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
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

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

**Unknown** state dynamics and reward function
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.

[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?
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?
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

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

3. Experiments
   (Our Contributions)

4. Concluding Remarks

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

- $\mathcal{X}$ is the *observable* context space with $|\mathcal{X}| = n$
- $S$ is the *latent* state space with $|S| = S$
- $A$ is the action space with $|A| = A$
- $p$ 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 $X = \bigcup_{s \in S} f^{-1}(s)$, $f^{-1}(s) := \{x \in X: f(x) = s\}$.
A **Block MDP** [Du et al., 2019] is defined by $\Phi = (\mathcal{X}, S, A, p, q, f)$

- $\mathcal{X}$ is the *observable* context space with $|\mathcal{X}| = n$
- $S$ is the *latent* state space with $|S| = S$
- $A$ is the action space with $|A| = A$
- $p$ 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

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

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

$\Phi = (p, q, f)$ is **unknown** to the learner.
Example Trajectories of Block MDPs

\[ \begin{align*}
  q(x|s_3) & \quad \text{Transition probability from } s_3 \\
  p(s_3|s_2, a) & \quad \text{Transition probability from } s_2 \\
  x & \quad \text{Current state} \\
  s_1, s_2, s_3 & \quad \text{States of the MDP}
\end{align*} \]
Example Trajectories of Block MDPs

Model vs Observations
**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 $\rho \sim \mathcal{U}(A)$ (*no generative model!*):

<table>
<thead>
<tr>
<th></th>
<th>$(h = 1)$</th>
<th>$(h = 2)$</th>
<th>$\ldots$</th>
<th>$(h = H)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(t = 1)$</td>
<td>$(x_1^{(1)}, a_1^{(1)})$</td>
<td>$(x_2^{(1)}, a_2^{(1)})$</td>
<td>$\ldots$</td>
<td>$(x_H^{(1)}, a_H^{(1)})$</td>
</tr>
<tr>
<td>$(t = 2)$</td>
<td>$(x_1^{(2)}, a_1^{(2)})$</td>
<td>$(x_2^{(2)}, a_2^{(2)})$</td>
<td>$\ldots$</td>
<td>$(x_H^{(2)}, a_H^{(2)})$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(t = T)$</td>
<td>$(x_1^{(T)}, a_1^{(T)})$</td>
<td>$(x_2^{(T)}, a_2^{(T)})$</td>
<td>$\ldots$</td>
<td>$(x_H^{(T)}, a_H^{(T)})$</td>
</tr>
</tbody>
</table>

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!*):

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

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:

\[
\left( x_h^{(t)}, a_h^{(t)} \right)_{h \in [H], t \in [T]}
\]

Observations

\[ \rightarrow \]

Clustering algorithm

\[ \rightarrow \]

Decoding function

**Clustering error:** the number of misclassified contexts

\[
\mathcal{E}(\hat{f}) = \min_{\sigma} \bigcup_{s \in S} \hat{f}^{-1}(\sigma(s)) \setminus f^{-1}(s)
\]

\[
|\mathcal{E}(\hat{f})| = \min_{\sigma} \left| \bigcup_{s \in S} \hat{f}^{-1}(\sigma(s)) \setminus f^{-1}(s) \right|
\]
2. Theoretical Results
[Jedra et al., 2022]
Theorem 1 Under certain regularity assumptions, any algorithm that is $\beta$-locally better-than-random in $\Phi$ must satisfy

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

(1)

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

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].
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**

\[
\{(x_h^{(t)}, a_h^{(t)})_{t \in [T], h \in [H]}\} \rightarrow \text{Matrix estimation} \rightarrow (\hat{N}_a, \Gamma_a)_{a \in \mathcal{A}}
\]

\[
(\hat{N}_a, \Gamma_a)_{a \in \mathcal{A}} \rightarrow \text{S-rank approximation} \rightarrow (\hat{M}_a)_{a \in \mathcal{A}}
\]

\[
(\hat{M}_a)_{a \in \mathcal{A}} \quad (\hat{M}_a^\top)_{a \in \mathcal{A}} \rightarrow \text{Aggregation} \rightarrow \hat{M}
\]

\[
\hat{M} \rightarrow \ell_1\text{-weighted K-medians} \rightarrow \hat{f}_1
\]
Latent State Decoding Algorithm

