On Solving Local Minimax Optimization: A Follow-the-Ridge Approach

Yuanhao Wang[*1], **Guodong Zhang**[*2,3], **Jimmy Ba**[2,3]
[1]IIIS, Tsinghua University, [2]University of Toronto, [3]Vector Institute
yuanhao-16@mails.tsinghua.edu.cn, {gdzhang,jba}@cs.toronto.edu

Abstract

Many tasks in modern machine learning can be formulated as finding equilibria in *sequential* games. In particular, two-player zero-sum sequential games, also known as minimax optimization, have received growing interest. It is tempting to apply gradient descent to solve minimax optimization given its popularity in supervised learning. However, we note that naive application of gradient descent fails to find local minimax – the analogy of local minima in minimax optimization, since the fixed points of gradient dynamics might not be local minimax. In this paper, we propose *Follow-the-Ridge* (FR), an algorithm that locally converges to and only converges to local minimax. We show theoretically that the algorithm addresses the limit cycling problem around fixed points, and is compatible with preconditioning and *positive* momentum. Empirically, FR solves quadratic minimax problems and improves GAN training on simple tasks.

1 Introduction

We consider differentiable *sequential* games with two players: a leader who can commit to an action, and a follower who responds after observing the leader's action. Particularly, we focus on the zero-sum case of this problem which is also known as minimax optimization, *i.e.*,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{y}\in\mathbb{R}^m} f(\mathbf{x},\mathbf{y}).$$

Applications of minimax optimization include generative adversarial networks (GANs) [9, 1], adversarial training [14] and primal-dual reinforcement learning [6, 3]. In these machine learning applications, finding the global minimax may be intractable due to non-convex landscapes. It is then natural to consider finding the local surrogate, known as local minimax [11].

The vanilla algorithm for minimax optimization is gradient descent-ascent (GDA), *i.e.*, both players take a gradient update simultaneously. GDA is known to suffer from two drawbacks. First, it has undesirable convergence properties: it fails to converge to some local minimax, and can converge to fixed points that are not local minimax [11, 5]. Second, GDA exhibits strong rotation around fixed points, which requires using very small learning rates [16, 2] to converge.

In this paper, we propose *Follow-the-Ridge* (FR), an algorithm for minimax optimization that addresses both issues. We summarize our main contributions as follows:

- We prove that FR has exact local convergence to local minimax points. Previously, this property is only known to be satisfied when the leader moves infinitely slower than the follower in gradient descent-ascent [11, 7].
- We show that FR addresses the limit cycling problem around fixed points and hence allows us to use a much larger learning rate.

[*]These two authors contributed equally.

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada. Full version available at https://arxiv.org/abs/1910.07512.

- We prove that our algorithm – FR is compatible with standard acceleration techniques such as preconditioning and *positive* momentum, which can speed up convergence significantly.
- We further show that our algorithm can be adapted to non-zero-sum games (general Stackelberg games [7, 23]) with similar theoretical guarantees.
- Finally, we demonstrate empirically the effectiveness of our algorithm in both quadratic minimax problems and GAN training.

2 Preliminaries

We consider sequential games with two players where one player is deemed the *leader* and the other the *follower*. We assume that the leader's action is $\mathbf{x} \in \mathbb{R}^n$, and the follower's action is $\mathbf{y} \in \mathbb{R}^m$. The leader aims to minimize the cost function $f(\mathbf{x}, \mathbf{y})$ while the follower aims at maximizing $f(\mathbf{x}, \mathbf{y})$. The only assumption we make on the cost function is the following.

Assumption 1. *f is twice differentiable. $\nabla^2_{\mathbf{yy}} f$ is invertible (i.e., non-singular).*

The global solution to $\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ is an action pair $(\mathbf{x}^*, \mathbf{y}^*)$, such that $\mathbf{y}^*$ is the global optimal response to $\mathbf{x}^*$ for the follower, and that $\mathbf{x}^*$ is the global optimal action for the leader assuming the follower always play the global optimal response. However, finding this global solution is often intractable especially when both players are parameterized by deep neural networks; therefore, we follow [11] and take *local minimax* as the local surrogate.

Definition 1 (local minimax)**.** $(\mathbf{x}^*, \mathbf{y}^*)$ is a local minimax for $f(\mathbf{x}, \mathbf{y})$ if (1) $\mathbf{y}^*$ is a local maximum of $f(\mathbf{x}^*, \cdot)$; (2) $\mathbf{x}^*$ is a local minimum of $\phi(\mathbf{x}) := f(\mathbf{x}, r(\mathbf{x}))$, where $r(\mathbf{x})$ is the implicit function defined by $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$ in a neighborhood of $\mathbf{x}^*$ with $r(\mathbf{x}^*) = \mathbf{y}^*$.

The notion of local minimax is truly a local property in that it is (almost) completely characterized by the derivatives of f at $(\mathbf{x}^*, \mathbf{y}^*)$. Specifically, let

$$\nabla f(\mathbf{x}^*, \mathbf{y}^*) = \begin{bmatrix} \nabla_{\mathbf{x}} f \\ \nabla_{\mathbf{y}} f \end{bmatrix}, \quad \nabla^2 f(\mathbf{x}^*, \mathbf{y}^*) = \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} & \mathbf{H}_{\mathbf{xy}} \\ \mathbf{H}_{\mathbf{yx}} & \mathbf{H}_{\mathbf{yy}} \end{bmatrix}.$$

Then $\nabla f(\mathbf{x}^*, \mathbf{y}^*) = 0$, $\mathbf{H}_{\mathbf{yy}} \preccurlyeq 0$ and $\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} \succcurlyeq 0$ is a **necessary** condition for being a local minimax. Meanwhile, $\nabla f(\mathbf{x}^*, \mathbf{y}^*) = 0$, $\mathbf{H}_{\mathbf{yy}} \prec 0$, and $\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} \succ 0$ is a **sufficient** condition for being a local minimax.

Local minimax is a necessary condition for being a *local Nash* [20], which is the proper local solution concept for *simultaneous* games. However, the stable limit points of GDA, roughly speaking the points GDA locally converges to, are a different superset of local Nash [11]. The relation between the three sets of points is illustrated in Fig. 1.

