Agentic Subjective Q-Learning Equilibrium

Anonymous authors
Paper under double-blind review

Abstract

In many applications, agents/decision makers take part in systems with very complex dynamics and they respond by inevitably making incorrect modeling assumptions. In this context, we define the concept of Agentic Subjective Q-Learning Equilibrium as an equilibrium concept where each agent uses local/partial information in their learning algorithm, as if the partial information constitutes an approximate Markov model. A distinguishing feature of such a setup is that the exploration policy used for learning impacts the perceived model and there is a dual dependence of the induced cost on the agent policies: By noting an equivalence with empirical model learning, it follows that an exploration policy generates the sample path which induces a model (which depends on the exploration policy), and the model is used to obtain an optimal policy (for the learned model) either via reinforcement learning or empirical learning. This then leads to the question on existence of a fixed point equation involving learning and exploration. An agentic subjective learning equilibrium policy is thus defined as a policy which is self-confirming: the model induced by the policy has the policy as an optimal solution. We establish an existence result on equilibria critically building on continuous dependence of invariant measures on policies under a suitable control topology. We then present an associated learning/convergence theorem to ϵ-equilibria via policy revision dynamics. We show implications for symmetric dynamic games (including mean-field games), weakly acyclic games (including potential games), and generalized weakly acyclic games.

1 Subjective Equilibrium in Complex Environments

In many stochastic decision, game, and control problems, one does not know the true dynamics or the cost structure, and may wish to use past data to obtain an asymptotically near optimal solution (that is, via *learning* from past data). In many problems involving engineering and applied sciences, as well as social sciences, economics, and health sciences, one may not know whether the problem studied can be formulated as a Markov Decision Process (MDP), or a partially observable Markov Decision Process (POMDP) or a multi-agent system where other agents are present or not. There are many practical settings where one indeed works with data and does not know the possibly very complex structure under which the data is generated and tries to respond to the environment.

This phenomenon arises also when autonomous agents are involved in the decision making process, mapping data to decisions and outcomes via learning algorithms.

A common practical and reasonable response is to view the system as if it is a Markovian system, with a perceived state and action (which may or may not define a genuine controlled Markov chain and therefore, the MDP assumption may not hold in actuality), and arrive at corresponding solutions via some learning algorithm. Such a setup has been studied in Singh et al. (1994), Szepesvàri & Smart (2004); Melo et al. (2008), and including the recent studies Kara & Yüksel (2024) Chandak et al. (2024) Dong et al. (2022) Kara & Yüksel (2023).

In this paper, we focus on consistency and equilibrium properties arising from such learning algorithms, and in addition we consider a multi-agent setup involving such a complex system where each agent only has local and partial information and a limited model information.

Consider N Agents or Decision Makers, denoted with DM i, $i \in \mathcal{N} := \{1, \cdots, N\}$, with corresponding policies $\gamma^i \in \Gamma^i, i \in \mathcal{N}$, and (possibly subjective) evaluations of these policies leading to cost functions, denoted with $J^i : \prod_{i \in \mathcal{N}} \Gamma^i \to \mathbb{R}$. A policy profile $\{\gamma^{*,i}, i \in \mathcal{N}\}$ would then be an equilibrium if

$$\min_{\gamma^i \in \Gamma^i} J^i(\gamma^i, \gamma^{*,-i}) = J^i(\gamma^{*,i}, \gamma^{*,-i})$$

In our paper, each agent will aim to minimize a discounted cost criterion. However, the agents apply the criterion in an agentic sense only implicitly via running a learning algorithm; an equivalent interpretation is that the agents can also explicitly (via empirical learning and modeling) under a learned model, with its corresponding cost and transition kernel, optimize a discounted cost criterion. Further discussion on these interpretations will be provided later in the paper.

Accordingly, our analysis can thus be viewed in the context of subjective equilibria and the various forms in which such a solution concept may appear. Different from the Bayesian subjective equilibria of Kalai & Lehrer (1995), where subjective priors of agents are updated individually with common information history and conditions for convergence are established under restrictive absolute continuity conditions due to the best-response constraint, in this paper we consider subjectivity due to incorrect modeling assumptions. Such an equilibrium may arise in cases of imperfect modeling information, or as a result of algorithmic learning (as in reinforcement learning).

Relevant references on such a subjective equilibrium concept include Fudenberg & Levine (1993); Adlakha et al. (2015); Weintraub et al. (2011); Wiszniewska-Matyszkiel (2017); Dudebout & Shamma (2012; 2014); Kalai & Lehrer (1995); Wiszniewska-Matyszkiel (2014). Fudenberg & Levine (1993) presents an equilibrium concept, self-confirming equilibrium, where the otherwise possibly subjective beliefs are correct given the realized data, where the data, as in Kalai & Lehrer (1995), is commonly shared with perfect recall and the beliefs are on the opponent strategies. The notion of equilibrium introduced in Adlakha et al. (2015), called stationary equilibrium, is based on the assumption that the players view the population states (which play the role of the environment variables in this paper) as deterministic and constant. The notion of oblivious equilibrium utilized in Weintraub et al. (2011) is analogous to the stationary equilibrium in Adlakha et al. (2015). The notion of belief distorted Nash equilibrium introduced in Wiszniewska-Matyszkiel (2017) for a deterministic system requires players' (probabilistic) beliefs to concentrate on the actual deterministic play path. The notion of empirical evidence equilibrium introduced in Dudebout & Shamma (2012; 2014) for stochastic games (with finite state and control spaces) is also conceptually similar as it requires a consistency between the objective and the subjective probabilities of the signal variable (environment variable in this paper). There is also a separate body of work in the literature on dynamic games with a continuum of players, which includes private as well as global state variables; see Wiszniewska-Matyszkiel (2014) and the references therein.

