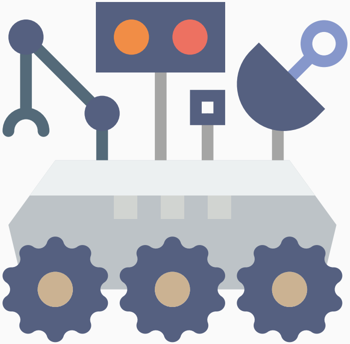
Optimal Policy Identification



Wouter M. Koolen

Centrum Wiskunde & Informatica

University of Twente

RL Seminar
University of Twente
Friday 9th June, 2023

Today we are looking at

A. Al Marjani and A. Proutiere (2021). “Adaptive sampling for best policy identification in Markov decision processes”. In: **International Conference on Machine Learning**. PMLR, pp. 7459–7468

Outline

1. Setup

Markov Decision Process

Finite sets of states S and **actions** A. Rewards bounded in $[0, 1]$.

Markov Decision Process

Finite sets of states S and **actions** A. Rewards bounded in $[0, 1]$.

MDP $\phi = (p_\phi, q_\phi)$ specified by

- dynamics $p_\phi(s'|s, a)$ and
- rewards $q_\phi(r|s, a)$.

Markov Decision Process

Finite sets of states S and **actions** A. Rewards bounded in $[0, 1]$.

MDP $\phi = (p_\phi, q_\phi)$ specified by

- dynamics $p_\phi(s'|s, a)$ and
- rewards $q_\phi(r|s, a)$.

We will write $r_\phi(s, a)$ for the mean reward (of doing a in s under ϕ).

Policies and their Performance

A **policy** is map $\pi : S \to A$.

Executing a policy π from state s under ϕ gives a sequence $(s_t^\pi)_{t\geq 0}$ with

$$s_0^\pi = s \qquad \text{and} \qquad s_{t+1}^\pi \sim p_\phi\left(\cdot \middle| s_t^\pi, \pi(s_t^\pi)\right)$$

Policies and their Performance

A **policy** is map $\pi : S \to A$.

Executing a policy π from state s under ϕ gives a sequence $(s_t^\pi)_{t \geq 0}$ with

$$s_0^\pi = s \qquad \text{and} \qquad s_{t+1}^\pi \sim p_\phi\left(\cdot \middle| s_t^\pi, \pi(s_t^\pi)\right)$$

Let's introduce discount factor $\gamma \in (0, 1)$.

Value function of policy π in ϕ:

$$V_\phi^\pi(s) = \mathbb{E}_\phi \left[\sum_{t=0}^{\infty} \gamma^t r_\phi(s_t^\pi, \pi(s_t^\pi)) \middle| s_0^\pi = s \right]$$

Policies and their Performance

A **policy** is map $\pi : S \to A$.

Executing a policy π from state s under ϕ gives a sequence $(s_t^\pi)_{t \geq 0}$ with

$$s_0^\pi = s \qquad \text{and} \qquad s_{t+1}^\pi \sim p_\phi\left(\cdot \middle| s_t^\pi, \pi(s_t^\pi)\right)$$

Let's introduce discount factor $\gamma \in (0, 1)$.

Value function of policy π in ϕ:

$$V_\phi^\pi(s) = \mathbb{E}_\phi\left[\sum_{t=0}^{\infty} \gamma^t r_\phi\left(s_t^\pi, \pi(s_t^\pi)\right) \middle| s_0^\pi = s\right]$$

Optimal policy for ϕ is $\pi^*(\phi) \coloneqq \arg\max_{\pi : S \to A} V_\phi^\pi$ (NB: not a scalar(!))

Problem in this talk

Problem

Given any unknown MDP ϕ, identify its optimal policy $\pi^*(\phi)$ from interactive exploration.

We want

- Reliability: output policy is indeed optimal
- **Efficiency**: as few samples as possible

Interactive Exploration: Interaction with a generative model

Fix unknown MDP ϕ.

