Journal of Machine Learning Research 20 (2019) 1-40 Submitted 1/19; Published 2/19

Differentiable Game Mechanics

Alistair Letcher* AHP.LETCHER@GMAIL.COM
University of Oxford

David Balduzzi* DBALDUZZI@GOOGLE.COM
DeepMind

Sébastien Racanière SRACANIERE@GOOGLE.COM
DeepMind

James Martens JAMESMARTENS@GOOGLE.COM
DeepMind

Jakob Foerster JAKOBFOERSTER@GMAIL.COM
University of Oxford

Karl Tuyls KARLTUYLS@GOOGLE.COM
DeepMind

Thore Graepel THORE@GOOGLE.COM
DeepMind

Editor: Kilian Weinberger

Abstract

Deep learning is built on the foundational guarantee that gradient descent on an objective function converges to local minima. Unfortunately, this guarantee fails in settings, such as generative adversarial nets, that exhibit multiple interacting losses. The behavior of gradient-based methods in games is not well understood – and is becoming increasingly important as adversarial and multi-objective architectures proliferate. In this paper, we develop new tools to understand and control the dynamics in n-player differentiable games.

The key result is to decompose the game Jacobian into two components. The first, symmetric component, is related to potential games, which reduce to gradient descent on an implicit function. The second, antisymmetric component, relates to *Hamiltonian games*, a new class of games that obey a conservation law akin to conservation laws in classical mechanical systems. The decomposition motivates *Symplectic Gradient Adjustment* (SGA), a new algorithm for finding stable fixed points in differentiable games. Basic experiments show SGA is competitive with recently proposed algorithms for finding stable fixed points in GANs – while at the same time being applicable to, and having guarantees in, much more general cases.

Keywords: game theory, generative adversarial networks, deep learning, classical mechanics, hamiltonian mechanics, gradient descent, dynamical systems

1. Introduction

A significant fraction of recent progress in machine learning has been based on applying gradient descent to optimize the parameters of neural networks with respect to an objective function. The objective functions are carefully designed to encode particular tasks such as

supervised learning. A basic result is that gradient descent converges to a local minimum of the objective function under a broad range of conditions (Lee et al., 2017). However, there is a growing set of algorithms that do not optimize a single objective function, including: generative adversarial networks (Goodfellow et al., 2014; Zhu et al., 2017), proximal gradient TD learning (Liu et al., 2016), multi-level optimization (Pfau and Vinyals, 2016), synthetic gradients (Jaderberg et al., 2017), hierarchical reinforcement learning (Wayne and Abbott, 2014; Vezhnevets et al., 2017), intrinsic curiosity (Pathak et al., 2017; Burda et al., 2019), and imaginative agents (Racanière et al., 2017). In effect, the models are trained via games played by cooperating and competing modules.

The time-average of iterates of gradient descent, and other more general no-regret algorithms, are guaranteed to converge to coarse correlated equilibria in games (Stoltz and Lugosi, 2007). However, the dynamics do not converge to Nash equilibria – and do not even stabilize in general (Mertikopoulos et al., 2018; Papadimitriou and Piliouras, 2018). Concretely, cyclic behaviors emerge even in simple cases, see example 1.

This paper presents an analysis of the second-order structure of game dynamics that allows to identify two classes of games, potential and Hamiltonian, that are easy to solve separately. We then derive *symplectic gradient adjustment*[1] (SGA), a method for finding stable fixed points in games. SGA's performance is evaluated in basic experiments.

1.1. Background and Problem Description

Tractable algorithms that converge to Nash equilibria have been found for restricted classes of games: potential games, two-player zero-sum games, and a few others (Hart and Mas-Colell, 2013). Finding Nash equilibria can be reformulated as a nonlinear complementarity problem, but these are 'hopelessly impractical to solve' in general (Shoham and Leyton-Brown, 2008) because the problem is PPAD hard (Daskalakis et al., 2009).

Players are primarily neural nets in our setting. For computational reasons we restrict to gradient-based methods, even though game-theorists have considered a much broader range of techniques. Losses are not necessarily convex in *any* of their parameters, so Nash equilibria are not guaranteed to exist. Even leaving existence aside, finding Nash equilibria in nonconvex games is analogous to, but much harder than, finding global minima in neural nets – which is not realistic with gradient-based methods.

There are at least three problems with gradient-based methods in games. Firstly, the potential existence of cycles (recurrent dynamics) implies there are no convergence guarantees, see example 1 below and Mertikopoulos et al. (2018). Secondly, even when gradient descent converges, the rate of convergence may be too slow in practice because 'rotational forces' necessitate extremely small learning rates, see Figure 4. Finally, since there is no single objective, there is no way to measure progress. Concretely, the losses obtained by the generator and the discriminator in GANs are not useful guides to the quality of the images generated. Application-specific proxies have been proposed, for example the inception score for GANs (Salimans et al., 2016), but these are of little help during training. The inception score is domain specific and is no substitute for looking at samples. This paper tackles the first two problems.

1. Source code is available at `https://github.com/deepmind/symplectic-gradient-adjustment`.

1.2. Outline and Summary of Main Contributions

1.2.1. The Infinitesimal Structure of Games

We start with the basic case of a zero-sum bimatrix game: example 1. It turns out that the dynamics under simultaneous gradient descent can be reformulated in terms of Hamilton's equations. The cyclic behavior arises because the dynamics live on the level sets of the Hamiltonian. More directly useful, gradient descent on the Hamiltonian converges to a Nash equilibrium.

Lemma 1 shows that the Jacobian of any game decomposes into symmetric and antisymmetric components. There are thus two 'pure' cases corresponding to when the Jacobian is symmetric and anti-symmetric. The first case, known as potential games (Monderer and Shapley, 1996), have been intensively studied in the game-theory literature because they are exactly the games where gradient descent *does* converge.

The second case, Hamiltonian[2] games, were not studied previously, probably because they coincide with zero-sum games in the bimatrix case (or constant-sum, depending on the constraints). Zero-sum and Hamiltonian games differ when the losses are not bilinear or when there are more than two players. Hamiltonian games are important because (i) they are easy to solve and (ii) general games combine potential-like and Hamiltonian-like dynamics. Unfortunately, the concept of a zero-sum game is too loose to be useful when there are many players: any n-player game can be reformulated as a zero-sum $(n+1)$-player game where $\ell_{n+1} = -\sum_{i=1}^{n} \ell_i$. In this respect, zero-sum games are as complicated as general-sum games. In contrast, Hamiltonian games are much simpler than general-sum games. Theorem 4 shows that Hamiltonian games obey a conservation law – which also provides the key to solving them, by gradient descent on the conserved quantity.

1.2.2. Algorithms

The general case, neither potential nor Hamiltonian, is more difficult and is therefore the focus of the remainder of the paper. Section 3 proposes *symplectic gradient adjustment (SGA)*, a gradient-based method for finding stable fixed points in general games. Appendix A contains TensorFlow code to compute the adjustment. The algorithm computes two Jacobian-vector products, at a cost of two iterations of backprop. SGA satisfies a few natural desiderata explained in Section 3.1: $(D1)$ it is compatible with the original dynamics; and it is guaranteed to find stable equilibria in $(D2)$ potential and $(D3)$ Hamiltonian games.

For general games, correctly picking the *sign* of the adjustment (whether to add or subtract) is critical since it determines the behavior near stable and unstable equilibria. Section 2.4 defines stable equilibria and contrasts them with local Nash equilibria. Theorem 10 proves that SGA converges locally to stable fixed points for sufficiently small parameters (which we quantify via the notion of an additive condition number). While strong, this may be impractical or slow down convergence significantly. Accordingly, Lemma 11 shows how to set the sign so as to be attracted towards stable equilibria and repelled from unstable ones. Correctly aligning SGA allows higher learning rates and faster, more robust convergence, see Theorem 15. Finally, Theorem 17 tackles the remaining class of saddle fixed points by proving that SGA locally avoids strict saddles for appropriate parameters.

2. Lu (1992) defined an unrelated notion of Hamiltonian game.

1.2.3. Experiments

We investigate the empirical performance of SGA in four basic experiments. The first experiment shows how increasing alignment allows higher learning rates and faster convergence, Figure 4. The second set of experiments compares SGA with optimistic mirror descent on two-player and four-player games. We find that SGA converges over a much wider range of learning rates.

The last two sets of experiments investigate mode collapse, mode hopping and the related, less well-known problem of boundary distortion identified in Santurkar et al. (2018). Mode collapse and mode hopping are investigated in a setup involving a two-dimensional mixture of 16 Gaussians that is somewhat more challenging than the original problem introduced in Metz et al. (2017). Whereas simultaneous gradient descent completely fails, our symplectic adjustment leads to rapid convergence – slightly improved by correctly choosing the sign of the adjustment.

Finally, boundary distortion is studied using a 75-dimensional spherical Gaussian. Mode collapse is not an issue since there the data distribution is unimodal. However, as shown in Figure 10, a vanilla GAN with RMSProp learns only one of the eigenvalues in the spectrum of the covariance matrix, whereas SGA approximately learns all of them.

The appendix provides some background information on differential and symplectic geometry, which motivated the developments in the paper. The appendix also explores what happens when the analogy with classical mechanics is pushed further than perhaps seems reasonable. We experiment with assigning units (in the sense of masses and velocities) to quantities in games, and find that type-consistency yields unexpected benefits.

1.3. Related Work

Nash (1950) was only concerned with existence of equilibria. Convergence in two-player games was studied in Singh et al. (2000). WoLF (Win or Learn Fast) converges to Nash equilibria in two-player two-action games (Bowling and Veloso, 2002). Extensions include weighted policy learning (Abdallah and Lesser, 2008) and GIGA-WoLF (Bowling, 2004). Infinitesimal Gradient Ascent (IGA) is a gradient-based approach that is shown to converge to pure Nash equilibria in two-player two-action games. Cyclic behaviour may occur in case of mixed equilibria. Zinkevich (2003) generalised the algorithm to n-action games called GIGA. Optimistic mirror descent approximately converges in two-player bilinear zero-sum games (Daskalakis et al., 2018), a special case of Hamiltonian games. In more general settings it converges to coarse correlated equilibria.

Convergence has also been studied in various n-player settings, see Rosen (1965); Scutari et al. (2010); Facchinei and Kanzow (2010); Mertikopoulos and Zhou (2016). However, the recent success of GANs, where the players are neural networks, has focused attention on a much larger class of *nonconvex* games where comparatively little is known, especially in the n-player case. Heusel et al. (2017) propose a two-time scale methods to find Nash equilibria. However, it likely scales badly with the number of players. Nagarajan and Kolter (2017) prove convergence for some algorithms, but under very strong assumptions (Mescheder et al., 2018). Consensus optimization (Mescheder et al., 2017) is closely related to our proposad algorithm, and is extensively discussed in Section 3. A variety of game-theoretically or

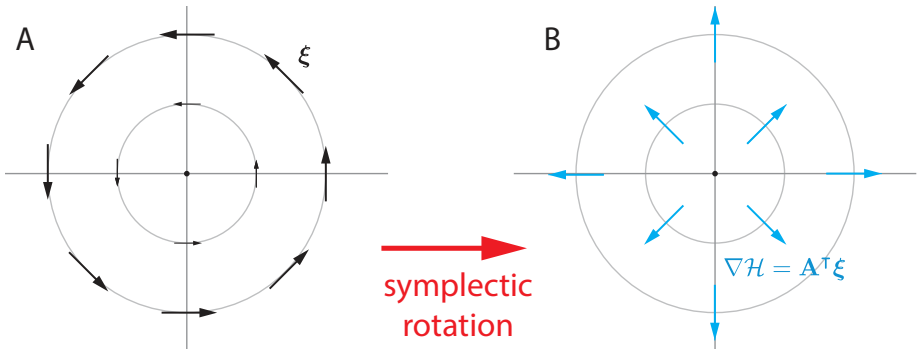

Figure 1: *A minimal example of Hamiltonian mechanics.* Consider a game where $\ell_1(x, y) = xy$, $\ell_2(x, y) = -xy$, and the dynamics are given by $\boldsymbol{\xi}(x, y) = (y, -x)$. The game is a special case of example 1. **(A)** The dynamics $\boldsymbol{\xi}$ cycle around the origin since they live on the level sets of the Hamiltonian $\mathcal{H}(x, y) = \frac{1}{2}(x^2 + y^2)$. **(B)** Gradient descent on the Hamiltonian $\mathcal{H}$ converges to the Nash equilibrium of the game, at the origin $(0, 0)$. Note that $\mathbf{A}^\intercal\boldsymbol{\xi} = (x, y) = \nabla\mathcal{H}$.

minimax motivated modifications to vanilla gradient descent have been investigated in the context of GANs, see Mertikopoulos et al. (2019); Gidel et al. (2018).

Learning with opponent-learning awareness (LOLA) infinitesimally modifies the objectives of players to take into account their opponents' goals (Foerster et al., 2018). However, Letcher et al. (2019) recently showed that LOLA modifies fixed points and thus fails to find stable equilibria in general games.

Symplectic gradient adjustment was independently discovered by Gemp and Mahadevan (2018), who refer to it as "crossing-the-curl". Their analysis draws on powerful techniques from variational inequalities and monotone optimization that are complementary to those developed here—see for example Gemp and Mahadevan (2016, 2017); Gidel et al. (2019). Using techniques from monotone optimization, Gemp and Mahadevan (2018) obtained more detailed and stronger results than ours, in the more particular case of Wasserstein LQ-GANs, where the generator is linear and the discriminator is quadratic (Feizi et al., 2017; Nagarajan and Kolter, 2017).

Network zero-sum games are shown to be Hamiltonian systems in Bailey and Piliouras (2019). The implications of the existence of invariant functions for games is just beginning to be understood and explored.

1.3.1. Notation

Dot products are written as $\mathbf{v}^\intercal\mathbf{w}$ or $\langle\mathbf{v}, \mathbf{w}\rangle$. The angle between two vectors is $\theta(\mathbf{v}, \mathbf{w})$. Positive definiteness is denoted $\mathbf{S} \succ 0$.

2. The Infinitesimal Structure of Games

In contrast to the classical formulation of games, we do not constrain the parameter sets to the probability simplex or require losses to be convex in the corresponding players'

parameters. Our motivation is that we are primarily interested in use cases where players are interacting neural nets such as GANs (Goodfellow et al., 2014), a situation in which results from classical game theory do not straightforwardly apply.

Definition 1 (differentiable game)
A differentiable game consists in a set of players $[n] = \{1, \ldots, n\}$ *and corresponding twice continuously differentiable losses* $\{\ell_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$. *Parameters are* $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_n) \in \mathbb{R}^d$ *where* $\sum_{i=1}^n d_i = d$. *Player* i *controls* $\mathbf{w}_i \in \mathbb{R}^{d_i}$, *and aims to minimize its loss.*

It is sometimes convenient to write $\mathbf{w} = (\mathbf{w}_i, \mathbf{w}_{-i})$ where $\mathbf{w}_{-i}$ concatenates the parameters of all the players other than the i^{th}, which is placed out of order by abuse of notation.

The *simultaneous gradient* is the gradient of the losses with respect to the parameters of the respective players:

$$\boldsymbol{\xi}(\mathbf{w}) = (\nabla_{\mathbf{w}_1} \ell_1, \ldots, \nabla_{\mathbf{w}_n} \ell_n) \in \mathbb{R}^d.$$

By the **dynamics** of the game, we mean following the *negative* of the vector field, $-\boldsymbol{\xi}$, with infinitesimal steps. There is no reason to expect $\boldsymbol{\xi}$ to be the gradient of a *single* function in general, and therefore no reason to expect the dynamics to converge to a fixed point.

2.1. Hamiltonian Mechanics

Hamiltonian mechanics is a formalism for describing the dynamics in classical physical systems, see Arnold (1989); Guillemin and Sternberg (1990). The system is described via canonical coordinates $(\mathbf{q}, \mathbf{p})$. For example, $\mathbf{q}$ often refers to position and $\mathbf{p}$ to momentum of a particle or particles.

The Hamiltonian of the system $\mathcal{H}(\mathbf{q}, \mathbf{p})$ is a function that specifies the total energy as a function of the generalized coordinates. For example, in a closed system the Hamiltonian is given by the sum of the potential and kinetic energies of the particles. The time evolution of the system is given by Hamilton's equations:

$$\frac{d\mathbf{q}}{dt} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}} \quad \text{and} \quad \frac{d\mathbf{p}}{dt} = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}}.$$

An importance consequence of the Hamiltonian formalism is that the dynamics of the physical system—that is, the trajectories followed by the particles in phase space—live on the level sets of the Hamiltonian. In other words, the total energy is conserved.

2.2. Hamiltonian Mechanics in Games

The next example illustrates the essential problem with gradients in games and the key insight motivating our approach.