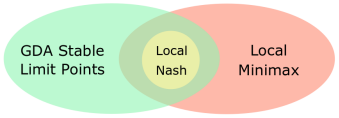

Figure 1: Relation between local Nash, local minimax and gradient descent-ascent.

3 Follow the Ridge

We propose a novel algorithm, which we call *Follow-the-Ridge* (FR), for minimax optimization. The algorithm modifies gradient descent-ascent by applying an asymmetric preconditioner. The update rule is described below.

Algorithm 1: Follow-the-Ridge (FR). Differences from gradient descent-ascent are shown in blue.

for $t = 1, ..., T$ **do**
- $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$ ▷ gradient descent
- $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t + \eta_{\mathbf{y}} \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) + \eta_{\mathbf{x}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)$ ▷ modified gradient ascent

The main intuition behind FR is the following. Suppose that $\mathbf{y}_t$ is a local minimum of $f(\mathbf{x}_t, \cdot)$. Let $r(\mathbf{x})$ be the implicit function defined by $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$ around $(\mathbf{x}_t, \mathbf{y}_t)$, *i.e.*, a ridge. By definition, a local minimax always lie on a ridge; hence, it is intuitive to follow the ridge's direction during learning. However, because $\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) = 0$, one step of gradient descent-ascent will take $(\mathbf{x}_t, \mathbf{y}_t)$

to $(\mathbf{x}_t - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} f, \mathbf{y}_t)$, which is off the ridge. In other words, gradient descent-ascent tends to drift away from the ridge. The correction term we introduce is

$$\nabla_{\mathbf{x}} r(\mathbf{x}) \left(-\eta_{\mathbf{x}} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right) = \eta_{\mathbf{x}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} \nabla_{\mathbf{x}} f.$$

It would bring $\mathbf{y}_t$ to $\mathbf{y}_t + \nabla_{\mathbf{x}} r(\mathbf{x})(\mathbf{x}_{t+1} - \mathbf{x}_t) \approx r(\mathbf{x}_{t+1})$, thereby encouraging both players to stay along the ridge. When $(\mathbf{x}_t, \mathbf{y}_t)$ is not on a ridge yet, we expect the $-\eta_{\mathbf{x}} \nabla_{\mathbf{x}} f$ term and the $\eta_{\mathbf{x}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} \nabla_{\mathbf{x}} f$ term to move parallel to the ridge, while the $\eta_{\mathbf{y}} \nabla_{\mathbf{y}} f$ term brings $(\mathbf{x}_t, \mathbf{y}_t)$ closer to the ridge. Our main theoretical result is the following theorem.

Theorem 1 (Exact local convergence)**.** *The Jacobian of FR has only real eigenvalues at fixed points. With a suitable learning rate, all strictly stable fixed points of FR are local minimax, and all local minimax are stable fixed points of FR.*

The proof is mainly based on the following observation. Let $c = \eta_{\mathbf{y}} / \eta_{\mathbf{x}}$. Then at a fixed point $(\mathbf{x}^*, \mathbf{y}^*)$, the Jacobian of FR is given by

$$\mathbf{J} = \mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{I} \\ -\mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} & \mathbf{H}_{\mathbf{xy}} \\ -c\mathbf{H}_{\mathbf{yx}} & -c\mathbf{H}_{\mathbf{yy}} \end{bmatrix},$$

which is similar to

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} \\ \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \mathbf{J} \begin{bmatrix} \mathbf{I} \\ -\mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} = \mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} & \mathbf{H}_{\mathbf{xy}} \\ & -c\mathbf{H}_{\mathbf{yy}} \end{bmatrix}.$$

Therefore, the eigenvalues of $\mathbf{J}$ are those of $\mathbf{I} + \eta_{\mathbf{y}} \mathbf{H}_{\mathbf{yy}}$ and those of $\mathbf{I} - \eta_{\mathbf{x}} (\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}})$, which are all real. Moreover, when $(\mathbf{x}^*, \mathbf{y}^*)$ is a local minimax, one can show that the spectral radius of the Jacobian satisfies $\rho(\mathbf{J}) \leq 1$, *i.e.*, $(\mathbf{x}^*, \mathbf{y}^*)$ is a stable limit point. On the other hand, when $\rho(\mathbf{J}) < 1$, $(\mathbf{x}^*, \mathbf{y}^*)$ must be a local minimax.

4 Extending the Algorithm

We now discuss several extension of FR that preserves the theoretical guarantees.

Preconditioning: To speed up the convergence, it is often desirable to apply a preconditioner on the gradients that compensates for the curvature. For FR, the preconditioned variant is given by

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \begin{bmatrix} \mathbf{I} \\ -\mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \eta_{\mathbf{x}} \mathbf{P}_1 \nabla_{\mathbf{x}} f \\ -\eta_{\mathbf{y}} \mathbf{P}_2 \nabla_{\mathbf{y}} f \end{bmatrix} \tag{1}$$

We can show that with *any* constant positive definite preconditioners $\mathbf{P}_1$ and $\mathbf{P}_2$, the local convergence behavior of Algorithm 1 remains exact.

Momentum: Another important technique in optimization is momentum, which speeds up convergence significantly both in theory and in practice [19, 21]. We show that momentum can be incorporated into FR, which gives the following update rule:

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \begin{bmatrix} \mathbf{I} \\ -\mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \eta_{\mathbf{x}} \nabla_{\mathbf{x}} f \\ -\eta_{\mathbf{y}} \nabla_{\mathbf{y}} f \end{bmatrix} + \gamma \begin{bmatrix} \mathbf{x}_t - \mathbf{x}_{t-1} \\ \mathbf{y}_t - \mathbf{y}_{t-1} \end{bmatrix}. \tag{2}$$

Because all of the Jacobian eigenvalues are real, we can show that momentum speeds up local convergence in a similar way it speeds up single objective minimization.

