Adaptive Sampling for Best Policy Identification in MDPs

Aymen Al Marjani[1]
joint work with Alexandre Proutiere[2]

[1]ENS de Lyon

[2]KTH Royal Institute of Technology

38^{th} International Conference on Machine Learning

Outline

Motivation

How many samples does it take to learn an optimal policy in RL ?

Infinite horizon discounted MDPs

$\phi =< \mathcal{S}, \mathcal{A}, p_\phi, q_\phi, \gamma >$

1. $\mathcal{S}, \mathcal{A}$: **Finite** state and action spaces.

Infinite horizon discounted MDPs

$\phi =< \mathcal{S}, \mathcal{A}, p_\phi, q_\phi, \gamma >$

1. $\mathcal{S}, \mathcal{A}$: **Finite** state and action spaces.
2. After choosing action a at state s the agent:
    - receives reward $R(s, a) \sim q_\phi(.|s, a)$ and mean $r(s, a) \triangleq \mathbb{E}_{q(.|s,a)}[R(s, a)]$.
    - makes transition to $s' \sim p_\phi(.|s, a)$.

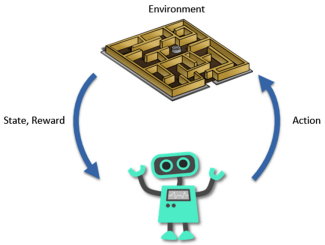


Figure: src:packtpub

Infinite horizon discounted MDPs

$\phi =< \mathcal{S}, \mathcal{A}, p_\phi, q_\phi, \gamma >$

1. $\mathcal{S}, \mathcal{A}$: **Finite** state and action spaces.
2. After choosing action a at state s the agent:
   - receives reward $R(s, a) \sim q_\phi(.|s, a)$ and mean $r(s, a) \triangleq \mathbb{E}_{q(.|s,a)}[R(s, a)]$.
   - makes transition to $s' \sim p_\phi(.|s, a)$.
   - For simplicity, we assume q with support in $[0, 1]$.

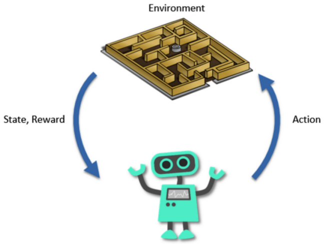


Figure: src:packtpub

Best Policy Identification

$\phi =< \mathcal{S}, \mathcal{A}, p_\phi, q_\phi, \gamma >$

- $\gamma \in [0, 1)$ is the discount factor.

Best Policy Identification

$\phi =< \mathcal{S}, \mathcal{A}, p_\phi, q_\phi, \gamma >$

- $\gamma \in [0, 1)$ is the discount factor.
- Identify a policy $\pi : \mathcal{S} \to \mathcal{A}$ maximizing the total discounted reward:

$$\pi^\star_\phi \in \arg\max_\pi \; V^\pi_\phi(s) = \mathbb{E}_\phi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t^\pi, \pi(s_t^\pi)) \middle| \; s_0 = s \right]$$

- **δ-PC algorithm:** $\mathbb{P}_\phi(\widehat{\pi}^\star_\tau \neq \pi^\star) \leq \delta$.

Best Policy Identification

$\phi =< \mathcal{S}, \mathcal{A}, p_\phi, q_\phi, \gamma >$

- $\gamma \in [0, 1)$ is the discount factor.
- Identify a policy $\pi : \mathcal{S} \to \mathcal{A}$ maximizing the total discounted reward:

$$\pi^\star_\phi \in \arg\max_\pi \; V^\pi_\phi(s) = \mathbb{E}_\phi\left[\sum_{t=0}^{\infty} \gamma^t R(s^\pi_t, \pi(s^\pi_t)) \middle| \; s_0 = s \right]$$

- δ-**PC algorithm:** $\mathbb{P}_\phi(\widehat{\pi}^\star_\tau \neq \pi^\star) \leq \delta$.
- Identify $\pi^\star$ using minimum number of samples!

Assumptions

- **Assumption 1:** $\pi^\star \triangleq \pi^\star_\phi$ is unique.

Assumptions

- **Assumption 1:** $\pi^\star \triangleq \pi^\star_\phi$ is unique.
- **Generative Model:** The agent has access to a simulator. At round t, she agent can query a sample *any* pair (s_t, a_t). She then observes $(R_t, s'_t) \sim q_\phi(.|s_t, a_t) \otimes p_\phi(.|s_t, a_t)$. Next, she can choose *any* other pair (s_{t+1}, a_{t+1}) *independently of her previous state*.

