POMDP Planning by Marginal-MAP Probabilistic Inference in Generative Models

Igor Kiselev and Pascal Poupart
David R. Cheriton School of Computer Science, University of Waterloo
200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada
{ipkiselev, ppoupart}@uwaterloo.ca

ABSTRACT

While most current POMDP planning methods have focused on the development of scalable approximate algorithms, they often neglect the important aspect of solution quality and sacrifice performance guarantees to improve efficiency. In contrast, we propose a novel approach to optimize POMDP controllers by probabilistic inference while obtaining bounds on solution quality as follows: 1) re-formulate the original POMDP problem as a task of marginal-MAP (maximum a posteriori) inference in a novel single-$\mathcal{DBN}$ model, 2) define a dual representation of the marginal-MAP problem and derive a Bayesian variational approximation framework to obtain the approximate solution and an upper bound, and 3) design hybrid message-passing algorithms to solve a POMDP problem by approximate variational marginal-MAP inference in the equivalent single-$\mathcal{DBN}$ model.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent agents, Multiagent systems*

General Terms

Algorithms, Theory, Performance, Experimentation

Keywords

Planning under uncertainty, Probabilistic Inference, POMDPs

1. INTRODUCTION

Partially observable Markov decision processes (POMDPs) provide a natural framework for planning under uncertainty. Recent work by Toussaint et al. [15, 16] showed that it is possible to optimize controllers by maximizing the likelihood of rewards in a certain equivalent inference problem. This approach was subsequently extended and generalized to continuous [5], average reward [11], hierarchical [14], reinforcement learning [17] and multi-agent [7] domains. So far, planning as inference relies on the conversion of a POMDP model into a mixture of dynamic Bayesian networks ($\mathcal{DBN}$s). This mixture of $\mathcal{DBN}$s presents some hurdles in practice since most inference techniques are designed do work on a single graphical model and therefore must be adapted to work with mixtures of graphical models. Our first contribution is a new technique to convert POMDPs into a single $\mathcal{DBN}$ in which maximizing the likelihood of a random value variable is equivalent to optimizing a POMDP controller. The simplification to a single graphical model opens the door to a wider range of inference techniques. Our second contribution is the formulation of a Marginal-MAP (Maximum A Posteriori) inference problem for policy optimization. In the novel single-$\mathcal{DBN}$ model, we maximize over the policy variables while marginalizing the remaining variables. We solve this mixed ("max-sum") inference problem by a proposed message passing. Our third contribution is the formulation of message-passing rules for Marginal-MAP inference in general cyclic factor graphs, which extends [8]'s message passing rules for pairwise graphs. The message-passing rules can be instantiated to find an approximation or an upper bound on the value of the best controller.

2. BACKGROUND

The challenge of planning problems in partially observable settings is to find a control policy for selecting actions when the precise state of the environment is unknown and the agent can only perceive partial observations, which convey incomplete information about the world's state. A partially observable Markov decision process (POMDP) provides a framework for sequential decision making under uncertainty, and is formally defined by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, b^0, \gamma \rangle$, where: 1) $\mathcal{S}$ is a finite set of all states s; 2) $\mathcal{A}$ is a set of actions a; 3) $\mathcal{O}$ is a set of observations o; 4) $T(s' \mid s, a)$ defines the *transition function*; 5) $O(o' \mid s', a)$ defines the *observation function*; 6) $R(r \mid s, a) \in \mathbb{R}$ is the immediate *reward function*; 7) $b^0(s) \in \Pr(s)$ is the *initial belief state* of the environment; 8) $\gamma \in [0, 1)$ is the *discount factor* at each time step that measures the relative importance of immediate and future rewards. The agent goal is to find an optimal control policy η^* that maximizes the expected discounted infinite-horizon reward: $\mathcal{V}^\eta = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t; \eta\right]$.

We can represent POMDP control policies compactly by restricting the space of control policies being considered and representing the control policy explicitly as a stochastic finite-state controller ($\mathcal{FSC}$). A controller $\eta = \langle \mathcal{N}, \pi, \lambda \rangle$ can encode a stochastic policy with three distributions (controller parameters): $p(N_0)$ (initial), $\pi = p(A_t \mid N_t)\ \forall t$ (action selection distribution) and $\lambda = p(N_t \mid N_{t-1}, O_{t-1})\ \forall t$ (controller successor distribution). Thus, a policy encoded by a controller is executed by performing the action associated with each node and by following the edge associated with each observation received.

Several techniques have been proposed to optimize controllers of a given size, including gradient ascent [10], branch and bound [3], bounded policy iteration [12], stochastic local search [2], non-convex quadratically constrained optimization [1] and expectation maximization (EM) [16]. The last approach makes a key observation: planning in the space of controllers can be transformed into an equivalent inference problem.

More specifically, POMDP controller optimization can be formulated as a parameter estimation problem with respect to a mixture of dynamic Bayesian networks ($\mathcal{DBN}$s). Since the correlations between the rewards at different time steps are irrelevant in planning, the problem can be decomposed into a series of finite horizon $\mathcal{DBN}$s with a single reward variable at the last time step. Figure 1 shows the $\mathcal{DBN}$s for one step, two steps and t steps. The parameters π and λ define respectively the conditional distributions $\Pr(A_t \,|\, N_t)$ and $\Pr(N_{t+1} \,|\, N_t, O_t)$ in each $\mathcal{DBN}$. Since all nodes in a $\mathcal{DBN}$ must be random variables, the reward variables are binary variables with conditional distributions: $\Pr(\hat{R}_t = \text{true} \,|\, S_t, A_t) = (R(S_t, A_t) - R_{\min})/((R_{\max} - R_{\min}))$, where $R_{\max} = \max_{s,a} R(s,a)$ and $R_{\min} = \min_{s,a} R(s,a)$. This effectively normalizes the rewards between 0 and 1, which allows us to treat them as probabilities. Since the value $\mathcal{V}^\eta = \mathbb{E}\,[\sum_t \gamma^t R(s_t, a_t) \,|\, \eta]$ of a controller η is the expected sum of discounted rewards, we can combine the $\mathcal{DBN}$s into a mixture with probabilities induced from discounting. The discount factor γ can be interpreted as the probability with which the process continues at each step. Hence, we can combine the $\mathcal{DBN}$s into a mixture such that the probability of the t-step $\mathcal{DBN}$ is proportional to the likelihood γ^t that the process will last at least t steps. Toussaint et al. [16] showed that the value of a controller is proportional to the mixture likelihood that the reward variables are true: $\mathcal{V}^\eta \propto \sum_t \gamma^t \Pr(\hat{R}_t = \text{true} \,|\, \eta)$. An optimal controller achieves the highest mixture of reward likelihoods. The search for an optimal controller can be formulated as a parameter estimation problem that can be tackled by expectation maximization (EM) [16]. This approach of policy optimization by likelihood maximization is appealing as it allows for exploiting the factored structure in a controller architecture and for taking advantage of natural structural constraints of planning problems. Unfortunately, due to the non-convex nature of the optimization problem in partially observable domains, the EM algorithm is not guaranteed to converge to a global optimum and may get stuck in arbitrarily bad sub-optimal configurations. Escape techniques have been developed to circumvent local optima [13], but there is still no performance guarantee.

