# Bayesian reinforcement learning and partially observable Markov decision processes

Christos Dimitrakakis

EPFL

November 6, 2013

# Summary of previous developments

- Probability and utility.
- Making decisions under uncertainty.
- Updating probabilies
- Optimal experiment design
- Markov decision processes
- Stochastic algorithms for Markov decision processes.
- MDP Approximations.
- Bayesian reinforcement learning

# Summary of previous developments

- Probability and utility.
- Making decisions under uncertainty.
- Updating probabilies
- Optimal experiment design
- Markov decision processes
- Stochastic algorithms for Markov decision processes.
- MDP Approximations.
- Bayesian reinforcement learning

## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

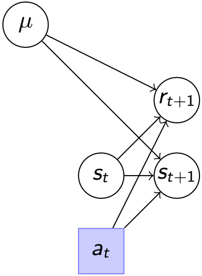
## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

## Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in \mathcal{S}$.
- Take action $a_t \in \mathcal{A}$.
- Receive reward $r_t \in \mathbb{R}$.



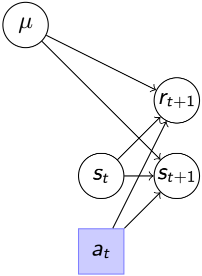
## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

## Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in \mathcal{S}$.
- Take action $a_t \in \mathcal{A}$.
- Receive reward $r_t \in \mathbb{R}$.




## The optimal policy for a given $w$

- Find policy $\pi : \mathcal{S} \to \mathcal{A}$ maximising the utility $U = \sum_t r_t$ in expectation.

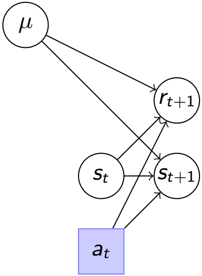
## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

## Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in \mathcal{S}$.
- Take action $a_t \in \mathcal{A}$.
- Receive reward $r_t \in \mathbb{R}$.




## The optimal policy for a given $w$

- Find policy $\pi : \mathcal{S} \to \mathcal{A}$ maximising the utility $U = \sum_t r_t$ in expectation.
- When $w$ is known, use standard algorithms, such as value or policy iteration. However this is

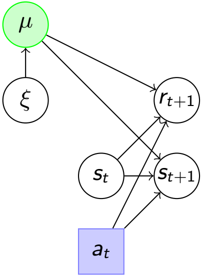
## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

### Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in \mathcal{S}$.
- Take action $a_t \in \mathcal{A}$.
- Receive reward $r_t \in \mathbb{R}$.




### The optimal policy for a given $w$

- Find policy $\pi : \mathcal{S} \to \mathcal{A}$ maximising the utility $U = \sum_t r_t$ in expectation.
- When $w$ is known, use standard algorithms, such as value or policy iteration. However this is contrary to the problem definition!

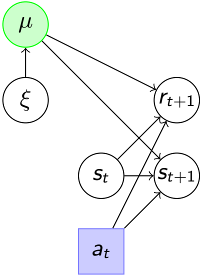
## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

## Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in \mathcal{S}$.
- Take action $a_t \in \mathcal{A}$.
- Receive reward $r_t \in \mathbb{R}$.




## Bayesian RL: Use a subjective belief $\xi(\mu)$

$$\mathbb{E}(U \mid \pi, \xi)$$

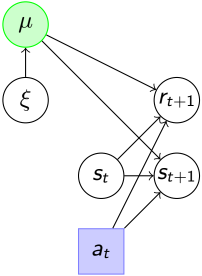
## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

### Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in \mathcal{S}$.
- Take action $a_t \in \mathcal{A}$.
- Receive reward $r_t \in \mathbb{R}$.




### Bayesian RL: Use a subjective belief $\xi(\mu)$

$$\mathbb{E}(U \mid \pi, \xi) = \sum_{\mu} \mathbb{E}(U \mid \pi, \mu)\xi(\mu)$$

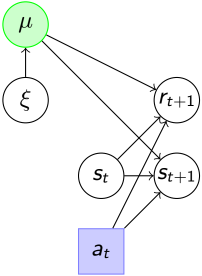
## The reinforcement learning problem

Learning to act in an unknown environment, by interaction and reinforcement.