Protocol (generative model)

for $t = 1, 2, \ldots, \tau$

- Learner picks state s_t and action a_t
- Learner observes reward $r_t \sim q_\phi(\cdot|s_t, a_t)$ and successor state $s'_t \sim p_\phi(\cdot|s_t, a_t)$

Learner recommends policy $\hat{\pi}$.

Interactive Exploration: Interaction with a generative model

Fix unknown MDP ϕ.

Protocol (generative model)

for $t = 1, 2, \ldots, \tau$

- Learner picks state s_t and action a_t
- Learner observes reward $r_t \sim q_\phi(\cdot|s_t, a_t)$ and successor state $s'_t \sim p_\phi(\cdot|s_t, a_t)$

Learner recommends policy $\hat{\pi}$.

NB: generative model different from **navigation**, where $s_{t+1} = s'_t$.

Interactive Exploration: Interaction with a generative model

Fix unknown MDP ϕ.

Protocol (generative model)

for $t = 1, 2, \ldots, \tau$

- Learner picks state s_t and action a_t
- Learner observes reward $r_t \sim q_\phi(\cdot|s_t, a_t)$ and successor state $s_t' \sim p_\phi(\cdot|s_t, a_t)$

Learner recommends policy $\hat{\pi}$.

NB: generative model different from **navigation**, where $s_{t+1} = s_t'$.

Definition (Correctness)

Fix confidence $\delta \in (0, 1)$. A learner is δ-correct if $\mathbb{P}_\phi(\hat{\pi} \neq \pi^*(\phi)) \leq \delta$.

Interactive Exploration: Interaction with a generative model

Fix unknown MDP ϕ.

Protocol (generative model)

for $t = 1, 2, \ldots, \tau$

- Learner picks state s_t and action a_t
- Learner observes reward $r_t \sim q_\phi(\cdot|s_t, a_t)$ and successor state $s'_t \sim p_\phi(\cdot|s_t, a_t)$

Learner recommends policy $\hat{\pi}$.

NB: generative model different from **navigation**, where $s_{t+1} = s'_t$.

Definition (Correctness)

Fix confidence $\delta \in (0, 1)$. A learner is δ-correct if $\mathbb{P}_\phi\left(\hat{\pi} \neq \pi^*(\phi)\right) \leq \delta$.

Quest: efficiency

Minimise sample complexity $\phi \mapsto \mathbb{E}_\phi[\tau]$ over all δ-correct learners.

Rolling our own : concentration and perturbation

Uniform sampling: sample each pair (s, a) for n times. Concentration results say

- Estimate $r_\phi(s, a)$ up to precision $1/\sqrt{n}$.
- Estimate $p_\phi(s'|s, a)$ up to precision $1/\sqrt{n}$ for each s'.

Rolling our own : concentration and perturbation

Uniform sampling: sample each pair (s, a) for n times. Concentration results say

- Estimate $r_\phi(s, a)$ up to precision $1/\sqrt{n}$.
- Estimate $p_\phi(s'|s, a)$ up to precision $1/\sqrt{n}$ for each s'.

The value function of π satisfies the recurrence

$$V_\phi^\pi(s) \; = \; r_\phi(s, \pi(s)) + \gamma \sum_{s'} p_\phi\left(s' \middle| s, \pi(s)\right) V_\phi^\pi(s')$$

That is

$$V_\phi^\pi \; = \; \left(I - \gamma \sum_{s,s'} p_\phi\left(s' \middle| s, \pi(s)\right) e_s e_{s'}^\intercal\right)^{-1} \sum_s e_s r_\phi(s, \pi(s))$$

A perturbation argument gives $\|V_\phi^\pi - V_{\hat{\phi}}^\pi\|_\infty \leq O\left(\frac{1}{\sqrt{n}(1-\gamma)^2}\right)$.

Rolling our own : concentration and perturbation

Uniform sampling: sample each pair (s, a) for n times. Concentration results say

- Estimate $r_\phi(s, a)$ up to precision $1/\sqrt{n}$.
- Estimate $p_\phi(s'|s, a)$ up to precision $1/\sqrt{n}$ for each s'.

