A Search Algorithm for Latent Variable Models with Unbounded Domains

Michael Chiang and **David Poole**
{mchc,poole}@cs.ubc.ca
http://www.cs.ubc.ca/~poole
http://www.cs.ubc.ca/~mchc

Abstract

This paper concerns learning and prediction with probabilistic models where the domain sizes of latent variables have no *a priori* upper-bound. Current approaches represent prior distributions over latent variables by stochastic processes such as the Dirichlet process, and rely on Monte Carlo sampling to estimate the model from data. We propose an alternative approach that searches over the domain size of latent variables, and allows arbitrary priors over the their domain sizes. We prove error bounds for expected probabilities, where the error bounds diminish with increasing search scope. The search algorithm can be truncated at *any time*. We empirically demonstrate the approach for topic modelling of text documents.

Introduction

Latent variables are important for building models that can compactly represent a domain. However, often we do not know the domain size of latent variables *a priori*. Rather than fixing the sizes of each latent variable, we want to have a distribution over the sizes of these variables. The standard representation is in terms of nonparametric statistical models [Teh et al., 2006; Aldous, 1983; Antoniak, 1974; Ferguson, 1973; Griffiths and Ghahramani, 2006; Rasmussen, 2006], where prior distributions for variables with unbounded domains are represented as stochastic processes, e.g. Dirichlet process [Ferguson, 1973], and the posterior model estimated via Markov chain Monte Carlo sampling [Neal, 1998]. In this paper we present an alternative way to represent and compute probabilistic models with latent variable. Our approach permits arbitrary distributions over the domain sizes, and learning is done with a search algorithm over domain sizes which outputs a bound on the posterior probabilities.

We represent the domain size of latent variables as random variables called **size variables**. Size variable domains are the set of positive integers. Each latent variable with unknown domain size has a distinguished size variable as a parent. Let $\mathbf{Y},\mathbf{Z},\mathbf{S}$ be the sets of observed, latent, and *size* random variables respectively. Define a **size configuration** to be an assignment of a value to each size variable.

For learning, our approach uses a sightly different joint probability model to those based on stochastic processes; we model the joint distribution as $p(\mathbf{Y},\mathbf{Z} \mid \mathbf{S})\, p(\mathbf{S})$. Learning involves searching over the domain of $\mathbf{S}$, and employing marginal inference to compute the conditional $p(\mathbf{Y},\mathbf{Z} \mid \mathbf{s})$ for each size configuration $\mathbf{s}$ visited. The main contributions of this approach are: (i) that the prior $p(\mathbf{S})$ can be any distribution over positive integers such as the Poisson or geometric distribution, or a domain-specific distribution[1]; (ii) different subsets of latent variables in the model can be tied to different priors which can be interdependent; (iii) the search algorithm can accept an arbitrary number of size variables, and is accompanied by expected likelihood bounds that monotonically converge as more size configurations in the domain of $\mathbf{S}$ are enumerated.

In our experiments, we consider topic modelling of text documents, where we search over the number of latent topics, and demonstrate the optimisation of the expected likelihood bounds computed by the search algorithm.

Preliminaries

Upper-case letters denote random variables, e.g. X, and their values represented in lower-case, e.g. x. Boldface is used for sets of random variables and values, e.g. $\mathbf{X}$ and $\mathbf{x}$. Where it is clear from context, we abbreviate the assignment $X = x$ as x. In this paper we consider only discrete random variables (the domain of random variables are discrete sets). The domain (set of possible values) of a variable X is written as $dom(X)$. For a set of variables $\mathbf{X}$, the domain is $dom(\mathbf{X})$ – the cross product of domains of variables in $\mathbf{X}$.

Nonparametric Bayesian Models

Models that are not captured by a fixed set of parameters are generally referred to as *nonparametric models*. Existing nonparametric models typically adopt the factorisation $p(\mathbf{Y} \mid \mathbf{Z},\mathbf{S})\, p(\mathbf{Z},\mathbf{S})$ of the joint distribution, where the prior $p(\mathbf{Z},\mathbf{S})$ is often described by a stochastic process. A well-known example of this is the Dirichlet process mixture model [Antoniak, 1974]; a mixture model with an *a priori*

[1] One difference from DPs, is that if we have, for example, 5 latent classes, these classes are uniformly distributed in that any instance is equally likely to be in any class. In DPs, the classes will not be equally likely. It is an empirical question as to which of these prior assumptions are more appropriate, and the answer may differ from domain to domain.

unbounded number of component distributions that are generated according to a Dirichlet process [Ferguson, 1973].

Related models based on other stochastic processes such as the Chinese restaurant process [Aldous, 1983] have also been studied, and have been used widely for infinite mixture models, as well as infinite relational models [Kemp et al., 2006] for relational data. Another example is the Indian Buffet process for infinite feature models [Griffiths and Ghahramani, 2006]. Extensions of the Dirichlet process have also been proposed, e.g. hierarchical Dirichlet process [Teh et al., 2006] where a hierarchy of Chinese restaurant processes generate the data.

