Tight last-iterate convergence rates for no-regret learning in multi-player games

Noah Golowich*
MIT CSAIL
nzg@mit.edu

Sarath Pattathil
MIT EECS
sarathp@mit.edu

Constantinos Daskalakis†
MIT CSAIL
costis@csail.mit.edu

October 27, 2020

Abstract

We study the question of obtaining last-iterate convergence rates for no-regret learning algorithms in multi-player games. We show that the optimistic gradient (OG) algorithm with a constant step-size, which is no-regret, achieves a last-iterate rate of $O(1/\sqrt{T})$ with respect to the gap function in smooth monotone games. This result addresses a question of Mertikopoulos & Zhou (2018), who asked whether extra-gradient approaches (such as OG) can be applied to achieve improved guarantees in the multi-agent learning setting. The proof of our upper bound uses a new technique centered around an adaptive choice of potential function at each iteration. We also show that the $O(1/\sqrt{T})$ rate is tight for all p-SCLI algorithms, which includes OG as a special case. As a byproduct of our lower bound analysis we additionally present a proof of a conjecture of Arjevani et al. (2015) which is more direct than previous approaches.

1 Introduction

In the setting of *multi-agent online learning* ([SS11, CBL06]), K players interact with each other over time. At each time step t, each player $k \in \{1, \dots, K\}$ chooses an *action* $\mathbf{z}_k^{(t)}$; $\mathbf{z}_k^{(t)}$ may represent, for instance, the bidding strategy of an advertiser at time t. Player k then suffers a *loss* $\ell_t(\mathbf{z}_k^{(t)})$ that depends on both player k's action $\mathbf{z}_k^{(t)}$ and the actions of all other players at time t (which are absorbed into the loss function $\ell_t(\cdot)$). Finally, player k receives some *feedback* informing them of how to improve their actions in future iterations. In this paper we study gradient-based feedback, meaning that the feedback is the vector $\mathbf{g}_k^{(t)} = \nabla_{\mathbf{z}_k} \ell_t(\mathbf{z}_k^{(t)})$.

A fundamental quantity used to measure the performance of an online learning algorithm is the *regret* of player k, which is the difference between the total loss of player k over T time steps and the loss of the best possible action in hindsight: formally, the regret at time T is $\sum_{t=1}^T \ell_t(\mathbf{z}_k^{(t)}) - \min_{\mathbf{z}_k} \sum_{t=1}^T \ell_t(\mathbf{z}_k)$. An algorithm is said to be *no-regret* if its regret at time T grows sub-linearly with T for an adversarial choice of the loss functions ℓ_t. If all agents playing a game follow no-regret learning algorithms to choose their actions, then it is well-known that the empirical frequency of their actions converges to a *coarse correlated equilibrium (CCE)* ([MV78, CBL06]). In turn, a substantial body of work (e.g., [CBL06, DP09, EDMN09, CD11, VZ13, KKDB15, BTHK15, MP17, MZ18, KBTB18]) has focused on establishing for which classes of games or learning algorithms this convergence to a CCE can be strengthened, such as to convergence to a *Nash equilibrium (NE)*.

However, the type of convergence guaranteed in these works generally either applies only to the time-average of the joint action profiles, or else requires the sequence of learning rates to converge to 0. Such guarantees leave substantial room for improvement: a statement about the average of the joint action profiles

*Supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship.
†Supported by NSF Awards IIS-1741137, CCF-1617730 and CCF-1901292, by a Simons Investigator Award, and by the DOE PhILMs project (No. DE-AC05-76RL01830).

Table 1: Known last-iterate convergence rates for learning in smooth monotone games with perfect gradient feedback (i.e., *deterministic* algorithms). We specialize to the 2-player 0-sum case in presenting prior work, since some papers in the literature only consider this setting. Recall that a game $\mathcal{G}$ has a γ-*singular value lower bound* if for all $\mathbf{z}$, all singular values of $\partial F_{\mathcal{G}}(\mathbf{z})$ are $\geq \gamma$. ℓ, Λ are the Lipschitz constants of $F_{\mathcal{G}}, \partial F_{\mathcal{G}}$, respectively, and $c, C > 0$ are absolute constants where c is sufficiently small and C is sufficiently large. Upper bounds in the left-hand column are for the EG algorithm, and lower bounds are for a general form of 1-SCLI methods which include EG. Upper bounds in the right-hand column are for algorithms which are implementable as online no-regret learning algorithms (e.g., OG or online gradient descent), and lower bounds are shown for two classes of algorithms containing OG and online gradient descent, namely p-SCLI algorithms for general $p \geq 1$ (recall for OG, $p = 2$) as well as those satisfying a 2-step linear span assumption (see [IAGM19]). The reported upper and lower bounds are stated for the total gap function (Definition 3); leading constants and factors depending on distance between initialization and optimum are omitted.

	Deterministic	
Game class	Extra gradient	Implementable as no-regret
μ-strongly monotone	*Upper:* $\ell\left(1-\frac{c\mu}{\ell}\right)^T$ [MOP19b, EG] *Lower:* $\mu\left(1-\frac{C\mu}{\ell}\right)^T$ [AMLJG19, 1-SCLI]	*Upper:* $\ell\left(1-\frac{c\mu}{\ell}\right)^T$ [MOP19b, OG] *Lower:* $\mu\left(1-\frac{C\mu}{\ell}\right)^T$ [IAGM19, 2-step lin. span] *Lower:* $\mu\left(1-\sqrt[p]{\frac{C\mu}{\ell}}\right)^T$ [ASSS15, IAGM19, p-SCLI]
Monotone, γ-sing. val. low. bnd.	*Upper:* $\ell\left(1-\frac{c\gamma^2}{\ell^2}\right)^T$ [AMLJG19, EG] *Lower:* $\gamma\left(1-\frac{C\gamma^2}{\ell^2}\right)^T$ [AMLJG19, 1-SCLI]	*Upper:* $\ell\left(1-\frac{c\gamma^2}{\ell^2}\right)^T$ [AMLJG19, OG] *Lower:* $\gamma\left(1-\frac{C\gamma}{\ell}\right)^T$ [IAGM19, 2-step lin. span] *Lower:* $\gamma\left(1-\sqrt[p]{\frac{C\gamma}{\ell}}\right)^T$ [ASSS15, IAGM19, p-SCLI]
λ-cocoercive	Open	*Upper:* $\frac{1}{\lambda\sqrt{T}}$ [LZMJ20, Online grad. descent]
Monotone	*Upper:* $\frac{\ell+\Lambda}{\sqrt{T}}$ [GPDO20, EG] *Lower:* $\frac{\ell}{\sqrt{T}}$ [GPDO20, 1-SCLI]	*Upper:* $\frac{\ell+\Lambda}{\sqrt{T}}$ (**Theorem 5**, OG) *Lower:* $\frac{\ell}{\sqrt{T}}$ (**Theorem 7**, p-SCLI, lin. coeff. matrices)

fails to capture the game dynamics over time ([MPP17]), and both types of guarantees use newly acquired information with decreasing weight, which, as remarked by [LZMJ20], is very unnatural from an economic perspective.[1] Therefore, the following question is of particular interest ([MZ18, LZMJ20, MPP17, DISZ17]):

Can we establish last-iterate rates if all players act according to a no-regret learning algorithm with constant step size? ($\star$)

We measure the proximity of an action profile $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_K)$ to equilibrium in terms of the *total gap function* at $\mathbf{z}$ (Definition 3): it is defined to be the sum over all players k of the maximum decrease in cost player k could achieve by deviating from its action $\mathbf{z}_k$. [LZMJ20] took initial steps toward addressing ($\star$), showing that if all agents follow the *online gradient descent* algorithm, then for all λ-*cocoercive games*, the action profiles $\mathbf{z}^{(t)} = (\mathbf{z}_1^{(t)}, \ldots, \mathbf{z}_K^{(t)})$ will converge to equilibrium in terms of the total gap function at a rate of $O(1/\sqrt{T})$. Moreover, linear last-iterate rates have been long known for smooth *strongly-monotone* games ([Tse95, GBV+18, LS18, MOP19b, AMLJG19, ZMM+20]), a sub-class of λ-cocoercive games. Unfortunately, even λ-cocoercive games exclude many important classes of games, such as bilinear games, which are the adaptation of matrix games to the unconstrained setting. Moreover, this shortcoming is not merely an artifact of the analysis of [LZMJ20]: it has been observed (e.g. [DISZ17, GBV+18]) that in bilinear games, the players' actions in online gradient descent not only fail to converge, but diverge to infinity. Prior work on last-iterate convergence rates for these various subclasses of monotone games is summarized in Table 1 for the case of perfect gradient feedback; the setting for noisy feedback is summarized in Table 2 in Appendix A.4.

[1] In fact, even in the adversarial setting, standard no-regret algorithms such as FTRL ([SS11]) need to be applied with decreasing step-size in order to achieve sublinear regret.

1.1 Our contributions

In this paper we answer ($\star$) in the affirmative for all *monotone games* (Definition 1) satisfying a mild smoothness condition, which includes smooth λ-cocoercive games and bilinear games. Many common and well-studied classes of games, such as zero-sum polymatrix games ([BF87, DP09, CCDP16]) and its generalization zero-sum socially-concave games ([EDMN09]) are monotone but are not in general λ-cocoercive. Hence our paper is the first to prove last-iterate convergence in the sense of ($\star$) for the unconstrained version of these games as well. In more detail, we establish the following:

- We show in Theorem 5 and Corollary 6 that the actions taken by learners following the *optimistic gradient (OG)* algorithm, which is no-regret, exhibit last-iterate convergence to a Nash equilibrium in smooth, monotone games at a rate of $O(1/\sqrt{T})$ in terms of the global gap function. The proof uses a new technique which we call *adaptive potential functions* (Section 3.1) which may be of independent interest.
- We show in Theorem 7 that the rate $O(1/\sqrt{T})$ cannot be improved for any algorithm belonging to the class of p-SCLI algorithms (Definition 5), which includes OG.

The OG algorithm is closely related to the *extra-gradient (EG)* algorithm ([Kor76, Nem04]),[2] which, at each time step t, assumes each player k has an oracle $\mathcal{O}_k$ which provides them with an additional gradient at a slightly different action than the action $\mathbf{z}_k^{(t)}$ played at step t. Hence EG does not naturally fit into the standard setting of multi-agent learning. One could try to "force" EG into the setting of multi-agent learning by taking actions at odd-numbered time steps t to simulate the oracle $\mathcal{O}_k$, and using the even-numbered time steps to simulate the actions $\mathbf{z}_k^{(t)}$ that EG actually takes. Although this algorithm exhibits last-iterate convergence at a rate of $O(1/\sqrt{T})$ in smooth monotone games when all players play according to it [GPDO20], it is straightforward to see that it is *not* a no-regret learning algorithm, i.e., for an adversarial loss function the regret can be linear in T (see Proposition 10 in Appendix A.3).

Nevertheless, due to the success of EG at solving monotone variational inequalities, [MZ18] asked whether similar techniques to EG could be used to speed up last-iterate convergence to Nash equilibria. Our upper bound for OG answers this question in the affirmative: various papers ([CYL+12, RS12, RS13, HIMM19]) have observed that OG may be viewed as an approximation of EG, in which the previous iteration's gradient is used to simulate the oracle $\mathcal{O}_k$. Moreover, our upper bound of $O(1/\sqrt{T})$ applies in many games for which the approach used in [MZ18], namely Nesterov's dual averaging ([Nes09]), either fails to converge (such as bilinear games) or only yields asymptotic rates with decreasing learning rate (such as smooth strictly monotone games). Proving last-iterate rates for OG has also been noted as an important open question in [HIMM19, Table 1]. At a technical level, the proof of our upper bound (Theorem 5) uses the proof technique in [GPDO20] for the last-iterate convergence of EG as a starting point. In particular, similar to [GPDO20], our proof proceeds by first noting that some iterate $\mathbf{z}^{(t^*)}$ of OG will have gradient gap $O(1/\sqrt{T})$ (see Definition 2; this is essentially a known result) and then showing that for all $t \geq t^*$ the gradient gap only increases by at most a constant factor. The latter step is the bulk of the proof, as was the case in [GPDO20]; however, since each iterate of OG depends on the previous two iterates and gradients, the proof for OG is significantly more involved than that for EG. We refer the reader to Section 3.1 and Appendix B for further details.

