in the tree

•**Theorem:** for any completely exchangeable hierarchical model, there exists an alternate set of latent parameters $\tau_v \in T$ of at most countable dimension, and a function $f : T \to [0,1]^\infty$ such that $E[f(\tau_c) \mid \tau_v] = f(\tau_v)$

•Therefore, every completely exchangeable model can be realized using our construction
- •But the reparameterization in terms of τ might be inconvenient computationally

Tractability of Inference

•To perform inference, we need to compute the posterior over $\varphi(X)$ given just some prefix $v_1,v_2,\ldots,v_d$ of the path for X



To perform efficient inference, we need to sample $\phi(X) \mid \theta_{v,4}$.

•If $X \sim H(\varphi(X))$, just need sufficient statistics for H

•For discrete models (e.g. $H(\varphi) = \mathrm{Multinomial}(\varphi)$), $E[\varphi]$ is a sufficient statistic

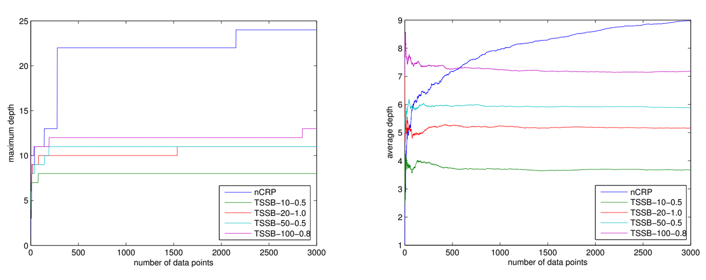
•Then the computation is easy: by the martingale condition, $E[f(\theta_c) \mid \theta_v] = f(\theta_v)$, so $E[\varphi \mid \theta_v] = f(\theta_v)$

Comparison to Tree-Structured Stick Breaking

•The main alternative proposal for Bayesian hierarchies is tree-structured stick-breaking

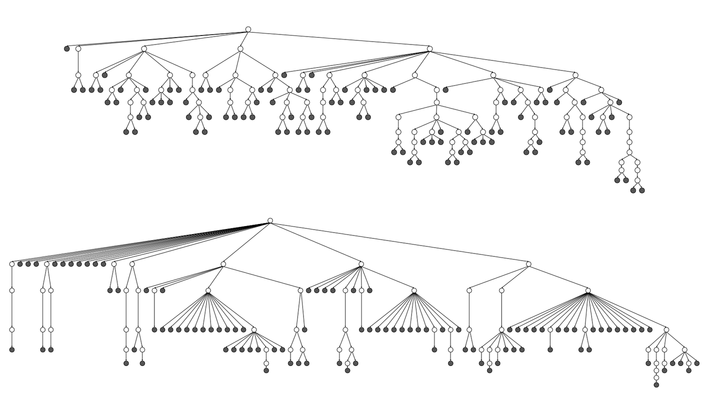
•To demonstrate the desirability of our construction, we perform an empirical comparison of the nCRP and TSSB
- •A theoretical analysis is given in the paper

•Comparison 1: depth of the tree as a function of data size



Note that the depth of the nCRP grows with the data, but the depth of TSSB does not.

•Comparison 2: samples from the prior for |Data|=100



Top: nCRP, bottom: TSSB; note that TSSB is very wide and shallow.