Convergence of regularized agent-state-based Q-learning in POMDPs

Amit Sinha, Matthieu Geist, Aditya Mahajan

Abstract— **In this paper, we present a framework to understand the convergence of commonly used Q-learning reinforcement learning algorithms in practice. Two salient features of such algorithms are: (i) the Q-table is recursively updated using an agent state (such as the state of a recurrent neural network) which is not a belief state or an information state and (ii) policy regularization is often used to encourage exploration and stabilize the learning algorithm. We investigate the simplest form of such Q-learning algorithms which we call regularized agent-state-based Q-learning (RASQL) and show that it converges under mild technical conditions to the fixed point of an appropriately defined regularized MDP, which depends on the stationary distribution induced by the behavioral policy. We also show that a similar analysis continues to work for a variant of RASQL that learns periodic policies. We present numerical examples to illustrate that the empirical convergence behavior matches with the proposed theoretical limit.**

I. Introduction

Reinforcement learning (RL) is a useful paradigm in learning optimal control policies via simulation when the system model is not available or when the system is too large to explicitly solve the dynamic program. The simplest setting is the fully-observed setting of Markov decision processes (MDP), where the controller has access to the environment state. Most existing theoretical RL results on convergence of learning algorithms and their rates of convergence and regret bounds, etc. are established for the MDP setting.

However, in many real-world applications, such as autonomous driving, robotics, healthcare, finance, and others, the controller does not have access to the environment state; rather, it has a partial observation of the environment state. So these applications need to be modeled as a partially observable Markov decision process (POMDP) rather than a MDP.

When the system model is known, the POMDP model can be converted into an MDP by considering the controller's belief on the state of the environment (also called the belief state) as an information state [1]–[3]. However, such a reduction does not work in the RL setting because the belief state depends on the system model, which is unknown. Nonetheless, there have been several empirical works which show that standard RL algorithms for MDPs continue to work for POMDPs if one uses "frame stacking" (i.e., use the last few observations as a state) or recurrent neural networks [4]–[7]. In recent years, considerable progress has been made in understanding the properties of such algorithms but a complete theoretical understanding is still lacking.

A common way to model such RL algorithms for POMDPs is to consider the state of the controller as an agent state [8]. Such agent-state-based-controllers have also been considered in the planning setting as they can be simpler to implement than belief-state based controllers. See [9] for an overview.

A challenge in understanding the convergence of agent-state-based RL algorithms for POMDPs is that an agent state is not an information state. So, it is not possible to write a dynamic programming decomposition based on the agent state. So, one cannot follow the typical proof techniques used to evaluate the convergence of RL algorithms for MDPs (where RL algorithms can be viewed as stochastic approximation variant of MDP algorithms such as value iteration and policy iteration to compute the optimal policy).

There is a good understanding of the convergence of agent-state-based Q-learning (ASQL) for POMDPs [10]–[12] (which is related to Q-learning for non-Markovian environments [13], [14]). There is also some work on understanding the convergence of actor-critic algorithms for POMDPs [3], [15]. However, most practical RL algorithms for POMDPs use some form of policy regularization, while most theoretical analysis is restricted to the unregularized setting.

Regularization adds an auxiliary loss to the per-step rewards. This loss typically depends on the policy but may also depend on the value function. Regularization is commonly used in RL algorithms for various reasons, such as entropy regularization to encourage exploration [16]–[18] and improve generalization [19], KL-regularization to constrain the policy updates to be similar to a prior policy [20], [21], and others. Unified theory for different facets of regularization in MDPs is provided in [22], [23].

Based on the various benefits of regularization in RL for MDPs, it is also commonly used in RL for POMDPs [3], [5], [16], [21], [24]–[27]. However, the recent theoretical analysis of RL for POMDPs discussed above do not consider regularization. The objective of this paper to to present initial results on understanding regularization in RL for POMDPs.

There is some recent work on understanding regularization in POMDPs but they either consider the role of entropy regularization in POMDP solvers (when the model information is known) [28], [29], or consider regularization of the belief distribution [30] or observation distribution [31]. These results do not directly provide an understanding of the role of regularization in RL for POMDPs.

In this paper, we revisit Q-learning for POMDPs when the learning agent is using an agent state and using policy regularization. Our main contribution is to show that in

A. Sinha and A. Mahajan are with the Department of Electrical and Computer Engineering, McGill University, Canada. Emails: amit.sinha@mail.mcgill.ca, aditya.mahajan@mcgill.ca. Their work was supported in part by in part by a grant from Google's Institutional Research Program in collaboration with Mila. M. Geist is with the Earth Species Project. Email: matthieu@earthspecies.org

this setting, Q-learning converges under mild technical conditions. We characterize the converged limit in terms of the model parameters and choice of behavioral policy used in Q-learning. Recently, it has been argued that periodic policies may perform better when considering agent-state-based POMDPs [12]. We show that our analysis extends to a periodic version of regularized Q-learning as well.

Notation: We use uppercase letters to denote random variables (e.g. S, A, etc.), lowercase letters to denote their realizations (e.g. s, a, etc.) and calligraphic letters to denote sets (e.g. S, A; etc.). Subscripts (e.g. S_t, A_t, etc.) denote variables at time t. Similarly, $S_{1:t}$ denotes the collection of random variables from time 1 to t. $\Delta(\mathsf{S})$ denotes the space of probability measures on a set S; $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ denote the probability of an event and the expectation of a random variable, respectively; and $\mathbb{1}(\cdot)$ denotes the indicator function. $|\mathsf{S}|$ denotes the number of elements in S (when it is a finite set). $\mathbb{R}$ denotes real numbers. $[L]$ denotes the set of integers from 0 to $L-1$, where $L \in \mathbb{Z}^+$. $[\![\ell]\!]$ denotes (ℓ mod L).

II. Background

A. Legendre-Fenchel transform (convex conjugate)

We start with a short review of convex conjugates and Legendre-Fenchel transforms [32], which are an important tool to understand regularization in MDPs [23].

Definition 1 For a strongly convex function $\Omega\colon \mathbb{R}^n \to \mathbb{R}$, its convex conjugate $\Omega^\star\colon \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\Omega^\star(q) = \max_{p \in \mathbb{R}^n} \big\{ \langle p, q \rangle - \Omega(p) \big\}.$$

The mapping $\Omega \mapsto \Omega^*$ is the Legendre-Fenchel transform.

The following is a useful property of the Legendre-Fenchel transform for regularized MDPs:

Lemma 1 (Based on [33], [34]) *Let Δ be a simplex in $\mathbb{R}^n$ and $\Omega\colon \Delta \to \mathbb{R}$ be twice differentiable and a strongly convex function. Let $\Omega^\star\colon \mathbb{R}^n \to \mathbb{R}$ be the Legendre-Fenchel transform of Ω. Then, $\nabla\Omega^\star$ is Lipschitz and satisfies*

$$\nabla\Omega^\star(q) = \arg\max_{p \in \Delta} \big\{ \langle p, q \rangle - \Omega(p) \big\}.$$

In Markov decision processes, one often regularizes the policy. Below we describe some of the commonly used policy regularizers. For the purpose of the discussion below, let A be a finite set (later we will take A to be the set of actions of an MDP, but for now we can consider it as a generic set).

1) **Entropy regularization** uses the regularizer $\Omega\colon \Delta(\mathsf{A}) \to \mathbb{R}$ given by $\Omega(p) = \frac{1}{\beta} \sum_{a \in \mathsf{A}} p(a) \ln p(a)$ where $\beta \in \mathbb{R}_{>0}$ is a parameter. Its convex conjugate $\Omega^\star\colon \mathbb{R}^{|\mathsf{A}|} \to \mathbb{R}$ is given by $\Omega^\star(q) = \frac{1}{\beta} \ln\big(\sum_{a \in \mathsf{A}} \exp(\beta q(a))\big)$. Furthermore, from Lemma 1, we get that the argmax in the definition of convex conjugate is achieved by

$$p^\star(a) = \frac{\exp(\beta q(a))}{\sum_{a' \in \mathsf{A}} \exp(\beta q(a'))}.$$

2) **KL regularization** uses the regularizer $\Omega\colon \Delta(\mathsf{A}) \to \mathbb{R}$ given by $\Omega(p) = \frac{1}{\beta} \sum_{a \in \mathsf{A}} p(a) \ln(p(a)/p_{\text{REF}}(a))$, where $\beta \in \mathbb{R}_{>0}$ is a parameter and $p_{\text{REF}} \in \Delta(\mathsf{A})$ is a reference distribution. Its convex conjugate $\Omega^\star\colon \mathbb{R}^{|\mathsf{A}|} \to \mathbb{R}$ is given by $\Omega^\star(q) = \frac{1}{\beta} \ln\big(\sum_{a \in \mathsf{A}} p_{\text{REF}}(a) \exp(\beta q(a))\big)$. Furthermore, from Lemma 1, we get that the argmax in the definition of convex conjugate is achieved by

$$p^\star(a) = \frac{p_{\text{REF}}(a) \exp(\beta q(a))}{\sum_{a' \in \mathsf{A}} p_{\text{REF}}(a') \exp(\beta q(a'))}.$$

B. Regularized MDPs

In this section, we provide a brief review of regularized Markov decision processes (MDPs), which are a generalization of standard MDPs with an additional "regularization cost" at each stage.

