MIRROR DESCENT IN SADDLE-POINT PROBLEMS: GOING THE EXTRA (GRADIENT) MILE

PANAYOTIS MERTIKOPOULOS*, HOUSSAM ZENATI‡, BRUNO LECOUAT‡, CHUAN-SHENG FOO‡, VIJAY CHANDRASEKHAR‡¶, AND GEORGIOS PILIOURAS§

Abstract. Owing to their connection with generative adversarial networks (GANs), saddle-point problems have recently attracted considerable interest in machine learning and beyond. By necessity, most theoretical guarantees revolve around convex-concave problems; however, making theoretical inroads towards efficient GAN training crucially depends on moving beyond this classic framework. To make piecemeal progress along these lines, we analyze the widely used mirror descent (MD) method in a class of non-monotone problems – called *coherent* – whose solutions coincide with those of a naturally associated variational inequality. Our first result is that, under *strict* coherence (a condition satisfied by all strictly convex-concave problems), MD methods converge globally; however, they may fail to converge even in simple, bilinear models. To mitigate this deficiency, we add on an "extra-gradient" step which we show stabilizes MD methods by looking ahead and using a "future gradient". These theoretical results are subsequently validated by numerical experiments in GANs.

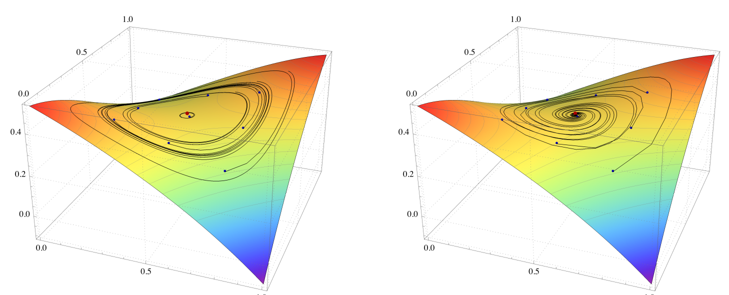

Figure 1: Mirror descent (MD) in a non-monotone saddle-point problem with value function $f(x_1, x_2) = (x_1 - 1/2)(x_2 - 1/2) + \frac{1}{3}\exp(-(x_1 - 1/4)^2 - (x_2 - 3/4)^2)$. Left: vanilla MD; right: extra-gradient MD.

* Univ. Grenoble Alpes, CNRS, Inria, LIG, 38000 Grenoble, France.

‡ Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore.

¶ Nanyang Technological University, 50 Nanyang Ave, Singapore.

§ Singapore University of Technology and Design, 8 Somapah Road, Singapore.

E-mail addresses: panayotis.mertikopoulos@imag.fr, houssam_zenati@i2r.a-star.edu.sg, bruno_lecouat@i2r.a-star.edu.sg, foocs@i2r.a-star.edu.sg, vijay@i2r.a-star.edu.sg, georgios@sutd.edu.sg.

P. Mertikopoulos was partially supported by the French National Research Agency (ANR) grant ORACLESS (ANR–16–CE33–0004–01). G. Piliouras would like to acknowledge SUTD grant SRG ESD 2015 097, MOE AcRF Tier 2 Grant 2016-T2-1-170 and NRF fellowship NRF-NRFF2018-07. This work was partly funded by the deep learning 2.0 program at A*STAR. .

1. Introduction

The surge of recent breakthroughs in artificial intelligence has sparked significant interest in solving optimization problems that are universally considered hard. Such problems are now viewed not only as abstract mathematical questions, but as concrete roadblocks on our way to create robust systems that adapt to complex and unforeseen circumstances. Accordingly, the need for an effective theory has two different sides: First, a deeper understanding that could help demystify the reasons behind the success and/or failures of different training algorithms; second, theoretical advances can inspire concrete algorithmic tweaks leading to empirical performance gains.

Deep learning has been an area of AI where theory has provided a significant boost. For instance, deep learning involves non-convex loss functions for which finding even local optima is NP-hard; nevertheless, elementary techniques such as gradient descent (and other first-order methods) seem to work fairly well in practice. For this class of problems, recent theoretical results have indeed provided useful insights: using tools from the theory of dynamical systems, Lee et al. [2016, 2017] and Panageas and Piliouras [2017] showed that a wide variety of first-order methods (including gradient descent and mirror descent) almost always avoid saddle points. Understanding the geometry of these problems has also helped in providing strong convergence results, as in the work of Du et al. [2017] and Jin et al. [2017]. More generally, the optimization and machine learning communities alike have dedicated significant effort in understanding the geometry of non-convex landscapes by searching for properties which could be leveraged for efficient training. For example, the well-known "strict saddle" property has been shown to hold in a wide range of salient objective functions: low-rank matrix factorization [Bhojanapalli et al., 2016, Ge et al., 2016, 2017], principal component analysis (PCA), a fourth-order tensor factorization [Ge et al., 2015], dictionary learning [Sun et al., 2017a,b], phase retrieval [Sun et al., 2016], and many other models.

On the other hand, concurrent, *adversarial* deep learning models are nowhere near as well understood, especially in the case of generative adversarial networks (GANs) [Goodfellow et al., 2014]. Even though GANs comprise a natural playground for optimization and game theory, and despite the immense amount of scrutiny they have recently received, our theoretical understanding cannot boast similar breakthroughs. To make matters worse, GANs are notoriously hard to train and standard optimization methods often fail to converge. Because of this, a considerable corpus of work has been devoted to exploring and enhancing the stability properties of GANs. In particular, different techniques that have been explored include the use of better-behaved cost metrics such as the Wasserstein distance [Arjovsky et al., 2017], combining the Wasserstein distance with a norm penalty on the critic's gradient [Gulrajani et al., 2017], different activation functions in different layers, feature matching, minibatch discrimination, etc. [Radford et al., 2015, Salimans et al., 2016].

A key observation in this context is that first-order methods may fail to converge even in the toy class of bilinear, zero-sum games like Rock-Paper-Scissors and Matching Pennies [Mertikopoulos et al., 2018, Mescheder et al., 2018]. This basic failure has inspired a handful of recent successful approaches, e.g., consensus optimization [Mescheder et al., 2017], optimistic mirror descent [Daskalakis et al., 2018], symplectic gradient adjustment [Balduzzi et al., 2018], etc. Other failure modes which are unique to GANs have also been revealed thanks to a deeper theoretical analysis of first-order methods in simple models [Arjovsky and Bottou, 2017, Li et al., 2017]. The common theme that arises from this line of work is that, to obtain a principled methodology for training GANs, it is beneficial to start by establishing improvements for a training method in a more restricted setting (often bilinear), and then test whether the gains in performance carry over to more demanding settings.

Following these theoretical breadcrumbs, we focus on a class of possibly non-monotone saddle-point problems that we call *coherent*, and whose solutions coincide with those of a naturally associated variational inequality (VI). Motivated by their prolific success in solving convex-concave zero-sum games, we first study the long-run behavior of the *mirror descent* (MD) class of algorithms in coherent saddle-point problems. On the plus side, we show that if a problem is *strictly* coherent (a condition that is satisfied by strictly convex-concave problems, but not bilinear ones), MD converges almost surely, even with noisy gradient information (Theorem 3.3). On the other hand, if the problem is *not* strictly coherent (e.g., if it is a finite zero-sum game), MD methods provably spiral away from the problem's solutions and cycle in perpetuity (for a schematic illustration, see Fig. 2). Thus, taken by themselves, MD methods do not appear suitable for training convoluted deep learning models.

Inspired by the work of Nemirovski [2004] on monotone VIs, we mitigate this deficiency by including an "extra-gradient" step which allows the algorithm to look ahead and take a prox step along a "future" gradient. This approach was independently pursued in a very recent preprint by Gidel et al. [2018] who introduced an extra "gradient reuse" mechanism for reducing the cost of gradient computations and established the convergence of the algorithm's so-called "ergodic average" in convex-concave saddle-point problems. However, given that averaging offers no tangible benefits beyond convex-concave problems, we focus on the method's "last iterate" and we show it converges in all coherent saddle-point problems; moreover, the distance to a solution decreases *monotonically*, so each iterate is better than the previous one (Theorem 4.1). Since the extra-gradient step has the same computational complexity as an ordinary gradient step, this property is particularly appealing for deep learning models and GAN training. In particular, it suggests that a cheap, extra-gradient add-on can lead to significant performance gains when applied to existing state-of-the-art methods (such as consensus and symplectic gradient adjustment). This theoretical prediction is validated by our numerial experiments in Section 5.

2. Problem setup and preliminaries

2.1. Saddle-point problems.

Consider a saddle-point problem of the general form

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} f(x_1, x_2), \tag{SP}$$

where each feasible region $\mathcal{X}_i$, $i = 1, 2$, is a compact convex subset of a finite-dimensional normed space $\mathcal{V}_i \equiv \mathbb{R}^{d_i}$, and $f : \mathcal{X} \equiv \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ denotes the problem's objective function. From a game-theoretic standpoint, (SP) can be seen as a *zero-sum game* between two optimizing agents (or *players*): Player 1 (the *minimizer*) seeks to incur the least possible loss, while Player 2 (the *maximizer*) seeks to obtain the highest possible reward – both given by $f(x_1, x_2)$.

