

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

YA-PING HSIEH\*, PANAYOTIS MERTIKOPOULOS<sup>◊,‡</sup>, AND VOLKAN CEVHER\*

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

1. INTRODUCTION

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

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

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

*Where does the algorithm converge to?* (★)

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

---

\* 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.

The authors are grateful to Thomas Pethick for his help in the numerical simulation of adaptive methods. This research was partially supported by the COST Action CA16228 “European Network for Game Theory” (GAMENET), the Army Research Office under grant number W911NF-19-1-0404, the Swiss National Science Foundation (SNSF) under grant number 200021\_178865 / 1, the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement n° 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).

**Related work.** This question has attracted significant interest in the machine learning literature because of its potential implications to generative adversarial networks [32], robust reinforcement learning [74], and other models of adversarial training [56]. In this broad setting, it has become empirically clear that the joint training of two neural networks (NNs) is fundamentally more difficult than that of a *single* NN of similar size and architecture. The latter task boils down to successfully finding a (good) local minimum of a non-convex function, so it is instructive to revisit (★) in the context of *non-convex minimization*.

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

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

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

This well-documented failure of SGDA has led to an extensive literature that is impossible to survey here. As a purely indicative – and highly incomplete – list, we mention the works of Daskalakis et al [21], Gidel et al [30], Mertikopoulos et al [59] and Mokhtari et al [62], who studied how these failures can be overcome in *deterministic* bilinear problems by means of an *extra-gradient* step (or an optimistic proxy thereof). By contrast, in *stochastic* problems, the convergence of optimistic / extra-gradient methods is compromised unless additional, tailor-made mitigation mechanisms are put in place – such as variance reduction [16, 40] or variable step-size schedules [38]. This shows that the convergence of min-max training methods can be particularly fragile, even in simple, bilinear problems.

Beyond the class of convex-concave problems analyzed above, another vigorous thread of research has focused on the *local analysis* of a min-max optimization algorithm close to the game’s critical points – typically subject to a second-order sufficient condition; cf. Adolphs et al [2], Daskalakis and Panageas [20], Fiez and Ratliff [25], Grimmer et al [33, 34], Heusel et al [36], Mazumdar et al [58], Nagarajan and Kolter [64]. The global analysis is much more challenging and requires strong structural assumptions such as variational coherence [59] and/or the existence of a Minty-type solution [53]. In the absence of such conditions, Flokas et al [27, 28] showed that periodic and/or Poincaré recurrent behavior may persist in deterministic, continuous-time min-max dynamics.

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

**Our contributions.** Our first contribution is to provide a unified framework for a comprehensive selection of first- and zeroth-order min-max optimization methods (including SGDA, proximal point methods, optimistic / extra-gradient schemes, their alternating variants, etc.). The principal ingredients of our approach are twofold: (i) a generalized Robbins–Monro (RM) template that is wide enough to include all the above algorithms; and (ii) an analytic framework leveraging the ordinary differential equation (ODE) method of stochastic approximation [7, 50]. Based on these two elements, we prove a precise version

of the following general principle: *the long-run behavior of all generalized RM methods can be mapped to the study of **the same**, mean-field dynamical system.*

In more detail, we show that the limit points of all generalized RM schemes belong to an *internally chain-transitive* (ICT) set of these mean dynamics. The notion of an ICT set is central in the study of dynamical systems [9, 13, 19] and, in some cases, they are easy to characterize: in minimization problems (and possibly up to a “hidden” transformation in the spirit of 27), the dynamics’ ICT sets are the function’s critical points. As such, in this case, we recover *exactly* the min-min landscape of SGD – but for an *entire family* of algorithms, not just SGD.

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

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

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

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

Our results indicate a steep, qualitative increase in difficulty when passing from min-min to min-max problems, in line with concurrent works by Daskalakis et al [22] and Letcher [52]. In plain terms, Daskalakis et al [22] proved the impossibility of attaining a critical point in polynomial time in deterministic, constrained min-max games. In a similar spirit, the concurrent work of Letcher [52] showed that there are min-max games where all “reasonable” deterministic algorithms may fail to converge. By contrast, our paper focuses on the occurrence of *spurious convergence phenomena* with probability 1 in *stochastic* algorithms. In addition, our avoidance result (Theorem 3) can be seen as a stochastic counterpart of the “reasonableness” requirement of Letcher [52], thereby enriching the applicability of the results therein. Taken together, these works and our own provide a complementing look into the fundamental limits of min-max optimization algorithms.

## 2. SETUP AND PRELIMINARIES

Throughout our paper, we focus on general unconstrained problems with  $\mathcal{X} = \mathbb{R}^{d_x}$ ,  $\mathcal{Y} = \mathbb{R}^{d_y}$ , and  $\Phi$  assumed  $C^1$  and Lipschitz. To avoid unnecessary notation, we will let

$z = (x, y)$ ,  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$  and  $d = d_{\mathcal{X}} + d_{\mathcal{Y}}$ . In addition, we will write

$$V(z) \equiv (V_x(x, y), V_y(x, y)) := (-\nabla_x \Phi(x, y), \nabla_y \Phi(x, y)) \quad (1)$$

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

$$\langle V(z), z \rangle \leq 0 \quad \text{for all sufficiently large } z. \quad (2)$$

This condition is a weaker version of standard coercivity conditions in the literature [6], it is satisfied by all convex-concave problems (including bilinear ones) and, importantly, it does not impose any growth requirements on the elements of  $V$  (as standard coercivity conditions do). We discuss it further in [Appendix A](#).

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

From an algorithmic standpoint, we will focus exclusively on the black-box optimization paradigm [68] with *stochastic first-order oracle* (SFO) feedback. Algorithms with a more complicated feedback structure, such as a best-response oracle [26, 41, 65] or based on mixed-strategy sampling [23, 39, 43], are not considered in this work.

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

$$\mathbf{V}(z; \omega) = V(z) + \mathbf{U}(z; \omega) \quad (\text{SFO})$$

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

$$\forall z \in \mathcal{Z}, \quad \mathbb{E}[\mathbf{U}(z; \omega)] = 0 \quad \text{and} \quad \mathbb{E}[\|\mathbf{U}(z; \omega)\|^2] \leq \sigma^2. \quad (3)$$

These will be our blanket assumptions throughout.

### 3. CORE ALGORITHMIC FRAMEWORK

**3.1. The Robbins–Monro template.** Much of our analysis will revolve around iterative algorithms that can be cast as generalized Robbins–Monro algorithms [77] of the general form

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

where:

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

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

$$V_n = V(Z_n) + W_n \quad (4)$$

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

$$W_n = U_n + b_n \quad (5)$$

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

$$a) \text{ Bias: } B_n = \mathbb{E}[\|b_n\| | \mathcal{F}_n] \quad (6a)$$

$$b) \text{ Variance: } \sigma_n^2 = \mathbb{E}[\|U_n\|^2 | \mathcal{F}_n] \quad (6b)$$

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

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

▼ **Algorithm 1** (Stochastic gradient descent/ascent). The basic stochastic gradient descent/ascent (SGDA) algorithm – also known as the *Arrow–Hurwicz* method [4] – queries an SFO and proceeds as:

$$Z_{n+1} = Z_n + \gamma_n \mathbf{V}(Z_n; \omega_n), \quad (\text{SGDA})$$

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  $W_n = U_n = \mathbf{U}(Z_n; \omega_n)$  and  $b_n = 0$ . ▲

▼ **Algorithm 2** (Proximal point method). The (deterministic) *proximal point method* (PPM) [79] is an implicit update rule of the form:

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

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

▼ **Algorithm 3** (Stochastic extra-gradient). Since (PPM) is only implicitly defined, one can rarely run it in practice. Nonetheless, it is possible to approximate (PPM) by locally querying two (stochastic) gradients at each iteration [67]. This can be achieved by the *stochastic extra-gradient* (SEG):

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

To recast (SEG) in the Robbins–Monro framework, simply take  $W_n = \mathbf{V}(Z_n^+; \omega_n^+) - V(Z_n)$ , i.e.,  $U_n = \mathbf{U}(Z_n^+; \omega_n^+)$  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 [75]:

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

Popov’s extra-gradient has been rediscovered several times and is more widely known as the optimistic gradient (OG) method in the machine learning literature [17, 21, 37, 76]. In unconstrained problems, (OG/PEG) turns out to be equivalent to a number of other existing methods, including “extrapolation from the past” [30] and reflected gradient [57]. Its Robbins–Monro representation is obtained by setting  $W_n = \mathbf{V}(Z_n^+; \omega_n) - V(Z_n)$ , i.e.,  $U_n = \mathbf{U}(Z_n^+; \omega_n)$  and  $b_n = V(Z_n^+) - V(Z_n)$ . ▲

▼ **Algorithm 5** (Kiefer–Wolfowitz). When first-order feedback is unavailable, a popular alternative is to obtain gradient information of  $\Phi$  via zeroth-order observations [54]. This idea can be traced back to the seminal work of Kiefer and Wolfowitz [45] and the subsequent development of the simultaneous perturbation stochastic approximation (SPSA) method by Spall [83]. 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,  $\omega_n$  is drawn uniformly at random from the composite basis  $\Omega = \mathcal{E}_X \cup \mathcal{E}_Y$  of  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ , and the “ $\pm$ ” sign is equal to  $-1$  if  $\omega_n \in \mathcal{E}_X$  and  $+1$  if  $\omega_n \in \mathcal{E}_Y$ . Viewed this way, the interpretation of (SPSA) as a Robbins–Monro method is immediate; furthermore, a straightforward calculation (that we defer to Appendix B.3) shows that the sequence of gradient estimators  $V_n$  in (SPSA) has  $B_n = \mathcal{O}(\delta_n)$  and  $\sigma_n^2 = \mathcal{O}(1/\delta_n^2)$ . ▲

Further examples that can be cast in the RM framework include the negative momentum method [31], generalized OG schemes [63], the Chambolle–Pock algorithm [15], the “prediction method” of Yadav et al [86], and centripetal acceleration [72]; the analysis is similar and we omit the details. Certain scalable second-order methods can also be viewed as RM schemes, but the driving vector field  $V$  is no longer the gradient field of  $\Phi$ ; we discuss this in the supplement.

**3.3. Alternating updates and moving averages.** There are two extremely common heuristics for practitioners in applying min-max algorithms to real applications: alternating and averaging. An *alternating* algorithm for (SP) updates the  $x$  and  $y$  variables sequentially (instead of simultaneously as in Section 3.2). An *averaged* algorithm takes the next state as a convex combination of  $Z_n$  and  $Z_{n+1}$  in (RM), cf. [44].

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

**Lemma 1.** *Let  $Z_{n+1} = Z_n + \gamma_n[V(Z_n) + W_n]$  be an RM scheme where  $W_n = U_n + b_n$  as in (5). Then its  $\alpha$ -averaged version (where  $0 < \alpha < 1$ ), defined as*

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

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

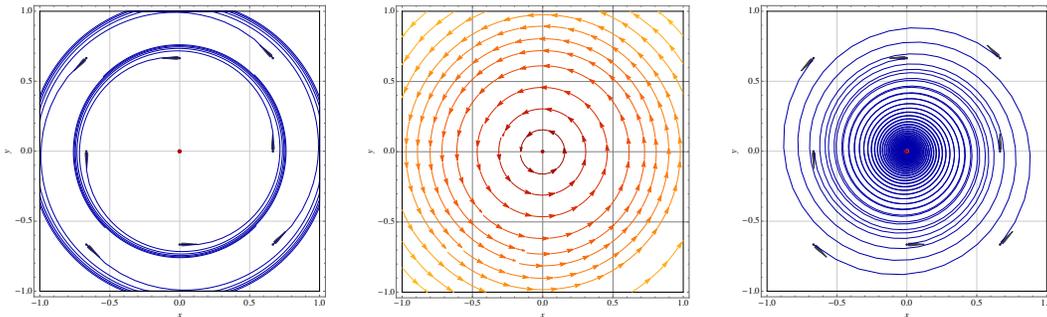
*Remark 3.1.* Lemma 1 can be easily adapted to the scenario where one only averages either the  $X_n$  or  $Y_n$  variable.

**Lemma 2.** *Let  $Z_{n+1} = Z_n + \gamma_n[V(Z_n) + W_n]$  be an RM scheme where  $W_n = U_n + b_n$  as in (5). Then its alternating version, defined as*

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

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

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



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

*Remark 3.2.* One can easily generalize Lemma 2 to the “ $(k_1, k_2)$ -RM schemes” where one performs  $k_1$  updates for  $x$  and then  $k_2$  updates for  $y$  (here  $k_1, k_2 \in \mathbb{N}$  are arbitrary but fixed). The resulting scheme will still be an RM scheme. In particular, our framework captures the popular  $(k_1, k_2) = (1, 5)$  variant of (SGDA) used in the seminal works of Goodfellow et al [32] and Arjovsky et al [3]. In view of Lemmas 1–2, Remark 3.2, and a simple calculation (see (B.18)), all of our results on first-order methods (e.g., Algorithms 1–4) apply also to their averaging/alternating and the more general  $(k_1, k_2)$  versions.

