ECE 1508: Reinforcement Learning

Chapter 2: Model-based RL

Ali Bereyhi

ali.bereyhi@utoronto.ca

Department of Electrical and Computer Engineering

University of Toronto

Fall 2025

Bellman Equation: *Backup Diagram*

Bellman equation gives an

interesting visualization for values and action-values

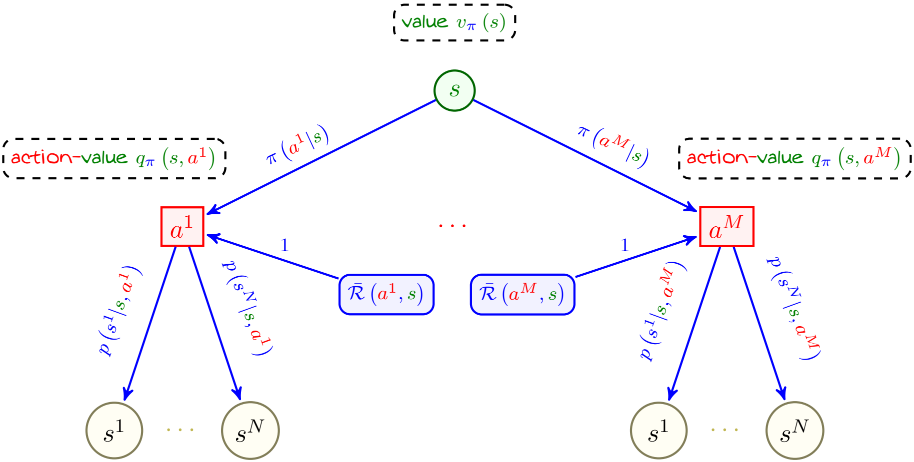
which can be shown in the so-called *backup diagram*

For simplicity, we consider $\gamma = 1$ in the backup diagram

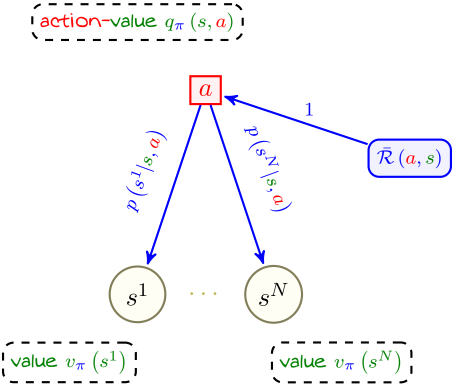
- *Each circle node is a state and carries the value of the state*
- *Each square node is an action and carries the action-value of the pair*
- *Each edge is a transition and carries a probability*
- *As we pass from leaves to root*
  - *Value of each node multiplies to its probability on the edge*
  - *They add up when they meet at a parent node*

    ↳ *This makes the value of the parent node*

Backup Diagram: *For Given Policy*



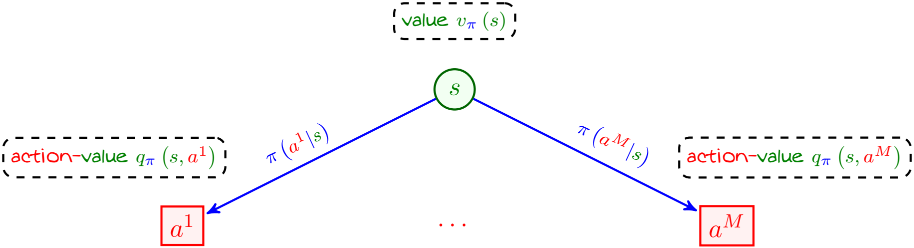
Backup Diagram: *For Given Policy*



Let's look at it part by part: *first we pass from leaves to action parent*

$$q_\pi(s,a) = \bar{\mathcal{R}}(s,a) + \sum_{n=1}^{N} v_\pi(s^n)\, p(s^n|s,a)$$

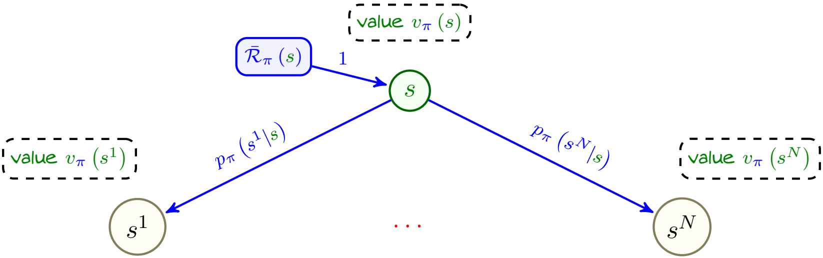
Backup Diagram: *For Given Policy*



Then, we pass from action parents to the root state

$$v_\pi(s) = \sum_{m=1}^{M} \pi(a^m|s)\, q_\pi(s, a^m)$$

Backup Diagram: *For Given Policy*



We could also have its alternative form expected over actions

$$v_\pi(s) = \bar{\mathcal{R}}_\pi(s) + \sum_{n=1}^{N} p_\pi(s^n|s)\, v_\pi(s^n)$$

