On Combining Graph-based Variance Reduction schemes

Vibhav Gogate
Computer science & Engineering
University of Washington, Seattle
Seattle, WA 98195, USA.
vgogate@cs.washington.edu

Rina Dechter
School of Information and Computer Sciences
University of California, Irvine
Irvine, CA 92697, USA.
dechter@ics.uci.edu

Abstract

In this paper, we consider two variance reduction schemes that exploit the structure of the primal graph of the graphical model: *Rao-Blackwellised w-cutset sampling* and *AND/OR sampling*. We show that the two schemes are orthogonal and can be combined to further reduce the variance. Our combination yields a new family of estimators which trade time and space with variance. We demonstrate experimentally that the new estimators are superior, often yielding an order of magnitude improvement over previous schemes on several benchmarks.

1 Introduction

Importance sampling (Rubinstein, 1981) is a general scheme which can be used to approximate various weighted counting tasks defined over graphical models such as computing the probability of evidence in a Bayesian network, computing the partition function of a Markov network and counting the number of solutions of a constraint network. The main idea is to transform the weighted counts or summation into an expectation using a special distribution called the proposal distribution, generate samples from the proposal and estimate the weighted counts by a weighted average (also called the sample mean) over the generated samples. It is well known that the quality of estimation is highly dependent on the variance of the sample mean and therefore significant research has focused on reducing its variance (Liu, 2001).

In this paper, we consider two graph-based variance reduction schemes in the context of graphical models: the *Rao-Blackwellised w-cutset sampling scheme* (Bidyuk and Dechter, 2007) and the *AND/OR sampling scheme* (Gogate and Dechter, 2008). Based on the Rao-Blackwell theorem (Casella and Robert, 1996) and w-cutset conditioning (Dechter, 1990), the w-cutset sampling scheme combines sampling with exact inference. The idea is to sample only a subset $\mathbf{C}$ of variables, called the *w*-cutset and exactly marginalize out the remaining variables conditioned on each sampled assignment. The AND/OR sampling scheme, on the other hand, reduces variance by exploiting conditional independencies uncovered by the AND/OR tree or graph (Dechter and Mateescu, 2007) to derive a different sample mean from the same set of input samples. Previously in (Gogate and Dechter, 2008), we considered two alternative AND/OR sample means: one based on AND/OR tree which has the same time and space complexity as the conventional OR tree approach but has smaller variance and the second based on AND/OR graph which is more expensive to compute but has the smallest variance.

The main idea in this paper is to combine these two schemes by performing AND/OR tree or graph sampling over the w-cutset variables and exact inference over the remaining variables conditioned on each sampled assignment. We show that this yields new sample means, which have smaller variance than the sample means of AND/OR sampling and w-cutset sampling. However, they are more expensive to compute both time and space wise and thus there is a trade-off.

We conducted extensive experimental evaluation of all the new schemes proposed on several benchmark probabilistic and deterministic networks. Our results show that as the networks get larger and harder, exploiting more decomposition improves the accuracy of the estimates as a function of time. In particular, the scheme that exploits the most decomposition, the AND/OR w-cutset graph sampling scheme is superior to all the other schemes.

The rest of the paper is organized as follows. In Section 2, we present notation and background. Section 3 describes AND/OR w-cutset tree sampling and Section 4 describes AND/OR w-cutset graph sampling. Complexity versus variance trade-offs are discussed in Section 5. Experiments

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

are described in Section 6 and Section 7 concludes.

2 Background

We start by presenting notation and preliminaries on graphical models. Then we present an overview of importance sampling, *w*-cutset sampling and AND/OR sampling.

We denote variables by upper case letters (e.g. $X, Y, \ldots$) and values of variables by lower case letters (e.g. $x, y, \ldots$). Sets of variables are denoted by bold upper case letters, (e.g. $\mathbf{X} = \{X_1, \ldots, X_n\}$) while sets of values are denoted by bold lower case letters (e.g. $\mathbf{x} = \{x_1, \ldots, x_n\}$). We denote by $\mathbf{D}_i$ the set of possible values of X_i (also called as the domain of X_i). $\sum_{\mathbf{x}\in\mathbf{X}}$ denotes the sum over the possible values of variables in $\mathbf{X}$, namely, $\sum_{x_1\in X_1}\sum_{x_2\in X_2}\cdots\sum_{x_n\in X_n}$. The expected value $\mathbb{E}_Q[X]$ of a random variable X with respect to a distribution Q is defined as: $\mathbb{E}_Q[X] = \sum_{x\in X} xQ(x)$. The variance $V_Q[X]$ of X is defined as: $V_Q[X] = \sum_{x\in X}(x - \mathbb{E}_Q[X])^2$.

Definition 1 (Graphical models). *(Pearl, 1988) A graphical model is a three-tuple $\mathscr{M} = \langle \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{F}\rangle$ where $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ is a set of random variables, $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n\}$ is a set of domains where $\boldsymbol{D}_i$ is the domain of X_i and $\boldsymbol{F} = \{F_1, , \ldots, F_m\}$ is a set of non-negative real valued functions where each F_i is defined over a subset of variables $\boldsymbol{S}_i \subset \boldsymbol{X}$, called its scope. A graphical model represents a joint distribution over $\boldsymbol{X}$ given by: $P_{\mathscr{M}}(\mathbf{x}) = \frac{1}{Z}\prod_{i=1}^m F_i(\mathbf{x})$ where Z is a normalization constant given by: $Z = \sum_{\mathbf{x}\in\boldsymbol{X}}\prod_{i=1}^m F_i(\mathbf{x})$. We will often refer to Z as weighted counts. The **primal graph** of a graphical model is an undirected graph which has variables as its vertices and an edge between any two variables which are included in the scope of a function.*

We will focus on the query of computing the weighted counts Z. It is easy to show that the weighted counts specialize to the probability of evidence of a Bayesian network, the partition function of a Markov network and the number of solutions of a constraint network.