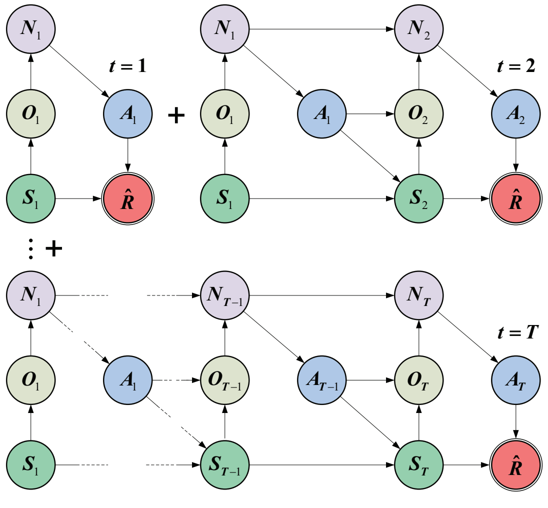


Figure 1: Modeling the POMDP value function as mixture-$\mathcal{DBN}$ model of a reward likelihood $\hat{R}$ of finite-time $\mathcal{DBN}$s.

3. SINGLE-DBN GENERATIVE MODEL

As a compelling alternative to the mixture of $\mathcal{DBN}$s, we developed a novel single-$\mathcal{DBN}$ model for planning by *marginal-MAP inference*, which allows us to adapt a Bayesian variational framework to approximate and bound the optimal value of any deterministic controller. For the rest of the paper, we will restrict ourselves to deterministic controllers (i.e., controllers with degenerate distributions that select unique actions and successor nodes). While stochastic controllers can achieve higher value, previous work has shown that optimal stochastic controllers are often nearly deterministic [4] and it may be easier to find a good controller when the space is restricted to deterministic controllers [3].

Deterministic controllers can be parameterized by a set $\boldsymbol{\theta} = \boldsymbol{\pi} \cup \boldsymbol{\lambda}$ of categorical variables, where $\boldsymbol{\pi} = \{\pi_n\}_{\forall n}$ and $\boldsymbol{\lambda} = \{\lambda_{no}\}_{\forall no}$. Here, $\pi_n \in \mathcal{A}$ indicates the action to be executed in node n and $\lambda_{no} \in \mathcal{N}$ indicates the successor node after receiving observation o in node n. Fig. 2 shows a dynamic Bayesian network that includes a controller, parameterized by $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, for which policy optimization is equivalent to marginal-MAP inference. In addition to the boolean reward variables $\hat{R}_t$, we also have boolean variables V_t and D_t whose probability of being true is proportional to the cumulative value and cumulative discount at time step t. We set $\Pr(V_t \,|\, V_{t-1}, R_t, D_t) = \psi(R_t, D_t) + \phi(V_{t-1})$ such that

$$\psi(R_t, D_t) = \begin{cases} (1-k) & \text{, if } R_t = D_t = \text{true} \\ 0 & \text{, otherwise} \end{cases}$$

$$\phi(V_{t-1}) = \begin{cases} k & \text{, if } V_{t-1} = \text{true} \\ 0 & \text{, otherwise} \end{cases}$$

Here $k \in (0, 1)$ is a scaling factor to ensure that probabilities are not greater than 1 when adding ψ and ϕ together. We can interpret $\psi(R_t = \text{true}, D_t = \text{true})$ as computing a probability proportional to the immediate reward at the current time step. Similarly, we can interpret $\phi(V_{t-1} = \text{true})$ as computing a probability proportional to the discounted cumulative reward up to the previous time step. The addition of ϕ and ψ yields a probability for $V_t = \text{true}$ that is proportional to the discounted accumulated reward up to the current time step. We also set

$$\Pr(D_t = \text{true} \,|\, D_{t-1}) = \begin{cases} k \cdot \gamma & \text{, if } D_{t-1} = \text{true} \\ 0 & \text{, otherwise} \end{cases}$$

This ensures that when we start with $\Pr(D_1 = \text{true}) = 1$, then $\Pr(D_t = \text{true})$ is proportional to γ^{t-1}. The following theorem confirms that $\Pr(V_T = \text{true} \,|\, \eta)$ is a positive affine transformation of the value function $\mathcal{V}_T^\eta$. It follows that a solution to the marginal-MAP inference problem $\max_\eta \Pr(V_T = \text{true} \,|\, \eta)$ is an optimal controller.

THEOREM 1. *The following equation holds:*

$$\Pr(V_T = \text{true} \,|\, \eta) = c_1\, \mathbb{E}\,[\sum_{t=1}^{T} \gamma^{t-1} R(S_t, A_t) \,|\, \eta] + c_2 \;\forall T$$

, where $c_1 \in \Re^+$ and $c_2 \in \Re$ are constants.

PROOF. We first prove by induction that $\Pr(V_T = \text{true} \,|\, \eta) = (1-k) \cdot k^{T-1} \cdot \sum_{t=1}^{T} \gamma^{t-1} \Pr(\hat{R}_T = \text{true} \,|\, S_T, A_T, \eta)\ \forall T$. Base case:

$$\Pr(V_1 = \text{true} \,|\, \eta) = (1-k)\Pr(\hat{R}_1 = \text{true} \,|\, S_1, A_1, \eta)$$

Assume by induction that $\Pr(V_{T-1} = \text{true} \mid \eta) = (1-k)k^{T-2}\sum_{t=1}^{T-1}\gamma^{t-1}\Pr(\hat{R}_t = \text{true} \mid S_t, A_t, \eta)$ then

$$\begin{aligned}
\Pr(V_T = \text{true} \mid \eta) &= k\Pr(V_{T-1} = \text{true} \mid \eta) \\
&\quad + (1-k)\Pr(D_T = \text{true})\Pr(R_T = \text{true} \mid S_T, A_T, \eta) \\
&= k(1-k)k^{T-2}\sum_{t=1}^{T-1}\gamma^{t-1}\Pr(\hat{R}_t = \text{true} \mid S_t, A_t, \eta) \\
&\quad + (1-k)(k\gamma)^{T-1}\Pr(R_T = \text{true} \mid S_T, A_T, \eta) \\
&= (1-k)k^{T-1}\sum_{t=1}^{T}\gamma^{t-1}\Pr(\hat{R}_t = \text{true} \mid S_t, A_t, \eta)
\end{aligned}$$