To further motivate our approach, let us note that many recent theoretical contributions in multi-agent learning have focused on structured classes of stochastic games, such as zero-sum games (Sayin et al., 2021; 2022) and n-player stochastic teams and their generalizations (Ding et al., 2022; Fox et al., 2022; Leonardos et al., 2022; Mguni et al., 2021; Unlu & Sayin, 2023; Zhang et al., 2022). Furthermore, in some of the literature on multi-agent learning, various information sharing structures are assumed, such as full state observability (as assumed by (Daskalakis et al., 2020) among others) or action-sharing among all agents (as assumed by (Littman & Szepesvári, 1996; Hu & Wellman, 2003) and (Littman, 2001)), which are appropriate in some settings but are unrealistic in many large-scale, decentralized systems. One issue with designing learning algorithms that use global information about the local states and actions of all players is that such algorithms do not scale with the number of players (Wang et al., 2020). However, in large decentralized systems, agents have limited information about the overall system; and policies, actions, and local states of other agents may be unobservable or difficult to estimate using local data. Independent learners (Matignon et al., 2009; 2012) Yongacoglu et al. (2024) are a class of multi-agent learning algorithms that are characterized by intentional obliviousness to the strategic environment: in the independent learning paradigm, agents use only local information and run single-agent learning algorithms, treating their environment as though it were a fully observed (single-agent) MDP. Such intentional obliviousness is used to mitigate the computational burden at each agent. Each independent learning agent makes a simplifying assumption about its environ-

ment to obtain tractable, scalable algorithms in complex multi-agent settings. This approach motivates the equilibrium concept we study and our analysis in the paper.

Contributions.

- We define the concept of Agentic Subjective Q-Learning Equilibrium, first introduced in (Yongacoglu et al., 2024, Definition 16) under a more restrictive context, as an equilibrium concept for reinforcement learning in complex environments where each agent uses partial information in their reinforcement learning algorithm, as if the partial information constitutes a Markov model (which in actuality is not). This is presented in Definition 3.1. We also present an equivalent model learning interpretation, where the reinforcement learning converges to a solution which would have been a solution to a Markov Decision Problem with empirically learned transition kernels and cost functions if the agents were able to interpret what their algorithm is computing as an explicit optimization problem.
- In our model, through its induced invariant probability measure, the exploration policy affects the learned solution (and, equivalently, learned model) due to the general state space considered and the resulting non-Markovian approximate state process. Accordingly, a distinguishing aspect of the proposed equilibrium concept is that learning and best-responding are consistent under equilibrium.
- In Theorem 4.1 we present an existence result for Agentic Subjective Q-Learning Equilibrium. This addresses an open question noted in (Yongacoglu et al., 2024, Section J.2) and (Kara & Yüksel, 2024, Section 3.5).
- In Theorem 5.1 we present an associated convergence theorem to equilibria. In the standard Borel setup, we show that a revision policy is guaranteed to converge to an ϵ-Agentic Subjective Q-Learning Equilibrium with arbitrarily high probability. We show that, while the existence result obtained is very general, convergence to equilibria requires some further structure, notably the existence of satisficing paths under a policy revision dynamics: We discuss several examples which satisfy both existence and convergence; these include symmetric games such as mean-field games (with arbitrary local and partial information as long as symmetry is preserved), single agent stochastic control, team games, potential games, weakly acyclic games, and generalized weakly acyclic games.

2 Towards Agentic Equilibria: Convergence of Single Agent Q-Learning under a Non-Markovian Asymptotically Ergodic Environment and Equivalence with Empirical Model Learning

Consider the following general model with N agents,

$$X_{t+1} = F(X_t, U_t^1, U_t^2, \cdots, U_t^N, W_t), \tag{1}$$

where W_t is an i.i.d. noise process, X_t is the $\mathbb{X}$-valued state process and U_t^m is the $\mathbb{U}^m$-valued action of Agent m, $m = 1, 2, \cdots, N$ at time t. We assume that F is Borel measurable where $\mathbb{X}$ is a Borel subset of a Polish space with metric d, and therefore the hidden state X_t, can represent a very general model. We will use the notation $\mathbf{U}_t := U_t^1, U_t^2, \ldots, U_t^N$. We assume that the action sets $\mathbb{U}^i$ are finite.

This hidden state, however, is not known to the agents, nor is the possibly very complex dynamical description defined in the above by F and the noise W_t. The agents do not have knowledge even on the presence of other agents. Each agent has access to some local information, but each agent perceives some state to be Markovian or nearly Markovian under their perceived model. In particular, the environment involving their perceived state is non-Markovian; but agents perceive it to be Markovian with a pseudo-state.

In particular, let

$$Y_{t+1}^i = \rho^i(X_t),$$

where $\rho^i : \mathbb{X} \to \mathbb{Y}^i$ with $\mathbb{Y}^i$ finite and ρ^i Borel measurable.

As noted, each Agent i agent perceives Y_t^i to be Markovian (due to modeling inaccuracies, bias, or computational limitations) and applies Q-learning with Y_t^i as the state and applies: $U_t^i = \gamma_t^i(Y_t^i)$.

Observe that the hidden state can be quite general and abstract and therefore the model above is very general and captures many real-world applications, both in technical and engineering systems, as well as with regard to applications in social sciences, economics and psychology.

In the following, we let the perceived state structure be so that it corresponds to a quantization of the state variable X_t: The state space $\mathbb{X}$ is quantized such that for disjoint $\{B_k^i\}_{k=1}^{M_i}$ with $\cup_{k=1}^{M_i} B_k^i = \mathbb{X}$, we define a finite set $\mathbb{S}^i = \{y_1^i, \ldots, y_{M_i}^i\}$ and write

$$y_k^i = \rho^i(x), \text{ if } x \in B_k^i, \quad k = 1, \cdots, M_i.$$

where ρ^i is the quantization map for Agent i.

To gain some critical insight and motivation, we review the case with a single agent in the following: In the single-agent setting, the above setup leads to an instance where the non-Markovian dynamics is perceived as a Markovian one. In this case, under mild conditions, the Quantized Q-learning algorithm Kara et al. (2023) (see also (Kara & Yüksel, 2024, Theorem 2.1)) converges to a fixed point which is the solution of the standard value iteration algorithm corresponding to an approximate MDP under ergodicity conditions on the overall dynamics.

2.1 Convergence of Single Agent Q-Learning under a Non-Markovian Asymptotically Ergodic Environment

Let there be a single agent who applies an exploration policy γ and collects realizations of state, action, and stage-wise cost under this policy:

$$X_0, U_0, c(X_0, U_0), X_1, U_1, c(X_1, U_1), \ldots.$$

Let $\mathcal{T}(dx_1|x, u)$ denote the transition kernel of the model.

The agent updates his Q-functions defined only for state-action pairs in $\mathbb{Y} \times \mathbb{U}$ as follows: for $t \geq 0$, if $(X_t, U_t) = (x, u) \in \mathbb{X} \times \mathbb{U}$, then

$$\begin{aligned} Q_{t+1}(\rho(x), u) = (1 - \alpha_t(\rho(x), u))\, Q_t(\rho(x), u) \\ + \alpha_t(\rho(x), u) \left(c(x, u) + \beta \min_{v \in \mathbb{U}} Q_t(\rho(X_{t+1}), v) \right), \end{aligned} \tag{2}$$

that is, for any true value of the state, we use its representative state from the finite set $\mathbb{Y}$ when updating the Q-function.