Consider a Markov decision process (MDP) with state $s_t \in \mathsf{S}$, control action $a_t \in \mathsf{A}$, where all sets are finite. The system operates in discrete time. The initial state $s_1 \sim \rho$ and for any time $t \in \mathbb{N}$, we have

$$\mathbb{P}(s_{t+1} \mid s_{1:t}, a_{1:t}) = \mathbb{P}(s_{t+1} \mid s_t, a_t) =: P(s_{t+1} \mid s_t, a_t),$$

where P is a probability transition matrix. The system yields a reward $R_t = r(s_t, a_t) \in [0, R_{\max}]$. The rewards are discounted by a factor $\gamma \in [0, 1)$.

Consider a policy $\pi : \mathsf{S} \to \Delta(\mathsf{A})$. Let $\Omega\colon \Delta(\mathsf{A}) \to \mathbb{R}$ be a strongly convex function that is used as a policy regularizer. Then, the *regularized performance* of policy π is given by

$$J_\pi^\Omega := \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} \big[r(s_t, a_t) - \Omega(\pi(\cdot \mid s_t)) \big] \;\middle|\; s_1 \sim \rho \right],$$

where the notation $\mathbb{E}^\pi$ means that the expectation is taken with the joint measure on the system variables induced by the policy π.

The objective in a regularized MDP is to find a policy π that maximizes the regularized performance J_π^Ω defined above. A key step in understanding the optimal solution of the regularized MDP is to define the regularized Bellman operator $\mathcal{B}^\Omega$ on the space of real-valued functions on $\mathsf{S} \times \mathsf{A}$ as follows. For any $Q : \mathsf{S} \times \mathsf{A} \to \mathbb{R}$,

$$\mathcal{B}^\Omega Q(s, a) = r(s, a) + \gamma \sum_{s' \in \mathsf{S}} P(s' \mid s, a) \Omega^\star(Q(s', \cdot)),$$

where $\Omega^\star$ is the Legendre-Fenchel transform of Ω.

Proposition 1 (Based on [23]) *The following results hold:*

1) *The operator $\mathcal{B}^\Omega$ is a contraction and therefore has a unique fixed point, which we denote by Q^Ω. By definition,*

$$Q^\Omega(s, a) = r(s, a) + \gamma \sum_{s' \in \mathsf{S}} P(s' \mid s, a) \Omega^\star(Q^\Omega(s', \cdot)).$$

2) *Define the policy $\pi^{\Omega,*}\colon \mathsf{S} \to \Delta(\mathsf{A})$ as follows: for any $s \in \mathsf{S}$,*

$$\begin{aligned} \pi^{\Omega,\star}(\cdot \mid s) &= \nabla\Omega^\star(Q^\Omega(s, \cdot)) \\ &= \arg\max_{\xi \in \Delta(\mathsf{A})} \left\{ \sum_{a \in \mathsf{A}} \xi(a) Q^\Omega(s, a) - \Omega(\xi) \right\} \end{aligned}$$

where the last equality follows from Lemma 1. Then, the policy $\pi^{\Omega,\star}$ is optimal for maximizing the regularized performance J_π^Ω over the set of all policies.

III. System model and regularized Q-learning for POMDPs

A. Model for POMDPs

Consider a partially observable Markov decision process (POMDP) with state $s_t \in \mathsf{S}$, control action $a_t \in \mathsf{A}$, and output $y_t \in \mathsf{Y}$, where all sets are finite. The system operates in discrete time with the dynamics given as follows. The initial state $s_1 \sim \rho$ and for any time $t \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{P}(s_{t+1}, y_{t+1} \mid s_{1:t}, y_{1:t}, a_{1:t}) &= \mathbb{P}(s_{t+1}, y_{t+1} \mid s_t, a_t) \\ &=: P(s_{t+1}, y_{t+1} \mid s_t, a_t),\end{aligned}$$

where P is a probability transition matrix. In addition, at each time the system yields a reward $r_t = r(s_t, a_t) \in [0, R_{\max}]$. The rewards are discounted by a factor $\gamma \in [0, 1)$.

Let $\vec{\boldsymbol{\pi}} = (\vec{\pi}_1, \vec{\pi}_2, \dots)$ denote any (history dependent and possibly randomized) policy, i.e., under policy $\vec{\boldsymbol{\pi}}$ the action at time t is chosen as $a_t \sim \vec{\pi}_t(y_{1:t}, a_{1:t-1})$. The performance of policy $\vec{\boldsymbol{\pi}}$ is given by

$$J_{\vec{\boldsymbol{\pi}}} \coloneqq \mathbb{E}^{\vec{\boldsymbol{\pi}}}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \;\middle|\; s_1 \sim \rho\right],$$

where the notation $\mathbb{E}^{\vec{\boldsymbol{\pi}}}$ means that the expectation is taken with the joint measure on the system variables induced by the policy $\vec{\boldsymbol{\pi}}$.

The objective is to find a (history dependent and possibly randomized) policy $\vec{\boldsymbol{\pi}}$ to maximize $J_{\vec{\boldsymbol{\pi}}}$. When the system model is known, the above POMDP model can be converted to a fully observed Markov decision process (MDP) by considering the controller's posterior belief on the system state as an information state [1], [2]. However, when the system model is not known, it is not possible to run reinforcement learning (RL) algorithms on the belief-state MDP because the belief depends on the system model. For that reason, in RL for POMDPs it is often assumed that the controller is an agent-state-based controller.

Definition 2 (Agent state) An agent state is a model-free recursively updateable function of the history of observations and actions. In particular, let Z denote the agent state space. Then, the agent state is a process $\{z_t\}_{t\geq 0}$, $z_t \in \mathsf{Z}$, which starts with some initial value z_0, and is then recursively computed as

$$z_{t+1} = \phi(z_t, y_{t+1}, a_t), \quad t \geq 0 \tag{1}$$

where ϕ is a pre-specified agent-state update function.

Some examples of agent-state-based controllers are: (i) a finite memory controller, which chooses the actions based on the previous k observations; (ii) a finite state controller, which effectively filters the possible histories to values from a finite set Z. We refer the reader to [9] for a detailed review of agent-state-based policies in POMDPs.

We use $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ to denote an agent-state-based policy,[1] i.e., a policy where the action at time t is given by $a_t \sim \pi_t(z_t)$. An agent-state-based policy is said to be **stationary** if for all t and t', we have $\pi_t(a \mid z) = \pi_{t'}(a \mid z)$ for all $(z, a) \in \mathsf{Z} \times \mathsf{A}$.

If the agent state is an information state, then MDP-based RL algorithms can directly be applied to find optimal stationary solutions [3]. However, in general, an agent state is not an information state, as is the case in frame-stacking or when using recurrent neural networks. In such settings, the dynamics of the agent state process is non Markovian and the standard dynamic programming based argument does not work. It is possible to find the optimal policy by viewing the POMDP with an agent-state-based controller as a decentralized control problem and using the designer's approach [35] to compute an optimal agent-state-based policy, as is done in [9], but such an approach is intractable for all but small toy problems.

The Q-learning algorithms for POMDPs maintain a Q-table based on the agent states and actions and update the Q-values based on the samples generated by the environment. Since the agent state is non Markovian, it is not clear if such an iterative scheme converges, and if so, to what value. In the next section, we present a formal model for agent state based Q-learning when the agent also uses policy regularization.

B. Regularized agent-state-based Q-learning for POMDPs

In this section we describe regularized agent-state-based Q-learning (RASQL), which is an online off-policy learning approach in which the agent acts according to a fixed behavioral policy to generate a sample path $(z_1, a_1, r_1, z_2, \dots)$ of agent states, actions, and rewards observed by a learning agent. We assume that the sampled rewards $r_t = r(s_t, a_t)$ are available to the the agent during the learning process.

The learning agent uses a policy regularizer $\Omega\colon \Delta(\mathsf{A}) \to \mathbb{R}$ and maintains a regularized Q-table, which is arbitrarily initialized and then recursively updated as follows:

$$\begin{aligned}Q_{t+1}(z, a) &= Q_t(z, a) \\ &\quad + \alpha(z, a)\left[r_t + \gamma \Omega^\star(Q_t(z_{t+1}, \cdot)) - Q_t(z, a)\right], \end{aligned} \tag{2}$$

where the learning rate sequence $\{\alpha_t(z, a)\}_{t\geq 1}$ is chosen such that $\alpha_t(z, a) = 0$ whenever $(z, a) \neq (z_t, a_t)$. For instance, if the policy regularizer is the entropy regularizer, then the above iteration corresponds to an agent-state-based version of soft-Q-learning [36]. The "greedy" policy at each time is given by $\pi_t(\cdot \mid z) = \nabla\Omega^\star(Q_t(z, \cdot))$. Thus, for entropy regularization, it would correspond to soft-max based on Q_t.