The proof of our lower bound for p-SCLI algorithms, Theorem 7, reduces to a question about the spectral radius of a family of polynomials. In the course of our analysis we prove a conjecture by [ASSS15] about such polynomials; though the validity of this conjecture is implied by each of several independent results in the literature (e.g., [AS16, Nev93]), our proof is more direct than previous ones.

Lastly, we mention that our focus in this paper is on the unconstrained setting, meaning that the players' losses are defined on all of Euclidean space. We leave the constrained setting, in which the players must project their actions onto a convex constraint set, to future work.

1.2 Related work

Multi-agent learning in games. In the constrained setting, many papers have studied conditions under which the action profile of no-regret learning algorithms, often variants of Follow-The-Regularized-Leader

[2] EG is also known as *mirror-prox*, which specifically refers to its generalization to general Bregman divergences.

(FTRL), converges to equilibrium. However, these works all assume either a learning rate that decreases over time ([MZ18, ZMB+17, ZMA+18, ZMM+17]), or else only apply to specific types of *potential games* ([KKDB15, KBTB18, PPP17, KPT09, CL16, BEDL06, PP14]), which significantly facilitates the analysis of last-iterate convergence.[3]

Such potential games are in general incomparable with monotone games, and do not even include finite-state two-player zero sum games (i.e., *matrix games*). In fact, [BP18] showed that the actions of players following FTRL in two-player zero-sum matrix games *diverge* from interior Nash equilibria. Many other works ([HMC03, MPP17, KLP11, DFP+10, BCM12, PP16]) establish similar non-convergence results in both discrete and continuous time for various types of monotone games, including zero-sum polymatrix games. Such non-convergence includes chaotic behavior such as Poincaré recurrence, which showcases the insufficiency of on-average convergence (which holds in such settings) and so is additional motivation for the question ($\star$).

Monotone variational inequalities & OG. The problem of finding a Nash equilibrium of a monotone game is exactly that of finding a solution to a monotone variational inequality (VI). OG was originally introduced by [Pop80], who showed that its iterates converge to solutions of monotone VIs, without proving explicit rates.[4] It is also well-known that the *averaged* iterate of OG converges to the solution of a monotone VI at a rate of $O(1/T)$ ([HIMM19, MOP19a, RS13]), which is known to be optimal ([Nem04, OX19, ASM+20]). Recently it has been shown ([DP18, LNPW20]) that a modification of OG known as optimistic multiplicative-weights update exhibits last-iterate convergence to Nash equilibria in two-player zero-sum monotone games, but as with the unconstrained case ([MOP19a]) non-asymptotic rates are unknown. To the best of our knowledge, the only work proving last-iterate convergence rates for general smooth monotone VIs was [GPDO20], which only treated the EG algorithm, which is not no-regret. There is a vast literature on solving VIs, and we refer the reader to [FP03] for further references.

2 Preliminaries

Throughout this paper we use the following notational conventions. For a vector $\mathbf{v} \in \mathbb{R}^n$, let $\|\mathbf{v}\|$ denote the Euclidean norm of $\mathbf{v}$. For $\mathbf{v} \in \mathbb{R}^n$, set $\mathcal{B}(\mathbf{v}, R) := \{\mathbf{z} \in \mathbb{R}^n : \|\mathbf{v} - \mathbf{z}\| \leq R\}$; when we wish to make the dimension explicit we write $\mathcal{B}_{\mathbb{R}^n}(\mathbf{v}, R)$. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ let $\|\mathbf{A}\|_\sigma$ denote the spectral norm of $\mathbf{A}$.

We let the set of K players be denoted by $\mathcal{K} := \{1, 2, \dots K\}$. Each player k's actions $\mathbf{z}_k$ belong to their *action set*, denoted $\mathcal{Z}_k$, where $\mathcal{Z}_k \subseteq \mathbb{R}^{n_k}$ is a convex subset of Euclidean space. Let $\mathcal{Z} = \prod_{k=1}^K \mathcal{Z}_k \subseteq \mathbb{R}^n$, where $n = n_1 + \cdots + n_K$. In this paper we study the setting where the action sets are unconstrained (as in [LZMJ20]), meaning that $\mathcal{Z}_k = \mathbb{R}^{n_k}$, and $\mathcal{Z} = \mathbb{R}^n$, where $n = n_1 + \cdots + n_K$. The *action profile* is the vector $\mathbf{z} := (\mathbf{z}_1, \dots, \mathbf{z}_K) \in \mathcal{Z}$. For any player $k \in \mathcal{K}$, let $\mathbf{z}_{-k} \in \prod_{k' \neq k} \mathcal{Z}_{k'}$ be the vector of actions of all the other players. Each player $k \in \mathcal{K}$ wishes to minimize its *cost function* $f_k : \mathcal{Z} \to \mathbb{R}$, which is assumed to be twice continuously differentiable. The tuple $\mathcal{G} := (\mathcal{K}, (\mathcal{Z}_k)_{k=1}^K, (f_k)_{k=1}^K)$ is known as a *continuous game*.

At each time step t, each player k plays an action $\mathbf{z}_k^{(t)}$; we assume the feedback to player k is given in the form of the gradient $\nabla_{\mathbf{z}_k} f_k(\mathbf{z}_k^{(t)}, \mathbf{z}_{-k}^{(t)})$ of their cost function with respect to their action $\mathbf{z}_k^{(t)}$, given the actions $\mathbf{z}_{-k}^{(t)}$ of the other players at time t. We denote the concatenation of these gradients by $F_{\mathcal{G}}(\mathbf{z}) := (\nabla_{\mathbf{z}_1} f_1(\mathbf{z}), \dots, \nabla_{\mathbf{z}_K} f_K(\mathbf{z})) \in \mathbb{R}^n$. When the game $\mathcal{G}$ is clear, we will sometimes drop the subscript and write $F : \mathcal{Z} \to \mathbb{R}^n$.

Equilibria & monotone games. A *Nash equilibrium* in the game $\mathcal{G}$ is an action profile $\mathbf{z}^* \in \mathcal{Z}$ so that for each player k, it holds that $f_k(\mathbf{z}_k^*, \mathbf{z}_{-k}^*) \leq f_k(\mathbf{z}_k', \mathbf{z}_{-k}^*)$ for any $\mathbf{z}_k' \in \mathcal{Z}_k$. Throughout this paper we study *monotone* games:

[3] In *potential games*, there is a canonical choice of potential function whose local minima are equivalent to being at a Nash equilibrium. The lack of existence of a natural potential function in general monotone games is a significant challenge in establishing last-iterate convergence.

[4] Technically, the result of [Pop80] only applies to two-player zero-sum monotone games (i.e., finding the saddle point of a convex-concave function). The proof readily extends to general monotone VIs ([HIMM19]).

Definition 1 (Monotonicity; [Ros65])**.** The game $\mathcal{G} = (\mathcal{K}, (\mathcal{Z}_k)_{k=1}^K, (f_k)_{k=1}^K)$ is *monotone* if for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, it holds that $\langle F_{\mathcal{G}}(\mathbf{z}') - F_{\mathcal{G}}(\mathbf{z}), \mathbf{z}' - \mathbf{z} \rangle \geq 0$. In such a case, we say also that $F_{\mathcal{G}}$ is a monotone operator.

The following classical result characterizes the Nash equilibria in monotone games:

Proposition 1 ([FP03])**.** *In the unconstrained setting, if the game $\mathcal{G}$ is monotone, any Nash equilibrium $\mathbf{z}^*$ satisfies $F_{\mathcal{G}}(\mathbf{z}^*) = \mathbf{0}$. Conversely, if $F_{\mathcal{G}}(\mathbf{z}) = \mathbf{0}$, then $\mathbf{z}$ is a Nash equilibrium.*

In accordance with Proposition 1, one measure of the proximity to equilibrium of some $\mathbf{z} \in \mathcal{Z}$ is the norm of $F_{\mathcal{G}}(\mathbf{z})$:

Definition 2 (Gradient gap function)**.** Given a monotone game $\mathcal{G}$ with its associated operator $F_{\mathcal{G}}$, the *gradient gap function* evaluated at $\mathbf{z}$ is defined to be $\|F_{\mathcal{G}}(\mathbf{z})\|$.

It is also common ([MOP19a, Nem04]) to measure the distance from equilibrium of some $\mathbf{z} \in \mathcal{Z}$ by adding the maximum decrease in cost that each player could achieve by deviating from their current action $\mathbf{z}_k$:

Definition 3 (Total gap function)**.** Given a monotone game $\mathcal{G} = (\mathcal{K}, (\mathcal{Z}_k)_{k=1}^K, (f_k)_{k=1}^K)$, compact subsets $\mathcal{Z}_k' \subseteq \mathcal{Z}_k$ for each $k \in \mathcal{K}$, and a point $\mathbf{z} \in \mathcal{Z}$, define the *total gap function* at $\mathbf{z}$ with respect to the set $\mathcal{Z}' := \prod_{k=1}^K \mathcal{Z}_k'$ by $\mathrm{TGap}_{\mathcal{G}}^{\mathcal{Z}'}(\mathbf{z}) := \sum_{k=1}^K \left(f_k(\mathbf{z}) - \min_{\mathbf{z}_k' \in \mathcal{Z}_k'} f_k(\mathbf{z}_k', \mathbf{z}_{-k}) \right)$. At times we will slightly abuse notation, and for $F := F_{\mathcal{G}}$, write $\mathrm{TGap}_F^{\mathcal{Z}'}$ in place of $\mathrm{TGap}_{\mathcal{G}}^{\mathcal{Z}'}$.

As discussed in [GPDO20], it is in general impossible to obtain meaningful guarantees on the total gap function by allowing each player to deviate to an action in their entire space $\mathcal{Z}_k$, which necessitates defining the total gap function in Definition 3 with respect to the compact subsets $\mathcal{Z}_k'$. We discuss in Remark 4 how, in our setting, it is without loss of generality to shrink $\mathcal{Z}_k$ so that $\mathcal{Z}_k = \mathcal{Z}_k'$ for each k. Proposition 2 below shows that in monotone games, the gradient gap function upper bounds the total gap function:

Proposition 2. *Suppose $\mathcal{G} = (\mathcal{K}, (\mathcal{Z}_k)_{k=1}^K, (f_k)_{k=1}^K)$ is a monotone game, and compact subsets $\mathcal{Z}_k' \subset \mathcal{Z}_k$ are given, where the diameter of each $\mathcal{Z}_k'$ is upper bounded by $D > 0$. Then*

$$\mathrm{TGap}_{\mathcal{G}}^{\mathcal{Z}'}(\mathbf{z}) \leq D\sqrt{K} \cdot \|F_{\mathcal{G}}(\mathbf{z})\|.$$

For completeness, a proof of Proposition 2 is presented in Appendix A.

Special case: convex-concave min-max optimization. Since in a two-player zero-sum game $\mathcal{G} = (\{1,2\}, (\mathcal{Z}_1, \mathcal{Z}_2), (f_1, f_2))$ we must have $f_1 = -f_2$, it is straightforward to show that $f_1(\mathbf{z}_1, \mathbf{z}_2)$ is convex in $\mathbf{z}_1$ and concave in $\mathbf{z}_2$. Moreover, it is immediate that Nash equilibria of the game $\mathcal{G}$ correspond to saddle points of f_1; thus a special case of our setting is that of finding saddle points of convex-concave functions ([FP03]). Such saddle point problems have received much attention recently since they can be viewed as a simplified model of generative adversarial networks (e.g., [GBV+18, DISZ17, CGFLJ19, GHP+18, YSX+17]).

