Pathological Properties of Deep Bayesian Hierarchies

Jacob Steinhardt and Zoubin Ghahramani

Overview

- In hierarchical models, we need a distribution over the latent parameters at each node
- Common solution: recursively draw from a distribution such as a Dirichlet process, beta process, Pitman-Yor process, etc.
- We show that for DP, BP, and GammaP, **this won't work for deep hierarchies**
- But!...Pitman-Yor is okay

Pathologies of the Gamma Distribution for Small Parameters

- For small settings of the parameters, samples from a gamma distribution can end up very close to zero.
- Lemma 1: If $y \sim \text{Gamma}(cx,c)$, and $cx \leq 1$, then

$$\mathbb{P}\left[y \leq \frac{1}{c}2^{-\frac{1}{cx}}\right] \geq \frac{1}{2e}$$

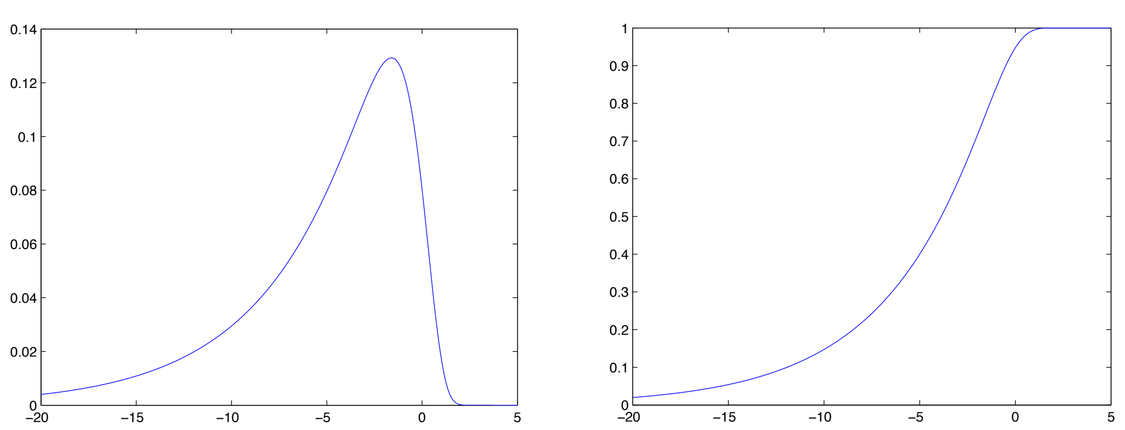


Figure 1: The pdf (left) and cdf (right) of log(Y), where Y ~ Gamma(0.2,1.0). Note the relatively large amount of probability mass placed on values as small as exp(-20).

- So, we should avoid choosing such small parameters. But **for deep hierarchies, this turns out to be unavoidable!**
- Gamma, beta, and Dirichlet sequences all decay towards 0 or 1 at a rate governed by a **tower of exponentials:** 1/e^(e^(e^(e^(...)))).

Why Call This Behavior Pathological?

- <u>Practically</u>: if the parameters converge extremely rapidly, then posterior inference is extremely sensitive to parameter values deep in the tree, which are too small to represent accurately on a computer
  - The difference between a parameter value of 0, $10^{-1000000000}$, and 10^{-100} matters significantly to the conditional distribution of a parameter 3 levels up
- <u>Philosophically</u>: as Bayesians, we would never report confidences as high as exp(exp(...(1))), so ours models should not, either.

Convergence of Martingale Sequences

- Consider the following sequences (thought of as parameters on a path down a hierarchy):

