
An Empirical Study of Hierarchical Dirichlet Process Priors for Grammar Induction

Kewei Tu

Department of Computer Science
Iowa State University
Ames, IA 50011, USA
tukw@cs.iastate.edu

Vasant Honavar

Department of Computer Science
Iowa State University
Ames, IA 50011, USA
honavar@cs.iastate.edu

1 Introduction

In probabilistic grammar induction, to avoid overfitting, simplicity priors are often used [1–4], which favor smaller grammars and penalize more complex grammars (the Occam’s Razor principle). The most used such prior is Solomonoff’s *universal probability distribution* $2^{-l(G)}$, where $l(G)$ is the description length of the grammar G .

Recently, there has been significant progress in applying nonparametric Bayesian models to machine learning problems. This kind of models do not assume a fixed model size (the number of parameters and/or hidden variables), but instead determine the model size from the data, with a preference towards a smaller model. One type of the nonparametric Bayesian models, the *hierarchical Dirichlet process* (HDP) [5], defines a distribution over an interrelated group of categorical distributions. Therefore, HDP seems to be a suitable prior for the transition probabilities of a probabilistic grammar. The advantage of using HDP as the prior is that, it can be naturally incorporated into the graphical model of the grammar, so many sophisticated inference algorithms, like variational Bayesian methods and Gibbs sampling, can be applied for grammar induction. So far a number of grammar models based on the HDP prior have been proposed, for both hidden Markov models (HMM) [5] and probabilistic context-free grammars (PCFG) [6, 7].

It is, however, still not clear how well the HDP prior favors smaller grammars. In particular, it is unclear how the prior probability of a grammar defined by HDP changes with the description length of the grammar, and what the relationship is between the HDP prior and the universal probability distribution that is widely used in previous grammar induction work. In addition, we wonder how the parameters of HDP affect its behavior. To answer these questions, we conducted a number of experiments as described in the rest of the paper.

2 Hierarchical Dirichlet Process

For the purpose of grammar induction, the hierarchical Dirichlet process (HDP) can be described as follows, using the *stick breaking representation*. First we define the stick breaking distribution $\text{GEM}(\alpha)$. To sample an infinite dimensional vector $\beta = (\beta_1, \beta_2, \dots) \sim \text{GEM}(\alpha)$, we first generate an infinite sequence of real numbers $\beta'_1, \beta'_2, \dots$ between 0 and 1 from a beta distribution: $\beta'_i \sim \text{Beta}(1, \alpha)$ for $i = 1, 2, \dots$; then β_i is defined to be $\beta'_i \prod_{j < i} (1 - \beta'_j)$. It is easy to see that $\sum_{i=1}^{\infty} \beta_i = 1$, so β can be interpreted as a probability distribution over positive integers. This construction procedure of β can be seen as iteratively breaking off portions from a unit length stick, each time according to a proportion sampled from a beta distribution.

Given the vector β , we can generate a set of categorical distributions over positive integers, each with an infinite dimensional vector ϕ_i as the parameter, s.t. $\phi_i \sim \text{DP}(\gamma, \beta)$ which is defined as follows. First we sample an infinite dimensional vector $\pi \sim \text{GEM}(\gamma)$, and sample an infinite

sequence of positive integers l_1, l_2, \dots from the categorical distribution defined by β ; then for any j , ϕ_{ij} is defined to be $\sum_{k \in L_j} \pi_k$ where $L_j := \{k | l_k = j\}$.

The above procedure defines how a set of categorical distributions over positive integers can be sampled from an HDP, which has two parameters α and γ . One important property of HDP is that, the categorical distribution sampled from it tends to put most probability mass to the first few outcomes. In other words, although an HDP assumes an infinite number of outcomes, only a small number of them are significant.

In grammar induction, since we do not know the number of nonterminals of the target grammar, we can assume a countably infinite number of them, each having a transition probability distribution over a countably infinite number of possible nonterminal productions¹, and HDP can be used here as the prior of these infinite number of distributions. Because of the property of HDP mentioned above, only a finite number of the nonterminals would be significant. For more details of using HDP priors for grammar induction, please see [5] for learning HMM and [6, 7] for learning PCFG.