If the $\Omega^\star(Q_t(z_{t+1}, \cdot))$ term in (2) is replaced by $\max_{a' \in \mathsf{A}} Q_t(z_{t+1}, a')$, the iteration in RASQL corresponds to agent-state-based Q-learning (ASQL):

$$\begin{aligned}Q_{t+1}(z_t, a_t) &= Q_t(z_t, a_t) \quad + \\ &\alpha_t(z_t, a_t)\left[r_t + \gamma \max_{a' \in \mathsf{A}} Q_t(z_{t+1}, a') - Q_t(z_t, a_t)\right].\end{aligned}$$

[1] We use $\vec{\boldsymbol{\pi}}$ to denote history dependent policies and $\boldsymbol{\pi}$ to denote agent-state-based policies.

The convergence of ASQL and its variations have been recently studied in [11], [12], [14]. However, the analysis of ASQL does not include regularization. The main result of this paper is to characterize the convergence of RASQL.

IV. Main result

We impose the following standard assumptions on the model.

Assumption 1 For all (z, a), the learning rates $\{\alpha_t(z,a)\}_{t\geq 1}$ are measurable with respect to the sigma-algebra generated by $(z_{1:t}, a_{1:t})$ and satisfy $\alpha_t(z,a) = 0$ if $(z,a) \neq (z_t, a_t)$. Moreover, $\sum_{t\geq 1} \alpha_t(z,a) = \infty$ and $\sum_{t\geq 1}(\alpha_t(z,a))^2 < \infty$, almost surely.

Assumption 2 The behavior policy μ is such that the Markov chain $\{(S_t, Y_t, Z_t, A_t)\}_{t\geq 1}$ converges to a limiting distribution ζ_μ, where $\sum_{(s,y)} \zeta_\mu(s,y,z,a) > 0$ for all (z,a) (i.e., all (z,a) are visited infinitely often).

Assumption 1 is the standard assumption for convergence of stochastic approximation algorithms [37]. Assumption 2 ensures persistence of excitation and is a standard assumption in convergence analysis of Q-learning [10]–[12], [38], [39].

For ease of notation, we will continue to use ζ_μ to denote the marginal and conditional distributions w.r.t. ζ_μ. In particular, for marginals we use $\zeta_\mu(y,z,a)$ to denote $\sum_{s\in\mathsf{S}} \zeta_\mu(s,y,z,a)$ and so on; for conditionals, we use $\zeta_\mu(s|z,a)$ to denote $\zeta_\mu(s,z,a)/\zeta_\mu(z,a)$ and so on. Note that $\zeta_\mu(s,z,y,a) = \zeta_\mu(s,z)\mu(a|z)P(y|s,a)$. Thus, we have that $\zeta_\mu(s|z,a) = \zeta_\mu(s|z)$.

The key idea to characterize the convergence behavior is the following. Given the limiting distribution ζ_μ, we can define an MDP with state space Z, action space A, and per-step reward $r_\mu \colon \mathsf{Z}\times\mathsf{A} \to \mathbb{R}$ and dynamics $P_\mu \colon \mathsf{Z}\times\mathsf{A} \to \Delta(\mathsf{Z})$ given as follows:

$$\begin{aligned} r_\mu(z,a) &\coloneqq \sum_{s\in\mathsf{S}} r(s,a)\zeta_\mu(s \mid z), \\ P_\mu(z'|z,a) &\coloneqq \sum_{(s,y')\in\mathsf{S}\times\mathsf{Y}} \mathbb{1}_{\{z'=\phi(z,y',a)\}} P(y'|s,a)\zeta_\mu(s|z). \end{aligned} \tag{3}$$

Now consider a regularized version of this MDP, where we regularize the policy using Ω. Let Q_μ denote the fixed point of the regularized Bellman operator corresponding to this regularized MDP, i.e., Q_μ is the unique fixed point of the following (see the discussion in Sec. II-B):

$$Q_\mu(z,a) = r_\mu(z,a) + \gamma \sum_{z'\in\mathsf{Z}} P_\mu(z' \mid z,a)\Omega^\star(Q_\mu(z',\cdot)). \tag{4}$$

Then, our main result is the following:

Theorem 1 *Under Assumptions 1 and 2, the RASQL iteration (2) converges to Q_μ almost surely.*

Proof The proof is given in appendix A.

Remark 1 Note that Proposition 1 implies that the "greedy" regularized policy with respect to the limit point of $\{Q_t\}_{t\geq 1}$ is given by $\pi^*(\cdot \mid z) = \nabla\Omega^\star(Q_\mu(z,\cdot))$, which typically lies in the interior of $\Delta(\mathsf{A})$ for each z. Thus, the greedy policy is stochastic. This is a big advantage of RASQL compared to ASQL because in ASQL, the greedy policy corresponding to the limit point of the Q-learning iteration is deterministic. As shown in [40] (also see [9], [12]), in general for POMDPs with agent-state-based controllers, stochastic stationary policies can outperform deterministic stationary policies.

V. Regularized periodic Q-learning

The idea of periodic Q-learning has been explored in [12]. They show that periodic policies can perform better than stationary policies when the agent state is not an information state. Regularized Q-learning can be generalized by regularized periodic Q-learning, since taking the period $L = 1$ reproduces the stationary setting.

Consider the convergence properties when we consider the following regularized periodic agent-state-based Q-learning (RePASQL) update for $\ell \in [L]$.

$$\begin{aligned} Q^\ell_{t+1}(z,a) &= Q^\ell_t(z,a) \\ &\quad + \alpha^\ell_t(z,a)\left[r_t + \gamma\Omega^\star(Q^{[\![\ell+1]\!]}_t(z',\cdot)) - Q^\ell_t(z,a)\right]. \end{aligned} \tag{5}$$

Assumption 3 For all (ℓ, z, a), the learning rates $\{\alpha^\ell_t(z,a)\}_{t\geq 1}$ are measurable with respect to the sigma-algebra generated by $(z_{1:t}, a_{1:t})$ and satisfy $\alpha^\ell_t(z,a) = 0$ if $(\ell,z,a) \neq ([\![t]\!], z_t, a_t)$. Moreover, $\sum_{t\geq 1} \alpha^\ell_t(z,a) = \infty$ and $\sum_{t\geq 1}(\alpha^\ell_t(z,a))^2 < \infty$, almost surely.

Assumption 4 The behavior/exploration policy $\mu = \{\mu^\ell\}_{\ell\in[L]}$ is such that the Markov chain $\{(S_t, Y_t, Z_t, A_t)\}_{t\geq 1}$ converges to a limiting periodic distribution ζ^ℓ_μ, where $\sum_{(s,y)} \zeta^\ell_\mu(s,y,z,a) > 0$ for all (ℓ,z,a) (i.e., all (ℓ,z,a) are visited infinitely often).

By considering this limiting distribution w.r.t. the original model's rewards and dynamics, we can construct an artificial MDP on the agent state for each $\ell \in [L]$, which has the following rewards and dynamics:

$$\begin{aligned} r^\ell_\mu(z,a) &\coloneqq \sum_{s\in\mathsf{S}} r(s,a)\zeta^\ell_\mu(s \mid z), \\ P^\ell_\mu(z'|z,a) &\coloneqq \sum_{(s,y')\in\mathsf{S}\times\mathsf{Y}} \mathbb{1}_{\{z'=\phi(z,y',a)\}} P(y'|s,a)\zeta^\ell_\mu(s|z). \end{aligned} \tag{6}$$

Now we can extend the same techniques used in regularized MDPs II-B to this by defining a regularized Bellman operator $\mathcal{B}^\ell_\mu$ on an arbitrary Q-function $Q \in \mathbb{R}^{|\mathsf{Z}|\times|\mathsf{A}|}$ as follows:

$$\mathcal{B}^\ell_\mu Q(z,a) = r^\ell_\mu(z,a) + \gamma \sum_{z'\in\mathsf{Z}} P^\ell_\mu(z' \mid z,a)\Omega^\star(Q(z',\cdot)).$$

Next define the composition of the sequence of L Bellman operators corresponding to cycle ℓ as is done in [12].

$$\mathbb{B}^\ell_\mu = \mathcal{B}^\ell_\mu \mathcal{B}^{[\![\ell+1]\!]}_\mu \cdots \mathcal{B}^{[\![\ell+L-1]\!]}_\mu.$$

Then we can apply Prop. 1 to $\mathbb{B}^\ell_\mu$. In addition, considering the periodicity of the operators, the same approach followed in [12] can be used to show that $\mathbb{B}^\ell_\mu$ is a contraction and therefore has a unique fixed point denoted by Q^ℓ_μ which is given by

$$Q^\ell_\mu(z,a) = r^\ell_\mu(z,a) + \gamma \sum_{z'\in\mathsf{Z}} P^\ell_\mu(z' \mid z,a) V^{[\![\ell+1]\!]}_\mu(z'). \tag{7}$$

Theorem 2 *Under Assumptions 3 and 4, the RePASQL iteration (5) converges to* $\{Q_\mu^\ell\}_{\ell\in[L]}$ *almost surely.*

PROOF The proof is given in appendix B.