The approach of [Blei and Jordan, 2005] bears similarity to that of this paper, with truncated search of the number of Dirichlet process mixture components whilst maintaining a variational lower-bound on the probability of evidence. Another search and bound approach for Dirichlet process mixture was given by [Daume III, 2007], where search is guided by a heuristic estimate of the path cost, similar to A* search. Our approach provides upper and lower bounds on the expected likelihood of evidence, which are not derived using variational approximations, and the accompanying search procedure does not rely on heuristics.

Expected Likelihood and Witnesses

In this paper we model $p(\mathbf{Y},\mathbf{Z},\mathbf{S})$ in the form $p(\mathbf{Y},\mathbf{Z}\mid\mathbf{S})\,p(\mathbf{S})$, parametrised in θ. The parametrisation θ specifies both the distribution over the size variables, and the distribution of the other variables given the size variables. The model in question is

$$p(\mathbf{Y},\mathbf{Z}\mid\mathbf{S},\theta)\,p(\mathbf{S}\mid\theta) \tag{1}$$

Size variables in $\mathbf{S}$ can have unbounded domains, thus θ may not have a finite representation. Given evidence $\mathbf{y}$, we show how finitely-representable approximations of the joint probability $p(\mathbf{y},\mathbf{Z}\mid\mathbf{S})p(\mathbf{S})$ can be computed by evaluating a finite subset of $dom(\mathbf{S})$ (and this is done by search below).

For clarity, we assume in the remainder of this paper that all latent variables in $\mathbf{Z}$ have *a priori* unbounded domains, although in general some latent variables may have finite domains. Results derived apply directly to the case when some latent variables fixed finite domains.

Bounds on Expected Likelihoods

Given evidence $\mathbf{y}$, we propose to generate some size configurations, where for each configuration $\mathbf{s}$ generated a model of $p(\mathbf{y},\mathbf{Z}\mid\mathbf{s})$ is computed. Using the likelihoods induced by these models and the prior probabilities of the remaining unvisited configurations, we bound expected likelihoods whose exact values involve averaging over all size configurations.

First we want to be able to compare configurations of size variables:

Definition 1. *Suppose* $\mathbf{S}=\langle S_1,\dots,S_k\rangle$ *is a tuple of the size variables in a model. Let* $\mathbf{a}=\langle a_1,\dots,a_k\rangle$ *and* $\mathbf{b}=\langle b_1,\dots,b_k\rangle$ *be two configurations of size variables* $\mathbf{S}$*, where for each i,* $a_i\in dom(S_i)$ *and* $b_i\in dom(S_i)$*. Configuration* $\mathbf{a}$ ***precedes*** $\mathbf{b}$*, written as* $\mathbf{a}\preceq\mathbf{b}$*, if* $\forall i, a_i\le b_i$*.*

For each $\mathbf{s}\in dom(\mathbf{S})$ and a parametrization θ, let $\theta_\mathbf{s}$ be parameters of the model whose size variables have been assigned the value $\mathbf{s}$. The **expected likelihood** of data $\mathbf{y}$ given $\theta_\mathbf{s}$ is then

$$\begin{aligned} p(\mathbf{y}\mid\mathbf{s},\theta_\mathbf{s}) &= \sum_{\mathbf{z}\in\Gamma_\mathbf{s}} p(\mathbf{y},\mathbf{z}\mid\mathbf{s},\theta_\mathbf{s}) \\ &= \sum_{\mathbf{z}\in\Gamma_\mathbf{s}} p(\mathbf{y}\mid\mathbf{z},\mathbf{s},\theta_\mathbf{s})\,p(\mathbf{z}\mid\mathbf{s},\theta_\mathbf{s}) \end{aligned} \tag{2}$$

where $\Gamma_\mathbf{s}$ is the domain of $\mathbf{Z}$ where the domain size of each $Z\in\mathbf{Z}$ is specified in $\mathbf{s}$.

For two size configuration $\mathbf{s}$ and $\mathbf{t}$ where $\mathbf{s}\preceq\mathbf{t}$, we have the following proposition:

Proposition 1. *Let* $\mathbf{s}$ *and* $\mathbf{t}$ *be two configurations of size variables* $\mathbf{S}$ *such that* $\mathbf{s}\preceq\mathbf{t}$*, and* $\mathbf{y}$ *be values for variables* $\mathbf{Y}$*, then for each parametrisation* $\theta_\mathbf{s}$*, there is a parametrization* $\theta_\mathbf{t}$ *such that*

$$p(\mathbf{y}\mid\mathbf{s},\theta_\mathbf{s})\le p(\mathbf{y}\mid\mathbf{t},\theta_\mathbf{t})\le 1 \tag{3}$$

Proof. Let the domain of latent variables under $\mathbf{s}$ and $\mathbf{t}$ be $\Gamma_\mathbf{s}$ and $\Gamma_\mathbf{t}$ respectively. Given $\theta_\mathbf{s}$, we can choose parameters $\theta_\mathbf{t}$ such that

$$\begin{aligned} &\forall\mathbf{z}\in\Gamma_\mathbf{s}, p(\mathbf{y},\mathbf{z}\mid\mathbf{t},\theta_\mathbf{t}) = p(\mathbf{y},\mathbf{z}\mid\mathbf{s},\theta_\mathbf{s}) \\ &\forall\mathbf{z}\in\Gamma_\mathbf{t}-\Gamma_\mathbf{s}, p(\mathbf{z}\mid\mathbf{t},\theta_\mathbf{t}) = 0 \end{aligned}$$

