Improved Exploration in Factored Average-Reward MDPs

Talebi, Sadegh; Jonsson, Anders; Maillard, Odalric-Ambrym

Published in:
Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)

Publication date:
2021

Document version
Publisher’s PDF, also known as Version of record

Citation for published version (APA):
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 \textit{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, \textit{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 \textit{DBN-UCRL} is upper bounded by the sum of regret of the base MDPs. We demonstrate, through numerical experiments on standard environments, that \textit{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: \textit{Exploration}, i.e. discovering the effects of actions on the environment, and \textit{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 \textit{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 \( \mathcal{A} = \otimes_{i=1}^{n} \mathcal{A}_i \) 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 \( Z_i \subset \{1, \ldots, n\} \), termed the \textit{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 \textit{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 Dean, 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
basis functions (Guestrin et al. 2003; Dolgov and Durfee 2006; Szita and Lörincz 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; Srebril et al. 2007; Dukl 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-regret criterion, in an intermediate setting where the underlying structure of the FMDP 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).

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 Lörincz (2009) study RL in discounted FMDPs presenting DBN-UCRL 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 \( \tilde{O}(D \sum_{i=1}^m N_i^{1/2}) \) after \( T \) steps. (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 2.2). 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 \( \tilde{O}(\sum_{i=1}^m N_i^{1/2}) \) for a rich class of structures, where \( H \) denotes the fixed episode length. In the average-regret 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 \( \tilde{O}(D \sum_{i=1}^m N_i^{1/2}) \). 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 (Srebril 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 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 in terms of regret or sample complexity from both theoretical and empirical standpoints.

DBN-UCRL, a novel algorithm for average-reward RL in FMDPs, assuming a known factorization structure.
We study a learning task in a finite MDP $\mathcal{M}$. The state-space is $\mathcal{X} = \{s_1, \ldots, s_m\}$ with scopes $Z_i^1, \ldots, Z_i^p$ if and only if for all $R \in \mathcal{R}$ and $x \in \mathcal{X}$, there exist $\{R_i \in \mathcal{R}_i \mid i \in [\mathcal{E}]\}$ such that any realization $r \sim R(x)$ implies $r = \sum_{i=1}^{|\mathcal{E}|} r[i] = \sum_{i=1}^{|\mathcal{E}|} r[i] \sim R_i(x[Z_i^1])$. Furthermore, let us define $r^{col} = \frac{1}{|\mathcal{E}|} \sum_{i=1}^{|\mathcal{E}|} r[i]$. 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 $\mathcal{X} = \bigotimes_{i=1}^n X_i$ with scopes $Z_i^1, \ldots, Z_i^p$ if and only if for all $R \in \mathcal{R}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$, there exist $\{P_i \in \mathcal{P}_i \mid i \in [\mathcal{E}]\}$ such that $P(s|x) = \prod_{i=1}^{|\mathcal{E}|} P_i(s[i]|x[Z_i^1])$.

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 $G_r = \{(X_i)_{i \in [n]}, (Z_i^1), \ldots, Z_i^p\}$, and $G_p = \{(X_i)_{i \in [n]}, (S_i)_{i \in [n]}, (Z_i^p)_{i \in [m]}\}$. We compactly represent an FMDP with structure $G = G_r \cup G_p$ by a tuple $M = \{(P_i)_{i \in [m]}, (R_i)_{i \in [\mathcal{E}]}, G\}$, and let $G(M)$ denotes its corresponding structure. We finally introduce the set $\mathcal{M}_G$ of all FMDPs with structure $G$:

$$\mathcal{M}_G = \left\{ M = (P, R, G) : P \in \mathcal{P}_{\mathcal{X}, S}(G_p) \text{ and } R \in \mathcal{R}_{\mathcal{X}, \mathcal{S}, [0,1]}(G_r) \right\},$$

where $\mathcal{P}_{\mathcal{X}, S}(G_p)$ (resp. $\mathcal{R}_{\mathcal{X}, \mathcal{S}, [0,1]}(G_r)$) 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) $B = (V, E, T)$, where $V$ is a set of $m$ discrete variables $\{v_i\}_{i \in [m]}$ and $t$ continuous variables $\{u_i\}_{i \in [n]}$, duplicated on two timeslices. $E$ is a set of edges between the two timeslices, and $T$ is a set of conditional probability tables (CPTs). In this case, $S_i = D(v_i)$ is the domain of variable $v_i$, $i \in [m]$, $X[Z_i^1]$ (resp. $X[Z_i^p]$) are the elements used to index the rows of the CPT of variable $v_i$ (resp. $u_i$), and $X[Z_i^1]$ (resp. $X[Z_i^p]$) distinguishes between elements of $S_k$, $k \in [m]$, if and only if $(v_k, v_i) \in E$ (resp. $(v_k, u_i) \in E$). In this context, $G$ is the structure of the DBN $B$ while $P$ and $R$ are the parameters of the CPT in $T$. By a slight abuse of terminology, we refer to $G(M)$ 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 state-action space, such as decision trees, but we do not consider such representations in the present paper.
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 $S_1$ is independent of factors $S_3$ and $S_4$. Furthermore, $\mathcal{X} = \{a, b, c\}^4 \times A$, $\mathcal{X}'[Z_1^p] = \{a, b, c\}^2 \times A$, $\mathcal{X}'[Z_2^p] = \{a, b, c\}^3 \times A$, $\mathcal{X}'[Z_3^p] = \{a, b, c\} \times A$, and $\mathcal{X}'[Z_4^p] = \{a, b, c\}^2 \times A$.