We assume that the learning rates satisfy:

$$\alpha_t(\rho(x), u) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{\rho(X_k) = \rho(x), U_k = u\}}}$$

and that under the stationary exploration policy, each $\rho(x)$ and u is visited infinitely often.

Theorem 2.1 (Kara et al. (2023); Kara & Yüksel (2024; 2023))**.** *Let under any stationary exploration policy γ, x_t be a positive Harris recurrent Markov chain. Suppose that the exploration policy charges every action at each state and the invariant measure under this exploration policy places a positive measure on every (non-empty) open set. Then, $Q_t^i \to Q^{*,i}$ almost surely, where*

$$Q^*(\hat{x}, u) = C^*(\hat{x}, u) + \beta \sum_{\hat{x}_1} P^*(\hat{x}_1|\hat{x}, u) \min_v Q^{*,i}(\hat{x}_1, v) \tag{3}$$

where for $y, z \in \mathbb{Y}$,

$$C^*(y,u) := \int_{\mathbb{X}} c(x,u)\pi^*_\gamma(dx|x \in \rho^{-1}(y)) = \int_{\rho^{-1}(y)} c(x,u)\hat{\pi_\gamma}^*(dx)$$
$$P^*(z|y,u) := \int_{\mathbb{X}} \mathcal{T}(x_1 \in \rho^{-1}(z)|x,u)\pi^*_\gamma(dx|z \in \rho^{-1}(y)) = \int_{\rho^{-1}(y)} \mathcal{T}(\rho^{-1}(z)|x,u)\hat{\pi_\gamma}^*(dx). \tag{4}$$

such that $\hat{\pi_\gamma}^*$ *is the invariant measure on the* X_t *process conditioned on the set* $\rho^{-1}(y)$ *with* $\hat{\pi_\gamma}^*(A) = \frac{\pi^*_\gamma(A)}{\pi^*_\gamma(\rho^{-1}(y))}$ *for all* $A \subset \rho^{-1}(y)$ *and* π^*_γ *being the invariant measure under the exploration policy* γ.

2.2 Equivalence with Empirical Model Learning under Markovian Modeling Assumption

Let the exploration policy γ given in the quantized Q-learning algorithm give rise to the invariant probability measure π^*_γ. The limiting Q-function $Q^*(y,u)$ in the discussion above corresponds to the optimal Q-function of an approximate MDP defined over the quantized state space $\mathbb{Y}$. The effective cost $C^*(y,u)$ is the average cost over the bin B_y weighted by the invariant distribution π^*_γ conditioned on bin B_y:

$$C^*(y,u) = \mathbb{E}_{x\sim\pi^*_\gamma|x\in B_y}[c(x,u)] = \int_{B_y} \frac{\pi^*_\gamma(dx)}{\pi^*_\gamma(B_y)} c(x,u). \tag{5}$$

Observe that the above is, see e.g. (Kara & Yüksel, 2024, Theorem 2.1), equal to the almost sure limit of the empirical expression on the right hand side below:

$$C^*(y,u) = \lim_{N\to\infty} \frac{\sum_{k=0}^{N-1} c(X_k,U_k)1_{\{X_k\in B_y, U_k=u\}}}{\sum_{k=0}^{N-1} 1_{\{X_k\in B_y, U_k=u\}}} \tag{6}$$

Similarly, the effective transition probability $P^*(y'|y,u)$ represents the probability of transitioning from bin B_y to bin $B_{y'}$ under action u, averaged over the invariant distribution:

$$P^*(y'|y,u) = \mathbb{P}_{x\sim\pi^*_\gamma|x\in B_y}[q(X_{t+1}) = y'|X_t = x, U_t = u] = \int_{B_y} \frac{\pi^*_\gamma(dx)}{\pi^*_\gamma(B_y)} \mathcal{T}(B_{y'}|x,u). \tag{7}$$

Likewise, by (Kara & Yüksel, 2024, Theorem 2.1), the above is the almost sure empirical limit of of the right hand side below:

$$P^*(y'|y,u) = \lim_{N\to\infty} \frac{\sum_{k=0}^{N-1} 1_{\{X_{k+1}\in B_{y'}\}}1_{\{X_k\in B_y, U_k=u\}}}{\sum_{k=0}^{N-1} 1_{\{X_k\in B_y, U_k=u\}}} \tag{8}$$

In our paper, each agent thus aims to minimize a discounted cost criterion. However, in view of the discussion above, the agents apply the criterion, implicitly via running the learning algorithm above in (2), or explicitly (via incorrect empirical modeling) with corresponding cost and transition kernel estimates (6)-(8). The criterion to be minimized is then

$$J^i(\gamma^i, \gamma^{-i}) = E^i[\sum_{k=0}^{\infty} \beta^k C^*(\hat{x}_t, u^i_t)], \tag{9}$$

where E^i denotes the transition kernels incorrectly modeled by (8).

An interpretation of the above result then is that one can first obtain the approximate model given with (5-7) by forcing the data into a Markovian model for both the empirical cost estimate (6) and empirical transition kernel estimate (8), and then solve the MDP as if this empirically constructed model is the actual one, instead of running Q-learning whose limit is then optimal precisely for this learned/empirically constructed model.

- Suppose that under any stationary policy the induced Markov chain is positive Harris recurrent and the invariant measure places a positive value to each bin: this ensures that by the ergodic theorem for Harris recurrent Markov chains (Meyn & Tweedie, 1993, Theorem 17.1.7) or (Hernández-Lerma & Lasserre, 2003, Theorem 4.2.13), each bin is visited infinitely often almost surely. This assumption may be relaxed under only unique ergodicity, though the initial state is to be restricted in such a setup.
- For a deterministic stationary policy γ, where certain (y, u) pairs are not realized, empirical learning still applies. However in this case if $y \mapsto u$ is realized with probability 0, $C^*(y,u)$ and $P^*(y'|y,u)$ are not well-defined in the above (6)-(8). In this case, the model, however, is such that when the state is y, the action u would not be an admissible control action. Such a situation may violate the continuity process in learning as policies converge; this will be discussed further in Section 4. Accordingly, experimentation in actions is critical.