To obtain a solution of (SP), we will focus on incremental processes that exploit the individual loss/reward gradients of f (assumed throughout to be at least C^1-smooth). Since the individual gradients of f will play a key role in our analysis, we will encode them in a single vector as

$$g(x) = (g_1(x), g_2(x)) = (\nabla_{x_1} f(x_1, x_2), -\nabla_{x_2} f(x_1, x_2)), \tag{2.1}$$

and, following standard conventions, we will treat $g(x)$ as an element of $\mathcal{Y} \equiv \mathcal{V}^*$, the dual of the ambient space $\mathcal{V} \equiv \mathcal{V}_1 \times \mathcal{V}_2$.[1]

2.2. **Variational inequalities and coherence.** Most of the literature on saddle-point problems has focused on the case where f is *convex-concave* (or *semi-convex*), i.e., when $f(x_1, x_2)$ is convex in x_1 and concave in x_2. Under this assumption, von Neumann's [1928] celebrated minmax theorem shows that the order of the min and max operators in (SP) can be interchanged, i.e.,

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} f(x_1, x_2) = \max_{x_2 \in \mathcal{X}_2} \min_{x_1 \in \mathcal{X}_1} f(x_1, x_2). \tag{2.2}$$

In game theory, this is known as the *value* of (SP) and any profile $x^* = (x_1^*, x_2^*) \in \mathcal{X}$ that realizes it is a *Nash equilibrium* (NE): Player 1 cannot further decrease their incurred loss if Player 2 plays x_2^*; similarly, Player 2 cannot further increase their reward if Player 1 sticks to x_1^*.

When f is convex-concave, it is also well known that solutions of (SP) can be characterized equivalently as solutions of the associated (Minty) variational inequality:

$$\langle g(x), x - x^* \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}. \tag{VI}$$

For an in-depth discussion of the links between the formulations (SP) and (VI), see Facchinei and Pang [2003].[2] We only note here that the equivalence between (SP) and (VI) extends well beyond the realm of convex-concave problems: it trivially includes all quasi-convex-concave objectives (where Sion's minmax theorem applies), as well as problems with no semi-convexity properties whatsoever. For instance, the solutions of

$$\min_{x_1 \in [-1,1]} \max_{x_2 \in [-1,1]} x_1^2(1 + x_2^2) \tag{2.3}$$

are all points of the form $(0, t)$ for some $t \in [-1, 1]$. It is easy to check that these are precisely the solutions of (VI), despite the fact that the problem's objective is (strictly) convex-*convex*.

One of our main goals in this paper is to study saddle-point problems that are not necessarily convex-concave. Of course, obtaining global convergence in this case is a daunting task (NP-hard in general), so we will focus on a class of problems which at least preserve the equivalence between (SP) and (VI). We call such problems *coherent:*

Definition 2.1. We say that (SP) is *coherent* if every saddle-point of f is a solution of the associated variational inequality problem (VI) and vice versa. If, in addition, (VI) holds as a strict inequality whenever x is not a saddle-point of f, (SP) will be called *strictly coherent*.

The notion of coherence will play a central part in our considerations, so a few remarks are in order. First, it is possible to relax the equivalence between the two formulations of the problem – (SP) and (VI) – by positing that only *some* of the solutions of (SP) can be harvested from (VI). Our analysis still goes through in this case but, to keep things simple, we do not pursue this relaxation. More importantly, we could also consider a "localized" equivalence between (SP) and (VI) in the sense that local solutions of (SP) coincide with local solutions of (VI). However, since we are chiefly interested in global convergence, we defer this localization analysis to the future.

[1]There are several ways to induce a norm on the product space $\mathcal{V}$ from the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on the component spaces $\mathcal{V}_1$ and $\mathcal{V}_2$. The most convenient for our purposes is to take $\|x\|^2 = \|x_1\|^2 + \|x_2\|^2$.

[2]For completeness, we also prove in Appendix A that (SP) and (VI) are equivalent if f is convex-concave.

Finally, regarding the distinction between coherence and *strict* coherence, we show in Appendix A that (SP) is strictly coherent when f is strictly convex-concave. Typical examples of problems that are coherent but not strictly coherent comprise bilinear objectives with an interior equilibrium: for instance, $f(x_1, x_2) = x_1 x_2$ with $x_1, x_2 \in [-1, 1]$ is coherent (since f is convex-concave) but $\langle g(x), x\rangle = x_1 x_2 - x_2 x_1 = 0$ for all $x_1, x_2 \in [-1, 1]$, indicating that the problem is not strictly coherent. On the contrapositive side, strict coherence *does not* in any way imply a strictly convex structure and/or a unique solution to (SP): for instance, the problem (2.3) is strictly coherent, even though it is convex-convex and does not admit a unique solution.

3. The saddle-point mirror descent algorithm

3.1. **Mirror descent.** Motivated by its prolific success in solving convex programs and monotone variational inequalities, our starting point will be the well-known *mirror descent* (MD) method of Nemirovski and Yudin [1983]. Several variants of the method exist, ranging from dual averaging [Nesterov, 2009, Xiao, 2010] to Nesterov's [2007] dual extrapolation add-on, and many others. For concreteness, we focus here on the original incarnation of the method, suitably adapted to saddle-point problems; for a survey, see Hazan [2012] and Bubeck [2015].

The basic idea of mirror descent is to generate a new state variable x^+ from some starting state x by taking a "mirror step" along a gradient-like vector y. To do this, let $h\colon \mathcal{X} \to \mathbb{R}$ be a continuous and K-strongly convex *distance-generating function* (DGF) on $\mathcal{X}$, i.e.,

$$h(tx + (1-t)x') \leq t h(x) + (1-t)h(x') - \frac{1}{2}Kt(1-t)\|x' - x\|^2, \tag{3.1}$$

for all $x, x' \in \mathcal{X}$ and all $t \in [0, 1]$. In terms of smoothness (and in a slight abuse of notation), we also assume that the subdifferential of h admits a *continuous selection*, i.e., a continuous function $\nabla h\colon \operatorname{dom}\partial h \to \mathcal{Y}$ such that $\nabla h(x) \in \partial h(x)$ for all $x \in \operatorname{dom}\partial h$.[3] Then, following Bregman [1967], h generates a pseudo-distance on $\mathcal{X}$ via the relation

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x\rangle, \tag{3.2}$$

for all $p \in \mathcal{X}$, $x \in \operatorname{dom}\partial h$.

This pseudo-distance is known as the *Bregman divergence.* As we show in Appendix B, it satisfies the basic property $D(p, x) \geq \frac{1}{2}K\|x - p\|^2$, so the convergence of a sequence X_n to some target point p can be verified by showing that $D(p, X_n) \to 0$. On the other hand, $D(p, x)$ may fail to be symmetric and/or satisfy the triangle inequality, so it is not a true distance function per se. Moreover, the level sets of $D(p, x)$ may fail to form a neighborhood basis of p, so the convergence of X_n to p does not necessarily imply that $D(p, X_n) \to 0$; we provide an example of this (admittedly pathological) behavior in Appendix B. For technical reasons, it will be convenient to assume that such phenomena do not occur, i.e., that $D(p, X_n) \to 0$ whenever $X_n \to p$. This mild regularity condition is known in the literature as "Bregman reciprocity" [Chen and Teboulle, 1993, Kiwiel, 1997], and it will be our standing assumption in what follows (note also that it holds trivially for both Examples 3.1 and 3.2 below).

Now, as with standard Euclidean distances, the Bregman divergence generates an associated *prox-mapping* defined as

$$P_x(y) = \operatorname*{arg\,min}_{x' \in \mathcal{X}} \{\langle y, x - x'\rangle + D(x', x)\}, \tag{3.3}$$

[3] Recall here that the subdifferential of h at $x \in \mathcal{X}$ is defined as $\partial h(x) \equiv \{y \in \mathcal{Y} : h(x') \geq h(x) + \langle y, x' - x\rangle \text{ for all } x \in \mathcal{V}\}$, with the standard convention that $h(x) = +\infty$ for all $x \in \mathcal{V} \setminus \mathcal{X}$.

Algorithm 1: saddle-point mirror descent (SPMD)

```
Require: K-strongly convex regularizer h: 𝒳 → ℝ, step-size sequence γ_n > 0
choose X ∈ dom ∂h                                    # initialization
for n = 1, 2, ... do
    oracle query at X returns g                      # gradient feedback
    set X ← P_X(−γ_n g)                              # new state
end for
return X
```

for all $x \in \operatorname{dom} \partial h$ and all $y \in \mathcal{Y}$. In analogy with the Euclidean case (discussed below), the prox-mapping (3.3) produces a feasible point $x^+ = P_x(y)$ by starting from $x \in \operatorname{dom} \partial h$ and taking a step along a dual (gradient-like) vector $y \in \mathcal{Y}$. In this way, we obtain the *saddle-point mirror descent* (SPMD) algorithm

$$X_{n+1} = P_{X_n}(-\gamma_n \hat{g}_n), \tag{SPMD}$$

where γ_n is a variable step-size sequence and $\hat{g}_n$ is the calculated value of the gradient vector $g(X_n)$ at the n-th stage of the algorithm (for a pseudocode implementation, see Section 3.1).

For concreteness, two widely used examples of prox-mappings are as follows:

Example 3.1 (Euclidean projections). When $\mathcal{X}$ is endowed with the L^2 norm $\|\cdot\|_2$, the archetypal prox-function is the (square of the) norm itself, i.e., $h(x) = \frac{1}{2}\|x\|_2^2$. In that case, $D(p, x) = \frac{1}{2}\|x - p\|^2$ and the induced prox-mapping is

$$P_x(y) = \Pi(x + y), \tag{3.4}$$

with $\Pi(x) = \arg\min_{x' \in \mathcal{X}} \|x' - x\|^2$ denoting the ordinary Euclidean projection onto $\mathcal{X}$.

Example 3.2 (Entropic regularization). When $\mathcal{X}$ is a d-dimensional simplex, a widely used DGF is the (negative) Gibbs–Shannon entropy $h(x) = \sum_{j=1}^d x_j \log x_j$. This function is 1-strongly convex with respect to the L^1 norm [Shalev-Shwartz, 2011] and the associated pseudo-distance is the Kullback–Leibler divergence $D_{\mathrm{KL}}(p, x) = \sum_{j=1}^d p_j \log(p_j/x_j)$, leading to the prox-mapping

$$P_x(y) = \frac{(x_j \exp(y_j))_{j=1}^d}{\sum_{j=1}^d x_j \exp(y_j)} \tag{3.5}$$

for all $x \in \mathcal{X}^\circ$, $y \in \mathcal{Y}$. The resulting update $x \leftarrow P_x(y)$ is known in the literature as the *multiplicative weights* (MW) algorithm [Arora et al., 2012], and is one of the centerpieces for learning in zero-sum games [Freund and Schapire, 1999], adversarial bandits [Auer et al., 1995], etc.

