Foundations of Reinforcement Learning

Model-based RL with simulators

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2023

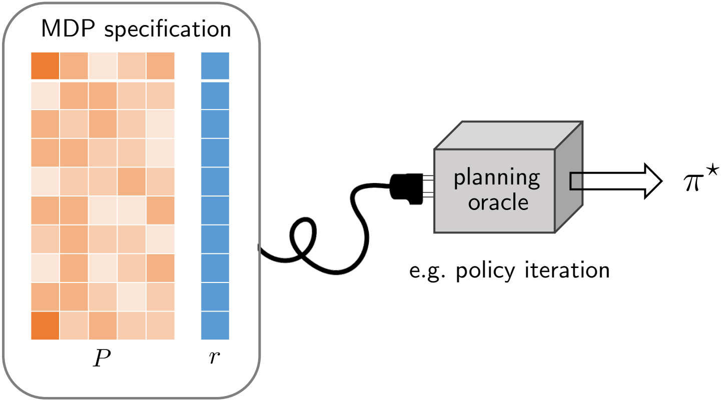
Outline

RL with a generative model

Model-based policy evaluation

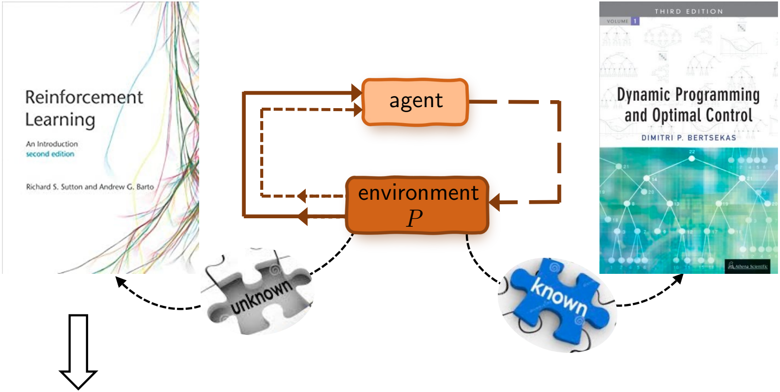
Model-based policy learning

Planning: when the MDP model is known



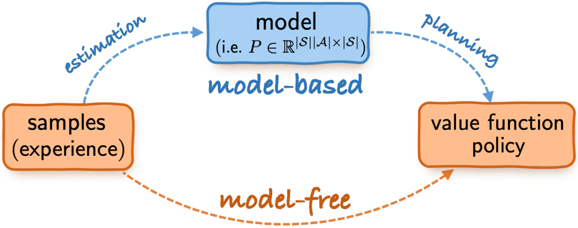
Planning: find the optimal policy $\pi^\star$ given MDP specification

Reinforcement learning (RL)



Learning: learn a desired policy from samples w/o model specification

Two approaches to RL



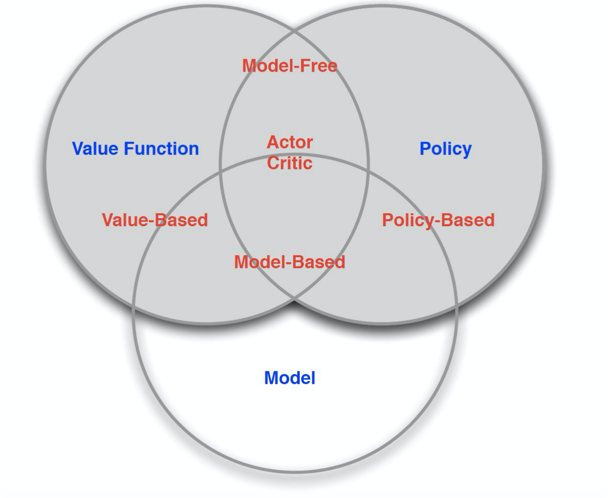
Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on empirical $\widehat{P}$

Model-free approach

— learning w/o constructing model explicitly

A taxonomy of RL approaches



—Credit: David Silver's slide

RL with a generative model

Motivation: study sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials

autonomous driving

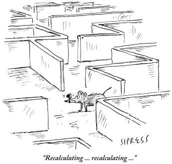
online ads

Understand and design of sample-efficient RL algorithms!

