Improved Exploration in Factored Average-Reward MDPs

Mohammad Sadegh Talebi
Department of Computer Science
University of Copenhagen

Anders Jonsson
ICT Department
Universitat Pompeu Fabra

Odalric-Ambrym Maillard
Univ. Lille, Inria, CNRS, Centrale Lille
UMR 9189 – CRIStAL, F-59000 Lille, France

Abstract

We consider a regret minimization task under the average-reward criterion in an unknown Factored Markov Decision Process (FMDP). More specifically, we consider an FMDP where the state-action space $\mathcal{X}$ and the state-space $\mathcal{S}$ admit the respective factored forms of $\mathcal{X} = \otimes_{i=1}^{n} \mathcal{X}_i$ and $\mathcal{S} = \otimes_{i=1}^{m} \mathcal{S}_i$, and the transition and reward functions are factored over $\mathcal{X}$ and $\mathcal{S}$. Assuming known factorization structure, we introduce a novel regret minimization strategy inspired by the popular UCRL2 strategy, called DBN-UCRL, which relies on Bernstein-type confidence sets defined for individual elements of the transition function. We show that for a generic factorization structure, DBN-UCRL achieves a regret bound, whose leading term strictly improves over existing regret bounds in terms of the dependencies on the size of $\mathcal{S}_i$'s and the involved diameter-related terms. We further show that when the factorization structure corresponds to the Cartesian product of some base MDPs, the regret of DBN-UCRL is upper bounded by the sum of regret of the base MDPs. We demonstrate, through numerical experiments on standard environments, that DBN-UCRL enjoys a substantially improved regret empirically over existing algorithms that have frequentist regret guarantees.

1 INTRODUCTION

In reinforcement learning (RL), an agent repeatedly interacts with an unknown environment in order to maximize its cumulative reward. A typical model of the environment is a Markov decision process (MDP): In each time step, the agent observes a state, takes an action and receives a reward before transiting to the next state. To achieve its objective, the agent has to estimate the parameters of the MDP from experience and learn a policy that maps states to actions. While doing so, the agent faces a choice between two basic strategies: *Exploration*, i.e. discovering the effects of actions on the environment, and *exploitation*, i.e. using its current knowledge to maximize reward in the short term.

Most model-based RL algorithms treat the state as a black box. In many practical cases, however, the environment exhibits structure that can be exploited to learn more efficiently. A common form of such structure is *factorization*. In a Factored MDP (FMDP), (see, e.g., Boutilier et al. (1999)), the state-space $\mathcal{S} = \otimes_{i=1}^{m} \mathcal{S}_i$ and action space $\mathcal{A} = \otimes_{i=1}^{n-m} \mathcal{A}_i$ are composed of m and $n-m$ individual factors, respectively. In this context, a state-action pair $x = (s, a) \in \mathcal{X} := \mathcal{S} \times \mathcal{A}$ is a tuple of n factor values. Each state factor $\mathcal{S}_i$ has its own transition function P_i, and the new factor value of $\mathcal{S}_i$, as a result of applying action a in state s, only depends on a small subset of the factors in $\mathcal{S} \times \mathcal{A}$. For $\mathcal{S}_i$, the set $Z_i \subset \{1, \dots, n\}$, termed the *scope* of $\mathcal{S}_i$, collects the indices of relevant factors for $\mathcal{S}_i$. Then P_i only depends on $\mathcal{X}[Z_i] := \otimes_{i \in Z_i} \mathcal{X}_i \subset \mathcal{X}$. Namely, $\mathcal{S}_i$ is conditionally independent of factors with indices outside Z_i. This *conditional independence structure* can be exploited to compactly represent the parameters of an FMDP. (We present a complete definition of FMDPs in Section 2.)

In this paper we consider the problem of regret minimization in FMDPs. Regret measures how much more reward the agent could have obtained using the best stationary policy, compared to the actual reward obtained. To achieve low regret, the agent must carefully balance exploration and exploitation: An agent that explores too much will not accumulate enough reward, while the one exploiting too much may fail to discover high-reward regions of the state-space.

Related Work. Factored state representations have been used since the early days of artificial intelligence (Fikes and Nilsson, 1971). In RL, factored states were first proposed as part of Probabilistic STRIPS (Boutilier and Dearden, 1994). When the FMDP structure and parameters are known, researchers have proposed two main approaches for efficiently learning a policy. The first approach consists in maintaining and updating a structured representation of the policy (Boutilier et al. (1999); Poupart et al. (2002); Degris et al. (2006); Raghavan et al. (2015)), whereas the second is to perform linear function approximation over a set of

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

basis functions (Guestrin et al. (2003); Dolgov and Durfee (2006); Szita and Lőrincz (2008)). However, only a few theoretical guarantees for these approaches exist. When the FMDP structure and parameters are unknown, several authors have proposed algorithms for structure learning (Kearns and Koller (1999b); Strehl et al. (2007); Diuk et al. (2009); Chakraborty and Stone (2011); Hallak et al. (2015); Guo and Brunskill (2018); Rosenberg and Mansour (2020)). Many of these algorithms admit PAC-type guarantees on their sample complexities. The focus of this paper is RL in an FMDP under the average-reward criterion, in an intermediate setting where the underlying structure of the FDMP is *known*, while actual reward and transition distributions are *unknown*. There is a rich and growing literature on average-reward RL in finite non-factored MDPs, where several algorithms with theoretical regret guarantees are presented (e.g., Burnetas and Katehakis (1997); Jaksch et al. (2010); Bartlett and Tewari (2009); Fruit et al. (2018); Talebi and Maillard (2018); Zhang and Ji (2019); QIAN et al. (2019); Bourel et al. (2020); Wei et al. (2020)) [1].

Despite such a rich literature in non-factored MDPs, RL in FMDPs has received relatively less attention, and only a few algorithms with performance guarantees in terms of regret or sample complexity are known. Among a few existing works, Kearns and Koller (1999a); Szita and Lőrincz (2009) study RL in discounted FMDPs presenting DBN-E^3 and FOIM, respectively. In the regret setting, Osband and Van Roy (2014b) present the first algorithms with provably sublinear regret, and are followed very recently by Xu and Tewari (2020); Tian et al. (2020); Chen et al. (2021); Rosenberg and Mansour (2020). Except (Rosenberg and Mansour, 2020), all these works assume a known structure. In the episodic setting, Osband and Van Roy (2014b) present Factored-UCRL achieving a regret of $\widetilde{\mathcal{O}}(D\sum_{i=1}^m \sqrt{S_i|\mathcal{X}[Z_i]|T})$ after T steps.[2] (To simplify the presentation, in this section we assume that the reward and transition functions have the same scope sets.) Here, D denotes the diameter of the FMDP (for a precise definition, see the footnote in Section 2). Tian et al. (2020) present two algorithms, F-EULER and F-UCBVI, which are extensions of UCBVI-CH (Gheshlaghi Azar et al., 2017) and EULER (Zanette and Brunskill, 2019) to FMDPs, respectively. In particular, F-EULER achieves a minimax-optimal regret of $\widetilde{\mathcal{O}}(\sum_{i=1}^m \sqrt{H|\mathcal{X}[Z_i]|T})$ for a rich class of structures, where H denotes the fixed episode length. In the average-reward setting, Xu and Tewari (2020) present two oracle-efficient algorithms, DORL and PSRL, which admit efficient implementations when an efficient oracle exists. DORL achieves a regret of $\widetilde{\mathcal{O}}(D\sum_{i=1}^m \sqrt{S_i|\mathcal{X}[Z_i]|T})$. The main objective in (Xu and Tewari, 2020) is to design a computationally efficient algorithm (with sublinear regret), for when an efficient oracle exists. RL in FMDPs with unknown structure are seldom studied in the literature. To the best of our knowledge, (Rosenberg and Mansour, 2020) is the only work presenting an algorithm with provable regret in FMDPs without any prior knowledge of the structure. The presented algorithm, SLF-UCRL, combines the structure learning method of (Strehl et al., 2007) with DORL (Xu and Tewari, 2020). Thus, it is oracle-efficient, like DORL. In contrast to (Xu and Tewari, 2020) and (Rosenberg and Mansour, 2020), we do not address the problem of efficient planning in FMDPs and instead aim for statistical efficiency from both theoretical and empirical standpoints.

We finally mention that some papers, notably (Zimmert and Seldin, 2018), study regret minimization in factored bandit problems, where the action-space is a Cartesian product of some atomic sets. Following Osband and Van Roy (2014a), recent literature on FMDPs (including the present paper) consider a factored action-space, which includes the Cartesian product as a special case. Nonetheless, the key feature making FMDPs suitable to model large decision problems is their factored dynamics. (In practice, the action-space may not be factored.) More importantly, the key challenge of RL in generic FMDPs is due to the factored transition function, for which the technical tools developed for factored bandit problems could not be directly used. We also stress that the factored bandit model of (Zimmert and Seldin, 2018) assumes a more restricted feedback than the bandit version of the FMDP model studied here.

Outline and Contributions. We introduce in Section 3 DBN-UCRL, a novel algorithm for average-reward RL in FMDPs, assuming a known factorization structure. DBN-UCRL is a model-based algorithm maintaining confidence sets for transition and reward functions. Specifically, it maintains tight Bernstein-type confidence sets for $P_i, i = 1, \dots, m$, in contrast to L_1-type confidence sets used in DORL and UCRL-Factored. On the theoretical side, we derive finite-time regret upper bounds for DBN-UCRL demonstrating the potential gain of using such confidence sets in terms of regret: For generic structures, we report a regret upper bound (in Theorem 1) scaling as $\widetilde{\mathcal{O}}\big(\sum_{i=1}^m \sqrt{\sum_{(s,a)\in\mathcal{X}[Z_i]} D_{i,s}^2 K_{i,s,a} T}\big)$, where $D_{i,s}$ is a notion of diameter termed *factored diameter* (Definition 4) and $K_{i,s,a}$ denotes the number of next-states for P_i under (s,a). DBN-UCRL achieves a strictly smaller regret than existing ones: (i) In contrast to previous bounds that depend on the (global) diameter D of the FMDP, this bound depends on the factored diameter which is tighter and problem-dependent; (ii) it improves the dependency of the regret on S_i to $K_{i,s,a}$. The factored diameter is always smaller than D: There exist cases, as illustrated in Section 4, where D may scale as $S := |\mathcal{S}|$, whereas $D_{i,s}$ could scale as $\max_a K_{i,s,a}$. Hence, $D_{i,s}$ could be exponentially (in m) smaller than D. Our second result concerns specific structures in the form of Cartesian products of some base MDPs. Theorem 2 shows that in Cartesian products, DBN-UCRL incurs the *sum* of regret of each underlying base MDP. This latter result signif-

[1] Besides this growing line of research, some papers study RL in episodic MDPs; see, e.g., (Gheshlaghi Azar et al., 2017; Dann et al., 2017).

[2] The notation $\widetilde{O}(.)$ hides poly-logarithmic terms in T.

icantly improves over previous regret bounds for the product case that were unable to establish a fully localized regret bound. This includes the bounds of (Osband and Van Roy, 2014b) and (Xu and Tewari, 2020) that would still depend on the global diameter D of the FMDP in this case. This leads to a term that is exponentially smaller (in m) than D. In Section 5, through numerical experiments, we show that on standard environments DBN-UCRL significantly outperforms other state-of-the-art algorithms that have frequentist regret guarantees.

Notations. We introduce some notations that will be used throughout. Given sets $\mathcal{X}$ and $\mathcal{S}$, let $\mathcal{R}_{\mathcal{X},[0,1]}$ be the set of all reward functions on $\mathcal{X}$ with image bounded in $[0,1]$, and let $\mathcal{P}_{\mathcal{X},\mathcal{S}}$ be the set of all transition functions from $\mathcal{X}$ to $\mathcal{S}$, i.e. $P \in \mathcal{P}_{\mathcal{X},\mathcal{S}}$ satisfies: For all $x \in \mathcal{X}$, $P(\cdot|x)$ is a probability distribution over $\mathcal{S}$, i.e. $P(s|x) \geq 0$ for all $s \in \mathcal{S}$ and $\sum_{s\in\mathcal{S}} P(s|x) = 1$. For a distribution q, $\mathrm{supp}(q)$ denotes the support set of q. For $n \in \mathbb{N}$, let $[n] := \{1,\dots,n\}$. $\mathbb{I}\{\cdot\}$ denotes the indicator function of an event.

2 PROBLEM FORMULATION

We study a learning task in a finite MDP $M = (\mathcal{S}, \mathcal{A}, P, R)$ under the average-reward criterion, where $\mathcal{S}$ denotes the set of states with cardinality S, $\mathcal{A}$ denotes the set of actions (available at each state) with cardinality A, and P and R denote the transition and reward functions, respectively. Choosing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ results in a transition to a state $s' \sim P(\cdot|s,a)$ and a reward drawn from $R(s,a)$, with mean $\mu(s,a)$. We assume that M is an FMDP, namely its transition and reward functions admit some conditional independence structure, as detailed below.

Factored Representations. To formally describe the factored structure, we introduce a few notations and definitions that are standard in the literature on FMDPs (see, e.g., Szita and Lőrincz (2009); Osband and Van Roy (2014b)). We begin by introducing the *scope* operator for a factored set.

Definition 1 (Scope Operator for a Factored Set $\mathcal{X}$) *Let $\mathcal{X} = \otimes_{i=1}^n \mathcal{X}_i$ be a finite factored set. For any subset of indices $Z \subseteq [n]$, we define $\mathcal{X}[Z] := \otimes_{i\in Z}\mathcal{X}_i$. Moreover, for any $x \in \mathcal{X}$, we let $x[Z] \in \mathcal{X}[Z]$ denote the value of the variables $x_i \in \mathcal{X}_i$ with indices $i \in Z$. For $i \in [n]$, we will write $x[i]$ as a shorthand for $x[\{i\}]$.*

An FMDP is represented by a tuple $M = \big(\{\mathcal{S}_i\}_{i\in[m]}, \{\mathcal{X}_i\}_{i\in[n]}, \{P_i\}_{i\in[m]}, \{Z_i^p\}_{i\in[m]}, \{R_i\}_{i\in[\ell]}, \{Z_i^r\}_{i\in[\ell]}\big)$, where $\mathcal{S}_i$ is the i-th state factor, $\mathcal{X}_i$ is the i-th state-action factor, P_i is the transition function associated with $\mathcal{S}_i$, R_i is the i-th reward function, and Z_i^p (resp. Z_i^r) denotes the scope set of P_i (resp. R_i) for $\mathcal{X} = \otimes_{i=1}^n \mathcal{X}_i$. The state-space is $\mathcal{S} = \otimes_{i=1}^m \mathcal{S}_i$ and the state-action space is $\mathcal{X} = \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_m \otimes \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_{n-m} = \mathcal{S} \times \mathcal{A}$, where $\mathcal{A} = \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_{n-m}$ is the (possibly factored) action-space.[3]

[3]Some authors have studied compact representations of the state-action space, such as decision trees, but we do not consider such representations in the present paper.

Definition 2 (Factored Reward Functions) *The class $\mathcal{R}$ of reward functions is factored over $\mathcal{X} = \otimes_{i=1}^n \mathcal{X}_i$ with scopes $Z_1^r, \dots, Z_\ell^r$ if and only if for all $R \in \mathcal{R}$ and $x \in \mathcal{X}$, there exist $\{R_i \in \mathcal{R}_{\mathcal{X}[Z_i^r],[0,1]}\}_{i\in[\ell]}$ such that any realization $r \sim R(x)$ implies $r = \sum_{i=1}^\ell r[i]$ with $r[i] \sim R_i(x[Z_i^r])$. Furthermore, let us define $r^{col} = \frac{1}{\ell}\sum_{i=1}^\ell r[i]$.*

Without loss of generality, we assume that the rewards of each factor are bounded in $[0,1]$. Note that the collected reward r^{col} is, by definition, bounded in $[0,1]$ too.

