Symmetry-Aware Marginal Density Estimation

Mathias Niepert
Computer Science & Engineering
University of Washington
Seattle, WA 98195-2350, USA

Abstract

The Rao-Blackwell theorem is utilized to analyze and improve the scalability of inference in large probabilistic models that exhibit symmetries. A novel marginal density estimator is introduced and shown both analytically and empirically to outperform standard estimators by several orders of magnitude. The developed theory and algorithms apply to a broad class of probabilistic models including statistical relational models considered not susceptible to lifted probabilistic inference.

Introduction

Many successful applications of artificial intelligence research are based on large probabilistic models. Examples include Markov logic networks (Richardson and Domingos 2006), conditional random fields (Lafferty, McCallum, and Pereira 2001) and, more recently, deep learning architectures (Hinton, Osindero, and Teh 2006; Bengio and LeCun 2007; Poon and Domingos 2011). Especially the models one encounters in the statistical relational learning (SRL) literature often have joint distributions spanning millions of variables and features. Indeed, these models are so large that, at first sight, inference and learning seem daunting. For numerous of these models, however, scalable approximate and, to a lesser extend, exact inference algorithms do exist. Most notably, there has been a strong focus on lifted inference algorithms, that is, algorithms that group indistinguishable variables and features during inference. For an overview we refer the reader to (Kersting 2012). Lifted algorithms facilitate efficient inference in numerous large probabilistic models for which inference is NP-hard in principle.

We are concerned with the estimation of marginal probabilities based on a finite number of sample points. We show that the feasibility of inference and learning in large and highly symmetric probabilistic models can be explained with the Rao-Blackwell theorem from the field of statistics. The theory and algorithms do not directly depend on the syntactical nature of the relational models such as arity of predicates and number of variables per formula but only on the given automorphism group of the probabilistic model, and are applicable to classes of probabilistic models much broader than the class of statistical relational models.

Consider an experiment where a coin is flipped n times. While a frequentist would assume the flips to be i.i.d., a Bayesian typically makes the weaker assumption of *exchangeability* – that the probability of an outcome sequence only depends on the number of "heads" in the sequence and not on their order. Under the non-i.i.d. assumption, a possible corresponding graphical model is the fully connected graph with n nodes and high treewidth. The actual number of parameters required to specify the distribution, however, is only $n+1$, one for each sequence with $0 \leq k \leq n$ "heads." Bruno de Finetti was the first to realize that such a sequence of random variables can be (re-)parameterized as a unique mixture of $n+1$ independent urn processes (de Finetti 1938). It is this notion of a parameterization as a mixture of urn processes that is at the heart of our work. A direct application of de Finetti's results, however, is often impossible since not all variables are exchangeable in realistic probabilistic models.

Motivated by the intuition of exchangeability, we show that *arbitrary* model symmetries allow us to re-paramterize the distribution as a mixture of independent urn processes where each urn consists of isomorphic joint assignments. Most importantly, we develop a novel Rao-Blackwellized estimator that implicitly estimates the fewer parameters of the simpler mixture model and, based on these, computes the marginal densities. We identify situations in which the application of the Rao-Blackwell estimator is tractable. In particular, we show that the Rao-Blackwell estimator is always linear-time computable for single-variable marginal density estimation. By invoking the Rao-Blackwell theorem, we show that the mean squared error of the novel estimator is at least as small as that of the standard estimator and strictly smaller under non-trivial symmetries of the probabilistic model. Moreover, we prove that for estimates based on sample points drawn from a Markov chain $\mathcal{M}$, the bias of the Rao-Blackwell estimator is governed by the mixing time of the quotient Markov chain whose convergence behavior is superior to that of $\mathcal{M}$.

We present empirical results verifying that the Rao-Blackwell estimator always outperforms the standard estimator by up to several orders of magnitude, irrespective of the model structure. Indeed, we show that the results of the novel estimator resemble those typically observed in lifted inference papers. For the first time such a performance is shown for an SRL model with a transitivity formula.

Background

We review some concepts from group and estimation theory.

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. For a permutation group $\mathfrak{G}$ acting on Ω and a sequence $\mathbf{A} = \langle \alpha_1, ..., \alpha_k \rangle \in \Omega^k$ we write $\mathbf{A}^g$ to denote the image $\langle \alpha_1{}^g, ..., \alpha_k{}^g \rangle$ of $\mathbf{A}$ under $\mathfrak{g}$. Moreover, we write $\mathbf{A}^{\mathfrak{G}}$ to denote the orbit of the sequence $\mathbf{A}$.

Point Estimation Let $s_1, ..., s_N$ be N sample points drawn from some distribution P. An estimator $\hat{\theta}_N$ of a parameter θ is a function of $s_1, ..., s_N$. The bias of an estimator is defined by $\texttt{bias}(\hat{\theta}_N) := \mathbb{E}[\hat{\theta}_N - \theta]$ and the variance by $\mathrm{Var}(\hat{\theta}_N) := \mathbb{E}[(\hat{\theta}_N - \mathbb{E}(\hat{\theta}_N))^2]$, where $\mathbb{E}$ is the expectation with respect to P, the distribution that generated the data. We say that $\hat{\theta}_N$ is unbiased if $\texttt{bias}(\hat{\theta}_N) = 0$. The quality of an estimator is often assessed with the mean squared error (MSE) defined by $\mathrm{MSE}[\hat{\theta}_N] := \mathbb{E}[(\hat{\theta}_N - \theta)^2] = \mathrm{Var}(\hat{\theta}_N) + \texttt{bias}(\hat{\theta}_N)^2$.

Theorem 1 (Rao-Blackwell). *Let $\hat{\theta}$ be an estimator with $\mathbb{E}[\hat{\theta}^2] < \infty$ and T a sufficient statistic both for θ, and let $\hat{\theta}^* := \mathbb{E}[\hat{\theta} \mid T]$. Then, $\mathrm{MSE}[\hat{\theta}^*] \leq \mathrm{MSE}[\hat{\theta}]$. Moreover, $\mathrm{MSE}[\hat{\theta}^*] < \mathrm{MSE}[\hat{\theta}]$ unless $\hat{\theta}^*$ is a function of $\hat{\theta}$.*

Finite Markov chains A finite Markov chain $\mathcal{M}$ defines a random walk on elements of a finite set Ω. For all $x, y \in \Omega$, $Q(x, y)$ is the chain's probability to transition from x to y, and $Q^t(x, y) = Q_x^t(y)$ the probability of being in state y after t steps if the chain starts at x. A Markov chain is irreducible if for all $x, y \in \Omega$ there exists a t such that $Q^t(x, y) > 0$ and aperiodic if for all $x \in \Omega$, $\gcd\{t \geq 1 \mid Q^t(x, x) > 0\} = 1$. An irreducible and aperiodic chain converges to its unique stationary distribution and is called ergodic.

