Lifted Probabilistic Inference for Asymmetric Graphical Models

Guy Van den Broeck
Department of Computer Science
KU Leuven, Belgium
guy.vandenbroeck@cs.kuleuven.be

Mathias Niepert
Computer Science and Engineering
University of Washington, Seattle
mniepert@cs.washington.edu

Abstract

Lifted probabilistic inference algorithms have been successfully applied to a large number of symmetric graphical models. Unfortunately, the majority of real-world graphical models is asymmetric. This is even the case for relational representations when evidence is given. Therefore, more recent work in the community moved to making the models symmetric and then applying existing lifted inference algorithms. However, this approach has two shortcomings. First, all existing over-symmetric approximations require a relational representation such as Markov logic networks. Second, the induced symmetries often change the distribution significantly, making the computed probabilities highly biased. We present a framework for probabilistic sampling-based inference that only uses the induced approximate symmetries to propose steps in a Metropolis-Hastings style Markov chain. The framework, therefore, leads to improved probability estimates while remaining unbiased. Experiments demonstrate that the approach outperforms existing MCMC algorithms.

Introduction

Probabilistic graphical models are successfully used in a wide range of applications. Inference in these models is intractable in general and, therefore, approximate algorithms are mostly applied. However, there are several probabilistic graphical models for which inference is tractable due to (conditional) independences and the resulting low treewidth (Darwiche 2009; Koller and Friedman 2009). Examples of the former class of models are chains and tree models. More recently, the AI community has uncovered additional statistical properties based on symmetries of the graphical model that render inference tractable (Niepert and Van den Broeck 2014). In the literature, approaches exploiting these symmetries are often referred to as lifted or symmetry-aware inference algorithms (Poole 2003; Kersting 2012).

While lifted inference algorithms perform well for highly symmetric graphical models, they depend heavily on the presence of symmetries and perform worse for asymmetric models due to their computational overhead. This is especially unfortunate as numerous real-world graphical models are not symmetric. To bring the achievements of the lifted inference community to the mainstream of machine learning and uncertain reasoning it is crucial to explore ways to apply ideas from the lifted inference literature to inference problems in asymmetric graphical models.

Recent work has introduced methods to generate symmetric approximations of probabilistic models (Van den Broeck and Darwiche 2013; Venugopal and Gogate 2014; Singla, Nath, and Domingos 2014). All of these approaches turn approximate symmetries, that is, symmetries that "almost" hold in the probabilistic models, into perfect symmetries, and proceed to apply lifted inference algorithms to the symmetrized model. These approaches were shown to perform well but are also limited in a fundamental way. The introduction of artificial symmetries results in marginal probabilities that are different from the ones of the original model. The per variable Kullback-Leibler divergence, a measure often used to assess the performance of approximate inference algorithms, might improve when these symmetries are induced but it is possible that the marginals the user actually cares about are highly biased. Of course, this is a potential problem in applications. For instance, consider a medical application where one queries the probability of diseases given symptoms. A symmetric approximation may perform well in terms of the KL divergence but might skew the probabilities of the most probable diseases to become equal. A major argument for graphical models is the need to detect subtle differences in the posterior, which becomes impossible when approximate symmetries skew the distribution.

To apply lifted inference to asymmetric graphical models we propose a completely novel approach. As in previous approaches, we compute a symmetric approximation of the original model but leverage the symmetrized model to compute a proposal distribution for a Metropolis-Hastings chain. The approach combines a base MCMC algorithm such as the Gibbs sampler with the Metropolis chain that performs jumps in the symmetric model. The novel framework allows us to utilize work on approximate symmetries such as color passing algorithms (Kersting et al. 2014) and low-rank Boolean matrix approximations (Van den Broeck and Darwiche 2013) while producing unbiased probability estimates. We identify properties of an approximate symmetry group that make it suitable for the novel lifted Metropolis-Hastings approach.

We conduct experiments where, for the first time, lifted

inference is applied to graphical models with no exact symmetries and no color-passing symmetries, and where every random variable has distinct soft evidence. The framework, therefore, leads to improved probability estimates while remaining unbiased. Experiments demonstrate that the approach outperforms existing MCMC algorithms.

Background

We review some background on concepts and methods used throughout the remainder of the paper.

Group Theory

A group is an algebraic structure $(\mathfrak{G}, \circ)$, where $\mathfrak{G}$ is a set closed under a binary associative operation $\circ$ with an identity element and a unique inverse for each element. We often write $\mathfrak{G}$ rather than $(\mathfrak{G}, \circ)$. A permutation group acting on a set Ω is a set of bijections $\mathfrak{g} : \Omega \to \Omega$ that form a group. Let Ω be a finite set and let $\mathfrak{G}$ be a permutation group acting on Ω. If $\alpha \in \Omega$ and $\mathfrak{g} \in \mathfrak{G}$ we write $\alpha^{\mathfrak{g}}$ to denote the image of α under $\mathfrak{g}$. A cycle $(\alpha_1\ \alpha_2\ ...\ \alpha_n)$ represents the permutation that maps α_1 to α_2, α_2 to α_3,..., and α_n to α_1. Every permutation can be written as a product of disjoint cycles. A generating set R of a group is a subset of the group's elements such that every element of the group can be written as a product of finitely many elements of R and their inverses.