Definition 3 (Factored Transition Functions) *The class $\mathcal{P}$ of transition functions is factored over $\mathcal{X} = \otimes_{i=1}^n \mathcal{X}_i$ and $\mathcal{S} = \otimes_{i=1}^m \mathcal{S}_i$ with scopes $Z_1^p, \dots, Z_m^p$ if and only if for all $P \in \mathcal{P}$ and $x \in \mathcal{X}$ and $s \in \mathcal{S}$, there exist $\{P_i \in \mathcal{P}_{\mathcal{X}[Z_i^p],\mathcal{S}_i}\}_{i\in[m]}$ such that $P(s|x) = \prod_{i=1}^m P_i\big(s[i]\big|x[Z_i^p]\big)$.*

In order to clarify the presentation of confidence sets in the subsequent sections, we further introduce the following more compact representation of an FMDP. Let $\mathcal{G}_r = \big(\{\mathcal{X}_i\}_{i\in[n]}; \{Z_i^r\}_{i\in[\ell]}\big)$ and $\mathcal{G}_p = \big(\{\mathcal{X}_i\}_{i\in[n]}, \{\mathcal{S}_i\}_{i\in[m]}; \{Z_i^p\}_{i\in[m]}\big)$. We compactly represent an FMDP with structure $\mathcal{G} = \mathcal{G}_r \cup \mathcal{G}_p$ by a tuple $M = \big(\{P_i\}_{i\in[m]}, \{R_i\}_{i\in[\ell]}; \mathcal{G}\big)$, and let $\mathcal{G}(M)$ denotes its corresponding structure. We finally introduce the set $\mathbb{M}_{\mathcal{G}}$ of all FMDPs with structure $\mathcal{G}$:

$$\mathbb{M}_{\mathcal{G}} = \Big\{M = (P,R;\mathcal{G}) : P \in \mathcal{P}^{\mathrm{fac}}_{\mathcal{X},\mathcal{S}}(\mathcal{G}_p) \text{ and } R \in \mathcal{R}^{\mathrm{fac}}_{\mathcal{X},[0,1]}(\mathcal{G}_r)\Big\},$$

where $\mathcal{P}^{\mathrm{fac}}_{\mathcal{X},\mathcal{S}}(\mathcal{G}_p)$ (resp. $\mathcal{R}^{\mathrm{fac}}_{\mathcal{X},[0,1]}(\mathcal{G}_r)$) denotes the set of transition (resp. reward) functions satisfying Definition 3 (resp. Definition 2).

Remark 1 *An FMDP M can be represented by a Dynamic Bayesian Network (DBN) $\mathcal{B} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$, where $\mathcal{V}$ is a set of m discrete variables $\{v_i\}_{i\in[m]}$ and ℓ continuous variables $\{u_j\}_{j\in[\ell]}$, duplicated on two timeslices, $\mathcal{E}$ is a set of edges between the two timeslices, and $\mathcal{T}$ is a set of conditional probability tables (CPTs). In this case, $\mathcal{S}_i = \mathcal{D}(v_i)$ is the domain of variable v_i, $i \in [m]$, $\mathcal{X}[Z_i^p]$ (resp. $\mathcal{X}[Z_j^r]$) are the elements used to index the rows of the CPT of variable v_i (resp. u_j), and $\mathcal{X}[Z_i^p]$ (resp. $\mathcal{X}[Z_j^r]$) distinguishes between elements of $\mathcal{S}_k$, $k \in [m]$, if and only if $(v_k, v_i) \in \mathcal{E}$ (resp. $(v_k, u_j) \in \mathcal{E}$). In this context, $\mathcal{G}$ is the structure of the DBN $\mathcal{B}$ while P and R are the parameters of the CPTs in $\mathcal{T}$. By a slight abuse of terminology, we refer to $\mathcal{G}(M)$ as the DBN structure of the FMDP M.*

To help understand our notations, we provide an example of an FMDP, whose conditional independence structure is represented using the DBN shown in Figure 1. The state-space has $m = 4$ factors. For simplicity, we assume that all state factors are identical and equal to $\{a, b, c\}$ and that the action-space is non-factored. Nodes on the left-hand side of

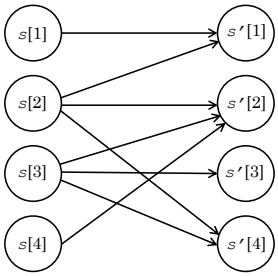

Figure 1: An example of a DBN characterizing the conditional independence structure of an FMDP.

the DBN correspond to the current state s, whereas those on the right-hand side represent the next state s', with each node representing a random variable corresponding to the value of a factor. For this DBN, the scopes are given by $Z_1^p = \{1,2\}$, $Z_2^p = \{2,3,4\}$, $Z_3^p = \{3\}$, and $Z_4^p = \{2,3\}$. Hence, for example, the resulting value of factor $\mathcal{S}_1$ is independent of factors $\mathcal{S}_3$ and $\mathcal{S}_4$. Furthermore, $\mathcal{X} = \{a,b,c\}^4 \times \mathcal{A}$, $\mathcal{X}[Z_1^p] = \{a,b,c\}^2 \times \mathcal{A}$, $\mathcal{X}[Z_2^p] = \{a,b,c\}^3 \times \mathcal{A}$, $\mathcal{X}[Z_3^p] = \{a,b,c\} \times \mathcal{A}$, and $\mathcal{X}[Z_4^p] = \{a,b,c\}^2 \times \mathcal{A}$.

Regret Minimization in FMDPs. We consider a finite FMDP $M = \big(\{P_i\}_{i\in[m]}, \{R_i\}_{i\in[\ell]}; \mathcal{G}\big)$ and the following RL task. An agent interacts with M for T rounds, starting in an initial state $s_1 \in \mathcal{S}$ chosen by Nature. At each time t, the agent is in state $s_t = (s_t[1], \ldots, s_t[m])$ and chooses an action a_t based on its observations so far. Let $x_t = (s_t, a_t)$ denote the state-action pair of the agent at time t. Then, (i) it receives a reward vector $r_t = (r_t[1], \ldots, r_t[\ell])$, where for each $i \in [\ell]$, $r_t[i] \sim R_i(x_t[Z_i^r])$; and (ii) Nature decides a next state $s_{t+1} = (s_{t+1}[1], \ldots, s_{t+1}[m])$ where for each $i \in [m]$, $s_{t+1}[i] \sim P_i(\cdot|x_t[Z_i^p])$. Let $r_t^{\text{col}} = \frac{1}{\ell}\sum_{i\in[\ell]} r_t[i]$ denote the normalized collected reward at time t.

The goal of the agent is to maximize the cumulative reward $\sum_{t=1}^T r_t^{\text{col}} = \sum_{t=1}^T \frac{1}{\ell}\sum_{i\in[\ell]} r_t[i]$, where T denotes the time horizon. We assume that the agent has a perfect knowledge about $\mathcal{G}$, but knows neither the transition function P nor the reward function R. It therefore has to learn them by trying different actions and recording the realized rewards and state transitions. The performance of the agent can be assessed resorting to the notion of regret. Following Jaksch et al. (2010), we define the regret of a learning agent (or algorithm) $\mathbb{A}$, after T steps and starting from an initial state $s_1 \in \mathcal{S}$, as:

$$\mathfrak{R}(T, \mathbb{A}, s_1) = Tg^\star(s_1) - \sum_{t=1}^T r_t^{\text{col}},$$

where $g^\star$ denotes the long-term average-reward (or gain) of M, in terms of r^{col}, starting from state s_1; we refer to (Puterman, 2014) for further details. Alternatively, the objective of the agent is to minimize the regret, which calls for balancing exploration and exploitation. In this paper we consider communicating MDPs, for which the gain does not depend on s_1, that is, $g^\star(s_1) = g^\star$ for all $s_1 \in \mathcal{S}$. We therefore define: $\mathfrak{R}(T, \mathbb{A}) = Tg^\star - \sum_{t=1}^T r_t^{\text{col}}$. The class of communicating MDPs arguably captures a big class of RL tasks of practical interest, and most literature on regret minimization in the average-reward setting has developed algorithms for this class. A notable property of communicating MDPs is having a finite *diameter*, as formalized in (Jaksch et al., 2010).[4]

3 The DBN-UCRL Algorithm

3.1 Confidence Sets for Factored MDPs

We begin with introducing empirical estimates and confidence sets used by DBN-UCRL. Throughout this section, for each given $Z \subseteq [n]$ and $x \in \mathcal{X}[Z]$, we let $\mathrm{N}(t, x; Z) := \max(\sum_{t'=1}^{t-1} \mathbb{I}\{x_{t'}[Z] = x\}, 1)$ denote the number of visits to x up to time t.

Empirical Estimates for Factored Representation. Let us consider time $t \geq 1$ and recall that $x_t = (s_t, a_t)$. For $i \in [m]$, we define the shorthand notation $N_{i,t}^p(x) := \mathrm{N}(t, x; Z_i^p)$. Likewise, for $i \in [\ell]$, we define $N_{i,t}^r(x) := \mathrm{N}(t, x; Z_i^r)$. We then introduce the following empirical estimates of transition and reward functions. Given $i \in [m]$ and $x \in \mathcal{X}[Z_i^p]$, we let $\widehat{P}_{i,t}(\cdot|x)$ be the empirical estimate of $P_i(\cdot|x)$ built using $N_{i,t}^p(x)$ i.i.d. samples from $P_i(\cdot|x)$:

$$\widehat{P}_{i,t}(y|x) := \frac{1}{N_{i,t}^p(x)} \sum_{t'=1}^{t-1} \mathbb{I}\{x_{t'}[Z_i^p] = x, s_{t'+1}[i] = y\}.$$

Similarly, given $i \in [\ell]$ and $x \in \mathcal{X}[Z_i^r]$, we define $\widehat{\mu}_{i,t}(x)$ as the empirical estimate of $R_i(x)$ built using $N_{i,t}^r(x)$ i.i.d. samples from $R_i(x)$:

$$\widehat{\mu}_{i,t}(x) := \frac{1}{N_{i,t}^r(x)} \sum_{t'=1}^{t-1} r_{t'}[i]\mathbb{I}\{x_{t'}[Z_i^r] = x\}.$$

Confidence Sets. We first define the confidence set for the reward function. For each $i \in [\ell]$ and $x \in \mathcal{X}[Z_i^r]$, we introduce the following entry-wise confidence set:

$$c_{t,\delta,i}(x) = \Bigg\{ q \in [0,1] : |\widehat{\mu}_{i,t}(x) - q| \leq \sqrt{\frac{2\widehat{\sigma}_{i,t}^2(x)}{N_{i,t}^r(x)}\beta_{N_{i,t}^r(x)}(\delta)} + \frac{7\beta_{N_{i,t}^r(x)}(\delta)}{3N_{i,t}^r(x)} \Bigg\},$$

where $\widehat{\sigma}_{i,t}^2(x)$ denotes the empirical variance of the reward function $R_i(x)$ built using $N_{i,t}^r(x)$ i.i.d. samples from $R_i(x)$, and for $n \in \mathbb{N}$ and $\delta \in (0,1)$, we define

$$\beta_n(\delta) := \eta \log\Big(\frac{\log(n)\log(\eta n)}{\log^2(\eta)\delta}\Big),$$

[4] Given an MDP M, the diameter $D := D(M)$ is defined as $D(M) := \max_{s\neq s'} \min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}[T^\pi(s,s')]$, where $T^\pi(s,s')$ denotes the number of steps it takes to get to s' starting from s and following policy π (Jaksch et al., 2010).

with $\eta = 1.12$. (In fact, any choice of $\eta > 1$ is valid, however $\eta = 1.12$ yields a small bound.[5]) The definition of the confidence set $c_{t,\delta,i}(x)$ is obtained using an empirical Bernstein concentration inequality (see, e.g., Maurer and Pontil (2009)), modified using a peeling technique to handle arbitrary random stopping times.[6] We also note that in the definition of $\beta_n(\delta)$, This leads us to define the following confidence set for mean rewards: For $x \in \mathcal{X}$,

$$\mathcal{C}^r_{t,\delta}(x) = \left\{\mu' \in \mathcal{R}^{\text{fac}}_{\mathcal{X},[0,1]}(\mathcal{G}_r) : \forall i \in [\ell], \mu'_i(x[Z^r_i]) \in c_{t,\delta_i,i}(x[Z^r_i])\right\},$$

where $\delta_i = \delta(\ell|\mathcal{X}[Z^r_i]|)^{-1}$.

As for the transition function, we define for each $i \in [m]$, $x \in \mathcal{X}[Z^p_i]$, and $y \in \mathcal{S}_i$ the following confidence set:

$$C_{t,\delta,i}(x,y) = \Big\{q \in [0,1] : |\widehat{P}_{i,t}(y|x) - q| \leq \sqrt{\frac{2q(1-q)}{N^p_{i,t}(x)}\beta_{N^p_{i,t}(x)}(\delta)} + \frac{\beta_{N^p_{i,t}(x)}(\delta)}{3N^p_{i,t}(x)}\Big\}.$$

This confidence set comes from a Bernstein concentration inequality as above.[7] Finally, we define the confidence set for P as follows: For $x \in \mathcal{X}$,

$$\mathcal{C}^p_{t,\delta}(x) = \Big\{P' \in \mathcal{P}^{\text{fac}}_{\mathcal{X},\mathcal{S}}(\mathcal{G}_p) : \forall i \in [m], \forall y \in \mathcal{S}_i, P'_i(y|x[Z^p_i]) \in C_{t,\delta_i,i}(x[Z^p_i], y)\Big\},$$

where $\delta_i = \delta(2mS_i|\mathcal{X}[Z^p_i]|)^{-1}$. We therefore define the following set of FMDPs that are plausible at time t:

$$\mathcal{M}_{t,\delta} = \Big\{M' = (\mathcal{S}, \mathcal{A}, P', R') \in \mathbb{M}_{\mathcal{G}(M)} : \mu'(x) \in \mathcal{C}^r_{t,\delta}(x) \text{ and } P'(\cdot|x) \in \mathcal{C}^p_{t,\delta}(x),\ \forall x \in \mathcal{X}\Big\}.$$

By construction of the confidence sets, the set $\mathcal{M}_{t,\delta}$ contains the true FMDP with high probability, and *uniformly* for all time horizons T: Formally, $\mathbb{P}\big(\exists t \in \mathbb{N}, M \notin \mathcal{M}_{t,\delta}\big) \leq 2\delta$. (We present a formal proof of this fact in Appendix B.)

3.2 DBN-UCRL: Pseudo-code

DBN-UCRL receives the structure $\mathcal{G}(M)$ of the true FMDP M as input. In order to implement the optimistic principle, DBN-UCRL considers the set $\mathcal{M}_{t,\delta}$ of plausible FMDPs and aims to compute the optimal policy $\overline{\pi}^+_t$ among all policies in all plausible FMDPs in $\mathcal{M}_{t,\delta}$, that is $\overline{\pi}^+_t = \operatorname{argmax}_{\pi:\mathcal{S}\to\mathcal{A}} \max\{g^M_\pi : M \in \mathcal{M}_{t,\delta}\}$, where g^M_π denotes the gain of policy π in M. This maximization can be solved approximately by the Extended Value Iteration (EVI) algorithm that builds a near-optimal policy π^+_t and an FMDP $\widetilde{M}_t$ such that $g^{\widetilde{M}_t}_{\pi^+_t} \geq \max_{\pi, M\in\mathcal{M}_{t,\delta}} g^M_\pi - \frac{1}{\sqrt{t}}$. Similarly to UCRL2 and its variants, DBN-UCRL proceeds in internal episodes $k = 1, 2, \ldots$, where a near-optimistic policy π^+_t is computed only at the starting time of each episode. Letting t_k denote the starting time of episode k, the algorithm computes $\pi^+_k := \pi^+_{t_k}$ and applies it until $t = t_{k+1} - 1$, where t_{k+1} is the first time step in which the number of observations gathered on some reward factor or transition factor within episode k is doubled. This event writes a bit differently in the FMDP setup. Namely, the sequence $(t_k)_{k\geq 1}$ is defined as follows: $t_1 = 1$, and for each $k > 1$

$$t_k = \min\Big\{t > t_{k-1} : \max\Big\{\max_{x\in\mathcal{X}[Z^p_i], i\in[m]} \frac{\nu^p_{i,t_{k-1}:t}(x)}{N^p_{i,t_{k-1}}(x)}, \max_{x\in\mathcal{X}[Z^r_i], i\in[\ell]} \frac{\nu^r_{i,t_{k-1}:t}(x)}{N^r_{i,t_{k-1}}(x)}\Big\} \geq 1\Big\},$$

where $\nu^p_{i,t_1:t_2}(x)$ (resp. $\nu^r_{i,t_1:t_2}(x)$) denotes the number of observations of $x \in \mathcal{X}[Z^p_i]$ (resp. of $x \in \mathcal{X}[Z^r_i]$) between time t_1 and t_2. The pseudo-code of DBN-UCRL is provided in Algorithm 1, which uses EVI (Algorithm 2) and InnerMax (Algorithm 3) as subroutines.

