Posterior sampling and Bayesian bootstrap: sample complexity and regret bounds

Alexey Naumov, Daniil Tiapkin

HDI Lab

HSE University

November 1, 2022

1. Lecture 1: Introduction to stochastic multi-armed bandits
   - regret
   - Exploration-exploitation dilemma
   - Explore-First Algorithm
   - Optimism in the Face of Uncertainty

2. MDP Basics. Policy Evaluation
   - Markov Decision Process
   - Policy's quality
   - Value iteration
   - TD learning
   - Stochastic approximation

3. Learning a (near-) optimal policy
   - SARSA
   - Q-learning

4. From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses

References

- A book for reading on the plane, on the beach, on the train etc:

  Richard Sutton and Andrew Barto, Reinforcement Learning: An Introduction. Second edition. MIT (Sutton and Barto [2018]);
- For strong people:

  Csaba Szepesvári, Algorithms for reinforcement learning (Szepesvári [2010]);
- For very strong people:

  Martin Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley (Puterman [2014]);

  Vivek Borkar, Stochastic Approximation. A Dynamical Systems Viewpoint, Cambridge University Press, 2008 (Borkar [2008]).

Lecture 1: Introduction to stochastic multi-armed bandits

- Stochastic Multi-armed bandits: basic concepts.
- Concept of regret bounds.
- Exploration-exploitation trade-off.
- Recap: Hoeffding inequality.
- Exploration first algorithm.
- Optimism in face of uncertainty: Upper confidence bound algorithm (UCB).

Stochastic Multi-Armed Bandit Problem

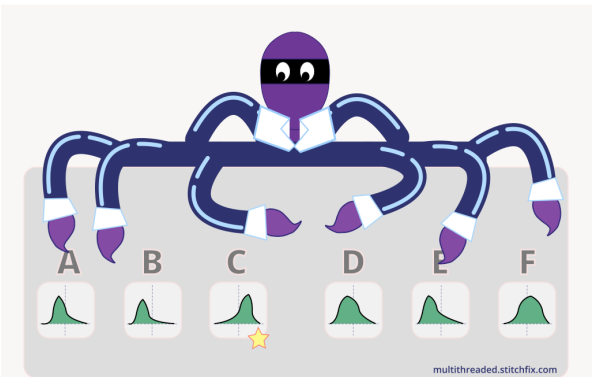
Stochastic Multi-Armed Bandit Problem

Given K possible actions $\mathcal{A}$ (a.k.a. *arms*), each arm a has its underlying distribution of rewards $\mathcal{D}_a$. The goal of the algorithm (a.k.a. *agent*) is to find an arm a that maximizes expectation of an observed reward $\mu(a) = \mathbb{E}[\mathcal{D}_a]$ during T rounds of interaction.

In each round $t \in [T]$:

- Agent picks arm $a_t \in \mathcal{A}$;
- Agent receives reward $r_t \sim \mathcal{D}_{a_t}$ for a chosen arm a_t.

All rewards generated by a single arm assumed to be independent and identically distributed (IID). For simplicity we assume bounded reward $r_t \in [0, 1]$.



Regret

Set of notations:

- The mean reward is $\mu(a) := \mathbb{E}[\mathcal{D}_a]$;
- The best reward is $\mu^\star := \max_{a \in \mathcal{A}} \mu(a)$;
- The difference $\Delta(a) := \mu^\star - \mu(a)$ describes how worse the arm a compared to $\mu^\star$; we call $\Delta(a)$ as a *gap* of arm a;
- An optimal arm $a^\star$ is an arm with $\mu(a^\star) = \mu^\star$ or, equivalently, $\Delta(a^\star) = 0$. It may not be unique!

We define a performance measure as a *cumulative regret* (or just regret) at round T[1]

$$\mathfrak{R}^T := \sum_{t=1}^{T} \mu^\star - \mu(a_t) = T\mu^\star - \sum_{t=1}^{T} \mu(a_t).$$

[1] In the literature this quantity is often called *pseudo-regret*.

Exploration-exploitation dilemma

On each round t there is a choice: we need to search an information on rarely used arms (exploration) or just act according to the arm with best estimated mean (exploitation).

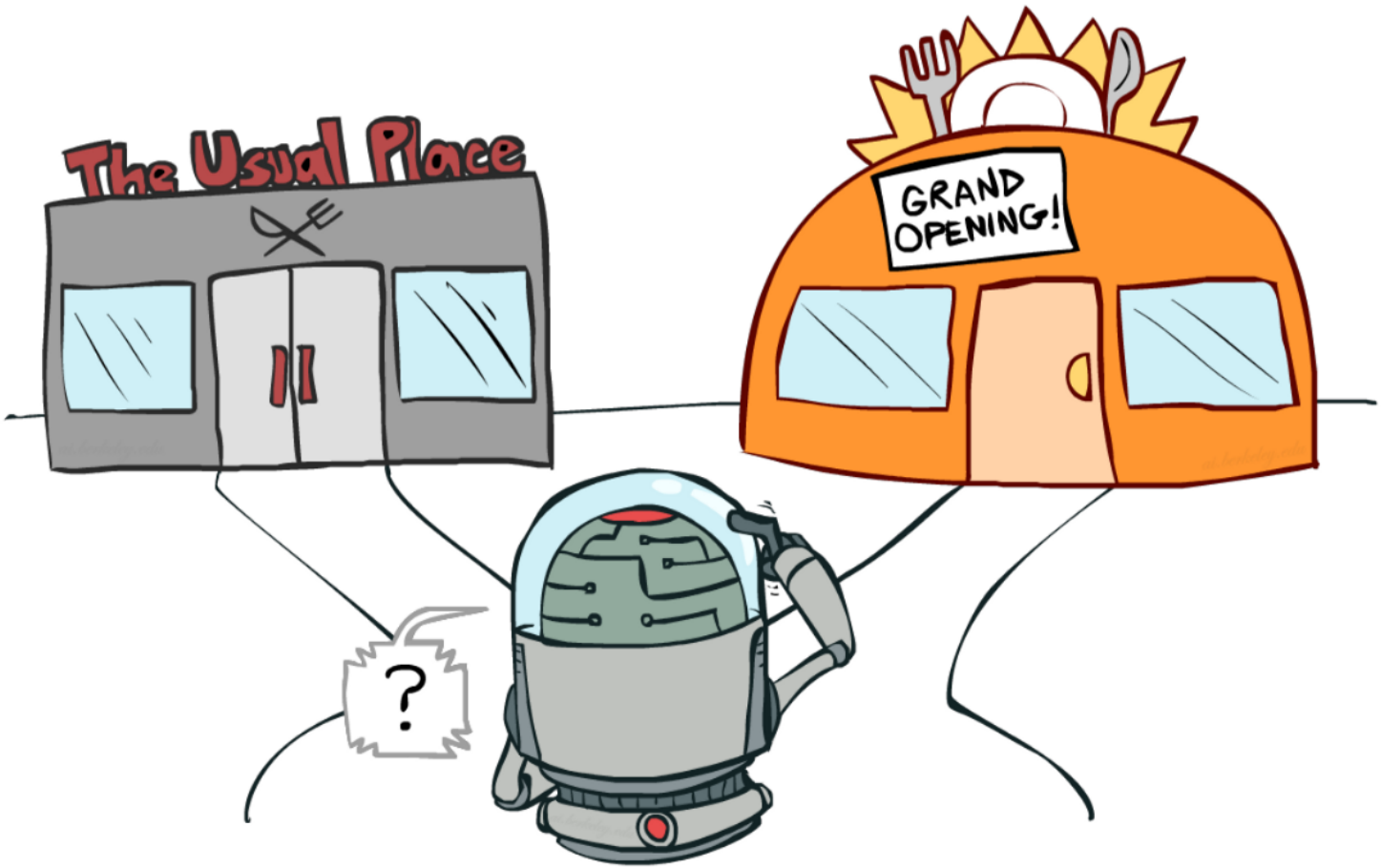


Figure: Image source: UC Berkeley Intro to AI course

Explore-First Algorithm

- Exploration phase: Try each arm N times;
- Select arm $\hat{a}$ with the highest average reward;
- Exploitation phase: play arm $\hat{a}$ in all remaining $T - NK$ rounds.

