Finding First-Order Nash Equilibria of Zero-Sum Games with the Regularized Nikaido-Isoda Function

Ioannis Tsaknakis **Mingyi Hong**
Department of Electrical and Computer Engineering
University of Minnesota, Twin Cities
Minneapolis, MN 55455

Abstract

Efficiently finding First-order Nash Equilibria (FNE) in zero-sum games can be challenging, even in a two-player setting. This work proposes an algorithm for finding the FNEs of a two-player zero-sum game, in which the local cost functions can be non-convex, and the players only have access to local stochastic gradients. The proposed approach is based on reformulating the problem of interest as minimizing the Regularized Nikaido-Isoda (RNI) function. We show that the global minima of the RNI correspond to the set of FNEs, and that for certain classes of non-convex games the RNI minimization problem becomes convex. Moreover, we introduce a first-order (stochastic) optimization method, and establish its convergence to a neighborhood of a stationary solution of the RNI objective. The key in the analysis is to properly control the bias between the local stochastic gradient and the true one. Although the RNI function has been used in analyzing convex games, to our knowledge, this is the first time that the properties of the RNI formulation have been exploited to find FNEs for non-convex games in a stochastic setting.

1 INTRODUCTION

In this work we consider the following *two player zero-sum games*, in which one player (the "min" player) aims to minimize the objective (payoff) $f(\mathbf{x}, \mathbf{y})$, while the other one (the "max" player) aims to minimize the negative of the objective (that is, to maximize it):

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}^*), \quad \mathbf{y}^* = \arg\min_{\mathbf{y}\in\mathcal{Y}} -f(\mathbf{x}^*, \mathbf{y}). \quad (1)$$

Here $f(\mathbf{x}, \mathbf{y})$ can be a usual deterministic function, or can take the form of an expectation over certain realizations of the objective, i.e., $f(\mathbf{x}, \mathbf{y}) := \mathbb{E}[F(\mathbf{x}, \mathbf{y}; w)]$, where w is some random variable. Moreover, the solution $(\mathbf{x}^*, \mathbf{y}^*)$ of the above problem is called the *(global) Nash equilibrium (NE)* of the game (Jin et al., 2020).

Recently, problem (1) has received renewed interest, since it has found many applications in machine learning, such as reinforcement learning (Cai et al., 2019), adversarial learning (Madry et al., 2017; Shafahi et al., 2018) and Generative Adversarial Networks (GAN) (Goodfellow et al., 2014; Arjovsky et al., 2017). However, in these applications, the objective f is typically *non-convex with respect to (w.r.t)* $\mathbf{x}$ *and non-concave w.r.t* $\mathbf{y}$, making it extremely difficult to find the NEs of the resulting games (e.g., the problem is NP-hard in the general case (Jin et al., 2020)). As a result, one has to resort to a relaxation of the above solution concept. To be more precise, a useful alternative to the NE are the solutions that satisfy the first-order stationarity conditions of the min and the max problem in (1). We refer to those as *First-order Nash equilibria (FNE)* (Pang and Scutari, 2011; Nouiehed et al., 2019), and formally define them below as solutions $(\mathbf{x}^*, \mathbf{y}^*)$ that satisfy the following conditions[1]:

$$\langle \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \ \forall \mathbf{x} \in \mathcal{X}, \quad (2)$$

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{y} - \mathbf{y}^* \rangle \leq 0, \ \forall \mathbf{y} \in \mathcal{Y}. \quad (3)$$

Despite such a simple description, the problem of finding FNEs in nonconvex-nonconcave min-max games is challenging, and there has been various research investigating this matter (Pang and Scutari, 2011; Nouiehed et al., 2019; Ostrovskii et al., 2020). Among others, a

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] Note that in unconstrained games (i.e., $\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \mathbb{R}^m$) the FNE conditions for (1) reduce to the stationarity conditions $\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*) = 0, \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*) = 0$.

key challenge here is that it is difficult to find a well-defined merit function, which not only captures the essence of problem (2)-(3), but is also easy to optimize. The major objective of this paper is to find a simple but meaningful way to re-formulate the FNE problem from an optimization perspective, and design efficient (stochastic) algorithms.

1.1 Related Work

Merit Function-Based Approaches for (F)NE. The well-known Nikaido-Isoda (NI) function (Nikaidô et al., 1955) has been used in literature to construct merit functions for finding NEs (Uryas' ev and Rubinstein, 1994). Specifically, for a N-player game (not necessarily zero-sum), with $\mathbf{x}_i$, $\mathcal{X}_i$ and f_i as the strategy, the strategy space and the payoff of the i player, respectively, the NI function is defined as

$$R(\mathbf{x}) = \sum_{i=1}^{N} \left(f_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \inf_{\mathbf{y} \in \mathcal{X}_i} f_i(\mathbf{y}, \mathbf{x}_{-i}) \right), \quad (4)$$

where $\mathbf{x}_{-i} = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots \mathbf{x}_N)$. Nonetheless, the minimization problem in (4) can be potentially difficult to solve and as a result a regularized extension of the original NI function was subsequently proposed, the Regularized Nikaido-Isoda (RNI) function. The RNI is defined as

$$S(\mathbf{x}) = \sum_{i=1}^{N} \left(f_i(\mathbf{x}_i, \mathbf{x}_{-i}) - \min_{\mathbf{y} \in \mathcal{X}_i} \{ f_i(\mathbf{y}, \mathbf{x}_{-i}) + \frac{\alpha}{2} \|\mathbf{y} - \mathbf{x}_i\|^2 \} \right), \quad (5)$$

and it has been used in the context of Nash equilibria (Gürkan and Pang, 2009) and generalized Nash equilibria games (Von Heusinger and Kanzow, 2009a,b; Qu and Zhao, 2013). However, note that in the works that use the RNI formulation the analysis is restricted to the case where the objectives (utilities) are convex functions.

Moreover, another approach is offered by the Gradient Nikaido-Isoda (GNI) (Raghunathan et al., 2019) function, where the min problem in (4) is solved approximately with the application of one gradient step, that is the respective function takes the form

$$Q(\mathbf{x}) = \sum_{i=1}^{N} \left(f_i(\mathbf{x}_i, \mathbf{x}_{-i}) - f_i(\mathbf{x}_i - \alpha \nabla_{\mathbf{x}_i} f_i(\mathbf{x}), \mathbf{x}_{-i}) \right). \quad (6)$$

The GNI formulation offers a simple way to address the minimization problem in (4), however note that the objective Q does not necessarily possess the Lipschitz gradient property (when only the Lipschitz gradient property of f is assumed). This property is crucial for showing convergence to stationary solutions, and in fact in the work of Raghunathan et al. (2019) it is stated in the theorems as an assumption.

We would like to mention that in all the above works that involve an NI-type function, a deterministic setting is considered. This is an important drawback since a stochastic setting (i.e., stochastic objective and algorithm) is crucial in large scale machine learning applications.

A closely related approach to the NI-based ones is provided by the Hamiltonian function. Specifically, we can reformulate the problem of finding the FNEs of (1), with $\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \mathbb{R}^m$ (in this case the FNEs are the stationary solutions of (1)), to the problem of minimizing the following objective:

$$H(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\|^2 + \frac{1}{2} \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|^2. \quad (7)$$

In the work of Abernethy et al. (2019) it is shown that the gradient descent method on H, that is Hamiltonian Gradient Descent (HGD), exhibits last iterate convergence to a stationary point of f, in certain problem classes that may include some nonconvex-nonconcave min-max games. Moreover, in the paper of Loizou et al. (2020) the Hamiltonian approach is extended to a stochastic setting, where the objective is expressed as a finite-sum. However, note that the above formulation cannot be applied directly to constrained problems.

