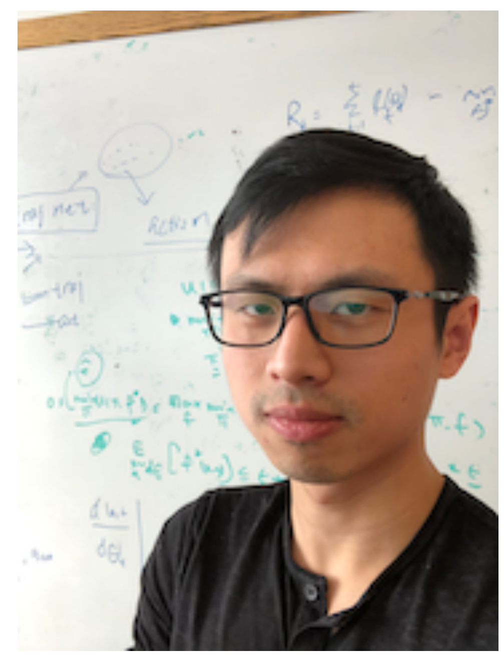
Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage

10/25, 2021 @RL Theory Seminar

Masatoshi Uehara, Wen Sun

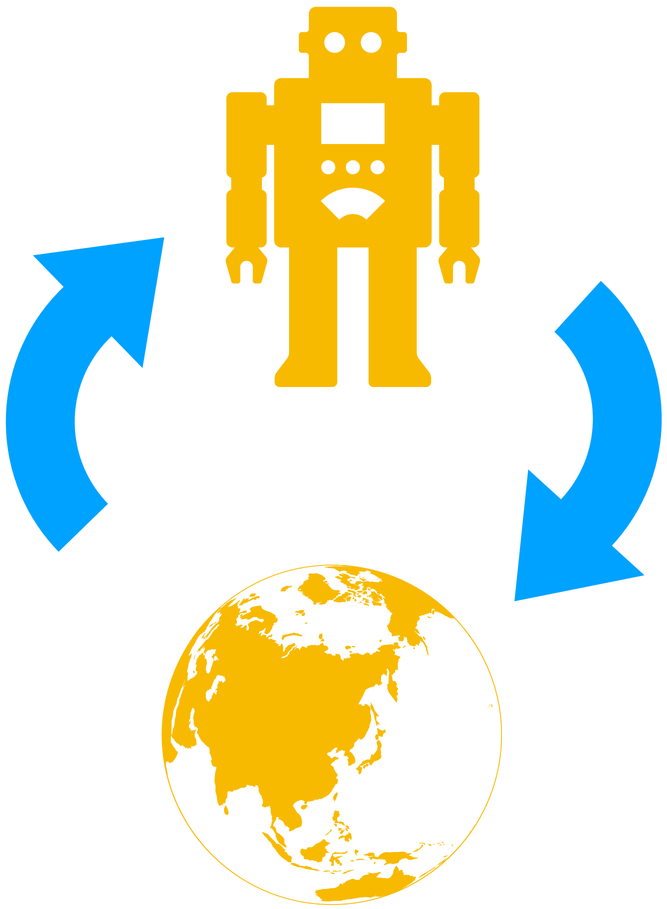


Cornell University
Department of Computer Science

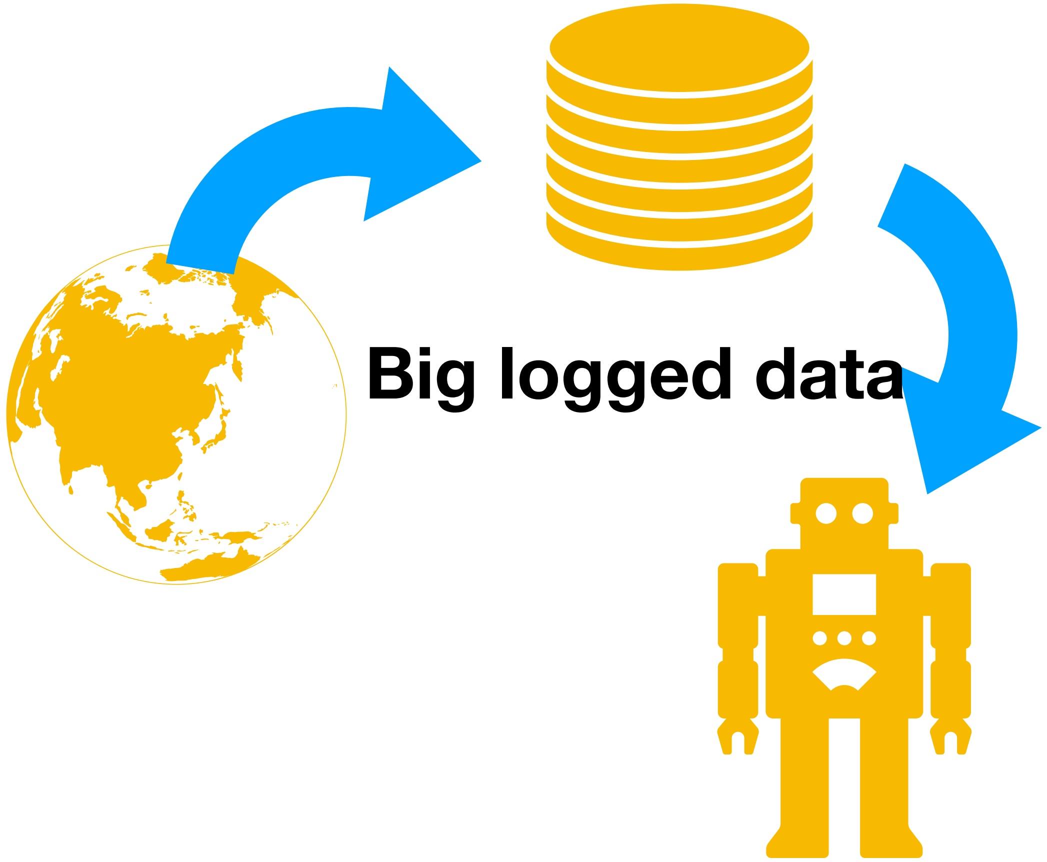
- 1: Overview
- 2: Preliminary
- 3: Pessimistic Model-based Offline RL
- 4: Examples with Refined Concentrability Coefficients

