Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

A Novel Single-$\mathcal{DBN}$ Generative Model for Optimizing POMDP Controllers by Probabilistic Inference

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

As a promising alternative to using standard (often intractable) planning techniques with Bellman equations, we propose an interesting method of optimizing POMDP controllers by probabilistic inference in a novel equivalent single-$\mathcal{DBN}$ generative model. Our inference approach to POMDP planning allows for (1) for application of various techniques for probabilistic inference in single graphical models, and (2) for exploiting the factored structure in a controller architecture to take advantage of natural structural constrains of planning problems and represent them compactly. Our contributions can be summarized as follows: (1) we designed a novel single-$\mathcal{DBN}$ generative model that ensures that the task of probabilistic inference is equivalent to the original problem of optimizing POMDP controllers, and (2) we developed several inference approaches to approximate the value of the policy when exact inference methods are not tractable to solve large-size problems with complex graphical models.

Inference in $\mathcal{R}$-mixture model of reward likelihood

The challenge of planning problems in partially observable settings (POMDP) 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. We can represent POMDP control policies compactly by restricting the space of control policies being considered and representing the policy explicitly as a stochastic finite-state controller ($\mathcal{FSC}$). However, the task of optimizing controllers is a notoriously difficult problem as the search space of all possible controller parameters is exponentially large. To address the scalability issues of solving large-scale planning problems, the community has been making significant progress in developing approximate planning algorithm to solve increasingly large problems.

As a promising alternative to using standard (often intractable) planning techniques with Bellman equations, an interesting method of optimizing POMDP controllers by probabilistic inference in the equivalent mixture-$\mathcal{DBN}$ generative model with exponentiated rewards as observation likelihoods has been previously proposed (Toussaint, Charlin, and Poupart 2008). Particularly, the artificial $\mathcal{R}$-mixture ("reward" mixture) model with broken correlations between the reward variables at consecutive time periods is defined in a special way, where for each time period there is only one length-T mixture-$\mathcal{DBN}$ component, modeling a single binary stochastic reward with $\Pr(R = \text{true} \,|\, A_T, S_T, T = t) \propto r(a_t, s_t)$, emitted only at its final termination step T (from the last state and action A_T and S_T) with $\Pr(T = t) = \gamma^t$. The fact that the observations of rewards at two different time periods of the $\mathcal{R}$-mixture model are now treated by separate mixture-$\mathcal{DBN}$ components allows us (1) to treat the decision policy of the Markovian model as a free *hidden policy parameter* $\theta \in [0, 1]$ of the $\mathcal{R}$-mixture model to be optimized, (2) to define artificial marginal likelihood of observing a single hypothetical evidence of a binary stochastic reward $L(\theta \,|\, R = \text{true})$ as a special representation of the *objective function*, and (3) to demonstrate that the control policies computed by probabilistic inference in $\mathcal{R}$-mixture model are optimal with respect to the original expected cumulative future reward: $\max L(\theta \,|\, R = \text{true}) \propto \max \mathcal{V}(\pi(\mathcal{A}, \mathcal{S}))$. The original task of optimizing POMDP controllers can now be approached by maximizing this objective likelihood of observing reward in a mixture-$\mathcal{DBN}$ generative model.

Novel single-$\mathcal{DBN}$ model for policy optimization by maximizing a single-observation likelihood

As a compelling alternative to modeling the value function by the composite $\mathcal{R}$-mixture model of the reward likelihood, we developed a novel single-$\mathcal{DBN}$ "$V - D$" generative model, which now allows (1) for application of various techniques for probabilistic inference in single graphical models as usual (planning by maximizing a single observation likelihood), and (2) for exploiting the factored structure in a controller architecture and for taking advantage of natural structural constrains of planning problems.

Particularly, to ensure the equivalence to the original problem of policy optimization, we additionally introduce two new sets of special nodes into the single-$\mathcal{DBN}$ graphical model, namely "D" (discount factor) and "V" (partial value or reward accumulation) binary random variables (figure 1). Thus, the discount factor "D" node represents a binary random variable, which $\Pr(D = \text{true})$ is set to be proportional to the discount rate for the current time step. Similarly, the special reward accumulation node "V"