Since $\Pr(\hat{R}_T = \text{true} \mid S_T, A_T, \eta)$ is a positive affine transformation of $\mathbb{E}\left[R(S_T, A_T \mid \eta\right]$, the theorem follows. □

Corollary 1. *The controller $\eta^* = \arg\max_\eta \Pr(V_T = \text{true} \mid \eta)$ is optimal for the corresponding POMDP problem.*

Proof. By Theorem 1 and since $c_1 > 0$, it follows that

$$\arg\max_\eta \Pr(V_T = \text{true} \mid \eta) = \arg\max_\eta \mathbb{E}\left[\sum_{t=1}^{T}\gamma^{t-1}R(S_t, A_t) \mid \eta\right]$$

□

Interestingly, by casting the problem of optimizing POMDP controllers as a task of marginal-MAP probabilistic inference in the space of finite control policies, we can achieve a computational complexity reduction from PSPACE-complete to $\mathsf{NP}^{\mathsf{PP}}$-complete in comparison to solving the POMDP models ("search" vs. "dynamic programming"). It should be noted here that we propose to search for the best control policy in the restricted space of controllers that has a certain size (limit the search space with a fixed number of nodes), but not the best policy of the arbitrary size, as we cannot guarantee the optimal policy in a full sense for all problems. This way, there may be policies that are better, but have a larger size than considered by our approach.

Further, such a reformulation of the policy optimization task as a marginal-MAP inference problem allows for adapting a Bayesian variational framework to approximate the marginal-MAP inference and to obtain bounded algorithmic performance guarantees.

4. PLANNING BY MMAP INFERENCE

4.1 Summary of variational $\mathcal{MMAP}$ approach

To approach the task of policy optimization by marginal-MAP inference, we derived the Bayesian variational framework and developed mixed-product message-passing algorithms to (a) approximate the marginal-MAP inference, and (b) compute the upper bound of its solution specifically for general factor graphs with cycles as it is required for our case to optimize POMDP controllers by marginal-MAP inference in the proposed single-$\mathcal{DBN}$ model.

We propose to approach the original task of marginal-MAP inference by defining its dual variational representation and replacing the inference with an equivalent continuous optimization over variational distributions, which can be summarized as follows.

(1) We transform the original marginal-MAP problem Φ^{MMAP} into its dual variational form $\Phi^{MMAP}_{q_\tau}$ in order to further derive the tractable approximations and variational algorithms for estimating and bounding the marginal-MAP solution (Section 4.2): Φ^{MMAP}_{bethe}(approximate) $\propto$ $\Phi^{MMAP}_{q_\tau}$(exact) $\leq$ Φ^{MMAP}_{ttrw}(bound).

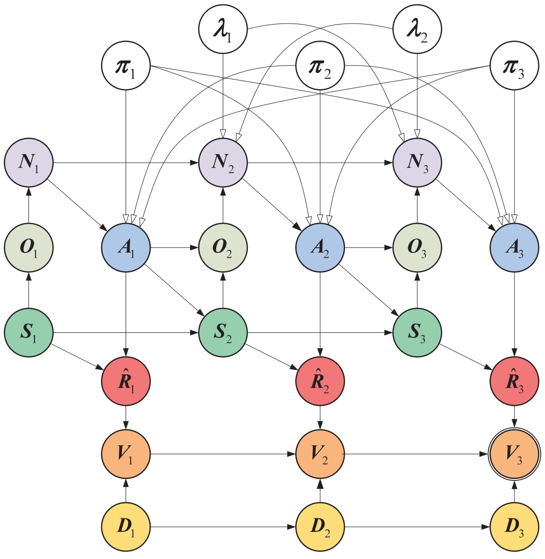


Figure 2: Single-$\mathcal{DBN}$ "$V - \mathcal{D}$" model for POMDP planning

(2) To solve the equivalent variational problems (Φ^{MMAP}_{bethe}, Φ^{MMAP}_{ttrw}) approximately and obtain performance guarantees, we developed the mixed-product message-passing algorithms to compute a set of optimal marginals (τ^*_i, τ^*_f) and the optimal configuration $\{X^*_B\}$ (Section 4.3).

(3) In order to obtain an upper bound Φ^{MMAP}_{ttrw} on the global optimum and guarantee its tractability, we propose a method to decompose the original factor graph with cycles into a special combination of its "AB" trees with a convex combination of their tractable distributions, where "sum" variables are denoted by "A" and "max" variables are denoted by "B" (Section 4.3). We also designed a new method for computing valid weights of "B" factors, based on the concave approximation of their entropy with "double counting numbers" (free energy "convexifying"), to convexity of the objective TTRW free energy function [20].

We are the first to our knowledge who derived these variational problems for factor graphs with cycles, and developed hybrid "mixed-product" algorithms to solve POMDP planning problems approximately with performance bounds. As a reference, we used previous work on variational marginal-MAP, developed for pairwise MRF models only [8], and work on TRW BP for "max" and "sum"-inference [18].

4.2 Variational approximation framework

We proposed a *Bayesian variational framework* and developed variational algorithms specifically for general factor graphs with cycles, as it is required in our case for marginal-MAP inference in single-$\mathcal{DBN}$ inference models. Particularly, we develop new variational inference algorithms to compute an approximate solution of the marginal-MAP problem and its upper bound on the global optimum by (1) representing the marginal-MAP problem by its special exponential form (convex but intractable), (2) translating this exponential marginal-MAP into a desirable dual variational form (*differentiable* objective function) with the unique optimal solution to further obtain tractable approximations, and (3) deriving the desired tractable approximate variational problems (free energies) and a special message-passing algorithm to solve the optimization problem efficiently (optimal MMAP configurations $\{X_B^*\}$, approximate estimate Φ_{bethe}^{MMAP} and the upper bound Φ_{ttrw}^{MMAP} for the original MMAP problem).

