
Nearly Optimal Latent State Decoding in Block MDPs

Yassir Jedra^{*1}

Junghyun Lee^{*2}

Alexandre Proutière¹

Se-Young Yun²

¹School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

²Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea

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

Abstract

We consider the problem of model estimation in episodic Block MDPs. In these MDPs, the decision maker has access to rich observations or contexts generated from a small number of latent states. We are interested in estimating the latent state decoding function (the mapping from the observations to latent states) based on data generated under a fixed behavior policy. We derive an information-theoretical lower bound on the error rate for estimating this function and present an algorithm approaching this fundamental limit. In turn, our algorithm also provides estimates of all the components of the MDP. We apply our results to the problem of learning near-optimal policies in the reward-free setting. Based on our efficient model estimation algorithm, we show that we can infer a policy converging (as the number of collected samples grows large) to the optimal policy at the best possible asymptotic rate. Our analysis provides necessary and sufficient conditions under which exploiting the block structure yields improvements in the sample complexity for identifying near-optimal policies. When these conditions are met, the sample complexity in the minimax reward-free setting is improved by a multiplicative factor n , where n is the number of contexts.

1 Introduction

In Reinforcement Learning, leveraging succinct representations of the system state is empirically known to considerably accelerate the search for near-optimal policies, see, e.g., [Laskin et al. \(2020\)](#); [Guo et al. \(2020\)](#); [Stooke et al. \(2021\)](#) and references therein. The design of RL algorithms

with provable performance guarantees that learn and exploit such representations remains largely open.

In this paper, we address this challenge for a specific class of models, namely episodic Block MDPs (BMDPs). In these MDPs, introduced in [Krishnamurthy et al. \(2016\)](#) and since then widely studied and motivated (see [Du et al. \(2019\)](#); [Zhang et al. \(2022b\)](#) and references therein), the decision maker has, in each round, access to rich observations, referred to as *contexts*, generated from a small number of *latent* states. More precisely, to each context x corresponds a unique latent state $s = f(x)$ where f is referred to as the latent state decoding function. If the decision maker selects control action a , the system moves from context x to context y with probability $q(y|s')p(s'|s, a)$ where $s = f(x)$ and $s' = f(y)$. The emission distributions q , the latent state transition rates p and the decoding function f are initially unknown. Intuitively, if we could learn the latent state decoding function, we would be able to efficiently summarize the environment with a low-dimensional state space, and hence devise RL algorithms that can quickly learn near-optimal policies. In this work, our objective is to substantiate this intuition. Specifically, we aim at answering the following questions: (i) How fast can we learn the latent state decoding function, as well as p and q ? (ii) What optimal gains in terms of sample complexity can we expect when exploiting the existing yet initially unknown structure?

Most existing studies on representation learning in BMDPs ([Jiang et al., 2017](#); [Dann et al., 2018](#); [Du et al., 2019](#); [Misra et al., 2020](#); [Foster et al., 2021](#); [Zhang et al., 2022b](#)) address these questions using the function approximation framework. Specifically, the latent state decoding function or some of its functionals is assumed to belong to a parametrized class of functions. This assumption somehow introduces an additional structure in the BMDP. Now, the performance guarantees obtained for algorithms learning within this class of functions depend on the cardinality or complexity of this class. For classes with moderate complexity, these algorithms are able to learn near-optimal policies quickly. However, as it turns out, in absence of any prior knowledge of the class of functions that could contain the true latent state decoding function, these algorithms cannot

^{*}Equal contributions. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

exploit the block structure (the class becomes too complex and the guarantees are not better than those achieved in plain tabular MDPs – refer to §6 for details). Moreover, to exploit the function approximation framework, the algorithms proposed in the aforementioned work rely on strong computational oracles (e.g., ERM, MLE). In this paper, we depart from the function approximation framework and analyze scenarios where no additional structure is imposed on the BMDP. For these scenarios, we wish to understand how to optimally exploit the block structure alone to speed up the convergence of learning algorithms.

Contributions. Our main results concern the estimation of the latent state decoding function in BMDPs using data generated from a fixed behavior policy. We apply our results to the problem of learning near-optimal policies in the reward-free learning setting, where the reward function is revealed after the data has been collected. More precisely, our contributions are as follows.

1. *Learning the latent state decoding function*

(a) We first derive information-theoretical lower bounds on the latent state decoding error rates. When the BMDP $\Phi = (p, q, f)$ satisfies some regularity assumptions, we establish that the expected number $\mathbb{E}_\Phi[|\mathcal{E}|]$ of contexts for which we do not infer the corresponding latent state correctly must be (roughly) greater than $n \exp(-\frac{TH}{n} I(\Phi))$ where T is the number of episodes in the collected data, H the duration of an episode, n the size of the context space and $I(\Phi)$ a well-defined and non-negative rate function. This lower bound provides conditions on T , H , n , and Φ under which one can hope to estimate the block structure accurately.

(b) We present a structure estimation algorithm whose performance approaches our lower bound. Its design is inspired by spectral methods typically used for inference tasks in the Stochastic Block Model (Abbe, 2018), Degree-Corrected Block Model (Gao et al., 2018) and Block Markov Chain (Sanders et al., 2020). To analyze its performance, we develop new concentration inequalities for functionals of Markov chains with *restarts* (restarts are needed to model the episodic nature of the MDP).

2. *Learning near-optimal policies in the reward-free setting*

In the reward-free setting, the performance of an RL algorithm is quantified through $\Delta(\mathcal{R}) = \sup_{r \in \mathcal{R}} \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r))$ where $V^*(r)$ and $V^{\hat{\pi}_r}(r)$ are the values of the optimal policy and of that of the policy $\hat{\pi}_r$ returned by the algorithm for the (possibly context-dependent) rewards r , and where \mathcal{R} is a given class of reward functions.

(a) We present lower bounds on the number of observed episodes required to infer ϵ -optimal policies or more precisely to ensure that $\Delta(\mathcal{R}) \leq \epsilon$. These lower bounds hold even if the data can be collected in an adaptive manner. In the *minimax* setting where \mathcal{R} contains all possible bounded reward functions, this sample complexity lower bound is

$TH = \Omega(\frac{n}{\epsilon^2})$. In the *reward-specific* setting where \mathcal{R} reduces to a single function r , the lower bound becomes $TH = \Omega(n \log(\frac{1}{\epsilon}) + \frac{1}{\epsilon^2})$ (the first term corresponds to the data required to learn the block structure accurately, and the second term to the data required to learn an ϵ -optimal policy given the block structure). These lower bounds quantify the gains achieved when exploiting the block structure: in the minimax and reward-specific settings, without any structure, these bounds would be $\Omega(\frac{n^2}{\epsilon^2})$ and $\Omega(\frac{n}{\epsilon^2})$, respectively.

(b) We study the performance of RL algorithms built upon our structure estimation algorithm using data generated with a fixed behavior policy. We show that these algorithms learn ϵ -optimal policies at the fastest rate possible. More precisely, when $\epsilon = o(1)$ (as n grows large), their sample complexities match (up to logarithmic factors) our lower bounds in the minimax and reward-specific settings.

2 Models and Objectives

2.1 Episodic block MDPs

A Block MDP (BMDP) Φ is defined by $(\mathcal{S}, \mathcal{X}, \mathcal{A}, \mu, p, q, f, H)$ where \mathcal{S} , \mathcal{X} , \mathcal{A} denote the sets of hidden latent states, observed contexts, and actions, respectively. Let S, n, A be their cardinalities. $H \geq 2$ is the length of each episode and μ denotes the distribution of the initial context $x_1 \in \mathcal{X}$. p represents the latent state time-homogenous dynamics: $p(s'|s, a)$ is the probability of moving from latent state s to s' given that the learner selects action a . The *decoding function* f maps each context x to a unique latent state $f(x) \in \mathcal{S}$. When the system is in the latent state s , the learner observes a context x drawn according to the *emission distribution* $q(\cdot|s)$ with support $f^{-1}(s) \subset \mathcal{X}$. Naturally, the decoding function f induces a partition of \mathcal{X} into *clusters* indexed by s , and is initially unknown to the learner. To simplify the notation, we write $\Phi = (p, q, f)$ to specify the dynamics and the structure of the BMDP. We further denote \mathbb{P}_Φ and \mathbb{E}_Φ as the probability measure and the expectation for observations generated under the model Φ , respectively.

2.2 Latent state decoding and model estimation

In this paper, our main objective is an offline model estimation: given some dataset, the learner wishes to estimate the latent state decoding function f and the transition dynamics p and q of the BMDP.

The learner is given a dataset of transitions: $\mathcal{D} = \{(x_1^{(t)}, a_2^{(t)}, \dots, x_{H-1}^{(t)}, a_{H-1}^{(t)}, x_H^{(t)})_{t \in [T]}\}$. The dataset has been generated using the uniform behavior policy $\rho \sim \mathcal{U}(\mathcal{A})$ ¹. The use of passive (also referred to as memoryless)

¹This assumption can be relaxed. For further details, we refer the reader to Appendix K.

behavior policy is common to derive theoretical guarantees (Azizzadenesheli et al., 2016a,b), and more importantly, to accommodate practical offline RL applications (such as healthcare) where active data collection strategies are impractical or dangerous (Levine et al., 2020). Also note that recently, Xiao et al. (2022) showed for passive data collection in offline RL, the uniform policy is the best behavior policy. An additional important remark is that we consider policy-induced data collection (Xiao et al., 2022), i.e., we do not assume access to any generative model.

The objective of the learner is to build \hat{f} , an estimate of the latent decoding function. Since the latent states are not observed, \hat{f} is only defined up to a permutation of these states, as in any clustering task (Abbe, 2018; Yun and Proutière, 2019; Sanders et al., 2020). The accuracy of \hat{f} is measured through the cardinality of the set \mathcal{E} of misclassified contexts defined as $\mathcal{E} = \mathcal{E}_\nu$, where

$$\mathcal{E}_\sigma = \bigcup_{s \in \mathcal{S}} \hat{f}^{-1}(\sigma(s)) \setminus f^{-1}(s), \quad \sigma \in \Upsilon(\mathcal{S}), \quad (1)$$

$\nu \in \operatorname{argmin}_{\sigma \in \Upsilon(\mathcal{S})} |\mathcal{E}_\sigma|$, and $\Upsilon(\mathcal{S})$ denotes the set of permutations over \mathcal{S} . We say that the clusters are recovered *asymptotically accurately* if $\mathbb{E}[|\mathcal{E}|] = o(n)$ and *asymptotically exactly* if $\mathbb{E}[|\mathcal{E}|] = o(1)$. The lower and upper bounds of the error rates have been derived for various block models. Such clustering guarantees for BMDPs have *not* been studied, and they are highly nontrivial for the following reasons. First, because we deal with Markovian data, standard concentration techniques used in clustering Stochastic Block Models (SBMs) and their extensions (Abbe, 2018; Yun and Proutière, 2019; Gao et al., 2018) do not apply. Then, we have controllable actions that can change the transition probability, even between the same pair of states, which prohibits us from using the clustering algorithms of Block Markov Chains (BMCs; Sanders et al. (2020)) directly.

Using \hat{f} , the latent state transition probabilities and emission distribution are estimated using the *plug-in estimators*, which are very effective despite their simplicity (Duan et al., 2020; Ren et al., 2021; Xiao et al., 2021). Their precise forms are provided in Section 4.2. The accuracies of these estimates, \hat{p} and \hat{q} , are assessed through the ℓ_1 -distances: for all $(s, s', a) \in \mathcal{S}^2 \times \mathcal{A}$, $\|\hat{p}(\nu(\cdot)|\nu(s), a) - p(\cdot|s, a)\|_1$, and $\|\hat{q}(\cdot|\nu(s)) - q(\cdot|s)\|_1$. Without loss of generality, for simplicity, we assume that $\nu = \text{Id}$, i.e., $\nu(s) = s$ for all s .

2.3 Assumptions

We impose the following assumptions, which allow us to invoke and translate theoretical tools and ideas from various literature on clustering in block models (Abbe, 2018; Gao et al., 2018; Yun and Proutière, 2019; Sanders et al., 2020):

Assumption 1 (Asymptotics). S, A, p are independent of n

Assumption 2 (Linear context cluster sizes). *The sizes of the clusters grow linearly with n . Specifically, there exists*

$\alpha = (\alpha_s)_{s \in \mathcal{S}}$ independent of n such that (i) $\alpha_s > 0$ for all $s \in \mathcal{S}$, and (ii) $\sum_s \alpha_s = 1$ and $|f^{-1}(s)| = \alpha_s n$.

Assumption 3 (η -regularity). *There exists $\eta > 1$, independent of n , such that:*

- (i) $\max_{s_1, s_2 \in \mathcal{S}} \frac{\alpha_{s_1}}{\alpha_{s_2}} \leq \eta$,
- (ii) $\max_{a \in \mathcal{A}} \max_{s_1, s_2, s_3 \in \mathcal{S}} \left\{ \frac{p(s_2|s_1, a)}{p(s_3|s_1, a)}, \frac{p(s_1|s_2, a)}{p(s_1|s_3, a)} \right\} \leq \eta$,
- (iii) $\max_{s \in \mathcal{S}} \max_{x, y \in f^{-1}(s)} \frac{q(x|s)}{q(y|s)} \leq \eta$.

Assumption 4 (Uniform initial context). $\mu \sim \mathcal{U}(\mathcal{X})$.

Assumptions 1, 2, and 3(i), which express some kind of homogeneity in the number and sizes of the clusters, are imposed in the majority of the clustering literature, see Abbe (2018); Gao et al. (2018); Yun and Proutière (2019); Sanders et al. (2020). Such homogeneity is required to establish theoretical guarantees for clustering tasks.

Assumption 3(ii) has been used for the clustering task in BMCs (Sanders et al., 2020) to ensure a minimum level of separability between clusters. A similar assumption is made in SBMs (Yun and Proutière, 2016, 2019; Gao et al., 2018). In Appendix L, we discuss how to relax Assumption 3(ii) and only assume that the transition kernel p is aperiodic and communicating.

Assumption 3 (iii) has been considered in the Degree Corrected Block Model (DCBM; Gao et al. (2018)), where the heterogeneity via degree-correction of each cluster was allowed, but limited.

Assumption 4 is not crucial but simplifies many statements in the proofs. It ensures that all contexts can be reached with positive probability. We can replace it with $\mu(x) = \Theta(\frac{1}{n})$, and given Assumption 3, even remove it, as long as $H \geq 3$.

3 Fundamental Lower Bound of Latent State Decoding

To derive instance-specific lower bounds on the clustering error rates, we leverage change-of-measure arguments (Lai and Robbins, 1985) where we pretend that the observations are generated by a BMDP model Ψ that is slightly different than the true model Φ . To obtain instance-specific lower bounds, we always need to restrict the attention to a class of *adaptive* algorithms. Indeed, an algorithm always returning $\hat{f} = f$ would have no error for the BMDP $\Phi = (p, q, f)$ but would fail for any other BMDP with $\tilde{\Phi} = (p, q, \tilde{f})$ with \tilde{f} different than f . Here, we restrict our attention to the wide class of so-called β -locally better-than-random clustering algorithms.

Definition 1 (β -locally better-than-random clustering algorithms). *A clustering algorithm is β -locally better-than-random for the BMDP $\Phi = (p, q, f)$ if for any BMDP $\tilde{\Phi} =$*

$(p, \tilde{q}, \tilde{f})$ with $\max_{y: f(y)=\tilde{f}(y)} \max_s |q(y|s) - \tilde{q}(y|s)| \leq \beta$ and $|\{y \in \mathcal{X} : f(y) \neq \tilde{f}(y)\}| \leq 1$, for all $x \in \mathcal{X}$, x is misclassified with probability at most $1 - \frac{1}{5}$.

Considering locally better-than-random algorithms is not restrictive, as algorithms with reasonable performance should indeed adapt to the BMDP they are facing, or in other words, such algorithms behave similarly when the considered BMDP model changes very slightly.

We derive a lower bound for the clustering error rate of *each context* $x \in \mathcal{X}$ characterized by a rate function $I(x; \Phi)$ defined as follows. Consider an alternate BMDP $\Psi^{(x,j)} = (p, \tilde{q}, \tilde{f})$ obtained from the true model Φ by changing the latent state of context x from $f(x)$ to j , for $j \neq f(x)$. The alternate latent decoding function \tilde{f} is defined as $\tilde{f}(y) = f(y)\mathbb{1}_{y \neq x} + j\mathbb{1}_{y=x}$ for all $y \in \mathcal{X}$. The alternate emission function \tilde{q} remains unchanged for latent states different than $f(x)$ and j : for any $s \neq f(x), j$, $\tilde{q}(\cdot|s) = q(\cdot|s)$. Under $\Psi^{(x,j)}$, x has an emission probability $\tilde{q}(x|j) = cq(x|i)$ for $c > 0$ to be chosen later, and we define for all $y \in f^{-1}(i)$, $\tilde{q}(y|i) = q(y|i)/(1 - q(x|i))$ and for all $y \in f^{-1}(j)$, $\tilde{q}(y|j) = q(y|j)(1 - cq(x|i))$. Note that with this definition, $\tilde{q}(\cdot|s)$ is a well-defined probability distribution over $\tilde{f}^{-1}(s)$, for all $s \in \mathcal{S}$.

We now define the rate function $I(x; \Phi)$:

$$I(x; \Phi) = \min_{j \neq f(x)} \inf_{c > 0} I_j(x; c, \Phi), \quad (2)$$

where $I_j(x; c, \Phi)$ is (almost)² equal to $\frac{H}{n}$ times the expected log-likelihood ratio of the observations over one episode made under $\Psi^{(x,j)}$ and Φ , and is ∞ when c is such that Ψ_j is not well-defined. Importantly, the rate function $I(x; \Phi)$ does not scale with n , and we can identify necessary and sufficient conditions for the *clusterability* of x , i.e., for $I(x; \Phi) > 0$. In Appendix D.3.4, we provide illustrative examples of $I(x; \Phi)$ for clusterable and non-clusterable BMDPs.

In the next theorem, whose proof is presented in Appendix D, we present our lower bounds for the clustering error rates.

Theorem 1. *Let Φ be a BMDP satisfying Assumptions 1,2,3,4. Consider a clustering algorithm that is β -locally better-than-random for Φ with $\beta \geq \frac{2S\eta^2}{n}$, when applied to the data gathered over T episodes, each of length H , using ρ . Then, for all $x \in \mathcal{X}$,*

$$\mathbb{P}_\Phi[x \in \mathcal{E}] \geq \frac{1}{2\eta S} \exp\left(-\frac{TH}{n} I(x; \Phi)(1 + o(1))\right).$$

As a consequence, we have that

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

²Refer to Appendix D for the exact expression of $I_j(x; c, \Phi)$.

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

The above theorem provides necessary conditions for the existence of algorithms classifying a given context x asymptotically accurately, i.e., $\mathbb{P}_\Phi[x \in \mathcal{E}] \rightarrow 0$ as $n \rightarrow \infty$. These conditions are $I(x; \Phi) > 0$ and $TH = \omega(n)$. In Appendix D.3, we show that $I(x; \Phi) = 0$ if and only if there exists $j \neq f(x)$ and $c > 0$ such that for all (s, a) , $p(s|f(x), a) = p(s|j, a)$ and $p(f(x)|s, a) = cp(j|s, a)$. Equivalently, these conditions state that the transition rates to and out of the latent states $f(x)$ and j are identical, in which case, of course, we cannot determine whether x belongs to $f(x)$ or j . We note that, thanks to Assumption 3, if $I(x; \Phi) > 0$, then $I(y; \Phi) > 0$ for all y satisfying $f(y) = f(x)$. In words, the condition for accurate classification of a context only depends on its corresponding latent state.

We further deduce, from Theorem 1, necessary conditions for the existence of algorithms recovering the clusters asymptotically accurately, i.e., $\mathbb{E}_\Phi[|\mathcal{E}|] = o(n)$. These conditions are $I(\Phi) > 0$ and $TH = \omega(n)$. Note that $I(\Phi) > 0$ if and only if for all x such that $I(x; \Phi) > 0$ (we need to classify each context asymptotically accurately if we wish to do the same over all the latent states). Similarly, necessary conditions for the existence of an asymptotically exact clustering algorithm, i.e., $\mathbb{E}_\Phi[|\mathcal{E}|] = o(1)$, are $I(\Phi) > 0$ and $TH - \frac{n \log(n)}{I(\Phi)} = \omega(1)$. In particular, TH must be larger than $n \log(n)$. In the *critical* regime where $TH = n \log(n)$, the necessary condition for exact recovery is $I(\Phi) > 1$. These regimes are reminiscences of the clustering guarantees for SBMs (Abbe, 2018; Yun and Proutière, 2019), DCBMs (Gao et al., 2018), and BMCs (Sanders et al., 2020).

Remark 1. *We can extend Theorem 1 to scenarios where the behavior policy p is adaptive, in which we just obtain a different rate function $I'(\Phi) > I(\Phi)$; see Appendix K.*

4 Latent State Decoding and Model Estimation Algorithms

In this section, we present our algorithms to recover the latent state decoding function f as well as the parameters (p, q) defining the BMDP dynamics. We also analyze their performance and sketch the elements of the analyses.

4.1 Algorithms

The latent decoding function f of the given BMDP is estimated using a two-step procedure. The first step consists in leveraging spectral methods to obtain rough estimates of the latent state decoding function or clusters and is described in Algorithm 1. The second step iteratively improves the clusters using log-likelihoods of each cluster and is presented in Algorithm 2. At the end of the second step, we get accurate estimates of the latent state decoding function f , from

Algorithm 1: Initial Spectral Clustering

Input: $\mathcal{D} = \{(x_1^{(t)}, a_2^{(t)}, \dots, x_{H-1}^{(t)}, a_{H-1}^{(t)}, x_H^{(t)})_{t \in [T]}\}$

- 1 **for** $a \in \mathcal{A}$ **do**
- 2 **for all** $(x, y) \in \mathcal{X}^2$,

$$\hat{N}_a(x, y) \leftarrow \sum_{t,h} \mathbb{1}[(x_h^{(t)}, a_h^{(t)}, x_{h+1}^{(t)}) = (x, a, y)]$$
- 3 $\Gamma_a \leftarrow \mathcal{X}$ after removing $\lfloor n \exp(- (TH/nA) \log(TH/nA)) \rfloor$ contexts with the highest $\hat{N}_a(x) = \sum_y \hat{N}_a(x, y)$;
- 4 $\hat{N}_{a, \Gamma_a} \leftarrow (\hat{N}_a(x, y) \mathbb{1}_{\{(x,y) \in \Gamma_a\}})_{x,y \in \mathcal{X}}$;
- 5 $\hat{M}_a \leftarrow$ rank- S approximation of \hat{N}_{a, Γ_a} ;
- 6 **end**
- 7 $\hat{M} \leftarrow [(\hat{M}_1)^\top \ \dots \ (\hat{M}_A)^\top \ \hat{M}_1 \ \dots \ \hat{M}_A]$;
- 8 Normalize the rows of \hat{M} by the ℓ_1 -norm;
- 9 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)

which we use the plug-in estimator to estimate the transition dynamics p and q .

Our two-step procedure is inspired by the clustering algorithm designed for DCBMs (Gao et al., 2018) and BMCs (Sanders et al., 2020), but there are significant differences mainly due to the fact that the data consists of episodes of a controlled Markov chain (the selected actions do matter) and to the presence of non-uniform emission distributions.

Initial spectral clustering step. As most clustering algorithms with optimality guarantees (von Luxburg, 2007), our procedure starts with a spectral decomposition of matrices built from the data \mathcal{D} , observed from the underlying BMDP model. We wish to distinguish whether a context belongs to one cluster or another whenever this is statistically possible, from the observations corresponding to at least one of the actions. Hence, for each action $a \in \mathcal{A}$, we build from the data an observation matrix \hat{N}_a and inspect its spectral properties. We compute the rank- S approximation \hat{M}_a of each \hat{N}_a after *trimming*³. We then *concatenate*⁴ these matrices in \hat{M} , a $(n \times 2nA)$ matrix, whose x -th row contains all the information we have for context x (transitions starting from

³In the sparse regime ($TH = o(n \log n)$), trimming is necessary to remove contexts with too many observations that would perturb the empirical spectral distribution (Feige and Ofek, 2005; Keshavan et al., 2010a; Yun and Proutière, 2016; Sanders et al., 2020). In the dense regime ($TH = \Omega(n \log n)$), such a trimming procedure is not necessary (Sanders and Senen-Cerda, 2023).

⁴For large BMDPs, concatenation may not be feasible due to memory or computational limitations. One possible workaround is to utilize random linear combination, as done in Yun and Proutière (2016) for clustering labeled SBMs; this is left as future work.

Algorithm 2: Iterative Likelihood Improvement

Input: Initial cluster estimates \hat{f}_1 and $\{\hat{N}_a\}_{a \in \mathcal{A}}$

- 1 **for** $\ell = 1$ **to** $L = \lfloor \log(nA) \rfloor$ **do**
- 2 **for all** $(s, j, a) \in \mathcal{S}^2 \times \mathcal{A}$,

$$\hat{p}_\ell(s|j, a) \leftarrow \frac{\hat{N}_a(\hat{f}_\ell^{-1}(j), \hat{f}_\ell^{-1}(s))}{\hat{N}_a(\hat{f}_\ell^{-1}(j), \mathcal{X})}, \quad (3)$$
- 3 $\hat{p}_\ell^{bwd}(s, a|j) \leftarrow \frac{\hat{N}_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(j))}{\sum_{\bar{a} \in \mathcal{A}} \hat{N}_{\bar{a}}(\mathcal{X}, \hat{f}_\ell^{-1}(j))}, \quad (4)$
- 3 **for all** $x \in \mathcal{X}$, $\hat{f}_{\ell+1}(x) \leftarrow \operatorname{argmax}_{j \in \mathcal{S}} \mathcal{L}^{(\ell)}(x, j)$,
- 4 **where**

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

Output: $\hat{f} = \hat{f}_{L+1}$ (final estimate of the decoding function)

x and ending in x when selecting action a for any possible a). Applying the S -median algorithm to the rows of \hat{M} yields our initial cluster estimate encoded in \hat{f}_1 .

Iterative likelihood improvement step. This step takes as input our initial cluster estimates \hat{f}_1 as well as the counters of transitions $\hat{N} = (\hat{N}_a(x, y))_{a,x,y}$ (see Algorithm 1). For any $X, Y \subseteq \mathcal{X}$, we denote by $\hat{N}_a(X, Y)$ the number of observed transitions, when action a is chosen, from any context $x \in X$ to any context $y \in Y$. In the ℓ -th iteration, we first use our current cluster estimates \hat{f}_ℓ and \hat{N} to estimate the empirical latent transition $\hat{p}_\ell(s|j, a)$ and the empirical backward latent transition $\hat{p}_\ell^{bwd}(s, a|j)$ (\hat{p}^{bwd} is in a sense similar to the backward probability vector used in Du et al. (2019)). Using $\hat{p}_\ell, \hat{p}_\ell^{bwd}, \hat{f}_\ell$, and \hat{N}_a 's, we compute for each context x the log-likelihood $\mathcal{L}^{(\ell)}(x, j)$ of the event that x is assigned to the latent state j . x is then re-assigned to the cluster with the highest $\mathcal{L}^{(\ell)}(x, j)$, and we update $\hat{f}_{\ell+1}$ accordingly. We perform $L = \lfloor \log(nA) \rfloor$ iterations and then output the final estimated decoding function $\hat{f} = \hat{f}_{L+1}$.

Estimating p and q . For simplicity of the analysis, we decouple the estimation of f and that of p and q . Precisely, we first estimate \hat{f} via Algorithms 1 and 2 using the first $\lfloor T/2 \rfloor$ episodes. We then use the remaining $\lfloor T/2 \rfloor$ episodes to get the estimates \hat{p} and \hat{q} using the plug-in estimators: for each $(s, s', a) \in \mathcal{S}^2 \times \mathcal{A}$ and $x \in \hat{f}^{-1}(s)$,

$$\begin{cases} \hat{p}(s'|s, a) = \frac{\hat{N}_a(\hat{f}^{-1}(s), \hat{f}^{-1}(s'))}{\hat{N}_a(\hat{f}^{-1}(s), \mathcal{X})}, \\ \hat{q}(x|s) = \frac{\sum_{t=1}^{\lfloor T/2 \rfloor} \sum_{h=1}^H \mathbb{1}[(x_h^{(t)}, \hat{f}(x_h^{(t)})) = (x, s)]}{\sum_{t=1}^{\lfloor T/2 \rfloor} \sum_{h=1}^H \mathbb{1}[\hat{f}(x_h^{(t)}) = s]} \end{cases} \quad (5)$$

If the denominator in some of the above fractions is 0, then

the corresponding estimate is set to be 0. Here, we leverage \hat{f} (from the first half of the data) but use the second half of the data to *redefine* the number of transitions $\hat{N}_a(x, y)$ for all (x, y, a) .

4.2 Performance guarantees

The following theorem, whose proof is deferred to Appendix F, provides a performance guarantee for Algorithm 1, the initial spectral clustering algorithm. It states that under the necessary conditions for asymptotically accurate cluster recovery, the algorithm indeed misclassifies only a vanishing proportion of contexts.

Theorem 2. *Assume that $TH = \omega(n)$ and $I(\Phi) > 0$. Then after Algorithm 1, the following holds w.h.p.⁵:*

$$\frac{|\mathcal{E}|}{n} = \mathcal{O}\left(\sqrt{\frac{nSA}{TH}}\right).$$

The next theorem, whose proof is presented in Appendix G, provides the clustering error rate guarantee after the iterative likelihood improvement steps (Algorithm 2), as well as the estimation error upper bounds for \hat{p} and \hat{q} :

Theorem 3. *Assume that $TH = \omega(n)$ and $I(\Phi) > 0$. Then, after Algorithm 2,*

(i) *there exists a universal constant $C = \text{poly}(\eta) > 0$ such that the following holds w.h.p.:*

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

(ii) *Let us denote $E_p = \|\hat{p}(\cdot|s, a) - p(\cdot|s, a)\|_1$ and $E_q = \|\hat{q}(\cdot|s) - q(\cdot|s)\|_1$. For all $(s, s', a) \in \mathcal{S}^2 \times \mathcal{A}$, we have w.h.p.,*

$$E_p = \mathcal{O}\left(\sqrt{\frac{S^3 A^2 \log(nSA)}{TH}} + \frac{SA|\mathcal{E}|}{n}\right),$$

$$E_q = \mathcal{O}\left(\sqrt{\frac{Sn}{TH}} + \frac{S|\mathcal{E}|}{n}\right).$$

Theorem 3 (i) refines the result of Theorem 2 and shows that our two-step procedure approaches the optimal recovery rate identified in Theorem 1 up to the constant C in the exponential. Moreover, it indicates that our algorithm exactly recovers the clusters if $TH - \frac{n \log(n)}{CI(x; \Phi)} = \omega(1)$ for all x and $I(\Phi) > 0$.

4.3 Sketches of the proofs

The proofs of Theorems 2 and 3 rely on two concentration inequalities presented in Appendix E and F. The first inequality is a novel Bernstein-type inequality for functionals of

Markov chains with restarts (to account for the episodic nature of the MDP). This contrasts with existing concentration results for Markov chains that concern a single trajectory of the chain starting in a stationary regime (Avrachenkov et al., 2013, 2018; Paulin, 2015), which we do not assume here. The second inequality characterizes how the observation matrix \hat{N}_a for a given action a concentrates around its mean. The main challenge for the analysis of \hat{N}_a stems from the Markovian dependence in its entries, and is addressed using similar techniques as in Sanders et al. (2020); Sanders and Senen-Cerda (2023).

5 Reward-Free RL in BLock MDPs

5.1 Setting and objectives

A policy π for the BMDP Φ is a collection of H mappings $\pi_h : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{A})$ ($\mathcal{P}(\mathcal{A})$ is the set of probability distributions on \mathcal{A}). We denote by $\pi_h(a|x)$ the probability of selecting a when the context x is observed in stage h . The value of a policy π is defined as the expected reward accumulated over an episode $V^\pi(r) = \mathbb{E}_{\pi, \Phi}[\sum_{h=1}^H r_h(x_h^\pi, a_h^\pi)]$, where x_h^π and a_h^π are the context and the selected action under π in stage h . The reward function $r = (r_h)_{h \in [H]}$ is non-stationary (it depends on the stage h), deterministic, and depends on contexts, not only on the corresponding latent states. We assume that the reward $r_h(x, a)$ gathered in (context, action) pair (x, a) in stage h lies in $[0, 1]$. We denote by π_r^* an optimal policy for the reward r , and by $V^*(r)$ its value.

In this section, we study the offline reward-free RL task (Yin and Wang, 2021). Specifically, the learner does not know nor utilize the reward function in the BMDP until the model estimation is completed (Jin et al., 2020). After that, the reward function r is revealed. The learner then computes $\hat{\pi}_r$, the optimal policy for the MDP $\hat{\Phi}$ with reward function r . This procedure is sometimes referred to as the *plug-in algorithm*. The performance of this model-based RL algorithm is characterized by $\Delta(\mathcal{R}) = \sup_{r \in \mathcal{R}} (V^*(r) - V^{\hat{\pi}_r}(r))$ for a given class \mathcal{R} of reward functions. The case where \mathcal{R} includes all possible reward functions is referred to as the *minimax reward* setting, whereas the case where \mathcal{R} reduces to a single reward function r to as the *reward-specific* setting.

5.2 Lower bounds

We derive lower bounds on the sample complexity for identifying ϵ -optimal policies in both the minimax reward and reward-specific settings. These lower bounds hold even when *adaptive* behavior policies are used to gather the data. As shown later in this section, the performance of the plug-in algorithm based on the model estimation procedure presented in previous sections almost matches these lower bounds, even if the behavior policy is restricted to

⁵w.h.p. means that with probability tending to 1 as $n \rightarrow \infty$.

being passive.

For the minimax setting, we first define

$$\Lambda(\Phi) = \max_{\mathbf{v} \in [-1, 1]^S} \frac{1}{S} \sum_{s=1}^S \max_{a_1, a_2} \langle p(\cdot|s, a_1) - p(\cdot|s, a_2), \mathbf{v} \rangle,$$

where $p(\cdot|s, a) = (p(s'|s, a))_{s' \in \mathcal{S}}$ is a vector of length S . Note that as $p(s'|s, a)$ does not depend on n for all (s, s') , $\Lambda(\Phi)$ is strictly positive and does not depend on n .

Theorem 4. [Minimax reward setting] Consider a BMDP Φ such that $\Lambda(\Phi) > 0$. Any algorithm that guarantees $\sup_r \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) \leq \epsilon$ with probability at least $1/2$ requires $TH = \Omega(\frac{n\Lambda(\Phi)}{\epsilon^2})$ samples.

The proof of the above theorem, presented in Appendix I, reveals that to get a minimax reward-free guarantee, the minimal sample complexity is mainly dictated by the estimation of the emission distributions q . This contrasts with the case where we target reward-specific guarantees. In this case, the minimal sample complexity is limited by the block structure estimation as shown in the following theorem, whose proof is deferred to Appendix I.

Theorem 5. [Reward-specific setting] Let $\epsilon = o(1)$. Consider an algorithm with the following guarantees: for any BMDP Φ satisfying Assumptions 1-4 and $I(\Phi) > 0$, and for any reward function r initially revealed to the algorithm, $\mathbb{E}_\Phi[\frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r))] \leq \epsilon$. Then the algorithm requires $TH = \Omega(n \log(\frac{1}{\epsilon}) + \frac{SA}{\epsilon^2})$ samples.

The first term $n \log(1/\epsilon)$ in the sample complexity lower bound may be interpreted as the number of samples required to learn the block structure accurately, and the second term to the data required to learn an ϵ -optimal policy given the block structure. We establish in the next subsection that our model-based method achieves the limits predicted in the above theorems. In addition, note that our method does not require any adaptive exploration procedure. And for the reward-specific guarantees, it does not even require any prior knowledge of the reward function (this knowledge is assumed for the lower bound in Theorem 5).

5.3 Performance of the plug-in algorithm

Consider the plug-in algorithm computing the optimal policy $\hat{\pi}_r$ for the estimated model $\hat{\Phi} = (\hat{p}, \hat{q}, f)$ and the reward function r . In the minimax reward setting, we have the following performance guarantees (their proof is presented in Appendix J).

Theorem 6 (Minimax reward setting). Consider a BMDP Φ satisfying Assumptions 1-4. Further assume that $TH = \omega(n)$ and $I(\Phi) > 0$. Then we have w.h.p.,

$$\sup_r \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) = \mathcal{O} \left(\sqrt{\frac{nS^2 A^2 \log(SAH)}{TH}} \right).$$

The above minimax guarantees match those predicted by the lower bounds presented in Theorem 4: $\sup_r \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r))$ scales as $\sqrt{\frac{n}{TH}}$ (up to a multiplicative constant $S\sqrt{A}$ and logarithmic factors). Further note that the gap between the value of the optimal policy and that of $\hat{\pi}_r$ decreases as H increases. This is the first time such decay is proved (for general MDPs, the gap should scale at least as $\frac{1}{\sqrt{T}}$ (Jin et al., 2020); we get a gap scaling as $\frac{1}{\sqrt{TH}}$ due the specific nature of our BMDPs and more specifically, the fact that they enjoy a bounded mixing time).

The next theorem (see Appendix J for the proof) provides performance guarantees in the reward-specific setting.

Theorem 7. [Reward-specific setting] Let C be the constant introduced in Theorem 3(i). Under the assumptions of Theorem 6, we have for any reward function r , w.h.p.

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

The upper bound shown in Theorem 7 consists of two terms. The first scaling as $\sqrt{1/T}$ corresponds to the error made when learning the optimal policy assuming the block structure has been accurately inferred. The second term, scaling as $e^{-D \frac{TH}{n}}$ for some $D > 0$, corresponds to the error made due to mistakes in the block structure estimation. These two terms also match the lower bound derived in Theorem 4: there, we have shown that $\frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r))$ should scale at least as $\sqrt{1/T}$ and $e^{-D \frac{TH}{n}}$. Finally note that in the case our algorithm recovers the clusters asymptotically exactly, i.e., when $TH - \frac{n \log(n)}{C'I(x; \Phi)} = \omega(1)$ for all x , then we can remove the second term in our upper bound, and hence prove that $\frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r))$ scales as $\sqrt{1/T}$.

Remark 2. The condition that $I(\Phi) > 0$ is closely related to the previous separability notions considered in BMDP literature (Du et al., 2019; Misra et al., 2020). In Appendix D.3.5, we show that our separability condition is strictly weaker than the γ -separability (Du et al., 2019) and kinematic separability (Misra et al., 2020) in that ours encompass a larger set of “separable” BMDPs. Additionally, unlike the previous works where the separability condition was imposed as assumptions, we have naturally derived our separability condition from an information-theoretic argument (e.g., from the proof of Theorem 1).

6 Related Work

Structure recovery in block models. The problem of learning the latent state decoding function is in a sense similar to the cluster recovery problem in SBM (Holland et al., 1983); see Abbe (2018) for a brief survey and references

therein. In the SBM, the learner observes a random graph whose edges are drawn independently of each other and with probabilities that depend only on the cluster ids of the two corresponding vertices. From this observation, the objective is to infer the initial clusters. Information-theoretical lower bounds and optimal algorithms have been proposed for this simple block model (Yun and Proutière, 2016; Abbe, 2018), and extensions, e.g., DCBM (Gao et al., 2018; Dasgupta et al., 2004; Karrer and Newman, 2011). For BMDPs, the data consists of trajectories of a controlled Markov chain. This implies that we deal with correlated samples, which significantly complicates the cluster recovery task. Dealing with Markovian data has been investigated in the case of *uncontrolled* Markov chains: in Zhu et al. (2022); Zhang and Wang (2020); Duan et al. (2019), the authors analyze scenarios under which the transition kernel of the Markov chain exhibits a low-rank structure. Closer to our problem, Sanders et al. (2020) studies cluster recovery in BMC. The main differences with our model are that in Sanders et al. (2020), (i) the Markov chain is uncontrolled (there are no control actions); (ii) more importantly, the emission distributions are known and uniform for each latent state (i.e., within each context cluster, the contexts are indistinguishable); (iii) finally, the observations come from a single long trajectory of the Markov chain, which simplifies the analysis (since it can leverage existing concentration results for Markov chains in the stationary regime (Paulin, 2015)).

Reinforcement learning in BMDPs. There have been considerable research efforts recently towards the design of efficient RL algorithms for BMDPs (Jiang et al., 2017; Dann et al., 2018; Du et al., 2019; Misra et al., 2020; Foster et al., 2021; Zhang et al., 2022b), as well as low-rank MDPs (Sun et al., 2019; Agarwal et al., 2020; Modi et al., 2021; Uehara et al., 2022). All the aforementioned studies (both for block and low-rank MDPs) rely on rich function approximators. When applied to our BMDP problem formulation, this means that the block structure can be accurately represented through a function belonging to some function class \mathcal{F} . When the reward function is fixed and given (similar to our reward-specific setting), the sample complexity for identifying an ϵ -optimal policy is proved to scale at most as $\text{poly}(S, A, H) \log |\mathcal{F}| / \epsilon^2$ where S , A and H denotes the number of latent states, of actions and of rounds per episode, respectively. However, without any prior or any additional structural assumption imposed in BMDPs, \mathcal{F} should correspond to the set of all possible assignments of contexts to latent states, so that $\log |\mathcal{F}| = n \log(S)$. As a consequence, the algorithms presented in the aforementioned papers do not provably exploit the block structure. Indeed, for tabular MDPs (without any block structure), there are algorithms with sample complexity scaling as $\text{poly}(H)An/\epsilon^2$, see Menard et al. (2021). We should also mention that the use of function approximation in Block MDPs and low-rank MDPs requires strong computational oracle assumptions

(ERM or MLE estimators on the function class Υ). The computational issues for RL with function approximation have been recently discussed in many papers including Kane et al. (2022); Golowich et al. (2022); Zhang et al. (2022a); Sam et al. (2022); Shah et al. (2020).

Feng et al. (2020) follow a different path towards efficient algorithms compared to the aforementioned studies, but assume that the learner has access to an unsupervised learning oracle that outputs the correct decoding function with high probability after $\text{poly}(S)$ observed samples. Finally, it is worth mentioning Azzadenesheli et al. (2016b), an early work where the authors analyze the regret of online algorithms using spectral methods to infer the structure. However, their regret guarantees scale as (and is valid only after) n^2 rounds.

To the best of our knowledge, this paper is the first to analyze when and how a block structure in MDPs can be learned and exploited without the assumption that this structure can be represented using function approximation and without resorting to any kind of oracles.

Reward-free RL The last part of this paper deals with reward-free RL. Such RL task has been extensively studied recently for tabular MDPs, see e.g., Jin et al. (2020); Kaufmann et al. (2021); Menard et al. (2021). In tabular episodic MDP with n states, the sample complexity of algorithms leading to ϵ -optimal policies for any reward function (in the minimax setting) is $\text{poly}(H)n^2A/\epsilon^2$. If the algorithm needs to return an ϵ -optimal policy for a given reward function, the sample complexity becomes $\text{poly}(H)nA/\epsilon^2$ (Menard et al., 2021). By exploiting the block structure and under appropriate conditions, we show in this paper that we can significantly reduce the sample complexity in both scenarios (e.g. by a factor n in the minimax setting).

Our work is also related to the so-called offline RL problems. For example, Xiao et al. (2022) study a learning task where the reward is specified at the very beginning and is given as part of the static data \mathcal{D} (the authors referred to this setting as *batch policy optimization (BPO)*).

7 Numerical Experiments

Setting. To show the efficacy of our clustering algorithm, we consider a simple synthetic η -regular block MDP with $n = 100$, $S = 2$, and $A = 3$. The latent transition matrix of each action is given as

$$P_1 = \begin{bmatrix} \frac{\eta}{1+\eta} & \frac{1}{1+\eta} \\ \frac{1}{1+\eta} & \frac{\eta}{1+\eta} \end{bmatrix}, P_2 = \begin{bmatrix} \frac{1}{1+\eta} & \frac{\eta}{1+\eta} \\ \frac{\eta}{1+\eta} & \frac{1}{1+\eta} \end{bmatrix}, P_3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

where $\eta > 1$ is the parameter determining the hardness of our BMDP instance and is pre-determined. Lastly, we use the uniform behavior policy to generate the trajectories.

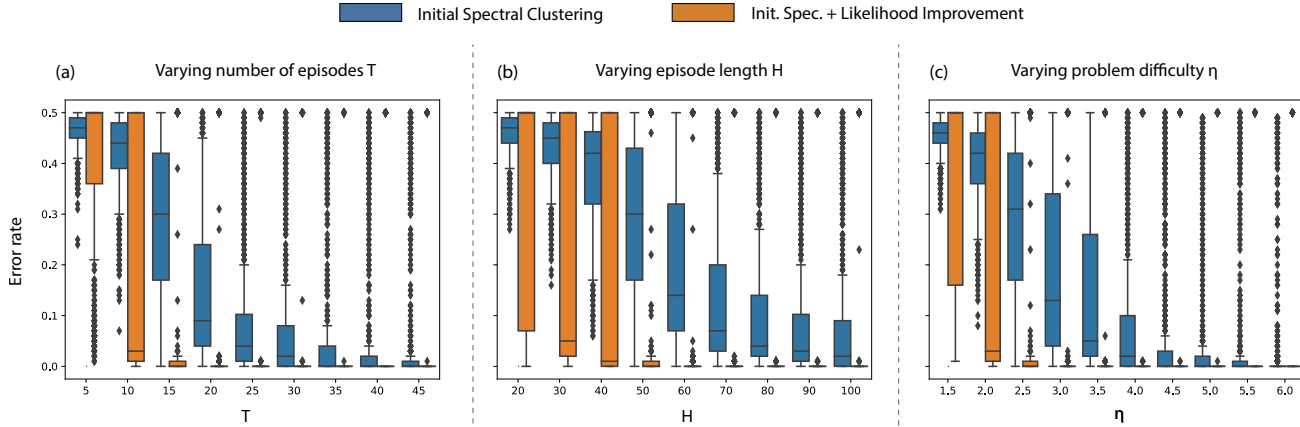


Figure 1: The clustering error rates for various choices of (a) T 's, (b) H 's, and (c) η 's.

This particular example shows that it is *necessary* to consider all actions via concatenation in the initial spectral clustering, as 1. playing the third action does not provide any useful information for clustering, and 2. considering the “marginalized” Markov chain, i.e., a single Markov chain with the transition matrix $\frac{1}{3}(P_1 + P_2 + P_3)$ also renders clustering impossible.

Results. For T , H , and η , we considered three settings where for each, we vary one parameter while fixing the other two. All experiments are repeated 1000 times, and the codes are available in our GitHub repository⁶.

The results are shown in Figure 1, where we show all the outliers for the sake of transparency. First note that with sufficiently many observations, we obtain exact clustering using initial clustering and likelihood improvement. In (a) and (b), the BMDP instance is fixed, and we observe a decay in the clustering error rate as the number of observations increases. In (c), the BMDP instance changes with the number of observations fixed, and as expected, we also observe a decay in the clustering error rate as η increases, i.e., as the instance gets easier.

We believe that the outliers are there because the initial spectral clustering sometimes results in poor initialization for the likelihood improvement step. This is not contradictory to our guarantees because they hold in probability in the limit of $TH = \omega(n \log n)$ as $n \rightarrow \infty$. Also, one can note that increasing T is more effective than increasing H , even though theoretically, the guarantees depend only on TH . We believe that this is due to a finite-size effect. Lastly, observe that in all experiments, there seems to be a sharp phase transition after which the algorithm consistently results in exact clustering. This is consistent with our lower bound (Theorem 1), which suggests an asymptotic phase transition at $TH = \omega(n)$ and $TH = \omega(n \log n)$.

⁶<https://github.com/nick-jhlee/optimal-block-mdp>

In Appendix M, we provide the precise experimental details as well as additional results suggesting that our algorithm is somewhat robust to random corruption.

8 Conclusion and Future Work

In this paper, we address learning problems in episodic BMDPs. We provide, for the first time, information-theoretical lower bounds on the latent state decoding error rate, as well as on the sample complexity for near-optimal policy identification in the reward-free setting (valid even for algorithms with adaptive exploration). We also devise simple algorithms that approach these fundamental limits. Importantly, by exploiting the block structure, we demonstrate that we can significantly accelerate the search for near-optimal policies (in most cases, by a factor n , the size of the context space). This is also verified empirically in a synthetic BMDP environment.

In this paper, we mostly restricted the analysis to the case where the data is generated using a fixed behavior policy. There are cases where active exploration is more desirable (Jin et al., 2020; Kaufmann et al., 2021; Zhang et al., 2021a; Tarbouriech and Lazaric, 2019; Tarbouriech et al., 2020). In Appendix K, we already provide some discussions on how to extend our theoretical results to such active behavior policies. We plan to make these results and their proofs precise as well as conduct numerical experiments to assess the practical relevance of active exploration.

Another interesting research direction would be to apply to BMDPs the concept of δ -significant latent state introduced in Jin et al. (2020) (later extended in Zhang et al. (2021a)) for tabular MDPs, which allows the analysis of RL tasks when some states are hard to reach. Using this concept, we might be able to further relax the η -regularity assumption (we have already proposed an interesting relaxation in Appendix L). Then we would have to recover the context to latent state mapping for significant latent states only.

