arXiv:2002.00057v2 [cs.LG] 7 Jul 2020

Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems

Noah Golowich[*] NZG@MIT.EDU
Department of EECS, Massachusetts Institute of Technology

Sarath Pattathil SARATHP@MIT.EDU
Department of EECS, Massachusetts Institute of Technology

Constantinos Daskalakis[†] COSTIS@CSAIL.MIT.EDU
Department of EECS, Massachusetts Institute of Technology

Asuman Ozdaglar ASUMAN@MIT.EDU
Department of EECS, Massachusetts Institute of Technology

Abstract

In this paper we study the smooth convex-concave saddle point problem. Specifically, we analyze the last iterate convergence properties of the Extragradient (EG) algorithm. It is well known that the ergodic (averaged) iterates of EG converge at a rate of $O(1/T)$ (Nemirovski (2004)). In this paper, we show that the last iterate of EG converges at a rate of $O(1/\sqrt{T})$. To the best of our knowledge, this is the first paper to provide a convergence rate guarantee for the last iterate of EG for the smooth convex-concave saddle point problem. Moreover, we show that this rate is tight by proving a lower bound of $\Omega(1/\sqrt{T})$ for the last iterate. This lower bound therefore shows a quadratic separation of the convergence rates of ergodic and last iterates in smooth convex-concave saddle point problems.

Keywords: Minimax optimization, Extragradient algorithm, Last iterate convergence

1. Introduction

In this paper we study the following saddle-point problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y}), \tag{1}$$

where the function f is smooth, convex in $\mathbf{x}$, and concave in $\mathbf{y}$. This problem is equivalent (Facchinei and Pang (2003)) to finding a global saddle point of the function f, i.e., a point $(\mathbf{x}^*, \mathbf{y}^*)$ such that:

$$f(\mathbf{x}^*,\mathbf{y}) \le f(\mathbf{x}^*,\mathbf{y}^*) \le f(\mathbf{x},\mathbf{y}^*) \quad \forall\, \mathbf{x}\in\mathbb{R}^m, \mathbf{y}\in\mathbb{R}^n. \tag{2}$$

The saddle point problem (1) arises in many fields. Besides its central importance in Game Theory, Online Learning and Convex Programming, it has recently found application in the study of

[*] Supported by a Fannie & John Hertz Foundation Fellowship, an MIT Akamai Fellowship, and an NSF Graduate Fellowship.

[†] Supported by NSF Awards IIS-1741137, CCF-1617730 and CCF-1901292, by a Simons Investigator Award, by the DOE PhILMs project (No. DE-AC05-76RL01830), by the DARPA award HR00111990021, by a Google Faculty award, and by the MIT Frank Quick Faculty Research and Innovation Fellowship.

© 2020 N. Golowich, S. Pattathil, C. Daskalakis & A. Ozdaglar.

generative adversarial networks (GANS) (e.g. Goodfellow et al. (2014); Arjovsky et al. (2017)), adversarial examples (e.g. Madry et al. (2019)), robust optimization (e.g. Ben-Tal et al. (2009)), and reinforcement learning (e.g. Du et al. (2017); Dai et al. (2018)).

The convex-concave minimax problem (1) is a special case of a *monotone variational inequality* (see Section 2), which has been studied since the 1960s (Hartman and Stampacchia (1966); Browder (1965); Lions and Stampacchia (1967); Brezis and Sibony (1968); Sibony (1970)). Several first-order iterative algorithms to approximate the solution to a monotone variational inequality, including the Proximal Point (PP) algorithm (Martinet (1970); Rockafellar (1976)), the extragradient (EG) algorithm (Korpelevich (1976)) and optimistic gradient descent-ascent (OGDA) (Popov (1980)), have been studied. It is known that the optimal rate of convergence for first-order methods for solving monotone variational inequalities (and thus (1)) is $O(1/T)$, and this rate is achieved by both the EG and OGDA algorithms (Nemirovski (2004); Mokhtari et al. (2019a); Hsieh et al. (2019); Monteiro and Svaiter (2010); Auslender and Teboulle (2005); Tseng (2008)). However, such convergence guarantees are only known for the *averaged (ergodic)* iterates: in particular, if $(\mathbf{x}_t, \mathbf{y}_t)$ are the iterates generated by the EG or OGDA algorithm for the convex-concave problem (1), the convergence rate of $O(1/T)$ is known for $(\bar{\mathbf{x}}^{(T)}, \bar{\mathbf{y}}^{(T)}) := (\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^{(t)}, \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}^{(t)})$.

The EG and OGDA algorithms have additionally received significant recent attention due to their ability to improve the training dynamics in GANs (Chavdarova et al. (2019); Gidel et al. (2018a,b); Liang and Stokes (2018); Yadav et al. (2017); Daskalakis et al. (2017)). In the saddle point formulation of GANs, given by (1), the parameters $\mathbf{x}$ and $\mathbf{y}$ correspond to parameters of the generator and the discriminator, which are usually represented by neural networks, and therefore the function f is *not* convex-concave. The goal in such a case is to find a point $(\mathbf{x}^*, \mathbf{y}^*)$ which satisfies a saddle-point property such as (2) *locally*. However, since f is not convex-concave, few, if any, theoretical guarantees are known for the averaged iterates $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T)$; indeed, in practice, the *last iterates* $(\mathbf{x}^{(T)}, \mathbf{y}^{(T)})$ typically have reasonably good performance.

Several works including Korpelevich (1976); Facchinei and Pang (2003); Mertikopoulos et al. (2018) prove that, in the convex-concave case, $\lim_{T\to\infty}(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) = (\mathbf{x}^*, \mathbf{y}^*)$ where $(\mathbf{x}^{(T)}, \mathbf{y}^{(T)})$ are the iterates of EG or OGDA, but they do not establish an upper bound on the *convergence rate* of the quality of the solution $(\mathbf{x}^{(T)}, \mathbf{y}^{(T)})$ to that of $(\mathbf{x}^*, \mathbf{y}^*)$. Such a convergence rate is known for the *best iterate* among $(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(T)}, \mathbf{y}^{(T)})$ for each $T \in \mathbb{N}$ (Facchinei and Pang (2003); Monteiro and Svaiter (2010); Mertikopoulos et al. (2018)), but not on the *last iterate* $(\mathbf{x}^{(T)}, \mathbf{y}^{(T)})$. Finally, in the case that f is *strongly* convex-*strongly* concave, linear convergence rates on the distance between the last iterate and the global min-max point (namely, $\|(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) - (\mathbf{x}^{(*)}, \mathbf{y}^{(*)})\|$) are known (Tseng (1995); Gidel et al. (2018a); Liang and Stokes (2018); Mokhtari et al. (2019b)), but to the best of our knowledge, before our work there were no known convergence rates for the last iterate of EG in the absence of strong convexity. In this paper, we prove the following tight last-iterate convergence guarantees for the EG algorithm in the unbounded setting for different termination criteria including the primal-dual gap and Hamiltonian:

Theorem 1 (Last iterate rate for EG; informal version of Theorem 10) *The EG algorithm has a last-iterate convergence rate of $O(1/\sqrt{T})$ for monotone variational inequalities satisfying first and second order smoothness; this convergence holds when measured with respect to either the square root of the Hamiltonian (Definition 3) or the primal-dual gap (Definition 4).*

Theorem 2 shows that the rate of Theorem 1 is tight. Moreover, it establishes a quadratic separation between the last iterate of the extragradient algorithm (which converges at a rate of $O(1/\sqrt{T})$) and the averaged iterate (which converges at a rate of $O(1/T)$).

Theorem 2 (Lower bound for 1-SCLIs; informal version of Theorem 9) *The $O(1/\sqrt{T})$ last-iterate upper bound of Theorem 1 is tight for all 1-stationary canonical linear iterative methods (which includes EG; see Definition 5).*

1.1. Related work

Upper bounds on last-iterate convergence rates. Motivated by applications in GANs, several recent papers have focused on proving last-iterate convergence guarantees for various min-max optimization algorithms. Linear convergence rates have been established for EG, OGDA and several of their variants, in the bilinear case, where $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{M}\mathbf{y} + \mathbf{b}_1^\top \mathbf{x} + \mathbf{b}_2^\top \mathbf{y}$ (Daskalakis et al. (2017); Liang and Stokes (2018); Gidel et al. (2018a); Mokhtari et al. (2019b); Peng et al. (2019); Zhang and Yu (2020)). Azizian et al. (2019) establishes a similar linear convergence rate for EG, OGDA, and consensus optimization (Mescheder et al. (2017)) applied to general convex-concave f in the case that a global lower bound of $\gamma > 0$ is known on the singular values of the Jacobian of $\begin{pmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{pmatrix}$. Daskalakis and Panageas (2018) study the bilinear case where $\mathbf{x}, \mathbf{y}$ are constrained to lie in the simplex and show that the iterates of the optimistic hedge algorithm converge to a global saddle point, without providing any rates of convergence.

Abernethy et al. (2019) proved linear last-iterate convergence rates for Hamiltonian gradient descent when f belongs to a class of 'sufficiently bilinear' (possibly nonconvex-nonconcave) problems. Although their result does generalize the strongly convex-strongly concave and bilinear cases, it does not include the full generality of the convex-concave setting; moreover, as it requires computing derivatives of the Hamiltonian $\|\nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2$, it is a *second order* method. Hsieh et al. (2019) proved local linear convergence rates of OGDA to local saddle points in the neighborhood of which f is strongly convex-strongly concave. Azizian et al. (2020) describe a class of convex-concave functions for which first-order algorithms such as EG can be accelerated locally (with linear rates). Several recent works (Gidel et al. (2018a,b); Bailey et al. (2019)) analyze alternating gradient descent-ascent and show that the iterates neither converge or diverge, but rather cycle infinitely in a bounded set. Finally, there are several works (Shamir and Zhang (2013); Jain et al. (2019); Ge et al. (2019)) in the literature on *non-smooth* convex minimization that compare the convergence of the last iterate and the averaged iterate; the algorithms considered in these papers require decaying step-sizes in order to achieve last-iterate convergence, and so are not directly comparable to our results

Lower bounds. Using lower bounds for non-smooth convex minimization (Nemirovsky (1992)) as a black box, Nemirovski (2004) gives a lower bound of $\Omega(1/T)$ for first-order methods for the smooth convex-concave saddle point problem; this is achieved by, for instance, the EG algorithm with averaged iterates. Ouyang and Xu (2019) gave a direct proof of this fact, and extended it to the case where $\mathbf{x}, \mathbf{y}$ are affinely constrained. The lower bounds of (Nemirovski (2004); Ouyang and Xu (2019)) rely on Krylov subspace techniques, and therefore only apply in the case where $T \leq n$, where n is the dimension of the problem. Azizian et al. (2019); Ibrahim et al. (2019) amend this issue of dimension-dependence using the *canonical linear iterative (CLI) algorithm* framework of Arjevani and Shamir (2016). The lower bounds in these papers focus primarily on the smooth and

strongly-convex strongly-concave case, and proceed by lower bounding the spectral radius of the operator corresponding to a single iteration of a CLI algorithm. Independently Zhang et al. (2019) developed similar lower bounds for the strongly-convex strongly-concave case.

A significant conceptual hurdle in establishing the tight lower bound of $\Omega(1/\sqrt{T})$ in Theorem 2 is that averaging the iterates of EG produces the asymptotically faster rate of $O(1/T)$. Thus, the framework for our lower bound must rule out such averaging schemes; we do so by proving lower bounds for *stationary* CLI (i.e., SCLI) algorithms, i.e., the iterations are time invariant. The class of SCLI algorithms for which our lower bound applies is essentially the same as that of (Azizian et al., 2019, Theorem 5).

Outline In Section 2 we formally define the problem considered in this paper and introduce some notation. In Section 3, we derive a lower bound for the last iterate of 1-SCLI algorithms, of which EG is a special case, establishing Theorem 2. In Section 4, we derive an upper bound for the last iterate of the EG algorithm under first and second-order smoothness assumptions, establishing Theorem 1.

2. Preliminaries

Notation. Lowercase boldface (e.g., $\mathbf{v}$) denotes a vector and uppercase boldface (e.g., $\mathbf{A}$) denotes a matrix. We use $\|\mathbf{v}\|$ to denote the Euclidean norm of vector $\mathbf{v}$. Throughout this paper we will be considering a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, for convex domains $\mathcal{X} \subseteq \mathbb{R}^{n_\mathbf{x}}, \mathcal{Y} \subseteq \mathbb{R}^{n_\mathbf{y}}$, for some $n_\mathbf{x}, n_\mathbf{y} \in \mathbb{N}$. Write $n = n_\mathbf{x} + n_\mathbf{y}$. We will often write $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ and $\mathbf{z} := (\mathbf{x}, \mathbf{y})$ as the concatenation of the vectors $\mathbf{x}, \mathbf{y}$. The gradient of f with respect to $\mathbf{x}$ and $\mathbf{y}$ at $(\mathbf{x}_0, \mathbf{y}_0)$ are denoted by $\nabla_\mathbf{x} f(\mathbf{x}_0, \mathbf{y}_0)$ and $\nabla_\mathbf{y} f(\mathbf{x}_0, \mathbf{y}_0)$, respectively. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\|\mathbf{A}\|_\sigma$ denotes its spectral norm, i.e., the largest singular value of $\mathbf{A}$. For symmetric matrices $\mathbf{A}, \mathbf{B}$, we write $\mathbf{A} \preceq \mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is positive semidefinite (PSD). The *diameter* of $\mathcal{Z} \subset \mathbb{R}^n$ is $\sup_{\mathbf{z}, \mathbf{z}' \in \mathcal{Z}} \|\mathbf{z} - \mathbf{z}'\|$. For a vector $\mathbf{z} \in \mathbb{R}^n$ and $D > 0$, let $\mathcal{B}(\mathbf{z}, D)$ denote the Euclidean ball centered at $\mathbf{z}$ with radius D. For a complex number $w \in \mathbb{C}$, write $\Re(w), \Im(w)$, respectively, to denote the real and imaginary parts of w; thus $w = \Re(w) + i\Im(w)$.