In order to derive the desirable variational (dual) form of the marginal-MAP problem with the unique optimal solution, we use the convexity properties of the exponential form of the marginal-MAP problem Φ_θ^{MMAP} and KL-divergence $D_{KL}(q_\tau \,||\, p) = \sum_x q_\tau \log(q_\tau(x_A \,|\, x_B)/p(x_A \,|\, x_B)) \geq 0$. The KL-divergence is the measure of dissimilarity between the variational distribution $q_\tau(x)$ and the true joint distribution $p(x)$. Therefore, when $D_{KL}(q_\tau \,||\, p) = 0$ then $q_\tau(x) = p(x)$. Thus, we obtain the following dual form of the variational marginal-MAP:

$$\Phi_{q_\tau}^{MMAP} = \max_{\tau \in M} \mathbb{E}_{\ q_\tau}[\theta(x)] + H(X_A \,|\, X_B, q_\tau) = \max_{\tau \in M} F_{mix}, \tag{1}$$

where: (1) $M = \{\tau : \exists q_\tau(x), \text{s.t } \tau_f(x_f) = \sum_{x \setminus x_f} q_\tau(x)\}$ is the marginal polytope with factor nodes marginals $\tau_f(x_f)$ of valid variational distributions q_τ; (2) q_τ is a valid variational distribution for marginals $\tau_f \in M$, when it is consistent with model observations, has the same form as a true joint and maximizes entropy $H(X_A \,|\, X_B, q_\tau)$; (3) $\mathbb{E}_{\ q_\tau}[\theta(x)] = \sum_x q_\tau(x)\theta(x)$ is the expected energy for the variational distribution q_τ; (4) $H(X_A \,|\, X_B, q_\tau) = -\sum_x q_\tau(x) \log q_\tau(x_A \,|\, x_B)$ is the conditional entropy for the variational distribution q_τ, (5) the exponential parameter vector $\theta(x) = \sum_{f_i \in F} \log f_i(X_{f_i})$ is introduced to transform the original joint distribution into its exponential form $P(x) = \exp[\theta(x)]$, (6) the set of sum-nodes $x_i \in X_A$ and the set of max-variables $x_i \in X_B$.

Thus, the variational form translates the original marginal-MAP inference problem into the problem of continuous optimization of the truncated free energy F_{mix} over the marginals defined by the marginal polytope M. The solution of this problem is the variational distribution $q_{\tau^*}(x)$, which probability mass function concentrates on the optimal set $\{X_B^*\}$.

There are two primary challenges associated with the variational representation of the marginal-MAP problem: (1) the set of constraints M (marginal polytope) is extremely difficult to characterize in an explicit manner, and (2) the optimization of the truncated free energy F_{mix} is computationally very complex since F_{mix} has an indirect dependency on the marginals from the marginal polytope M. Although we can not solve this variational problem exactly, it allows us to derive tractable approximations and new variational algorithms to estimate and bound the marginal-MAP solution for general cyclic factor graphs.

In order to obtain tractable approximations from the exact variational problem $\Phi_{q_\tau}^{MMAP}$, we relax a complex marginal polytope M with a large set of constraints to a simple convex polytope L. The derived approximate variational problems are tractable since their dual objective function has a differentiable and explicit form and it is optimized over marginals from a set of simple linear constraints. Particularly, we substitute the maximization of the indirect objective function $\max_{\tau \in M} F_{mix}(q_\tau, \theta)$ over the complex marginal polytope M with the maximization of the explicit approximate function $\max_{\tau \in L} F_{bethe}^{MMAP}(\tau, \theta)$ (and F_{ttrw}^{MMAP}) over locally consistent marginals from a convex set L with simple linear constraints: $L = \{\tau_i, \tau_{f_i} : \tau_i, \tau_{f_i} \geq 0; \sum_{x_i} \tau_i = 1; \sum_{X_{f_i} \setminus x_j} \tau_{f_i} = \tau_j\}$. We apply these results to obtain the truncated Bethe free energy F_{bethe}^{MMAP}, which optimization over convex set L provides an approximate marginal-MAP solution for cyclic factor graphs:

$$\begin{cases} \Phi_{q_\tau}^{MMAP} \approx \left[\Phi_{bethe}^{MMAP} = \max_{\tau \in L} F_{bethe}^{MMAP}\right] \\ F_{bethe}^{MMAP} = \mathbb{E}_{\ \tau}[\theta(x)] + \sum_{i \in A} H_i(\tau) - \sum_{f_i \setminus f_i^B} I_{f_i}(\tau) \end{cases} \tag{2}$$

where: (1) $\tau_i(x_i), \tau_{f_i}(X_{f_i}) \in L$ are locally consistent marginals of variable and factor nodes ($i \in V, f_i \in F$); (2) $\mathbb{E}_{\ \tau}[\theta(x)] = \sum_{f_i} \sum_{X_{f_i}} \tau_{f_i}(X_{f_i})\theta_{f_i}(X_{f_i})$ is the expected energy for marginals $\tau_i, \tau_{f_i} \in L$; (3) $H_i(\tau) = -\sum_{x_i} \tau_i(x_i) \log \tau_i(x_i)$ is the entropy of marginals of sum variable nodes $x_i \in X_A$; (4) $I_{f_i}(\tau) = \sum_{X_{f_i}} \tau_{f_i}(X_{f_i}) \log\left[\frac{\tau_{f_i}(X_{f_i})}{\prod_{x_i \in X_{f_i}} \tau_i(x_i)}\right]$ is the mutual information of "A" and "AB" factors, where each of these factors either contains only sum variables or sum variables together with max variables $f_i^A, f_i^{AB} \in F$, and factors $f_i^B(X_B)$ with max variables only are excluded. The truncated Bethe free energy is not convex to guarantee a global optimality of the solution. However, our experiments demonstrated that it provides a high quality approximation even for complex cyclic factor graphs.

To compute a variational upper bound for marginal-MAP inference, we derive a new objective function: truncated tree-reweighted free energy F_{ttrw}^{MMAP}. Optimization of the truncated tree-reweighted free energy over marginals from the set L allows to compute an upper bound of the MMAP problem for our single-$\mathcal{DBN}$ inference model. It is important to highlight that F_{ttrw}^{MMAP} is convex by its construction and the resulting upper bound is globally optimal.

To guarantee the *tractability* of the resulting upper bound, we developed a method for splitting a factor graph with cycles into a tractable combination of "AB" trees, which special structure provides an effective elimination order (Section 4.3). Additionally, to ensure global optimality of the upper bound, we define certain *convexity* constrains for the weights $\boldsymbol{\rho}$ of "AB" trees such that resulting decomposition of the original graph (parameter vector θ_T) into a linear combination of tractable distributions $\{\theta_T\}$ is convex.

$$\begin{cases} \Phi_{q_\tau}^{MMAP} \leq \left[\Phi_{ttrw}^{MMAP} = \max_{\tau \in L} F_{ttrw}^{MMAP}\right] \\ F_{ttrw}^{MMAP} = \mathbb{E}_{\ \tau}[\theta(x)] + \sum_{i \in A} H_i(\tau) - \sum_{f_i \setminus f_i^B} \rho_{f_i} I_{f_i}(\tau) \end{cases} \tag{3}$$

where: $\rho_{f_i} = \sum_{T \ni f_i} \rho_T$ is the factor appearance probability, and a set of weights assigned to the "AB" trees is defined as: $\boldsymbol{\rho} = \{\rho_T : \rho_T \geq 0, \sum_T \rho_T = 1, \sum_T \rho_T \theta_T = \theta\}$.