Choosing $\theta_\mathbf{t}$ accordingly results in a $\theta_\mathbf{t}$ that nests $\theta_\mathbf{s}$, and yields the equivalence $p(\mathbf{y}\mid\mathbf{t},\theta_\mathbf{t})=p(\mathbf{y}\mid\theta_\mathbf{s},\mathbf{s})$. Any choice of $\theta_\mathbf{t}$ that yields a lower expected likelihood can be discarded, since nesting is always possible given $\theta_\mathbf{s}$. Thus, choosing a $\theta_\mathbf{t}$ that nests $\theta_\mathbf{s}$ yields the inequality $p(\mathbf{y}\mid\theta_\mathbf{s},\mathbf{s})\le p(\mathbf{y}\mid\theta_\mathbf{t},\mathbf{t})$. The upper-bound is true by definition of probability. □

We want to generate bounds for the **full expected likelihood** – the probability of $\mathbf{y}$ given parametrisation θ:

$$p(\mathbf{y}\mid\theta)=\sum_\mathbf{s} p(\mathbf{s}\mid\theta)\sum_{\mathbf{z}\in\Gamma_\mathbf{s}} p(\mathbf{y},\mathbf{z}\mid\mathbf{s},\theta) \tag{4}$$

Like the expected likelihood of Eq. (2), the full expected likelihood is also an infinite sum. We approximate the full expected likelihood by partially computing the sum, evaluating only sumands corresponding to a finite subset of configurations $\mathbf{G}$ of $dom(\mathbf{S})$; called the **generated set** in this paper. For each element $\mathbf{s}$ of $\mathbf{G}$, we assume that we have derived parameters $\theta_\mathbf{s}$. Let $\theta_\mathbf{G}$ be the set of parametrisations such that the parameters for each element $\mathbf{s}$ of $\mathbf{G}$ are given by $\theta_\mathbf{s}$. For each configuration $\mathbf{t}\notin\mathbf{G}$ the parameters of $\theta_\mathbf{t}$ are such that for all $\mathbf{s}\in\mathbf{G}$, where $\mathbf{s}\preceq\mathbf{t}$, $p(\mathbf{y}\mid\mathbf{s},\theta_\mathbf{s})\le p(\mathbf{y}\mid\mathbf{t},\theta_\mathbf{t})$. The existence of $\theta_\mathbf{t}$ is given by Prop. 1. For example, the maximum likelihood choice for $\theta_\mathbf{t}$ is a special case (maximum *a posteriori* parametrisations can also be used with careful consideration of pseudo counts), although any choice that improves upon the expected likelihood for a preceding configuration in $\mathbf{G}$ is valid.

We use search to construct $\mathbf{G}$, and using models corresponding to configurations in $\mathbf{G}$ we bound the remaining probability mass. To do so, we make use of the notion of *witnesses*[2]:

Definition 2. *Let $\mathbf{G} \subset dom(\mathbf{S})$ be a generated set of size variable configurations. A **witness function** is a function f_W mapping from $dom(\mathbf{S}) - \mathbf{G}$ into $\mathbf{G}$, such that for all $\mathbf{c} \in dom(\mathbf{S}) - \mathbf{G}$, $f_W(c) \preceq c$. The **witness set** for witness function f_W is the range of f_W.*

Note that given $\mathbf{G}$, there may be many possible witness functions for each assignment not in $\mathbf{G}$. We wish to keep the best witness functions – i.e. those that yield the highest expected likelihoods – for all assignments not in $\mathbf{G}$. As such, the witness set need not include all of $\mathbf{G}$.

The idea of our search approach is that maintains a generated set of size assignments for which it have computed the parameters. For every size assignment not generated, it uses a witness for that configuration to bound the probability. The simplest instance of this idea uses a single witness for all non-generated configurations. For a witness set $\mathbf{W}$, the **min-witness** configuration is

$$\mathbf{w}_\perp = \arg\min_{\mathbf{w}\in\mathbf{W}} p\left(\mathbf{y} \mid \mathbf{w}, \theta_\mathbf{w}\right) \tag{5}$$

which provides an underestimate of the likelihood of the data for all unvisited configurations.