2.1 Importance Sampling

Importance sampling (Rubinstein, 1981; Liu, 2001) is a general Monte Carlo simulation technique which can be used for estimating various statistics of a given target distribution such as $P_{\mathscr{M}}$. Since it is often hard to sample from $P_{\mathscr{M}}$, the main idea is to generate samples from another easy-to-simulate distribution Q called the proposal (or importance) distribution and then estimate various statistics over $P_{\mathscr{M}}$ by a weighted average over the samples. Following (Cheng and Druzdzel, 2000), we assume that the proposal distribution is specified in a factored product form (namely a Bayesian network): $Q(\mathbf{X}) = \prod_{i=1}^n Q_i(X_i|X_1,\ldots,X_{i-1}) = \prod_{i=1}^n Q_i(X_i|\mathbf{Y_i})$ along an ordering $o = (X_1,\ldots,X_n)$ of variables, where $\mathbf{Y}_i \subseteq \{X_1,\ldots,X_{i-1}\}$. The cardinality of the set $\mathbf{Y}_i$ is assumed to be bounded by a constant.

Next, we show how the weighted counts can be estimated using importance sampling. Given a (importance) proposal distribution $Q(\mathbf{X})$ satisfying $\prod_{i=1}^m F_i(\mathbf{x}) > 0 \Rightarrow Q(\mathbf{x}) > 0$, we can rewrite Z as follows:

$$Z = \sum_{\mathbf{x}\in\mathbf{X}} \frac{\prod_{i=1}^m F_i(\mathbf{x})}{Q(\mathbf{x})} Q(\mathbf{x}) = \mathbb{E}_Q\left[\frac{\prod_{i=1}^m F_i(\mathbf{x})}{Q(\mathbf{x})}\right] \quad (1)$$

Given independent and identically distributed (i.i.d.) samples $(\mathbf{x}^1,\ldots,\mathbf{x}^N)$ generated from Q, we can estimate Z by:

$$\widehat{Z} = \frac{1}{N}\sum_{k=1}^N \frac{\prod_{i=1}^m F_i(\mathbf{x}^k)}{Q(\mathbf{x}^k)} = \frac{1}{N}\sum_{k=1}^N w(\mathbf{x}^k) \quad (2)$$

where $w(\mathbf{x}^k) = \frac{\prod_{i=1}^m F_i(\mathbf{x}^k)}{Q(\mathbf{x}^k)}$ is the weight of sample $\mathbf{x}^k$. It is easy to see that $\mathbb{E}_Q[\widehat{Z}] = Z$, namely it is unbiased. The variance of the weights is given by:

$$V_Q[w(\mathbf{x})] = \sum_{\mathbf{x}\in\mathbf{X}} (w(\mathbf{x}) - Z)^2 Q(\mathbf{x}) \quad (3)$$

Note that $\widehat{Z}$ is itself a random variable and its variance is given by:

$$V_Q[\widehat{Z}] = \frac{1}{N}\sum_{\mathbf{x}\in\mathbf{X}} (w(\mathbf{x}) - Z)^2 Q(\mathbf{x}) = \frac{V_Q[w(\mathbf{x})]}{N} \quad (4)$$

Because the mean-squared error of an unbiased estimate such as $\widehat{Z}$ is equal to its variance, we would like the variance of $\widehat{Z}$ to be as small as possible. Based on Equation 4, we can reduce the variance by decreasing the variance of the weights or by increasing the number of samples or both. Next, we present two schemes that use graph decompositions to either reduce $V_Q[w(\mathbf{x})]$ or increase N.

2.2 Rao-Blackwellised w-cutset Importance Sampling

The *w*-cutset sampling framework is based on a graph concept called w-cutset and the Rao-Blackwell theorem (Casella and Robert, 1996).

Theorem 1 (Rao-Blackwell Theorem). *Let $F(\boldsymbol{Y},\boldsymbol{Z})$ be a function and $Q(\boldsymbol{Y},\boldsymbol{Z})$ be a proposal distribution then, $V_Q\left[\frac{F(\boldsymbol{y},\boldsymbol{z})}{Q(\boldsymbol{y},\boldsymbol{z})}\right] \geq V_Q\left[\frac{F(\boldsymbol{y})}{Q(\boldsymbol{y})}\right]$ where $Q(\boldsymbol{y}) = \sum_{\boldsymbol{z}} Q(\boldsymbol{y},\boldsymbol{z})$ and $F(\boldsymbol{y}) = \sum_{\boldsymbol{z}} F(\boldsymbol{y},\boldsymbol{z})$.*

Definition 2 (w-cutset). *Given a graph $G(\boldsymbol{X},\boldsymbol{E})$ and a constant w, a w-cutset is a sub-set of variables $\boldsymbol{C} \subseteq \boldsymbol{X}$ such that after removing $\boldsymbol{C}$, the treewidth (see for e.g., (Dechter, 1999) for a definition of treewidth) of the remaining graph is bounded by w. A cycle cutset is a 1-cutset of G. $\boldsymbol{X}$ is a 0-cutset.*

In w-cutset sampling, we sample only the variables in the *w*-cutset $\mathbf{C}$ and perform exact computations (e.g. using bucket elimination (Dechter, 1999)) on the remaining variables $\mathbf{R} = \mathbf{X} \setminus \mathbf{C}$ given a sample $\mathbf{C} = \mathbf{c}$. Because the time

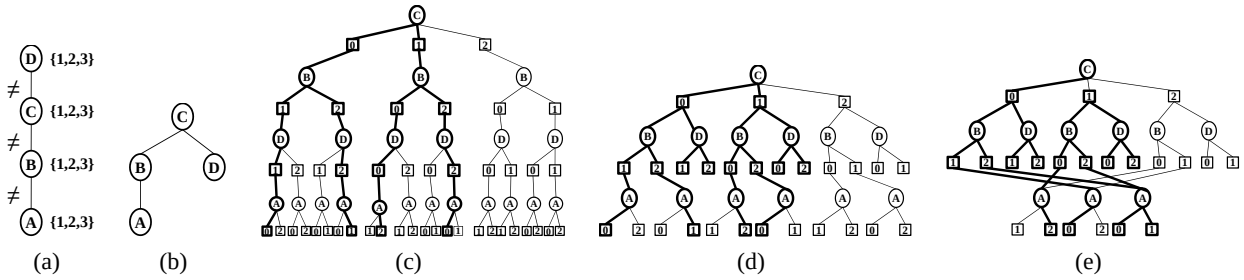

Figure 1: (a) A 3-coloring problem, (b) Pseudo-tree (c) OR tree (d) AND/OR tree (e) AND/OR graph

