The Limits of Min-Max Optimization Algorithms: Convergence to Spurious Non-Critical Sets

Ya-Ping Hsieh [1] **Panayotis Mertikopoulos** [2 3] **Volkan Cevher** [4]

Abstract

Compared to ordinary function minimization problems, min-max optimization algorithms encounter far greater challenges because of the existence of periodic cycles and similar phenomena. Even though some of these behaviors can be overcome in the convex-concave regime, the general case is considerably more difficult. With this in mind, we take an in-depth look at a comprehensive class of state-of-the art algorithms and prevalent heuristics in *non-convex / non-concave* problems, and we establish the following general results: *a*) generically, the algorithms' limit points are contained in the *internally chain-transitive* (ICT) sets of a common, mean-field system; *b*) the attractors of this system also attract the algorithms in question with arbitrarily high probability; and *c*) all algorithms avoid the system's unstable sets with probability 1. On the surface, this provides a highly optimistic outlook for min-max algorithms; however, we show that there exist *spurious attractors* that do not contain *any* stationary points of the problem under study. In this regard, our work suggests that existing min-max algorithms may be subject to inescapable convergence failures. We complement our theoretical analysis by illustrating such attractors in simple, two-dimensional, almost bilinear problems.

1. Introduction

Consider a min-max optimization – or *saddle-point* – problem of the form

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y). \tag{SP}$$

Given an algorithm for solving (SP), it is then natural to ask:

Where does the algorithm converge to? ($\star$)

The goal of our paper is to treat ($\star$) in a general non-convex / non-concave setting and to provide answers for a comprehensive array of state-of-the-art algorithms.

Related work. This question has attracted significant interest in the machine learning literature because of its potential implications to generative adversarial networks (Goodfellow et al., 2014), robust reinforcement learning (Kamalaruban et al., 2020; Pinto et al., 2017), and other models of adversarial training (Madry et al., 2018). In this broad setting, it has become empirically clear that the joint training of two neural networks (NNs) is fundamentally more difficult than that of a *single* NN of similar size and architecture. The latter task boils down to successfully finding a (good) local minimum of a non-convex function, so it is instructive to revisit ($\star$) in the context of *non-convex minimization*.

In this case, the existing convergence theory for *stochastic gradient descent* (SGD) – the "gold standard" for deep NN training – can be informally summed up as follows:

1. SGD always converges to critical points.
2. SGD does not converge to strict saddle points or other spurious solutions.

These results could be seen as plausible expectations for algorithmic proposals to solve (SP). Unfortunately however, there are well-known examples of simple *bilinear* min-max games where stochastic gradient descent/ascent (SGDA), the min-max analogue of SGD, leads to recurrent orbits that do not contain *any* critical point of Φ. Such *spurious convergence* phenomena arise from the min-max structure of (SP) and have no counterpart in minimization problems.

This well-documented failure of SGDA has led to an extensive literature that is impossible to survey here. As a purely indicative – and highly incomplete – list, we mention the works of Daskalakis et al. (2018), Gidel et al. (2019a), Mertikopoulos et al. (2019) and Mokhtari et al. (2019a), who studied how these failures can be overcome in *deterministic* bilinear problems by means of an *extra-gradient* step (or an optimistic proxy thereof). By contrast, in *stochastic* prob-

[1]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [2]Univ. Grenoble Alpes, CNRS, Inria, LIG, Grenoble, France. [3]Criteo AI Lab. [4]Ecole Polytechnique Fedérale de Lausanne, Switzerland. Correspondence to: Ya-Ping Hsieh <yaping.hsieh@inf.ethz.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

lems, the convergence of optimistic / extra-gradient methods is compromised unless additional, tailor-made mitigation mechanisms are put in place – such as variance reduction (Chavdarova et al., 2019; Iusem et al., 2017) or variable step-size schedules (Hsieh et al., 2020). This shows that the convergence of min-max training methods can be particularly fragile, even in simple, bilinear problems.

Beyond the class of convex-concave problems analyzed above, another vigorous thread of research has focused on the *local analysis* of a min-max optimization algorithm close to the game's critical points – typically subject to a second-order sufficient condition; cf. Adolphs et al. (2019); Daskalakis & Panageas (2018); Fiez & Ratliff (2020); Grimmer et al. (2020a;b); Heusel et al. (2017); Hsieh et al. (2019a); Mazumdar et al. (2020); Nagarajan & Kolter (2017). The global analysis is much more challenging and requires strong structural assumptions such as variational coherence (Mertikopoulos & Zhou, 2019; Mertikopoulos et al., 2019) and/or the existence of a Minty-type solution (Liu et al., 2019a; Zhou et al., 2017; 2020). In the absence of such conditions, Flokas et al. (2019; 2021) showed that periodic and/or Poincaré recurrent behavior may persist in deterministic, continuous-time min-max dynamics.

From a practical viewpoint, these studies have led to a broad array of sophisticated algorithmic proposals for solving min-max games; we review many of these algorithms in Section 3. However, a central question that remains unanswered is whether it is theoretically plausible to expect a qualitatively different behavior relative to SGDA in the full spectrum of non-convex / non-concave games. Our work aims to provide concrete answers to this question.

Our contributions. Our first contribution is to provide a unified framework for a comprehensive selection of first- and zeroth-order min-max optimization methods (including SGDA, proximal point methods, optimistic / extra-gradient schemes, their alternating variants, etc.). The principal ingredients of our approach are twofold: (*i*) a generalized Robbins–Monro (RM) template that is wide enough to include all the above algorithms; and (*ii*) an analytic framework leveraging the ordinary differential equation (ODE) method of stochastic approximation (Benaïm, 1999; Kushner & Yin, 1997). Based on these two elements, we prove a precise version of the following general principle: *the long-run behavior of all generalized RM methods can be mapped to the study of* ***the same****, mean-field dynamical system.*

In more detail, we show that the limit points of all generalized RM schemes belong to an *internally chain-transitive* (ICT) set of these mean dynamics. The notion of an ICT set is central in the study of dynamical systems (Benaïm & Hirsch, 1996; Bowen, 1975; Conley, 1978) and, in some cases, they are easy to characterize: in minimization problems (and possibly up to a "hidden" transformation in the spirit of Flokas et al., 2019), the dynamics' ICT sets are the function's critical points. As such, in this case, we recover *exactly* the min-min landscape of SGD – but for an *entire family* of algorithms, not just SGD.

