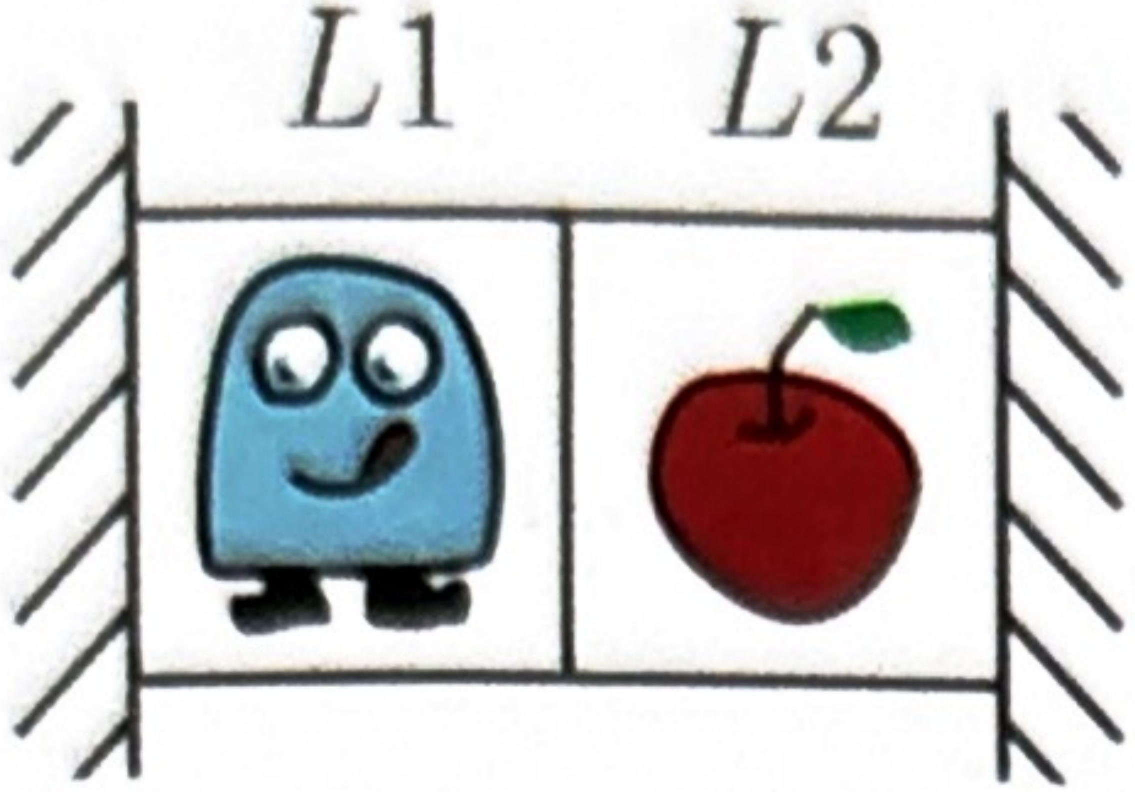
Bellman Equation

What if agent acts stochastically?

Jaewon Choi, 2nd of April, 2025

Today's Goal: Small Grid World



Environment: deterministic

Action: Stochastic

Continuous task
(not single episode)