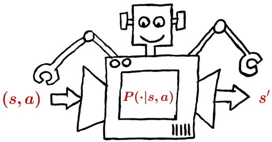
Data source in RL

Exploration

offline RL



online RL



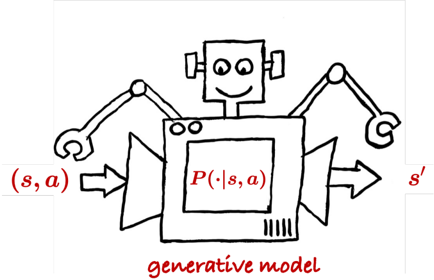
generative model

The capability of exploration increases from left to right.

This lecture: generative model / simulator

RL with a generative model / simulator

— [Kearns and Singh, 1999]

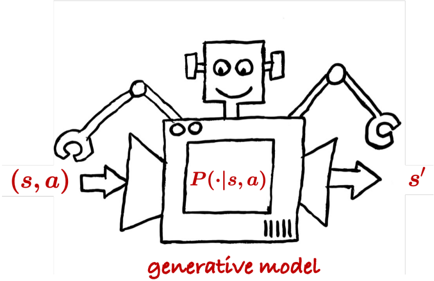


Protocol: for any state-action pair (s, a), we can probe the simulator to output the next state s'.

We focus on the transition kernel and assume the reward is known or fixed, since the transition kernel captures the harder aspect of the problem.

RL with a generative model / simulator

— [Kearns and Singh, 1999]

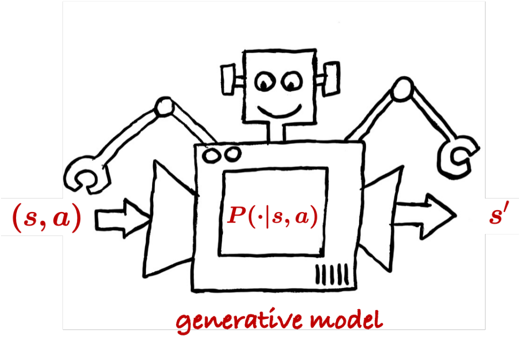


For each state-action pair (s,a), collect N samples

$$\{(s,a,s'_{(i)})\}_{1\le i\le N}$$

Question: How many samples are necessary and sufficient to solve the RL problem without worrying about exploration?

Model estimation under the generative model

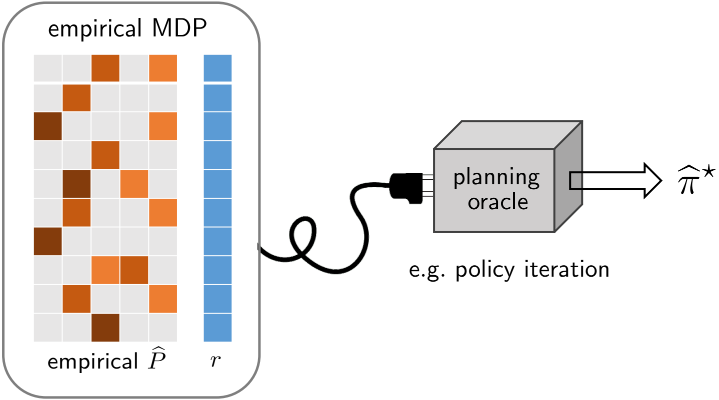


For each (s,a), collect N independent samples $\{(s,a,s'_{(i)})\}_{1\le i\le N}$

Empirical estimates: estimate $\widehat{P}(s'|s,a)$ by $\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{s'_{(i)}=s'\}}_{\text{empirical frequency}}$

Model-based (plug-in) estimator

— [Azar et al., 2013, Pananjady and Wainwright, 2020, Agarwal et al., 2020]



Run planning algorithms based on the *empirical* MDP

Questions

ℓ_∞-**sample complexity:** how many samples are required for

1) evaluate an $\underbrace{\varepsilon\text{-accurate policy}}_{\forall s:\ \left|\widehat{V}^{\pi}(s)-V^{\pi}(s)\right|\le\varepsilon}$?