We define a relation $\sim$ on Ω with $\alpha \sim \beta$ if and only if there is a permutation $\mathfrak{g} \in \mathfrak{G}$ such that $\alpha^{\mathfrak{g}} = \beta$. The relation partitions Ω into equivalence classes which we call orbits. We call this partition of Ω the orbit partition induced by $\mathfrak{G}$. We use the notation $\alpha^{\mathfrak{G}}$ to denote the orbit $\{\alpha^{\mathfrak{g}} \mid \mathfrak{g} \in \mathfrak{G}\}$ containing α.

Symmetries of Graphical Models

Symmetries of a set of random variables and graphical models have been formally defined in the lifted and symmetry-aware probabilistic inference literature with concepts from group theory (Niepert 2012b; Bui, Huynh, and Riedel 2013).

Definition 1. Let $\mathbf{X}$ be a set of discrete random variables with distribution π and let Ω be the set of states (configurations) of $\mathbf{X}$. We say that a permutation group $\mathfrak{G}$ acting on Ω is an automorphism group for $\mathbf{X}$ if and only if for all $\mathbf{x} \in \Omega$ and all $g \in \mathfrak{G}$ we have that $\pi(\mathbf{x}) = \pi(\mathbf{x}^g)$.

Note that we do not require the automorphism group to be maximal, that is, it can be a subgroup of a different automorphism group for the same set of random variables. Moreover, note that the definition of an automorphism group is independent of the particular representation of the probabilistic model. For particular representations, there are efficient algorithms for computing the automorphism groups exploiting the structure of relational and propositional graphical models (Niepert 2012b; 2012a; Bui, Huynh, and Riedel 2013).

Most probabilistic models are asymmetric. For instance, the Ising model which is used in numerous applications, is asymmetric if we assume an external field as it leads to different unary potentials. However, we can make the model symmetric simply by assuming a constant external field. Figure 1 depicts this situation. The framework we propose in this paper will take advantage of such an over-symmetric model without biasing the probability estimates.

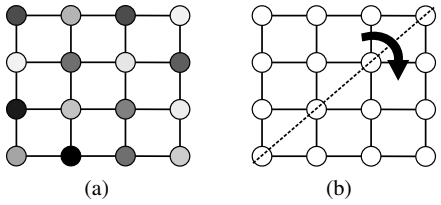


Figure 1: A ferromagnetic Ising model with constant interaction strength. In the presence of an external field, that is, when the variables have different unary potentials, the probabilistic model is asymmetric (a). However, the model is rendered symmetric by assuming a constant external field (b). In this case, the symmetries of the model are generated by the reflection and rotation automorphisms.

Exploiting Symmetries for Lifted Inference

The advent of high-level representations of probabilistic graphical models, such as plate models and relational representations (Getoor and Taskar 2007; De Raedt et al. 2008), have motivated a new class of *lifted inference* algorithms (Poole 2003). These algorithms exploit the high-level structure and symmetries to speed up inference (Kersting 2012). Surprisingly, they perform tractable inference even in the absence of conditional independencies.

Our current understanding of exact lifted inference is that syntactic properties of relational representations permit efficient lifted inference (Van den Broeck 2011; Jaeger and Van den Broeck 2012; Van den Broeck 2013; Gribkoff, Van den Broeck, and Suciu 2014). The Appendix will review such representations, and Markov logic in particular. More recently, it has been shown that (partial) exchangeability as a statistical property can explain numerous results in this literature (Niepert and Van den Broeck 2014). Indeed, there are deep connections between automorphisms and exchangeability (Niepert 2012b; 2013; Bui, Huynh, and Riedel 2013; Bui, Huynh, and de Salvo Braz 2012). Moreover, the (fractional) automorphisms of the graphical model representation have been related to lifted inference and exploited for more efficient inference (Niepert 2012b; Bui, Huynh, and Riedel 2013; Noessner, Niepert, and Stuckenschmidt 2013; Mladenov and Kersting 2013). In particular, there are a number of sampling algorithms that take advantage of symmetries (Venugopal and Gogate 2012; Gogate, Jha, and Venugopal 2012). However, these approaches expect a relational representation and require the model to be symmetric.

Finite Markov Chains

Given a finite set Ω a *Markov chain* defines a random walk $(\mathbf{x}_0, \mathbf{x}_1, ...)$ on elements of Ω with the property that the conditional distribution of $\mathbf{x}_{n+1}$ given $(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_n)$ depends only on $\mathbf{x}_n$. For all $\mathbf{x}, \mathbf{y} \in \Omega$, $P(\mathbf{x} \to \mathbf{y})$ is the chain's probability to transition from $\mathbf{x}$ to $\mathbf{y}$, and $P^t(\mathbf{x} \to \mathbf{y}) = P^t_{\mathbf{x}}(\mathbf{y})$ the probability of being in state $\mathbf{y}$ after t steps if the chain

starts at state $\mathbf{x}$. We often refer to the conditional probability matrix P as the *kernel* of the Markov chain. A Markov chain is *irreducible* if for all $\mathbf{x}, \mathbf{y} \in \Omega$ there exists a t such that $P^t(\mathbf{x} \rightarrow \mathbf{y}) > 0$ and *aperiodic* if for all $\mathbf{x} \in \Omega$, $\gcd\{t \geq 1 \mid P^t(\mathbf{x} \rightarrow \mathbf{x}) > 0\} = 1$.

Theorem 1. *Any irreducible and aperiodic Markov chain has exactly one stationary distribution.*

A distribution π on Ω is reversible for a Markov chain with state space Ω and transition probabilities P, if for every $\mathbf{x}, \mathbf{y} \in \Omega$

$$\pi(\mathbf{x})P(\mathbf{x} \rightarrow \mathbf{y}) = \pi(\mathbf{y})P(\mathbf{y} \rightarrow \mathbf{x}).$$

We say that a Markov chain is reversible if there exists a reversible distribution for it. The AI literature often refers to reversible Markov chains as Markov chains satisfying the detailed balance property.

