Finite-sample Bounds for Marginal MAP

Qi Lou
University of California, Irvine
Irvine, CA 92697, USA
qlou@ics.uci.edu

Rina Dechter
University of California, Irvine
Irvine, CA 92697, USA
dechter@ics.uci.edu

Alexander Ihler
University of California, Irvine
Irvine, CA 92697, USA
ihler@ics.uci.edu

Abstract

Marginal MAP is a key task in Bayesian inference and decision-making, and known to be very challenging in general. In this paper, we present an algorithm that blends heuristic search and importance sampling to provide anytime finite-sample bounds for marginal MAP along with predicted MAP solutions. We convert bounding marginal MAP to a surrogate task of bounding a series of summation problems of an augmented graphical model, and then adapt dynamic importance sampling [Lou *et al.*, 2017b], a recent advance in bounding the partition function, to provide finite-sample bounds for the surrogate task. Those bounds are guaranteed to be tight given enough time, and the values of the predicted MAP solutions will converge to the optimum. Our algorithm runs in an anytime/anyspace manner, which gives flexible trade-offs between memory, time, and solution quality. We demonstrate the effectiveness of our approach empirically on multiple challenging benchmarks in comparison with some state-of-the-art search algorithms.

1 INTRODUCTION

Probabilistic graphical models, including Bayesian networks and Markov random fields, provide a framework for representing and reasoning with probabilistic and deterministic information [Dechter, 2013; Dechter *et al.*, 2010; Darwiche, 2009]. Typical inference queries in graphical models include *maximum a posteriori* (MAP) that aims to find an assignment of MAP (or MAX) variables with the highest value, the *partition function* that is the normalizing constant ensuring a proper probability measure over all variables, and marginal MAP (MMAP) that generalizes the aforementioned two tasks by maximizing over a subset of variables with the remaining variables marginalized, which arises in many scenarios such as latent variable models [Ping *et al.*, 2014] and decision-making tasks [Kiselev and Poupart, 2014].

MMAP has complexity NP^{PP} [Park, 2002], commonly believed to be more challenging than either max inference (NP-complete [Darwiche, 2009]) or sum inference (#P-hard [Valiant, 1979]), and can be intractable even for tree-structured models [Park, 2002]. Because of the inherent difficulty of MMAP, recent works on MMAP often focus on approximate schemes. Among these, approximations with deterministic or probabilistic guarantees are of particular interest because they quantify bounds on the approximation errors. We also prefer approaches with an anytime behavior because they allow users to trade off computational resources with solution quality.

Approaches that offer deterministic bounds are typically based on search or variational methods. Some early works [Park and Darwiche, 2003; Yuan and Hansen, 2009] in search solve MMAP exactly based on depth-first branch and bound. Marinescu *et al.* [2014] outperformed its predecessors by introducing AND/OR search spaces and high-quality variational heuristics; this was further improved using best-first search variants, including weighted heuristic search [Lee *et al.*, 2016b], and alternating depth-first and best-first AND/OR search (AAOBF [Marinescu *et al.*, 2017]). However, these methods typically require regular evaluation of internal summation problems when traversing the MAP space; when these internal sums are difficult, the search process may stall completely. One way to avoid this issue is to unify the summation with the MAP search in a single, best-first search framework (UBFS [Lou *et al.*, 2018]), which allows the bounds to improve as the summation is performed, and switch to other MAP configurations when appropriate. However, another promising approach is to make use of probabilistic bounds (e.g., Lou *et al.* [2017b]), which hold with a user-selected probability, and can be significantly faster

and tighter than deterministic bounds. However, since each MAP configuration is associated with an independent summation problem, comparing MAP configurations using probabilistic bounds must compensate for the presence of many uncertain tests (in effect, a multiple hypothesis testing problem), and is thus non-trivial to adapt to the MAP search, which may contain exponentially many such configurations.

Variational methods [Wainwright and Jordan, 2008] offer another class of deterministic bounds for MMAP. However, these bounds are often not anytime (e.g., [Liu and Ihler, 2013]), and those, such as [Ping *et al.*, 2015], are often not "any-space", meaning that their quality depends heavily on the available memory and may not continue to improve without more. Other types of algorithms can provide anytime deterministic bounds for MMAP as well, for example, one based on factor set elimination [Mauá and de Campos, 2012]; however, the factor sets that it maintains tend to grow very large, which limits its practical use to problems with relatively small induced widths (see Marinescu *et al.* [2017] or Lou *et al.* [2018]).

Some Monte Carlo approaches are able to provide probabilistic bounds; for example, Xue *et al.* [2016] proposes a random hashing based algorithm that provides a constant factor approximation. However, this approach can have difficulty on large scale problem instances (see [Lou *et al.*, 2018]). Other Monte Carlo methods may have no bound guarantees at all, e.g., those based on Markov chain Monte Carlo [Yuan *et al.*, 2004; Doucet *et al.*, 2002].

To some extent, the intrinsic hardness of MMAP arises from the non-commutativity of the sum and max operations. One natural idea to alleviate this issue is to convert the mixed inference task to a pure sum or a pure max one first. For example, Cheng *et al.* [2012] constructs an explicit factorized approximation of the marginalized distribution using a form of approximate variable elimination, which results in a structured MAP problem.

Our Contributions. In this paper, we present an approach that provides anytime finite-sample bounds (i.e., they hold with probability $1-\delta$ for some confidence parameter δ) for MMAP, that enjoys the benefits of both heuristic search and importance sampling. Briefly speaking, we follow Doucet *et al.* [2002] to construct an augmented graphical model from the original model by replicating the marginalized variables and potential functions. From this augmented model, we derive a sequence of decreasing summation objectives that bound the MMAP optimum raised to some fixed power. Then, we adapt dynamic importance sampling [Lou *et al.*, 2017b] to bound these summation objectives and provide finite-sample bounds of the MMAP optimum.

Our framework has several key advantages: 1) it provides anytime probabilistic upper and lower bounds that are guaranteed to be tight given enough time. 2) it is able to predict high-quality MAP solutions whose values converge to the optimum; the exploration-exploitation trade-off of searching MAP solutions is controlled by the number of replicates of the marginalized variables. 3) it runs in an anytime/anyspace manner, which gives flexible trade-offs between memory, time, and solution quality.

2 BACKGROUND

Let $X = (X_1, \ldots, X_N)$ be a vector of random variables, where each X_i takes values in a discrete domain $\mathcal{X}_i$; we use lower case letters, e.g. $x_i \in \mathcal{X}_i$, to indicate a value of X_i, and x to indicate an assignment of X. A graphical model over X consists of a set of factors $\mathcal{F} = \{f_\alpha(X_\alpha) \mid \alpha \in \mathcal{I}\}$, where each factor (a.k.a. potential function) f_α is defined on a subset $X_\alpha = \{X_i \mid i \in \alpha\}$ of X, called its scope.