3 Relation between HDP Priors and Description Length

In previous work of probabilistic grammar induction, the description length of a grammar $l(G)$ plays an important role in defining the prior probability of the grammar. For example, one of the most widely used prior is the universal probability distribution $P(G) \propto 2^{-l(G)}$ [1, 2, 4]. In this paper, to compute $l(G)$ we simply count the number of grammar rules in a grammar. Notice that for a probabilistic grammar, we may have grammar rules with negligible probabilities. Such rules are usually ignored when computing $l(G)$.

In this section, we try to find out the relation between the prior probability of the transition rules of a grammar defined by HDP, and the description length of the transition rules. Let vector ϕ_i be the transition probabilities of the nonterminal i , and let Φ be the set of these transition probability vectors. Based on the definition of HDP in Section 2, the prior probability is

$$P_{\text{HDP}}(\Phi) = \int \text{GEM}(\beta | \alpha) \prod_i \text{DP}(\phi_i | \gamma, \beta) d\beta \quad (1)$$

We can not find a closed form of this probability, so we resort to experimental methods.

First we generated a set of grammars of different sizes. The grammar formalism we used is HMM, but we believe the results we got would also hold for PCFG. Specifically, we set the number of nonterminals to be 5, 10, 15 or 20; for each nonterminal, we sampled the number of possible transitions based on a Gaussian distribution with the mean being either the total or a half of the number of nonterminals, and then we randomly selected the nonterminals it can change to, while made sure that each nonterminal can be reached from the first nonterminal which is the start symbol; finally we generated the transition probabilities from a uniform distribution. Since $\text{DP}(\phi)$ is undefined if ϕ contains zeros, we assigned a very small probability $\epsilon = 10^{-6}$ to transition rules not present in the grammar. To make it a fair comparison between grammars of different sizes, we also added virtual nonterminals into each grammar to make the total number of nonterminals to be always $K = 20$, and transition probabilities were generated for the virtual nonterminals in the same way as the real nonterminals. These virtual nonterminals cannot be reached from the start symbol, so they are not a part of the actual grammar. We did not generate the emission rules because they are irrelevant to the HDP prior.

On each of the generated grammars, we used importance sampling [8] to evaluate Equation 1. To make it possible, we truncated β at $K = 20$, so that $\beta_K = 1 - \sum_{i=1}^{K-1} \beta_i$ and $\beta_i = 0$ for $i > K$ (truncation is typical in inference of Dirichlet process with the stick breaking representation [6]). Because of the truncation, in Equation 1 DP is degenerated to a Dirichlet distribution of dimension K , and ϕ_i becomes a K -dimensional vector. The proposal distribution of β for importance sampling is a Dirichlet distribution with all the parameters being 1 (i.e., uniform over the simplex).

Figure 1 shows the relation between the log HDP prior probabilities of grammars and their description length l , when $\alpha = 1$ and $\gamma = 2$ (marked by plus signs in Fig.1). It can be seen that, the

¹For an HMM, the number of possible productions is the number of nonterminals; for a PCFG in the Chomsky normal form, the number of possible productions is the square of the number of nonterminals.

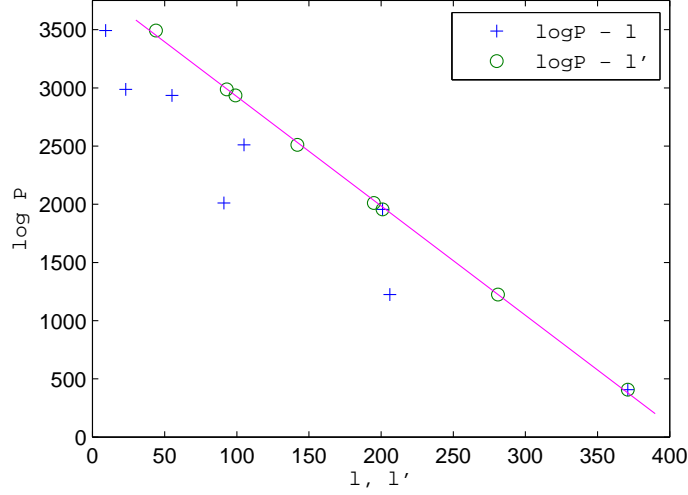


Figure 1: The relation between the log HDP prior probabilities and the description length when $\alpha = 1$ and $\gamma = 2$. l is the description length with only the real nonterminals counted; l' is the description length with both the real and the virtual nonterminals counted. The line is the least squares linear fit of the $\log P-l'$ data points.

relation between the two is not perfectly linear, but still smaller grammars tend to have exponentially higher prior probabilities. So HDP does seem to be an approximation of the universal probability distribution, and thus a decent simplicity prior for grammar induction.

