Abstract
Computing Nash equilibrium (NE) of multiplayer games has witnessed renewed interest due to recent advances in generative adversarial networks. However, computing equilibrium efficiently is challenging. To this end, we introduce the Gradient-based Nikaido-Isoda (GNI) function which serves: (i) as a merit function, vanishing only at the first-order stationary points of each player’s optimization problem, and (ii) provides error bounds to a stationary Nash point. Gradient descent is shown to converge sublinearly to a first-order stationary point of the GNI function. For the particular case of bilinear min-max games and multi-player quadratic games the GNI function is convex. Hence, the application of gradient descent in this case yields linear convergence to an NE (when one exists). In our numerical experiments we observe that the GNI formulation always converges to the first-order stationary point of each player’s optimization problem.

International Conference on Machine Learning (ICML)
Game Theoretic Optimization via Gradient-based Nikaido-Isoda Function

Arvind U. Raghunathan 1  Anoop Cherian 1  Devesh K. Jha 1

Abstract
Computing Nash equilibrium (NE) of multiplayer games has witnessed renewed interest due to recent advances in generative adversarial networks. However, computing equilibrium efficiently is challenging. To this end, we introduce the Gradient-based Nikaido-Isoda (GNI) function which serves: (i) as a merit function, vanishing only at the first-order stationary points of each player’s optimization problem, and (ii) provides error bounds to a stationary Nash point. Gradient descent is shown to converge sublinearly to a first-order stationary point of the GNI function. For the particular case of bilinear min-max games and multi-player quadratic games the GNI function is convex. Hence, the application of gradient descent in this case yields linear convergence to an NE (when one exists). In our numerical experiments we observe that the GNI formulation always converges to the first-order stationary point of each player’s optimization problem.

1. Introduction
In this work, we consider the general $N$-player game:

\[
\begin{align*}
\text{Find } \mathbf{x}^* &= (x_1^*, \ldots, x_N^*) \text{ s.t.} \\
x_i^* &= \arg \min_{x \in \mathbb{R}^{n_i}} \sum_{i=1}^{N} f_i(x) 
\end{align*}
\]

where $x_i \in \mathbb{R}^{n_i}, n = \sum_{i=1}^{N} n_i, f_i : \mathbb{R}^n \rightarrow \mathbb{R}, x = (x_1, \ldots, x_N) \in \mathbb{R}^n$ denotes the collection of all $x_j$’s, while $x_i$ denotes the collection of all $x_j$’s except for index $i$, i.e. $x_i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N) \in \mathbb{R}^{n-n_i}$. Observe that the choice of $x_i$ are specified when performing the minimization in (1) for player $i$.

A point $\mathbf{x}^*$ satisfying (1) is called a Nash Equilibrium.

\[1\] All authors are with Mitsubishi Electric Research Labs (MERL), Cambridge, MA. Correspondence to: Arvind U. Raghunathan <raghunathan@merl.com>, Anoop Cherian <cherian@merl.com>, Devesh K. Jha <jha@merl.com>.

function converges to a stationary Nash point (§4). In addition, if each of the player’s objective is convex in the player’s variables \( (x_i) \) then the algorithm converges to the NE point as long as one exists (§4). A secant approximation is provided to simplify the computation of the gradient of the GNI function and the convergence of the modified algorithm is also analyzed (§5). Numerical experiments in §6 show that the proposed algorithm is effective in converging to stationary Nash points of the games.

We believe our proposed GNI formulation could be an effective approach for training GANs. However, we emphasize that the focus of this paper is to provide a rigorous analysis of the GNI formulation for games and explore its properties in a non-stochastic setting. The adaptation of our proposed formulations to a stochastic setting (which is the typical framework commonly used in GANs) will need additional results, which will be explored in a future paper.

2. Related Work

Nash Equilibrium (NE) computation, a key area in algorithmic game theory, has seen a number of developments since the pioneering work of John von Neumann (Basar & Olsder, 1999). It is well known that the Nash equilibrium problem can be reformulated as a variational inequality problem, VIP for short, see, for example, (Facchinei & Pang, 2003a). The VIP is a generalization of the first-order optimality condition in \( S^{NP} \) to the case where the decision variables of player i’s \( z_i \) are constrained to be in a convex set. Facchinei & Kanzow (2010) proposed penalty methods for the solution of generalized Nash equilibrium problems (Nash equilibrium problems with joint constraints). Iusem et al. (2017) provides a detailed analysis of the extragradient algorithm for stochastic pseudomonotone variational inequalities (corresponding to games with polyconvex costs).

Nash Equilibrium computation has found renewed interest due to the emergence of Generative Adversarial Networks (GANs). It has been observed that the alternating stochastic gradient descent (SGD) is oscillatory when training GANs (Goodfellow, 2016). Several papers proposed to modify the GAN formulation in order to stabilize the convergence of the iterates. These include non-saturating GAN formulation of (Goodfellow et al., 2014; Fedus et al., 2018), the DCGAN formulation (Radford et al., 2015), the gradient penalty formulation for WGANs (Gulrajani et al., 2017). The authors in (Yadav et al., 2017) proposed a momentum based step on the generator in the alternating SGD for convex-concave saddle point problems. Daskalakis et al. (Daskalakis et al., 2018) proposed the optimistic mirror descent (OMD) algorithm, and showed convergence for bilinear games and dexterity of the gradient descent iterates. In a subsequent work, Daskalakis et al. (Daskalakis & Panageas, 2018) analyzed the limit points of gradient descent and OMD, and showed that the limit points of OMD is a superset of alternating gradient descent. Mertikopoulos et al. (2019) generalized and extended the work of Daskalakis et al. (2018) for bilinear games. Li et al. (2017) dualize the GAN objective to reformulate it as a maximization problem and Mescheder et al. (2017) add the norm of the gradient in the objective. The norm of the gradient is shown to locally stabilize the gradient descent iterations in Nagarajan & Kolter (2017). Gidel et al. (2018) formulate the GAN equilibrium as a VIP and propose an extrapolation technique to prevent oscillations. The authors show convergence of stochastic algorithm under the assumption of monotonicity of VIP, which is stronger than the convex-concave assumption in min-max games. Finally, the convergence of stochastic gradient descent in non-convex games has also been studied in Bervoets et al. (2018); Mertikopoulos & Zhou (2019).