Theorem 2. *For local minimax $(\mathbf{x}^*, \mathbf{y}^*)$, let $\alpha = \min\left\{\lambda_{min}(-\mathbf{H}_{\mathbf{yy}}), \lambda_{min}(\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}} \mathbf{H}_{\mathbf{yy}}^{-1} \mathbf{H}_{\mathbf{yx}})\right\}$, $\beta = \rho\left(\nabla^2 f(\mathbf{x}^*, \mathbf{y}^*)\right)$, $\kappa := \beta / \alpha$. Then FR converges asymptotically to $(\mathbf{x}^*, \mathbf{y}^*)$ with a rate $\Omega(\kappa^{-2})$; FR with a momentum parameter of $\gamma = 1 - \Theta\left(\kappa^{-1}\right)$ converges asymptotically with a rate $\Omega(\kappa^{-1})$.*[2]

This is in contrast to gradient descent-ascent, whose complex Jacobian eigenvalues prevent the use of positive momentum. Instead, negative momentum may be more preferable [8], which does not achieve the same level of acceleration.

Non-zero-sum games: In non-zero-sum sequential games, the notion of equilibrium is captured by *Stackelberg equilibrium*, a generalization of minimax. Similarly, local Stackelberg equilibrium can be defined as an extension of local minimax [7]. Applications of finding Stackelberg equilibrium include hyperparameter optimization [13]. For non-zero-sum games, we can show that a simple variant of FR converges exactly to local Stackelberg equilibria (see Appendix D.2).

[2] By a rate a, we mean that one iteration shortens the distance toward the fixed point by a factor of $(1 - a)$; hence the larger the better.

5 Experiments

Our experiments had two main aims: (1) to test if FR is able to converge and only converge to local minimax, and (2) to test the effectiveness of FR in standard machine learning applications.

5.1 Quadratic Problems

We compare FR with gradient descent-ascent (GDA), optimistic mirror descent (OGDA) [4], extragradient (EG) [12], symplectic gradient adjustment (SGA) [2] and consensus optimization (CO) [16] on two simple quadratic problems:

$$g_1(x, y) = -4x^2 - y^2 + 5xy$$

$$g_2(\mathbf{x}, \mathbf{y}) = x_1^2 + 2x_1y_1 + \frac{1}{2}y_1^2 - \frac{1}{2}x_2^2 + 2x_2y_2 - y_2^2.$$

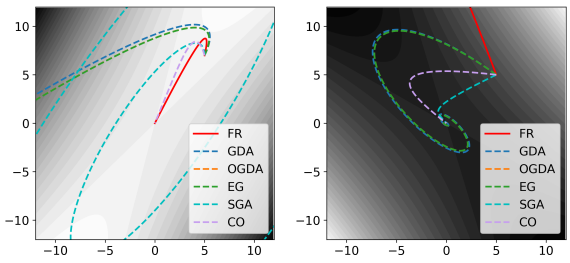

Figure 2: Trajectory of FR and other algorithms in quadratic problems. **Left:** for g_1, $(0, 0)$ is local minimax. **Right:** for g_2, $(0, 0)$ is **NOT** local minimax. The contours are for the function value. For g_2, we plotted the coordinates x_1 and y_1 and the function value on $x_2 = y_2 = 0$.

It can be seen in Fig. 2 that when running in g_1, where $(0, 0)$ is a local minimax, only FR and CO converge to it; all other method diverges. On the other hand, in g_2, where $(0, 0)$ is not a local minimax, all algorithms except for FR converges to this undesirable fixed point. This suggests that even on extremely simple instances, existing algorithms can either fail to converge to a desirable fixed point or converge to bad fixed points, whereas FR always exhibits desirable behavior.

5.2 Generative Adversarial Networks

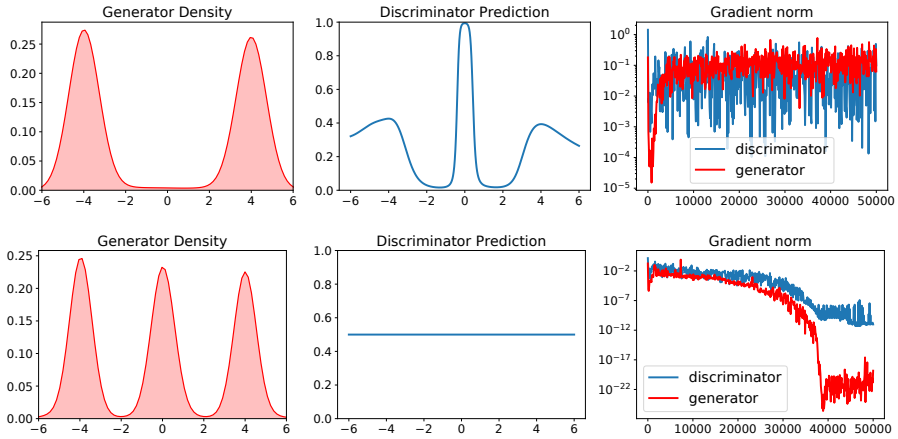

Figure 3: Top: GDA; **Bottom:** FR.

We further compared FR with GDA on GAN training to illustrate the importance the Ridge gradient term. Particularly, we focused on mixtures of Gaussian problem and used the original saturating loss. To satisfy the non-singular Hessian assumption, we add L_2 regularization (0.0002) to the discriminator. For both generator and discriminator, we used 3-layers MLP with 64 hidden units each layer where tanh activations was used. By default, RMSprop [22] was used in all our experiments while the learning rate was tuned for GDA. As our FR involves the computation of Hessian inverse which is computational prohibitive, we instead used conjugate gradient [15, 17] to solve the linear system. To be specific, instead of solving $\mathbf{H_{yy}z} = \mathbf{H_{yx}}\nabla_{\mathbf{x}} f$ directly, we solved $\mathbf{H_{yy}^2 z} = \mathbf{H_{yy}H_{yx}}\nabla_{\mathbf{x}} f$ to ensure that the problem is well-posed since $\mathbf{H_{yy}^2}$ is always positive semidefinite. For all experimental details, we refer readers to Appendix E.2.

