Consistency of Multi-Agent Batch Reinforcement Learning

Clemens Possnig

Vancouver School of Economics, University of British Columbia

April 29, 2023

Working Paper

Link to current version

Abstract. This paper provides asymptotic results for a class of model-free actor-critic batch - reinforcement learning algorithms in the multi-agent setting. At each period, each agent faces an estimation problem (the critic, e.g. a value function), and a policy updating problem. The estimation step is done by parametric function estimation based on a batch of past observations. Agents have no knowledge of each others incentives and policies. I give sufficient conditions on the environment, growth rate of the batch-size and speed of their policy-stepsizes, so that each agent's parametric function estimator is consistent in the following sense: For large t, the optimal empirical parameter θ_t is close to a true population parameter θ_t^*, depending on t only through the current period's policy profile. In other words, agents can learn to best respond in spite of the nonstationarity in their problem due to the multi-agent environment.

These sufficient conditions are useful in the asymptotic analysis of multi-agent learning, e.g. in the application of long-run characterisations using stochastic approximation techniques.

Keywords. Multi-Agent Reinforcement Learning, Batch-Reinforcement Learning, Consistency.

I thank Nima Haghpanah, Vadim Marmer, and Kevin Song for helpful discussions. I thank the participants at EC 22, GTA22, CORS/INFORMS 22, and CETC 22 for insightful comments.

1. Introduction

This paper develops asymptotic results for the multi-agent reinforcement learning (MARL) setting which will help analyse what behaviors can be learned by algorithms that interact with one another.

Reinforcement Learning (RL) algorithms are updating rules meant for the learning of optimal policies or value functions for a given problem. Such algorithms are commonly used to solve Markov decision problems. In general, RL updating rules move policies towards actions that have performed well in the past (i.e., such actions are *reinforced*), and away from actions that perform poorly, based on some objective function. Commonly, a RL agent estimates a value function, and updates policies based on that value function. If estimates converge to the correct value function, policies will usually also converge to optimal policies, and learning is successful. For a thorough introduction to RL see Sutton and Barto (2018). Multiple recent surveys on MARL and related theoretical results exist, notably Zhang, Yang, and Başar (2021), and Hernandez-Leal, Kartal, and Taylor (2019).

Recent years have brought significant advancements in the literature on multi-agent reinforcement learning (MARL). Such algorithms have proven successful in various strategic settings, such as the games of Go and Poker, and also autonomous driving. Despite these widespread successes, the performance of multi-agent learning algorithms is commonly verified only empirically through simulations, while theoretical results are relatively lacking. The aim of this paper is to provide novel theoretical results that are useful in determining the asymptotic behavior of multi-agent systems.

This paper considers the setting of actor-critic batch RL, which involves the mixture of offline (training performance measures on a batch of observations) and online (updating policies during play) approaches (c.f. Busoniu et al. (2017), Chapter 3).

A main problem when establishing theoretical results for the MARL setting is the inherent nonstationarity of the environment faced by each agent. This comes from the fact that each agent's observations are drawn from distributions dependent on each other agent's policies, which themselves are moving over time. At the same time, an agent needs to find an optimal decision at any given period, were for optimality only the current policies of their opponents matter. This introduces the problem commonly referred ot as 'tracking a moving target': To find a best response, agents need to estimate a value function based on their opponent's *current* policy, but can only use data generated from their opponent's *past* strategies.

The batch-setting we study allows a useful solution to this issue. The name refers to the fact that the historical data used for estimating the performance measure is only a most recent window (the batch) of past observations, not the full available set of observations.

Akin to the idea of two-timescale approaches, which are well known in the literature (c.f. Borkar (2009), Chapter 6), it will be true that the batch each agent uses to train their performance measure grows at a speed that is slower than the convergence rate of each agent's policy-stepsize. This motivates the intuition that the most recent observations made by each agent are generated from distributions that are quite similar. Once that is true, I apply techniques developed and used in econometric theory due to Newey and McFadden (1994) to show that tracking the moving target becomes feasible. The Batch-RL setting is in contrast to more commonly known online-only RL schemes, which at every period t incorporate only the new information that has been accrued to adapt their performance measure estimator (see for example stochastic gradient descent methods for parametric Q-estimation in Sutton and Barto (2018)). The method of batch-learning is computationally more costly at each period, since a separate optimization routine is run at every period. However, we will see that this can put us at an advantage when it comes to estimation and consistency given nonstationarity.

The setting I study is one of discrete state spaces but interval action spaces, as it is motivated by the oligopoly-competition setting as studied in Possnig (2022). This implies the requirement of using function approximation in the estimation step. Results carry over to the finite actions and states case, modulo adjusted notation. I do not make assumptions on the strategic nature of the interaction each agent faces, i.e. I make no requirement on the game being played to be zero-sum, cooperative or otherwise as is commonly done in the MARL setting. This paper focuses on results on a more fundamental level: we are only concerned with giving guarantees for the function approximator of each agent to be well-behaved in an appropriate sense. Once this can be verified, other techniques such as stochastic approximation can be applied to paint a full picture of the asymptotic behavior of the policy-profile process implied by the MARL updating scheme. In Possnig (2022), I give such an analysis for Markov games of discrete states and interval actions under the assumption that function approximators are well-behaved in the sense developed in this paper.

To the best of my knowledge, this is the first paper providing MARL asymptotic consistency analysis for the actor-critic batch RL agents as considered here. Two-timescale approaches as defined in Borkar (2009), chapter 6 have a connection to this paper along the intuitions used to to tackle nonstationarity. Perolat, Piot, and Pietquin (2018) construct a stochastic approximation result for the two-timescale actor-critic scheme in the discrete state-action setting, while Perkins and Leslie (2013) consider asynchronous two-timescale schemes allowing for multi-valued updates. For a more thorough discussion on recent advancements in the MARL literature, consider Zhang, Yang, and Başar (2021).

The technique of using the advantages of 'forgetfulness' (i.e. a small batch of recent observations to be used in estimation) in the face of nonstationarity is not novel in this paper. The literature on multi-armed bandit learning under nonstationarity using this idea has seen multiple recent advancements, e.g. Cheung, Simchi-Levi, and Zhu (2020), which consider finite single agent settings and focus on regret bounds. Zhou et al. (n.d.) provide a regret bound analaysis of an RL with linear function approximation in nonstationary environments. Khetarpal et al. (2022) give a thorough survey on the literature of reinforcement learning approaches to nonstationarity.

This paper is organized as follows: Section 2 gives the consistency result in full generality. Section 3 shows how the result applies to common learning rules such as actor-critic Q learning and gradient learning, and finishes with a Corollary that shows how the results developed here apply to Assumption 3 in Possnig (2022). All proofs are in the appendix.

