Bandit Online Learning of Nash Equilibria in Monotone Games

Tatiana Tatarenko and Maryam Kamgarpour, IEEE Member

***Abstract*—We address online bandit learning of Nash equilibria in multi-agent convex games. We propose an algorithm whereby each agent uses only obtained values of her cost function at each joint played action, lacking any information of the functional form of her cost or other agents' costs or strategies. In contrast to past work where convergent algorithms required strong monotonicity, we prove that the algorithm converges to a Nash equilibrium under mere monotonicity assumption. The proposed algorithm extends the applicability of bandit learning in several games including zero-sum convex games with possibly unbounded action spaces, mixed extension of finite-action zero-sum games, as well as convex games with linear coupling constraints.**

***Index Terms*—learning in games, monotone bandit learning**

I. Introduction

Game theory is a powerful framework to address optimization and learning of multiple interacting agents referred to as players. Such multi-agent problems arise in several application domains including traffic networks, internet, auctions and adversarial learning. In a multi-agent setting, the notion of Nash equilibrium captures a desirable solution as it exhibits stability. Namely, no player has an incentive to unilaterally deviate from this solution. Nash equilibrium is also consistent with a notion referred to as rationality - each player aims to optimize her own cost function. Given these theoretical justifications for Nash equilibria, from the viewpoint of learning the question is whether players can learn their Nash equilibrium strategies with limited information about the game. In particular, in several application domains, each player might not know the functional form of her objective. For example, the travel times of edges in a traffic network or the market constraints in an auction are unknown a priori and depend in non-trivial ways on actions and objectives of other players. However, by playing the game, a player can have access to the so-called online bandit information feedback, namely, she can receive payoff information of her objective (zeroth-order oracle) for any feasible joint actions taken by all the players. Thus, we focus on how do players learn Nash equilibria given online bandit information.

Bandit learning in games has been mainly explored in finite action settings. It is known that if each player uses a no-regret algorithm, the time average of the sequence of played actions converges to a coarse-correlated equilibrium - a relaxed notion of equilibrium which encompasses Nash equilibria as well as possibly non-rationalizable strategies. The convergence of time-averaged sequence of plays in no-regret algorithms to a mixed strategy Nash equilibrium can be established for special games such as two-player zero-sum game. However, the convergence of time-averaged actions does not imply the actual sequence of plays also converges to a Nash equilibrium. This issue was recently re-examined in light of progresses in bandit convex optimization. In particular, [13] showed that if each player uses a fairly general model of no-regret algorithm (FTRL) in continuous-time, the actual sequence of plays may not converge to a mixed strategy Nash equilibrium in a two-player zero-sum game and it might exhibit a non-vanishing cycle[1]. The work of [2] analyzed this non-convergence through the lens of Hamiltonian and potential of a game and proposed convergent algorithms assuming access to exact first and second order oracles (gradients and Hessians) of the players' objectives. More recently, motivated by the Hamiltonian analogy of zero-sum games, [1] extended the analysis of [13] to discrete-time setting and showed that sequential gradient descent can overcome the divergence of discretized simultaneous gradient descent in these games.

The mixed extension of a finite action game falls into category of convex games. In such games, each player's objective is convex in her own decision variable for any fixed action of other players (in mixed extension of finite action games, this objective is linear in each player's action). Furthermore, the strategy sets are convex (simplexes in mixed extension games). It is known that the Nash equilibria in convex differentiable games correspond to the solution set of a variational inequality problem. Hence, one can use algorithms for solving variational inequalities to compute Nash equilibria. This connection has been used in two lines of works addressing bandit learning of Nash equilibria in convex games. The works in [18, 4] showed that no-regret learning can converge to Nash equilibria in certain class of convex games. Both works leveraged the idea of one-point estimation of the gradient of the player's cost function using the bandit payoff information. Whereas [4] performed this gradient estimation by perturbing the played actions with a point sampled from the uniform distribution on the unit sphere, motivated by Stoke's theorem [6, 9], the approach in [18] formed these gradient estimates using Normal distribution and smoothing motivated by [20, 14]. In both cases, the convergence was proven by appropriately choosing the stepsizes to tradeoff the bias and variance of the resulting estimation terms in the stochastic approximation procedure.

T. Tatarenko (tatiana.tatarenko@rmr.tu-darmstadt.de) is with the Control Methods and Robotics Lab Technical University Darmstadt, Darmstadt, Germany 64283, M. Kamgarpour (maryamk@ece.ubc.ca) is with Electrical and Computer Engineering, UBC, Vancouver, Canada.

M. Kamgarpour gratefully acknowledges ERC Starting Grant CONENE.

[1]The FTRL algorithms explored had access to more than bandit feedback as each player could observe the gradient of its mixed extension cost function at each time step.

The above works on learning Nash equilibria rely on strict monotonicity of the game mapping. The game mapping is the stacked vector of gradient of each player's objective with respect to her actions. In case the game is potential, the game mapping is symmetric and thus, the game mapping corresponds to gradient of a single function, the so-called potential function of the game. Hence, finding equilibria can be cast as a single-objective optimization and bandit algorithms can be readily applied to learn Nash equilibria. However, in general noncooperative games, this game mapping need not be symmetric. In such cases, a necessary and sufficient condition for convergence of no-regret algorithms such as those explored in [18, 4, 13] is strict monotonicity of the game mapping.

The strict monotonicity of the game mapping is a stringent condition. The non-convergence issue and in particular cyclic behavior discussed in [13, 2, 1] is due to the fact that the game mapping in several games do not satisfy strict monotonicity. Zero-sum games for example are only monotone. This class of games have been widely adopted in robust optimization and control, in models of perfect competition and more recently in adversarial training and deep learning. In addition to zero-sum games, generalized Nash equilibria problems, that is, Nash equilibria with coupling constraints can at best have monotone extended game mappings due to the coupling constraints. The generalized Nash equilibria problems arise in several domains where a resource must be shared between agents and this can only be formulated as a hard rather than soft constraints [5]. The relevance of games with merely monotone mapping, motivate our paper on learning Nash equilibria under bandit feedback in this game class.

Given the connection of Nash equilibria of convex games with solution set of variational inequality (VI) problem, a natural starting point for learning Nash equilibria of merely monotone games is to search for algorithms for solving the corresponding VIs. This topic is well-explored in [16, Chapter 12] and several approaches including extra-gradient and Tikhonov regularization for finding solution of merely monotone or pseudo-monotone VIs are proposed. All the proposed approaches though require at least first-order (gradient) feedback. The challenge with generalizing these approaches to zeroth-order (bandit) feedback is that they require coordination among the players and are not suitable in the bandit feedback setting because a player cannot sample her objective functions at different points while ensuring actions of other players remains fixed. For example, the extra-gradient algorithm in [12] remedies the cyclic behavior of trajectories in monotone game. However, it is only applicable to exact first-order feedback oracle. On the other hand, standard Tikhonov regularization requires a double iterative procedure, where the players would solve a regularized VI corresponding to a regularized game mapping in each inner iteration. Here, solving the inner VI itself would require an iterative algorithm and it is not clear how the players should coordinate stopping time for the inner algorithm in addition to setting the regularization parameter.

Motivated by generalizing the algorithms for learning Nash equilibria, in our recent work [19] we proposed an approach to learn Nash equilibria in merely monotone games. Our approach was inspired by the single timescale Tikhonov regularization algorithm for solving stochastic variational inequalities [11]. In contrast to the above work that assumed access to noisy gradients, we considered the single-point estimation of gradients using the bandit feedback. The main contribution was to appropriately control the bias and variance introduced in the single-point estimation along with the Tikhonov regularization parameter to ensure convergence. Our proposed algorithm was not applicable to online learning because the played actions were not bound to lie in the feasible set. Our current paper fills this gap by providing a convergent algorithm while ensuring the query points do lie in the feasible set.

Our contributions in this paper are as follows. We develop, to our best knowledge, the first bandit approach for online learning Nash equilibria in convex games with merely monotone game mappings. In doing so, we propose a novel single time-scale double-regularization - this double regularization corresponds to both the Tikhonov regularization and at the same time the projection of the actions onto a shrunk feasible set. By properly managing the interplay between the choice of the regularization sequence, the radius of shrinkage for the feasible set and the stepsize we ensure that bias and variance of the resulting stochastic approximation vanish with appropriate rates. The choice of parameters are stated in Assumption 6, whereas the main convergence result is stated in our Theorem 2. In terms of the proof techniques, there are few main novelties leading to this theorem: first, showing that the doubly regularized Tikhonov sequence and a single-time scale approach to solve it, remain bounded and converges to the least-norm solution of the variational inequality (see Theorem 3 and Proposition 1, respectively); second, showing that iterates of our algorithm remain bounded (see Lemma 5). These results enable us to use well-established results on boundedness and convergence of stochastic processes to our setup. In summary, we prove convergence in probability of the sequence of actions to a Nash equilibria in monotone games.

Notations. The set $\{1, \dots, N\}$ is denoted by $[N]$. Boldface is used to distinguish between vectors in a multi-dimensional space and scalars. Given N vectors $\boldsymbol{x}^i \in \mathbb{R}^d$, $i \in [N]$, $(\boldsymbol{x}^i)_{i=1}^N := (\boldsymbol{x}^{1\top}, \dots, \boldsymbol{x}^{N\top})^\top \in \mathbb{R}^{Nd}$; $\boldsymbol{x}^{-i} := (\boldsymbol{x}^1, \dots, \boldsymbol{x}^{i-1}, \boldsymbol{x}^{i+1}, \dots, \boldsymbol{x}^N) \in \mathbb{R}^{(N-1)d}$. $\mathbb{R}^d_+$ and $\mathbb{Z}_+$ denote, respectively, vectors from $\mathbb{R}^d$ with non-negative coordinates and non-negative whole numbers. The standard inner product on $\mathbb{R}^d$ is denoted by $(\cdot,\cdot)$: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, with associated norm $\|\boldsymbol{x}\| := \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$. Given some matrix $A \in \mathbb{R}^{d\times d}$, $A \succeq (\succ) 0$, if and only if $\boldsymbol{x}^\top A \boldsymbol{x} \geq (>) 0$ for all $\boldsymbol{x} \neq 0$. We use the big-O notation, that is, the function $f(x) : \mathbb{R} \to \mathbb{R}$ is $O(\mathbf{g}(x))$ as $x \to a$, $f(x) = O(g(x))$ as $x \to a$, if $\lim_{x\to a} \frac{|f(x)|}{|g(x)|} \leq K$ for some positive constant K. And with a slight abuse of notation, we write $f(x) \leq O(g(x))$ as we estimate certain bounds. We say that a function $f(\boldsymbol{x})$ grows not faster than a function $g(\boldsymbol{x})$ as $\boldsymbol{x} \to \infty$, if there exists a positive constant Q such that $f(\boldsymbol{x}) \leq g(\boldsymbol{x})$ $\forall \boldsymbol{x}$ with $\|\boldsymbol{x}\| \geq Q$. For $x \in \mathbb{R}^n$ and convex closed set $C \subset \mathbb{R}^n$, $\text{Proj}_C x$ denotes the projection of x onto C.

Definition 1: A mapping $\boldsymbol{M} : \mathbb{R}^d \to \mathbb{R}^d$ is *monotone* over $X \subseteq \mathbb{R}^d$, if $(\boldsymbol{M}(\boldsymbol{x}) - \boldsymbol{M}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}) \geq 0$ for every $\boldsymbol{x}, \boldsymbol{y} \in X$.

what about Q-learning on network games?

II. Problem Formulation

Consider a game $\Gamma(N, \{A_i\}, \{J_i\})$ with N players, the sets of players' actions $A_i \subseteq \mathbb{R}^d$, $i \in [N]$, and the cost (objective) functions $J_i : \boldsymbol{A} \to \mathbb{R}$, where $\boldsymbol{A} = A_1 \times \ldots \times A_N$ denotes the set of joint actions. We restrict the class of games as follows.

Assumption 1: The game under consideration is *convex*. Namely, for all $i \in [N]$ the set A_i is convex and closed, the cost function $J_i(\boldsymbol{a}^i, \boldsymbol{a}^{-i})$ is defined on $\mathbb{R}^{Nd}$, continuously differentiable in $\boldsymbol{a}$ and convex in $\boldsymbol{a}^i$ for fixed $\boldsymbol{a}^{-i}$.

