# Complexity of stochastic branch and bound methods for belief tree search in Bayesian reinforcement learning

Christos Dimitrakakis

Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

23 Jan 2010

# Reinforcement learning

# Reinforcement learning

## Definition (The Reinforcement Learning Problem)

Learning how to act in an environment solely by **interaction** and **reinforcement**.

## Characteristics

- The environment is unknown.
- Data is collected by the agent through interaction.
- The optimal actions are hinted at via reinforcement (scalar rewards).

## Applications: Sequential Decision Making tasks

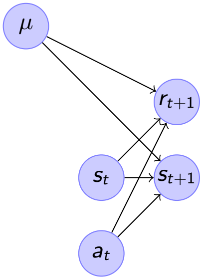
- Control, robotics, etc.
- Scheduling, network routing.
- Game playing, co-ordination of multiple agents.
- Relation to biological learning.

# Markov decision processes

## Markov decision processes (MDP)

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

- We observe state $s_t \in \mathcal{S}$.
- We take action $a_t \in \mathcal{A}$.
- We receive a reward $r_t \in \mathbb{R}$.




## Model

$$\mathbb{P}_\mu(s_{t+1} | s_t, a_t) \qquad \text{(Transition distribution)}$$

$$\mathbb{P}_\mu(r_{t+1} | s_t, a_t) \qquad \text{(Reward distribution)}$$

# Markov decision processes (MDPs)

## The agent

The agent is defined by its policy $\pi$.

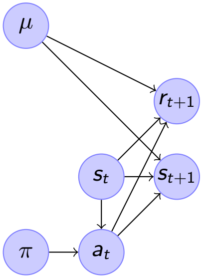
$$\mathbb{P}_\pi(a_t | s_t)$$

## Controlling the environment

We wish to find $\pi$ maximising the expected total future reward

$$\mathbb{E}_{\mu,\pi} \sum_{t=1}^{T} r_t \qquad \text{(utility)}$$

to the horizon $T$ .



# Markov decision processes (MDPs)

## The agent

The agent is defined by its policy $\pi$.

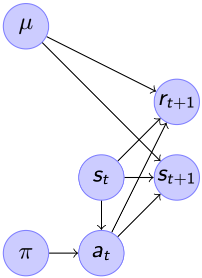
$$\mathbb{P}_\pi(a_t|s_t)$$

## Controlling the environment

We wish to find $\pi$ maximising the expected total future reward

$$\mathbb{E}_{\mu,\pi} \sum_{t=1}^{T} \gamma^t r_t \qquad \text{(utility)}$$

to the horizon $T$ with discount factor $\gamma \in (0, 1]$.



# Value functions

## State value function

$$V^{\pi}_{t,\mu}(s) \triangleq \mathbb{E}_{\pi,\mu}\left(\sum_{k=1}^{T} \gamma^k r_{t+k} \middle| s_t = s\right)$$

How good a state is under the policy $\pi$ for the environment $\mu$.

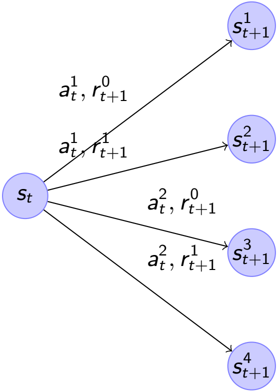
$$\pi^*(\mu) : V^{\pi^*(\mu)}_{t,\mu}(s) \geq V^{\pi}_{t,\mu}(s) \quad \forall \pi, t, s \qquad \text{(optimal policy)}$$

The optimal policy $\pi^*$ dominates all other policies $\pi$ everywhere in $\mathcal{S}$.

$$V^*_{t,\mu}(s) \triangleq V^{\pi^*(\mu)}_{t,\mu}(s), \qquad \text{(optimal value function)}$$

The optimal value function $V^*$ is the value function of the optimal policy $\pi^*$.

# When the environment $\mu$ is known




## Iterative/offline methods

- Estimate the optimal value function $V^*$ (i.e. with backwards induction on all $\mathcal{S}$).
- Iteratively improve $\pi$ (i.e. with policy iteration) to obtain $\pi^*$.

## Online methods

- Forward search followed by backwards induction (on subset of $\mathcal{S}$).