2. Consistency

We begin by giving a general consistency result on the estimation-step for batch-RL algorithms. We assume there are n algorithmic agents, a finite state space S, a compact interval action space $A^i \subset \mathbb{R}$, with $A = \times_i A^i$, a twice differentiable, bounded payoff function $u^i : A \times S \mapsto \mathbb{R}$. We define the state transition probability to be $P_{ss'}[a] \geq 0$ for all $a \in A$ and $s, s' \in S$, where throughout we will maintain an assumption of irreducibility stated below.

We assume that each agent follows a batch-RL algorithm to update their policies $\rho_t^i : S \mapsto A^i$ over time. Let the resulting compact policy profile space be called Γ.

Assumption 1. *For all $\rho \in \Gamma$, the Markov chain induced by $P_{ss'}[\rho(s)]$ is irreducible and aperiodic.*[1]

We maintain this assumption throughout the paper.

The updates are done using a parametric estimator of an underlying performance measure; this can be e.g. a Q-value function as discussed in Section 3. In this section the statement will be given in general terms, agnostic to the exact definition of the performance measure. All we need is that the parametric estimator can be expressed as the minimizer of a loss function.

We will call the parametric estimator $F^i(\rho^i, \theta^i)$. We assume that F^i is continuous in both arguments for all i, and that parameters $\theta \in \Theta$ for a set $\Theta \subset \mathbb{R}^m$ compact.

The consistency result this paper will establish is of the following form: each agent will use data generated from interactions with each other over time to estimate their parameter

[1] For Definitions see e.g. Appendix A in Puterman (2014)

vector θ^i, while ρ_t^i are being updated concurrently. As a result, it is likely that an optimal θ_t^{i*} moves with time also, generating a moving-target problem. We will then prove:

Under suitable Assumptions, each agent's estimated θ_t^i behaves in the following way:

$$\|\theta_t^i - \theta_t^{i*}\| \to_P 0,$$

as $t \to \infty$, and also that in a sense that will become apparent, θ_t^{i*} depends on time t *only* through current period's policy profile ρ_t. This result is desirable as RL agents commonly face an issue of computing policies optimal with respect to the current distributional environment they face, but have only access to data generated from past distributional environments. This issue is absent in the single-agent stationary Markov Decision Problem, but salient in the multi-agent learning of focus here.

First, let $\mathcal{Z} \subset \mathbb{R}^d$ be a space of observations used in the construction of the loss function. Each period, a realization $Z_t \in \mathcal{Z}$ is generated after each algorithm chooses their actions. For example, $Z_t^i = \langle s_t, a_t^i, u_t^i, s_{t+1}\rangle$ would be the typical tuple of current state s_t, current action, payoff, and next state observed at the end of each period by a model-free[2] algorithm. We define a bounded function $\ell(Z, \theta) \in \mathcal{U} \subset \mathbb{R}$, Lipschitz in both arguments as the basic building block of the loss function.

Each algorithm uses only a batch of the most recent observations to construct their empirical loss function. Define a sequence $0 < K_t < t$ with $K_t \in \mathbb{N}$ such that $K_t \to \infty$ with t, and let

$$W_t = \{k : t - K_t + 1 \leq k \leq t\},$$

be the batch of periods used in the constuction of the loss function. We define $\underline{W}_t = t - K_t + 1$ as the first period of the batch. Then

$$L_t(\theta) = \frac{1}{K_t} \sum_{k \in W_t} \ell(z_k, \theta),$$

is the empirical loss. Then we define

$$\theta_t \in \arg\min_{\theta \in \Theta} L_t(\theta),$$

as the empirical parametric minimizer. Let $\boldsymbol{\rho}_t = \{\rho_k\}_{\underline{W}_t \leq k \leq t}$ denote batch-sequences of variables. Our first assumption is on the smoothness of the loss function and the behavior of its conditional expectation:

Assumption 2. *There exists a function $\phi(\rho, s, \theta) \in \mathbb{R}$ Lipschitz in the first and third arguments with*

[2] Model-free algorithms estimate their performance measure without a model of their environment. See Sutton and Barto (2018), e.g. Chapter 6.

(1)

$$\mathbb{E}[\ell(Z_t,\theta) \mid \boldsymbol{\rho}_t, \boldsymbol{s}_t] = \phi(\rho_t, s_t, \theta).$$

(2)

$$\lim_{t\to\infty} \frac{1}{K_t} \sum_{k\in W_t} \mathbb{E}C_1(Z_k) < \infty,$$

$$\lim_{t\to\infty} \frac{1}{K_t} \sum_{k\in W_t} \mathbb{E}C_2(\rho_k, s_k) < \infty,$$

and

$$\sup_{\theta\in\Theta} \max_{s\in S} C_3(\theta, s) < \infty,$$

Where $C_1(Z), C_2(\rho,s), C_3(\theta,s)$ *are bounded, nonnegative functions that exist by the Lipschitz properties of* ℓ, ϕ *so that:*

$$\begin{aligned} |\ell(Z,\theta) - \ell(Z,\theta')| &\le C_1(Z)\|\theta-\theta'\|, \\ |\phi(\rho,s,\theta) - \phi(\rho,s,\theta')| &\le C_2(\rho,s)\|\theta-\theta'\|, \\ |\phi(\rho,s,\theta) - \phi(\rho',s,\theta)| &\le C_3(\theta,s)\|\rho-\rho'\|. \end{aligned}$$

Assumption 2 (1) is a distributional assumption that can be satisfied if Z_t is Markov given current policy profile ρ_t and state s_t, as will be discussed in Section 3.

Now we are ready to state the actor-critic updating schemes studied in this paper.

Definition 1. *For each agent,* ρ^i_t *is updated in the following way:*

$$\rho^i_{t+1} = \rho^i_t + \alpha_t\big[F^i(\rho^i_t, \theta^i_t) + M^i_{t+1}\big],$$

where $F^i(\rho^i_t, \theta^i_t)$ *is the bounded parametric function to estimate the population objective,* α_t *is a decreasing stepsize sequence satisfying the Robbins-Monro condition:*

$\alpha_t \to 0$ *with*

$$\sum_{t=0}^{\infty} \alpha_t = \infty; \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty,$$

and M^i_{t+1} *is an almost surely bounded martingale-difference noise based on an increasing sequence of sigma algebras* $\mathcal{G}_t$*.*

The next Assumption will ensure that the data used by the loss functions appropriately adjusts for the fact that a moving target has to be followed:

Assumption 3. *Assume that*

$$K_t \alpha_{t-K_t} \to 0,$$

as $t \to \infty$*.*

Define $\mu_{\rho_t}(s) \in (0,1)$ for every s as the unique invariant state distribution if policy profile ρ_t where played forever, which exists by our irreducibility assumption 1. Further, define $\boldsymbol{\rho}_{\underline{W}_t:k} = \{\rho_j\}_{\underline{W}_t \le j \le k}$ as a truncated sequence of policy profiles. Then let

$$\lambda_k(s, \boldsymbol{\rho}_{\underline{W}_t:k}, s_{\underline{W}_t}) = \mathbf{P}\big(s_k = s \mid \boldsymbol{\rho}_{\underline{W}_t:k}, s_{\underline{W}_t}\big),$$

be the likelihood of reaching state s in period $k \in W_t$, if over periods $\underline{W}_t, ..., k$, $\boldsymbol{\rho}_t$ is the policy profile sequence played, and $s_{\underline{W}_t}$ is the initial state in the first batch-period. Also let $\lambda_k(s, \rho, s_t)$ be the counterpart where $\rho_l = \rho$ in all periods $\underline{W}_t, ..., k$.

