Chapter 3: The Reinforcement Learning Problem (Markov Decision Processes, or MDPs)

Objectives of this chapter:

- \Box present Markov decision processes—an idealized form of the AI problem for which we have precise theoretical results
- \Box introduce key components of the mathematics: value functions and Bellman equations

Agent and environment interact at discrete time steps: $t = 0, 1, 2, 3$, discrete time steps: $t = 0, 1, 2, 3, \ldots$ discrete time stens: $t = 0, 1, 2, 3$ t discrete time steps: $t = 0, 1, 2, 3, ...$ Agent and environment interact at discrete time steps: $t = 0, 1, 2, 3, \ldots$ some the environment include the the environment $S \in$ \mathbb{R}^2 \mathbb{S}

Agent observes state at step *t*: $S_t \in$ produces action at step $t: A_t \in \mathcal{A}(S_t)$ gets resulting reward: $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ and resulting next state: $S_{t+1} \in S^+$ $\mathcal S$ $S_{\rm t}$ set of all states, including the terminal states, including the terminal states, i.e., $\frac{1}{2}$ \mathbb{R} $S_t \in \mathcal{S}$ $A(S_t)$ 3^+ $\frac{z_i - z_i}{z_i - z_i}$ produces action at step $t: A_t \in \mathcal{A}(S_t)$ S^+ $S \in \mathcal{S}$ $p_t = 0$
 $q_t = 0$ $\mathcal{A}(S_t)$ $R \subset \mathbb{R}$ $\frac{1}{2}$. \sum_{t} Agent observes state at step t: S_t and resulting next state: $S_{t+1} \in S^+$ is the set of actions available in state *St*. One time step later, in part as a consequence of its action, the agent receives a numerical *reward*, *Rt*+1 2 R ⇢ R, where R is the set of possible rewards, and finds itself in a new state, *St*+1.

$$
\cdots \hspace{1cm} \dots \hspace{1cm} \overbrace{S_t} \hspace{1cm} \overbrace{A_t}^{\bullet \hspace{1.3cm} R_{t+1}} \hspace{1.5cm} \overbrace{S_{t+1}}^{R_{t+1}} \hspace{1.5cm} \overbrace{A_{t+1}}^{R_{t+2}} \hspace{1.5cm} \overbrace{S_{t+2}}^{R_{t+2}} \hspace{1.5cm} \overbrace{A_{t+2}}^{R_{t+3}} \hspace{1.5cm} \overbrace{S_{t+3}}^{R_{t+3}} \hspace{1.5cm} \overbrace{A_{t+3}}^{R_{t+4}} \hspace{1.5cm} \cdots
$$

Markov Decision Processes *finite Markov decision process (finite MDP)*. Finite MDPs are particularly important

- \Box If a reinforcement learning task has the Markov Property, it is basically a **Markov Decision Process (MDP)**. *, r|s, a*) *. , Rt*+1 = *r | S^t* =*s, A^t* =*a* T_{res} medically a Markoy Decision Process (MDP) basically a **wialkov decision process** (wid**p**).
- \Box If state and action sets are finite, it is a **finite MDP**. we present in the rest of the rest of the rest of the environment is a finite MDP. $G = 3.6$, one can compute as specified by (3.6), one can compute any $\frac{1}{2}$. **d** If state and action sets are finite, it is a **finite MDP**.
- \Box To define a finite MDP, you need to give: **T** To define a finite MDP, you need to give:
- ! **state and action sets** A particular finite \mathcal{L} one-step dynamics of the environment. Given any state and action *s* and *a*, pairs, *p*(*s* 0 *x*_{**e**} and action sets $\frac{1}{2}$
- **r** one-step "dynamics" These quantities completely specifically specifically specifically dynamics of a finite M

$$
p(s',r|s,a) = \mathbf{Pr}\{S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a\}
$$