### Markov decision processes (MDP)

We are in some environment $\mu$, where at each time step $t$:

- Observe state $s_t \in \mathcal{S}$.
- Take action $a_t \in \mathcal{A}$.
- Receive reward $r_t \in \mathbb{R}$.




### Bayesian RL: Use a subjective belief $\xi(\mu)$

Not actually easy as $\pi$ must now map from complete histories to actions.

$$U_\xi^* = \max_\pi \mathbb{E}(U \mid \pi, \xi) = \max_\pi \sum_\mu \mathbb{E}(U \mid \pi, \mu)\xi(\mu)$$

Planning must take into account future learning

# Updating the belief

## Example

When the number of MDPs is finite

## Exercise

*Another practical scenario is when we have an independent belief over the transition probabilities of each state-action pair. Consider the case where we have n states and k actions. Similar to the product-prior in the bandit exercise of exercise set 4, we assign a probability (density)* $\xi_{s,a}$ *to the probability vector* $\boldsymbol{\theta}_{(s,a)} \in \mathbb{S}^n$*. We can then define our joint belief on the* $(nk) \times n$ *matrix* $\boldsymbol{\Theta}$ *to be*

$$\xi(\boldsymbol{\Theta}) = \prod_{s \in \mathcal{S}, a \in \mathcal{A}} \xi_{s,a}(\boldsymbol{\theta}_{(s,a)}).$$

*Derive the updates for a product-Dirichlet prior on transitions and a product-Normal-Gamma prior on rewards.*
*What is the meaning of using a Normal-Wishart prior on rewards?*

# The expected MDP heuristic

1. For a given belief $\xi$, calculate the expected MDP:

$$\bar{\mu}_\xi \triangleq \mathbb{E}_\xi \, \mu.$$

2. Calculate the optimal memoryless policy for $\bar{\mu}_\xi$:

$$\pi^*(\bar{\mu}_\xi) \in \arg\max_{\pi \in \Pi_1} V^\pi_{\bar{\mu}_\xi},$$

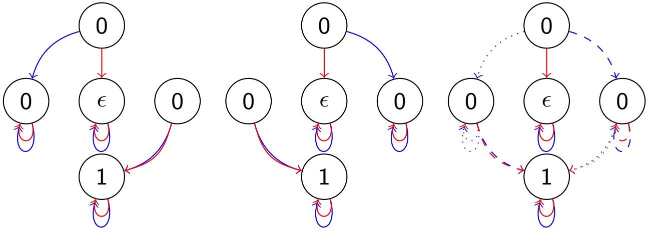
where $\Pi_1 = \left\{\pi \in \Pi \;\middle|\; \mathbb{P}_\pi(a_t \mid s^t, a^{t-1}) = \mathbb{P}_\pi(a_t \mid s_t)\right\}$.

3. Execute $\pi^*(\bar{\mu}_\xi)$.

**Problem**

Unfortunately, this approach may be far from the optimal policy in $\Pi_1$.

# Counterexample[1]




(a) $\mu_1$ (b) $\mu_2$ (c) $\bar{\mu}_\xi$

Figure: $\mathcal{M} = \{\mu_1, \mu_2\}$, $\xi(\mu_1) = \theta$, $\xi(\mu_2) = 1 - \theta$, deterministic transitions.

- For $T \to \infty$, the $\bar{\mu}_\xi$-optimal policy is not optimal in $\Pi_1$ if:

$$\epsilon < \frac{\gamma\theta(1-\theta)}{1-\gamma}\left(\frac{1}{1-\gamma\theta} + \frac{1}{1-\gamma(1-\theta)}\right)$$

- In this example, $\bar{\mu}_\xi \notin \mathcal{M}$.
- For smooth beliefs, $\bar{\mu}_\xi$ is close to $\hat{\mu}^*_\xi$.

[1] Based on one by Remi Munos

# Counterexample for $\hat{\mu}^*_\xi \triangleq \arg\max_\mu \xi(\mu)$

MDP set $\mathcal{M} = \{\mu_i \mid i = 1, \ldots, n\}$ with $\mathcal{A} = \{0, \ldots, n\}$. In all MDPs, $a_0$ gives you a reward of $\epsilon$ and the MDP terminates. In the $i$-th MDP, all other actions give you a reward of 0 apart from the $i$-th action which gives you a reward of 1.

