RLEMMC Theory Cheat Sheet

Orhan Sönmez

August 4, 2017

1 Problem Definition

1.1 Markov Decision Processes

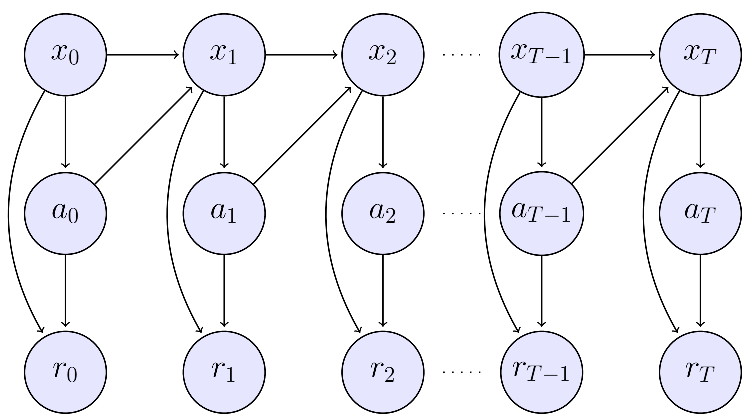

Figure 1: Graphical model of an MDP with an horizon of T.

Init Model	$x_0 \sim P(x_0)$	
Action Selection	$a_t \sim P(a_t \vert x_t; \pi)$	$t = 0, ..., T$
Transition Model	$x_{t+1} \sim P(x_{t+1} \vert x_t, a_t)$	$t = 0, ..., T-1$
Reward Function	$r_t \sim P(r_t \vert x_t, a_t)$	$t = 0, ..., T$

Full Joint Distribution of an MDP:

$$P(x_{0:T}, a_{0:T}; \pi) = P(x_0)\Big[\prod_{t=0}^{T-1} P(a_t|x_t;\pi)P(x_{t+1}|x_t, a_t)\Big]P(a_T|x_T;\pi)$$

where $P(r_t|x_t, a_t) = \delta_{R(x_t,a_t)}(r_t)$.

1.2 Optimal Policy

Value Function $$V(\pi) = \left\langle \sum_{t=0}^{T} \gamma^t r_t \right\rangle_{P(x_{0:T},a_{0:T};\pi)}$$

Optimal Policy $$\pi^* = \arg\max_{\pi} V(\pi)$$

2 Probabilistic Approach [2]

2.1 Mixture Model

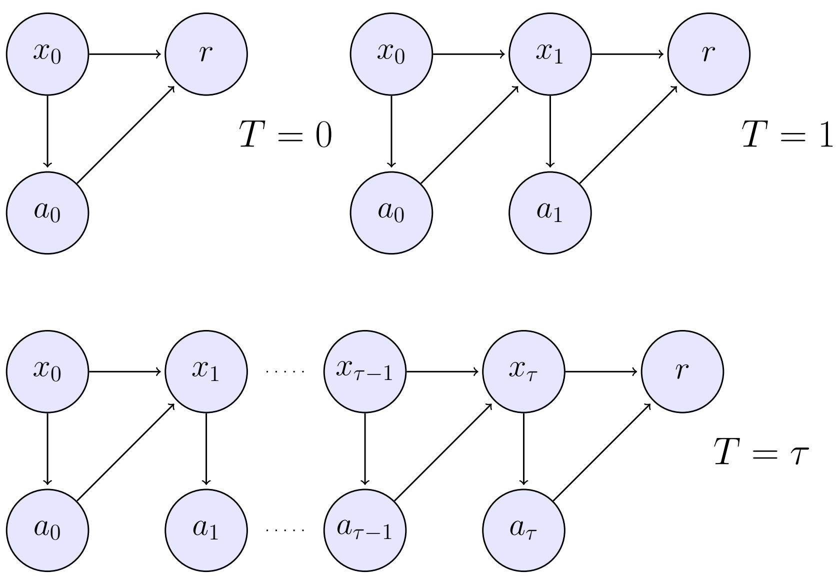

Figure 2: Graphical model of the mixture model

Full Joint Distribution of a Finite-time MDP:

$$P(r, x_{0:T}, a_{0:T}|T;\pi) = P(x_0)\Big[\prod_{t=0}^{T-1} P(a_t|x_t;\pi)P(x_{t+1}|x_t,a_t)\Big]P(a_T|x_T;\pi)P(r|x_t,a_t)$$

Full Mixture of Finite-Time MDPs:

$$P(r, x_{0:T}, a_{0:T}, T;\pi) = P(r, x_{0:T}, a_{0:T}|T;\pi)P(T)$$

where $P(T) = \gamma^T(1-\gamma)$.