Optimistic gradient (OG) algorithm. In the *optimistic gradient (OG)* algorithm, each player k performs the following update:

$$\mathbf{z}_k^{(t+1)} := \mathbf{z}_k^{(t)} - 2\eta_t \mathbf{g}_k^{(t)} + \eta_t \mathbf{g}_k^{(t-1)}, \tag{OG}$$

where $\mathbf{g}_k^{(t)} = \nabla_{\mathbf{z}_k} f_k(\mathbf{z}_k^{(t)}, \mathbf{z}_{-k}^{(t)})$ for $t \geq 0$. The following essentially optimal regret bound is well-known for the OG algorithm, when the actions of the other players $\mathbf{z}_{-k}^{(t)}$ (often referred to as the *environment*'s actions) are adversarial:

Proposition 3. *Assume that for all $\mathbf{z}_{-k}$ the function $\mathbf{z}_k \mapsto f_k(\mathbf{z}_k, \mathbf{z}_{-k})$ is convex. Then the regret of OG with learning rate $\eta_t = O(D/L\sqrt{t})$ is $O(DL\sqrt{T})$, where $L = \max_t \|\mathbf{g}_k^{(t)}\|$ and $D = \max\{\|\mathbf{z}_k^*\|, \max_t \|\mathbf{z}_k^{(t)}\|\}$.*

In Proposition 3, $\mathbf{z}_k^*$ is defined by $\mathbf{z}_k^* \in \arg\min_{\mathbf{z}_k \in \mathcal{Z}_k} \sum_{t'=0}^t f_k(\mathbf{z}_k, \mathbf{z}_{-k}^{(t')})$. The assumption in the proposition that $\|\mathbf{z}_k^{(t)}\| \leq D$ may be satisfied in the unconstrained setting by projecting the iterates onto the region $\mathcal{B}(0, D) \subset \mathbb{R}^{n_k}$, for some $D \geq \|\mathbf{z}_k^*\|$, without changing the regret bound. The implications of this modification to (OG) are discussed further in Remark 4.

3 Last-iterate rates for OG via adaptive potential functions

In this section we show that in the unconstrained setting (namely, that where $\mathcal{Z}_k = \mathbb{R}^{n_k}$ for all $k \in \mathcal{K}$), when all players act according to OG, their iterates exhibit last-iterate convergence to a Nash equilibrium. Our convergence result holds for games $\mathcal{G}$ for which the operator $F_{\mathcal{G}}$ satisfies the following smoothness assumption:

Assumption 4 (Smoothness). *For a monotone operator $F : \mathcal{Z} \to \mathbb{R}^n$, assume that the following first and second-order Lipschitzness conditions hold, for some $\ell, \Lambda > 0$:*

$$\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \qquad \|F(\mathbf{z}) - F(\mathbf{z}')\| \le \ell \cdot \|\mathbf{z} - \mathbf{z}'\| \tag{1}$$

$$\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \qquad \|\partial F(\mathbf{z}) - \partial F(\mathbf{z}')\|_\sigma \le \Lambda \cdot \|\mathbf{z} - \mathbf{z}'\|. \tag{2}$$

Here $\partial F : \mathcal{Z} \to \mathbb{R}^{n \times n}$ denotes the Jacobian of F.

Condition (1) is entirely standard in the setting of solving monotone variational inequalities ([Nem04]); condition (2) is also very mild, being made for essentially all second-order methods (e.g., [ALW19, Nes06]).

By the definition of $F_{\mathcal{G}}(\cdot)$, when all players in a game $\mathcal{G}$ act according to (OG) with constant step size η, then the action profile $\mathbf{z}^{(t)}$ takes the form

$$\mathbf{z}^{(-1)}, \mathbf{z}^{(0)} \in \mathbb{R}^n, \qquad \mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - 2\eta F_{\mathcal{G}}(\mathbf{z}^{(t)}) + \eta F_{\mathcal{G}}(\mathbf{z}^{(t-1)}) \quad \forall t \ge 0. \tag{3}$$

The main theorem of this section, Theorem 5, shows that under the OG updates (3), the iterates converge at a rate of $O(1/\sqrt{T})$ to a Nash equilibrium with respect to the gradient gap function:

Theorem 5 (Last-iterate convergence of OG). *Suppose $\mathcal{G}$ is a monotone game so that $F_{\mathcal{G}}$ satisfies Assumption 4. For some $\mathbf{z}^{(-1)}, \mathbf{z}^{(0)} \in \mathbb{R}^n$, suppose there is $\mathbf{z}^* \in \mathbb{R}^n$ so that $F_{\mathcal{G}}(\mathbf{z}^*) = 0$ and $\|\mathbf{z}^* - \mathbf{z}^{(-1)}\| \le D, \|\mathbf{z}^* - \mathbf{z}^{(0)}\| \le D$. Then the iterates $\mathbf{z}^{(T)}$ of OG (3) for any $\eta \le \min\left\{\frac{1}{150\ell}, \frac{1}{1711D\Lambda}\right\}$ satisfy:*

$$\|F_{\mathcal{G}}(\mathbf{z}^{(T)})\| \le \frac{60D}{\eta\sqrt{T}} \tag{4}$$

By Proposition 2, we immediately get a bound on the total gap function at each time T:

Corollary 6 (Total gap function for last iterate of OG). *In the setting of Theorem 5, let $\mathcal{Z}'_k := \mathcal{B}(\mathbf{z}_k^{(0)}, 3D)$ for each $k \in \mathcal{K}$. Then, with $\mathcal{Z}' = \prod_{k \in \mathcal{K}} \mathcal{Z}'_k$,*

$$\mathrm{TGap}_{\mathcal{G}}^{\mathcal{Z}'}(\mathbf{z}^{(T)}) \le \frac{180KD^2}{\eta\sqrt{T}}. \tag{5}$$

We made no attempt to optimize the consants in Theorem 5 and Corollary 6, and they can almost certainly be improved.

Remark 4 (Bounded iterates). Recall from the discussion following Proposition 3 that it is necessary to project the iterates of OG onto a compact ball to achieve the no-regret property. As our guiding question ($\star$) asks for last-iterate rates achieved by a no-regret algorithm, we should ensure that such projections are compatible with the guarantees in Theorem 5 and Corollary 6. For this we note that [MOP19a, Lemma 4(b)] showed that for the dynamics (3) without constraints, for all $t \ge 0$, $\|\mathbf{z}^{(t)} - \mathbf{z}^*\| \le 2\|\mathbf{z}^{(0)} - \mathbf{z}^*\|$. Therefore, as long as we make the very mild assumption of a known a priori upper bound $\|\mathbf{z}^*\| \le D/2$ (as well as $\|\mathbf{z}_k^{(-1)}\| \le D/2$, $\|\mathbf{z}_k^{(0)}\| \le D/2$), if all players act according to (3), then the updates (3) remain unchanged if we project onto the constraint sets $\mathcal{Z}_k := \mathcal{B}(\mathbf{0}, 3D)$ at each time step t. This observation also serves as motivation for the compact sets $\mathcal{Z}'_k$ used in Corollary 6: the natural choice for $\mathcal{Z}'_k$ is $\mathcal{Z}_k$ itself, and by restricting $\mathcal{Z}_k$ to be compact, this choice becomes possible.

3.1 Proof overview: adaptive potential functions

In this section we sketch the idea of the proof of Theorem 5; full details of the proof may be found in Appendix B. First we note that it follows easily from results of [HIMM19] that OG exhibits *best-iterate*

convergence, i.e., in the setting of Theorem 5 we have, for each $T > 0$, $\min_{1 \le t \le T} \|F_{\mathcal{G}}(\mathbf{z}^{(t)})\| \le O(1/\sqrt{T})$.[5] The main contribution of our proof is then to show the following: if we choose t^* so that $\|F_{\mathcal{G}}(\mathbf{z}^{(t^*)})\| \le O(1/\sqrt{T})$, then for all $t' \ge t^*$, we have $\|F_{\mathcal{G}}(\mathbf{z}^{(t')})\| \le O(1) \cdot \|F_{\mathcal{G}}(\mathbf{z}^{(t^*)})\|$. This was the same general approach taken in [GPDO20] to prove that the extragradient (EG) algorithm has last-iterate convergence. In particular, they showed the stronger statement that $\|F_{\mathcal{G}}(\mathbf{z}^{(t)})\|$ may be used as an approximate potential function in the sense that it only increases by a small amount each step:

$$\|F_{\mathcal{G}}(\mathbf{z}^{(t'+1)})\| \underbrace{\le}_{t' \ge 0} (1 + \|F(\mathbf{z}^{(t')})\|^2) \cdot \|F_{\mathcal{G}}(\mathbf{z}^{(t')})\| \underbrace{\le}_{t' \ge t^*} (1 + O(1/T)) \cdot \|F_{\mathcal{G}}(\mathbf{z}^{(t')})\|. \tag{6}$$

However, their approach relies crucially on the fact that for the EG algorithm, $\mathbf{z}^{(t+1)}$ depends only on $\mathbf{z}^{(t)}$. For the OG algorithm, it is possible that (6) fails to hold, even when $F_{\mathcal{G}}(\mathbf{z}^{(t)})$ is replaced by the more natural choice of $(F_{\mathcal{G}}(\mathbf{z}^{(t)}), F_{\mathcal{G}}(\mathbf{z}^{(t-1)}))$.[6]

Instead of using $\|F_{\mathcal{G}}(\mathbf{z}^{(t)})\|$ as a potential function in the sense of (6), we propose instead to track the behavior of $\|\tilde{F}^{(t)}\|$, where

$$\tilde{F}^{(t)} := F_{\mathcal{G}}(\mathbf{z}^{(t)} + \eta F_{\mathcal{G}}(\mathbf{z}^{(t-1)})) + \mathbf{C}^{(t-1)} \cdot F_{\mathcal{G}}(\mathbf{z}^{(t-1)}) \in \mathbb{R}^n, \tag{7}$$

and the matrices $\mathbf{C}^{(t-1)} \in \mathbb{R}^{n \times n}$ are defined recursively *backwards*, i.e., $\mathbf{C}^{(t-1)}$ depends directly on $\mathbf{C}^{(t)}$, which depends directly on $\mathbf{C}^{(t+1)}$, and so on. For an appropriate choice of the matrices $\mathbf{C}^{(t)}$, we show that $\tilde{F}^{(t+1)} = (I - \eta \mathbf{A}^{(t)} + \mathbf{C}^{(t)}) \cdot \tilde{F}^{(t)}$, for some matrix $\mathbf{A}^{(t)} \approx \partial F_{\mathcal{G}}(\mathbf{z}^{(t)})$. We then show that for $t \ge t^*$, it holds that $\|I - \eta \mathbf{A}^{(t)} + \mathbf{C}^{(t)}\|_\sigma \le 1 + O(1/T)$, from which it follows that $\|\tilde{F}^{(t+1)}\| \le (1 + O(1/T)) \cdot \|\tilde{F}^{(t)}\|$. This modification of (6) is enough to show the desired upper bound of $\|F_{\mathcal{G}}(\mathbf{z}^{(T)})\| \le O(1/\sqrt{T})$.

To motivate the choice of $\tilde{F}^{(t)}$ in (7) it is helpful to consider the simple case where $F(\mathbf{z}) = \mathbf{A}\mathbf{z}$ for some $\mathbf{A} \in \mathbb{R}^{n \times n}$, which was studied by [LS18]. Simple algebraic manipulations using (3) (detailed in Appendix B) show that, for the matrix $\mathbf{C} := \frac{(I + (2\eta\mathbf{A})^2)^{1/2} - I}{2}$, we have $\tilde{F}^{(t+1)} = (I - \eta\mathbf{A} + \mathbf{C})\tilde{F}^{(t)}$ for all t. It may be verified that we indeed have $\mathbf{A}^{(t)} = \mathbf{A}$ and $\mathbf{C}^{(t)} = \mathbf{C}$ for all t in this case, and thus (7) may be viewed as a generalization of these calculations to the nonlinear case.