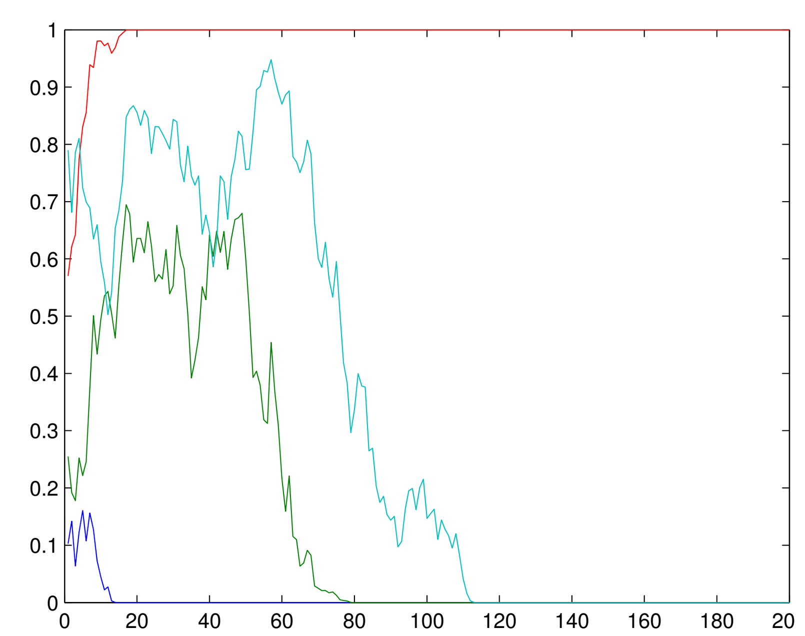
$$\theta_{n+1} \mid \theta_n \sim \text{DP}(c\theta_n), \qquad \theta_{n+1} \mid \theta_n \sim \text{BP}(c\theta_n),$$
$$\theta_{n+1} \mid \theta_n \sim \text{GammaP}(c\theta_n), \qquad \theta_{n+1} \mid \theta_n \sim \text{PYP}(c\theta_n)$$

- All have the property that $E[\theta_{n+1} \mid \theta_n] = \theta_n$.
  - Called the *martingale property*
  - Philosophically desirable because it means that information is preserved as we move down the hierarchy
- **Theorem (Doob):** All non-negative martingale sequences have a limit with probability 1.

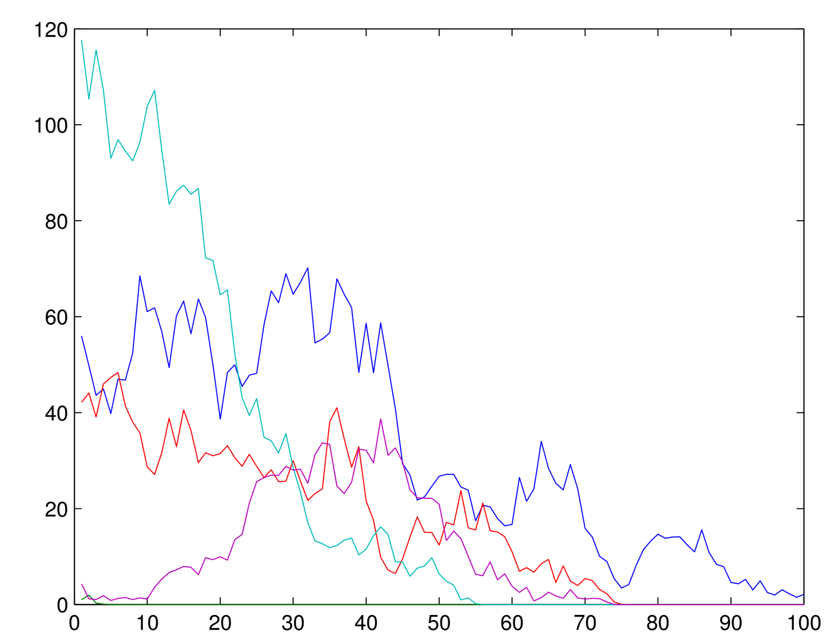
Computing the Limit

- The limiting variance of the distributions in a martingale must be 0, which implies:
  - θ converges to a single atom (DP and PYP)
  - All masses converge to 0 or 1 (beta process)
  - θ converges to 0 (gamma process)
- **DP, BP, and GammaP all involve draws from a gamma random variable, so we will necessarily run into the pathology described in Lemma 1!**
- See Example 3 for a martingale that can converge to an arbitrary value in [0,1] (also used in Solution 2)

