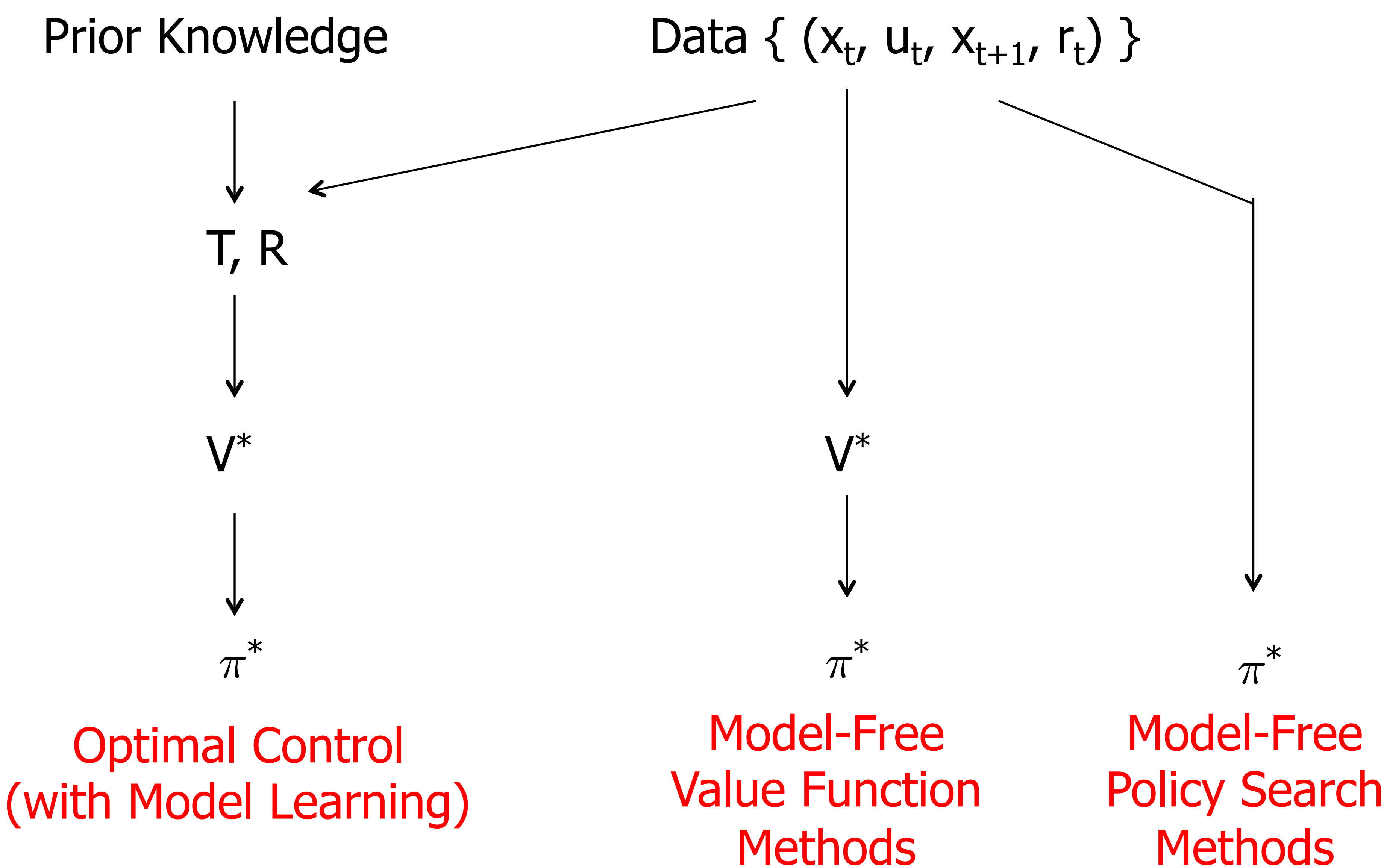
ICRA 2012 Tutorial on Reinforcement Learning
3a Optimal Control

Pieter Abbeel
UC Berkeley

Jan Peters
TU Darmstadt

A Reinforcement Learning Ontology



Outline

- Optimal Control

  =

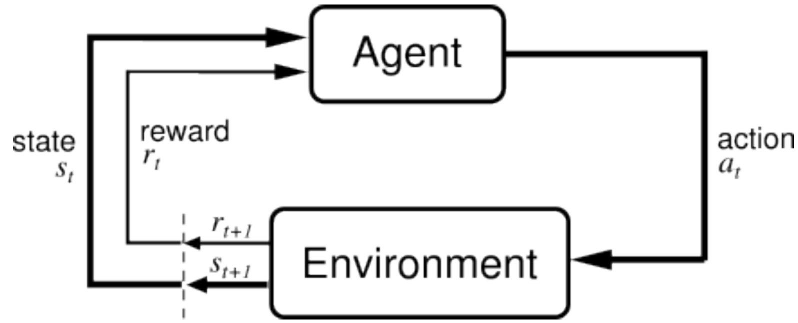
  given an MDP (S, A, T, R, γ, H)

  find the optimal policy π^*

- Exact Methods:
  - Value Iteration
  - Policy Iteration
  - Linear Programming

For now: discrete state-action spaces as they are simpler to get the main concepts across. Will consider continuous spaces in the next session!

Markov Decision Process (X, U, T, R, γ, H)



Given

- X: set of states
- U: set of actions
- T: $T(x,u,x') = P(x_{t+1} = x' \mid x_t = x, u_t = u)$
- R: $R(x,u)$ = reward for ($x_t = x$, $u_t = u$)
- $\gamma \in [0,1]$, discount factor
- H: horizon over which the agent will act

Goal:

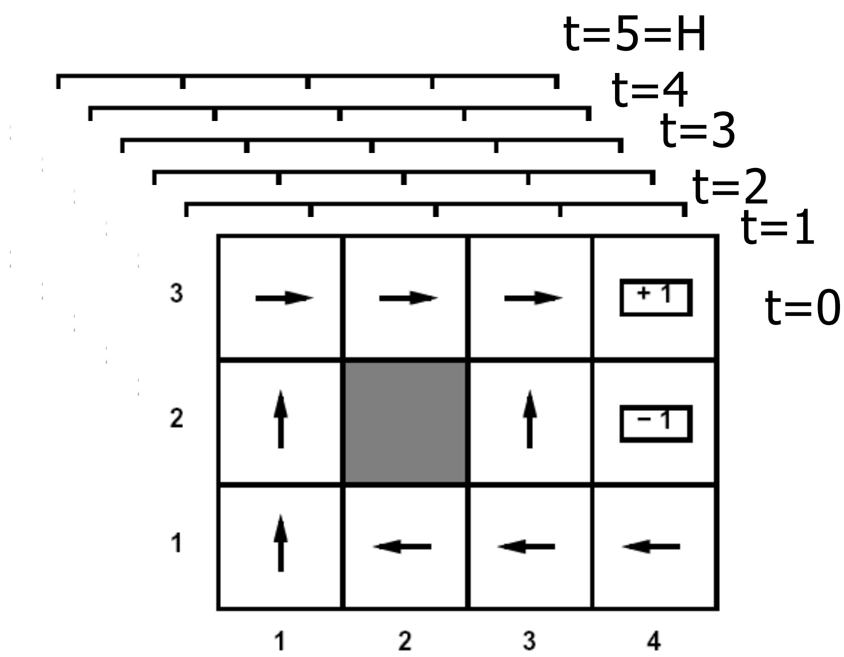
- Find π : X x {0, 1, …, H} → U that maximizes expected sum of rewards, i.e.,

$$\pi^* = \arg\max_{\pi} \mathrm{E}[\sum_{t=0}^{H} R(X_t, U_t) | \pi]$$

Solving MDPs

- An optimal policy π^*: X x 0:H → A

  UA policy π gives an action for each state for each time



  - An optimal policy maximizes expected sum of rewards

- Contrast: If deterministic, want an optimal plan, or sequence of actions

Value Iteration

- Idea:
  - $$V_i^*(x) = \max_{\pi_{H-i:H-1}} \mathrm{E}[\sum_{t=H-i}^{H-1} R_t(X_t, U_t) | \pi_{H-i:H}, X_{H-i} = x]$$

  = the expected sum of rewards accumulated when starting from state x and acting optimally for a horizon of i steps

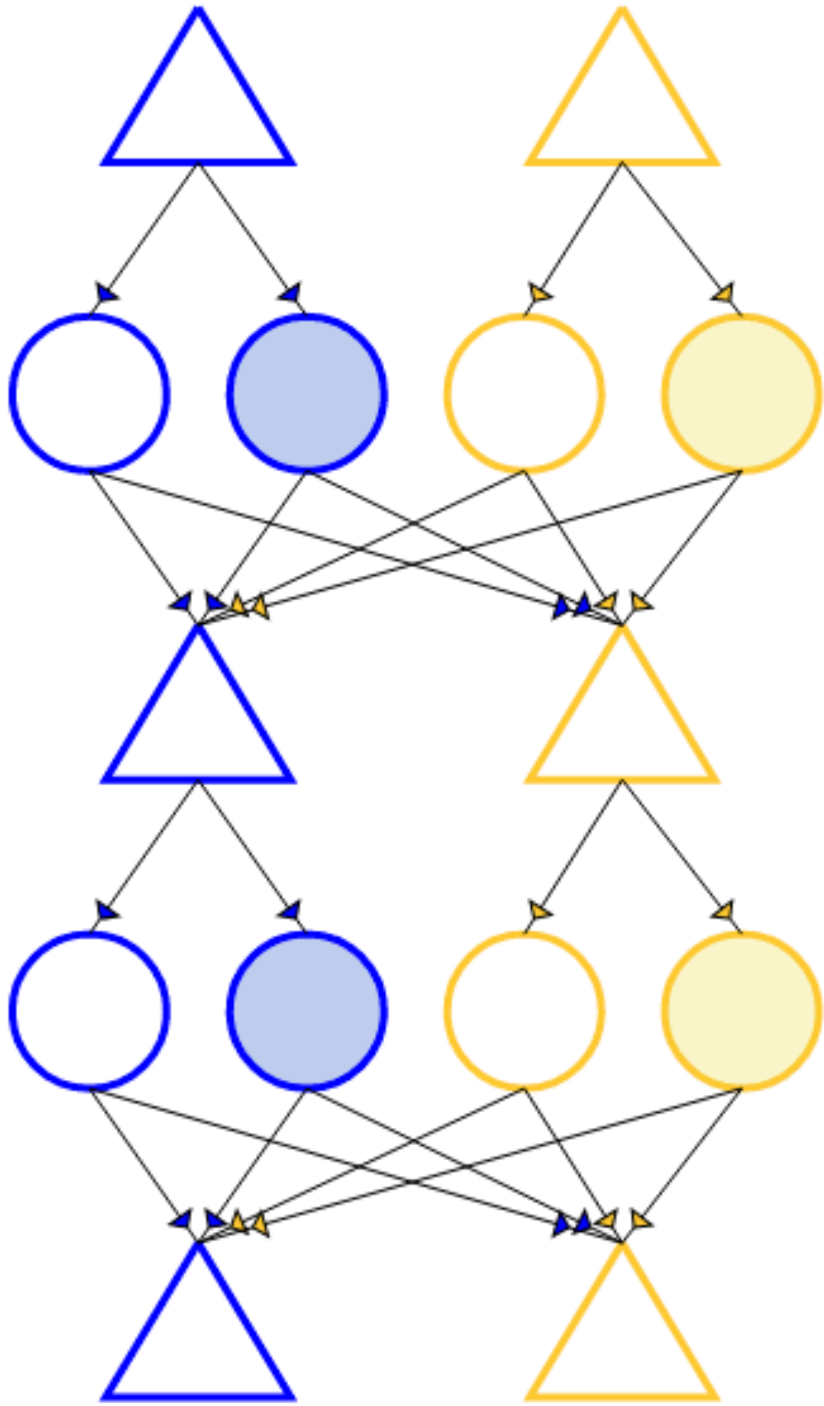
- Algorithm:
  - Start with $V_0^*(s) = 0$ for all s.
  - For i=1, … , H

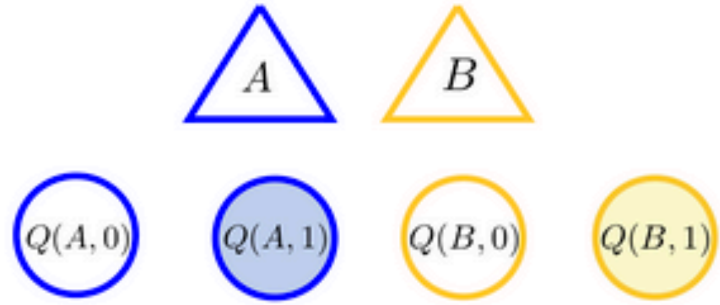
    For x $\in$ S:

    For u $\in$ A: $Q_{i+1}^*(x,u) \leftarrow \sum_{x'} T(x,a,x') \left[R(x,u) + \gamma V_i^*(x')\right]$

    $V_{i+1}^*(x) \leftarrow \max_u Q(x,u)$

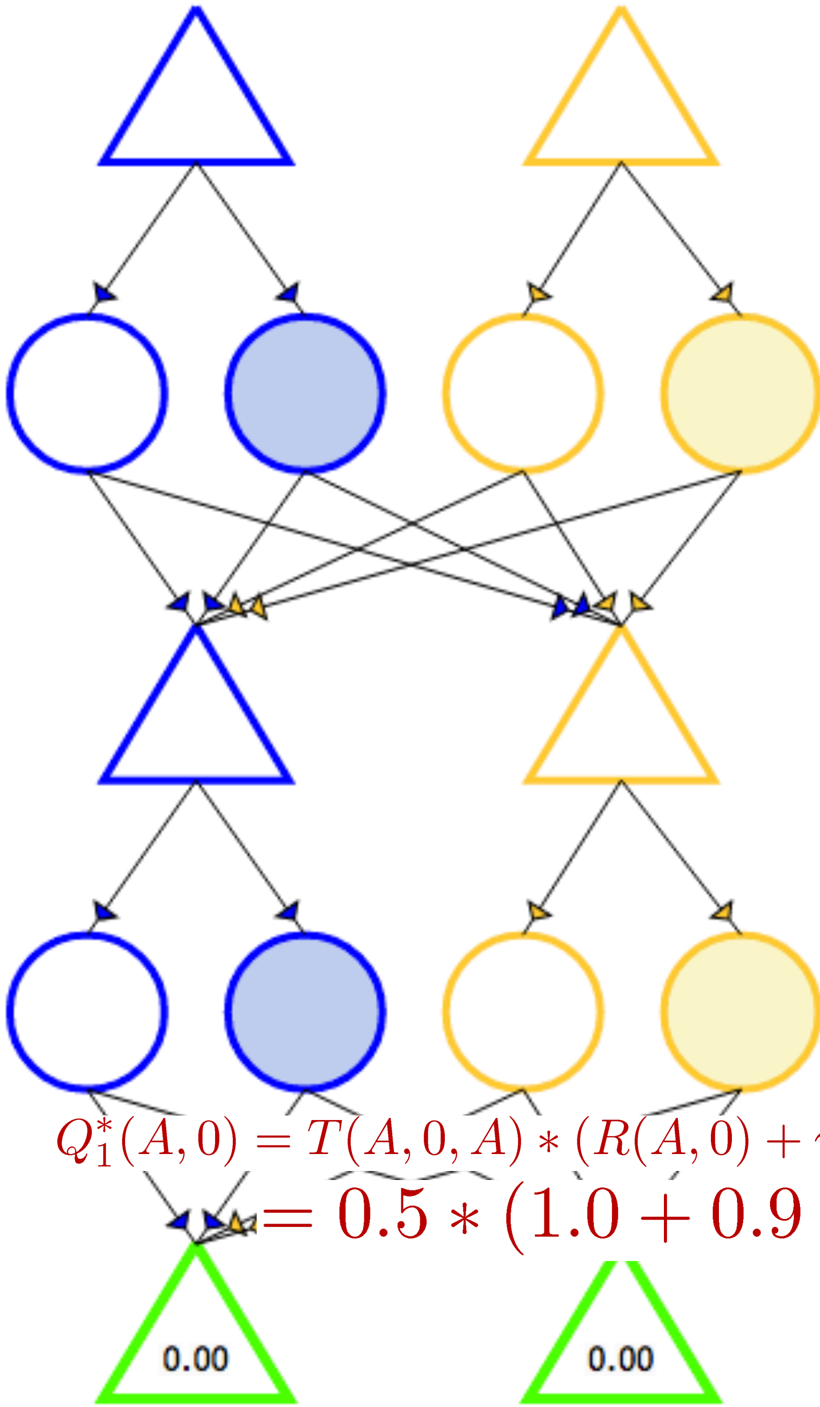
    $\pi_{i+1}^*(x) \leftarrow \arg\max_u Q(x,u)$

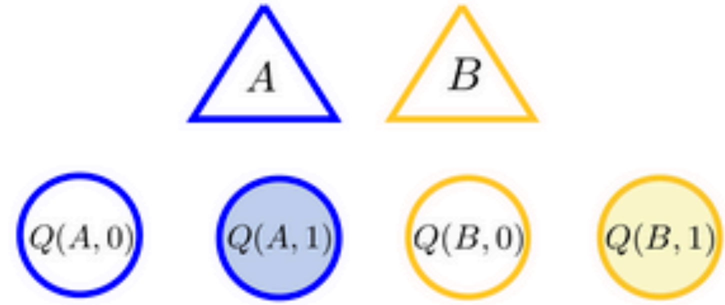




x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
B	1	A	0.40	1.0
B	1	B	0.60	1.0

$$V_0^*(A) = V_0^*(B) = 0$$

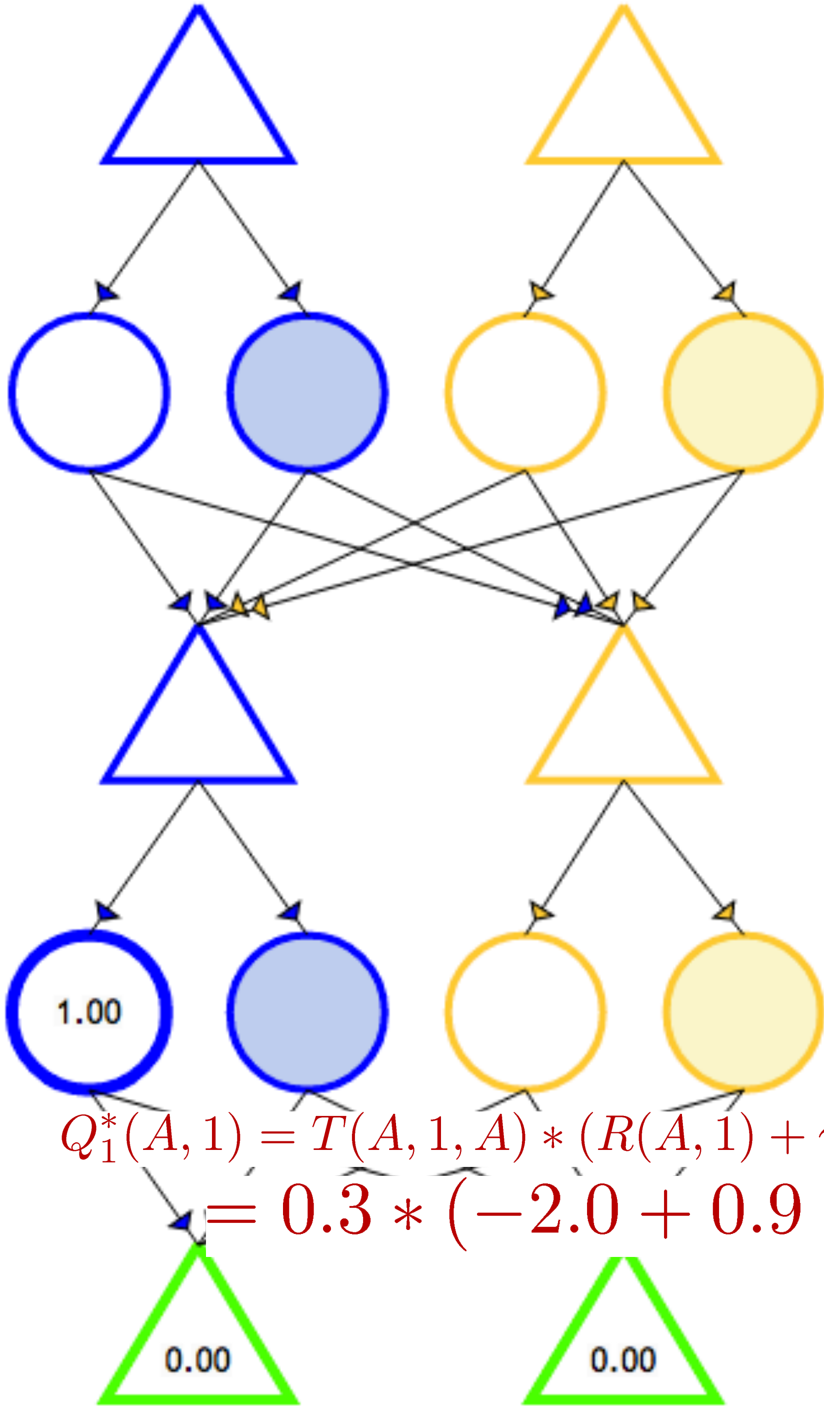


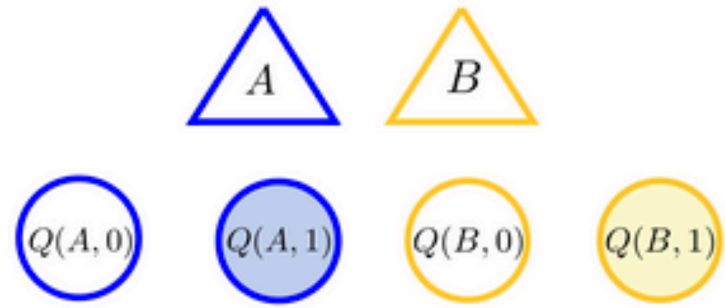


x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	1.0

$$Q_1^*(A,0) = T(A,0,A)*(R(A,0)+\gamma V_0^*(A)) + T(A,0,B)*(R(A,0)+\gamma V_0^*(B))$$

$$= 0.5*(1.0+0.9*0) + 0.5*(1.0+0.9*0)$$

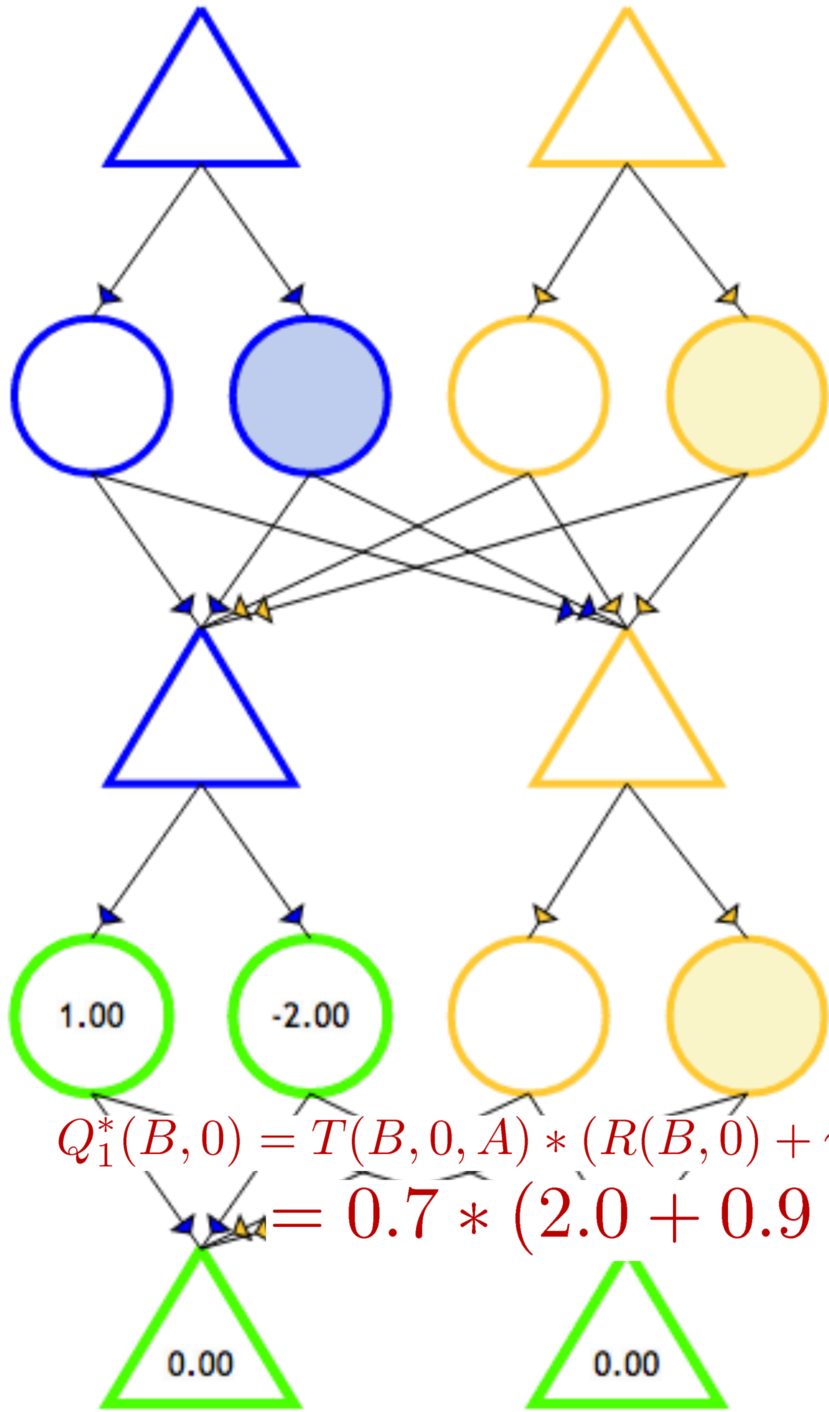


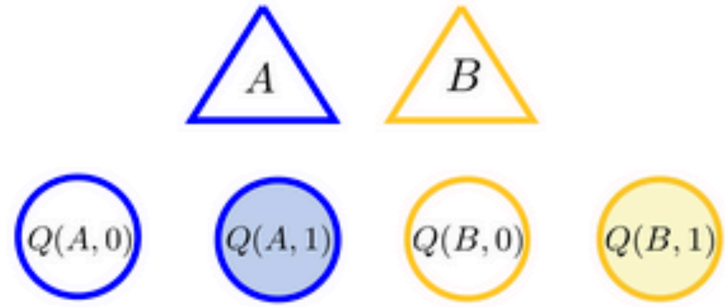


x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	1.0

$$Q_1^*(A,1) = T(A,1,A) * (R(A,1) + \gamma V_0^*(A)) + T(A,1,B) * (R(A,1) + \gamma V_0^*(B))$$