We associate an undirected graph $\mathcal{G} = (V, E)$, or *primal graph*, with $\mathcal{F}$, where each node $i \in V$ corresponds to a variable X_i and we connect two nodes, $(i, j) \in E$, iff $\{i, j\} \subseteq \alpha$ for some α. Then,

$$f(x) = \prod_{\alpha \in \mathcal{I}} f_\alpha(x_\alpha)$$

defines an unnormalized probability measure over X.

Let X_{M} be a subset of X called MAX variables, and $X_{\mathrm{S}} = X \setminus X_{\mathrm{M}}$ SUM variables. The MMAP task seeks an assignment $x^\star_{\mathrm{M}}$ of X_{M} with the largest marginal probability:

$$x^\star_{\mathrm{M}} = \operatorname*{argmax}_{x_{\mathrm{M}}} \pi(x_{\mathrm{M}}) \tag{1}$$

where

$$\pi(x_{\mathrm{M}}) = \sum_{x_{\mathrm{S}}} f(x).$$

If X_{M} is an empty set, the MMAP task reduces to computing the normalizing constant (a.k.a. partition function); if X_{S} is empty, it becomes the standard MAP inference task. We use $\mathcal{X}_{\mathrm{M}}$ to denote the MAP space, i.e., the Cartesian product of all $\mathcal{X}_i$'s where X_i is a MAX variable. We will assume in the sequel that $x^\star_{\mathrm{M}}$ is unique for convenience, though our algorithm and analysis still hold without this assumption.

2.1 AND/OR Search Spaces

An AND/OR search space is a generalization of the standard ("OR") search space, that enables us to exploit conditional independence structure during search [Dechter and

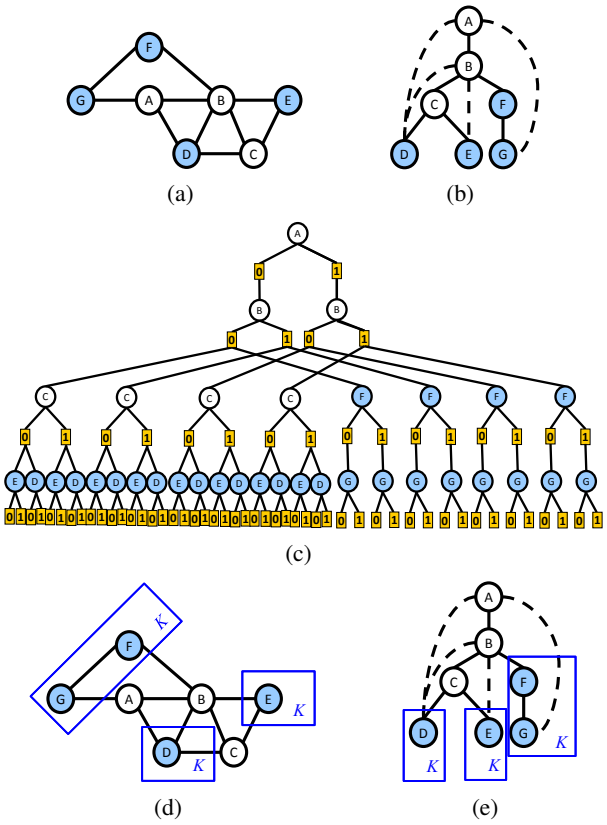


Figure 1: (a) A primal graph of a graphical model over 7 variables (A, B, C are MAX variables and D, E, F, G are SUM variables) with unary and pairwise potential functions. (b) A valid pseudo tree for the primal graph. (c) An AND/OR search tree guided by the pseudo tree. (d) An augmented model created by replicating SUM variables and factors of the model in (a). Plate notations used here. (e) A valid pseudo tree for the augmented model.

Mateescu, 2007]. The AND/OR search space for a graphical model is defined relative to a *pseudo tree* that captures problem decomposition along a fixed search order.

Definition 1 (pseudo tree). *A pseudo tree of a primal graph $\mathcal{G} = (V, E)$ is a directed tree $\mathcal{T} = (V, E')$ sharing the same set of nodes as $\mathcal{G}$. The tree edges E' form a subset of E, and each edge $(i, j) \in E \setminus E'$ are required to be a "back edge", i.e., the path from the root of $\mathcal{T}$ to j passes through i (denoted $i \leq j$).*

If a tree node of a pseudo tree corresponds to a MAX variable in the associated graphical model of the pseudo tree, we call it MAX node, otherwise we call it SUM node. A pseudo tree is called *valid* for an MMAP task if there is *no* MAX variable descended from any SUM variable. Thus, all MAX variables of a valid pseudo tree form a subtree (assuming a dummy MAX root) that contains the root. We assume valid pseudo trees in the sequel.

Guided by a pseudo tree, we can construct an AND/OR search tree consisting of alternating levels of OR and AND nodes for a graphical model. Each OR node s is associated with a variable, which we lightly abuse notation to denote X_s; the children of s, $ch(s)$, are AND nodes corresponding to the possible values of X_s. If an OR node is associated with some MAX variable, it is called OR-MAX node. Notions of OR-SUM, AND-MAX, AND-SUM nodes are defined analogously. The root $\emptyset$ of the AND/OR search tree corresponds to the root of the pseudo tree. Let $pa(c) = s$ indicate the parent of c in the AND/OR tree, and $an(c) = \{n \mid n \leq c\}$ indicate the ancestors of c (including itself) in the tree.

In an AND/OR tree, any AND node c corresponds to a partial configuration $x_{\leq c}$ of X, defined by its assignment and that of its ancestors: $x_{\leq c} = x_{\leq p} \cup \{X_s = x_c\}$, where $s = pa(c)$, $p = pa(s)$. For completeness, we also define $x_{\leq s}$ for any OR node s, which is the same as that of its AND parent, i.e., $x_{\leq s} = x_{\leq pa(s)}$. For any node n, the corresponding variables of $x_{\leq n}$ are denoted as $X_{\leq n}$. Let $de(X_n)$ be the set of variables below X_n in the pseudo tree; we define $X_{>n} = de(X_n)$ if n is an AND node; $X_{>n} = de(X_n) \cup \{X_n\}$ if n is an OR node.

We also associate a weight w_c with each AND node, defined to be the product of all factors f_α that are instantiated at c but not before:

$$w_c = \prod_{\alpha \in \mathcal{I}_c} f_\alpha(x_\alpha), \quad \mathcal{I}_c = \{\alpha \mid X_{pa(c)} \in X_\alpha \subseteq X_{an(c)}\}.$$

Example. Fig. 1(a) shows the primal graph of a pairwise model. Variables A, B, C are MAX variables, and the rest SUM. Fig. 1(b) shows one valid pseudo tree of the model. Fig. 1(c) shows the AND/OR search tree that respects the pseudo tree.

