The limits of min-max optimization algorithms: Convergence to spurious non-critical sets

Ya-Ping Hsieh, Panayotis Mertikopoulos, Volkan Cevher

THE LIMITS OF MIN-MAX OPTIMIZATION ALGORITHMS: CONVERGENCE TO SPURIOUS NON-CRITICAL SETS

YA-PING HSIEH*, PANAYOTIS MERTIKOPOULOS◇,♯, AND VOLKAN CEVHER*

Abstract. Compared to minimization problems, the min-max landscape in machine learning applications is considerably more convoluted because of the existence of cycles and similar phenomena. Such oscillatory behaviors are well-understood in the convex-concave regime, and many algorithms are known to overcome them. In this paper, we go beyond the convex-concave setting and we characterize the convergence properties of a wide class of zeroth-, first-, and (scalable) second-order methods in non-convex/non-concave problems. In particular, we show that these state-of-the-art min-max optimization algorithms may converge with arbitrarily high probability to attractors that are in no way min-max optimal or even stationary. Spurious convergence phenomena of this type can arise even in two-dimensional problems, a fact which corroborates the empirical evidence surrounding the formidable difficulty of training GANs.

1. Introduction

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

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

where $\mathcal{X}$, $\mathcal{Y}$ are subsets of a Euclidean space and $\Phi\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ may be non-convex/non-concave. Given an algorithm for solving (SP), the following fundamental questions arise:

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

The goal of this paper is to provide concrete answers to ($\star$) and to study their practical implications for a wide array of existing methods.

Min-max problems of this type have found widespread applications in machine learning in the context of generative adversarial networks (GANs) [32], robust reinforcement learning [72], and other models of adversarial training [51]. In this broad setting, it has become empirically clear that the joint training of two neural networks (NNs) with competing objectives is fundamentally more difficult than training a *single* NN of similar size and architecture. The latter task boils down to successfully finding a (good) local minimum

* LIONS, École Polytechnique Fédérale de Lausanne (EPFL).
◇ Univ. Grenoble Alpes, CNRS, Inria, LIG, 38000, Grenoble, France.
♯ Criteo AI Lab.
E-mail addresses: ya-ping.hsieh@epfl.ch, panayotis.mertikopoulos@imag.fr, volkan.cevher@epfl.ch.

2020 *Mathematics Subject Classification.* Primary 90C47, 91A26, 62L20; secondary 90C26, 91A05, 37N40.
Key words and phrases. Min-max optimization; internally chain transitive sets; Robbins-Monro algorithms; spurious attractors.
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° 725594 - time-data), and 2019 Google Faculty Research Award. P. Mertikopoulos is also grateful for financial support by the French National Research Agency (ANR) under grant no. ANR–16–CE33–0004–01 (ORACLESS).

of a non-convex function, so it is instructive to revisit ($\star$) in the context of (non-convex) minimization problems.

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

(1) Bounded trajectories of SGD always converge to a set of critical points [12, 49, 50].

(2) The limits of SGD do not contain saddle points or other spurious solutions [15, 29, 68].

At first glance, these positive results might raise high expectations for solving (SP). Unfortunately, one can easily find counterexamples with very simple *bilinear* games of the form $\Phi(x,y) = x^\top Ay$: naïvely applying stochastic gradient descent/ascent (SGDA) methods in this case leads to recurrent orbits that do not contain *any* critical point of Φ. Such a phenomenon has no counterpart in non-convex minimization, and is fundamentally tied to the min-max structure of (SP).

The failure of SGDA in bilinear games has been studied extensively [1, 4, 30, 31, 46, 56, 59, 69, 77, 81, 82], leading to more sophisticated schemes such as stochastic extra-gradient (SEG) methods and their variants [19, 25, 30, 36, 57]. Meanwhile, to bypass such globally oscillatory issues, another thread of research [2, 24, 34, 37, 47, 53, 54, 57, 60, 66, 74] has shifted its attention to *local analysis*. Essentially, these works either analyze the algorithmic behaviors only "sufficiently close" to critical points, or impose stringent assumptions on Φ (such as "coherence" [57] or the existence of solutions to a Minty variational inequality [47]) to ensure the equivalence between global and local convergence.

Although these studies have certainly led to fruitful results, the realm beyond bilinear games and (locally) idealized objectives remains somewhat unexplored (with a few exceptions that we discuss in detail below). In particular, a convergence theory for general non-convex/non-concave problems is still lacking.

Our contributions. In this paper, we aim to bridge this gap by providing precise answers to ($\star$) for a wide range of min-max optimization algorithms that can be seen as *generalized Robbins–Monro (RM) schemes* [76]. Mirrorring the minimization perspective, we prove that, for any such algorithm $\mathcal{A}$:

(1) Bounded trajectories of $\mathcal{A}$ always converge to an *internally chain-transitive* (ICT) set.

(2) Trajectories of $\mathcal{A}$ may converge with arbitrarily high probability to spurious attractors that contain *no* critical point of Φ.

The most critical implication of our theory is that one can reduce the long-term behavior of a training algorithm to its associated ICT sets, a notion deeply rooted in the study of dynamical systems [6, 8, 9, 14, 23] that formalizes the idea of "discrete limits of continuous flows"; cf. Section 4. As an example, in minimization problems, one can prove that the ICT sets of SGD consist solely of components of critical points; on the other hand, we show that ICT sets in min-max optimization can exhibit drastically more complicated structures, even when $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. In particular, we establish the following negative results:

- An ICT set may contain (almost) *globally attracting limit cycles*, and the algorithms designed to eliminate periodic orbits in bilinear games *cannot escape them*. This observation corroborates the persistence of non-convergent behaviors in GAN training, and suggests that bilinear games may be insufficient as models for such applications.
- There exist *unstable* critical points whose neighborhood contains an (almost) *globally stable* ICT set. Therefore, in sharp contrast to minimization problems, "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, existing algorithms are repelled from a desirable solution with high probability, even if initialized arbitrarily close to it.

Finally, we provide numerical illustrations of the above, which further show that common practical tweaks (such as averaging or adaptive algorithms) also fail to address these problematic cases.

Further related work. To our knowledge, the convergence to non-critical sets in (SP) has only been systematically studied in a few settings. Besides the bilinear games alluded to above, other instances include the "almost bilinear games" [1] and deterministic gradient descent/ascent (GDA) applied to "hidden bilinear games" [28]. In contrast to these works, our framework does not impose any structural assumption and requires only mild regularity of Φ, and our results apply to many existing methods beyond (S)GDA; cf. Section 3. The generality of our approach is made possible by foundational results in dynamical systems [6, 8], which have not been exploited before in the context of min-max optimization, and have only recently been applied to learning in games with the aim of showing convergence to (local) Nash equilibria [9, 10, 13, 16, 17, 22, 53, 55, 70, 71].

Upon completion of our paper (two weeks prior to the actual submission date), we discovered a preprint by Letcher [45] whose motivation is similar to our own. The focus of [45] is on providing counterexamples that rule out the convergence of deterministic "reasonable" and "global" algorithms. There are two major distinctions that make our approaches complementary: [45] focuses on the impossibility of *desirable* convergence guarantees in a purely *determinist*c setting; in contrast, our paper focuses squarely on the occurrence of *undesirable convergence* phenomena with probability 1 in *stochastic* algorithms. Taken together, the work [45] and our own paint a fairly complete picture of the fundamental limits of min-max optimization algorithms.

2. Setup and preliminaries

We focus on general problems of the form (SP) with $\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}$, $\mathcal{Y} = \mathbb{R}^{d_\mathcal{Y}}$, and Φ assumed C^1. To ease notation, we will denote $z = (x, y)$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $d = d_\mathcal{X} + d_\mathcal{Y}$. In addition, we will write

$$V(z) \equiv (V_x(x,y), V_y(x,y)) \coloneqq (-\nabla_x \Phi(x,y), \nabla_y \Phi(x,y)) \tag{1}$$

for the (min-max) gradient field of Φ, and we will assume that V is Lipschitz. In some cases we will also require V to be C^1 and we will write $J(z)$ for its Jacobian; this additional assumption will be stated explicitly whenever invoked.

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. When V is C^1, any local solution is a *stable* critical point [37], i.e., $\nabla_x^2 \Phi(x^*, y^*) \succeq 0$ and $\nabla_y^2 \Phi(x^*, y^*) \preceq 0$.

From an algorithmic standpoint, we will focus exclusively on the black-box optimization paradigm [64] with *stochastic first-order oracle* (SFO) feedback; algorithms with a more complicated feedback structure (such as a best-response oracle [27, 37, 61]) or based on mixed-strategy sampling [26, 35] are not considered in this work. In detail, 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 models, system state observations in reinforcement

learning, etc.). Regarding this error term, we will assume throughout that it is zero-mean and sub-Gaussian:

$$\mathbb{E}[\mathsf{U}(z;\omega)] = 0 \quad \text{and} \quad \mathbb{P}(\|\mathsf{U}(z;\omega)\| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}} \tag{2}$$

for some $\sigma > 0$ and all $z \in \mathcal{Z}$. The sub-Gaussian tail assumption is standard in the literature [38, 63–65], and it can be further relaxed with little loss of generality to finite variance $\mathbb{E}[\|\mathsf{U}(z;\omega)\|^2] \leq \sigma^2$. To streamline our discussion, we will present our results in the sub-Gaussian regime and we will rely on a series of remarks to explain any modifications required for different assumptions on U.

3. Core algorithmic framework

3.1. **The Robbins–Monro template.** Much of our analysis will focus on iterative algorithms that can be cast in the abstract Robbins–Monro framework of stochastic approximation [76]:

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

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 a generalized error term (described in detail below).

(3) γ_n is the step-size (a hyperparameter, typically of the form $\gamma_n \propto 1/n^p$, $p \geq 0$).

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

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

so V_n can be seen as a noisy estimate 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 of the error 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:

$$a) \quad \textit{Bias:} \qquad B_n = \|b_n\| \tag{5a}$$

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

The precise behavior of B_n and σ_n^2 will be examined on a case-by-case basis below.

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 the general template (RM) above.

▼ **Algorithm 1** (Stochastic gradient descent/ascent)**.** The basic SGDA algorithm – also known as the *Arrow–Hurwicz* method [3] – 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** (Alternating stochastic gradient descent/ascent)**.** A common variant of SGDA, is to *alternate* the updates of the min/max variables, resulting in the *alternating stochastic gradient descent/ascent* (alt-SGDA) method:

$$\begin{aligned} X_{n+1} &= X_n + \gamma_n \mathsf{V}_x(X_n, Y_n; \omega_n) &&= X_n + \gamma_n[V_x(X_n, Y_n) + U_{x,n}] \\ Y_{n+1} &= Y_n + \gamma_n \mathsf{V}_y(X_{n+1}, Y_n; \omega_n^+) &&= Y_n + \gamma_n[V_y(X_{n+1}, Y_n) + U_{y,n}] \end{aligned} \tag{alt-SGDA}$$

where ω_n, ω_n^+ $(n = 1, 2, \dots)$ are sequences of i.i.d. random seeds, $U_{x,n} \coloneqq \mathsf{U}_x(X_n, Y_n; \omega_n)$, and $U_{y,n} \coloneqq \mathsf{U}_y(X_{n+1}, Y_n; \omega_n^+)$. The RM representation of (alt-SGDA) is obtained by taking $Z_n = (X_n, Y_n)$, $b_n = (0, V_y(X_{n+1}, Y_n) - V_y(X_n, Y_n))$, and $U_n = (U_{x,n}, U_{y,n})$. ▲

▼ **Algorithm 3** (Stochastic extra-gradient)**.** Going beyond (SGDA), the (stochastic) extra-gradient algorithm exploits the following principle [38, 41, 62]: given a "base" state Z_n, the algorithm queries the oracle at Z_n to generate a *leading* state Z_n^+ and then updates Z_n with oracle information from Z_n^+. Assuming SFO feedback as above, this process may be described as follows:

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

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 [73]:

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

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 $b_n = V(Z_n^+) - V(Z_n)$.