Regarding the gradient input sequence $\hat{g}_n$ of (SPMD), we assume that it is obtained by querying a *first-order oracle*, i.e., a black-box feedback mechanism which outputs an estimate of $g(X_n)$ when called at X_n. This oracle could be either *perfect*, returning $\hat{g}_n = g(X_n)$ for all n, or *imperfect*, providing noisy gradient estimations.[4] On that account,

[4]The reason for this is that, depending on the application at hand, gradients might be difficult to compute directly e.g., because they require huge amounts of data, the calculation of an unknown expectation, etc.

we will make the following blanket assumptions for the gradient feedback sequence $\hat{g}_n$:

$$\begin{aligned} &a)\ \textit{Unbiasedness:} && \mathbb{E}[\hat{g}_n \mid \mathcal{F}_n] = g(X_n). \\ &b)\ \textit{Finite mean square:} && \mathbb{E}[\|\hat{g}_n\|_*^2 \mid \mathcal{F}_n] \leq G^2 \quad \text{for some finite } G \geq 0. \end{aligned} \tag{3.6}$$

In the above, $\|y\|_* \equiv \sup\{\langle y, x\rangle : x \in \mathcal{V}, \|x\| \leq 1\}$ denotes the dual norm on $\mathcal{Y}$ while $\mathcal{F}_n$ represents the history (natural filtration) of the generating sequence X_n up to stage n (inclusive). Since $\hat{g}_n$ is generated randomly from X_n at stage n, it is obviously not $\mathcal{F}_n$-measurable, i.e., $\hat{g}_n = g(X_n) + U_{n+1}$, where U_n is an adapted martingale difference sequence with $\mathbb{E}[\|U_{n+1}\|_*^2 \mid \mathcal{F}_n] \leq \sigma^2$ for some finite $\sigma \geq 0$. Clearly, when $\sigma = 0$, we recover the exact gradient feedback framework $\hat{g}_n = g(X_n)$.

3.2. **Convergence analysis.** We are now in a position to state our main results for (SPMD). When the objective of (SP) is convex-concave, it is customary to consider the so-called *ergodic average*

$$\bar{X}_n = \frac{\sum_{k=1}^n \gamma_k X_k}{\sum_{k=1}^n \gamma_k}, \tag{3.7}$$

or some other average of the sequence X_n where the objective is sampled. The reason for this averaging is that convexity guarantees – via Jensen's inequality and gradient monotonicity – that a regret-based analysis of (SPMD) can lead to explicit rates for the convergence of $\bar{X}_n$ to the solution set of (SP) [Bubeck, 2015, Nesterov, 2009]. Beyond convex-concave problems however, this is no longer the case: in general, averaging provides no tangible benefits in a non-monotone setting, so we need to examine the convergence properties of the generating sequence X_n of (SPMD) directly.

The starting point of our analysis is the following stabilization result:

Proposition 3.1. *Suppose that* (SP) *is coherent and* (SPMD) *is run with a gradient oracle satisfying* (3.6) *and a variable step-size γ_n such that $\sum_{n=1}^\infty \gamma_n^2 < \infty$. If $x^* \in \mathcal{X}$ is a solution of* (SP)*, the Bregman divergence $D(x^*, X_n)$ converges (a.s.) to a random variable $D(x^*)$ with $\mathbb{E}[D(x^*)] < \infty$.*

Proposition 3.1 can be seen as a "dichotomy principle": it shows that the Bregman divergence is an asymptotic constant of motion, so (SPMD) either converges to a saddle-point x^* (if $D(x^*) = 0$) or to some nonzero level set of the Bregman divergence (with respect to x^*). In this way, Proposition 3.1 rules out more complicated chaotic or aperiodic behaviors that may arise in general – for instance, as in the analysis of Palaiopanos et al. [2017] for the long-run behavior of the multiplicative weights algorithm in two-player games. However, unless this limit value can be somehow predicted (or estimated) in advance, this result cannot be easily applied. To that end, we turn below to a different question, namely whether (SPMD) gets arbitrarily close to a solution of (SP):

Proposition 3.2. *Suppose that* (SP) *is strictly coherent and* (SPMD) *is run with a gradient oracle satisfying* (3.6) *and a step-size γ_n such that $\sum_{n=1}^\infty \gamma_n = \infty$ and $\sum_{n=1}^\infty \gamma_n^2 < \infty$. Then, with probability* 1*, there exists a (possibly random) solution x^* of* (SP) *such that* $\liminf_{n\to\infty} D(x^*, X_n) = 0$.

In words, Proposition 3.2 shows that (SPMD) admits a subsequence converging to a solution of (SP). This, in view of the dichotomy principle stated above, yields the following convergence result:

Theorem 3.3. *Suppose that* (SP) *is strictly coherent and* (SPMD) *is run with a gradient oracle satisfying* (3.6) *and a variable step-size sequence γ_n such that $\sum_{n=1}^\infty \gamma_n = \infty$ and $\sum_{n=1}^\infty \gamma_n^2 < \infty$. Then, with probability* 1*, X_n converges to a solution of* (SP).

Proof. Proposition 3.2 shows that, with probability 1, there exists a (possibly random) solution x^* of (SP) such that $\liminf_{n\to\infty} \|X_n - x^*\| = 0$ and, hence, $\liminf_{n\to\infty} D(x^*, X_n) = 0$ (by Bregman reciprocity). Since $\lim_{n\to\infty} D(x^*, X_n)$ exists with probability 1 (by Proposition 3.1), it follows that $\lim_{n\to\infty} D(x^*, X_n) = \liminf_{n\to\infty} D(x^*, X_n) = 0$, i.e., X_n converges to x^*. ∎

As an immediate corollary of Theorem 3.3 and the fact that strictly convex-concave problems are also strictly coherent (see Appendix B for the details), we obtain:

Corollary 3.4. *Suppose that f is strictly convex-concave and* (SPMD) *is run with a gradient oracle satisfying* (3.6) *and a variable step-size γ_n such that $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$. Then, with probability* 1*, X_n converges to the* (*necessarily unique*) *solution of* (SP).

On the other hand, if (SP) is coherent but not *strictly* so, (SPMD) may cycle indefinitely instead of converging. In Appendix C, we show that this non-convergent behavior may arise even in very simple models: symmetric two-player games with two pure strategies per player. This failure of mirror descent is due to the fact that, witout a mitigating mechanism in place, a "blind" first-order step could overshoot and lead to an outwards spiral, even with a *vanishing* step-size. This phenomenon becomes even more pronounced in GANs where it can lead to mode collapse and/or cycles between different modes. The next two sections address precisely these issues.

4. Mirror descent with an extra-gradient step

In convex-concave problems, averaging as in (3.7) resolves cycling phenomena by generating an auxiliary sequence that gravitates towards the "center of mass" of the driving sequence X_n (which orbits interior solutions). However, this technique cannot be employed in non-monotone problems because Jensen's inequality does not hold there. In view of this, we take an alternative approach based on an anticipative technique that performs an "extra-gradient" step (possibly outside the convex hull of generated states), and uses the obtained information to "amortize" the next prox step.

The seed of this "extra-gradient" idea dates back to Korpelevich [1976], and has since found wide applications in optimization theory and beyond – for a survey, see Bubeck [2015], Facchinei and Pang [2003], Hazan [2012] and references therein. In a nutshell, given a state x, the method generates an intermediate, "waiting" state $\hat{x} = P_x(-\gamma g(x))$ by taking a prox step as usual. However, instead of continuing from $\hat{x}$, the method samples $g(\hat{x})$ and goes back to the *original* state x in order to generate a new state $x^+ = P_x(-\gamma g(\hat{x}))$. In other words, instead of taking a prox step using the incumbent gradient, the extra-gradient method "looks ahead" and takes a prox step along a "future" gradient.

Based on this heuristic, we obtain the *extra-gradient mirror descent* (EGMD) algorithm

$$\begin{aligned} X_{n+1/2} &= P_{X_n}(-\gamma_n \hat{g}_n) \\ X_{n+1} &= P_{X_n}(-\gamma_n \hat{g}_{n+1/2}) \end{aligned} \tag{EGMD}$$

where, in obvious notation, $\hat{g}_n$ and $\hat{g}_{n+1/2}$ represent gradient oracle queries at the incumbent and intermediate states X_n and $X_{n+1/2}$ respectively (for a pseudocode implementation, see Algorithm 2).

To the best of our knowledge, the closest antecedents to this algorithm are the mirror-prox/dual extrapolation methods of Nemirovski [2004], Juditsky et al. [2011] and Nesterov [2007]. These algorithms essentially have the same iteration structure as (EGMD) but, at the end of their runtime, they return the ergodic average (3.7) of the sequence X_n, not its "last iterate". As we explained above, this kind of averaging is helpful in convex-concave

Algorithm 2: extra-gradient mirror descent (EGMD)

Require: K-strongly convex regularizer $h\colon \mathcal{X} \to \mathbb{R}$, step-size sequence $\gamma_n > 0$

choose $X \in \operatorname{dom} \partial h$ # initialization
for $n = 1, 2, \dots$ **do**
  oracle query at X returns g # gradient feedback
  set $X^+ \leftarrow P_X(-\gamma_n g)$ # waiting state
  oracle query at X^+ returns g^+ # gradient feedback
  set $X \leftarrow P_X(-\gamma_n g^+)$ # new state
end for
return X

problems but, beyond this case, it is not clear whether it offers any tangible benefits: in more general problems, X_n appears to be the most natural solution candidate. Our first convergence result below justifies this intuition:

Theorem 4.1. *Suppose that* (SP) *is coherent and g is L-Lipschitz continuous. If* (EGMD) *is run with exact gradient input and γ_n such that $0 < \lim_{n\to\infty} \gamma_n \leq \sup_n \gamma_n < K/L$, the sequence X_n converges monotonically to a solution x^* of* (SP)*, i.e., $D(x^*, X_n)$ decreases monotonically to 0.*

Theorem 4.1 already shows that the extra-gradient step plays a crucial role in stabilizing (SPMD): not only does (EGMD) converge in problems where (SPMD) provably fails (e.g., zero-sum finite games), but this convergence is, in fact, monotonic. In other words, at each iteration, (EGMD) comes closer to a solution of (SP), whereas (SPMD) may spiral outwards, towards higher and higher values of the Bregman divergence, ultimately converging to a limit cycle (cf. Proposition 3.1). This phenomenon can be seen very clearly in Fig. 2, and also in the detailed theoretical analysis we provide in Appendix C for finite zero-sum games.

