2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

On the Convergence of TD-Learning on Markov Reward Processes with Hidden States

Mohsen Amiri and Sindri Magnússon

Abstract— **We investigate the convergence properties of Temporal Difference (TD) Learning on Markov Reward Processes (MRPs) with new structures for incorporating hidden state information. In particular, each state is characterized by both observable and hidden components, with the assumption that the observable and hidden parts are statistically independent. This setup differs from Hidden Markov Models and Partially Observable Markov Decision Models, in that here it is not possible to infer the hidden information from the state observations. Nevertheless, the hidden state influences the MRP through the rewards, rendering the reward sequence non-Markovian. We prove that TD learning, when applied only on the observable part of the states, converges to a fixed point under mild assumptions on the step-size. Furthermore, we characterize this fixed point in terms of the statistical properties of both the Markov chains representing the observable and hidden parts of the states. Beyond the theoretical results, we illustrate the novel structure on two application setups in communications. Furthermore, we validate our results through experimental evidence, showcasing the convergence of the algorithm in practice.**

I. Introduction

Reinforcement learning (RL) has emerged as an important paradigm in machine learning, unlocking solutions for a vast array of problems, from optimizing operations in industrial settings to developing strategies for complex games [1]. Its essence lies in the ability to learn optimal actions by interacting with an environment, a capability that draws parallels with how humans and animals learn from their experiences. Within RL, Temporal Difference (TD) learning stands as a central technique. Bridging the divide between dynamic programming methods and Monte Carlo methodologies, TD learning provides a mechanism to learn from direct experiences, which is particularly beneficial in scenarios where crafting an explicit environmental model is either challenging or not feasible.

While TD learning has been successful in many applications, it often operates on the premise of complete state observations [1]–[3]. This assumption may not always hold, especially in more intricate environments characterized by hidden or latent states. These states introduce an added layer of complexity due to inherent partial observability, which challenges the process of value estimation. This has motivated various studies of TD learning with hidden state information.

M. Amiri and S. Magnússon are with the Department of Computer and System Science, Stockholm University, 11419 Stockholm, Sweden (e-mail: mohsen.amiri@dsv.su.se; sindri.magnusson@dsv.su.se).

This work was partially supported by the Swedish Research Council through grant 2020-03607 and in part by Sweden's Innovation Agency (Vinnova).

In previous research on TD learning with hidden states, there is a common assumption that there exists a correlation between the hidden states and the observed data. This assumption is often derived from the framework of the Hidden Markov Model (HMM) [4]–[6]. Bridging the gap between the traditional TD learning framework and the existence of hidden states with correlated observations presents a promising avenue for more robust and accurate value estimation in partially observable systems. Within the framework of realization theory for HMMs, two concomitant yet distinctly delineated quandaries manifest: the formulation of an HMM based on empirical data, and the prospective streamlining of a preexisting model, when feasible, into a more compact equivalent form. Regarding the first one, [7] suggests including non-consecutive correlations to extend HMM parameter estimation without significantly increasing computational cost. The contribution in [8] addresses two core challenges in HMMs: the HMM partial realization problem, which deals with characterizing minimal-order HMMs based on finite sequences of joint densities and learning HMMs from finite output observations of stochastic processes. Regarding the second one, the research work in [9] presents an algebraic approach to model reduction HMMs, ensuring that the reduced model maintains its HMM characteristics. It delves into the algebraic structures and their representations in a more comprehensive manner.

As researchers tackle these challenges in the realm of hidden states, the exploration of hidden states and their relationships with observable data extends beyond TD learning and HMMs, presenting a rich field of research with broad applications. This exploration coincides with the advancements in Partially Observable Markov Decision Processes (POMDPs), derived from HMMs [10], [11], extend the Markov Decision Process (MDP) framework by acknowledging the inherent imperfection in decision makers' ability to fully observe the world state. Since POMDPs often face the "curse of dimensionality" with their large state and action spaces, the contribution in [12] addresses the "curse of ambiguity," stemming from the challenges in precisely quantifying and defining exact transition probabilities. In [13], they introduced a methodology for synthesizing policies that fulfill a linear temporal logic formula within a POMDP. Beyond advancements in enhancing POMDPs, there is a plethora of applications across various domains. The survey by Lauri et al. [14] aims to connect POMDP model development with its utilization in robot decision tasks. This endeavor involves aligning task characteristics with the mathematical and algorithmic aspects to enable effective modeling and

problem-solving.

In all of the work discussed above, there is a key assumption that it is possible to estimate the full state information from observable states. However, in some applications, there is a hidden state that is independent of the observable states while still influencing the reward. For example, in communications, it is common that unknown Markov environmental noise influences communication performance.

The key contribution of this paper is to study the convergence properties of TD learning on Markov Reward Processes (MRPs) with a novel structure for incorporating hidden state information. In particular, each state is characterized by both observable and hidden components, with the assumption that the observable and hidden parts are statistically independent. We prove that when TD learning is applied to the observable part of the states, it converges to a new optimal cost-to-go function that can be expressed analytically based on the statistical properties of the observable and hidden Markov chains. We illustrate our novel MRP structure on two application setups in communications. Furthermore, we validate our results through experimental evidence, showcasing the convergence of the algorithm in practice. In this paper, we focus on examining this new structure for MRP. However, our upcoming research will explore decision-making in MDPs within a similar structure.

The paper is organized as follows. Section II gives an explanation of the basic requirements that are necessary for understanding the main contributions of this study. Section III introduces the Markov Reward Process (MRP) with hidden states, elucidating its conceptual framework and highlighting two exemplifications of its practical utility. In Section IV, we delve into the Temporal Difference (TD) learning algorithm tailored to this model and proffer formal demonstrations pertaining to the estimation of cost-to-go. Section VI is dedicated to the investigation of simulation outcomes and the scrutiny of the convergence behavior of cost-to-go estimations towards theoretical underpinnings. Finally, Section VII summarizes our findings, highlights the main contributions, and suggests potential areas for future investigation.