In contrast to existing approaches, the GNI approach does not assume monotonicity in the game formulations. The GNI approach is also closely related to the idea of minimizing residuals (Facchinei & Pang, 2003a,b).

3. Gradient-based Nikaido-Isoda Function

The Nikaido-Isoda (NI) function introduced in (Nikaido & Isoda, 1955) is defined as

\[
\psi(x) = \sum_{i=1}^{N} \left( f_i(x) - \inf_{\hat{x} \in \mathbb{R}^n : \hat{x}_{-i} = x_{-i}} f_i(\hat{x}) \right).
\]

From the definition of NI function \( \psi(x) \), it is easy to show that \( \psi(x) \geq 0 \) for all \( x \in \mathbb{R}^n \). Further, \( \psi(x) = 0 \) is the global minimum which is only achieved if the NE point \( x^* = (x_1^*, \ldots, x_N^*) \) occurs at points where \( x_i^* \) are global minimizers of the respective optimization problems in (1). A number of papers (Uryasev & Rubinstein, 1994; von Heusinger & Kanzow, 2009a,b) have proposed algorithms that minimize \( \psi(x) \) to compute NE points. However, the infimum needed to compute \( \psi_i(x) \) can be prohibitive for all but a handful of functions. For bilinear min-max games (i.e., \( f_1(x) = x_1^T x_2 = -f_2(x) \)), the infimum is unbounded below and the approach of minimizing NI fails. To rectify this recent papers have proposed regularized variants (von Heusinger & Kanzow, 2009b). However, the cost of globally minimizing the nonlinear function can still be prohibitive.

To rectify the shortcoming of the NI function, we introduce the Gradient-based Nikaido-Isoda (GNI) function

\[
V(x; \eta) = \sum_{i=1}^{N} f_i(x) - f_i(y(x; i, \eta)) \quad (2)
\]
where $y_j(x; i, \eta) = \begin{cases} x_i - \eta \nabla_i f_i(x), & \text{if } j = i \\ x_j, & \text{otherwise}. \end{cases}$

where $\nabla_i f(x)$ denotes the derivative of function $f$ w.r.t. $x_i$.

The GNI function is obtained by replacing the infimum in the NI function for player $i$ with a point $y(x; i, \eta)$ in the steepest descent direction. This provides a local measure of decrease that can be obtained in the objective for player $i$. The point $y(x; i, \eta)$ is similar in spirit to the Cauchy point that is used in trust-region methods (Nocedal & Wright, 2006). We will show that any point satisfying $V_i(x; \eta) = 0$ also satisfies $\nabla_i f_i(x) = 0$. To show this, we first provide bounds on $V_i(x; \eta)$ in terms of the distance from first-order optimality conditions for each of the players.

We make the following standing assumption.

**Assumption 1.** The functions $f_i$ are at least twice continuously differentiable and gradients of $f_i$ (i.e., $\nabla_i f_i$) are Lipschitz continuous with constant $L_f$.

**Lemma 1.** \[ \frac{1}{2} \| \nabla_i f_i(x) \|^2 \leq V_i(x; \eta) \leq \frac{\eta^2}{2} \| \nabla_i f_i(x) \|^2 \]

for all $x \in \mathbb{R}^n$ and $0 < \eta \leq \frac{1}{L_f}$.

**Proof.** Using the Taylor’s series expansion of $f_i$ around $x$ and substituting for $y(x; i, \eta)$, we obtain

\[
\begin{align*}
 f_i(y(x; i, \eta)) &= f_i(x) - \eta \nabla_i f_i(x) \|^2 \\
 &+ \frac{\eta^2}{2} \int_0^1 \nabla_i f_i(x(t)) \nabla_i^2 f_i(x(t)) dt
\end{align*}
\]

where $x(t) = x_{i-1} - \eta \nabla_i f_i(x)$ and $x_{i-1}$ is $x_{i-1}$ for $j \neq i$. From the Lipschitz continuity of the gradient of $f_i$, we have that $-L_f I_i \leq \nabla_i^2 f_i(x(t)) \leq L_f I_i$, where $I_i$ is the $(n_i \times n_i)$ identity matrix. Substituting in the above and using $\eta \leq \frac{1}{L_f}$ yields the claim.

We now state our main result relating the zeros of $V_i(x; \eta)$ and the first-order critical points of the players’ optimization problems.

**Theorem 1.** The global minimizers of $V(x; \eta)$ are all stationary Nash points, i.e., \{ $x^* \mid V(x^*; \eta) = 0$ \} $\leq \mathcal{S}^{SNF}$ for all $0 \leq \eta \leq \frac{1}{L_f}$. If the individual functions $f_i$ are convex, then the global minimizers of $V(x; \eta)$ are precisely the set $\mathcal{S}^{NE}$.

**Proof.** The nonnegativity of $V(x; \eta)$ follows from Lemma 1. Further, $V(x; \eta) = 0$ if and only if $\nabla_i f_i(x) = 0$. This proves the claim. The second claim follows by noting that $\mathcal{S}^{NE} = \mathcal{S}^{SNF}$, if the functions $f_i$ are convex.

Theorem 1 shows that the function $V(x; \eta)$ can be employed as a merit function for obtaining a stationary Nash point.

When $f_i(x)$ are non-convex, the convergence to first-order point is possibly the best that one can hope for.