Finding Optimal Values

+ *Well! Bellman lets us compute value of a given policy. But, how can we find the optimal value? It doesn't seem to solve this problem!*
– We can in fact use it to directly find the optimal values!
+ *That sounds a bit weird!*
– Once we know the *optimality constraint*, it doesn't anymore

Optimal Value: *Optimality Constraint*

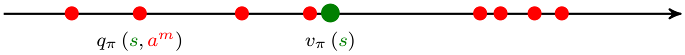
In Assignment 1, you show that *for any state we have*

$$v_\pi(s) = \sum_{m=1}^{M} q_\pi(s, a^m)\, \pi(a^m|s)$$

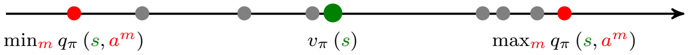
Now, recall that *policy is a conditional distribution meaning that*

$$0 \leqslant \pi(a^m|s) \leqslant 1$$

We can think of it as



Optimal Value: *Optimality Constraint*



It is hence obvious that

$$\min_m q_\pi(s, a^m) \leqslant v_\pi(s) \leqslant \max_m q_\pi(s, a^m)$$

We can use this simple fact to find a constraint on optimal values

Optimal Value: *Optimality Constraint*

If our policy is the optimal policy; then, we should have

$$v_\star(s) = \text{maximum possible value} = \max_m q_\star(s, a^m)$$

+ *But, can we guarantee that we can achieve such value?*
– Sure! We can set an optimal policy to

$$\pi^\star(a^m|s) = \begin{cases} 1 & m = \operatorname*{argmax}_m q_\star(s, a^m) \\ 0 & m \neq \operatorname*{argmax}_m q_\star(s, a^m) \end{cases}$$

+ *But, they are both in terms of $q_\star(s, a^m)$! We don't have the optimal action-values!*
– Sure! But, we could say that optimal values must satisfy this constraint: *if not, they cannot be optimal*

Optimal Value: *Optimality Constraint*

Optimality Constraint

Optimal value at each state s satisfies the following identity

$$v_\star(s) = \max_m q_\star(s, a^m)$$

and is achieved if we set the policy to

$$\pi^\star(a^m|s) = \begin{cases} 1 & m = \operatorname*{argmax}_m q_\star(s, a^m) \\ 0 & m \neq \operatorname*{argmax}_m q_\star(s, a^m) \end{cases}$$

which is an optimal policy

+ *But, how can we relate this constraint to Bellman equation?*
- *Let's see!*

Optimal Value: *Bellman Equation*

We know from Bellman equation II for action-value function that

$$q_\pi(s, a) = \bar{\mathcal{R}}(s, a) + \gamma \sum_{n=1}^{N} v_\pi(s^n) p(s^n | s, a)$$

If we play with optimal policy: *we are going to have same identity*

$$q_\star(s, a) = \bar{\mathcal{R}}(s, a) + \gamma \sum_{n=1}^{N} v_\star(s^n) p(s^n | s, a)$$

We now substitute it in *optimality constraint*

$$v_\star(s) = \max_m \left[\bar{\mathcal{R}}(s, a^m) + \gamma \sum_{n=1}^{N} v_\star(s^n) p(s^n | s, a^m) \right]$$

Optimal Value: *Bellman Equation*

This is again a recursive equation that

does not depend on any policy!