Assumption 4.

(1) Assume for all t, $\lambda_k(s, \rho, s_t)$ and μ_ρ are Lipschitz in ρ with Liptschitz constants bounded uniformly over S.

(2) There exists $c_P > 0$ and $1 \le k < \infty$ such that for all $s', s \in S$

$$\inf_{\rho \in \Gamma} \mathbf{P}[s_k = s' \mid s_0 = s, \rho] \ge c_P.$$

Assumption 4 (2) is slightly stronger than our irreducibility assumption on the Markov chain over s. It ensures that in the asymptotic analysis we can safely assume $\lambda_k > 0$ for t large enough.

Next, define

$$\Lambda_t(s, \boldsymbol{\rho}_t, s_{\underline{W}_t}) = \frac{1}{K_t} \sum_{k \in W_t} \lambda_k(s, \boldsymbol{\rho}_{\underline{W}_t:k}, s_{\underline{W}_t}),$$

and let $\ell(Z_k, \theta)_s = \ell(Z_k, \theta)\mathbf{1}\{s_k = s\}$ for all $s \in S$.

The population counterpart to $L_t(\theta)$ is then defined as

$$L_t^*(\theta, \boldsymbol{\rho}_t) = \sum_{s \in S} \Lambda_t(s, \boldsymbol{\rho}_t, s_{\underline{W}_t}) \frac{\frac{1}{K_t}\sum_{k \in W_t} \mathbb{E}\big[\ell(Z_k, \theta)_s \mid \boldsymbol{\rho}_k\big]}{\Lambda_t(s, \boldsymbol{\rho}_t, s_{\underline{W}_t})}.$$

Next, define

$$L_t^*(\theta, \rho_t) = \sum_{s \in S} \Lambda_t(s, \rho_t, s_{\underline{W}_t}) \frac{\frac{1}{K_t}\sum_{k \in W_t} \mathbb{E}\big[\ell(Z_k, \theta)_s \mid \rho_t\big]}{\Lambda_t(s, \rho_t, s_{\underline{W}_t})},$$

as the population loss in the case where in all periods $k \in W_t$, ρ_t is the policy profile played.

The t-limit of this loss function will play an important role in our results:

Lemma 1. *Suppose Assumptions 1, 2, and 4 hold. Then*

$$\lim_{t\to\infty} \sup_{\theta\in\Theta,\rho\in\Gamma} \left\| L_t^*(\theta,\rho) - \sum_{s\in S} \mu_{\rho_t}(s)\phi(\rho,s,\theta) \right\| = 0.$$

Proof. All proofs can be found in the Appendix. □

From now on, we define

$$L_\infty^*(\theta,\rho_t) = \sum_{s\in S} \mu_{\rho_t}(s)\phi(\rho_t,s,\theta),$$

and

$$\theta^*(\rho_t) = \arg\min_{\theta\in\Theta} L_\infty^*(\theta,\rho_t),$$

as the optimal population parameter. Notice that θ^* is a random variable given that ρ_t is a random variable.

The next Assumption ensures that for any trajectories, there is a unique minimizer θ^*. Define $B(x,\varepsilon)$ as the ε-ball centered at x.

Assumption 5 (Identification)**.** *For any ρ, any $\varepsilon > 0$ and $\theta \notin B(\theta^*(\rho),\varepsilon)$, there exists $\delta > 0$ such that:*

$$L_\infty^*(\theta,\rho) \geq L_\infty^*(\theta^*(\rho),\rho) + \delta.$$

We can now state the main result of this paper:

Theorem 1. *Impose Assumptions 1 - 5. Then for any sequence ρ_t satisfying Definition 1, and for any $\varepsilon > 0$,*

$$\mathbf{P}\big(\|\theta_t - \theta^*(\rho_t)\| > \varepsilon\big) \to 0,$$

as $t \to \infty$.

This result is useful in the following sense: in general the function approximation parameter vector θ_t will depend on the whole policy profile trajectory $\boldsymbol{\rho}_t$. Given that opponent's policies are moving over time, this can result in a quite hard to interpret estimator and can lead to bad performance of the iteration ρ_t^i. However, the Assumptions taken in the Theorem ensure that in fact, θ_t will, for large enough t, depend on the trajectory of policy profiles only through the most *current* period t. Thus, the resulting loss function behaves as if each agent knew their opponent's current policy, and sampled observations from that policy to estimate their loss function.

Furthermore, the limiting population loss L_∞^* can represent desirable population loss functions commonly used in the literature, as will be shown in the next section. This will allow to make much more accurate predictions about future behavior of opponents, and therefore better performance of the algorithm as will be seen in Section 3.

3. Applications

Given the setup defined in the previous section, a valid performance measure would be based on the commonly used action-value function $Q_i^* : S \times A^i \mapsto \mathbb{R}$. Given a reward function $u^i : A^i \times S \mapsto \mathbb{R}$, it is defined implicitly as

$$Q_i^*(s,a) = u^i(a,s) + \delta \mathbb{E}\big[\max_{a' \in A} Q_i^*(s',a') \mid a, s\big]. \tag{1}$$

An extensive literature of reinforcement learning theory has focused on estimating this function. In what follows, individual identifiers i will dropped whenever the tradeoff between rigor and ease of notation favors the latter. In the single agent setting, where states evolve according to a controlled markov chain, many convergence results exist for estimators of Q^*, starting with the seminal results in Watkins (1989). See Sutton and Barto (2018) for a thorough exposition of learning algorithms related to Q^*.

