Optimal Policy Identification

Wouter M. Koolen

RL Seminar
University of Twente
Friday 9th June, 2023
Today we are looking at

1. Setup
Finite sets of **states** $S$ and **actions** $A$. Rewards bounded in $[0, 1]$. 
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 $\phi$).
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 $\phi$).
A **policy** is map $\pi : S \rightarrow A$.

**Executing** a policy $\pi$ from state $s$ under $\phi$ gives a sequence $(s_0^\pi)_t$ with $t \geq 0$ with

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

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

**Value function of policy $\pi$ in $\phi$**:

$$V_\pi(\phi)(s) = \mathbb{E}_\phi \, \sum_{t=0}^{\infty} \gamma^t r_\phi(s_t^\pi, \pi(s_t^\pi))$$
A policy is map $\pi : S \rightarrow A$.

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

$$s^\pi_0 = s \quad \text{and} \quad s^\pi_{t+1} \sim p_\phi(\cdot | s^\pi_t, \pi(s^\pi_t))$$

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

Value function of policy $\pi$ in $\phi$:

$$V^\pi_\phi(s) = \mathbb{E}_\phi \left[ \sum_{t=0}^{\infty} \gamma^t r_\phi(s^\pi_t, \pi(s^\pi_t)) \mid s^\pi_0 = s \right]$$
A **policy** is map $\pi : S \to A$.

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

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

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

Value function of policy $\pi$ in $\phi$:

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

**Optimal policy** for $\phi$ is $\pi^*(\phi) := \arg \max_{\pi : S \to A} V_\phi^\pi$ (NB: not a scalar(!))
Problem in this talk

**Problem**

Given any unknown MDP $\phi$, 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 $\phi$.

**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 $\phi$.

**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 $\phi$.

**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 $\delta$-correct if $\mathbb{P}_\phi (\hat{\pi} \neq \pi^*(\phi)) \leq \delta$. 
Interactive Exploration: Interaction with a generative model

Fix unknown MDP $\phi$.

**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 $\delta$-correct if $\mathbb{P}_\phi(\hat{\pi} \neq \pi^*(\phi)) \leq \delta$.

**Quest: efficiency**

Minimise sample complexity $\phi \mapsto \mathbb{E}_\phi[\tau]$ over all $\delta$-correct learners.
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 \(s'\).
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 \(\pi\) satisfies the recurrence

\[
V^\pi_\phi(s) = r_\phi(s, \pi(s)) + \gamma \sum_{s'} p_\phi(s'|s, \pi(s)) V^\pi_\phi(s')
\]

That is

\[
V^\pi_\phi = \left( I - \gamma \sum_{s,s'} p_\phi(s'|s, \pi(s)) e_s e_{s'}^T \right)^{-1} \sum_s e_s r_\phi(s, \pi(s))
\]

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

\[
V^\pi_\phi(s) = r_\phi(s, \pi(s)) + \gamma \sum_{s'} p_\phi(s'|s, \pi(s)) V^\pi_\phi(s')
\]

That is

\[
V^\pi_\phi = \left( I - \gamma \sum_{s, s'} p_\phi(s'|s, \pi(s)) e_s e_{s'}^T \right)^{-1} \sum_s e_s r_\phi(s, \pi(s))
\]

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

Obtain \(\epsilon\)-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 $\epsilon$-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!**
Theorem (Azar, Munos, and Kappen, 2013)

Can identify an $\epsilon$-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 $\phi$ are **easy**?
Instance-Optimal vs Worst-case Optimal

**Theorem (Azar, Munos, and Kappen, 2013)**

*Can identify an $\epsilon$-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 $\phi$ are **easy**?

Can we have instance-dependent / instance-optimal results?
In MDP $\phi$, the average total evidence collected against MDP $\psi$ is

$$\text{KL}_{\phi|\psi}((s_t, a_t, s'_t, r_t)_{t=1}^{\mathcal{T}}, \hat{\pi})$$

$$= \sum_{s, a} \mathbb{E}_{\phi}[N_{s,a}(\tau)] \left\{ \text{KL}(p_{\phi}(\cdot|s, a)||p_{\psi}(\cdot|s, a)) + \text{KL}(q_{\phi}(\cdot|s, a)||q_{\psi}(\cdot|s, a)) \right\}$$

$$= \mathbb{E}_{\phi}[\tau] \sum_{s, a} \frac{\mathbb{E}_{\phi}[N_{s,a}(\tau)]}{\mathbb{E}_{\phi}[\tau]} \left\{ \text{KL}(p_{\phi}(\cdot|s, a)||p_{\psi}(\cdot|s, a)) + \text{KL}(q_{\phi}(\cdot|s, a)||q_{\psi}(\cdot|s, a)) \right\}$$
In MDP $\phi$, the average total evidence collected against MDP $\psi$ is

