Preliminary Empirical Analyses of Clustering in Block MDPs

KSC 2022 Oral Session #8

Junghyun Lee, Se-Young Yun

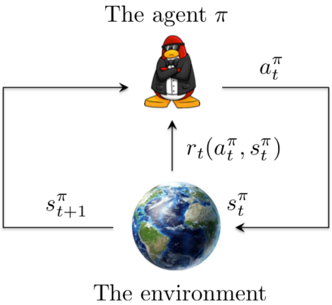
December 22, 2022

Kim Jaechul Graduate School of AI, KAIST

KAIST AI Graduate School of AI

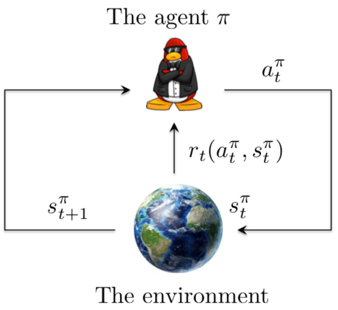
Optimization and Statistical Inference LAB

Learning optimal sequential behaviour / control from interacting with the environment



Unknown state dynamics and reward function

Learning optimal sequential behaviour / control from interacting with the environment



Unknown state dynamics and reward function

1. **Best policy identification** – sample complexity (e.g., [Azar et al., 2013])
2. Online learning – regret (e.g., [Jaksch et al., 2010])

Motivation

- Tabular MDPs (S states, A actions, $p(\cdot|, s, a)$, $r(s, a)$) are not (really) learnable – sample complexity always scales as SA.

Motivation

- Tabular MDPs (S states, A actions, $p(\cdot|, s, a)$, $r(s, a)$) are not (really) learnable – sample complexity always scales as SA.
- RL algorithms need to learn and exploit as much as possible any underlying structure.

Motivation

- Tabular MDPs (S states, A actions, $p(\cdot|, s, a)$, $r(s, a)$) are not (really) learnable – sample complexity always scales as SA.
- RL algorithms need to learn and exploit as much as possible any underlying structure.
- Our considered setting: the **rich observation** MDP 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.

Motivation

- Tabular MDPs (S states, A actions, $p(\cdot|, s, a)$, $r(s, a)$) are not (really) learnable – sample complexity always scales as SA.
- RL algorithms need to learn and exploit as much as possible any underlying structure.
- Our considered setting: the **rich observation** MDP 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.

[Jedra et al., 2022]: (with some regularity assumptions) complete characterization of clustering (and reward-free RL) in block MDPs

Motivation

- Tabular MDPs (S states, A actions, $p(\cdot|, s, a)$, $r(s, a)$) are not (really) learnable – sample complexity always scales as SA.
- RL algorithms need to learn and exploit as much as possible any underlying structure.
- Our considered setting: the **rich observation** MDP 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.

[Jedra et al., 2022]: (with some regularity assumptions) complete characterization of clustering (and reward-free RL) in block MDPs

Empirically, how well does the clustering algorithm of [Jedra et al., 2022] work?

Outline

1. Block MDPs

Contexts, Latent States, and Transition Dynamics

A **Block MDP** [Du et al., 2019] is defined by $\Phi = (\mathcal{X}, \mathcal{S}, \mathcal{A}, p, q, f)$

- $\mathcal{X}$ is the *observable* context space with $|\mathcal{X}| = n$
- $\mathcal{S}$ is the *latent* state space with $|\mathcal{S}| = S$
- $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$
- p is the transition kernel of *latent* dynamics: $p(s'|s,a)$
- q denote the *emission probabilities*: $q(x|s)$ (prob. of x if the new latent state is s)
- f is the **decoding function**: $f(x)$ is the *cluster* or *latent state* of context x and satisfies
$$f(x) = s \iff q(x|s) > 0.$$

Contexts, Latent States, and Transition Dynamics

A **Block MDP** [Du et al., 2019] is defined by $\Phi = (\mathcal{X}, \mathcal{S}, \mathcal{A}, p, q, f)$

- $\mathcal{X}$ is the *observable* context space with $|\mathcal{X}| = n$
- $\mathcal{S}$ is the *latent* state space with $|\mathcal{S}| = S$
- $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$
- p is the transition kernel of *latent* dynamics: $p(s'|s,a)$
- q denote the *emission probabilities*: $q(x|s)$ (prob. of x if the new latent state is s)
- f is the **decoding function**: $f(x)$ is the *cluster* or *latent state* of context x and satisfies
$$f(x) = s \iff q(x|s) > 0.$$