Theorem

For $N = \mathcal{O}(T\sqrt{\log(T)}/K)^{2/3}$ Explore-First Algorithm achieves

$$\mathbb{E}[\mathfrak{R}^T] \leq \mathcal{O}\left(T^{2/3}(K\log(T))^{1/3}\right).$$

Proof

Theorem (Hoeffding bound)

Let $X_1, \ldots, X_n$ be a sequence of IID random variables supported in $[a, b]$. Then

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - n\,\mathbb{E}[X_1]\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{n(b-a)^2}\right)$$

ML friendly: For any $\delta \in (0, 1)$ with probability at least $1 - \delta$ the following bound holds

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right| \leq (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}$$

See Vershynin [2019] for intro to concentration of measure.

Proof (1)

- Define an average reward after exploration phase for action a as $\widehat{\mu}(a)$. Let us define so-called *clean event* $\mathcal{E}$ as follows

$$\mathcal{E} = \{\forall a \in \mathcal{A} : |\widehat{\mu}(a) - \mu(a)| \leq \beta\}$$

  for $\beta = \sqrt{\log(T^4)/(2N)} = \sqrt{2\log(T)/N}$.
- By the Hoeffding inequality for any fixed a

$$\mathbb{P}[|\widehat{\mu}(a) - \mu(a)| > \beta] \leq \frac{2}{T^4}.$$

- Note that $\overline{\mathcal{E}} = \bigcup_{a \in \mathcal{A}} \{|\widehat{\mu}(a) - \mu(a)| > \beta\}$, therefore we may apply union bound

$$\mathbb{P}[\overline{\mathcal{E}}] \leq \sum_{a \in \mathcal{A}} \mathbb{P}[|\widehat{\mu}(a) - \mu(a)| > \beta] \leq \frac{2K}{T^4}. \tag{1}$$

Proof (2)

- Define $\hat{a} = \arg\max_{a \in \mathcal{A}} \widehat{\mu}(a)$. Assume that $\hat{a} \neq a^\star$. In this case we have that under the clean event $\mathcal{E}$

$$\mu(\hat{a}) + \beta \geq \widehat{\mu}(\hat{a}) \geq \widehat{\mu}(a^\star) \geq \mu^\star - \beta.$$

- Therefore

$$\Delta(\hat{a}) = \mu^\star - \mu(\hat{a}) \leq 2\beta. \tag{2}$$

- Let us derive a regret bound

$$\mathbb{E}[\mathfrak{R}^T] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta(a_t)\right] = \underbrace{\mathbb{E}\left[\sum_{t=1}^{NK} \Delta(a_t)\right]}_{\text{exploration phase}} + \underbrace{\mathbb{E}\left[\sum_{t=NK+1}^{T} \Delta(\hat{a})\right]}_{\text{exploitation phase}}.$$

Proof (3)

- In the first phase we have only a trivial regret bound NK. For the second phase we divide our expectation into two parts: with and without clean event.

$$\mathbb{E}\left[\sum_{t=NK+1}^{T}\Delta(\hat{a})\right] = \underbrace{\mathbb{E}\left[\sum_{t=NK+1}^{T}\Delta(\hat{a}) \mid \mathcal{E}\right]}_{\leq 2T\beta}\underbrace{\mathbb{P}[\mathcal{E}]}_{\leq 1} + \underbrace{\mathbb{E}\left[\sum_{t=NK+1}^{T}\Delta(\hat{a}) \mid \overline{\mathcal{E}}\right]}_{\leq T}\underbrace{\mathbb{P}[\overline{\mathcal{E}}]}_{\leq 2K/T^4}.$$

- Therefore, we have

$$\mathbb{E}[\mathfrak{R}^T] \leq NK + 2T\sqrt{\frac{2\log(T)}{N}} + \frac{2K}{T^3}.$$

Let us optimize the upper bound over N. The optimal value is $N^\star = (T\sqrt{\log(2T)/K})^{2/3}$ and in this case we derive claimed regret bound (assuming that $K \leq T$)

$$\mathbb{E}[\mathfrak{R}^T] \leq 3T^{2/3}(K\log(2T))^{1/3} + T^{-2} = \mathcal{O}(T^{2/3}(K\log(T))^{1/3}).$$

Optimism in the Face of Uncertainty (OFU)

Define $\overline{\mu}_t(a)$ as a so-called *upper-confidence bound* for arm a that means that with high probability we have $\overline{\mu}_t(a) \geq \mu(a)$. Hoeffding inequality tells us that this upper confidence bound could be defined in the form

$$\overline{\mu}_t(a) = \widehat{\mu}_t(a) + \underbrace{\beta_t(a)}_{\textit{Exploration bonus}} \triangleq \frac{1}{n_t(a)} \sum_{t:a_t=a} r_t + \sqrt{\frac{2\log(T)}{n_t(a)}},$$

where $n_t(a)$ is a number of times when the arm a was picked up to a timestamp t.
Define algorithm UCB-1 as follows

- In each round t pick $a_t = \arg\max_{a \in \mathcal{A}} \overline{\mu}_t(a)$.

Why does it make sense? There is basically two reasons to choose arm a on the round t:

- Arm a has a high mean reward $\widehat{\mu}_t(a)$ that means that it is likely to have a high mean reward $\mu(a)$;
- Arm a has a large confidence interval $\beta_t(a)$ that means that this arms is not explored properly.

Optimism in the Face of Uncertainty (OFU)

Theorem

Algorithm UCB-1 achieves $\mathbb{E}[\mathfrak{R}^T] = \widetilde{\mathcal{O}}(\sqrt{KT})$, *where* $\widetilde{\mathcal{O}}(f(x))$ *is an upper bound on* $f(x)$ *up to constant and poly-logarithmic factors for sufficiently large* x.

Proof (1)

- Let us define an *optimistic event*

$$\mathcal{E}_{\text{opt}} = \{\forall t \in [T], \forall a \in \mathcal{A} : |\widehat{\mu}_t(a) - \mu(a)| \leq \beta_t(a)\}.$$

- Unfortunately, it is rather hard to get guarantees for this event directly due to the random number of arm pulls inside the definition of $\widehat{\mu}_t(a)$.
- To overcome this issue, let us imagine *a reward tape*: an $1 \times T$ table filled with IID sampled reward from $\mathcal{D}_a$. Then for j-th choice of arm a we will think not as about a new sample from $\mathcal{D}_a$ but as about a selecting j-th element on this tape. Let us call $v_j(a)$ as a mean reward over first j elements of this tape.
- It is clear that

$$\mathcal{E} \triangleq \left\{\forall j \in [T], \forall a \in \mathcal{A} : |v_j(a) - \mu(a)| \leq \sqrt{\frac{\log(2T)}{j}}\right\} \subseteq \mathcal{E}_{\text{opt}}.$$

Proof (2)

For each separate j, a we have

$$\mathbb{P}\left[|v_j(a) - \mu(a)| > \sqrt{\frac{2\log(T)}{j}}\right] \leq \frac{2}{T^4}$$

thus, by union the bound argument (similar as in Explore-First algorithm) and assuming that $K \leq T$

$$\mathbb{P}[\mathcal{E}_{\text{opt}}] \geq \mathbb{P}[\mathcal{E}] \geq 1 - \frac{2K}{T^3} \geq 1 - \frac{2}{T^2}.$$

Proof (3)

Decompose regret depending on the event $\mathcal{E}_{\text{opt}}$

$$\mathbb{E}[\mathfrak{R}^T] = \mathbb{E}[\mathfrak{R}^T|\mathcal{E}_{\text{opt}}]\,\mathbb{P}[\mathcal{E}_{\text{opt}}] + \mathbb{E}\left[\mathfrak{R}^T|\overline{\mathcal{E}_{\text{opt}}}\right]\mathbb{P}\left[\overline{\mathcal{E}_{\text{opt}}}\right] \leq \mathbb{E}[\mathfrak{R}^T|\mathcal{E}_{\text{opt}}] + 2\,T^{-1}.$$

