MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Game Theoretic Optimization via Gradient-based Nikaido-Isoda Function

Raghunathan, Arvind; Cherian, Anoop; Jha, Devesh K.

TR2019-045 June 28, 2019

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)

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

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 multi-player 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:

$$\begin{aligned} &\text{Find } \boldsymbol{x}^* = (x_1^\star, \ldots, x_N^\star) \text{ s.t.} \\ &x_i^\star = \arg \min_{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x}_{-i} = \boldsymbol{x}^*_{-i}} f_i(\boldsymbol{x}) \end{aligned} \tag{1}$$

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

A point $\boldsymbol{x}^\star$ satisfying (1) is called a *Nash Equilibrium* (NE). We denote by $\mathcal{S}^{NE}$ the set of all NE points, *i.e.,* $\mathcal{S}^{NE} = \{\boldsymbol{x}^\star \,|\, (1) \text{ holds}\}$. In the absence of convexity for the functions f_i we may not be able to obtain a minimizer in (1) and have to settle for a first-order stationary point. Accordingly, define $\mathcal{S}^{SNP}$ to be the set of all *Stationary Nash Points*, *i.e.,* $\mathcal{S}^{SNP} = \{\boldsymbol{x}^\star \,|\, \nabla_i f_i(\boldsymbol{x}^\star) = 0; \forall\, i = 1, \ldots, N\}$ where $\nabla_i f$ denotes the derivative of function f *w.r.t.* x_i.

There has been renewed interest in Nash equilibrium computation for games owing to the success of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). GANs have been successful in learning probability distributions and have found application in tasks including image-to-image translation (Isola et al., 2016), domain adaptation (Tzeng et al., 2017), probabilistic inference (Dumoulin et al., 2016; Mescheder et al., 2017) among others. Despite their popularity, GANs are known to be difficult to train. In order to stabilize training recent approaches have resorted to carefully designed models, either by adapting an architecture (Radford et al., 2015) or by selecting an easy-to-optimize objective function (Salimans et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017).

The Nikaido-Isoda (NI) function (Nikaido & Isoda, 1955) (formally introduced in §3) is popular in equilibrium computation (Uryasev & Rubinstein, 1994; Contreras et al., 2004; Facchinei & Kanzow, 2007; von Heusinger & Kanzow, 2009a;b) and often used as a merit function for NE. The evaluation of the NI function requires optimizing each player's problem globally which can be intractable for non-convex objectives.

In this paper, we introduce *Gradient-based Nikaido-Isoda* (GNI) function which allows us to computationally simplify the original NI formulation. Instead of computing a globally optimal solution, every player can locally improve their objectives using the steepest descent direction. The proposed GNI function simplifies the original NI formulation by relaxing the requirement on optimizing individual player's objective globally. We prove that GNI is a valid merit function for multi-player games and vanishes only at the first-order stationary points of each player's optimization problem (§3). The GNI function is shown to be locally stable in a neighborhood of a stationary Nash point (§3.1) and convex when the player's objective function is quadratic (§3.2). The gradient descent algorithm applied to the GNI

[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>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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 $\mathcal{S}^{SNP}$ to the case where the decision variables of player i's x_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 pseudoconvex 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 divergence 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

$$\psi(\boldsymbol{x}) = \sum_{i=1}^{N} \underbrace{\left(f_i(\boldsymbol{x}) - \inf_{\hat{\boldsymbol{x}} \in \mathbb{R}^n : \hat{\boldsymbol{x}}_{-i} = \boldsymbol{x}_{-i}} f_i(\hat{\boldsymbol{x}}) \right)}_{=:\psi_i(\boldsymbol{x})}.$$

From the definition of NI function $\psi(\boldsymbol{x})$, it is easy to show that $\psi(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Further, $\psi(\boldsymbol{x}) = 0$ is the global minimum which is only achieved if the NE point $\boldsymbol{x}^\star = (x_1^*, \ldots, x_N^\star)$ 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 $\psi(\boldsymbol{x})$ to compute NE points. However, the infimum needed to compute $\psi_i(\boldsymbol{x})$ can be prohibitive for all but a handful of functions. For bilinear min-max games (i.e., $f_1(\boldsymbol{x}) = x_1^T x_2 = -f_2(\boldsymbol{x})$), 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

$$V(\boldsymbol{x}; \eta) = \sum_{i=1}^{N} \underbrace{f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}(\boldsymbol{x}; i, \eta))}_{=:V_i(\boldsymbol{x};\eta)} \tag{2}$$

$$\text{where } y_j(\boldsymbol{x}; i, \eta) = \begin{cases} x_i - \eta \nabla_i f_i(\boldsymbol{x}), \text{ if } j = i \\ x_j, \text{ otherwise.} \end{cases}$$

where $\nabla_i f(\boldsymbol{x})$ 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 $\boldsymbol{y}(\boldsymbol{x}; i, \eta)$ in the steepest descent direction. This provides a local measure of decrease that can be obtained in the objective for player i. The point $\boldsymbol{y}(\boldsymbol{x}; i, \eta)$ 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 $V_i(\boldsymbol{x}; \eta) = 0$ also satisfies $\nabla_i f_i(\boldsymbol{x}) = 0$. To show this, we first provide bounds on $V_i(\boldsymbol{x}; \eta)$ 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., ∇f_i) are Lipschitz continuous with constant L_f.*