Example 1 (Conservation of energy in a zero-sum unconstrained bimatrix game)
Zero-sum games, where $\sum_{i=1}^n \ell_i \equiv 0$, *are well-studied. The zero-sum game*

$$\ell_1(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\intercal \mathbf{A} \mathbf{y} \quad \textit{and} \quad \ell_2(\mathbf{x}, \mathbf{y}) = -\mathbf{x}^\intercal \mathbf{A} \mathbf{y}$$

has a Nash equilibrium at $(\mathbf{x}, \mathbf{y}) = (\mathbf{0}, \mathbf{0})$. *The simultaneous gradient* $\boldsymbol{\xi}(\mathbf{x}, \mathbf{y}) = (\mathbf{A}\mathbf{y}, -\mathbf{A}^\intercal \mathbf{x})$ *rotates around the Nash, see Figure 1.*

The matrix $\mathbf{A}$ *admits singular value decomposition (SVD)* $\mathbf{A} = \mathbf{U}^\intercal\mathbf{D}\mathbf{V}$. *Changing to coordinates* $\mathbf{u} = \mathbf{D}^{\frac{1}{2}}\mathbf{U}\mathbf{x}$ *and* $\mathbf{v} = \mathbf{D}^{\frac{1}{2}}\mathbf{V}\mathbf{y}$ *gives* $\ell_1(\mathbf{u},\mathbf{v}) = \mathbf{u}^\intercal\mathbf{v}$ *and* $\ell_2(\mathbf{u},\mathbf{v}) = -\mathbf{u}^\intercal\mathbf{v}$. *Introduce the* Hamiltonian

$$\mathcal{H}(\mathbf{u},\mathbf{v}) = \frac{1}{2}\left(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2\right) = \frac{1}{2}\left(\mathbf{x}^\intercal\mathbf{U}^\intercal\mathbf{D}\mathbf{U}\mathbf{x} + \mathbf{y}^\intercal\mathbf{V}^\intercal\mathbf{D}\mathbf{V}\mathbf{y}\right).$$

Remarkably, the dynamics can be reformulated via Hamilton's equations in the coordinates given by the SVD of $\mathbf{A}$*:*

$$\boldsymbol{\xi}(\mathbf{u},\mathbf{v}) = \left(\frac{\partial\mathcal{H}}{\partial\mathbf{v}}, -\frac{\partial\mathcal{H}}{\partial\mathbf{u}}\right).$$

The vector field $\boldsymbol{\xi}$ *cycles around the equilibrium because* $\boldsymbol{\xi}$ *conserves the Hamiltonian's level sets (i.e.* $\langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle = 0$*). However,* **gradient descent on the Hamiltonian converges to the Nash equilibrium.** *The remainder of the paper explores the implications and limitations of this insight.*

Papadimitriou and Piliouras (2016) recently analyzed the dynamics of Matching Pennies (essentially, the above example) and showed that the cyclic behavior covers the entire parameter space. The Hamiltonian reformulation directly explains the cyclic behavior via a conservation law.

2.3. The Generalized Helmholtz Decomposition

The **Jacobian** of a game with dynamics $\boldsymbol{\xi}$ is the $(d \times d)$-matrix of second-derivatives $\mathbf{J}(\mathbf{w}) := \nabla_{\mathbf{w}} \cdot \boldsymbol{\xi}(\mathbf{w})^\intercal = \left(\frac{\partial\xi_\alpha(\mathbf{w})}{\partial w_\beta}\right)_{\alpha,\beta=1}^{d}$, where $\boldsymbol{\xi}_\alpha(\mathbf{w})$ is the α^{th} entry of the d-dimensional vector $\boldsymbol{\xi}(\mathbf{w})$. Concretely, the Jacobian can be written as

$$\mathbf{J}(\mathbf{w}) = \begin{pmatrix} \nabla^2_{\mathbf{w}_1}\ell_1 & \nabla^2_{\mathbf{w}_1,\mathbf{w}_2}\ell_1 & \cdots & \nabla^2_{\mathbf{w}_1,\mathbf{w}_n}\ell_1 \\ \nabla^2_{\mathbf{w}_2,\mathbf{w}_1}\ell_2 & \nabla^2_{\mathbf{w}_2}\ell_2 & \cdots & \nabla^2_{\mathbf{w}_2,\mathbf{w}_n}\ell_2 \\ \vdots & & & \vdots \\ \nabla^2_{\mathbf{w}_n,\mathbf{w}_1}\ell_n & \nabla^2_{\mathbf{w}_n,\mathbf{w}_2}\ell_n & \cdots & \nabla^2_{\mathbf{w}_n}\ell_n \end{pmatrix}$$

where $\nabla^2_{\mathbf{w}_i,\mathbf{w}_j}\ell_k$ is the $(d_i \times d_j)$-block of 2^{nd}-order derivatives. The Jacobian of a game is a square matrix, but not necessarily symmetric. Note: Greek indices α, β run over d parameter dimensions whereas Roman indices i, j run over n players.

Lemma 1 (generalized Helmholtz decomposition)
The Jacobian of any vector field decomposes uniquely into two components $\mathbf{J}(\mathbf{w}) = \mathbf{S}(\mathbf{w}) + \mathbf{A}(\mathbf{w})$ *where* $\mathbf{S} \equiv \mathbf{S}^\intercal$ *is symmetric and* $\mathbf{A} + \mathbf{A}^\intercal \equiv 0$ *is antisymmetric.*

Proof Any matrix decomposes uniquely as $\mathbf{M} = \mathbf{S} + \mathbf{A}$ where $\mathbf{S} = \frac{1}{2}(\mathbf{M} + \mathbf{M}^\intercal)$ and $\mathbf{A} = \frac{1}{2}(\mathbf{M} - \mathbf{M}^\intercal)$ are symmetric and antisymmetric. The decomposition is preserved by orthogonal change-of-coordinates: given orthogonal matrix $\mathbf{P}$, we have $\mathbf{P}^\intercal\mathbf{M}\mathbf{P} = \mathbf{P}^\intercal\mathbf{S}\mathbf{P} + \mathbf{P}^\intercal\mathbf{A}\mathbf{P}$ since the terms remain symmetric and antisymmetric. Applying the decomposition to the Jacobian yields the result. ■

The connection to the classical Helmholtz decomposition in calculus is sketched in appendix B. Two natural classes of games arise from the decomposition:

Definition 2 *A game is a* ***potential game*** *if the Jacobian is symmetric, i.e. if* $\mathbf{A}(\mathbf{w}) \equiv 0$. *It is a* ***Hamiltonian game*** *if the Jacobian is antisymmetric, i.e. if* $\mathbf{S}(\mathbf{w}) \equiv 0$.

Potential games are well-studied and easy to solve. Hamiltonian games are a new class of games that are also easy to solve. The general case is more difficult, see Section 3.

2.4. Stable Fixed Points (SFPs) vs Local Nash Equilibria (LNEs)

There are (at least) two possible solution concepts in general differentiable games: stable fixed points and local Nash equilibria.

Definition 3 *A point* $\mathbf{w}$ *is a* ***local Nash equilibrium*** *if, for all* i*, there exists a neighborhood* U_i *of* $\mathbf{w}_i$ *such that* $\ell_i(\mathbf{w}_i', \mathbf{w}_{-i}) \geq \ell_i(\mathbf{w}_i, \mathbf{w}_{-i})$ *for* $\mathbf{w}_i' \in U_i$.

We introduce *local* Nash equilibria because finding *global* Nash equilibria is unrealistic in games involving neural nets. Gradient-based methods can reliably find local—but not global—optima of nonconvex objective functions (Lee et al., 2016, 2017). Similarly, gradient-based methods cannot be expected to find global Nash equilibria in nonconvex games.

Definition 4 *A fixed point* $\mathbf{w}^*$ *with* $\boldsymbol{\xi}(\mathbf{w}^*) = 0$ *is* **stable** *if* $\mathbf{J}(\mathbf{w}^*) \succeq 0$ *and* $\mathbf{J}(\mathbf{w}^*)$ *is invertible,* **unstable** *if* $\mathbf{J}(\mathbf{w}^*) \prec 0$ *and a* **strict saddle** *if* $\mathbf{J}(\mathbf{w}^*)$ *has an eigenvalue with negative real part. Strict saddles are a subset of unstable fixed points.*

The definition is adapted from Letcher et al. (2019), where conditions on the Jacobian hold *at* the fixed point; in contrast, Balduzzi et al. (2018a) imposed conditions on the Jacobian in a *neighborhood* of the fixed point. We motivate this concept as follows.

Positive semidefiniteness, $\mathbf{J}(\mathbf{w}^*) \succeq 0$, is a minimal condition for any reasonable notion of stable fixed point. In the case of a single loss ℓ, the Jacobian of $\boldsymbol{\xi} = \nabla\ell$ is the Hessian of ℓ, i.e. $\mathbf{J} = \nabla^2\ell$. Local convergence of gradient descent on single functions cannot be guaranteed if $\mathbf{J}(\mathbf{w}^*) \not\succeq 0$, since such points are strict saddles. These are almost always avoided by Lee et al. (2017), so this semidefinite condition must hold.

Another viewpoint is that invertibility and positive semidefiniteness of the Hessian together imply *positive definiteness*, and the notion of stable fixed point specializes, in a one-player game, to local minima that are detected by the second partial derivative test. These minima are precisely those which gradient-like methods provably converge to. Stable fixed points are defined by analogy, though note that invertibility and semidefiniteness do *not* imply positive definiteness in n-player games since $\mathbf{J}$ may not be symmetric.

Finally, it is important to impose only positive *semi*-definiteness to keep the class as large as possible. Imposing strict positivity would imply that the origin is not an SFP in the cyclic game $\ell_1 = xy = -\ell_2$ from Example 1, while clearly deserving of being so.

Remark 1 *The conditions* $\mathbf{J}(\mathbf{w}^*) \succeq 0$ *and* $\mathbf{J}(\mathbf{w}^*) \prec 0$ *are equivalent to the conditions on the symmetric component* $\mathbf{S}(\mathbf{w}^*) \succeq 0$ *and* $\mathbf{S}(\mathbf{w}^*) \prec 0$ *respectively, since*

$$\mathbf{u}^\intercal \mathbf{J} \mathbf{u} = \mathbf{u}^\intercal \mathbf{S} \mathbf{u} + \mathbf{u}^\intercal \mathbf{A} \mathbf{u} = \mathbf{u}^\intercal \mathbf{S} \mathbf{u}$$

for all $\mathbf{u}$*, by antisymmetry of* $\mathbf{A}$. *This equivalence will be used throughout.*

Stable fixed points and local Nash equilibria are both appealing solution concepts, one from the viewpoint of optimisation by analogy with single objectives, and the other from game theory. Unfortunately, neither is a subset of the other:

Example 2 (stable $\not\Longrightarrow$ local Nash)
Let $\ell_1(x,y) = x^3 + xy$ and $\ell_2(x,y) = -xy$. Then

$$\boldsymbol{\xi}(x,y) = \begin{pmatrix} 3x^2 + y \\ -x \end{pmatrix} \quad \textit{and} \quad \mathbf{J}(x,y) = \begin{pmatrix} 6x & 1 \\ -1 & 0 \end{pmatrix}.$$

There is a stable fixed point with invertible Hessian at $(x,y) = (0,0)$, since $\boldsymbol{\xi}(0,0) = 0$ and $\mathbf{J}(0,0) \succeq 0$ invertible. However any neighbourhood of $x = 0$ contains some small $\epsilon > 0$ for which $\ell_1(-\epsilon, 0) = -\epsilon^3 < 0 = \ell_1(0,0)$, so the origin is not a local Nash equilibrium.

Example 3 (local Nash $\not\Longrightarrow$ stable)
Let $\ell_1(x,y) = \ell_2(x,y) = xy$. Then

$$\boldsymbol{\xi}(x,y) = \begin{pmatrix} y \\ x \end{pmatrix} \quad \textit{and} \quad \mathbf{J}(x,y) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

There is a fixed point at $(x,y) = (0,0)$ which is a local (in fact, global) Nash equilibrium since $\ell_1(0,y) = 0 \geq \ell_1(0,0)$ and $\ell_2(x,0) = 0 \geq \ell_2(0,0)$ for all $x, y \in \mathbb{R}$. However $\mathbf{J} = \mathbf{S}$ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -1 < 0$, so $(0,0)$ is not a stable fixed point.

In Example 3, the Nash equilibrium is a *saddle point* of the common loss $\ell = xy$. Any algorithm that converges to Nash equilibria will thus converge to an undesirable saddle point. This rules out local Nash equilibrium as a solution concept for our purposes. Conversely, Example 2 emphasises the better notion of stability whereby player 1 may have a local incentive to deviate from the origin *immediately*, but would later be punished for doing so since the game is locally dominated by the $\pm xy$ terms, whose only 'resolution' or 'stable minimum' is the origin (see Example 1).

2.5. Potential Games

Potential games were introduced by Monderer and Shapley (1996). It turns out that our definition of potential game above coincides with a special case of the potential games of Monderer and Shapley (1996), which they refer to as exact potential games.

Definition 5 (classical definition of potential game)
A game is a potential game if there is a single potential function $\phi : \mathbb{R}^d \to \mathbb{R}$ and positive numbers $\{\alpha_i > 0\}_{i=1}^n$ such that

$$\phi(\mathbf{w}_i', \mathbf{w}_{-i}) - \phi(\mathbf{w}_i'', \mathbf{w}_{-i}) = \alpha_i \Big(\ell_i(\mathbf{w}_i', \mathbf{w}_{-i}) - \ell_i(\mathbf{w}_i'', \mathbf{w}_{-i}) \Big)$$

for all i and all $\mathbf{w}_i', \mathbf{w}_i'', \mathbf{w}_{-i}$, see Monderer and Shapley (1996).

Lemma 2 *A game is a potential game iff $\alpha_i \nabla_{\mathbf{w}_i} \ell_i = \nabla_{\mathbf{w}_i} \phi$ for all i, which is equivalent to*

$$\alpha_i \nabla^2_{\mathbf{w}_i \mathbf{w}_j} \ell_i = \alpha_j \nabla^2_{\mathbf{w}_i \mathbf{w}_j} \ell_j = \alpha_j \left(\nabla^2_{\mathbf{w}_j \mathbf{w}_i} \ell_j \right)^\intercal \quad \forall i, j. \tag{1}$$

Proof See Monderer and Shapley (1996). ■

Corollary 3 *If $\alpha_i = 1$ for all i then Equation* (1) *is equivalent to requiring that the Jacobian of the game is symmetric.*

Proof In an exact potential game, the Jacobian coincides with the Hessian of the potential function ϕ, which is necessarily symmetric. ■

Monderer and Shapley (1996) refer to the special case where $\alpha_i = 1$ for all i as an ***exact potential game***. We use the shorthand 'potential game' to refer to exact potential games in what follows.

Potential games have been extensively studied since they are one of the few classes of games for which Nash equilibria can be computed (Rosenthal, 1973). For our purposes, they are games where simultaneous gradient descent on the losses corresponds to gradient descent on a single function. It follows that descent on $\boldsymbol{\xi}$ converges to a fixed point that is a local minimum of ϕ or a saddle.

2.6. Hamiltonian Games

Hamiltonian games, where the Jacobian is antisymmetric, are a new class games. They are related to the harmonic games introduced in Candogan et al. (2011), see Section B.4. An example from Balduzzi et al. (2018b) may help develop intuition for antisymmetric matrices:

Example 4 (antisymmetric structure of tournaments)
Suppose n competitors play one-on-one and that the probability of player i beating player j is p_{ij}. Then, assuming there are no draws, the probabilities satisfy $p_{ij} + p_{ji} = 1$ and $p_{ii} = \frac{1}{2}$. The matrix $\mathbf{A} = \left(\log \frac{p_{ij}}{1-p_{ij}}\right)_{i,j=1}^{n}$ of logits is then antisymmetric. Intuitively, antisymmetry reflects a hyperadversarial *setting where all pairwise interactions between players are zero-sum.*

Hamiltonian games are closely related to zero-sum games.

Example 5 (an unconstrained bimatrix game is zero-sum iff it is Hamiltonian)
Consider bimatrix game with $\ell_1(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\intercal \mathbf{P}\mathbf{y}$ and $\ell_2(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\intercal \mathbf{Q}\mathbf{y}$, but where the parameters are not *constrained to the probability simplex. Then $\boldsymbol{\xi} = (\mathbf{P}\mathbf{y}, \mathbf{Q}^\intercal \mathbf{x})$ and the Jacobian components have block structure*

$$\mathbf{A} = \frac{1}{2}\begin{pmatrix} 0 & \mathbf{P} - \mathbf{Q} \\ (\mathbf{Q} - \mathbf{P})^\intercal & 0 \end{pmatrix} \quad \textit{and} \quad \mathbf{S} = \frac{1}{2}\begin{pmatrix} 0 & \mathbf{P} + \mathbf{Q} \\ (\mathbf{P} + \mathbf{Q})^\intercal & 0 \end{pmatrix}$$

The game is Hamiltonian iff $\mathbf{S} = 0$ iff $\mathbf{P} + \mathbf{Q} = 0$ iff $\ell_1 + \ell_2 = 0$.

However, in general there are Hamiltonian games that are *not* zero-sum and vice versa.

Example 6 (Hamiltonian game that is not zero-sum)
Fix constants a and b and suppose players 1 and 2 minimize losses

$$\ell_1(x, y) = x(y - b) \;\; \textit{and} \;\; \ell_2(x, y) = -(x - a)y$$

with respect to x and y respectively.

Example 7 (zero-sum game that is not Hamiltonian)
Players 1 and 2 minimize

$$\ell_1(x, y) = x^2 + y^2 \qquad \ell_2(x, y) = -(x^2 + y^2).$$

The game actually has potential function $\phi(x, y) = x^2 - y^2$.

Hamiltonian games are quite different from potential games. In a Hamiltonian game there is a Hamiltonian function $\mathcal{H}$ that specifies a conserved quantity. In potential games the dynamics *equal* $\nabla\phi$; in Hamiltonian games the dynamics are *orthogonal* to $\nabla\mathcal{H}$. The orthogonality implies the conservation law that underlies the cyclic behavior in example 1.

Theorem 4 (conservation law for Hamiltonian games)
Let $\mathcal{H}(\mathbf{w}) := \frac{1}{2}\|\boldsymbol{\xi}(\mathbf{w})\|_2^2$. If the game is Hamiltonian then

i) $\nabla\mathcal{H} = \mathbf{A}^\intercal\boldsymbol{\xi}$ *and*

ii) $\boldsymbol{\xi}$ ***preserves the level sets of*** $\mathcal{H}$ *since* $\langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle = 0$.

iii) If the Jacobian is invertible and $\lim_{\|\mathbf{w}\|\to\infty}\mathcal{H}(\mathbf{w}) = \infty$ *then gradient descent on* $\mathcal{H}$ *converges to a stable fixed point.*

Proof Direct computation shows $\nabla\mathcal{H} = \mathbf{J}^\intercal\boldsymbol{\xi}$ for any game. The first statement follows since $\mathbf{J} = \mathbf{A}$ in Hamiltonian games.

