Nearly Optimal Latent State Decoding in Block MDPs

(KSC 2023 Workshop - Advances in Bandits and Bayesian Optimization)

Yassir Jedra[*], **Junghyun Lee**[†], Alexandre Proutière[*] and <u>Se-Young Yun</u>[†]

December 6, 2023

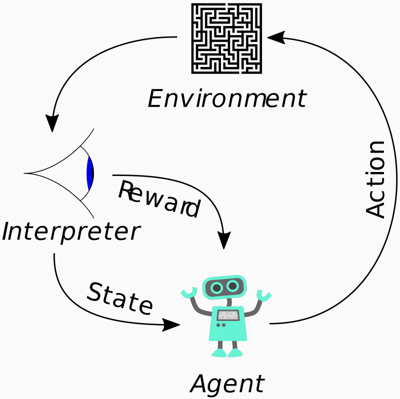
[*]Division of Decision and Control System, KTH Royal Institute of Technology
[†]Kim Jaechul Graduate School of AI, KAIST




Motivation

Reinforcement Learning

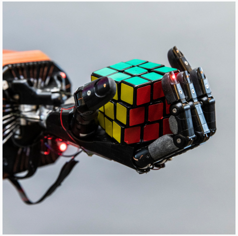
Learning optimal sequential behaviour/control from interacting with the environment

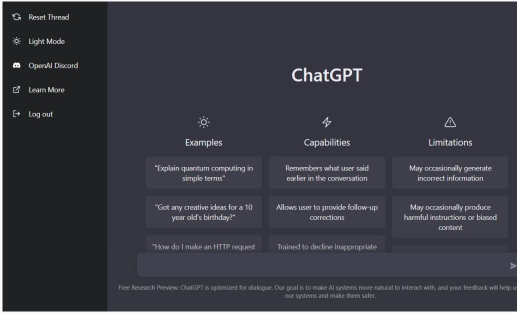


- *Unknown* state dynamics and rewards
- *Extremely large* state and action spaces

Numerous successes!

AlphaGo (Silver et al., 2016), robotic arm manipulation (Andrychowicz et al., 2020), flight manoeuvres (Abbeel et al., 2010), chatGPT (OpenAI, 2023), etc





- Many problems in reality are highly structured. What sort of structure in RL problems can enable fast learning? Can we learn the structure efficiently?

- Many problems in reality are highly structured. What sort of structure in RL problems can enable fast learning? Can we learn the structure efficiently?
- In this talk we focus on the **rich observation** (Krishnamurthy et al., 2016; Du et al., 2019; Zhang et al., 2022) setting where

- Many problems in reality are highly structured. What sort of structure in RL problems can enable fast learning? Can we learn the structure efficiently?
- In this talk we focus on the **rich observation** (Krishnamurthy et al., 2016; Du et al., 2019; Zhang et al., 2022) setting where
  - → The decision maker has access to high dimensional *contexts*;
  - → The dynamics depend on *unobserved* low dimensional *latent states* only;
  - → The mapping between contexts and latent states is unknown

- How can the decision maker exploit the underlying structure?
- What improvements in the sample complexity can we expect?

Our Contributions

- First instance-specific lower bound on the clustering error of BMDPs
- *Computationally efficient (oracle-free)* clustering algorithm with near-optimal upper bound on the clustering error as well as estimation of the dynamics (p, q)
- Implication of near-optimal clustering to offline, reward-free RL in BMDPs:
    - Improved sample complexities (lower bound and upper bound)

Block MDPs

Context, Latent States, and Dynamics

A Block MDP is denoted by $\Phi = (\mathcal{X}, \mathcal{S}, \mathcal{A}, p, q, f)$. The following are **unknown** to the learner:

- p is *transition kernel of the latent dynamics*: $p(s'|s, a)$
- q denotes the *emission probabilities*: $q(x|s')$ (prob. of emitting x at the latent state s')
- $f : \mathcal{X} \to \mathcal{S}$ is the *decoding function*: $f(x) = s \iff q(x|s) > 0$

Context, Latent States, and Dynamics

A Block MDP is denoted by $\Phi = (\mathcal{X}, \mathcal{S}, \mathcal{A}, p, q, f)$. The following are **unknown** to the learner:

- p is *transition kernel of the latent dynamics*: $p(s'|s,a)$
- q denotes the *emission probabilities*: $q(x|s')$ (prob. of emitting x at the latent state s')
- $f : \mathcal{X} \to \mathcal{S}$ is the *decoding function*: $f(x) = s \iff q(x|s) > 0$

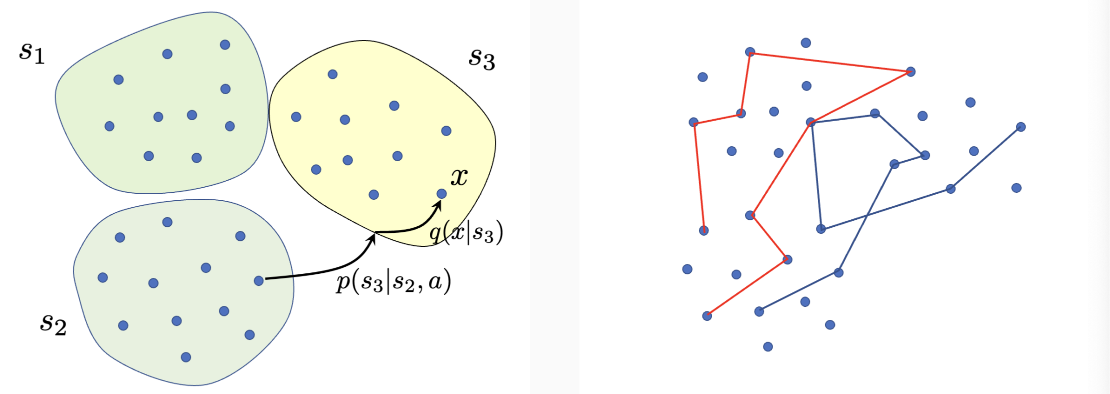
$\longrightarrow$ Assumption 0. The clusters do not overlap: $\forall s \neq s'$, $q(\cdot|s) \cap q(\cdot|s') = \emptyset$

$\longrightarrow$ Assumption 1. S, A, p are independent of n.