We assume throughout this paper that the function $f(\mathbf{x}, \mathbf{y})$ is twice differentiable. To the function $f : \mathcal{Z} \to \mathbb{R}$ we associate an operator $F_f : \mathcal{Z} \to \mathbb{R}^n$, defined by $F_f(\mathbf{x}, \mathbf{y}) := \begin{pmatrix} \nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}) \end{pmatrix}$. We usually omit the subscript when the function f is clear. It is well-known (Facchinei and Pang (2003)) that if f is convex-concave, then F is *monotone*, meaning that for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, we have $\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq 0$. In this case, it is well-known (Facchinei and Pang (2003)) that a point $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{Z}$ satisfies the global saddle point property (2) if and only if

$$\langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0 \quad \forall \mathbf{z} \in \mathcal{Z}. \tag{3}$$

Finding a point $\mathbf{z}^*$ satisfying (3) is known as the *variational inequality* problem corresponding to F. To measure the quality of a solution $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ for the saddle point problem (1) or equivalently the variational inequality (3) given by a function f, two measures are typically used in the literature (see, e.g., Nemirovski (2004); Monteiro and Svaiter (2010); Mokhtari et al. (2019a)). The first is the *Hamiltonian*, which is equal to the squared norm of the gradient of f at $(\mathbf{x}, \mathbf{y})$.

Definition 3 (Hamiltonian) *For a function $f : \mathcal{Z} \to \mathbb{R}$, the* **Hamiltonian**[1] *of f at $(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ is:*

$$\mathrm{Ham}_f(\mathbf{x}, \mathbf{y}) := \|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|^2 = \|F_f(\mathbf{z})\|^2.$$

Note that if $(\mathbf{x}, \mathbf{y})$ is a global saddle point of (1), then $\mathrm{Ham}_f(\mathbf{x}, \mathbf{y}) = 0$.

The second quality measure of $(\mathbf{x}, \mathbf{y})$ is the *primal-dual gap*, which measures the amount by which $\mathbf{y}$ fails to maximize $f(\mathbf{x}, \cdot)$ and by which $\mathbf{x}$ fails to minimize $f(\cdot, \mathbf{y})$.

Definition 4 (Primal-Dual Gap) *For $f : \mathcal{Z} \to \mathbb{R}$, and some convex region $\mathcal{X}' \times \mathcal{Y}' \subseteq \mathcal{Z}$, the* **primal-dual gap** *at $(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ with respect to $\mathcal{X}' \times \mathcal{Y}'$ is:*

$$\mathrm{Gap}_f^{\mathcal{X}' \times \mathcal{Y}'}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}' \in \mathcal{Y}'} f(\mathbf{x}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{X}'} f(\mathbf{x}', \mathbf{y}). \tag{4}$$

When the set $\mathcal{X}' \times \mathcal{Y}'$ is clear from context, we shall write $\mathrm{Gap}_f(\mathbf{x}, \mathbf{y})$.

As we work in the unconstrained setting, usually we will have $\mathcal{Z} = \mathbb{R}^n$. In such a case, we cannot obtain meaningful guarantees on the primal-dual gap with respect to the set $\mathcal{X}' \times \mathcal{Y}' = \mathcal{Z} = \mathbb{R}^n$, since the gap may be infinite, if, for instance, f is bilinear. Thus, in the unconstrained setting, it is necessary to restrict $\mathcal{X}' \times \mathcal{Y}'$ to be a compact set; following (Mokhtari et al. (2019a)), for our upper bounds, we will usually consider the primal-dual gap with respect to the set $\mathcal{X}' \times \mathcal{Y}'$ for $\mathcal{X}' = \mathcal{B}(\mathbf{x}^*, D), \mathcal{Y}' = \mathcal{B}(\mathbf{y}^*, D)$ for some $D > 0$. As highlighted in (Mokhtari et al. (2019a)), the iterates $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ of many convergent first-order algorithms, including EG and PP, lie in $\mathcal{B}(\mathbf{x}^*, O(\|\mathbf{x}^{(0)} - \mathbf{x}^*\|)) \times \mathcal{B}(\mathbf{y}^*, O(\|\mathbf{y}^{(0)} - \mathbf{y}^*\|))$. Thus, choosing $D = O(\|\mathbf{x}^* - \mathbf{x}^{(0)}\| + \|\mathbf{y}^* - \mathbf{y}^{(0)}\|)$ ensures that the set $\mathcal{X}' \times \mathcal{Y}'$ contains the convex hull of all the iterates $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$.

3. Lower bound for first-order 1-SCLI algorithms

In this section we prove lower bounds for the convergence of a broad range of first order algorithms including the EG algorithm for the convex-concave problem saddle point problem (1). The class of "hard functions" we use to prove our lower bounds are simply bilinear (and thus convex-concave) functions of the form:

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{M} \mathbf{y} + \mathbf{b}_1^\top \mathbf{x} + \mathbf{b}_2^\top \mathbf{y}, \tag{5}$$

where $\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}, \mathbf{y} \in \mathbb{R}^{n/2}$ for some even $n \in \mathbb{N}$, and $\mathbf{M} \in \mathbb{R}^{n/2 \times n/2}$ is a square matrix. Then the monotone operator $F = F_f : \mathbb{R}^n \to \mathbb{R}^n$ corresponding to f is of the form

$$F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b} \quad \text{where} \quad \mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}, \mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{M} \\ -\mathbf{M}^\top & \mathbf{0} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ -\mathbf{b}_2 \end{pmatrix}. \tag{6}$$

Remark. We will assume that the first iterate $\mathbf{z}^{(0)}$ of all 1-SCLIs considered in this paper is $\mathbf{0} \in \mathbb{R}^n$; this assumption is without loss of generality, since we can modify f by applying a translation of $\mathbf{x}, \mathbf{y}$ in (5) to make this assumption hold for any given $\mathcal{A}$.

For $L, D > 0$, we denote the set of L-Lipschitz operators F of the form in (6) for which $\mathbf{M}$, and therefore, $\mathbf{A}$, is of full rank, and for which $\|\mathbf{A}^{-1}\mathbf{b}\| = D$, by $\mathcal{F}^{\mathrm{bil}}_{n,L,D}$. The parameter D represents

1. Often there is an additional factor of $\frac{1}{2}$ multiplying $\|F_f(\mathbf{z})\|^2$ in the definition of the Hamiltonian (see, e.g., Abernethy et al. (2019)), but for simplicity we opt to drop this factor. We do not use any physical interpretation of the Hamiltonian in this paper.

the distance between the initialization (namely, $\mathbf{0}$) and the optimal point $\mathbf{z}^*$, and also measures the diameter of the balls $\mathcal{X}, \mathcal{Y}$ with respect to which the primal-dual gap is computed for our lower bounds. As discussed in the previous section, this choice of $\mathcal{X}, \mathcal{Y}$ is motivated by the fact that for many convergent algorithms such as EG and PP, the iterates never leave $\mathcal{X}, \mathcal{Y}$. (We also use the same convention for our upper bounds.) For $F \in \mathcal{F}^{\text{bil}}_{n,L,D}$, letting $f : \mathbb{R}^n \to \mathbb{R}$ be such that $F = F_f$, there is a unique global min-max point for f, which is given by $\mathbf{z}^* = -\mathbf{A}^{-1}\mathbf{b}$.

Now we are ready to introduce the class of optimization algorithms we consider, namely 1-stationary canonical linear iterative algorithms:

Definition 5 (**1-SCLI algorithms, Arjevani et al. (2015), Definition 1)** *An algorithm $\mathcal{A}$ producing iterates $\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \ldots, \in \mathbb{R}^n$ with access to a monotone first order oracle F is called a* **1-stationary canonical linear iterative** (**1- SCLI**) *optimization algorithm*[2] *over $\mathbb{R}^n$ if when $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$ for some $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{b} \in \mathbb{R}^n$, the iterates $\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \ldots$ take the form*

$$\mathbf{z}^{(t)} = \mathbf{C}_0(\mathbf{A})\mathbf{z}^{(t-1)} + \mathbf{N}(\mathbf{A})\mathbf{b}, \quad t \geq 1, \tag{7}$$

for some mappings $\mathbf{C}_0, \mathbf{N} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ and initial vector $\mathbf{z}^{(0)} \in \mathbb{R}^n$.

When we wish to show the dependence of the iterates $\mathbf{z}^{(t)}$ on the monotone mapping F of (6) explicitly, we shall write $\mathbf{z}^{(t)}(F)$.

Notice that EG with constant step size $\eta > 0$, is a 1-SCLI, as its updates given an operator F of the form in (6) are of the form

$$\mathbf{z}^{(t)} = \mathbf{z}^{(t-1)} - \eta(\mathbf{A}(\mathbf{z}^{(t-1)} - \eta(\mathbf{A}\mathbf{z}^{(t-1)} + \mathbf{b})) + \mathbf{b}) = (I - (\eta\mathbf{A}) + (\eta\mathbf{A})^2)\mathbf{z}^{(t-1)} - (I - \eta\mathbf{A})\eta\mathbf{b}. \tag{8}$$

In contrast to minimization, in which it is natural to measure the quality of the iterates $\mathbf{z}^{(t)}$ via the function value, there are multiple quality measures, including the Hamiltonian $\mathrm{Ham}_f(\cdot)$ (Definition 3) and the primal-dual gap $\mathrm{Gap}_f(\cdot)$ (Definition 4), for the setting of min-max optimization. We will refer to such a quality measure as *loss function*, formalized as a mapping $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$; note that $\mathcal{L}$ in general depends on F.

Definition 6 (Iteration complexity, Arjevani et al. (2015)) *Fix $L, D > 0$, and let $\mathcal{A}$ be a 1-SCLI algorithm for the saddle point problem for f as in (5), whose description may depend on L, D. Suppose, for each $F \in \mathcal{F}^{\text{bil}}_{n,L,D}$, $\mathcal{A}$ produces iterates $\mathbf{z}^{(t)}(F) \in \mathbb{R}^n$ and suppose an objective (loss) function $\mathcal{L}_F : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is given. Then the* **iteration complexity** *of $\mathcal{A}$ at time T and loss functions $\mathcal{L}_F$, denoted $\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}; T)$, is defined as follows:*

$$\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}; T) := \sup_{F \in \mathcal{F}^{\text{bil}}_{n,L,D}} \left\{ \mathcal{L}_F(\mathbf{z}^{(T)}(F)) \right\}. \tag{9}$$

Definition 6 is slightly different from other definitions of iteration complexity in the literature on convex minimization (Arjevani et al. (2015); Nemirovsky (1992)), in that $\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}; T)$ is often replaced with the potentially larger quantity $\sup_{t\geq T}\{\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}; t)\}$. However, since our goal in this section is to prove *lower bounds* on the iteration complexity, our results in terms of (9) are stronger than those with this alternative definition of iteration complexity.[3]

Finally, we formalize the following convergence property of 1-SCLIs:

2. The "1" in "1-SCLI" denotes that $\mathbf{z}^{(t)}$ depends only on the previous iterate $\mathbf{z}^{(t-1)}$.
3. This additional strength of our results rules out an algorithm which achieves small loss at iteration T for any function F, but has large loss at some iteration $T' > T$. This additional strength to our lower bound could be useful given the cyclical nature of the iterates of many min-max algorithms.

Definition 7 (Consistency, Arjevani et al. (2015), Definition 3) *A 1-SCLI optimization algorithm $\mathcal{A}$ is* **consistent with respect to an invertible matrix $\mathbf{A}$** *if for any $\mathbf{b} \in \mathbb{R}^n$, the iterates $\mathbf{z}^{(t)}$ of $\mathcal{A}$ converge to $-\mathbf{A}^{-1}\mathbf{b}$. $\mathcal{A}$ is called* **consistent** *if it is consistent with respect to all (full-rank) $\mathbf{A}$ of the form (6).*

We shall need the following consequence of consistency.

Lemma 8 (Arjevani et al. (2015), Theorem 5) *If a 1-SCLI optimization algorithm $\mathcal{A}$ is consistent with respect to $\mathbf{A}$, then*

$$\mathbf{C}_0(\mathbf{A}) = I + \mathbf{N}(\mathbf{A})\mathbf{A}. \tag{10}$$

3.1. 1-SCLI lower bound

In this section we state Theorem 9, which gives a lower bound on the convergence rate of 1-SCLIs for convex-concave functions by considering functions f of the form (5).

Theorem 9 (Iteration complexity lower bounds) *Let $\mathcal{A}$ be a consistent 1-SCLI[4] and suppose that the inversion matrix $\mathbf{N}(\cdot)$ of $\mathcal{A}$ is a polynomial of degree at most $k-1$ with real-valued coefficients for some $k \in \mathbb{N}$, and let $L, D > 0$. Then the following iteration complexity lower bounds hold:*

1. *For $F \in \mathcal{F}^{\text{bil}}_{n,L,D}$, set $\mathcal{L}^{\text{Ham}}_F(\mathbf{z}) = \|F(\mathbf{z})\|^2$. Then $\text{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\text{Ham}}; T) \geq \frac{L^2D^2}{20Tk^2}$.*
2. *For $F \in \mathcal{F}^{\text{bil}}_{n,L,D}$, set $\mathcal{L}^{\text{Gap}}_F(\mathbf{z}) = \sup_{\mathbf{y}':\|\mathbf{y}'-\mathbf{y}^*\|\leq D} f(\mathbf{x},\mathbf{y}') - \inf_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}^*\|\leq D} f(\mathbf{x}',\mathbf{y})$. Then $\text{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\text{Gap}}; T) \geq \frac{LD^2}{k\sqrt{20T}}$.*
3. *For $F = F_f \in \mathcal{F}^{\text{bil}}_{n,L,D}$, set $\mathcal{L}^{\text{Func}}_F(\mathbf{z}) = |f(\mathbf{x},\mathbf{y}) - f(\mathbf{x}^*,\mathbf{y}^*)|$. Then*

$$\max\left\{\text{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\text{Func}}; T), \text{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\text{Func}}; 2T)\right\} \geq \frac{LD^2}{36k\sqrt{T}}.$$

It will follow from Theorem 10 that the dependence on L, D, and T of the lower bounds in Theorem 9 is tight; in Proposition 15, we show additionally that the inverse linear dependence on k is also tight, at least for $T = 1$.