Lemma 1. $\frac{\eta}{2}\|\nabla_i f_i(\boldsymbol{x})\|^2 \le V_i(\boldsymbol{x}; \eta) \le \frac{3\eta}{2}\|\nabla_i f_i(\boldsymbol{x})\|^2$ *for all $\boldsymbol{x} \in \mathbb{R}^n$ and $0 < \eta \le \frac{1}{L_f}$.*

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

$$f_i(\boldsymbol{y}(\boldsymbol{x}; i, \eta)) = f_i(\boldsymbol{x}) - \eta\|\nabla_i f_i(\boldsymbol{x})\|^2 + \eta^2 \int_0^1 \nabla_i f_i(\boldsymbol{x})^T \nabla_i^2 f_i(\hat{\boldsymbol{x}}(t)) \nabla_i f_i(\boldsymbol{x}) t dt$$

where $\hat{x}_i(t) = x_i - t\eta\nabla_i f_i(\boldsymbol{x})$ and $\hat{x}_j(t) = x_j$, for $j \ne i$. From the Lipschitz continuity of the gradient of f_i, we have that $-L_f I_i \preceq \nabla_i^2 f_i(\hat{\boldsymbol{x}}(t)) \preceq L_f I_i$, where I_i is the $(n_i \times n_i)$ identity matrix. Substituting in the above and using $\eta \le \frac{1}{L_f}$ 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's optimization problems.

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

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

Theorem 1 shows that the function $V(\boldsymbol{x}; \eta)$ can be employed as a merit function for obtaining a stationary Nash point. When $f_i(\boldsymbol{x})$ are non-convex, the convergence to first-order point is possibly the best that one can hope for.

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

$$\nabla V_i(\boldsymbol{x}; \eta) = \nabla f_i(\boldsymbol{x}) - \left(I - \eta \nabla^2 f_i(\boldsymbol{x}) E_i\right) \nabla f_i(\boldsymbol{y}(\boldsymbol{x}; i, \eta)) \tag{3}$$

where $E_i = F_i F_i^T$ with $F_i \in \mathbb{R}^{n \times n_i}$ defined as $F_i^T = \left[\mathbf{0}_{n_i \times \sum_{j=1}^{i-1} n_j} \quad I_i \quad \mathbf{0}_{n_i \times \sum_{j=i+1}^{N} n_j}\right]$, $I \in \mathbb{R}^{n \times n}$, and $I_i \in \mathbb{R}^{n_i \times n_i}$ are identity matrices. The Hessian of $V_i(\boldsymbol{x}; \eta)$ is given by

$$\begin{aligned}\nabla^2 V_i(\boldsymbol{x}; \eta) &= \nabla^2 f_i(\boldsymbol{x}) + \eta \nabla^3 f_i(\boldsymbol{x})[E_i \nabla f_i(\boldsymbol{y}_i(\boldsymbol{x}; i, \eta)] \\ &- (I - \eta \nabla^2 f_i(x) E_i) \nabla^2 f_i(\boldsymbol{y}(\boldsymbol{x}; i, \eta))(I - \eta E_i \nabla^2 f_i(\boldsymbol{x}))\end{aligned} \tag{4}$$

where $\nabla^3 f_i(\boldsymbol{x})[d] = \lim_{\alpha \to 0} \frac{\nabla^2 f_i(\boldsymbol{x} + \alpha d) - \nabla^2 f_i(\boldsymbol{x})}{\alpha}$ 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 decent algorithm of Rakhlin and Sridharan (Rakhlin & Sridharan, 2013) was shown to be convergent. Daskalakis and Pangeas (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 $V(\boldsymbol{x}; \eta)$ is positive semidefinite. This ensures that the points in $\mathcal{S}^{SNP}$ are all stable limit points for the gradient descent algorithm on $V(\boldsymbol{x}; \eta)$.

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

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

$$\begin{aligned}\nabla^2 V_i(\boldsymbol{x}^\star;\eta) &= 2\eta\nabla^2 f_i(\boldsymbol{x}^\star)E_i\nabla^2 f_i(\boldsymbol{x}^\star)\\ &\quad - \eta^2\nabla^2 f_i(\boldsymbol{x}^\star)E_i\nabla^2 f_i(\boldsymbol{x}^\star)E_i\nabla^2 f_i(\boldsymbol{x}^\star)\\ &= \eta\nabla^2 f_i(\boldsymbol{x}^\star)(2E_i - \eta E_i\nabla^2 f_i(\boldsymbol{x}^\star)E_i)\nabla^2 f_i(\boldsymbol{x}^\star),\end{aligned} \tag{5}$$

From the Lipschitz continuity of $f_i(\boldsymbol{x})$ we have that $\nabla^2 f_i(\boldsymbol{x}^\star) \preceq L_f I_n$. Substituting into (5), we obtain

$$\begin{aligned}\nabla^2 V_i(\boldsymbol{x}^\star;\eta) &\succeq \eta\nabla^2 f_i(\boldsymbol{x}^\star)(2E_i - (\eta L_f)E_i^2)\nabla^2 f_i(\boldsymbol{x}^\star)\\ &\succeq \eta\nabla^2 f_i(\boldsymbol{x}^\star)E_i\nabla^2 f_i(\boldsymbol{x}^\star)\end{aligned}$$