We provide the expression for the gradient and Hessian of $V_i(x; \eta)$ next. These expressions follow from the chain rule of differentiation. The gradient of $V_i(x; \eta)$ is

\[
\begin{align*}
\nabla V_i(x; \eta) &= \nabla f_i(x) - (I - \eta \nabla^2 f_i(x) E_i) \nabla f_i(y(x; i, \eta)) \\
&= \nabla f_i(x) - (I - \eta \nabla^2 f_i(x) E_i) \nabla f_i(y(x; i, \eta))
\end{align*}
\]

where $E_i = F_i F_i^T$ with $F_i \in \mathbb{R}^{n_i \times n_i}$ defined as $F_i = \begin{bmatrix} 0_{n_i \times n_i} & E_i \\ I_i & 0_{n_i \times n_i} \end{bmatrix}$, $I_i \in \mathbb{R}^{n_i \times n_i}$ and $I_i \in \mathbb{R}^{n_i \times n_i}$ are identity matrices. The Hessian of $V_i(x; \eta)$ is given by

\[
\begin{align*}
\nabla^2 V_i(x; \eta) &= \nabla^2 f_i(x) + \eta \nabla^3 f_i(x) [E_i \nabla f_i(y(x; i, \eta))] \\
&- (I - \eta \nabla^2 f_i(x) E_i) \nabla^2 f_i(y(x; i, \eta))(I - \eta E_i \nabla^2 f_i(x))
\end{align*}
\]

where $\nabla^3 f_i(x) d = \lim_{t \rightarrow 0} \frac{\nabla^2 f_i(x+od)}{d}$ is the action of the third derivative along the direction $d$. These expressions will come useful in our analysis to follow.

### 3.1. GNI is Locally Stable

GAN formulations typically result in objective functions $f_i(x)$ that are not convex. Nagarajan and Kolter (Nagarajan & Kolter, 2017) showed that the gradient descent for min-max games is not stable for Wasserstein GANs. This is due to the concave-concave nature of Wasserstein GAN around stationary Nash points (Nagarajan & Kolter, 2017). Daskalakis et al. (Daskalakis et al., 2018) showed that the gradient descent diverges for simple bilinear min-max games, while the optimistic gradient descent algorithm of Rakhlin and Sridharan (Rakhlin & Sridharan, 2013) was shown to be convergent. Daskalakis and Panageas (Daskalakis & Panageas, 2018) further analyzed the limit points of gradient descent and optimistic gradient descent using dynamical systems theory.

In this section, we show that at every stationary Nash point, the Hessian of $V(x; \eta)$ is positive semidefinite. This ensures that the points in $\mathcal{S}^{SNF}$ are all stable limit points for the gradient descent algorithm on $V(x; \eta)$.

**Lemma 2.** For $0 \leq \eta \leq \frac{1}{L_f}$, $\nabla V^2(x^*; \eta) = \sum_{i=1}^n \nabla^2 V_i(x^*; \eta)$ is positive semidefinite for all $x^* \in \mathcal{S}^{SNF}$.

**Proof.** Let $x^* \in \mathcal{S}^{SNF}$. Since $\nabla_i f_i(x^*) = 0$, we have that $y(x^*; i, \eta) = x^*$ and $\nabla^3 f_i(x^*) [E_i \nabla f_i(y(x; i, \eta))] = 0$. Substituting in the expression for $\nabla^2 V_i(x; \eta)$ in (4)
simplifying, we obtain

\[ \nabla^2 V(\mathbf{x}^*; \eta) = 2\eta \nabla^2 f_i(\mathbf{x}^*) E_i \nabla^2 f_i(\mathbf{x}^*) - \nabla^2 f_i(\mathbf{x}^*) E_i \nabla^2 f_i(\mathbf{x}^*) \]

\[ = \eta \nabla^2 f_i(\mathbf{x}^*) (2E_i - \eta E_i \nabla^2 f_i(\mathbf{x}^*) E_i) \nabla^2 f_i(\mathbf{x}^*) \]

(5)

From the Lipschitz continuity of \( f_i(\mathbf{x}) \) we have that \( \nabla^2 f_i(\mathbf{x}^*) \leq L_f E_i \). Substituting into (5), we obtain

\[ \nabla^2 V(\mathbf{x}^*; \eta) \geq \eta \nabla^2 f_i(\mathbf{x}^*) (2E_i - (\eta L_f) E_i) \nabla^2 f_i(\mathbf{x}^*) \]

\[ \geq \eta \nabla^2 f_i(\mathbf{x}^*) E_i \nabla^2 f_i(\mathbf{x}^*) \]

where the final simplification follows from \( \eta L_f \leq 1 \) and \( E_i^2 = E_i \). The claim follows from the positive semidefiniteness of \( \nabla^2 f_i(\mathbf{x}^*) E_i \nabla^2 f_i(\mathbf{x}^*) \). Since \( \nabla^2 V(\mathbf{x}^*; \eta) \) is the sum of positive semidefinite matrices the claim holds.

3.2. Convexity Properties of GNI: An Example

In this section, we present an example NE reformulation of a (non-) convex game using the GNI setup. Suppose the player’s objective is quadratic, i.e., \( f_i(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T Q_i \mathbf{x} + \mathbf{r}_i^T \mathbf{x} \). Then, the GNI function is

\[ V_i(\mathbf{x}) = f_i(\mathbf{x}) - f_i(\mathbf{x} - \eta E_i (Q_i \mathbf{x} + \mathbf{r}_i)) \]

\[ = \frac{1}{2} \mathbf{x}^T \left( Q_i - \tilde{Q}_i^T \tilde{Q}_i \right) \mathbf{x} + \eta \mathbf{r}_i^T E_i Q_i (I + \tilde{Q}_i) \mathbf{x} \]

\[ + \frac{1}{2} \eta \mathbf{r}_i^T (2E_i - \eta E_i Q_i E_i) \mathbf{r}_i \]

(6)

where \( \tilde{Q}_i = (I - \eta E_i Q_i) \). Suppose \( \|Q_i\| \leq L_f \) and let \( \eta \leq \frac{1}{2L_f} \), then

\[ (Q_i - \tilde{Q}_i^T \tilde{Q}_i) = \eta (Q_i - E_i Q_i) (2I - \eta E_i Q_i) E_i Q_i \geq 0 \]  

(7)

where the positive semidefiniteness holds since for all \( u \neq 0 \) \( u^T (Q_i - E_i Q_i) (2I - \eta E_i Q_i) E_i Q_i u = (Q_i - E_i Q_i)^2 (2I - \eta E_i Q_i) E_i Q_i u \geq 0 \). Hence, when \( f_i(\mathbf{x}) \) is quadratic, the GNI function is a convex, quadratic function. Note that the convexity of GNI function holds regardless of the convexity of the original function \( f_i(\mathbf{x}) \). However, for general nonlinear functions \( f_i(\mathbf{x}) \), the GNI function \( V_i(\mathbf{x}) \) does not preserve convexity.