Adaptive potential functions. In general, a *potential function* $\Phi(F_{\mathcal{G}}, \mathbf{z})$ depends on the problem instance, here taken to be $F_{\mathcal{G}}$, and an element $\mathbf{z}$ representing the current state of the algorithm. Many convergence analyses from optimization (e.g., [BG17, WRJ18], and references therein) have as a crucial element in their proofs a statement of the form $\Phi(F_{\mathcal{G}}, \mathbf{z}^{(t+1)}) \lesssim \Phi(F_{\mathcal{G}}, \mathbf{z}^{(t)})$. For example, for the iterates $\mathbf{z}^{(t)}$ of the EG algorithm, [GPDO20] (see (6)) used the potential function $\Phi(F_{\mathcal{G}}, \mathbf{z}^{(t)}) := \|F_{\mathcal{G}}(\mathbf{z}^{(t)})\|$.

Our approach of controlling the the norm of the vectors $\tilde{F}^{(t)}$ defined in (7) can also be viewed as an instantion of the potential function approach: since each iterate of OG depends on the previous two iterates, the state is now given by $\mathbf{v}^{(t)} := (\mathbf{z}^{(t-1)}, \mathbf{z}^{(t)})$. The potential function is given by $\Phi_{\mathrm{OG}}(F_{\mathcal{G}}, \mathbf{v}^{(t)}) := \|\tilde{F}^{(t)}\|$, where $\tilde{F}^{\top}$ is defined in (7) and indeed only depends on $\mathbf{v}^{(t)}$ once $F_{\mathcal{G}}$ is fixed since $\mathbf{v}^{(t)}$ determines $\mathbf{z}^{(t')}$ for all $t' \ge t$ (as OG is deterministic), which in turn determine $\mathbf{C}^{(t-1)}$. However, the potential function Φ_{OG} is quite unlike most other choices of potential functions in optimization (e.g., [BG17]) in the sense that it depends *globally* on $F_{\mathcal{G}}$: For any $t' > t$, a local change in $F_{\mathcal{G}}$ in the neighborhood of $\mathbf{v}^{(t')}$ may cause a change in $\Phi_{\mathrm{OG}}(F_{\mathcal{G}}, \mathbf{v}^{(t)})$, *even if* $\|\mathbf{v}^{(t)} - \mathbf{v}^{(t')}\|$ *is arbitrarily large*. Because $\Phi_{\mathrm{OG}}(F_{\mathcal{G}}, \mathbf{v}^{(t)})$ adapts to the behavior of $F_{\mathcal{G}}$ at iterates later on in the optimization sequence, we call it an *adaptive potential function*. We are not aware of any prior works using such adaptive potential functions to prove last-iterate convergence results, and we believe this technique may find additional applications.

4 Lower bound for convergence of p-SCLIs

The main result of this section is Theorem 7, stating that the bounds on last-iterate convergence in Theorem 5 and Corollary 6 are tight when we require the iterates $\mathbf{z}^{(T)}$ to be produced by an optimization algorithm

[5] In this discussion we view η, D as constants.

[6] For a trivial example, suppose that $n = 1$, $F_{\mathcal{G}}(\mathbf{z}) = \mathbf{z}$, $\mathbf{z}^{(t')} = \delta > 0$, and $\mathbf{z}^{(t'-1)} = 0$. Then $\|(F_{\mathcal{G}}(\mathbf{z}^{(t')}), F_{\mathcal{G}}(\mathbf{z}^{(t'-1)}))\| = \delta$ but $\|(F_{\mathcal{G}}(\mathbf{z}^{(t'+1)}), F_{\mathcal{G}}(\mathbf{z}^{(t')}))\| > \delta\sqrt{2 - 4\eta}$.

satisfying a particular formal definition of "last-iterate convergence". Notice that that we cannot hope to prove that they are tight for *all* first-order algorithms, since the averaged iterates $\bar{\mathbf{z}}^{(T)} := \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}^{(t)}$ of OG satisfy $\mathrm{TGap}_{\mathcal{G}}^{\mathcal{Z}'}(\bar{\mathbf{z}}^{(T)}) \leq O\left(\frac{D^2}{\eta T}\right)$ [MOP19a, Theorem 2]. Similar to [GPDO20], we use *p-stationary canonical linear iterative methods (p-SCLIs)* to formalize the notion of "last-iterate convergence". [GPDO20] only considered the special case $p = 1$ to establish a similar lower bound to Theorem 7 for a family of last-iterate algorithms including the extragradient algorithm. The case $p > 1$ leads to new difficulties in our proof since even for $p = 2$ we must rule out algorithms such as Nesterov's accelerated gradient descent ([Nes75]) and Pólya's heavy-ball method ([Pol87]), a situation that did not arise for $p = 1$.

Definition 5 (p-SCLIs [ASSS15, ASM+20])**.** An algorithm $\mathcal{A}$ is a *first-order p-stationary canonical linear iterative algorithm (p-SCLI)* if, given a monotone operator F, and an arbitrary set of p initialization points $\mathbf{z}^{(0)}, \mathbf{z}^{(-1)}, \ldots, \mathbf{z}^{(-p+1)} \in \mathbb{R}^n$, it generates iterates $\mathbf{z}^{(t)}$, $t \geq 1$, for which

$$\mathbf{z}^{(t)} = \sum_{j=0}^{p-1} \alpha_j \cdot F(\mathbf{z}^{(t-p+j)}) + \beta_j \cdot \mathbf{z}^{(t-p+j)}, \tag{8}$$

for $t = 1, 2, \ldots$, where $\alpha_j, \beta_j \in \mathbb{R}$ are any scalars.[7]

From (3) it is evident that OG with constant step size η is a 2-SCLI with $\beta_1 = 1, \beta_0 = 0, \alpha_1 = -2\eta, \alpha_0 = \eta$. Many standard algorithms for convex function minimization, including gradient descent, Nesterov's accelerated gradient descent (AGD), and Pólya's Heavy Ball method, are of the form (8) as well. We additionally remark that several variants of SCLIs (and their non-stationary counterpart, CLIs) have been considered in recent papers proving lower bounds for min-max optimization ([AMLJG19, IAGM19, ASM+20]).

For simplicity, we restrict our attention to monotone operators F arising as $F = F_{\mathcal{G}} : \mathbb{R}^n \to \mathbb{R}^n$ for a two-player zero-sum game $\mathcal{G}$ (i.e., the setting of min-max optimization). For simplicity suppose that n is even and for $\mathbf{z} \in \mathbb{R}^n$ write $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n/2}$. Define $\mathcal{F}^{\mathrm{bil}}_{n,\ell,D}$ to be the set of ℓ-Lipschitz operators $F : \mathbb{R}^n \to \mathbb{R}^n$ of the form $F(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))^\top$ for some bilinear function $f : \mathbb{R}^{n/2} \times \mathbb{R}^{n/2} \to \mathbb{R}$, with a unique equilibrium point $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$, which satisfies $\mathbf{z}^* \in \mathcal{D}_D := \mathcal{B}_{\mathbb{R}^{n/2}}(\mathbf{0}, D) \times \mathcal{B}_{\mathbb{R}^{n/2}}(\mathbf{0}, D)$. The following Theorem 7 uses functions in $\mathcal{F}^{\mathrm{bil}}_{n,\ell,D}$ as "hard instances" to show that the $O(1/\sqrt{T})$ rate of Corollary 5 cannot be improved by more than an *algorithm-dependent* constant factor.

Theorem 7 (Algorithm-dependent lower bound for p-SCLIs)**.** *Fix $\ell, D > 0$, let $\mathcal{A}$ be a p-SCLI, and let $\mathbf{z}^{(t)}$ denote the tth iterate of $\mathcal{A}$. Then there are constants $c_{\mathcal{A}}, T_{\mathcal{A}} > 0$ so that the following holds: For all $T \geq T_{\mathcal{A}}$, there is some $F \in \mathcal{F}^{\mathrm{bil}}_{n,\ell,D}$ so that for some initialization $\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(-p+1)} \in \mathcal{D}_D$ and $T' \in \{T, T+1, \ldots, T+p-1\}$, it holds that $\mathrm{TGap}_F^{\mathcal{D}_{2D}}(\mathbf{z}^{(T')}) \geq \frac{c_{\mathcal{A}} \ell D^2}{\sqrt{T}}$.*

We remark that the order of quantifiers in Theorem 7 is important: if instead we first fix a monotone operator $F \in \mathcal{F}^{\mathrm{bil}}_{n,\ell,D}$ corresponding to some bilinear function $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{M} \mathbf{y}$, then as shown in [LS18, Theorem 3], the iterates $\mathbf{z}^{(T)} = (\mathbf{x}^{(T)}, \mathbf{y}^{(T)})$ of the OG algorithm will converge at a rate of $e^{-O\left(\frac{\sigma_{\min}(\mathbf{M})^2}{\sigma_{\max}(\mathbf{M})^2} \cdot T\right)}$, which eventually becomes smaller than the sublinear rate of $1/\sqrt{T}$.[8] Such "instance-specific" bounds are complementary to the minimax perspective taken in this paper.

We briefly discuss the proof of Theorem 7; the full proof is deferred to Appendix C. As in prior work proving lower bounds for p-SCLIs ([ASSS15, IAGM19]), we reduce the problem of proving a lower bound on $\mathrm{TGap}_{\mathcal{G}}^{\mathcal{D}_D}(\mathbf{z}^{(t)})$ to the problem of proving a lower bound on the supremum of the spectral norms of a family of polynomials (which depends on $\mathcal{A}$). Recall that for a polynomial $p(z)$, its *spectral norm* $\rho(p(z))$ is the maximum norm of any root. We show:

Proposition 8. *Suppose $q(z)$ is a degree-p monic real polynomial such that $q(1) = 0$, $r(z)$ is a polynomial of degree $p-1$, and $\ell > 0$. Then there is a constant $C_0 > 0$, depending only on $q(z), r(z)$ and ℓ, and some $\mu_0 \in (0, \ell)$, so that for any $\mu \in (0, \mu_0)$,*

$$\sup_{\nu \in [\mu, \ell]} \rho(q(z) - \nu \cdot r(z)) \geq 1 - C_0 \cdot \frac{\mu}{\ell}.$$

[7] We use slightly different terminology from [ASSS15]; technically, the p-SCLIs considered in this paper are those in [ASSS15] with *linear coefficient matrices*.

[8] $\sigma_{\min}(\mathbf{M})$ and $\sigma_{\max}(\mathbf{M})$ denote the minimum and maximum singular values of $\mathbf{M}$, respectively. The matrix $\mathbf{M}$ is assumed in [LS18] to be a square matrix of full rank (which holds for the construction used to prove Theorem 7).

The proof of Proposition 8 uses elementary tools from complex analysis. The fact that the constant C_0 in Proposition 8 depends on $q(z), r(z)$ leads to the fact that the constants $c_{\mathcal{A}}, T_{\mathcal{A}}$ in Theorem 7 depend on $\mathcal{A}$. Moreover, we remark that this dependence cannot be improved from Proposition 8, so removing it from Theorem 7 will require new techniques:

Proposition 9 (Tightness of Proposition 8)**.** *For any constant $C_0 > 0$ and $\mu_0 \in (0, \ell)$, there is some $\mu \in (0, \mu_0)$ and polynomials $q(z), r(z)$ so that $\sup_{\nu \in [\mu, \ell]} \rho(q(z) - \nu \cdot r(z)) < 1 - C_0 \cdot \mu$. Moreover, the choice of the polynomials is given by*

$$q(z) = \ell(z - \alpha)(z - 1), \qquad r(z) = -(1 + \alpha)z + \alpha \qquad \text{for} \qquad \alpha := \frac{\sqrt{\ell} - \sqrt{\mu}}{\sqrt{\ell} + \sqrt{\mu}}. \tag{9}$$

The choice of polynomials $q(z), r(z)$ in (9) are exactly the polynomials that arise in the p-SCLI analysis of Nesterov's AGD [ASSS15]; as we discuss further in Appendix C, Proposition 8 is tight, then, even for $p = 2$, because acceleration is possible with a 2-SCLI. As byproducts of our lower bound analysis, we additionally obtain the following:

- Using Proposition 8, we show that any p-SCLI algorithm must have a rate of at least $\Omega_{\mathcal{A}}(1/T)$ for smooth convex function minimization (again, with an algorithm-dependent constant).[9] This is slower than the $O(1/T^2)$ error achievable with Nesterov's AGD with a time-varying learning rate.
- We give a direct proof of the following statement, which was conjectured by [ASSS15]: for polynomials q, r in the setting of Proposition 8, for any $0 < \mu < \ell$, there exists $\nu \in [\mu, \ell]$ so that $\rho(q(z) - \nu \cdot r(z)) \geq \frac{\sqrt{\ell/\mu} - 1}{\sqrt{\ell/\mu} + 1}$. Using this statement, for the setting of Theorem 7, we give a proof of an *algorithm-independent* lower bound $\mathrm{TGap}_F^{\mathcal{D}_D}(\mathbf{z}^{(t)}) \geq \Omega(\ell D^2/T)$. Though the algorithm-independent lower bound of $\Omega(\ell D^2/T)$ has already been established in the literature, even for non-stationary CLIs (e.g., [ASM+20, Proposition 5]), we give an alternative proof from existing approaches.