For the second statement, the directional derivative is $D_{\boldsymbol{\xi}}\mathcal{H} = \langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle = \boldsymbol{\xi}^\intercal\mathbf{A}^\intercal\boldsymbol{\xi}$ where $\boldsymbol{\xi}^\intercal\mathbf{A}^\intercal\boldsymbol{\xi} = (\boldsymbol{\xi}^\intercal\mathbf{A}^\intercal\boldsymbol{\xi})^\intercal = \boldsymbol{\xi}^\intercal\mathbf{A}\boldsymbol{\xi} = -(\boldsymbol{\xi}^\intercal\mathbf{A}^\intercal\boldsymbol{\xi})$ since $\mathbf{A} = -\mathbf{A}^\intercal$ by anti-symmetry. It follows that $\boldsymbol{\xi}^\intercal\mathbf{A}^\intercal\boldsymbol{\xi} = 0$.

For the third statement, gradient descent on $\mathcal{H}$ will converge to a point where $\nabla\mathcal{H} = \mathbf{J}^\intercal\boldsymbol{\xi}(\mathbf{w}) = 0$. If the Jacobian is invertible then clearly $\boldsymbol{\xi}(\mathbf{w}) = 0$. The fixed-point is stable since $0 \equiv \mathbf{S} \succeq 0$ in a Hamiltonian game, recall remark 1. ■

In fact, $\mathcal{H}$ is a Hamiltonian function for the game dynamics, see appendix B for a concise explanation. We use the notation $\mathcal{H}(\mathbf{w}) = \frac{1}{2}\|\boldsymbol{\xi}(\mathbf{w})\|^2$ throughout the paper. However, $\mathcal{H}$ can only be interpreted as a Hamiltonian function for $\boldsymbol{\xi}$ when the game is Hamiltonian.

There is a precise mapping from Hamiltonian games to symplectic geometry, see appendix B. Symplectic geometry is the modern formulation of classical mechanics (Arnold, 1989; Guillemin and Sternberg, 1990). Recall that periodic behaviors (e.g. orbits) often arise in classical mechanics. The orbits lie on the level sets of the Hamiltonian, which expresses the total energy of the system.

3. Algorithms

We have seen that fixed points of potential and Hamiltonian games can be found by descent on $\boldsymbol{\xi}$ and $\nabla\mathcal{H}$ respectively. This Section tackles finding stable fixed points in general games.

3.1. Finding Stable Fixed Points

There are two classes of games where we know how to find stable fixed points: potential games where $\boldsymbol{\xi}$ converges to a local minimum and Hamiltonian games where $\nabla\mathcal{H}$, which is orthogonal to $\boldsymbol{\xi}$, finds stable fixed points.

In the general case, the following desiderata provide a set of reasonable properties for an adjustment $\boldsymbol{\xi}_\lambda$ of the game dynamics. Recall that $\theta(\mathbf{u}, \mathbf{v})$ is the angle between the vectors $\mathbf{u}$ and $\mathbf{v}$.

3.1.1. Desiderata

To find stable fixed points, an adjustment $\boldsymbol{\xi}_\lambda$ to the game dynamics should satisfy

D1. *compatible*[3] *with game dynamics:* $\langle \boldsymbol{\xi}_\lambda, \boldsymbol{\xi} \rangle = \alpha_1 \cdot \|\boldsymbol{\xi}\|^2$;

D2. *compatible with potential dynamics:*
if the game is a potential game then $\langle \boldsymbol{\xi}_\lambda, \nabla\phi \rangle = \alpha_2 \cdot \|\nabla\phi\|^2$;

D3. *compatible with Hamiltonian dynamics:*
If the game is Hamiltonian then $\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle = \alpha_3 \cdot \|\nabla\mathcal{H}\|^2$;

D4. *attracted to stable equilibria:*
in neighborhoods where $\mathbf{S} \succ 0$, require $\theta(\boldsymbol{\xi}_\lambda, \nabla\mathcal{H}) \leq \theta(\boldsymbol{\xi}, \nabla\mathcal{H})$;

D5. *repelled by unstable equilibria:*
in neighborhoods where $\mathbf{S} \prec 0$, require $\theta(\boldsymbol{\xi}_\lambda, \nabla\mathcal{H}) \geq \theta(\boldsymbol{\xi}, \nabla\mathcal{H})$.

for some $\alpha_1, \alpha_2, \alpha_3 > 0$.

Desideratum $D1$ does not guarantee that players act in their own self-interest—this requires a stronger positivity condition on dot-products with subvectors of $\boldsymbol{\xi}$, see Balduzzi (2017). Desiderata $D2$ and $D3$ imply that the adjustment behaves correctly in potential and Hamiltonian games respectively.

To understand desiderata $D4$ and $D5$, observe that gradient descent on $\mathcal{H} = \frac{1}{2}\|\boldsymbol{\xi}\|^2$ will find local minima that are fixed points of the dynamics. However, we specifically wish to converge to stable fixed points. Desideratum $D4$ and $D5$ require that the adjustment improves the rate of convergence to stable fixed points (by finding a steeper angle of descent), and avoids unstable fixed points.

More concretely, desiderata $D4$ can be interpreted as follows. If $\boldsymbol{\xi}$ points at a stable equilibrium then we require that $\boldsymbol{\xi}_\lambda$ points *more* towards the equilibrium (i.e. has smaller angle). Conversely, desiderata $D5$ requires that if $\boldsymbol{\xi}$ points away then the adjustment should point *further* away.

The unadjusted dynamics $\boldsymbol{\xi}$ satisfies all the desiderata except $D3$.

3. Two nonzero vectors are compatible if they have positive inner product.

3.2. Consensus Optimization

Since gradient descent on the function $\mathcal{H}(\mathbf{w}) = \frac{1}{2}\|\boldsymbol{\xi}\|^2$ finds stable fixed points in Hamiltonian games, it is natural to ask how it performs in general games. If the Jacobian $\mathbf{J}(\mathbf{w})$ is invertible, then $\nabla\mathcal{H} = \mathbf{J}^\intercal\boldsymbol{\xi} = 0$ iff $\boldsymbol{\xi} = 0$. Thus, gradient descent on $\mathcal{H}$ converges to fixed points of $\boldsymbol{\xi}$.

However, there is no guarantee that descent on $\mathcal{H}$ will find a *stable* fixed point. Mescheder et al. (2017) propose *consensus optimization*, a gradient adjustment of the form

$$\boldsymbol{\xi} + \lambda \cdot \mathbf{J}^\intercal\boldsymbol{\xi} = \boldsymbol{\xi} + \lambda \cdot \nabla\mathcal{H}.$$

Unfortunately, consensus optimization can converge to unstable fixed points even in simple cases where the 'game' is to minimize a single function:

Example 8 (consensus optimization can converge to a global maximum) *Consider a potential game with losses $\ell_1(x, y) = \ell_2(x, y) = -\frac{\kappa}{2}(x^2 + y^2)$ with $\kappa \gg 0$. Then*

$$\boldsymbol{\xi} = -\kappa \cdot \begin{pmatrix} x \\ y \end{pmatrix} \quad \textit{and} \quad \mathbf{J} = -\begin{pmatrix} \kappa & 0 \\ 0 & \kappa \end{pmatrix}$$

Note that $\|\boldsymbol{\xi}\|^2 = \kappa^2(x^2 + y^2)$ and

$$\boldsymbol{\xi} + \lambda \cdot \mathbf{J}^\intercal\boldsymbol{\xi} = \kappa(\lambda\kappa - 1) \cdot \begin{pmatrix} x \\ y \end{pmatrix}.$$

Descent on $\boldsymbol{\xi} + \lambda \cdot \mathbf{J}^\intercal\boldsymbol{\xi}$ converges to the global maximum $(x, y) = (0, 0)$ unless $\lambda < \frac{1}{\kappa}$.

Although consensus optimization works well in two-player zero-sum, it cannot be considered a candidate algorithm for finding stable fixed points in general games since it fails in the basic case of potential games. Consensus optimization only satisfies desiderata $D3$ and $D4$.

3.3. Symplectic Gradient Adjustment

The problem with consensus optimization is that it can perform worse than gradient descent on potential games. Intuitively, it makes bad use of the symmetric component of the Jacobian. Motivated by the analysis in Section 2, we propose symplectic gradient adjustment, which takes care to only use the antisymmetric component of the Jacobian when adjusting the dynamics.

Proposition 5 *The **symplectic gradient adjustment (SGA)***

$$\boldsymbol{\xi}_\lambda := \boldsymbol{\xi} + \lambda \cdot \mathbf{A}^\intercal\boldsymbol{\xi}.$$

satisfies $D1$—$D3$ for $\lambda > 0$, with $\alpha_1 = 1 = \alpha_2$ and $\alpha_3 = \lambda$.

Proof First claim: $\lambda \cdot \boldsymbol{\xi}^\intercal\mathbf{A}^\intercal\boldsymbol{\xi} = 0$ by anti-symmetry of $\mathbf{A}$. Second claim: $\mathbf{A} \equiv 0$ in a potential game, so $\boldsymbol{\xi}_\lambda = \boldsymbol{\xi} = \nabla\phi$. Third claim: $\langle\boldsymbol{\xi}_\lambda, \nabla\mathcal{H}\rangle = \langle\boldsymbol{\xi}_\lambda, \mathbf{J}^\intercal\boldsymbol{\xi}\rangle = \langle\boldsymbol{\xi}_\lambda, \mathbf{A}^\intercal\boldsymbol{\xi}\rangle = \lambda \cdot \boldsymbol{\xi}^\intercal\mathbf{A}\mathbf{A}^\intercal\boldsymbol{\xi} = \lambda \cdot \|\nabla\mathcal{H}\|^2$ since $\mathbf{J} = \mathbf{A}$ by assumption. ■

Note that desiderata $D1$ and $D2$ are true even when $\lambda < 0$. This will prove useful, since example 9 shows that it may be necessary to pick negative λ near $\mathbf{S} \prec 0$. Section 3.5 shows how to also satisfy desiderata $D4$ and $D5$.

3.4. Convergence

We begin by analysing convergence of SGA near stable equilibria. The following lemma highlights that the interaction between the symmetric and antisymmetric components is important for convergence. Recall that two matrices $\mathbf{A}$ and $\mathbf{S}$ *commute* iff $[\mathbf{A}, \mathbf{S}] := \mathbf{AS} - \mathbf{SA} = \mathbf{0}$. That is, $\mathbf{A}$ and $\mathbf{S}$ commute iff $\mathbf{AS} = \mathbf{SA}$. Intuitively, two matrices commute if they have the same preferred coordinate system.

Lemma 6 *If $\mathbf{S} \succeq 0$ is symmetric positive semidefinite and $\mathbf{S}$ commutes with $\mathbf{A}$ then $\boldsymbol{\xi}_\lambda$ points towards stable fixed points for non-negative λ:*

$$\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle \geq 0 \text{ for all } \lambda \geq 0.$$

Proof First observe that $\boldsymbol{\xi}^\intercal \mathbf{AS}\boldsymbol{\xi} = \boldsymbol{\xi}^\intercal \mathbf{S}^\intercal \mathbf{A}^\intercal \boldsymbol{\xi} = -\boldsymbol{\xi}^\intercal \mathbf{SA}\boldsymbol{\xi}$, where the first equality holds since the expression is a scalar, and the second holds since $\mathbf{S} = \mathbf{S}^\intercal$ and $\mathbf{A} = -\mathbf{A}^\intercal$. It follows that $\boldsymbol{\xi}^\intercal \mathbf{AS}\boldsymbol{\xi} = 0$ if $\mathbf{SA} = \mathbf{AS}$. Finally rewrite the inequality as

$$\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle = \langle \boldsymbol{\xi} + \lambda \cdot \mathbf{A}^\intercal \boldsymbol{\xi}, \mathbf{S}\boldsymbol{\xi} + \mathbf{A}^\intercal \boldsymbol{\xi} \rangle = \boldsymbol{\xi}^\intercal \mathbf{S}\boldsymbol{\xi} + \lambda \boldsymbol{\xi}^\intercal \mathbf{AA}^\intercal \boldsymbol{\xi} \geq 0$$

since $\boldsymbol{\xi}^\intercal \mathbf{AS}\boldsymbol{\xi} = 0$ and by positivity of $\mathbf{S}$, λ and $\mathbf{AA}^\intercal$. ■

The lemma suggests that in general the *failure* of $\mathbf{A}$ and $\mathbf{S}$ to commute should be important for understanding the dynamics of $\boldsymbol{\xi}_\lambda$. We therefore introduce the **additive condition number** κ to upper-bound the worst-case noncommutativity of $\mathbf{S}$, which allows to quantify the relationship between $\boldsymbol{\xi}_\lambda$ and $\nabla\mathcal{H}$. If $\kappa = 0$, then $\mathbf{S} = \sigma \cdot \mathbf{I}$ commutes with *all* matrices. The larger the additive condition number κ, the larger the *potential* failure of $\mathbf{S}$ to commute with other matrices.

Theorem 7 *Let $\mathbf{S}$ be a symmetric matrix with eigenvalues $\sigma_{max} \geq \cdots \geq \sigma_{min}$. The **additive condition number**[4] of $\mathbf{S}$ is $\kappa := \sigma_{max} - \sigma_{min}$. If $\mathbf{S} \succeq 0$ is positive semidefinite with additive condition number κ then $\lambda \in (0, \frac{4}{\kappa})$ implies*

$$\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle \geq 0.$$

If $\mathbf{S}$ is negative semidefinite, then $\lambda \in (0, \frac{4}{\kappa})$ implies

$$\langle \boldsymbol{\xi}_{-\lambda}, \nabla\mathcal{H} \rangle \leq 0.$$

The inequalities are strict if $\mathbf{J}$ is invertible.

Proof We prove the case $\mathbf{S} \succeq 0$; the case $\mathbf{S} \preceq 0$ is similar. Rewrite the inequality as

$$\begin{aligned} \langle \boldsymbol{\xi} + \lambda \cdot \mathbf{A}^\intercal \boldsymbol{\xi}, \nabla\mathcal{H} \rangle &= (\boldsymbol{\xi} + \lambda \cdot \mathbf{A}^\intercal \boldsymbol{\xi})^\intercal \cdot (\mathbf{S} + \mathbf{A}^\intercal)\boldsymbol{\xi} \\ &= \boldsymbol{\xi}^\intercal \mathbf{S}\boldsymbol{\xi} + \lambda \boldsymbol{\xi}^\intercal \mathbf{AS}\boldsymbol{\xi} + \lambda \boldsymbol{\xi}^\intercal \mathbf{AA}^\intercal \boldsymbol{\xi} \end{aligned}$$

4. The condition number of a positive definite matrix is $\frac{\sigma_{\max}}{\sigma_{\min}}$.

Let $\beta = \|A^\intercal \boldsymbol{\xi}\|$ and $\tilde{\mathbf{S}} = \mathbf{S} - \sigma_{\min} \cdot \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Then

$$\boldsymbol{\xi}^\intercal \mathbf{S} \boldsymbol{\xi} + \lambda \boldsymbol{\xi}^\intercal \mathbf{A} \mathbf{S} \boldsymbol{\xi} + \lambda \cdot \beta^2 \geq \boldsymbol{\xi}^\intercal \tilde{\mathbf{S}} \boldsymbol{\xi} + \lambda \boldsymbol{\xi}^\intercal \mathbf{A} \tilde{\mathbf{S}} \boldsymbol{\xi} + \lambda \cdot \beta^2$$

since $\boldsymbol{\xi}^\intercal \mathbf{S} \boldsymbol{\xi} \geq \boldsymbol{\xi}^\intercal \tilde{\mathbf{S}} \boldsymbol{\xi}$ by construction and $\boldsymbol{\xi}^\intercal \mathbf{A} \tilde{\mathbf{S}} \boldsymbol{\xi} = \boldsymbol{\xi}^\intercal \mathbf{A} \mathbf{S} \boldsymbol{\xi} - \sigma_{\min} \boldsymbol{\xi}^\intercal \mathbf{A} \boldsymbol{\xi} = \boldsymbol{\xi}^\intercal \mathbf{A} \mathbf{S} \boldsymbol{\xi}$ because $\boldsymbol{\xi}^\intercal \mathbf{A} \boldsymbol{\xi} = 0$ by the anti-symmetry of $\mathbf{A}$. It therefore suffices to show that the inequality holds when $\sigma_{\min} = 0$ and $\kappa = \sigma_{\max}$.