Importantly, the truncated Bethe free energy $F_{bethe}^{MMAP}(\tau, \theta)$ equals to the exact truncated free energy $F_{mix}(q_\tau, \theta)$ for tree-structured graphs. This property allows us to derive the optimal tree-based variational bound Φ_{ttrw}^{MMAP} based on the tree-based exponential general upper bound $\psi(\theta^T, \rho)$. In particular, we obtain the optimal upper bound for the marginal-MAP problem Φ_{ttrw}^{MMAP} in two steps. On the first step, we obtain a general upper bound in the exponential form $\psi(\theta^T, \rho)$ (inequality 1.2), by splitting the parameter vector θ in the exponential form of the MMAP problem (equality 1.1) into a linear combination of parameter vectors θ^T of tractable distributions ("AB" trees), and then by applying the Jensen's inequality:

$$\Phi_\theta^{MMAP} \overset{1.1)}{=} \Phi(\sum_T \rho_T \theta^T) \overset{1.2)}{\leq} \sum_T \rho_T \Phi(\theta^T) = \psi(\theta^T, \rho) \tag{4}$$

On the second step, we finally obtain the optimal upper bound Φ_{ttrw}^{MMAP}, by proving that optimization of the exponential convex function $\psi(\theta^T,\rho)$ by θ^T with the fixed parameters ρ is equivalent to solving the variational optimization problem, where we maximize the convex objective function $F_{ttrw}^{MMAP}(\tau;\theta)$ over the convex set L (equality 2.1):

$$\begin{cases} \min_{\theta^T} \psi(\theta^T,\rho) \overset{2.1)}{=} \max_{\tau\in L} F_{ttrw}^{MMAP}(\tau,\theta) \\ \max_{\tau\in M} F_{mix}(q_\tau,\theta) \overset{2.2)}{\leq} \max_{\tau\in L} F_{ttrw}^{MMAP}(\tau,\theta) \end{cases}$$

In order to prove this equivalence, we apply the Lagrangian dual method to optimize the convex exponential function $\psi(\theta^T,\rho)$ by distributions θ^T with linear constrains $\boldsymbol{\rho}$. Importantly, linear constrains $\boldsymbol{\rho}$ of the θ decomposition are enforced by Lagrangian multipliers and ensure that optimal pseudo-marginals agree across all "AB" trees $\{\theta^T\}$. We finally obtain a desirable variational optimization problem Φ_{ttrw}^{MMAP} with differentiable and convex objective function $F_{ttrw}^{MMAP}(\tau;\theta)$, which computes the globally optimal upper bound, by combining the results of this optimization $\min_{\theta^T} \psi(\theta^T,\rho)$ together with the tree-based truncated Bethe free energies for each "A-B" tree.

4.3 Mixed-product variational MMAP algorithms

We developed a new "mixed" message-passing algorithm specifically for the factor graphs with cycles, which allows to compute the marginal-MAP solution efficiency by executing both the maximization and marginalization operations within one message-propagation sweep.

In short, we derived our variational "mixed" message-passing approach by (1) introducing an approximate free energy F'_{mix} of a special generalized form to obtain a dual differentiable objective function, (2) obtaining the generalized message-update rules, based on the method of Lagrangian multipliers applied to the generalized free energy, (3) transforming the generalized message-update rules into the final "mixed" message-passing rules by minimizing the approximation error, which appears due to the generalization of the variational objective function with the weighted entropy of "B" nodes, and finally (4) computing both the estimate and the upper bound of the marginal-MAP problem by applying the derived mixed-product rules to optimize the variational truncated Bethe (F_{bethe}^{MMAP}) and TTRW free energies (F_{ttrw}^{MMAP}) respectively. Importantly, we additionally discuss methods for computing factors weights, which have critical impact on the validity of the resulting upper bound when optimizing the TTRW free energy with the mix-product algorithm.

In order to derive the message-passing rules for optimizing the approximate truncated free energy, we follow a similar approach previously introduced for sum-inference in MRF models [18]. According to this existing approach, a message-passing scheme for optimizing the TRW free-energy of sum-inference problems can be derived based on the method of Lagrangian multipliers. Unfortunately, we cannot apply this method directly to approximate a free energy of the marginal-MAP inference since we have to remove the entropy of "B" nodes from the variational objective function as it is required for the variational form of the marginal-MAP problem. Thus, if we attempt to formulate the dual Lagrangian function using the same approach, it will not cover "max" marginals $\tau_{i^B}, \tau_{f_i^B}$ and related constraints. In order to overcome this problem, we propose to "add back" the entropy of "max" nodes to the truncated free energy with a specific temperature coefficient ϵ. In this case, it can be proved that such an approximate variational problem approaches the exact marginal-MAP when we enforce ($\epsilon \to 0^+$) [8].

Thus, we "add back" the entropy of max nodes to the truncated Bethe and TTRW free energies, and formulate the generalized variational problem for factor graphs as follows:

$$\max_{\tau\in L} F'_{mix} = \max_{\tau\in L} \mathbb{E}_{\tau}[\theta(x)] + \sum_{i\in V} w_i H_i(\tau) - \sum_{f_i\in F} w_{f_i} I_{f_i}(\tau), \tag{5}$$

where weights w_i, w_{f_i} are strictly positive and depend on the type of the variable nodes (sum-node $i \in A$ or max-node $i \in B$) and the type of the factor nodes ("B"-factors or "A", "AB"-factors):

$$\begin{cases} w_i = 1 \text{ for any } i \in A \\ w_i = \epsilon \text{ for any } i \in B \\ w_{f_i} = \rho_{f_i} \text{ for any } f_i \in F^A, f_i \in F^{AB} \\ w_{f_i} = \epsilon\cdot\rho_{f_i} \text{ for any } f_i \in F^B \end{cases} \tag{6}$$

Thus, we obtained the generalized free energy, which can now be converted into either a truncated Bethe free energy or a TTRW free energy by setting the factor weights ρ_{f_i}. An important property of this generalized dual variational problem is its differentiable objective function, which covers all factor and variable nodes. This property allows us to further apply the Lagrangian multipliers method and derive the mixed-product algorithm, which fixed points correspond to the optimal solution of the approximate variational problem. We obtain the optimal marginals τ_j^*, $\tau_{f_i}^*$ by taking derivatives of the Lagrangian dual function $\mathcal{L}(\tau,\lambda)$ (omitted due to space limitation) with respect to marginals τ_j, τ_{f_i} and introducing "messages" from factors to variables as $\mu_{f_i\to x_j}(x_j) = \exp(\lambda_{f_i x_j}(x_j))$:

$$\tau_j^*(x_j) \propto \prod_{f_i\in N(j)} \left[\mu_{f_i\to x_j}(x_j)\right]^{1/w_i} \tag{7}$$

$$\tau_{f_i}^*(X_{f_i}) \propto \prod_{j\in N(f_i)} \left[\tau_j^*(x_j)\cdot\left(\frac{f_i(X_{f_i})}{\mu_{f_i\to x_j}(x_j)}\right)^{1/w_{f_i}}\right] \tag{8}$$