Acknowledgments

We would like to thank Laura Schmid for helping with polishing the figures, and also the anonymous reviewers for their helpful comments and suggestions. Junghyun Lee and Se-Young Yun are supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korean government(MSIT) (No.2022-0-00311, Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects; No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)). Yasir Jedra and Alexandre Proutière are supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

- Emmanuel Abbe. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018. 2, 3, 4, 7, 8
- Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 20095–20107. Curran Associates, Inc., 2020. 8
- Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutière. Navigating to the Best Policy in Markov Decision Processes. In *Advances in Neural Information Processing Systems*, volume 34, pages 25852–25864. Curran Associates, Inc., 2021. 95, 96
- Esraa M. Al-Sharoa, Bara’ M. Ababneh, and Mahmood A. Alkhasawneh. Robust Community Detection in Graphs. *IEEE Access*, 9:118757–118770, 2021. 100
- Konstantin Avrachenkov, Alexey Piunovskiy, and Yi Zhang. Markov Processes with Restart. *Journal of Applied Probability*, 50(4):960–968, 2013. 6
- Konstantin Avrachenkov, Alexey Piunovskiy, and Yi Zhang. Hitting Times in Markov Chains with Restart and their Application to Network Centrality. *Methodology and Computing in Applied Probability*, 20(4):1173–1188, Dec 2018. 6
- Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement Learning of POMDPs using Spectral Methods. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 193–256, Columbia University, New York, New York, USA, 23–26 Jun 2016a. PMLR. 3, 97
- Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement Learning in Rich-Observation MDPs using Spectral Methods. In *Proceedings of the 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2016b. 3, 8
- 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. 17
- Patrick Billingsley. Statistical Methods in Markov Chains. *The Annals of Mathematical Statistics*, 32(1):12 – 40, 1961. 17
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. 39, 40
- Pierre Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, volume 31 of *Texts in Applied Mathematics*. Springer, 2 edition, 2020. 21, 22
- Yudong Chen, Xiaodong Li, and Jiaming Xu. Convexified modularity maximization for degree-corrected stochastic block models. *The Annals of Statistics*, 46(4):1573 – 1602, 2018. 46, 48, 50
- Imre Csiszar and Zsolt Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Transactions on Information Theory*, 52(3): 1007–1016, 2006. 33
- 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. 1, 8
- Anirban Dasgupta, John E. Hopcroft, and Frank McSherry. Spectral analysis of random graphs with skewed degree distributions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 602–610, 2004. 8
- Eric V. Denardo. Periods of Connected Networks and Powers of Nonnegative Matrices. *Mathematics of Operations Research*, 2(1):20–24, 1977. 97
- Roland L. Dobrushin. Central Limit Theorem for Nonstationary Markov Chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956a. 21
- Roland L. Dobrushin. Central Limit Theorem for Nonstationary Markov Chains. II. *Theory of Probability & Its Applications*, 1(4):329–383, 1956b. 21
- 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. 1, 5, 7, 8, 35
- Yaqi Duan, Tracy Ke, and Mengdi Wang. State Aggregation Learning from Markov Transition Data. In *Advances*

- in *Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. [8](#)
- Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-Optimal Off-Policy Evaluation with Linear Function Approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2701–2709. PMLR, 13–18 Jul 2020. [3](#), [17](#)
- Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005. [5](#), [51](#)
- Fei Feng, Ruosong Wang, Wotao Yin, Simon S Du, and Lin Yang. Provably Efficient Exploration for Reinforcement Learning Using Unsupervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 22492–22504. Curran Associates, Inc., 2020. [8](#)
- 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. [1](#), [8](#)
- 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. [2](#), [3](#), [4](#), [5](#), [8](#), [46](#), [48](#), [49](#)
- Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in Observable POMDPs, without Computationally Intractable Oracles. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. [8](#)
- Zhaohan Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-Bastien Grill, Florent Althé, Remi Munos, and Mohammad Gheshlaghi Azar. Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3875–3886. PMLR, 13–18 Jul 2020. [1](#)
- Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. [7](#)
- 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. [1](#), [8](#)
- 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 2020. [6](#), [7](#), [8](#), [9](#), [17](#), [82](#), [89](#), [95](#)
- Daniel Kane, Sihan Liu, Shachar Lovett, and Gaurav Mahajan. Computational-statistical gap in reinforcement learning. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1282–1302. PMLR, 02–05 Jul 2022. [8](#)
- Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physics Review E*, 83:016107, Jan 2011. [8](#)
- Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive Reward-Free Exploration. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 16–19 Mar 2021. [8](#), [9](#), [17](#)
- Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix Completion from Noisy Entries. *Journal of Machine Learning Research*, 11(69):2057–2078, 2010a. [5](#), [51](#)
- Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix Completion From a Few Entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010b. [56](#), [57](#)
- 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. [1](#)
- Tse L. Lai and Herbert Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. [3](#), [24](#)
- Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 13–18 Jul 2020. [1](#)
- David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2 edition, 2017. [43](#), [44](#)
- Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv preprint arXiv:2005.01643*, 2020. [3](#), [17](#)
- Allen Liu and Ankur Moitra. Minimax Rates for Robust Community Detection. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 823–831, 2022. [100](#)

- Aymen Al Marjani and Alexandre Proutière. Adaptive Sampling for Best Policy Identification in Markov Decision Processes. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7459–7468. PMLR, 18–24 Jul 2021. [95](#)
- 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. [8](#), [17](#), [95](#)
- 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. [1](#), [7](#), [8](#), [36](#)
- Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free Representation Learning and Exploration in Low-rank MDPs. *arXiv preprint arXiv:2102.07035v2*, 2021. [8](#), [89](#)
- Andrei V. Novikov. PyClustering: Data Mining Library. *Journal of Open Source Software*, 4(36):1230, 2019. [100](#)
- Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1 – 32, 2015. [6](#), [8](#), [38](#)
- 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. [3](#), [17](#)
- Tyler Sam, Yudong Chen, and Christina Lee Yu. Overcoming the Long Horizon Barrier for Sample-Efficient Reinforcement Learning with Latent Low-Rank Structure. *arXiv preprint arXiv:2206.03569*, 2022. [8](#)
- Jaron Sanders and Albert Senen–Cerdea. Spectral norm bounds for block markov chain random matrices. *Stochastic Processes and their Applications*, 158:134–169, 2023. [5](#), [6](#), [17](#), [39](#), [51](#)
- Jaron Sanders and Alexander Van Werde. Singular value distribution of dense random matrices with block markovian dependence. *Stochastic Processes and their Applications*, 2023. [17](#)
- Jaron Sanders, Alexandre Proutière, and Se-Young Yun. Clustering in Block Markov Chains. *The Annals of Statistics*, 48(6):3488 – 3512, 2020. [2](#), [3](#), [4](#), [5](#), [6](#), [8](#), [17](#), [24](#), [33](#), [39](#), [51](#), [60](#), [65](#), [70](#)
- Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample Efficient Reinforcement Learning via Low-Rank Matrix Estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 12092–12103. Curran Associates, Inc., 2020. [8](#)
- Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling Representation Learning from Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 18–24 Jul 2021. [1](#)
- Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 25–28 Jun 2019. [8](#)
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, 2 edition, 2018. [17](#)
- Jean Tarbouriech and Alessandro Lazaric. Active Exploration in Markov Decision Processes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 974–982. PMLR, 16–18 Apr 2019. [9](#), [95](#), [100](#)
- Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirota, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active Model Estimation in Markov Decision Processes. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 03–06 Aug 2020. [9](#)
- Jean Tarbouriech, Matteo Pirota, Michal Valko, and Alessandro Lazaric. A Provably Efficient Sample Collection Strategy for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 7611–7624. Curran Associates, Inc., 2021. [95](#)
- Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2009. [82](#), [84](#)
- Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation Learning for Online and Offline RL in Low-rank MDPs. In *International Conference on Learning Representations*, 2022. [8](#), [17](#)
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. [54](#)
- Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007. [5](#)
- Alexander Van Werde, Albert Senen–Cerdea, Gianluca Kosmella, and Jaron Sanders. Detection and Evaluation

- of Clusters within Sequential Data. *arXiv preprint arXiv:2204.13534*, 2022. 17
- Geoffrey Wolfer and Aryeh Kontorovich. Minimax Learning of Ergodic Markov Chains. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 904–930. PMLR, 22–24 Mar 2019. 17
- Geoffrey Wolfer and Aryeh Kontorovich. Statistical estimation of ergodic Markov chain kernel over discrete state space. *Bernoulli*, 27(1):532 – 553, 2021. 17
- Tianhao Wu, Yunchang Yang, Simon Du, and Liwei Wang. On Reinforcement Learning with Adversarial Corruption and Its Application to Block MDP. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11296–11306. PMLR, 18–24 Jul 2021. 100
- Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. On the Optimality of Batch Policy Optimization Algorithms. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11362–11371. PMLR, 18–24 Jul 2021. 3
- Chenjun Xiao, Ilbin Lee, Bo Dai, Dale Schuurmans, and Csaba Szepesvari. The Curse of Passive Data Collection in Batch Reinforcement Learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8413–8438. PMLR, 28–30 Mar 2022. 3, 8, 17
- Ming Yin and Yu-Xiang Wang. Asymptotically Efficient Off-Policy Evaluation for Tabular Reinforcement Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3948–3958. PMLR, 26–28 Aug 2020. 17
- 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. 6, 17
- Se-Young Yun and Alexandre Proutière. Community Detection via Random and Adaptive Sampling. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 138–175, Barcelona, Spain, 13–15 Jun 2014. PMLR. 24, 95, 100
- Se-Young Yun and Alexandre Proutière. Optimal Cluster Recovery in the Labeled Stochastic Block Model. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 3, 5, 8, 24, 60
- Se-Young Yun and Alexandre Proutière. Optimal Sampling and Clustering in the Stochastic Block Model. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3, 4, 95, 100
- Andrea Zanette and Emma Brunskill. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 09–15 Jun 2019. 95
- Anru Zhang and Mengdi Wang. Spectral State Compression of Markov Processes. *IEEE Transactions on Information Theory*, 66(5):3202–3231, 2020. 8
- Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making Linear MDPs Practical via Contrastive Representation Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26447–26466. PMLR, 17–23 Jul 2022a. 8
- 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 2022b. 1, 8, 89
- Zihan Zhang, Simon Du, and Xiangyang Ji. Near Optimal Reward-Free Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12402–12412. PMLR, 18–24 Jul 2021a. 9, 17
- Zihan Zhang, Xiangyang Ji, and Simon Du. Is Reinforcement Learning More Difficult Than Bandits? A Near-optimal Algorithm Escaping the Curse of Horizon. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4528–4531. PMLR, 15–19 Aug 2021b. 17
- Ziwei Zhu, Xudong Li, Mengdi Wang, and Anru Zhang. Learning Markov Models Via Low-Rank Optimization. *Operations Research*, 70(4):2384–2398, 2022. 8

Contents

1	Introduction	1
2	Models and Objectives	2
2.1	Episodic block MDPs	2
2.2	Latent state decoding and model estimation	2
2.3	Assumptions	3
3	Fundamental Lower Bound of Latent State Decoding	3
4	Latent State Decoding and Model Estimation Algorithms	4
4.1	Algorithms	4
4.2	Performance guarantees	6
4.3	Sketches of the proofs	6
5	Reward-Free RL in BLock MDPs	6
5.1	Setting and objectives	6
5.2	Lower bounds	6
5.3	Performance of the plug-in algorithm	7
6	Related Work	7
7	Numerical Experiments	8
8	Conclusion and Future Work	9
A	Further Related Work	17
B	Notations	18
C	Markov Chains induced in BMDPs and their Mixing Times	19
C.1	Regularity and stationary distributions	19
C.2	Bounds on multi-hop transition probabilities	21
C.3	Mixing times	21
D	Proof of Theorem 1 – Fundamental Lower Bound of Latent State Decoding	24
D.1	Proof of Proposition 8	26
D.2	Proof of Proposition 9	28
D.3	Properties of the rate function $I(x; \Phi)$	30
D.3.1	Necessary and sufficient conditions for $I(x; \Phi) = 0$	30
D.3.2	Alternative KL-form of the rate function $I(x; \Phi)$	31

D.3.3	Asymptotics of m_ρ	33
D.3.4	Examples of rate function $I(x; \Phi)$	34
D.3.5	Relation to other BMDP separability notions	35
E	Bernstein-type Concentration for Markov Chains with Restarts and Applications	38
E.1	Concentration of Markov chains with restarts	38
E.2	Proof of Theorem 9	38
E.3	Towards concentration inequalities in BMDPs	39
E.4	Supporting lemmas for Theorem 9	39
F	Proof of Theorem 2 – Initial Spectral Clustering	45
F.1	Algorithm, preliminaries, and notations	45
F.2	Proof of Theorem 2	46
F.3	Separability	46
F.4	Weighted K-medians clustering	48
F.5	S -rank approximation	51
F.6	Analysis of the trimmed random matrix	51
G	Proof of Theorem 3 (i) – Iterative Likelihood Improvement	60
G.1	Proof of Proposition 20 – Bounding $ \mathcal{H}^{\mathbb{G}} $	60
G.2	Proof of Proposition 21 – Bounding $ \mathcal{E}_{\mathcal{H}}^{(\ell)} $	62
G.3	Postponed proofs – Bounding $ \mathcal{H}_1^{\mathbb{G}} $	64
G.3.1	Proof of Proposition 22: concentration for the rate function I	64
G.3.2	Proof of Lemma 20	66
G.4	Postponed proofs - Bounding $ \mathcal{E}_{\mathcal{H}}^{(\ell)} $	68
G.4.1	Proof of Lemma 21: bounding E_1, E_2, U	68
G.4.2	Intermediate estimation errors for p and p^{bwd}	70
H	Proof of Theorem 3 (ii) – Estimation of the Latent Transitions and Emission Probabilities	73
H.1	Preliminaries on the estimators	73
H.2	Proof of Theorem 3 (ii)	73
H.3	Estimation error decomposition lemmas	74
H.4	Concentration bounds	74
H.5	Proofs – Estimation error decompositions	75
H.6	Proofs – Concentration bounds	77
I	Proofs of Theorems 4 and 5 – Lower Bounds for Reward-Free RL	82
I.1	Proof of Theorem 4 – minimax setting	82
I.2	Proof of Theorem 5 – reward-specific setting	83

J	Proofs of Theorems 6 and 7 – Upper Bounds for Reward-Free RL	85
J.1	Preliminaries and notations	85
J.2	Proof of Theorem 6	85
J.3	Proof of Theorem 7	86
J.4	Value difference lemmas	87
J.5	Concentration bounds on averaged optimal value functions	88
J.6	Proofs – Value difference lemmas	89
J.7	Proofs – Concentration bounds on averaged optimal value functions	93
K	Towards Optimal Adaptive Exploration in BMDPs	95
K.1	Lower bound on the latent state decoding error rate	95
K.2	Adaptive latent state decoding algorithms	96
L	Beyond the η-regularity Assumption	97
L.1	Relaxing Assumption 3(ii)	97
L.2	Bernstein-type concentration bounds for Markov chains with restarts.	97
L.3	On the lower bound for the latent state decoding error rate	99
L.4	On the performance guarantee of the initial spectral clustering (Algorithm 1)	99
L.5	On the guarantee of the iterative likelihood improvement (Algorithm 2)	99
L.6	On the guarantees for reward-free RL	99
M	Deferred Experimental Details	100
M.1	Missing details for Section 7	100
M.2	Additional experiment: clustering with random corruptions	100

A Further Related Work

Estimation of (controlled) Markov chains The plug-in estimator used for the offline reward-free RL has been studied extensively in the context of estimating the transition matrix of some underlying finite, ergodic Markov chain, which is of both theoretical and practical importance. For uncontrolled Markov chains, Billingsley (1961) was the first to consider this as a non-parametric estimation problem. Recently, Wolfer and Kontorovich (2019, 2021) established the PAC-type minimax sample complexity bound w.r.t. any metric. There has also been some recent interest in obtaining improved guarantees or improving the estimator itself when a certain structural assumption is imposed on the Markov chain. As mentioned previously, Sanders et al. (2020) introduced and studied the problem of clustering and estimation of the transition kernel (and the emission probabilities) of BMCs; further spectral analyses have been done in Sanders and Senen-Cerda (2023); Sanders and Van Werde (2023). It is also worthwhile to mention Werde et al. (2022), in which a thorough evaluation of the BMC model in a realistic setting has been done, showing that indeed the BMC model as well as the proposed clustering algorithm can provide meaningful insights into exploratory data analyses and more.

However, all the aforementioned works are for uncontrolled Markov chains, whereas most applications in RL and control theory deal with *controlled* Markov chains (e.g., action in RL literature). Although the efficacy of the plug-in estimators in offline RL is known (Duan et al., 2020; Yin and Wang, 2020; Ren et al., 2021), the precise characterization of the non-parametric estimation of *controlled* Markov chain in an offline setting has been an open question. Banerjee et al. (2022) first tackled this by showing that the usual plug-in estimator (per action) is minimax optimal in the PAC sense and provides results for various situations such as Markov chain with restarts and offline policy evaluation (Sutton and Barto, 2018).

Offline, reward-free RL The main focus of the first half is on the statistically tight characterization of clustering, as well as guarantees on how well the transition dynamics can be estimated. In RL, this is often referred to as *model estimation*. As the learner cannot interact further with the environment other than the given trajectories that have no reward information and are generated via some fixed, passive behavior policy ρ , our setting naturally extends to the intersection of (model-based) offline RL and reward-free RL. To the best of our knowledge, we are the first to consider offline reward-free RL on BMDPs.

Offline RL is a framework in which the learner employs a static dataset \mathcal{D} collected by some behavior policy ρ and performs training *without* any further interaction with the environment (Levine et al., 2020). Such a framework is especially useful when the data collection is expensive or dangerous in cases such as healthcare. Several recent works (Yin and Wang, 2021; Ren et al., 2021; Zhang et al., 2021b; Uehara et al., 2022; Xiao et al., 2022) studied the statistical efficiency of offline RL.

Reward-free RL (or reward-free exploration), pioneered by Jin et al. (2020), is the framework in which the goal is to estimate the optimal policy under *any* reward function that is revealed to the learner *after* a single exploration phase. Thus the main focus is to explore well such that the collected data has sufficient coverage. Recently, this has been extensively studied theoretically for tabular MDPs, see e.g., Jin et al. (2020); Kaufmann et al. (2021); Menard et al. (2021); Zhang et al. (2021a); Yin and Wang (2021).

For a fair comparison, we consider a tabular episodic MDP with n states such that every context-action pair can be reached with probability $\Theta(1/nA)$. we now recall the state-of-the-art sample complexity results for *offline reward-free* settings, in which the (later revealed) rewards can be *non-stationary*. If the algorithm needs to return an ϵ -optimal policy for any reward functions (reward-free setting), Yin and Wang (2021) showed the bound $\tilde{\mathcal{O}}\left(\frac{H^2 n^2 A}{\epsilon^2}\right)$, which matches the known lower bound up to logarithmic factors (Jin et al., 2020). If the algorithm needs to return an ϵ -optimal policy for a single reward function (reward-specific setting), Ren et al. (2021) showed⁷ the bound $\tilde{\mathcal{O}}\left(\frac{nA}{\epsilon^2} + \frac{H^2 n^2 A}{\epsilon}\right)$, as well as nearly matching minimax lower bound $\Omega\left(\frac{nA}{\epsilon^2}\right)$. The reason for the extra H terms compared to Ren et al. (2021) is because they consider the total bounded reward assumption ($\sum_h r_h \in [0, 1]$), and thus to compare to our setting in which we assume bounded reward ($r_h \in [0, 1]$), we need to replace n with nH , ϵ with ϵH , and multiply by H in their guarantees. By exploiting the block structure and under appropriate conditions (e.g., when $\epsilon = o(1)$), we significantly improve the sample complexities in all cases by a factor of n , and especially in reward-specific setting, exponentially in ϵ for a certain regime of ϵ .

⁷Technically, they considered the so-called *offline policy optimization* scenario in which instead of the reward function being revealed to the learner, the trajectories outputted by some (potentially unknown) behavior policy contain the rewards. Thus, the learner has to also estimate the underlying reward function, while in our setting, the true reward function is revealed, i.e., estimation of the reward function is not necessary. Despite this difference, one can immediately conclude from a quick inspection of their proofs that their upper and lower bounds still hold for our reward-specific setting.

B Notations

Generic notations	
$\mathbf{1}$	Column vector with all entries equal to 1
$\ \cdot\ _1$	The ℓ_1 norm on vectors
$\ \cdot\ _2$	The ℓ_2 norm on vectors
$\ \cdot\ _\infty$	The infinity norm on vectors
$\ \cdot\ $	Operator norm on matrices
$\ \cdot\ _F$	Frobenius norm on matrices
$x \vee y$	To mean $\max(x, y)$
$x \wedge y$	To mean $\min(x, y)$
\mathbb{S}^{d-1}	Unit sphere in \mathbb{R}^d
$[K]$	For a given integer K , denotes the set $\{1, \dots, K\}$
$\mathcal{P}(\mathcal{Z})$	The set of probability distributions over \mathcal{Z}
η -regular	A discrete probability ν in $\mathcal{P}(\mathcal{Z})$ is said η -regular if $\max_{z_1, z_2 \in \mathcal{Z}} \frac{\nu(z_1)}{\nu(z_2)} \leq \eta$
$\mathcal{P}(\mathcal{Z}, \eta)$	The set of probability distributions over \mathcal{Z} that are η -regular
$d_{TV}(\cdot, \cdot)$	The total variation distance between probability measures
$\text{KL}(\cdot \ \cdot)$	The Kullback-Liebler divergence between probability measures
$f(n) \sim g(n)$	To mean $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 1$
$f(n) \asymp g(n)$	To mean there exists $c, C > 0$ such that for all $n \geq 1$, $cg(n) \leq f(n) \leq Cg(n)$.
$f(n) \gtrsim g(n)$	To mean there exists $c > 0$ such that for all $n \geq 1$, $f(n) \geq cg(n)$.
$f(n) \lesssim g(n)$	To mean there exists $C > 0$ such that for all $n \geq 1$, $f(n) \leq Cg(n)$.

Block MDPs	
n	Number of contexts
S	Number of latent states
A	Number of actions
$f(x)$	Latent state of context x
$p(s' s, a)$	Transition probability from latent state s to s' with action a
$q(x s)$	Emission probability for context x given the latent state s
$P(y x, a)$	Transition probability from context x to y with action a
$\rho(a x)$	Behavior policy (probability of choosing action a at context x)

C Markov Chains induced in BMDPs and their Mixing Times

In this appendix, we analyze some of the Markov chains induced by the dynamics in a BMDP. Specifically, we study the Markov chain capturing the dynamics of the context, the Markov chain representing the evolution of the (action, next context) pair, and finally the Markov chain whose state is the triple (context, action, next context). To do that, we first define the *contextual transition kernel* $P(y|x, a) = p(f(y)|f(x), a)q(y|f(y))$. We denote by MC_0, MC_1, MC_2 these chains, respectively. Under the (memoryless) behavior policy $\rho \sim \mathcal{U}(\mathcal{A})$, the transition kernels and the initial state distributions of these chains are: for all $(x, x', y, y') \in \mathcal{X}^4$, and all $(a, b) \in \mathcal{A}^2$,

$$\begin{cases} P_0(y|x) = \sum_a \rho(a|x)P(y|x, a), & \mu_0(x) = \mu(x), \\ P_1((b, y)|(a, x)) = \rho(b|x)P(y|x, b), & \mu_1(a, x) = \sum_{y \in \mathcal{X}} \mu(y)\rho(a|y)P(x|y, a), \\ P_2((y, b, y')|(x, a, x')) = \mathbb{1}_{y=x'}\rho(b|y)P(y'|y, b), & \mu_2(x, a, x') = \mu(x)\rho(a|x)P(x'|x, a). \end{cases}$$

Inspired by our Assumption 3 (ii), we introduce the notion of η -regular Markov chains:

Definition 2 (η -regular Markov chain). *A time-homogenous Markov chain $(Z_t)_{t \geq 1}$ with finite state space \mathcal{Z} and transition kernel $P(z'|z) = \mathbb{P}[Z_{h+1} = z'|Z_h = z]$ is η -regular for some $\eta \geq 1$ if and only if*

$$\max_{(x, y, z) \in \mathcal{Z}^3} \max \left\{ \frac{P(y|x)}{P(z|x)}, \frac{P(x|y)}{P(x|z)} \right\} \leq \eta.$$

We remark that Assumption 3 implies that: for all (s, s', a) and all $x \in f^{-1}(s)$,

$$\frac{1}{\eta S} \leq \alpha_s \leq \frac{\eta}{S}, \quad \frac{1}{\eta S} \leq p(s'|s, a) \leq \frac{\eta}{S}, \quad \frac{1}{\eta \alpha_s n} \leq q(x|s) \leq \frac{\eta}{\alpha_s n}.$$

This observation will be instrumental in all the proofs presented in the remainder of the Appendix.

C.1 Regularity and stationary distributions

We start with the following:

Proposition 1. *For all $(x, y, a) \in \mathcal{X}^2 \times \mathcal{A}$,*

$$\frac{1}{\eta^3 n} \leq P(y|x, a) \leq \frac{\eta^3}{n} \tag{6}$$

The following two propositions provide basic properties of the Markov chains MC_0 and MC_1 .

Proposition 2. *Under Assumption 3, the Markov chain MC_0 is η^3 -regular, and irreducible aperiodic. Let Π_0 denote its stationary distribution. We have: for all $(x, y) \in \mathcal{X}^2$,*

$$\frac{1}{\eta^3 n} \leq P_0(y|x) \leq \frac{\eta^3}{n}, \quad \frac{1}{\eta^3 n} \leq \Pi_0(x) \leq \frac{\eta^3}{n}. \tag{7}$$

Proposition 3. *Under Assumptions 3 and 4, the Markov chain MC_1 is η^3 -regular, and irreducible aperiodic. Let Π_1 denote its stationary distribution. We have: for all $(x, y) \in \mathcal{X}^2$, and all $(a, b) \in \mathcal{A}^2$,*

$$\frac{1}{\eta^3 n A} \leq \mu_1(a, x) \leq \frac{\eta^3}{n A}, \quad \frac{1}{\eta^3 n A} \leq P_1((b, y)|(a, x)) \leq \frac{\eta^3}{n A}, \quad \frac{1}{\eta^3 n A} \leq \Pi_1(a, x) \leq \frac{\eta^3}{n A}. \tag{8}$$

In addition, for all $(a, x) \in \mathcal{A} \times \mathcal{X}$, $\Pi_1(a, x) = \sum_{y \in \mathcal{X}} \Pi_0(y)\rho(a|y)P(x|y, a)$.

One may easily check that the Markov chain MC_2 is not η -regular for any η , but our analyses will actually leverage the properties of Markov chains with kernel P_2^2 . More precisely, we introduce the following two Markov chains: $MC_{2, \text{odd}}$ and $MC_{2, \text{even}}$. As the name suggests, they correspond to $Z_h = (X_{2h-1}, A_{2h-1}, X_{2h})$ and $Z_h = (X_{2h}, A_{2h}, X_{2h+1})$ (for $h \geq 1$), respectively. These chains share the same transition kernel, defined as follows:

$$P_2^2((y, b, y')|(x, a, x')) := (P_2)^2((y, b, y')|(x, a, x'))$$

$$\begin{aligned}
 &= \sum_{\tilde{x}, \tilde{a}, \tilde{y}} P_2((y, b, y') | (\tilde{x}, \tilde{a}, \tilde{y})) P_2((\tilde{x}, \tilde{a}, \tilde{y}) | (x, a, x')) \\
 &= \sum_{\tilde{a} \in \mathcal{A}} \rho(b|y) P(y'|y, b) \rho(\tilde{a}|x') P(y|x', \tilde{a}).
 \end{aligned} \tag{9}$$

The two chains have different initial distributions:

$$\mu_{2, \text{odd}}(x, a, x') = \mu(x) \rho(a|x) P(x'|x, a), \tag{10}$$

$$\mu_{2, \text{even}}(x, a, x') = \sum_{y, b} \mu(y) \rho(b|y) P(x|y, b) \rho(a|x) P(x'|x, a). \tag{11}$$

Proposition 4. *Under Assumptions 3 and 4, the Markov chains $MC_{2, \text{odd}}$ and $MC_{2, \text{even}}$ are η^3 -regular, and irreducible aperiodic. The two Markov chains share the same stationary distribution, denoted by Π_2 . We have: for all $(x, x', y, y') \in \mathcal{X}^2$, and all $(a, b) \in \mathcal{A}^2$,*

$$\frac{1}{\eta^3 n^2 A} \leq \mu_{2, \text{odd}}(x, a, x') \leq \frac{\eta^3}{n^2 A}, \quad \frac{1}{\eta^6 n^2 A} \leq \mu_{2, \text{even}}(x, a, x') \leq \frac{\eta^6}{n^2 A}, \tag{12}$$

$$\frac{1}{\eta^6 n^2 A} \leq P_2^2((y, b, y') | (x, a, x')) \leq \frac{\eta^6}{n^2 A}, \quad \frac{1}{\eta^6 n^2 A} \leq \Pi_2(x, a, x') \leq \frac{\eta^6}{n^2 A}. \tag{13}$$

In addition, for all $(x, x', a) \in \mathcal{X}^2 \times \mathcal{A}$, $\Pi_2(x, a, x') = \Pi_0(x) \rho(a|x) P(x'|x, a)$.

The proofs of the above propositions are straightforward. We just justify the expressions of the stationary distributions Π_1 and Π_2 .

Proof. (Stationary distribution of Π_1) It suffices to show that $\sum_{y \in \mathcal{X}} \Pi_0(y) \rho(b|y) P(z|y, b)$ satisfies the balance equation for the MC_1 :

$$\begin{aligned}
 \sum_{(z, b) \in \mathcal{X} \times \mathcal{A}} \left[\sum_{y \in \mathcal{X}} \Pi_0(y) \rho(b|y) P(z|y, b) \right] P_1(x, a|z, b) &= \sum_{(z, b) \in \mathcal{X} \times \mathcal{A}} \left[\sum_{y \in \mathcal{X}} \Pi_0(y) \rho(b|y) P(z|y, b) \right] \rho(a|z) P(x|z, a) \\
 &= \sum_{z \in \mathcal{X}} \rho(a|z) P(x|z, a) \sum_{y \in \mathcal{X}} \Pi_0(y) \left[\sum_{b \in \mathcal{A}} \rho(b|y) P(z|y, b) \right] \\
 &= \sum_{z \in \mathcal{X}} \rho(a|z) P(x|z, a) \sum_{y \in \mathcal{X}} \Pi_0(y) P_0(z|y) \\
 &= \sum_{z \in \mathcal{X}} \rho(a|z) P(x|z, a) \Pi_0(z).
 \end{aligned}$$

□

Proof. (Stationary distribution of Π_2) Again we show that $\Pi_0(x) \rho(a|x) P(x'|x, a)$ satisfies the balance equations of the Markov chain MC_2 :

$$\begin{aligned}
 &\sum_{(y, y', b) \in \mathcal{X}^2 \times \mathcal{A}} [\Pi_0(y) \rho(b|y) P(y'|y, b)] P_2^2((x, a, x') | (y, b, y')) \\
 &= \sum_{(y, y', b)} [\Pi_0(y) \rho(b|y) P(y'|y, b)] \sum_{\tilde{a} \in \mathcal{A}} \rho(a|x) P(x'|x, a) \rho(\tilde{a}|y') P(x|y', \tilde{a}) \\
 &= \rho(a|x) P(x'|x, a) \sum_{(y, y', b)} \Pi_0(y) \rho(b|y) P(y'|y, b) \underbrace{\sum_{\tilde{a} \in \mathcal{A}} \rho(\tilde{a}|y') P(x|y', \tilde{a})}_{= P_0(x|y')} \\
 &= \rho(a|x) P(x'|x, a) \sum_{(y, y')} \Pi_0(y) P_0(y'|y) P_0(x|y') \\
 &= \rho(a|x) P(x'|x, a) \Pi_0(x).
 \end{aligned}$$

□

C.2 Bounds on multi-hop transition probabilities

We can establish that the bounds for P presented in the above three propositions hold for P^h for any $h \geq 1$. To this aim, we use the following generic lemma:

Lemma 1. *Let \mathcal{Z} be a finite state space. For a row-stochastic matrix P , the following holds for all $x, y \in \mathcal{Z}$ and all $h \geq 1$:*

$$\min_{x, y \in \mathcal{Z}} P(x, y) \leq P^h(x, y) \leq \max_{x, y \in \mathcal{Z}} P(x, y). \quad (14)$$

Proof. One important observation is that P^h is also row-stochastic, for any $h \geq 1$. We only prove the upper bound, as the lower bound follows in the exact same manner.

$h = 1$ is trivial, and thus let $h \geq 2$. Then, for any $(x, y) \in \mathcal{Z} \times \mathcal{Z}$

$$P^h(x, y) = \sum_{z \in \mathcal{Z}} P(x, z) P^{h-1}(z, y) \leq \left(\max_{x, z \in \mathcal{Z}} P(x, z) \right) \sum_{z \in \mathcal{Z}} P^{h-1}(z, y) = \max_{x, z \in \mathcal{Z}} P(x, z).$$

□

Combining the results of the above lemma and those of Propositions 3 and 4, we simply deduce:

Corollary 1. *For all $(x, a, x'), (y, b, y') \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$ and $h \geq 1$,*

$$\frac{1}{\eta^3 n} \leq (P_0)^h(y|x) \leq \frac{\eta^3}{n}, \quad (15)$$

$$\frac{1}{\eta^3 n A} \leq (P_1)^h((b, y)|(a, x)) \leq \frac{\eta^3}{n A}, \quad (16)$$

$$\frac{1}{\eta^6 n^2 A} \leq (P_2)^h((y, b, y')|(x, a, x')) \leq \frac{\eta^6}{n^2 A}. \quad (17)$$

C.3 Mixing times

Consider an irreducible aperiodic Markov chain with initial distribution μ , transition kernel P , and stationary distribution Π . Its mixing time is defined as $\inf\{h \geq 1 : d_{TV}(\mu P^h, \Pi) \leq 1/4\}$. The following proposition provides upper bounds of the Markov chains MC_0 , MC_1 , $MC_{2,odd}$ and $MC_{2,even}$.

Proposition 5. *Assume that $\rho(\cdot|x) \sim \mathcal{U}(A)$. Under Assumptions 3 and 4, we have:*

- (i) *The mixing times of MC_0 is upper bounded by $2\eta^2$.*
- (i) *The mixing times of MC_1 is upper bounded by $2\eta^2$.*
- (ii) *The mixing times of $MC_{2,odd}$ and $MC_{2,even}$ are both upper bounded by $\eta^2 + 1$.*

As described below, to obtain our tight upper bounds for the mixing times, instead of simply using the loose (regularity) bounds for the transition kernels, we use more sophisticated arguments.

The proof of the above results relies on Dobrushin's ergodic coefficient:

Definition 3 (Dobrushin (1956a,b)). *For any row-stochastic matrix P , define the Dobrushin's ergodic coefficient $\delta(P)$ as follows:*

$$\delta(P) := \frac{1}{2} \max_{x, y \in \mathcal{X}} \sum_{z \in \mathcal{X}} |P(z|x) - P(z|y)|. \quad (18)$$

The Dobrushin's ergodic coefficient can be equivalently rewritten (Exercise 4.4.12 of Bremaud (2020)) as follows:

$$\delta(P) = 1 - \min_{x, y \in \mathcal{X}} \sum_{z \in \mathcal{X}} (P(z|x) \wedge P(z|y)). \quad (19)$$

Now, $\delta(P)$ is directly related to the convergence rate of the stationary distribution:

Theorem 8 (Theorem 4.3.15 of [Bremaud \(2020\)](#)). For any h ,

$$d_{TV}(\mu P^h, \Pi) \leq (\delta(P))^h d_{TV}(\mu, \Pi) \quad (20)$$

and

$$d_{TV}(P^{h+1}(z, \cdot), \Pi) \leq (\delta(P))^h d_{TV}(P(z, \cdot), \Pi) \quad (21)$$

Proof of Proposition 5 (i) We analyze the mixing time $t_{\text{mix}, MC_0}(\varepsilon) = \inf\{h \geq 1 : d_{TV}(\mu P_0^h, \Pi_0) \leq \varepsilon\}$, and apply the results to $\varepsilon = 1/4$. We prove that:

$$t_{\text{mix}, MC_0}(\varepsilon) \leq \eta^2 \log \frac{1}{\varepsilon}. \quad (22)$$

Proof. For any $x, y, z \in \mathcal{X}$,

$$\begin{aligned} P_0(z|x) \wedge P_0(z|y) &= \left(\sum_{a \in \mathcal{A}} \rho(a|x) P(z|x, a) \right) \wedge \left(\sum_{a \in \mathcal{A}} \rho(a|y) P(z|y, a) \right) \\ &= \frac{1}{A} \left(\sum_{a \in \mathcal{A}} P(z|x, a) \right) \wedge \left(\sum_{a \in \mathcal{A}} P(z|y, a) \right) \\ &\geq \frac{1}{A} \left(\sum_{a \in \mathcal{A}} \frac{1}{\eta \alpha_{f(z)} n} p(f(z)|f(x), a) \right) \wedge \left(\sum_{a \in \mathcal{A}} \frac{1}{\eta \alpha_{f(z)} n} p(f(z)|f(y), a) \right) \\ &\geq \frac{1}{\eta \alpha_{f(z)} n A} \left(\sum_{a \in \mathcal{A}} \frac{1}{\eta} p(f(z)|f(y), a) \right) \wedge \left(\sum_{a \in \mathcal{A}} p(f(z)|f(y), a) \right) \\ &\geq \frac{1}{\eta^2 \alpha_{f(z)} n A} \sum_{a \in \mathcal{A}} p(f(z)|f(y), a) \end{aligned}$$

Then the Dobrushin's coefficient can be bounded as follows:

$$\begin{aligned} \delta(P_0) &= 1 - \min_{x, y \in \mathcal{X}} \sum_{z \in \mathcal{X}} (P_0(z|x) \wedge P_0(z|y)) \\ &\leq 1 - \min_{x, y \in \mathcal{X}} \sum_{z \in \mathcal{X}} \frac{1}{\eta^2 \alpha_{f(z)} n A} \sum_{a \in \mathcal{A}} p(f(z)|f(y), a) \\ &= 1 - \frac{1}{\eta^2 A} \min_{y \in \mathcal{X}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \sum_{z \in f^{-1}(s)} \frac{1}{n \alpha_s} p(s|f(y), a) \\ &= 1 - \frac{1}{\eta^2}. \end{aligned}$$

Thus, we have:

$$d_{TV}(P_0^h(z, \cdot), \Pi) \leq (\delta(P_0))^h d_{TV}(\mu_0, \Pi_0) \leq \left(1 - \frac{1}{\eta^2}\right)^h d_{TV}(\mu_0, \Pi_0) \leq e^{-\frac{h}{\eta^2}} d_{TV}(\mu_0, \Pi_0).$$

Note that $d_{TV}(\mu_0, \Pi_0) = \frac{1}{2} \sum_{z \in \mathcal{X}} |\mu(z) - \Pi(z)| \leq \frac{1}{2} \sum_{z \in \mathcal{X}} (\mu_0(z) + \Pi_0(z)) = 1$. In summary, we have $d_{TV}(P_0^h(z, \cdot), \Pi) \leq \varepsilon$ whenever $h \geq \eta^2 \log \frac{1}{\varepsilon}$. This completes the proof. \square

Proof of Proposition 5 (ii) The proof follows from exactly the same arguments as those used in the proof of (i).

Proof of Proposition 5 (iii) We start with the following lemma, which relates the power of P_2 to the power of P_1 :

Lemma 2. $P_2^{h+1}(y', a', z|x, a, y) = \sum_{\tilde{a} \in \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) P_1^h((\tilde{a}, y')|(a, y)).$

Proof. We proceed by induction.

- For $h = 1$:

$$\begin{aligned} P_2^2((y', a', z)|(x, a, y)) &= \sum_{(\tilde{x}, \tilde{a}, \tilde{y}) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}} P_2((y', a', z)|(\tilde{x}, \tilde{a}, \tilde{y})) P_2((\tilde{x}, \tilde{a}, \tilde{y})|(x, a, y)) \\ &= \sum_{\tilde{a} \in \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) P_1((\tilde{a}, y')|(a, y)). \end{aligned}$$

- For $h \geq 2$:

$$\begin{aligned} P_2^{h+1}((y', a', z)|(x, a, y)) &= \sum_{(\tilde{x}, \tilde{a}, \tilde{y}) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}} P_2((y', a', z)|(\tilde{x}, \tilde{a}, \tilde{y})) P_2^h((\tilde{x}, \tilde{a}, \tilde{y})|(x, a, y)) \\ &= \sum_{(\tilde{x}, \tilde{a}) \in \mathcal{X} \times \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) \sum_{\tilde{a}' \in \mathcal{A}} P_1((\tilde{a}, y')|(\tilde{a}', \tilde{x})) P_1^{h-1}((\tilde{a}', \tilde{x})|(a, y)) \\ &= \sum_{\tilde{a} \in \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) \sum_{(\tilde{x}, \tilde{a}') \in \mathcal{X} \times \mathcal{A}} P_1((\tilde{a}, y')|(\tilde{a}', \tilde{x})) P_1^{h-1}((\tilde{a}', \tilde{x})|(a, y)) \\ &= \sum_{\tilde{a} \in \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) P_1^h((\tilde{a}, y')|(a, y)). \end{aligned}$$

□

Lemma 3. For all $h \geq 1$ and all $(x, a, y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$,

$$d_{TV} \left((P_2)^{h+1}(\cdot|(x, a, y)), \Pi_2 \right) \leq d_{TV} \left((P_1)^h(\cdot|(a, y)), \Pi_1 \right). \quad (23)$$

Proof. Noting that $\Pi_2(y', a', z) = \sum_{\tilde{a} \in \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) \Pi_1(\tilde{a}, y')$, we have:

$$\begin{aligned} &2d_{TV} \left(P_2^{h+1}(\cdot|x, a, y), \Pi_2 \right) \\ &= \sum_{(y', a', z) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}} \left| P_2^{h+1}(y', a', z|x, a, y) - \Pi_2(y', a', z) \right| \\ &\stackrel{(*)}{=} \sum_{(y', a', z) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}} \left| \sum_{\tilde{a} \in \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) P_1^h((\tilde{a}, y')|(a, y)) - \sum_{\tilde{a} \in \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) \Pi_1(\tilde{a}, y') \right| \\ &\leq \sum_{(y', \tilde{a}) \in \mathcal{X} \times \mathcal{A}} \left| P_1^h((\tilde{a}, y')|(a, y)) - \Pi_1(\tilde{a}, y') \right| \sum_{(z, a') \in \mathcal{X} \times \mathcal{A}} P_1((a', z)|(\tilde{a}, y')) \\ &= 2d_{TV} \left(P_1^h(\cdot|(a, y)), \Pi_1 \right), \end{aligned}$$

where $(*)$ follows from Lemma 2. □

We now complete the proof of Proposition 5 (iii): from Lemma 3, we have that for all $h \geq 1$ and all $(x, a, y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$,

$$d_{TV} \left((P_2^2)^h(\cdot|(x, a, y)), \Pi_2 \right) \leq d_{TV} \left((P_1)^{2h-1}(\cdot|(a, y)), \Pi_1 \right).$$

In particular, this implies that $t_{\text{mix}, MC_2} \leq \frac{t_{\text{mix}, MC_1} + 1}{2} \leq \frac{t_{\text{mix}, MC_1}}{2} + 1 \leq \eta^2 + 1$.

D Proof of Theorem 1 – Fundamental Lower Bound of Latent State Decoding

The proof is based upon an appropriate *change-of-measure* argument (Lai and Robbins, 1985). The clustering error rate lower bound in SBMs (Yun and Proutière, 2014, 2016) and Block Markov Chains (Sanders et al., 2020) also leveraged a change-of-measure argument, but different than ours. More precisely, our confusing model is constructed by first fixing a specific context and moving it to some other cluster. Of course, the q 's (emission probabilities) are changed appropriately. We note that since each cluster is inhomogeneous (the emission distributions are not uniform), we derive a clustering error rate lower bound for each context $x \in \mathcal{X}$.

The confusing model. Denote the T observed trajectories as $\mathcal{T} = \{\mathcal{T}^{(t)}\}_{t=1}^T$. Fix a context x and denote by $i = f(x)$. Note that $\{x \in \mathcal{E}\} = \{\hat{f}(x) \neq i\}$ is the event that x is mis-classified, where we recall that \mathcal{E} is the set of mis-classified contexts under ρ and our chosen algorithm \mathcal{A} . Let Φ be the true BMDP model, induced by (p, q, f) , from which \mathcal{T} is ‘‘actually’’ generated. We define the confusing BMDP model by moving x from its original cluster i to some other cluster $j \neq i$, which will be determined later on. More precisely, let $\Psi^{(x,j)}$ be the confusing model, induced by $(p, \tilde{q}, \tilde{f})$, where $\tilde{f}(x) = j$ and $\tilde{f} \equiv f$ on all other contexts and \tilde{q} , the context emission distribution of $\Psi^{(x,j)}$, is defined as follows:

$$\tilde{q}(x|j) = cq(x|i), \quad \tilde{q}(y|j) = (1 - cq(x|i))q(y|j), \quad y \in f^{-1}(j) \setminus \{x\}, \quad (24)$$

$$\tilde{q}(z|i) = \frac{q(z|i)}{1 - q(x|i)}, \quad z \in f^{-1}(i) \setminus \{x\}, \quad (25)$$

$$\tilde{q}(y|s) = q(y|s), \quad y \in f^{-1}(s), s \in \mathcal{S} \setminus \{i, j\}. \quad (26)$$

Here, $c \geq 0$ is to be chosen later. We now provide the possible values taken by c , so that $\Psi^{(x,j)}$ is a possible BMDP compatible with the β -locality assumption (for $\beta \gtrsim \frac{1}{n}$) at Φ . First of all, for \tilde{q} to be a well-defined probability distribution, we must have that $\tilde{q} \in [0, 1]$. From the regularity condition on q , we have that $0 \leq c \leq \frac{1}{\eta^2} \frac{n}{S}$. Now, from the β -locality, we must have that

$$\left| \frac{q(y|i)}{1 - q(x|i)} - q(y|i) \right| \leq \beta, \quad |(1 - cq(x|i))q(z|j) - q(z|j)| \leq \beta.$$

The first inequality is trivially true for $\beta \geq 2\eta^2 \frac{S}{n}$, with sufficiently large n , or precisely speaking, with $n \geq 2\eta^2 S$. The second inequality is true for $0 \leq c \leq \frac{2}{\eta^2} \frac{n}{S}$. Overall, the domain of c is given as $0 \leq c \leq \frac{1}{\eta^2} \frac{n}{S}$. This completes the description of $\Psi^{(x,j)}$. To simplify the notation, we use Ψ to represent $\Psi^{(x,j)}$ in the remaining of the proof.

Log-likelihood ratio and its connection to the error rate. The log-likelihood ratio of the observed trajectories under Φ and Ψ is

$$\mathcal{L} = \log \frac{\mathbb{P}_\Psi[\{\mathcal{T}^{(t)}\}_{t=1}^T]}{\mathbb{P}_\Phi[\{\mathcal{T}^{(t)}\}_{t=1}^T]} = \sum_{t=1}^T \underbrace{\log \frac{\mathbb{P}_\Psi[\mathcal{T}^{(t)}|\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(t-1)}]}{\mathbb{P}_\Phi[\mathcal{T}^{(t)}|\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(t-1)}]}}_{\triangleq \mathcal{L}^{(t)}}, \quad (27)$$

with by convention $\mathcal{T}^{(-1)} = \emptyset$ and $\mathcal{T}^{(t)} = (x_1^{(t)}, a_1^{(t)}, \dots, x_{H-1}^{(t)}, a_{H-1}^{(t)}, x_H^{(t)}, a_H^{(t)})$. In our case in which the policies are non-adaptive, the conditioning is meaningless i.e. $\mathbb{P}[\mathcal{T}^{(t)}|\mathcal{T}^{(1)}, \dots, \mathcal{T}^{(t-1)}] = \mathbb{P}[\mathcal{T}^{(t)}]$. The conditioning will be important when considering adaptive behavior policies (refer to Appendix K). The following proposition relates \mathcal{L} to the classification error rate of x .

Proposition 6. *For β -locally better-than-random clustering algorithm that outputs \hat{f} given some dataset (of trajectories), we have:*

$$\begin{aligned} (i) \quad & \mathbb{P}_\Psi[\hat{f}(x) \neq f(x)] \geq \frac{1}{S}, \\ (ii) \quad & \varepsilon_x = \mathbb{P}_\Phi[x \in \mathcal{E}] \geq \frac{1}{2S} \exp\left(-\mathbb{E}_\Psi[\mathcal{L}] - \sqrt{2S \text{Var}_\Psi[\mathcal{L}]}\right). \end{aligned}$$

Proof. The proof is analogous to that of Proposition 4 of Sanders et al. (2020).

Proof of (i)

From the definition of β -locally better-than-random algorithm (Definition 1),

$$\mathbb{P}_\Psi[\hat{f}(x) \neq f(x)] \geq \mathbb{P}_\Psi[\hat{f}(x) = \tilde{f}(x)] = 1 - \mathbb{P}_\Psi[\hat{f}(x) \neq \tilde{f}(x)] \geq 1 - \left(1 - \frac{1}{S}\right) = \frac{1}{S},$$

where the first inequality follows from the observation that $\{\hat{f}(x) = \tilde{f}(x)\} \subset \{\hat{f}(x) \neq f(x)\}$.

Proof of (ii)

For simplicity, we denote the event $\xi_x = \{\hat{f}(x) \neq f(x)\}$. First, consider any function $R(n, T) : \mathbb{N}_+^2 \rightarrow \mathbb{R}$. We have:

$$\mathbb{P}_\Psi[\mathcal{L} \leq R(n, T)] = \mathbb{P}_\Psi[\mathcal{L} \leq R(n, T), \xi_x^c] + \mathbb{P}_\Psi[\mathcal{L} \leq R(n, T), \xi_x].$$

From (i), we deduce:

$$\mathbb{P}_\Psi[\mathcal{L} \leq R(n, T), \xi_x^c] \leq \mathbb{P}_\Psi[\xi_x^c] = 1 - \mathbb{P}_\Psi[\xi_x] \leq 1 - \frac{1}{S}.$$

From our change-of-measure,

$$\mathbb{P}_\Psi[\mathcal{L} \leq R(n, T), \xi_x] \leq e^{R(n, t)} \mathbb{P}_\Phi[\mathcal{L} \leq R(n, T), \xi_x] \leq e^{R(n, T)} \mathbb{P}_\Phi[\xi_x].$$

Combining the above results, we obtain:

$$\mathbb{P}_\Psi[\mathcal{L} \geq R(n, T)] \geq 1 - \left(1 - \frac{1}{S}\right) - e^{R(n, T)} \mathbb{P}_\Phi[\xi_x].$$

Specify $R(n, T) = \log \frac{1}{2S} + \log \frac{1}{\mathbb{P}_\Phi[\xi_x]}$ and apply Chebyshev's inequality to get

$$\mathbb{P}_\Psi \left[\mathcal{L} \geq \mathbb{E}_\Psi[\mathcal{L}] + \sqrt{2S \text{Var}_\Psi[\mathcal{L}]} \right] \leq \frac{1}{2S} \leq \mathbb{P}_\Psi \left[\mathcal{L} \geq \log \frac{1}{2S} - \log \mathbb{P}_\Phi[\xi_x] \right],$$

from which the result follows. □

The rate (or divergence) function and its connection to the log-likelihood ratio. We first introduce the following divergence or rate function:

$$\begin{aligned} I_j^{(t)}(x; c, \Phi) &:= n \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left\{ cq(x|f(x))p(j|s, a)m_\rho^{\Psi, (t)}(s, a) \log \frac{cp(j|s, a)}{p(f(x)|s, a)} \right. \\ &\quad + cq(x|f(x))m_\rho^{\Psi, (t)}(j, a)p(s|j, a) \log \frac{p(s|j, a)}{p(s|f(x), a)} \\ &\quad \left. + (1 - cq(x|f(x))p(j|s, a))m_\rho^{\Psi, (t)}(s, a) \log \frac{1 - cq(x|f(x))p(j|s, a)}{1 - q(x|f(x))p(f(x)|s, a)} \right\}, \end{aligned} \quad (28)$$

where $m_\rho^{\Psi, (t)}(s, a)$ denotes the expected proportion of rounds spent in (latent state, action) pair (s, a) under policy ρ and model Ψ , in the t -th episode:

$$m_\rho^{\Psi, (t)}(s, a) := \frac{1}{H-1} \sum_{h=1}^{H-1} \mathbb{P}_\Psi[\tilde{f}(x_h^{(t)}) = s, a_h^{(t)} = a]. \quad (29)$$

Note that since the behavior policy is not changing from one episode to the other, $m_\rho^{\Psi, (t)}(s, a)$ and hence $I_j^{(t)}(x; c, \Phi)$ do not depend on t . However, we keep separating the different episodes, so that the analysis will remain valid under adaptive behavior policies; see Appendix K for more detailed discussions.

Next, we define $I_j(x; c, \Phi) := \frac{1}{T} \sum_{t=1}^T I_j^{(t)}(x; c, \Phi)$, and finally, the rate function:

$$I(x; \Phi) := \min_{j: j \neq f(x)} \inf_{c > 0} I_j(x; c, \Phi). \quad (30)$$

Note that the choice of the "most" confusing cluster j^* as well as the "optimal" c^* depends on x , i.e., such choices can (and will differ, generally) for each $x \in \mathcal{X}$.

We now consider two cases; when $I(x; \Phi) = 0$ and when $I(x; \Phi) > 0$.

Proposition 7. *If $I(x; \Phi) = 0$, then $\mathbb{P}_\Phi[\hat{f}(x) \neq f(x)] \geq \alpha_{\min}$.*

Proof. Proposition 10 (proved in Section D.3) asserts that $I(x; \Phi) = 0$ iff there exists some $j \neq f(x)$ and $c > 0$ such that

1. $p(f(x)|s, a) = cp(j|s, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$,
2. $p(s|f(x), a) = p(s|j, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$.

We now show that when $I(x; \Phi) = 0$, we can construct a different BMDP model in which clusters $i = f(x)$ and j are merged to a single cluster, yet the likelihoods for any given observation(trajectory) are the same for this alternate model or the true model. This then implies that the clustering error rate of the algorithm is at least α_{\min} .

Our original model is driven by the contextual transition kernel $P(y|x, a) = p(f(y)|f(x), a)q(y|f(y))$ for any $(x, y, a) \in \mathcal{X}^2 \times \mathcal{A}$. For our alternate model, let k be the merged cluster index of $f(x)$ and j , and first denote $\tilde{\mathcal{S}} = \mathcal{S} \setminus \{f(x), j\}$. Then the alternate set of latent states is $\mathcal{S} = \mathcal{S}' \cup \{k\}$. We then define f', p' , and q' as follows:

1. $p'(s|k, a) = p(s|j, a)$ and $q'(\cdot|s) = q(\cdot|s)$ for any $(s, a) \in \tilde{\mathcal{S}} \times \mathcal{A}$,
2. $p'(k|s, a) = (1 + c)p(j|s, a)$ for any $(s, a) \in \times \mathcal{A}$,
3. $q'(y|k) = \frac{c}{1+c}q(y|i)$ and $q'(z|k) = \frac{1}{1+c}q(z|j)$ for any $y \in f^{-1}(i)$ and any $z \in f^{-1}(j)$,
4. $p'(k|k, a) = (1 + c)p(j|i, a)$.

Then it is straightforward to check that 1. p' and q' 's are all well-defined probability distributions over their respective supports and more importantly, 2. the alternate contextual transition kernel $P'(y|x, a) = p'(f'(y)|f'(x), a)q'(y|f'(y))$ is *identical* to that of P . As $P' = P$, for any given observation(trajectory), the likelihoods under true and alternate models are exactly the same, and we are done. \square

Now, assuming that $I(x; \Phi) > 0$ i.e. $I_j(x; c, \Phi) > 0$ for any $j \neq f(x)$ and $c > 0$, the next propositions assert that $\mathbb{E}_\Psi[\mathcal{L}]$ is precisely the leading term with $I_j(x; c, \Phi)$ to be defined later, and $\text{Var}_\Psi[\mathcal{L}]$ is negligible:

Proposition 8. $\mathbb{E}_\Psi[\mathcal{L}] \leq \frac{TH}{n}(I_j(x; c, \Phi) + \mathcal{O}(\frac{1}{n}))$.

Proposition 9. $\text{Var}_\Psi[\mathcal{L}] \leq \mathcal{O}(\frac{TH}{n})$.

Combining all the above results will complete the proof of Theorem 1. Denote by (j^*, c^*) such that $I(x; \Phi) = I_{j^*}(x; c^*, \Phi)$. Now Propositions 6, 7, 8 and 9 applied to (j^*, c^*) imply:

$$\mathbb{P}_\Phi[\hat{f}(x) \neq f(x)] \geq \frac{1}{2\eta S} \exp\left(-I(x; \Phi) \frac{TH}{n} (1 + o(1))\right).$$

D.1 Proof of Proposition 8

Here we compute $\mathbb{E}_\Psi[\mathcal{L}]$. We fix our attention to episode t and focus on $\mathcal{L}^{(t)}$. For full generality, we consider a more general form of behavior policy that can change per episode i.e. $\rho = (\rho^{(t)})_{t \in [T]}$. Recall that $x_h^{(t)}$ is the context observed in step h of this episode. Note that the corresponding latent state $s_h^{(t)}$ may depend on the model Φ or Ψ considered. Further recall that the transition kernels under the two models are:

$$\begin{aligned} P^\Phi(y|x, a) &= q(y|f(y))p(f(y)|f(x), a), \\ P^\Psi(y|x, a) &= \tilde{q}(y|\tilde{f}(y))p(\tilde{f}(y)|\tilde{f}(x), a). \end{aligned}$$

Then the likelihoods of the observed trajectory on the t -episode under both models are:

$$\begin{aligned} \mathbb{P}_\Phi[\mathcal{T}^{(t)}] &= \left[\mu(x_1^{(t)}) \prod_{h=1}^H \rho^{(t)}(a_h^{(t)} | x_h^{(t)}) \right] \left[\prod_{h=1}^{H-1} P^\Phi(x_{h+1}^{(t)} | x_h^{(t)}, a_h^{(t)}) \right], \\ \mathbb{P}_\Psi[\mathcal{T}^{(t)}] &= \left[\mu(x_1^{(t)}) \prod_{h=1}^H \rho^{(t)}(a_h^{(t)} | x_h^{(t)}) \right] \left[\prod_{h=1}^{H-1} P^\Psi(x_{h+1}^{(t)} | x_h^{(t)}, a_h^{(t)}) \right]. \end{aligned}$$

We deduce that:

$$\mathbb{E}_\Psi[\mathcal{L}^{(t)}] = \mathbb{E}_\Psi \left[\sum_{h=1}^{H-1} \log \frac{P^\Psi(x_{h+1}|x_h^{(t)}, a_h^{(t)})}{P^\Phi(x_{h+1}|x_h^{(t)}, a_h^{(t)})} \right] = \sum_{(z,a,y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}} \mathbb{E}_\Psi \left[\hat{N}_a^{(t)}(z,y) \right] \log \frac{P^\Psi(y|z,a)}{P^\Phi(y|z,a)},$$

where $\hat{N}_a^{(t)}(z,y) := \sum_{h=1}^{H-1} \text{indicator}[x_h^{(t)} = z, a_h^{(t)} = a, x_{h+1}^{(t)} = y]$. Note that the dependency on the behavior policy $\rho^{(t)}$ becomes implicit in $\hat{N}_a^{(t)}(z,y)$. In what follows, we use the following notations: $N_a^{\Psi,(t)}(z,y) := \mathbb{E}_\Psi[\hat{N}_a^{(t)}(z,y)]$, and for any subsets X, Y of the set of contexts \mathcal{X} , $N_a^{\Psi,(t)}(X,Y) := \sum_{z \in X, y \in Y} N_a^{\Psi,(t)}(z,y)$.