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 $\nabla^2 f_i(\boldsymbol{x}^\star)E_i\nabla^2 f_i(\boldsymbol{x}^\star)$. Since $\nabla^2 V(\boldsymbol{x}^*;\eta)$ 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.,* $f_i(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}_i\boldsymbol{x} + \boldsymbol{r}_i^T\boldsymbol{x}$. Then, the GNI function is

$$\begin{aligned}&V_i(\boldsymbol{x}) = f_i(x) - f_i(\boldsymbol{x} - \eta E_i(\boldsymbol{Q}_i\boldsymbol{x} + \boldsymbol{r}_i)) \qquad (6)\\ &= \frac{1}{2}\boldsymbol{x}^T\left(\boldsymbol{Q}_i - \widehat{\boldsymbol{Q}}_i^T\boldsymbol{Q}_i\widehat{\boldsymbol{Q}}_i\right)\boldsymbol{x} + \eta\boldsymbol{r}_i^T E_i\boldsymbol{Q}_i(I + \widehat{\boldsymbol{Q}}_i)\boldsymbol{x}\\ &\quad + \frac{1}{2}\eta\boldsymbol{r}_i^T(2E_i - \eta E_i\boldsymbol{Q}_i E_i)\boldsymbol{r}_i\end{aligned}$$

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

$$(\boldsymbol{Q}_i - \widehat{\boldsymbol{Q}}_i^T\boldsymbol{Q}_i\widehat{\boldsymbol{Q}}_i) = \eta(\boldsymbol{Q}_i E_i)(2I - \eta\boldsymbol{Q}_i)(E_i\boldsymbol{Q}_i) \succeq 0 \tag{7}$$

where the positive semidefiniteness holds since for all $u \neq 0$ $u^T(\boldsymbol{Q}_i E_i)(2I - \eta\boldsymbol{Q}_i)(E_i\boldsymbol{Q}_i)u = (\boldsymbol{Q}_i E_i u)^T(2I - \eta\boldsymbol{Q}_i)(\boldsymbol{Q}_i E_i u) \geq 0$. Hence, when $f_i(\boldsymbol{x})$ 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 $f_i(\boldsymbol{x})$. However, for general nonlinear functions $f_i(\boldsymbol{x})$, the GNI function $V_i(\boldsymbol{x})$ does not preserve convexity.

4. Descent Algorithm for GNI

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

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \rho\nabla V(\boldsymbol{x}^k;\eta) \text{ for } k = 0, 1, 2, \ldots \tag{8}$$

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

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

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

$$\begin{aligned}V(\boldsymbol{x}^{k+1};\eta) &\leq V(\boldsymbol{x}^k;\eta) + \nabla V(\boldsymbol{x}^k;\eta)^T(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)\\ &\quad + \frac{L_V}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2\\ &\leq V(\boldsymbol{x}^k,\eta) - \rho(1 - \frac{\rho L_V}{2})\|\nabla V(\boldsymbol{x};\eta)\|^2\\ &\leq V(\boldsymbol{x}^k;\eta) - \frac{\overline{\alpha}}{2L_V}\|\nabla V(\boldsymbol{x};\eta)\|^2\end{aligned} \tag{9}$$

where $\overline{\alpha} = \alpha(2-\alpha)$. Telescoping the sum for $k = 0, ..., K$, we obtain

$$V(\boldsymbol{x}^{K+1};\eta) \leq V(\boldsymbol{x}^0) - \frac{\overline{\alpha}}{2L_V}\sum_{k=0}^{K}\|V(\boldsymbol{x}^K;\eta)\|^2. \tag{10}$$

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

$$\begin{aligned}&\frac{\overline{\alpha}}{2L_V}\sum_{k=0}^{K}\|\nabla V(\boldsymbol{x}^K;\eta)\|^2 \leq V(\boldsymbol{x}^0) - V(\boldsymbol{x}^{K+1}) \leq V(\boldsymbol{x}^0)\\ &\implies \frac{\overline{\alpha}}{2L_V}\min_{k\in\{0,\ldots,K\}}\|\nabla V(\boldsymbol{x}^k;\eta)\|^2 \leq \frac{V(\boldsymbol{x}^0)}{K+1}.\end{aligned}$$

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

$$V(\boldsymbol{x}^{k+1};\eta) \leq \left(1 - \overline{\alpha}\frac{\mu}{L_V}\right)V(\boldsymbol{x}^k;\eta) \tag{11}$$

which proves the claim on linear convergence of $\{V(\boldsymbol{x}^k;\eta)\}$ to 0. By Theorem 1, $\{\boldsymbol{x}^k\}$ converges to $\boldsymbol{x}^\star \in \mathcal{S}^{SNP}$. □

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 $f_i(\boldsymbol{x})$'s to be quadratic. For this setting §3.2 showed that GNI function $V_i(\boldsymbol{x})$ is a convex quadratic function. This proves that $V_i(\boldsymbol{x};\eta)$ 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 $V(\boldsymbol{x};\eta) \le \frac{1}{2\mu}\|\nabla V(\boldsymbol{x};\eta)\|^2$ holds. Hence, we can state the following stronger result for quadratic objective functions.