VI. NUMERICAL EXAMPLE

In this section, we present an example to highlight the salient features of our results. First, we describe the POMDP model.

A. POMDP model

Consider a POMDP with $\mathsf{S} = \{0, 1, 2, 3\}, \mathsf{A} = \{0, 1\}, \mathsf{Y} = \{0, 1\}$ and $\gamma = 0.9$. The start state distribution is given by

$$\rho(s) = \begin{bmatrix} 0.3, 0.0, 0.2, 0.5 \end{bmatrix}$$

Now consider the reward and transitions when $a = 0$:

$$r(s, 0) = (1 - \gamma) \times \begin{bmatrix} 0.6, 0.0, 0.5, -0.3 \end{bmatrix}$$

$$P(s' \mid s, 0) = \begin{bmatrix} 0.0 & 0.6 & 0.4 & 0.0 \\ 0.8 & 0.0 & 0.2 & 0.0 \\ 0.7 & 0.3 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.8 \end{bmatrix}.$$

Note that s, s' (state, next state) corresponds to the rows, columns of P, respectively. Next, when $a = 1$

$$r(s, 1) = (1 - \gamma) \times \begin{bmatrix} 0.1, -0.3, -0.2, 0.5 \end{bmatrix}$$

$$P(s' \mid s, 1) = \begin{bmatrix} 0.8 & 0.2 & 0.0 & 0.0 \\ 0.4 & 0.0 & 0.6 & 0.0 \\ 0.0 & 0.8 & 0.2 & 0.0 \\ 0.1 & 0.7 & 0.2 & 0.0 \end{bmatrix}.$$

Finally, we have the observations function which maps $s = \{0, 3\}$ to $y = 0$ and $s = \{1, 2\}$ to $y = 1$.

B. Regularized agent-state-based Q-learning (RASQL) experiment

For the purpose of providing a simple illustration in this example, we fix the agent state to be the observation of the agent, i.e., $z_t = y_t$. However, in general the theoretical results hold for the general agent-state update rule given in (1). Consider the following fixed exploration policy:

$$\mu(a \mid z) = \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix}.$$

Note that z, a (observation, action) corresponds to the rows, columns of μ, respectively.

Using μ, we run 25 random seeds on the given POMDP and we perform the RASQL update (2) with a regularization coefficient $(\beta) = 1.0$ for 10^5 timesteps/iterations. We plot the median and quartiles from 25 seeds of the iterates $\{Q_t(z, a)\}_{t\geq 1}$ for each (z, a) pair as well as their corresponding theoretical limits $Q_\mu(z, a)$ (computed using Theorem 1) are shown in Fig. 1. The salient features of these results are as follows:

- RASQL converges to the theoretical limit predicted by Theorem 1.
- The limit Q_μ depends on the exploration policy μ.

Thus, it can be seen from this example that we can precisely characterize the limits of convergence when using regularized Q-learning with an agent-state-based representation.

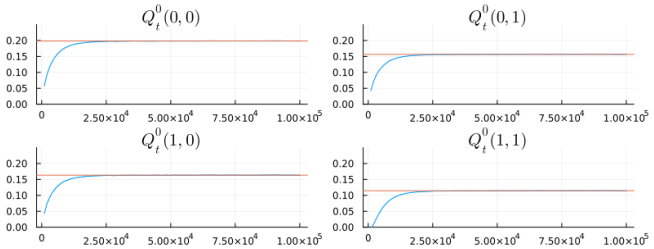

Fig. 1: RASQL convergence: Q-values vs. number of iterations. Blue: RASQL iterates, Red: Theoretical limit from Theorem 1.

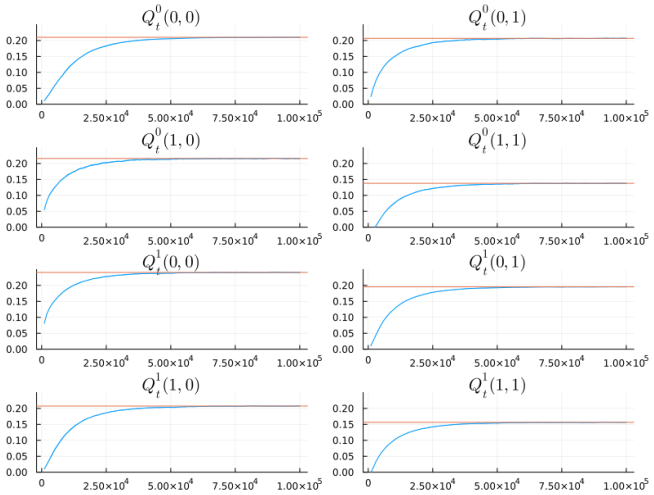

Fig. 2: RePASQL convergence: Q-values vs. number of iterations. Blue: RePASQL iterates, Red: Theoretical limit from Theorem 2

C. Regularized periodic agent-state-based Q-learning (RePASQL) experiment

Similar to the RASQL experiment, we fix the agent state to be the observation of the agent, i.e., $z_t = y_t$. Consider the following fixed periodic exploration policy for period $L = 2$:

$$\mu^0(a \mid z) = \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix}, \quad \mu^1(a \mid z) = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}.$$

Using μ^ℓ, we run 25 random seeds on the given POMDP and we perform the RePASQL update (5) with a regularization coefficient $(\beta) = 1.0$ for 10^5 timesteps/iterations. We plot the median and quartiles from 25 seeds of the iterates $\{Q_t^\ell(z, a)\}_{t\geq 1}$ for each (ℓ, z, a) pair as well as their corresponding theoretical limits $Q_\mu^\ell(z, a)$ (computed using Theorem 2) are shown in Fig. 2. The salient features of these results are as follows:

- RePASQL converges to the theoretical limit predicted by Theorem 2.
- The limits $\{Q_\mu^\ell\}_{\ell\in[L]}$ depend on the periodic exploration policy $\{\mu^\ell\}_{\ell\in[L]}$.

Thus, it can be seen from this example that we can precisely characterize the limits of convergence.

VII. Conclusions

In this work, we present theoretical results on the convergence of regularized agent-state-based Q-learning (RASQL) under some standard assumptions from the literature. In particular, we show that: 1) RASQL converges and 2) we characterize the solution that RASQL converges to as a function of the model parameters and the choice of exploration policy. We illustrate these ideas on a small POMDP example and show that the Q-learning iterates of RASQL matches with the calculated theoretical limit. We also generalize these ideas to the periodic setting and demonstrate the theoretical and empirical convergence of RePASQL. Thus, in doing so we are able to understand how regularization works when combined with Q-learning for POMDPs that have an agent state that is not an information state.

A noteworthy issue with RASQL/RePASQL is that it inherits the limitations of its predecessor approaches of ASQL and PASQL. In particular, while we are able to prove convergence and characterize the converged solution in RASQL/RePASQL, we cannot guarantee the convergence to the optimal agent-state-based solution and this largely depends on the choice of exploration policy and the POMDP dynamics. Even so, seeing how regularization is an important component in several empirical works concerning POMDPs with agent states that are not an information state, we find it useful to establish some useful theoretical properties on the convergence of such algorithms.

References

[1] K. J. Åström, "Optimal control of Markov processes with incomplete state information I," *Journal of Mathematical Analysis and Applications*, vol. 10, pp. 174–205, 1965.

[2] R. D. Smallwood and E. J. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Operations Research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[3] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems," *J. Mach. Learn. Res.*, vol. 23, no. 12, pp. 1–83, 2022.

[4] M. J. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs." in *AAAI Fall Symposia*, vol. 45, 2015, p. 141.

[5] M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson, "Deep variational reinforcement learning for POMDPs," in *Int. Conf. Mach. Learn.* PMLR, 2018, pp. 2117–2126.

[6] P. Zhu, X. Li, P. Poupart, and G. Miao, "On improving deep reinforcement learning for POMDPs," *arXiv:1704.07978*, 2017.

[7] L. Meng, R. Gorbet, and D. Kulić, "Memory-based deep reinforcement learning for POMDPs," in *Int. Conf. Intell. Robots Syst.* IEEE, 2021, pp. 5619–5626.

[8] S. Dong, B. Van Roy, and Z. Zhou, "Simple agent, complex environment: Efficient reinforcement learning with agent states," *J. Mach. Learn. Res.*, vol. 23, no. 255, pp. 1–54, 2022.

[9] A. Sinha and A. Mahajan, "Agent-state based policies in POMDPs: Beyond belief-state MDPs," *Conference on Decision and Control*, 2024.

[10] T. Jaakkola, S. Singh, and M. Jordan, "Reinforcement learning algorithm for partially observable Markov decision problems," in *Adv. Neural Inf. Process. Syst.*, vol. 7. MIT Press, 1994, pp. 345–352.