A. Notation

To represent non-random vectors and matrices, we use lowercase and uppercase bold letters, respectively. Given a vector $\mathbf{v} \in \mathbb{R}^n$ and a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, the notation $\mathbf{v}(i)$ refers to the i-th entry of $\mathbf{v}$, and $\mathbf{A}(i,j)$ refers to the entry in the i-th row and j-th column of $\mathbf{A}$. For a vector $\mathbf{x} \in \mathbb{R}^n$ we denote by $\texttt{Diag}(\mathbf{x})$ the $n \times n$ diganoal matrix with $\mathbf{x}$ on its diagonal. We use calligraphy to represent sets. For a probability distribution $p(\cdot)$ we use the notation $X \sim p(\cdot)$ to indicate that X is a random variable sampled from the $p(\cdot)$. For a random variable $X \sim p(\cdot)$, we denote by $\mathbf{Pr}\left[X \in \mathcal{X}\right]$ denote the probability of X being in the set $\mathcal{X}$. We denote the Total Variation between distributions using the notation, i.e, for distributions $\mu(\cdot)$ and $\kappa(\cdot)$ on $\mathcal{X}$ we have

$$||\mu - \kappa||_{\text{TV}} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \kappa(x)|.$$

II. Preliminaries

A. Markov Chains

We begin by outlining key properties of Markov chains, drawing from Chapter 1 of [15]. A Markov chain is a pair $(\mathcal{S}, p(\cdot))$ where $\mathcal{S}$ is a finite state space and $p(\cdot)$ is a transition function. More formally,

$$p : \mathcal{S} \times \mathcal{S} \to [0,1]$$

is the state transition probability function where $p(s'|s)$ indicates the probability of transitioning from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$. In particular, for a given $s \in \mathcal{S}$ we have that $p(s'|s) \geq 0$ for all $s' \in \mathcal{S}$ and

$$\sum_{s'\in\mathcal{S}} p(s'|s) = 1.$$

It will be useful to introduce the Markov chain transition matrix representation $\mathbf{P} \in \mathbb{R}^{n\times n}$, where

$$\mathbf{P}(s,s') = p(s'|s).$$

A specific sequence or trajectory of the states is represented by the sequence of random variables:

$$S_0, S_1 \ldots, S_k, \ldots, \tag{1}$$

where $\mathbf{Pr}\left[S_{k+1} = s'|S_k = s\right] = p(s'|s)$. For $t \in \mathbb{N}$, the notation

$$p^t(s'|s) := \mathbf{Pr}\left[S_{k+t} = s'|S_k = s\right],$$

denotes the probability of transitioning to state s' after t time steps, given that the initial state was s. It is easily verified that for any k we have $p^k(s'|s) = \mathbf{P}^k(s',s)$ where $\mathbf{P}^k(i,j)$ refers to the entry in the i-th row and the j-th column in the matrix $\mathbf{P}^k$.

A Markov chain $(\mathcal{S}, p(\cdot))$ is said to be *irreducible* if for any two states $s, s' \in \mathcal{S}$ there exists $k \in \mathbb{N}$ such that $p^k(s'|s) > 0$. For each state $s \in \mathcal{S}$ define the set

$$\mathcal{T}(s) = \{t \geq 1 | p^t(s,s) > 0\}.$$

The *period* of a state $s \in \mathcal{S}$ is defined to be the greatest common divisor of $\mathcal{T}(s)$. In an irreducible Markov chain, all states share the same period, which is subsequently referred to as the period of the entire Markov chain. The chain is called *aperiodic* if all states have the period 1. The following results will be useful [15]:

Proposition 1: Suppose that $(\mathcal{S}, p(\cdot))$ is an irreducible and aperiodic Markov chain. Then there is a unique distribution, $\boldsymbol{\pi} \in \mathbb{R}^{|\mathcal{S}|}$ satisfying $\boldsymbol{\pi}(s) > 0$, for all $s \in \mathcal{S}$, and

$$\sum_{s\in\mathcal{S}} \boldsymbol{\pi}(s) = 1, \quad \text{and} \quad \boldsymbol{\pi} = \mathbf{P}^{\mathrm{T}}\boldsymbol{\pi}.$$

Moreover, we have for each $s, s' \in \mathcal{S}$ that

$$\boldsymbol{\pi}(s') = \lim_{k\to\infty} p^k(s'|s).$$

We call $\boldsymbol{\pi}$ the *invariant distribution* of the Markov chain.

Proposition 2: A Markov chain $(\mathcal{S}, p(\cdot))$ is irreducible and aperiodic if and only if there exists a $K \in \mathbb{N}$ such that $p^k(s'|s) > 0$ for all $s, s' \in \mathcal{S}$ and $k \geq K$.

B. Markov Reward Process (MRP)

A Markov Reward Process (MRP) is a tuple $M = (\mathcal{S}, p(\cdot), \mathbf{r}, \gamma)$ where $\mathcal{S}$ is a finite state space; $p(\cdot)$ is the Markov chain transition function; $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ is a reward vector/function where $\mathbf{r}(s)$ signifying the immediate reward for occupying state $s \in \mathcal{S}$; and $\gamma \in [0, 1]$ is a discount factor that determines the weight we assign to immediate rewards relative to those observed later.

A specific sequence or trajectory of the MRP is represented by the sequence of random variables:

$$S_0, R_0, S_1, R_1 \ldots, S_k, R_k, \ldots \tag{2}$$

where $k \in \mathbb{N}$ is a time index and $R_k = \mathbf{r}(S_k)$. A central task in an MRP is value estimation, i.e., computing the cost-to-go of each state $s \in \mathcal{S}$. Formally, we denote the cost-to-go by the vector $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$ where

$$\mathbf{v}(s) = \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k R_k | S_0 = s\right].$$

It is well known that [16]

$$\mathbf{v} = (\mathbf{I} - \gamma \mathbf{P})^{-1}\mathbf{r}.$$

However, in many real-life situations, we do not know the transition probability or the rewards of the MRP. Instead, we might only have access to a sample trajectory in Equation (2).

C. Temporal Difference (TD) Learning

A well-regarded stochastic method to estimate the value $\mathbf{v}$ from a sample trajectory is the Temporal Difference (TD) evaluation algorithm. The algorithm is based on iteratively updating an estimation $\mathbf{v}^k \in \mathbb{R}^n$ of $\mathbf{v}$ based on each sample (S_k, R_k, S_{k+1}) from the MRP trajectory. In particular, we can take any initialization $\mathbf{v}_0 \in \mathbb{R}^{|\mathcal{S}|}$. After that, for each k, we update the $k+1$-th estimate as follows

$$\mathbf{v}_{k+1}(s) = \begin{cases} \mathbf{v}_k(s) + \alpha_k(R_k + \gamma \mathbf{v}_k(S_{k+1}) - \mathbf{v}_k(s)) & \text{if } s = S_k \\ \mathbf{v}_k(s) & \text{if } s \neq S_k \end{cases}$$

where $\alpha_k > 0$ is a step-size. It is well known that this algorithm's iterates converge to the cost-to-go under appropriate step size selection:

Assumption 1: The step sizes α_k are non-negative, deterministic, and satisfy

$$\sum_{k=0}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{3}$$

Under this assumption on the step-size and given that the $(\mathcal{S}, p(\cdot))$ is irreducible and aperiodic, it is well known that the algorithm converges to the following fixed point with probability one [16]:

$$\lim_{k \to \infty} \mathbf{v}_k = \mathbf{v} = (\mathbf{I} - \gamma \mathbf{P})^{-1}\mathbf{r}. \tag{4}$$

III. Markov Reward Process with Hidden States

In this paper, we study MRPs $M = (\mathcal{S}, p(\cdot), \mathbf{R}, \gamma)$ with hidden states. Specifically, the state space $\mathcal{S}$ is of the form $\mathcal{S} = \mathcal{O} \times \mathcal{H}$, where $\mathcal{O}$ represents the observable part of the state and $\mathcal{H}$ represents the hidden part of the state, a part that remains unknown to the agent. Distinct from the conventional partially observable MRPs (or MDPs), in our setup, it is infeasible to deduce any information about the complete state $s = (o, h) \in \mathcal{O} \times \mathcal{H}$ from the observable states o. To be precise, we examine a scenario where the observable states and hidden states are independent of each other. This is encapsulated in the subsequent assumption.