$$
\text{KL}_{\phi|\psi}((s_t, a_t, s'_t, r_t)_{t=1}^{T}, \hat{\pi})
= \sum_{s,a} \mathbb{E}_{\phi}[N_{s,a}(\tau)] \left\{ \text{KL}(p_{\phi}(\cdot|s,a)\|p_{\psi}(\cdot|s,a)) + \text{KL}(q_{\phi}(\cdot|s,a)\|q_{\psi}(\cdot|s,a)) \right\}
= \mathbb{E}_{\phi}[\tau] \sum_{s,a} \frac{\mathbb{E}_{\phi}[N_{s,a}(\tau)]}{\mathbb{E}_{\phi}[\tau]} \left\{ \text{KL}(p_{\phi}(\cdot|s,a)\|p_{\psi}(\cdot|s,a)) + \text{KL}(q_{\phi}(\cdot|s,a)\|q_{\psi}(\cdot|s,a)) \right\}
$$

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

$$
\text{KL}_{\phi|\psi}((s_t, a_t, s'_t, r_t)_{t=1}^{T}, \hat{\pi}) \geq \text{KL}_{\phi|\psi}(1_{\hat{\pi} = \pi^*(\phi)}) \geq \text{KL}(1 - \delta, \delta) \approx \ln \frac{1}{\delta}
$$
In MDP $\phi$, the average total evidence collected against MDP $\psi$ is

$$
KL_{\phi \mid \psi}((s_t, a_t, s'_t, r_t)_{t=1}^{\tau}, \hat{\pi}) = \sum_{s,a} E_{\phi}[N_{s,a}(\tau)] \left\{ KL(p_\phi(\cdot | s, a) \| p_\psi(\cdot | s, a)) + KL(q_\phi(\cdot | s, a) \| q_\psi(\cdot | s, a)) \right\}
$$

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

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

So all in all

$$
E_{\phi}[\tau] \geq \frac{\ln \frac{1}{\delta}}{\max_{\omega \in \Delta SA} \min_{\psi: \pi^*(\psi) \neq \pi^*(\phi)} \sum_{s,a} w_{s,a} \left\{ KL(p_\phi(\cdot | s, a) \| p_\psi(\cdot | s, a)) + KL(q_\phi(\cdot | s, a) \| q_\psi(\cdot | s, a)) \right\}}
$$
Can we have a single algorithm so that for all $\phi$,

$$
\mathbb{E}_\phi [\tau] \leq \max_{\mathbf{w} \in \Delta_{SA}} \min_{\psi : \pi^*(\psi) \neq \pi^*(\phi)} \sum_{s,a} w_{s,a} \left\{ KL(p_\phi(\cdot|s,a)\|p_\psi(\cdot|s,a)) + KL(q_\phi(\cdot|s,a)\|q_\psi(\cdot|s,a)) \right\} + \ln \frac{1}{\delta}
$$
Can we have a single algorithm so that for all $\phi$,

$$
\mathbb{E}_\phi[\tau] \leq \max_{\psi \in \Delta_{SA}} \min_{\pi \neq \pi^*(\phi)} \sum_{s,a} w_{s,a} \left\{ \text{KL}(p_\phi(\cdot|s,a) \| p_\psi(\cdot|s,a)) + \text{KL}(q_\phi(\cdot|s,a) \| q_\psi(\cdot|s,a)) \right\} + \text{tiny}
$$

**Track and Stop**

For $t = 1, 2, \ldots$
- Form estimate $\hat{\phi}_t$ of MDP
- Compute oracle weights $w^*(\hat{\phi}_t)$.
- Track $w^*$ 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 $w^*(\hat{\phi}_t) \to w^*(\phi)$ and we stop at optimal time + tiny.
The instance dependent problem complexity of MDPs is (apparently)

\[
\frac{1}{\max_{\mathbf{w} \in \Delta_{SA}} \min_{\psi: \pi^*(\psi) \neq \pi^*(\phi)} \sum_{s, a} w_{s, a} \left\{ \text{KL} (p_{\phi} (\cdot | s, a) \| p_{\psi} (\cdot | s, a)) + \text{KL} (q_{\phi} (\cdot | s, a) \| q_{\psi} (\cdot | s, a)) \right\}}
\]

as lower and upper bounds match.

Algorithmics **not settled**.

Current algorithms target relaxations instead.
Thanks!