5 Discussion

In this paper we proved tight last-iterate convergence rates for smooth monotone games when all players act according to the optimistic gradient algorithm, which is no-regret. We believe that there are many fruitful directions for future research. First, it would be interesting to obtain last-iterate rates in the case that each player's actions is constrained to the simplex and they use the *optimistic multiplicative weights update (OMWU)* algorithm. [DP18, LNPW20] showed that OMWU exhibits last-iterate convergence, but non-asymptotic rates remain unknown even for the case that $F_{\mathcal{G}}(\cdot)$ is linear, which includes finite-action polymatrix games. Next, it would be interesting to determine whether Theorem 5 holds if (2) is removed from Assumption 4; this problem is open even for the EG algorithm ([GPDO20]). Finally, it would be interesting to extend our results to the setting where players receive noisy gradients (i.e., the stochastic case). As for lower bounds, it would be interesting to determine whether an algorithm-independent lower bound of $\Omega(1/\sqrt{T})$ in the context of Theorem 7 could be proven for stationary p-SCLIs. As far as we are aware, this question is open even for convex minimization (where the rate would be $\Omega(1/T)$).

Acknowledgements

We thank Yossi Arjevani for a helpful conversation.

References

[Ahl79] L.V. Ahlfors. *Complex Analysis*. McGraw-Hill, 1979.

[9] [AS16] claimed to prove a similar lower bound for stationary algorithms in the setting of smooth convex function minimization; however, as we discuss in Appendix C, their results only apply to the strongly convex case, where they show a linear lower bound.

[ALW19] 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.

[AMLJG19] 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.

[AS16] 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.

[ASM+20] Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating Smooth Games by Manipulating Spectral Shapes. *arXiv:2001.00602 [cs, math, stat]*, January 2020. arXiv: 2001.00602.

[ASSS15] 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.

[BCM12] Maria-Florina Balcan, Florin Constantin, and Ruta Mehta. The Weighted Majority Algorithm does not Converge in Nearly Zero-sum Games. In *ICML Workshop on Markets, Mechanisms, and Multi-Agent Models*, 2012.

[BEDL06] Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: on convergence to nash equilibria of regret-minimizing algorithms in routing games. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing - PODC '06*, page 45, Denver, Colorado, USA, 2006. ACM Press.

[BF87] LM Bregman and IN Fokin. Methods of determining equilibrium situations in zero-sum polymatrix games. *Optimizatsia*, 40(57):70–82, 1987.

[BG17] Nikhil Bansal and Anupam Gupta. Potential-Function Proofs for First-Order Methods. *arXiv:1712.04581 [cs, math]*, December 2017. arXiv: 1712.04581.

[BP18] James P. Bailey and Georgios Piliouras. Multiplicative Weights Update in Zero-Sum Games. In *Proceedings of the 2018 ACM Conference on Economics and Computation - EC '18*, pages 321–338, Ithaca, NY, USA, 2018. ACM Press.

[BTHK15] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *Journal of Artificial Intelligence Research*, 53:659–697, August 2015.

[CBL06] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge; New York, 2006. OCLC: 70056026.

[CCDP16] Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-Sum Polymatrix Games: A Generalization of Minmax. *Mathematics of Operations Research*, 41(2):648–655, May 2016.

[CD11] Yang Cai and Constantinos Daskalakis. On minmax theorems for multiplayer games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete algorithms*, pages 217–234. SIAM, 2011.

[CGFLJ19] 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.

[CL16] Po-An Chen and Chi-Jen Lu. Generalized mirror descents in congestion games. *Artificial Intelligence*, 241:217–243, December 2016.

[CYL+12] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online Optimization with Gradual Variations. In *Proceedings of the 25th Annual Conference on Learning Theory*, page 20, 2012.

[DFP+10] Constantinos Daskalakis, Rafael Frongillo, Christos H. Papadimitriou, George Pierrakos, and Gregory Valiant. On Learning Algorithms for Nash Equilibria. In *Algorithmic Game Theory*, volume 6386, pages 114–125. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[DISZ17] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. *arXiv:1711.00141 [cs, stat]*, October 2017. arXiv: 1711.00141.

[DP09] Constantinos Daskalakis and Christos H Papadimitriou. On a network generalization of the minmax theorem. In *International Colloquium on Automata, Languages, and Programming*, pages 423–434. Springer, 2009.

[DP18] 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.

[EDMN09] Eyal Even-Dar, Yishay Mansour, and Uri Nadav. On the convergence of regret minimization dynamics in concave games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 523–532, 2009.

[FOP20] Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An Optimal Multistage Stochastic Gradient Method for Minimax Problems. *arXiv:2002.05683 [cs, math, stat]*, February 2020. arXiv: 2002.05683.

[FP03] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer series in operations research. Springer, New York, 2003.

[GBV+18] 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 2018. arXiv: 1802.10551.

[GHP+18] 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 2018. arXiv: 1807.04740.

[GPDO20] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems. In *arXiv: 2002.00057*, 2020.

[HIMM19] 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.

[HIMM20] Yu-Guan Hsieh, Franck Iutzeler, Jerome Malick, and Panayotis Mertikopoulos. Explore Aggressively, Update Conservatively: Stochastic Extragradient Methods with Variable Stepsize Scaling. *arXiv:2003.10162*, page 27, 2020.

[HJ12] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge ; New York, 2nd ed edition, 2012.

[HMC03] Sergiu Hart and Andreu Mas-Colell. Uncoupled Dynamics Do Not Lead to Nash Equilibrium. *THE AMERICAN ECONOMIC REVIEW*, 93(5):7, 2003.

[IAGM19] 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.

[KBTB18] Walid Krichene, Mohamed Chedhli Bourguiba, Kiet Tlam, and Alexandre Bayen. On Learning How Players Learn: Estimation of Learning Dynamics in the Routing Game. *ACM Trans. Cyber-Phys. Syst.*, 2(1):6:1–6:23, January 2018.

[KKDB15] Syrine Krichene, Walid Krichene, Roy Dong, and Alexandre Bayen. Convergence of heterogeneous distributed learning in stochastic routing games. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 480–487, Monticello, IL, September 2015. IEEE.

[KLP11] Robert Kleinberg, Katrina Ligett, and Georgios Piliouras. Beyond the Nash Equilibrium Barrier. page 15, 2011.

[Kor76] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matem. Metody*, 12(4):747–756, 1976.

[Koz09] Victor Kozyakin. On accuracy of approximation of the spectral radius by the Gelfand formula. *Linear Algebra and its Applications*, 431:2134–2141, 2009.

[KPT09] Robert Kleinberg, Georgios Piliouras, and Eva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games: extended abstract. In *Proceedings of the 41st annual ACM symposium on Symposium on theory of computing - STOC '09*, page 533, Bethesda, MD, USA, 2009. ACM Press.

[KUS19] Aswin Kannan, Uday, and V. Shanbhag. Pseudomonotone Stochastic Variational Inequality Problems: Analysis and Optimal Stochastic Approximation Schemes. *Computational Optimization and Applications*, 74:669–820, 2019.

[LBJM+20] Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic Hamiltonian Gradient Methods for Smooth Games. *arXiv:2007.04202 [cs, math, stat]*, July 2020. arXiv: 2007.04202.

[LNPW20] Qi Lei, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. *arXiv:2002.06768 [cs, stat]*, February 2020. arXiv: 2002.06768.

[LS18] 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.

[LZMJ20] Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael I. Jordan. Finite-Time Last-Iterate Convergence for Multi-Agent Learning in Games. *arXiv:2002.09806 [cs, math, stat]*, February 2020. arXiv: 2002.09806.

[MKS+19] Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting Stochastic Extragradient. *arXiv:1905.11373 [cs, math]*, May 2019. arXiv: 1905.11373.

[MOP19a] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence Rate of $\mathcal{O}(1/k)$ for Optimistic Gradient and Extra-gradient Methods in Smooth Convex-Concave Saddle Point Problems. *arXiv:1906.01115 [cs, math, stat]*, June 2019. arXiv: 1906.01115.

[MOP19b] 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 2019. arXiv: 1901.08511.

[MP17] Barnabé Monnot and Georgios Piliouras. Limits and limitations of no-regret learning in games. *The Knowledge Engineering Review*, 32, 2017.

[MPP17] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. *arXiv:1709.02738 [cs]*, September 2017. arXiv: 1709.02738.

[MV78] H. Moulin and J. P. Vial. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *Int J Game Theory*, 7(3):201–221, September 1978.

[MZ18] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *arXiv:1608.07310 [cs, math]*, January 2018. arXiv: 1608.07310 version: 2.

[Nem04] 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.

[Nes75] M. C. Nesterov. *Introductory Lectures on Convex Programming*. North-Holland, 1975.

[Nes06] Yurii Nesterov. Cubic Regularization of Newton's Method for Convex Problems with Constraints. *SSRN Electronic Journal*, 2006.

[Nes09] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Math. Program.*, 120(1):221–259, August 2009.

[Nev93] Olavi Nevanlinna. *Convergence of Iterations for Linear Equations*. Birkhäuser Basel, Basel, 1993.

[OX19] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, August 2019.

[PB16] Balamurugan Palaniappan and Francis Bach. Stochastic Variance Reduction Methods for Saddle-Point Problems. In *Proceedings of the 30th International Conferencne on Neural Information Processing Systems*, pages 1416–1424, 2016.

[Pol87] Boris T. Polyak. *Introduction to optimization.*, volume 1. Optimization Software, 1987.

[Pop80] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, November 1980.

[PP14] Ioannis Panageas and Georgios Piliouras. Average Case Performance of Replicator Dynamics in Potential Games via Computing Regions of Attraction. *arXiv:1403.3885 [cs, math]*, 2014. arXiv: 1403.3885.

[PP16] Christos Papadimitriou and Georgios Piliouras. From Nash Equilibria to Chain Recurrent Sets: Solution Concepts and Topology. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science - ITCS '16*, pages 227–235, Cambridge, Massachusetts, USA, 2016. ACM Press.

[PPP17] Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative Weights Update with Constant Step-Size in Congestion Games: Convergence, Limit Cycles and Chaos. *arXiv:1703.01138 [cs]*, March 2017. arXiv: 1703.01138.

[Ros65] J.B. Rosen. Existence and Uniqueness of Equilibrium Points for Concave N-Person Games. *Econometrica*, 33(3):520–534, 1965.

[RS12] Alexander Rakhlin and Karthik Sridharan. Online Learning with Predictable Sequences. *arXiv:1208.3728 [cs, stat]*, August 2012. arXiv: 1208.3728.

[RS13] Alexander Rakhlin and Karthik Sridharan. Optimization, Learning, and Games with Predictable Sequences. *arXiv:1311.1869 [cs]*, November 2013. arXiv: 1311.1869.