Algorithm 1 DBN-UCRL

Input: Structure $\mathcal{G}$, confidence parameter δ
Initialize: For all $i \in [m], x \in \mathcal{X}[Z^p_i]$, set $N^p_{i,0}(x) = 0$. For all $i \in [\ell], x \in \mathcal{X}[Z^r_i]$, set $N^r_{i,0}(x) = 0$. Set $t_0 = 0$, $t = 1$, $k = 1$.
for episodes $k = 1, 2, \ldots$ **do**
  Set $t_k = t$
  Compute empirical estimates $\{\widehat{\mu}_{i,t_k}(x)\}_{i\in[\ell],x\in\mathcal{X}[Z^r_i]}$ and $\{\widehat{P}_{i,t_k}(\cdot|x)\}_{i\in[m],x\in\mathcal{X}[Z^p_i]}$
  Compute $\pi^+_k = \texttt{EVI}\Big(\mathcal{M}_{t_k,\delta}, \frac{1}{\sqrt{t_k}}\Big)$ – see Algorithm 2
  Set $\nu^p_{i,k}(x) = 0$ for all $i \in [m]$ and $x \in \mathcal{X}[Z^p_i]$
  Set $\nu^r_{i,k}(x) = 0$ for all $i \in [\ell]$ and $x \in \mathcal{X}[Z^r_i]$
  continue = True
  while continue **do**
    Observe the current state s_t, play action $a_t = \pi^+_k(s_t)$, and observe reward $r_t = (r_t[1], \ldots, r_t[\ell])$. Set $x_t = (s_t, a_t)$
    Set $\begin{cases}\nu^p_{i,k}(x_t[Z^p_i]) = \nu^p_{i,k}(x_t[Z^p_i]) + 1, & i \in [m]\\ \nu^r_{i,k}(x_t[Z^r_i]) = \nu^r_{i,k}(x_t[Z^r_i]) + 1, & i \in [\ell]\end{cases}$
    continue $= \bigwedge_{i\in[m]}\big(\nu^r_{i,k}(x_t[Z^r_i]) < N^r_{i,t_k}(x_t[Z^r_i])\big) \wedge \bigwedge_{i\in[\ell]}\big(\nu^p_{i,k}(x_t[Z^p_i]) < N^p_{i,t_k}(x_t[Z^p_i])\big)$
    Set $t = t + 1$
  end while
  Set $\begin{cases}N^p_{i,t_k}(x) = N^p_{i,t_{k-1}}(x) + \nu^p_{i,k-1}(x), & i \in [m], x \in \mathcal{X}[Z^p_i]\\ N^r_{i,t_k}(x) = N^r_{i,t_{k-1}}(x) + \nu^r_{i,k-1}(x), & i \in [\ell], x \in \mathcal{X}[Z^r_i]\end{cases}$
end for

[5] The optimal η is obtained by optimizing $\beta_n(\delta)$ over η. The optimal η will depend on n, but as it turns out, the optimal η can be approximated well by a constant function $\eta = 1.12$ since $\beta_n(\delta)$ grows very slowly with n.

[6] We refer the interested reader to (Maillard, 2019) for the generic proof technique behind this result.

[7] We note that Bourel et al. (2020) define a similar Bernstein-type confidence set for the transition function of tabular (and non-factored) MDPs.

Algorithm 2 $\texttt{EVI}(\mathcal{M}, \varepsilon)$

Let $u_0 \equiv 0, u_{-1} \equiv -\infty, n = 0$
while $\max_s(u_n(s)-u_{n-1}(s)) - \min_s(u_n(s)-u_{n-1}(s)) > \varepsilon$ **do**
  Compute
  For all (s,a), compute $\tilde{\mu}(\cdot|s,a) = \max\{\mu'(s,a) : \mu' \in \mathcal{C}^r\}$.
  For all (s,a), compute $\widetilde{P}_n(\cdot|s,a)$ using $\texttt{InnerMax}(u_n, \mathcal{C}^p)$ – See Algorithm 3.
  Update $\begin{cases} u_{n+1}(s) = \max_{a\in\mathcal{A}}\left(\frac{\tilde{\mu}(s,a)}{\ell} + \sum_{y\in\mathcal{S}} \widetilde{P}_n(y|s,a)u_n(y)\right) \\ \pi^+_{n+1}(s) \in \operatorname{argmax}_{a\in\mathcal{A}}\left(\frac{\tilde{\mu}(s,a)}{\ell} + \sum_{y\in\mathcal{S}} \widetilde{P}_n(y|s,a)u_n(y)\right) \end{cases}$
  $n = n+1$
end while

Algorithm 3 $\texttt{InnerMax}(u, \mathcal{C}^p)$

Enumerate $\mathcal{S} = \{s_1, s_2, \ldots, s_S\}$ such that $u(s_1) \geq \ldots \geq u(s_S)$
Compute $\begin{cases} P_i^+(y[i]) \in \max\{q(y[i]) : q \in \mathcal{C}^p\}, & i \in [m] \\ P_i^-(y[i]) \in \min\{q(y[i]) : q \in \mathcal{C}^p\}, & i \in [m] \end{cases}$
Set $\begin{cases} P^+(y) = \prod_{i=1}^m P_i^+(y[i]) \\ P^-(y) = \prod_{i=1}^m P_i^-(y[i]) \end{cases}$
For $j \in [S]$, set $q(s_j) = \prod_{i=1}^m P_i^-(s_j[i])$
$l = 1$
while $\sum_{j=1}^S q(s_j) > 1$ and $l \leq S$ **do**
  $q' = \prod_{i=1}^m P_i^+(s_l[i])$
  $q(s_l) = \max\left(q(l) + \min\left(1 - \sum_{j=1}^S q(s_j), q' - q(s_l)\right), 1\right)$
  $l = l+1$
end while
Output q

4 DBN-UCRL: REGRET ANALYSIS

In this section, we present high-probability and finite-time regret upper bounds for DBN-UCRL. Our main result, presented in Theorem 1, is a regret upper bound assuming a generic structure $\mathcal{G}$. It is followed by Theorem 2 stating a substantially refined bound when the underlying structure $\mathcal{G}$ admits a Cartesian product form. Before presenting the theorems, we introduce a new notion of connectivity in FMDPs.

Factored Diameter. Theorem 1 relates the regret of DBN-UCRL to a new notion of connectivity in FDMPs, which we call the *factored diameter*. To present its definition, for $i \in [m]$ and $x \in \mathcal{X}[Z_i^p]$, we introduce: $\mathcal{K}_{i,x} := \text{supp}(P_i(\cdot|x))$ and $K_{i,x} := |\mathcal{K}_{i,x}|$.

Definition 4 (Factored Diameter) *The* factored diameter *of an FMDP M along factor $i \in [m]$ and for $u \in \mathcal{S}[Z_i^p]$, denoted by $D_{i,u} = D_{i,u}(M)$, is defined as*

$$D_{i,u} = \max_{s\in\mathcal{S}: s[Z_i^p]=u} \max_{s_1,s_2\in\mathcal{L}_s} \min_{\pi} \mathbb{E}[T^\pi(s_1,s_2)],$$

where $\mathcal{L}_s := \otimes_{i=1}^m (\cup_{a\in\mathcal{A}[Z_i^p]} \mathcal{K}_{i,s[Z_i^p],a})$, and where for s_1, s_2 with $s_1 \neq s_2$, $T^\pi(s_1,s_2)$ denotes the number of steps it takes to reach s_2 starting from s_1 by following policy π.

The notion of factored diameter refines that of diameter D of Jaksch et al. (2010) (see Section 2). It also extends the notion of local diameter introduced in (Bourel et al., 2020) for tabular MDPs to FMDPs, in the sense that in the absence of factorization (i.e., when the DBN is a complete graph), the factored diameter coincides with the local diameter in (Bourel et al., 2020). The factored diameter takes into account the joint support sets of factors, and is therefore a problem-dependent refinement of the global diameter D. In particular, for all i and $u \in \mathcal{S}[Z_i^p]$, $D_{i,u} \leq D$. Interestingly, there exist cases where $D_{i,y} \ll D$, as we illustrate through an example below, which is motivated by a multi-agent RL scenario.

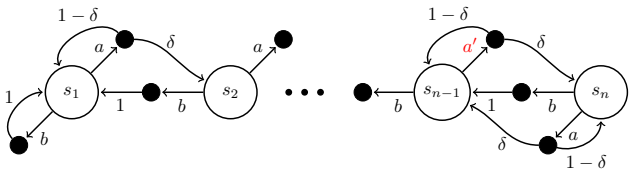

Figure 2: An Illustrative Example for the Factored Diameter.

Consider 2 agents, each independently interacting with an instance of the n-state MDP shown in Figure 2. Each agent $i \in \{1,2\}$ occupies a state in $\mathcal{S}_i = \{s_1, s_2, \ldots, s_n\}$ with $n > 2$, and the transition function P_i is defined according to the MDP shown in the figure. In each state $s \neq s_{n-1}$, each agent has access to two actions a and b. Only when both agents are *simultaneously* in s_{n-1}, they have access to an extra action a', which causes each agent to stochastically (but independently) transit to a high-reward state s_n — For instance, this could be relevant in scenarios where cooperation yields higher rewards. This scenario can be modeled using an FMDP, whose state-space is $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ and whose (state-dependent) action-space is: $\mathcal{A}_s = \{a,b\} \times \{a,b\}$ for all $s \neq (s_{n-1}, s_{n-1})$, and $\mathcal{A}_{(s_{n-1},s_{n-1})} = \{a',b\} \times \{a',b\}$.

For the case of a single agent, observe that $D = \frac{n-1}{\delta}$, as it takes $\frac{n-1}{\delta}$ steps in expectations to reach s_n from s_1 (namely, the worst-case travel time in $\mathcal{S}_i$). In the considered FMDP, it is easy to verify that $D = \left(\frac{n-1}{\delta}\right)^2$. The factored diameter here is at most $\frac{4}{\delta^2}$. Indeed, for $s = (s_i, s_j)$, we have $\mathcal{L}_s = \{s_{(i-1)\vee 1}, s_i, s_{(i+1)\wedge n}\} \times \{s_{(j-1)\vee 1}, s_j, s_{(j+1)\wedge n}\}$,[8] and one can verify that for any two states $u, v \in \mathcal{L}_s$, it takes at most $\frac{4}{\delta^2}$ steps in expectation to reach u starting from v – For details see Appendix E.

In this example, the factored diameter is smaller than D by a factor of $\mathcal{O}(n^2)$. Now, if we extend this to the case of m agents, the ratio of D and the factored diameter would be $\mathcal{O}(n^m)$. This further implies that DBN-UCRL achieves a much sharper regret than DORL and UCRL-Factored in these cases, whose regret bounds depend on D. Let us remark that such a massive reduction is a consequence of using Bernstein confidence sets for transition function that

[8] We use shorthands $a \vee b = \max\{a,b\}$ and $a \wedge b = \min\{a,b\}$.

takes into account the support of P_i. We however stress that in contrast to non-factored MDPs where a corresponding local diameter is straightforward to define (as done in (Bourel et al., 2020)), the task in FMDPs involves technical challenges for the decomposition of transition function along factors. To carefully exploit the gain of using Bernstein confidence intervals for P_i, we rely on the following factored deviation lemma, which is a refined variant of Lemma 1 in (Osband and Van Roy, 2014b), and whose proof is reported in Appendix A:

Lemma 1 *Let P and P' be two probability measures defined over $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_m$ such that for all $y = (y_1, \ldots, y_m) \in \mathcal{S}$: $P(y) = \prod_{i=1}^m P_i(y_i)$ and $P'(y) = \prod_{i=1}^m P'_i(y_i)$. Assume that for all $i \in [m]$, there exist $\xi_i > 0$ and $\xi'_i > 0$ such that $|(P'_i - P_i)(y_i)| \le \sqrt{P_i(y_i)\xi_i} + \xi'_i$, for all $y_i \in \mathcal{S}_i$. Then, for any function $f : \mathcal{S} \to \mathbb{R}_+$,*

$$\sum_{y\in\mathcal{S}} |(P-P')(y)| f(y) \le 3 \max_{y\in\mathcal{S}} f(y) \sum_{i=1}^m \xi'_i S_i + \max_{y \in \otimes_{i=1}^m \mathrm{supp}(P_i)} f(y) \sum_{i=1}^m \sum_{y_i \in \mathcal{S}_i} \sqrt{P_i(y_i)\xi_i}\,.$$

Regret Bound for Generic Structure. The following theorem presents a high-probability regret bound for DBN-UCRL under a generic and known structure:

Theorem 1 (Regret of DBN-UCRL) *Uniformly over all $T \ge 3$, with probability higher than $1-\delta$, it holds that*

$$\mathfrak{R}(\textit{DBN-UCRL}, T) \le \mathcal{O}\Big(c(M)\sqrt{T \log\big(\log(T)/\delta\big)} + D\Big(S_i \sum_{i=1}^m |\mathcal{X}[Z_i^p]| + \sum_{i=1}^{\ell} |\mathcal{X}[Z_i^r]|\Big) \log(T) \log\big(\log(T)/\delta\big)\Big),$$

with $c(M) = \sum_{i\in[m]} \sqrt{\sum_{(s,a)\in\mathcal{X}[Z_i^p]} D_{i,s}^2 (K_{i,s,a} - 1)} + \sum_{i\in[\ell]} \sqrt{|\mathcal{X}[Z_i^r]|} + D$.

In comparison, the regret of both UCRL-Factored and DORL satisfies $\widetilde{\mathcal{O}}(D \sum_{i=1}^m \sqrt{S_i |\mathcal{X}[Z_i^p]| T})$. The regret bound of DBN-UCRL improves over these regret bounds as for all $i \in [m]$ and $(s,a) \in \mathcal{X}[Z_i^p]$, we have $K_{i,s,a} \le S_i$ and $D_{i,s} \le D$. In view of $D_{i,s} \ll D$ in some FMDPs, this improvement can be substantial in some domains. We also demonstrate through numerical experiments on standard environments that DBN-UCRL is significantly superior to existing algorithms that admit frequentist regret guarantees. We finally note that Xu and Tewari (2020) presented another measure called the *factored span*, and present an algorithm following REGAL (Bartlett and Tewari, 2009), whose regret scales with the factored span (and not D). However, by design the presented algorithm crucially relies on knowing an upper bound on the factored span. The notions of factored diameter and factored span are not directly comparable. We however remark that the bound in Theorem 1 is achieved without any prior knowledge on the diameter.

The proof of Theorem 1 is provided in Appendix C. Similarly to most UCRL2-style algorithms, the proof of this theorem follows the machinery of the regret analysis in (Jaksch et al., 2010). However, to account for the underlying factored structure, as in the regret analyses in (Osband and Van Roy (2014b); Xu and Tewari (2020)), the proof decomposes the regret across factors. The algorithms in Osband and Van Roy (2014b); Xu and Tewari (2020) both rely on L_1-type confidence sets, as in UCRL2 (Jaksch et al., 2010), and their corresponding regret analyses proceed by decomposing an L_1 distance between two probability distributions to the sum of L_1 distances over various factors. This decomposition necessarily involves the global diameter D in the leading term of regret. In contrast, DBN-UCRL relies on Lemma 1, which carefully exploits the benefit of using the Bernstein-style confidence sets.

Regret Bound for Cartesian Products. We now focus on a structure $\mathcal{G}$ that can be represented as a Cartesian product, so that the true FMDP M can be seen as a Cartesian product of some base MDPs. Let the true FMDP M be a Cartesian product of m base MDPs, $M_i, i = 1, \ldots, m$, each with state-space $\mathcal{S}_i$, action-space $\mathcal{A}_i$, and diameter D_i.