3 Agentic Subjective Q-Learning Equilibrium

Definition 3.1. *A collection of policies $\gamma = \{\gamma^1, \cdots, \gamma^N\}$ is an agentic subjective Q-learning equilibrium if:*

- *Each agent has an approximate state space $\mathbb{Y}^i = \rho^i(\mathbb{X})$ with $|\mathbb{Y}^i| < \infty$ such that*

$$\mathbb{X} \ni x \mapsto \rho^i(x) = \hat{x} \in \mathbb{Y}^i,$$

and fixed point $Q^{*,i}$

$$Q^{*,i}(\hat{x}, u) = C^{*,i}(\hat{x}, u) + \beta \sum_{\hat{x}_1} P^{*,i}(\hat{x}_1|\hat{x}, u) \min_v Q^{*,i}(\hat{x}_1, v) \quad \hat{x} \in \mathbb{Y}^i, u \in \mathbb{U}^i \tag{10}$$

where for $B = (\rho^i)^{-1}(\hat{x})$ and $B_1 = (\rho^i)^{-1}(\hat{x}_1)$,

$$\begin{aligned} C^{*,i}(\hat{x}, u) &:= \int_{\mathbb{X}} c^i(x, \mathbf{u}) P(dx|\hat{x}) = \int_B c^i(x, \mathbf{u}) \hat{\pi}^{*,B}_\gamma(dx) \\ P^{*,i}(\hat{x}_1|\hat{x}, \mathbf{u}) &:= \int_{\mathbb{X}} \mathcal{T}(B_1|x, \mathbf{u}) P(dx|\hat{x}) = \int_B \mathcal{T}(B_1|x, \mathbf{u}) \hat{\pi}^{*,B}_\gamma(dx). \end{aligned} \tag{11}$$

such that $\hat{\pi}^{*,B}_\gamma$ is the invariant measure of the x_t process normalized for the set B with $\hat{\pi}^*(A) = \frac{\pi^*(A)}{\pi^*(B)}$ for all $A \subset B$ and π^*_γ is the invariant measure under γ.

- *For all $i = 1, \cdots, N$:*

$$\min_{u \in \mathbb{U}^i} Q^{*,i}(\hat{x}, u) = Q^{*,i}(\hat{x}, \gamma^i(\hat{x})) \tag{12}$$

Equivalently, and alternatively, each agent finds $\gamma^i : \mathbb{Y}^i \to \mathbb{U}^i$ by solving (9) under the empirically learned model leading to (11).

In particular, the best response solves the approximate MDP's fixed point equation (3) and the best response leads to the invariant measure defining the empirical MDP. Observe that $\hat{\pi}^*$ is an invariant measure obtained under the exploration policy. These two have to be compatible.

In particular, each agent thus aims to minimize a discounted cost criterion: However, in view of the discussion above, the agents optimize a discounted cost criterion either

- implicitly (with an agentic reinforcement learning motivation) via running the learning algorithm above in (2), or
- explicitly (with a subjective modeling motivation) by optimizing the discounted cost (9) (for the empirically learned model) with corresponding cost and transition kernel estimates (6)-(8).

That is, with the policies applied under such an equilibrium policy, the algorithm above should preserve the policies.

The discussion in Section 2.2 is then a justification on why we call this equilibrium *subjective*: Similar to Arslan & Yüksel (2023); Fudenberg & Levine (1993); Wiszniewska-Matyszkiel (2017); Dudebout & Shamma (2012; 2014), here the agents perceive and then fit the environment and data into a Markovian model, and then arrive at their best response maps. If the response maps are consistent with the induced model in that the agent policies generate the data fitted into induced models which then in turn generate the same policies via best response maps of the agents, then we arrive at a subjective equilibrium.

However, so far, the discussion does not touch on the convergence, therefore the equilibrium should be viewed only as a mathematical construction involving (15)-(11) and (12). The implication involving a collection of agents who perceive their quantized states as actual states and run their Q-learning algorithm will be studied further below.

4 Existence of Agentic Subjective Q-Learning Equilibrium

In this section, we establish the existence of such an agentic equilibrium as defined in Definition 3.1.

Assumption 4.1. *Let the original Markov model (1) be such that the induced kernel $\mathcal{T}$ is so that the family of conditional probability measures $\{\mathcal{T}(dx_{t+1}|x,\mathbf{u}), x \in \mathbb{X}, \mathbf{u} \in \prod_m \mathbb{U}^m\}$ admit densities $f_{x,\mathbf{u}}$ with respect to a reference measure ψ and these densities are bounded and equicontinuous over $x \in \mathbb{X}, \mathbf{u} \in \prod_m \mathbb{U}^m$. Furthermore, under any collection of stationary agentic policies, the induced Markov chain is positive Harris recurrent and admits a unique invariant probability measure.*

Observe that for the case when $\mathbb{X}, \mathbb{U}^i$ are finite for all $i \in \mathcal{N}$, irreducibility of the induced Markov model under any collection of stationary policies would satisfy Assumption 4.1.

In the following, we review a critical supporting result. Let $\mathbb{S}$ and $\mathbb{Y}$ be Borel sets in complete separable and metric spaces. The set of stochastic kernels from $\mathbb{S} \to \mathbb{Y}$, also called regular conditional probability measures, is the set

$$\Gamma_R = \Big\{\gamma : \gamma \text{ is a measurable function from } \mathbb{S} \text{ to } \mathcal{P}(\mathbb{Y})\Big\}, \tag{13}$$

where $\mathcal{P}(\mathbb{Y})$ is endowed with the Borel σ-algebra generated by the weak convergence topology.

A mathematically versatile topology on such stochastic kernels is the following Young topology, see Young (1937); Balder (1997; 1988); Florescu & Godet-Thobie (2012) for several properties.

Definition 4.1. *[Young Topology on kernels at reference (input) measure μ.] (Yüksel, 2024, Definition 3.1) Let μ be a σ-finite measure. A sequence of kernels $\gamma_n \to \gamma \in \Gamma_S$ under Young topology at input μ, if the joint measure $(\mu\gamma_n) \to (\mu\gamma)$ in the following sense: for every measurable and bounded $g : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ with $g(x,\cdot) : u \mapsto g(x,u)$ continuous and $\int \mu(dx) \sup_{u\in\mathbb{U}} |g(x,u)| < \infty$,*

$$\int \mu(dx)\Big(\gamma_n(du|x)g(x,u)\Big) \to \int \mu(dx)\Big(\gamma(du|x)g(x,u)\Big) \tag{14}$$

A key supporting result is the following (Yüksel, 2024, Theorem 4.1). If the setup involves a single agent and if the randomized stationary policy space Γ_S is endowed by the Young topology at input ψ (Yüksel, 2024, Definition 3.2), then the invariant measure induced by a policy is continuous (weakly and in total variation) on Γ_S. In our setup, since the policy spaces, under their respective finite models, for each agent is finite dimensional, such agent-wise policy convergence implies the convergence of the joint control policy from $x \to \mathcal{P}(\prod_{m=1}^N \mathbb{U}^m)$ both pointwise as well as under Young topology, as we discuss in the proof of the theorem below.

