On a Discrete Dirichlet Model

Arthur Choi and Adnan Darwiche
University of California, Los Angeles
{*aychoi, darwiche*}*@cs.ucla.edu*

Abstract

The Dirichlet distribution is a statistical model that is deeply entrenched in the theory and practice of learning and reasoning in Bayesian networks. While it is mathematically convenient in certain situations, it imposes computational challenges in others, such as in Bayesian parameter learning under incomplete data. We consider in this paper a discretized variant of the continuous Dirichlet distribution. We show first how one can model such a distribution compactly as a Bayesian network, which can be used as a sub-model in other Bayesian networks. We analyze some of its theoretical properties, relating the discrete variant to the original continous density. We further show how to represent and perform exact inference efficiently in this model. We finally discuss some implications that a discrete Dirichlet model may have in enabling the design of more sophisticated models, and in enabling new ways to reason about them.

1 Introduction

Bayesian networks are prevalent in the artificial intelligence and computer science communities, and in these domains, the natural model is often the one over discrete variables. Driven by the need to reason in these models, a significant body of research has developed, giving rise to effective and powerful algorithms for both exact and approximate inference in discrete Bayesian networks (Darwiche, 2009).

In contrast, for the task of learning Bayesian networks, one typically requires continuous variables representing, for example, possible network parametrizations. The Dirichlet is a popular model for such distributions, and has certainly been the most influential distribution in Bayesian learning (DeGroot, 1970; Heckerman, 1998). While they can be mathematically convenient in certain cases, the use of the Dirichlet distributions can pose a computational barrier in other situations. In particular, the Dirichlet distribution is the conjugate prior for the parameters of a multinomial distribution, and in the task of Bayesian parameter learning, a Dirichlet prior leads to a Dirichlet posterior, given complete data. However, in the case of incomplete data, the posteriors are in general no longer Dirichlet. To compute these posteriors, we must marginalize over the hidden variables, leading to a mixture of Dirichlets, which is both analytically and computationally prohibitive.[1] In such cases, we may appeal to variational approximations or Gibbs sampling, or otherwise settle for maximum a Posteriori (MAP) parameter estimates (Heckerman, 1998).

Considering the vast body of research on reasoning in discrete models, and considering further the increasing availability of computational resources (in the form of many-core and massively parallel architectures, distributed computing, cloud computing, etc.), posing such problems in fully-discrete approximations may become a compelling alternative. Towards this end, we consider in this paper a discretized variant of the Dirichlet distribution. A naive discretization of this domain would enumerate a set of candidate distributions, which may be coming from a high-dimensional space. In con-

[1] Drawing samples from a Dirichlet distribution is another difficulty, where in practice, approaches based on rejection sampling are often used. This is another case where a discrete Dirichlet has been considered as an alternative to the continuous one (Matsui et al., 2010).

trast, we propose a natural but compact representation of this domain that can be encoded directly as a Bayesian network, allowing it to be embedded naturally in other Bayesian network models. We further show that this discrete Dirichlet sub-model is further amenable to exact inference, via a specialized belief propagation procedure which we describe.

We also analyze the theoretical properties of this discrete Dirichlet model, relating it to the original continuous distribution. We conclude by discussing some of the advantages, in terms of both modeling and reasoning, that present themselves by assuming a discrete representation of Dirichlet priors.

2 Preliminaries

Let X be a random variable taking on k possible values x_i. Let the distribution over X be parametrized by a vector $\theta_X = (\theta_{x_1}, \ldots, \theta_{x_k})$, where each θ_{x_i} is a probability and $\sum_i \theta_{x_i} = 1$. We shall refer to θ_X as a parameter set and each θ_{x_i} as a parameter. Given a specific parameter set θ_X we thus have the probability $Pr(X{=}x_i \mid \theta_X{=}(\theta_{x_1}, \ldots, \theta_{x_k})) = \theta_{x_i}$. Note that when the symbol θ_{x_i} appears in the context of θ_X, it refers to the i-th component of the parameter set θ_X .