Assumption 2: The transition probability of moving from state $s = (o, h)$ to $s' = (o', h')$ can be decomposed as

$$p(s'|s) = p_{\mathrm{O}}(o'|o) p_{\mathrm{H}}(h'|h). \tag{5}$$

where $p_{\mathrm{O}} : \mathbb{R} \times \mathcal{O} \to [0, 1]$ and $p_{\mathrm{H}} : \mathbb{R} \times \mathcal{H} \to [0, 1]$ are, respectively, the transition distributions for the observable and hidden states.

This assumption implies that the sequence of observable states O_k and hidden states H_k are independent and can be represented with independent Markov chains $(\mathcal{O}, p_{\mathrm{O}}(\cdot))$ and $(\mathcal{H}, p_{\mathrm{H}}(\cdot))$. However, the reward function still depends on both observable and hidden states. This dependency is captured in the reward matrix $\mathbf{R} \in \mathbb{R}^{|\mathcal{O}| \times |\mathcal{H}|}$, where for $s = (o, h) \in \mathcal{O} \times \mathcal{H}$, $\mathbf{R}(o, h)$ signifies the immediate reward for occupying observable state $o \in \mathcal{O}$ provided that we are in the hidden state $h \in \mathcal{H}$.

To our knowledge, MRP (or Markov Decision Processes) with similar hidden states with the structure in Assumption 2 have not been studied before. However, it is easy to envision application examples where there is a hidden state independent of the observations. In this regard, the sets of observable and hidden states can be modeled as follows:

- The set of observable states:

$$\mathcal{O} = \{1, 2, \ldots, O\}$$

- The set of hidden states:

$$\mathcal{H} = \{1, 2, \ldots, H\}$$

where O is the number of observable states, and H is the number of hidden states. In the following, we will provide two examples from communications below.

A. Example I: Communication Over a Noisy Channel

In communication systems and wireless sensor networks, nodes often transmit packages periodically over a noise channel. In such systems, both the package size and the channel noise might change at each time period, and both can be modeled with Markov chains. Usually, the package size is known, and the corresponding Markov chain could be modeled with the observable Markov chain $(\mathcal{O}, p_{\mathrm{O}}(\cdot))$. The channel noise is, on the other hand, often not known and could be modeled with the hidden state Markov chain $(\mathcal{H}_{\mathrm{H}}, p_{\mathrm{h}}(\cdot))$. For illustration, we might consider a simple system where both Markov chains have only two states.

For example, with $\mathcal{O} = \{\texttt{SMALL}, \texttt{LARGE}\}$ and $\mathcal{H} = \{\texttt{LOW}, \texttt{HIGH}\}$ where the package size is either SMALL or LARGE and the channel noise is either LOW or HIGH.

To increase the reliability of the packet communications, error correction codes are commonly used to detect and correct the errors on the received packets. The level of error correction can be tuned, but the size of packets and the level of channel noise influence the code's performance [17]. This is due to the fact that error correction codes can detect and correct a limited number of errors. For example, if a big packet is transmitted over a channel with a high noise level, the error probability could be increased, affecting the error correction procedure. For a given error correction code, the performance can be measured based on observing the success of the communicated packets. This is modeled in the reward $R(o, h)$, a function of both the packet size (the observable state o) and the channel noise (the hidden state h). Ideally, we would like to be able to estimate the accumulated rewards so we know the value of different error-correction strategies.

B. Example II: Resource Allocation in Networks

In intelligent networks, a common task is to optimally allocate hardware resources for client services [18]. For example, in a host center, there might be O servers that can be used to provide services to up to H clients over multiple time periods. It takes time to shut down or start servers, so the number of servers per time step is stochastic but depends on the number of servers available during the previous time step, i.e., the number of servers is a Markov chain $(\mathcal{O}, p_{\mathrm{O}}(\cdot))$. Similarly, it is often natural to model the number of clients per time step as a Markov chain $(\mathcal{H}, p_{\mathrm{H}}(\cdot))$. The number of servers is typically known by the host center, while the number of clients is often not known. Thus, they are naturally modeled, respectively, with observable and hidden Markov chains.

It is costly to keep servers running, so ideally, the number of servers should be kept down while trying to service the demand. The service performance can thus be measured in a reward $R(o, h)$, which depends both on the number of servers and clients. A key goal for the host center is to find the optimal decision-making policy for scheduling the number of servers. In this paper, we look at the first step towards that problem, finding the accumulated reward for a fixed policy.

IV. Algorithm and Main Results

For MRPs with hidden states, it is challenging to estimate the full value function. This complexity arises due to our lack of knowledge about the hidden part of the states, and given the independence between observable and hidden states, their estimation becomes futile. Our sole source of information comes from the trajectory of observed states and rewards, represented as:

$$O_0, R_0, O_1, R_1 \ldots, O_k, R_k, \ldots, \tag{6}$$

where $R_k = \mathbf{R}(O_k, H_k)$. Conversely, the hidden states:

$$H_0, H_1, \ldots, H_k, \ldots \tag{7}$$

remain elusive and are beyond the scope of utilization.

Our main contribution is to show that the TD algorithm discussed above for standard MPR can still be applied even if we only use it on the observable part of the states. In particular, we consider the following TD algorithm:

$$\mathbf{v}_{k+1}(o) = \begin{cases} \mathbf{v}_k(o) + \alpha_k(R_k + \gamma \mathbf{v}_k(O_{k+1}) - \mathbf{v}_k(o)) & \text{if } o = O_k \\ \mathbf{v}_k(o) & \text{if } o \neq O_k \end{cases} \tag{8}$$

where $\mathbf{v}_0 \in \mathbb{R}^{|\mathcal{O}|}$ is some initialization. We will show that even under unknown hidden states, this algorithm still converges to a fixed point. However, this fixed point is different from the fixed point of the traditional TD algorithm in Equation (4), as it will depend on some statistical properties of the hidden states dynamics. To derive our results, we need the following assumption on the observable and hidden Markov chains.