Assumption 2: The mapping $\boldsymbol{M} : \mathbb{R}^{Nd} \to \mathbb{R}^{Nd}$, referred to as the *game mapping*, defined by

$$\boldsymbol{M}(\boldsymbol{a}) = (\nabla_{\boldsymbol{a}^i} J_i(\boldsymbol{a}^i, \boldsymbol{a}^{-i}))_{i=1}^N = (\boldsymbol{M}_1(\boldsymbol{a}), \ldots, \boldsymbol{M}_N(\boldsymbol{a}))^\top,$$
$$\text{where } \boldsymbol{M}_i(\boldsymbol{a}) = (M_{i,1}(\boldsymbol{a}), \ldots, M_{i,d}(\boldsymbol{a}))^\top, \text{ and}$$
$$M_{i,k}(\boldsymbol{a}) = \frac{\partial J_i(\boldsymbol{a})}{\partial a_k^i}, \quad \boldsymbol{a} \in \boldsymbol{A}, \quad i \in [N], \quad k \in [d], \tag{1}$$

is *monotone on* $\boldsymbol{A}$ (see Definition 1).

We consider a *Nash equilibrium* in game $\Gamma(N, \{A_i\}, \{J_i\})$ as a stable solution outcome because it represents a joint action from which no player has any incentive to unilaterally deviate.

Definition 2: A point $\boldsymbol{a}^* \in \boldsymbol{A}$ is called a *Nash equilibrium* if for any $i \in [N]$ and $\boldsymbol{a}^i \in A_i$

$$J_i(\boldsymbol{a}^{i*}, \boldsymbol{a}^{-i*}) \leq J_i(\boldsymbol{a}^i, \boldsymbol{a}^{-i*}).$$

Our goal is to learn such a stable action in a game through designing a payoff-based algorithm. To do so, we first connect existence of Nash equilibria for $\Gamma(N, \{A_i\}, \{J_i\})$ with solution set of a corresponding variational inequality problem.

Definition 3: Consider a mapping $\boldsymbol{T}(\cdot)$: $\mathbb{R}^d \to \mathbb{R}^d$ and a set $Y \subseteq \mathbb{R}^d$. The *solution set* $SOL(Y, \boldsymbol{T})$ *to the variational inequality problem* $VI(Y, \boldsymbol{T})$ is the set of vectors $\boldsymbol{y}^* \in Y$ such that $(\boldsymbol{T}(\boldsymbol{y}^*), \boldsymbol{y} - \boldsymbol{y}^*) \geq 0$, $\forall \boldsymbol{y} \in Y$.

Theorem 1: ([16, Proposition 1.4.2]) Given a game $\Gamma(N, \{A_i\}, \{J_i\})$ with game mapping $\boldsymbol{M}$, suppose that the action sets $\{A_i\}$ are closed and convex, the cost functions $\{J_i\}$ are continuously differentiable in $\boldsymbol{a}$ and convex in $\boldsymbol{a}^i$ for every fixed $\boldsymbol{a}^{-i}$ on the interior of $\boldsymbol{A}$. Then, some vector $\boldsymbol{a}^* \in \boldsymbol{A}$ is a Nash equilibrium in Γ, if and only if $\boldsymbol{a}^* \in SOL(\boldsymbol{A}, \boldsymbol{M})$.

It follows that under Assumptions 1 and 2 for a game with mapping $\boldsymbol{M}$, any solution of $VI(\boldsymbol{A}, \boldsymbol{M})$ is also a Nash equilibrium in such games and vice versa. While $\Gamma(N, \{A_i\}, \{J_i\})$ under Assumptions 1 and 2 might admit a Nash equilibrium, these two assumptions alone do not guarantee uniqueness of a Nash equilibrium. More restrictive assumptions for uniqueness are needed, for example, strong monotonicity of the game mapping or compactness of the action sets [16]. Here, we do not restrict our attention to such cases. However, to have a meaningful discussion, we do assume existence of at least one Nash equilibrium in the game.

Assumption 3: The set $SOL(\boldsymbol{A}, \boldsymbol{M})$ is not empty.

Corollary 1: Let $\Gamma(N, \{A_i\}, \{J_i\})$ be a game with game mapping $\boldsymbol{M}$ for which Assumptions 1, 2, and 3 hold. Then, there exists at least one Nash equilibrium in Γ. Moreover, any Nash equilibrium in Γ belongs to the set $SOL(\boldsymbol{A}, \boldsymbol{M})$.

The following additional assumptions are needed for convergence of the proposed payoff-based algorithm to a Nash equilibrium (see proofs of Lemma 5 and Theorem 2).

Assumption 4: Each element $\boldsymbol{M}_i$ of the game mapping $\boldsymbol{M} : \mathbb{R}^{Nd} \to \mathbb{R}^{Nd}$, defined in Assumption (2) is Lipschitz continuous on $\mathbb{R}^d$ with a Lipschitz constant L_i.

Assumption 5: Each cost function $J_i(\boldsymbol{a})$, $i \in [N]$ grows at most polynomially in $\boldsymbol{a}$ as $\|\boldsymbol{a}\| \to \infty$. Moreover, in the case of unbounded joint action set $\boldsymbol{A}$, each continuous cost function $J_i(\boldsymbol{a})$, $i \in [N]$ grows at most linearly in $\boldsymbol{a}$ as $\|\boldsymbol{a}\| \to \infty$.

Remark 1: Note that if the set $\boldsymbol{A}$ is unbounded, Assumption 5 is equivalent to each cost function $J_i(\boldsymbol{a})$, $i \in [N]$, being Lipschitz continuous on $\mathbb{R}^{Nd}$ with some constant l_i Thus, in both bounded and unbounded $\boldsymbol{A}$, we denote $l = \max_{i \in [N]} l_i$ as the uniform upper bound of the mapping $\boldsymbol{M}$ over $\boldsymbol{A}$.

For the development and analysis of our algorithms, we use the following well-established and easy to verify result.

Lemma 1: Consider a mapping $\boldsymbol{T}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ and a convex closed set $Y \subseteq \mathbb{R}^d$. Given $\theta > 0$,

$$\boldsymbol{y}^* \in \text{SOL}(Y, \boldsymbol{T}) \iff \boldsymbol{y}^* = \text{Proj}_Y(\boldsymbol{y}^* - \theta \boldsymbol{T}(\boldsymbol{y}^*)). \tag{2}$$

III. Payoff-Based Algorithm

Given online payoff-based information, also referred to as online bandit or zeroth-order oracle information, each agent has access to its current action, referred to as its state and denoted by $\mathbf{x}^i(t) = (x_1^i(t), \ldots, x_d^i(t))^\top \in \mathbb{R}^d$, and plays the action $\boldsymbol{a}^i(t) = \text{Proj}_{A_i}(\mathbf{x}^i(t))$ at iteration t. After that the cost value $\hat{J}_i(t)$ at the joint action $\boldsymbol{a}(t) = (\boldsymbol{a}^1(t), \ldots, \boldsymbol{a}^N(t)) \in \boldsymbol{A}$, $\hat{J}_i(t) = J_i(\boldsymbol{a}(t))$ is revealed to each agent i. Given these pieces of information, in the proposed algorithm each agent i "mixes" its next state $\mathbf{x}^i(t+1)$. Namely, it chooses $\mathbf{x}^i(t+1)$ randomly according to the multidimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}^i(t+1) = (\mu_1^i(t+1), \ldots, \mu_d^i(t+1))^\top, \sigma(t+1))$ with the following density function:

$$p_i(\boldsymbol{x}^i; \boldsymbol{\mu}^i(t+1), \sigma_{t+1}) = p_i(x_1^i, \ldots, x_d^i; \boldsymbol{\mu}^i(t+1), \sigma_{t+1})$$
$$= \frac{1}{(\sqrt{2\pi}\sigma_{t+1})^d} \exp\left\{ -\sum_{k=1}^d \frac{(x_k^i - \mu_k^i(t+1))^2}{2\sigma_{t+1}^2} \right\}.$$

The initial value of the means $\boldsymbol{\mu}^i(0)$, $i \in [N]$, can be set to any finite value. The successive means are updated as follows:

$$\boldsymbol{\mu}^i(t+1) = \text{Proj}_{(1-r_t)A_i}\big[\boldsymbol{\mu}^i(t)$$
$$- \gamma_t \sigma_t^2 \left(\hat{J}_i(t) \frac{\mathbf{x}^i(t) - \boldsymbol{\mu}^i(t)}{\sigma_t^2} + \epsilon_t \boldsymbol{\mu}^i(t) \right) \big]. \tag{3}$$

In the above, $(1 - r_t)A_i = \{\boldsymbol{x} \in A_i : \text{dist}(\boldsymbol{x}, \partial A_i) \geq r_t\}$ and $0 < r_t < 1$, is a time-dependent shrinkage parameter, γ_t is the stepsize parameter and $\epsilon_t > 0$ is a regularization parameter. The convergence of the algorithm is dependent on the interplay of these parameters and the variance term $\sigma_t > 0$.

The difference between the proposed approach and that of [18] is due to the additional term ϵ_t in (3). In the absence of this term the algorithm would converge only if the game mapping is *strictly* monotone (see [18, Theorem 2] and counterexamples in [7, 13]). Moreover, in distinction from [19], in the bandit online feedback considered here, players can only evaluate their costs over their feasible action set A_i and not over the whole $\mathbb{R}^{Nd}$, necessitating the additional projection term $\boldsymbol{a}^i(t) = \text{Proj}_{A_i}(\mathbf{x}^i(t))$ and the shrinkage radius r_t. As such, the previous convergence analysis does not apply.

Before stating the convergence result, let us provide insight into the procedure defined by Equation (3) by deriving an analogy to a regularized stochastic gradient algorithm.

Let $p(\boldsymbol{x};\boldsymbol{\mu},\sigma) = \prod_{i=1}^{N} p_i(x_1^i,\dots,x_d^i;\boldsymbol{\mu}^i,\sigma)$ denote the density function of the joint distribution of agents' states. Given $\sigma > 0$, for any $i \in [N]$ define $\tilde{J}_i : \mathbb{R}^{Nd} \to \mathbb{R}$ as

$$\tilde{J}_i(\boldsymbol{\mu}^1,\dots,\boldsymbol{\mu}^N,\sigma) = \int_{\mathbb{R}^{Nd}} J_i(\boldsymbol{x}) p(\boldsymbol{x};\boldsymbol{\mu},\sigma) d\boldsymbol{x}. \tag{4}$$

Thus, $\tilde{J}_i$, $i \in [N]$ is the ith player's cost function in mixed strategies. Let $\boldsymbol{\mu}(t) = (\boldsymbol{\mu}^1(t),\dots,\boldsymbol{\mu}^N(t))$ and for $i \in [N]$, define $\tilde{\boldsymbol{M}}_i(\cdot) = (\tilde{M}_{i,1}(\cdot),\dots,\tilde{M}_{i,d}(\cdot))^\top$ as the d-dimensional mapping with the following elements:

$$\tilde{M}_{i,k}(\boldsymbol{\mu},\sigma) = \frac{\partial \tilde{J}_i(\boldsymbol{\mu},\sigma)}{\partial \mu_k^i}, \text{ for } k \in [d]. \tag{5}$$

Our first lemma below shows that the second term inside the projection in (3) is a sample of the gradient of agent i's cost function in mixed strategies.

Lemma 2: Under Assumptions 1 and 5

$$\tilde{M}_{i,k}(\boldsymbol{\mu}(t),\sigma_t) = \mathrm{E}\{J_i(\mathbf{x}^1(t),\dots,\mathbf{x}^N(t))\frac{x_k^i(t)-\mu_k^i(t)}{\sigma_t^2}| \\ x_k^i(t) \sim \mathcal{N}(\mu_k^i(t),\sigma_t), i \in [N], k \in [d]\}. \tag{6}$$

Moreover, $\tilde{M}_{i,k}(\boldsymbol{\mu},\sigma)$ is bounded for any $\boldsymbol{\mu} \in \boldsymbol{A}$.
The proof of this Lemma is very similar to that of [19] and is provided in Appendix B.