Regret Minimization in FMDPs. We consider a finite FMDP $M = (\{P_i\}_{i \in [m]}, \{R_i\}_{i \in [\ell]}: \mathcal{X})$ and the following RL task. An agent interacts with $M$ for $T$ rounds, starting in an initial state $s_1 \in S$ 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 $i \in [\ell]$, $r_t[i] \sim R_t(x_t[Z_i^p])$; and (ii) Nature decides a next state $s_{t+1} = (s_{t+1}[1], \ldots, s_{t+1}[m])$ where for each $i \in [m]$, $s_{t+1}[i] \sim P_{t+1}(\cdot|x'_t[Z_i^p])$. Let $r^\text{col}_t = \frac{1}{T} \sum_{i \in [\ell]} r_t[i]$ denote the normalized collected reward at time $t$.

The goal of the agent is to maximize the long-term cumulative reward $\sum_{t=1}^T r^\text{col}_t$, which is 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) $\hat{A}$, after $T$ steps and starting from an initial state $s_1 \in S$ as:

$$\mathcal{R}(T, A, s_1) = T g^*(s_1) - \sum_{t=1}^T r^\text{col}_t,$$

where $g^*$ denotes the long-term average-reward (or gain) of $M$, in terms of $r^\text{col}$, 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, $g^*(s_1) = g^*$ for all $s_1 \in S$. We therefore define: $\mathcal{R}(T, A) = T g^* - \sum_{t=1}^T r^\text{col}_t$. 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].

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 $Z \subseteq [m]$ and $x \in \mathcal{X}[Z]$, we let $N(t, x; Z) := \max(\sum_{t'=1}^{t-1} \{x_{t'}[Z] = x\}, 1)$ 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 $N_{t,i}(x) := N(t, x; Z_i^p)$. Likewise, for $i \in [\ell]$, we define $N_{t,i}(x) := N(t, x; Z_i^p)$. 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 $\hat{P}_{t,i}(x|\cdot)$ be the empirical estimate of $P_t(\cdot|x)$ built using $N_{t,i}(x)$ i.i.d. samples from $P_t(\cdot|x)$:

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

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

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

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

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

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

$$\beta_n(\delta) := \eta \log \left( \frac{\log(n) \log(qn)}{\log(\eta \delta)} \right),$$

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

The definition of the confidence set $C_{t,\delta}(x)$ 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). We also note that in the definition of $\beta_n(\delta)$, this leads us to define the following confidence set for mean rewards: For $x \in \mathcal{X}$,

$$C^*_{t,\delta}(x) = \left\{ \mu' \in \mathcal{R}_{\mathcal{X},[0,1]}(G_t) : \forall i \in [t], \mu'_i(x[Z_i^t]) \in C_{t,\delta}(x[Z_i^t]) \right\},$$

where $\delta_i = \delta(t,|\mathcal{X}|^{-1})^{-1}$. As for the transition function, we define for each $i \in [m]$, $x \in \mathcal{X}[Z_i^t]$, and $y \in \mathcal{S}_i$ the following confidence set:

$$C_{t,\delta}(x,y) = \left\{ q \in [0,1] : \tilde{P}_{t,i}(x|y) - q \right\} \leq \left\{ \frac{2q(1-q)}{N^{P'}_{t,i}(x)} \beta_{N^{P'}_{t,i}(x)}(\delta) + \frac{\beta_{N^{P}_{t,i}(x)}(\delta)}{3N^{P}_{t,i}(x)} \right\},$$

$$t_k = \min \left\{ t > t_{k-1} : \max \left\{ \frac{\nu_{t,i,k-1:t-1}(x)}{N_{t,i,k-1}(x)}, \nu_{t,i,k-1:t}(x) \right\} \geq 1 \right\},$$

where $\nu_{t,i,k-1:t-1}(x)$ (resp. $\nu_{t,i,t-1:t}(x)$) denotes the number of observations of $x \in \mathcal{X}[Z_i^{t-1}]$ (resp. of $x \in \mathcal{X}[Z_i^t]$) between time $t$ and $t_{k-1}$. 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 $G$, confidence parameter $\delta$

**Initialize:** For all $i \in [m]$, $x \in \mathcal{X}[Z_i^t]$, set $N_{i,t,0}(x) = 0$. For all $i \in [t], x \in \mathcal{X}[Z_i^t]$, set $N_{i,t,0}(x) = 0$. Set $t_0 = t = 1, k = 1$.