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 [20, 25, 36, 75]. In unconstrained min-max optimization, (OG/PEG) turns out to be equivalent to a number of other existing methods, including "extrapolation from the past" [30], reflected gradient [52], and the "prediction method" of [80]. ▲

▼ **Algorithm 5** (Kiefer–Wolfowitz)**.** When first-order feedback is unavailable, a popular alternative is to obtain gradient information of Φ via zeroth-order observations [48]. This idea can be traced back to the seminal work of Kiefer and Wolfowitz [39] and the subsequent development of the simultaneous perturbation stochastic approximation (SPSA) method by Spall [78]. 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 the supplement) shows that the sequence of gradient estimators V_n in (SPSA) has $B_n = \mathcal{O}(\delta_n)$ and $\sigma_n^2 = \mathcal{O}(1/\delta_n^2)$. ▲

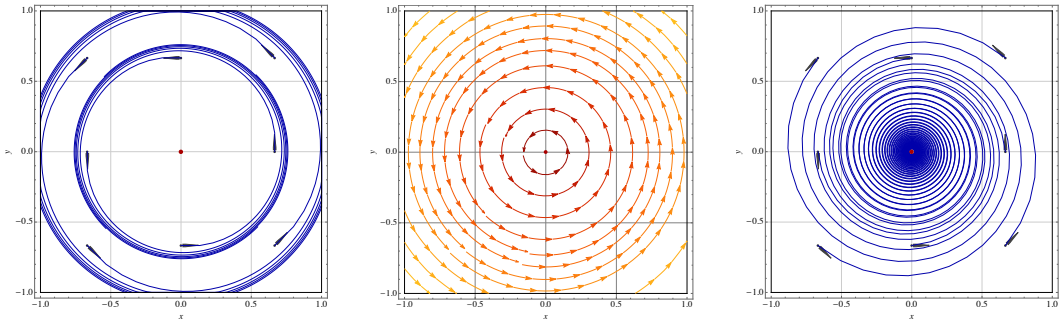


Figure 1: Comparison of different RM schemes for bilinear games $\Phi(x,y) = xy$, $x, y \in \mathbb{R}$. From left to right: (*a*) gradient descent/ascent; (*b*) the mean dynamics (MD); (*c*) extra-gradient.

Further examples that can be cast in the general framework (RM) include the negative momentum method [31], generalized OG schemes [59], and centripetal acceleration [69]; the analysis is similar and we omit the details. Certain scalable second-order methods can also be viewed as Robbins–Monro schemes, but the driving vector field V is no longer the gradient field of Φ; we discuss this in Example 5.3 and the supplement.

4. Convergence analysis

4.1. **Continuous vs. discrete time.** The main idea of our approach will be to treat (RM) as a noisy discretization of the *mean dynamics*

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

This is motivated by the fact that $\dot{z}(t)$ can be seen as the continuous-time limit of the finite difference quotient $(Z_{n+1} - Z_n)/\gamma_n$: in this way, if the error term W_n in (RM) is sufficiently well-behaved, it is plausible to expect that the iterates of (RM) and the solutions of (MD) eventually come together. This approach has proved very fruitful when the mean dynamics (MD) comprise a *gradient system*, i.e., $V = -\nabla f$ for some (possibly non-convex) $f\colon \mathcal{Z} \to \mathbb{R}$. In this case (and modulo mild assumptions), the systems (RM) and (MD) both converge to the critical set of f, see e.g., [11, 12, 42, 43, 49].

On the other hand, the min-max landscape is 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). 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, trajectories of (SEG) 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, cf. [25, 30, 57]. 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 can (MD) accurately predict the long-run behavior of an RM method?

The rest of this section is devoted to providing precise answers to these questions.

4.2. **Stochastic approximation.** 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

$$Z(t) = Z_n + \frac{t-\tau_n}{\tau_{n+1}-\tau_n}(Z_{n+1}-Z_n) \tag{6}$$

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" due to Benaïm and Hirsch [7, 8]:

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

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|Z(t+h) - \Theta_h(Z(t))\| = 0. \tag{7}$$

This comparison criterion is due to Benaïm and Hirsch [8] 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. To state it, we will make the following assumptions:

$$\lim_{n\to\infty} B_n = 0, \tag{A1}$$

$$\sum_{n=1}^{\infty} \gamma_n^2 \sigma_n^2 < \infty, \tag{A2}$$

both assumed to hold with probability 1. Under these blanket requirements, we have:

Theorem 1. *Suppose that* (RM) *is run with a step-size policy* γ_n *such that* $\sum_n \gamma_n = \infty$, $\lim_n \gamma_n = 0$, *and Assumptions* (A1)–(A2) *hold. Then, with probability* 1, *one of the following holds: a)* Z_n *is an APT of* (MD)*; or b)* Z_n *is unbounded* (*and hence, non-convergent*).

A key challenge in proving Theorem 1 is that Assumptions (A1) and (A2) allow for very general error processes W_n in (RM), including cases where W_n is non-zero-mean ($b_n \neq 0$) and/or unbounded, either with positive probability or in all its moments (e.g., $\sup_n \mathbb{E}[\|W_n\|^q] = \infty$ for all $q \geq 2$). Because of this, earlier foundational results on asymptotic pseudotrajectories [6, 8] do not apply, and we need to employ a series of direct (sub)martingale convergence arguments to control the quadratic variation of Z_n. The precise argument relies on a pathwise version of the Burkholder–Davis–Gundy (BDG) maximal inequality [33], but the details are fairly involved so we defer them to the supplement.

4.3. **Applications and examples.** Applying Theorem 1 requires verifying Assumptions (A1) and (A2) for the algorithmic framework of Section 3. However, even though the noise $\mathsf{U}(z;\omega)$ in (SFO) is assumed zero-mean and sub-Gaussian, this *does not imply* that the generalized error term $W_n = U_n + b_n$ in Algorithms 1–5 enjoys the same guarantees. For example, the RM representation of Algorithms 2–4 has non-zero bias, while Algorithm 5 exhibits both non-zero bias *and* unbounded variance (the latter behaving as $\mathcal{O}(1/\delta_n^2)$ with $\delta_n \to 0$ as $n \to \infty$).

In the following proposition we prove that, for a wide range of parameters, Algorithms 1–5 indeed generate asymptotic pseudotrajectories of (MD).

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* $A/n \leq \gamma_n \leq B/(\log n)^{1+\varepsilon}$ *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, with probability 1, *one of the following holds: a)* Z_n *is an APT of* (MD)*; or b)* Z_n *is unbounded.*

Remark 4.1. We note that the requirements for (SFO) are closely linked to the assumptions for γ_n: for instance, one can remove the sub-Gaussian tail and impose only that $\mathsf{U}(z;\omega)$ in (SFO) is bounded in L^q for some $q \geq 2$, and the conclusion of Proposition 1 still holds as long as $\sum_n \gamma_n^{1+q/2} < \infty$.

We conclude this discussion with a remark on the boundedness clause for Z_n in Theorem 1 and Proposition 1. Clearly, if Z_n is unbounded, it cannot converge to a solution of (SP), so we need not go further in examining the failure of (RM) as a solution method. Still, for completeness, we provide in the supplement a coercivity condition for Φ which guarantees that Z_n is bounded with probability 1.

4.4. **Convergence analysis.** To proceed, it is important to recall 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 an ensemble of periodic orbits. To account for this considerably richer landscape, we will need some more notions from the theory of dynamical systems:

Definition 2. Let $\mathcal{S}$ be a nonempty compact subset of $\mathcal{Z}$. We then say that:

a) $\mathcal{S}$ is *invariant* if $\Theta_t(\mathcal{S}) \subseteq \mathcal{S}$ for all $t \geq 0$.

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

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

Heuristically, ICT sets are characterized by the property that any two points in such a set may be joined by a piecewise continuous chain of arbitrarily long segments of orbits of (MD) broken by arbitrarily small jump discontinuities. As such, they account for a wide range of invariant sets of (MD), ranging from stationary points and periodic orbits (cf. Fig. 1), to homoclinic loops (trajectories that join a unstable critical point to itself), limit cycles (isolated periodic orbits), and many others.

Our next result shows that, *with probability* 1, *any limit point of* (RM) *lies in an ICT set of* Φ:

Theorem 2. *Suppose that* (RM) *is run with a step-size sequence* γ_n *such that* $\sum_n \gamma_n = \infty$, $\lim_n \gamma_n = 0$. *If Assumptions* (A1) and (A2) *hold, then, with probability* 1, *we have: a)* Z_n *converges to an ICT set of* Φ*; or b)* Z_n *is unbounded* (*and hence, non-convergent*).

Corollary 1. *Let* Z_n *be a sequence generated by any of the Algorithms 1–5 with parameters as in Proposition 1. If* Z_n *is bounded, then, with probability* 1, *it converges to an ICT set of* Φ.

The proof of Theorem 2 builds on a series of deep results in [8]; see the supplement. In plain terms, the theorem asserts that any trajectory of (RM) is either unbounded or eventually converges to an ICT set, which is "infinitely close" to the long-term orbits of the mean dynamics (MD). In particular, it rules out *any other type of asymptotic behavior* (convergent or non-convergent).

In gradient systems – i.e., when $V = -\nabla f$ for some $f\colon \mathcal{Z} \to \mathbb{R}$ – the only ICT sets of (MD) are connected sets of critical points of f (for a detailed statement and proof, see the supplement). As a result, we can effortlessly conclude that any RM scheme exhibits the same asymptotic behavior in minimization problems: they converge to connected components of critical points of f.

At the other end of the spectrum, in the bilinear objective $\Phi(x, y) = xy$, we show in the supplement that *any* tuple $(x, y) \in \mathbb{R}^2$ belongs to an ICT set of Φ. The most crucial implication of this observation is that although there exist many non-critical convergent sets in bilinear games, *none of these can be an attractor*: for any bounded region $\mathcal{S}$, there always exists $z \notin \mathcal{S}$ such that, no matter how close z is to $\mathcal{S}$, the mean dynamics (MD) initialized at z will stay at a positive distance from $\mathcal{S}$.

Importantly, in the full space of min-max problems, the two settings described above are both outliers: mixing a gradient system with a bilinear component can give rise to *isolated periodic attractors* (limit cycles) and other forms of attracting ICT sets that cannot be observed in either gradient systems *or* bilinear games. Indeed, our final result in this section shows that, while (SEG) and/or (OG/PEG) might be capable of eliminating periodic orbits in bilinear games [4, 25, 30, 46, 57], these methods fail to escape *spurious (i.e., non-critical) attractors* arising in generic non-convex/non-concave objectives (see also Example 5.1 for a visual illustration). The formal statement is as follows:

Theorem 3. *Let $\mathcal{S}$ be an attractor of (MD) and fix some confidence level $\alpha > 0$. If γ_n is small enough and Assumptions (A1) and (A2) hold, there exists a neighborhood $\mathcal{U}$ of $\mathcal{S}$, independent of α, such that $\mathbb{P}(Z_n \text{ converges to } \mathcal{S} \mid Z_1 \in \mathcal{U}) \geq 1 - \alpha$.*

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

As we show in the next section, Corollary 2 can have catastrophic implications for the convergence of min-max optimization algorithms.

5. Spurious attractors: Illustrations and examples

We now provide concrete examples of attracting ICT sets consisting *entirely* of non-critical points. For illustration purposes, we focus on the simple case $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ with polynomial objectives; of course, all examples below can be suitably generalized to higher dimensions. Despite their rudimentary character, these examples already reveal many unexpected phenomena that are unknown in the context of non-convex minimization (or convex-concave saddle-point problems).