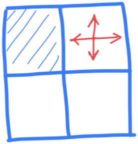
Bellman Optimality Equation

The optimal value function $v_\star(s)$ *satisfies*

$$v_\star(s) = \max_m \left[\bar{\mathcal{R}}(s, a^m) + \gamma \sum_{n=1}^{N} v_\star(s^n)\, p(s^n | s, a^m) \right]$$

We can again treat it as a fixed-point equation and solve it for $v_\star(s)$

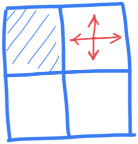
Example: *Dummy Grid World*



Let's find optimal values for our dummy grid world: we first find $\bar{\mathcal{R}}(s, a)$

$$\bar{\mathcal{R}}(0, a) = 0 \quad \bar{\mathcal{R}}(1, 0) = -1 \quad \bar{\mathcal{R}}(2, 0) = -0.5 \quad \bar{\mathcal{R}}(3, 0) = -1$$
$$\bar{\mathcal{R}}(1, 1) = -1 \quad \bar{\mathcal{R}}(2, 1) = -0.5 \quad \bar{\mathcal{R}}(3, 1) = -0.5$$
$$\bar{\mathcal{R}}(1, 2) = -0.5 \quad \bar{\mathcal{R}}(2, 2) = -1 \quad \bar{\mathcal{R}}(3, 2) = -0.5$$
$$\bar{\mathcal{R}}(1, 3) = -0.5 \quad \bar{\mathcal{R}}(2, 3) = -1 \quad \bar{\mathcal{R}}(3, 3) = -1$$

Example: *Dummy Grid World*

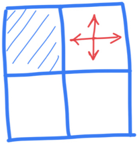


We next write down Bellman equations

1. *Since $s = 0$ is a terminal state we know that $v_\star(0) = 0$*
2. *Now, let's consider $s = 1$*

$$\begin{aligned} p(0|1,0) &= 1 \\ p(1|1,0) &= 0 \\ p(2|1,0) &= 0 \\ p(3|1,0) &= 0 \end{aligned} \rightsquigarrow \sum_{\bar{s}=0}^{4} v_\star(\bar{s})\, p(\bar{s}|1,0) = v_\star(0) = 0$$

Example: *Dummy Grid World*

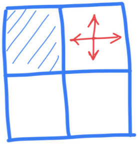


We next write down Bellman equations

1. *Since* $s = 0$ *is a terminal state we know that* $v_\star(0) = 0$
2. *Now, let's consider* $s = 1$

$$\begin{aligned} p(0|1,1) &= 0 \\ p(1|1,1) &= 0 \\ p(2|1,1) &= 0 \\ p(3|1,1) &= 1 \end{aligned} \rightsquigarrow \sum_{\bar{s}=0}^{4} v_\star(\bar{s})\, p(\bar{s}|1,1) = v_\star(3)$$

Example: *Dummy Grid World*

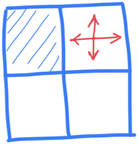


We next write down Bellman equations

1. *Since* $s = 0$ *is a terminal state we know that* $v_\star(0) = 0$
2. *Now, let's consider* $s = 1$

$$\begin{aligned} p(0|1,2) &= 0 \\ p(1|1,2) &= 1 \\ p(2|1,2) &= 0 \\ p(3|1,2) &= 0 \end{aligned} \rightsquigarrow \sum_{\bar{s}=0}^{4} v_\star(\bar{s})\, p(\bar{s}|1,2) = v_\star(1)$$

Example: *Dummy Grid World*

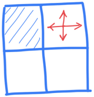


We next write down Bellman equations

1. *Since* $s = 0$ *is a terminal state we know that* $v_\star(0) = 0$
2. *Now, let's consider* $s = 1$

$$\begin{aligned} p(0|1,3) &= 0 \\ p(1|1,3) &= 1 \\ p(2|1,3) &= 0 \\ p(3|1,3) &= 0 \end{aligned} \rightsquigarrow \sum_{\bar{s}=0}^{4} v_\star(\bar{s})\, p(\bar{s}|1,3) = v_\star(1)$$

Example: *Dummy Grid World*

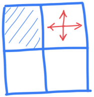


We next write down Bellman equations

1. *Since* $s = 0$ *is a terminal state we know that* $v_\star(0) = 0$
2. *Now, let's consider* $s = 1$

$$
\begin{aligned}
v_\star(1) &= \max_m \bar{\mathcal{R}}(1, a^m) + \sum_{\bar{s}=0}^{4} v_\star(\bar{s})\, p(\bar{s}|1, a^m) \\
&= \max\left\{-1, -1 + v_\star(3), -0.5 + v_\star(1), -0.5 + v_\star(1)\right\}
\end{aligned}
$$