Min-Max Optimization Based Approaches for (F)NE. There are some recent works in the literature of min-max optimization problems that analyze and develop algorithms for nonconvex-(strongly) concave problems (Nouiehed et al., 2019; Lu et al., 2020; Lin et al., 2020a,b; Ostrovskii et al., 2020) or special cases of nonconvex-nonconcave problems (Nouiehed et al., 2019; Liu et al., 2018; Yang et al., 2020; Liu et al., 2019; Grimmer et al., 2020). To be more precise, Liu et al. (2018, 2019) assumes that the respective Minty variational inequality problem has a solution. In the work of Nouiehed et al. (2019) the main assumption is that the game's objective is non-convex in $\mathbf{x}$ and satisfies the Polyak-Lojasiewicz (PL) condition in $\mathbf{y}$, while Yang et al. (2020) assumes that the PL condition holds for both $\mathbf{x}$ and $\mathbf{y}$. Also, in the paper of Grimmer et al. (2020) the convergence of a proximal point algorithm to a stationary point is studied for a problem with objective $f(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) + \mathbf{x}^T Q \mathbf{y} - g(\mathbf{y})$, non-convex functions h, g, and different "sizes" of the coupling term $\mathbf{x}^T Q \mathbf{y}$. The common characteristic of all these works, which they provide theoretical convergence guarantees to first order stationary solutions (i.e., FNEs or stationary points in unconstrained problems), is that they restrict their analysis only to special classes of non-convex games.

In addition, there are works that develop min-max algorithms motivated by applications in machine learning, such as the works by Chavdarova et al. (2020) and Gidel et al. (2018). The former paper develops

a stochastic variance-reduced extragradient algorithm for the min-max game that arises in GAN training. Also, the latter work treats the FNE finding problem from a variational inequality (VI) perspective, and extends known VI methods for use in GAN problems. Finally, there is a set of works (Adolphs et al., 2018; Daskalakis and Panageas, 2018; Mescheder et al., 2017) where the authors study the behavior of min-max algorithms around local NEs, however this approach can only be utilized for establishing local convergence to these points.

1.2 Contributions

Finding first-order stationary solutions is a tractable problem in non-convex optimization problems, under mild assumptions. However, this is not the case in non-convex min-max games where it is observed that standard algorithms (such as gradient descent-ascent with constant stepsize) can potentially fail to attain solutions that satisfy the first order stationarity conditions (i.e., FNEs), even in simple problems (e.g. bilinear) (Daskalakis et al., 2017). Therefore, finding FNEs in non-convex games remains an important problem. For this reason in this work we focus on the latter concept and consider a reduction of the problem of finding the FNEs of the min-max game (1) to a minimization problem. Moreover, we propose a stochastic first-order algorithm for the minimization problem and show convergence to a neighborhood of a stationary solution. The contributions of this work can be summarized as follows :

- Using the RNI function (5) we formulate an optimization problem whose solution set (i.e., its global minima) is equal to the set of FNEs of the two player zero-sum game (1). Although the RNI objective has been utilized before in convex games, to the best of our knowledge, this is the first time that it is used and analyzed in a *non-convex* and *stochastic* setting. Moreover, the power of the proposed formulation is highlighted by the fact that a number of games, such as certain classes of strongly convex, bilinear, and non-convex ones, correspond to a convex minimization problem after the reduction is performed. Finally, among the attractive features of this formulation are the following: 1) its gradient expression does not require the use of the Hessian, 2) it can be directly utilized for games over compact constraint sets.

- We introduce a *stochastic algorithm* for the proposed minimization problem. Note that the stochastic setting we consider is more realistic, since in many applications the large amount of data or the nature of the objective necessitate the use of that type of algorithms. Moreover, we observe that the form of the objective leads naturally to a biased gradient estimator. This characteristic makes the respective analysis more complicated, however by properly controlling the magnitude of the bias term we manage to show convergence to a first-order stationary point of the RNI objective, up to some additive error terms.

Char.\Form.	*RNI*	*GNI*	*Ham.*
No Hessian Required	**YES**	NO	NO
Constraints	**YES**	NO	NO
Lipschitz gradient	**YES**	NO	NO
Stochastic algorithm	**YES**	NO	**YES**
Unbiased Estimator	NO	—	**YES**

Table 1: Summary of the characteristics of the three formulations described above, that is the Gradient Nikaido-Isoda (GNI) (Raghunathan et al., 2019), the Hamiltonian (Ham.) (Abernethy et al., 2019; Loizou et al., 2020) and the Regularized Nikaido-Isoda (RNI) (this work). Starting from the top row we note the following: 1) whether the gradient formula requires the use of the Hessian, 2) whether the formulations can be used for min-max games with constraints, 3) assuming that the objective f of game (1) is Lipschitz gradient, and without any additional assumptions on f, whether the respective formulation also possesses the Lipschitz gradient property 4) whether a stochastic algorithm is available for the respective formulation, 5) if a stochastic algorithm is available, whether the respective gradient estimators are unbiased. The cells that exhibit the desired behavior are highlighted.

In order to highlight the main advantages of this work, we provide in table 1 a comparison of the Gradient Nikaido-Isoda (6), the Hamiltonian (7) and the Regularized Nikaido-Isoda (this work) formulations, along a number of key characteristics, such as the existence of a stochastic algorithm.

2 REGULARIZED NIKAIDO-ISODA (RNI) APPROACH FOR FNE

2.1 RNI Formulation

To begin with, we consider games where the following assumptions hold.

Assumption 1. *The game* (1) *satisfies the following assumptions:*

1. *The objective $f(\mathbf{x}, \mathbf{y})$ is a two times continuously differentiable function.*

2. *The sets $\mathcal{X}$ and $\mathcal{Y}$ are non-empty, convex and compact.*

3. *The function f has Lipschitz continuous gradients in both $\mathbf{x}$ and $\mathbf{y}$, that is for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \mathbf{y}_1, \mathbf{y}_2 \in$*

$\mathcal{Y}$ *it holds that*

$$\begin{aligned}&\|\nabla_{\mathbf{x}} f(\mathbf{x}_1,\mathbf{y}_1)-\nabla_{\mathbf{x}} f(\mathbf{x}_2,\mathbf{y}_2)\|\\&\leq L_x\left(\|\mathbf{x}_1-\mathbf{x}_2\|+\|\mathbf{y}_1-\mathbf{y}_2\|\right),\\&\|\nabla_{\mathbf{y}} f(\mathbf{x}_1,\mathbf{y}_1)-\nabla_{\mathbf{y}} f(\mathbf{x}_2,\mathbf{y}_2)\|\\&\leq L_y\left(\|\mathbf{x}_1-\mathbf{x}_2\|+\|\mathbf{y}_1-\mathbf{y}_2\|\right).\end{aligned}$$

Then, we define the following auxiliary functions:

$$\Phi_x(\mathbf{x},\mathbf{y})=\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y})+\frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\},\tag{8}$$

$$\Phi_y(\mathbf{x},\mathbf{y})=\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x},\mathbf{z})+\frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}.\tag{9}$$