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

$$\Phi(x, y) = xy + \varepsilon\phi(y), \tag{8}$$

where $\varepsilon > 0$ and $\phi(y) = \frac{1}{2}y^2 - \frac{1}{4}y^4$. This problem admits an unstable critical point at the origin; further, using a general criterion provided in the supplement, one can prove, for ε small enough, the existence of an *attracting* ICT set $\mathcal{S}$ in a neighborhood of the circle $\{z : \|z\|^2 = 4/3\}$. Thus, any of the RM schemes of Section 3 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. Because of this, care needs to be taken when interpreting algorithms that

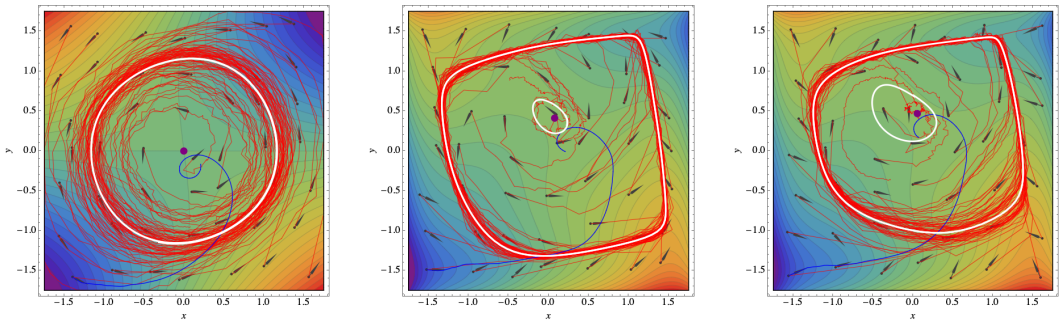


Figure 2: Spurious limits of min-max optimization algorithms. From left to right: (*a*) (SEG) for (8) with $\varepsilon = 0.01$; (*b*) "forsaken solutions" of (SEG); (*c*) "forsaken solutions" of symplectic gradient adjustment (SGA). The red curves present trajectories with different initialization; non-critical ICT sets are depicted in white; the blue curves represent an time-averaged sample orbit.

are characterized as "locally avoiding unstable critical points", since they might be incapable of escaping their neighborhoods. ▲

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

$$\Phi(x, y) = x(y - 0.5) + \phi(x) - \phi(y) \tag{9}$$

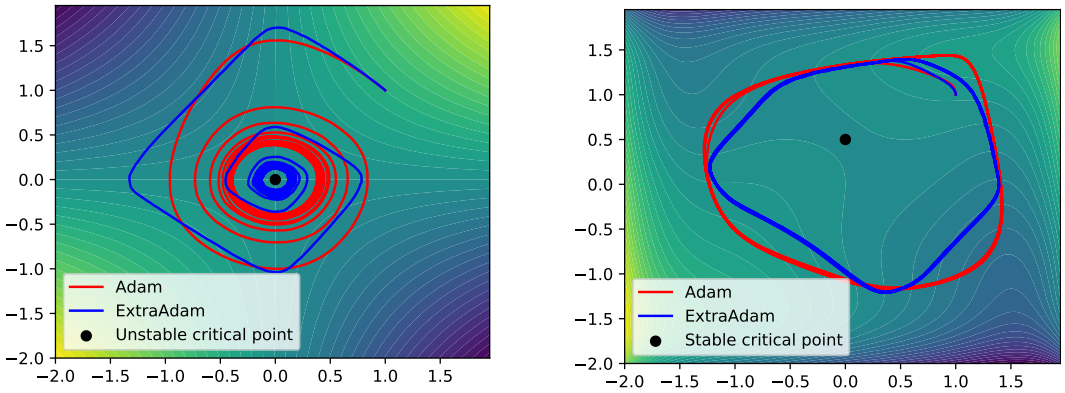
where $\phi(z) = \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$. This problem has a desirable min-max solution at $(x^*, y^*) = (0, 0.5)$. However, we prove in the supplement that there exist *two* spurious limit cycles that do not contain *any* critical point of Φ. Worse, the limit cycle closer to (x^*, y^*) is *unstable* and repels any trajectory that comes close to the solution; see Fig. 2(b) for an illustration for (SEG). Solutions that are "shielded" by spurious limit cycles in this way are unlikely to be visited by existing algorithms; to the best of our knowledge, no research has been conducted to tackle such problematic cases. ▲

▼ **Example 5.3** (Second-order methods)**.** Thanks to the efficient implementation of Hessian-gradient multiplications [67], a popular second-order method for min-max optimization in machine learning is the *Hamiltonian descent* method [1]. The idea is simply to run SGD on $f = \|\nabla\Phi\|^2/2$, giving

$$Z_{n+1} = Z_n - \gamma_n J(Z_n)\nabla\Phi(Z_n). \tag{HD}$$

As a (discretized) gradient system, our theory in Section 4 shows that (HD) does not possess ICT sets other than critical points. However, a serious issue of (HD) is that it ignores the *sign* of gradients, i.e., it does not distinguish between minimization and maximization. For this reason, it has mostly been used as a *gradient penalty* scheme by mixing (HD) (or its variants) with (SGDA), giving rise to a number of other second-order methods such as *symplectic gradient adjustment* (SGA) [5] and *consensus optimization* (ConO) [58]. As in Section 3, one can cast these algorithms as RM schemes with $V(Z_n)$ replaced by $(I - \lambda J(Z_n))V(Z_n)$, where λ is the regularization parameter. The analysis can then proceed as in Section 4 by replacing (MD) with the appropriate continuous system.

Fig. 2(c) shows the spurious convergence of SGA with $\lambda = 0.2$ applied to (9). The ICT sets of SGA are only slightly different from Algorithms 1–5 and, in a certain precise sense, are perturbations thereof (so they suffer the same symptoms); see the supplement for more algorithms and details. ▲



(a) Adaptive algorithms for (8). **(b)** Adaptive algorithms for (9).

We conclude with two remarks of a practical nature. First, Fig. 2 shows that the common tweak of *averaging* the iterates can force the trajectories to halt at non-critical points, and this convergence is by no means min-max optimal. To our knowledge, this provides the first explicit instances where training can get stuck even with non-vanishing gradients, a phenomenon often observed in training GANs.

Second, in Figs. 3a–3b, we report the behaviors of popular *adaptive algorithms* in training GANs, including Adam [40] and its extra-gradient variant [30], both with hyperparameters set to the default values in PyTorch. The result reveals a worrisome trend: while both Adam and ExtraAdam are able to somewhat mitigate the cycling of (8), this nonetheless comes at the price of converging to the *unstable* critical point $(0, 0)$ (which is in fact a local max-min, the opposite of a desirable solution). On the other hand, as all RM schemes, both adaptive methods fail to reach the "forsaken" solutions in Example 5.2.

Finally, we stress that the purpose of examining these practical tweaks is *not* to prove that they will always fail (we have not performed extensive hyperparameter search). Rather, our aim is to point out that they cannot consistently serve as off-the-shelf solutions to the pathological ICT sets, and thus warrant a novel approach in studying min-max optimization problems.

Appendix A. Asymptotic pseudotrajectories

In this appendix, we discuss how the algorithms discussed in Section 3 fit within the general stochastic approximation framework of Section 4.2. Specifically, we prove the general conditions of Theorem 1 and Proposition 1 which guarantee that Algorithms 1–5 generate asymptotic pseudotrajectories of the mean dynamics (MD).

A.1. **Generalities and preliminaries.** Before doing so, we will require some background material on asymptotic pseudotrajectories. Following Benaïm and Hirsch [8] and Benaïm [6], we first recall the definition of the "effective time" $\tau_n = \sum_{k=1}^{n} \gamma_k$ as the time that has elapsed at the n-th iteration of the discrete-time process Z_n; recall also the definition (6) of the continuous-time interpolation $Z(t)$ of Z_n as

$$Z(t) = Z_n + \frac{t - \tau_n}{\tau_{n+1} - \tau_n}(Z_{n+1} - Z_n) \tag{6}$$

We will further require the "continuous-to-discrete" correspondence

$$M(t) = \sup\{n \geq 1 : t \geq \tau_n\} \tag{A.1}$$

which measures the number of iterations required for the effective time τ_n of the process to reach the timestamp t; for future use, we also define the quantity

$$M_n \equiv M_n(T) = M(\tau_n + T). \tag{A.2}$$

Finally, given an arbitrary sequence A_n, we will denote its piecewise constant interpolation as

$$\overline{A}(t) = A_n \quad \text{for all } t \in [\tau_n, \tau_{n+1}],\ n \geq 1. \tag{A.3}$$

Using this notation, the (affinely) interpolated process $Z(t)$ can be expressed in integral form as

$$Z(t) = Z(0) + \int_0^t [V(\overline{Z}(s)) + \overline{W}(s)]\, ds \tag{A.4}$$

where W_n denotes the generalized error term of (RM).

With all this in hand, Benaïm [6, Prop. 4.1] provides the following general condition for $Z(t)$ to be an APT of the mean dynamics (7):

Proposition A.1. *Suppose that $Z(t)$ is bounded and satisfies the general condition*

$$\lim_{t\to\infty} \Delta(t;T) = 0 \quad \textit{for all } T > 0, \tag{A.5}$$

where

$$\Delta(t;T) = \sup_{0\leq h\leq T} \left\| \textstyle\int_t^{t+h} \overline{W}(s)\, ds \right\|. \tag{A.6}$$

Then, $Z(t)$ is an APT of (MD).

A.2. **Proof of Theorem 1.** Our proof of Theorem 1 revolves around the direct verification of the requirement (A.5) of Proposition A.1 via the use of maximal inequalities and martingale limit theory.[1] For convenience, we restate the theorem below in full:

Theorem 1. *Suppose that (RM) is run with a step-size policy γ_n such that $\sum_n \gamma_n = \infty$, $\lim_n \gamma_n = 0$, and Assumptions (A1)–(A2) hold. Then, with probability 1, one of the following holds: a) Z_n is an APT of (MD); or b) Z_n is unbounded (and hence, non-convergent).*

Proof. Our proof relies on the Burkholder–Davis–Gundy (BDG) inequality [18, 33] which bounds the maximal value of a martingale S_n via its quadratic variation as

$$c_2\, \mathbb{E}\left[\sum_{k=1}^n (S_k - S_{k-1})^2\right] \leq \mathbb{E}\left[\max_{k=1,\dots,n} |S_k|^2\right] \leq C_2\, \mathbb{E}\left[\sum_{k=1}^n (S_k - S_{k-1})^2\right], \tag{BDG}$$

where $c_2, C_2 > 0$ are universal constants. As such, applying (BDG) to the martingale $S_m = \sum_{k=n}^m \gamma_k U_k$ (after an appropriate shift of the starting time), we get

$$\begin{aligned}\mathbb{E}\left[\sup_{n\leq m\leq M_n} \left\|\sum_{k=n}^m \gamma_k U_k\right\|^2\right] &\leq C_2\, \mathbb{E}\left[\sum_{k=n}^{M_n} \gamma_k^2 \|U_k\|^2\right] \\ &= C_2 \sum_{k=n}^{M_n} \gamma_k^2 \sigma_k^2 = C_2 \int_{\tau_n}^{\tau_n+T} \overline{\gamma}^2(s)\overline{\sigma}^2(s)\, ds,\end{aligned} \tag{A.7}$$

where $M_n = M_n(T) = M(\tau_n + T)$ is defined as in (A.2). Now, mimicking (A.6), let

$$\Delta_0(t;T) = \sup_{0\leq h\leq T} \left\| \int_t^{t+h} \overline{U}(s)\, ds \right\|. \tag{A.8}$$

[1] Benaïm [6] provides a set of sufficient conditions for (A.5) to hold when $Z(t)$ is generated by a RM scheme with $B_n = 0$ and $\sup_n \sigma_n < \infty$; however, our setting requires a more general treatment.

so our previous bound shows that

$$\mathbb{E}[\Delta_0(t;T)^2] \le C_2 \int_t^{t+T} \overline{\gamma}^2(s)\overline{\sigma}^2(s)\, ds. \tag{A.9}$$

We will proceed to show that $\lim_{t\to\infty} \Delta_0(t;T) = 0$ for all $T > 0$ by considering the sequence of intervals $[kT,(k+1)T]$ and using the Borel-Cantelli lemma to show that $\Delta_0(kT;T) \to 0$ as $k \to \infty$. Indeed, we have

$$\sum_{k=1}^{\infty} \mathbb{E}[\Delta_0(kT;T)^2] \le C_2 \int_0^{\infty} \overline{\gamma}^2(s)\overline{\sigma}^2(s)\, ds = C_2 \sum_{n=1}^{\infty} \gamma_n^2\sigma_n^2 < \infty \tag{A.10}$$

with the last step following from Assumption (A2). Then, if we consider the event $\mathcal{E}_k(\varepsilon) = \{\Delta_0(kT;T) > \varepsilon\}$, Chebysev's inequality gives

$$\sum_{k=1}^{\infty} \mathbb{P}(\mathcal{E}_k(\varepsilon)) \le \frac{\sum_{k=1}^{\infty} \mathbb{E}[\Delta_0(kT;T)^2]}{\varepsilon^2} < \infty, \tag{A.11}$$

and hence, by the Borel-Cantelli lemma, we get

$$\mathbb{P}\left(\limsup_{k\to\infty} \mathcal{E}_k(\varepsilon)\right) = 0. \tag{A.12}$$

This shows that, with probability 1, we have $\Delta_0(kT;T) \le \varepsilon$ for all but a finite number of k; put differently, the event $\mathcal{E}(\varepsilon) = \{\Delta_0(kT;T) \text{ occurs infinitely often}\} = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} \mathcal{E}_k(\varepsilon)$ has $\mathbb{P}(\mathcal{E}(\varepsilon)) = 0$. Therefore, as a union of probability zero events, we have

$$\mathbb{P}\left(\liminf_{k\to\infty} \Delta_0(kT;T) > 0\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} \mathcal{E}(1/n)\right) \le \sum_{n=1}^{\infty} \mathbb{P}(\mathcal{E}(1/n)) = 0, \tag{A.13}$$

i.e., $\Delta_0(kT;T) \to 0$ with probability 1.