The total variation distance d_{tv} of the Markov chain from its stationary distribution π at time t with initial state x is defined by

$$d_{\mathsf{tv}}(Q_x^t, \pi) = \frac{1}{2} \sum_{y \in \Omega} |Q^t(x, y) - \pi(y)|.$$

For $\varepsilon > 0$, let $\tau_x(\varepsilon)$ denote the least value T such that $d_{\mathsf{tv}}(Q_x^t, \pi) \leq \varepsilon$ for all $t \geq T$. The *mixing time* $\tau(\varepsilon)$ is defined by $\tau(\varepsilon) = \max\{\tau_x(\varepsilon) \mid x \in \Omega\}$.

Related Work

There are numerous lifted inference algorithms such as lifted variable elimination (Poole 2003), lifted belief propagation (Singla and Domingos 2008; Kersting, Ahmadi, and Natarajan 2009), first-order knowledge compilation (Van den Broeck 2011), and lifted variational inference (Choi and Amir 2012). Probabilistic theorem proving applied to a clustering of the relational model was used to lift the Gibbs sampler (Venugopal and Gogate 2012). Recent work exploits automorphism groups of probabilistic models for more efficient probabilistic inference (Bui, Huynh, and Riedel 2012; Niepert 2012). Orbital Markov chains (Niepert 2012) are a class of Markov chains that implicitly operate on the orbit partition of the assignment space and do not invoke the Rao-Blackwell theorem.

Rao-Blackwellized (RB) estimators have been used for inference in Bayesian networks (Doucet et al. 2000; Bidyuk and Dechter 2007) and latent Dirichlet allocation (Teh, Newman, and Welling 2006) with application in robotics (Stachniss, Grisetti, and Burgard 2005) and activity recognition (Bui, Venkatesh, and West 2002). The RB theorem and estimator are important concepts in statistics (Gelfand and Smith 1990; Casella and Robert 1996).

Symmetry-Aware Point Estimation

An automorphism group of a probabilistic model is a group whose elements are permutations of the probabilistic model's random variables $\mathbf{X}$ that leave the joint distribution $P(\mathbf{X})$ invariant. There is a growing interest in computing and utilizing automorphism groups of probabilistic models for more efficient inference algorithms (Bui, Huynh, and Riedel 2012; Niepert 2012). The line of research is primarily motivated by the highly symmetric nature of statistical relational models and provides a complementary view on lifted probabilistic inference. Here, we will not be concerned with deriving automorphism groups of probabilistic models but with developing algorithms that utilize these permutation groups for efficient marginal density estimation. Hence, we always assume a given automorphism group $\mathfrak{G}$ of the probabilistic model under consideration.

We begin by deriving a re-parameterization of the joint distribution in the presence of symmetries that generalizes the mixture of independent urn processes parameterization for finitely exchangeable variables (Diaconis and Freedman 1980). All random variables are assumed to be discrete.

Let $\mathbf{X} = \langle X_1, ..., X_n \rangle$ be a finite sequence of discrete random variables with joint distribution $P(\mathbf{X})$, let $\mathfrak{G}$ be an automorphism group of $\mathbf{X}$, and let $\mathcal{O}$ be an orbit partition of the assignment space induced by $\mathfrak{G}$. Please note that for any $\mathbf{x}, \mathbf{x}' \in O \in \mathcal{O}$ we have $P(\mathbf{x}) = P(\mathbf{x}')$. For a subsequence $\hat{\mathbf{X}}$ of $\mathbf{X}$ and an orbit $O \in \mathcal{O}$ we write $P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid O)$ for the marginal density $P(\hat{\mathbf{X}} = \hat{\mathbf{x}})$ conditioned on O. Thus,

$$P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid O) = \frac{1}{|O|} \sum_{\mathbf{x} \in O} \mathbb{I}_{\{\mathbf{x}\langle\hat{\mathbf{X}}\rangle = \hat{\mathbf{x}}\}},$$
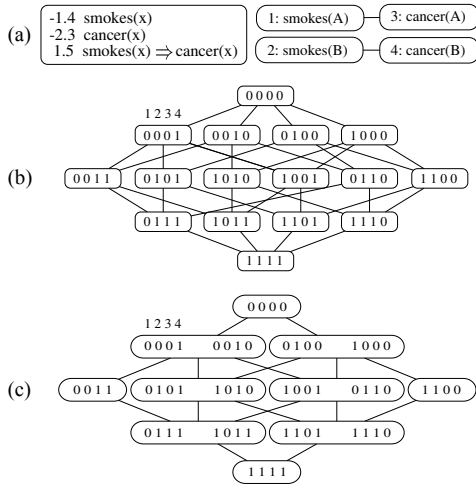
Figure 1: Illustration of the orbit partition of the assignment space induced by the renaming automorphism group $\{(A\ B), ()\}$. A renaming automorphism is a permutation of constants that forms an isomorphism between two graphical models. (a) An MLN with three formulas and the grounding for two constants A and B; (b) the state space of a Gibbs chain with non-zero transitions indicated by lines and without self-arcs; (c) the lumped state space of the quotient Markov chain which has 10 instead of 16 states. The joint distribution can be expressed as a mixture of draws from the orbits.

where $\mathbb{I}$ is the indicator function and $\mathbf{x}\langle\hat{\mathbf{X}}\rangle$ the assignment within $\mathbf{x}$ to the variables in the sequence $\hat{\mathbf{X}}$. We can now (re-)parameterize the marginal density as a mixture of independent orbit distributions

$$P(\hat{\mathbf{X}} = \hat{\mathbf{x}}) = \sum_{O \in \mathcal{O}} P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid O)P(O),$$

where $P(O) = \sum_{\mathbf{x} \in O} P(\mathbf{X} = \mathbf{x})$. For instance, the joint distribution of the Markov logic network in Figure 1(a) can be parameterized as a mixture of the distributions for the 10 orbits depicted in Figure 1(c).