Q-learning algorithms are algorithms used to estimate Q^*, and compute an optimal policy based on it. A class of algorithms fitting our Batch-RL framework would be versions of *Fitted Q-Iteration* (FQI), (Ernst, Geurts, and Wehenkel (2005), and Busoniu et al. (2017) Chapter 3 for a general discussion) as will be introduced below. Such iterations are meant to minimize parameters based on what is often called the squared Bellman-loss:

$$\ell(Z_t, \theta) = \Big[u_t + \delta \max_{a'} Q(s_{t+1}, a', \theta) - Q(s_t, a_t, \theta)\Big]^2, \tag{2}$$

where we let $Z_t = \langle s_t, a_t, u_t, s_{t+1} \rangle$. Suppose that each agent samples actions using a randomized policy $\vec{\rho}_t^{\,i}$ based on their iteration policy ρ_t^i. We will assume for simplicity here that $\mathbb{E}\vec{\rho}_t^{\,i} = \rho_t^i$, with full support on A^i for all states. In that case, we have

$$\mathbb{E}\big[\ell(Z_t,\theta) \mid \boldsymbol{\rho}_t, \boldsymbol{s}_t\big] = \mathbb{E}\big[\ell(Z_t,\theta) \mid \rho_t, s_t\big],$$

by the Markov property, and thus Assumption 2 (1) is satisfied. If we then impose Assumptions 1- 5, we get that empirical minimizer θ_t approaches, in probability,

$$\theta^*(\rho_t) \in \arg\min_{\theta \in \Theta} L_\infty^*(\theta, \rho_t),$$

where $L_\infty^*(\theta, \rho_t)$ is the mean-squared Bellman loss:

$$\sum_{s \in S} \mu_{\rho_t}(s) \mathbb{E}\Big[\big(u_t + \delta \max_{a'} Q(s_{t+1}, a', \theta) - Q(s_t, a_t, \theta)\big)^2 \mid \rho_t, s_t = s\Big]. \tag{3}$$

The mean-squared Bellman loss represents a desirable population loss commonly studied in the literature (see for example Sutton and Barto (2018), e.g. Chapters 9, 11).

Two important examples of algorithms based on such a loss function then are Actor-Critic Q- learning, for which

$$F^i(\rho_t^i, \theta_t^i) = \big\{\arg\max_{a \in A_i} Q^i(s, a, \theta_t^i)\big\}_{s \in S}, \tag{4}$$

and actor-critic gradient learning, where gradient here refers to a gradient in policies:

$$F^i(\rho_t^i, \theta_t^i) = \Big\{\frac{\partial Q^i(s, a, \theta_t^i)}{\partial a}\Big\}_{s \in S}, \tag{5}$$

where $\{\}_{s \in S}$ is to be understood as stacking a vector over $s \in S$.

Now we can verify that the result given in Theorem 1 can be applied to show that a given algorithm falls into the class developed in Possnig (2022). For a fixed opponent profile ρ_t^{-i}, we can define

$$Q_i^*(s, a, \rho_t^{-i}) = u^i(a, \rho_t^{-i}(s), s) + \delta \mathbb{E}\big[\max_{a' \in A} Q_i^*(s', a', \rho_t^{-i}) \mid a, s, \rho_t^{-i}(s)\big],$$

the action-value function in a repeated game where opponents play ρ_t^{-i} forever. The next result will show that for sufficient conditions given in this paper, Assumption 3 in Possnig (2022) holds. For convenience, we re-state this assumption here:

Assumption 6 (Assumption 3 in Possnig (2022))**.** *Define*

$$\chi_t^i \equiv \sup_{(s,a) \in S \times A,} \big\|Q^i(s, a; \theta_t) - Q_i^*(s, a, \rho_t^{-i}) - g^i(s, a, \rho_t^{-i})\big\|,$$

where for each i there exists a bounded function $g^i(s, a, \rho^{-i})$, $\mathcal{C}^2$ in a, ρ^{-i} such that

(i) There is an increasing sequence of σ-algebras $\{\mathcal{F}_t\}_{t \geq 0}$ such that

$$\chi_t^i - \mathbb{E}[\chi_t^i]$$

is a martingale difference sequence given $\mathcal{F}_t$,

(ii)

$$\sup_t \mathbb{E}\big[(\chi_t^i)^2\big] < \infty,$$

(iii) There exists a sequence $\zeta_t \geq \alpha_t$ with $\zeta_t \to 0$ such that

$$\lim_{t \to \infty} \Big\| \sum_{k=t}^{\infty} \zeta_k \mathbb{E}[\chi_k^i] \Big\| = 0.$$

Theorem 1 then allows us to conclude the following, under some additional regularity restrictions:

Lemma 2. *Suppose Assumptions 1 - 5 are satisfied, and algorithms update according to Definition 1. Write*

$$g^i(s, a, \rho_t) = Q_i^*(s, a, \rho_t^{-i}) - Q^i(s, a, \theta^*(\rho_t)).$$

Assume that, for all ρ_t and i

(i) $u^i(a, s)$ *is bounded below and above.*

(ii) $Q_i(s, a, \theta)$ *is twice differentiable in θ for all s, a.*

(iii) $L_\infty^{i*}(\theta, \rho_t)$ *is twice differentiable in θ on a small neighborhood $\mathcal{N}$ of $\theta^{i*}(\rho_t)$,*

(iv) $\frac{\partial}{\partial \theta} L_t^i(\theta^{i*}(\rho_t)) = O_P(n^{-\frac{1}{2}})$,

(v)

$$\sup_{\theta \in \Theta} \left\| \frac{\partial^2}{\partial \theta \partial \theta'} L_t^i(\theta) - B^i(\theta) \right\| = o_P(1),$$

for some nonstochastic matrix $B^i(\theta)$ such that $B^i()$ is continuous and positive definite at $\theta^{i*}(\rho_t)$.

Then Assumption 3 in Possnig (2022) holds.

4. Conclusion

This paper gives sufficient conditions on the payoff structures, state evolution, and hyperparameters of batch-RL algorithms so that their batch-estimation procedure has a tractable analytical interpretation. The setting studied here is one of discrete states and interval action spaces. However, it is likely that an extension can be constructed for more general state spaces, which is subject of further study here.

The assumption throughout this paper is that each agent uses parametric function estimation in the classical sense, where the number of parameters is finite and *smaller* than the number of observations. This precludes the analysis of Deep RL methods, which by definition overparametrize. However, recent advancements in the convergence analysis of Deep RL for function approximation (Ramaswamy and Hullermeier (2021)) allow for optimism that an extension to this paper can be made that appropriately applies to Deep RL methods.

References

Borkar, Vivek S (2009). *Stochastic approximation: a dynamical systems viewpoint.* Vol. 48. Springer.

Busoniu, Lucian et al. (2017). *Reinforcement learning and dynamic programming using function approximators*. CRC press.

Cheung, Wang Chi, David Simchi-Levi, and Ruihao Zhu (2020). "Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism". In: *International Conference on Machine Learning*. PMLR, pp. 1843–1854.

Ernst, Damien, Pierre Geurts, and Louis Wehenkel (2005). "Tree-based batch mode reinforcement learning". In: *Journal of Machine Learning Research* 6, pp. 503–556.

Freedman, Ari (2017). "Convergence theorem for finite markov chains". In: *Proc. REU*.

Hernandez-Leal, Pablo, Bilal Kartal, and Matthew E Taylor (2019). "A survey and critique of multiagent deep reinforcement learning". In: *Autonomous Agents and Multi-Agent Systems* 33.6, pp. 750–797.

Hofbauer, Josef and William H Sandholm (2002). "On the global convergence of stochastic fictitious play". In: *Econometrica* 70.6, pp. 2265–2294.

Khetarpal, Khimya et al. (2022). "Towards continual reinforcement learning: A review and perspectives". In: *Journal of Artificial Intelligence Research* 75, pp. 1401–1476.