4. Descent Algorithm for GNI

Consider the gradient descent iteration minimizing \( V(\mathbf{x}; \eta) \)

\[ \mathbf{x}^{k+1} = \mathbf{x}^k - \rho \nabla V(\mathbf{x}^k; \eta) \quad \text{for} \quad k = 0, 1, 2, \ldots \]  

(8)

where \( \rho > 0 \) is a stepsize. The restrictions on \( \rho \), if any, are provided in subsequent discussions.

Theorem 2 proves sublinear convergence of \( \{\mathbf{x}^k\} \) to a stationary point of GNI function based on standard analysis. Linear convergence to a stationary point point is shown under the assumption of the Polyak-Lojasiewicz inequality (Lojasiewicz, 1963; Polyak, 1963; Karimi et al., 2018). Luo & Tseng (1993) employed similar error bound conditions in the context of descent algorithms of variational inequalities.

**Theorem 2.** Suppose \( \nabla V(\mathbf{x}) \) is \( L_V \)-Lipschitz continuous. Let \( \rho = \frac{\alpha}{L_V} \) for \( 0 < \alpha \leq 1 \). Then, the \( \{\mathbf{x}^k\} \) generated by (8) converges sublinearly to \( \mathbf{x}^* \) a first-order stationary point of \( V(\mathbf{x}; \eta) \), i.e., \( \nabla V(\mathbf{x}^*; \eta) = 0 \). If \( V(\mathbf{x}; \eta) \leq \frac{\alpha}{2L_V} \|\nabla V(\mathbf{x}; \eta)\|^2 \) then the sequence \( \{V(\mathbf{x}^k)\} \) converges linearly to 0, i.e., \( \{\mathbf{x}^k\} \) converges to \( \mathbf{x}^* \in S^{\text{SNP}} \).

**Proof.** From Lipschitz continuity of \( \nabla V(\mathbf{x}; \eta) \)

\[ V(\mathbf{x}^{k+1}; \eta) \leq V(\mathbf{x}^k; \eta) + \nabla V(\mathbf{x}^k; \eta)^T (\mathbf{x}^{k+1} - \mathbf{x}^k) \]

\[ + \frac{L_V}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \]

\[ \leq V(\mathbf{x}^k; \eta) - \rho (1 - \frac{\rho L_V}{2}) \|\nabla V(\mathbf{x}; \eta)\|^2 \]

\[ \leq V(\mathbf{x}^k; \eta) - \frac{\alpha}{2L_V} \|\nabla V(\mathbf{x}; \eta)\|^2 \]

(9)

where \( \alpha = \alpha(2 - \alpha) \). Telescoping the sum for \( k = 0, \ldots, K \), we obtain

\[ V(\mathbf{x}^{K+1}; \eta) \leq V(\mathbf{x}^0) - \frac{\alpha}{2L_V} \sum_{k=0}^{K} \|\nabla V(\mathbf{x}^k; \eta)\|^2 \]

(10)

Since \( V(\mathbf{x}; \eta) \) is bounded below by 0, we have that

\[ \frac{\alpha}{2L_V} \sum_{k=0}^{K} \|\nabla V(\mathbf{x}^k; \eta)\|^2 \leq V(\mathbf{x}^0) - V(\mathbf{x}^{K+1}) \leq V(\mathbf{x}^0) \]

\[ \implies \frac{\alpha}{2L_V} \min_{k \in \{0, \ldots, K\}} \|\nabla V(\mathbf{x}^k; \eta)\|^2 \leq \frac{V(\mathbf{x}^0)}{K+1} \]

This proves the claim on sublinear convergence to a first-order stationary point of \( V(\mathbf{x}; \eta) \). Suppose \( V(\mathbf{x}; \eta) \leq \frac{1}{2L_V} \|\nabla V(\mathbf{x}; \eta)\|^2 \) holds. Substituting in (9) obtain

\[ V(\mathbf{x}^{k+1}; \eta) \leq \left(1 - \frac{\alpha}{L_V}\frac{\mu}{\eta}\right) V(\mathbf{x}^k; \eta) \]

(11)

which proves the claim on linear convergence of \( \{V(\mathbf{x}^k; \eta)\} \) to 0. By Theorem 1, \( \{\mathbf{x}^k\} \) converges to \( \mathbf{x}^* \in S^{\text{SNP}} \).

4.1. Quadratic Objectives

In the following, we explore a popular setting of quadratic objective function and explore the implication of Theorem 2. Note that the bilinear case is a special case of the quadratic objective. Consider the \( f_i(\mathbf{x}) \)’s to be quadratic. For this setting \S 3.2 showed that GNI function \( V_i(\mathbf{x}) \) is a convex
quadratic function. This proves that \( V_i(x; \eta) \) has \((3L_f)\)-Lipschitz continuous gradient. It is well known that for a composition of a linear function with a strongly convex function, we have that Polyak-Łojasiewicz inequality holds (Luo & Tseng, 1993), i.e., there exists \( \mu > 0 \) such that \( V(x; \eta) \leq \frac{1}{2 \mu} \| \nabla V(x; \eta) \|^2 \) holds. Hence, we can state the following stronger result for quadratic objective functions.

**Corollary 1.** Suppose \( f_i(x) \) are quadratic and player convex, i.e. \( f_i(x) \) is convex in \( x_i \). Let \( \rho = \frac{1}{2L_fN} \). Then, the sequence \( \{x^k\} \) converges linearly to 0, i.e. \( \{x^k\} \) converges to \( x^* \in S^{NE} \).