2.2 SEARCH IN AND/OR SEARCH TREES

Finally, the purpose of the search tree is to compute some inference quantity for the model, such as the MAP optimum $\max_x f(x)$, the partition function $Z = \sum_x f(x)$ and the MMAP optimum $\max_{x_M} \sum_{x_S} f(x)$. To this end, we associate a "value" v_n with each node n in the AND/OR search tree, which represents the inference task's value on the unexpanded portion of the search space below node n. The value v_n can be defined recursively in terms of its children and grandchildren as follows. We first define $v_l = 1$ for any leaf (since no part of the model remains uninstantiated). Let n be a non-leaf node; for maximization tasks, we have

$$\text{Max:} \quad v_n = \begin{cases} \prod_{c \in ch(n)} v_c, & \text{if AND node } n. \\ \max_{c \in ch(n)} w_c v_c, & \text{if OR node } n. \end{cases}$$

while for summation, the recursion defining v_n for n is

$$\text{Sum:} \quad v_n = \begin{cases} \prod_{c \in ch(n)} v_c, & \text{if AND node } n. \\ \sum_{c \in ch(n)} w_c v_c, & \text{if OR node } n. \end{cases}$$

For MMAP tasks, the recursion for AND nodes is the same as the aforementioned tasks, while the recursion for OR nodes is more involved:

$$\text{MMAP}: \quad v_n = \begin{cases} \max\limits_{c \in ch(n)} w_c v_c, & \text{if OR-MAX node } n. \\ \sum\limits_{c \in ch(n)} w_c v_c, & \text{if OR-SUM node } n. \end{cases}$$

Any search algorithm for reasoning about the model can be thought of as maintaining upper (and/or lower) bounds on these quantities at each node. In particular, for heuristic search, we assume that we have a heuristic function h_n that gives upper (or lower) bound on v_n. These heuristics typically are more accurate deeper in the search tree, and therefore their updates can be propagated upwards to the root to yield tighter bounds to the overall inference value. Any search algorithm is then defined by the order of expansion of the search tree.

A typical example of this kind of search algorithms is AOBFS [Lou *et al.*, 2017a], a best-first search algorithm that can provide anytime upper (and/or lower) bounds for the summation task. Since AOBFS will be a component of our proposed algorithm, we briefly present some of its essence here. AOBFS maintains an explicit AND/OR search tree of visited nodes, denoted $\mathcal{S}$. For each node n in the AND/OR search tree, AOBFS maintains u_n, an upper bound on v_n, initialized via a pre-compiled heuristic $v_n \leq h_n^+$, and subsequently updated during search using information propagated from the frontier:

$$u_n = \begin{cases} \prod_{c \in ch(n)} u_c, & \text{if AND node } n. \\ \sum_{c \in ch(n)} w_c u_c, & \text{if OR node } n. \end{cases}$$

Thus, the upper bound at the root, $u_\emptyset$, is an anytime deterministic upper bound of the partition function. Note that this upper bound depends on the current search tree $\mathcal{S}$, so we write $U^{\mathcal{S}} = u_\emptyset$.

If all nodes below n have been visited, then $u_n = v_n$; we call n *solved* and can remove the subtree below n from memory. Hence we can partition the frontier nodes into two sets: solved frontier nodes, *SOLVED*($\mathcal{S}$), and unsolved ones, *OPEN*($\mathcal{S}$).

2.3 DYNAMIC IMPORTANCE SAMPLING

Our work can be viewed as a generalization of dynamic importance sampling (DIS) [Lou *et al.*, 2017b], a recent advance in bounding the partition function with finite-sample bounds (see also [Liu *et al.*, 2015]), which we briefly introduce here to make our paper self-contained.

DIS interleaves search with sampling: search, as it improves the deterministic upper bound of the partition function by expanding nodes in the AND/OR search tree, also induces a sequence of importance sampling proposal distributions with bounded importance weights that are unbiased estimators of the partition function. Meanwhile, samples are drawn independently from those improving proposal distributions. By averaging those importance weights based on their corresponding upper bounds, DIS constructs an unbiased estimator of the partition function Z with strong probabilistic guarantees.

To be more specific, DIS applies AOBFS with its "upper priority" to quickly drive down the deterministic upper bound. The current search tree $\mathcal{S}$ induces a proposal distribution $q^{\mathcal{S}}$; importance weights $f(x)/q^{\mathcal{S}}(x)$ are bounded by $U^{\mathcal{S}}$ and give an unbiased estimator of Z:

$$f(x)/q^{\mathcal{S}}(x) \leq U^{\mathcal{S}}, \qquad \mathbb{E}\left[f(x)/q^{\mathcal{S}}(x)\right] = Z.$$

Drawing a sample from $q^{\mathcal{S}}$ can be described as a "two-step" top-down sampling process from the root:

Step 1 For an internal node $n \in \mathcal{S}$: if it is an AND node, all its children are selected; if n is an OR node, one child $c \in ch(n)$ is randomly selected with probability $w_c u_c / u_n$.

Step 2 When an unsolved frontier node $n \in$ *OPEN*($\mathcal{S}$) is reached, draw a sample of its descendant variables $X_{>n}$ in the pseudo tree according to the mixture proposal $q(x_{>n}|x_{\leq n})$ derived from *weighted mini-bucket* (WMB, [Liu and Ihler, 2011]).

DIS introduces an unbiased estimator $\widehat{Z}$ of Z:

$$\widehat{Z} = \frac{\mathrm{HM}(\boldsymbol{U})}{N} \sum_{i=1}^{N} \frac{\widehat{Z}_i}{U_i}, \quad \mathrm{HM}(\boldsymbol{U}) = \Big[\frac{1}{N}\sum_{i=1}^{N} \frac{1}{U_i}\Big]^{-1}.$$

where $\{\widehat{Z}_i = f(x^i)/q^{\mathcal{S}_i}(x^i)\}_{i=1}^N$ are importance weights from samples $\{x^i | x^i \sim q^{\mathcal{S}_i}(x)\}$ with $\{\mathcal{S}_i\}$ the corresponding search trees, and $\{U_i = U^{\mathcal{S}_i}\}_{i=1}^N$ the corresponding upper bounds on the importance weights respectively. By defining

$$\Delta = \mathrm{HM}(\boldsymbol{U})\Big[\sqrt{\frac{2\widehat{\mathrm{Var}}(\{\widehat{Z}_i/U_i\}_{i=1}^N)\ln(2/\delta)}{N}} + \frac{7\ln(2/\delta)}{3(N-1)}\Big]$$

where $\widehat{\mathrm{Var}}(\{\widehat{Z}_i/U_i\}_{i=1}^N)$ is the unbiased empirical variance of $\{\widehat{Z}_i/U_i\}_{i=1}^N$, $\widehat{Z}$ enjoys the finite-sample guarantees: with probability at least $1-\delta$, $\widehat{Z}+\Delta$ and $\widehat{Z}-\Delta$ are upper and lower bounds of Z, respectively, i.e., $\Pr[Z \leq \widehat{Z}+\Delta] \geq 1-\delta$ and $\Pr[Z \geq \widehat{Z}-\Delta] \geq 1-\delta$.