Newey, Whitney K and Daniel McFadden (1994). "Large sample estimation and hypothesis testing". In: *Handbook of econometrics* 4, pp. 2111–2245.

Perkins, Steven and David S Leslie (2013). "Asynchronous stochastic approximation with differential inclusions". In: *Stochastic Systems* 2.2, pp. 409–446.

Perolat, Julien, Bilal Piot, and Olivier Pietquin (2018). "Actor-critic fictitious play in simultaneous move multistage games". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 919–928.

Possnig, Clemens (2022). "Reinforcement Learning and Collusion". URL: https://cjmpossnig.github.io/papers/jmp_CPossnig.pdf.

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

Ramaswamy, Arunselvan and Eyke Hullermeier (2021). "Deep Q-Learning: Theoretical Insights from an Asymptotic Analysis". In: *IEEE Transactions on Artificial Intelligence*.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Watkins, Christopher John Cornish Hellaby (1989). "Learning from delayed rewards". In:

Zhang, Kaiqing, Zhuoran Yang, and Tamer Başar (2021). "Multi-agent reinforcement learning: A selective overview of theories and algorithms". In: *Handbook of Reinforcement Learning and Control*, pp. 321–384.

Zhou, Huozhi et al. (n.d.). "Nonstationary Reinforcement Learning with Linear Function Approximation". In: *Transactions on Machine Learning Research* ().

Appendix A. Appendix

All proofs are given here.

A.1. Proof of Lemma 1

We can write

$$
\begin{aligned}
&\|L_t^*(\theta,\rho_t) - \sum_{s\in S}\mu_{\rho_t}(s)\phi(\rho_t,s,\theta)\| \\
&\qquad\le \sum_{s\in S}\left\|\Lambda_t(s,\rho_t,s_{\underline{W}_t}) - \mu_{\rho_t}(s)\right\|\left\|\frac{\frac{1}{K_t}\sum_{k\in W_t}\mathbb{E}\big[\ell(Z_k,\theta)_s \mid \rho_t\big]}{\Lambda_t(s,\rho_t,s_{\underline{W}_t})}\right\| \\
&\qquad+ \sum_{s\in S}\mu_{\rho_t}(s)\left\|\frac{\frac{1}{K_t}\sum_{k\in W_t}\mathbb{E}\big[\ell(Z_k,\theta)_s \mid \rho_t\big]}{\Lambda_t(s,\rho_t,s_{\underline{W}_t})} - \phi(\rho_t,s,\theta)\right\| \\
&\qquad\le \sum_{s\in S}R_{1,s,\rho_t,t}\left\|\frac{\frac{1}{K_t}\sum_{k\in W_t}\mathbb{E}\big[\ell(Z_k,\theta)_s \mid \rho_t\big]}{\Lambda_t(s,\rho_t,s_{\underline{W}_t})}\right\| + \max_{s\in S}R_{2,s,\rho_t,t},
\end{aligned}
$$

where

$$
\begin{aligned}
R_{1,s,\rho_t,t} &= \left\|\Lambda_t(s,\rho_t,s_{\underline{W}_t}) - \mu_{\rho_t}(s)\right\|, \\
R_{2,s,\rho_t,t} &= \left\|\frac{\frac{1}{K_t}\sum_{k\in W_t}\mathbb{E}\big[\ell(Z_k,\theta)_s \mid \rho_t\big]}{\Lambda_t(s,\rho_t,s_{\underline{W}_t})} - \phi(\rho_t,s,\theta)\right\|.
\end{aligned}
$$

First note that for all fixed $\rho\in\Gamma$,

$$|\Lambda_t(s,\rho,s_{\underline{W}_t}) - \mu_\rho(s)| \to 0$$

as $t\to\infty$, and independently from initial state $s_{\underline{W}_t}$, which follows from irreducibility (c.f. Freedman (2017), Theorem 4.9). Now to prove uniform convergence in ρ. Firstly, $\rho_t\in\Gamma$ compact. So for any fixed $\delta>0$, we can find an open cover of Γ of cardinality J_δ, using δ-balls centered at θ_j with $1\le j\le J_\delta$. Now write for all $s\in S$

$$H_t(\rho,s) = \frac{1}{K_t}\sum_{k\in W_t}\lambda_k(s,\rho,s_{\underline{W}_t}) - \mu_\rho(s).$$

Then

$$
\begin{aligned}
\sup_{\rho\in\Gamma}|H_t(\rho,s)| &\le \max_{1\le j\le J_\delta}\sup_{\rho\in B(\rho_j,\delta)}\big\{|H_t(\rho,s)-H_t(\rho_j,s)| + |H_t(\rho_j,s)|\big\} \\
&\le \sup_{\rho\in\Gamma}\sup_{\rho_1\in B(\rho,\delta)}|H_t(\rho,s)-H_t(\rho_1,s)| + \max_{1\le j\le J_\delta}|H_t(\rho_j,s)| \\
&= A_t + B_t,
\end{aligned}
$$

where

$$A_t = \sup_{\rho \in \Gamma} \sup_{\rho_1 \in B(\rho,\delta)} |H_t(\rho, s) - H_t(\rho_1, s)|,$$
$$B_t = \max_{1 \le j \le J_\delta} |H_t(\rho_j, s)|.$$

Pointwise convergence of $H_t(\rho, s)$ implies that $B_t \to 0$ as $t \to \infty$. Then,

$$\begin{aligned}
A_t &\le \sup_{\rho \in \Gamma} \sup_{\rho_1 \in B(\rho,\delta)} \frac{1}{K_t} \sum_{k \in W_t} |\lambda_k(s, \rho, s_{\underline{W}_t}) - \lambda_k(s, \rho_1, s_{\underline{W}_t})| + \sup_{\rho \in \Gamma} \sup_{\rho_1 \in B(\rho,\delta)} |\mu_\rho(s) - \mu_{\rho_1}(s)| \\
&\le \sup_{\rho \in \Gamma} \sup_{\rho_1 \in B(\rho,\delta)} D_1 \frac{1}{K_t} \sum_{k \in W_t} \|\rho - \rho_1\| + \sup_{\rho \in \Gamma} \sup_{\rho_1 \in B(\rho,\delta)} D_2 \|\rho - \rho_1\| \\
&\le \delta(D_1 + D_2),
\end{aligned}$$

where $0 < D_1, D_2 < \infty$ are the Lipschitz constants existing by Assumption 4. Thus for all $t \ge 1$, $A_t \to 0$ as $\delta \to 0$ (and recall that δ is picked arbitrarily), and the result follows:

$$\sup_{\rho \in \Gamma} |R_{1,s,\rho,t}| \to 0,$$

as $t \to \infty$.