[11] A. D. Kara and S. Yüksel, "Convergence of finite memory Q learning for POMDPs and near optimality of learned policies under filter stability," *Mathematics of Operations Research*, Nov. 2022.

[12] A. Sinha, M. Geist, and A. Mahajan, "Periodic agent-state based Q-learning for POMDPs," *Adv. Neural Inf. Process. Syst.*, 2024.

[13] A. Devran Kera and S. Yüksel, "Q-learning for stochastic control under general information structures and non-markovian environments," *Transactions on Machine Learning Research*, 2024.

[14] S. Chandak, P. Shah, V. S. Borkar, and P. Dodhia, "Reinforcement learning in non-Markovian environments," *Systems & Control Letters*, vol. 185, p. 105751, 2024.

[15] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1143–1166, 2003.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.

[17] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[18] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans, "Understanding the impact of entropy on policy optimization," in *Int. Conf. Mach. Learn.* PMLR, 2019, pp. 151–160.

[19] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *AAAI Conference on Artificial Intelligence*, 2018.

[20] J. Peters, K. Mulling, and Y. Altun, "Relative entropy policy search," in *AAAI Conference on Artificial Intelligence*, 2010, pp. 1607–1612.

[21] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Int. Conf. Mach. Learn.* PMLR, 2015, pp. 1889–1897.

[22] G. Neu, A. Jonsson, and V. Gómez, "A unified view of entropy-regularized Markov decision processes," *Adv. Neural Inf. Process. Syst.*, 2017.

[23] M. Geist, B. Scherrer, and O. Pietquin, "A theory of regularized Markov decision processes," in *Int. Conf. Mach. Learn.* PMLR, 2019.

[24] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *Int. Conf. Mach. Learn.* PMLR, 2019, pp. 2555–2565.

[25] M. Zhang, S. Vikram, L. Smith, P. Abbeel, M. Johnson, and S. Levine, "Solar: Deep structured representations for model-based reinforcement learning," in *Int. Conf. Mach. Learn.* PMLR, 2019, pp. 7444–7453.

[26] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *Int. Conf. Learn. Represent.*, 2020.

[27] T. Ni, B. Eysenbach, E. Seyedsalehi, M. Ma, C. Gehring, A. Mahajan, and P.-L. Bacon, "Bridging state and history representations: Understanding self-predictive RL," *Int. Conf. Learn. Represent.*, 2024.

[28] H. Delecki, M. Vazquez-Chanlatte, E. Yel, K. Wray, T. Arnon, S. Witwicki, and M. J. Kochenderfer, "Entropy-regularized point-based value iteration," *arXiv:2402.09388*, 2024.

[29] A. Somani, N. Ye, D. Hsu, and W. S. Lee, "DESPOT: Online POMDP planning with regularization," *Adv. Neural Inf. Process. Syst.*, 2013.

[30] T. L. Molloy and G. N. Nair, "Smoother entropy for active state trajectory estimation and obfuscation in POMDPs," *IEEE Transactions on Automatic Control*, vol. 68, no. 6, pp. 3557–3572, 2023.

[31] R. Zamboni, D. Cirino, M. Restelli, and M. Mutti, "The limits of pure exploration in POMDPs: When the observation entropy is enough," *Reinforcement Learning Journal*, 2024.

[32] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[33] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

[34] A. Mensch and M. Blondel, "Differentiable dynamic programming for structured prediction and attention," in *Int. Conf. Mach. Learn.* PMLR, 2018, pp. 3462–3471.

[35] A. Mahajan, "Sequential decomposition of sequential dynamic teams: applications to real-time communication and networked control systems," Ph.D. dissertation, U. Michigan, Ann Arbor, MI, 2008.

[36] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Int. Conf. Mach. Learn.* PMLR, 2017, pp. 1352–1361.

[37] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[38] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[39] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, pp. 185–202, 1994.

[40] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Learning without state-estimation in partially observable Markovian decision processes," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 284–292.

[41] L. Prashanth, S. Bhatnagar *et al.*, "Gradient-based algorithms for zeroth-order optimization," *Foundations and Trends® in Optimization*, vol. 8, no. 1–3, pp. 1–332, 2025.

APPENDIX

A. Proof of Theorem 1

The proof argument for Theorem 1 is similar to the proof argument given in [10]–[13].

Define an error function between the converged value and the Q-learning iteration $\Delta_{t+1} \coloneqq Q_{t+1} - Q_\mu$. Then, combine (2), (4) and (3) as follows for all (z, a).

$$\begin{aligned}\Delta_{t+1}(z,a) &= Q_{t+1}(z,a) - Q_\mu(z,a)\\ &= (1-\alpha_t(z,a))\Delta_t(z,a)\\ &\quad + \alpha_t(z,a)\left[U_t^0(z,a) + U_t^1(z,a) + U_t^2(z,a)\right], \quad (8)\end{aligned}$$

where

$$\begin{aligned}U_t^0(z,a) &\coloneqq [r(S_t,A_t) - r_\mu(z,a)]\,\mathbb{1}_{\{Z_t=z,A_t=a\}},\\ U_t^1(z,a) &\coloneqq \Big[\gamma\Omega^\star(Q_\mu(Z_{t+1},\cdot))\\ &\quad - \gamma\sum_{z'\in\mathsf{Z}} P_\mu(z' \mid z,a)\Omega^\star(Q_\mu(z',\cdot))\Big]\mathbb{1}_{\{Z_t=z,A_t=a\}},\\ U_t^2(z,a) &\coloneqq\\ &\quad [\gamma\Omega^\star(Q_t(Z_{t+1},\cdot)) - \gamma\Omega^\star(Q_\mu(Z_{t+1},\cdot))]\,\mathbb{1}_{\{Z_t=z,A_t=a\}}.\end{aligned}$$

Note that we are adding the term $\gamma\Omega^\star(Q_\mu(Z_{t+1},\cdot))\mathbb{1}_{\{Z_t=z,A_t=a\}}$ in $U_t^1(z,a)$ and subtracting it in $U_t^2(z,a)$. We can now view (8) as a linear system with state Δ_t and three inputs $U_t^0(z,a), U_t^1(z,a)$ and $U_t^2(z,a)$. Using the linearity, we can now split the state into three components $\Delta_{t+1} = X_{t+1}^0 + X_{t+1}^1 + X_{t+1}^2$, where the components evolve as follows for $i \in \{0,1,2\}$:

$$X_{t+1}^i(z,a) = (1-\alpha_t(z,a))X_t^i(z,a) + \alpha_t(z,a)U_t^i(z,a).$$

We will now separately show each $\|X_t^i\| \to 0$.

a) Convergence of component X_t^0

The proof for the convergence of component X_t^0 is similar to that given in [12].

b) Convergence of component X_t^1

The proof for the convergence of component X_t^1 is based on the argument given in [12]. Let W_t denote the tuple $(S_t, Z_t, A_t, S_{t+1}, Z_{t+1}, A_{t+1})$. Note that $\{W_t\}_{t\geq 1}$ is a Markov chain and converges to a limiting distribution $\bar{\zeta}_\mu$, where

$$\begin{aligned}&\bar{\zeta}_\mu(s,z,a,s',z',a')\\ &\quad = \zeta_\mu(s,z,a)\sum_{y'\in\mathsf{Y}} P(s',y'|s,a)\mathbb{1}_{\{z'=\phi(z,y',a)\}}\mu(a'|z').\end{aligned}$$

We use $\bar{\zeta}_\mu(s,z,a,\mathcal{S},\mathcal{Z},\mathcal{A})$ to denote the marginalization over the "future states" and a similar notation for other marginalizations. Note that $\bar{\zeta}_\mu(s,z,a,\mathcal{S},\mathcal{Z},\mathcal{A}) = \zeta_\mu(s,z,a)$.