Assumption 3: The Markov chains $M_{\mathrm{O}} = (\mathcal{O}, p_{\mathrm{O}}(\cdot))$ and $M_{\mathrm{H}} = (\mathcal{H}, p_{\mathrm{H}}(\cdot))$ are irreducible and aperiodic.

The assumption ensures that both Markov chains M_{O} and M_{H} have an invariant distribution (see our discussion in Section II-A). We can now establish the following convergence results for the TD algorithm under hidden states.

Theorem 1: Consider a MRP with hidden states and suppose that Assumption 2 and 3 hold true. Then the TD policy evaluation algorithm in Equation (8) under the step-size selection in Assumption 1 converges to the following point with probability one:

$$\lim_{k \to \infty} \mathbf{v}_k = (\mathbf{I} - \gamma \mathbf{P}_{\mathrm{O}})^{-1} \mathbf{R} \pi_{\mathrm{H}} \tag{9}$$

where π_{H} is the invariant distribution of the Markov chain M_{H}.

Proof: The proof is found in Section V. ■

The theorem asserts that the TD algorithm remains valid when applied exclusively to the observable parts of the states. In simpler terms, TD learning can be effective in situations where full access to the entire state information is lacking, as long as the observable and hidden states are independent of each other. Moreover, the theorem characterizes the fixed point with respect to the characteristics of both the observable and hidden Markov chains. In particular, the dependence on the hidden Markov chain appears in the term $\mathbf{R}\boldsymbol{\pi}_{\mathrm{H}}$, which is essentially an average of the columns of the reward matrix $\mathbf{R}$ weighted by the invariant distribution $\boldsymbol{\pi}_{\mathrm{H}}$.

V. Proof of Theorem 1

The proof builds on the following classic results for establishing convergence of stochastic algorithms, see, e.g., Proposition 4.8 in [1].

Proposition 3: Let $(\mathcal{X}, q(\cdot))$ be a finite state Markov chain with state space $\mathcal{X}$ and a state trajectory

$$X_0, X_1, \ldots, X_k, \ldots \ .$$

Consider the functions $\mathbf{A} : \mathcal{X} \to \mathbb{R}^{n \times n}$ and $\mathbf{b} : \mathcal{X} \to \mathbb{R}^n$ and the algorithm with iterates $\mathbf{v}_k \in \mathbb{R}^n$ that progress as

$$\mathbf{v}_{k+1} = \mathbf{v}_k + \gamma_k(\mathbf{A}(X_k)\mathbf{v}_k + \mathbf{b}(X_k)), \tag{10}$$

with step-size $\gamma_k > 0$ and initialization $\mathbf{v}_0 \in \mathbb{R}^n$. Suppose that the following holds:

a) The step-sizes γ_k are deterministic and satisfy Equation (3) in Assumption 1.
b) The Markov chain $(\mathcal{X}, q(\cdot))$ has an invariant (steady-state) distribution denoted by $\boldsymbol{\pi} \in [0,1]^{|\mathcal{X}|}$.
c) The matrix defined by $\mathbf{A} = \mathbf{E}_{X\sim\boldsymbol{\pi}}[\mathbf{A}(X)]$ exists and is negative-definite.
d) There exists a constant K such that $||\mathbf{A}|| \leq K$ and $||\mathbf{b}|| \leq K$ for all $X \in \mathcal{X}$, where $b = \mathbf{E}_{X\sim\boldsymbol{\pi}}[\mathbf{b}(X)]$.
e) There exist scalars $C \in \mathbb{R}_+$ and $\rho \in [0,1)$ such that

$$\|\mathbf{E}[\mathbf{A}(X_k)|X_0 = X] - \mathbf{A}\| \leq C\rho^k$$
$$\|\mathbf{E}[b(X_k)|X_0 = X] - b\| \leq C\rho^k$$

for all $k \in \mathbb{N}$ and $X \in \mathcal{X}$.

Then, the algorithm's iterate converges with probability one to the following point:

$$\lim_{k\to\infty} \mathbf{v}_k = -\mathbf{A}^{-1}\mathbf{b}.$$

The proof of our Theorem 1 is based on establishing that the TD algorithm in Equations (8) can be written in the form of the algorithm in Equation (10), and all the conditions of Proposition 3 hold true. To this end, set

$$\mathcal{X} = \{(o, o', h) \in \mathcal{O} \times \mathcal{O} \times \mathcal{H} |\ p_{\mathrm{O}}(o'|o) > 0\}$$

and define the transition distribution indicating the probability of going from state $x = (o_1, o_2, h) \in \mathcal{X}$ and $x' = (o_1', o_2', h') \in \mathcal{X}$ by

$$q(x'|x) = \begin{cases} p_{\mathrm{O}}(o_2'|o_2)p_{\mathrm{H}}(h'|h) & \text{if } \ o_2 = o_1' \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $(\mathcal{X}, q(\cdot))$ is a Markov chain. Moreover, the state trajectory $X_k = (O_k, O_{k+1}, H_k)$ generates the trajectories (6) and (7) for our MRP with hidden states. With this in mind, we can write the TD algorithm in Equations (8) as follows. For a sample $X_k = (O_k, O_{k+1}, H_k)$ define

$$\mathbf{A}(X_k) = \gamma \mathbf{e}_{\mathrm{O}}(O_k)\mathbf{e}_{\mathrm{O}}(O_{k+1})^{\mathrm{T}} - \mathbf{e}_{\mathrm{O}}(O_k)\mathbf{e}_{\mathrm{O}}(O_k^{\mathrm{T}}) \quad (11)$$
$$\mathbf{b}(X_k) = \mathbf{e}_{\mathrm{O}}(O_k)\mathbf{e}_{\mathrm{O}}(O_k)^{\mathrm{T}}\mathbf{R}\ \mathbf{e}_{\mathrm{H}}(H_k), \quad (12)$$

where $\mathbf{e}_{\mathrm{O}}(j) \in \mathbb{R}^{|\mathcal{O}|}$ and $\mathbf{e}_{\mathrm{H}}(j) \in \mathbb{R}^{|\mathcal{H}|}$ are all zero vectors except they have 1 in the j-th entry. It can be verified that the algorithm in Equation (10) with $\mathbf{A}(\cdot)$ and $\mathbf{b}(\cdot)$ from Equations (11) and (12) is equivalent to the TD algorithm in Equations (8).