We also plot the relation between the log prior probabilities and l' , the description length of grammars with the transition rules of virtual nonterminals counted in (marked by circles in Fig.1). Surprisingly, there is an almost perfect linear relation between the two. In other words, the HDP prior probability would be almost equivalent to the universal probability distribution if the description length was defined to count in grammar rules that cannot be reached from the start symbol.

This almost perfect fit can be explained as follows. With the truncation at K , Equation 1 can be rewritten as

$$\begin{aligned} P(\Phi) &= \int \alpha^{K-1} \beta_K^{\alpha-1} \prod_{i=1}^K \text{Dir}(\phi_i | \gamma \beta) d\beta \\ &= \alpha^{K-1} \int \beta_K^{\alpha-1} B(\gamma \beta)^{-K} \prod_{i=1}^K \prod_{j=1}^K \phi_{ij}^{\gamma \beta_j - 1} d\beta \end{aligned}$$

where $\text{Dir}(\cdot)$ is the Dirichlet distribution, and $B(\cdot)$ is the beta function. Notice that the transition probabilities in Φ (including those of virtual nonterminals) are either a very small value ϵ (for rules not actually present in the grammar, the set of which is denoted by E) or moderately large (for actual rules, the set of which is denoted by A). So the prior probability can be approximated as

$$\begin{aligned} P(\Phi) &\approx \alpha^{K-1} \int \beta_K^{\alpha-1} B(\gamma \beta)^{-K} \prod_{r \in E} \epsilon^{\gamma \bar{\beta} - 1} \prod_{r \in A} \phi_r^{\gamma \bar{\beta} - 1} d\beta \\ &\approx \alpha^{K-1} \epsilon^{(\gamma \bar{\beta} - 1)|E|} \bar{\phi}^{(\gamma \bar{\beta} - 1)|A|} \int \beta_K^{\alpha-1} B(\gamma \beta)^{-K} d\beta \end{aligned}$$

where $\bar{\beta}$ and $\bar{\phi}$ are the mean value of $\{\beta_j | j = 1, \dots, K\}$ and $\{\phi_r | r \in A\}$ respectively. Because $l' = |A| = K^2 - |E|$, we can get

$$\begin{aligned} \log P(\Phi) &\approx (\gamma \bar{\beta} - 1)(K^2 - l') \log \epsilon + (\gamma \bar{\beta} - 1)l' \log \bar{\phi} + (K - 1) \log \alpha \\ &\quad + \log \int \beta_K^{\alpha-1} B(\gamma \beta)^{-K} d\beta \end{aligned} \quad (2)$$

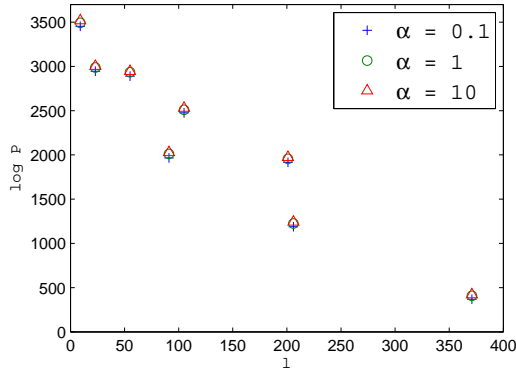


Figure 2: The effect of different α values when $\gamma = 2$.

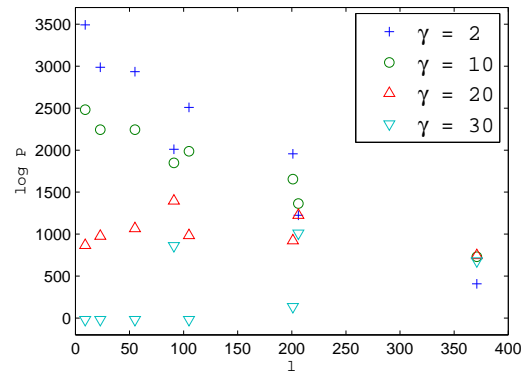


Figure 3: The effect of different γ values when $\alpha = 1$.

Therefore $\log P(\Phi)$ is approximately a linear function of l' . When $\alpha = 1$ and $\gamma = 2$, Equation 2 predicts a slope around 10.6, which is quite close to the slope of 9.4 of the linear fit.