## Dynamic programming (Backwards Induction)

$$V_t(s_t) = \sup_a \mathbb{E}_\mu[r_t | s_t, a] + \gamma \sum_i V_{t+1}(s_{t+1}^i) \mathbb{P}_\mu(s_{t+1}^i | s_t, a)$$

# When the environment $\mu$ is unknown

## Decision theoretic solkution using a probabilistic belief

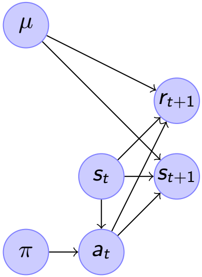
- Belief: A distribution over possible MDPs.
- Method: Take into account future beliefs when planning.
- Problem: The combined belief/MDP model is an infinite MDP.
- Goal: Efficient methods to approximately solve the infinite MDP.

# Near-optimal Bayesian RL

# Bayesian Reinforcement Learning

## Estimating the correct MDP

- The true $\mu$ is unknown, but we assume it is in $\mathcal{M}$.
- Maintain a belief $\xi_t(\mu)$ over all possible MDPs $\mu \in \mathcal{M}$.
- $\xi_0$ is our initial belief about $\mu \in \mathcal{M}$.




## The belief update

$$\xi_{t+1}(\mu) \triangleq \xi_t(\mu \mid s_{t+1}, r_{t+1}, s_t, a_t) \tag{1a}$$

$$= \frac{\mathbb{P}_\mu(s_{t+1}, r_{t+1} | s_t, a_t)\xi_t(\mu)}{\xi_t(s_{t+1}, r_{t+1} | s_t, a_t)}. \tag{1b}$$

# Exploration-exploitation trade-offs with Bayesian RL

## Exploration-exploitation trade-off

- We just described an estimation method.
- But, how should we behave while the MDP is not well estimated?
- The plausibility of different MDPs is important.

## Main idea

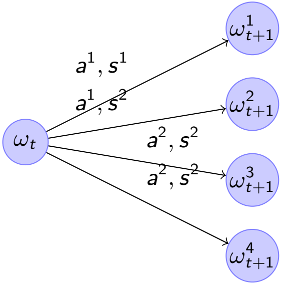
- Take knowledge gains into account when planning.
- Starting with the current belief, enumerate all possible future beliefs.
- Take the action that maximises the expected utility.

# Decision-theoretic solution

## Belief-Augmented Markov Decision Processes

- RL problems with state are expressed as MDPs.
- Under uncertainty there are two types of state variables: the environment's state $s_t$ and our belief state $\xi_t$.
- Augmenting the state space with the belief space allows us to represent RL under uncertainty as a *big* MDP with a hyperstate.
- This MDP can be solved with DP techniques (backwards induction).

# Belief tree




## Of interest

- (Pseudo)-tree structure.
- Hyperstate $\omega_t \triangleq (s_t, \xi_t)$.
- $\Omega_t \triangleq \{\omega_t^i : i = 1, 2, \dots\}$.

## The induced MDP $\nu$

$$\mathbb{P}_\nu(\omega_{t+1}^i | \omega_t, a_t) = \xi_t(s_{t+1}^i, r_{t+1}^i | s_t, a_t) = \int_{\mathcal{M}} \mathbb{P}_\mu(s_{t+1}^i, r_{t+1}^i | s_t, a_t) \xi_t(\mu) d\mu$$

## Backwards induction

$$V_t^*(\omega) = \sum_{\omega' \in \Omega_{t+1}} \xi_t(\omega' | \omega_t, a_t^*) [\mathbb{E}_{\xi_t}(r | \omega', \omega_t) + \gamma V_{t+1}^*(\omega')]$$

# The $n$-armed bandit problem

- Actions $\mathcal{A} = \{1, \ldots, n\}$.
- Expected reward $\mathbb{E}(r_t \mid a_t = i) = x_i$.
- Discount factor $\gamma \leq 1$ and/or horizon $T > 0$.
- If the expected rewards are unknown, what must we do?

## Decision-theoretic approach

- Assume $r_t \mid a_t = i \sim \psi(\theta_i)$, with $\theta_i \in \Theta$ unknown parameters.
- Define prior $\xi(\theta_1, \ldots, \theta_n)$.
- Select actions to maximise $\mathbb{E}_\xi \, U_t = \mathbb{E}_\xi \sum_{k=1}^{T-t} \gamma^k r_{t+k}$.