and space complexity of bucket elimination is exponential in the treewidth of the graph (Dechter, 1999), it is obvious that given a *w*-cutset, bucket elimination can be carried out efficiently in polynomial time (exponential in the constant *w*). Formally, given a proposal distribution $Q(\mathbf{C})$ defined over the *w*-cutset, and a set of samples $(\mathbf{c}^1,\ldots,\mathbf{c}^N)$ generated from Q, the *w*-cutset estimate of Z is given by:

$$\widehat{Z}_{wc} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{\mathbf{r}\in\mathbf{R}}\prod_{j=1}^{m}F_j(\mathbf{r},\mathbf{C}=\mathbf{c}^i)}{Q(\mathbf{c}^i)} \quad (5)$$

From the Rao-Blackwell theorem, it follows that the variance of $\widehat{Z}_{wc}$ is less than $\widehat{Z}$. *w*-cutset sampling generalizes importance sampling in the following sense. When $w = 0$, $\widehat{Z}_{wc}$ equals the conventional sample mean $\widehat{Z}$.

2.3 AND/OR Importance Sampling

AND/OR importance sampling (for more information, see (Gogate and Dechter, 2008)) is a generalization of importance sampling to AND/OR search spaces (Dechter and Mateescu, 2007). The main idea is to arrange the generated samples over an AND/OR tree or graph and then utilize conditional independencies to derive a larger set of virtual samples. The structure of the AND/OR tree is guided by a backbone pseudo-tree of Q defined below.

Definition 3 (pseudo-tree). *Given an undirected graph $G = (\boldsymbol{V},\boldsymbol{E}')$, a directed rooted tree $T = (\boldsymbol{V},\boldsymbol{E})$ defined on all its nodes is called pseudo tree if any arc of G which is not included in $\boldsymbol{E}$ is a back-arc, namely it connects a node to an ancestor in T.*

AND/OR search Tree Given a graphical model $\mathscr{M} = \langle\mathbf{X},\mathbf{D},\mathbf{F}\rangle$, its primal graph G and a backbone pseudo tree T of G, the associated AND/OR search tree S_T, has alternating levels of AND and OR nodes. The OR nodes are labeled X_i and correspond to the variables. The AND nodes are labeled by x_i and correspond to the value assignments in the domains of the variables. The structure of the AND/OR search tree is based on T. The root of the AND/OR search tree is an OR node labeled by the root of T. The children of an OR node X_i are AND nodes labeled with assignment x_i, which is consistent along the path from the root. The children of an AND node x_i are OR nodes labeled with the children of variable X_i in T. When the pseudo tree is a chain, the AND/OR search tree coincides with the regular OR search tree.

AND/OR search Tree Given a set of samples $\mathbf{S} = \{\mathbf{x}^1,\ldots,\mathbf{x}^N\}$, the AND/OR sample tree (similarly, the OR sample tree) is a subset of the full AND/OR search tree (OR search tree) from which all nodes not in $\mathbf{S}$ are removed. The set of unique samples of an AND/OR sample tree is equal to the set of its solution sub-trees, which is defined recursively as follows. A *solution sub-tree* contains the root node. For each OR node, it contains one of its children and for each AND node, it contains all of its children.

Example 1. *Figure 1(a) shows a primal graph of a 3-coloring problem over* 4 *variables. A possible pseudo tree is given in Figure 1(b). The full OR and AND/OR search trees are shown in Figures 1(c) and 1(d) respectively. Let us assume that we have generated the four samples: (1) (C=0,B=1,D=1,A=0), (2) (C=0,B=2,D=2,A=1), (3) (C=1,B=0,D=0,A=2) and (4) (C=1, B=2, D=2,A=0) from a proposal distribution that is defined along the topological order of the pseudo tree. The bold edges and nodes in Figures 1 (c) and (d) show these four samples arranged on an OR tree and an AND/OR tree respectively. One can verify that the* 4 *samples (solution sub-trees) over the OR sample tree correspond to* 8 *virtual samples over the AND/OR sample tree. The AND/OR sample tree includes for example the assignment (C=0,B=2,D=1,A=0) which does not appear in the OR sample tree.*

Because of this larger virtual sample size, we can prove that the variance of the sample mean computed over the AND/OR sample tree is smaller than or equal to the variance of the sample mean over the conventional OR sample tree (Gogate and Dechter, 2008). Note that the sample mean computed over an OR sample tree equals $\widehat{Z}$ and thus AND/OR sampling generalizes importance sampling.

3 AND/OR w-cutset Tree sampling

We now describe one of our main contributions in which we combine *w*-cutset sampling with AND/OR sampling. We illustrate the main idea in the following example.

Example 2. *Consider the primal graph in Figure 2 (a). The minimal cycle cutset contains the three nodes $\{A,B,C\}$. Given a proposal distribution $Q(A,B,C) = Q(A)$ $Q(B|A)$ $Q(C|A)$, in cycle cutset sampling, the variables A, B and C are sampled, as if they form a chain pseudo-tree shown in Figure 2(c) before executing bucket elimination*

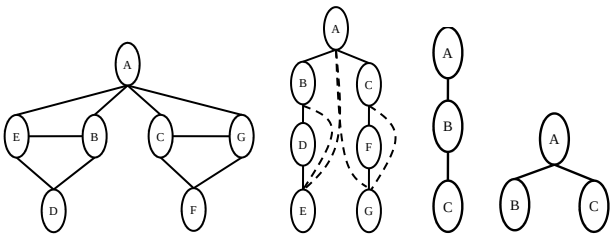

Figure 2: (a) an example primal graph (b) Pseudo tree (c) Start pseudo-tree of OR w-cutset tree sampling and (d) Start pseudo-tree of AND/OR w-cutset tree sampling.

on the remaining network defined by $\{D,E,F,G\}$ *given* $(A=a,B=b,C=c)$ *to compute the sample weight.*

However, after A is sampled, we see that the remaining sub-problem is split into two components and therefore we can organize the cycle cutset into two portions as in the (start) pseudo-tree of Figure 2(d) (Mateescu and Dechter, 2005). We can now arrange the generated samples on an AND/OR sample tree restricted over the cutset variables $\{A,B,C\}$ *and separately compute the weighted counts (using bucket elimination) over the networks defined by the two components* $\{D,E\}$ *given* $(A=a,B=b)$ *and* $\{F,G\}$ *given* $(A=a,C=c)$ *respectively.*

We now formalize the intuition in Example 2, defining a new sample mean called AND/OR *w*-cutset sample tree mean. We start with some required definitions.