2) learn an $\underbrace{\varepsilon\text{-optimal policy}}_{\forall s:\ V^{\hat{\pi}}(s)\ge V^{\star}(s)-\varepsilon}$?

Model-based policy evaluation

Minimax lower bound

Theorem 1 (minimax lower bound; [Pananjady and Wainwright, 2020])

Fix a policy π. For all $\varepsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

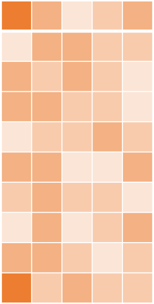
to achieve $\|\widehat{Q} - Q^{\pi}\|_{\infty} \leq \varepsilon$, where $\widehat{Q}$ is the output of any RL algorithm.

- Consider the relative accuracy $\varepsilon_{\mathsf{rel}}$ by setting $\varepsilon := \frac{\varepsilon_{\mathsf{rel}}}{1-\gamma}$, the lower bound can be equivalently expressed as

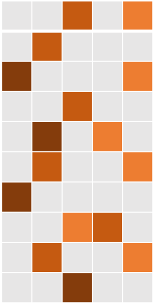
$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\varepsilon_{\mathsf{rel}}^2}\right).$$

- much smaller than the model dimension $|\mathcal{S}|^2|\mathcal{A}|$ — hint at the possibility of evaluating the policy without estimating the model reliably!

Challenges in the sample-starved regime



truth: $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$



empirical estimate: $\widehat{P}$

- Can't recover P faithfully if sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$!
- Can we trust our policy estimate when reliable model estimation is infeasible?

Recall: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}} \Big]$$

The value/Q function can be decomposed into two parts:

- immediate reward $\mathbb{E}\left[r(s,a)\right]$
- discounted value of at the successor state $\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V(s')$

Richard Bellman

Policy evaluation for state-action function

Matrix-vector representation:

$$Q^\pi(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')} [Q^\pi(s',a')]$$

$$\Updownarrow$$

$$Q^\pi = r + \gamma P^\pi Q^\pi$$

$$\Updownarrow$$

$$Q^\pi = (I - \gamma P^\pi)^{-1} r$$

- Here, P^π is the state-action transition matrix induced by π, namely,

$$P^\pi(s',a'|s,a) = P(s'|s,a)\pi(a'|s').$$

Sample complexity for plug-in policy evaluation

Model-based plug-in estimate:

$$\widehat{Q}^{\pi} = (I - \gamma \widehat{P}^{\pi})^{-1} r$$

Theorem 2 ([Pananjady and Wainwright, 2020])

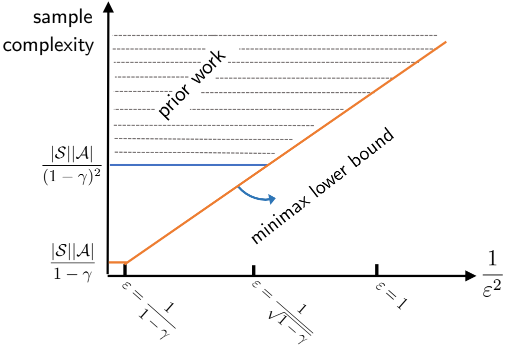
Fix any policy π. *For* $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$, *the plug-in estimator* $\widehat{Q}^{\pi}$ *obeys*

$$\|\widehat{Q}^{\pi} - Q^{\pi}\|_{\infty} \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \varepsilon^2}\Big).$$

A sample size barrier



- A sample size barrier $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$ appeared in prior works [Azar et al., 2013, Pananjady and Wainwright, 2020]

Refined analysis

Model-based plug-in estimate:

$$\widehat{Q}^{\pi} = (I - \gamma \widehat{P}^{\pi})^{-1} r$$

Theorem 3 ([Li et al., 2020])

Fix any policy π. *For* $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the plug-in estimator* $\widehat{Q}^{\pi}$ *obeys*

$$\|\widehat{Q}^{\pi} - Q^{\pi}\|_{\infty} \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \varepsilon^2}\Big)$$

- Minimax optimal for all ε [Azar et al., 2013, Pananjady and Wainwright, 2020].