The lemma above implies that had we used the term $J_i(\mathbf{x}^1(t),\dots,\mathbf{x}^N(t))\frac{x_k^i(t)-\mu_k^i(t)}{\sigma_t^2}$ in (3) we could perform a one-point estimation of the gradient of the cost functions in mixed strategies. In the bandit setting considered here, however, we use the term $J_i(\boldsymbol{a}^1(t),\dots,\boldsymbol{a}^N(t))\frac{x_k^i(t)-\mu_k^i(t)}{\sigma_t^2}$. Despite this difference, in the analysis (see (22), (23)) we show that the difference between these two terms converges to zero due to the shrinkage radius selection. Thus, algorithm (3) can be interpreted as a doubly regularized (due to r_r and ϵ_t) stochastic projection algorithms.

Following the above interpretation, our main result is Theorem 2 below where we show that by appropriately choosing the algorithm parameters, we can bound the bias and variance terms of the stochastic projection and consequently establish convergence of the iterates $\boldsymbol{\mu}(t)$ in (3) to a Nash equilibrium.

Assumption 6: Let $\beta_t = \gamma_t\sigma_t^2$ and choose $\gamma_t = \frac{1}{t^a}$, $\sigma_t = \frac{1}{t^b}$, $\epsilon_t = \frac{1}{t^c}$ and $r_t = \frac{1}{t^d}$ $a,b,c,d > 0$ respectively, such that
i) $\sum_{t=0}^{\infty}\beta_t = \infty$, $\sum_{t=0}^{\infty}\beta_t\epsilon_t = \infty$,
ii) $\sum_{t=0}^{\infty}\frac{(\epsilon_t-\epsilon_{t-1})^2}{\beta_t\epsilon_t^3} + \frac{(r_t-r_{t-1})^2}{\beta_t\epsilon_t^6} < \infty$,
iii) $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$, $\sum_{t=0}^{\infty}\beta_t\sigma_t < \infty$,
iv) $\lim_{t\to\infty}\frac{r_t}{\sigma_t} = \infty$, $\lim_{t\to\infty}\frac{r_t}{\epsilon_t} = 0$.

As an example for existence of parameters to satisfy Assumption 6, let $a_1 = \frac{5}{9}$, $a_2 = \frac{5}{27}$, $a_3 = \frac{1}{54}$, $a_4 = \frac{1}{6}$.

Theorem 2: Let the players in game $\Gamma(N,\{A_i\},\{J_i\})$ choose the states $\{\mathbf{x}^i(t)\}$ at time t according to the normal distribution $\mathcal{N}(\boldsymbol{\mu}^i(t),\sigma_t)$, where the mean $\boldsymbol{\mu}^i(0)$ is arbitrary and $\boldsymbol{\mu}^i(t)$ is updated as in (3). Under Assumptions 1-6, as $t \to \infty$, the mean vector $\boldsymbol{\mu}(t)$ converges almost surely to a Nash equilibrium $\boldsymbol{\mu}^* = \boldsymbol{a}^*$ of the game Γ and the joint action $\boldsymbol{a}(t)$ converges in probability to $\boldsymbol{a}^*$.

IV. Analysis of the Algorithm

To prove Theorem 2 we need to first establish boundedness of the iterates $\boldsymbol{\mu}(t)$ for the cases in which the action space is unbounded. Having established the boundedness, we can show that the limit of the iterates $\boldsymbol{\mu}(t)$ exists and is the minimum norm Nash equilibrium of the problem. This convergence is proven using existing results on convergence of a sequence of random variables (Lemma 10 on page 49 in [17]). For ease of reference, we provide the supporting statements used for boundedness [15, Theorem 2.5.2] and for convergence [17, Lemma 10 on page 49] of the iterates in Appendix A.

A. Characterizing the terms in the algorithm

We first show that algorithm (3) can be interpreted within the framework of well-studied Robbins-Monro stochastic approximations procedures [3], where the iterates are updated according to stochastic gradient descent. In our case, bias of the game mapping arises due to each player's one-point estimation of its gradient. However, in contrast to a stochastic approximation procedure, the game mapping is in general not gradient of a single function (as its derivative is not symmetric) unless the game is potential. Furthermore, there are additional terms in the algorithm iterates due to the projection of the query points onto the shrunk feasible set and the regularization required to handle the non-strictly monotone game mapping. Let us specify all these terms below.

Using the notation $\boldsymbol{M}_i(\cdot) = (M_{i,1}(\cdot),\dots,M_{i,d}(\cdot))$, we can rewrite the algorithm step in (3) in the following form:

$$\boldsymbol{\mu}^i(t+1) = \mathrm{Proj}_{(1-r_t)A_i}[\boldsymbol{\mu}^i(t) - \gamma_t\sigma_t^2 \\ \times \big(\boldsymbol{M}_i(\boldsymbol{\mu}(t)) + \boldsymbol{Q}_i(\boldsymbol{\mu}(t),\sigma_t) + \boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) \\ + \boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) + \epsilon_t)\boldsymbol{\mu}^i(t)\big)], \tag{7}$$

for all $i \in [N]$, where $\boldsymbol{Q}_i$, $\boldsymbol{R}_i$, $\boldsymbol{P}_i$ are

$$\boldsymbol{Q}_i(\boldsymbol{\mu}(t),\sigma_t) = \tilde{\boldsymbol{M}}_i(\boldsymbol{\mu}(t),\sigma_t) - \boldsymbol{M}_i(\boldsymbol{\mu}(t)),$$
$$\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) = \boldsymbol{F}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) - \tilde{\boldsymbol{M}}_i(\boldsymbol{\mu}(t),\sigma_t),$$
$$\boldsymbol{F}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) = J_i(\mathbf{x}(t))\frac{\mathbf{x}^i(t)-\boldsymbol{\mu}^i(t)}{\sigma_t^2},$$
$$\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) = \frac{\mathbf{x}^i(t)-\boldsymbol{\mu}^i(t)}{\sigma_t^2}(J_i(\boldsymbol{a}(t)) - J_i(\mathbf{x}(t))).$$

Above, $\boldsymbol{M}(\boldsymbol{\mu}(t)) = (\boldsymbol{M}_1(\boldsymbol{\mu}(t)),\dots,\boldsymbol{M}_N(\boldsymbol{\mu}(t)))$ corresponds to the gradient term in stochastic approximation procedures. The mapping $\tilde{\boldsymbol{M}}_i(\boldsymbol{\mu}(t))$ evaluated at $\boldsymbol{\mu}(t)$ is equivalent to the game mapping in mixed strategies [19]. That is,

$$\tilde{\boldsymbol{M}}_i(\boldsymbol{\mu}(t)) = \int_{\mathbb{R}^{Nd}} \boldsymbol{M}_i(\boldsymbol{x}) p(\boldsymbol{x};\boldsymbol{\mu}(t),\sigma_t) d\boldsymbol{x}. \tag{8}$$

Thus, the $\boldsymbol{Q}(\boldsymbol{\mu}(t),\sigma_t) = (\boldsymbol{Q}_1(\boldsymbol{\mu}(t),\sigma_t),\dots,\boldsymbol{Q}_N(\boldsymbol{\mu}(t),\sigma_t))$ can be interpreted as disturbance of the gradient. Furthermore, according to (6), we have[2]

$$\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) = \boldsymbol{F}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) \\ - \mathrm{E}_{\mathbf{x}(t)}\{\boldsymbol{F}_i(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t)\}, \quad i \in [N]. \tag{9}$$

[2]The notation $\mathrm{E}_{\mathbf{x}(t)}\{\cdot\}$ further is used to emphasize that the expectation is taken in respect to $\mathbf{x}(t)$ which has the normal distribution with the mean $\boldsymbol{\mu}(t)$ and the covariance matrix $\sigma(t)I$, where I is the identity matrix.

Thus, $\boldsymbol{R}$ below is a martingale difference

$$\boldsymbol{R}(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) = (\boldsymbol{R}_1(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t),\ldots, \boldsymbol{R}_N(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t)).$$

Finally, due to the projection of query points, the term

$$\boldsymbol{P}(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t) = (\boldsymbol{P}_1(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t),\ldots, \boldsymbol{P}_N(\mathbf{x}(t),\boldsymbol{\mu}(t),\sigma_t)),$$

is the vector of the difference between the gradient estimation based on the state $\mathbf{x}(t) \in \mathbb{R}^{Nd}$ and the played action $\boldsymbol{a}(t) \in \boldsymbol{A}$.

Our goal is to bound each of the terms above to ensure convergence and boundedness of the iterates. However, first, we need to account for the regularization terms ϵ_t, r_t.

B. Analyzing the modified Tikhonov sequence

In contrast to stochastic approximation algorithms and the proof in [18], we have an addition term $\epsilon_t\boldsymbol{\mu}(t)$ to be able to address merely monotone game mappings. As such, to bound $\boldsymbol{\mu}(t)$ we also relate the variations of the sequence $\boldsymbol{\mu}(t)$ to those of the *modified Tikhonov sequence* defined below. Let $\boldsymbol{y}(t) = (\boldsymbol{y}^1(t),\ldots,\boldsymbol{y}^N(t))$ denote the solution of the variational inequality $VI((1-r_t)\boldsymbol{A}, \boldsymbol{M}(\boldsymbol{y}) + \epsilon_t\boldsymbol{y})$, namely

$$\boldsymbol{y}(t) \in SOL((1-r_t)\boldsymbol{A}, \boldsymbol{M}(\boldsymbol{y}) + \epsilon_t\boldsymbol{y}). \tag{10}$$

The Tikhonov sequence corresponds to the solution of the variational inequality above with the $r_t = 0$. Thus, the sequence $\{\boldsymbol{y}(t)\}$ can be considered the *modified* Tikhonov sequence. Similar to the Tikhonov sequence, $\boldsymbol{y}(t)$ enjoys the following important property.

Theorem 3: Under Assumptions 2, 3, and 4, $\boldsymbol{y}(t)$ defined in (10) exists and is unique for each t. Moreover, for $\epsilon_t \downarrow 0$ and $r_t \to 0$ and given $\lim_{t\to\infty} \frac{r_t}{\epsilon_t} = 0$, $\boldsymbol{y}(t)$ is uniformly bounded and converges to the least norm solution of $VI(\boldsymbol{A}, \boldsymbol{M})$.

The significance of the above theorem is that if we can establish $\boldsymbol{\mu}(t)$ converges to $\boldsymbol{y}(t)$, then from 1 we can establish convergence to a Nash equilibrium for the game. To prove Theorem 3, first we establish some useful properties of projecting onto sets $(1-r_t)\boldsymbol{A}$, $t = 1, 2, \ldots$.

Lemma 3: For any $\mathbf{x} \in \mathbb{R}^{Nd}$ the following holds:

$$\|\text{Proj}_{(1-r_{t-1})\boldsymbol{A}}\mathbf{x} - \text{Proj}_{(1-r_t)\boldsymbol{A}}\mathbf{x}\| = O(|r_{t-1} - r_t|).$$

Please see the Appendix C for the proof of above Lemma.