Theorem 2. *Every reversible distribution for a Markov chain is also a stationary distribution for the chain.*

Markov Chains for Probability Estimation Numerous approximate inference algorithms for probabilistic graphical models draw sample points from a Markov chain whose stationary distribution is that of the probabilistic model, and use the sample points to estimate marginal probabilities. Sampling approaches of this kind are referred to as Markov chain Monte Carlo methods. We discuss the Gibbs sampler, a sampling algorithm often used in practice.

Let $\mathbf{X}$ be a finite set of random variables with probability distribution π. The Markov chain for the *Gibbs sampler* is a Markov chain $\mathcal{M} = (\mathbf{x}_0, \mathbf{x}_1, ...)$ which, being in state $\mathbf{x}_t$, performs the following steps at time $t + 1$:

1. Select a variable $X \in \mathbf{X}$ uniformly at random;
2. Sample $\mathbf{x}'_{t+1}(X)$, the value of X in the state $\mathbf{x}'_{t+1}$, according to the conditional π-distribution of X given that all other variables take their values according to $\mathbf{x}_t$; and
3. Let $\mathbf{x}'_{t+1}(Y) = \mathbf{x}_t(Y)$ for all variables $Y \in \mathbf{X} \setminus \{X\}$.

The Gibbs chain is aperiodic and has π as a stationary distribution. If the chain is irreducible, then the marginal estimates based on sample points drawn from the chain are unbiased once the chain reaches the stationary distribution.

Two or more Markov chains can be combined by constructing mixtures and compositions of the kernels (Tierney 1994). Let P_1 and P_2 be the kernels for two Markov chains $\mathcal{M}_1$ and $\mathcal{M}_2$ both with stationary distribution π. Given a positive probability $0 < \alpha < 1$, a *mixture* of the Markov chains is a Markov chain where, in each iteration, kernel P_1 is applied with probability α and kernel P_2 with probability $1 - \alpha$. The resulting Markov chain has π as a stationary distribution. The following result relates properties of the individual chains to properties of their mixture.

Theorem 3 (Tierney 1994)**.** *A mixture of two Markov chains $\mathcal{M}_1$ and $\mathcal{M}_2$ is irreducible and aperiodic if at least one of the chains is irreducible and aperiodic.*

For a more in-depth discussion of combining Markov chains and the application to machine learning, we refer the interested reader to an overview paper (Andrieu et al. 2003).

Mixing Symmetric and Asymmetric Markov Chains

We propose a novel MCMC framework that constructs *mixtures* of Markov chains where one of the chains operates on the *approximate symmetries* of the probabilistic model. The framework assumes a base Markov chain $\mathcal{M}_\mathtt{B}$ such as the Gibbs chain, the MC-SAT chain (Poon and Domingos 2006), or any other MCMC algorithm. We then construct a mixture of the base chain and an Orbital Metropolis chain which exploits approximate symmetries for its proposal distribution. Before we describe the approach in more detail, let us first review Metropolis samplers.

Metropolis-Hastings Chains

The construction of a Metropolis-Hastings Markov chain is a popular general procedure for designing reversible Markov chains for MCMC-based estimation of marginal probabilities. Metropolis-Hastings chains are associated with a proposal distribution $Q(\cdot|\mathbf{x})$ that is utilized to *propose* a move to the next state given the current state $\mathbf{x}$. The closer the proposal distribution to the distribution π to be estimated, that is, the closer $Q(\mathbf{x} \mid \mathbf{x}_t)$ to $\pi(\mathbf{x})$ for large t, the better the convergence properties of the Metropolis-Hastings chain.

We first describe the Metropolis algorithm, a special case of the Metropolis-Hastings algorithm (Häggström 2002). Let $\mathbf{X}$ be a finite set of random variables with probability distribution π and let Ω be the set of states of the random variables. The Metropolis chain is governed by a transition graph $G = (\Omega, \mathbf{E})$ whose nodes correspond to states of the random variables. Let $\mathtt{n}(\mathbf{x})$ be the set of neighbors of state $\mathbf{x}$ in G, that is, all states reachable from $\mathbf{x}$ with a single transition. The Metropolis chain with graph G and distribution π has transition probabilities

$$P(\mathbf{x} \rightarrow \mathbf{y}) =$$

$$\begin{cases} \frac{1}{|\mathtt{n}(\mathbf{x})|} \min\left\{\frac{\pi(\mathbf{y})|\mathtt{n}(\mathbf{x})|}{\pi(\mathbf{x})|\mathtt{n}(\mathbf{y})|}, 1\right\}, & \text{if } x \text{ and } y \text{ are neighbors} \\ 0, & \text{if } x \neq y \text{ are not neighbors} \\ 1 - \sum\limits_{y' \in \mathtt{n}(\mathbf{x})} \frac{1}{|\mathtt{n}(\mathbf{x})|} \min\left\{\frac{\pi(\mathbf{y}')|\mathtt{n}(\mathbf{x})|}{\pi(\mathbf{x})|\mathtt{n}(\mathbf{y}')|}, 1\right\}, & \text{if } x = y. \end{cases}$$

Being in state $\mathbf{x}_t$ of the Markov chain $\mathcal{M} = (\mathbf{x}_0, \mathbf{x}_1, ...)$, the Metropolis sampler therefore performs the following steps at time $t + 1$:

1. Select a state $\mathbf{y}$ from $\mathtt{n}(\mathbf{x}_t)$, the neighbors of $\mathbf{x}_t$, uniformly at random;
2. Let $\mathbf{x}_{t+1} = \mathbf{y}$ with probability $\min\left\{\frac{\pi(\mathbf{y})|\mathtt{n}(\mathbf{x})|}{\pi(\mathbf{x})|\mathtt{n}(\mathbf{y})|}, 1\right\}$;
3. Otherwise, let $\mathbf{x}_{t+1} = \mathbf{x}_t$.

