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

Convergence Rate of Learning a Strongly Variationally Stable Equilibrium

Tatiana Tatarenko, Maryam Kamgarpour *Member, IEEE*

***Abstract*— We derive the rate of convergence to the globally strongly variationally stable Nash equilibrium in a convex game, for a zeroth-order learning algorithm. Though we do not assume strong monotonicity of the game, our rates for the one-point feedback, $O\left(\frac{Nd}{t^{1/2}}\right)$, and for the two-point feedback, $O\left(\frac{N^2d^2}{t}\right)$, match the best known rates for strongly monotone games under zeroth-order information.**

I. Introduction

Game-theoretic learning under zeroth-order information consists in deriving an algorithm for each player that uses only evaluations of the player's cost function and converges to an equilibrium of the game. The zeroth-order setting arises in applications in which each agent does not know the functional form of her objective or cannot readily compute its gradient, due to complex dependence of her cost on other players' actions. For example, price functions in electricity markets depend on consumption/production of all agents in non-trivial way [22], travel times in a routing or transportation network depend on routes taken by other agents and the link capacities [5]. In contrast, each agent can evaluate her objective at her chosen action, given actions played by other agents, and thus, obtain the zeroth-order information (cost function values).

The works [2], [18], [17] proposed zeroth-order learning algorithms for games over continuous action sets. The underlying idea in the proposed algorithms of the above works is developing a randomized sampling technique to estimate gradients of players' cost functions using the zeroth-order information, and to then use the estimated gradient in a stochastic gradient descent scheme. In general, convergence in zeroth-order learning is slow due to the high variance of the gradient estimators. Hence, it is relevant to establish optimal rates of convergence for this class of problems. Our goal in this paper is to estimate rate of convergence for zeroth-order learning algorithms in a specific class of games.

While convergence rates in zeroth-order convex optimization have been well-explored, less work has been dedicated to deriving convergence rates for zeroth-order learning in convex games. Under strong monotonicity assumption on the game pseudo-gradient, past work derived a rate of $O\left(\frac{1}{t^{1/3}}\right)$ in [2, Theorem 5.2], [17, Theorem 3], for the corresponding algorithms in each work. Recently, [3] demonstrated that the rate of $O\left(\frac{1}{t^{1/3}}\right)$ for the algorithm proposed in [2] is suboptimal. Indeed, a refined analysis technique of the same algorithm along with suitable choices of the stepsize and sampling distribution can ensure $O\left(\frac{1}{t^{1/2}}\right)$ [3]. Independently, [19] showed that the rate derived in [17, Theorem 3] for their proposed algorithm can be improved to $O\left(\frac{1}{t^{1/2}}\right)$. The rate of $O\left(\frac{1}{t^{1/2}}\right)$ is optimal as it matches the lower bound for the class of zeroth-order strongly convex smooth optimization under one-point feedback (a subset of the class of strongly convex games considered in the works mentioned above) [10]. Both [3], [19], require strong monotonicity of the game pseudo-gradient in their rate analysis.

A major recent interest in learning equilibria in convex games is on relaxing the requirement of monotonicity on the game pseudo-gradient. To this end, some works have relaxed the assumption of strongly monotone pseudo-gradients and considered games with pseudo-gradients which are restricted strongly monotone with respect to a Nash equilibrium [7], [12], [21]. However, the convergence rate of zeroth-order learning algorithms for such games have not been addressed. Further relaxing the monotonicity requirements, the work [13] considers the so-called (local/global) *variational stability* of an equilibrium. While the game monotonicity implies variational stability of the equilibria, an equilibrium can be variationally stable (VS) even when the game psuedo-gradient is not (strongly) monotone or restricted strongly monotone, see Examples 1 and 2[1].

It has been shown that the existence of a globally strongly VS Nash equilibrium is sufficient for convergence of the first-order learning algorithms proposed in [2], [13]. Given stochastic first-order information, [13] derived a convergence rate to the strongly VS Nash equilibrium. This rate was in terms of ergodic average of the sequence of played actions, a weaker notion of convergence than the last iterate of the played actions addressed in [2], [17], [3], [19]. Relaxing from strong to mere variational stability of an equilibrium, the work in [8] proposed an algorithm that converges to an interior mere VS equilibrium of a convex game under exact first-order feedback, i.e. knowledge of game pseudo-gradient, and characterized its convergence rate. Building on this, [9] addressed learning of a mere VS equilibrium with zeroth-order information. However, convergence rates were not established in that work.

Summarizing the above, the problem of characterizing the

This work has been accepted for publication in the proceedings of the European Control Conference 2024.

T. Tatarenko is with TU Darmstadt, Control Methods and Intelligent Systems Lab (e-mail: tatiana.tatarenko@tu-darmstadt.de).

M. Kamgarpour is with EPFL Sycamore Lab (e-mail: maryam.kamgarpour@epfl.ch).

[1]In contrast to strong monotonicity, which is a property of the game pseudo-gradient only, the conditions of variational stability and restricted strong monotonicity entail properties of the pseudo-gradient *and a Nash equilibrium* of the game. Accordingly, to establish these latter properties theoretically, one requires knowledge regarding the equilibrium point. This is a trade-off allowing for establishing convergence and its rate in games with non-(strongly) monotone pseudo-gradients.

rate of convergence of the iterates to the strongly variationally stable equilibrium under zeroth-order feedback to our knowledge was not addressed. Addressing this gap, our contributions are as follows.

- We derive the convergence rate of the zeroth-order gradient play to the strongly VS Nash equilibrium of a convex game using a one-point feedback as $O\left(\frac{1}{t^{1/2}}\right)$. This is the best rate (as a function of t) since it meets the known best bound in a subclass containing potential strongly monotone games given the same information setting (see Theorem 3 in [10]).
- We consider a two-point zeroth-order feedback model, motivated by the rate improvement achieved with the two-point feedback model in the zeroth-order optimization literature [4]. By adapting our randomized gradient estimation approach, we also improve the rate of convergence to the order of $O\left(\frac{1}{t}\right)$ which is the tight bound according to the literature on strongly convex stochastic optimization (see Theorem 2 in [1]).