Corollary 1. *Suppose $f_i(\boldsymbol{x})$ are quadratic and player convex, i.e. $f_i(\boldsymbol{x})$ is convex in x_i. Let $\rho = \frac{1}{3L_f N}$. Then, the sequence $\{V(\boldsymbol{x}^k)\}$ converges linearly to 0,* i.e. *$\{\boldsymbol{x}^k\}$ converges to $\boldsymbol{x}^\star \in \mathcal{S}^{NE}$.*

5. Modified Descent Algorithm for GNI

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

$$\begin{aligned}&\nabla^2 f_i(\boldsymbol{x})(\eta E_i \nabla f_i(\boldsymbol{x} - \eta E_i \nabla f_i(\boldsymbol{x})))\\ &\approx \nabla f_i(\boldsymbol{x} + \eta E_i \nabla f_i(\boldsymbol{x} - \eta E_i \nabla f_i(\boldsymbol{x}))) - \nabla f_i(\boldsymbol{x})\end{aligned} \quad (12)$$

Substituting (12) for the term involving the Hessian in $\nabla V_i(\boldsymbol{x};\eta)$ and simplifying obtain the direction $\nabla \widehat{V}_i(\boldsymbol{x};\eta)$:

$$\begin{aligned}\nabla \widehat{V}_i(\boldsymbol{x};\eta) = &\nabla f_i(\boldsymbol{x} + \eta E_i \nabla f_i(\boldsymbol{x} - \eta E_i \nabla f_i(\boldsymbol{x})))\\ &- \nabla f_i(\boldsymbol{x} - \eta E_i \nabla f_i(\boldsymbol{x}))\end{aligned} \quad (13)$$

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

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \rho \nabla \widehat{V}_i(\boldsymbol{x};\eta) \text{ for } k = 0,1,2,\ldots \quad (14)$$

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

$$\|\nabla \widehat{V}(\boldsymbol{x};\eta) - \nabla V(\boldsymbol{x};\eta)\| \le \tau \|\nabla V(\boldsymbol{x};\eta)\|, \quad (15)$$

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

Proof. Let $\nabla \widehat{V}(\boldsymbol{x}^k;\eta) = \nabla V(\boldsymbol{x}^k;\eta) + e^k$. From (15), $\|e^k\| \le \tau\|\nabla V(\boldsymbol{x}^k;\eta)\|$. Applying the triangle inequality to $\|\nabla \widehat{V}(\boldsymbol{x}^k;\eta)\|$ and use (15) obtain

$$\begin{aligned}\|\nabla \widehat{V}(\boldsymbol{x}^k)\| &\le \|\nabla V(\boldsymbol{x}^k;\eta) + \|e^k\|\\ &\le (1+\tau)\|\nabla V(\boldsymbol{x}^k;\eta)\|.\end{aligned} \quad (16)$$

The term $-(\nabla V(\boldsymbol{x}^k;\eta))^T(\nabla \widehat{V}(\boldsymbol{x}^k;\eta))$ can be upper bounded as

$$\begin{aligned}&-(\nabla V(\boldsymbol{x}^k;\eta))^T(\nabla \widehat{V}(\boldsymbol{x}^k;\eta))\\ =&-\|\nabla V(\boldsymbol{x}^k;\eta)\|^2 - (\nabla V(\boldsymbol{x}^k;\eta))^T e^k\\ \le&-\|\nabla V(\boldsymbol{x}^k;\eta)\|^2 + \|\nabla V(\boldsymbol{x}^k;\eta)\|\|e^k\|\\ \le&-(1-\tau)\|\nabla V(\boldsymbol{x}^k;\eta)\|^2\end{aligned} \quad (17)$$

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

$$\begin{aligned}&V(\boldsymbol{x}^{k+1};\eta)\\ \le& V(\boldsymbol{x}^k;\eta) + \nabla V(\boldsymbol{x}^k;\eta)^T(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)\\ &+ \frac{L_V}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2\\ \le& V(\boldsymbol{x}^k,\eta) - \rho(\nabla V(\boldsymbol{x}^k;\eta))^T(\nabla \widehat{V}(\boldsymbol{x}^k;\eta)) +\\ &\frac{L_V\rho^2}{2}\|\nabla \widehat{V}(\boldsymbol{x};\eta)\|^2\\ \le& V(\boldsymbol{x}^k,\eta) - \rho(1-\tau)\|\nabla V(\boldsymbol{x}^k;\eta)\|^2 +\\ &\frac{L_V\rho^2}{2}(1+\tau)^2\|\nabla V(\boldsymbol{x};\eta)\|^2\\ \le& V(\boldsymbol{x}^k;\eta) - \rho\left(1 - \tau - \frac{L_V\rho(1+\tau)^2}{2}\right)\|\nabla V(\boldsymbol{x};\eta)\|^2\\ \le& V(\boldsymbol{x}^k;\eta) - \frac{\overline{\alpha}}{2}\frac{(1-\tau)^2}{L_V(1+\tau)^2}\|\nabla V(\boldsymbol{x};\eta)\|^2\end{aligned} \quad (18)$$

