Optimization in Smooth Games

Abhilash, Karthik

February 15, 2021

Abstract

The report contains our understanding of two algorithms Consensus Optimization (CO)[1] and Symplectic Gradient Adjustment (SGA) [2]. The notations are introduced as and when they are used. We were not able to include everything we studied in the report. Our motivation for the project was to understand the intricacies involved in multi objective optimization.

1 Consensus optimization (CO)

1.1 Interpreting eigen values of the Hessian

Second derivative of a function carries information about the curvature of the function[3]. In traditional convex optimization, the second partial derivatives are continuous and the differential operator is commutative leading to a symmetric Hessian matrix ($\mathbf{H}$). Symmetric matrices have real eigen values and the eigen vectors orthogonal to each other can be found. The eigen vectors and eigen values of the Hessian matrix provide information about the curvature at a point in a function.

Let's consider a vector $\mathbf{u}$ s.t $||\mathbf{u}|| = 1$ the scalar value $\mathbf{u}^T\mathbf{H}\mathbf{u}$ provides the second derivative in the direction of u. If we let u be the eigen vector then the corresponding eigen value would be $\mathbf{u}^T\mathbf{H}\mathbf{u}$. Thus eigen vectors and corresponding eigen values of a Hessian describe the curvature at a point in x. The reason behind looking at an expression of the form $\mathbf{u}^T\mathbf{H}\mathbf{u}$ is because of the second term in the local quadratic approximation of a function at a point x close to x_0 $f(x) = f(x_0) + f^{'}(x)(x - x_0) + \frac{1}{2}(x - x_0)^T\mathbf{H}(x - x_0)$.

The problem with optimization in games is that the differential operator is not commutative leading to asymmetric Hessian and complex eigen spectra. We feel that interpretation or the intuition behind complex eigen spectra will help us better approach the problem of finding local nash equilibrium.

1.2 Introduction

In CO, the authors show that the reason behind the failure of SGD can be attributed to two below reasons.

- The Jacobian of the gradient vector field (Hessian) having purely eigen values.
- The eigen spectra having eigen values with big imaginary part.

CO alleviates the above problem to some extent leading to batter convergence. The following proposition plays a key role for convergence of CO (Proposition 3 in CO[1]).

Proposition 1: Let $F : \Omega \to \Omega$ be a continuously differential function on an open subset Ω of $\mathbb{R}^n$ and let $\bar{x}$ be so that.

- $F(\bar{x}) = \bar{x}$, and
- The absolute value of the eigen values of the Jacobian $F^{'}(\bar{x})$ are all smaller than 1.

Then there is an open neighbourhood U of $\bar{x}$ so that for all $x_0 \in U$, the iterates $F_k(x_0)$ converge to $\bar{x}$. The rate of convergence is at least linear. More precisely the error $||F^k(x_0) - \bar{x}||$ is in $O(|\lambda_{max}|^k)$ for $k \to \infty$ where λ_{max} is the eigen value of $F^{'}(\bar{x})$ with largest absolute eigen value.

Where the function F gives the next iterate taking current iterate as input.

The above proposition does not talk about the size of the neighbourhood and bounds the error by the eigen value at the fixed point. How do eigen values of intermediate iterates matter? or is the neighbourhood so small that the eigen value at the fixed point is sufficient to bound convergence rate?

1.3 Reason for failure of GD

The function F in numerics is of the form $F(x) = x + hG(x)$ where h is the step size or the learning rate. Finding a fixed point in this case corresponds to solving for $G(x) = 0$. In the case of GD, $G(x) = \boldsymbol{\xi}$, $\boldsymbol{\xi}$ here refers to the first derivative of the function we are trying to optimize.
The Jacobian of F is.

$$F^{'}(x) = I + hG^{'}(x) \tag{1}$$

Both $F^{'}(x)$ and $G^{'}(x)$ can be asymmetric leading to imaginary eigen values. The lemma 4 in CO shows that GD can converges to a stationary point only if the eigen values of the

Jacobian at that point have negative real part and the step size h is bounded as follows.

$$h < \frac{1}{|\Re(\lambda)|} \frac{2}{1 + (\frac{\Im(\lambda)}{\Re(\lambda)})^2} \tag{2}$$

If the above conditions are satisfied the absolute value of the eigen values of the Jacobian $F^{'}(x)$ fall in the unit circle and proposition 1 becomes applicable.