Notations. The set $\{1,\dots,N\}$ is denoted by $[N]$. We consider real normed space $\mathbb{R}^d$. The column vector $\boldsymbol{x}\in\mathbb{R}^d$ is denoted by $\boldsymbol{x}=(x^1,\dots,x^d)$. We use superscripts to denoted coordinates of vectors and the player-related functions and sets. We use the subscript j which takes values $j\in\{1,2\}$ to differentiate between particular terms in the proposed algorithm for one and two-point feedback models, respectively. For any function $f:K\to\mathbb{R}$, $K\subseteq\mathbb{R}^d$, $\nabla_{x^i}f(\boldsymbol{x})=\frac{\partial f(\boldsymbol{x})}{\partial x^i}$ is the partial derivative taken in respect to the x^ith variable (coordinate) in the vector argument $\boldsymbol{x}\in\mathbb{R}^d$. We use $\langle\cdot,\cdot\rangle$ to denote the inner product in $\mathbb{R}^d$. We use $\|\cdot\|$ to denote the Euclidean norm induced by the standard dot product in $\mathbb{R}^d$. A mapping $g:\mathbb{R}^d\to\mathbb{R}^d$ is said to be *strongly monotone* on $Q\subseteq\mathbb{R}^d$ with the constant η, if for any $u,v\in Q$, $\langle g(u)-g(v),u-v\rangle\geq\eta\|u-v\|^2$; *strictly monotone*, if the strict inequality holds for $\eta=0$ and $u\neq v$, and *merely monotone* if the inequality holds for $\eta=0$. We use $\mathrm{Proj}_{\Omega}v$ to denote the projection of $v\in E$ to a set $\Omega\subseteq E$. The mathematical expectation of a random value ξ is denoted by $\mathrm{E}\{\xi\}$. Its conditional expectation in respect to some σ-algebra $\mathcal{F}$ is denoted by $\mathrm{E}\{\xi|\mathcal{F}\}$. We use the big-$O$ notation, that is, the function $f(x):\mathbb{R}\to\mathbb{R}$ is $O(g(x))$ as $x\to a$ for some $a\in\mathbb{R}$, i.e. $f(x)=O(g(x))$ as $x\to a$, if $\lim_{x\to a}\frac{|f(x)|}{|g(x)|}\leq K$ for some positive constant K. We use the little-o notation, that is, the function $f(x):\mathbb{R}\to\mathbb{R}$ is $o(g(x))$ as $x\to a$ for some $a\in\mathbb{R}$, i.e. $f(x)=o(g(x))$ as $x\to a$, if $\lim_{x\to a}\frac{|f(x)|}{|g(x)|}=0$.

II. Game Setup and the Zeroth-Order Algorithm

Consider a game $\Gamma=\Gamma(N,\{A^i\},\{J^i\})$ 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[2]. Thus, each joint action is a vector $\boldsymbol{a}=(\boldsymbol{a}^1,\dots,\boldsymbol{a}^N)\in\boldsymbol{A}\subseteq\mathbb{R}^{Nd}$, where $\boldsymbol{a}^i=(a^{i,1},\dots,a^{i,d})\in\mathbb{R}^d$. We use the notation $\boldsymbol{a}=(\boldsymbol{a}^i,\boldsymbol{a}^{-i})$, where $\boldsymbol{a}^{-i}$ is actions of players not including player i.

Definition 1: A vector $\boldsymbol{a}^*=(\boldsymbol{a}^{*1},\dots,\boldsymbol{a}^{*N})\in\boldsymbol{A}$ is called a *Nash equilibrium* if $J^i(\boldsymbol{a}^{i*},\boldsymbol{a}^{-i*})\leq J^i(\boldsymbol{a}^i,\boldsymbol{a}^{-i*})$ for any $i\in[N]$ and $\boldsymbol{a}^i\in A^i$.

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 action sets A^i, $i\in[N]$, are compact.

Note that Assumptions 1 and 2 together imply the existence of a Nash equilibrium in the game Γ [15]. In a convex game, the Nash equilibrium can be characterized through the so-called *pseudo-gradient* of the game defined below.

Definition 2: The mapping $\boldsymbol{M}:\mathbb{R}^{Nd}\to\mathbb{R}^{Nd}$, referred to as the *pseudo-gradient* of the game $\Gamma(N,\{A^i\},\{J^i\})$, is defined by $\boldsymbol{M}(\boldsymbol{a})=(\nabla_{\boldsymbol{a}^i}J^i(\boldsymbol{a}^i,\boldsymbol{a}^{-i}))_{i=1}^N=(\boldsymbol{M}^1(\boldsymbol{a}),\dots,\boldsymbol{M}^N(\boldsymbol{a}))^\top$, where $\boldsymbol{M}^i(\boldsymbol{a})=(M^{i,1}(\boldsymbol{a}),\dots,M^{i,d}(\boldsymbol{a}))^\top$, $M^{i,k}(\boldsymbol{a})=\frac{\partial J^i(\boldsymbol{a})}{\partial a^{i,k}}$, $\boldsymbol{a}\in\boldsymbol{A}, i\in[N], k\in[d]$.

In a convex game, i.e. under Assumption 1, $\boldsymbol{a}^*$ is a Nash equilibrium if and only if $\langle\boldsymbol{M}(\boldsymbol{a}^*),\boldsymbol{a}-\boldsymbol{a}^*\rangle\geq 0,\ \forall\boldsymbol{a}\in\boldsymbol{A}$ (see, for example, [15]). However, this characterization alone is not sufficient to ensure convergence of learning algorithms using the idea of a (stochastic) gradient descent approach (called also gradient play) in convex games.In particular, in most past work on learning algorithms certain structural assumptions on the game pseudo-gradient, such as strong/strict monotonicity, or assumptions on the Nash equilibrium such as variational stability, are required to prove convergence of algorithms.

Definition 3: A Nash equilibrium $\boldsymbol{a}^*$ is globally *ν-strongly variationally stable (SVS)*, if $\langle\boldsymbol{M}(\boldsymbol{a}),\boldsymbol{a}-\boldsymbol{a}^*\rangle\geq\nu\|\boldsymbol{a}-\boldsymbol{a}^*\|^2$ for any $\boldsymbol{a}\in\boldsymbol{A}$ and some $\nu>0$.

In the definition above, if the inequality holds with $\nu=0$, then the Nash equilibrium at $\boldsymbol{a}^*$ is referred to as globally merely variationally stable. On the other hand, if the above properties hold only on a neighborhood $\mathbf{D}\subset\boldsymbol{A}$, then the Nash equilibrium is locally (strongly/merely) VS.