Proof: (of Theorem 3) Let $\boldsymbol{a}$ be the least norm solution of $VI(\boldsymbol{A}, \boldsymbol{M})$. Moreover, let $\boldsymbol{a}^p$ be the projection of $\boldsymbol{a}$ onto the set $(1-r_t)\boldsymbol{A}$. Next, let $\boldsymbol{y}(t)$ be the unique solution of the doubly regularized inequality, namely $\boldsymbol{y}(t) \in SOL((1-r_t)\boldsymbol{A}, \boldsymbol{M} + \epsilon_t I)$. Thus, we conclude that

$$(\boldsymbol{M}(\boldsymbol{a}), \boldsymbol{y}(t) - \boldsymbol{a}) \geq 0,$$

$$(\boldsymbol{M}(\boldsymbol{y}(t)) + \epsilon_t\boldsymbol{y}(t), \boldsymbol{a}^p - \boldsymbol{y}(t)) \geq 0.$$

Thus, taking into account monotonicity of $\boldsymbol{M}$, we obtain

$$\begin{aligned}
0 \leq& (\boldsymbol{M}(\boldsymbol{a}), \boldsymbol{y}(t) - \boldsymbol{a}) \\
&+ (\boldsymbol{M}(\boldsymbol{y}(t)) + \epsilon_t\boldsymbol{y}(t), \boldsymbol{a} - \boldsymbol{y}(t)) \\
&+ (\boldsymbol{M}(\boldsymbol{y}(t)) + \epsilon_t\boldsymbol{y}(t), \boldsymbol{a}^p - \boldsymbol{a}) \\
=& -(\boldsymbol{M}(\boldsymbol{a}) - \boldsymbol{M}(\boldsymbol{y}(t)), \boldsymbol{a} - \boldsymbol{y}(t)) + \epsilon_t(\boldsymbol{y}(t), \boldsymbol{a} - \boldsymbol{y}(t)) \\
&+ (\boldsymbol{M}(\boldsymbol{y}(t)) + \epsilon_t\boldsymbol{y}(t), \boldsymbol{a}^p - \boldsymbol{a}) \\
\leq& \epsilon_t(\boldsymbol{y}(t), \boldsymbol{a}) - \epsilon_t\|\boldsymbol{y}(t)\|^2 \\
&+ (\boldsymbol{M}(\boldsymbol{y}(t)), \boldsymbol{a}^p - \boldsymbol{a}) + \epsilon_t(\boldsymbol{y}(t), \boldsymbol{a}^p - \boldsymbol{a}).
\end{aligned}$$

Hence,

$$\begin{aligned}
\epsilon_t\|\boldsymbol{y}(t)\|^2 &\leq \epsilon_t(\boldsymbol{y}(t), \boldsymbol{a}) + (\boldsymbol{M}(\boldsymbol{y}(t)), \boldsymbol{a}^p - \boldsymbol{a}) \\
&\quad + \epsilon_t(\boldsymbol{y}(t), \boldsymbol{a}^p - \boldsymbol{a}) \\
&\leq \epsilon_t\|\boldsymbol{y}(t)\|\|\boldsymbol{a}\| + l\|\boldsymbol{a}^p - \boldsymbol{a}\| + \epsilon_t\|\boldsymbol{y}(t)\|\|\boldsymbol{a}^p - \boldsymbol{a}\| \\
&= \epsilon_t\|\boldsymbol{y}(t)\|\|\boldsymbol{a}\| + lO(r_t) + \epsilon_t\|\boldsymbol{y}(t)\|O(r_t),
\end{aligned}$$

where in the first inequality we used Remark 1 and in the second one we applied Lemma 3. Hence,

$$\|\boldsymbol{y}(t)\|^2 \leq \|\boldsymbol{y}(t)\|\|\boldsymbol{a}\| + lO\left(\frac{r_t}{\epsilon_t}\right) + \|\boldsymbol{y}(t)\|O(r_t).$$

By taking the upper limit $t \to \infty$ in the inequality above and due to the settings for ϵ_t and r_t, we obtain

$$\begin{aligned}
\overline{\lim}_{t\to\infty}[\|\boldsymbol{y}(t)\|^2] \leq& \|\boldsymbol{a}\|\overline{\lim}_{t\to\infty}\|\boldsymbol{y}(t)\| \\
&+ \overline{\lim}_{t\to\infty}\|\boldsymbol{y}(t)\|\overline{\lim}_{t\to\infty}O(r_t).
\end{aligned}$$

It implies that $\overline{\lim}_{t\to\infty}\|\boldsymbol{y}(t)\| \leq \|\boldsymbol{a}\|$, and, thus, the sequence $\|\boldsymbol{y}(t)\|$ is upper bounded. Moreover, any accumulation point of this sequence is bounded above by the Euclidean norm of the least-norm solution of $VI(\boldsymbol{A}, \boldsymbol{M})$. Hence, according to the fact that the function $\text{Proj}_{(1-r)\boldsymbol{A}}(\mathbf{x})$ is continuous in both r and $\mathbf{x}$, $\boldsymbol{y}(t)$ converges to the least norm solution of $VI(\boldsymbol{A}, \boldsymbol{M})$. ■

Since our goal now is to relate $\boldsymbol{\mu}(t)$ to $\boldsymbol{y}(t)$, aligned with procedure (3), we will now design a one-time scale approach to solving (10) as per (11) below:

$$\mathbf{z}(t+1) = \text{Proj}_{(1-r_t)\boldsymbol{A}}[\mathbf{z}(t) - \beta_t(\boldsymbol{M}(\mathbf{z}(t)) + \epsilon_t\mathbf{z}(t))], \tag{11}$$

where β_t is defined in Assumption 6. We show that the procedure above is a one time-scale algorithm and similarly to $\boldsymbol{y}(t)$, it converges to the least norm solution of $VI(\boldsymbol{A}, \boldsymbol{M})$.

Proposition 1: The sequence $\mathbf{z}(t)$ defined by (11) converges to the least norm solution of $VI(\boldsymbol{A}, \boldsymbol{M})$.

To prove the result above, we bound $\|\mathbf{z}(t+1) - \boldsymbol{y}(t)\|$ in terms of the previous terms in the sequence, namely, $\|\mathbf{z}(t) - \boldsymbol{y}(t-1)\|$ and show that [17, Lemma 10, page 49] on convergence of a random sequence applies. To do so though, first we need to bound the variations of $\boldsymbol{y}(t)$ as below.

Lemma 4: Under Assumptions 2, 4, and 6, the Tikhonov sequence $\boldsymbol{y}(t)$ defined in (10) satisfies

$$\begin{aligned}
&\|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2 = \\
&O\left(\frac{(\epsilon_t - \epsilon_{t-1})^2}{\epsilon_t^2} + \frac{(r_t - r_{t-1})^2}{\epsilon_t^5}\right).
\end{aligned}$$

Please see Appendix D for the proof.

In summary, the results of this section enabled us to prove Proposition 1. This proposition serves as the main new result in comparison to non-regularized stochastic gradient procedures in order to show almost-sure boundedness of $\|\boldsymbol{\mu}(t)\|$ and the convergence of the algorithm to a Nash equilibrium.

C. Boundedness of the iterates

Lemma 5: Let Assumptions 2-6 hold in $\Gamma(N,\{A_i\},\{J_i\})$ and $\boldsymbol{\mu}(t)$ be the vector updated in the run of the payoff-based algorithm (3). Then, $\Pr\{\sup_{t\geq 0}\|\boldsymbol{\mu}(t)\|<\infty\}=1$.

Proof: If the set $\boldsymbol{A}$ is compact, then, according to the updated in (7), the norm of the vector $\boldsymbol{\mu}_t$ is bounded for all t. So, let us consider the case of the unbounded $\boldsymbol{A}$.

Define $V(t,\boldsymbol{\mu})=\|\boldsymbol{\mu}-\mathbf{z}(t)\|^2$, where $\mathbf{z}(t)$ is given in (10). Consider the generating operator of the Markov process $\boldsymbol{\mu}(t)$

$$LV(t,\boldsymbol{\mu})=E[V(t+1,\boldsymbol{\mu}(t+1))\mid\boldsymbol{\mu}(t)=\boldsymbol{\mu}]-V(t,\boldsymbol{\mu}).$$

We aim to show that $LV(t,\boldsymbol{\mu})$ satisfies the following decay

$$LV(t,\boldsymbol{\mu})\leq-\alpha(t+1)\psi(\boldsymbol{\mu})+\phi(t)(1+V(t,\boldsymbol{\mu})),\qquad(12)$$

where ψ, ϕ and α are terms arising from $\boldsymbol{Q}$, $\boldsymbol{R}$ and $\boldsymbol{P}$ in (7). Our goal is to show that $\psi\geq 0$ on $\mathbb{R}^{Nd}$, $\phi(t)>0$, $\forall t$, $\sum_{t=0}^{\infty}\phi(t)<\infty$, $\alpha(t)>0$, $\sum_{t=0}^{\infty}\alpha(t)=\infty$. This combined with the boundedness of the iterates $\mathbf{z}(t)$ stated in Proposition 1 enable us to apply Theorem 2.5.2 in [15] to conclude almost sure boundedness of $\boldsymbol{\mu}(t)$.

From now on, for simplicity in notation, we omit the argument σ_t in the terms $\tilde{\boldsymbol{M}}$, $\boldsymbol{Q}$, and $\boldsymbol{R}$. In certain derivations, for the same reason we omit the time parameter t as well.

Let us analyze each term $i=1,\dots,N$ in

$$V(t+1,\boldsymbol{\mu}(t+1))=\sum_{i=1}^{N}\|\boldsymbol{\mu}^i(t+1)-\mathbf{z}^i(t+1)\|^2.$$

From the procedures for the update of $\boldsymbol{\mu}(t)$ and $\mathbf{z}(t)$ and the non-expansion property of the projection operator, we obtain

$$\begin{aligned}
&\|\boldsymbol{\mu}^i(t+1)-\mathbf{z}^i(t+1)\|^2\\
&\leq\|\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t)-\beta_t\big[\epsilon_t(\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))\\
&+(\boldsymbol{M}_i(\boldsymbol{\mu}(t))-\boldsymbol{M}_i(\mathbf{z}(t))+\boldsymbol{Q}_i(\boldsymbol{\mu}(t))+\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\\
&\qquad+\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\big]\|^2\\
&=\|\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t)\|^2\\
&\quad-2\beta_t(\boldsymbol{M}_i(\boldsymbol{\mu}(t))-\boldsymbol{M}_i(\mathbf{z}(t)),\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))\\
&\quad-2\beta_t\epsilon_t(\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t),\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))\\
&\quad-2\beta_t(\boldsymbol{Q}_i(\boldsymbol{\mu}(t))+\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)),\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))\\
&\quad-2\beta_t(\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)),\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))\\
&\quad+\beta_t^2\|\boldsymbol{G}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2,
\end{aligned}\qquad(13)$$

where, for ease of notation, we have defined

$$\begin{aligned}
\boldsymbol{G}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))=&\epsilon_t(\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))\\
&+\boldsymbol{M}_i(\boldsymbol{\mu}(t))-\boldsymbol{M}_i(\mathbf{z}(t))\\
&+\boldsymbol{Q}_i(\boldsymbol{\mu}(t))+\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\\
&+\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)).
\end{aligned}\qquad(14)$$

Note that the terms in $\|\boldsymbol{G}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2$ are given as

$$\begin{aligned}
&\|\boldsymbol{G}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2=\epsilon^2(t)\|\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t)\|^2\\
&+\|\boldsymbol{M}_i(\boldsymbol{\mu}(t))-\boldsymbol{M}_i(\mathbf{z}(t))\|^2\\
&+\|\boldsymbol{Q}_i(\boldsymbol{\mu}(t))\|^2+\|\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2+\|\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2\\
&+2(\boldsymbol{Q}_i(\boldsymbol{\mu}(t)),\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)))\\
&+2(\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)),\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)))\\
&+2(\boldsymbol{Q}_i(\boldsymbol{\mu}(t)),\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)))\\
&+2\epsilon_t(\boldsymbol{M}_i(\boldsymbol{\mu}(t))-\boldsymbol{M}_i(\mathbf{z}(t)),\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))\\
&+2(\epsilon_t(\boldsymbol{\mu}^i(t)-\mathbf{z}^i(t))+\boldsymbol{M}_i(\boldsymbol{\mu}(t))-\boldsymbol{M}_i(\mathbf{z}(t)),\\
&\qquad\boldsymbol{Q}_i(\boldsymbol{\mu}(t))+\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))+\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))),
\end{aligned}\qquad(15)$$

Thus, accounting for the above, for (9), which implies $\mathrm{E}\{\|\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|\|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}=0$ for any $\boldsymbol{\mu}$, and for the Cauchy-Schwarz inequality, we get from (13)

$$\begin{aligned}
&\mathrm{E}\{\|\boldsymbol{\mu}^i(t+1)-\mathbf{z}^i(t+1)\|^2|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\
&\leq(1-2\beta_t\epsilon_t)\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|^2\\
&\quad-2\beta_t(\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t)),\boldsymbol{\mu}^i-\mathbf{z}^i(t))\\
&\quad-2\beta_t(\boldsymbol{Q}_i(\boldsymbol{\mu}),\boldsymbol{\mu}^i-\mathbf{z}^i(t))\\
&\quad+2\beta_t\mathrm{E}\{\|\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|\|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|\\
&\quad+\beta_t^2\mathrm{E}\{\|\boldsymbol{G}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\
&\leq(1-2\beta_t\epsilon_t)\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|^2\\
&\quad-2\beta_t(\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t)),\boldsymbol{\mu}^i-\mathbf{z}^i(t))\\
&\quad+2\beta_t\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|\\
&\quad+2\beta_t\mathrm{E}\{\|\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|\|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|\\
&+\beta_t^2\epsilon^2(t)\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|^2\\
&+\beta_t^2[\|\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t))\|^2\\
&\quad+\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|^2\\
&\quad+\mathrm{E}\{\|\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2+\|\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\
&\quad+2\mathrm{E}\{(\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)),\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)))|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\
&\quad+2\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|\mathrm{E}\{\|\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)))\|\|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\
&\quad+2(\epsilon_t\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|+\|\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t))\|)\\
&\quad\times(\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|+\mathrm{E}\{\|\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|\|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\})].
\end{aligned}\qquad(16)$$

