Nearly Optimal Latent State Decoding in Block MDPs

Yassir Jedra * Junghyun Lee † Alexandre Proutiere * Se-Young Yun †

*KTH Royal Institute of Technology, †KAIST AI

{jedra, alepro}@kth.se, {jh_lee00, yunseyoung}@kaist.ac.kr

Contributions

- First theoretical analyses of model estimation (mainly clustering) and reward-free RL, *specific* to BMDPs.
- Our clustering algorithm is computationally tractable (no oracles required!)
- We depart from the function approximation framework, i.e., no additional structural assumption!

→ Previous works (e.g., [1]) depend on a-priori chosen function class to approximate the decoding function f

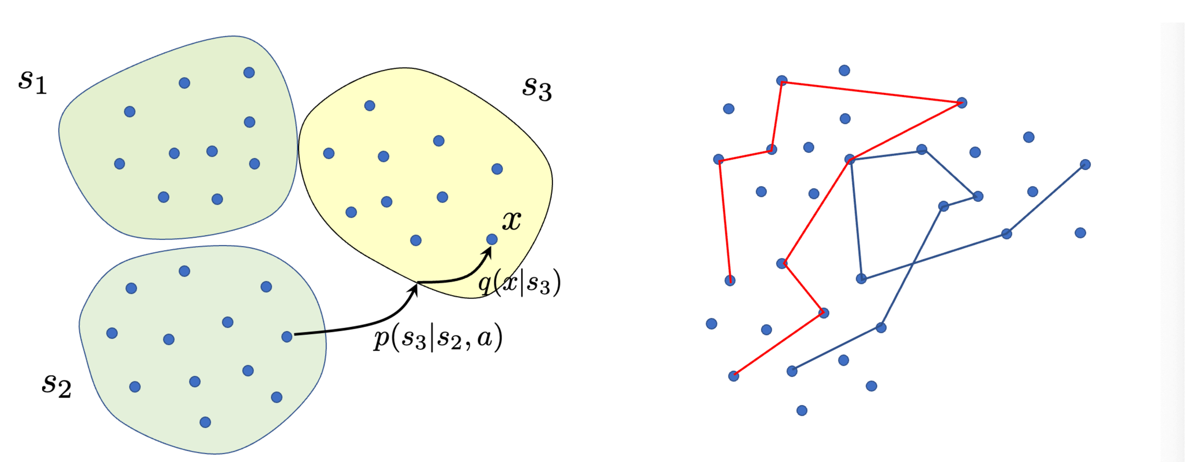
Block MDPs

BMDP Dynamics

In an episodic Block MDP (BMDP), the dynamics are defined by the tuple $(\mathcal{S}, \mathcal{X}, \mathcal{A}, p, q, f)$, where

- *Latent dynamics (unknown)* $p(s'|s,a)$
- *Emission distribution (unknown)* $q(x'|s')$
- *Decoding function (unknown)* $f : \mathcal{X} \to \mathcal{S}$

→ Assumption 1. The clusters are disjoint, i.e., for all $s \neq s'$, $f^{-1}(s) \cap f^{-1}(s') = \emptyset$.



Model vs. Observations

Learning Objective

Observations. T trajectories of length H, generated using the uniform behavior policy ρ:

$$\left\{(x_h^{(t)}, a_h^{(t)})_{h=1,\dots,H}\right\}_{t=1,\dots,T}$$

Objective. Find accurate estimates of f, p, and q.

We make the following assumptions

→ Assumption 2. The latent dynamics and emission distribution are η-regular, i.e., for any $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $x \in f^{-1}(s)$,

$$p(s'|s,a) = \Theta\left(\frac{1}{S}\right), \quad q(x|s) = \Theta\left(\frac{S}{n}\right).$$

→ Assumption 3. The cluster sizes $|f^{-1}(s)|$ grow linearly with $n = |\mathcal{X}|$.