Assumption 3: The Nash equilibrium in Γ is globally ν-strongly variationally stable with the constant ν.

If a game has a strongly variationally stable (SVS) Nash equilibrium, then the Nash equilibrium is unique [13, Proposition 2.5]. Furthermore, if a game has a strongly monotone pseudo-gradient, then its unique Nash equilibrium is strongly variationally stable. However, the converse statement is not true. The example below illustrates these definitions.

Example 1: Consider a 3-player game, where each player's action set is $[-1,2]\subset\mathbb{R}$. The cost of player i, for $i\in\{1,2,3\}$ is $J^i(a^1,a^2,a^3)=a^1a^2a^3+(a^i)^2$. Hence, the game pseudo-gradient is given by $\boldsymbol{M}(\boldsymbol{a})=(a^2a^3+2a^1,a^1a^3+2a^2,a^1a^2+2a^3)^\top$. It can be verified that there exists a Nash equilibrium at $\boldsymbol{a}^*=(0,0,0)$, since $\boldsymbol{M}(\boldsymbol{a}^*)=0$ and the game is convex. Furthermore, this Nash equilibrium is globally strongly VS with $\nu=1/2$, since

[2]For notation simplicity, we assume the dimension of each action set to be d. The algorithm and analysis readily generalize to the case of different dimensions d^i, $i\in[N]$.

$\langle \boldsymbol{M}(\boldsymbol{a}), \boldsymbol{a}-\boldsymbol{a}^*\rangle = 3a^1a^2a^3 + 2\sum_{i=1}^3 (a^i)^2 \geq \frac{\|\boldsymbol{a}-\boldsymbol{a}^*\|^2}{2}$ for any $\boldsymbol{a}\in\boldsymbol{A}=[-1,2]^3$.

Notice that the game is not monotone since the Jacobian of the $\boldsymbol{M}(\boldsymbol{a})$, given as

$$\nabla \boldsymbol{M}(\boldsymbol{a}) = \begin{pmatrix} 2 & a^3 & a^2 \\ a^3 & 2 & a^1 \\ a^2 & a^1 & 2 \end{pmatrix}$$

has a negative eigenvalue for $\boldsymbol{a}=(2,1,2)$. Furthermore, the restriction of the game to the action set $[-1,0]$ results in the same unique globally stable Nash equilibrium, but this time, this equilibrium will be on the boundary.

For further examples of games exhibiting variationally stable equilibria, please, see examples in [13], [9]. For examples of games from telecommunication or adversarial learning domains with VS Nash equilibria, please, see [9].

Remark 1: Recent works have addressed convergent procedures in games with *restricted strongly monotone* pseudo-gradients [7], [12], [21]. This property is formulated as follows: The pseudo-gradient $\boldsymbol{M}$ is called restricted strongly monotone in the game Γ possessing a Nash equilibrium $\boldsymbol{a}^*$, if $\langle \boldsymbol{M}(\boldsymbol{a}) - \boldsymbol{M}(\boldsymbol{a}^*), \boldsymbol{a}-\boldsymbol{a}^*\rangle \geq \nu\|\boldsymbol{a}-\boldsymbol{a}^*\|^2$ for some $\nu > 0$ and any $\boldsymbol{a}\in\boldsymbol{A}$. If some game satisfies this condition, the Nash equilibrium is unique [21]. It can be seen that, due to the inequality $\langle \boldsymbol{M}(\boldsymbol{a}^*), \boldsymbol{a}-\boldsymbol{a}^*\rangle \geq 0$ holding for any $\boldsymbol{a}\in\boldsymbol{A}$, the Nash equilibrium in a game with a restricted strongly monotone pseudo-gradient is necessarily strongly variationally stable. However, as Example 2 below demonstrates, existence of a strongly variationally stable Nash equilibrium does not imply that the pseudo-gradient is restricted strongly monotone. Thus, games with restricted strongly monotone pseudo-gradients are a subclass of games with strongly variationally stable Nash equilibria considered in this paper.

Example 2: Consider a 2-player game, where each player's action set is $[0,1]\subset\mathbb{R}$. The cost of each player i, for $i\in\{1,2\}$, is $J^i(a^1,a^2) = \frac{1}{4}((a^1)^2+(a^2)^2) - \frac{1}{4}a^1a^2 + 2\sqrt{1+a^1} + 2\sqrt{1+a^2}$. The unique minimizer of the equal cost functions and, thus, the unique Nash equilibrium of the game, is $\boldsymbol{a}^*=(0,0)$. The game pseudo-gradient is given by $\boldsymbol{M}(\boldsymbol{a}) = (\frac{1}{2}a^1 - \frac{1}{4}a^2 + \frac{1}{\sqrt{1+a^1}}, \frac{1}{2}a^2 - \frac{1}{4}a^1 + \frac{1}{\sqrt{1+a^2}})^\top$. It can be verified that $\langle \boldsymbol{M}(\boldsymbol{a}), \boldsymbol{a}-\boldsymbol{a}^*\rangle \geq \nu\|\boldsymbol{a}-\boldsymbol{a}^*\|^2$ with $\nu=\frac{1}{4}$. Thus, the game possesses the unique globally strongly VS Nash equilibrium. However, the pseudo-gradient of the game is not restricted strongly monotone, as $\langle \boldsymbol{M}(\boldsymbol{a}) - \boldsymbol{M}(\boldsymbol{a}^*), \boldsymbol{a}-\boldsymbol{a}^*\rangle < 0$ for $\boldsymbol{a}=(1,1)\in\boldsymbol{A}$.

III. Payoff-Based Learning Algorithm

The steps of the procedure run by each player are summarized in Algorithm 1. In particular, the one-point approach has already been proposed in [17], [19]. However, its convergence properties established in the above works was only for the case of strongly monotone games.

1) Algorithm iterates: Let us denote by $\boldsymbol{m}^i_j$, $j\in\{1,2\}$, some estimate of $\boldsymbol{M}^i$ in the pseudo-gradient of the game (see Definition 2). Here, $j=1$ denotes the one-point and $j=2$ denotes the two-point procedure estimate, respectively, and

Algorithm 1 One-point and two-point zeroth-order algorithm for learning Nash equilibria

Require: Action set $\boldsymbol{A}^i\subset\mathbb{R}^d$, the sequences $\{\sigma_t\},\{\gamma_t\}$, initial state $\boldsymbol{\mu}^i(0)$.

for $t=0,1,\dots$ **do**

Sample a query point $\boldsymbol{\xi}^i(t)$ according to probability density (2).