As shown in Fig. 3, GDA suffers from the "missing mode" problem and both discriminator and generator fail to converge as confirmed by the gradient norm plot. In contrast, the generator trained with FR successfully learns the true distribution with three modes while the discriminator is totally fooled by the generator. Interestingly, both two players reach much lower gradient norm with FR, indicating convergence.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. *arXiv preprint arXiv:1802.05642*, 2018.

[3] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. *arXiv preprint arXiv:1712.10285*, 2017.

[4] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[5] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9236–9246, 2018.

[6] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058, 2017.

[7] Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

[8] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Gabriel Huang, Rémi Le Priol, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. *CoRR*, abs/1807.04740, 2018.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, 2nd edition, 2013.

[11] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

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

[13] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122, 2015.

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[15] James Martens. Deep learning via hessian-free optimization. In *International Conference on Machine Learning*, pages 735–742, 2010.

[16] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835, 2017.

[17] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[18] Peter J Olver. Nonlinear systems. http://www-users.math.umn.edu/~olver/ln_/nls.pdf, 2015.

[19] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[20] Lillian J. Ratliff, Samuel A. Burden, and S. Shankar Sastry. On the characterization of local nash equilibria in continuous games. *IEEE Transactions on Automatic Control*, 61(8):2301–2307, 2016.

[21] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1139–1147, 2013.

[22] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[23] F. Zeuthen. Heinrich von stackelberg: Marktformen und gleichgewicht. julius springer. 1934 (138 s.). pris r. m. 9,60. *Nationalokonomisk Tidsskrift*, 3, 1935.

A Basic Properties of Local Minimax

First, we would like to mention some basic properties of local minimax for completeness. Most of the results are established in [11].

Proposition 1 (Necessary condition)**.** *Any local minimax* $(\mathbf{x}, \mathbf{y})$ *satisfies* $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = 0$, $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$, $\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \preccurlyeq 0$ *and* $\left[\nabla^2_{\mathbf{xx}} f - \nabla^2_{\mathbf{xy}} f \nabla^2_{\mathbf{yy}} f^{-1} \nabla^2_{\mathbf{yx}} f\right](\mathbf{x}, \mathbf{y}) \succcurlyeq 0$.

Proposition 2 (Sufficient condition)**.** *If* $(\mathbf{x}, \mathbf{y})$ *satisfies* $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = 0$, $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$, $\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \prec 0$ *and* $\left[\nabla^2_{\mathbf{xx}} f - \nabla^2_{\mathbf{xy}} f \nabla^2_{\mathbf{yy}} f^{-1} \nabla^2_{\mathbf{yx}} f\right](\mathbf{x}, \mathbf{y}) \succ 0$*, then* $(\mathbf{x}, \mathbf{y})$ *is a local minimax.*

B Proof of Theorem 1

Proof. First of all, note that FR's update rule can be rewritten as

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & c\mathbf{I} \end{bmatrix} \begin{bmatrix} \nabla_{\mathbf{x}} f \\ -\nabla_{\mathbf{y}} f \end{bmatrix}, \tag{3}$$

where $c := \eta_{\mathbf{y}}/\eta_{\mathbf{x}}$, and that $\begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & c\mathbf{I} \end{bmatrix}$ is always invertible. Therefore, the fixed points of FR is exactly those that satisfy $\nabla f(\mathbf{x}, \mathbf{y}) = 0$, *i.e.*, the first-order necessary condition of local minimax.

Now, consider a fixed point $(\mathbf{x}^*, \mathbf{y}^*)$. The Jacobian of FR's update rule at $(\mathbf{x}^*, \mathbf{y}^*)$ is given by

$$\mathbf{J} = \mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} & \mathbf{H}_{\mathbf{xy}} \\ c\mathbf{H}_{\mathbf{yx}} & c\mathbf{H}_{\mathbf{yy}} \end{bmatrix}.$$

Observe that $\mathbf{J}$ is similar to

$$\begin{aligned} &\begin{bmatrix} \mathbf{I} & \\ \mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \mathbf{J} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \\ =&\mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{I} & \\ \mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} & \mathbf{H}_{\mathbf{xy}} \\ c\mathbf{H}_{\mathbf{yx}} & c\mathbf{H}_{\mathbf{yy}} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \\ =&\mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{H}_{\mathbf{xy}} \\ & -c\mathbf{H}_{\mathbf{yy}} \end{bmatrix}, \end{aligned}$$

which is block diagonal. Therefore, the eigenvalues of J are exactly those of $\mathbf{I} + \eta_{\mathbf{y}}\mathbf{H}_{\mathbf{yy}}$ and those of $\mathbf{I} - \eta_{\mathbf{x}}(\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}})$, which are all real because both matrices are symmetric.

Moreover, suppose that

$$\eta_{\mathbf{x}} < \frac{2}{\max\left\{\rho(\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}}), c\rho(-\mathbf{H}_{\mathbf{yy}})\right\}},$$

where $\rho(\cdot)$ stands for spectral radius. In this case

$$-\mathbf{I} \prec \mathbf{I} + \eta_{\mathbf{y}}\mathbf{H}_{\mathbf{yy}}, \quad -\mathbf{I} \prec \mathbf{I} - \eta_{\mathbf{x}}(\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}}).$$

Therefore whether $\rho(\mathbf{J}) < 1$ depends on whether $-\mathbf{H}_{\mathbf{yy}}$ or $\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}}$ has negative eigenvalues. If $(\mathbf{x}^*, \mathbf{y}^*)$ is a local minimax, by the necessary condition, $\mathbf{H}_{\mathbf{yy}} \preccurlyeq 0$, $\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} \succcurlyeq 0$. It follows that the eigenvalues of $\mathbf{J}$ all fall in $(-1, 1]$. $(\mathbf{x}^*, \mathbf{y}^*)$ is thus a stable limit point of FR.