$\longrightarrow$ Assumption 2. $|f^{-1}(s)| = \alpha_s n$ for some $\alpha_s > 0$ s.t. $\sum_{s\in\mathcal{S}} \alpha_s = 1$.

$\longrightarrow$ Assumption 4. $\mu \sim \mathcal{U}(\mathcal{X})$, where μ is the distribution of the initial context.

Block MDPs



Model vs. observations

[Optional] Block MDPs vs. Linear MDPs

- **Linear structure**: $P(x'|x, a) = \phi(x, a)^\top \mu(x')$ with $\phi(x, a), \mu(x') \in \mathbb{R}^d$.

[1] means smaller in terms of sample complexity

[Optional] Block MDPs vs. Linear MDPs

- **Linear structure**: $P(x'|x,a) = \phi(x,a)^\top \mu(x')$ with $\phi(x,a), \mu(x') \in \mathbb{R}^d$.
- Block MDPs have a hidden linear structure in dimension $d = SA$:

$$\phi(x,a) = e_{(f(x),a)} \qquad \text{and} \qquad \mu(x')_{(s,a)} = q(x'|f(x'))p(f(x')|s,a)$$

[1] means smaller in terms of sample complexity

[Optional] Block MDPs vs. Linear MDPs

- **Linear structure**: $P(x'|x,a) = \phi(x,a)^\top \mu(x')$ with $\phi(x,a), \mu(x') \in \mathbb{R}^d$.
- Block MDPs have a hidden linear structure in dimension $d = SA$:

$$\phi(x,a) = e_{(f(x),a)} \qquad \text{and} \qquad \mu(x')_{(s,a)} = q(x'|f(x'))p(f(x')|s,a)$$

Linear MDPs	$\lesssim$[1]	Block MDPs	$\lesssim$	LowRank MDPs
μ is unknown ϕ is known		μ is unknown ϕ is unknown $\phi \in \mathcal{F}_{BMDP}$ $d = SA$		μ is unknown ϕ is unknown $\phi \in \mathcal{F}$

[1] means smaller in terms of sample complexity

[Optional] Block MDPs vs. Linear MDPs

- **Linear structure**: $P(x'|x,a) = \phi(x,a)^\top \mu(x')$ with $\phi(x,a), \mu(x') \in \mathbb{R}^d$.
- Block MDPs have a hidden linear structure in dimension $d = SA$:

$$\phi(x,a) = e_{(f(x),a)} \qquad \text{and} \qquad \mu(x')_{(s,a)} = q(x'|f(x'))p(f(x')|s,a)$$

Linear MDPs	$\lesssim$ [1]	Block MDPs	$\lesssim$	LowRank MDPs
μ is unknown ϕ is known		μ is unknown ϕ is unknown $\phi \in \mathcal{F}_{BMDP}$ $d = SA$		μ is unknown ϕ is unknown $\phi \in \mathcal{F}$

- **Linear structure in RL** (Jin et al., 2020b)

$$\underbrace{\text{Linear MDP}}_{P(x'|x,a)=\phi(x,a)^\top \mu(s')} + \underbrace{\text{Structured rewards}}_{r(x,a)=\phi(x,a)^\top \theta} \Longrightarrow \underbrace{\text{Q-function is linear}}_{Q^\pi(x,a)=\phi(x,a)^\top \xi^\pi}$$

[1] means smaller in terms of sample complexity

η-Regularity

$\longrightarrow$ Assumption 3. *(η-regularity)* There exists a $\eta > 1$ such that

$$(i) \quad \max_{s_1,s_2\in\mathcal{S}} \frac{\alpha_{s_1}}{\alpha_{s_2}} \le \eta \qquad (ii) \quad \max_{a\in\mathcal{A}} \max_{s_1,s_2,s_3\in\mathcal{S}} \frac{p(s_2|s_1,a)}{p(s_3|s_1,a)} \frac{p(s_1|s_2,a)}{p(s_1|s_2,a)} \le \eta$$

$$(iii) \quad \max_{s\in\mathcal{S}} \max_{x,y\in\mathcal{X}} \frac{q(x|s)}{q(y|s)} \le \eta \qquad (iv) \quad \max_{a_1,a_2\in\mathcal{A}} \max_{x,y\in\mathcal{X}} \frac{\pi(a_1|x)}{\pi(a_2|y)} \le \eta$$

$\longrightarrow$ Remark 1. similar to SBMs (Abbe, 2018), DCBMs (Gao et al., 2018), *Block Markov Chains* (Sanders et al., 2020), etc.

$\longrightarrow$ Remark 2. Assumption 3 assures that every context is visited sufficiently many times with uniform-like ρ. This can be relaxed to a weaker assumptions, e.g., aperiodic and communicating.

$\longrightarrow$ Remark 3. Without Assumption 3, there can exist some under-explored latent state, which unavoidably leads to constant error.

$\longrightarrow$ Remark 4. η controls *the mixing time* and scaling of *separation* between clusters!

Difference to Block Markov Chains (Sanders et al., 2020)

- Controllability of the Markov chains via action
- Possibly *nonuniform* emission probabilities at each latent state
- Doesn't necessarily start from stationary distribution (e.g., it may be that $H < t_{mix}$)
  - This is compensated by uniform initial distribution (Assumption 4)

Latent State Decoding (Clustering)

The Data

T trajectories of length H, $\left\{(x_h, a_h)_{h\in[H],t\in[T]}\right\}$ collected with some *memoryless*[2], behavior policy ρ.

$$\begin{array}{ccccc}
 & (h=1) & (h=2) & \dots & (h=H) \\
(t=1) & (x_1^{(1)}, a_1^{(1)}), & (x_2^{(1)}, a_2^{(1)}), & \dots, & (x_H^{(1)}, a_H^{(1)}) \\
(t=2) & (x_1^{(2)}, a_1^{(2)}), & (x_2^{(2)}, a_2^{(2)}), & \dots, & (x_H^{(2)}, a_H^{(2)}) \\
\vdots & & & & \\
(t=T) & (x_1^{(T)}, a_1^{(T)}), & (x_2^{(T)}, a_2^{(T)}), & \dots, & (x_H^{(T)}, a_H^{(T)})
\end{array}$$

$\rightarrow$ Remark. The data is *Markovian* across $[H]$ and *independent* across $[T]$.