Offline RL

Online RL



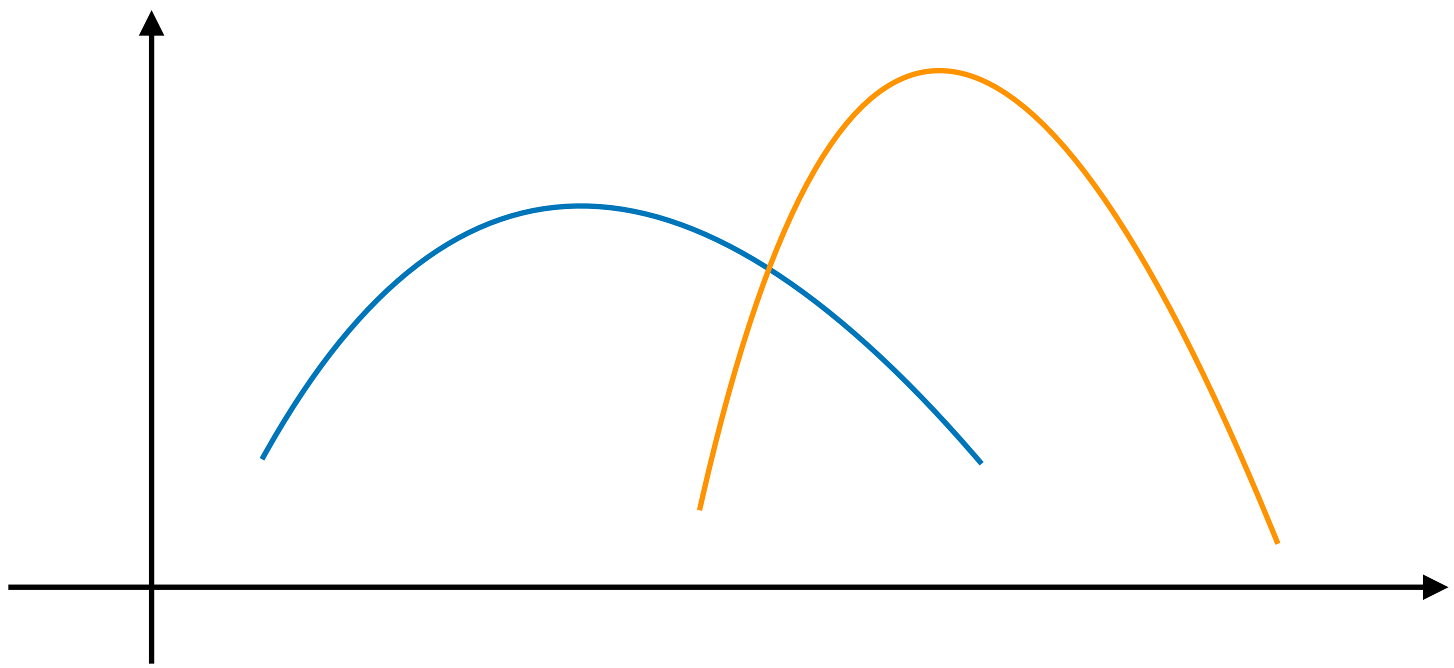
Offline RL



- We only have access to logged data.
- We want to learn high-quality polices from the logged data.

Question

- Unfortunately, the offline data is often not exploratory.
- Q. Can we still learn good policies when the offline data is not fully exploratory? (with realizability of the model)



Offline data: $\rho(s, a)$.

Distribution induced by a policy π**,** $d^{\pi}(s, a)$

Global vs. Partial Coverage

- Most of offline RL works assume global coverage. Under $\max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} < \infty \; \forall \pi$, they show the learned policy $\hat{\pi}$ can compete with the global optimal policy [MS, 2008].