Theorem 4.1. *[Existence of Agentic Subjective Q-Learning Equilibrium] Let Assumption 4.1 hold. Then, there exists an agentic Q-learning equilibrium, which is an equilibrium under the (subjectively quantized) Q-learning dynamics as given above.*

Proof. We recall the Kakutani-Fan-Glicksberg theorem: *(Aliprantis & Border, 2006, Corollary 17.55): Let K be a compact non-empty convex subset of a locally convex Hausdorff space and let the set-valued map $\phi : K \to 2^K$ have convex values and have a closed graph $G = \{(x, \phi(x)), x \in K\}$ (i.e., if $x_n \to x$, and if $\phi(x_n) \to z$, then $(x, z) \in G$). Then, ϕ admits a fixed point.*

In view of the Kakutani-Fan-Glicksberg, we state the following:

- The policy profile $\{\gamma^m, m = 1, \cdots, N\}$ is in the product space $\prod_{m=1}^N \Gamma^m$ and consists of stationary and randomized policies γ^m mapping the agentic states $\hat{x}^m$ to $\mathcal{P}(\mathbb{U}^m)$. This product space is finite dimensional, convex, and compact valued under the product topology, where for each local state the conditional probability is metrized under the weak topology.

- Given a policy profile $\{\gamma^m, m = 1, \cdots, N\}$ and its induced model via (4) the best response map

$$\mathbf{BR} : \prod_{m=1}^N \Gamma^m \ni \gamma = \{\gamma^m, m = 1, \cdots, N\} \mapsto \mathbf{BR}(\gamma) \in 2^{\prod_{m=1}^N \Gamma^m},$$

  defined as the optimal policies corresponding to the model (4) leading to (12), defines a best response map **BR** which is convex valued: the Q-learning algorithm is the fixed point equation solution and an optimal policy solves the dynamic programming equation with value given as (12): Thus, with

$$J_\beta^{*,i}(\hat{x}) = \min_{u \in \mathbb{U}^i} Q^{*,i}(\hat{x}, u)$$

  we have that an optimal (possibly randomized) policy solves and thefore has full mass on the set of actions which attain

$$\min_{u \in \mathbb{U}^i} \left(C^{*,i}(\hat{x}, u) + \beta \sum_{\hat{x}_1} P^{*,i}(\hat{x}_1 | \hat{x}, u) J_\beta^{*,i}(\hat{x}_1) \right) \qquad \hat{x} \in \mathbb{Y}^i, \tag{15}$$

  and therefore convex combinations of optimal policies are also best response policies.

- The map from policy spaces to invariant measure $\hat{\pi}$ and model (4) is continuous, building on (Yüksel, 2024, Theorem 4.1)[1]: In particular, by a generalized dominated convergence theorem (Serfozo, 1982, Theorem 3.5), the joint distribution

$$P^{\gamma_n^{-i}}(dx_1|x) \to P^{\gamma^{-i}}(dx_1|x),$$

  converges weakly as $\gamma_n^m(du|x) \to \gamma^m(du|x)$ pointwise in x. That is the product of the control policy measures induce a kernel $\mathbb{X} \to \mathcal{P}(\prod_{m=1}^N \mathbb{U}^m)$ which converges weakly pointwise in x and thus the joint policy $\gamma_n(du|x) \to \gamma(du|x)$ pointwise. By an application of the dominated convergence theorem applied to (14), convergence is then also under the Young topology (see Definition 4.1). Accordingly, under Assumption 4.1, we have that the invariant measures also converge (both weakly and in total variation); that is $\pi_{\gamma_n} \to \pi_\gamma$ where π_{γ_n} is the invariant measure under γ_n and π_γ is so under γ (Yüksel, 2024, Theorem 4.1).

  As a result, as policies converge their induced models (4) converge continuously in total variation, and accordingly the value functions $J_\beta^*(\hat{x})$ converge (Kara & Yüksel, 2020, Section 4).

  Therefore, any weak limit of the best response policies is also optimal as they solve the optimality equation (3) or (15). This ensures that the best response map is upper semi-continuous and the graph $G = \{\gamma, \mathbf{BR}(\gamma), \gamma \in \mathbf{\Gamma}\}$ noted above is closed.

Thus, existence of an equilibrium follows. □

[1]While this is not in general correct for standard Borel spaces where multiple agents are concerned, due to the finite model seen by each agent, the finite dimensional policies converge pointwise.

Remark 4.1. *Observe that in the above, we have a purely mathematical concept for equilibrium characterized by the limit equations (11) and best response maps (12). We have not yet discussed the conditions for Q-learning or empirical learning to converge to the limit and it is critical in the above proof that $C^{*,i}$ and $P^{*,i}$ in (11) are to be defined under every admissible stationary policy: Nonetheless, small perturbations in policies lead to small changes in values and invariant measures (Yüksel, 2024, Theorem 4.1), and this discussion motivates the algorithmically and practically consequential discussion in the following section. Consider an exploration policy γ which is deterministic where for a given state realization $\hat{x}$, the action u is not realized: Then, if u is not an admissible state under the model induced by the deterministic policy, a construction can be made where while under every small randomized perturbation of the policy γ, $(\hat{x}, u)$ would be an admissible pair and also an optimal pair for the best response, in the limit of the deterministic exploration policy γ, such a pair would be inadmissible: therefore the closed graph property would be violated.*

5 Convergence of Policy Revision Processes via Subjective ϵ-Satisficing Dynamics to ϵ-Agentic Subjective Q-Learning Equilibria

In this section, we present a learning algorithm, characterized by a policy revision process over play paths. This approach builds on the analysis presented in Arslan & Yüksel (2017) (see also Germano & Lugosi (2007) as a related work) for weakly acylic stochastic games which was subsequently generalized in view of *satisficing* policy revision paths Yongacoglu et al. (2023). In this approach, the agents are allowed to use constant policies for extended periods of time called *exploration phases*. As illustrated in Figure 1, the $k-$th exploration phase runs through times $t = t_k, \ldots, t_{k+1} - 1$, where

$$t_{k+1} = t_k + T_k \qquad (\text{with } t_0 = 0)$$

for some integer $T_k \in [1, \infty)$ denoting the length of the $k-$th exploration phase. During the $k-$th exploration phase, DMs use some constant policies $\pi_k^1, \ldots, \pi_k^N$ as their (baseline) policies with occasional experimentation.