In subsections V-A and V-B, we show that $(\mathcal{X}, q(\cdot))$ has an invariant distribution and establish that,

$$\mathbf{A} = \mathbf{E}_{X\sim\boldsymbol{\pi}(\cdot)}\mathbf{A}(X) = \gamma\mathbf{D}_{\mathrm{O}}\mathbf{P}_{\mathrm{O}} - \mathbf{D}_{\mathrm{O}} \quad (13)$$
$$\mathbf{b} = \mathbf{E}_{X\sim\boldsymbol{\pi}(\cdot)}\mathbf{b}(X) = \mathbf{D}_{\mathrm{O}}\mathbf{R}\boldsymbol{\pi}_{\mathrm{H}} \quad (14)$$

where $\mathbf{D}_{\mathrm{O}} = \mathtt{Diag}(\boldsymbol{\pi}_{\mathrm{O}})$, and we have that,

$$\begin{aligned} -\mathbf{A}^{-1}\mathbf{b} &= \left(\mathbf{D}_{\mathrm{O}}\left(\mathbf{I} - \gamma\mathbf{P}_{\mathrm{O}}\right)\right)^{-1}\mathbf{D}_{\mathrm{O}}\mathbf{R}\boldsymbol{\pi}_{\mathrm{H}} \\ &= \left(\mathbf{I} - \gamma\mathbf{P}_{\mathrm{O}}\right)^{-1}\mathbf{R}\boldsymbol{\pi}_{\mathrm{H}}. \end{aligned}$$

Therefore, if we can establish that the conditions a)-e) of Proposition 3 hold true, then we have proved Theorem 1. Note that condition a) trivially holds true; the step-size is chosen according to Assumption 1. In the following subsections, we prove that conditions b)-e) also hold true.

A. Condition b)

We now show that the Markov chain $(\mathcal{X}, q(\cdot))$ has an invariant distribution $\boldsymbol{\pi}$. From Proposition 1 in Section II-A above, it suffices to show that $(\mathcal{X}, q(\cdot))$ is irreducible and aperiodic. We now establish that this is indeed the case due to our Assumption 3.

Lemma 1: If Assumption 3 holds true, then the Markov chain $(\mathcal{X}, q(\cdot))$ is irreducible and aperiodic.

Proof: We follow the notation from Section II-A. By the independence of the hidden and observable state dynamics as described in Assumption 2, for any $x = (o_1, o_2, h)$ and $x' = (o_1', o_2', h')$ in $\mathcal{X}$ we have

$$q^k(x'|x) = p_{\mathrm{O}}^k(o_2'|o_2)p_{\mathrm{H}}^k(h'|h).$$

To prove the above equation, it is sufficient to write and develop the right side of the equation

$$\begin{aligned} &q^k(x'|x) = \mathbf{Pr}\left[X_k{=}x'|X_0{=}x\right] \\ &= \mathbf{Pr}\left[X_k{=}(o_1', o_2', h')|X_0{=}(o_1, o_2, h)\right] \\ &= \mathbf{Pr}\left[O_{k+1}{=}o_2', O_k{=}o_1', H_k = h'|O_1{=}o_2, O_0{=}o_1, H_0 = h\right]. \end{aligned}$$

By using the chain rule property we get that

$$\begin{aligned} &\mathbf{Pr}\left[O_{k+1}{=}o_2', O_k{=}o_1', H_k = h'|O_1{=}o_2, O_0{=}o_1, H_0 = h\right] \\ &= \mathbf{Pr}\left[O_{k+1}{=}o_2'|O_k{=}o_1', H_k = h', O_1{=}o_2, O_0{=}o_1, H_0 = h\right] \\ &\mathbf{Pr}\left[O_k{=}o_1'|H_k = h', O_1{=}o_2, O_0{=}o_1, H_0 = h\right] \\ &\mathbf{Pr}\left[H_k = h'|O_1{=}o_2, O_0{=}o_1, H_0 = h\right]. \end{aligned}$$

Then by Markov chain property and independence of hidden and observable states from Assumption 1 we have

$$\begin{aligned} &q^k(x'|x) = \mathbf{Pr}\left[O_{k+1}{=}o_2'|O_k{=}o_1'\right] \\ &\mathbf{Pr}\left[O_k = o_1'|O_1 = o_2\right]\mathbf{Pr}\left[H_k = h', |H_0 = h\right] \\ &= \mathbf{Pr}\left[O_{k+1} = o_2'|O_1 = o_2\right]\mathbf{Pr}\left[H_k{=}h', |H_0{=}h\right] \\ &= p_{\mathrm{O}}^k(o_2'|o_2)p_{\mathrm{H}}^k(h'|h), \end{aligned}$$

where the second equation is derived from the chain rule property. Now since both Markov chains $(\mathcal{O}, p_{\mathrm{O}}(\cdot))$ and $(\mathcal{H}, p_{\mathrm{H}}(\cdot))$ are irreducible and aperiodic (Assumption 3) we know from Proposition 2 that there exists $K \in \mathbb{N}$ such that $p_{\mathrm{O}}^k(o_2'|o_2) > 0$ and $p_{\mathrm{H}}^k(h'|h) > 0$ for all $k \geq K$ and $o \in \mathcal{O}$ and $h \in \mathcal{H}$. This, in turn, ensures $q^k(x'|x) > 0$ for all $k \geq K$ and $x, x' \in \mathcal{X}$. Therefore, by using the other direction of Proposition 2 we can conclude that $(\mathcal{X}, q(\cdot))$ is irreducible and aperiodic. ■

B. Conditions c) and d)

We start by proving that Equations (13) and (14) hold true. Note that for states $o, o' \in \mathcal{O}$ then $\mathbf{e}_{\mathrm{O}}(o)\mathbf{e}_{\mathrm{O}}(o)^{\mathrm{T}}$ is a $n \times n$ matrix that is everywhere zero except it has 1 on the diagonal element corresponding to state o. Similarly, $\mathbf{e}_{\mathrm{O}}(o)\mathbf{e}_{\mathrm{O}}(o')^{\mathrm{T}}$ is a $n \times n$ matrix that is everywhere zero except it has 1 on the row and column corresponding, respectively, to the states o and o'. Therefore, if we take $X = (O, O', H) \sim \boldsymbol{\pi}(\cdot)$, i.e., $O \sim \pi_{\mathrm{O}}(\cdot)$ and $O' \sim p_{\mathrm{O}}(\cdot|O = o)$, then we get

$$\mathbf{E}_{X\sim\boldsymbol{\pi}(\cdot)}\mathbf{e}_{\mathrm{O}}(O)\mathbf{e}_{\mathrm{O}}(O)^{\mathrm{T}} = \mathtt{Diag}(\boldsymbol{\pi}_{\mathrm{O}}) = \mathbf{D}_{\mathrm{O}} \tag{15}$$

$$\mathbf{E}_{X\sim\boldsymbol{\pi}(\cdot)}\mathbf{e}_{\mathrm{O}}(O)\mathbf{e}_{\mathrm{O}}(O')^{\mathrm{T}} = \mathbf{D}_{\mathrm{O}}\mathbf{P}_{\mathrm{O}}. \tag{16}$$

Equation (13) now follows directly from the definition of $\mathbf{A}(\cdot)$ in Equation (11). In the same way, Equation (14) is obtained by merging Equation (15) and the definition in Equation (12), given the independence between hidden and observable states. Additionally, $\mathbf{E}_{X\sim\boldsymbol{\pi}(\cdot)}\mathbf{e}_{\mathrm{H}}(h) = \boldsymbol{\pi}_{\mathrm{H}}$.