Theorem 2 (Regret of DBN-UCRL for Cartesian products) *With probability higher than $1-\delta$, for all $T \ge 3$,*

$$\mathfrak{R}(\textit{DBN-UCRL}, T) \le \mathcal{O}\Big(\sum_{i=1}^m c_i \sqrt{T \log\big(\log(T)/\delta\big)} + \sum_{i=1}^m D_i S_i A_i \log(T) \log\big(\log(T)/\delta\big)\Big)\,, \quad \textit{with}$$

$$c_i = \sqrt{\textstyle\sum_{i=1}^m \sum_{s\in\mathcal{S}_i, a\in\mathcal{A}_i} D_{i,s}^2 K_{i,s,a}} + \sum_{i=1}^m \sqrt{S_i A_i} + D_i.$$

This result asserts that in the case of Cartesian products, the regret of DBN-UCRL boils down to the sum of individual regret of m base MDPs, where each individual term corresponds to a fully local quantity, i.e. depending only on the properties of M_i. This bound significantly improves over previous regret bounds for the product case, which were unable to establish a fully localized regret bound. In particular, the bounds of (Osband and Van Roy, 2014b) and (Xu and Tewari, 2020) for this case would necessarily depend on the *global diameter of the FMDP*, which might scale as $\prod_{i=1}^m D_i$, whereas ours in Theorem 2 depends on *the local diameter of the local MDPs*. This would in turn imply an exponentially (in the number m of base MDPs) tighter regret bound. Finally, we mention that Xu and Tewari (2020) present a regret lower bound scaling as $\Omega(\sqrt{bLT})$ in FMDPs based on worst-case Cartesian products, where L is an upper bound on both $|\mathcal{X}[Z_i^p]|$ and $|\mathcal{X}[Z_i^r]|$, and b denotes the span of the optimal bias function. Our regret bounds do not contradict this lower bound as $b \le \sum_i D_{i,s}$ for any s.

Remark 2 *Cartesian products might seem specific, but admittedly they represent the extreme case of FMDPs, where the individual MDPs are independent of one another. Hence, they are used in existing works (e.g., Osband and Van Roy (2014b); Xu and Tewari (2020)) to establish* best-case bounds *on exploration. The resulting bounds are typically more explicit than their corresponding complicated bounds for generic FMDPs. This allowed us to establish a best-case bound depending only in* fully local *quantities, in contrast to existing bounds above for Cartesian products. We believe that there is still value in analysing these special cases, and that analysing intermediate cases (in which individual MDPs are only* weakly connected*) is an important avenue for future work.*

We finally note that Theorem 2 cannot be directly obtained from Theorem 1, and its proof crucially relies on the following lemma stating that in FMDPs with Cartesian structures, the value function can be decomposed into the sum of individual value functions of the base MDPs:[9]

Lemma 2 (VI for Cartesian products) *Consider VI with* $u_0(s) = 0$*, and for each* $n \geq 0$*,* $u_{n+1}(s) = \max_{a\in\mathcal{A}} \{m^{-1}\mu(s,a) + \sum_{y\in\mathcal{S}} P(y|s,a)u_n(y)\}$*. Then, for all* n*,* $u_n(s) = m^{-1}\sum_{i=1}^{m} u_n^{(i)}(s[i])$*, where* $(u_n^{(i)})_{n\geq 0}$ *is a sequence of VI on MDP* i*, that is* $u_0^{(i)}(x) = 0$ *and* $u_{n+1}^{(i)}(x) = \max_{a\in\mathcal{A}_i} \{\mu_i(x,a) + \sum_{y\in\mathcal{S}_i} P_i(y|x,a)u_n^{(i)}(y)\}$ *for all* $x \in \mathcal{S}_i$ *and* $n \geq 0$*.*

5 NUMERICAL EXPERIMENTS

In this section, we present results from numerical experiments with DBN-UCRL. We perform experiments with the algorithm in two domains: Two factored versions of *RiverSwim* (Strehl and Littman, 2008; Filippi et al., 2010), and the *SysAdmin* domain (Guestrin et al., 2003).

We consider two factored versions of RiverSwim. In the first one, we construct an FMDP by taking the Cartesian product of two RiverSwim instances, with 6 states each, and introduce additional reward for a single joint state to couple the two instances through the reward factors. This corresponds to $S = 36$ and $A = 4$ (and so, $|\mathcal{X}| = 144$). We shall refer to this domain as Two-Layer RiverSwim. We construct the second factored version of RiverSwim by coupling three RiverSwim instances with 4 states each, in a similar fashion. This results in an FMDP with $S = 64$ and $A = 8$ (and hence, $|\mathcal{X}| = 512$), which we call Three-Layer RiverSwim. Recall that $\mathcal{X} = \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_m \otimes \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_{n-m}$, i.e. each factor scope $\mathcal{X}[Z_i^p]$ (resp. $\mathcal{X}[Z_i^r]$) is the Cartesian product of a subset of state and action factors. In other words, the agent knows which subset of factors is relevant for transition factor P_i (resp. reward factor R_i), but does not have access to a compact representation, e.g. in the form of a decision tree. Having access to such a compact representation would improve the performance of the algorithm but makes a stronger assumption on the available prior knowledge.

We compare DBN-UCRL to the following three algorithms[10]: UCRL-Factored (Osband and Van Roy, 2014b), the previous state-of-the-art algorithm for FMDPs, which is the natural extension of UCRL2 (Jaksch et al., 2010) to FMDPs; PSRL-Factored (Osband and Van Roy, 2014b), which is an algorithm based on posterior sampling and is in fact a natural extension of PSRL (Osband et al., 2013) to episodic FMDPs; and UCRLB-peeling, an improved variant of UCRL2 and UCRL2B (Fruit et al., 2020) relying on the same confidence sets for reward and transition functions as in DBN-UCRL but ignoring the factored structure. In particular, the comparison against UCRLB-peeling indicates the gain achieved by taking into account the factored structure, whereas that against UCRL-Factored reveals the gain of (element-wise) Bernstein-type confidence sets over their counterparts derived using Hoeffding's and Weissman's concentrations. As far as we know, ours is the first full-scale empirical evaluation of regret minimization algorithms for FMDPs. We stress that among these algorithms, PSRL-Factored is only shown to guarantee a Bayesian regret bound, and to the best of our knowledge, its frequentist regret analysis is still open. (We also refer to (Xu and Tewari, 2020) for a Bayesian regret analysis of PSRL-Factored in the average-reward setting.) Finally, to ease its implementation, in our experiments we let PSRL-Factored have access to the reward function.

In our experiments, we set $\delta = 0.01$ and report for each domain the average results over 50-100 independent experiments (depending on the domain), along with 95% confidence intervals. Figure 3 shows the regret of various algorithms against time in Two-Layer RiverSwim. (Note the logarithmic scale on the y-axis.) As the figure reveals, the regret under DBN-UCRL significantly improves over that of UCRL-Factored and UCRLB-peeling, and remains competitive with PSRL-Factored. However, we observe that the regret under PSRL-Factored has a very large variance, which is in stark contrast to the other algorithms. Finally note that both DBN-UCRL and PSRL-Factored enjoy a short *burn-in* phase (compared to UCRLB-peeling and UCRL-Factored), after which the regret grows sublinearly with time.

Figure 4 displays the regret of various algorithms against time in Three-Layer RiverSwim. The regret under DBN-UCRL significantly improves over that of UCRL-Factored and UCRLB-peeling, but is considerably worse than that of PSRL-Factored. Again, we see that the regret under PSRL-Factored has a high variance, although its average regret is smaller than the rest.

We perform two experiments in the SysAdmin domain. This domain consists of N computer servers that are organized

[9] A similar results for the bias function of the FMDP in the case of Cartesian product was provided in Xu and Tewari (2020), but not for Value Iteration (VI).

[10] The code is made publicly available via `https://github.com/aig-upf/dbn-ucrl`.

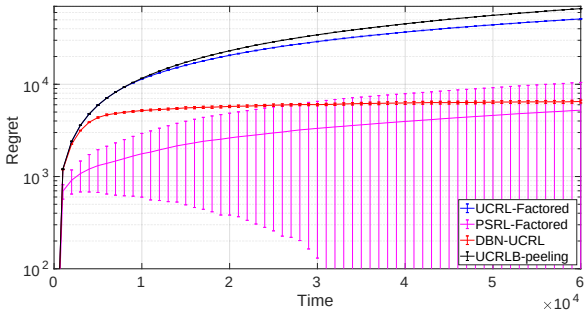

Figure 3: Regret in Two-Layer RiverSwim.

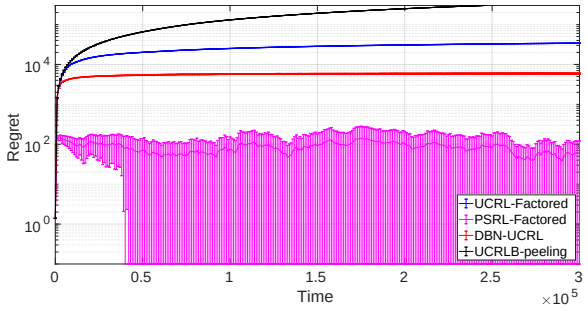

Figure 4: Regret in Three-Layer RiverSwim.

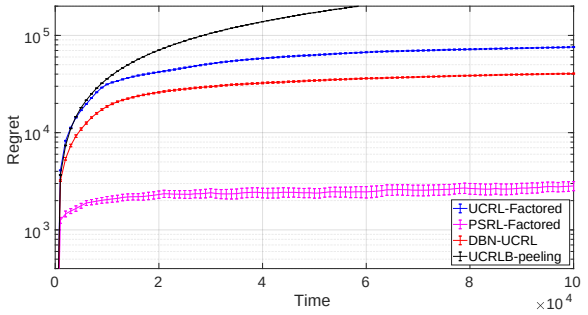

Figure 5: Regret in SysAdmin with the circular topology.

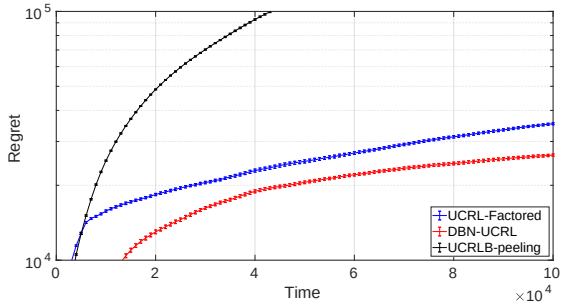

Figure 6: Regret in SysAdmin with the three-legged topology.

in a graph with a certain topology. Each server is represented by a binary variable that indicates whether or not it is working. At each time step, each server has a chance of failing, which depends on its own status and the status of the servers connected to it. There are $N+1$ actions: N actions for rebooting a server (after which it works with high probability) and an idle action. In previous work, researchers have performed experiments with two different topologies: A circular topology in which each server is connected to the next server in the circle, and a three-legged topology in which the servers are organized in a tree with three branches. In each topology, the status of each server depends on at most one other server.

Figures 5 and 6 show the regret of various algorithms in the SysAdmin domain for the two topologies, along with 95% confidence intervals. For each topology, $N = 7$, i.e. the circular topology has 7 servers arranged in a circle, and the three-legged topology has a root server and two servers on each of the three branches. Hence, the respective size of the state and action-space is $S = 2^7 = 128$ and $A = 8$, and so, $|\mathcal{X}| = 1024$. Note again the logarithmic scale on the y-axis. As in the other domains, DBN-UCRL clearly outperforms the other algorithms in terms of regret, but does worse than PSRL-Factored. (PSRL-Factored has a similar performance in both SysAdmin domains, so we only reported its regret for one.) Compared to the previous RiverSwim domains, the regret under PSRL-Factored has a much smaller variance.

In summary, in these experiments, DBN-UCRL significantly outperformed existing algorithms for which high-probability frequentist regret bounds exist. Furthermore, it incurred a worse average regret than PSRL-Factored in most domains, but the latter was shown to suffer from a large variance – In contrast to confident behavior of DBN-UCRL. We again remark that PSRL-Factored is only shown, to our knowledge, to guarantee a Bayesian regret bound, which is weaker than the corresponding high-probability frequentist regret bound.

6 CONCLUSIONS

We studied reinforcement learning under the average-reward criterion in a Factored Markov Decision Process (FMDP) with a known factorization structure, and introduced DBN-UCRL, an optimistic algorithm maintaining Bernstein-type confidence sets for individual elements of transition probabilities for each factor. We presented two high-probability regret bounds for DBN-UCRL, strictly improving existing regret bounds: The first one is valid for any factorization structure making appear the notion of factored diameter for FMDPs, whereas the second concerns structures taking the form of a Cartesian product. We also demonstrated through numerical experiments on standard environments that DBN-UCRL enjoys a significantly superior empirical regret than existing algorithms that admit frequentist regret guarantees. One interesting future direction is to derive regret lower bounds valid for FMDPs with a generic structure.

Acknowledgements

The authors would like to thank anonymous reviewers for their comments. Anders Jonsson is partially supported by Spanish grants PID2019-108141GB-I00 and PCIN-2017-082. Odalric-Ambrym Maillard is supported by CPER Nord-Pas-de-Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, Inria, Inria Scool, the French Agence Nationale de la Recherche (ANR) under grant ANR-16-CE40-0002 (the BADASS project), the MEL, the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF.

References

Peter L Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 35–42, 2009.

Hippolyte Bourel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1056–1066, 2020.

Craig Boutilier and Richard Dearden. Using abstractions for decision-theoretic planning with time constraints. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI)*, 1994.

Craig Boutilier, Thomas Dean, and Steve Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Doran Chakraborty and Peter Stone. Structure learning in ergodic factored MDPs without knowledge of the transition function's in-degree. In *Proceedings of the 28th International Conference on Machine Learning*, pages 737–744, 2011.

Xiaoyu Chen, Jiachen Hu, Lihong Li, and Liwei Wang. Efficient reinforcement learning in factored MDPs with application to constrained RL. In *International Conference on Learning Representations*, 2021.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30*, pages 5711–5721, 2017.

Thomas Degris, Olivier Sigaud, and Pierre Wuillemin. Learning the structure of factored Markov decision processes in reinforcement learning problems. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 257–264, 2006.

Carlos Diuk, Lihong Li, and Bethany Leffler. The adaptive k-Meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 249–256, 2009.

Dmitri Dolgov and Edmund Durfee. Symmetric approximate linear programming for factored MDPs with application to constrained problems. *Annals of Mathematics and Artificial Intelligence*, 47(3):273–293, 2006.

Richard Fikes and Nils Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. In *Proceedings of the 2nd International Joint Conference on Artificial Intelligence*, pages 608–620, 1971.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, 2010.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1578–1586, 2018.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of UCRL2 with empirical Bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272, 2017.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19: 399–468, 2003.

Zhaohan Daniel Guo and Emma Brunskill. Sample efficient learning with feature selection for factored MDPs. In *Proceedings of the 14th European Workshop on Reinforcement Learning*, 2018.

Assaf Hallak, François Schnitzler, Timothy Mann, and Shie Mannor. Off-policy model-based learning under unknown factored dynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 711–719, 2015.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 11:1563–1600, 2010.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, volume 16, pages 740–747, 1999a.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In *Proceedings of the 16th*

International Joint Conference on Artificial Intelligence, pages 740–747, 1999b.

Odalric-Ambrym Maillard. Mathematics of statistical sequential decision making. *Habilitation à Diriger des Recherches*, 2019.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *22nd Conference on Learning Theory (COLT)*, 2009.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems 27*, pages 1466–1474, 2014a.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014b.

Ian Osband, Dan Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26*, pages 3003–3011, 2013.

Pascal Poupart, Craig Boutilier, Relu Patrascu, and Dale Schuurmans. Piecewise linear value function approximation for factored MDPs. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI)*, pages 292–299, 2002.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Jian QIAN, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Exploration bonus for regret minimization in discrete and continuous average reward MDPs. In *Advances in Neural Information Processing Systems 32*, pages 4891–4900, 2019.

Aswin Raghavan, Roni Khardon, Prasad Tadepalli, and Alan Fern. Memory-efficient symbolic online planning for factored MDPs. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 732–741, 2015.

Aviv Rosenberg and Yishay Mansour. Oracle-efficient reinforcement learning in factored MDPs with unknown structure. *arXiv preprint arXiv:2009.05986*, 2020.

Alexander Strehl, Carlos Diuk, and Michael Littman. Efficient structure learning in factored-state MDPs. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI)*, pages 645–650, 2007.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74 (8):1309–1331, 2008.