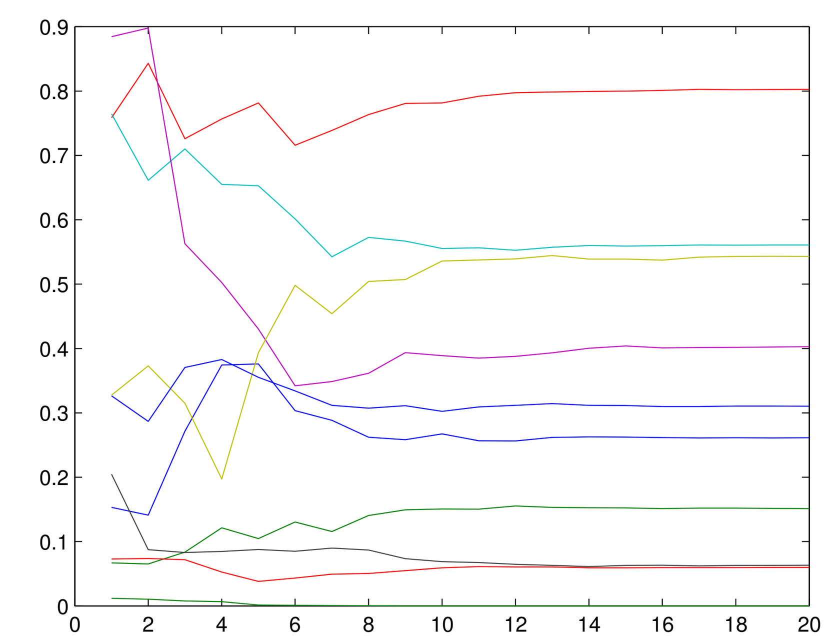
Examples of Martingale Sequences



Example 1: parameters of a hierarchical Beta process.
$\theta_{n+1} \mid \theta_n \sim \text{Beta}(50\theta_n, 50(1-\theta_n))$



Example 2: parameters of a hierarchical Gamma process.
$x_{n+1} \mid x_n \sim \text{Gamma}(x_n, 1)$



Example 3: a martingale given by $\theta_n = \alpha_n/(\alpha_n+\beta_n)$, where:
$\alpha_{n+1} \mid \alpha_n \sim \alpha_n + \text{Gamma}(\alpha_n, 1)$,
$\beta_{n+1} \mid \beta_n \sim \beta_n + \text{Gamma}(\beta_n, 1)$.
This construction can be used to rectify the problems with HBPs and HDPs.

Proving That the Decay Rate is a Tower of Exponentials

- **Theorem**: If $x_{n+1} \sim \text{Gamma}(c_n x_n, c_n)$, where $\{c_n\}$ is bounded, then $x_k \leq (\exp)^M(1)$ with probability 1-ε, where k = bM and b depends only on ε.
  - Note: $(\exp)^M$ means exponentiation composed M times
- Proof sketch: $x_{n+1} \ll x_n$ with non-negligible probability by Lemma 1, but the martingale property together with Markov's inequality bounds the probability that x_{n+1} is ever more than a constant greater than x_n.
- **Similar convergence properties (tower of exponentials) for Beta and Dirichlet.**

Naive Solution: Mixing with Noise

- Break martingale property and take, e.g., $\theta_{n+1} \sim \text{DP}(c[(1-\varepsilon)\theta_n + \varepsilon\mu_0])$, where μ_0 is some global base measure
- Issue: with N atoms, μ_0 places mass 1/N on some atom, so DP has at least one parameter as small as cε/N
  - Even more trouble with infinitely many atoms
- Forgets information after 1/ε steps