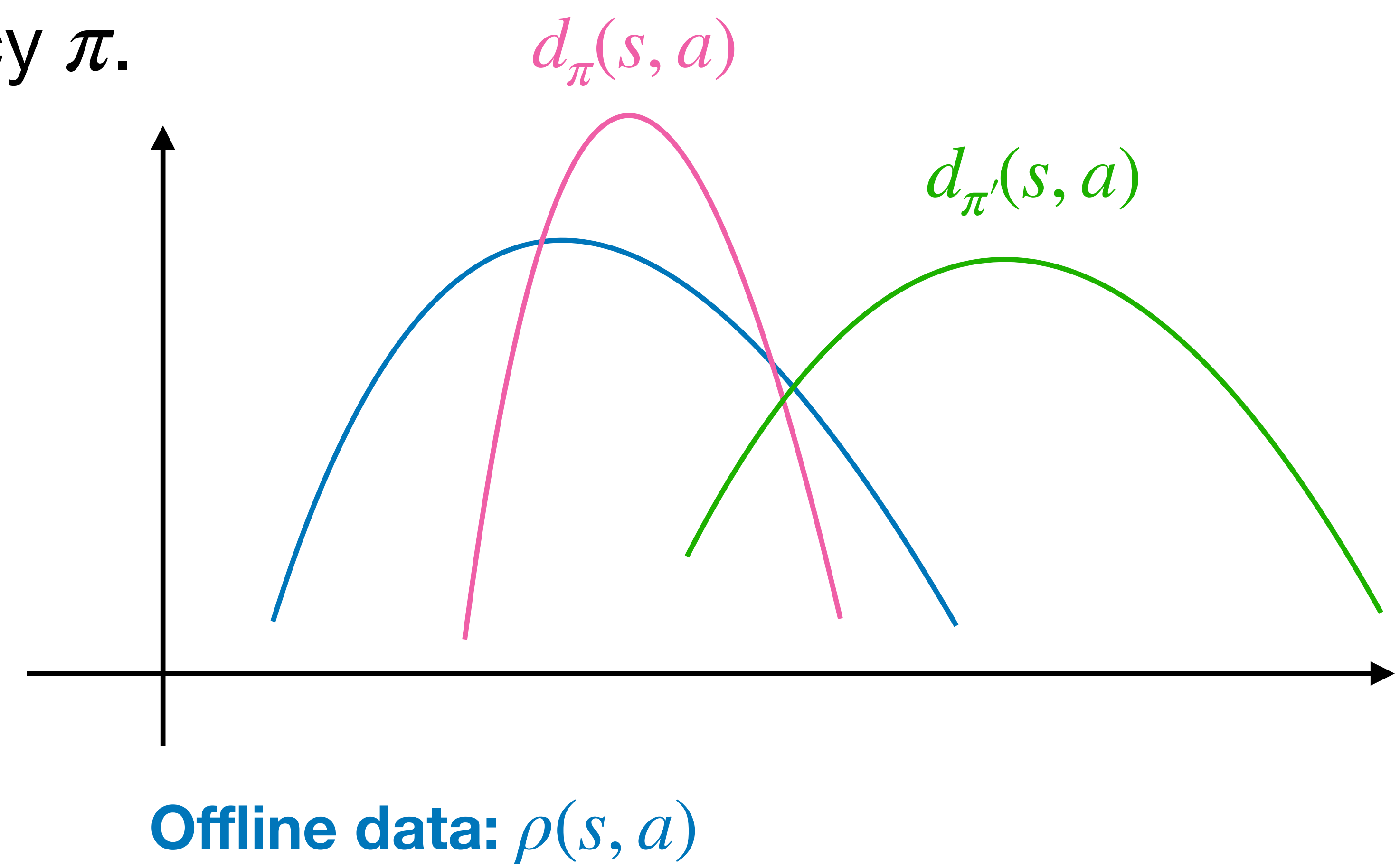
*$V^{\pi(P^{\star})} - V^{\hat{\pi}} = Small$. ($\pi(P^{\star})$ is the optimal policy. V^{π} is the policy value of π.)

- In this work, we want to show results under partial coverage. We want to show the output policy can compete with any polices $\pi \; \mathrm{s.t.} \max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} < \infty$.

*$V^{\pi} - V^{\hat{\pi}} = Small$ for any π covered by offline data.

Global vs. Partial Coverage

- Global coverage is not satisfied in the following. (π' is not covered by offline data)
- But, under partial coverage, we can still compete with a policy π.



What We Know So Far💭

- There are many works under global coverage [MS, 2008].
- In particular (linear) models, there exists a model-based algorithm under partial coverage [CUSKS21]. **But not for any models!**
- Several papers under partial coverage in the model-free setting [RZMIR21, JYW21, ZCZS21,XCJMA21,ZWB21], which assume completeness as well as realizability.

What We Show

- We propose a model-based offline RL algorithm CPPO. We show the PAC guarantee under partial coverage assuming the realizability of the model.

- This works for any MDPs 😄.

- When we have more structures, the density-ratio based partial coverage concept is refined.

  - Examples: linear mixture MDPs, KNRs, low-rank MDPs (models with unknown features), factored MDPs.

- 1: Overview
- 2: Preliminary
- 3: Pessimistic Model-based Offline RL
- 4: Examples with Refined Concentrability Coefficients

Notation

- MDP: $\langle \mathcal{S}, \mathcal{A}, r, P, \gamma, d_0 \rangle$. Discount factor $\gamma \in [0,1)$, $\mathcal{S}$: State space, $\mathcal{A}$: Action space.

Transition Dynamics	**Reward function**	**Initial distribution**
$P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$	$r : \mathcal{S} \times \mathcal{A} \to [0,1]$	$d_0 \in \Delta(\mathcal{S})$

- We have an offline dataset: $\mathcal{D} = \{s^{(i)}, a^{(i)}, s^{'(i)}\}_{i=1}^{n}$ following $(s, a) \sim \rho, s' \sim P^\star(s, a)$. ($P^\star$ is the true unknown transition density)

- $d^\pi = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t d_t^\pi$ is a state action discounted occupancy distribution under π and $P^\star$.

- V_P^π is an expected cumulative reward of π under P: $\mathrm{E}[\sum_{h=0}^{\infty} \gamma^h r_h \mid s_0 \sim d_0, a_0 \sim \pi(s_0), s_1 \sim P(s_0, a_0), \cdots]$.

Function Classes We Use

- We need two function classes:
  - Model class M ($\subset \{\mathscr{S} \times \mathscr{A} \to \Delta(\mathscr{S})\}$) to learn the true transition $P^\star$.
  - Policy class Π ($\subset \{\mathscr{S} \to \Delta(\mathscr{A})\}$). Throughout this presentation, this is the unrestricted policy class.

Model-based RL

Step 1: MLE. $\hat{P}_{\mathrm{MLE}} = \mathrm{argmax}_{P \in M} \sum_{i=1}^{n} \log P(s'^{(i)} \mid s^{(i)}, a^{(i)})$.

Step 2: Policy Optimization. $\hat{\pi} = \mathrm{argmax}_{\pi \in \Pi} V^{\pi}_{\hat{P}_{\mathrm{MLE}}}$.

- Under global coverage 😓 ($\max_{s,a} \frac{d^{\pi}(s,a)}{\rho(s,a)} \leq C, \forall \pi$) , the output can compete with the global optimal policy $\pi(P^\star)$ with $1-\delta$:
$V^{\pi(P^\star)}_{P^\star} - V^{\hat{\pi}}_{P^\star} = O((1-\gamma)^{-2}\sqrt{C \ln(|M|/\delta)/n})$.

- 1: Overview
- 2: Preliminary
- 3: Pessimistic Model-based Offline RL
- 4: Examples with Refined Concentrability Coefficients

Algorithm

CPPO: Constrained Pessimistic Policy Optimization

Step 1: MLE. $\hat{P}_{\mathrm{MLE}} = \mathrm{argmax}_{P \in M} \sum_{i=1}^{n} \log P(s^{\prime(i)} \mid s^{(i)}, a^{(i)})$.

