The Bellman Equation

Self-Adaptive & Learning Algorithms

Milan R. Rapaić

Chair of Automatic Control
Computing and Control Department
Faculty of Technical Sciences
University of Novi Sad
Novi Sad • Serbia

November 1, 2023

Outline

1. Finding Policy Values in the Deterministic Case
2. Finding Policy Values in the Stochastic Case
3. Finding Optimal Values and Optimal Policies

Bellman Equation & Dynamic Programming

What is it? What problem does it solve?

Bellman Equation

Bellman Equation (in its various forms) is the equation which, if solved, would specify optimal action of every possible state. Any policy satisfying the Bellman Equation is an optimal policy.

Dynamic Programming

The term dynamic programming is often used in somewhat different contexts, however in the present context it will be used solely to indicate a group of techniques for finding an optimal policy by solving Bellman Equation.

Finding Policy Values in the Deterministic Case

1. Finding Policy Values in the Deterministic Case
2. Finding Policy Values in the Stochastic Case
3. Finding Optimal Values and Optimal Policies

Deterministic Bellman Equation

... or at least the seed from which it will be derived

$$g_k = \sum_{i=0}^{T} \gamma^i r_{k+i} = r_0 + \sum_{i=1}^{T} \gamma^i r_{k+i} = r_0 + \gamma \sum_{i=0}^{T} \gamma^i r_{k+1+i}$$

$$g_k = r_k + \gamma g_{k+1}$$

$$g_\pi(s) = r + \gamma g_\pi(s^+)$$

A recursive formula for state values (Deterministic Case)

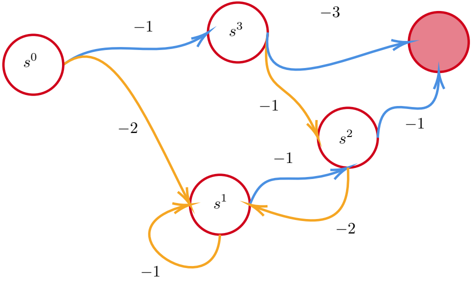
$$v_\pi(s) = h(s, \pi(s)) + \gamma v_\pi(f(s, \pi(s)))$$

Example

Write Bellman Equations and evaluate State Values

POLICY

Use blue action.



Bellman Equation

$$v^0 = -1 + \gamma v^3$$

$$v^1 = -1 + \gamma v^2$$

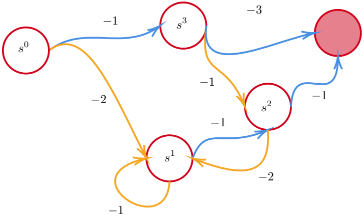
$$v^2 = -1 + \gamma v^{\text{term}}$$

$$v^3 = -3 + \gamma v^{\text{term}}$$

Bellman Equations for a given policy are always a set of linear equations!

Evaluation of State Values for a Given Policy

... by solving a linear matrix equation



$$\underbrace{\begin{bmatrix} v^0 \\ v^1 \\ v^2 \\ v^3 \end{bmatrix}}_{\mathbf{v}} = \gamma \underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} v^0 \\ v^1 \\ v^2 \\ v^3 \end{bmatrix}}_{\mathbf{v}} \underbrace{+\gamma \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} v^{\text{term}} + \begin{bmatrix} -1 \\ -1 \\ -1 \\ -3 \end{bmatrix}}_{\mathbf{r}}$$

Direct Solution

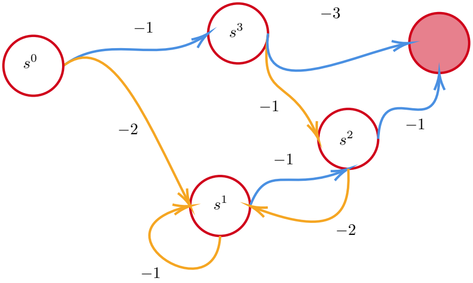
$$\mathbf{v} = (\mathbf{I} - \gamma \mathbf{A})^{-1} \mathbf{r}$$

Iterative Solution

$$\mathbf{v}^{k+1} = \gamma \mathbf{A} \mathbf{v}^k + \mathbf{r} \ , \quad \mathbf{v}^0 \in \text{rnd}$$