To make sure that the clusters do not overlap, we make the following assumption:

Assumption 0 $\forall s \neq s'$, $q(\cdot|s) \cap q(\cdot|s') = \emptyset$, which implies that

$$\mathcal{X} = \dot{\bigcup}_{s \in \mathcal{S}} f^{-1}(s), \quad f^{-1}(s) := \{x \in \mathcal{X} : f(x) = s\}.$$

Contexts, Latent States, and Transition Dynamics

A **Block MDP** [Du et al., 2019] is defined by $\Phi = (\mathcal{X}, \mathcal{S}, \mathcal{A}, p, q, f)$

- $\mathcal{X}$ is the *observable* context space with $|\mathcal{X}| = n$
- $\mathcal{S}$ is the *latent* state space with $|\mathcal{S}| = S$
- $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$
- p is the transition kernel of *latent* dynamics: $p(s'|s,a)$
- q denote the *emission probabilities*: $q(x|s)$ (prob. of x if the new latent state is s)
- f is the **decoding function**: $f(x)$ is the *cluster* or *latent state* of context x and satisfies

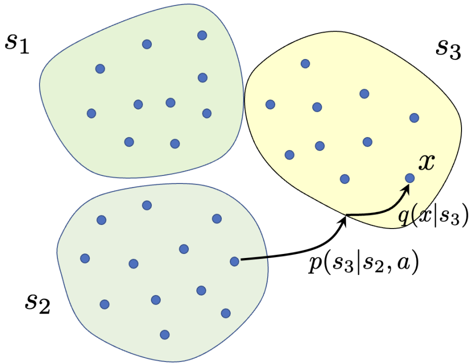
$$f(x) = s \iff q(x|s) > 0.$$

To make sure that the clusters do not overlap, we make the following assumption:

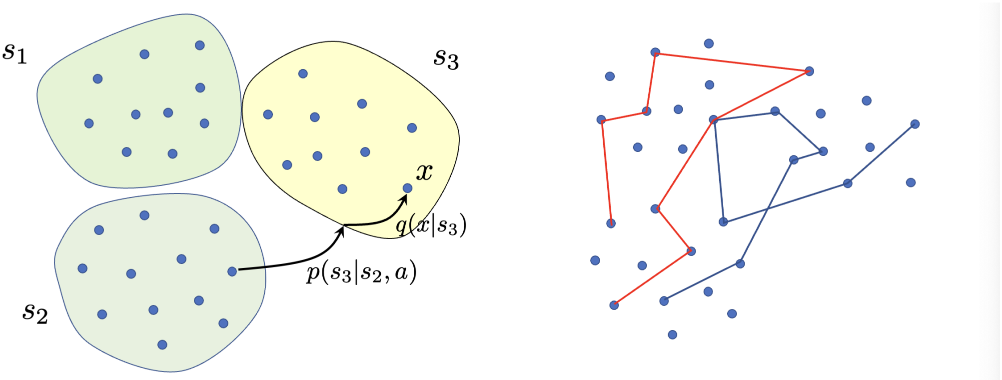
Assumption 0 $\forall s \neq s'$, $q(\cdot|s) \cap q(\cdot|s') = \emptyset$, which implies that

$$\mathcal{X} = \dot{\bigcup}_{s\in\mathcal{S}} f^{-1}(s), \quad f^{-1}(s) := \{x \in \mathcal{X} : f(x) = s\}.$$

$\Phi = (p, q, f)$ is **unknown** to the learner.


s_1
s_3
x
$q(x|s_3)$
$p(s_3|s_2, a)$
s_2

Example Trajectories of Block MDPs



Model vs Observations

The Data

Observations. T trajectories of length H, $\left\{(x_h, a_h)_{h\in[H],t\in[T]}\right\}$, obtained via *policy-induced* data collection with the uniform behavior(logging) policy $\rho \sim \mathcal{U}(\mathcal{A})$ (*no generative model!*):

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

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

Remark Use of fixed behavior policy is common to derive theoretical guarantees [Azizzadenesheli et al., 2016b, Azizzadenesheli et al., 2016a], and to accommodate practical *offline* RL applications [Levine et al., 2020]. [Xiao et al., 2022] showed that for passive data collection in batch RL, the uniform behavior policy is the best.