$$= 0.3 * (-2.0 + 0.9 * 0) + 0.7 * (-2.0 + 0.9 * 0)$$

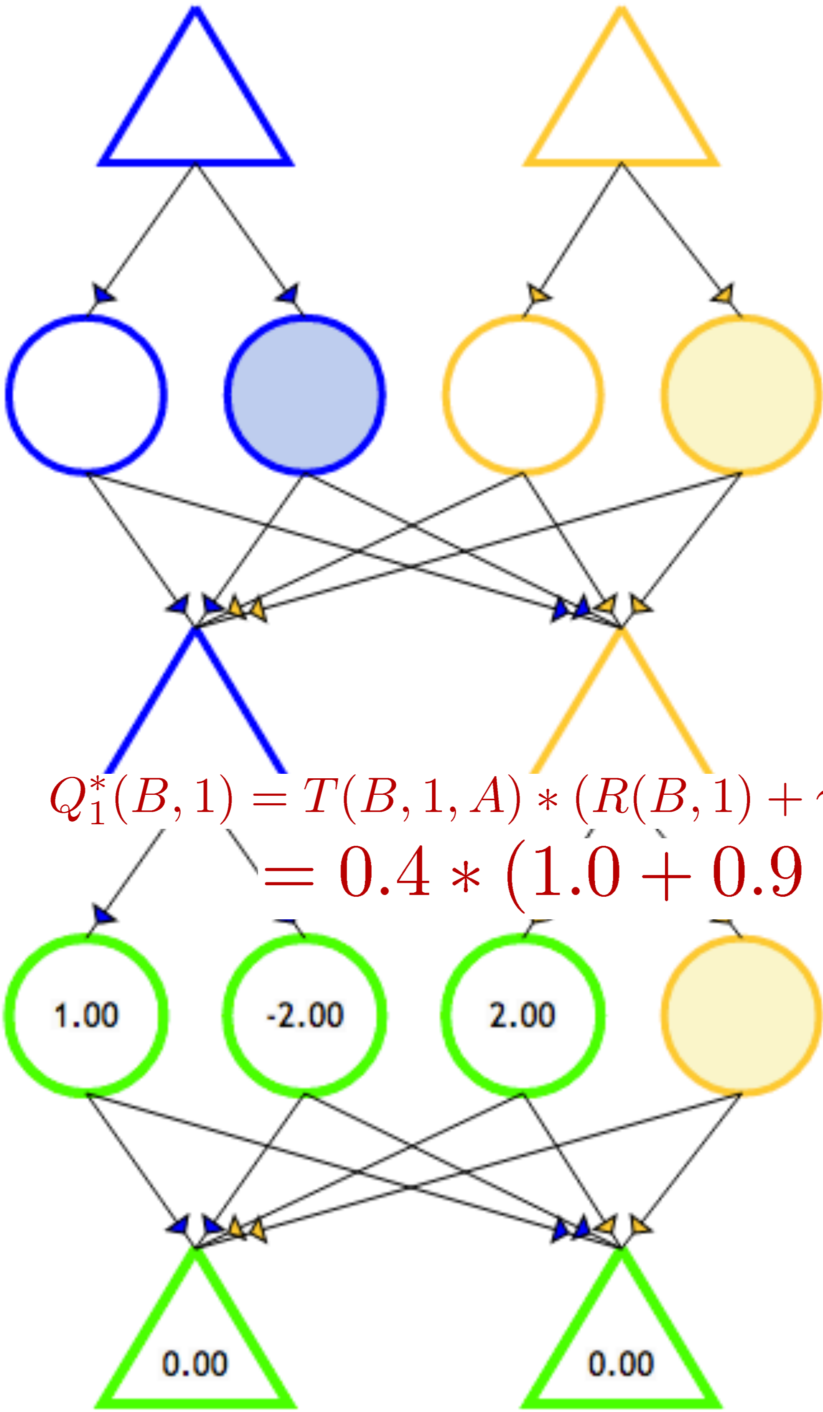


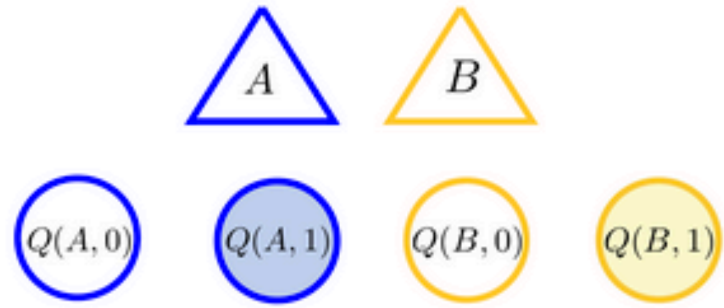


x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	1.0

$$Q_1^*(B,0) = T(B,0,A)*(R(B,0)+\gamma V_0^*(A)) + T(B,0,B)*(R(B,0)+\gamma V_0^*(B))$$

$$= 0.7*(2.0+0.9*0) + 0.3*(2.0+0.9*0)$$

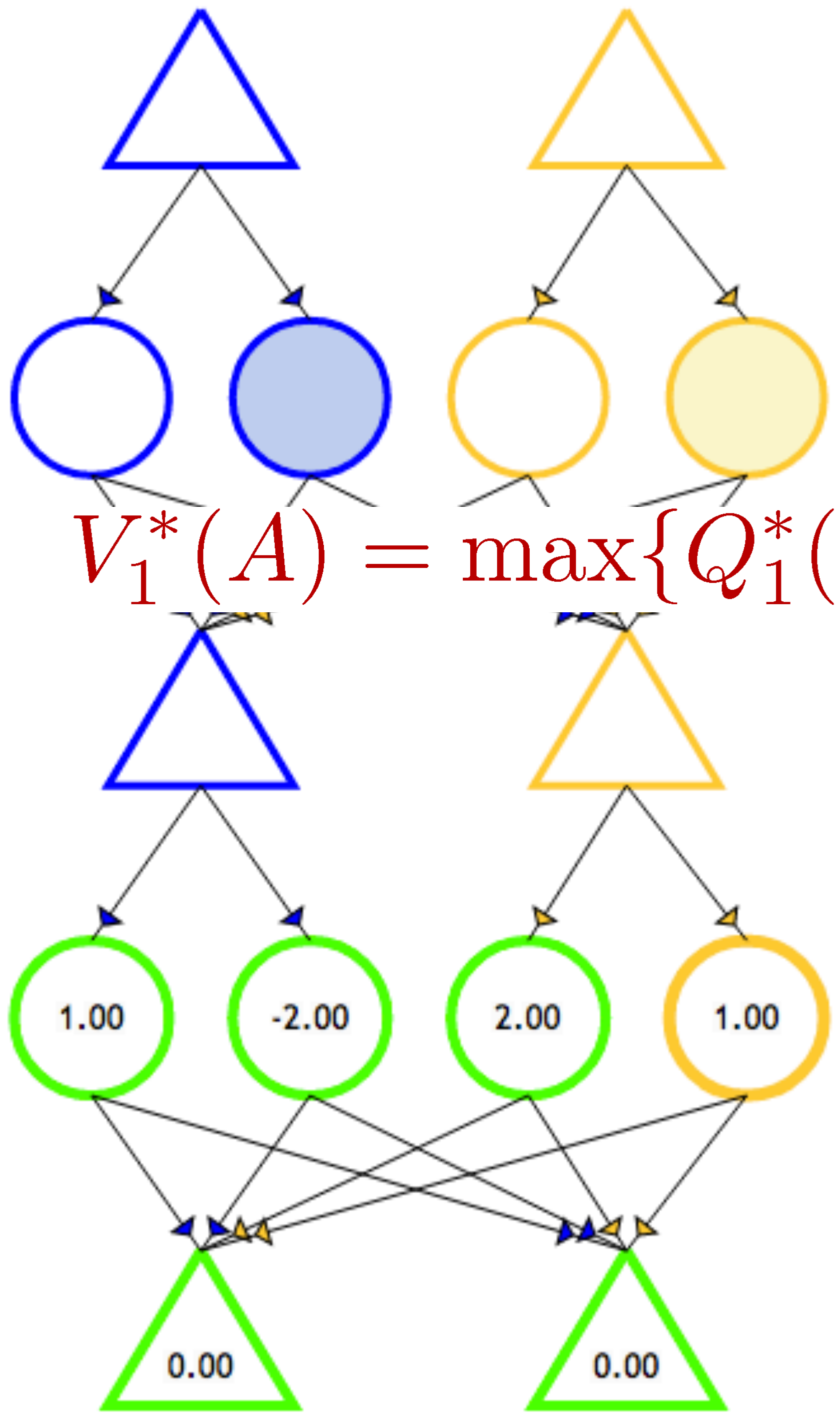


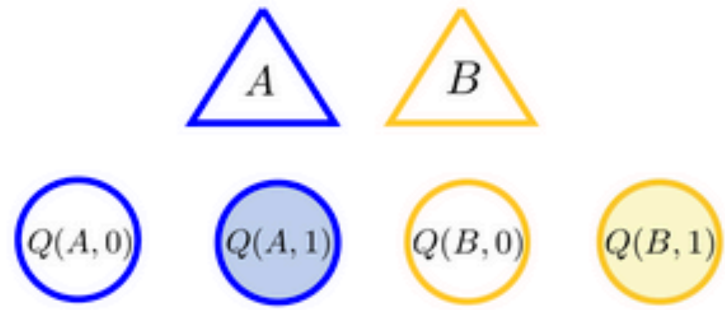


x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
B	1	A	0.40	1.0
B	1	B	0.60	1.0

$$Q_1^*(B,1) = T(B,1,A) * (R(B,1) + \gamma V_0^*(A)) + T(B,1,B) * (R(B,1) + \gamma V_0^*(B))$$