Define V_t as the value function associated with Q_t, i.e., $V_t(z) \coloneqq \Omega^\star(Q_t(z,\cdot))$. Fix $(z_\circ, a_\circ) \in \times\mathsf{Z}\times\mathsf{A}$ and define

$$\begin{aligned}h_P(W_t; z_\circ, a_\circ) = \Big[&\gamma V_\mu(Z_{t+1}) -\\ &\gamma\sum_{\bar{z}\in\mathsf{Z}} P_\mu(\bar{z}|z_\circ,a_\circ)V_\mu(\bar{z})\Big]\mathbb{1}_{\{Z_t=z_\circ,A_t=a_\circ\}}.\end{aligned}$$

Then the process $\{X_t^1(z,a)\}_{t\geq 1}$ is given by the stochastic iteration

$$\begin{aligned}X_{t+1}^1(z_\circ,a_\circ) &= (1-\alpha_t(z_\circ,a_\circ))X_t^1(z_\circ,a_\circ)\\ &\quad + \alpha_t(z_\circ,a_\circ)h_P(W_t; z_\circ, a_\circ).\end{aligned}$$

As argued earlier, the process $\{W_t\}_{t\geq 1}$ is a Markov chain. Due to Assm. 1, the learning rate $\alpha_t(z_\circ,a_\circ)$ is measurable with respect to the sigma-algebra generated by $(Z_{1:t}, A_{1:t})$ and is therefore also measurable with respect to the sigma-algebra generated by $W_{1:t}$. Thus, the learning rates $\{\alpha_t(z_\circ,a_\circ)\}_{t\geq 1}$ satisfy the conditions of Theorem 2.7 from [41]. Therefore, the theorem implies that $\{X_t^1(z_\circ,a_\circ)\}_{t\geq 1}$ converges a.s. to the following limit

$$\begin{aligned}\lim_{t\to\infty} X_t^1(z_\circ,a_\circ) &= \sum_{\substack{s,z,a\in\mathsf{S}\times\mathsf{Z}\times\mathsf{A}\\ s',z',a'\in\mathsf{S}\times\mathsf{Z}\times\mathsf{A}}} \bar{\zeta}_\mu(s,z,a,s',z',a')\\ &\quad h_P(s,z,a,s',z',a';z_\circ,a_\circ)\\ &= \gamma\left[\sum_{z'\in\mathsf{Z}}\bar{\zeta}_\mu(\mathcal{S},z_\circ,a_\circ,\mathcal{S},z',\mathcal{A})V_\mu(z')\right]\\ &\quad - \left[\gamma\bar{\zeta}_\mu(\mathcal{S},z_\circ,a_\circ,\mathcal{S},\mathcal{Z},\mathcal{A})\sum_{\bar{z}\in\mathsf{Z}}P_\mu(\bar{z}|z_\circ,a_\circ)V_\mu(\bar{z})\right]\\ &= 0\end{aligned}$$

where the last step follows from the fact that $\bar{\zeta}_\mu(\mathsf{S},z_\circ,a_\circ,\mathsf{S},\mathcal{Z},\mathcal{A}) = \zeta_\mu(z_\circ,a_\circ)$ and $\bar{\zeta}_\mu(\mathsf{S},z_\circ,a_\circ,\mathsf{S},z',\mathsf{A}) = \zeta_\mu(z_\circ,a_\circ)P_\mu(z'|z_\circ,a_\circ)$.

c) Convergence of component X_t^2

The convergence of the X_t^2 component is based on [11], [12] but requires some additional considerations due to the regularization term. We start by defining:

$$\begin{aligned}\pi_t(\cdot \mid z) &= \arg\max_{\xi\in\Delta(\mathsf{A})}\sum_{a\in\mathsf{A}}\xi(a)Q_t(z,a) - \Omega(\xi)\\ \pi^\star(\cdot \mid z) &= \arg\max_{\xi\in\Delta(\mathsf{A})}\sum_{a\in\mathsf{A}}\xi(a)Q_\mu(z,a) - \Omega(\xi).\end{aligned}$$

In the previous steps, we have shown that $\|X_t^i\| \to 0$ a.s., for $i \in \{0,1\}$. Thus, we have that $\|X_t^0 + X_t^1\| \to 0$ a.s. Arbitrarily fix an $\epsilon > 0$. Therefore, there exists a set Ω^1 of measure one and a constant $T(\omega,\epsilon)$ such that for $\omega \in \Omega^1$, all $t > T(\omega,\epsilon)$, and $(z,a) \in \times\mathsf{Z}\times\mathsf{A}$, we have

$$X_t^0(z,a) + X_t^1(z,a) < \epsilon. \quad (9)$$

Now pick a constant C such that

$$\kappa \coloneqq \gamma\left(1 + \frac{1}{C}\right) < 1 \quad (10)$$

Suppose for some $t > T(\omega,\epsilon)$, $\|X_t^2\| > C\epsilon$. Then, for $(z,a) \in \mathsf{Z}\times\mathsf{A}$,

$$\begin{aligned}U_t^2(z,a) &= \gamma V_t(Z_{t+1}) - \gamma V_\mu(Z_{t+1})\\ &= \gamma\Omega^\star(Q_t(Z_{t+1},\cdot)) - \gamma\Omega^\star(Q_\mu(Z_{t+1},\cdot))\\ &= \gamma\Bigg[\sum_{a\in\mathsf{A}}\pi_t(a \mid Z_{t+1})Q_t(Z_{t+1},a) - \Omega(\pi_t(\cdot \mid Z_{t+1}))-\\ &\quad \sum_{a\in\mathsf{A}}\pi^\star(a \mid Z_{t+1})Q_\mu(Z_{t+1},a) + \Omega(\pi^\star(\cdot \mid Z_{t+1}))\Bigg]\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \gamma \Bigg[\sum_{a \in \mathsf{A}} \pi_t(a \mid Z_{t+1}) Q_t(Z_{t+1}, a) - \Omega(\pi_t(\cdot \mid Z_{t+1})) - \\
&\quad \sum_{a \in \mathsf{A}} \pi_t(a \mid Z_{t+1}) Q_\mu(Z_{t+1}, a) + \Omega(\pi_t(\cdot \mid Z_{t+1})) \Bigg] \\
&\leq \gamma \sum_{a \in \mathsf{A}} \pi_t(a \mid Z_{t+1}) \big| Q_t(Z_{t+1}, a) - Q_\mu(Z_{t+1}, a) \big| \\
&\overset{(b)}{\leq} \gamma \|Q_t - Q_\mu\| = \gamma \|\Delta_t\| \\
&\leq \gamma \|X_t^0 + X_t^1\| + \gamma \|X_t^2\| \qquad (11a) \\
&\overset{(c)}{\leq} \gamma \epsilon + \gamma \|X_t^2\| \qquad (11b) \\
&\overset{(d)}{\leq} \gamma \left(1 + \frac{1}{C}\right) \|X_t^2\| \overset{(e)}{=} \kappa \|X_t^2\| \overset{(e)}{<} \|X_t^2\|, \qquad (11c)
\end{aligned}
$$

where (a) follows from the fact that we replace the argmax $\pi^\star$ with a different argument π_t in the second term, (b) follows from maximizing over all realizations of Z_{t+1} and $a \in \mathsf{A}$, (c) follows from (9), (d) follows from $\|X_t^2\| > C\epsilon$, (e) follows from (10). Thus, for any $t > T(\omega, \epsilon)$ and $\|X_t^2\| > C\epsilon$:

$$
\begin{aligned}
X_{t+1}^2(z,a) &= (1 - \alpha_t(z,a)) X_t^2(z,a) + \\
&\qquad \alpha_t(z,a) U_t^2(z,a) < \|X_t^2\| \\
\implies \|X_{t+1}^2\| &< \|X_t^2\|.
\end{aligned}
$$

Hence, when $\|X_t^2\| > C\epsilon$, it decreases monotonically with time. Hence, there are two possibilities: either

1) $\|X_t^2\|$ always remains above $C\epsilon$; or
2) it goes below $C\epsilon$ at some stage.

We consider these two possibilities separately.

Possibility (i): $\|X_t^2\|$ always remains above $C\epsilon$

We will now prove that $\|X_t^2\|$ cannot remain above $C\epsilon$ forever. The proof is by contradiction. Suppose $\|X_t^2\|$ remains above $C\epsilon$ forever. As argued earlier, this implies that $\|X_t^2\|$, $t \geq T(\omega, \epsilon)$, is a strictly decreasing sequence, so it must be bounded from above. Let $B^{(0)}$ be such that $\|X_t^2\| \leq B^{(0)}$ for all $t \geq T(\omega, \epsilon)$. Eq. (11c) implies that $\|U_t^2\| < \kappa B^{(0)}$. Then, we have for all $(z, a) \in \mathsf{Z} \times \mathsf{A}$ that

$$
\begin{aligned}
X_{t+1}^2(z,a) &\leq (1 - \alpha_t(z,a)) \|X_t^2\| + \alpha_t(z,a) \|U_t^2\| \\
&< (1 - \alpha_t(z,a)) \|X_t^2\| + \alpha_t(z,a) \kappa \|X_t^2\|
\end{aligned}
$$

which implies that $\|X_t^2\| \leq \|M_t^{(0)}\|$, where $\{M_t^{(0)}\}_{t \geq T(\omega,\epsilon)}$ is a sequence given by

$$
M_{t+1}^{(0)}(z,a) \leq (1 - \alpha_t(z,a)) M_t^{(0)}(z,a) + \alpha_t(z,a) \kappa B^{(0)}.
$$

Theorem 2.4 from [41] implies that $M_t^{(0)}(z,a) \to \kappa B^{(0)}$ and hence $\|M_t^{(0)}\| \to \kappa B^{(0)}$. Now pick an arbitrary $\bar{\epsilon} \in (0, (1-\kappa)C\epsilon)$. Thus, there exists a time $T^{(1)} = T^{(1)}(\omega, \epsilon, \bar{\epsilon})$ such that for all $t > T^{(1)}$, $\|M_t^{(0)}\| \leq B^{(1)} := \kappa B^{(0)} + \bar{\epsilon}$. Since $\|X_t^2\|$ is bounded by $\|M_t^{(0)}\|$, this implies that for all $t > T^{(1)}$, $\|X_t^2\| \leq B^{(1)}$ and, by (11c), $\|U_t^2\| \leq \kappa B^{(1)}$. By repeating the above argument, there exists a time $T^{(2)}$ such that for all $t \geq T^{(2)}$,

$$
\|X_t^2\| \leq B^{(2)} := \kappa B^{(1)} + \bar{\epsilon} = \kappa^2 B^{(0)} + \kappa \bar{\epsilon} + \bar{\epsilon},
$$

and so on. By (10), $\kappa < 1$ and $\bar{\epsilon}$ is chosen to be less than $C\epsilon$. So eventually, $B^{(m)} := \kappa^m B^{(0)} + \kappa^{m-1}\bar{\epsilon} + \cdots + \bar{\epsilon}$ must get below $C\epsilon$ for some m, contradicting the assumption that $\|X_t^2\|$ remains above $C\epsilon$ forever.