#### 4. CONVERGENCE ANALYSIS

**4.1. Overview: Continuous vs. discrete time.** The key in providing a unified treatment of all algorithms in Section 3 is the reduction of (RM) to the *mean dynamics*

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

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

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

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

As a result, we can rewrite Algorithm 3 as

$$\frac{Z_{n+1} - Z_n}{\gamma_n} = V(Z_n) + \mathcal{O}(\gamma_n). \quad (7)$$

If  $\gamma_n \searrow 0$ , we should then expect (7) to converge to (MD). More generally, if the error term  $W_n$  in (RM) is sufficiently well-behaved, we should expect the iterates of (RM) and the solutions of (MD) to eventually come together.

Connecting (RM) to (MD) has proved very fruitful when the latter comprises a *gradient system*, i.e.,  $V = -\nabla f$  for some (possibly non-convex)  $f: \mathcal{Z} \rightarrow \mathbb{R}$ : Modulo mild assumptions, the systems (RM) and (MD) are known to both converge to the critical set of  $f$  [10, 11, 48, 50, 55].

On the other hand, bona fide min-max problems are considerably more involved. The most widely known illustration is given by the bilinear objective  $\Phi(x, y) = xy$ : in this case (see Fig. 1), the trajectories (MD) comprise periodic orbits of perfect circles centered at the origin (the unique critical point of  $\Phi$ ). However, the behavior of different RM schemes can

vary wildly, even in the absence of noise ( $\sigma = 0$ ): trajectories of (SGDA) spiral outwards, each converging to an initialization-dependent periodic orbit; instead, (SEG) trajectories spiral inwards, eventually converging to the solution  $z^* = (0, 0)$ .

This particular difference between gradient and extra-gradient schemes has been well-documented in the literature [21, 30, 59]. More pertinent to our theory, it also raises several key questions:

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

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

**4.2. Connecting (RM) to (MD).** We begin by introducing a measure of “closeness” between the iterates of (RM) and the solution orbits of (MD). To do so, let  $\tau_n = \sum_{k=1}^n \gamma_k$  denote the “effective time” that has elapsed at the  $n$ -th iteration of (RM), and define the continuous-time interpolation  $Z(t)$  of  $Z_n$  as

$$Z(t) = Z_n + \frac{t - \tau_n}{\tau_{n+1} - \tau_n} (Z_{n+1} - Z_n) \quad (8)$$

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  $\Theta: \mathbb{R}_+ \times \mathcal{Z} \rightarrow \mathcal{Z}$  of (MD), which is simply the orbit of (MD) at time  $t \in \mathbb{R}_+$  with an initial condition  $z(0) = z \in \mathcal{Z}$ . We then have the following notion of “asymptotic closeness”:

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

$$\lim_{t \rightarrow \infty} \sup_{0 \leq h \leq T} \|Z(t+h) - \Theta_h(Z(t))\| = 0. \quad (9)$$

This comparison criterion is due to Benaïm and Hirsch [9] and it plays a central role in our analysis. In words, it simply posits that  $Z(t)$  eventually tracks the flow of (MD) with arbitrary accuracy over windows of arbitrary length; as a result, if  $Z_n$  is an APT of (MD), it is reasonable to expect its behavior to be closely correlated to that of (MD).

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

$$B_n \rightarrow 0 \text{ (a.s.)} \quad \text{and} \quad \sum_{n=1}^{\infty} \mathbb{E}[\gamma_n B_n] < \infty \quad (\text{A1})$$

$$\sum_{n=1}^{\infty} \mathbb{E}[\gamma_n^2 (1 + B_n^2 + \sigma_n^2)] < \infty \quad (\text{A2})$$

We then have:

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

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

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

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

- a) For first-order methods (*Algorithms 1–4*), the algorithm is run with SFO feedback satisfying (3) and a step-size  $\gamma_n$  such that  $A/n \leq \gamma_n \leq B/\sqrt{n(\log n)^{1+\varepsilon}}$  for some  $A, B, \varepsilon > 0$ .
- b) For zeroth-order methods (*Algorithm 5*), the algorithm is run with parameters  $\gamma_n$  and  $\delta_n$  such that  $\lim_n(\gamma_n + \delta_n) = 0$ ,  $\sum_n \gamma_n = \infty$ , and  $\sum_n \gamma_n^2/\delta_n^2 < \infty$  (e.g.,  $\gamma_n = 1/n$ ,  $\delta_n = 1/n^{1/3}$ ).

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

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

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

- (I) SGD converges to the function’s set of critical points [[11](#), [55](#)].
- (II) SGD avoids unstable critical points [[29](#), [60](#), [70](#)].

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

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

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

**Definition 2** ([7](#)). Let  $\mathcal{S}$  be a nonempty compact subset of  $\mathcal{Z}$ . Then:

- a)  $\mathcal{S}$  is *invariant* if  $\Theta_t(\mathcal{S}) = \mathcal{S}$  for all  $t \in \mathbb{R}$ .
- b)  $\mathcal{S}$  is *attracting* if it is invariant and there exists a compact neighborhood  $\mathcal{K}$  of  $\mathcal{S}$  such that  $\lim_{t \rightarrow \infty} \text{dist}(\Theta_t(z), \mathcal{S}) = 0$  uniformly in  $z \in \mathcal{K}$ .
- c)  $\mathcal{S}$  is *internally chain-transitive* (ICT) if it is invariant and  $\Theta|_{\mathcal{S}}$  admits no proper attractors in  $\mathcal{S}$ .

*Remark.* Equivalently, ICT sets can be viewed as “minimal connected periodic orbits up to arbitrarily small numerical errors”, cf. Benaïm [[7](#), Prop. 5.3]. The definition above is more convenient to work with because it provides the key insights in [Section 4.4.3](#) below.

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

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

**Corollary 1.** *Let  $Z_n$  be a sequence generated by any of the [Algorithms 1–5](#) with parameters as in [Proposition 1](#). Then  $Z_n$  converges almost surely to an ICT set of  $\Phi$ .*

4.4.2. *Avoidance of unstable points and sets.* Our next result provides an avoidance result for RM schemes in min-max optimization. In analogy with function minimization problems, we will focus on unstable *invariant sets* of (MD), i.e., invariant sets that admit a nontrivial unstable manifold (for an in-depth discussion and precise definition, see 81 and Appendix C.1).

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

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

$$\mathbb{P}(\lim_{n \rightarrow \infty} \text{dist}(Z_n, \mathcal{K}) = 0) = 0.$$

*Remark 4.1.* We note that Assumptions (i) and (ii) above are standard in the literature for avoidance results of SGD [7, 60, 70], and are significantly lighter than other “isotropic noise” assumptions that are common in the literature [29]. Specifically, even though Assumption (ii) looks somewhat obscure, it only posits that the noise is not degeneratively equal to zero along certain directions in space; for a more detailed discussion, see Appendix C.1. We also stress that neither of these assumptions is required for the rest of our paper.

4.4.3. *When do RM schemes behave the same?* So far, we have successfully generalized (I) and (II) to the context of (SP) as follows:<sup>1</sup>

- (I-SP) RM schemes always converge to ICT sets, and
- (II-SP) RM schemes always avoid invariant sets.

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

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

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

Importantly, the celebrated *Kupka-Smale theorem* [47, 82] asserts that systems with degenerate periodic orbits (such as bilinear games) occur “almost never” in the Baire category sense. More precisely, an arbitrarily small perturbation can fundamentally destroy the topological properties of their ICT sets and give rise to proper, non-trivial attractors; cf. Example 5.1. In contrast, systems with nontrivial attractors are known to be robust under

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

perturbations [81], and our final result in the section shows that it is precisely the *existence of nontrivial attractors* that makes the discrepancy of RM schemes disappear, at least locally.

**Theorem 4.** *Let  $\mathcal{S}$  be an attractor of (MD) and fix some confidence level  $\alpha > 0$ . If  $\gamma_n$  is small enough and Assumptions (A1)–(A2) hold, there exists a neighborhood  $\mathcal{U}$  of  $\mathcal{S}$ , independent of  $\alpha$ , such that  $\mathbb{P}(Z_n \text{ converges to } \mathcal{S}) \geq 1 - \alpha$  if  $Z_1 \in \mathcal{U}$ .*

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

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

On a positive note, since the Hartman-Grobman Theorem [78] implies that all critical points of  $\Phi$  with  $\Re\{\lambda(JV(z^*))\} < 0$  for all eigenvalues  $\lambda$  are attractors of (MD), Theorem 4 immediately yields:

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

Corollary 3 generalizes the local convergence of deterministic SGDA and SEG studied by Daskalakis and Panageas [20]. It also extends [37, Theorem 5] from (OG/PEG) to all generalized RM schemes.

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

## 5. SPURIOUS ATTRACTORS: ILLUSTRATIONS AND EXAMPLES

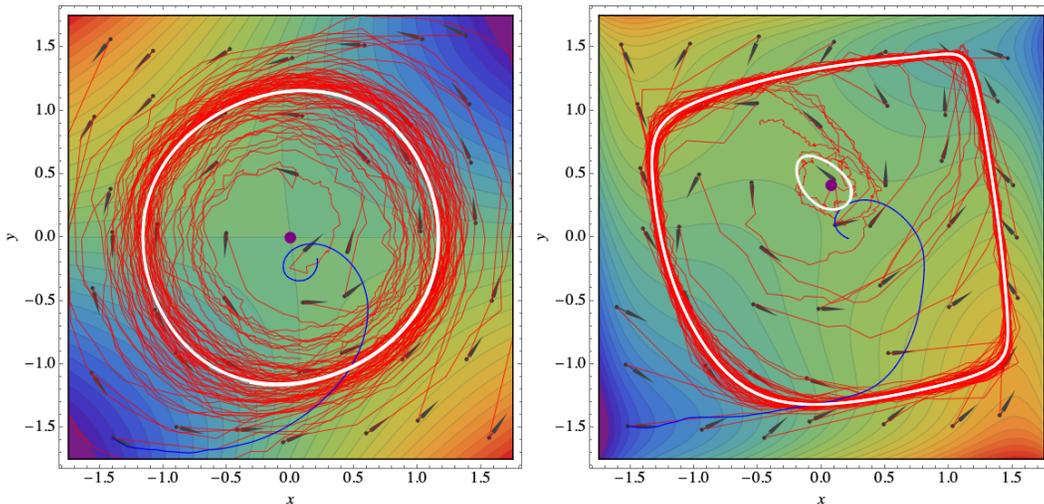
In this last section, we provide a range of simple examples that exhibit *spurious attractors* – i.e., attractors that consist entirely of non-critical points. For illustration purposes, we focus on the simple case  $\mathcal{X} = \mathcal{Y} = \mathbb{R}$  with polynomial objectives. In doing this, our goal is to highlight a number of issues that can arise in min-max optimization problems; whether limit cycles of this type occur in actual large-scale experiments – e.g., in GANs – is an open research question [52].

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

$$\Phi(x, y) = xy + \varepsilon\phi(y), \quad (10)$$

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

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



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

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

$$\Phi(x, y) = x(y - 0.45) + \phi(x) - \phi(y) \quad (11)$$

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

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

- (1) Instead of the “moving average” in Lemma 1, one can take the *ergodic average* ( $Z'_n = \frac{1}{n} \sum_{k=1}^n Z_k$ ) as is customary in convex-concave problems [42, 66]. We plot one such trajectory in Fig. 2 (the blue curves). Evidently, we see that ergodic average can force the algorithms to halt at non-critical points, and this convergence is by no means min-max optimal.
- (2) Many recent works attempt to address the cycling issues of min-max algorithms via incorporating *second-order* oracles. For completeness, we also study a range of popular second-order methods in Appendix D.3. Our analysis shows that these algorithms suffer similar symptoms as first-order schemes in our examples, cf. Figs. 4–5.
- (3) In addition to the diminishing step-size policies studied here, another common strategy in practice is to simply set  $\gamma_n$  to a *constant step-size*. While our analysis does not cover this setting, there exist several techniques in stochastic approximation to boost from our “almost surely” statements for  $\gamma_n \searrow 0$  to *concentration* or *high-probability* results when  $\gamma_n \equiv \gamma$  is small [12, 49, 50].

For completeness, in [Section 4.4](#) we examine various constant step-size RM schemes applied to [\(10\)](#) and [\(11\)](#). The outcome coincides with our intuition that these schemes should concentrate around the spurious attracting ICT sets, and hence exhibit similar behaviors as RM schemes with  $\gamma_n \searrow 0$ ; see [Fig. 6](#).

- (4) *Adaptive methods* such as Adam [\[46\]](#) are ubiquitous in GAN training. We study such methods in [Section 4.5](#): our results show that they fail solve the simple objectives [\(10\)](#) and [\(11\)](#). Moreover, some methods even show a potentially detrimental tendency of converging to *max-min points*, the exact opposite of desirable solutions; see [Fig. 7](#).