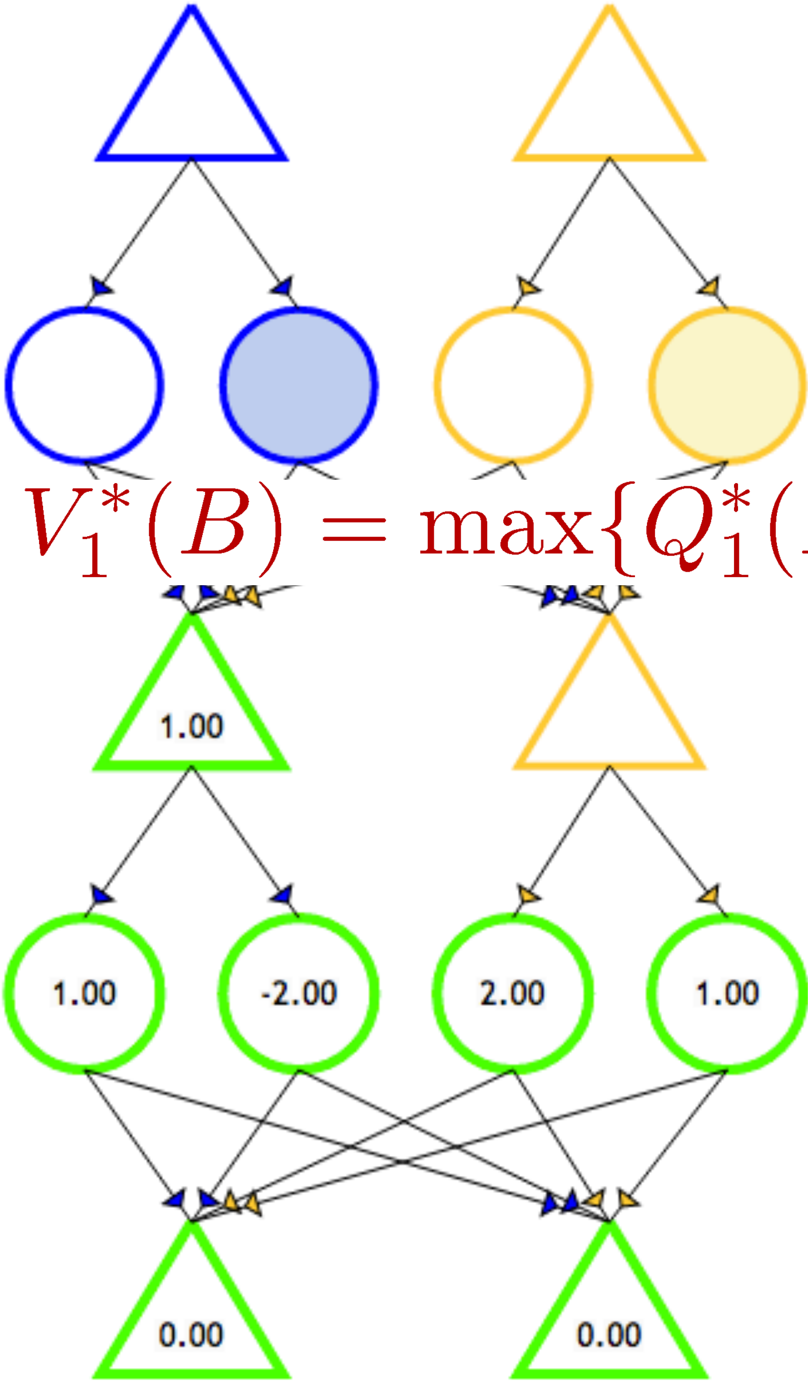
$$= 0.4 * (1.0 + 0.9 * 0) + 0.6 * (1.0 + 0.9 * 0)$$



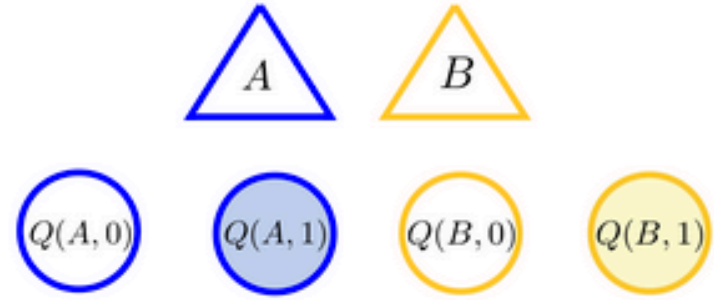


$$V_1^*(A) = \max\{Q_1^*(A,0), Q_1^*(A,1)\}$$

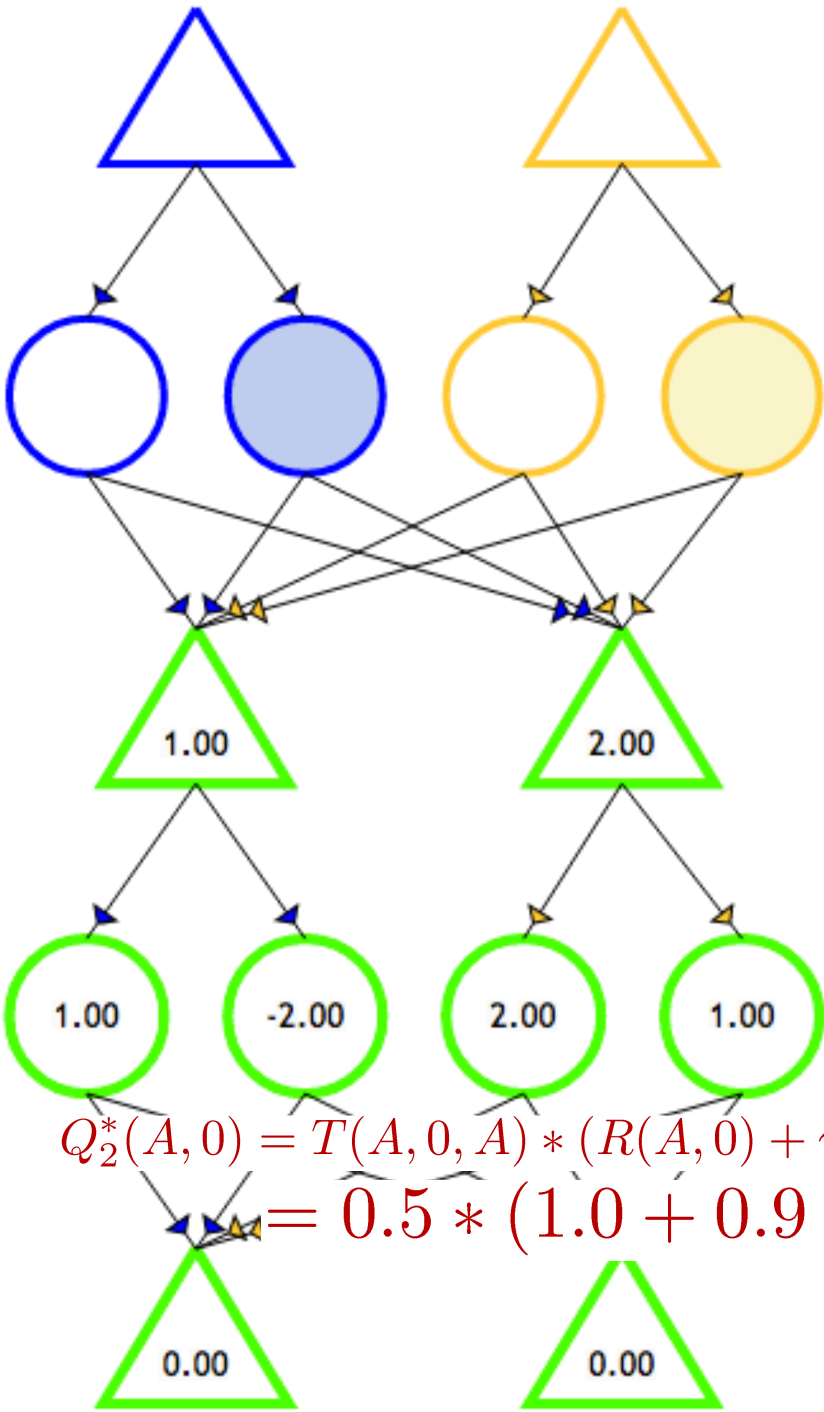
			$'(x, u, x')$	$R(x, u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
B	1	A	0.40	1.0
B	1	B	0.60	1.0

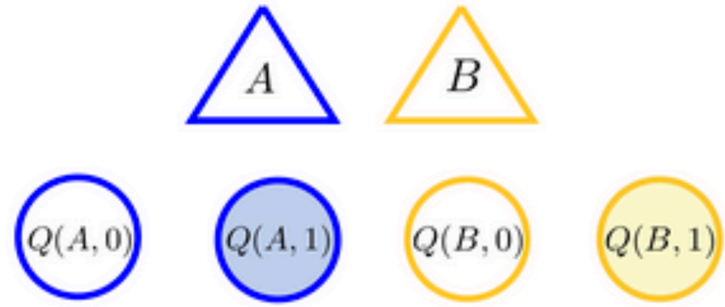


$$V_1^*(B) = \max\{Q_1^*(B,0), Q_1^*(B,1)\}$$



			$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
B	1	A	0.40	1.0
B	1	B	0.60	1.0

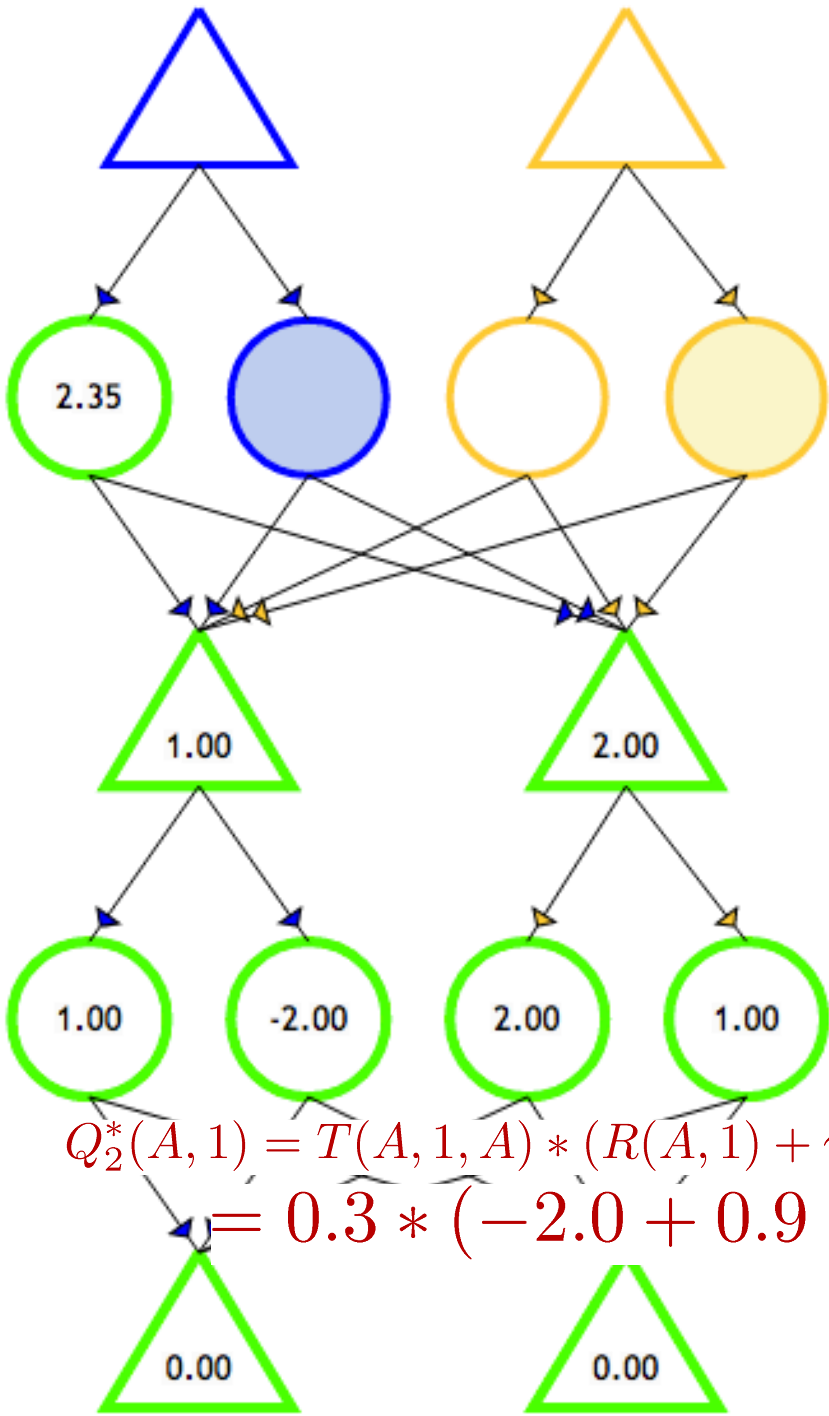


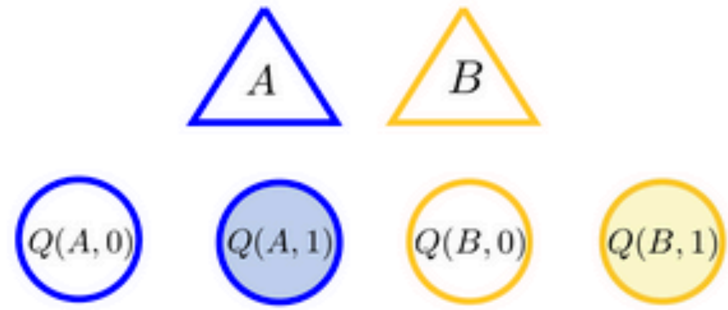


x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	1.0

$$Q_2^*(A,0) = T(A,0,A) * (R(A,0) + \gamma V_1^*(A)) + T(A,0,B) * (R(A,0) + \gamma V_1^*(B))$$