Thus, again it is sufficient to analyze the regret only under $\mathcal{E}_{\text{opt}}$. Let us start to provide bound on $\Delta(a_t)$. In this case we have

$$\mu^\star \underbrace{\leq}_{\text{since } \overline{\mu}_t \text{ is UCB}} \overline{\mu}_t(a^\star) \underbrace{\leq}_{\text{by greedy choice}} \overline{\mu}_t(a_t) \underbrace{\leq}_{\text{by overestimation of UCB}} \mu(a_t) + 2\beta_t(a_t).$$

Thus, we have

$$\Delta(a_t) = \mu^\star - \mu(a_t) \leq 2\sqrt{\frac{2\log(T)}{n_t(a)}}.$$

Proof (4)

- On the event $\mathcal{E}_{\text{opt}}$:

$$\mathfrak{R}^T = \widetilde{\mathcal{O}}\left(\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(a_t)}}\right) = \widetilde{\mathcal{O}}\left(\sum_{a\in\mathcal{A}} \sum_{k=1}^{n_T(a)} \frac{1}{\sqrt{k}}\right).$$

  where we used $\sum_{a\in\mathcal{A}} n_T(a) = T$.
- By the integral bound we have

$$\sum_{k=1}^{n_T(a)} \frac{1}{\sqrt{k}} \leq \int_1^{n_T(a)} \frac{1}{\sqrt{x}} \mathrm{d}x \leq 2\sqrt{n_T(a)}.$$

- Note that $f(x) = \sqrt{x}$ is concave function for positive x thus by Jensen's inequality

$$\sum_{a\in\mathcal{A}} \sqrt{n_T(a)} = K \sum_{a\in\mathcal{A}} \frac{\sqrt{n_T(a)}}{K} \leq K\sqrt{\sum_{a\in\mathcal{A}} \frac{n_T(a)}{K}} = \sqrt{TK}.$$

- Combining all estimates

$$\mathbb{E}[\mathfrak{R}^T] \leq \mathbb{E}[\mathfrak{R}^T|\mathcal{E}_{\text{opt}}] + 2T^{-1} = \widetilde{\mathcal{O}}\left(\sum_{t=1}^{T} \frac{1}{\sqrt{n_t(a_t)}}\right) = \widetilde{\mathcal{O}}(\sqrt{TK}).$$

Lecture 2: MDP Basics. Policy Evaluation

- Markov Decision process formalism.
- Definitions of value- and action-value functions, optimal policy and optimal value functions.
- Bellman optimality and Bellman expectation equations, existence of the optimal value function.
- Finite and infinite horizon.
- Policy and value iteration algorithms.
- Policy evaluation problem: $TD(0)$ and Monte-Carlo algorithms.

Markov Decision Process

We start from the case when H is finite (in this case $\gamma = 1$).

MDP

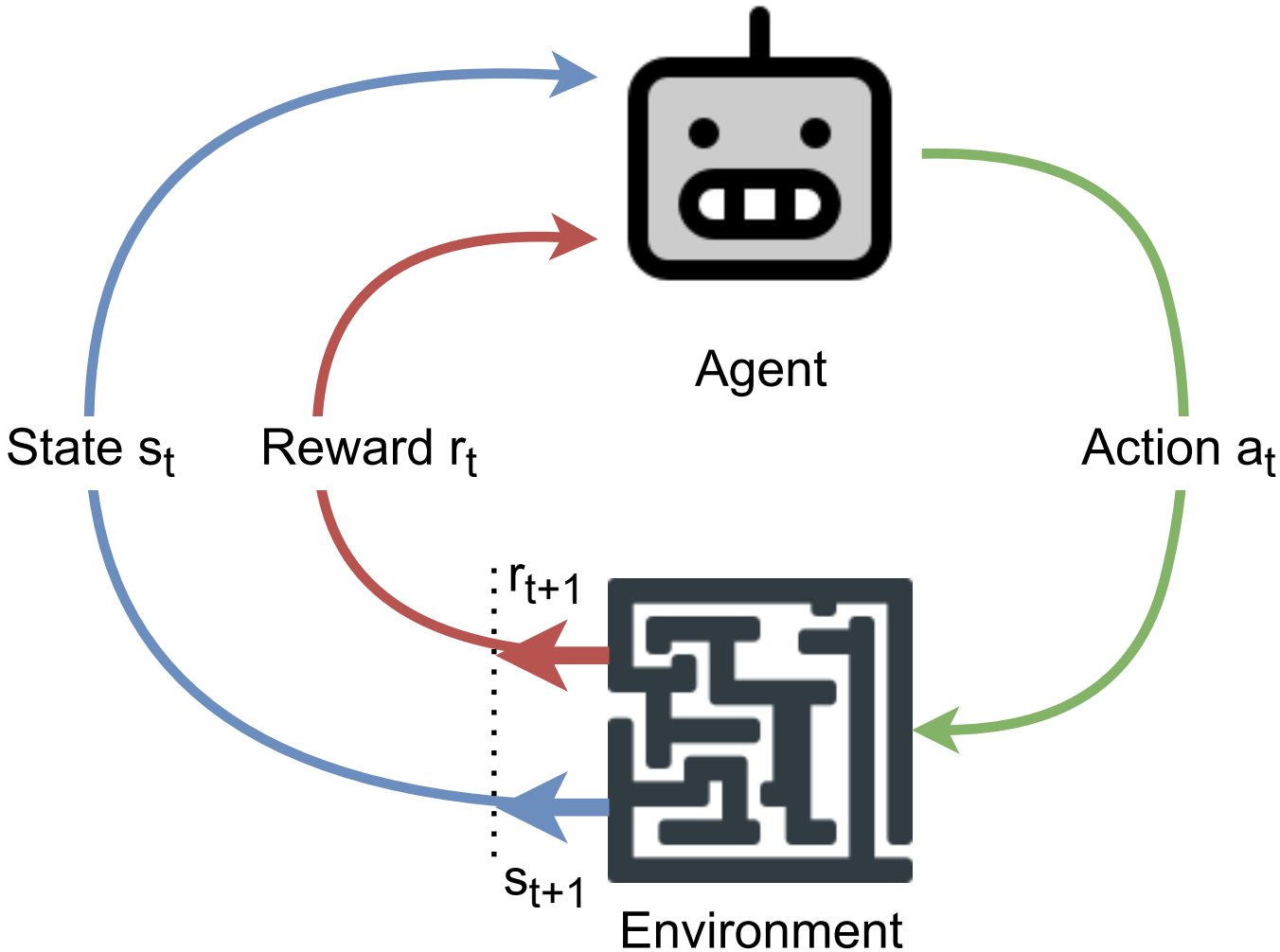
Tuple $(\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \mathrm{H})$ is called Markov Decision Process:

- S - state space. By $(S_k)_{k \geq 0}$ we denote a sequence of random states.
- A - action space. Let $(A_k)_{k \geq 0}$ be a sequence of random actions.
- Agent's policy $\pi_h(\cdot|s)$ is the distribution on A.
- Family of Markov transition kernels $(\mathrm{P}_h(s'|s,a))_{a \in \mathrm{A}}$:

$$\mathrm{P}_h(s'|s,a) := \mathbb{P}(S_{h+1} = s' | S_h = s, A_h = a)\,.$$

  (*For simplicity we assume that* P *doesn't depend on* h).
- *the reward distribution* $\mathrm{R}(\cdot|s,a)$ as a set of measures over $\mathbb{R}$ for any $(s,a) \in \mathrm{S} \times \mathrm{A}$ and *the immediate reward function* $r(s,a) = \mathbb{E}[\mathrm{R}(s,a)]$. The role of immediate reward function is similar to mean reward in bandits. (*For simplicity we assume that* $r(s,a)$ *is bounded in* $[0,1]$).