From the Lipschitz gradient assumption of f, it follows that f is an L_x-weakly-convex (in $\mathbf{x}$), L_y-weakly-concave (in $\mathbf{y}$) function. As a result, for $L>\max\{L_x,L_y\}$ the arguments of the problems in (8),(9) are strongly convex. This implies that the respective minimization problems have a unique solution (the existence of the solution is established by the compactness of the sets $\mathcal{X}$ and $\mathcal{Y}$) and thus $\Phi_x(\mathbf{x},\mathbf{y})$ and $\Phi_y(\mathbf{x},\mathbf{y})$ are well-defined. Moreover, using the above elements we define the RNI objective, that is

$$\begin{aligned}P(\mathbf{x},\mathbf{y})&=[f(\mathbf{x},\mathbf{y})-\Phi_x(\mathbf{x},\mathbf{y})]+[-f(\mathbf{x},\mathbf{y})-\Phi_y(\mathbf{x},\mathbf{y})]\\&=-\Phi_x(\mathbf{x},\mathbf{y})-\Phi_y(\mathbf{x},\mathbf{y})\\&=-\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y})+\frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\}\\&\quad-\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x},\mathbf{z})+\frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}.\end{aligned}\tag{10}$$

Then, the optimization problem we propose for finding the FNEs of (1) is

$$\min_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} P(\mathbf{x},\mathbf{y}).\tag{11}$$

Remark 1. *The RNI formulation can be used to reformulate constrained min-max games (over compact sets). This property highlights the flexibility of the RNI formulation, compared to the GNI* (6) *and Hamiltonian* (7) *ones, since the latter formulations cannot be directly used in constrained problems (for a proposed extension of the Hamiltonian method look at the supplementary material, sec. B). Moreover, note that the RNI objective remains well-defined even in the unconstrained case where we have the non-compact sets* $\mathcal{X}=\mathbb{R}^n,\mathcal{Y}=\mathbb{R}^m$*, due to the strong convexity of the involved problems.*

2.2 Properties of the RNI formulation

The RNI formulation has a number of properties that make it suitable for attaining FNEs of non-convex zero-sum games. In the next proposition we present those properties.

Proposition 1. *Suppose that Assumption 1 holds. Then, provided that* $L>\max\{L_x,L_y\}$ *the function* $P(\mathbf{x},\mathbf{y})$ *possesses the following properties:*

1. *The global minimum of* $P(\mathbf{x},\mathbf{y})$ *is* 0.
2. *A point* $(\mathbf{x}^*,\mathbf{y}^*)$ *is a FNE of* (1) *if and only if* $(\mathbf{x}^*,\mathbf{y}^*)$ *is a global minimum of* $P(\mathbf{x},\mathbf{y})$.

Proof. First of all, using the definitions of Φ_x and Φ_y we can see that, for every $\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}$, it holds that

$$\begin{aligned}\Phi_x(\mathbf{x},\mathbf{y})&=\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y})+\frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\}\\&\leq f(\mathbf{x},\mathbf{y})+\frac{L}{2}\|\mathbf{x}-\mathbf{x}\|^2=f(\mathbf{x},\mathbf{y})\\\Phi_y(\mathbf{x},\mathbf{y})&=\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x},\mathbf{z})+\frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}\\&\leq -f(\mathbf{x},\mathbf{y})+\frac{L}{2}\|\mathbf{y}-\mathbf{y}\|^2=-f(\mathbf{x},\mathbf{y}).\end{aligned}$$

Then, it follows that $P(\mathbf{x},\mathbf{y})=-\Phi_x(\mathbf{x},\mathbf{y})-\Phi_y(\mathbf{x},\mathbf{y})\geq -f(\mathbf{x},\mathbf{y})+f(\mathbf{x},\mathbf{y})\geq 0$. Moreover, note that the compactness of the sets $\mathcal{X}$ and $\mathcal{Y}$ and the twice differentiable objective f ensure the existence of an FNE of (1) (Nouiehed et al., 2019, Theorem 2.2, pg.3).

Now, suppose that $(\mathbf{x}^*,\mathbf{y}^*)$ is a FNE of (1). Then, we know that

$$\langle\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*),\mathbf{x}-\mathbf{x}^*\rangle\geq 0,\ \forall\mathbf{x}\in\mathcal{X},\tag{12}$$

$$\langle\nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*),\mathbf{y}-\mathbf{y}^*\rangle\leq 0,\ \forall\mathbf{y}\in\mathcal{Y}.\tag{13}$$

Also, note that the function

$$g(\mathbf{z})=f(\mathbf{z},\mathbf{y}^*)+\frac{L}{2}\|\mathbf{z}-\mathbf{x}^*\|^2$$

is strongly convex (for $L>\max\{L_x,L_y\}$) and additionally it holds that

$$\begin{aligned}&\langle\nabla_{\mathbf{x}} g(\mathbf{x}^*),\mathbf{x}-\mathbf{x}^*\rangle\\&=\langle\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*)+L(\mathbf{x}^*-\mathbf{x}^*),\mathbf{x}-\mathbf{x}^*\rangle\\&=\langle\nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*),\mathbf{x}-\mathbf{x}^*\rangle\geq 0,\forall\mathbf{x}\in\mathcal{X},\end{aligned}$$

where the last inequality follows from (12). As a result, $\mathbf{x}^*$ is a global minimum of g and thus $\Phi_x(\mathbf{x}^*,\mathbf{y}^*)=f(\mathbf{x}^*,\mathbf{y}^*)$. Following a similar reasoning we can show that $\Phi_y(\mathbf{x}^*,\mathbf{y}^*)=-f(\mathbf{x}^*,\mathbf{y}^*)$. Then, $P(\mathbf{x}^*,\mathbf{y}^*)=-f(\mathbf{x}^*,\mathbf{y}^*)+f(\mathbf{x}^*,\mathbf{y}^*)=0$. The latter equation combined with the fact that $P(\mathbf{x},\mathbf{y})\geq 0$, implies that $(\mathbf{x}^*,\mathbf{y}^*)$ is a global minimum of P. At the same time it implies that the global minimum of P is 0 (i.e., the lower bound in $P(\mathbf{x},\mathbf{y})\geq 0$ is attained), which proves property 1.

Next, consider the opposite direction, that is suppose that $(\mathbf{x}^*,\mathbf{y}^*)$ is a global minimum of P. Then, $P(\mathbf{x}^*,\mathbf{y}^*)=0\Rightarrow\Phi_x(\mathbf{x}^*,\mathbf{y}^*)=-\Phi_y(\mathbf{x}^*,\mathbf{y}^*)$. Moreover, we know that $\Phi_x(\mathbf{x}^*,\mathbf{y}^*)\leq f(\mathbf{x}^*,\mathbf{y}^*)$ and