The two reasons for the failure of GD now become apparent as both the real part of the eigen value and the ratio of imaginary part to real part are inversely related to the stepsize. If the above quantities are huge than the step size to be taken for convergence becomes small.

Another interesting thing to note is the fact that GD can converge (If convergence is possible) to points that need not be a local Nash Equilibrium. Work in [4] provides an example on which gradient descent converges to a min-min point. [4] addresses the above problem by using the second order information. In summary, the unstable points are avoided by adding a correction term to repel from such points.

1.4 Consensus Optimization (CO)

The objective now is to design a function F that satisfies the following properties.

1. The fixed points of F correspond to the stationary points of the $\boldsymbol{\xi}$. (i.e if $F(\bar{x}) = \bar{x}$ then $\boldsymbol{\xi} = 0$)
2. The second order conditions for negative semi definiteness should be met. (to guarantee the point is local nash equilibrium)
3. Ideally, the Jacobian of F, represented $F^{'}$ should have eigen spectra in the unit circle.

The function F proposed in CO is as follows.

$$F(x) = x + hA(x)\boldsymbol{\xi} \tag{3}$$

Where $A(x) = I - \gamma \mathbf{H}^T$, it is assumed that $\frac{1}{\gamma}$ is not an eigen value of $\mathbf{H}^T$ so that A is invertible.

Substituting the value of $A(x)$ the equation reduces to

$$F(x) = x + h(\boldsymbol{\xi} - \gamma \mathbf{H}^T \boldsymbol{\xi}) \tag{4}$$

Note that $\mathbf{H}^T \boldsymbol{\xi} = \nabla \mathcal{H}$ Where $\mathcal{H} = \frac{1}{2}||\boldsymbol{\xi}||^2$.

The Jacobian of F will be given by

$$F^{'}(x) = I + hA(x)\mathbf{H} \tag{5}$$

While taking the Jacobian of F, not sure why the derivative of $A(x)$ is not taken. $A(x)$ stays the same in the Jacobian as $A(x) = I - \gamma H^T$ as per the equations in [1].

Equation 5 reduces to the following form after substituting $A(x)$.

$$F^{'}(x) = I + h(\mathbf{H} - \gamma \mathbf{H}^T \mathbf{H}) \tag{6}$$

It is easy to verify that minimizing the hamiltonian $\mathcal{H}$ reduces the norm of the gradient making the gradient zero. The Hamiltonian Gradient Descent(HGD)[5] does the same. The above algorithm is shown to converge to other local minima. Hence not desirable. The term $\boldsymbol{\xi} - \gamma \nabla \mathcal{H}$ is in the direction of $\boldsymbol{\xi}$ and also performs descent along $\nabla \mathcal{H}$ with γ controlling the importance of the two terms.

- Property 1 is satisfied by the above adjustment as $\boldsymbol{\xi} - \gamma \nabla \mathcal{H} = 0$ when $\boldsymbol{\xi} = 0$.
- Property 2 is also satisfied as it can be shown that if $\mathbf{H}$ is negative semidefinite and invertible then $\mathbf{H} - \gamma \mathbf{H}^T \mathbf{H}$ is also negative semidefinite.
- Property 3 the following lemma from bounds the value of $1 + (\frac{\Im(\lambda)}{\Re(\lambda)})$ by choosing a suitable γ.