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

#### APPENDIX A. STABILIZATION OF RM SCHEMES

Our aim in this appendix will be to prove the stability of generalized RM schemes, namely that asymptotic pseudotrajectories generated by (RM) are bounded with probability 1. The key ingredient in our analysis is the weak asymptotic coercivity (WAC) condition [\(2\)](#), which, as we discussed in the main body of our paper, is a relaxation of the standard coercivity requirement

$$\lim_{\|z\| \rightarrow \infty} \frac{\langle V(z), z \rangle}{\|z\|} = -\infty. \quad (\text{A.1})$$

The hypothesis [\(A.1\)](#) is a mainstay in the analysis of monotone operators and variational inequalities [\[6, 24, 73\]](#). Roughly speaking, it states that the “radial component”

$$V_r(z) = \frac{\langle V(z), z \rangle}{\|z\|} \quad (\text{A.2})$$

of  $V(z)$  grows to  $-\infty$  as  $\|z\| \rightarrow \infty$ . In other words, any vector field that satisfies [\(A.1\)](#) has an inward-pointing component that grows infinitely large for large  $\|z\|$ .

In view of the above, the coercivity assumption [\(A.1\)](#) suggests that any process that takes successive steps along  $V(z)$  will be subject to an “inwards drift” towards regions with smaller norm, and this drift will be more and more pronounced the farther one moves away from the origin. On that account, [\(A.1\)](#) is a natural candidate for showing that RM processes based on  $V$  never escape to infinity. On the other hand, vector fields that do not have a strong radial component – such as the bilinear game field  $V(x, y) = (-y, x)$  which has  $V_r(x, y) = 0$  – are not covered by [\(A.1\)](#). In this regard, the WAC condition [\(2\)](#) provides an important relaxation of [\(A.1\)](#), because it only posits that the radial component of  $V(z)$  is asymptotically non-positive – or, more simply, that  $V(z)$  does not have a persistent outward-pointing component.

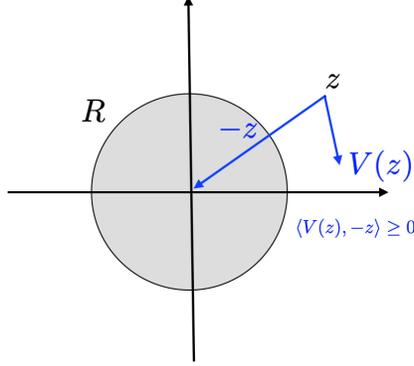
Before proving the stability of generalized RM schemes under [\(2\)](#), we provide below a series of important examples that satisfy the WAC condition [\(2\)](#):

- (1)  $V$  satisfies [\(A.1\)](#). Indeed, in this case, for all  $M > 0$ , there exists some  $R \equiv R(M)$  such that

$$\frac{\langle V(z), z \rangle}{R} \leq \frac{\langle V(z), z \rangle}{\|z\|} \leq -M < 0 \quad (\text{A.3})$$

whenever  $\|z\| \geq R$ , i.e., [\(2\)](#) holds

- (2)  $\Phi$  is convex-concave and it admits a critical point. By shifting the problem’s frame of reference if necessary, we can assume without loss of generality that  $z^* = 0$  is a



**Figure 3:** Schematic illustration of the weak asymptotic coercivity condition (2).

critical point of  $\Phi$ . Then, with  $\Phi$  assumed convex-concave, we readily get  $\langle V(z), z \rangle \leq \langle V(z^*), z - z^* \rangle = 0$ , i.e., (2) holds

The first item above justifies the terminology “weak asymptotic coercivity”; for a geometric illustration, see Fig. 3 below.

We now proceed to establish our main stability result for generalized RM schemes under the WAC condition (2):

**Proposition A.1.** *Suppose that  $V$  satisfies (2). Then, under Assumptions (A1)–(A2), the sequence  $Z_n$  generated by (RM) is bounded (a.s.).*

*Proof of Proposition A.1.* Our proof hinges on the introduction of a suitable “energy function” for (MD). To define it, recall that that  $\langle V(z), -z \rangle \geq 0$  whenever  $\|z\| \geq R$ . Then, with a fair amount of hindsight, fix some  $\lambda > 0$  and let

$$E(z) = \begin{cases} 0 & \text{if } \|z\| \leq R, \\ (\|z\| - R)^2/2 & \text{if } R \leq \|z\| \leq (1 + \lambda)R, \\ \lambda R\|z\| - \lambda(1 + \lambda/2)R^2 & \text{if } (1 + \lambda)R \leq \|z\|. \end{cases} \quad (\text{A.4})$$

By a direct calculation, we can verify the following:

- (1)  $E$  is continuously differentiable and its gradient is given by  $\nabla E(z) = \phi(\|z\|/R)z$  where  $\phi(u) = 0$  if  $u \leq 1$ ,  $\phi(u) = 1 - 1/u$  if  $1 \leq u \leq 1 + \lambda$ , and  $\phi(u) = \lambda/u$  if  $1 + \lambda \leq u$ .
- (2)  $E$  is negatively correlated to  $V$ , i.e.,  $\langle V(z), -\nabla E(z) \rangle \geq 0$  for all  $z \in \mathcal{Z}$ .
- (3)  $E$  is 1-smooth, i.e.,  $E(z') \leq E(z) + \langle \nabla E(z), z' - z \rangle + (1/2)\|z' - z\|^2$  for all  $z, z' \in \mathcal{Z}$ .

Then, letting  $E_n = E(Z_n)$  and  $\phi_n = \phi(\|Z_n\|/R)$ , we get

$$\begin{aligned} E_{n+1} &= E(Z_n - \gamma_n V_n) \leq E(Z_n) - \gamma_n \langle \nabla E(Z_n), V_n \rangle + \frac{\gamma_n^2}{2} \|V_n\|^2 \\ &\leq E_n - \gamma_n \phi_n \langle U_n + b_n, Z_n \rangle + \frac{3\gamma_n^2}{2} [\|V(Z_n)\|^2 + \|U_n\|^2 + \|b_n\|^2], \end{aligned} \quad (\text{A.5})$$

where the second line follows from the properties of  $E$ , the definition (4) of  $V_n$ , and the Cauchy-Schwarz inequality. Hence, conditioning on  $\mathcal{F}_n$  and taking expectations, we obtain:

$$\mathbb{E}[E_{n+1} | \mathcal{F}_n] \leq E_n + \gamma_n \phi_n \|Z_n\| B_n + \frac{3}{2} \gamma_n^2 [M^2 + B_n^2 + \sigma_n^2], \quad (\text{A.6})$$

where we made a second use of the Cauchy-Schwarz inequality in the term involving  $B_n$  and  $M$  is the Lipschitz constant of  $\Phi$ .

To proceed, let  $\varepsilon_n = \gamma_n \phi_n \|Z_n\| B_n + (3/2) \gamma_n^2 [M^2 + B_n^2 + \sigma_n^2]$  denote the “residual” term in (A.6), and consider the auxiliary process  $\tilde{E}_n = E_{n+1} + \sum_{k=n+1}^{\infty} \varepsilon_k$ . By (A.6), we have  $\mathbb{E}[\tilde{E}_n | \mathcal{F}_n] \leq E_n + \sum_{k=n}^{\infty} \varepsilon_n = \tilde{E}_{n-1}$ , i.e.,  $\tilde{E}_n$  is a supermartingale relative to  $\mathcal{F}_n$ . By the definition of  $\phi$ , we further have  $\phi(u) \leq \lambda/u$ , so  $\phi_n \|Z_n\| \leq \lambda R$  for all  $n$ . We thus get

$$\sum_{n=1}^{\infty} \varepsilon_n \leq \lambda R \sum_{n=1}^{\infty} \gamma_n B_n + \frac{3}{2} \sum_{n=1}^{\infty} \gamma_n^2 (M^2 + B_n^2 + \sigma_n^2) \quad (\text{A.7})$$

and hence, by Assumptions (A1) and (A2), we conclude that  $\mathbb{E}[\sum_n \varepsilon_n] < \infty$ . This shows that  $\mathbb{E}[\tilde{E}_n] \leq \mathbb{E}[\tilde{E}_1] < \infty$ , i.e.,  $\tilde{E}_n$  is uniformly bounded in  $L^1$ . Hence, by Doob’s submartingale convergence theorem [35, Theorem 2.5], it follows that  $\tilde{E}_n$  converges with probability 1 to some finite random limit  $\tilde{E}_\infty$ . In turn, since  $\sum_n \varepsilon_n < \infty$ , this implies that  $E_n = \tilde{E}_{n-1} - \sum_{k=n}^{\infty} \varepsilon_k$  also converges to some (random) finite limit (a.s.). From this we deduce that  $\limsup_n \|Z_n\| < \infty$ , as claimed. ■

## APPENDIX B. PROOF OF THEOREM 1 AND PROPOSITION 1

In this appendix, we discuss how the algorithms 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).

**B.1. Generalities and preliminaries.** Before doing so, we will require some background material on asymptotic pseudotrajectories. Following Benaïm and Hirsch [9] and Benaïm [7], 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 (8) of the continuous-time interpolation  $Z(t)$  of  $Z_n$  as

$$Z(t) = Z_n + \frac{t - \tau_n}{\tau_{n+1} - \tau_n} (Z_{n+1} - Z_n) \quad (8)$$

We will further require the “continuous-to-discrete” correspondence

$$M(t) = \sup\{n \geq 1 : t \geq \tau_n\} \quad (\text{B.1})$$

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

$$M_n \equiv M_n(T) = M(\tau_n + T). \quad (\text{B.2})$$

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

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

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

$$Z(t) = Z(0) + \int_0^t [V(\bar{Z}(s)) + \bar{W}(s)] ds \quad (\text{B.4})$$

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

With all this in hand, Benaïm [7, Prop. 4.1] provides the following general condition for  $Z(t)$  to be an APT of the mean dynamics (9):

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

$$\lim_{t \rightarrow \infty} \Delta(t; T) = 0 \quad \text{for all } T > 0, \quad (\text{B.5})$$

where

$$\Delta(t; T) = \sup_{0 \leq h \leq T} \left\| \int_t^{t+h} \overline{W}(s) ds \right\|. \quad (\text{B.6})$$

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

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

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

*Proof.* Since we have shown that  $Z_n$  remains bounded in Proposition A.1, it suffices to verify (B.5).

Our proof relies on the Burkholder–Davis–Gundy (BDG) inequality [14, 35] which bounds the maximal value of a martingale  $S_n$  via its quadratic variation as

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

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

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

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

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

so our previous bound shows that

$$\mathbb{E}[\Delta_0(t; T)^2] \leq C_2 \int_t^{t+T} \bar{\gamma}^2(s) \bar{\sigma}^2(s) ds. \quad (\text{B.9})$$

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

$$\sum_{k=1}^{\infty} \mathbb{E}[\Delta_0(kT; T)^2] \leq C_2 \int_0^{\infty} \bar{\gamma}^2(s) \bar{\sigma}^2(s) ds = C_2 \sum_{n=1}^{\infty} \gamma_n^2 \sigma_n^2 < \infty \quad (\text{B.10})$$

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

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

<sup>2</sup>Benaïm [7] provides a set of sufficient conditions for (B.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.

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

$$\mathbb{P}\left(\limsup_{k \rightarrow \infty} \mathcal{E}_k(\varepsilon)\right) = 0. \quad (\text{B.12})$$

This shows that, with probability 1, we have  $\Delta_0(kT; T) \leq \varepsilon$  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

$$\mathbb{P}\left(\liminf_{k \rightarrow \infty} \Delta_0(kT; T) > 0\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} \mathcal{E}(1/n)\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(\mathcal{E}(1/n)) = 0, \quad (\text{B.13})$$

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

Thus, going back to the requirements of [Proposition B.1](#), we get

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

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

**B.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:*

- For first-order methods ([Algorithms 1–4](#)), the algorithm is run with SFO feedback satisfying [\(3\)](#) and a step-size  $\gamma_n$  such that  $A/n \leq \gamma_n \leq B/\sqrt{n(\log n)^{1+\varepsilon}}$  for some  $A, B, \varepsilon > 0$ .*
- For zeroth-order methods ([Algorithm 5](#)), the algorithm is run with parameters  $\gamma_n$  and  $\delta_n$  such that  $\lim_n(\gamma_n + \delta_n) = 0$ ,  $\sum_n \gamma_n = \infty$ , and  $\sum_n \gamma_n^2 / \delta_n^2 < \infty$  (e.g.,  $\gamma_n = 1/n$ ,  $\delta_n = 1/n^{1/3}$ ).*

*Then  $Z_n$  is almost surely an APT of [\(MD\)](#).*

*Proof.* We first note that  $\sum_n \gamma_n^2 < \infty$  by our choice of step-sizes. Thus, in order to prove that  $\mathbb{E}[\sum_n \gamma_n B_n] < \infty$ ,  $\mathbb{E}[\sum_n \gamma_n^2 B_n^2] < \infty$ , and  $\mathbb{E}[\sum_n \gamma_n^2 \sigma_n^2] < \infty$ , it suffices to show  $\mathbb{E}[B_n] = \mathbb{E}[|b_n|] = \mathcal{O}(\gamma_n)$  and  $\mathbb{E}[\sigma_n^2] \leq \sigma^2$  for some constant  $\sigma$ .

We next proceed method-by-method:

**Algorithm 1: Stochastic gradient descent/ascent.** For [\(SGDA\)](#), we have  $W_n = U_n = \mathbf{U}(Z_n; \omega_n)$  and  $b_n = 0$ , so [Assumption \(A1\)](#) is satisfied automatically (since  $B_n = 0$ ). Our claim then follows from the stated assumptions for [\(SFO\)](#).