István Szita and András Lőrincz. Factored value iteration converges. *Acta Cybernetica*, 18(4):615–635, 2008.

István Szita and András Lőrincz. Optimistic initialization and greediness lead to polynomial time learning in factored MDPs. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1001–1008, 2009.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 770–805, 2018.

Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored Markov decision processes. In *Advances in Neural Information Processing Systems 33*, 2020.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10170–10180. PMLR, 2020.

Ziping Xu and Ambuj Tewari. Near-optimal reinforcement learning in factored MDPs: Oracle-efficient algorithms for the non-episodic setting. In *Advances in Neural Information Processing Systems 33*, 2020.

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*, pages 7304–7312, 2019.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems 32*, pages 2823–2832, 2019.

Julian Zimmert and Yevgeny Seldin. Factored bandits. In *Advances in Neural Information Processing Systems 31*, pages 2840–2849, 2018.

A PROOF OF FACTORIZATION LEMMA (LEMMA 1)

In this section, we prove Lemma 1 (restated below), which provides a bound on factored deviations, and can be seen as a refined variant of Lemma 1 in Osband and Van Roy (2014b).

Lemma 1 (Restated) *Let P and P' be two probability measures defined over $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_m$ such that for all $y = (y_1, \ldots, y_m) \in \mathcal{S}$: $P(y) = \prod_{i=1}^m P_i(y_i)$ and $P'(y) = \prod_{i=1}^m P'_i(y_i)$. Assume that for all $i \in [m]$, there exist positive numbers ξ_i and ξ'_i such that*

$$|(P'_i - P_i)(y_i)| \leq \sqrt{P_i(y_i)\xi_i} + \xi'_i\,, \quad \forall y_i \in \mathcal{S}_i.$$

Then, for any function $f : \mathcal{S} \to \mathbb{R}_+$, we have

$$\sum_{y\in\mathcal{S}} |(P - P')(y)|f(y) \leq \max_{y\in\otimes_{i=1}^m \mathrm{supp}(P_i)} f(y) \sum_{i=1}^m \sum_{y_i\in\mathcal{S}_i} \sqrt{P_i(y_i)\xi_i} + 3\max_{y\in\mathcal{S}} f(y) \sum_{i=1}^m \xi'_i S_i\,,$$

where for a distribution q, $\mathrm{supp}(q)$ denotes the support set of q.

Proof. We prove the lemma by induction on m. For $m = 2$, we have

$$\begin{aligned}\sum_{y\in\mathcal{S}} |(P - P')(y)|f(y) &\leq \sum_{y_1}\sum_{y_2} |P_1(y_1)P_2(y_2) - P'_1(y_1)P'_2(y_2)|f(y)\\ &\leq \sum_{y_1}\sum_{y_2} P_1(y_1)|P_2(y_2) - P'_2(y_2)|f(y) + \sum_{y_1}\sum_{y_2} P'_2(y_2)|P_1(y_1) - P'_1(y_1)|f(y)\,.\end{aligned}$$

The first term is bounded as:

$$\begin{aligned}\sum_{y_1}\sum_{y_2} P_1(y_1)|P_2(y_2) - P'_2(y_2)|f(y) &\leq \sum_{y_2} |P_2(y_2) - P'_2(y_2)| \max_{y_1\in\mathrm{supp}(P_1)} f(y) \underbrace{\sum_{y_1} P_1(y_1)}_{=1}\\ &\leq \sum_{y_2} \max_{y_1\in\mathrm{supp}(P_1)} f(y)\Big(\sqrt{P_2(y_2)\xi_2} + \xi'_2\Big)\\ &\leq \max_{y_1\in\mathrm{supp}(P_1), y_2\in\mathrm{supp}(P_2)} f(y) \sum_{y_2} \sqrt{P_2(y_2)\xi_2} + \xi'_2 S_2 \max_y f(y)\,.\end{aligned}$$

For the second term, we have:

$$\begin{aligned}&\sum_{y_1}\sum_{y_2} P'_2(y_2)|P_1(y_1) - P'_1(y_1)|f(y)\\ &\quad= \sum_{y_1} |P_1(y_1) - P'_1(y_1)|\Big(\sum_{y_2\in\mathrm{supp}(P_2)} P'_2(y_2)f(y) + \sum_{y_2\notin\mathrm{supp}(P_2)} P'_2(y_2)f(y)\Big)\\ &\quad\leq \sum_{y_1} |P_1(y_1) - P'_1(y_1)|\Big(\max_{y_2\in\mathrm{supp}(P_2)} f(y) \sum_{y_2\in\mathrm{supp}(P_2)} P'_2(y_2) + \max_y f(y) \sum_{y_2\notin\mathrm{supp}(P_2)} P'_2(y_2)\Big)\\ &\quad\overset{(a)}{\leq} \sum_{y_1} |P_1(y_1) - P'_1(y_1)|\Big(\max_{y_2\in\mathrm{supp}(P_2)} f(y) + \xi'_2 S_2 \max_y f(y)\Big)\\ &\quad\overset{(b)}{\leq} \sum_{y_1} \Big(\sqrt{P_1(y_1)\xi_1} + \xi'_1\Big) \max_{y_2\in\mathrm{supp}(P_2)} f(y) + 2\xi'_2 S_2 \max_y f(y)\\ &\quad\leq \max_{y_1\in\mathrm{supp}(P_1), y_2\in\mathrm{supp}(P_2)} f(y) \sum_{y_1} \sqrt{P_1(y_1)\xi_1} + (\xi'_1 S_1 + 2\xi'_2 S_2) \max_y f(y)\,,\end{aligned}$$

where (a) follows from the fact that $P'_2(x) \leq \xi'_2$ for all $x \notin \mathrm{supp}(P_2)$, and where (b) uses $\sum_{y_1} |(P_1 - P'_1)(y_1)| \leq 2$. Hence,

$$\sum_{y\in\mathcal{S}} |(P - P')(y)|f(y) \leq \max_{y_i\in\otimes_{i=1}^2 \mathrm{supp}(P_i)} f(y) \sum_{i=1}^2 \sum_{y_i} \sqrt{P_i(y_i)\xi_i} + 3\max_y f(y) \sum_{i=1}^2 \xi'_i S_i\,.$$

Now assume that the induction hypothesis is correct for $m > 2$:

$$\sum_{y\in\mathcal{S}} |(P-P')(y)|f(y) \le \max_{y_i\in\otimes_{i=1}^m \mathrm{supp}(P_i)} f(y) \sum_{i=1}^m \sum_{y_i} \sqrt{P_i(y_i)\xi_i} + 3\max_y f(y) \sum_{i=1}^m \xi_i' S_i \,.$$

We then show that the above holds for $m+1$. To this aim, we define shorthand $y_{1:m} := y_1, \dots, y_m$, and let $q(y_{1:m}) = P_1(y_1)\cdots P_m(y_m)$ and $q'(y_{1:m}) = P_1'(y_1)\cdots P_m'(y_m)$. We have:

$$\begin{aligned}\sum_y |(P-P')|(y)f(y) &= \sum_y \Big|P_{m+1}(y_{m+1})q(y_{1:m}) - P'_{m+1}(y_{m+1})q'(y_{1:m})\Big| f(y) \\ &\le \sum_{y_{1:m}}\sum_{y_{m+1}} q(y_{1:m})|(P_{m+1}-P'_{m+1})(y_{m+1})|f(y) + \sum_{y_{1:m}}\sum_{y_{m+1}} P'_{m+1}(y_{m+1})|(q-q')(y_{1:m})|f(y)\,.\end{aligned}$$

The first term is bounded as:

$$\begin{aligned}\sum_{y_{m+1}}\sum_{y_{1:m}} q(y_{1:m})|(P_{m+1}-P'_{m+1})(y_{m+1})|f(y) &\le \sum_{y_{m+1}} |(P_{m+1}-P'_{m+1})(y_{m+1})| \max_{y_{1:m}\in\otimes_{i=1}^m \mathrm{supp}(P_i)} f(y) \underbrace{\sum_{y_{1:m}} q(y_{1:m})}_{=1} \\ &\le \sum_{y_{m+1}} \Big(\sqrt{P_{m+1}(y_{m+1})\xi_{m+1}} + \xi'_{m+1}\Big) \max_{y_{1:m}\in\otimes_{i=1}^m \mathrm{supp}(P_i)} f(y) \\ &\le \max_{y\in\otimes_{i=1}^{m+1}\mathrm{supp}(P_i)} f(y) \sum_{y_{m+1}} \sqrt{P_{m+1}(y_{m+1})\xi_{m+1}} + \xi'_{m+1} S_{m+1} \max_y f(y)\,. \end{aligned} \tag{1}$$

The second term is bounded as follows:

$$\begin{aligned}&\sum_{y_{m+1}}\sum_{y_{1:m}} P'_{m+1}(y_{m+1})|(q-q')(y_{1:m})|f(y) \\ &\le \sum_{y_{1:m}} |(q-q')(y_{1:m})|\Big(\sum_{y_{m+1}\in\mathrm{supp}(P_{m+1})} P'_{m+1}(y_{m+1})f(y) + \sum_{y_{m+1}\notin\mathrm{supp}(P_{m+1})} P'_{m+1}(y_{m+1})f(y)\Big) \\ &\le \sum_{y_{1:m}} |(q-q')(y_{1:m})| \max_{y_{m+1}\in\mathrm{supp}(P_{m+1})} f(y) + 2\xi'_{m+1}S_{m+1}\max_y f(y)\,.\end{aligned} \tag{2}$$

Note that the induction hypothesis implies

$$\sum_{y_{1:m}} |(q-q')(y_{1:m})| \max_{y_{m+1}\in\mathrm{supp}(P_{m+1})} f(y) \le \max_{y_i\in\otimes_{i=1}^{m+1}\mathrm{supp}(P_i)} f(y) \sum_{i=1}^m \sum_{y_i} \sqrt{P_i(y_i)\xi_i} + 3\max_y f(y) \sum_{i=1}^m \xi_i' S_i\,.$$

Putting this together with (2), and combining with (1) yield the desired result:

$$\sum_y |(P-P')(y)|f(y) \le \max_{y_i\in\otimes_{i=1}^{m+1}\mathrm{supp}(P_i)} f(y) \sum_{i=1}^{m+1} \sum_{y_i} \sqrt{P_i(y_i)\xi_i} + 3\max_y f(y) \sum_{i=1}^{m+1} \xi_i' S_i\,,$$

thus concluding the proof. □

B CONCENTRATION INEQUALITIES

In this section, for the sake of completeness, we collect some concentration inequalities used when constructing the set $\mathcal{M}_{t,\delta}$ of plausible MDPs in DBN-UCRL. The first lemma provides a time-uniform Bernstein-type concentration inequality for bounded random variables:

Lemma 3 *Let $Z = (Z_t)_{t\in\mathbb{N}}$ be a sequence of random variables generated by a predictable process, and $\mathcal{F} = (\mathcal{F}_t)_t$ be its natural filtration. Assume for all $t \in \mathbb{N}$, $|Z_t| \le b$ and $\mathbb{E}[Z_s^2|\mathcal{F}_{s-1}] \le v$ for some positive numbers v and b. Let n be an integer-valued (and possibly unbounded) random variable that is $\mathcal{F}$-measurable. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left[\exists n\in\mathbb{N},\ \frac{1}{n}\sum_{t=1}^n Z_t \ge \sqrt{\frac{2\beta_n(\delta)v}{n}} + \frac{\beta_n(\delta)b}{3n}\right] \le \delta\,,$$

$$\mathbb{P}\left[\exists n\in\mathbb{N},\ \frac{1}{n}\sum_{t=1}^n Z_t \le -\sqrt{\frac{2\beta_n(\delta)v}{n}} - \frac{\beta_n(\delta)b}{3n}\right] \le \delta\,,$$

where $\beta_n(\delta) := \eta \log\left(\frac{\log(n)\log(\eta n)}{\delta \log^2(\eta)}\right)$*, with* $\eta > 1$ *being an arbitrary parameter.*

The next lemma presents a time-uniform concentration for i.i.d. random variables supported in $[0,1]$:

Lemma 4 *Let* $Z = (Z_t)_{t\in\mathbb{N}}$ *be a sequence of i.i.d. random variables bounded in* $[0,1]$*, with mean* μ*. Then, for all* $\delta \in (0,1)$*, it holds*

$$\mathbb{P}\left[\exists n \in \mathbb{N},\ \mu - \frac{1}{n}\sum_{t=1}^{n} Z_t \geq \sqrt{\frac{\beta_n(\delta)}{2n}}\right] \leq \delta\,,$$
$$\mathbb{P}\left[\exists n \in \mathbb{N},\ \mu - \frac{1}{n}\sum_{t=1}^{n} Z_t \leq -\sqrt{\frac{\beta_n(\delta)}{2n}}\right] \leq \delta\,,$$

where $\beta_n(\delta) := \eta \log\left(\frac{\log(n)\log(\eta n)}{\delta \log^2(\eta)}\right)$*, with* $\eta > 1$ *being an arbitrary parameter.*

Both Lemma 3 and Lemma 4 are derived from Lemma 2.4 in (Maillard, 2019).

We also present the following lemma implying that the set of MDPs $\mathcal{M}_{t,\delta}$ contains the true MDP by high probability:

Lemma 5 *For any FMDP with rewards in* $\mathcal{R}^{fac}_{\mathcal{X},[0,1]}(\mathcal{G}_r)$*, and transition function in* $\mathcal{P}^{fac}_{\mathcal{X},\mathcal{S}}(\mathcal{G}_p)$*, for all* $\delta \in (0,1)$*, it holds*

$$\mathbb{P}\Big(\exists t \in \mathbb{N}, x \in \mathcal{X}, \quad \mu(x) \notin \mathcal{C}^r_{t,\delta}(x) \ \ \textit{or} \ \ P(\cdot|x) \notin \mathcal{C}^p_{t,\delta}(x)\Big) \leq 2\delta\,.$$

In particular, for all $T \in \mathbb{N}$*:* $\mathbb{P}\big(\exists t \in \mathbb{N} : M \notin \mathcal{M}_{t,\delta}\big) \leq 2\delta$*.*

Proof. First note that for any $i \in [\ell]$ and $x \in \mathcal{X}[Z_i^r]$, by a time-uniform version of (Maurer and Pontil, 2009, Theorem 10): $\mathbb{P}\big(\exists t \in \mathbb{N} : \mu_i(x) \notin c_{t,\delta,i}(x)\big) \leq \delta$. Using union bounds and recalling that $\delta_i = \delta(\ell|\mathcal{X}[Z_i^r]|)^{-1}$, it then follows that

$$\begin{aligned}
\mathbb{P}\big(\exists t \in \mathbb{N}, x \in \mathcal{X}, \mu(x) \notin \mathcal{C}^r_{t,\delta}(x)\big) &\leq \mathbb{P}\big(\exists t \in \mathbb{N}, i \in [\ell], x \in \mathcal{X}[Z_i^r], \mu_i(x) \notin c_{t,\delta_i,i}(x)\big) \\
&\leq \sum_{i\in[\ell]} \sum_{x\in\mathcal{X}[Z_i^r]} \mathbb{P}\big(\exists t \in \mathbb{N} : \mu_i(x) \notin c_{t,\delta_i,i}(x)\big) \\
&\leq \sum_{i\in[\ell]} \sum_{x\in\mathcal{X}[Z_i^r]} \frac{\delta}{\ell|\mathcal{X}[Z_i^r]|} = \delta.
\end{aligned}$$

Now we consider the case of transition function. For any $i \in [m]$, $x \in \mathcal{X}[Z_i^m]$, and $y \in \mathcal{S}_i$, Lemma 3 implies:

$$\mathbb{P}\big(\exists t \in \mathbb{N} : P_i(y|x) \notin C_{t,\delta,i}(x,y)\big) \leq 2\delta\,.$$

Using union bounds gives

$$\begin{aligned}
\mathbb{P}\big(\exists t \in \mathbb{N}, x \in \mathcal{X},\, P(\cdot|x) \notin \mathcal{C}^p_{t,\delta}(x)\big) &\leq \mathbb{P}\big(\exists t \in \mathbb{N}, i \in [m], x \in \mathcal{X}[Z_i^p], y \in \mathcal{S}_i,\, P_i(y|x) \notin C_{t,\delta_i,i}(x,y)\big) \\
&\leq \sum_{i\in[m]} \sum_{x\in\mathcal{X}[Z_i^p]} \sum_{y\in\mathcal{S}_i} \mathbb{P}\big(\exists t \in \mathbb{N} : P_i(y|x) \notin C_{t,\delta_i,i}(x,y)\big) \\
&\leq \sum_{i\in[m]} \sum_{x\in\mathcal{X}[Z_i^p]} \sum_{y\in\mathcal{S}_i} \frac{2\delta}{2mS_i|\mathcal{X}[Z_i^p]|} = \delta.
\end{aligned}$$

Putting together and taking a union bound complete the proof. □

C REGRET ANALYSIS: GENERIC STRUCTURES (THEOREM 1)

In this section, we prove Theorem 1. Our proof follows similar lines as in the proof of (Jaksch et al., 2010, Theorem 2). Let $\delta \in (0,1)$. To simplify notations, let us define the shorthand $J_k := J_{t_k}$ for a generic measurable random variable J that is fixed within a given episode k and omit its dependence on δ (for example, $\mathcal{M}_k := \mathcal{M}_{t_k,\delta}$). Denote by $K(T)$ the number of

episodes initiated by the algorithm up to time T. Given a pair $x = (s, a)$, let $N_t(x)$ denote the number of times x is visited by the algorithm up to time t. Furthermore, let $n_k(x)$ denote the number of times x is sampled in a given episode k.

