Policy Optimization by Marginal-MAP Probabilistic Inference in Generative Models

(Extended Abstract)

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}@cs.uwaterloo.ca

ABSTRACT

While most current work in POMDP planning focus on the development of scalable approximate algorithms, existing techniques often neglect performance guarantees and sacrifice solution quality to improve efficiency. In contrast, our approach to optimizing POMDP controllers by probabilistic inference and obtaining bounded on solution quality can be summarized as follows: (1) re-formulate POMDP planning as a task of marginal-MAP "mix" (max-sum) inference with respect to a new single-$\mathcal{DBN}$ generative model, (2) define a dual representation of the MMAP problem and derive a Bayesian variational approximation framework with an upper bound, (3) and design hybrid message-passing algorithms to optimize a POMDP policy by approximate variational MMAP inference in the $\mathcal{DBN}$ generative 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

To address the scalability issues of solving increasingly large planning problems, the community has been making significant progress in developing scalable approximate planning algorithms. Unfortunately, solution quality is often sacrificed for scalability and there is a lack of performance guarantees. A promising approach to optimize POMDP controllers consists of viewing planning as an inference problem with respect to a mixture of dynamic Bayesian networks ($\mathcal{DBN}$s) [2]. This allows us to exploit the factored structure of the problem and to leverage recent advances in inference algorithms. Unfortunately, planning as inference does not change the fact that controller optimization is inherently non-convex. Hence, local optima are a major issue and existing techniques for planning as inference do not provide any performance guarantee. As a compelling alternative, we develop a single-$\mathcal{DBN}$ generative model for planning by marginal-MAP inference, which allows the application of a broader range of inference techniques since most inference techniques can be directly applied to a single graphical model, but not a mixture of graphical models. We also show how to adapt a Bayesian variational framework to marginal-MAP inference with upper bounds on solution quality.

2. PLANNING BY MMAP INFERENCE

2.1 Generative model for MMAP inference

We propose to solve POMDP problems by representing policies explicitly as deterministic finite-state controllers ($\mathcal{FSC}$s) and to optimize controllers by marginal-MAP inference. A controller encodes a policy with a set $\mathcal{N}$ of nodes n and 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. 1 shows a dynamic Bayesian network that includes a controller, parametrized by $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, for which policy optimization is equivalent to marginal-MAP inference. Here, S_t, A_t, O_t, N_t, R_t, V_t and D_t denote the state, action, observation, node, reward, value and discount variables at each time step t. Since this is a Bayesian network, all variables are random variables, including R_t, V_t and D_t. The key is to think about the conditional distributions of those variables as normalized versions of the immediate reward, cumulative value and cumulative discount in $[0, 1]$. As proposed by [2], R_t is a Boolean variable such that $\Pr(R_t = \text{true} \mid S_t, A_t) = [R(S_t, A_t) - R_{\min}]/[R_{\max} - R_{\min}]$, where $R_{\max} = \max_{s,a} R(s, a)$ and $R_{\min} = \min_{s,a} R(s, a)$. Instead of working with a mixture of $\mathcal{DBN}$s as done by [2], we introduce two additional Boolean variables V_t and D_t, which allow the entire POMDP to be converted in a single $\mathcal{DBN}$. We set $\Pr(V_t \mid V_{t-1}, R_t, D_t) = \psi(R_t, D_t) + \phi(V_{t-1})$, where $\psi(R_t, D_t) = (1 - k)$ when $R_t = D_t = \text{true}$ and 0 otherwise, and $\phi(V_{t-1}) = k$ when $V_{t-1} = \text{true}$ and 0 otherwise. We also set $\Pr(D_t = \text{true} \mid D_{t-1}) = k \cdot \gamma$ when $D_{t-1} = \text{true}$ and 0 otherwise. Here, $\gamma \in [0, 1]$ is the discount factor and $k \in (0, 1)$ is a scaling factor that ensures that probabilities are never greater than 1. At each step t, $\Pr(D_t = \text{true})$ is proportional to the cumulative discount γ^t, and $\Pr(V_t = \text{true})$ is proportional to the discounted sum

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

of rewards earned so far. For a planning horizon of T time steps, an optimal controller $\boldsymbol{\theta}^*$ can be obtained by computing $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \Pr(V_T = \text{true} \,|\, \boldsymbol{\theta})$. This optimization corresponds to a marginal-MAP inference problem since we are maximizing the decision variables $\boldsymbol{\theta}$ while summing out all random variables except for V_T, which is set to "true".