Analysis: crude idea

- We'll demonstrate a crude version based on Hoeffding's inequality.

$$Q^\pi = (I - \gamma P^\pi)^{-1} r, \qquad \widehat{Q}^\pi = (I - \gamma \widehat{P}^\pi)^{-1} r$$

Useful expansion:

$$\begin{aligned}
\widehat{Q}^\pi - Q^\pi &= (I - \gamma \widehat{P}^\pi)^{-1} r - (I - \gamma P^\pi)^{-1} r \\
&= (I - \gamma \widehat{P}^\pi)^{-1} \Big((I - \gamma P^\pi) - (I - \gamma \widehat{P}^\pi) \Big) Q^\pi \\
&= \gamma \big(I - \gamma \widehat{P}^\pi \big)^{-1} \big(\widehat{P}^\pi - P^\pi \big) Q^\pi \\
&= \gamma \big(I - \gamma \widehat{P}^\pi \big)^{-1} \big(\widehat{P} - P \big) V^\pi .
\end{aligned}$$

Analysis: Hoeffding's inequality

By Hoeffding's inequality and union bound, with probability at least $1-\delta$,

$$\left\|\big(\widehat{P}-P\big)V^{\pi}\right\|_{\infty} \le \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N(1-\gamma)^2}}.$$

Then, using $\big(I-\gamma\widehat{P}^{\pi}\big)^{-1} = \sum_{i=0}^{\infty}\gamma^i\big(\widehat{P}^{\pi}\big)^i$,

$$\begin{aligned}
\left\|\gamma\big(I-\gamma\widehat{P}^{\pi}\big)^{-1}\big(\widehat{P}-P\big)V^{\pi}\right\|_{\infty} &\le \gamma\sum_{i=0}^{\infty}\gamma^i\left\|\big(\widehat{P}^{\pi}\big)^i\big(\widehat{P}-P\big)V^{\pi}\right\|_{\infty} \\
&\le \gamma\sum_{i=0}^{\infty}\gamma^i\left\|\big(\widehat{P}-P\big)V^{\pi}\right\|_{\infty} \\
&\le \frac{\gamma}{1-\gamma}\left\|\big(\widehat{P}-P\big)V^{\pi}\right\|_{\infty} \le \gamma\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N(1-\gamma)^4}}.
\end{aligned}$$

Analysis: variance control + a peeling argument

- Better concentration with variance control: Bernstein's inequality
- Going beyond the 1st-order error expansion

$$\widehat{Q}^{\pi} - Q^{\pi} = \gamma \big(I - \gamma P^{\pi}\big)^{-1} \big(\widehat{P}^{\pi} - P^{\pi}\big) \widehat{Q}^{\pi}$$

Instead: higher-order expansion $\longrightarrow$ tighter control

$$\begin{aligned} \widehat{Q}^{\pi} - Q^{\pi} = \gamma \big(I - \gamma P^{\pi}\big)^{-1} \big(\widehat{P}^{\pi} - P^{\pi}\big) {\color{red}Q^{\pi}} + \\ + \gamma^2 \Big(\big(I - \gamma P^{\pi}\big)^{-1} \big(\widehat{P}^{\pi} - P^{\pi}\big) \Big)^2 {\color{red}Q^{\pi}} \\ + \gamma^3 \Big(\big(I - \gamma P^{\pi}\big)^{-1} \big(\widehat{P}^{\pi} - P^{\pi}\big) \Big)^3 {\color{red}Q^{\pi}} \\ + \dots \end{aligned}$$

Model-based policy learning

Minimax lower bound

Theorem 4 (minimax lower bound; [Azar et al., 2013])

For all $\varepsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, where $\widehat{Q}$ is the output of any RL algorithm.