Let us now simplify $\mathbb{E}_\Psi[\mathcal{L}^{(t)}]$. Note that the terms for z and y involved in $\mathbb{E}_\Psi[\mathcal{L}^{(t)}]$ are not equal to zero only if $x \in \{z, y\}$. There are three disjoint cases to consider. Recalling that $i = f(x)$,

Case 1. $z \neq x, y = x$:

$$\sum_{a \in \mathcal{A}} \sum_{z \neq x} N_a^{\Psi,(t)}(z,x) \log \frac{P^\Psi(x|z,a)}{P^\Phi(x|z,a)} = \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} N_a^{\Psi,(t)}(\tilde{f}^{-1}(s),x) \log \frac{cp(j|s,a)}{p(i|s,a)} - \sum_{a \in \mathcal{A}} N_a^{\Psi,(t)}(x,x) \log \frac{cp(j|j,a)}{p(i|j,a)}.$$

Case 2. $z = x, y \neq x$:

$$\begin{aligned} \sum_{a \in \mathcal{A}} \sum_{y \neq x} N_a^{\Psi,(t)}(x,y) \log \frac{P^\Psi(y|x,a)}{P^\Phi(y|x,a)} &= \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} N_a^{\Psi,(t)}(x, \tilde{f}^{-1}(s)) \log \frac{p(s|j,a)}{p(s|i,a)} \\ &\quad - \sum_{a \in \mathcal{A}} \left[N_a^{\Psi,(t)}(x, \tilde{f}^{-1}(i)) \log(1 - q(x|i)) + N_a^{\Psi,(t)}(x,x) \log \frac{p(j|j,a)}{p(j|i,a)} \right]. \end{aligned}$$

Case 3. $z = y = x$:

$$\sum_{a \in \mathcal{A}} N_a^{\Psi,(t)}(x,x) \log \frac{P^\Psi(x|x,a)}{P^\Phi(x|x,a)} = \sum_{a \in \mathcal{A}} N_a^{\Psi,(t)}(x,x) \log \frac{cp(j|j,a)}{p(i|i,a)}.$$

Combining the three cases yields:

$$\begin{aligned} \mathbb{E}_\Psi[\mathcal{L}^{(t)}] &= \frac{H-1}{n} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\frac{n}{H-1} N_a^{\Psi,(t)}(\tilde{f}^{-1}(s),x) \log \frac{cp(j|s,a)}{p(i|s,a)} + \frac{n}{H-1} N_a^{\Psi,(t)}(x, \tilde{f}^{-1}(s)) \log \frac{p(s|j,a)}{p(s|i,a)} \right] \\ &\quad + \underbrace{\frac{H-1}{n} \sum_{a \in \mathcal{A}} \left[\frac{n}{H-1} N_a^{\Psi,(t)}(x,x) \log \frac{p(i|j,a)p(j|i,a)}{p(i|i,a)p(j|j,a)} - \frac{n}{H-1} N_a^{\Psi,(t)}(x, \tilde{f}^{-1}(i)) \log(1 - q(x|i)) \right]}_{:= \Lambda_1^{(t)}}. \end{aligned}$$

From Proposition 14 (proved in Section D.3.3), it can be easily seen that $|\Lambda_1^{(t)}| = \mathcal{O}(1/n) = o(1)$.

Now we can relate $m_\rho^{\Psi,(t)}(s,a)$ to the $N_a^{\Psi,(t)}(s,x)$:

$$\begin{aligned} N_a^{\Psi,(t)}(s,x) &= (H-1)cq(x|i)p(j|s,a)m_\rho^{\Psi,(t)}(s,a), \\ N_a^{\Psi,(t)}(x,s) &= (H-1)cq(x|i)p(s|j,a)m_\rho^{\Psi,(t)}(j,a). \end{aligned}$$

Then, recalling the definition of $I_j^{(t)}(x; c, \Phi)$ (Eqn. (28)), we obtain:

$$\mathbb{E}_\Psi[\mathcal{L}^{(t)}] = \frac{H-1}{n} I_j^{(t)}(x; c, \Phi) + \frac{H-1}{n} \Lambda^{(t)},$$

where

$$\Lambda^{(t)} = \Lambda_1^{(t)} - \underbrace{\sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} n(1 - cq(x|i)p(j|s,a))m_\rho^{\Psi,(t)}(s,a) \log \frac{1 - cq(x|i)p(j|s,a)}{1 - q(x|i)p(i|s,a)}}_{:= \Lambda_2^{(t)}}.$$

We now conclude the proof of Proposition 8 by summing the above equalities over t . Recall that $I_j(x; c, \Phi) = \frac{1}{T} \sum_{t=1}^T I_j^{(t)}(x; c, \Phi)$ which implies:

$$\mathbb{E}_\Psi[\mathcal{L}] = \sum_{t=1}^T \mathbb{E}_\Psi[\mathcal{L}^{(t)}] = \frac{T(H-1)}{n} (I(x; \Phi) + \Lambda),$$

where $\Lambda := \frac{1}{T} \sum_{t=1}^T \Lambda^{(t)}$. The proof is completed by applying Lemma 4. \square

Lemma 4. $|\Lambda| = \mathcal{O}(1/n)$.

Proof. We show that for all t , $|\Lambda^{(t)}| = \mathcal{O}(1/n)$.

We first note from the fact that $|\log(1-x) + x| \leq x^2$ for $|x| = o(1)$, we have $\left| \log \frac{1-cp(j|s,a)q(x|i)}{1-p(i|s,a)q(x|i)} \right| = \mathcal{O}\left(\frac{1}{n^2}\right)$. Because $\sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} m_\rho^{\Psi, (t)}(s, a) = 1$, $|\Lambda_2^{(t)}|$ can be bounded as follows:

$$\left| \Lambda_2^{(t)} \right| \leq n \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} m_\rho^{\Psi, (t)}(s, a) \left| \log \frac{1-cp(j|s,a)q(x|i)}{1-p(i|s,a)q(x|i)} \right| = \mathcal{O}\left(\frac{1}{n}\right).$$

Hence: $|\Lambda^{(t)}| \leq |\Lambda_1^{(t)}| + |\Lambda_2^{(t)}| \leq \mathcal{O}(1/n)$. \square

D.2 Proof of Proposition 9

We now compute $\text{Var}_\Psi[\mathcal{L}]$. For our analysis, we can just compute the variance for the model Ψ constructed using the "optimal" choices j^* and c^* . As $\mathcal{L}^{(t)}$'s are independent, we have that $\text{Var}_\Psi[\mathcal{L}] = \sum_{t=1}^T \text{Var}_\Psi[\mathcal{L}^{(t)}]$, and thus we fix some t and compute $\text{Var}_\Psi[\mathcal{L}^{(t)}]$. To simplify the notations, we ignore the dependency on t throughout the proof. Denote $L_h = \log \frac{P^\Psi(x_{h+1}|x_h, a_h)}{P^\Phi(x_{h+1}|x_h, a_h)}$. We have:

$$\text{Var}_\Psi[\mathcal{L}^{(t)}] = \sum_{h, h'=1}^H \text{Cov}_\Psi[L_h, L_{h'}] = \sum_{h, h'=1}^H \{\mathbb{E}_\Psi[L_h L_{h'}] - \mathbb{E}_\Psi[L_h] \mathbb{E}_\Psi[L_{h'}]\}. \quad (31)$$

Note that if $I(x; \Phi) = 0$, then we know from previous subsection that $\text{Var}_\Psi[\mathcal{L}^{(t)}] = 0$ as well, and so let us assume that $I(x; \Phi) > 0$. Recall that $f(n) \asymp \tilde{f}(n)$ (resp. $f(n) \lesssim \tilde{f}(n)$) denotes $f(n) = \Theta(\tilde{f}(n))$ (resp. $f(n) = \mathcal{O}(\tilde{f}(n))$). We divide up the computation of the r.h.s. of Eqn. (31) into three parts:

Part 1. $h' = h$:

$$\begin{aligned} \text{Cov}_\Psi[L_h, L_h] &\leq \mathbb{E}_\Psi[L_h^2] \\ &= \sum_{(z,a) \in \mathcal{X} \times \mathcal{A}} \mathbb{P}_\Psi[x_h = z] \rho(a|z) \sum_{y \in \mathcal{X}} P^\Psi(y|z, a) \left(\log \frac{P^\Psi(y|z, a)}{P^\Phi(y|z, a)} \right)^2 \\ &\asymp \frac{1}{n^2} \sum_{a \in \mathcal{A}} \left\{ \rho(a|x) \sum_{y \in \mathcal{X} \setminus \{x\}} \left(\log \frac{P^\Psi(y|x, a)}{P^\Phi(y|x, a)} \right)^2 + \sum_{z \in \mathcal{X} \setminus \{x\}} \rho(a|z) \left(\log \frac{P^\Psi(x|z, a)}{P^\Phi(x|z, a)} \right)^2 \right\} \\ &\lesssim \frac{1}{n}, \end{aligned}$$

and thus $\sum_{h'=h} \text{Cov}_\Psi[L_h, L_{h'}] \lesssim \frac{H}{n}$.

Part 2. $h' = h + 1$:

We first note that for any $h, h' \in [H]$, $\mathbb{E}_\Psi[L_h] \asymp \mathbb{E}_\Psi[L_{h'}] \geq 0$ as

$$\mathbb{E}_\Psi[L_h] = \sum_{(z,a) \in \mathcal{X} \times \mathcal{A}} \mathbb{P}_\Psi[x_h = z] \rho(a|z) \sum_{y \in \mathcal{X}} P^\Psi(y|z, a) \log \frac{P^\Psi(y|z, a)}{P^\Phi(y|z, a)}$$

$$= \sum_{(z,a) \in \mathcal{X} \times \mathcal{A}} \mathbb{P}_\Psi[x_h = z] \rho(a|z) \text{KL} \left(P^\Psi(\cdot|z, a) \parallel P^\Phi(\cdot|z, a) \right) \geq 0$$

and $\mathbb{P}_\Psi[x_h = z] \asymp \mathbb{P}_\Psi[x_{h'} = z] \asymp \frac{1}{n}$. Hence

$$\text{Cov}_\Psi[L_h, L_{h'}] \asymp \mathbb{E}_\Psi[L_h L_{h'}] - \mathbb{E}_\Psi[L_h \wedge L_{h'}]^2 \leq \mathbb{E}_\Psi[L_h L_{h'}]. \quad (32)$$

Thus,

$$\begin{aligned} \text{Cov}_\Psi[L_h, L_{h+1}] &\lesssim \mathbb{E}_\Psi[L_h L_{h+1}] \\ &= \sum_{(z,a) \in \mathcal{X} \times \mathcal{A}} \mathbb{P}_\Psi[x_h = z] \rho(a|z) \sum_{y \in \mathcal{X}} P^\Psi(y|z, a) \\ &\quad \times \sum_{(a', z') \in \mathcal{A} \times \mathcal{X}} \rho(a'|y) P^\Psi(z'|y, a') \left(\log \frac{P^\Psi(y|z, a)}{P^\Phi(y|z, a)} \right) \left(\log \frac{P^\Psi(z'|y, a')}{P^\Phi(z'|y, a')} \right) \\ &\asymp \frac{1}{n^3} \sum_{a, a' \in \mathcal{A}} \left\{ \sum_{z, y, z' \in \mathcal{X}} \rho(a|z) \rho(a'|y) \left(\log \frac{P^\Psi(y|z, a)}{P^\Phi(y|z, a)} \right) \left(\log \frac{P^\Psi(z'|y, a')}{P^\Phi(z'|y, a')} \right) \right\} \lesssim \frac{1}{n}, \end{aligned}$$

and thus $\sum_{h'=h+1} \text{Cov}_\Psi[L_h, L_{h'}] \lesssim \frac{H}{n}$.

Part 3. $h' \geq h + 2$:

From $\text{Cov}_\Psi[L_h, L_{h'}] = \mathbb{E}_\Psi[L_h L_{h'}] - \mathbb{E}_\Psi[L_h] \mathbb{E}_\Psi[L_{h'}]$, we compute each term separately:

$$\begin{aligned} &\mathbb{E}_\Psi[L_h L_{h'}] \\ &= \sum_{\substack{y_h, y_{h+1}, y_{h'}, y_{h'+1} \in \mathcal{X} \\ b_h, b_{h'} \in \mathcal{A}}} \mathbb{P}_\Psi[x_h = y_h] \rho(b_h|y_h) P^\Psi(y_{h+1}|y_h, a_h) \\ &\quad \times \left(\sum_{\substack{y_{h+2}, \dots, y_{h'-1} \in \mathcal{X} \\ b_{h+1}, \dots, b_{h'-1} \in \mathcal{A}}} \prod_{j=h+1}^{h'-1} \rho(b_j|y_j) P^\Psi(y_{j+1}|y_j, b_j) \right) \\ &\quad \times \rho(b_{h'}|y_{h'}) P^\Psi(y_{h'+1}|y_{h'}, b_{h'}) \left(\log \frac{P^\Psi(y_{h+1}|y_h, b_h)}{P^\Phi(y_{h+1}|y_h, b_h)} \right) \left(\log \frac{P^\Psi(y_{h'+1}|y_{h'}, b_{h'})}{P^\Phi(y_{h'+1}|y_{h'}, b_{h'})} \right) \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}_\Psi[L_h] \mathbb{E}_\Psi[L_{h'}] \\ &= \left(\sum_{\substack{y_h, y_{h+1} \in \mathcal{X} \\ b_h \in \mathcal{A}}} \mathbb{P}_\Psi[x_h = y_h] \rho(b_h|y_h) P^\Psi(y_{h+1}|y_h, a_h) \left(\log \frac{P^\Psi(y_{h+1}|y_h, b_h)}{P^\Phi(y_{h+1}|y_h, b_h)} \right) \right) \\ &\quad \left(\sum_{\substack{y_{h'}, y_{h'+1} \in \mathcal{X} \\ b_{h'} \in \mathcal{A}}} \mathbb{P}_\Psi[x_{h'} = y_{h'}] \rho(b_{h'}|y_{h'}) P^\Psi(y_{h'+1}|y_{h'}, a_{h'}) \left(\log \frac{P^\Psi(y_{h'+1}|y_{h'}, b_{h'})}{P^\Phi(y_{h'+1}|y_{h'}, b_{h'})} \right) \right) \\ &= \sum_{\substack{y_h, y_{h+1}, y_{h'}, y_{h'+1} \in \mathcal{X} \\ b_h, b_{h'} \in \mathcal{A}}} \mathbb{P}_\Psi[x_h = y_h] \rho(a_h|y_h) P^\Psi(y_{h+1}|y_h, a_h) \mathbb{P}_\Psi[x_{h'} = y_{h'}] \rho(b_{h'}|y_{h'}) P^\Psi(y_{h'+1}|y_{h'}, b_{h'}) \\ &\quad \left(\log \frac{P^\Psi(y_{h+1}|y_h, b_h)}{P^\Phi(y_{h+1}|y_h, b_h)} \right) \left(\log \frac{P^\Psi(y_{h'+1}|y_{h'}, b_{h'})}{P^\Phi(y_{h'+1}|y_{h'}, b_{h'})} \right). \end{aligned}$$

Then,

$$\mathbb{E}_\Psi[L_h L_{h'}] - \mathbb{E}_\Psi[L_h] \mathbb{E}_\Psi[L_{h'}]$$

$$\begin{aligned}
 &= \sum_{\substack{y_h, y_{h+1}, y_{h'}, y_{h'+1} \in \mathcal{X} \\ b_h, b_{h'} \in \mathcal{A}}} \mathbb{P}_\Psi[x_h = y_h] \rho(b_h|y_h) P^\Psi(y_{h+1}|y_h, a_h) \rho(b_{h'}|y_{h'}) \\
 &\times P^\Psi(y_{h'+1}|y_{h'}, b_{h'}) Q_{h+1, h'}(y_{h+1}, y_{h'}) \left(\log \frac{P^\Psi(y_{h+1}|y_h, b_h)}{P^\Phi(y_{h+1}|y_h, b_h)} \right) \left(\log \frac{P^\Psi(y_{h'+1}|y_{h'}, b_{h'})}{P^\Phi(y_{h'+1}|y_{h'}, b_{h'})} \right),
 \end{aligned}$$

where

$$\begin{aligned}
 Q_{h+1, h'}(y_{h+1}, y_{h'}) &:= \sum_{\substack{y_{h+2}, \dots, y_{h'-1} \in \mathcal{X} \\ b_{h+1}, \dots, b_{h'-1} \in \mathcal{A}}} \prod_{j=h+1}^{h'-1} \rho(b_j|y_j) P^\Psi(y_{j+1}|y_j, b_j) - \mathbb{P}_\Psi[x_{h'} = y_{h'}] \\
 &= \mathbb{P}_\Psi[x_{h'} = y_{h'} | x_{h+1} = y_{h+1}] - \mathbb{P}_\Psi[x_{h'} = y_{h'}].
 \end{aligned}$$

We can relate $|Q_{h+1, h'}|$ to the mixing time of P^Ψ :

$$\begin{aligned}
 |Q_{h+1, h'}(y, z)| &= \left| (P^\Psi)^{h'-h-1}(y, z) - \mu(P^\Psi)^{h'-1}(z) \right| \\
 &\leq \left| (P^\Psi)^{h'-h-1}(y, z) - \Pi^\Psi(z) \right| + \left| \Pi^\Psi(z) - \mu(P^\Psi)^{h'-1}(z) \right| \\
 &\leq \sum_{z \in \mathcal{Z}} \left| (P^\Psi)^{h'-h-1}(y, z) - \Pi^\Psi(z) \right| + \sum_{z \in \mathcal{Z}} \left| \Pi^\Psi(z) - \mu(P^\Psi)^{h'-1}(z) \right| \\
 &\leq 2d_{TV}((P^\Psi)^{h'-h-1}(y, \cdot), \Pi^\Psi) + 2d_{TV}(\mu(P^\Psi)^{h'-1}, \Pi^\Psi) \\
 &\leq 2(\delta(P^\Psi))^{h'-h-2} d_{TV}(P^\Psi(y, \cdot), \Pi^\Psi) + 2(\delta(P^\Psi))^{h'-1} d_{TV}(\mu, \Pi^\Psi) \\
 &\lesssim (\delta(P^\Psi))^{h'-1}.
 \end{aligned}$$

Using the triangle inequality,

$$\begin{aligned}
 \left| \sum_{h' \geq h+2} \text{Cov}_\Psi[L_h, L_{h'}] \right| &\lesssim \frac{1}{n^3} \sum_{h' \geq h+2} \sum_{\substack{y_h, y_{h+1}, y_{h'}, y_{h'+1} \in \mathcal{X} \\ b_h, b_{h'} \in \mathcal{A}}} \rho(b_h|y_h) \rho(b_{h'}|y_{h'}) |Q(y_{h+1}, y_{h'})| \\
 &\quad \times \left| \left(\log \frac{P^\Psi(y_{h+1}|y_h, b_h)}{P^\Phi(y_{h+1}|y_h, b_h)} \right) \left(\log \frac{P^\Psi(y_{h'+1}|y_{h'}, b_{h'})}{P^\Phi(y_{h'+1}|y_{h'}, b_{h'})} \right) \right| \\
 &\lesssim \frac{1}{n} \sum_{h' \geq h+2} (\delta(P^\Psi))^{h'-1} \lesssim \frac{H}{n}.
 \end{aligned}$$

Finally, combining all above completes the proof of Proposition 9. \square

D.3 Properties of the rate function $I(x; \Phi)$

In this subsection, we provide properties of the rate function $I(x; \Phi)$ that are useful for the analysis. We also provide intermediate results used throughout our proofs.

D.3.1 Necessary and sufficient conditions for $I(x; \Phi) = 0$

Proposition 10. $I(x; \Phi) \geq 0$, and the equality holds iff there exist $j \in \mathcal{S}$ and $c > 0$ such that both of the following holds:

1. $p(f(x)|s, a) = cp(j|s, a)$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$,
2. $p(s|f(x), a) = p(s|j, a)$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$.

Before we prove the above proposition, we give its interpretation. It states that from the observations on x , we cannot distinguish whether x 's latent state is j or $f(x)$. Indeed, 1. represents the fact that the observations of transitions leading to context x are statistically equivalent in Φ and $\Psi^{(x, j)}$, and 2. represents the fact that the observations of transitions from context x are statistically equivalent in Φ and $\Psi^{(x, j)}$. The additional c takes into account the non-uniform emission probabilities of each context of each cluster.

Proof. $I(x; \Phi) = 0$ if and only if there exists j and c such that $I_j(x; c, \Phi) = 0$. We now show that $I_j(x; c, \Phi)$ is actually equal to a mixture of KL's. From there, the results will follow immediately. Let $m_\rho^\Psi(s, a) = m_\rho^{\Psi, (t)}(s, a)$.

Define $p_{k,a}^{\text{out}}(s) := p(s|k, a)$, and $r \in \mathcal{P}([2K] \times \mathcal{A})$ as

$$r(\bar{s}, a) := \begin{cases} m_\rho^\Psi(s, a)p(f(x)|s, a)q(x|f(x)) & (\bar{s} = 2s - 1), \\ m_\rho^\Psi(s, a)(1 - p(f(x)|s, a)q(x|f(x))) & (\bar{s} = 2s). \end{cases} \quad (33)$$

Analogously, we define

$$\tilde{r}(\bar{s}, a; c) := \begin{cases} m_\rho^\Psi(s, a)p(j|s, a)cq(x|f(x)) & (\bar{s} = 2s - 1), \\ m_\rho^\Psi(s, a)(1 - p(j|s, a)cq(x|f(x))) & (\bar{s} = 2s), \end{cases}. \quad (34)$$

Then it is easy to see that actually,

$$I_j(x; c, \Phi) = n\text{KL}(\tilde{r}(\cdot, \cdot; c) || r(\cdot, \cdot)) + ncq(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho^\Psi(f(x), a)\text{KL}(p_{j,a}^{\text{out}}(\cdot) || p_{f(x),a}^{\text{out}}(\cdot)). \quad (35)$$

The proposition then follows from the fact that $\text{KL}(p||q) \geq 0$ and $\text{KL}(p||q) = 0$ iff $p = q$ a.e. \square

D.3.2 Alternative KL-form of the rate function $I(x; \Phi)$

Next, we derive an alternative asymptotic KL-form for the rate function $I(x; \Phi)$ that will be useful later in the analysis of the algorithm. For simplicity, fix t . We denote $m_\rho := m_\rho^{\Phi, (t)}$. We first introduce the *alternative divergence*:

$$\tilde{I}(x; \Phi) := \min_{j: j \neq f(x)} \inf_{c > 0} \tilde{I}_j(x; c, \Phi), \quad (36)$$

where

$$\tilde{I}_j(x; c, \Phi) := n\text{KL}(p_{\Phi, x}^{\text{in}}(\cdot, \cdot) || p_{\Psi, x}^{\text{in}}(\cdot, \cdot; c)) + cnq(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho(f(x), a)\text{KL}(p_{f(x), a}^{\text{out}}(\cdot) || p_{j, a}^{\text{out}}(\cdot)). \quad (37)$$

and where $p_{k,a}^{\text{out}}(\cdot) := p(\cdot|k, a)$, and recalling that $\mathcal{S} = [K]$, for $\bar{s} \in [2K]$ and $a \in \mathcal{A}$,

$$p_{\Phi, x}^{\text{in}}(\bar{s}, a) := \begin{cases} m_\rho(s, a)p(f(x)|s, a)q(x|f(x)) & (\bar{s} = 2s - 1), \\ m_\rho(s, a)(1 - p(f(x)|s, a)q(x|f(x))) & (\bar{s} = 2s), \end{cases} \quad (38)$$

$$p_{\Psi, x}^{\text{in}}(\bar{s}, a; c) := \begin{cases} cm_\rho(s, a)p(j|s, a)q(x|f(x)) & (\bar{s} = 2s - 1), \\ m_\rho(s, a)(1 - cp(j|s, a)q(x|f(x))) & (\bar{s} = 2s), \end{cases}. \quad (39)$$

It's easy to check that $p_{k,a}^{\text{out}}(\cdot) \in \mathcal{P}(\mathcal{S})$ and $p_{\Phi, x}^{\text{in}}(\cdot, \cdot), p_{\Psi, x}^{\text{in}}(\cdot, \cdot; c) \in \mathcal{P}([2K] \times \mathcal{A})$. Note that compared to the actual divergence I , the order of Φ and Ψ is switched.

Roughly speaking, $p_{k,a}^{\text{out}}(s)$ describes the *outgoing* probabilities from the state-action pair (k, a) , and $p_{\Phi, x}^{\text{in}}(\bar{s}, a)$ (resp. $p_{\Psi, x}^{\text{in}}(\cdot, \cdot)$) describes the *inward* probabilities into the context x under Φ (resp. Ψ), up to order-wise negligible remainders. We then show the following of \tilde{I} :

Proposition 11. *For all $c > 0$, we have that $I_j(x; c, \Phi) = 0$ if and only if $\tilde{I}_j(x; c, \Phi) = 0$. Furthermore, we have that $\min(1, c, 1/c, 1/\eta)\tilde{I}_j(x; c, \Phi) \leq I_j(x; c, \Phi) \leq \max(1, c, 1/c, \eta)\tilde{I}_j(x; c, \Phi)$.*

Proof. The first part is trivial.

For the second part, we start with the following observations:

$$\frac{p_{\Psi, x}^{\text{in}}(\bar{s}, a)}{p_{\Phi, x}^{\text{in}}(\bar{s}, a; c)} = \begin{cases} c & (\bar{s} = 2s - 1), \\ \frac{1 - cp(j|s, a)q(x|f(x))}{1 - p(f(x)|s, a)q(x|f(x))} \simeq 1 & (\bar{s} = 2s), \end{cases} \\ \lesssim \max(c, 1),$$

and

$$\frac{p_{\Phi,x}^{in}(\bar{s}, a)}{p_{\Psi,x}^{in}(\bar{s}, a; c)} = \begin{cases} \frac{1}{c} & (\bar{s} = 2s - 1), \\ \frac{1 - p(f(x)|s, a)q(x|f(x))}{1 - cp(j|s, a)q(x|f(x))} \simeq 1 & (\bar{s} = 2s), \end{cases}$$

$$\lesssim \max(1/c, 1).$$

Also, we have that $m_\rho(s, a) \sim m_\rho^\Psi(s, a)$. From the characterization of I_j as a mixture of KL's (see proof of Proposition 10), we conclude by applying Lemma 5. \square

Proposition 12. *There exist nonnegative functions ψ_1, ψ_2 , independent of n , such that for all $c > 0$, $x \in \mathcal{X}$, $j \in \mathcal{S}$, and BMDP Φ (satisfying Assumptions 1-4), $\psi_1(p, \eta, c) \leq \tilde{I}_j(x; c, \Phi) \leq \psi_2(p, \eta, c)$. Consequently, $I_j(x; \Phi) = \inf_{c>0} I_j(x; c, \Phi)$ does not scale with n .*

Proof. For simplicity, denote $u_{s,a} = \frac{p(s|f(x), a)}{p(s|j, a)} \in [1/\eta, \eta]$ and $v_{s,a} = \frac{p(f(x)|s, a)}{p(j|s, a)} \in [1/\eta, \eta]$. Then,

$$\begin{aligned} & \tilde{I}_j(x; c, \Phi) \\ &= n\text{KL}\left(p_{\Phi,x}^{in}(\cdot, \cdot) \| p_{\Psi,x}^{in}(\cdot, \cdot; c)\right) + cnq(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho(f(x), a) \text{KL}\left(p_{f(x),a}^{out}(\cdot) \| p_{j,a}^{out}(\cdot)\right) \\ &\stackrel{(a)}{\leq} \frac{n}{2} \sum_{\bar{s}, a} \frac{\left(p_{\Phi,x}^{in}(\bar{s}, a) - p_{\Psi,x}^{in}(\bar{s}, a; c)\right)^2}{p_{\Psi,x}^{in}(\bar{s}, a; c)} + \frac{c}{2} nq(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho(f(x), a) \sum_s \frac{\left(p_{f(x),a}^{out}(s) - p_{j,a}^{out}(s)\right)^2}{p_{j,a}^{out}(s)} \\ &\leq \frac{n}{2} \sum_{s,a} m_\rho(s, a) q(x|f(x)) \left(\frac{1}{cp(j|s, a)} + \frac{q(x|f(x))}{1 - cp(j|s, a)q(x|f(x))} \right) (p(f(x)|s, a) - cp(j|s, a))^2 \\ &\quad + c\eta^2 S \frac{\eta^4}{SA} \sum_{s,a} p(s|j, a) \left(\frac{p(s|f(x), a)}{p(s|j, a)} - 1 \right)^2 \\ &\stackrel{(b)}{\leq} c\eta^7 \frac{1}{SA} \sum_{s,a} \left(\frac{v_{s,a}}{c} - 1 \right)^2 + c\eta^7 \frac{1}{SA} \sum_{s,a} (u_{s,a} - 1)^2 \\ &= \eta^7 \frac{1}{SA} \sum_{s,a} \left(\frac{v_{s,a}^2}{c} + c - 2v_{s,a} \right) + c\eta^7 \frac{1}{SA} \sum_{s,a} (u_{s,a} - 1)^2 \triangleq \psi_2(p, \eta, c) \end{aligned}$$

and

$$\begin{aligned} & \tilde{I}_j(x; c, \Phi) \\ &= n\text{KL}\left(p_{\Phi,x}^{in}(\cdot, \cdot) \| p_{\Psi,x}^{in}(\cdot, \cdot; c)\right) + cnq(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho(f(x), a) \text{KL}\left(p_{f(x),a}^{out}(\cdot) \| p_{j,a}^{out}(\cdot)\right) \\ &\stackrel{(a)}{\geq} \frac{n}{2} \sum_{\bar{s}, a} \frac{\left(p_{\Phi,x}^{in}(\bar{s}, a) - p_{\Psi,x}^{in}(\bar{s}, a; c)\right)^2}{p_{\Phi,x}^{in}(\bar{s}, a) \vee p_{\Psi,x}^{in}(\bar{s}, a; c)} + \frac{c}{2} nq(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho(f(x), a) \sum_s \frac{\left(p_{f(x),a}^{out}(s) - p_{j,a}^{out}(s)\right)^2}{p_{f(x),a}^{out}(s) \vee p_{j,a}^{out}(s)} \\ &\geq \frac{n}{2} \sum_{s,a} m_\rho(s, a) q(x|f(x)) \frac{(p(f(x)|s, a) - cp(j|s, a))^2}{p(f(x)|s, a) \vee cp(j|s, a)} \\ &\quad + \frac{n}{2} \sum_{s,a} m_\rho(s, a) q(x|f(x))^2 \frac{(p(f(x)|s, a) - cp(j|s, a))^2}{1 - q(x|f(x)) (p(f(x)|s, a) \wedge cp(j|s, a))} \\ &\quad + \frac{c}{2\eta^6} \frac{1}{A} \sum_{s,a} \frac{(p(s|f(x), a) - p(s|j, a))^2}{p(s|f(x), a) \vee p(s|j, a)} \\ &\stackrel{(b)}{\geq} \frac{1}{2\eta^6} \frac{1}{A} \sum_{s,a} \frac{(p(f(x)|s, a) - cp(j|s, a))^2}{p(f(x)|s, a) \vee cp(j|s, a)} + \frac{c}{2\eta^8} \frac{1}{SA} \sum_{s,a} (u_{s,a} - 1)^2 \end{aligned}$$

$$\geq \frac{1}{2\eta^7} \frac{c}{1 \vee \frac{\eta}{c}} \frac{1}{SA} \sum_{s,a} \left(\frac{v_{s,a}}{c} - 1 \right)^2 + \frac{c}{2\eta^8} \frac{1}{SA} \sum_{s,a} (u_{s,a} - 1)^2 \triangleq \psi_1(p, \eta, c)$$

where (a) follows from Lemma 6 and (b) follows from the observation that $0 \leq \frac{q(x|f(x))}{1 - cp(j|s,a)q(x|f(x))} = o(1)$. The last point follows from Proposition 11 and the inequality we proved above. \square

Lemma 5. *If p, q are (discrete) probability distributions with the same support \mathcal{Z} satisfying $\max_{z \in \mathcal{Z}} (p(z)/q(z), q(z)/p(z)) \leq \xi$ for some $\xi > 1$, then the following holds:*

$$\max \left(\frac{\text{KL}(p||q)}{\text{KL}(q||p)}, \frac{\text{KL}(q||p)}{\text{KL}(p||q)} \right) \leq \xi. \quad (40)$$

Proof. The statement then follows from the following lemma:

Lemma 6 (Lemma 19 of SM6.3 of Sanders et al. (2020) & Lemma 6.3 of Csiszar and Talata (2006)).

$$\frac{1}{2} \sum_{z \in \mathcal{Z}} \frac{(p(z) - q(z))^2}{p(z) \vee q(z)} \leq \text{KL}(p||q) \leq \frac{1}{2} \sum_{z \in \mathcal{Z}} \frac{(p(z) - q(z))^2}{q(z)}. \quad (41)$$

\square

D.3.3 Asymptotics of m_ρ

We provide properties of the quantity: $m_\rho(s, a) := \frac{1}{H-1} \sum_{h=1}^{H-1} \mathbb{P}_\Phi[f(x_h^{(t)}) = s, a_h^{(t)} = a]$ (there is no dependence in t). These properties are extensively used in the appendices, and they also hold for m_ρ^Ψ (up to some negligible remainders – actually, it can be easily seen that $m_\rho(s, a) \sim m_\rho^\Psi(s, a)$).

Proposition 13. *For $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$m_\rho(s, a) = \frac{1}{H-1} \sum_{h=1}^{H-1} \sum_{x \in f^{-1}(s)} \rho(a|x) \sum_{z \in \mathcal{X}} \mu(z) (P_0)^{h-1}(x|z), \quad (42)$$

where remember that $P_0 = \sum_{a \in \mathcal{A}} \rho(a|x) P(y|x, a)$, and by convention $\sum_z \mu(z) P_0^0(x|z) = \mu(x)$.

Proof. By definition,

$$\begin{aligned} m_\rho(s, a) &= \frac{1}{H-1} \sum_{h=1}^{H-1} \mathbb{P}_\Phi \left[f(x_h^{(t)}) = s, a_h^{(t)} = a \right] \\ &= \frac{1}{H-1} \sum_{h=1}^{H-1} \sum_{x \in f^{-1}(s)} \rho(a|x) \mathbb{P}_\Phi \left[x_h^{(t)} = x \right]. \end{aligned}$$

Now we have: $\mathbb{P}_\Phi \left[x_h^{(t)} = x \right] = \sum_z \mu(z) P_0^{h-1}(x|z)$ for $h \geq 1$. The result follows immediately. \square

Proposition 14. *The following holds for all $(s, a) \in \mathcal{S} \times \mathcal{A}$: with $\rho(a|s) := \frac{1}{n\alpha_s} \sum_{y \in f^{-1}(s)} \rho(a|y)$, $\rho \sim \mathcal{U}(\mathcal{A})$ and $\mu \sim \mathcal{U}(\mathcal{X})$,*

$$\frac{1}{\eta^4 SA} \leq \frac{1}{\eta^2} \alpha_s \rho(a|s) \leq m_\rho(s, a) \leq \eta^2 \alpha_s \rho(a|s) \leq \frac{\eta^4}{SA}. \quad (43)$$

i.e. $m_\rho(s, a) = \Theta\left(\frac{1}{SA}\right)$.

Proof. The result follows directly from Proposition 13 and from the results of Appendix C. \square

D.3.4 Examples of rate function $I(x; \Phi)$

For simplicity, we consider the following environment: $\mathcal{X} = \{x_1, \dots, x_{10}\}$, $\mathcal{S} = \{s_1, s_2\}$, $\mathcal{A} = \{a_1, a_2\}$, and $f : \mathcal{X} \rightarrow \mathcal{S}$ defined such that $f^{-1}(s_1) = \{x_1, x_3, x_5, x_7, x_9\}$ and $f^{-1}(s_2) = \{x_2, x_4, x_6, x_8, x_{10}\}$. We denote P_k to be the transition probability matrix corresponding to action a_k . Our exploration policy and initial distribution are set to be uniformly random, i.e., $\rho(\cdot|x) \sim \mathcal{U}(\mathcal{A})$ for all $x \in \mathcal{X}$ and $\mu(\cdot) \sim \mathcal{U}(\mathcal{X})$. As we only have two clusters, we have that $I(x; c, \Phi) = I_2(x; c, \Phi)$.

Furthermore, to clearly illustrate our change-of-measure argument, we fix the construction of alternate model Ψ as follows: change f to g such that $g(x_1) = s_2$ and rest stay the same. Also, we assume that q is uniform over its support (or its respective cluster), which implies that the alternate \tilde{q} is set as follows:

$$\tilde{q}(x_{2k-1}|s_1) = \frac{1}{4}, \quad \tilde{q}(x_1|s_2) = \frac{c}{5}, \quad \tilde{q}(x_{2k}|s_2) = \frac{5-c}{25}.$$

where $c \in (0, 5)$ is to be set later.

Lastly, we recall (from the proof of Proposition 10) that

$$I(x_1; c, \Phi) = n\text{KL}(\tilde{r}(\cdot, \cdot; c) || r(\cdot, \cdot)) + \frac{c}{5}n \sum_{a \in \mathcal{A}} m_\rho^\Psi(s_1, a) \text{KL}(p_{j,a}^{\text{out}}(\cdot) || p_{s_1,a}^{\text{out}}(\cdot)),$$

where $p_{k,a}^{\text{out}}(s) := p(s|k, a)$, and $r, \tilde{r} \in \mathcal{P}(\{s_1, \bar{s}_1, s_2, \bar{s}_2\} \times \mathcal{A})$ are defined as

$$r(\bar{s}, a) := \begin{cases} \frac{1}{5}m_\rho^\Psi(s, a)p(s_1|s_k, a) & (\bar{s} = s_k), \\ m_\rho^\Psi(s, a) \left(1 - \frac{1}{5}p(s_1|s_k, a)\right) & (\bar{s} = \bar{s}_k), \end{cases}$$

and

$$\tilde{r}(\bar{s}, a; c) := \begin{cases} \frac{c}{5}m_\rho^\Psi(s, a)p(s_2|s_k, a) & (\bar{s} = s_k), \\ m_\rho^\Psi(s, a) \left(1 - \frac{c}{5}p(s_2|s_k, a)\right) & (\bar{s} = \bar{s}_k), \end{cases}.$$

From Proposition 13, we have that

$$m_\rho^\Psi(s, a) = \frac{1}{20(H-1)} \sum_{h=1}^{H-1} \sum_{x \in g^{-1}(s)} \sum_{z \in \mathcal{X}} \left(\tilde{P}_0\right)^{h-1}(x|z),$$

where $\tilde{P}_0 = \frac{1}{2} \sum_{a \in \mathcal{A}} \tilde{P}(y|x, a)$ and $\tilde{P}(y|x, a) = p(g(y)|g(x), a)\tilde{q}(y|g(y))$.

For computational simplicity, we set the horizon length to be $H = 10$. All the computations were done using Mathematica.

$I(x_1; \Phi) > 0$ Consider the following instantiation:

$$P_1 = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

Then we have that

$$\begin{aligned} m_\rho^\Psi(s_1, a_1) &= m_\rho^\Psi(s_2, a_1) = \frac{73567181}{302330880} \approx 0.2433 \\ m_\rho^\Psi(s_1, a_2) &= m_\rho^\Psi(s_2, a_2) = \frac{77598259}{302330880} \approx 0.2567 \end{aligned}$$

With some more computations, we have the explicit form of the divergence I as follows:

$$\begin{aligned} I(x_1; c, \Phi) &= \frac{73567181}{4453496320} \left[2c \log 2 + (15 - 2c) \log \frac{15 - 2c}{14} + 3(10 - c) \log \frac{10 - c}{9} \right. \\ &\quad \left. + (15 - c) \log \frac{15 - c}{13} + 6c \log c \right] \end{aligned}$$

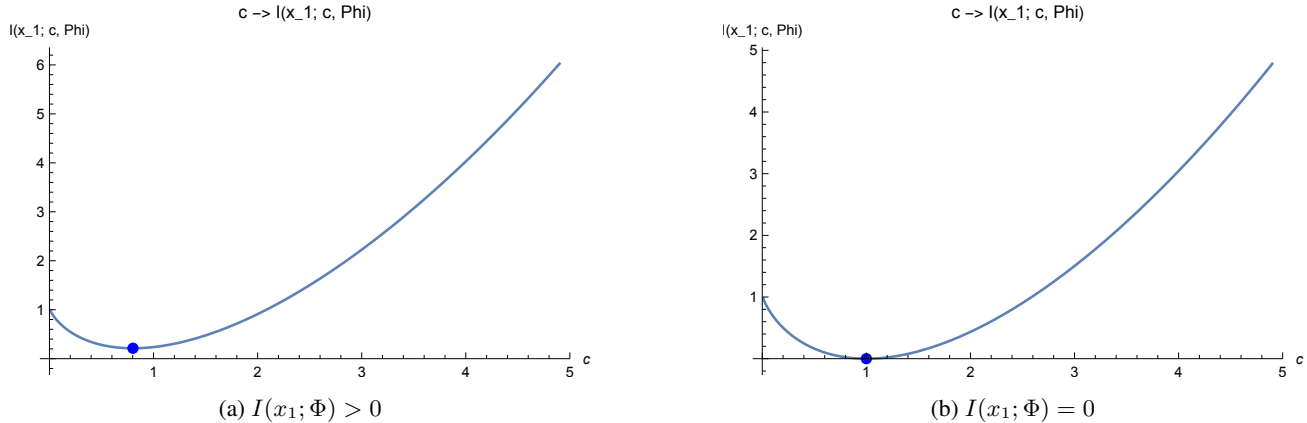


Figure 2: Plot of $I(x_1; c, \Phi)$ as function of c . Note that the domain of c is $(0, 5)$

and thus,

$$I(x_1; \Phi) = \inf_{c>0} I(x_1; c, \Phi) \approx 0.2127 > 0,$$

where the minimum is attained at $c \approx 0.8023$. We also provide a plot of $I(x_1; c, \Phi)$ as a function of c in Figure 2a.

$I(x_1; \Phi) = 0$ Consider the following instantiation:

$$P_1 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

Then we have that

$$m_\rho^\Psi(s_1, a_1) = \frac{11}{45} \approx 0.2444, \quad m_\rho^\Psi(s_1, a_2) = m_\rho^\Psi(s_2, a_2) = \frac{23}{90} \approx 0.2556$$

Again, with some more computations, we have that

$$I(x_1; c, \Phi) = \frac{44}{45} \left[(10 - c) \log \frac{10 - c}{9} + c \log c \right],$$

and thus,

$$I(x_1; \Phi) = \inf_{c>0} I(x_1; c, \Phi) = 0,$$

where the minimum is attained at $c = 1$. Again, we also provide a plot of $I(x_1; c, \Phi)$ as a function of c in Figure 2b.

D.3.5 Relation to other BMDP separability notions

$I(\Phi) > 0$ can be considered as a separability condition, as it implies that we can correctly “separate”(cluster) all contexts. In this section, we actually show that our notion of separability encompasses previously considered separability notions i.e. ours is the strongest.

γ -separability Du et al. (2019) considered a notion of γ -separability, in which the *backward probability vectors* of two different latent states should be sufficiently separated. Precisely, for $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ with $\text{supp}(\nu) = \mathcal{S} \times \mathcal{A}$, define

$$\mathbf{b}_\nu(s') = (b_\nu(s, a|s'))_{(s,a) \in \mathcal{S} \times \mathcal{A}}, \quad b_\nu(s, a|s') := \frac{p(s'|s, a)\nu(s, a)}{\sum_{\tilde{s}, \tilde{a}} p(s'|\tilde{s}, \tilde{a})\nu(\tilde{s}, \tilde{a})}.$$

Definition 4 (Assumption 3.2 of Du et al. (2019)). For $\gamma > 0$, a BMDP is said to be γ -separable if

$$\|\mathbf{b}_\nu(s') - \mathbf{b}_\nu(s'')\|_1 \geq \gamma, \quad \forall s', s'' \in \mathcal{S}, s' \neq s'' \tag{44}$$

If no such γ exists, then we say that the BMDP is γ -inseparable.

The following proposition shows that our notion of separability ($I(\Phi) > 0$) is *weaker* (in that it encompasses a broader range of BMDPs) than γ -separability:

Proposition 15. *If the BMDP Φ is γ -separable, then $I(\Phi) > 0$.*

Proof. We prove the contrapositive. Assume that $I(\Phi) = 0$. Then by Proposition 10, there exists some $j \in \mathcal{S}$ with $j \neq f(x)$ and $c > 0$ such that

$$p(s|f(x), a) = p(s|j, a) \text{ and } p(f(x)|s, a) = cp(j|s, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

WLOG fix some $\nu \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ with $\text{supp}(\nu) = \mathcal{S} \times \mathcal{A}$. Then, for arbitrary $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} b_\nu(s, a|f(x)) &= \frac{p(f(x)|s, a)\nu(s, a)}{\sum_{\tilde{s}, \tilde{a}} p(f(x)|\tilde{s}, \tilde{a})\nu(\tilde{s}, \tilde{a})} \\ &= \frac{cp(j|s, a)\nu(s, a)}{\sum_{\tilde{s}, \tilde{a}} cp(j|\tilde{s}, \tilde{a})\nu(\tilde{s}, \tilde{a})} \\ &= \frac{p(j|s, a)\nu(s, a)}{\sum_{\tilde{s}, \tilde{a}} p(j|\tilde{s}, \tilde{a})\nu(\tilde{s}, \tilde{a})} = b_\nu(s, a|j), \end{aligned}$$

which implies that $\mathbf{b}_\nu(f(x)) = \mathbf{b}_\nu(j)$. □

Kinematic separability Misra et al. (2020) considered a notion of kinematic inseparability, in which two contexts have the same forward and backward probabilities.

We recall that $P(y|x, a) = q(y|f(y))p(f(y)|f(x), a)$ is the context transition probability kernel. Similarly as above, we define the *contextual backward transition probability kernel*: given some $u \in \mathcal{P}(\mathcal{X} \times \mathcal{A})$ with $\text{supp}(u) = \mathcal{X} \times \mathcal{A}$,

$$P^{bwd}(x, a|y) := \frac{P(y|x, a)u(x, a)}{\sum_{x', a'} P(y|x', a')u(x', a')}. \quad (45)$$

Definition 5 (Definition 3 of Misra et al. (2020)). *Given a BMDP Φ , two contexts $x_1, x_2 \in \mathcal{X}$ are **kinematically inseparable** if for every distribution $u \in \mathcal{P}(\mathcal{X} \times \mathcal{A})$ with $\text{supp}(u) = \mathcal{X} \times \mathcal{A}$ the following holds: for all $(x, a) \in \mathcal{X} \times \mathcal{A}$,*

$$(C1) \quad P(x|x_1, a) = P(x|x_2, a),$$

$$(C2) \quad P^{bwd}(x, a|x_1) = P^{bwd}(x, a|x_2).$$

*If one of these conditions fails, then we say that x_1, x_2 are **kinematically separable**.*

The above definition can be equivalently rewritten using latent transition and emission probabilities:

Lemma 7. *The above conditions for x_1, x_2 being kinematically inseparable are equivalent to the following: for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$(C1') \quad p(s|f(x_1), a) = p(s|f(x_2), a),$$

$$(C2') \quad \frac{p(f(x_1)|s, a)}{\sum_{x', a'} p(f(x_1)|f(x'), a')u(x', a')} = \frac{p(f(x_2)|s, a)}{\sum_{x', a'} p(f(x_2)|f(x'), a')u(x', a')}.$$

Proof. Follows from straightforward computations. □

Misra et al. (2020) then extended the notion of kinematic separability, which is defined between two contexts, to the whole BMDP:

Definition 6 (Definition 4 of Misra et al. (2020)). *A BMDP Φ is in **canonical form** if for any $x_1, x_2 \in \mathcal{X}$ the following holds: $f(x_1) = f(x_2)$ if and only if x_1 and x_2 are kinematically inseparable. If this does not hold, then Φ is said to be not in canonical form.*

Finally, the following proposition shows that our notion of separability ($I(\Phi) > 0$) is *weaker* (in that it encompasses a broader range of BMDPs) than BMDP being in canonical form:

Proposition 16. *If the BMDP Φ is in canonical form, then $I(\Phi) > 0$.*

Proof. Again, we prove the contrapositive. Assume that $I(\Phi) = 0$. From Proposition 10, there exists some $j \in \mathcal{S}$ with $j \neq f(x)$ and $c > 0$ such that

$$p(s|f(x), a) = p(s|j, a) \text{ and } p(f(x)|s, a) = cp(j|s, a), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

First part implies (C1'). For the second part, WLOG fix some $u \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ with $\text{supp}(u) = \mathcal{S} \times \mathcal{A}$. Then, for arbitrary $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{p(f(x)|s, a)}{\sum_{x', a'} p(f(x)|f(x'), a')u(x', a')} = \frac{cp(j|s, a)}{\sum_{x', a'} cp(j|f(x'), a')u(x', a')} = \frac{p(j|s, a)}{\sum_{x', a'} p(j|f(x'), a')u(x', a')},$$

which precisely implies (C2'). By Lemma 7, we have that Φ is not in canonical form. □

E Bernstein-type Concentration for Markov Chains with Restarts and Applications

In this appendix, we present concentration results for Markov chains *with restarts*. These results will be crucial in the analysis of the performance of our algorithms.

E.1 Concentration of Markov chains with restarts

Consider T i.i.d. episodes of BMDP of length H generated under the behavior policy. Each episode corresponds to the trajectory of length H of a Markov chain. We are interested in deriving concentration results for these trajectories for any T and H . In particular, since H can possibly be limited, we cannot use either the long-term properties of the Markov chain or the assumption that the chain starts in its steady-state regime.

Our concentration results differ from most prior Bernstein-type concentration bounds for Markov chains (refer to [Paulin \(2015\)](#) and references therein). Indeed existing bounds hold for a single trajectory and assume that the Markov chain starts from its stationary distribution. Considering restarts from an arbitrary distribution induces a bias term that vanishes with the number of observations (as H grows large), and our results account for this bias.

Theorem 9. *Let $\{(X_h^{(t)})_{h=1}^H\}_{t \in [T]}$ be a collection of i.i.d. possibly time-inhomogeneous Markov chains over a finite state space \mathcal{Z} , with transition probability matrices $\{P_h\}_{h \geq 1}$ and initial distribution $\mu \in \mathcal{P}(\mathcal{Z})$. We assume that μ and P_h 's are η -regular (see [Appendix C](#)), and that each P_h admits a stationary distribution ν_h . Let $\{\phi_h : \mathcal{X} \rightarrow \mathbb{R}\}$ be a collection of bounded measurable real-valued functions. Then we have that for all $u \geq 0$,*

$$\mathbb{P} \left[\sum_{t=1}^T \sum_{h=1}^H \phi_h(X_h^{(t)}) - \mathbb{E}_\mu[\phi_h(X_h^{(t)})] > u \right] \leq \exp \left(- \frac{u^2}{2THV_{\mu,P,\phi} + \frac{2}{3}M_{P,\phi}u} \right), \quad (46)$$

where the variance and maximum deviation terms are defined as follows:

$$V_{\mu,P,\phi} := (1 + \sqrt{2\eta}(2\eta - 1))^2 \max_{z \in \mathcal{Z}} \max_{1 \leq \ell \leq h \leq H} \text{Var}_{P_{\ell-1}(z,\cdot)}[\phi_h], \quad (47)$$

$$M_{P,\phi} := (2\eta - 1) \max_{h \in [H]} \|\phi_h\|_\infty, \quad (48)$$

and where by convention $P_0(z, \cdot) = \mu(\cdot)$.

E.2 Proof of Theorem 9

All the supporting lemmas are presented and proved in [Section E.4](#). In this proof, we use the following notations. We define $\Pi_h := \mathbf{1}\nu_h$ (where $\mathbf{1}$ is a column vector with coordinates equal to 1 and ν_h is a row vector). Denote $\mathbb{E}_\nu[\phi] := \mathbb{E}_{X \sim \nu}[\phi(X)]$ and $\mathbb{E}_\nu[\phi(X_h)] := \mathbb{E}_{X_1 \sim \nu}[\phi(X_h)]$, and similarly for Var .

Without loss of generality, we assume that $\nu_h \phi_h := \mathbb{E}_{X \sim \nu_h}[\phi_h(X)] = 0$ for all $h \in [H]$. Indeed, if $\nu_h \phi_h \neq 0$, then we can write $\phi_h(X_h^{(t)}) - \mathbb{E}_\mu[\phi_h(X_h^{(t)})] = (\phi_h(X_h^{(t)}) - \nu_h \phi_h) - \mathbb{E}_\mu[(\phi_h(X_h^{(t)}) - \nu_h \phi_h)]$, then continue the proof using $\tilde{\phi}_h : x \mapsto \phi_h(x) - \nu_h \phi_h$ instead.

We start by obtaining an upper bound on the moment generating function of $\sum_{t=1}^T \sum_{h=1}^H \phi_h(X_h^{(t)}) - \mathbb{E}_\mu[\phi_h(X_h^{(t)})]$. Using the fact that the trajectories are independent across episodes $t \in [T]$ and applying [Lemma 8](#), we immediately obtain, for all $\lambda \geq 0$,

$$\begin{aligned} & \mathbb{E}_\mu \left[\exp \left(\lambda \left(\sum_{t=1}^T \sum_{h=1}^H \phi_h(X_h^{(t)}) - \mathbb{E}_\mu[\phi_h(X_h^{(t)})] \right) \right) \right] \\ &= \prod_{t=1}^T \mathbb{E}_\mu \left[\exp \left(\lambda \left(\sum_{h=1}^H \phi_h(X_h^{(t)}) - \mathbb{E}_\mu[\phi_h(X_h^{(t)})] \right) \right) \right] \leq \exp \left(\frac{THV_{\mu,P,\phi}}{M_{P,\phi}^2} \varphi(\lambda M_{P,\phi}) \right), \end{aligned}$$

where we recall that $\varphi(x) = e^x - x - 1$, and where definitions of $V_{\mu,P,\phi}$ and $M_{P,\phi}$ are given in [Lemmas 8](#). By Markov inequality, we have: for any $u \geq 0$,

$$\mathbb{P} \left(\sum_{t=1}^T \sum_{h=1}^H \phi_h(X_h^{(t)}) - \mathbb{E}_\mu[\phi_h(X_h^{(t)})] > u \right) \leq \inf_{\lambda \geq 0} \exp \left(\frac{THV_{\mu,P,\phi}}{M_{P,\phi}^2} \varphi(\lambda M_{P,\phi}) - \lambda u \right)$$

$$\begin{aligned} &\leq \exp\left(-\frac{THV_{\mu,P,\phi}}{M_{P,\phi}^2}\varphi^*\left(\frac{M_{P,\phi}u}{THV_{\mu,P,\phi}}\right)\right) \\ &\leq \exp\left(-\frac{u^2}{2THV_{\mu,P,\phi} + \frac{2}{3}M_{P,\phi}u}\right), \end{aligned}$$

where we introduced $\varphi^*(y) = (1+y)\log(1+y) - y$, the Fenchel dual⁸ of φ . The last inequality then follows from $\varphi^*(y) \geq \frac{y^2}{2+\frac{2}{3}y}$ (see Boucheron et al. (2013) for the proof of this elementary inequality). The result then follows from the upper bounds of $V_{\mu,P,\phi}$ and $M_{P,\phi}$ derived in Lemma 9. \square

E.3 Towards concentration inequalities in BMDPs

Next, we specify the results of Theorem 9 to the case of homogenous Markov chains. This will be instrumental in our analysis since indeed the Markov chains induced in BMDPs are homogenous. The results resemble the concentration results established by Sanders et al. (2020) (see Proposition 10 of their supplementary material SM1) and the subsequent improvements established by Sanders and Senen–Cerda (2023), but there are several key differences. One is that we keep track of the asymptotics in S and A , and another is that we consider restarts and have to deal with the absence of equilibrium assumption.

The next theorem is a direct application of Theorem 9 to homogenous Markov chains (e.g., BMDPs).