Moving on to *general* min-max problems, the structure of the dynamics' ICT sets could be considerably more complicated, so we provide two further, complementing results:

1. *With high probability, all generalized RM methods converge locally to attractors of the mean dynamics.*
2. *With probability* 1*, all generalized RM methods avoid the mean dynamics' unstable invariant sets.*

As far as we are aware, there are no results of comparable generality in the min-max optimization literature. From a high level, these theoretical contributions would seem to be analogous to existing results for SGD in minimization problems (i.e., that SGD converges to critical points while avoiding strict saddles). However, this similarity is only skin-deep: as we show by a range of concrete, *almost bilinear* examples, min-max optimization algorithms may encounter a series of immovable roadblocks. Specifically,

- An ICT set may contain a *globally attracting limit cycle*, and the range of algorithms under consideration cannot escape it – even though extra-gradient methods escape recurrent orbits in exact bilinear problems. This suggests that bilinear games may not be representative as a testbed for GAN training algorithms and heuristics.
- There exist *unstable* critical points whose neighborhood contains an (almost) *globally stable* ICT set. Therefore, in sharp contrast to minimization, "avoiding unstable critical points" *does not imply* "escaping unstable critical points" in min-max problems.
- There exist *stable* min-max points whose basin of attraction is "shielded" by an *unstable* ICT set. As a result, if run with non-negligible noise in the gradients, then, with high probability, existing algorithms are repelled away from the desirable solutions.

Our results indicate a steep, qualitative increase in difficulty when passing from min-min to min-max problems, in line with concurrent works by Daskalakis et al. (2020) and Letcher (2020). In plain terms, Daskalakis et al. (2020) proved the impossibility of attaining a critical point in polynomial time in deterministic, constrained min-max games. In a similar spirit, the concurrent work of Letcher (2020) showed that there are min-max games where all "reasonable" deterministic algorithms may fail to converge. By contrast, our paper focuses on the occurrence of *spurious convergence phenomena* with probability 1 in *stochastic* algorithms. In addition, our avoidance result (Theorem 3) can be seen as a stochastic counterpart of the "reasonableness" requirement of Letcher (2020), thereby enriching the

applicability of the results therein. Taken together, these works and our own provide a complementing look into the fundamental limits of min-max optimization algorithms.

2. Setup and preliminaries

Throughout our paper, we focus on general unconstrained problems with $\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}$, $\mathcal{Y} = \mathbb{R}^{d_\mathcal{Y}}$, and Φ assumed C^1 and Lipschitz. To simplify notation, we will let $z = (x, y)$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $d = d_\mathcal{X} + d_\mathcal{Y}$. In addition, we will write

$$V(z) \equiv (V_x(x,y), V_y(x,y)) \coloneqq (-\nabla_x \Phi(x,y), \nabla_y \Phi(x,y))$$

for the (min-max) gradient field of Φ, assumed here to be Lipschitz; in some cases we may also require V to be C^1 and write $JV(z)$ for its Jacobian. Finally, we will assume that V satisfies the weak asymptotic coercivity condition

$$\langle V(z), z\rangle \leq 0 \quad \text{for all sufficiently large } z. \tag{1}$$

This condition is a weaker version of standard coercivity conditions in the literature (Bauschke & Combettes, 2017), it is satisfied by all convex-concave problems (including bilinear ones) and, importantly, it does not impose any growth requirements on the elements of V (as standard coercivity conditions do). We discuss it further in Appendix A.

A *solution* of (SP) is a tuple $z^* = (x^*, y^*)$ with $\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*)$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$; likewise, a *local solution* of (SP) is a tuple (x^*, y^*) that satisfies this inequality locally. Finally, a state z^* with $V(z^*) = 0$ is said to be a *critical* (or *stationary*) *point* of Φ.

From an algorithmic standpoint, we will focus exclusively on the black-box optimization paradigm (Nesterov, 2004) with *stochastic first-order oracle* (SFO) feedback. Algorithms with a more complicated feedback structure, such as a best-response oracle (Fiez et al., 2019; Jin et al., 2019; Naveiro & Insua, 2019) or based on mixed-strategy sampling (Domingo-Enrich et al., 2020; Hsieh et al., 2019b; Kamalaruban et al., 2020), are not considered in this work.

Specifically, when called at $z = (x, y)$ with random seed $\omega \in \Omega$, an SFO returns a random vector $\mathsf{V}(z;\omega) \equiv (\mathsf{V}_x(z;\omega), \mathsf{V}_y(z;\omega))$ of the form

$$\mathsf{V}(z;\omega) = V(z) + \mathsf{U}(z;\omega) \tag{SFO}$$

where the error term $\mathsf{U}(z;\omega)$ captures all sources of uncertainty in the model (e.g., the selection of a minibatch in GAN training, system state observations in reinforcement learning, etc.). As is standard in the literature, we require $\mathsf{U}(z;\omega)$ to be zero-mean and finite-variance:

$$\forall z \in \mathcal{Z}, \quad \mathbb{E}[\mathsf{U}(z;\omega)] = 0 \text{ and } \mathbb{E}[\|\mathsf{U}(z;\omega)\|^2] \leq \sigma^2. \tag{2}$$

These will be our blanket assumptions throughout.

3. Core algorithmic framework

3.1. The Robbins–Monro template

Much of our analysis will revolve around iterative algorithms that can be cast as generalized Robbins–Monro algorithms (Robbins & Monro, 1951) of the general form

$$Z_{n+1} = Z_n + \gamma_n [V(Z_n) + W_n], \tag{RM}$$

where

1. $Z_n = (X_n, Y_n) \in \mathcal{Z}$ denotes the state of the algorithm at each stage $n = 1, 2, \dots$;
2. W_n is an abstract error term described in detail below;
3. γ_n is the method's step-size hyperparameter, and is typically of the form $\gamma_n \propto 1/n^p$ for some $p \geq 0$. Throughout the paper, we will always assume $\sum_n \gamma_n = \infty$ and $\lim_n \gamma_n = 0$.

In the above, the error term W_n is generated *after* Z_n; thus, by default, W_n is not adapted to the history $\mathcal{F}_n \coloneqq \mathcal{H}(Z_1, \dots, Z_n)$ of Z_n. For concision, we will also write

$$V_n = V(Z_n) + W_n \tag{3}$$

so V_n can be seen as a noisy estimator of $V(Z_n)$. In more detail, to differentiate between "random" (zero-mean) and "systematic" (non-zero-mean) errors in V_n it will be convenient to further decompose the error process W_n as

$$W_n = U_n + b_n \tag{4}$$

where $b_n = \mathbb{E}[W_n \,|\, \mathcal{F}_n]$ represents the systematic component and $U_n = W_n - b_n$ captures the random, zero-mean part. In view of all this, we will consider the following descriptors for W_n:

$$a) \quad \textit{Bias:} \qquad B_n = \mathbb{E}[\|b_n\| \,|\, \mathcal{F}_n]. \tag{5a}$$

$$b) \quad \textit{Variance:} \qquad \sigma_n^2 = \mathbb{E}[\|U_n\|^2 | \mathcal{F}_n]. \tag{5b}$$

Note that (conditioned on $\mathcal{F}_n$) both B_n and σ_n are random; this will play an important part in the sequel.

3.2. Specific algorithms

In the rest of this section, we discuss how a wide range of algorithms used in the literature can be seen as special instances of our general Robbins–Monro (RM) template.