The Data

Observations. T trajectories of length H, $\{(x_h, a_h)_{h\in[H],t\in[T]}\}$, obtained via *policy-induced* data collection with the uniform behavior(logging) policy $\rho \sim \mathcal{U}(\mathcal{A})$ (*no generative model!*):

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

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

Remark Use of fixed behavior policy is common to derive theoretical guarantees [Azizzadenesheli et al., 2016b, Azizzadenesheli et al., 2016a], and to accommodate practical *offline* RL applications [Levine et al., 2020]. [Xiao et al., 2022] showed that for passive data collection in batch RL, the uniform behavior policy is the best.

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

Clustering Error

Clustering algorithms:

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

Clustering error: the number of misclassified contexts

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

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

2. Theoretical Results [Jedra et al., 2022]

Fundamental Lower Bound on Total Clustering Error

Theorem 1 Under certain regularity assumptions, *any algorithm that is β-locally better-than-random in Φ must satisfy*

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

where $I(\Phi) := -\frac{n}{TH} \log\left(\frac{1}{2\eta S n} \sum_{x \in \mathcal{X}} \exp\left(-\frac{TH}{n} I(x;\Phi)\right)\right)$.

Proof.
Utilizes the change-of-measure argument [Lai and Robbins, 1985]. □

- $I(x;\Phi)$ is an *information-theoretic* quantity that quantifies the difficulty of clustering for *each* context $x \in \mathcal{X}$.
- $I(x;\Phi)$ is defined through an optimization problem (ugly expressions!) [Jedra et al., 2022].

Latent State Decoding Algorithm

A two-phase algorithm with performance matching the lower bound up to some universal constants.

Latent State Decoding Algorithm

A two-phase algorithm with performance matching the lower bound up to some universal constants.

- Phase 1

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

Latent State Decoding Algorithm

A two-phase algorithm with performance matching the lower bound up to some universal constants.

- Phase 1

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

- 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)

Phase 1: Spectral Clustering

Phase 1

- Construction matrices of observations

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

- Trimming

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

where Γ_a corresponds to the remaining context in $\mathcal{X}$ after removing $\left\lfloor n \exp\left(-\frac{TH}{nA} \log\left(\frac{TH}{nA}\right)\right)\right\rfloor$ contexts with the highest number of visits.

Initial Error Upper Bound after Spectral Clustering

Theorem 2 *(Initial Spectral Clustering) If $TH = \omega(n)$, and $I(\Phi) > 0$, $\hat{f}_1$ outputted from the initial spectral clustering satisfies*

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

($I(\Phi) > 0$ means clustering is possible in an information-theoretic sense)

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, $\hat{f}_{\ell+1}(x) \leftarrow \text{argmax}_{j \in \mathcal{S}} \mathcal{L}^{(\ell)}(x, j)$ 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: $\hat{f}$

Final Error Upper Bbound after Iterative Likelihood Improvement

Theorem 3 (i) *(Iterative Likelihood Improvement) If* $TH = \omega(n)$, *and* $I(\Phi) > 0$, $\hat{f}$ *outputted from the iterative likeliehood improvement started from* $\hat{f}_1$ *satisfies*

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

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

- The form of $\mathcal{L}$ is inspired by the derivation of the lower bound (*Theorem 1*).
- 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 $TH - \frac{n\log(n)}{CI(x;\Phi)} = \omega_n(1)$ for all $x \in \mathcal{X}$.

3. Experiments
(Our Contributions)

Setting

- Consider a simple synthetic BMDP with $n = 100$, $S = 2$, $A = 3$ with the latent transition matrix of each action given as

$$P_1 = \begin{bmatrix} 1/2 - \epsilon & 1/2 + \epsilon \\ 1/2 + \epsilon & 1/2 - \epsilon \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1/2 + \epsilon & 1/2 - \epsilon \\ 1/2 - \epsilon & 1/2 + \epsilon \end{bmatrix}, \quad P_3 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix},$$

  where $\epsilon \in [0, 1/2)$ is the parameter determining the hardness of our BMDP instance and is pre-determined.
- We use the uniform behavior policy to generate the trajectories.