Lemma 1. *Let $\mathbf{G}$ be a generated set of configurations. For all parametrisations $\theta \in \theta_\mathbf{G}$, the full expected likelihood $p\left(\mathbf{y} \mid \theta\right)$ is bound as follows*

$$F + p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta\right) H \leq p\left(\mathbf{y} \mid \theta\right) \leq F + H \tag{6}$$

where

$$\begin{aligned} F &= \sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right) \\ H &= \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p\left(\mathbf{t} \mid \theta\right) \end{aligned} \tag{7}$$

Proof. Suppose $\mathbf{t} \in dom(\mathbf{S}) - \mathbf{G}$ with a witness $\mathbf{w}$, by Prop. 1, there exist a $\theta_\mathbf{t}$ such that $p\left(\mathbf{y} \mid \mathbf{w}, \theta_\mathbf{w}\right) \leq p\left(\mathbf{y} \mid \mathbf{t}, \theta_\mathbf{t}\right)$. By Eq. (5), $p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta_{\mathbf{w}_\perp}\right) \leq p\left(\mathbf{y} \mid \mathbf{w}, \theta_\mathbf{w}\right)$ and so,

$$p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta_{\mathbf{w}_\perp}\right) \leq p\left(\mathbf{y} \mid \mathbf{t}, \theta_\mathbf{t}\right) \tag{8}$$

To bound the full expected likelihood, we split Eq. (4) according to $\mathbf{G}$ to yield a finite and infinite sum

$$\begin{aligned} p\left(\mathbf{y} \mid \theta\right) = &\sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right) \\ &+ \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p\left(\mathbf{t} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{t}, \theta\right) \end{aligned} \tag{9}$$

Using the parametrisation θ, and applying Eq. (8), the second summation (an infinite sum) can be bound:

$$\begin{aligned} p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta\right) &\sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p\left(\mathbf{t} \mid \theta\right) \\ \leq &\sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p\left(\mathbf{t} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{t}, \theta\right) \\ \leq &\sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p\left(\mathbf{t} \mid \theta\right) \end{aligned} \tag{10}$$

Finally, applying the bounds of Eq. (10) to the infinite sum in Eq. (9) directly yield Eq. (6). □

A key property of Lemma 1 is that the bounds converge monotonically with the size of $\mathbf{G}$ for two reasons:

- $p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta\right)$ is monotonically non-decreasing due to Prop. 1, and can be made to be increasing as long as not all of the data is fit perfectly
- the sum $H = \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p(\mathbf{t} \mid \theta)$ monotonically decreases as long as no size has a prior probability of zero.

For model selection and prediction we wish to estimate the posterior distribution over size configurations. Using bounds on the full expected likelihood from Lem. 1, we bound the posterior distribution over size configurations

Lemma 2. *Given a generated set of size variable assignments $\mathbf{G}$ and a parametrization $\theta \in \theta_\mathbf{G}$, for all $\mathbf{s} \in dom(\mathbf{S})$ it holds that*

$$\frac{p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right)}{F + H} \leq p\left(\mathbf{s} \mid \mathbf{y}, \theta\right) \leq \frac{p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right)}{F + p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta\right) H} \tag{11}$$

where F, H and $\mathbf{w}_\perp$ are defined in Lem. 1.

Proof. Using Bayes' rule,

$$p\left(\mathbf{s} \mid \mathbf{y}, \theta\right) = \frac{p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right)}{p\left(\mathbf{y} \mid \theta\right)}$$

Applying Lem. 1 to the denominator $p\left(\mathbf{y} \mid \theta\right)$ directly yields Eq. (11). □

Prediction

Let $\mathbf{y}'$ be some unobserved values we wish to predict, the predictive distribution is

$$p\left(\mathbf{y}' \mid \mathbf{y}, \theta\right) = \sum_{\mathbf{s}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) p\left(\mathbf{s} \mid \mathbf{y}, \theta\right) \tag{12}$$

which is an infinite sum with non-finite parametrisation θ. Let $\theta \in \theta_\mathbf{G}$, where $\mathbf{G}$ is defined prior to Def. 2. The posterior term $p(\mathbf{s} \mid \mathbf{y}, \theta)$ in Eq. (12) can be bound by Lem. 2, then Eq. (12) satisfies the bounds $A \leq p\left(\mathbf{s} \mid \mathbf{y}, \theta\right) \leq B$ where

$$\begin{aligned} A &= \sum_{\mathbf{s}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right)}{F + H} \\ B &= \sum_{\mathbf{s}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right)}{F + p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta\right) H} \end{aligned}$$

Here A and B are infinite sums. Given a generated set of size configuration $\mathbf{G}$, the sums can be split into finite and infinite components:

$$\begin{aligned} A = &\sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right)}{F + H} \\ &+ \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{t}, \mathbf{y}, \theta\right) \frac{p\left(\mathbf{t} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{t}, \theta\right)}{F + H} \\ B = &\sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p\left(\mathbf{s} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{s}, \theta\right)}{F + p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta\right) H} \\ &+ \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{t}, \mathbf{y}, \theta\right) \frac{p\left(\mathbf{t} \mid \theta\right) p\left(\mathbf{y} \mid \mathbf{t}, \theta\right)}{F + p\left(\mathbf{y} \mid \mathbf{w}_\perp, \theta\right) H} \end{aligned}$$

[2] We use the term witness in a similar manner to that defined for the witness algorithm for partially-observed Markov decision processes [Cassandra and Littman, 1994]. Every size configuration that is not generated (not in $\mathbf{G}$) can refer to a witness that testifies to its bounds.