Of course, except for very special cases, the monotonic convergence of X_n cannot hold when the gradient input to (EGMD) is imperfect: a single "bad" sample of $\hat{g}_n$ would suffice to throw X_n off-track. Nevertheless, in this case, we still have the following result:

Theorem 4.2. *Suppose that* (SP) *is strictly coherent and* (EGMD) *is run with a gradient oracle satisfying* (3.6) *and a variable step-size sequence γ_n such that $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$. Then, with probability 1, X_n converges to a solution of* (SP).

By this token, the step-size conditions of Theorem 4.2 should be contrasted to those of Theorem 4.1, which allow for constant step-sizes (provided they do not exceed K/L). This extra legroom is afforded by *a*) the lack of randomness (which obviates the summability requirement $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$; and *b*) the Lipschitz gradient condition (which guarantees sufficient descent at each step, provided that the step-size is not too big). Importantly, the maximum allowable step-size is also controlled by the strong convexity modulus of h, suggesting that the choice of distance-generating function can be fine-tuned further to allow for more aggressive step-size policies.

5. Experiments

For validation purposes, we evaluated the extra-gradient add-on in a highly multi-modal mixture of 16 Gaussians arranged in a 4×4 grid as in Metz et al. [2017]. The generator and discriminator have 6 fully connected layers with 384 neurons and Relu activations, and the generator generates 2-dimensional vectors. The output after {2000, 8000, 12000, 16000, 20000} iterations is shown in Table 1. The networks are trained with a specific optimization algorithm as in Balduzzi et al. [2018], Kingma and Ba [2014], Mescheder et al. [2017], Tieleman and Hinton [2012], and are compared to the version with look-ahead (an extra-gradient step). Learning rates and other hyper-parameters were chosen by an inspection of grid search results so as to enable a fair comparison between a method and its look-ahead version. Consensus and SGA are trained with RMSProp [Tieleman and Hinton, 2012].

Overall, the different optimization strategies without look-ahead exhibit mode collapse or oscillations throughout the training period (20000 iterations were chosen to evaluate the hopping behavior of the generator). In all cases, the look-ahead add-on exhibits consistently better behavior in learning the multi-modal distribution and, in particular, greatly reduces occurrences of cycling behavior.

Appendix A. Coherent saddle-point problems

We begin our discussion with some basic results on coherence:

Proposition A.1. *If f is convex-concave,* (SP) *is coherent. In addition, if f is strictly convex-concave,* (SP) *is strictly coherent.*

Proof. Let x^* be a solution point of (SP). Since f is convex-concave, first-order optimality gives

$$\langle g_1(x_1^*, x_2^*), x_1 - x_1^* \rangle = \langle \nabla_{x_1} f(x_1^*, x_2^*), x_1 - x_1^* \rangle \geq 0, \tag{A.1a}$$

and

$$\langle g_2(x_1^*, x_2^*), x_2 - x_2^* \rangle = \langle -\nabla_{x_2} f(x_1^*, x_2^*), x_2 - x_2^* \rangle \geq 0. \tag{A.1b}$$

Combining the two, we readily obtain the (Stampacchia) variational inequality

$$\langle g(x^*), x - x^* \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}. \tag{A.2}$$

In addition to the above, the fact that f is convex-concave also implies that $g(x)$ is *monotone* in the sense that

$$\langle g(x') - g(x), x' - x \rangle \geq 0 \tag{A.3}$$

for all $x, x' \in \mathcal{X}$ Bauschke and Combettes [2017]. Thus, setting $x' \leftarrow x^*$ in (A.3) and invoking (A.2), we get

$$\langle g(x), x - x^* \rangle \geq \langle g(x^*), x - x^* \rangle \geq 0, \tag{A.4}$$

i.e., (VI) is satisfied.

To establish the converse implication, focus for concreteness on the minimizer, and note that (VI) implies that

$$\langle g_1(x), x_1 - x_1^* \rangle \geq 0 \quad \text{for all } x_1 \in \mathcal{X}_1. \tag{A.5}$$

Now, if we fix some $x_1 \in \mathcal{X}_1$ and consider the function $\phi(t) = f(x_1^* + t(x_1 - x_1^*), x_2^*)$, the inequality (A.5) yields

$$\begin{aligned} \phi'(t) &= \langle g(x_1^* + t(x_1 - x_1^*), x_2^*), x_1 - x_1^* \rangle \\ &= \frac{1}{t} \langle g(x_1^* + t(x_1 - x_1^*), x_2^*), x_1^* + t(x_1 - x_1^* - x_1^* \rangle \geq 0, \end{aligned} \tag{A.6}$$

for all $t \in [0, 1]$. This implies that ϕ is nondecreasing, so $f(x_1, x_2^*) = \phi(1) \geq \phi(0) = f(x_1^*, x_2^*)$. The maximizing component follows similarly, showing that x^* is a solution of (SP) and, in turn, establishing that (SP) is coherent.

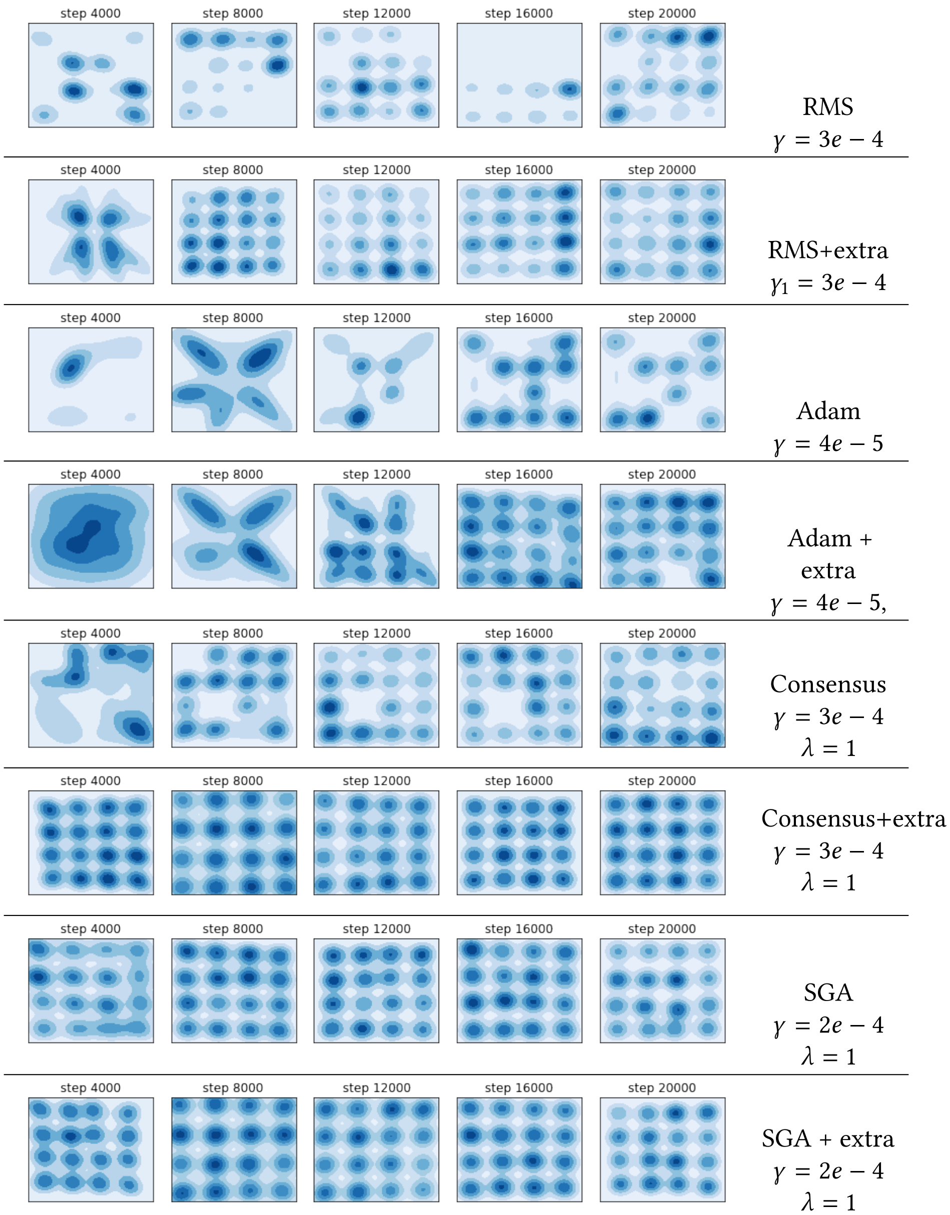

Table 1: Different algorithmic benchmarks: adding an extra-gradient step succeeds at learning the distributions and avoids cycling. λ refers to the regularization weight of Balduzzi et al. [2018], Mescheder et al. [2017]

For the strict part of the claim, the same line of reasoning shows that if $\langle g(x), x - x^* \rangle = 0$ for some x that is not a saddle-point of f, the function $\phi(t)$ defined above must be constant on $[0, 1]$, indicating in turn that f cannot be strictly convex-concave, a contradiction. ■

We proceed to show that the solution set of a coherent saddle-point problem is closed (we will need this regularity result in the convergence analysis of Appendix C):

Lemma A.2. *Let $\mathcal{X}^*$ denote the solution set of* (SP)*. If* (SP) *is coherent, $\mathcal{X}^*$ is closed.*

Proof. Let x^*_n, $n = 1, 2, \dots$, be a sequence of solutions of (SP) converging to some limit point $x^* \in \mathcal{X}$. To show that $\mathcal{X}^*$ is closed, it suffices to show that $x^* \in \mathcal{X}$.