[2] Our discussions can be partially extended to a more general history-dependent behavior policy.

The Data

T trajectories of length H, $\left\{(x_h, a_h)_{h\in[H],t\in[T]}\right\}$ collected with some *memoryless*[2], behavior policy ρ.

$$
\begin{array}{ccccc}
 & (h=1) & (h=2) & \dots & (h=H) \\
(t=1) & (x_1^{(1)}, a_1^{(1)}), & (x_2^{(1)}, a_2^{(1)}), & \dots, & (x_H^{(1)}, a_H^{(1)}) \\
(t=2) & (x_1^{(2)}, a_1^{(2)}), & (x_2^{(2)}, a_2^{(2)}), & \dots, & (x_H^{(2)}, a_H^{(2)}) \\
\vdots & & & & \\
(t=T) & (x_1^{(T)}, a_1^{(T)}), & (x_2^{(T)}, a_2^{(T)}), & \dots, & (x_H^{(T)}, a_H^{(T)})
\end{array}
$$

$\rightarrow$ Remark. The data is *Markovian* across $[H]$ and *independent* across $[T]$.

From this data, can we identify f in an optimal and computationally efficient manner?

[2]Our discussions can be partially extended to a more general history-dependent behavior policy.

A clustering algorithm $\mathcal{A}$ would do the following

$$\underbrace{\left(x_h^{(t)}, a_h^{(t)}\right)_{h\in[H],t\in[T]}}_{\text{Observations}} \longrightarrow \underbrace{\mathcal{A}}_{\text{Clustering algorithm}} \longrightarrow \underbrace{\hat{f}}_{\text{Decoding function}}$$

Number of misclassified contexts. *(up to permutation σ)*

$$\mathcal{E}(\hat{f}) := \min_{\sigma} \bigcup_{s\in\mathcal{S}} \hat{f}^{-1}(\sigma(s))\backslash f^{-1}(s)$$

$$|\mathcal{E}(\hat{f})| := \min_{\sigma} \left| \bigcup_{s\in\mathcal{S}} \hat{f}^{-1}(\sigma(s))\backslash f^{-1}(s) \right|$$

Objective. Output $\hat{f}$ that minimizes $|\mathcal{E}(\hat{f})|$.

Remark. We only care about the asymptotic dependencies on n, T, H.

Fundamental Lower Bound of Latent State Decoding

→ Definition 1. A clustering algorithm $\mathcal{A}$ is said β-locally better-than-random in $\tilde{\Phi}$ if the following holds:

$$\forall \widetilde{\Phi} \in \mathcal{V}_\beta(\Phi), \qquad \mathbb{P}_{\widetilde{\Phi}}\left(x \in \mathcal{E}(\hat{f})\right) \leq 1 - \frac{1}{S}$$

The β-neighborhood of Φ, $\mathcal{V}_\beta(\Phi)$ is defined as follows:

$$\mathcal{V}_\beta(\Phi) = \left\{ \tilde{\Phi} : \begin{cases} \max_{y \in \mathcal{X} : f(y) = \tilde{f}(y)} \max_{s \in \mathcal{S}} |q(y|s) - \tilde{q}(y|s)| \leq \beta, \\ \left| y \in \mathcal{X} : f(y) \neq \tilde{f}(y) \right| \leq 1 \end{cases} \right\}$$

β-locally better-than-random have reasonable performance and are stable to small model perturbations; see our paper (Jedra et al., 2023) for more details.

Theorem 1. Any algorithm that is β-locally better-than-random in Φ must satisfy

$$\forall x \in \mathcal{X}, \qquad \mathbb{P}_\Phi\left(x \in \mathcal{E}(\hat{f})\right) \gtrsim \exp\left(-\frac{TH}{n} I(x;\Phi)(1 + o_n(1))\right)$$

where $n = |\mathcal{X}|$, and $I(x;\Phi)$ is an information-theoretic constant specific to Φ.

Consequently, any such algorithm must also satisfy:

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

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

Proof based on the change-of-measure argument (Lai and Robbins, 1985).

Some Remarks on $I(x;\Phi)$ and $I(\Phi)$

- $I(x;\Phi)$ is defined through an optimization problem (Ugly expressions!)
- $I(x;\Phi)$ is independent of n, T, H.
- **Context x in the BMDP instance Φ with *small* $I(x;\Phi)$ is harder to cluster.**
  - If $I(x;\Phi) > 0$, then $I(y;\Phi) > 0$ for all y s.t. $f(y) = f(x)$.
  - $I(x;\Phi) = 0$ if and only if the transition rates to and out of the latent states f(x) and j are identical[3].
  - $I(\Phi) > 0$ if and only if $\min_{x\in\mathcal{X}} I(x;\Phi) > 0$.
- Assumption 3 (η-regularity) is crucial, as without it, we may have very "heterogeneous" BMDP with $I(x;\Phi)$ varying significantly, even in the same cluster.

[3] There exists $j \neq f(x)$ and $c > 0$ s.t. $p(\cdot|f(x), a) = p(\cdot|j, a)$ and $p(f(x)|\cdot, a) = cp(\cdot|j, a)$.

Some Remarks on $I(x;\Phi)$ and $I(\Phi)$

- $I(x;\Phi)$ is defined through an optimization problem (Ugly expressions!)
- $I(x;\Phi)$ is independent of n, T, H.
- **Context x in the BMDP instance Φ with *small* $I(x;\Phi)$ is harder to cluster.**
  - If $I(x;\Phi) > 0$, then $I(y;\Phi) > 0$ for all y s.t. $f(y) = f(x)$.
  - $I(x;\Phi) = 0$ if and only if the transition rates to and out of the latent states f(x) and j are identical[3].
  - $I(\Phi) > 0$ if and only if $\min_{x\in\mathcal{X}} I(x;\Phi) > 0$.
- Assumption 3 (η-regularity) is crucial, as without it, we may have very “heterogeneous” BMDP with $I(x;\Phi)$ varying significantly, even in the same cluster.