Theorem 10. *Let $\{(X_h^{(t)})_{h=1}^H\}_{t \in [T]}$ be a collection of i.i.d. time-homogeneous Markov chains over a finite state space \mathcal{Z} , with transition probability matrix P and initial distribution $\mu \in \mathcal{P}(\mathcal{Z})$. We assume that μ and P are η -regular, and that P admits a stationary distribution ν . Let $\phi : \mathcal{X} \rightarrow \mathbb{R}$ be a bounded measurable real-valued function. Then we have that for all $u \geq 0$,*

$$\mathbb{P}\left[\sum_{t=1}^T \sum_{h=1}^H \phi(X_h^{(t)}) - \mathbb{E}_\mu[\phi(X_h^{(t)})] > u\right] \leq \exp\left(-\frac{u^2}{2THV_{\mu,P,\phi} + \frac{2}{3}M_{P,\phi}u}\right), \quad (49)$$

where the variance and maximum deviation terms are defined as follows:

$$V_{\mu,P,\phi} := (1 + \sqrt{2}\eta(2\eta - 1))^2 \max\left(\text{Var}_\mu[\phi], \max_{z \in \mathcal{Z}} \text{Var}_{P(z,\cdot)}[\phi]\right), \quad (50)$$

$$M_{P,\phi} := (2\eta - 1)\|\phi\|_\infty. \quad (51)$$

We can further simplify the bound depending on the choice of ρ :

- For any u satisfying $\|\phi\|_\infty u = o(THV_{\mu,P,\phi})$,

$$\mathbb{P}\left(\left|\sum_{t=1}^T \sum_{h=1}^H \phi(X_h^{(t)}) - \mathbb{E}_\mu[\phi(X_h^{(t)})]\right| > u\right) \leq 2 \exp\left(-\frac{u^2}{2THV_{\mu,P,\phi}}\right)$$

- For any u satisfying $\|\phi\|_\infty u = \omega(THV_{\mu,P,\phi})$,

$$\mathbb{P}\left(\left|\sum_{t=1}^T \sum_{h=1}^H \phi(X_h^{(t)}) - \mathbb{E}_\mu[\phi(X_h^{(t)})]\right| > u\right) \leq 2 \exp\left(-\frac{u}{\frac{2}{3}M_{P,\phi}}\right)$$

E.4 Supporting lemmas for Theorem 9

Let $(X_h)_{h=1}^H$ be a (fixed) Markov chain over a finite state space \mathcal{Z} with transition probability matrices $\{P_h\}_{h \geq 1}$. We further assume that each P_h admits a stationary distribution, denoted by ν_h , and let $\Pi_h := \mathbf{1}\nu_h$. We denote by $P_0 := \mu \in \mathcal{P}(\mathcal{Z})$ its initial distribution ($X_1 \sim \mu$). We assume that μ and P_h 's are η -regular. Let $\{\phi_h : \mathcal{X} \rightarrow \mathbb{R}\}$ be a collection of bounded measurable real-valued functions. Lastly, we denote $\mathbb{E}_\nu[\phi] := \mathbb{E}_{X \sim \nu}[\phi(X)]$ and $\mathbb{E}_\nu[\phi(X_h)] = \mathbb{E}_{X_1 \sim \nu}[\phi(X_h)]$, and similarly for Var , and denote $\|\phi\|_\nu := \sqrt{\mathbb{E}_\nu[\phi(X)^2]}$. Furthermore, we regard any probability measures over \mathcal{Z} as a $|\mathcal{Z}|$ -dimensional row vector and all real-valued measurable functions over \mathcal{Z} as a $|\mathcal{Z}|$ -dimensional column vector⁹.

⁸Recall its definition: $\varphi^*(y) := \sup_x yx - \varphi(x)$.

⁹With this notation, note that for $\phi : \mathcal{Z} \rightarrow \mathbb{R}$, $\mathbb{E}_\mu[\phi] = \mu\phi$.

Lemma 8. For all $\lambda > 0$, we have

$$\mathbb{E}_\mu \left[\exp \left(\lambda \left(\sum_{h=1}^H \phi_h(X_h) - \mathbb{E}_\mu[\phi_h(X_h)] \right) \right) \right] \leq \exp \left(\frac{HV_{\mu,P,\phi}}{M_{P,\phi}^2} \varphi(\lambda M_{P,\phi}) \right),$$

where $\varphi : x \mapsto e^x - x - 1$. The variance term and maximum deviation term are defined, respectively, as follows:

$$V_{\mu,P,\phi} := \max_{z \in \mathcal{Z}, \ell \in [H]} \text{Var}_{P_{\ell-1}(z, \cdot)} \left[\sum_{h=\ell}^H \left(\prod_{i=\ell}^{h-1} (P_i - \Pi_i) \right) \phi_h \right], \quad (52)$$

$$M_{P,\phi} := \max_{\ell \in [H]} \left\| \sum_{h=\ell}^H \left(\prod_{i=\ell}^{h-1} (P_i - \Pi_i) \right) \phi_h \right\|_\infty. \quad (53)$$

Proof. For notational convenience, let us introduce, for all $h \geq 1$, the $|\mathcal{Z}|$ -dimensional random row vector $\delta_h = (\mathbb{1}\{X_h = z\})_{z \in \mathcal{Z}}$. Then observe that we may write, via a telescoping sum, for all $h \geq 1$,

$$\begin{aligned} & \phi_h(X_h) - \mathbb{E}_\mu[\phi_h(X_h)] \\ &= \delta_1 \left(\prod_{i=1}^{h-1} P_i \right) \phi_h - \mu \left(\prod_{i=1}^{h-1} P_i \right) \phi_h + \sum_{\ell=1}^{h-1} \delta_{\ell+1} \left(\prod_{i=\ell+1}^{h-1} P_i \right) \phi_h - \delta_\ell \left(\prod_{i=\ell}^{h-1} P_i \right) \phi_h \\ &= (\delta_1 - \mu) \left(\prod_{i=1}^{h-1} P_i \right) \phi_h + \sum_{\ell=1}^{h-1} (\delta_{\ell+1} - \delta_\ell P_\ell) \left(\prod_{i=\ell+1}^{h-1} P_i \right) \phi_h. \end{aligned}$$

To further ease notations, we introduce $Z_1 = \delta_1 - \mu$ and $Z_\ell = \delta_\ell - \delta_{\ell-1} P_{\ell-1}$ for $\ell \geq 2$, and $P_{\ell \rightarrow h} = \prod_{i=\ell}^{h-1} P_i$, with the convention that $P_{\ell \rightarrow \ell} = I$. With that, we may write $\phi_h(X_h) - \mathbb{E}_\mu[\phi_h(X_h)] = \sum_{\ell=1}^h Z_\ell P_{\ell \rightarrow h} \phi_h$. One important observation is that $Z_\ell \mathbf{1} = 0$.

On the other hand, from the fact that $(P_i - \Pi_i) \Pi_j = 0$ for all $i, j \geq 1$, we have that

$$P_{\ell \rightarrow h} = \prod_{i=\ell}^{h-1} P_i = \prod_{i=\ell}^{h-1} (P_i - \Pi_i) + \Pi_\ell \prod_{j=\ell+1}^{h-1} P_j = \prod_{i=\ell}^{h-1} (P_i - \Pi_i) + \Pi_\ell P_{\ell+1 \rightarrow h}. \quad (54)$$

Since $\Pi_j = \mathbf{1}\nu_j$, we have that $\Pi_\ell P_{\ell+1 \rightarrow h} = \mathbf{1}\xi$ for some $\xi \in \mathbb{R}^{1 \times |\mathcal{Z}|}$, i.e.,

$$Z_\ell (\Pi_\ell P_{\ell+1 \rightarrow h}) \phi_h = Z_\ell \mathbf{1} \xi \phi_h = 0, \quad \forall h > \ell \geq 1.$$

Thus, introducing $\Delta_{\ell \rightarrow h} = \prod_{i=\ell}^{h-1} (P_i - \Pi_i)$, with the convention that $\Delta_{\ell \rightarrow \ell} = I$, we may finally write

$$\sum_{h=1}^H \phi_h(X_h) - \mathbb{E}_\mu[\phi_h(X_h)] = \sum_{h=1}^H \sum_{\ell=1}^h Z_\ell \Delta_{\ell \rightarrow h} \phi_h = \sum_{\ell=1}^H Z_\ell \left(\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} \phi_h \right). \quad (55)$$

Now, we are ready to upper bound the moment-generating function of the LHS. The convenience of Eqn. (55) is that $(Z_h)_{h \geq 1}$ is adapted to the filtration generated by the Markov chain $(X_h)_{h \geq 1}$, $\mathbb{E}_\mu[Z_h | X_{h-1}] = 0$ for all $h \geq 2$, and $\mathbb{E}_\mu[Z_1] = 0$. Using a standard Bernstein type upper bound technique (Boucheron et al., 2013), we have for all $\ell \geq 2$, $\lambda > 0$,

$$\mathbb{E}_\mu \left[\exp \left(\lambda Z_\ell \left(\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} \phi_h \right) \right) \middle| X_{\ell-1} \right] \leq \exp \left(\frac{\text{Var}_{P_{\ell-1}(X_{\ell-1}, \cdot)} \left[\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} \phi_h \right]}{C^2} \varphi(\lambda C) \right),$$

where $\varphi : x \mapsto e^x - x - 1$, and C can be any positive constant that verifies $\| \sum_{h=\ell}^H \Delta_{\ell \rightarrow h} f_h \|_\infty \leq C$. Introducing $V_{\mu,P,\phi}$ and $M_{P,\phi}$ as defined in the statement of this Lemma, we see that for all $\ell \geq 2$,

$$\mathbb{E}_\mu \left[\exp \left(\lambda Z_\ell \left(\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} \phi_h \right) \right) \middle| X_{\ell-1} \right] \leq \exp \left(\frac{V_{\mu,P,\phi}}{M_{\mu,P,\phi}^2} \varphi(\lambda M_{\mu,P,\phi}) \right).$$

Similarly, we also have

$$\mathbb{E}_\mu \left[\exp \left(\lambda Z_1 \left(\sum_{h=1}^H \Delta_{\ell \rightarrow h} \phi_h \right) \right) \right] \leq \exp \left(\frac{V_{\mu, P, \phi}}{M_{\mu, P, \phi}^2} \varphi(\lambda M_{\mu, P, \phi}) \right).$$

Now using the fact that $(Z_h)_{h \geq 1}$ is adapted to the filtration generated by the Markov chain $(X_h)_{h \geq 1}$, we obtain via a peeling argument that

$$\mathbb{E}_\mu \left[\exp \left(\lambda \left(\sum_{h=1}^H \phi_h(X_h) - \mathbb{E}_\Phi[\phi_h(X_h)] \right) \right) \right] \leq \exp \left(\frac{H V_{\mu, P, \phi}}{M_{\mu, P, \phi}^2} \varphi(\lambda M_{\mu, P, \phi}) \right), \quad \forall \lambda > 0.$$

□

The following lemma provides simple bounds for $V_{\mu, P, \phi}$ and $M_{P, \phi}$.

Lemma 9. For all $\ell \in [H]$ and all $z \in \mathcal{Z}$,

$$\text{Var}_{P_{\ell-1}(z, \cdot)} \left[\sum_{h=\ell}^H \left(\prod_{i=\ell}^{h-1} (P_i - \Pi_i) \right) \phi_h \right] \leq (1 + \sqrt{2\eta}(2\eta - 1))^2 \max_{\ell \leq h \leq H} \text{Var}_{P_{\ell-1}(z, \cdot)}[\phi_h] \quad (56)$$

and

$$\left\| \sum_{h=\ell}^H \left(\prod_{i=\ell}^{h-1} (P_i - \Pi_i) \right) \phi_h \right\|_\infty \leq (2\eta - 1) \max_{\ell \leq h \leq H} \|\phi_h\|_\infty. \quad (57)$$

Proof. Recall that $\Delta_{\ell \rightarrow h} = \prod_{i=\ell}^{h-1} (P_i - \Pi_i)$ for $h > \ell \geq 1$, with the convention that $\Delta_{\ell \rightarrow \ell} = I$. Finally, remember that $P_0(z, \cdot) := \mu(\cdot)$.

Proof of Eqn. (56) (Variance term)

$$\begin{aligned} \text{Var}_\mu \left[\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} \phi_h \right] &= \text{Var}_\mu \left[\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} (\phi_h - \mu \phi_h \mathbf{1}) \right] \\ &\leq \mathbb{E}_{X \sim \mu} \left[\left(\left(\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} (\phi_h - \mu \phi_h \mathbf{1}) \right) (X) \right)^2 \right] \\ &= \left\| \sum_{h=\ell}^H \Delta_{\ell \rightarrow h} (\phi_h - \mu \phi_h \mathbf{1}) \right\|_\mu^2 \\ &\leq \left(\sum_{h=\ell}^H \|\Delta_{\ell \rightarrow h} (\phi_h - \mu \phi_h \mathbf{1})\|_\mu \right)^2 \end{aligned}$$

Now, we note that for $h > \ell \geq 1$, we have:

$$\begin{aligned} \|\Delta_{\ell \rightarrow h} (\phi_h - \mu \phi_h \mathbf{1})\|_\mu^2 &\leq \|\Delta_{\ell \rightarrow h} (\phi_h - \mu \phi_h \mathbf{1})\|_\infty^2 \\ &= \max_{x \in \mathcal{Z}} \left| \sum_y \Delta_{\ell \rightarrow h}(x, y) (\phi_h - \mu \phi_h \mathbf{1})(y) \right|^2 \\ &\leq \max_{x \in \mathcal{Z}} \left| \sum_y \frac{\Delta_{\ell \rightarrow h}(x, y)^2}{\mu(y)} \right| \|\phi_h - \mu \phi_h \mathbf{1}\|_\mu^2 \\ &\leq \left(\max_{x, y \in \mathcal{Z}} \left| \frac{\Delta_{\ell \rightarrow h}(x, y)}{\mu(y)} \right| \right) \left(\max_{x \in \mathcal{Z}} \sum_{y \in \mathcal{Z}} |\Delta_{\ell \rightarrow h}(x, y)| \right) \|\phi_h - \mu \phi_h \mathbf{1}\|_\mu^2 \end{aligned}$$

$$\begin{aligned}
 &\leq \left(\max_{x,y \in \mathcal{Z}} \left| \frac{\Delta_{\ell \rightarrow h}(x,y)}{\mu(y)} \right| \right) \left(\max_{g \in \mathbb{R}^{|\mathcal{Z}|}: \|g\|_\infty \leq 1} \|\Delta_{\ell \rightarrow h} g\|_\infty \right) \|\phi_h - \mu \phi_h \mathbf{1}\|_\mu^2 \\
 &\leq 2 \left(\max_{x,y \in \mathcal{Z}} \left| \frac{\Delta_{\ell \rightarrow h}(x,y)}{\mu(y)} \right| \right) \left(1 - \frac{1}{\eta} \right)^{h-\ell} \|\phi_h - \mu \phi_h \mathbf{1}\|_\mu^2,
 \end{aligned}$$

where the last inequality follows from Lemma 10.

As μ and $(P_h)_{h \geq 1}$ satisfy the η -regularity property, so does $P_{\ell \rightarrow h}(x, \cdot)$ and $\Pi_\ell P_{\ell+1 \rightarrow h}$ with the same parameter η , for all $x \in \mathcal{Z}$ (this can be easily proven using induction on the number of multiplicands $h - \ell$). Thus, from Eqn. (54), $\Delta_{\ell \rightarrow h}(x, \cdot) = P_{\ell \rightarrow h}(x, \cdot) - \Pi_\ell P_{\ell+1 \rightarrow h}$ is also η -regular, which implies that

$$\max_{x,y \in \mathcal{Z}} \left| \frac{\Delta_{\ell \rightarrow h}(x,y)}{\mu(y)} \right| \leq \eta |\mathcal{Z}| \left(\frac{\eta}{|\mathcal{Z}|} - \frac{1}{\eta |\mathcal{Z}|} \right) = \eta^2 - 1,$$

which in turn gives

$$\|\Delta_{\ell \rightarrow h}(\phi_h - \mu \phi_h \mathbf{1})\|_\mu^2 \leq 2(\eta^2 - 1) \left(1 - \frac{1}{\eta} \right)^{h-\ell} \|\phi_h - \mu \phi_h \mathbf{1}\|_\mu^2.$$

Noting that $\|\phi_h - \mu \phi_h \mathbf{1}\|_\mu^2 = \text{Var}_\mu[\phi_h]$, we finally obtain

$$\begin{aligned}
 \text{Var}_\mu \left[\sum_{h=\ell}^H \Delta_{\ell \rightarrow h} \phi_h \right] &\leq \left(\sum_{h=\ell}^H \|\Delta_{\ell \rightarrow h}(\phi_h - \mu \phi_h \mathbf{1})\|_\mu \right)^2 \\
 &\leq \left(\sqrt{\text{Var}_\mu[\phi_\ell]} + \sum_{h=\ell+1}^H \sqrt{2(\eta^2 - 1) \left(1 - \frac{1}{\eta} \right)^{h-\ell}} \sqrt{\text{Var}_\mu[\phi_h]} \right)^2 \\
 &\leq \max_{\ell \leq h \leq H} \text{Var}_\mu[\phi_h] \left(1 + \sqrt{2(\eta^2 - 1)} \sum_{h=\ell+1}^H \left(1 - \frac{1}{\eta} \right)^{(h-\ell)/2} \right)^2 \\
 &\leq (1 + \sqrt{2}\eta(2\eta - 1))^2 \max_{\ell \leq h \leq H} \text{Var}_\mu[\phi_h].
 \end{aligned}$$

Proof of Eqn. (57) (Maximum deviation term)

$$\begin{aligned}
 \left\| \sum_{h=\ell}^H \Delta_{\ell \rightarrow h} \phi_h \right\|_\infty &\leq \sum_{h=\ell}^H \|\Delta_{\ell \rightarrow h} \phi_h\|_\infty \\
 &= \|\phi_\ell\|_\infty + \sum_{h=\ell+1}^H \|\Delta_{\ell \rightarrow h} \phi_h\|_\infty \\
 &\leq \|\phi_\ell\|_\infty + 2 \sum_{h=\ell+1}^H \left(1 - \frac{1}{\eta} \right)^{h-\ell} \|\phi_h\|_\infty \\
 &\leq (1 + 2(\eta - 1)) \max_{h \in [H]} \|\phi_h\|_\infty \\
 &= (2\eta - 1) \max_{h \in [H]} \|\phi_h\|_\infty,
 \end{aligned}$$

where we used the triangle inequality in the first inequality and Lemma 10 in the third inequality. \square

The following lemma is used above in the proof of Lemma 9.

Lemma 10. *Let $(P_h)_{h=1}^{H-1}$ be transition matrices (i.e. row-stochastic) over \mathcal{Z} , with each P_h having a stationary distribution ν_h . Let $g \in \mathbb{R}^{\mathcal{Z}}$ be such that $\|g\|_\infty < \infty$ and let us denote $\Pi_h := \mathbf{1}\nu_h$. Then*

$$\left\| \prod_{h=1}^{H-1} (P_h - \Pi_h) g \right\|_\infty \leq 2 \left(\prod_{h=1}^{H-1} \delta(P_h) \right) \|g\|_\infty,$$

where $\delta(P_h) = \max_{x,y \in \mathcal{Z}} d_{TV}(P_h(x, \cdot), P_h(y, \cdot))$ is the Dobrushin's coefficient of P_h . In particular, if we assume that P_h 's are η -regular, it follows that

$$\left\| \prod_{h=1}^{H-1} (P_i - \Pi_i) g \right\|_{\infty} \leq 2 \left(1 - \frac{1}{\eta}\right)^{H-1} \|g\|_{\infty}.$$

Proof. First, we observe that

$$\begin{aligned} \left\| \prod_{h=1}^{H-1} (P_i - \Pi_i) f \right\|_{\infty} &\stackrel{(a)}{=} \left\| \left(\prod_{h=1}^{H-1} P_i - \Pi_1 \prod_{i=2}^{H-1} P_i \right) g \right\|_{\infty} \\ &= \max_{x \in \mathcal{Z}} \left| \sum_{y \in \mathcal{Z}} \left(\prod_{h=1}^{H-1} P_i \right) (x, y) g(y) - \sum_{y \in \mathcal{Z}} \left(\Pi_1 \prod_{h=2}^{H-1} P_i \right) (x, y) g(y) \right| \\ &\leq 2 \max_{x \in \mathcal{Z}} d_{TV} \left(\left(\prod_{h=1}^{H-1} P_i \right) (x, \cdot), \left(\Pi_1 \prod_{h=2}^{H-1} P_i \right) (x, \cdot) \right) \|g\|_{\infty} \end{aligned}$$

where (a) follows from the fact that $(P_i - \Pi_i)\Pi_j = 0$ for all $i, j \geq 1$. Recall that $P_{\ell \rightarrow h} = \prod_{i=\ell}^{h-1} P_i$ $h \geq \ell \geq 1$, with the convention that $P_{\ell \rightarrow \ell} = I$. Then we may simply write

$$\left\| \prod_{h=1}^{H-1} (P_i - \Pi_i) g \right\|_{\infty} \leq 2 \max_{x \in \mathcal{Z}} d_{TV} (P_{1 \rightarrow H}(x, \cdot), \Pi_1 P_{2 \rightarrow H}(x, \cdot)) \|g\|_{\infty}. \quad (58)$$

Next, we formulate two claims to complete the proof:

Claim 1.

$$\max_{x \in \mathcal{Z}} d_{TV} (P_{1 \rightarrow H}(x, \cdot), \Pi_1 P_{2 \rightarrow H}(x, \cdot)) \leq \delta(P_{1 \rightarrow H}). \quad (59)$$

Claim 2.

$$\delta(P_{\ell \rightarrow h}) \leq \delta(P_{\ell \rightarrow k}) \delta(P_{k \rightarrow h}), \quad \forall h \geq k \geq \ell. \quad (60)$$

Here, as stated in the Lemma, $\delta(P) := \max_{x,y \in \mathcal{Z}} d_{TV}(P(x, \cdot), P(y, \cdot))$ is the Dobrushin's coefficient.

Assuming that the two above claims hold, from (58), we have:

$$\begin{aligned} \left\| \prod_{h=1}^{H-1} (P_i - \Pi_i) g \right\|_{\infty} &\leq 2 \max_{x \in \mathcal{Z}} d_{TV} (P_{1 \rightarrow H}(x, \cdot), \Pi_1 P_{2 \rightarrow H}(x, \cdot)) \|g\|_{\infty} \\ &\leq 2 \delta(P_{1 \rightarrow H}) \|f\|_{\infty} \\ &\leq 2 (\delta(P_{1 \rightarrow 2}) \times \delta(P_{2 \rightarrow 3}) \times \dots \times \delta(P_{H-1 \rightarrow H})) \|g\|_{\infty} = 2 \left(\prod_{h=1}^{H-1} \delta(P_h) \right) \|g\|_{\infty}, \end{aligned}$$

where we noted that $P_{h \rightarrow h+1} = P_h$.

Finally, we provide proof of the above two claims. Both are simple adaptations of the proofs of Lemma 4.10 and Lemma 4.11 of [Levin and Peres \(2017\)](#), respectively. We provide these proofs for completeness.

Proof of Claim 1

By definition,

$$\begin{aligned} \max_{x \in \mathcal{Z}} d_{TV} (P_{1 \rightarrow H}(x, \cdot), \Pi_1 P_{2 \rightarrow H}(x, \cdot)) &= \max_{x \in \mathcal{Z}} \max_{A \subseteq \mathcal{Z}} |P_{1 \rightarrow H}(x, A) - \Pi_1 P_{2 \rightarrow H}(x, A)| \\ &= \max_{x \in \mathcal{Z}} \max_{A \subseteq \mathcal{Z}} |P_{1 \rightarrow H}(x, A) - \Pi_1 P_{1 \rightarrow H}(x, A)| \quad (\Pi_1 P_1 = \Pi_1) \end{aligned}$$

$$\begin{aligned}
 &= \max_{x \in \mathcal{Z}} \max_{A \subseteq \mathcal{Z}} \left| P_{1 \rightarrow H}(x, A) - \sum_{y \in \mathcal{Z}} \nu_1(y) P_{1 \rightarrow H}(y, A) \right| \\
 &= \max_{x \in \mathcal{Z}} \max_{A \subseteq \mathcal{Z}} \left| \sum_{y \in \mathcal{Z}} \nu_1(y) (P_{1 \rightarrow H}(x, A) - P_{1 \rightarrow H}(y, A)) \right| \\
 &\leq \max_{x \in \mathcal{Z}} \sum_{y \in \mathcal{Z}} \nu_1(y) \max_{A \subseteq \mathcal{Z}} |(P_{1 \rightarrow H}(x, A) - P_{1 \rightarrow H}(y, A))| \\
 &= \max_{x \in \mathcal{Z}} \sum_{y \in \mathcal{Z}} \nu_1(y) d_{TV}(P_{1 \rightarrow H}(x, \cdot), P_{1 \rightarrow H}(y, \cdot)) \\
 &\leq \max_{x, y} d_{TV}(P_{1 \rightarrow H}(x, \cdot), P_{1 \rightarrow H}(y, \cdot)).
 \end{aligned}$$

Proof of Claim 2

Let $h \geq k \geq \ell \geq 1$ and fix $x, y \in \mathcal{Z}$. From Proposition 4.7 of [Levin and Peres \(2017\)](#), there exists a coupling (X_k, Y_k) of $P_{\ell \rightarrow k}(x, \cdot)$ and $P_{\ell \rightarrow k}(y, \cdot)$ such that

$$d_{TV}(P_{\ell \rightarrow k}(x, \cdot), P_{\ell \rightarrow k}(y, \cdot)) = \mathbb{P}(X_k \neq Y_k).$$

We also have $P_{\ell \rightarrow h}(x, w) = \sum_{z \in \mathcal{Z}} P_{\ell \rightarrow k}(x, z) P_{k \rightarrow h}(z, w) = \mathbb{E}[P_{k \rightarrow h}(X_k, w)]$ and similarly $P_{\ell \rightarrow h}(y, w) = \mathbb{E}[P_{k \rightarrow h}(Y_k, w)]$, for all $w \in \mathcal{Z}$. Thus, for any $A \subseteq \mathcal{Z}$, we have

$$\begin{aligned}
 |P_{\ell \rightarrow h}(x, A) - P_{\ell \rightarrow h}(y, A)| &= |\mathbb{E}[P_{k \rightarrow h}(X_k, A)] - \mathbb{E}[P_{k \rightarrow h}(Y_k, A)]| \\
 &= |\mathbb{E}[\mathbb{1}\{X_k \neq Y_k\}(P_{k \rightarrow h}(X_k, A) - P_{k \rightarrow h}(Y_k, A))]| \\
 &\leq \mathbb{E}[\mathbb{1}\{X_k \neq Y_k\} |P_{k \rightarrow h}(X_k, A) - P_{k \rightarrow h}(Y_k, A)|] \\
 &\leq \mathbb{E}[\mathbb{1}\{X_k \neq Y_k\} d_{TV}(P_{k \rightarrow h}(X_k, \cdot), P_{k \rightarrow h}(Y_k, \cdot))] \\
 &\leq \mathbb{P}(X_k \neq Y_k) \max_{x', y' \in \mathcal{Z}} d_{TV}(P_{k \rightarrow h}(x', \cdot), P_{k \rightarrow h}(y', \cdot)) \\
 &= d_{TV}(P_{\ell \rightarrow k}(x, \cdot), P_{\ell \rightarrow k}(y, \cdot)) \max_{x', y' \in \mathcal{Z}} d_{TV}(P_{k \rightarrow h}(x', \cdot), P_{k \rightarrow h}(y', \cdot))
 \end{aligned}$$

Finally, maximizing both sides over $A \subseteq \mathcal{Z}$ and $x, y \in \mathcal{Z}$ yields the desired result. \square

F Proof of Theorem 2 – Initial Spectral Clustering

In this appendix, we present the proof of Theorem 2. We further clarify the details of our initial spectral clustering algorithm and provide complementing proofs and comments.

F.1 Algorithm, preliminaries, and notations

Notations. We introduce notations used extensively throughout this appendix. We recall that $\hat{N}_a(x, y) = \sum_{t,h} \hat{N}_{a,t,h}(x, y)$ for all $x, y \in \mathcal{X}$, $a \in \mathcal{A}$, where we denote $\hat{N}_{a,t,h}(x, y) \triangleq \mathbb{1}\{(x_h^{(t)}, a_h^{(t)}, x_{h+1}^{(t)}) = (x, a, y)\}$ for all $t \in [T]$, $h \in [H]$, and use the short hand $\sum_{t,h} = \sum_{t=1}^T \sum_{h=1}^H$. Furthermore, when writing \hat{N}_a and $\hat{N}_{a,t,h}$, we will think of these as matrices in $\mathbb{R}^{n \times n}$. Finally, we will also define for all $a \in \mathcal{A}$, $P_a \triangleq (P(y|x, a))_{x,y \in \mathcal{X}}$.

Trimming. In the trimming step, for each $a \in \mathcal{A}$, we recall that Γ_a is defined as a subset of contexts constructed from \mathcal{X} , by removing γ contexts with the highest number of visits. More precisely,

$$\forall a \in \mathcal{A}, \quad \Gamma_a = \left\{ x \in \mathcal{X} : \left| \left\{ z \in \mathcal{X} : \sum_y \hat{N}_a(x, y) < \sum_y \hat{N}_a(z, y) \right\} \right| \geq \gamma \right\}$$

where we choose $\gamma = \lfloor n \exp\left(-\frac{TH}{nA} \log\left(\frac{TH}{nA}\right)\right) \rfloor$. The reason for this choice will be apparent in our analysis later on (see proof of Proposition 19). Then, the trimmed matrix is simply defined as $N_{a,\Gamma_a} = (\hat{N}_a(x, y) \mathbb{1}\{x, y \in \Gamma_a\})_{x,y \in \mathcal{X}}$.

Two-step conditioning. We further define for each $a \in \mathcal{A}$, $\tilde{N}_a \triangleq \sum_{t,h} \tilde{N}_{a,t,h}$, where we set, for $t \in [T]$, $3 \leq h \leq H$, $\tilde{N}_{a,t,h} \triangleq \mathbb{E}_\mu[\hat{N}_{a,t,h} | \hat{N}_{a,t,h-2}]$, $\tilde{N}_{a,t,2} \triangleq \mathbb{E}_\mu[\hat{N}_{a,t,2}]$, and $\tilde{N}_{a,t,1} \triangleq \mathbb{E}_\mu[\hat{N}_{a,t,1}]$. In fact, we note that

$$\begin{aligned} \forall h \geq 3, \quad \tilde{N}_{a,t,h} &= \text{diag}(P_0(x_{t,h-1}, \cdot)) \text{diag}(\rho(a|\cdot)) P_a, \\ \tilde{N}_{a,t,2} &= \text{diag}(\mu P_0) \text{diag}(\rho(a|\cdot)) P_a \\ \tilde{N}_{a,t,1} &= \text{diag}(\mu) \text{diag}(\rho(a|\cdot)) P_a \end{aligned}$$

where recall that $P_0 \triangleq (\sum_{a \in \mathcal{A}} P(y|x, a) \rho(a|x))_{x,y \in \mathcal{X}}$. Introducing $(\tilde{N}_a)_{a \in \mathcal{A}}$ is crucial in the analysis of the concentration of the trimmed matrix.

S-rank approximations. As described in our algorithm, for each $a \in \mathcal{A}$, we build a matrix \hat{M}_a that is an S -rank approximation of \hat{N}_{a,Γ_a} . More precisely, the procedure is as follows: (i) for each $a \in \mathcal{A}$, via an SVD decomposition, we can write $\hat{N}_{a,\Gamma_a} = U_a^\top \text{diag}(\sigma_{a,1}, \dots, \sigma_{a,n}) V_a^\top$ where U_a, V_a are two $n \times n$ orthonormal matrices, and $\sigma_{a,1} \geq \sigma_{a,2} \geq \dots \geq \sigma_{a,n}$ are the singular values of \hat{N}_{a,Γ_a} ; (ii) we obtain an S -rank approximation of \hat{N}_{a,Γ_a} by setting all but the first S singular values to zero, that is $\hat{M}_a = U_a^\top \text{diag}(\sigma_{a,1}, \dots, \sigma_{a,S}, 0, \dots, 0) V_a$.

Aggregation of information across actions. In order to fully exploit all the observations we gather, we aggregate the information across different actions by stacking together the obtained S -rank approximation matrices $(\hat{M}_a)_{a \in \mathcal{A}}$ to form a fat matrix \hat{M} of size $n \times 2nA$. More precisely, we write

$$\hat{M} = [(\hat{M}_1)^\top \quad \dots \quad (\hat{M}_A)^\top \quad \hat{M}_1 \quad \dots \quad \hat{M}_A].$$

Importantly, as shown later in this appendix, our random matrix \hat{M} will concentrate around the matrix \tilde{N} , an $n \times 2nA$ matrix, defined as follows

$$\tilde{N} \triangleq [(\tilde{N}_1)^\top \quad \dots \quad (\tilde{N}_A)^\top \quad \tilde{N}_1 \quad \dots \quad \tilde{N}_A].$$

As we shall see, the motivation behind our aggregation procedure stems from the fact that analyzing \tilde{N} gives rise to a separability quantity that is tightly related to the rate function $I(\Phi)$ that appears in our lower bound.

Weighted K-medians clustering. Finally, we run a weighted K-medians clustering on the rows of \hat{M} . The procedure consists of the following two steps:

- (i) we re-weigh or normalize in an ℓ_1 sense the rows of \hat{M} , by setting for all $x \in \mathcal{X}$ such that $\hat{M}(x, \cdot) \neq 0$, $\tilde{M}(x, \cdot) = \hat{M}(x, \cdot) / \|\hat{M}(x, \cdot)\|_1$ and define $\mathcal{X}_0 = \{x \in \mathcal{X} : \hat{M}(x, \cdot) = 0\}$. We further define for any $x \in \mathcal{X}$, $\tilde{R}(x, \cdot) = \tilde{N}(x, \cdot) / \|\tilde{N}(x, \cdot)\|_1$, which will be useful in the analysis.
- (ii) for some $\epsilon > 0$, we solve the following $(1 + \epsilon)$ -K-medians optimization problem on $\mathcal{X} \setminus \mathcal{X}_0$: find $\{\hat{f}(x)\}_{x \in \mathcal{X} \setminus \mathcal{X}_0}$ in $\mathcal{S}^{\mathcal{X} \setminus \mathcal{X}_0}$ such that:

$$\begin{aligned} & \sum_{s \in \mathcal{S}} \min_{u_s \in \mathbb{R}^n} \sum_{x \in \mathcal{X} \setminus \mathcal{X}_0 : \hat{f}(x) = s} \|\hat{M}(x, \cdot)\|_1 \|\tilde{M}(x, \cdot) - u_s\|_1 \leq \\ \text{Spectral} \quad & (1 + \epsilon) \min_{f \in \mathcal{S}^{\mathcal{X}}} \sum_{s \in \mathcal{S}} \min_{u_s \in \mathbb{R}^n} \sum_{x \in \mathcal{X} \setminus \mathcal{X}_0 : f(x) = s} \|\hat{M}(x, \cdot)\|_1 \|\tilde{M}(x, \cdot) - u_s\|_1 \end{aligned} \quad (61)$$

and for all $x \in \mathcal{X}_0$, we set $\hat{f}(x) = 0$.

We note that step (ii) can be executed efficiently (see e.g. [Chen et al. \(2018\)](#); [Gao et al. \(2018\)](#)).

F.2 Proof of Theorem 2

Here we state a more precise version of Theorem 2 and provide its proof.

Theorem 11. *Assume that $TH = \Omega(n)$ and $I(\Phi) > 0$. The clustering error rate of the initial spectral clustering algorithm satisfies:*

$$\mathbb{P} \left(\frac{|\mathcal{E}|}{n} \leq \text{poly}(\eta) \left(1 + \frac{(2 + \epsilon)}{\tilde{J}(\Phi, \rho)} \right) \sqrt{\frac{SAn}{TH}} \right) \geq 1 - \frac{2}{n} - 2e^{-n} - 2e^{-\frac{TH}{2nA}}.$$

Observe that when $TH = \omega(n)$, then $1 - \frac{2}{n} - 2e^{-n} - 2e^{-\frac{TH}{2nA}} \xrightarrow{n \rightarrow \infty} 1$, which justifies the claim of Theorem 2 that $\frac{|\mathcal{E}|}{n} = \mathcal{O} \left(\sqrt{\frac{SAn}{TH}} \right)$ with high probability.

Proof of Theorem 11. We know from Proposition 18 that the weighted K-medians clustering algorithm ensures that

$$\frac{|\mathcal{E}|}{n} \leq \left(2 + \eta^2 + \frac{4(2 + \epsilon)\eta^2}{\tilde{J}(\Phi, \rho)} \right) \frac{2\eta^5 n \sqrt{A}}{TH} \|\hat{M} - \hat{N}\|_F.$$

Next, by construction of \hat{M} , using Lemma 12, we have

$$\frac{|\mathcal{E}|}{n} \leq \left(2 + \eta^2 + \frac{4(2 + \epsilon)\eta^2}{\tilde{J}(\Phi, \rho)} \right) \frac{8\eta^5 n A \sqrt{S}}{TH} \max_{a \in \mathcal{A}} \|\hat{N}_{a, \Gamma_a} - \tilde{N}_a\|.$$

Finally, applying Proposition 19, we immediately obtain

$$\mathbb{P} \left(\frac{|\mathcal{E}|}{n} \leq \text{poly}(\eta) \left(1 + \frac{(2 + \epsilon)}{\tilde{J}(\Phi, \rho)} \right) \sqrt{\frac{SAn}{TH}} \right) \geq 1 - \frac{2}{n} - 2e^{-n} - 2e^{-\frac{TH}{2nA}}.$$

□

F.3 Separability

Definition 7. *For a given BMDP Φ , we define the separability quantity as follows*

$$J(\Phi, \rho) \triangleq \min_{\substack{\nu \in \mathcal{P}(\mathcal{X}, n^3), \\ x, y \in \mathcal{X} : f(x) \neq f(y)}} \frac{1}{2A} \sum_{a \in \mathcal{A}} J_1(\nu, x, y; \Phi, \rho, a) + J_2(\nu, x, y; \Phi, \rho, a)$$

where

$$J_1(\nu, x, y; \Phi, \rho, a) = \sum_{z \in \mathcal{X}} nA |\nu(z)\rho(a|z)P_a(z, x) - \nu(z)\rho(a|z)P_a(z, y)|,$$

$$J_2(\nu, x, y; \Phi, \rho, a) = \sum_{z \in \mathcal{X}} nA |\nu(x)\rho(a|x)P_a(x, z) - \nu(y)\rho(a|y)P_a(y, z)|,$$

and where $\mathcal{P}(\mathcal{X}, \eta^3) = \{\nu : \nu \text{ is a probability distribution over } \mathcal{X}, \max_{z_1, z_2 \in \mathcal{X}} \frac{\nu(z_1)}{\nu(z_2)} \leq \eta^3\}$, namely, the set of all η^3 -regular probability distributions.

It is not difficult to verify that under Assumptions 1-4, we have that $J(\Phi, \rho) \leq \text{poly}(\eta)$.

Proposition 17 (Separability property). *Under Assumptions 1-4, the matrix \tilde{N} satisfies the following: for all $x, y \in \mathcal{X}$ such that $f(x) \neq f(y)$, we have:*

$$\|\tilde{N}(x, \cdot) - \tilde{N}(y, \cdot)\|_1 \geq \frac{2TH}{n} J(\Phi, \rho). \quad (62)$$

Consequently, it also holds that:

$$\|\tilde{R}(x, \cdot) - \tilde{R}(y, \cdot)\|_1 \geq \tilde{J}(\Phi, \rho) \quad (63)$$

for some $\tilde{J}(\Phi, \rho) = \text{poly}(1/\eta)J(\Phi, \rho)$. Furthermore, when $I(\Phi) > 0$, then $J(\Phi, \rho) > 0$.

Proof of Proposition 17. Let $a \in \mathcal{A}$, and $x, y \in \mathcal{X}$ such that $f(x) \neq f(y)$. First, to ease notations, we introduce for all $z \in \mathcal{X}$,

$$\nu(z) = \frac{1}{T(H-1)} \left(\sum_{t=1}^T \left(\mu(z) + (\mu P_0)(z) + \sum_{h=3}^{H-1} P_0(x_{h-1}^{(t)}, z) \right) \right),$$

Clearly, $\nu = (\nu(z))_{z \in \mathcal{X}}$ is a probability distribution over \mathcal{X} , and moreover, one can easily see that under Assumptions 2-4, ν is η^3 -regular. More precisely, we also have, for all $z \in \mathcal{X}$,

$$\frac{1}{\eta^3 n} \leq \nu(z) \leq \frac{\eta^3}{n}.$$

(Proof of (62)). First, noting that $\tilde{N}_a(z, x) = TH\nu(z)\rho(a|z)P_a(z, x)$, we have that

$$\begin{aligned} \|\tilde{N}_a(\cdot, x) - \tilde{N}_a(\cdot, y)\|_1 &= \sum_{z \in \mathcal{X}} \left| \tilde{N}_a(z, x) - \tilde{N}_a(z, y) \right| \\ &= \frac{TH}{nA} \sum_{z \in \mathcal{X}} nA |\nu(z)\rho(a|z)P_a(z, x) - \nu(z)\rho(a|z)P_a(z, y)| \\ &\geq \frac{TH}{nA} J_1(\nu, x, y; \Phi, \rho, a). \end{aligned}$$

Similarly, we have:

$$\begin{aligned} \|\tilde{N}_a(x, \cdot) - \tilde{N}_a(y, \cdot)\|_1 &= \sum_{z \in \mathcal{X}} \left| \tilde{N}_a(x, z) - \tilde{N}_a(y, z) \right| \\ &= \frac{TH}{nA} \sum_{z \in \mathcal{X}} nA |\nu(x)\rho(a|x)P_a(x, z) - \nu(y)\rho(a|y)P_a(y, z)| \\ &\geq \frac{TH}{nA} J_2(\nu, x, y; \Phi, \rho, a). \end{aligned}$$

Therefore, we obtain

$$\|\tilde{N}(x, \cdot) - \tilde{N}(y, \cdot)\|_1 = \sum_{a \in \mathcal{A}} \|\tilde{N}_a(x, \cdot) - \tilde{N}_a(y, \cdot)\|_1 + \|\tilde{N}_a(\cdot, x) - \tilde{N}_a(\cdot, y)\|_1$$

$$\begin{aligned} &\geq \frac{TH}{nA} \sum_{a \in \mathcal{A}} J_1(\nu, x, y; \Phi, \rho, a) + J_2(\nu, x, y; \Phi, \rho, a) \\ &\geq \frac{2TH}{n} J(\Phi, \rho). \end{aligned}$$

(*Proof of (63)*). Next, under Assumption 3, in view of Proposition 2, we have for all $t \geq 1, h \geq 1$,

$$\frac{1}{\eta^5 n^2 A} \leq \min_{x, y \in \mathcal{X}} \tilde{N}_{a, t, h}(x, y) \leq \max_{x, y \in \mathcal{X}} \tilde{N}_{a, t, h}(x, y) \leq \frac{\eta^5}{n^2 A},$$

which leads to

$$\begin{aligned} \frac{1}{\eta^5 n A} &\leq \max_{x, y \in \mathcal{X}} \left(\|\tilde{N}_{a, t, h}(x, \cdot)\|_1, \|\tilde{N}_{a, t, h}(\cdot, y)\|_1 \right) \leq \frac{\eta^5}{n A} \\ \frac{1}{\eta^5 n A} &\leq \min_{x, y \in \mathcal{X}} \left(\|\tilde{N}_{a, t, h}(x, \cdot)\|_1, \|\tilde{N}_{a, t, h}(\cdot, y)\|_1 \right) \leq \frac{\eta^5}{n A} \end{aligned}$$

and gives that

$$\frac{2TH}{\eta^5 n} \leq \min \left(\|\tilde{N}(x, \cdot)\|_1, \|\tilde{N}(y, \cdot)\|_1 \right) \leq \frac{2\eta^5 TH}{n}. \quad (64)$$

Therefore, recalling the definition of \tilde{R} , we have for all $x, y \in \mathcal{X}$, such that $f(x) \neq f(y)$,

$$\|\tilde{R}(x, \cdot) - \tilde{R}(y, \cdot)\|_1 \geq \frac{\|\tilde{N}(x, \cdot) - \tilde{N}(y, \cdot)\|_1}{\min(\|\tilde{N}(x, \cdot)\|_1, \|\tilde{N}(y, \cdot)\|_1)} \geq \tilde{J}(\Phi, \rho) = \frac{1}{\eta^5} J(\Phi, \rho).$$

(*Proving $I(\Phi) > 0 \implies J(\Phi, \rho) > 0$*). Now let us prove that if $J(\Phi, \rho) = 0$, then $I(\Phi) = 0$. Assume that $J(\Phi, \rho) = 0$, then there exist $x, y \in \mathcal{X}$ with $f(x) \neq f(y)$, and $\nu \in \mathcal{P}(\mathcal{X}, \eta^3)$ such that for all $a \in \mathcal{A}$,

$$\sum_{z \in \mathcal{X}} nA |\nu(z)\rho(a|z)P_a(z, x) - \nu(z)\rho(a|z)P_a(z, y)| = 0, \quad (65)$$

$$\sum_{z \in \mathcal{X}} nA |\nu(x)\rho(a|x)P_a(x, z) - \nu(y)\rho(a|y)P_a(y, z)| = 0. \quad (66)$$

Now observe that:

1. from Eqn. (65), we can immediately deduce that for all $a \in \mathcal{A}, z \in \mathcal{X}, P_a(z, x) = P_a(z, y)$. This entails that

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \quad q(x, f(x))p(f(x)|s, a) = q(y, f(y))p(f(y)|s, a); \quad (67)$$

2. from Eqn. (66), it must be the case that for all $a \in \mathcal{A}, z \in \mathcal{X}, \nu(x)\rho(a|x)P_a(x, z) = \nu(y)\rho(a|y)P_a(y, z)$. By summing over $z \in \mathcal{X}$, we deduce that $\nu(x)\rho(a|x) = \nu(y)\rho(a|y) > 0$ because $\nu \in \mathcal{P}(\mathcal{X}, \eta^3)$. This entails that

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \quad p(s|f(x), a) = p(s|f(y), a). \quad (68)$$

Now in view of Proposition 10, we observe that (67) and (68) imply that $\min_{x \in \mathcal{X}} I(x; \Phi) = 0$, which in turn implies that $I(\Phi) = 0$. □

F.4 Weighted K-medians clustering

The solution to (61). The solution to the $(1+\epsilon)$ -K-medians optimization (61) on $\mathcal{X} \setminus \mathcal{X}_0$ can be obtained efficiently (e.g., see Chen et al. (2018); Gao et al. (2018) from which we took inspiration). Here, we recall that $\mathcal{X}_0 = \{x \in \mathcal{X} : \|\hat{M}(x, \cdot)\|_1 = 0\}$. Thus, let us denote $\{\hat{f}(x)\}_{x \in \mathcal{X} \setminus \mathcal{X}_0}$ in $\mathcal{S}^{\mathcal{X} \setminus \mathcal{X}_0}$ and $\hat{u}_1, \dots, \hat{u}_S \in \mathbb{R}^n$ such a solution. We further set $\hat{f}(x) = 1$, for all $x \in \mathcal{X}_0$, and define $\hat{U}(x, \cdot) = \hat{u}_{\hat{f}(x)}$ for all $x \in \mathcal{X}$. With this, observe that by definition of \hat{f} and $\hat{U} = (\hat{U}(x, \cdot))_{x \in \mathcal{X}}$, we have

$$\sum_{x \in \mathcal{X}} \|\hat{M}(x, \cdot)\|_1 \|\tilde{M}(x, \cdot) - \hat{U}(x, \cdot)\|_1 \leq (1+\epsilon) \sum_{x \in \mathcal{X}} \|\hat{M}(x, \cdot)\|_1 \|\tilde{M}(x, \cdot) - \tilde{R}(x, \cdot)\|_1. \quad (69)$$

Note that the choice of ϵ is irrelevant in our analysis and can be viewed as a constant hyperparameter.

Linking $|\mathcal{E}|$ to the geometry of the rows of \hat{U} and \tilde{R} . Next, we will need the technical Lemma 11 which relates the number of misclassified contexts $|\mathcal{E}|$ to that of the geometry of points, namely the rows of \tilde{R} and \hat{U} in our case, provided some separability condition is satisfied for \tilde{R} . The statement of the lemma is valid for any norm $\|\cdot\|$.

Lemma 11 (Lemma 6 in Gao et al. (2018)). *Assume that $\min_{x,y \in \mathcal{X}: f(x) \neq f(y)} \|\tilde{R}(x, \cdot) - \tilde{R}(y, \cdot)\| \geq 2\xi$ for some $\xi > 0$. Then it holds that*

$$|\mathcal{E}| = \left| \min_{\sigma \in \Upsilon(\mathcal{S})} \bigcup_{s \in \mathcal{S}} \hat{f}^{-1}(\sigma(s)) \setminus f^{-1}(s) \right| \leq |\mathcal{X}_0| + (\eta^2 + 1)|\mathcal{X}_1| \quad (70)$$

where we define $\mathcal{X}_1 = \{x \in \mathcal{X} \setminus \mathcal{X}_0 : \|\hat{U}(x, \cdot) - \tilde{R}(x, \cdot)\|_1 \geq \xi\}$.

The proof of Lemma 11 is borrowed from Gao et al. (2018). The statement we provide here differs slightly from that provided by Gao et al. (2018) as they considered the *Degree Corrected Block Model* while in our case we consider BMDPs. However, the proof is essentially the same and is provided here for completeness (postponed end of this subsection).

Now, invoking Proposition 17, we have under the assumption that $I(\Phi) > 0$, for all $x, y \in \mathcal{X}$ such that $f(x) \neq f(y)$, $\|\tilde{R}(x, \cdot) - \tilde{R}(y, \cdot)\|_1 \geq \tilde{J}(\Phi, \rho)$. Thus, applying Lemma 11 specialised to the ℓ_1 norm, we immediately obtain that

$$|\mathcal{E}| \leq |\mathcal{X}_0| + (\eta^2 + 1)|\mathcal{X}_1| \quad (71)$$

where $\mathcal{X}_1 = \{x \in \mathcal{X} \setminus \mathcal{X}_0 : \|\hat{U}(x, \cdot) - \tilde{R}(x, \cdot)\|_1 \geq \tilde{J}(\Phi, \rho)/2\}$.

Starting from our regularity Assumption 3, observe that

$$\frac{|\mathcal{X}_0|}{n} \leq \frac{\eta^5}{2TH} \sum_{x \in \mathcal{X}_0} \|\tilde{N}(x, \cdot)\|_1 \quad \text{and} \quad \frac{|\mathcal{X}_1|}{n} \leq \frac{\eta^5}{2TH} \sum_{x \in \mathcal{X}_1} \|\tilde{N}(x, \cdot)\|_1.$$

Bounding $\sum_{x \in \mathcal{X}_0} \|\tilde{N}(x, \cdot)\|_1$. We have:

$$\begin{aligned} \sum_{x \in \mathcal{X}_0} \|\tilde{N}(x, \cdot)\|_1 &= \sum_{x \in \mathcal{X}_0} \|\hat{M}(x, \cdot) - \tilde{N}(x, \cdot)\|_1 \\ &\leq \sum_{x \in \mathcal{X}} \|\hat{M}(x, \cdot) - \tilde{N}(x, \cdot)\|_1 \\ &\leq n\sqrt{A} \|\hat{M} - \tilde{N}\|_F, \end{aligned}$$

where the first equality holds by definition of \mathcal{X}_0 , and the last inequality holds by the equivalence of norms on matrices (essentially using Cauchy-Schwarz inequality).

Bounding $\sum_{x \in \mathcal{X}_1} \|\tilde{N}(x, \cdot)\|_1$. First, by triangle inequality,

$$\begin{aligned} \sum_{x \in \mathcal{X}_1} \|\tilde{N}(x, \cdot)\|_1 &\leq \sum_{x \in \mathcal{X}_1} \left(\|\tilde{N}(x, \cdot) - \hat{M}(x, \cdot)\|_1 + \|\hat{M}(x, \cdot)\|_1 \right) \\ &\leq n\sqrt{A} \|\hat{M} - \tilde{N}\|_F + \sum_{x \in \mathcal{X}_1} \|\hat{M}(x, \cdot)\|_1, \end{aligned}$$

where again for the second inequality, we used the equivalence between norms on matrices. Next, we have:

$$\begin{aligned} \sum_{x \in \mathcal{X}_1} \|\hat{M}(x, \cdot)\|_1 &\stackrel{(a)}{\leq} \frac{2}{\tilde{J}(\Phi, \rho)} \sum_{x \in \mathcal{X}_1} \|\hat{M}(x, \cdot)\|_1 \|\hat{U}(x, \cdot) - \tilde{R}(x, \cdot)\|_1 \\ &\stackrel{(b)}{\leq} \frac{2}{\tilde{J}(\Phi, \rho)} \sum_{x \in \mathcal{X}_1} \|\hat{M}(x, \cdot)\|_1 \left(\|\tilde{M}(x, \cdot) - \hat{U}(x, \cdot)\|_1 + \|\tilde{M}(x, \cdot) - \tilde{R}(x, \cdot)\|_1 \right) \\ &\stackrel{(c)}{\leq} \frac{2(2 + \epsilon)}{\tilde{J}(\Phi, \rho)} \sum_{x \in \mathcal{X}_1} \|\hat{M}(x, \cdot)\|_1 \|\tilde{M}(x, \cdot) - \tilde{R}(x, \cdot)\|_1 \\ &\stackrel{(d)}{\leq} \frac{4(2 + \epsilon)}{\tilde{J}(\Phi, \rho)} \sum_{x \in \mathcal{X}_1} \frac{\|\hat{M}(x, \cdot)\|_1 \|\tilde{M}(x, \cdot) - \tilde{N}(x, \cdot)\|_1}{\max(\|\hat{M}(x, \cdot)\|_1, \|\tilde{N}(x, \cdot)\|_1)} \end{aligned}$$

$$\begin{aligned}
 &\stackrel{(e)}{\leq} \frac{4(2+\epsilon)}{\tilde{J}(\Phi, \rho)} \sum_{x \in \mathcal{X}_1} \|\tilde{M}(x, \cdot) - \tilde{N}(x, \cdot)\|_1 \\
 &\stackrel{(f)}{\leq} \frac{4(2+\epsilon)}{\tilde{J}(\Phi, \rho)} n\sqrt{A} \|\hat{M} - \hat{N}\|_F,
 \end{aligned}$$

where in (a) we used the definition of \mathcal{X}_1 , in (b) we used the triangular inequality, in (c) we used the definition \hat{U} and (69), in (d) we used the elementary inequality that $\|\|x\|_1^{-1}x - \|y\|_1^{-1}y\|_1 \leq \frac{2\|x-y\|_1}{\max(\|x\|_1, \|y\|_1)}$ (this can be shown by triangular inequality and reverse triangular inequality), in (e) we used $\frac{\|x\|_1}{\max(\|x\|_1, \|y\|_1)} \leq 1$, and finally in (f) we used equivalence of norms between matrices. In summary, we get:

$$\sum_{x \in \mathcal{X}_1} \|\tilde{N}(x, \cdot)\|_1 \leq \left(1 + \frac{4(2+\epsilon)}{\tilde{J}(\Phi, \rho)}\right) n\sqrt{A} \|\hat{M} - \hat{N}\|_F$$

Thus, we have just proved that:

$$\frac{|\mathcal{X}_0|}{n} \leq \frac{\eta^5 n\sqrt{A}}{2TH} \|\hat{M} - \hat{N}\|_F, \tag{72}$$

$$\frac{|\mathcal{X}_1|}{n} \leq \left(1 + \frac{4(2+\epsilon)}{\tilde{J}(\Phi, \rho)}\right) \frac{\eta^5 n\sqrt{A}}{2TH} \|\hat{M} - \hat{N}\|_F. \tag{73}$$

Now, in view of the inequalities (71), (72), (73), we have established the following Proposition.