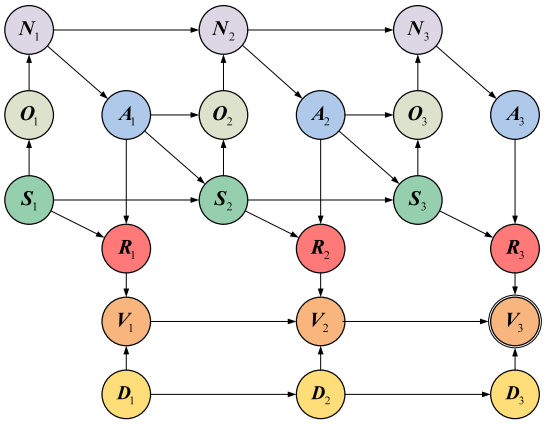

Figure 1: Single-$\mathcal{DBN}$ model for POMDP planning

represents a binary stochastic variable, which $\mathcal{CPT}$ is defined by a special additive representation such that it is proportional to the sum of all discounted rewards, accumulated till the current time step, and it is equal to one only if maximum discounted reward is received now and accumulated before: $\Pr(V_i = \text{true} \,|\, V_{i-1}, R_i, D_{i-1}) = \Theta(R_i, D_{i-1}) + \Psi(V_{i-1})$. Importantly, we formally prove that the task of finding the optimal control policy π^* of the original planning problem by maximizing its value function is equivalent to the task of finding the optimal policy parameter θ by maximizing the likelihood of observing the single evidence at the last time step of the single-$\mathcal{DBN}$ "$V - D$" inference model: $\max_{\boldsymbol{\theta}} P(V^{\boldsymbol{\theta}}_{t=H+1} = \text{true}) \propto \max_{\pi} \mathcal{V}(\pi, t = t_0)$ (refer to supplementary materials for details). Establishing this equivalence makes it further possible to maximize the objective function and compute the optimal policy parameter θ approximately using the Expectation-Maximization algorithm by varying the distributions $p(N_0), \pi_t, \lambda_t$ encoding the control policy.

Inference algorithms for computing optimal policy

To solve a planning task by probabilistic inference in the single-$\mathcal{DBN}$ inference "$V - D$" models, we implemented several inference methods to approximate the value of the policy due to the fact that exact methods are not tractable for complex (large-size) graphical model with cycles: (1) We initially developed the exact frequency-based EM algorithm with the integrated forward-backward procedure (based on Baum-Welch algorithm) to optimize policies of POMDP problems (for directed models). (2) We investigated some existing alternatives to the exact forward-backward algorithm by analyzing such exact methods as sum-product and belief-propagation algorithms (for undirected factor-graph models) (Murphy 2002). (3) We applied the Boyen-Koller approximation and found that specifically for our single-$\mathcal{DBN}$ inference "$V - D$" model, the resulting approximation may not provide a desirable complexity reduction in comparison to exact inference methods. (4) We implemented a frequency-based EM algorithm, where we applied a special Loopy Belief Propagation (LBP) algorithm to find an estimate for the maximum likelihood of the model parameter and to approximate the value of the policy. Conducted experimental evaluation demonstrates that in some cases divergence may occur due to deterministic dependencies in small cycles of a general LBP algorithm. Thus, we developed a modified version of the LBP algorithm with a special belief update protocol for our single-$\mathcal{DBN}$ inference model to overcome the cycling problems and to approximate the computation of the intractable E-step posterior. (5) We additionally proposed a different approximate inference scheme and implemented a version of the Factored Frontier (FF) algorithm (clique-tree) (Murphy and Weiss 2001), which approximation assumptions are more aggressive. The developed FF algorithm implements a forward-inference procedure that computes exact marginals in the next time period subject to a factored approximation of the previous time period. We also derived the backward-pass propagation procedure in addition to existing definition of the forward-pass propagation rules. The developed FF algorithm has the advantage of exploiting the factored structure of the involved $\mathcal{DBN}$s to derive the formulas for these marginals.

Conclusions and future directions

Due to the fact that exact inference methods (EM algorithm) for policy optimization by likelihood maximization are not tractable to solve large-size problems with complex graphical model with cycles, we implemented several alternative inference methods to approximate the value of the control policy by computing the intractable E-step posterior approximately. The proposed approaches to policy optimization by probabilistic inference are evaluated on several POMDP benchmark problems and the performance of the implemented approximation algorithms is compared. Nevertheless, conducted experimental work demonstrates that since planning by inference in partially observable stochastic domains essentially tackles a non-convex optimization problem, the important problem still remains that approximation algorithms do not get rid of local optima issues and in some cases they can still stuck at a sub-optimal configuration (stationary point). To address this issue, we outline our ongoing work on developing alternative approximate inference methods for policy optimization that can provide bounded algorithmic performance guarantees and estimate the quality of approximation (Kiselev and Poupart 2014).

References

Kiselev, I., and Poupart, P. 2014. Policy optimization by marginal-MAP probabilistic inference in generative models. In *AAMAS 2014*. IFAAMAS.

Murphy, K. P., and Weiss, Y. 2001. The factored frontier algorithm for approximate inference in DBNs. In Breese, J. S., and Koller, D., eds., *UAI*, 378–385. Morgan Kaufmann.

Murphy, K. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Dissertation, UC Berkeley, Computer Science Division.

Toussaint, M.; Charlin, L.; and Poupart, P. 2008. Hierarchical POMDP controller optimization by likelihood maximization. In *UAI*, 562–570. AUAI Press.