By applying Lemma 4 and noting that $r_t[i] \in [0, 1]$, we deduce that

$$\mathfrak{R}(T) = \sum_{t=1}^{T} g^\star - \sum_{t=1}^{T} \frac{1}{\ell} \sum_{i=1}^{\ell} r_t[i] \leq \sum_{x \in \mathcal{X}} N_T(x) \Big(g^\star - \frac{1}{\ell} \sum_{i=1}^{\ell} \mu_i(x[Z_i^r]) \Big) + \sqrt{T \beta_T(\delta)/2}\,,$$

with probability at least $1 - \delta$. We have

$$\begin{aligned} \sum_{x} N_T(x) \Big(g^\star - \frac{1}{\ell} \sum_{i=1}^{\ell} \mu_i(x) \Big) &= \sum_{k=1}^{K(T)} \sum_{x} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\{x_t = x\} \Big(g^\star - \frac{1}{\ell} \sum_{i=1}^{\ell} \mu_i(x[Z_i^r]) \Big) \\ &= \sum_{k=1}^{K(T)} \sum_{x} n_k(x) \Big(g^\star - \frac{1}{\ell} \sum_{i=1}^{\ell} \mu_i(x[Z_i^r]) \Big)\,. \end{aligned}$$

Introducing $\Delta_k := \sum_{x \in \mathcal{X}} n_k(x) \big(g^\star - \frac{1}{\ell} \sum_{i=1}^{\ell} \mu_i(x[Z_i^r]) \big)$ for $1 \leq k \leq K(T)$, we get

$$\mathfrak{R}(T) \leq \sum_{k=1}^{K(T)} \Delta_k + \sqrt{T \beta_T(\delta)/2}\,,$$

with probability at least $1 - \delta$. A given episode k is called *good* if $M \in \mathcal{M}_k$ (that is, the set of plausible MDPs contains the true model), and *bad* otherwise. By Lemma 5, for all T, and for all episodes $k = 1, \ldots, K(T)$, the set $\mathcal{M}_k$ contains the true MDP with probability higher than $1 - 2\delta$. As a consequence, with probability at least $1 - 2\delta$, $\sum_{k=1}^{K(T)} \Delta_k \mathbb{I}\{M \notin \mathcal{M}_k\} = 0$.

To upper bound regret in good episodes, we closely follow (Jaksch et al., 2010) and decompose the regret to control the deviation of optimistic transition and reward functions from their true values. Consider a good episode k (hence, $M \in \mathcal{M}_k$). By choosing π_k^+ and $\widetilde{M}_k$, we get that

$$g_k := g_{\pi_k^+}^{\widetilde{M}_k} \geq g^\star - \frac{1}{\sqrt{t_k}}\,,$$

so that with probability greater than $1 - \delta$,

$$\Delta_k \leq \sum_{x \in \mathcal{X}} n_k(x) \Big(g_k - \frac{1}{\ell} \sum_{i=1}^{\ell} \mu_i(x[Z_i^r]) \Big) + \sum_{x \in \mathcal{X}} \frac{n_k(x)}{\sqrt{t_k}}\,. \tag{3}$$

Using the same argument as in the proof of (Jaksch et al., 2010, Theorem 2), the value function $u_{l,k}$ computed by `EVI` at iteration l satisfies: $\max_s u_{l,k}(s) - \min_s u_{l,k}(s) \leq D$. The convergence criterion of `EVI` implies

$$|u_{l+1,k}(s) - u_{l,k}(s) - g_k| \leq \frac{1}{\sqrt{t_k}}\,, \qquad \forall s \in \mathcal{S}\,. \tag{4}$$

Using the Bellman operator on the optimistic MDP, we have:

$$u_{l+1,k}(s) = \frac{1}{\ell} \sum_{i=1}^{\ell} \widetilde{\mu}_{i,k}((s, \pi_k^+(s))[Z_i^r]) + \sum_{s'} \widetilde{P}_k(s'|s, \pi_k^+(s)) u_{l,k}(s')\,.$$

Substituting this into (4) gives

$$\Big| \Big(g_k - \frac{1}{\ell} \sum_{i=1}^{\ell} \widetilde{\mu}_{i,k}((s, \pi_k^+(s))[Z_i^r]) \Big) - \Big(\sum_{s'} \widetilde{P}_k(s'|s, \pi_k^+(s)) u_{l,k}(s') - u_{l,k}(s) \Big) \Big| \leq \frac{1}{\sqrt{t_k}}\,, \qquad \forall s \in \mathcal{S}\,. \tag{5}$$

Now returning to (3), we can write

$$\Delta_k \leq \sum_{x} n_k(x) \Big(g_k - \frac{1}{\ell} \sum_{i=1}^{\ell} \widetilde{\mu}_{i,k}(x[Z_i^r]) \Big) + \frac{1}{\ell} \sum_{x} n_k(x) \sum_{i=1}^{\ell} \big(\widetilde{\mu}_{i,k}(x[Z_i^r]) - \mu_i(x[Z_i^r]) \big) + \sum_{x} \frac{n_k(x)}{\sqrt{t_k}}\,,$$

which, using (5), can be simplified as

$$\Delta_k \le \sum_s n_k(s,\pi_k^+(s))\Big(\sum_{s'}\widetilde{P}_k(s'|s,\pi_k^+(s))u_{l,k}(s') - u_{l,k}(s)\Big) + \frac{1}{\ell}\sum_x n_k(x)\sum_{i=1}^{\ell}(\widetilde{\mu}_{i,k}-\mu_i)(x[Z_i^r]) + 2\sum_x \frac{n_k(x)}{\sqrt{t_k}}$$

Defining $\mathbf{g}_k = g_k\mathbf{1}$, $\widetilde{\mathbf{P}}_k := \big(\widetilde{P}_k\big(s'|s,\pi_k^+(s)\big)\big)_{s,s'}$, and $n_k := \big(n_k\big(s,\pi_k^+(s)\big)\big)_s$, we can rewrite the above inequality as:

$$\Delta_k \le n_k(\widetilde{\mathbf{P}}_k - \mathbf{I})u_{l,k} + \frac{1}{\ell}\sum_x n_k(x)\sum_{i=1}^{\ell}(\widetilde{\mu}_{i,k}-\mu_i)(x[Z_i^r]) + 2\sum_x \frac{n_k(x)}{\sqrt{t_k}}\,.$$

Similarly to (Jaksch et al., 2010), we define $w_k(s) := u_{l,k}(s) - \frac{1}{2}(\min_s u_{l,k}(s) + \max_s u_{l,k}(s))$ for all $s \in \mathcal{S}$. Then, in view of the fact that $\widetilde{\mathbf{P}}_k$ is row-stochastic, we obtain

$$\Delta_k \le n_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k + \frac{1}{\ell}\sum_x n_k(x)\sum_{i=1}^{\ell}(\widetilde{\mu}_{i,k}-\mu_i)(x[Z_i^r]) + 2\sum_x \frac{n_k(x)}{\sqrt{t_k}}\,. \tag{6}$$

The second term in the right-hand side can be upper bounded as follows: $M \in \mathcal{M}_k$ implies that

$$\begin{aligned}\sum_{i=1}^{\ell}(\widetilde{\mu}_{i,k}-\mu_i)(x[Z_i^r]) &\le 2\sum_{i=1}^{\ell}\sqrt{\frac{2\widehat{\sigma}^2_{i,k}(x[Z_i^r])}{N^r_{i,t}(x[Z_i^r])}\beta_{N^r_{i,k}(x[Z_i^r])}\Big(\frac{\delta}{\ell|\mathcal{X}[Z_i^r]|}\Big)} + 2\sum_{i=1}^{\ell}\frac{7}{3N^r_{i,t}(x[Z_i^r])}\beta_{N^r_{i,t}(x[Z_i^r])}\Big(\frac{\delta}{\ell|\mathcal{X}[Z_i^r]|}\Big)\\ &\le \sum_{i=1}^{\ell}\sqrt{\frac{2}{N^r_{i,k}(x[Z_i^r])}\beta_T\Big(\frac{\delta}{\ell|\mathcal{X}[Z_i^r]|}\Big)} + \frac{14}{3}\sum_{i=1}^{\ell}\frac{1}{N^r_{i,k}(x[Z_i^r])}\beta_T\Big(\frac{\delta}{\ell|\mathcal{X}[Z_i^r]|}\Big)\,,\end{aligned}$$

where we have used $\widehat{\sigma}^2_{i,k}(x[Z_i^r]) \le \frac{1}{4}$ and $1 \le N^r_{i,k}(x[Z_i^r]) \le T$ in the last inequality. Furthermore, since $t_k \ge \max_{i\in[\ell]} N^r_{i,k}(x[Z_i^r])$, we have

$$\sum_x \frac{n_k(x)}{\sqrt{t_k}} \le \sum_x\sum_{i=1}^{\ell}\frac{n_k(x)}{\sqrt{N^r_{i,k}(x[Z_i^r])}}\,.$$

Putting together, and denoting $\beta_{T,i} := \beta_T\big(\frac{\delta}{\ell|\mathcal{X}[Z_i^r]|}\big)$, we obtain

$$\begin{aligned}\Delta_k &\le n_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k + \frac{\sqrt{2}}{\ell}\sum_{i=1}^{\ell}\sqrt{\beta_{T,i}}\sum_{x\in\mathcal{X}}\frac{n_k(x)}{\sqrt{N^r_{i,k}(x[Z_i^r])}} + \frac{14}{3\ell}\sum_{i=1}^{\ell}\beta_{T,i}\sum_{x\in\mathcal{X}}\frac{n_k(x)}{N^r_{i,k}(x[Z_i^r])}\\ &\le n_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k + \frac{\sqrt{2}}{\ell}\sum_{i=1}^{\ell}\sqrt{\beta_{T,i}}\sum_{x\in\mathcal{X}[Z_i^r]}\frac{\nu^r_{i,k}(x)}{\sqrt{N^r_{i,k}(x[Z_i^r])}} + \frac{14}{3\ell}\sum_{i=1}^{\ell}\beta_{T,i}\sum_{x\in\mathcal{X}[Z_i^r]}\frac{\nu^r_{i,k}(x)}{N^r_{i,k}(x[Z_i^r])}\,,\end{aligned} \tag{7}$$

where the last inequality follows from Lemma 10, stated and proven in Section C.3.

In what follows, we derive an upper bound on $n_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k$. Similarly to (Jaksch et al., 2010), we consider the following standard decomposition:

$$n_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k = \underbrace{n_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)w_k}_{L_1(k)} + \underbrace{n_k(\mathbf{P}_k - \mathbf{I})w_k}_{L_2(k)}\,.$$

The following lemmas provide upper bounds on $L_1(k)$ and $L_2(k)$:

Lemma 6 *Consider a good episode k. Then,*

$$L_1(k) \le 3\sum_{i=1}^{m}\sqrt{\beta'_{T,i}}\sum_{x=(s,a)\in\mathcal{X}[Z_i^p]}\nu^p_k(x)D_{i,s}\sqrt{\frac{K_{i,x}-1}{N^p_{i,k}(x)}} + 7\sum_{i=1}^{m}DS_i\beta'_{T,i}\sum_{x\in\mathcal{X}[Z_i^p]}\frac{\nu^p_{i,k}(x)}{N^p_{i,k}(x)}\,,$$

where $\beta'_{T,i} := \beta_T\big(\frac{\delta}{2mS_i|\mathcal{X}[Z_i^p]|}\big)$.

Lemma 7 *For all T, it holds with probability at least $1-\delta$,*

$$\sum_{k=1}^{K(T)} L_2(k)\mathbb{I}\{M\in\mathcal{M}_k\} \le D\sqrt{2T\beta_T(\delta)} + DK(T)\,.$$

Applying Lemmas 6 and 7, and summing over all good episodes, we obtain the following bound that holds with probability higher than $1-2\delta$, uniformly over all $T\in\mathbb{N}$:

$$\begin{aligned}
\sum_{k=1}^{K(T)} \Delta_k \mathbb{I}\{M\in\mathcal{M}_k\} &\le \sum_{k=1}^{K(T)} (L_1(k)+L_2(k)) + \frac{\sqrt{2}}{\ell}\sum_{i=1}^{\ell}\sum_{x\in\mathcal{X}[Z_i^r]} \frac{\sqrt{\beta_{T,i}}\nu_{i,k}^r(x)}{\sqrt{N_{i,k}^r(x[Z_i^r])}} + \frac{14}{3\ell}\sum_{i=1}^{\ell}\sum_{x\in\mathcal{X}[Z_i^r]}\frac{\beta_{T,i}\nu_{i,k}^r(x)}{N_{i,k}^r(x[Z_i^r])} \\
&\le 3\sum_{i=1}^{m}\sqrt{\beta'_{T,i}}\sum_{x=(s,a)\in\mathcal{X}[Z_i^p]} D_{i,s}\sqrt{K_{i,x}-1}\frac{\nu_{i,k}^p(x)}{\sqrt{N_{i,k}^p(x)}} + 7\sum_{i=1}^{m} DS_i\beta'_{T,i}\sum_{x\in\mathcal{X}[Z_i^p]}\frac{\nu_{i,k}^p(x)}{N_{i,k}^p(x)} \\
&+ \frac{\sqrt{2}}{\ell}\sum_{i=1}^{\ell}\sqrt{\beta_{T,i}}\sum_{x\in\mathcal{X}[Z_i^r]}\frac{\nu_{i,k}^r(x)}{\sqrt{N_{i,k}^r(x[Z_i^r])}} + \frac{14}{3\ell}\sum_{i=1}^{\ell}\beta_{T,i}\sum_{x\in\mathcal{X}[Z_i^r]}\frac{\nu_{i,k}^r(x)}{N_{i,k}^r(x[Z_i^r])} \\
&+ 2\sum_{i=1}^{\ell}\sum_{x\in\mathcal{X}[Z_i^r]}\beta'_{T,i}\frac{n_k(x)}{\sqrt{N_{i,k}^r(x[Z_i^r])}} + D\sqrt{2T\beta_T(\delta)} + DK(T)\,. \qquad (8)
\end{aligned}$$

To simplify the above bound, we provide the following lemma:

Lemma 8 *Consider a set $\mathcal{X}'$, and for $x\in\mathcal{X}'$, let $\nu_k(x)$ (resp. $N_k(x)$) denote the number of times x is oberserved in episode k (resp. before episode k starts). We have:*

$$(i)\qquad \sum_{x\in\mathcal{X}'}\sum_{k=1}^{K(T)}\frac{\nu_k(x)\alpha(x)}{\sqrt{N_k(x)}} \le \big(\sqrt{2}+1\big)\sqrt{T\sum_{x\in\mathcal{X}'}\alpha(x)}\,.$$

$$(ii)\qquad \sum_{x\in\mathcal{X}'}\sum_{k=1}^{K(T)}\frac{\nu_k(x)}{N_k(x)} \le 2|\mathcal{X}'|\log\big(\tfrac{T}{|\mathcal{X}'|}\big) + |\mathcal{X}'|\,.$$