Note that the proposal distribution $Q(\cdot|\mathbf{x})$ is simply the uniform distribution on the set of $\mathbf{x}$'s neighbors. It is straight-forward to show that π is a stationary distribution for the Metropolis chain by showing that π is a reversible distribution for it (Häggström 2002).

Now, the performance of the Metropolis chain hinges on the structure of the graph G. We would like the graph structure to facilitate global moves between high probability

modes, as opposed to the local moves typically performed by MCMC chains. To design such a graph structure, we take advantage of approximate symmetries in the model.

Orbital Metropolis Chains

We propose a novel class of *orbital* Metropolis chains that move between approximate symmetries of a distribution. The approximate symmetries form an automorphism group $\mathfrak{G}$. We will discuss approaches to obtain such an automorphism group in Section . Here, we introduce a novel Markov chain that takes advantage of the approximate symmetries.

Given a distribution π over random variables $\mathbf{X}$ with state space Ω, and a permutation group $\mathfrak{G}$ acting on Ω, the orbital Metropolis chain $\mathcal{M}_{\mathtt{S}}$ for $\mathfrak{G}$ performs the following steps:

1. Select a state $\mathbf{y}$ from $\mathbf{x}_t^{\mathfrak{G}}$, the orbit of $\mathbf{x}_t$, uniformly at random;
2. Let $\mathbf{x}_{t+1} = \mathbf{y}$ with probability $\min\left\{\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}, 1\right\}$;
3. Otherwise, let $\mathbf{x}_{t+1} = \mathbf{x}_t$.

Note that a permutation group acting on Ω partitions the states into disjoint orbits. The orbital Metropolis chain simply moves between states in the same orbit. Hence, two states in the same orbit have the same number of neighbors and, thus, the expressions cancel out in line 2 above. It is straight-forward to show that the chain $\mathcal{M}_{\mathtt{S}}$ is reversible and, hence, that it has π as a stationary distribution. However, the chain is *not* irreducible as it never moves between states that are not symmetric with respect to the permutation group $\mathfrak{G}$. In the binary case, for example, it cannot reach states with a different Hamming weight from the initial state.

Lifted Metropolis-Hastings

To obtain an irreducible Markov chain that exploits approximate symmetries, we construct a mixture of (a) some base chain $\mathcal{M}_{\mathtt{B}}$ with stationary distribution π for which we know that it is irreducible and aperiodic; and (b) an orbital Metropolis chain $\mathcal{M}_{\mathtt{S}}$. We can prove the following theorem.

Theorem 4. *Let $\mathbf{X}$ be a set of random variables with distribution π and approximate automorphisms $\mathfrak{G}$. Moreover, let $\mathcal{M}_{\mathtt{B}}$ be an aperiodic and irreducible Markov chain with stationary distribution π, and let $\mathcal{M}_{\mathtt{S}}$ be the orbital Metropolis chain for $\mathbf{X}$ and $\mathfrak{G}$. The mixture of $\mathcal{M}_{\mathtt{B}}$ and $\mathcal{M}_{\mathtt{S}}$ is aperiodic, irreducible, and has π as its unique stationary distribution.*

The mixture of the base chain and the orbital Metropolis chain has several advantages. First, it exploits the approximate symmetries of the model which was shown to be advantageous for marginal probability estimation (Van den Broeck and Darwiche 2013). Second, the mixture of Markov chains performs wide ranging moves via the orbital Metropolis chain, exploring the state space more efficiently and, therefore, improving the quality of the probability estimates. Figure 2 depicts the state space and the transition graph of (a) the Gibbs chain and (b) the mixture of the Gibbs chain and an orbital Metropolis chain. It illustrates that the mixture is able to more freely move about the state space by jumping between orbit states. For instance, moving from state 0110 to 1001 would require 4 steps of the Gibbs chain

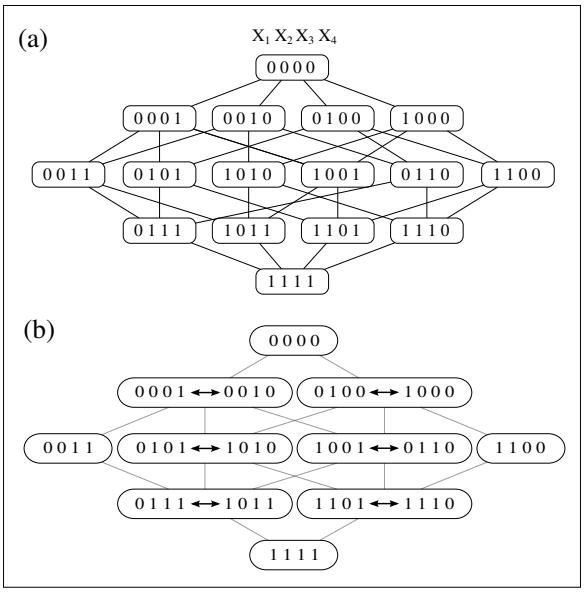


Figure 2: The state space (self-arcs are omitted) of (a) the Gibbs chain for four binary random variables and (b) the orbit partition of its state space induced by the permutation group generated by the permutation $(X_1\ X_2)(X_3\ X_4)$. The permutations are approximate symmetries, derived from an over-symmetric approximation of the original model. The Gibbs chain proposes moves to states whose Hamming distance to the current state is at most 1. The orbital Metropolis chain, on the other hand, proposes moves between orbit elements which have a Hamming distance of up to 4. The mixture of the two chains leads to faster convergence while maintaining an unbiased stationary distribution.