Proposition 18. *Assume that $I(\Phi) > 0$, and that Assumptions 1-4 hold, then the weighted K-medians clustering algorithm ensures that*

$$\frac{|\mathcal{E}|}{n} \leq \left(2 + \eta^2 + \frac{4(2+\epsilon)\eta^2}{\tilde{J}(\Phi, \rho)}\right) \frac{2\eta^5 n\sqrt{A}}{TH} \|\hat{M} - \hat{N}\|_F.$$

Proof of Lemma 11. For each $s \in \mathcal{S}$, define

$$\mathcal{C}_s := \left\{x \in f^{-1}(s) \cap (\mathcal{X} \setminus \mathcal{X}_0) : \|\hat{U}(x, \cdot) - \tilde{R}(x, \cdot)\|_1 < \xi\right\}.$$

By construction, we have that $\bigcup_{s \in \mathcal{S}} \mathcal{C}_s = \mathcal{X} \setminus (\mathcal{X}_0 \cup \mathcal{X}_1)$, and that $\mathcal{C}_s \cup \mathcal{C}_{s'} = \emptyset$ whenever $s \neq s'$. Also, by assumption, it is easy to see that $\hat{f}(x) \neq \hat{f}(y)$ if x, y are in different \mathcal{C}_s 's. Following Chen et al. (2018), we partition \mathcal{X} into three groups:

$$\begin{aligned}
 R_1 &:= \{s \in \mathcal{S} : \mathcal{C}_s = \emptyset\}, \\
 R_2 &:= \left\{s \in \mathcal{S} : \mathcal{C}_s \neq \emptyset, \forall x, y \in \mathcal{C}_s \hat{f}(x) = \hat{f}(y)\right\}, \\
 R_3 &:= \left\{s \in \mathcal{S} : \mathcal{C}_s \neq \emptyset, \exists x \neq y \in \mathcal{C}_s \hat{f}(x) \neq \hat{f}(y)\right\}.
 \end{aligned}$$

By the definition of R_2 , we observe that the contexts in $\bigcup_{s \in R_2} \mathcal{C}_s$ have the same partition induced by f and \hat{f} i.e. they can be considered to be correctly classified, up to a permutation. Thus,

$$|\mathcal{E}| \leq |\mathcal{X}_0 \cup \mathcal{X}_1| + \left| \bigcup_{s \in R_3} \mathcal{C}_s \right| \leq |\mathcal{X}_0| + |\mathcal{X}_1| + \left| \bigcup_{s \in R_3} \mathcal{C}_s \right|.$$

Note that for each $s \in \mathcal{S}$, \mathcal{C}_s contains at least two different cluster indices given by \hat{f} i.e.

$$|R_1| + |R_2| + |R_3| = S \geq |R_2| + 2|R_3|,$$

which then implies that $|R_1| \geq |R_3|$. We now conclude the proof by noting that

$$\left| \bigcup_{s \in R_3} \mathcal{C}_s \right| \leq |R_3| \frac{\eta n}{S}$$

$$\begin{aligned}
&\leq \eta |R_1| \frac{n}{S} \\
&\leq \eta^2 \left| \bigcup_{s \in R_1} f^{-1}(s) \right| \\
&\leq \eta^2 |\mathcal{X}_1|.
\end{aligned}$$

□

F.5 S -rank approximation

Lemma 12 allows us to control the error in Frobenius norm between \hat{M} and \tilde{N} , which is necessary to bound the number of misclassified contexts to that in operator norm between the trimmed matrices \hat{N}_{a,Γ_a} and \tilde{N}_a for which we are able to provide a concentration bound.

Lemma 12. *Under Assumption 1-4, after Algorithm 1, we have:*

$$\|\hat{M} - \tilde{N}\|_F \leq 4\sqrt{SA} \max_{a \in \mathcal{X}} \|\hat{N}_{a,\Gamma_a} - \tilde{N}_a\|$$

Proof of Lemma 12. We have

$$\begin{aligned}
\|\hat{M} - \tilde{N}\|_F^2 &= 2 \sum_{a \in \mathcal{A}} \|\hat{M}_a - \tilde{N}_a\|_F^2 \\
&\stackrel{(a)}{\leq} 4S \sum_{a \in \mathcal{A}} \|\hat{M}_a - \tilde{N}_a\|^2 \\
&\stackrel{(b)}{\leq} 4S \sum_{a \in \mathcal{A}} \left(\|\hat{M}_a - \hat{N}_{a,\Gamma_a}\| + \|\hat{N}_{a,\Gamma_a} - \tilde{N}_a\| \right)^2 \\
&\stackrel{(c)}{\leq} 4S \sum_{a \in \mathcal{A}} \left(\sigma_{|S|+1}(\hat{N}_{a,\Gamma_a}) + \|\hat{N}_{a,\Gamma_a} - \tilde{N}_a\| \right)^2 \\
&\stackrel{(d)}{\leq} 4S \sum_{a \in \mathcal{A}} \left(\sigma_{|S|+1}(\tilde{N}_a) + 2\|\hat{N}_{a,\Gamma_a} - \tilde{N}_a\| \right)^2 \\
&\stackrel{(e)}{\leq} 16SA \max_{a \in \mathcal{A}} \|\hat{N}_{a,\Gamma_a} - \tilde{N}_a\|^2
\end{aligned}$$

where inequality (a) follows from $\|M\|_F \leq \sqrt{\text{rank}(M)}\|M\|$ and the facts that $\text{rank}(\hat{M}_a) \leq S$ by construction and $\tilde{N}_a = LP_a$ for some random matrix L , and thus $\text{rank}(\tilde{N}_a) \leq \text{rank}(P_a) \leq S$ by the structure of the P_a . Inequality (b) follows by triangular inequality. Inequality (c) follows by construction of \hat{M}_a for all $a \in \mathcal{A}$. Inequality (d) follows using Weyl's inequality, and finally inequality (e) follows by noting again that \tilde{N}_a is at most of rank S , thus $\sigma_{S+1}(\tilde{N}_a) = 0$. □

F.6 Analysis of the trimmed random matrix

This subsection is devoted to the derivation of concentration results for the matrices \hat{N}_{a,Γ_a} obtained after trimming the observation matrices. These concentration results are central to the performance analysis of our algorithms. The proof techniques used here draw inspiration from those used in SBMs, BMCs, and matrix completion problems (Feige and Ofek, 2005; Keshavan et al., 2010a; Sanders et al., 2020; Sanders and Senen-Cerda, 2023). We adapt these techniques to our setting. The proof relies on the analysis of *light* and *heavy* couples, an involved net argument, where the analysis of the *heavy* couples relies on the so-called *discrepancy property* (Feige and Ofek, 2005; Keshavan et al., 2010a). Furthermore, the observations upon which the matrices \hat{N}_{a,Γ_a} are built are not independent but rather possess a Markovian nature. To tackle this challenge, we follow a similar reasoning to that in Sanders et al. (2020). Our setting is however different since we have to accommodate for restarts in episodic Block MDPs. This is done thanks to our new concentration bound (see Appendix E).

We state below the main result of this subsection.

Proposition 19. *For all $a \in \mathcal{A}$, the following holds:*

$$\mathbb{P} \left(\max_{a \in \mathcal{A}} \|\hat{N}_{a,\Gamma_a} - \tilde{N}_a\| \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right) \geq 1 - \frac{2}{n} - 2e^{-n} - 2e^{-\frac{TH}{2nA}} \quad (74)$$

Proof of Proposition 19. First, we express $\|\hat{N}_{a,\Gamma_a} - \tilde{N}_a\|$ using the variational form of the operator norm, then use the triangle inequality to obtain

$$\left\| \hat{N}_{a,\Gamma_a} - \tilde{N}_a \right\| = \sup_{u,v \in \mathbb{S}^{n-1}} u^\top \left(\hat{N}_{a,\Gamma_a} - \tilde{N}_a \right) v \leq T_1 + T_2 + T_3,$$

where \mathbb{S}^{n-1} denotes the unit sphere in \mathbb{R}^n , and T_1, T_2 and T_3 are defined as follows:

$$\begin{aligned} T_1 &\triangleq \left| \sup_{u,v \in \mathbb{S}^{n-1}} \sum_{(x,y) \in \mathcal{L} \cap \mathcal{K}^c} u_x v_y \hat{N}_a(x,y) \right|, \\ T_2 &\triangleq \left| \sup_{u,v \in \mathbb{S}^{n-1}} \sum_{(x,y) \in \mathcal{L}} u_x v_y \hat{N}_a(x,y) - u^\top \tilde{N}_a v \right|, \\ T_3 &\triangleq \left| \sup_{u,v \in \mathbb{S}^{n-1}} \sum_{(x,y) \in \mathcal{H} \cap \mathcal{K}} u_x v_y \hat{N}_a(x,y) \right|, \end{aligned}$$

with

$$\begin{aligned} \mathcal{L} &\triangleq \{(x,y) : x,y \in \mathcal{X} \text{ and } |u_x v_y| < m\} && \text{(light couples)} \\ \mathcal{K} &\triangleq \{(x,y) : x,y \in \mathcal{X} \text{ and } x,y \in \Gamma_a\} && \text{(non-trimmed couples)} \\ \mathcal{H} &\triangleq \{(x,y) : x,y \in \mathcal{X} \text{ and } |u_x v_y| \geq m\} && \text{(heavy couples)} \end{aligned}$$

where we set $m = \frac{1}{n} \sqrt{\frac{TH}{nA}}$. This choice will appear suitable for our analysis. We recall that number of trimmed nodes is exactly $\gamma = \lfloor n \exp(-\frac{TH}{nA} \log(\frac{TH}{nA})) \rfloor$. By Lemma 13, Lemma 16 and Lemma 18, that the terms T_1, T_2 , and T_3 satisfy the following high probability bounds whenever $TH = \Omega(n)$:

$$\mathbb{P} \left(T_1 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right) \geq 1 - \frac{e^{-n}}{A}, \quad (75)$$

$$\mathbb{P} \left(T_2 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right) \geq 1 - \frac{e^{-n}}{A}, \quad (76)$$

$$\mathbb{P} \left(T_3 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right) \geq 1 - \frac{2}{An} - 2e^{-\frac{TH}{nA}}. \quad (77)$$

The desired result follows from the above concentration results. Indeed, we first note that the event

$$\left\{ T_1 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right\} \cap \left\{ T_2 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right\} \cap \left\{ T_3 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right\}$$

is a subset of the event

$$\left\{ \left\| \hat{N}_{a,\Gamma_a} - \tilde{N}_a \right\| \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right\}.$$

Thus, using the union bound, we conclude that:

$$\mathbb{P} \left(\left\| \hat{N}_{a,\Gamma_a} - \tilde{N}_a \right\| \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right) \geq 1 - \frac{2}{An} - \frac{2e^{-n}}{A} - 2e^{-TH/nA}$$

which further implies, applying the union bound once more, that

$$\mathbb{P} \left(\max_{a \in \mathcal{A}} \left\| \hat{N}_{a,\Gamma_a} - \tilde{N}_a \right\| \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}} \right) \geq 1 - \frac{2}{n} - 2e^{-n} - 2e^{-TH/2nA},$$

for $TH = \Omega(n)$ where $\Omega(\cdot)$ hides a dependence in $A \log(A)$ (where e used the fact that $TH \geq 2nA \log(A)$ gives $e^{-\frac{TH}{nA} + \log(A)} \leq e^{-\frac{TH}{2nA}}$). \square

Bounding the contribution of the light couples. In Lemmas 13 and 16, we obtain bounds on the terms that depend on the light couples, that is when we sum over $(x, y) \in \mathcal{L}$.

We start by bounding T_1 :

Lemma 13 (Bounding T_1). *We have for all $TH = \Omega(n)$,*

$$\mathbb{P} \left(T_1 > \text{poly}(\eta) \left(1 + \frac{\log(A)}{n} \right) \sqrt{\frac{TH}{nA}} \right) \leq e^{-n - \log(A)}.$$

Proof of Lemma 13. First, let us observe that T_1 can be upper bounded as follows:

$$\begin{aligned} T_1 &\stackrel{(a)}{\leq} \sup_{u, v \in \mathbb{S}^{n-1}} \sum_{(x, y) \in \mathcal{L} \cap \mathcal{K}^c} |u_x v_y| \hat{N}_a(x, y) \\ &\stackrel{(b)}{\leq} m \sum_{(x, y) \in \mathcal{K}^c} \hat{N}_a(x, y) \\ &\leq m \hat{N}_a(\mathcal{X}, \Gamma_a^c) \\ &\leq m \max_{\mathcal{Y}: |\mathcal{Y}|=\gamma} \hat{N}_a(\mathcal{X}, \mathcal{Y}) \end{aligned}$$

where in inequality (a) we used the triangle inequality, in (b) we used the fact that $(x, y) \in \mathcal{L}$ (i.e., light couples). Next, we know from Lemma 14 that for all $TH \geq 2n|\mathcal{A}|$, for all $\mathcal{Y} \subseteq \mathcal{X}$ such that $|\mathcal{Y}| = \gamma$, for all $u > 0$, we have

$$\max_{\mathcal{Y}: |\mathcal{Y}|=\gamma} \mathbb{P}(\hat{N}_a(\mathcal{X}, \mathcal{Y}) > \text{poly}(\eta)n(1 + \max(\sqrt{u}, u))) \leq e^{-nu}.$$

Therefore, we have by union bound

$$\begin{aligned} \mathbb{P}(T_1 > \text{poly}(\eta)mn(1 + \max(\sqrt{u}, u))) &\leq \sum_{\mathcal{Y}: |\mathcal{Y}|=\gamma} \mathbb{P}(\hat{N}_a(\mathcal{X}, \mathcal{Y}) > \text{poly}(\eta)n(1 + \max(\sqrt{u}, u))) \\ &\leq e^{-n(u - \log(2))} \end{aligned}$$

where we used the upper bound $|\{\mathcal{Y} \subseteq \mathcal{X} : |\mathcal{Y}| = \gamma\}| \leq 2^n$. To conclude, we simply recall that $m = \frac{1}{n} \sqrt{\frac{TH}{nA}}$, plug in its value in the final concentration, and choose u sufficiently large. \square

Lemma 14. *Let $\mathcal{Y} \subseteq \mathcal{X}$ such that $|\mathcal{Y}| = \lfloor n \exp(-\frac{TH}{nA} \log(\frac{TH}{nA})) \rfloor$. Assume that $TH \geq 2nA$, then for all $u > 0$, we have*

$$\mathbb{P} \left(\left| \hat{N}_a(\mathcal{X}, \mathcal{Y}) - \mathbb{E}_\mu[\hat{N}_a(\mathcal{X}, \mathcal{Y})] \right| > \text{poly}(\eta)n \max(\sqrt{u}, u) \right) \leq 2e^{-nu},$$

Furthermore, we have

$$\mathbb{P} \left(\hat{N}_a(\mathcal{X}, \mathcal{Y}) > \text{poly}(\eta)n(1 + \max(\sqrt{u}, u)) \right) \leq 2e^{-nu}.$$

Proof of Lemma 14. The result follows immediately from Theorem 9. Indeed, consider the induced restarted Markov chain $(x_{h+1}^{(t)}, a_h^{(t)})_{t \geq 1, h \geq 1}$, referred to as MC_1 in Appendix C. By Proposition 4, the transition kernel and initial distribution of MC_1 are both η^3 -regular. Introducing $\phi : (y, b) \mapsto \mathbb{1}\{b = a, y \in \mathcal{Y}\}$, we see that $\hat{N}_a(\mathcal{X}, \mathcal{Y}) = \sum_{t=1}^T \sum_{h=1}^H \phi(x_{h+1}^{(t)}, a_h^{(t)})$. Therefore applying Theorem 9, we obtain for all $u > 0$

$$\begin{aligned} \mathbb{P} \left(\left| \hat{N}_a(\mathcal{X}, \mathcal{Y}) - \mathbb{E}_\mu[\hat{N}_a(\mathcal{X}, \mathcal{Y})] \right| > u \right) &\leq 2 \exp \left(- \frac{u^2}{2(1 + \sqrt{2}\eta^3(2\eta^3 - 1))^2 TH \frac{\gamma}{nA} \eta^3 + \frac{2}{3}(2\eta^3 - 1)u} \right) \\ &\leq 2 \exp \left(- \frac{1}{2} \min \left(\frac{nAu^2}{c_1 TH \gamma}, \frac{u}{c_2} \right) \right), \end{aligned}$$

where $c_1 = (1 + \sqrt{2}\eta^3(2\eta^3 - 1))^2\eta^3$ and $c_2 = \frac{1}{3}(2\eta^3 - 1)$. Reparameterizing by $u' = \frac{1}{2n} \min\left(\frac{nAu^2}{c_1TH\gamma}, \frac{u}{c_2}\right)$ yields

$$\mathbb{P}\left(|\hat{N}_a(\mathcal{X}, \mathcal{Y}) - \mathbb{E}_\mu[\hat{N}_a(\mathcal{X}, \mathcal{Y})]| > \max\left(\sqrt{2c_1nu'}\frac{\gamma TH}{nA}, 2c_2nu'\right)\right) \leq 2\exp(-nu')$$

Recalling that $\gamma = \lfloor n \exp(-\frac{TH}{nA} \log(\frac{TH}{nA})) \rfloor$, we can easily verify that $\gamma \leq \frac{3n^2A}{TH}$ (which follows from the elementary inequalities that $x \log(x) + 1 \geq x$ and $e^{-x} \leq \frac{1}{x}$ for $x > 0$). Thus, we may simply write that

$$\mathbb{P}\left(|\hat{N}_a(\mathcal{X}, \mathcal{Y}) - \mathbb{E}_\mu[\hat{N}_a(\mathcal{X}, \mathcal{Y})]| > nc \max(\sqrt{u'}, u')\right) \leq 2\exp(-nu'),$$

where $c = \max(\sqrt{6c_1}, 2c_2) = \text{poly}(\eta)$. The additional bound follows from the fact that $\mathbb{E}[\hat{N}_a(\mathcal{X}, \mathcal{Y})] \leq 3\eta^4n$. \square

The following lemma is a standard net argument, which will be useful for bounding T_2 :

Lemma 15 (ϵ -net argument on the unit sphere). *Let W be an $n \times n$ random matrix, for all $\epsilon \in (0, 1/2)$, let \mathcal{N}_ϵ an ϵ -net of the unit sphere with respect to the Euclidian distance $\|\cdot\|_2$ and with minimal cardinality. Then, for any $\rho > 0$,*

$$\mathbb{P}(\|W\| > \rho) \leq \left(\frac{2}{\epsilon} + 1\right)^{2n} \max_{u, v \in \mathcal{N}_\epsilon} \mathbb{P}(u^\top W v > (1 - 2\epsilon)\rho).$$

Proof of Lemma 15. The existence of \mathcal{N}_ϵ is shown in Corollary 4.2.13 of [Vershynin \(2018\)](#), which states that $|\mathcal{N}_\epsilon| \leq \left(\frac{2}{\epsilon} + 1\right)^n$. The statement then follows from Exercise 4.4.3 of [Vershynin \(2018\)](#), which relates the spectral norm of a random matrix to that of maximizing its rectangular form over \mathcal{N}_ϵ . \square

We now bound T_2 :

Lemma 16 (Bounding T_2). *We have, for $n \geq 2 \log(A)$,*

$$\mathbb{P}\left(T_2 > \text{poly}(\eta) \sqrt{\frac{TH}{nA}}\right) \leq 2e^{-n - \log(A)}.$$

Proof of Lemma 16. Our first step is to split the sum based on the parity of the time steps per episode. More precisely, we write $T_2 \leq T_2^{\text{even}} + T_2^{\text{odd}}$ and define

$$\begin{aligned} T_2^{\text{even}} &= \sup_{u, v \in \mathbb{S}^{n-1}} \sum_{t=1}^T \sum_{h=1}^{\lfloor \frac{H-1}{2} \rfloor} \left(\sum_{(x, y) \in \mathcal{L}} u_x v_y \hat{N}_{a, t, 2h}(x, y) \right) - u^\top \tilde{N}_{a, t, 2h} v, \\ T_2^{\text{odd}} &= \sup_{u, v \in \mathbb{S}^{n-1}} \sum_{t=1}^T \sum_{h=0}^{\lfloor \frac{H}{2} \rfloor - 1} \left(\sum_{(x, y) \in \mathcal{L}} u_x v_y \hat{N}_{a, t, 2h+1}(x, y) \right) - u^\top \tilde{N}_{a, t, 2h+1} v. \end{aligned}$$

We note that T_2^{even} and T_2^{odd} are expressed as supremum over the unit sphere. To proceed we will deploy a net argument, but first let us define for all $u, v \in \mathbb{S}^{n-1}$,

$$\begin{aligned} T_2^{\text{even}}(u, v) &= \sum_{t=1}^T \sum_{h=1}^{\lfloor \frac{H-1}{2} \rfloor} \left(\sum_{(x, y) \in \mathcal{L}} u_x v_y \hat{N}_{a, t, 2h+1}(x, y) \right) - u^\top \tilde{N}_{a, t, 2h+1} v, \\ T_2^{\text{odd}}(u, v) &= \sum_{t=1}^T \sum_{h=0}^{\lfloor \frac{H}{2} \rfloor - 1} \left(\sum_{(x, y) \in \mathcal{L}} u_x v_y \hat{N}_{a, t, 2h+1}(x, y) \right) - u^\top \tilde{N}_{a, t, 2h+1} v. \end{aligned}$$

The analysis of $T_2^{\text{even}}(u, v)$ and $T_2^{\text{odd}}(u, v)$ will be the same, therefore, and without loss of generality, we only show how to obtain a concentration bound for $T_2^{\text{odd}}(u, v)$. We start by computing its moment-generating function. First, let $\lambda > 0$, $h \geq 3$, we have:

$$\mathbb{E}_\mu \left[\exp \left(\lambda \left(\sum_{(x, y) \in \mathcal{L}} u_x v_y \hat{N}_{a, t, h}(x, y) \right) - \lambda u^\top \tilde{N}_{a, t, h} v \right) \middle| \tilde{N}_{a, t, h-2} \right]$$

$$\begin{aligned}
 &= \left(\sum_{(x,y) \in \mathcal{X} \times \mathcal{X}} (\mathbb{1}\{(x,y) \notin \mathcal{L}\} + \mathbb{1}\{(x,y) \in \mathcal{L}\} e^{\lambda u_x v_y}) \tilde{N}_{a,t,h}(x,y) \right) \exp(-\lambda u^\top \tilde{N}_{a,t,h} v) \\
 &= \left(1 + \sum_{(x,y) \in \mathcal{L}} \tilde{N}_{a,t,h}(x,y) (e^{\lambda u_x v_y} - 1) \right) \exp(-\lambda u^\top \tilde{N}_{a,t,h} v) \\
 &\stackrel{(a)}{\leq} \exp \left(\sum_{(x,y) \in \mathcal{L}} \tilde{N}_{a,t,h}(x,y) \left(\lambda u_x v_y + \frac{e^{\lambda |u_x v_y|}}{2} (\lambda u_x v_y)^2 \right) - u^\top \tilde{N}_{a,t,h} v \right) \\
 &\stackrel{(b)}{\leq} \exp \left(-\lambda \left(\sum_{(x,y) \in \mathcal{L}^c} u_x v_y \tilde{N}_{a,t,h}(x,y) \right) + \frac{\lambda^2 e^{\lambda m}}{2} \left(\sum_{(x,y) \in \mathcal{L}} (u_x v_y)^2 \tilde{N}_{a,t,h}(x,y) \right) \right) \\
 &\stackrel{(c)}{\leq} \exp \left(\lambda \frac{\eta^5}{n^2 A m} + \lambda^2 e^{\lambda m} \frac{\eta^5}{2 n^2 A} \right)
 \end{aligned}$$

where in the inequality (a), we use the elementary inequalities $1+x \leq \exp(x)$, then $e^{\lambda u_x v_y} - 1 \leq \lambda u_x v_y + \frac{e^{\lambda |u_x v_y|}}{2} (\lambda u_x v_y)^2$, in the inequality (b), we used $|u_x v_y| \leq m$ for all $(x,y) \in \mathcal{L}$, and in inequality (c), we used the fact that $\max_{x,y \in \mathcal{X}} \tilde{N}_{a,t,h}(x,y) \leq \frac{\eta^5}{n^2 A}$ along with

$$\left| \sum_{(x,y) \in \mathcal{L}^c} u_x v_y \tilde{N}_{a,t,h}(x,y) \right| \leq \frac{\max_{(x,y) \in \mathcal{L}} |\tilde{N}_{a,t,h}(x,y)|}{\min_{(i,j) \in \mathcal{L}^c} |u_i v_j|} \sum_{(i,j) \in \mathcal{L}^c} |u_i v_j|^2 \leq \frac{\eta^5}{n^2 A m}$$

and

$$\sum_{(x,y) \in \mathcal{L}} (u_x v_y)^2 \tilde{N}_{a,t,h}(x,y) \leq \frac{\eta^5}{n^2 A}.$$

Therefore, using a peeling argument, we obtain, for all $\lambda > 0$,

$$\begin{aligned}
 \mathbb{E}_\mu \left[\exp \left(\lambda \sum_{t=1}^T \sum_{h=0}^{\lfloor \frac{H}{2} \rfloor - 1} \left(\sum_{(x,y) \in \mathcal{L}} u_x v_y \hat{N}_{a,t,h}(x,y) \right) - \lambda u^\top \tilde{N}_{a,t,h} v \right) \right] \\
 \leq \exp \left(\lambda \frac{\eta^5 T H}{2 n^2 A m} + \lambda^2 e^{\lambda m} \frac{\eta^5 T H}{4 n^2 A} \right).
 \end{aligned}$$

Using Markov inequality and reparameterizing by $\lambda = z/m$, we obtain

$$\begin{aligned}
 \mathbb{P} \left(T_2^{\text{odd}}(u, v) > n m \eta^5 (\rho + 1) \right) &\leq \inf_{z > 0} \exp \left(z \frac{\eta^5 T H}{n^2 A m^2} + z^2 e^z \frac{\eta^5 T H}{2 n^2 A m^2} - z n \eta^5 (\rho + 1) \right) \\
 &\leq \inf_{z > 0} \exp \left(-n \eta^5 (z \rho - z^2 e^z / 2) \right),
 \end{aligned}$$

where the last inequality follows by plugging in the value of $m = \frac{1}{n} \sqrt{\frac{T H}{n A}}$. Thus, at the end, say for $\rho \in [0, \kappa e^\kappa]$, after optimizing for $z \in (0, \kappa)$, we obtain

$$\mathbb{P} \left(T_2^{\text{odd}}(u, v) > \sqrt{\frac{T H}{n A}} \left(\eta^{5/2} e^{\kappa/2} \sqrt{\rho} + \eta^5 \right) \right) \leq e^{-n \rho}.$$

Finally using an ϵ -net argument with $\epsilon = 1/4$ (Lemma 15), we get, for all $\kappa > 0$, $\rho \in (0, \kappa e^\kappa - 10)$,

$$\mathbb{P} \left(T_2^{\text{odd}} > 2 \sqrt{\frac{T H}{n |A|}} \left(\eta^{5/2} e^{\kappa/2} \sqrt{\rho + 10} + \eta^5 \right) \right) \leq e^{-n \rho}$$

and similarly

$$\mathbb{P}\left(T_2^{\text{even}} > 2\sqrt{\frac{TH}{n|A|}}\left(\eta^{5/2}e^{\kappa/2}\sqrt{\rho+10}+\eta^5\right)\right) \leq e^{-n\rho},$$

which finally, by union bound, gives

$$\mathbb{P}\left(T_2 > 4\sqrt{\frac{TH}{nA}}\left(\eta^{5/2}e^{\kappa/2}\sqrt{\rho+10}+\eta^5\right)\right) \leq 2e^{-n\rho}.$$

The final statement follows by choosing $\rho = 1$ and κ large enough. \square

Bounding the contribution of the heavy couples. The analysis of T_3 relies on the discrepancy property which is satisfied by the trimmed matrix. This property will be defined in the proof of Lemma 18 and remains crucial in order to obtain high probability bounds in the regime when $TH = \omega(n)$ and $TH = \mathcal{O}(n \log(n))$.

Lemma 17 will be required for the analysis of T_3 .

Lemma 17. *We have*

$$\mathbb{P}\left(\max_{x \in \Gamma_a} \hat{N}_a(x, \mathcal{X}) \leq \text{poly}(\eta) \frac{TH}{nA}\right) \geq 1 - 2 \exp\left(-\frac{TH}{nA}\right).$$

Proof of Lemma 17. The proof follows from an immediate application of Theorem 9. In fact, the proof is identical to that of Lemma 14. Therefore we refer the reader to that proof and omit it. \square

Let us now present the analysis of T_3 :

Lemma 18 (Bounding T_3). *For $TH = \Omega(n)$, we have*

$$\mathbb{P}\left(T_3 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}}\right) \geq 1 - \frac{2}{nA} - 2e^{-\frac{TH}{nA}}.$$

Proof of Lemma 18. Again, it will be convenient to write $T_3 \leq T_3^{\text{even}} + T_3^{\text{odd}}$, where

$$T_3^{\text{even}} = \sup_{u, v \in \mathbb{S}^{n-1}} \sum_{t=1}^T \sum_{h=1}^{\lfloor \frac{H-1}{2} \rfloor} \sum_{(x, y) \in \mathcal{H} \cap \mathcal{K}} u_x v_y \hat{N}_{a, t, 2h}(x, y),$$

$$T_3^{\text{odd}} = \sup_{u, v \in \mathbb{S}^{n-1}} \sum_{t=1}^T \sum_{h=0}^{\lfloor \frac{H}{2} \rfloor - 1} \sum_{(x, y) \in \mathcal{H} \cap \mathcal{K}} u_x v_y \hat{N}_{a, t, 2h+1}(x, y).$$

Then without loss of generality, we may focus on T_3^{odd} , as the analysis of T_3^{even} will follow similarly. In the sparse regime when we do not have enough observations, we cannot unfortunately use a standard argument that combines a uniform concentration bound with a net argument. Instead, we will use the so-called discrepancy property.

Discrepancy property. First, in order to declutter notations, we shall denote the random matrix $Q = \left(\sum_{t=1}^T \sum_{h=0}^{\lfloor H/2 \rfloor - 1} \hat{N}_{a, t, 2h+1}(x, y) \mathbb{1}\{x, y \in \Gamma_a\}\right)_{x, y \in \mathcal{X}}$, and for all $\mathcal{I}, \mathcal{J} \subseteq \mathcal{X}$ define the quantity $e(\mathcal{I}, \mathcal{J}) \triangleq \sum_{(x, y) \in \mathcal{I} \times \mathcal{J}} Q(x, y)$. Now, to obtain the desired result, we will follow a similar approach as that used in Keshavan et al. (2010b) which relies on showing that the matrix Q satisfies the so-called *discrepancy property*. We say that a random matrix Q satisfies the *discrepancy property*, if there exist $\xi_1, \xi_2 > 0$ such that for all $\mathcal{I}, \mathcal{J} \subseteq \mathcal{X}$, the following holds:

- (i) $\frac{e(\mathcal{I}, \mathcal{J}) n^2 A}{|\mathcal{I}| |\mathcal{J}| TH} \leq \xi_1$
- (ii) $e(\mathcal{I}, \mathcal{J}) \log\left(\frac{e(\mathcal{I}, \mathcal{J}) n^2 A}{|\mathcal{I}| |\mathcal{J}| TH}\right) \leq \xi_2 \max(|\mathcal{I}|, |\mathcal{J}|) \log\left(\frac{n}{\max(|\mathcal{I}|, |\mathcal{J}|)}\right)$

We know from Remark 4.5 in Keshavan et al. (2010b), that if the matrix Q satisfies the discrepancy property, then for some given ϵ -net \mathcal{N}_ϵ of the unit sphere \mathbb{S}^{n-1} , there exists an absolute constant $C > 0$ such that

$$\sup_{u,v \in \mathcal{N}_\epsilon} \sum_{(i,j) \in \mathcal{H}} u_x v_y Q(i,j) \leq \sqrt{\frac{TH}{nA}}$$

Now to conclude, it remains to show that Q satisfies the discrepancy property with high probability. Let $\mathcal{I}, \mathcal{J} \subseteq \mathcal{X}$, we may assume w.l.o.g that $|\mathcal{J}| \geq |\mathcal{I}|$. We distinguish between two cases:

→ *Case 1: if $|\mathcal{J}| \geq n/5$.* First, because of trimming we have the average bounded degree property which follows from Lemma 17

$$\mathbb{P} \left(\max_{x \in \Gamma_a} \hat{N}_a(x, \mathcal{X}) \leq \text{poly}(\eta) \frac{TH}{nA} \right) \geq 1 - 2 \exp \left(-\frac{TH}{nA} \right).$$

Thus, with probability at least $1 - 2 \exp \left(-\frac{TH}{nA} \right)$, we have

$$e(\mathcal{I}, \mathcal{J}) \leq |\mathcal{I}| \max_{x \in \Gamma_a} e(x, \mathcal{X}) \leq \frac{5C|\mathcal{I}||\mathcal{J}|TH}{n^2A}$$

which leads to

$$\frac{e(\mathcal{I}, \mathcal{J})n^2|\mathcal{A}|}{|\mathcal{I}||\mathcal{J}|} \leq 5C$$

whenever $TH = \Omega(n)$.

→ *Case 2: if $|\mathcal{J}| \leq n/5$.* We start by defining

$$\mu(\mathcal{I}, \mathcal{J}) \triangleq \mathbb{E} \left[\sum_{x \in \mathcal{I}, y \in \mathcal{J}} \sum_{t=1}^T \sum_{h=0}^{\lfloor H/2 \rfloor - 1} \hat{N}_{a,t,2h+1}(x, y) \right]$$

and note that in view of Assumptions 1-4, it can be easily verified that $\frac{TH|\mathcal{I}||\mathcal{J}|}{2\eta^5 n^2 A} \leq \mu(\mathcal{I}, \mathcal{J}) \leq \frac{\eta^5 TH|\mathcal{I}||\mathcal{J}|}{2n^2 A}$. Then letting $\kappa_\star(|\mathcal{I}|, |\mathcal{J}|) = \max\{\kappa_0, t^\star(\mathcal{I}, \mathcal{J})\}$ where κ_0 is chosen large enough, and $t^\star(\mathcal{I}, \mathcal{J})$ is defined as the constant t satisfying

$$t \log(t) = \frac{1}{\mu(\mathcal{I}, \mathcal{J})} \xi_2 |\mathcal{J}| \log \left(\frac{n}{|\mathcal{J}|} \right).$$

Next, we define the event

$$\mathcal{E} = \bigcap_{\mathcal{I}, \mathcal{J} \subset \mathcal{X}: |\mathcal{I}| \leq |\mathcal{J}| \leq n/5} \{e(\mathcal{I}, \mathcal{J}) \leq \kappa_\star(\mathcal{I}, \mathcal{J}) \mu(\mathcal{I}, \mathcal{J})\}.$$

We claim that whenever the event \mathcal{E} holds then Q satisfies either condition (i) or (ii) of the discrepancy property. Indeed, let $\mathcal{I}, \mathcal{J} \subseteq \mathcal{X}$, such that $|\mathcal{I}| \leq |\mathcal{J}| \leq n/5$, if $\kappa_\star(|\mathcal{I}|, |\mathcal{J}|) = \kappa_0$, then under \mathcal{E} , it holds that $e(\mathcal{I}, \mathcal{J}) \leq \kappa_0 \mu(\mathcal{I}, \mathcal{J}) \leq \frac{\kappa_0 \eta^5 TH |\mathcal{I}||\mathcal{J}|}{2n^2 A}$, which clearly means that property (i) is satisfied. On the other hand, if $\kappa_\star(|\mathcal{I}|, |\mathcal{J}|) = t^\star(\mathcal{I}, \mathcal{J})$, then under the event \mathcal{E} , we have $\frac{\eta^5 e(\mathcal{I}, \mathcal{J})}{2\mu(\mathcal{I}, \mathcal{J})} \log \left(\frac{\eta^5 e(\mathcal{I}, \mathcal{J})}{2\mu(\mathcal{I}, \mathcal{J})} \right) \leq \eta^5 t^\star(\mathcal{I}, \mathcal{J})/2 \log(\eta^5 t^\star(\mathcal{I}, \mathcal{J})/2) \leq \eta^5 t^\star(\mathcal{I}, \mathcal{J}) \log(t^\star(\mathcal{I}, \mathcal{J})) \leq \frac{\eta^5}{\mu(\mathcal{I}, \mathcal{J})} \xi_2 |\mathcal{J}| \log \left(\frac{n}{|\mathcal{J}|} \right)$ by monotonicity of $t \log(t)$ and when choosing κ_0 large enough so that $t_\star(\mathcal{I}, \mathcal{J}) \geq \eta^5/2$ ($\kappa_0 \geq \eta^5/2$). This implies that $e(\mathcal{I}, \mathcal{J}) \log \left(\frac{e(\mathcal{I}, \mathcal{J})n^2 A}{TH|\mathcal{I}||\mathcal{J}|} \right) \leq \eta^5 \xi_2 |\mathcal{J}| \log \left(\frac{n}{|\mathcal{J}|} \right)$. Thus, the property (ii) is satisfied in this case. It remains to show that the event \mathcal{E} holds with high probability. We have by union bound

$$\begin{aligned} \mathbb{P}(\mathcal{E}^c) &\leq \sum_{\mathcal{I}, \mathcal{J} \in \mathcal{X}: |\mathcal{I}| \leq |\mathcal{J}| \leq n/5} \mathbb{P}(e(\mathcal{I}, \mathcal{J}) > \kappa_\star(\mathcal{I}, \mathcal{J}) \mu(\mathcal{I}, \mathcal{J})) \\ &\stackrel{(a)}{\leq} \sum_{\mathcal{I}, \mathcal{J} \in \mathcal{X}: |\mathcal{I}| \leq |\mathcal{J}| \leq n/5} \exp \left(-\frac{1}{2} \kappa_\star(\mathcal{I}, \mathcal{J}) \log(\kappa_\star(\mathcal{I}, \mathcal{J})) \mu(\mathcal{I}, \mathcal{J}) \right) \\ &\stackrel{(b)}{\leq} \sum_{\mathcal{I}, \mathcal{J} \in \mathcal{X}: |\mathcal{I}| \leq |\mathcal{J}| \leq n/5} \exp \left(-\frac{1}{2} \xi_2 |\mathcal{J}| \log \left(\frac{n}{|\mathcal{J}|} \right) \right) \end{aligned}$$

$$\begin{aligned}
 &\leq \sum_{1 \leq i \leq j \leq n/5} 2 \binom{n}{i} \binom{n}{j} \exp(-\xi_2 j \log(n/j)) \\
 &\leq \sum_{1 \leq i \leq j \leq n/5} 2 \exp((4 - \xi_2) j \log(n/j))
 \end{aligned}$$

where in inequality (a), we applied Lemma 19, and in the inequality (b), we use the definition of $\kappa_*(\mathcal{I}, \mathcal{J})$ and finally chose $\xi_2 > 4$. In particular, choosing $\xi_2 = 7 + \log(A)/\log(n)$ ensures that

$$\mathbb{P}(\mathcal{E}^c) \leq \frac{1}{nA}.$$

We have just shown that

$$\mathbb{P}(Q \text{ satisfies the discrepancy property}) \geq 1 - \frac{1}{nA} - e^{-\frac{TH}{nA}}.$$

This implies that:

$$\mathbb{P}\left(T_3^{odd} \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}}\right) \geq 1 - \frac{1}{nA} - e^{-\frac{TH}{nA}}.$$

We can show the same for T_3^{even} . Therefore, we can conclude that for $n \gtrsim A$,

$$\mathbb{P}\left(T_3 \leq \text{poly}(\eta) \sqrt{\frac{TH}{nA}}\right) \geq 1 - \frac{2}{nA} - 2e^{-\frac{TH}{nA}}.$$

□

Lemma 19. *Let $k \geq e^2 \eta^8$, then*

$$\mathbb{P}(e(\mathcal{I}, \mathcal{J}) \geq k\mu(\mathcal{I}, \mathcal{J})) \leq \exp\left(-\frac{1}{2}k \log(k)\mu(\mathcal{I}, \mathcal{J})\right).$$

Proof of Lemma 19. We can easily verify see, from Assumptions 1-4, that $\frac{TH|\mathcal{I}||\mathcal{J}|}{2\eta^5 n^2 A} \leq \mu(\mathcal{I}, \mathcal{J}) \leq \frac{\eta^5 TH|\mathcal{I}||\mathcal{J}|}{2n^2 A}$. Next, we compute the moment-generating function of $e(\mathcal{I}, \mathcal{J})$. Let $\lambda > 0$, we have, via a peeling argument,

$$\begin{aligned}
 \mathbb{E}[\exp(\lambda e(\mathcal{I}, \mathcal{J}))] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \sum_{t=1}^T \sum_{h=0}^{\lfloor H/2 \rfloor - 1} \hat{N}_{a,t,2h+1}(x, y)\right)\right] \\
 &\leq \prod_{t=1}^T \prod_{h=1}^{\lfloor H/2 \rfloor - 1} \left(1 + \frac{\eta^2 |\mathcal{I}||\mathcal{J}|}{n^2 A} e^\lambda\right) \\
 &\leq \prod_{t=1}^T \prod_{h=1}^{\lfloor H/2 \rfloor - 1} \exp\left(\frac{\eta^2 |\mathcal{I}||\mathcal{J}|}{n^2 A} e^\lambda\right) \\
 &\leq \exp\left(\frac{\eta^2 TH|\mathcal{I}||\mathcal{J}|}{2n^2 A} e^\lambda\right).
 \end{aligned}$$

Now by Markov inequality, we have:

$$\begin{aligned}
 \mathbb{P}(e(\mathcal{I}, \mathcal{J}) \geq k\mu(\mathcal{I}, \mathcal{J})) &\leq \inf_{\lambda > 0} \mathbb{E}[\exp(\lambda(e(\mathcal{I}, \mathcal{J}) - k\mu(\mathcal{I}, \mathcal{J})))] \\
 &\leq \inf_{\lambda > 0} \exp\left(\frac{\eta^2 TH|\mathcal{I}||\mathcal{J}|}{2n^2 A} e^\lambda - \lambda k\mu(\mathcal{I}, \mathcal{J})\right) \\
 &= \exp\left(k\mu(\mathcal{I}, \mathcal{J}) \left(1 - \log\left(k\mu(\mathcal{I}, \mathcal{J}) \frac{2n^2 A}{\eta^2 TH|\mathcal{I}||\mathcal{J}|}\right)\right)\right)
 \end{aligned}$$

$$\begin{aligned} &\leq \exp\left(k\mu(\mathcal{I}, \mathcal{J})\left(1 - \log\left(\frac{k}{\eta^4}\right)\right)\right) \\ &\leq \exp\left(-\frac{1}{2}k \log(k)\mu(\mathcal{I}, \mathcal{J})\right) \end{aligned}$$

where in the last inequality we used $k \geq e^2\eta^8$. This concludes the proof of the claim. \square

G Proof of Theorem 3 (i) – Iterative Likelihood Improvement

The likelihood improvement steps are inspired by our lower bound. Specifically, the set of misclassified contexts is divided into two sets: the first set corresponds to "well-behaved" contexts where the empirical lower bound divergence is high, meaning that these contexts are likely to be classified accurately; the second set corresponds to the other contexts.

For the proof, we introduce the following notation: for $x \in \mathcal{X}$ and $j \in \mathcal{S}$,

$$\hat{I}_j(x; \Phi) \triangleq \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left(\hat{N}_a(f^{-1}(s), x) \log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} + \hat{N}_a(x, f^{-1}(s)) \log \frac{p(s|f(x), a)}{p(s|j, a)} \right), \quad (78)$$

where $\tilde{c}_j := \frac{\sum_{s, a} m_\rho(s, a) p(j|s, a)}{\sum_{s, a} m_\rho(s, a) p(f(x)|s, a)}$. It can be easily shown that $1/\eta \leq \tilde{c}_j \leq \eta$. Moreover, defining $p^{bwd}(s, a|j) := \frac{m_\rho(s, a) p(j|s, a)}{\sum_{\bar{s} \in \mathcal{S}} \sum_{\bar{a} \in \mathcal{A}} m_\rho(\bar{s}, \bar{a}) p(j|\bar{s}, \bar{a})}$, we have that $\frac{p^{bwd}(s, a|f(x))}{p^{bwd}(s, a|j)} = \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)}$.

Following Yun and Proutière (2016) and Sanders et al. (2020), we start by defining the subset of "well-behaved" contexts:

Definition 8 (Well-behaved contexts). *The set of well-behaved contexts $\mathcal{H} \subset \mathcal{X}$ is the largest subset of $\Gamma \triangleq \bigcup_{a \in \mathcal{A}} \Gamma_a$ with the following properties: for $x \in \mathcal{H}$,*

(H1) For all $j \neq f(x)$,

$$\hat{I}_j(x; \Phi) \geq \frac{1}{4\eta^2} \frac{TH}{n} I(x; \Phi), \quad (79)$$

(H2)

$$\sum_{a \in \mathcal{A}} \left\{ \hat{N}_a(x, \mathcal{X} \setminus \mathcal{H}) + \hat{N}_a(\mathcal{X} \setminus \mathcal{H}, x) \right\} \leq 2 \left(\log \frac{TH}{n} \right)^2. \quad (80)$$

Let $\mathcal{E}^{(\ell)}$ be the set of misclassified contexts after the ℓ -th iteration of Algorithm 2, and let $\mathcal{E}_{\mathcal{H}}^{(\ell)} := \mathcal{E}^{(\ell)} \cap \mathcal{H}$. The basic idea is to show that $\mathcal{E}_{\mathcal{H}}^{(\ell)}$ vanishes for $\ell = \lfloor \log(nA) \rfloor$, and then we obtain a worst-case upper bound for the error rate by simply setting all vertices in \mathcal{H}^c to be misclassified.

In Section G.1, we show the following:

Proposition 20. *If $I(\Phi) > 0$, then for some universal constant $C > 0$, we have, w.h.p.*

$$\left| \mathcal{E}_{\mathcal{H}^c}^{(\ell)} \right| \leq \left| \mathcal{H}^c \right| = \mathcal{O} \left(\sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right) \right). \quad (81)$$

In Section G.2, we show the following:

Proposition 21. *If $I(\Phi) > 0$, then w.h.p.,*

$$\left| \mathcal{E}_{\mathcal{H}}^{(\ell)} \right| = 0 \quad \text{when} \quad \ell = \log(nA). \quad (82)$$

From the above results and the fact that $|\mathcal{E}^{(\ell)}| = \left| \mathcal{E}_{\mathcal{H}^c}^{(\ell)} \right| + \left| \mathcal{E}_{\mathcal{H}}^{(\ell)} \right|$, we conclude the proof of Theorem 3.

G.1 Proof of Proposition 20 – Bounding $|\mathcal{H}^c|$

First, note that

$$\frac{|\Gamma^c|}{n} = \frac{\left| \bigcap_{a \in \mathcal{A}} \Gamma_a^c \right|}{n} \leq \frac{\min_{a \in \mathcal{A}} |\Gamma_a^c|}{n} = \exp \left(-\frac{TH}{nA} \right) \xrightarrow{n \rightarrow \infty} 0, \quad (83)$$

i.e., the number of contexts *not* in Γ is negligible, and thus we only need to worry about $\mathcal{H}^c \cap \Gamma$. From now on, all the contexts are assumed to be in Γ .

The proof consists of two parts.

The set of $x \in \mathcal{X}$ such that (H1) does not hold is bounded

Denote $\mathcal{H}_1 \triangleq \{x \in \mathcal{X} : \text{(H1) holds}\}$. To bound the cardinality of such set, we start by bounding $|\mathcal{H}_1^c|$, which follows from the following concentration result, whose proof is postponed to Section G.3.1:

Proposition 22 (Concentration regarding the rate function I). *For any $x \in \mathcal{X}$ and $j \in \mathcal{S}$,*

$$\mathbb{P} \left[\sum_{t=1}^T \sum_{h=1}^{H-1} \phi_j(\tilde{x}_h^{(t)}) < \frac{1}{4\eta^2} \frac{TH}{n} I(x; \Phi) \right] \leq 2 \exp \left(-2C \frac{TH}{n} I(x; \Phi) \right), \quad (84)$$

where $C > 0$ is an universal constant, $\tilde{x}_h^{(t)} := (x_h^{(t)}, a_h^{(t)}, x_{h+1}^{(t)})$, and

$$\begin{aligned} \phi_j(\tilde{x}_h^{(t)}) \triangleq & \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left(\mathbb{1} \left[x_h^{(t)} = x, a_h^{(t)} = a, x_{h+1}^{(t)} \in f^{-1}(s) \right] \log \frac{p(s|f(x), a)}{\tilde{c}_j p(s|j, a)} \right. \\ & \left. + \mathbb{1} \left[x_h^{(t)} \in f^{-1}(s), a_h^{(t)} = a, x_{h+1}^{(t)} = x \right] \log \frac{p(f(x)|s, a)}{p(j|s, a)} \right). \end{aligned}$$

Then by observing that $\hat{I}_j(x; \Phi) = \sum_{t=1}^T \sum_{h=1}^{H-1} \phi_j(\tilde{x}_h^{(t)})$, from Proposition 22,

$$\begin{aligned} \mathbb{E} \left[|\mathcal{H}_1^c| \right] &= \sum_{x \in \mathcal{X}} \mathbb{P}_\Phi \left[\exists j \neq f(x) \text{ s.t. } \hat{I}_j(x; \Phi) < \frac{1}{4\eta^2} \frac{TH}{n} I(x; \Phi) \right] \\ &\leq \sum_{x \in \mathcal{X}} \sum_{j \neq f(x)} \mathbb{P}_\Phi \left[\hat{I}_j(x; \Phi) < \frac{1}{4\eta^2} \frac{TH}{n} I(x; \Phi) \middle| x, j \right] \\ &\leq (S-1) \sum_{x \in \mathcal{X}} \exp \left(-2C \frac{TH}{n} I(x; \Phi) \right). \end{aligned}$$

We then conclude using Markov inequality:

$$\begin{aligned} \mathbb{P} \left[|\mathcal{H}_1^c| \geq \sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right) \right] &\leq \frac{\mathbb{E} \left[|\mathcal{H}_1^c| \right]}{\sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right)} \\ &\leq (S-1) \sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right) \rightarrow 0, \end{aligned}$$

where the \rightarrow holds in the limit $n \rightarrow \infty$ when $I(\Phi) > 0$ and at least $TH = \omega(n)$.

Final construction

Now consider the following iterative constructions of sets $\{Z(t)\}_{t \geq 0}$:

1. $Z(0) = \mathcal{H}_1^c$;
2. $Z(t) = Z(t-1) \cup \{z(t)\}$, where $z(t)$ does not satisfy (H2) w.r.t. $Z(t-1)$, i.e.,

$$\sum_{a \in \mathcal{A}} \left\{ \hat{N}_a(z(t), Z(t-1)) + \hat{N}_a(Z(t-1), z(t)) \right\} > 2 \left(\log \frac{TH}{n} \right)^2;$$

3. If such $z(t)$ does not exist, stop and let t^* be the total number of iterations. We lastly define $Z := Z(t^*)$.

First, observe that if $x \in Z^c$, then x satisfies (H1) and (H2), i.e., $x \in \mathcal{H}$. By the maximality of \mathcal{H} , $Z^c \subseteq \mathcal{H}$, which implies that, w.h.p.,

$$|\mathcal{H}^c| \leq |Z| \leq |\mathcal{H}_1^c| + t^* \leq \sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right) + t^*.$$

Thus it suffices to bound t^* . Define $s \triangleq \lceil 2 \sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right) \rceil$.

Obviously we have that $|Z(0)| \leq \frac{s}{2}$. We consider two cases:

- $s = 0$, i.e., $Z(0) = \emptyset$. Then, we have that for all $x \in \mathcal{X}$,

$$\sum_{a \in \mathcal{A}} \left\{ \hat{N}_a(x, \emptyset) + \hat{N}_a(\emptyset, x) \right\} = 0 \leq 2 \left(\log \frac{TH}{n} \right)^2.$$

This means that $t^* = 0$.

- $s \geq 1$. By construction, we have that $|Z(t)| \leq \frac{s}{2} + t$. We then bound $\hat{N}(Z(t), Z(t)) := \sum_{a \in \mathcal{A}} \hat{N}_a(Z(t), Z(t))$ as follows:

$$\begin{aligned} \hat{N}(Z(t), Z(t)) &= \hat{N}(z(t), Z(t-1)) + \hat{N}(Z(t-1), z(t)) + \hat{N}(Z(t-1), Z(t-1)) + \hat{N}(z(t), z(t)) \\ &> \hat{N}(Z(t-1), Z(t-1)) + 2 \left(\log \frac{TH}{n} \right)^2. \end{aligned}$$

Unfolding the recursion gives $\hat{H}(Z(t-1), Z(t-1)) > 2t \left(\log \frac{TH}{n} \right)^2$.

We now claim that $t^* < \frac{s}{2}$ with high probability. To show this, we proceed by contradiction and suppose that $t^* \geq \frac{s}{2}$. Then when $t = \frac{s}{2}$, we have that $|Z(s/2)| \leq s$ and $\hat{N}(Z(s/2), Z(s/2)) \geq s \left(\log \frac{TH}{n} \right)^2$. However, the following lemma, whose proof is presented in Section G.3.2, shows that this event does *not* happen with high probability:

Lemma 20. *Assume $s \geq 1$. Then, with high probability, no $W \subset \mathcal{X}$ with cardinality s satisfying $\hat{N}(W, W) \geq s \log \left(\frac{TH}{n} \right)^2$. Precisely,*

$$\mathbb{P} \left[\exists W \subset \mathcal{X} : |W| = s, \hat{N}(W, W) < s \left(\log \frac{TH}{n} \right)^2 \right] \leq 2 \exp \left(-\frac{1}{8} \frac{TH}{n} \log \frac{TH}{n} \right). \quad (85)$$

In summary, we have that $t^* < \frac{s}{2}$. We deduce that: w.h.p.,

$$|\mathcal{H}^{\mathcal{C}}| \leq \sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right) + \frac{s}{2} = 2 \sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right).$$

G.2 Proof of Proposition 21 – Bounding $|\mathcal{E}_{\mathcal{H}}^{(\ell)}|$

From the algorithm, we must have that

$$E \triangleq \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell)}} \left[\mathcal{L}^{(\ell)}(x, \hat{f}_{\ell+1}(x)) - \mathcal{L}^{(\ell)}(x, f(x)) \right] \geq 0, \quad (86)$$

where we recall that

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

with

$$\hat{p}_{\ell}(j|s, a) = \frac{\hat{N}_a(\hat{f}_{\ell}^{-1}(j), \hat{f}_{\ell}^{-1}(s))}{\hat{N}_a(\hat{f}_{\ell}^{-1}(j), \mathcal{X})}, \quad \hat{p}_{\ell}^{bwd}(s, a|j) = \frac{\hat{N}_a(\hat{f}_{\ell}^{-1}(s), \hat{f}_{\ell}^{-1}(j))}{\sum_{\bar{a} \in \mathcal{A}} \hat{N}_{\bar{a}}(\mathcal{X}, \hat{f}_{\ell}^{-1}(j))}.$$

We can decompose E as $E = E_1 + E_2 + U$, where

$$E_1 = \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \left\{ \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\hat{N}_a(x, f^{-1}(s)) \log \frac{p(s|\hat{f}_{\ell+1}(x), a)}{p(s|f(x), a)} \right. \right. \quad (87)$$

$$\left. \left. + \hat{N}_a(f^{-1}(s), x) \log \frac{p^{bwd}(s, a|\hat{f}_{\ell+1}(x))}{p^{bwd}(s, a|f(x))} \right] \right\}, \quad (88)$$

$$E_2 = \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \left\{ \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\left(\hat{N}_a(x, \hat{f}_\ell^{-1}(s)) - \hat{N}_a(x, f^{-1}(s)) \right) \log \frac{p(s|\hat{f}_{\ell+1}(x), a)}{p(s|f(x), a)} \right. \right. \\ \left. \left. + \left(\hat{N}_a(\hat{f}_\ell^{-1}(s), x) - \hat{N}_a(f^{-1}(s), x) \right) \log \frac{p^{bwd}(s, a|j)}{p^{bwd}(s, a|f(x))} \right] \right\}, \quad (89)$$

and $U = E - E_1 - E_2$.