Fundamental Limit

Lower Bound on Clustering Error

Theorem 1. *(Informal) For any BMDP Φ and any good clustering algorithm, the number of misclassified contexts $|\mathcal{E}|$ must satisfy*

$$\mathbb{E}_\Phi\left[\left|\mathcal{E}(\hat{f})\right|\right] \geq n \exp\left(-\frac{TH}{n} I(\Phi)(1 + o_n(1))\right)$$

with $I(\Phi) = -\frac{n}{TH}\log\left(\frac{C}{n}\sum_{x\in\mathcal{X}}\exp\left(-\frac{TH}{n}I(x;\Phi)\right)\right)$

→ Remark 1. The instance dependent constant $I(x;\Phi)$ measures the hardness of clustering context x, and $I(\Phi)$ the overall hardness of the instance.

→ Remark 2. The proof is based on the *change-of-measure* argument [2].

- $|\mathcal{E}| = o(n)$ (asymptotically accurate clustering) only if $TH = \omega(n)$ and $I(\Phi) > 0$
- $|\mathcal{E}| = o(1)$ (asymptotically exact clustering) only if $TH - \frac{n\log n}{I(\Phi)} = \omega(1)$ and $I(\Phi) > 0$

Latent State Decoding

Clustering Algorithm

Our algorithm runs in two phases sketched below:

- Phase 1 *(Initial Spectral Clustering)*

$\{(x_h^{(t)}, a_h^{(t)})_{h\in[H]}\}_{t\in[T]} \to$ Matrix Estimation $\to (\widehat{N}_{a,\Gamma_a})_{a\in\mathcal{A}}$

$(\widehat{N}_{a,\Gamma_a})_{a\in\mathcal{A}} \to$ S-Rank Approximation $\to (\hat{M}_a)_{a\in\mathcal{A}}$

$(\hat{M}_a)_{a\in\mathcal{A}}, (\hat{M}_a^\top)_{a\in\mathcal{A}} \to$ Aggregation $\to \hat{M}$

$\hat{M} \to$ ℓ_1-weighted K-medians $\to \hat{f}_1$

- Phase 2 *(Improvement)*

$\hat{f}_1 \to$ Iterative Likelihood Improvement $\to \hat{f}$

→ Remark 3. This is inspired by various literature on structure recovery in block models, e.g., [3].

Theoretical Guarantee on Initial Phase

Theorem 2. *Provided $TH = \omega(n)$, and $I(\Phi) > 0$, then we have*

$$\frac{|\mathcal{E}(\hat{f}_1)|}{n} \leq \mathcal{O}\left(\frac{nSA}{TH}\right) \quad w.h.p.$$

→ $|\mathcal{E}| = o(n)$ if $TH = \omega(n)$ and $I(\Phi) > 0$

Theoretical Guarantee after Improvement

Theorem 3.1. *If $TH = \omega(n)$ and $I(\Phi) > 0$, then w.h.p.,*

$$|\mathcal{E}(\hat{f})| \lesssim \sum_{x\in\mathcal{X}} \exp\left(-C\frac{TH}{n} I(x;\Phi)\right).$$

where $1/C = \mathrm{poly}(\eta)$.

→ $|\mathcal{E}| = o(1)$ if $TH - \frac{n\log n}{CI(x;\Phi)} = \omega(1)$ for all $x \in \mathcal{X}$ and $I(\Phi) > 0$.

Theoretical Guarantee on Model Estimation

We also provide guarantees on the plug-in estimators for the BMDP dynamics, $\hat{p}$ and $\hat{q}$:

Theorem 3.2. *The following holds w.h.p.: for all $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$d_{TV}(p(\cdot|s,a), \hat{p}(\cdot|s,a)) \lesssim \sqrt{\frac{S^3A^2\log(nSA)}{TH}} + \frac{SA|\mathcal{E}(\hat{f})|}{n}$$

$$d_{TV}(q(\cdot|s), \hat{q}(\cdot|s)) \lesssim \sqrt{\frac{Sn}{TH}} + \frac{S|\mathcal{E}(\hat{f})|}{n}$$

→ Both estimation errors are of order $o(1)$ if $TH = \omega(n)$ and $I(\Phi) > 0$.