Further, we ensure that these marginals are valid ($\tau_i, \tau_{f_i} \in L$) by deriving such message-update rules that enforce the normalization, non-negativity and marginalization constraints of the locally consistent set L. Note that we explicitly enforce normalization and non-negativity constraints ($\sum_{x_i}\tau_i = 1, \tau_i, \tau_{f_i} \geq 0$), and we apply marginalization constraints ($\sum_{X_{f_i}\setminus x_j}\tau_{f_i}(X_{f_i}) = \tau_j(x_j)$) to the equations (7) and (8), which yield the update rules for the *general messages* (omitted due to space limitation). Finally, we derive the final mixed message-passing scheme for general factor graphs with cycles. In particular, in addition to the regular sum-product message rule, we also obtain max-product and special argmax message rules by enforcing $\epsilon \to 0^+$ for weights of "max" variable nodes and "B" factors (expression 6) in the generalized message-update rules and applying zero temperature limit formula together with properties of limits near infinity. This special argmax message is particularly important, since it allows to decode the joint optimal configuration for all max variable during the message propagation.

Finally, we derive the following mixed message-passing scheme for general factor graphs with cycles. We can compute the marginal-MAP solution effectively by propagating the proposed mixed-product messages since the proposed algorithm allows to solve "sum" and "max" problems simultaneously, and to decode the optimal configuration of the "max" from the simple node beliefs during the message propagation with the "argmax" messages.

Messages from variables to factors:

$$\mu_{x_k \to f i}(x_k) = \prod_{f_h \in \text{neighbours}(x_k)} \mu_{f_h \to x_k}(x_k)$$

Messages from "A" factors to variables:

$$\mu_{f \to x_j}(x_j) = \left[\sum_{\sim x_j} \prod_{k \neq j} \left\{ \mu_{x_k \to f}(x_k) \cdot \left(\frac{f(X_f)}{\mu_{f \to x_k}(x_k)} \right)^{1/\rho_f} \right\} \right]^{\rho_f}$$

Messages from "B" factors to variables:

$$\mu_{f \to x_j}(x_j) = \max_{\sim x_j} \prod_{k \neq j} \left\{ \left(\mu_{x_k \to f}(x_k) \right)^{\rho_f} \cdot \left(\frac{f(X_f)}{\mu_{f \to x_k}(x_k)} \right) \right\}$$

Messages from "AB" factors to variables:

$$\mu_{f \to x_j}(x_j) = \left[\sum_{\{x_A, x_B^* \backslash x_j\}} \prod_{k \neq j} \left\{ \mu_{x_k \to f}(x_k) \left(\frac{f(X_f)}{\mu_{f \to x_k}(x_k)} \right)^{1/\rho_f} \right\} \right]^{\rho_f}$$

where x_A denotes the "sum-out" random variables and x_B^* denotes the "max-out" decision variables with their domain restricted to the values that maximize $\prod_{f_h} \mu_{f_h \to x_k}(x_k)$.

Thus, we compute the marginal-MAP solution effectively by propagating the proposed mixed-product messages since the proposed algorithm allows to solve "sum" and "max" problems simultaneously, and to decode the optimal configuration of the "max" from the simple node beliefs during the message propagation with the "argmax" messages. It is important to highlight, that although we compute the marginal-MAP configurations locally for each "max" variable, we continue message-propagation and continually reassess configurations for all "max" nodes until its convergence, where at the final convergent point each "local" marginal-MAP solution is based on the joint set of optimal configurations for all other "max" variables $x_B^*, x_B \in B$.

It is also important that the mixed-product algorithm allows to compute not only the optimal configuration of "max" variables, but also a TTRW upper bound and the Bethe approximation for the marginal-MAP problem (where all factors weights are set as $\rho_f = 1$). In particular, the set of "optimal" messages obtained at the convergent point of the mixed-product algorithm can be used to find the optimal mixed-marginals, which are required to further compute optimal objective functions (free energies $F_{ttrw}^{MMAP}, F_{bethe}^{MMAP}$).

However, to obtain a valid upper bound by optimizing the TTRW free energy with the mixed-product algorithm, we additionally developed (1) a new method for computing "A","AB" factors weights, based on the decomposition of cyclic graphs into a convex and tractable combination of "AB"-trees, and (2) a new method for computing the provably convex weights for "B" factors, based on the approximation of convex free energies with "double counting numbers" [20].

"AB" trees for the TTRW mixed-product algorithm To guarantee the *tractability* and *global optimality* of the resulting upper bound, we developed a method for splitting an original factor graph with cycles (exponential parameter vector θ) into a convex, tractable combination of "AB" trees (tractable distributions $\{\theta_T\}$), such that the weights assigned to "AB" trees ($\{\rho_T\}$) satisfy convexity constraints ($\rho_T \geq 0, \sum_T \rho_T = 1$) and θ-decomposition constraints ($\sum_T \rho_T \theta_T = \theta$), and a structure of each "AB" tree guarantees an effective elimination order for the marginal-MAP inference. Further, this weighted combination of "AB" trees allows to compute a weight of each "A", "AB" factor as a factor appearance probability across all of its trees: $\{\rho_f = \sum_{T: f \in F^T} \rho_T\}$.

We formulate the following rules to construct each "AB" tree, which special structure provides an effective elimination order for the marginal-MAP inference to guarantee the tractability of the resulting upper bound.

Rules for constructing "AB" trees for factor graphs:

- Identify all connected components of the sub-graph with only sum-nodes and sum-factors (G^A).
- Construct a spanning forest in G^A by adding at most one "AB" factor to each connected component of G^A, and remove edges and factors to eliminate cycles.
- Validate that there are no two "AB" factors connected to the same sum node or a sub-tree with sum nodes and sum factors. If validation fails, remove one of such "AB" factors from the "AB" tree.

The following algorithm summarizes the proposed results for splitting a factor graph with cycles into a tractable, convex combination of "AB" trees and computing optimized "A", "AB" factors weights. Optimization of factor decomposition weights is based on a full coverage of "A","AB" factors with a minimum number of "AB" trees. This property allows us to effectively compute an upper bound for the marginal-MAP problem and additionally to avoid possible numerical issues when the number of "AB" trees in a graph decomposition is large. This method allows to compute an upper bound for the marginal-MAP solution effectively by optimizing the TTRW free energy with the mixed-product algorithm.