`/* Simultaneously and similarly, other players sample their query points` $\boldsymbol{\xi}^{-i}(t)$. `/*`

One-point scheme: Perform the one-point gradient estimate according to (3).

Two-point scheme: Observe additional $J^{i0}(t) = J^i(\boldsymbol{\mu}^1(t),\dots,\boldsymbol{\mu}^N(t))$ and perform the two-point gradient estimate according to (4).

Update the state according to (1). `/* Simultaneously and similarly, other players update their states` $\boldsymbol{\mu}^{-i}(t+1)$. `/*`

will be detailed in the next subsection. The proposed method to update player i's so-called state $\boldsymbol{\mu}^i$ is as follows:

$$\boldsymbol{\mu}^i(t+1) = \mathrm{Proj}_{A^i}[\boldsymbol{\mu}^i(t) - \gamma_t \boldsymbol{m}^i_j(t)], \quad (1)$$

where $\boldsymbol{\mu}^i(0)\in\mathbb{R}^{Nd}$ is an arbitrary finite value and γ_t is the step size or the learning rate. The step size γ_t needs to be chosen based on the bias and variance of the pseudo-gradient estimates $\boldsymbol{m}^i_j$. The term $\boldsymbol{m}^i_j(t)$, $j\in\{1,2\}$, is obtained using the payoff-based feedback as described below.

Remark 2: While our algorithm is similar to [2], [17], the reason for being able to establish the stronger result compared to these past works is our new analysis technique. In particular, due to the lack of strong monotonicity, we develop a new approach to estimate the distance between the algorithm iterates and the Nash equilibrium. This approach is based on proving that the Nash equilibrium $\boldsymbol{a}^*$ stays "almost" strongly variationally stable with respect to the pseudo-gradient in the *mixed strategies*, a property we establish in Proposition 1.

2) Gradient estimation in one and two-point settings: We estimate the unknown gradients using the randomizing sampling technique. In particular, we use the Gaussian distribution for sampling inspired by [23], [14]. Since this distribution has an unbounded support, we need the following assumption on the cost functions' behavior at infinity.

Assumption 4: Each function $J^i(\boldsymbol{x}) = O(\exp\{\|\boldsymbol{x}\|^\alpha\})$ as $\|\boldsymbol{x}\|\to\infty$, where $\alpha<2$.

Given $\boldsymbol{\mu}^i(t)$, let player i sample the random vector $\boldsymbol{\xi}^i(t)$ according to the multidimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}^i(t) = (\mu^{i,1}(t),\dots,\mu^{i,d}(t))^\top, \sigma_t)$ with the following density function:

$$p^i(\boldsymbol{x}^i;\boldsymbol{\mu}^i(t),\sigma_t) = \frac{1}{(\sqrt{2\pi}\sigma_t)^d} e^{-\sum_{k=1}^d \frac{(x^{i,k}-\mu^{i,k}(t))^2}{2\sigma_t^2}}. \quad (2)$$

According to the algorithm's setting, the cost value at the query point $\boldsymbol{\xi}(t) = (\boldsymbol{\xi}^1(t),\dots,\boldsymbol{\xi}^N(t))\in\mathbb{R}^{Nd}$, denoted by $J^i(t) := J^i(\boldsymbol{\xi}(t))$, is revealed to each player i. In the one-point setting, player i then estimates her local gradient $\frac{\partial J^i}{\partial \boldsymbol{\mu}^i}$ evaluated at the point of the joint state $\boldsymbol{\mu}(t) = (\boldsymbol{\mu}^1(t), \ldots, \boldsymbol{\mu}^N(t))$ as follows:

$$\boldsymbol{m}_1^i(t) = J^i(t)\frac{\boldsymbol{\xi}^i(t) - \boldsymbol{\mu}^i(t)}{\sigma_t^2}. \tag{3}$$

In the two-point setting at each iteration t, each player i makes two queries: a query corresponding to playing randomly chosen $\boldsymbol{\xi}^i(t)$, $i \in [N]$; and another query of the cost function at $\boldsymbol{\mu}^i(t)$, $i \in [N]$. Hence, there is an extra piece of information available to each player, namely the cost function value at the mean (state) vector $\boldsymbol{\mu}(t)$: $J^{i0}(t) := J^i(\boldsymbol{\mu}(t))$. Then each player uses the following estimation of the local gradient $\frac{\partial J^i}{\partial \boldsymbol{\mu}^i}$ at the point $\boldsymbol{\mu}(t)$:

$$\boldsymbol{m}_2^i(t) = (J^i(t) - J^{i0}(t))\frac{\boldsymbol{\xi}^i(t) - \boldsymbol{\mu}^i(t)}{\sigma_t^2}. \tag{4}$$

Remark 3: Observe that the both estimations $\boldsymbol{m}_j^i(t)$, $j = 1,2$, above can be performed on the feasible set $\boldsymbol{A}$. One can set $\boldsymbol{m}_1^i(t) = J^i(\mathrm{Proj}_{\boldsymbol{A}}\boldsymbol{\xi}(t))\frac{\boldsymbol{\xi}^i(t)-\boldsymbol{\mu}^i(t)}{\sigma_t^2}$ $\boldsymbol{m}_2^i(t) = (J^i(\mathrm{Proj}_{\boldsymbol{A}}\boldsymbol{\xi}(t)) - J^i(\boldsymbol{\mu}(t))\frac{\boldsymbol{\xi}^i(t)-\boldsymbol{\mu}^i(t)}{\sigma_t^2}$, using, thus, cost values at feasible actions, namely at the points $\mathrm{Proj}_{\boldsymbol{A}}\boldsymbol{\xi}(t)$ and $\boldsymbol{\mu}(t) \in \boldsymbol{A}$. However, to guarantee convergence of the algorithm to the Nash equilibrium, an adjustment of the updates in (1) is required. An extra parameter needs to be introduced to project the $\boldsymbol{\mu}(t)$'s on a shrunk set and this parameter has to be balanced with both the step size γ_t and the variance σ_t, see [19] for details of this analysis in the case of monotone games, and [6] for similar consideration in zeroth-order online optimization.