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

$$\overline{\alpha}\frac{(1-\tau)^2}{2L_V(1+\tau)^2}\min_{k\in\{0,\ldots,K\}}\|\nabla V(\boldsymbol{x}^k;\eta)\|^2 \le \frac{V(\boldsymbol{x}^0)}{K+1}.$$

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

$$V(\boldsymbol{x}^{k+1};\eta) \le \left(1 - \overline{\alpha}\frac{\mu(1-\tau)^2}{L_V(1+\tau)^2}\right)V(\boldsymbol{x}^k;\eta) \quad (19)$$

which proves the claim on linear convergence of $\{V(\boldsymbol{x}^k;\eta)\}$ to 0. By Theorem 1, $\{\boldsymbol{x}^k\}$ converges to $\boldsymbol{x}^\star \in \mathcal{S}^{SNP}$. $\square$

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:

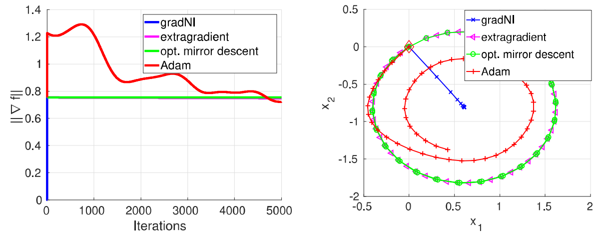
We consider the following two-player game:

$$f_1(x) = x_1^T Q x_2 + q_1^T x_1 + q_2^T x_2 = -f_2(x), \tag{20}$$

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 $\|\nabla f(\boldsymbol{x}^k)\| = \|(\nabla_1 f_1(\boldsymbol{x}^k), \ldots, \nabla_N f_N(\boldsymbol{x}^k))\|$. 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 $x_1, x_2 \in \mathbb{R}^1$, and plot them for every 100-th iteration starting from the same initial point (shown by the red-diamond). Interestingly,

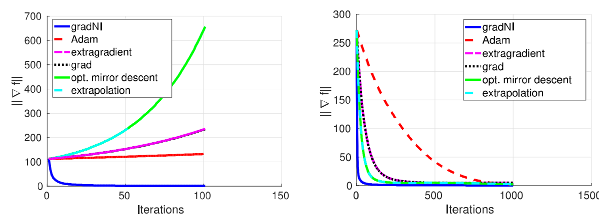
[1]Other values of η did not seem to result in stable descent.



(a) (b)

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).



(a) non-convex QP (b) convex-QP

Figure 2. Convergence of GNI against other methods for Quadratic games. (a) Non-convex QP with indefinite Q matrices for each player, (b) convex QP with semi-definite Q matrices.

6.2. Two-Player Quadratic Games:

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

$$f_i(x) = \frac{1}{2} x^T Q_i x + r_i^T x, \text{ for } i = 1, 2 \tag{21}$$

where $Q_i \in \mathbb{R}^{n \times n}$ is symmetric. We consider cases when each Q_i is indefinite (i.e., non-convex QP) and positive semi-definite. As with the bilinear case, all the QP payoffs result in convex GNI reformulations. We used 20-d data, the same stepsizes $\eta = \max_i(\|Q_i\|)$ 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 QP 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.

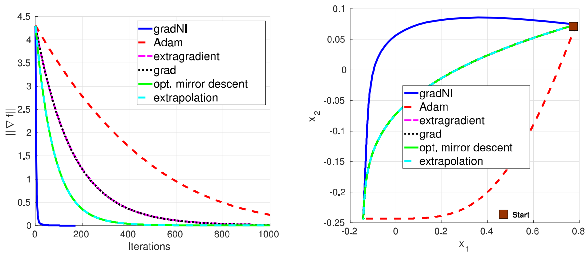


Figure 3. Convergence of GNI against other methods on a convex $1-$d quadratic game. Left: the convergence achieved by different algorithms. Right: the trajectories of the two players to the NE.

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:

$$\begin{aligned} f_1 &= \log(1+\exp(\theta x_1)) + \log(1+\exp(x_1 x_2)) \\ f_2 &= -\log(1+\exp(x_1 x_2)), \end{aligned} \tag{22}$$

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 Lipscitz 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).

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 exposit 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

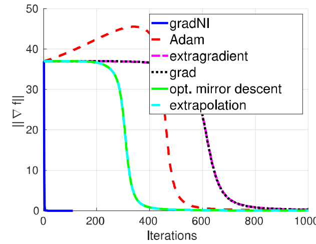


Figure 4. Convergence of GNI against other methods on the Dirac-Delta GAN.

L. While, our proposed setting is not a neural network, it allows to understand the behavior of GNI when other non-linearities arising from the layers of a neural network are absent, and thereby study GNI in isolation.

Experimental Setup: The payoff functions are:

$$\begin{aligned} f_1 &= -\mathbb{E}_{\theta\sim P_r}\log\left(x_1^T\theta\right) - \mathbb{E}_{z\sim P_z}\log\left(1 - x_1^T \operatorname{diag}(x_2)\, z\right), \\ f_2 &= -\mathbb{E}_{z\sim P_z}\log\left(x_2^T \operatorname{diag}(x_1)\, z\right), \end{aligned} \tag{23}$$

where P_r and P_z are the real and the noise data distributions, the latter being the standard normal distribution $N(0, I)$. The operator diag returns a diagonal matrix with its argument as its diagonal. We consider two cases for P_r: (i) $P_r = N(\mu, I)$ for a mean μ and (ii) $P_r = N(\mu, \Sigma)$ 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.