• there is also: T there is also. filter the probabilities as also, one can compute a specified by \mathbf{S} , one can compute any the might else one mig

$$
p(s'|s, a) \doteq \Pr\{S_{t+1} = s' \mid S_t = s, A_t = a\} = \sum_{r \in \mathcal{R}} p(s', r|s, a)
$$

$$
r(s, a) \doteq \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a)
$$

Policy at step $t = \pi_t$

 a mapping from states to action probabilities $\pi_t(a \mid s) = \text{probability that } A_t = a \text{ when } S_t = s$

Special case - *deterministic policies*: π *t*(*s*) = the action taken with prob=1 when $S_t = s$

- \Box Reinforcement learning methods specify how the agent changes its policy as a result of experience.
- \Box Roughly, the agent's goal is to get as much reward as it can over the long run.

The Markov Property

- \Box By "the state" at step *t*, the book means whatever information is available to the agent at step *t* about its environment.
- \Box The state can include immediate "sensations," highly processed sensations, and structures built up over time from sequences of sensations. *3.5. THE MARKOV PROPERTY* 55
- □ Ideally, a state should summarize past sensations so as to retain all "essential" information, i.e., it should have the **Markov Property**: $\mathbf{q} = \mathbf{q} + \mathbf{q} + \mathbf{q}$ $\frac{1}{2}$ $\frac{1}{2}$ *A^t*¹, *Rt*, *St*, *At*. If the state signal has the *Markov property*, on the other , and all possible values of the past events: *S*0, *A*0, *R*1, ..., *S^t*¹, e it should have the **Markov** action representations at *t*, in which case the environment's dynamics can be

$$
\mathbf{Pr}\{R_{t+1} = r, S_{t+1} = s' \mid S_0, A_0, R_1, \dots, S_{t-1}, A_{t-1}, R_t, S_t, A_t\} =
$$

$$
p(s', r | s, a) = \mathbf{Pr}\{R_{t+1} = r, S_{t+1} = s' \mid S_t, A_t\}
$$

 \Box for all $s' \in \mathcal{S}^+, r \in \mathcal{R}$, and all histories S_0 , A_0 , R_1 , ..., S_{t-1} , A_t α represented α is the environment case the environment of α and α be entirely dynamics can be entirely dynamics can be entirely dynamics of α and α for all *r*, *s*⁰ , *St*, and *At*. In other words, a state signal has the Markov property, $s' \in \mathcal{S}^+, r \in \mathcal{R}$, and all histories $S_0, A_0, R_1, ..., S_{t-1}, A_{t-1}, R_t, S_t, A_t$.

The Meaning of Life (goals, rewards, and returns)

- The objective in RL is to maximize long-term future reward
- That is, to choose A_t so as to maximize $\ R_{t+1}, R_{t+2}, R_{t+3}, \ldots$
- But what exactly should be maximized?
- The <u>discounted *return* at time t</u>: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$ $\gamma \in [0,1)$ the *discount rate*

- The objective in RL is to maximize long-term future reward
- That is, to choose A_t so as to maximize $\ R_{t+1}, R_{t+2}, R_{t+3}, \ldots$
- But what exactly should be maximized?
- The <u>discounted *return* at time t</u>: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$ $\gamma \in [0,1)$ the *discount rate*

- The objective in RL is to maximize long-term future reward
- That is, to choose A_t so as to maximize $\ R_{t+1}, R_{t+2}, R_{t+3}, \ldots$
- But what exactly should be maximized?
- The <u>discounted *return* at time t</u>: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$ $\gamma \in [0,1)$ the *discount rate*

- The objective in RL is to maximize long-term future reward
- That is, to choose A_t so as to maximize $\ R_{t+1}, R_{t+2}, R_{t+3}, \ldots$
- But what exactly should be maximized?
- The <u>discounted *return* at time t</u>: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$ $\gamma \in [0,1)$ the *discount rate*