Step 2: Solve constrained Optimization.

$\hat{\pi} = \mathrm{argmax}_{\pi \in \Pi} \min_{P \in M_{\mathscr{D}}} V_P^{\pi}$ **where**

$$M_{\mathscr{D}} = \left\{ P \mid P \in M, \frac{1}{n} \sum_{i=1}^{n} \| \hat{P}_{\mathrm{MLE}}(\cdot \mid s^{(i)}, a^{(i)}) - P(\cdot \mid s^{(i)}, a^{(i)}) \|_1^2 \leq \xi \right\}.$$

*** ξ is a hyperparamter**

- Search for the least favorable model in terms of V_P^{π} that is feasible w.r.t the constraint.
- Why? Pessimistic principle (being conservative on uncovered regions) is employed.

Model-based Concentrability Coefficient

[Definition] Model-based concentrability coefficient:

$$C_\pi^\dagger = \sup_{P' \in M} \frac{\mathrm{E}_{(s,a)\sim d^\pi}[\|P'(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1^2]}{\mathrm{E}_{(s,a)\sim \rho}[\|(P'(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1^2]}.$$

- Smaller than the density ratio: $C_\pi^\dagger \leq \max_{s,a} d^\pi(s,a)/\rho(s,a)$.
- Adaptive to model classes. If the model class is small, $C_\pi^\dagger$ is small either.

Guarantee of CPPO

[PAC Bound for CPPO] Suppose $P^\star \in M$. (by choosing ξ properly)
With probability $1-\delta$,

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = O\left((1-\gamma)^{-2} \sqrt{C_{\pi^*}^{\dagger} \ln(|M|/\delta)/n} \right).$$

- The output can simultaneously compete with any comparator polices satisfying partial coverage $C_{\pi^*}^{\dagger} < \infty$.
- Even if π^* is the optimal policy $\pi(P^\star)$, $C_{\pi(P^\star)}^{\dagger} < \infty$ is still weaker than the global coverage ($\max_{s,a} d^\pi(s,a)/\rho(s,a) < \infty, \forall \pi$) .
- When $|\mathscr{M}|$ is infinite, we can still use localized Rademacher complexities.

Derivation

- Define $\hat{V}^{\pi} = \min_{P \in M_D} V^{\pi}_{P}$. then, $\hat{\pi} = \mathrm{argmax}_{\pi} \hat{V}^{\pi}$.
- We can show $P^{\star} \in M_D$ in high probability.
- We have $\hat{V}^{\pi} \leq V^{\pi}_{P^{\star}}, \forall \pi \in \Pi$ (Pessimism).
- $$V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} = V^{\pi^*}_{P^\star} - \hat{V}^{\pi^*} + \hat{V}^{\pi^*} - V^{\hat{\pi}}_{P^\star} \leq V^{\pi^*}_{P^\star} - \hat{V}^{\pi^*} + \hat{V}^{\hat{\pi}} - V^{\hat{\pi}}_{P^\star} \leq V^{\pi^*}_{P^\star} - \hat{V}^{\pi^*}.$$

  Definition of $\hat{\pi}$. **Pessimism.**

- Finally, use performance difference lemma. Done 👍

Model free vs. Model-based

- The error in CPPO does not include $|\Pi|$. As a result, the policy class Π can be unrestricted. More strongly, we can compete with any history dependent policies.

- [XCJMA21] shows the PAC guarantee under partial coverage, realizability and Bellman completeness of Q-function class for any policy in Π, i.e. , $\mathcal{T}^{\pi}\mathcal{Q} \subset \mathcal{Q}$.

  *** $\mathcal{T}^{\pi}$ is the Bellman operator for a policy π.**

  - Thus, Π needs to be generally restricted .
  - It cannot compete with history dependent policies.

Comparison to Existing Pessimistic Algorithms

- CPPO use the MLE guarantee:
$\mathrm{E}_{(s,a)\sim\rho}[\|\hat{P}_{\mathrm{MLE}}(\cdot \mid s,a) - P^{\star}(\cdot \mid s,a)\|_1^2] \lesssim \sqrt{\ln|M|/\delta)/n}\,.$

- For linear models, [CUSKS21, JYW21] (existing offline RL papers using negative bonus terms) use
$\mathrm{Distance}(\hat{P}(\cdot \mid s,a), P^{\star}(\cdot \mid s,a))^2 \lesssim \mathrm{Poly}(1/n, \ln(1/\delta), \cdots), \forall (s,a)\,.$

- Average error (over offline data) guarantees are weaker than pointwise error guarantees 😩

- But average error guarantees are enough for the pessimism and obtained for any nonlinear models 😆

- **1: Overview**
- **2: Preliminary**
- **3: Pessimistic Model-based Offline RL**
- **4: Examples with Refined Concentrability Coefficients**