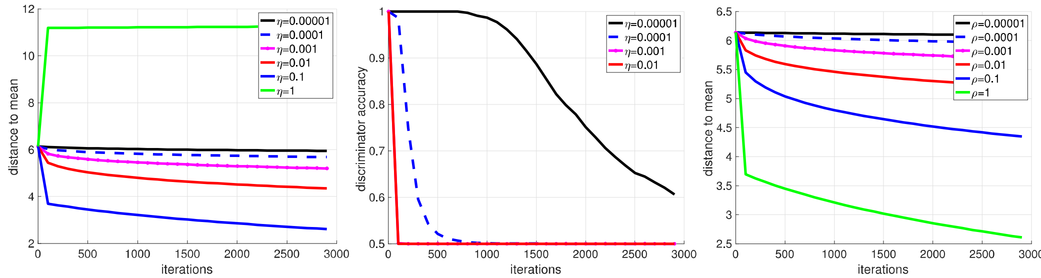
Evaluation Metrics: To evaluate the performance on various hyper-parameters of GNI, we define two metrics: (i) *discriminator-accuracy*, and (ii) the *distance-to-mean*. The discriminator-accuracy measures how well the learned discriminator classifies the two distributions, defined as:

$$d_{acc} = \frac{1}{2M}\sum_{i=1}^{M}\mathcal{I}(x_1^T\theta_i \geq \zeta) + \mathcal{I}\left(x_1^T \operatorname{diag}(x_2^i) z_i \leq (1-\zeta)\right),$$

where $\mathcal{I}$ 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_{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:

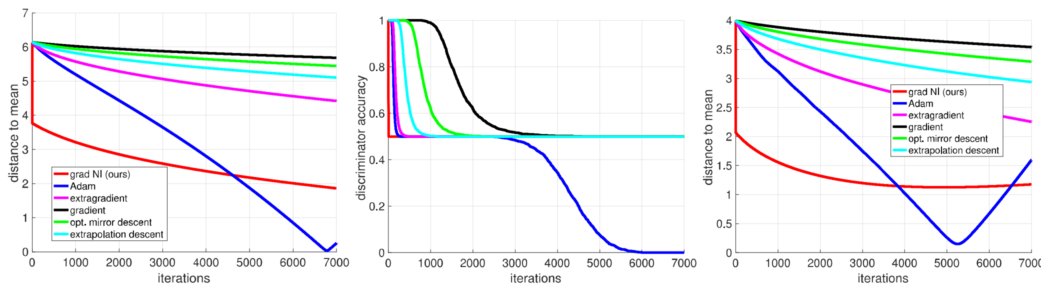
$$d_{mean} = \|\mathbb{E}_{z\sim P_z}\operatorname{diag}(x_2)z - \mathbb{E}_{\theta\sim P_r}\theta\| \tag{24}$$

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,



(a) Varying η, $\rho = 1$ (b) varying η, $\rho = 1$ (c) varying ρ, $\eta = 0.1$

Figure 5. Study of the influence of the step sizes (ρ and η) on the convergence of GNI reformulations for the linear GAN game.



(a) Generator, $P_r = N(\mu, I)$ (b) Discriminator, $P_r = N(\mu, I)$ (c) Generator, $P_r = N(\mu, \Sigma)$

Figure 6. Convergence of GNI against other methods on the linear GAN two-player game. The real-data distribution is sampled from $N(\mu, I)$ for (a) and (b), while we use $N(\mu, \Sigma)$ for (c), where $\Sigma = \text{diag}(\xi), \xi \sim U(0, 1]$. Note that, when the optimization converges, the discriminator is expected to be confused between the real and fake data distributions (i.e., classification accuracy is 0.5).

we vary η and ρ 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 η 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, $\eta > 1$ leads to a generator breakdown (Figure 5(a)). Similarly, a small value of ρ, 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 ρ 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.

Comparisons to Other Algorithms: In Figures 6(a) and (b), we plot the distance-to-mean and discriminator-accuracy of linear GAN using $\eta = 0.1$ and $\rho = 1$, and compare it to all other descent schemes. Interestingly, we found that Adam shows a different pattern of convergence, with the distance-to-mean steadily decreasing to zero; on close inspection (Figure 6(b)), we see that the discriminator-accuracy simultaneously goes to zero as well, suggesting the non-optimality of the descent. In contrast, our GNI converges quickly. In Figure 6(c), we plot the convergence when using a real data distribution $P_r = N(\mu, \Sigma)$, where $\mu \sim U(0, 1)^d$; a d-dimensional uniform distribution and Σ is a randomly-sampled diagonal covariance matrix. The descent in this general setting also looks similar to the one in Figure 6(a).

7. Conclusions

We presented a novel formulation for Nash equilibrium computation in multi-player games by introducing the Gradient-based Nikaido-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

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint*, arXiv:1701.07875, 2017.

Basar, T. and Olsder, G. J. *Dynamic noncooperative game theory*, volume 23. Siam, 1999.