Thus, going back to the requirements of Proposition A.1, we get

$$\begin{aligned}
\Delta(kT;T) &= \sup_{0\le h\le T} \left\| \int_{kT}^{kT+h} \overline{W}(t)\, dt \right\| = \sup_{0\le h\le T} \left\| \int_{kT}^{kT+h} [\overline{U}(t) + \overline{b}(t)]\, dt \right\| \\
&\le \Delta_0(kT;T) + \sup_{0\le h\le T} \int_{kT}^{kT+h} \overline{B}(t)\, dt. \\
&\le \Delta_0(kT;T) + T \max_{0\le h\le T} \overline{B}(kT+h).
\end{aligned} \tag{A.14}$$

Given that $\lim_{k\to\infty} B_k = 0$, the above shows that $\Delta(kT;T) \to 0$ as $k \to \infty$. Moreover, for all $t \in [kT,(k+1)T]$, we have $\Delta(t;T) \le 2\Delta(kT;T) + \Delta((k+1)T;T)$ so $\Delta(t;T) \to 0$ with probability 1. With $T > 0$ arbitrary, we conclude that (A.5) holds with probability 1, and our claim follows from Proposition A.1. ∎

To proceed, it will be convenient to consider a stronger version of Assumption (A2):

$$\mathbb{P}(\|U_n\| \ge t \,|\, \mathcal{F}_n) \le 2e^{-\frac{t^2}{2\sigma^2}} \tag{A2′}$$

for some $\sigma \ge 0$ and all $n = 1, 2, \dots$, $t \ge 0$. Some of the RM schemes presented in Section 3 will verify this stronger criterion; see Appendix A.3 below.

Under this assumption, we obtain the following generalization of a criterion due to Benaïm and Hirsch [8]:

Proposition A.2. *Suppose that* (RM) *is run with a step-size policy* γ_n *such that* $A/n \le \gamma_n \le B/(\log n)^{1+\varepsilon}$ *for some* $B, \varepsilon > 0$. *If Assumptions* (A1) and (A2′) *hold, then, with probability* 1, *a)* Z_n *is an APT of* (MD)*; or b)* Z_n *is unbounded* (*and hence, non-convergent*).

Proof. As in the proof of Theorem 1, our approach will hinge on the proviso (A.5) of Proposition A.1 and, in particular, controlling the quantity $\Delta_0(t;T)$ defined in (A.8). We proceed step-by-step:

Step 1: A union bound for the tails of $\sup_{n\leq m\leq M_n}\|\sum_{k=n}^m \gamma_k U_k\|$. Up to a multiplicative constant that depends only on the dimension of the problem, we can assume without loss of generality that $\|\cdot\|$ is the sup-norm $\|z\| = \max_i |z_i|$. In this case, we have $\|z\|\geq t$ if and only if there exists a basis vector e_i of $\mathbb{R}^d$ such that $\langle z, e_i\rangle \geq t$ or $\langle z, e_i\rangle \leq -t$. We thus get the union bound

$$\begin{aligned}\mathbb{P}\left(\sup_{n\leq m\leq M_n}\left\|\sum_{k=n}^m \gamma_k U_k\right\| \geq t\right) &\leq \sum_{i=1}^d \mathbb{P}\left(\sup_{n\leq m\leq M_n}\sum_{k=n}^m \langle \gamma_k U_k, e_i\rangle \geq t\right)\\ &\quad + \sum_{i=1}^d \mathbb{P}\left(\sup_{n\leq m\leq M_n}\sum_{k=n}^m \langle \gamma_k U_k, -e_i\rangle \geq t\right).\end{aligned} \tag{A.15}$$

In view of this, we will focus below on the tail probability $\mathbb{P}(\sup_{n\leq m\leq M_n}\sum_{k=n}^m\langle\gamma_k U_k, z\rangle)$ for arbitrary $z\in\mathbb{R}^d$.

Step 2: Exponential tail concentration. By standard arguments, Assumption (A2′) is equivalent to asking that

$$\mathbb{E}[\exp(\langle z, U_n\rangle)\,|\,\mathcal{F}_n] \leq \exp(\sigma^2\|z\|^2/2). \tag{A.16}$$

With this reformulation in mind, consider the process

$$Q_n(z) = \exp\left(\sum_{k=1}^n \langle z, \gamma_k U_k\rangle - \frac{\sigma^2}{2}\sum_{k=1}^n \gamma_k^2\|z\|^2\right). \tag{A.17}$$

Then, by construction

$$\begin{aligned}\mathbb{E}[Q_n(z)\,|\,\mathcal{F}_n] &= \mathbb{E}\left[\exp\left(\sum_{k=1}^n\langle z,\gamma_k U_k\rangle - \frac{\sigma^2}{2}\sum_{k=1}^n\gamma_k^2\|z\|^2\right)\,\middle|\,\mathcal{F}_n\right]\\ &= Q_{n-1}(z)\,\mathbb{E}\left[\exp\left(\langle z,\gamma_n U_n\rangle - \frac{\sigma^2}{2}\gamma_n^2\|z\|^2\right)\,\middle|\,\mathcal{F}_n\right] \leq Q_{n-1}(z),\end{aligned} \tag{A.18}$$

i.e., $Q_n(z)$ is a supermartingale relative to $\mathcal{F}_n$.[2] Moreover, we have:

$$\begin{aligned}\mathbb{P}\left(\sup_{n\leq m\leq M_n}\sum_{k=n}^m\langle\gamma_k U_k, z\rangle \geq \alpha\right) &= \mathbb{P}\left(\sup_{n\leq m\leq M_n}\frac{Q_m(z)}{Q_n(z)}\exp\left(\frac{\sigma^2}{2}\sum_{k=n}^m\gamma_k^2\|z\|^2\right)\geq\exp(\alpha)\right)\\ &= \mathbb{P}\left(\sup_{n\leq m\leq M_n}\frac{Q_m(z)}{Q_n(z)}\exp\left(\frac{\sigma^2}{2}\sum_{k=n}^{M_n}\gamma_k^2\|z\|^2\right)\geq\exp(\alpha)\right)\\ &= \mathbb{P}\left(\sup_{n\leq m\leq M_n}\frac{Q_m(z)}{Q_n(z)}\geq\exp\left(\alpha-\frac{\sigma^2}{2}\sum_{k=n}^{M_n}\gamma_k^2\|z\|^2\right)\right)\\ &\leq \mathbb{E}\left[\sup_{n\leq m\leq M_n}\frac{Q_m(z)}{Q_n(z)}\right]\cdot\exp\left(\frac{\sigma^2}{2}\sum_{k=n}^{M_n}\gamma_k^2\|z\|^2-\alpha\right)\\ &\leq \exp\left(\frac{\sigma^2}{2}\sum_{k=n}^{M_n}\gamma_k^2\|z\|^2-\alpha\right)\end{aligned} \tag{A.19}$$

[2]Recall here that, by the definition of the filtration $\mathcal{F}_n$, U_n is $\mathcal{F}_{n+1}$-measurable but not $\mathcal{F}_n$-measurable.

where we used Markov's inequality in the last step and the fact that $Q_n(z)$ is a submartingale in the penultimate one. Thus, letting $\Sigma = \sigma^2 \sum_{k=n}^{M_n} \gamma_k^2 \|z\|^2$ and taking $z \leftarrow (t/\Sigma)e_i$, $t \leftarrow t^2/\Sigma$, we get

$$\mathbb{P}\left(\sup_{n \leq m \leq M_n} \sum_{k=n}^{m} \langle \gamma_k U_k, e_i \rangle \geq t \right) \leq \exp\left(-\frac{\sigma^2 t^2}{2 \sum_{k=n}^{M_n} \gamma_k^2} \right). \tag{A.20}$$

Step 3: Closing the gap. By assumption, $\sum_{k=n}^{M_n} \gamma_n^2 \leq T\gamma_n^2 \leq T/(\log n)^{2+2\varepsilon}$. Hence

$$\exp\left(-\frac{\sigma^2 t^2}{2 \sum_{k=n}^{M_n} \gamma_k^2} \right) \leq \exp\left(-\frac{\sigma^2}{2} \frac{(\log n)^{2+2\varepsilon}}{T} \right) = n^{-\frac{\sigma^2}{2} \frac{(\log n)^{1+2\varepsilon}}{T}}. \tag{A.21}$$

Therefore

$$\mathbb{P}\left(\sup_{n \leq m \leq M_n} \left\| \sum_{k=n}^{m} \gamma_k U_k \right\| \geq t \right) \leq \frac{C_2'}{n^2} \tag{A.22}$$

for some suitable constant $C_2' > 0$. With notation as in the proof of Theorem 1, this implies that

$$\sum_{k=1}^{\infty} \mathbb{P}(\Delta_0(kT; T) \leq \alpha) = \mathcal{O}\left(\sum_{k=1}^{\infty} \frac{1}{k^2} \right) < \infty. \tag{A.23}$$

Thus, by applying the Borel-Cantelli lemma as in the proof of Theorem 1, we conclude that $\Delta_0(kT; T) \to 0$ with probability 1. The rest of the arguments required to show that $\lim_{t \to 0} \Delta(t; T) = 0$ for all T follow the lines of the proof of Theorem 1, so we omit them. ∎

A.3. **Proof of Proposition 1.** We are now in a position to prove that the generalized RM schemes presented in Section 3 comprise asymptotic pseudotrajectories of the mean dynamics (MD). For convenience, we state the relevant result below:

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 $A/n \leq \gamma_n \leq B/(\log n)^{1+\varepsilon}$ 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, with probability 1, one of the following holds: a) Z_n is an APT of (MD); or b) Z_n is unbounded.

Proof. We proceed method-by-method:

Algorithm 1: Stochastic gradient descent/ascent. For (SGDA), we have $W_n = U_n = \mathsf{U}(\omega_n)$ and $b_n = 0$, so Assumption (A1) is satisfied automatically (since $B_n = 0$). Moreover, under the stated assumptions for (SFO), U_n is sub-Gaussian, so our claim follows from Proposition A.2.

Algorithm 2: Alternating stochastic gradient descent/ascent. For (alt-SGDA), we have $b_n = (0, V_y(X_{n+1}, Y_n) - V_y(X_n, Y_n))$, and $U_n = (U_{x,n}, U_{y,n})$. Under the stated assumptions for (SFO), U_n satisfies Assumption (A2′), so we are left to show that Assumption (A1) holds, i.e., that $b_n \to 0$. To that end, since V is Lipschitz, we have

$$\|b_n\| = \|V_y(X_{n+1}, Y_n) - V_y(X_n, Y_n)\| \leq L\|X_{n+1} - X_n\|, \tag{A.24}$$

where L denotes the Lipschitz modulus of V. Hence, by the definition of (alt-SGDA), we get

$$\|b_n\| \leq \gamma_n L\|V_y(X_{n+1},Y_n) + U_{y,n}\| \leq \gamma_n L\|V_y(X_{n+1},Y_n)\| + \gamma_n L\|U_{y,n}\| \tag{A.25}$$

If Z_n is bounded, we also have $\sup_n\|V_y(X_{n+1},Y_n)\| < \infty$, so the first term above vanishes as $n \to \infty$ (recall that $\lim_n \gamma_n = 0$). As for the second, we have

$$\mathbb{P}(\|U_n\| \geq \log n) \leq 2e^{-(\log n)^2/(2\sigma^2)} = 2n^{-\log n/(2\sigma^2)} \tag{A.26}$$

In turn, this implies that $\sum_{n=1}^{\infty}\mathbb{P}(\|U_n\| \geq \log n) < \infty$ so, by the Borel-Cantelli lemma, we have $\|U_n\| = \mathcal{O}(\log n)$ with probability 1. Hence, by our assumptions for the method's step-size, we get

$$\gamma_n\|U_{y,n}\| \leq \gamma_n\|U_n\| = \mathcal{O}\left(\frac{\log n}{(\log n)^{1+\varepsilon}}\right) = \mathcal{O}\left(\frac{1}{(\log n)^{\varepsilon}}\right) \tag{A.27}$$

i.e., $B_n \to 0$ with probability 1. Our claim then follows from Proposition A.2.