**Algorithm 2: Proximal point method.** For (PPM), we have  $U_n = 0$  and

$$\begin{aligned} \|b_n\| &= \|V(Z_n) - V(Z_{n+1})\| \\ &\leq L\|Z_n - Z_{n+1}\| \\ &= \gamma_n L \|V(Z_n)\| \\ &\leq \gamma_n L M = \mathcal{O}(\gamma_n). \end{aligned}$$

where  $L$  and  $M$  are the Lipschitz constant of  $V$  and  $\Phi$ , respectively.

**Algorithm 3: Stochastic extra-gradient.** For (SEG), we have  $U_n = \mathbf{U}(Z_n^+; \omega_n^+)$  and  $b_n = V(Z_n^+) - V(Z_n)$  so that  $\mathbb{E}[\sigma_n^2] \leq \sigma^2$  by (SFO). To verify Assumption (A1), by the definition of (SEG), we have

$$\begin{aligned} \|b_n\| &= \|V(Z_n^+) - V(Z_n)\| \leq L\|Z_n^+ - Z_n\| \\ &= \gamma_n \|\mathbf{V}(Z_n; \omega_n)\| = \gamma_n L \|V(Z_n) + \mathbf{U}(Z_n; \omega_n)\| \\ &\leq \gamma_n L \|V(Z_n)\| + \gamma_n L \|\mathbf{U}(Z_n; \omega_n)\|. \end{aligned} \quad (\text{B.15})$$

Since  $\Phi$  is assumed to be Lipschitz and  $\mathbf{U}(Z_n; \omega_n)$  finite variance, taking the expectation on both sides of the above shows  $\mathbb{E}[B_n] = \mathcal{O}(\gamma_n)$ . It remains to verify that  $\lim_{n \rightarrow \infty} B_n = 0$  with probability 1. Now, by Chebyshev's inequality and (SFO), we have

$$\mathbb{P}\left(\|\mathbf{U}(Z_n; \omega_n)\| \geq \sqrt{n \log^{1+\frac{\varepsilon}{2}} n}\right) \leq \frac{\sigma^2}{n \log^{1+\frac{\varepsilon}{2}} n} \quad (\text{B.16})$$

where  $\varepsilon$  is the same as in our choice of step-size in Proposition 1. In turn, this implies that

$$\sum_{n=2}^{\infty} \mathbb{P}\left(\|\mathbf{U}(Z_n; \omega_n)\| \geq \sqrt{n \log^{1+\frac{\varepsilon}{2}} n}\right) < \infty$$

so, by the Borel-Cantelli lemma, we have  $\|\mathbf{U}(Z_n; \omega_n)\| = \mathcal{O}\left(\sqrt{n \log^{1+\frac{\varepsilon}{2}} n}\right)$  with probability 1. Hence, by our assumptions for the method's step-size, we get

$$\gamma_n \|\mathbf{U}(Z_n; \omega_n)\| = \mathcal{O}\left(\frac{\sqrt{n \log^{1+\frac{\varepsilon}{2}} n}}{\sqrt{n \log^{1+\varepsilon} n}}\right) = \mathcal{O}\left(\frac{1}{\log^{\frac{\varepsilon}{4}} n}\right) \quad (\text{B.17})$$

so that, in view of (B.15),  $B_n \rightarrow 0$  with probability 1.

**Algorithm 4: Optimistic gradient.** For (OG/PEG), we have  $U_n = \mathbf{U}(Z_n; \omega_n^+)$  and  $b_n = V(Z_n^+) - V(Z_n)$ , so  $\mathbb{E}[\sigma_n^2] = \sigma^2$  again holds by (SFO). The bias term can then be bounded exactly as in the case of Algorithm 3.

**Algorithms 1–4: Alternating RM schemes (alt-RM).** We now show that the alternating version of Algorithms 1–4 still constitute an APT of (MD).

By Lemma 2, we know that the alternating version of an RM scheme is another RM scheme with the same noise and new bias satisfying:

$$\begin{aligned} \|b'_n\| &\leq \|b_n\| + \|\mathbf{V}_y(X_{n+1}, Y_n) - \mathbf{V}_y(X_n, Y_n)\| \\ &\leq \|b_n\| + L\|X_{n+1} - X_n\| \\ &\leq \|b_n\| + \gamma_n L (\|V(Z_n)\|) + \|b_n\| + \|U_n\| \end{aligned} \quad (\text{B.18})$$

by the definition of an RM scheme. Since  $\gamma_n L \|b_n\| = o(\|b_n\|)$ , the rest is the same as Algorithm 3.

We also note that (B.18) can be applied recursively to show that the bias term  $b_n^{(k_1, k_2)}$  of any  $(k_1, k_2)$  version of RM schemes satisfy

$$\|b_n^{(k_1, k_2)}\| \leq (k_1 + k_2 - 1) \left( \|b_n\| + \gamma_n L (\|V(Z_n)\| + \|b_n\| + \|U_n\|) \right), \quad (\text{B.19})$$

thus enjoying the same properties as the vanilla alternating (1, 1)-RM schemes in view of (B.18).

**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

$$b_n = \mathbb{E}[V_n | \mathcal{F}_n] - V(Z_n) \quad (\text{B.20})$$

with

$$V_n = \pm(d/\delta_n) \Phi(Z_n + \delta_n \omega_n) \omega_n \quad (\text{B.21})$$

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

$$v_{i,n} = \mathbb{E}[V_{i,n} | \mathcal{F}_n] \quad (\text{B.22})$$

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

$$v_{i,n} = \pm \frac{d}{\delta_n} \cdot \frac{1}{2d} [\Phi(Z_n + \delta_n e_i) - \Phi(Z_n - \delta_n e_i)] \quad (\text{B.23})$$

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

$$v_{i,n} = \pm \partial_i \Phi(\tilde{E}_n) = V_{i,n}(\tilde{E}_n). \quad (\text{B.24})$$

Since  $V$  is Lipschitz continuous, it follows that

$$|v_{i,n} - V_{i,n}(Z_n)| = |V_{i,n}(\tilde{E}_n) - V_{i,n}(Z_n)| \leq L \|\tilde{E}_n - Z_n\| = \mathcal{O}(\delta_n) \quad (\text{B.25})$$

since  $\tilde{E}_n \in [Z_n - \delta_n e_i, Z_n + \delta_n e_i]$ . 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  $\gamma_n$  and  $\delta_n$ , Assumption (A2) is satisfied and our claim follows from Theorem 1. ■

## APPENDIX C. CONVERGENCE ANALYSIS: PROOF OF THEOREMS 3–4

With all this preliminary work in hand, we are finally in a position to prove Theorems 3–4.

The heavy lifting for Theorem 2 is already 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. Theorems 3 and 4 on the other hand require a completely different set of techniques and involve a much finer analysis of the process in hand.

**C.1. Avoidance of unstable periodic orbits.** While the proof of Theorem 3 is highly technical, the high-level intuition for its conclusion is crystal clear: Assume that we are given an *unstable* critical point  $z^*$ . Then, by the stable manifold theorem [81], the set of all initializations such that the flow of (MD) converges to  $z^*$  is of measure 0 in  $\mathcal{Z}$ . Consequently, if the noise process  $\{U_n\}$  is such that it has “non-negligible” magnitude in the unstable directions near  $z^*$ , then it is plausible that the RM scheme should escape  $z^*$  along these directions. Assumption (ii) in Theorem 3 quantifies exactly the magnitude of noise for which we can formalize this heuristic argument.

Throughout this section we assume that we are given:

- A  $(d - m)$ -dimensional embedded submanifold  $\mathcal{S} \subset \mathbb{R}^d$  where  $1 \leq m \leq d$  and  $d - m$  is to be understood as the dimension of the unstable manifold.
- A nonempty compact set  $\mathcal{K} \subset \mathcal{S}$  invariant under  $\Theta := \{\Theta_t : t \in \mathbb{R}_+\}$ .
- We also assume that  $\mathcal{S}$  is  $C^2$  is locally invariant: there exists a neighborhood  $\mathcal{U}$  of  $\mathcal{K}$  in  $\mathbb{R}^d$  and a positive time  $t_0$  such that

$$\Theta_t(\mathcal{U} \cap \mathcal{S}) \subset \mathcal{S} \quad (\text{C.1})$$

for all  $|t| \leq t_0$ .

We further assume that for every point  $z \in \mathcal{K}$ , we have

$$\mathbb{R}^d = T_z \mathcal{S} \oplus \mathcal{E}_z^u \quad (\text{C.2})$$

where

- (i)  $z \rightarrow \mathcal{E}_z^u$  is a continuous map from  $\mathcal{K}$  into the Grassmanian manifold  $G(m, d)$  of  $m$  planes in  $\mathbb{R}^d$ .
- (ii)  $D\Theta_t(z)\mathcal{E}_z^u = \mathcal{E}_{\Theta_t(z)}^u$  for all  $t \in \mathbb{R}, z \in \mathcal{E}_z^u$ .
- (iii) There exist  $\lambda, C > 0$  such that for all  $z \in \mathcal{K}, w \in \mathcal{E}_z^u$  and  $t \geq 0$

$$\|D\Theta_t(z)w\| \geq Ce^{\lambda t}\|w\|. \quad (\text{C.3})$$

We call any  $\mathcal{K}$  satisfying the above an *unstable invariant set*. As a simple illustration we show:

**Lemma C.1.** *If  $z^*$  is a critical point of  $\Phi$  with any eigenvalue  $\lambda$  of  $JV(z^*)$  such that  $\Re\{\lambda(JV(z^*))\} > 0$ . Then  $z^*$  verifies all the assumptions of an unstable invariant set.*

*As a corollary,  $Z_n$  generated by any of the Algorithms 1–4 in Theorem 3 avoids  $z^*$  almost surely.*

*Proof.* All the requirements for an unstable invariant set are readily verified by the Stable Manifold Theorem [78]. The lemma follows by noting the the dimension for the unstable manifold is greater than or equal to 1; see [78, Chap 5].  $\blacksquare$

A further justification of these technical assumptions is the following: Suppose  $\mathcal{K}$  is a periodic orbit. We then say that  $\mathcal{K}$  is (linearly) unstable if 1 is a Floquet multiplier of  $\mathcal{K}$  and some multipliers have modulus strictly greater than 1 [84]. If the vector field  $V$  is assumed to be  $C^2$ , then a classical result in dynamical systems (see e.g., [81]) states that  $\mathcal{K}$  verifies all the above assumptions.

We now proceed to the proof of Theorem 3. For ease of reading we reformulate its statements in the following more convenient form:

**Theorem 2.** *Let  $\mathcal{K}$  be an unstable invariant set of  $V$ . Assume that*

- (i) *There exists  $K > 0$  such that  $\|U_n\| \leq K$  for all  $n$ .*
- (ii)  *$\gamma_n$  is as in Proposition 1.*
- (iii) *There exists a neighborhood  $\mathcal{U}(\mathcal{K})$  of  $\mathcal{K}$  and  $b > 0$  such that for all unit vector  $v \in \mathbb{R}^d$*

$$\mathbb{E}[\langle U_{n+1}, v \rangle^+ | \mathcal{F}_n] \geq b \mathbf{1}_{\{Z_n \in \mathcal{U}(\mathcal{K})\}}.$$

- (iv) *The vector field  $V$  is  $C^2$ .*

*Then  $Z_n$  generated by any of the Algorithms 1–4 satisfies*

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

proof of [Theorem 3](#). We first note that, for our choice of  $\gamma_n$ ,

$$\lim_{n \rightarrow \infty} \frac{\gamma_n}{\sqrt{\sum_{k=n}^{\infty} \gamma_k^2}} = 0 \quad (\text{C.4})$$

so  $\gamma_n = o\left(\sqrt{\sum_{k=n}^{\infty} \gamma_k^2}\right)$ . This fact will be used in the proof when we invoke [Lemma C.3](#) below with  $\varepsilon_n = \mathcal{O}(\gamma_n)$  and  $\alpha_n = \sum_{k=n}^{\infty} \gamma_k^2$  therein.

**Helper lemmas.** We will need some technical lemmas. The first one is a deep result by Benaïm and Hirsch [\[8\]](#), which asserts the existence of a local Lyapunov function near the unstable periodic orbits.

For a right-differentiable function  $\eta : \mathbb{R}^d \rightarrow \mathbb{R}$  we define its right derivative  $D\eta$  applied to a vector  $h \in \mathbb{R}^d$  by

$$D\eta(z)h = \lim_{t \rightarrow 0^+} \frac{\eta(z+th) - \eta(z)}{t}. \quad (\text{C.5})$$

If  $\eta$  is differentiable, then [\(C.5\)](#) is simply  $\langle \nabla \eta(z), h \rangle$ .