Evaluation of State Values for a Given Policy

... by backtracking from the terminal state



- choose $v^{\text{term}} = 0$
- Find all states from which the terminal state can be reached directly (s^2 and s^3)

$$v^2 = -1 + \gamma v^{\text{term}} = -1$$

$$v^3 = -3 + \gamma v^{\text{term}} = -3$$

- Find a state from which it is possible to directly reach a state of known value (s^0 and s^1)

$$v^0 = -1 + \gamma v^3 = -3.7$$

$$v^1 = -1 + \gamma v^2 = -1.9$$

Bellman Equation

... for the state-action value of an arbitrary given policy in the deterministic case

$$g_k = r_k + \gamma g_{k+1}$$

$$q_\pi(s_k, a_k) = r_k + \gamma q_\pi(s_{k+1}, a_{k+1})$$

A recursive formula for state-action values (Deterministic Case)

$$q_\pi(s, a) = h(s, a) + \gamma q_\pi(f(s, a), \pi(f(s, a)))$$

Note that this is still a set of linear equations (although in more unknown variables compared to the equations used to evaluate state values).

Finding Policy Values in the Stochastic Case

1 Finding Policy Values in the Deterministic Case

2 Finding Policy Values in the Stochastic Case

3 Finding Optimal Values and Optimal Policies

Bellman Equation for evaluating State-Values

... of a given policy in the stochastic case

$$g_\pi(s) = \mathbb{E}_\pi\left\{\sum_{i=0}^{T}\gamma^i R_{k+i}\big|S_0 = s\right\} = \mathbb{E}_\pi\left\{R_0 + \gamma\sum_{i=1}^{T}\gamma^{i-1}R_i\big|S_0 = s\right\}$$

$$g_\pi(s) = \sum_{a\in\mathcal{A}}\pi(a|s)\mathbb{E}_\pi\left\{R_0 + \gamma\sum_{i=1}^{T}\gamma^{i-1}R_i\big|S_0 = s, A_0 = a\right\}$$

$$g_\pi(s) = \sum_{a\in\mathcal{A}}\pi(a|s)\sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}}p(s^+,r|s,a)\mathbb{E}_\pi\left\{R_0 + \gamma\sum_{i=1}^{T}\gamma^{i-1}R_i\big|S_0 = s, A_0 = a, S_1 = s^+, R_0 = r\right\}$$

$$g_\pi(s) = \sum_{a\in\mathcal{A}}\pi(a|s)\sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}}p(s^+,r|s,a)\left[r + \gamma\mathbb{E}_\pi\left\{\sum_{i=1}^{T}\gamma^{i-1}R_i\big|S_1 = s^+\right\}\right]$$

A recursive formula for state values (Stochastic Case)

$$v_\pi(s) = \sum_{a\in\mathcal{A}}\pi(a|s)\sum_{\substack{s^+\in\mathcal{S}\\ r\in\mathcal{R}}}p(s^+,r|s,a)\left[r + \gamma v_\pi(s^+)\right]$$

Bellman Equation for evaluating State-Action values

... of a given policy in the stochastic case

$$q_\pi(s,a) = \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_0 = s, A_0 = a \right\}$$

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_0 = s, A_0 = a, S_1 = s^+, R_0 = r \right\}$$

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \mathbb{E}_\pi \left\{ \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_1 = s^+ \right\} \right]$$

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \sum_{a^+ \in \mathcal{A}} \pi(a^+|s^+) \mathbb{E}_\pi \left\{ \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_1 = s^+, A_1 = a^+ \right\} \right]$$

A recursive formula for state values (Stochastic Case)

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \sum_{a^+ \in \mathcal{A}} \pi(a^+|s^+) q_\pi(s^+, a^+) \right]$$

Finding Optimal Values and Optimal Policies

1. Finding Policy Values in the Deterministic Case
2. Finding Policy Values in the Stochastic Case
3. Finding Optimal Values and Optimal Policies