▼ **Algorithm 1** (Stochastic gradient descent/ascent)**.** The widely used stochastic gradient descent/ascent (SGDA) algorithm – also known as the *Arrow–Hurwicz* method (Arrow et al., 1958) – queries an SFO and proceeds as:

$$Z_{n+1} = Z_n + \gamma_n \mathsf{V}(Z_n; \omega_n), \tag{SGDA}$$

where $\omega_n \in \Omega$ ($n = 1, 2, \dots$) is an independent and identically distributed (i.i.d.) sequence of oracle seeds. As such, (SGDA) admits a straightforward RM representation by taking $W_n = U_n = \mathsf{U}(Z_n; \omega_n)$ and $b_n = 0$. ▲

▼ **Algorithm 2** (Proximal point method)**.** The (deterministic) *proximal point method* (PPM) (Rockafellar, 1976) is an implicit update rule of the form:

$$Z_{n+1} = Z_n + \gamma_n V(Z_{n+1}). \tag{PPM}$$

The RM representation of (PPM) is obtained by taking $W_n = b_n = V(Z_{n+1}) - V(Z_n)$ and $U_n = 0$. ▲

▼ **Algorithm 3** (Stochastic extra-gradient)**.** Since (PPM) is only implicitly defined, one can rarely run it in practice. Nonetheless, it is possible to approximate (PPM) by locally querying two (stochastic) gradients at each iteration (Nemirovski, 2004b). This can be achieved by the *stochastic extra-gradient* (SEG):

$$\begin{aligned} Z_n^+ &= Z_n + \gamma_n \mathsf{V}(Z_n; \omega_n), \\ Z_{n+1} &= Z_n + \gamma_n \mathsf{V}(Z_n^+; \omega_n^+). \end{aligned} \tag{SEG}$$

To recast (SEG) in the Robbins–Monro framework, simply take $W_n = \mathsf{V}(Z_n^+; \omega_n^+) - V(Z_n)$, i.e., $U_n = \mathsf{U}(Z_n^+; \omega_n^+)$ and $b_n = V(Z_n^+) - V(Z_n)$. ▲

▼ **Algorithm 4** (Optimistic gradient / Popov's extra-gradient)**.** Compared to (SGDA), the scheme (SEG) involves two oracle queries per iteration, which is considerably more costly. An alternative iterative method with a single oracle query per iteration was proposed by Popov (1980):

$$\begin{aligned} Z_n^+ &= Z_n + \gamma_n \mathsf{V}(Z_{n-1}^+; \omega_{n-1}), \\ Z_{n+1} &= Z_n + \gamma_n \mathsf{V}(Z_n^+; \omega_n). \end{aligned} \tag{OG/PEG}$$

Popov's extra-gradient has been rediscovered several times and is more widely known as the optimistic gradient (OG) method in the machine learning literature (Chiang et al., 2012; Daskalakis et al., 2018; Hsieh et al., 2019a; Rakhlin & Sridharan, 2013). In unconstrained problems, (OG/PEG) turns out to be equivalent to a number of other existing methods, including "extrapolation from the past" (Gidel et al., 2019a) and reflected gradient (Malitsky & Tam, 2020). Its Robbins–Monro representation is obtained by setting $W_n = \mathsf{V}(Z_n^+; \omega_n) - V(Z_n)$, i.e., $U_n = \mathsf{U}(Z_n^+; \omega_n)$ and $b_n = V(Z_n^+) - V(Z_n)$. ▲

▼ **Algorithm 5** (Kiefer–Wolfowitz)**.** When first-order feedback is unavailable, a popular alternative is to obtain gradient information of Φ via zeroth-order observations (Liu et al., 2019b). This idea can be traced back to the seminal work of Kiefer & Wolfowitz (1952) and the subsequent development of the simultaneous perturbation stochastic approximation (SPSA) method by Spall (1992). In our setting, this leads to the recursion:

$$\begin{aligned} V_n &= \pm(d/\delta_n)\, \Phi(Z_n + \delta_n \omega_n)\, \omega_n, \\ Z_{n+1} &= Z_n + \gamma_n V_n, \end{aligned} \tag{SPSA}$$

where $\delta_n \searrow 0$ is a vanishing "sampling radius" parameter, ω_n is drawn uniformly at random from the composite basis $\Omega = \mathcal{E}_\mathcal{X} \cup \mathcal{E}_\mathcal{Y}$ of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and the "$\pm$" sign is equal to -1 if $\omega_n \in \mathcal{E}_\mathcal{X}$ and $+1$ if $\omega_n \in \mathcal{E}_\mathcal{Y}$. Viewed this way, the interpretation of (SPSA) as a Robbins–Monro method is immediate; furthermore, a straightforward calculation (that we defer to Appendix B.3) shows that the sequence of gradient estimators V_n in (SPSA) has $B_n = \mathcal{O}(\delta_n)$ and $\sigma_n^2 = \mathcal{O}(1/\delta_n^2)$. ▲

Further examples that can be cast in the RM framework include the negative momentum method (Gidel et al., 2019b), generalized OG schemes (Mokhtari et al., 2019b), the Chambolle-Pock algorithm (Chambolle & Pock, 2011), the "prediction method" of Yadav et al. (2017), and centripetal acceleration (Peng et al., 2020); the analysis is similar and we omit the details. Certain scalable second-order methods can also be viewed as RM schemes, but the driving vector field V is no longer the gradient field of Φ; we discuss this in the supplement.

3.3. Alternating updates and moving averages

There are two extremely common heuristics for practitioners in applying min-max algorithms to real applications: alternating and averaging. An *alternating* algorithm for (SP) updates the x and y variables sequentially (instead of simultaneously as in Section 3.2). An *averaged* algorithm takes the next state as a convex combination of Z_n and Z_{n+1} in (RM), cf. Karras et al. (2018).

An important feature of our framework is that it captures alternating and averaged algorithms in a seamless manner. Indeed, introducing alternating updates or a moving average in RM schemes results in another RM scheme:

Lemma 1. *Let $Z_{n+1} = Z_n + \gamma_n[V(Z_n) + W_n]$ be an RM scheme where $W_n = U_n + b_n$ as in (4). Then its α-averaged version (where $0 < \alpha < 1$), defined as*

$$\begin{aligned} Z'_{n+1} &= Z_n + \gamma_n[V(Z_n) + W_n], \\ Z_{n+1} &= \alpha Z'_{n+1} + (1-\alpha) Z_n, \end{aligned} \tag{avg-RM}$$

is also an RM scheme: $Z_{n+1} = Z_n + \alpha\gamma_n[V(Z_n) + W_n]$.

Remark 3.1. Lemma 1 can be easily adapted to the scenario where one only averages either the X_n or Y_n variable.