$\Phi_y(\mathbf{x}^*,\mathbf{y}^*) \leq -f(\mathbf{x}^*,\mathbf{y}^*)$. Therefore, $f(\mathbf{x}^*,\mathbf{y}^*) \geq \Phi_x(\mathbf{x}^*,\mathbf{y}^*) = -\Phi_y(\mathbf{x}^*,\mathbf{y}^*) \geq f(\mathbf{x}^*,\mathbf{y}^*)$, which implies that $\Phi_x(\mathbf{x}^*,\mathbf{y}^*) = f(\mathbf{x}^*,\mathbf{y}^*)$ and $\Phi_y(\mathbf{x}^*,\mathbf{y}^*) = -f(\mathbf{x}^*,\mathbf{y}^*)$. Thus, using the strong convexity of the problems involved in Φ_x and Φ_y we conclude that

$$\begin{aligned}
&\mathbf{x}^* = \arg\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y}^*) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}^*\|^2\} \Rightarrow \\
&\langle \nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*) + L(\mathbf{x}^*-\mathbf{x}^*), \mathbf{x}-\mathbf{x}^*\rangle \geq 0 \\
&\Rightarrow \langle \nabla_{\mathbf{x}} f(\mathbf{x}^*,\mathbf{y}^*), \mathbf{x}-\mathbf{x}^*\rangle \geq 0, \forall \mathbf{x}\in\mathcal{X}. \\
&\mathbf{y}^* = \arg\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x}^*,\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}^*\|^2\} \Rightarrow \\
&\langle -\nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*) + L(\mathbf{y}^*-\mathbf{y}^*), \mathbf{y}-\mathbf{y}^*\rangle \geq 0 \\
&\Rightarrow \langle \nabla_{\mathbf{y}} f(\mathbf{x}^*,\mathbf{y}^*), \mathbf{y}-\mathbf{y}^*\rangle \leq 0, \forall \mathbf{y}\in\mathcal{Y}.
\end{aligned}$$

Consequently, $(\mathbf{x}^*,\mathbf{y}^*)$ is an FNE of (1). □

Notice that computing the gradient of $P(\mathbf{x},\mathbf{y})$ is not straightforward, due to the existence of the "arg min" operator. However, under certain assumptions we can obtain the following result.

Lemma 1. *Suppose that Assumption 1 holds. Then, provided that $L > \max\{L_x, L_y\}$, the gradients of P are*

$$\nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) = L(\overline{\mathbf{x}}-\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}}), \tag{14}$$
$$\nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) = L(\overline{\mathbf{y}}-\mathbf{y}) - \nabla_{\mathbf{y}} f(\overline{\mathbf{x}},\mathbf{y}), \tag{15}$$

where $\overline{\mathbf{x}} = \overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) = \arg\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\}$ and $\overline{\mathbf{y}} = \overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) = \arg\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x},\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}$.

Proof. To begin with, we define the functions

$$\begin{aligned}
g_x(\mathbf{z},\mathbf{x},\mathbf{y}) &= -f(\mathbf{z},\mathbf{y}) - \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2 \\
g_y(\mathbf{z},\mathbf{x},\mathbf{y}) &= f(\mathbf{x},\mathbf{z}) - \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2.
\end{aligned}$$

Thus, we have $\Phi_x(\mathbf{x},\mathbf{y}) = -\max_{\mathbf{z}\in\mathcal{X}}\{g_x(\mathbf{z},\mathbf{x},\mathbf{y})\}$ and $\Phi_y(\mathbf{x},\mathbf{y}) = -\max_{\mathbf{z}\in\mathcal{Y}}\{g_y(\mathbf{z},\mathbf{x},\mathbf{y})\}$. Moreover, note that the sets $\mathcal{X}$ and $\mathcal{Y}$ are compact, and the functions g_x and g_y are differentiable. Also, the Lipschitz gradient property of f implies that

$$\begin{aligned}
-L_x I \preceq \nabla^2_{\mathbf{xx}} f(\mathbf{x},\mathbf{y}) \preceq L_x I \\
-L_y I \preceq \nabla^2_{\mathbf{yy}} f(\mathbf{x},\mathbf{y}) \preceq L_y I
\end{aligned}$$

and as a result for $L > \max\{L_x, L_y\}$ the functions g_x and g_y are strongly concave in $\mathbf{z}$. The last property implies that the problems $\max_{\mathbf{z}\in\mathcal{X}}\{g_x(\mathbf{z},\mathbf{x},\mathbf{y})\}$ and $\max_{\mathbf{z}\in\mathcal{Y}}\{g_y(\mathbf{z},\mathbf{x},\mathbf{y})\}$ admit unique solutions, which we denote with

$$\begin{aligned}
\overline{\mathbf{x}} = \overline{\mathbf{x}}(\mathbf{x},\mathbf{y}) &= \arg\max_{\mathbf{z}\in\mathcal{X}}\{g_x(\mathbf{z},\mathbf{x},\mathbf{y})\} \\
&= \arg\min_{\mathbf{z}\in\mathcal{X}}\{f(\mathbf{z},\mathbf{y}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\} \\
\overline{\mathbf{y}} = \overline{\mathbf{y}}(\mathbf{x},\mathbf{y}) &= \arg\max_{\mathbf{z}\in\mathcal{Y}}\{g_y(\mathbf{z},\mathbf{x},\mathbf{y})\} \\
&= \arg\min_{\mathbf{z}\in\mathcal{Y}}\{-f(\mathbf{x},\mathbf{z}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\},
\end{aligned}$$

respectively.

Then, Danskin's theorem (Bernhard and Rapaport, 1995), (Barazandeh and Razaviyayn, 2020, Theorem 1) implies that

$$\begin{aligned}
\nabla\Phi_x(\mathbf{x},\mathbf{y}) &= \begin{bmatrix}\nabla_{\mathbf{x}}\Phi_x(\mathbf{x},\mathbf{y}) \\ \nabla_{\mathbf{y}}\Phi_x(\mathbf{x},\mathbf{y})\end{bmatrix} = -\begin{bmatrix}\nabla_{\mathbf{x}} g_x(\overline{\mathbf{x}},\mathbf{x},\mathbf{y}) \\ \nabla_{\mathbf{y}} g_x(\overline{\mathbf{x}},\mathbf{x},\mathbf{y})\end{bmatrix} \\
&= -\begin{bmatrix} L(\overline{\mathbf{x}}-\mathbf{x}) \\ -\nabla_{\mathbf{y}} f(\overline{\mathbf{x}},\mathbf{y})\end{bmatrix},
\end{aligned}$$

$$\begin{aligned}
\nabla\Phi_y(\mathbf{x},\mathbf{y}) &= \begin{bmatrix}\nabla_{\mathbf{x}}\Phi_y(\mathbf{x},\mathbf{y}) \\ \nabla_{\mathbf{y}}\Phi_y(\mathbf{x},\mathbf{y})\end{bmatrix} = -\begin{bmatrix}\nabla_{\mathbf{x}} g_y(\overline{\mathbf{y}},\mathbf{x},\mathbf{y}) \\ \nabla_{\mathbf{y}} g_y(\overline{\mathbf{y}},\mathbf{x},\mathbf{y})\end{bmatrix} \\
&= -\begin{bmatrix}\nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}}) \\ L(\overline{\mathbf{y}}-\mathbf{y})\end{bmatrix}.
\end{aligned}$$

Combining the above we obtain

$$\begin{aligned}
\nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) &= -\nabla_{\mathbf{x}}\Phi_x(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}}\Phi_y(\mathbf{x},\mathbf{y}) \\
&= L(\overline{\mathbf{x}}-\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x},\overline{\mathbf{y}}) \\
\nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) &= -\nabla_{\mathbf{y}}\Phi_y(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}}\Phi_x(\mathbf{x},\mathbf{y}) \\
&= L(\overline{\mathbf{y}}-\mathbf{y}) - \nabla_{\mathbf{y}} f(\overline{\mathbf{x}},\mathbf{y}).
\end{aligned}$$

This concludes the proof. □

Remark 2. *The main results of the RNI, GNI (Raghunathan et al., 2019) and Hamiltonian (Abernethy et al., 2019) formulations are derived under the assumption that f is a two times differentiable function. However, differently from the GNI and the Hamiltonian formulations, the gradient formula of the RNI objective does not require the use of the Hessian of f.*

Moreover, among the advantages of this reformulation is that it reduces the FNE finding problem of certain strongly convex, bilinear and even some non-convex min-max games (with some "strong coupling" condition) to a convex optimization problem.

