The Gambler's problem and beyond

Baoxiang Wang

Aug 12, 2020, CUHKSZ

Based on joint works with Shuai Li, Jiajin Li, and Siu On Chan

School of Data Science, CUHK Shenzhen

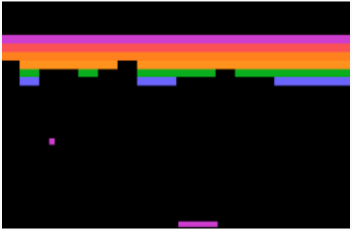
Outline

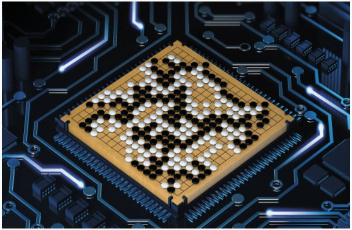
Outline of this talk

- Background on reinforcement learning and positioning this work
- Solving the Gambler's problem - The question #1 in the RL text book [SB18]
- What does it imply for reinforcement learning

Outline

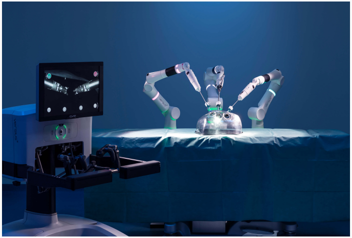
Reinforcement learning and sequential decisions





Milestones, in chronological order: Breakout in Atari 2600, AlphaGo and AlphaZero, Libratus and DeepStack, and AlphaStar

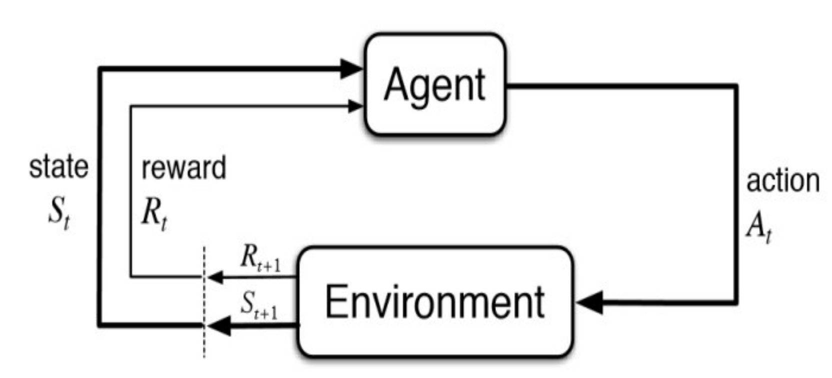
Reinforcement learning and sequential decisions



Applications: humanoid simulation, robot surgeon, robotics, and autonomous driving

Reinforcement learning and sequential decisions

Reinforcement learning: To model and learn sequential agent-environment interaction from reinforces



Sutton and Barto. Reinforcement learning: An introduction. 2018.

Connections to other areas

Cognitive science

- RL discusses the interaction between action and perception of an agent, while cognitive science studies that of humans.
- Cognitive science concepts are heavily adopted

Optimal control

- RL targets mostly model-free learning. To learn only from the reward signals *tabula rasa* without knowing the environment
- Optimal control is based on the model instead

Online learning

- RL is contextual multi-arm bandit with an additional dynamic: The action will impact the environment.

Outline

Position of this work

- This work is solves a sequential decision problem by analysis (not by learning algorithms)
- Technically, this work can be categorized into optimal control (to solve policy) and dynamical systems (to solve value)
- Despite these, it is a description of the optimum of the sequential decision processes and the corresponding learning problems

Outline

Markov decision processes (MDP) - Formulation

- RL is formulated as MDP - tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho_0, \gamma)$
  - $\mathcal{S} \subseteq \mathbb{R}^m$ state space, $\mathcal{A} \subseteq \mathbb{R}^n$ action space, $\rho_0 \in \Delta(\mathcal{S})$ the initial state distribution[1]
  - $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ environment transition probability function
  - $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathbb{R})$ reward function
  - $\gamma \in [0, 1]$ unnormalized discount factor
- The MDP follows
  $a_t \sim \pi(a|s_t), r_t \sim \mathcal{R}(s_t, a_t), s_{t+1} \sim \mathcal{T}(s_t, a_t), t = 0, 1, 2, \ldots$
- The objective is to learn the policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$
- To maximize the expected return

$$R_T = \sum_{0 \leq t \leq T} \gamma^t r_t, \quad J = \mathbb{E}[R_\infty] = \mathbb{E}\Big[\sum_{t \geq 0} \gamma^t r_t \Big| s_0 \sim \rho_0, \pi\Big]$$