3 OUR ALGORITHM

In this section, we introduce our algorithm, the general idea of which is to first bound the mixed inference objective with a series of sum inference objectives whose finite-sample bounds can be established by generalizing DIS, and then translate the bounds back to those of the original objective in which we are interested.

3.1 AN AUGMENTED GRAPHICAL MODEL

We first introduce an augmented graphical model which connects the MMAP optimum to a series of summation tasks. The augmented graphical model is built from the original model by replicating the SUM variables and the factors. Note that the idea of introducing an augmented space on which we perform inference is adopted from Doucet *et al.* [2002].

Let $X_{\text{aug}} = (X_{\text{M}}, X_{\text{S}}^1, \ldots, X_{\text{S}}^K)$ be all the variables of the augmented model where $X_{\text{S}}^1, \ldots, X_{\text{S}}^K$ are K replicates of the SUM variables X_{S}. The overall function f_{aug} of the augmented model is defined as

$$f_{\text{aug}}(x_{\text{aug}}) = \prod_{k=1}^{K} f(x_{\text{M}}, x_{\text{S}}^k).$$

Thus, the partition function of the augmented model is

$$Z_{\text{aug}} = \sum_{x_{\text{M}}, x_{\text{S}}^1, \ldots, x_{\text{S}}^K} \prod_{k=1}^{K} f(x_{\text{M}}, x_{\text{S}}^k) = \sum_{x_{\text{M}}} \pi^K(x_{\text{M}}).$$

Considering that $\pi^K(x_{\text{M}}^\star)$ (see (1)) is the largest term in the sum on the r.h.s., we have

$$Z_{\text{aug}}/|\mathcal{X}_{\text{M}}| \leq \pi^K(x_{\text{M}}^\star) \leq Z_{\text{aug}}, \tag{2}$$

that is to say,

$$(Z_{\text{aug}}/|\mathcal{X}_{\text{M}}|)^{1/K} \leq \pi(x_{\text{M}}^\star) \leq Z_{\text{aug}}^{1/K},$$

where $|\mathcal{X}_{\text{M}}|$ is the size of the MAP space. The above inequalities are actually well-known boundedness relations between the ∞-norm and p-norms of the Euclidean space $\mathbb{R}^{|\mathcal{X}_{\text{M}}|}$. These bounds are monotonic in K, i.e., they improve as K increases, and become tight as K goes to infinity. In other words, K acts as a "reverse temperature" parameter. The lower bound is negatively impacted by the domain sizes of the MAX variables, which can be quite loose if $|\mathcal{X}_{\text{M}}|$ is large compared to the scale of K.

The significance of (2) is that it connects the MMAP optimum to a summation quantity Z_{aug} that can be easily approximated using Monte Carlo methods such as importance sampling.

Example. Fig. 1(d) shows an augmented graphical model created from the model of Fig. 1(a). Fig. 1(e) shows one valid pseudo tree for the augmented model.

3.2 MIXED DYNAMIC IMPORTANCE SAMPLING

A straightforward idea is to apply DIS to bound Z_{aug} whose finite-sample bounds can then be translated to those of $\pi(x_{\text{M}}^\star)$. However, several key issues remain to be addressed for this idea to work well.

The first issue is about how to adapt DIS to the augmented model in an efficient manner. Since the augmented model might have many more variables compared to the original model, a naïve construction of AND/OR trees leads to an excessively large search space. Note that any X_{S}^k in the augmented model is an identical copy of X_{S}; we thus do not necessarily distinguish those X_{S} copies during search. That is to say, when search instantiates those factors involving SUM variables, it behaves as usual but takes into account the effect of replication when using information propagated from SUM nodes. We can also apply an analogous idea to construct weighted mini-bucket (WMB) [Liu and Ihler, 2011] heuristics to ensure that they are still compatible with the new search process. In a nutshell, search for the augmented model can enjoy the same complexity as that for the original model.

Meanwhile, the proposal distribution $q_{\text{aug}}^{\mathcal{S}}(x_{\text{aug}})$ associated with a search tree $\mathcal{S}$ has a decomposition property:

$$q_{\text{aug}}^{\mathcal{S}}(x_{\text{aug}}) = q_{\text{aug}}^{\mathcal{S}}(x_{\text{M}}) \prod_{k=1}^{K} q_{\text{aug}}^{\mathcal{S}}(x_{\text{S}}^k | x_{\text{M}}), \tag{3}$$

with $q_{\text{aug}}^{\mathcal{S}}(x_{\text{S}}^k | x_{\text{M}})$ are identical conditional distributions. Its importance weights also share the boundedness property:

$$f_{\text{aug}}(x_{\text{aug}})/q_{\text{aug}}^{\mathcal{S}}(x_{\text{aug}}) \leq U_{\text{aug}}^{\mathcal{S}},$$

where $U_{\text{aug}}^{\mathcal{S}}$ is the upper bound associated with $\mathcal{S}$. Note that sampling from $q_{\text{aug}}^{\mathcal{S}}$ can also be done via a two-step sampling procedure analogous to that in DIS.

One point worth mentioning is that

$$\pi(x_{\text{M}}) = \mathbb{E}[f(x_{\text{M}}, x_{\text{S}}^k)/q_{\text{aug}}^{\mathcal{S}}(x_{\text{S}}^k | x_{\text{M}})]$$

implies that we can estimate the value of each sampled MAP configuration x_{M} along the way.