Example: *Dummy Grid World*



We next write down Bellman equations

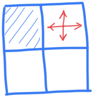
1. *Since* $s = 0$ *is a terminal state we know that* $v_\star(0) = 0$
2. *Now, let's consider* $s = 1$

$$v_\star(1) = \max\{-1, -1 + v_\star(3), -0.5 + v_\star(1), -0.5 + v_\star(1)\}$$

3. *Similarly, we have for* $s = 2$

$$v_\star(2) = \max\{-0.5 + v_\star(2), -0.5 + v_\star(2), -1 + v_\star(3), -1\}$$

Example: *Dummy Grid World*



We next write down Bellman equations

1. *Since* $s = 0$ *is a terminal state we know that* $v_\star(0) = 0$
2. *Now, let's consider* $s = 1$

$$v_\star(1) = \max\{-1, -1 + v_\star(3), -0.5 + v_\star(1), -0.5 + v_\star(1)\}$$

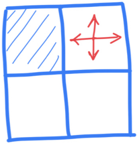
3. *Similarly, we have for* $s = 2$

$$v_\star(2) = \max\{-0.5 + v_\star(2), -0.5 + v_\star(2), -1 + v_\star(3), -1\}$$

4. *Finally for* $s = 3$, *we have*

$$v_\star(3) = \max\{-1 + v_\star(2), -0.5 + v_\star(3), -0.5 + v_\star(3), -1 + v_\star(1)\}$$

Example: *Dummy Grid World*

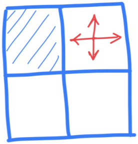


After sorting out the Bellman equations, we get

$$v_\star(1) = \max\{-1, -1 + v_\star(3), -0.5 + v_\star(1)\}$$
$$v_\star(2) = \max\{-1, -1 + v_\star(3), -0.5 + v_\star(2)\}$$
$$v_\star(3) = \max\{-1 + v_\star(2), -0.5 + v_\star(3), -1 + v_\star(1)\}$$

We should now solve this system of equations

Example: *Dummy Grid World*



We first note that

$$\max\left\{-1, -1 + v_\star(3), -0.5 + v_\star(1)\right\} \neq -0.5 + v_\star(1)$$

Proof: *Assume that*

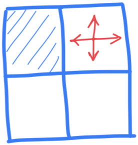
$$\max\left\{-1, -1 + v_\star(3), -0.5 + v_\star(1)\right\} = -0.5 + v_\star(1)$$

Then, we have

$$v_\star(1) - 0.5 + v_\star(1) \rightsquigarrow 0 = -0.5$$

impossible!

Example: *Dummy Grid World*



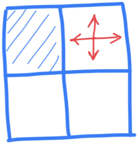
For the same reason, we have

$$\max\left\{-1, -1 + v_\star(3), -0.5 + v_\star(2)\right\} \neq -0.5 + v_\star(2)$$
$$\max\left\{-1 + v_\star(2), -0.5 + v_\star(3), -1 + v_\star(1)\right\} \neq -0.5 + v_\star(3)$$

So the equations reduce to

$$v_\star(1) = \max\left\{-1, -1 + v_\star(3)\right\} = v_\star(2)$$
$$v_\star(2) = \max\left\{-1, -1 + v_\star(3)\right\} = v_\star(1)$$
$$v_\star(3) = \max\left\{-1 + v_\star(2), -1 + v_\star(1)\right\} = -1 + v_\star(1)$$

Example: *Dummy Grid World*



Thus, we should only solve

$$v_\star(1) = \max\{-1, -1 + v_\star(3)\}$$
$$v_\star(3) = -1 + v_\star(1)$$

It is again easy to see that $\max\{-1, -1 + v_\star(3)\} \neq -1 + v_\star(3)$*; therefore,*

$$v_\star(1) = v_\star(2) = -1 \rightsquigarrow v_\star(3) = -2$$

Well! This is what we expected!

From Optimal Values to *Optimal Policy*

+ *What is the benefit then? It only finds optimal value, but we are looking for optimal policy!*
  - We can actually back-track optimal policy, once we have optimal value

---

The idea is quite simple:

1. *We can find optimal values from Bellman optimality equations*
2. *We could then find the optimal action-values*
3. *We finally get the optimal policy from optimal action-values*