Since $\mathbf{S}$ is positive semidefinite, there exists an upper-triangular square-root matrix T such that $\mathbf{T}^\intercal \mathbf{T} = \mathbf{S}$ and so $\boldsymbol{\xi}^\intercal \mathbf{S} \boldsymbol{\xi} = \|\mathbf{T} \boldsymbol{\xi}\|^2$. Further,

$$|\boldsymbol{\xi}^\intercal \mathbf{A} \mathbf{S} \boldsymbol{\xi}| \leq \|\mathbf{A}^\intercal \boldsymbol{\xi}\| \cdot \|\mathbf{T}^\intercal \mathbf{T} \boldsymbol{\xi}\| \leq \sqrt{\sigma_{\max}} \cdot \|\mathbf{A}^\intercal \boldsymbol{\xi}\| \cdot \|\mathbf{T} \boldsymbol{\xi}\|.$$

since $\|\mathbf{T}\|_2 = \sqrt{\sigma_{\max}}$. Putting the observations together obtains

$$\begin{aligned}
\|\mathbf{T}\boldsymbol{\xi}\|^2 + \lambda(\|\mathbf{A}\boldsymbol{\xi}\|^2 - \langle \mathbf{A}\boldsymbol{\xi}, \mathbf{S}\boldsymbol{\xi} \rangle) &\geq \|\mathbf{T}\boldsymbol{\xi}\|^2 + \lambda(\|\mathbf{A}\boldsymbol{\xi}\|^2 - \|\mathbf{A}\boldsymbol{\xi}\| \ \|\mathbf{S}\boldsymbol{\xi}\| \\
&\geq \|\mathbf{T}\boldsymbol{\xi}\|^2 + \lambda \|\mathbf{A}\boldsymbol{\xi}\|(\|\mathbf{A}\boldsymbol{\xi}\| - \|\mathbf{S}\boldsymbol{\xi}\|) \\
&\geq \|\mathbf{T}\boldsymbol{\xi}\|^2 + \lambda \|\mathbf{A}\boldsymbol{\xi}\|(\|\mathbf{A}\boldsymbol{\xi}\| - \sqrt{\sigma_{max}} \|\mathbf{T}\boldsymbol{\xi}\|)
\end{aligned}$$

Set $\alpha = \sqrt{\lambda}$ and $\eta = \sqrt{\sigma_{max}}$. We can continue the above computation

$$\begin{aligned}
\|\mathbf{T}\boldsymbol{\xi}\|^2 + \lambda(\|\mathbf{A}\boldsymbol{\xi}\|^2 - \langle \mathbf{A}\boldsymbol{\xi}, \mathbf{S}\boldsymbol{\xi} \rangle) &\geq \|\mathbf{T}\boldsymbol{\xi}\|^2 + \alpha^2 \|\mathbf{A}\boldsymbol{\xi}\|(\|\mathbf{A}\boldsymbol{\xi}\| - \eta \|\mathbf{T}\boldsymbol{\xi}\|) \\
&= \|\mathbf{T}\boldsymbol{\xi}\|^2 + \alpha^2 \|\mathbf{A}\boldsymbol{\xi}\|^2 - \alpha^2 \|\mathbf{A}\boldsymbol{\xi}\| \eta \|\mathbf{T}\boldsymbol{\xi}\| \\
&= (\|\mathbf{T}\boldsymbol{\xi}\| - \alpha \|\mathbf{A}\boldsymbol{\xi}\|)^2 + 2\alpha \|\mathbf{A}\boldsymbol{\xi}\| \ \|\mathbf{T}\boldsymbol{\xi}\| - \alpha^2 \eta \|A\boldsymbol{\xi}\| \ \|\mathbf{T}\boldsymbol{\xi}\| \\
&= (\|\mathbf{T}\boldsymbol{\xi}\| - \alpha \|\mathbf{A}\boldsymbol{\xi}\|)^2 + \|\mathbf{A}\boldsymbol{\xi}\| \ \|\mathbf{T}\boldsymbol{\xi}\|(2\alpha - \alpha^2 \eta)
\end{aligned}$$

Finally, $2\alpha - \alpha^2 \eta > 0$ for any α in the range $(0, \frac{2}{\eta})$, which is to say, for any $0 < \lambda < \frac{4}{\sigma_{max}}$. The kernel of $\mathbf{S}$ and the kernel of $\mathbf{T}$ coincide. If $\boldsymbol{\xi}$ is in the kernel of $\mathbf{A}$, resp. $\mathbf{T}$, it cannot be in the kernel of $\mathbf{T}$, resp. $\mathbf{A}$ and the term $(\|\mathbf{T}\boldsymbol{\xi}\| - \alpha \|\mathbf{A}\boldsymbol{\xi}\|)^2$ is positive. Otherwise, the term $\|\mathbf{A}\boldsymbol{\xi}\| \|\mathbf{T}\boldsymbol{\xi}\|$ is positive. ■

The theorem above guarantees that SGA always points in the direction of stable fixed points for λ sufficiently small. This does not technically guarantee convergence; we use Ostrowski's theorem to strengthen this formally. Applying Ostrowski's theorem will require taking a more abstract perspective by encoding the adjusted dynamics into a differentiable map $F : \Omega \to \mathbb{R}^d$ of the form $F(\mathbf{w}) = \mathbf{w} - \alpha \boldsymbol{\xi}_\lambda(\mathbf{w})$.

Theorem 8 (Ostrowski) *Let $F : \Omega \to \mathbb{R}^d$ be a continuously differentiable map on an open subset $\Omega \subseteq \mathbb{R}^d$, and assume $\mathbf{w}^* \in \Omega$ is a fixed point. If all eigenvalues of $\nabla F(\mathbf{w}^*)$ are strictly in the unit circle of $\mathbb{C}$, then there is an open neighbourhood U of $\mathbf{w}^*$ such that for all $\mathbf{w}_0 \in U$, the sequence $F^k(\mathbf{w}_0)$ of iterates of F converges to $\mathbf{w}^*$. Moreover, the rate of convergence is at least linear in k.*

Proof This is a standard result on fixed-point iterations, adapted from Ortega and Rheinboldt (2000, 10.1.3). ■

Corollary 9 *A matrix* $\mathbf{M}$ *is called positive stable if all its eigenvalues have positive real part. Assume* $\mathbf{w}^*$ *is a fixed point of a differentiable game such that* $(\mathbf{I} + \lambda\mathbf{A}^\intercal)\mathbf{J}(\mathbf{w}^*)$ *is positive stable for* λ *in some set* Λ*. Then SGA converges locally to* $\mathbf{w}^*$ *for* $\lambda \in \Lambda$ *and* $\alpha > 0$ *sufficiently small.*

Proof Let $X = (\mathbf{I} + \lambda\mathbf{A}^\intercal)$. By definition of fixed points, $\boldsymbol{\xi}(\mathbf{w}^*) = 0$ and so

$$\nabla[X\boldsymbol{\xi}](\mathbf{w}^*) = \nabla X(\mathbf{w}^*)\boldsymbol{\xi}(\mathbf{w}^*) + X(\mathbf{w}^*)\nabla\boldsymbol{\xi}(\mathbf{w}^*) = X\mathbf{J}(\mathbf{w}^*)$$

is positive stable by assumption, namely has eigenvalues $a_k + ib_k$ with $a_k > 0$. Writing $F(\mathbf{w}) = \mathbf{w} - \alpha X\boldsymbol{\xi}(\mathbf{w})$ for the iterative procedure given by SGA, it follows that

$$\nabla F(\mathbf{w}^*) = \mathbf{I} - \alpha\nabla[X\boldsymbol{\xi}](\mathbf{w}^*)$$

has eigenvalues $1 - \alpha a_k - i\alpha b_k$, which are in the unit circle for small α. More precisely,

$$\begin{aligned} |1 - \alpha a_k - i\alpha b_k|^2 < 1 \quad &\iff \quad 1 - 2\alpha a_k + \alpha^2 a_k^2 + \alpha^2 b_k^2 < 1 \\ &\iff \quad 0 < \alpha < \frac{2a_k}{a_k^2 + b_k^2} \end{aligned}$$

which is always possible for $a_k > 0$. Hence $\nabla F(\mathbf{w}^*)$ has eigenvalues in the unit circle for $0 < \alpha < \min_k 2a_k/(a_k^2 + b_k^2)$, and we are done by Ostrowski's Theorem since $\mathbf{w}^*$ is a fixed point of F. ■

Theorem 10 *Let* $\mathbf{w}^*$ *be a stable fixed point and* κ *the additive condition number of* $\mathbf{S}(\mathbf{w}^*)$*. Then SGA converges locally to* $\mathbf{w}^*$ *for all* $\lambda \in (0, \frac{4}{\kappa})$ *and* $\alpha > 0$ *sufficiently small.*

Proof By Theorem 5 and the assumption that $\mathbf{w}^*$ is a stable fixed point with invertible Jacobian, we know that

$$\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle = \langle (\mathbf{I} + \lambda\mathbf{A}^\intercal)\boldsymbol{\xi}, \mathbf{J}^\intercal\boldsymbol{\xi} \rangle > 0$$

for $\lambda \in (0, \frac{4}{\kappa})$. The proof does not rely on any particular property of $\boldsymbol{\xi}$, and can trivially be extended to the claim that

$$\langle (\mathbf{I} + \lambda\mathbf{A}^\intercal)\mathbf{u}, \mathbf{J}^\intercal\mathbf{u} \rangle > 0$$

for all non-zero vectors $\mathbf{u}$. In particular this can be rewritten as

$$\mathbf{u}^\intercal\mathbf{J}(\mathbf{I} + \lambda\mathbf{A}^\intercal)\mathbf{u}\rangle > 0\,,$$

which implies positive definiteness of $\mathbf{J}(\mathbf{I} + \lambda\mathbf{A}^\intercal)$. A positive definite matrix is positive stable, and any matrices AB and BA have identical spectrum. This implies also that $(\mathbf{I} + \lambda\mathbf{A}^\intercal)\mathbf{J}$ is positive stable, and we are done by the corollary above. ■

We conclude that SGA converges to an SFP if λ is small enough, where 'small enough' depends on the additive condition number.

3.5. Picking $\text{sign}(\lambda)$

This section explains desiderata $D4$—$D5$ and shows how to pick $\text{sign}(\lambda)$ to speed up convergence towards stable and away from unstable fixed points. In the example below, almost any choice of positive λ results in convergence to an unstable equilibrium. The problem arises from the combination of a weak repellor with a strong rotational force.

Example 9 (failure case for $\lambda > 0$)
Suppose $\epsilon > 0$ is small and

$$\ell_1(x, y) = -\frac{\epsilon}{2}x^2 - xy \;\; \textit{and} \;\; \ell_2(x, y) = -\frac{\epsilon}{2}y^2 + xy$$

with an unstable equilibrium at $(0, 0)$. The dynamics are

$$\boldsymbol{\xi} = \epsilon \cdot \begin{pmatrix} -x \\ -y \end{pmatrix} + \begin{pmatrix} -y \\ x \end{pmatrix} \quad \textit{with} \quad \mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and

$$\mathbf{A}^\intercal \boldsymbol{\xi} = \begin{pmatrix} x \\ y \end{pmatrix} + \epsilon \begin{pmatrix} -y \\ x \end{pmatrix}$$

Finally observe that

$$\boldsymbol{\xi} + \lambda \cdot \mathbf{A}^\intercal \boldsymbol{\xi} = (\lambda - \epsilon) \cdot \begin{pmatrix} x \\ y \end{pmatrix} + (1 + \epsilon\lambda) \cdot \begin{pmatrix} -y \\ x \end{pmatrix}$$

which converges to the unstable equilibrium if $\lambda > \epsilon$.

We now show how to pick the sign of λ to avoid unstable equilibria. First, observe that $\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle = \boldsymbol{\xi}^\intercal(\mathbf{S} + \mathbf{A})^\intercal \boldsymbol{\xi} = \boldsymbol{\xi}^\intercal \mathbf{S} \boldsymbol{\xi}$. It follows that for $\boldsymbol{\xi} \neq 0$:

$$\begin{cases} \text{if} \quad \mathbf{S} \succeq 0 \quad & \text{then } \langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle \geq 0; \\ \text{if} \quad \mathbf{S} \prec 0 \quad & \text{then } \langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle < 0. \end{cases} \tag{2}$$

A criterion to probe the positive/negative definiteness of $\mathbf{S}$ is thus to check the sign of $\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle$. The dot product can take any value if $\mathbf{S}$ is neither positive nor negative (semi-)definite. The behavior near saddle points will be explored in Section 3.7.

Recall that desiderata $D4$ requires that, if $\boldsymbol{\xi}$ points at a stable equilibrium then we require that $\boldsymbol{\xi}_\lambda$ points *more* towards the equilibrium (i.e. has smaller angle). Conversely, desiderata $D5$ requires that, if $\boldsymbol{\xi}$ points away then the adjustment should point *further* away. More formally,

Definition 6 *Let $\mathbf{u}$ and $\mathbf{v}$ be two vectors. The **infinitesimal alignment** of $\boldsymbol{\xi}_\lambda := \mathbf{u} + \lambda \cdot \mathbf{v}$ with a third vector $\mathbf{w}$ is*

$$\text{align}(\boldsymbol{\xi}_\lambda, \mathbf{w}) := \frac{d}{d\lambda} \left\{ \cos^2 \theta_\lambda \right\}_{|\lambda=0} \;\; \textit{for} \;\; \theta_\lambda := \theta(\boldsymbol{\xi}_\lambda, \mathbf{w}).$$

If $\mathbf{u}$ and $\mathbf{w}$ point the same way, $\mathbf{u}^\intercal \mathbf{w} > 0$, then align > 0 when $\mathbf{v}$ bends $\mathbf{u}$ further toward $\mathbf{w}$, see Figure 2A. Otherwise align > 0 when $\mathbf{v}$ bends $\mathbf{u}$ away from $\mathbf{w}$, see Figure 2B.

The following lemma allows us to rewrite the infinitesimal alignment in terms of known (computable) quantities, from which we can deduce the correct choice of λ.

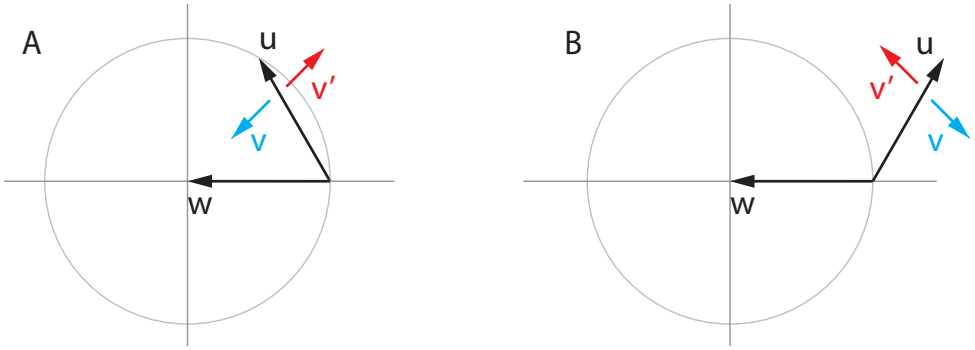

Figure 2: *Infinitesimal alignment* between $\mathbf{u} + \lambda\mathbf{v}$ and $\mathbf{w}$ is positive (cyan) when small positive λ either: **(A)** pulls $\mathbf{u}$ toward $\mathbf{w}$, if $\mathbf{w}$ and $\mathbf{u}$ have angle $< 90°$; or **(B)** pushes $\mathbf{u}$ away from $\mathbf{w}$ if their angle is $> 90°$. Conversely, the infinitesimal alignment is negative (red) when small positive λ either: **(A)** pushes $\mathbf{u}$ away from $\mathbf{w}$ when their angle is acute or **(B)** pulls $\mathbf{u}$ toward $\mathbf{w}$ when their angle is obtuse.

Algorithm 1 Symplectic Gradient Adjustment

Input: losses $\mathcal{L} = \{\ell_i\}_{i=1}^n$, weights $\mathcal{W} = \{\mathbf{w}_i\}_{i=1}^n$
$\boldsymbol{\xi} \leftarrow \big[\texttt{gradient}(\ell_i, \mathbf{w}_i) \ \texttt{for} \ (\ell_i, \mathbf{w}_i) \in (\mathcal{L}, \mathcal{W})\big]$
$\mathbf{A}^\intercal\boldsymbol{\xi} \leftarrow \texttt{get\_sym\_adj}(\mathcal{L}, \mathcal{W})$ // appendix A
if `align` **then**
  $\nabla\mathcal{H} \leftarrow \big[\texttt{gradient}(\frac{1}{2}\|\boldsymbol{\xi}\|^2, \mathbf{w}) \ \texttt{for} \ \mathbf{w} \in \mathcal{W})\big]$
  $\lambda \leftarrow \texttt{sign}\Big(\frac{1}{d}\langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle\langle\mathbf{A}^\intercal\boldsymbol{\xi}, \nabla\mathcal{H}\rangle + \epsilon\Big)$ // $\epsilon = \frac{1}{10}$
else
  $\lambda \leftarrow 1$
end if
Output: $\boldsymbol{\xi} + \lambda \cdot \mathbf{A}^\intercal\boldsymbol{\xi}$ // plug into any optimizer

Lemma 11 *When $\boldsymbol{\xi}_\lambda$ is the symplectic gradient adjustment,*

$$\text{sign}\Big(\text{align}(\boldsymbol{\xi}_\lambda, \nabla\mathcal{H})\Big) = \text{sign}\Big(\langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle \cdot \langle\mathbf{A}^\intercal\boldsymbol{\xi}, \nabla\mathcal{H}\rangle\Big).$$

Proof Observe that

$$\cos^2\theta_\lambda = \left(\frac{\langle\boldsymbol{\xi}_\lambda, \nabla\mathcal{H}\rangle}{\|\boldsymbol{\xi}_\lambda\| \cdot \|\nabla\mathcal{H}\|}\right)^2 = \frac{\langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle + 2\lambda\langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle\langle\mathbf{A}^\intercal\boldsymbol{\xi}, \nabla\mathcal{H}\rangle + O(\lambda^2)}{\big(\|\boldsymbol{\xi}\|^2 + O(\lambda^2)\big) \cdot \|\nabla\mathcal{H}\|^2}$$

where the denominator has no linear term in λ because $\boldsymbol{\xi} \perp \mathbf{A}^\intercal\boldsymbol{\xi}$. It follows that the sign of the infinitesimal alignment is

$$\text{sign}\left\{\frac{d}{d\lambda}\cos^2\theta_\lambda\right\} = \text{sign}\Big\{\langle\boldsymbol{\xi}, \nabla\mathcal{H}\rangle\langle\mathbf{A}^\intercal\boldsymbol{\xi}, \nabla\mathcal{H}\rangle\Big\}$$

as required. ■

Intuitively, computing the sign of $\langle \boldsymbol{\xi}, \nabla \mathcal{H} \rangle$ provides a check for stable and unstable fixed points. Computing the sign of $\langle \mathbf{A}^\intercal \boldsymbol{\xi}, \nabla \mathcal{H} \rangle$ checks whether the adjustment term points towards or away from the nearby fixed point. Putting the two checks together yields a prescription for the sign of λ, as follows.