Moreover, following the same steps in the proof of (Jaksch et al., 2010, Proposition 18) to upper bound the number of episodes, we obtain

$$K(T) = \mathcal{O}\Big(\Big[\sum_{i=1}^{m}|\mathcal{X}[Z_i^p]| + \sum_{i=1}^{\ell}|\mathcal{X}[Z_i^r]|\Big]\log(T)\Big)\,.$$

Putting everything together, it holds that with probability at least $1-5\delta$,

$$\begin{aligned}
\mathfrak{R}(T) &\le 8\sum_{i=1}^{m}\sqrt{\beta'_{T,i}\sum_{(s,a)\in\mathcal{X}[Z_i^p]}D_{i,s}^2(K_{i,s,a}-1)T} + \frac{4}{\ell}\sum_{i=1}^{\ell}\sqrt{\beta_{T,i}|\mathcal{X}[X_i^r]|T} \\
&+ \frac{6}{\ell}\sum_{i=1}^{\ell}\beta_{T,i}|\mathcal{X}[Z_i^r]|\log\big(\tfrac{T}{|\mathcal{X}[Z_i^r]|}\big) + 14\sum_{i=1}^{m}DS_i\beta'_{T,i}|\mathcal{X}[Z_i^p]|\log\big(\tfrac{T}{|\mathcal{X}[Z_i^p]|}\big) \\
&+ \Big(D\sqrt{2}+\sqrt{\tfrac{1}{2}}\Big)\sqrt{T\beta_T(\delta)} + \mathcal{O}\Big(D\Big[\sum_{i=1}^{m}|\mathcal{X}[Z_i^p]| + \sum_{i=1}^{\ell}|\mathcal{X}[Z_i^r]|\Big]\log(T)\Big)\,.
\end{aligned}$$

Noting that $\beta_T, \beta_{T,i}, \beta'_{T,i} = \mathcal{O}(\log(\log(T)/\delta))$ gives the desired result and completes the proof. □

C.1 Proof of Lemma 6

Recall that $L_1(k) = \sum_{s\in\mathcal{S}} n_k(s,\pi_k^+(s)) \sum_{y\in\mathcal{S}} \big(\widetilde{P}_k - P_k\big)\big(y|s,\pi_k^+(s)\big) w_k(y)$. Fix $s\in\mathcal{S}$, and introduce $x^k := (s,\pi_k^+(s))$. We have

$$\sum_{y\in\mathcal{S}} \big(\widetilde{P}_k - P_k\big)\big(y|s,\pi_k^+(s)\big) w_k(y) \le \underbrace{\sum_{y\in\mathcal{S}} \big|\big(\widehat{P}_k - P_k\big)\big(y|x^k\big)\big| |w_k(y)|}_{F_1} + \underbrace{\sum_{y\in\mathcal{S}} \big|\big(\widetilde{P}_k - \widehat{P}_k\big)\big(y|x^k\big)\big| |w_k(y)|}_{F_2} .$$

To upper bound F_1, recalling the definition $\mathcal{K}_{i,x} := \mathrm{supp}(P_i(\cdot|x))$, let us define

$$H_k(x^k) := \max_{y\in\otimes_{i=1}^m \mathcal{K}_{i,x^k[Z_i^p]}} |w_k(y)| ,$$

Now an application of Lemma 1 gives:

$$F_1 \le H_k(x^k) \sum_{i=1}^m \sqrt{\frac{2\beta'_{T,i}}{N_{i,k}^p(x^k[Z_i^p])}} \sum_{y[i]\in\mathcal{S}_i} \sqrt{P_i(1-P_i)(y[i]|x^k[Z_i^p])} + \sum_{i=1}^m \frac{DS_i\beta'_{T,i}}{2N_{i,k}^p(x^k[Z_i^p])} .$$

where we used that $\max_{y\in\mathcal{S}} |w_k(y)| \le \frac{D}{2}$ since $M\in\mathcal{M}_k$ (following a similar argument as in (Jaksch et al., 2010)), and where we used the shorthand $\beta'_{T,i} := \beta_T(\frac{\delta}{2mS_i|\mathcal{X}[Z_i^p]|})$.

By Cauchy-Schwarz, we get

$$\begin{aligned}
\sum_{y[i]\in\mathcal{S}_i} \sqrt{P_i(1-P_i)(y|x^k[Z_i^p])} &= \sum_{y[i]\in\mathcal{K}_{i,x^k[Z_i^p]}} \sqrt{P_i(1-P_i)(y[i]|x^k[Z_i^p])} \\
&\le \sqrt{\sum_{y[i]\in\mathcal{K}_{i,x^k[Z_i^p]}} P_i(y[i]|x^k[Z_i^p])} \sqrt{\sum_{y[i]\in\mathcal{K}_{i,x^k[Z_i^p]}} (1-P_i)(y[i]|x^k[Z_i])} \\
&\le \sqrt{K_{i,x^k[Z_i^p]} - 1} ,
\end{aligned}$$

so that

$$F_1 \le H_k(x^k) \sum_{i=1}^m \sqrt{\frac{2\beta'_{T,i}}{N_{i,k}^p(x^k[Z_i^p])}} \sqrt{K_{i,x^k[Z_i^p]} - 1} + \sum_{i=1}^m \frac{DS_i\beta'_{T,i}}{2N_{i,k}^p(x^k[Z_i^p])} .$$

To upper bound F_2, we will need the following lemma:

Lemma 9 ((Bourel et al., 2020)) *Consider x and y satisfying $|x-y| \le \sqrt{2y(1-y)\zeta} + \zeta/3$. Then,*

$$\sqrt{y(1-y)} \le \sqrt{x(1-x)} + 2.4\sqrt{\zeta} .$$

Applying this lemma twice, we have that for any y and x, if $|\widetilde{P}_{i,k} - \widehat{P}_{i,k}|(y|x) \le \sqrt{2\widetilde{P}_{i,k}(1-\widetilde{P}_{i,k})(y|x)\zeta} + \zeta'/3$, where ζ and ζ' come from the definition of $C_{t,\delta,i}(x,y)$, then

$$|\widetilde{P}_{i,k} - \widehat{P}_{i,k}|(y|x) \le \sqrt{2P_i(1-P_i)(y|x)\zeta} + 4\zeta'.$$

Hence, an application of Lemma 1 gives

$$F_2 \le H_k(x^k) \sum_{i=1}^m \sqrt{\frac{2\beta'_{T,i}}{N_{i,k}^p(x^k[Z_i^p])}} \sqrt{K_{i,x^k[Z_i^p]} - 1} + 6\sum_{i=1}^m \frac{DS_i\beta'_{T,i}}{N_{i,k}^p(x^k[Z_i^p])} .$$

Putting together, we get

$$L_1(k) \le 3\sum_{x\in\mathcal{X}} n_k(x) H_k(x) \sum_{i=1}^m \sqrt{\frac{\beta'_{T,i}}{N_{i,k}^p(x[Z_i^p])}} \sqrt{K_{i,x[Z_i^p]} - 1} + 7\sum_{x\in\mathcal{X}} \nu_k^p(x) \sum_{i=1}^m \frac{DS_i\beta'_{T,i}}{N_{i,k}^p(x[Z_i^p])} .$$

To control the right-hand side, we further show that given $x \in \mathcal{X}$,

$$H_k(x) \le \max_{u=(s,a):u[Z_i^p]=x} \max_{y \in \mathcal{L}_s} |w_k(y)| \le D_{i,s[Z_i^p]}, \quad \forall i \in [m].$$

where $\mathcal{L}_s := \otimes_{i=1}^m (\cup_{a \in \mathcal{A}[Z_i^p]} \mathcal{K}_{i,s[Z_i^p],a})$

The first inequality holds by the definition of H_k. To verify the second claim, we note that similarly to (Jaksch et al., 2010), we can combine all MDPs in $\mathcal{M}_k$ to form a single MDP $\widetilde{\mathcal{M}}_k$ with continuous action space $\mathcal{A}'$. In this extended MDP, in each state $s \in \mathcal{S}$, and for each $a \in \mathcal{A}$, there is an action in $\mathcal{A}'$ with mean $\tilde{\mu}(s,a)$ and transition $\widetilde{P}(\cdot|s,a)$ satisfying the definition of the set of plausible MDPs. Similarly to the arguments in (Jaksch et al., 2010), we recall that $u_{l,k}(s)$ amounts to the total expected l-step reward of an optimal non-stationary l-step policy starting in state s on the MDP $\widetilde{\mathcal{M}}_k$ with extended action set. Recall that we are in a case where $M \in \mathcal{M}_k$. This implies that the local diameter of factor i and state s of this extended MDP is at most $D_{i,s[Z_i^p]}$, since the actions of the true MDP are contained in the continuous action set of the extended MDP $\widetilde{\mathcal{M}}_k$. Let

$$\mathcal{B}_i := \Big\{ y \in \mathcal{S} : y[i] \in \cup_{a' \in \mathcal{A}[Z_i^p]} \mathcal{K}_{i,s[Z_i^p],a'} \text{ and } y[j] \in \mathcal{S}_j, i \neq j \Big\}.$$

Now, if there were states $s_1, s_2 \in \mathcal{B}_i$ with $u_{l,k}(s_1) - u_{l,k}(s_2) > D_{i,s[Z_i^p]}$, then an improved value for $u_{l,k}(s_1)$ could be achieved by the following non-stationary policy: First follow a policy which moves from s_1 to s_2 most quickly, which takes at most $D_{i,s[Z_i^p]}$ steps on average. Then follow the optimal l-step policy for s_2. We thus have $u_{l,k}(s_1) \ge u_{l,k}(s_2) - D_{i,s[Z_i^p]}$, since at most $D_{i,s[Z_i^p]}$ rewards of the policy for s_2 are missed. This is a contradiction, and so the claim follows.

Hence,

$$\begin{aligned} L_1(k) &\le 3 \sum_{i=1}^m \sum_{x \in \mathcal{X}} n_k(x) D_{i,s[Z_i^p]} \sqrt{\frac{\beta'_{T,i}}{N^p_{i,k}(x[Z_i^p])}} \sqrt{(K_{i,x[Z_i^p]} - 1)} + 7 \sum_{i=1}^m \sum_{x \in \mathcal{X}} n_k(x) \frac{D S_i \beta'_{T,i}}{N^p_{i,k}(x[Z_i^p])} \\ &\le 3 \sum_{i=1}^m \sum_{x=(s,a) \in \mathcal{X}[Z_i^p]} \nu^p_{i,k}(x) D_{i,s} \sqrt{\frac{\beta'_{T,i}(K_{i,x} - 1)}{N^p_{i,k}(x)}} + 7 \sum_{i=1}^m D S_i \beta'_{T,i} \sum_{x \in \mathcal{X}[Z_i^p]} \frac{\nu^p_{i,k}(x)}{N^p_{i,k}(x)}, \end{aligned}$$

where the last inequality follows from Lemma 10. □

C.2 Proof of Lemma 7

The proof follows similar steps as in the proof of (Jaksch et al., 2010, Theorem 2). Here, we collect all necessary arguments for completeness. Let us define the sequence $(X_t)_{t \ge 1}$ with $X_t := (P(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}}) w_{k(t)} \mathbb{I}\{M \in \mathcal{M}_{k(t)}\}$, for all t, where $k(t)$ denotes the episode containing step t. For any good k (i.e., $M \in \mathcal{M}_k$), as established in the proof of (Jaksch et al., 2010, Theorem 2), it holds that:

$$L_2(k) = \nu_k(\mathbf{P}_k - \mathbf{I}) w_k = \sum_{t=t_k}^{t_{k+1}-1} (P(\cdot|s_t,a_t) - \mathbf{e}_{s_t}) w_k = \sum_{t=t_k}^{t_{k+1}-1} X_t + w_k(s_{t+1}) - w_k(s_t) \le \sum_{t=t_k}^{t_{k+1}-1} X_t + D\,,$$

so that $\sum_{k=1}^{K(T)} L_2(k) \le \sum_{t=1}^T X_t + K(T) D$. Moreover, if k is a good episode, as shown in (Jaksch et al., 2010, Theorem 2), $|X_t| \le \|P(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}}\|_1 \frac{D}{2} \le D$. Further, $\mathbb{E}[X_t | s_1, a_1, \ldots, s_t, a_t] = 0$, so that $(X_t)_t$ is martingale difference sequence with $|X_t| \le D$. Therefore, applying Lemma 4 gives:

$$\mathbb{P}\Big(\exists T : \sum_{t=1}^T X_t \ge D \sqrt{2T \beta_T(\delta)} \Big) \le \delta\,.$$

Putting this together with the above bound on $\sum_{k=1}^{K(T)} L_2(k)$ gives the desired result. □

C.3 Other Supporting Lemmas

C.3.1 Proof of Lemma 8

Observe that for any sequence of numbers $z_1, z_2, \ldots, z_n$ with $0 \le z_k \le Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$, it holds

$$\sum_{k=1}^n \frac{z_k}{\sqrt{Z_{k-1}}} \le \big(\sqrt{2}+1\big)\sqrt{Z_n} \quad \text{and} \quad \sum_{k=1}^n \frac{z_k}{Z_{k-1}} \le 2\log(Z_n) + 1\,.$$

We refer to (Jaksch et al., 2010, Lemma 19) and (Talebi and Maillard, 2018, Lemma 24) for proof of these facts. The assertion of the lemma then easily follows by these facts and using Jensen's inequality. □

C.3.2 Lemma 10 and Its Proof

Lemma 10 *Let $Z_1, \dots, Z_m \subseteq [n]$, and for $x \in \mathcal{X}$, let $\nu_k(x)$ denote the local counts for some episode k. Then, for any i and any positive function $\alpha : \mathcal{X}[Z_i] \to \mathbb{R}$, we have:*

$$\sum_i \sum_{x \in \mathcal{X}} \nu_k(x)\alpha(x[Z_i]) \le \sum_i \sum_{x' \in \mathcal{X}[Z_i]} \nu_k(x')\alpha(x')\,.$$

Proof. We have:

$$\begin{aligned}
\sum_i \sum_{x \in \mathcal{X}} \nu_k(x)\alpha(x[Z_i]) &= \sum_i \sum_{x \in \mathcal{X}} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\{x_{t'} = x\}\alpha(x[Z_i]) \\
&= \sum_i \sum_{x' \in \mathcal{X}[Z_i]} \sum_{x'' \in \mathcal{X}[[n]\setminus Z_i]} \sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\{x_{t'}[[n] \setminus Z_i] = x''\}\mathbb{I}\{x_{t'}[Z_i] = x\}\alpha(x[Z_i]) \\
&= \sum_i \sum_{x' \in \mathcal{X}[Z_i]} \alpha(x') \sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\{x_{t'}[Z_i] = x\} \underbrace{\sum_{x'' \in \mathcal{X}[[n]\setminus Z_i]} \mathbb{I}\{x_{t'}[[n] \setminus Z_i] = x''\}}_{\le 1} \\
&\le \sum_i \sum_{x' \in \mathcal{X}[Z_i]} \nu_k(x')\alpha(x')\,.
\end{aligned}$$

□

C.3.3 Proof of Lemma 9

The proof is provided in (Bourel et al., 2020) and is presented here for completeness. By Taylor's expansion, we have

$$\begin{aligned}
y(1-y) &= x(1-x) + (1-2x)(y-x) - (y-x)^2 \\
&= x(1-x) + (1-x-y)(y-x) \\
&\le x(1-x) + |1-x-y|\left(\sqrt{2y(1-y)\zeta} + \tfrac{1}{3}\zeta\right) \\
&\le x(1-x) + \sqrt{2y(1-y)\zeta} + \tfrac{1}{3}\zeta\,.
\end{aligned}$$

Using the fact that $a \le b\sqrt{a} + c$ implies $a \le b^2 + b\sqrt{c} + c$ for nonnegative numbers a, b, and c, we get

$$\begin{aligned}
y(1-y) &\le x(1-x) + \tfrac{1}{3}\zeta + \sqrt{2\zeta\left(x(1-x) + \tfrac{1}{3}\zeta\right)} + 2\zeta \\
&\le x(1-x) + \sqrt{2\zeta x(1-x)} + 3.15\zeta \\
&= \left(\sqrt{x(1-x)} + \sqrt{\tfrac{1}{2}\zeta}\right)^2 + 2.65\zeta\,, \qquad (9)
\end{aligned}$$

where we have used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ valid for all $a, b \ge 0$. Taking square-root from both sides and using the latter inequality give the desired result:

$$\sqrt{y(1-y)} \le \sqrt{x(1-x)} + \sqrt{\tfrac{1}{2}\zeta} + \sqrt{2.65\zeta} \le \sqrt{x(1-x)} + 2.4\sqrt{\zeta}\,.$$