Lemma 1[1]: Assume $A \in \mathbb{R}^{n \times n}$ is negative semidefinite. Let $q(\gamma)$ be the maximum value of $\frac{\Im(\lambda)}{\Re(\lambda)}$ (possibly infinite) with respect to λ where λ denotes the eigen values of $A - \gamma A^T A$ and $\Re(\lambda)$ and $\Im(\lambda)$ denote thier real and imaginary part respectively. Moreover, assume that A is invertible with $|Av| \geq \rho v$ for $\rho > 0$ and let

$$c = \min_{v \in \mathbb{S}(\mathbb{C}^n)} \frac{|\bar{v}^T (A + A^T) v|}{|\bar{v}^T (A - A^T) v|} \tag{7}$$

Where $\mathbb{S}(\mathbb{C}^n)$ denotes the unit sphere in $\mathbb{C}^n$. Then,

$$q(\gamma) \leq \frac{1}{c + 2\rho^2 \gamma} \tag{8}$$

Proof Let $v \in \mathbb{C}^n$ s.t $||v|| \neq 0$ be any eigen vector of $B = A - \gamma A^T A$ and $\lambda \in \mathbb{C}$ the corresponding eigen value. We can assume w.l.o.g that $||v|| = 1$.
Then

$$\lambda = \lambda \bar{v}^T v = \bar{v}^T B v \tag{9}$$

This implies that

$$\Re(\lambda) = \frac{\lambda + \bar{\lambda}}{2} = \frac{1}{2} \bar{v}^T (B + B^T) v \tag{10}$$

similarly

$$\Im(\lambda) = \frac{\lambda - \bar{\lambda}}{2i} = \frac{1}{2i} \bar{v}^T (B - B^T) v \tag{11}$$

Consequently we have

$$\frac{|\Im\lambda|}{|\Re\lambda|} = \frac{|\bar{v}^T(B - B^T)v|}{|\bar{v}^T(B + B^T)v|} \tag{12}$$

and thus

$$q(\gamma) \leq \max_{v \in \mathbb{S}(\mathbb{C}^n)} \frac{|\bar{v}^T(B - B^T)v|}{|\bar{v}^T(B + B^T)v|} \tag{13}$$

However, we have

$$B - B^T = A - A^T \tag{14}$$

$$B + B^T = A + A^T - 2\gamma A^T A \tag{15}$$

Not sure of the reason behind $AA^T = A^T A$ substituting the above values in (13).

$$\frac{|\bar{v}^T(B - B^T)v|}{|\bar{v}^T(B + B^T)v|} = \frac{|\bar{v}^T(A - A^T)v|}{|\bar{v}^T(A + A^T)v| + 2\gamma||Av||^2} \tag{16}$$

Since $||Av||^2 \geq \rho||v||^2 = \rho$ thus.

$$q(\gamma) \leq \frac{1}{c + 2\rho^2\gamma} \tag{17}$$

In CO the problem of restricting the eigen spectrum to the unit circle at stationary points is not fully solved. The proof of Lemma 1 bounds only the ratio of imaginary part to real part. The Proposition 1 requires the eigen spectrum to be in the unit circle. An interesting problem to explore would be to see if there are other adjustements to bound or restrict the eigen spectrum to unit circle leading to better convergence.

2 Symplectic Gradient Adjustment

2.1 Preliminaries

SGA[2] solves for n player differentiable games where each agent has an objective. Loss is taken as objective and every agent minimizes their respective loss. So, it can be termed as multi-objective optimization. For the sake of convenience, we will follow the same notations as in the paper.

Definition 1. A game is a set of $[n] = \{1, 2, 3...n\}$ players with action set $\mathbf{w} = (\mathbf{w_1}, \mathbf{w_2}...\mathbf{w_n}) \in \mathbb{R}^d$ where $d = \sum_{i=1}^{n} d_i$. Each agent i has an associated continuously twice differentiable loss $l_i : \mathbb{R}^d \to \mathbb{R}$ which they are trying to minimize.