The Dirichlet distribution is the one typically used to specify a prior probability density over parameter sets θ_X:

$$\rho(\theta_X{=}(\theta_{x_1}, \ldots, \theta_{x_k})) = \eta \prod_{i=1}^{k} [\theta_{x_i}]^{\psi_{x_i}-1} \qquad (1)$$

where the exponents ψ_{x_i} are hyper-parameters of the Dirichlet, and η is a normalizing constant:

$$\eta = 1/\int \prod_{i=1}^{k} [\theta_{x_i}]^{\psi_{x_i}-1}\ d\theta_X = \frac{\Gamma(\sum_{i=1}^{k} \psi_{x_i})}{\prod_{i=1}^{k} \Gamma(\psi_{x_i})}$$

where Γ is the gamma function.

We next present a discrete approximation of the Dirichlet distribution and discuss some of its properties and advantages in later sections.

3 A Discretized Dirichlet

Suppose we discretize the space of a parameter set $\theta_X = (\theta_{x_1}, \ldots, \theta_{x_k})$ so that each parameter θ_{x_i} takes a value from a finite set Ω_{x_i} of probabilities. Let Ω_X be the values of parameter set $\theta_X = (\theta_{x_1}, \ldots, \theta_{x_k})$ such that each $\theta_{x_i} \in \Omega_{x_i}$ and $\theta_{x_1} + \ldots + \theta_{x_k} = 1$. We can now define a discrete analogue of the Dirichlet distribution:

$$Pr(\theta_X{=}(\theta_{x_1}, \ldots, \theta_{x_k})) = \beta \prod_{i=1}^{k} [\theta_{x_i}]^{\psi_{x_i}-1} \qquad (2)$$

for $\theta_X \in \Omega_X$ where β is a normalizing constant. The exponents ψ_{x_i} are the hyper-parameters of the now discrete Dirichlet distribution.[2]

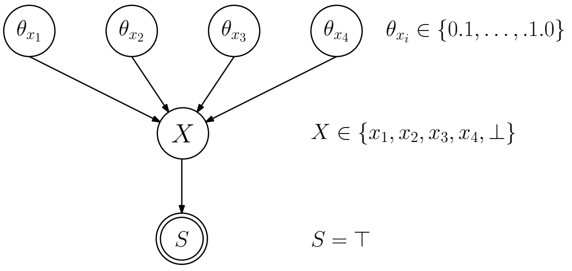
θ_{x_1}
θ_{x_2}
θ_{x_3}
θ_{x_4}
$\theta_{x_i} \in \{0.1, \ldots, .1.0\}$
X
$X \in \{x_1, x_2, x_3, x_4, \perp\}$
S
$S = \top$

Figure 1: A Bayesian network model for a discrete Dirichlet distribution.

3.1 A Bayesian Network Micro-Model

We present now a Bayesian network micro-model of the discrete analogue of the Dirichlet we just described. This network fragment can be embedded in any other discrete Bayesian network model, and may be augmented further as dictated by any prior knowledge available.

Let X be a variable with k values $x_1, \ldots, x_k$, and that we want to specify a prior over its distribution. Figure 1 illustrates an example of our model. We model each parameter θ_{x_i} of the parameter set $\theta_X = (\theta_{x_1}, \ldots, \theta_{x_k})$ as marginally independent root variables. These parameter variables serve as the prior for the CPT parameters of variable X. We include another observed variable S that enforces a constraint that the parameter set θ_X normalizes to sum to one.

[2] Note that others have also investigated discretizations of the Dirichlet for other purposes, such as (Matsui et al., 2010), who were interested in drawing samples from a discrete Dirichlet. In contrast, we propose next an explicit Bayesian network representation, and in Section 5 we provide an exact inference algorithm for it.

First, each root variable θ_{x_i} has values in Ω_{x_i}, having the following CPT:

$$Pr(\theta_{x_i}) = \alpha_{x_i}[\theta_{x_i}]^{\psi_{x_i}-1},$$

where α_{x_i} is a normalizing constant:

$$\alpha_{x_i} = 1/\sum_{\theta_{x_i}\in\Omega_{x_i}} [\theta_{x_i}]^{\psi_{x_i}-1}$$

Next, variable X has values in $\{x_1, \ldots, x_k, \perp\}$, where $\perp$ is a new distinguished state capturing invalid distributions that do not sum to one. Its CPT is as follows:

- If $\theta_{x_1} + \ldots + \theta_{x_k} = 1$, then

$$\begin{aligned} &Pr(X{=}x_i \mid \theta_{x_1}, \ldots, \theta_{x_k}) = \theta_{x_i} \text{ for all } x_i \\ &Pr(X{=}\perp \mid \theta_{x_1}, \ldots, \theta_{x_k}) = 0 \end{aligned}$$