The main idea is to create a stationary environment over each exploration phase so that DMs can almost accurately learn their optimal Q-factors corresponding to the constant policies used during each exploration phase and update their policies; see Figure 2.

This paradigm has been adopted under two types of policy updates: (i) *Best response dynamics with inertia* for weakly acyclic games Arslan & Yüksel (2017) considered for the case where each agent has access to the global state but only local actions (requiring typically deterministic policies), and (ii) a variation of it which is referred to as *satisficing paths* dynamics Yongacoglu et al. (2024; 2023) which assumes that the agents have access to a variety of information states and the policies may be randomized. Arslan & Yüksel (2017) established convergence involving pure strategies for finite games which are weakly acyclic (including stochastic team) problems, via best-response or better-response dynamics with inertia. Yongacoglu et al. (2023; 2024) introduced the paradigm of satisficing paths and showed applicability to two classes of stochastic dynamic games: Two-player games and symmetric games. In particular, a policy revision process is called an ϵ-satisficing one if the performance of the policy is within $\epsilon > 0$ of the optimal policy:

Definition 5.1. *Let $\epsilon \geq 0$ and T^i be a policy update rule for Agent $i \in \mathcal{N}$. The policy update rule T^i is said to be ϵ-satisficing if, for any $\boldsymbol{\pi} = (\pi^i, \boldsymbol{\pi}^{-i}) \in \boldsymbol{\Gamma}_S$, we have that $\pi^i \in \mathrm{BR}^i_\epsilon(\boldsymbol{\pi}^{-i})$ implies $T^i(\boldsymbol{\pi}) = \pi^i$.*

A policy revision process is called ϵ-satisficing if it is associated with a collection of policy update rules $\mathbf{T} = \{T^i\}_{i \in \mathcal{N}}$ such that T^i is ϵ-satisficing for each Agent $i \in \mathcal{N}$.

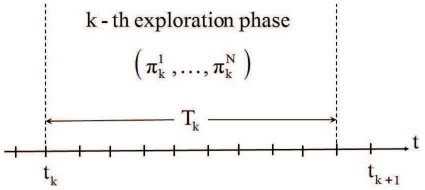

Figure 1: An illustration of the $k-$th exploration phase.

Definition 5.2. *Let $\epsilon \geq 0$. A (possibly finite) sequence $(\boldsymbol{\pi})_{k\geq 0}$ of stationary joint policies is called an ϵ-satisficing path if, for every $k \geq 0$ and $i \in \mathcal{N}$, $\pi_k^i \in \mathrm{BR}_\epsilon^i(\boldsymbol{\pi}_k^{-i})$ implies $\pi_{k+1}^i = \pi_k^i$.*

Definition 5.3. *Let $\epsilon \geq 0$ and let $\boldsymbol{\Pi} \subseteq \boldsymbol{\Gamma}_S$ be a subset of stationary joint policies. A game is said to have the* **ϵ-satisficing** *paths property within $\boldsymbol{\Pi}$ if for every $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, there exists an ϵ-satisficing path $(\boldsymbol{\pi}_t)_{t\geq 0}$ and an integer $K = K(\boldsymbol{\pi})$, such that (i) $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$, (ii) $\boldsymbol{\pi}_t \in \boldsymbol{\Pi}$ for every $t \geq 0$, and (iii) $\boldsymbol{\pi}_K \in \boldsymbol{\Gamma}_S^{\epsilon-eq}$ (which is the set of ϵ-equilibrium stationary policies).*

The revision process is called **subjective ϵ-satisficing**, if the ϵ-optimality holds under the subjectively evaluated criteria, as is done in our paper.

In view of this discussion, consider the following algorithm, which considers the satisficing paths paradigm.

Algorithm 1: Independent Learning with Local Information and Convergence to Agentic Subjective Q-Learning Equilibria

Set Parameters
- $\Pi^i \subset \Gamma_S^i$: a fine quantization of Γ_S^i consisting of policies charging each action with positive probability
- $\{T_k\}_{k\geq 0}$: a sequence in $\mathbb{N}$ of learning phase lengths
  - set $t_0 = 0$ and $t_{k+1} = t_k + T_k$ for all $k \geq 0$.
- $e^i \in (0,1)$: random policy updating probability
- $d^i \in (0,\infty)$: tolerance level for sub-optimality

Initialize $\pi_0^i \in \Pi^i$ (arbitrary), $\widehat{Q}_0^i = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{U}}$, $\widehat{J}_0^i = 0 \in \mathbb{R}^{\mathbb{Y}}$

for $k \geq 0$ (k^{th} exploration phase)

  for $t = t_k, t_k+1, \ldots, t_{k+1}-1$

    Observe $y_t^i = \varphi^i(\mathbf{x}_t)$

    Select $u_t^i \sim \pi_k^i(\cdot|y_t^i)$

    Observe $c_t^i := c(x_t^i, \mu(\cdot|\mathbf{x}_t), u_t^i)$ and y_{t+1}^i

    Set $n_t^i = \sum_{\tau=t_k}^t \mathbf{1}\{(y_\tau^i, u_\tau^i) = (y_t^i, u_t^i)\}$

    Set $m_t^i = \sum_{\tau=t_k}^t \mathbf{1}\{y_\tau^i = y_t^i\}$

    $\widehat{Q}_{t+1}^i(y_t^i, u_t^i) = \left(1 - \frac{1}{n_t^i}\right)\widehat{Q}_t^i(y_t^i, u_t^i) + \frac{1}{n_t^i}\left[c_t^i + \beta \min_{a^i} \widehat{Q}_t^i(y_{t+1}^i, a^i)\right]$

    $\widehat{J}_{t+1}^i(y_t^i) = \left(1 - \frac{1}{m_t^i}\right)\widehat{J}_t^i(y_t^i) + \frac{1}{m_t^i}\left[c_t^i + \beta \widehat{J}_t^i(y_{t+1}^i)\right]$

  if $\widehat{J}_{t_{k+1}}^i(y) \leq \min_{a^i} \widehat{Q}_{t_{k+1}}^i(y, a^i) + \epsilon + d^i \ \forall y \in \mathbb{Y}$, **then**

    $\pi_{k+1}^i = \pi_k^i$

  else

    $\pi_{k+1}^i \sim (1-e^i)\delta_{\pi_k^i} + e^i \mathrm{Unif}(\Pi^i)$

  Reset $\widehat{J}_{t_{k+1}}^i = 0 \in \mathbb{R}^{\mathbb{Y}}$ and $\widehat{Q}_{t_{k+1}}^i = 0 \in \mathbb{R}^{\mathbb{Y}\times\mathbb{U}}$

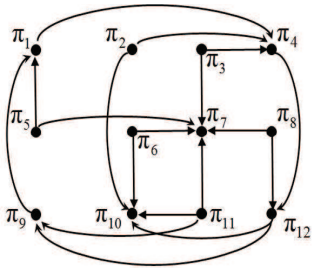

Figure 2: The response graph of a stochastic game.