We next prove that $\mathbf{A}$ is negative definite. To that end, we show that $\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} < 0$ for all $\mathbf{w} \in \mathbb{R}^{|\mathcal{O}|} \setminus \{0\}$. In particular, we have that

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} &= \mathbf{w}^{\mathrm{T}}\left(\gamma\mathbf{D}_{\mathrm{O}}\mathbf{P}_{\mathrm{O}} - \mathbf{D}_{\mathrm{O}}\right)\mathbf{w} \\ &= \gamma\mathbf{w}^{\mathrm{T}}\mathbf{D}_{\mathrm{O}}\mathbf{P}_{\mathrm{O}}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{D}_{\mathrm{O}}\mathbf{w}. \end{aligned} \tag{17}$$

Let $\mathbf{D}_{\mathrm{O}}^{1/2} \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{O}|}$ be the diagonal matrix whose entries are the element-wise square roots of the corresponding elements in $\mathbf{D}_{\mathrm{O}}$. Then, by the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\mathbf{w}^{\mathrm{T}}\mathbf{D}_{\mathrm{O}}\mathbf{P}_{\mathrm{O}}\mathbf{w} &= \left(\mathbf{D}_{\mathrm{O}}^{1/2}\mathbf{w}\right)^{\mathrm{T}}\mathbf{D}_{\mathrm{O}}^{1/2}\mathbf{P}_{\mathrm{O}}\mathbf{w} \\ &\leq ||\mathbf{D}_{\mathrm{O}}^{1/2}\mathbf{w}||_2||\mathbf{D}_{\mathrm{O}}^{1/2}\mathbf{P}_{\mathrm{O}}\mathbf{w}||_2. \end{aligned} \tag{18}$$

By considering the norm

$$||\mathbf{w}||_{\mathbf{D}_{\mathrm{O}}} = \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{D}_{\mathrm{O}}\mathbf{w}}$$

and using that $||\mathbf{D}_{\mathrm{O}}^{1/2}\mathbf{w}||_2 = ||\mathbf{w}||_{\mathbf{D}_{\mathrm{O}}}$ for all $\mathbf{w}$ we have that

$$\mathbf{w}^{\mathrm{T}}\mathbf{D}_{\mathrm{O}}\mathbf{P}_{\mathrm{O}}\mathbf{w} \leq ||\mathbf{w}||_{\mathbf{D}_{\mathrm{O}}}||\mathbf{P}_{\mathrm{O}}\mathbf{w}||_{\mathbf{D}_{\mathrm{O}}}.$$

It is easily verified that $||\mathbf{P}_{\mathrm{O}}\mathbf{w}||_{\mathbf{D}_{\mathrm{O}}} \leq ||\mathbf{w}||_{\mathbf{D}_{\mathrm{O}}}$ for all $\mathbf{w} \in \mathbb{R}^{|\mathcal{O}|}$, see, e.g., Lemma 1 in [16]. This, together with Equations (17) and (18) ensures that

$$\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} \leq \gamma||\mathbf{w}||^2_{\mathbf{D}_{\mathrm{O}}} - ||\mathbf{w}||^2_{\mathbf{D}_{\mathrm{O}}} = (\gamma - 1)||\mathbf{w}||^2_{\mathbf{D}_{\mathrm{O}}}.$$

Since $\gamma < 1$, it follows that $\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} < 0$ for all $\mathbf{w} \in \mathbb{R}^{|\mathcal{O}|}$.

Finally, we establish that there exists $K \in \mathbb{R}$ such that $||\mathbf{A}|| \leq K$ and $||\mathbf{b}|| \leq K$. To that end, note that the state space $\mathcal{X}$ is finite, thus $\mathbf{A}(x)$ and $\mathbf{b}(x)$ can only take finite values, and must thus be bounded for all $x \in \mathcal{X}$, i.e., there exists $K \in \mathbb{R}$ such that $\mathbf{A}(x) \leq K$ for all $x \in \mathcal{X}$. This means that $||\mathbf{A}|| = ||\mathbf{E}_{X\sim\pi}[\mathbf{A}(X)]|| \leq K$ and $||\mathbf{b}|| = ||\mathbf{E}_{X\sim\pi}[\mathbf{b}(X)]|| \leq K$, so $\mathbf{A}$ and $\mathbf{b}$ are bounded.

C. Condition e)

From Lemma 1 proved above, the Markov chain $(\mathcal{X}, q(\cdot))$ is both irreducible and aperiodic. Therefore, by the Convergence Theorem for Markov chains, see, e.g., Theorem 4.9 in Chapter 4 in [15], there exist $\alpha \in (0, 1)$ and $C > 0$ such that for all $x \in \mathcal{X}$ we have

$$\max_{x\in\mathcal{X}} ||q^k(\cdot|x) - \boldsymbol{\pi}||_{\mathrm{TV}} \leq C\alpha^k \quad \text{for all} \quad n \in \mathbb{N}.$$

Therefore, recalling from above that there exists $K \in \mathbb{R}$ such that $||A(x)|| \leq K$ for all $x \in \mathcal{X}$, we have

$$\begin{aligned}&\|E[\mathbf{A}(X_k)|X_0{=}x_0]{-}\mathbf{A}\| = \left\|\sum_{x\in\mathcal{X}} \mathbf{A}(x)(q^k(x|x_0){-}\boldsymbol{\pi}(x))\right\| \\ &\leq \sum_{x\in\mathcal{X}} \|\mathbf{A}(x)\||q^k(x|x_0) - \boldsymbol{\pi}(x)| \\ &\leq K\sum_{x\in\mathcal{X}} |q^k(x|x_0) - \boldsymbol{\pi}(x)| = 2K||q^k(\cdot|x_0) - \boldsymbol{\pi}||_{\mathrm{TV}} \\ &\leq 2KC\alpha^k.\end{aligned}$$

Therefore, the first inequality in part e) of Proposition 3 is established. The second inequality follows similarly

$$\begin{aligned}\|E[\mathbf{b}(X_k)|X_0 = x_0] - \mathbf{b}\| &\leq K\sum_{x\in\mathcal{X}} |q^k(x|x_0) - \boldsymbol{\pi}(x)| \\ &\leq 2KC\alpha^k.\end{aligned}$$

As a result, both inequalities of part e) are established, which concludes the proof.