- If otherwise $\theta_{x_1} + \ldots + \theta_{x_k} \neq 1$, then

$$\begin{aligned} &Pr(X{=}x_i \mid \theta_{x_1}, \ldots, \theta_{x_k}) = 0 \text{ for all } x_i \\ &Pr(X{=}\perp \mid \theta_{x_1}, \ldots, \theta_{x_k}) = 1 \end{aligned}$$

Finally, variable S has two values: $\top$ representing valid parameter sets, and $\perp$ representing invalid parameter sets. Its CPT is as follows:

$$\begin{aligned} Pr(S{=}\top \mid X{=}x_i) &= 1 \\ Pr(S{=}\perp \mid X{=}x_i) &= 0 \end{aligned}$$

for all x_i, and

$$\begin{aligned} Pr(S{=}\top \mid X{=}\perp) &= 0 \\ Pr(S{=}\perp \mid X{=}\perp) &= 1 \end{aligned}$$

The observation $S{=}\top$ represents a constraint that parameter sets θ_X must be valid, forcing invalid parameter sets to have probability zero.

In the extended paper, we prove the following key property of our proposed micro-model:

$$Pr(\theta_{x_1}, \ldots, \theta_{x_k} \mid S{=}\top) = \beta \prod_{i=1}^{k} [\theta_{x_i}]^{\psi_{x_i}-1} \quad (3)$$

which is precisely the distribution given in Equation 2. This basically shows that once we condition our micro-model on the observation $S{=}\top$, the model induces a discrete Dirichlet distribution on parameter sets. We examine the properties of this distribution further in Section 4, where we show how it resembles the continuous Dirichlet distribution in key aspects.

Before we proceed, we make some observations. First, observe that a naive discretization would enumerate a finite set of parameter sets $\theta_X = \{\theta_{x_1}, \ldots, \theta_{x_k}\}$, which require an intractably large selection to achieve a reasonable coverage of the domain Ω_X. Although our variable X has k parents, one for each parameter θ_{x_i}, and the CPT has a number of entries that is exponential in k, our representation remains compact as we may leave the CPT of X defined implicitly. Moreover, as our micro-model has a polytree structure, we can in fact perform inference efficiently in this implicit representation, as we show in Section 5.

3.2 An Example

Consider again our example from Figure 1. We have assumed a random variable X with four states $x_1, \ldots, x_4$, and a distribution over X has four parameters in its parameter set: $\theta_X = (\theta_{x_1}, \ldots, \theta_{x_4})$. Suppose then that each parameter can take on one of the $n = 10$ values in $\theta_{x_i} = \{0.1, 0.2, \ldots, 1.0\}$. Indeed, we can think of $\frac{1}{n}$ as a discretization granularity, and define parameter domains such as $\Omega_{x_i} = \{\frac{a}{n} \mid a \in \{1, 2, \ldots, n\}\}$. Using an exponent $\psi_{x_i} = 1$, we have the prior distribution over parameter values $Pr(\theta_{x_i}{=}p) = 0.1$ for all $p \in \Omega_{x_i}$. Using an exponent $\psi_{x_i} = 2$, we have the prior distribution $Pr(\theta_{x_i}{=}p) = \alpha \cdot p$ where $\alpha = \sum_{a=1}^{10} \frac{a}{10} = 5.5$.

4 Properties of the Discrete Dirichlet

In this section, we examine the properties of the discrete Dirichlet distribution of Equation 3, comparing it to the original continuous distribution of Equation 1. In particular, one of our goals here is to confirm that the discrete Dirichlet distribution behaves in a natural way, as compared to the continuous version that it is based on. When we make use of the original continuous Dirichlet, we refer to a density ρ, and when we make use of the discrete version, we refer to a distribution Pr.

Expected Value. The first property of our micro-model is its ability to explicitly represent the expected value of each parameter θ_{x_j}, which is known as the *Bayesian estimate* (Heckerman, 1998). In particular, we prove the following in the extended paper:

$$\begin{aligned}\mathrm{Ex}[\theta_{x_j}] &= Pr(X{=}x_j \mid S{=}\top) \\ &= \sum_{\theta_X \in \Omega_X} \theta_{x_j} \cdot Pr(\theta_X \mid S{=}\top)\end{aligned}$$

which shows that we can recover the Bayesian estimates of our parameter set θ_X directly from the distribution of variable X conditioned on observation $S{=}\top$.