Example 1.

1. *Let $f(\mathbf{x},\mathbf{y}) = h(\mathbf{x}) + \mathbf{x}^T Q\mathbf{y} - g(\mathbf{y})$. Then, consider the following cases:*
   - *(a) (Constrained) Strongly-convex strongly-concave objective : $h(\mathbf{x})$ and $g(\mathbf{x})$ are strongly convex functions.*
   - *(b) (Unconstrained) Bilinear objective : $h(\mathbf{x}) = g(\mathbf{y}) = 0$; $\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \mathbb{R}^m$.*

(c) (Unconstrained) Strongly-convex strongly-convex objective : $h(\mathbf{x})$ is a strongly convex function, $g(\mathbf{y})$ is a strongly concave function; $\mathcal{X} = \mathbb{R}^n, \mathcal{Y} = \mathbb{R}^m$.

For each of the above problem classes, the RNI reformulated objective (10) *is a convex function.*

2. *Consider an unconstrained objective $f(\mathbf{x}, \mathbf{y}), \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ with $n = m$,*

$$\sigma^2_{min}(\nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y})) \geq \lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \quad (16)$$

$$\sigma^2_{max}(\nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y})) \leq \Lambda, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m, \quad (17)$$

and for which the following conditions are satisfied:

$$a) \ -L_x I \preceq \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) \preceq L_x I, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \quad (18)$$

$$b) \ -L_y I \preceq \nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \preceq L_y I, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \quad (19)$$

$$c) \ \lambda \geq \max \left\{ \frac{L + L_y}{L + L_x} \left(L_x^2 + 2LL_x \right), \frac{L + L_x}{L + L_y} \left(L_y^2 + 2LL_y \right) \right\}, \quad (20)$$

where $L > max\{L_x, L_y\}$ is the parameter of the RNI formulation (10)*. In addition, assume that $L_x = L_y = \widetilde{L}$, while we set $L = 2\widetilde{L}$ in* (10)*. Then, if*

$$\left[\lambda - 5\widetilde{L}^2 \right]^2 - 144\widetilde{L}^2 \Lambda > 0 \quad (21)$$

holds the RNI reformulated objective (10) *is a strongly convex function.*

Note that there exist nonconvex-nonconcave min-max games that belong in the above problem class. For instance, the (non-convex) quadratic function $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^T A \mathbf{x} + \mathbf{x}^T Q \mathbf{y} + \frac{1}{2}\mathbf{y}^T B \mathbf{y} + \mathbf{c}^T \mathbf{x} + \mathbf{d}^T \mathbf{y} + \mathbf{e}$, $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n$ with $-\widetilde{L} I \preceq A \preceq \widetilde{L} I$, $-\widetilde{L} I \preceq B \preceq \widetilde{L} I$, and for which the inequalities (21) *and $\lambda = \sigma^2_{min}(Q) \geq 5\widetilde{L}^2$ hold, satisfies the above conditions.*

Proof. See supplementary material, sec. A.1. □

Finally, a key property that is needed in order to establish convergence to stationary solutions is the Lipschitz gradient property of the objective.

Proposition 2. *Suppose that Assumption 1 holds. Then, provided that $L > \max\{L_x, L_y\}$, P has Lipschitz continuous gradients in $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, that is*

$$\|\nabla P(\mathbf{z}_1) - \nabla P(\mathbf{z}_2)\| \leq \overline{L} \|\mathbf{z}_1 - \mathbf{z}_2\|, \ \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{X} \times \mathcal{Y},$$

with constant $\overline{L} = \overline{L}_x + \overline{L}_y$, $\overline{L}_x = L + L_x + \frac{L^2 + LL_y}{L - L_x} + \frac{L_x L_y + LL_y}{L - L_y}$, $\overline{L}_y = L + L_y + \frac{L^2 + LL_x}{L - L_y} + \frac{L_x L_y + LL_x}{L - L_x}$.

Proof. See supplementary material, sec. A.1. □

Remark 3. *The Lipschitz gradient property of $f(\mathbf{x}, \mathbf{y})$, along with the compactness of the constraint sets, suffice to ensure the Lipschitz gradient property for the reformulated objective $P(\mathbf{x}, \mathbf{y})$. On the contrary, for the GNI* (6) *and Hamiltonian* (7) *formulations more assumptions are required in order to establish the same property. For instance, in the work of Abernethy et al. (2019), in addition to the Lipschitz gradient property of f, it is also assumed that f possesses bounded gradients and its Jacobian (i.e., the gradient of the vector $\xi = (\nabla_{\mathbf{x}} f, -\nabla_{\mathbf{y}} f)$) is Lipschitz.*

3 PROPOSED ALGORITHM AND THEORETICAL ANALYSIS

3.1 Preliminaries

In this section we study the proposed formulation in a stochastic setting, that is we assume that the objective f is expressed as $f(\mathbf{x}, \mathbf{y}) = \mathbb{E}_w[F(\mathbf{x}, \mathbf{y}; w)]$, where $F(\mathbf{x}, \mathbf{y}; w)$ is the stochastic oracle at $(\mathbf{x}, \mathbf{y})$ and w is a random variable drawn from some distribution $\mathcal{W}$. Before we proceed, we provide below the basic assumptions for problem (1) and the stochastic oracle F that will hold in the following analysis.

Assumption 2. *The objective*

$$f(\mathbf{x}, \mathbf{y}) = \mathbb{E}_w[F(\mathbf{x}, \mathbf{y}; w)], w \sim \mathcal{W},$$

of the game (1) *satisfies the following assumptions :*

1. *The function $F(\mathbf{x}, \mathbf{y}; w)$ is a two times continuously differentiable function.*
2. *The sets $\mathcal{X}$ and $\mathcal{Y}$ are non-empty, convex and compact with diameter D.*
3. *The function F has Lipschitz continuous gradients in both $\mathbf{x}$ and $\mathbf{y}$, that is for every $w \sim \mathcal{W}$ and for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$*

$$\|\nabla_{\mathbf{x}} F(\mathbf{x}_1, \mathbf{y}_1; w) - \nabla_{\mathbf{x}} F(\mathbf{x}_2, \mathbf{y}_2; w)\| \leq \leq L_x \left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\| \right),$$

$$\|\nabla_{\mathbf{y}} F(\mathbf{x}_1, \mathbf{y}_1; w) - \nabla_{\mathbf{y}} F(\mathbf{x}_2, \mathbf{y}_2; w)\| \leq \leq L_y \left(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\| \right).$$

Remark 4. *The assumptions 1 and 3 given above imply that the function f also satisfies assumptions 1 and 3, respectively, in Assumption 1. Then, the Lipschitz gradient constants of f are L_x and L_y w.r.t. $\mathbf{x}$ and $\mathbf{y}$, respectively.*

Assumption 3. *The stochastic oracle of the function $f(\mathbf{x}, \mathbf{y}) = \mathbb{E}_w[F(\mathbf{x}, \mathbf{y}; w)], w \sim \mathcal{W}$, and its gradient satisfy, for every $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$, the following assumptions:*

1. $\mathbb{E}[\nabla_{\mathbf{x}} F(\mathbf{x},\mathbf{y};w)] = \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$,
   $\mathbb{E}[\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y};w)] = \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$.
2. $\mathbb{E}[\|\nabla_{\mathbf{x}} F(\mathbf{x},\mathbf{y};w) - \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|^2] \leq \sigma^2$,
   $\mathbb{E}[\|\nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y};w) - \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|^2] \leq \sigma^2$.