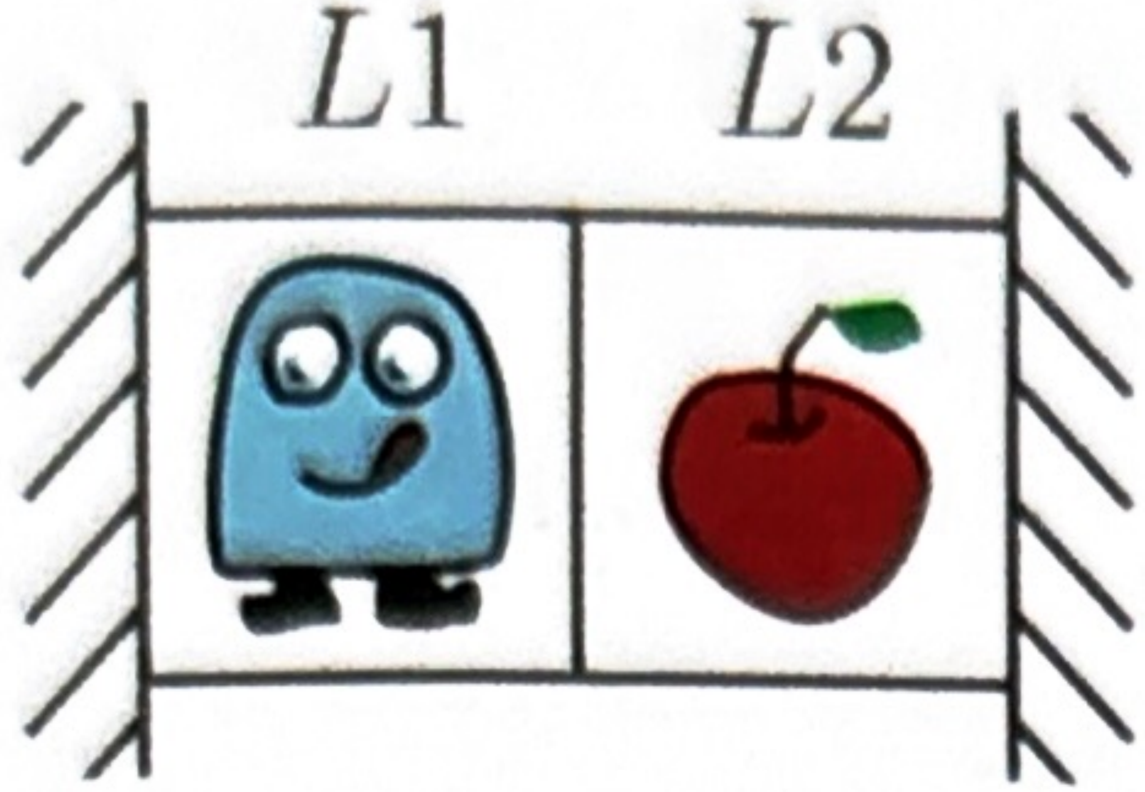
$$q_*(s, a) = ?$$

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

One more assumption: Stochastic action

- Agent moves left or right in 50% probability
- What is our expectation of return at s? Infinite? Zero? Specific number?

$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots$$



ANDREW G. BARTO AND RICHARD S. SUTTON IMAGE CREDITS:ACM

AI pioneers scoop Turing Award for reinforcement learning work

Richard E. Bellman

American applied mathematician (1920-1984)

- Inventor of
  - Bellman Equation and Dynamic Programming
  - Curse of dimensionality
  - Bellman-Ford algorithm
    - Shortest path finding with negative weighted edges

Goal: Optimal Bellman Equation

$$q_*(s, a) = \mathbb{E}\left[r(s, a, s') + \gamma \max_{a'} q_*(s', a')\right]$$

Primer: Probability and Expectation

$$\mathbb{E}[x] = \sum_{x} x \cdot p(x)$$

Primer: Joint Probability

$$p(x, y) = p(x)p(y \mid x)$$

Primer: Expectation of Joint Probability

$$\mathbb{E}[r(x, y)] = \sum_x \sum_y p(x, y) r(x, y)$$

$$= \sum_x \sum_y p(x) p(y \mid x) r(x, y)$$

Recall: Return

$$
\begin{aligned}
G_t &= R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \\
&= R_t + \gamma(R_{t+1} + \gamma R_{t+1} + \cdots) \\
&= R_t + \gamma G_{t+1}
\end{aligned}
$$

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

State-value function

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$$

State-value function over policy Pi

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t \,|\, S_t = s] \\
&= \mathbb{E}_\pi[R_t + \gamma G_{t+1} \,|\, S_t = s] \\
&= \mathbb{E}_\pi[R_t \,|\, S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s]
\end{aligned}
$$

Linearity!