Whilst the expected likelihood $p(\mathbf{y} \mid \mathbf{s}, \theta)$ and the prediction $p(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta)$ are known for all $\mathbf{s} \in \mathbf{G}$, they must be approximated for all $\mathbf{s} \in dom(\mathbf{S}) - \mathbf{G}$. The prediction term can be approximated by our prior belief on $\mathbf{y}'$, and the expected likelihood $p(\mathbf{y}' \mid \mathbf{t}, \theta)$ can be bound using Prop. 1 with the witness likelihood $p(\mathbf{y} \mid \mathbf{w}_\perp, \theta)$. As such, $A_1 \le A \le A_2$, where

$$A_1 = \sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p(\mathbf{s} \mid \theta)\, p(\mathbf{y} \mid \mathbf{s}, \theta)}{F+H} + \frac{p(\mathbf{y}' \mid \theta)\, p(\mathbf{y} \mid \mathbf{w}_\perp, \theta)}{F+H} \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p(\mathbf{t} \mid \theta)$$

$$A_2 = \sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p(\mathbf{s} \mid \theta)\, p(\mathbf{y} \mid \mathbf{s}, \theta)}{F+H} + \frac{p(\mathbf{y}' \mid \theta)}{F+H} \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p(\mathbf{t} \mid \theta)$$

Similarly, $B_1 \le B \le B_2$ where

$$B_1 = \sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p(\mathbf{s} \mid \theta)\, p(\mathbf{y} \mid \mathbf{s}, \theta)}{F+p(\mathbf{y} \mid \mathbf{w}_\perp, \theta) H} + \frac{p(\mathbf{y}' \mid \theta)\, p(\mathbf{y} \mid \mathbf{w}_\perp, \theta)}{F+p(\mathbf{y} \mid \mathbf{w}_\perp, \theta) H} \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p(\mathbf{t} \mid \theta)$$

$$B_2 = \sum_{\mathbf{s}\in\mathbf{G}} p\left(\mathbf{y}' \mid \mathbf{s}, \mathbf{y}, \theta\right) \frac{p(\mathbf{s} \mid \theta)\, p(\mathbf{y} \mid \mathbf{s}, \theta)}{F+p(\mathbf{y} \mid \mathbf{w}_\perp, \theta) H} + \frac{p(\mathbf{y}' \mid \theta)}{F+p(\mathbf{y} \mid \mathbf{w}_\perp, \theta) H} \sum_{\mathbf{t}\in dom(\mathbf{S})-\mathbf{G}} p(\mathbf{t} \mid \theta)$$

Finally, the predictive distribution (Eq. (12)) can be bound as follows:

$$A_1 + B_1 \le p\left(\mathbf{y}' \mid \mathbf{y}, \theta\right) \le A_2 + B_2 \tag{13}$$

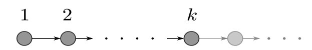
Note that as more configurations are generated, the infinite sums diminish due to the decreasing mass of the size configuration prior, and that the denominators approach $p(\mathbf{y} \mid \theta)$. As such, the predictive distribution bounds converge to the exact expression given by Eq. (12).

A Witness Algorithm for Learning

Our approach to estimating $p(\mathbf{Y}, \mathbf{Z} \mid \mathbf{S})p(\mathbf{S})$ is to search over $dom(\mathbf{S})$, and for each size configuration $\mathbf{s} \in dom(\mathbf{S})$ generated, compute a model of $p(\mathbf{Y}, \mathbf{Z} \mid \mathbf{s})$. Every step of the search process adds to the generated set $\mathbf{G}$, and using witnesses in $\mathbf{G}$ and Lem. 1 we compute bounds on the full expected likelihood $p(\mathbf{y} \mid \theta)$. This approach to probability computation is related to that proposed in [Poole, 1993].

As long as the minimal configuration (assigning 1 to every size variable) has been generated, there are no restrictions on which domain sizes are to be generated. However, we wish the full expected likelihood bounds (Lem. 1) to be as tight as possible. Since the expected likelihood induced by the min-witness configuration controls the lower-bound of the full expected likelihood, we can always select a min-witness to improve upon by expanding a successor configuration for which that min-witness is a witness, and choose parameters that improve the expected likelihood, if possible[3].

Before listing the algorithm, we illustrate the basic idea with two examples. The first is a univariate case (Fig. 1) where $\mathbf{S} = \{S\}$. Figure 1 illustrates a search up to $S = k$, at



$\mathbf{W}$	1	2	$\cdots$	k
$\mathbf{w}_\perp$	1	2	$\cdots$	k
$\mathbf{G}$	1	1, 2	$\cdots$	$1, 2, \ldots, k$

Figure 1: A trace of Alg. 1 for the case where there is only one size variable, i.e. $\mathbf{S} = \{S\}$. The witness set, min-witness, and the generated configurations are listed for every step of the search.

which point the generated set $\mathbf{G}$ contains all size configurations up to $\mathbf{S} = \{k\}$ (the set of parameters computed thus far is $\theta_1, \ldots, \theta_k$). The witness set contains only $S = k$, where all previously generated configurations are pruned from the witness, but remain in the generated set, set as $S = k$ is sufficient to witness all configurations greater than k, and is the min-witness. Expand upon the min-witness, and by Prop. 1, the lower-bound of the full expected likelihood is non-decreasing.