Possibility (ii): $\|X_t^2\|$ goes below $C\epsilon$ at some stage

Suppose that there is some $t > T(\omega, \epsilon)$ such that $\|X_t^2\| < C\epsilon$. Then (11a), (11b) and (10) imply that

$$
\|U_t^2\| \leq \gamma \|X_t^0 + X_t^1\| + \gamma \|X_t^2\| \leq \gamma\epsilon + \gamma C\epsilon < C\epsilon.
$$

Therefore,

$$
X_{t+1}^2(z,a) \leq (1 - \alpha_t(z,a)) \|X_t^2\| + \alpha_t(z,a) \|U_t^2\| < C\epsilon
$$

where the last inequality uses the fact that both $\|U_t^2\|$ and $\|X_{t+1}^2\|$ are both below $C\epsilon$. Thus, we have that

$$
X_{t+1}^2(z,a) < C\epsilon.
$$

Hence, once $\|X_{t+1}^2\|$ goes below $C\epsilon$, it stays there.

d) Implication

We have show that for sufficiently large $t > T(\omega, \epsilon)$, $X_t^2(z,a) < C\epsilon$. Since ϵ is arbitrary, this means that for all realizations $\omega \in \Omega^1$, $\|X_t^2\| \to 0$. Thus,

$$
\lim_{t \to \infty} \|X_t^2\| = 0, \quad a.s. \qquad (12)
$$

Putting everything together

Recall that we initially defined $\Delta_t = Q_t - Q_\mu$ and we split $\Delta_t = X_t^0 + X_t^1 + X_t^2$. Steps a) and b) together show that $\|X_t^0 + X_t^1\| \to 0$, a.s. and Step c) (12) shows us that $\|X_t^2\| \to 0$, a.s. Thus, by the triangle inequality,

$$
\lim_{t \to \infty} \|\Delta_t\| \leq \lim_{t \to \infty} \|X_t^0 + X_t^1\| + \lim_{t \to \infty} \|X_t^2\| = 0,
$$

which establishes that $Q_t \to Q_\mu$, a.s.

B. Proof of Theorem 2

The proof follows a similar style used in [12]. Define an error function between the converged value and the Q-learning iteration $\Delta_{t+1}^\ell := Q_{t+1}^\ell - Q_\mu^\ell$. Then, combine (5), (7) and (6) as follows for all (z, a).

$$
\begin{aligned}
\Delta_{t+1}^\ell(z,a) &= Q_{t+1}^\ell(z,a) - Q_\mu^\ell(z,a) \\
&= (1 - \alpha_t(z,a)) \Delta_t^\ell(z,a) \\
&\quad + \alpha_t(z,a) \left[U_t^{\ell,0}(z,a) + U_t^{\ell,1}(z,a) + U_t^{\ell,2}(z,a) \right] \qquad (13)
\end{aligned}
$$

where

$$
\begin{aligned}
U_t^{\ell,0}(z,a) &:= \left[r(S_t, A_t) - r_\mu^\ell(z,a) \right] \mathbb{1}_{\{Z_t = z, A_t = a\}}, \\
U_t^{\ell,1}(z,a) &:= \Big[\gamma \Omega^\star(Q_\mu^{[\![\ell+1]\!]}(Z_{t+1}, \cdot)) \\
&\quad - \gamma \sum_{z' \in \mathsf{Z}} P_\mu^\ell(z' \mid z, a) \Omega^\star(Q_\mu^{[\![\ell+1]\!]}(z', \cdot)) \Big] \mathbb{1}_{\{Z_t = z, A_t = a\}}, \\
U_t^{\ell,2}(z,a) &:= \Big[\gamma \Omega^\star(Q_t^{[\![\ell+1]\!]}(Z_{t+1}, \cdot)) \\
&\quad - \gamma \Omega^\star(Q_\mu^{[\![\ell+1]\!]}(Z_{t+1}, \cdot)) \Big] \mathbb{1}_{\{Z_t = z, A_t = a\}}.
\end{aligned}
$$

Note that we are adding the term $\gamma \Omega^\star(Q_\mu^{[\![\ell+1]\!]}(Z_{t+1}, \cdot)) \mathbb{1}_{\{Z_t = z, A_t = a\}}$ in $U_t^{\ell,1}(z,a)$ and

subtracting it in $U_t^{\ell,2}(z,a)$. We can now view (13) as a linear system with state Δ_t^ℓ and three inputs $U_t^{\ell,0}(z,a), U_t^{\ell,1}(z,a)$ and $U_t^{\ell,2}(z,a)$. Using the linearity, we can now split the state into three components $\Delta_{t+1}^\ell = X_{t+1}^{\ell,0} + X_{t+1}^{\ell,1} + X_{t+1}^{\ell,2}$, where the components evolve as follows for $i \in \{0,1,2\}$:

$$X_{t+1}^{\ell,i}(z,a) = (1-\alpha_t(z,a))X_t^{\ell,i}(z,a) + \alpha_t(z,a)U_t^{\ell,i}(z,a).$$

We will now separately show each $\|X_t^{\ell,i}\| \to 0$.

a) Convergence of component $X_t^{\ell,0}$

The proof for the convergence of component $X_t^{\ell,0}$ is similar to that given in [12]. The only difference from the RASQL proof of Theorem 1 is that the convergence has to established for each $\ell \in [L]$ in $\|X_t^{\ell,i}\| \to 0$. Note that this case is identical to the periodic case of [12], since the component $\|X_t^{\ell,i}\|$ does not involve any of the regularized terms.
The main result that is applied here is proposition 4 from [12], which establishes the exact convergence of $\|X_t^{\ell,i}\|$ when the underlying Markov chain is periodic.

b) Convergence of component $X_t^{\ell,1}$

The proof for the convergence of component $X_t^{\ell,1}$ is based on the argument given in [12]. Let W_t denote the tuple $(S_t, Z_t, A_t, S_{t+1}, Z_{t+1}, A_{t+1})$. Note that $\{W_t\}_{t\geq 1}$ is a periodic Markov chain and converges to a periodic limiting distribution $\bar{\zeta}_\mu^\ell$, where

$$\begin{aligned}&\bar{\zeta}_\mu^\ell(s,z,a,s',z',a')\\ &\quad= \zeta_\mu^\ell(s,z,a)\sum_{y'\in\mathsf{Y}} P(s',y'|s,a)\mathbb{1}_{\{z'=\phi(z,y',a)\}}\mu(a'|z').\end{aligned}$$

We use $\bar{\zeta}_\mu^\ell(s,z,a,\mathcal{S},\mathcal{Z},\mathcal{A})$ to denote the marginalization over the "future states" and a similar notation for other marginalizations. Note that $\bar{\zeta}_\mu^\ell(s,z,a,\mathcal{S},\mathcal{Z},\mathcal{A}) = \zeta_\mu^\ell(s,z,a)$. Define $V_t^{[\![\ell+1]\!]}$ as the value function associated with $Q_t^{[\![\ell+1]\!]}$, i.e., $V_t^{[\![\ell+1]\!]}(z) := \Omega^\star(Q_t^{[\![\ell+1]\!]}(z,\cdot))$. Fix $(z_\circ, a_\circ) \in \times\mathsf{Z}\times\mathsf{A}$ and define

$$\begin{aligned}h_P(W_t;\ell,z_\circ,a_\circ) = \Big[&\gamma V_\mu^{[\![\ell+1]\!]}(Z_{t+1}) -\\ &\gamma\sum_{\bar{z}\in\mathsf{Z}} P_\mu^\ell(\bar{z}|z_\circ,a_\circ)V_\mu^{[\![\ell+1]\!]}(\bar{z})\Big]\mathbb{1}_{\{Z_t=z_\circ, A_t=a_\circ\}}.\end{aligned}$$

Then the process $\{X_t^{\ell,1}(z,a)\}_{t\geq 1}$ is given by the stochastic iteration

$$\begin{aligned}X_{t+1}^{\ell,1}(z_\circ,a_\circ) = (1-\alpha_t^\ell(z_\circ,a_\circ))&X_t^{\ell,1}(z_\circ,a_\circ)\\ &+\alpha_t^\ell(z_\circ,a_\circ)h_P(W_t;\ell,z_\circ,a_\circ).\end{aligned}$$