Algorithm 3: Stochastic extra-gradient. For (SEG), we have $U_n = \mathsf{U}(Z_n^+;\omega_n^+)$ and $b_n = V(Z_n^+) - V(Z_n)$, so Assumption (A2′) holds by default. For Assumption (A1), arguing as in the case of Algorithm 2 above, we have

$$\begin{aligned}\|b_n\| = \|V(Z_n^+) - V(Z_n)\| &\leq L\|Z_n^+ - Z_n\| \\ &= \gamma_n\|\mathsf{V}(\omega_n)\| = \gamma_n L\|V(Z_n) + \mathsf{U}(\omega_n)\| \\ &\leq \gamma_n L\|V(Z_n)\| + \gamma_n L\|\mathsf{U}(\omega_n)\|,\end{aligned} \tag{A.28}$$

Thus, by Proposition A.2, we conclude that Z_n is an APT of (MD).

Algorithm 4: Optimistic gradient. For (OG/PEG), we have $U_n = \mathsf{U}(\omega_n^+)$ and $b_n = V(Z_n^+) - V(Z_n)$. so Assumption (A2′) again holds by default. The bias term can then be bounded exactly as in the case of Algorithm 3, so our APT claim follows again by Proposition A.2.

Algorithm 5: Simultaneous perturbation stochastic approximation. Because of the algorithm's different oracle structure (zeroth- vs. first-order feedback), the analysis of (SPSA) is different. We begin with the algorithm's bias term, given here by

$$b_n = \mathbb{E}[V_n \,|\, \mathcal{F}_n] - V(Z_n) \tag{A.29}$$

with

$$V_n = \pm(d/\delta_n)\,\Phi(Z_n + \delta_n\omega_n)\,\omega_n \tag{A.30}$$

denoting the method's one-shot SPSA estimator. To bound it, let

$$v_{i,n} = \mathbb{E}[V_{i,n} \,|\, \mathcal{F}_n] \tag{A.31}$$

denote the i-th component of $V_n \in \mathbb{R}^d$ after having averaged out the choice of the random seed ω_n (which, by default, is not $\mathcal{F}_n$-measurable). We then have

$$v_{i,n} = \pm\frac{d}{\delta_n}\cdot\frac{1}{2d}\big[\Phi(Z_n + \delta_n e_i) - \Phi(Z_n - \delta_n e_i)\big] \tag{A.32}$$

where, as per our discussion in Section 3, the "$\pm$" sign is equal to -1 if $e_i \in \mathcal{E}_{\mathcal{X}}$ and $+1$ if $e_i \in \mathcal{E}_{\mathcal{Y}}$. Then, by the mean value theorem, there exists some $\tilde{Z}_n$ in the line segment $\big[Z_n - \delta_n e_i, Z_n + \delta_n e_i\big]$ such that

$$v_{i,n} = \pm\partial_i\Phi(\tilde{Z}_n) = V_{i,n}(\tilde{Z}_n). \tag{A.33}$$

Since V is Lipschitz continuous, it follows that

$$|v_{i,n} - V_{i,n}(Z_n)| = \big|V_{i,n}(\tilde{Z}_n) - V_{i,n}(Z_n)\big| \leq L\|\tilde{Z}_n - Z_n\| = \mathcal{O}(\delta_n) \tag{A.34}$$

since $\tilde{Z}_n \in \big[Z_n - \delta_n e_i, Z_n + \delta_n e_i\big]$. Finally, for the oracle's variance, we have $\|V_n\|^2 = \mathcal{O}(1/\delta_n^2)$ by construction so, under the stated assumptions for γ_n and δ_n, Assumption (A2) is satisfied and our claim follows from Theorem 1. ∎

We conclude this appendix with a simple coercivity criterion which guarantees that the iterates of an iterative method of the general form (RM) remain bounded:

Proposition A.3. *Suppose that V satisfies the coercivity condition*

$$\liminf_{\|z\|\to\infty} \frac{\langle V(z), z\rangle}{\|z\|^2} < 0. \tag{A3}$$

Then, under Assumptions (A1) and (A2), the sequence Z_n generated by (RM) is bounded (a.s.).

Corollary 3. *Under Assumptions (A1)–(A3), the iterates Z_n of (RM) comprise an APT of (MD).*

Proof. To begin, observe that, under Assumption (A3), the quadratic penalty function $E(z) = \sum_i z_i^2/2$ is a Lyapunov function for (MD) as $\|z\| \to \infty$. Indeed, by Assumption (A3), there exists some $R > 0$ such that, whenever $\|z\| \geq R$, we have

$$\frac{dE}{dt} = \langle \nabla E(z), \dot{z}\rangle = \langle \nabla E(z), V(z)\rangle \leq -\frac{\kappa}{2}\|z\|^2 \tag{A.35}$$

where $\kappa = -\liminf_{\|z\|\to\infty}\langle V(z), z\rangle/\|z\|^2 > 0$.[3] This shows that trajectories of (MD) cannot escape to infinity so it is plausible to expect the same to hold for (RM).

Our proof of this fact follows a direct stabilization technique due to Kushner and Yin [43]. Specifically, going back to (RM), a simple expansion gives

$$\begin{aligned} E(Z_{n+1}) &= E(Z_n) + \gamma_n\langle V_n, Z_n\rangle + \frac{1}{2}\gamma_n^2\|V_n\|^2 \\ &\leq E(Z_n) + \gamma_n\langle V(Z_n), Z_n\rangle + \gamma_n\langle W_n, Z_n\rangle + \gamma_n^2\|V_n\|^2 \end{aligned} \tag{A.36}$$

Hence, taking (conditional) expectations, we obtain:

$$\mathbb{E}[E(Z_{n+1}) \,|\, \mathcal{F}_n] \leq E(Z_n) + \gamma_n\langle V(Z_n) + b_n, Z_n\rangle + \gamma_n^2\,\mathbb{E}[\|V_n\|^2 \,|\, \mathcal{F}_n]. \tag{A.37}$$

To proceed, note that, by Assumptions (A1) and (A2), we have

$$\mathbb{E}\left[\sum_{n=1}^{\infty} \gamma_n^2\|V_n^2\|\,\mathbb{1}_{\{\|Z_n\|\leq R\}}\right] < \infty, \tag{A.38}$$

while, otherwise

$$\mathbb{E}\big[\|V_n\|^2 \,\big|\, \mathcal{F}_n\big] \leq C\big(\sigma_n^2 + (\kappa/2)\|Z_n\|^2\big) \quad \text{whenever } \|Z_n\| \geq R. \tag{A.39}$$

Consider now the process

$$S_n = \mathbb{E}\Big[\textstyle\sum_{k\geq n} \gamma_k^2\|V_n\|^2\,\mathbb{1}_{\{\|Z_k\|\leq R\}} \,\Big|\, \mathcal{F}_n\Big] \tag{A.40}$$

and let $E_n = E(Z_n) + S_n$. By definition, E_n is non-negative; moreover, by (A.36), we get

$$\mathbb{E}[E_{n+1} - E_n \,|\, \mathcal{F}_n] \leq -\frac{\kappa\gamma_n}{2}\|Z_n\|^2 + \frac{C\gamma_n^2}{2}\|Z_n\|^2. \tag{A.41}$$

Since $\gamma_n \to 0$, it follows that E_n is eventually a supermartingale: specifically, if $n_0 = \sup\{n : C\gamma_n > \kappa\}$ (with the standard convention $\sup\varnothing = -\infty$), we have $\mathbb{E}[E_{n+1} \,|\, \mathcal{F}_n] \leq E_n$ for all

[3] In the above and throughout this proof, we assume that $\|\cdot\|$ is the ordinary Euclidean norm on $\mathbb{R}^d$; this assumption is only made for notational convenience and to avoid carrying around many multiplicative constants.

$n \geq n_0$. Since $\mathbb{E}[E_{n_0}] < \infty$, Doob's submartingale convergence theorem subsequently implies that E_n converges with probability 1 to some non-negative random variable E_∞. Since $S_n \to 0$ with probability 1 (by Assumption (A2)), we conclude that $\|Z_n\| = (2/\kappa)E(Z_n) \to (2/\kappa)E_\infty$ (a.s.), and our claim follows. ∎

Appendix B. Convergence analysis

With all this preliminary work in hand, we are finally in a position to prove Theorems 2 and 3. The heavy lifting for the former is provided by the fact that, under the requirements of Theorem 1 and/or Proposition 1, Z_n is an APT of the mean dynamics (MD), so it inherits its limit structure. The latter requires completely different techniques and involves a much finer analysis of the process in hand.

B.1. Convergence to ICTs.

We begin with Theorem 2, which we restate below for convenience:

Theorem 2. *Suppose that* (RM) *is run with a step-size sequence* γ_n *such that* $\sum_n \gamma_n = \infty$, $\lim_n \gamma_n = 0$. *If Assumptions* (A1) and (A2) *hold, then, with probability* 1, *we have: a)* Z_n *converges to an ICT set of* Φ*; or b)* Z_n *is unbounded* (*and hence, non-convergent*).

Proof. We consider two cases. First, if Z_n is unbounded, there is nothing to show. Otherwise, if Z_n is bounded, Theorem 2 shows that it is an APT of the mean dynamics (MD). Now, let $\mathcal{L} = \bigcap_{t\geq 0} \mathrm{cl}(Z(t,\infty))$ be the limit set of $Z(t)$, i.e., the set of limit points of convergent sequences $Z(t_n)$ with $\lim_n t_n = \infty$. Our claim then follows by the limit set theorem of Benaïm and Hirsch [8, Theorem 8.2]. ∎

As we discussed in the main part of our paper, the ICT sets of Φ may exhibit a wide variety of structural properties (limit cycles, heteroclinic networks, etc.). As a complement to this, we show below that, in *gradient* systems ($V = -\nabla f$ for some $f\colon \mathcal{Z} \to \mathbb{R}$), ICT sets can only be compoments of equilibria. Specifically, building on a general result by Benaïm [6], we have:

Proposition B.1. *Suppose that* $V(z) = -\nabla f(z)$ *for some* C^d*-smooth potential function* $f\colon \mathcal{Z} \to \mathbb{R}$ *with a compact critical set* $\mathrm{crit}(f) = \{z^* : \nabla f(z^*) = 0\}$. *Then, every ICT set* $\mathcal{S}$ *of* (MD) *is contained in* $\mathrm{crit}(f)$*; moreover,* f *is constant on* $\mathcal{S}$. *In particular, any ICT set of* (MD) *consists solely of critical points of* f.

Proof. Under the stated conditions, the critical set $\mathcal{Z}^* := \mathrm{crit}(f)$ of f coincides with the set of rest points of (MD). Moreover, by Sard's theorem [44], $f(\mathcal{Z}^*)$ has zero Lebesgue measure and hence empty interior. Our claim then follows from Proposition 6.4 of Benaïm [6]. ∎

As another elementary illustration in addition to the gradient systems, one can show that for bilinear games $\Phi(x,y) = xy$, the ICT sets are annular regions of the form $\{z : r \leq \|z\| \leq R,\ 0 \leq r \leq R\}$. This can be easily seen by considering the widely known Hamiltonian function $H(x,y) = x^2 + y^2$, which satisfies $\dot{H} = 0$ provided (x,y) follows (MD). An immediate consequence of this fact is that *any* point on $\mathbb{R}^2$ lies in some ICT set of (MD), which further implies that there is no bounded attracting region, i.e., attractors.