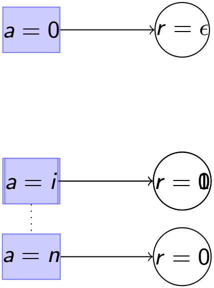



Figure: The MDP $\mu_i$.

- The $\xi$-optimal policy takes action $i$ iff $\xi(\mu_i) \geq \epsilon$, otherwise takes action 0.
- The $\hat{\mu}^*_\xi$-optimal policy takes $a = \arg\max_i \xi(\mu_i)$.

# Policy evaluation

## Expected utility of a policy $\pi$ for a belief $\xi$

$$V_\xi^\pi \triangleq \mathbb{E}(U \mid \xi, \pi) \tag{2.1}$$
$$= \int_{\mathcal{M}} \mathbb{E}(U \mid \mu, \pi) \, \mathrm{d}\xi(\mu) \tag{2.2}$$
$$= \int_{\mathcal{M}} V_\mu^\pi \, \mathrm{d}\xi(\mu) \tag{2.3}$$

## Bayesian Monte-Carlo policy evaluation

**input** policy $\pi$, belief $\xi$
**for** $k = 1, \ldots, K$ **do**
  $\mu_k \sim \xi$.
  $\boldsymbol{v}_k = V_{\mu_k}^\pi$
**end for**
$\boldsymbol{u} = \frac{1}{K} \sum_{k=1}^K \boldsymbol{v}_k$.
**return** $\boldsymbol{u}$.

## Upper bounds on the utility for a belief $\xi$

$$V_\xi^* \triangleq \sup_\pi \mathbb{E}(U \mid \xi, \pi) = \sup_\pi \int_{\mathcal{M}} \mathbb{E}(U \mid \mu, \pi) \, \mathrm{d}\xi(\mu) \tag{2.4}$$

$$\leq \int_{\mathcal{M}} \sup_\pi V_\mu^\pi \, \mathrm{d}\xi(\mu) = \int_{\mathcal{M}} V_\mu^* \, \mathrm{d}\xi(\mu) \triangleq V_\xi^+ \tag{2.5}$$

## Bayesian Monte-Carlo upper bound

**input** policy $\pi$, belief $\xi$
**for** $k = 1, \ldots, K$ **do**
  $\mu_k \sim \xi$.
  $\boldsymbol{v}_k = V_{\mu_k}^*$
**end for**
$\boldsymbol{u}^* = \frac{1}{K} \sum_{k=1}^K \boldsymbol{v}_k$.
**return** $\boldsymbol{u}^*$.