Next we discuss the assumptions made on $\mathcal{A}$ in Theorem 9. First we remark that consistency is a standard assumption made in the literature on SCLIs and is satisfied by virtually every SCLI used in practice (see, e.g., Arjevani et al. (2015); Azizian et al. (2019); Ibrahim et al. (2019)). Moreover, if $\mathcal{A}$ is not consistent, then a lower bound of $\Omega(1)$ holds on $\sup_{t\geq T}\{\text{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}, t)\}$ for $\mathcal{L} \in \{\mathcal{L}^{\text{Ham}}, \mathcal{L}^{\text{Gap}}\}$ (though the constant may depend on $\mathcal{A}$): to see this, let $\mathbf{A}$ be some full-rank matrix and $\mathbf{b} \in \mathbb{R}^n$ be so that the iterates $\mathbf{z}^{(t)}$ of $\mathcal{A}$ do not converge to $-\mathbf{A}^{-1}\mathbf{b}$. Since $\mathbf{A}$ is full-rank, neither of $\text{Ham}_f(\mathbf{z}^{(t)}), \text{Gap}_f(\mathbf{z}^{(t)})$ converge to 0.

The assumption in Theorem 9 that $\mathbf{N}(\mathbf{A})$ is a polynomial in $\mathbf{A}$ of degree at most $k-1$ is essentially the same as the one made in (Azizian et al., 2019, Theorem 5), which also studied 1-SCLIs (though in the strongly convex case, deriving linear lower bounds). We remark that *some* assumption on $\mathbf{N}(\mathbf{A})$ is necessary, as the choice $\mathbf{C}_0(\mathbf{A}) = \mathbf{0}, \mathbf{N}(\mathbf{A}) = -\mathbf{A}^{-1}$ leads to $\mathbf{z}^{(t)} = -\mathbf{A}^{-1}\mathbf{b} = \mathbf{z}^*$ for all $t \geq 1$. The assumption of the polynomial dependence of $\mathbf{N}(\mathbf{A})$ on $\mathbf{A}$ may be motivated by the fact that, as noted in Azizian et al. (2019), it includes many known first order 1-SCLI methods, including:

4. More generally, $\mathcal{A}$ may be any 1-SCLI so that (10) holds.

- k-extrapolation methods, in which the single "extra" gradient step in EG is replaced by $k \geq 1$ steps (see (Azizian et al., 2019, Eqn. 13));
- Cyclic Richardson iterations (Opfer and Schober (1984)), in which a single update from $\mathbf{z}^{(t)}$ to $\mathbf{z}^{(t+1)}$ consists of a sequence of k gradient updates with different step-sizes $\eta_1, \ldots, \eta_k$ (so that the step sizes cycle between $\eta_1, \ldots, \eta_k$),

and combinations of the above with varying step-sizes. In particular, Theorem 9 applies to the EG algorithm with constant step size; thus, in light of the fact that the averaged iterates $\bar{\mathbf{z}}_T$ of EG have primal-dual gap bounded by $O\left(\frac{D^2 L}{T}\right)$ ((Mokhtari et al., 2019a, Theorem 3)), Theorem 9 establishes a quadratic gap (in T) in the convergence rate between the averaged and last iterates of EG.[5] Below we provide the proof of item 1 of Theorem 9; the proofs of items 2 and 3 are deferred to Appendix A.1.

Proof (of item 1 of Theorem 9) We claim that for all $t \geq 0$,

$$\mathbf{z}^{(t)} = (\mathbf{C}_0(\mathbf{A})^t - I) \cdot \mathbf{A}^{-1}\mathbf{b}. \tag{11}$$

To see that (11) holds, we argue by induction. The base case is trivial since $\mathbf{z}^{(0)} = \mathbf{0}$. For the inductive hypothesis, note that

$$\begin{aligned}\mathbf{z}^{(t+1)} =& \mathbf{C}_0(\mathbf{A}) \cdot (\mathbf{C}_0(\mathbf{A})^t - I) \cdot \mathbf{A}^{-1}\mathbf{b} + \mathbf{N}(\mathbf{A})\mathbf{b} \\ =& \mathbf{C}_0(\mathbf{A}) \cdot (\mathbf{C}_0(\mathbf{A})^t - I) \cdot \mathbf{A}^{-1}\mathbf{b} + (\mathbf{C}_0(\mathbf{A}) - I) \cdot \mathbf{A}^{-1}\mathbf{b} = (\mathbf{C}_0(\mathbf{A})^{t+1} - I) \cdot \mathbf{A}^{-1}\mathbf{b},\end{aligned}$$

where the second equality uses consistency of $\mathcal{A}$ and Lemma 8.

From (11) it follows that

$$\mathrm{Ham}_f(\mathbf{z}^{(t)}) = \|\mathbf{A}(\mathbf{C}_0(\mathbf{A})^t - I)\mathbf{A}^{-1}\mathbf{b} + \mathbf{b}\|^2 = \|\mathbf{A}\mathbf{C}_0(\mathbf{A})^t\mathbf{A}^{-1}\mathbf{b}\|^2 = \|\mathbf{C}_0(\mathbf{A})^t\mathbf{b}\|^2, \tag{12}$$

where (12) follows from the fact that $\mathbf{C}_0(\mathbf{A})$ is a polynomial in $\mathbf{A}$ with scalar coefficients, and therefore $\mathbf{A}$ and $\mathbf{C}_0(\mathbf{A})$ commute.

Next we describe the choice of $\mathbf{A}, \mathbf{b}$: given a dimension $n \in \mathbb{N}$, Lipschitz constant $L > 0$ and a diameter parameter $D > 0$, for some $\nu \in (0, L)$ (to be specified later), we set

$$\mathbf{M} = \nu I \in \mathbb{R}^{n/2 \times n/2}, \quad \mathbf{b}_1 = \mathbf{b}_2 = \begin{pmatrix} \nu D/\sqrt{n} \\ \vdots \\ \nu D/\sqrt{n} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{M} \\ -\mathbf{M}^\top & \mathbf{0} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ -\mathbf{b}_2 \end{pmatrix}. \tag{13}$$

From our choice of $\mathbf{A}$ and the fact that $\|\mathbf{A}^{-1}\mathbf{b}\| = \nu^{-1}\|\mathbf{b}\|$ for all $\mathbf{b} \in \mathbb{R}^n$, it follows from (12) and $\mathbf{z}^{(0)} = \mathbf{0}$ that

$$\frac{\mathrm{Ham}_f(\mathbf{z}^{(t)})}{\|\mathbf{z}^{(0)} - \mathbf{z}^*\|^2} = \frac{\|\mathbf{C}_0(\mathbf{A})^t\mathbf{b}\|^2}{\|\mathbf{A}^{-1}\mathbf{b}\|^2} = \frac{\nu^2\|\mathbf{C}_0(\mathbf{A})^t\mathbf{b}\|^2}{\|\mathbf{b}\|^2}. \tag{14}$$

5. Note that the upper bounds of Mokhtari et al. (2019a) for EG actually apply to the averages of $\mathbf{z}_{t+1/2} = \mathbf{z}_t - \eta F(\mathbf{z}_t)$ as opposed to the averages of $\mathbf{z}_t$. This does not cause a problem for the separation since our lower bound on $\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{z}_T)$ (with $\mathcal{Z} = \mathcal{B}(\mathbf{x}^*, D) \times \mathcal{B}(\mathbf{y}^*, D)$) can be easily extended to a lower bound on $\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{z}_{T+1/2})$ as long as $\eta < 1/L$ by noting that for F_f L-smooth, $\|F_f(\mathbf{z}_{T+1/2})\| = \|F_f(\mathbf{z}_T - \eta F_f(\mathbf{z}_T))\| \geq (1 - \eta L)\|F_f(\mathbf{z}_T)\|$, and for the functions f used in the proof of Theorem 9 (see (5)), we have $\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{z}) = D\|F_f(\mathbf{z})\|$ for all $\mathbf{z} \in \mathbb{R}^n$.

Recall the assumption that $\mathbf{N}(\mathbf{A})$ is a polynomial in $\mathbf{A}$ of degree $k-1$ with scalar coefficients. Moreover, by consistency, we have $\mathbf{C}_0(\mathbf{A}) = I + \mathbf{N}(\mathbf{A})\mathbf{A}$, so $\mathbf{C}_0(\mathbf{A})$ is a polynomial in $\mathbf{A}$ of degree k with scalar coefficients. Thus we may write $\mathbf{C}_0(\mathbf{A}) = q_{0,0}I + q_{0,1} \cdot \mathbf{A} + \cdots + q_{0,k} \cdot \mathbf{A}^k$, where $q_{0,0}, \ldots, q_{0,k} \in \mathbb{R}$ and $q_{0,0} = 1$. Write

$$q_0(y) := q_{0,0} + q_{0,1}y + \cdots + q_{0,k}y^k$$

for $y \in \mathbb{C}$. It is easily verified that $\mathbf{A}$ has $n/2$ eigenvalues equal to νi and $n/2$ eigenvalues equal to $-\nu i$. Therefore, by the spectral mapping theorem (see, e.g., (Lax, 2007, Theorem 4)), $\mathbf{C}_0(\mathbf{A})$ has $n/2$ eigenvalues equal to each of $q_0(\nu i)$ and $q_0(-\nu i) = \overline{q_0(\nu i)}$. Notice that our choice of $\mathbf{A}$ in (13) is normal;[6] hence $\mathbf{C}_0(\mathbf{A})$ is normal as well, meaning the magnitudes of its eigenvalues are equal to its singular values. In particular, all singular values of $\mathbf{C}_0(\mathbf{A})$ are equal to $|q_0(\nu i)|$. Thus, for any vector $\mathbf{b} \in \mathbb{R}^n$, $\|\mathbf{C}_0(\mathbf{A}) \cdot \mathbf{b}\| = |q_0(\nu i)| \cdot \|\mathbf{b}\|$. It follows that

$$\begin{aligned}
\sup_{\nu \in (0,L]} \frac{\nu^2 \|\mathbf{C}_0(\mathbf{A})^t \mathbf{b}\|^2}{\|\mathbf{b}\|^2} &= \sup_{\nu \in (0,L]} \nu^2 |q_0(\nu i)|^{2t} \\
&\geq \sup_{\nu \in (0,L]} \nu^2 \left| \sum_{0 \leq k' \leq \lfloor k/2 \rfloor} (-1)^{k'} q_{0,2k'} \cdot \nu^{2k'} \right|^{2t} && (15) \\
&= \sup_{y \in (0,L^2]} y \cdot \left| \sum_{0 \leq k' \leq \lfloor k/2 \rfloor} (-1)^{k'} q_{0,2k'} \cdot y^{k'} \right|^{2t} \\
&> \frac{L^2}{20tk^2}, && (16)
\end{aligned}$$

where (16) follows from Lemma 13 (see Section A.2). The desired bound in item 1 of the theorem statement follows from (14) with $t = T$ and the fact that $\|\mathbf{A}^{-1}\mathbf{b}\| = D$. ■

4. Upper bound for extragradient

In this section, we discuss upper bounds for the last iterate of the Extragradient (EG) algorithm. The updates of EG algorithm can be written as:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^{(t+1/2)}, \mathbf{y}^{(t+1/2)}), \qquad \mathbf{y}^{(t+1)} = \mathbf{y}^{(t)} + \eta \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1/2)}, \mathbf{y}^{(t+1/2)})$$

where

$$\mathbf{x}^{(t+1/2)} = \mathbf{x}^{(t)} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \qquad \mathbf{y}^{(t+1/2)} = \mathbf{y}^{(t)} + \eta \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$$

This algorithm can be succinctly written in terms of the operator $F = F_f$, and the concatenated vector $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ as:

$$\mathbf{z}^{(t+1/2)} = \mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)}) \qquad \mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t+1/2)})$$

Let $\partial F \in \mathbb{R}^{n \times n}$ denote the matrix of partial derivatives of F; in particular, $(\partial F)_{i,j} = \frac{\partial F_i(\mathbf{z})}{\partial \mathbf{z}_j}$. Our upper bound on convergence rates makes use of the following two assumptions, namely of the Lipschitzness of F and ∂F:

6. A matrix $\mathbf{A}$ is *normal* if and only if there exists a unitary matrix $\mathbf{U}$ so that $\mathbf{U}\mathbf{A}\mathbf{U}^*$ is diagonal. It is known that if $\mathbf{A}$ is normal, then the magnitudes of its eigenvalues are equal to its singular values.

Assumption 1 *For some $L > 0$, the operator F is L-**Lipschitz**, i.e., for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, we have that $\|F(\mathbf{z}) - F(\mathbf{z}')\| \leq L\|\mathbf{z} - \mathbf{z}'\|$.*

In the case that $F = F_f$, the assumption that F is L-Lipschitz is simply a smoothness assumption on f.

Assumption 2 *For some $\Lambda > 0$, the operator F has a Λ-**Lipschitz derivative**, i.e., for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, we have that $\|\partial F(\mathbf{z}) - \partial F(\mathbf{z}')\|_\sigma \leq \Lambda\|\mathbf{z} - \mathbf{z}'\|$.*

Assumption 2 is standard in the literature on *second-order* optimization, both in the minimax setting (see, e.g., (Abernethy et al., 2019, Definition 2.5)) and in the setting of minimization (see, e.g., Nesterov (2006)). Even for first-order algorithms, we believe that Assumption 2 is necessary to obtain a $O(1/\sqrt{T})$ last-iterate convergence rate for convex-concave saddle point optimization, and leave a proof (or disproof) of this fact as an open problem.

In this section our goal is to prove the following theorem.