3) Properties of the gradient estimators: We provide insight into the procedure defined by Equation (1) by deriving an analogy to a stochastic gradient algorithm. Denote $p(\boldsymbol{x};\boldsymbol{\mu},\sigma) = \prod_{i=1}^N p^i(x^{i,1},\ldots,x^{i,d};\boldsymbol{\mu}^i,\sigma)$ as the density function of the joint distribution of players' query points $\boldsymbol{\xi}$, given some state $\boldsymbol{\mu} = (\boldsymbol{\mu}^1,\ldots,\boldsymbol{\mu}^N)$. For any $\sigma > 0$ and $i \in [N]$ define $\tilde{J}_i : \mathbb{R}^{Nd} \to \mathbb{R}$ as $\tilde{J}_i(\boldsymbol{\mu},\sigma) = \int_{\mathbb{R}^{Nd}} J^i(\boldsymbol{x})p(\boldsymbol{x};\boldsymbol{\mu},\sigma)d\boldsymbol{x}$. Thus, $\tilde{J}_i$, $i \in [N]$, is the ith player's cost function in the mixed strategies, where the strategies are sampled from the Gaussian distribution with the density function $p(\boldsymbol{x};\boldsymbol{\mu},\sigma)$. For $i \in [N]$ define $\tilde{\boldsymbol{M}}^{i,\sigma}(\cdot) = (\tilde{M}^{i,1,\sigma}(\cdot),\ldots,\tilde{M}^{i,d,\sigma}(\cdot))^\top$ as the d-dimensional mapping with the following elements:

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

Furthermore, let $\boldsymbol{R}_j^i$, $j = 1,2$, denote:

$$\boldsymbol{R}_j^i(\boldsymbol{\xi}(t),\boldsymbol{\mu}(t),\sigma_t) = \boldsymbol{m}_j^i(t) - \tilde{\boldsymbol{M}}^{i,(t)}(\boldsymbol{\mu}(t),\sigma_t), \text{ where}$$
$$\boldsymbol{m}_j^i(t) = \begin{cases} J^i(\boldsymbol{\xi}(t))\frac{\boldsymbol{\xi}^i(t)-\boldsymbol{\mu}^i(t)}{\sigma_t^2}, \text{ if } j = 1, \\ (J^i(\boldsymbol{\xi}(t)) - J^i(\boldsymbol{\mu}(t)))\frac{\boldsymbol{\xi}^i(t)-\boldsymbol{\mu}^i(t)}{\sigma_t^2}, \text{ if } j = 2, \end{cases} \tag{6}$$

where, to simplify notations, we used $\tilde{\boldsymbol{M}}^{(t)} = \tilde{\boldsymbol{M}}^{\sigma_t}$. With the above definitions, the update rule (1) is equivalent to:

$$\boldsymbol{\mu}^i(t+1) = \mathrm{Proj}_{A^i}[\boldsymbol{\mu}^i(t) - \gamma_t(\tilde{\boldsymbol{M}}^{i,(t)}(\boldsymbol{\mu}(t)) + \boldsymbol{R}_j^i(\boldsymbol{\xi}(t),\boldsymbol{\mu}(t),\sigma_t))]. \tag{7}$$

Recall that the cases $j = 1$ and $j = 2$ above correspond to the one-point and two-point gradient estimators, respectively.

We now show that $\tilde{\boldsymbol{M}}^{i,(t)}$ is equal to $\boldsymbol{M}^i$ in expectation and the term $\boldsymbol{R}_j^i$ has a zero-mean. Thus, we can interpret (1) as a stochastic gradient descent procedure.

Let $\mathcal{F}_t$ be the σ-algebra generated by the random variables $\{\boldsymbol{\mu}(k),\boldsymbol{\xi}(k)\}_{k\le t}$. First, we demonstrate in the next lemma that the mapping $\tilde{\boldsymbol{M}}^{(t)} = (\tilde{\boldsymbol{M}}^{1,(t)},\ldots,\tilde{\boldsymbol{M}}^{N,(t)})$ evaluated at $\boldsymbol{\mu}(t)$ is equivalent to the pseudo-gradient in mixed strategies, that is,

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

Moreover, this lemma proves that the conditional expectation of the terms defined in (3) and (4), namely, $\boldsymbol{m}^i(t) = (m_j^{i,1}(t),\ldots,m_j^{i,d}(t)) \in \mathbb{R}^d$, $j = 1,2$, is equal to $\tilde{\boldsymbol{M}}_i$.

Lemma 1: Given Assumptions 1 and 4, $\tilde{M}^{i,k,(t)}(\boldsymbol{\mu}(t)) = \mathrm{E}\{M^{i,k}(\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^N)|\xi^{i,k} \sim \mathcal{N}(\mu^{i,k}(t),\sigma_t), i \in [N], k \in [d]\} = \mathrm{E}\{m_j^{i,k}(t)|\mathcal{F}_t\}$.
The proof of this result is very similar to that of Lemma 1 in [18] for the one-point feedback, and its extension to [19] for two-point feasible feedback. For the sake of notation simplicity, let us use $\boldsymbol{R}(t) = \boldsymbol{R}(\boldsymbol{\xi}(t),\boldsymbol{\mu}(t),\sigma_t)$. Our second lemma below characterizes the variance of the term $\boldsymbol{R}(t)$.

Lemma 2: Under Assumptions 1 and 4, as $\sigma_t \to 0$, $\mathrm{E}\{\|\boldsymbol{R}_j^i(t)\|^2|\mathcal{F}_t\} = O\left(\frac{d}{\sigma_t^2}\right)$, if $j = 1$, and $\mathrm{E}\{\|\boldsymbol{R}_j^i(t)\|^2|\mathcal{F}_t\} = O(Nd^2)$, if $j = 2$.
The proof of this result is similar to one of Lemma 1 in [18].

IV. Convergence Rate of the Algorithm

We will provide the analysis of Algorithm 1 in the cases $j = 1,2$ (one-point and two-point feedback) under the following smoothness assumption.

Assumption 5: We use one of the following assumptions on pseudo-gradient $\boldsymbol{M} : \mathbb{R}^{Nd} \to \mathbb{R}^{Nd}$ (see Definition 2), depending on the one-point or two-point feedback model.

1) The pseudo-gradient $\boldsymbol{M} : \mathbb{R}^{Nd} \to \mathbb{R}^{Nd}$ is twice differentiable over $\mathbb{R}^{Nd}$.
2) The pseudo-gradient $\boldsymbol{M} : \mathbb{R}^{Nd} \to \mathbb{R}^{Nd}$ is Lipschitz continuous over $\mathbb{R}^{Nd}$.