The value function of π satisfies the recurrence

$$V_\phi^\pi(s) \;=\; r_\phi(s, \pi(s)) + \gamma \sum_{s'} p_\phi\left(s' \middle| s, \pi(s)\right) V_\phi^\pi(s')$$

That is

$$V_\phi^\pi \;=\; \left(I - \gamma \sum_{s,s'} p_\phi\left(s' \middle| s, \pi(s)\right) e_s e_{s'}^\intercal\right)^{-1} \sum_s e_s r_\phi(s, \pi(s))$$

A perturbation argument gives $\|V_\phi^\pi - V_{\hat{\phi}}^\pi\|_\infty \leq O\left(\frac{1}{\sqrt{n}(1-\gamma)^2}\right)$.

Obtain ϵ-optimal policy in using $O\left(\frac{SA}{\epsilon^2(1-\gamma)^4} \ln(SA)\right)$ samples.

Instance-Optimal vs Worst-case Optimal

Theorem (Azar, Munos, and Kappen, 2013)

Can identify an ϵ-optimal policy in samples

$$O\left(\frac{SA}{\epsilon^2(1-\gamma)^3}\ln(SA)\right)$$

And no algorithm can do uniformly better. Worst-case optimal!

Instance-Optimal vs Worst-case Optimal

Theorem (Azar, Munos, and Kappen, 2013)

Can identify an ϵ-optimal policy in samples

$$O\left(\frac{SA}{\epsilon^2(1-\gamma)^3}\ln(SA)\right)$$

And no algorithm can do uniformly better. Worst-case optimal!

But maybe some MDPs ϕ are **easy**?

Instance-Optimal vs Worst-case Optimal

Theorem (Azar, Munos, and Kappen, 2013)

Can identify an ϵ-optimal policy in samples

$$O\left(\frac{SA}{\epsilon^2(1-\gamma)^3}\ln(SA)\right)$$

And no algorithm can do uniformly better. Worst-case optimal!

But maybe some MDPs ϕ are **easy**?

Can we have instance-dependent / instance-optimal results?

Discrimination

In MDP ϕ, the average total evidence collected against MDP ψ is

$$
\begin{aligned}
&\mathsf{KL}_{\phi|\psi}\big((s_t, a_t, s'_t, r_t)_{t=1}^{\tau}, \hat{\pi}\big) \\
&= \sum_{s,a} \mathbb{E}_\phi[N_{s,a}(\tau)] \Big\{ \mathsf{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathsf{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\} \\
&= \mathbb{E}_\phi[\tau] \sum_{s,a} \frac{\mathbb{E}_\phi[N_{s,a}(\tau)]}{\mathbb{E}_\phi[\tau]} \Big\{ \mathsf{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathsf{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\}
\end{aligned}
$$

Discrimination

In MDP ϕ, the average total evidence collected against MDP ψ is

$$\begin{aligned}
&\mathrm{KL}_{\phi|\psi}\big((s_t, a_t, s'_t, r_t)_{t=1}^{\tau}, \hat{\pi}\big) \\
&= \sum_{s,a} \mathbb{E}_\phi[N_{s,a}(\tau)] \Big\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\} \\
&= \mathbb{E}_\phi[\tau] \sum_{s,a} \frac{\mathbb{E}_\phi[N_{s,a}(\tau)]}{\mathbb{E}_\phi[\tau]} \Big\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\}
\end{aligned}$$

Moreover, if Learner is δ-correct and $\pi^*(\phi) \neq \pi^*(\psi)$ then

$$\mathrm{KL}_{\phi|\psi}\big((s_t, a_t, s'_t, r_t)_{t=1}^{\tau}, \hat{\pi}\big) \;\geq\; \mathrm{KL}_{\phi|\psi}\big(\mathbf{1}_{\hat{\pi}=\pi^*(\phi)}\big) \;\geq\; \mathrm{KL}(1-\delta, \delta) \;\approx\; \ln \frac{1}{\delta}$$