$$= 0.5 * (1.0 + 0.9 * 1) + 0.5 * (1.0 + 0.9 * 2)$$

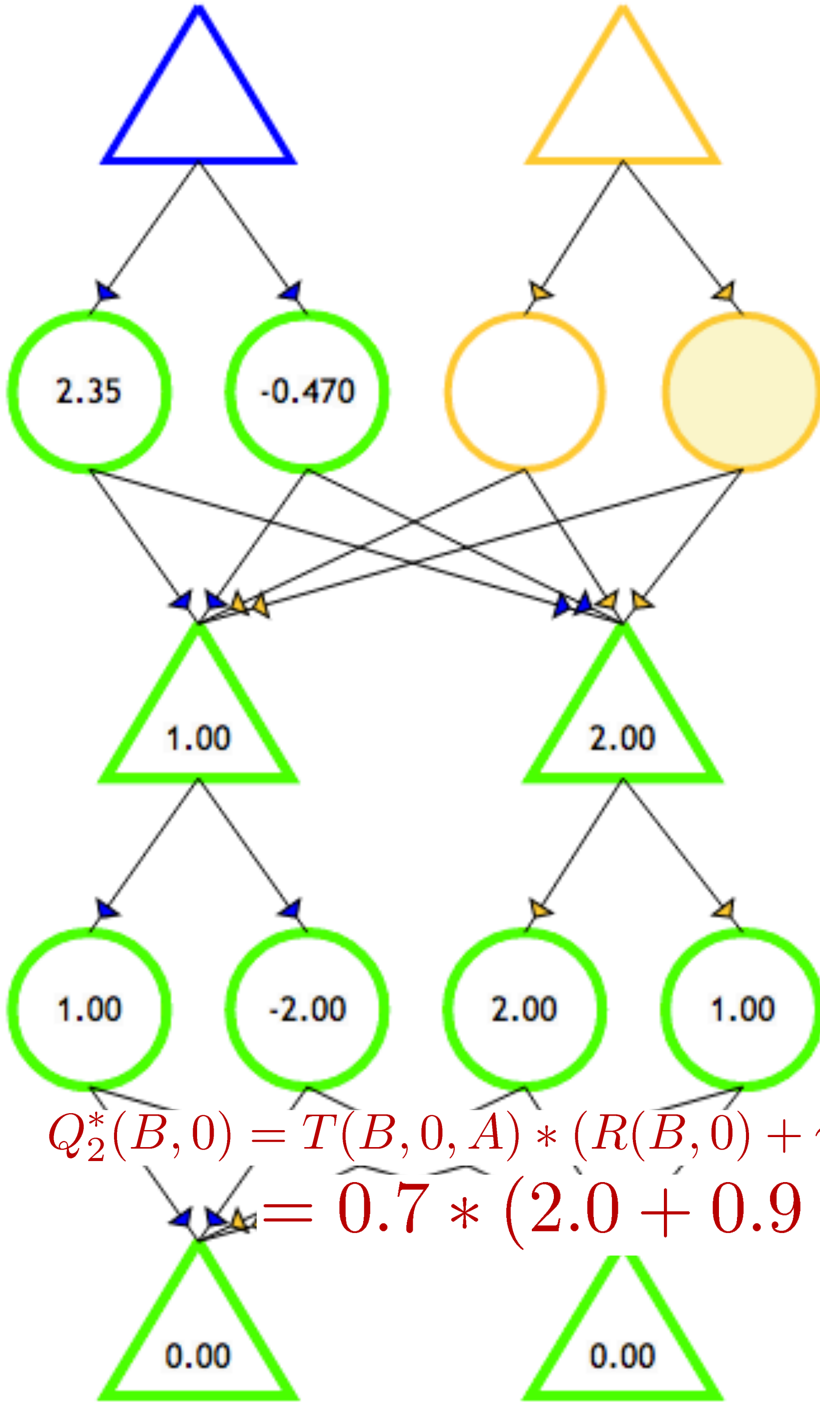


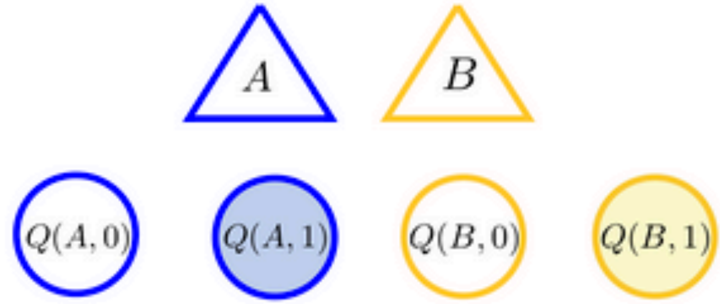


x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	1.0

$$Q_2^*(A,1) = T(A,1,A) * (R(A,1) + \gamma V_1^*(A)) + T(A,1,B) * (R(A,1) + \gamma V_1^*(B))$$

$$= 0.3 * (-2.0 + 0.9 * 1) + 0.7 * (-2.0 + 0.9 * 2)$$

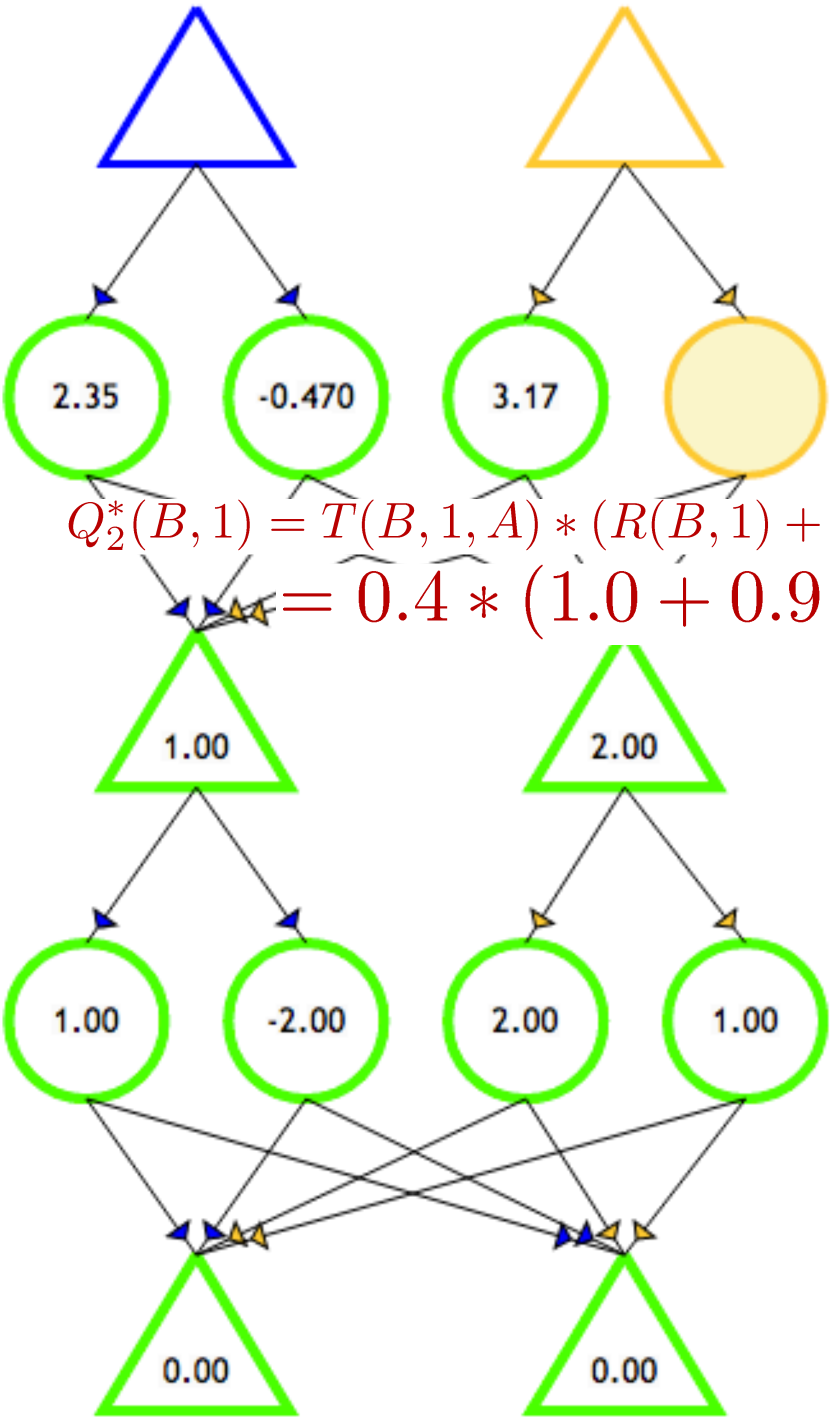




x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	1.0

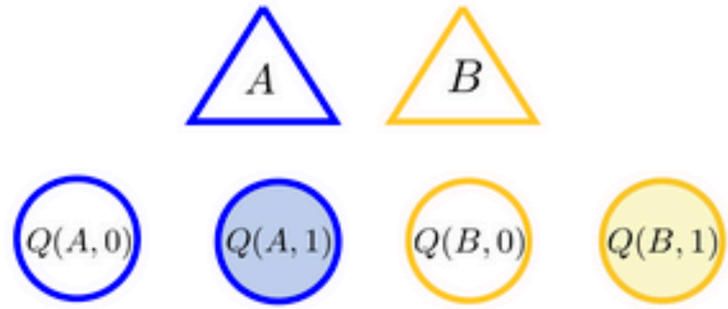
$$Q^*_2(B,0) = T(B,0,A) * (R(B,0) + \gamma V^*_1(A)) + T(B,0,B) * (R(B,0) + \gamma V^*_1(B))$$

$$= 0.7 * (2.0 + 0.9 * 1) + 0.3 * (2.0 + 0.9 * 2)$$

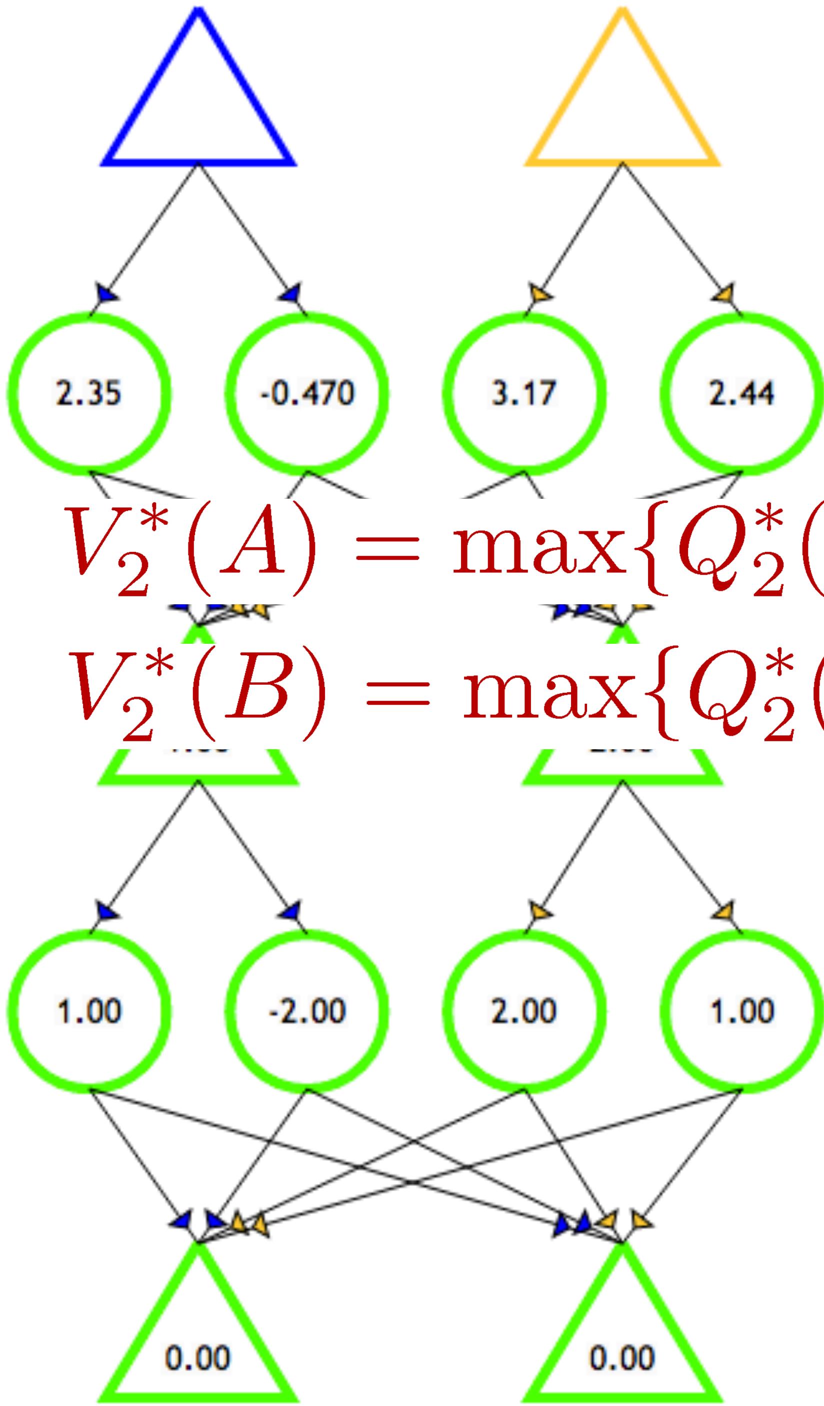


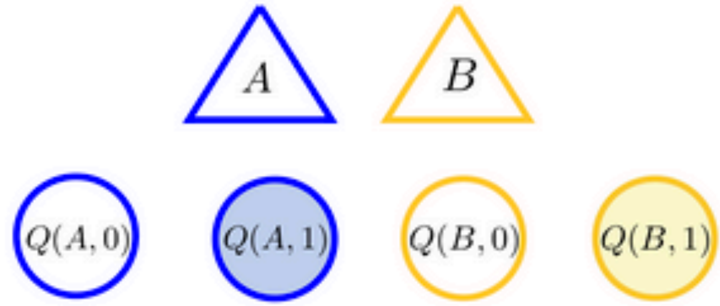
$$Q_2^*(B,1) = T(B,1,A)*(R(B,1)+\gamma V_1^*(A)) + T(B,1,B)*(R(B,1)+\gamma V_1^*(B))$$

$$= 0.4*(1.0+0.9*1)+0.6*(1.0+0.9*2)$$



				u)
				1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
B	1	A	0.40	1.0
B	1	B	0.60	1.0