Note that in the case of one-point feedback ($j = 1$), twice differentiability implies that the Jacobian of the pseudo-gradient is Lipschitz continuous over a compact set. This latter condition was employed in [3] for deriving the convergence rate under the strong monotonicity assumption. While do not assume strong monotonicity, we require twice differentiability for the one-point feedback case.

Theorem 1: Let the states $\boldsymbol{\mu}^i(t)$, $i \in [N]$, evolve according to Algorithm 1 with the gradient estimators $\boldsymbol{m}_j^i(t)$, $j = 1,2$. Let Assumptions 1–4 hold. Let Assumption 5.1) hold for $j = 1$ and Assumption 5.2) hold for $j = 2$. Let the step size parameter in the procedure be chosen as follows: $\gamma_t = \frac{c}{t}$ with $c \ge \frac{1}{\nu}$, where ν is the constant from Assumption 3. Moreover, let $\sigma_t = \frac{a}{t^{\frac{1}{4}}}$, if $j = 1$, and $\sigma_t = \frac{a}{t^s}$, if $j = 2$, where $a > 0$ and $s \ge 1$. Then, the joint state

$\boldsymbol{\mu}(t)$ converges almost surely to the unique Nash equilibrium $\boldsymbol{\mu}^* = \boldsymbol{a}^*$ of the game Γ, whereas the joint query point $\boldsymbol{\xi}(t)$ converges in probability to $\boldsymbol{a}^*$. Moreover,

$$\mathrm{E}\|\boldsymbol{\mu}(t) - \boldsymbol{a}^*\|^2 = \begin{cases} O\left(\frac{Nd}{t^{1/2}}\right), & \text{if } j = 1, \\ O\left(\frac{N^2 d^2}{t}\right), & \text{if } j = 2. \end{cases}$$

We will base our analysis on the algorithm's representation in (7). Thus, in this subsection we exploit the properties of the term $\tilde{\boldsymbol{M}}^{i,(t)}(\boldsymbol{\mu}(t))$ therein. Let us now focus on the mapping $\tilde{\boldsymbol{M}}^{(t)}(\cdot) = (\tilde{\boldsymbol{M}}^{1,(t)}(\cdot), \ldots, \tilde{\boldsymbol{M}}^{N,(t)}(\cdot))$, where, as before, for any $\boldsymbol{\mu} \in \mathbb{R}^{Nd}$, $\tilde{\boldsymbol{M}}^{i,(\sigma_t)}(\boldsymbol{\mu}) = \tilde{\boldsymbol{M}}^{i,(t)}(\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}^i} \tilde{J}^i(\boldsymbol{\mu}, \sigma_t) = \int_{\mathbb{R}^{Nd}} \boldsymbol{M}^i(\boldsymbol{x}) p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma_t) d\boldsymbol{x}$ for some given σ_t (see definition in (5) and also the property (8)). We emphasize that such mapping is the pseudo-gradient in the mixed strategies, given that the joint action is generated by the normal distribution with the density $p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma)$.

A technical novelty in deriving the convergence in the absence of strong monotonicity is Proposition 1 below. It states that under the made assumption the Nash equilibrium $\boldsymbol{a}^*$ stays "almost" strongly variationally stable with respect to the pseudo-gradient in the mixed strategies.

Proposition 1: Let Assumptions 2, 3, 4, and 5 hold. Then $\langle \tilde{\boldsymbol{M}}^{(t)}(\boldsymbol{\mu}), \boldsymbol{\mu} - \boldsymbol{a}^* \rangle \geq -O(Nd\sigma_t^2) + \nu \|\boldsymbol{\mu} - \boldsymbol{a}^*\|^2$ for any $\boldsymbol{\mu} \in \boldsymbol{A}$.

Proof: We focus on each term in the following sum representation of the dot-product $\langle \tilde{\boldsymbol{M}}^{(t)}(\boldsymbol{\mu}), \boldsymbol{\mu} - \boldsymbol{a}^* \rangle$:

$$\langle \tilde{\boldsymbol{M}}^{(t)}(\boldsymbol{\mu}), \boldsymbol{\mu} - \boldsymbol{a}^* \rangle = \sum_{i=1}^{N} \sum_{k=1}^{d} \tilde{M}^{i,k,(t)}(\boldsymbol{\mu})(\mu^{i,k} - a^{*,i,k}). \tag{9}$$

Due to Assumption 5, we can use the following Taylor approximation for the elements $M^{i,k}$, $i \in [N]$, $k \in [d]$, of the mapping $\boldsymbol{M}$ around some point $\boldsymbol{\mu} \in \boldsymbol{A}$ (see Definition 2):

$$\begin{aligned} M^{i,k}(\boldsymbol{x}) = & M^{i,k}(\boldsymbol{\mu}) + \langle \nabla M^{i,k}(\boldsymbol{\mu}), \boldsymbol{x} - \boldsymbol{\mu} \rangle \\ & + \langle \nabla^2 M^{i,k}(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{x} - \boldsymbol{\mu} \rangle, \end{aligned} \tag{10}$$

where $\tilde{\boldsymbol{x}} = \boldsymbol{\mu} + \theta(\boldsymbol{x} - \boldsymbol{\mu})$ for some $\theta \in [0, 1]$. Taking into account the fact that $\tilde{M}^{i,k,(t)}(\boldsymbol{\mu}) = \int_{\mathbb{R}^{Nd}} M^{i,k}(\boldsymbol{x}) p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma_t) d\boldsymbol{x}$, we obtain

$$\begin{aligned} & \tilde{M}^{i,k,(t)}(\boldsymbol{\mu})(\mu^{i,k} - a^{*,i,k}) \\ & = \left[\int_{\mathbb{R}^{Nd}} (M^{i,k}(\boldsymbol{x}) - M^{i,k}(\boldsymbol{\mu})) p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma_t) d\boldsymbol{x} \right] (\mu^{i,k} - a^{*,i,k}) \\ & \quad + M^{i,k}(\boldsymbol{\mu})(\mu^{i,k} - a^{*,i,k}) \\ & = \left[\int_{\mathbb{R}^{Nd}} \langle \nabla^2 M^{i,k}(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{x} - \boldsymbol{\mu} \rangle p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma_t) d\boldsymbol{x} \right] \\ & \qquad \times (\mu^{i,k} - a^{*,i,k}) + M^{i,k}(\boldsymbol{\mu})(\mu^{i,k} - a^{*,i,k}), \end{aligned} \tag{11}$$