Let us first recall the standard estimator used in most sampling approaches. After collecting N sample points $s_1, ..., s_N$ the standard estimator for the marginal density $\theta := P(\hat{\mathbf{X}} = \hat{\mathbf{x}})$ is defined as

$$\hat{\theta}_N := \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{s_i\langle\hat{\mathbf{X}}\rangle = \hat{\mathbf{x}}\}}. \qquad (1)$$

Now, the *symmetry-aware Rao-Blackwell* estimator for N sample points $s_1, ..., s_N$ is defined as

$$\hat{\theta}_N^{\mathtt{rb}} := \frac{1}{N} \sum_{i=1}^{N} P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid s_i^{\mathfrak{G}}), \qquad (2)$$

where $\mathfrak{G}$ is the given automorphism group that induces $\mathcal{O}$. Hence, the *unbiased* Rao-Blackwell estimator integrates out the joint assignments of each orbit. We will prove that the mean squared error of the Rao-Blackwell estimator is less

than or equal to that of the standard estimator. First, however, we want to investigate under what conditions we can efficiently compute the conditional density of equation (2). To this end, we establish a connection between the orbit of the subsequence $\hat{\mathbf{X}}$ under the automorphism group $\mathfrak{G}$ and the orbit partition of the assignment space induced by $\mathfrak{G}$[1].

Definition 2. *Let $\mathbf{X}$ be a finite sequence of random variables with joint distribution $P(\mathbf{X})$, let $\mathfrak{G}$ be an automorphism group of $\mathbf{X}$, let $\hat{\mathbf{X}}$ be a subsequence of $\mathbf{X}$, let $\mathsf{Val}(\mathbf{X})$ be the assignment space of $\mathbf{X}$, and let $s \in \mathsf{Val}(\mathbf{X})$. The* orbit Hamming weight *of s with respect to the marginal assignment $\hat{\mathbf{X}} = \hat{\mathbf{x}}$ is defined by*

$$\mathsf{H}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}^{\mathfrak{G}}(s) := \sum_{\mathbf{A} \in \hat{\mathbf{X}}^{\mathfrak{G}}} \mathbb{I}_{\{s\langle\mathbf{A}\rangle = \hat{\mathbf{x}}\}}.$$

Based on this definition, we state a lemma which allows us to compute the density of equation (2) in closed form, without having to enumerate all of the orbit's elements.

Lemma 3. *Let $\mathbf{X}$ be a finite sequence of random variables with joint distribution $P(\mathbf{X})$, let $\mathfrak{G}$ be an automorphism group of $\mathbf{X}$, let $\hat{\mathbf{X}}$ be a subsequence of $\mathbf{X}$, and let $s \in \mathsf{Val}(\mathbf{X})$. Then,*

$$P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid s^{\mathfrak{G}}) = \frac{\mathsf{H}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}^{\mathfrak{G}}(s)}{|\hat{\mathbf{X}}^{\mathfrak{G}}|} = \mathbb{E}[\hat{\theta}_N \mid \mathsf{H}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}^{\mathfrak{G}}(s)].$$

The following example demonstrates the application of the lemma to the special case of single-variable marginal density estimation for the MLN in Figure 1.

Example 4. *Let us assume we want to estimate the marginal density $P(smokes(A)=1)$ of the MLN in Figure 1(a). Since $\mathfrak{G} = \{(smokes(A)\ smokes(B))(cancer(A)\ cancer(B)), ()\}$ we have that $\langle smokes(A)\rangle^{\mathfrak{G}}=\{\langle smokes(A)\rangle, \langle smokes(B)\rangle\}$. Given the sample point $s = \langle 1, 0, 1, 0\rangle$ we have that $\mathsf{H}_{\langle smokes(A)\rangle=\langle 1\rangle}^{\mathfrak{G}}(s) = 1$ and $P(smokes(A)=1 \mid s^{\mathfrak{G}}) = \frac{1}{2}$.*

Thus, given a sample point s, the marginal density conditioned on an orbit of the assignment space is computable in closed form using the *orbit Hamming weight* of s with respect to the marginal assignment since it is a *sufficient statistic* for the marginal density. If the probabilistic model exhibits symmetries, then the Rao-Blackwell estimator's MSE is less than or equal to that of the standard estimator.

Theorem 5. *Let $\mathbf{X}$ be a finite sequence of random variables with joint distribution $P(\mathbf{X})$, let $\mathfrak{G}$ be an automorphism group of $\mathbf{X}$ given by a generating set R, let $\hat{\mathbf{X}}$ be a subsequence of $\mathbf{X}$, and let $\theta := P(\hat{\mathbf{X}} = \hat{\mathbf{x}})$ be the marginal density to be estimated. The Rao-Blackwell estimator $\hat{\theta}_N^{\mathtt{rb}}$ has the following properties:*

(a) Its worst-case time complexity is $O(R|\hat{\mathbf{X}}^{\mathfrak{G}}|+N|\hat{\mathbf{X}}^{\mathfrak{G}}|)$;

(b) $MSE[\hat{\theta}_N^{\mathtt{rb}}] \leq MSE[\hat{\theta}_N]$.

The inequality of (b) is strict if there exists a joint assignment s with non-zero density and $0 < \mathsf{H}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}^{\mathfrak{G}}(s) < |\hat{\mathbf{X}}^{\mathfrak{G}}| > 1$.

---

[1] Please note the *two different types* of orbit partitions discussed here. One results from $\mathfrak{G}$ acting on the assignment space the other from $\mathfrak{G}$ acting on sequences of random variables.

For single-variable density estimation the worst-case time complexity of the Rao-Blackwell estimator is $O(R|\mathbf{X}| + N|\mathbf{X}|)$ and, therefore, linear both in the number of variables and the number of sample points. For most symmetric models, the inequality of Theorem 5(b) is strict and the Rao-Blackwell estimator outperforms the standard estimator, a behavior we will verify empirically.

Please note that in the special case of single-variable marginal density estimation the RB estimator is identical to the estimator that averages the identically distributed variables located in the same orbit. The advantages of utilizing the Rao-Blackwell theory are (1) it directly provides conditions for which the inequality of Theorem 5(b) is strict; (2) it generalizes the single-variable case to marginals spanning multiple variables; (3) it allows us to investigate the *completeness* of an estimator with respect to a given automorphism group; and (4) it provides the link to the quotient Markov chain in the MCMC setting and its superior convergence behavior presented in the following section.

The Rao-Blackwell estimator is unbiased if the drawn sample points are independent. Since it is often only practical to collect sample points from a Markov chain, the bias for a finite number of N points will depend on the chain's mixing behavior. We will show that if there are non-trivial model symmetries and if we are using the Rao-Blackwell estimator, we only need to worry about the mixing behavior of the Markov chain whose state space is the orbit partition.