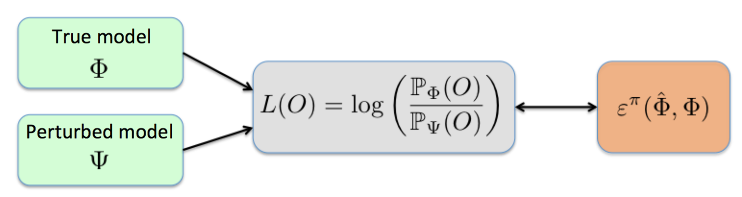
Importantly, the *necessary* conditions for the algorithm to be

asymptotically accurate ($\mathbb{E}_\Phi[|\mathcal{E}|] = o(n)$): $I(\Phi) > 0$ and $TH = \omega(n)$

asymptotically exact ($\mathbb{E}_\Phi[|\mathcal{E}|] = o(1)$): $I(\Phi) > 0$ and $TH - \frac{n \log n}{I(\Phi)} = \omega_n(1)$

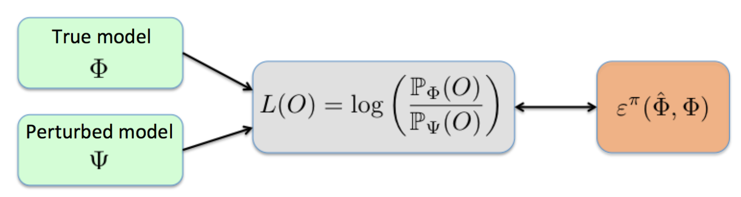
[3] There exists $j \neq f(x)$ and $c > 0$ s.t. $p(\cdot|f(x), a) = p(\cdot|j, a)$ and $p(f(x)|\cdot, a) = cp(\cdot|j, a)$.

[Optional] Proof of Theorem 1: Change-of-Measure Argument



Let $\mathcal{A}$ be an algorithm.

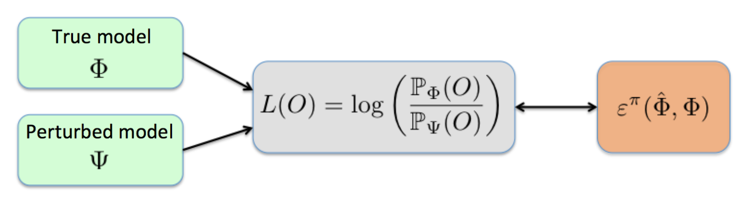
[Optional] Proof of Theorem 1: Change-of-Measure Argument



Let $\mathcal{A}$ be an algorithm.

1. Select a perturbed model Ψ.

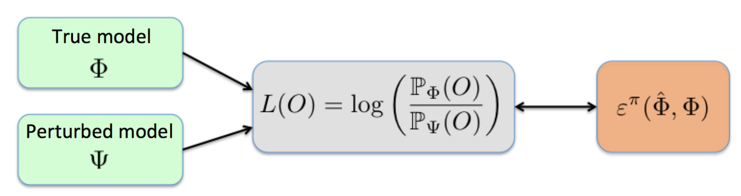
[Optional] Proof of Theorem 1: Change-of-Measure Argument



Let $\mathcal{A}$ be an algorithm.

1. Select a perturbed model Ψ.
2. Relate the log-likelihood ratio of observations under Φ and Ψ to the performance metrics: $\varepsilon^{\mathcal{A}}(\hat{\Phi}, \Phi) \leftrightarrow L(O)$.

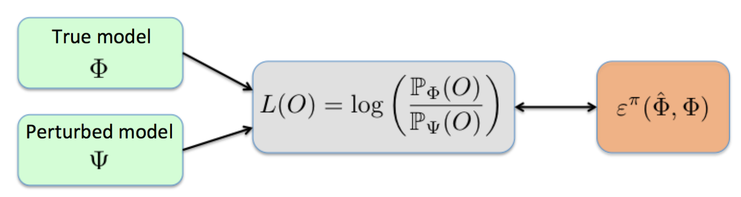
[Optional] Proof of Theorem 1: Change-of-Measure Argument



Let $\mathcal{A}$ be an algorithm.

1. Select a perturbed model Ψ.
2. Relate the log-likelihood ratio of observations under Φ and Ψ to the performance metrics: $\varepsilon^{\mathcal{A}}(\hat{\Phi}, \Phi) \leftrightarrow L(O)$.
3. Change-of-measure argument: $\mathbb{E}_\Phi[L(O)] \geq KL(\mathbb{P}_\Phi(A), \mathbb{P}_\Psi(A))$.

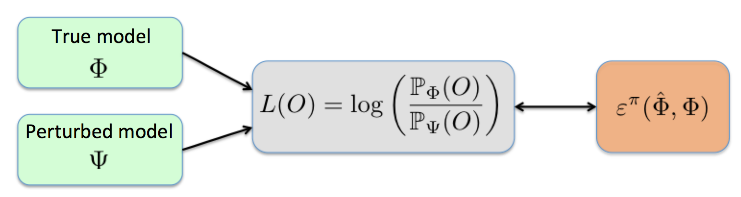
[Optional] Proof of Theorem 1: Change-of-Measure Argument



Let $\mathcal{A}$ be an algorithm.

1. Select a perturbed model Ψ.
2. Relate the log-likelihood ratio of observations under Φ and Ψ to the performance metrics: $\varepsilon^{\mathcal{A}}(\hat{\Phi}, \Phi) \leftrightarrow L(O)$.
3. Change-of-measure argument: $\mathbb{E}_\Phi[L(O)] \geq KL(\mathbb{P}_\Phi(A), \mathbb{P}_\Psi(A))$.
4. "Good" algorithm: $KL(\mathbb{P}_\Phi(A), \mathbb{P}_\Psi(A)) \geq G(\Phi, \Psi, T)$ (T observed trajectories).

[Optional] Proof of Theorem 1: Change-of-Measure Argument



Let $\mathcal{A}$ be an algorithm.

