

Inference strategies for solving semi-Markov decision processes

Matthew Hoffman Peter Carbonetto Nando de Freitas Arnaud Doucet

Department of Computer Science
University of British Columbia
Vancouver, B.C., Canada V6T 1Z4

Semi-Markov decision processes (SMDPs) generalize standard MDPs to domains where time is not discretized equally between every set of states and actions [3]. Instead we can define a jump-Markov process where the amount of time spent in each state is a stochastic random variable. This formulation gives us an intuitive way to reason about actions where it is also necessary to take into account how long these actions will take to perform.

Formally we can define an SMDP as a continuous-time controlled stochastic process $(x(t), u(t))$ consisting, respectively, of states and actions at every point in time t where state transitions occur at random *arrival times* T_n . In particular, the process is stationary in between jumps, i.e. $x(t) = x_n$ and $u(t) = u_n$ for all $t \in [t_n, t_{n+1})$. Here each of the state/action pairs (x_n, u_n) evolve according to the same model as a standard MDP, i.e.

- an initial state model $\mu(x_0)$,
- a state transition model $f(x_{n+1}|x_n, u_n)$,
- and finally a stochastic policy $\pi_\theta(u_n|x_n)$.

The amount of time spent in each state is controlled by the *sojourn times* S_n , distributed as $T(s_n|x_n, u_n)$, where the arrival times are given by $t_n = t_{n-1} + s_n$ and $t_0 = 0$. See Fig. 1 for an illustration of this process. The objective of this problem is then to optimize the expected, discounted reward with respect to the policy parameters θ .

In this work we show how to formalize this problem as one of inference (based on work in [1, 4]), and in particular provide an analytic, model-based solution which is similar to learning in mixture-models. As in earlier work, this connection is made by interpreting the discount factors as a probability distribution over

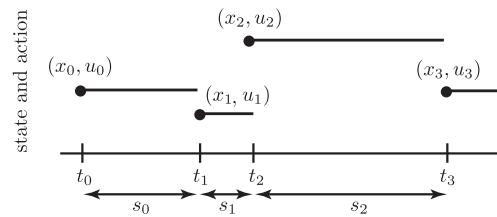


Figure 1: Relationship between arrival times t_n , sojourn times s_n , and the system state (x_n, u_n) .

random trajectory lengths, but in a departure from earlier methods we show that it is not possible to formulate this problem as a mixture-model. This occurs because the discounting terms for an SMDP are a function of the amount of time spent in the world and thus the probability over path lengths will depend on all possible sojourn times. We instead show how to properly formulate this problem by working directly with the *joint distribution* over trajectories and trajectory lengths.

In situations where the analytic solution is intractable, or where the models are unknown, we also provide an approximate sample-based solution using a forward sweep through the SMDP, similar to [2, 5]. Finally, we will give preliminary examples of these methods, as well as discuss possible extensions to more general time-dependent decision processes.

References

- [1] M. Hoffman, N. de Freitas, A. Doucet, and J. Peters. An expectation maximization algorithm for continuous Markov decision processes with arbi-

- trary reward. In *Artificial Intelligence and Statistics*, 2009.
- [2] J. Kober and J. Peters. Policy search for motor primitives in robotics. In *Advances in Neural Information Processing Systems*, 2008.
 - [3] M. Puterman. *Markov decision processes*. Wiley-Interscience, 1994.
 - [4] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state Markov decision processes. In *Proceedings of the International Conference on Machine Learning*, 2006.
 - [5] N. Vlassis and M. Toussaint. Model-free reinforcement learning as mixture learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.