VI. Numerical Experiments

We now evaluate the algorithm in simulation by considering problems similar to the two application examples in sections III-A and III-B. We demonstrate the algorithm's convergence towards its theoretically defined limit as specified in Equation (9). In both instances, the algorithm is executed 100 times, allowing for a comprehensive evaluation of its average performance. You can find the code for implementing these two examples by following the link provided below[1].

For investigating the employment of TD learning for the example in Section III-A, we consider the Markov chains with the following observable and hidden transition and reward matrices:

$$P_{\mathrm{O}} = \begin{pmatrix}0.3 & 0.7\\ 0.2 & 0.8\end{pmatrix} \; P_{\mathrm{H}} = \begin{pmatrix}0.9 & 0.1\\ 0.8 & 0.2\end{pmatrix} \; R = \begin{pmatrix}0 & -1\\ -1 & 0\end{pmatrix}$$

To compute the true fixed point, we note that the invariant distribution of the hidden Markov chain can be computed from P_{H}, it is $\pi_h = [8/9, 1/9]^{\mathrm{T}}$. Note that none of the matrices above is known by the algorithm. We experimentally tuned the step-size as $\alpha_k = 10/k$, $\gamma = 0.5$, and run the algorithm for 5000 iterations.

In Figure 1, we visualize the algorithm's convergence for each individual state. The figures clearly indicate that the cost-to-go for each observable state steadily approaches

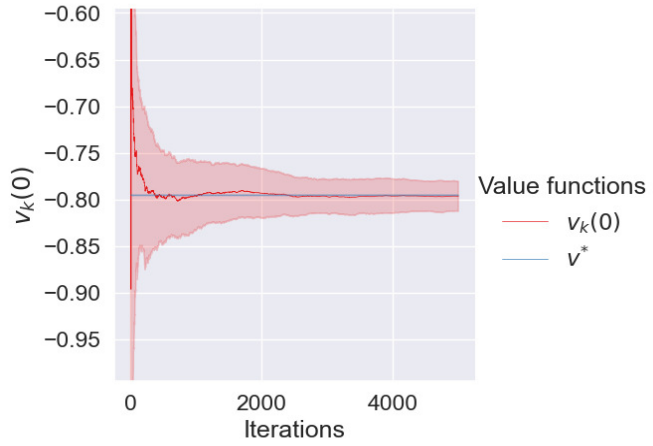
[1] https://github.com/mohsen1amiri/MRP-with-Hidden-States/tree/main.

the theoretical limit value as postulated in Theorem 1. This convergence trend becomes evident after approximately 5000 iterations of the algorithm delineated in Equation (8).

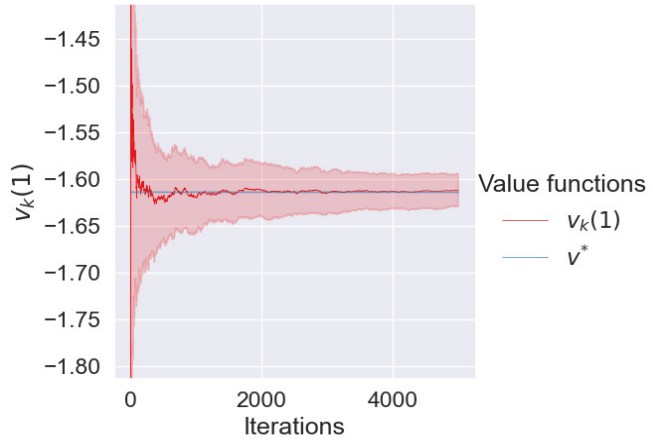
Figure 2 illustrates the convergence of the algorithm to the limit $\mathbf{v}^* = (\mathbf{I} - \gamma \mathbf{P}_{\mathrm{O}})^{-1} \mathbf{R} \pi_{\mathrm{H}}$. In particular, we illustrate the convergence to the proportional error

$$\frac{\|\mathbf{v}_k - \mathbf{v}^*\|_2}{\|\mathbf{v}^*\|_2}. \tag{19}$$

The figure shows that the algorithm converges, and the error approaches zero as the number of iterations increases. After roughly 4000 iterations, the proportional estimation error has almost reached 1%.



(a) The estimation of the cost-to-go function for observable state 0 in each algorithm iteration.



(b) The estimation of the cost-to-go function for observable state 1 in each algorithm iteration.

Fig. 1: The estimation of the cost-to-go function of the observable states, for example III-A using the algorithm in Equation (8). The solid red line shows the mean value across all iterations and the shaded red area shows the standard deviation. The blue line represents the optimal cost-to-go function.

We next investigate the application of TD learning for the example in Section III-B, with $O = 10$ and $H = 11$. The transition matrices for the observable and hidden Markov

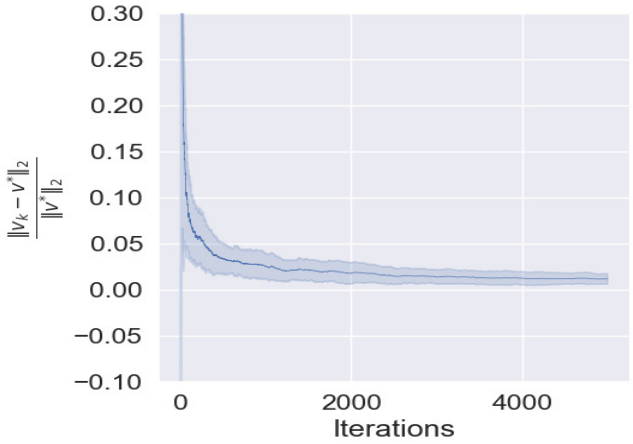


Fig. 2: The average behavior of convergence of the difference (Equation (19)) in the cost-to-go vector estimation in each iteration with the theoretical limit for the observable states in example III-A's setup. Here, the M is 100 and shows the number of times that the algorithm is repeated. The solid blue line shows the mean value across all iterations and the shaded blue area shows the standard deviation.

chains are randomly generated.The reward matrix is defined as

$$\mathbf{R}(o,h) = \begin{cases} 1 & \text{if } o = h \text{ or } (o = 10 \text{ and } h = 11) \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

The step size is $\alpha_k = 10/k$, $\gamma = 0.5$, and we run the algorithm for 50000 iterations.