but is possible in one step with the mixture of chains. The larger the size of the automorphism groups, the more densely connected is the transition graph. Since the moves of the orbital Metropolis chain are between approximately symmetric states of the random variables, it does not suffer from the problem of most proposals being rejected. We will be able to verify this hypothesis empirically.

The general Lifted Metropolis-Hastings framework can be summarized as follows.

1. Obtain an approximate automorphism group $\mathfrak{G}$;
2. Run the following mixture of Markov chains:
   (a) With probability $0 < \alpha < 1$, apply the kernel of the base chain $\mathcal{M}_{\mathtt{B}}$;
   (b) Otherwise, apply the kernel of the orbital Metropolis chain $\mathcal{M}_{\mathtt{S}}$ for $\mathfrak{G}$.

Note that the proposed approach is a generalization of lifted MCMC (Niepert 2013; 2012b), essentially using it as a subroutine, and that all MH proposals are accepted if $\mathfrak{G}$ is an automorphism group of the original model. Moreover, note that the framework allows one to combine multiple orbital Metropolis chains with a base chain.

Approximate Symmetries

The Lifted Metropolis-Hastings algorithm assumes that a permutation group $\mathfrak{G}$ is given, representing the approximate symmetries. We now discuss several approaches to the computation of such an automorphism group. While it is not possible to go into technical detail here, we will provide pointers to the relevant literature.

There exist several techniques to compute the *exact symmetries* of a graphical model and construct $\mathfrak{G}$; see (Niepert 2012b; Bui, Huynh, and Riedel 2013). The color refinement algorithm is also well-studied in lifted inference (Kersting et al. 2014). It can find (exact) orbits of random variables for a slightly weaker notion of symmetry, called fractional automorphism. These techniques all require some form of exact symmetry to be present in the model.

Detecting *approximate symmetries* is a problem that is largely open. One key idea is that of an *over-symmetric approximations* (OSAs) (Van den Broeck and Darwiche 2013). Such approximations are derived from the original model by rendering the model more symmetric. After the computation of an over-symmetric model, we can apply existing tools for exact symmetry detection. Indeed, the exact symmetries of an approximate model are approximate symmetries of the exact model. These symmetrization techniques are indispensable to our algorithm.

Relational Symmetrization Existing symmetrization techniques operate on relational representations, such as Markov logic networks (MLNs). The full paper reviews MLNs and shows a web page classification model. Relational models have numerous symmetries. For example, swapping the web pages A and B does not change the MLN. This permutation of constants induces a permutations of random variables (e.g., between $\texttt{Page}(\mathsf{A}, \mathsf{Faculty})$ and $\texttt{Page}(\mathsf{B}, \mathsf{Faculty})$). Unfortunately, hard and soft evidence breaks symmetries, even in highly symmetric relational models (Van den Broeck and Darwiche 2013). When the variables $\texttt{Page}(\mathsf{A}, \mathsf{Faculty})$ and $\texttt{Page}(\mathsf{B}, \mathsf{Faculty})$ get assigned distinct soft evidence, the symmetry between A and B is removed, and lifted inference breaks down.[1] Similarly, when the `Link` relation is given, its graph is unlikely to be symmetric (Erdős and Rényi 1963), which in turn breaks the symmetries in the MLN. These observations motivated research on OSAs. Van den Broeck and Darwiche (2013) propose to approximate binary relations, such as `Link`, by a low-rank Boolean matrix factorization. Venugopal and Gogate (2014) cluster the constants in the domain of the MLN. Singla, Nath, and Domingos (2014) present a message-passing approach to clustering similar constants.

Propositional Symmetrization A key property of our LMH algorithm is that it operates at the propositional level, regardless of how the graphical model was generated. It also means that the relational symmetrization approaches outlined above are inadequate in the general case. Unfortunately, we are not aware of any work on OSAs of propositional graphical models. However, some existing techniques provide a promising direction. First, basic clustering can group together similar potentials. Second, the low-rank Boolean matrix factorization used for relational approximations can be applied to any graph structure, including graphical models. Third, color passing techniques for exact symmetries operate on propositional models (Kersting, Ahmadi, and Natarajan 2009; Kersting et al. 2014). Combined with early stopping, they can output approximate variable orbits.

From OSAs to Automorphisms Given an OSA of our model, we need to compute an automorphism group $\mathfrak{G}$ from it. The obvious choice is to compute the exact automorphisms from the OSA. While this works in principle, it may not be optimal. Let us first consider the following two concepts. When a group $\mathfrak{G}$ operates on a set Ω, only a subset of the elements in Ω can actually be mapped to an element other than itself. When Ω is the set of random variables, we call these elements the *moved variables*. When Ω is the set of potentials in a probabilistic graphical model, we call these the *moved potentials*. It is clear that we want $\mathfrak{G}$ to move many random variables, as this will create the largest jumps and improve the mixing behavior. However, each LMH step comes at a cost: in the second step of the algorithm, the probability of the proposed approximately-symmetric state $\pi(\mathbf{y})$ is estimated. This requires the re-evaluation of all potentials that are moved by $\mathfrak{G}$. Thus, the time complexity of an orbital Metropolis step is linear in the number of moved potentials. It will therefore be beneficial to construct *subgroups* of the automorphism group of the OSA and, in particular, ones that move many variables and few potentials. The full paper discusses a heuristic to construct such subgroups.