In the above, we have two iterations: In one, the limit of $\widehat{J}^i_t(\cdot)$, is the (subjectively, under the induced model) evaluated cost under the policy corresponding to the assumed model; and the limit of $\min_{a^i} \widehat{Q}^i_{t_{k+1}}(\cdot, a^i)$ gives the subjectively computed optimal cost via the limit $\widehat{Q}^i$. As in (Yongacoglu et al., 2024, Section 5 and Appendix J), subjective ϵ-satisficing dynamics possess the same graph theoretic properties as (objective) satisficing.

Theorem 5.1. *[Convergence to Subjective Q-Learning Equilibria] If the Q-learning play path admits a subjective ϵ-satisficing path, Algorithm 1 converges to policy profiles in an ϵ subjective Q-learning equilibrium with arbitrarily high probability (by taking the exploration length satisfy $T_k \geq T, k \in \mathbb{N}$ for sufficiently large T) from any initial policy profile.*

Proof. Under Assumption 4.1, during exploration, a small perturbation leads to a small change in both induced cost as well as the invariant measure used for exploration.

As all agents use constant policies throughout any particular exploration phase, each agent indeed faces an ergodic environment in each exploration phase. Therefore, if the length of each exploration phase is long enough and the soft-policy inducing experimentation probabilities are small enough, each agent can learn its corresponding subjective Q-factors, $\widehat{Q}^i$, in each exploration phase with arbitrarily high probability. Likewise, the subjectively evaluated cost $\widehat{J}^i$ can also be learned with arbitrarily high accuracy.

This thus allows each agent to accurately, with arbitrarily high probability, compute $\widehat{Q}^i$ and $\widehat{J}^i$ at the end of each exploration phase. By allowing each agent to update its policy π^i_k with some probability $e^i \in (0, 1)$ when not ϵ-subjectively satisfied, the resulting policy adjustment process approximates a satisficing reply process.

This then induces a policy revision process for which equilibria serve as near-absorbing sets which are also accessible (with a positive transition probability bounded from below uniformly over all initial policy profiles): The result is then a corollary of the analysis in (Yongacoglu et al., 2024, Theorem 25), given the existence of a subjective equilibrium under the subjective ϵ-satisficing condition. □

Examples.

- **Symmetric Games** (Yongacoglu et al., 2023, Theorem 3.6) showed that symmetric normal form games admit (ϵ-)satisficing paths *provided that they admit ϵ-Nash equilibria.* Furthermore, identical quantization of information preserves the symmetry and also the existence of satisficing paths. Therefore, existence of equilibria presented in Theorem 4.1 together with the satisficing paths lead to the applicability of Theorem 5.1. In particular symmetric Markov games with symmetric local state information (with identical quantization maps) satisfy such conditions.

- **Mean Field Games** Mean-field games constitute a significant special case of symmetric games. In this context, Yongacoglu et al. (2024) developed a comprehensive analysis of convergence and decentralized learning. The analysis in our current paper is in fact motivated by the study of mean-field games, in view of an open question of equilibrium existence and convergence for agentic subjective Q-learning equilibria (the mean-field case allow for tailored arguments on existence). Our analysis in this paper shows that, mean-field games with only local state information as well as mean-field games with local state and compressed mean-field information available at the agents do admit satisficing paths to equilibria since the set of equilibria is non-empty.

- **Team Games, Weakly Acyclic Games, and Generalized Weakly Acyclic Games** If the quantized games are generalized weakly acyclic, then convergence to equilibria holds as in Arslan & Yüksel (2017). In particular, ϵ-satisficing paths exist for such problems as long as equilibria exist: For team problems involving agents with identical cost functions, if the quantizers are such that the diameter of each bin is small enough, then ϵ-satisficing paths always exist since the approximate MDP is near optimal when compared with the original MDP by Saldi et al. (2017; 2024).

- **Single Agent Case** In the case with single agent stochastic control, by definition satisficing dynamics to equilibria exist: the policy revision process is ϵ-satisficing. Note that the policy space is

compact, and the induced cost is continuous in the (quantized) policy space; accordingly for every $\epsilon > 0$ one can discretize the space of randomized stationary policies which leads to an ϵ perturbation in the realized expected costs. Therefore, learning among a finitely many policies is feasible. This does, therefore, also provide a recipe for policy learning in MDPs even though there are more efficient algorithms for such a setting. Accordingly, iterative best responding is guaranteed to converge to an equilibrium with high probability.

- **A Remark on Subjective vs. Objective Equilibria** Note that the analysis here is not with respect to an exact ϵ-equilibrium, but only a subjective one. For Markov games with Borel spaces, Saldi et al. (2024) showed that such a subjective equilibrium is in fact also ϵ-equilibrium under high-rate quantization/state approximation and further technical conditions on the kernels which would ensure that either the value functions are either sufficiently regular (such as in team games or zero-sum games) or the transition kernels are sufficiently regular (such as admitting total variation continuity).

6 Conclusion

In this paper, we introduced the concept of agentic subjective Q-Learning equilibrium, where agents apply Q-learning under an incorrect state space modeling assumption. An equivalent empirical learning interpretation is also presented. For such problems, a distinguishing feature is that the exploration policy itself impacts the perceived model and therefore there is a dual dependence on the agent policies. We established an existence result on such equilibria. We then presented an associated convergence theorem to ϵ-equilibria for a large class of stochastic dynamic games.

References

S. Adlakha, R. Johari, and G.Y. Weintraub. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory*, 156:269–316, 2015.

C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis*. Berlin, Springer, 3rd ed., 2006.