1. Select a perturbed model Ψ.
2. Relate the log-likelihood ratio of observations under Φ and Ψ to the performance metrics: $\varepsilon^{\mathcal{A}}(\hat{\Phi}, \Phi) \leftrightarrow L(O)$.
3. Change-of-measure argument: $\mathbb{E}_\Phi[L(O)] \geq KL(\mathbb{P}_\Phi(A), \mathbb{P}_\Psi(A))$.
4. "Good" algorithm: $KL(\mathbb{P}_\Phi(A), \mathbb{P}_\Psi(A)) \geq G(\Phi, \Psi, T)$ (T observed trajectories).
5. Maximize $G(\Phi, \Psi, T)$ over the choice of Ψ.

Near-Optimal Latent State Decoding

Algorithm

We propose an algorithm that has a matching upper bound up to some universal constants. The algorithm runs in two phases:

- Phase 1

$$
\begin{array}{ccccc}
\{(x_h^{(t)}, a_h^{(t)})_{t\in[T],h\in[H]}\} & \longrightarrow & \text{Matrix estimation} & \longrightarrow & (\hat{N}_{a,\Gamma_a})_{a\in\mathcal{A}} \\
(\hat{N}_{a,\Gamma_a})_{a\in\mathcal{A}} & \longrightarrow & \text{S-rank approximation} & \longrightarrow & \left(\hat{M}_a\right)_{a\in\mathcal{A}} \\
(\hat{M}_a)_{a\in\mathcal{A}} \quad (\hat{M}_a^\top)_{a\in\mathcal{A}} & \longrightarrow & \text{Aggregation} & \longrightarrow & \hat{M} \\
\hat{M} & \longrightarrow & \ell_1\text{-weighted K-medians} & \longrightarrow & \hat{f}_1
\end{array}
$$

- Phase 2

$$
\hat{f}_1 \longrightarrow \text{Iterative Likelihood Improvement} \longrightarrow \hat{f}
$$

Phase 1: Spectral Clustering

Algorithm 1: Initial Spectral Clustering

Input: T episodes $\{x_1^{(t)}, a_2^{(t)}, \ldots, x_{H-1}^{(t)}, a_{H-1}^{(t)}, x_H^{(t)}\}_{t\in[T]}$ generated by a behavior policy π

for $a \in \mathcal{A}$ **do**

  for all (x,y), $\hat{N}_a(x,y) \leftarrow \sum_{t,h} \mathbb{1}[(x_h^{(t)}, a_h^{(t)}, x_{h+1}^{(t)}) = (x,a,y)]$;

  $\Gamma_a \leftarrow \mathcal{X}$ after removing $\lfloor n \exp\left(-(TH/nA)\log(TH/nA)\right)\rfloor$ contexts with the highest number of visits i.e. those with the highest $\hat{N}_a(x) = \sum_y \hat{N}_a(x,y)$;

  $\hat{N}_{a,\Gamma_a} \leftarrow (\hat{N}_a(x,y)\mathbb{1}_{\{(x,y)\in\Gamma_a\}})_{x,y\in\mathcal{X}}$;

  $\hat{M}_a \leftarrow$ rank-S approximation of $\hat{N}_{a,\Gamma_a}$;

end

$\hat{M} \leftarrow \begin{bmatrix}(\hat{M}_1)^\top & \cdots & (\hat{M}_A)^\top & \hat{M}_1 & \cdots & \hat{M}_A\end{bmatrix}$;

Normalize the rows of $\hat{M}$ by the ℓ_1-norm;

Obtain $\hat{f}_1$ by applying the K-medians algorithm to the rows of $\hat{M}$;

Output: $\hat{f}_1$ (initial estimate of the decoding function)

- Empirical observation matrices:

$$\hat{N}_a(x,y) = \sum_{t,h} \mathbf{1}\left\{(x_h^{(t)}, a_h^{(t)}, a_{h+1}^{(t)}) = (x,a,y)\right\}$$

- Trimming (Regularization)

$$\hat{N}_{a,\Gamma_a}(x,y) = \hat{N}_a(x,y)\mathbf{1}\left\{(x,y) \in \Gamma_a \times \Gamma_a\right\}$$

where $\Gamma_a \subseteq \mathcal{X}$ is obtained by *trimming* $\left\lfloor n \exp\left(-\frac{TH}{nA}\log\left(\frac{TH}{nA}\right)\right)\right\rfloor$ contexts x with the highest number of visits of (x,a).

- Empirical observation matrices:

$$\hat{N}_a(x,y) = \sum_{t,h} \mathbf{1}\left\{\left(x_h^{(t)}, a_h^{(t)}, a_{h+1}^{(t)}\right) = (x,a,y)\right\}$$

- Trimming (Regularization)

$$\hat{N}_{a,\Gamma_a}(x,y) = \hat{N}_a(x,y)\mathbf{1}\left\{(x,y) \in \Gamma_a \times \Gamma_a\right\}$$

where $\Gamma_a \subseteq \mathcal{X}$ is obtained by *trimming* $\left\lfloor n \exp\left(-\frac{TH}{nA}\log\left(\frac{TH}{nA}\right)\right)\right\rfloor$ contexts x with the highest number of visits of (x,a).

Proposition 19. *(Markovian matrix concentration)*

$$\mathbb{P}\left(\max_{a\in\mathcal{A}} \|\hat{N}_{a,\Gamma_a} - \widetilde{N}_a\| \lesssim \mathrm{poly}(\eta)\sqrt{\frac{TH}{nA}}\right) \geq 1 - \mathcal{O}\left(\frac{1}{n} + e^{-\frac{TH}{nA}}\right)$$

- Proof inspired by Feige and Ofek (2005); Keshavan et al. (2010); Le et al. (2017); Sanders et al. (2020); Sanders and Senen–Cerda (2023).
- **Key point:** Bernstein concentration bounds for Markov chains with restarts!
  - Slightly generalizes the Markovian Bernstein concentration of Paulin (2015).

Theorem 2. *(Misclassification error of Phase 1)* Provided $TH = \omega(n)$, and $I(\Phi) > 0$, then we have

$$\frac{|\mathcal{E}(\hat{f}_1)|}{n} \leq \mathcal{O}\left(\frac{nSA}{TH}\right) = o(1) \qquad w.h.p.$$

$\rightarrow$ asymptotically accurate clustering!