The simultaneous gradient is the gradient of the losses w.r.t their weights given below.

$$\boldsymbol{\xi}(\mathbf{w}) = (\nabla_{w_1} l_1, \nabla_{w_2} l_2, .., \nabla_{w_n} l_n)$$

.
Hessian of the game is a $d \times d$ matrix of second order derivatives.

$$\mathbf{H}(\mathbf{w}) = \begin{pmatrix} \nabla^2_{w_1} l_1 & \nabla^2_{w_1,w_2} l_1 & \dots & \nabla^2_{w_1,w_n} l_1 \\ \nabla^2_{w_2,w_1} l_2 & \nabla^2_{w_2} l_2 & \dots & \nabla^2_{w_2,w_n} l_2 \\ \vdots & & & \vdots \\ \nabla^2_{w_n,w_1} l_n & \nabla^2_{w_n,w_2} l_n & \dots & \nabla^2_{w_n} l_n \end{pmatrix}$$

One thing to note here is that, the Hessian $\mathbf{H}$ need not be symmetric and so can have imaginary eigen values. By using the fact that any matrix can be decomposed into a symmetric component $\mathbf{S}$ and an anti-symmetric component $\mathbf{A}$, concept of Helmholtz decomposition is generalized where $\mathbf{S}$ and $\mathbf{A}$ represent the curl-free part and rotational component respectively.

$$\mathbf{S} = \frac{\mathbf{H} + \mathbf{H^T}}{2}, \mathbf{A} = \frac{\mathbf{H} - \mathbf{H^T}}{2} \implies \mathbf{H} = \mathbf{S} + \mathbf{A}$$

In such a case, we can consider two solution concepts for a game, stable fixed points and local Nash equilibrium.

Definition 2. A fixed point $\mathbf{w}^*$ with $\boldsymbol{\xi}(\mathbf{w}^*) = 0$ is defined as **stable** if $\mathbf{S}(\mathbf{w}) \succcurlyeq 0$ and unstable if $\mathbf{S}(\mathbf{w}) \prec 0$ for $\mathbf{w}$ in a neighborhood on $\mathbf{w}^*$.

Note: As we are looking at minimization of loss, the standard second order constraint for a stable fixed point would be $\mathbf{H}(\mathbf{w}) \succcurlyeq 0$. But, positive definiteness of $\mathbf{S}$ implies the positive definiteness of the $\mathbf{H}$ because of equation(H=S+A) and the anti-symmetric nature of $\mathbf{A}$.

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 $l_i(\mathbf{w_i}, \mathbf{w_i^*}) \geq l_i(\mathbf{w_i^*}, \mathbf{w_i^*})$ for $\mathbf{w_i} \in U_i$.

2.2 Potential and Hamiltonian Games

Definition 4. A game is a **potential game** if $\mathbf{A} \equiv 0$ and a **Hamiltonian game** if $\mathbf{S} \equiv 0$.

Although, this definition of a potential game is not completely general, it is defined as such for simplicity. For this analysis, potential games are games that have a single potential function ϕ and descent on its simultaneous gradient $\boldsymbol{\xi}$ converges to either a fixed point that is local minimum of ϕ or a saddle.

Theorem 1. Let $\mathcal{H} := \frac{1}{2}||\boldsymbol{\xi}||^2$ as defined in CO. If the game is Hamiltonian then,

(i) $\nabla\mathcal{H} = \mathbf{A^T}\boldsymbol{\xi}$.

(ii) $\boldsymbol{\xi}$ preserves level sets of $\mathcal{H}$, that is dot product $\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle = 0$.

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

Proof:

(i) As seen in CO proof and by direct computation, $\nabla\mathcal{H} = \mathbf{H^T}\boldsymbol{\xi} = \mathbf{A^T}\boldsymbol{\xi}$ as $\mathbf{S} \equiv 0$ in Hamiltonian games.

(ii) From (i), $\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle = \langle \boldsymbol{\xi}, \mathbf{A^T}\boldsymbol{\xi} \rangle = \boldsymbol{\xi}^T\mathbf{A^T}\boldsymbol{\xi} = (\boldsymbol{\xi}^T\mathbf{A^T}\boldsymbol{\xi})^T = -\boldsymbol{\xi}^T\mathbf{A^T}\boldsymbol{\xi} = 0$ as A is anti-symmetric.