On the other hand, when $(\mathbf{x}^*, \mathbf{y}^*)$ is a strictly stable limit point, $\rho(\mathbf{J}) < 1$. It follows that both $\mathbf{H}_{\mathbf{yy}}$ and $\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}}$ must be positive definite. By the sufficient conditions of local minimax, $(\mathbf{x}^*, \mathbf{y}^*)$ is a local minimax. □

C Proof of Theorem 2

Consider a general discrete dynamical system $\mathbf{z}_{t+1} \leftarrow g(\mathbf{z}_t)$. Let $\mathbf{z}^*$ be a fixed point of $g(\cdot)$. Let $\mathbf{J}(\mathbf{z})$ denote the Jacobian of $g(\cdot)$ at $\mathbf{z}$. Similar results can be found in many texts; see, for instance, Theorem 2.12 [18].

Proposition 3 (Local convergence from Jacobian eigenvalue)**.** *If $\rho(\mathbf{J}(\mathbf{z}^*)) = 1 - \Delta < 1$, then there exists a neighborhood U of $\mathbf{z}^*$ such that for any $\mathbf{z}_0 \in U$,*

$$\|\mathbf{z}_t - \mathbf{z}^*\|_2 \leq C\left(1 - \frac{\Delta}{2}\right)^t \|\mathbf{z}_0 - \mathbf{z}^*\|_2,$$

where C is some constant.

Proof. By Lemma 5.6.10 [10], since $\rho(\mathbf{J}(\mathbf{z}^*)) = 1 - \Delta$, there exists a matrix norm $\|\cdot\|$ induced by vector norm $\|\cdot\|$ such that $\|\mathbf{J}(\mathbf{z}^*)\| < 1 - \frac{3\Delta}{4}$. Now consider the Taylor expansion of $g(\mathbf{z})$ at the fixed point $\mathbf{z}^*$:

$$g(\mathbf{z}) = g(\mathbf{z}^*) + \mathbf{J}(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*) + R(\mathbf{z} - \mathbf{z}^*),$$

where the remainder term satisfies

$$\lim_{\mathbf{z}\to\mathbf{z}^*} \frac{R(\mathbf{z} - \mathbf{z}^*)}{\|\mathbf{z} - \mathbf{z}^*\|} = 0.$$

Therefore, we can choose $0 < \delta$ such that whenever $\|\mathbf{z} - \mathbf{z}^*\| < \delta$, $\|R(\mathbf{z} - \mathbf{z}^*)\| \leq \frac{\Delta}{4}\|\mathbf{z} - \mathbf{z}^*\|$. In this case,

$$\begin{aligned}\|g(\mathbf{z}) - g(\mathbf{z}^*)\| &\leq \|\mathbf{J}(\mathbf{z}^*)(\mathbf{z} - \mathbf{z}^*)\| + \|R(\mathbf{z} - \mathbf{z}^*)\| \\ &\leq \|\mathbf{J}(\mathbf{z}^*)\|\|\mathbf{z} - \mathbf{z}^*\| + \frac{\Delta}{4}\|\mathbf{z} - \mathbf{z}^*\| \\ &\leq \left(1 - \frac{\Delta}{2}\right)\|\mathbf{z} - \mathbf{z}^*\|.\end{aligned}$$

In other words, when $\mathbf{z}_0 \in U = \{\mathbf{z} |\ \|\mathbf{z} - \mathbf{z}^*\| < \delta\}$,

$$\|\mathbf{z}_t - \mathbf{z}^*\| \leq \left(1 - \frac{\Delta}{2}\right)^t \|\mathbf{z}_0 - \mathbf{z}^*\|.$$

By the equivalence of finite dimensional norms, there exists constants $c_1, c_2 > 0$ such that

$$\forall \mathbf{z}, \quad c_1\|\mathbf{z}\|_2 \leq \|\mathbf{z}\| \leq c_2\|\mathbf{z}\|_2.$$

Therefore

$$\|\mathbf{z}_t - \mathbf{z}^*\|_2 \leq \frac{c_2}{c_1}\left(1 - \frac{\Delta}{2}\right)^t \|\mathbf{z}_0 - \mathbf{z}^*\|_2.$$

□

In other words, the rate of convergence is given by the gap between $\rho(\mathbf{J})$ and 1. We now prove Theorem 2 using this view.

proof of Theorem 2. In the following proof we use $\|\cdot\|$ to denote the standard spectral norm. It is not hard to see that $\lambda_{max}(-\mathbf{H_{yy}}) \leq \rho(\nabla^2 f(\mathbf{x}^*, \mathbf{y}^*)) = \beta$ and $\|\mathbf{H_{xy}}\| \leq \beta$. Also,

$$\lambda_{max}(\mathbf{H_{xx}} - \mathbf{H_{xy}}\mathbf{H_{yy}^{-1}}\mathbf{H_{yx}}) \leq \|\mathbf{H_{xx}}\| + \|\mathbf{H_{xy}}\|^2 \cdot \|\mathbf{H_{yy}^{-1}}\| \leq \beta + \frac{\beta^2}{\alpha} = (1 + \kappa)\beta.$$

Therefore we choose our learning rate to be $\eta_{\mathbf{x}} = \eta_{\mathbf{y}} = \frac{1}{2\kappa\beta}$. In this case, the eigenvalues of the Jacobian of FR without momentum all fall in $\left[0, 1 - \frac{1}{2\kappa^2}\right]$. Using Proposition 3, we can show that FR locally converges with a rate of $\Omega(\kappa^{-2})$.

Now, let us focus on FR with Polyak's momentum:

$$\begin{bmatrix}\mathbf{x}_{t+1} \\ \mathbf{y}_{t+1}\end{bmatrix} \leftarrow \begin{bmatrix}\mathbf{x}_t \\ \mathbf{y}_t\end{bmatrix} - \eta_{\mathbf{x}}\begin{bmatrix}\mathbf{I} & \\ -\mathbf{H_{yy}^{-1}}\mathbf{H_{yx}} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\nabla_{\mathbf{x}} f \\ -\nabla_{\mathbf{y}} f\end{bmatrix} + \gamma\begin{bmatrix}\mathbf{x}_t - \mathbf{x}_{t-1} \\ \mathbf{y}_t - \mathbf{y}_{t-1}\end{bmatrix}. \tag{4}$$