Expected Value: An Exact Case. In general, the expected value of a parameter θ_{x_j} from the discrete Dirichlet is only an approximation for that of the continuous Dirichlet. However, in the special case that the Dirichlet exponents are all the same exponent ψ, the expected value is the same under both distributions:

$$\mathrm{Ex}[\theta_{x_j}] = \frac{\psi_{x_j}}{\sum_{i=1}^{k} \psi_{x_i}} = \frac{\psi}{k \cdot \psi} = \frac{1}{k}$$

which yields a parameter set θ_X that is uniform.[3] Note that this holds even in the case the discretization does not admit a uniform parameter set (i.e., $\frac{1}{k} \notin \Omega_{x_j}$); we will see an example of this in Section 5.1.

Observed Data. Suppose now that we have a data set $\mathcal{D}$ where we have observed N cases, with N_i cases observed for each value x_i of variable X. In the continuous case, we have the posterior Dirichlet:

$$\rho(\theta_X \mid \mathcal{D}) \propto \prod_{i=1}^{k} [\theta_{x_i}]^{N_i + \psi_{x_i} - 1}$$

which is a Dirichlet with exponents $N_i + \psi_{x_i}$. In the discretized case, assume that we have replicated instances of the X variable, each sharing the same parents in parent set θ_X. In the extended paper, we show the following for the discrete case

$$Pr(\theta_X \mid \mathcal{D}) \propto \prod_{i=1}^{k} [\theta_{x_i}]^{N_i + \psi_{x_i} - 1}$$

which is a discrete Dirichlet with exponents $N_i + \psi_{x_i}$, therefore, resembling the continuous distribution in this case.

Remember at this point that if we were using the Dirichlet for Bayesian parameter learning under incomplete data, the posterior is in general not Dirichlet. Analogously, if we had used a discrete Dirichlet. On the other hand, the posterior in this latter case is still a discrete distribution, which leaves us with more options in terms of performing exact and approximate inference, as we shall discuss in the next section.

5 Inference in a Discrete Dirichlet

In Section 3, we proposed an efficient representation for a discrete Dirichlet distribution, assuming that the CPT of variable X is implicitly defined. Taking advantage of the fact that the network fragment is a polytree, and that we can leave the CPT of X implicit, we propose a belief propagation (BP) algorithm for exact inference in our sub-model. The corresponding message update equations can then be used in a general belief propagation procedure for performing (possibly approximate) inference in a Bayesian network with discrete Dirichlet submodels. The inference procedure we describe may also be used in other inference frameworks, which we will discuss later in this section.

Our presentation will be similar in spirit to inference using noisy-or CPTs. The noisy-or model also has a compact representation that is linear in the number of parents, and has an efficient belief propagation procedure for performing inference (Pearl, 1988, Section 4.3.2).

First, consider the following forms for the message updates required for performing belief propagation (BP) (Pearl, 1988):

$$\pi_X(\theta_{x_j}) = Pr(\theta_{x_j}) \tag{4}$$

$$\lambda_X(\theta_{x_j}) = \sum_{(\theta_{x_1}, \ldots, \theta_{x_j}, \ldots, \theta_{x_k}) \in \Omega_X} \prod_{i \neq j} \pi_X(\theta_{x_i}) \tag{5}$$

[3] To sketch the proof, we note that if $\psi_{x_i} = \psi_{x_j}$, then θ_{x_i} and θ_{x_j} are not otherwise distinguishable, so it must be that the distribution $Pr(\theta_{x_i} \mid S{=}\top)$ is equivalent to the distribution $Pr(\theta_{x_j} \mid S{=}\top)$. By Equation 7, it follows that $\mathrm{Ex}[\theta_{x_i}] = \mathrm{Ex}[\theta_{x_j}]$ for all x_i, x_j.