We proceed estimating the terms in the inequality above. Due to Assumption 4, we conclude that

$$\begin{aligned}
&\|\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t))\|^2\leq L_i^2\|\boldsymbol{\mu}-\mathbf{z}(t)\|^2=O(V(t,\boldsymbol{\mu}))\\
&(\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t)),\boldsymbol{\mu}^i-\mathbf{z}^i(t))\\
&\leq\|\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t))\|\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|\\
&\leq L_i\|\boldsymbol{\mu}-\mathbf{z}(t)\|\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\|=O(V(t,\boldsymbol{\mu})).
\end{aligned}\qquad(17)$$

Let us analyze the terms containing the disturbance of gradient, namely $\boldsymbol{Q}_i$, in Equation (15). Since $\boldsymbol{Q}_i(\boldsymbol{\mu}(t))=\tilde{\boldsymbol{M}}_i(\boldsymbol{\mu}(t))-\boldsymbol{M}_i(\boldsymbol{\mu}(t))$, due to Assumption 2 and Equation (8), we obtain

$$\begin{aligned}
\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|&=\|\int_{\mathbb{R}^{Nd}}[\boldsymbol{M}_i(\boldsymbol{x})-\boldsymbol{M}_i(\boldsymbol{\mu})]p(\boldsymbol{x};\boldsymbol{\mu},\sigma_t)d\boldsymbol{x}\|\\
&\leq\int_{\mathbb{R}^{Nd}}\|\boldsymbol{M}_i(\boldsymbol{x})-\boldsymbol{M}_i(\boldsymbol{\mu})\|p(\boldsymbol{x};\boldsymbol{\mu},\sigma_t)d\boldsymbol{x}\\
&\leq\int_{\mathbb{R}^{Nd}}L_i\|\boldsymbol{x}-\boldsymbol{\mu}\|p(\boldsymbol{x};\boldsymbol{\mu},\sigma_t)d\boldsymbol{x}\\
&\leq\int_{\mathbb{R}^{Nd}}L_i\left(\sum_{i=1}^{N}\sum_{k=1}^{d}|x_k^i-\mu_k^i|\right)p(\boldsymbol{x};\boldsymbol{\mu},\sigma_t)d\boldsymbol{x}\\
&=O(\sigma_t),
\end{aligned}\qquad(18)$$

where the last equality is due to the fact that the first central absolute moment of a random variable with a normal

distribution $\mathcal{N}(\mu,\sigma)$ is $O(\sigma)$. The estimation above imply, in particular, that for any $\boldsymbol{\mu}\in\boldsymbol{A}$

$$\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|\|\boldsymbol{\mu}^i-\mathbf{z}^i(t)\| = O(\sigma_t)(1+V(t,\boldsymbol{\mu})) \tag{19}$$

$$\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|\|M_i(\boldsymbol{\mu})-M_i(\mathbf{z}(t))\| \le L_i\|\boldsymbol{Q}_i(\boldsymbol{\mu})\|\|\boldsymbol{\mu}-\mathbf{z}(t)\| = O(\sigma_t)(1+V(t,\boldsymbol{\mu})). \tag{20}$$

We bound the martingale term $\|\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2$.

$$\begin{aligned}&\mathrm{E}\{\|\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\ &\le\sum_{k=1}^{d}\int_{\mathbb{R}^{Nd}} {J_i}^2(\boldsymbol{x})\frac{(x_k^i-\mu_k^i)^2}{\sigma^4(t)}p(\boldsymbol{\mu},\boldsymbol{x})d\boldsymbol{x}\\ &\le\frac{f_i(\boldsymbol{\mu},\sigma_t)}{\sigma^4(t)}=\frac{O(1+V(t,\boldsymbol{\mu}))}{\sigma^4(t)},\end{aligned} \tag{21}$$

where the first inequality is due to the fact that $\mathrm{E}(\xi-\mathrm{E}\xi)^2\le\mathrm{E}\xi^2$ and taking into account (9), the second inequality is due to Assumption 5, with $f_i(\boldsymbol{\mu},\sigma_t)$ being a quadratic function of $\boldsymbol{\mu}$ and σ_t, $i\in[N]$ (see Appendix E for more details).

We proceed estimating the terms containing $P_t(\mathbf{x}_t,\boldsymbol{\mu})$. For any $\boldsymbol{\mu}\in\boldsymbol{A}$ we have

$$\begin{aligned}&\mathrm{E}\{\|\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t))\|^2|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\ &=\mathrm{E}\frac{\|\mathbf{x}^i(t)-\boldsymbol{\mu}^i\|^2|J_i(\mathbf{x}(t))-J_i(\boldsymbol{a}(t))|^2}{\sigma_t^4}\\ &=\Pr\{\mathbf{x}(t)\in\mathbb{R}^{Nd}\setminus\boldsymbol{A}\}\mathrm{E}\frac{\|\mathbf{x}^i(t)-\boldsymbol{\mu}^i\|^2|J_i(\mathbf{x}(t))-J_i(\boldsymbol{a}(t))|^2}{\sigma_t^4}\\ &\le\Pr\{\mathbf{x}(t)\in\mathbb{R}^{Nd}\setminus\boldsymbol{A}\}\mathrm{E}l^2\frac{\|\mathbf{x}^i(t)-\boldsymbol{\mu}^i\|^2\|\mathbf{x}(t)-\boldsymbol{\mu}\|^2}{\sigma_t^4}\\ &=k_0\Pr\{\mathbf{x}(t)\in\mathbb{R}^{Nd}\setminus\boldsymbol{A}\},\text{ for some }k_0>0,\end{aligned} \tag{22}$$

where the inequality is due to Assumption (5) implying $|J_i(\mathbf{x}(t))-J_i(\boldsymbol{a}(t))|^2\le l^2\|\mathbf{x}(t)-\boldsymbol{a}(t)\|^2$, and furthermore because $\|\mathbf{x}(t)-\boldsymbol{a}(t)\|^2\le\|\mathbf{x}(t)-\boldsymbol{\mu}\|^2$ for $\boldsymbol{a}(t)=\mathrm{Proj}_{\boldsymbol{A}}\mathbf{x}(t)$.

Next, let us estimate $\Pr\{\mathbf{x}(t)\in\mathbb{R}^{Nd}\setminus\boldsymbol{A}\}$. The idea is that due to the fact that $\mathbf{x}(t)$ is sampled from a Gaussian distribution with mean $\boldsymbol{\mu}(t)$, $\mathbf{x}(t)$ concentrates around its mean $\boldsymbol{\mu}(t)$ with high probability. Since the mean is projected onto a shrunk version of the set $\boldsymbol{A}$, namely, $(1-r_t)\boldsymbol{A}$, by appropriately tuning r_t and σ_t we can ensure that $\mathbf{x}(t)$ stays within the original feasible set with high probability. Let $\mathcal{O}_{r_t}(\boldsymbol{\mu})=\{\boldsymbol{y}\in\mathbb{R}^{Nd}|\|\boldsymbol{y}-\boldsymbol{\mu}\|^2<r_t^2\}$ denote the r_t-neighborhood of the point $\boldsymbol{\mu}$. Hence, $\sup_{\boldsymbol{y}\notin\mathcal{O}_{r_t}(\boldsymbol{\mu})}-\|\boldsymbol{y}-\boldsymbol{\mu}\|^2=-r_t^2$. Then, taking into account the fact that $\mathcal{O}_{r_t}(\boldsymbol{\mu})$ is contained in $\boldsymbol{A}$ and $r_t<1$, we obtain that for any t and any bounded $\sigma>\sigma_t$:

$$\begin{aligned}&\Pr\{\mathbf{x}(t)\in\mathbb{R}^{Nd}\setminus\boldsymbol{A}\}\le\Pr\{\mathbf{x}(t)\in\mathbb{R}^{Nd}\setminus\mathcal{O}_{r_t}(\boldsymbol{\mu})\}\\ &=\int_{\boldsymbol{y}\notin\mathcal{O}_{r_t}(\boldsymbol{\mu})}\frac{1}{(2\pi)^{Nd/2}\sigma_t^{Nd}}\exp\left\{-\frac{\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{2\sigma_t^2}\right\}d\boldsymbol{y}\\ &=\int_{\boldsymbol{y}\notin\mathcal{O}_{r_t}(\boldsymbol{\mu})}\exp\left\{-\|\boldsymbol{y}-\boldsymbol{\mu}\|^2\left(\frac{1}{2\sigma_t^2}-\frac{1}{2\sigma^2}\right)\right\}\\ &\qquad\times\frac{\sigma^{Nd}}{\sigma_t^{Nd}}\frac{1}{(2\pi)^{Nd/2}\sigma^{Nd}}\exp\left\{-\frac{\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right\}d\boldsymbol{y}\\ &\le\exp\left\{-r_t\left(\frac{1}{2\sigma_t^2}-\frac{1}{2\sigma^2}\right)\right\}\frac{\sigma^{Nd}}{\sigma_t^{Nd}}\\ &\quad\times\int_{\boldsymbol{y}\notin\mathcal{O}_{r_t}(\boldsymbol{\mu})}\frac{1}{(2\pi)^{Nd/2}\sigma^{Nd}}\exp\left\{-\frac{\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right\}d\boldsymbol{y}\\ &\le k_2\frac{e^{-\frac{r_t^2}{2\sigma_t^2}}}{\sigma_t^{Nd}}\end{aligned} \tag{23}$$

for some finite $k_2>0$. The last inequality holds because

$$\int_{\boldsymbol{y}\notin\mathcal{O}_{r_t}(\boldsymbol{\mu})}\frac{1}{(2\pi)^{Nd/2}\sigma^{Nd}}\exp\left\{-\frac{\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right\}d\boldsymbol{y}\le 1$$

and, thus, there exists $0<k_3<\infty$:

$$\int_{\boldsymbol{y}\notin\mathcal{O}_{r_t}(\boldsymbol{\mu})}\frac{e^{\frac{r_t^2}{2\sigma^2}}\sigma^n}{(2\pi)^{n/2}\sigma^{Nd}}\exp\left\{-\frac{\|\boldsymbol{y}-\boldsymbol{\mu}\|^2}{2\sigma^2}\right\}d\boldsymbol{y}\le k_3.$$

From (16) it now remains to bound the term

$$\mathrm{E}\{(\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)),\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)))|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}.$$

According to definitions of $\boldsymbol{P}_i$ and $\boldsymbol{R}_i$, Remark 1, and the Cauchy-Schwarz inequality,

$$\begin{aligned}&\mathrm{E}\{(\boldsymbol{R}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)),\boldsymbol{P}_i(\mathbf{x}(t),\boldsymbol{\mu}(t)))|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\ &=\mathrm{E}(J_i(\mathbf{x}(t))\frac{\mathbf{x}^i(t)-\boldsymbol{\mu}^i}{\sigma_t^2}-\tilde{\boldsymbol{M}}_i(\boldsymbol{\mu},\sigma_t),\\ &\qquad\frac{\mathbf{x}^i(t)-\boldsymbol{\mu}^i}{\sigma_t^2}(J_i(\boldsymbol{a}(t))-J_i(\mathbf{x}(t))))\\ &\le\|\tilde{\boldsymbol{M}}_i(\boldsymbol{\mu},\sigma_t)\|\mathrm{E}\|P_i(\mathbf{x}_t,\boldsymbol{\mu})\|\\ &-\mathrm{E}\left\{\frac{J_i(\mathbf{x}(t))(J_i(\boldsymbol{a}(t))-J_i(\mathbf{x}(t)))\|\mathbf{x}^i(t)-\boldsymbol{\mu}^i\|^2}{\sigma_t^4}\right\}\\ &\le\|\tilde{\boldsymbol{M}}_i(\boldsymbol{\mu},\sigma_t)\|\mathrm{E}\|P_i(\mathbf{x}_t,\boldsymbol{\mu})\|\\ &\qquad+\Pr\{\mathbf{x}(t)\in\mathbb{R}^{Nd}\setminus\boldsymbol{A}\}\times\\ &\times\mathrm{E}\left\{\frac{|J_i(\mathbf{x}(t))|\|\mathbf{x}(t)-\boldsymbol{\mu}\|\|\mathbf{x}^i(t)-\boldsymbol{\mu}^i\|^2}{\sigma_t^4}\right\}.\end{aligned} \tag{24}$$