MDP



Markov Decision Process (MDP)

- Note that $(S_k)_{k\geq 0}$ is a Markov chain (MC) with Markov kernels

$$\mathrm{P}_h^\pi(s'|s) = \sum_{a\in \mathrm{A}} \mathrm{P}(s'|s,a)\pi_h(a|s)$$

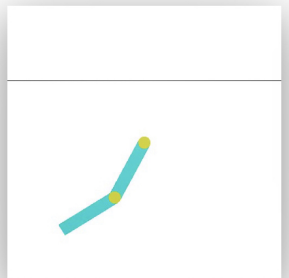
- Path distribution: for some $T \in \mathbb{N}$

$$\begin{aligned}\mathbb{P}(A_0 = a_0, S_1 = s_1, \ldots, , S_T = s_T, A_T = a_T | S_0 = s_0) \\ = \pi_0(a_0|s_0) \prod_{k=1}^{T} \mathrm{P}(s_k|s_{k-1}, a_{k-1})\pi_k(a_k|s_k)\end{aligned}$$

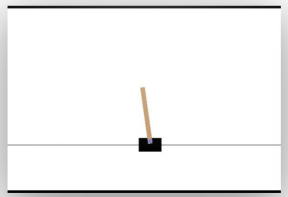
OpenAI Gym: classical control (https://gym.openai.com)

Classic control

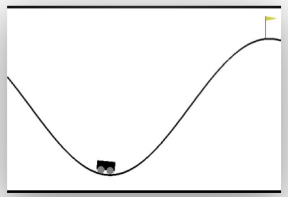
Control theory problems from the classic RL literature.



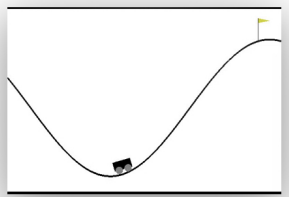
Acrobot-v1
Swing up a two-link robot.



CartPole-v1
Balance a pole on a cart.



MountainCar-v0
Drive up a big hill.



MountainCarContinuous-v0
Drive up a big hill with continuous control.

Pendulum-v0
Swing up a pendulum.

OpenAI

How to measure policy's quality?

Value Function

Value function, associated with the policy $\pi = (\pi_1, \ldots, \pi_H)$, is defined as

$$V_h^{\pi}(s) := \mathbb{E}\left[\sum_{k=h}^{H} r_k | S_h = s\right]$$

Here for all $k \geq h$ $r_k \sim \mathrm{R}(\cdot|s_k, a_k), S_{k+1} \sim \mathrm{P}_k(\cdot|s_k, a_k), A_k \sim \pi_k(\cdot|s_k)$.

Optimal value function

The *optimal value function* at step h and state $s \in \mathrm{S}$

$$V_h^{\star}(s) = \sup_{\pi} V_h^{\pi}(s).$$

Action-value function

Action-value function

The action-value function $Q_h^\pi : \mathrm{S} \to \mathbb{R}$ is an expectation of return of agent then it start at step h, state s_h and selects a prescribed action a_h. In other words, for all $h \geq 1$ and $s \in \mathrm{S}$ and $a \in \mathrm{A}$

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H} r_t \mid S_h = s, A_h = a\right],$$

where for all $t \geq h$ $r_t \sim \mathrm{R}(\cdot|s_t, a_t), S_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t), A_{t+1} \sim \pi_t(\cdot|s_{t+1})$

Optimal action-value function

The *optimal action-value function* at step h and state $s \in \mathrm{S}$

$$Q_h^\star(s, a) = \sup_\pi Q_h^\pi(s, a).$$

MC notations

For a kernel $\mathrm{P}(\cdot|s,a)$ we may define its action on any (measurable) function $f : \mathrm{S} \to \mathbb{R}$ as follows

$$\mathrm{P}\, f(s,a) = \int_{\mathrm{S}} f(s')\, \mathrm{P}(\mathrm{d}s'|s,a).$$

In the case of finite MDP this formula simplifies to

$$\mathrm{P}\, f(s,a) = \sum_{s' \in \mathrm{S}} f(s') \cdot \mathrm{P}(s'|s,a).$$

In the case then distribution $\mathrm{P}(\cdot|s,a)$ has a density $p(s'|s,a)$ we have

$$\mathrm{P}\, f(s,a) = \int_{\mathrm{S}} f(s') p(s'|s,a) \mathrm{d}s'.$$

Bellman equations

Theorem (Bellman equations)

Fix a finite-horizon MDP $M = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \mathrm{H})$ *and policy* π*. Let* r *be the immediate reward function of* M*. Then* V_h^π *and* Q_h^π *satisfy Bellman equations*

$$Q_h^\pi(s,a) = r(s,a) + \mathrm{P}\, V_{h+1}^\pi(s,a), \qquad \forall (s,a,h) \in \mathrm{S} \times \mathrm{A} \times [\mathrm{H}]$$

$$V_h^\pi(s) = \sum_{a \in \mathrm{A}} Q_h^\pi(s,a)\pi_h(a|s) \qquad \forall (s,h) \in \mathrm{S} \times [\mathrm{H}]$$

$$V_{H+1}^\pi(s) = 0 \qquad \forall s \in \mathrm{S}$$

In the case of deterministic policies π_h Bellman equation on V_h^π could be simplified as follows

$$V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)).$$

Proof (1)

Without loss of generality assume that rewards $r_t \sim \mathrm{R}(\cdot|s_t, a_t)$ are deterministic and equal to $r(s_t, a_t)$. Then by definition and tower property of conditional expectation

$$
\begin{aligned}
Q_h^\pi(s,a) &= \mathbb{E}\left[\sum_{t=h}^{H} r(S_t, A_t) \middle| S_h = s, A_h = a\right] \\
&= r(s,a) + \mathbb{E}\left[\sum_{t=h+1}^{H} r(S_t, A_t) \middle| S_h = s, A_h = a\right] \\
&= r(s,a) + \mathbb{E}\left[\mathbb{E}\left[\sum_{t=h+1}^{H} r(S_t, A_t) \middle| S_{h+1}\right] \middle| S_h = s, A_h = a\right] \\
&= r(s,a) + \mathbb{E}\left[V_{h+1}^\pi(S_{h+1}) \middle| S_h = s, A_h = a\right] \\
&= r(s,a) + \mathsf{P}\, V_{h+1}^\pi(s,a).
\end{aligned}
$$

Proof (2)

Next we provide second Bellman equation

$$
\begin{aligned}
V_h^\pi(s,a) &= \mathbb{E}\left[\sum_{t=h}^{H} r(S_t, A_t) \middle| S_h = s\right] && \text{(tower property)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=h}^{H} r(S_t, A_t) \middle| S_h, A_h\right] \middle| S_h = s\right] && \text{(definition of } Q_h^\pi\text{)} \\
&= \mathbb{E}\left[Q_h^\pi(s, A_h) \middle| S_h = s\right] && (A_h \sim \pi_h(\cdot|s)) \\
&= \sum_{a \in \mathcal{A}} Q_h^\pi(s,a)\pi_h(a|s).
\end{aligned}
$$

Policy improvement

Theorem (Policy improvement theorem)

Let $M = (\mathsf{S}, \mathsf{A}, \mathsf{P}, \mathrm{R}, \mathsf{H})$ *be a finite-horizon MDP and* π *be a fixed policy. Define* $\hat{\pi}$ *as a discrete greedy policy to* $Q^{\pi}(s, a)$*, i.e.*

$$\hat{\pi}_h(s) := \arg\max_{a \in \mathcal{A}} Q_h^{\pi}(s, a).$$

Then for any $(s, h) \in \mathsf{S} \times [\mathsf{H}]$ *we have* $V_h^{\hat{\pi}}(s) \geq V_h^{\pi}(s)$*.*

Proof

Backward induction over $h = H + 1, \ldots, 1$.