Lemma 2. *Let $Z_{n+1} = Z_n + \gamma_n[V(Z_n) + W_n]$ be an RM scheme where $W_n = U_n + b_n$ as in (4). Then its alternating version, defined as*

$$\begin{aligned} X_{n+1} &= X_n + \gamma_n[V_x(X_n, Y_n) + W_{x,n}], \\ Y_{n+1} &= Y_n + \gamma_n\,[V_y(X_{n+1}, Y_n) + W_{y,n}], \end{aligned} \tag{alt-RM}$$

is also an RM scheme: $Z_{n+1} = Z_n + \gamma_n[V(Z_n) + U_n + b'_n]$ where

$$b'_n = b_n + \begin{bmatrix} 0 \\ V_y(X_{n+1}, Y_n) - V_y(X_n, Y_n) \end{bmatrix}.$$

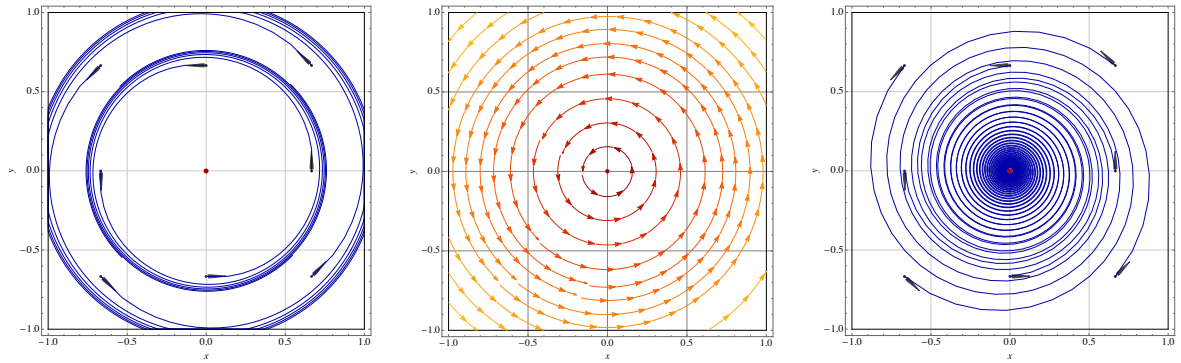

Figure 1: Comparison of different RM schemes for bilinear games $\Phi(x, y) = xy$, $x, y \in \mathbb{R}$. From left to right: (*a*) gradient descent/ascent; (*b*) the mean dynamics (MD); (*c*) extra-gradient.

Remark 3.2. One can easily generalize Lemma 2 to the "(k_1, k_2)-RM schemes" where one performs k_1 updates for x and then k_2 updates for y (here $k_1, k_2 \in \mathbb{N}$ are arbitrary but fixed). The resulting scheme will still be an RM scheme. In particular, our framework captures the popular $(k_1, k_2) = (1, 5)$ variant of (SGDA) used in the seminal works of Goodfellow et al. (2014) and Arjovsky et al. (2017). In view of Lemmas 1–2, Remark 3.2, and a simple calculation (see (B.18)), *all of our results on first-order methods (e.g., Algorithms 1–4) also apply to their averaging/alternating and the more general (k_1, k_2) versions.*

4. Convergence analysis

4.1. Overview: Continuous vs. discrete time

The key in providing a unified treatment of all algorithms in Section 3 is the reduction of (RM) to the *mean dynamics*

$$\dot{z}(t) = V(z(t)). \tag{MD}$$

To see why (MD) can capture the limiting behavior of a vast family of RM schemes beyond GDA, let us illustrate the high-level intuition on the deterministic version of Algorithm 3 ($U_n = 0$).

Since Φ and V are assumed to be Lipschitz (say with constants M and L), we see that the bias term in Algorithm 3 satisfies

$$\begin{aligned}\|b_n\| = \|V(Z_n^+) - V(Z_n)\| &\leq L\|Z_n^+ - Z_n\| \\ &= \gamma_n L\|V(Z_n)\| \leq \gamma_n LM = \mathcal{O}(\gamma_n).\end{aligned}$$

As a result, we can rewrite Algorithm 3 as

$$\frac{Z_{n+1} - Z_n}{\gamma_n} = V(Z_n) + \mathcal{O}(\gamma_n). \tag{6}$$

If $\gamma_n \searrow 0$, we should then expect (6) to converge to (MD). More generally, if the error term W_n in (RM) is sufficiently well-behaved, we should expect the iterates of (RM) and the solutions of (MD) to eventually come together.

Connecting (RM) to (MD) has proved very fruitful when the latter comprises a *gradient system*, i.e., $V = -\nabla f$ for some (possibly non-convex) $f\colon \mathcal{Z} \to \mathbb{R}$: Modulo mild assumptions, the systems (RM) and (MD) are known to both converge to the critical set of f (Benveniste et al., 1990; Bertsekas & Tsitsiklis, 2000; Kushner & Clark, 1978; Kushner & Yin, 1997; Ljung, 1977).

On the other hand, bona fide min-max problems are considerably more involved. The most widely known illustration is given by the bilinear objective $\Phi(x, y) = xy$: in this case (see Fig. 1), the trajectories (MD) comprise periodic orbits of perfect circles centered at the origin (the unique critical point of Φ; cf. Mertikopoulos et al., 2018). However, the behavior of different RM schemes can vary wildly, even in the absence of noise ($\sigma = 0$): trajectories of (SGDA) spiral outwards, each converging to an initialization-dependent periodic orbit; instead, (SEG) trajectories spiral inwards, eventually converging to the solution $z^* = (0, 0)$.

This particular difference between gradient and extra-gradient schemes has been well-documented in the literature (Daskalakis et al., 2018; Gidel et al., 2019a; Mertikopoulos et al., 2019). More pertinent to our theory, it also raises several key questions:

1. *What is the precise link between RM methods and the mean dynamics* (MD)*?*
2. *When does* (MD) *yield accurate predictions for the long-run behavior of an RM method?*

Below, we devote Sections 4.2–4.3 to the first question, and Section 4.4 to the second.

4.2. Connecting (RM) to (MD)

We begin by introducing a measure of "closeness" between the iterates of (RM) and the solution orbits of (MD). To do so, let $\tau_n = \sum_{k=1}^{n} \gamma_k$ denote the "effective time" that has elapsed at the n-th iteration of (RM), and define the

continuous-time interpolation $Z(t)$ of Z_n as

$$Z(t) = Z_n + \frac{t-\tau_n}{\tau_{n+1}-\tau_n}(Z_{n+1}-Z_n) \tag{7}$$

for all $t \in [\tau_n, \tau_{n+1}]$, $n \geq 1$. To compare $Z(t)$ to the solution orbits of (MD), we will further consider the *flow* $\Theta\colon \mathbb{R}_+ \times \mathcal{Z} \to \mathcal{Z}$ of (MD), which is simply the orbit of (MD) at time $t \in \mathbb{R}_+$ with an initial condition $z(0) = z \in \mathcal{Z}$. We then have the following notion of "asymptotic closeness":

Definition 1. $Z(t)$ is an *asymptotic pseudotrajectory* (APT) of (MD) if, for all $T > 0$, we have:

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|Z(t+h) - \Theta_h(Z(t))\| = 0. \tag{8}$$

This comparison criterion is due to Benaïm & Hirsch (1996) and it plays a central role in our analysis. In words, it simply posits that $Z(t)$ eventually tracks the flow of (MD) with arbitrary accuracy over windows of arbitrary length; as a result, if Z_n is an APT of (MD), it is reasonable to expect its behavior to be closely correlated to that of (MD).