□

D REGRET ANALYSIS: CARTESIAN PRODUCTS

D.1 Proof of Lemma 2

Lemma 2 (Restated) *Consider VI with $u_0(s) = 0$, and for each $n \geq 0$, $u_{n+1}(s) = \max_{a\in\mathcal{A}} \big\{m^{-1}\mu(s,a) + \sum_{y\in\mathcal{S}} P(y|s,a)u_n(y)\big\}$. Then, for all n, $u_n(s) = m^{-1}\sum_{i=1}^m u_n^{(i)}(s[i])$, where $(u_n^{(i)})_{n\geq 0}$ is a sequence of VI on MDP i, that is $u_0^{(i)}(x) = 0$ and $u_{n+1}^{(i)}(x) = \max_{a\in\mathcal{A}_i}\big\{\mu_i(x,a) + \sum_{y\in\mathcal{S}_i} P_i(y|x,a)u_n^{(i)}(y)\big\}$ for all $x \in \mathcal{S}_i$ and $n \geq 0$.*

Proof. We prove the lemma by induction on n. Consider $u_0(s) = 0$. For $n = 1$, we have:

$$u_1(s) = \max_{a\in\mathcal{A}} \frac{\mu(s,a)}{m} = \max_{a[1],\dots,a[m]} \sum_i \frac{\mu_i(s[i],a[i])}{m} = \frac{1}{m}\sum_i \max_{a[i]\in\mathcal{A}_i} \mu_i(s[i],a[i]) = \frac{1}{m}\sum_i u_1^{(i)}\,.$$

Now assume that the induction hypothesis is correct for $n > 1$, that is, $u_n(s) = m^{-1}\sum_{i=1}^k u_n^{(i)}(s[i])$. We would like to show that $u_{n+1}(s) = m^{-1}\sum_{i=1}^k u_{n+1}^{(i)}(s[i])$. We have

$$\begin{aligned} u_{n+1}(s) &= \max_{a\in\mathcal{A}} \left\{\mu(s,a)/m + \sum_{y\in\mathcal{S}} P(y|s,a)u_n(y)\right\} \\ &= \max_{a[1],\dots,a[m]} \left\{\frac{1}{m}\sum_i \mu_i(s[i],a[i]) + \sum_{y_1\in\mathcal{S}^{(1)}} \dots \sum_{y_m\in\mathcal{S}_m} \prod_{i=1}^m P_i(y_i|s[i],a[i]) \sum_{j=1}^m \frac{u_n^{(j)}(y_j)}{m}\right\}. \end{aligned}$$

We have

$$\begin{aligned} \sum_{y_1\in\mathcal{S}_1} \dots \sum_{y_m\in\mathcal{S}_m} \prod_{i=1}^m P_i\big(y_i|s[i],a[i]\big) \sum_{j=1}^m u_n^{(j)}(y_j) &= \sum_{j=1}^m \sum_{y_j\in\mathcal{S}_j} P_j(y_j|s[j],a[j])u_n^{(j)}(y_j) \sum_{y\in\otimes_{i\neq j}\mathcal{S}_i} \prod_{i\neq j} P_i(y_i|s[i],a[i]) \\ &= \sum_{j=1}^m \sum_{y_j\in\mathcal{S}_j} P_j(y_j|s[j],a[j])u_n^{(j)}(y_j)\,. \end{aligned}$$

Hence,

$$\begin{aligned} mu_{n+1}(s) &= \max_{a[1],\dots,a[m]} \left\{\sum_i \mu_i(s[i],a[i]) + \sum_{i=1}^m \sum_{y_i\in\mathcal{S}_i} P_i(y_i|s[i],a[i])u_n^{(i)}(y_i)\right\} \\ &= \sum_{i=1}^m \max_{a[i]\in\mathcal{A}_i} \left\{\mu_i(s[i],a[i]) + \sum_{y_i\in\mathcal{S}_i} P_i(y_i|s[i],a[i])u_n^{(i)}(y_i)\right\} \\ &= \sum_{i=1}^m u_{n+1}^{(i)}(s[i])\,, \end{aligned}$$

thus concluding the lemma. □

D.2 Proof of Theorem 2

Without loss of generality, we assume $\mathcal{G}_r = \mathcal{G}_p$ (and in particular, $\ell = m$), and further assume that $(Z_i^r)_{i\in[\ell]}$ (and so, $(Z_i^p)_{i\in[m]}$) forms a partition of $[n]$ – Hence, with no loss of generality, the FMDP is assumed to be the product of m base MDPs. The proof can be straightforwardly extended to the case where $\mathcal{G}_r \neq \mathcal{G}_p$, at the expense of more complicated and tedious notations.

As $\mathcal{G}_r = \mathcal{G}_p$, in what follows we omit the dependence of scope sets or various quantities on r and p. We first make the following observation, which follows by straightforward calculations: $g^\star = \frac{1}{m}\sum_{i=1}^m g_i^\star$, where $g_i^\star$ denotes the optimal gain in M_i. We have

$$\mathfrak{R}(T) = Tg^\star - \sum_{t=1}^T \frac{1}{m}\sum_{i=1}^m r_t[i] = \frac{1}{m}\sum_{i=1}^m \big(Tg_i^\star - \sum_{t=1}^T r_t[i]\big)\,.$$

Following similar steps as in the proof of Theorem 1, we obtain that with probability at least $1-\delta$,

$$\mathfrak{R}(T) \leq \sum_{k=1}^{K(T)} \Delta_k + \sqrt{T\beta_T(\delta)/2}\,,$$

where Δ_k is defined similarly to the proof of Theorem 1. Now consider a good episode k. The corresponding optimistic policy π_k^+ and $\widetilde{M}_k$ satisfy (as in the proof of Theorem 1): $g_k := g_{\pi_k^+}^{\widetilde{M}_k} \geq g^\star - \frac{1}{\sqrt{t_k}}$. In particular, by construction $\widetilde{M}_k \in \mathbb{M}_{\mathcal{G}(M)}$, and thus, $g_k = \frac{1}{m}\sum_{i=1}^m g_k^{(i)}$, where $g_k^{(i)}$ denotes the gain of the restriction of policy π_k^+ to the base MDP i in $\widetilde{M}_k$. We therefore have:

$$\Delta_k \leq \sum_{x\in\mathcal{X}} \frac{n_k(x)}{m} \sum_{i=1}^m \big(g_k^{(i)} - \mu_i(x[Z_i])\big) + \sum_{x\in\mathcal{X}} \frac{n_k(x)}{\sqrt{t_k}}\,.$$

Leveraging the arguments in the proof of Theorem 1, and applying Lemma 2,[11] we observe that the value function $u_{l,k}$ computed by `EVI` at iteration l satisfies:

$$\begin{aligned} u_{l+1,k}(s) &= \frac{1}{m}\sum_{i=1}^m \widetilde{\mu}_{i,k}((s,\pi_k^+(s))[Z_i]) + \sum_{s'} \widetilde{P}_k(s'|s,\pi_k^+(s))u_{l,k}(s') \\ &= \frac{1}{m}\sum_{i=1}^m \Big(\widetilde{\mu}_{i,k}((s,\pi_k^+(s))[Z_i]) + \sum_{y_i\in\mathcal{S}_i} \widetilde{P}_{i,k}(y_i|(s,\pi_k^+(s))[Z_i])u_{l,k}^{(i)}(y_i)\Big)\,. \end{aligned}$$

Now the stopping criterion of `EVI` implies:

$$\Big|\sum_{i=1}^m \Big(g_{i,k} - \widetilde{\mu}_{i,k}((s,\pi_k^+(s))[Z_i]) - \sum_{y_i\in\mathcal{S}_i} \widetilde{P}_{i,k}(y_i|(s,\pi_k^+(s))[Z_i])u_{i,k}^{(l)}(y_i) + u_{i,k}^{(l)}(s[i])\Big)\Big| \leq \frac{m}{\sqrt{t_k}}\,, \qquad \forall s\in\mathcal{S}\,,$$

which, after following similar steps as in the proof of Theorem 1, gives

$$\begin{aligned} \Delta_k &\leq \frac{1}{m}\sum_x n_k(x)\sum_{i=1}^m \Big(g_k^{(i)} - \widetilde{\mu}_{i,k}(x[Z_i])\Big) + \frac{1}{m}\sum_x n_k(x)\sum_{i=1}^m \Big(\widetilde{\mu}_{i,k}(x[Z_i]) - \mu_i(x[Z_i])\Big) + 2\sum_x \frac{n_k(x)}{\sqrt{t_k}} \\ &\leq \frac{1}{m}\sum_x n_k(x)\sum_{i=1}^m \Big(g_k^{(i)} - \widetilde{\mu}_{i,k}(x[Z_i])\Big) + \frac{1}{m}\sum_x n_k(x)\sum_{i=1}^m \Big(\widetilde{\mu}_{i,k}(x[Z_i]) - \mu_i(x[Z_i])\Big) \\ &\quad + 2\sum_x\sum_{i=1}^m \frac{n_k(x)}{\sqrt{N_{i,k}(x[Z_i])}} \\ &\leq \frac{1}{m}\sum_{i=1}^m \sum_{x\in\mathcal{X}[Z_i]} \nu_{i,k}(x)\Big(g_k^{(i)} - \widetilde{\mu}_{i,k}(x)\Big) + \frac{1}{m}\sum_x\sum_{i=1}^m n_k(x)\frac{2\beta'_{T,i}}{\sqrt{N_{i,k}(x[Z_i])}} + 2\sum_{i=1}^m\sum_{x\in\mathcal{X}[Z_i]} \frac{\nu_{i,k}(x)}{\sqrt{N_{i,k}(x)}} \\ &\leq \frac{1}{m}\sum_{i=1}^m \nu_{i,k}(\widetilde{\mathbf{P}}_{i,k} - \mathbf{I})u_{l,k}^{(i)} + \frac{1}{m}\sum_{i=1}^m \beta'_{T,i}\sum_{x\in\mathcal{X}[Z_i]} \frac{\nu_{i,k}(x)}{\sqrt{N_{i,k}(x)}} + 2\sum_{i=1}^m\sum_{x\in\mathcal{X}[Z_i]} \frac{\nu_{i,k}(x)}{\sqrt{N_{i,k}(x)}}\,. \end{aligned}$$

Here, we used the fact that $t_k \geq \max_{i\in[m]} N_{i,k}(x[Z_i])$, and applied Lemma 10.

Now defining for each i, $w_{i,k}(s) := u_{l,k}^{(i)}(s) - \frac{1}{2}(\min_s u_{l,k}^{(i)}(s) + \max_s u_{l,k}^{(i)}(s))$ for all $s\in\mathcal{S}_i$, we arrive at $\Delta_k \leq \frac{1}{m}\sum_{i=1}^m \widetilde{\Delta}_{i,k}$ with

$$\widetilde{\Delta}_{i,k} := \nu_{i,k}(\widetilde{\mathbf{P}}_{i,k} - \mathbf{I})u_{l,k}^{(i)} + \beta'_{T,i}\sum_{x\in\mathcal{X}[Z_i]} \frac{\nu_{i,k}(x)}{\sqrt{N_{i,k}(x)}} + 2\sum_{x\in\mathcal{X}[Z_i]} \frac{\nu_{i,k}(x)}{\sqrt{N_{i,k}(x)}}\,.$$

In other words, the regret is upper bounded by a quantity, which only depends on the properties of MDP M_i. Following exact same arguments as in the rest of proof of Theorem 1 (or similarly, those in the proof of Thoerem 1 in (Bourel et al., 2020)) gives the desired result. □

[11] We stress that Lemma 2 applies to `EVI` as well, as the inner maximization of `EVI` is guaranteed to return a factored transition function.

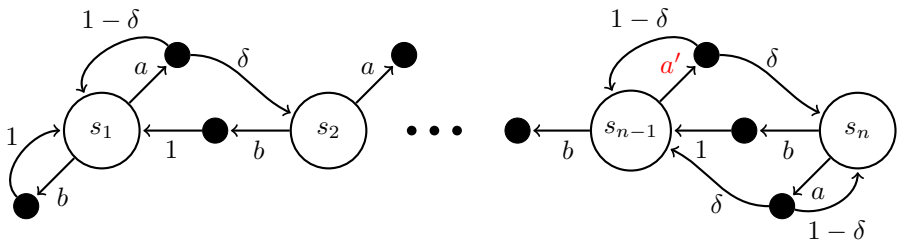

Figure 7: Global vs. Factored Diameter

E DETAILS OF THE EXAMPLE FOR THE FACTORED DIAMETER

Let us first consider a single agent. In this case, by definition, action a' is absent. The form of the transition function here allows us to derive closed-form expressions for the notions of diameters. In particular, the (global) diameter is $D = \frac{n-1}{\delta}$, as it takes $\frac{n-1}{\delta}$ steps in expectations to reach s_n from s_1 (this is the worst-case shortest path between any pair of states). Moreover, the local diameter (Bourel et al., 2020) is upper bounded by $2/\delta$: For any state s, this is the worst-case shortest path between any pair of states taken among the possible next-states of s.

Now we consider the case of 2 agents, where each agent independently interacts with an instance of the n-state MDP shown in Figure 7. Each agent $i \in \{1, 2\}$ occupies a state in $\mathcal{S}_i = \{s_1, s_2, \dots, s_n\}$ with $n > 2$, and the transition function P_i is defined according to the MDP shown in the figure. In each state $s \neq s_{n-1}$, each agent has access to two actions a and b. Only when both agents are *simultaneously* in s_{n-1}, they have access to an extra action a', which causes each agent to stochastically (but independently) transit to a high-reward state s_n — For instance, this could be relevant in scenarios where cooperation yields higher rewards. This scenario can be modeled using an FMDP as follows. The state-space is $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ and the action-space is state-dependent: in each state $s \neq (s_{n-1}, s_{n-1})$, $\mathcal{A}_s = \{a, b\} \times \{a, b\}$, whereas in $s = (s_{n-1}, s_{n-1})$, $\mathcal{A}_s = \{a', b\} \times \{a', b\}$.

The state space of the full MDP (modeling the interactions of both agents with the environment) is denoted by $\mathcal{S}$:

$$\mathcal{S} = \Big\{ s = (s[1], s[2]) : s[1] \in \mathcal{S}_1, s[2] \in \mathcal{S}_2 \Big\}$$

It thus has $|\mathcal{S}_1| \times |\mathcal{S}_2| = n^2$ states. For example, the possible transitions at the state s, where $s[1] = s_1$ and $s[2] = s_1$ would be as follows. Under action (a, a): the next is (s_1, s_1) w.p. $(1-\delta)^2$, or (s_2, s_1) w.p. $\delta(1-\delta)$, or (s_1, s_2) w.p. $\delta(1-\delta)$, or (s_2, s_2) w.p. δ^2. Under action (b, b), the next state is (s_1, s_1) w.p. 1. Under action (b, a) the next state is (s_1, s_2) w.p. δ or (s_1, s_1) w.p. $1-\delta$. Finally, under action (a, b), the next state is (s_2, s_1) w.p. δ or (s_1, s_1) w.p. $1-\delta$.

In the full MDP, it is easy to verify that $D = \left(\frac{n-1}{\delta}\right)^2$. But the local diameter of the full MDP (which corresponds to the factored diameter of the FMDP) is at most $\frac{4}{\delta^2}$. Recalling the definition of the factored diameter, it is straightforward to see that, for $s = (s_i, s_j)$, we have $\mathcal{L}_s = \{s_{(i-1)\vee 1}, s_i, s_{(i+1)\wedge n}\} \times \{s_{(j-1)\vee 1}, s_j, s_{(j+1)\wedge n}\}$ — We use shorthands $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. So one can verify that for any two states $u, v \in \mathcal{L}_s$, it takes at most $\frac{4}{\delta^2}$ steps in expectation to reach u starting from v.