Here, $\pi_X(\theta_{x_j})$ is the message passed from parameter node θ_{x_j} to its child X, and $\lambda_X(\theta_{x_j})$ is the message that X passes to its parent θ_{x_j}. Using these messages, we can compute the posterior marginals:

$$Pr(\theta_{x_j}|S=\top) \propto \pi_X(\theta_{x_j})\lambda_X(\theta_{x_j}) \qquad (6)$$

$$Pr(X=x_j|S=\top) = \sum_{\theta_{x_j}\in\Omega_{x_j}} \theta_{x_j} Pr(\theta_{x_j}|S=\top) \quad (7)$$

The key computational component in this procedure is for the message $\lambda_X(\theta_{x_j})$, which requires an efficient evaluation of terms of the following form:

$$f(I,p) = \sum_{\theta_{X_I}\in\Omega^p_{X_I}} \prod_{i\in I} \pi_X(\theta_{x_i}) \qquad (8)$$

Here, I is an index set $I \subseteq \{1,\ldots,k\}$ that selects a subset of the states x_i of variable X. Moreover, parameter set θ_{X_I} contains the selection of parameters θ_{x_i} for each index $i \in I$. Finally, $\Omega^p_{X_I}$ denotes the domain of parameter sets θ_{X_I} that sum to p, i.e., $\theta_{X_I} \in \Omega^p_{X_I}$ iff $\sum_{i\in I}\theta_{x_i} = p$. For the case of Equation 5, $I = \{1,\ldots,k\} \setminus j$ and $p = 1 - \theta_{x_j}$.

We sketch in Appendix A how to compute these summations efficiently. More specifically, suppose that we have k parameters θ_{x_k}, and we have n possible parameter values, i.e., $|\Omega_{x_i}| = n$. We sketch in the Appendix how all messages $\lambda_X(\theta_{x_j})$ can be computed in time $O(k \cdot n^2)$, which is polynomial and avoids the exponential (in k) computation required by standard belief propagation. In an extended version of the paper, we show how one can compute BP messages when this model is embedded in a network where X has parents and children. The computation of these messages are similar in spirit, and also rely primarily on Equation 8.

To conclude this section, we remark that the inference equations we have identified in this section (together with more general ones in an extended version) can be used to perform inference in a general Bayesian network that has discrete Dirichlet sub-models embedded in it. If this Bayesian network is also a polytree, the equations can be used to perform exact inference using belief propagation, where we apply updates according to Equations 4 and 5 for messages passed along edges in our discrete Dirichlet sub-model. Analogously, we can perform approximate inference in a network that is not a polytree, by using loopy belief propagation (Yedidia et al., 2003). These exact computations may also be used in approximate inference frameworks based on performing exact inference in approximate networks, such as variational approximations (Jaakkola, 2001), and generalizations of belief propagation based on structural relaxations (Choi and Darwiche, 2006; Choi and Darwiche, 2009). The latter approach, in particular, could assume a structural relaxation where discrete Dirichlet sub-models are independent (where the relaxation is later compensated for). In such an approach, one needs only to perform exact inference independently in each discrete Dirichlet sub-model.

5.1 Examples

Consider again our example from Figure 1 and Section 3.2, where we are now interested in the distribution $Pr(\theta_X \mid S=\top)$ over parameter sets, and the expected parameter values $\mathrm{Ex}[\theta_X]$ implied by it. Assuming we have Dirichlet exponents $\psi_{x_i} = 1$ for all four x_i, we have the following distribution over parameter values:

θ_{x_i}	$Pr(\theta_{x_i} \mid S=\top)$	θ_{x_i}	$Pr(\theta_{x_i} \mid S=\top)$
0.1	33.33%	0.6	3.57%
0.2	25.00%	0.7	1.19%
0.3	17.86%	0.8	0.00%
0.4	11.90%	0.9	0.00%
0.5	7.14%	1.0	0.00%

for all four x_i. Note that since each parameter θ_{x_i} must be at least 0.1, and there are four parameters θ_{x_i}, it is not possible for a parameter to have a value of 0.8, 0.9 or 1.0. The expected parameter values (the Bayesian estimates) are:

$$\mathrm{Ex}[\theta_{x_i}] = \sum_{\theta_X\in\Omega_X} \theta_{x_i} \cdot Pr(\theta_X \mid S=\top) = 0.25$$

for all x_i, which is the uniform distribution, and also the expected parameter values given by the original continuous Dirichlet.