Note that $\|\tilde{\boldsymbol{M}}_i(\boldsymbol{\mu},\sigma_t)\|$ is bounded from Lemma 2), and that

$$\mathrm{E}\left\{\frac{|J_i(\mathbf{x}(t))|\|\mathbf{x}(t)-\boldsymbol{\mu}\|\|\mathbf{x}^i(t)-\boldsymbol{\mu}^i\|^2}{\sigma_t^4}\right\}=\frac{h_i(\boldsymbol{\mu},\sigma_t)}{\sigma_t^4}, \tag{25}$$

where $h_i(\boldsymbol{\mu},\sigma_t)$ is a quadratic function of $\boldsymbol{\mu}$ and σ_t, $i\in[N]$ (see Appendix E for more details). Hence, due to the choice of the parameters r_t and σ_t (in particular, Assumption 6 d)) and the estimations in (22)- (24), we conclude that the terms containing $\boldsymbol{P}_i$ are dominated by other terms in the inequality in (16). Thus, by inserting (17)-(21) into (16), we obtain

$$\begin{aligned}&\mathrm{E}\{\|\boldsymbol{\mu}^i(t+1)-\mathbf{z}^i(t+1)\|^2|\boldsymbol{\mu}(t)=\boldsymbol{\mu}\}\\ &\le(1-2\beta_t\epsilon_t)\|\boldsymbol{\mu}^i-\mathbf{z}^i(t+1)\|^2\\ &\quad-2\beta_t(\boldsymbol{M}_i(\boldsymbol{\mu})-\boldsymbol{M}_i(\mathbf{z}(t)),\boldsymbol{\mu}^i-\mathbf{z}^i(t))\\ &\quad+2\beta_tO(\sigma_t)(1+V(t,\boldsymbol{\mu}))\\ &\quad+\beta_t^2\epsilon^2(t)V(t,\boldsymbol{\mu})\\ &\quad+O(\gamma_t^2)(1+V(t,\boldsymbol{\mu})),\end{aligned} \tag{26}$$

where in the last inequality we used the fact that $\epsilon_t\to0$ (Assumption 6 a)), $\gamma_t\to0$, and $\sigma_t\to0$ for all $i\in[N]$ as $t\to\infty$ (Assumption 6 c), d)). Thus, taking into account Assumption 6 c), d) and (26), we obtain

$$\mathrm{E}[\|\boldsymbol{\mu}(t+1)-\mathbf{z}(t+1)\|^2|\boldsymbol{\mu}(t)=\boldsymbol{\mu}]$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \mathrm{E}[\|\boldsymbol{\mu}^i(t+1) - \mathbf{z}^i(t+1)\|^2 | \boldsymbol{\mu}(t) = \boldsymbol{\mu}] \\
&\leq (1 - 2\epsilon_t\beta_t)\|\boldsymbol{\mu} - \mathbf{z}(t)\|^2 - 2\beta_t(\boldsymbol{M}(\boldsymbol{\mu}) - \boldsymbol{M}(\mathbf{z}(t)), \boldsymbol{\mu} - \mathbf{z}(t)) \\
&\quad + O(\beta_t\sigma_t + \gamma_t^2)(1 + V(t, \boldsymbol{\mu})). \qquad (27)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
LV(t, \boldsymbol{\mu}) \leq & -2\beta_t(\boldsymbol{M}(\boldsymbol{\mu}) - \boldsymbol{M}(\mathbf{z}(t)), \boldsymbol{\mu} - \mathbf{z}(t)) \\
& + O(\beta_t\sigma_t + \gamma_t^2)(1 + V(t, \boldsymbol{\mu})). \qquad (28)
\end{aligned}
$$

According to Assumption 6 b)-c), $\sum_{t=0}^{\infty} \beta_t\sigma_t + \gamma_t^2 < \infty$. Furthermore, from Assumption 6 a) $\sum_{t=0}^{\infty} \beta_t = \infty$. Taking into account this, (28), and monotonicity of $\boldsymbol{M}$ implying

$$
(\boldsymbol{M}(\boldsymbol{\mu}) - \boldsymbol{M}(\mathbf{z}(t)), \boldsymbol{\mu} - \mathbf{z}(t)) \geq 0, \ \forall t, \ \forall \boldsymbol{\mu} \in \boldsymbol{A}, \qquad (29)
$$

we conclude that $LV(t, \boldsymbol{\mu})$ satisfies the decay needed for the application of Theorem 2.5.2 in [15] and consequently, $\boldsymbol{\mu}(t)$ is finite almost surely for any $t \in \mathbb{Z}_+$ irrespective of $\boldsymbol{\mu}(0)$. ■

D. Convergence to Nash equilibrium

We will use the bound estimations in the previous section to prove convergence of the algorithm. In particular, we use Inequality (27), which bounds the decay of the sequence $\mathrm{E}[\|\boldsymbol{\mu}(t+1) - \mathbf{z}(t)\|^2 | \boldsymbol{\mu}(t)]$ in terms of $\|\boldsymbol{\mu} - \mathbf{z}(t+1)\|^2$. We will show that this decay satisfies the conditions for applying Lemma 10 in [17]. From this, it can readily be inferred that random variables $\|\boldsymbol{\mu}(t) - \mathbf{z}(t)\|$ converge to zero.

First, however, let us verify that even in the compact action case, Inequalities (27), (28) hold.

Remark 2: If the set $\boldsymbol{A}$ is compact, due to Assumption 5, the inequality (21) can be replaced by

$$
\mathrm{E}\{\|\boldsymbol{R}_i(\mathbf{x}(t), \boldsymbol{\mu}(t))\|^2 | \boldsymbol{\mu}(t) = \boldsymbol{\mu}\} = O\left(\frac{1}{\sigma_t^4}\right).
$$

Moreover, the inequalities (22) and (24) hold for the case of the bounded set $\boldsymbol{A}$. Indeed, due to polynomial behavior of $J_i(\mathbf{x}(t))$ for large $\mathbf{x}(t)$, the terms $\mathrm{E}\frac{\|\mathbf{x}^i(t) - \boldsymbol{\mu}^i\|^2 |J_i(\mathbf{x}(t)) - J_i(\boldsymbol{a}(t))|^2}{\sigma_t^4}$ and $\mathrm{E}\left\{\frac{|J_i(\mathbf{x}(t))(J_i(\boldsymbol{a}(t)) - J_i(\mathbf{x}(t)))| \|\mathbf{x}^i(t) - \boldsymbol{\mu}^i\|^2}{\sigma_t^4}\right\}$ are upper bounded by some constants. Thus, for the bounded set $\boldsymbol{A}$, the inequality (27) can be rewritten as

$$
\begin{aligned}
&\mathrm{E}[\|\boldsymbol{\mu}(t+1) - \mathbf{z}(t+1)\|^2 | \boldsymbol{\mu}(t) = \boldsymbol{\mu}] \\
&\leq (1 - 2\epsilon_t\beta_t)\|\boldsymbol{\mu} - \mathbf{z}(t)\|^2 - 2\beta_t(\boldsymbol{M}(\boldsymbol{\mu}) - \boldsymbol{M}(\mathbf{z}(t)), \boldsymbol{\mu} - \mathbf{z}(t)) \\
&\quad + O(\beta_t\sigma_t + \gamma_t^2).
\end{aligned}
$$

Proof: (of Theorem 2) Note that we can rewrite (27) as:

$$
\begin{aligned}
&\mathrm{E}[\|\boldsymbol{\mu}(t+1) - \mathbf{z}(t+1)\|^2 | \mathcal{F}_t] \\
&\leq (1 - 2\epsilon_t\beta_t)\|\boldsymbol{\mu}(t) - \mathbf{z}(t)\|^2 + O(\gamma_t^2 + \beta_t\sigma_t), \qquad (30)
\end{aligned}
$$

where $\mathcal{F}_t$ is the σ-algebra generated by the random variables $\{\mathbf{x}(k), \boldsymbol{\mu}(k)\}_{k=0}^t$. In (30) we used (29) and Lemma 5.

From Assumption 6, and the choices of γ_t, σ_t, ϵ_t, we get $O(\gamma_t^2 + \beta_t\sigma_t) = O(\frac{1}{t^n})$, $\epsilon_t\beta_t = \frac{1}{t^m}$, with $n > 1$, $m \leq 1$. Thus,

$$
\lim_{t\to\infty} \frac{O(\gamma_t^2 + \beta_t\sigma_t)}{\epsilon_t\beta_t} = 0.
$$

Assumption 6 d), the fact that $\sum_{t=0}^{\infty} \gamma_t^2 + \beta_t\sigma_t < \infty$ and the above result in the decay (30) imply that we can apply Lemma 10 in [17] to the sequence $\|\boldsymbol{\mu}(t+1) - \mathbf{z}(t+1)\|^2$ to conclude its almost sure convergence to 0 as $t \to \infty$. Next, by taking into account Theorem 3 and Theorem 1, we obtain that

$$
\Pr\{\lim_{t\to\infty} \boldsymbol{\mu}(t) = \boldsymbol{a}^*\} = 1,
$$

where $\boldsymbol{a}^*$ is the least norm Nash equilibrium in the game $\Gamma(N, \{A_i\}, \{J_i\})$. Finally, Assumption 6 implies that $\lim_{t\to\infty} \sigma_t = 0$. Taking into account that $\mathbf{x}(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \sigma_t)$ and $\lim_{t\to\infty} \|\boldsymbol{a}(t) - \mathbf{x}(t)\| = 0$, we conclude that $\boldsymbol{a}(t)$ converges weakly to a Nash equilibrium $\boldsymbol{a}^* = \boldsymbol{\mu}^*$. Moreover, according to Portmanteau Lemma [10], this convergence is also in probability.

■

V. Discussion

In the proposed algorithm convergence is established under mild conditions as strict monotonicity of the game mapping is not implied. This significantly extends the applicability of bandit online learning. For example, the zero-sum game considered in [13] with an interior Nash equilibrium satisfies the assumption of our theorem. Whereas all the follow-the-regularized leader (FTRL) learning approach fails to converge in a simple zero-sum game (such as matching penny), our doubly regularized approach can resolve this problem. In general, examples of games that satisfy assumptions above include mixed extensions of zero-sum games, Cournot competition, continuous-action congestion games and convex potential games. On the other hand, mixed extensions of non-zero sum games do not satisfy the monotonicity assumption in general.

In accordance with the payoff-based information structure, the parameters $\gamma_t, \sigma_t, \epsilon_t, r_t$ are independent of the problem data including the Lipschitz constant of the game mapping or the constraint sets. Below, we further specify feasible choices of the parameters to ensure convergence.

