IOWA STATE UNIVERSITY

Artificial Intelligence Research Laboratory
Department of Computer Science

An Empirical Study of Hierarchical Dirichlet Process Priors for Grammar Induction

Kewei Tu and Vasant Honavar

1. Introduction

In probabilistic grammar induction, to avoid overfitting, simplicity priors are often used, which favor smaller grammars. An example of simplicity priors is Solomonoff's universal probability distribution $P(G) \propto 2^{-l(G)}$, where $l(G)$ is the description length of the grammar G.

The Hierarchical Dirichlet process (HDP) [Teh, et al., 2006] has recently been used as a prior for the transition probabilities of a probabilistic grammar [Teh, et al., 2006; Liang, et al, 2007; Finkel, et al, 2007].

- It is a kind of nonparametric Bayesian model.
- It can be naturally incorporated into the graphical model of the grammar, so many sophisticated inference algorithms can be used for grammar induction.

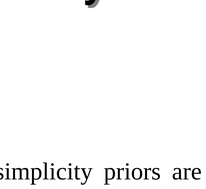
We want to find out

- how the HDP prior probability of a grammar changes with the description length of the grammar (compared with the universal probability distribution)
- how the parameters of HDP affect its behavior

2. Hierarchical Dirichlet Process

To sample a discrete distribution from an HDP (with two parameters α and γ):

- First, sample an infinite dimensional vector from the stick breaking distribution $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots) \sim \mathrm{GEM}(\alpha)$ defined as follows.



$$\beta_i' \sim \mathrm{Beta}(1, \alpha) \quad \text{for } i = 1, 2, \ldots$$

$$\beta_i = \beta_i' \prod_{j<i} (1 - \beta_j') \quad \text{for } i = 1, 2, \ldots$$

- Then, sample an infinite dimensional vector from the Dirichlet process $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots) \sim \mathrm{DP}(\gamma, \boldsymbol{\beta})$ defined as follows.

$$\boldsymbol{\pi} \sim \mathrm{GEM}(\gamma)$$

$$l_i \sim \boldsymbol{\beta} \quad \text{for } i = 1, 2, \ldots$$

$$\phi_i = \sum_{j \in L_i} \pi_j \quad \text{for } i = 1, 2, \ldots, \text{ where } L_i := \{ j \mid l_j = i \}$$

This vector defines a discrete distribution over positive integers.

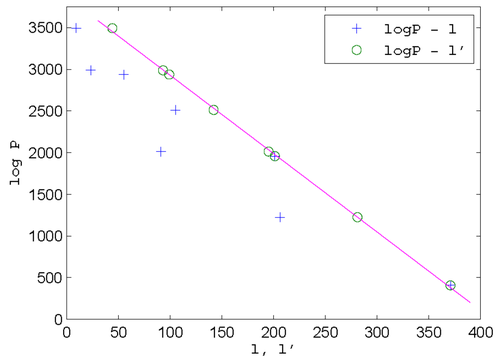
Discrete distributions sampled from HDP tend to put most probability mass to the first few outcomes. When they are used as the transition probabilities of a grammar, this means only a small number of nonterminals would be significant.

3. Relation between HDP Priors and Description Length

We generated a set of probabilistic grammars (HMM) of different sizes (i.e., different numbers of nonterminals and transition rules).

- Since HDP is undefined if any transition probability is zero, we assigned a 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. These virtual nonterminals cannot be reached from the start symbol, so they are not a part of the actual grammar.

To compute the description length, we simply counted the number of grammar rules. Since there is no closed form of the HDP prior probability, we approximate it by importance sampling [Andrieu, et al, 2003] with $\boldsymbol{\beta}$ truncated at K=20.



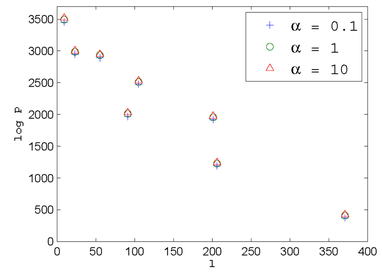
The relation between the log HDP prior probabilities and the description length when α=1 and γ=2. l is the description length with only the real nonterminals counted; l' is the description length with both the real and the virtual nonterminals counted. The red line is the least squares linear fit of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

The result shows that, HDP does tend to give exponentially higher prior probabilities to smaller grammars, so it can be seen as an approximation of the universal probability distribution.

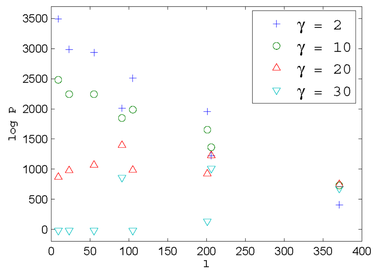
However, surprisingly the relation between the log HDP prior and the description length that takes into account the transition rules of virtual nonterminals is almost perfectly linear. See our paper for a theoretical explanation.

This reveals a possible problem of the HDP prior for grammar induction: the transition probabilities of **any** nonterminal play an **equal** role, 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. It is unclear whether this poses a real problem in probabilistic grammar induction using HDP.

3.1 Effect of Parameters



The effect of different α values when γ=2

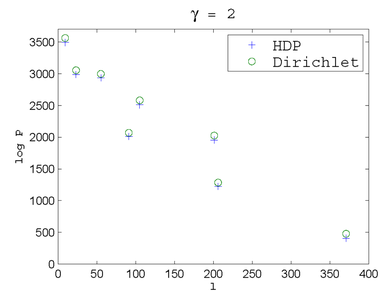


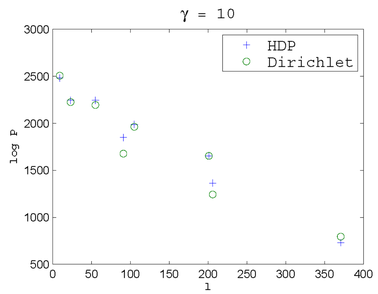
The effect of different γ values when α=1

The influence of α is rather small. The value of γ, however, significantly changes the "slope" of the relation between the log HDP prior and the description length. This implies that in grammar induction γ should be set to a small value.

4. Comparison of HDP and Dirichlet Priors

We compared the HDP prior with the Dirichlet prior (with every parameter set to γ/K).





When γ is small, the Dirichlet prior is not very different from the HDP prior; when $\gamma \geq 20$ (not shown here, please see our paper), the two are different but still have similar trends. Since we shall use a small γ in grammar induction, this raises the question as to why one should use (truncated) HDP in grammar induction instead of the much simpler Dirichlet prior.

5. Discussion

All the experiments were done with a truncated HDP, which is an approximation of HDP. Also, Bayesian inference is often used for grammar induction with HDP, which finds the posterior of grammars instead of a single best grammar. So, it would be interesting to study whether the findings of this paper still hold for non-truncated HDP and for grammar induction with Bayesian inference.