As another example, if we have Dirichlet exponents $\psi_{x_i} = 2$, we have the following distribution over parameter values:

θ_{x_i}	$Pr(\theta_{x_i} \mid S = \top)$	θ_{x_i}	$Pr(\theta_{x_i} \mid S = \top)$
0.1	26.92%	0.6	2.10%
0.2	29.37%	0.7	0.41%
0.3	22.03%	0.8	0.00%
0.4	13.05%	0.9	0.00%
0.5	6.12%	1.0	0.00%

and again we have the parameter estimates $\text{Ex}[\theta_X] = (25\%, 25\%, 25\%, 25\%)$, as would be given by the continuous Dirichlet. Since this is a small example, we can also compute the MAP estimates $\text{argmax}_{\theta_X} Pr(\theta_X \mid S = \top)$, using a generic MAP algorithm, such as (Park and Darwiche, 2004). For the continuous Dirichlet, the expected value $\text{Ex}[\theta_X]$ with respect to density $\rho(\theta_X)$ is equivalent to the MAP estimates $\text{argmax}_{\theta_X} \rho(\theta_X)$, in this case. The discrete Dirichlet's MAP estimates are not unique here: there are $\binom{4}{2} = 6$ MAP estimates for the discrete Dirichlet, each having two parameters $\theta_{x_i} = 20.0\%$ and two parameters $\theta_{x_j} = 30.0\%$. This is due, however, to the particular discretization we used which cannot represent a uniform distribution. If we use, e.g., 20 discrete states, the discrete Dirichlet has a unique and uniform MAP estimate.

Suppose now we have exponents $(1, 2, 3, 4)$. The continuous Dirichlet yields expected parameter values $\text{Ex}[\theta_X] = (10\%, 20\%, 30\%, 40\%)$. Varying the number of discrete states n used in our discretization, we arrive at the following, increasingly accurate parameter estimates from the discrete Dirichlet (compared to the continuous ones):

n=10	(15.05%, 20.11%, 27.86%, 36.98%)
n=20	(12.38%, 19.77%, 29.08%, 38.77%)
n=50	(10.92%, 19.84%, 29.67%, 39.56%)
n=100	(10.46%, 19.91%, 29.84%, 39.79%)
n=1000	(10.05%, 19.99%, 29.98%, 39.98%)

Note that by $n = 47$ (not shown), the maximum absolute error is less than 1%.

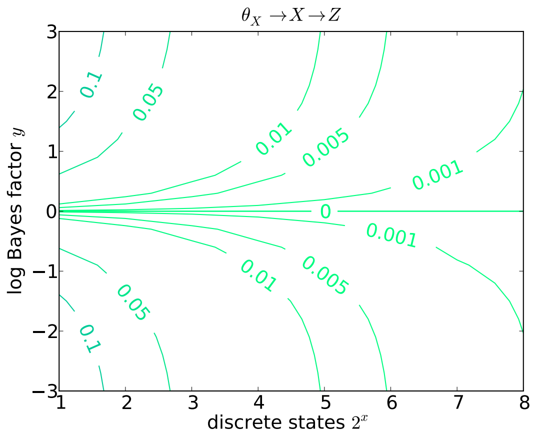

Figure 2: The error introduced by discretization. On the x-axis, we vary the number of discrete states (in powers of 2), and on the y-axis we vary the strength of the link $X \to Z$.

Consider now a network $\theta_X \to X \to Z$ where θ_X is a continuous Dirichlet, and where we have observed $Z = z$. Note that if X is unobserved, the posterior $\rho(\theta_X \mid Z = z)$ is in general only a mixture of Dirichlet's, which are generally unwieldy, both analytically and computationally (Heckerman, 1998). In contrast, if we represent this network using our discrete Dirichlet, each variable θ_{x_i} in our parameter set is just another discrete variable in a discrete Bayesian network, which we can approach using many of the tools and algorithms available to us, such as belief propagation, as we described earlier. In this sense, the discrete Dirichlet yields an approximation to the above mixture of Dirichlet's.