In the second example there are two size variables. Our search algorithm proceeds in a two-dimensional space as illustrated in Fig. 2. Starting with the initial configura-

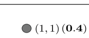
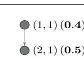
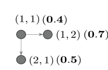
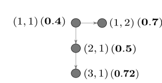
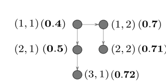
		$\mathbf{W}$	$\mathbf{w}_\perp$	$\mathbf{G}$
expand $(1,1)$		$(1,1)$	$(1,1)$	$(1,1)$
expand $(2,1)$		$(1,1)$ $(2,1)$	$(1,1)$	$(1,1)$ $(2,1)$
expand $(1,2)$ prune $(1,1)$		$(1,2)$ $(2,1)$	$(2,1)$	$(1,1)$ $(2,1)$ $(1,2)$
expand $(3,1)$ prune $(2,1)$		$(1,2)$ $(3,1)$	$(1,2)$	$(1,1)$ $(2,1)$ $(1,2)$ $(3,1)$
expand $(2,2)$		$(1,2)$ $(3,1)$ $(2,2)$	$(1,2)$	$(1,1)$ $(2,1)$ $(1,2)$ $(3,1)$ $(2,2)$

Figure 2: A trace of Alg. 1 for the case where $|\mathbf{S}| = 2$. At each step (from top to bottom), configurations are annotated with their expected likelihood scores in bold. The witness set, min-witness, and the generated sets are shown in columns on the right of the search graphs.

tion $\mathbf{S} = (1, 1)$, the two-dimensional search successively expands successors of the min-witness (with expected likelihood shown in bold next to the configurations). By Prop. 1,

[3] It is possible to choose parameters that improve the likelihood as long as the data is not fit perfectly.

successor configurations have non-decreasing expected likelihood scores. In this example, where configuration $(1,2)$ is expanded in the third step, $(1,1)$ is rendered redundant as $(2,1)$ or $(1,2)$ are can witness all unexpanded configurations previously witnessed by $(1,1)$, and are of higher score than that of $(1,1)$. The configuration $(2,1)$ can be pruned from the witness set after $(3,1)$ is expanded. All of the configurations for which $(3,1)$ is a predecessor can use $(2,1)$ or $(3,1)$ as their witnesses.

The above examples follow Alg. 1, with function $\texttt{choose\_config}(\mathbf{W})$ as one that returns a successor of the min-witness in $\mathbf{W}$. The function $\texttt{prune\_witnesses}(\mathbf{W})$ maintains a minimal set of witnesses, by removing any witness $\mathbf{w}$ when there are higher scoring witnesses in $\mathbf{W}$ that can witness the same unexplored configurations as $\mathbf{w}$. Algorithm 1 details the proposed procedure.

input : Data $\mathbf{Y} = \mathbf{y}$
output: $\tilde{\theta}, \mathbf{B}$

$\mathbf{G} = \{\mathbf{w}_0\}$, where $\mathbf{w}_0 = \{x_i : x_i = 1, i = 1, \ldots, |\mathbf{S}|\}$
$\mathbf{W} = \{\mathbf{w}_0\}$
$\mathbf{B} = \{B_0\}$ /* Marg. likhd. bounds (Lem. (1)) */
while *not terminate* **do**
  $\mathbf{s} = \texttt{choose\_config}(\mathbf{W})$
  Compute $\theta_{\mathbf{s}}$ s.t. $p\left(\mathbf{y} \mid \mathbf{s}, \theta_{\mathbf{s}}\right) \geq p\left(\mathbf{y} \mid \mathbf{w}_{\perp}, \theta_{\mathbf{w}_{\perp}}\right)$
  Compute marg.likhd bounds B /* Lem. (1) */
  Add $\mathbf{s}$ to $\mathbf{G}$, $\theta_{\mathbf{s}}$ to $\tilde{\theta}$, and B to $\mathbf{B}$
  $\mathbf{W} \leftarrow \texttt{prune\_witnesses}(\mathbf{W})$
end

Algorithm 1: Witness algorithm

Termination of Alg. 1 can be done at any time; that is, when the width of the expected likelihood bounds (during training) is sufficiently small. Alternatively, one can truncate the search according to the available resources, e.g. computing space and/or time.

Subroutines used for computing parameters $\theta_{\mathbf{s}}$ for each configuration $\mathbf{s}$ visited can be chosen according to the problem in question. For instance, in searching over the number of topics of a topic model, any algorithm for computing finite topic models can be used at each search step. Our empirical demonstration pertains to topic models.

Topic Modelling

A well-studied topic model is Latent Dirichlet allocation (LDA) [Blei et al., 2003]. In LDA, a text document consisting of a set of words is modelled as a distribution over a set of K latent *topics*, where each word in the document is generated from a K-mixture of multinomial distributions corresponding to K topics. The mixing distribution over topics is specified as a K-dimensional Dirichlet distribution. LDA parameters are estimated by inferring and marginalising out the latent topics, e.g. by variational expectation-maximisation [Blei et al., 2003]. Whilst LDA assumes a fixed value for K, K is a size variable in this experiment, and it is the only size variable. We apply Alg. 1 for searching over values of K and compute the expected liklelihood bounds described by Lem. 1 and Lem. 2.