Bervoets, S., Bravo, M., and Faure, M. Learning with minimal information in continuous games. *arXiv*, arXiv:1806.11506, 2018. URL https://arxiv.org/abs/1806.11506.

Contreras, J., Klusch, M., and Krawczyk, J. Numerical solutions to nash-cournot equilibria in coupled constraint electricity markets. *IEEE Transactions on Power Systems*, 19:195206, 2004.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. *arXiv preprint*, arXiv:1807.03907v1, 2018.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *arXiv preprint*, arXiv:1711.00141v2, 2018.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. Adversarially learned inference. *arXiv preprint*, arXiv:1606.00704, 2016.

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

Facchinei, F. and Kanzow, C. Penalty methods for the solution of generalized nash equilibrium problems. *SIAM Journal on Optimization*, 20(5):2228–2253, 2010.

Facchinei, F. and Pang, J.-S. *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I*. Springer, New York, NY, 2003a.

Facchinei, F. and Pang, J.-S. *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II*. Springer, New York, NY, 2003b.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., and Goodfellow, I. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. In *ICLR*, 2018.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint*, arXiv:1802.10551v3, 2018.

Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint*, arXiv:1701.00160, 2016.

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

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of Wasserstein GANs. *arXiv preprint*, arXiv:1704.00028, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, arXiv:1611.07004, 2016.

Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. *arXiv preprint*, arXiv:1608.04636v3, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Korpelevich, G. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Li, Y., Schwing, A., K.-C.!Wang, and Zemel, R. Dualizing gans. In *NIPS*, 2017.

Łojasiewicz, S. A topological property of real analytic subsets (in french). *Coll. du CNRS, Les équations aux dérivées partielles*, pp. 8789, 1963.

Luo, Z.-Q. and Tseng, P. Error bounds and convergence analysis of feasible descent methods: A general approach. *Ann. Oper. Res.*, pp. 157178, 1993.

Mertikopoulos, P. and Zhou, Z. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1–2):465–507, 2019.

Mertikopoulos, P., Zenati, H., Lecouat, B., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representation (ICLR)*, 2019.

Mescheder, L., Nowozin, S., and Geiger, A. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint*, arXiv:1701.04722, 2017.

Nagarajan, V. and Kolter, J. Gradient descent gan optimization is locally stable. *arXiv preprint*, arXiv:1706.04156v3, 2017.

Nikaido, H. and Isoda, K. Note on noncooperative convex games. *Pacific Journal of Mathematics*, 5(1):807815, 1955.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2006.

Polyak, B. T. Gradient methods for minimizing functionals (in russian). *Zh. Vychisl. Mat. Mat. Fiz.*, pp. 643653, 1963.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint*, 2015.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *COLT 2013 - The 26th Annual Conference on Learning Theory*, pp. 993–1019, 2013.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. *CoRR*, abs/1606.03498, 2016.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. *arXiv preprint*, arXiv:1702.05464, 2017.

Uryasev, S. and Rubinstein, R. On relaxation algorithms in computation of noncooperative equilibria. *IEEE Transactions on Automatic Control*, 39(6):12631267, 1994.

von Heusinger, A. and Kanzow, C. Optimization reformulations of the generalized nash equilibrium problem using nikaido-isoda-type functions. *Comput. Optim. Appl.*, 43 (3):353377, 2009a.

von Heusinger, A. and Kanzow, C. Relaxation methods for generalized nash equilibrium problems with inexact line search. *Journal of Optimization Theory and Applications*, 143(1):159183, 2009b.

Yadav, A., Shah, S., Xu, Z., Jacobs, D., and Goldstein, T. Stabilizing adversarial nets with prediction methods. *arXiv preprint*, arXiv:1705.07364, 2017.

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 $\Phi(\boldsymbol{x}) = \frac{1}{2}\sum_{i=1}^{2}\|\nabla_i f_i(\boldsymbol{x})\|^2$. We can state a result that is analogous to Theorem 1.

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

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

$$\begin{aligned}\nabla\Phi(\boldsymbol{x}) &= \nabla F(\boldsymbol{x})F(\boldsymbol{x})\\ &= \begin{bmatrix}\nabla^2_{11}f_1(\boldsymbol{x}) & \nabla^2_{12}f_2(\boldsymbol{x})\\ \nabla^2_{21}f_1(\boldsymbol{x}) & \nabla^2_{22}f_2(\boldsymbol{x})\end{bmatrix}\begin{bmatrix}\nabla_1 f_1(\boldsymbol{x})\\ \nabla_2 f_2(\boldsymbol{x})\end{bmatrix}.\end{aligned} \tag{1}$$