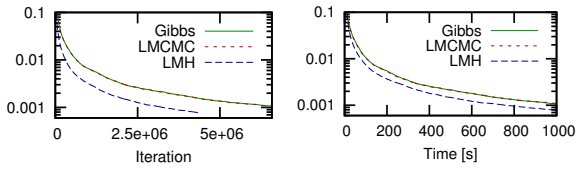
Experiments

The LMH algorithm is implemented in the GAP algebra system which provides basic algorithms for automorphism groups such as the product replacement algorithm that allows one to sample uniformly from orbits of states (Niepert 2012b).

For our first experiments, we use the standard WebKB data set, consisting of web pages from four computer science departments (Craven and Slattery 2001). The data has information about approximately 800 words that appear on 1000 pages, 7 page labels and links between web pages. There are 4 folds, one for each university. We use the standard MLN structure for the WebKB domain, which has MLN formulas of the form shown above, but for all combinations of labels and words, adding up to around 5500 first-order MLN formulas. We learn the MLN parameters using Alchemy.

We consider a collective classification setting, where we are given the link structure and the word content of each web page, and want to predict the page labels. We run Gibbs sampling and the Lifted MCMC algorithm (Niepert 2012b), and show the average KL divergence between the estimated and true marginals in Figures 3 and 4. When true marginals are not computable, we used a very long run of a Gibbs sampler for the gold standard marginals. Since every web page contains a unique set of words, the evidence on the word content creates distinct soft evidence on the page labels. Moreover, the link structure is largely asymmetric and, therefore, there

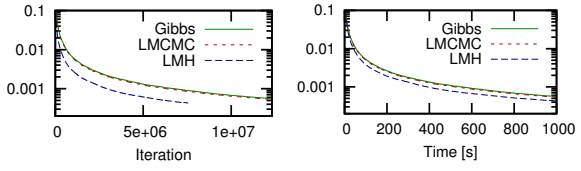
[1] Solutions to this problem exist if the soft evidence is on a single unary relation (Bui, Huynh, and de Salvo Braz 2012)



(a) Texas - Iterations (b) Texas - Time

(c) Washington - Iterations (d) Washington - Time

Figure 3: WebKB - KL Divergence of Texas and Washington



(a) Cornell - Iterations (b) Cornell - Time

(c) Wisconsin - Iterations (d) Wisconsin - Time

Figure 4: WebKB - KL Divergence of Cornell and Wisconsin

are no exploitable exact symmetries and Lifted MCMC coincides with Gibbs sampling. Next we construct an OSA using a rank-5 approximation of the link structure (Van den Broeck and Darwiche 2013) and group the potential weights into 6 clusters. From this OSA we construct a set of automorphisms that is efficient for LMH (see Appendix). Figures 3 and 4 show that the LMH chain, with mixing parameter $\alpha = 4/5$, has a lower KL divergence than Gibbs and Lifted MCMC vs. the number of iterations. Note that there is a slight overhead to LMH because the orbital Metropolis chain is run between base chain steps. Despite this overhead, LMH outperforms the baselines as a function of time. The orbital Metropolis chain accepts approximately 70% of its proposals.

Figure 5 illustrates the effect of running Lifted MCMC on OSA, which is the current state-of-the-art approach for asymmetric models. As expected, the drawn sample points produce biased estimates. As the quality of the approximation increases, the bias reduces, but so do the speedups. LMH does not suffer from a bias. Moreover, we observe that its performance is stable across different OSAs (not depicted).

We also ran experiments for two propositional models

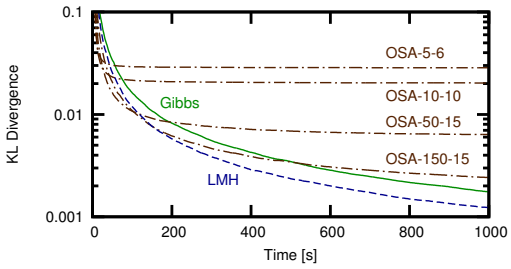
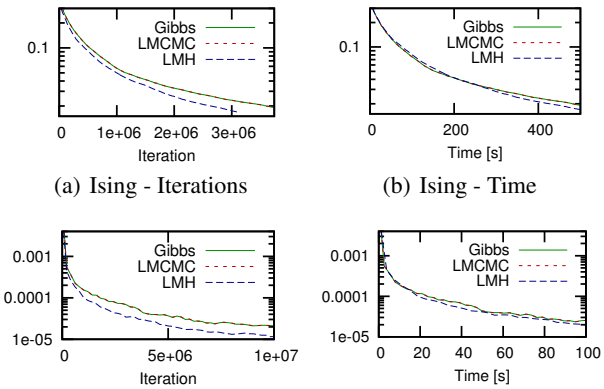


Figure 5: LMH vs. over-symmetric approximations (OSA) on WebKB Washington. OSA-r-c denotes binary evidence of Boolean rank r and c clusters of formula weights.