Phase 2: Iterative Likelihood Improvement

Algorithm 2: Iterative Likelihood Improvement

Input: Initial cluster estimates $\hat{f}_1$ and T episodes $\{x_1^{(t)}, a_2^{(t)}, \ldots, x_{H-1}^{(t)}, a_{H-1}^{(t)}, x_H^{(t)}\}_{t\in[T]}$

for $\ell = 1$ *to* $L = \lfloor \log(nA) \rfloor$ **do**

for all (s, j, a), $\hat{p}_\ell(s|j, a) \leftarrow \frac{\hat{N}_a(\hat{f}_\ell^{-1}(j), \hat{f}_\ell^{-1}(s))}{\hat{N}_a(\hat{f}_\ell^{-1}(j), \mathcal{X})}$ and $\hat{p}_\ell^{bwd}(s, a|j) \leftarrow \frac{\hat{N}_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(j))}{\sum_{\tilde{a}\in\mathcal{A}} \hat{N}_{\tilde{a}}(\mathcal{X}, \hat{f}_\ell^{-1}(j))}$;

for all x, $\hat{f}_{\ell+1}(x) \leftarrow \mathrm{argmax}_{j\in\mathcal{S}} \mathcal{L}^{(\ell)}(x, j)$ where

$$\mathcal{L}^{(\ell)}(x, j) = \sum_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \left[\hat{N}_a(x, \hat{f}_\ell^{-1}(s)) \log \hat{p}_\ell(s|j, a) + \hat{N}_a(\hat{f}_\ell^{-1}(s), x) \log \hat{p}_\ell^{bwd}(s, a|j) \right];$$

end

$\hat{f} \leftarrow \hat{f}_{L+1}$;

Output: $\hat{f}$

- The form of $\mathcal{L}^{(\ell)}$ is inspired by the derivation of the lower bound.

Theorem 3. *(Final misclassification error)* If $TH = \omega(n)$ and $I(\Phi) > 0$, then

$$\frac{|\mathcal{E}(\hat{f})|}{n} = \mathcal{O}\left(\frac{1}{n}\sum_{x\in\mathcal{X}} \exp\left(-C'\frac{TH}{n} I(x;\Phi)\right)\right)$$

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

- If $\hat{f}_1$ is sufficiently good (*Theorem 2*), then the likelihood iterations are contractive and convergence to the optimal f is guaranteed with high probability.

→ Exact clustering when $TH - \frac{n\log(n)}{C' I(x;\Phi)} = \omega_n(1)$ for all $x \in \mathcal{X}$

  - Compare with the necessary condition from the lower bound: $TH - \frac{n \log n}{I(\Phi)} = \omega_n(1)$

Model estimation.

With the final estimated $\hat{f}$, the plug-in estimators give a good estimate of the transition dynamics:

Theorem 3. For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

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

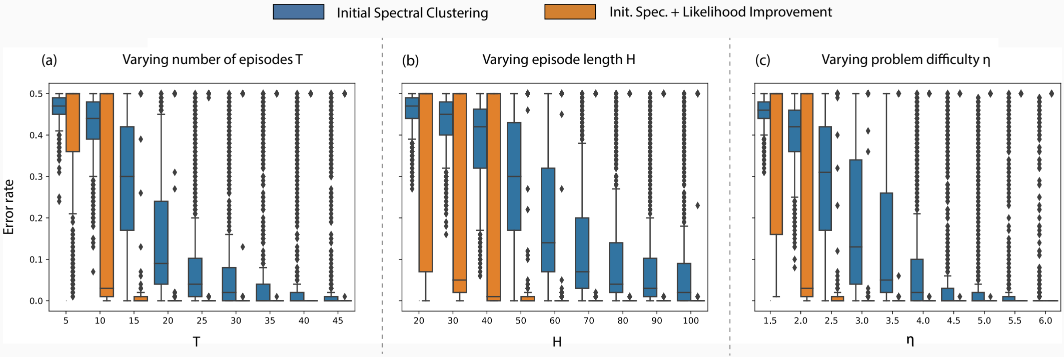
$$d_{TV}(q(\cdot|s), \hat{q}(\cdot|s)) \lesssim \sqrt{\frac{Sn}{TH}} + \frac{S|\mathcal{E}(\hat{f})|}{n}$$

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

$\longrightarrow$ Remark. We don't know whether this rate is (minimax) optimal for BMDPs. It would be interesting to see whether recent works on Markov chain estimation (Wolfer and Kontorovich, 2021; Banerjee et al., 2022) can give some insights.

Experiments on Synthetic BMDP Environments

We consider a BMDP environment where η-regularity holds.

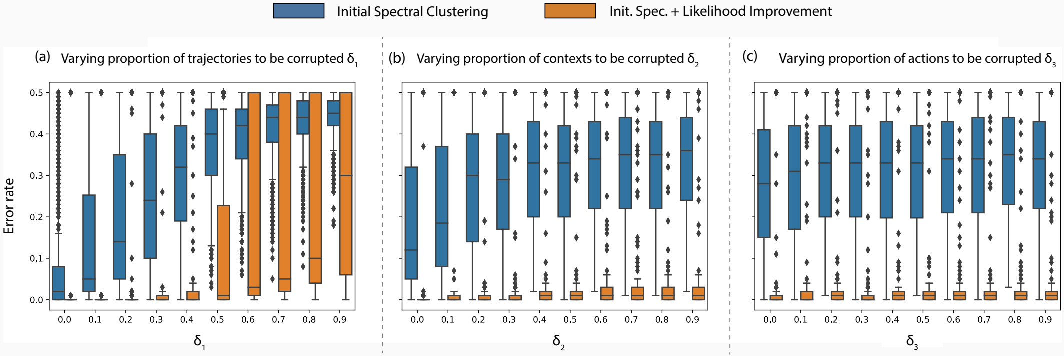


We plot the clustering error against T, H and η.

See Lee and Yun (2022) for more details.