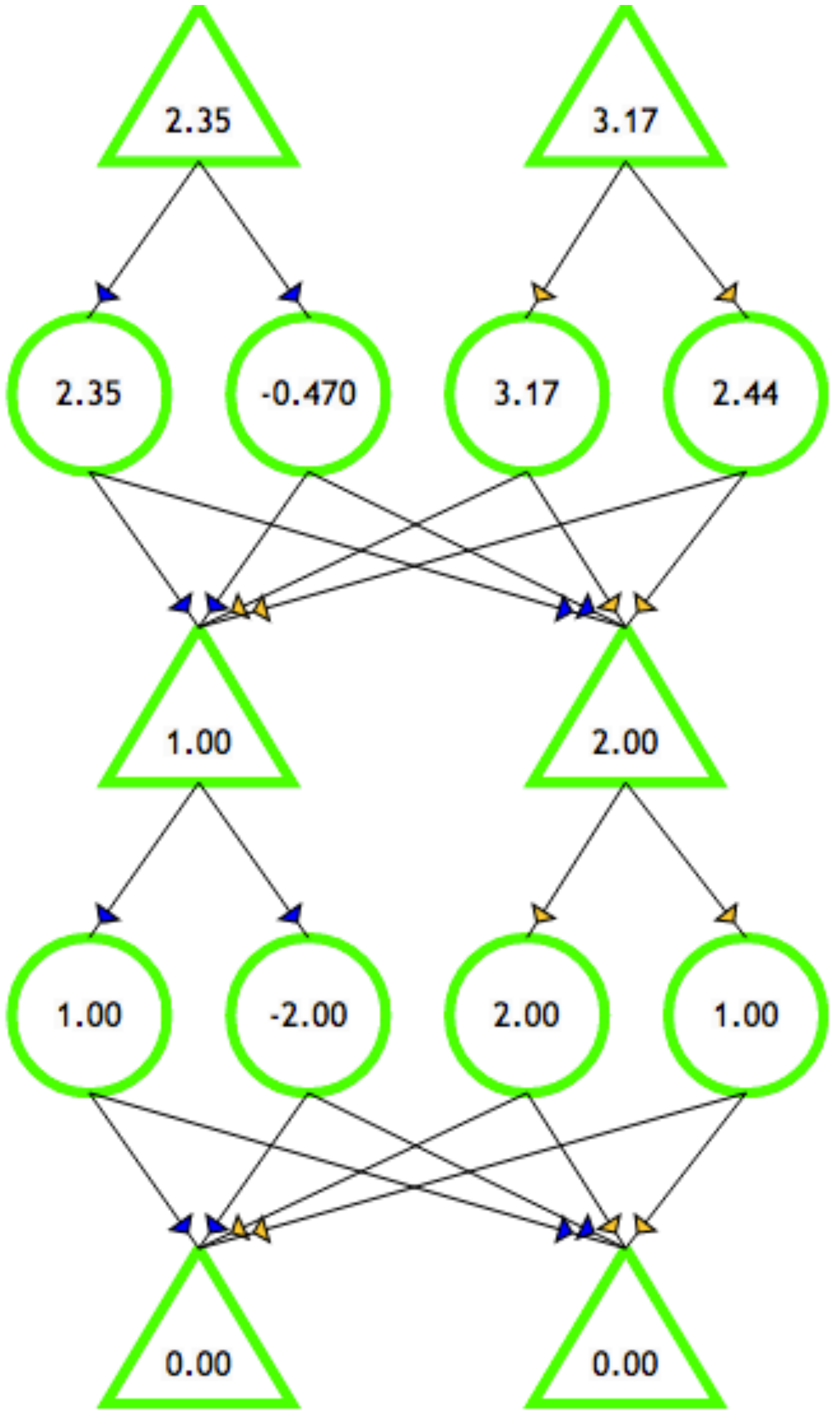


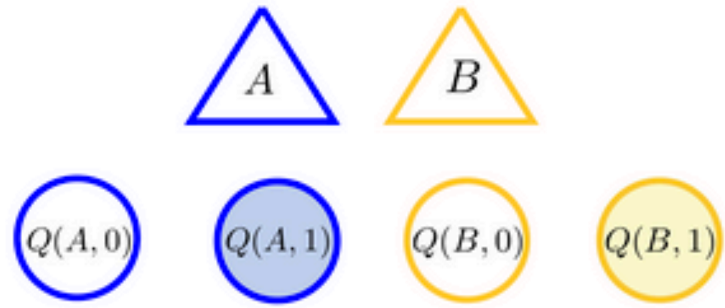


$$V_2^*(A) = \max\{Q_2^*(A,0), Q_1^*(A,1)\}$$

$$V_2^*(B) = \max\{Q_2^*(B,0), Q_1^*(B,1)\}$$

			(x,u,x')	$R(x,u)$
A	0	A	0.50	1.0
			0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
B	1	A	0.40	1.0
B	1	B	0.60	1.0





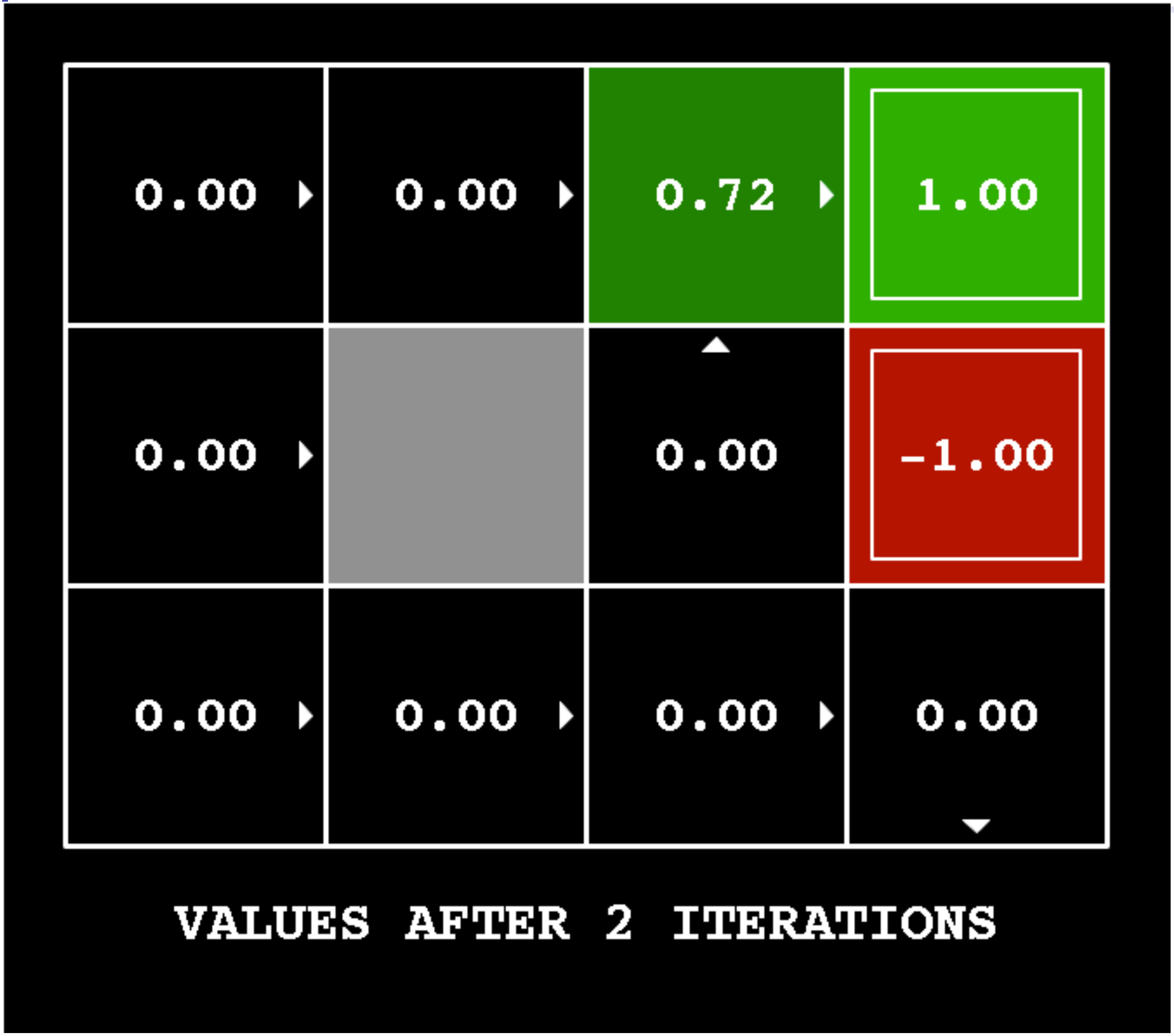
x	u	x'	$T(x,u,x')$	$R(x,u)$
A	0	A	0.50	1.0
A	0	B	0.50	1.0
A	1	A	0.30	-2.0
A	1	B	0.70	-2.0
B	0	A	0.70	2.0
B	0	B	0.30	2.0
B	1	A	0.40	1.0
B	1	B	0.60	1.0