Our first result below makes this link precise. Consider an RM scheme which satisfies

$$B_n \to 0 \text{ (a.s.)} \quad \text{and} \quad \textstyle\sum_{n=1}^{\infty} \mathbb{E}[\gamma_n B_n] < \infty, \tag{A1}$$
$$\textstyle\sum_{n=1}^{\infty} \mathbb{E}[\gamma_n^2(1 + B_n^2 + \sigma_n^2)] < \infty. \tag{A2}$$

We then have:

Theorem 1. *Suppose that Assumptions (A1)–(A2) hold. Then Z_n is an APT of (MD) w.p.*1.

4.3. Applications and examples

Of course, applying Theorem 1 to a specific algorithm (e.g., as in Section 3) would first require verifying Assumptions (A1)–(A2). However, even though the noise $\mathsf{U}(z;\omega)$ in (SFO) is assumed zero-mean and finite-variance, this *does not imply* that the error term $W_n = U_n + b_n$ in Algorithms 2–5 enjoys the same guarantees. For example, the RM representation of Algorithms 2–4 has non-zero bias, while Algorithm 5 has non-zero bias *and* unbounded variance (the latter behaving as $\mathcal{O}(1/\delta_n^2)$ with $\delta_n \to 0$).

In the following proposition we prove that Algorithms 1–5 generate asymptotic pseudotrajectories of (MD) for the typical range of hyperparameters used to ensure almost sure convergence of stochastic first-order methods.

Proposition 1. *Let Z_n be a sequence generated by any of the Algorithms 1–5. Assume further that*

a) For first-order methods (Algorithms 1–4), the algorithm is run with SFO feedback satisfying (2) and a step-size γ_n such that $A/n \leq \gamma_n \leq B/\sqrt{n(\log n)^{1+\varepsilon}}$ for some $A, B, \varepsilon > 0$.

b) For zeroth-order methods (Algorithm 5), the algorithm is run with parameters γ_n and δ_n such that $\lim_n(\gamma_n + \delta_n) = 0$, $\sum_n \gamma_n = \infty$, and $\sum_n \gamma_n^2/\delta_n^2 < \infty$ (e.g., $\gamma_n = 1/n$, $\delta_n = 1/n^{1/3}$).

Then Z_n is almost surely an APT of (MD).

4.4. The limit sets of RM schemes

The APT results in Sections 4.2–4.3 can be heuristically interpreted as: "RM schemes eventually behave as some orbits of (MD)." We now further ask: What are *the* candidate limit orbits of (MD) for RM schemes?

To shed some light on the question, let us recall that, in nonconvex *minimization* problems, stochastic gradient descent (SGD) enjoys the following properties:

(I) SGD converges to the function's set of critical points (Bertsekas & Tsitsiklis, 2000; Ljung, 1977).

(II) SGD avoids unstable critical points (Ge et al., 2015; Mertikopoulos et al., 2020; Pemantle, 1990).

This leads to the following "law of the excluded middle": generically, the only solution candidates left for SGD are stable critical points, i.e., the local minimizers of the problem's minimization objective.

In the remaining of this section, we will assimilate (I) and (II) in the context of RM schemes applied to (SP).

4.4.1. The long-run limit of RM schemes

We first focus on generalizing (I) for min-max optimization. To proceed, recall first that critical points alone cannot capture the broad spectrum of algorithmic behaviors when (MD) is not a gradient system: already in Fig. 1 we see a critical point surrounded by *spurious* periodic orbits. In addition, in dynamical systems many other spurious convergence phenomena are known, such as homoclinic loops, limit cycles, or chaos. To account for this considerably richer landscape, we will need some definitions from the theory of dynamical systems.

Definition 2 (Benaïm, 1999)**.** Let $\mathcal{S}$ be a nonempty compact subset of $\mathcal{Z}$. Then:

a) $\mathcal{S}$ is *invariant* if $\Theta_t(\mathcal{S}) = \mathcal{S}$ for all $t \in \mathbb{R}$.

b) $\mathcal{S}$ is *attracting* if it is invariant and there exists a compact neighborhood $\mathcal{K}$ of $\mathcal{S}$ such that $\lim_{t\to\infty} \operatorname{dist}(\Theta_t(z), \mathcal{S}) = 0$ uniformly in $z \in \mathcal{K}$.

c) $\mathcal{S}$ is *internally chain-transitive* (ICT) if it is invariant and $\Theta|_{\mathcal{S}}$ admits no proper attractors in $\mathcal{S}$.

Remark. Equivalently, ICT sets can be viewed as "minimal connected periodic orbits up to arbitrarily small numerical errors", cf. Benaïm (1999, Prop. 5.3). The definition above is more convenient to work with because it provides the key

insights in Section 4.4.3 below.

Our next result shows that, with probability 1, all limit points of (RM) lie in these "approximate periodic orbits":

Theorem 2. *If Assumptions (A1)–(A2) hold, then Z_n converges almost surely to an ICT set of Φ.*

Corollary 1. *Let Z_n be a sequence generated by any of the Algorithms 1–5 with parameters as in Proposition 1. Then Z_n converges almost surely to an ICT set of Φ.*

4.4.2. Avoidance of unstable points and sets

Our next result provides an avoidance result for RM schemes in min-max optimization. In analogy with function minimization problems, we will focus on unstable *invariant sets* of (MD), i.e., invariant sets that admit a nontrivial unstable manifold (for an in-depth discussion and precise definition, see Shub, 1987 and Appendix C.1).

In generic minimization problems, these are precisely the sets of strict saddle points of the function being minimized. However, since general min-max problems do *not* comprise a gradient system, (MD) could exhibit a plethora of unstable sets, not containing any stationary points of Φ (e.g., periodic orbits, heteroclinic networks, etc.). On account of the above, our result below is stated in terms of invariant *sets* – and not only points. For convenience, we will assume that V is C^2 and γ_n is as in Proposition 1.