Symmetry-Aware MCMC

Whenever we collect sample points from a Markov chain, the efficiency of an estimator is influenced by (a) the mixing behavior of the Markov chain and (b) the variance of the estimator under the assumption that the Markov chain has reached stationarity, that is, the asymptotic variance (Neal 2004). That the Rao-Blackwell estimator's asymptotic variance is at least as low as that of the standard estimator is a corollary of Theorem 5. We show that the same is true for the bias that is caused by the fact that we collect a finite number of sample points from Markov chains which never *exactly* reach stationarity.

A lumping of a Markov chain is a partition of its state space which is possible under certain conditions on the transition probabilities of the original Markov chain (Buchholz 1994; Derisavi, Hermanns, and Sanders 2003).

Definition 6. *Let $\mathcal{M}$ be an ergodic Markov chain with transition matrix Q, stationary distribution π, and state space Ω, and let $\mathcal{C} = \{C_1, ..., C_n\}$ be a partition of the state space. If for all $C_i, C_j \in \mathcal{C}$ and all $s_i', s_i'' \in C_i$*

$$Q'(C_i, C_j) := \sum_{s_j \in C_j} Q(s_i', s_j) = \sum_{s_j \in C_j} Q(s_i'', s_j)$$

then we say that $\mathcal{M}$ is ordinary lumpable with respect to $\mathcal{C}$. If, in addition, $\pi(s_i') = \pi(s_i'')$ for all $s_i', s_i'' \in C_i$ and all $C_i \in \mathcal{C}$ then $\mathcal{M}$ is exactly lumpable with respect to $\mathcal{C}$. The Markov chain $\mathcal{M}'$ with state space $\mathcal{C}$ and transition matrix Q' is called the quotient chain of $\mathcal{M}$ with respect to $\mathcal{C}$.

Every finite ergodic Markov chain is exactly lumpable with respect to an orbit partition of its state space. The fol-lowing theorem states this and the convergence behavior of the quotient Markov chain in relation to the original Markov chain (cf. (Boyd et al. 2005)).

Proposition 7. *Let $\mathcal{M}$ be an ergodic Markov chain and let $\mathcal{O}$ be an orbit partition of its state space. Then, the Markov chain $\mathcal{M}$ is exactly lumpable with respect to $\mathcal{O}$. If $\mathcal{M}$ is reversible, then the quotient Markov chain $\mathcal{M}'$ with respect to $\mathcal{O}$ is also reversible. Moreover, the mixing time of $\mathcal{M}'$ is smaller than or equal to the mixing time of $\mathcal{M}$.*

Example 8. *Figure 1(b) depicts the state space of the Gibbs chain for the MLN shown in Figure 1(a). The constants renaming automorphism group $\{(A\ B), ()\}$ acting on the sets of constants leads to the automorphism group $\{(smokes(A)\ smokes(B))(cancer(A)\ cancer(B)), ()\}$ on the ground level. This permutation group acting on the state space of the Gibbs chain induces an orbit partition which is the state space of the quotient Markov chain (see Figure 1(c)).*

The explicit construction of the state space of a quotient Markov chain is intractable. Given an automorphism group $\mathfrak{G}$, merely counting the number of equivalence classes of the orbit partition of the assignment space induced by $\mathfrak{G}$ is known to be a #P-complete problem (Goldberg 2001). Nevertheless, if the Rao-Blackwell estimator is utilized, one can draw the sample points from the original Markov chain while analyzing the convergence behavior of the quotient Markov chain of the original chain.

Theorem 9. *Let $\mathbf{X}$ be a finite sequence of random variables with joint distribution $P(\mathbf{X})$, let $\mathcal{M}$ be an ergodic Markov chain with stationary distribution P, and let $\mathcal{O}$ be an orbit partition of $\mathcal{M}$'s state space. Let $\hat{\theta}_N^{\mathtt{rb}}$ be the Rao-Blackwell estimator for N sample points $s_{T+1}, ..., s_{T+N}$ collected from $\mathcal{M}$, after discarding the first T sample points. Then, $|\mathtt{bias}(\hat{\theta}_N^{\mathtt{rb}})| \leq \epsilon$ if $T \geq \tau'(\epsilon)$, where $\tau'(\epsilon)$ is the mixing time of the quotient Markov chain of $\mathcal{M}$ with respect to $\mathcal{O}$.*

Hence, if one wants to make sure that the absolute value of the bias of the Rao-Blackwell estimator is smaller than a given $\epsilon > 0$, one only needs a burn-in period consisting of $\tau'(\epsilon)$ simulation steps, where $\tau'(\epsilon)$ is the mixing time of the quotient Markov chain. Existing work on analyzing the influence of symmetries in random walks has shown that it is often more convenient to investigate the mixing behavior of the quotient Markov chain (Boyd et al. 2005). In the context of marginal density estimation, Markov chains implicitly operating on the orbit partition of the assignment space were shown to have better mixing behavior (Niepert 2012).

In summary, whenever probabilistic models exhibit non-trivial symmetries we can have the best of both worlds. The bias owed to the fact that we are collecting a finite number of sample points from a Markov chain as well as the asymptotic variance (Neal 2004) of the Rao-Blackwell estimator are at least as small as those of the standard estimator. The more symmetric the probabilistic model the larger the reduction in mean squared error.
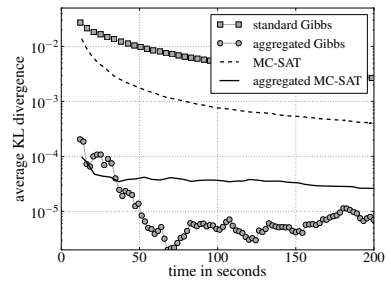
We now present the experimental results for several large probabilistic models, both relational and non-relational.
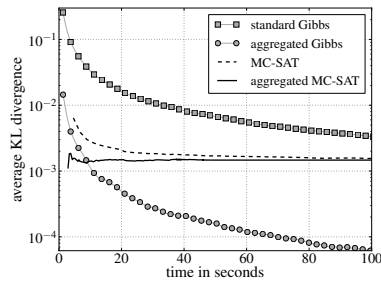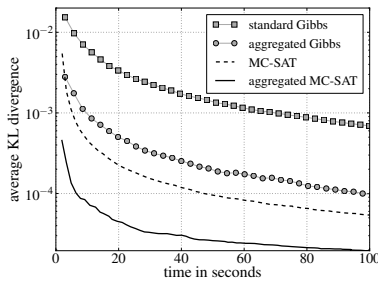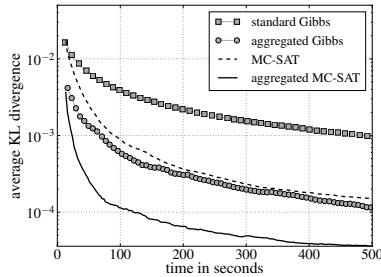
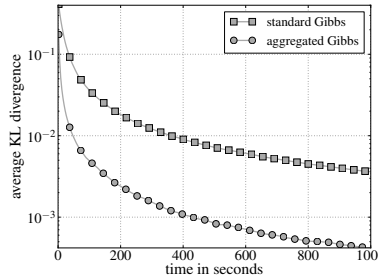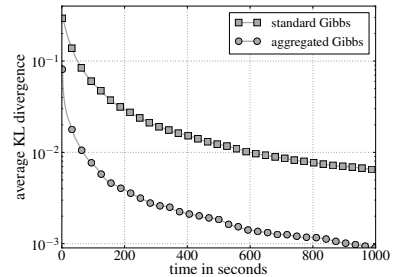|		
(a) The asthma-smokes-cancer MLN with 50 people and 10% evidence.	(b) The smokes-cancer MLN with 50 people and 10% evidence.	(c) The smokes-cancer MLN with 50 people, no evidence, and transitivity.
(d) The smokes-cancer MLN with 50 people, 10% evidence, and transitivity.	(e) The 2-coloring 100×100 grid model with weight 0.2.	(f) The 2-coloring 100×100 grid model with hard constraints.