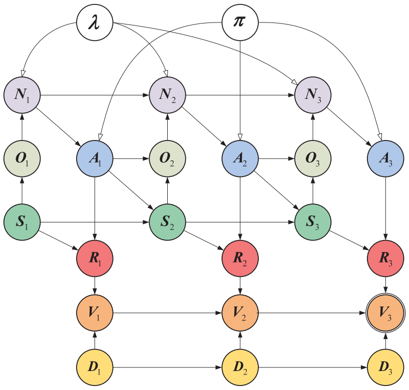


Figure 1: Single-$\mathcal{DBN}$ model for POMDP planning

2.2 Variational approximation framework

To opimize controllers by marginal-MAP inference, we extend previous approaches for variational sum-inference [3] and marginal-MAP inference [1], which were developed for pair-wise Markov random fields only. To our knowledge, we are the first to derive the following Bayesian variational framework and hybrid "mixed-product" message-passing algorithms to (1) approximate marginal-MAP inference, and (2) compute an upper bound of its solution for general factor graphs with cycles as it is required to optimize POMDP controllers. A tractable approximation to marginal-MAP inference consists of optimizing a truncated version of the Bethe free energy while an upper bound can be obtained by optimizing a truncated tree reweighted (TTRW) variational form. In both cases, message-passing algorithms can be derived to perform the optimization.

Consider a factor graph representation of the $\mathcal{DBN}$ in Fig. 1 where factors correspond to the conditional distributions. Variables will be referred to by x and factors by f. Let "A" be the set of factors that include only variables to be summed out (i.e., random variables: S_t, A_t, O_t, R_t, V_t, D_t), "B" the set of factors that include only variables to be maximized (i.e., decision variables: π_n, λ_{no}), and "AB" be the set of factors that include a mix of random and decision variables. A TTRW upper bound is obtained by decomposing the original factor graph into a weighted convex combination of spanning trees that each allow tractable inference. Let $\mathcal{T}$ denote a tree, which includes a subset of the factors of the original factor graph. We denote by $\rho_{\mathcal{T}}$ the probability associated with $\mathcal{T}$ in the convex combination. Thus, in the message passing rules below, when ρ_f is the appearance probability of factor f in the convex combination of trees (i.e. $\rho_f = \sum_{\{\mathcal{T}|f\in\mathcal{T}\}}$), the result is a TTRW upper bound on marginal-MAP and when $\rho_f = 1\,\forall f$ the result is a Bethe approximation of marginal-MAP. The following rules define the messages μ to be passed between variables and factors.

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:

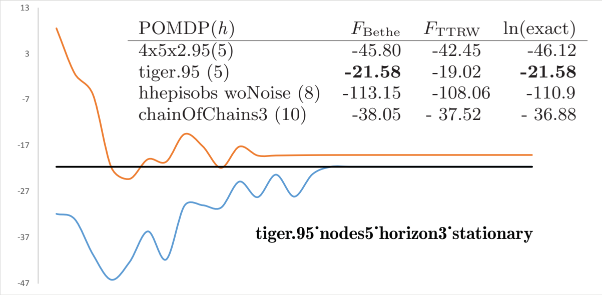
$$\mu_{f\to x_j}(x_j) = \max_{\sim x_j}\prod_{k\neq j}\left\{\left(\mu_{x_k\to f}(x_k)\right)^{\rho_f}\cdot\left(\frac{f(X_f)}{\mu_{f\to x_k}(x_k)}\right)\right\}$$

Messages from "AB" factors to variables:

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

Here, 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 $\prod_{f_h}\mu_{f_h\to x_k}(x_k)$.

Fig. 2 compares the log probability proportional to the sum of discounted rewards of the policy found by exact marginal-MAP to the Bethe approximation and the TTRW upper bound on a set of benchmark POMDPs with planning horizon h in parentheses. The graph shows how the log probability of each algorithm converges as the number of iterations of message passing increases.



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	-108.06	-110.9
chainOfChains3 (10)	-38.05	- 37.52	- 36.88

Figure 2: Results for several POMDP benchmark problems

3. CONCLUSION

We demonstrated that the optimization of POMDP controllers can be casted as a marginal-MAP inference problem in a novel single-$\mathcal{DBN}$ generative 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. Future work will focus on the optimization of the weight vector ρ to achieve tighter upper bounds and to scale the algorithms to problems with longer planning horizons.

4. REFERENCES

[1] 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.

[2] M. Toussaint, L. Charlin, and P. Poupart. Hierarchical POMDP controller optimization by likelihood maximization. In D. A. McAllester and P. Myllymaki, editors, *UAI*, pages 562–570. AUAI Press, 2008.

[3] 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.