Learning: be specific!

Two kinds of guarantees:

- **Minimax** over a set of MDPs Φ:

$$\inf_{\mathbb{A}:\delta\text{-PC}} \sup_{\phi\in\Phi} \mathbb{E}_{\phi,\mathbb{A}}[\tau_\delta]$$

- **Instance-specific:** For a given ϕ:

$$\inf_{\mathbb{A}:\delta\text{-PC}} \mathbb{E}_{\phi,\mathbb{A}}[\tau_\delta]$$

Learning: be specific!

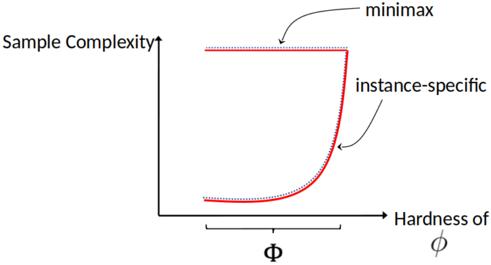
Two kinds of guarantees:

- **Minimax** over a set of MDPs Φ:

$$\inf_{\mathbb{A}:\delta\text{-PC}} \sup_{\phi\in\Phi} \mathbb{E}_{\phi,\mathbb{A}}[\tau_\delta]$$

- **Instance-specific:** For a given ϕ:

$$\inf_{\mathbb{A}:\delta\text{-PC}} \mathbb{E}_{\phi,\mathbb{A}}[\tau_\delta]$$



Learning: be specific!

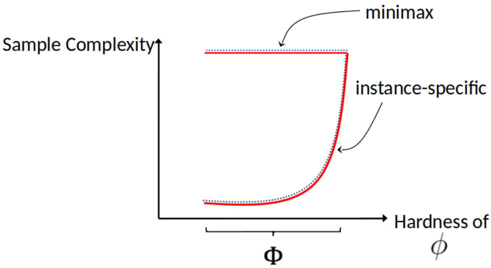
Two kinds of guarantees:

- **Minimax** over a set of MDPs Φ:

$$\inf_{\mathbb{A}:\delta\text{-PC}} \sup_{\phi\in\Phi} \mathbb{E}_{\phi,\mathbb{A}}[\tau_\delta]$$

- **Instance-specific:** For a given ϕ:

$$\inf_{\mathbb{A}:\delta\text{-PC}} \mathbb{E}_{\phi,\mathbb{A}}[\tau_\delta]$$



- We seek algorithms that can adapt to the hardness of the instance.

Information-Theoretical lower bound

Define:

- The set of alternative MDPs $\mathsf{Alt}(\phi) = \{\psi : \ \pi^\star \text{ is not optimal in } \psi\}$.
- Σ the simplex of $\mathbb{R}^{SA}$.
- $\mathrm{KL}_{\phi|\psi}(s,a) = \mathsf{KL}(q_\phi(s,a),\ q_\psi(s,a)) + \mathsf{KL}(p_\phi(s,a),\ p_\psi(s,a))$

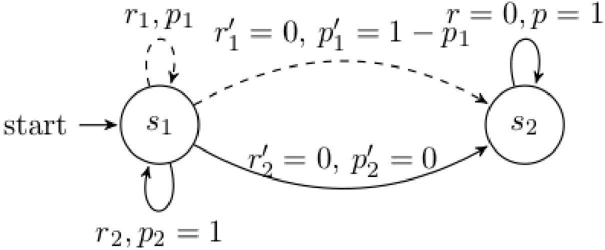
Proposition 1

The sample complexity of any δ-PC algorithm satisfies: for any ϕ with a unique optimal policy,

$$\mathbb{E}_\phi[\tau] \geq T^\star(\phi)\log(1/2.4\delta),$$

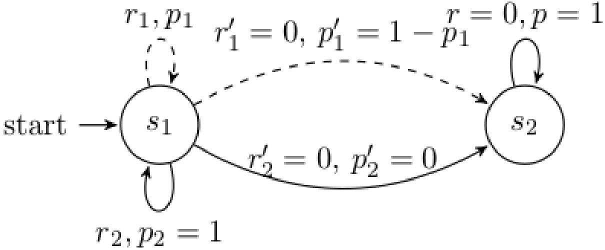
$$\text{where } T^\star(\phi)^{-1} = \sup_{\omega\in\Sigma}\ \inf_{\psi\in\mathsf{Alt}(\phi)} \sum_{s,a} \omega_{sa}\mathrm{KL}_{\phi|\psi}(s,a). \tag{1}$$

IT Lower bound: Hard to solve !