3.2 Algorithm

In order to solve problem (11) we propose a first-order stochastic optimization method, the *Regularized Nikaido-Isoda Stochastic Gradient Descent (RNI-SGD)* algorithm. This algorithm has access to the stochastic oracle F and it uses the following mini-batch estimators in order to approximate the values of the objective and its gradient at a given point $(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$,

$$\widetilde{f}(\mathbf{x},\mathbf{y};\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} F(\mathbf{x},\mathbf{y};w_i) \tag{22}$$

$$\widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x},\mathbf{y};\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{x}} F(\mathbf{x},\mathbf{y};w_i), \tag{23}$$

$$\widetilde{\nabla}_{\mathbf{y}} f(\mathbf{x},\mathbf{y};\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{y}} F(\mathbf{x},\mathbf{y};w_i), \tag{24}$$

where $\mathbf{w} = (w_1,\dots,w_n)$ with $w_i \sim \mathcal{W}, \forall i$. Using the above mini-batch estimators, the stochastic estimators for the gradient of P admit the form

$$\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) = L(\widetilde{\mathbf{x}} - \mathbf{x}) + \widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x},\widetilde{\mathbf{y}};\widetilde{\mathbf{w}}), \tag{25}$$

$$\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) = L(\widetilde{\mathbf{y}} - \mathbf{y}) - \widetilde{\nabla}_{\mathbf{y}} f(\widetilde{\mathbf{x}},\mathbf{y};\widetilde{\mathbf{w}}), \tag{26}$$

where $\widetilde{\mathbf{x}} = \widetilde{\mathbf{x}}(\mathbf{x},\mathbf{y}) = \arg\min_{\mathbf{z}\in\mathcal{X}}\{\tilde{f}(\mathbf{z},\mathbf{y};\mathbf{w}) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}\|^2\}$, $\widetilde{\mathbf{y}} = \widetilde{\mathbf{y}}(\mathbf{x},\mathbf{y}) = \arg\min_{\mathbf{z}\in\mathcal{Y}}\{-\widetilde{f}(\mathbf{x},\mathbf{z};\mathbf{w}) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}\|^2\}$, and the elements of $\mathbf{w},\widetilde{\mathbf{w}}$ are sampled for $\mathcal{W}$.

Informally, the RNI-SGD algorithm works by performing at each iteration one projected (stochastic) gradient descent step on $P(\mathbf{x},\mathbf{y})$, (w.r.t both $\mathbf{x}$ and $\mathbf{y}$), using the gradients in (25). Therefore, before this step the solutions of the subproblems involved in the computation of $\nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y})$ and $\nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y})$ are needed. We assume that these subproblems are solved to a given accuracy using known methods, such as the projected gradient descent method. Note that this is a reasonable assumption since those problems are tractable strongly convex tasks. Finally, the complete description of the RNI-SGD algorithm is provided on Algorithm 1.

3.3 Theoretical Analysis

The expressions in Assumption 3, that is the unbiasedness of the oracle F and its bounded variance are standard in literature and directly imply the same properties for the mini-batch estimators (22)-(24) (see Lemma 2 in sec, A.2). However, note that the stochastic gradient estimator of $P(\mathbf{x},\mathbf{y})$ (25)-(26), which is formulated by plugging the estimators (22)-(24) into its gradient formula (14), is biased.

Algorithm 1 Regularized Nikaido-Isoda Stochastic Gradient Descent (RNI-SGD)

Input : $\mathbf{x}^0, \mathbf{y}^0, \alpha, \delta_x, \delta_y, L$
for $r = 0,\dots,T-1$ **do**
  Sample $\mathbf{w}^r, \widetilde{\mathbf{w}}^r \sim \mathcal{W}$
  Find $\widehat{\mathbf{x}}^r$ s.t $\|\widehat{\mathbf{x}}^r - \widetilde{\mathbf{x}}^r\| \leq \delta_x$
   with $\widetilde{\mathbf{x}}^r = \arg\min_{\mathbf{z}\in\mathcal{X}}\{\widetilde{f}(\mathbf{z},\mathbf{y}^r;\mathbf{w}^r) + \frac{L}{2}\|\mathbf{z}-\mathbf{x}^r\|^2\}$
  Find $\widehat{\mathbf{y}}^r$ s.t $\|\widehat{\mathbf{y}}^r - \widetilde{\mathbf{y}}^r\| \leq \delta_y$
   with $\widetilde{\mathbf{y}}^r = \arg\min_{\mathbf{z}\in\mathcal{Y}}\{-\widetilde{f}(\mathbf{x}^r,\mathbf{z};\mathbf{w}^r) + \frac{L}{2}\|\mathbf{z}-\mathbf{y}^r\|^2\}$
  $\mathbf{x}^{r+1} = proj_{\mathcal{X}}\left(\mathbf{x}^r - \alpha[L(\widehat{\mathbf{x}}^r - \mathbf{x}^r) + \widetilde{\nabla}_{\mathbf{x}} f(\mathbf{x}^r,\widehat{\mathbf{y}}^r;\widetilde{\mathbf{w}}^r)]\right)$
  $\mathbf{y}^{r+1} = proj_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha[L(\widehat{\mathbf{y}}^r - \mathbf{y}^r) - \widetilde{\nabla}_{\mathbf{y}} f(\widehat{\mathbf{x}}^r,\mathbf{y}^r;\widetilde{\mathbf{w}}^r)]\right)$
end for
Output : $\mathbf{x}^T, \mathbf{y}^T$

Proposition 3. *Suppose that Assumption 2 and 3 hold. For the stochastic gradient estimator of P defined in* (25)-(26) *and for every $\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}$ and $w\sim\mathcal{W}$, it holds that*

$$\mathbb{E}[\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x},\mathbf{y})] = \nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) + \mathbf{e}_x,$$
$$\mathbb{E}[\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x},\mathbf{y})] = \nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) + \mathbf{e}_y,$$
$$\text{with } \|\mathbf{e}_x\| \leq \frac{L\sigma}{(L-L_x)\sqrt{n}} + \frac{L_x\sigma}{(L-L_y)\sqrt{n}}$$
$$\|\mathbf{e}_y\| \leq \frac{L\sigma}{(L-L_y)\sqrt{n}} + \frac{L_y\sigma}{(L-L_x)\sqrt{n}}.$$

Moreover, for every $\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}$, we have that

$$\mathbb{E}[\|\widetilde{\nabla}_{\mathbf{x}} P(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} P(\mathbf{x},\mathbf{y})\|^2] \leq \widetilde{\sigma}_x^2,$$
$$\mathbb{E}[\|\widetilde{\nabla}_{\mathbf{y}} P(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} P(\mathbf{x},\mathbf{y})\|^2] \leq \widetilde{\sigma}_y^2,$$

where $\widetilde{\sigma}_x^2 = \frac{\sigma^2}{n}\left(\frac{2L^2}{(L-L_x)^2} + \frac{4L_x^2}{(L-L_y)^2} + 1\right)$ *and* $\widetilde{\sigma}_y^2 = \frac{\sigma^2}{n}\left(\frac{2L^2}{(L-L_y)^2} + \frac{4L_y^2}{(L-L_x)^2} + 1\right)$.