Another issue is that if Z_{aug} is much larger than $\pi^K(x_{\text{M}}^\star)$, even high-quality bounds of Z_{aug} might not result in reasonably good bounds of $\pi(x_{\text{M}}^\star)$, let alone those bounds will never be tight in general for $\pi(x_{\text{M}}^\star)$ with a finite K. One way to alleviate this issue is based on the following key observation: for any subset $\mathcal{A}$ of $\mathcal{X}_{\text{M}}$ that contains $x_{\text{M}}^\star$, we have

$$Z_{\text{aug}}^{\mathcal{A}}/|\mathcal{A}| \leq \pi^K(x_{\text{M}}^\star) \leq Z_{\text{aug}}^{\mathcal{A}}, \tag{4}$$

Algorithm 1 Mixed Dynamic Importance Sampling

Require: Control parameters K, N_d, N_l; confidence parameter δ; memory budget, time budget.
Ensure: $\widehat{Z}_{\text{aug}}$, Δ, $\text{HM}(\boldsymbol{U})$, $\text{HM}(\boldsymbol{U}/|\boldsymbol{\mathcal{A}}|)$.

```
Construct WMB heuristics for the augmented model.
Initialize S ← {∅} with the root ∅.
while within the time budget
                    // update S, U^S, A^S during search.
    if within the memory budget
        Expand N_d nodes via AOBFS (Alg. 1 of Lou et al. [2017a]) with its "upper priority").
    else
        Expand N_d nodes via depth-first search.
    end if
    Draw N_l samples from q^S_aug (see (3)).
    After drawing each sample:
        Update N, Ẑ_aug, HM(U), HM(U/|A|), Var-hat, Δ via (5), (6), (11), (12).
end while
```

where

$$Z^{\mathcal{A}}_{\text{aug}} = \sum_{x_{\text{M}} \in \mathcal{A}} \pi^K(x_{\text{M}}).$$

The above inequalities tell us that if we know an instantiation of X_{M} is not optimal, we can mute its contribution to Z_{aug} and use the resulting smaller summation quantity to bound $\pi^K(x^\star_{\text{M}})$.

This observation enables pruning during search: any node ruled out from being associated with the optimal configuration can be removed from memory. Such pruning is particularly useful to prune MAX nodes: for any AND-MAX node with its sub-problem beneath solved, if it holds the highest value among its siblings, all its siblings (solved or not) and their descendants can be pruned immediately. Thus, as pruning proceeds along with search, $\mathcal{A}$ shrinks towards $\{x^\star_{\text{M}}\}$. We use $\mathcal{A}^{\mathcal{S}}$ to denote the remaining MAP space associated with the search tree $\mathcal{S}$.

Note that when we approach the memory limit, we switch the default best-first search to a depth-first search (DFS) that is also compatible with the sampling procedure, and leads to a complete search algorithm with the capability to identify $x^\star_{\text{M}}$ and its value given enough time. By interleaving search and sampling, we derive our algorithm named *mixed dynamic importance sampling* (MDIS) and present it in Alg. 1.

Remarks on Alg. 1.

1) K as the number of replicates of the SUM variables controls the exploration-exploitation trade-off. When K is small, we draw a small number of samples for the SUM variables in each iteration, which allows us to evaluate each sampled MAP configuration fast, however introduces more randomness when assessing the MAP configuration; when K is large, we have more accurate estimate of a MAP configuration being sampled, but also slow down exploration of the MAP space.

2) To predict MAP solutions in an anytime manner, one can simply choose the one with the highest estimated value among those configurations that have been sampled.

3.2.1 Finite-sample Bounds for Marginal MAP

In MDIS, each sample not only comes from a different proposal distribution but also gives importance weights corresponding to a different expectation, which is more complicated than in DIS.

Let $\{x^i_{\text{aug}}\}^N_{i=1}$ be a series of samples drawn via Alg. 1, with $\{\mathcal{S}_i\}$ the corresponding search trees, $\{\widehat{Z}^i_{\text{aug}} = f_{\text{aug}}(x^i_{\text{aug}})/q^{\mathcal{S}_i}_{\text{aug}}(x^i_{\text{aug}})\}^N_{i=1}$ the corresponding importance weights, and $\{U_i = U^{\mathcal{S}_i}_{\text{aug}}\}^N_{i=1}$ the corresponding upper bounds associated with those search trees respectively. We denote $\mathcal{A}_i$ as the MAP space preserved in $\mathcal{S}_i$. Thus,

$$\mathbb{E}\left[\widehat{Z}^i_{\text{aug}}\right] = Z^{\mathcal{A}_i}_{\text{aug}}.$$

That is to say, the importance weights have different (in fact, decreasing) expectations; this differs from the case of DIS where any importance weight has the same expectation (the partition function). We propose an estimate $\widehat{Z}_{\text{aug}}$ whose expectation is again an upper bound of $\pi^K(x^\star_{\text{M}})$ in the following way:

$$\widehat{Z}_{\text{aug}} = \frac{\text{HM}(\boldsymbol{U})}{N} \sum_{i=1}^{N} \frac{\widehat{Z}^i_{\text{aug}}}{U_i}, \tag{5}$$

where

$$\text{HM}(\boldsymbol{U}) = \Big[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{U_i}\Big]^{-1} \tag{6}$$

is the harmonic mean of the upper bounds $\{U_i\}^N_{i=1}$. Thus, $\widehat{Z}_{\text{aug}}$ upweights the terms $\widehat{Z}^i_{\text{aug}}$ whose expectations are closer to $\pi^K(x^\star_{\text{M}})$. The expectation of $\widehat{Z}_{\text{aug}}$ is

$$\mathbb{E}\left[\widehat{Z}_{\text{aug}}\right] = \frac{\text{HM}(\boldsymbol{U})}{N} \sum_{i=1}^{N} \frac{Z^{\mathcal{A}_i}_{\text{aug}}}{U_i}.$$

$\mathbb{E}\left[\widehat{Z}_{\text{aug}}\right]$ is a convex combination of $\{Z^{\mathcal{A}_i}_{\text{aug}}\}^N_{i=1}$ with coefficients $\{\frac{\text{HM}(\boldsymbol{U})}{NU_i}\}^N_{i=1}$, shrinking towards $\pi^K(x^\star_{\text{M}})$ as search proceeds.