B.2. Convergence to attractors.

We now proceed with the analysis of RM schemes in the presence of an attractor; the relevant result is Theorem 3:

Theorem 3. *Let* $\mathcal{S}$ *be an attractor of* (MD) *and fix some confidence level* $\alpha > 0$. *If* γ_n *is small enough and Assumptions* (A1) and (A2) *hold, there exists a neighborhood* $\mathcal{U}$ *of* $\mathcal{S}$, *independent of* α, *such that* $\mathbb{P}(Z_n \text{ converges to } \mathcal{S} \mid Z_1 \in \mathcal{U}) \geq 1 - \alpha$.

Because of the generality of our assumptions, the proof of Theorem 3 requires a range of completely different arguments and techniques. We illustrate the main steps of our technical trajectory below:

(1) The first crucial component of our proof is to establish an energy function for (RM) in a neighborhood of $\mathcal{S}$. To do this, we rely on Conley's decomposition theorem (the so-called "fundamental theorem of dynamical systems") which states that the mean dynamics (MD) are "gradient-like" in a neighborhood of an attractor, i.e., they admit a (local) Lyapunov function.

(2) Because of the noise in (RM), the evolution of E along the trajectories of (RM) could present *signifcant* jumps: in particular, a single "bad" realization of the noise could carry Z_n out of the basin of attraction of $\mathcal{S}$, possibly never to return. A major difficulty here is that the driving vector field V is *not* assumed bounded, so it is not straightforward to establish proper control over the error terms of (RM). However, we show that, with high probability (and, in particular, with probability at least $1-\alpha$), the aggregation of these errors remains controllably small; this is the most technically challenging part of our argument and it unfolds in a series of lemmas below.

(3) Conditioning on the above, we will show that, with probability at least $1-\alpha$, the value of the trajectory's energy cannot grow more than a token threshold ε; as a result, if (RM) is initialized close to $\mathcal{S}$, it will remain in a neighborhood thereof for all n (again, with probability at least $1-\alpha$).

(4) Thanks to this "stochastic Lyapunov stability" result, we can regain control of the variance of the process and use martingale limit and maximal inequality arguments to show that Z_n converges to $\mathcal{S}$.

In the rest of this section, we make this roadmap precise via a series of technical lemmas and intermediate results.

A local energy function for (RM). We begin by providing a suitable (local) energy function for (MD). Indeed, since $\mathcal{S}$ is an attractor, there exists a compact neighborhood $\mathcal{K}$ of $\mathcal{S}$, called the *fundamental neighborhood* of $\mathcal{S}$, and having the defining property that $\operatorname{dist}(\Theta_t(z),\mathcal{S}) \to 0$ as $t\to\infty$ uniformly in $z\in\mathcal{K}$. Since all trajectories of (MD) that start in $\mathcal{K}$ converge to $\mathcal{S}$, there are no other non-trivial invariant sets in $\mathcal{K}$ except $\mathcal{S}$. As a result, with $\mathcal{K}$ compact, Conley's decomposition theorem for dynamical systems [23] shows that there exists a smooth Lyapunov – or "energy" – function $E\colon \mathcal{K}\to\mathbb{R}$ such that (i) $E(z)\geq 0$ with equality if and only if $z\in\mathcal{S}$; and (ii) $\dot{E}(z) \coloneqq \langle \nabla E(z), V(z)\rangle < 0$ for all $z\in\mathcal{K}\setminus\mathcal{S}$ (implying in particular that $E(\Theta_t(z))$ is strictly decreasing in t whenever $z\in\mathcal{K}\setminus\mathcal{S}$).

In the discrete-time context of (RM), the energy $E_n \coloneqq E(Z_n)$ of Z_n may fail to be decreasing (strictly or otherwise). However, a simple Taylor expansion with Lagrange remainder yields the basic energy bound

$$E_{n+1} \leq E_n + \gamma_n \langle \nabla E(Z_n), V(Z_n)\rangle + \gamma_n \xi_n + \gamma_n \psi_n + \gamma_n^2 \theta_n^2, \tag{B.1}$$

where the error terms ξ_n, ψ_n and θ_n are defined as

$$\xi_n = \langle \nabla E(Z_n), U_n\rangle \tag{B.2a}$$

$$\psi_n = B_n \|\nabla E(Z_n)\| + \gamma_n \beta B_n^2 \tag{B.2b}$$

$$\theta_n^2 = \beta \|V(Z_n) + U_n\|^2 \tag{B.2c}$$

with β denoting the strong smoothness modulus of E over the compact set $\mathcal{K}$. Clearly, each of these error terms can be positive, so E_n may fail to be decreasing; we discuss how these errors can be controlled below.

Error control. We begin by encoding the aggregation of the error terms in (B.1) as

$$M_n = \sum_{k=1}^{n} \gamma_k \xi_k \tag{B.3a}$$

and

$$S_n = \sum_{k=1}^{n} [\gamma_k \psi_k + \gamma_k^2 \theta_k^2] \tag{B.3b}$$

Since $\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] = 0$, we have $\mathbb{E}[M_n \,|\, \mathcal{F}_n] = M_{n-1}$, so M_n is a martingale; likewise, $\mathbb{E}[S_n \,|\, \mathcal{F}_n] \geq S_{n-1}$, so S_n is a submartingale. Interestingly, even though M_n appears more "balanced" as an error (because ξ_n is zero-mean), it is more difficult to control because the variance of its increments is

$$\mathbb{E}[|\gamma_n \xi_n|^2 \,|\, \mathcal{F}_n] = \gamma_n^2 \, \mathbb{E}[|\langle \nabla E(Z_n), U_n \rangle|^2 \,|\, \mathcal{F}_n], \tag{B.4}$$

so the jumps of M_n can become arbitrarily big if Z_n escapes $\mathcal{K}$ (which is the event we are trying to discount in the first place). On that account, we will instead bound the total error increments by *conditioning* everything on the event that Z_n remains within $\mathcal{K}$.

To make this precise, consider the "mean square" error process

$$R_n = M_n^2 + S_n \tag{B.5}$$

and the indicator events

$$\mathcal{E}_n \equiv \mathcal{E}_n(\mathcal{K}) = \{Z_n \in \mathcal{K} \text{ for all } k = 1, 2, \dots, n\} \tag{B.6}$$

$$\mathcal{H}_n \equiv \mathcal{H}_n(\varepsilon) = \{R_k \leq \varepsilon \text{ for all } k = 1, 2, \dots, n\}, \tag{B.7}$$

with the convention $\mathcal{E}_0 = \mathcal{H}_0 = \Omega$. Moving forward, with significant hindsight, we will choose ε small enough so that

$$\{z \in \mathcal{Z} : E(z) \leq 2\varepsilon + \sqrt{\varepsilon}\} \subseteq \mathcal{K}. \tag{B.8}$$

and we will assume that Z_1 is initialized in a neighborhood $\mathcal{U} \subseteq \mathcal{K}$ such that

$$\mathcal{U} \subseteq \{z \in \mathcal{Z} : E(z) \leq \varepsilon\} \tag{B.9}$$

We then have the following estimates:

Lemma B.1. *Suppose that $Z_1 \in \mathcal{U}$ and Assumptions (A1) and (A2) hold. Then*

(1) $\mathcal{E}_{n+1} \subseteq \mathcal{E}_n$ and $\mathcal{H}_{n+1} \subseteq \mathcal{H}_n$.

(2) $\mathcal{H}_{n-1} \subseteq \mathcal{E}_n$.

(3) Consider the "bad realization" event

$$\begin{aligned} \tilde{\mathcal{H}}_n &:= \mathcal{H}_{n-1} \setminus \mathcal{H}_n = \mathcal{H}_{n-1} \cap \{R_n > \varepsilon\} \\ &= \{R_k \leq \varepsilon \text{ for } k = 1, 2, \dots, n-1 \text{ and } R_n > \varepsilon\}, \end{aligned} \tag{B.10}$$

and let $\tilde{R}_n = R_n \mathbb{1}_{\mathcal{H}_{n-1}}$ denote the cumulative error subject to the noise being "small" until time n. Then:

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_{n-1}] + \gamma_n G B_n + \gamma_n^2 [2\beta G^2 + (2\beta + G^2)\sigma_n^2 + \beta B_n^2] - \varepsilon \, \mathbb{P}(\tilde{\mathcal{H}}_{n-1}), \tag{B.11}$$

where $G^2 = \sup_{z \in \mathcal{K}} \{\|\nabla E(z)\|^2 + \|V(z)\|^2\}$ and, by convention, $\tilde{\mathcal{H}}_0 = \varnothing$, $\tilde{R}_0 = 0$.

Proof. The first claim is obvious. For the second, we proceed inductively:

(1) For the base case $n = 1$, we have $\mathcal{E}_1 = \{Z_1 \in \mathcal{K}\} \supseteq \{Z_1 \in \mathcal{U}\} = \Omega$ (recall that Z_1 is initialized in $\mathcal{U} \subseteq \mathcal{K}$). Since $\mathcal{H}_0 = \Omega$, our claim follows.

(2) Inductively, suppose that $\mathcal{H}_{n-1} \subseteq \mathcal{E}_n$ for some $n \geq 1$. To show that $\mathcal{H}_n \subseteq \mathcal{E}_{n+1}$, suppose that $R_k \leq \varepsilon$ for all $k = 1, 2, \dots, n$. Since $\mathcal{H}_n \subseteq \mathcal{H}_{n-1}$, this implies that $\mathcal{E}_n$ also occurs, i.e., $Z_k \in \mathcal{K}$ for all $k = 1, 2, \dots, n$; as such, it suffices to show that $Z_{n+1} \in \mathcal{K}$.

To do so, given that $Z_k \in \mathcal{U} \subseteq \mathcal{K}$ for all $k = 1, 2, \dots n$, the bound (B.1) gives

$$E_{k+1} \leq E_k + \gamma_n \xi_n + \gamma_n \psi_n + \gamma_n^2 \theta_n^2, \quad \text{for all } k = 1, 2, \dots n, \tag{B.12}$$

and hence, after telescoping over $k = 1, 2, \dots, n$, we get

$$E_{n+1} \leq E_1 + M_n + S_n \leq E_1 + \sqrt{R_n} + R_n \leq \varepsilon + \sqrt{\varepsilon} + \varepsilon = 2\varepsilon + \sqrt{\varepsilon}. \tag{B.13}$$

We conclude that $E(Z_{n+1}) \leq 2\varepsilon + \sqrt{\varepsilon}$, i.e., $Z_{n+1} \in \mathcal{K}$, as required for the induction.