- For $h = H = 1$ value functions of all policies are equal to zero, thus we are done.
- Step of induction: First we show that $Q_h^\pi(s,a) \leq Q_h^{\hat{\pi}}(s,a)$ for all $(s,a) \in \mathsf{S} \times \mathsf{A}$. By Bellman equations and induction hypothesis

$$Q_h^\pi(s,a) = r(s,a) + \mathsf{P}\, V_{h+1}^\pi(s,a) \leq r(s,a) + \mathsf{P}\, V_{h+1}^{\hat{\pi}}(s,a) = Q_h^{\hat{\pi}}(s,a).$$

- Then since $\pi_h(\mathrm{d}a|s)$ is a probability measure then

$$V_h^\pi(s) = \int_{\mathcal{A}} Q_h^\pi(s,a)\pi_h(\mathrm{d}a|s) \leq \max_{a \in \mathcal{A}} Q_h^\pi(s,a) \leq \max_{a \in \mathcal{A}} Q_h^{\hat{\pi}}(s,a) = V_h^{\hat{\pi}}(s).$$

Policy Iteration

Greedy policies

This theorem tell us that it is enough to consider only greedy policies Π_{greedy} when we are taking supremum over all policies Π in the definition of optimal value and action-value functions.

Algorithm 1: Policy Iteration for finite-horizon MDPs

Input: MDP $M = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \mathrm{H})$ and the immediate reward function r, iterations budget T
Initialize: π^0 as some set of policies;
for $t \in [T]$ **do**
  Compute $Q_h^{\pi^t}$ by solving Bellman equations (see Theorem 8);
  Find π^{t+1} as a greedy policy to $Q_h^{\pi^t}$.
end for
Output: estimate of optimal policy π^T.

Convergence

What about $T \to \infty$?

Optimal Bellman equations

Theorem (Optimal Bellman equations)

Fix a finite-horizon MDP $M = (\mathsf{S}, \mathsf{A}, \mathsf{P}, \mathrm{R}, \mathsf{H})$. Let r be the immediate reward function of M and assume that r is bounded. Then optimal value and action-value functions satisfies a similar optimal Bellman equations

$$Q_h^\star(s,a) = r(s,a) + \mathsf{P}\, V_{h+1}^\star(s,a) \qquad \forall (s,a,h) \in \mathsf{S} \times \mathsf{A} \times [\mathsf{H}]$$
$$V_h^\star(s) = \max_{a \in \mathsf{A}} Q_h^\star(s,a) \qquad \forall (s,h) \in \mathsf{S} \times [\mathsf{H}]$$
$$Q_{H+1}^\star(s,a) = V_{H+1}^\star(s) = 0.$$

Proof

By Bellman equations we have

$$Q_h^\star(s,a) = \sup_{\pi \in \Pi_{\text{greedy}}} Q_h^\pi(s,a) = r(s,a) + \sup_{\pi \in \Pi_{\text{greedy}}} \mathrm{P}\, V_{h+1}^\pi(s,a).$$

Since V^π are bounded for any π, then by Beppo-Levi theorem

$$\sup_{\pi \in \Pi_{\text{greedy}}} \mathrm{P}\, V_{h+1}^\pi(s,a) = \mathrm{P}\left[\sup_{\pi \in \Pi_{\text{greedy}}} V_{h+1}^\pi\right](s,a) = \mathrm{P}\, V^\star(s,a).$$

To prove the second statement we use Bellman equations and greedy policies

$$\begin{aligned} V_h^\star(s) &= \sup_{\pi \in \Pi_{\text{greedy}}} V_h^\pi(s) = \sup_{\pi \in \Pi_{\text{greedy}}} \max_{a \in \mathcal{A}} Q_h^\pi(s,a) = \max_{a \in \mathcal{A}} \sup_{\pi \in \Pi_{\text{greedy}}} Q_h^\pi(s,a) \\ &= \max_{a \in \mathcal{A}} Q_h^\star(s,a). \end{aligned}$$

Value Iteration

How to compute optimal value and action-value functions

Optimal Bellman equations gives us an alternative way to compute optimal value-function and optimal policy using *dynamic programming* directly.

Algorithm 2: Value Iteration for finite-horizon MDPs

Input: MDP $M = (\mathsf{S}, \mathsf{A}, \mathsf{P}, \mathrm{R}, \mathsf{H})$ and the immediate reward function r;
Initialize: $Q_{H+1}(s,a) = 0, \;\; V_{H+1}(s) = 0$;
for $h = H, H-1, \ldots, 1$ **do**

$$
\begin{aligned}
Q_h(s,a) &:= r(s,a) + \mathsf{P}\, V_{h+1}(s,a) && \forall (s,a) \in \mathsf{S} \times \mathsf{A}; \\
V_h(s) &= \max_{a \in \mathsf{A}} Q_h(s,a) && \forall s \in \mathsf{S}; \\
\pi_h(s) &= \arg\max_{a \in \mathsf{A}} Q_h(s,a) && \forall s \in \mathsf{S}.
\end{aligned}
$$

end for
Output: optimal policy π.

Frame Title

Corollary

Let $M = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \mathrm{H})$ be a finite-horizon MDP with $|\mathrm{A}| < \infty$. Then an optimal policy $\pi^s tar$ exists and could be computed using Value Iteration algorithm. Moreover, the policy computed by Value Iteration is greedy.

Here we may observe the main difficulties with this algorithm.

Problems

- It requires full knowledge of the model P and immediate reward function r;
- It computes value and action-value functions for all states that is impossible for $|\mathrm{S}| = \infty$.

This is a reason why finite MDPs are called *tabular*: for them it is possible to handle full table of Q-values.

MDP: infinite case

MDP: infinite case

Let $\gamma \in (0, 1]$ be the discount factor. Tuple $(\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \gamma)$ is called Markov Decision Process:

- S - state space. By $(S_k)_{k \geq 0}$ we denote a sequence of random states.
- A - action space. Let $(A_k)_{k \geq 0}$ be a sequence of random actions.
- Agent's policy $\pi(\cdot|s)$ is the distribution on A.
- Family of Markov transition kernels $(\mathrm{P}(s'|s, a))_{a \in \mathrm{A}}$:

$$\mathrm{P}(s'|s, a) := \mathbb{P}(S_k = s'|S_{k-1} = s, A_{k-1} = a)\,.$$

- *the reward distribution* $\mathrm{R}(\cdot|s, a)$ as a set of measures over $\mathbb{R}$ for any $(s, a) \in \mathrm{S} \times \mathrm{A}$ and *the immediate reward function* $r(s, a) = \mathbb{E}[\mathrm{R}(s, a)]$. The role of immediate reward function is similar to mean reward in bandits. (*For simplicity we assume that* $r(s, a)$ *is bounded in* $[0, 1]$).

Value and action-value functions

Value function

The value function $V^\pi : \mathrm{S} \to \mathbb{R}$ is an expectation of *discounted* return of agent then it start at state s_0. In other words, for all $h \geq 1$ and $s \in \mathrm{S}$

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid S_0 = s\right],$$

where for all $t \geq 0$ $r_t \sim \mathrm{R}(\cdot|s_t, a_t), S_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t), A_t \sim \pi(\cdot|s_t)$.

Action-value function

The action-value function $Q_h^\pi : \mathrm{S} \to \mathbb{R}$ is an expectation of *discounted* return of agent then it start at state s_0 and selects a prescribed action a_0. In other words, for all $s \in \mathrm{S}$ and $a \in \mathrm{A}$

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid S_0 = s, A_0 = a\right],$$

where for all $t \geq h$ $r_t \sim \mathrm{R}(\cdot|s_t, a_t), S_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t), a_{t+1} \sim \pi_t(\cdot|s_{t+1})$.