Theorem 10 *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone operator that is L-Lipschitz (Assumption 1) and has Λ-Lipschitz derivative (Assumption 2). Fix some $\mathbf{z}^{(0)} \in \mathbb{R}^n$, and suppose there is $\mathbf{z}^* \in \mathbb{R}^n$ so that $F(\mathbf{z}^*) = 0$ and $\|\mathbf{z}^* - \mathbf{z}^{(0)}\| \leq D$. If the extragradient algorithm with step size $\eta \leq \min\left\{\frac{5}{\Lambda D}, \frac{1}{30L}\right\}$ is initialized at $\mathbf{z}^{(0)}$, then its iterates $\mathbf{z}^{(T)}$ satisfy*

$$\|F(\mathbf{z}^{(T)})\| \leq \frac{2D}{\eta\sqrt{T}}. \tag{17}$$

If moreover $\mathcal{Z} = \mathcal{B}(\mathbf{x}^*, D) \times \mathcal{B}(\mathbf{y}^*, D)$ and $F(\cdot) = F_f(\cdot) = \begin{pmatrix} \nabla_{\mathbf{x}} f(\cdot) \\ -\nabla_{\mathbf{y}} f(\cdot) \end{pmatrix}$ for a convex-concave function f, then

$$\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) = \max_{\mathbf{y}' \in \mathcal{B}(\mathbf{y}^*, D)} f(\mathbf{x}^{(T)}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{B}(\mathbf{x}^*, D)} f(\mathbf{x}', \mathbf{y}^{(T)}) \leq \frac{2\sqrt{2}D^2}{\eta\sqrt{T}} \tag{18}$$

for all $T \in \mathbb{N}$.

4.1. Proximal point algorithm

Before proving Theorem 10, we briefly discuss similar convergence bounds for an "idealized" version of EG, namely the proximal point (PP) algorithm (see Monteiro and Svaiter (2010); Mokhtari et al. (2019a)). The updates of the PP algorithm are given by $\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t+1)})$. As shown in Mokhtari et al. (2019a), the ergodic iterates of PP and EG have the same rate of convergence (for a constant step size η); moreover, Mokhtari et al. (2019b) showed that the EG algorithm can be viewed as an approximation of the PP algorithm for bilinear functions. It is natural to wonder whether the same rate of $O(1/\sqrt{T})$ of Theorem 10 applies to the PP algorithm as well. This is indeed the case, even without the assumption of F having Λ-Lipschitz derivatives and F being L-Lipschitz. The proof of this (Theorem 20) is provided in Appendix C, and it relies on $\|F(\mathbf{z}^{(t)}\|$ decreasing monotonically.

4.2. Proof of Theorem 10

The proof of Theorem 10 proceeds by first using the well-known fact (Facchinei and Pang (2003); Mertikopoulos et al. (2018); Mokhtari et al. (2019a)) that for any $T \in \mathbb{N}$, there is some $t^* \in \{1, 2, \ldots, T\}$ so that the t^*th iterate $\mathbf{z}^{(t^*)} = (\mathbf{x}^{(t^*)}, \mathbf{y}^{(t^*)})$ obtains the upper bound in (17), namely that $\|F(\mathbf{z}^{(t^*)})\| \leq \frac{2D}{\eta\sqrt{T}}$ [7]; this step relies only on L-Lipschitzness of F (Assumption 1). The bulk of the proof is then to use Assumption 2 to show that $\|F(\mathbf{z}^{(t)})\|$ does not increase much above $\|F(\mathbf{z}^{(t^*)}\|$ for all $t^* < t \leq T$, from which (17) follows. Finally (18) is an immediate consequence of (17) and the fact that F is convex-concave.

Proof (of Theorem 10). Recall that the iterates of the extragradient algorithm are given by

$$\mathbf{z}^{(t+1/2)} = \mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)}), \qquad \mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t+1/2)}).$$

By Lemma 5(b) in Mokhtari et al. (2019a), we have that for any $T > 0$,

$$\sum_{t=0}^{T-1} \eta^2 \|F(\mathbf{z}^{(t)})\|^2 = \sum_{t=0}^{T-1} \|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1/2)}\|^2 \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{1 - \eta^2 L^2} \leq \frac{D^2}{1 - \eta^2 L^2}.$$

Thus there is some $t^* \in \{0, 1, 2, \ldots, T-1\}$ so that

$$\|F(\mathbf{z}^{(t^*)})\|^2 \leq \frac{D^2}{T\eta^2(1 - \eta^2 L^2)}. \tag{19}$$

Next we show that for each $t \in \{1, 2, \ldots, T-1\}$, $\|F(\mathbf{z}^{(t+1)}\|^2$ is not much greater than $\|F(\mathbf{z}^{(t)}\|^2$. To do so we need two lemmas; the first, Lemma 11, uses Assumption 2 to write each $F(\mathbf{z}^{(t+1)})$ in terms of $F(\mathbf{z}^{(t)})$.

Lemma 11 *For all $\mathbf{z} \in \mathcal{Z}$, there are some matrices $\mathbf{A}_\mathbf{z}, \mathbf{B}_\mathbf{z}$ so that $\mathbf{A}_\mathbf{z} + \mathbf{A}_\mathbf{z}^\top$ and $\mathbf{B}_\mathbf{z} + \mathbf{B}_\mathbf{z}^\top$ are PSD and*

$$F(\mathbf{z} - \eta F(\mathbf{z} - \eta F(\mathbf{z}))) = F(\mathbf{z}) - \eta \mathbf{A}_\mathbf{z} F(\mathbf{z}) + \eta^2 \mathbf{A}_\mathbf{z} \mathbf{B}_\mathbf{z} F(\mathbf{z}). \tag{20}$$

and

$$\|\mathbf{A}_\mathbf{z} - \mathbf{B}_\mathbf{z}\|_\sigma \leq \frac{\eta\Lambda}{2} \|F(\mathbf{z}) - F(\mathbf{z} - \eta F(\mathbf{z}))\|, \quad \|\mathbf{A}_\mathbf{z}\|_\sigma \leq L, \quad \|\mathbf{B}_\mathbf{z}\|_\sigma \leq L. \tag{21}$$

The proof of Lemma 11 is provided in Section B.2. Next, Lemma 12 will be used to upper bound the norm of the right-hand side of (20).

Lemma 12 *Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ are matrices so that $\mathbf{A} + \mathbf{A}^\top$ and $\mathbf{B} + \mathbf{B}^\top$ are PSD and $\|\mathbf{A}\|_\sigma, \|\mathbf{B}\|_\sigma \leq 1/30$. Then $\|I - \mathbf{A} + \mathbf{A}\mathbf{B}\|_\sigma \leq \sqrt{1 + 26\|\mathbf{A} - \mathbf{B}\|_\sigma^2}$.*

The proof of Lemma 12 is deferred to Section B.3.

7. This is immediate for the Proximal Point algorithm

By Lemma 11 and Lemma 12 with $\mathbf{A} = \eta \mathbf{A}_{\mathbf{z}^{(t)}}, \mathbf{B} = \eta \mathbf{B}_{\mathbf{z}^{(t)}}$, we have that, as long as $\eta < 1/(30L)$,

$$\begin{aligned}
\|F(\mathbf{z}^{(t+1)})\|^2 &\le \|I - \eta \mathbf{A}_{\mathbf{z}^{(t)}} + \eta^2 \mathbf{A}_{\mathbf{z}^{(t)}} \mathbf{B}_{\mathbf{z}^{(t)}}\|_\sigma^2 \cdot \|F(\mathbf{z}^{(t)})\|^2 \\
&\le (1 + 26\eta^2 \|\mathbf{A}_{\mathbf{z}^{(t)}} - \mathbf{B}_{\mathbf{z}^{(t)}}\|^2) \cdot \|F(\mathbf{z}^{(t)})\|^2 \\
&\le (1 + 7\eta^4 \Lambda^2 \cdot \|F(\mathbf{z}^{(t)}) - F(\mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)}))\|^2) \cdot \|F(\mathbf{z}^{(t)})\|^2 \\
(F \text{ is } L\text{-Lipschitz}) &\le (1 + 7\eta^4 \Lambda^2 \cdot \eta^2 L^2 \|F(\mathbf{z}^{(t)})\|^2) \cdot \|F(\mathbf{z}^{(t)})\|^2 \\
&\le (1 + (\eta^4 \Lambda^2 / 100) \cdot \|F(\mathbf{z}^{(t)})\|^2) \cdot \|F(\mathbf{z}^{(t)})\|^2.
\end{aligned}$$

Next we will prove by induction that for all $t \in \{t^*, t^*+1, \ldots, T\}$, we have that $\|F(\mathbf{z}^{(t)})\|^2 \le \frac{2D^2}{\eta^2 T}$. The base case is immediate by (19). To see the inductive step, note that if for all $t' \in \{t^*, \ldots, t\}$, $\|F(\mathbf{z}^{(t')})\|^2 \le \frac{2D^2}{\eta^2 T}$, then

$$\begin{aligned}
\|F(\mathbf{z}^{(t+1)})\|^2 &\le \|F(\mathbf{z}^{(t)})\|^2 \cdot \left(1 + \frac{\Lambda^2 \eta^2 D^2}{50T}\right) \\
&\le \|F(\mathbf{z}^{(t^*)})\|^2 \cdot \left(1 + \frac{\Lambda^2 \eta^2 D^2}{50T}\right)^{t+1-t^*} \\
(\text{since } \eta < 1/(30L)) &\le \frac{D^2}{\eta^2 T(1 - 1/900)} \cdot \left(1 + \frac{\Lambda^2 \eta^2 D^2}{50T}\right)^T \le \frac{2D^2}{\eta^2 T},
\end{aligned}$$

where the last inequality holds as long as $\Lambda^2 \eta^2 D^2 / 50 \le \frac{1}{2}$, or equivalently, $\eta \le \frac{5}{\Lambda D}$. In particular, we get that $\|F(\mathbf{z}^{(T)})\| \le \frac{2D}{\eta\sqrt{T}}$. If $F(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{pmatrix}$, for some convex-concave function f, then, writing $\mathcal{X} = \mathcal{B}(\mathbf{x}^*, D), \mathcal{Y} = \mathcal{B}(\mathbf{y}^*, D)$, we have

$$\begin{aligned}
&\max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}^{(T)}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{X}} f(\mathbf{x}', \mathbf{y}^{(T)}) \\
&\le \max_{\mathbf{y}' \in \mathcal{Y}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}), \mathbf{y}' - \mathbf{y}^{(T)} \rangle + \max_{\mathbf{x}' \in \mathcal{X}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}), \mathbf{x}^{(T)} - \mathbf{x}' \rangle \\
&= \max_{\mathbf{z}' \in \mathcal{Z}} \langle F(\mathbf{z}^{(T)}), \mathbf{z}^{(T)} - \mathbf{z}' \rangle \le \|F(\mathbf{z}^{(T)})\| \cdot D\sqrt{2} \le \frac{2\sqrt{2} D^2}{\eta\sqrt{T}}.
\end{aligned}$$

■

5. Conclusion and Future Work

In this paper we establish a $O(1/\sqrt{T})$ upper bound on the primal-dual gap for the Tth iterate of EG, and show that this is tight among 1-SCLI algorithms. This is slower than the primal-dual gap of $O(1/T)$ for the average of the first T iterates of EG (Nemirovski (2004)). An interesting direction for future work is to determine if there is a provable benefit to averaging in the nonconvex-nonconcave case. Some experimental work has suggested that such a benefit to averaging exists (Yazıcı et al. (2019)); moreover, averaging is effective even for large-scale GANs (Brock et al. (2019)).

Another direction for future work is to extend the lower bound of Theorem 9 (or prove a stronger upper bound) for algorithms with decaying step-sizes, which correspond to *non-stationary* CLIs.

Such a question is only nontrivial for the case of 1-CLIs, as the averaged iterates of extragradient can be written as the iterates of a particular 2-CLI[8], and the $O(DL^2/T)$ rate of convergence of the averaged iterates of extragradient is known to be optimal (Nemirovski (2004)). Towards this question, we show in Section D.2 that, in contrast to the case for non-smooth convex minimization (Jain et al. (2019); Shamir and Zhang (2013)), any choice of decaying step-size for the EG algorithm cannot improve the $\Omega(1/\sqrt{T})$ lower bound from Theorem 9.

Acknowledgments

We are grateful to Aleksander Madry for helpful suggestions and to anonymous reviewers for helpful comments on the paper.

References

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

Yossi Arjevani and Ohad Shamir. On the Iteration Complexity of Oblivious First-Order Optimization Algorithms. *arXiv:1605.03529 [cs, math]*, May 2016. arXiv: 1605.03529.

Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On Lower and Upper Bounds for Smooth and Strongly Convex Optimization Problems. *arXiv:1503.06833 [cs, math]*, March 2015. arXiv: 1503.06833.

Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein GANs. *arXiv:1701.07875*, 2017.

Alfred Auslender and Marc Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical Programming*, 104(1):39–68, September 2005.

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A Tight and Unified Analysis of Extragradient for a Whole Spectrum of Differentiable Games. *arXiv:1906.05945 [cs, math, stat]*, June 2019. arXiv: 1906.05945.

Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. *arXiv preprint arXiv:2001.00602*, 2020.

James P. Bailey, Gauthier Gidel, and Georgios Piliouras. Finite Regret and Cycles with Fixed Step-Size via Alternating Gradient Descent-Ascent. *arXiv:1907.04392 [cs, math]*, July 2019. arXiv: 1907.04392.

A. Ben-Tal, Laurent El Ghaoui, and A. S. Nemirovskiĭ. *Robust optimization*. Princeton series in applied mathematics. Princeton University Press, Princeton, 2009.

Haïm Brezis and Moïse Sibony. Méthodes d'approximation et di'itération pour les opérateurs monotones. *Archive for Rational Mechanics and Analysis*, 28(1):59–82, 1968.