Figure 2: Plots of average KL divergence versus time in seconds of the two MCMC algorithms with the standard estimator (standard) and the Rao-Blackwell estimator (aggregated) for various probabilistic models.

Experiments

The aim of the empirical investigation is twofold. First, we want to verify the efficiency of the novel Rao-Blackwell estimator when applied as a post-processing step to the output of state-of-the-art sampling algorithms. Second, we want to test the hypothesis that the efficiency gains of the novel estimator on standard SRL models are similar empirically to those of state-of-the-art lifted inference algorithms.

For the SRL models we computed the orbit partitions of the variables based on the model's renaming automorphisms (Bui, Huynh, and Riedel 2012). As discussed earlier, renaming automorphisms are computable in time linear in the domain size. We applied GAP (GAP 2012) to compute the variables' orbit partition. For all non SRL models we computed the automorphism group and the orbit partitions as in (Niepert 2012) using the graph automorphism algorithm SAUCY (Darga, Sakallah, and Markov 2008) and the GAP system, respectively. Overall, the computation of the orbit partitions of the models' variables took less than one second for each of the probabilistic models we considered.

We conducted experiments with several benchmark Markov logic networks, a statistical relational language general enough to capture numerous types of graphical models (Richardson and Domingos 2006). Here, we used (a) the asthma-smokes-cancer MLN (Venugopal and Gogate 2012)

with 10% evidence[2]; (b) the "Friends & Smokers" MLN exactly as specified in (Singla and Domingos 2008) with 10% evidence; and the "Friends & Smokers" MLN with the transitivity formula on the friends relation having weight 1.0, (c) without and (d) with 10% evidence. Each of the models had between 10 and 100 objects in the domain, leading to log-linear models with 10^2-10^4 variables and 10^2-10^6 features. We used WFOMC (Van den Broeck 2011), to compute the exact single-variable marginals of the asthma MLN. For all other MLNs, existing exact lifted inference algorithm were unable to compute single-variable densities. In these cases, we performed several very long runs (burn-in 1 day; overall 5 days) of a Gibbs sampler guaranteed to be ergodic and made sure that state-of-the-art MCMC diagnostics indicated convergence (Brooks and Gelman 1998).

We executed our implementation of the standard Gibbs sampler and ALCHEMY's implementation of the MC-SAT algorithm (Poon and Domingos 2006) on the MLNs based on 10 separate runs, without a burn-in period. For each sampling algorithm we computed the marginal densities with the standard estimator and the Rao-Blackwell estimator, respectively, which we implemented in the GAP programming language[3]. Figure 2 depicts, for each MLN, the av-

---

[2] For a random 10% of all people it is known (a) whether they smoke or not and (b) who 10 of their friends are.

[3] https://code.google.com/p/lifted-mcmc/

(a) Average time for the smokes-cancer MLN.

(b) Number of sample points for the smokes-cancer MLN.

(c) Average time for the smokes-cancer MLN with transitivity.

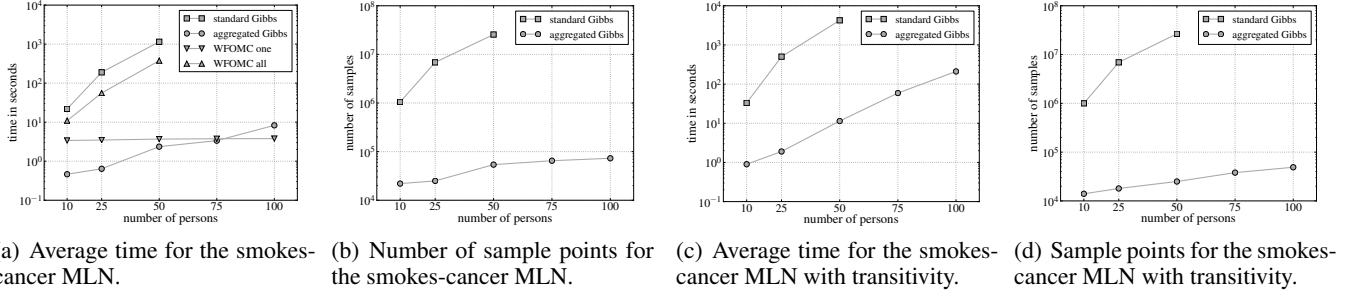(d) Sample points for the smokes-cancer MLN with transitivity.

Figure 3: Average time and number of sample points, respectively, needed to achieve an average KL divergence of < 0.0001.

erage Kullback-Leibler divergence[4] between the estimated and precomputed true single-variable marginals of the non-evidence variables plotted against the absolute running time of the algorithms in seconds.

The Rao-Blackwell estimator improves the density estimates by at least an order of magnitude and, in the absence of evidence, even up to four orders of magnitude relative to the standard estimator. The improvement of the empirical results is independent of the relational structure of the MLNs. For the MLN with a transitivity formula on the `friends` relation, generally considered a problematic and as of now not domain-liftable model, the results are as pronounced as for the MLNs known to be domain-liftable.