- holds for both value-based and policy-based algorithms.
- much smaller than the model dimension $|\mathcal{S}|^2|\mathcal{A}|$ — hint at the possibility of solving RL without estimating the model reliably!

Sample complexity for learning $Q^\star$

Theorem 5 ([Azar et al., 2013])

For any $0 < \varepsilon \leq 1$*, the optimal Q-function* $\widehat{Q}$ *of the empirical MDP achieves*

$$\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$$

with sample complexity at most $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$.

- matches with the minimax lower bound whenever $\varepsilon \in (0, 1]$.
- **Question:** Does it imply a near minimax-optimal policy $\widehat{\pi}$?

From Q-function to policy

Lemma 6 ([Singh and Yee, 1994])

Let the greedy policy w.r.t. $\widehat{Q}$ *be* $\widehat{\pi}$*, then*

$$V^\star - V^{\widehat{\pi}} \le \frac{2}{1-\gamma}\|Q^\star - \widehat{Q}\|_\infty.$$

$$\boxed{\|\widehat{Q} - Q^\star\|_\infty \le \epsilon} \xrightarrow{\widehat{\pi} = \mathrm{Greedy}(\widehat{Q})} \boxed{V^\star - V^{\widehat{\pi}} \le \frac{\epsilon}{1-\gamma}}$$

This **error amplification** has consequences in sample complexities.

- To reach ϵ-optimality, the greedy policy of a minimax-optimal Q-function estimator needs

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}\right)$$

samples invoking the above naive argument. Need refined arguments!

Theory of model-based policy learning

Theorem 7 ([Agarwal et al., 2020])

For any $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$, *the optimal policy* $\widehat{\pi}^\star$ *of the empirical MDP achieves*

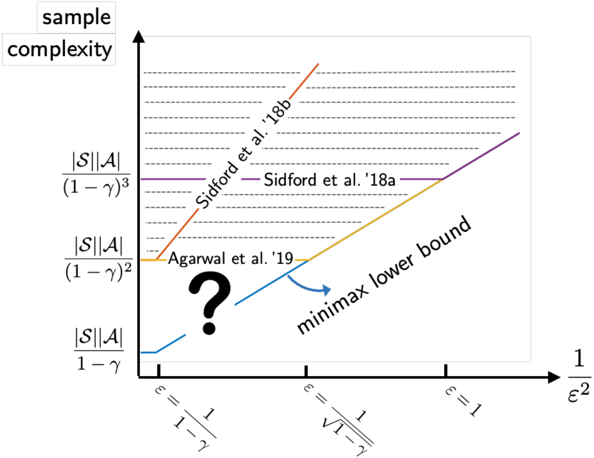
$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- Matches with the lower bound $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ [Azar et al., 2013] when $\varepsilon \in (0, \frac{1}{\sqrt{1-\gamma}}]$.

A sample complexity barrier

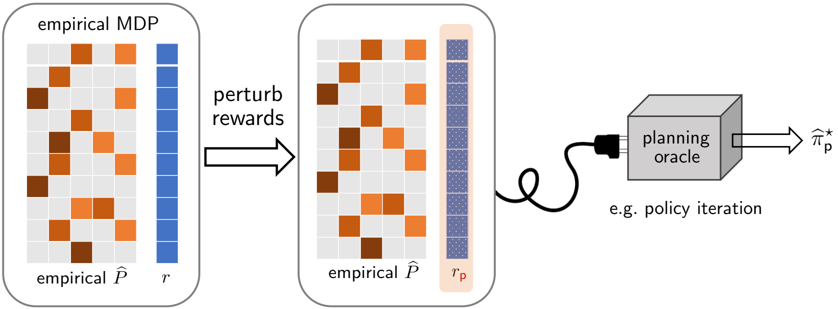


All prior theory requires $\underbrace{\text{sample size} > \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}}_{\text{sample size barrier}}$

Is it possible to close the gap?