Next, note that

$$\phi(\rho_t, s, \theta) = \frac{\mathbb{E}[\ell(Z_k, \theta)_s | \rho_t]}{P(s_k = s, \rho_t)},$$

with $P(s_k = s, \rho_t) = \sum_{s' \in S} \mu_{\rho_t}(s') \lambda_k(s, \rho_t, s')$ is the stationary expected value of λ_k over initial states s'. By Assumption 1, for all fixed ρ, $\lim_{k \to \infty} \lambda_k(s, \rho, s') = \lim_{k \to \infty} P(s_k = s, \rho) = \mu_\rho(s)$. Then we get

$$\begin{aligned}
R_{2,s,\rho_t,t} &\le \left\| \mathbb{E}\left[\ell(Z_t, \theta)_s \mid \rho_t\right] \right\| \left\| \frac{1}{\Lambda_t(s, \rho_t, s_{\underline{W}_t})} - \frac{1}{P(s_k = s, \rho_t)} \right\| && (6) \\
&\le D_3 D_4 \|\Lambda_t(s, \rho_t, s_{\underline{W}_t}) - P(s_k = s, \rho_t)\|, && (7)
\end{aligned}$$

where $0 < D_3 < \infty$ is an upper bound on $\|\phi(\rho_t, s, \theta)\|$ following from the boundedness of the loss function, and $0 < D_4 < \infty$ is an upper bound on $\frac{1}{\Lambda_t(s,\rho_t,s_{\underline{W}_t})} \frac{1}{P(s_k = s, \rho_t)}$ which follows from irreducibility and Assumption 4, which implies that both fractions cannot diverge. Finally, the last term in (6) converges to zero uniformly over ρ by an argument analogous to the convergence of $H_t(\rho, s)$.

Finally, since

$$\left\| \frac{\frac{1}{K_t}\sum_{k\in W_t}\mathbb{E}\big[\ell(Z_k,\theta)_s \mid \rho_t\big]}{\Lambda_t(s,\rho_t,s_{\underline{W}_t})} \right\| \leq D_3 D_4,$$

where the last bound is independent of θ, convergence of $\|L_t^*(\theta,\rho_t)-\sum_{s\in S}\mu_{\rho_t}(s)\phi(\rho_t,s,\theta)\|$ is uniform over $\theta\in\Theta$ and $\rho\in\Gamma$.
■

A.2. Proof of Theorem 1

The following Lemma will help prove the result. From now on, we drop the i-superscript whenever possible.

Lemma 3. *Impose Assumptions 1, 2, and 4.*
For all $\varepsilon>0$,

$$\mathbf{P}\Big(\sup_{\theta\in\Theta}\|L_t(\theta)-L_t^*(\theta,\boldsymbol{\rho}_t))\|>\varepsilon\Big)\to 0,$$

as $t\to\infty$.

Proof. We first show pointwise convergence of $\|L_t(\theta)-L_t^*(\theta,\boldsymbol{\rho}_t)\|$.
We can write

$$L_t(\theta)=\sum_{s\in S}\frac{n_t(s)}{K_t}\frac{\sum_{k\in W_t}\ell(Z_k,\theta)\mathbf{1}\{s_k=s\}}{n_t(s)},$$

where

$$n_t(s)=\sum_{k\in W_t}\mathbf{1}\{s_k=s\}.$$

First we show, for all $s\in S$

$$\Big|\frac{n_t(s)}{K_t}-\Lambda_t(s,\boldsymbol{\rho}_t,s_{\underline{W}_t})\Big|\to_P 0,$$

as $t\to\infty$. For this, define

$$V_{1,t}=\mathbb{E}\Big[\Big(\frac{n_t(s)}{K_t}-\Lambda_t(s,\boldsymbol{\rho}_t,s_{\underline{W}_t})\Big)^2\Big],$$

and let $d_t=\mathbf{1}\{s_k=s\}-\lambda_k(s,\boldsymbol{\rho}_{\underline{W}_t:k},s_{\underline{W}_t})$.

$$V_{1,t}=\frac{1}{K_t^2}\sum_{k\in W_t}\mathbb{E}\Big[d_k^2\Big]+\frac{1}{K_t^2}\sum_{k,k'\in W_t|k\neq k'}\mathbb{E}\Big[d_kd_{k'}\Big],$$

where the second term

$$\frac{1}{K_t^2}\sum_{k,k'\in W_t|k\neq k'}\mathbb{E}\Big[d_k d_{k'}\Big]=\frac{1}{K_t^2}\mathbb{E}\sum_{k,k'\in W_t|k\neq k'}\mathbb{E}\Big[d_k d_{k'}\mid \boldsymbol{Z}_{k\vee k'},\boldsymbol{\rho}_{\underline{W}_t:(k\wedge k')},s_{k\wedge k'}\Big]=0,$$

since by by definition of d_t and Assumption 1, states form a controlled markov chain and thus

$$\begin{aligned}\mathbb{E}\Big[\mathbb{E}\Big[d_k d_{k'}\mid \boldsymbol{Z}_{k\vee k'},\boldsymbol{\rho}_{\underline{W}_t:(k\wedge k')},s_{k\wedge k'}\Big]\Big]&=\mathbb{E}\Big[d_{k\vee k'}\mathbb{E}\Big[d_{k\wedge k'}\mid \boldsymbol{Z}_{k\vee k'},\boldsymbol{\rho}_{\underline{W}_t:(k\wedge k')},s_{k\wedge k'}\Big]\Big]\\&=\mathbb{E}\Big[d_{k\vee k'}\mathbb{E}\Big[d_{k\wedge k'}\mid \rho_{k\wedge k'},s_{k\wedge k'}\Big]\Big]=0.\end{aligned}$$

It follows that $V_{1,t}\to 0$ as $K_t\to\infty$. We can then apply Chebyshev's inequality and the first result follows:

$$\Big|\frac{n_t(s)}{K_t}-\Lambda_t(s,\boldsymbol{\rho}_t,s_{\underline{W}_t})\Big|=o_P(1).$$

Similarly, let $h_t(\theta,s)=\ell(Z_t,\theta)\mathbf{1}\{s_t=s\}-\mathbb{E}[\ell(Z_t,\theta)_s|\boldsymbol{\rho}_t]$. Define

$$V_{2,t,s}=\mathbb{E}\Big[\big(\frac{1}{K_t}\sum_{k\in W_t}h_k(\theta,s)\big)^2\Big],$$

then by an argument analogous to above, using Assumption 2 and boundedness of l we can conclude that $V_{2,t,s}\to 0$ as $t\to\infty$ for all s. By Assumption 4 we have that $\frac{n_t(s)}{K_t}>0$ with probability approaching 1 with t. Thus we can apply the continuous mapping theorem to arrive at the result: for all $\theta\in\Theta$,

$$\|L_t(\theta)-L_t^*(\theta,\boldsymbol{\rho}_t)\|=o_P(1).$$

The rest of the proof is based on Newey and McFadden (1994)'s proof of their Theorem 2.1, but we have to adapt to the fact that we face a random population objective L^* due to the randomness of $\boldsymbol{\rho}_t$.