**Lemma C.2.** *There exists a compact neighborhood  $\mathcal{U}(\mathcal{K})$  of  $\mathcal{K}$ , positive numbers  $l, \beta > 0$ , and a map  $\eta : \mathcal{U}(\mathcal{K}) \rightarrow \mathbb{R}$  such that:*

- (i)  $\eta$  is  $C^2$  on  $\mathcal{U}(\mathcal{K}) \setminus \mathcal{S}$ .
- (ii) For all  $z \in \mathcal{U}(\mathcal{K}) \cap \mathcal{S}$ ,  $\eta$  admits a right derivative  $D\eta(z) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  which is Lipschitz, convex and positively homogeneous.
- (iii) There exists  $k > 0$  and a neighborhood  $\mathcal{W} \subset \mathbb{R}^d$  of 0 such that for all  $z \in \mathcal{U}(\mathcal{K})$  and  $v \in \mathcal{W}$

$$\eta(z+v) \geq \eta(z) + D\eta(z)v - k\|v\|^2. \quad (\text{C.6})$$

- (iv) There exists  $c_1 > 0$  such that for all  $z \in \mathcal{U}(\mathcal{K}) \setminus \mathcal{S}$

$$\|\nabla \eta(z)\| \geq c_1 \quad (\text{C.7})$$

and for all  $z \in \mathcal{U}(\mathcal{K}) \cap \mathcal{S}$  and  $v \in \mathbb{R}^d$

$$\langle D\eta(z), v \rangle \geq c_1\|v - D\Pi(z)v\|. \quad (\text{C.8})$$

- (v) For all  $z \in \mathcal{U}(\mathcal{K}) \cap \mathcal{S}$ ,  $u \in T_z\mathcal{S}$  and  $v \in \mathbb{R}^d$

$$D\eta(z)(u+v) = D\eta(z)v. \quad (\text{C.9})$$

- (vi) For all  $z \in \mathcal{U}(\mathcal{K})$  we have

$$D\eta(z)V(z) \geq \beta\eta(z). \quad (\text{C.10})$$

The second lemma we need is a probabilistic estimate from [\[71\]](#).

**Lemma C.3.** *Let  $S_n$  be a nonnegative stochastic process,  $S_n = S_0 + \sum_{k=1}^n X_k$  where  $X_n$  is  $\mathcal{F}_n$ -measurable. Let  $\alpha_n := \sum_{k=n}^{\infty} \gamma_k^2$ .*

*Assume there exist a sequence  $0 \leq \varepsilon_n = o(\sqrt{\alpha_n})$ , constants  $a_1, a_2 > 0$  and an integer  $N_0$  such that for all  $n \geq N_0$ ,*

- (i)  $|X_n| = o(\sqrt{\alpha_n})$ .
- (ii)  $\mathbb{1}_{\{S_n > \varepsilon_n\}} \mathbb{E}[X_{n+1} | \mathcal{F}_n] \geq 0$ .
- (iii)  $\mathbb{E}[S_{n+1}^2 - S_n^2 | \mathcal{F}_n] \geq a_1 \gamma_n^2$ .
- (iv)  $\mathbb{E}[X_{n+1}^2 | \mathcal{F}_n] \leq a_2 \gamma_n^2$ .

*Then  $\mathbb{P}(\lim_{n \rightarrow \infty} S_n = 0) = 0$ .*

Armed with [Lemmas C.2](#) and [C.3](#) we are now ready to prove [Theorem 3](#).

Let  $N \in \mathbb{N}$ . Assume  $z_N \in \mathcal{U}(\mathcal{K})$  where  $\mathcal{U}(\mathcal{K})$  is the neighborhood given by [Lemma C.2](#). Define  $T$  as

$$T := \inf\{k \geq N : Z_n \notin \mathcal{U}(\mathcal{K})\}. \quad (\text{C.11})$$

Evidently,  $T$  is a stopping time adaptive to  $\mathcal{F}_n$ . Thus, proving [Theorem 3](#) amounts to showing  $\mathbb{P}(T < \infty) = 1$ . Without loss of generality we may assume  $N = 0$ .

Define two sequences of random variables  $\{X_n\}_{n \geq 2}$  and  $\{S_n\}$  as

$$X_{n+1} = (\eta(Z_{n+1}) - \eta(Z_n)) \mathbf{1}_{\{n \leq T\}} + \gamma_n \mathbf{1}_{\{n > T\}}, \quad (\text{C.12a})$$

$$S_0 = \eta(Z_0), \quad S_n = S_0 + \sum_{k=2}^n X_k. \quad (\text{C.12b})$$

Note that  $S_n \geq 0$  (a.s.) for every  $n$ . Our proof will revolve around verifying [Lemma C.3\(i\)–\(iv\)](#).

**Verifying [Lemma C.3\(i\)](#) and [\(iv\)](#)**. By Lipschitz continuity of  $\eta$  we know that

$$\begin{aligned} \|\eta(Z_n) - \eta(Z_{n+1})\| &\leq L' \|Z_n - Z_{n+1}\| \\ &= \gamma_n \|V(Z_n) + U_n + b_n\| \end{aligned} \quad (\text{C.13})$$

where  $L'$  is the Lipschitz constant of  $\eta$ . We have seen in the proof of [Proposition 1](#) that  $\|b_n\| = \mathcal{O}(\gamma_n(\|V(Z_n)\| + \|U_n\|))$ . By [Proposition A.1](#) and [Assumption \(i\)](#) in [Theorem 3](#), we then have  $|X_{n+1}| = \mathcal{O}(\gamma_n) = o(\sqrt{\alpha_n})$  which implies both [Lemma C.3\(i\)](#) and [\(iv\)](#).

**Verifying [Lemma C.3\(ii\)](#)**. Let  $k' = k\|V\| + K$  where  $k$  is given by [Lemma C.2\(iii\)](#) and  $\|V\| := \sup\{V(z) : z \in \mathcal{U}(\mathcal{K})\}$  and  $K$  is the uniform bound of  $U_n$ . If  $n \leq T$ , using [Lemma C.2\(ii\)](#), [\(iii\)](#), [\(v\)](#) and [\(vi\)](#) we have

$$\begin{aligned} \eta(Z_{n+1}) - \eta(Z_n) &\geq \gamma_n \text{D}\eta(Z_n)(U_n + b_n + V(Z_n)) - k\gamma_n^2(\|V\| + \|U_n\| + \|b_n\|)^2 \\ &\geq \gamma_n \beta \eta(Z_n) + \gamma_n \text{D}\eta(Z_n)U_n + \gamma_n \text{D}\eta(Z_n)b_n - 2k'\gamma_n^2 - 2k\gamma_n^2\|b_n\|^2. \end{aligned} \quad (\text{C.14})$$

By the same calculation leading up to [\(B.15\)](#), [Assumption \(i\)](#) in [Theorem 3](#), and the Lipschitz continuity of  $\Phi$ , there exists a constant  $c' > 0$  such that the bias sequence for [Algorithms 1–4](#) can be bounded as  $-\|b_n\| \geq -c'\gamma_n$  (a.s.). Combining this with the Lipschitz continuity of  $\eta$ , we can merge the last three terms in [\(C.14\)](#) as

$$\eta(Z_{n+1}) - \eta(Z_n) \geq \gamma_n \beta \eta(Z_n) + \gamma_n \text{D}\eta(Z_n)U_n - 2k''\gamma_n^2 \quad (\text{C.15})$$

for some constant  $k'' > 0$ . Thus

$$\mathbf{1}_{\{n \leq T\}} \mathbb{E}[X_{n+1} | \mathcal{F}_n] \geq \mathbf{1}_{\{n \leq T\}} [\gamma_n \beta \eta(Z_n) - 2k''\gamma_n^2 + \gamma_n \mathbb{E}[\text{D}\eta(Z_n)U_n | \mathcal{F}_n]]. \quad (\text{C.16})$$

By [Lemma C.2\(ii\)](#) again, we have

$$\mathbb{E}[\text{D}\eta(Z_n)U_n | \mathcal{F}_n] \geq \text{D}\eta(Z_n) \mathbb{E}[U_n | \mathcal{F}_n] = 0 \quad (\text{C.17})$$

since we have assumed noise to be zero mean. Combining [\(C.16\)](#) and [\(C.17\)](#), we then get

$$\mathbf{1}_{\{n \leq T\}} \mathbb{E}[X_{n+1} | \mathcal{F}_n] \geq \mathbf{1}_{\{n \leq T\}} [\gamma_n \beta \eta(Z_n) - 2k''\gamma_n^2]. \quad (\text{C.18})$$

If  $n > T$ ,  $X_{n+1} = \gamma_n$  so trivially

$$\mathbf{1}_{\{n \leq T\}} \mathbb{E}[X_{n+1} | \mathcal{F}_n] \geq 0. \quad (\text{C.19})$$

Combining [\(C.18\)](#) with [\(C.19\)](#), we see that [Lemma C.3\(ii\)](#) is satisfied with  $\varepsilon_n = \frac{k''}{\beta} \gamma_n$ .

**Verifying Lemma C.3(iii)** . We begin by observing that

$$\mathbb{E}[S_{n+1}^2 - S_n^2 | \mathcal{F}_n] = \mathbb{E}[X_{n+1}^2 | \mathcal{F}_n] + 2S_n \mathbb{E}[X_{n+1} | \mathcal{F}_n]. \quad (\text{C.20})$$

If  $S_n \geq \varepsilon_n$ , then the right-hand side of (C.20) is non-negative by Lemma C.3(ii) that we just verified above. If  $S_n < \varepsilon_n$ , (C.18) with (C.19) imply that  $S_n \mathbb{E}[X_{n+1} | \mathcal{F}_n] \geq -\varepsilon_n k'' \gamma_n^2 = -\mathcal{O}(\gamma_n^3)$ . In other words, (C.20) can be rewritten as

$$\mathbb{E}[S_{n+1}^2 - S_n^2 | \mathcal{F}_n] \geq \mathbb{E}[X_{n+1}^2 | \mathcal{F}_n] - \mathcal{O}(\gamma_n^3). \quad (\text{C.21})$$

Below, we shall prove that  $\mathbb{E}[X_{n+1}^2 | \mathcal{F}_n] \geq b_1 \gamma_n^2$  for some  $b_1 > 0$  and  $n$  large enough. Combining this with (C.21) proves Lemma C.3(iii).

From (C.15), we deduce

$$\mathbb{1}_{\{n \leq T\}} [\mathbb{E}[X_{n+1}^+ | \mathcal{F}_n] - (\gamma_n \mathbb{E}[(D\eta(Z_n)U_n)^+ | \mathcal{F}_n] - k'' \gamma_n^2)] \geq 0. \quad (\text{C.22})$$

Invoking Lemma C.2(iv) and Assumption (ii) in Theorem 3, we see that

$$\mathbb{1}_{\{n \leq T\} \cap \{Z_n \notin \mathcal{S}\}} \left( \mathbb{E}[(D\eta(Z_n)U_n)^+ | \mathcal{F}_n] - c_1 b \right) \geq 0 \quad (\text{C.23})$$

If  $Z_n \in \mathcal{S}$ , we can choose a unit vector  $v_n \in \ker(I - D\Pi(Z_n))^\perp$  where  $\Pi$  denotes the projection operator onto  $\mathcal{S}$ . By the definition of  $v_n$ , we have  $\langle U_n, v_n \rangle = \langle U_n - D\Pi(Z_n)U_n, v_n \rangle$ . Let  $\mathcal{H} = \{n \leq T\} \cap \{Z_n \notin \mathcal{S}\}$ . By Lemma C.2(iv), Cauchy-Schwartz, and Assumption (ii) of Theorem 3 we get

$$\begin{aligned} \mathbb{1}_{\mathcal{H}} \mathbb{E}[(D\eta(Z_n)U_n)^+ | \mathcal{F}_n] &\geq c_1 \mathbb{1}_{\mathcal{H}} \mathbb{E}[\|U_n - D\Pi(Z_n)U_n\| | \mathcal{F}_n] \\ &\geq c_1 \mathbb{1}_{\mathcal{H}} \mathbb{E}[\langle U_n - D\Pi(Z_n)U_n, v_n \rangle^+ | \mathcal{F}_n] \\ &= c_1 \mathbb{1}_{\mathcal{H}} \mathbb{E}[\langle U_n, v_n \rangle^+ | \mathcal{F}_n] \\ &\geq c_1 b \mathbb{1}_{\mathcal{H}}. \end{aligned} \quad (\text{C.24})$$

Combining Eqs. (C.19) and (C.22)–(C.24) then gives

$$\mathbb{E}[X_{n+1}^+ | \mathcal{F}_n] \geq \gamma_n c_1 b - k'' \gamma_n^2. \quad (\text{C.25})$$

On the other hand, we always have  $\mathbb{E}[X_{n+1}^2 | \mathcal{F}_n] \geq \mathbb{E}[X_{n+1}^+ | \mathcal{F}_n]$  by Jensen. It then follows that  $\mathbb{E}[X_{n+1}^2 | \mathcal{F}_n] \geq b_1 \gamma_n^2$  for some  $b_1 > 0$  and large enough  $n$  as desired.

**Closing the gap.** We have now verified Lemma C.3(i)–(iv). Thus, Lemma C.3 concludes that

$$\mathbb{P}(\lim_{n \rightarrow \infty} S_n = 0) = 0. \quad (\text{C.26})$$

We will use (C.26) to show that  $T < \infty$  (a.s.).