**for** episodes $k = 1, 2, \ldots$ **do**

Set $t_k = t$

Compute empirical estimates $\{\hat{\mu}, t_k(x)\}_{x \in \mathcal{X}[Z_i^t]}$ and $\{\tilde{P}_{t,i}(x|y)\}_{i \in [m], x \in \mathcal{X}[Z_i^t]}$

Compute $\pi_k^* = EVI \left( \mathcal{M}_{t,\delta}, \frac{1}{\sqrt{T}} \right)$ — see Algorithm 2

Set $\nu_{t,i,k-1:t}(x) = 0$ for all $i \in [m]$ and $x \in \mathcal{X}[Z_i^t]$

Set $\nu_{t,i,k-1:t}(x) = 0$ for all $i \in [t]$ and $x \in \mathcal{X}[Z_i^t]$

continue = True

while continue **do**

Observe the current state $s_t$, play action $a_t = \pi_k^*(s_t)$, and observe reward $r_t = (r_{t[1]}, \ldots, r_{t[t]})$.

Set $x_t = (s_t, a_t)$

Set $\nu_{t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) = \nu_{t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) + 1$, $i \in [m]$

Set $\nu_{t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) = \nu_{t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) + 1$, $i \in [t]$

continue = $\bigwedge_{i \in [m]} \left( \nu_{t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) < N_{i,t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) \right)$ and $\bigwedge_{i \in [t]} \left( \nu_{t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) < N_{i,t,i,k-1:t}(x_{t[i]}, Z_{t[i]}^t) \right)$

Set $t = t + 1$

end while

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

Let $u_0 \equiv 0$, $u_{-1} \equiv -\infty$, $n = 0$

while $\max_a(u_n(s) - u_{n-1}(s)) - \min_a(u_n(s) - u_{n-1}(s)) > \varepsilon$

do

Compute

For all $(s, a)$, compute $\tilde{\mu}(\cdot|s, a) = \max[\mu'(|s, a)|: \mu' \in C']$.

For all $(s, a)$, compute $\tilde{P}_n(|s, a)$ using $\text{InnerMax}(u_n, C')$

- See Algorithm 3

Update

\[
\begin{align*}
&\text{argmax}_a \quad u_{n+1}(s) = \max_{a \in A} \left( \frac{\tilde{\mu}(s, a)}{\kappa} + \sum_{s' \in S} \tilde{P}_n(y|s, a)u_n(y) \right) \\
&\pi_{n+1}(s) = \arg\max_{a \in A} \left( \frac{\tilde{\mu}(s, a)}{\kappa} + \sum_{s' \in S} \tilde{P}_n(y|s, a)u_n(y) \right)
\end{align*}
\]

$n = n + 1$

end while

Algorithm 3 $\text{InnerMax}(u, C')$

Enumerate $S = \{s_1, s_2, \ldots, s_n\}$ such that $u(s_1) \geq \cdots \geq u(s_n)$

Compute

\[
\begin{align*}
&P^+_i(y[i]) = \max_{q_i(y[i]) : q_i \in C'}, \quad i \in [m] \\
&P^-_i(y[i]) = \min_{q_i(y[i]) : q_i \in C'}, \quad i \in [m]
\end{align*}
\]

Set \( P_i(y) = \prod_{l=1}^{\infty} P^+_l(y[i]) \) \( P_i(y) = \prod_{l=1}^{\infty} P^-_l(y[i]) \)

For $j \in [S], \text{set } q(s_j) = \prod_{l=1}^{\infty} P^+_l(s_j[i])$

while $\sum_{l=1}^{\infty} q(s_j) > 1$ and $l \leq S$

\[
q' = \prod_{l=1}^{\infty} P^+_l(s_l[i])
\]

$q(s_j) = \max(q(l) + \min\{1 - \sum_{j=1}^{l} q(s_j), q' - q(s_j)\}, 1)
\]

$l = l + 1$

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 FMDPs, which we call the factored diameter. To present its definition, for $i \in [m]$ and $x \in A[Z]^i$, 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 $u \in S[Z]^i$, denoted by $D_{i,u} = D_i(u, M)$, is defined as

\[
D_{i,u} = \max_{s \in S[Z]^i} \max_{x \in A[S]^i} \min_{\pi} \mathbb{E}[T^\pi(s_1, s_2)],
\]

where $\mathcal{L}_s := \{s_{(i-1)\vee 1}, s_{(i+1)\wedge n}\} \times \{s_{(j-1)\vee 1}, s_{(j+1)\wedge n}\}$ and one can verify that for any two states $u, v \in \mathcal{L}_s$, it takes at most $\frac{1}{2}$ steps in expectation to reach $u$ starting from $v$.

