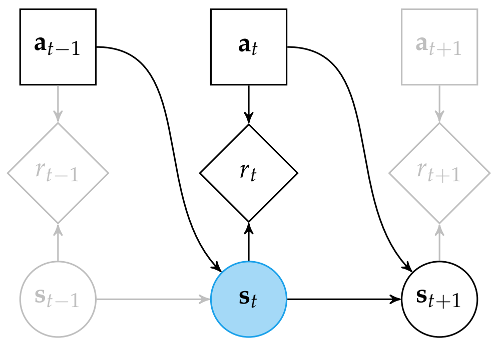
Markov Decision Process: Chain Rule

Robert Moss
Department of Computer Science
MOSSR@CS.STANFORD.EDU
Stanford University

1. MDP Graphical Model

The graphical model of a *Markov decision process* (MDP) illustrates the dependencies between states $\mathbf{s}$, actions $\mathbf{a}$, and rewards r at time t. The *Markov property* holds if the conditional probability distribution of future states dependents only on the current state and action, and not the previous sequence (i.e. trajectory) of states and actions.



A trajectory τ is a sequence of states and actions up to some time T, where $\tau = \langle \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T \rangle$. The probability of a particular trajectory using a policy π_θ that is parameterized by θ is denoted $p_\theta(\tau)$, where

$$\underbrace{p_\theta(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)}_{p_\theta(\tau)} = p(\mathbf{s}_1) \prod_{t=1}^{T} \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t). \tag{1}$$

Using the Markov property of an MDP, we can derive equation (1) using the chain rule. Recall that the definition of conditional probability for two events E and F can be written as $P(E, F) = P(E \mid F)P(F)$, which we call the *chain rule*. So the general form of the chain rule for n events can be written as:

$$P(E_1, E_2, \ldots, E_n) = P(E_1)P(E_2 \mid E_1) \cdots P(E_n \mid E_1, E_2, \ldots, E_{n-1}) \tag{2}$$

$$= \prod_{i=1}^{n} P\left(E_i \,\middle|\, \bigcap_{k=1}^{i-1} E_k\right) \tag{3}$$

Now recall that $\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)$ is the probability of taking action $\mathbf{a}_t$ from state $\mathbf{s}_t$ using policy π parameterized by θ. Now we can derive equation (1):

$$\begin{aligned}
p_\theta(\tau) &= p_\theta(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T) && \text{(definition of trajectory)}\\
&= p(\mathbf{s}_1)\pi_\theta(\mathbf{a}_1 \mid \mathbf{s}_1) \cdots p(\mathbf{s}_T \mid \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_{T-1}, \mathbf{a}_{T-1})\pi_\theta(\mathbf{a}_T \mid \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T) && \text{(chain rule)}\\
&= p(\mathbf{s}_1)\pi_\theta(\mathbf{a}_1 \mid \mathbf{s}_1) \cdots p(\mathbf{s}_T \mid \mathbf{s}_{T-1}, \mathbf{a}_{T-1})\pi_\theta(\mathbf{a}_T \mid \mathbf{s}_T) && \text{(Markov property)}\\
&= p(\mathbf{s}_t) \prod_{t=1}^{T} \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) && \text{(generalized for time } T\text{)}
\end{aligned}$$