# Bounds on $V^*_\xi \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

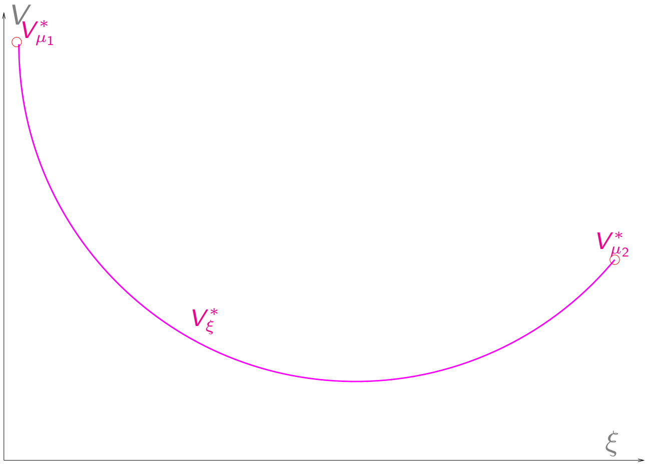



Figure: A geometric view of the bounds

# Bounds on $V^*_\xi \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

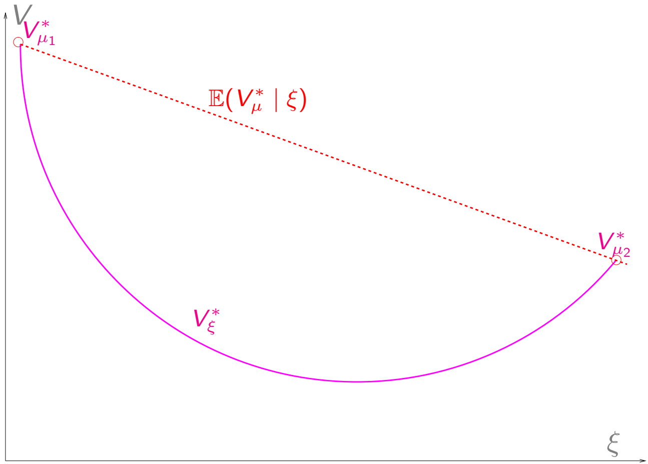



Figure: A geometric view of the bounds

# Bounds on $V^*_\xi \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

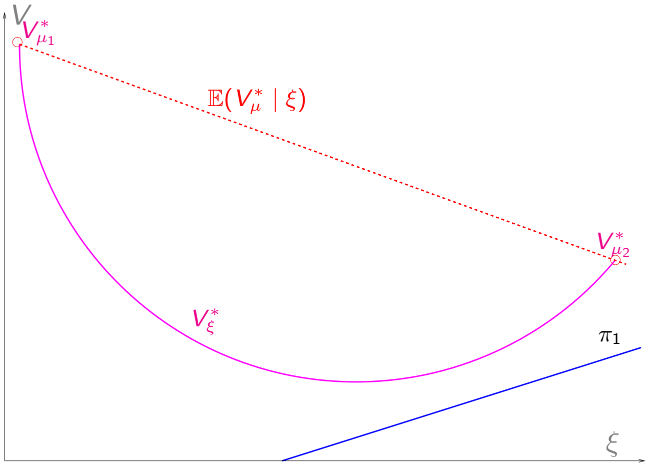



Figure: A geometric view of the bounds

# Bounds on $V^*_\xi \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

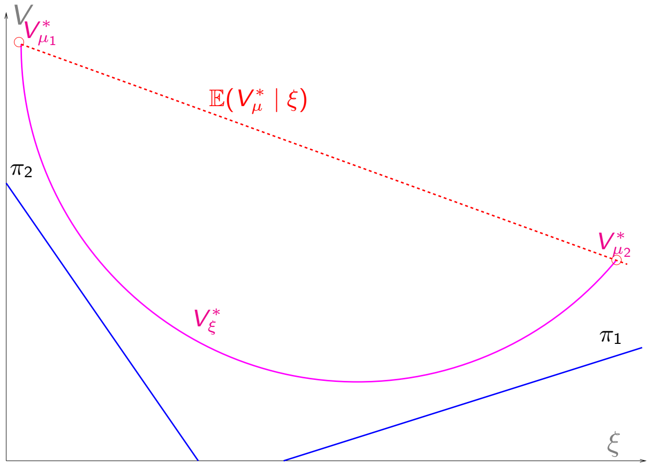



Figure: A geometric view of the bounds

# Bounds on $V^*_\xi \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

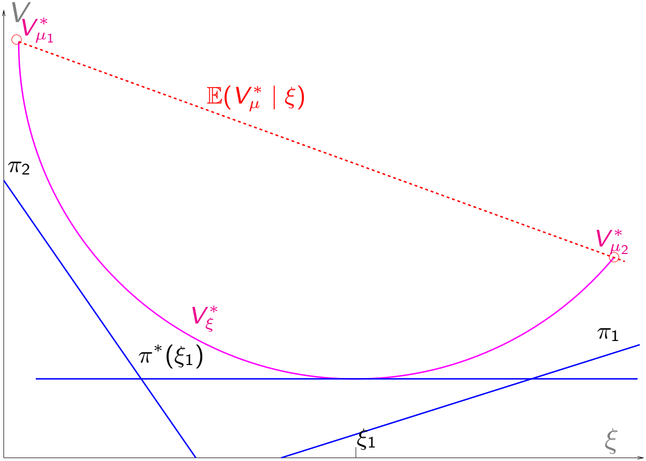



Figure: A geometric view of the bounds

# Bounds on $V^*_\xi \triangleq \max_\pi \mathbb{E}(U \mid \pi, \xi)$

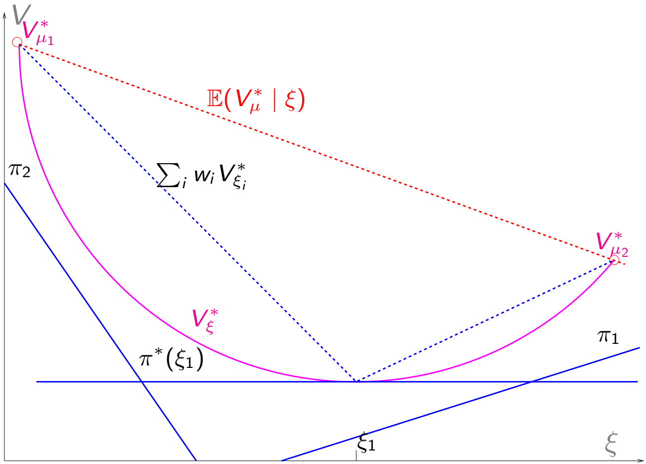
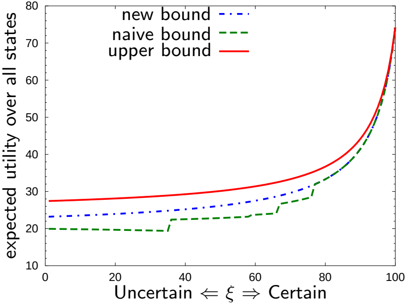



Figure: A geometric view of the bounds

# Better lower bounds [? ]



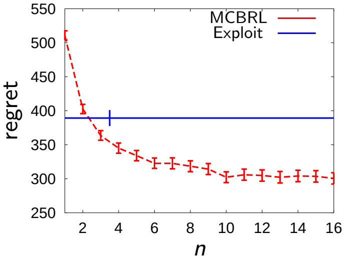

## Main idea: maximisation in memoryless policies

- Then we can assume a fixed belief.
- Backwards induction on $n$ MDPs
- This improves the naive lower bound.

$$Q^{\pi}_{\xi,t}(s,a) \triangleq \int_{\mathcal{M}} \left\{ \bar{R}_{\mu}(s,a) + \gamma \int_{\mathcal{S}} V^{\pi}_{\mu,t+1}(s') \, \mathrm{d}\mathcal{T}^{s,a}_{\mu}(s') \right\} \mathrm{d}\xi(\mu) \tag{2.6}$$