The better fit of l' instead of the real description length l reveals a possible problem of the HDP prior when used in grammar induction. From Equation 1 we can see that, the transition probabilities ϕ_i of *any* nonterminal i plays an *equal* role in the formula, even if the nonterminal can not be reached from the start symbol (with a nonnegligible probability) and thus will not be used by the grammar. Notice that this problem still exists even if we do not truncate HDP. It is still unclear to us whether this poses a real problem in probabilistic grammar induction using HDP, because firstly, HDP is a prior, which is not so important if we have enough data; and secondly, while this problem is over individual grammars, in reality Bayesian inference (either MCMC [5, 7] or Variational methods [6]) is usually used to find the posterior of grammars instead of a single best grammar.

3.1 Effect of Parameters

We also studied how the relation between the HDP prior probabilities and the description length changes with different values of the two parameters α and γ , as shown in Figure 2 and 3 respectively.

The influence of α is rather small. The log probabilities with a larger α is only slightly (but consistently) higher than those with a smaller α . This is not surprising considering that α is only indirectly related to Φ . From Equation 2, it can also be seen that the log prior probability is dominated by the first term, which does not contain α .

The changing of γ , on the other hand, significantly affects the relation between the log prior probabilities and the description length. It can be seen that, γ controls how much and in which direction the HDP prior probabilities change with the description length. With $\gamma < 20$, larger description length leads to lower prior probabilities (as preferred in grammar induction). With γ around 20, the prior probabilities do not change much with the description length. With even larger γ , we see some grammars with larger description length now have higher prior probabilities, while for smaller grammars there seems to be a floor effect of the prior probabilities. This observation can be roughly explained as follows. Based on Equation 2, $\gamma\bar{\beta} - 1$ is the slope of the linear function of l' ; since $\bar{\beta}$ is around $1/K$, the slope changes its sign when γ is around K , which is 20 in our experiments. Although l is always smaller than l' , it does not change the shape of the function too much. This observation suggests that in grammar induction γ should be set to a value less than K , and that smaller γ leads to stronger simplicity bias.

4 Comparison of HDP and Dirichlet Priors

A Dirichlet distribution with very small parameters (less than 1) also generates categorical distributions that put most probability mass to a small number of outcomes, so it can also be used as