Experiments on Synthetic BMDP Environments

We now consider a BMDP environment where η-regularity does not hold.



We plot the clustering error against some corruption parameters δ_1, δ_2 and δ_3. ($\delta_1 T$ trajectories, $\delta_2 n$ contexts, $\delta_3 A$ actions corrupted)

See Lee and Yun (2023) for more details.

From Clustering to Offline, Reward-Free RL

RL Preliminaries. A Block MDP $\Phi = (\mathcal{X}, \mathcal{S}, \mathcal{A}, p, q, f, H)$

- Deterministic rewards $r \in \mathcal{R}$ such that

$$\forall h \in [H], \forall (x, a) \in \mathcal{X} \times \mathcal{A}, \qquad r_h(x, a) \in [0, 1]$$

- Value function of a policy $\pi = (\pi_h)_{h \in [H]}$,

$$V^\pi(r) = \mathbb{E}_\Phi \left[\sum_{h=1}^{H} r_h(x_h, \pi_h(x_h)) \right]$$

- Optimal policy $\pi^\star(r)$ and it value $V^\star(r)$

$$\pi^\star(r) \in \arg\max_{\pi \in \Pi} V^\pi(r) \qquad \text{and} \qquad V^\star(r) = V^{\pi^\star(r)}(r)$$

In offline, reward-free RL (Jin et al., 2020a; Ren et al., 2021; Yin and Wang, 2021), the setup is as follows:

1. **Estimation phase.** From the given data $(x_h^{(t)}, a_h^{(t)})_{h\in[H],t\in[T]}$, estimate the (B)MDP $\hat{\Phi}$;
2. **Planning phase.** From the revealed reward function $(r_h)_{h\in[H]}$, compute $\hat{\pi}$ the optimal policy for $(\hat{\Phi}, r)$.

In offline, reward-free RL (Jin et al., 2020a; Ren et al., 2021; Yin and Wang, 2021), the setup is as follows:

1. **Estimation phase.** From the given data $(x_h^{(t)}, a_h^{(t)})_{h\in[H],t\in[T]}$, estimate the (B)MDP $\hat{\Phi}$;
2. **Planning phase.** From the revealed reward function $(r_h)_{h\in[H]}$, compute $\hat{\pi}$ the optimal policy for $(\hat{\Phi}, r)$.

Objectives. Find a model estimation procedure so that

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

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

with the best decay rates $\varepsilon(T,H,n)$ in T, H, n. Here, $\mathcal{R}$ is the set of all possible reward functions.

Lower Bounds

Theorem 6. *(minimax reward)* Let Φ be a BMDP such that $I(\Phi) > 0$, then any algorithm that guarantees

$$\mathbb{P}\left(\sup_{r \in \mathcal{R}} \frac{1}{H} V^{\star}(r) - V^{\hat{\pi}}(r) < \varepsilon\right) > \frac{1}{2},$$

requires $TH = \Omega\left(\frac{n\Lambda(\Phi)}{\varepsilon^2}\right)$ samples, where $\Lambda(\Phi)$ is some well-defined quantity[4] that does not depend on n, T, H.

[4] Precisely, $\Lambda(\Phi) = \max_{v \in [-1,1]^S} \frac{1}{S} \sum_{s=1}^{S} \max_{a_1, a_2} \langle p(\cdot|s, a_1) - p(\cdot|s, a_2), v \rangle$, taken from Jin et al. (2020a).

Lower Bounds

Theorem 6. *(minimax reward)* Let Φ be a BMDP such that $I(\Phi) > 0$, then any algorithm that guarantees

$$\mathbb{P}\left(\sup_{r\in\mathcal{R}} \frac{1}{H} V^{\star}(r) - V^{\hat{\pi}}(r) < \varepsilon\right) > \frac{1}{2},$$

requires $TH = \Omega\left(\frac{n\Lambda(\Phi)}{\varepsilon^2}\right)$ samples, where $\Lambda(\Phi)$ is some well-defined quantity[4] that does not depend on n, T, H.

- Gain over tabular MDPs *(no structure)*. For minimax reward setting in tabular MDPs, the lower bound (Menard et al., 2021; Yin and Wang, 2021) is $\Omega(\frac{H^3 A n^2}{\epsilon^2})$
- Improvement of order n and H^3

[4]Precisely, $\Lambda(\Phi) = \max_{v\in[-1,1]^S} \frac{1}{S}\sum_{s=1}^{S} \max_{a_1,a_2}\langle p(\cdot|s,a_1) - p(\cdot|s,a_2), v\rangle$, taken from Jin et al. (2020a).

Lower Bounds

Theorem 7. *(reward specific)* Let Φ be a block MDP such that $I(\Phi) > 0$, then for all $r \in \mathcal{R}$ initially revealed to the algorithm, for the algorithm to satisfy

$$\frac{1}{H}\mathbb{E}_{\Phi}[V^{\star}(r) - V^{\hat{\pi}}(r)] \leq \varepsilon,$$

requires $TH = \Omega\left(\frac{n}{I(\Phi)} \log\left(\frac{1}{\varepsilon}\right) + \frac{SA}{\varepsilon^2}\right)$ samples.

Lower Bounds

Theorem 7. *(reward specific)* Let Φ be a block MDP such that $I(\Phi) > 0$, then for all $r \in \mathcal{R}$ initially revealed to the algorithm, for the algorithm to satisfy

$$\frac{1}{H}\mathbb{E}_{\Phi}[V^{\star}(r) - V^{\hat{\pi}}(r)] \leq \varepsilon,$$

requires $TH = \Omega\left(\frac{n}{I(\Phi)}\log\left(\frac{1}{\varepsilon}\right) + \frac{SA}{\varepsilon^2}\right)$ samples.