Suppose  $T = \infty$ . Then  $X_{n+1} = \eta(Z_{n+1}) - \eta(Z_n)$  and  $S_n = \eta(Z_n)$  by (C.12a)–(C.12b), and  $\{Z_n\}$  remains in  $\mathcal{U}(\mathcal{K})$  by definition of the stopping time  $T$ . Theorem 2 then asserts that the limit set  $L(\{Z_n\})$  of  $\{Z_n\}$  is a nonempty compact invariant subset of  $\mathcal{U}(\mathcal{K})$ , so that for all  $z' \in L(\{Z_n\})$  and  $t \in \mathbb{R}$ ,  $\Theta_t(z') \in \mathcal{U}(\mathcal{K})$ . But then Lemma C.2(iv) implies that  $\eta(\Theta_t(z')) \geq e^{\beta t} \eta(z')$  for all  $t > 0$ , forcing  $\eta(z')$  to be zero. In other words, we have  $L(\{Z_n\}) \subset \mathcal{S}$ , which implies  $S_n = \eta(Z_n) \rightarrow 0$ . By (C.26), this event occurs with probability 0, thus showing that  $T$  is finite almost surely. ■

**C.2. Convergence to ICTs.** We now prove Theorem 2, which we restate below for convenience:

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

*Proof.* By [Theorem 1](#),  $Z_n$  generates APT of the mean dynamics [\(MD\)](#). Now, let  $\mathcal{L} = \bigcap_{t \geq 0} \text{cl}(Z(t, \infty))$  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 [[9](#), Theorem 8.2]. ■

As we discussed in the main part of our paper, the ICT sets of  $\Phi$  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  $f: \mathcal{Z} \rightarrow \mathbb{R}$ ), ICT sets can only be components of equilibria. Specifically, building on a general result by Benaïm [[7](#)], we have:

**Proposition C.1.** *Suppose that  $V(z) = -\nabla f(z)$  for some  $C^d$ -smooth potential function  $f: \mathcal{Z} \rightarrow \mathbb{R}$  with a compact critical set  $\text{crit}(f) = \{z^* : \nabla f(z^*) = 0\}$ . Then, every ICT set  $\mathcal{S}$  of [\(MD\)](#) is contained in  $\text{crit}(f)$ ; 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  $\mathcal{Z}^* := \text{crit}(f)$  of  $f$  coincides with the set of rest points of [\(MD\)](#). Moreover, by Sard’s theorem [[51](#)],  $f(\mathcal{Z}^*)$  has zero Lebesgue measure and hence empty interior. Our claim then follows from Proposition 6.4 of Benaïm [[7](#)]. ■

As another elementary illustration in addition to the gradient systems, one can show that for bilinear games  $\Phi(x, y) = xy$ , the ICT sets the whole space  $\mathbb{R}^2$ . 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.

**C.3. Convergence to attractors.** We now proceed with the analysis of RM schemes in the presence of an attractor; the relevant result is [Theorem 4](#):

**Theorem 4.** *Let  $\mathcal{S}$  be an attractor of [\(MD\)](#) and fix some confidence level  $\alpha > 0$ . If  $\gamma_n$  is small enough and [Assumptions \(A1\)–\(A2\)](#) hold, there exists a neighborhood  $\mathcal{U}$  of  $\mathcal{S}$ , independent of  $\alpha$ , such that  $\mathbb{P}(Z_n \text{ converges to } \mathcal{S}) \geq 1 - \alpha$  if  $Z_1 \in \mathcal{U}$ .*

Because of the generality of our assumptions, the proof of [Theorem 4](#) 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 *significant* 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  $\varepsilon$ ; 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 energy function for (MD). Indeed, since  $\mathcal{S}$  is an attractor of (MD), there exists a compact neighborhood  $\mathcal{K}$  of  $\mathcal{S}$ , called the *fundamental neighborhood* of  $\mathcal{S}$ , with the property that  $\text{dist}(\Theta_t(z)tz, \mathcal{S}) \rightarrow 0$  as  $t \rightarrow \infty$  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  $\mathcal{S}$ . Hence, with  $\mathcal{K}$  compact, Conley's decomposition theorem [19] shows that there exists a strongly smooth Lyapunov – or “energy” – function  $E: \mathcal{K} \rightarrow \mathbb{R}$  such that (i)  $E(z) \geq 0$  with equality if and only if  $z \in \mathcal{S}$ ; and (ii)  $\dot{E}(z) := \langle \nabla E(z), V(z) \rangle < 0$  for all  $z \in \mathcal{K} \setminus \mathcal{S}$  (implying in particular that  $E(\Theta_t(z)tz)$  is strictly decreasing in  $t$  whenever  $z \in \mathcal{K} \setminus \mathcal{S}$ ).

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

$$E_{n+1} \leq E_n + \gamma_n \langle \nabla E(Z_n), V(Z_n) \rangle + \gamma_n \xi_n + \gamma_n \psi_n + \gamma_n^2 \theta_n^2, \quad (\text{C.27})$$

where the error terms  $\xi_n$ ,  $\psi_n$  and  $\theta_n$  are defined as

$$\xi_n = \langle \nabla E(Z_n), U_n \rangle \quad (\text{C.28a})$$

$$\psi_n = B_n \|\nabla E(Z_n)\| + \gamma_n \beta B_n^2 \quad (\text{C.28b})$$

$$\theta_n^2 = \beta \|V(Z_n) + U_n\|^2 \quad (\text{C.28c})$$

with  $\beta$  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 (C.27) as

$$M_n = \sum_{k=1}^n \gamma_k \xi_k \quad (\text{C.29a})$$

and

$$S_n = \sum_{k=1}^n [\gamma_k \psi_k + \gamma_k^2 \theta_k^2] \quad (\text{C.29b})$$

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  $\xi_n$  is zero-mean), it is more difficult to control because the variance of its increments is

$$\mathbb{E}[|\gamma_n \xi_n|^2 | \mathcal{F}_n] = \gamma_n^2 \mathbb{E}[|\langle \nabla E(Z_n), U_n \rangle|^2 | \mathcal{F}_n], \quad (\text{C.30})$$

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

$$R_n = M_n^2 + S_n \quad (\text{C.31})$$

and the indicator events

$$\mathcal{E}_n \equiv \mathcal{E}_n(\mathcal{K}) = \{Z_k \in \mathcal{K} \text{ for all } k = 1, 2, \dots, n\} \quad (\text{C.32})$$

$$\mathcal{H}_n \equiv \mathcal{H}_n(\varepsilon) = \{R_k \leq \varepsilon \text{ for all } k = 1, 2, \dots, n\}, \quad (\text{C.33})$$

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

$$\{z \in \mathcal{Z} : E(z) \leq 2\varepsilon + \sqrt{\varepsilon}\} \subseteq \mathcal{K} \quad (\text{C.34})$$

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

$$\mathcal{U} \subseteq \{z \in \mathcal{Z} : E(z) \leq \varepsilon\}. \quad (\text{C.35})$$

We then have the following estimates:

**Lemma C.4.** *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

$$\begin{aligned} \tilde{\mathcal{H}}_n &:= \mathcal{H}_{n-1} \setminus \mathcal{H}_n = \mathcal{H}_{n-1} \cap \{R_n > \varepsilon\} \\ &= \{R_k \leq \varepsilon \text{ for } k = 1, 2, \dots, n-1 \text{ and } R_n > \varepsilon\}, \end{aligned} \quad (\text{C.36})$$

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

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}[\tilde{R}_{n-1} + \gamma_n G B_n + \gamma_n^2 [2\beta G^2 + (2\beta + G^2)\sigma_n^2 + \beta B_n^2]] - \varepsilon \mathbb{P}(\tilde{\mathcal{H}}_{n-1}), \quad (\text{C.37})$$

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

*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 (C.27) gives

$$E_{k+1} \leq E_k + \gamma_n \xi_n + \gamma_n \psi_n + \gamma_n^2 \theta_n^2, \quad \text{for all } k = 1, 2, \dots, n, \quad (\text{C.38})$$

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

$$E_{n+1} \leq E_1 + M_n + S_n \leq E_1 + \sqrt{R_n} + R_n \leq \varepsilon + \sqrt{\varepsilon} + \varepsilon = 2\varepsilon + \sqrt{\varepsilon}. \quad (\text{C.39})$$

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

$$\begin{aligned} R_n &= (M_{n-1} + \gamma_n \xi_n)^2 + S_{n-1} + \gamma_n \psi_n + \gamma_n^2 \theta_n^2 \\ &= R_{n-1} + 2\gamma_n \xi_n M_{n-1} + \gamma_n^2 \xi_n^2 + \gamma_n \psi_n + \gamma_n^2 \theta_n^2, \end{aligned} \quad (\text{C.40})$$

so, after taking expectations:

$$\mathbb{E}[R_n | \mathcal{F}_n] = R_{n-1} + 2M_{n-1}\gamma_n \mathbb{E}[\xi_n | \mathcal{F}_n] + \mathbb{E}[\gamma_n^2 \xi_n^2 + \gamma_n \psi_n + \gamma_n^2 \theta_n^2 | \mathcal{F}_n] \geq R_{n-1} \quad (\text{C.41})$$

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

$$\begin{aligned}\tilde{R}_n &= R_{n-1} \mathbb{1}_{\mathcal{H}_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}} \\ &= R_{n-1} \mathbb{1}_{\mathcal{H}_{n-2}} - R_{n-1} \mathbb{1}_{\tilde{\mathcal{H}}_{n-1}} + (R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}}, \\ &= \tilde{R}_{n-1} + (R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}} - R_{n-1} \mathbb{1}_{\tilde{\mathcal{H}}_{n-1}},\end{aligned}\tag{C.42}$$

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, (C.40) yields

$$R_n - R_{n-1} = 2M_{n-1}\gamma_n\xi_n + \gamma_n^2\xi_n^2 + \gamma_n\psi_n + \gamma_n^2\theta_n^2\tag{C.43}$$

so

$$\mathbb{E}[(R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}}] = 2\mathbb{E}[\gamma_n M_{n-1} \xi_n \mathbb{1}_{\mathcal{H}_{n-1}}]\tag{C.44a}$$

$$+ \mathbb{E}[\gamma_n^2 \xi_n^2 \mathbb{1}_{\mathcal{H}_{n-1}}]\tag{C.44b}$$

$$+ \mathbb{E}[(\gamma_n \psi_n + \gamma_n^2 \theta_n^2) \mathbb{1}_{\mathcal{H}_{n-1}}].\tag{C.44c}$$

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 (C.44a), we have:

$$\mathbb{E}[M_{n-1} \xi_n \mathbb{1}_{\mathcal{H}_{n-1}}] = \mathbb{E}[M_{n-1} \mathbb{1}_{\mathcal{H}_{n-1}} \mathbb{E}[\xi_n | \mathcal{F}_n]] = 0.\tag{C.45}$$

- (2) The term (C.44b) is where the reduction to  $\mathcal{H}_{n-1}$  kicks in; indeed:

$$\begin{aligned}\mathbb{E}[\xi_n^2 \mathbb{1}_{\mathcal{H}_{n-1}}] &= \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \mathbb{E}[|\langle \nabla E(Z_n), U_n \rangle|^2 | \mathcal{F}_n]] \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \|\nabla E(Z_n)\|^2 \mathbb{E}[|U_n|^2 | \mathcal{F}_n]] && \{\text{by Cauchy-Schwarz}\} \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{E}_n} \|\nabla E(Z_n)\|^2 \mathbb{E}[|U_n|^2 | \mathcal{F}_n]] && \{\text{because } \mathcal{H}_{n-1} \subseteq \mathcal{E}_n\} \\ &\leq G^2 \sigma_n^2, && \{\text{by Eq. (6b)}\}\end{aligned}$$

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

- (3) Finally, for the term (C.44c), we have:

$$\mathbb{E}[\theta_n^2 \mathbb{1}_{\mathcal{H}_{n-1}}] \leq 2\beta \mathbb{E}[\|V(Z_n)\|^2 \mathbb{1}_{\mathcal{E}_n} + |U_n|^2] \leq 2\beta(G^2 + \sigma_n^2),\tag{C.46}$$

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

$$\mathbb{E}[\psi_n \mathbb{1}_{\mathcal{H}_{n-1}}] \leq \mathbb{E}[GB_n + \gamma_n \beta B_n^2].\tag{C.47}$$

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

$$\mathbb{E}[(R_n - R_{n-1}) \mathbb{1}_{\mathcal{H}_{n-1}}] \leq \mathbb{E}[\gamma_n GB_n + \gamma_n^2 [2\beta G^2 + (2\beta + G^2)\sigma_n^2 + \beta B_n^2]].\tag{C.48}$$

Going back to (C.42), we have  $R_{n-1} > \varepsilon$  if  $\tilde{\mathcal{H}}_{n-1}$  occurs, so the last term becomes

$$\mathbb{E}[R_{n-1} \mathbb{1}_{\tilde{\mathcal{H}}_{n-1}}] \geq \varepsilon \mathbb{E}[\mathbb{1}_{\tilde{\mathcal{H}}_{n-1}}] = \varepsilon \mathbb{P}(\tilde{\mathcal{H}}_{n-1}).\tag{C.49}$$

Our claim then follows by combining Eqs. (C.42), (C.46), (C.47) and (C.49).  $\blacksquare$

**Containment probability.** Lemma C.4 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 C.2.** *Fix some confidence threshold  $\alpha > 0$ . If (RM) is run with sufficiently small  $\gamma_n$  satisfying the conditions of Proposition 1, then*

$$\mathbb{P}(\mathcal{H}_n | Z_1 \in \mathcal{U}) \geq 1 - \alpha \quad \text{for all } n = 1, 2, \dots\tag{C.50}$$