Value Iteration in Gridworld

noise = 0.2, γ =0.9, two terminal states with R = +1 and -1

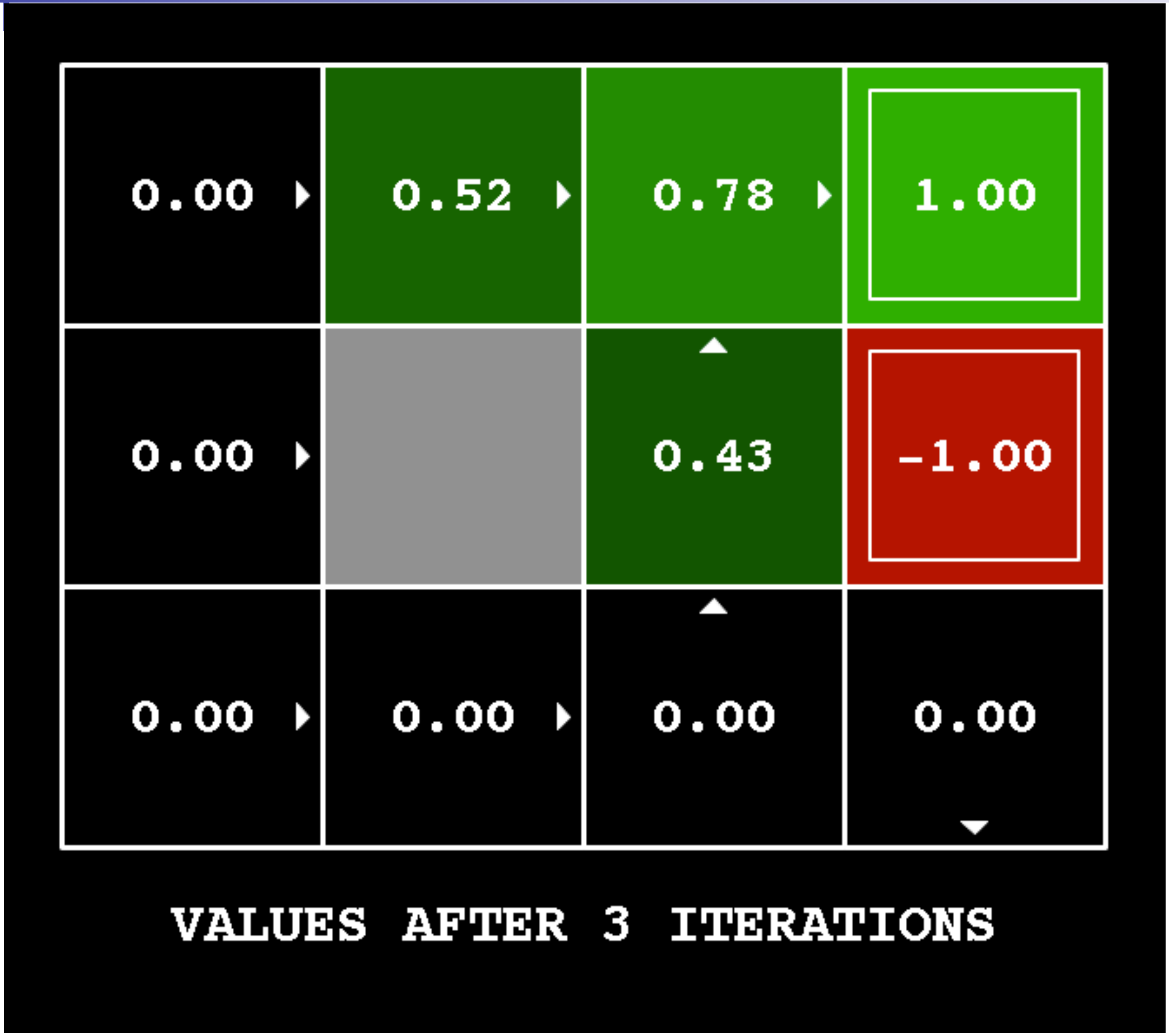
Value Iteration in Gridworld

noise = 0.2, γ =0.9, two terminal states with R = +1 and -1



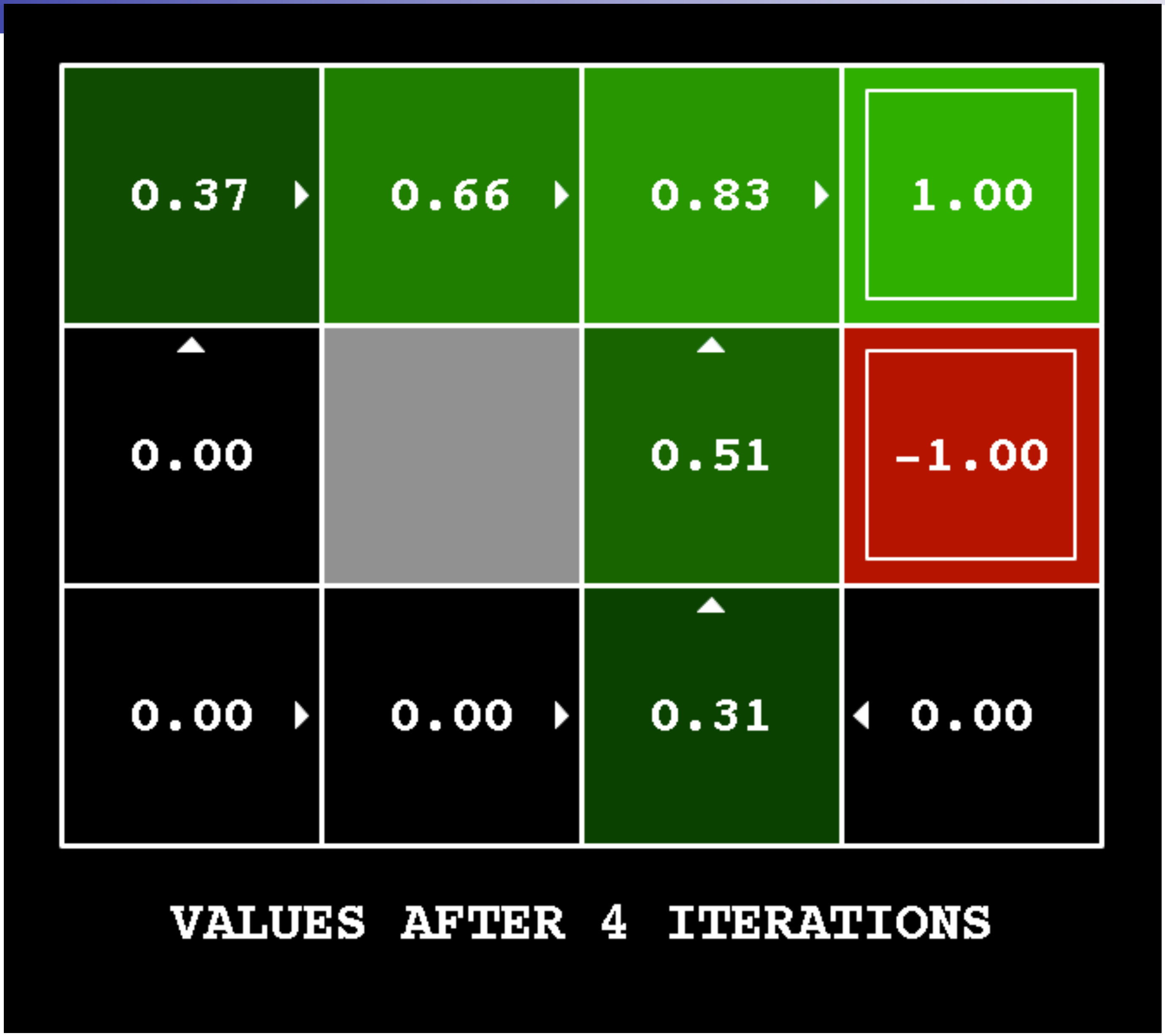
Value Iteration in Gridworld

noise = 0.2, γ =0.9, two terminal states with R = +1 and -1



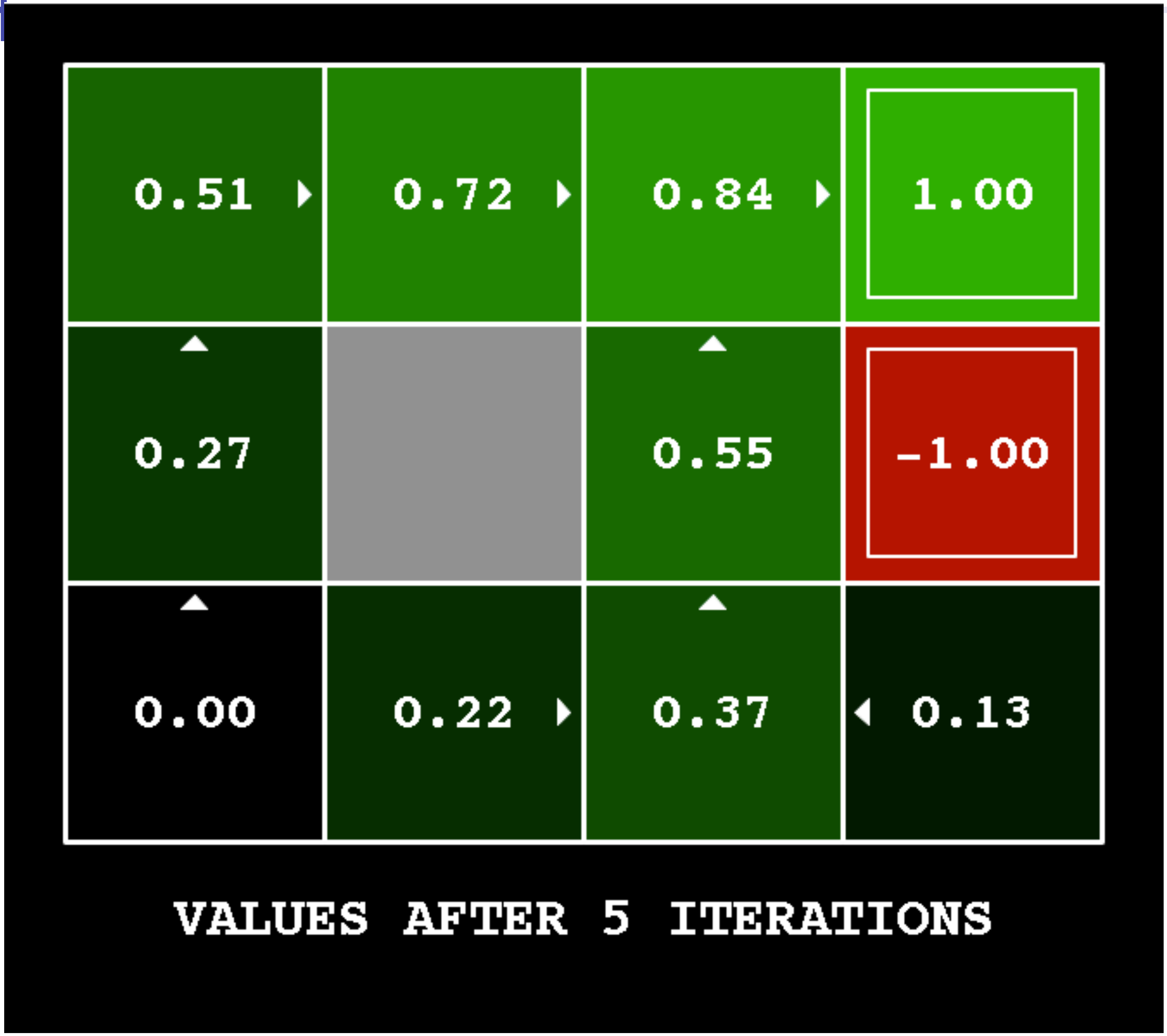
Value Iteration in Gridworld

noise = 0.2, γ =0.9, two terminal states with R = +1 and -1



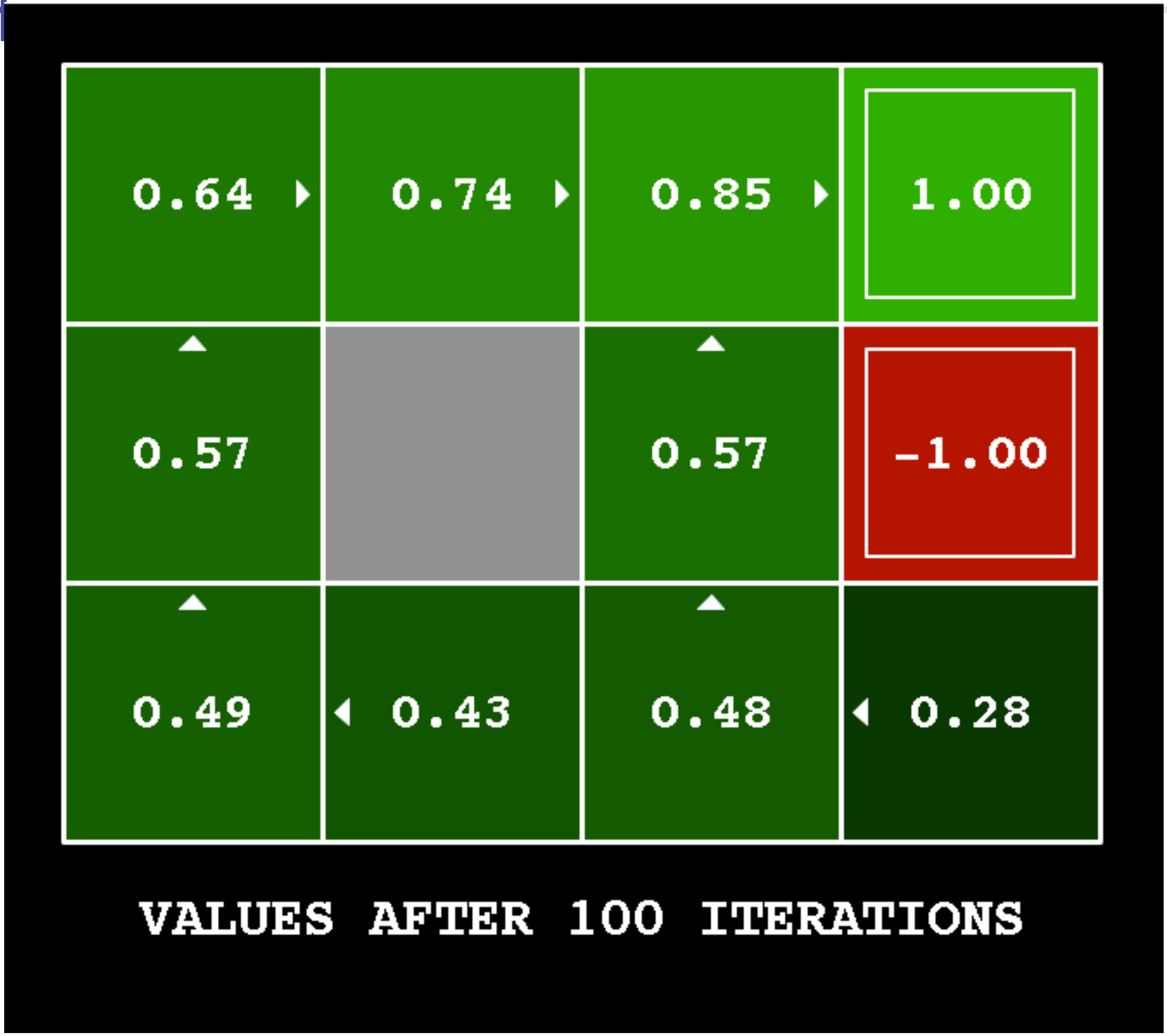
Value Iteration in Gridworld

noise = 0.2, γ =0.9, two terminal states with R = +1 and -1



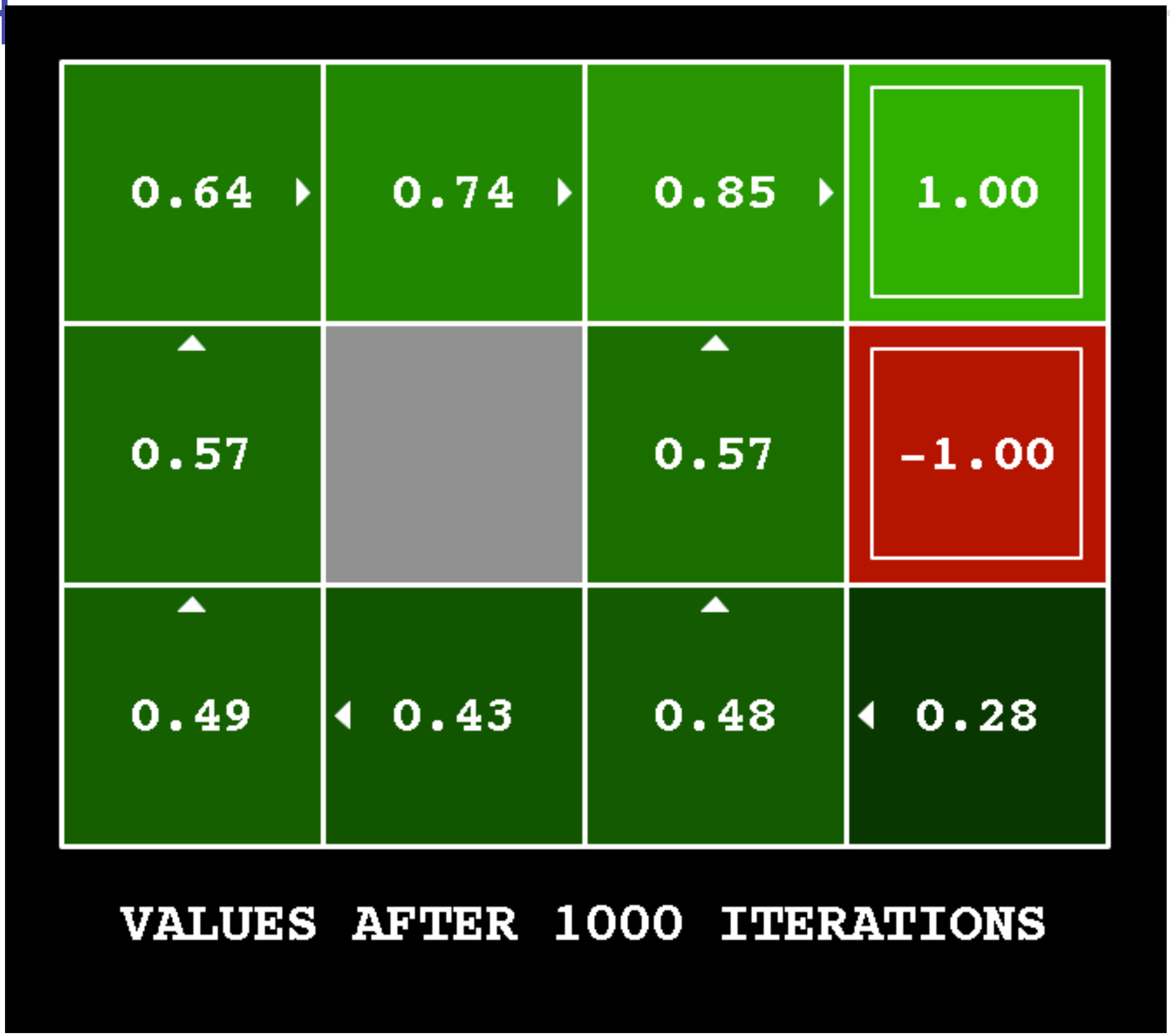
Value Iteration in Gridworld

noise = 0.2, γ =0.9, two terminal states with R = +1 and -1

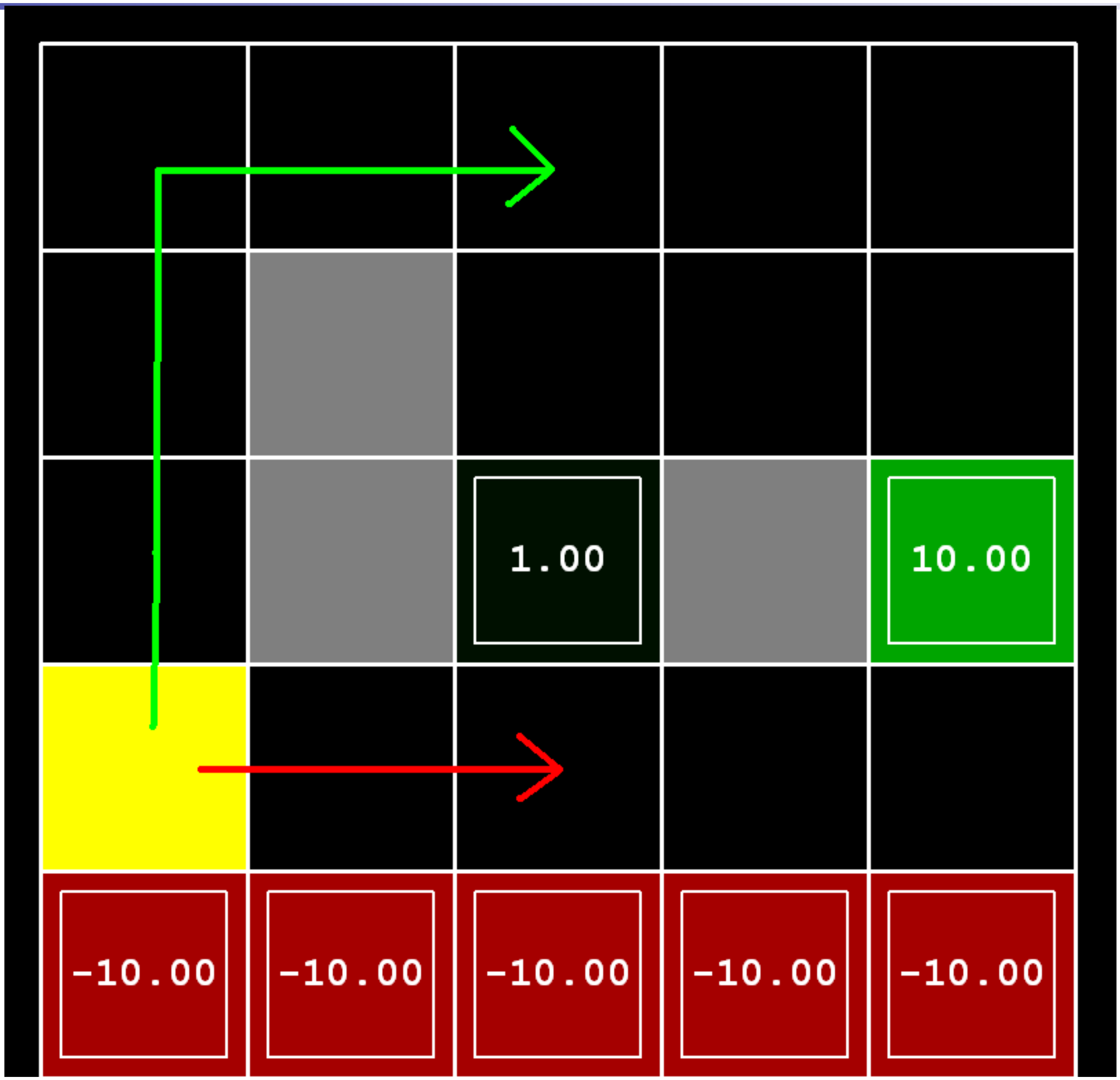


Value Iteration in Gridworld

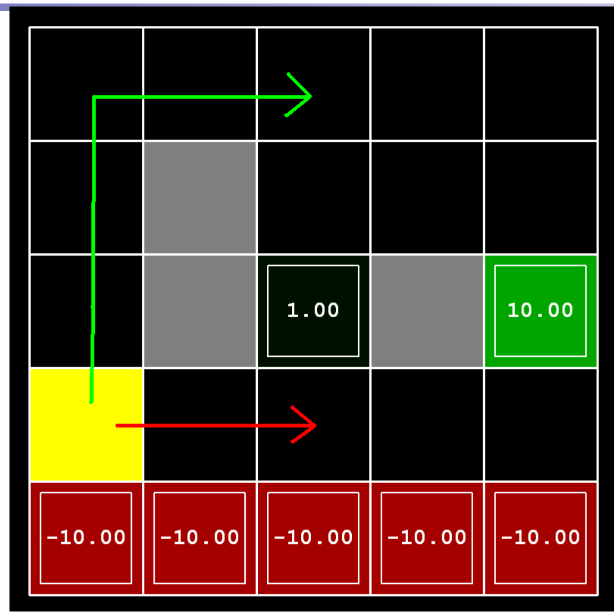
noise = 0.2, γ =0.9, two terminal states with R = +1 and -1



Exercise 1: Effect of discount, noise



Exercise 1: Effect of discount, noise



(a) Prefer the close exit (+1), risking the cliff (-10)

(b) Prefer the close exit (+1), but avoiding the cliff (-10)

(c) Prefer the distant exit (+10), risking the cliff (-10)

(d) Prefer the distant exit (+10), avoiding the cliff (-10)

Exercise 2

→ value iteration step through

Value Iteration Convergence

Theorem. Value iteration converges. At convergence, we have found the optimal value function V* for the discounted infinite horizon problem, which satisfies the Bellman equations

$$\forall x \in S: \quad V^*(x) = \max_u \sum_{x'} T(x, u, x') \left[R(x, u) + \gamma V^*(x')\right]$$

- Now we know how to act for infinite horizon with discounted rewards!
  - Run value iteration till convergence.