- Alt(ϕ) and $\mathrm{Alt}_{s_1 a_1}(\phi)$ are not convex.

IT Lower bound: Hard to solve !



- Alt(ϕ) and $\mathrm{Alt}_{s_1 a_1}(\phi)$ are not convex.
- $\Longrightarrow$ The sub-problem $\inf\limits_{\psi \in \mathrm{Alt}(\phi)} \sum\limits_{s,a} \omega_{sa} \mathrm{KL}_{\phi|\psi}(s, a)$ is non-convex.

IT Lower bound: MDP vs MAB

	MAB	**MDP**
Parameters	$\mu_1 > \ldots \geq \mu_K$	$\big(r(s,a), p(s,a)\big)_{s,a}$
Objective	Identify $a^\star = \underset{a \in [K]}{\arg\max}\ \mu_a$	Identify $\pi^\star = \underset{\pi}{\arg\max}\ (I - \gamma P_\pi)^{-1} r_\pi$
Alternative instances	$\bigcup_{a \neq 1} \{\lambda : \ \lambda_a > \lambda_1\}$ union of convex sets	$\bigcup_{s, a \neq \pi^\star(s)} \{\psi : \ Q_\psi^{\pi^\star}(s,a) > V_\psi^{\pi^\star}(s)\}$ Not union of convex sets
IT lower bound	Easy to solve	Hard to solve

Upper bound: Idea

Define the characteristic time: $T(\phi, \omega)^{-1} \triangleq \inf_{\psi \in \mathrm{Alt}(\phi)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{\phi|\psi}(s, a).$

Upper bound: Idea

Define the characteristic time: $T(\phi, \omega)^{-1} \triangleq \inf_{\psi \in \mathrm{Alt}(\phi)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{\phi|\psi}(s, a)$.

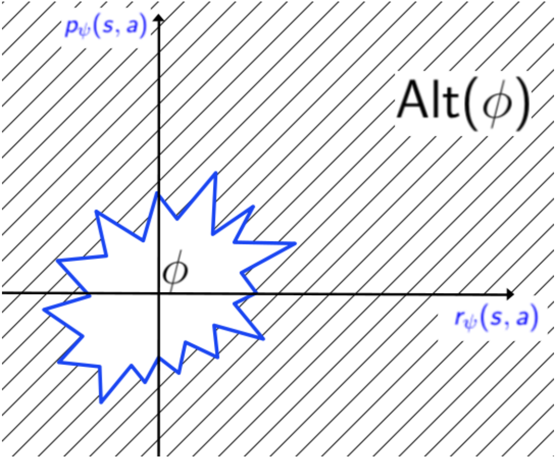


Figure: Alt(ϕ): Non-convex boundary

Upper bound: Idea

Define the characteristic time: $T(\phi, \omega)^{-1} \triangleq \inf_{\psi \in \mathrm{Alt}(\phi)} \sum_{s,a} \omega_{sa} \mathrm{KL}_{\phi|\psi}(s, a)$.

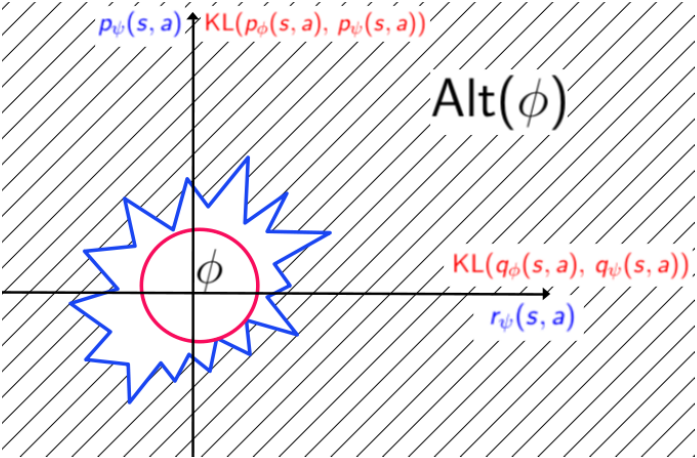


Figure: KL Ball

Upper bound: Instance-specific quantities

Define:

- The sub-optimality gap: $\Delta_{sa} = V_\phi^\star(s) - Q_\phi^\star(s,a)$.
- The minimum gap $\Delta_{\min} = \min_{s,a \neq \pi^\star(s)} \Delta_{sa}$.
- The variance of the value function $\mathrm{Var}_{(s,a)}[V_\phi^\star] = \mathbb{V}_{s' \sim p_\phi(.|s,a)}[V_\phi^\star(s)]$.
- The span of the value function $\mathrm{sp}[V_\phi^\star] = \max_s V_\phi^\star(s) - \min_s V_\phi^\star(s)$.

Upper bound of the characteristic time

Theorem 1 (Upper bound of minimal sample complexity)

For all vectors ω in the simplex:

$$T(\phi,\omega) \leq U(\phi,\omega) \triangleq \max_{s,a\neq\pi^\star(s)} \frac{T_1(s,a;\phi)+T_2(s,a;\phi)}{\omega_{sa}} + \frac{T_3(\phi)+T_4(\phi)}{\min\limits_{s} \omega_{s,\pi^\star(s)}},$$

$$\text{where} \begin{cases} T_1(s,a;\phi) = \dfrac{2}{\Delta_{sa}^2}, \\ T_2(s,a;\phi) = \max\left(\dfrac{16\mathrm{Var}_{(s,a)}[V_\phi^\star]}{\Delta_{sa}^2}, \dfrac{6\mathrm{sp}[V_\phi^\star]^{4/3}}{\Delta_{sa}^{4/3}}\right), \\ T_3(\phi) = \dfrac{2}{[\Delta_{\min}(\phi)(1-\gamma)]^2}, \\ T_4(\phi) \leq \dfrac{27}{\Delta_{\min}(\phi)^2(1-\gamma)^3} = \mathcal{O}\left(\dfrac{\text{Minimax lower bound}}{SA}\right) \end{cases}$$

KLB-TS: Sampling rule

- The optimal weights minimizing the upper-bound program:

$$\overline{\omega}(\phi) = \arg\inf_{\omega \in \Sigma} \max_{(s,a): a \neq \pi^\star(s)} \frac{T_1(s,a;\phi) + T_2(s,a;\phi)}{\omega_{sa}} + \frac{T_3(\phi) + T_4(\phi)}{\min_s \omega_{s,\pi^\star(s)}}$$

are easy to compute !

KLB-TS: Sampling rule

- The optimal weights minimizing the upper-bound program:

$$\overline{\omega}(\phi) = \arg\inf_{\omega \in \Sigma} \max_{(s,a): a \neq \pi^\star(s)} \frac{T_1(s,a;\phi) + T_2(s,a;\phi)}{\omega_{sa}} + \frac{T_3(\phi) + T_4(\phi)}{\min_s \omega_{s,\pi^\star(s)}}$$

  are easy to compute !

- Ensures that $\mathbb{P}_\phi \left(\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \lim_{t \to \infty} \frac{N_{sa}(t)}{t} = \overline{\omega}_{s,a}(\phi) \right) = 1.$

KLB-TS: stopping rule

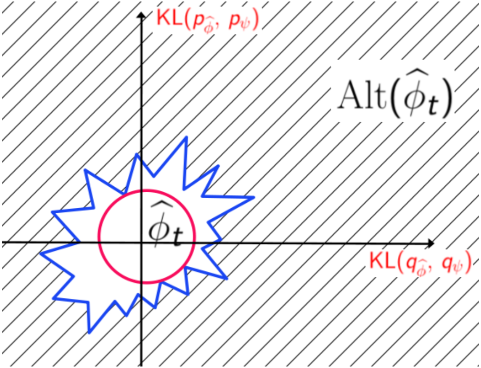


Figure: KL-Ball Stopping rule

KLB-TS: stopping rule

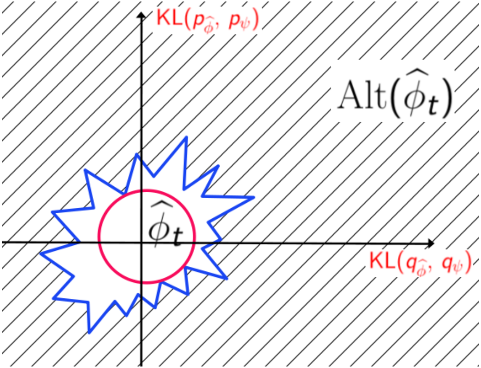


Figure: KL-Ball Stopping rule

- We ensure that ϕ falls within the KL-ball with probability $1-\delta$, using PAC bounds on the KL divergence..