Next Questions

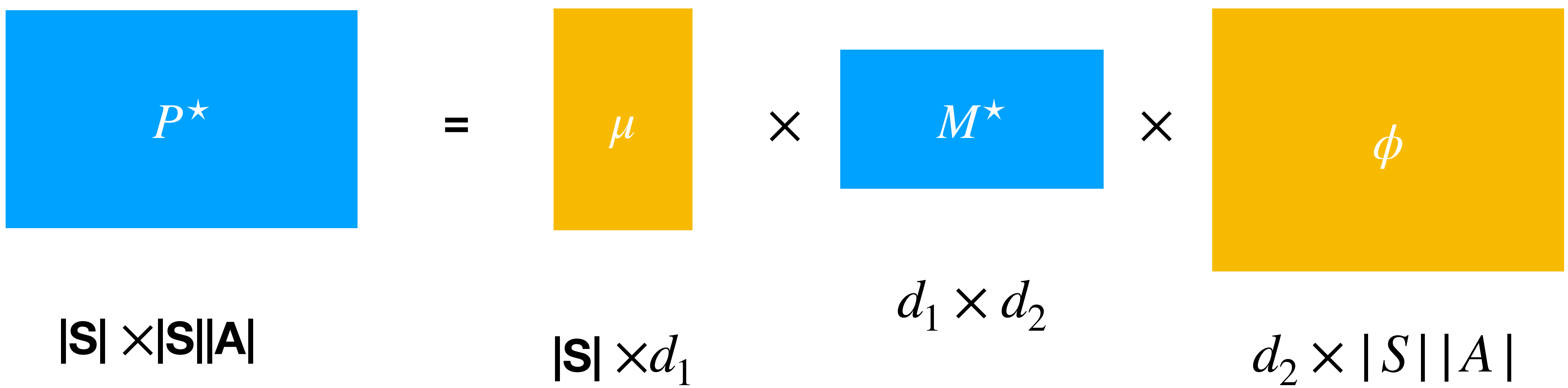
- $C^{\dagger}_{\pi^*}$ is very abstract. Can we replace it with more interpretable quantities? (and tighter than the density ratio.)

- To see them, we analyze on four models:
  - Linear mixture MDPs (including linear MDPs),
  - KNRs (generalization of LQRs),
  - Low-rank MDPs (with unknown features),
  - Factored MDPs.

1:Linear MDPs

Definition: Linear MDPs [YW20]

The true $P^\star$ is $\mu^\top(s')M^\star\phi(s,a)$ (Unknown $M^\star \in \mathbb{R}^{d_1 \times d_2}$) given feature vectors $\phi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_2}$, $\mu(s) : \mathcal{S} \to \mathbb{R}^{d_1}$.



1:Linear MDPs

Definition: Linear MDPs [YW20]

The true $P^\star$ is $\mu^\top(s')M^\star\phi(s,a)$ (Unknown $M^\star \in \mathbb{R}^{d_1\times d_2}$) given feature vectors $\phi(s,a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{d_2}$, $\mu(s) : \mathcal{S} \to \mathbb{R}^{d_1}$.

[Concentrability Coefficient for Linear MDPs]

$$\bar{C}_{\pi^*} = \sup_{x\in\mathbb{R}^d} \frac{x^\top \mathrm{E}_{(s,a)\sim d^{\pi^\star}}[\phi(s,a)\phi(s,a)^\top]x}{x^\top \mathrm{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top]x}.$$

- Smaller than the density ratio, i.e., $\bar{C}_{\pi^*} \le \max_{s,a} d^{\pi^*}(s,a)/\rho(s,a)$.
- If $\bar{C}_{\pi^*}$ is small, this implies the offline data sufficiently covers the subspace that the comparator policy π^* visits measured by $\phi(s,a)$.
- In tabular MDPs, $\bar{C}_{\pi^*} = \max_{s,a} d^{\pi^*}(s,a)/\rho(s,a)$.

1:Linear MDPs

[PAC Bound for CPPO] Suppose $P^\star \in M$. With probability $1-\delta$,
$\forall \pi^*; V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} = \tilde{O}((1-\gamma)^{-2}\sqrt{\bar{C}_{\pi^*} d^2 \ln(1/\delta)/n})$.

- Partial coverage is refined as $\bar{C}_{\pi^*} < \infty$.
- $\mathrm{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top]$ can be singular. (Previous works assume the non-singularity.)

1: Linear Mixture MDPs

Definition: Linear Mixture MDPs [AJSWY20,MJTS 20]

The true $P^\star$ is ${\theta^\star}^\top \psi(s,a,s')$ given a feature vector $\psi(s,a,s') : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.

- Linear MDPs belong to linear mixture MDPs.

- Define pseudo feature vectors: $\psi_V(s,a) = \int \psi(s,a,s') V(s') \mathrm{d}(s')$

[Concentrability Coefficient for Linear Mixture MDPs]

$$\bar{C}_{\pi^*,\mathrm{mix}} = \sup_{P \in Z_{P^\star}} \sup_{x \in \mathbb{R}^d} \frac{x^\top \mathrm{E}_{(s,a)\sim d^{\pi^\star}}[\psi_{V_P^{\pi^*}}(s,a)\psi_{V_P^{\pi^*}}(s,a)^\top] x}{x^\top \mathrm{E}_{(s,a)\sim \rho}[\psi_{V_P^{\pi^*}}(s,a)\psi_{V_P^{\pi^*}}(s,a)^\top] x},$$

where $Z_{P^\star} = \{P : \mathrm{E}_{(s,a)\sim\rho}[TV(P(\cdot \mid s,a), P^\star(\cdot \mid s,a)^2] \leq \xi\}$.

- $\bar{C}_{\pi^*,\mathrm{mix}}$ is defined for varying feature vectors $\psi_{V_P^{\pi^*}}$.

- In linear MDPs, $\bar{C}_{\pi^*,\mathrm{mix}}$ reduces to $\bar{C}_{\pi^*}$.

1: Linear Mixture MDPs

[PAC Bound for CPPO] Suppose $P^\star \in M$. With probability $1-\delta$,
$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = \tilde{O}((1-\gamma)^{-2}\sqrt{\bar{C}_{\pi^*,\mathrm{mix}} d^2 \ln(1/\delta)/n}) .$$

- Partial coverage concept is refined as $\bar{C}_{\pi^*,\mathrm{mix}} < \infty$.

2 KNRs

Definition: Kerneliazed nonlinear regulators. The true $P^\star$ is a Gaussian distribution $\mathcal{N}(W^\star \phi(s,a), \mathrm{I})$ ($W^\star \in \mathbb{R}^{d_S \times d}$) given a feature vector $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. ($d_S$ is a dimension of S)

- Include LQRs.
- Include RKHS models (GPs) .

[Concentrability Coefficient for KNRs]

$$\bar{C}_{\pi^*} = \sup_{x \in \mathbb{R}^d} \frac{x^\top \mathrm{E}_{(s,a)\sim d^{\pi^\star}}[\phi(s,a)\phi(s,a)^\top]x}{x^\top \mathrm{E}_{(s,a)\sim \rho}[\phi(s,a)\phi(s,a)^\top]x}.$$

- This is exactly the same as the one in linear MDPs.