- The objective in RL is to maximize long-term future reward
- That is, to choose A_t so as to maximize $\ R_{t+1}, R_{t+2}, R_{t+3}, \ldots$
- But what exactly should be maximized?
- The <u>discounted *return* at time t</u>: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$ $\gamma \in [0,1)$ the *discount rate*

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0, 1)
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

*G*¹ = 130 *G*⁰ = 118

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 =
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

*G*¹ = 130 *G*⁰ = 118

 $G_4 = 0$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4=0 \quad G_3=
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

 $G_4 = 0$ $G_3 = 16$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

 $G_4 = 0$ $G_3 = 16$ $G_2 =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 =
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

*G*¹ = 130 *G*⁰ = 118

 $G_4 = 0$ $G_3 = 16$ $G_2 = -4$ $G_1 = 4$ $G_0 =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

*G*¹ = 130 *G*⁰ = 118

 $G_4 = 0$ $G_3 = 16$ $G_2 = -4$ $G_1 = 4$ $G_0 = 3$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

*G*¹ = 130 *G*⁰ = 118

 $G_4 = 0$ $G_3 = 16$ $G_2 = -4$ $G_1 = 4$ $G_0 = 3$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

 $G =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

*G*¹ = 130 *G*⁰ = 118

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

$$
G=\frac{1}{1-\gamma}
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0, 1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s. 1

$$
G = \frac{1}{1 - \gamma} = 2
$$

*G*¹ = 130 *G*⁰ = 118

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
 $G_2 =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0, 1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s. 1

$$
G = \frac{1}{1 - \gamma} = 2
$$

*G*¹ = 130 *G*⁰ = 118

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
 $G_2 = 26$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

*G*¹ = 130 *G*⁰ = 118

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
 $G_2 = 26$ $G_1 =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

*G*¹ = 130 *G*⁰ = 118

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
 $G_2 = 26$ $G_1 = 26$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

*G*¹ = 130 *G*⁰ = 118

• Suppose $\gamma = 0.5$ and the reward sequence is

 $R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13,$ and so on, all 13s $G_2 = 26$ $G_1 = 26$ $G_0 =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

*G*¹ = 130 *G*⁰ = 118

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
 $G_2 = 26$ $G_1 = 26$ $G_0 = 14$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0,1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

• Suppose $\gamma = 0.5$ and the reward sequence is

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s

$$
G_2 = 26 \quad G_1 = 26 \quad G_0 = 14
$$

• And if $\,\gamma=0.9?$

 $G_1 =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0, 1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

• Suppose $\gamma = 0.5$ and the reward sequence is

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
 $G_2 = 26$ $G_1 = 26$ $G_0 = 14$

• And if $\,\gamma=0.9?$

 $G_1 = 130$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0, 1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

• Suppose $\gamma = 0.5$ and the reward sequence is

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
 $G_2 = 26$ $G_1 = 26$ $G_0 = 14$

• And if $\,\gamma=0.9?$ $G_1 = 130$ $G_0 =$

$$
G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \qquad \gamma \in [0, 1)
$$

 $R_1 = 1, R_2 = 6, R_3 = -12, R_4 = 16$, then zeros for R_5 and later

• What are the following returns?

$$
G_4 = 0 \quad G_3 = 16 \quad G_2 = -4 \quad G_1 = 4 \quad G_0 = 3
$$

• Suppose $\gamma = 0.5$ and the reward sequence is all 1s.

$$
G = \frac{1}{1 - \gamma} = 2
$$

• Suppose $\gamma = 0.5$ and the reward sequence is

$$
R_1 = 1, R_2 = 13, R_3 = 13, R_4 = 13
$$
, and so on, all 13s
\n $G_2 = 26$ $G_1 = 26$ $G_0 = 14$
\n• And if $\gamma = 0.9$?