Indeed, given that (SP) is coherent, every solution thereof satisfies (VI), so we have $\langle g(x), x - x^*_n\rangle \geq 0$ for all $x \in \mathcal{X}$. With $x^*_n \to x^*$ as $n \to \infty$, it follows that

$$\langle g(x), x - x^*\rangle = \lim_{n\to\infty} \langle g(x), x - x^*_n\rangle \geq 0 \quad \text{for all } x \in \mathcal{X}, \tag{A.7}$$

i.e., x^* satisfies (VI). By coherence, this implies that x^* is a solution of (SP), as claimed. ■

Appendix B. Properties of the Bregman divergence

In this appendix, we provide some auxiliary results and estimates that are used throughout the convergence analysis of Appendix C. Some of the results we present here (or close variants thereof) are not new [see e.g., Juditsky et al., 2011, Nemirovski et al., 2009]. However, the hypotheses used to obtain them vary wildly in the literature, so we provide all the necessary details for completeness.

To begin, recall that the Bregman divergence associated to a K-strongly convex distance-generating function $h\colon \mathcal{X} \to \mathbb{R}$ is defined as

$$D(p, x) = h(p) - h(x) - \langle \nabla h(x), p - x\rangle \tag{B.1}$$

with $\nabla h(x)$ denoting a continuous selection of $\partial h(x)$. The induced prox-mapping is then given by

$$\begin{aligned} P_x(y) &= \operatorname*{arg\,min}_{x' \in \mathcal{X}} \{\langle y, x - x'\rangle + D(x', x)\} \\ &= \operatorname*{arg\,max}_{x' \in \mathcal{X}} \{\langle y + \nabla h(x), x'\rangle - h(x')\} \end{aligned} \tag{B.2}$$

and is defined for all $x \in \operatorname{dom} \partial h$, $y \in \mathcal{Y}$ (recall here that $\mathcal{Y} \equiv \mathcal{V}^*$ denotes the dual of the ambient vector space $\mathcal{V}$). In what follows, we will also make frequent use of the convex conjugate $h^*\colon \mathcal{Y} \to \mathbb{R}$ of h, defined as

$$h^*(y) = \max_{x \in \mathcal{X}} \{\langle y, x\rangle - h(x)\}. \tag{B.3}$$

By standard results in convex analysis [Rockafellar, 1970, Chap. 26], h^* is differentiable on $\mathcal{Y}$ and its gradient satisfies the identity

$$\nabla h^*(y) = \operatorname*{arg\,max}_{x \in \mathcal{X}} \{\langle y, x\rangle - h(x)\}. \tag{B.4}$$

For notational convenience, we will also write

$$Q(y) = \nabla h^*(y) \tag{B.5}$$

and we will refer to $Q\colon \mathcal{Y} \to \mathcal{X}$ as the *mirror map* generated by h. All these notions are related as follows:

Lemma B.1. *Let h be a distance-generating function on $\mathcal{X}$. Then, for all $x \in \operatorname{dom} \partial h$, $y \in \mathcal{Y}$, we have:*

a) $x = Q(y) \iff y \in \partial h(x)$. (B.6a)

b) $x^+ = P_x(y) \iff \nabla h(x) + y \in \partial h(x^+) \iff x^+ = Q(\nabla h(x) + y)$. (B.6b)

Finally, if $x = Q(y)$ and $p \in \mathcal{X}$, we have

$$\langle \nabla h(x), x - p \rangle \leq \langle y, x - p \rangle. \tag{B.7}$$

Remark. By (B.6b), we have $\partial h(x^+) \neq \varnothing$, i.e., $x^+ \in \operatorname{dom} \partial h$. As a result, the update rule $x \leftarrow P_x(y)$ is *well-posed*, i.e., it can be iterated in perpetuity.

Proof of Lemma B.1. For (B.6a), note that x solves (B.3) if and only if $y - \partial h(x) \ni 0$, i.e., if and only if $y \in \partial h(x)$. Similarly, comparing (B.2) with (B.3), it follows that x^+ solves (B.2) if and only if $\nabla h(x) + y \in \partial h(x^+)$, i.e., if and only if $x^+ = Q(\nabla h(x) + y)$.

For (B.7), by a simple continuity argument, it suffices to show that the inequality holds for interior $p \in \mathcal{X}^\circ$. To establish this, let

$$\phi(t) = h(x + t(p - x)) - [h(x) + \langle y, x + t(p - x) \rangle]. \tag{B.8}$$

Since h is strongly convex and $y \in \partial h(x)$ by (B.6a), it follows that $\phi(t) \geq 0$ with equality if and only if $t = 0$. Since $\psi(t) = \langle \nabla h(x + t(p - x)) - y, p - x \rangle$ is a continuous selection of subgradients of ϕ and both ϕ and ψ are continuous on $[0, 1]$, it follows that ϕ is continuously differentiable with $\phi' = \psi$ on $[0, 1]$. Hence, with ϕ convex and $\phi(t) \geq 0 = \phi(0)$ for all $t \in [0, 1]$, we conclude that $\phi'(0) = \langle \nabla h(x) - y, p - x \rangle \geq 0$, which proves our assertion. ■

We continue with some basic bounds on the Bregman divergence before and after a prox step. The basic ingredient for these bounds is a generalization of the (Euclidean) law of cosines which is known in the literature as the "three-point identity" [Chen and Teboulle, 1993]:

Lemma B.2. *Let h be a distance-generating function on $\mathcal{X}$. Then, for all $p \in \mathcal{X}$ and all $x, x' \in \operatorname{dom} \partial h$, we have*

$$D(p, x') = D(p, x) + D(x, x') + \langle \nabla h(x') - \nabla h(x), x - p \rangle. \tag{B.9}$$

Proof. By definition, we have:

$$\begin{aligned} D(p, x') &= h(p) - h(x') - \langle \nabla h(x'), p - x' \rangle \\ D(p, x) &= h(p) - h(x) - \langle \nabla h(x), p - x \rangle \\ D(x, x') &= h(x) - h(x') - \langle \nabla h(x'), x - x' \rangle. \end{aligned} \tag{B.10}$$

Our claim then follows by adding the last two lines and subtracting the first. ■

With this identity at hand, we have the following series of upper and lower bounds:

Proposition B.3. *Let h be a K-strongly convex distance-generating function on $\mathcal{X}$, fix some $p \in \mathcal{X}$, and let $x^+ = P_x(y)$ for $x \in \operatorname{dom} \partial h$, $y \in \mathcal{Y}$. We then have:*

$$D(p, x) \geq \frac{K}{2} \|x - p\|^2. \tag{B.11a}$$

$$D(p, x^+) \leq D(p, x) - D(x^+, x) + \langle y, x^+ - p \rangle \tag{B.11b}$$

$$\leq D(p, x) + \langle y, x - p \rangle + \frac{1}{2K} \|y\|_*^2 \tag{B.11c}$$

Proof of (B.11a). By the strong convexity of h, we get

$$h(p) \geq h(x) + \langle \nabla h(x), p - x \rangle + \frac{K}{2} \|p - x\|^2 \tag{B.12}$$

so (B.11a) follows by gathering all terms involving h and recalling the definition of $D(p, x)$. ■

Proof of B.11b and (B.11c). By the three-point identity (B.9), we readily obtain

$$D(p, x) = D(p, x^+) + D(x^+, x) + \langle \nabla h(x) - \nabla h(x^+), x^+ - p \rangle. \tag{B.13}$$

In turn, this gives

$$\begin{aligned} D(p, x^+) &= D(p, x) - D(x^+, x) + \langle \nabla h(x^+) - \nabla h(x), x^+ - p \rangle \\ &\leq D(p, x) - D(x^+, x) + \langle y, x^+ - p \rangle, \end{aligned} \tag{B.14}$$

where, in the last step, we used (B.7) and the fact that $x^+ = P_x(y)$, so $\nabla h(x) + y \in \partial h(x^+)$. The above is just (B.11b), so the first part of our proof is complete.

For (B.11c), the bound (B.14) gives

$$D(p, x^+) \leq D(p, x) + \langle y, x - p \rangle + \langle y, x^+ - x \rangle - D(x^+, x). \tag{B.15}$$

Therefore, by Young's inequality [Rockafellar, 1970], we get

$$\langle y, x^+ - x \rangle \leq \frac{K}{2} \|x^+ - x\|^2 + \frac{1}{2K} \|y\|_*^2, \tag{B.16}$$

and hence

$$\begin{aligned} D(p, x^+) &\leq D(p, x) + \langle y, x - p \rangle + \frac{1}{2K} \|y\|_*^2 + \frac{K}{2} \|x^+ - x\|^2 - D(x^+, x) \\ &\leq D(p, x) + \langle y, x - p \rangle + \frac{1}{2K} \|y\|_*^2, \end{aligned} \tag{B.17}$$

with the last step following from Lemma B.1 applied to x in place of p. ■

The first part of Proposition B.3 shows that X_n converges to p if $D(p, X_n) \to 0$. However, as we mentioned in the main body of the paper, the converse may fail: in particular, we could have $\liminf_{n\to\infty} D(p, X_n) > 0$ even if $X_n \to p$. To see this, let $\mathcal{X}$ be the L^2 ball of $\mathbb{R}^d$ and take $h(x) = -\sqrt{1 - \|x\|_2^2}$. Then, a straightforward calculation gives

$$D(p, x) = \frac{1 - \langle p, x \rangle}{\sqrt{1 - \|x\|_2^2}} \tag{B.18}$$

whenever $\|p\|_2 = 1$. The corresponding level sets $L_c(p) = \{x \in \mathbb{R}^d : D(p, x) = c\}$ of $D(p, \cdot)$ are given by the equation

$$1 - \langle p, x \rangle = c\sqrt{1 - \|x\|_2^2}, \tag{B.19}$$

which admits p as a solution for all $c \geq 0$ (so p belongs to the closure of $L_c(p)$ even though $D(p, p) = 0$ by definition). As a result, under this distance-generating function, it is possible to have $X_n \to p$ even when $\liminf_{n\to\infty} D(p, X_n) > 0$ (simply take a sequence X_n that converges to p while remaining on the same level set of D). As we discussed in the main body of the paper, such pathologies are discarded by the Bregman reciprocity condition

$$D(p, X_n) \to 0 \quad \text{whenever} \quad X_n \to p. \tag{B.20}$$

This condition comes into play at the very last part of the proofs of **????**; other than that, we will not need it in the rest of our analysis.