8. We refer the reader to Section D.1 for the verification of this statement.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Felix E. Browder. Nonlinear monotone operators and convex sets in Banach spaces. *Bulletin of the American Mathematical Society*, 71(5):780–785, 1965.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing Noise in GAN Training with Variance Reduced Extragradient. *arXiv:1904.08598 [cs, math, stat]*, April 2019. arXiv: 1904.08598.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent Reinforcement Learning with Nonlinear Function Approximation. *arXiv:1712.10285 [cs]*, June 2018. arXiv: 1712.10285.

Constantinos Daskalakis and Ioannis Panageas. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. *arXiv:1807.04252 [cs, math, stat]*, July 2018. arXiv: 1807.04252.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. *arXiv:1711.00141 [cs, stat]*, October 2017. arXiv: 1711.00141.

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic Variance Reduction Methods for Policy Evaluation. *arXiv:1702.07944 [cs, math, stat]*, June 2017. arXiv: 1702.07944.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer series in operations research. Springer, New York, 2003.

Rong Ge, Sham M. Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near-optimal geometrically decaying step size procedure for least squares. In *Neural Information Processing Systems*, 2019.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. *arXiv:1802.10551 [cs, math, stat]*, February 2018a. arXiv: 1802.10551.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Remi Lepriol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative Momentum for Improved Game Dynamics. *arXiv:1807.04740 [cs, stat]*, July 2018b. arXiv: 1807.04740.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

Philip Hartman and Guido Stampacchia. On some non-linear elliptic differential-functional equations. *Acta Mathematica*, 115(0):271–310, 1966.

Yu-Guan Hsieh, Franck Iutzeler, Jérome Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *arXiv:1908.08465 [cs, math]*, August 2019. arXiv: 1908.08465.

Adam Ibrahim, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear Lower Bounds and Conditioning of Differentiable Games. *arXiv:1906.07300 [cs, math, stat]*, October 2019. arXiv: 1906.07300.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the Last Iterate of SGD Information Theoretically Optimal. *arXiv:1904.12443 [cs, math]*, May 2019. arXiv: 1904.12443.

GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Peter D. Lax. *Linear Algebra and Its Applications*. Wiley-Interscience, Hoboken, NJ, second edition, 2007.

Tengyuan Liang and James Stokes. Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks. *arXiv:1802.06132 [cs, stat]*, February 2018. arXiv: 1802.06132.

J. L. Lions and G. Stampacchia. Variational inequalities. *Communications on Pure and Applied Mathematics*, 20(3):493–519, 1967.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]*, September 2019. arXiv: 1706.06083.

B. Martinet. Brève communication. Régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 4 (R3):154–158, 1970.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv:1807.02629 [cs, math, stat]*, July 2018. arXiv: 1807.02629.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The Numerics of GANs. *arXiv:1705.10461 [cs]*, May 2017. arXiv: 1705.10461.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Proximal point approximations achieving a convergence rate of $\mathcal{O}(1/k)$ for smooth convex-concave saddle point problems: Optimistic gradient and extra-gradient methods. *arXiv:1906.01115 [cs, math, stat]*, June 2019a. arXiv: 1906.01115.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. *arXiv:1901.08511 [cs, math, stat]*, January 2019b. arXiv: 1901.08511.

Renato D. C. Monteiro and Benar Fux Svaiter. On the Complexity of the Hybrid Proximal Extragradient Method for the Iterates and the Ergodic Mean. *SIAM Journal on Optimization*, 20: 2755–2787, 2010.

Arkadi 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, January 2004.

A.S Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, June 1992.

Yurii Nesterov. Cubic Regularization of Newton's Method for Convex Problems with Constraints. *SSRN Electronic Journal*, 2006.

Gerhard Opfer and Glenn Schober. Richardson's iteration for nonsymmetric matrices. *Linear Algebra and its Applications*, 58:343–361, April 1984.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, August 2019.

Wei Peng, Yuhong Dai, Hui Zhang, and Lizhi Cheng. Training GANs with Centripetal Acceleration. *arXiv:1902.08949 [cs, stat]*, February 2019. arXiv: 1902.08949.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical Notes*, 28(5):845–848, 1980.

R. Tyrrell Rockafellar. Monotone Operators and the Proximal Point Algorithm*. *SIAM Journal on Control and Optimization*, 14(5), 1976.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79, 2013.

Moïse Sibony. Méthodes itératives pour les équations et inéquations aux dérivés partielles non-linéares de type monotone. *Calcolo*, 7:65 – 183, 1970.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, June 1995.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008.

Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing Adversarial Nets With Prediction Methods. *arXiv:1705.07364 [cs]*, May 2017. arXiv: 1705.07364.

Yasin Yazıcı, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. The unusual effectiveness of averaging in GAN training. In *International Conference on Learning Representations*, 2019.

Guojun Zhang and Yaoliang Yu. Convergence of gradient methods on bilinear zero-sum games. In *International Conference on Learning Representations*, 2020.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On Lower Iteration Complexity Bounds for the Saddle Point Problems. *arXiv:1912.07481 [math]*, December 2019. arXiv: 1912.07481.

Appendix A. Proofs for Theorem 9

A.1. Proof of items 2 and 3 of Theorem 9

Proof (of items 2 and 3 of Theorem 9) We begin with item 2, namely the lower bound on the primal-dual gap. The choice of $\mathbf{M}, \mathbf{A}, \mathbf{b}_1, \mathbf{b}_2$ (which depend on $\nu \in (0, L]$) is exactly the same as for item 1, and is given in (13). Write $\mathcal{Z} := \mathcal{B}(\mathbf{x}^*, D) \times \mathcal{B}(\mathbf{y}^*, D)$. Next we compute $\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{z}^{(t)})$ in a similar manner to the Hamiltonian in (12). The components of the primal-dual gap $\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{x}, \mathbf{y})$ for a given point $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n$ are given as follows:

$$\begin{aligned}
\max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}') &= \mathbf{b}_1^\top \mathbf{x} + \max_{\mathbf{y}' : \|\mathbf{y}' - \mathbf{y}^*\| \le D} \langle \mathbf{y}', \mathbf{M}^\top \mathbf{x} + \mathbf{b}_2 \rangle \\
&= D\|\mathbf{M}^\top \mathbf{x} + \mathbf{b}_2\| + \langle \mathbf{y}^*, \mathbf{M}^\top \mathbf{x} + \mathbf{b}_2 \rangle + \mathbf{b}_1^\top \mathbf{x} \\
&= D\|\mathbf{M}^\top \mathbf{x} + \mathbf{b}_2\| + \langle -\mathbf{M}^{-1}\mathbf{b}_1, \mathbf{M}^\top \mathbf{x} + \mathbf{b}_2 \rangle + \langle \mathbf{b}_1, \mathbf{x} \rangle \\
&= D\|\mathbf{M}^\top \mathbf{x} + \mathbf{b}_2\| - \langle \mathbf{M}^{-1}\mathbf{b}_1, \mathbf{b}_2 \rangle. \\
-\min_{\mathbf{x}' \in \mathcal{X}} f(\mathbf{x}', \mathbf{y}) &= -\mathbf{b}_2^\top \mathbf{y} - \min_{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}^*\| \le D} \langle \mathbf{x}', \mathbf{M}\mathbf{y} + \mathbf{b}_1 \rangle \\
&= D\|\mathbf{M}\mathbf{y} + \mathbf{b}_1\| - \langle \mathbf{x}^*, \mathbf{M}\mathbf{y} + \mathbf{b}_1 \rangle - \mathbf{b}_2^\top \mathbf{y} \\
&= D\|\mathbf{M}\mathbf{y} + \mathbf{b}_1\| - \langle -(\mathbf{M}^\top)^{-1}\mathbf{b}_2, \mathbf{M}\mathbf{y} + \mathbf{b}_1 \rangle - \langle \mathbf{b}_2, \mathbf{y} \rangle \\
&= D\|\mathbf{M}\mathbf{y} + \mathbf{b}_1\| + \langle \mathbf{M}^{-1}\mathbf{b}_1, \mathbf{b}_2 \rangle.
\end{aligned}$$

Thus

$$\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{X}} f(\mathbf{x}', \mathbf{y}) = D\|\mathbf{M}^\top \mathbf{x} + \mathbf{b}_2\| + D\|\mathbf{M}\mathbf{y} + \mathbf{b}_1\| = D\|\mathbf{A}\mathbf{z} + \mathbf{b}\|, \tag{22}$$

and so

$$\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) = D\|\mathbf{C}_0(\mathbf{A})^t \mathbf{b}\|. \tag{23}$$

From (23) we have

$$\frac{\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})}{\|\mathbf{z}^{(0)} - \mathbf{z}^*\|^2} = \frac{D\|\mathbf{C}_0(\mathbf{A})^t \mathbf{b}\|}{\|\mathbf{A}^{-1}\mathbf{b}\|^2} = \frac{\nu\|\mathbf{C}_0(\mathbf{A})^t \mathbf{b}\|}{\|\mathbf{b}\|}. \tag{24}$$

The desired bound in item 2 of the theorem statement follows from (24), (16), and the fact that $\|\mathbf{A}^{-1}\mathbf{b}\| = D$.

Next we turn to convergence in function value (item 3 of the theorem). First note that

$$\begin{aligned}
& f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - f(\mathbf{x}^*, \mathbf{y}^*) \\
=& (\mathbf{x}^{(t)})^\top \mathbf{M}\mathbf{y}^{(t)} + \langle \mathbf{x}^{(t)}, \mathbf{b}_1 \rangle + \langle \mathbf{y}^{(t)}, \mathbf{b}_2 \rangle - (\mathbf{x}^*)^\top \mathbf{M}\mathbf{y}^* - \langle \mathbf{x}^*, \mathbf{b}_1 \rangle - \langle \mathbf{y}^*, \mathbf{b}_2 \rangle \\
=& \langle \mathbf{x}^{(t)} - \mathbf{x}^*, \mathbf{M}(\mathbf{y}^{(t)} - \mathbf{y}^*) \rangle - 2\langle \mathbf{x}^*, \mathbf{M}\mathbf{y}^* \rangle + \langle \mathbf{x}^*, \mathbf{M}\mathbf{y}^* \rangle + \langle \mathbf{y}^*, \mathbf{M}^\top \mathbf{x}^* \rangle \\
=& \langle \mathbf{x}^{(t)} - \mathbf{x}^*, \mathbf{M}(\mathbf{y}^{(t)} - \mathbf{y}^*) \rangle \\
=& \nu \cdot \sum_{i=1}^{n/2} (\mathbf{x}_i^{(t)} - \mathbf{x}_i^*) \cdot (\mathbf{y}_i^{(t)} - \mathbf{y}_i^*),
\end{aligned} \tag{25}$$

where we have used that $\mathbf{y}^* = -\mathbf{M}^{-1}\mathbf{b}_1, \mathbf{x}^* = -(\mathbf{M}^\top)^{-1}\mathbf{b}_2$.

Note that the diagonalization of $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{U}\cdot\mathrm{diag}(\nu i, \cdots, \nu i, -\nu i, \cdots -\nu i)\cdot\mathbf{U}^{-1}, \qquad \mathbf{U} = \frac{1}{\sqrt{2}}\cdot\begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ & & \vdots & & & & \vdots & \\ i & 0 & \cdots & 0 & -i & 0 & \cdots & 0 \\ 0 & i & \cdots & 0 & 0 & -i & \cdots & 0 \\ & & \vdots & & & & \vdots & \end{pmatrix}.$$

Since $\mathbf{U}$ is unitary, it follows from (11) that

$$\begin{aligned}
&\mathbf{z}^{(t)} - \mathbf{z}^* \\
=&\mathbf{C}_0(\mathbf{A})^t\mathbf{A}^{-1}\mathbf{b} \\
=&\nu^{-1}\mathbf{U}\cdot\mathrm{diag}(q_0(\nu i), \cdots, q_0(\nu i), q_0(-\nu i), \cdots, q_0(-\nu i))^t\cdot\mathrm{diag}(-i, \cdots, -i, i, \cdots, i)\cdot\mathbf{U}^{-1}\mathbf{b} \\
=&\frac{D}{\sqrt{2n}}\cdot\mathbf{U}\cdot(q_0(\nu i)^t(1-i), \ldots, q_0(\nu i)^t(1-i), q_0(-\nu i)^t(1+i), \ldots, q_0(-\nu i)^t(1+i))^\top \\
=&\frac{D}{\sqrt{n}}\cdot(\Re(q_0(\nu i)^t(1-i)), \ldots, \Re(q_0(\nu i)^t(1-i)), -\Im(q_0(\nu i)^t(1-i)), \ldots, -\Im(q_0(\nu i)^t(1-i))).
\end{aligned}$$

Now let us write $q_0(\nu i) = |q_0(\nu i)|\cdot e^{i\theta(\nu)}$, where $\theta(\nu)\in[0, 2\pi)$. It folllows from (25) that

$$\begin{aligned}
&f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - f(\mathbf{x}^*, \mathbf{y}^*) \\
&\qquad =\nu\sum_{i=1}^{n/2}(\mathbf{x}_i^{(t)} - \mathbf{x}_i^*)\cdot(\mathbf{y}_i^{(t)} - \mathbf{y}_i^*) \\
&\qquad =\nu D^2\cdot\frac{1}{2}\cdot|q_0(\nu i)|^{2t}(\cos(t\theta(\nu)) + \sin(t\theta(\nu)))\cdot(\cos(t\theta(\nu)) - \sin(t\theta(\nu))) \\
&\qquad =\nu D^2\cdot\frac{1}{2}\cdot|q_0(\nu i)|^{2t}\cdot\cos(2t\theta(\nu)) \\
&\qquad =\nu D^2\cdot\frac{1}{2}\cdot\Re(q_0(\nu i)^{2t}). \qquad (26)
\end{aligned}$$

Now fix some T. It follows in a manner identical to (16), using Lemma 13, that there is some ν_* with $\nu_*^2\in[L^2/(40Tk^2), L^2]$ so that $\nu_*^2\cdot|q_0(\nu_* i)|^{8T}\geq\frac{L^2}{80Tk^2}$, which implies $\nu_*\cdot|q_0(\nu_* i)|^{4T}\geq\frac{L}{\sqrt{80T}k}$. We claim that also $\nu_*\cdot|q_0(\nu_* i)|^{2T}\geq\frac{L}{\sqrt{80T}k}$. If $|q_0(\nu_* i)|\geq 1$, this is immediate from $\nu_*\geq L/(\sqrt{40T}k)$; otherwise, this follows from $|q_0(\nu_* i)|^{2T}\geq|q_0(\nu_* i)|^{4T}$. To complete the proof we consider two cases:

Case 1. If $|\Re(q_0(\nu_* i)^{2T})|\geq\frac{1}{2}\cdot|q_0(\nu_* i)^{2T}|$, then by (26) $|f(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) - f(\mathbf{x}^*, \mathbf{y}^*)|\geq\frac{LD^2}{\sqrt{1280T}k}$ (where f is so that ν in (13) is set to ν_*), and we get that $\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\mathrm{Func}}; T)\geq\frac{LD^2}{\sqrt{1280T}k}\geq\frac{LD^2}{36k\sqrt{T}}$.