## Multi-MDP Backwards Induction

MMBI$\mathcal{M}, \xi, \gamma, T$<br>
Set $V_{\mu,T+1}(s) = 0$ for all $s \in \mathcal{S}$.<br>
**for** $t = T, T-1, \ldots, 0$ **do**<br>
    **for** $s \in \mathcal{S}, a \in \mathcal{A}$ **do**<br>
        Calculate $Q_{\xi,t}(s,a)$ from (2.6) using $\{V_{\mu,t+1}\}$ .<br>
    **end for**<br>
    **for** $s \in \mathcal{S}$ **do**<br>
        $a^*_{\xi,t}(s) \in \arg\max_{a \in \mathcal{A}} Q_{\xi,t}(s,a)$.<br>
        **for** $\mu \in \mathcal{M}$ **do**<br>
            $V_{\mu,t}(s) = Q_{\mu,t}(s, a^*_{\xi,t}(s))$.<br>
        **end for**<br>
    **end for**<br>
**end for**




## MCBRL: Application to Bayesian RL

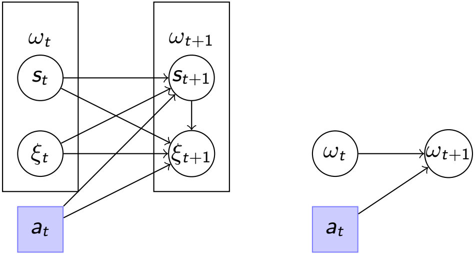
1. For $i = 1, \ldots$
2. At time $t_i$, sample $n$ MDPs from $\xi_{t_i}$.
3. Calculate best memoryless policy $\pi_i$ wrt the sample.
4. Execute $\pi_i$ until $t = t_{i+1}$.

## Relation to other work

- For $n = 1$, this is equivalent to the Thompson sampling used by Strens [**?** ].
- Unlike BOSS [**?** ] it does not become more optimistic as $n$ increases.
- BEETLE[**? ?** ] is a belief-sampling approach.
- Furmston and Barber [**?** ] use approximate inference to estimate policies.

# Generalisations

- Policy search for improving lower bounds.
- Search enlarged class of policies
- Examine all history-based policies.