Optimal value and action-value functions

Optimal value and action-value functions

The *optimal value function* at state $s \in \mathrm{S}$:

$$V^\star(s) = \sup_\pi V^\pi(s).$$

The *optimal action-value function* at state $s \in \mathrm{S}$ and action $a \in \mathrm{A}$:

$$Q^\star(s, a) = \sup_\pi Q^\pi(s, a).$$

Frame Title

Theorem (Bellman equations: discounted MDP)

Fix a discounted MDP $M = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \gamma)$ and policy π. Let r be the immediate reward function of M. Then V^π and Q^π satisfies Bellman equations

$$Q^\pi(s,a) = r(s,a) + \gamma\, \mathrm{P}\, V^\pi(s,a), \qquad \forall (s,a) \in \mathrm{S} \times \mathrm{A}$$

$$V^\pi(s) = \sum_{a \in \mathcal{A}} Q^\pi(s,a)\pi(a|s) \qquad \forall s \in \mathrm{S}$$

Optimal policy and optimal value function

The following result holds (see, e.g., Puterman [2014]):

Theorem

When the reward function is bounded, one can always find a *deterministic* Markov policy that is optimal. Moreover, the optimal value function $V^\star := V^{\pi^\star}$ satisfies the *Bellman optimality equation*:

$$Q^\star(s,a) = r(s,a) + \mathrm{P}\, V^\star(s,a) \qquad \forall (s,a) \in \mathrm{S} \times \mathrm{A}$$
$$V^\star(s) = \max_{a \in \mathrm{A}} Q^\star(s,a) \qquad \forall s \in \mathrm{S}$$

Take greedy policy

$$\pi^\star(s) = \arg\max_{a \in \mathrm{A}} Q^\star(s,a)$$

Two Problems for MDPs

- Policy evaluation: compute V^π given fixed policy $\pi \in \Pi$.
- Policy improvement: compute or approximate some optimal policy $\pi^\star$, solve control problem.

Both could be solved with Bellman equations using fixed point iteration...

BUT

Even if transition model $\mathrm{P}(\cdot|s, a)$ is known, the expectation in the right part is often intractable!

Policy iteration and improvement in the tabular case

Assume that we know transition matrix $\mathrm{P}(\cdot|s,a)$. The value function can be represented as a vector $V^\pi \in \mathbb{R}^{|\mathrm{S}|}$.

Algorithm 3: Value iteration

Input: MDP $M = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \mathrm{H})$, immediate reward function r, policy π and number of steps T

Initialize: V_0^π;

for $k \in [T]$ **do**

$Q_k^\pi(s,a) = r(s,a) + \gamma\, \mathrm{P}\, V_{k-1}^\pi(s,a)$

$V_k^\pi(s) = \sum_{a \in \mathcal{A}} Q_k^\pi(s,a)\pi(a|s)$

end for

Output: estimate V_T^π, Q_T^π.

Due to the Banach's fixed point theorem,
$\|V_{\pi,k} - V_\pi\|_\infty \leq \gamma^k \|V_{\pi,0} - V_\pi\|_\infty$, provided that $\|V_{\pi,0}\|_\infty < \infty$.

Policy improvement

Algorithm 4: Policy improvement

Input: MDP $M = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \mathrm{H})$, immediate reward function r, policy π and number of steps T

Initialize: π_0, $V_0^{\pi_0}$;

for $k \in [T]$ **do**

$Q^{\pi_{k-1}}(s, a) = \text{Value\_Iteration}(\pi_{k-1})$;

$\pi_k(s) = \arg\max_{a \in \mathrm{A}} Q^{\pi_{k-1}}(s, a)$

end for

Output: estimate π_T.

Policy iteration and improvement in the tabular case

Problems

- Computational problems: Even in tabular case $|S| < \infty$ may be extremely large (see chess,...);
- Algorithmic problems: In infinite case the iterative procedure is intractable or cannot be correctly built.

TD(0) in the tabular case

Fix π and recall:

$$V^\pi(s) = \sum_{a \in \mathrm{A}} \pi(a|s) \Big\{ r(s,a) + \gamma \sum_{s' \in \mathrm{S}} V^\pi(s') \, \mathrm{P}(s'|s,a) \Big\}.$$

Suppose that we observe a sequence of states $(S_k)_{k \geq 0}$ (generated according to P^π) and let $r(s) = \sum_{a \in \mathrm{A}} r(s,a)\pi(a|s)$.

Algorithm 5: Policy iteration

Input: MDP $M = (\mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{R}, \mathrm{H})$, immediate reward function r, policy π and number of steps T
Initialize: V_0, s_0;
for $k \in [T]$ **do**
  Simulate $S_{k+1} \sim \mathrm{P}^\pi(\cdot|S_k)$;
  $\delta_{k+1} = r(S_k) + \gamma V_k(S_{k+1}) - V_k(S_k)$ temporal difference error;
  $V_{k+1}(s) = V_k(s) + \alpha_{k+1} \delta_{k+1} \mathbb{1}_{s=S_k}$
end for
Output: estimate V_T^π.

Step-size sequence $(\alpha_k, k \geq 0)$ are chosen by the user.
This algorithm converges to $\hat{V}$:
$F\hat{V}(s) := r(s) + \gamma \, \mathbb{E}[\hat{V}(S_1)|S_0 = s] - \hat{V}(s) = 0$. Clearly, $\hat{V} = V_\pi$.

Stochastic Approximation

- Consider the problem of finding $\theta^\star \in \mathbb{R}^d$ such that

$$f(\theta^\star) = 0\,.$$

- Only stochastic samples of $f(\theta)$ are revealed, e.g., $F(\theta; Z_n)$, such that

$$\mathbb{E}[F(\theta; Z_n)] = f(\theta) \quad \text{or, at least,} \quad \lim_{n\to\infty} \mathbb{E}[F(\theta; Z_n)] = f(\theta).$$

- Such algorithms are called *stochastic approximation (SA)* schemes to a fixed point equation:

$$\theta_{n+1} = \theta_n + \alpha_n F(\theta_n; Z_n).$$

Robbins and Monro [1951]

- The simplest instance of the problem corresponds to the Linear Stochastic Approximation (LSA)
- Compare with the standard 'Euler scheme' for numerically approximating a trajectory of the o.d.e. $\dot\theta(t) = f(\theta(t))$

$$\theta_{t+1} = \theta_t + \alpha f(\theta_t)$$

TD(0) with function approximation

- We consider approximation of the value function $V^\pi : \mathsf{S} \to \mathbb{R}$ using a parameterized function $V : \mathsf{S} \times \mathbb{R}^{\mathsf{d}} \to \mathbb{R}$, $(s, \theta) \mapsto V(s, \theta)$ where θ is a vector of parameters.
- Minimize the mean-squared error (MSE):

$$\mathrm{MSE}(\theta) = \sum_{s \in \mathsf{S}} (V^\pi(s) - V(s, \theta))^2.$$

- Suppose that we observe a sequence of states $(S_k)_{k \geq 0}$ (generated according to P^π), and that at time k, the vector of parameters is denoted as θ_k.
- Use gradient descent

$$\begin{aligned} \theta_{k+1} &= \theta_k - \frac{\alpha}{2} \nabla (V^\pi(S_k) - V(S_k, \theta_k))^2 \\ &= \theta_k + \alpha (V^\pi(S_k) - V(S_k, \theta_k)) \nabla V(S_k, \theta_k). \end{aligned} \tag{3}$$

TD(0) with function approximation

- Recall that

$$V^\pi(s) = \sum_{a \in \mathsf{A}} \pi(a|s) \Big\{ r(s,a) + \gamma \sum_{s' \in \mathsf{S}} V^\pi(s')\, \mathsf{P}(s'|s,a) \Big\}.$$

- Replace $V^\pi(S_k)$ by its estimator:

$$V^\pi(S_k) \approx r(S_k) + \gamma V(S_{k+1}, \theta_k).$$

- Rewrite (3) in the following way

$$\theta_{k+1} = \theta_k + \alpha \{ r(S_k) + \gamma V(S_{k+1}, \theta_k) - V(S_k, \theta_k) \} \nabla V(S_k, \theta_k). \quad (4)$$

- Linear function approximation:

$$V(s, \theta) = \theta^\top \psi(s),$$

where $\psi(s) = [\psi^1(s), \ldots, \psi^d(s)]^\top$. The vector $\psi(s)$ is referred to as the *feature vector* associated to the state $s \in \mathsf{S}$.