[RVV16] Lorenzo Rosasco, Silvia Villa, and Bang Cong Vũ. A Stochastic forward-backward splitting method for solving monotone inclusions in Hilbert spaces. *Journal of Optimization Theory and Applications*, 169:388–406, 2016. arXiv: 1403.7999.

[SS11] Shai Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2011.

[Tse95] 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.

[VZ13] Yannick Viossat and Andriy Zapechelnyuk. No-regret Dynamics and Fictitious Play. *Journal of Economic Theory*, 148(2):825–842, March 2013. arXiv: 1207.0660.

[WRJ18] Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A Lyapunov Analysis of Momentum Methods in Optimization. *arXiv:1611.02635 [cs, math]*, March 2018. arXiv: 1611.02635.

[YSX+17] 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.

[ZMA+18] Zhengyuan Zhou, Panayotis Mertikopoulos, Susan Athey, Nicholas Bambos, Peter W Glynn, and Yinyu Ye. Learning in Games with Lossy Feedback. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5134–5144. Curran Associates, Inc., 2018.

[ZMB+17] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter W Glynn, and Claire Tomlin. Countering Feedback Delays in Multi-Agent Learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6171–6181. Curran Associates, Inc., 2017.

[ZMM+17] Zhengyuan Zhou, Panayotis Mertikopoulos, Aris L. Moustakas, Nicholas Bambos, and Peter Glynn. Mirror descent learning in continuous games. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5776–5783, Melbourne, Australia, December 2017. IEEE.

[ZMM+20] Zhengyuan Zhou, Panayotis Mertikopoulos, Aris L Moustakas, Nicholas Bambos, and Peter Glynn. Robust Power Management via Learning and Game Design. *Operations Research*, 2020.

A Additional preliminaries

A.1 Proof of Proposition 2

Proof of Proposition 2. Fix a game $\mathcal{G}$, and let $F = F_{\mathcal{G}} : \mathcal{Z} \to \mathbb{R}^n$. Monoticity of F gives that for any fixed $\mathbf{z}_{-k} \in \prod_{k' \neq k} \mathcal{Z}_{k'}$, for any $\mathbf{z}_k, \mathbf{z}'_k \in \mathcal{Z}_k$, we have

$$\langle F(\mathbf{z}'_k, \mathbf{z}_{-k}) - F(\mathbf{z}_k, \mathbf{z}_{-k}), (\mathbf{z}'_k, \mathbf{z}_{-k}) - (\mathbf{z}_k, \mathbf{z}_{-k}) \rangle = \langle \nabla_{\mathbf{z}_k} f_k(\mathbf{z}'_k, \mathbf{z}_{-k}) - \nabla_{\mathbf{z}_k}(\mathbf{z}_k, \mathbf{z}_{-k}), \mathbf{z}'_k - \mathbf{z}_k \rangle \geq 0.$$

Since f_k is continuously differentiable, [Nes75, Theorem 2.1.3] gives that f_k is convex. Thus

$$f_k(\mathbf{z}_k, \mathbf{z}_{-k}) - \min_{\mathbf{z}'_k \in \mathcal{Z}'_k} f_k(\mathbf{z}'_k, \mathbf{z}_{-k}) \leq \langle \nabla_{\mathbf{z}_k} f_k(\mathbf{z}_k, \mathbf{z}_{-k}), \mathbf{z}_k - \mathbf{z}'_k \rangle \leq \|\nabla_{\mathbf{z}_k} f_k(\mathbf{z}_k, \mathbf{z}_{-k})\| \cdot D.$$

Summing the above for $k \in \mathcal{K}$ and using the definition of the total and gradient gap functions, as well as Cauch-Schwarz, gives that $\mathrm{TGap}_{\mathcal{G}}^{\mathcal{Z}'}(\mathbf{z}) \leq D \cdot \sum_{k=1}^{K} \|\nabla_{\mathbf{z}_k} f_k(\mathbf{z})\| \leq D\sqrt{K}\|F(\mathbf{z})\|$. ☐

A.2 Optimistic gradient algorithm

In this section we review some additional background about the optimistic gradient algorithm in the setting of no-regret learning. The starting point is *online gradient descent*; player k following online gradient descent produces iterates $\mathbf{z}_k^{(t)} \in \mathcal{Z}_k$ defined by $\mathbf{z}_k^{(t+1)} = \mathbf{z}_k^{(t)} - \eta_t \mathbf{g}_k^{(t)}$, where $\mathbf{g}_k^{(t)} = \nabla_{\mathbf{z}_k} f_k(\mathbf{z}_k^{(t)}, \mathbf{z}_{-k}^{(t)})$ is player k's gradient given its action $\mathbf{z}_k^{(t)}$ and the other players' actions $\mathbf{z}_{-k}^{(t)}$ at time t. Online gradient descent is a no-regret algorithm (in particular, it satisfies the same regret bound as OG in Proposition 3); it is also closely related to the *follow-the-regularized-leader (FTRL)* ([SS11]) algorithm from online learning.[10]

The *optimistic gradient (OG)* algorithm ([RS13, DISZ17]) is a modification of online gradient descent, for which player k performs the following update:

$$\mathbf{z}_k^{(t+1)} := \mathbf{z}_k^{(t)} - 2\eta_t \mathbf{g}_k^{(t)} + \eta_t \mathbf{g}_k^{(t-1)}, \tag{OG}$$

where again $\mathbf{g}_k^{(t)} = \nabla_{\mathbf{z}_k} f_k(\mathbf{z}_k^{(t)}, \mathbf{z}_{-k}^{(t)})$ for $t \geq 0$. As way of intuition behind the updates (OG), [DISZ17] observed that OG is closely related to the *optimistic follow-the-regularized-leader (OFTRL)* algorithm from online learning: OFTRL augments the standard FTRL update by using the gradient $\mathbf{g}_k^{(t)}$ at time t as a prediction for the gradient at time $t+1$. When the actions $\mathbf{z}_{-k}^{(t)}$ of the other players are predictable in the sense that they do not change quickly over time, then such a prediction using $\mathbf{g}_k^{(t)}$ is reasonably accurate and can improve the speed of convergence to an equilibrium ([RS13]).

A.3 Linear regret for extragradient algorithm

In this section we review the definition of the extragradient (EG) algorithm, and show that if one attempts to implement it in the setting of online multi-agent learning, then it is not a no-regret algorithm. Given a monotone game $\mathcal{G}$ and its corresponding monotone operator $F_{\mathcal{G}} : \mathcal{Z} \to \mathbb{R}^n$ and an initial point $\mathbf{u}^{(0)} \in \mathbb{R}^n$ the EG algorithm attempts to find a Nash equilibrium $\mathbf{z}^*$ (i.e., a point satisfying $F_{\mathcal{G}}(\mathbf{z}^*) = 0$) by performing the updates:

$$\mathbf{u}^{(t)} = \Pi_{\mathcal{Z}}(\mathbf{u}^{(t-1)} - \eta F_{\mathcal{G}}(\mathbf{z}^{(t-1)})), \qquad t \geq 1 \tag{10}$$

$$\mathbf{z}^{(t)} = \Pi_{\mathcal{Z}}(\mathbf{u}^{(t)} - \eta F_{\mathcal{G}}(\mathbf{u}^{(t)})), \qquad t \geq 0, \tag{11}$$

where $\Pi_{\mathcal{Z}}(\cdot)$ denotes Euclidean projection onto the convex set $\mathcal{Z}$. Assuming $\mathcal{Z}$ contains a sufficiently large ball centered at $\mathbf{z}^*$, this projection step has no effect for the updates shown above when all players perform EG updates (see Remark 4); the projection is typically needed, however, for the adversarial setting that we proceed to discuss in this section (e.g., as in Proposition 3).

It is easy to see that the updates (10) and (11) can be rewritten as $\mathbf{u}^{(t)} = \Pi_{\mathcal{Z}}(\mathbf{u}^{(t-1)} - \eta F_{\mathcal{G}}(\Pi_{\mathcal{Z}}(\mathbf{u}^{(t-1)} - \eta F_{\mathcal{G}}(\mathbf{u}^{(t-1)}))))$. Note that these updates are somewhat similar to those of OG when expressed as (23) and

[10] In particular, they are equivalent in the unconstrained setting when the learning rate η_t is constant.

(24), with $\mathbf{w}^{(t)}$ in (23) and (24) playing a similar role to $\mathbf{u}^{(t)}$ in (10) and (11). A key difference is that the iterate $\mathbf{u}^{(t)}$ is needed to update $\mathbf{z}^{(t)}$ in (11), whereas this is not true for the update to $\mathbf{z}^{(t)}$ in (23). Since in the standard setting of online multi-agent learning, agents can only see gradients corresponding to actions they play, in order to implement the above EG updates in this setting, we need two timesteps for every timestep of EG. In particular, the agents will play actions $\mathbf{v}^{(t)}$, $t \geq 0$, where $\mathbf{v}^{(2t)} = \mathbf{u}^{(t)}$ and $\mathbf{v}^{(2t+1)} = \mathbf{z}^{(t)}$ for all $t \geq 0$. Recalling that $F_{\mathcal{G}}(\mathbf{z}) = (\nabla_{\mathbf{z}_1} f_1(\mathbf{z}), \ldots, \nabla_{\mathbf{z}_K} f_K(\mathbf{z}))$, this means that player $k \in [K]$ performs the updates

$$\mathbf{v}_k^{(2t)} = \Pi_{\mathcal{Z}_k}(\mathbf{v}_k^{(2t-2)} - \eta \nabla_{\mathbf{z}_k} f_k(\mathbf{v}_k^{(2t-1)}, \mathbf{z}_{-k}^{(2t-1)})), \qquad t \geq 1 \tag{12}$$

$$\mathbf{v}_k^{(2t+1)} = \Pi_{\mathcal{Z}_k}(\mathbf{v}_k^{(2t)} - \eta \nabla_{\mathbf{z}_k} f_k(\mathbf{v}_k^{(2t)}, \mathbf{v}_{-k}^{(2t)})), \qquad t \geq 0, \tag{13}$$

where $\mathbf{v}_k^{(0)} = \mathbf{u}_k^{(0)}$. Unfortunately, as we show in Proposition 10 below, in the setting when the other players' actions $\mathbf{z}_{-k}^{(t)}$ are adversarial (i.e., players apart from k do not necessarily play according to EG), the algorithm for player k given by the EG updates (12) and (13) can have linear regret, i.e., is not a no-regret algorithm. Thus the EG algorithm is insufficient for answering our motivating question ($\star$).