In the context of Example III-B, we apply the established experimental setup to examine the convergence behavior of the algorithm for an individual observable state, as depicted in Figure 3. Evidently, the cost-to-go function consistently approaches its theoretical limit through the iterative execution of the algorithm. On a broader scale, Figure 4 illustrates the average behavior of the criterion expressed in Equation (19), which converges to zero. This signifies that the cost-to-go vector, encompassing the cost-to-go values for all observable states, systematically converges to the specified theoretical limit. Notably, the limited standard deviation observed in these simulations implies the absence of any divergence during the execution. It's worth noting that in this particular example, the convergence process proceeds at a relatively slower pace, owing to the presence of numerous observable states.

These simulation results collectively establish that running the algorithm delineated in Equation (8) for MRPs with hidden states, without leveraging information from these hidden states, culminates in the convergence of the theoretical limit. This theoretical limit is calculated based on the invariant distribution of the hidden states' Markov chain, and the results offer compelling evidence of this convergence phenomenon.

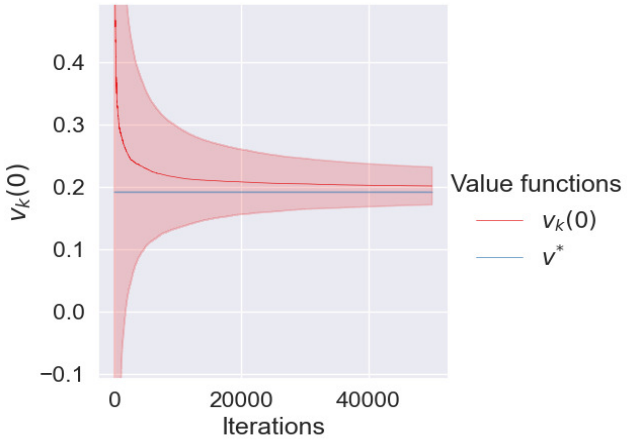


Fig. 3: The algorithm convergence for an instance observable state of the example III-B. The solid red line shows the mean value across all iterations and the shaded red area shows the standard deviation. The blue line represents the optimal cost-to-go function.

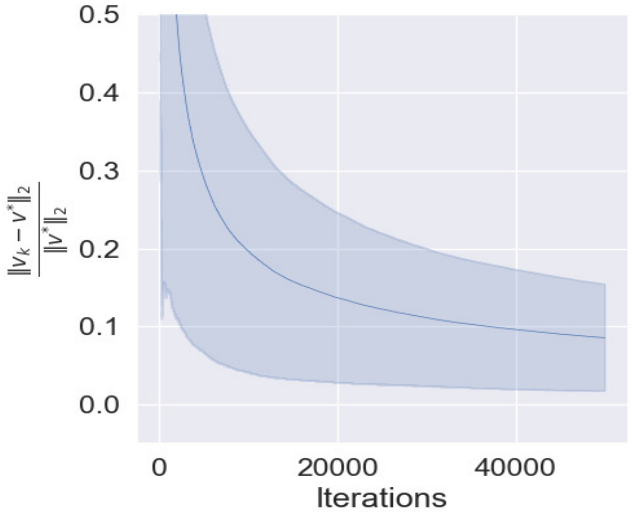


Fig. 4: The average behavior of convergence of the difference (Equation (19)) in the cost-to-go vector estimation in each iteration with the theoretical limit for the observable states in example III-B's setup. Here, the M is 100 and shows the number of times that the algorithm is repeated. The solid blue line shows the mean value across all iterations and the shaded blue area shows the standard deviation.

VII. Conclusion

Reinforcement learning, with TD learning as its prominent algorithm, has emerged as a powerful paradigm in machine learning, bridging dynamic programming and Monte Carlo methods for knowledge acquisition in challenging environments. However, the assumption of complete state observations is limiting, particularly in situations involving hidden states. This contribution introduces a novel model for MRPs, segregating states into observable and hidden components, with their independence assumed. Despite a lack of information about hidden states, the reward signal intricately depends on both observable and hidden components. The study achieves an analytical representation of the cost-to-go function, revealing the surprising accuracy of the TD learning algorithm in estimating it. Beyond theory, we validate the model in practical scenarios and highlight its potential for addressing complex problems. Future work will explore the application of similar structures as we considered here MRP to more general Markov Decision Processes.

References

[1] Dimitri Bertsekas and John N Tsitsiklis, *Neuro-dynamic programming*, Athena Scientific, 1996.

[2] Rayadurgam Srikant and Lei Ying, "Finite-time error bounds for linear stochastic approximation andtd learning," in *Conference on Learning Theory*. PMLR, 2019, pp. 2803–2830.

[3] Jalaj Bhandari, Daniel Russo, and Raghav Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Conference on learning theory*. PMLR, 2018, pp. 1691–1692.

[4] Olivier Cappé, Eric Moulines, and Tobias Rydén, "Inference in hidden markov models," in *Proceedings of EUSFLAT conference*, 2009, pp. 14–16.

[5] Sean R Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[6] Bhavya Mor, Sunita Garhwal, and Ajay Kumar, "A systematic review of hidden markov models and their applications," *Archives of computational methods in engineering*, vol. 28, pp. 1429–1448, 2021.

[7] Robert Mattila, Cristian Rojas, Eric Moulines, Vikram Krishnamurthy, and Bo Wahlberg, "Fast and consistent learning of hidden markov models by incorporating non-consecutive correlations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6785–6796.

[8] Qingqing Huang, Rong Ge, Sham Kakade, and Munther Dahleh, "Minimal realization problems for hidden markov models," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1896–1904, 2015.

[9] Tommaso Grigoletto and Francesco Ticozzi, "Algebraic reduction of hidden markov models," *IEEE Transactions on Automatic Control*, 2023.

[10] Karl J Astrom et al., "Optimal control of markov processes with incomplete state information," *Journal of mathematical analysis and applications*, vol. 10, no. 1, pp. 174–205, 1965.

[11] Michael L Littman, "A tutorial on partially observable markov decision processes," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 119–125, 2009.

[12] Soroush Saghafian, "Ambiguous partially observable markov decision processes: Structural results and applications," *Journal of Economic Theory*, vol. 178, pp. 1–35, 2018.

[13] Maxime Bouton, Jana Tumova, and Mykel J Kochenderfer, "Point-based methods for model checking in partially observable markov decision processes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 10061–10068.

[14] Mikko Lauri, David Hsu, and Joni Pajarinen, "Partially observable markov decision processes in robotics: A survey," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 21–40, 2022.

[15] David A Levin and Yuval Peres, *Markov chains and mixing times*, vol. 107, American Mathematical Soc., 2017.

[16] John N Tsitsiklis and Benjamin Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, 1997.

[17] John G Proakis, *Digital communications*, McGraw-Hill, Higher Education, 2008.

[18] V V Vinothina, R Sridaran, and Padmavathi Ganapathi, "A survey on resource allocation strategies in cloud computing," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 6, 2012.