G. Arslan and S. Yüksel. Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62:1545 – 1558, 2017.

G. Arslan and S. Yüksel. Subjective equilibria under beliefs of exogenous uncertainty for dynamic games. *SIAM Journal on Control and Optimization*, 61(3):1038–1062, 2023.

E. J. Balder. Consequences of denseness of dirac young measures. *Journal of Mathematical Analysis and Applications*, 207(2):536–540, 1997.

E.J. Balder. Generalized equilibrium results for games with incomplete information. *Mathematics of Operations Research*, 13(2):265–276, 1988.

S. Chandak, V.S. Borkar, and P. Dodhia. Reinforcement learning in non-markovian environments. *Systems & Control Letters*, 185:105751, 2024.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Tenth Innovative Applications of Artificial Intelligence Conference, Madison, Wisconsin*, pp. 746–752, 1998.

Anne Condon. On algorithms for simple stochastic games. *Advances in Computational Complexity Theory*, 13:51–72, 1990.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33:5527–5540, 2020.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pp. 5166–5220. PMLR, 2022.

S. Dong, B. van Roy, and Z. Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *The Journal of Machine Learning Research*, 23(1):11627–11680, 2022.

Nicolas Dudebout and Jeff S Shamma. Empirical evidence equilibria in stochastic games. In *51st IEEE Conference on Decision and Control*, pp. 5780–5785, 2012.

Nicolas Dudebout and Jeff S Shamma. Exogenous empirical-evidence equilibria in perfect-monitoring repeated games yield correlated equilibria. In *53rd IEEE Conference on Decision and Control*, pp. 1167–1172, 2014.

L.C. Florescu and C. Godet-Thobie. *Young measures and compactness in measure spaces*. Walter de Gruyter, 2012.

Roy Fox, Stephen M. McAleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in Markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4414–4425. PMLR, 2022.

D. Fudenberg and D.K. Levine. Self-confirming equilibrium. *Econometrica: Journal of the Econometric Society*, pp. 523–545, 1993.

F. Germano and G. Lugosi. Global nash convergence of foster and young's regret testing. *Games and Economic Behavior*, 60(1):135–154, 2007.

O. Hernández-Lerma and J. B. Lasserre. *Markov Chains and Invariant Probabilities*. Birkhäuser, Basel, 2003.

J. Hu and M. P. Wellman. Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, **4**:1039–1069, 2003.

Ehud Kalai and Ehud Lehrer. Subjective games and equilibria. *Games and Economic Behavior*, 8(1):123–163, 1995.

A.D Kara and S. Yüksel. Robustness to incorrect system models in stochastic control. *SIAM Journal on Control and Optimization*, 58(2):1144–1182, 2020.

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*, 48(4):2066–2093, 2023.

A.D. Kara and S. Yüksel. Q-learning for stochastic control under general information structures and non-markovian environments. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=1Yp6xpTV55`. Featured Certification.

A.D Kara, N. Saldi, and S. Yüksel. Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity. *Journal of Machine Learning Research*, pp. 1–34, 2023.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. In *International Conference on Learning Representations*, 2022.

Michael L. Littman. Friend-or-foe Q-learning in general-sum games. In *International Conference on Machine Learning*, volume 1, pp. 322–328, 2001.

Michael L. Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *International Conference on Machine Learning*, volume 96, pp. 310–318, 1996.

Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Coordination of independent learners in cooperative Markov games. *HAL preprint hal-00370889*, 2009.

Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *Knowledge Engineering Review*, 27(1):1–31, 2012.

F. C. Melo, S. P. Meyn, and I. M. Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pp. 664–671, 2008.

S. P. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

David Mguni, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, and Jun Wang. Learning in nonzero-sum stochastic games with potentials. In *International Conference on Machine Learning*, pp. 7688–7699. PMLR, 2021.

N. Saldi, S. Yüksel, and T. Linder. On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. *Mathematics of Operations Research*, 42(4):945–978, 2017.

Naci Saldi, Gurdal Arslan, and Serdar Yuksel. Existence of ϵ-nash equilibria in nonzero-sum borel stochastic games and equilibria of quantized models. *arXiv preprint arXiv:2411.10805*, 2024.

M. Sayin, K. Zhang, D. Leslie, T. Başar, and A. Ozdaglar. Decentralized q-learning in zero-sum markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334, 2021.

Muhammed O. Sayin, Francesca Parise, and Asuman Ozdaglar. Fictitious play in zero-sum stochastic games. *SIAM Journal on Control and Optimization*, 60(4):2095–2114, 2022.

Sandip Sen, Mahendra Sekaran, and John Hale. Learning to coordinate without sharing information. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pp. 426–431, 1994.

R. Serfozo. Convergence of Lebesgue integrals with varying measures. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 380–402, 1982.

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

C. Szepesvàri and W.D. Smart. Interpolation-based q-learning. 2004.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337, 1993.

Onur Unlu and Muhammed O. Sayin. Episodic Logit-Q dynamics for efficient learning in stochastic teams. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 1985–1990. IEEE, 2023.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Breaking the curse of many agents: Provable mean embedding Q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pp. 10092–10103. PMLR, 2020.

Gabriel Y Weintraub, C Lanier Benkard, and Benjamin Van Roy. Industry dynamics: Foundations for models with an infinite number of firms. *Journal of Economic Theory*, 146(5):1965–1994, 2011.

Agnieszka Wiszniewska-Matyszkiel. Open and closed loop Nash equilibria in games with a continuum of players. *Journal of Optimization Theory and Applications*, 160(1):280–301, 2014.

Agnieszka Wiszniewska-Matyszkiel. Redefinition of belief distorted Nash equilibria for the environment of dynamic games with probabilistic beliefs. *Journal of Optimization Theory and Applications*, 172(3): 984–1007, 2017.

B. Yongacoglu, G. Arslan, and S. Yüksel. Satisficing paths and independent multi-agent reinforcement learning in stochastic games. *SIAM Journal on Mathematics of Data Science (arXiv:2110.04638)*, 2023.

B. Yongacoglu, G. Arslan, and S. Yüksel. Mean-field games with finitely many players: Independent learning and subjectivity. *Journal of Machine Learning Research*, 25(419):1–69, 2024.

L.C. Young. Generalized curves and the existence of an attained absolute minimum in the calculus of variations. *Comptes Rendus de la Societe des Sci. et des Lettres de Varsovie*, 30:212–234, 1937.

S. Yüksel. On Borkar and Young relaxed control topologies and continuous dependence of invariant measures on control policy. *SIAM Journal on Control and Optimization*, 62(4):2367–2386, 2024.

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in Markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935, 2022.