2 KNRs

[PAC Bound for CPPO]

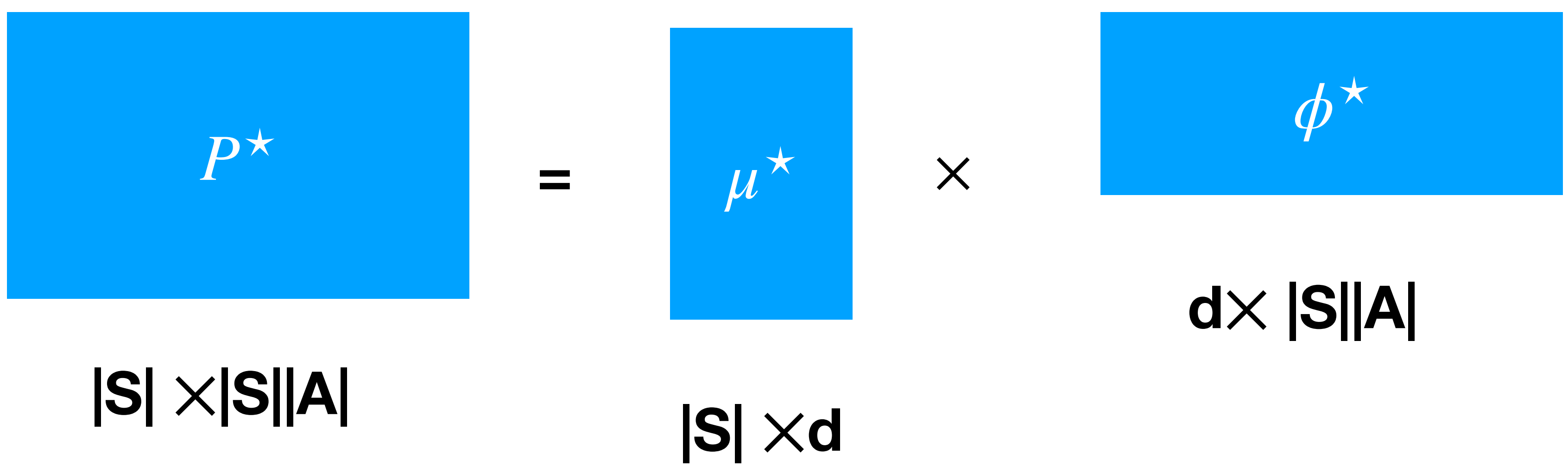
Let $\Sigma_\rho = \mathrm{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top]$. Suppose $P^\star \in M$. With $1-\delta$,

$$\forall \pi^*; V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} = \tilde{O}\left((1-\gamma)^{-2}\mathrm{rank}(\Sigma_\rho)^3\sqrt{d_S \bar{C}_{\pi^*}\ln(1/\delta)/n}\right).$$

- Partial coverage concept is refined as $\bar{C}_{\pi^*} < \infty$.
- Σ_ρ can be singular!! The error depends on $\mathrm{rank}[\Sigma_\rho]$ but not d.
- d can be infinite. Formally, extended to the infinite-dimensional setting, $P^\star = \mathcal{N}(g^\star(s,a), I)$ where $g^\star$ is an element of RKHS.

3: Low-rank MDPs

Definition: Low-rank MDPs [JKALS17, AKKS20]. The true $P^\star$ is $\mu^\star(s')^\top \phi^\star(s,a)$. Both $\mu^\star(\cdot), \phi^\star(\cdot)$ are unknown features. ($\mu^\star : \mathcal{S} \to \mathbb{R}^d, \phi^\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$)



- Features are unknown 😩. We set the function classes $\mu^\star \in \Psi, \phi^\star \in \Phi$.

- Low-rank MDPs include latent variable models, block MDPs and linear MDPs.

3: Low-rank MDPs

Definition: Low-rank MDPs [JKALS17,AKKS20]. The true $P^\star$ is $\mu^\star(s')^\top \phi^\star(s,a)$. Both $\mu^\star(\cdot), \phi^\star(\cdot)$ are unknown features. ($\mu^\star : \mathcal{S} \to \mathbb{R}^d, \phi^\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$)

[Concentrability Coefficient for Low-rank MDPs]

$$\bar{C}_{\pi^*,\phi^*} = \sup_{x \in \mathbb{R}^d} \frac{x^\top \mathrm{E}_{(s,a)\sim d^{\pi^\star}}[\phi^\star(s,a)\phi^\star(s,a)^\top] x}{x^\top \mathrm{E}_{(s,a)\sim \rho}[\phi^\star(s,a)\phi^\star(s,a)^\top] x}$$

- Looks similar to the one in linear MDPs and KNRs 🤔 $C^\dagger_{\pi^*,\phi^\star}$ depends on the only true feature $\phi^\star$ but not on other features.

3: Low-rank MDPs

[PAC Bound for CPPO] Let $\Sigma_{\rho,\phi^\star} = \mathrm{E}_{(s,a)\sim\rho}[\phi^\star(s,a)\phi^\star(s,a)^\top]$.
Suppose $P^\star \in M$. with $1-\delta$,
$$\forall \pi^*; V^{\pi^*}_{P^\star} - V^{\hat{\pi}}_{P^\star} = \tilde{O}((1-\gamma)^{-2}\sqrt{\bar{C}_{\pi^*,\phi^\star}\mathrm{rank}(\Sigma_{\rho,\phi^\star})\ln(|M|/\delta)/n}).$$

- Partial coverage concept is refined as $\bar{C}_{\pi^*,\phi^\star} < \infty$.
- Error depends on $\mathrm{rank}(\Sigma_{\rho,\phi^\star})$ instead of d.
- Previous related work on sparse linear MDPs ([HDLSW20]) assumes the non-singularity of $\Sigma_{\rho,\phi}$ for any $\phi \in \Phi$.