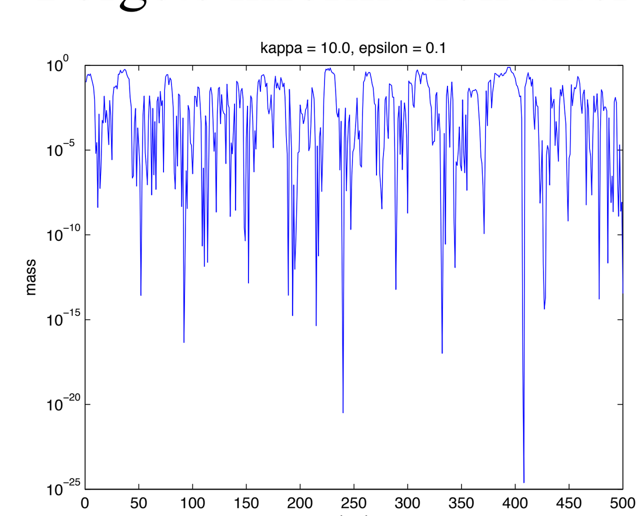


Figure 2: The mass assigned to an atom for a hierarchical Dirichlet process with noise mixed in. Here we have parameters c = 10.0, ε = 0.1, and μ_0 a uniform distribution over 10 atoms.

Solution 1: Pitman-Yor Processes

- Pitman-Yor processes have the following consistency property: if $G_1 \mid G_0 \sim \text{PYP}(\alpha,d_0,G_0)$, and $G_2 \mid G_1 \sim \text{PYP}(\alpha d_1,d_1,G_1)$, then $G_2 \mid G_0 \sim \text{PYP}(\alpha d_1,d_0 d_1,G_0)$.
- In general, $G_n \mid G_0 \sim \text{PYP}(\alpha d_1 \dots d_n, d_0 \dots d_n, G_0)$. If $G_0(\{p\}) = \varepsilon$, then $G_n(\{p\})$ is approximately

$$\left(\frac{\alpha\epsilon}{d_0+\alpha\epsilon}\right)^{\frac{1}{d_0\cdots d_n}}$$

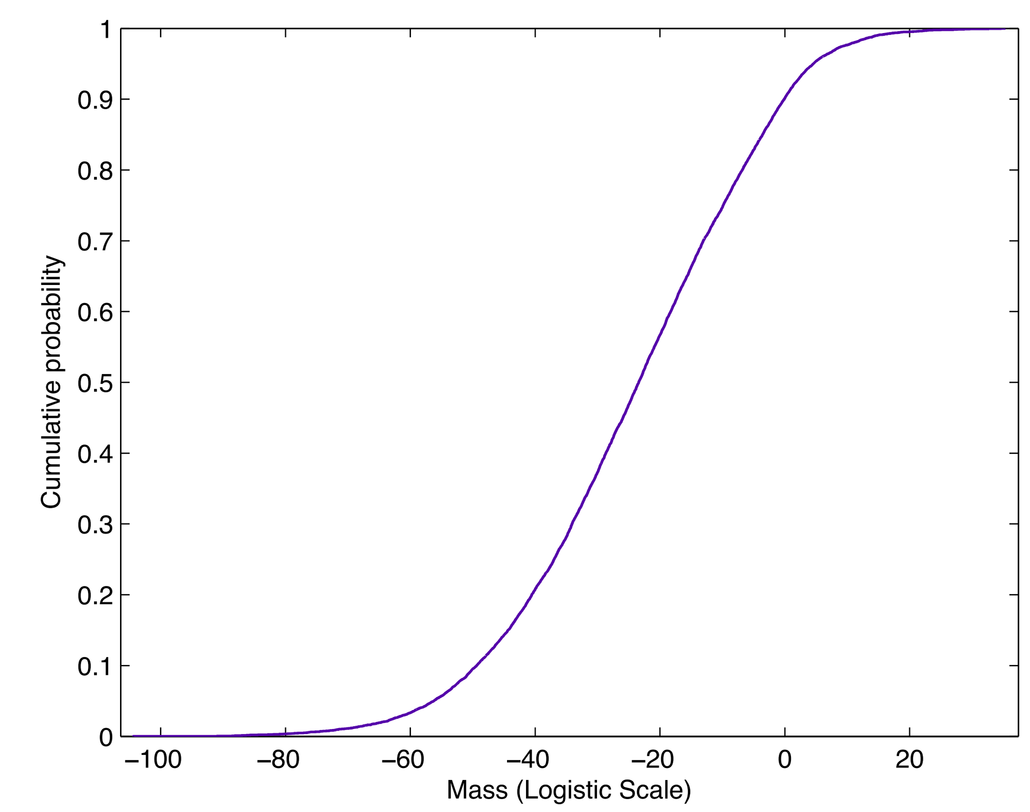


Figure 3: The cdf of the mass placed on an atom of base mass 0.1, for a draw from PYP(0.1,0.05).