(a) Ising - Iterations (b) Ising - Time

(c) Chimera - Iterations (d) Chimera - Time

Figure 6: KL Divergences for the propositional models.

that are frequently used in real world applications. The first model is a 100x100 ferromagnetic Ising model with constant interaction strength and external field (see Figure 1(a) for a 4x4 version). Due to the different potentials induced by the external field, the model has no symmetries. We use the model without external field to compute the approximate symmetries. The automorphism group representing these symmetries is generated by the rotational and reflectional symmetries of the grid model (see Figure 1(b)). As in the experiments with the relational models, we used the mixing parameter $\alpha = 4/5$ for the LMH algorithm. Figure 6(c) and (d) depicts the plots of the experimental results. The LMH algorithm performs better with respect to the number of iterations and, to a lesser extent, with respect to time.

We also ran experiments on the Chimera model which has recently received some attention as it was used to assess the performance of quantum annealing (Boixo et al. 2013). We used exactly the model as described in Boixo et al. (2013). This model is also asymmetric but can be made symmetric by assuming that all pairwise interactions are identical. The KL divergence vs. number of iterations and vs. time in seconds is plotted in Figure 6(a) and (b), respectively. Similar to the results for the Ising model, LMH outperforms Gibbs and LMCMC both with respect to the number of iterations and wall clock time. In summary, the LMH algorithm outperforms standard sampling algorithms on these propo-

sitional models in the absence of any symmetries. We used very simple symmetrization strategies for the experiments. This demonstrates that the LMH framework is powerful and allows one to design state-of-the-art sampling algorithms.

Conclusions

We have presented a Lifted Metropolis-Hastings algorithms capable of mixing two types of Markov chains. The first is a non-lifted base chain, and the second is an orbital Metropolis chain that moves between approximately symmetric states. This allows lifted inference techniques to be applied to asymmetric graphical models.

Acknowledgments This work was supported by the Research Foundation-Flanders (FWO-Vlaanderen).

References

Andrieu, C.; de Freitas, N.; Doucet, A.; and Jordan, M. I. 2003. An introduction to MCMC for machine learning. *Machine Learning* 50(1-2):5–43.

Boixo, S.; Rønnow, T. F.; Isakov, S. V.; Wang, Z.; Wecker, D.; Lidar, D. A.; Martinis, J. M.; and Troyer, M. 2013. Quantum annealing with more than one hundred qubits. *Nature Physics* 10(3):218–224.

Bui, H.; Huynh, T.; and de Salvo Braz, R. 2012. Exact lifted inference with distinct soft evidence on every object. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI)*.

Bui, H.; Huynh, T.; and Riedel, S. 2013. Automorphism groups of graphical models and lifted variational inference. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.

Craven, M., and Slattery, S. 2001. Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning Journal* 43(1/2):97–119.

Darwiche, A. 2009. *Modeling and reasoning with Bayesian networks*. Cambridge University Press.

De Raedt, L.; Frasconi, P.; Kersting, K.; and Muggleton, S., eds. 2008. *Probabilistic inductive logic programming: theory and applications*. Berlin, Heidelberg: Springer-Verlag.

Erdős, P., and Rényi, A. 1963. Asymmetric graphs. *Acta Mathematica Hungarica* 14(3):295–315.

Getoor, L., and Taskar, B. 2007. *Introduction to Statistical Relational Learning*. The MIT Press.

Gogate, V.; Jha, A. K.; and Venugopal, D. 2012. Advances in lifted importance sampling. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI)*.

Gribkoff, E.; Van den Broeck, G.; and Suciu, D. 2014. Understanding the complexity of lifted inference and asymmetric weighted model counting. In *Proceedings of UAI*.

Häggström, O. 2002. *Finite Markov chains and algorithmic applications*. London Mathematical Society student texts. Cambridge University Press.

Jaeger, M., and Van den Broeck, G. 2012. Liftability of probabilistic inference: Upper and lower bounds. In *Proceedings of the 2nd International Workshop on Statistical Relational AI*.

Kersting, K.; Ahmadi, B.; and Natarajan, S. 2009. Counting belief propagation. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 277–284.

Kersting, K.; Mladenov, M.; Garnett, R.; and Grohe, M. 2014. Power iterated color refinement. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1904–1910.

Kersting, K. 2012. Lifted probabilistic inference. In *Proceedings of European Conference on Artificial Intelligence (ECAI)*.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models*. The MIT Press.

Mladenov, M., and Kersting, K. 2013. Lifted inference via k-locality. In *Proceedings of the 3rd International Workshop on Statistical Relational AI*.

Niepert, M., and Van den Broeck, G. 2014. Tractability through exchangeability: A new perspective on efficient probabilistic inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2467–2475.

Niepert, M. 2012a. Lifted probabilistic inference: An mcmc perspective. In *Proceedings of the 2nd International Workshop on Statistical Relational AI (StaRAI)*.

Niepert, M. 2012b. Markov chains on orbits of permutation groups. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Niepert, M. 2013. Symmetry-aware marginal density estimation. In *Proceedings of the 27th Conference on Artificial Intelligence (AAAI)*.

Noessner, J.; Niepert, M.; and Stuckenschmidt, H. 2013. RockIt: Exploiting Parallelism and Symmetry for MAP Inference in Statistical Relational Models. In *Proceedings of the 27th Conference on Artificial Intelligence (AAAI)*.

Poole, D. 2003. First-order probabilistic inference. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 985–991.

Poon, H., and Domingos, P. 2006. Sound and efficient inference with probabilistic and deterministic dependencies. In *Proceedings of the 21st Conference on Artificial Intelligence (AAAI)*, 458–463.