A 10-fold cross-validation experiment was carried out with the Cora text corpus[McCallum et al., 2000]. Let $\mathbf{D}_{train}$ be the training documents, and $\mathbf{D}_{test}$ be the held-out documents. For each fold, the search was carried out for $K = 1, \ldots 150$ topics. For each size configuration $K = k$, we estimate parameters θ_k of a k-topic LDA model using a variational expectation maximisation algorithm, modified so that a k-topic LDA model can be learned by seeding on a $n < k$-topic LDA model [Blei et al., 2003][4]. (Note that we could have simply computed a k-topic LDA by random restarts until one with a better expected likelihood than those for $n < k$-topics, but our approach allows us to evaluate the worst case of our algorithm; the nested parameters case.) It outputs an approximation of the expected likelihood $p(\mathbf{D}_{train} \mid k, \theta_k)$ is an underestimate given the variational approximation in the algorithm.

At each step $K = k$, the witness configuration is $K = k$, and the witness set consists of only $K = k$, and our learned parametrisation θ includes all LDA parameters up to and including θ_k. Our prior distribution over the number of topics $p(K)$ is a Poisson distribution with parameter λ, where we evaluate $\lambda = 5, 10, 40$ for this experiment. We use a uniform prior for the value of test documents. The bounds on the full expected log-likelihood $\log p(\mathbf{D}_{train} \mid \theta)$ as stated in Lem. 1 – cross-validated over the 10 folds – are shown in Fig. 3.

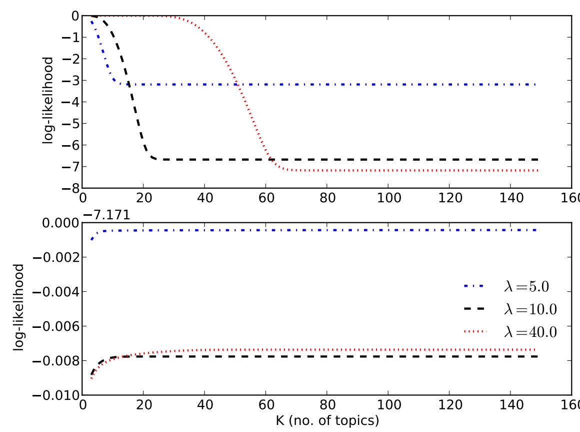


Figure 3: Bounds on $\log p(\mathbf{D}_{train} \mid \theta)$ (Lem. 1), evaluated over $K = 1, \ldots 150$ LDA topics, using a Poisson prior over K with parameters $\lambda = 5, 10, 40$.

Figure 3 shows that Lem. 1 bounds on the likelihood $p(\mathbf{D}_{train} \mid \theta)$ converges at a rate controlled by the prior distribution, where a smaller λ yields faster convergence. Although the lower-bound is increasing, it does so slowly. This is attributed to the LDA inference algorithm we use, where a $k+1$-topic LDA model seeded on a k-topic LDA model yields diminishing improvements as k increases. Next we show bounds on the posterior distribution $\log p(K \mid \mathbf{D}_{train}, \theta)$, following Lem. 2.

[4] The base code is available at www.cs.princeton.edu/ blei/lda-c

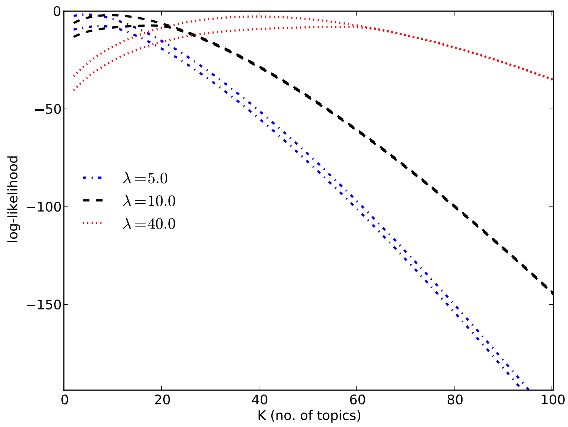


Figure 4: Bounds on $\log p(K \mid \mathbf{D}_{train}, \theta)$ (Lem. 2), where K is the number of LDA topics. Different Poisson priors over K are used, corresponding to Poisson parameters $\lambda = 5, 10, 40$. The values of K in this experiment range from 1 to 150, but this figure only shows up to $K = 100$ topics.