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

- **Phase 1**

  $\{(x_{ht}^{(t)}, a_{ht}^{(t)})_{t \in [T], h \in [H]}\}$ → Matrix estimation → $(\hat{N}_a, \Gamma_a)_{a \in A}$

  $(\hat{N}_a, \Gamma_a)_{a \in A}$ → S-rank approximation → $(\hat{M}_a)_{a \in A}$

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

  $\hat{M}$ → $\ell_1$-weighted K-medians → $\hat{f}_1$

- **Phase 2**

  $\hat{f}_1$ → Iterative Likelihood Improvement → $\hat{f}$
Algorithm 1: Initial Spectral Clustering

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

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

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

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

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

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

**end**

$\hat{M} \leftarrow [(\hat{M}_1)^T \cdots (\hat{M}_A)^T \hat{M}_1 \cdots \hat{M}_A]$;

Normalize the rows of $\hat{M}$ by the $\ell_1$-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

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

- Trimming

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

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

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

($I(\Phi) > 0$ means clustering is possible in an information-theoretic sense)
Algorithm 2: Iterative Likelihood Improvement

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

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

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

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

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

**end**

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

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

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

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

- The form of $L$ 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 $\epsilon$ (difficulty of the BMDP instance)

<table>
<thead>
<tr>
<th>$H$</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.0</td>
</tr>
<tr>
<td>30</td>
<td>0.1</td>
</tr>
<tr>
<td>40</td>
<td>0.2</td>
</tr>
<tr>
<td>50</td>
<td>0.3</td>
</tr>
<tr>
<td>60</td>
<td>0.4</td>
</tr>
<tr>
<td>70</td>
<td>0.5</td>
</tr>
<tr>
<td>80</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0.0</td>
</tr>
<tr>
<td>100</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$T$</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>0.1</td>
</tr>
<tr>
<td>15</td>
<td>0.2</td>
</tr>
<tr>
<td>20</td>
<td>0.3</td>
</tr>
<tr>
<td>25</td>
<td>0.4</td>
</tr>
<tr>
<td>30</td>
<td>0.5</td>
</tr>
<tr>
<td>35</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.0</td>
</tr>
<tr>
<td>45</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.15</td>
<td>0.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>0.0</td>
</tr>
<tr>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>0.45</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Figure 1:** Sensitivity of clustering performance on various levels of $T$, $H$, $\epsilon$. 
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 $\delta_1$ ($\delta_1 T$ trajectories corrupted), $\delta_2$ ($\delta_2 H$ contexts corrupted), and $\delta_3$ ($\delta_3 H$ actions corrupted).

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?
• 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].

**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 Proutièere, 2016])
- Empirical and theoretical exploration to adaptive adversaries [Liu and Moitra, 2022] and methods to mitigate them [Yun and Proutièere, 2019, Tarbouriech et al., 2020].
- Beyond Block structure $\rightarrow$ Low Rank.
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.

**Open problems:**

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

\[
\phi(x, a) = e(f(x), a), \quad \mu(x')_{(s,a)} = q(x'|f(x))p(f(x')|s, a).
\]

Linear MDPs \( \leq \) Block MDPs \( \leq \) Low Rank MDPs

- \( \mu \) is unknown
- \( \phi \) is unknown
- \( \mu \) is unknown
- \( \phi \) is unknown
- \( \mu \) is unknown
- \( \phi \) is unknown
- \( \phi \in \mathcal{F}_{BM}\)
- \( \phi \in \mathcal{F} \)

Linear structure in RL:

\[
\begin{align*}
\underbrace{\text{Linear MDP}}_{P(x'|x, a) = \phi(x, a)\top \mu(s')} + \underbrace{\text{Structured rewards}}_{r(x, a) = \phi(x, a)\top \theta} & \implies \underbrace{\text{Q-function is linear}}_{Q^\pi(x, a) = \phi(x, a)\top \xi^\pi} \\
\end{align*}
\]
References


**On Oracle-Efficient PAC RL with Rich Observations.**


**Provably efficient RL with Rich Observations via Latent State Decoding.**


**Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective.**

Near-optimal Regret Bounds for Reinforcement Learning.

Nearly Optimal Latent State Decoding in Block MDPs.

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.

Asymptotically Efficient Adaptive Allocation Rules.

Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open
Problems.
**Minimax Rates for Robust Community Detection.**  
In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science*.

**Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning.**  

**PyClustering: Data Mining Library.**  

**Active Model Estimation in Markov Decision Processes.**  