Setting

- Consider a simple synthetic BMDP with $n = 100$, $S = 2$, $A = 3$ with the latent transition matrix of each action given as

$$P_1 = \begin{bmatrix} 1/2 - \epsilon & 1/2 + \epsilon \\ 1/2 + \epsilon & 1/2 - \epsilon \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1/2 + \epsilon & 1/2 - \epsilon \\ 1/2 - \epsilon & 1/2 + \epsilon \end{bmatrix}, \quad P_3 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix},$$

  where $\epsilon \in [0, 1/2)$ is the parameter determining the hardness of our BMDP instance and is pre-determined.
- We use the uniform behavior policy to generate the trajectories.

It is *necessary* to consider all actions via concatenation in the initial spectral clustering!

Observation 1. Playing the third action does not provide any useful information for clustering, as the latent transition probabilities are all the same.

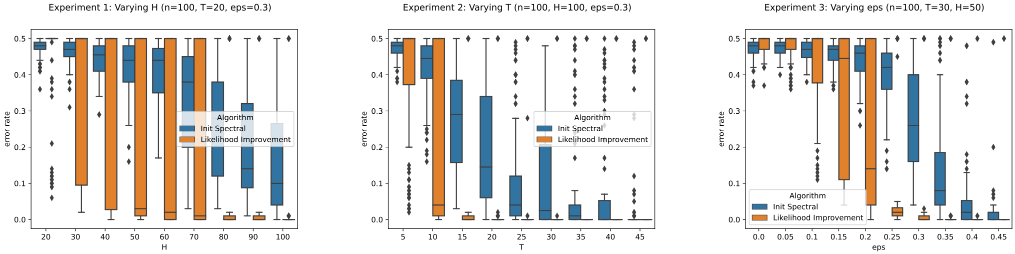
Observation 2. Considering the "marginalized" Markov chain, i.e., a single Markov chain with average transition matrix $\frac{1}{3}(P_1 + P_2 + P_3)$, renders clustering impossible.

Implementation

- The whole algorithm was implemented using Python
- For initial spectral clustering, we use the pyclustering [Novikov, 2019] for the K-median clustering
- All experiments were repeated 100 times to ensure statistical significance, and the results are shown via error bar/scatter plots

Experiment #1. Non-corrupted Setting

Vary H (length of episodes), T (number of episodes), and ϵ (difficulty of the BMDP instance)



(a) Varying H **(b)** Varying T **(c)** Varying ϵ

Figure 1: Sensitivity of clustering performance on various levels of T, H, ϵ.

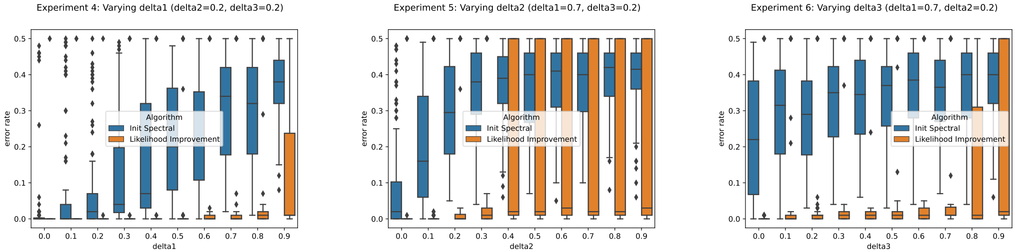
Experiment #2. Randomly corrupted Setting

- Is the algorithm robust to corruption in the given dataset?
- We fix $T = 30$, $H = 100$, and $\epsilon = 0.35$.

Experiment #2. Randomly corrupted Setting

- Is the algorithm robust to corruption in the given dataset?
- We fix $T = 30$, $H = 100$, and $\epsilon = 0.35$.

Vary δ_1 ($\delta_1 T$ trajectories corrupted), δ_2 ($\delta_2 H$ *contexts* corrupted), and δ_3 ($\delta_3 H$ *actions* corrupted)



(a) Varying δ_1 **(b)** Varying δ_2 **(c)** Varying δ_3

Figure 2: Sensitivity of clustering performance on various (random) corruption levels of $\delta_1, \delta_2, \delta_3$.

Some Observations

- A phase transition happening, from which **exact clustering** is observed
  - consistent with the Kesten-Stigum bound of clustering in binary SBM [Abbe, 2018], and even the asymptotic phase transition of BMDP [Jedra et al., 2022]
  - Difference between effect of T and H; can we (theoretically) quantify this in finite-sample regime?

Some Observations

- A phase transition happening, from which **exact clustering** is observed
  - consistent with the Kesten-Stigum bound of clustering in binary SBM [Abbe, 2018], and even the asymptotic phase transition of BMDP [Jedra et al., 2022]
  - Difference between effect of T and H; can we (theoretically) quantify this in finite-sample regime?