Proposition 12 *Desiderata D4—D5 are satisfied for λ such that $\lambda \cdot \langle \boldsymbol{\xi}, \nabla \mathcal{H} \rangle \cdot \langle \mathbf{A}^\intercal \boldsymbol{\xi}, \nabla \mathcal{H} \rangle \geq 0$.*

Proof If we are in a neighborhood of a stable fixed point then $\langle \boldsymbol{\xi}, \nabla \mathcal{H} \rangle \geq 0$. It follows by Lemma 11 that $\text{sign}\left(\text{align}(\boldsymbol{\xi}_\lambda), \nabla \mathcal{H})\right) = \text{sign}\left(\langle \mathbf{A}^\intercal \boldsymbol{\xi}, \nabla \mathcal{H} \rangle\right)$ and so choosing $\text{sign}(\lambda) = \text{sign}\left(\langle \mathbf{A}^\intercal \boldsymbol{\xi}, \nabla \mathcal{H} \rangle\right)$ leads to the angle between $\boldsymbol{\xi}_\lambda$ and $\nabla \mathcal{H}$ being smaller than the angle between $\boldsymbol{\xi}$ and $\nabla \mathcal{H}$, satisfying desideratum $D4$. The proof for the unstable case is similar. ■

3.5.1. Alignment and Convergence Rates

Gradient descent is also known as the method of steepest descent. In general games, however, $\boldsymbol{\xi}$ does not follow the steepest path to fixed points due to the 'rotational force', which forces lower learning rates and slows down convergence.

The following lemma provides some intuition about alignment. The idea is that, the smaller the cosine between the 'correct direction' $\mathbf{w}$ and the 'update direction' $\boldsymbol{\xi}$, the smaller the learning rate needs to be for the update to stay in a unit ball, see Figure 3.

Lemma 13 (Alignment Lemma)
If $\mathbf{w}$ and $\boldsymbol{\xi}$ are unit vectors with $0 < \mathbf{w}^\intercal \boldsymbol{\xi}$ then $\|\mathbf{w} - \eta \cdot \boldsymbol{\xi}\| \leq 1$ for $0 \leq \eta \leq 2\mathbf{w}^\intercal \boldsymbol{\xi} = 2\cos\theta(\mathbf{w}, \boldsymbol{\xi})$. In other words, ensuring that $\mathbf{w} - \eta\boldsymbol{\xi}$ is closer to the origin than $\mathbf{w}$ requires smaller learning rates η as the angle between $\mathbf{w}$ and $\boldsymbol{\xi}$ gets larger.

Proof Check $\|\mathbf{w} - \eta \cdot \boldsymbol{\xi}\|^2 = 1 + \eta^2 - 2\eta \cdot \mathbf{w}^\intercal \boldsymbol{\xi} \leq 1$ iff $\eta^2 \leq 2\eta \cdot \mathbf{w}^\intercal \boldsymbol{\xi}$. The result follows. ■

The next lemma is a standard technical result from the convex optimization literature.

Lemma 14 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex Lipschitz smooth function satisfying $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L \cdot \|\mathbf{y} - \mathbf{x}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Then*

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{L}{2} \cdot \|\mathbf{y} - \mathbf{x}\|^2$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Proof See Nesterov (2004). ■

Finally, we show that increasing alignment helps speed convergence:

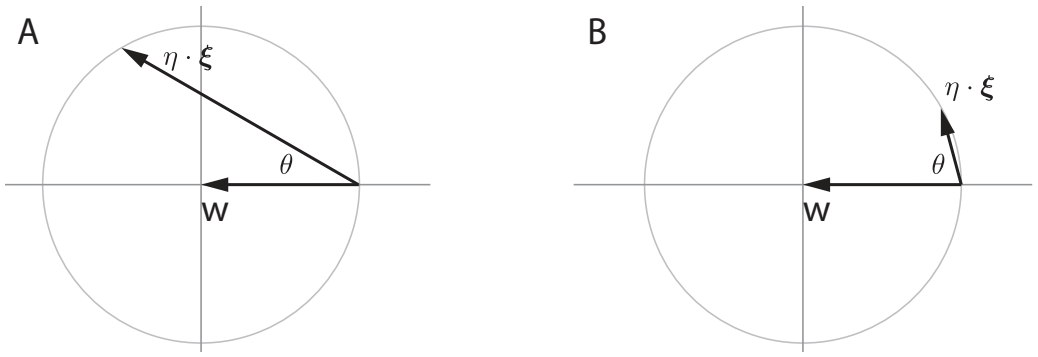

Figure 3: *Alignment and learning rates.* The larger $\cos\theta$, the larger the learning rate η that can be applied to unit vector $\boldsymbol{\xi}$ without $\mathbf{w} + \eta \cdot \boldsymbol{\xi}$ leaving the unit circle.

Theorem 15 *Suppose f is convex and Lipschitz smooth with $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \cdot \|\mathbf{x} - \mathbf{y}\|$. Let $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \mathbf{v}$ where $\|\mathbf{v}\| = \|\nabla f(\mathbf{w}_t)\|$. Then the optimal step size is $\eta^* = \frac{\cos\theta}{L}$ where $\theta := \theta(\nabla f(\mathbf{w}_t), \mathbf{v})$, with*

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) - \frac{\cos^2\theta}{2L} \cdot \|\nabla f(\mathbf{w}_t)\|^2.$$

The proof of Theorem 15 adapts Lemma 14 to handle the angle arising from the 'rotational force'.

Proof By the Lemma 14,

$$\begin{aligned} f(\mathbf{y}) &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\ &= f(\mathbf{x}) - \eta \cdot \langle \nabla f, \boldsymbol{\xi} \rangle + \eta^2 \frac{L}{2} \cdot \|\boldsymbol{\xi}\|^2 \\ &= f(\mathbf{x}) - \eta \cdot \langle \nabla f, \boldsymbol{\xi} \rangle + \eta^2 \frac{L}{2} \cdot \|\nabla f\|^2 \\ &= f(\mathbf{x}) - \eta(\alpha - \frac{\eta}{2}L) \cdot \|\nabla f\|^2 \end{aligned}$$

where $\alpha := \cos\theta$. Solve

$$\min_\eta \Delta(\eta) = \min_\eta \left\{ -\eta(\alpha - \frac{\eta}{2}L) \right\}$$

to obtain $\eta^* = \frac{\alpha}{L}$ and $\Delta(\eta^*) = -\frac{\alpha^2}{2}L$ as required. ■

Increasing the cosine with the steepest direction improves convergence. The alignment computation in algorithm 1 chooses λ to be positive or negative such that $\boldsymbol{\xi}_\lambda$ is bent towards stable (increasing the cosine) and away from unstable fixed points. Adding a small $\epsilon > 0$ to the computation introduces a weak bias towards stable fixed points.

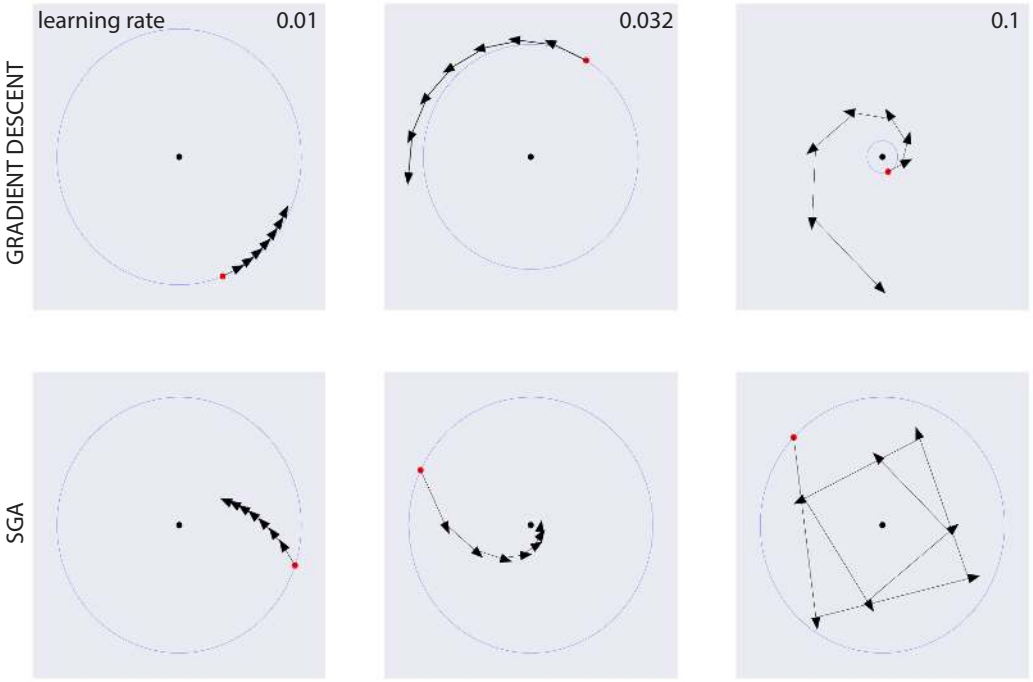

Figure 4: SGA allows faster and more robust convergence to stable fixed points than vanilla gradient descent in the presence of 'rotational forces', by bending the direction of descent towards the fixed point. Note the gradient descent diverges extremely rapidly in the top-right panel, which has a different scale from the other panels.

3.6. Aligned Consensus Optimization

The stability criterion in Equation (2) also provides a simple way to prevent consensus optimization from converging to unstable equilibria. **Aligned consensus optimization** is

$$\boldsymbol{\xi} + |\lambda| \cdot \operatorname{sign}\Big(\langle \boldsymbol{\xi}, \nabla \mathcal{H} \rangle\Big) \cdot \mathbf{J}^{\intercal} \boldsymbol{\xi}, \tag{3}$$

where in practice we set $\lambda = 1$. Aligned consensus optimization satisfies desiderata *D3—D5*. However, it behaves strangely in potential games. Multiplying by the Jacobian is the 'inverse' of Newton's method since for potential games the Jacobian of $\boldsymbol{\xi}$ is the Hessian of the potential function. Multiplying by the Hessian increases the gap between small and large eigenvalues, increasing the (usual, multiplicative) condition number and slows down convergence. Nevertheless, consensus optimization works well in GANs (Mescheder et al., 2017), and aligned consensus may improve performance, see experiments below.

Dropping the first term $\boldsymbol{\xi}$ from Equation (3) yields a simpler update that also satisfies *D*3—*D*5. However, the resulting algorithm performs poorly in experiments (not shown), perhaps because it is attracted to saddles.

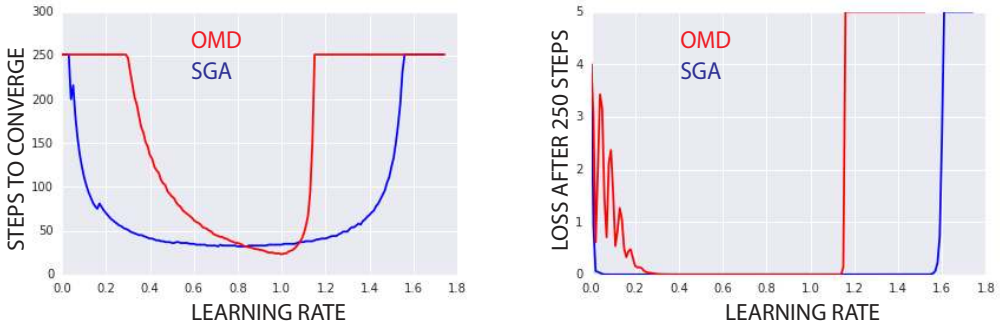

Figure 5: Comparison of SGA with optimistic mirror descent. The plots sweep over learning rates in range $[0.01, 1.75]$, with $\lambda = 1$ throughout for SGA. **(Left):** iterations to convergence, with maximum value of 250 after which the run was interrupted. **(Right):** average absolute value of losses over the last 10 iterations, 240-250, with a cutoff at 5.

3.7. Avoiding Strict Saddles

How does SGA behave near saddles? We show that Symplectic Gradient Adjustment locally avoids strict saddles, provided that λ and α are small and parameters are initialized with (arbitrarily small) noise. More precisely, let $\mathbf{F}(\mathbf{w}) = \mathbf{w} - \alpha\boldsymbol{\xi}_\lambda(\mathbf{w})$ be the iterative optimization procedure given by SGA. Then every strict saddle $\mathbf{w}^*$ has a neighbourhood U such that $\{\mathbf{w} \in U \mid \mathbf{F}^n(\mathbf{w}) \to \mathbf{w}^* \text{ as } n \to \infty\}$ has measure zero for small $\alpha > 0$ and λ.

Intuitively, the Taylor expansion around a strict saddle $\mathbf{w}^*$ is locally dominated by the Jacobian at $\mathbf{w}^*$, which has a negative eigenvalue. This prevents convergence to $\mathbf{w}^*$ for random initializations of $\mathbf{w}$ near $\mathbf{w}^*$. The argument is made rigorous using the Stable Manifold Theorem following Lee et al. (2017).

Theorem 16 (Stable Manifold Theorem)
Let $\mathbf{w}^*$ be a fixed point for the C^1 local diffeomorphism $F : U \to \mathbb{R}^d$, where U is a neighbourhood of $\mathbf{w}^*$ in $\mathbb{R}^d$. Let $E^s \oplus E^u$ be the generalized eigenspaces of $\nabla F(\mathbf{w}^*)$ corresponding to eigenvalues with $|\sigma| \leq 1$ and $|\sigma| > 1$ respectively. Then there exists a local stable center manifold W with tangent space E^s at $\mathbf{w}^*$ and a neighbourhood B of $\mathbf{w}^*$ such that $F(W) \cap B \subset W$ and $\cap_{n=0}^{\infty} F^{-n}(B) \subset W$.

Proof See Shub (2000). ■

It follows that if $\nabla F(\mathbf{w}^*)$ has at least one eigenvalue $|\sigma| > 1$ then E^u has dimension at least 1. Since W has tangent space E^s at $\mathbf{w}^*$ with codimension at least one, we conclude that W has measure zero. This is central to proving that the set of nearby initial points which converge to a given strict saddle $\mathbf{w}^*$ has measure zero. Since $\mathbf{w}$ is initialized randomly, the following theorem is obtained.

Theorem 17 *SGA locally avoids strict saddles almost surely, for $\alpha > 0$ and λ small.*

Proof Let $\mathbf{w}^*$ a strict saddle and recall that SGA is given by

$$F(\mathbf{w}) = \mathbf{w} - \alpha(\mathbf{I} - \alpha\mathbf{J})\boldsymbol{\xi}(\mathbf{w})\,.$$

All terms involved are continuously differentiable and we have

$$\nabla F(\mathbf{w}^*) = \mathbf{I} - \alpha(\mathbf{I} - \alpha\mathbf{J})\mathbf{J}(\mathbf{w}^*)$$

by assumption that $\boldsymbol{\xi}(\mathbf{w}^*) = 0$. Since all terms except $\mathbf{I}$ are of order at least α, $\nabla F(\mathbf{w}^*)$ is invertible for all α sufficiently small. By the inverse function theorem, there exists a neighbourhood U of $\mathbf{w}^*$ such that F is has a continuously differentiable inverse on U. Hence F restricted to U is a C^1 diffeomorphism with fixed point $\mathbf{w}^*$.

By definition of strict saddles, $\mathbf{J}(\mathbf{w}^*)$ has an eigenvalue with negative real part. It follows by continuity that $(\mathbf{I} - \alpha\mathbf{J})\mathbf{J}(\mathbf{w}^*)$ also has an eigenvalue $a + ib$ with $a < 0$ for α sufficiently small. Finally,

$$\nabla F(\mathbf{w}^*) = \mathbf{I} - \alpha(\mathbf{I} - \alpha\mathbf{J})\mathbf{J}(\mathbf{w}^*)$$

has an eigenvalue $\sigma = 1 - \alpha a - i\alpha b$ with

$$|\sigma| = 1 - 2\alpha a + \alpha^2(a^2 + b^2) \geq 1 - 2\alpha a > 1\,.$$

It follows that E^s has codimension at least one, implying in turn that the local stable set W has measure zero. We can now prove that

$$Z = \{\mathbf{w} \in U \mid \lim_{n\to\infty} F^n(\mathbf{w}) = \mathbf{w}^*\}$$

has measure zero, or in other words, that local convergence to $\mathbf{w}^*$ occurs with zero probability. Let B the neighbourhood guaranteed by the Stable Manifold Theorem, and take any $\mathbf{w} \in Z$. By definition of convergence there exists $N \in \mathbb{N}$ such that $F^{N+n}(\mathbf{w}) \in B$ for all $n \in \mathbb{N}$, so that

$$F^N(\mathbf{w}) \in \cap_{n\in\mathbb{N}}^{\infty} F^{-n}(B) \subset W$$

by the Stable Manifold Theorem. This implies that $\mathbf{w} \in F^{-N}(W)$, and by extension $\mathbf{w} \in \cup_{n\in\mathbb{N}} F^{-n}(W)$. Since $\mathbf{w}$ was arbitrary, we obtain the inclusion

$$Z \subseteq \cup_{n\in\mathbb{N}} F^{-n}(W)\,.$$

Now F^{-1} is C^1, hence locally Lipschitz and thus preserves sets of measure zero, so that $F^{-n}(W)$ has measure zero for each n. Countable unions of measure zero sets are still measure zero, so we conclude that Z also has measure zero. In other words, SGA converges to $\mathbf{w}^*$ with zero probability upon random initialization of $\mathbf{w}$ in U. ■

Unlike stable and unstable fixed points, it is unclear how to avoid strict saddles using only alignment, that is, independently from the size of λ.