### 5. Modified Descent Algorithm for GNI

The evaluation of the gradient \( \nabla V(x; \eta) \) requires the computation of the Hessian of the functions \( f_i(x) \) (see (3)) which can be prohibitive to compute. A close examination of the expression of \( \nabla^2 V(x; \eta) \) in (3) reveals that we only require the action of the Hessian in a particular direction, i.e. \( \nabla f_i(x - \eta E_i \nabla f_i(x)) \). This immediately suggests the use of an approximation for this term inspired by secant methods (Nocedal & Wright, 2006)

\[
\nabla^2 f_i(x)(\eta E_i, \nabla f_i(x - \eta E \nabla f_i(x))) \approx \nabla f_i(x + \eta E_i \nabla f_i(x - \eta E \nabla f_i(x))) - \nabla f_i(x) \quad (12)
\]

Substituting (12) for the term involving the Hessian in \( \nabla V_i(x; \eta) \) and simplifying the direction \( \nabla V_i(x; \eta) \)

\[
\nabla V_i(x; \eta) = \nabla f_i(x + \eta E_i \nabla f_i(x - \eta E \nabla f_i(x))) - \nabla f_i(x - \eta E \nabla f_i(x)) \quad (13)
\]

Substituting (12) in the gradient descent iteration (9), we obtain the modified iteration

\[
x^{k+1} = x^k - \rho \nabla V_i(x; \eta) \quad \text{for} \quad k = 0, 1, 2, \ldots \quad (14)
\]

where \( \nabla V_i(x; \eta) = \sum_{i=1}^{N} \nabla V_i(x; \eta) \). We assume that the following bound on the error in the approximation

\[
\| \nabla V(x; \eta) - \nabla V(x; \eta) \| \leq \tau \| \nabla V(x; \eta) \|, \quad (15)
\]

for some \( \tau \in (0, 1) \). Such a bound on the error in the gradients has also been used in Luo & Tseng (1993).

**Theorem 3.** Suppose \( \nabla V(x) \) is \( L_V \)-Lipschitz continuous. Let \( \rho = \alpha \frac{1 - \tau}{2L_V(1 + \tau)} \) for \( 0 < \alpha \leq 1 \) and (15) holds. Then, the \( \{x^k\} \) generated by (14) converges sublinearly to \( x^* \) a first-order stationary point of \( V(x; \eta) \), i.e., \( \nabla V(x^*; \eta) = 0 \). If \( V(x; \eta) \leq \frac{1}{2 \mu} \| \nabla V(x; \eta) \|^2 \), then the sequence \( \{x^k\} \) converges linearly to 0, i.e., \( \{x^k\} \) converges to \( x^* \in S^{NE} \).

**Proof.** Let \( \nabla V(x^k; \eta) = \nabla V(x^k; \eta) + e^k \). From (15), \( \| e^k \| \leq \tau \| \nabla V(x^k; \eta) \| \). Applying the triangle inequality to \( \| \nabla V(x^k; \eta) \| \) and use (15) obtain

\[
\| \nabla V(x^k) \| \leq \| \nabla V(x^k; \eta) \| + \| e^k \| \leq (1 + \tau) \| \nabla V(x^k; \eta) \|. \quad (16)
\]

The term \( - (\nabla V(x^k; \eta))^T (\nabla V(x^k; \eta)) \) can be upper bounded as

\[
- (\nabla V(x^k; \eta))^T (\nabla V(x^k; \eta)) \approx - \| \nabla V(x^k; \eta) \|^2 - (\nabla V(x^k; \eta))^T e^k \leq - \| \nabla V(x^k; \eta) \|^2 + \| \nabla V(x^k; \eta) \| \| e^k \| \leq (1 - \tau) \| \nabla V(x^k; \eta) \|^2 \quad (17)
\]

where the final inequality follows from (15). From Lipschitz continuity of \( \nabla V(x; \eta) \)

\[
V(x^{k+1}; \eta) \leq V(x^k; \eta) + \nabla V(x^k; \eta)^T (x^{k+1} - x^k) + \frac{L_V}{2} \| x^{k+1} - x^k \|^2 \leq V(x^k; \eta) - \rho (1 - \tau) \| \nabla V(x^k; \eta) \|^2 + \frac{L_V \rho^2}{2} \| \nabla V(x; \eta) \|^2 \leq V(x^k; \eta) - \rho (1 - \tau) \frac{L_V \rho (1 + \tau)^2}{2} \| \nabla V(x; \eta) \|^2 \leq V(x^k; \eta) - \frac{\alpha}{2} \frac{(1 - \tau)^2}{L_V (1 + \tau)} \| \nabla V(x; \eta) \|^2 \quad (18)
\]

where \( \alpha = \alpha(2 - \alpha) \), the third inequality is obtained by substituting (16) and (17), and the final inequality follows from the definition of \( \rho \) in the statement of the theorem. By similar arguments to those in Theorem 2 obtain

\[
\alpha \left( 1 - \tau \right)^2 \min_{k \in \{0, \ldots, K\}} \| \nabla V(x^k; \eta) \|^2 \leq \frac{L_V}{K + 1} V(x^0; \eta). \quad (19)
\]

This proves the claim on sublinear convergence to a first-order stationary point of \( V(x; \eta) \). Suppose \( V(x; \eta) \leq \frac{1}{2 \mu} \| \nabla V(x; \eta) \|^2 \) holds. Substituting in (18) obtain

\[
V(x^{k+1}; \eta) \leq \left( 1 - \alpha \frac{L_V(1 - \tau)^2}{L_V(1 + \tau)^2} \right) V(x^k; \eta) \quad \text{for} \quad k = 0, \ldots, K - 1.
\]

which proves the claim on linear convergence of \( \{V(x^k; \eta)\} \) to 0. By Theorem 1, \( \{x^k\} \) converges to \( x^* \in S^{NE} \).

The approximation in (12) is in fact exact when the function \( f_i(x) \) is quadratic. Consequently, the claims on the convergence of the iterates continue to hold when the iterates are generated by (14).
6. Experiments

In this section, we present several empirical results on simulated data demonstrating the effectiveness of the proposed GNI formulation. To demonstrate the correctness of our theoretical results, we show numerical results on several simple game settings with known equilibrium. Specifically, we consider the following payoff functions: i) bilinear two-player games, ii) quadratic games with convex and non-convex payoffs, iii) linear GAN using a Dirac delta generator, and iv) a more general linear GAN with linear generator and discriminator. We compare our descent algorithm against several popular choices such as (i) gradient descent, (ii) gradient descent with Adam-style updates (Kingma & Ba, 2014), (iii) optimistic mirror descent (Rakhlin & Sridharan, 2013; Daskalakis et al., 2018), (iv) the extrapolation scheme (Gidel et al., 2018), and (v) the extra-gradient method (Korpelevich, 1976). For all these methods, we either follow the standard hyperparameter settings (e.g., in Adam), or we find the hyperparameters that lead to the best convergence. For each of these games, we observe convergence of the proposed algorithm to stationary Nash points and contrast the quality of solutions against what can be theoretically guaranteed. As discussed in Section 3.2, the quadratic and bilinear cases lead to convex GNI function and thus, the game always converges to a NE. Refer to supplementary materials for extra experiments. Below, we detail each of the game settings.