We also conducted experiments with non-SRL models to investigate the efficiency of the approach on graphical models. We executed the Gibbs sampler with and without using the Rao-Blackwell estimator on a 100×100 2-coloring grid model with binary random variables. The symmetries of the model are the reflection and rotation automorphisms of the 2-dimensional square grid. Figure 2(e) depicts the plot of the average KL divergence against the running time in seconds, where each pairwise factor between neighboring variables X, Y was defined as $\exp(0.2)$ if $X \neq Y$, and 1 otherwise. Figure 2(f) depicts the plot of the same grid model except that the pairwise factors were defined as 1 if $X \neq Y$, and 0 otherwise. The results clearly demonstrate the superior performance of the Rao-Backwell estimator even for probabilistic models with a smaller number of symmetries.

In addition, we analyzed the impact of the domain size on the estimator performance for (a) domain-liftable MLNs and (b) MLNs not liftable by any state-of-the-art exact lifted inference algorithm. We used the "Friends & Smokers" MLN without evidence; and with and without the transitivity formula on the `friends` relation. The MLN without transitivity is a standard benchmark for lifted algorithms whereas MLNs with transitivity are considered difficult and no exact lifted inference algorithm exists for such MLNs as of now. Figures 3(a)&(c) depict the time needed to achieve an average KL divergence of less than 10^{-4} plotted against the domain size of the models without and with transitivity. The increase in runtime is far less pronounced with the Rao-

Blackwell estimator. The plots resemble those often shown in lifted inference papers where an algorithm that can lift a model is contrasted with one that cannot. The increase in runtime is slightly higher for the model with transitivity but this is owed to the size increase of each variable's Markov blanket and, thus, the time needed for each Gibbs sampler step. Figures 3(b)&(d) plots the sample size required to achieve an average KL divergence of less than 10^{-4} against the domain size. Interestingly, the number of sample points is almost identical for the model with and without transitivity, indicating that the advantage of the Rao-Blackwell estimator is independent of the model's formulas.

In Figure 3(a) we plot the results of WFOMC for compiling a first-order circuit and computing (a) one single-variable marginal and (b) all single-variable marginals. WFOMC has constant runtime for exactly computing one single-variable marginal density. The Rao-Blackwell estimation for all of the model's variables scales sub-linearly and is more efficient than repeated calls to WFOMC. While we do not need to run WFOMC once per single-variable marginal density if the variables are first partitioned into sets of variables with identical marginal densities (de Salvo Braz, Amir, and Roth 2005), the results demonstrate that the symmetry-aware estimator scales comparably to exact lifted inference algorithms on domain-liftable models and that its runtime is polynomial in the domain size of the MLNs.

Discussion

A Rao-Blackwell estimator was developed and shown, both analytically and empirically, to have lower mean squared error under non-trivial model symmetries. The presented theory provides a novel perspective on the notion of lifted inference and the underlying reasons for the feasibility of marginal density estimation in large but highly symmetric probabilistic models. For the first time, the applicability of such an approach does not directly depend on the properties of the relational structure such as the arity of predicates and the type of formulas but only on the given *evidence* and the corresponding automorphism group of the model. We believe the theoretical and empirical insights to be of great interest to the machine learning community and that the presented work might contribute to a deeper understanding of lifted inference algorithms.

---

[4]We computed both MSE and average KL divergence but omitted the qualitatively identical MSE results due to space constraints.

Acknowledgments

References

[Bengio and LeCun 2007] Bengio, Y., and LeCun, Y. 2007. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press.

[Bidyuk and Dechter 2007] Bidyuk, B., and Dechter, R. 2007. Cutset sampling for bayesian networks. *J. Artif. Intell. Res.* 28:1–48.

[Blackwell 1947] Blackwell, D. 1947. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics* 18(1):105–110.

[Boyd et al. 2005] Boyd, S. P.; Diaconis, P.; Parrilo, P. A.; and Xiao, L. 2005. Symmetry analysis of reversible markov chains. *Internet Mathematics* 2(1):31–71.

[Brooks and Gelman 1998] Brooks, S. P., and Gelman, A. 1998. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* 7(4):434–455.

[Buchholz 1994] Buchholz, P. 1994. Exact and ordinary lumpability in finite markov chains. *Journal of Applied Probability* 31(1):59–75.

[Bui, Huynh, and Riedel 2012] Bui, H. H.; Huynh, T. N.; and Riedel, S. 2012. Automorphism groups of graphical models and lifted variational inference. *CoRR* abs/1207.4814.

[Bui, Venkatesh, and West 2002] Bui, H. H.; Venkatesh, S.; and West, G. 2002. Policy recognition in the abstract hidden markov model. *J. Artif. Int. Res.* 17(1):451–499.

[Casella and Robert 1996] Casella, G., and Robert, C. P. 1996. Rao-blackwellisation of sampling schemes. *Biometrika* 83(1):81–94.

[Choi and Amir 2012] Choi, J., and Amir, E. 2012. Lifted relational variational inference. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 196–206.

[Darga, Sakallah, and Markov 2008] Darga, P. T.; Sakallah, K. A.; and Markov, I. L. 2008. Faster symmetry discovery using sparsity of symmetries. In *Proceedings of the 45th annual Design Automation Conference*, 149–154.

[de Finetti 1938] de Finetti, B. 1938. Sur la condition d'équivalence partielle. In *Colloque consacré a la theorie des probabilités*, volume VI. Hermann, Paris. 5–18. English translation in R. Jeffrey (ed.), pp. 193–205.

[de Salvo Braz, Amir, and Roth 2005] de Salvo Braz, R.; Amir, E.; and Roth, D. 2005. Lifted first-order probabilistic inference. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1319–1325.

[Derisavi, Hermanns, and Sanders 2003] Derisavi, S.; Hermanns, H.; and Sanders, W. H. 2003. Optimal state-space lumping in markov chains. *Inf. Process. Lett.* 87(6):309–315.

[Diaconis and Freedman 1980] Diaconis, P., and Freedman, D. 1980. De finetti's generalizations of exchangeability. In *Studies in Inductive Logic and Probability*, volume II.

[Doucet et al. 2000] Doucet, A.; Freitas, N. d.; Murphy, K. P.; and Russell, S. J. 2000. Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 176–183.

[GAP 2012] The GAP Group. 2012. *GAP – Groups, Algorithms, and Programming, Version 4.5.7.*

[Gelfand and Smith 1990] Gelfand, A. E., and Smith, A. F. M. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85(410):398–409.