Definition 4 (AND/OR w-cutset, start and full pseudo trees). *(Mateescu and Dechter, 2005) Given a pseudo tree* $T(\boldsymbol{V},\boldsymbol{E})$*, a directed rooted tree* $T'(\boldsymbol{V'},\boldsymbol{E'})$ *where* $\boldsymbol{V'} \subseteq \boldsymbol{V}$ *and* $\boldsymbol{E'} \subseteq \boldsymbol{E}$ *is a start pseudo tree of* T *if it has the same root as* T *and is a connected sub-graph of* T*.* T *is called the* full pseudo tree *of its start pseudo tree* T'*. An AND/OR w-cutset is a pair* $\langle T',\boldsymbol{K}\rangle$ *where* $\boldsymbol{K}$ *is a w-cutset and* $T'(\boldsymbol{K},\boldsymbol{E'})$ *is a start pseudo tree defined over* $\boldsymbol{K}$*.*

Example 3. *The pseudo tree given in Figure 2(d) is a start pseudo tree of the (full) pseudo tree given in Figure 2(b). The AND/OR w-cutset is the set of variables* $\{A,B,C\}$ *together with the start pseudo tree of Figure 2(d).*

Definition 5 ((Arc-labeled) AND/OR w-cutset sample tree). *Given a graphical model* $\mathscr{M} = \langle \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{F} \rangle$*, an AND/OR w-cutset* $\langle T',\boldsymbol{K}\rangle$*, a full pseudo tree* T *of* T'*, a proposal distribution along the pseudo tree* $Q(\boldsymbol{X})$*, and a sequence of samples* $\mathbf{S}$ *over* T'*, an AND/OR w-cutset sample tree* S_{AOWT} *is a subset of full AND/OR search tree over* $\boldsymbol{K}$ *w.r.t.* T' *from which all assignments not in* $\boldsymbol{S}$ *are removed.*

A path from the root of S_{AOWT} *to a node n is denoted by* π_n*. If n is an OR node labeled with* X_i *or an AND node labeled with* x_i*, the path will be denoted by* $\pi_n(X_i)$ *or* $\pi_n(x_i)$ *respectively. The assignment sequence along the path* π_n*, denoted by* $A(\pi_n)$ *is the set of assignments associated with the sequence of AND nodes along* π_n*. Namely,* $A(\pi_n(X_i)) = \{x_1,\ldots,x_{i-1}\}$ *and* $A(\pi_n(x_i)) = \{x_1,\ldots,x_i\}$*.*

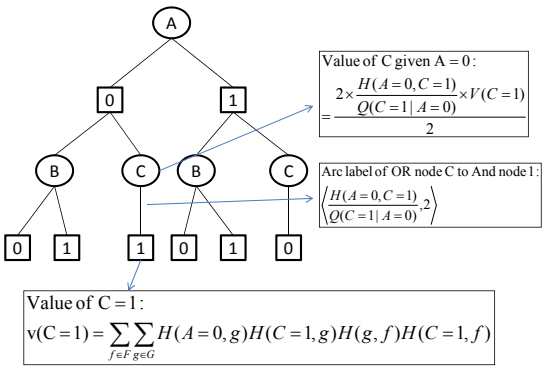

Figure 3: Figure demonstrating computation of arc-labels and node values over an AND/OR w-cutset sample tree.

Each arc from an OR node n labeled by X_i *to an AND node m labeled by* x_i *is labeled with a pair* $\langle w(n,m),\#(n,m)\rangle$*.* $w(n,m)$ *is called the weight of the arc and is given by* $w(n,m) = \frac{B_{T,X_i}(x_i,A(\pi_n))}{Q_i(x_i|A(\pi_n))}$ *where* $B_{T,X_i}(x_i,A(\pi_n))$ *is the product of all functions in* $\boldsymbol{F}$ *that mention* X_i *but do not mention any variables that are descendants of* X_i *in* T*.* $\#(n,m)$ *is the frequency of the arc. It equals the number of times the partial assignment* $A(\pi_m)$ *occurs in* $\boldsymbol{S}$*.*

Note that an OR w-cutset sample tree is an AND/OR w-cutset sample tree based on a chain start pseudo tree.

Definition 6 (AND/OR w-cutset sample tree mean). *Given an AND/OR w-cutset sample tree* S_{AOWT}*, the value of a node n, denoted by* $v(n)$ *is defined recursively as follows. The value of leaf AND node l is given by:*

$$v(l) = \sum_{\boldsymbol{u}\in lpath_T(X_i)} \prod_{X_j\in lpath_T(X_i)} B_{T,X_j}(\boldsymbol{u},A(\pi_l)) \qquad (6)$$

where $lpath_T(X_i)$ *is the set of variables along the path from* X_i *to the leaf l in the full pseudo tree* T*. If n is an internal AND node then:* $v(n) = \prod_{n'\in chi(n)} v(n')$ *and if n is an internal OR node then,*

$$v(n) = \frac{\sum_{n'\in chi(n)} \#(n,n') \times w(n,n') \times v(n')}{\sum_{n'\in chi(n)} \#(n,n')}$$

where $chi(n)$ *is the set of child nodes of node n in* S_{AOWT}*. The AND/OR w-cutset sample tree mean is the value of the root node of* S_{AOWT}*.*

Example 4. *Figure 3 shows an AND/OR w-cutset sample tree corresponding to the four samples shown in Figure 3 w.r.t. the start pseudo tree shown in Figure 2(d). We assume that all the functions in our graphical model are pairwise. Namely, we have functions corresponding to each edge: H(A,B), H(A,E), ... , H(F,G). The arc-labels and values of a few arcs and nodes are shown in Figure 3. The AND/OR w-cutset sample mean is the value of the node A.*

The AND/OR w-cutset sample tree mean generalizes the sample means of importance sampling, w-cutset impor-

tance sampling and AND/OR tree importance sampling in the following sense.