6.1. Bi-Linear Two-player Game:

We consider the following two-player game:

\[
f_1(x) = x_1^T Q x_2 + q_1^T x_1 + q_2^T x_2 = -f_2(x),
\]

where \( f_1 \) and \( f_2 \) are the player’s payoff functions—a setting explored in (Gidel et al., 2018). The GNI for this game leads to a convex objective. For GNI, we use a step-size \( \eta = 1/L \), where \( L = \|Q\| \), and \( \rho = 0.01 \), while for other methods we use a stepsize of \( \eta = 0.001^1 \). The methods are initialized randomly—the initialization is seen to have little impact on the convergence of GNI, however changed drastically for that of others.

In Figure 1(a), we plot the gradient convergence (using 10-d data). In this plot (and all subsequent plots of gradient convergence), the norm of the gradient \( \|\nabla f(x^k)\| = \|\langle \nabla_1 f_1(x^k), \ldots, \nabla_N f_N(x^k) \rangle\| \). We see that GNI converges linearly. However, other methods, such as gradient descent and mirror descent iterates diverge, while the extragradient and Adam are seen to converge slowly. To understand the descent better, in Figure 1(b), we use \( x_1, x_2 \in \mathbb{R}^3 \), and plot them for every 100-th iteration starting from the same initial point (shown by the red-diamond). Interestingly,

\(^1\)Other values of \( \eta \) did not seem to result in stable descent.

Figure 1. (a) shows GNI against other methods for bilinear min-max game. (b) shows convergence trajectories for 1-dimensional players. For (b), the initial point is shown in red diamond.

we find that the extragradient and mirror-descent methods show a circular trajectory, while Adam (with \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \)) takes a spiral convergence path. GNI takes a more straight trajectory steadily decreasing to optima (shown by the blue straight line).

6.2. Two-Player Quadratic Games:

We consider two-player games ( multiplayer extensions are trivial) with the payoff functions:

\[
f_i(x) = \frac{1}{2} x^T Q_i x + r_i^T x, \text{ for } i = 1, 2
\]

where \( Q_i \in \mathbb{R}^{n \times n} \) is symmetric. We consider cases when each \( Q_i \) is indefinite (i.e., non-convex QU) and positive semi-definite. As with the bilinear case, all the QU payoffs result in convex GNI reformulations. We used 20-d data, the same stepizes \( \eta = \max_i(\|Q_i\|) \) and \( \rho = 0.01 \) for GNI, while using \( \eta = 10^{-4} \) for other methods. The players are initialized from \( N(0, I) \).

In Figure 2, we compare the descent on these quadratic games. We find that the competitive methods are difficult to optimize for the non-convex QU and almost all of them diverge, except Adam which converges slowly. GNI is found to converge to the stationary Nash point (as it is convex— in 3.2). For the convex case, all methods are found to
converge. To gain insight, we plot the convergence trajectory for a 1-d convex quadratic game (i.e., \(x_1, x_2 \in \mathbb{R}^1\)) in Figure 3. The initializations are random for both players and the parameters are equal. We see that all schemes follow similar trajectories, except for Adam and GNI – all converging to the same point.

![Figure 3. Convergence of GNI against other methods on a convex 1-d quadratic game. Left: the convergence achieved by different algorithms. Right: the trajectories of the two players to the NE.](image)

**6.3. Dirac Delta GAN**

This is a one-dimensional GAN explored in (Gidel et al., 2018). In this case, the real data is assumed to follow a Dirac delta distribution (with a spike at say point \(-2\)). The payoff functions for the two players are:

\[
\begin{align*}
    f_1 &= \log(1 + \exp(\theta x_1)) + \log(1 + \exp(x_1 x_2)) \\
    f_2 &= -\log(1 + \exp(x_1 x_2)),
\end{align*}
\]

where \(\theta \in \mathbb{R}^1\) is the location of the delta spike. Unlike other game settings described above, we do not have an analytical formula to find the Lipschitz constant for the payoffs. To this end, we did an empirical estimate (more details to follow). We used \(L = 2\), \(\eta = \rho = 1/L\) and initialized all players uniformly from \([0, 4]\).

Figure 4 shows the comparison of the convergence of the dirac delta GAN game to a stationary Nash point. The GNI achieves faster convergence than all other methods, albeit having a non-convex reformulation in contrast to the bilinear and QP cases discussed above. The game has multiple local solutions and the schemes may converge to varied points depending on their initialization (see supplementary material for details).

![Figure 4. Convergence of GNI against other methods on the Dirac-Delta GAN.](image)

**6.4. Linear GAN**

We now introduce a more general GAN setup — a variant of the non-saturating GAN described in (Goodfellow, 2016), however using a linear generator and discriminator. We designed this experiment to serve two key goals: (i) to exposize the influence of the GNI hyperparameters in a more general GAN setting, and (ii) show the performance of GNI on a setting for which it is harder to estimate a Lipschitz constant

\[
f_1 = -E_{\theta \sim P_\theta} \log (x_1^T \theta) - E_{x_1 \sim P_x} \log (1 - x_1^T \text{diag}(x_2)) z,
\]

\[
f_2 = -E_{x_2 \sim P_x} \log (x_2^T \text{diag}(x_1)) z,
\]

where \(P_\theta\) and \(P_x\) are the real and the noise data distributions, the latter being the standard normal distribution \(N(0, I)\). The operator \(\text{diag}\) returns a diagonal matrix with its argument as its diagonal. We consider two cases for \(P_x\): (i) \(P_x = N(\mu, I)\) for a mean \(\mu\) and (ii) \(P_x = N(\mu, \Sigma)\) for a covariance matrix \(\Sigma \in \mathbb{R}^{d \times d}\). In our experiments to follow, we use \(\mu = 2e, e\) being a \(d\)-dimensional vector \((d = 10)\) of all ones. We initialized \(x_1 = x_2 = e/d\) for all the methods.