(iii) Convergence from gradient descent on $\mathcal{H}$ leads to a point where $\nabla\mathcal{H} = \mathbf{H^T}\boldsymbol{\xi} = 0$. As $\mathbf{H} \not\equiv 0 \implies \boldsymbol{\xi} = 0$ which is a fixed point. In an Hamiltonian game, we have $\mathbf{S} \equiv 0 \implies \mathbf{S} \succcurlyeq 0 \implies \mathbf{H} \succcurlyeq 0$ which makes it stable from Def 2.

Hamiltonian games have a function $\mathcal{H}$ that specifies the quantity preserved by the dynamics. In such games, $\boldsymbol{\xi}$ preserves the level sets of $\mathcal{H}$ which implies that dot product $\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle = 0$. This causes the dynamics of the game to rotate around the equilibrium.

With this, we understand that descent on $\boldsymbol{\xi}$ converges in case of potential games and descent on $\nabla\mathcal{H}$ converges in case of Hamiltonian games. But games in general are a mix of both.

The following desiderata provides properties that any aligned gradient $\boldsymbol{\xi}_\lambda$ should satisfy to perform well in games.

2.3 Desiderata

- *D1*: Compatible with game dynamics $\langle \boldsymbol{\xi}_\lambda, \boldsymbol{\xi} \rangle = \alpha_1 \cdot ||\boldsymbol{\xi}||^2$
- *D2*: Compatible with potential dynamics $\langle \boldsymbol{\xi}_\lambda, \nabla\phi \rangle = \alpha_2 \cdot ||\boldsymbol{\xi}||^2$
- *D3*: Compatible with Hamiltonian dynamics $\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle = \alpha_3 \cdot ||\boldsymbol{\xi}||^2$
- *D4*: Attracted to stable equilibria $\theta(\boldsymbol{\xi}_\lambda, \nabla\mathcal{H}) \leq \theta(\boldsymbol{\xi}, \nabla\mathcal{H})$ where $\mathbf{S} \succ 0$.
- *D5*: Repelled by unstable equilibria $\theta(\boldsymbol{\xi}_\lambda, \nabla\mathcal{H}) \geq \theta(\boldsymbol{\xi}, \nabla\mathcal{H})$ where $\mathbf{S} \prec 0$.

where $\alpha_1, \alpha_2, \alpha_3 > 0$ and θ is the angle between the argument vectors.

Here, two non zero vectors are said to be compatible if their dot products are positive. It means that after the adjustment, the gradient should be generally in the same direction as the other component in the inner product.

Considering the simultaneous gradient $\boldsymbol{\xi}$, it satisfies *D1, D2* by Theorem 1 and we know that it converges in potential games. It does not satisfy *D3* as we have that $\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle = 0$. The reason why $\boldsymbol{\xi}$ satisfies *D4, D5* we believe was explained by [6] and it has something to do with stable manifolds which we are yet to understand.

Consensus Optimization considered in the first part of this report has an adjustment of the form $\boldsymbol{\xi_{CO}} = \boldsymbol{\xi} + \cdot\mathbf{H}^T\boldsymbol{\xi}$. It satisfies only *D3, D4*.

A symplectic adjustment of the form

$$\boldsymbol{\xi}_\lambda = \boldsymbol{\xi} + \lambda \cdot \mathbf{A^T}\boldsymbol{\xi}$$

is introduced which satisfies *D1, D2, D3*. We can see that the difference between this and CO is mainly that it has only the anti-symmetric component of $\mathbf{H}$ in the alignment. This seems reasonable because we do not need information from the potential component for the alignment on the gradient to converge on potential games.

Proof: *D1* is satisfied by **Theorem 1** as $\boldsymbol{\xi}^T\mathbf{A^T}\boldsymbol{\xi} = 0$. *D2* is satisfied as $\mathbf{A} \equiv 0$ in potential games. We have to note here that *D3* is satisfied only if $\lambda > 0$ as $\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle = \langle \boldsymbol{\xi} + \lambda\mathbf{A^T}\boldsymbol{\xi}, \mathbf{A^T}\boldsymbol{\xi} \rangle = \lambda||\nabla\mathcal{H}||^2$.