  - This produces V*, which in turn tells us how to act, namely following:

$$\pi^*(x) = \arg\max_{u \in A} \sum_{x'} T(x, u, x')[R(x, u) + \gamma V^*(x')]$$

- Note: the infinite horizon optimal policy is stationary, i.e., the optimal action at a state s is the same action at all times. (Efficient to store!)

Convergence and Contractions

- Define the max-norm: $||U|| = \max_s |U(s)|$

- Theorem: For any two approximations U and V

$$||U_{i+1} - V_{i+1}|| \leq \gamma \, ||U_i - V_i||$$

  - I.e. any distinct approximations must get closer to each other, so, in particular, any approximation must get closer to the true U and value iteration converges to a unique, stable, optimal solution

- Theorem:

$$||V_{i+1} - V_i|| < \epsilon, \Rightarrow ||V_{i+1} - V^*|| < 2\epsilon\gamma/(1-\gamma)$$

  - I.e. once the change in our approximation is small, it must also be close to correct

Policy Evaluation

- Recall value iteration iterates:

$$V_{i+1}^*(x) \leftarrow \max_u \sum_{x'} T(x,u,x')[R(x,u) + \gamma V_i^*(x')]$$

- Policy evaluation:

$$V_{i+1}^\pi(x) \leftarrow \sum_{x'} T(x,\pi(x),x')[R(x,\pi(x)) + \gamma V_i^\pi(x')]$$

  - At convergence:

$$\forall x \quad V^\pi(x) = \sum_{x'} T(x,\pi(x),x')[R(x,\pi(x)) + \gamma V^\pi(x')]$$

Exercise 3

Consider a stochastic policy $\mu(u|x)$, where $\mu(u|x)$ is the probability of taking action u when in state x. Which of the following is the correct value iteration update to perform policy evaluation for this stochastic policy?

1. $V^{\mu}_{i+1}(x) \leftarrow \max_u \sum_{x'} T(x,u,x')(R(x,u) + \gamma V^{\mu}_i(x'))$
2. $V^{\mu}_{i+1}(x) \leftarrow \sum_{x'} \sum_u \mu(u|x) T(x,u,x')(R(x,u) + \gamma V^{\mu}_i(x'))$
3. $V^{\mu}_{i+1}(x) \leftarrow \sum_u \mu(u|x) \max_{x'} T(x,u,x')(R(x,u) + \gamma V^{\mu}_i(x'))$