Let us denote $e^{(\ell)} = |\mathcal{E}_{\mathcal{H}}^{(\ell)}|$. In Section G.4.1, we derive bounds on U, E_1, E_2 :

Lemma 21. *Assume that $I(\Phi) > 0$ and $TH = \omega(n)$. Then the following holds w.h.p.:*

$$-E_1 = \Omega \left(e^{(\ell+1)} \frac{TH}{n} \right). \quad (90)$$

$$U = \mathcal{O} \left(e^{(\ell+1)} SA \left(\frac{e^{(\ell)} TH}{n nA} + S \sqrt{\frac{TH}{nA}} \right) \right) \quad (91)$$

$$|E_2| \leq F_1 + F_2 + F_3, \quad (92)$$

where

$$F_1 = \mathcal{O} \left(\frac{TH}{n} \frac{e^{(\ell)}}{n} e^{(\ell+1)} \right), \quad F_2 = \mathcal{O} \left(\sqrt{e^{(\ell+1)} e^{(\ell)} \frac{THA}{n}} \right),$$

$$F_3 = \mathcal{O} \left(e^{(\ell+1)} \left(\log \frac{TH}{n} \right)^2 \right).$$

With the above lemma, we can now quantify the minimum number of iterations ℓ for $e^{(\ell)}$ to vanish:

Proposition 23. *If $I(\Phi) > 0$, $TH = \omega(n)$ and $\frac{e^{(1)}}{n} = o(1)$ w.h.p., then after $\ell \geq \log(nA)$ iterations of the likelihood improvement, $e^{(\ell)} = 0$ w.h.p..*

Proof. From $E \geq 0$ and $I(\Phi) > 0$, the following holds a.s.:

$$-E_1 \leq F_1 + F_2 + F_3 + |U|.$$

From Lemma 21, we have that w.h.p.

$$e^{(\ell+1)} \frac{TH}{n} = \mathcal{O} \left(\frac{TH}{n} \frac{e^{(\ell)}}{n} e^{(\ell+1)} + \sqrt{e^{(\ell+1)} e^{(\ell)} \frac{THA}{n}} + e^{(\ell+1)} \left(\log \frac{TH}{n} \right)^2 \right. \\ \left. + e^{(\ell+1)} SA \left(\frac{e^{(\ell)} TH}{n nA} + S \sqrt{\frac{TH}{nA}} \right) \right).$$

With the given assumptions, we have that w.h.p.

$$\frac{e^{(\ell+1)}}{e^{(\ell)}} \leq \mathcal{O} \left(\frac{nA}{TH} \right).$$

We can readily see that when $\ell = \lfloor \log(nA) \rfloor$, $e^{(\ell)} = 0$ w.h.p., and we are done. \square

G.3 Postponed proofs – Bounding $|\mathcal{H}_1^G|$

G.3.1 Proof of Proposition 22: concentration for the rate function I

Let $x \in \mathcal{X}, j \in \mathcal{S}$. We first compute $\|\phi\|_\infty$:

$$\begin{aligned} |\phi(\tilde{x}_h^{(t)})| &\leq \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left(\mathbb{1} [x_h^{(t)} = x, a_h^{(t)} = a, x_{h+1}^{(t)} \in f^{-1}(s)] \left| \log \frac{p(s|f(x), a)}{p(s|j, a)} \right| \right. \\ &\quad \left. + \mathbb{1} [x_h^{(t)} \in f^{-1}(s), a_h^{(t)} = a, x_{h+1}^{(t)} = x] \left| \log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} \right| \right) \\ &\leq 3 \log \eta. \end{aligned}$$

We now compute ϕ^2 in closed form:

$$\begin{aligned} &\phi(X, A, Y)^2 \\ &= \sum_{(a, s), (\tilde{a}, \tilde{s})} \left(\mathbb{1} [X = x, A = a, Y \in f^{-1}(s)] \log \frac{p(s|f(x), a)}{p(s|j, a)} \right. \\ &\quad \left. + \mathbb{1} [X \in f^{-1}(s), A = a, Y = x] \log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} \right) \\ &\quad \left(\mathbb{1} [X = x, A = \tilde{a}, Y \in f^{-1}(\tilde{s})] \log \frac{p(\tilde{s}|f(x), \tilde{a})}{p(\tilde{s}|j, \tilde{a})} \right. \\ &\quad \left. + \mathbb{1} [X \in f^{-1}(\tilde{s}), A = \tilde{a}, Y = x] \log \frac{p(f(x)|\tilde{s}, \tilde{a})}{\tilde{c}_j p(j|\tilde{s}, \tilde{a})} \right) \\ &= \sum_{a, s} \left[\mathbb{1} [X = x, A = a, Y \in f^{-1}(s)] \left(\log \frac{p(s|f(x), a)}{p(s|j, a)} \right)^2 \right. \\ &\quad \left. + \mathbb{1} [X \in f^{-1}(s), A = a, Y = x] \left(\log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} \right)^2 \right] \\ &\quad + 2 \underbrace{\sum_{a \in \mathcal{A}} \mathbb{1} [X = x, A = a, Y = x] \left(\log \frac{p(f(x)|f(x), a)}{p(f(x)|j, a)} \right) \left(\log \frac{p(f(x)|f(x), a)}{\tilde{c}_j p(j|f(x), a)} \right)}_{\triangleq G} \\ &\leq \sum_{a, s} \left[\mathbb{1} [X = x, A = a, Y \in f^{-1}(s)] \left(\log \frac{p(s|f(x), a)}{p(s|j, a)} \right)^2 \right. \\ &\quad \left. + \mathbb{1} [X \in f^{-1}(s), A = a, Y = x] \left(\log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} \right)^2 \right] + G. \end{aligned}$$

As $\text{Var}[\phi] \leq \mathbb{E}[\phi^2]$ (for any given probability measure), it suffices to derive an upper bound for $\mathbb{E}_\nu[\phi^2]$ for $\nu \in \{\mu_{\text{odd}}, \mu_{\text{even}}, P_2(\cdot|x', a', y')\}$, where we recall that $\mathbb{E}_\nu[\phi^2] = \mathbb{E}_{X \sim \nu}[\phi(X)^2]$. Observe that for any choice of ν , $\mathbb{E}_\mu[|G|] = \mathcal{O}(\frac{1}{n^2})$, possibly up to some factors involving S, A .

We first consider μ_{odd} . Recalling the definitions of $p^{\text{in}}, p^{\text{out}}, m_\rho$ (Section D.3.2),

$$\begin{aligned} &\mathbb{E}_{\mu_{\text{odd}}}[\phi^2] \\ &\leq \sum_{a, s} \left[q(x|f(x)) m_\rho(f(x), a) p(s|f(x), a) \left(\log \frac{p(s|f(x), a)}{p(s|j, a)} \right)^2 \right. \\ &\quad \left. + m_\rho(s, a) p(f(x)|s, a) q(x|f(x)) \left(\log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} \right)^2 \right] + \mathcal{O}\left(\frac{1}{n^2}\right) \\ &\stackrel{(i)}{\leq} 2 \left[q(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho(f(x), a) \left(\max_{s \in \mathcal{S}} \frac{p_{f(x), a}^{\text{out}}(s) \vee p_{j, a}^{\text{out}}(s)}{p_{f(x), a}^{\text{out}}(s) \wedge p_{j, a}^{\text{out}}(s)} \right)^2 \text{KL}(p_{f(x), a}^{\text{out}} \| p_{j, a}^{\text{out}}) \right. \end{aligned}$$

$$\begin{aligned}
 & + \left(\max_{\bar{s} \in [2K]} \frac{p_{\Phi,x}^{in}(\bar{s}, a) \vee p_{\Psi,x}^{in}(\bar{s}, a; \tilde{c}_j)}{p_{\Phi,x}^{in}(\bar{s}, a) \wedge p_{\Psi,x}^{in}(\bar{s}, a; \tilde{c}_j)} \right)^2 \text{KL}(p_{\Phi,x}^{in}(\cdot, \cdot) \| p_{\Psi,x}^{in}(\cdot, \cdot; \tilde{c}_j)) \Big] + \mathcal{O}\left(\frac{1}{n^2}\right) \\
 & \stackrel{(ii)}{\leq} 2\eta^2 \left[q(x|f(x)) \sum_{a \in \mathcal{A}} m_\rho(f(x), a) \text{KL}(p_{f(x),a}^{out} \| p_{j,a}^{out}) + \text{KL}(p_{\Phi,x}^{in}(\cdot, \cdot) \| p_{\Psi,x}^{in}(\cdot, \cdot; \tilde{c}_j)) \right] + \mathcal{O}\left(\frac{1}{n^2}\right) \\
 & \stackrel{(iii)}{\leq} \frac{2\eta^3}{n} \tilde{I}_j(x; \tilde{c}_j, \Phi) + \mathcal{O}\left(\frac{1}{n^2}\right) \stackrel{(iv)}{\leq} \frac{2\eta^4}{n} I_j(x; \tilde{c}_j, \Phi) + \mathcal{O}\left(\frac{1}{n^2}\right),
 \end{aligned}$$

where (i) follows from Lemma 22, (ii) and (iii) follow from the facts that p^{out} is η -regular and $1/\eta \leq \tilde{c}_j \leq \eta$, and (iv) follows from Proposition 11.

Similarly, we can bound $\mathbb{E}_{\mu_{even}}[\phi^2] \leq \frac{2\eta^6}{n} I_j(x; \tilde{c}_j, \Phi) + \mathcal{O}\left(\frac{1}{n^2}\right)$ and $\mathbb{E}_{P_2^2(\cdot|x',a',y')}[\phi^2] \leq \frac{2\eta^4}{n} I_j(x; \tilde{c}_j, \Phi) + \mathcal{O}\left(\frac{1}{n^2}\right)$.

Lemma 22 (Lemma 19 of SM6.3 of Sanders et al. (2020)). *When $\sum_{z \in \mathcal{Z}} p(z) = \sum_{z \in \mathcal{Z}} q(z) = 1$ and $\text{supp}(p) = \text{supp}(q) = \mathcal{Z}$, then the following holds:*

$$\sum_{z \in \mathcal{Z}} p(z) \left(\log \frac{p(z)}{q(z)} \right)^2 \leq 2 \left(\max_{z \in \mathcal{Z}} \frac{p(z) \vee q(z)}{p(z) \wedge q(z)} \right)^2 \text{KL}(p \| q).$$

In summary, we have:

$$\begin{aligned}
 M_{P,\phi} &= 3(2\eta^3 - 1) \log \eta \triangleq C_1, \\
 V_{\mu,P,\phi} &\leq \frac{C_2}{n} \tilde{I}_j(x; \tilde{c}_j, \Phi),
 \end{aligned}$$

with $C_2 \triangleq 2(1 + \sqrt{2}\eta^3(2\eta^3 - 1))^2 \eta^4$.

Recalling the definition of ϕ ,

$$\begin{aligned}
 & \mathbb{E}_{x_1^{(t)} \sim \mu} \left[\sum_{t,h} \phi(\tilde{X}_h^{(t)}) \right] \\
 &= \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left(N_a(x, f^{-1}(s)) \log \frac{p(s|f(x), a)}{p(s|j, a)} + N_a(f^{-1}(s), x) \log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} \right) \\
 &= \frac{TH}{n} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left(q(x|f(x)) m_\rho(f(x), a) p(s|f(x), a) \log \frac{p(s|f(x), a)}{p(s|j, a)} \right. \\
 & \quad \left. + m_\rho(s, a) p(f(x)|s, a) q(x|f(x)) \log \frac{p(f(x)|s, a)}{\tilde{c}_j p(j|s, a)} \right) \\
 & \stackrel{(i)}{\geq} \frac{TH}{\eta n} \tilde{I}_j(x; \tilde{c}_j, \Phi) (1 - o(1)) \\
 & \stackrel{(ii)}{\geq} \frac{TH}{2\eta^2 n} I_j(x; \tilde{c}_j, \Phi),
 \end{aligned}$$

where (i) follows from $\tilde{c}_j \eta \geq 1$ and the definition of \tilde{I} , and (ii) follows from Proposition 11.

Recalling that $I(x; \Phi) \leq I_j(x; c, \Phi)$ for all $j \in \mathcal{S}$ and $c > 0$, we conclude by applying our Bernstein-type concentration (Theorem 10):

$$\begin{aligned}
 & \mathbb{P} \left[\sum_{t,h} \phi(\tilde{X}_h^{(t)}) < \frac{1}{4\eta^2} \frac{TH}{n} I(x; \Phi) \right] \\
 & \leq \mathbb{P} \left[\sum_{t,h} \phi(\tilde{X}_h^{(t)}) < \frac{1}{4\eta^2} \frac{TH}{n} I_j(x; \tilde{c}_j, \Phi) \right]
 \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{P} \left[\sum_{t,h} \left(\phi \left(\tilde{X}_h^{(t)} \right) - \mathbb{E}_\mu \left[\phi \left(\tilde{X}_h^{(t)} \right) \right] \right) < -\frac{1}{4\eta^2} \frac{TH}{n} I_j(x; \tilde{c}_j, \Phi) \right] \\
 &\leq \mathbb{P} \left[\sum_{t,h} \left(\phi \left(\tilde{X}_{2h}^{(t)} \right) - \mathbb{E}_\mu \left[\phi \left(\tilde{X}_{2h}^{(t)} \right) \right] \right) < -\frac{1}{8\eta^2} \frac{TH}{n} I_j(x; \tilde{c}_j, \Phi) \right] \\
 &\quad + \mathbb{P} \left[\sum_{t,h} \left(\phi \left(\tilde{X}_{2h-1}^{(t)} \right) - \mathbb{E}_\mu \left[\phi \left(\tilde{X}_{2h-1}^{(t)} \right) \right] \right) < -\frac{1}{8\eta^2} \frac{TH}{n} I_j(x; \tilde{c}_j, \Phi) \right] \\
 &\leq 2 \exp \left(-\frac{\left(\frac{1}{8\eta^2} \frac{TH}{n} I_j(x; \tilde{c}_j, \Phi) \right)^2}{2TH \frac{C_2}{n} I_j(x; \tilde{c}_j, \Phi) + \frac{2}{3} C_1 \frac{1}{8\eta^2} \frac{TH}{n} I_j(x; \tilde{c}_j, \Phi)} \right) \\
 &= 2 \exp \left(-2C \frac{TH}{n} I_j(x; \tilde{c}_j, \Phi) \right) \leq 2 \exp \left(-2C \frac{TH}{n} I(x; \Phi) \right),
 \end{aligned}$$

where $C \triangleq \frac{1}{256\eta^4 C_2 + \frac{32}{3}\eta^2 C_1}$.

G.3.2 Proof of Lemma 20

Let $W \subset \mathcal{X}$ be any subset of size $s \geq 1$. As done previously, we split the summation into two parts:

$$\hat{N}(W, W) = \hat{N}^{even}(W, W) + \hat{N}^{odd}(W, W), \quad (93)$$

where $\tilde{H} \triangleq \lfloor H/2 \rfloor$, $I_{h,t} = \sum_{a \in \mathcal{A}} \mathbb{1} \left[X_h^{(t)} \in W, A_h^{(t)} = a, X_{h+1}^{(t)} \in W \right] = \mathbb{1} \left[X_h^{(t)} \in W, X_{h+1}^{(t)} \in W \right]$,

$$\hat{N}^{even}(W, W) \triangleq \sum_{t=1}^T \sum_{h=1}^{\tilde{H}} I_{2h,t}, \quad \hat{N}^{odd}(W, W) \triangleq \sum_{t=1}^T \sum_{h=1}^{\tilde{H}+1} I_{2h-1,t}. \quad (94)$$

Again, we exploit the conditional independency structure:

$$\begin{aligned}
 \mathbb{E} \left[e^{\theta \hat{N}^{even}(W, W)} \right] &= \mathbb{E} \left[\prod_{t=1}^T \prod_{h=1}^{\tilde{H}} e^{\theta I_{2h,t}} \right] \\
 &= \prod_{t=1}^T \mathbb{E} \left[\prod_{h=1}^{\tilde{H}-1} e^{\theta I_{2h,t}} \mathbb{E} \left[e^{\theta I_{2\tilde{H},t}} \mid \{I_{2h,t}\}_{h=1}^{\tilde{H}-1} \right] \right] \\
 &= \prod_{t=1}^T \mathbb{E} \left[\prod_{h=1}^{\tilde{H}-1} e^{\theta I_{2h,t}} \mathbb{E} \left[e^{\theta I_{2\tilde{H},t}} \mid \mathbb{1}[X_{2\tilde{H}-1} \in W] \right] \right] \\
 &= \prod_{t=1}^T \mathbb{E} \left[\prod_{h=1}^{\tilde{H}-1} e^{\theta I_{2h,t}} \left\{ \mathbb{E} \left[e^{\theta I_{2\tilde{H},t}} \mid X_{2\tilde{H}-1} \in W \right] \mathbb{1}[X_{2\tilde{H}-1} \in W] \right. \right. \\
 &\quad \left. \left. + \mathbb{E} \left[e^{\theta I_{2\tilde{H},t}} \mid X_{2\tilde{H}-1} \notin W \right] \mathbb{1}[X_{2\tilde{H}-1}^{(t)} \notin W] \right\} \right] \\
 &\leq \prod_{t=1}^T \mathbb{E} \left[\prod_{h=1}^{\tilde{H}-1} e^{\theta I_{2h,t}} \left\{ \mathbb{E} \left[e^{\theta I_{2\tilde{H},t}} \mid X_{2\tilde{H}-1} \in W \right] \mathbb{1}[X_{2\tilde{H}-1} \in W] + \mathbb{1}[X_{2\tilde{H}-1}^{(t)} \notin W] \right\} \right] \\
 &\stackrel{(*)}{\leq} \prod_{t=1}^T \mathbb{E} \left[\prod_{h=1}^{\tilde{H}-1} e^{\theta I_{2h,t}} \left(1 + e^{\theta \frac{s^2 \eta^6}{n^2}} \right) \right] \\
 &\leq \dots
 \end{aligned}$$

$$\begin{aligned} &\leq \left(1 + e^\theta \frac{s^2 \eta^6}{n^2}\right)^{T\tilde{H}} \\ &\leq \exp\left(e^\theta \frac{T\tilde{H}s^2\eta^6}{n^2}\right) \leq \exp\left(e^\theta \frac{THs^2\eta^6}{2n^2}\right), \end{aligned}$$

where (*) follows from

$$\begin{aligned} \mathbb{E}\left[e^{\theta I_{2\tilde{H},t}} \mid X_{2\tilde{H}-1} \in W\right] &= e^\theta \mathbb{P}\left[X_{2\tilde{H}} \in W, X_{2\tilde{H}+1} \in W \mid X_{2\tilde{H}-1} \in W\right] + 1 \\ &= 1 + e^\theta \frac{\sum_{x,y,z \in W} \mathbb{P}[X_{2\tilde{H}-1} = x, X_{2\tilde{H}} = y, X_{2\tilde{H}+1} = z]}{\sum_{x \in W} \mathbb{P}[X_{2\tilde{H}-1} = x]} \\ &= 1 + e^\theta \frac{\sum_{x,y,z \in W} \mathbb{P}[X_{2\tilde{H}-1} = x] \mathbb{P}[X_{2\tilde{H}} = y \mid X_{2\tilde{H}-1} = x] \mathbb{P}[X_{2\tilde{H}+1} = z \mid X_{2\tilde{H}} = y]}{\sum_{x \in W} \mathbb{P}[X_{2\tilde{H}-1} = x]} \\ &\leq 1 + e^\theta \frac{s^2 \left(\frac{\eta^3}{n}\right)^2 \sum_{x \in W} \mathbb{P}[X_{2\tilde{H}-1} = x]}{\sum_{x \in W} \mathbb{P}[X_{2\tilde{H}-1} = x]} = 1 + e^\theta \frac{s^2 \eta^6}{n^2}. \end{aligned}$$

By Markov inequality, we have:

$$\begin{aligned} \mathbb{P}\left[\hat{N}_a^{even}(W, W) \geq \frac{s}{2} \left(\log \frac{TH}{n}\right)^2\right] &\leq \inf_{\theta \geq 0} \frac{\mathbb{E}\left[\exp\left(\theta \hat{N}_a^{even}(W, W)\right)\right]}{\exp\left(\theta \frac{s}{2} \left(\log \frac{TH}{n}\right)^2\right)} \\ &\leq \inf_{\theta \geq 0} \exp\left(e^\theta \frac{THs^2\eta^6}{2n^2} - \theta \frac{s}{2} \left(\log \frac{TH}{n}\right)^2\right) \\ &\stackrel{(i)}{\leq} \exp\left(-\frac{TH}{n} s \left(\frac{1}{2} \log \frac{TH}{n} - e^{\frac{TH}{\log \frac{TH}{n}}} \frac{\eta^6 s}{2n}\right)\right) \\ &\stackrel{(ii)}{\leq} \exp\left(-\frac{s}{4} \frac{TH}{n} \log \frac{TH}{n}\right), \end{aligned}$$

where (i) follows from choosing $\theta = \frac{TH}{\log \frac{TH}{n}}$, and (ii) follows from a simple calculation:

$$\begin{aligned} e^{\frac{TH}{\log \frac{TH}{n}}} \frac{\eta^6 s}{2n} &\simeq \frac{1}{n} \sum_{x \in \mathcal{X}} \exp\left(\frac{TH}{\log \frac{TH}{n}} - C \frac{TH}{n} I(x; \Phi)\right) \\ &= \frac{1}{n} \sum_{x \in \mathcal{X}} \exp\left(-\frac{TH}{n} \left(CI(x; \Phi) - \frac{1}{\log \frac{TH}{n}}\right)\right) = o(1) \leq \frac{1}{4} \log \frac{TH}{n}, \end{aligned}$$

where we recall the definition of s and our assumptions that $\sum_{x \in \mathcal{X}} \exp(-C \frac{TH}{n} I(x; \Phi)) > 0$, and $TH = \omega(n)$.

Similarly, we also have that

$$\mathbb{P}\left[\hat{N}_a^{odd}(W, W) \geq \frac{s}{2} \log \left(\frac{TH}{n}\right)^2\right] \leq \exp\left(-\frac{s}{4} \frac{TH}{n} \log \frac{TH}{n}\right).$$

It follows from a simple union bound (combinatorial) argument that

$$\begin{aligned} &\mathbb{E}\left[\left|\left\{W : \hat{N}(W, W) \geq s \left(\log \frac{TH}{n}\right)^2, |W| = s\right\}\right|\right] \\ &\leq \mathbb{E}\left[\left|\left\{W : \hat{N}^{even}(W, W) \geq \frac{s}{2} \log \left(\frac{TH}{n}\right)^2, |W| = s\right\}\right|\right] \end{aligned}$$

$$\begin{aligned}
 & + \mathbb{E} \left[\left| \left\{ W : \hat{N}^{odd}(W, W) \geq \frac{s}{2} \log \left(\frac{TH}{n} \right)^2, |W| = s \right\} \right| \right] \\
 & \leq 2 \left(\frac{en}{s} \right)^s \exp \left(-\frac{s}{4} \frac{TH}{n} \log \frac{TH}{n} \right) \\
 & = 2 \exp \left(-\frac{s}{4} \left(\frac{TH}{n} \log \frac{TH}{n} - \log \frac{en}{s} \right) \right) \\
 & \stackrel{(i)}{\leq} 2 \exp \left(-\frac{s}{8} \frac{TH}{n} \log \frac{TH}{n} \right) \stackrel{(ii)}{\leq} 2 \exp \left(-\frac{1}{8} \frac{TH}{n} \log \frac{TH}{n} \right),
 \end{aligned}$$

where (i) follows from $\log \frac{en}{s} \leq \frac{1}{2} \frac{TH}{n} \log \frac{TH}{n}$, and (ii) follows from $s \geq 1$.

Finally, by Markov inequality, we conclude that

$$\mathbb{P} \left[\left| \left\{ W : \hat{N}(W, W) \geq s \left(\log \frac{TH}{n} \right)^2, |W| = s \right\} \right| \geq 1 \right] \leq 2 \exp \left(-\frac{1}{8} \frac{TH}{n} \log \frac{TH}{n} \right) \rightarrow 0.$$

□

G.4 Postponed proofs - Bounding $|\mathcal{E}_{\mathcal{H}}^{(\ell)}|$

G.4.1 Proof of Lemma 21: bounding E_1, E_2, U

We bound each term separately.

Proof of (90) - Lower bounding $-E_1$

The lower bound for $-E_1$ can be obtained by recalling that the form is asymptotically the same to that of the rate function:

$$\begin{aligned}
 -E_1 & = \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \left\{ \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\hat{N}_a(x, f^{-1}(s)) \log \frac{p(s|f(x), a)}{p(s|\hat{f}_{\ell+1}(x), a)} \right. \right. \\
 & \quad \left. \left. + \hat{N}_a(f^{-1}(s), x) \log \frac{p^{bwd}(f(x)|s, a)}{p^{bwd}(\hat{f}_{\ell+1}(x)|s, a)} \right] \right\} \\
 & = \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \hat{I}_{\hat{f}_{\ell+1}(x)}(x; \Phi) \\
 & \stackrel{(i)}{=} \Omega \left(\frac{TH}{n} \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} I(x; \Phi) \right) \stackrel{(ii)}{=} \Omega \left(e^{(\ell+1)} \frac{TH}{n} \right),
 \end{aligned}$$

where (i) follows from the condition (H1) for the well-defined contexts (Definition 8), and (ii) follows from our assumption that $I(\Phi) > 0$, which implies that $I(x; \Phi) > 0$ for all $x \in \mathcal{X}$ (see Section 3).

Proof of (91) - Upper bounding U

We again rewrite U as $U = U^{in} + U^{out}$, with

$$\begin{aligned}
 U^{in} & \triangleq \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \left\{ \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\hat{N}_a(x, \hat{f}_{\ell+1}^{-1}(s)) \left(\log \frac{\hat{p}_{\ell}(s|\hat{f}_{\ell+1}(x), a)}{\hat{p}_{\ell}(s|f(x), a)} - \log \frac{p(s|\hat{f}_{\ell+1}(x), a)}{p(s|f(x), a)} \right) \right] \right\} \\
 & = \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \left\{ \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\hat{N}_a(x, \hat{f}_{\ell+1}^{-1}(s)) \left(\log \frac{\hat{p}_{\ell}(s|\hat{f}_{\ell+1}(x), a)}{p(s|\hat{f}_{\ell+1}(x), a)} - \log \frac{\hat{p}_{\ell}(s|f(x), a)}{p(s|f(x), a)} \right) \right] \right\},
 \end{aligned}$$

and

$$U^{out} \triangleq \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \left\{ \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\hat{N}_a(x, \hat{f}_{\ell+1}^{-1}(s)) \left(\log \frac{\hat{p}_{\ell}^{bwd}(\hat{f}_{\ell+1}(x), a|s)}{\hat{p}_{\ell}^{bwd}(f(x), a|s)} - \log \frac{p^{bwd}(\hat{f}_{\ell+1}(x), a|s)}{p^{bwd}(f(x), a|s)} \right) \right] \right\}$$

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

We conclude by Lemma 25 and the triangle inequality: w.h.p.

$$\begin{aligned} |U| &\leq |U^{in}| + |U^{out}| \\ &\leq \mathcal{O} \left(e^{(\ell+1)} S \frac{TH}{n} \left(\frac{e^{(\ell)}}{n} + S \sqrt{\frac{nA}{TH}} \right) \right). \end{aligned}$$

Proof of (92) - Upper bounding E_2

From the regularity assumptions and the triangle inequality, we first have that

$$|E_2| \lesssim \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \left\{ \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left[\left| \hat{N}_a(x, \hat{f}_{\ell}^{-1}(s)) - \hat{N}_a(x, f^{-1}(s)) \right| + \left| \hat{N}_a(\hat{f}_{\ell}^{-1}(s), x) - \hat{N}_a(f^{-1}(s), x) \right| \right] \right\}.$$

We bound the first summation (the second summation follows the exact same argument): by the triangle inequality,

$$\begin{aligned} &\sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left| \hat{N}_a(x, \hat{f}_{\ell}^{-1}(s)) - \hat{N}_a(x, f^{-1}(s)) \right| \\ &\leq \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left| \hat{N}_a(x, \hat{f}_{\ell}^{-1}(s) \cap \mathcal{H}) - \hat{N}_a(x, f^{-1}(s) \cap \mathcal{H}) \right| \\ &\quad + \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left(\hat{N}_a(x, \hat{f}_{\ell}^{-1}(s) \cap \mathcal{H}^c) + \hat{N}_a(x, f^{-1}(s) \cap \mathcal{H}^c) \right) \\ &= \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left| \hat{N}_a(x, \hat{f}_{\ell}^{-1}(s) \cap \mathcal{H}) - \hat{N}_a(x, f^{-1}(s) \cap \mathcal{H}) \right| \\ &\quad + 2 \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \hat{N}_a(x, \mathcal{X} \setminus \mathcal{H}). \end{aligned}$$

From the construction of \mathcal{H} (specifically (H2)), the second sum can be further bounded as follows:

$$2 \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \hat{N}_a(x, \mathcal{X} \setminus \mathcal{H}) \leq 4e^{(\ell+1)} \left(\log \frac{TH}{n} \right)^2.$$

The first sum is bounded as follows:

$$\begin{aligned} &\sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left| \hat{N}_a(x, \hat{f}_{\ell}^{-1}(s) \cap \mathcal{H}) - \hat{N}_a(x, f^{-1}(s) \cap \mathcal{H}) \right| \\ &= \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left| \sum_{y \in (\hat{f}_{\ell}^{-1}(s) \cap \mathcal{H}) \setminus (f^{-1}(s) \cap \mathcal{H})} \hat{N}_a(x, y) - \sum_{y \in (f^{-1}(s) \cap \mathcal{H}) \setminus (\hat{f}_{\ell}^{-1}(s) \cap \mathcal{H})} \hat{N}_a(x, y) \right| \\ &\leq \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \sum_{y \in (\hat{f}_{\ell}^{-1}(s) \cap \mathcal{H}) \Delta (f^{-1}(s) \cap \mathcal{H})} \hat{N}_a(x, y) \\ &= 2 \sum_{x \in \mathcal{E}_{\mathcal{H}}^{(\ell+1)}} \sum_{a \in \mathcal{A}} \sum_{y \in \mathcal{E}_{\mathcal{H}}^{(\ell)}} \hat{N}_a(x, y) = 2 \sum_{a \in \mathcal{A}} \hat{N}_a \left(\mathcal{E}_{\mathcal{H}}^{(\ell+1)}, \mathcal{E}_{\mathcal{H}}^{(\ell)} \right). \end{aligned}$$

Next, we will use the following lemma related to the spectral norm of matrices:

Lemma 23 (Lemma 20 of SM6.4 of Sanders et al. (2020)). *For any matrix $B \in \mathbb{R}^{n \times n}$ and any subsets $E, F \subset [n]$, we have $\sum_{r \in E} \sum_{c \in F} B(r, c) = \mathbf{1}_E^\top B \mathbf{1}_F$, where $\mathbf{1}_E$ and $\mathbf{1}_F$ are column vector such that $\mathbf{1}_E(x) = \mathbb{1}[x \in E]$ for $x \in [n]$. Furthermore, we have that $\mathbf{1}_E^\top B \mathbf{1}_F \leq \|B\| \sqrt{|E||F|}$.*

Thus,

$$\begin{aligned} \sum_{a \in \mathcal{A}} \hat{N}_a \left(\mathcal{E}_{\mathcal{H}}^{(\ell+1)}, \mathcal{E}_{\mathcal{H}}^{(\ell)} \right) &= \sum_{a \in \mathcal{A}} N_a \left(\mathcal{E}_{\mathcal{H}}^{(\ell+1)}, \mathcal{E}_{\mathcal{H}}^{(\ell)} \right) + \sum_{a \in \mathcal{A}} (\hat{N}_a - N_a) \left(\mathcal{E}_{\mathcal{H}}^{(\ell+1)}, \mathcal{E}_{\mathcal{H}}^{(\ell)} \right) \\ &\leq \mathcal{O} \left(\frac{TH}{n^2} e^{(\ell+1)} e^{(\ell)} + \sum_{a \in \mathcal{A}} \left\| \hat{N}_a - N_a \right\|_2 \sqrt{e^{(\ell+1)} e^{(\ell)}} \right). \end{aligned}$$

From Proposition 19, we have that for each $a \in \mathcal{A}$, w.h.p.

$$\begin{aligned} \left\| \hat{N}_a - N_a \right\|_2 &\leq \left\| \hat{N}_a - \text{Trim}_{\Gamma_a} \left(\hat{N}_a \right) \right\|_2 + \left\| \text{Trim}_{\Gamma_a} \left(\hat{N}_a \right) - N_a \right\|_2 \\ &\leq \mathcal{O} \left(\left\| \hat{N}_a - \text{Trim}_{\Gamma_a} \left(\hat{N}_a \right) \right\|_F + \sqrt{\frac{TH}{nA}} \right) \\ &\leq \mathcal{O} \left(\sqrt{\left(\frac{TH}{n^2 A} \right)^2 n e^{-\frac{TH}{nA}} + \frac{TH}{nA}} \right) \leq \mathcal{O} \left(\sqrt{\frac{TH}{nA}} \right). \end{aligned}$$

In summary, we have, w.h.p.,

$$|E_2| = \mathcal{O} \left(\underbrace{\frac{TH}{n} \frac{e^{(\ell)}}{n} e^{(\ell+1)}}_{\triangleq F_1} + \underbrace{\sqrt{e^{(\ell+1)} e^{(\ell)} \frac{THA}{n}}}_{\triangleq F_2} + \underbrace{e^{(\ell+1)} \left(\log \frac{TH}{n} \right)^2}_{\triangleq F_3} \right).$$

□

G.4.2 Intermediate estimation errors for p and p^{bwd}

In this subsection, we bound the *estimation errors of p and p^{bwd}* during the improvement steps.

One important remark is that p and p^{bwd} can be *precisely* written as ratios of expected numbers of observations of transitions (even without the stationarity assumption):

Lemma 24. *For all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$p(s'|s, a) = \frac{N_a(f^{-1}(s), f^{-1}(s'))}{N_a(f^{-1}(s), \mathcal{X})}, \quad p^{bwd}(s, a|s') = \frac{N_a(f^{-1}(s), f^{-1}(s'))}{\sum_{\tilde{a} \in \mathcal{A}} N_{\tilde{a}}(\mathcal{X}, f^{-1}(s'))}. \quad (95)$$

Proof. Both follow from a simple chain of computations:

$$\begin{aligned} \frac{N_a(f^{-1}(s), f^{-1}(s'))}{N_a(f^{-1}(s), \mathcal{X})} &= \frac{TH \sum_{y \in f^{-1}(s')} m_\rho(s, a) p(s'|s, a) q(y|s')}{TH \sum_{z \in \mathcal{X}} m_\rho(s, a) p(f(z)|s, a) q(z|f(z))} \\ &= p(s'|s, a) \frac{1}{\sum_{\tilde{s} \in \mathcal{S}} p(\tilde{s}|s, a)} = p(s'|s, a). \end{aligned}$$

and

$$\begin{aligned} \frac{N_a(f^{-1}(s), f^{-1}(s'))}{\sum_{\tilde{a} \in \mathcal{A}} N_{\tilde{a}}(\mathcal{X}, f^{-1}(s'))} &= \frac{TH \sum_{y \in f^{-1}(s')} m_\rho(s, a) p(s'|s, a) q(y|s')}{TH \sum_{\tilde{s} \in \mathcal{S}} \sum_{\tilde{a} \in \mathcal{A}} m_\rho(\tilde{s}, \tilde{a}) p(s'|\tilde{s}, \tilde{a})} \\ &= \frac{m_\rho(s, a) p(s'|s, a)}{\sum_{\tilde{s} \in \mathcal{S}} \sum_{\tilde{a} \in \mathcal{A}} m_\rho(\tilde{s}, \tilde{a}) p(s'|\tilde{s}, \tilde{a})} = p^{bwd}(s, a|s'). \end{aligned}$$

□

Now the intermediate error bound for p :

Lemma 25. *After ℓ rounds of improvement, the following holds: if $e^{(\ell)}$ at least satisfies Theorem 2, then for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, w.h.p.*

$$\left| \log \frac{\hat{p}_\ell(s'|s, a)}{p(s'|s, a)} \right| \leq \left| \frac{\hat{p}_\ell(s'|s, a) - p(s'|s, a)}{p(s'|s, a)} \right| = \mathcal{O} \left(S \left(\frac{e^{(\ell)}}{n} + S \sqrt{\frac{nA}{TH}} \right) \right) \quad (96)$$

and

$$\left| \log \frac{\hat{p}_\ell^{bwd}(s'|s, a)}{p^{bwd}(s'|s, a)} \right| \leq \left| \frac{\hat{p}_\ell^{bwd}(s'|s, a) - p^{bwd}(s'|s, a)}{p^{bwd}(s'|s, a)} \right| = \mathcal{O} \left(S \left(\frac{e^{(\ell)}}{n} + S \sqrt{\frac{nA}{TH}} \right) \right). \quad (97)$$

Proof. We start with p_ℓ . From the inequalities $\frac{x}{1+x} \leq \log(1+x) \leq x$ for $x > -1$ and Lemma 24, we have that

$$\begin{aligned} \left| \log \frac{\hat{p}_\ell(s'|s, a)}{p(s'|s, a)} \right| &\leq \left| \frac{\hat{p}_\ell(s'|s, a) - p(s'|s, a)}{p(s'|s, a)} \right| \\ &= \left| \frac{N_a(f^{-1}(s), \mathcal{X})}{N_a(f^{-1}(s), f^{-1}(s'))} \frac{\hat{N}_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{\hat{N}_a(\hat{f}_\ell^{-1}(s), \mathcal{X})} - 1 \right| \\ &= \left| \underbrace{\frac{N_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{N_a(f^{-1}(s), f^{-1}(s'))} \frac{N_a(f^{-1}(s), \mathcal{X})}{N_a(\hat{f}_\ell^{-1}(s), \mathcal{X})}}_{\text{left ratios}} \underbrace{\frac{\hat{N}_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{N_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))} \frac{N_a(\hat{f}_\ell^{-1}(s), \mathcal{X})}{\hat{N}_a(\hat{f}_\ell^{-1}(s), \mathcal{X})}}_{\text{right ratios}} - 1 \right|. \end{aligned}$$

The "left ratios" capture the clustering error, both of which are concentrated around 1 with high probability. Denoting $V \triangleq f^{-1}(s)$, $\hat{V} \triangleq \hat{f}_\ell^{-1}(s)$, $W \triangleq f^{-1}(s')$, $\hat{W} \triangleq \hat{f}_\ell^{-1}(s')$, we first compute an upper bound for $|N_a(\hat{V}, \hat{W}) - N_a(V, W)|$:

$$\begin{aligned} |N_a(\hat{V}, \hat{W}) - N_a(V, W)| &= \left| \left(N_a(V \setminus \hat{V}, W) + N_a(V \cap \hat{V}, \hat{W} \setminus W) \right) \right. \\ &\quad \left. - \left(N_a(\hat{V} \setminus V, \hat{W}) + N_a(V \cap \hat{V}, W \setminus \hat{W}) \right) \right| \\ &\leq N_a(V \cap \hat{V}, W \Delta \hat{W}) + N_a(V \setminus \hat{V}, W) + N_a(\hat{V} \setminus V, \hat{W}) \\ &\leq N_a(V, \mathcal{E}^{(\ell)}) + N_a(\mathcal{E}^{(\ell)}, W) + N_a(\mathcal{E}^{(\ell)}, \hat{W}) \\ &\leq \mathcal{O} \left(\frac{TH}{n^2 A} \frac{n}{S} e^{(\ell)} \right) = \mathcal{O} \left(\frac{TH}{nSA} e^{(\ell)} \right), \end{aligned}$$

where Δ is the symmetric difference operator.

Now we compute the asymptotics of the first left ratio:

$$\begin{aligned} \left| \frac{N_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{N_a(f^{-1}(s), f^{-1}(s'))} - 1 \right| &= \left| \frac{1}{N_a(f^{-1}(s), f^{-1}(s'))} \left(N_a(\hat{V}, \hat{W}) - N_a(V, W) \right) \right| \\ &\leq \mathcal{O} \left(\frac{1}{\frac{n^2}{S^2} \frac{TH}{n^2 A}} \frac{TH}{nSA} e^{(\ell)} \right) = \mathcal{O} \left(\frac{S e^{(\ell)}}{n} \right). \end{aligned}$$

Observe that the same bound also holds for the other left ratio.

The "right ratios" are readily bounded using Proposition 24, provided at the end of this appendix: w.h.p.

$$\begin{aligned} \left| \frac{\hat{N}_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{N_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))} - 1 \right| &\leq \mathcal{O} \left(\frac{1}{\frac{n^2}{S^2} \frac{TH}{n^2 A}} \sqrt{\frac{nTH}{A}} \right) \\ &= \mathcal{O} \left(S^2 \sqrt{\frac{nA}{TH}} \right). \end{aligned}$$

Similarly, the same bound also holds for the other right ratio, and combining them all gives our result.

We now turn to p_ℓ^{bwd} . Using similar reasoning as previous, we first have that

$$\begin{aligned} \left| \log \frac{\hat{p}_\ell^{bwd}(s, a|s')}{p^{bwd}(s, a|s')} \right| &\leq \left| \frac{\hat{p}_\ell^{bwd}(s, a|s') - p^{bwd}(s, a|s')}{p^{bwd}(s, a|s')} \right| \\ &= \left| \frac{\sum_{\bar{a} \in \mathcal{A}} N_{\bar{a}}(\mathcal{X}, f^{-1}(s'))}{N_a(f^{-1}(s), f^{-1}(s'))} \frac{\hat{N}_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{\sum_{\bar{a} \in \mathcal{A}} \hat{N}_{\bar{a}}(\mathcal{X}, \hat{f}_\ell^{-1}(s'))} - 1 \right| \\ &= \left| \frac{N_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{N_a(f^{-1}(s), f^{-1}(s'))} \frac{\sum_{\bar{a} \in \mathcal{A}} N_{\bar{a}}(\mathcal{X}, f^{-1}(s'))}{\sum_{\bar{a} \in \mathcal{A}} N_{\bar{a}}(\mathcal{X}, \hat{f}_\ell^{-1}(s'))} \right. \\ &\quad \left. \times \frac{\hat{N}_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))}{N_a(\hat{f}_\ell^{-1}(s), \hat{f}_\ell^{-1}(s'))} \frac{\sum_{\bar{a} \in \mathcal{A}} N_{\bar{a}}(\mathcal{X}, \hat{f}_\ell^{-1}(s'))}{\sum_{\bar{a} \in \mathcal{A}} \hat{N}_{\bar{a}}(\mathcal{X}, \hat{f}_\ell^{-1}(s'))} - 1 \right|. \end{aligned}$$

All ratios in the above can be bounded as those involved in p_ℓ . This completes the proof. \square

Concentration of \hat{N}_a around N_a We now provide the concentration result relating N_a and \hat{N}_a , for any subsets $E, F \subset \mathcal{X}$, used in the discussions above:

Proposition 24 (Concentration of N_a and \hat{N}_a over all possible subsets). *There exists an absolute constant $c > 0$ such that for any $a \in \mathcal{A}$,*

$$\mathbb{P} \left[\max_{E, F \subset \mathcal{X}} \left| \hat{N}_a(E, F) - N_a(E, F) \right| \geq c \sqrt{\frac{nTH}{A}} \right] \leq 4 \exp(-2n(1 - \log 2)). \quad (98)$$

Proof. Let $E, F \subset \mathcal{X}$, and let $\phi(X, A, Y) = \mathbb{1}[X \in E, A = a, Y \in F]$. We have that $\hat{N}_a(E, F) = \sum_{t, h} \phi(X_{h-1}^{(t)}, A_{h-1}^{(t)}, X_h^{(t)})$ and $\|\phi\|_\infty = 1$. As for the proof of other concentration results for BMDPs (e.g. Proposition 22), we consider $MC_{2, odd}$ and $MC_{2, even}$. We first have that $\mathbb{E}_{\mu_{odd}}[\phi], \mathbb{E}_{\mu_{even}}[\phi], \mathbb{E}_{P_2^{(\cdot|x', a', y')}}[\phi] \leq \frac{\eta^5}{A} \triangleq p$ for all $(x', a', y') \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$. From this, the variances for all cases are bounded as follows:

$$\text{Var}[\phi] = \mathbb{E}[\phi] (1 - \mathbb{E}[\phi]) \leq \mathbb{E}[\phi] \leq p,$$

implying that $V_{\mu, P, \phi} \leq (1 + \sqrt{2}\eta^3(2\eta^3 - 1))^2 p$.

Thus for any $u = o(THp)$,

$$\begin{aligned} &\mathbb{P} \left[\left| \sum_{t, h} \phi(X_h^{(t)}) - \mathbb{E}_\mu[\phi(X_h^{(t)})] \right| > u \right] \\ &\leq \mathbb{P} \left[\left| \sum_{t, h} \phi(X_{2h}^{(t)}) - \mathbb{E}_\mu[\phi(X_{2h}^{(t)})] \right| > \frac{u}{2} \right] + \mathbb{P} \left[\left| \sum_{t, h} \phi(X_{2h+1}^{(t)}) - \mathbb{E}_\mu[\phi(X_{2h+1}^{(t)})] \right| > \frac{u}{2} \right] \\ &\leq 4 \exp \left(-\frac{u^2}{2TH(1 + \sqrt{2}\eta^3(2\eta^3 - 1))^2 p} \right). \end{aligned}$$

Choose $u = c\sqrt{\frac{nTH}{A}} = o(THp)$ with $c^2 \geq 4(1 + \sqrt{2}\eta^3(2\eta^3 - 1))^2 \eta^5$. Then,

$$\begin{aligned} \mathbb{P} \left[\left| \hat{N}_a(E, F) - N_a(E, F) \right| \geq c\sqrt{\frac{nTH}{A}} \right] &\leq 4 \exp \left(-\frac{c^2 \frac{nTH}{A}}{2TH(1 + \sqrt{2}\eta^3(2\eta^3 - 1))^2 \frac{\eta^5}{A}} \right) \\ &\leq 4 \exp(-2n). \end{aligned}$$

We conclude by taking the union bound over all possible pairs $(E, F) \subset \mathcal{X}$:

$$\mathbb{P} \left[\max_{E, F \subset \mathcal{X}} \left| \hat{N}_a(E, F) - N_a(E, F) \right| \geq c\sqrt{\frac{nTH}{A}} \right] \leq 2^{2n} 4 \exp(-2n) \leq 4 \exp(-2(1 - \log 2)n).$$

\square

H Proof of Theorem 3 (ii) – Estimation of the Latent Transitions and Emission Probabilities

In this appendix, we establish concentration results on the *final* estimation error of the latent state transitions rates and the emission probabilities.

H.1 Preliminaries on the estimators

Estimation under \hat{f} . Let us recall, that given an estimated decoding function \hat{f} , we estimate the latent state transition probabilities \hat{p} and the emission probabilities \hat{q} as follows:

$$\begin{aligned} \forall s, s' \in \mathcal{S}, \forall a \in \mathcal{A}, \quad \hat{p}(s'|s, a) &= \frac{\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) = s, a_h^{(t)} = a, \hat{f}(x_{h+1}^{(t)}) = s'\}}{\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) = s, a_h^{(t)} = a\}}, \\ \forall x \in \mathcal{X}, \forall s \in \mathcal{S}, \quad \hat{q}(x|s) &= \frac{\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) = s, x_h^{(t)} = x\}}{\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) = s\}}. \end{aligned}$$

Here, we use the short hand $\sum_{t,h} \triangleq \sum_{t=\lfloor T/2 \rfloor + 1}^T \sum_{h=1}^H$. Without explicitly mentioning it, we set $\hat{p}(s'|s, a) = 0$ (resp. $\hat{q}(x|s) = 0$) whenever $\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) = s, a_h^{(t)} = a\} = 0$ (resp. $\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) = s\} = 0$), and as we shall see this will not happen with high probability.

Estimation under the true f . We will denote the estimates of the latent transition probabilities and emission probabilities under the true clustering function f by \hat{p}_f and \hat{q}_f , respectively. They are defined as follows:

$$\begin{aligned} \forall s, s' \in \mathcal{S}, \forall a \in \mathcal{A} \quad \hat{p}_f(s'|s, a) &= \frac{\sum_{t,h} \mathbb{1}\{f(x_h^{(t)}) = s, a_h^{(t)} = a, f(x_{h+1}^{(t)}) = s'\}}{\sum_{t,h} \mathbb{1}\{f(x_h^{(t)}) = s, a_h^{(t)} = a\}}, \\ \forall x \in \mathcal{X}, \forall s \in \mathcal{S}, \quad \hat{q}_f(x|s) &= \frac{\sum_{t,h} \mathbb{1}\{f(x_h^{(t)}) = s, x_h^{(t)} = x\}}{\sum_{t,h} \mathbb{1}\{f(x_h^{(t)}) = s\}}. \end{aligned}$$

Again, without explicitly mentioning it, we set $\hat{p}_f(s'|s, a) = 0$ (resp. $\hat{q}_f(x|s) = 0$) whenever $\sum_{t,h} \mathbb{1}\{f(x_h^{(t)}) = s, a_h^{(t)} = a\} = 0$ (resp. $\sum_{t,h} \mathbb{1}\{f(x_h^{(t)}) = s\} = 0$), and as we shall see this will not happen with high probability.

Notations. To declutter notations, we introduce for all $t \in [T], h \in [H], \forall x \in \mathcal{X}, \forall X \subseteq \mathcal{X}, \forall a \in \mathcal{A}, \delta_{t,h,a}(x) \triangleq \mathbb{1}\{x_h^{(t)} = x, a_h^{(t)} = a\}, \delta_{t,h}(x) \triangleq \mathbb{1}\{x_h^{(t)} = x\}, \delta_{t,h,a}(X) \triangleq \mathbb{1}\{x_h^{(t)} \in X, a_h^{(t)} = a\}$ and $\delta_{t,h}(X) \triangleq \mathbb{1}\{x_h^{(t)} \in X\}$. We will further write $\hat{p}(s, a), p(s, a), \hat{q}(s), q(s)$ instead of $\hat{p}(\cdot|s, a), p(\cdot|s, a), \hat{q}(\cdot|s), q(\cdot|s)$

H.2 Proof of Theorem 3 (ii)

Here we present the precise statement of Theorem 3 - (ii), as Proposition 25.

Proposition 25. *Under Assumptions 1-4, the estimators \hat{p} and \hat{q} satisfy:*

(i) for all $TH = \Omega(\log(n))$, the event

$$\max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 \leq \text{poly}(\eta) SA \left(\sqrt{\frac{S + \log(nSA)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \quad (99)$$

holds with probability at least $1 - \frac{4}{n} - e^{-\frac{TH}{n}}$.

(ii) for all $TH = \Omega(n)$, the event

$$\max_{s \in \mathcal{S}} \|\hat{q}(s) - q(s)\|_1 \leq \text{poly}(\eta) S \left(\sqrt{\frac{n}{TH}} + \frac{|\mathcal{E}|}{n} \right) \quad (100)$$

holds with probability at least $1 - \frac{4}{n} - e^{-\frac{TH}{n}}$

Proof of Proposition 25. The proof is an immediate consequence of the estimation error decomposition Lemmas 26 and 27, and the concentration bounds provided in Lemma 28, Lemma 29, Lemma 30, and Lemma 31. \square

Proposition 26. *Under Assumptions 1-4, the estimator and \hat{q} satisfy: for all $V \in \mathbb{R}^n$, for all $\rho > 0$, for $TH \geq \text{poly}(\eta)S(\rho + \log(S))$,*

$$\mathbb{P} \left(\max_{s \in \mathcal{S}} |(\hat{q}(s) - q(s))V| \leq \text{poly}(\eta) \|V\|_\infty S \left(\sqrt{\frac{\rho + \log(S)}{TH}} + \frac{|\mathcal{E}|}{n} + \frac{\rho}{TH} \right) \right) \geq 1 - 4e^{-\rho}$$

Proof of Proposition 26. The proof is an immediate consequence of the estimation error decomposition Lemma 27, and the concentration bounds provided in Lemma 28, Lemma 29, and Lemma 31. \square

H.3 Estimation error decomposition lemmas

A key step in the proof of Theorem 3 - (ii) is to establish Lemmas 26 and 27. These lemmas allow us to obtain upper bounds on the estimation error of \hat{p} and \hat{q} that only depend on the estimation error of \hat{p}_f and \hat{q}_f and the total number of misclassified nodes $\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\}$ where here we use the slight abuse of notations that $\sum_{t,h} = \sum_{t=\lfloor T/2 \rfloor + 1}^T \sum_{h=1}^{H+1}$.

Here, we state Lemma 26 which will serve in the analysis of the estimation error of \hat{p} .

Lemma 26 (First Error Decomposition). *The estimator \hat{p} satisfies the following error decomposition: for all $s \in \mathcal{S}, a \in \mathcal{A}$, provided that $\hat{N}_a(f^{-1}(s)) \neq 0$, we have*

$$\|\hat{p}(s, a) - p(s, a)\|_1 \leq \|\hat{p}_f(s, a) - p_f(s, a)\|_1 + \frac{6 \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\}}{\hat{N}_a(f^{-1}(s))}. \quad (101)$$

Next, we state Lemma 27 which will serve to analyze the estimation error of \hat{q} .

Lemma 27 (Second Error Decomposition). *The estimator \hat{q} satisfies the following error decomposition: for all $s \in \mathcal{S}$, provided that $\hat{N}(f^{-1}(s)) \neq 0$, we have*

$$\|\hat{q}(s) - q(s)\|_1 \leq \|\hat{q}_f(s) - q_f(s)\|_1 + \frac{4 \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\}}{\hat{N}(f^{-1}(s))}. \quad (102)$$

Moreover, for any $V \in \mathbb{R}^n$, it holds

$$\left| (\hat{q}(s) - q(s))V \right| \leq \left| (\hat{q}_f(s) - q_f(s))V \right| + \frac{4 \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} \|V\|_\infty}{\hat{N}(f^{-1}(s))}. \quad (103)$$

H.4 Concentration bounds

Lemma 28. *Under Assumptions 1-4, we have for all $\rho > 0$,*

$$\mathbb{P} \left(\frac{1}{TH} \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} \leq \text{poly}(\eta) \left(\frac{|\mathcal{E}|}{n} + \frac{\rho}{TH} \right) \right) \geq 1 - e^{-\rho} \quad (104)$$

Consequently, we have:

$$\mathbb{P} \left(\frac{1}{TH} \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} \leq \text{poly}(\eta) \frac{|\mathcal{E}|}{n} \right) \geq 1 - e^{-\frac{\rho}{n}} \quad (105)$$

Lemma 29. *Under Assumptions 1-4, for all $\rho > 0$, we have:*

(i) for all $TH \geq \text{poly}(\eta)S(\rho + \log(S))$

$$\mathbb{P} \left(\min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s)) \geq \text{poly} \left(\frac{1}{\eta} \right) \frac{TH}{S} \right) \geq 1 - e^{-\rho}. \quad (106)$$

(ii) For all $TH \geq \text{poly}(\eta)SA(\rho + \log(SA))$, we obtain

$$\mathbb{P} \left(\min_{s \in \mathcal{S}, a \in \mathcal{A}} \hat{N}_a(f^{-1}(s)) \geq \text{poly} \left(\frac{1}{\eta} \right) \frac{TH}{SA} \right) \leq e^{-\rho}. \quad (107)$$

An immediate consequence of Lemma 29 is that for $TH = \Omega(\log(n))$, we have

$$\mathbb{P} \left(\min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s)) \geq \text{poly} \left(\frac{1}{\eta} \right) \frac{TH}{S} \right) \geq 1 - \frac{1}{n}$$

and

$$\mathbb{P} \left(\min_{s \in \mathcal{S}, a \in \mathcal{A}} \hat{N}_a(f^{-1}(s)) \geq \text{poly} \left(\frac{1}{\eta} \right) \frac{TH}{SA} \right) \geq 1 - \frac{1}{n}.$$

where $\Omega(\cdot)$ hides a dependence on $\text{poly}(\eta)SA \log(SA)$.