This is a dynamical system on the augmented space of $(\mathbf{x}_t, \mathbf{y}_t, \mathbf{x}_{t-1}, \mathbf{y}_{t-1})$. Let

$$\mathbf{J}_1 := \mathbf{I} - \eta_{\mathbf{x}}\begin{bmatrix}\mathbf{I} & \\ -\mathbf{H_{yy}^{-1}}\mathbf{H_{yx}} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{H_{xx}} & \mathbf{H_{xy}} \\ -\mathbf{H_{yx}} & -\mathbf{H_{yy}}\end{bmatrix}$$

be the Jacobian of the original FR at a fixed point $(\mathbf{x}^*, \mathbf{y}^*)$. Then the Jacobian of Polyak's momentum at $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{x}^*, \mathbf{y}^*)$ is

$$\mathbf{J}_2 := \begin{bmatrix} \gamma\mathbf{I} + \mathbf{J}_1 & -\gamma\mathbf{I} \\ \mathbf{I} & 0 \end{bmatrix}.$$

The spectrum of $\mathbf{J}_2$ is given by solutions to

$$\det\left(\lambda\mathbf{I} - \mathbf{J}_2\right) = \det\left((\lambda^2 - \gamma\lambda + \gamma)\mathbf{I} - \gamma\mathbf{J}_1\right) = 0.$$

In other words, an eigenvalue r of $\mathbf{J}_1$ corresponds to two eigenvalues of $\mathbf{J}_2$ given by the roots of $\lambda^2 - (\gamma + r)\lambda + \gamma = 0$. For our case, let us choose $\gamma = 1 + \frac{1}{2\kappa^2} - \frac{\sqrt{2}}{\kappa}$. Then for any $r \in \left[0, 1 - \frac{1}{2\kappa^2}\right]$,

$$(r+\gamma)^2 - 4\gamma \leq \left(1 - \frac{1}{2\kappa^2} + \gamma\right)^2 - 4\gamma = 0.$$

Therefore the two roots of $\lambda^2 - (\gamma + r)\lambda + \gamma = 0$ must be imaginary, and their magnitude are exactly $\sqrt{\gamma}$. Since $\sqrt{\gamma} \leq 1 - \frac{1-\gamma}{2} \leq 1 - \frac{1}{2\sqrt{2}\kappa}$, we now know that $\rho(\mathbf{J}_2) \leq 1 - \frac{1}{2\sqrt{2}\kappa}$. Using Proposition 3, we can see that FR with momentum locally converge with a rate of $\Omega(\kappa^{-1})$. □

D Proofs for Section 4

D.1 Preconditioning

Recall that the preconditioned variant of FR is given by

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \eta_{\mathbf{x}}\mathbf{P}_1\nabla_{\mathbf{x}} f \\ -\eta_{\mathbf{y}}\mathbf{P}_2\nabla_{\mathbf{y}} f \end{bmatrix}. \tag{5}$$

We now prove that preconditioning does not effect the local convergence properties.

Proposition 4. *If A is a symmetric real matrix, B is symmetric and positive definite, then the eigenvalues of AB are all real, and AB and A have the same number of positive, negative and zero eigenvalues.*

Proof. AB is similar to and thus has the same eigenvalues as $B^{\frac{1}{2}}AB^{\frac{1}{2}}$, which is symmetric and has real eigenvalues. Since $B^{\frac{1}{2}}AB^{\frac{1}{2}}$ is congruent to A, they have the same number of positive, negative and zero eigenvalues (see Theorem 4.5.8 [10]). □

Proposition 5. *Assume that $\mathbf{P}_1$ and $\mathbf{P}_2$ are positive definite. The Jacobian of (5) has only real eigenvalues at fixed points. With a suitable learning rate, all strictly stable fixed points of (5) are local minimax, and all local minimax are stable fixed points of (5).*

Proof. First, observe that both $\begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{P}_1 & \\ & \mathbf{P}_2 \end{bmatrix}$ are both always invertible. Hence fixed points of (5) are exactly stationary points. Let $c := \eta_{\mathbf{y}}/\eta_{\mathbf{x}}$. Note that the Jacobian of (5) is given by

$$\mathbf{J} = \mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{P}_1 & \\ & \mathbf{P}_2 \end{bmatrix} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} & \mathbf{H}_{\mathbf{xy}} \\ c\mathbf{H}_{\mathbf{yx}} & c\mathbf{H}_{\mathbf{yy}} \end{bmatrix},$$

which is similar to

$$\begin{aligned} &\begin{bmatrix} \mathbf{I} & \\ \mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \mathbf{J} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \\ =&\mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{P}_1 & \\ & \mathbf{P}_2 \end{bmatrix} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}} & \mathbf{H}_{\mathbf{xy}} \\ & -c\mathbf{H}_{\mathbf{yy}} \end{bmatrix}. \end{aligned}$$

Therefore the eigenvalues of $\mathbf{J}$ are exactly those of $\mathbf{I} - \eta_{\mathbf{x}}\mathbf{P}_1\left(\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}}\right)$ and $\mathbf{I} + \eta_{\mathbf{y}}\mathbf{P}_2\mathbf{H}_{\mathbf{yy}}$. By Proposition 4, the eigenvalues of both matrices are all real. When the learning rates are small enough, *i.e.*, when

$$\eta_{\mathbf{x}} < \frac{2}{\max\left\{\rho\left(\mathbf{P}_1(\mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{H}_{\mathbf{yy}}^{-1}\mathbf{H}_{\mathbf{yx}})\right), c\rho(-\mathbf{P}_2\mathbf{H}_{\mathbf{yy}})\right\}},$$

whether $\rho(\mathbf{J}) \leq 1$ solely depends on whether $\mathbf{P}_1\left(\mathbf{H_{xx}} - \mathbf{H_{xy}}\mathbf{H_{yy}^{-1}}\mathbf{H_{yx}}\right)$ and $-\mathbf{P}_2\mathbf{H_{yy}}$ have negative eigenvalues. By Proposition 4, the number of positive, negative and zero eigenvalues of the two matrices are the same as those of $\mathbf{H_{xx}} - \mathbf{H_{xy}}\mathbf{H_{yy}^{-1}}\mathbf{H_{yx}}$ and $-\mathbf{H_{yy}}$ respectively. Therefore the proposition follows from the same argument as in Theorem 1. □

D.2 Non-zero-sum Stackelberg Games

A non-zero-sum Stackelberg games is formulated as follows. There is a leader, whose action is $\mathbf{x} \in \mathbb{R}^n$, and a follower, whose action is $\mathbf{y} \in \mathbb{R}^m$. The leader's cost function is given by $f(\mathbf{x}, \mathbf{y})$ while the follower's is given by $g(\mathbf{x}, \mathbf{y})$. The generalization of minimax in non-zero-sum Stackelberg games is *Stackelberg equilibrium*.