Algorithm to compute weights of "A", "AB" factors:

- Identify all connected components of the sub-graph with only sum-nodes and sum-factors (G^A).
- Identify all "AB" factors that share the same sub-set of sum-nodes in its scope, and form a set F'_{AB} with these factors.
- For "AB" tree T_i select the "AB" factor f_i^{AB} from the set F'_{AB} and construct the tree $T'_{A_{AB}}$ with this "AB" factor and all connected components of G^A, which include sum-nodes from the scope of this factor.
- Select "AB" factors $f^{AB} \in F_{AB} \backslash F'_{AB}$ and construct trees with each of these factors and remaining sum-nodes and sum-factors from $G^A \backslash T'_{A_{AB}}$.
- Add not-covered trees or separate sum-nodes from G^A to obtain a *spanning forest* in the sub-graph G^A.
- Check if any two "AB" factors in the resulting "AB" tree are connected by a sum-node or share the same tree of G^A. If such factors exist, then remove one of them from the "AB" tree and add it back to the set F'_{AB}.
- Build the next "AB" tree $\{T_{i+1}\}$ by selecting the "AB" factor f_{i+1} from F'_{AB} and performing the steps above. Continue building "AB" trees until all "A","AB" factors are covered.

- Compute the weight for each "AB" tree as: $\rho_T = \frac{1}{N}$, where N is the total number of "AB" trees.
- Compute the "A","AB" factors weights: $\rho_f = \sum_{T:f \in F^T} \rho_T$, $f \in F^A, F^{AB}$

However, in order to compute the upper bound with the mixed-product algorithm, we additionally need to apply the TTRW approximation to the generalized free energy (equation 5), which also requires the entropy of "B" nodes to be valid and provably concave [6]. Therefore, we further propose a new method for computing the weights of "B" factors, which provides a provably concave and valid approximation for the entropy of "B" nodes.

Rules for computing "B" factors decomposition weights In order to guarantee the convergence of the TTRW mixed-product algorithm to the globally optimal upper bound, it is critical to ensure the convexity of the variational objective function, which requires the entropy of "B" nodes to be concave and the weights of "B" factors to satisfy convexity and validity constraints. In order to address this problem, we designed a new method for computing the weights of "B" factors ($\rho_f, f \in F^B$), based on (1) applying the concave approximation to the entropy of "B" nodes (with certain double counting numbers) of the TTRW variational problem (H_B^{TTRW}) [20], [19], and (2) ensuring that the resulting factor weights ρ_{f_B} satisfy a constraint on valid counting of variable and factor nodes [9].

In order to construct a concave approximation of entropy H_B^{TTRW}, we introduce its concave decomposition based on the method of using double counting numbers, which satisfy certain convexity constraints [20]. The double counting numbers c_f, c_i define a linear combination of entropies over individual variables and factor nodes:

$$H(x) = \sum_f c_f H_f(x_f) + \sum_i c_i H_i(x_i) \qquad (9)$$

where the double counting numbers c_f, c_i satisfy the convexity constraint if there are exist the non-negative numbers c_{ff}, c_{ii}, c_{if} such that:

$$\begin{cases} c_f = c_{ff} + \sum_{i:i \in N(f)} c_{if} \\ c_i = c_{ii} - \sum_{f:i \in N(f)} c_{if} \\ c_{if} \geq 0, c_{ff} \geq 0, c_{ii} \geq 0 \end{cases}$$

In order to obtain a concave entropy form, which we can further use to derive conditions for the concavity of the H_B^{TTRW}, we replace c_f, c_i with the expressions from the convexity constraints in the entropy decomposition (equation 9):

$$H = \sum_{f,i:i \in N(f)} c_{if}(H_f - H_i) + \sum_f c_{ff} H_f + \sum_i c_{ii} H_i \quad (10)$$

Now we can prove that the entropy H_B^{TTRW} is concave with any $\rho_{f_j} > 0$, by transforming it into a special form which corresponds to the concave entropy form (expression 10) with $c_{ii} = 1, c_{ff} = 0, c_{if} = \rho_{f_j} > 0$:

$$H_B^{TTRW} = \sum_{i \in B} H_i + \sum_{f_j \in F^B} \sum_{i:i \in N(f_j)} \rho_{f_j} \big(H_{f_j}(x_{f_j}) - H_i(x_i)\big)$$

Thus, we identified that the double counting numbers in the concave decomposition of the entropy H_B^{TTRW} satisfy convexity constraints with any non-negative ρ_{f_j}. However, this concave entropy H_B^{TTRW} is valid only if the double counting numbers ensure a "valid counting" of factor and variable nodes [9]: $\{c_i = (1 - \sum_{f:i \in N(f)} c_f), c_f = 1\}$. We apply this requirement to the c_f, c_i in the concave decomposition of H_B^{TTRW}, and obtain that the counting of variables is always valid with our counting numbers, while the counting of factors is valid only when $\sum_{i:i \in N(f)} \rho_f = 1$:

$$\begin{cases} c_{if} = \rho_{f_j}, c_{ff} = 0, c_{ii} = 1 : \\ c_f = \sum_{i:i \in N(f)} \rho_f = 1 \\ c_i = 1 - \sum_{f:i \in N(f)} \rho_f = 1 - \sum_{f:i \in N(f)} c_f \end{cases}$$

Further, we use this factor counting constraint to derive the rule for computing such weights of "B" factors which guarantee the concavity of the "B" nodes and convexity of the generalized TTRW objective function:

$$c_f = \sum_{i:i \in N(f)} \rho_f = 1 \Rightarrow \rho_f = \frac{1}{d_f}, d_f = \sum_{i:i \in N(f)} 1$$

, where d_f is the number of variables in scope of factor f. Indeed, for any factor f, the equality holds and $\rho_f \in (0, 1]$ since each factor includes at minimum one variable. Therefore, this method for computing "B" factors weights provides a valid and provably concave entropy H_B^{TTRW}, and can be effectively used for optimizing the TTRW free energy by the mixed-product algorithm on general factor graphs. Thus, our new methods allow to compute valid weights for all factors, and guarantee tractability of the mixed-product algorithm and its convergence to the globally optimal upper bound of a marginal-MAP problem for general cyclic factor graphs with cycles.

