Stochastic Dynamic Optimisation

Connie Trojan

12th May 2022

Markov Decision Processes

- A **Markov decision process** (MDP) is a sequential decision-making process

- The decision maker or **agent** is in some state S from a finite **state space** $\mathcal{S}$ and must select some action A from a finite action set $\mathcal{A}(S)$.

- After taking an action, the process moves to the next time step, transitioning randomly to some new state S' according to fixed transition probabilities $\mathbb{P}(S' \mid S, A)$ and awarding a reward $R(S, A)$ to the agent.

Aims

- Find the decision rule or **policy** $\pi : \mathcal{S} \to \mathcal{A}$ that maximises expected reward in some sense.
- For tasks that can continue indefinitely, "maximising total reward" might not be a well defined objective.
- Might maximise the rate at which reward is accumulated via the long run **limiting average reward**:

$$\lim_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} R(S_t, A_t) \, .$$

- **Discounted reward**: maximise the total reward when a discount factor $\gamma \in [0, 1)$ is applied to future rewards.

$$\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) .$$

- The use of discounting might be motivated by a mechanism like inflation, by which rewards earned in the future are less valuable than those earned immediately.

- In MDPs, an optimal policy for the limiting average case can always be obtained by solving the discounted case for γ sufficiently close to 1.

The Bellman Optimality Equations

- It is always possible to find a **stationary** optimal policy.
- The expected values of each state under an optimal policy are uniquely defined by the **Bellman Optimality Equations**:

$$v(S) = \max_{A \in \mathcal{A}(S)} \left\{ R(S, A) + \gamma \mathbb{E} \left(v(S') \,|\, S, A \right) \right\}$$

- We can find these values by linear programming: find the smallest value vector that is $\geq$ the RHS for all states and actions.
- Also have algorithms like value and policy iteration that are guaranteed to find an optimal policy in a finite number of iterations.

Example: Blackjack

- On their turn, players have two choices: **hit** (be dealt another card from the deck) or **stick** (stop drawing cards).
- Assuming that the cards are dealt from a deck that is sufficiently large, card draws are i.i.d. and each player plays independently against the dealer.
- The state space is defined by the player's current total, the dealer's card, and whether or not the player has a useable ace.
- We also have three possible terminal states: `WIN`, `DRAW`, and `LOSE` with the player receiving a reward of 1, 0, or -1 respectively on transition.

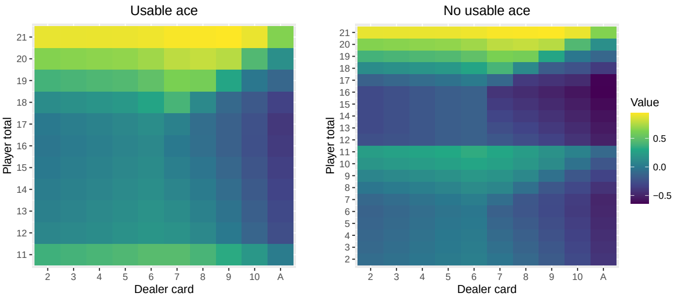

Figure: State values in blackjack

The state values can be found efficiently by linear programming since the state and action spaces are relatively small, $|\mathcal{S}| = 344$ and $|\mathcal{A}| = 2$.

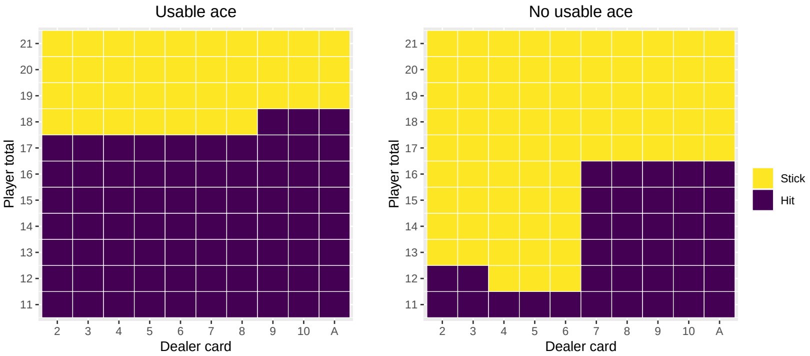

Figure: Optimal blackjack strategy

Can calculate expected reward as the expected value of your starting state: -0.047. True optimal strategy - don't play at all.