Definition 2 (Stackelberg equilibrium)**.** $(\mathbf{x}^*, \mathbf{y}^*)$ is a (global) Stackelberg equilibrium if $\mathbf{y}^* \in R(\mathbf{x}^*)$, and $\forall \mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y} \in R(\mathbf{x})} g(\mathbf{x}, \mathbf{y}),$$

where $R(\mathbf{x}) := \arg\min g(\mathbf{x}, \cdot)$ is the best response set for the follower.

Similarly, local minimax is generalized to local Stackelberg equilibrium, defined as follows.

Definition 3. $(\mathbf{x}^*, \mathbf{y}^*)$ is a local Stackelberg equilibrium if

1. $\mathbf{y}^*$ is a local minimum of $g(\mathbf{x}^*, \cdot)$;
2. Let $r(\mathbf{x})$ be the implicit function defined by $\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}) = 0$ in a neighborhood of $\mathbf{x}^*$ with $r(\mathbf{x}^*) = \mathbf{y}^*$. Then $\mathbf{x}^*$ is a local minimum of $\phi(\mathbf{x}) := f(\mathbf{x}, r(\mathbf{x}))$.

For local Stackelberg equilibrium, we have similar necessary conditions and sufficient conditions. For simplicity, we use the following notation when it is clear from the context

$$\nabla^2 f(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \mathbf{H_{xx}} & \mathbf{H_{xy}} \\ \mathbf{H_{yx}} & \mathbf{H_{yy}} \end{bmatrix}, \ \nabla^2 g(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \mathbf{G_{xx}} & \mathbf{G_{xy}} \\ \mathbf{G_{yx}} & \mathbf{G_{yy}} \end{bmatrix}.$$

Proposition 6 (Necessary conditions)**.** *Any local Stackelberg equilibrium satisfies* $\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}) = 0$*,* $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \mathbf{G_{xy}}\mathbf{G_{yy}^{-1}}\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$*,* $\nabla^2_{\mathbf{yy}} g(\mathbf{x}, \mathbf{y}) \succcurlyeq 0$ *and*

$$\mathbf{H_{xx}} - \mathbf{H_{xy}}\mathbf{G_{yy}^{-1}}\mathbf{G_{yx}} - \nabla_{\mathbf{x}}\left(\mathbf{G_{xy}}\mathbf{G_{yy}^{-1}}\nabla_{\mathbf{y}} f\right) + \nabla_{\mathbf{y}}\left(\mathbf{G_{xy}}\mathbf{G_{yy}^{-1}}\nabla_{\mathbf{y}} f\right)\mathbf{G_{yy}^{-1}}\mathbf{G_{yx}} \succcurlyeq 0.$$

Proposition 7 (Sufficient conditions)**.** *If* $(\mathbf{x}, \mathbf{y})$ *satisfy* $\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}) = 0$*,* $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) - \mathbf{G_{xy}}\mathbf{G_{yy}^{-1}}\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = 0$*,* $\nabla^2_{\mathbf{yy}} g(\mathbf{x}, \mathbf{y}) \succ 0$ *and*

$$\mathbf{H_{xx}} - \mathbf{H_{xy}}\mathbf{G_{yy}^{-1}}\mathbf{G_{yx}} - \nabla_{\mathbf{x}}\left(\mathbf{G_{xy}}\mathbf{G_{yy}^{-1}}\nabla_{\mathbf{y}} f\right) + \nabla_{\mathbf{y}}\left(\mathbf{G_{xy}}\mathbf{G_{yy}^{-1}}\nabla_{\mathbf{y}} f\right)\mathbf{G_{yy}^{-1}}\mathbf{G_{yx}} \succ 0.$$

then $(\mathbf{x}, \mathbf{y})$ *is a local Stackelberg equilibrium.*

Henceforth we will use $D_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ to denote $\nabla_{\mathbf{x}} f - \mathbf{G_{xy}}\mathbf{G_{yy}^{-1}}\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. The non-zero-sum version of Follow-the-Ridge is given by

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \begin{bmatrix} \mathbf{I} & \\ -\mathbf{G_{yy}^{-1}}\mathbf{G_{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \eta_{\mathbf{x}} D_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) \\ \eta_{\mathbf{y}} \nabla_{\mathbf{y}} g(\mathbf{x}_t, \mathbf{y}_t) \end{bmatrix}. \tag{6}$$

Just as the zero-sum version of FR converges exactly to local minimax, we can show that the non-zero-sum version of FR converges exactly to local Stackelberg equilibria.

Theorem 3. *The Jacobian of (6) has only real eigenvalues at fixed points. With a suitable learning rate, all strictly stable fixed points of (6) are local Stackelberg equilibria, and all local Stackelberg equilibria are stable fixed points of (6).*

Proof. Let $c := \eta_{\mathbf{y}}/\eta_{\mathbf{x}}$. Note that $\begin{bmatrix} \mathbf{I} & \\ -\mathbf{G_{yy}^{-1}}\mathbf{G_{yx}} & \mathbf{I} \end{bmatrix}$ is always invertible. Therefore, the fixed points of (6) are exactly those that satisfy $D_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = 0$ and $\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}) = 0$, i.e. the first-order necessary condition for local Stackelberg equilibria.