Proposition 1. *The sample mean obtained via conventional importance sampling is equal to the OR 0-cutset sample tree mean. The w-cutset sample mean is equal to the OR w-cutset sample tree mean. The AND/OR sample tree mean defined in (Gogate and Dechter, 2008) is equal to the AND/OR 0-cutset sample tree mean.*

Using simple algebraic manipulations, we can prove that:

Theorem 2. *The AND/OR w-cutset sample tree mean is an unbiased estimate of the weighted counts Z.*

Using the Rao-Blackwell theorem and AND/OR theory, we can prove the following two theorems showing the superiority of our hybrid scheme over its individual components.

Theorem 3. *Given $w \geq 0$, the variance of AND/OR w-cutset sample tree mean is less than or equal to the variance of OR w-cutset sample tree mean.*

Theorem 4. *Given $w > 0$, the variance of AND/OR w-cutset sample tree mean is less than or equal to the variance of AND/OR 0-cutset sample tree mean.*

4 AND/OR graph w-cutset sampling

Next, we define a more powerful sample mean by moving from AND/OR trees to AND/OR graphs (Dechter and Mateescu, 2007). An AND/OR-tree may contain nodes that root identical sub-trees. When such unifiable nodes are merged, the tree becomes a graph and its size becomes smaller. Some unifiable nodes can be identified using contexts defined below.

Definition 7 (**Context**). *Given a pseudo-tree $T(\boldsymbol{V},\boldsymbol{E})$, the context of a node $X_i \in \boldsymbol{V}$ is the set of ancestors of X_i, that are connected to X_i and descendants of X_i.*

Example 5. *For illustration, the bold nodes in Figure 1(e) show* 8 *virtual samples (solution sub-trees) of the AND/OR sample tree of Figure 1(d) arranged on an AND/OR sample graph by merging context unifiable nodes (based on the conditional independence assertion that A is independent of C given B). One can verify that the* 8 *virtual samples on the AND/OR sample tree correspond to* 12 *virtual samples (solution sub-trees) on the AND/OR sample graph. The AND/OR sample graph includes for example the sample $(C = 0, B = 2, D = 1, A = 0)$ which is not present in the AND/OR sample tree. Due to an increase in the virtual sample size, the variance of AND/OR sample graph mean is smaller than (or equal to) that of AND/OR sample tree mean (Gogate and Dechter, 2008).*

The main idea in AND/OR *w*-cutset graph estimation is to store all the generated samples on an AND/OR *w*-cutset graph instead of an AND/OR *w*-cutset tree and then compute a new sample mean over the AND/OR *w*-cutset sample graph, which will have smaller variance. Formally,

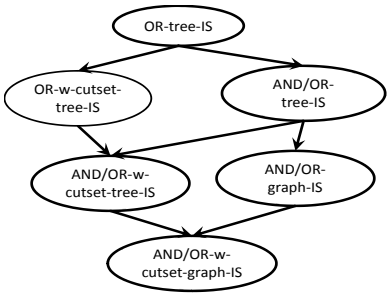

Figure 4: Variance Hierarchy

Definition 8 (AND/OR *w*-cutset sample graph and mean)**.** *Given an AND/OR w-cutset sample tree S_{AOWT}, the AND/OR w-cutset sample graph S_{AOWG} is obtained from S_{AOWT} by merging all nodes based on context. The AND/OR w-cutset sample graph mean is the AND/OR sample mean computed over S_{AOWG}.*

Theorem 5. *Given $w \geq 0$, the variance of the AND/OR w-cutset sample graph mean is less than or equal to the variance of AND/OR w-cutset sample tree mean. Given $w > 0$, the variance of the AND/OR w-cutset sample graph mean is less than or equal to the variance of AND/OR* 0*-cutset sample graph mean.*

5 Complexity and Variance Hierarchy

Theorems 3, 4 and 5 along with the Rao-Blackwell theorem help us establish the variance hierarchy shown in Figure 4. The main assumption is that all sample means are based on the same set of samples and the same full and start pseudo trees. Semantically, given a *w*-cutset where $0 < w < t^*$ (t^* is the treewidth), the directed arcs in Figure 4 indicate that the variance of the child node is less than (or equal to) the variance of the parent. We see that the variance of AND/OR sample tree mean is incomparable with w-cutset sample mean. We also see that AND/OR w-cutset sample graph mean has the lowest variance.

We summarize the complexity of computing various sample means in the following theorem:

Theorem 6. *Given a graphical model $\mathscr{M}$ having n variables, an AND/OR w-cutset $\langle T', \boldsymbol{X} \rangle$ of $\mathscr{M}$, a full pseudo tree T of T' and a proposal distribution Q defined along the pseudo-tree T, let h be the height and t^* be the maximum context size (treewidth) of T and let c be the size of the w-cutset. Given N samples generated i.i.d. from Q, the complexity of computing AND/OR w-cutset tree and graph sample means is given in the following table:*

Sample mean	*Time Complexity*	*Space Complexity*
OR 0-cutset tree	$O(nN)$	$O(1)$
AND/OR 0-cutset tree	$O(nN)$	$O(h)$
AND/OR 0-cutset	$O(nNt^*)$	$O(nN)$
OR w-cutset tree	$O(cN+(n-c)Nexp(w))$	$O((n-c)exp(w))$
AND/OR w-cutset tree	$O(cN+(n-c)Nexp(w))$	$O(h+(n-c)exp(w))$
AND/OR w-cutset graph	$O(cNt^*+(n-c)Nexp(w))$	$O(cN+(n-c)exp(w))$

From Theorem 6, we see that variance reduction comes

at an extra computational cost. In particular, as we move down the variance hierarchy, the time and space complexity of the schemes typically increases.

6 Experimental Results

In this section, we demonstrate empirically that the AND/OR w-cutset tree and graph sampling schemes are superior in terms of accuracy to OR w-cutset sampling and AND/OR (0-cutset) tree and graph sampling.

