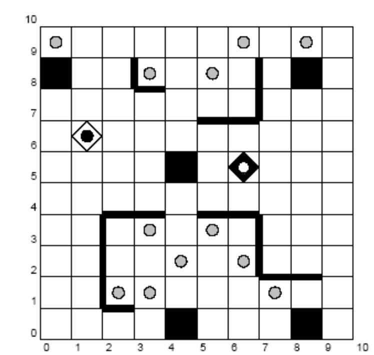
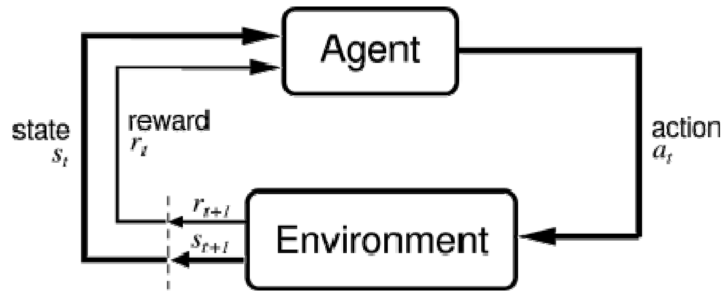
Markov Decision Processes, Dynamic Programming, and Reinforcement Learning in R



Jeffrey Todd Lins
Thomas Jakobsen
Saxo Bank A/S

jtl@saxobank.com, tj@saxobank.com

Agent-Environment Interface



Source: Sutton & Barto, 2001

Markov Decision Process

We define a Markov Decision Process as a tuple $(\mathcal{S}, \mathcal{A}, T, R)$ where

- $\mathcal{S}$ is a finite set of states
- $\mathcal{A}$ is a finite set of actions
- $T : \mathcal{S} \times \mathcal{A} \rightarrow \Pi(\mathcal{S})$ is the transition model giving a probability distribution over all states for ending in a future state, s', given that an agent takes action, a in state s.
- $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is a real-valued reward function yielding the immediately expected reward for taking each action in each state.

Dynamic Programming

- Deterministic Policy
$$\pi : \mathcal{S} \rightarrow \mathcal{A}$$
- Stochastic Policy
$$\pi : \mathcal{S} \rightarrow \Pi(\mathcal{A})$$
- State Value Function
$$V_\pi(s) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_t = s\right],$$
where $0 < \gamma < 1$ is a discount factor that controls how much influence future rewards have, and r_t is the reward received at time t.
- State-Action Value Function
$$Q_\pi(s, a) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_t = s, a_t = a\right].$$

Bellman Equation

- Bellman Equation

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') Q_\pi(s',a).$$

or in matrix notation

$$Q_\pi = R + \gamma \mathbf{T}\mathbf{\Pi}_\pi Q_\pi.$$

- Now we consider that an optimal policy, π^*, will satisfy

$$\pi^* = \arg\max_\pi E\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi\Big].$$

Bellman Optimality Equation

- It can now be shown that there is a policy,

$$\pi^* = \arg\max_a E\Big[\sum_{s'} T(s,a,s') V(s') \mid \pi\Big]$$

that is optimal for the value in a *subsequent* state.

- Bellman Optimality Equation

$$Q_{\pi^*}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_a Q_{\pi^*}(s',a).$$

or in matrix notation

$$Q_{\pi^*} = R + \gamma \mathbf{T}\mathbf{\Pi}_{\max} Q_{\pi^*}.$$

Value Iteration

// $\mathcal{S}$: States
// $\mathcal{A}$: Actions
// $\mathcal{T}$: Transition Model
// $\mathcal{R}$: Reward Function
// γ : Discount Factor
// Q_0 : Initial State-Action Value Function
// ϵ : Stopping criterion
$Q' \leftarrow Q_0$
repeat
  $Q \leftarrow Q'$
  $Q^\pi \leftarrow \mathcal{R} + \gamma \mathbf{T}^{\max} Q$
until $|Q - Q'| < \epsilon$
$\forall s \in \mathcal{S}, \pi'(s) \leftarrow \arg\max_a Q'(s,a)$
return π

Policy Iteration

// $\mathcal{S}$: States
// $\mathcal{A}$: Actions
// $\mathcal{T}$: Transition Model
// $\mathcal{R}$: Reward Function
// γ : Discount Factor
// π_0 : Initial Policy

$\pi' \leftarrow \pi_0$
repeat
  $\pi \leftarrow \pi'$
  $Q^\pi \leftarrow (\mathbf{I} - \gamma \mathbf{T}^\pi)^{-1} \mathcal{R}$
  $\forall s \in \mathcal{S}, \pi'(s) \leftarrow \arg\max_a Q^\pi(s,a)$
until $\pi \leftarrow \pi'$
return π

Reinforcement Learning

- **Temporal Difference (TD) Learning**(Sutton, 1988) yields the state value function, V_π, for a fixed policy, given a sample set (s, a, r, s')

$$\widehat{V}_{t+1}(s) = \widehat{V} + \alpha\Big[r + \gamma \widehat{V}_t(s') - \widehat{V}_t(s)\Big]$$

.