Bellman Equation

... for the state value of the optimal policy in the deterministic case

Since optimal policy is a policy, the recursive formula for computing v_π with $\pi \equiv \pi^*$ must hold

$$v_{\pi^*}(s) = h(s, \pi^*(s)) + \gamma v_{\pi^*}(f(s, \pi^*(s)))$$

Since v^* is optimal, the right hand side must be maximal among all possible policies

$$v_{\pi^*}(s) = \max_{\pi \in \mathscr{P}} \{g(s, \pi(s)) + \gamma v_{\pi^*}(f(s, \pi(s)))\}$$

Deterministic Bellman Equation for State Values

$$v^*(s) = \max_{a \in \mathcal{A}} \{g(s, a) + \gamma v^*(f(s, a))\}$$

Bellman Equation

... for the state value of the optimal policy in the stochastic case

Since optimal policy is a policy, the recursive formula for computing $v^* \equiv v_{\pi^*}$ holds

$$v_{\pi^*}(s) = \sum_{a \in \mathcal{A}} \pi^*(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v_{\pi^*}(s^+)\right]$$

Since v^* is optimal, the right hand side must be maximal among all possible policies

$$v_{\pi^*}(s) = \max_{\pi \in \mathscr{P}} \sum_{a \in \mathcal{A}} \pi^*(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v_{\pi^*}(s^+)\right]$$

Stochastic Bellman Equation for State Values

$$v^*(s) = \max_{a \in \mathcal{A}} \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v^*(s^+)\right]$$

Bellman Equation

... for the state-action values of the optimal policy in the deterministic case

Since optimal policy is a policy, the recursive formula for computing q_π with $\pi \equiv \pi^*$ must hold

$$q_{\pi^*}(s,a) = h(s,a) + \gamma q_{\pi^*}(f(s,a), \pi^*(f(s,a)))$$

Since v^* is optimal, the right hand side must be maximal among all possible policies

$$q_{\pi^*}(s,a) = \max_{\pi \in \mathscr{P}} \{h(s,a) + \gamma q_\pi(f(s,a), \pi(f(s,a)))\}$$

Deterministic Bellman Equation for State-Action Values

$$q^*(s,a) = h(s,a) + \gamma \max_{a^+ \in \mathcal{A}} q^*(f(s,a), a^+)$$

Bellman Equation

... for the state-action values of the optimal policy in the stochastic case

Since optimal policy is a policy, the recursive formula for computing q_π with $\pi \equiv \pi^*$ must hold

$$q_{\pi^*}(s,a) = \sum_{\substack{s^+ \in \mathscr{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \sum_{a^+ \in \mathcal{A}} \pi^*(a^+|s^+) q_{\pi^*}(s^+, a^+) \right]$$

Since v^* is optimal, the right hand side must be maximal among all possible policies

$$q_{\pi^*}(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \max_{\pi \in \mathscr{P}} \sum_{a^+ \in \mathcal{A}} \pi^*(a^+|s^+) q_{\pi^*}(s^+, a^+) \right]$$

Stochastic Bellman Equation for State-Action Values

$$q^*(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \max_{a^+ \in \mathcal{A}} q^*(s^+, a^+) \right]$$

Bellman Equation

... and its various beautiful forms :)

Bellman Equation for State Values

$$v^*(s) = \max_{a \in \mathcal{A}} \{g(s,a) + \gamma v^*(f(s,a))\}$$

$$v^*(s) = \max_{a \in \mathcal{A}} \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r | s, a) \left[r + \gamma v^*(s^+)\right]$$

Bellman Equation for State-Action Values

$$q^*(s,a) = h(s,a) + \gamma \max_{a^+ \in \mathcal{A}} q^*(f(s,a), a^+)$$

$$q^*(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r | s, a) \left[r + \gamma \max_{a^+ \in \mathcal{A}} q^*(s^+, a^+)\right]$$

Due to the appearance of the `max` operator, the Bellman equation is a set of coupled nonlinear equations. Direct, closed-form solution is no longer possible, not even in principle.

Further Reading

For further reading please consult [Sutton and Barto, 2018], **Chapter 4**.

References I

Sutton, R. S. and Barto, A. G. (2018).
Reinforcement Learning: An Introduction.
The MIT Press, second edition.