# Bernoulli example

Consider $n$ Bernoulli bandits with unknown parameters $\theta_i$, $i = 1, \ldots, n$ such that

$$r_t \mid a_t = i \sim \mathcal{B}ern(\theta_i), \qquad \mathbb{E}(r_t \mid a_t = i) = \theta_i. \tag{2}$$

We model our belief for each bandit's parameter $\theta_i$ as a Beta distribution $\mathcal{B}eta(\alpha_i, \beta_i)$, with density $f(\theta \mid \alpha_i, \beta_i)$ so that

$$\xi(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} f(\theta_i \mid \alpha_i, \beta_i).$$

Recall that the posterior of a Beta prior is also a Beta. Let $k_{t.i} \triangleq \sum_{k=1}^{t} \mathbb{I}\, a_k = i$ be the number of times we played arm $i$ and $\hat{r}_{t.i} \triangleq \frac{1}{k_{t,i}} \sum_{k=1}^{t} r_t \, \mathbb{I}\, a_k = i$ be the empirical reward of arm $i$ at time $t$. . Then, the posterior distribution for the parameter of arm $i$ is

$$\xi_t = \mathcal{B}eta(\alpha_i + k_{t,i}\hat{r}_{t,i}, \beta_i + k_{t,i}(1 - \hat{r}_{t,i}))$$

Since $r_t \in \{0, 1\}$ the possible states of our belief given some prior are $\mathbb{N}^{2n}$.

# Belief states

- The state of the bandit problem is the state of our belief.
- A sufficient statistic for our belief is the number of times we played each bandit and the total reward from each bandit.
- Thus, our state at time $t$ is entirely described our priors $\alpha, \beta$ (the initial state) and the vectors

$$k_t = (k_{t,1}, \ldots, k_{t,i}) \tag{3}$$
$$\hat{r}_t = (\hat{r}_{t,1}, \ldots, \hat{r}_{t,i}). \tag{4}$$

- At any time $t$, we can calculate the probability of observing $r_t = 1$ or $r_t = 0$ if we pull arm $i$ as:

$$\xi_t(r_t = 1 \mid a_t = i) = \frac{\alpha_i + k_{t,i}\hat{r}_{t,i}}{\alpha_i + \beta_i + k_{t,i}}$$

- The next state is well-defined and depends only on the current state.
- Thus, the $n$-armed bandit problem is an MDP.

# Experiment design

## Example

Consider $k$ treatments to be administered to $T$ volunteers. Each volunteer can only be used once. At the $t$-th trial, we perform some experiment $a_t \in \{1, \ldots, k\}$ and obtain a reward $r_t = 1$ if the result is successful and 0 otherwise. If simply randomise trials, then we will obtain a much lower number of successes than if we solve the bandit MDP.

## Example

We are given a hypothesis set $H = \{h_1, h_2\}$, a prior $\psi_0$ on $H$, a decision set $D = \{d_1, d_2\}$ and a loss function $L : D \times H \to \mathbb{R}$. We can choose from a set of $k$ possible experiments to be performed over $T$ trials. At the $t$-th trial, we choose experiment $a_t \in \{1, \ldots, k\}$ and observe outcome $x_t \in \mathcal{X}$. Our posterior is $\psi_t(h) = \psi_0(h \mid a_1, \ldots, a_t, x_1, \ldots, x_t)$¿ The reward is $r_t = 0$ for $t < T$ and

$$r_T = -\min_{d \in D} \mathbb{E}_{\psi_T}(L \mid d).$$

The process is a $T$-horizon MDP, which can be solved with standard backwards induction.

# Tree properties

## Tree depth

(Naive) Error at depth $k$: $\epsilon \propto \frac{\gamma^k}{1-\gamma}$.

## Branching factor

$$\phi = |\mathcal{R}| \cdot |\mathcal{A}| \cdot |\mathcal{S}|$$

## Practical methods to handle the tree

- Lookeahead up to fixed time $T$.
- In some cases, closed-form solutions (i.e. Gittins indices)
- Pruning or sparse expansion.
- Value function approximations.