- Gain over tabular MDPs *(no structure)*. For reward specific setting in tabular MDPs, the lower bound is $\Omega(\frac{HAn}{\epsilon^2})$ with matching upper bound (Menard et al., 2021; Ren et al., 2021).
- The gain is $\Omega\left(n\log(\frac{1}{\varepsilon}) + \frac{1}{\varepsilon^2}\right)$ vs. $\Omega(\frac{Hn}{\varepsilon^2})$
  - ex) If $\varepsilon = 1/\sqrt{n}$, then $\Omega(n\log n)$ vs. $\Omega(Hn^2)$, i.e., improvement by a factor of $Hn/\log n$

Upper Bounds

Efficient Clustering + Planning $\implies$ ***Minimax optimality***

Theorem 8. Under our efficient clustering method with an additional planner we achieve

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

$$\frac{1}{H} \left| V^\star(r) - V^{\hat{\pi}}(r) \right| = \mathcal{O}\left(\sqrt{\frac{S^3 A^2 H \log(SAHn)}{T}} + \frac{SH^2}{n} \sum_{x \in \mathcal{X}} \exp\left(-\frac{TH}{n} I(x;\Phi) \right) \right)$$

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

- These (nearly) match our lower bounds.

Conclusion

Concluding Remarks

Related work: use function approximations and optimization oracles to approximate the latent state decoding function (Jiang et al., 2017; Dann et al., 2018; Du et al., 2019; Misra et al., 2020; Foster et al., 2021; Zhang et al., 2022).

- Sample complexity scaling as $\log|\mathcal{F}|/\varepsilon^2$ where $\mathcal{F}$ is the class of approximation functions;
- Without any further assumption, $\log|\mathcal{F}| \approx n$, and no gain vs tabular MDP!
- Intractable algorithm (in principle) due to the dependency on oracles.

Concluding Remarks

Related work: use function approximations and optimization oracles to approximate the latent state decoding function (Jiang et al., 2017; Dann et al., 2018; Du et al., 2019; Misra et al., 2020; Foster et al., 2021; Zhang et al., 2022).

- Sample complexity scaling as $\log |\mathcal{F}|/\varepsilon^2$ where $\mathcal{F}$ is the class of approximation functions;
- Without any further assumption, $\log |\mathcal{F}| \approx n$, and no gain vs tabular MDP!
- Intractable algorithm (in principle) due to the dependency on oracles.

Our contributions: First instance-specific lower and near-optimal *efficient* clustering algorithm for BMDPs, as well as order-optimal sample complexities in offline, reward-free RL.

Concluding Remarks

Related work: use function approximations and optimization oracles to approximate the latent state decoding function (Jiang et al., 2017; Dann et al., 2018; Du et al., 2019; Misra et al., 2020; Foster et al., 2021; Zhang et al., 2022).

- Sample complexity scaling as $\log |\mathcal{F}|/\varepsilon^2$ where $\mathcal{F}$ is the class of approximation functions;
- Without any further assumption, $\log |\mathcal{F}| \approx n$, and no gain vs tabular MDP!
- Intractable algorithm (in principle) due to the dependency on oracles.

Our contributions: First instance-specific lower and near-optimal *efficient* clustering algorithm for BMDPs, as well as order-optimal sample complexities in offline, reward-free RL.

Future Directions:

- No clever exploration scheme, can we be adaptive and do better?
- Interleaved estimation and exploration?
- Removing/Relaxing Assumption 3 (η-regularity)
- BMDP with corruptions?
- Beyond block structures $\rightarrow$ low-rank, hierarchical, latent MDPs...etc.

Thank you for your attention!

Paper link (pmlr)

References

Emmanuel Abbe. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Pieter Abbeel, Adam Coates, and Andrew Y. Ng. Autonomous Helicopter Aerobatics through Apprenticeship Learning. *The International Journal of Robotics Research*, 29(13):1608–1639, 2010. doi: 10.1177/0278364910371999. URL `https://doi.org/10.1177/0278364910371999`.

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. doi: 10.1177/0278364919887447. URL `https://doi.org/10.1177/0278364919887447`.

Imon Banerjee, Harsha Honnappa, and Vinayak Rao. Offline Estimation of Controlled Markov Chains: Minimax Nonparametric Estimators and Sample Efficiency. *arXiv preprint arXiv:2211.07092*, 2022.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On Oracle-Efficient PAC RL with Rich Observations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with Rich Observations via Latent State Decoding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 09–15 Jun 2019.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Dylan Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2059–2059. PMLR, 15–19 Aug 2021.

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153 – 2185, 2018.

Yassir Jedra, Junghyun Lee, Alexandre Proutière, and Se-Young Yun. Nearly Optimal Latent State Decoding in Block MDPs. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2805–2904. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/jedra23a.html.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-Free Exploration for Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879. PMLR, 13–18 Jul 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020b.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix Completion from Noisy Entries. *Journal of Machine Learning Research*, 11(69):2057–2078, 2010.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC Reinforcement Learning with Rich Observations. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

Can M. Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

Junghyun Lee and Se-Young Yun. Preliminary Empirical Analyses of Clustering in Block MDPs. In *Korea Software Congress 2022, Jeju Island, Republic of Korea, Dec 20 - Dec 23, 2022*, 2022.

Junghyun Lee and Se-Young Yun. (Further) Empirical Analyses of Clustering in Block MDPs. *submitted to KIISE Journal of Computing*, 2023.

Pierre Menard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7599–7608. PMLR, 18–24 Jul 2021.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6961–6971. PMLR, 13–18 Jul 2020.

OpenAI. GPT-3.5 Language Model. `https://www.openai.com`, 2023.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1 – 32, 2015.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly Horizon-Free Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 15621–15634. Curran Associates, Inc., 2021.

Jaron Sanders and Albert Senen–Cerda. Spectral norm bounds for block markov chain random matrices. *Stochastic Processes and their Applications*, 158:134–169, 2023.

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

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. doi: 10.1038/nature16961. URL `https://doi.org/10.1038/nature16961`.

Geoffrey Wolfer and Aryeh Kontorovich. Statistical estimation of ergodic Markov chain kernel over discrete state space. *Bernoulli*, 27(1):532 – 553, 2021.

Ming Yin and Yu-Xiang Wang. Optimal Uniform OPE and Model-based Offline Reinforcement Learning in Time-Homogeneous, Reward-Free and Task-Agnostic Settings. In *Advances in Neural Information Processing Systems*, volume 34, pages 12890–12903. Curran Associates, Inc., 2021.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26517–26547. PMLR, 17–23 Jul 2022.