The Hessian of the function $\Phi(\boldsymbol{x})$ is

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

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

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

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

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

Proof. From Lipschitz continuity of $\nabla\Phi(\boldsymbol{x})$

$$\begin{aligned}\Phi(\boldsymbol{x}^{k+1}) \le&\Phi(\boldsymbol{x}^k) + \nabla\Phi(\boldsymbol{x}^k)^T(\boldsymbol{x}^{k+1} - x^k)\\ &+ \frac{L_\Phi}{2}\|x^{k+1} - x^k\|^2\\ \le&\Phi(\boldsymbol{x}^k,) - \rho(1 - \frac{\rho L_\Phi}{2})\|\nabla\Phi(\boldsymbol{x})\|^2\\ \le&\Phi(\boldsymbol{x}^k) - \frac{\overline{\alpha}}{2L_\Phi}\|\nabla\Phi(\boldsymbol{x})\|^2\end{aligned} \tag{4}$$

where $\overline{\alpha} = \alpha(2-\alpha)$. Telescoping the sum and $k = 0, ..., K$ obtain

$$\Phi(\boldsymbol{x}^{K+1}) \le \Phi(\boldsymbol{x}^0) - \frac{\overline{\alpha}}{2L_\Phi}\sum_{k=0}^{K}\|\Phi(\boldsymbol{x}^K)\|^2. \tag{5}$$

Since $\Phi(\boldsymbol{x})$ is bounded below by 0 we have that

$$\begin{aligned}&\frac{\overline{\alpha}}{2L_\Phi}\sum_{k=0}^{K}\|\nabla\Phi(\boldsymbol{x}^K)\|^2 \le \Phi(\boldsymbol{x}^0) - \Phi(\boldsymbol{x}^{K+1}) \le \Phi(\boldsymbol{x}^0)\\ \implies &\frac{\overline{\alpha}}{2L_\Phi}\min_{k\in\{0,\dots,K\}}\|\nabla\Phi(\boldsymbol{x}^k)\|^2 \le \frac{\Phi(\boldsymbol{x}^0)}{K+1}.\end{aligned}$$

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

$$\Phi(\boldsymbol{x}^{k+1}) \le \left(1 - \overline{\alpha}\frac{\mu}{L_\Phi}\right)\Phi(\boldsymbol{x}) \tag{6}$$

which proves the claim on linear convergence. ∎

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

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

  $$\|\nabla\Phi(x;\eta)\|^2 \ge (\beta)^2\|F(x)\|^2 = 2\beta^2\Phi(x). \tag{7}$$

  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 V 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 ∇f) and thus we skip plots for ∇f.

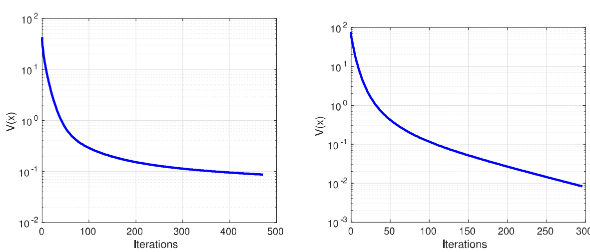


Figure 1. Convergence rate for bilinear and convex quadratic games using the GNI method. Left: Decay of V function for Bilinear Game. Right: Decay of V function for Strongly-convex Quadratic Game.

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 stepsizes $\eta = \max_i(Q_i)$ 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, ∇f goes to zero as V goes to zero.

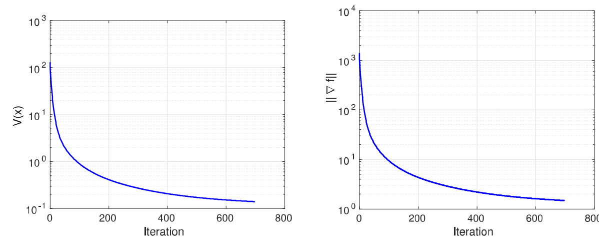


Figure 2. Convergence of GNI method for non-convex quadratic game setting shown on a semi-log plot for clarity. Left: Decay of V function. Right: Decay of the ∇f.

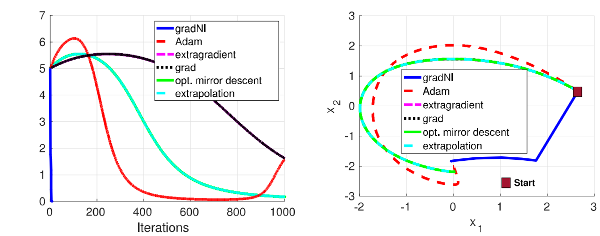


Figure 3. Convergence of GNI against other methods on the Dirac-Delta GAN. Left: Convergence of different methods seen by the decay of ∇f. Right: Trajectory of the two players to the optima.

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 ∇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 on 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.

Algorithm	GNI	Adam	ExGrad	Grad	OMD	ExPol
Mean Error	2.11	0.64	2.17	2.18	1.87	1.87
Mean number of Iterations	77	3048	10000	10000	10000	10000

Figure 4. Error Statistics for GNI compared against other techniques for the Dirac Delta GAN

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.

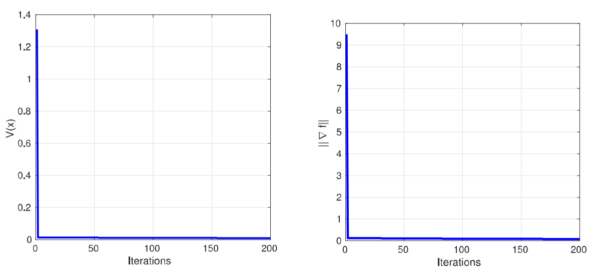


Figure 5. Convergence of V function and $||\nabla f||$ for the Linear GAN discussed in the main paper. Left: Decay of V function. Right: Decay of the ∇f.