Case 2. In the other case that $|\Re(q_0(\nu_* i)^{2T})|\leq\frac{1}{2}\cdot|q_0(\nu_* i)^{2T}|$, we have $2T\theta(\nu_*)\in[\pi/3, 2\pi/3]\cup[-2\pi/3, -\pi/3]$. Hence $4T\theta(\nu_*)\in[2\pi/3, 4\pi/3]$, and so $|\Re(q_0(\nu_* i)^{4T})|\geq\frac{1}{2}\cdot|q_0(\nu_* i)^{4T}|$. By (26) and the fact that $\nu_*\cdot|q_0(\nu i)|^{4T}\geq\frac{L}{\sqrt{80T}k}$, it follows that in this case we have $\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\mathrm{Func}}; 2T)\geq\frac{LD^2}{\sqrt{1280T}k}$. ■

A.2. Supplementary lemmas for Theorem 9

Lemma 13 below is similar to the bounds derived in (Nemirovsky, 1992, Section 2.3.B), but it achieves a better dependence on t; in particular, if the bounds in (Nemirovsky, 1992, Section 2.3.B) are used in a black-box manner, one would instead get a lower bound of $\Omega(L/t^2k^2)$ in (27).

Lemma 13 *Fix some $k, t \in \mathbb{N}$, $L > 0$. Let $r(y) \in \mathbb{R}[y]$ be a polynomial with real-valued coefficients of degree at most k, such that $r(0) = 1$. Then*

$$\sup_{y\in(0,L]} y \cdot |r(y)|^t \geq \sup_{y\in[L/(20tk^2),L]} y \cdot |r(y)|^t > \frac{L}{40tk^2}. \tag{27}$$

Proof Set $\mu := \frac{L}{20tk^2}$. Then $\sqrt{L/\mu} - 1 = \sqrt{20tk^2} - 1 \geq \sqrt{12t} \cdot k$. By Lemma 14 we have that

$$\sup_{y\in[\mu,L]} y \cdot |r(y)|^t \geq \frac{L}{20tk^2} \cdot \left(1 - \frac{6k^2}{(\sqrt{L/\mu} - 1)^2}\right)^t \geq \frac{L}{20tk^2} \cdot (1 - 1/(2t))^t \geq \frac{L}{40tk^2}.$$

■

Lemma 14 *Fix some $k \in \mathbb{N}$ and $L > \mu > 0$ such that $k \leq \sqrt{L/\mu} - 1$. Let $r(y) \in \mathbb{R}[y]$ be a polynomial with real-valued coefficients of degree at most k, such that $r(0) = 1$. Then*

$$\sup_{y\in[\mu,L]} |r(y)| > 1 - \frac{6k^2}{(\sqrt{L/\mu} - 1)^2}. \tag{28}$$

Lemma 14 is very similar to the combination of Lemmas 5 and 12 in Azizian et al. (2019), but has a superior dependence on k. In particular, we could use (Azizian et al., 2019, Lemmas 5 & 12) to conclude that a lower bound of $1 - k^3 \cdot \frac{4\mu}{L\pi}$ holds in (28), which is smaller than $1 - \frac{6k^2}{(\sqrt{L/\mu}-1)^2}$ for sufficiently large k (e.g., $k > 10$). We also remark that the proof of Lemma 14 is much simpler than that of (Azizian et al., 2019, Lemmas 5 & 12), though the proofs use similar techniques.

Proof (of Lemma 14). Let $T_k(y)$ denote the Chebyshev polynomial of the first kind of degree k; it is characterised by the property that:

$$T_k\left(\cos\left(\frac{j\pi}{k}\right)\right) = (-1)^j, \qquad j = 0, 1, \ldots, k, \tag{29}$$

which turns out to be equivalent to the property that

$$T_k\left(\frac{1}{2} \cdot \left(z + \frac{1}{z}\right)\right) = \frac{1}{2} \cdot \left(z^k + \frac{1}{z^k}\right), \qquad \forall z \in \mathbb{C}. \tag{30}$$

It follows immediately from (30), that for k odd, T_k is an odd function, and for k even, T_k is an even function.

Let $q(y) = \frac{T_k\left(\frac{2y-(\mu+L)}{L-\mu}\right)}{T_k\left(\frac{L+\mu}{L-\mu}\right)}$. Then $q(0) = 1$. Using (29) and the fact that $r(0) = q(0) = 1$, it was shown in (Arjevani and Shamir, 2016, Lemma 2) that

$$\sup_{y\in[\mu,L]} |r(y)| \geq \sup_{y\in[\mu,L]} |q(y)|.$$

Let $\kappa = L/\mu$. From (29) we have that $\sup_{y\in[\mu,L]} |q(y)| \geq \frac{1}{T_k\left(\frac{L+\mu}{L-\mu}\right)} = \frac{1}{T_k\left(\frac{\kappa+1}{\kappa-1}\right)}$ (in fact, equality holds). At this we depart from the proof of (Arjevani and Shamir, 2016, Lemma 2), noting simply that a tighter lower bound on $\frac{1}{T_k\left(\frac{\kappa+1}{\kappa-1}\right)}$ than the one shown in (Arjevani and Shamir, 2016, Lemma 2) holds when $k^2 \ll \kappa$. In particular, since $\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 2 \cdot \frac{\kappa+1}{\kappa-1}$, (30) gives that

$$\begin{aligned}
T_k\left(\frac{\kappa+1}{\kappa-1}\right) =& \frac{1}{2} \cdot \left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right) \\
<& \frac{1}{2} \cdot \left(\left(1 + \frac{2k}{\sqrt{\kappa}-1} + \frac{(2k)^2}{(\sqrt{\kappa}-1)^2}\right) + \left(1 - \frac{2k}{\sqrt{\kappa}+1} + \frac{(2k)^2}{(\sqrt{\kappa}+1)^2}\right)\right) \qquad (31) \\
\leq& \frac{1}{2} \cdot \left(2 + \frac{4k}{\kappa-1} + \frac{8k^2}{(\sqrt{\kappa}-1)^2}\right) \\
\leq& 1 + \frac{2k+4k^2}{(\sqrt{\kappa}-1)^2}. \qquad (32)
\end{aligned}$$

Above (31) follows from the fact $k \leq \sqrt{\kappa} - 1$ and that for $-2 \leq yk \leq 2$, we have that

$$(1+y)^k \leq \exp(yk) \leq 1 + yk + 2y^2k^2.$$

From (32) it follows that

$$\frac{1}{T_k\left(\frac{\kappa+1}{\kappa-1}\right)} > 1 - \frac{2k+4k^2}{(\sqrt{\kappa}-1)^2} \geq 1 - \frac{6k^2}{(\sqrt{\kappa}-1)^2}.$$

■

A.3. Tightness of dependence on the degree k

The below proposition establishes that the inverse linear dependence on k in Theorem 9 is tight:

Proposition 15 *Then there is a consistent 1-SCLI $\mathcal{A}$ whose inversion matrix $\mathbf{N}(\cdot)$ is a polynomial of degree at most $k-1$ so that*

$$\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\mathrm{Ham}}; 1) \leq O\left(\frac{L^2D^2}{k^2}\right), \ \max\left\{\mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\mathrm{Gap}}; 1), \mathrm{IC}_{n,L,D}(\mathcal{A}, \mathcal{L}^{\mathrm{Func}}; 1)\right\} \leq O\left(\frac{LD^2}{k}\right). \qquad (33)$$

Iteration complexities are defined with respect to $\mathcal{F}^{\mathrm{bil}}_{n,L,D}$ as in Definition 6.

Remark. We note that the upper bounds in Proposition 15 hold more generally with respect to any monotone linear operator $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$. We stick with the class $\mathcal{F}^{\mathrm{bil}}_{n,L,D}$ of operators corresponding to a bilinear function f (as in Definition 6) to simplify notation.

Proof (of Proposition 15) Consider a monotone operator $F = F_f \in \mathcal{F}^{\mathrm{bil}}_{n,L,D}$ of the form $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$. For $t \geq 0$, let $\mathbf{w}^{(t)}$ be the iterates obtained by running extragradient on the monotone operator F starting at $\mathbf{w}^{(0)} = \mathbf{0}$ and with step size $\eta = 1/(2L)$. Letting

$$\mathbf{C}_0(\mathbf{A}) := I - \eta\mathbf{A} + (\eta\mathbf{A})^2, \qquad \mathbf{N}(\mathbf{A}) := -\eta(I - \eta\mathbf{A})\mathbf{b},$$

by (8), we can write

$$\mathbf{w}^{(t)} = (\mathbf{C}_0(\mathbf{A})^t + \mathbf{C}_0(\mathbf{A})^{t-1} + \cdots + \mathbf{C}_0(\mathbf{A}) + I) \cdot \mathbf{N}(\mathbf{A})\mathbf{b}.$$

For any $T > 0$, denote the averaged iterates up to time T by $\bar{\mathbf{w}}^{(T)} := \frac{\mathbf{w}^{(0)}+\cdots+\mathbf{w}^{(T)}}{T+1}$. Also write $\bar{\mathbf{w}}^{(T)} = (\hat{\mathbf{x}}^{(T)}, \hat{\mathbf{y}}^{(T)})$. By (Mokhtari et al., 2019a, Theorem 3), we have, with $\mathcal{X} := \mathcal{B}(\mathbf{x}^*, D), \mathcal{Y} := \mathcal{B}(\mathbf{y}^*, D)$,

$$\mathcal{L}_F^{\text{Gap}}(\bar{\mathbf{w}}^{(T)}) = \max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}^{(T)}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}^{(T)}) \leq O\left(\frac{D^2 L}{T}\right).$$

It follows immediately from the fact that f is convex-concave that also $\mathcal{L}^{\text{Func}}(\bar{\mathbf{w}}^{(T)}) \leq O\left(\frac{D^2 L}{T}\right)$. By (22) it follows that $\mathcal{L}_F^{\text{Ham}}(\bar{\mathbf{w}}^{(T)}) = \|\mathbf{A}\bar{\mathbf{w}}^{(T)} + \mathbf{b}\|^2 \leq O\left(\frac{D^2 L^2}{T^2}\right)$.

Next fix $T = \lfloor \frac{k-1}{2} \rfloor$, and define the polynomial

$$\mathbf{N}'(\mathbf{A}) := \frac{(\mathbf{C}_0(\mathbf{A})^T + 2\mathbf{C}_0(\mathbf{A})^{T-1} + \cdots + (T+1)I) \cdot \mathbf{N}(\mathbf{A})}{T+1},$$

so that $\mathbf{N}'(\mathbf{A})$ is a polynomial in $\mathbf{A}$ with real-valued coefficients of degree at most $2T + 1 \leq k - 1$. Also let $\mathbf{C}_0'(\mathbf{A}) := I + \mathbf{N}'(\mathbf{A}) \cdot \mathbf{A}$. It is immediate from the definition of $\mathbf{N}'(\cdot)$ that $\mathbf{w}^{(T)} = \mathbf{N}'(\mathbf{A}) \cdot \mathbf{b}$, and so $\mathbf{w}^{(T)}$ is the first iterate (namely, $\mathbf{z}^{(1)}$, as in (7)) in the consistent 1-SCLI defined by $\mathbf{C}_0'(\cdot), \mathbf{N}'(\cdot)$.

■

Appendix B. Proofs of Lemmas 11 and 12

B.1. Preliminary lemmas

Before proving Lemmas 11 and 12 we state a few simple lemmas.

