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

Arvind U. Raghunathan Anoop Cherian Devesh K. Jha

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 $\rho > 0$

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

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 $\rho = \frac{1}{L_\Phi}$. Then, the $\{x^k\}$ generated by (3) converges sublinearly to $x^\star$ a first-order critical 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 $\nabla\Phi(\boldsymbol{x})$

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

Telescoping the sum and $k = 0, ..., K$ obtain

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

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

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

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

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

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. Additional 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 Bilinear and Quadratic Games:

We provide plots that suggest linear convergence rate for bilinear and strongly-convex quadratic games as was described in the main paper. 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 close to zero and then it slows down as the gradient of V starts to vanish (suggested by Theorem 2 in the main paper). 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.

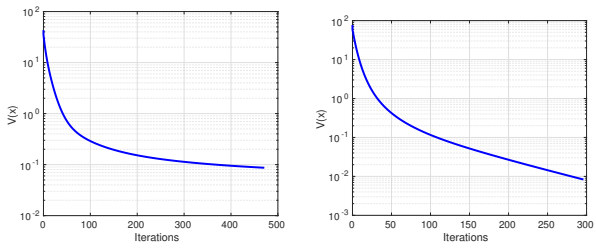

Figure 1. Convergence rate for bilinear and convex quadratic games using the GNI method. Left: Decay of V function for Bilinear Game. Right: Decay of V function for Strongly-convex Quadratic Game.

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.

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

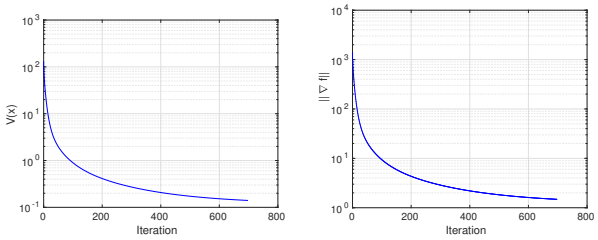

Figure 2. Convergence of GNI method for non-convex quadratic game setting shown on a semi-log plot for clarity. Left: Decay of V function. Right: Decay of the ∇f.

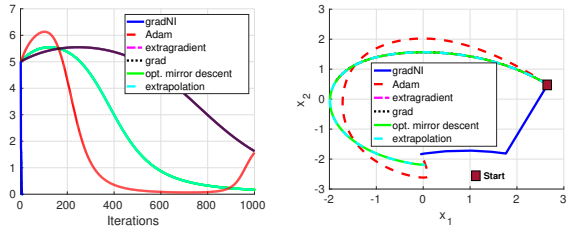

Figure 3. Convergence of GNI against other methods on the Dirac-Delta GAN. Left: Convergence of different methods seen by the decay of ∇f. Right: Trajectory of the two players to the optima.

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.

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

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

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 SNP. Additionally, since all SNPs are NEs in this particular setting, the GNI converges to a NE.

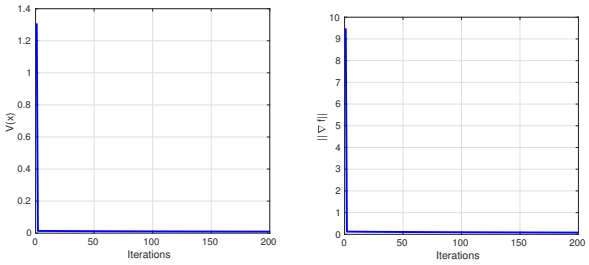

Figure 5. Convergence of V function and $||\nabla f||$ for the Linear GAN discussed in the main paper. Left: Decay of V function. Right: Decay of the ∇f.