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:

$$\begin{aligned} \mathbb{P}(\tilde{\mathcal{H}}_n) &= \mathbb{P}(\mathcal{H}_{n-1} \setminus \mathcal{H}_n) = \mathbb{P}(\mathcal{H}_{n-1} \cap \{R_n > \varepsilon\}) \\ &= \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \times \mathbb{1}_{\{R_n > \varepsilon\}}] \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{H}_{n-1}} \times (R_n/\varepsilon)] \\ &= \mathbb{E}[\tilde{R}_n]/\varepsilon \end{aligned} \tag{C.51}$$

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 (C.37) yields

$$\mathbb{E}[\tilde{R}_n] \leq \mathbb{E}\left[\tilde{R}_0 + G \sum_{k=1}^n \gamma_k B_k + \sum_{k=1}^n \gamma_k^2 \varrho_k^2\right] - \varepsilon \sum_{k=1}^n \mathbb{P}(\tilde{\mathcal{H}}_{k-1}) \tag{C.52}$$

where we set  $\varrho_n^2 = 2\beta G^2 + (2\beta + G^2)\sigma_n^2 + \beta B_n^2$ . Hence, combining (C.51) and (C.52) and invoking Assumptions (A1) and (A2), we get  $\sum_{k=1}^n \mathbb{P}(\tilde{\mathcal{H}}_k) \leq \frac{1}{\varepsilon} \mathbb{E}[\sum_{k=1}^n [\gamma_k G B_k + \gamma_k^2 \varrho_k^2]] \leq \Gamma/\varepsilon$  for some  $\Gamma > 0$ . Now, by choosing  $\gamma_n$  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

$$\mathbb{P}\left(\bigcup_{k=1}^n \tilde{\mathcal{H}}_k\right) = \sum_{k=1}^n \mathbb{P}(\tilde{\mathcal{H}}_k) \leq \alpha \tag{C.53}$$

and hence:

$$\mathbb{P}(\mathcal{H}_n) = \mathbb{P}\left(\bigcap_{k=1}^n \tilde{\mathcal{H}}_k^c\right) \geq 1 - \alpha, \tag{C.54}$$

as claimed.  $\blacksquare$

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

*Proof of Theorem 4.* By Proposition C.2, if  $Z_n$  is initialized within the neighborhood  $\mathcal{U}$  defined in (C.35), we have  $\mathbb{P}(Z_n \in \mathcal{K} \mid Z_1 \in \mathcal{U}) \geq 1 - \alpha$  (note also that the neighborhood  $\mathcal{U}$  is independent of the required confidence level  $\alpha$ ). Since  $\mathcal{K}$  is compact, if  $Z_n \in \mathcal{K}$  for all  $n$ , we conclude by Theorem 1 that the continuous-time interpolation  $Z(t)$  of  $Z_n$  is an APT of (MD).

Now, if we write  $\mathcal{L} = \bigcap_{t \geq 0} \text{cl}(Z(t, \infty))$  for the limit set of  $Z(t)$ , we will have  $\mathcal{K} \cap \mathcal{L} \neq \emptyset$  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  $\{\Theta_t(z)tz : 0 \leq t \leq T, z \in \mathcal{K}\}$ . Since points in  $\mathcal{L} \cap \mathcal{K}$  are attracted to  $\mathcal{S}$  under (MD) and  $\mathcal{L}$  is invariant under (MD), we conclude that  $\mathcal{L} \cap \mathcal{S} \neq \emptyset$ . 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 hence  $Z_n$  – converges to  $\mathcal{S}$ , and our proof is complete.  $\blacksquare$

#### APPENDIX D. OMITTED DETAILS FOR SECTION 5

**D.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 (10); cf. Lemma D.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 D.1.** *Let  $\phi(y) = \sum_k a_k y^k$  be an analytic function such that*

$$\sum_k a_{2k} k h^{2k} \prod_{i=1}^k \frac{2i-1}{2i} = 0 \quad (\text{D.1})$$

*has a solution with  $h > 0$ . Then, for small enough  $\varepsilon$ , 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:

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

The most important tool of the proof is the *Abelian integral* [18]:

$$I(h) := - \oint_{\gamma_h} \phi' dx \quad (\text{AI})$$

where  $h > 0$  is a parameter and  $\gamma_h$  is a family of ovals defined as in (2.3) of [18].

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

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

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 (D.1) holds. By Theorem 2.4 in [18], the solution  $h^*$  of  $I(h^*) = 0$  then implies the existence of a limit cycle in a neighborhood of the oval  $\gamma_{h^*} := \{z : \|z\| = h^*\}$ .  $\blacksquare$

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

**D.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 (11) reads:

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

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

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

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

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

since  $|x| \leq \sqrt{\frac{4}{3}}$  on  $\{r \geq 0 : r^2 = \frac{4}{3}\}$ , whence  $\dot{r} > 0$  on  $\{r \geq 0 : r^2 = \frac{4}{3}\}$ . Likewise, one can check that  $\dot{r} < 0$  on  $\{r \geq 0 : r^2 = 2\}$ , and that there is no stationary point in the region  $\mathcal{S} := \{r \geq 0 : \frac{4}{3} \leq r^2 \leq 2\}$ . By the Poincaré-Bendixson theorem [85], there exists at least a limit cycle in  $\mathcal{S}$ .

Finally, it is easy to see that  $(x^*, y^*) \simeq (0, 0.49)$  is a stable critical point of (11). 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.

**D.3. Second-order methods.** 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.

▼ **Example D.1** (Second-order methods). Thanks to the efficient implementation of Hessian-gradient multiplications [69], 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

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

As a (discretized) gradient system, our theory in Section 4 shows that (HD) does not possess ICT sets other than critical points of  $f$ . However, a serious issue of (HD) is that it ignores the *sign* of gradients, i.e., it does not distinguish between minimization and maximization. As such, 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) [61]. As in Section 3, one can cast these algorithms as RM schemes with  $V(Z_n)$  replaced by  $(I - \lambda JV(Z_n))V(Z_n)$ , where  $\lambda$  is the regularization parameter. The analysis can then proceed as in Section 4 by replacing (MD) with the appropriate continuous-time systems.

Fig. 4(a) shows the spurious convergence of SGA with  $\lambda = 0.2$  applied to (11). 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). ▲

We now discuss how to model second-order methods as RM schemes. We will showcase on the *consensus optimization* (ConO):

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

where  $\lambda > 0$  is the regularization parameter. Recalling the efficient implementation scheme of Hessian-gradient multiplication [69], 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  $JV(z; \omega')$  of the form

$$JV(z; \omega') = JV(z)V(z) + U'(z; \omega') \quad (\text{SSO})$$

where  $U'(z; \omega')$  is assumed to be unbiased and sub-Gaussian as in (3). With these assumptions, one can then proceed exactly as in Appendix B.3 for the Algorithms 1–4 cases to show that ConO, and its alternating version, give rise to asymptotic pseudotrajectories of the continuous-time dynamics:

$$\dot{z}(t) = \left( I - \lambda JV(z(t)) \right) V(z(t)).$$

Fig. 4(b) demonstrates that the spurious ICT sets of ConO for (11) is similar to that of SGA.

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

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

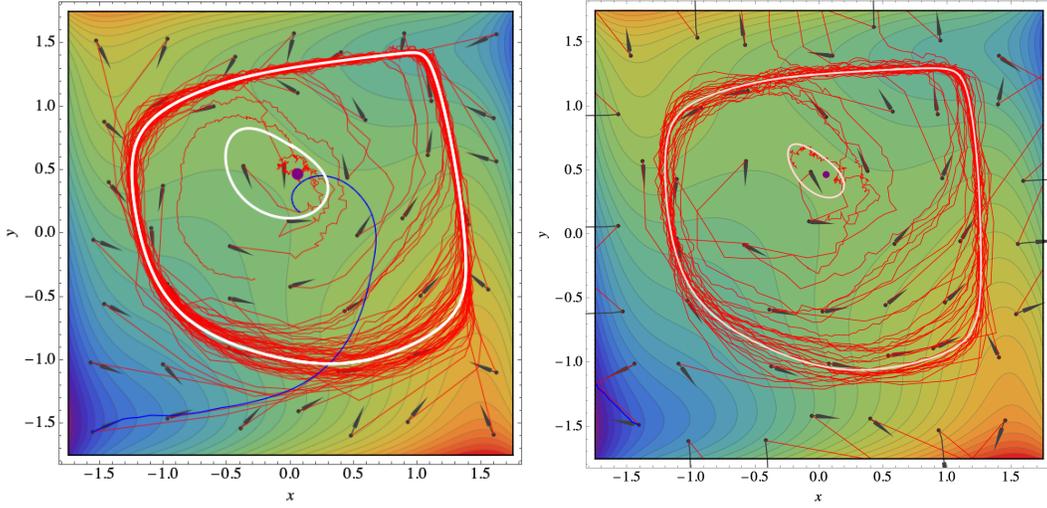
As explained in Example D.1, it is undesirable to set a large number of  $\lambda$ , since then we are essentially treating min max and max min as the same problem. However, if  $\lambda$  is small, then the structure stability of *hyperbolic* orbits (which holds for any stable/unstable ICT sets) implies that any stable (unstable) ICT set of (MD) remains stable (unstable) under perturbations [85]. We therefore expect the ICT sets of various second-order algorithms in Example D.1 be to similar to that of first-order RM schemes.

In addition, we have included yet another second-order method, the *Competitive Gradient Descent* (CGD) [80], 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.

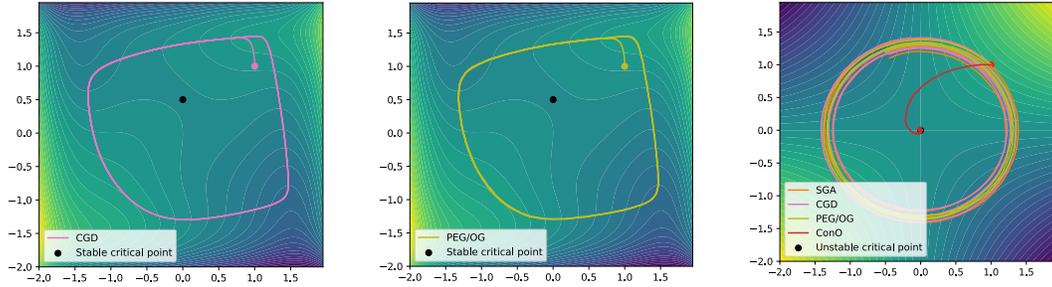
Finally, we report the behavior of various algorithms applied to the “almost bilinear game” (10) in Fig. 5(c). In this case, all algorithms fail to escape the spurious ICT set, with the sole exception of ConO. Intriguingly, ConO converges to the *unstable* critical point. A plausible explanation of this phenomenon is provided by [1], where it is shown that the Hamiltonian descent (HD) 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.

**4.4. Constant step-sizes.** We report in Fig. 6 the behaviors of constant step-size RM schemes. In accord with our intuition, these schemes exhibit concentration behaviors around the attractors.

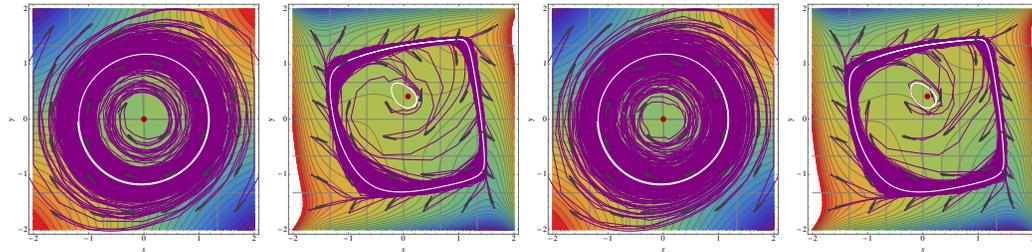
**4.5. Adaptive methods.** We report in Fig. 7 the behaviors of popular *adaptive algorithms* for min-max optimization, including Adam [46] and its extra-gradient variant [30], both set to default hyperparameter values in PyTorch. The result reveals a potentially dangerous trend: while both Adam and ExtraAdam are able to somewhat mitigate cycling phenomena, this comes at the cost of converging to the *max-min* point  $(0, 0)$  of (10). In other words, the algorithm has converged, but to a very bad solution point – an observation which, in the terminology of Letcher [52], would mean that Adam is not a “reasonable” algorithm.



**Figure 4:** Spurious limits of second-order algorithms. From left to right: (a) SGA with  $\lambda = 0.2$  applied to (11); (b) ConO with  $\lambda = 0.2$  applied to (11).