- The gradient of the approximate value function in such case is

$$\nabla V(s, \theta) = \psi(s).$$

TD(0) with function approximation

- Define for $k \geq 0$, $Z_k = [S_{k-1}, S_k]^\top$. We may rewrite (4) as

$$\theta_{k+1} = \{\mathsf{I} - \alpha \mathbf{A}(Z_{k+1})\}\,\theta_k + \alpha \mathbf{b}(Z_{k+1}),$$

where $\mathbf{A}(z)$ is a $d \times d$ matrix given for $z = [s, s']^\top \in \mathsf{S}^2$ by

$$\mathbf{A}(z) = \psi(s)\{\psi(s) - \gamma\psi(s')\}^\top, \tag{5}$$

and $\mathbf{b}(z)$ is a $d \times 1$ vector given by

$$\mathbf{b}(z) = \overline{\mathrm{R}}(s)\psi(s). \tag{6}$$

- Note that $\{Z_k\}_{k\geq 0}$ is a Markov chain on the state-space

$$\mathsf{Z} = \{z = (s_0, s_1) \in \mathsf{S}^2, \mathsf{P}^\pi(\mathsf{s}_0, \mathsf{s}_1) > 0\}. \tag{7}$$

- The transition matrix of this Markov chain is given, for any $(s_0, s_1), (s_0', s_1') \in \mathsf{Z}$, by

$$\mathsf{P}(s_0, s_1; s_0', s_1') = \delta_{s_1, s_0'}\, \mathsf{P}^\pi(s_1, s_1'),$$

  where $\delta_{u,v}$ is the Kronecker symbol.
- It is easily seen that this P has a unique invariant distribution given by

$$\bar{\pi}(s_0, s_1) = \bar{\pi}_0(s_0)\, \mathsf{P}^\pi(s_0, s_1),$$

  where $\bar{\pi}_0$ is stationary distribution of P^π (we assume that $\bar{\pi}_0$ exists)

Linear Stochastic Approximation

- Given $\bar{\mathbf{A}} \in \mathbb{R}^{d \times d}$ and $\bar{\mathbf{b}} \in \mathbb{R}^d$, we aim at finding $\theta^\star \in \mathbb{R}^d$, which is a solution of $\bar{\mathbf{A}}\theta^\star = \bar{\mathbf{b}}$
- Our analysis is based on noisy observations $\{(\mathbf{A}(Z_n), \mathbf{b}(Z_n))\}_{n \in \mathbb{N}}$. Here $\mathbf{A} : \mathsf{Z} \to \mathbb{R}^{d \times d}$, $\mathbf{b} : \mathsf{Z} \to \mathbb{R}^d$ are measurable functions, and $(Z_k)_{k \in \mathbb{N}}$ is
  - either an i.i.d. sequence taking values in a state space $(\mathsf{Z}, \mathcal{Z})$ with distribution π satisfying $\mathbb{E}[\mathbf{A}(Z_1)] = \bar{\mathbf{A}}$ and $\mathbb{E}[\mathbf{b}(Z_1)] = \bar{\mathbf{b}}$;
  - or a Z-valued ergodic Markov chain with unique invariant distribution π, such that $\lim_{n \to +\infty} \mathbb{E}[\mathbf{A}(Z_n)] = \bar{\mathbf{A}}$ and $\lim_{n \to +\infty} \mathbb{E}[\mathbf{b}(Z_n)] = \bar{\mathbf{b}}$.

  For a fixed step size $\alpha > 0$, burn-in period $n_0 \in \mathbb{N}$, and initialization θ_0, consider the sequences of estimates $\{\theta_n\}_{n \in \mathbb{N}}, \{\bar{\theta}_{n_0,n}\}_{n \geq n_0+1}$ given by

$$\begin{aligned} \theta_k &= \theta_{k-1} - \alpha\{\mathbf{A}(Z_k)\theta_{k-1} - \mathbf{b}(Z_k)\}\,, \quad k \geq 1, \\ \bar{\theta}_{n_0,n} &= (n - n_0)^{-1} \textstyle\sum_{k=n_0}^{n-1} \theta_k\,, \quad n \geq n_0 + 1\,. \end{aligned} \tag{8}$$

Linear Stochastic Approximation

Set

$$\tilde{\mathbf{A}}(z) = \mathbf{A}(z) - \bar{\mathbf{A}}\,, \quad \tilde{\mathbf{b}}(z) = \mathbf{b}(z) - \bar{\mathbf{b}}\,, \quad \varepsilon(z) = \tilde{\mathbf{A}}(z)\theta^\star - \tilde{\mathbf{b}}(z)\,,$$

and denote by $\Gamma_{1:n}^{(\alpha)}$ the product of random matrices

$$\Gamma_{m:n}^{(\alpha)} = \prod_{i=m}^{n}(\mathrm{I} - \alpha\mathbf{A}(Z_i))\,, \quad m, n \in \mathbb{N}^*, \quad m \leq n\,.$$

(8) implies the following decomposition

$$\theta_n - \theta^\star = \tilde{\theta}_n^{(\mathsf{tr})} + \tilde{\theta}_n^{(\mathsf{fl})},$$

where $\tilde{\theta}_n^{(\mathsf{tr})}$ is a transient term (reflecting the forgetting of initial condition) and $\tilde{\theta}_n^{(\mathsf{fl})}$ is a fluctuation term (reflecting misadjustement noise)

$$\tilde{\theta}_n^{(\mathsf{tr})} = \Gamma_{1:n}^{(\alpha)}\{\theta_0 - \theta^\star\}\,, \quad \tilde{\theta}_n^{(\mathsf{fl})} = -\alpha\sum_{j=1}^{n}\Gamma_{j+1:n}^{(\alpha)}\varepsilon(Z_j)\,.$$

A cornerstone of the theoretical analysis is a tight bound for $\mathbb{E}^{1/p}[\|\Gamma_{m:n}^{(\alpha)}\|^p]$ under some assumptions on the matrix $\bar{\mathbf{A}}$.

Exponential stability

Exponential stability of $\{\mathbf{A}(Z_i)\}_{i\in\mathbb{N}^*}$ (see Guo and Ljung [1995], Ljung [2002])

For $q \geq 1$, there exist $\mathsf{a}_q, \mathsf{C}_q > 0$ and $\alpha_{\infty,q} < \infty$ such that, for any step size $\alpha \leq \alpha_{\infty,q}$, $m, n \in \mathbb{N}$, $m < n$,

$$\mathbb{E}[\|\Gamma_{m:n}^{(\alpha)}\|^q] \leq \mathsf{C}_q \exp\left(-\mathsf{a}_q \alpha (n-m)\right) .$$

Intuitively, $\Gamma_{m:n}^{(\alpha)} \approx (\mathsf{I} - \alpha\bar{\mathbf{A}})^{n-m}$, for $m, n \in \mathbb{N}$, $m \leq n$, under the assumption that $-\bar{\mathbf{A}}$ is Hurwitz, i.e., for any eigenvalue λ of $\bar{\mathbf{A}}$, we have $\mathrm{Re}(\lambda) > 0$.

Theorem

Assume that $-\bar{\mathbf{A}}$ is Hurwitz. There exists a unique symmetric positive definite matrix Q satisfying the Lyapunov equation $\bar{\mathbf{A}}^\top Q + Q\bar{\mathbf{A}} = \mathsf{I}$. In addition, setting

$$a = \|Q\|^{-1}/2\,, \quad \text{and} \quad \alpha_\infty = (1/2)\|\bar{\mathbf{A}}\|_Q^{-2}\|Q\|^{-1} \wedge \|Q\|\,,$$

for any $\alpha \in [0, \alpha_\infty]$, it holds that $\|\mathsf{I} - \alpha\bar{\mathbf{A}}\|_Q^2 \leq 1 - a\alpha$, and $\alpha a \leq 1/2$.