Lemma 16 *(Nesterov (2006)) If $\mathcal{Z} \subset \mathbb{R}^n$ and $F : \mathcal{Z} \to \mathbb{R}^n$ is monotone, then for any $\mathbf{z}, \mathbf{w} \in \mathbb{R}^n$, $\mathbf{z}^\top(\partial F(\mathbf{w}))\mathbf{z} \geq 0$. Equivalently, $\partial F(\mathbf{w}) + \partial F(\mathbf{w})^\top$ is PSD.*

Lemma 17 *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$ be any square matrices. Then*

$$\mathbf{X}\mathbf{X}^\top \preceq 2\mathbf{Y}\mathbf{Y}^\top + 2\|\mathbf{X} - \mathbf{Y}\|_\sigma^2 \cdot I. \tag{34}$$

Proof (of Lemma 17) Note that for any real numbers x, y we have that $x^2 = (y + (x - y))^2 \leq 2y^2 + 2(x - y)^2$. It follows that for any vector $\mathbf{v} \in \mathbb{R}^n$,

$$\begin{aligned}\|\mathbf{X}^\top\mathbf{v}\|^2 &\leq 2\|\mathbf{Y}^\top\mathbf{v}\|^2 + 2\|\mathbf{X}^\top\mathbf{v} - \mathbf{Y}^\top\mathbf{v}\|^2 \\ &\leq 2\|\mathbf{Y}^\top\mathbf{v}\|^2 + 2\|\mathbf{X}^\top - \mathbf{Y}^\top\|_\sigma^2\|\mathbf{v}\|^2,\end{aligned}$$

which establishes (34). ■

Lemma 18 *Let $\mathbf{S}, \mathbf{R} \in \mathbb{R}^{n\times n}$ be (symmetric) PSD matrices. Then*

$$\mathbf{S}\mathbf{R} + \mathbf{R}\mathbf{S} \preceq 4\mathbf{S}^2 + 4\|\mathbf{S} - \mathbf{R}\|_\sigma^2 \cdot I. \tag{35}$$

Proof (of Lemma 18) Note that for any real numbers r, s we have that $rs \leq 2s^2 + 2(r-s)^2$. It follows that for any $\mathbf{v} \in \mathbb{R}^n$,

$$2\langle \mathbf{R}\mathbf{v}, \mathbf{S}\mathbf{v}\rangle \leq 4\|\mathbf{S}\mathbf{v}\|^2 + 4\|\mathbf{R}\mathbf{v} - \mathbf{S}\mathbf{v}\|^2 \leq 4\|\mathbf{S}\mathbf{v}\|^2 + 4\|\mathbf{S} - \mathbf{R}\|_\sigma^2 \|\mathbf{v}\|^2.$$

■

B.2. Proof of Lemma 11

Proof (of Lemma 11). Since F is continuously differentiable, by the fundamental theorem of calculus, for all $\mathbf{z}$,

$$F(\mathbf{z} - \eta F(\mathbf{z})) = F(\mathbf{z}) - \int_0^1 \partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z})) \cdot \eta F(\mathbf{z}) d\alpha,$$

so if we set

$$\mathbf{B}_\mathbf{z} = \int_0^1 \partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z})) d\alpha,$$

then we have $F(\mathbf{z} - \eta F(\mathbf{z})) = F(\mathbf{z}) - \eta \mathbf{B}_\mathbf{z} F(\mathbf{z})$. Again using the fundamental theorem of calculus,

$$F(\mathbf{z} - \eta F(\mathbf{z} - \eta F(\mathbf{z}))) = F(\mathbf{z}) - \eta \int_0^1 \partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z} - \eta F(\mathbf{z})))F(\mathbf{z} - \eta F(\mathbf{z})) d\alpha.$$

Then if we set

$$\mathbf{A}_\mathbf{z} = \int_0^1 \partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z} - \eta F(\mathbf{z}))) d\alpha,$$

then

$$\begin{aligned} F(\mathbf{z} - \eta F(\mathbf{z} - \eta F(\mathbf{z}))) =& F(\mathbf{z}) - \eta \mathbf{A}_\mathbf{z} F(\mathbf{z} - \eta F(\mathbf{z})) \\ =& F(\mathbf{z}) - \eta \mathbf{A}_\mathbf{z}(F(\mathbf{z}) - \eta \mathbf{B}_\mathbf{z} F(\mathbf{z})) \\ =& F(\mathbf{z}) - \eta \mathbf{A}_\mathbf{z} F(\mathbf{z}) + \eta^2 \mathbf{A}_\mathbf{z}\mathbf{B}_\mathbf{z} F(\mathbf{z}). \end{aligned}$$

Note that $\mathbf{A}_\mathbf{z}, \mathbf{B}_\mathbf{z}$ have spectral norms at most L and $\mathbf{A}_\mathbf{z} + \mathbf{A}_\mathbf{z}^\top, \mathbf{B}_\mathbf{z} + \mathbf{B}_\mathbf{z}^\top$ are PSD since the same is true of the matrices $\partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z} - \eta F(\mathbf{z})))$ and $\partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z}))$ (here we are using Lemma 16). Finally, since F has a Λ-smooth Jacobian, we have that

$$\begin{aligned} \|\mathbf{A}_\mathbf{z} - \mathbf{B}_\mathbf{z}\|_\sigma \leq& \int_0^1 \|\partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z})) - F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z} - \eta F(\mathbf{z})))\|_\sigma d\alpha \\ \leq& \int_0^1 (1-\alpha)\eta\Lambda \|F(\mathbf{z}) - F(\mathbf{z} - \eta F(\mathbf{z}))\| d\alpha \\ =& \frac{\eta\Lambda}{2}\|F(\mathbf{z}) - F(\mathbf{z} - \eta F(\mathbf{z}))\|. \end{aligned}$$

■

B.3. Proof of Lemma 12

Proof (of Lemma 12). Set $L_0 = \max\{\|\mathbf{A}\|_\sigma, \|\mathbf{B}\|_\sigma\}$. We wish to show that

$$(I - \mathbf{A} + \mathbf{AB})(I - \mathbf{A} + \mathbf{AB})^\top \preceq I \cdot (1 + 26\|\mathbf{A} - \mathbf{B}\|_\sigma^2),$$

or equivalently that

$$(\mathbf{A} + \mathbf{A}^\top) - (\mathbf{AB} + \mathbf{B}^\top\mathbf{A}^\top) - \mathbf{AA}^\top + (\mathbf{ABA}^\top + \mathbf{AB}^\top\mathbf{A}^\top) - \mathbf{ABB}^\top\mathbf{A}^\top \succeq -26\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I.$$

Notice that $\mathbf{ABA}^\top + \mathbf{AB}^\top\mathbf{A}^\top \succeq 0$ since for any vector $\mathbf{v} \in \mathbb{R}^n$, we have $\mathbf{v}^\top\mathbf{A}(\mathbf{B}+\mathbf{B}^\top)\mathbf{A}^\top\mathbf{v} \geq 0$ as $\mathbf{B} + \mathbf{B}^\top \succeq 0$. Moreover, since $\mathbf{BB}^\top \preceq L_0^2 \cdot I$, we have that for any $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}^\top\mathbf{ABB}^\top\mathbf{A}^\top\mathbf{v} \leq L_0^2 \cdot \mathbf{v}^\top\mathbf{AA}^\top\mathbf{v}$, and so $\mathbf{ABB}^\top\mathbf{A}^\top \preceq L_0^2 \cdot \mathbf{AA}^\top$. Thus it suffices to show

$$(\mathbf{A} + \mathbf{A}^\top) - (\mathbf{AB} + \mathbf{B}^\top\mathbf{A}^\top) - (1 + L_0^2) \cdot \mathbf{AA}^\top \succeq -26\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I. \tag{36}$$

Next write $\mathbf{M} := (\mathbf{A} - \mathbf{A}^\top)/2, \mathbf{S} := (\mathbf{A} + \mathbf{A}^\top)/2, \mathbf{N} := (\mathbf{B} - \mathbf{B}^\top)/2, \mathbf{R} := (\mathbf{B} + \mathbf{B}^\top)/2$. Then $\mathbf{R}, \mathbf{S}$ are positive semi-definite and $\mathbf{M}, \mathbf{N}$ are anti-symmetric (i.e., $\mathbf{M}^\top = -\mathbf{M}, \mathbf{N}^\top = -\mathbf{N}$). Also note that $\|\mathbf{R} - \mathbf{S}\|_\sigma \leq \|\mathbf{A} - \mathbf{B}\|_\sigma$ and $\|\mathbf{M} - \mathbf{N}\|_\sigma \leq \|\mathbf{A} - \mathbf{B}\|_\sigma$. Then we have:

$$\begin{aligned}
\mathbf{AA}^\top &= (\mathbf{M} + \mathbf{S})(\mathbf{M}^\top + \mathbf{S}^\top) = \mathbf{MM}^\top + \mathbf{MS} + \mathbf{SM}^\top + \mathbf{SS} \\
\mathbf{AB} &= (\mathbf{M} + \mathbf{S})(\mathbf{N} + \mathbf{R}) = \mathbf{MN} + \mathbf{MR} + \mathbf{SN} + \mathbf{SR} \\
&= -\mathbf{MN}^\top + \mathbf{MR} - \mathbf{SN}^\top + \mathbf{SR} \\
\mathbf{B}^\top\mathbf{A}^\top &= (\mathbf{N}^\top + \mathbf{R}^\top)(\mathbf{M}^\top + \mathbf{S}^\top) = \mathbf{N}^\top\mathbf{M}^\top + \mathbf{N}^\top\mathbf{S} + \mathbf{RM}^\top + \mathbf{RS} \\
&= -\mathbf{NM}^\top - \mathbf{NS} + \mathbf{RM}^\top + \mathbf{RS}.
\end{aligned}$$

Next, note that for any vector $\mathbf{v} \in \mathbb{R}^n$ and any real number $\epsilon > 0$, we have

$$\begin{aligned}
\langle \mathbf{v}, (\mathbf{MS} + \mathbf{SM}^\top)\mathbf{v} \rangle &= 2\langle \mathbf{Sv}, \mathbf{M}^\top\mathbf{v} \rangle \\
&= 2\sum_{j=1}^n (\mathbf{Sv})_j \cdot (\mathbf{M}^\top\mathbf{v})_j \\
\text{(Young's inequality)} \quad &\leq 2\sum_{j=1}^n \frac{\epsilon \cdot (\mathbf{M}^\top\mathbf{v})_j^2}{2} + \frac{(\mathbf{Sv})_j^2}{2\epsilon} \\
&= \epsilon \cdot \|\mathbf{M}^\top\mathbf{v}\|_2^2 + \frac{\|\mathbf{Sv}\|_2^2}{\epsilon}.
\end{aligned}$$

Thus $\mathbf{MS} + \mathbf{SM}^\top \preceq \epsilon \cdot \mathbf{MM}^\top + \frac{\mathbf{S}^2}{\epsilon}$. Replacing $\mathbf{M}$ with $-\mathbf{N}$ gives that for all $\epsilon > 0$, $-\mathbf{NS} - \mathbf{SN}^\top \preceq \epsilon \cdot \mathbf{NN}^\top + \frac{\mathbf{S}^2}{\epsilon}$, and replacing $\mathbf{S}$ with $\mathbf{R}$ gives that for all $\epsilon > 0$, $\mathbf{MR} + \mathbf{RM}^\top \preceq \epsilon \cdot \mathbf{MM}^\top + \frac{\mathbf{R}^2}{\epsilon}$. Hence

$$\begin{aligned}
&(1 + L_0^2) \cdot \mathbf{AA}^\top + \mathbf{AB} + \mathbf{B}^\top\mathbf{A}^\top \\
&\quad \preceq (1 + L_0^2) \cdot \left((1 + \epsilon)\mathbf{MM}^\top + (1 + \frac{1}{\epsilon})\mathbf{S}^2 \right) - \mathbf{MN}^\top - \mathbf{NM}^\top + \mathbf{SR} + \mathbf{RS} \\
&\qquad + \epsilon \cdot \mathbf{NN}^\top + \frac{\mathbf{S}^2}{\epsilon} + \epsilon \cdot \mathbf{MM}^\top + \frac{\mathbf{R}^2}{\epsilon}.
\end{aligned}$$

By Lemma 17 with $\mathbf{X} = \mathbf{R}, \mathbf{Y} = \mathbf{S}$ and Lemma 18, we have that

$$\begin{aligned}\mathbf{SR} + \mathbf{RS} + \epsilon \cdot \mathbf{NN}^\top + \frac{\mathbf{S}^2}{\epsilon} + \epsilon \cdot \mathbf{MM}^\top + \frac{\mathbf{R}^2}{\epsilon} \\ \preceq \epsilon \cdot (\mathbf{NN}^\top + \mathbf{MM}^\top) + \left(4 + \frac{1}{\epsilon} + \frac{2}{\epsilon}\right) \cdot \mathbf{S}^2 + \left(4 + \frac{2}{\epsilon}\right) \cdot \|\mathbf{R} - \mathbf{S}\|_\sigma^2 \cdot I,\end{aligned}$$

so

$$\begin{aligned}(1 + L_0^2) \cdot \mathbf{AA}^\top + \mathbf{AB} + \mathbf{B}^\top \mathbf{A}^\top \\ \preceq \epsilon \cdot \mathbf{NN}^\top + ((1 + L_0^2)(1 + \epsilon) + \epsilon)\mathbf{MM}^\top - \mathbf{NM}^\top - \mathbf{MN}^\top \\ + \left((1 + L_0^2)\left(1 + \frac{1}{\epsilon}\right) + 4 + \frac{3}{\epsilon}\right)\mathbf{S}^2 + \left(4 + \frac{2}{\epsilon}\right) \cdot \|\mathbf{A} - \mathbf{B}\|_\sigma^2 \cdot I.\end{aligned} \tag{37}$$

Next, note that as long as $5\epsilon + 2L_0^2 + 2\epsilon L_0^2 \leq 1$, we have that

$$\begin{aligned}\epsilon \cdot \mathbf{NN}^\top + ((1 + L_0^2)(1 + \epsilon) + \epsilon)\mathbf{MM}^\top - \mathbf{NM}^\top - \mathbf{MN}^\top \\ \preceq^{(*1)} (\epsilon + 2 \cdot (2\epsilon + L_0^2 + \epsilon L_0^2)) \cdot \mathbf{NN}^\top + \mathbf{MM}^\top - \mathbf{NM}^\top - \mathbf{MN}^\top \\ + 2 \cdot (2\epsilon + L_0^2 + \epsilon L_0^2) \cdot \|\mathbf{M} - \mathbf{N}\|_\sigma^2 I \\ \preceq \mathbf{NN}^\top + \mathbf{MM}^\top - \mathbf{NM}^\top - \mathbf{MN}^\top + (1 - \epsilon)\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I \\ = (\mathbf{N} - \mathbf{M})(\mathbf{N}^\top - \mathbf{M}^\top) + (1 - \epsilon)\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I \\ \preceq \|\mathbf{N} - \mathbf{M}\|_\sigma^2 I + (1 - \epsilon)\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I \\ \preceq (2 - \epsilon)\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I.\end{aligned} \tag{38}$$

where $(^*1)$ follows from Lemma 17 with $\mathbf{X} = \mathbf{M}, \mathbf{Y} = \mathbf{N}$. Moreover, as long as

$$L_0\left((1 + L_0^2)(1 + (1/\epsilon)) + 4 + (3/\epsilon)\right) \leq 2,$$

since $\|\mathbf{S}\|_\sigma \leq \|\mathbf{A}\|_\sigma \leq L_0$, we have that

$$\left((1 + L_0^2)\left(1 + \frac{1}{\epsilon}\right) + 4 + \frac{3}{\epsilon}\right)\mathbf{S}^2 \preceq 2\mathbf{S} = \mathbf{A} + \mathbf{A}^\top. \tag{39}$$

Combining (37), (38), and (39) gives that

$$(1 + L_0^2) \cdot \mathbf{AA}^\top + \mathbf{AB} + \mathbf{B}^\top \mathbf{A}^\top \preceq \left(6 + \frac{2}{\epsilon}\right)\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I + \mathbf{A} + \mathbf{A}^\top,$$

which is equivalent to (36) as long as $\epsilon = 1/10$.