Finding Optimal Policy: *Back-Tracking from Optimal Values*

We could summarize this approach algorithmically as follows

```
OptimBackTrack():
```

for $n = 1 : N$ **do**

    Solve Bellman equation $v_\star \left(s^n\right) = \max_m \bar{\mathcal{R}} \left(s^n, a^m\right) + \gamma \mathbb{E} \left\{ v_\star \left(\bar{S}\right) | s^n, a^m \right\}$

end for

for $n = 1 : N$ **do**

    for $m = 1 : M$ **do**

        Compute action-value $q_\star \left(s^n, a^m\right) = \bar{\mathcal{R}} \left(s^n, a^m\right) + \gamma \mathbb{E} \left\{ v_\star \left(\bar{S}\right) | s^n, a^m \right\}$

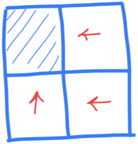
    end for

    Compute optimal policy via optimality constraint

$$\pi^\star \left(a^m | s\right) = \begin{cases} 1 & m = \operatorname*{argmax}_m q_\star \left(s, a^m\right) \\ 0 & m \neq \operatorname*{argmax}_m q_\star \left(s, a^m\right) \end{cases}$$

end for

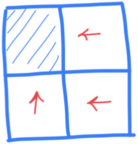
Example: *Dummy Grid World*



Let's find optimal policy at state $s = 1$ in our dummy grid world: first, we write

$$\begin{bmatrix} q_\star(1,0) \\ q_\star(1,1) \\ q_\star(1,2) \\ q_\star(1,3) \end{bmatrix} = \begin{bmatrix} \bar{\mathcal{R}}(1,0) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|1,0) \\ \bar{\mathcal{R}}(1,1) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|1,1) \\ \bar{\mathcal{R}}(1,2) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|1,2) \\ \bar{\mathcal{R}}(1,3) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|1,3) \end{bmatrix} = \begin{bmatrix} -1+0 \\ -1-2 \\ -0.5-1 \\ -0.5-1 \end{bmatrix} = \begin{bmatrix} -1 \\ -3 \\ -1.5 \\ -1.5 \end{bmatrix}$$

Example: *Dummy Grid World*



The optimal policy at state $s = 1$ is then given by

$$\pi^\star(a|1) = \begin{cases} 1 & a = \operatorname*{argmax}_a q_\star(1, a) \\ 0 & a \neq \operatorname*{argmax}_a q_\star(1, a) \end{cases} = \begin{cases} 1 & a = 0 \\ 0 & a \neq 0 \end{cases}$$

Well! We know that this is optimal in this problem!

Finding Optimal Policy: *Back-Tracking from Optimal Values*

+ *Wait a moment! Does that mean that our optimal policy is always deterministic? But, you said it could be also random!*
– Well! In some cases we could find random optimal policies as well!

If $q_\star(s, a^m)$ has a single maximizer; then,

optimal policy $\pi^\star(a^m|s)$ is deterministic

But, if it has multiple maximizers

optimal policy $\pi^\star(a^m|s)$ can also be random

Finding Optimal Policy: *General Form*

Generic Optimal Policy

Assume that $m^1, \ldots, m^J$ are all maximizers of $q_\star(s, a^m)$; then, policy

$$\pi^\star(a^m|s) = \begin{cases} p_1 & m = m^1 \\ \vdots & \\ p_J & m = m^J \\ 0 & m \notin \left\{m^1, \ldots, m^J\right\} \end{cases}$$

for any $p_1, \ldots, p_J$ that satisfy

$$\sum_{j=1}^{J} p_j = 1$$

is an optimal policy

Finding Optimal Policy

+ *But, why are all such policies optimal?*
– Well! We could look back at the optimality constraint

With any policy $\pi^\star(a|s)$ of the form given in the last slide, we have

$$v_{\pi^\star}(s) = \sum_{m=1}^{M} \pi^\star(a^m|s)\, q_{\pi^\star}(s, a^m) = \sum_{j=1}^{J} p_j q_{\pi^\star}\left(s, a^{m^j}\right) + 0$$

$$= \sum_{j=1}^{J} p_j \max_m q_{\pi^\star}(s, a^m) = \max_m q_{\pi^\star}(s, a^m) \sum_{j=1}^{J} p_j = \max_m q_{\pi^\star}(s, a^m)$$

which is the optimality constraint! It's intuitive, because

> *If we have multiple options for next action that give us same maximal value; then, we could randomly pick any of them*

Finding Optimal Policy

+ *But, still we could have a deterministic optimal* policy *in such cases! Right?!*
– Sure! *We could always have a deterministic optimal* policy*!*

Deterministic Optimal Policy

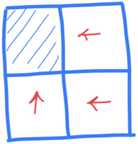
With known MDP for the environment, there exists at least one deterministic optimal policy