[Goldberg 2001] Goldberg, L. A. 2001. Computation in permutation groups: counting and randomly sampling orbits. In *Surveys in Combinatorics*. Cambridge University Press. 109–143.

[Hinton, Osindero, and Teh 2006] Hinton, G. E.; Osindero, S.; and Teh, Y. W. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18(7):1527–1554.

[Holt, Eick, and O'Brien 2005] Holt, D. F.; Eick, B.; and O'Brien, E. A. 2005. *Handbook of Computational Group Theory*. Chapman and Hall.

[Kersting, Ahmadi, and Natarajan 2009] 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 2012] Kersting, K. 2012. Lifted probabilistic inference. In *Proceedings of the 20th European Conference on Artificial Intelligence*, 33–38.

[Lafferty, McCallum, and Pereira 2001] Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289.

[Levin, Peres, and Wilmer 2008] Levin, D. A.; Peres, Y.; and Wilmer, E. L. 2008. *Markov Chains and Mixing Times*. American Mathematical Society.

[Neal 2004] Neal, R. M. 2004. Improving asymptotic variance of mcmc estimators: Non-reversible chains are better. Technical Report 0406, University of Toronto.

[Niepert 2012] Niepert, M. 2012. Markov chains on orbits of permutation groups. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 624–633.

[Poole 2003] Poole, D. 2003. First-order probabilistic inference. In *Proceedings of the 18th Joint Conference on Artificial Intelligence*, 985–991.

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

[Poon and Domingos 2011] Poon, H., and Domingos, P. 2011. Sum-product networks: A new deep architecture. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 337–346.

[Richardson and Domingos 2006] Richardson, M., and Domingos, P. 2006. Markov logic networks. *Machine Learning* 62(1-2).

[Singla and Domingos 2008] Singla, P., and Domingos, P. 2008. Lifted first-order belief propagation. In *Proceedings of the 23rd national conference on Artificial intelligence*, 1094–1099.

[Stachniss, Grisetti, and Burgard 2005] Stachniss, C.; Grisetti, G.; and Burgard, W. 2005. Proceedings of the international conference on robotics and automation. In *ICRA*, 655–660.

[Teh, Newman, and Welling 2006] Teh, Y. W.; Newman, D.; and Welling, M. 2006. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, 1353–1360.

[Van den Broeck 2011] Van den Broeck, G. 2011. On the completeness of first-order knowledge compilation for lifted probabilistic inference. In *Neural Information Processing Systems*, 1386–1394.

[Venugopal and Gogate 2012] Venugopal, D., and Gogate, V. 2012. On lifting the gibbs sampling algorithm. In *Advances in Neural Information Processing Systems 25*. 1664–1672.

Proof of Lemma 3

Lemma. *Let $\mathbf{X}$ be a finite sequence of random variables with joint distribution $P(\mathbf{X})$, let $\mathfrak{G}$ be an automorphism group of $\mathbf{X}$, let $\hat{\mathbf{X}}$ be a subsequence of $\mathbf{X}$, let $\hat{\mathbf{X}}^{\mathfrak{G}}$ be the orbit of $\hat{\mathbf{X}}$, and let $s \in \mathsf{Val}(\mathbf{X})$. Then,*

$$P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid s^{\mathfrak{G}}) = \frac{\mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}(s)}{|\hat{\mathbf{X}}^{\mathfrak{G}}|} = \mathbb{E}[\hat{\theta}_N \mid \mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}(s)].$$

Proof. Let $\mathfrak{G}_s := \{\mathfrak{g} \in \mathfrak{G} \mid s^{\mathfrak{g}} = s\}$ be the stabilizer subgroup of s. Then,

$$P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid s^{\mathfrak{G}}) = \frac{|\{\mathfrak{g} \in \mathfrak{G} \mid s^{\mathfrak{g}}\langle\hat{\mathbf{X}}\rangle = \hat{\mathbf{x}}\}|}{|\mathfrak{G}|}$$

since for each $x \in s^{\mathfrak{G}}$ we have that $|\{\mathfrak{g} \in \mathfrak{G} \mid s^{\mathfrak{g}} = x\}| = |\mathfrak{G}_s|$ by the orbit stabilizer theorem. For each $\mathbf{A} \in \hat{\mathbf{X}}^{\mathfrak{G}}$ let $\mathfrak{G}^{\mathbf{A}} := \{\mathfrak{g} \in \mathfrak{G} \mid \mathbf{A}^{\mathfrak{g}} = \hat{\mathbf{X}}\}$. Again, by the orbit stabilizer theorem, we have that $|\mathfrak{G}^{\mathbf{A}}| = |\mathfrak{G}_{\hat{\mathbf{X}}}|$ for each $\mathbf{A} \in \hat{\mathbf{X}}^{\mathfrak{G}}$, where $\mathfrak{G}_{\hat{\mathbf{X}}}$ is the stabilizer subgroup of $\hat{\mathbf{X}}$. Hence,

$$\frac{|\{\mathfrak{g} \in \mathfrak{G} \mid s^{\mathfrak{g}}\langle\hat{\mathbf{X}}\rangle = \hat{\mathbf{x}}\}|}{|\mathfrak{G}|} = \frac{\mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}(s)|\mathfrak{G}_{\hat{\mathbf{X}}}|}{|\mathfrak{G}|} = \frac{\mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}(s)}{|\hat{\mathbf{X}}^{\mathfrak{G}}|}.$$

Hence, $\mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}$ is a sufficient statistic for the marginal density $P(\hat{\mathbf{X}} = \hat{\mathbf{x}})$. Moreover, we have that

$$P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid s^{\mathfrak{G}}) = \mathbb{E}[\hat{\theta}_N \mid \mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}(s)].$$

This concludes the proof. □

Proof of Theorem 5

Theorem. *Let $\mathbf{X}$ be a finite sequence of random variables with joint distribution $P(\mathbf{X})$, let $\mathfrak{G}$ be an automorphism group of $\mathbf{X}$ given by R generators, let $\hat{\mathbf{X}}$ be a subsequence of $\mathbf{X}$, let $\hat{\mathbf{X}}^{\mathfrak{G}}$ be the orbit of $\hat{\mathbf{X}}$, and let $\theta := P(\hat{\mathbf{X}} = \hat{\mathbf{x}})$ be the marginal density to be estimated. The Rao-Blackwell estimator $\hat{\theta}^{\mathrm{rb}}_N$ has the following properties:*

(a) Its worst-case time complexity is $O(R|\hat{\mathbf{X}}^{\mathfrak{G}}|+N|\hat{\mathbf{X}}^{\mathfrak{G}}|)$;

(b) $MSE[\hat{\theta}^{\mathrm{rb}}_N] \leq MSE[\hat{\theta}_N]$.