We evaluate the quality of this approximation in Figure 2, which depicts the absolute maximum error seen in the discrete Dirichlet's parameter estimates as compared to those obtained from the mixture of continuous Dirichlets. We assume here that variable X has two states and that we have observed only $Z = z$. We vary: (1) on the horizontal axis, the number of discrete states n used for the parameter domains $\Omega_{x_i} = \{\frac{a}{n} \mid a \in \{1, 2, \ldots, n\}\}$; and (2) on the vertical axis, the strength of link $X \to Z$:

$$\log \frac{Pr(Z = z | X = x_1)}{Pr(Z = z | X = x_2)}$$

which is the log Bayes factor for the event $X = x_1$ and observation $Z = z$. Note that for the continuous Dirichlet, we have only one hidden variable X, so it is still tractable to enumerate over all cases $X = x_i$ to compute $\rho(\theta_X \mid Z = z)$.

In Figure 2, we plot the contours where the number of discrete states n and the log Bayes factor yield a maximum absolute error of E, for different errors E. We used Dirichlet exponents $\psi_{x_1} = 1$ and $\psi_{x_2} = 1$. We observe a few general trends. First, as we increase the number of discrete states n, we find the error decreases, as expected. Second, as we increase the strength of the link $X \to Z$, we find that the error tends to increase. Third, if the link $X \to Z$ is vacuous, the discrete Dirichlet's parameter estimates are exact (they recover the uniform distribution). We note also that using only $2^5 = 32$ discrete states, the errors E for the range of y considered are below 1%.

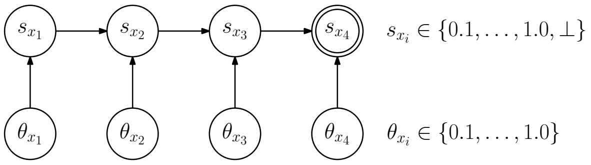

Figure 3: Another micro-model, enforcing the constraint that parameters θ_{x_i} sum to one. We maintain the cumulative sum of the parameters θ_{x_i} in the variables s_{x_i}. We clamp the end of the chain s_{x_k} to 1 as evidence. The state $\perp$ now indicates a sum that has surpassed the value 1.

6 Discussion

Continuous distributions (such as Dirichlet and logistic normal distributions) have been an integral part of learning Bayesian networks. The use of continuous distributions, however, can limit both the scalability and representational power of these models. On scalability, these distributions constrain the class of algorithms one can use to perform learning and inference with the models. On the representational side, they provide restrictions on what can be modeled as the result must fit into one of the known distributions (such as Dirichlet).

A sound and principled procedure for designing purely, or more, discrete models could potentially broaden the use and scope of learning. Topics models are a particularly relevant example (Blei and Lafferty, 2009), where there is significant interest in augmenting and designing new models, to enable new analysis and queries. One of the challenges, however, is that one generally needs to design new algorithms for learning and inference when one is dealing with a new or augmented model. In contrast, consider two points: (1) practitioners are already in general well-versed in discrete modeling, and would more easily be able to incorporate prior knowledge for their particular applications (Niculescu, 2005), and (2) there is a great body of research devoted to reasoning in discrete Bayesian networks that we can immediately take advantage of.

In this paper, we have laid some of the groundwork for a discrete model of the Dirichlet distribution, targeting the longer-term goals of developing compelling alternatives for learning and modeling Bayesian networks. Given the Bayesian network micro-model for the Bayesian network, and the exact and efficient algorithm for reasoning in it, we are now in a position to start developing new learning algorithms, formulated in terms of performing (approximate) inference in a meta-network for Bayesian parameter learning (Darwiche, 2009), as we discussed in Section 5.

A Inference: Proof Sketch

We sketch here how to efficiently compute the messages $\lambda_X(\theta_{x_j})$ of Equation 5, which is the central computational component for performing exact inference in the discrete Dirichlet submodel. Our approach is based on message passing in an augmented network where every node has at most two parents. Consider first a network where we marginalize out variable X:

$$\begin{aligned} Pr(S{=}\top \mid \theta_X) &= \sum_X Pr(S{=}\top \mid X) Pr(X \mid \theta_X) \\ &= \sum_{X \neq \perp} Pr(X \mid \theta_X) \end{aligned}$$

since $Pr(S{=}\top \mid X{=}\perp) = 0$, and 1 otherwise for all $X{=}x_i$. If $\theta_X \in \Omega_X$ (it sums to one), then

$$\sum_{i=1}^{k} Pr(X{=}x_i \mid \theta_X) = \sum_{i=1}^{k} \theta_{x_i} = 1$$

and zero otherwise (when $\theta_X \notin \Omega_X$). Variable S now has parameter nodes θ_{x_i} as direct parents. We now augment the variable S, which enforces the sum-to-one constraint, into a chain that enforces this constraint by accumulating the sum of the parameters θ_{x_i}; see Figure 3. Here, $Pr(s_{x_i} \mid s_{x_{i-1}}, \theta_{x_i}) = 1$ iff $s_{x_i} = s_{x_{i-1}} + \theta_{x_i}$, and $Pr(s_{x_1} \mid \theta_{x_1}) = 1$ iff $s_{x_1} = \theta_{x_1}$.