- **Q-Learning**(Watkins, 1989) yields an optimal policy, π^*, by an approximation of Q_{π^*}, for a fixed policy, given a sample set (s, a, r, s')

$$\widehat{Q}_{t+1}(s,a) = \widehat{Q} + \alpha\Big[r + \gamma \widehat{Q}_t(s', a') - \widehat{Q}_t(s,a)\Big]$$

Temporal Difference Learning

// D : Samples (s, a, r, s')
// $\mathcal{A}$: Actions
// α_0 : Initial Learning Rate
// γ : Discount Factor
// V_0 : Initial Value Function
// π :Policy

$\tilde{V} \leftarrow V_0, \alpha \leftarrow \alpha_0, t \leftarrow 0$
for $(s, a, r, s') \in D(\pi)$ **do**
  $\tilde{V}(s) \leftarrow \tilde{V}(s) + \alpha(r + \gamma \tilde{V}(s') - \tilde{V}(s,))$
  $\alpha \leftarrow \sigma(\alpha, \alpha_0, t)$
  $t \leftarrow t + 1$
end for

return $\tilde{V}$

Q-Learning

// D : Samples (s, a, r, s')
// $\mathcal{A}$: Actions
// α_0 : Initial Learning Rate
// γ : Discount Factor
// Q_0 : Initial State-Action Value Function
// π : Exploration Policy

$\tilde{Q} \leftarrow Q_0, \alpha \leftarrow \alpha_0, t \leftarrow 0$
for $(s, a, r, s') \in D(\pi, Q)$ **do**
  $\tilde{Q}(s,a) \leftarrow \tilde{Q}(s,a) + \alpha(r + \gamma \max \tilde{Q}(s', a') - \tilde{Q}(s,a))$
  $\alpha \leftarrow \sigma(\alpha, \alpha_0, t)$
  $t \leftarrow t + 1$
end for

return $\tilde{Q}$

Linear Architectures

Linear Approximation Architectures

- Basis Functions: $\phi(s, a)$
- Weights: w_i
- k: column vector of size $|\mathcal{S}||\mathcal{A}|$

$$Q_{\pi,w}(s,a) = \sum_{i=1}^{k} \phi_i(s,a) w_{i,\pi} = \phi(s,a)^\top w_\pi.$$

Least Squares TD Learning

// D : Samples (s, a, r, s')
// $\mathcal{A}$: Actions
// k : Number of basis functions
// γ : Discount Factor
// V_0 : Initial Value Function
// π :Policy

$\mathbf{A} \leftarrow 0, b \leftarrow 0$
for $(s, a, r, s') \in D(\pi)$ **do**
  $\tilde{A} \leftarrow \tilde{A} + \phi(s)(\phi(s) - \phi(s'))$
  $\tilde{b} \leftarrow \tilde{b} + \phi(s) r$
  $w_\pi \leftarrow \tilde{A}^{-1}\tilde{b}$
end for

return $\tilde{w}_\pi$

Examples of RL in Finance

Performance Functions and Reinforcement Learning for Trading Systems and Portfolios.

John Moody, Lizhong Wu, Yuansong Liao & Matthew Saffell. Journal of Forecasting, Volume 17, Pages 441-470, 1998.

Intraday FX trading: Reinforcement learning vs evolutionary learning.

M. A. H. Dempster, T. W. Payne, & V. S. Romahi. Working Paper No. 23/01, Judge Institute of Management, University of Cambridge, 2001.

Advantages of RL in R

- Vectorized Programming
- Flexible, Interactive Simulation Environment
- Wide Range of Possibilities for Linear Basis Functions
- Interface to Existing Packages: HMMs, SVMs, GAs, Neural Networks

References

Richard Sutton and Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press, Cambridge, Massachusetts, 1998.

Michail G. Lagoudakis and Ronald Parr. “Least-Squares Policy Iteration,” *Journal of Machine Learning Research*, 4, 2003, pp. 1107-1149.