(a) The complete MDP model

(b) Compact form of the model

Figure: Belief-augmented MDP

## The augmented MDP

The optimal policy for the augmented MDP is the $\xi$-optimal for the original problem.

$$P(s_{t+1} \in S \mid \xi_t, s_t, a_t) \triangleq \int_S P_\mu(s_{t+1} \in S \mid s_t, a_t) \, \mathrm{d}\xi_t(\mu) \tag{2.7}$$

$$\xi_{t+1}(\cdot) = \xi_t(\cdot \mid s_{t+1}, s_t, a_t) \tag{2.8}$$

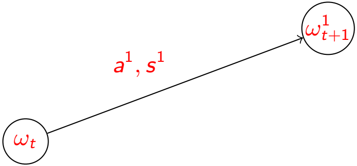
# Belief-augmented MDP tree structure

Consider an MDP family $\mathcal{M}$ with $\mathcal{A} = \{a^1, a^2\}$, $\mathcal{S} = \{s^1, s^2\}$.


$$\omega_t = (s_t, \xi_t)$$

# Belief-augmented MDP tree structure

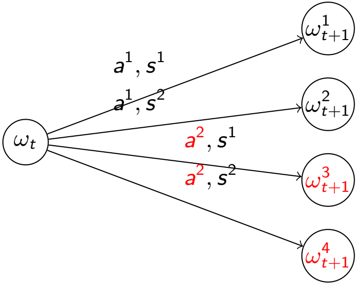
Consider an MDP family $\mathcal{M}$ with $\mathcal{A} = \{a^1, a^2\}$, $\mathcal{S} = \{s^1, s^2\}$.




$$\omega_t = (s_t, \xi_t)$$

# Belief-augmented MDP tree structure

Consider an MDP family $\mathcal{M}$ with $\mathcal{A} = \{a^1, a^2\}$, $\mathcal{S} = \{s^1, s^2\}$.




$$\omega_t = (s_t, \xi_t)$$

# Branch and bound

## Value bounds

Let upper and lower bounds $\boldsymbol{q}^+$ and $\boldsymbol{q}^-$ such that:

$$\boldsymbol{q}^+(\omega, a) \geq Q^*(\omega, a) \geq \boldsymbol{q}^-(\omega, a) \tag{2.9}$$

$$\boldsymbol{v}^+(\omega) = \max_{a \in \mathcal{A}} Q^+(\omega, a), \qquad \boldsymbol{v}^-(\omega) = \max_{a \in \mathcal{A}} Q^-(\omega, a). \tag{2.10}$$

$$\boldsymbol{q}^+(\omega, a) = \sum_{\omega'} p(\omega' \mid \omega, a) \left[ r(\omega, a, \omega') + V^+(\omega') \right] \tag{2.11}$$

$$\boldsymbol{q}^-(\omega, a) = \sum_{\omega'} p(\omega' \mid \omega, a) \left[ r(\omega, a, \omega') + V^-(\omega') \right] \tag{2.12}$$

## Remark

*If $\boldsymbol{q}^-(\omega, a) \geq \boldsymbol{q}^+(\omega, b)$ then $b$ is sub-optimal at $\omega$.*

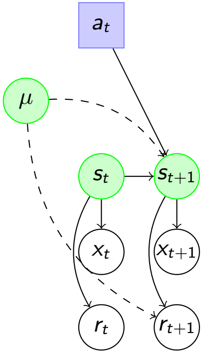
## Stochastic branch and bound for belief tree search [? ? ]

- (Stochastic) Upper and lower bounds on the values of nodes (via Monte-Carlo sampling)
- Use upper bounds to expand tree, lower bounds to select final policy.
- Sub-optimal branches are quickly discarded.

## Partially observable Markov decision processes (POMDP)

When acting in $\mu$, each time step $t$:

- The system state $s_t \in \mathcal{S}$ is not observed.
- We receive an observation $x_t \in \mathcal{X}$ and a reward $r_t \in \mathcal{R}$.
- We take action $a_t \in \mathcal{A}$.
- The system transits to state $s_{t+1}$.