For our third claim, note first that

$$\begin{aligned} R_n &= (M_{n-1} + \gamma_n \xi_n)^2 + S_{n-1} + \gamma_n \psi_n + \gamma_n^2 \theta_n^2 \\ &= R_{n-1} + 2\gamma_n \xi_n M_{n-1} + \gamma_n^2 \xi_n^2 + \gamma_n \psi_n + \gamma_n^2 \theta_n^2, \end{aligned} \tag{B.14}$$

so, after taking expectations:

$$\mathbb{E}[R_n \,|\, \mathcal{F}_n] = R_{n-1} + 2M_{n-1}\gamma_n \,\mathbb{E}[\xi_n \,|\, \mathcal{F}_n] + \mathbb{E}[\gamma_n^2 \xi_n^2 + \gamma_n \psi_n + \gamma_n^2 \theta_n^2 \,|\, \mathcal{F}_n] \geq R_{n-1} \tag{B.15}$$

i.e., R_n is a submartingale. To proceed, let $\tilde{R}_n = R_n \mathbb{1}_{\mathcal{H}_{n-1}}$ so

$$\begin{aligned} \tilde{R}_n &= R_{n-1} \mathbb{1}_{\mathcal{H}_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}} \\ &= R_{n-1} \mathbb{1}_{\mathcal{H}_{n-2}} - R_{n-1} \mathbb{1}_{\tilde{\mathcal{H}}_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}}, \\ &= \tilde{R}_{n-1} + (R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}} - R_{n-1} \mathbb{1}_{\tilde{\mathcal{H}}_{n-1}}, \end{aligned} \tag{B.16}$$

where we used the fact that $\mathcal{H}_{n-1} = \mathcal{H}_{n-2} \setminus \tilde{\mathcal{H}}_{n-1}$ so $\mathbb{1}_{\mathcal{H}_{n-1}} = \mathbb{1}_{\mathcal{H}_{n-2}} - \mathbb{1}_{\tilde{\mathcal{H}}_{n-1}}$. Then, (B.14) yields

$$R_n - R_{n-1} = 2M_{n-1}\gamma_n \xi_n + \gamma_n^2 \xi_n^2 + \gamma_n \psi_n + \gamma_n^2 \theta_n^2 \tag{B.17}$$

so

$$\mathbb{E}[(R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}}] = 2\,\mathbb{E}[\gamma_n M_{n-1} \xi_n \mathbb{1}_{\mathcal{H}_{n-1}}] \tag{B.18a}$$

$$+ \mathbb{E}[\gamma_n^2 \xi_n^2 \mathbb{1}_{\mathcal{H}_{n-1}}] \tag{B.18b}$$

$$+ \mathbb{E}[(\gamma_n \psi_n + \gamma_n^2 \theta_n^2) \mathbb{1}_{\mathcal{H}_{n-1}}] \tag{B.18c}$$

However, since $\mathcal{H}_{n-1}$ and M_{n-1} are both $\mathcal{F}_n$-measurable, we have the following estimates:

(1) For the noise term in (B.18a), we have:

$$\mathbb{E}[M_{n-1} \xi_n \mathbb{1}_{\mathcal{H}_{n-1}}] = \mathbb{E}[M_{n-1} \mathbb{1}_{\mathcal{H}_{n-1}} \mathbb{E}[\xi_n \,|\, \mathcal{F}_n]] = 0. \tag{B.19}$$

(2) The term (B.18b) is where the reduction to $\mathcal{H}_{n-1}$ kicks in; indeed:

$$\begin{aligned} \mathbb{E}[\xi_n^2 \mathbb{1}_{\mathcal{H}_{n-1}}] &= \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \mathbb{E}[|\langle \nabla E(Z_n), U_n \rangle|^2 \,|\, \mathcal{F}_n]] \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \|\nabla E(Z_n)\|^2 \,\mathbb{E}[\|U_n\|^2 \,|\, \mathcal{F}_n]] && \{\text{by Cauchy–Schwarz}\} \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{E}_n} \|\nabla E(Z_n)\|^2 \,\mathbb{E}[\|U_n\|^2 \,|\, \mathcal{F}_n]] && \{\text{because } \mathcal{H}_{n-1} \subseteq \mathcal{E}_n\} \\ &\leq G^2 \sigma_n^2, && \{\text{by Eq. (5b)}\} \end{aligned}$$

where $G^2 = \sup_{z \in \mathcal{K}} \{\|\nabla E(z)\|^2 + \|V(z)\|^2\}$.

(3) Finally, for the term (B.18c), we have:

$$\mathbb{E}[\theta_n^2 \mathbb{1}_{\mathcal{H}_{n-1}}] \leq 2\beta \,\mathbb{E}[\|V(Z_n)\|^2 \mathbb{1}_{\mathcal{E}_n} + \|U_n\|^2] \leq 2\beta(G^2 + \sigma_n^2), \tag{B.20}$$

where we used the fact that $\mathbb{1}_{\mathcal{H}_{n-1}} \leq \mathbb{1}_{\mathcal{E}_n} \leq 1$. Likewise,

$$\mathbb{E}[\psi_n \mathbb{1}_{\mathcal{H}_{n-1}}] \leq G B_n + \gamma_n \beta B_n^2. \tag{B.21}$$

Thus, putting together all of the above, we obtain:

$$\mathbb{E}[(R_n - R_{n-1})\,\mathbb{1}_{\mathcal{H}_{n-1}}] \leq \gamma_n G B_n + \gamma_n^2[2\beta G^2 + (2\beta + G^2)\sigma_n^2 + \beta B_n^2]. \tag{B.22}$$

Going back to (B.16), we have $R_{n-1} > \varepsilon$ if $\tilde{\mathcal{H}}_{n-1}$ occurs, so the last term becomes

$$\mathbb{E}[R_{n-1}\,\mathbb{1}_{\tilde{\mathcal{H}}_{n-1}}] \geq \varepsilon\,\mathbb{E}[\mathbb{1}_{\tilde{\mathcal{H}}_{n-1}}] = \varepsilon\,\mathbb{P}(\tilde{\mathcal{H}}_{n-1}). \tag{B.23}$$

Our claim then follows by combining Eqs. (B.16), (B.20), (B.21) and (B.23). ∎

Containment probability. Lemma B.1 is the key to showing that Z_n remains close to $\mathcal{S}$ with high probability: we formalize this in a final intermediate result below.

Proposition B.2. *Fix some confidence threshold $\alpha > 0$. If (RM) is run with sufficiently small γ_n satisfying the conditions of Proposition 1, then*

$$\mathbb{P}(\mathcal{H}_n \mid Z_1 \in \mathcal{U}) \geq 1 - \alpha \quad \text{for all } n = 1, 2, \dots \tag{B.24}$$

i.e., Z remains within the basin of attraction $\mathcal{K}$ of $\mathcal{S}$ with probability at least $1 - \alpha$.

Proof. We begin by bounding the probability of the "bad realization" event $\tilde{\mathcal{H}}_n = \mathcal{H}_{n-1} \setminus \mathcal{H}_n$. Indeed, if $Z_1 \in \mathcal{U}$, we have:

$$\begin{aligned}
\mathbb{P}(\tilde{\mathcal{H}}_n) &= \mathbb{P}(\mathcal{H}_{n-1} \setminus \mathcal{H}_n) = \mathbb{P}(\mathcal{H}_{n-1} \cap \{R_n > \varepsilon\}) \\
&= \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \times \mathbb{1}_{\{R_n > \varepsilon\}}] \\
&\leq \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \times (R_n/\varepsilon)] \\
&= \mathbb{E}[\tilde{R}_n]/\varepsilon
\end{aligned} \tag{B.25}$$

where, in the second-to-last line, we used the fact that $R_n \geq 0$ (so $\mathbb{1}_{\{R_n > \varepsilon\}} \leq R_n/\varepsilon$). Telescoping (B.11) yields

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_0] + G\sum_{k=1}^{n} \gamma_k B_k + \sum_{k=1}^{n} \gamma_k^2 \varrho_k^2 - \varepsilon \sum_{k=1}^{n} \mathbb{P}(\tilde{\mathcal{H}}_{k-1}) \tag{B.26}$$

where we set $\varrho_n^2 = 2\beta G^2 + (2\beta + G^2)\sigma_n^2 + \beta B_n^2$. Hence, combining (B.25) and (B.26) and invoking Assumptions (A1) and (A2), we get $\sum_{k=1}^{n} \mathbb{P}(\tilde{\mathcal{H}}_k) \leq \frac{1}{\varepsilon}\sum_{k=1}^{n}[\gamma_k G B_k + \gamma_k^2 \varrho_k^2] \leq \Gamma/\varepsilon$ for some $\Gamma > 0$. Now, by choosing γ_n sufficiently small, we can ensure that $\Gamma/\varepsilon < \alpha$; therefore, given that the events $\tilde{\mathcal{H}}_k$ are disjoint for all $k = 1, 2, \dots$, we get

$$\mathbb{P}\left(\bigcup_{k=1}^{n} \tilde{\mathcal{H}}_k\right) = \sum_{k=1}^{n} \mathbb{P}(\tilde{\mathcal{H}}_k) \leq \alpha \tag{B.27}$$

and hence:

$$\mathbb{P}(\mathcal{H}_n) = \mathbb{P}\left(\bigcap_{k=1}^{n} \tilde{\mathcal{H}}_k^{\mathsf{c}}\right) \geq 1 - \alpha, \tag{B.28}$$

as claimed. ∎

Convergence with high probability. We are finally in a position to prove the convergence of generalized RM algorithms:

Proof of Theorem 3. By Proposition B.2, if Z_n is initialized within the neighborhood $\mathcal{U}$ defined in (B.9), we have $\mathbb{P}(Z_n \in \mathcal{K} \,|\, Z_1 \in \mathcal{U}) \geq 1 - \alpha$ (note also that the neighborhood $\mathcal{U}$ is independent of the required confidence level α). Since $\mathcal{K}$ is compact, if $Z_n \in \mathcal{K}$ for all n, we conclude by Theorem 1 that the continuous-time interpoloation $Z(t)$ of Z_n is an APT of (MD).

Now, if we write $\mathcal{L} = \bigcap_{t\geq 0} \mathrm{cl}(Z(t,\infty))$ for the limit set of $Z(t)$, we have $\mathcal{K} \cap \mathcal{L} \neq \varnothing$ by the compactness of $\mathcal{K}$ and the fact that $Z_n \in \mathcal{K}$ for all $n \geq 1$; moreover, $\mathcal{L}$ is itself compact as a closed subset of the compact set $\{\Theta_t(z) : 0 \leq t \leq T, z \in \mathcal{K}\}$. Since points in $\mathcal{L} \cap \mathcal{K}$ are a fortiori attracted to $\mathcal{S}$ under (MD) and $\mathcal{L}$ is invariant under (MD), we conclude that $\mathcal{L} \cap \mathcal{S} \neq \varnothing$. However, since $\mathcal{L}$ is internally chain-transitive (by Theorem 2) and internally chain-transitive sets do not contain any proper attractors, we conclude that $\mathcal{L} \subseteq \mathcal{S}$. This shows that $Z(t)$ – and, by consequence, Z_n – converges to $\mathcal{S}$, as claimed. ∎

Appendix C. Omitted proofs for Section 5

C.1. **A general criterion for spurious ICT sets in almost bilinear games.** We first provide a generic criterion for the existence of spurious ICT sets in almost bilinear games (8); cf. Lemma C.1. We then verify that the perturbation $\phi(y) = \frac{1}{2}y^2 - \frac{1}{4}y^4$ employed in Example 5.1 indeed satisfies the required conditions.

Lemma C.1. *Let $\phi(y) = \sum_k a_k y^k$ be an analytic function such that*

$$\sum_k a_{2k} k h^{2k} \prod_{i=1}^{k} \frac{2i-1}{2i} = 0 \tag{C.1}$$

has a solution with $h > 0$. Then, for small enough ε, there is an ICT set of mean dynamics (MD) with objective $\Phi(x,y) = xy + \varepsilon\phi(y)$ such that it does not contain any critical point.

Proof. Recall the mean dynamics (MD):

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

In the case of $\Phi(x,y) = xy + \varepsilon\phi(y)$, (MD) reads:

$$\begin{cases} \dot{x} = -y \\ \dot{y} = x + \varepsilon\phi'(y) \end{cases}. \tag{C.2}$$

The most important tool of the proof is the *Abelian integral* [21]:

$$I(h) \coloneqq -\oint_{\gamma_h} \phi' \mathrm{d}x \tag{AI}$$

where $h > 0$ is a parameter and γ_h is a family of ovals defined as in (2.3) of [21].

Suppose $\phi(y) = a_k y^k$, so that $\phi'(y) = k a_k y^{k-1}$. We choose $\gamma_h = \{z : \|z\| = h\}$. Then, using the polar coordinate representation, we get

$$\begin{aligned} I(h) &= -\oint_{\gamma_h} \phi' \mathrm{d}x \\ &= k a_k \int_0^{2\pi} h^k \sin^k(\theta) \mathrm{d}\theta \\ &= k a_k \cdot \begin{cases} 0 & \text{if } k \text{ is odd,} \\ 2\pi h^k \prod_{i=1}^{\frac{k}{2}} \frac{2i-1}{2i} & \text{if } k \text{ is even.} \end{cases} \end{aligned} \tag{C.3}$$

Since contour integrals are linear in the integrands, when $\phi(y) = \sum_k a_k y^k$ in (AI), we have

$$I(h) = 4\pi \sum_k a_{2k} k h^{2k} \prod_{i=1}^{k} \frac{2i-1}{2i}.$$

Therefore, $I(h) = 0$ if and only if (C.1) holds. By Theorem 2.4 in [21], the solution h^* of $I(h^*) = 0$ then implies the existence of a limit cycle in a neighborhood of the oval $\gamma_{h^*} \coloneqq \{z : \|z\| = h^*\}$. ∎

Finally, it is easy to verify that for $\phi(y) = \frac{1}{2}y^2 - \frac{1}{4}y^4$, the condition (C.1) is satisfied with $h^* = \sqrt{\frac{4}{3}}$, thus implying the existence of a spurious ICT set near the neighborhood of $\{z : \|z\| = \sqrt{\frac{4}{3}}\}$.