[1] $\Delta(\cdot)$ denotes the set of all random variables over the input space

Markov decision processes - Learning algorithms

- Action-value function $Q(s,a)$: $\mathbb{E}[R_\infty]$ condition on initial s, a

$$Q(s,a) = \mathbb{E}[R_\infty] = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r_t \Big| s_0 = s, a_0 = a, \pi\Big]$$

$$v(s) = \mathbb{E}_a[Q(s,a)] = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r_t \Big| s_0 = s, \pi\Big]$$

- Can be learned temporal-difference (TD) methods
    - TD(0) by Monte-Carlo sampling
    - TD(1) by Bellman recursive property
- Alternatively, learning by policy gradient

$$\nabla_\pi \mathbb{E}[R_\infty|\pi] = \mathbb{E}_{\pi(a|s)}[\nabla_\pi \log \pi(a|s) Q(s,a)] \tag{1}$$

Outline

The Gambler's problem

The Gambler's problem is an early example in the RL textbook by Sutton and Barto [SB18, SB98]

- The gambler starts with $s \leq 1$ capital (state)
- At each round bets a, $0 < a \leq s$ (action) and

  receives $\begin{cases} 0 & \text{with probability constant } p > 0.5, \\ 2a & \text{with probability } 1 - p. \end{cases}$

- Target capital is $1.$ Game terminates at $s = 1$ or $s = 0$

What is the probability of reaching the target, under the best a (the optimal state-value function $v(s)$)?

The Gambler's problem

Some additional notes on the problem

1. The problem looks very simple (but deceptively!). It's in fact the most simple RL setting in the book apart from bandits.
2. The original problem starts with capital n, bets only integers, and targets capital N. We solve both the original and the continuous versions.
3. Numerically estimated in the book by the value iteration algorithm. Strange patterns have been observed.

The Gambler's problem

- Recall MDP formulation - tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho_0, \gamma)$
  - $\mathcal{S} = [0, 1]$ state space, $\mathcal{A} = (0, \min(s, 1-s)]$ action space, $\rho_0 \in \Delta(\mathcal{S})$ an arbitrary initial state distribution, $\gamma \in [0, 1]$ arbitrary
  - $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, $\mathcal{T}(s, a)$ is $s - a$ and $s + a$ w.p. $p > 0.5$ and $1 - p$, respectively
  - $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, $\mathcal{R}(1, \cdot) = 1$.
- The MDP follows
  $a_t \sim \pi(a|s_t), r_t \sim \mathcal{R}(s_t, a_t), s_{t+1} \sim \mathcal{T}(s_t, a_t), t = 0, 1, 2, \ldots$
- The MDP terminates when $s \in \{0, 1\}$

The Gambler's problem

Some additional notes on the MDP

1. The MDP is stationary: Termination only on terminate states. Optimal policy/value does not need to depend on t. The Bellman equation is stringently satisfied
2. The MDP is stationary. Fewer results apply to the continuous settings and some known properties do not extend to continuous MDPs
3. At least one deterministic policy is optimal in MDPs so we can wlog restrict π to be deterministic.

Outline

The optimal value function

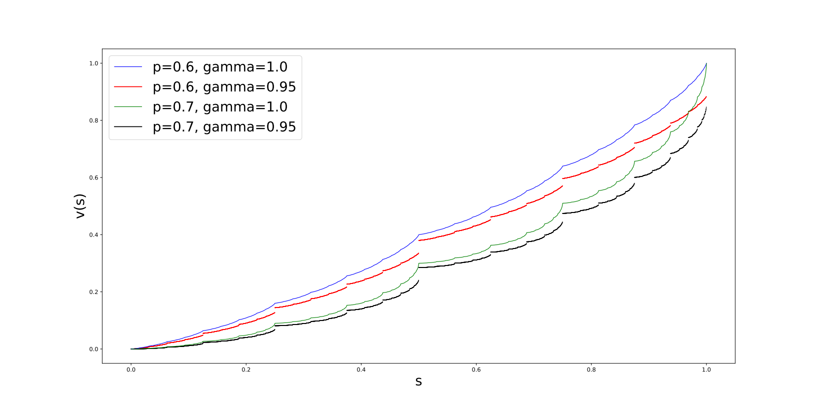
Theorem 12. $v(s) = \sum_{i=1}^{\infty} (1-p)\gamma^i b_i \prod_{j=1}^{i-1}((1-p) + (2p-1)b_j)$ is the optimal state-value function for any $0 \leq \gamma \leq 1$ and $p > 0.5$, where $s = 0.b_1 b_2 \ldots b_\ell \ldots_{(2)}$ is the binary representation of the state $0 \leq s < 1$.
(p: probability of losing a bet. γ: constant discount factor)