4 value functions

- All theoretical objects, mathematical ideals (expected values)
- Distinct from their estimates:

 $V_t(s)$ $Q_t(s,a)$

Values are *expected* returns

• The value of a state, given a policy:

 $v_{\pi}(s) = \mathbb{E}\{G_t \mid S_t = s, A_{t:\infty} \sim \pi\}$ $v_{\pi}: S \to \Re$

- The value of a state-action pair, given a policy: $q_{\pi}(s, a) = \mathbb{E}\{G_t \mid S_t = s, A_t = a, A_{t+1:\infty} \sim \pi\}$ $q_{\pi}: S \times A \rightarrow \Re$
- The optimal value of a state:

$$
v_*(s) = \max_{\pi} v_{\pi}(s) \qquad v_* : \mathcal{S} \to \mathcal{R}
$$

• The optimal value of a state-action pair:

$$
q_*(s, a) = \max_{\pi} q_{\pi}(s, a) \qquad q_* : S \times A \to \Re
$$

- Optimal policy: π_* is an optimal policy if and only if $\pi_*(a|s) > 0$ only where $q_*(s, a) = \max_b q_*(s, b)$ $\forall s \in \mathcal{S}$
	- in other words, π_* is optimal iff it is *greedy* wrt q_*

What policy is optimal? $A: left$ B. Right C. Other If $x=0$? $IF X = .99$ It $\gamma = \frac{1}{2}$?

Gridworld

- ❐ Actions: north, south, east, west; deterministic.
- \Box If would take agent off the grid: no move but reward $=-1$
- \Box Other actions produce reward = 0, except actions that move agent out of special states A and B as shown.

State-value function for equiprobable random policy; $\gamma = 0.9$

Golf

- \Box State is ball location
- \Box Reward of -1 for each stroke until the ball is in the hole
- \Box Value of a state?
- ❐ Actions:
	- · putt (use putter)
	- ! driver (use driver)
- ❐ putt succeeds anywhere on the green

Optimal Value Functions $S₁$

 \Box For finite MDPs, policies can be partially ordered: Capital letters are used for random variables and major algorithm variables. Lower case of the values of $\mathbf y$ ordered:

 $\pi \geq \pi'$ if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in S$ $\mathcal S$ Ω letters are used for Ω

 \Box There are always one or more policies that are better than or equal to all the others. These are the **optimal policies**. We denote them all π_* . α or α A(*s*) set of actions possible in state *s* \mathbf{u} \mathbf{f} uan or \mathbf{f} $\overline{\text{res}}$. $\overline{\text{ve}}$

 $v_*(s) = \max_{\tau} v_{\tau}(s)$ for all $s \in$ π **n** Optimal policies share the same optimal state-value function: □ Optimal policies also share the same optimal action-value *S^t* state at *t* $\mathcal S$ **ue runction:**

function: *A^t* action at *t*

$$
q_*(s,a) = \max_{\pi} q_{\pi}(s,a)
$$
 for all $s \in S$ and $a \in A$

This is the expected return for taking action *a* in state *s* and thereafter following an optimal policy. *T* final time state *s* and state *s* **S***t* state at θ *t* state at θ *t* state at θ *t* state at θ s_{total} sincludes, including the terminal states, s_{total} *t* discrete time step

s state

Why Optimal State-Value Functions are Useful

Any policy that is greedy with respect to v_* is an optimal policy.

Therefore, given v_* , one-step-ahead search produces the long-term optimal actions.

E.g., back to the gridworld:

 $c)$ π_*

Optimal Value Function for Golf <u>E l'unction for qui</u>

- \Box We can hit the ball farther with driver than with putter, but with less accuracy ϵ ball farther with dr ive $\overline{1}$
- $\Box q_*(s, \text{driver})$ gives the value or using driver first, then using whichever actions are best

Given q_* , the agent does not even have to do a one-step-ahead search:

$$
\pi_*(s) = \arg\max_a q_*(s, a)
$$

Value Functions x 4

Bellman Equations x 4

Bellman Equation for a Policy π

The basic idea:

$$
G_{t} = R_{t+1} + \gamma R_{t+2} + \gamma^{2} R_{t+3} + \gamma^{3} R_{t+4} + \cdots
$$