Discrimination

In MDP ϕ, the average total evidence collected against MDP ψ is

$$\begin{aligned}
&\mathrm{KL}_{\phi|\psi}\big((s_t, a_t, s_t', r_t)_{t=1}^{\tau}, \hat{\pi}\big) \\
&= \sum_{s,a} \mathbb{E}_\phi[N_{s,a}(\tau)] \Big\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\} \\
&= \mathbb{E}_\phi[\tau] \sum_{s,a} \frac{\mathbb{E}_\phi[N_{s,a}(\tau)]}{\mathbb{E}_\phi[\tau]} \Big\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\}
\end{aligned}$$

Moreover, if Learner is δ-correct and $\pi^*(\phi) \neq \pi^*(\psi)$ then

$$\mathrm{KL}_{\phi|\psi}\big((s_t, a_t, s_t', r_t)_{t=1}^{\tau}, \hat{\pi}\big) \ \geq\ \mathrm{KL}_{\phi|\psi}\big(\mathbf{1}_{\hat{\pi}=\pi^*(\phi)}\big) \ \geq\ \mathrm{KL}(1-\delta, \delta) \ \approx\ \ln \frac{1}{\delta}$$

So all in all

$$\mathbb{E}_\phi[\tau] \ \geq\ \frac{\ln \frac{1}{\delta}}{\max\limits_{\boldsymbol{w} \in \triangle_{SA}} \min\limits_{\psi : \pi^*(\psi) \neq \pi^*(\phi)} \sum_{s,a} w_{s,a} \Big\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\}}$$

Optimal Algorithm (Track-and-Stop template)

Can we have a single algorithm so that for all ϕ,

$$\mathbb{E}_\phi[\tau] \le \frac{\ln \frac{1}{\delta}}{\max\limits_{\boldsymbol{w} \in \triangle_{SA}} \min\limits_{\psi : \pi^*(\psi) \neq \pi^*(\phi)} \sum\limits_{s,a} w_{s,a} \Big\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\}} + \text{tiny?}$$

Optimal Algorithm (Track-and-Stop template)

Can we have a single algorithm so that for all ϕ,

$$\mathbb{E}_\phi[\tau] \le \frac{\ln \frac{1}{\delta}}{\max\limits_{\boldsymbol{w} \in \triangle_{SA}} \min\limits_{\psi : \pi^*(\psi) \neq \pi^*(\phi)} \sum\limits_{s,a} w_{s,a} \left\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \right\}} + \text{tiny?}$$

Track and Stop

For $t = 1, 2, \ldots$

- Form estimate $\hat{\phi}_t$ of MDP
- Compute oracle weights $\boldsymbol{w}^*(\hat{\phi}_t)$.
- Track $\boldsymbol{w}^*$ with sub-linear forced exploration.
- Stop/recommend using GLRT (generalised likelihood ratio test).

Analysis: as $t \to \infty$, then $\hat{\phi}_t \to \phi$ hence $\boldsymbol{w}^*(\hat{\phi}_t) \to \boldsymbol{w}^*(\phi)$ and we stop at optimal time + tiny.

Conclusion

The instance dependent problem complexity of MDPs is (apparently)

$$\frac{1}{\max_{\boldsymbol{w} \in \triangle_{SA}} \min_{\psi : \pi^*(\psi) \neq \pi^*(\phi)} \sum_{s,a} w_{s,a} \Big\{ \mathrm{KL}\big(p_\phi(\cdot|s,a) \big\| p_\psi(\cdot|s,a)\big) + \mathrm{KL}\big(q_\phi(\cdot|s,a) \big\| q_\psi(\cdot|s,a)\big) \Big\}}$$

as lower and upper bounds match.

Algorithmics **not settled**.

Current algorithms target relaxations instead.

Thanks!

References i

Al Marjani, A. and A. Proutiere (2021). "Adaptive sampling for best policy identification in Markov decision processes". In: **International Conference on Machine Learning**. PMLR, pp. 7459–7468.

Azar, M. G., R. Munos, and H. J. Kappen (2013). "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model". In: **Machine learning** 91, pp. 325–349.