4. Experiments

We compare SGA with simultaneous gradient descent, optimistic mirror descent (Daskalakis et al., 2018) and consensus optimization (Mescheder et al., 2017) in basic settings.

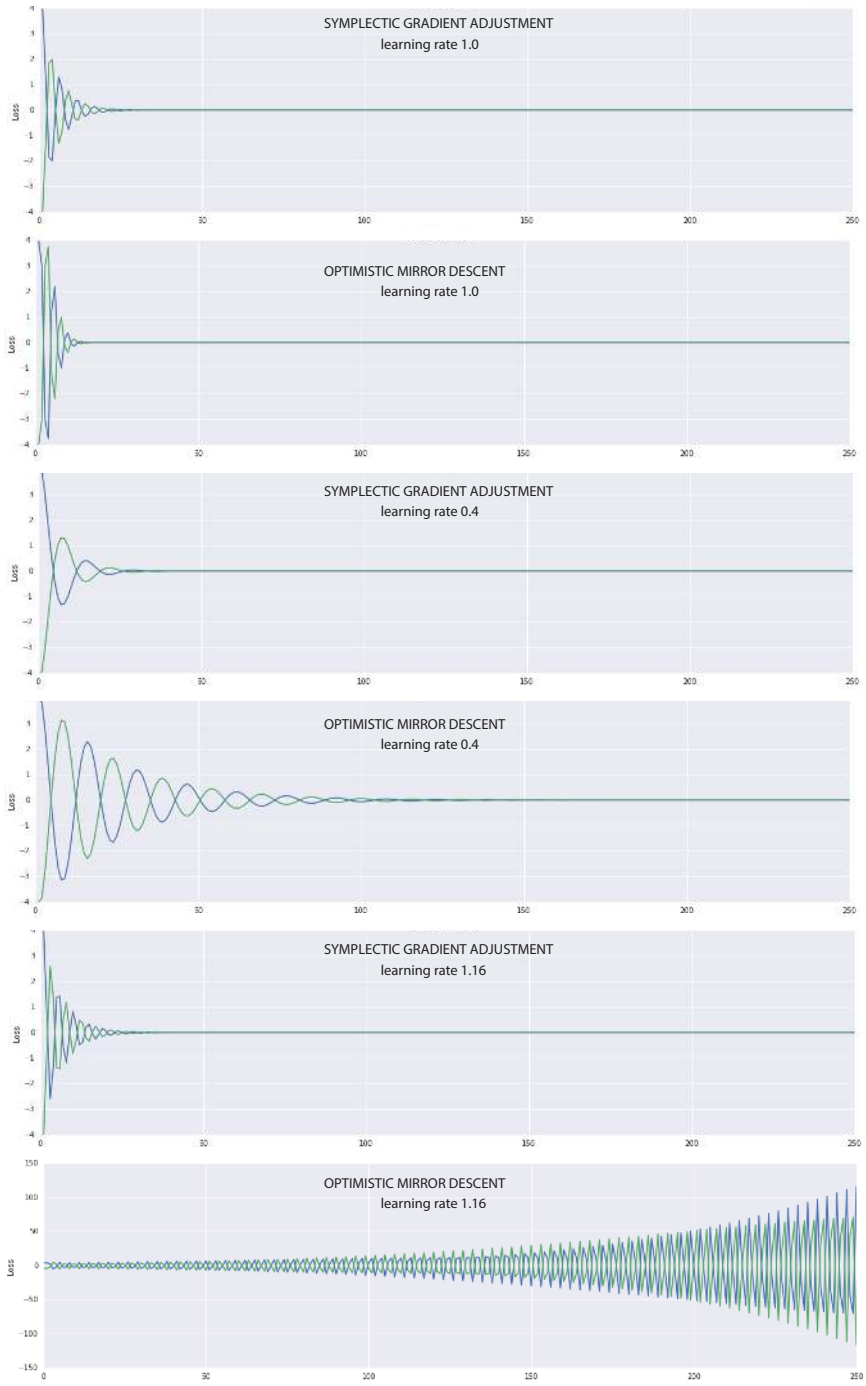

Figure 6: Individual runs on zero-sum bimatrix game in Section 4.2.

4.1. Learning rates and alignment

We investigate the effect of SGA when a weak attractor is coupled to a strong rotational force:

$$\ell_1(x, y) = \frac{1}{2}x^2 + 10xy \quad \text{and} \quad \ell_2(x, y) = \frac{1}{2}y^2 - 10xy$$

Gradient descent is extremely sensitive to the choice of learning rate η, top row of Figure 4. As η increases through $\{0.01, 0.032, 0.1\}$ gradient descent goes from converging extremely slowly, to diverging slowly, to diverging rapidly. SGA yields faster, more robust convergence. SGA converges faster with learning rates $\eta = 0.01$ and $\eta = 0.032$, and only starts overshooting the fixed point for $\eta = 0.1$.

4.2. Basic adversarial games

Optimistic mirror descent is a family of algorithms that has nice convergence properties in games (Rakhlin and Sridharan, 2013; Syrgkanis et al., 2015). In the special case of optimistic *gradient* descent the updates are

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \cdot \boldsymbol{\xi}_t - \eta \cdot (\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t-1}).$$

Figure 5 compares SGA with optimistic gradient descent (OMD) on a zero-sum bimatrix game with $\ell_{1/2}(\mathbf{w}_1, \mathbf{w}_2) = \pm \mathbf{w}_1^\intercal \mathbf{w}_2$. The example is modified from Daskalakis et al. (2018) who also consider a linear offset that makes no difference. A run is taken to have converged if the average absolute value of losses on the last 10 iterations is < 0.01; we end each experiment after 250 steps.

The left panel shows the number of steps to convergence (when convergence occurs) over a range of learning rates. OMD's peak performance is better than SGA, where the red curve dips below the blue. Howwever, we find that SGA converges—and does so faster—for a much wider range of learning rates. OMD diverges for learning rates not in the range [0.3, 1.2]. Simultaneous gradient descent oscillates without converging (not shown). The right panel shows the average performance of OMD and SGA on the last 10 steps. Once again, here SGA consistently performs better over a wider range of learning rates. Individual runs are shown in Figure 6.

4.2.1. OMD and SGA on a Four-Player Game

Figure 7 shows time to convergence (using the same convergence criterion as above) for optimistic mirror descent and SGA. The games are constructed with four players, each of which controls one parameter. The losses are

$$\ell_1(w, x, y, z) = \frac{\epsilon}{2}w^2 + wx + wy + wz$$
$$\ell_2(w, x, y, z) = -wx + \frac{\epsilon}{2}x^2 + xy + xz$$

$$\ell_3(w, x, y, z) = -wy - xy + \frac{\epsilon}{2}y^2 + yz$$
$$\ell_4(w, x, y, z) = -wz - xz - yz + \frac{\epsilon}{2}z^2,$$

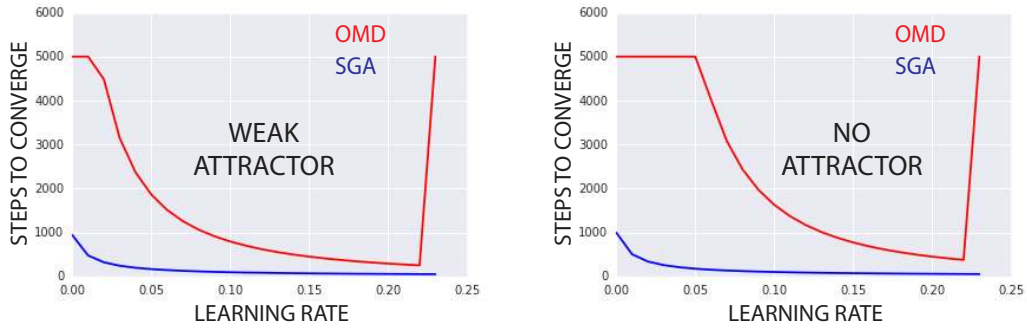

Figure 7: Time to convergence of OMD and SGA on two 4-player games. Times are cutoff after 5000 iterations. **Left panel:** Weakly positive definite $\mathbf{S}$ with $\epsilon = \frac{1}{100}$. **Right panel:** Symmetric component is identically zero.

where $\epsilon = \frac{1}{100}$ in the left panel and $\epsilon = 0$ in the right panel. The antisymmetric component of the game Jacobian is

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 \end{pmatrix}$$

and the symmetric component is

$$\mathbf{S} = \epsilon \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

OMD converges considerably slower than SGA across the full range of learning rates. It also diverges for learning rates > 0.22. In contrast, SGA converges more quickly and robustly.

4.3. Learning a two-dimensional mixture of Gaussians

We apply SGA to a basic Generative Adversarial Network setup adapted from Metz et al. (2017). Data is sampled from a highly multimodal distribution designed to probe the tendency of GANs to collapse onto a subset of modes during training. The distribution is a mixture of 16 Gaussians arranged in a 4×4 grid. Figure 8 shows the probability distribution that is sampled to train the generator and discriminator. The generator and discriminator networks both have 6 ReLU layers of 384 neurons. The generator has two output neurons; the discriminator has one.

Figure 9 shows results after $\{2000, 4000, 6000, 8000\}$ iterations. The networks are trained under RMSProp. Learning rates were chosen by visual inspection of grid search results at iteration 8000. More precisely, grid search was over learning rates {1e-5, 2e-5,5e-5, 8e-5, 1e-4, 2e-4, 5e-4} and then a more refined linear search over [8e-5, 2e-4]. Simultaneous gradient descent and SGA are shown in the figure.

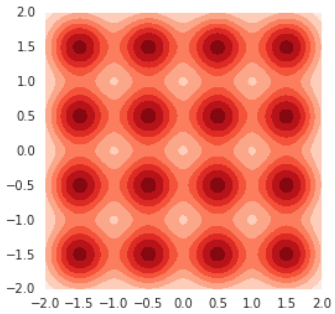

Figure 8: Ground truth for GAN experiments on a two-dimensional mixture of 16 Gaussians.

The last two rows of Figure 9 show the performance of consensus optimization without and with alignment. Introducing alignment slightly improves speed of convergence (second column) and final result (fourth column), although intermediate results in third column are ambiguous.

Simultaneous gradient descent exhibits mode collapse followed by mode hopping in later iterations (not shown). Mode hopping is analogous to the cycles in example 1. Unaligned SGA converges to the correct distribution; alignment speeds up convergence slightly. Consensus optimization performs similarly in this GAN example. However, consensus optimization can converge to local maxima even in potential games, recall example 8.

4.4. Learning a high-dimensional unimodal Gaussian

Mode collapse is a well-known phenomenon in GANs. A more subtle phenomenon, termed boundary distortion, was identified in Santurkar et al. (2018). Boundary distortion is a form of covariate shift where the generator fails to model the true data distribution.

Santurkar *et al* demonstrate boundary distortion using data sampled from a 75-dimensional unimodal Gaussian with spherical covariate matrix. Mode collapse is not a problem in this setting because the data distribution is unimodal. Nevertheless, they show that vanilla GANs fail to learn most of the spectrum of the covariate matrix.

Figure 10 reproduces their result. Panel A shows the ground truth: all 75 eigenvalues are equal to 1.0. Panel B shows the spectrum of the covariance matrix of the data generated by a GAN trained with RMSProp. The GAN concentrates on a single eigenvalue and essentially ignores the remaining 74 eigenvalues. This is similar to, but more extreme than, the empirical results obtained in Santurkar et al. (2018). We emphasize that the problem is not mode collapse, since the data is unimodal (although, it's worth noting that most of the mass of a high-dimensional Gaussian lies on the "shell").

Finally, panel C shows the spectrum of the covariance matrix of the data sampled from a GAN trained via SGA. The GAN approximately learns all the eigenvalues, with values ranging between 0.6 and 1.5.

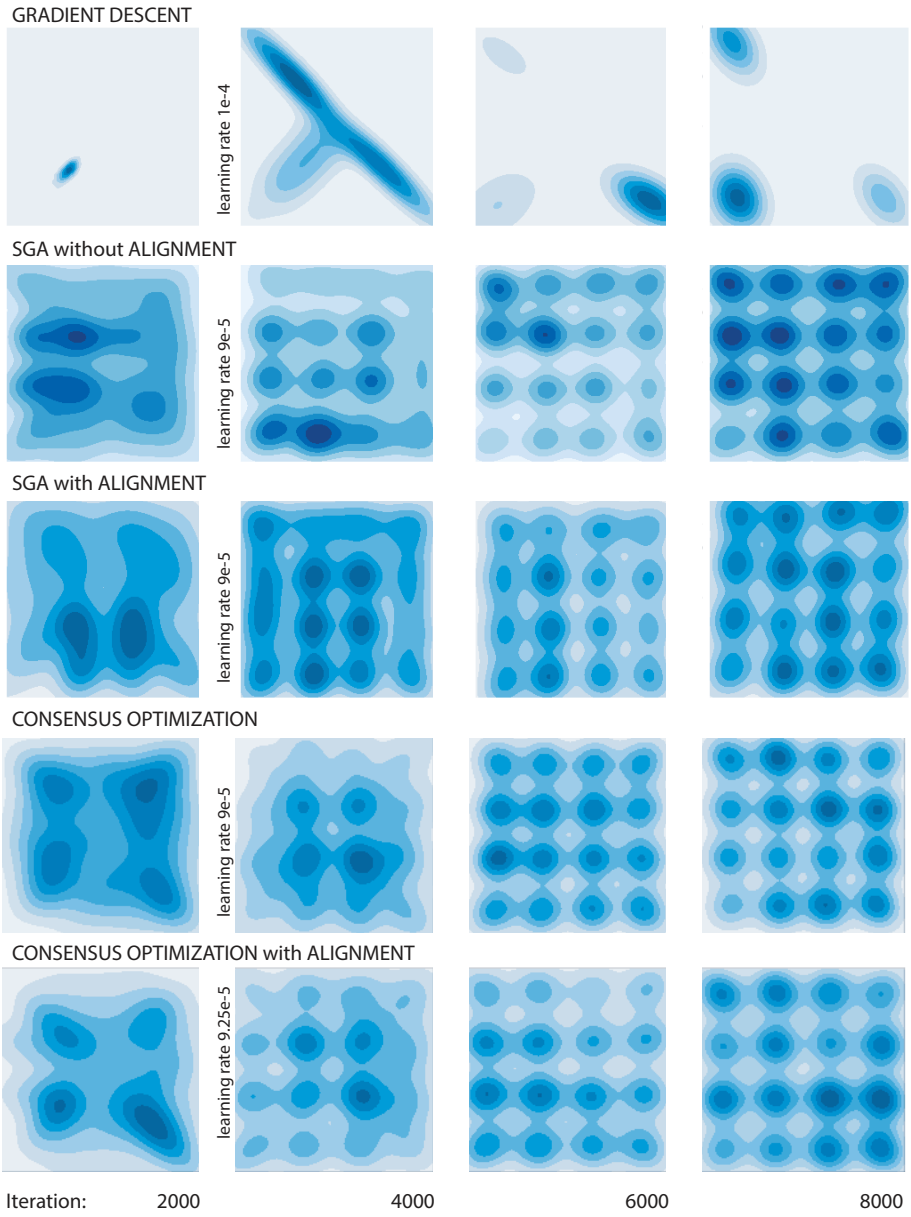

Figure 9: **First row:** Simultaneous gradient descent suffers from mode collapse and in later iterations (not shown) mode hopping. **Second and third rows:** vanilla SGA converges smoothly to the ground truth (Figure 8). SGA with alignment converges slightly faster. **Fourth and fifth rows:** Consensus optimization without and with alignment.

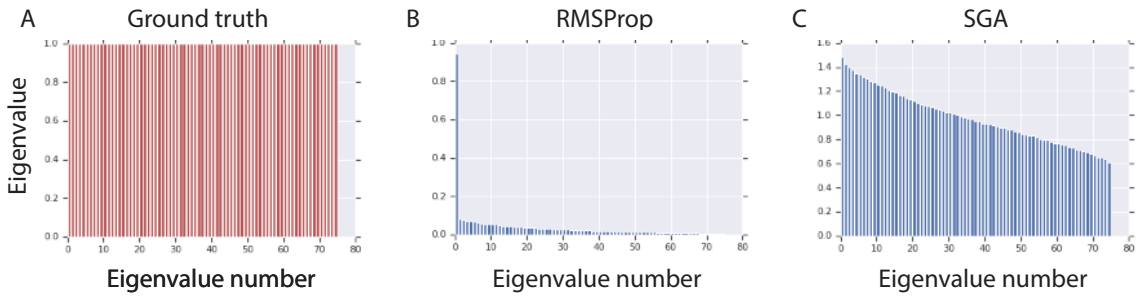

Figure 10: **Panel A:** The ground truth is a 75 dimensional spherical Gaussian whose covariance matrix has all eigenvalues equal to 1.0. **Panel B:** A vanilla GAN trained with RMSProp approximately learns the first eigenvalue, but essentially ignores all the rest. **Panel C:** Applying SGA results in the GAN approximately learning all 75 eigenvalues, although the range varies from 0.6 to 1.5.

5. Discussion

Modern deep learning treats differentiable modules like plug-and-play lego blocks. For this to work, at the very least, we need to know that gradient descent will find local minima. Unfortunately, gradient descent does *not* necessarily find local minima when optimizing multiple interacting objectives. With the recent proliferation of algorithms that optimize more than one loss, it is becoming increasingly urgent to understand and control the dynamics of interacting losses. Although there is interesting recent work on two-player adversarial games such as GANs, there is essentially no work on finding stable fixed points in more general games played by interacting neural nets.