Lemma 6: A sufficient condition on $0 < a_1, a_2, a_3, a_4 < 1$ for satisfying Assumption 6 is as follows:

i) $a_1 + 2a_2 < 1, \quad a_1 + 2a_2 + a_3 < 1.$

ii) $a_1 + 2a_2 + a_3 < 1, \quad a_1 + 2a_2 + 6a_3 - 2a_4 < 1.$

iii) $2a_1 > 1, \quad a_1 + 3a_2 > 1.$

iv) $a_3 < a_4 < a_2.$

Proof: The series $\sum_{t=0}^{\infty} 1/t^m$ converges for $m > 1$ and diverges otherwise. Thus, the statements i), iii), iv) above follow. To show statement ii), let us consider the term $(\epsilon_t - \epsilon_{t-1})^2$ in the first summand of b), namely, $\sum_{t=0}^{\infty} \frac{(\epsilon_t - \epsilon_{t-1})^2}{\beta_t\epsilon_t^3}$:

$$
\begin{aligned}
&(\epsilon_t - \epsilon_{t-1})^2 = (t^{-a_3} - (t-1)^{-a_3})^2 \quad (\text{multiply by } \frac{t^{2a_3}}{t^{2a_3}}) \\
&= \left((1 - 1/t)^{-a_3} - 1\right)^2 / t^{2a_3} \qquad (\text{do Taylor approximation}) \\
&= (1 + a_3/t + O(t^{-2}) - 1)^2 / t^{2a_3} = O(t^{-2-2a_3}).
\end{aligned}
$$

Combining the above with the denominator $\beta_t\epsilon_t^3$, we obtain that $\sum_t \frac{(\epsilon_t - \epsilon_{t-1})^2}{\beta_t\epsilon_t^3}$ converges if $a_3 1 + 2a_2 + a_3 < 1$. Repeating the same analysis for $\frac{(r_t - r_{t-1})^2}{\beta_t\epsilon_t^6}$, we obtain $\sum_t \frac{(r_t - r_{t-1})^2}{\beta_t\epsilon_t^6}$ converges if $a_1 + 2a_2 + 6a_3 - 2a_4 < 1$ and ii) is verified.

■

VI. Conclusions

We designed an algorithm for learning Nash equilibria in convex games with monotone game mappings using online bandit feedback information. Our algorithm relied on a suitable double regularization to handle non-strictly monotone game maps as well as feasibility of the queried actions (online setting). The implications of our result is that players can learn Nash equilibria in several monotone games such as finite action zero-sum games, infinite action zero-sum convex games and convex games with linear coupling constraints. Several points remain open and are topic of our current study. These include showing that our algorithm is no-regret, unifying different sampling approaches to perform one-point estimation of the game mapping for bandit learning in games, and analyzing the convergence rate of the algorithm.

References

[1] J. P. Bailey, G. Gidel, and G. Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Conference on Learning Theory*, pages 391–407, 2020.

[2] D Balduzzi, S Racaniere, J Martens, J Foerster, K Tuyls, and T Graepel. The mechanics of n-player differentiable games. In *ICML*, volume 80, pages 363–372. JMLR. org, 2018.

[3] B. Bharath and V. S. Borkar. Stochastic approximation algorithms: Overview and recent trends. *Sadhana*, 24(4):425–452, 1999.

[4] M. Bravo, D. Leslie, and P. Mertikopoulos. Bandit learning in concave n-person games. In *Advances in Neural Information Processing Systems*, pages 5661–5671, 2018.

[5] F. Facchinei and C. Kanzow. Generalized Nash equilibrium problems. *4OR*, 5(3):173–210, 2007.

[6] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[7] S. Grammatico. Comments on distributed robust adaptive equilibrium computation for generalized convex games (automatica 63(2016) 82-91). *Automatica*, 97:186 – 188, 2018.

[8] J.-B. Hiriart-Urruty and C. Lemarchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer, 2001.

[9] R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.

[10] A. Klenke. *Probability theory: a comprehensive course*. Springer, London, 2008.

[11] J. Koshal, A Nedić, and U. Shanbhag. Single timescale regularized stochastic approximation schemes for monotone nash games under uncertainty. In *IEEE Conference on Decision and Control*, pages 231–236, 2010.

[12] P. Mertikopoulos, B. Lecouat, H. Zenati, Ch.-Sh. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[13] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717, 2018.

[14] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, April 2017.

[15] M. B. Nevelson and R. Z. Khasminskii. *Stochastic approximation and recursive estimation [translated from the Russian by Israel Program for Scientific Translations ; translation edited by B. Silver]*. American Mathematical Society, 1973.

[16] J.-S. Pang and F. Facchinei. *Finite-dimensional variational inequalities and complementarity problems : vol. 1*. Springer series in operations research. Springer, New York, Berlin, Heidelberg, 2003.

[17] B. T. Poljak. *Introduction to optimization*. Optimization Software, 1987.

[18] T. Tatarenko and M. Kamgarpour. Learning generalized nash equilibria in a class of convex games. *IEEE Transactions on Automatic Control*, 2018. to appear. URL: https://arxiv.org/abs/1703.04113.

[19] T. Tatarenko and M. Kamgarpour. Learning nash equilibria in monotone games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3104–3109, 2019.

[20] A. L. Thathachar and P. S. Sastry. *Networks of Learning Automata: Techniques for Online Stochastic Optimization*. Springer US, 2003.

[21] V.A. Zorich and R. Cooke. *Mathematical Analysis II*. Mathematical Analysis. Springer, 2004.

Appendix

The appendix provides supporting theorems and proof of certain lemmas and statements.

A. Supporting Theorems

Let $\{\mathbf{X}(t)\}_t$, $t \in \mathbb{Z}_+$, be a discrete-time Markov process on some state space $E \subseteq \mathbb{R}^d$, namely $\mathbf{X}(t) = \mathbf{X}(t,\omega) : \mathbb{Z}_+ \times \Omega \to E$, where Ω is the sample space of the probability space on which the process $\mathbf{X}(t)$ is defined. The transition function of this chain, namely $\Pr\{\mathbf{X}(t+1) \in \Gamma | \mathbf{X}(t) = \mathbf{X}\}$, is denoted by $P(t, \mathbf{X}, t+1, \Gamma)$, $\Gamma \subseteq E$.

Definition 4: The operator L defined on the set of measurable functions $V : \mathbb{Z}_+ \times E \to \mathbb{R}$, $\mathbf{X} \in E$, by

$$\begin{aligned} LV(t,\mathbf{X}) &= \int P(t,\mathbf{X},t+1,dy)[V(t+1,y) - V(t,\mathbf{X})] \\ &= E[V(t+1,\mathbf{X}(t+1)) \mid \mathbf{X}(t) = \mathbf{X}] - V(t,\mathbf{X}), \end{aligned}$$

is called a *generating operator* of a Markov process $\{\mathbf{X}(t)\}_t$. Next, we formulate the following theorem for discrete-time Markov processes, which is proven in [15], Theorem 2.5.2.

Theorem 4: Consider a Markov process $\{\mathbf{X}(t)\}_t$ and suppose that there exists a function $V(t,\mathbf{X}) \geq 0$ such that $\inf_{t\geq 0} V(t,\mathbf{X}) \to \infty$ as $\|\mathbf{X}\| \to \infty$ and

$$LV(t,\mathbf{X}) \leq -\alpha(t+1)\psi(t,\mathbf{X}) + f(t)(1+V(t,\mathbf{X})),$$

where $\psi \geq 0$ on $\mathbb{R}\times\mathbb{R}^d$, $f(t)>0$, $\sum_{t=0}^{\infty} f(t) < \infty$. Let $\alpha(t)$ be such that $\alpha(t)>0$, $\sum_{t=0}^{\infty}\alpha(t)=\infty$. Then, almost surely $\sup_{t\geq 0}\|\mathbf{X}(t,\omega)\| = R(\omega) < \infty$.

The following result related to the convergence of the stochastic process is proven in Lemma 10 (page 49) in [17].

Theorem 5: Let $v_0,\ldots,v_k$ be a sequence of random variables, $v_k \geq 0$, $\mathrm{E}v_0 < \infty$ and let

$$\mathrm{E}\{v_{k+1}|\mathcal{F}_k\} \leq (1-\alpha_k)v_k + \beta_k,$$

where $\mathcal{F}_k$ is the σ-algebra generated by the random variables $\{v_0,\ldots,v_k\}$, $0<\alpha_k<1$, $\sum_{k=0}^{\infty}\alpha_k=\infty$, $\beta_k\geq 0$, $\sum_{k=0}^{\infty}\beta_k<\infty$, $\lim_{k\to\infty}\frac{\beta_k}{\alpha_k}=0$. Then $v_k\to 0$ almost surely, $\mathrm{E}v_k\to 0$ as $k\to\infty$.

B. Proof of Lemma 2

Proof: We verify that the differentiation under the integral sign in (4) is justified. It can then readily be verified that (6) holds, by taking the differentiation inside the integral. A sufficient condition for differentiation under the integral is that the integral of the formally differentiated function with respect to μ_k^i converges uniformly, whereas the differentiated function is continuous (see [21], Chapter 17). By formally differentiating the function under the integral sign and omitting the arguments t, we obtain

$$\frac{1}{\sigma^2}\int_{\mathbb{R}^{Nd}} J_i(\boldsymbol{x})(x_k^i-\mu_k^i)p(\boldsymbol{\mu},\boldsymbol{x},\sigma)d\boldsymbol{x}. \tag{31}$$

Given Assumption 1, $J_i(\boldsymbol{x})(x_k^i-\mu_k^i)p(\boldsymbol{\mu},\boldsymbol{x},\sigma)$ is continuous. Thus, it remains to check that the integral of this function converges uniformly with respect to any $\boldsymbol{\mu}\in\boldsymbol{A}$. If $\boldsymbol{A}$ is bounded, then the conclusion follows from the polynomial behavior of the function J_i on the infinity.

Now, we move to the case of the unbounded set $\boldsymbol{A}$. To this end, we can write the Taylor expansion of the function J_i around the point $\boldsymbol{\mu}(i,k)\in\mathbb{R}^{Nd}$ with the coordinates $\mu(i,k)_k^i=\mu_k^i$ and $\mu(i,k)_m^j=x_m^j$ for any $j\neq i$, $m\neq k$, in the integral (31):

$$\begin{aligned}
&\int_{\mathbb{R}^{Nd}} J_i(\boldsymbol{x})(x_k^i-\mu_k^i)p(\boldsymbol{\mu},\boldsymbol{x},\sigma)d\boldsymbol{x}\\
&=\int_{\mathbb{R}^{Nd}}[J_i(\boldsymbol{\mu}(i,k))\\
&\quad+\frac{\partial J_i(\boldsymbol{\eta}(\boldsymbol{x},\boldsymbol{\mu}))}{\partial x_k^i}(x_k^i-\mu_k^i)](x_k^i-\mu_k^i)p(\boldsymbol{\mu},\boldsymbol{x},\sigma)d\boldsymbol{x}\\
&=\int_{\mathbb{R}^{Nd}}\frac{\partial J_i(\boldsymbol{\eta}(\boldsymbol{x},\boldsymbol{\mu}))}{\partial x_k^i}(x_k^i-\mu_k^i)^2p(\boldsymbol{\mu},\boldsymbol{x},\sigma)d\boldsymbol{x}\\
&=\int_{\mathbb{R}^{Nd}}\frac{\partial J_i(\boldsymbol{\eta}_1(\boldsymbol{y},\boldsymbol{\mu}))}{\partial x_k^i}(y_k^i)^2p(\mathbf{0},\boldsymbol{y},\sigma)d\boldsymbol{y},
\end{aligned}$$

where $\boldsymbol{\eta}(\boldsymbol{x},\boldsymbol{\mu}) = \boldsymbol{\mu}(i,k)+\theta(\boldsymbol{x}-\boldsymbol{\mu}(i,k))$, $\theta\in(0,1)$, $\boldsymbol{y}=\boldsymbol{x}-\boldsymbol{\mu}(i,k)$, $\boldsymbol{\eta}_1(\boldsymbol{y},\boldsymbol{\mu})=\boldsymbol{\mu}(i,k)+\theta\boldsymbol{y}$. The uniform convergence of the integral above and, in particular, its boundedness, follows from the fact[3] that under Assumption 5, $\frac{\partial J_i(\boldsymbol{\eta}_1(\boldsymbol{y},\boldsymbol{\mu}))}{\partial x_k^i}\leq l_k^i$ for some positive constant l_k^i and for all $i\in[N]$ and $k\in[d]$. Thus, $|\frac{\partial J_i(\boldsymbol{\eta}_1(\boldsymbol{y},\boldsymbol{\mu}))}{\partial x_k^i}(y_k^i)^2p(\mathbf{0},\boldsymbol{y},\sigma)|\leq h(\boldsymbol{y}) = l(y_k^i)^2p(\mathbf{0},\boldsymbol{y},\sigma)$,where $\int_{\mathbb{R}^{Nd}}h(\boldsymbol{y})d\boldsymbol{y}<\infty$. ■

[3]see the basic sufficient condition using majorant [21], Chapter 17.2.3.