6.1 Experimental Setup

The strength of AND/OR w-cutset estimates is that the samples on which the estimates are based upon can be generated using any importance sampling scheme. Therefore, in order to demonstrate the impact of our new schemes in a non-trivial setting, we generate samples using state-of-the-art techniques such as IJGP-IS (Gogate and Dechter, 2005) and IJGP-SampleSearch (Gogate and Dechter, 2007).

IJGP-IS uses the output of Iterative Join graph propagation (IJGP) (Dechter et al., 2002; Mateescu et al., 2010) to compute a proposal distribution because it was shown to yield good approximation to the posterior distribution $P_{\mathscr{M}}$ (Yedidia et al., 2004; Dechter et al., 2002; Yuan and Druzdzel, 2006). IJGP is a generalized belief propagation scheme parameterized by a constant *i*, called the *i*-bound, yielding a class of algorithms $(IJGP(i))$ whose complexity is exponential in *i*, that allow a trade-off between accuracy and complexity. As *i* increases, accuracy generally increases. When *i* equals the treewidth of the graphical model, $IJGP(i)$ is exact. We use a *i*-bound of 5 and set the number of iterations to 10 in all our experiments to ensure that IJGP terminates in a reasonable amount of time. On benchmarks which have strong deterministic relationships (specifically the linkage and coloring instances), we use IJGP-based SampleSearch specialized to handle the rejection problem (Gogate and Dechter, 2007).

We experimented with the following schemes: (a) OR tree importance sampling (or-tree-IS) (b) AND/OR tree importance sampling (ao-tree-IS), (c) AND/OR graph importance sampling (ao-graph-IS) (d) OR tree w-cutset importance sampling (or-wc-tree-IS) (e) AND/OR w-cutset tree importance sampling (ao-wc-tree-IS) and (f) AND/OR w-cutset graph importance sampling (ao-wc-graph-IS). The last two are the new schemes.

We used the min-fill ordering to generate the pseudo-trees. We set the *w* of *w*-cutset to 5, again to ensure that the bucket elimination component of w-cutset sampling does not run out of memory and terminates in a reasonable amount of time. We generated the w-cutset using a greedy scheme outlined in (Bidyuk and Dechter, 2004). This scheme requires a tree-decomposition as input, which was generated using the min-fill ordering. Note that the underlying scheme for generating the samples, namely the proposal distribution is identical in all the schemes.

All of our experiments were run on linux servers, each with dual 2.4Ghz processors and 2GB of memory. We experimented with three sets of benchmarks: (a) the grid networks, (b) the Linkage networks and (c) 4-coloring problems. We organize the results in two subsections. In the next subsection, we present results on instances for which the exact weighted counts are known and in subsection 6.3 we present results on instances for which the exact counts are not known. The reason for this separation is the difference in the evaluation criteria used.

6.2 Results on instances for which the exact weighted counts are known

Table 1 shows the results. For each instance, in column 2, we report the number of variables (n), average domain size (k), the number of evidence nodes (c) (or constraints for the graph coloring problem) and treewidth (t^*). The third column reports the exact value of the weighted counts. Columns 4-9 report the sample mean output by various schemes after 1hr of CPU time.

Grid networks Our first problem domain is that of partially deterministic $s \times s$ grid networks, available from the authors of Cachet (Sang et al., 2005). The last node in the grid network is called the sink node whose marginal probability is to be determined. Given a parameter called the *deterministic ratio*, a fraction of the functions in a grid are made deterministic by randomly filling them with 0 or 1. In Table 1, the instances are designated as $p-s$ where p is the deterministic ratio expressed as a percentage and s is the size of the grid. We observe that AND/OR w-cutset graph and tree schemes (ao-wc-graph-IS and ao-wc-tree-IS) are better than the other schemes, with the AND/OR *w*-cutset graph scheme being the best performing scheme. *w*-cutset importance sampling (or-wc-tree-IS) is slightly worse than AND/OR tree and graph schemes which do not use a w-cutset (ao-graph-IS and ao-tree-IS). Pure importance sampling (or-tree-IS) is the worst performing scheme.

Linkage Networks The linkage instances are generated by converting a pedigree to a Bayesian network (Fishelson and Geiger, 2003).

The *BN*_69 to *BN*_77 instances were used in the UAI 2006 evaluation (Bilmes and Dechter, 2006). We observe that on 6 out of the 9 instances, AND/OR *w*-cutset graph scheme (ao-wc-graph-IS) is more accurate than the AND/OR *w*-cutset tree scheme (ao-wc-tree-IS) which in turn is substantially more accurate than the OR-*w*-cutset tree scheme (or-wc-tree-IS). Pure importance sampling (or-tree-IS) is the worst performing scheme. On an average, we observe that the AND/OR *w*-cutset graph scheme (ao-wc-graph-IS) is the most accurate scheme.