In this example, the factored diameter is smaller than $D$ by a factor of $O(n^2)$. Now, if we extend this to the case of $m$ agents, the ratio of $D$ and the factored diameter would be $O(n^m)$. 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.

---

**We use shorthands $a \lor b = \max\{a, b\}$ and $a \land b = \min\{a, b\}$.**
We finally note that Xu and Tewari (2020) presented another \( R \) diameter and factored span are not directly comparable. We upper bound on the factored span. The notions of factored measure called the DBN-UCRL improvement can be substantial in some domains. We also bound of \( \tilde{\Theta} \) in Appendix A: 

\[
\sum_{y \in \mathcal{S}} (P - P') (y) f(y) \leq 3 \max_{y \in \mathcal{S}} f(y) \sum_{i=1}^{m} \xi_i \mathcal{S}_i + \max_{y \in \Phi} f(y) \sum_{i=1}^{m} \sum_{y_i \in \mathcal{S}_i} \sqrt{P_i(y_i) \xi_i}.
\]

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 \( T \geq 3 \), with probability higher than \( 1 - \delta \), it holds that

\[
\mathcal{R}(DBN-UCRL, T) \leq O(c(M) \sqrt{T \log (\log(T)/\delta)}
+ D(S) \sum_{i=1}^{m} |\mathcal{X}[Z_i]| + \sum_{i=1}^{m} D_i A_i \log(T) \log (\log(T)/\delta)),
\]

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

In comparison, the regret of both UCRL-Factored and DORL satisfies \( \mathcal{O}(D \sum_{i=1}^{m} \sqrt{S_i |\mathcal{X}[Z_i]|} T) \). The regret bound of DBN-UCRL improves over these regret bounds as for all \( i \in [m] \) and \( (s,a) \in \mathcal{X}[Z_i] \), we have \( K_{i,s,a} \leq S_i \) and \( D_{i,s} \leq D \). In view of \( D_{i,s} \ll D \) in some FMDPs, this improvement can be substantial in many 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 \( S_i \), action-space \( A_i \), and diameter \( D_i \).

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

\[
\mathcal{R}(DBN-UCRL, T) \leq O\left( \sum_{i=1}^{m} c_i \sqrt{T \log (\log(T)/\delta)} + \sum_{i=1}^{m} D_i A_i \log(T) \log (\log(T)/\delta) \right),
\]

with \( c_i = \sum_{s \in S_i, a \in A_i} D_i^2 K_{i,s,a} + \sum_{i=1}^{m} \sqrt{S_i A_i} + D_i \).

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 \( \sum_{i=1}^{m} D_i \), 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 \( \Omega(\sqrt{bLT}) \) for some \( b \) in FMDPs based on worst-case Cartesian products, where \( L \) is an upper bound on both \( |\mathcal{X}[Z_i]| \) and \( |\mathcal{X}[Z_i]| \), and \( b \) denotes the span of the optimal bias function. Our regret bounds do not contradict this lower bound as \( b \leq \sum_i D_i, s \) 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.

**Lemma 2 (VI for Cartesian products)** Consider VI with $u_0(s) = 0$, and for each $n \geq 0$, $u_{n+1}(s) = \max_{a \in A} \left\{ m^{-1} \mu(s, a) + \sum_{y \in S} P(y|s, a) u_n(y) \right\}$. Then, for all $n$, $u_n(s) = m^{-1} \sum_{i=1}^{m} u^{(i)}(s^{(i)})$, where $(u^{(i)}(s))_{i=0}^{n}$ is a sequence of VI on MDP $i$, that is $u_0^{(i)}(x) = 0$ and $u_{n+1}^{(i)}(x) = \max_{a \in A_i} \left\{ \mu_i(x, a) + \sum_{s \in S_i} P_i(y|x, a) u_i^{(i)}(y) \right\}$ for all $x \in S_i$ 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, $|X| = 144$). 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, $|X| = 512$), which we call Three-Layer RiverSwim. Recall that $X = \bigotimes_{i=1}^{m} S_i \otimes \bigotimes_{i=m+1}^{n} A_i \otimes \cdots \otimes A_{n-m}$, i.e., each factor scope $X|Z_i^n$ (resp. $X|Z_i^m$) 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: 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 Weisstein’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 displays 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 UCRL2-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, though 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...
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 3 and 5 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, \textsc{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, \textsc{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 \textsc{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 \textsc{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 \textsc{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 \textsc{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


Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In Proceedings of the 16th
Mohammad Sadegh Talebi, Anders Jonsson, Odalric-Ambrym Maillard

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