Proposition 10. *There is a set $\mathcal{Z} = \prod_{k=1}^K \mathcal{Z}_k$ together with a convex, 1-Lipschitz, and 1-smooth function $f_1 : \mathcal{Z} \to \mathbb{R}$ so that for an adversarial choice of $\mathbf{z}_{-k}^{(t)}$, the EG updates (12) and (13) produce a sequence $\mathbf{v}_k^{(t)}$, $0 \leq t \leq T$ with regret $\Omega(T)$ with respect to the sequence of functions $\mathbf{v}_k \mapsto f_k(\mathbf{v}_k, \mathbf{v}_{-k}^{(t)})$ for any $T > 0$.*

Proof. We take $K = 1, k = 1, n = 2, \mathcal{Z}_1 = \mathcal{Z}_2 = [-1, 1]$, and $f_1 : \mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathbb{R}$ to be $f_1(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{v}_1 \cdot \mathbf{v}_2$, where $\mathbf{v}_1, \mathbf{v}_2 \in [-1, 1]$. Consider the following sequence of actions $\mathbf{v}_2^{(t)}$ of player 2:

$$\mathbf{v}_2^{(t)} = 1 \text{ for } t \text{ even}; \qquad \mathbf{v}_2^{(t)} = 0 \text{ for } t \text{ odd}.$$

Suppose that player 1 initializes at $\mathbf{v}_1^{(0)} = 0$. Then for all $t \geq 0$, we have

$$\nabla_{\mathbf{z}_1} f_1(\mathbf{v}_1^{(2t-1)}, \mathbf{v}_2^{(2t-1)}) = \mathbf{v}_2^{(2t-1)} = 0 \qquad \forall t \geq 1$$
$$\nabla_{\mathbf{z}_1} f_1(\mathbf{z}_1^{(2t)}, \mathbf{v}_2^{(2t)}) = \mathbf{v}_2^{(2t)} = 1 \qquad \forall t \geq 0.$$

It follows that for $t \geq 0$ we have $\mathbf{v}_1^{(2t)} = 0$ and $\mathbf{v}_1^{(2t+1)} = \max\{-\eta, -1\}$. Hence for any $T \geq 0$ we have $\sum_{t=0}^{T-1} f_1(\mathbf{v}_1^{(t)}, \mathbf{v}_2^{(t)}) = 0$ whereas

$$\min_{\mathbf{v}_1 \in \mathcal{Z}_1} \sum_{t=0}^{T-1} f_1(\mathbf{v}_1, \mathbf{v}_2^{(t)}) = -\lceil T/2 \rceil,$$

(with the optimal point $\mathbf{v}_1$ being $\mathbf{v}_1^* = -1$) so the regret is $\lceil T/2 \rceil$. ∎

A.4 Prior work on last-iterate rates for noisy feedback

In this section we present Table 2, which exhibits existing last-iterate convergence rates for gradient-based learning algorithms in the case of noisy gradient feedback (i.e., it is an analogue of Table 1 for noisy feedback, leading to stochastic algorithms). We briefly review the setting of noisy feedback: at each time step t, each player k plays an action $\mathbf{z}_k^{(t)}$, and receives the feedback

$$\mathbf{g}_k^{(t)} := \nabla_{\mathbf{z}_k} f_k(\mathbf{z}_k^{(t)}, \mathbf{z}_{-k}^{(t)}) + \xi_k^{(t)},$$

where $\xi_k^{(t)} \in \mathbb{R}^{n_k}$ is a random variable satisfying:

$$\mathbb{E}[\xi_k^{(t)} | \mathcal{F}^{(t)}] = \mathbf{0}, \tag{14}$$

where $\mathcal{F} = (\mathcal{F}^{(t)})_{t \geq 0}$ is the filtration given by the sequence of σ-algebras $\mathcal{F}^{(t)} := \sigma(\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(t)})$ generated by $\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(t)}$. Additionally, it is required that the variance of $\xi_k^{(t)}$ be bounded; we focus on

the following two possible boundedness assumptions:

$$\mathbb{E}[\|\xi_k^{(t)}\|^2|\mathcal{F}^{(t)}] \le \sigma_t^2 \tag{Abs}$$

$$\text{or} \qquad \mathbb{E}[\|\xi_k^{(t)}\|^2|\mathcal{F}^{(t)}] \le \tau_t\|F_{\mathcal{G}}(\mathbf{z}^{(t)})\|^2, \tag{Rel}$$

where $\sigma_t > 0$ and $\tau_t > 0$ are sequences of positive reals (typically taken to be decreasing with t). Often it is assumed that σ_t is the same for all t, in which case we write $\sigma = \sigma_t$. Noise model (Abs) is known as *absolute random noise*, and (Rel) is known as *relative random noise* [LZMJ20]. The latter is only of use in the unconstrained setting in which the goal is to find $\mathbf{z}^*$ with $F_{\mathcal{G}}(\mathbf{z}^*) = \mathbf{0}$. While we restrict Table 2 to 1st order methods, we refer the reader also to the recent work of [LBJM+20], which provides last-iterate rates for stochastic Hamiltonian gradient descent, a 2nd order method, in "sufficiently bilinear" games.

As can be seen in Table 2, there is no work to date proving last-iterate rates for general smooth monotone games. We view the problem of extending the results of this paper and of [GPDO20] to the stochastic setting (i.e., the bottom row of Table 2) as an interesting direction for future work.

Table 2: Known upper bounds on last-iterate convergence rates for learning in smooth monotone games with noisy gradient feedback (i.e., *stochastic* algorithms). Rows of the table are as in Table 1; ℓ, Λ are the Lipschitz constants of $F_{\mathcal{G}}, \partial F_{\mathcal{G}}$, respectively, and $c > 0$ is a sufficiently small absolute constant. The right-hand column contains algorithms implementable as online no-regret learning algorithms: stochastic optimistic gradient (Stoch. OG) or stochastic gradient descent (SGD). The left-hand column contains algorithms not implementable as no-regret algorithms, which includes stochastic extragradient (Stoch. EG), stochastic forward-backward (FB) splitting, double stepsize extragradient (DSEG), and stochastic variance reduced extragradient (SVRE). SVRE only applies in the *finite-sum setting*, which is a special case of (Abs) in which f_k is a sum of m individual loss functions $f_{k,i}$, and a noisy gradient is obtained as $\nabla f_{k,i}$ for a random $i \in [m]$. Due to the stochasticity, many prior works make use of a step size η_t that decreases with t; we make note of whether this is the case (*"η_t decr."*) or whether the step size η_t can be constant (*"η_t const."*). For simplicity of presentation we assume $\Omega(1/t) \le \{\tau_t, \sigma_t\} \le O(1)$ for all $t \ge 0$ in all cases for which σ_t, τ_t vary with t. Reported bounds are stated for the total gap function (Definition 3); leading constants and factors depending on distance between initialization and optimum are omitted.

	Stochastic	
Game class	Not implementable as no-regret	Implementable as no-regret
μ-strongly monotone	*(Abs):* $\frac{\sigma\ell}{\mu\sqrt{T}}$ [PB16, Stoch. FB splitting, η_t decr.] (See also [RVV16, MKS+19]) *(Abs):* $\frac{\ell(\sigma+\ell)}{\mu\sqrt{T}}$ [KUS19, Stoch. EG, η_t decr.] *Finite-sum:* $\ell\left(1 - c\min\{\frac{1}{m}, \frac{\mu}{\ell}\}\right)^T$ [CGFLJ19, SVRE, η_t const.] (See also [PB16])	*(Abs):* $\frac{\sigma\ell}{\mu\sqrt{T}}$ [HIMM19, Stoch. OG, η_t decr.] (See also [FOP20])
Monotone, γ-sing. val. low. bnd.	*(Abs), (Rel):* Stoch. EG may not convg. [CGFLJ19, HIMM20] *(Abs):* $\frac{\ell^2\sigma}{\gamma^{3/2}\sqrt[6]{T}}$ [HIMM20, DSEG, η_t decr.]	Open
λ-cocoercive	Open	*(Rel):* $\frac{1}{\lambda\sqrt{T}} + \sqrt{\frac{\sum_{t\le T}\tau_t}{T}}$ [LZMJ20, SGD, η_t const.] *(Abs):* $\frac{\sqrt{\sum_{t\le t}(t+1)\sigma_t^2}}{\lambda\sqrt{T}}$ [LZMJ20, SGD, η_t const.]
Monotone	Open	Open

B Proofs for Section 3

In this section we prove Theorem 5. In Section B.1 we show that OG exhibits *best-iterate convergence*, which is a simple consequence of prior work. In Section B.1 we begin to work towards the main contribution of this

work, namely showing that best-iterate convergence implies last iterate convergence, treating the special case of linear monotone operators $F(\mathbf{z}) = \mathbf{A}\mathbf{z}$. In Section B.3 we introduce the adaptive potential function for the case of general smooth monotone operators F, and finally in Section B.4, using this choice of adaptive potential function, we prove Theorem 5. Some minor lemmas used throughout the proof are deferred to Section B.5.

B.1 Best-iterate convergence

Throughout this section, fix a monotone game $\mathcal{G}$ satisfying Assumption 4, and write $F = F_{\mathcal{G}}$, so that F is a monotone operator (Definition 1). Recall that the OG algorithm with constant step size $\eta > 0$ is given by:

$$\mathbf{z}^{(-1)}, \mathbf{z}^{(0)} \in \mathbb{R}^n, \qquad \mathbf{z}^{(t+1)} = \mathbf{z}^{(t)} - 2\eta F(\mathbf{z}^{(t)}) + \eta F(\mathbf{z}^{(t-1)}) \ \ \forall t \geq 0. \tag{15}$$

In Lemma 11 we observe that *some* iterate $\mathbf{z}^{(t^*)}$ of OG has small gradient gap.

Lemma 11. *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone operator that is ℓ-Lipschitz. Fix some $\mathbf{z}^{(0)}, \mathbf{z}^{(-1)} \in \mathbb{R}^n$, and suppose there is $\mathbf{z}^* \in \mathbb{R}^n$ so that $F(\mathbf{z}^*) = 0$ and $\max\{\|\mathbf{z}^* - \mathbf{z}^{(0)}\|, \|\mathbf{z}^* - \mathbf{z}^{(-1)}\| \leq D$. Then the iterates $\mathbf{z}^{(t)}$ of OG for any $\eta < \frac{1}{\ell\sqrt{10}}$ satisfy:*

$$\min_{0 \leq t \leq T-1} \|F(\mathbf{z}^{(t)})\| \leq \frac{4D}{\eta\sqrt{T} \cdot \sqrt{1 - 10\eta^2\ell^2}}. \tag{16}$$

More generally, we have, for any $S \geq 0$ with $S < T/3$,

$$\min_{0 \leq t \leq T-S} \max_{0 \leq s < S} \|F(\mathbf{z}^{(t+s)})\| \leq \frac{6D}{\eta\sqrt{T/S} \cdot \sqrt{1 - 10\eta^2\ell^2}}. \tag{17}$$

Proof. For all $t \geq 1$, define $\mathbf{w}^{(t)} = \mathbf{z}^{(t)} + \eta F(\mathbf{z}^{(t-1)})$. Equation (B.4) of [HIMM19] gives that for each $t \geq 0$, $\mathbf{z} \in \mathbb{R}^n$

$$\|\mathbf{w}^{(t+1)} - \mathbf{z}\|^2 \leq \|\mathbf{w}^{(t)} - \mathbf{z}\|^2 - 2\eta\langle F(\mathbf{z}^{(t)}), \mathbf{z}^{(t)} - \mathbf{z}\rangle + \eta^2\ell^2\|\mathbf{z}^{(t)} - \mathbf{z}^{(t-1)}\|^2 - \|\eta F(\mathbf{z}^{(t-1)})\|^2.$$

Choosing $\mathbf{z} = \mathbf{z}^*$, using that $\langle F(\mathbf{z}^{(t)}), \mathbf{z}^{(t)} - \mathbf{z}^*\rangle \geq 0$, and applying Young's inequality gives that for $t \geq 1$,

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{z}^*\|^2 \leq& \|\mathbf{w}^{(t)} - \mathbf{z}^*\|^2 + \eta^2\ell^2\|2\eta F(\mathbf{z}^{(t-1)}) - \eta F(\mathbf{z}^{(t-2)})\|^2 - \|\eta F(\mathbf{z}^{(t-1)})\|^2 \\ \leq& \|\mathbf{w}^{(t)} - \mathbf{z}^*\|^2 + (\eta^2\ell^2) \cdot 8\eta^2\|F(\mathbf{z}^{(t-1)})\|^2 + (\eta^2\ell^2) \cdot 2\eta^2\|F(\mathbf{z}^{(t-2)})\|^2 - \eta^2\|F(\mathbf{z}^{(t-1)})\|^2.\end{aligned}$$

Summing the above equation for $1 \leq t \leq T-1$ gives

$$\eta^2 \cdot \left((1 - 8\eta^2\ell^2)\sum_{t=0}^{T-2}\|F(\mathbf{z}^{(t)})\|^2 - 2\eta^2\ell^2\sum_{t=-1}^{T-3}\|F(\mathbf{z}^{(t)})\|^2\right) \leq \|\mathbf{w}^{(1)} - \mathbf{z}^*\|^2 - \|\mathbf{w}^{(T-1)} - \mathbf{z}^*\|^2.$$

Since $\|\mathbf{w}^{(1)} - \mathbf{z}^*\| \leq 3D$, $\|F(\mathbf{z}^{(-1)})\| \leq D\ell$, and $2\eta^2\ell^2 \leq 1$, it follows that

$$\min_{0 \leq t \leq T-2} \|F(\mathbf{z}^{(t)})\| \leq \frac{4D}{\eta\sqrt{T-1} \cdot \sqrt{1 - 10\eta^2\ell^2}}.$$

The desired result (16) follows by substituting $T+1$ for T.