The proof follows a similar logic as the proof of Lemma 1. First define

$$H_t(\theta,\boldsymbol{\rho}_t)=\frac{1}{K_t}\sum_{k\in W_t}h_k(\theta),$$

where we drop the dependence on state s since the statement holds for any s and there are finitely many.

Take any $\varepsilon>0$ and any $\delta>0$. Let $B(x,\delta)$ denote the δ ball centered at x. Then by compactness of Θ, we can construct a finite open cover of Θ with cardinality $J_\delta<\infty$ using open balls $B(\theta_j, \delta)$. Now note that

$$
\begin{aligned}
&\mathbf{P}\Big(\sup_{\theta\in\Theta} \|H_t(\theta, \boldsymbol{\rho}_t)\| > 2\varepsilon\Big) \\
&\qquad\qquad\qquad \le \mathbf{P}\Big(\max_{1\le j\le J_\delta} \sup_{\theta\in B(\theta_j,\delta)} \big\{\|H_t(\theta, \boldsymbol{\rho}_t) - H_t(\theta_j, \boldsymbol{\rho}_t)\| + \|H_t(\theta_j, \boldsymbol{\rho}_t)\|\big\} > 2\varepsilon\Big) \\
&\qquad\qquad\qquad \le \mathbf{P}\Big(\sup_{\theta\in\Theta} \sup_{\theta_1\in B(\theta,\delta)} \|H_t(\theta, \boldsymbol{\rho}_t) - H_t(\theta_1, \boldsymbol{\rho}_t)\| + \max_{1\le j\le J_\delta} \|H_t(\theta_j, \boldsymbol{\rho}_t)\| > 2\varepsilon\Big) \\
&\qquad\qquad\qquad \le A_t + B_t,
\end{aligned}
$$

where

$$
\begin{aligned}
A_t &= \mathbf{P}\Big(\sup_{\theta\in\Theta} \sup_{\theta_1\in B(\theta,\delta)} \|H_t(\theta, \boldsymbol{\rho}_t) - H_t(\theta_1, \boldsymbol{\rho}_t)\| > \varepsilon\Big), \\
B_t &= \mathbf{P}\Big(\max_{1\le j\le J_\delta} \|H_t(\theta_j, \boldsymbol{\rho}_t)\| > \varepsilon\Big).
\end{aligned}
$$

The second term must converge to zero by pointwise convergence as proved before, since

$$
B_t \le \sum_{1\le j\le J_\delta} \mathbf{P}\Big(\|H_t(\theta_j, \boldsymbol{\rho}_t)\| > \varepsilon\Big) \to 0
$$

as $t \to \infty$. Now define

$$
Y_\delta = \sup_{\theta\in\Theta} \sup_{\theta_1\in B(\theta,\delta)} \frac{1}{K_t} \sum_{k\in W_t} \|\ell(Z_k, \theta) - \ell(Z_k, \theta_1)\|,
$$

and

$$
\tilde{Y}_\delta = \sup_{\theta\in\Theta} \sup_{\theta_1\in B(\theta,\delta)} \frac{1}{K_t} \sum_{k\in W_t} \|\mathbb{E}\big[\big(\ell(Z_k, \theta) - \ell(Z_k, \theta_1)\big) \mid \boldsymbol{\rho}_k\big]\|.
$$

Then note that

$$
A_t \le \mathbf{P}\big(Y_\delta + \tilde{Y}_\delta > \varepsilon\big) \le \frac{1}{\varepsilon}\mathbb{E}\Big[Y_\delta + \tilde{Y}_\delta\Big], \tag{8}
$$

by Markov's inequality. Finally, note that

$$
\begin{aligned}
\mathbb{E}Y_\delta &\le \frac{1}{K_t} \sum_{k\in W_t} \mathbb{E} \sup_{\theta\in\Theta} \sup_{\theta_1\in B(\theta,\delta)} \|\ell(Z_k, \theta) - \ell(Z_k, \theta_1)\| \\
&\le \frac{1}{K_t} \sum_{k\in W_t} \mathbb{E} \sup_{\theta\in\Theta} \sup_{\theta_1\in B(\theta,\delta)} C_1(Z_k)\|\theta - \theta_1\| \le \frac{1}{K_t} \sum_{k\in W_t} \mathbb{E}C_1(Z_k)\delta,
\end{aligned}
$$

where the second to last inequality follows from Assumption 2 and the Lipschitz property of $\ell(Z, \theta)$. Thus, we get

$$
\lim_{t\to\infty} \mathbb{E}Y_\delta \le \lim_{t\to\infty} \frac{1}{K_t} \sum_{k\in W_t} \mathbb{E}C_1(Z_k)\delta,
$$

where the right hand side vanishes as $\delta \to 0$ by Assumption 2. We can make an analogous argument to show that $\lim_{t\to\infty} \mathbb{E}\tilde{Y}_\delta \to 0$ as $\delta \to 0$. It follows that $A_t \to 0$ as $t \to \infty$ and $\delta \to 0$ by the bound given in (8). The result follows, since $H_t(\theta)$ is the only factor in $L_t(\theta) - L_t^*(\theta, \boldsymbol{\rho}_t)$ that depends on θ:

We can write

$$\begin{aligned}
&\|L_t(\theta) - L_t^*(\theta, \boldsymbol{\rho}_t)\| \\
&\qquad \le \sum_{s\in S} \Big\| \frac{n_t(s)}{K_t} - \Lambda_t(s, \boldsymbol{\rho}_t, s_{\underline{W}_t}) \Big\| \Big\| \frac{K_t}{n_t(s)} \frac{1}{K_t} \sum_{k\in W_t} \ell(Z_k, \theta)_s \Big\| \\
&\qquad + \max_{s\in S} \Big\| \frac{K_t}{n_t(s)} - \frac{1}{\Lambda_t(s, \boldsymbol{\rho}_t, s_{\underline{W}_t})} \Big\| \Big\| \frac{1}{K_t} \sum_{k\in W_t} \ell(Z_k, \theta)_s \Big\| \\
&\qquad + \max_{s\in S} \Big\| \frac{1}{\Lambda_t(s, \boldsymbol{\rho}_t, s_{\underline{W}_t})} \frac{1}{K_t} \sum_{k\in W_t} h_k(\theta, s) \Big\|.
\end{aligned}$$

There first two terms converge uniformly in θ to zero by our first arguments in this proof, due to the boundedness assumption on l and Assumption 4. Only the last term depends on $h_t(\theta, s)$, the uniform convergence of which has been shown above.

$\square$

Lemma 4. *Impose Assumptions 1 - 4. Then for all $\varepsilon > 0$*

$$\mathbf{P}\Big(\sup_{\theta\in\Theta} \big\| L_t^*(\theta, \boldsymbol{\rho}_t) - L_t^*(\theta, \rho_t) \big\| > \varepsilon \Big) \to 0,$$

as $t \to \infty$.