Richardson, M., and Domingos, P. 2006. Markov logic networks. *Machine learning* 62(1-2):107–136.

Singla, P.; Nath, A.; and Domingos, P. 2014. Approximate lifting techniques for belief propagation. In *Proceedings of AAAI*.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics* 22(4):1701–1728.

Van den Broeck, G., and Darwiche, A. 2013. On the complexity and approximation of binary evidence in lifted inference. In *NIPS*, 2868–2876.

Van den Broeck, G. 2011. On the completeness of first-order knowledge compilation for lifted probabilistic inference. In *Advances in Neural Information Processing Systems 24 (NIPS),*, 1386–1394.

Van den Broeck, G. 2013. *Lifted Inference and Learning in Statistical Relational Models*. Ph.D. Dissertation, KU Leuven.

Venugopal, D., and Gogate, V. 2012. On lifting the gibbs sampling algorithm. In *Advances in Neural Information Processing Systems*, 1664–1672.

Venugopal, D., and Gogate, V. 2014. Evidence-based clustering for scalable inference in markov logic. In *Proceedings of the Conference on Machine Learning and Knowledge Discovery in Databases*, 258–273.

Markov Logic Networks

We first introduce some standard concepts from first-order logic. An *atom* $P(t_1, \ldots, t_n)$ consists of a predicate P/n of arity n followed by n argument terms t_i, which are either *constants*, $\{\mathsf{A}, \mathsf{B}, \ldots\}$ or *logical variables* $\{\mathsf{x}, \mathsf{y}, \ldots\}$. A formula combines atoms with connectives (e.g., $\wedge$, $\Leftrightarrow$). A formula is *ground* if it contains no logical variables. The groundings of a formula are obtained by instantiating the variables with particular constants.

Many statistical relational languages have been proposed in recent years (Getoor and Taskar 2007; De Raedt et al. 2008). One such language is *Markov logic networks* (MLN) (Richardson and Domingos 2006). An MLN is a set of tuples (w, f), where w is a real number representing a weight and f is a formula in first-order logic. Consider for example the MLN

$$1.3 \quad \texttt{Page}(\mathsf{x}, \mathsf{Faculty}) \Rightarrow \texttt{HasWord}(\mathsf{x}, \mathsf{Hours})$$
$$1.5 \quad \texttt{Page}(\mathsf{x}, \mathsf{Faculty}) \wedge \texttt{Link}(\mathsf{x}, \mathsf{y}) \Rightarrow \texttt{Page}(\mathsf{y}, \mathsf{Course})$$

which states that faculty web pages are more likely to contain the word "hours", and that faculty pages are more likely to link to course pages.

Given a domain of constants $\mathbf{D}$, a first-order MLN Δ induces a *grounding*, which is the MLN obtained by replacing each formula in Δ with all its groundings. For the domain $\mathbf{D} = \{\mathsf{A}, \mathsf{B}\}$ (i.e., two pages), the MLN represents the following grounding.

$$1.3 \quad \texttt{Page}(\mathsf{A}, \mathsf{Faculty}) \Rightarrow \texttt{HasWord}(\mathsf{A}, \mathsf{Hours})$$
$$1.3 \quad \texttt{Page}(\mathsf{B}, \mathsf{Faculty}) \Rightarrow \texttt{HasWord}(\mathsf{B}, \mathsf{Hours})$$
$$1.5 \quad \texttt{Page}(\mathsf{A}, \mathsf{Faculty}) \wedge \texttt{Link}(\mathsf{A}, \mathsf{B}) \Rightarrow \texttt{Page}(\mathsf{B}, \mathsf{Course})$$
$$1.5 \quad \texttt{Page}(\mathsf{B}, \mathsf{Faculty}) \wedge \texttt{Link}(\mathsf{B}, \mathsf{A}) \Rightarrow \texttt{Page}(\mathsf{A}, \mathsf{Course})$$
$$1.5 \quad \texttt{Page}(\mathsf{A}, \mathsf{Faculty}) \wedge \texttt{Link}(\mathsf{A}, \mathsf{A}) \Rightarrow \texttt{Page}(\mathsf{A}, \mathsf{Course})$$
$$1.5 \quad \texttt{Page}(\mathsf{B}, \mathsf{Faculty}) \wedge \texttt{Link}(\mathsf{B}, \mathsf{B}) \Rightarrow \texttt{Page}(\mathsf{B}, \mathsf{Course})$$

This grounding has ten random variables, yielding a distribution over 2^{10} possible worlds. The weight of each world is the product of the expressions $\exp(w)$, where (w, f) is a ground MLN formula and f is satisfied by the world.

Approximate Automorphism Heuristic

Given an OSA, we construct a set of approximate automorphisms as follows. First, we compute the exact automorphisms $\mathfrak{G}_1$ of the OSA. Second, we compute the variable orbits of $\mathfrak{G}_1$, grouping together all variables that can be mapped into each other. Then, for every orbit O, we construct a set of automorphisms as follows. We greedily search for a $O' \subseteq O$ such that the symmetric group $\mathfrak{G}_{O'}$ on O' maximizes the ratio between the number of moved variables (i.e., $|O'|$) and the number of moved potentials, while keeping the number of moved potentials bounded by a constant K. This guarantees that $\mathfrak{G}_{O'}$ yields an efficient orbital Metropolis chain. Finally, we remove O' from O and recurse until O is empty. From this set of symmetric groups $\mathfrak{G}_{O'}$, we construct a set of orbital Metropolis chains, each with it own set of moved potentials.