Proof. See supplementary material, sec. A.2. □

The use of biased stochastic estimators is usually something we would like to avoid. However, the bound we provide above for the norm of the bias, between the true gradient and the stochastic gradient estimator, will allow us to bypass this obstacle, and show convergence.

Furthermore, in this work we consider a problem with constraints and as a result we measure the distance of the iterate at iteration r from a first-order stationary

point using the following optimality criterion

$$\mathbf{G}_a^r = \frac{1}{\alpha}\begin{bmatrix}\mathbf{x}^r - \text{proj}_{\mathcal{X}}\left(\mathbf{x}^r - \alpha\nabla_{\mathbf{x}}P(\mathbf{x}^r,\mathbf{y}^r)\right)\\ \mathbf{y}^r - \text{proj}_{\mathcal{Y}}\left(\mathbf{y}^r - \alpha\nabla_{\mathbf{y}}P(\mathbf{x}^r,\mathbf{y}^r)\right)\end{bmatrix}. \quad (27)$$

Then, $(\mathbf{x}^r, \mathbf{y}^r)$ is an ϵ-stationary point of P if it holds that $\|\mathbf{G}_a^r\| \leq \epsilon$. Notice that when $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$ (and for $\alpha = 1$) the respective conditions reduce to the standard stationarity gap of (unconstrained) non-convex optimization, that is $\|\nabla P(\mathbf{x}^r, \mathbf{y}^r)\| \leq \epsilon$.

Ideally, we would like to attain a global minimum of $P(\mathbf{x}, \mathbf{y})$ (which would ensure that we found an FNE of the game (1)), however the non-convex nature of the objective prevents us from achieving that in the general case. Therefore, we show convergence of the RNI-SGD algorithm to a neighborhood of a first-order stationary point of $P(\mathbf{x}, \mathbf{y})$.

Theorem 1. *Suppose that Assumption 2 and 3 hold. In addition, assume that the gradients of f are bounded, that is $\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\| \leq c_x$ and $\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\| \leq c_y$, for every $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$. We run Algorithm 1 for T iterations, with constant stepisize $0 \leq \alpha < 2/3\overline{L}$, $L > \max\{L_x, L_y\}$ and for given parameters δ_x, δ_y. Then, we have*

$$\frac{1}{T}\sum_{r=0}^{T-1}\mathbb{E}[\|\mathbf{G}_\alpha^r\|^2] \leq \frac{\mathbb{E}[P^0]/\overline{a}}{T} + \frac{3\overline{L}\alpha^2}{2\overline{\alpha}}\delta^2 + \frac{c\alpha}{\overline{\alpha}}\delta + \frac{c\alpha}{\overline{\alpha}}\hat{\sigma} + \frac{3\overline{L}\alpha^2}{\overline{\alpha}}\widetilde{\sigma}^2,$$

where $\overline{\alpha} = \alpha - \frac{3\overline{L}\alpha^2}{2}$, $\delta = 2(L + L_y)\delta_x + 2(L + L_x)\delta_y$, $\hat{\sigma} = \widetilde{\sigma}_x + \widetilde{\sigma}_y$, $\widetilde{\sigma}^2 = 2(\widetilde{\sigma}_x^2 + \widetilde{\sigma}_y^2)$ and $c = 4LD + c_x + c_y$.

Proof. See supplementary material A.2. □

The above theorem implies that Algorithm 1 converges to a neighborhood around a stationary point (in the sense of definition (27)) at a rate of $\frac{1}{T}$, up to some additive error terms, due to the inexact solution of the problems in (8),(9) and the stochastic nature of the objective. Although the above result does not ensure convergence to an FNE of (1), in the general non-convex case (since we cannot guarantee convergence to a global minimum of P), it is more general than the results presented in other related works (and which might offer stronger guarantees). For instance, the results provided by Abernethy et al. (2019) apply only to specific problem classes, such as bilinear or "sufficiently bilinear" games. Also, in the work of Raghunathan et al. (2019) the respective results hold under the assumption that the reformulated objective (6) is Lipschitz gradient, since the latter property does not follow directly from the respective property of f. Finally, differently than the above works our convergence results can be applied to problems with constraints.

Moreover, note that there are interesting problem classes, (for instance, some of the examples provided in Example 1), that make the objective $P(\mathbf{x}, \mathbf{y})$ a strongly convex function. In that case theorem 1 implies convergence of Algorithm 1, in the objective value sense, to a neighborhood around a global min of $P(\mathbf{x}, \mathbf{y})$, for which we know that it corresponds to an FNE of (1).

Another interesting setting arises in the special case where the objective f is deterministic. Then, two out of the four additive terms in theorem 1 become zero (the ones involving $\widetilde{\sigma}$ and $\hat{\sigma}$), while by solving problems (8),(9) very accurately (this is attainable since both require the solution of a strongly convex problem) the other two additive terms (which involve δ) can become very small. In other words, we still have convergence to a neighborhood around a stationary point, but its diameter can become very small (and we can control how small).

4 NUMERICAL EXPERIMENTS

In this section we conduct a number of experiments in order to illustrate the utility of the proposed method and compare it with other relevant algorithms. Specifically, we consider the following algorithms (with constant stepsize): i) Gradient Descent-Ascent (GDA), ii) Optimistic Gradient Descent-Ascent (OGDA)(Daskalakis and Panageas, 2018), iii) Extragradient method (EG)(Mokhtari et al., 2019b), iv) Regularized Nikaido-Isoda-Stochastic Gradient Descent (RNI-SGD) (proposed method). Moreover, we consider the problem of finding a stationary/FNE point of a finite sum objective $f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{10} f_i(\mathbf{x}, \mathbf{y})$, where the f_is are either bilinear or quadratic functions. The dimension of the problem is $n = m = 5$, i.e., $\mathbf{x}, \mathbf{y} \in \mathbb{R}^5$, unless otherwise stated. Finally, the performance measure we are using is the distance of the iterates from the stationary point.

In the case of the bilinear objective we have

$$f_i(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T Q_i \mathbf{y},$$

where $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^n$. The entries of the matrix Q_i are generated at random from a normal Gaussian distribution, and the stepsizes are selected after a few trials, in order to guarantee convergence of both algorithms. In figure 1a we plot the trajectory of the iterates of EG and RNI-SGD for a simple bilinear problem of dimension $n = m = 1$. Also, we set $L = 10 \cdot \max\{L_x, L_y\}$ for RNI-SGD, while the subproblems (8),(9) are solved using 5 steps of the projected gradient descent algorithm. Note that our aim in this experiment is not to compare the algorithms in terms of convergence speed, but rather to illustrate the general behavior (trajectory) of RNI-SGD compared to other classical algorithms.

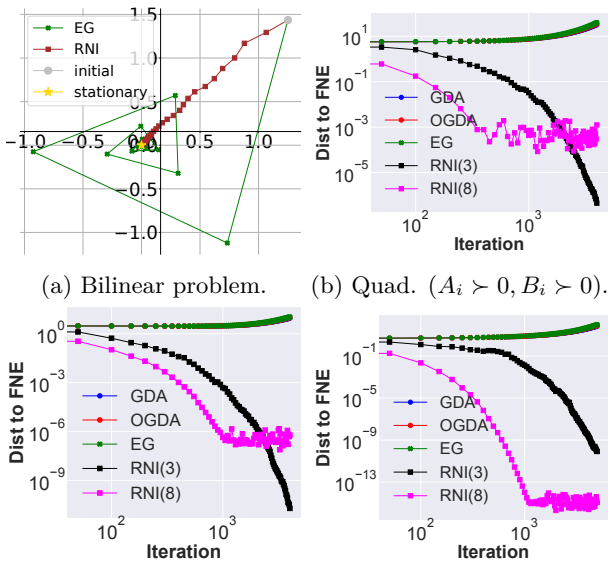
Indeed, notice that contrary to the behavior of the EG algorithm, RNI-SGD approaches the stationary point following a direct path. Finally, we also tested GDA and OGDA in the same experiment and noticed that the former cycles around the stationary point (without converging), while the latter behaves similar to EG (we omitted them in the presentation in order for the plot to be more legible).