Lemma 30. Under Assumptions 1-4, for $TH = \Omega(\log(n))$, we have

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \left\| \hat{p}_f(s, a) - p(s, a) \right\|_1 \leq \text{poly}(\eta)SA \sqrt{\frac{S + \log(nSA)}{TH}} \right) \leq 1 - \frac{3}{n}. \quad (108)$$

Lemma 31. Under Assumptions 1-4, we obtain:

(i) for all $\rho > 0$, for all $V \in \mathbb{R}^n$, for all $TH \geq \text{poly}(\eta)S(\rho + \log(S))$,

$$\mathbb{P} \left(\max_{s \in \mathcal{S}} \left| (\hat{q}_f(s) - q(s))V \right| \leq \text{poly}(\eta) \|V\|_\infty S \sqrt{\frac{\rho + \log(S)}{TH}} \right) \geq 1 - 3e^{-\rho}. \quad (109)$$

(ii) For $TH = \Omega(n)$,

$$\mathbb{P} \left(\max_{s \in \mathcal{S}} \left\| \hat{q}_f(s) - q(s) \right\|_1 \leq \text{poly}(\eta)S \sqrt{\frac{n}{TH}} \right) \geq 1 - \frac{3}{n}. \quad (110)$$

H.5 Proofs – Estimation error decompositions

Proof of Lemma 26. Let $U \in \mathbb{R}^S$ such that $\|U\|_\infty \leq 1$. We wish to relate the estimation error of $|(\hat{p}(s, a) - p(s, a))U|$ to that $|(\hat{p}_f(s, a) - p(s, a))U|$ and the number of misclassified nodes $|\mathcal{E}|$. First, we start by writing

$$\hat{N}_a(f^{-1}(s))(\hat{p}(s, a) - p(s, a))U = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$$

where we define

$$\begin{aligned} \Delta_1 &= \left(\hat{N}_a(f^{-1}(s)) - \hat{N}_a(\hat{f}^{-1}(s)) \right) (\hat{p}(s, a) - p(s, a))U \\ \Delta_2 &= \sum_{s' \in \mathcal{S}} \left(\hat{N}_a(f^{-1}(s), f^{-1}(s')) - p(s'|s, a) \hat{N}_a(f^{-1}(s)) \right) U(s') \\ \Delta_3 &= \sum_{s' \in \mathcal{S}} \left(\left(\hat{N}_a(\hat{f}^{-1}(s), f^{-1}(s')) \right. \right. \\ &\quad \left. \left. - \hat{N}_a(f^{-1}(s), f^{-1}(s')) \right) - p(s'|s, a) \left(\hat{N}_a(\hat{f}^{-1}(s)) - \hat{N}_a(f^{-1}(s)) \right) \right) U(s') \\ \Delta_4 &= \sum_{s' \in \mathcal{S}} \left(\hat{N}_a(\hat{f}^{-1}(s), f^{-1}(s')) - \hat{N}_a(\hat{f}^{-1}(s), \hat{f}^{-1}(s')) \right) U(s'). \end{aligned}$$

Bounding Δ_1 . The term Δ_1 can be bounded as follows:

$$|\Delta_1| \leq \left| \sum_{t, h} \delta_{t, h, a}(\hat{f}^{-1}(s)) - \delta_{t, h, a}(f^{-1}(s)) \right| \|\hat{p}(s, a) - p(s, a)\|_1$$

$$\begin{aligned}
 &\leq 2 \sum_{t,h} \left| \delta_{t,h,a}(\hat{f}^{-1}(s)) - \delta_{t,h,a}(f^{-1}(s)) \right| \\
 &\leq 2 \sum_{t,h} \sum_{x \in \mathcal{X}} \mathbb{1}\{\hat{f}(x) \neq f(x)\} \mathbb{1}\{x_h^{(t)} = x\} \max \left(\mathbb{1}\{f(x) = s\}, \mathbb{1}\{\hat{f}(x) = s\} \right) \\
 &\leq 2 \sum_{t,h} \sum_{x \in \mathcal{X}} \mathbb{1}\{\hat{f}(x) \neq f(x)\} \mathbb{1}\{x_h^{(t)} = x\}
 \end{aligned}$$

where in the second inequality, we used the fact that $\|\hat{p}(s, a) - p(s, a)\|_1 \leq 2$.

Bounding Δ_2 . We observe that

$$\begin{aligned}
 |\Delta_2| &= \hat{N}_a(f^{-1}(s)) \left| \sum_{s' \in \mathcal{S}} \left(\hat{p}_f(s'|s, a) - p(s'|s, a) \right) U(s') \right| \\
 &\leq \hat{N}_a(f^{-1}(s)) \|\hat{p}_f(s, a) - p(s, a)\|_1.
 \end{aligned}$$

Bounding Δ_3 . We bound Δ_3 as follows

$$\begin{aligned}
 \Delta_3 &= \left| \sum_{s' \in \mathcal{S}} \sum_{t,h} \left(\delta_{t,h+1}(f^{-1}(s')) - p(s'|s, a) \right) \left(\delta_{t,h,a}(\hat{f}^{-1}(s)) - \delta_{t,h,a}(f^{-1}(s)) \right) U(s') \right| \\
 &\leq \sum_{s' \in \mathcal{S}} \sum_{t,h} \left| \left(\delta_{t,h+1}(f^{-1}(s')) - p(s'|s, a) \right) U(s') \right| \left| \delta_{t,h,a}(\hat{f}^{-1}(s)) - \delta_{t,h,a}(f^{-1}(s)) \right| \\
 &\leq \sum_{t,h} 2 \left| \delta_{t,h,a}(\hat{f}^{-1}(s)) - \delta_{t,h,a}(f^{-1}(s)) \right| \\
 &\leq \sum_{t,h} 2 \sum_{x \in \mathcal{X}} \mathbb{1}\{f(x) \neq \hat{f}(x)\} \mathbb{1}\{x_{t,h} = x\}.
 \end{aligned}$$

Bounding Δ_4 . We bound Δ_4 as follows

$$\begin{aligned}
 \Delta_4 &= \left| \sum_{t,h} \sum_{s' \in \mathcal{S}} \left(\delta_{t,h+1}(f^{-1}(s')) - \delta_{t,h+1}(\hat{f}^{-1}(s')) \right) \delta_{t,h,a}(\hat{f}^{-1}(s)) \right| \\
 &\leq \sum_{t,h} \sum_{s' \in \mathcal{S}} \left| \delta_{t,h+1}(f^{-1}(s')) - \delta_{t,h+1}(\hat{f}^{-1}(s')) \right| \delta_{t,h,a}(\hat{f}^{-1}(s)) \\
 &\leq \sum_{t,h} \sum_{s' \in \mathcal{S}} \left| \delta_{t,h+1}(f^{-1}(s')) - \delta_{t,h+1}(\hat{f}^{-1}(s')) \right| \\
 &\leq \sum_{t,h} \sum_{s' \in \mathcal{S}} \mathbb{1}\{\hat{f}(x_{h+1}^{(t)}) \neq f(x_{h+1}^{(t)})\} \max(\mathbb{1}\{f(x_{h+1}^{(t)}) = s'\}, \mathbb{1}\{\hat{f}(x_{h+1}^{(t)}) = s'\}) \\
 &\leq 2 \sum_{t,h} \mathbb{1}\{\hat{f}(x_{h+1}^{(t)}) \neq f(x_{h+1}^{(t)})\}.
 \end{aligned}$$

Finally, we conclude by writing

$$\|\hat{p}(s, a) - p(s, a)\|_1 \leq \|\hat{p}(s, a) - p(s, a)\|_1 + \frac{6 \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\}}{\hat{N}_a(f^{-1}(s))}$$

provided $\hat{N}_a(f^{-1}(s)) > 0$, where we slightly abuse notations and use $\sum_{t,h} = \sum_{t=\lfloor T/2 \rfloor + 1}^T \sum_{h=1}^{H+1}$. \square

Proof of Lemma 27. Let $V \in \mathbb{R}^n$, we start relating the estimation error of $|(\hat{q}(s) - q(s))V|$ to that of $|(\hat{q}_f(s) - q(s))V|$ and the misclassification error. Let us also recall that

$$\forall x \in \mathcal{X}, s \in \mathcal{S}, \quad \hat{q}(x|s) = \frac{\hat{N}(\{x\} \cap \hat{f}^{-1}(s))}{\hat{N}(\hat{f}^{-1}(s))} \quad \text{and} \quad \hat{q}_f(x|s) = \frac{\hat{N}(\{x\} \cap f^{-1}(s))}{\hat{N}(f^{-1}(s))}$$

whenever $\hat{N}(\hat{f}^{-1}(s)), \hat{N}(f^{-1}(s)) > 0$, otherwise the estimates are set to zero. Now, we have

$$\begin{aligned}
 \hat{N}(\hat{f}^{-1}(s)) \left| (\hat{q}(s) - q(s))V \right| &= \left| \sum_{x \in \mathcal{X}} \left(\hat{N}(\{x\} \cap \hat{f}^{-1}(s)) - q(x|s) \hat{N}(\hat{f}^{-1}(s)) \right) V(x) \right| \\
 &\leq \left| \sum_{x \in \mathcal{X}} \left(\hat{N}(\{x\} \cap f^{-1}(s)) - q(x|s) \hat{N}(f^{-1}(s)) \right) V(x) \right| \\
 &\quad + \left| \sum_{t,h} \sum_{x \in \mathcal{X}} (\delta_{t,h}(x) - q(x|s)) \left(\delta_{t,h}(\hat{f}^{-1}(s)) - \delta_{t,h}(f^{-1}(s)) \right) V(x) \right| \\
 &\leq \left| \sum_{x \in \mathcal{X}} \left(\hat{N}(\{x\} \cap f^{-1}(s)) - q(x|s) \hat{N}(f^{-1}(s)) \right) V(x) \right| \\
 &\quad + \left| \sum_{t,h} \left(\sum_{x \in \mathcal{X}} |\delta_{t,h}(x) - q(x|s)| \right) \left| \delta_{t,h}(\hat{f}^{-1}(s)) - \delta_{t,h}(f^{-1}(s)) \right| \|V\|_\infty \right| \\
 &\leq \left| \sum_{x \in \mathcal{X}} \left(\hat{N}(\{x\} \cap f^{-1}(s)) - q(x|s) \hat{N}(f^{-1}(s)) \right) V(x) \right| \\
 &\quad + \left| \sum_{t,h} \left(\sum_{x \in \mathcal{X}} |\delta_{t,h}(x) - q(x|s)| \right) \left(\sum_{x \in \mathcal{X}} \delta_{t,h}(x) \mathbb{1}\{\hat{f}(x) \neq f(x)\} \right) \|V\|_\infty \right| \\
 &\leq \left| \sum_{x \in \mathcal{X}} \left(\hat{N}(\{x\} \cap f^{-1}(s)) - q(x|s) \hat{N}(f^{-1}(s)) \right) V(x) \right| \\
 &\quad + 2 \sum_{x \in \mathcal{X}} \left(\sum_{t,h} \delta_{t,h}(x) \right) \mathbb{1}\{\hat{f}(x) \neq f(x)\} \|V\|_\infty
 \end{aligned}$$

where we used at the end we used the fact that $\|\delta_{t,h}(\cdot) - q(\cdot|s)\|_\infty \leq 1$. Thus, provided that $\hat{N}(\hat{f}^{-1}(s)) > 0$ and $\hat{N}(f^{-1}(s)) > 0$, we obtain

$$\hat{N}(\hat{f}^{-1}(s)) \left| (\hat{q}(s) - q(s))V \right| \leq \hat{N}(f^{-1}(s)) \left| (\hat{q}_f(s) - q(s))V \right| + 2|\mathcal{E}| \max_{x \in \mathcal{X}} \hat{N}(x) \|V\|_\infty$$

where we recall that $|\mathcal{E}| = \sum_{x \in \mathcal{X}} \mathbb{1}\{\hat{f}(x) \neq f(x)\}$. Furthermore,

$$\begin{aligned}
 \hat{N}(f^{-1}(s)) \left| (\hat{q}(s) - q(s))V \right| &\leq \hat{N}(f^{-1}(s)) \left| (\hat{q}_f(s) - q(s))V \right| + 2|\mathcal{E}| \max_{x \in \mathcal{X}} \hat{N}(x) \|V\|_\infty \\
 &\quad + \left(\hat{N}(\hat{f}^{-1}(s)) - \hat{N}(f^{-1}(s)) \right) \left| (\hat{q}(s) - q(s))V \right| \\
 &\leq \hat{N}(f^{-1}(s)) \left| (\hat{q}_f(s) - q(s))V \right| + 4|\mathcal{E}| \max_{x \in \mathcal{X}} \hat{N}(x) \|V\|_\infty
 \end{aligned}$$

where we used the fact $\|\hat{q}(s) - q(s)\|_1 \leq 2$, and $|\hat{N}(\hat{f}^{-1}(s)) - \hat{N}(f^{-1}(s))| \leq |\mathcal{E}| \max_{x \in \mathcal{X}} \hat{N}(x)$. Therefore, by taking the supremum over V such that $\|V\|_\infty \leq 1$, we finally obtain, provided $\hat{N}(f^{-1}(s)) > 0$, that

$$\|\hat{q}(s) - q(s)\|_1 \leq \|\hat{q}_f(s) - q(s)\|_1 + \frac{4|\mathcal{E}| \max_{x \in \mathcal{X}} \hat{N}(x)}{\hat{N}(f^{-1}(s))}.$$

□

H.6 Proofs – Concentration bounds

Proof of Lemma 28. In view of Proposition 2, we can easily verify that for all $t > \lfloor T/2 \rfloor, h \in [H]$,

$$\mathbb{E}_\mu[\mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} | \hat{f}] \leq \max_{x \in \mathcal{X}} \mathbb{E}_\mu[\mathbb{1}\{x_h^{(t)} = x\} | \hat{f}] |\mathcal{E}| \leq \max_{x \in \mathcal{X}} \mu^\top P_0^{h-1}(x) |\mathcal{E}| \leq \frac{\eta^2 |\mathcal{E}|}{n}.$$

Thus, $\mathbb{E}_\mu[\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} | \hat{f}] \leq \frac{\eta^2 TH |\mathcal{E}|}{2n}$. Now, we may apply Theorem 9 conditionally on \hat{f} , which gives for all $u > 0$ (possibly depending on \hat{f})

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1} \left\{ \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} > \frac{\eta^2 TH |\mathcal{E}|}{2n} + u \right\} \middle| \hat{f} \right] \\ & \leq \mathbb{E} \left[\mathbb{1} \left\{ \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} > \mathbb{E}_\mu \left[\sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} \middle| \hat{f} \right] + u \right\} \middle| \hat{f} \right] \\ & \leq \exp \left(-\frac{u^2}{8\eta^{10} \frac{TH |\mathcal{E}|}{n} + \frac{4}{3} \eta^2 u} \right) \\ & \leq \exp \left(-\min \left(\frac{nu^2}{8\eta^{10} TH |\mathcal{E}|}, \frac{u}{2\sqrt{2}\eta^5} \right) \right). \end{aligned}$$

Reparametrizing by $z = \frac{nu}{TH |\mathcal{E}|}$ gives

$$\begin{aligned} & \mathbb{E} \left[\mathbb{1} \left\{ \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} > \frac{\eta^2 TH |\mathcal{E}|}{2n} + \frac{TH |\mathcal{E}|}{n} z \right\} \middle| \hat{f} \right] \\ & \leq \exp \left(-\frac{TH |\mathcal{E}|}{n} \min \left(\frac{z^2}{8\eta^{10}}, \frac{z}{2\sqrt{2}\eta^5} \right) \right). \end{aligned}$$

Further reparametrizing $\rho = \frac{TH |\mathcal{E}|}{n} \min \left(\frac{z^2}{8\eta^{10}}, \frac{z}{2\sqrt{2}\eta^5} \right)$ gives

$$\mathbb{E} \left[\mathbb{1} \left\{ \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} > \frac{\eta^2 TH |\mathcal{E}|}{2n} + 2\sqrt{2}\eta^5 \max \left\{ \sqrt{\frac{TH |\mathcal{E}| \rho}{n}}, \rho \right\} \right\} \middle| \hat{f} \right] \leq e^{-\rho}.$$

Finally, noting that $\max \left\{ \sqrt{\frac{TH |\mathcal{E}| \rho}{n}}, \rho \right\} \leq \frac{TH |\mathcal{E}|}{n} + \rho$, we conclude that for all $\rho > 0$

$$\mathbb{P} \left(\frac{1}{TH} \sum_{t,h} \mathbb{1}\{\hat{f}(x_h^{(t)}) \neq f(x_h^{(t)})\} > \left(\frac{\eta^2}{2} + s\sqrt{2}\eta^5 \right) \frac{|\mathcal{E}|}{n} + 2\sqrt{2}\eta^5 \frac{\rho}{TH} \right) \leq e^{-\rho}.$$

□

Proof of Lemma 29. In view of Proposition 2, we know that P_0 is η^3 -regular. We can easily verify that for all $t > \lfloor T/2 \rfloor + 1, h \in [H]$,

$$\begin{aligned} \mathbb{E}_\mu[\mathbb{1}\{f(x_{t,h}) = s\}] & \geq \min_{x \in \mathcal{X}} \mathbb{E}_\mu[\mathbb{1}\{x_h^{(t)} = x\}] \alpha_s n = \min_{x \in \mathcal{X}} \mu^\top(P_0)^{h-1}(x) \alpha_s n \geq \frac{1}{\eta^4 S} \\ \mathbb{E}_\mu[\mathbb{1}\{f(x_{t,h}) = s\}] & \leq \max_{x \in \mathcal{X}} \mathbb{E}_\mu[\mathbb{1}\{x_h^{(t)} = x\}] \alpha_s n = \max_{x \in \mathcal{X}} \mu^\top(P_0)^{h-1}(x) \alpha_s n \leq \frac{\eta^4}{S}. \end{aligned}$$

Thus, we have $\frac{TH}{2\eta^4 S} \leq \mathbb{E}_\mu[\hat{N}(f^{-1}(s))] \leq \frac{TH\eta^4}{2S}$. Now, applying Theorem 9, we can immediately obtain that for all $u > 0$,

$$\begin{aligned} \mathbb{P} \left(\frac{TH}{2\eta^3 S} - u > \hat{N}(f^{-1}(s)) \right) & \leq \mathbb{P}(\mathbb{E}_\mu[\hat{N}(f^{-1}(s))] > \hat{N}(f^{-1}(s)) + u) \\ & \leq \exp \left(-\frac{u^2}{8\eta^{10} \frac{TH}{S} + \frac{4}{3} \eta^2 u} \right) \\ & \leq \exp \left(-\min \left(\frac{Su^2}{8\eta^{11} TH}, \frac{u}{2\sqrt{2}\eta^{11/2}} \right) \right). \end{aligned}$$

Using a union bound gives us

$$\mathbb{P}\left(\frac{TH}{2\eta^3 S} - u > \min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s))\right) \leq \exp\left(-\min\left(\frac{Su^2}{8\eta^{11}TH}, \frac{u}{2\sqrt{2}\eta^{11/2}}\right) + \log(S)\right).$$

Reparametrizing by $\rho = \min\left(\frac{Su^2}{8\eta^{11}TH}, \frac{u}{2\sqrt{2}\eta^{11/2}}\right) - \log(S)$, yields

$$\mathbb{P}\left(\frac{TH}{2\eta^3 S} - 2\sqrt{2}\eta^{11/2} \max\left\{\sqrt{\frac{TH}{S}}(\rho + \log(S)), \rho + \log(S)\right\} > \min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s))\right) \leq e^{-\rho}.$$

Thus, for all $TH \geq 8\eta^{17}S(\log(\rho) + \log(S))$, we obtain

$$\mathbb{P}\left(\frac{TH}{4\eta^3 S} > \min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s))\right) \leq e^{-\rho}.$$

Choosing $\rho = \log(n)$, we obtain that for $TH = \Omega(\log(n))$,

$$\mathbb{P}\left(\frac{TH}{4\eta^3 S} > \min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s))\right) \leq \frac{1}{n}$$

where $\Omega(\cdot)$ hides a dependence in $\text{poly}(\eta)S \log(S)$.

Following, a similar proof with the only exception that we use instead the fact P_1 is η^2 -regular, we obtain that for $TH = \Omega(\log(n))$,

$$\mathbb{P}\left(\frac{TH}{4\eta^3 SA} > \min_{s \in \mathcal{S}, a \in \mathcal{A}} \hat{N}_a(f^{-1}(s))\right) \leq \frac{1}{n}$$

where $\Omega(\cdot)$ hides a dependence in $\text{poly}(\eta)SA \log(SA)$. □

Lemma 32 (ϵ -net argument for ℓ_1 norm). *Let q be a d -dimensional random vector, and $\epsilon \in (0, 1)$. Furthermore, let \mathcal{N} be an ϵ -net of the unit ball with respect to $\|\cdot\|_\infty$, with minimal cardinality. Then, for all $\rho > 0$, we have*

$$\mathbb{P}\left(\|q\|_1 > \frac{\rho}{1-\epsilon}\right) \leq \left(\frac{3}{\epsilon}\right)^d \max_{V \in \mathcal{N}} \mathbb{P}(q^\top V > \rho).$$

Proof of Lemma 30. Let $U \in \mathbb{R}^{\mathcal{S}}$ such that $\|U\|_\infty \leq 1$. We have

$$\hat{N}_a(f^{-1}(s)) \left| (\hat{p}_f(s, a) - p(s, a))U \right| = \left| \sum_{t, h} \sum_{s' \in \mathcal{S}} (\delta_{t, h+1}(f^{-1}(s')) - p(s'|s, a))U(s') \delta_{t, h, a}(f^{-1}(s)) \right|.$$

We note by Hoeffding's lemma that for all $\lambda > 0$, we have

$$\mathbb{E}\left[\exp\left(\lambda\left(\sum_{s' \in \mathcal{S}} U(s')(\delta_{t, h+1}(f^{-1}(s')) - p(s'|s, a))\delta_{t, h, a}(f^{-1}(s))\right)\right)\middle|s_h^{(t)}\right] \leq \exp\left(\frac{\lambda^2}{2}\right).$$

Using the above inequality alongside a peeling argument, we obtain

$$\mathbb{E}\left[\exp\left(\lambda\left(\sum_{t, h} \sum_{s' \in \mathcal{S}} U(s')(\delta_{t, h+1}(f^{-1}(s')) - p(s'|s, a))\delta_{t, h, a}(f^{-1}(s))\right)\right)\right] \leq \exp\left(\frac{TH\lambda^2}{2}\right).$$

Now using Markov's inequality and optimizing over $\lambda > 0$, we obtain that for all $\rho > 0$

$$\mathbb{P}\left(\hat{N}_a(f^{-1}(s))(\hat{p}_f(s, a) - p(s, a))U > \rho\right) \leq \exp\left(-\frac{\rho^2}{2TH}\right). \quad (111)$$

This gives via a standard union bound

$$\mathbb{P} \left(\min_{s \in \mathcal{S}, a \in \mathcal{A}} \hat{N}_a(f^{-1}(s)) \max_{s \in \mathcal{S}, a \in \mathcal{A}} \left| (\hat{p}_f(s, a) - p(s, a))U \right| > \rho \right) \leq 2 \exp \left(-\frac{\rho^2}{2TH} + \log(SA) \right).$$

Reparametrizing by $\rho = \sqrt{TH(\rho' + \log(SA))}$, we may write for all $\rho' > 0$,

$$\mathbb{P} \left(\min_{s \in \mathcal{S}, a \in \mathcal{A}} \hat{N}_a(f^{-1}(s)) \max_{s \in \mathcal{S}, a \in \mathcal{A}} \left| (\hat{p}_f(s, a) - p(s, a))U \right| > \sqrt{TH(\rho' + \log(SA))} \right) \leq 2e^{-\rho'}.$$

Using Lemma 29, we obtain that for all $TH \geq \text{poly}(\eta)SA(\rho' + \log(SA))$,

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \left| (\hat{p}_f(s, a) - p(s, a))U \right| > \text{poly}(\eta)SA \sqrt{\frac{\rho' + \log(SA)}{TH}} \right) \leq 3e^{-\rho'}.$$

Now, we apply an ϵ -net argument using Lemma 32 with $\epsilon = 1/2$ to obtain

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \left\| \hat{p}_f(s, a) - p(s, a) \right\|_1 > \text{poly}(\eta)SA \sqrt{\frac{\rho' + \log(SA)}{TH}} \right) \leq 3e^{-\rho' + \log(6)S}.$$

Reparametrizing by $\rho'' = \rho' + \log(6)S$

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \left\| (\hat{p}_f(s, a) - p(s, a)) \right\|_1 > \text{poly}(\eta)SA \sqrt{\frac{\rho' + \log(SA) + S}{TH}} \right) \leq 3e^{-\rho''}.$$

Choosing $\rho'' = \log(n)$, we finally obtain for $TH = \Omega(\log(n))$

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \left\| \hat{p}_f(s, a) - p(s, a) \right\|_1 > \text{poly}(\eta)SA \sqrt{\frac{S + \log(nSA)}{TH}} \right) \leq \frac{3}{n}$$

where $\Omega(\cdot)$ hides a dependence of order $\text{poly}(\eta)SA \log(SA)$. □

Proof of Lemma 31. Let $V \in \mathbb{R}^n$ such that $\|V\|_\infty < \infty$.

Proof of (i). We have

$$\hat{N}(f^{-1}(s)) \left| (\hat{q}_f(s) - q(s))V \right| = \left| \sum_{t,h} \sum_{x \in \mathcal{X}} (\delta_{t,h}(x) - q(x|s)) \mathbf{1}\{s_{t,h} = s\} V(x) \right|.$$

We note by Hoeffding's lemma that for all $\lambda > 0$, we have

$$\mathbb{E} \left[\exp \left(\lambda \left(\sum_{x \in \mathcal{X}} (\delta_{t,h}(x) - q(x|s)) \mathbf{1}\{s_{t,h} = s\} V(x) \right) \right) \middle| s_{t,h} \right] \leq \exp \left(\frac{\lambda^2 \|V\|_\infty^2}{2} \right).$$

Using the above inequality alongside a peeling argument, we immediately obtain

$$\mathbb{E} \left[\exp \left(\lambda \left(\sum_{t,h} \sum_{x \in \mathcal{X}} (\delta_{t,h}(x) - q(x|s)) \mathbf{1}\{s_{t,h} = s\} V(x) \right) \right) \right] \leq \exp \left(\frac{TH \lambda^2 \|V\|_\infty^2}{2} \right).$$

Now using Markov's inequality and optimizing over $\lambda > 0$, we obtain that for all $\rho > 0$

$$\mathbb{P} \left(\hat{N}(f^{-1}(s)) (\hat{q}_f(s) - q(s))V > \rho \right) \leq \exp \left(-\frac{\rho^2}{2TH \|V\|_\infty^2} \right). \tag{112}$$

This gives via a standard union bound

$$\mathbb{P} \left(\min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s)) \max_{s \in \mathcal{S}} \left| (\hat{q}_f(s) - q(s))V \right| > \rho \right) \leq 2 \exp \left(-\frac{\rho^2}{2TH \|V\|_\infty^2} + \log(S) \right).$$

Reparametrizing by $\rho = \|V\|_\infty \sqrt{TH(\rho' + \log(S))}$, we may write for all $\rho' > 0$,

$$\mathbb{P} \left(\min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s)) \max_{s \in \mathcal{S}} \left| (\hat{q}_f(s) - q(s))V \right| > \|V\|_\infty \sqrt{TH(\rho' + \log(S))} \right) \leq 2e^{-\rho'}.$$

Using Lemma 29, we obtain that for all $TH \geq \text{poly}(\eta)S(\rho' + \log(S))$,

$$\mathbb{P} \left(\max_{s \in \mathcal{S}} \left| (\hat{q}_f(s) - q(s))V \right| > \text{poly}(\eta) \|V\|_\infty S \sqrt{\frac{\rho' + \log(S)}{TH}} \right) \leq 3e^{-\rho'}.$$

Proof of (ii) We start from the inequality (112) and apply an ϵ -net argument using Lemma 32 with $\epsilon = 1/2$ to obtain

$$\mathbb{P} \left(\hat{N}(f^{-1}(s)) \left\| \hat{q}_f(s) - q(s) \right\|_1 > 2\rho \right) \leq \exp \left(-\frac{\rho^2}{2TH} + \log(6)n \right).$$

Then, using a union bound we obtain

$$\mathbb{P} \left(\min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s)) \max_{s \in \mathcal{S}} \left\| \hat{q}_f(s) - q(s) \right\|_1 > 2\rho \right) \leq \exp \left(-\frac{\rho^2}{2TH\|V\|_\infty^2} + \log(6)n + \log(S) \right).$$

Reparametrizing by $\rho = \sqrt{2TH(\rho' + \log(6)n + \log(S))}$

$$\mathbb{P} \left(\min_{s \in \mathcal{S}} \hat{N}(f^{-1}(s)) \max_{s \in \mathcal{S}} \left\| \hat{q}_f(s) - q(s) \right\|_1 > 2\sqrt{2TH(\rho' + \log(6)n + \log(S))} \right) \leq e^{-\rho'}.$$

Choosing $\rho' = \log(n)$ and applying Lemma 29, we can obtain for $TH = \Omega(\log(n))$

$$\mathbb{P} \left(\max_{s \in \mathcal{S}} \left\| \hat{q}_f(s) - q(s) \right\|_1 > \text{poly}(\eta) \frac{S\sqrt{2TH(n + \log(n) + \log(S))}}{TH} \right) \leq \frac{3}{n}.$$

□

I Proofs of Theorems 4 and 5 – Lower Bounds for Reward-Free RL

In this appendix, we provide the proofs of the sample complexity lower bounds in the (offline) reward-free RL. On a high level, both rely on a specific construction of alternate BMDPs.

I.1 Proof of Theorem 4 – minimax setting

Recall that

$$\Lambda(\Phi) = \max_{v \in [-1, 1]^S} \frac{1}{S} \sum_{s=1}^S \max_{a_1, a_2} \sum_{s'=1}^S (p(s'|s, a_1) - p(s'|s, a_2)) v_{s'}.$$

In the following, we denote by v^* the S -dimensional vector achieving the maximum leading to $\Lambda(\Phi)$, and by $a_{1,s}^*$ and $a_{2,s}^*$ the two actions achieving this maximum for latent state s . Such choices exist as the domains $[-1, 1]^S$ and \mathcal{A} are both compact.

We start by identifying a necessary condition for $\sum_{x \in f^{-1}(s)} |q(x|s) - \hat{q}(x|s)| \leq \frac{\varepsilon}{\Lambda(\Phi)}$, where \hat{q} would be estimated from the data.

From Lemma D.6 in Jin et al. (2020), for a given transition vector $q(x|s)$, we can define $M_s = e^{\Omega(n/S)}$ transition vectors $\{q^{(i)}(x|s)\}_{1 \leq i \leq M_s}$ such that

- $|q(x|s) - q^{(i)}(x|s)| = \frac{8\varepsilon}{|f^{-1}(s)|\Lambda(\Phi)}$ for all $x \in f^{-1}(s)$,
- $\sum_{x \in f^{-1}(s)} |q^{(i)}(x|s) - q^{(j)}(x|s)| \geq \frac{2\varepsilon}{\Lambda(\Phi)}$.

Let $q^{(0)}(x|s) = q(x|s)$. From Lemma 5, for all $i, j \in \{0, 1, \dots, M_s\}$ such that $i \neq j$,

$$\text{KL}(q^{(i)}(\cdot|s), q^{(j)}(\cdot|s)) = \mathcal{O}(\varepsilon^2 / \Lambda(\Phi)^2).$$

Let $\mathbb{P}^{(0)}$ be the model defined by $q(\cdot|1), \dots, q(\cdot|S)$ and $\mathbb{P}^{(v)}$ with $v \in [M_1] \times \dots \times [M_S]$ be the model defined by $q^{(v_1)}(\cdot|1), \dots, q^{(v_S)}(\cdot|S)$. Let $N^\pi(s)$ be the expected number of visits to the s -th cluster under policy π , then

$$\text{KL}(\mathbb{P}^{(v)}, \mathbb{P}^{(v')}) = \sum_{s=1}^S N^\pi(s) \text{KL}(q^{(v_s)}(\cdot|s), q^{(v'_s)}(\cdot|s)) = \mathcal{O}(TH \frac{\varepsilon^2}{\Lambda(\Phi)^2}).$$

When $\frac{TH \frac{\varepsilon^2}{\Lambda(\Phi)^2}}{\sum_{s=1}^S \log M_s} = \mathcal{O}(1)$, from Theorem 2.5 in Tsybakov (2009),

$$\sum_{x \in f^{-1}(s)} |q(x|s) - \hat{q}(x|s)| \geq \frac{\varepsilon}{\Lambda(\Phi)} \quad \text{for all } s \quad \text{with probability } \frac{1}{2}.$$

We now show that we can design a reward function r such that $\frac{1}{H}(V^*(r) - V^{\hat{\pi}_r}(r)) > \varepsilon$ when $\sum_{x \in f^{-1}(s)} |q(x|s) - \hat{q}(x|s)| \geq \frac{\varepsilon}{\Lambda(\Phi)}$ for all s .

For this proof, let $V_h(x)$ be the value of x from step h and $\hat{V}_h(x)$ be the estimated value using \hat{q} .

We first design r_H as follows. Pick an action a_H that will be optimal in all context x for both model (p, q) and (p, \hat{q}) . Hence, $V_H(x) = \hat{V}_H(x) = r_H(x, a_H)$. Now we choose the reward function such that $r_H(x, a_H) = \frac{1+v_s^*}{2}$ if $q(x|s) \geq \hat{q}(x|s)$ and $r_H(x, a_H) = \frac{1-v_s^*}{2}$ if $q(x|s) < \hat{q}(x|s)$. Then,

$$\begin{aligned} & \frac{1}{S} \sum_{s=1}^S \sum_{s'=1}^S (p(s'|s, a_{1,s}^*) - p(s'|s, a_{2,s}^*)) \left(\sum_{x \in f^{-1}(s')} (q(x|s') - \hat{q}(x|s')) V_H(x) \right) \\ & \geq \frac{1}{S} \sum_{s=1}^S \sum_{s'=1}^S (p(s'|s, a_{1,s}^*) - p(s'|s, a_{2,s}^*)) \frac{\varepsilon}{\Lambda(\Phi)} \frac{v_s^*}{4} = \frac{\varepsilon}{4}. \end{aligned}$$

We then set r_{H-1} so that

- $r_{H-1}(x, a) = 0$ for all $a \notin \{a_{1,s}^*, a_{2,s}^*\}$ for all $x \in f^{-1}(s)$;
- $a_{1,s}^*$ is the optimal action of $x \in f^{-1}(s)$ while $a_{2,s}^*$ is the optimal action under \hat{q} at time $H - 1$, and so that

$$\begin{aligned} & r_{H-1}(x, a_{1,s}^*) + \sum_{s'=1}^S p(s'|s, a_{1,s}^*) \sum_{x' \in f^{-1}(s')} q(x'|s') V_H(x') \\ & - \left(r_{H-1}(x, a_{2,s}^*) + \sum_{s'=1}^S p(s'|s, a_{2,s}^*) \sum_{x' \in f^{-1}(s')} q(x'|s') V_H(x') \right) \\ & = \sum_{s'=1}^S (p(s'|s, a_{1,s}^*) - p(s'|s, a_{2,s}^*)) \frac{\varepsilon}{\Lambda(\Phi)} \frac{v_s^*}{8} \end{aligned}$$

Then, although

$$\begin{aligned} & \frac{1}{S} \sum_{s=1}^S \left(r_{H-1}(x, a_{1,s}^*) + \sum_{s'=1}^S p(s'|s, a_{1,s}^*) \sum_{x' \in f^{-1}(s')} q(x'|s') V_H(x') \right) \\ & - \frac{1}{S} \sum_{s=1}^S \left(r_{H-1}(x, a_{2,s}^*) + \sum_{s'=1}^S p(s'|s, a_{2,s}^*) \sum_{x' \in f^{-1}(s')} q(x'|s') V_H(x') \right) = \frac{\varepsilon}{8}, \end{aligned}$$

the error on \hat{q} makes the algorithm play $a_{2,s}^*$ instead of $a_{1,s}^*$ for all s , which induces $\Omega(\varepsilon)$ loss with respect to the optimal policy.

Analogously, at every h , we can design r_h so that $a_{1,s}^*$ is the optimal action of $x \in f^{-1}(s)$ while $a_{2,s}^*$ is the optimal action under \hat{q} at time h and loose ε value. Therefore, there exist r_1, \dots, r_H such that

$$V_h^*(r) - V_h^{\hat{\pi}_r}(r) = \Omega((H - h)\varepsilon), \quad \text{for all } h.$$

Therefore, $\max_r \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) \geq \varepsilon$, when $\sum_{x \in f^{-1}(s)} |q(x|s) - \hat{q}(x|s)| \geq \frac{\varepsilon}{\Lambda(\Phi)}$ for all s . \square

I.2 Proof of Theorem 5 – reward-specific setting

We first establish the first term of the lower bound derived in Theorem 5. Consider a block MDP model such that every latent state s has the same size $\frac{n}{S}$ and a unique optimal action a_s , defined as follows.

Rewards. We consider a simple reward model such that every context state in $s \in \{1, \dots, \lfloor S/2 \rfloor\}$ has $r_h(x, a) = 1$ for all a and h and every context state in $s \in \{\lfloor S/2 \rfloor + 1, \dots, S\}$ has $r_h(x, a) = 0$ for all a and h . Then, policies should visit $s \in \{1, \dots, \lfloor S/2 \rfloor\}$ as many as possible to maximize the value. Note that from the reward information, we know whether $f(x) \in \{1, \dots, \lfloor S/2 \rfloor\}$ or not. However, we do not have any prior knowledge from the reward about the exact membership among $\{1, \dots, \lfloor S/2 \rfloor\}$ or among $\{\lfloor S/2 \rfloor + 1, \dots, S\}$.

Transitions. The transition $\sum_{v \in \{1, \dots, \lfloor S/2 \rfloor\}} p(v|s, a_s) = 3/4$ and $\sum_{v \in \{1, \dots, \lfloor S/2 \rfloor\}} p(v|s, a) = 1/4$ for $a \neq a_s$ and $q(x|s) = \frac{S}{n}$ for all s and $x \in f^{-1}(s)$. One can easily check that p and q satisfy $I(x; \Phi) > 0$. We design the transition model so that playing $a \neq a_s$ causes $\frac{\lfloor S/2 \rfloor}{2S}$ expected loss at the next time slot.

Policy and clustering. From any given policy, we can simply design a clustering algorithm such that x is classified to s when x selects a_s with probability at least $1/2$ at a given time step h . From the clustering lower bound derived in Theorem 1, to have less than εn misclassified context states, it is necessary to collect $TH = \Omega(n \log(1/\varepsilon))$ samples. Therefore, when $TH = \mathcal{O}(n \log(1/\varepsilon))$, every policy has to play the best policy a_s with probability at most $1/2$ from at least εn context states, which makes

$$\frac{1}{H} (V^*(r) - V^{\hat{\pi}}(r)) = \Omega(\varepsilon).$$

To justify the second term of our lower bound derived in Theorem 5, we consider another Block MDP model. This model is similar to the previous model but with a slightly different transition kernel. The transitions are defined as follows:

$q(x|s) = \frac{\xi}{n}$ for all s and $x \in f^{-1}(s)$ and $\sum_{v \in \{1, \dots, \lfloor S/2 \rfloor\}} p(v|s, a_s) = 1/2 + \varepsilon$ while $\sum_{v \in \{1, \dots, \lfloor S/2 \rfloor\}} p(v|s, a) = 1/2$ for $a \neq a_s$. We also constraint p so that $|p(v|s, a) - p(v|s, a')| \leq \frac{4\varepsilon}{5}$ for all s, v , and a, a' . Therefore, the model is designed so that failing to identify a_s ends up with ε loss (since when a policy plays $a \neq a_s$, the policy loses ε at the next time slot) and

$$\text{KL}(p(\cdot|s, a), p(\cdot|s, a')) = \mathcal{O}(\varepsilon^2) \quad \text{for all } a \neq a',$$

which comes from Lemma 5.

We now find the necessary condition to correctly identify a_s . From Theorem 2.5 in [Tsybakov \(2009\)](#), every policy fails to find a_s with probability at least $1/2$ when $TH = \mathcal{O}(SA/\varepsilon^2)$. We thus have $\mathbb{E} \left[\frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) \right] \geq \varepsilon$ with $TH = \mathcal{O}(SA/\varepsilon^2)$. □

J Proofs of Theorems 6 and 7 – Upper Bounds for Reward-Free RL

In this appendix, we provide proof of the reward-free guarantees of our algorithms. We start by introducing concepts and notations extensively used in our proofs. Then we provide the proofs of Theorem 6 and 7 which are fairly similar but with subtle differences. On a high level, the first step of these proofs is to use a value difference lemma to decouple the estimation error \hat{p} and \hat{q} from the clustering error due to the estimation of \hat{f} . In the second step, we use the specific concentration results for each setting to control the estimation error of \hat{p} and \hat{q} . In the final step, we invoke our upper bound on the proportion of misclassified nodes, Theorem 3, to conclude.

J.1 Preliminaries and notations

Transitions and value functions. Under the BMDP $\Phi = (p, q, f)$, we will denote the transition probabilities from the rich observations by P where $P(y|x, a) = q(y, f(y))p(f(y)|f(x), a)$ for all $x, y \in \mathcal{X}$, $a \in \mathcal{A}$. Additionally, for a given reward r and for all $h \in [H]$, we define the value function, of a policy π , under BMDP Φ , at step h , by

$$V_h^\pi(x) = \mathbb{E}_\Phi \left[\sum_{k=h}^H r_k(x_k^\pi, a_k^\pi) \mid x_h = x \right]$$

where the dependence on r is omitted for simplicity. Observe that we may simply write $V^\pi(r) = \mathbb{E}_\mu[V_1^\pi(x_1)]$. Furthermore, we note that such value functions satisfy the following recursions:

$$\forall h \in [H], \forall x \in \mathcal{X}, \quad V_h^\pi(x) = r_h(x, \pi_h(x)) + P(x_h, \pi_h(x))V_{h+1}^\pi$$

where we use the convention $V_{H+1}^\pi = 0$ and assume implicitly that the policy π is deterministic for simplicity. Here, the notation $P(x_h, \pi_h(x))V_{h+1}^\pi$ means $\sum_{y \in \mathcal{X}} P(y|x_h, \pi_h(x))V_{h+1}^\pi(y)$.

Empirical BMDP. Through our estimation procedures we obtain the estimates \hat{p} , \hat{q} , and \hat{f} . With these, we will denote the empirical BMDP by $\hat{\Phi} = (\hat{p}, \hat{q}, \hat{f})$. The context transition probabilities under $\hat{\Phi}$ will be denoted by \hat{P} . Additionally, for a given reward r , the value function under policy π , will be denoted by $\hat{V}^\pi(r)$, and at step h , by \hat{V}_h^π for all $h \in [H]$.

True BMDP under inaccurate clustering. We will also have to use the BMDP $\tilde{\Phi} = (p, q, \hat{f})$ in our analysis. The context transition probabilities under $\tilde{\Phi}$ will be denoted by \tilde{P} . Additionally, for a given reward r , the value function under policy π , will be denoted by $\tilde{V}^\pi(r)$, and at step h , by \tilde{V}_h^π for all $h \in [H]$.

J.2 Proof of Theorem 6

Here we restate Theorem 6.

Theorem 6 (Minimax reward setting). *Consider a BMDP Φ satisfying Assumptions 1-4. Further assume that $TH = \omega(n)$ and $I(\Phi) > 0$. Then we have, w.h.p.,*

$$\sup_r \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) = \mathcal{O} \left(\sqrt{\frac{nS^2A^2 \log(SAH)}{TH}} \right).$$

Proof of Theorem 6. The proof is an immediate application of Theorem 3 - (i) and Proposition 27. Indeed, let us define the events

$$E_1 = \left\{ \frac{|\mathcal{E}|}{n} \leq \frac{1}{n} \sum_{x \in \mathcal{X}} \exp \left(-C' I(x; \Phi) \frac{TH}{n} \right) \right\},$$

$$E_2 = \left\{ \frac{1}{H} \sup_r V^*(r) - V^{\hat{\pi}_r}(r) \leq \text{poly}(\eta) \left(\sqrt{\frac{S^2A^2n \log(SA)}{TH}} + \frac{SA|\mathcal{E}|}{n} \right) \right\}.$$

Under the event $E_1 \cap E_2$, we have

$$\frac{1}{H} \sup_r V^*(r) - V^{\hat{\pi}_r}(r) \leq \text{poly}(\eta) \left(\sqrt{\frac{S^2A^2n \log(SA)}{TH}} + SAe^{-C' \frac{TH}{n} \min_{x \in \mathcal{X}} I(x; \Phi)} \right)$$

$$\begin{aligned} &\leq \text{poly}(\eta) \left(\sqrt{\frac{S^2 A^2 n \log(SA)}{TH}} + \frac{SA n}{C' \min_{x \in \mathcal{X}} I(x; \Phi) TH} \right) \\ &= \mathcal{O} \left(\sqrt{\frac{S^2 A^2 n \log(SA)}{TH}} \right) \end{aligned}$$

where we used the fact that $\exp(-C' \min_{x \in \mathcal{X}} I(x; \Phi) \frac{TH}{n}) \leq \frac{n}{C' \min_{x \in \mathcal{X}} I(x; \Phi) TH}$. Therefore, by union bound,

$$\begin{aligned} \mathbb{P} \left(\frac{1}{H} \sup_r V^*(r) - V^{\hat{\pi}_r}(r) = \mathcal{O} \left(\sqrt{\frac{S^2 A^2 n \log(SA)}{TH}} \right) \right) &\geq 1 - \mathbb{P}(E_1^c \cup E_2^c) \\ &\geq 1 - \mathbb{P}(E_1^c) - \mathbb{P}(E_2^c). \end{aligned}$$

Next, in view of Theorem 3 - (i) and Proposition 27, we know that $\mathbb{P}(E_1^c) \xrightarrow{n \rightarrow \infty} 0$ and $\mathbb{P}(E_2^c) \xrightarrow{n \rightarrow \infty} 0$ for $TH = \omega(n)$. This concludes the proof. \square

Proposition 27 (Minimax setting). *Under Assumptions 1-4, for all reward functions r , provided $TH = \Omega(n)$, we have:*

$$\mathbb{P} \left(\frac{1}{H} \sup_r V^*(r) - V^{\hat{\pi}_r}(r) \leq \text{poly}(\eta) \left(\sqrt{\frac{S^2 A^2 n \log(SA)}{TH}} + \frac{SA |\mathcal{E}|}{n} \right) \right) \geq 1 - \frac{14}{n} - e^{-\frac{TH}{n}}.$$

Proof of Proposition 27. We start by applying Lemma 33, which ensures the following decomposition

$$\frac{1}{H} \sup_r V^*(r) - V^{\hat{\pi}_r}(r) \leq 2\eta^2 \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 + \max_{s \in \mathcal{S}} \|\hat{q}(s) - q(s)\|_1 \right) + \frac{4\eta^4 |\mathcal{E}|}{n}.$$

Then, we have from Proposition 25 that for all $TH = \Omega(n)$,

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 \leq \text{poly}(\eta) SA \left(\sqrt{\frac{S + \log(nSA)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) \geq 1 - \frac{4}{n} - e^{-\frac{TH}{n}}$$

and

$$\mathbb{P} \left(\max_{s \in \mathcal{S}} \|\hat{q}(s) - q(s)\|_1 \leq \text{poly}(\eta) S \left(\sqrt{\frac{n}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) \geq 1 - \frac{4}{n} - e^{-\frac{TH}{n}}.$$

Finally, combining the above three inequalities yields the result. \square

J.3 Proof of Theorem 7

Here we restate Theorem 7 in a more precise way.

Theorem 7 (Reward-specific setting). *Let C be the constant introduced in Theorem 3(i). Under the assumptions of Theorem 6, we have for any reward function r , w.h.p.*

$$\begin{aligned} \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) &= \mathcal{O} \left(\sqrt{\frac{S^3 A^2 H \log(SAHn)}{T}} \right. \\ &\quad \left. + \frac{SH^2}{n} \sum_{x \in \mathcal{X}} \exp \left(-C \frac{TH}{n} I(x; \Phi) \right) \right). \end{aligned}$$

Proof of Theorem 7. The proof is an immediate application of Theorem 3 - (i) and Proposition 28. Indeed, let us define the events

$$E_1 = \left\{ \frac{|\mathcal{E}|}{n} \leq \frac{1}{n} \sum_{x \in \mathcal{X}} \exp \left(-C' \frac{TH}{n} I(x; \Phi) \right) \right\},$$

$$E_2 = \left\{ \frac{1}{H} V^*(r) - V^{\hat{\pi}_r}(r) \leq \text{poly}(\eta) \left(\sqrt{\frac{S^3 A H^4 \log(SAHn)}{TH}} + \frac{SH^2 |\mathcal{E}|}{n} \right) \right\}.$$

Proving (i). We note that under the event $E_1 \cap E_2$, we have

$$\frac{1}{H} V^*(r) - V^{\hat{\pi}_r}(r) \leq \text{poly}(\eta) \left(\sqrt{\frac{S^3 A^2 \log(SAHn)}{TH}} + \frac{SAH^2 \sum_{x \in \mathcal{X}} \exp(-C' \frac{TH}{n} I(x; \Phi))}{n} \right).$$

Next, in view of Theorem 3 - (i) and Proposition 28, we know that $\mathbb{P}(E_1^c) \xrightarrow[n \rightarrow \infty]{} 0$ and $\mathbb{P}(E_2^c) \xrightarrow[n \rightarrow \infty]{} 0$ for $TH = \omega(n)$. This completes the proof of (i).

Proving (ii). We start by noting that when $TH - \frac{n \log(n)}{C' I(x; \Phi)} = \omega(1)$ for all $x \in \mathcal{X}$, then we have $|\mathcal{E}| < 1$ which simply implies that $|\mathcal{E}| = 0$ (i.e., we recover the clusters exactly). Thus, under $E_1 \cap E_2$, we have in this case,

$$\frac{1}{H} V^*(r) - V^{\hat{\pi}_r}(r) = \mathcal{O} \left(\sqrt{\frac{S^3 A^2 \log(SAHn)}{TH}} \right).$$

And again, in view of Theorem 3 - (i) and Proposition 28, we know that $\mathbb{P}(E_1^c) \xrightarrow[n \rightarrow \infty]{} 0$ and $\mathbb{P}(E_2^c) \xrightarrow[n \rightarrow \infty]{} 0$ when $TH - \frac{n \log(n)}{C' I(x; \Phi)} = \omega(1)$ for all $x \in \mathcal{X}$. This concludes the proof of (ii). \square

Proposition 28 (Reward-specific setting). *Under Assumptions 1-4, we have for all reward functions r , provided $TH = \Omega(n)$,*

$$\mathbb{P} \left(\frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) \leq \text{poly}(\eta) \left(\sqrt{\frac{S^3 A^2 H^4 \log(SAHn)}{TH}} + \frac{SH^2 |\mathcal{E}|}{n} \right) \right) \geq 1 - \frac{14}{n} - e^{-\frac{TH}{n}}.$$

Proof of Proposition 28. We start by applying Lemma 34, which ensures the following decomposition

$$\begin{aligned} \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) &\leq \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \hat{V}_{h+1}^{\hat{\pi}_r} \right| + 2H \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 \\ &\quad + \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \tilde{V}_{h+1}^{\pi^*} \right| + \frac{6\eta^2 H |\mathcal{E}|}{n}. \end{aligned}$$

From Lemma 38, we have the following concentration bounds that hold as long as $TH = \Omega(n)$,

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, h \in [H]} \left| (\hat{q}(s) - q(s)) \hat{V}_h^{\hat{\pi}_r} \right| > \text{poly}(\eta) SH \left(\sqrt{\frac{SA \log(SAHn)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) \leq \frac{4}{n}$$

and

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, h \in [H]} \left| (\hat{q}(s) - q(s)) \tilde{V}_h^{\pi^*} \right| > \text{poly}(\eta) SH \left(\sqrt{\frac{SA \log(SAHn)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) \leq \frac{4}{n}.$$

Additionally from Proposition 25 we also have for $TH = \Omega(\log(n))$

$$\mathbb{P} \left(\|\hat{p}(s, a) - p(s, a)\|_1 > \text{poly}(\eta) SA \left(\sqrt{\frac{S + \log(nSA)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) > \frac{6}{n} + e^{-\frac{TH}{n}}.$$

The final result follows from combining the above four inequalities. \square

J.4 Value difference lemmas

A crucial step towards obtaining our reward-free guarantees is to establish value difference lemmas. These lemmas must account for the clustering error and whether we are in the minimax setting or reward-specific setting. We state and prove such lemmas (Lemma 33 and Lemma 34).

Lemma 33 (First Value Difference Lemma). *For any reward function r , let $\hat{\pi}_r$ be the optimal policy for the empirical model $\hat{\Phi}$. Under Assumptions 1-4 for the true BMDP Φ , we have:*

$$\frac{1}{H} \sup_r V^*(r) - V^{\hat{\pi}_r}(r) \leq 2\eta^2 \left(\max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 + \max_{s \in \mathcal{S}} \|\hat{q}(s) - q(s)\|_1 \right) + \frac{4\eta^4 |\mathcal{E}|}{n}.$$

Lemma 34 (Second Value Difference Lemma). *Let r be some reward function and let $\hat{\pi}_r$ be the optimal policy for the empirical model $\hat{\Phi}$. Under Assumptions 1-4 for the true BMDP Φ , we have:*

$$\begin{aligned} \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) &\leq \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \hat{V}_{h+1}^{\hat{\pi}_r} \right| + 2H \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 \\ &\quad + \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \tilde{V}_{h+1}^{\pi^*} \right| + \frac{6\eta^2 H |\mathcal{E}|}{n}. \end{aligned}$$

Lemma 33 (resp. Lemma 34) is used to establish the reward-free guarantee in the *minimax* (resp. *reward-specific*) setting. We note that Lemma 34 has a worse dependence in H , than that of Lemma 33. This is due to the fact that $\hat{\Phi}$ does not necessarily satisfy Assumptions 1-4 which are essential to get such improvement in H . Therefore, it may appear strange to use Lemma 34. It is however useful in order to obtain an improvement of order n in the *reward-specific* setting in contrast with the *minimax* one, as will be apparent in our concentration bounds. The proofs of these lemmas rely on Lemma 35, Lemma 36, and Lemma 37.