Problem-name	$\langle n,k,c,t^* \rangle$	Exact	or-tree-IS	or-wc-tree-IS	ao-tree-IS	ao-graph-IS	ao-wc-tree-IS	ao-wc-graph-IS
Grids								
50-18-5	$\langle 324,2,1,18 \rangle$	0.4137	0.31	0.64	0.27	**0.422**	0.401	**0.422**
50-19-5	$\langle 361,2,1,19 \rangle$	0.2209	0.243	0.2301	0.244	0.23	0.2244	**0.2217**
50-20-5	$\langle 400,2,1,20 \rangle$	0.5692	0.34	0.333	0.542	0.567	0.571	**0.569**
75-22-5	$\langle 484,2,1,22 \rangle$	0.437	0.695	0.452	0.572	0.493	0.446	**0.435**
75-23-5	$\langle 529,2,1,23 \rangle$	0.348	0.25	0.380	0.19	0.299	0.331	**0.350**
75-26-5	$\langle 676,2,1,26 \rangle$	0.264	0.079	0.1432	0.124	0.19	0.184	**0.24**
90-34-5	$\langle 1156,2,1,34 \rangle$	0.0859	0.044	0.0452	0.0449	0.0516	0.0557	**0.082**
90-38-5	$\langle 1444,2,1,38 \rangle$	0.141	0.08	0.123	0.19	0.183	**0.143**	**0.143**
90-42-5	$\langle 1764,2,1,42 \rangle$	0.654	0.42	0.81	0.576	0.511	0.74	**0.67**
Linkage								
BN_69	$\langle 777,7,78,36 \rangle$	5.28E-054	3.31E-55	3.01E-55	2.58E-55	2.66E-55	3.00E-55	**3.15E-54**
BN_70	$\langle 2315,5,159,35 \rangle$	2.00E-71	6.77E-76	1.10E-75	9.50E-76	4.81E-75	3.22E-75	**2.81E-73**
BN_71	$\langle 1740,6,202,35 \rangle$	5.12E-111	1.89E-118	9.78E-114	4.27E-117	1.32E-113	**4.63E-112**	2.36E-112
BN_72	$\langle 2155,6,252,33 \rangle$	4.21E-150	7.35E-155	3.28E-153	5.49E-154	1.81E-150	**2.59E-150**	1.71E-150
BN_73	$\langle 2140,5,216,42 \rangle$	2.26E-113	1.17E-126	6.39E-122	5.33E-126	1.70E-118	**2.33E-117**	8.08E-118
BN_74	$\langle 749,6,66,32 \rangle$	3.75E-45	1.58E-47	2.13E-46	4.57E-47	2.00E-46	2.08E-46	**2.22E-46**
BN_75	$\langle 1820,5,155,32 \rangle$	5.88E-91	5.39E-97	1.09E-95	2.58E-98	3.19E-95	9.59E-95	**1.21E-91**
BN_76	$\langle 2155,7,169,37 \rangle$	4.93E-110	1.02E-121	6.93E-117	3.03E-119	1.23E-117	1.56E-115	**1.69E-112**
BN_77	$\langle 1020,9,135,22 \rangle$	6.88E-79	3.57E-87	3.46E-82	3.20E-86	2.42E-85	3.63E-84	**1.92E-81**
pedigree13	$\langle 1077,3,0,31 \rangle$	5.44E-32	1.83E-44	5.57E-35	1.83E-44	6.75E-34	3.99E-32	**4.03E-32**
pedigree34	$\langle 1160,3,0,28 \rangle$	5.89E-65	5.80E-78	1.24E-69	7.20E-74	8.19E-70	1.65E-64	**1.77E-64**
pedigree44	$\langle 811,3,0,25 \rangle$	3.36E-64	3.15E-66	2.00E-65	3.58E-66	2.30E-64	**3.21E-64**	**3.21E-64**
pedigree50	$\langle 514,3,0,17 \rangle$	1.32E-23	3.32E-24	3.44E-25	1.65E-24	1.14E-23	**1.38E-23**	**1.38E-23**
pedigree51	$\langle 1152,3,0,35 \rangle$	1.33E-74	9.95E-82	4.38E-78	1.24E-79	1.44E-76	6.75E-75	**6.86E-75**
pedigree7	$\langle 1068,4,0,30 \rangle$	1.5E-65	5.91E-71	7.73E-69	2.49E-72	2.42E-67	7.28E-66	**7.77E-66**
pedigree9	$\langle 1118,3,0,25 \rangle$	3.43E-79	2.12E-81	2.81E-85	2.56E-81	7.86E-79	**1.36E-79**	1.27E-79

Table 1: Table showing the sample means output by various schemes on 3 sets of benchmark graphical models. Each algorithm was run for 1 hr. The best results are highlighted by bold in each row.

On the pedigree linkage instances, we observe that the ao-wc-graph-IS, ao-wc-tree-IS and ao-graph-IS schemes are more accurate than the other schemes, often outperforming them by an order of magnitude. ao-tree-IS is better than or-tree-IS on most instances while or-wc-tree-IS is usually better than ao-tree-IS and or-tree-IS. ao-wc-graph-IS is the best performing scheme.

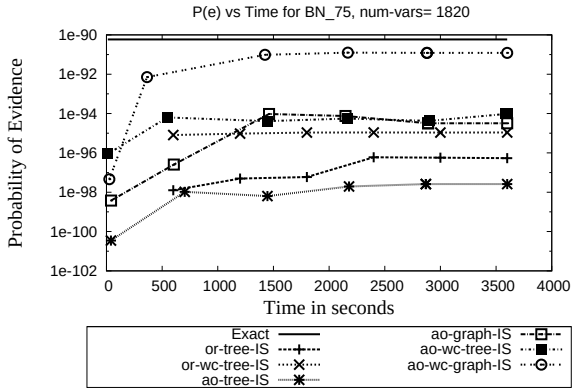

Figure 5: P(e) as a function of time for BN_75 instance.

Finally, Figure 5 shows how the probability of evidence $P(\mathbf{e})$ changes with time for one of Linkage instances. Here, we can clearly see the superior any-time performance of AND/OR-based *w*-cutset schemes over the other schemes. We observed similar behavior on other instances.

6.3 Results on instances for which the exact weighted counts are not known

When exact results are not available, evaluating the performance of approximate schemes is problematic because the quality of the approximation, namely how close the approximation is to the exact, cannot be measured. To allow a comparison on such hard instances we evaluate the power of the various sampling schemes for yielding good lower-bound approximations whose quality can be compared (the higher the better) even when the exact solution is not available. Specifically, when the exact weighted counts are not known, we compare the lower bounds obtained by combining the sample means output by various schemes with the Markov inequality based lower bounding scheme presented in (Gogate et al., 2007). Such lower bounding schemes, see also (Gomes et al., 2007), take as input: (a) a set of unbiased sample means and (b) a real number $0 < \alpha < 1$, and output a lower bound on the weighted counts that is correct with probability greater than α. Formally,

Theorem 7. *(Gomes et al., 2007; Gogate et al., 2007) Let $\widehat{Z}_1, \widehat{Z}_2, \ldots, \widehat{Z}_r$ be the unbiased sample means over "r" independent runs of a sampling scheme. Let $0 < \alpha < 1$ be a constant and let $\beta = \left(\frac{1}{1-\alpha}\right)^{\frac{1}{r}}$. Then, $Z_{lb} = \frac{1}{\beta} \times min_{i=1}^{r} \widehat{Z}_r$ is a lower bound on Z with probability greater than α.*

In our experiments, we set $\alpha = 0.99$ and $r = 5$, namely, we run each algorithm five times and each lower bound is correct with probability greater than 0.99.