In the quadratic case we have the objective

$$f_i(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^T A_i \mathbf{x} + \mathbf{x}^T Q_i \mathbf{y} + \frac{1}{2}\mathbf{y}^T B_i \mathbf{y},$$

where $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^n$. On this objective we consider the following three cases: i) $A_i \succ 0, B_i \succ 0$ (str. convex-str. convex), ii) $A_i \prec 0, B_i \prec 0$ (str. concave-str. concave), iii) $A_i \prec 0, B_i \succ 0$ (str. concave-str. convex). Notice that all the above problems are potentially difficult since they involve at least one difficult subproblem; especially case iii which involves solving two difficult problems (i.e., minimize a concave function and maximize a convex one).

In figures 1b,1c,1d, we plot the results for the cases i, ii, iii of the quadratic objective, respectively. In these experiments the RNI-SGD algorithm solves the subproblems (8),(9) using 50 steps of the projected gradient descent algorithm. Also, the value of L is selected as $L = 1.05 \cdot \max\{L_x, L_y\}$ in cases i,ii and as $L = 1.5 \cdot \max\{L_x, L_y\}$ in case iii. It should be noted that these problems are more challenging than the bilinear case and (strongly) convex-concave problems (for which it is established in practice that OGDA and EG converge to a NE (Mokhtari et al., 2019b)), and as a result the GDA, OGDA and EG methods diverge. On the other hand RNI-SGD approaches a stationary point of f, as predicted by theory.



(a) Bilinear problem. (b) Quad. ($A_i \succ 0, B_i \succ 0$).

(c) Quad. ($A_i \prec 0, B_i \prec 0$). (d) Quad. ($A_i \prec 0, B_i \succ 0$).

Figure 1: The results of the experiments for bilinear and quadratic objectives; the plots of GDA, OGDA, EG behave similarly and so they are indistinguishable. These results (with the exception of the trajectory plots) are averaged over 5 independent runs. In the quadratic objective the RNI-SGD algorithm is tested with batch size 3 and 8, while the batch size of the stochastic versions of the rest of the algorithms is set to 8. Also, the stepsize is selected in all cases after trials, in order to ensure that the algorithms approach the stationary point sufficiently fast, if that is possible.

5 CONCLUSION

In this paper we use the RNI objective as a merit function for finding the FNEs of two-player zero-sum games over compact constraint sets, and propose a first-order algorithm, the RNI-SGD, for solving the respective optimization problem. The key characteristics of our setting is the fact that we consider games with *non-convex and stochastic* objectives. Under this setting we manage to show convergence of RNI-SGD to a neighborhood around a stationary solution of the RNI objective. Moreover, we identify nonconvex min-max games whose corresponding RNI reformulations are convex functions. In the future we would be interested to see if this approach has other benefits to offer in addressing the problem of finding FNEs or more generally (local) Nash equilibria in non-convex games. Finally, another interesting direction for future research is the study of the performance of the proposed approach in machine learning problems (e.g. generative adversarial networks, adversarial learning).

Acknowledgments

M. Hong and I. Tsaknakis are supported by NSF Award CIF-1910385.

References

Abernethy, J., Lai, K. A., and Wibisono, A. (2019). Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*.

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. (2018). Local saddle point optimization: A curvature exploitation approach. *arXiv preprint arXiv:1805.05751*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223.

Barazandeh, B. and Razaviyayn, M. (2020). Solving non-convex non-differentiable min-max games

using proximal gradient method. *arXiv preprint arXiv:2003.08093*.

Bernhard, P. and Rapaport, A. (1995). On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear Anal.*, 24(8):1163–1181.

Cai, Q., Hong, M., Chen, Y., and Wang, Z. (2019). On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. (2020). Reducing noise in gan training with variance reduced extragradient. *arXiv preprint arXiv:1904.08598*.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2017). Training gans with optimism. *arXiv preprint arXiv:1711.00141*.

Daskalakis, C. and Panageas, I. (2018). The limit points of (optimistic) gradient descent in min-max optimization. In *32nd Annual Conference on Neural Information Processing Systems*.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2018). A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Grimmer, B., Lu, H., Worah, P., and Mirrokni, V. (2020). The landscape of nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*.

Gürkan, G. and Pang, J.-S. (2009). Approximations of nash equilibria. *Mathematical Programming*, 117(1-2):223–253.

Hogben, L. (2013). *Handbook of Linear Algebra*. Discrete Mathematics and Its Applications. CRC Press.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Jin, C., Netrapalli, P., and Jordan, M. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR.

Lin, T., Jin, C., and Jordan, M. (2020a). On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR.

Lin, T., Jin, C., and Jordan, M. I. (2020b). Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR.

Liu, M., Mroueh, Y., Ross, J., Zhang, W., Cui, X., Das, P., and Yang, T. (2019). Towards better understanding of adaptive gradient algorithms in generative adversarial nets.

Liu, M., Rafique, H., Lin, Q., and Yang, T. (2018). First-order convergence theory for weakly-convex-weakly-concave min-max problems. *arXiv preprint arXiv:1810.10207*.

Loizou, N., Berard, H., Jolicoeur-Martineau, A., Vincent, P., Lacoste-Julien, S., and Mitliagkas, I. (2020). Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. (2020). Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mescheder, L., Nowozin, S., and Geiger, A. (2017). The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2019a). Convergence rate of o(1/k) for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. *arXiv preprint arXiv:1906.01115*.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2019b). A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*.

Nikaidô, H., Isoda, K., et al. (1955). Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5(Suppl. 1):807–815.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14934–14942.

Ostrovskii, D. M., Lowy, A., and Razaviyayn, M. (2020). Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*.

Pang, J.-S. and Scutari, G. (2011). Nonconvex games with side constraints. *SIAM Journal on Optimization*, 21(4):1491–1522.

Qu, B. and Zhao, J. (2013). Methods for solving generalized nash equilibrium. *Journal of Applied Mathematics*, 2013.

Raghunathan, A. U., Cherian, A., and Jha, D. K. (2019). Game theoretic optimization via gradient-based nikaido-isoda function. *arXiv preprint arXiv:1905.05927*.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113.

Uryas' ev, S. and Rubinstein, R. Y. (1994). On relaxation algorithms in computation of noncooperative equilibria. *IEEE Transactions on Automatic Control*, 39(6):1263–1267.

Von Heusinger, A. and Kanzow, C. (2009a). Optimization reformulations of the generalized nash equilibrium problem using nikaido-isoda-type functions. *Computational Optimization and Applications*, 43(3):353–377.

Von Heusinger, A. and Kanzow, C. (2009b). Relaxation methods for generalized nash equilibrium problems with inexact line search. *Journal of Optimization Theory and Applications*, 143(1):159–183.

Yang, J., Kiyavash, N., and He, N. (2020). Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33.