The generalized Helmholtz decomposition provides a powerful new perspective on game dynamics. A key feature is that the analysis is indifferent to the number of players. Instead, it is the interplay between the simultaneous gradient $\boldsymbol{\xi}$ on the losses and the symmetric and antisymmetric matrices of second-order terms that guides algorithm design and governs the dynamics under gradient adjustments.

Symplectic gradient adjustment is a straightforward application of the generalized Helmholtz decomposition. It is unlikely that SGA is the best approach to finding stable fixed points. A deeper understanding of the interaction between the potential and Hamiltonian components will lead to more effective algorithms. Reinforcement learning algorithms that optimize multiple objectives are increasingly common, and second-order terms are difficult to estimate in practice. Thus, first-order methods that do not use Jacobian-vector products are of particular interest.

5.0.1. Gamification

Finally, it is worth raising a philosophical point. In this paper we are concerned with finding stable fixed points (because, for example, they yield pleasing samples in GANs). We are not concerned with the losses of the players *per se*. The gradient adjustments may lead to a

player acting against its own self-interest by increasing its loss. We consider this acceptable insofar as it encourages convergence to a stable fixed point. The players are but a means to an end.

We have argued that stable fixed points are a more useful solution concept than local Nash equilibria for our purposes. However, neither is entirely satisfactory, and the question "What is the right solution concept for neural games?" remains open. In fact, it likely has many answers. The intrinsic curiosity module introduced by Pathak et al. (2017) plays two objectives against one another to drive agents to search for novel experiences. In this case, converging to a fixed point is precisely what is to be avoided.

It is remarkable—to give a few examples sampled from many—that curiosity, generating photorealistic images, and image-to-image translation (Zhu et al., 2017) can be formulated as games. What else can games do?

Acknowledgments

We thank Guillaume Desjardins and Csaba Szepesvari for useful comments.

Appendix A. TensorFlow Code to Compute SGA

Source code is available at `https://github.com/deepmind/symplectic-gradient-adjustment`. Since computing the symplectic adjustment is quite simple, we include an explicit description here for completeness.

The code requires a list of n losses, `Ls`, and a list of variables for the n players, `xs`. The function `fwd_gradients` which implements forward mode auto-differentiation is in the module `tf.contrib.kfac.utils`.

```
% compute Jacobian-vector product Jv
def jac_vec(ys, xs, vs):
    return fwd_gradients(ys, xs, grad_xs = vs, stop_gradients = xs)

% compute Jacobian^T-vector product J^T v
def jac_tran_vec(ys, xs, vs):
  dydxs = tf.gradients(ys, xs, grad_ys = vs, stop_gradients = xs)
  return [tf.zeros_like(x) if dydx is None
          else dydx for (x, dydx) in zip(xs, dydxs)]

% compute Symplectic Gradient Adjustment A^T ξ
def get_sym_adj(Ls, xs):
  % compute game dynamics ξ
  xi = [tf.gradients(ℓ, x)[0] for (ℓ, x) in zip(Ls, xs)]
  J_xi = jac_vec(xi, xs, xi)
  Jt_xi = jac_tran_vec(xi, xs, xi)
  % compute A^T ξ = 1/2 (J^T ξ − J ξ)
```

```
At_xi = [(jt−j)/2 for (j, jt) in zip(J_xi, Jt_xi)]
return At_xi
```

Appendix B. Helmholtz, Hamilton, Hodge, and Harmonic Games

This section explains the mathematical connections with the Helmholtz decomposition, symplectic geometry and the Hodge decomposition. The discussion is *not* necessary to understand the main text. It is also not self-contained. The details can be found in textbooks covering differential and symplectic geometry (Arnold, 1989; Guillemin and Sternberg, 1990; Bott and Tu, 1995).

B.1. The Helmholtz Decomposition

The classical Helmholtz decomposition states that any vector field $\boldsymbol{\xi}$ in 3-dimensions is the sum of curl-free (gradient) and divergence-free (infinitesimal rotation) components:

$$\boldsymbol{\xi} = \underbrace{\nabla\phi}_{\text{gradient component}} + \underbrace{\text{curl}(\mathbf{B})}_{\text{rotational component}} \qquad \Big[\ \text{curl}(\bullet) := \nabla \times (\bullet)\Big]$$

We explain the link between curl and the antisymmetric component of the game Jacobian. Recall that gradients of functions are actually differential 1-forms, not vector fields. Differential 1-forms and vector fields on a manifold are canonically isomorphic once a Riemannian metric has been chosen. In our case, we are implicitly using the Euclidean metric. The antisymmetric matrix $\mathbf{A}$ is the differential 2-form obtained by applying the exterior derivative d to the 1-form $\boldsymbol{\xi}$.

In 3-dimensions, the Hodge star operator is an isormorphism from differential 2-forms to vector fields, and the curl can be reformulated as $\text{curl}(\bullet) = *d(\bullet)$. In claiming $\mathbf{A}$ is analogous to curl, we are simply dropping the Hodge-star operator.

Finally, recall that the Lie algebra of infinitesimal rotations in d-dimensions is given by antisymmetric matrices. When $d = 3$, the Lie algebra can be represented as vectors (three numbers specify a 3×3 antisymmetric matrix) with the $\times$-product as Lie bracket. In general, the antisymmetric matrix $\mathbf{A}$ captures the infinitesimal tendency of $\boldsymbol{\xi}$ to rotate at each point in the parameter space.

B.2. Hamiltonian Mechanics

A symplectic form ω is a closed nondegenerate differential 2-form. Given a manifold with a symplectic form, a vector field $\boldsymbol{\xi}$ is **Hamiltonian vector field** if there exists a function $\mathcal{H} : M \to \mathbb{R}$ satisfying

$$\omega(\boldsymbol{\xi}, \bullet) = d\mathcal{H}(\bullet) = \langle \nabla\mathcal{H}, \bullet \rangle. \tag{4}$$

The function is then referred to as the Hamiltonian function of the vector field. In our case, the antisymmetric matrix $\mathbf{A}$ is a closed 2-form because $\mathbf{A} = d\boldsymbol{\xi}$ and $d \circ d = 0$. It may however be degenerate. It is therefore a presymplectic form (Bottacin, 2005).

Setting $\omega = \mathbf{A}$, Equation (4) can be rewritten in our notation as

$$\underbrace{\omega(\boldsymbol{\xi}, \bullet)}_{\mathbf{A}^\intercal\boldsymbol{\xi}} = \underbrace{d\mathcal{H}(\bullet)}_{\nabla\mathcal{H}},$$

justifying the terminology 'Hamiltonian'.

B.3. The Hodge Decomposition

The exterior derivative $d_k : \Omega^k(M) \to \Omega^{k+1}(M)$ is a linear operator that takes differential k-forms on a manifold M, $\Omega^k(M)$, to differential $k+1$-forms, $\Omega^{k+1}(M)$. In the case $k = 0$, the exterior derivative is the gradient, which takes 0-forms (that is, functions) to 1-forms. Given a Riemannian metric, the adjoint of the exterior derivative δ goes in the opposite direction. Hodge's Theorem states that k-forms on a compact manifold decompose into a direct sum over three types:

$$\Omega^k(M) = d\Omega^{k-1}(M) \oplus \text{Harmonic}^k(M) \oplus \delta\Omega^{k+1}(M).$$

Setting $k = 1$, we recover a decomposition that closely resembles the generalized Helmholtz decomposition:

$$\underbrace{\Omega^1(M)}_{\text{1-forms}} = \underbrace{d\Omega^0(M)}_{\text{gradients of functions}} \oplus \text{Harm}^k(M) \oplus \underbrace{\delta\Omega^2(M)}_{\text{antisymmetric component}}$$

The harmonic component is isomorphic to the de Rham cohomology of the manifold—which is zero when $k = 1$ and $M = \mathbb{R}^n$.

Unfortunately, the Hodge decomposition does not straightforwardly apply to the case when $M = \mathbb{R}^n$, since $\mathbb{R}^n$ is not compact. It is thus unclear how to relate the generalized Helmholtz decomposition to the Hodge decomposition.

B.4. Harmonic and Potential Games

Candogan et al. (2011) derive a Hodge decomposition for games that is closely related in spirit to our generalized Helmholtz decomposition—although the details are quite different. Candogan et al. (2011) work with classical games (probability distributions on finite strategy sets). Their losses are multilinear, which is easier than our setting, but they have constrained solution sets, which is harder in many ways. Their approach is based on combinatorial Hodge theory (Jiang et al., 2011) rather than differential and symplectic geometry. Finding a best-of-both-worlds approach that encompasses both settings is an open problem.

Appendix C. Type Consistency

The next two sections carefully work through the units in classical mechanics and two-player games respectively. The third section briefly describes a use-case for type consistency.

C.1. Units in Classical Mechanics

Consider the well-known Hamiltonian

$$\mathcal{H}(p, q) = \frac{1}{2}\left(\kappa \cdot q^2 + \frac{1}{\mu} \cdot p^2\right)$$

where q is position, $p = \mu \cdot \dot{q}$ is momentum, μ is mass, κ is surface tension and $\mathcal{H}$ measures energy. The units (denoted by τ) are

$$\tau(q) = m \qquad \tau(p) = \tfrac{kg \cdot m}{s}$$

$$\tau(\kappa) = \tfrac{kg}{s^2} \qquad \tau(\mu) = kg$$

where m is meters, kg is kilograms and s is seconds. Energy is measured in joules, and indeed it is easy to check that $\tau(\mathcal{H}) = \frac{kg \cdot m^2}{s^2}$.

Note that the units for differentation by x are $\tau(\frac{\partial}{\partial x}) = \frac{1}{\tau(x)}$. For example, differentiating by time has units $\frac{1}{s}$. Hamilton's equations state that $\dot{q} = \frac{\partial H}{\partial p} = \frac{1}{\mu} \cdot p$ and $\dot{p} = -\frac{\partial \mathcal{H}}{\partial q} = -\kappa \cdot q$ where

$$\tau(\dot{q}) = \tfrac{m}{s} \qquad \tau(\dot{p}) = \tfrac{kg \cdot m}{s^2}$$

$$\tau\left(\tfrac{\partial}{\partial q}\right) = \tfrac{1}{m} \qquad \tau\left(\tfrac{\partial}{\partial p}\right) = \tfrac{s}{kg \cdot m}$$

The resulting flow describing the dynamics of the system is

$$\boldsymbol{\xi} = \dot{q} \cdot \frac{\partial}{\partial q} + \dot{p} \cdot \frac{\partial}{\partial p} = \frac{1}{\mu} p \cdot \frac{\partial}{\partial q} - \kappa q \cdot \frac{\partial}{\partial p}$$

with units $\tau(\boldsymbol{\xi}) = \frac{1}{s}$. Hamilton's equations can be reformulated more abstractly via symplectic geometry. Introduce the symplectic form

$$\omega = dq \wedge dp \quad \text{with units} \quad \tau(\omega) = \frac{kg \cdot m^2}{s}.$$

Observe that contracting the flow with the Hamiltonian obtains

$$\iota_{\boldsymbol{\xi}} \omega = \omega(\boldsymbol{\xi}, \bullet) = dH = \frac{\partial \mathcal{H}}{\partial q} \cdot dq + \frac{\partial \mathcal{H}}{\partial p} \cdot dp$$

with units $\tau(d\mathcal{H}) = \tau(\mathcal{H}) = \frac{kg \cdot m^2}{s^2}$.

C.1.1. Losses in Classical Mechanics

Although there is no notion of "loss" in classical mechanics, it is useful (for the next section) to keep pushing the formal analogy. Define the "losses"

$$\ell_1(q, p) = \frac{1}{\mu} \cdot qp \quad \text{and} \quad \ell_2(q, p) = -\kappa \cdot qp \tag{5}$$

with units $\tau(\ell_1) = \frac{m^2}{s}$ and $\tau(\ell_2) = \frac{kg^2 \cdot m^2}{s^3}$. The Hamiltonian dynamics can then be recovered game-theoretically by differentiating ℓ_1 and ℓ_2 with respect to q and p respectively. It is easy to check that

$$\boldsymbol{\xi} = \frac{\partial \mathcal{H}}{\partial p} \frac{\partial}{\partial q} - \frac{\partial \mathcal{H}}{\partial q} \frac{\partial}{\partial p} = \frac{\partial \ell_1}{\partial q} \frac{\partial}{\partial q} + \frac{\partial \ell_2}{\partial p} \frac{\partial}{\partial p}.$$

C.1.2. The Duality between Vector Fields and Differential Forms

Finally recall that the symplectic form in games was not "pulled out of thin air" as $\omega = dq \wedge dp$, but rather derived as $\omega = d\boldsymbol{\xi}^\flat$, where $\boldsymbol{\xi}^\flat$ is the differential form corresponding to the vector field $\boldsymbol{\xi}$ under the musical isomorphism $\flat : TM \to T^*M$.

It is instructive to compute $\boldsymbol{\xi}^\flat$ in the case of a classical mechanical system and see what happens. Naively, we would guess that the musical isomorphism is $\left(\frac{\partial}{\partial q}\right)^\flat = dq$ and $\left(\frac{\partial}{\partial p}\right)^\flat = dp$. However, applying the naive musical isomorphism to $\boldsymbol{\xi}$ to get

$$\boldsymbol{\xi}^\flat = \frac{\partial \ell_1}{\partial q} \cdot dq + \frac{\partial \ell_2}{\partial p} \cdot dp$$

results in a *type violation* because

$$\tau\left(\frac{\partial \ell_1}{\partial q} \cdot dq\right) = \tau(\ell_1) = \frac{m^2}{s}$$

whereas

$$\tau\left(\frac{\partial \ell_2}{\partial p} \cdot dp\right) = \tau(\ell_2) = \frac{kg^2 \cdot m^2}{s^3}$$

and we cannot add objects with different types.

To correct the type inconsistency, define the musical isomorphism as

$$\left(\frac{\partial}{\partial q}\right)^\flat = \frac{\mu}{2} \cdot dq \quad \text{and} \quad \left(\frac{\partial}{\partial p}\right)^\flat = \frac{1}{2\kappa} \cdot dp$$

with inverse

$$(dq)^\sharp = \frac{2}{\mu} \cdot \frac{\partial}{\partial q} \quad \text{and} \quad (dp)^\sharp = 2\kappa \cdot \frac{\partial}{\partial p}.$$

The correction terms in the direction $\flat : TM \to T^*M$ invert the coupling terms κ and $\frac{1}{\mu}$ that were originally introduced into the Hamiltonian for physical reasons. Applying the corrected musical isomorphism to $\boldsymbol{\xi}$ yields

$$\boldsymbol{\xi}^\flat = \frac{\mu}{2} \cdot \frac{\partial f}{\partial q} \cdot dq + \frac{1}{2\kappa} \cdot \frac{\partial g}{\partial p} \cdot dp = \frac{1}{2}\left(p \cdot dq - q \cdot dp\right).$$

The two terms of $\boldsymbol{\xi}^\flat$ then have coherent types

$$\begin{aligned} \tau\left(\tfrac{\partial \ell_1}{\partial q} \cdot \mu \cdot dq\right) &= \tfrac{m}{s} \cdot kg \cdot m = \tfrac{kg \cdot m^2}{s} \\ \tau\left(\tfrac{\partial \ell_2}{\partial p} \cdot \tfrac{1}{\kappa} \cdot dp\right) &= \tfrac{kg \cdot m}{s^2} \cdot \tfrac{s^2}{kg} \cdot \tfrac{kg \cdot m}{s} = \tfrac{kg \cdot m^2}{s} \end{aligned}$$

as required. The associated two form is

$$\omega := d\boldsymbol{\xi}^\flat = -\left(\mu \cdot \frac{\partial^2 f}{\partial q \partial p} - \frac{1}{\kappa} \cdot \frac{\partial^2 g}{\partial q \partial p}\right) dq \wedge dp = -dq \wedge dp$$

which recovers the symplectic form (up to sign) with units $\tau(\omega) = \frac{kg \cdot m^2}{s}$ as required. Finally, observe that

$$\begin{aligned}\langle \boldsymbol{\xi}, \boldsymbol{\xi}^\flat \rangle &= \frac{1}{2}\left\langle \frac{p}{\mu} \cdot \frac{\partial}{\partial q} - \kappa q \cdot \frac{\partial}{\partial p}, p \cdot dq - q \cdot dp \right\rangle \\ &= \frac{1}{2}\left(\kappa \cdot q^2 + \frac{1}{\mu} \cdot p^2\right) = \mathcal{H}(p, q)\end{aligned}$$

recovering the Hamiltonian.

C.2. Units in Two-Player Games

Without loss of generality let $\mathbf{w} = (\mathbf{x}; \mathbf{y})$ where we refer to $\mathbf{x}$ as position and $\mathbf{y}$ as momentum so that $\tau(\mathbf{x}) = m$ and $\tau(\mathbf{y}) = \frac{kg \cdot m}{s}$. The aim of this section is to check type-consistency under these, rather arbitrarily assigned, units. Since we are considering a game, we do not require that $\mathbf{x}$ and $\mathbf{y}$ have the same dimension—even though this would necessarily be the case for a physical system. The goal is to verify that units can be consistently assigned to games.