Theorem 3. *Let $\mathcal{K}$ be an unstable invariant set of (MD) (e.g., an unstable critical point or unstable periodic orbit). Assume further that the noise in (RM) satisfies: (i) $\sup_n \|U_n\| < \infty$ with probability 1; and (ii) $\inf_{z:\|z\|=1} \mathbb{E}[\langle U_n, z\rangle_+ \,|\, \mathcal{F}_n] > 0$. Then the sequence Z_n generated by any of the Algorithms 1–4 satisfies*

$$\mathbb{P}\left(\lim_{n\to\infty} \operatorname{dist}(Z_n, \mathcal{K}) = 0\right) = 0.$$

Remark 4.1. We note that Assumptions (*i*) and (*ii*) above are standard in the literature for avoidance results of SGD (Benaïm, 1999; Mertikopoulos et al., 2020; Pemantle, 1990), and are significantly lighter than other "isotropic noise" assumptions that are common in the literature (Ge et al., 2015). Specifically, even though Assumption (*ii*) looks somewhat obscure, it only posits that the noise is not degeneratively equal to zero along certain directions in space; for a more detailed discussion, see Appendix C.1. We also stress that neither of these assumptions is required for the rest of our paper.

4.4.3. When do RM schemes behave the same?

So far, we have successfully generalized (I) and (II) to the context of (SP) as follows:[1]

(I-SP) RM schemes always converge to ICT sets, and

(II-SP) RM schemes always avoid invariant sets.

Nonetheless, (I-SP) and (II-SP) still fail to explain the distinct behaviors of RM schemes in bilinear objectives: Why does SGDA converge only to periodic orbits, while deterministic SEG only to critical points? Or, more generally,

> *Are different RM schemes more likely to exhibit different convergence topologies – e.g., cycles vs. critical points – in generic min-max problems?*

Our next result takes a closer look at *attracting* ICT sets and provides a generically negative answer to this question. To set the stage, suppose we want to apply (I-SP) to the bilinear objective $\Phi(x, y) = xy$. Stricto sensu, (I-SP) does not apply in this case since Φ is not Lipschitz. However, Fig. 1(*b*) shows (and we rigorously prove in Appendix C.2) that *any* tuple $(x, y) \in \mathbb{R}^2$ belongs to an ICT set of Φ, so Theorem 2 holds trivially. This in turn implies that *the only attractor for Φ is trivially the whole space* $\mathbb{R}^2$, since Definition 2-*b*) is never satisfied for any $\mathcal{S} \subsetneq \mathbb{R}^2$.

Importantly, the celebrated *Kupka-Smale theorem* (Kupka, 1963; Smale, 1963) asserts that systems with degenerate periodic orbits (such as bilinear games) occur "almost never" in the Baire category sense. More precisely, an arbitrarily small perturbation can fundamentally destroy the topological properties of their ICT sets and give rise to proper, non-trivial attractors; cf. Example 5.1. In contrast, systems with nontrivial attractors are known to be robust under perturbations (Shub, 1987), and our final result in the section shows that it is precisely the *existence of nontrivial attractors* that makes the discrepancy of RM schemes disappear (at least locally).

Theorem 4. *Let $\mathcal{S}$ be an attractor of (MD) and fix some confidence level $\alpha > 0$. If γ_n is small enough (cf. Appendix C.3 for the precise bound) and Assumptions (A1)–(A2) hold, there exists a neighborhood $\mathcal{U}$ of $\mathcal{S}$, independent of α, such that*

$$\mathbb{P}(Z_n \text{ converges to } \mathcal{S} \mid Z_1 \in \mathcal{U}) \geq 1 - \alpha. \tag{9}$$

Corollary 2. *Let Z_n be a sequence generated by any of the Algorithms 1–5 with sufficiently small γ_n satisfying the conditions of Proposition 1. If $Z_1 \in \mathcal{U}$, then $\mathbb{P}(Z_n \text{ converges to } \mathcal{S}) \geq 1 - \alpha$.*

In short, Theorem 4 asserts that any non-degenerate ICT set dictates the local convergence of *all* RM schemes under the general Assumptions (A1)–(A2).

On a positive note, since the Hartman-Grobman Theorem (Robinson, 1998) implies that all critical points of Φ with $\Re\{\lambda(JV(z^*))\} < 0$ for all engenvalues λ are attractors of (MD), Theorem 4 immediately yields:

[1] To see why this is really a generalization, simply note that the only ICT sets of $V = -\nabla f$ are connected critical points of f; cf. Proposition C.1.

Corollary 3. *Let z^* be a critical point of Φ such that $\Re\{\lambda(JV(z^*))\} < 0$ for all engenvalues of $JV(z^*)$. Then all RM schemes satisfying Assumptions (A1)–(A2) locally converge to z^* with high probability.*

Corollary 3 generalizes the local convergence of deterministic SGDA and SEG studied by Daskalakis & Panageas (2018). It also extends (Hsieh et al., 2019a, Theorem 5) from (OG/PEG) to all generalized RM schemes.

On the flip side, however, Theorem 4 also bears an undesirable consequence: it implies that many RM schemes designed to improve SGDA (e.g., Algorithms 2–4) may in fact be trapped by *spurious ICT sets* in exactly the same way as SGDA. Thus, even though many of these algorithms have been motivated by their appealing properties in bilinear games, it is not clear whether they offer any significant advantages beyond the convex-concave case. We examine this issue in detail in the next section.

5. Spurious attractors: Illustrations and examples

In this last section, we provide a range of simple examples that exhibit *spurious attractors* – i.e., attractors that consist entirely of non-critical points. For illustration purposes, we focus on the simple case $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ with polynomial objectives. In doing this, our goal is to highlight a number of issues that can arise in min-max optimization problems; whether limit cycles of this type occur in actual large-scale experiments – e.g., in GANs – is an open research question (Letcher, 2020).

▼ **Example 5.1** (Almost bilinear $\not\approx$ bilinear, instability $\not\approx$ escape)**.** Consider an arbitrarily small perturbation of a bilinear game:

$$\Phi(x, y) = xy + \varepsilon\phi(y), \tag{10}$$

where $\varepsilon > 0$ and $\phi(y) = \frac{1}{2}y^2 - \frac{1}{4}y^4$. There is an unstable critical point at the origin; further, Lemma D.1 asserts, for small ε, the existence of an *attracting* ICT set $\mathcal{S}$ in a neighborhood of the circle $\{z : \|z\|^2 = 4/3\}$. By Corollary 2, any RM scheme of Section 3 thus gets trapped by $\mathcal{S}$; see Fig. 2(*a*) for an illustration for (SEG).