Algorithm: Guarantees

Theorem 3

KLB-TS has a sample complexity τ_δ satisfying:
for all $\delta \in (0,1)$, $\mathbb{E}_\phi[\tau_\delta]$ is finite and $\limsup_{\delta \to 0} \frac{\mathbb{E}_\phi[\tau_\delta]}{\log(1/\delta)} \leq 4U(\phi)$, where:

$$
\begin{aligned}
U(\phi) &\triangleq \inf_{\omega} U(\phi, \omega) \\
&= \mathcal{O}\Bigg(S \min\left(\frac{\mathrm{Var}^\star_{\max}[V^\star_\phi]}{\Delta^2_{\min}(1-\gamma)^2}, \frac{1}{\Delta^2_{\min}(1-\gamma)^3} \right) \\
&\quad + \sum_{s, a \neq \pi^\star(s)} \frac{1 + \mathrm{Var}_{(s,a)}[V^\star_\phi]}{\Delta^2_{sa}} \Bigg)
\end{aligned}
$$

Experiments

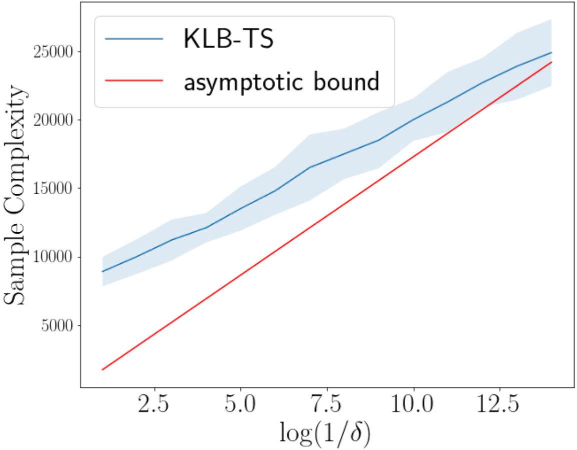


Figure: Asymptotic bound: S=A=2, $\gamma = 0.5$.

Experiments

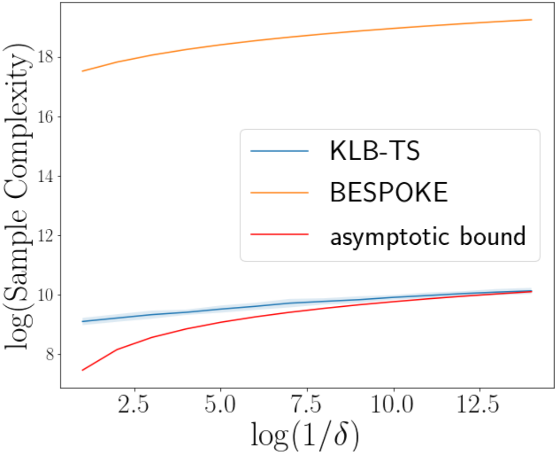


Figure: KLB-TS vs. BESPOKE. S=A=2, $\gamma = 0.5$.

Experiments

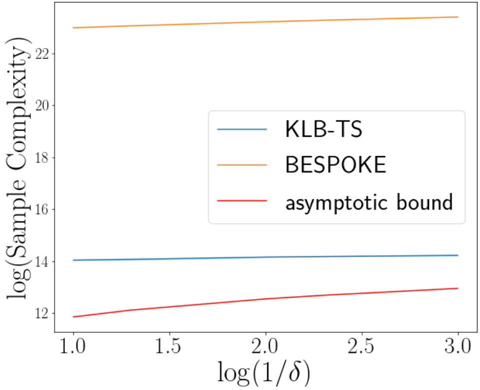


Figure: KLB-TS vs. BESPOKE. $S = 5, A = 10, \gamma = 0.7$.

Conclusion

1. Contrary to MAB, IT lower bound is hard to solve for MDPs.
2. We can derive problem-specific surrogates which :
    - Are *explicit*, depending on functionals of the MDP.
    - Have a corresponding allocation that is easy to compute.
3. Can be used to devise (Asymptocically) Matching algorithm.
4. First step towards understanding problem-specific ε-optimal policy identification.

Thanks !