C. Proof of Lemma 3

Proof: Without loss of generality, assume $\mathbf{x}\notin(1-r_{t-1})\boldsymbol{A}$ (otherwise, $\|\mathrm{Proj}_{(1-r_{t-1})\boldsymbol{A}}\mathbf{x}-\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}\mathbf{x}\|=0$). Due to convexity of the set $\boldsymbol{A}\subset\mathbb{R}^{Nd}$ there exists a convex function $g:\mathbb{R}^{Nd}\to\mathbb{R}$ such that $\boldsymbol{A}=\{\mathbf{x}: g(\mathbf{x})\leq 0\}$, whereas $(1-r_t)\boldsymbol{A}=\{\mathbf{x}: g(\mathbf{x})\leq -r_t\}$ for any t. Moreover, since $\mathrm{Proj}_{(1-r_{t-1})\boldsymbol{A}}\mathbf{x} = \mathrm{Proj}_{(1-r_{t-1})\boldsymbol{A}}\{\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}\mathbf{x}\}$, we have $\|\mathrm{Proj}_{(1-r_{t-1})\boldsymbol{A}}\mathbf{x}-\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}\mathbf{x}\| = d$, where

$$\begin{aligned}
d = \min_{\mathbf{y}}&\|\mathbf{y}-\mathbf{x}'\|,\ \mathbf{x}'=\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}\mathbf{x},\\
&\text{s.t. } g(\mathbf{y})=-r_{t-1}.
\end{aligned}$$

The optimization problem has a solution $\mathbf{y}^*$ for which the gradient of the corresponding Lagrangian is zero, namely

$$\frac{(\mathbf{y}^*-\mathbf{x}')}{\|\mathbf{y}^*-\mathbf{x}'\|}+\lambda\nabla g(\mathbf{y}^*)=\mathbf{0},$$

where $\lambda>0$ is the dual multiplier of the problem under consideration. Notice that due to Assumption 1 and the choice of r_1 the Slator's condition for the constraints $g(\mathbf{x})\leq -r_t$ holds for all t. Hence, for any $\mathbf{x}\in\mathbb{R}^{Nd}$ there exists a constant $\Lambda>0$ such that $\lambda<\Lambda$ (see [8]). Thus, we conclude that

$$\nabla g(\mathbf{y}^*) = -\frac{(\mathbf{y}^*-\mathbf{x}')}{\lambda\|\mathbf{y}^*-\mathbf{x}'\|}.$$

Next, due to convexity of the function g,

$$\begin{aligned}
g(\mathbf{x}') &\geq g(\mathbf{y}^*)+(\nabla g(\mathbf{y}^*),\mathbf{x}'-\mathbf{y}^*)\\
&= -r_{t-1}+\frac{\|\mathbf{y}^*-\mathbf{x}'\|^2}{\lambda|\mathbf{y}^*-\mathbf{x}'\|}\geq -r_{t-1}+\frac{\|\mathbf{y}^*-\mathbf{x}'\|}{\Lambda}.
\end{aligned}$$

Thus, taking into account that $g(\mathbf{x}')\leq -r_t$, we obtain

$$d=\|\mathbf{y}^*-\mathbf{x}'\|\leq\Lambda(r_{t-1}-r_t)=O(|r_{t-1}-r_t|).$$

■

D. Proof of Lemma 4

Proof: Let us use $\theta=\epsilon_t^{3/2}$ in Lemma 1 to express $\boldsymbol{y}(t)$ as $\boldsymbol{y}(t)=\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}[\boldsymbol{y}(t)-\epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t))+\epsilon_t\boldsymbol{y}(t))]$. Using this equivalence, the triangle inequality and non-expansion of projection operator we obtain

$$\begin{aligned}
&\|\boldsymbol{y}(t)-\boldsymbol{y}(t-1)\|\\
&\leq\|\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}[\boldsymbol{y}(t)-\epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t))+\epsilon_t\boldsymbol{y}(t))]\\
&-\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}[\boldsymbol{y}(t-1)-\epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t-1))\\
&+\epsilon_{t-1}\boldsymbol{y}(t-1))]\|+\|\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}[\boldsymbol{y}(t-1)\\
&-\epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t-1))+\epsilon_{t-1}\boldsymbol{y}(t-1))]\\
&-\mathrm{Proj}_{(1-r_{t-1})\boldsymbol{A}}[\boldsymbol{y}(t-1)-\\
&\quad\epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t-1))+\epsilon_{t-1}\boldsymbol{y}(t-1))]\|\\
&\leq\|(1-\epsilon_t^{3/2}\epsilon_t)(\boldsymbol{y}(t)-\boldsymbol{y}(t-1))
\end{aligned}$$

$$
\begin{aligned}
&- \epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t)) - \boldsymbol{M}(\boldsymbol{y}(t-1)))\\
&+ \boldsymbol{y}(t-1)\epsilon_t^{3/2}(\epsilon_{t-1}-\epsilon_t)\|\\
&+ \|\mathrm{Proj}_{(1-r_t)\boldsymbol{A}}[\tilde{\boldsymbol{y}}(t-1)] - \mathrm{Proj}_{(1-r_{t-1})\boldsymbol{A}}[\tilde{\boldsymbol{y}}(t-1)]\|,
\end{aligned}
$$

where $\tilde{\boldsymbol{y}}(t-1) = \boldsymbol{y}(t-1) - \epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t-1)) + \epsilon_{t-1}\boldsymbol{y}(t-1))$. Next, due to Lemma 3 we have for any $\theta_t > 0$ and $\kappa_t > 0$

$$
\begin{aligned}
&\|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2 \leq (1+\theta_t)\|(1-\epsilon_t^{3/2}\epsilon_t)(\boldsymbol{y}(t) - \boldsymbol{y}(t-1))\\
&- \epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t)) - \boldsymbol{M}(\boldsymbol{y}(t-1)))\\
&+ \boldsymbol{y}(t-1)\epsilon_t^{3/2}(\epsilon_{t-1}-\epsilon_t)\|^2 + (1+1/\theta_t)O((r_t - r_{t-1})^2)\\
&\leq (1+\theta_t)(1+\kappa_t)\|(1-\epsilon_t^{3/2}\epsilon_t)(\boldsymbol{y}(t) - \boldsymbol{y}(t-1))\\
&- \epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t)) - \boldsymbol{M}(\boldsymbol{y}(t-1)))\|^2\\
&+ (1+1/\kappa_t)\|\boldsymbol{y}(t-1)\|^2\epsilon_t^3(\epsilon_{t-1}-\epsilon_t)^2\\
&+ (1+1/\theta_t)O((r_t - r_{t-1})^2).
\end{aligned}
$$

Furthermore, there exists T such that for all $t > T$

$$
\begin{aligned}
&\|(1-\epsilon_t^{3/2}\epsilon_t)(\boldsymbol{y}(t) - \boldsymbol{y}(t-1))\\
&\qquad - \epsilon_t^{3/2}(\boldsymbol{M}(\boldsymbol{y}(t)) - \boldsymbol{M}(\boldsymbol{y}(t-1)))\|^2\\
&\leq (1-\epsilon_t^{3/2}\epsilon_t)^2\|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2\\
&\qquad + \epsilon_t^3\|\boldsymbol{M}(\boldsymbol{y}(t)) - \boldsymbol{M}(\boldsymbol{y}(t-1))\|^2\\
&\leq (1 - 2\epsilon_t^{3/2}\epsilon_t + \epsilon_t^3(\epsilon_t^2 + L^2))\|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2\\
&\leq (1-\epsilon_t^{3/2}\epsilon_t)\|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2,
\end{aligned}
$$

where the first inequality is due to monotonicity of the mapping $\boldsymbol{M}$, the second one is as it is Lipschitz continuous, and the third one is due to the fact that $\epsilon_t^3(\epsilon_t^2 + L^2) \leq \epsilon_t^{3/2}\epsilon_t$ for sufficiently large t (since $\epsilon_t \to 0$, see Assumption 6). Thus,

$$
\begin{aligned}
&\|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2 \leq (1+\theta_t)(1+\kappa_t)(1-\epsilon_t^{3/2}\epsilon_t)\times\\
&\quad \|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2 + (1+1/\kappa_t)M_{\boldsymbol{y}}^2\epsilon_t^3(\epsilon_{t-1}-\epsilon_t)^2\\
&+ (1+1/\theta_t)O((r_t - r_{t-1})^2),
\end{aligned}
$$

where $M_{\boldsymbol{y}}^2$ is the uniform upper bound of the norm of the sequence $\boldsymbol{y}(t)$. By rearranging the terms above, we obtain

$$
\begin{aligned}
(1 - (1+\theta_t)(1+\kappa_t)(1-\epsilon_t^{3/2}\epsilon_t))\|\boldsymbol{y}(t) - \boldsymbol{y}(t-1)\|^2&\\
\leq (1+1/\kappa_t)M_{\boldsymbol{y}}^2\epsilon_t^3(\epsilon_{t-1}-\epsilon_t)^2&\\
+ (1+1/\theta_t)O((r_t - r_{t-1})^2).&
\end{aligned}
$$

We conclude the proof by noticing that, according to the choice of $\epsilon_t^{3/2}$ and ϵ_t and by taking $\kappa_t = \theta_t = \frac{1}{4}\epsilon_t^{3/2}\epsilon_t$, we obtain,

$$
\begin{aligned}
(1 - (1+\theta_t)(1+\kappa_t)(1-\epsilon_t^{3/2}\epsilon_t)) &\geq \epsilon_t^{3/2}\epsilon_t - \theta_t - \kappa_t - \theta_t\kappa_t\\
&\geq 0.4\epsilon_t^{3/2}\epsilon_t.
\end{aligned}
$$

■

E. Verification of Equations (21) *and* (25)

Due to Assumption 5 there exists a compact set $\mathbb{S} \subset \mathbb{R}^{Nd}$ such that for any $\boldsymbol{x} \notin \mathbb{S}$

$$
J_i(\boldsymbol{x}) \leq (\boldsymbol{c}, \boldsymbol{x}) + b_0
$$

for some $\boldsymbol{c} = (c_1^1, \ldots, c_d^1, \ldots, c_1^N, \ldots, c_d^N) \in \mathbb{R}^{Nd}$ and $b_0 \in \mathbb{R}$. Thus, for some positive S, d_1 and d_2 we get

$$
\begin{aligned}
&\int_{\mathbb{R}^{Nd}} {J_i}^2(\boldsymbol{x})\frac{(x_k^i - \mu_k^i)^2}{\sigma_t^4}p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x}\\
&\leq \int_{\mathbb{S}} {J_i}^2(\boldsymbol{x})\frac{(x_k^i - \mu_k^i)^2}{\sigma_t^4}p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x}\\
&\qquad + \int_{\mathbb{R}^{Nd}\setminus\mathbb{S}} [d_1\|x\|^2 + d_2]\frac{(x_k^i - \mu_k^i)^2}{\sigma_t^4}p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x}\\
&\leq \frac{S}{\sigma_t^4} + \int_{\mathbb{R}^{Nd}} [d_1\|x\|^2 + d_2]\frac{(x_k^i - \mu_k^i)^2}{\sigma_t^4}p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x}\\
&= \frac{S}{\sigma_t^4} + \frac{f_i(\boldsymbol{\mu}, \sigma_t)}{\sigma_t^4},
\end{aligned}
$$

$f_i(\boldsymbol{\mu}, \sigma_t)$ being a quadratic function of $\boldsymbol{\mu}$ and σ_t. The last equality is due to the fact that

$$
\begin{aligned}
&\int_{\mathbb{R}^{Nd}} (x_k^i - \mu_k^i)^2 p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x} = \sigma_t^2,\\
&\int_{\mathbb{R}^{Nd}} (x_k^i)^2(x_k^i - \mu_k^i)^2 p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x}\\
&\quad \leq \int_{\mathbb{R}^{Nd}} [2(x_k^i - \mu_k^i)^2 + (\mu_k^i)^2)](x_k^i - \mu_k^i)^2 p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x}\\
&\quad = 2\sigma_t^4 + (\mu_k^i)^2)\sigma_t^2,\\
&\int_{\mathbb{R}^{Nd}} (x_m^j)^2(x_k^i - \mu_k^i)^2 p(\boldsymbol{\mu}, \boldsymbol{x})d\boldsymbol{x}\\
&\quad = (\sigma_t^2 + (\mu_m^j)^2)\sigma_t^2|.
\end{aligned}
$$

Analogously one obtains the estimation (25).