Consider a quadratic two player game of the form

$$\ell_1(\mathbf{w}) = \frac{1}{2}\begin{pmatrix}\mathbf{x}^\intercal & \mathbf{y}^\intercal\end{pmatrix}\begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{pmatrix}\begin{pmatrix}\mathbf{x} \\ \mathbf{y}\end{pmatrix} + \begin{pmatrix}\mathbf{x}^\intercal & \mathbf{y}^\intercal\end{pmatrix}\begin{pmatrix}\mathbf{b}_1 \\ \mathbf{b}_2\end{pmatrix}$$

and

$$\ell_2(\mathbf{w}) = \frac{1}{2}\begin{pmatrix}\mathbf{x}^\intercal & \mathbf{y}^\intercal\end{pmatrix}\begin{pmatrix}\mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{pmatrix}\begin{pmatrix}\mathbf{x} \\ \mathbf{y}\end{pmatrix} + \begin{pmatrix}\mathbf{x}^\intercal & \mathbf{y}^\intercal\end{pmatrix}\begin{pmatrix}\mathbf{d}_1 \\ \mathbf{d}_2\end{pmatrix}$$

We restrict to quadratic games since our methods only involve first and second derivatives. We assume the matrices $\mathbf{A}$ and $\mathbf{C}$ are symmetric without loss of generality so that, for example, $\mathbf{A}_{12} = \mathbf{A}_{21}^\intercal$. Adding constant terms to ℓ_1 and ℓ_2 makes no difference to the analysis so they are omitted.

By Equation (5), the units for ℓ_1 and ℓ_2 should be $\frac{m^2}{s}$ and $\frac{kg \cdot m^2}{s^3}$ respectively. We can therefore derive the correct units for each of the components of the quadratic losses as

$$\underbrace{\begin{pmatrix}m & \frac{kg \cdot m}{s}\end{pmatrix}\begin{pmatrix}\frac{1}{s} & \frac{1}{kg} \\ \frac{1}{kg} & \frac{s}{kg^2}\end{pmatrix}\begin{pmatrix}m \\ \frac{kg \cdot m}{s}\end{pmatrix}}_{\mathbf{w}^\intercal \mathbf{A} \mathbf{w}} + \underbrace{\begin{pmatrix}m & \frac{kg \cdot m}{s}\end{pmatrix}\begin{pmatrix}\frac{m}{s} \\ \frac{m}{kg}\end{pmatrix}}_{\mathbf{w}^\intercal \mathbf{b}}$$

for ℓ_1 and

$$\underbrace{\begin{pmatrix}m & \frac{kg \cdot m}{s}\end{pmatrix}\begin{pmatrix}\frac{kg^2}{s^3} & \frac{kg}{s^2} \\ \frac{kg}{s^2} & \frac{1}{s}\end{pmatrix}\begin{pmatrix}m \\ \frac{kg \cdot m}{s}\end{pmatrix}}_{\mathbf{w}^\intercal \mathbf{C} \mathbf{w}} + \underbrace{\begin{pmatrix}m & \frac{kg \cdot m}{s}\end{pmatrix}\begin{pmatrix}\frac{kg^2 \cdot m}{s^3} \\ \frac{kg \cdot m}{s^2}\end{pmatrix}}_{\mathbf{w}^\intercal \mathbf{d}}$$

for ℓ_2. It follows from a straightforward computation that the vector field $\boldsymbol{\xi} = \frac{\partial \ell_1}{\partial \mathbf{x}}\frac{\partial}{\partial \mathbf{x}} + \frac{\partial \ell_2}{\partial \mathbf{y}}\frac{\partial}{\partial \mathbf{y}}$ has type $\tau(\boldsymbol{\xi}) = \frac{1}{s}$ as required.

The presymplectic form $\omega = d\boldsymbol{\xi}^\flat$ makes use of the musical isomorphism $\flat : T^M \to T^*M$. As in Section C.1, if we naively define $(\frac{\partial}{\partial \mathbf{x}})^\flat = d\mathbf{x}$ and $(\frac{\partial}{\partial \mathbf{y}})^\flat = d\mathbf{y}$ then

$$\boldsymbol{\xi}^\flat = \frac{\partial \ell_1}{\partial \mathbf{x}} \cdot d\mathbf{x} + \frac{\partial \ell_2}{\partial \mathbf{y}} \cdot d\mathbf{y}$$

which is type inconsistent because $\tau(\frac{\partial \ell_1}{\partial \mathbf{x}} \cdot d\mathbf{x}) = \frac{m^2}{s}$ and $\tau(\frac{\partial \ell_2}{\partial \mathbf{y}} \cdot d\mathbf{y}) = \frac{kg^2 \cdot m^2}{s^3}$.

C.2.1. Type-Consistency via SVD

It is necessary, as in Section C.1, to correct the naive musical isomorphism by taking into account the coupling constants for the mixed position-momentum terms. In the classical setup the coupling constants were the scalars $\frac{1}{\mu}$ and κ, whereas in a game they are the off-diagonal blocks $\mathbf{A}_{12}$ and $\mathbf{C}_{12}$.

Apply singular value decomposition to factorize

$$\mathbf{A}_{12} = \mathbf{U}_\mathbf{A}^\intercal \mathbf{D}_\mathbf{A} \mathbf{V}_\mathbf{A} \quad \text{and} \quad \mathbf{C}_{12} = \mathbf{U}_\mathbf{C}^\intercal \mathbf{D}_\mathbf{C} \mathbf{V}_\mathbf{C}$$

where the entries of the diagonal matrices have types $\tau(\mathbf{D}_\mathbf{A}) = \frac{1}{kg}$ and $\tau(\mathbf{D}_\mathbf{C}) = \frac{kg}{s^2}$, and *the types of the orthogonal matrices* $\mathbf{U}$ *and* $\mathbf{V}$ *are pure scalars.* The diagonal matrices $\mathbf{D}_\mathbf{A}$ and $\mathbf{D}_\mathbf{C}$ have the same types as $\frac{1}{\mu}$ and κ in the classical system since they play the same coupling role.

Extending the procedure adopted in the Section C.1, fix the type-inconsistency by defining the musical isomorphisms as

$$\left(\frac{\partial}{\partial \mathbf{x}}\right)^\flat = \mathbf{U}_\mathbf{A}^\intercal \mathbf{D}_\mathbf{A}^{-1} \mathbf{U}_\mathbf{A} \cdot d\mathbf{x}$$

and

$$\left(\frac{\partial}{\partial \mathbf{y}}\right)^\flat = \mathbf{V}_\mathbf{C}^\intercal \mathbf{D}_\mathbf{C}^{-1} \mathbf{V}_\mathbf{C} \cdot d\mathbf{y}.$$

Alternatively, the isomorphisms can be computed by noting that $\mathbf{U}_\mathbf{A}^\intercal \mathbf{D}_\mathbf{A}^{-1} \mathbf{U}_\mathbf{A} = (\sqrt{\mathbf{A}_{12}\mathbf{A}_{21}})^{-1}$ and $\mathbf{V}_\mathbf{C}^\intercal \mathbf{D}_\mathbf{C}^{-1} \mathbf{V}_\mathbf{C} = (\sqrt{\mathbf{C}_{21}\mathbf{C}_{12}})^{-1}$.

The dual isomorphism $\sharp : T^*M \to TM$ is then

$$(d\mathbf{x})^\sharp = \mathbf{U}_\mathbf{A}^\intercal \mathbf{D}_\mathbf{A} \mathbf{U}_\mathbf{A} \cdot \frac{\partial}{\partial \mathbf{x}} \text{ and } (d\mathbf{y})^\sharp = \mathbf{V}_\mathbf{C}^\intercal \mathbf{D}_\mathbf{C} \mathbf{V}_\mathbf{C} \cdot \frac{\partial}{\partial \mathbf{y}}$$

If

$$\boldsymbol{\xi} = \begin{pmatrix} \mathbf{A}_{12}\mathbf{y} + \mathbf{b}_1 \\ \mathbf{C}_{21}\mathbf{x} + \mathbf{d}_2 \end{pmatrix}$$

then it follows that

$$\boldsymbol{\xi}^\flat = \begin{pmatrix} \mathbf{U}_\mathbf{A}^\intercal \mathbf{D}_\mathbf{A}^{-1} \mathbf{U}_\mathbf{A} \mathbf{b}_1 + \mathbf{U}_\mathbf{A}^\intercal \mathbf{U}_\mathbf{A} \mathbf{y} \\ \mathbf{V}_\mathbf{C}^\intercal \mathbf{D}_\mathbf{C}^{-1} \mathbf{V}_\mathbf{C} \mathbf{d}_2 + \mathbf{V}_\mathbf{C}^\intercal \mathbf{V}_\mathbf{C} \mathbf{x} \end{pmatrix}$$

with associated closed two form

$$\omega_\tau = d\boldsymbol{\xi}^\flat = -\left(\mathbf{U}_\mathbf{A}^\intercal \mathbf{V}_\mathbf{A} - \mathbf{U}_\mathbf{C}^\intercal \mathbf{V}_\mathbf{C}\right) d\mathbf{x} \wedge d\mathbf{y}.$$

where the notation ω_τ emphasizes that the two-form is type-consistent.

C.3. What Does Type-Consistency Buy?

Example 10 *Consider the loss functions*

$$f(x, y) = xy \quad \textit{and} \quad g(x, y) = 2xy,$$

with $\boldsymbol{\xi} = (y, 2x)$*. There is no function* $\phi : \mathbb{R}^2 \to \mathbb{R}$ *such that* $\nabla\phi = \boldsymbol{\xi}$*. However, there is a family of functions* $\phi_\alpha(x, y) = \alpha \cdot xy$ *which satisfies*

$$\left\langle \boldsymbol{\xi}, \nabla\phi_\alpha \right\rangle = \alpha \cdot (x^2 + 2y^2) \geq 0 \quad \textit{for all } \alpha > 0.$$

Although $\boldsymbol{\xi}$ is not a potential field, there is a family of functions on which $\boldsymbol{\xi}$ performs gradient descent—albeit with coordinate-wise learning rates that may not be optimal. The vector field $\boldsymbol{\xi}$ arguably does not require adjustment. This kind of situation often arises when the learning rates of different parameters are set adaptively during training of neural nets, by rescaling them by positive numbers.

The vanilla and type-consistent 1-forms corresponding to $\boldsymbol{\xi}$ are, respectively,

$$\boldsymbol{\xi}^\flat = y \cdot dx + 2x \cdot dy \quad \text{and} \quad \boldsymbol{\xi}_\tau^\flat = y \cdot dx + x \cdot dy$$

with

$$\omega = d\boldsymbol{\xi}_{\text{non}}^\flat = dx \wedge dy \quad \text{and} \quad \omega_\tau = d\boldsymbol{\xi}_\tau^\flat = 0.$$

It follows that the *type-consistent* symplectic gradient adjustment is zero. Type-consistency 'detects' that no gradient adjustment is needed in example 10.

References

S Abdallah and V R Lesser. A multiagent reinforcement learning algorithm with non-linear dynamics. *J. Artif. Intell. Res.*, 33:521–549, 2008.

V Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 1989.

J Bailey and G Piliouras. Multiagent learning in network zero-sum games is a Hamiltonian system. In *AAMAS*, 2019.

D Balduzzi. Strongly-typed agents are guaranteed to interact safely. In *ICML*, 2017.

D Balduzzi, S Racanière, J Martens, J Foerster, K Tuyls, and T Graepel. The mechanics of n-player differentiable games. In *ICML*, 2018a.

D Balduzzi, K Tuyls, J Perolat, and T Graepel. Re-evaluating evaluation. In *NeurIPS*, 2018b.

R Bott and L Tu. *Differential Forms in Algebraic Topology*. Springer, 1995.

F Bottacin. A Marsden-Weinstein reduction theorem for presymplectic manifolds. 2005.

M Bowling and M Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

M Bowling. Convergence and no-regret in multiagent learning. In *NeurIPS*, pages 209–216, 2004.

Y Burda, H Edwards, D Pathak, A Storkey, T Darrell, and A Efros. Large-scale study of curiosity-driven learning. In *ICLR*, 2019.

O Candogan, I Menache, A Ozdaglar, and P A Parrilo. Flows and decompositions of games: harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503, 2011.

C Daskalakis, P W Goldberg, and C Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM J. Computing*, 39(1):195–259, 2009.

C Daskalakis, A Ilyas, V Syrgkanis, and H Zeng. Training GANs with optimism. In *ICLR*, 2018.

F Facchinei and C Kanzow. Generalized Nash equilibrium problems. *Annals of Operations Research*, 175(1):177–211, 2010.

S Feizi, C Suh, F Xia, and D Tse. Understanding GANs: the LQG setting. In *arXiv:1710.10793*, 2017.

J N Foerster, R Y Chen, M Al-Shedivat, S Whiteson, P Abbeel, and I Mordatch. Learning with opponent-learning awareness. In *AAMAS*, 2018.

I Gemp and S Mahadevan. Online monotone optimization. In *arXiv:1608.07888*, 2016.

I Gemp and S Mahadevan. Online monotone games. In *arXiv:1710.07328*, 2017.

I Gemp and S Mahadevan. Global convergence to the equilibrium of GANs using variational inequalities. In *arXiv:1808.01531*, 2018.

G Gidel, R A Hemmat, M Pezeshki, R Lepriol, G Huang, S Lacoste-Julien, and I Mitliagkas. Negative momentum for improved game dynamics. In *arXiv:1807.04740*, 2018.

G Gidel, H Berard, G Vignoud, P Vincent, and Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

I J Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, and Y Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

V Guillemin and S Sternberg. *Symplectic Techniques in Physics*. Cambridge University Press, 1990.

S Hart and A Mas-Colell. *Simple Adaptive Strategies: From Regret-Matching to Uncoupled Dynamics*. World Scientific, 2013.

M Heusel, H Ramsauer, T Unterthiner, B Nessler, G Klambauer, and S Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *NeurIPS*, 2017.

M Jaderberg, W M Czarnecki, S Osindero, O Vinyals, A Graves, and K Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In *ICML*, 2017.

X Jiang, L Lim, Y Yao, and Y Ye. Statistical ranking and combinatorial Hodge theory. *Math. Program., Ser. B*, 127:203–244, 2011.

J D Lee, M Simchowitz, M I Jordan, and B Recht. Gradient descent converges to minimizers. In *COLT*, 2016.

JD Lee, I Panageas, G Piliouras, M Simchowitz, MI Jordan, and B Recht. First-order methods almost always avoid saddle points. In *arXiv:1710.07406*, 2017.

A Letcher, J Foerster, D Balduzzi, T Rocktäschel, and S Whiteson. Stable opponent shaping in differentiable games. In *ICLR*, 2019.

B Liu, J Liu, M Ghavamzadeh, S Mahadevan, and M Petrik. Proximal gradient temporal difference learning algorithms. In *IJCAI*, 2016.

X Lu. Hamiltonian games. *Journal of Combinatorial Theory, Series B*, 55:18–32, 1992.

P Mertikopoulos and Z Zhou. Learning in games with continuous action sets and unknown payoff functions. In *arXiv:1608.07310*, 2016.

P Mertikopoulos, C Papadimitriou, and G Piliouras. Cycles in adversarial regularized learning. In *SODA*, 2018.

P Mertikopoulos, H Zenati, B Lecouat, C Foo, V Chandrasekhar, and G Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*, 2019.

L Mescheder, S Nowozin, and A Geiger. The numerics of GANs. In *NeurIPS*. 2017.

L Mescheder, A Geiger, and S Nowozin. Which training methods for GANs do actually converge? In *ICML*, 2018.

L Metz, B Poole, D Pfau, and J Sohl-Dickstein. Unrolled generative adversarial networks. In *ICLR*, 2017.

D Monderer and L S Shapley. Potential games. *Games and Economic Behavior*, 14:124–143, 1996.

V Nagarajan and J Z Kolter. Gradient descent GAN optimization is locally stable. In *NeurIPS*, 2017.

J Nash. Equilibrium points in n-person games. *PNAS*, 36(1):48–49, 1950.

Y Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.

J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics, 2000.

C Papadimitriou and G Piliouras. From Nash equilibria to chain recurrent sets: solution concepts and topology. In *ITCS*, 2016.

C Papadimitriou and G Piliouras. From Nash equilibria to chain recurrent sets: an algorithmic solution concept for game theory. *Entropy*, 20, 2018.

D Pathak, P Agrawal, A A Efros, and T Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. In *arXiv:1610.01945*, 2016.

S Racanière, T Weber, D P Reichert, L Buesing, A Guez, D J Rezende, A P Badia, O Vinyals, N Heess, Y Li, R Pascanu, P Battaglia, D Hassabis, D Silver, and D Wierstra. Imagination-augmented agents for deep reinforcement learning. In *NeurIPS*, 2017.

S Rakhlin and K Sridharan. Optimization, learning, and games with predictable sequences. In *NeurIPS*, 2013.

J B Rosen. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 33(3):520–534, 1965.

R W Rosenthal. A class of games possessing pure-strategy Nash equilibria. *Int J Game Theory*, 2:65–67, 1973.

T Salimans, I Goodfellow, W Zaremba, V Cheung, A Radford, and X Chen. Improved techniques for training GANs. In *NeurIPS*, 2016.

S Santurkar, L Schmidt, and A Madry. A classification-based study of covariate shift in GAN distributions. In *ICML*, 2018.

G Scutari, D P Palomar, F Facchinei, and J Pang. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, pages 35–49, 2010.

Y Shoham and K Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

M Shub. *Global Stability of Dynamical Systems*. Springer, 2000.

S Singh, M Kearns, and Y Mansour. Nash convergence of gradient dynamics in general-sum games. In *UAI*, 2000.

G Stoltz and G Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59:187–208, 2007.

V Syrgkanis, A Agarwal, H Luo, and R E Schapire. Fast convergence of regularized learning in games. In *NeurIPS*, 2015.

A Vezhnevets, S Osindero, T Schaul, N Heess, M Jaderberg, D Silver, and K Kavukcuoglu. FeUdal networks for hierarchical reinforcement learning. In *ICML*, 2017.

G Wayne and L F Abbott. Hierarchical control using networks trained with higher-level forward models. *Neural Computation*, (26), 2014.

J Zhu, T Park, P Isola, and A A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017.

M Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.