Finally, for the analysis of the EGMD algorithm, we will need to relate prox steps taken along different directions:

Proposition B.4. *Let h be a K-strongly convex distance-generating function on $\mathcal{X}$ and fix some $p \in \mathcal{X}$, $x \in \operatorname{dom} \partial h$. Then:*

a) For all $y_1, y_2 \in \mathcal{Y}$, we have:

$$\|P_x(y_2) - P_x(y_1)\| \leq \frac{1}{K}\|y_2 - y_1\|_*, \tag{B.21}$$

i.e., P_x is $(1/K)$-Lipschitz.

b) In addition, letting $x_1^+ = P_x(y_1)$ and $x_2^+ = P_x(y_2)$, we have:

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) + \langle y_2, x_1^+ - p\rangle + [\langle y_2, x_2^+ - x_1^+\rangle - D(x_2^+, x)] &\text{(B.22a)}\\ &\leq D(p, x) + \langle y_2, x_1^+ - p\rangle + \frac{1}{2K}\|y_2 - y_1\|_*^2 - \frac{K}{2}\|x_1^+ - x\|^2. &\text{(B.22b)}\end{aligned}$$

Proof. We begin with the proof of the Lipschitz property of P_x. Indeed, for all $p \in \mathcal{X}$, (B.7) gives

$$\langle \nabla h(x_1^+) - \nabla h(x) - y_1, x_1^+ - p\rangle \leq 0, \tag{B.23a}$$

and

$$\langle \nabla h(x_2^+) - \nabla h(x) - y_2, x_2^+ - p\rangle \leq 0. \tag{B.23b}$$

Therefore, setting $p \leftarrow x_2^+$ in (B.23a), $p \leftarrow x_1^+$ in (B.23b) and rearranging, we obtain

$$\langle \nabla h(x_2^+) - \nabla h(x_1^+), x_2^+ - x_1^+\rangle \leq \langle y_2 - y_1, x_2^+ - x_1^+\rangle. \tag{B.24}$$

By the strong convexity of h, we also have

$$K\|x_2^+ - x_1^+\|^2 \leq \langle \nabla h(x_2^+) - \nabla h(x_1^+), x_2^+ - x_1^+\rangle. \tag{B.25}$$

Hence, combining (B.24) and (B.25), we get

$$K\|x_2^+ - x_1^+\|^2 \leq \langle y_2 - y_1, x_2^+ - x_1^+\rangle \leq \|y_2 - y_1\|_*\|x_2^+ - x_1^+\|, \tag{B.26}$$

and our assertion follows.

For the second part of our claim, the bound (B.11b) of Proposition B.3 applied to $x_2^+ = P_x(y_2)$ readily gives

$$\begin{aligned} D(p, x_2^+) &\leq D(p, x) - D(x_2^+, x) + \langle y_2, x_2^+ - p\rangle \\ &= D(p, x) + \langle y_2, x_1^+ - p\rangle + [\langle y_2, x_2^+ - x_1^+\rangle - D(x_2^+, x)] \end{aligned} \tag{B.27}$$

thus proving (B.22a). To complete our proof, note that (B.11b) with $p \leftarrow x_2^+$ gives

$$D(x_2^+, x_1^+) \leq D(x_2^+, x) + \langle y_1, x_1^+ - x_2^+\rangle - D(x_1^+, x), \tag{B.28}$$

or, after rearranging,

$$D(x_2^+, x) \geq D(x_2^+, x_1^+) + D(x_1^+, x) + \langle y_1, x_2^+ - x_1^+\rangle. \tag{B.29}$$

We thus obtain

$$\begin{aligned} \langle y_2, x_2^+ - x_1^+\rangle - D(x_2^+, x) &\leq \langle y_2 - y_1, x_2^+ - x_1^+\rangle - D(x_2^+, x_1^+) - D(x_1^+, x) \\ &\leq \frac{\|y_2 - y_1\|_*^2}{2K} + \frac{K}{2}\|x_2^+ - x_1^+\|^2 - \frac{K}{2}\|x_2^+ - x_1^+\|^2 - \frac{K}{2}\|x_1^+ - x\|^2 \\ &\leq \frac{1}{2K}\|y_2 - y_1\|_*^2 - \frac{K}{2}\|x_1^+ - x\|^2, \end{aligned} \tag{B.30}$$

where we used Young's inequality and (B.11a) in the second inequality. The bound (B.22b) then follows by substituting (B.30) in (B.27). ∎

Appendix C. Analysis of the saddle-point mirror descent algorithm

We begin by recalling the definition of the saddle-point mirror descent algorithm. With notation as in the previous section, the algorithm is defined via the recursive scheme

$$X_{n+1} = P_{X_n}(-\gamma_n \hat{g}_n), \tag{SPMD}$$

where γ_n is a variable step-size sequence and $\hat{g}_n$ is the calculated value of the gradient vector $g(X_n)$ at the n-th stage of the algorithm. As we discussed in the main body of the paper, the gradient input sequence $\hat{g}_n$ of (SPMD) is assumed to satisfy the standard oracle assumptions

a) *Unbiasedness:* $\quad \mathbb{E}[\hat{g}_n \mid \mathcal{F}_n] = g(X_n)$.

b) *Finite mean square:* $\quad \mathbb{E}[\|\hat{g}_n\|_*^2 \mid \mathcal{F}_n] \leq G^2$ for some finite $G \geq 0$.

where $\mathcal{F}_n$ represents the history (natural filtration) of the generating sequence X_n up to stage n (inclusive).

With all this at hand, our first step is to prove Proposition 3.1:

Proof of Proposition 3.1. Let $D_n = D(x^*, X_n)$ for some solution x^* of (SP). Then, by Proposition B.3, we have

$$\begin{aligned} D_{n+1} = D(x^*, P_{X_n}(-\gamma_n \hat{g}_n)) &\leq D(x^*, X_n) - \gamma_n \langle \hat{g}_n, X_n - x^* \rangle + \frac{\gamma_n^2}{2K}\|\hat{g}_n\|^2 \\ &= D_n - \gamma_n \langle g(X_n), X_n - x^* \rangle - \gamma_n \langle U_{n+1}, X_n - x^* \rangle + \frac{\gamma_n^2}{2K}\|\hat{g}_n\|_*^2 \\ &\leq D_n + \gamma_n \xi_{n+1} + \frac{\gamma_n^2}{2K}\|\hat{g}_n\|_*^2, \end{aligned} \tag{C.1}$$

where, in the last line, we set $\xi_{n+1} = -\langle U_{n+1}, X_n - x^* \rangle$ and we invoked the assumption that (SP) is coherent. Thus, conditioning on $\mathcal{F}_n$ and taking expectations, we get

$$\mathbb{E}[D_{n+1} \mid \mathcal{F}_n] \leq D_n + \mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] + \frac{\gamma_n^2}{2K}\,\mathbb{E}[\|\hat{g}_n\|_*^2 \mid \mathcal{F}_n] \leq D_n + \frac{G^2}{2K}\gamma_n^2, \tag{C.2}$$

where we used the oracle assumptions (3.6) and the fact that X_n is $\mathcal{F}_n$-measurable (by definition).

Now, letting $R_n = D_n + (2K)^{-1} G^2 \sum_{k=n}^{\infty} \gamma_k^2$, the estimate (C.1) gives

$$\mathbb{E}[R_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[D_{n+1} \mid \mathcal{F}_n] + \frac{G^2}{2K}\sum_{k=n+1}^{\infty}\gamma_k^2 \leq D_n + \frac{G^2}{2K}\sum_{k=n}^{\infty}\gamma_k^2 = R_n, \tag{C.3}$$

i.e., R_n is an $\mathcal{F}_n$-adapted supermartingale. Since $\sum_{n=1}^{\infty}\gamma_n^2 < \infty$, it follows that

$$\mathbb{E}[R_n] = \mathbb{E}[\mathbb{E}[R_n \mid \mathcal{F}_{n-1}]] \leq \mathbb{E}[R_{n-1}] \leq \cdots \leq \mathbb{E}[R_1] \leq \mathbb{E}[D_1] + \frac{G^2}{2K}\sum_{n=1}^{\infty}\gamma_n^2 < \infty, \tag{C.4}$$

i.e., R_n is uniformly bounded in L^1. Thus, by Doob's convergence theorem for supermartingales [Hall and Heyde, 1980, Theorem 2.5], it follows that R_n converges (a.s.) to some finite random variable R_∞ with $\mathbb{E}[R_\infty] < \infty$. In turn, by inverting the definition of R_n, this shows that D_n converges (a.s.) to some random variable $D(x^*)$ with $\mathbb{E}[D(x^*)] < \infty$, as claimed. ∎

We now turn to the proof of existence of a convergent subsequence of (SPMD):

Proof of Proposition 3.2. We begin with the technical observation that the solution set $\mathcal{X}^*$ of (SP) is closed – and hence, compact (cf. Lemma A.2 in Appendix A). Clearly, if $\mathcal{X}^* = \mathcal{X}$, there is nothing to show; hence, without loss of generality, we may assume in what follows that $\mathcal{X}^* \neq \mathcal{X}$.