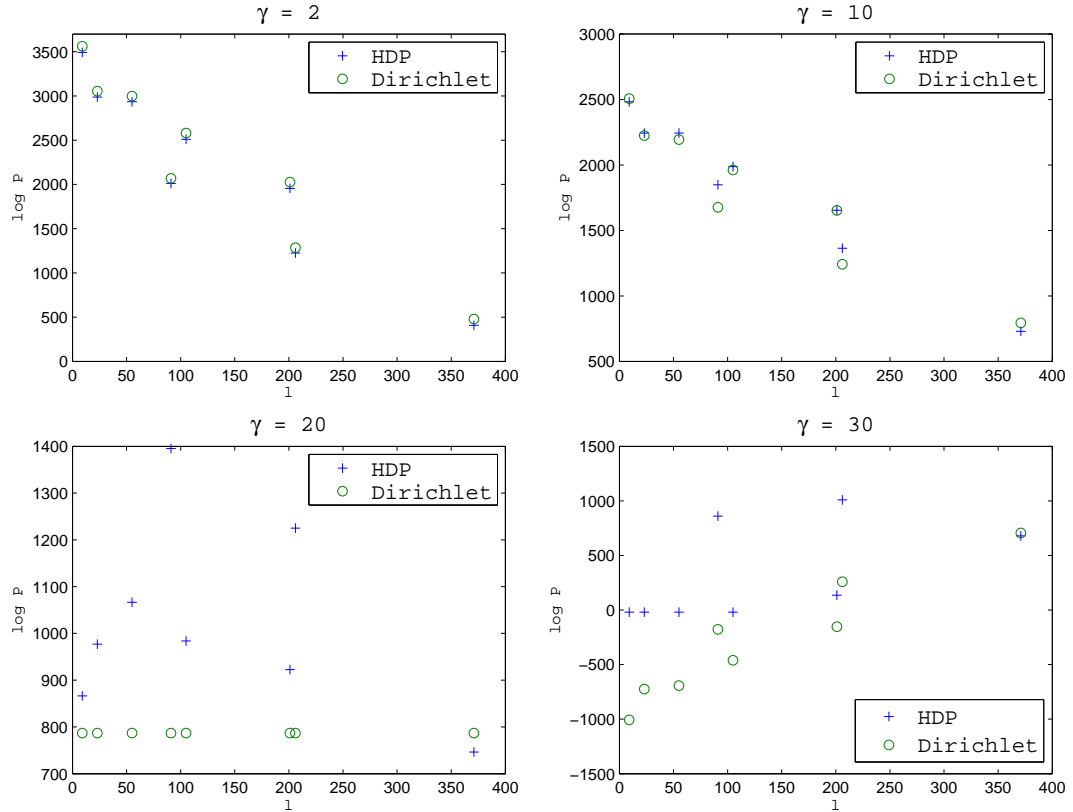


Figure 4: Comparison of the HDP and Dirichlet priors with different values of γ .

a simplicity prior. Indeed, if the parameters of a K -dimensional Dirichlet distribution are γ/K , then when K gets large, the Dirichlet prior becomes approximately equivalent to a Dirichlet process prior with γ as the concentration parameter [9]. In this section we compare the HDP prior with the Dirichlet prior.

For each grammar generated in Section 3, we computed the Dirichlet prior probability of the transition probabilities.

$$P_{\text{Dir}}(\Phi) = \prod_{i=1}^K \text{Dir}(\phi_i | \frac{\gamma}{K}, \dots, \frac{\gamma}{K})$$

Figure 4 shows the experimental results with different values of γ . When $\gamma = 2$ and 10, the Dirichlet prior is not very different from the HDP prior; when $\gamma = 20$ and 30, the two are different but still have similar trend. Notice that, as discussed in the previous section, HDP is suitable for grammar induction when $\gamma < K$. Since the HDP and Dirichlet priors are shown to be similar when $\gamma < K$, this raises the question of why one should use (truncated) HDP in grammar induction instead of the much simpler Dirichlet prior.

Another observation of the Dirichlet prior is that, when we plotted it against l' (the description length with virtual nonterminals counted in; figures not shown here), we found an almost perfect linear relation for any value of γ (while for HDP, the relation is perfectly linear only when γ is very small). This can be explained in a similar way as we did for HDP in the previous section.

5 Summary and Discussion

Hierarchical Dirichlet process (HDP) has recently been used as a prior distribution for probabilistic grammar induction. We conducted experiments to find out how the (truncated) HDP prior distribution is related to the description length of grammars, because description length is widely used

to define priors (e.g., the universal probability distribution) in previous grammar induction work. We find that, with proper parameter setting, the HDP prior does tend to assign exponentially higher probabilities to smaller grammars, but not as strictly as the universal probability distribution does. This discrepancy is because HDP takes into account nonterminals that are unreachable from the start symbol, which might be problematic for grammar induction. We also studied the effect of parameters of HDP, which provides some guideline of suitable parameter values when using HDP in grammar induction. We then compared the HDP prior with the Dirichlet prior, and found the two quite similar when using parameters suitable for grammar induction.

It shall be noted that, all the experiments were done with a truncated HDP, which is an approximation of HDP. Also, in practice Bayesian inference is often used for grammar induction with HDP, which finds the posterior of grammars instead of a single best grammar. Therefore, it would be interesting to study whether the findings of this paper still hold for non-truncated HDP and for grammar induction beyond point estimation.

References

- [1] Andreas Stolcke and Stephen M. Omohundro. Inducing probabilistic grammars by Bayesian model merging. In *ICGI*, pages 106–118, 1994.
- [2] Stanley F. Chen. Bayesian grammar induction for language modeling. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, 1995.
- [3] Pat Langley and Sean Stromsten. Learning context-free grammars with a simplicity bias. In *ECML*, pages 220–228, 2000.
- [4] Kewei Tu and Vasant Honavar. Unsupervised learning of probabilistic context-free grammar using iterative biclustering. In *Proceedings of 9th International Colloquium on Grammatical Inference (ICGI 2008)*, LNCS 5278, 2008.
- [5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- [6] Percy Liang, Slav Petrov, Michael I. Jordan, and Dan Klein. The infinite pcfg using hierarchical Dirichlet processes. In *Proceedings of EMNLP-CoNLL*, pages 688–697, 2007.
- [7] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. The infinite tree. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 272–279. Association for Computational Linguistics, June 2007.
- [8] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.
- [9] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269–284, 2002.