According to (4), since $\pi^K(x^\star_{\text{M}}) \le Z^{\mathcal{A}_i}_{\text{aug}}$, we know

$$\pi^K(x^\star_{\text{M}}) \le \mathbb{E}\left[\widehat{Z}_{\text{aug}}\right], \tag{7}$$

and from $Z_{\text{aug}}^{\mathcal{A}_i} \leq |\mathcal{A}_i|\pi^K(x_{\text{M}}^\star)$, we know

$$\mathbb{E}\left[\widehat{Z}_{\text{aug}}\right] \leq \pi^K(x_{\text{M}}^\star)\frac{\text{HM}(\boldsymbol{U})}{N}\sum_{i=1}^{N}\frac{|\mathcal{A}_i|}{U_i}. \quad (8)$$

By combining (6), (7), and (8), we derive two-sided bounds for $\pi^K(x_{\text{M}}^\star)$ involving $\mathbb{E}\left[\widehat{Z}_{\text{aug}}\right]$:

$$\frac{\sum_{i=1}^{N} 1/U_i}{\sum_{i=1}^{N}|\mathcal{A}_i|/U_i}\,\mathbb{E}\left[\widehat{Z}_{\text{aug}}\right] \leq \pi^K(x_{\text{M}}^\star) \leq \mathbb{E}\left[\widehat{Z}_{\text{aug}}\right]. \quad (9)$$

From the above, we can see that the bounds get tight only when $\mathcal{A}_i$ approaches $\{x_{\text{M}}^\star\}$. To be concise, we re-arrange the L.H.S. of (9) to derive:

$$\frac{\text{HM}(\boldsymbol{U}/|\boldsymbol{\mathcal{A}}|)}{\text{HM}(\boldsymbol{U})}\,\mathbb{E}\left[\widehat{Z}_{\text{aug}}\right] \leq \pi^K(x_{\text{M}}^\star) \leq \mathbb{E}\left[\widehat{Z}_{\text{aug}}\right], \quad (10)$$

where

$$\text{HM}(\boldsymbol{U}/|\boldsymbol{\mathcal{A}}|) = \Big[\frac{1}{N}\sum_{i=1}^{N}\frac{|\mathcal{A}_i|}{U_i}\Big]^{-1} \quad (11)$$

is the harmonic mean of $\{U_i/|\mathcal{A}_i|\}_{i=1}^N$.

Considering $\widehat{Z}_{\text{aug}}^i$ are independent, and $\mathbb{E}\,\widehat{Z}_{\text{aug}}/\,\text{HM}(\boldsymbol{U})$, $\widehat{Z}_{\text{aug}}/\,\text{HM}(\boldsymbol{U})$, $\widehat{Z}_{\text{aug}}^i/U_i$ are all within the interval $[0,1]$, we can apply an empirical Bernstein bound [Maurer and Pontil, 2009] to derive finite-sample bounds on $\mathbb{E}\,\widehat{Z}_{\text{aug}}$ and translate those bounds to $\pi(x_{\text{M}}^\star)$ based on (10).

Theorem 1. For any $\delta \in (0,1)$, we define

$$\Delta = \text{HM}(\boldsymbol{U})\Big(\sqrt{\frac{2\widehat{\text{Var}}\ln(2/\delta)}{N}} + \frac{7\ln(2/\delta)}{3(N-1)}\Big), \quad (12)$$

where $\widehat{\text{Var}}$ is the unbiased empirical variance of $\{\widehat{Z}_{\text{aug}}^i/U_i\}_{i=1}^N$. Then, the following probabilistic bounds hold for $\pi(x_{\text{M}}^\star)$:

$$\Pr\left[\pi(x_{\text{M}}^\star) \leq (\widehat{Z}_{\text{aug}}+\Delta)^{\frac{1}{K}}\right] \geq 1-\delta,$$
$$\Pr\left[\pi(x_{\text{M}}^\star) \geq \Big(\frac{(\widehat{Z}_{\text{aug}}-\Delta)\,\text{HM}(\boldsymbol{U}/|\boldsymbol{\mathcal{A}}|)}{\text{HM}(\boldsymbol{U})}\Big)^{\frac{1}{K}}\right] \geq 1-\delta,$$

i.e., $(\widehat{Z}_{\text{aug}}+\Delta)^{\frac{1}{K}}$ and $\big(\frac{(\widehat{Z}_{\text{aug}}-\Delta)\,\text{HM}(\boldsymbol{U}/|\boldsymbol{\mathcal{A}}|)}{\text{HM}(\boldsymbol{U})}\big)^{\frac{1}{K}}$ are upper and lower bounds of $\pi(x_{\text{M}}^\star)$ with probability at least $1-\delta$, respectively.

Note that it is possible that $\widehat{Z}_{\text{aug}} - \Delta < 0$ early on; if so, we may replace $\widehat{Z}_{\text{aug}} - \Delta$ with any non-trivial lower bound of Z_{aug}. In the experiments, we use $\delta\widehat{Z}_{\text{aug}}$, a $(1-\delta)$ probabilistic bound by the Markov inequality [Gogate and Dechter, 2011]. We can replace $\widehat{Z}_{\text{aug}}+\Delta$ with the current deterministic upper bound if the latter is tighter.

4 EXPERIMENTS

We evaluate our proposed approach (MDIS) against two baseline methods on five benchmarks. The baselines include UBFS [Lou *et al.*, 2018], a unified best-first search algorithm that emphasizes rapidly tightening the upper bound, and AAOBF [Marinescu *et al.*, 2017], a best-first/depth-first hybrid search algorithm that balances upper bound quality with generating and evaluating potential solutions. These two are state-of-the-art algorithms for anytime upper and lower bounds respectively. We do not compare to XOR_MMAP [Xue *et al.*, 2016] and AFSE [Mauá and de Campos, 2012] due to their limitations to relatively easy problem instances as shown in Lou *et al.* [2018].

Three benchmarks are formed by problem instances from recent UAI competitions: `grid`- 50 grid networks with size no smaller than 25 by 25, `promedas`- 50 medical diagnosis expert systems, `protein`- 44 instances made from the "small" protein side-chains of [Yanover and Weiss, 2002]. Since the original UAI instances are pure MAP tasks, we generate MMAP instances by randomly selecting 10% of the variables as MAP variables. The fourth benchmark is `planning`, formed by 15 instances from probabilistic conformant planning with a finite-time horizon [Lee *et al.*, 2016a]. On these four benchmarks, we compare anytime bounds. Some statistics of the four benchmarks are shown in Table 1. These benchmarks are selected to illustrate different problem characteristics; for example, `protein` instances are relatively small but high cardinality, while `planning` instances have more variables and higher induced width, but lower cardinality. The fifth benchmark, which we will describe in detail later, is created from an image denoising model in order to evaluate quality of the predicted MAP solutions.

The time budget is set to 1 hour for the experiments on the first four benchmarks. We allot 4GB memory to all algorithms, with 1GB extra memory to AAOBF for caching. For our experiments, we use the weighted mini-bucket [Liu and Ihler, 2011] heuristics, whose memory usage is roughly controlled by an *ibound* parameter. For a given memory budget, we first compute the largest ibound that fits in memory, then use the remaining memory for search. Since all the competing algorithms use weighted mini-bucket heuristics, the same ibound is shared during heuristic construction. We set $N_d = 100$ and $N_l = 1$ (see Alg. 1) as suggested by the experimental results in Lou *et al.* [2017b]. We set $\delta = 0.025$. All implementations are in C/C++ courtesy of the original authors.