Assume now ad absurdum that, with positive probability, the sequence X_n generated by (SPMD) admits no limit points in $\mathcal{X}^*$. Conditioning on this event, and given that $\mathcal{X}^*$ is compact, there exists a (nonempty) compact set $\mathcal{C} \subset \mathcal{X}$ such that $\mathcal{C} \cap \mathcal{X}^* = \varnothing$ and $X_n \in \mathcal{C}$ for all sufficiently large n. Moreover, given that (SP) is strictly coherent, we have $\langle g(x), x - x^* \rangle > 0$ whenever $x \in \mathcal{C}$ and $x^* \in \mathcal{X}^*$. Therefore, by the continuity of g and the compactness of $\mathcal{X}^*$ and $\mathcal{C}$, there exists some $a > 0$ such that

$$\langle g(x), x - x^* \rangle \geq a \quad \text{for all } x \in \mathcal{C}, x^* \in \mathcal{X}. \tag{C.5}$$

To proceed, fix some $x^* \in \mathcal{X}^*$ and let $D_n = D(x^*, X_n)$. Then, telescoping (C.1) yields the estimate

$$D_{n+1} \leq D_1 - \sum_{k=1}^{n} \gamma_k \langle g(X_n), X_n - x^* \rangle + \sum_{k=1}^{n} \gamma_k \xi_k + \sum_{k=1}^{n} \frac{\gamma_k^2}{2K} \|\hat{g}_n\|_*^2, \tag{C.6}$$

where, as in the proof of Proposition 3.1, we set $\xi_{n+1} = \langle U_{n+1}, X_n - x^* \rangle$. Subsequently, letting $\tau_n = \sum_{k=1}^{n} \gamma_k$ and using (C.5), we obtain

$$D_{n+1} \leq D_1 - \tau_n \left[a - \frac{\sum_{k=1}^{n} \gamma_k \xi_{k+1}}{\tau_n} - \frac{(2K)^{-1} \sum_{k=1}^{n} \gamma_k^2 \|\hat{g}_k\|_*^2}{\tau_n} \right]. \tag{C.7}$$

By the unbiasedness hypothesis of (3.6) for U_n, we have $\mathbb{E}[\xi_{n+1} \,|\, \mathcal{F}_n] = \langle \mathbb{E}[U_{n+1} \,|\, \mathcal{F}_n], X_n - x^* \rangle = 0$ (recall that X_n is $\mathcal{F}_n$-measurable by construction). Moreover, since U_n is bounded in L^2 and γ_n is ℓ^2 summable (by assumption), it follows that

$$\begin{aligned}
\sum_{n=1}^{\infty} \gamma_n^2 \, \mathbb{E}[\xi_{n+1}^2 \,|\, \mathcal{F}_n] &\leq \sum_{n=1}^{\infty} \gamma_n^2 \|X_n - x^*\|^2 \, \mathbb{E}[\|U_{n+1}\|_*^2 \,|\, \mathcal{F}_n] \\
&\leq \operatorname{diam}(\mathcal{X})^2 \sigma^2 \sum_{n=1}^{\infty} \gamma_n^2 < \infty.
\end{aligned} \tag{C.8}$$

Therefore, by the law of large numbers for martingale difference sequences [Hall and Heyde, 1980, Theorem 2.18], we conclude that $\tau_n^{-1} \sum_{k=1}^{n} \gamma_k \xi_{k+1}$ converges to 0 with probability 1.

Finally, for the last term of (C.6), let $S_{n+1} = \sum_{k=1}^{n} \gamma_k^2 \|\hat{g}_k\|_*^2$. Since $\hat{g}_k$ is $\mathcal{F}_n$-measurable for all $k = 1, 2, \dots, n-1$, we have

$$\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_n] = \mathbb{E}\left[\sum_{k=1}^{n-1} \gamma_k^2 \|\hat{g}_k\|_*^2 + \gamma_n^2 \|\hat{g}_n\|_*^2 \,\middle|\, \mathcal{F}_n \right] = S_n + \gamma_n^2 \, \mathbb{E}[\|\hat{g}_n\|_*^2 \,|\, \mathcal{F}_n] \geq S_n, \tag{C.9}$$

i.e., S_n is a submartingale with respect to $\mathcal{F}_n$. Furthermore, by the law of total expectation, we also have

$$\mathbb{E}[S_{n+1}] = \mathbb{E}[\mathbb{E}[S_{n+1} \,|\, \mathcal{F}_n]] \leq G^2 \sum_{k=1}^{n} \gamma_n^2 \leq G^2 \sum_{k=1}^{\infty} \gamma_n^2 < \infty, \tag{C.10}$$

so S_n is bounded in L^1. Hence, by Doob's submartingale convergence theorem [Hall and Heyde, 1980, Theorem 2.5], we conclude that S_n converges to some (almost surely finite) random variable S_∞ with $\mathbb{E}[S_\infty] < \infty$, implying in turn that $\lim_{n\to\infty} S_{n+1}/\tau_n = 0$ (a.s.).

Applying all of the above, the estimate (C.6) gives $D_{n+1} \leq D_1 - a\tau_n/2$ for sufficiently large n, so $D(x^*, X_n) \to -\infty$, a contradiction. Going back to our original assumption, this shows that, with probability 1, at least one of the limit points of X_n must lie in $\mathcal{X}^*$, as claimed. ■

The two propositions above comprise the main building blocks for the proof of Theorem 3.3. We proceed with the negative result hinted at in the main body of the paper, namely that (SPMD) may fail to converge even in simple, *finite* games.

To define the class of two-player finite games, let $\mathcal{A}_i = \{1, \dots, A_i\}$, $i = 1, 2$, be two finite sets of *pure strategies*, and let $\mathcal{X}_i = \Delta(\mathcal{A}_i)$ denote the set of *mixed strategies* of player i. A *finite, two-player zero-sum game* is then defined by a matrix $M \in \mathbb{R}^{\mathcal{A}_1 \times \mathcal{A}_2}$ so that the loss of Player 1 and the reward of Player 2 in the mixed strategy profile $x = (x_1, x_2) \in \mathcal{X}$ are concurrently given by

$$f(x_1, x_2) = x_1^\top M x_2 \tag{C.11}$$

Then, writing $\Gamma \equiv \Gamma(\mathcal{A}_1, \mathcal{A}_2, M)$ for the resulting game, we have:

Proposition C.1. *Let Γ be a two-player zero-sum game with an interior equilibrium x^*. If $X_1 \neq x^*$ and (SPMD) is run with exact gradient input ($\sigma^2 = 0$), we have $\lim_{n\to\infty} D(x^*, X_n) > 0$. If, in addition, $\sum_{n=1}^\infty \gamma_n^2 < \infty$, $\lim_{n\to\infty} D(x^*, X_n)$ is finite.*

Remark. Note that non-convergence does not require any summability assumptions on γ_n.

In words, Proposition C.1 states that (SPMD) does not converge in finite zero-sum games with a unique interior equilibrium and exact gradient input: instead, X_n cycles at positive Bregman distance from the game's Nash equilibrium. Heuristically, the reason for this behavior is that, for small $\gamma \to 0$, the incremental step $V_\gamma(x) = P_x(-\gamma g(x)) - x$ of (SPMD) is essentially tangent to the level set of $D(x^*, \cdot)$ that passes through x.[5] For finite $\gamma > 0$, things are even worse because $V_\gamma(x)$ points noticeably away from x, i.e., towards higher level sets of D. As a result, the "best-case scenario" for (SPMD) is to orbit x^* (when $\gamma \to 0$); in practice, for finite γ, the algorithm takes small outward steps throughout its runtime, eventually converging to some limit cycle farther away from x^*.

We make this intuition precise below (for a schematic illustration, see also Fig. 2 above):

Proof of Proposition C.1. Write $v_1(x) = -Mx_2$ and $v_2(x) = x_1^\top M$ for the players' payoff vectors under the mixed strategy profile $x = (x_1, x_2)$. By construction, we have $g(x) = -(v_1(x), v_2(x))$. Furthermore, since x^* is an interior equilibrium of f, elementary game-theoretic considerations show that $v_1(x^*)$ and $v_2(x^*)$ are both proportional to the constant vector of ones. We thus get

$$\begin{aligned}\langle g(x), x - x^*\rangle &= \langle v_1(x), x_1 - x_1^*\rangle + \langle v_2(x), x_2 - x_2^*\rangle \\ &= -x_1^\top M x_2 + (x_1^*)^\top M x_2 + x_1^\top M x_2 - x_1^\top M x_2^* \\ &= 0,\end{aligned} \tag{C.12}$$

where, in the last line, we used the fact that x^* is interior.

Now, the evolution of the Bregman divergence under (SPMD) satisfies the identity

$$\begin{aligned}D(x^*, X_{n+1}) &= D(x^*, X_n) + D(X_n, X_{n+1}) + \gamma_n \langle g(X_n), X_n - x^*\rangle \\ &= D(x^*, X_n) + D(X_n, X_{n+1}) \\ &\geq D(x^*, X_n),\end{aligned} \tag{C.13}$$

[5]This observation was also the starting point of Mertikopoulos et al. [2018] who showed that "following the regularized leader" (FTRL) in *continuous time* exhibits a similar cycling behavior in zero-sum games with an interior equilibrium.