Expectation of Reward

$$\mathbb{E}_\pi[R_t \mid S_t = s]$$

Expectation of Reward

$$\mathbb{E}_{\pi}[R_t \mid S_t = s]$$

$$\mathbb{E}[x] = \sum_{x} x \cdot p(x)$$

Expectation of Reward

$$\mathbb{E}_\pi[R_t \mid S_t = s]$$

$$\mathbb{E}[x] = \sum_x x \cdot p(x)$$

$$r(s, a, s')$$

Expectation of Reward

$$\pi(a \mid s)p(s' \mid s, a)$$

$$\mathbb{E}_\pi[R_t \mid S_t = s] \qquad \mathbb{E}[x] = \sum_x x \cdot p(x)$$

$$r(s, a, s')$$

Expectation of Reward

$$\pi(a \mid s)p(s' \mid s, a)$$

$$\mathbb{E}_\pi[R_t \mid S_t = s] \qquad \mathbb{E}[x] = \sum_x x \cdot p(x)$$

$$= \Sigma_a \Sigma_{s'} \pi(a \mid s)p(s' \mid s, a)r(s, a, s')$$

$$r(s, a, s')$$

Expectation of Reward

$$\pi(a \mid s)p(s' \mid s, a)$$

$$\mathbb{E}_\pi[R_t \mid S_t = s] \qquad\qquad \mathbb{E}[x] = \sum_x x \cdot p(x)$$

$$= \Sigma_a \Sigma_{s'} \pi(a \mid s) p(s' \mid s, a) r(s, a, s') \qquad\qquad r(s, a, s')$$

$$= \Sigma_a \pi(a \mid s) \Sigma_{s'} p(s' \mid s, a) r(s, a, s')$$

Expectation of Future Returns

$$\gamma \mathbb{E}_{\pi}[G_{t+1} \mid S_t = s]$$

Expectation of Future Returns

$$\gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s]$$

$$v_\pi(s) = \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s]$$

Expectation of Future Returns

$$
\begin{aligned}
\mathbb{E}_\pi[G_{t+1} \,|\, S_t = s] &= \Sigma_{a,s'} \pi(a \,|\, s) p(s' \,|\, s, a) \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] \\
&= \Sigma_{a,s'} \pi(a \,|\, s) p(s' \,|\, s, a) v_\pi(s')
\end{aligned}
$$

Summing up…

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[R_t \,|\, S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s] \\
&= \Sigma_{a,s'} \pi(a \,|\, s) p(s' \,|\, s, a) r(s, a, s') + \gamma \Sigma_{a,s'} \pi(a \,|\, s) p(s' \,|\, s, a) v_\pi(s') \\
&= \Sigma_{a,s'} \pi(a \,|\, s) p(s' \,|\, s, a) r(s, a, s') + \gamma v_\pi(s')
\end{aligned}
$$

“Bellman Equation”: Infinite series to system of equations

NEW: action-value function

Or, Q-function!

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

NEW: action-value function

Or, Q-function!

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

$$v_\pi(s) = \Sigma_a \pi(a \mid s) q_\pi(s, a)$$

Bellman Equation for Q-function

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_t \,|\, S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s, A_t = a]
\end{aligned}
$$

Bellman Equation for Q-function

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_t \,|\, S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \Sigma_{s'} p(s' \,|\, s, a) r(s, a, s') + \gamma \Sigma_{s'} p(s' \,|\, s, a) \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s']
\end{aligned}
$$

Bellman Equation for Q-function

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_t \,|\, S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \Sigma_{s'} p(s' \,|\, s, a) r(s, a, s') + \gamma \Sigma_{s'} p(s' \,|\, s, a) \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] \\
&= \Sigma_{s'} p(s' \,|\, s, a) \{ r(s, a, s') + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] \}
\end{aligned}
$$

Bellman Equation for Q-function

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_t \,|\, S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \Sigma_{s'} p(s' \,|\, s,a) r(s,a,s') + \gamma \Sigma_{s'} p(s' \,|\, s,a) \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] \\
&= \Sigma_{s'} p(s' \,|\, s,a) \{ r(s,a,s') + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] \} \\
&= \Sigma_{s'} p(s' \,|\, s,a) \{ r(s,a,s') + \gamma v_\pi(s') \}
\end{aligned}
$$