- The answer is surprisingly complicated despite the problem being simple
- Describing $v(s)$ using elementary functions is not possible

Plots and characterizations

x-axis: Initial capital (state); y-axis: Probability of winning (value function)

Characterizations: Fractal; self-similar; derivative is either zero or infinity; not written as elementary functions



The solution of optimal value function

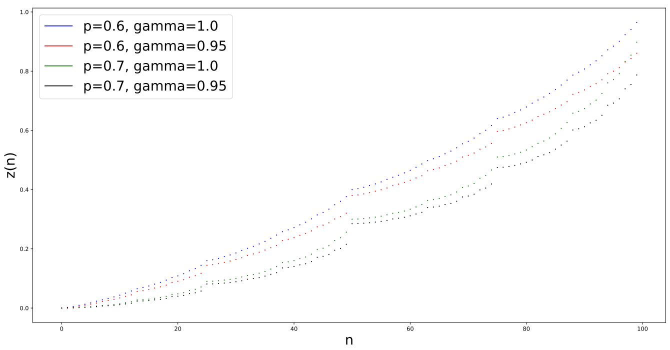
Proposition 1. The optimal value function $z(n)$ is $v(n/N)$ in the discrete setting of the Gambler's problem, where $v(\cdot)$ is the optimal value function under the continuous case defined in Theorem 12.

Corollary 13. The policy $\pi(s) = \min(s, 1-s)$ is (Blackwell) optimal in both the discrete and the continuous cases.

Discrete plots

Discrete problem value function is exactly the continuous problem value function evaluated at discrete points.

This is the "strange pattern" in Sutton and Barto's book.



Bellman equation TD(0)

The Bellman equation of the Gambler's problem is $f(0) = 0$, $f(1) = 1$,

$$f(s) = \max_{0<a\leq\min\{s,1-s\}} (1-p)\gamma f(s+a) + p\gamma f(s-a)$$

for some real function $f : [0,1] \to \mathbb{R}$.

Theorem 22. Let $\gamma = 1$, $p > 0.5$. $f(s)$ solves the Bellman equation if and only if either

- $f(s)$ is $v(s)$ defined in Theorem 12, or
- $f(0) = 0$, $f(1) = 1$, and $f(s) = C$ for all $0 < s < 1$, for some constant $C \geq 1$.

In a difficult case, this problem explores the most fundamental arguments in probabilities and math - the belief of axioms.

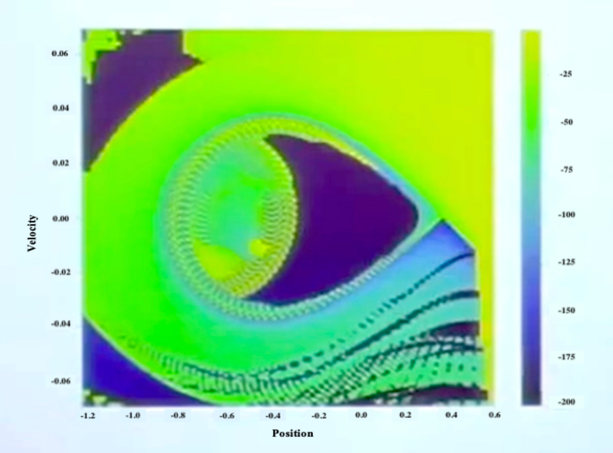
Theorem 27. Let $\gamma = 1$ and $p = 0.5$. A real function $f(s)$ satisfies the Bellman equation if and only if either

- $f(s) = C's + B'$ on $s \in (0, 1)$, for some constants $C' + B' \geq 1$, or
- $f(s)$ is some non-constructive, not Lebesgue measurable function under Axiom of Choice.