2.2 Equivalent Probabilistic Inference Problem

Likelihood on the Mixture Model:

$$\mathcal{L}(\pi) = P(r=1;\pi)$$

where $r \in [0,1]$.

Equivalent Problem:

$$L(\pi) = (1-\gamma)V(\pi)$$

3 RLEMMC

3.1 Expectation-Maximization Update Rule: [2]

$$\pi^{(k)} \leftarrow \arg\max_{\pi^{(k)}} \big\langle \log P(r=1, x_{0:T}, a_{0:T}, T;\pi^{(k)})\big\rangle_{P(x_{0:T},a_{0:T},T|r=1;\pi^{(k-1)})}$$

in order to maximize $\mathcal{L}(\pi)$.

3.2 Monte Carlo E-Step

Monte Carlo E-step Approximation:

$$\big\langle \log P(r=1, x_{0:T}, a_{0:T}, T;\pi^{(k)})\big\rangle_{P(x_{0:T},a_{0:T},T|r=1;\pi^{(k-1)})} \approx \sum_{s=1}^{S} \log P(r=1, x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)};\pi^{(k)})$$

where sample $(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}) \sim P(x_{0:T}, a_{0:T}, T|r=1;\pi^{(k-1)})$.

3.3 M-step

$$\pi^{(k)} \leftarrow \arg\max_{\pi} \frac{1}{S} \sum_{s=1}^{S} \sum_{t=0}^{T^{(s)}} \log P(a_t^{(s)} | x_t^{(s)}; \pi^{(k)}) \qquad \text{(Terms related with } \pi^{(k)}\text{)} \tag{1}$$

means that any policy $\pi^{(k)}$ such that $\forall t \forall s\ \pi^{(k)}(x_t^{(s)}) = a_t^{(s)}$ is a maximizer.

Algorithm:

Algorithm 1 RLEMMC

Initial Policy $\pi \leftarrow$ Uniform policy
while π not converged **do**
  E-step: Sample $x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}$ from posterior $P(x_{0:T}, a_{0:T}, T | r = 1; \pi)$
  M-step: Policy Learning using samples $x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}$
end while
Optimal Policy: $\pi^* \leftarrow \pi$

4 Importance Sampling E-step [1]

$$\begin{aligned} &\textbf{Target Distribution} \qquad P(x_{0:T}, a_{0:T}, T | r = 1; \pi) \qquad \text{(Posterior)} \\ &\textbf{Proposal Distribution} \qquad P(x_{0:T}, a_{0:T}, T; \pi) \qquad \text{(Prior)} \end{aligned}$$

Bayes Rule

$$P(x_{0:T}, a_{0:T}, T | r = 1; \pi) \propto P(r = 1 | x_{0:T}, a_{0:T}, T; \pi) P(x_{0:T}, a_{0:T}, T; \pi)$$

Weight Function

$$\begin{aligned} w(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}) &= \frac{P(r = 1 | x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}; \pi) P(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}; \pi)}{P(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}; \pi)} \\ &= P(r = 1 | x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}; \pi) \end{aligned}$$

Normalized Weight Function

$$W(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)}) = \frac{w(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)})}{\sum_s w(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)})}$$

Resampling Resampling with weights $W(x_{0:T}^{(s)}, a_{0:T}^{(s)}, T^{(s)})$ to have an unweighted Monte Carlo estimate.

References

[1] Orhan Sönmez and A. Taylan Cemgil. Modele Dayal Pekitirme ile Örenme için Önem Örneklemesi (Importance Sampling for Model-Based Reinforcement Learning). In *Proceedings of 20th IEEE Signal Processing ve Communication Applications Conference (SIU)*, 2012.

[2] Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 945–952, New York, New York, USA, 2006. ACM.