---

In the nutshell: if we know the complete state and its transition model

- *We always can find a deterministic optimal policy*
- *We might have multiple deterministic optimal policies*
  ↳ *In that case, we are going to have also random optimal policies*

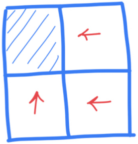
Example: *Dummy Grid World*



Let's find optimal policy at state $s = 3$ in our dummy grid world: first, we write

$$\begin{bmatrix} q_\star(3,0) \\ q_\star(3,1) \\ q_\star(3,2) \\ q_\star(3,3) \end{bmatrix} = \begin{bmatrix} \bar{\mathcal{R}}(3,0) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|3,0) \\ \bar{\mathcal{R}}(3,1) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|3,1) \\ \bar{\mathcal{R}}(3,2) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|3,2) \\ \bar{\mathcal{R}}(3,3) + \sum_{\bar{s}} v_\star(\bar{s})\, p(\bar{s}|3,3) \end{bmatrix} = \begin{bmatrix} -1-1 \\ -0.5-2 \\ -0.5-2 \\ -1-1 \end{bmatrix} = \begin{bmatrix} -2 \\ -2.5 \\ -2.5 \\ -2 \end{bmatrix}$$

Example: *Dummy Grid World*

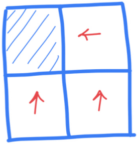


The optimal policy at state $s = 3$ is then given by

$$\pi^\star(a|3) = \begin{cases} 1 & a = \operatorname*{argmax}_a q_\star(3, a) \\ 0 & a \neq \operatorname*{argmax}_a q_\star(3, a) \end{cases} = \begin{cases} 1 & a = 0 \\ 0 & a \neq 0 \end{cases}$$

This is obviously optimal in this problem!

Example: *Dummy Grid World*

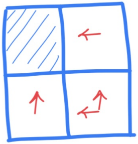


The optimal policy at state $s = 3$ is then given by

$$\pi^\star(a|3) = \begin{cases} 1 & a = \operatorname*{argmax}_a q_\star(3,a) \\ 0 & a \neq \operatorname*{argmax}_a q_\star(3,a) \end{cases} = \begin{cases} 1 & a = 3 \\ 0 & a \neq 3 \end{cases}$$

This is obviously optimal in this problem!

Example: *Dummy Grid World*

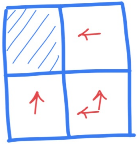


The optimal policy at state $s = 3$ is then given by

$$\pi^\star(a|3) = \begin{cases} 0.5 & a = 0 \\ 0 & a = 1, 2 \\ 0.5 & a = 3 \end{cases}$$

This is also optimal in this problem!

Example: *Dummy Grid World*

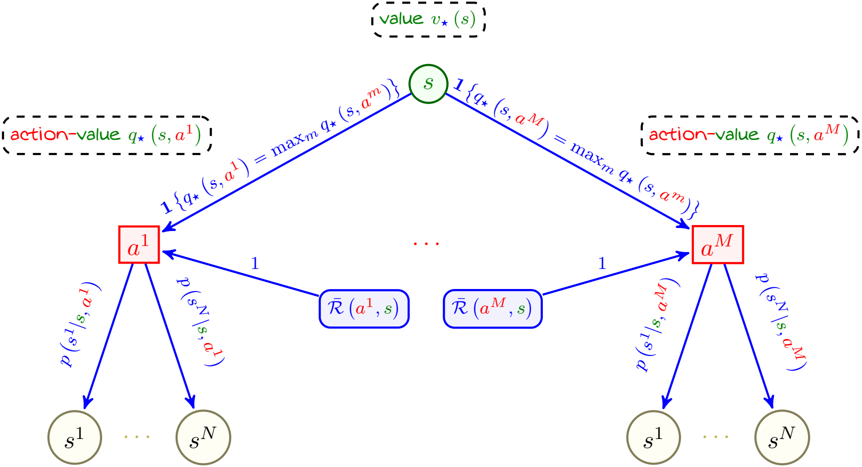


The optimal policy at state $s = 3$ *is then given by*

$$\pi^\star(a|3) = \begin{cases} 0.2 & a = 0 \\ 0 & a = 1, 2 \\ 0.8 & a = 3 \end{cases}$$

This is also optimal in this problem!

Backup Diagram: *For Optimal Policy*



Here, we assume $q_\star(s, a^m)$ has one maximizer $\equiv$ optimal policy is deterministic