Lemma 35. *Under Assumptions 1-4 for the true BMDP Φ , for all rewards r , and any policy π , it holds that*

$$|V^\pi(r) - \hat{V}^\pi(r)| \leq H \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \hat{V}_{h+1}^\pi \right| + \frac{\eta^2 H |\mathcal{E}|}{n}.$$

Alternatively, we also have

$$|V^\pi(r) - \hat{V}^\pi(r)| \leq H \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^\pi \right| + \frac{\eta^2 H |\mathcal{E}|}{n}.$$

Lemma 36. *Under Assumption 1-2, for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, we have*

$$\|\tilde{P}(x, a) - \hat{P}(x, a)\|_1 \leq \frac{2\eta^2 |\mathcal{E}|}{n} + \max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1 + \max_{s \in \mathcal{S}} \|\hat{q}(s) - q(s)\|_1.$$

More precisely, for any $V \in \mathbb{R}^n$, we have

$$\left| (\tilde{P}(x, a) - \hat{P}(x, a)) V \right| \leq \left(\frac{2\eta^2 |\mathcal{E}|}{n} + \max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1 \right) \|V\|_\infty + \max_{s \in \mathcal{S}} \left| (\hat{q}(s) - q(s)) V \right|.$$

In Lemma 37, we establish that the centered value function under the regularity assumption satisfies an upper bound that is horizon free. This has to do with the fact that all the transition matrices have mixing times that are uniformly bounded by η^2 . In fact, this result is easily generalizable to MDPs with finite state-action spaces that are communicating and aperiodic.

Lemma 37. *Under Assumptions 1-4 for the true BMDP Φ . Then, for all $h \in [H]$, the value function of a policy π at step h , V_h^π satisfies, for all $x \in \mathcal{X}$, $a \in \mathcal{A}$*

$$\left\| V_{h+1}^\pi - P(x, a) V_{h+1}^\pi \mathbf{1} \right\|_\infty \leq (2\eta^2 - 1).$$

J.5 Concentration bounds on averaged optimal value functions

The following lemma is a key technical result to establish a guarantee for the *reward-free* setting which does not suffer a linear dependence in n .

Lemma 38. *Assuming that \hat{f} is estimated using the first $\lfloor TH/2 \rfloor$ observations, and that \hat{p}, \hat{q} are estimated using the observations $\lfloor TH/2 \rfloor$ with \hat{f} . Then, under Assumptions 1-4, provided $TH = \Omega(n)$, we have*

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, h \in [H]} \left| (\hat{q}(s) - q(s)) \hat{V}_h^{\hat{\pi}_r} \right| > \text{poly}(\eta) SH \left(\sqrt{\frac{SA \log(HSn)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) \leq \frac{4}{n}$$

and

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, h \in [H]} \left| (\hat{q}(s) - q(s)) \tilde{V}_h^{\pi^*} \right| > \text{poly}(\eta) SH \left(\sqrt{\frac{SA \log(HSn)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) \leq \frac{4}{n}.$$

The challenge in obtaining the first concentration bound comes from the fact that $\hat{q}(s)$ and $\hat{V}^{\hat{\pi}_r}$ are dependent on each other. Fortunately, $\hat{V}^{\hat{\pi}_r}$ has a special structure that can be characterized; see Lemma 39. Lemma 39 says that actually, the optimal value function under the BMDP model possesses a linear structure. This observation stems from the remark that the transition matrices in the BMDP are low rank (see Zhang et al. (2022b)). This low-rank structure can then be leveraged to ensure that, actually the value function of any policy satisfies the representation stated in the Lemma 39 (See for example, Proposition 2.3. in Jin et al. (2020) or Lemma 1 in Modi et al. (2021)).

Lemma 39. *Let Φ be a BMDP, then, for all $h \in [H]$, there exists $\theta_h \in \mathbb{R}^{SA}$, such that $\|\theta_h\|_2 \leq \sqrt{SA}(H-h)$ such that the value function can be expressed as follows:*

$$\forall x \in \mathcal{X}, \quad V_h^*(x) = \max_{a \in \mathcal{A}} \{r_h(x, a) + \psi_f(x, a)^\top \theta_h\}$$

where for all $x \in \mathcal{X}, a \in \mathcal{A}$, $\psi_f(x, a)$ is an SA -dimensional column vector in $\{0, 1\}^{SA}$ such that $\psi_f(x, a)(s, b) = \mathbb{1}\{(f(x), a) = (s, b)\}$ (f is the latent state decoding function of Φ).

Now using the representation of the optimal value function stated in Lemma 39, we consider, for a given decoding function f , the set of all possible optimal value functions at step h as we vary the latent transitions p and emission probabilities q . More precisely, we define a set that contains all such functions as follows:

$$\mathcal{V}_h^*(f; r) := \left\{ V : \exists \theta_h : \|\theta_h\|_2 \leq M, \forall x \in \mathcal{X}, V(x) = \min \left\{ \max_{a \in \mathcal{A}} \{r_h(x, a) + \psi_f(x, a)^\top \theta_h\}, H \right\} \right\}$$

with $M = \sqrt{SA}(H-h)$. Lemma 40 shows that we can construct a ϵ -net of $\mathcal{V}_h^*(f; r)$ with a cardinality that grows exponentially only in SA and not n (see Lemma D.6 in Jin et al. (2020)). This property is crucial to obtain an error rate that is independent of n . However, it is also important to note that such a net will still depend on the given reward functions r and the clustering function f . It is for this reason that we won't be able to use such a net argument in the *minimax* setting, and also why we split our budget of episodes into two parts in the design of our algorithm.

Lemma 40. *Let f be any latent state decoding function and r be any reward function, then there exists ϵ -net $\mathcal{N}_\epsilon(f; r)$ of $\mathcal{V}_h^*(f; r)$ with respect to $\|\cdot\|_\infty$, such that*

$$|\mathcal{N}_\epsilon(f; r)| \leq \left(1 + \frac{2\sqrt{SA}(H-h)}{\epsilon} \right)^{SA}.$$

J.6 Proofs – Value difference lemmas

Proof of Lemma 33 and Lemma 34. The starting point for the proof of Lemma 33 and Lemma 34 is the same. We recall that for any given reward r , the policy $\hat{\pi}_r$ is optimal under $\hat{\Phi}$, thus, and in particular, $\hat{V}^{\pi_r^*}(r) - \hat{V}^{\hat{\pi}_r}(r) < 0$. From this observation, we immediately obtain

$$\begin{aligned} V^*(r) - V^{\hat{\pi}_r}(r) &= V^*(r) - \hat{V}^{\pi_r^*}(r) + \hat{V}^{\pi_r^*}(r) - \hat{V}^{\hat{\pi}_r}(r) + \hat{V}^{\hat{\pi}_r}(r) - V^{\hat{\pi}_r}(r) \\ &\leq V^*(r) - \hat{V}^{\pi_r^*}(r) + \hat{V}^{\hat{\pi}_r}(r) - V^{\hat{\pi}_r}(r). \end{aligned} \quad (113)$$

Now, let us introduce an intermediary model $\tilde{\Phi} = (p, q, \hat{f})$, and denote \tilde{V}^π the value function of policy π under the model $\tilde{\Phi}$.

(Proof of Lemma 33). We start from (113) to write

$$V^*(r) - V^{\hat{\pi}_r}(r) \leq 2 \max_{\pi} \left| V^\pi(r) - \hat{V}^\pi(r) \right|.$$

Then, an immediate application of Lemma 35 gives us

$$V^*(r) - V^{\hat{\pi}_r}(r) \leq 2H \max_{\pi} \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^\pi \right| + \frac{2\eta^2 H |\mathcal{E}|}{n}.$$

We then observe that

$$\begin{aligned}
 & \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^\pi \right| \stackrel{(a)}{=} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \left(\tilde{V}_{h+1}^\pi - \tilde{P}(x, a) \tilde{V}_{h+1}^\pi \mathbf{1} \right) \right| \\
 & \stackrel{(b)}{\leq} \|\tilde{P}(x, a) - \hat{P}(x, a)\|_1 \|\tilde{V}_{h+1}^\pi - \tilde{P}(x, a) \tilde{V}_{h+1}^\pi \mathbf{1}\|_\infty \\
 & \stackrel{(c)}{\leq} 2\eta^2 \|\tilde{P}(x, a) - \hat{P}(x, a)\|_1 \\
 & \stackrel{(d)}{\leq} 2\eta^2 \left(\frac{2\eta^2 |\mathcal{E}|}{n} + \max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1 + \max_{s \in \mathcal{S}} \|\hat{q}(s) - q(s)\|_1 \right)
 \end{aligned}$$

where (a) follows by simply adding and subtracting a constant, (b) follows from Holder's inequality, (c) follows from applying Lemma 37 (we recall that $\mathbf{1} \in \mathbb{R}^n$ and has all its entries equal to 1), and (d) follows from applying Lemma 36. Therefore, we finally obtain that

$$V^*(r) - V^{\hat{\pi}_r}(r) \leq 2\eta^2 H \left(\max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1 + \max_{s \in \mathcal{S}} \|\hat{q}(s) - q(s)\|_1 \right) + \frac{4H\eta^4 |\mathcal{E}|}{n},$$

which leads to the desired solution by taking the supremum over all possible values of rewards r .

(Proof of Lemma 34). Again, we start from (113) to write

$$V^*(r) - V^{\hat{\pi}_r}(r) \leq \underbrace{\left| V^{\hat{\pi}_r}(r) - \hat{V}^{\hat{\pi}_r}(r) \right|}_{\text{Term 1}} + \underbrace{\left| V^{\pi^*}(r) - \hat{V}^{\pi^*}(r) \right|}_{\text{Term 2}}.$$

In contrast to the proof of Lemma 33, we will have to analyze each of the above terms separately to take advantage of the fact that $\hat{\pi}_r$ is optimal under $\hat{\Phi}$, thus exploiting the fact that $\hat{V}^{\hat{\pi}_r}$ has a special structure.

(Bounding term 1). First, an immediate application of Lemma 35 gives

$$\left| V^{\hat{\pi}_r}(r) - \hat{V}^{\hat{\pi}_r}(r) \right| \leq H \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \hat{V}_{h+1}^{\hat{\pi}_r} \right| + \frac{\eta^2 H |\mathcal{E}|}{n}$$

where we took the supremum over all possible optimal value functions at all steps $h \in [H]$. Then, applying Lemma 36, we also have for all $h \in [H]$, $V \in \mathcal{V}_{h+1}^*(\hat{f}, r)$,

$$\left| (\tilde{P}(x, a) - \hat{P}(x, a)) \hat{V}_{h+1}^{\hat{\pi}_r} \right| \leq \max_{s \in \mathcal{S}} \left| (\hat{q}(s) - q(s)) \hat{V}_{h+1}^{\hat{\pi}_r} \right| + H \max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1 + \frac{2\eta^2 H |\mathcal{E}|}{n}$$

where we also used the fact that $\|V\|_\infty \leq H$. Therefore, combining the above bounds gives

$$\begin{aligned}
 \frac{1}{H} \left| V^{\hat{\pi}_r}(r) - \hat{V}^{\hat{\pi}_r}(r) \right| & \leq \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \hat{V}_{h+1}^{\hat{\pi}_r} \right| \\
 & \quad + H \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 + \frac{3\eta^2 H |\mathcal{E}|}{n}.
 \end{aligned}$$

(Bounding term 2). Next, again, by applying Lemma 35, we obtain

$$\left| V^{\pi^*}(r) - \hat{V}^{\pi^*}(r) \right| \leq H \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^{\pi^*} \right| + \frac{\eta^2 H |\mathcal{E}|}{n}.$$

Then, applying Lemma 36, we also have for all $h \in [H]$,

$$\left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^{\pi^*} \right| \leq \max_{s \in \mathcal{S}} \left| (\hat{q}(s) - q(s)) \tilde{V}_{h+1}^{\pi^*} \right| + H \max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1 + \frac{2\eta^2 H |\mathcal{E}|}{n}$$

where we also used the fact that $\|V\|_\infty \leq H$. Therefore, we obtain

$$\frac{1}{H} \left| V^{\pi^*}(r) - \hat{V}^{\pi^*}(r) \right| \leq \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \tilde{V}_{h+1}^{\pi^*} \right|$$

$$+ H \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 + \frac{3\eta^2 H |\mathcal{E}|}{n}.$$

Now combining the bounds on terms 1 and 2, obtain

$$\begin{aligned} \frac{1}{H} (V^*(r) - V^{\hat{\pi}_r}(r)) &\leq \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \hat{V}_{h+1}^{\hat{\pi}_r} \right| + \max_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]} \left| (\hat{q}(s) - q(s)) \tilde{V}_{h+1}^{\pi^*} \right| \\ &\quad + 2H \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|\hat{p}(s, a) - p(s, a)\|_1 + \frac{6\eta^2 H |\mathcal{E}|}{n}. \end{aligned}$$

□

Proof of Lemma 35. We start by writing

$$|V^\pi(r) - \hat{V}^\pi(r)| \leq |V^\pi(r) - \tilde{V}^\pi(r)| + |\tilde{V}^\pi(r) - \hat{V}^\pi(r)|.$$

Now, Let us define for all $h \in [H]$, $x \in \mathcal{X}$,

$$\begin{aligned} E_h^\pi(x) &= |V_h^\pi(x) - \tilde{V}_h^\pi(x)| \\ \tilde{E}_h^\pi(x) &= |\tilde{V}_h^\pi(x) - \hat{V}_h^\pi(x)| \end{aligned}$$

and note that

$$|V^\pi(r) - \hat{V}^\pi(r)| \leq \mathbb{E}_{x_1} [E_1^\pi(x_1)] + \mathbb{E}_{x_1} [\tilde{E}_1^\pi(x_1)].$$

Step 1 – (Bounding $E_1^\pi(x)$) First, we have for all $h \in [H]$, $x \in \mathcal{X}$,

$$\begin{aligned} E_h^\pi(x) &= \left| P(x, \pi_h(x)) V_{h+1}^\pi - \tilde{P}(x, \pi_h(x)) \tilde{V}_{h+1}^\pi \right| \\ &\leq \left| (P(x, \pi_h(x)) - \tilde{P}(x, \pi_h(x))) V_{h+1}^\pi \right| + \tilde{P}(x, \pi_h(x)) \left| V_{h+1}^\pi - \tilde{V}_{h+1}^\pi \right| \\ &\leq \left| (P(x, \pi_h(x)) - \tilde{P}(x, \pi_h(x))) V_{h+1}^\pi \right| + \tilde{P}(x, \pi_h(x)) E_{h+1}^\pi \\ &\leq \mathbf{1}\{\hat{f}(x) \neq f(x)\} \max_{y \in \mathcal{X}} \left| V_{h+1}^\pi(y) - P(x, \pi_h(x)) V_{h+1}^\pi \right| + \tilde{P}(x, \pi_h(x)) E_{h+1}^\pi \\ &\leq 2\eta^2 \mathbf{1}\{\hat{f}(x) \neq f(x)\} + \tilde{P}(x, \pi_h(x)) E_{h+1}^\pi \end{aligned}$$

where we used the facts that (i) $\tilde{P}(x, a) = P(x, a)$ whenever $\hat{f}(x) = f(x)$, (ii) P is η^2 -regular, (iii) Lemma 37 to obtain $\max_{y \in \mathcal{X}} |V_{h+1}^\pi(y) - P(x, \pi_h(x)) V_{h+1}^\pi| \leq 2\eta^2 - 1$, and setting $E_{H+1}^\pi = 0$. We conclude after iterating the above recursion that

$$\begin{aligned} \mathbb{E}_{x_1} [E_1^\pi(x_1)] &\leq \mathbb{E}_{\tilde{P}} \left[\sum_{h=1}^H \mathbf{1}\{f(x_h^\pi) \neq \hat{f}(x_h^\pi)\} \right] \\ &\leq \max_{x \in \mathcal{X}} \tilde{N}(x) |\mathcal{E}| \\ &\leq \frac{\eta^2 H |\mathcal{E}|}{n} \end{aligned}$$

where $\tilde{N}(x) = \mathbb{E}_{\tilde{P}} [\sum_{h=1}^H \mathbf{1}\{x_h^\pi = x\}]$, and by η^2 -regularity of \tilde{P} , we can easily verify that $\tilde{N}(x) \leq \frac{\eta^2 H}{n}$.

Step 2 – (Bounding $\tilde{E}_1^\pi(x)$) First, we verify that \tilde{E}_h^π satisfies the following recursion

$$\begin{aligned} \tilde{E}_h^\pi(x) &= \left| \tilde{P}(x, \pi_h(x)) \tilde{V}_{h+1}^\pi - \hat{P}(x, \pi_h(x)) \hat{V}_{h+1}^\pi \right| \\ &\leq \left| (\tilde{P}(x, \pi_h(x)) - \hat{P}(x, \pi_h(x))) \tilde{V}_{h+1}^\pi \right| + \hat{P}(x, \pi_h(x)) \left| \tilde{V}_{h+1}^\pi - \hat{V}_{h+1}^\pi \right| \\ &\leq \left| (\tilde{P}(x, \pi_h(x)) - \hat{P}(x, \pi_h(x))) \tilde{V}_{h+1}^\pi \right| + \hat{P}(x, \pi_h(x)) \tilde{E}_{h+1}^\pi \end{aligned}$$

$$\leq \max_{x \in \mathcal{X}, a \in \mathcal{A}} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^\pi \right| + \hat{P}(x, \pi_h(x)) \tilde{E}_{h+1}^\pi$$

with $\tilde{E}_{H+1}^\pi = 0$. By iterating the above recursion, we then obtain

$$\begin{aligned} \mathbb{E}_{x_1} [\tilde{E}_1^\pi(x_1)] &\leq \sum_{h=1}^H \max_{x \in \mathcal{X}, a \in \mathcal{A}} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^\pi \right| \\ &\leq H \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \tilde{V}_{h+1}^\pi \right|. \end{aligned}$$

Similarly, we can also obtain

$$\mathbb{E}_{x_1} [\tilde{E}_1^\pi(x_1)] \leq H \max_{x \in \mathcal{X}, a \in \mathcal{A}, h \in [H]} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) \hat{V}_{h+1}^\pi \right|.$$

This completes the proof of the lemma. \square

Proof of Lemma 36. First, we will start by introducing two intermediate probability distributions: $\tilde{P}_1(x, a) = (q(y, \hat{f}(y))p(\hat{f}(y)|f(x), a))_{y \in \mathcal{X}}$, and $\tilde{P}_2(x, a) = (\hat{q}(y, \hat{f}(y))p(\hat{f}(y)|f(x), a))_{y \in \mathcal{X}}$. We may then write

$$\begin{aligned} \left| (\tilde{P}(x, a) - \hat{P}(x, a)) V \right| &\leq \left| (\tilde{P}(x, a) - \tilde{P}_1(x, a)) V \right| \\ &\quad + \left| (\tilde{P}_1(x, a) - \tilde{P}_2(x, a)) V \right| \\ &\quad + \left| (\tilde{P}_2(x, a) - \hat{P}(x, a)) V \right|. \end{aligned}$$

Note that $\tilde{P}(y|x, a) = \tilde{P}_1(y|x, a)$ whenever $f(y) = \hat{f}(y)$, thus

$$\begin{aligned} \left| (\tilde{P}(x, a) - \tilde{P}_1(x, a)) V \right| &\leq \|V\|_\infty \sum_{y \in \mathcal{X}} |\tilde{P}(y|x, a) - \tilde{P}_1(y|x, a)| \\ &\leq \|V\|_\infty \sum_{y \in \mathcal{X}} \mathbb{1}\{\hat{f}(y) \neq f(y)\} \max_{y \in \mathcal{X}} q(y|s') p(s'|s, a) \\ &\leq \frac{\eta^2 |\mathcal{E}| \|V\|_\infty}{n}. \end{aligned}$$

Next, we have

$$\begin{aligned} \left| (\tilde{P}_1(x, a) - \tilde{P}_2(x, a)) V \right| &\leq \left| \sum_{y \in \mathcal{X}} (q(y|\hat{f}(y)) - \hat{q}(y|\hat{f}(y))) p(\hat{f}(y)|\hat{f}(x), a) V(y) \right| \\ &\leq \sum_{s \in \mathcal{S}} p(s|\hat{f}(x), a) \left| \sum_{y \in \hat{f}^{-1}(s)} (q(y|s) - \hat{q}(y|s)) V(y) \right| \\ &\leq \sum_{s \in \mathcal{S}} p(s|\hat{f}(x), a) \left| \sum_{y \in \mathcal{X}} (q(y|s) - \hat{q}(y|s)) V(y) \right| \\ &\quad + \sum_{s \in \mathcal{S}} p(s|\hat{f}(x), a) \left| \sum_{y \in \mathcal{X}} \mathbb{1}\{\hat{f}(y) \neq f(y)\} q(y|s) V(y) \right| \\ &\leq \max_{s \in \mathcal{S}} \left| \sum_{y \in \mathcal{X}} (q(y|s) - \hat{q}(y|s)) V(y) \right| + \frac{\eta^2 |\mathcal{E}| \|V\|_\infty}{n} \end{aligned}$$

where we used that, by construction, $\hat{q}(y|s) = 0$ for all $s \notin \hat{f}^{-1}(y)$. Finally, we have

$$\left| (\tilde{P}_2(x, a) - \hat{P}(x, a)) V \right| = \left| \sum_{y \in \mathcal{X}} \hat{q}(y|\hat{f}(y)) (p(\hat{f}(y)|\hat{f}(x), a) - \hat{p}(\hat{f}(y)|\hat{f}(x), a)) V(y) \right|$$

$$\begin{aligned}
 &\leq \left| \sum_{s' \in \mathcal{S}} (p(s'|\hat{f}(x), a) - \hat{p}(s'|\hat{f}(x), a)) \sum_{y \in \hat{f}^{-1}(s)} \hat{q}(y|s)V(y) \right| \\
 &\leq \sum_{s' \in \mathcal{S}} |p(s'|\hat{f}(x), a) - \hat{p}(s'|\hat{f}(x), a)| \max_{s' \in \mathcal{S}} \left| \sum_{y \in \hat{f}^{-1}(s)} \hat{q}(y|s)V(y) \right| \\
 &\leq \|p(\hat{f}(x), a) - \hat{p}(\hat{f}(x), a)\|_1 \|V\|_\infty \\
 &\leq \|V\|_\infty \max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1
 \end{aligned}$$

To conclude, we obtain

$$\begin{aligned}
 \left| (\tilde{P}(x, a) - \hat{P}(x, a))V \right| &\leq \frac{2\eta^2 \|V\|_\infty}{n} + \max_{s \in \mathcal{S}} \|\hat{p}(s, a) - p(s, a)\|_1 \|V\|_\infty \\
 &\quad + \max_{s \in \mathcal{S}} \left| \sum_{y \in \mathcal{X}} (q(y|s) - \hat{q}(y|s))V(y) \right|
 \end{aligned}$$

□

Proof of Lemma 37. For simplicity, let us denote $g_h(x) = r_h(x, \pi_h(x))$ and $\mathbb{E}_h[\cdot] = \mathbb{E}^\pi[\cdot | x_h = y]$. Then we note that

$$V_{h+1}^\pi(x) - P(y, \pi_h(y))V_{h+1}^\pi = \mathbb{E}_h \left[\sum_{\ell=h}^H g_\ell(x_\ell) - \mathbb{E}_h[g_\ell(x_\ell)] \right]$$

Since the transitions $(P(\cdot, \pi_h(\cdot)))_{h \in \mathcal{H}}$ are η^2 -regular, we obtain that

$$\max_{x \in \mathcal{X}} \left| V_{h+1}^\pi(x) - P(y, \pi_h(y))V_{h+1}^\pi \right| \leq (2\eta^2 - 1)$$

□

J.7 Proofs – Concentration bounds on averaged optimal value functions

Proof of Lemma 38. First, let us note that $\hat{V}_{h+1}^{\hat{\pi}_r} \in \mathcal{V}_{h+1}^*(\hat{f}; r)$. Let \mathcal{N} be an ϵ -net of $\mathcal{V}^*(\hat{f}; r)$ with respect to $\|\cdot\|_\infty$. We have:

$$\begin{aligned}
 \left| (\hat{q}(s) - q(s))\hat{V}_h^{\hat{\pi}_r} \right| &\leq \sup_{V \in \mathcal{V}_h^*(\hat{f}; r)} \left| (\hat{q}(s) - q(s))V \right| \\
 &\leq \max_{V \in \mathcal{N}} \left| (\hat{q}(s) - q(s))\tilde{V} \right| + \sup_{V \in \mathcal{V}_h^*(\hat{f}; r)} \max_{\tilde{V} \in \mathcal{N}} \left| (\hat{q}(s) - q(s))(V - \tilde{V}) \right| \\
 &\leq \max_{V \in \mathcal{N}} \left| (\hat{q}(s) - q(s))\tilde{V} \right| + \|\hat{q}(s) - q(s)\|_1 \epsilon \\
 &\leq \max_{V \in \mathcal{N}} \left| (\hat{q}(s) - q(s))\tilde{V} \right| + 2\epsilon.
 \end{aligned}$$

Let us recall that the construction of \hat{q} and \hat{p} only uses half the subsequent budget of $TH/2$ observations. Thus, we have

$$\begin{aligned}
 \mathbb{P}\left(\left| (\hat{q}(s) - q(s))\hat{V}_h^{\hat{\pi}_r} \right| > \rho \right) &\leq \mathbb{P}\left(\max_{V \in \mathcal{N}} \left| (\hat{q}(s) - q(s))\tilde{V} \right| + 2\epsilon > \rho \right) \\
 &= \mathbb{E} \left[\mathbb{E} \left[\mathbf{1} \left\{ \max_{V \in \mathcal{N}} \left| (\hat{q}(s) - q(s))\tilde{V} \right| + 2\epsilon > \rho \right\} \middle| \hat{f} \right] \right] \\
 &\leq \mathbb{E} \left[\mathbb{E} \left[\sum_{\tilde{V} \in \mathcal{N}} \mathbf{1} \left\{ \left| (\hat{q}(s) - q(s))\tilde{V} \right| + 2\epsilon > \rho \right\} \middle| \hat{f} \right] \right] \\
 &\leq \mathbb{E} \left[|\mathcal{N}| \max_{\tilde{V} \in \mathcal{N}} \mathbb{E} \left[\mathbf{1} \left\{ \left| (\hat{q}(s) - q(s))\tilde{V} \right| + 2\epsilon > \rho \right\} \middle| \hat{f} \right] \right].
 \end{aligned}$$

Applying Proposition 26, we obtain for $TH \geq \text{poly}(\eta)S(\rho' + \log(S))$,

$$\mathbb{E} \left[\mathbb{1} \left\{ \left| (\hat{q}(s) - q(s))\tilde{V} \right| + 2\epsilon > \text{poly}(\eta)SH \left(\sqrt{\frac{\rho' + \log(S)}{TH}} + \frac{|\mathcal{E}|}{n} + \frac{\rho'}{TH} \right) + 2\epsilon \right\} \middle| \hat{f} \right] \leq 4e^{-\rho'}.$$

Additionally, Lemma 40 ensures that $|\mathcal{N}| \leq \left(1 + \frac{H\sqrt{SA}}{\epsilon}\right)^{SA}$. Therefore,

$$\mathbb{E} \left[|\mathcal{N}| \max_{\tilde{V} \in \mathcal{N}} \mathbb{E} \left[\mathbb{1} \left\{ \left| (\hat{q}(s) - q(s))\tilde{V} \right| + 2\epsilon > \rho \right\} \middle| \hat{f} \right] \right] \leq 4 \left(1 + \frac{\sqrt{SA}(H-h)}{\epsilon}\right)^{SA} e^{-\rho'}$$

where we set $\rho = \text{poly}(\eta)SH \left(\sqrt{\frac{\rho' + \log(S)}{TH}} + \frac{|\mathcal{E}|}{n} + \frac{\rho'}{TH} \right) + 2\epsilon$. Further, reparametrizing by $\rho'' = \rho' + SA \log\left(1 + \frac{2H\sqrt{SA}}{\epsilon}\right)$, we obtain

$$\mathbb{P} \left(\left| (\hat{q}(s) - q(s))\hat{V}_h^{\hat{\pi}_r} \right| > \rho \right) \leq 4e^{-\rho''}$$

where

$$\rho = \text{poly}(\eta)SH \left(\sqrt{\frac{\rho'' + SA \log\left(1 + \frac{2H\sqrt{SA}}{\epsilon}\right) + \log(S)}{TH}} + \frac{|\mathcal{E}|}{n} + \frac{\rho'}{TH} \right) + 2\epsilon$$

Now choosing $\epsilon = \sqrt{\frac{SA}{n}}$, we obtain that

$$\begin{aligned} \rho &\leq \text{poly}(\eta)SH \left(\sqrt{\frac{\rho'' + SA \log(Hn) + \log(S)}{TH}} + \frac{|\mathcal{E}|}{n} + \frac{\rho'' + SA \log(Hn)}{TH} \right) + 2\sqrt{\frac{SA}{n}} \\ &\leq \text{poly}(\eta)SH \left(\sqrt{\frac{\rho'' + SA \log(HSn)}{TH}} + \frac{|\mathcal{E}|}{n} + \frac{\rho'' + SA \log(HSn)}{TH} \right) + 2\sqrt{\frac{SA}{TH}} \end{aligned}$$

where the second inequality holds as long as $TH = \Omega(n)$. Thus, choosing $\rho'' = \frac{1}{HSn}$ gives that for $TH = \Omega(n)$,

$$\mathbb{P} \left(\left| (\hat{q}(s) - q(s))\hat{V}_h^{\hat{\pi}_r} \right| > \text{poly}(\eta)SH \left(\sqrt{\frac{SA \log(HSn)}{TH}} + \frac{|\mathcal{E}|}{n} \right) + 2\sqrt{\frac{SA}{TH}} \right) \leq \frac{4}{HSn}$$

where we also used the fact that $\sqrt{\frac{SA \log(HSn)}{TH}} \geq \frac{SA \log(HSn)}{TH}$ for $TH = \Omega(\log(n))$. Therefore, using a union bound, we finally obtain that

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, h \in [H]} \left| (\hat{q}(s) - q(s))\hat{V}_h^{\hat{\pi}_r} \right| > \text{poly}(\eta)SH \left(\sqrt{\frac{SA \log(HSn)}{TH}} + \frac{|\mathcal{E}|}{n} \right) \right) \leq \frac{4}{n}$$

provided $TH = \Omega(n)$. Similarly, but this time without going through a net argument since \tilde{V}_h^* conditionally on \hat{f} is no more random, we can obtain that

$$\mathbb{P} \left(\max_{s \in \mathcal{S}, h \in [H]} \left| (\hat{q}(s) - q(s))V_h^* \right| > \text{poly}(\eta)SH \left(\sqrt{\frac{SA \log(HSn)}{TH}} + \frac{SH|\mathcal{E}|}{n} \right) \right) \leq \frac{4}{n}$$

for $TH = \Omega(n)$. □

K Towards Optimal Adaptive Exploration in BMDPs

We have so far considered scenarios where the T episodes were generated under a fixed behavior policy ρ . In this section, we briefly discuss how our results could be extended to the case where the behavior policy is (history-)adaptive. Note that for reward-free RL, our lower bounds derived for the sample complexity of identifying near-optimal policies (see Theorem 4 and 5) hold even for adaptive exploration policies; surprisingly, Theorem 6 and 7 establish that even using uniform behavior policy, we can match these lower bounds. However, we believe that using adaptive behavior policies would actually yield significant performance improvements.

As before, we denote by $\{\mathcal{T}^{(\tau)}\}_{\tau=1}^T$ the data collected during the T episodes. In case of an adaptive exploration process, the policy at episode t , $\rho^{(t)}$, used to generate $\mathcal{T}^{(t)}$ may be chosen depending on the previously observed data history $\{\mathcal{T}^{(\tau)}\}_{\tau=1}^{t-1}$.

Before we present ways to extend our analysis to the case of adaptive behavior policies, we provide here a few pointers towards existing studies on the adaptive design of the exploration process for various inference tasks, especially for reward-free RL. In tabular MDP, adaptive exploration for identifying near-optimal policies has extensively been studied with both minimax or problem-specific guarantees and in the reward-specific and reward-free settings, see e.g. Zanette and Brunskill (2019); Jin et al. (2020); Marjani and Proutière (2021); Al Marjani et al. (2021); Menard et al. (2021) and references therein. Recently, Tarbouriech and Lazaric (2019); Tarbouriech et al. (2021) developed a framework to learn the model or a policy with prescribed frequencies at which (state, action) pairs are visited. All the aforementioned studies are restricted to general tabular MDPs. For BMDPs, as already discussed in Section 6, most of the proposed algorithms use adaptive policies in settings where the reward is already known and cluster recovery *isn't* the main objective. Indeed, the problem of estimating the latent state decoding function in BMDPs has some similarities with the cluster recovery problem in SBMs, DCBMs, and BMCs. However, even for SBMs, there isn't much work on adaptive sampling strategies; refer to Yun and Proutière (2014, 2019) for preliminary results.

K.1 Lower bound on the latent state decoding error rate

We apply the same change-of-measure argument to derive this lower bound. Fix a context x . As the initial part of our lower bound proof is applicable to adaptive policies as well (from Appendix D), we first recall them here.

We start with the lower bound of the individual error rate:

$$\varepsilon_x = \mathbb{P}_\Phi[x \in \mathcal{E}] \geq \frac{1}{2S} \exp\left(-\mathbb{E}_\Psi[\mathcal{L}] - \sqrt{2\eta S \text{Var}_\Psi[\mathcal{L}]}\right).$$

We can further establish the connection between $\mathbb{E}_\Psi[\mathcal{L}]$ and the rate function $I_j(x; c, m_\rho^\Psi, \Phi)$ defined as:

$$\begin{aligned} I_j^{(t)}(x; c, m_\rho^\Psi, \Phi) &:= n \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \left\{ cq(x|f(x))p(j|s, a)m_\rho^{\Psi, (t)}(s, a) \log \frac{cp(j|s, a)}{p(f(x)|s, a)} \right. \\ &\quad + cq(x|f(x))m_\rho^{\Psi, (t)}(j, a)p(s|j, a) \log \frac{p(s|j, a)}{p(s|f(x), a)} \\ &\quad \left. + (1 - cq(x|f(x))p(j|s, a))m_\rho^{\Psi, (t)}(s, a) \log \frac{1 - cq(x|f(x))p(j|s, a)}{1 - q(x|f(x))p(f(x)|s, a)} \right\}, \end{aligned}$$

where $m_\rho^{\Psi, (t)}(s, a)$ denotes the expected proportion of rounds spent in (latent state, action) pair (s, a) under policy ρ and model Ψ , in the t -th episode:

$$m_\rho^{\Psi, (t)}(s, a) := \frac{1}{H-1} \sum_{h=1}^{H-1} \mathbb{P}_\Psi[\tilde{f}(x_h^{(t)}) = s, a_h^{(t)} = a].$$

We add the explicit dependence on m_ρ^Ψ for the rate function to remember that we can design the adaptive policy ρ based on the desirable m_ρ^Ψ .

We also have:

$$\mathbb{E}_\Psi[\mathcal{L}] \leq \frac{TH}{n} (I_j(x; c, m_\rho^\Psi, \Phi) + \mathcal{O}\left(\frac{1}{n}\right)) \quad \text{and} \quad \text{Var}_\Psi[\mathcal{L}] \leq \mathcal{O}\left(\frac{TH}{n}\right).$$

By defining

$$I(x; m_\rho^\Psi, \Phi) := \min_{j: j \neq f(x)} \min_{c > 0} I_j(x; c, m_\rho^\Psi, \Phi), \quad (114)$$

we can show, as previously, that under the adaptive policy ρ ,

$$\varepsilon_x = \mathbb{P}_\Phi[x \in \mathcal{E}] \geq \frac{1}{2\eta S} \exp\left(-\frac{TH}{n} I(x; m_\rho^\Psi, \Phi)(1 + o(1))\right).$$

We deduce that the expected number of misclassified contexts under the adaptive policy ρ is lower bounded as follows:

$$\mathbb{E}_\Phi[|\mathcal{E}|] \gtrsim n \exp\left(-\frac{TH}{n} I(m_\rho^\Phi, \Phi)(1 + o(1))\right),$$

where \gtrsim hides additional (polynomial) dependency on η, S, A , and

$$I(m_\rho^\Phi, \Phi) = -\frac{n}{TH} \log\left(\frac{1}{2\eta S n} \sum_x \exp\left(-\frac{TH}{n} I(x; m_\rho^\Phi, \Phi)\right)\right).$$

Note that we could replace m_ρ^Ψ by m_ρ^Φ in the above definition, because Φ and Ψ only differ from a single context. The formal justification is left for future work. Now, the difference with scenarios without adaptive exploration policies is that here the learner can design policies to optimize m_ρ^Φ . We denote by \mathcal{M} the set of possible expected proportions of rounds spent in the various (s, a) pairs. \mathcal{M} can be intricate to characterize for small T but when T grows large, we can proceed as in [Al Marjani et al. \(2021\)](#) and write the *navigation* constraints $m \in \mathcal{M}$ should asymptotically satisfy.

We now define the *adaptive rate function*:

$$I'(\Phi) := \max_{m \in \mathcal{M}} I(m, \Phi). \tag{115}$$

Finally, we can state the following preliminary result:

Theorem 12. *Let Φ be a BMDP satisfying Assumptions 1,2,3,4. Consider a clustering algorithm that is β -locally better-than-random for Φ with $\beta \geq \frac{2S\eta^2}{n}$, when applied to the data gathered over T episodes, each of length H , using **adaptive** ρ . Then, we have that*

$$\mathbb{E}_\Phi[|\mathcal{E}|] \gtrsim n \exp\left(-\frac{TH}{n} I'(\Phi)(1 + o(1))\right).$$

K.2 Adaptive latent state decoding algorithms

An interesting interpretation of the adaptive lower bound is that $m^*(\Phi) \in \arg \max_{m \in \mathcal{M}} I(m, \Phi)$ represents an optimal way of navigating the BMDP to minimize the proportion of misclassified context. We can leverage this observation to design an adaptive algorithm. We can proceed in the following steps:

1. Using δT episodes, explore the BMDP uniformly at random and run Algorithm 1 and 2 to obtain the initial estimate $\hat{\Phi}_1 = (\hat{f}_1, \hat{p}_1, \hat{q}_1)$
2. Compute $\hat{m} \in \arg \max_{m \in \mathcal{M}(\hat{\Phi}_1)} \tilde{I}(x; m, \hat{\Phi})$ for each $x \in \mathcal{X}$ (see Appendix D.3.2)
3. Using the remaining $(1 - \delta)T$ episodes, adaptively explore the BMDP such that the expected numbers of visits to each (estimated latent state-action) pair match \hat{m} ;
4. Based upon the new samples, run the cluster improvement algorithm to output improved $\hat{\Phi} = (\hat{f}, \hat{p}, \hat{q})$.

We emphasize that the above algorithm is not complete, and it just provides good design principles.

L Beyond the η -regularity Assumption

In this section, we describe how to recover our results under a weaker assumption than Assumption 3(ii), namely, only assuming that the dynamics of the latent state are ergodic.

L.1 Relaxing Assumption 3(ii)

We may relax Assumption 3(ii) in the following way as done for instance in [Azizzadenesheli et al. \(2016a\)](#).

Assumption 5. *The Markov chains corresponding to the latent transition matrices, $p_a = (p(s'|s, a))_{s, s' \in \mathcal{S}}$ for all $a \in \mathcal{A}$, are aperiodic and communicating.*

A direct consequence of Assumption 5 is that we may define a regularity parameter for the k -step transition probabilities, which again will allow us to quantify the mixing properties of the latent dynamics in our BMDPs, namely those described by the transition matrices $p_a = (p(s'|s, a))_{s, s' \in \mathcal{S}}$ for all $a \in \mathcal{A}$. Additionally, aperiodicity may be further relaxed by simply requiring that a stationary distribution exists instead.

Lemma 41. *Let P be the transition matrix corresponding to a communicating and aperiodic Markov chain over a finite state space \mathcal{Z} . Then, setting $k = |\mathcal{Z}|^2$, we have:*

(i) *there exists $\eta \geq 1$, such that*

$$\eta \geq \max_{z_1, z_2, z_3 \in \mathcal{Z}} \left\{ \frac{P^k(z_1|z_3)}{P^k(z_2|z_3)}, \frac{P^k(z_1|z_2)}{P^k(z_1|z_3)} \right\};$$

(ii) *the Dobrushin's coefficient of P^k is non-trivially bounded i.e., $\delta(P^k) \leq 1 - 1/\eta$;*

(iii) *the mixing time of P satisfies $t_{\text{mix}}(\varepsilon) \leq |\mathcal{Z}|^2(\eta \log(1/\varepsilon) + 1)$.*

Proof sketch of Lemma 41. To establish (i) we simply need to verify that for $k = |\mathcal{Z}|^2$, $P^k(z_1|z_2) > 0$ for all $z_1, z_2 \in \mathcal{Z}$. Because the Markov chain is communicating, there exists $k_0 \geq 0$, such that for all $k \geq k_0$, the k -step transition probabilities are strictly positive. Corollary 1 of [Denardo \(1977\)](#) implies that $k_0 \leq |\mathcal{Z}|^2$. Equipped with (i), the proof of (ii) follows immediately from Definition 3 and the proof of (iii) follows similarly as in the proof of Proposition 5. \square

Remark 3. *In view of Lemma 41, under Assumption 5, we can easily verify most of our results in Appendix C generalize. Most importantly, for all $a \in \mathcal{A}$, the S^2 -step transition $P_a^{S^2}$ will be η -regular, even though the state space \mathcal{X} is of size $n \gg S$. Consequently, the S^2 -step transitions of the induced Markov chains under our model will also inherit the η -regularity.*

L.2 Bernstein-type concentration bounds for Markov chains with restarts.

The extension of the concentration bound presented in Theorem 10 for the time-homogeneous Markov chain immediately follows under the assumption that the Markov chain is aperiodic and communicating.

Theorem 13. *Let $\{(X_h^{(t)})_{h=1}^H\}_{t \in [T]}$ be a collection of i.i.d. time-homogeneous Markov chains over a finite state space \mathcal{Z} , with transition probability matrix P and initial distribution $\mu \in \mathcal{P}(\mathcal{Z})$. We assume that μ and P^k are η -regular, and that P admits a stationary distribution ν . Let $(\phi_h)_{h \geq 1}$ be sequence of mapping from \mathcal{X} to \mathbb{R} , bounded and measurable. Then we have that for all $u \geq 0$,*

$$\mathbb{P} \left(\sum_{t=1}^T \sum_{h=1}^H \phi_h(X_h^{(t)}) - \mathbb{E}_\mu[\phi_h(X_h^{(t)})] > u \right) \leq k \exp \left(- \frac{u^2}{2THV_{\mu, P, \phi} + \frac{2}{3}M_{P, \phi}u} \right)$$

where

$$V_{\mu, P, \phi} \leq \text{poly}(\eta)k^2 \max_{h \geq 1, i \in \{0, \dots, k-1\}} \{ \text{Var}_{P^k(x, \cdot)}[\phi_h], \text{Var}_{\mu P^i}[\phi_h] \},$$

$$M_{P, \phi} \leq \text{poly}(\eta)k \max_{h \geq 1} \|\phi_h\|^2.$$

Proof sketch of Theorem 13. In fact, the proof steps are similar to those of the proof of Theorem 9 with some extra steps. First, recall from the proof of Lemma 7 that we may write

$$\sum_{h=1}^H \phi_h(X_h) - \mathbb{E}_\mu[\phi_h(X_h)] = \sum_{\ell=1}^H Z_\ell \left(\sum_{h=\ell}^H (P - \Pi)^{h-\ell} \phi_h \right) = \sum_{i=0}^{k-1} S_i$$

with

$$S_i = \sum_{\ell=1}^{H/k} Z_{\ell k+i} \left(\sum_{h=\ell k+i}^H (P - \Pi)^{h-\ell} \phi_h \right)$$

where we assume for simplicity that H/k is an integer. We note that for each $i \in \{0, \dots, k-1\}$, S_i is a martingale. We will analyze each S_i separately and then combine our concentration bounds via a union bound. Now let i be fixed, following a similar argument as in the proof of Theorem 9, we can conclude by upper bounding the terms

$$V_{\mu, P, \phi} = \max_{z \in \mathcal{Z}, \ell k+i \in [H]} \left\{ \text{Var}_{P^k(z, \cdot)} \left[\sum_{h=\ell}^H (P - \Pi)^{h-\ell} \phi_h \right], \text{Var}_{\mu^{P^i}(z, \cdot)} \left[\sum_{h=\ell k+i}^H (P - \Pi)^{h-\ell} \phi_h \right] \right\}$$

$$M_{P, \phi} = \left\| \sum_{h=\ell k+i}^H (P - \Pi)^{h-\ell} \phi_h \right\|_\infty.$$

We recall that we used Lemma 9 and Lemma 10 for that. Now, we shall provide the key steps that need to be adapted in order to obtain similar lemmas. First, we can establish, following a similar reasoning as in Lemma 9 that for any η -regular distribution μ , we have

$$V_\mu \left[\sum_{h=\ell}^H (P - \Pi)^{h-\ell} \phi_h \right] \leq \left(\sum_{h=\ell}^H \|(P - \Pi)^{h-\ell} (\phi_h - \mu \phi_h \mathbf{1})\|_\infty \right)^2$$

$$\leq \left(\sum_{h=0}^{\infty} k \max_{i \in [k]} \|(P - \Pi)^{kh} (\phi_{hk+i} - \mu \phi_{hk+i} \mathbf{1})\|_\infty \right)^2$$

where we recall here that $k = |\mathcal{Z}|^2$, upper bound naively the sum by the infinite sum, then split the sum into blocks of size k to obtain the final inequality. Next, we can easily verify that for all $h \geq 1$, P^{kh} , Π and μ are all η -regular since P^k is η -regular. Thus, since $(P - \Pi)^k = P^k - \Pi$, we can obtain in a similar fashion as in the proof of Lemma 9, we have

$$\max_{x, y \in \mathcal{Z}} \left| \frac{(P - \Pi)^{kh}(x, y)}{\mu(x)} \right| \leq \eta^2 - 1$$

Thus, we have for all $h \geq 1$,

$$\max_{i \in [k]} \|(P - \Pi)^{kh} (\phi_{hk+i} - \mu \phi_{hk+i} \mathbf{1})\|_\infty \leq 2(\eta^2 - 1) \delta(P^k) \max_{h \geq 1} \|\phi_h - \mu \phi_h \mathbf{1}\|_\mu$$

$$\leq 2(\eta^2 - 1) \left(1 - \frac{1}{\eta}\right)^h \max_{t \geq 1} \|\phi_t - \mu \phi_t \mathbf{1}\|_\mu$$

where in an intermediate step we use

$$\|(P - \Pi)^{kh} g\|_\infty \leq 2\delta(P^{kh}) \|g\|_\infty \leq 2(\delta(P^k))^h \|g\|_\infty \leq 2 \left(1 - \frac{1}{\eta}\right)^h \|g\|_\infty$$

for all g , such that $\|g\|_\infty < \infty$, where again the proof follows similarly as in the proof of Lemma 10. Therefore, we obtain at the end that for $\ell \in [H]$,

$$V_{\mu, P, \Phi} \leq \text{poly}(\eta) k^2 \max_{h \geq 1} \text{Var}_\mu[\phi_h]$$

$$M_{P, \Phi} \leq \text{poly}(\eta) k \max_{h \geq 1} \|\phi_h\|_\infty$$

□

Now, to further extend the above concentration to time-inhomogeneous Markov chains, we need a somewhat stronger assumption on the sequence of the transition matrices $(P_h)_{h \geq 1}$. More precisely, along with the assumption that all P_h 's are communicating and aperiodic, we also need to impose that

$$\forall m \geq 1, \quad \lim_{n \rightarrow \infty} \delta \left(\prod_{i=m}^n P_i \right) = 0$$

where we recall that $\delta(P)$ denotes the Dobrushin coefficient of the stochastic matrix P (see Definition 3 in Appendix C).

One observation is that our results for BMDPs do not require the use of a concentration for a time-inhomogeneous Markov chain, except for our horizon-free guarantee in Theorem 6, which only needs the bounds established in Lemma 9 and Lemma 10. Extension to the time-inhomogeneous case is left to future work.

L.3 On the lower bound for the latent state decoding error rate

To extend our lower bound (Theorem 1), we have to be careful when using the change-of-measure argument. More precisely, when defining the alternative models, we have to make sure that the absolute continuity of the dynamics under the true model, with respect to the dynamics of the perturbed models, holds. This will result in technical manipulations that will require a new definition of $I(x; \Phi)$ and $I(\Phi)$.

L.4 On the performance guarantee of the initial spectral clustering (Algorithm 1)

The algorithm of the initial clustering will remain the same and we expect that its performance guarantees will still hold under the relaxed assumption. As for the analysis, equipped with Theorem 13, we can immediately obtain all the required concentration results to prove Theorem 2.

Adapting the analysis of the concentration of the trimmed matrix. This will be the main part that requires changes in the proofs. First, all the concentration results used in proving Theorem 2 that relied on Theorem 9 are actually very easily modified by instead using Theorem 13. We will only pay an additional price for the dependence on S (which will not affect our results in terms of scaling in n). Second, our proof relied on the fact that the underlying Markov chains will visit each context $x \in \mathcal{X}$ with a probability that is roughly of order $1/n$ within two steps. That is why we used a two-step conditioning (See Appendix F.1). This can be easily fixed by instead using $(S^2 + 1)$ -step conditioning. To this aim, we may define for all a, t, h , $\tilde{N}_{a,t,h} = \mathbb{E}[\hat{N}_{a,t,h} | \hat{N}_{a,t,h-S^2-1}]$, and analyze the matrix differences

$$\forall r \in [S^2], \quad \sum_{t=1}^T \sum_{\ell=1}^{H/S^2} \hat{N}_{a,t,\ell S^2+r} - \tilde{N}_{a,t,\ell S^2+r}$$

whenever this is required in the proof instead of the odd and even terms, and then conclude by union bound. This will only come at a cost of $\text{poly}(S) \ll n$ in the new concentration bounds (see Appendix F.6). Finally, we will obtain the same results up to some minor technical details.

Separability, S -rank approximation, and ℓ_1 -weighted-K-means. These parts in the proof of Theorem 2 will practically remain unchanged, except for the separability section for which we need to take into account the $(S^2 + 1)$ -step conditioning. Again, we will obtain the same results up to minor technical details.

L.5 On the guarantee of the iterative likelihood improvement (Algorithm 2)

Algorithm 2 corresponding to the iterative likelihood improvement is inspired by our lower bound. Therefore, the expression of the updates in the algorithm might change depending on how the definition of $I(\Phi)$ and $I(x; \Phi)$ change. As for the analysis, we expect that the proof of Theorem 3 will not change much except for the requirement of new concentration bounds which again will be immediately obtained from Theorem 13.

L.6 On the guarantees for reward-free RL

For the reward-specific setting, the horizon-free result of Theorem 6 may not be preserved, unless we are able to recover the claims of Lemmas 9 and 10 with time-inhomogeneous Markov chains. On the other hand, the results of Theorem 6 and Theorem 7 will still hold with perhaps slightly worse dependencies in S, A, H , but not in n and T .

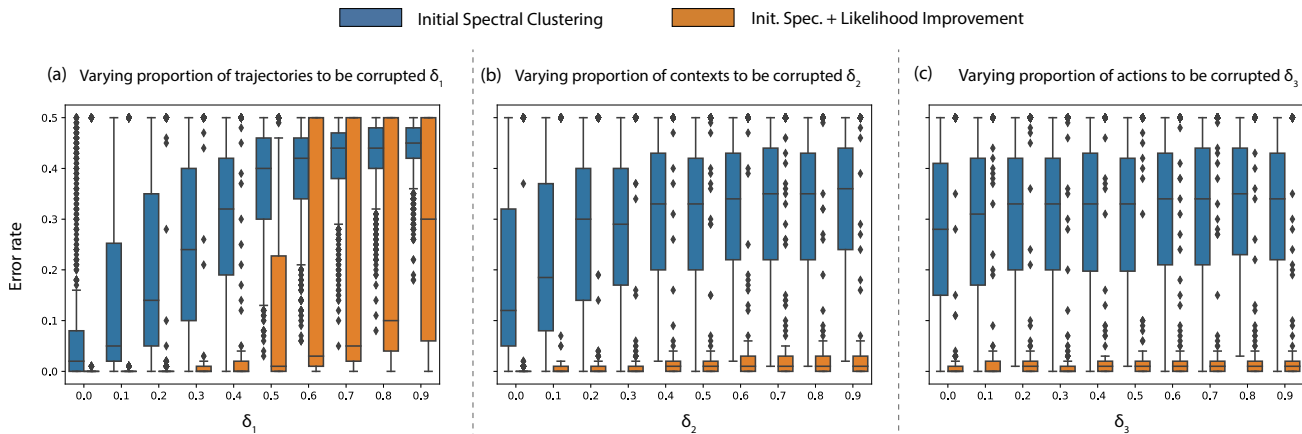


Figure 3: The clustering error rates with corruptions for various choices of (a) δ_1 's, (b) δ_2 's, and (c) δ_3 's.

M Deferred Experimental Details

M.1 Missing details for Section 7

The k -median clustering in Algorithm 1 is implemented using the `pyclustering` package (Novikov, 2019). In each setting, each parameter was varied as follows: $T \in \{5, 10, \dots, 45\}$ with $H = 100$ and $\eta = 4.0$; $H \in \{20, 30, \dots, 100\}$ with $T = 30$ and $\eta = 4.0$; $\eta \in \{1.5, 2.0, \dots, 6.0\}$ with $T = 30$ and $H = 100$.

M.2 Additional experiment: clustering with random corruptions

In this additional experiment (same setting as in Section 7), we show how robust the clustering algorithm is to corruption. For simplicity, we consider an oblivious adversary that chooses $\delta_1 T$ trajectories as well as $\delta_2 n$ contexts and $\delta_3 A$ actions (for each selected trajectory) to corrupt, uniformly at random. Contexts are corrupted by replacing them with other contexts in different clusters, and actions are corrupted by replacing them with other actions. Over all experiments, we fix $T = 30$, $H = 100$, and $\eta = 4.0$ vary each corruption ratio $\delta_1, \delta_2, \delta_3$ separately by fixing the other two ratios. Precisely, each ratio was varied as follows: $\delta_i \in \{0, 0.1, \dots, 0.9\}$ with $\delta_j = \delta_k = 0.4$, for $\{i, j, k\} = \{1, 2, 3\}$.

Figure 3 shows the results for clustering with corruption. As expected, the error rate increases (on average) as the corruption level becomes more severe. Note that for our example, the performance of our clustering is particularly robust to δ_2 (proportion of corrupted contexts) and δ_3 (proportion of corrupted actions). Surprisingly, it can be seen that up to a certain corruption level, our algorithm recovers near-exact clustering.

For future work, it would be interesting to theoretically and empirically investigate the effect of oblivious and adaptive adversarial corruption on clustering in BMDPs. Empirically, one could start by constructing various other types of BMDP environments and similarly test out our algorithms. Theoretically, one may take inspiration from recent progress on robust community detection (Al-Sharoa et al., 2021; Liu and Moitra, 2022) or robust RL (Wu et al., 2021). Another future direction is to make the algorithm provably these corruptions. One possible approach is to utilize adaptive clustering schemes (Yun and Proutière, 2014, 2019) or active model estimation (Tarbouriech and Lazaric, 2019).