As mentioned earlier, the process $\{W_t\}_{t\geq 1}$ is a periodic Markov chain. From the periodic Markov chain result of proposition 4 from [12], we have that: $\{X_t^{\ell,1}(z_\circ,a_\circ)\}_{t\geq 1}$ converges a.s. to the following periodic limits

$$\begin{aligned}&\lim_{t\to\infty} X_t^{\ell,1}(z_\circ,a_\circ) = \sum_{\substack{s,z,a\in\mathsf{S}\times\mathsf{Z}\times\mathsf{A}\\ s',z',a'\in\mathsf{S}\times\mathsf{Z}\times\mathsf{A}}} \bar{\zeta}_\mu^\ell(s,z,a,s',z',a')\\ &\quad h_P(s,z,a,s',z',a';\ell,z_\circ,a_\circ)\\ &= \gamma\Bigg[\sum_{z'\in\mathsf{Z}}\bar{\zeta}_\mu^\ell(\mathcal{S},z_\circ,a_\circ,\mathcal{S},z',\mathcal{A})V_\mu^{[\![\ell+1]\!]}(z')\Bigg]\\ &\quad-\Bigg[\gamma\bar{\zeta}_\mu^\ell(\mathcal{S},z_\circ,a_\circ,\mathcal{S},\mathcal{Z},\mathcal{A})\sum_{\bar{z}\in\mathsf{Z}}P_\mu^\ell(\bar{z}|z_\circ,a_\circ)V_\mu^{[\![\ell+1]\!]}(\bar{z})\Bigg]\\ &= 0\end{aligned}$$

where the last step follows from the fact that $\bar{\zeta}_\mu^\ell(\mathsf{S},z_\circ,a_\circ,\mathsf{S},\mathcal{Z},\mathcal{A}) = \zeta_\mu^\ell(z_\circ,a_\circ)$ and $\bar{\zeta}_\mu^\ell(\mathsf{S},z_\circ,a_\circ,\mathsf{S},z',\mathsf{A}) = \zeta_\mu^\ell(z_\circ,a_\circ)P_\mu^\ell(z'|z_\circ,a_\circ)$.

c) Convergence of component $X_t^{\ell,2}$

The convergence of the $X_t^{\ell,2}$ component is based on [11], [12] but requires some additional considerations due to the regularization term. We start by defining:

$$\begin{aligned}\pi_t^\ell(\cdot\mid z) &= \arg\max_{\xi\in\Delta(\mathsf{A})}\sum_{a\in\mathsf{A}}\xi(a)Q_t^{[\![\ell+1]\!]}(z,a) - \Omega(\xi)\\ \pi^{\ell,\star}(\cdot\mid z) &= \arg\max_{\xi\in\Delta(\mathsf{A})}\sum_{a\in\mathsf{A}}\xi(a)Q_\mu^{[\![\ell+1]\!]}(z,a) - \Omega(\xi).\end{aligned}$$

In the previous steps, we have shown that $\|X_t^{\ell,i}\| \to 0$ a.s., for $i \in \{0,1\}$. Thus, we have that $\|X_t^{\ell,0} + X_t^{\ell,1}\| \to 0$ a.s. Arbitrarily fix an $\epsilon > 0$. Therefore, there exists a set Ω^1 of measure one and a constant $T(\omega,\epsilon)$ such that for $\omega \in \Omega^1$, all $t > T(\omega,\epsilon)$, and $(z,a) \in \times\mathsf{Z}\times\mathsf{A}$, we have

$$X_t^{\ell,0}(z,a) + X_t^{\ell,1}(z,a) < \epsilon. \tag{14}$$

Now pick a constant C such that

$$\kappa := \gamma\left(1+\frac{1}{C}\right) < 1 \tag{15}$$

Suppose for some $t > T(\omega,\epsilon)$, $\|X_t^{\ell,2}\| > C\epsilon$. Then, for $(\ell,z,a) \in L\times\mathsf{Z}\times\mathsf{A}$,

$$\begin{aligned}&U_t^{\ell,2}(z,a) = \gamma V_t^{[\![\ell+1]\!]}(Z_{t+1}) - \gamma V_\mu^{[\![\ell+1]\!]}(Z_{t+1})\\ &= \gamma\Omega^\star(Q_t^{[\![\ell+1]\!]}(Z_{t+1},\cdot)) - \gamma\Omega^\star(Q_\mu^{[\![\ell+1]\!]}(Z_{t+1},\cdot))\\ &= \gamma\Bigg[\sum_{a\in\mathsf{A}}\pi_t^\ell(a\mid Z_{t+1})Q_t^{[\![\ell+1]\!]}(Z_{t+1},a) - \Omega(\pi_t^\ell(\cdot\mid Z_{t+1})) -\\ &\quad\sum_{a\in\mathsf{A}}\pi^{\ell,\star}(a\mid Z_{t+1})Q_\mu^{[\![\ell+1]\!]}(Z_{t+1},a) + \Omega(\pi^{\ell,\star}(\cdot\mid Z_{t+1}))\Bigg]\\ &\overset{(a)}{\leq}\gamma\Bigg[\sum_{a\in\mathsf{A}}\pi_t^\ell(a\mid Z_{t+1})Q_t^{[\![\ell+1]\!]}(Z_{t+1},a) - \Omega(\pi_t^\ell(\cdot\mid Z_{t+1})) -\\ &\quad\sum_{a\in\mathsf{A}}\pi_t^\ell(a\mid Z_{t+1})Q_\mu^{[\![\ell+1]\!]}(Z_{t+1},a) + \Omega(\pi_t^\ell(\cdot\mid Z_{t+1}))\Bigg]\\ &\leq\gamma\sum_{a\in\mathsf{A}}\pi_t^\ell(a\mid Z_{t+1})\big|Q_t^{[\![\ell+1]\!]}(Z_{t+1},a) - Q_\mu^{[\![\ell+1]\!]}(Z_{t+1},a)\big|\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} \gamma \|Q_t^{[\![\ell+1]\!]} - Q_\mu^{[\![\ell+1]\!]}\| = \gamma \|\Delta_t^\ell\| \\
&\leq \gamma \|X_t^{\ell,0} + X_t^{\ell,1}\| + \gamma \|X_t^{\ell,2}\| \\
&\overset{(c)}{\leq} \gamma\epsilon + \gamma \|X_t^{\ell,2}\| \\
&\overset{(d)}{\leq} \gamma \left(1 + \frac{1}{C}\right) \|X_t^{\ell,2}\| \overset{(e)}{=} \kappa \|X_t^{\ell,2}\| \overset{(e)}{<} \|X_t^{\ell,2}\|,
\end{aligned}
$$

where (a) follows from the fact that we replace the argmax $\pi^{\ell,\star}$ with a different argument π_t^ℓ in the second term, (b) follows from maximizing over all realizations of Z_{t+1} and $a \in \mathsf{A}$, (c) follows from (14), (d) follows from $\|X_t^{\ell,2}\| > C\epsilon$, (e) follows from (15). Thus, for any $t > T(\omega, \epsilon)$ and $\|X_t^{\ell,2}\| > C\epsilon$:

$$
\begin{aligned}
X_{t+1}^{\ell,2}(z,a) = (1 - \alpha_t^\ell(z,a)) X_t^{\ell,2}(z,a) + \\
\alpha_t^\ell(z,a) U_t^{\ell,2}(z,a) < \|X_t^{\ell,2}\| \\
\implies \|X_{t+1}^{\ell,2}\| < \|X_t^{\ell,2}\|.
\end{aligned}
$$

Hence, when $\|X_t^{\ell,2}\| > C\epsilon$, it decreases monotonically with time. Hence, there are two possibilities: either

1) $\|X_t^{\ell,2}\|$ always remains above $C\epsilon$; or
2) it goes below $C\epsilon$ at some stage.

These two cases must be considered separately. The proof follows exactly the same steps in the proof of theorem 1 given in appendix A, which finally gives us:

$$
\lim_{t\to\infty} \|X_t^{\ell,2}\| = 0, \quad a.s. \tag{17}
$$

Putting everything together Recall that we initially defined $\Delta_t^\ell = Q_t^\ell - Q_\mu^\ell$ and we split $\Delta_t^\ell = X_t^{\ell,0} + X_t^{\ell,1} + X_t^{\ell,2}$. Steps $a)$ and $b)$ together show that $\|X_t^{\ell,0} + X_t^{\ell,1}\| \to 0$, a.s. and Step $c)$ (17) shows us that $\|X_t^{\ell,2}\| \to 0$, a.s. Thus, by the triangle inequality,

$$
\lim_{t\to\infty} \|\Delta_t\| \leq \lim_{t\to\infty} \|X_t^{\ell,0} + X_t^{\ell,1}\| + \lim_{t\to\infty} \|X_t^{\ell,2}\| = 0,
$$

which establishes that $Q_t^\ell \to Q_\mu^\ell$, a.s.