This example brings two issues of existing studies to light. First, it shows that "almost bilinear games" can still trap many methods for solving exact bilinear games. Second, in contrast to minimization problems, the region around an unstable critical point can in fact be fully stable. Thus, one has to be careful when interpreting algorithms that "locally avoid unstable critical points", since they might be incapable of escaping their neighborhoods. ▲

▼ **Example 5.2** ("Forsaken" $\min\max$ solutions)**.** Suppose we apply Algorithms 1–5 to the objective

$$\Phi(x, y) = x(y - 0.45) + \phi(x) - \phi(y) \tag{11}$$

where $\phi(z) = \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$. This problem has a desirable $(x^*, y^*) \simeq (0.08, 0.4)$. However, as we show in Appendix D.2, there exist *two* spurious limit cycles that do not contain *any* critical point of Φ. Worse, the limit cycle closer to the solution is *unstable* and repels any trajectory that comes close to the solution; see Fig. 2(*b*) for an illustration for (SEG). As a result, the "shielded" solution is highly unlikely to be discovered by existing algorithms, even though it is perfectly stable. ▲

We conclude the paper by further examining several important settings that are not covered by our theory:

Ergodic averages. Instead of the "moving average" in Lemma 1, one can take the *ergodic average* ($Z'_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$) as is customary in convex-concave problems (Juditsky et al., 2011; Nemirovski, 2004a). We plot one such trajectory in Fig. 2 (the blue curves). Evidently, we see that ergodic average can force the algorithms to halt at non-critical points, and this convergence is by no means min-max optimal.

Scalable second-order methods. Many recent works attempt to address the cycling issues of min-max algorithms via incorporating *second-order* oracles. For completeness, we also study a range of popular second-order methods in Appendix D.3. Our analysis shows that these algorithms suffer similar symptoms as first-order schemes in our examples, cf. Figs. 5–6 in the supplement.

Constant step-size implementations. In addition to the diminishing step-size policies studied here, another common strategy in practice is to simply set γ_n to a *constant step-size*. While our analysis does not cover this setting, there exist several techniques in stochastic approximation to boost from our "almost surely" statements for $\gamma_n \searrow 0$ to *concentration* or *high-probability* results when $\gamma_n \equiv \gamma$ is small (Borkar, 2008; Kushner & Huang, 1981; Kushner & Yin, 1997).

For completeness, in Section D.4 we examine various constant step-size RM schemes applied to (10) and (11). The outcome coincides with our intuition that these schemes should concentrate around the spurious attracting ICT sets, and hence exhibit similar behaviors as RM schemes with $\gamma_n \searrow 0$; see Fig. 7.

Adaptive methods. Adaptive methods such as Adam (Kingma & Ba, 2014) are ubiquitous in GAN training. We study such methods in Section D.5 and provide an illustration in Fig. 3: our results show that they fail to solve the simple objectives (10) and (11). Moreover, some methods

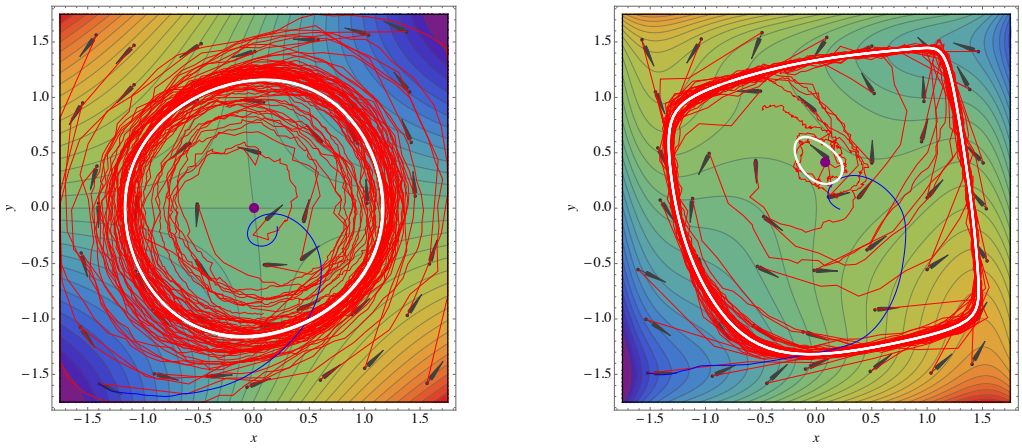

Figure 2: Spurious limits of min-max optimization algorithms. From left to right: (*a*) (SEG) for (10) with $\varepsilon = 0.01$; (*b*) "forsaken solutions" of (SEG); The red curves present trajectories with different initialization; non-critical ICT sets are depicted in white; the blue curves represent an time-averaged sample orbit.

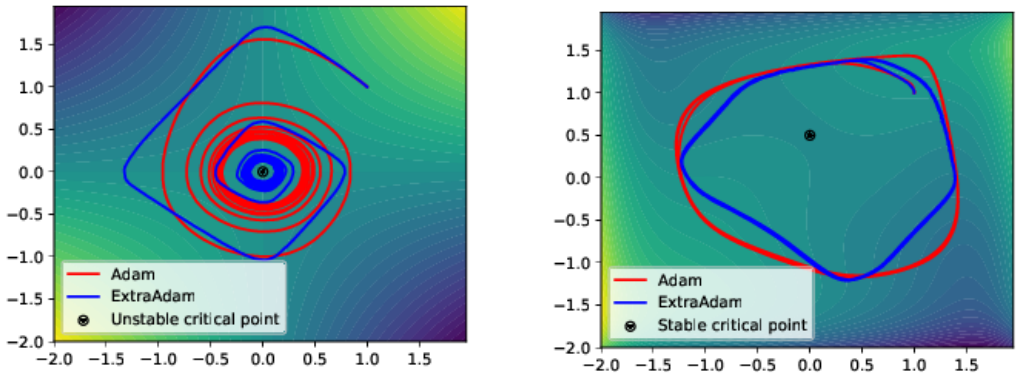

Figure 3: The limits of adaptive algorithms in (*a*) almost bilinear games, as per (10); and (*b*) games with a forsaken solution, as per (11).

even show a potentially detrimental tendency of converging to *max-min points*, the exact opposite of desirable solutions; see Fig. 3.

6. Concluding remarks

The generalized RM template captures a wide array of existing min-max algorithms, and the machinery of stochastic approximation allows us to derive a series of new convergence results, both desirable and undesirable. Our numerical experiments suggest that spurious limits also arise in a range of other algorithms and practical tweaks that are not covered by our theory; providing a rigorous theoretical statement and proof of these last observations would be a fruitful direction for future research.

In closing, we should also clarify that these illustrations are *not* meant to suggest that the algorithms and practical tweaks discussed above are always doomed, or that they comprise the principal cause of failure in GAN training. However, we do believe that they constitute an important cautionary tale to the effect that, in min-max problems, *convergence does not imply optimality* – or even *stationarity*.

Acknowledgements

The authors are grateful to Thomas Pethick for his help in the numerical simulation of adaptive methods. This research was partially supported by the COST Action CA16228 "European Network for Game Theory" (GAMENET), the Army Research Office under grant number W911NF-19-1-0404, the Swiss National Science Foundation (SNSF) under grant number 200021_178865 / 1, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^o 725594 - timedata), and 2019 Google Faculty Research Award. P. Mertikopoulos is grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the grants ORACLESS and ALIAS.

References

Abernethy, J., Lai, K. A., and Wibisono, A. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 486–495, 2019.