where in the last equality we used (10) and the fact that $\int_{\mathbb{R}^{Nd}} \langle \nabla M^{i,k}(\boldsymbol{\mu}), \boldsymbol{x} - \boldsymbol{\mu} \rangle p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma_t) d\boldsymbol{x} = 0$. We note that $\int_{\mathbb{R}^{Nd}} \langle \nabla^2 M^{i,k}(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{x} - \boldsymbol{\mu} \rangle p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma_t) d\boldsymbol{x} = \mathrm{E}\{\langle \nabla^2 M^{i,k}(\tilde{\boldsymbol{\xi}})(\boldsymbol{\xi} - \boldsymbol{\mu}), \boldsymbol{\xi} - \boldsymbol{\mu} \rangle\}$, given that $\boldsymbol{\xi}$ has the Gaussian distribution with the density function $p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma_t)$ and $\tilde{\boldsymbol{\xi}} = \boldsymbol{\mu} + \theta(\boldsymbol{\xi} - \boldsymbol{\mu})$. Next,

$$\begin{aligned} & \mathrm{E}\{\langle \nabla^2 M^{i,k}(\tilde{\boldsymbol{\xi}})(\boldsymbol{\xi} - \boldsymbol{\mu}), \boldsymbol{\xi} - \boldsymbol{\mu} \rangle\} \\ & \geq -\mathrm{E}\{\|\nabla^2 M^{i,k}(\tilde{\boldsymbol{\xi}})\| \|\boldsymbol{\xi} - \boldsymbol{\mu}\|^2\} \\ & \geq -\mathrm{E}\{\|\nabla^2 M^{i,k}(\tilde{\boldsymbol{\xi}})\|\} \mathrm{E}\{\|\boldsymbol{\xi} - \boldsymbol{\mu}\|^2\} = -O(Nd\sigma_t^2), \end{aligned} \tag{12}$$

where the first and the second inequalities are due to the Cauchy-Schwarz and the Hölder's ones respectively, and in the equality we use Lemma 5 in [19] stating finiteness of $\mathrm{E}\{\|\nabla^2 M^{i,k}(\tilde{\boldsymbol{\xi}})\|$ given Assumption (4). Combining (9)-(12), we conclude that $\langle \tilde{\boldsymbol{M}}^{(t)}(\boldsymbol{\mu}), \boldsymbol{\mu} - \boldsymbol{a}^* \rangle \geq -O(Nd\sigma_t^2) + \nu \|\boldsymbol{\mu} - \boldsymbol{a}^*\|^2$, where we used $\boldsymbol{\mu}, \boldsymbol{a}^* \in \boldsymbol{A}$ and compactness of $\boldsymbol{A}$ (Assumption 2) and $\langle \boldsymbol{M}(\boldsymbol{\mu}), \boldsymbol{\mu} - \boldsymbol{a}^* \rangle \geq \nu \|\boldsymbol{\mu} - \boldsymbol{a}^*\|^2$ (Assumption 3). ■

We are now equipped to provide the proof of Theorem 1.

Proof: (of Theorem 1) Let us notice that due to the theorem's conditions and the particular choice $\sigma_t \to 0$, as $t \to \infty$, Proposition1 hold.

We consider $\|\boldsymbol{\mu}(t+1) - \boldsymbol{a}^*\|^2$. We aim to bound the growth of $\|\boldsymbol{\mu}(t+1) - \boldsymbol{a}^*\|^2$ in terms of $\|\boldsymbol{\mu}(t) - \boldsymbol{a}^*\|^2$ and, thus, to obtain the convergence rate of the sequence $\|\boldsymbol{\mu}(t+1) - \boldsymbol{a}^*\|^2$. By analyzing each term in the following sum $\|\boldsymbol{\mu}(t+1) - \boldsymbol{a}^*\|^2 = \sum_{i=1}^{N} \|\boldsymbol{\mu}^i(t+1) - \boldsymbol{a}^{i*}\|^2$ and applying Lemmas 1 and 2, we obtain[3]

$$\begin{aligned} & \mathrm{E}\{\|\boldsymbol{\mu}^i(t+1) - \boldsymbol{a}^{i*}\|^2 | \mathcal{F}_t\} \leq \|\boldsymbol{\mu}^i(t) - \boldsymbol{a}^{i*}\|^2 \\ & \quad - 2\gamma_t \langle \tilde{\boldsymbol{M}}^{i,(t)}(\boldsymbol{\mu}(t)), \boldsymbol{\mu}^i(t) - \boldsymbol{a}^{i*} \rangle + h_0(t), \end{aligned} \tag{13}$$

where $h_0(t) = O\left(\frac{d\gamma_t^2}{\sigma_t^2}\right)$, if $j = 1$ and $h_0(t) = O(Nd^2\gamma_t^2)$, if $j = 2$. Thus, applying Proposition 1 in the case $j = 1$, and using the relation $\|\tilde{\boldsymbol{M}}^{i,(t)}(\boldsymbol{\mu}(t)) - \boldsymbol{M}^i(\boldsymbol{\mu}(t))\| = O(Nd\sigma_t)$ (see (20) in [18] for the proof) in the case $j = 2$, we conclude from (13) by summing up the inequalities over $i = 1, \ldots, N$,

$$\mathrm{E}\{\|\boldsymbol{\mu}(t+1) - \boldsymbol{a}^*\|^2 | \mathcal{F}_t\} \leq (1 - \nu\gamma_t) \|\boldsymbol{\mu}(t) - \boldsymbol{a}^*\|^2 + h_1(t),$$

where $h_1(t) = O\left(\frac{Nd\gamma_t^2}{\sigma_t^2} + Nd\gamma_t\sigma_t^2\right)$, if $j = 1$, and $h_1(t) = O(N^2d^2\gamma_t^2 + Nd\gamma_t\sigma_t)$, if $j = 2$. Thus, given the settings for the parameters $\gamma_t = \frac{c}{t}$ with $c \geq \frac{1}{\nu}$ and σ_t, we conclude that $\boldsymbol{\mu}(t)$ converges to $\boldsymbol{a}^*$ almost surely (see Lemma 10 in Chapter 2.2 [16]). Taking into account that $\boldsymbol{\xi}(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \sigma_t)$ and $\sigma_t \to 0$ as $t \to \infty$, we conclude that $\boldsymbol{\xi}(t)$ converges weakly to a Nash equilibrium $\boldsymbol{a}^*$. Moreover, according to the Portmanteau Lemma [11], this convergence is also in probability. Next, by taking the full expectation of the both sides and applying the Chung's lemma (see Lemma 4 in Chapter 2.2 [16]) to the resulting inequality, we obtain the result. ■