Technical assumptions

Assumption A1

1. $\{Z_k\}_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ with distribution π.
2. $\mathsf{C}_A = \sup_{z\in\mathsf{Z}} \|\mathbf{A}(z)\| \vee \sup_{z\in\mathsf{Z}} \|\tilde{\mathbf{A}}(z)\| < \infty$ and the matrix $-\bar{\mathbf{A}}$ is Hurwitz
3. There exists $\mathsf{C}_\varepsilon < +\infty$, such that for any $z \in \mathsf{Z}$, $\|\varepsilon(z)\| \le \mathsf{C}_\varepsilon \sqrt{\mathrm{Tr}\, \Sigma_\varepsilon}$, where
$$\Sigma_\varepsilon = \int_{\mathsf{Z}} \varepsilon(z)\varepsilon(z)^\top \mathrm{d}\pi(z)\,.$$

A1 and exponential bounds

- We show that under only A1, for fixed $\alpha > 0$, $\lim_{n\to+\infty} \mathbb{E}[\|\theta_n - \theta^\star\|^p] = \infty$ for $p \geq \bar{p}(\alpha)$;
- Exponential high probability bounds for $\|\theta_n - \theta^\star\|$ are not possible.
- Consider the one-dimensional instance of LSA problem $(2q-1)\theta^\star = 0$, $q > 1/2$;
- Let $\theta_0 > 0$, $\mathbf{b}_n = 0$, and

$$\mathbf{A}_n = \begin{cases} 1 & \text{with probability } q\,, \\ -1 & \text{with probability } 1-q\,. \end{cases}$$

- $\theta^\star = 0$, and LSA recursion is simply

$$\theta_n = \prod_{k=1}^{n} (1 - \alpha \mathbf{A}_k)\theta_0\,.$$

- If $\alpha \in (0,1)$ is fixed, for any $p > \bar{p}(q,\alpha)$, we have $\lim_{n\to+\infty} \mathbb{E}\left[|\theta_n|^p\right] = \infty$, while $\theta_n \xrightarrow{W} 0$.

Exponential stability, IID case

Let

$$\kappa_{\mathsf{Q}} = \lambda_{\max}(Q)/\lambda_{\min}(Q)\,, \quad b_Q = \sqrt{\kappa_{\mathsf{Q}}}\,\mathsf{C}_A\,,$$
$$\alpha_{q,\infty} = \alpha_\infty \wedge \mathsf{c}_{\mathbf{A}}\,/q\,, \quad \mathsf{c}_{\mathbf{A}} = a/\{2b_Q^2\}\,.$$

Theorem

Assume A1. For any $p, q \in \mathbb{N}$, $2 \le p \le q$, $\alpha \in (0, \alpha_{q,\infty}]$ and $n \in \mathbb{N}^*$, it holds

$$\mathbb{E}^{1/p}\left[\|\Gamma_{1:n}^{(\alpha)}\|^p\right] \le \sqrt{\kappa_{\mathsf{Q}}}\, d^{1/q}(1 - a\alpha + (q-1)b_Q^2\alpha^2)^{n/2}\,.$$

HP bound for the LSA error $\|\theta_n - \theta^\star\|$ in IID case

Theorem

Assume A1 and fix $\delta \in (0,1)$. Then, for any $\theta_0 \in \mathbb{R}^d$, sample size $n \in \mathbb{N}$ satisfying

$$n/\log n \geq (a/4)\left\{\alpha_\infty^{-1} \vee a^{-1}(1+\log d)\log\left(2\mathrm{e}/\delta\right)\right\},$$

and step size $\alpha = 4\log n/(an)$, it holds with probability at least $1-\delta$, that

$$\|\theta_n - \theta^\star\| \leq 4\mathrm{e}D_2\sqrt{\frac{\{\mathrm{Tr}\,\Sigma_\varepsilon\}\log n \log(2\mathrm{e}/\delta)}{n}} + \frac{2\mathrm{e}\kappa_{\mathrm{Q}}^{1/2}\|\theta_0 - \theta^\star\|}{n}.$$

Here $\alpha_\infty, a, \kappa_{\mathrm{Q}}$ are some constants and $\Sigma_\varepsilon = \int_{\mathsf{Z}} \varepsilon(z)\varepsilon(z)^\top \mathrm{d}\pi(z)$

See papers: Durmus et al. [2021a,b]. For Polyak-Ruppert averaging see Durmus et al. [2022].

Learning a (near-) optimal policy

SARSA-Algorithm

Recall the Bellman expectation equation for Q-function:

$$Q^\pi(s,a) = r(s,a) + \gamma\, \mathbb{E}\Big[Q^\pi(S_1, A_1) | S_0 = s, A_0 = a \Big].$$

Idea: the best immediate action is $a = \arg\max_{a' \in \mathrm{A}} Q_\pi(s, a')$ (Exploitation).
Now use Robbins-Monro method!

Algorithm 6: SARSA algorithm

Input: Q_0
Set $\pi = \pi_{Q_0}$;
for $k = 1, 2, \ldots$ **do**
  Sample $S_k, A_k, S_{k+1}, A_{k+1}$ with $A_\ell = \pi(S_\ell)$;
  Update Q: $Q_k(S_k, A_k) =$
    $Q_{k-1}(S_k, A_k) + \alpha_k(r(S_k, A_k) + \gamma Q_{k-1}(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k))$;
  Update Policy. $\pi = \pi_{Q_k}$.

This is an **on-policy** algorithm: we update π_{Q_k} with samples from π_{Q_k}.

Expected SARSA

Instead of estimate

$$r(S_k, A_k) + \gamma Q_{k-1}(S_{k+1}, A_{k+1})$$

for the right part of the Bellman equation, use another. Exploit another policy π_b

$$r(S_k, A_k) + \gamma \sum_{a \in \mathrm{A}} \pi_b(a|S_k) Q_{k-1}(S_{k+1}, a).$$

where $A \sim \pi_b(\cdot|_k)$. So we get **off-policy** algorithm: we use another (fixed) policy π_b to update the estimate of greedy π_{Q_k}.

Q-Learning

We might recall Bellman's optimality equation for $Q^\star$-function:

$$Q^\star(s,a) = r(s,a) + \gamma \sum_{s' \in \mathsf{S}} \max_{a' \in \mathsf{A}} Q^\star(s',a')\, \mathsf{P}(s'|s,a)$$

and apply Robbins-Monro algorithm to this equation in $Q^\star$.

Algorithm 7: Q-learning algorithm

Input: Q_0
Set $\pi = \pi_{Q_0}$;
for $k = 1, 2, \ldots$ **do**
  Sample $S_k, A_k, S_{k+1}, A_{k+1}$ with $a_\ell = \pi(S_\ell)$;
  Update Q: $Q_k(S_k, A_k) =$
  $Q_{k-1}(S_k, A_k) + \alpha_k(r(S_k, A_k) + \gamma \max_{a \in \mathsf{A}} Q_{k-1}(S_{k+1}, a) - Q_{k-1}(S_k, A_k))$;
  Update Policy. $\pi = \pi_{Q_k}$.

This is an **off-policy** algorithm since we update the estimate of $Q_\star$ with samples from other policies π_{Q_k}.

Thank you!

References I

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize. In *NeurIPS*, 2021a.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1711–1752. PMLR, 15–19 Aug 2021b. URL https://proceedings.mlr.press/v134/durmus21a.html.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time High-probability Bounds for Polyak-Ruppert Averaged Iterates of Linear Stochastic Approximation. *arXiv e-prints*, art. arXiv:2207.04475, July 2022.

L. Guo and L. Ljung. Exponential stability of general tracking algorithms. *IEEE Transactions on Automatic Control*, 40(8):1376–1387, 1995.

Lennart Ljung. Recursive identification algorithms. *Circuits, Systems and Signal Processing*, 21(1): 57–68, 2002.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Roman Vershynin. High-dimensional probability. 2019. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.