Anytime bounds for individual instances. Fig. 2 shows the anytime behavior of all the methods on instances from four benchmarks. In terms of lower bounds,

Table 1: Statistics of the four evaluated benchmarks. The first three benchmarks are formed by problem instances from recent UAI competitions, where 10% of variables are randomly selected as MAX variables. "avg. ind. width of sum" in the last row stands for the average induced width of the internal summation problems.

	grid	promedas	protein	planning
# instances	50	50	44	15
avg. # variables	1248.20	982.10	109.55	1122.33
avg. % of MAX vars	10%	10%	10%	12%
avg. # of factors	1248.20	994.76	394.64	1127.67
avg. max domain size	2.00	2.00	81.00	3.00
avg. max scope	3.00	3.00	2.00	5.00
avg. induced width	124.82	108.14	15.84	165.00
avg. pseudo tree depth	228.92	158.78	33.52	799.33
avg. ind. width of sum	43.44	40.32	10.20	49.67

Table 2: Number of instances that an algorithm achieves the best *lower bounds* at each timestamp (1 min, 10 min, and 1 hour) for each benchmark. The best for each setting is bolded. Entries for UBFS are blank because UBFS does not provide lower bounds.

	grid	promedas	protein	planning
# instances	50	50	44	15
Timestamp: 1min/10min/1hr				
MDIS (K=5)	**47/44/45**	**32/34/31**	**31/27/28**	**14/13/13**
MDIS (K=10)	3/2/1	4/5/6	11/13/14	1/2/2
UBFS	-/-/-	-/-/-	-/-/-	-/-/-
AAOBF	0/4/4	16/21/24	2/4/4	0/0/0

Table 3: Number of instances that an algorithm achieves the best *upper bounds* at each timestamp (1 min, 10 min, and 1 hour) for each benchmark. The best for each setting is bolded.

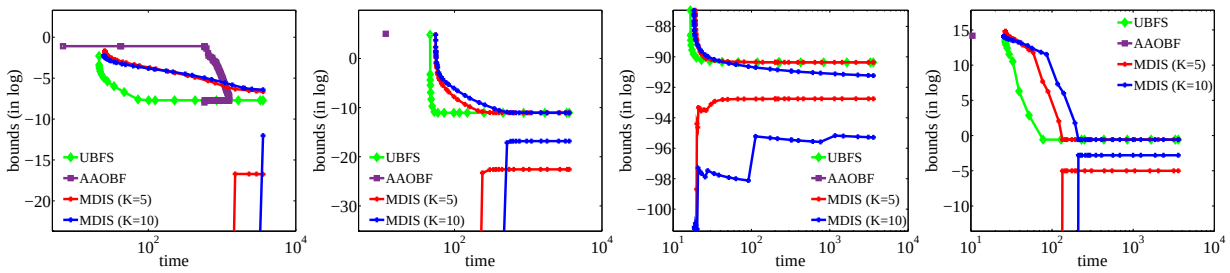
	grid	promedas	protein	planning
# instances	50	50	44	15
Timestamp: 1min/10min/1hr				
MDIS (K=5)	0/0/0	9/12/13	5/9/15	1/1/1
MDIS (K=10)	0/0/0	10/13/14	9/10/13	1/2/3
UBFS	**50/50/50**	**50/50/50**	**36/32/26**	**14/14/13**
AAOBF	0/0/1	2/4/6	2/2/2	1/1/1

our approach can always provide decent lower bounds even when the internal summation problems are quite challenging, while AAOBF may not work well since it relies on exact evaluation of those internal summation problems, e.g., on those shown in Fig. 2(b)-2(d). When the internal summation problems are relatively easy, their exact evaluation is cheap; thus AAOBF might perform better than ours. Fig. 2(a) gives a typical example. In terms of upper bounds, our bound quality is often eventually comparable to UBFS, e.g., Fig. 2(b)-2(d). UBFS typically performs better than MDIS early on, while MDIS quickly catches up and becomes comparable. Improvement in AAOBF on upper bounds also requires fast exact evaluation of the internal summation problems, which might not be possible in many cases. So, AAOBF is usually not as competitive as the other two methods on upper bounds.

Anytime bounds across benchmarks. We present the anytime performance across the four benchmarks in Table 2 and 3 where we compare anytime bounds at three different timestamps: 1 minute, 10 minutes and 1 hour. From Table 2, we can observe that MDIS with K=5 is dominant at any of these timestamp/benchmark combinations for lower bounds. MDIS with K=10 performs less well, perhaps because it requires more time to draw one full sample compared to when K=5, leading the empirical Bernstein lower bounds to kick in relatively late; this phenomenon can be also observed in all the plots in Fig. 2. UBFS provides the best upper bounds as shown in Table 3. However, our algorithm generally performs better than AAOBF in terms of upper bounds.

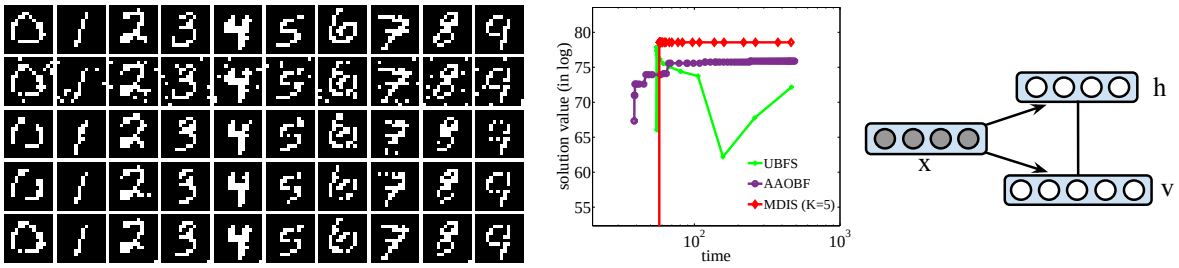
Empirical evaluation of solution quality. To evaluate the MAP solution quality predicted by our algorithm, we create an image denoising task from the MNIST database[1] of handwritten digits [LeCun *et al.*, 1998]. We binarize each image, resize it to 14 by 14, and then randomly flip 5% of the pixels to generate a corrupt one. We train a conditional restricted Boltzmann machine (CRBM) [Mnih *et al.*, 2011] model with 64 hidden units and 196 visible units using mixed-product BP [Ping and Ihler, 2017; Liu and Ihler, 2013] for the denoising task. The resulting graphical model thus has 64 SUM variables and 196 MAX variables. Fig. 3(c) gives an illustration of this model. The advantage of this model is that we can easily evaluate any MAP configuration since the internal summation problem only contains singleton potentials; thus this model favors AAOBF since AAOBF is able to evaluate MAP configurations at a very low cost. We set K to 5 and runtime to 10 minutes for convenience. We test on 100 images with 10 images per digit. Fig. 3(a) compares the denoising results among all the algorithms for one instance per digit. Fig. 3(b) gives an example of the quality of the predicted MAP solutions of our algorithm. In general, the quality of predicted MAP solutions for our algorithm are better than the other two baselines in 51 of 100 instances, which is generally as good as AAOBF (47/100) despite the model being well-suited to AAOBF. A possible reason is that our algorithm is able to traverse the MAP space very quickly and get cheap stochastic estimates of the most promising MAP solutions.