Model-based plug-in estimator + perturbation

— *[Li et al., 2020]*



Planning based on the *empirical* MDP with *slightly perturbed rewards*

Refined theory of model-based policy learning

Theorem 8 ([Li et al., 2020])

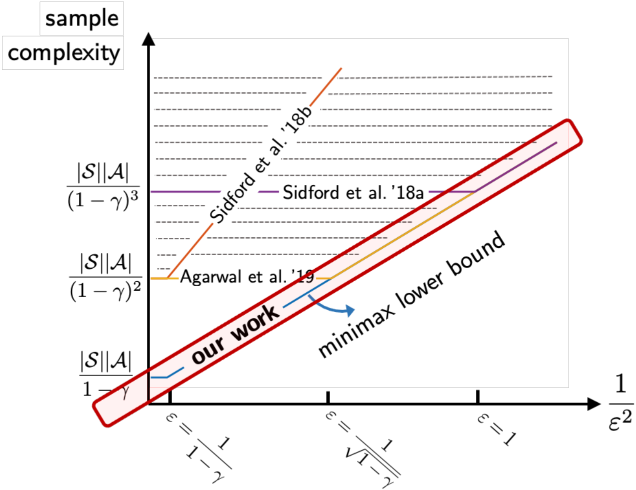
For any $0 < \varepsilon \le \frac{1}{1-\gamma}$*, the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of the perturbed empirical MDP achieves*

$$\|V^{\widehat{\pi}_{\mathrm{p}}^{\star}} - V^{\star}\|_{\infty} \le \varepsilon$$

with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- $\widehat{\pi}_{\mathrm{p}}^{\star}$: obtained by empirical VI or PI within $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$ iterations
- **Minimax lower bound:** $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ [Azar et al., 2013]



Model-based RL is nearly minimax optimal!

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V^\pi = (I - \gamma P_\pi)^{-1} r$
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \gamma \widehat{P}_\pi)^{-1} r$
- $\pi^\star$: optimal policy w.r.t. true value function
- $\widehat{\pi}^\star$: optimal policy w.r.t. empirical value function
- $V^\star := V^{\pi^\star}$: optimal values under true models
- $\widehat{V}^\star := \widehat{V}^{\widehat{\pi}^\star}$: optimal values under empirical models

Proof ideas

Elementary decomposition:

$$V^\star - V^{\widehat{\pi}^\star} = \left(V^\star - \widehat{V}^{\pi^\star}\right) + \left(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\right) + \left(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\right)$$
$$\leq \left(V^{\pi^\star} - \widehat{V}^{\pi^\star}\right) + 0 + \left(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\right)$$

- **Step 1:** control $V^\pi - \widehat{V}^\pi$ for a fixed π
  (**Bernstein inequality** + **high-order decomposition**)
- **Step 2:** extend it to control $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$ ($\widehat{\pi}^\star$ depends on samples)
  (**decouple statistical dependency via leave-one-out analysis and reward perturbation**)

References I

Agarwal, A., Kakade, S., and Yang, L. F. (2020).
Model-based reinforcement learning with a generative model is minimax optimal.
Conference on Learning Theory, pages 67–83.

Azar, M. G., Munos, R., and Kappen, H. J. (2013).
Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model.
Machine learning, 91(3):325–349.

Kearns, M. J. and Singh, S. P. (1999).
Finite-sample convergence rates for Q-learning and indirect algorithms.
In *Advances in neural information processing systems*, pages 996–1002.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2020).
Breaking the sample size barrier in model-based reinforcement learning with a generative model.
arXiv preprint arXiv:2005.12900.

Pananjady, A. and Wainwright, M. J. (2020).
Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning.
IEEE Transactions on Information Theory, 67(1):566–585.

Singh, S. P. and Yee, R. C. (1994).
An upper bound on the loss from approximate optimal-value functions.
Machine Learning, 16:227–233.