4. Factored MDPs

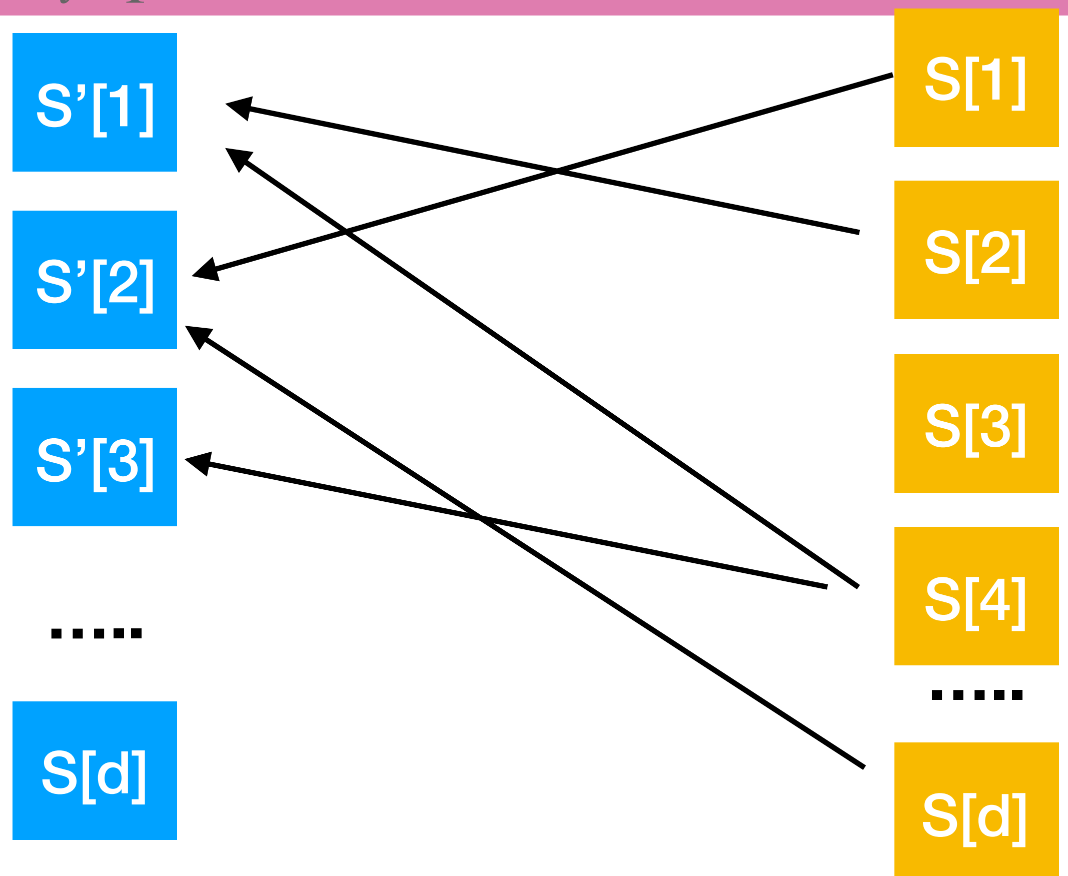
Definition: Factored (tabular) MDPs. $\mathcal{S} = \mathcal{O}^d$

$$P^\star(s' \mid s, a) = \prod_{i=1}^{d} P^\star(s'[i] \mid s[\mathrm{Pa}_i], a)\,.$$ Denote $\mathcal{S}_i = \mathcal{O}^{|\mathrm{Pa}_i|}$.

$\mathrm{Pa}_1 = [2,4]$

$\mathrm{Pa}_2 = [1,d]$

$\mathrm{Pa}_3 = [4]$



- Factored MDPs are governed by $O(\sum_i |\mathcal{O}|^{|\mathrm{Pa}_i|})$ parameters.
- Non-factored MDPs are governed by $O(|\mathcal{O}|^d)$ parameters.
- When $|\mathrm{Pa}_i| << d$, the difference is huge.
- Our goal is leveraging this factored structure.

4. Factored MDPs

Definition: Factored (tabular) MDPs. $\mathcal{S} = \mathcal{O}^d$

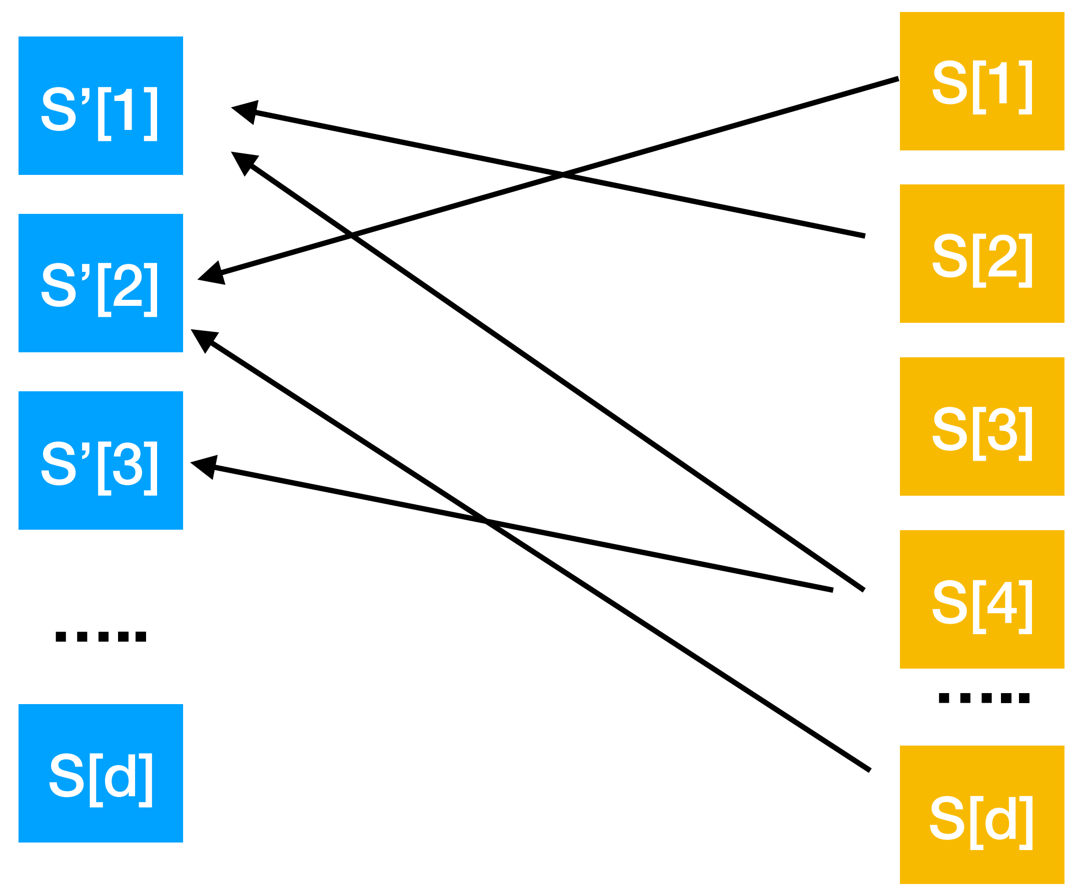
$$P^\star(s' \mid s, a) = \prod_{i=1}^{d} P^\star(s'[i] \mid s[\mathrm{Pa}_i], a)\,.$$ Denote $\mathcal{S}_i = \mathcal{O}^{|\mathrm{Pa}_i|}$.