[1] http://yann.lecun.com/exdb/mnist/



(a) `grid`/90-25-5 (b) `promedas`/or_chain_48.fg (c) `protein`/1skz (d) `planning`/bw2_2_4_4

Figure 2: Anytime bounds for MMAP on instances from four benchmarks. The max domain sizes of those instances from (a)-(d) are 2, 2, 81, 3 respectively, and the induced widths of the internal summation problems are 25, 28, 8, 24 respectively. Curves for some bounds may be (partially) missing because they are not in a reasonable scope. UBFS only provides upper bounds. The time limit is 1 hour.



(a) (b) (c) CRBM

Figure 3: (a) Image denoising results for one instance per digit. The first row is for the ground truth images. The second row is for the noisy inputs created from the ground truth by randomly flipping 5% pixels. Below the first two rows are denoised images from UBFS, AAOBF, MDIS (K=5) respectively. (b) An example on MAP solution quality comparison. (c) Illustration of the conditional restricted Boltzmann machine (CRBM) model used for the image denoising task. When conditioned on an input "X", this model has a bipartite graph structure between hidden units "h" (SUM variables) and visible units "v" (MAX variables).

5 CONCLUSION

In this paper, we propose an approach that provides anytime finite-sample upper and lower bounds for MMAP, which enjoys the merits of both heuristic search and importance sampling. Our approach is particularly useful for problem instances whose internal summation problems are challenging. It predicts high-quality MAP solutions along with their estimated values. It runs in an anytime/anyspace manner, which gives flexible trade-offs between memory, time, and solution quality.

Acknowledgements

We thank all the reviewers for their helpful feedback. We also thank Wei Ping for assistance with the experiments.

This work is sponsored in part by NSF grants IIS-1526842 and IIS-1254071, the U.S. Air Force (Contract FA9453-16-C-0508), and DARPA (Contract W911NF-18-C-0015).

References

Qiang Cheng, Feng Chen, Jianwu Dong, Wenli Xu, and Alexander Ihler. Approximating the sum operation for marginal-MAP inference. In *AAAI*, pages 1882–1887, 2012.

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

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

Rina Dechter, Hector Geffner, and Joseph Y Halpern. *Heuristics, Probability and Causality. A Tribute to Judea Pearl*. College Publications, 2010.

Rina Dechter. Reasoning with probabilistic and deterministic graphical models: Exact algorithms. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 7(3):1–191, 2013.

Arnaud Doucet, Simon J. Godsill, and Christian P. Robert. Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12(1):77–84, 2002.

Vibhav Gogate and Rina Dechter. Sampling-based lower bounds

for counting queries. *Intelligenza Artificiale*, 5(2):171–188, 2011.

Igor Kiselev and Pascal Poupart. Policy optimization by marginal-MAP probabilistic inference in generative models. In *AAMAS*, pages 1611–1612, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Junkyu Lee, Radu Marinescu, and Rina Dechter. Applying search based probabilistic inference algorithms to probabilistic conformant planning: Preliminary results. In *ISAIM*, 2016.

Junkyu Lee, Radu Marinescu, Rina Dechter, and Alexander Ihler. From exact to anytime solutions for marginal MAP. In *AAAI*, pages 3255–3262, 2016.

Qiang Liu and Alexander Ihler. Bounding the partition function using Hölder's inequality. In *ICML*, pages 849–856, 2011.

Qiang Liu and Alexander Ihler. Variational algorithms for marginal MAP. *Journal of Machine Learning Research*, 14(1):3165–3200, 2013.

Qiang Liu, John W. Fisher, III, and Alexander Ihler. Probabilistic variational bounds for graphical models. In *NIPS*, pages 1432–1440, 2015.

Qi Lou, Rina Dechter, and Alexander Ihler. Anytime anyspace AND/OR search for bounding the partition function. In *AAAI*, pages 860–867, 2017.

Qi Lou, Rina Dechter, and Alexander Ihler. Dynamic importance sampling for anytime bounds of the partition function. In *NIPS*, pages 3198–3206, 2017.

Qi Lou, Rina Dechter, and Alexander Ihler. Anytime anyspace and/or best-first search for bounding marginal MAP. In *AAAI*, 2018.

Radu Marinescu, Rina Dechter, and Alexander Ihler. AND/OR search for marginal MAP. In *UAI*, pages 563–572, 2014.

Radu Marinescu, Junkyu Lee, Alexander Ihler, and Rina Dechter. Anytime best+ depth-first search for bounding marginal MAP. In *AAAI*, pages 3775–3782, 2017.

Denis Mauá and Cassio de Campos. Anytime marginal maximum a posteriori inference. In *ICML*, pages 1395–1402, 2012.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *COLT*, 2009.

Volodymyr Mnih, Hugo Larochelle, and Geoffrey Hinton. Conditional restricted Boltzmann machines for structured output prediction. In *UAI*, pages 514–522, 2011.

James Park and Adnan Darwiche. Solving MAP exactly using systematic search. In *UAI*, pages 459–468, 2003.

James Park. MAP complexity results and approximation methods. In *UAI*, pages 388–396, 2002.

Wei Ping and Alex Ihler. Belief propagation in conditional RBMs for structured prediction. In *AISTATS*, pages 1141–1149, 2017.

Wei Ping, Qiang Liu, and Alexander Ihler. Marginal structured SVM with hidden variables. In *ICML*, pages 190–198, 2014.

Wei Ping, Qiang Liu, and Alexander Ihler. Decomposition bounds for marginal MAP. In *NIPS*, pages 3267–3275, 2015.

L.G. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189 – 201, 1979.

M.J. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Yexiang Xue, Zhiyuan Li, Stefano Ermon, Carla P. Gomes, and Bart Selman. Solving marginal MAP problems with NP oracles and parity constraints. In *NIPS*, pages 1127–1135, 2016.

Chen Yanover and Yair Weiss. Approximate inference and protein-folding. In *NIPS*, pages 1457–1464, 2002.

Changhe Yuan and Eric A. Hansen. Efficient computation of jointree bounds for systematic MAP search. In *IJCAI*, pages 1982–1989, 2009.

Changhe Yuan, Tsai-Ching Lu, and Marek J. Druzdzel. Annealed MAP. In *UAI*, pages 628–635, 2004.