Bellman Equation for Q-function

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_t + \gamma G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_t \,|\, S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_t = s, A_t = a] \\
&= \Sigma_{s'} p(s' \,|\, s,a) r(s,a,s') + \gamma \Sigma_{s'} p(s' \,|\, s,a) \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] \\
&= \Sigma_{s'} p(s' \,|\, s,a) \{ r(s,a,s') + \gamma \mathbb{E}_\pi[G_{t+1} \,|\, S_{t+1} = s'] \} \\
&= \Sigma_{s'} p(s' \,|\, s,a) \{ r(s,a,s') + \gamma v_\pi(s') \} \\
&= \Sigma_{s'} p(s' \,|\, s,a) \{ r(s,a,s') + \gamma \Sigma_{a'} \pi(a' \,|\, s') q_\pi(s',a') \}
\end{aligned}
$$

Bellman Optimal Equation of $v(s)$

$$
\begin{aligned}
v_\pi(s) &= \Sigma_{a,s'}\pi(a\,|\,s)p(s'\,|\,s,a)r(s,a,s') + \gamma v_\pi(s') \\
&= \Sigma_a\pi(a\,|\,s)\Sigma_{s'}p(s'\,|\,s,a)\{r(s,a,s') + \gamma v_\pi(s')\}
\end{aligned}
$$

Bellman Optimal Equation of $v(s)$

$$
\begin{aligned}
v_\pi(s) &= \Sigma_{a,s'} \pi(a \mid s) p(s' \mid s, a) r(s, a, s') + \gamma v_\pi(s') \\
&= \Sigma_a \pi(a \mid s) \Sigma_{s'} p(s' \mid s, a) \{ r(s, a, s') + \gamma v_\pi(s') \}
\end{aligned}
$$

$$
v_*(s) = \Sigma_a \pi_*(a \mid s) \Sigma_{s'} p(s' \mid s, a) \{ r(s, a, s') + \gamma v_*(s') \}
$$

Bellman Optimal Equation of $v(s)$

$$v_\pi(s) = \Sigma_{a,s'}\pi(a\,|\,s)p(s'\,|\,s,a)r(s,a,s') + \gamma v_\pi(s')$$

$$= \Sigma_a \pi(a\,|\,s)\Sigma_{s'}p(s'\,|\,s,a)\{r(s,a,s') + \gamma v_\pi(s')\}$$

$$v_*(s) = \Sigma_a \pi_*(a\,|\,s)\Sigma_{s'}p(s'\,|\,s,a)\{r(s,a,s') + \gamma v_*(s')\}$$

$$v_*(s) = \max_a \Sigma_{s'}p(s'\,|\,s,a)\{r(s,a,s') + \gamma v_*(s')\}$$

Bellman Optimal Equation of $q(s, a)$

$$q_\pi(s,a) = \Sigma_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma \Sigma_{a'} \pi(a' \mid s') q_\pi(s',a')\}$$

Bellman Optimal Equation of $q(s, a)$

$$q_\pi(s, a) = \Sigma_{s'} p(s' | s, a)\{r(s, a, s') + \gamma \Sigma_{a'} \pi(a' | s') q_\pi(s', a')\}$$

$$q_*(s, a) = \Sigma_{s'} p(s' | s, a)\{r(s, a, s') + \gamma \Sigma_{a'} \pi_*(a' | s') q_*(s', a')\}$$

Bellman Optimal Equation of $q(s, a)$

$$q_\pi(s,a) = \Sigma_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma \Sigma_{a'} \pi(a' \mid s') q_\pi(s',a')\}$$

$$q_*(s,a) = \Sigma_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma \Sigma_{a'} \pi_*(a' \mid s') q_*(s',a')\}$$

$$q_*(s,a) = \Sigma_{s'} p(s' \mid s,a)\{r(s,a,s') + \gamma \max_{a'} q_*(s',a')\}$$

Optimal policy from $q_*(s, a)$

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

Summary

- Bellman equation and Bellman optimal equation
  - Infinite sum to system of linear equation!
- Finding optimal policy is the ultimate goal of reinforcement learning
  - It’s easy to find the optimal policy if we can figure out v_{*}(s) of it

Thanks!