V. Numerical Example

We consider Example 1 in Section II. Notice that the game satisfies Assumptions 1–5. In the plots below, we show the convergence of the algorithm using the one-point and two-point feedback. The first plot considers the strategy set $[-1, 2]$, whereas the second plot considers the strategy set $[0, 1]$. The parameters were set to $c = 1$, $a = 1$ in both one-point and two-point settings, and $s = 1$ for the case of the two-point setting. The initial state $\boldsymbol{\mu}(0)$ was chosen from standard normal distribution. As predicted by

[3] The full proof can be found in [20].

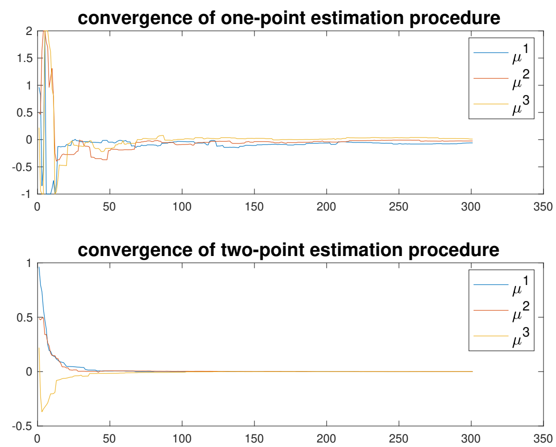

Fig. 1. Convergence of the algorithm for the game of Example 1 with strategy sets as $[-1, 2]$.

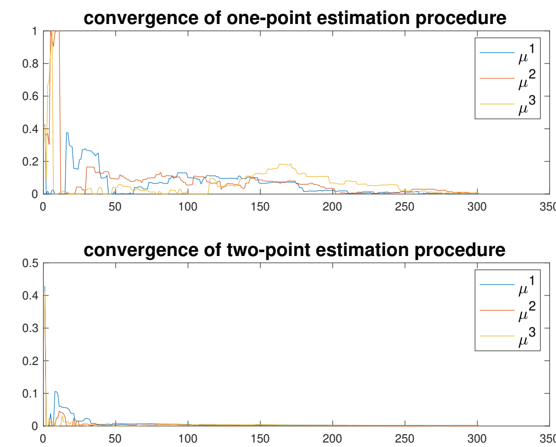

Fig. 2. Convergence of the algorithm for the game of Example 1 with strategy sets as $[0, 1]$.

the theory, in both one-point and two-point settings, the proposed algorithms converge to the unique SVS equilibrium of the game, despite lack of monotonicity. The two-point feedback results in much faster convergence rate, as also predicted.

VI. Conclusion

We established the convergence rate of the zeroth-order gradient play to the globally strongly VS Nash equilibrium of a convex game. In both the one-point and two-point setting, our rates of $O\left(\frac{Nd}{t^{1/2}}\right)$ and $O\left(\frac{N^2d^2}{t}\right)$ are (as functions on t) as they match the best rates established for the subclass of strongly monotone games. An open question is the lower bound for the convergence rate of zeroth-order learning in convex games with respect to the problem dimension. Moreover, it will be interesting to further relax the assumptions so as to establish the convergence and derive the convergence rate of a zeroth-order gradient play to a merely VS equilibrium or to equilibria in non-convex games.

References

[1] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, June 2010.

[2] 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.

[3] D. Drusvyatskiy, M. Fazel, and L. J. Ratliff. Improved rates for derivative free gradient play in strongly monotone games. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 3403–3408. IEEE, 2022.

[4] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61:2788–2806, 2015.

[5] P. S. Dutta, N. R. Jennings, and L. Moreau. Cooperative information sharing to improve distributed learning in multi-agent systems. *CoRR*, abs/1109.5712, 2011.

[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] B. Franci and S. Grammatico. Forward–backward algorithms for stochastic nash equilibrium seeking in restricted strongly and strictly monotone games. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 221–226, 2021.

[8] B. Gao and L. Pavel. Second-order mirror descent: exact convergence beyond strictly stable equilibria in concave games. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 948–953. IEEE, 2021.

[9] B. Gao and L. Pavel. Bandit learning with regularized second-order mirror descent. In *Conference on Decision and Control (CDC)*, pages 5731–5738. IEEE, 2022.

[10] K. G. Jamieson, R. Nowak, and B. Recht. Query complexity of derivative-free optimization. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

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

[12] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, i: Operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[13] P. Mertikopoulos and Z. Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173:465–507, 2019.

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

[15] J.-S. Pang and F. Facchinei. *Finite-dimensional variational inequalities and complementarity problems : vol. 2*. Springer series in operations research. Springer, New York, Berlin, Heidelberg, 2003.

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

[17] T. Tatarenko and M. Kamgarpour. Learning generalized nash equilibria in a class of convex games. *IEEE Transactions on Automatic Control*, 64(4):1426–1439, 2019.

[18] T. Tatarenko and M. Kamgarpour. Learning nash equilibria in monotone games. In *EEE Conference on Decision and Control*, pages 3104–3109. IEEE, 2019.

[19] T. Tatarenko and M. Kamgarpour. On the rate of convergence of payoff-based algorithms to nash equilibrium in strongly monotone games. *ArXiv*, arXiv:2202.11147, 2022.

[20] T. Tatarenko and M. Kamgarpour. Convergence rate of learning a strongly variationally stable equilibrium. *ArXiv*, arXiv:2304.02355, 2024.

[21] T. Tatarenko, W. Shi, and A. Nedić. Geometric convergence of gradient play algorithms for distributed nash equilibrium seeking. *IEEE Transactions on Automatic Control*, 66(11):5342–5353, 2021.

[22] A. C. Tellidou and A. G. Bakirtzis. Agent-based analysis of capacity withholding and tacit collusion in electricity markets. *IEEE Transactions on Power Systems*, 22(4):1735–1742, Nov 2007.

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