Outline

Implications (1) - Generalization

- Similar observations of chaos in other RL problems (e.g. Mountain Car, as below)
- Results and characterizations apply to RL in general



Implications (2) - Fractal and self-similarity

2. The value function is non-smooth on any interval
    - Modern deep reinforcement learning (incorrectly) assume the value function to be smooth to use neural networks.
    - **Proposition 19 and 20.** Using N-bin discretization incurs at least $O(1/N)$ approximation error. Using L-Lipschitz function has at least $O(1/L)$ error.
    - Revisit state and value representation

The state-of-the-art algorithm, soft actor-critic [HZAL18, HTAL17], learns a smooth surrogate instead of the optimal function. It achieves the state of the art by unintentionally avoiding the optimality.

Implications (3) - Singularity

3. Singularity means a function's derivation takes either zero or infinity, on its entire interval $(0, 1)$.
   - Remark: The curve still goes from $(0, 0)$ to $(1, 1)$, counter-intuitively
   - Algorithmically this denies the access to $\partial v(s)/\partial s$ and $\partial Q(s, a)/\partial a$ [LHP+15, GLT+17, HWS+15, FA12, Fai08, PYFW19, LJL+18], including famous DDPG and Dyna

Their are many more algorithms than what I can enumerate. The code will always return a *gradient* when called but it will depend on the discrete gradient rather than what the algorithm expect.

Implications (4) - Q-learning

4. The Q-learning algorithm minimizes the Bellman equation. We do not know which point it will converge to.

Optimization and approximation algorithms might prefer a large constant function than the desired optimal value function.

In fact, original Q-learning rarely works in continuous spaces and people did not know why. DeepMind made it work by combination of tricks while biasing the objective.

Implicated future works

- Long-term research goal of the line: To understand the sequential decision problem.
- Foundations will help us characterize and understand the problem itself instead of the methods, which then drives better algorithm designs.
- Implied future works
    - Improving state and value function approximation, as now that we know why previous methods suffer from errors;
    - Improving Q-learning's convergence, as we know why it did not behave as desired.

Outline

Dynamical systems

Let $f : [0,1] \to \mathbb{R}$ be a real function. For $f(s)$ to be the optimal value function, the Bellman equation for the non-terminal and terminal states are ($\mathcal{A}(s) = (0, \min\{s, 1-s\}]$, $s \in (0,1)$)

$$f(s) = \max_{a \in \mathcal{A}(s)} p\gamma\, f(s-a) + (1-p)\gamma\, f(s+a) \text{ for any } s \in (0,1), \quad \text{(A)}$$

and

$$f(0) = 0, \quad f(1) = 1. \quad \text{(B)}$$

Dynamical systems

The bounded version of the problem leads to the optimal value function.

$$0 \leq \gamma \leq 1,\ p > 0.5,\ f(s) \leq 1 \text{ for all } s,\ f(s) \text{ is continuous on } s = 0. \tag{X}$$

The unbounded version of the problem leads to the solutions of the Bellman equation.

$$0 \leq \gamma \leq 1,\ p > 0.5. \tag{Y}$$

The corner case of $\gamma = 1$, $p = 0.5$ is difficult and exceptional

$$\gamma = 1,\ p = 0.5,\ f(s) \text{ is unbounded.} \tag{Z}$$

Outline

The optimal value function

Recall that

Theorem 12. $v(1) = 1$ and

$$v(s) = \sum_{i=1}^{\infty} (1-p)\gamma^i b_i \prod_{j=1}^{i-1} ((1-p) + (2p-1)b_j)$$

for any $0 \leq \gamma < 1$ is the optimal state-value function, where $s = 0.b_1 b_2 \ldots b_\ell \ldots_{(2)}$ is the binary representation of the state $0 \leq s < 1$.
($p > 0.5$: probability of losing a bet. γ: constant discount factor)

Theorem (lemma 3, Monotonicity)

Let $\gamma = 1$ *and* $p > 0.5$*. If a real function* $f(s)$ *satisfies (AB) then* $f(s)$ *is monotonically increasing on* $[0, 1)$*.*

Proof sketch.

If otherwise there exists $s_1 < s_2$ and $f(s_1) > f(s_2)$, then by induction we obtain

$$f(s_2 + k2^{-\log(k)}\Delta s) - f(s_2) \leq -kp^{\log(k)}\Delta f.$$

By the arbitrarity of k this indicates the non-existence of $f(s_2 + k2^{-\ell}\Delta s)$. $\square$

The optimal value function (ABX)

Theorem (lemma 4, Continuity)

Let $\gamma = 1$ *and* $p \geq 0.5$*. If a real function* $f(s)$ *is monotonically increasing on* $(0, 1]$ *and it satisfies (AB), then* $f(s)$ *is continuous on* $(0, 1]$*.*

Proof sketch.