= $R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^{2} R_{t+4} + \cdots)$
= $R_{t+1} + \gamma G_{t+1}$

So:

$$
v_{\pi}(s) = E_{\pi} \{ G_{t} | S_{t} = s \}
$$

$$
= E_{\pi} \{ R_{t+1} + \gamma v_{\pi} (S_{t+1}) | S_{t} = s \}
$$

Or, without the expectation operator:

$$
v_{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s',r} p(s',r|s,a) \Big[r + \gamma v_{\pi}(s') \Big]
$$

R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction 20

More on the Bellman Equation

$$
v_{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s',r} p(s',r|s,a) \Big[r + \gamma v_{\pi}(s') \Big]
$$

This is a set of equations (in fact, linear), one for each state. The value function for π is its unique solution.

Backup diagrams:

Gridworld

- ❐ Actions: north, south, east, west; deterministic.
- \Box If would take agent off the grid: no move but reward $=-1$
- \Box Other actions produce reward = 0, except actions that move agent out of special states A and B as shown.

State-value function for equiprobable random policy; $\gamma = 0.9$

Bellman Optimality Equation for v_* \bm{r} - \bm{r} \bm{r} and \bm{r} \bm{r} \bm{r} \bm{r} \bm{r} \bm{r} inan Opumanty Equation i ⇡(*a|s*) *p*(*s*⁰ *, r|s, a*) *r* + *v*⇡(*s*⁰)
)
) *,* (2)

The value of a state under an optimal policy must equal the expected return for the best action from that state:

$$
v_*(s) = \max_a q_{\pi_*}(s, a)
$$

\n
$$
= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]
$$

\n
$$
= \max_a \sum_{s',r} p(s',r|s,a) [r + \gamma v_*(s')].
$$

\nThe relevant backup diagram:
\n
$$
\sum_{s'} p(s',r|s,a) \sum_{s'} p(s
$$

 v_* is the unique solution of this system of nonlinear equations.

Bellman Optimality Equation for *q** $\overline{\mathbf{a}}$

The last two equations are two forms of the Bellman optimality equations $\mathcal{L}_{\mathcal{A}}$

$$
q_*(s, a) = \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \Big| S_t = s, A_t = a\Big]
$$

=
$$
\sum_{s', r} p(s', r|s, a) \Big[r + \gamma \max_{a'} q_*(s', a')\Big].
$$

The relevant backup diagram: *a* \mathbf{h}

 q_* is the unique solution of this system of nonlinear equations.

Why Optimal State-Value Functions are Useful

Any policy that is greedy with respect to v_* is an optimal policy.

Therefore, given v_* , one-step-ahead search produces the long-term optimal actions.

E.g., back to the gridworld:

 v_* **c**) π_*

Solving the Bellman Optimality Equation

- \Box Finding an optimal policy by solving the Bellman Optimality Equation requires the following:
	- accurate knowledge of environment dynamics;
	- ! we have enough space and time to do the computation;
	- the Markov Property.
- \Box How much space and time do we need?
	- ! polynomial in number of states (via dynamic programming methods; Chapter 4),
	- ! BUT, number of states is often huge (e.g., backgammon has about 10^{20} states).
- \Box We usually have to settle for approximations.
- \Box Many RL methods can be understood as approximately solving the Bellman Optimality Equation.

Summary

- ❐ Agent-environment interaction
	- **States**
	- **E** Actions
	- **E** Rewards
- Policy: stochastic rule for selecting actions
- Return: the function of future rewards agent tries to maximize
- \Box Episodic and continuing tasks
- Markov Property
- **Markov Decision Process**
	- **Transition probabilities**
	- **Expected rewards**

 \Box Value functions

- State-value function for a policy
- Action-value function for a policy
- **Optimal state-value function**
- **Optimal action-value function**
- ❐ Optimal value functions
- ❐ Optimal policies
- **Bellman Equations**
- The need for approximation