Consider now a message passed from s_{x_i} to $s_{x_{i+1}}$, for some $1 < i \leq k$:

$$\begin{aligned} &\pi_{s_{x_{i+1}}}(s_{x_i}) \\ &= \sum_{s_{x_{i-1}}} \sum_{\theta_{x_i}} Pr(s_{x_i} | s_{x_{i-1}}, \theta_{x_i}) \pi_{s_{x_i}}(s_{x_{i-1}}) \pi_{s_{x_i}}(\theta_{x_i}) \\ &= \sum_{s_{x_{i-1}} + \theta_{x_i} = s_{x_i}} \pi_{s_{x_i}}(s_{x_{i-1}}) \pi_{s_{x_i}}(\theta_{x_i}) \qquad (9) \end{aligned}$$

since $Pr(s_{x_i} \mid s_{x_{i-1}}, \theta_{x_i}) = 0$ if $s_{x_{i-1}} + \theta_{x_i} \neq s_{x_i}$. By recursively substituting this equation for messages $\pi_{s_{x_i}}(s_{x_{i-1}})$, we find that:

$$\pi_{s_{x_{i+1}}}(s_{x_i}) = \sum_{\theta_{x_1} + \cdots + \theta_{x_i} = s_{x_i}} \prod_{j=1}^{i} \pi_{s_{x_j}}(\theta_{x_j})$$

which is a summation of the form in Equation 8. We can then compute Equation 8 for a given I and p by permuting the variables in our network so that indices I appear at the head of the chain. To compute all of the messages of Equation 5, however, we need only perform message-passing once in the network of Figure 3, since one can show that the messages $\lambda_{s_{x_i}}(\theta_{x_i})$ will be the messages $\lambda_X(\theta_{x_i})$ that we desire.

Computing a message takes (at most) $O(n)$ time for each of its $O(n)$ entries (as in Equation 9). There are $2k - 1$ edges in this model, so to compute all messages of Equation 8, we require only $O(k \cdot n^2)$ time.

References

David M. Blei and John D. Lafferty. 2009. Topic models. In Ashok Srivastava and Mehran Sahami, editors, *Text Mining: Classification, Clustering, and Applications*, chapter 4, pages 71–93. Chapman and Hall/CRC.

Arthur Choi and Adnan Darwiche. 2006. An edge deletion semantics for belief propagation and its practical impact on approximation quality. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 1107–1114.

Arthur Choi and Adnan Darwiche. 2009. Relax then compensate: On max-product belief propagation and more. In *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*, pages 351–359.

Adnan Darwiche. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press.

Morris H. DeGroot. 1970. *Optimal Statistical Decisions*. McGraw-Hill.

David Heckerman. 1998. A tutorial on learning with Bayesian networks. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. MIT Press.

Tommi Jaakkola. 2001. Tutorial on variational approximation methods. In D. Saad and M. Opper, editors, *Advanced Mean Field Methods*, chapter 10, pages 129–160. MIT Press.

Tomomi Matsui, Mitsuo Motoki, Naoyuki Kamatani, and Shuji Kijima. 2010. Polynomial time approximate or perfect samplers for discretized Dirichlet distribution. *Japan Journal of Industrial and Applied Mathematics*. To appear (currently published online).

Radu Stefan Niculescu. 2005. *Exploiting Parameter Domain Knowledge for Learning in Bayesian Networks*. Ph.D. thesis, Carnegie Mellon University.

James Park and Adnan Darwiche. 2004. A differential semantics for jointree algorithms. *Artificial Intelligence*, 156:197–216.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, California.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. 2003. Understanding belief propagation and its generalizations. In Gerhard Lakemeyer and Bernhard Nebel, editors, *Exploring Artificial Intelligence in the New Millennium*, chapter 8, pages 239–269. Morgan Kaufmann.