The inequality of (b) is strict if there exists a joint assignment s with non-zero density and $0 < \mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}(s) < |\hat{\mathbf{X}}^{\mathfrak{G}}| > 1$.

Proof. We first construct the set $\hat{\mathbf{X}}^{\mathfrak{G}}$ once, which has a worst-case time complexity of $R|\hat{\mathbf{X}}^{\mathfrak{G}}|$ (Holt, Eick, and O'Brien 2005). For each sample point, we have to access an array representing the values of the sample point at most $|\hat{\mathbf{X}}^{\mathfrak{G}}|$ times. This allows us, for each sample point s, to compute $P(\hat{\mathbf{X}} = \hat{\mathbf{x}} \mid s^{\mathfrak{G}})$ in time $O(|\hat{\mathbf{X}}^{\mathfrak{G}}|)$ by Lemma 3.

Since $\mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}$ is a sufficient statistic for θ and $\hat{\theta}^{\mathrm{rb}}_N = \mathbb{E}[\hat{\theta}_N \mid \mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}]$ by Lemma 3, statement (b) follows from the Rao-Blackwell theorem (Blackwell 1947). If there exists a joint assignment s with non-zero density and $0 < \mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}(s) < |\hat{\mathbf{X}}^{\mathfrak{G}}| > 1$, then $\hat{\theta}_N$ is not a function of $\mathsf{H}^{\mathfrak{G}}_{\hat{\mathbf{X}}=\hat{\mathbf{x}}}$ and the inequality is strict (Blackwell 1947). □

Proof of Theorem 9

Theorem. *Let $\mathbf{X}$ be a finite sequence of random variables with joint distribution $P(\mathbf{X})$, let $\mathcal{M}$ be an ergodic Markov chain with stationary distribution P, and let $\mathcal{O}$ be an orbit partition of $\mathcal{M}$'s state space. Let $\hat{\theta}_N^{\mathrm{rb}}$ be the Rao-Blackwell estimator for N sample points $s_{T+1}, ..., s_{T+N}$ collected from $\mathcal{M}$, after discarding the first T sample points. Then, $|\mathtt{bias}(\hat{\theta}_N^{\mathrm{rb}})| \leq \epsilon$ if $T \geq \tau'(\epsilon)$, where $\tau'(\epsilon)$ is the mixing time of the quotient Markov chain of $\mathcal{M}$ with respect to $\mathcal{O}$.*

Proof. For a subsequence $\hat{\mathbf{X}}$ of $\mathbf{X}$, let $\xi := \hat{\mathbf{X}} = \hat{\mathbf{x}}$ be the marginal assignment whose density θ is to be estimated, let $\mathsf{Val}(\mathbf{X})$ be the assignment space of $\mathbf{X}$, and let $S = \{s_{T+1}, ..., s_{T+N}\}$ be the multiset of sample points collected from $\mathcal{M}$, after discarding the first T sample points. Since $\mathcal{O}$ is a partition of the assignment space, we have that

$$\hat{\theta}_N^{\mathrm{rb}} = \frac{1}{N} \sum_{s \in S} P(\xi \mid s^{\mathfrak{G}}) = \sum_{O \in \mathcal{O}} P(\xi \mid O) \frac{1}{N} \sum_{s \in S} \mathbb{I}_{\{s \in O\}}.$$

Hence,

$$\mathbb{E}[\hat{\theta}_N^{\mathrm{rb}}] = \sum_{O \in \mathcal{O}} P(\xi \mid O) \mathbb{E}[\mathbb{I}_{\{s \in O\}}],$$

where $\mathbb{E}[\mathbb{I}_{\{s \in O\}}]$ is the expectation of some sample point being located in the orbit O. $\mathbb{E}[\mathbb{I}_{\{s \in O\}}]$ defines a probability distribution over the space $\mathcal{O}$. If the sample points are independent, then $\mathbb{E}[\mathbb{I}_{\{s \in O\}}] = P(O)$, for all $O \in \mathcal{O}$, and the estimator is unbiased. Since we collect sample points from a Markov chain we will often have that $\mathbb{E}[\mathbb{I}_{\{s \in O\}}] \neq P(O)$.

By the assumptions and Proposition 7, the Markov chain $\mathcal{M}$ is exactly lumpable with respect to $\mathcal{O}$ and, hence, for all states x of $\mathcal{M}$, all $t \in \{1, 2, ...\}$, and all orbits $O \in \mathcal{O}$, we have that $\sum_{o \in O} Q^t(x, o) = Q'^t(x^{\mathfrak{G}}, O)$, where $Q^t(x, o)$ is the probability of the Markov chain $\mathcal{M}$ being in state o after t simulation steps if the chain starts in state x. In addition, we start collecting sample points after $T \geq \tau'(\epsilon)$ simulation steps and, thus,

$$\frac{1}{2} \sum_{O \in \mathcal{O}} |\mathbb{E}[\mathbb{I}_{\{s \in O\}}] - P(O)| \leq$$

$$\max_{x \in \mathsf{Val}(\mathbf{X})} \left\{ \frac{1}{2} \sum_{O \in \mathcal{O}} |Q'^T(x^{\mathfrak{G}}, O) - P(O)| \right\} \leq \epsilon.$$

Finally, $|\mathtt{bias}(\hat{\theta}_N^{\mathrm{rb}})| = |\mathbb{E}[\hat{\theta}_N^{\mathrm{rb}} - \theta]| =$

$$\left| \sum_{O \in \mathcal{O}} P(\xi \mid O) \mathbb{E}[\mathbb{I}_{\{s \in O\}}] - \sum_{O \in \mathcal{O}} P(\xi \mid O) P(O) \right| =$$

$$\left| \sum_{O \in \mathcal{O}} P(\xi \mid O)(\mathbb{E}[\mathbb{I}_{\{s \in O\}}] - P(O)) \right| \leq$$

$$\sum_{\substack{O \in \mathcal{O} \\ \mathbb{E}[\mathbb{I}_{\{s \in O\}}] \geq P(O)}} (\mathbb{E}[\mathbb{I}_{\{s \in O\}}] - P(O)) =$$

$$\frac{1}{2} \sum_{O \in \mathcal{O}} |\mathbb{E}[\mathbb{I}_{\{s \in O\}}] - P(O)| \leq \epsilon.$$

The last equality follows from a known identity of the total variation distance (Levin, Peres, and Wilmer 2008). $\qquad\square$