Finally, note that as long as $L_0 \leq 1/30$, the choice $\epsilon = 1/10$ satisfies $5\epsilon + 2L_0^2 + 2\epsilon L_0^2 \leq 1$ and $L_0((1 + L_0^2)(1 + 1/\epsilon) + 4 + 3/\epsilon) \leq 2$, completing the proof.

■

Appendix C. Proof of Last Iterate convergence of Proximal Point

We first prove the following lemma which shows that the Hamiltonian decreases each iteration of the proximal point algorithm:

Lemma 19 *Suppose that $F : \mathbb{R}^D \to \mathbb{R}^D$ is a monotone operator. Then*

$$\|F(\mathbf{x})\|^2 \leq \|F(\mathbf{x} + \eta F(\mathbf{x}))\|^2.$$

Proof By monotonicity of F we have that, for $\eta > 0$, $\langle F(\mathbf{x}), F(\mathbf{x} + \eta F(\mathbf{x})) - F(\mathbf{x})\rangle \geq 0$. Now note that

$$\begin{aligned}
&\|F(\mathbf{x} + \eta F(\mathbf{x}))\|^2 - \|F(\mathbf{x})\|^2 \\
&= 2\langle F(\mathbf{x}), F(\mathbf{x} + \eta F(\mathbf{x})) - F(\mathbf{x})\rangle + \|F(\mathbf{x} + \eta F(\mathbf{x})) - F(\mathbf{x})\|^2 \\
&\geq 0.
\end{aligned}$$

■

Theorem 20 gives an analogue of Theorem 10 for the PP algorithm. Given Lemma 19, its proof is essentially immediate given prior results in the literature (see, e.g., Mokhtari et al. (2019a); Monteiro and Svaiter (2010)), but we reproduce the entire proof for completeness.

Theorem 20 *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone operator. Fix some $\mathbf{z}^{(0)} \in \mathbb{R}^n$, and suppose there is $\mathbf{z}^* \in \mathbb{R}^n$ so that $F(\mathbf{z}^*) = 0$ and $\|\mathbf{z}^* - \mathbf{z}^{(0)}\| \leq D$. If the proximal point algorithm with any step size $\eta > 0$ is initialized at $\mathbf{z}^{(0)}$, then its iterates $\mathbf{z}^{(T)}$ satisfy*

$$\|F(\mathbf{z}^{(t)})\| \leq \frac{D}{\eta\sqrt{T}}.$$

If moreover $\mathcal{Z} = \mathcal{B}(\mathbf{x}^*, D) \times \mathcal{B}(\mathbf{y}^*, D)$ and $F(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{pmatrix}$ for a convex-concave function f, then it follows that

$$\mathrm{Gap}_f^{\mathcal{Z}}(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) = \max_{\mathbf{y}' \in \mathcal{B}(\mathbf{y}^*, D)} f(\mathbf{x}^{(T)}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{B}(\mathbf{x}^*, D)} f(\mathbf{x}', \mathbf{y}^{(T)}) \leq \frac{\sqrt{2}D^2}{\eta\sqrt{T}}.$$

(Here $\mathbf{z}^{(T)} = (\mathbf{x}^{(T)}, \mathbf{y}^{(T)})$.)

Proof Recall that the iterates of the proximal point algorithm are defined by

$$\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t+1)}).$$

It is easy to see that the following equality holds at all iterations of the proximal point algorithm: for all $\mathbf{z} \in \mathbb{R}^D$,

$$\langle F(\mathbf{z}^{(t+1)}), \mathbf{z}^{(t+1)} - \mathbf{z}\rangle = \frac{1}{2\eta}\left(\|\mathbf{z}^{(t)} - \mathbf{z}\|^2 - \|\mathbf{z}^{(t+1)} - \mathbf{z}\|^2 - \|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1)}\|^2\right).$$

Setting $\mathbf{z} = \mathbf{z}^*$, so that $\langle F(\mathbf{z}'), \mathbf{z}' - \mathbf{z}^*\rangle \geq 0$ for all $\mathbf{z}'$, it follows that for any $T > 0$,

$$\sum_{t=0}^{T-1} \frac{\eta}{2}\|F(\mathbf{z}^{(t+1)})\|^2 \leq \sum_{t=0}^{T-1} \frac{1}{2\eta}\left(\|\mathbf{z}^{(t)} - \mathbf{z}\|^2 - \|\mathbf{z}^{(t+1)} - \mathbf{z}\|^2\right) \leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}\|^2 \leq \frac{1}{2\eta}D^2.$$

(The last inequality follows since $\mathbf{z}, \mathbf{z}^* \in \mathcal{Z}$, and the diameter of $\mathcal{Z}$ is at most D.) Thus, there exists some $t^* \in \{1, 2, \ldots, T\}$ so that

$$\|F(\mathbf{z}^{(t^*)})\|^2 \leq \frac{D^2}{\eta^2 T}. \tag{40}$$

Next, Lemma 19 with $\mathbf{x} = \mathbf{z}^{(t+1)}$ gives that for each $t \geq 0$,

$$\|F(\mathbf{z}^{(t+1)})\|^2 \leq \|F(\mathbf{z}^{(t+1)} + \eta \mathbf{z}^{(t+1)})\|^2 \leq \|F(\mathbf{z}^{(t)})\|^2.$$

Thus

$$\|F(\mathbf{z}^{(T)})\| \leq \frac{D}{\eta\sqrt{T}}.$$

If $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $F(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{pmatrix}$, for some convex-concave function f, then

$$\begin{aligned}
&\max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}^{(T)}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{X}} f(\mathbf{x}', \mathbf{y}^{(T)}) \\
&\qquad = \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}^{(T)}, \mathbf{y}') - f(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) - \min_{\mathbf{x}' \in \mathcal{X}} (f(\mathbf{x}', \mathbf{y}^{(T)}) - f(\mathbf{x}^{(T)}, \mathbf{y}^{(T)})) \\
&\qquad \leq \max_{\mathbf{y}' \in \mathcal{Y}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}), \mathbf{y}' - \mathbf{y}^{(T)} \rangle + \max_{\mathbf{x}' \in \mathcal{X}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}^{(T)}, \mathbf{y}^{(T)}), \mathbf{x}^{(T)} - \mathbf{x}' \rangle \\
&\qquad = \max_{z' \in \mathcal{Z}} \langle F(\mathbf{z}^{(T)}), \mathbf{z}^{(T)} - \mathbf{z}' \rangle \\
&\qquad \leq \|F(\mathbf{z}^{(T)})\| \cdot D\sqrt{2} \\
&\qquad \leq \frac{\sqrt{2} D^2}{\eta\sqrt{T}}.
\end{aligned}$$

■

Appendix D. CLIs with time-varying coefficients

D.1. Averaged EG iterates are a 2-CLI with time-varying coefficients

In this section we show that the averaged iterates of extra-gradient can be written as the iterates of a particular 2-CLI[9] with time-varying coefficients. Let $\mathbf{z}^{(t)}$, $t \geq 0$ be the iterates of extragradient and $\mathbf{v}^{(t)} := \frac{\sum_{t'=0}^{t} \mathbf{z}^{(t)}}{t+1}$. Then $\mathbf{z}^{(t)} = (t+1)\mathbf{v}^{(t)} - t\mathbf{v}^{(t-1)}$, and so

$$\begin{aligned}
(t+2)\mathbf{v}^{(t+1)} &= (t+1)\mathbf{v}^{(t)} + \mathbf{z}^{(t+1)} \\
&= (t+1) \cdot \mathbf{v}^{(t)} + \mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)} - \eta F(\mathbf{z}^{(t)})) \\
&= 2(t+1)\mathbf{v}^{(t)} - t\mathbf{v}^{(t-1)} - \eta F((t+1)\mathbf{v}^{(t)} - t\mathbf{v}^{(t-1)} - \eta F((t+1)\mathbf{v}^{(t)} - t\mathbf{v}^{(t-1)})).
\end{aligned}$$

Writing $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$ (i.e., restricting to the setting in which CLIs are defined) gives

$$\mathbf{v}^{(t+1)} = \left(2I - \eta\mathbf{A} + (\eta\mathbf{A})^2\right) \cdot \frac{t+1}{t+2}\mathbf{v}^{(t)} - \left(I - \eta\mathbf{A} + (\eta\mathbf{A})^2\right) \frac{t}{t+2}\mathbf{v}^{(t-1)} + \eta(-I + \eta\mathbf{A})\mathbf{b},$$

which is a 2-CLI.

9. See (Arjevani and Shamir, 2016, Definition 2) for a definition of 2-CLIs.

D.2. Lower bound for last iterate of EG with time-varying coefficients

In this section we consider the iterates of the EG algorithm with time-varying coefficients, defined as follows:

$$\mathbf{z}^{(t+1/2)} = \mathbf{z}^{(t)} - \eta_t F(\mathbf{z}^{(t)}) \qquad \mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - \eta_t F(\mathbf{z}^{(t+1/2)}), \tag{41}$$

where $\eta_t > 0$ is a sequence of time-varying coefficients.

As in Theorem 9, for given n, L, D and $F = F_f \in \mathcal{F}^{\text{bil}}_{n,L,D}$ with Nash equilibrium $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$, set $\mathcal{Z} = \mathcal{B}(\mathbf{x}^*, D) \times \mathcal{B}(\mathbf{y}^*, D)$, so that for $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, $\text{Gap}^{\mathcal{Z}}_f(\mathbf{z}) = \sup_{\mathbf{y}':\|\mathbf{y}'-\mathbf{y}^*\|\le D} f(\mathbf{x}, \mathbf{y}') - \inf_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}^*\|\le D} f(\mathbf{x}', \mathbf{y})$.

Proposition 21 *Fix any $n \in \mathbb{N}$ and $L, D > 0$. For $t \in \mathbb{N}$, let η_t be a sequence with $\eta_t \in (0, 1/L)$ for each t. Let $\mathbf{z}^{(t)} = (\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ denote the iterates of EG with step-sizes η_t, as in (41). Then there is some $F = F_f \in \mathcal{F}^{\text{bil}}_{n,L,D}$ such that $\text{Gap}^{\mathcal{Z}}_f(\mathbf{z}^{(T)}) \ge \frac{LD^2}{4\sqrt{T}}$.*

Proof The proof closely parallels that of Theorem 9. In particular, we choose $\mathbf{A}, \mathbf{b}$ as in (13), with $\nu \in (0, L)$ to be specified below. Set $\mathbf{C}_{0,t}(\mathbf{A}) := I - \eta_t \mathbf{A} + (\eta_t \mathbf{A})^2$. Then the update (41) can be written as

$$\mathbf{z}^{(t)} = \mathbf{C}_{0,t}(\mathbf{A}) \cdot \mathbf{z}^{(t-1)} - (I - \eta_t \mathbf{A}) \cdot \eta_t \mathbf{b},$$

We claim that for $t \ge 0$,

$$\mathbf{A}\mathbf{z}^{(t)} + \mathbf{b} = \prod_{t'=0}^{t-1} \mathbf{C}_{0,t'}(\mathbf{A}) \cdot \mathbf{b}. \tag{42}$$

To see (42), we argue by induction, noting that the base case is immediate since $\mathbf{z}^{(0)} = \mathbf{0}$, and for the inductive step:

$$\begin{aligned}
\mathbf{A}\mathbf{z}^{(t+1)} + \mathbf{b} &= \mathbf{A}\mathbf{C}_{0,t}(\mathbf{A})\mathbf{z}^{(t-1)} - \mathbf{A}(I - \eta_t \mathbf{A})\eta_t \mathbf{b} + \mathbf{b} \\
&= \mathbf{A}\mathbf{C}_{0,t}(\mathbf{A})\mathbf{z}^{(t-1)} + (\mathbf{C}_{0,t}(\mathbf{A}) - I)\mathbf{b} + \mathbf{b} \\
&= \mathbf{C}_{0,t}(\mathbf{A}) \cdot (\mathbf{A}\mathbf{z}^{(t-1)} + \mathbf{b}).
\end{aligned}$$

Thus, by (24), we have

$$\text{Gap}^{\mathcal{Z}}_f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) = \frac{D^2 \nu \left\| \prod_{t'=0}^{t-1} \mathbf{C}_{0,t'}(\mathbf{A}) \cdot \mathbf{b} \right\|}{\|\mathbf{b}\|}. \tag{43}$$

Let us choose $\nu = L/\sqrt{T}$. It is straightforward to check that the singular values of $\prod_{t'=0}^{t-1} \mathbf{C}_{0,t'}(\mathbf{A})$, which are equal to the magnitudes of its eigenvalues, are all equal to

$$\begin{aligned}
\left| \prod_{t'=0}^{t-1} (1 - \eta_t \nu i + (\eta_t \nu i)^2) \right| &= \prod_{t'=0}^{t-1} \left| (1 - \eta_t^2 \nu^2) + \eta_t \nu i \right| \\
&\ge \prod_{t'=0}^{t-1} (1 - \eta_t^2 \nu^2) \ge (1 - 1/T)^T \ge 1/4
\end{aligned}$$

for $T \ge 2$. For any nonzero choice of $\mathbf{b}$, (43) thus gives that $\text{Gap}^{\mathcal{Z}}_f(\mathbf{z}^{(t)}) \ge \frac{D^2 L}{4\sqrt{T}}$. ■