# Bounds on node values

Fortunately, we can obtain bounds on the value of any node $\omega = (s, \xi)$. Let $\pi^*(\mu)$ be the optimal policy for $\mu$:

## Lower bound

$$V^*(\omega) \geq \mathbb{E}_\xi \, V_\mu^{\pi^*(\bar{\mu}_\xi)}(s),$$

where $\bar{\mu}_\xi \triangleq \mathbb{E}_\xi \, \mu$ is the mean MPD.
The optimal policy must be at least as good as any stationary policy,

## Upper bound

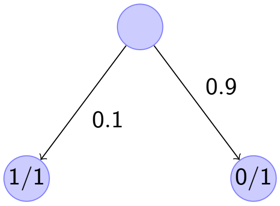
$$\mathbb{E}_\xi \max_\pi V_\mu^\pi(s) \geq V^*(\omega)$$

The optimal policy cannot do better than the policy which learns the correct model at the next time-step.

## Estimating the Bounds for some hyperstate $\omega = (s, \xi)$

$$\int V_\mu(s) \xi(\mu) \, \mathrm{d}\mu \approx \frac{1}{n} \sum_{i=1}^n \hat{v}_i, \qquad v_i = V_{\mu_i}(s), \mu_i \sim \xi.$$

# Stochastic branch and bound



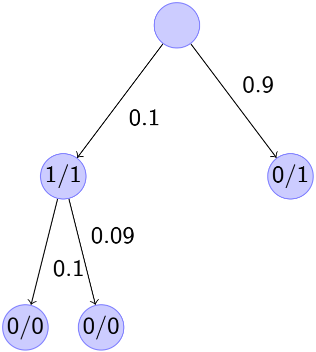

## Main idea

- Sample once from all leaf nodes.
- Expand the node with the highest mean upper bound.
- We quickly discover overoptimistic bounds.
- Unexplored leaf nodes accumulate samples.

## Hierarchical variant

- Sample children instead of leafs.
- Average bounds along path to avoid degeneracy.

# Stochastic branch and bound



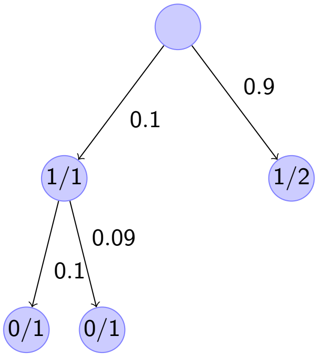

## Main idea

- Sample once from all leaf nodes.
- Expand the node with the highest mean upper bound.
- We quickly discover overoptimistic bounds.
- Unexplored leaf nodes accumulate samples.

## Hierarchical variant

- Sample children instead of leafs.
- Average bounds along path to avoid degeneracy.

# Stochastic branch and bound




## Main idea

- Sample once from all leaf nodes.
- Expand the node with the highest mean upper bound.
- We quickly discover overoptimistic bounds.
- Unexplored leaf nodes accumulate samples.

## Hierarchical variant

- Sample children instead of leafs.
- Average bounds along path to avoid degeneracy.

# Complexity results

Let $\Delta$ be the value difference between two branches and $\beta = V_{\max} - V_{\min}$.
If $N$ times an optimal branch is sampled without being expanded:

$$\mathbb{P}(N > n) \leq \exp(-2\beta^{-2} n^2 \Delta^2)$$

If $K$ is the number of times a sub-optimal branch will be expanded then $\mathbb{P}(K > k)$, for $k > k_0 = \log_\gamma \Delta/\beta$

## Stochastic branch and bound 1

$$\mathcal{O}\left(\exp\{-2\beta^{-2}[(k - k_0)\Delta^2]\}\right)$$

## Stochastic branch and bound 2

$$\tilde{\mathcal{O}}\left(\exp\{-2(k - k_0)^2(1 - \gamma^2)\right)$$

# Summary

## Results

- Development of upper and lower bounds for belief tree values
- Application to efficient tree expansion
- Complexity bounds for tree expansion

## Future work

- Sparse sampling and smoothness property to reduce branching factor
- Can we get regret bounds via posterior concentration?
- Extend approach to non-parametrics ...

# Questions?

Thank you for your attention.