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

Arrow, K. J., Hurwicz, L., and Uzawa, H. *Studies in linear and non-linear programming*. Stanford University Press, 1958.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pp. 354–363, 2018.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, NY, USA, 2 edition, 2017.

Benaïm, M. Dynamics of stochastic approximation algorithms. In Azéma, J., Émery, M., Ledoux, M., and Yor, M. (eds.), *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pp. 1–68. Springer Berlin Heidelberg, 1999.

Benaïm, M. and Hirsch, M. W. Dynamics of Morse-Smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6): 1005–1030, December 1995.

Benaïm, M. and Hirsch, M. W. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

Benveniste, A., Métivier, M., and Priouret, P. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Borkar, V. S. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press and Hindustan Book Agency, 2008.

Bowen, R. Omega limit sets of Axiom A diffeomorphisms. *Journal of Differential Equations*, 18:333–339, 1975.

Burkholder, D. L. Distribution function inequalities for martingales. *Annals of Probability*, 1(1):19–42, 1973.

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. Reducing noise in GAN training with variance reduced extragradient. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Chiang, C.-K., Yang, T., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., and Zhu, S. Online optimization with gradual variations. In *COLT '12: Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Christopher, C. and Li, C. *Limit cycles of differential equations*. Springer Science & Business Media, 2007.

Conley, C. C. *Isolated Invariant Set and the Morse Index*. American Mathematical Society, Providence, RI, 1978.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pp. 9236–9246, 2018.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with optimism. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

Daskalakis, C., Skoulakis, S., and Zampetakis, M. The complexity of constrained min-max optimization. *arXiv preprint arXiv:2009.09623*, 2020.

Domingo-Enrich, C., Jelassi, S., Mensch, A., Rotskoff, G., and Bruna, J. A mean-field analysis of two-player zero-sum games. *arXiv preprint arXiv:2002.06277*, 2020.

Facchinei, F. and Pang, J.-S. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

Fiez, T. and Ratliff, L. Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation. *arXiv preprint arXiv:2009.14820*, 2020.

Fiez, T., Chasnov, B., and Ratliff, L. J. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Flokas, L., Vlatakis-Gkaragkounis, E. V., and Piliouras, G. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Flokas, L., Vlatakis-Gkaragkounis, E.-V., and Piliouras, G. Solving min-max optimization with hidden structure via gradient descent ascent. *arXiv preprint arXiv:2101.05248*, 2021.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points — Online stochastic gradient for tensor decomposition. In *COLT '15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019a.

Gidel, G., Hemmat, R. A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1802–1811, 2019b.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

Grimmer, B., Lu, H., Worah, P., and Mirrokni, V. The landscape of nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*, 2020a.

Grimmer, B., Lu, H., Worah, P., and Mirrokni, V. Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems. *arXiv preprint arXiv:2010.10628*, 2020b.

Hall, P. and Heyde, C. C. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural*

information processing systems, pp. 6626–6637, 2017.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 6936–6946, 2019a.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pp. 2810–2819, 2019b.

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

Jin, C., Netrapalli, P., and Jordan, M. I. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Kamalaruban, P., Huang, Y.-T., Hsieh, Y.-P., Rolland, P., Shi, C., and Cevher, V. Robust reinforcement learning via adversarial training with langevin dynamics. *arXiv preprint arXiv:2002.06063*, 2020.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

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

Kupka, I. Contributiona la théorie des champs génériques. *Contributions to differential equations*, 2:457–484, 1963.

Kushner, H. J. and Clark, D. S. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer, 1978.

Kushner, H. J. and Huang, H. Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105, 1981.

Kushner, H. J. and Yin, G. G. *Stochastic approximation algorithms and applications*. Springer-Verlag, New York, NY, 1997.

Lee, J. M. *Introduction to Smooth Manifolds*. Number 218 in Graduate Texts in Mathematics. Springer-Verlag, New York, NY, 2003.

Letcher, A. On the impossibility of global convergence in multi-loss optimization. *arXiv preprint arXiv:2005.12649*, 2020.

Liu, M., Mroueh, Y., Ross, J., Zhang, W., Cui, X., Das, P., and Yang, T. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019a.

Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and Obelilly, U.-M. Min-max optimization without gradients: Convergence and applications to adversarial ml. *arXiv preprint arXiv:1909.13806*, 2019b.

Ljung, L. Analysis of recursive stochastic algorithms. *IEEE Trans. Autom. Control*, 22(4):551–575, August 1977.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

Malitsky, Y. and Tam, M. K. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

Mazumdar, E., Ratliff, L. J., and Sastry, S. S. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

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

Mertikopoulos, P., Papadimitriou, C. H., and Piliouras, G. Cycles in adversarial regularized learning. In *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of gans. In *Advances in Neural Information Processing Systems*, pp. 1825–1835, 2017.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. https://arxiv.org/abs/1901.08511v2, 2019a.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019b.

Nagarajan, V. and Kolter, J. Z. Gradient descent gan optimization is locally stable. In *Advances in neural information processing systems*, pp. 5585–5595, 2017.

Naveiro, R. and Insua, D. R. Gradient methods for solving stackelberg games. In *International Conference on Algorithmic DecisionTheory*, pp. 126–140. Springer, 2019.

Nemirovski, A. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004a.

Nemirovski, A. S. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004b.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Number 87 in Applied Optimization. Kluwer Academic Publishers, 2004.

Pearlmutter, B. A. Fast exact multiplication by the hessian. *Neural computation*, 6(1):147–160, 1994.

Pemantle, R. Nonconvergence to unstable points in urn models and stochastic aproximations. *Annals of Probability*, 18(2):

698–712, April 1990.

Pemantle, R. Vertex-reinforced random walk. *Probability Theory and Related Fields*, 92:117–136, 1992.

Peng, W., Dai, Y.-H., Zhang, H., and Cheng, L. Training gans with centripetal acceleration. *Optimization Methods and Software*, pp. 1–19, 2020.

Phelps, R. R. *Convex Functions, Monotone Operators and Differentiability*. Lecture Notes in Mathematics. Springer-Verlag, 2 edition, 1993.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017.

Popov, L. D. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Robinson, C. *Dynamical systems: stability, symbolic dynamics, and chaos*. CRC press, 1998.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14 (5):877–898, 1976.

Schäfer, F. and Anandkumar, A. Competitive gradient descent. In *Advances in Neural Information Processing Systems*, pp. 7623–7633, 2019.

Shub, M. *Global Stability of Dynamical Systems*. Springer-Verlag, Berlin, 1987.

Smale, S. Stable manifolds for differential equations and diffeomorphisms. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 17(1-2):97–116, 1963.

Spall, J. C. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control*, 37(3):332–341, March 1992.

Teschl, G. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

Wiggins, S. *Introduction to applied nonlinear dynamical systems and chaos*, volume 2. Springer Science & Business Media, 2003.

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

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S. P., and Glynn, P. W. Stochastic mirror descent for variationally coherent optimization problems. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S. P., and Glynn, P. W. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020.