Pathological Properties of Deep Bayesian Hierarchies

Jacob Steinhardt
Department of Mathematics
Massachusetts Institute of Technology
jsteinha@mit.edu

Zoubin Ghahramani
Department of Engineering
University of Cambridge
zoubin@eng.cam.ac.uk

There has been much interest lately in combining nonparametric distributions together to form a hierarchy (Neal, 2003; Teh et al., 2004; Thibaux and Jordan, 2007; Thibaux, 2008; Blei et al., 2010; Adams et al., 2010; Knowles and Ghahramani, 2011). A common theme in most of these models is to sample a gamma process (or some related process such as a Dirichlet or beta process) recursively at each level with mean given by the parent. We show that this strategy leads to severe pathologies in the generative model — for instance, in a hierarchical Dirichlet process, all of the mass converges to a single atom, and the total mass on all other atoms decays at a rate of

$$\frac{1}{e^{e^{\cdot^{\cdot^{\cdot^{e}}}}}},$$

where the height of the tower of exponentials is proportional to the depth of the hierarchy; hierarchical beta and gamma processes act similarly. Such behavior is highly undesirable from a practical perspective. It also breaks the Bayesian commitment to writing down priors that reflect our beliefs, since surely no one would expect their model parameters to decay so quickly. Even worse, posterior predictions at higher levels of the hierarchy end up being sensitive to miniscule variations in parameters deeper in the hierarchy, which is again undesirable, both practically and philosophically.

While the most extreme form of these pathologies (the decay rate given by a tower of exponentials) appears to be previously undiscussed in the literature, the simpler phenomenon of decaying to a single atom has been noticed (Adams et al., 2010). A common strategy for fixing this problem is to mix the distribution over each parameter with some base measure. For instance, the latent parameters of a hierarchical Dirichlet process (Teh et al., 2004) are usually generated using the relationship $\theta_c \sim \mathrm{DP}(\kappa\theta_v)$, where θ_v is the parameter at a node v, θ_c is the parameter at its child, DP stands for Dirichlet process, and κ is the concentration parameter. To prevent θ from converging to an atom for deep hierarchies, one may instead use the relationship $\theta_c \sim \mathrm{DP}(\kappa[(1-\epsilon)\theta_v + \epsilon\theta_0])$, where θ_0 is some global base measure. We show that while these models avoid some of the worst of our exhibited pathologies, they are still unsuitable for a large variety of situations.

We note again that while the preceding two paragraphs used the hierarchical Dirichlet process as an example, our analysis is not limited to the Dirichlet process. We use general techniques based on martingales, employing Doob's martingale convergence theorem to demonstrate convergence to an atom and using more refined techniques to compute the actual rate of convergence as well as to analyze the case where the distribution is mixed with some base measure. These techniques allow us to show that both hierarchical gamma and beta processes also exhibit pathological behavior very similar to that of the Dirichlet process.

Having identified problems common across most of the existing hierarchical model classes, and having shown that the most straightforward solution is unsatisfactory, we then propose two solutions of our own, and analyze their pros and cons. Our first solution is a fairly general construction such that the parameters do not converge to a single atom. This is particularly desirable in situations such as topic modeling, where the leaves of the hierarchy correspond to topics, which inherently induce a distribution over words rather than representing a single word (in this case the words are the atoms of the distribution). In some situations, however, convergence to an atom may actually be desirable, for instance if the leaves represent a single datum and we want to allow subtrees to contain information that is as specific as necessary. We therefore also show how to replace Dirichlet

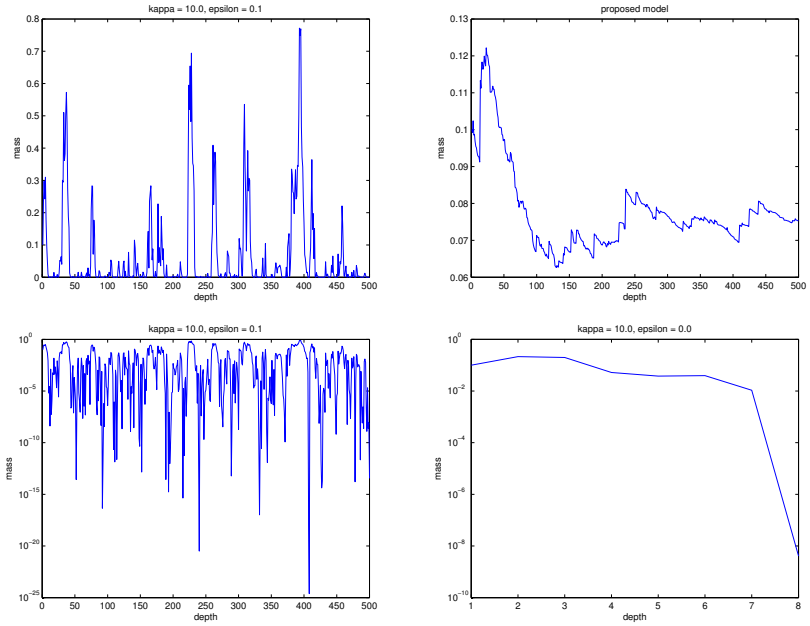

Figure 1: Sampled values of the mass of an atom in a hierarchical Dirichlet process as a function of depth. Top-left: a standard hierarchical Dirichlet process with $\kappa = 10.0$, $\epsilon = 0.1$. Note that the mass spends a lot of time near zero. Top-right: our first proposed change to the hierarchical Dirichlet process. Note that the mass now converges to a nonzero value. Bottom-left: the top-left graph on a semilogarithmic scale, which demonstrates that the mass regularly drops to values below 10^{-7}. Bottom-right: a demonstration of what happens if we just use the model $\theta_c \sim \mathrm{DP}(\kappa\theta_v)$ (in other words, if we set ϵ to zero). We only show up to a depth of 8 because at higher depths the mass is smaller than 10^{-308} and therefore rounded to 0 on a computer.

processes with a suitable parameterization of the Pitman-Yor process (Gasthaus and Teh, 2011), such that the parameters converge to a single atom at a controlled rate. This parameterization has the further advantage that it is the marginal distribution of a continuous-time stochastic process, and can therefore be incorporated into the framework of Dirichlet diffusion trees (Neal, 2003; Knowles and Ghahramani, 2011).

References

Ryan P. Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-structured stick breaking for hierarchical data. *Advanced in Neural Information Processing Systems*, 23, 2010.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2), Jan 2010.

Jan Gasthaus and Yee Whye Teh. Improvements to the sequence memoizer. *Advances in Neural Information Processing Systems*, 2011.

David Knowles and Zoubin Ghahramani. Pitman-yor diffusion trees. *Uncertainty in Artificial Intelligence*, 27, 2011.

R M Neal. Density modeling and clustering using dirichlet diffusion trees. In *Bayesian Statistics 7*, pages 619–629, 2003.

Y W Teh, M I Jordan, M J Beal, and D M Blei. Hierarchical dirichlet processes. Technical Report 653, 2004.

Romain Thibaux. *Nonparametric Bayesian Models for Machine Learning*. PhD thesis, University of California, Berkeley, 2008.

Romain Thibaux and Michael I. Jordan. Hierarchical beta processes and the indian buffet process. *AISTATS*, 2007.