Reliable and Scalable Variational Inference for the Hierarchical Dirichlet Process

Michael C. Hughes, Dae Il Kim, and Erik B. Sudderth
mhughes@cs.brown.edu daeil@cs.brown.edu sudderth@cs.brown.edu
Dept. of Computer Science, Brown University, Providence, RI, USA.

Abstract

We introduce a new variational inference objective for hierarchical Dirichlet process admixture models. Our approach provides novel and scalable algorithms for learning nonparametric topic models of text documents and Gaussian admixture models of image patches. Improving on the point estimates of topic probabilities used in previous work, we define full variational posteriors for all latent variables and optimize parameters via a novel surrogate likelihood bound. We show that this approach has crucial advantages for data-driven learning of the number of topics. Via merge and delete moves that remove redundant or irrelevant topics, we learn compact and interpretable models with less computation. Scaling to millions of documents is possible using stochastic or memoized variational updates.

1 INTRODUCTION

Bayesian nonparametric models are increasingly applied to data with rich hierarchical structure, such as words within documents (Teh et al., 2006) or patches within images (Sudderth et al., 2008). *Hierarchical Dirichlet process* (HDP) admixture models provide a natural way to discover shared clusters, or *topics*, in grouped data. The HDP prior expects the number of topics to smoothly grow as more examples appear, making it attractive for analyzing big datasets.

While there are numerous existing inference algorithms for the HDP, all suffer from some combination of inability to scale to large datasets, vulnerability to poor local optima, or the need for external specification of the target model complexity. Simple Markov chain Monte Carlo (MCMC) samplers (Teh et al., 2006) can dynamically add or remove topics, but are computationally demanding with more than a few thousand documents and may take a long time to mix from a poor initialization. Collapsed variational methods (Teh et al., 2008) are based on a sophisticated family of marginal likelihood bounds, but lead to challenging optimization problems and sensitivity to initialization. Stochastic variational methods (Wang et al., 2011) and streaming methods (Broderick et al., 2013) are by design more scalable, but are easily trapped at a fixed point near a poor initialization. More recent variational algorithms have dynamically inserted or removed topics to escape local optima, but either lack guarantees for improving whole-data model quality (Bryant and Sudderth, 2012) or rely on slow-to-mix Gibbs sampler steps (Wang and Blei, 2012).

We develop a scalable HDP learning algorithm that enables reliable selection of the number of active topics. After reviewing HDP admixtures in Sec. 2, we develop a novel variational bound (Sec. 3) that captures posterior uncertainty in topic appearance probabilities, and leads to sensible model selection behavior (see Fig. 2). Sec. 4 then develops novel stochastic (Hoffman et al., 2013) and memoized (Hughes and Sudderth, 2013) variational inference algorithms for the HDP. The memoized approach supports merge and delete moves (Sec. 5) that remove redundant or irrelevant topics, leading to compact and interpretable models. Sec. 6 demonstrates faster and more accurate learning of HDP models for documents and images.

2 HDP ADMIXTURE MODELS

Consider data partitioned into D exchangeable groups $x = \{x_1 \dots x_D\}$, for example documents or images. Each group d contains N_d tokens $x_d = \{x_{d1}, \dots x_{dN_d}\}$, for example words or small pixel patches. For large datasets we divide groups into B predefined *batches*, where $\mathcal{D}_b$ is the set of documents in batch b.

Appearing in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. Copyright 2015 by the authors.