Implications on Reward-Free RL

Preliminaries

Learning setup. In offline reward-free RL (ORF-RL), the setup is as follows:

→ Estimation phase. From the data $(x_h^{(t)}, a_h^{(t)})_{h\in[H],t\in[T]}$, estimate the MDP $\hat{\Phi}$;

→ Planning phase From the revealed reward function $r = (r_h)_{h\in[H]}$, compute $\hat{\pi}$ the optimal policy for $(\hat{\Phi}, r)$.

Objective. Find a model estimation procedure with the optimal decay rates $\varepsilon(T,H,n)$ in T, H, n.

→ Minimax setting:

$$\mathbb{P}\left(\sup_{r\in\mathcal{R}} V^\star(r) - V^{\hat{\pi}}(r) \leq \varepsilon(T,H,n)\right) \geq 1 - o_n(1)$$

→ Reward-specific setting:

$$\sup_{r\in\mathcal{R}} \mathbb{P}\left(V^\star(r) - V^{\hat{\pi}}(r) \leq \varepsilon(T,H,n)\right) \geq 1 - o_n(1)$$

($\mathcal{R}$ is the set of all possible reward functions)

Lower Bounds

Theorem 4. *(Minimax setting) For any BMDP Φ with $\Lambda(\Phi) > 0$, any algorithm satisfying $\mathbb{P}\left[\sup_{r\in\mathcal{R}}\frac{1}{H}V^\star(r) - V^{\hat{\pi}}(r) < \epsilon\right] \geq \frac{1}{2}$ requires $TH \gtrsim \frac{n\Lambda(\Phi)}{\epsilon^2}$, where $\Lambda(\Phi)$ doesn't depend on n.*

→ depends on the **estimation of** q, *not* the estimation of block structure!

Theorem 5. *(Reward-specific setting) Let $\epsilon = o(1)$ and $r \in \mathcal{R}$. For any BMDP Φ with $I(\Phi) > 0$, any algorithm satisfying $\frac{1}{H}V^\star(r) - V^{\hat{\pi}}(r) < \epsilon$ requires $TH \gtrsim n\log(\frac{1}{\epsilon}) + \frac{SA}{\epsilon^2}$.*

→ depends on the **estimation of block structure**!

Near-Matching Upper Bounds

Efficient Clustering + Planning $\implies$ Optimality!

Theorem 6, 7. *Under our efficient clustering method with an additional planner, we achieve*

$$\sup_{r\in\mathcal{R}}\frac{1}{H}\left|V^\star(r) - V^{\hat{\pi}}(r)\right| \lesssim \sqrt{\frac{nS^2A^2\log(SAH)}{TH}},$$

$$\frac{1}{H}\left|V^\star(r) - V^{\hat{\pi}}(r)\right| \lesssim \sqrt{\frac{S^3A^2H\log(SAHn)}{T}} + \frac{SH^2}{n}|\mathcal{E}(\hat{f})|$$

w.h.p., provided $TH = \omega(n)$ and $I(\Phi) > 0$.

→ Minimax setting: provided it holds that $TH = \omega(n)$, we have the following gains over the tabular setting

Block MDPs $\sqrt{\frac{n}{TH}}$ vs. Tabular MDPs $\sqrt{\frac{n^2}{TH}}$

→ Reward-specific setting: provided it holds that $TH = \omega(n\log(n))$, ignoring dependencies on H, we have the following gains over the tabular setting

Block MDPs $\sqrt{\frac{1}{T}}$ vs. Tabular MDPs $\sqrt{\frac{n}{T}}$

References

[1] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun.
Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning Approach.
In *ICML*, 2022.

[2] Tse L. Lai and Herbert Robbins.
Asymptotically Efficient Adaptive Allocation Rules.
Advances in Applied Mathematics, 6(1):4–22, 1985.

[3] Jaron Sanders, Alexandre Proutière, and Se-Young Yun.
Clustering in Block Markov Chains.
The Annals of Statistics, 48(6):3488 – 3512, 2020.