Solution 2: Adding Inertia

- Instead of $x_{n+1} \sim \text{Gamma}(c_n x_n, c_n)$, have, e.g., $d_n \sim \text{Gamma}(c_n x_n, c_n)$, and $x_{n+1} = (1-a_n)x_n + a_n d_n$.
- Still a martingale, even for Dirichlet
- Rate of decay controlled by the sequence a_n

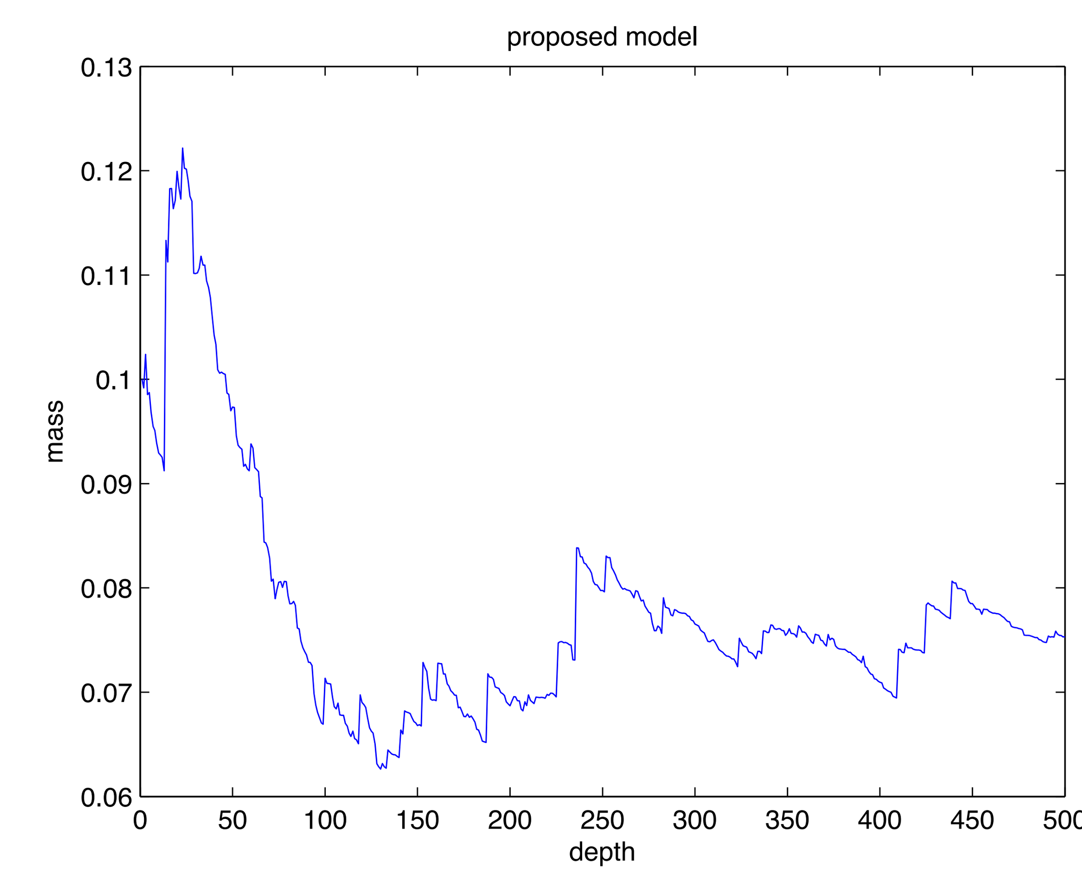


Figure 4: the mass of an atom on an inertia-added hierarchical Beta process. The sequence θ_n is generated as:
$\alpha_{n+1} = \alpha_n + \text{Gamma}(\alpha_n/\theta_n, 5)$
$\beta_{n+1} = \beta_n + \text{Gamma}(\beta_n/\theta_n, 5)$
$\theta_{n+1} = \alpha_{n+1}/(\alpha_{n+1}+\beta_{n+1})$