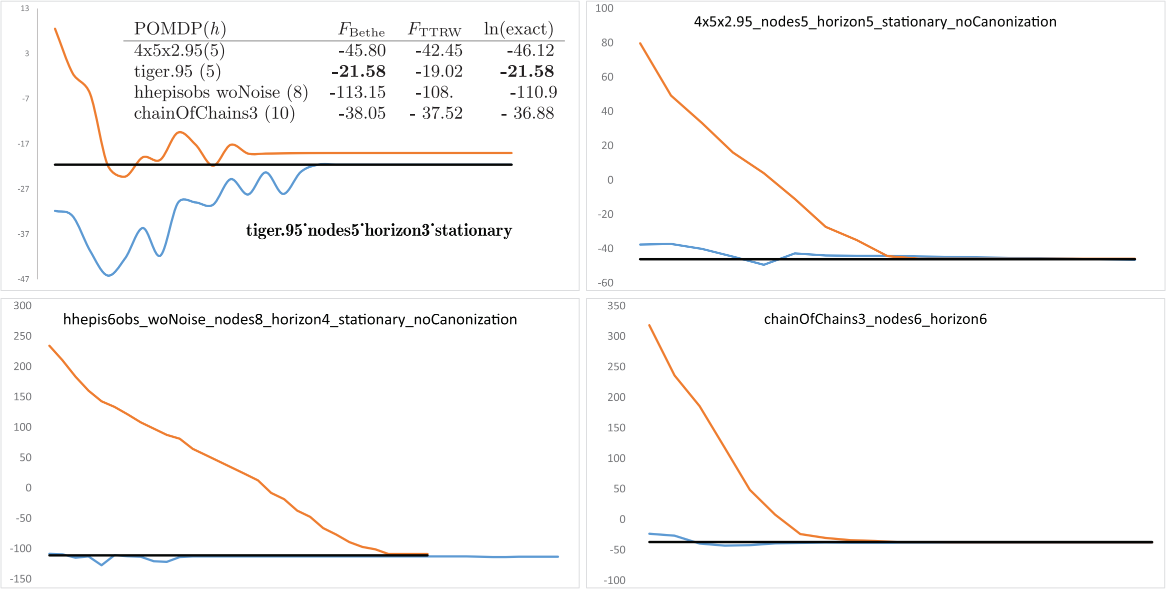
5. EXPERIMENTAL RESULTS

We evaluated our methods using several POMDP benchmark problems and compared their performance to the SamIam Bayesian solver. The Table in Figure 3 compares the log probability (positive affine transformation of the discounted rewards) of the policy found by exact marginal-MAP to the Bethe approximation and the TTRW upper bound. The planning horizon h for each problem is written in parentheses besides the name of each problem. The experimental analysis demonstrates that the truncated Bethe free energy provides a high quality approximation even for planning problems with complex cyclic factor graphs. The four graphs demonstrate convergence of approximate (blue) and upper bound (red) algorithms to exact value (black) of $\ln \Pr(V_T = true|\eta)$ as the number of iterations of message passing increases. Additionally, the experimental results reveal that for some problems (tiger, chain) we can maintain fairly tight upper bounds on the optimal value function, obtained by our variational inference algorithms.

6. CONCLUSION

We demonstrated that the original task of optimizing POMDP controllers can be casted as a problem of marginal-MAP inference in a novel single-$\mathcal{DBN}$ model, which can be solved by a variational hybrid "mixed-product" algorithm to obtain an approximate solution and its upper bound. The proposed approach is evaluated on several POMDP benchmark problems and the performance of the implemented variational algorithms is compared to SamIam Bayesian solver. Our future work is to optimize the weights to achieve better accuracy and to scale the algorithms to problems with larger

Figure 3: Performance of the variational algorithms is evaluated on several POMDP benchmark problems



POMDP(h)	F_{Bethe}	F_{TTRW}	ln(exact)
4x5x2.95(5)	-45.80	-42.45	-46.12
tiger.95 (5)	**-21.58**	-19.02	**-21.58**
hhepisobs woNoise (8)	-113.15	-108.	-110.9
chainOfChains3 (10)	-38.05	- 37.52	- 36.88

planning horizons. Additionally, further research is being conducted on extending the proposed single-$\mathcal{DBN}$ inference model to Dec-POMDP planning in multiagent settings.

7. REFERENCES

[1] C. Amato, D. Bernstein, and S. Zilberstein. Solving POMDPs using quadratically constrained linear programs. In *IJCAI*, pages 2418–2424, 2007.

[2] D. Braziunas and C. Boutilier. Stochastic local search for POMDP controllers. In *AAAI*, pages 690–696, 2004.

[3] M. Grzes, P. Poupart, and J. Hoey. Isomorph-free branch and bound search for finite state controllers. In *IJCAI*, 2013.

[4] E. A. Hansen. Sparse stochastic finite-state controllers for pomdps. In *UAI*, pages 256–263, 2008.

[5] M. Hoffman, H. Kueck, A. Doucet, and N. de Freitas. New inference strategies for solving Markov decision processes using reversible jump MCMC. In *UAI*, 2009.

[6] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[7] A. Kumar and S. Zilberstein. Anytime planning for decentralized POMDPs using expectation maximization. In *UAI*, pages 294–301. AUAI Press, 2010.

[8] Q. Liu and A. T. Ihler. Variational algorithms for marginal MAP. In F. G. Cozman and A. Pfeffer, editors, *UAI*, pages 453–462. AUAI Press, 2011.

[9] O. Meshi, A. Jaimovich, A. Globerson, and N. Friedman. Convexifying the bethe free energy. *arXiv:1205.2624 [cs]*, May 2012.

[10] N. Meuleau, K.-E. Kim, L. Kaelbling, and A. Cassandra. Solving POMDPs by searching the space of finite policies. In *UAI*, pages 417–426, 1999.

[11] J. Pajarinen and J. Peltonen. Expectation maximization for average reward decentralized pomdps. In *ECMLPKDD*, 2013.

[12] P. Poupart and C. Boutilier. Bounded finite state controllers. In *NIPS*, 2003.

[13] P. Poupart, T. Lang, and M. Toussaint. Analyzing and escaping local optima in planning as inference for partially observable domains. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *ECML/PKDD (2)*, volume 6912 of *Lecture Notes in Computer Science*, pages 613–628. Springer, 2011.

[14] M. Toussaint, L. Charlin, and P. Poupart. Hierarchical POMDP controller optimization by likelihood maximization. In *UAI*, 2008.

[15] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state Markov decision processes. In *ICML*, 2006.

[16] M. Toussaint, A. Storkey, and S. Harmeling. Expectation maximisation methods for solving (PO)MDPs and optimal control problems. In *Bayesian Time Series Models*, pages 388–413. Cambridge University Press, 2011.

[17] N. Vlassis and M. Toussaint. Model free reinforcement learning as mixture learning. In *ICML*, 2009.

[18] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335, 2005.

[19] Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *UAI*, 2007.

[20] J. Yedidia, W. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, July 2005.