**Evaluation Metrics:** To evaluate the performance on various hyper-parameters of GNI, we define two metrics: (i) discriminator-accuracy, and (ii) the distance-to-mean. The discriminator-accuracy measures how well the learned discriminators classify the two distributions, defined as:

\[
d_{\text{acc}} = \frac{1}{2M} \sum_{i=1}^{M} \mathbb{I}(x_1^T \theta_i \geq \zeta) + \mathbb{I}(x_2^T \text{diag}(x_2)z_i \leq (1 - \zeta)),
\]

where \(\mathbb{I}\) is the indicator function, \(M\) is the number of data points sampled from the respective distributions, and \(\zeta \in [0, 1]\) is a threshold for the indicator function. We use \(\zeta = 0.7\). While \(d_{\text{acc}}\) measures the quality of the discriminator learned, it does not tell us anything on the convergence of the generator. To this end, we present another measure to evaluate the generator; specifically, the distance-to-mean, that computes the distance of the generated distribution from the first moment of the true distribution, defined as:

\[
d_{\text{mean}} = \|E_{x_2 \sim P_x} \text{diag}(x_2)z - E_{\theta \sim P_\theta} \theta\|
\]

**Hyper-parameter Study:** The goal of this experiment is to analyze the descent trajectory of GNI-based gradient descent when the hyper-parameters are changed. To this end,
we vary $\eta$ and $\rho$ separately in the range $10^{-5}$ to $10$ in multiples of $10$, while keeping the other parameter fixed (we use $\eta = 0.1$ and $\rho = 1$ as the base settings). In Figure 5, we plot the discriminator-accuracy and distance-to-mean against GNI iterations for the generator and discriminator separately. From Figures 5(a) and (b), it appears that higher value of $\eta$ biases the descents on the generator and discriminator separately. For example, $\eta \geq 0.01$ leads to a sharp descent to the optimal solution of the discriminator, however, $\eta \geq 1$ leads to a generator breakdown (Figure 5(a)). Similarly, a small value of $\rho$, such as $\rho < 10^{-5}$ shows high distance-to-mean, i.e., generator is weak, while $\rho = 1$ leads to good descents for both the generator and the discriminator. We found that a higher $\rho$ leads to unstable descent, skewing the plots and thus not shown. In short, we found that making the discriminator quickly converge to its optimum could lead to a better convergence trajectory for the generator for this linear GAN setup using the GNI scheme.

We presented a novel formulation for Nash equilibrium computation in multi-player games by introducing the Gradient-based Nikaio-Isoda (GNI) function. The GNI formulation for games allows individual players to locally improve their objectives using steepest descent while preserving local stability and convergence guarantees. We showed that the GNI function is a valid merit function for multi-player games and presented an approximate descent algorithm. We compared our method against several popular descent schemes on multiple game settings and empirically demonstrated that our method outperforms all other techniques. Future research will explore the GNI method in stochastic settings, that may enable their applicability to GAN optimization.
References


Game Theoretic Optimization via Gradient-based Nikaido-Isoda Function
Supplementary Materials

1. Residual Minimization

Lemma 1 (in the main paper) also suggests another possible function for minimization, namely \( \Phi(x) = \frac{1}{2} \sum_{i=1}^{2} \| \nabla_i f_i(x) \|^2 \). We can state a result that is analogous to Theorem 1.

**Theorem 4.** The global minimizers of \( \Phi(x) \) are all first-order NE points, i.e., \( \{ x^* | \Phi(x^*) = 0 \} = S^{SNE} \). If the individual functions \( f_i \) are convex then the global minimizers of \( \Phi(x) \) are precisely the set \( S^{SNE} \).

Denote by \( F(x) = \begin{bmatrix} \nabla_1 f_1(x) \\ \nabla_2 f_2(x) \end{bmatrix} \) the vector function of the first-order stationary conditions for each of the players. So \( \Phi(x) = \frac{1}{2} \| F(x) \|^2 \). The gradient of \( \Phi(x) \) is given by

\[
\nabla \Phi(x) = \nabla F(x) F(x)^T = \begin{bmatrix} \nabla_1^2 f_1(x) & \nabla_2^2 f_1(x) \\ \nabla_2^2 f_2(x) & \nabla_1^2 f_2(x) \end{bmatrix} \begin{bmatrix} \nabla_1 f_1(x) \\ \nabla_2 f_2(x) \end{bmatrix}. \tag{1}
\]

The Hessian of the function \( \Phi(x) \) is

\[
\nabla^2 \Phi(x) = \left( \sum_{j=1}^{n} F_j(x) \nabla^2 F_j(x) + \nabla F(x) \nabla F(x)^T \right). \tag{2}
\]

Consider the gradient descent iteration for minimizing \( \Phi(x) \) with stepsize \( \rho > 0 \)

\[
x^{k+1} = x^k - \rho \nabla \Phi(x^k). \tag{3}
\]

We can state the following convergence result for the gradient descent iterations.

**Theorem 5.** Suppose \( \nabla \Phi(x) \) is \( L_\Phi \)-Lipschitz continuous. Let \( \rho = \frac{\alpha}{L_\Phi} \) for \( 0 < \alpha \leq 1 \). Then, the \( \{ x^k \} \) generated by (3) converges sublinearly to \( x^* \) a first-order stationary point of \( \Phi(x) \), \( \nabla \Phi(x^*) = 0 \). If \( \Phi(x) \leq \frac{1}{\alpha} \| \nabla \Phi(x) \|^2 \) then the sequence \( \{ x^k \} \) converges linearly to a \( x^* \in S^{SNE} \).