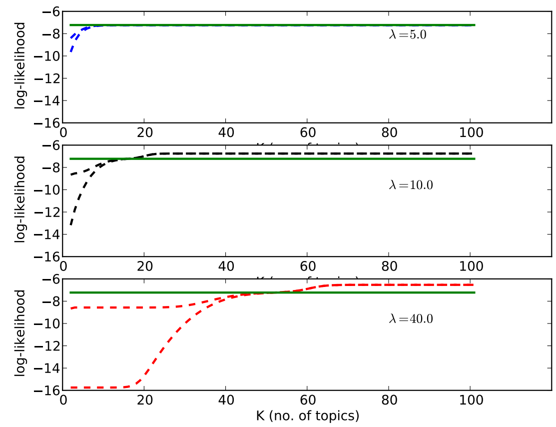


Figure 5: Comparison of Bounds on $\log p(\mathbf{D}_{test} \mid \mathbf{D}_{train}, \theta)$, shown against $p(\mathbf{D}_{test} \mid \mathbf{D}_{train}, k^*, \theta_{k^*})$ where K is the number of LDA topics, and where k^* is the number of topics for the best LDA model. The sub-plots correspond to different Poisson priors over K, for parameter values $\lambda = 5, 10, 40$. Higher log-likelihood indicates greater accuracy.

Depending on the prior distribution used, the bounds on the posterior log-likelihood of the number of topics converge at different points, where a smaller λ value again produces earlier convergence. We can use the posterior bound width to decide when to truncate the search. Searching beyond the point of convergence (up to some error threshold) likely yields models that over-fit the data, whilst truncating prematurely may yield poor fitting models.

Finally, our likelihood bounds and posterior bounds allow us to make predictions by combining all LDA models generated during search (Eq. (13)). In Fig. 5 we show bounds on the likelihood of unseen data $\mathbf{D}_{test}$ for the combined predictor, against that of the best LDA (with 148 topics). Here, the best LDA is one that with achieves the best test set accuracy, not training set accuracy. This approach intentionally biases the experiment in favour of fixed-size LDAs. Accuracy is given in log-likelihood[5].

The result of Fig. (5) shows that the combined LDA model, once converged, can achieve greater accuracy than the single best LDA model in prediction; this concurs with results from similar experiments in [Teh et al., 2006]. The effect of the prior distribution on the convergence is again evident, where smaller λ yields faster convergence. However, the steady-state log-likelihood can be improved by choosing a λ that favours a higher number of topics, albeit increasing λ indefinitely likely yields diminishing returns.

Conclusion

This paper presents a search-based approach as an alternative to current stochastic process based methods for probabilistic models with *a priori* unbounded dimensionality. Our algorithm searches in the domain of latent variable sizes, and for each size configuration visited, learns parameters for a model conditioned on the size. The method allows arbitrary prior distributions over size of latent variables, and maintains bounds on the expected likelihood of data at each step of the search. The bounds diminish as the scope of search expands, and thus search can be terminated at any time. We demonstrate our approach in the domain of text modelling, searching over the number of topics in a latent Dirichlet allocation model, and using LDA's parameter learning algorithm as a subroutine to the search. We demonstrate how the bounds could be used to guide the learning process, and on test data we showed that predictions obtained by combining all LDA models generated during search can yield close to, if not better accuracy compared to the single best LDA found by cross-validation.

In the future directions it would be interesting to evaluate the proposed approach on more complex models with multiple size variables, e.g. we may want to perform clustering for collaborative filtering models, where the number of clusters for users and items may be different and moreover interdependent. An extension that allows the prior distribution over size variables to be uncertain is another interesting step, which would allow for automated optimization of the size prior given data.

Acknowledgements

The authors would like to thank Emtiyaz Khan, Mark Crowley, David Buchman, and the anonymous reviwers for their valuable feedback.

[5] We show our results in log-likelihood, which is monotonic in *perplexity*, a standard measure of accuracy in the topic modelling and information retrieval community.

Bibliography

Aldous, D. 1983. Exchangeability and related topics. In *l'École d'été de probabilités de Saint-Flour, XIII*. Springer.

Antoniak, C. E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. of Stat.* 2:1152–1174.

Blei, D. M., and Jordan, M. I. 2005. Variational inference for Dirichlet process mixtures. *Bayesian Analysis* 1:121–144.

Blei, D. M.; Ng, A. Y.; Jordan, M. I.; and Lafferty, J. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:2003.

Cassandra, A., and Littman, M. 1994. Acting optimally in partially observed stochastic domains. In *UAI 1994*, 1023 – 1028.

Daume III, H. 2007. Fast search for Dirichlet process mixture models. In *AIStats*.

Ferguson, T. 1973. Bayesian analysis of some nonparametric problems. *Ann. of Stat.* 1(2):209–230.

Griffiths, T. L., and Ghahramani, Z. 2006. Infinite latent feature models and the Indian buffet process. In *NIPS 2006*.

Kemp, C.; Tenenbaum, J. B.; Griffiths, T. L.; Yamada, T.; and Ueda, N. 2006. Learning systems of concepts with an infinite relational model. In *AAAI 2006*.

McCallum, A.; Nigam, K.; Rennie, J.; and Seymore, K. 2000. Automating the construction of internet portals with machine learning. *J. Inform. Retrieval* 3:127–163.

Neal, R. M. 1998. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 9815, Dept. of Stat., University of Toronto.

Poole, D. 1993. Average-case analysis of a search algorithm for estimating prior and posterior probabilities in Bayesian networks with extreme probabilities. In *IJCAI 1993*, 606–612.

Rasmussen, C. E. 2006. *Gaussian processes for machine learning*. MIT Press.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* 101.