- Why the outliers?
  : the initial spectral clustering *sometimes* results in poor initialization for the likelihood improvement step.
  - Not contradictory to the results of [Jedra et al., 2022], which hold w.h.p. as $n \to \infty$.

4. Concluding Remarks

Concluding Remarks

Related work: All previous works provide experiments on only the downstream RL task (i.e., regret, value gap...etc) [Jiang et al., 2017, Dann et al., 2018, Du et al., 2019, Misra et al., 2020, Foster et al., 2021, Zhang et al., 2022].

Concluding Remarks

Related work: All previous works provide experiments on only the downstream RL task (i.e., regret, value gap...etc) [Jiang et al., 2017, Dann et al., 2018, Du et al., 2019, Misra et al., 2020, Foster et al., 2021, Zhang et al., 2022].

Our contributions: Preliminary empirical analyses of two-phase clustering algorithm [Jedra et al., 2022] for synthetic block MDP problems.

Concluding Remarks

Related work: All previous works provide experiments on only the downstream RL task (i.e., regret, value gap...etc) [Jiang et al., 2017, Dann et al., 2018, Du et al., 2019, Misra et al., 2020, Foster et al., 2021, Zhang et al., 2022].

Our contributions: Preliminary empirical analyses of two-phase clustering algorithm [Jedra et al., 2022] for synthetic block MDP problems.

Open problems:

- More memory-efficient clustering algorithm? (e.g., via random linear combination [Yun and Proutiére, 2016])
- Empirical and theoretical exploration to adaptive adversaries [Liu and Moitra, 2022] and methods to mitigate them [Yun and Proutiére, 2019, Tarbouriech et al., 2020].
- Beyond Block structure $\rightarrow$ Low Rank.

[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 linear structure in dimension $d = SA$:

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

Linear MDPs	$\leq$	Block MDPs	$\leq$	Low Rank 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:

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

References

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

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.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. (2016a).
Reinforcement Learning in Rich-Observation MDPs using Spectral Methods.
In *Proceedings of the 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. (2016b).
Reinforcement Learning of POMDPs using Spectral Methods.
In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 193–256, Columbia University, New York, New York, USA. PMLR.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2018).
On Oracle-Efficient PAC RL with Rich Observations.
In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. (2019).
Provably efficient RL with Rich Observations via Latent State Decoding.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR.

Foster, D., Rakhlin, A., Simchi-Levi, D., and Xu, Y. (2021).
Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective.
In Belkin, M. and Kpotufe, S., editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2059–2059. PMLR.

Jaksch, T., Ortner, R., and Auer, P. (2010).

Near-optimal Regret Bounds for Reinforcement Learning.
Journal of Machine Learning Research, 11(51):1563–1600.

Jedra, Y., Lee, J., Proutiére, A., and Yun, S.-Y. (2022).
Nearly Optimal Latent State Decoding in Block MDPs.
arXiv preprint arXiv:2208:08480.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017).
Contextual decision processes with low Bellman rank are PAC-learnable.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.

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

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020).
Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.
arXiv preprint arXiv:2005.01643.

Liu, A. and Moitra, A. (2022).
Minimax Rates for Robust Community Detection.
In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science*.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. (2020).
Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning.
In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6961–6971. PMLR.

Novikov, A. V. (2019).
PyClustering: Data Mining Library.
Journal of Open Source Software, 4(36):1230.

Tarbouriech, J., Shekhar, S., Pirotta, M., Ghavamzadeh, M., and Lazaric, A. (2020).
Active Model Estimation in Markov Decision Processes.
In Peters, J. and Sontag, D., editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR.

Xiao, C., Lee, I., Dai, B., Schuurmans, D., and Szepesvari, C. (2022).
The Curse of Passive Data Collection in Batch Reinforcement Learning.
In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8413–8438. PMLR.

Yun, S.-Y. and Proutiére, A. (2016).
Optimal Cluster Recovery in the Labeled Stochastic Block Model.
In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Yun, S.-Y. and Proutiére, A. (2019).
Optimal Sampling and Clustering in the Stochastic Block Model.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, X., Song, Y., Uehara, M., Wang, M., Agarwal, A., and Sun, W. (2022).
Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning Approach.

In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26517–26547. PMLR.