**Proof.** From Lipschitz continuity of \( \nabla \Phi(x) \)

\[
\Phi(x^{k+1}) \leq \Phi(x^k) + \nabla \Phi(x^k)^T (x^{k+1} - x^k) + \frac{L_\Phi}{2} \| x^{k+1} - x^k \|^2 \\
\leq \Phi(x^k) - \rho (1 - \frac{\rho L_\Phi}{2}) \| \nabla \Phi(x) \|^2 \tag{4}
\]

where \( \alpha = \alpha(2 - \alpha) \). Telescoping the sum and \( k = 0, \ldots, K \) obtain

\[
\Phi(x^{K+1}) \leq \Phi(x^0) - \frac{\alpha}{2 L_\Phi} \sum_{k=0}^{K} \| \Phi(x^k) \|^2. \tag{5}
\]

Since \( \Phi(x) \) is bounded below by 0 we have that

\[
\frac{\alpha}{2 L_\Phi} \sum_{k=0}^{K} \| \nabla \Phi(x^k) \|^2 \leq \Phi(x^0) - \Phi(x^{K+1}) \leq \Phi(x^0)
\]

This proves the claim on sublinear convergence to a first-order stationary point of \( \Phi(x) \). Suppose \( \Phi(x) \leq \frac{1}{\alpha} \| \nabla \Phi(x) \|^2 \) holds. Substituting in (4) obtain

\[
\Phi(x^{k+1}) \leq \left( 1 - \frac{\alpha \mu}{L_\Phi} \right) \Phi(x) \tag{6}
\]

which proves the claim on linear convergence. \( \square \)

In the following we provide specific conditions under which the bound \( \Phi(x) \leq \frac{1}{\alpha^2} \| \nabla \Phi(x) \|^2 \) holds.

- Suppose the function \( f_i \) are quadratic then the discussion following Theorem 3 applies.
- Suppose the function \( F(x) \) is strongly monotone, \( (F(x) - F(\hat{x}))^T (x - \hat{x}) \geq \beta \| x - \hat{x} \|^2 \). This implies that the \( f_i(x) \) are \( \beta \)-strongly convex. Then, it follows that \( \nabla F(x) \geq \beta I_n \) for all \( x \in \mathbb{R}^n \). This also provides the following bound

\[
\| \nabla \Phi(x; \eta) \|^2 \geq (\beta)^2 \| F(x) \|^2 = 2 \beta^2 \Phi(x). \tag{7}
\]

Hence, \( \mu = \beta^2 \).
2. Extra Experiments

In this Section, we present more empirical results for the four different games that were discussed in the main paper which help understand the convergence behavior of the proposed method. More concretely, the results validate the results for convergence rate and the quality of solutions for the different games discussed in the main paper.

2.1. Convergence Rate for Quadratic Games

We provide plots that suggest linear convergence rate for strongly-convex quadratic games as was described in the main paper. Since bilinear games are a special case of the quadratic games, we also show results for bilinear games. For both cases we use 20-d variables for both players which are initialized arbitrarily. From the plots shown in Figure 1, we observe that the function decays linearly to zero (suggested by Theorem 2 in the main paper). We observe that the convergence slows down close to $V = 0$. It is noted that the guarantees for linear convergence are for the $V$ function (and not for $\nabla f$) and thus we skip plots for $\nabla f$.

![Figure 1. Convergence rate for bilinear and convex quadratic games using the GNI method. Left: Decay of $V$ function for Bilinear Game. Right: Decay of $V$ function for Strongly-convex Quadratic Game.](image1)

2.2. Two-Player Quadratic Games

We describe an experiment for non-convex two-player quadratic games with indefinite $Q$ matrices for both players. We show the decay of the gradient and the $V$ function for the GNI formulation. The other optimization algorithms are seen to be diverging for the indefinite cases (as was shown in the main text) and thus are not shown here. We used 50-d data, the same stepsize $\eta = \max_i (Q_i)$ and $\rho = 0.01$ for GNI. The methods are initialized randomly from $N(0, I)$. For clarification, we show the plot on log scale. As can be observed from the plots in Figure 2, $\nabla f$ goes to zero as $V$ goes to zero.

![Figure 2. Convergence of GNI method for non-convex quadratic game setting shown on a semi-log plot for clarity. Left: Decay of $V$ function. Right: Decay of $\nabla f$.](image2)

2.3. Dirac Delta GAN

In this section, we show another experiment for the Dirac Delta GAN that was discussed in the main text. All the parameters for all optimizers are kept constant as in the main text for Dirac Delta GAN. In Figure 3, we see the convergence of $\nabla f$ as well as the trajectories followed by the two players to the NE. All the methods converge to the same optima; however, the GNI converges faster than any other method. As observed in the convex quadratic case, we see all descent methods following the same trajectory except for the GNI and Adam. However, it was observed that the GNI and the other algorithms do not converge to the same solution when initialized arbitrarily. To investigate this, we perform an experiment where the game was initialized randomly from 1000 initial conditions in a square region in $[-4, 4] \times [-4, 4]$. The error from the ground truth was computed after 10000 iterations or up to convergence (the minimum of two). Results of the experiment are shown as a table in Figure 4. It is observed that the game doesn’t converge to the known ground truth for the game– Adam is able to get closest to the ground truth while GNI converges to a stationary Nash point much faster than all other algorithms. This behavior could be explained by recalling that GNI is using $V$ function to descend and thus, converges to the closest stationary Nash point where $V$ vanishes.

![Figure 3. Convergence of GNI against other methods on the Dirac-Delta GAN. Left: Convergence of different methods seen by the decay of $\nabla f$. Right: Trajectory of the two players to the optima.](image3)
2.4. Linear GAN

We also show some additional results for the Linear GAN which suggests convergence of the proposed method to a NE. The second derivative of the objective function for both the players is positive semidefinite (see Equation (23) in the main paper) indicating all stationary points are minimas. In the following plots in Figure 5, we show the convergence of the $V$ function and the $||\nabla f||$ for the GNI formulation. We observe very fast convergence for both the $V$ and the $||\nabla f||$ indicating convergence to a stationary Nash point. Additionally, since all stationary Nash points are NEs in this particular setting, the GNI converges to a NE.

Figure 5. Convergence of $V$ function and $||\nabla f||$ for the Linear GAN discussed in the main paper. Left: Decay of $V$ function. Right: Decay of the $\nabla f$. 