Stochastic Games

- A **stochastic game** is the multi-agent version of an MDP, where there is more than one agent or **player**, and the state transitions and rewards can depend on all of their choices.
- They can also be seen as the sequential, stochastic generalisation of a **matrix game**.
- The key objective is to identify **Nash equilibrium** policies, defined as pairs of strategies where neither player can get a better expected reward by unilaterally changing their strategy.

Key Questions

- Do stationary Nash equilibrium strategies exist?
- Are they "optimal"?
- How can we find them?

Key Questions (without easy answers)

- Do stationary Nash equilibrium strategies exist? **(not always)**
- Are they "optimal"? **(not always)**
- How can we find them? **(can be complicated)**

When do stationary equilibrium strategies exist?

- Nash equilibrium stationary strategies always exist for stochastic games with **discounted** rewards.
- Shapley's theorem for 2-player zero-sum SGs: The value $v_t(S)$ of starting the game in state S at time t is the value of the matrix game $\Gamma(S)$ with rewards:

$$[\Gamma(S)]_{i,j} = R(S, A_i^1, A_j^2) + \gamma \sum_{S' \in \mathcal{S}} \mathbb{P}\left(S' \mid S, A_i^1, A_j^2\right) v(S') \,.$$

- For limiting average reward, stationary NE strategies do **not** always exist.

Example: The Big Match

- Two-player zero-sum stochastic game with limiting average reward. Has 3 states with the reward matrices:

$$\Gamma(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Gamma(2) = \begin{pmatrix} 0 \end{pmatrix}, \quad \Gamma(3) = \begin{pmatrix} 1 \end{pmatrix}.$$

- In state 1, players can choose either 0 or 1. If both make the same choice, player 1 receives a reward of 1 from player 2.
- If player 1 chooses 0, they **stay** in state 1. If player 1 chooses 1, they are **absorbed** in state 2 if player 2 chose 0, or 3 otherwise.
- There are no stationary NE strategies.

What is "optimal" play?

- If all of the games are **zero-sum**, all Nash equilibria will have the same value. If we can find a stationary NE then we are done.
- Otherwise, it is possible to have multiple (stationary) Nash equilibria with different values.
- Might be necessary to relax objective of finding stationary equilibrium strategies: e.g. **cyclic equilibria**.

Example: Iterated Prisoner's Dilemma

- The **prisoner's dilemma** has the following payoffs:

	You co-operate	You betray
Opponent co-operates	-1	0
Opponent betrays	-3	-2

- Unique Nash equilibrium: (betray, betray).
- When the game is repeated indefinitely (SG with one state), this is the only stationary Nash equilibrium.
- Many non-stationary NE strategies have better average/discounted reward: e.g. **tit-for-tat**.

How can we find stationary equilibria?

- Can for example derive an algorithm analogous to value iteration from Shapley's theorem.
- However, there is no guarantee of being able to find a stationary equilibrium policy from the game data in a finite number of iterations.
- It is possible for stochastic games to lack the **ordered field property**: a game with rational data could have an optimal strategy with irrational entries.
- Some classes of SG with this property are known, e.g. single and switching controller games.

Further Research

- When do rational stochastic games have the ordered field property?
- Multi-agent reinforcement learning - how can policies be learned from interaction with a system whose dynamics are not known?
- Partial observability - what can be done if we only have partial information about the state?

Any Questions?

References

Powell, W. B. *Approximate Dynamic Programming : Solving the Curses of Dimensionality.*
Wiley series in probability and statistics. J. Wiley Sons, Hoboken, N.J., 2nd edition, 2011.

Filar, J. and Vrieze, K. *Competitive Markov Decision Processes.*
Springer, New York, NY, 1997.

Zhang, K., Yang, Z., and Başar, T. *Multi-agent reinforcement learning: A selective overview of theories and algorithms.*
In Handbook of Reinforcement Learning and Control, pages 321–384. Springer International Publishing, Cham, 2021.