## Definition

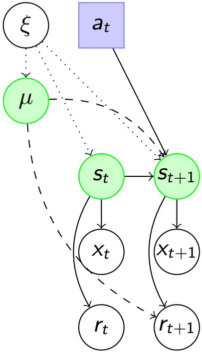
Partially observable Markov decision process (POMDP) A POMDP $\mu \in \mathcal{M}_P$ is a tuple $(\mathcal{X}, \mathcal{S}, \mathcal{A}, P)$ where $\mathcal{X}$ is an observation space, $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, and $P$ is a conditional distribution on observations, states and rewards. The following Markov property holds:

$$\mathbb{P}_\mu(s_{t+1}, r_t, x_t \mid s_t, a_t, \ldots) = P(s_{t+1} \mid s_t, a_t) P(x_t \mid s_t) P(r_t \mid s_t) \tag{3.1}$$

## Partially observable Markov decision processes (POMDP)

When acting in $\mu$, each time step $t$:

- The system **state** $s_t \in \mathcal{S}$ is not observed.
- We receive an **observation** $x_t \in \mathcal{X}$ and a **reward** $r_t \in \mathcal{R}$.
- We take **action** $a_t \in \mathcal{A}$.
- The system transits to state $s_{t+1}$.




### Definition

Partially observable Markov decision process (POMDP) A POMDP $\mu \in \mathcal{M}_P$ is a tuple $(\mathcal{X}, \mathcal{S}, \mathcal{A}, P)$ where $\mathcal{X}$ is an observation space, $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, and $P$ is a conditional distribution on observations, states and rewards. The following Markov property holds:

$$\mathbb{P}_\mu(s_{t+1}, r_t, x_t \mid s_t, a_t, \ldots) = P(s_{t+1} \mid s_t, a_t) P(x_t \mid s_t) P(r_t \mid s_t) \tag{3.1}$$

# Belief state in POMDPs when $\mu$ is known

If $\mu$ defines starting state probabilities, then the belief is not subjective

## Belief $\xi$

For any distribution $\xi$ on $\mathcal{S}$, we define:

$$\xi(s_{t+1} \mid a_t, \mu) \triangleq \int_{\mathcal{S}} P_\mu(s_{t+1} \mid s_t a_t) \, \mathrm{d}\xi(s_t) \tag{3.2}$$

## Belief update

$$\xi_t(s_{t+1} \mid x_{t+1}, r_{t+1}, a_t, \mu) = \frac{P_\mu(x_{t+1}, r_{t+1} \mid s_{t+1}) \xi_t(s_{t+1} \mid a_t, \mu)}{\xi_t(x_{t+1} \mid a_t, \mu)} \tag{3.3}$$

$$\xi_t(s_{t+1} \mid a_t, \mu) = \int_{\mathcal{S}} P_\mu(s_{t+1} \mid s_t, a_t, \mu) \, \mathrm{d}\xi_t(s_t) \tag{3.4}$$

$$\xi_t(x_{t+1} \mid a_t, \mu) = \int_{\mathcal{S}} P_\mu(x_{t+1} \mid s_{t+1}) \, \mathrm{d}\xi_t(s_{t+1} \mid a_t, \mu) \tag{3.5}$$

### Example

If $\mathcal{S}, \mathcal{A}, \mathcal{X}$ are finite, and then we can define

- $\partial_t(j) = P(x_t \mid s_t = j)$
- $\boldsymbol{A}_t(i, j) = P(s_{t+1} = j \mid s_t = i, a_t)$.
- $\boldsymbol{b}_t(i) = \xi_t(s_t = i)$

We can then use Bayes theorem:

$$\boldsymbol{b}_{t+1} = \frac{\mathrm{diag}(\boldsymbol{p}_{t+1}) \boldsymbol{A}_t \boldsymbol{b}_t}{\boldsymbol{p}_{t+1}^\top \boldsymbol{A}_t \boldsymbol{b}_t}, \tag{3.6}$$

# When the POMDP $\mu$ is unknown

$$\xi(\mu, s^t \mid x^t, a^t) \propto P_\mu(x^t \mid s^t, a^t) P_\mu(s^t \mid a^t) \xi(\mu) \tag{3.7}$$

## Cases

- Finite $\mathcal{M}$.
- Finite $\mathcal{S}$
- General case

## Strategies for POMDPs

- Bayesian RL on POMDPs? EXP inference and planning
- Approximations and stochastic methods.
- Policy search methods.