Proof. For any $\theta \in \Theta$ we can write

$$\begin{aligned}
Y_t(\theta, \boldsymbol{\rho}_t) \equiv \|L_t^*(\theta, \boldsymbol{\rho}_t) - L_t^*(\theta, \rho_t)\| &\le \frac{1}{K_t} \sum_{k\in W_t} \|\mathbb{E}\big[\ell(Z_k, \theta) \mid \rho_k\big] - \mathbb{E}\big[\ell(Z_k, \theta) \mid \rho_t\big]\| \\
&\le C_4 \frac{1}{K_t} \sum_{k\in W_t} \|\rho_k - \rho_t\| \le C_4 C_5 \frac{1}{K_t} \sum_{k\in W_t} \sum_{l=k}^{t} \alpha_l,
\end{aligned}$$

with $0 < C_4 < \infty$ being the bound on C_3 given by Assumption 2 and $0 < C_5 < \infty$ being the bound resulting from $F(\rho_t, \theta_t) + M_{t+1}$ being almost surely bounded given $\mathcal{G}_t$ for all t. Since α_t is decreasing, we have

$$\frac{1}{K_t} \sum_{k\in D_t} \sum_{l=k}^{t} \alpha_l \le K_t \alpha_{t-K_t},$$

and the last term vanishes by Assumption 3. Since the last term is independent of θ, we can use Markov's inequality with

$$\mathbb{E}\big[\sup_{\theta\in\Theta} Y_t(\theta, \boldsymbol{\rho}_t)\big] \leq C_4 C_5 K_t \alpha_{t-K_t},$$

and the conclusion follows. $\square$

Using Lemmas 1, 3, and 4, we conclude that, for all $\varepsilon > 0$,

$$\mathbf{P}\Big(\sup_{\theta\in\Theta} \big\|L_t(\theta) - L^*_\infty(\theta, \rho_t))\big\| > \varepsilon\Big) \to 0,$$

as $t \to \infty$.

As a last step we can prove the convergence of θ_t. By Assumption 5,

$$\begin{aligned}
\mathbf{P}\big(\theta_t \notin B(\theta^*(\rho_t), \varepsilon)\big) &\leq \mathbf{P}\big(L^*_\infty(\theta_t, \boldsymbol{\rho}_t) - L^*_\infty(\theta^*(\rho_t), \rho_t) \geq \delta\big) \\
&= \mathbf{P}\Big(L^*_\infty(\theta_t, \rho_t) - L_t(\theta_t) + L_t(\theta_t) - L^*_\infty(\theta^*(\rho_t), \rho_t) \geq \delta\Big) \\
&\leq \mathbf{P}\Big(L^*_\infty(\theta_t, \rho_t) - L_t(\theta_t) + L_t(\theta^*(\rho_t)) - L^*_\infty(\theta^*(\rho_t), \rho_t) \geq \delta\Big) \\
&\leq \mathbf{P}\Big(2\sup_{\theta\in\Theta} \|L_t(\theta) - L^*_\infty(\theta, \rho_t)\| \geq \delta\Big),
\end{aligned}$$

where the second-to-last inequality follows from Assumption 5. The result follows. It follows that we can write $F(\rho_t, \theta_t) = F(\rho_t, \theta^*(\rho_t)) + o_P(1)$ as a function approximator that depends on policy profiles only through the *current period's* profile ρ_t, and not some weighted average of past profiles. ■

A.3. Proof of Lemma 2

Firstly, one can prove $\chi_t \to_P 0$ as $t \to \infty$ given that $\theta_t \to_P \theta^*(\rho_t)$, using arguments analogous to the proof of Lemma 1, given point (ii). This can be done with the following argument: From point (i), we get that one can bound χ_t by a function linear in $\|\theta_t - \theta^*(\rho_t)\|$. Convergence in probability of χ_t follows, with also the rate of convergence of χ_t being bounded by the convergence rate of θ_t.

Assumptions (iii)-(v) are classical assumptions used in the asymptotic analysis of extremum estimators, usually to determine asymptotic normality. In this case, these assumptions give us that $\theta_t = O_P(n^{-\frac{1}{2}})$, via the standard taylor approximation argument applied to $L_t(\theta_t)$.

Now, square-integrability of χ_t follows from boundedness of Q, Q^*, which in turn follows from boundedness of u. Square-integrability gives us that convergence in probability of χ_t implies convergence in the mean also. Finally, the rate of convergence of $\mathbb{E}[\chi_t]$ can be approximated by the rate of convergence of θ_t: As mentioned before,

$$\chi_t \leq C\|\theta_t - \theta^*(\rho_t)\|,$$

so that, using $\theta_t - \theta^*(\rho_t) = O_P(n^{-\frac{1}{2}})$, we have that for all $\varepsilon > 0$ there exists an $0 < M_\varepsilon < \infty$ such that

$$\mathbf{P}\Big(\chi_t \geq M_\varepsilon n^{-\frac{1}{2}}\Big) \leq \varepsilon.$$

We can thus write,

$$\big|\mathbb{E}\big[\chi_t\big]\big| \leq M_\varepsilon n^{-\frac{1}{2}} + \varepsilon C_2,$$

where $C_2 = \mathbb{E}\big[\chi_t^2\big]$ since we used the Cauchy-Schwartz inequality on the second term. This bound is sufficient for our purposes, as ε can be chosen to be a decreasing function of t, at a arbitrarily slow rate. This can be seen by looking more closely at the exact use of Assumption 3 (iii) in Possnig (2022): This point is only required in the bounding exercise applied in Proposition 7 of Possnig (2022). Importantly, this bounding exercise is used to make Proposition 1.6 in Hofbauer and Sandholm (2002) applicable (see the last line of the proof). First introduce notation used in Possnig (2022):

$$\tau_0 = 0; \quad \tau_n = \sum_{k=1}^{n} \alpha_k; \quad m(t) = \sup\{k \geq 0 : t \geq \tau_k\}.$$

Then one can write, letting $\mu_k = \mathbb{E}\chi_k$, for any $\varepsilon > 0$,

$$\begin{aligned}\left|\sum_{k=n}^{m(\tau_n+T)} \alpha_{k+1}\mu_{k+1}\right| &\leq M_\varepsilon \sum_{k=n}^{m(\tau_n+T)} \alpha_{k+1}k^{-\frac{1}{2}} + \varepsilon \sum_{k=n}^{m(\tau_n+T)} \alpha_{k+1} \\ &= M_\varepsilon \sum_{k=n}^{m(\tau_n+T)} \alpha_{k+1}k^{-\frac{1}{2}} + \varepsilon T,\end{aligned}$$

where the equality is due to the definition of τ_n. To conclude Proposition 1.6 in Hofbauer and Sandholm (2002), it is necessary to take $n \to \infty$. As mentioned above, if one now takes ε_n as a function that decreases in n to zero, one can find a rate slow enough so that M_{ε_n} will diverge slowly enough for the whole term to vanish. Such a rate can always be found since ε_n's rate is arbitrary.