Now, consider a fixed point $(\mathbf{x}, \mathbf{y})$. The Jacobian of (6) at $(\mathbf{x}, \mathbf{y})$ is given by

$$\mathbf{J} = \mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} - \nabla_{\mathbf{x}}(\mathbf{G}_{\mathbf{xy}}\mathbf{G}_{\mathbf{yy}}^{-1}\nabla_{\mathbf{y}} f) & \mathbf{H}_{\mathbf{xy}} - \nabla_{\mathbf{y}}(\mathbf{G}_{\mathbf{xy}}\mathbf{G}_{\mathbf{yy}}^{-1}\nabla_{\mathbf{y}} f) \\ c\mathbf{G}_{\mathbf{yx}} & c\mathbf{G}_{\mathbf{yy}} \end{bmatrix}.$$

Observe that

$$\begin{aligned}
&\begin{bmatrix} \mathbf{I} & \\ \mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \mathbf{J} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \\
=&\mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} - \nabla_{\mathbf{x}}(\mathbf{G}_{\mathbf{xy}}\mathbf{G}_{\mathbf{yy}}^{-1}\nabla_{\mathbf{y}} f) & \mathbf{H}_{\mathbf{xy}} - \nabla_{\mathbf{y}}(\mathbf{G}_{\mathbf{xy}}\mathbf{G}_{\mathbf{yy}}^{-1}\nabla_{\mathbf{y}} f) \\ c\mathbf{G}_{\mathbf{yx}} & c\mathbf{G}_{\mathbf{yy}} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \\ -\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \\
=&\mathbf{I} - \eta_{\mathbf{x}} \begin{bmatrix} \mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}} - \nabla_{\mathbf{x}}(\square) + \nabla_{\mathbf{y}}(\square)\,\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}} & \mathbf{H}_{\mathbf{xy}} - \nabla_{\mathbf{y}}(\square) \\ 0 & c\mathbf{G}_{\mathbf{yy}} \end{bmatrix},
\end{aligned}$$

where $\square$ is a shorthand for $\mathbf{G}_{\mathbf{xy}}\mathbf{G}_{\mathbf{yy}}^{-1}\nabla_{\mathbf{y}} f$. Let

$$\widetilde{\mathbf{H}}_{\mathbf{xx}} := \mathbf{H}_{\mathbf{xx}} - \mathbf{H}_{\mathbf{xy}}\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}} - \nabla_{\mathbf{x}}(\square) + \nabla_{\mathbf{y}}(\square)\,\mathbf{G}_{\mathbf{yy}}^{-1}\mathbf{G}_{\mathbf{yx}}.$$

We can now see that the eigenvalues of $\mathbf{J}$ are exactly those of $\mathbf{I} - \eta_{\mathbf{x}}\widetilde{\mathbf{H}}_{\mathbf{xx}}$ and those of $\mathbf{I} - \eta_{\mathbf{y}}\mathbf{G}_{\mathbf{yy}}$. It follows that all eigenvalues of J are real.[3] Suppose that

$$\eta_{\mathbf{x}} < \frac{2}{\max\{\rho(\widetilde{\mathbf{H}}_{\mathbf{xx}}), c\rho(\mathbf{G}_{\mathbf{yy}})\}}.$$

In that case, if $(\mathbf{x}, \mathbf{y})$ is a local Stackelberg equilibrium, then from the second-order necessary condition, both $\widetilde{\mathbf{H}}_{\mathbf{xx}}$ and $\mathbf{G}_{\mathbf{yy}}$ are positive semidefinite. As a result, all eigenvalues of $\mathbf{J}$ would be in $(-1, 1]$. This suggests that $(\mathbf{x}, \mathbf{y})$ is a stable fixed point.

On the other hand, if $(\mathbf{x}, \mathbf{y})$ is a strictly stable fixed point, then all eigenvalues of $\mathbf{J}$ fall in $(-1, 1)$, which suggests that $\widetilde{\mathbf{H}}_{\mathbf{xx}} \succ 0$ and $\mathbf{G}_{\mathbf{yy}} \succ 0$. By the sufficient condition, $(\mathbf{x}, \mathbf{y})$ is a local Stackelberg equilibrium. □

E Experimental Details

E.1 Quadratic Problems

The algorithms we compared with are

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) \end{bmatrix}, \tag{GDA}$$

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - 2\eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) \end{bmatrix} + \eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \end{bmatrix}, \tag{OGDA}$$

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_t - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}_t - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)) \end{bmatrix}, \tag{EG}$$

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \eta \begin{bmatrix} \mathbf{I} & -\lambda\mathbf{H}_{\mathbf{xy}} \\ \lambda\mathbf{H}_{\mathbf{yx}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) \end{bmatrix}, \tag{SGA}$$

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{y}_{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) \end{bmatrix} - \gamma\eta\nabla \left\| \nabla f(\mathbf{x}_t, \mathbf{y}_t) \right\|^2. \tag{CO}$$

We used a learning rate of $\eta = 0.01$ for all algorithms, $\lambda = 0.1$ for SGA and $\gamma = 0.1$ for CO.

E.2 GAN Model with Mixture of Gaussian

Dataset. The Mixture of Gaussian dataset is composed of 5,000 points sampled independently from the following distribution $p_{\mathcal{D}}(x) = \frac{1}{3}\mathcal{N}(-4, 0.01) + \frac{1}{3}\mathcal{N}(0, 0.01) + \frac{1}{3}\mathcal{N}(4, 0.01)$ where $\mathcal{N}(\mu, \sigma^2)$ is the probability density function of a 1D-Gaussian distribution with mean μ and variance σ^2. The

[3] $\widetilde{\mathbf{H}}_{\mathbf{xx}}$ is always symmetric.

latent variables $\mathbf{z} \in \mathbb{R}^4$ are sampled from a standard Normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Because we want to use full-batch methods, we sample 5,000 points that we re-use for each iteration during training.

Neural Networks Architecture. Both the generator and discriminator are 2 hidden layer neural networks with 64 hidden units and Tanh activations.

Other Hyperparameters. For FR, we used conjugate gradient in the inner-loop to approximately invert the Hessian. In practice, we used 5 (10 and 20 also works well) CG iterations. Since the loss surface is highly non-convex (let alone quadratic), we added damping term to stabilize the training. Specifically, we followed Levenberg-Marquardt style heuristic adopted in [15]. For both generator and discriminator, we used learning rate 0.0002.