- Introduce $\bar{C}^{[j]}_{\pi^*,\infty} = \max_{s_j \in \mathcal{S}_j, a \in \mathcal{A}} \frac{d^\pi(s_j, a)}{\rho(s_j, a)}$, $\nu(s_j, a) = \sum_{s :\in \mathcal{S}, s[\mathrm{Pa}_j] = s_j} \nu(s, a)$
- $\bar{C}^{[j]}_{\pi^*,\infty}$ is the marginal density ratio over each component.

$\mathrm{Pa}_1 = [2,4]$

$\mathrm{Pa}_2 = [1,d]$

$\mathrm{Pa}_3 = [4]$



Example

$$\bar{C}^{[1]}_{\pi^*,\infty} = \max_{s_1 \in \{s[2], s[4]\}, a} \frac{d^\pi(s_1, a)}{\rho(s_1, a)}\,.$$

4. Factored MDPs

[Concentrability Coefficient for Factored MDPs]

$$\bar{C}_{\pi^*,\infty} = \max_{j\in[1,\cdots,d]} \bar{C}^{j}_{\pi^*,\infty}$$

- $\bar{C}^{[j]}_{\pi^*,\infty}$ is smaller than the global density ratio $\max_{s\in\mathcal{S},a\in\mathcal{A}} \frac{d^{\pi}(s,a)}{\rho(s,a)}$ for any $j \in [1,\cdots,d]$.

- Thus, $\bar{C}_{\pi^*,\infty}$ is smaller than the global density ratio $\max_{s\in\mathcal{S},a\in\mathcal{A}} \frac{d^{\pi}(s,a)}{\rho(s,a)}$.

4. Factored MDPs

[PAC Bound for CPPO] Suppose $P^\star \in M$. With probability $1-\delta$,

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = \tilde{O}((1-\gamma)^{-2}\sqrt{d\bar{C}_{\pi^*,\infty}\sum_i |\mathcal{O}|^{\mathrm{Pa}_i}\ln(1/\delta)/n}).$$

- Partial coverage concept is refined as $\bar{C}_{\pi^*,\infty} < \infty$.
- This formally demonstrates the benefit of the factored structure in terms of the coverage condition.

Disclaimer

- We claim CPPO works for any MDPs. What does it mean?
- Any MDPs where the MLE has valid statistical guarantees.
- CPPO does not work on (different) linear MDPs [JYWJ20] and linear Bellman complete MDPs 😩.
- But, by taking a model-based perspective on them and modifying CPPO, we can still ensure the PAC guarantee under partial coverage.

Conclusion

- CPPO has the PAC guarantee under partial coverage assuming the realizability of the model. This works for any MDPs.
- Partial coverage concept is tailored to each model:
  - KNRs, linear mixture MDPs: relative condition numbers.
  - Low-rank MDPs: relative condition numbers defined on the true unknown features.
  - Factored MDPs: density ratios considering the factored structures.

Future Directions

- Computationally efficient algorithm which has PAC guarantee under partial coverage.
- Lower bound results.
- Bayesian algorithms.

References

- [CUSKS21] Mitigating covariate shift in imitation learning via offline data without great coverage. 202
- [MS 08] Finite-time bounds for fitted value iteration. ´ Journal of Machine Learning Research, 9(May):815–857, 2008.
- [XCJMA21] Bellman-consistent pessimism for offline reinforcement learning. arXiv preprint arXiv:2106.06926, 2021.
- [ZWB 21] Provable benefits of actor-critic methods for offline reinforcement learning. arXiv preprint arXiv: 2108.08812, 2021.
- [RZMJR21] Bridging offline reinforcement learning and imitation learning: A tale of pessimism. arXiv preprint arXiv:2103.12021, 2021.
- [JYW21] Is pessimism provably efficient for offline rl? arXiv preprint arXiv:2012.15085, 2021.
- [ZCZS21] Corruption-robust offline reinforcement learning. arXiv preprint arXiv:2106.06630, 2021.
- [YW20] Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In Proceedings of the 37th International Conference on Machine Learning, pages 10746–10756, 2020.
- [AJSWY20] Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474. PMLR, 2020.
- [AKKS20] Flambe: Structural complexity and representation learning of low rank mdps. In Advances in Neural Information Processing Systems, volume 33, pages 20095–20107, 2020.
- [HDLSW20] Sparse feature selection makes batch reinforcement learning more sample efficient. In International Conference on Machine Learning, pages4063–4073. PMLR, 2021.