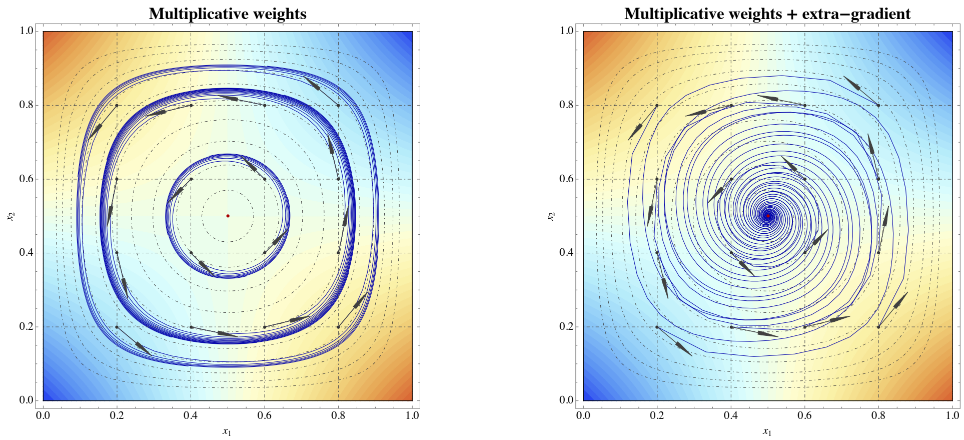

Figure 2: Trajectories of mirror descent in a zero-sum game of Matching Pennies, with and without an extra-gradient step. The color gradient represents the contours of the objective, $f(x_1, x_2) = (x_1 - 1/2)(x_2 - 1/2)$.

where, in the penultimate line, we used the fact that $\langle g(x), x - x^* \rangle = 0$ for all $x \in \mathcal{X}$. This shows that $D(x^*, X_n)$ is nondecreasing, so its limit exists (albeit possibly infinite).

For our second claim, arguing as above and using (B.11c), we get

$$\begin{aligned} D(x^*, X_{n+1}) &\leq D(x^*, X_n) + \gamma_n \langle g(X_n), X_n - x^* \rangle + \frac{\gamma_n^2}{2K} \|g(X_n)\|_*^2 \\ &\leq D(x^*, X_n) + \frac{\gamma_n^2 G^2}{2K} \end{aligned} \tag{C.14}$$

with $G = \max_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} \|(-M x_2, x_1^\top M)\|_*$. Telescoping this last bound yields

$$\sup_n D(x^*, X_n) \leq D(x^*, X_1) + \sum_{k=1}^{\infty} \frac{\gamma_n^2 G^2}{2K} < \infty, \tag{C.15}$$

so $D(x^*, X_n)$ is also bounded from above. Therefore, with $D(x^*, X_n)$ nondecreasing, bounded from above and $D(x^*, X_1) > 0$, it follows that $\lim_{n\to\infty} D(x^*, X_n) > 0$, as claimed. ■

Appendix D. Analysis of the extra-gradient mirror descent algorithm

We now turn to the *extra-gradient mirror descent* (EGMD) algorithm, as defined by the recursion

$$\begin{aligned} X_{n+1/2} &= P_{X_n}(-\gamma_n \hat{g}_n) \\ X_{n+1} &= P_{X_n}(-\gamma_n \hat{g}_{n+1/2}) \end{aligned} \tag{EGMD}$$

with X_1 initialized arbitrarily in $\operatorname{dom} \partial h$, and $\hat{g}_n$, $\hat{g}_{n+1/2}$ representing gradient oracle queries at the incumbent and intermediate states X_n and $X_{n+1/2}$ respectively.

The heavy lifting for our analysis is provided by Proposition B.4, which leads to the following crucial descent lemma:

Lemma D.1. *Suppose that* (SP) *is coherent and g is L-Lipschitz continuous. With notation as above and exact gradient input ($\sigma = 0$), we have*

$$D(x^*, X_{n+1}) \leq D(x^*, X_n) - \frac{1}{2}\left(K - \frac{\gamma_n^2 L^2}{K}\right) \|X_{n+1/2} - X_n\|^2, \tag{D.1}$$

for every solution x^* of (SP).

Proof. Substituting $x \leftarrow X_n$, $y_1 \leftarrow -\gamma_n g(X_n)$, and $y_2 \leftarrow -\gamma_n g(X_{n+1/2})$ in Proposition B.4, we obtain the estimate:

$$\begin{aligned} D(x^*, X_{n+1}) &\leq D(x^*, X_n) - \gamma_n \langle g(X_{n+1/2}), X_{n+1/2} - x^* \rangle \\ &\quad + \frac{\gamma_n^2}{2K} \|g(X_{n+1/2}) - g(X_n)\|_*^2 - \frac{K}{2} \|X_{n+1/2} - X_n\|^2 \\ &\leq D(x^*, X_n) + \frac{\gamma_n^2 L^2}{2K} \|X_{n+1/2} - X_n\|^2 - \frac{K}{2} \|X_{n+1/2} - X_n\|^2, \end{aligned} \tag{D.2}$$

where, in the last line, we used the fact that x^* is a solution of (SP)/(VI), and that g is L-Lipschitz. ■

We are now finally in a position to prove Theorem 4.1 (reproduced below for convenience):

Theorem. *Suppose that* (SP) *is coherent and g is L-Lipschitz continuous. If* (EGMD) *is run with exact gradient input and a step-size sequence γ_n such that*

$$0 < \lim_{n\to\infty} \gamma_n \leq \sup_n \gamma_n < K/L, \tag{D.3}$$

the sequence X_n converges monotonically to a solution x^* of (SP)*, i.e., $D(x^*, X_n)$ is non-increasing and converges to 0.*

Proof. Let x^* be a solution of (SP). Then, by the stated assumptions for γ_n, Lemma D.1 yields

$$D(x^*, X_{n+1}) \leq D(x^*, X_n) - \frac{1}{2} K(1 - \alpha^2) \|X_{n+1/2} - X_n\|^2, \tag{D.4}$$

where $\alpha \in (0, 1)$ is such that $\gamma_n^2 < \alpha K/L$ for all n (that such an α exists is a consequence of the assumption that $\sup_n \gamma_n < K/L$). This shows that $D(x^*, X_n)$ is non-decreasing for every solution x^* of (SP).

Now, telescoping (D.1), we obtain

$$D(x^*, X_{n+1}) \leq D(x^*, X_1) - \frac{1}{2} \sum_{k=1}^{n} \left(K - \frac{\gamma_k^2 L^2}{K} \right) \|X_{k+1/2} - X_k\|^2, \tag{D.5}$$

and hence:

$$\sum_{k=1}^{n} \left(1 - \frac{\gamma_k^2 L^2}{K^2} \right) \|X_{k+1/2} - X_k\|^2 \leq \frac{2}{K} D(x^*, X_1). \tag{D.6}$$

With $\sup_n \gamma_n < K/L$, the above estimate readily yields $\sum_{n=1}^{\infty} \|X_{n+1/2} - X_n\|^2 < \infty$, which in turn implies that $\|X_{n+1/2} - X_n\| \to 0$ as $n \to \infty$.

By the compactness of $\mathcal{X}$, we further infer that X_n admits an accumulation point $\hat{x}$, i.e., there exists a subsequence n_k such that $X_{n_k} \to \hat{x}$ as $k \to \infty$. Since $\|X_{n_k+1/2} - X_{n_k}\| \to 0$, this also implies that $X_{n_k+1/2}$ converges to $\hat{x}$ as $k \to \infty$. Then, by the Lipschitz continuity of the prox-mapping (cf. Proposition B.4), we readily obtain

$$\hat{x} = \lim_{k\to\infty} X_{n_k+1/2} = \lim_{k\to\infty} P_{X_{n_k}}(X_{n_k} - \gamma_{n_k} g(X_{n_k})) = P_{\hat{x}}(\hat{x} - \gamma g(\hat{x})), \tag{D.7}$$

i.e., $\hat{x}$ is a solution of (VI) – and, hence, (SP). Since $D(\hat{x}, X_n)$ is nonincreasing and $\liminf_{n\to\infty} D(\hat{x}, X_n) = 0$ (by the Bregman reciprocity requirement), we conclude that $\liminf_{n\to\infty} D(\hat{x}, X_n) = 0$, i.e., X_n converges to $\hat{x}$. Since $\hat{x}$ is a solution of (SP), our proof is complete. ■

Our last remaining result concerns the convergence of (EGMD) in strictly coherent problems with an imperfect oracle satisfying assumptions (3.6) (Theorem 4.1 in the main body of the paper). The key observation here is that, by the descent estimate (D.1), Eqs. (C.1) and (C.6) continue to hold, so the proof of Propositions 3.1 and 3.2 goes through essentially unchanged. The proof of Theorem 4.2 then follows in exactly the same way as the proof of Theorem 3.3 earlier in Appendix C. To avoid needless repetition, we omit the details.

References

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 214–223, 2017. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, 1995.

D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The Mechanics of n-Player Differentiable Games. In *ICML*, 2018.

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

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–358, 2015.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, August 1993.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *ICLR*, 2018.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Barnabas Poczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. *arXiv preprint arXiv:1705.10412*, 2017.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer, 2003.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

G. Gidel, H. Berard, P. Vincent, and S. Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. *ArXiv e-prints*, February 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5769–5779. Curran Associates, Inc., December 2017. URL https://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans. arxiv: 1704.00028.

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

Elad Hazan. A survey: The convex optimization approach to regret minimization. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, pages 287–304. MIT Press, 2012.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

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

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 12 2014.

Krzysztof C. Kiwiel. Free-steering relaxation methods for problems with strictly convex costs and linear constraints. *Mathematics of Operations Research*, 22(2):326–349, 1997.

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

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Jerry Li, Aleksander Madry, John Peebles, and Ludwig Schmidt. Towards understanding the dynamics of generative adversarial networks. *arXiv preprint arXiv:1706.09884*, 2017.

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

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? https://arxiv.org/abs/1801.04406, 2018.

Lars M. Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *NIPS*, pages 1823–1833, 2017.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *ICLR Proceedings*, 2017.

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

Arkadi Semen Nemirovski and David Berkovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, NY, 1983.

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

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

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

Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *Innovations of Theoretical Computer Science (ITCS)*, 2017.

A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv e-prints*, November 2015.

Ralph Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pages 2234–2242, 2016.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE, 2016.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017a.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017b.

T. Tieleman and G. Hinton. Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning. 2012.

John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. Translated by S. Bargmann as "On the Theory of Games of Strategy" in A. Tucker and R. D. Luce, editors, *Contributions to the Theory of Games IV*, volume 40 of *Annals of Mathematics Studies*, pages 13-42, 1957, Princeton University Press, Princeton.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, October 2010.