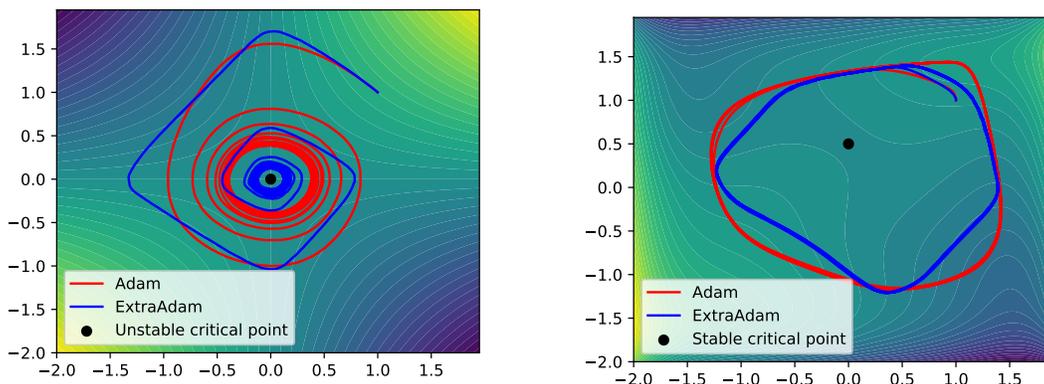


**Figure 5:** Spurious limits of min-max optimization algorithms from the same initialization. From left to right: (a) CGD for (11); (b) (OG/PEG) for (11); (c) Algorithms for (10).



**Figure 6:** RM schemes with constant step-size  $\gamma_n = 0.01$  under the same initializations. From left to right: (a) (SGDA) for (10) with  $\varepsilon = .1$ ; (b) (SGDA) for (11); (c) (SEG) for (10) with  $\varepsilon = .1$ ; (d) (SEG) for (11).

Moreover, as all RM schemes, both adaptive methods fail to reach the “forsaken” solutions in Example 5.2.



**Figure 7:** Adaptive algorithm. From left to right: (a) Adaptive algorithms for (10); (b) Adaptive algorithms for (11).

## REFERENCES

- [1] Abernethy J, Lai KA, Wibisono A (2019) Last-iterate convergence rates for min-max optimization. arXiv preprint arXiv:190602027
- [2] Adolphs L, Daneshmand H, Lucchi A, Hofmann T (2019) Local saddle point optimization: A curvature exploitation approach. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp 486–495
- [3] Arjovsky M, Chintala S, Bottou L (2017) Wasserstein gan. arXiv preprint arXiv:170107875
- [4] Arrow KJ, Hurwicz L, Uzawa H (1958) Studies in linear and non-linear programming. Stanford University Press
- [5] Balduzzi D, Racaniere S, Martens J, Foerster J, Tuyls K, Graepel T (2018) The mechanics of n-player differentiable games. In: International Conference on Machine Learning, pp 354–363
- [6] Bauschke HH, Combettes PL (2017) Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, New York, NY, USA
- [7] Benaïm M (1999) Dynamics of stochastic approximation algorithms. In: Azéma J, Émery M, Ledoux M, Yor M (eds) Séminaire de Probabilités XXXIII, Lecture Notes in Mathematics, vol 1709, Springer Berlin Heidelberg, pp 1–68
- [8] Benaïm M, Hirsch MW (1995) Dynamics of Morse-Smale urn processes. Ergodic Theory and Dynamical Systems 15(6):1005–1030
- [9] Benaïm M, Hirsch MW (1996) Asymptotic pseudotrajectories and chain recurrent flows, with applications. Journal of Dynamics and Differential Equations 8(1):141–176
- [10] Benveniste A, Métivier M, Priouret P (1990) Adaptive Algorithms and Stochastic Approximations. Springer
- [11] Bertsekas DP, Tsitsiklis JN (2000) Gradient convergence in gradient methods with errors. SIAM Journal on Optimization 10(3):627–642
- [12] Borkar VS (2008) Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press and Hindustan Book Agency
- [13] Bowen R (1975) Omega limit sets of Axiom A diffeomorphisms. Journal of Differential Equations 18:333–339
- [14] Burkholder DL (1973) Distribution function inequalities for martingales. Annals of Probability 1(1):19–42
- [15] Chambolle A, Pock T (2011) A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of mathematical imaging and vision 40(1):120–145
- [16] Chavdarova T, Gidel G, Fleuret F, Lacoste-Julien S (2019) Reducing noise in GAN training with variance reduced extragradient. In: NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems

- [17] Chiang CK, Yang T, Lee CJ, Mahdavi M, Lu CJ, Jin R, Zhu S (2012) Online optimization with gradual variations. In: COLT '12: Proceedings of the 25th Annual Conference on Learning Theory
- [18] Christopher C, Li C (2007) Limit cycles of differential equations. Springer Science & Business Media
- [19] Conley CC (1978) Isolated Invariant Set and the Morse Index. American Mathematical Society, Providence, RI
- [20] Daskalakis C, Panageas I (2018) The limit points of (optimistic) gradient descent in min-max optimization. In: Advances in Neural Information Processing Systems, pp 9236–9246
- [21] Daskalakis C, Ilyas A, Syrgkanis V, Zeng H (2018) Training GANs with optimism. In: ICLR '18: Proceedings of the 2018 International Conference on Learning Representations
- [22] Daskalakis C, Skoulakis S, Zampetakis M (2020) The complexity of constrained min-max optimization. arXiv preprint arXiv:200909623
- [23] Domingo-Enrich C, Jelassi S, Mensch A, Rotskoff G, Bruna J (2020) A mean-field analysis of two-player zero-sum games. arXiv preprint arXiv:200206277
- [24] Facchinei F, Pang JS (2003) Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer Series in Operations Research, Springer
- [25] Fiez T, Ratliff L (2020) Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation. arXiv preprint arXiv:200914820
- [26] Fiez T, Chasnov B, Ratliff LJ (2019) Convergence of learning dynamics in stackelberg games. arXiv preprint arXiv:190601217
- [27] Flokas L, Vlatakis-Gkaragkounis EV, Piliouras G (2019) 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
- [28] Flokas L, Vlatakis-Gkaragkounis EV, Piliouras G (2021) Solving min-max optimization with hidden structure via gradient descent ascent. arXiv preprint arXiv:210105248
- [29] Ge R, Huang F, Jin C, Yuan Y (2015) Escaping from saddle points — Online stochastic gradient for tensor decomposition. In: COLT '15: Proceedings of the 28th Annual Conference on Learning Theory
- [30] Gidel G, Berard H, Vignoud G, Vincent P, Lacoste-Julien S (2019) A variational inequality perspective on generative adversarial networks. In: ICLR '19: Proceedings of the 2019 International Conference on Learning Representations
- [31] Gidel G, Hemmat RA, Pezeshki M, Le Priol R, Huang G, Lacoste-Julien S, Mitliagkas I (2019) Negative momentum for improved game dynamics. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp 1802–1811
- [32] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: NIPS '14: Proceedings of the 28th International Conference on Neural Information Processing Systems
- [33] Grimmer B, Lu H, Worah P, Mirrokni V (2020) The landscape of nonconvex-nonconcave minimax optimization. arXiv preprint arXiv:200608667
- [34] Grimmer B, Lu H, Worah P, Mirrokni V (2020) Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems. arXiv preprint arXiv:201010628
- [35] Hall P, Heyde CC (1980) Martingale Limit Theory and Its Application. Probability and Mathematical Statistics, Academic Press, New York
- [36] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637
- [37] Hsieh YG, Iutzeler F, Malick J, Mertikopoulos P (2019) On the convergence of single-call stochastic extra-gradient methods. In: NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp 6936–6946
- [38] Hsieh YG, Iutzeler F, Malick J, Mertikopoulos P (2020) Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In: NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems
- [39] Hsieh YP, Liu C, Cevher V (2019) Finding mixed nash equilibria of generative adversarial networks. In: International Conference on Machine Learning, pp 2810–2819
- [40] Iusem AN, Jofré A, Oliveira RI, Thompson P (2017) Extragradient method with variance reduction for stochastic variational inequalities. SIAM Journal on Optimization 27(2):686–724

- [41] Jin C, Netrapalli P, Jordan MI (2019) What is local optimality in nonconvex-nonconcave minimax optimization? arXiv preprint arXiv:190200618
- [42] Juditsky A, Nemirovski A, Tauvel C (2011) Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems* 1(1):17–58
- [43] Kamalaruban P, Huang YT, Hsieh YP, Rolland P, Shi C, Cevher V (2020) Robust reinforcement learning via adversarial training with langevin dynamics. arXiv preprint arXiv:200206063
- [44] Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of gans for improved quality, stability, and variation. In: *International Conference on Learning Representations*
- [45] Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics* 23(3):462–466
- [46] Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
- [47] Kupka I (1963) Contribution à la théorie des champs génériques. *Contributions to differential equations* 2:457–484
- [48] Kushner HJ, Clark DS (1978) *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer
- [49] Kushner HJ, Huang H (1981) Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization* 19(1):87–105
- [50] Kushner HJ, Yin GG (1997) *Stochastic approximation algorithms and applications*. Springer-Verlag, New York, NY
- [51] Lee JM (2003) *Introduction to Smooth Manifolds*. No. 218 in *Graduate Texts in Mathematics*, Springer-Verlag, New York, NY
- [52] Letcher A (2020) On the impossibility of global convergence in multi-loss optimization. arXiv preprint arXiv:200512649
- [53] Liu M, Mroueh Y, Ross J, Zhang W, Cui X, Das P, Yang T (2019) Towards better understanding of adaptive gradient algorithms in generative adversarial nets. arXiv preprint arXiv:1912.11940
- [54] Liu S, Lu S, Chen X, Feng Y, Xu K, Al-Dujaili A, Hong M, Obelilily UM (2019) Min-max optimization without gradients: Convergence and applications to adversarial ml. arXiv preprint arXiv:1909.13806
- [55] Ljung L (1977) Analysis of recursive stochastic algorithms. *IEEE Trans Autom Control* 22(4):551–575
- [56] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. In: *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*
- [57] Malitsky Y, Tam MK (2020) A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization* 30(2):1451–1472
- [58] Mazumdar E, Ratliff LJ, Sastry SS (2020) On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science* 2(1):103–131
- [59] Mertikopoulos P, Lecouat B, Zenati H, Foo CS, Chandrasekhar V, Piliouras G (2019) Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In: *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*
- [60] Mertikopoulos P, Hallak N, Kavis A, Cevher V (2020) On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems* 33
- [61] Mescheder L, Nowozin S, Geiger A (2017) The numerics of gans. In: *Advances in Neural Information Processing Systems*, pp 1825–1835
- [62] Mokhtari A, Ozdaglar A, Pattathil S (2019) A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: proximal point approach. <https://arxiv.org/abs/1901.08511v2>
- [63] Mokhtari A, Ozdaglar A, Pattathil S (2019) A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. arXiv preprint arXiv:190108511
- [64] Nagarajan V, Kolter JZ (2017) Gradient descent gan optimization is locally stable. In: *Advances in neural information processing systems*, pp 5585–5595
- [65] Naveiro R, Insua DR (2019) Gradient methods for solving stackelberg games. In: *International Conference on Algorithmic Decision Theory*, Springer, pp 126–140
- [66] Nemirovski A (2004) 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

- [67] Nemirovski AS (2004) 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
- [68] Nesterov Y (2004) *Introductory Lectures on Convex Optimization: A Basic Course*. No. 87 in *Applied Optimization*, Kluwer Academic Publishers
- [69] Pearlmutter BA (1994) Fast exact multiplication by the hessian. *Neural computation* 6(1):147–160
- [70] Pemantle R (1990) Nonconvergence to unstable points in urn models and stochastic approximations. *Annals of Probability* 18(2):698–712
- [71] Pemantle R (1992) Vertex-reinforced random walk. *Probability Theory and Related Fields* 92:117–136
- [72] Peng W, Dai YH, Zhang H, Cheng L (2020) Training gans with centripetal acceleration. *Optimization Methods and Software* pp 1–19
- [73] Phelps RR (1993) *Convex Functions, Monotone Operators and Differentiability*, 2nd edn. *Lecture Notes in Mathematics*, Springer-Verlag
- [74] Pinto L, Davidson J, Sukthankar R, Gupta A (2017) Robust adversarial reinforcement learning. In: *ICML '17: Proceedings of the 34th International Conference on Machine Learning*
- [75] Popov LD (1980) 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
- [76] Rakhlin A, Sridharan K (2013) Online learning with predictable sequences. In: *COLT '13: Proceedings of the 26th Annual Conference on Learning Theory*
- [77] Robbins H, Monro S (1951) A stochastic approximation method. *Annals of Mathematical Statistics* 22:400–407
- [78] Robinson C (1998) *Dynamical systems: stability, symbolic dynamics, and chaos*. CRC press
- [79] Rockafellar RT (1976) Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization* 14(5):877–898
- [80] Schäfer F, Anandkumar A (2019) Competitive gradient descent. In: *Advances in Neural Information Processing Systems*, pp 7623–7633
- [81] Shub M (1987) *Global Stability of Dynamical Systems*. Springer-Verlag, Berlin
- [82] Smale S (1963) Stable manifolds for differential equations and diffeomorphisms. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze* 17(1-2):97–116
- [83] Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans Autom Control* 37(3):332–341
- [84] Teschl G (2012) *Ordinary differential equations and dynamical systems*, vol 140. *American Mathematical Soc.*
- [85] Wiggins S (2003) *Introduction to applied nonlinear dynamical systems and chaos*, vol 2. Springer Science & Business Media
- [86] Yadav A, Shah S, Xu Z, Jacobs D, Goldstein T (2017) Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:170507364*