- Otherwise we construct a series of points $s_1, s_2, \ldots$ around the discontinuity.
- Repeatedly applying condition (A) shows that the series $f(s_1), f(s_2), \ldots$ is unbounded, which contradicts with the monotonicity. □

Theorem (Lemma 2, Uniqueness under existence)

Let $f(s) : [0, 1] \to \mathbb{R}$ be a real function. If $v(s)$ and $f(s)$ both satisfy (ABX), then $v(s) = f(s)$ for all $0 \leq s \leq 1$.

Proof sketch.

- Find a point s_0 that maximizes $v(s_0) - f(s_0)$ then derive contradiction under $v(s_0) - f(s_0) > 0$.
- Show the existence of s_0 via the continuity of $v(s)$ and $f(s)$. $\square$

Theorem (lemma 11, Feasibility of $v(s)$)

$v(s)$ *is a solution of the system (ABX).*

Proof sketch.

Let $v'(s) = \max_{a \in \mathcal{A}(s)} p\gamma\ v(s-a) + (1-p)\gamma\ v(s+a)$.

- $v(s) = v'(s)$ on the dyadic rationals $\bigcup_{\ell \geq 1} G_\ell$.
- $v(s)$ and $v'(s)$ are continuous for any s if there does not exist an $\ell \geq 1$ such that $s \in G_\ell$.
- Since $\bigcup_{\ell \geq 1} G_\ell$ is dense and compact on $(0,1)$, $v(s) = v'(s)$ holds whenever both $v(s)$ and $v'(s)$ are continuous at s.
- Thus $v(s) = v'(s)$ on the complement of $\bigcup_{\ell \geq 1} G_\ell$. $\square$

The optimal value function (ABX)

Theorem (Theorem 12, The optimal value function)

Let $0 \leq \gamma \leq 1$ *and* $p > 0.5$*. Under the continuous setting of the Gambler's problem, the optimal state-value function is* $v(1) = 1$ *and* $v(s) = \sum_{i=1}^{\infty}(1-p)\gamma^i b_i \prod_{j=1}^{i-1}((1-p)+(2p-1)b_j)$ *for* $0 \leq s < 1$*.*

Proof.

- The optimal value function solves (ABX).
- $v(s)$ solves (ABX).
- There can be only one function who solves (ABX). □

Thank you

Baoxiang Wang (bxiangwang@gmail.com)

School of Data Science, The Chinese University of Hong Kong, Shenzhen

August 12, 2020. CUHKSZ.

[FA12] Michael Fairbank and Eduardo Alonso.
Value-gradient learning.
In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2012.

[Fai08] Michael Fairbank.
Reinforcement learning by value gradients.
arXiv preprint arXiv:0803.3539, 2008.

[GLT+17] Shixiang Shane Gu, Timothy Lillicrap, Richard E Turner, Zoubin Ghahramani, Bernhard Schölkopf, and Sergey Levine.
Interpolated policy gradient: Merging on-policy and off-policy gradient estimation for deep reinforcement learning.
In *Advances in neural information processing systems*, pages 3846–3855, 2017.

[HTAL17] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine.
Reinforcement learning with deep energy-based policies.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361. JMLR. org, 2017.

[HWS+15] Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa.
Learning continuous control policies by stochastic value gradients.

In *Advances in Neural Information Processing Systems*, pages 2944–2952, 2015.

[HZAL18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine.
Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.
arXiv preprint arXiv:1801.01290, 2018.

[LHP+15] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra.
Continuous control with deep reinforcement learning.
arXiv preprint arXiv:1509.02971, 2015.

[LJL+18] Sungsu Lim, Ajin Joseph, Lei Le, Yangchen Pan, and Martha White.
Actor-expert: A framework for using action-value methods in continuous action spaces.
arXiv preprint arXiv:1810.09103, 2018.

[PYFW19] Yangchen Pan, Hengshuai Yao, Amir-massoud Farahmand, and Martha White.
Hill climbing on value estimates for search-control in dyna.
arXiv preprint arXiv:1906.07791, 2019.

[SB98] Richard S Sutton and Andrew G Barto.
Reinforcement learning: An introduction.
MIT press Cambridge, 1998.

[SB18] Richard S Sutton and Andrew G Barto.
Reinforcement learning: An introduction.
MIT press, 2018.