4-Coloring Problems Our final domain is that of 4-coloring problems generated using Joseph Culberson's flat graph coloring generator (available at http://www.cs.ualberta.ca/∼joe/Coloring/). Here, the weighted counting task is equivalent to counting the number of solutions. From Table 2, we observe that in most cases, the AND/OR schemes are better than pure importance sampling (or-tree-IS) and *w*-cutset sampling

Problem-name	$\langle n,k,c,t^* \rangle$	Exact	or-tree-IS	or-wc-tree-IS	ao-tree-IS	ao-graph-IS	ao-wc-tree-IS	ao-wc-graph-IS
4-Coloring								
4-coloring1	$\langle 100,4,200,71 \rangle$		1.57E+37	1.98E+37	1.54E+37	2.18E+37	2.30E+37	**1.78E+38**
4-coloring2	$\langle 100,4,250,95 \rangle$		4.35E+27	9.86E+28	5.26E+29	6.00E+29	8.29E+28	**1.02E+30**
4-coloring3	$\langle 200,4,400,144 \rangle$		1.61E+70	4.63E+70	1.21E+72	1.76E+72	4.89E+70	**2.59E+72**
4-coloring4	$\langle 200,4,500,191 \rangle$		2.230E+62	4.82E+62	5.75E+63	1.21E+64	8.65E+64	**6.07E+65**
4-coloring5	$\langle 300,4,600,304 \rangle$		1.12E+97	1.21E+99	1.65E+100	1.61E+100	1.72E+102	**1.28E+104**
4-coloring6	$\langle 300,4,750,338 \rangle$		1.30E+88	9.01E+90	5.82E+88	1.01E+91	1.13E+91	**1.75E+91**

Table 2: Table showing the **lower bounds on the weighted counts with 99% confidence** obtained by various schemes for graph coloring benchmarks. The exact weighted counts for these instances are not known. Each algorithm was run 5 times, each run was 1 hr yielding 5 sample means. We use $\alpha = 0.99$, and combined these sample means using Theorem 7 to yield a lower bound on Z.

(or-wc-tree-IS). AND/OR *w*-cutset graph sampling yields the highest lower bound for all the instances.

7 Summary and Future work

The paper presents *AND/OR w-cutset sampling*, a general and unifying framework for developing and analyzing graph-based variance reduction schemes. Our generalization yields two new schemes called AND/OR w-cutset tree sampling and AND/OR w-cutset graph sampling, which have smaller variance than other schemes proposed in literature. Our experimental evaluation shows that our new schemes are often more accurate than other schemes, when all are given an identical time-bound and therefore they should be always preferred.

Several avenues remain for future work, such as: (a) designing good proposal distributions over the w-cutset variables, (b) developing variance reduction schemes that take advantage of local structure such as context specific independence (Boutilier et al., 1996) and combining them with AND/OR w-cutset sampling and (c) developing sequential versions of AND/OR w-cutset sampling.

Acknowledgements

This work was partially supported by the NSF grant IIS-0713118 and the NIH grant 5R01HG004175-03.

References

Bidyuk, B. and Dechter, R. (2004). On finding minimal w-cutset problem. In *UAI*, pages 43–50.

Bidyuk, B. and Dechter, R. (2007). Cutset sampling for Bayesian networks. *Journal of Artificial Intelligence Research*, 28:1–48.

Bilmes, J. and Dechter, R. (2006). Evaluation of Probabilistic Inference Systems of UAI'06. Available online at http://ssli.ee.washington.edu/ bilmes/uai06InferenceEvaluation/.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context-specific independence in Bayesian networks. In *UAI*, pages 115–123.

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

Cheng, J. and Druzdzel, M. J. (2000). AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research*, 13:155–188.

Dechter, R. (1990). Enhancement schemes for constraint processing: Backjumping, learning and cutset decomposition. *Artificial Intelligence*, 41:273–312.

Dechter, R. (1999). Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, 113:41–85.

Dechter, R., Kask, K., and Mateescu, R. (2002). Iterative Join Graph propagation. In *UAI*, pages 128–136.

Dechter, R. and Mateescu, R. (2007). AND/OR search spaces for graphical models. *Artificial Intelligence*, 171(2-3):73–106.

Fishelson, M. and Geiger, D. (2003). Optimizing exact genetic linkage computations. In *Proceedings of RECOMB*, pages 114–121.

Gogate, V., Bidyuk, B., and Dechter, R. (2007). Studies in lower bounding probability of evidence using the Markov inequality. In *UAI*, pages 141–148.

Gogate, V. and Dechter, R. (2005). Approximate inference algorithms for hybrid Bayesian networks with discrete constraints. In *UAI*, pages 209–216.

Gogate, V. and Dechter, R. (2007). Samplesearch: A scheme that searches for consistent samples. *AISTATS*, pages 147–154.

Gogate, V. and Dechter, R. (2008). AND/OR Importance Sampling. In *UAI*, pages 212–219.

Gomes, C. P., Hoffmann, J., Sabharwal, A., and Selman, B. (2007). From sampling to model counting. In *IJCAI*, pages 2293–2299.

Liu, J. (2001). *Monte-Carlo strategies in scientific computing*. Springer-Verlag, New York.

Mateescu, R. and Dechter, R. (2005). AND/OR Cutset Conditioning. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 230–235.

Mateescu, R., Kask, K., Gogate, V., and Dechter, R. (2010). Join-Graph Propagation Algorithms. *Journal of Artificial Intelligence Research*, 37:279–329.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. John Wiley & Sons Inc.

Sang, T., Beame, P., and Kautz, H. (2005). Performing Bayesian inference by weighted model counting. In *AAAI*, pages 475–481.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2004). Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312.

Yuan, C. and Druzdzel, M. J. (2006). Importance sampling algorithms for Bayesian networks: Principles and performance. *Mathematical and Computer Modelling*, 43(9-10):1189–1207.