C.2. **Proof of spurious ICT sets in Example 5.2.** We show the existence of two spurious ICT sets in Example 5.2.

The mean dynamics (MD) for (9) reads:

$$\begin{cases} \dot{x} = -(y - 0.5) - \frac{1}{2}x + 2x^3 - x^5 \\ \dot{y} = x - \frac{1}{2}y + 2y^3 - y^5 \end{cases}. \tag{C.4}$$

Define $r^2 \coloneqq x^2 + y^2$. Then straightforward calculations show that:

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}r^2 &= x\dot{x} + y\dot{y} \\ &= -x(y - 0.5) - \frac{1}{2}x^2 + 2x^4 - x^6 + xy - \frac{1}{2}y^2 + 2y^4 - y^6 \\ &= 0.5x - \frac{1}{2}r^2 + 2r^4 - r^6 + 3x^4y^2 + 3x^2y^4 - 4x^2y^2 \\ &= 0.5x - \frac{1}{2}r^2 + 2r^4 - r^6 + x^2y^2\left(3r^2 - 4\right). \end{aligned} \tag{C.5}$$

Substituting the value $r^2 = \frac{4}{3}$ into (C.5), we get

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}r^2 &= 0.5x + \frac{1}{2}\cdot\frac{4}{3} + 2\cdot\frac{16}{9} - \frac{64}{27} \\ &= 0.5x + \frac{14}{27} \\ &> 0 \end{aligned}$$

since $|x| \le \sqrt{\frac{4}{3}}$ on $\{r \ge 0 : r^2 = \frac{4}{3}\}$, whence $\dot{r} > 0$ on $\{r \ge 0 : r^2 = \frac{4}{3}\}$. Likewise, one can check that $\dot{r} < 0$ on $\{r \ge 0 : r^2 = 2\}$, and that there is no stationary point in the region $\mathcal{S} \coloneqq \{r \ge 0 : \frac{4}{3} \le r^2 \le 2\}$. By the Poincaré-Bendixson theorem [79], there exists at least a limit cycle in $\mathcal{S}$.

Finally, it is easy to see that $(x^*, y^*) = (0, 0.5)$ is a stable critical point of (9). Since the region $\mathcal{S}$ is trapping, Poincaré's index theorem then dictates that there exists at least another unstable limit cycle inside $\mathcal{S}$, establishing the claim.

C.3. **Second-order methods in Example 5.3 as perturbations.** In this section, we discuss how to cast existing second-order methods as an RM scheme with different driving vector fields, and show that their ICT sets are similar to the first-order methods under practical settings.

We will showcase on the *consensus optimization* (ConO):

$$Z_{n+1} = Z_n + \gamma_n(I - \lambda J(Z_n))V(Z_n) \tag{ConO}$$

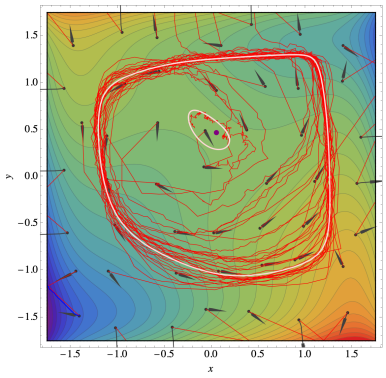


Figure 4: ConO with $\lambda = 0.2$ applied to (9).

where $\lambda > 0$ is the regularization parameter. Recalling the efficient implementation scheme of Hessian-gradient multiplication [67], we make the following assumption on the *stochastic second-order oracles* (SSO): when called at $z = (x, y)$ with random seed $\omega' \in \Omega$, an SSO returns a random vector $\mathsf{JV}(z; \omega')$ of the form

$$\mathsf{JV}(z; \omega') = J(z)V(z) + \mathsf{U}'(z; \omega') \tag{SSO}$$

where $\mathsf{U}'(z; \omega')$ is assumed to be unbiased and sub-Gaussian as in (2). With these assumptions, one can then proceed exactly as in Appendix A.3 for the (SGDA) and (alt-SGDA) cases to show that ConO, and its alternating version, give rise to asymptotic pseudotrajectories of the continuous-time dynamics:

$$\dot{z}(t) = \Big(I - \lambda J(z(t))\Big) V(z(t)).$$

Similarly, one can show (under appropriate assumptions of the oracles) the continuous-time dynamics of *symplectic gradient adjustment* (SGA) is

$$\dot{z}(t) = \left(I - \lambda \left(\frac{J(z(t)) - J(z(t))^\top}{2}\right)\right) V(z(t)).$$

As explained in Example 5.3, it is undesirable to set a large number of λ, since then we are essentially treating $\min\max$ and $\max\min$ as the same problem. However, if λ is small, then by continuity, any stable (unstable) ICT set of (MD) remains stable (unstable) under perturbations [79]. We therefore expect the ICT sets of various second-order algorithms in Example 5.3 be to similar to that of first-order RM schemes.

C.4. **Further comparisons.** This section includes further comparison of the ICT sets of various algorithms, and show that these existing methods all suffer from the spurious convergence depicted in Section 5.

First, Fig. 4 demonstrates that the spurious ICT sets of ConO for (9) is similar to that of SGA; cf. Fig. 2(c).

Second, we have included yet another second-order method, the *Competitive Gradient Descent* (CGD) [77], in Fig. 5(a). For ease of comparison, we run (OG/PEG) with the same initialization in Fig. 5(b). As is evident from the figure, both algorithms perform similarly and converge straight to the spurious ICT set.

Finally, we report the bahvior of various algorithms applied to the "almost bilinear game" (8). In this case, all algorithms fail to escape the spurious ICT set, with the sole exception of ConO. Intriguingly, ConO converges to the *unstable* critical point. A plausible explanation of this phenomenon is provided by [1], where it is shown that the Hamiltonian descent (HD)

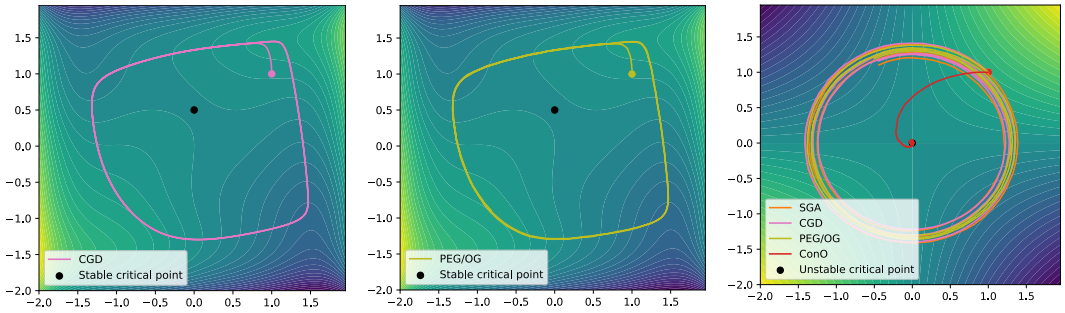


Figure 5: Spurious limits of min-max optimization algorithms from the same initialization. From left to right: (*a*) CGA for (9); (*b*) (OG/PEG) for (9); (*c*) Algorithms for (8).

converges to critical points for any almost bilinear game. Therefore, it is not surprising that ConO, being a mixture of SGDA and HD, also enjoys similar guarantees. Such a convergence is nonetheless highly undesirable in our example, echoing the concern that gradient penalty schemes cannot distinguish (local) min max from max min.

References

[1] Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

[2] Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 486–495, 2019.

[3] Kenneth Joseph Arrow, Leonid Hurwicz, and Hirofumi Uzawa. *Studies in linear and non-linear programming*. Stanford University Press, 1958.

[4] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of extragradient for a whole spectrum of differentiable games. *arXiv preprint arXiv:1906.05945*, 2019.

[5] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363, 2018.

[6] Michel Benaïm. Dynamics of stochastic approximation algorithms. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, editors, *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer Berlin Heidelberg, 1999.

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

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

[9] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

[10] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions, part II: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

[11] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

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

[13] Sebastian Bervoets, Mario Bravo, and Mathieu Faure. Learning with minimal information in continuous games. https://arxiv.org/abs/1806.11506, 2018.

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

[15] Odile Brandière and Marie Duflo. Les algorithmes stochastiques contournent-ils les pièges ? *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 32(3):395–427, 1996.

[16] Mario Bravo and Panayotis Mertikopoulos. On the robustness of learning in games with stochastically perturbed payoff observations. *Games and Economic Behavior*, 103, John Nash Memorial issue:41–66, May 2017.

[17] Mario Bravo, David S. Leslie, and Panayotis Mertikopoulos. Bandit learning in concave N-person games. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[18] Donald Lyman Burkholder. Distribution function inequalities for martingales. *Annals of Probability*, 1 (1):19–42, 1973.

[19] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[20] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT '12: Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

[21] Colin Christopher and Chengzhi Li. *Limit cycles of differential equations*. Springer Science & Business Media, 2007.

[22] Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[23] Charles Cameron Conley. *Isolated Invariant Set and the Morse Index*. American Mathematical Society, Providence, RI, 1978.

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

[25] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[26] Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. *arXiv preprint arXiv:2002.06277*, 2020.

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

[28] Lampros Flokas, Emmanouil Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. 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.

[29] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — Online stochastic gradient for tensor decomposition. In *COLT '15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

[30] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

[31] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019.

[32] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014.

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

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[35] Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pages 2810–2819, 2019.

[36] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6936–6946, 2019.

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

[38] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

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

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

[42] Harold J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer, 1978.

[43] Harold J. Kushner and G. G. Yin. *Stochastic approximation algorithms and applications*. Springer-Verlag, New York, NY, 1997.

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

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

[46] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915, 2019.

[47] Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019.

[48] Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Minyi Hong, and Una-May Obelilly. Min-max optimization without gradients: Convergence and applications to adversarial ml. *arXiv preprint arXiv:1909.13806*, 2019.

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

[50] Lennart Ljung. *System Identification Theory for the User*. Prentice Hall, Englewood Cliffs, NJ, 1986.

[51] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. 2018.

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

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

[54] Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

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

[56] Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

[57] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. 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.

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

[59] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

[60] Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In *Advances in neural information processing systems*, pages 5585–5595, 2017.

[61] Roi Naveiro and David Ríos Insua. Gradient methods for solving stackelberg games. In *International Conference on Algorithmic DecisionTheory*, pages 126–140. Springer, 2019.

[62] Arkadi Semen Nemirovski. 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, 2004.

[63] Arkadi Semen Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

[65] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[66] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14905–14916, 2019.

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

[68] Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic aproximations. *Annals of Probability*, 18(2):698–712, April 1990.

[69] Wei Peng, Yu-Hong Dai, Hui Zhang, and Lizhi Cheng. Training gans with centripetal acceleration. *Optimization Methods and Software*, pages 1–19, 2020.

[70] Steven Perkins and David S. Leslie. Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems*, 2(2):409–446, 2012.

[71] Steven Perkins, Panayotis Mertikopoulos, and David S. Leslie. Mixed-strategy learning with continuous action sets. *IEEE Trans. Autom. Control*, 62(1):379–384, January 2017.

[72] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org, 2017.

[73] Leonid Denisovich Popov. 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.

[74] Arvind Raghunathan, Anoop Cherian, and Devesh Jha. Game theoretic optimization via gradient-based nikaido-isoda function. In *International Conference on Machine Learning*, pages 5291–5300, 2019.

[75] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*, 2013.

[76] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[77] Florian Schäfer and Anima Anandkumar. Competitive gradient descent. In *Advances in Neural Information Processing Systems*, pages 7623–7633, 2019.

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

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

[80] Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017.

[81] Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[82] Guojun Zhang and Yaoliang Yu. Convergence of gradient methods on bilinear zero-sum games, 2019.