Choosing sign of λ is important to satisfy *D4,D5* which is predominantly determined by the expression $\texttt{sign}(\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle \langle \mathbf{A}^T\boldsymbol{\xi}, \nabla\mathcal{H} \rangle)$. Intuitively, the first expression determines the stability of the fixed point and second expression gives the direction of alignment w.r.t the gradient of Hamiltonian. The optimal thing we want to do is to descend on the gradient of Hamiltonian and reach a stable point.

$\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle = \boldsymbol{\xi}^T\mathbf{H}^T\boldsymbol{\xi} = \boldsymbol{\xi}^T\mathbf{S}^T\boldsymbol{\xi} + \boldsymbol{\xi}^T\mathbf{A}^T\boldsymbol{\xi} = \boldsymbol{\xi}^T\mathbf{S}^T\boldsymbol{\xi}$. From this and definition of stability, positive semi-definiteness of the expression $\langle \boldsymbol{\xi}, \nabla\mathcal{H} \rangle$ implies the stability of the fixed point being approached, and negative definiteness implies that the fixed point being approached is unstable. The sign of dot product $\langle \mathbf{A}^T\boldsymbol{\xi}, \nabla\mathcal{H} \rangle$ gives the general direction of the adjustment w.r.t $\nabla\mathcal{H}$. If its positive, it means that the adjustment takes us more towards the fixed point and negative sign implies that we would move away from the point. Picking the $\texttt{sign}(\lambda)$ based on these two expressions therefore would satisfy the desiderata *D4,D5*.

Ultimately, we can think of gradient $\boldsymbol{\xi}$ and an aligned gradient $\boldsymbol{\xi_\lambda}$ as this.

- Descending on $\boldsymbol{\xi}$ leads to convergence in potential games
- But this $\boldsymbol{\xi}$ cycles in case of purely Hamiltonian games $\implies \boldsymbol{\xi} \perp \nabla\mathcal{H}$.
- So an alignment is introduced to descend on these Hamiltonian level sets preserved by unaligned $\boldsymbol{\xi}$. So $\boldsymbol{\xi}_\lambda$ is no more perpendicular to $\nabla\mathcal{H}$. (*D3* takes care of this)
- But we have to make sure that this alignment does not interfere with $\boldsymbol{\xi}$'s descent on potential games. (*D2* takes care of this)
- At the same time, we have to make sure that the point which $\nabla\mathcal{H}$ is taking us towards is a stable point. (*D4* takes care of this)
- If its unstable, we need the alignment has to take us away from the point. (*D5* takes care of this)

Things in this paper that are not completely clear to us.

- `The convergence:` If we look at the Hamiltonian dynamics of the aligned gradient in general games, $\langle \boldsymbol{\xi}_\lambda, \nabla\mathcal{H} \rangle$, this is positive only when $\mathbf{S}$ and $\mathbf{A}$ commute. So the values that λ can take are limited by this condition. Formal convergence is proved by using Ostrowski's theorem, which we believe is analogous to **Proposition 3** discussed in CO. Informally, it states that if the eigen values of the Jacobian at fixed point are less than or equal to 1, then there exists a neighborhood around the fixed point from where, the iterates converge to the fixed point.
- `Strict Saddles:` Desiderata only handles stable and unstable points. Regarding strict saddles, they claim that the negative eigen value of the Jacobian of saddle point dominates the Taylor expansion, which prevents the convergence to strict saddles. They formalize the proof using Stable Manifold Theorem.
- `The right solution concept between local nash and stable fixed points(SFP)?` [7] says that in SGA, agents might act against oneself prioritizing stability over individual loss. This we believe comes from the fact that SGA adopts SFP as their solution concept. One question to ask here is that, is there a way to avoid these non Nash stable points? Inspired by [4].

References

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

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

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[4] Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach, 2018.

[5] Jacob Abernethy, Kevin A. Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization, 2019.

[6] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points, 2017.

[7] Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games, 2018.