To obtain (17), we break $\{0, 1, \ldots, T-2\}$ into $\lfloor (T-1)/S \rfloor$ windows of S consecutive time steps each. Then there must be some $t \in \{0, \ldots, T-2-(S-1)\}$ so that

$$\sum_{s=0}^{S-1}\|F(\mathbf{z}^{(t+s)})\|^2 \leq \frac{(4D)^2}{\eta^2(1 - 10\eta^2\ell^2)\lfloor (T-1)/S \rfloor},$$

from which (17) follows since $S < T/3$. ∎

In the remainder of this section we present our main technical contribution in the context of Theorem 5, showing that for a fixed T, the last iterate $\mathbf{z}^{(T)}$ does not have gradient gap $\|F(\mathbf{z}^{(T)})\|$ much larger than $\min_{1 \leq t \leq T} \max_{0 \leq s \leq 2} \|F(\mathbf{z}^{(t+s)})\|$.

B.2 Warm-up: different perspective on the linear case

Before treating the case where F is a general smooth monotone operator, we first explain our proof technique for the case that $F(\mathbf{z}) = \mathbf{A}\mathbf{z}$ for some matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. This case is covered by [LS18, Theorem 3][11]; the discussion here can be viewed as an alternative perspective on this prior work.

Assume that $F(\mathbf{z}) = \mathbf{A}\mathbf{z}$ for some $\mathbf{A} \in \mathbb{R}^{n\times n}$ throughout this section. Let $\mathbf{z}^{(t)}$ be the iterates of OG, and define

$$\mathbf{w}^{(t)} = \mathbf{z}^{(t)} + \eta F(\mathbf{z}^{(t-1)}) = \mathbf{z}^{(t)} + \eta\mathbf{A}\mathbf{z}^{(t-1)}. \tag{18}$$

Thus the updates of OG can be written as

$$\mathbf{z}^{(t)} = \mathbf{w}^{(t)} - \eta F(\mathbf{z}^{(t-1)}) = \mathbf{w}^{(t)} - \eta\mathbf{A}\mathbf{z}^{(t-1)} \tag{19}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta F(\mathbf{z}^{(t)}) = \mathbf{w}^{(t)} - \eta\mathbf{A}\mathbf{z}^{(t)}. \tag{20}$$

The *extra-gradient* (EG) algorithm is the same as the updates (19), (20), except that in (19), $F(\mathbf{z}^{(t-1)})$ is replaced with $F(\mathbf{w}_t)$. As such, OG in this context is often referred to as *past extragradient (PEG)* [HIMM19]. Many other works have also made use of this interpretation of OG, e.g., [RS12, RS13, Pop80].

Now define

$$\mathbf{C} = \frac{(I + (2\eta\mathbf{A})^2)^{1/2} - I}{2} = \eta^2\mathbf{A}^2 + O((\eta\mathbf{A})^4), \tag{21}$$

where the square root of $I + (2\eta\mathbf{A})^2$ may be defined via the power series $\sqrt{I - \mathbf{X}} := \sum_{j=0}^{\infty} \mathbf{X}^k(-1)^k\binom{1/2}{k}$. It is easy to check that $\mathbf{C}$ is well-defined as long as $\eta \le O(1/\ell) \le O(1/\|\mathbf{A}\|_\sigma)$, and that $\mathbf{C}\mathbf{A} = \mathbf{A}\mathbf{C}$. Also note that $\mathbf{C}$ satisfies

$$\mathbf{C}^2 + \mathbf{C} = \eta^2\mathbf{A}^2. \tag{22}$$

Finally set

$$\tilde{\mathbf{w}}^{(t)} = \mathbf{w}^{(t)} + \mathbf{C}\mathbf{z}^{(t-1)},$$

so that $\tilde{\mathbf{w}}^{(t)}$ corresponds (under the PEG interpretation of OG) to the iterates $\mathbf{w}^{(t)}$ of EG, plus an "adjustment" term, $\mathbf{C}\mathbf{z}^{(t)}$, which is $O((\eta\mathbf{A})^2)$. Though this adjustment term is small, it is crucial in the following calculation:

$$\begin{aligned}
\tilde{\mathbf{w}}^{(t+1)} &= \mathbf{w}^{(t+1)} + \mathbf{C}\mathbf{z}^{(t)} \\
&=^{(20)} \mathbf{w}^{(t)} - \eta\mathbf{A}\mathbf{z}^{(t)} + \mathbf{C}\mathbf{z}^{(t)} \\
&=^{(19)} \mathbf{w}^{(t)} + (\mathbf{C} - \eta\mathbf{A})(\mathbf{w}^{(t)} - \eta\mathbf{A}\mathbf{z}^{(t-1)}) \\
&= (I - \eta\mathbf{A} + \mathbf{C})\mathbf{w}^{(t)} + (\eta^2\mathbf{A}^2 - \eta\mathbf{A}\mathbf{C})\mathbf{z}^{(t-1)} \\
&=^{(22)} (I - \eta\mathbf{A} + \mathbf{C})(\mathbf{w}^{(t)} + \mathbf{C}\mathbf{z}^{(t-1)}) \\
&= (I - \eta\mathbf{A} + \mathbf{C})\tilde{\mathbf{w}}^{(t)}.
\end{aligned}$$

Since $\mathbf{C}, \mathbf{A}$ commute, the above implies that $F(\tilde{\mathbf{w}}^{(t+1)}) = (I - \eta\mathbf{A} + \mathbf{C})F(\tilde{\mathbf{w}}^{(t)})$. Monotonicity of F implies that for $\eta = O(1/\ell)$, we have $\|I - \eta\mathbf{A} + \mathbf{C}\|_\sigma \le 1$. It then follows that $\|F(\tilde{\mathbf{w}}^{(t+1)})\| \le \|F(\tilde{\mathbf{w}}^{(t)})\|$, which establishes that the last iterate is the best iterate.

B.3 Setting up the adaptive potential function

We next extend the argument of the previous section to the smooth convex-concave case, which will allow us to prove Theorem 5 in its full generality. Recall the PEG formulation of OG introduced in the previous section:

$$\mathbf{z}^{(t)} = \mathbf{w}^{(t)} - \eta F(\mathbf{z}^{(t-1)}) \tag{23}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta F(\mathbf{z}^{(t)}), \tag{24}$$

[11] Technically, [LS18] only considered the case where $\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{M} \\ -\mathbf{M}^\top & \mathbf{0} \end{pmatrix}$ for some matrix $\mathbf{M}$, which corresponds to min-max optimization for bilinear functions, but their proof readily extends to the case we consider in this section.

where again $\mathbf{z}^{(t)}$ denote the iterates of OG (15).

As discussed in Section 3.1, the adaptive potential function is given by $\|\tilde{F}^{(t)}\|$, where

$$\tilde{F}^{(t)} := F(\mathbf{w}^{(t)}) + \mathbf{C}^{(t-1)} \cdot F(\mathbf{z}^{(t-1)}) \in \mathbb{R}^n, \tag{25}$$

for some matrices $\mathbf{C}^{(t)} \in \mathbb{R}^{n\times n}$, $-1 \le t \le T$, to be chosen later. Then:

$$\begin{aligned}
\tilde{F}^{(t+1)} &= F(\mathbf{w}^{(t+1)}) + \mathbf{C}^{(t)} \cdot F(\mathbf{z}^{(t)}) \\
&=^{(24)} F(\mathbf{w}^{(t)} - \eta F(\mathbf{z}^{(t)})) + \mathbf{C}^{(t)} \cdot F(\mathbf{z}^{(t)}) \\
&= F(\mathbf{w}^{(t)}) - \eta \mathbf{A}^{(t)} F(\mathbf{z}^{(t)}) + \mathbf{C}^{(t)} \cdot F(\mathbf{z}^{(t)}) \\
&=^{(23)} F(\mathbf{w}^{(t)}) + (\mathbf{C}^{(t)} - \eta \mathbf{A}^{(t)}) \cdot F(\mathbf{w}^{(t)} - \eta F(\mathbf{z}^{(t-1)})) \\
&= F(\mathbf{w}^{(t)}) + (\mathbf{C}^{(t)} - \eta \mathbf{A}^{(t)}) \cdot (F(\mathbf{w}^{(t)}) - \eta \mathbf{B}^{(t)} F(\mathbf{z}^{(t-1)})) \\
&= (I - \eta \mathbf{A}^{(t)} + \mathbf{C}^{(t)}) \cdot F(\mathbf{w}^{(t)}) + \eta(\eta \mathbf{A}^{(t)} - \mathbf{C}^{(t)}) \mathbf{B}^{(t)} \cdot F(\mathbf{z}^{(t-1)}),
\end{aligned} \tag{26}$$

where

$$\begin{aligned}
\mathbf{A}^{(t)} &:= \int_0^1 \partial F(\mathbf{w}^{(t)} - (1-\alpha)\eta F(\mathbf{z}^{(t)}))d\alpha \\
\mathbf{B}^{(t)} &:= \int_0^1 \partial F(\mathbf{w}^{(t)} - (1-\alpha)\eta F(\mathbf{z}^{(t-1)}))d\alpha.
\end{aligned}$$

(Recall that $\partial F(\cdot)$ denotes the Jacobian of F.) We state the following lemma for later use:

Lemma 12. *For each t, $\mathbf{A}^{(t)} + (\mathbf{A}^{(t)})^\top, \mathbf{B}^{(t)} + (\mathbf{B}^{(t)})^\top$ are PSD, and $\|\mathbf{A}^{(t)}\|_\sigma \le \ell, \|\mathbf{B}^{(t)}\|_\sigma \le \ell$. Moreover, it holds that*

$$\begin{aligned}
\|\mathbf{A}^{(t)} - \mathbf{B}^{(t)}\|_\sigma &\le \frac{\eta\Lambda}{2}\|F(\mathbf{z}^{(t)}) - F(\mathbf{z}^{(t-1)})\| \\
\|\mathbf{A}^{(t)} - \mathbf{A}^{(t+1)}\|_\sigma &\le \Lambda\|\mathbf{w}^{(t)} - \mathbf{w}^{(t+1)}\| + \frac{\eta\Lambda}{2}\|F(\mathbf{z}^{(t)}) - F(\mathbf{z}^{(t+1)})\| \\
\|\mathbf{B}^{(t)} - \mathbf{B}^{(t+1)}\|_\sigma &\le \Lambda\|\mathbf{w}^{(t)} - \mathbf{w}^{(t+1)}\| + \frac{\eta\Lambda}{2}\|F(\mathbf{z}^{(t-1)}) - F(\mathbf{z}^{(t)})\|.
\end{aligned}$$

Proof. For all $\mathbf{z} \in \mathbb{R}^n$, monotonicity of F gives that $\partial F(\mathbf{z}) + \partial F(\mathbf{z})^\top$ is PSD, which means that so are $\mathbf{A}^{(t)} + (\mathbf{A}^{(t)})^\top, \mathbf{B}^{(t)} + (\mathbf{B}^{(t)})^\top$. Similarly, (1) gives that for all $\mathbf{z} \in \mathbb{R}^n$, $\|\partial F(\mathbf{z})\|_\sigma \le \ell$, from which we get $\|\mathbf{A}^{(t)}\|_\sigma \le \ell, \|\mathbf{B}^{(t)}\|_\sigma \le \ell$ by the triangle inequality.

The remaining three inequalities are an immediate consequence of the triangle inequality and the fact that ∂F is Λ-Lipschitz (Assumption 4). ☐

Now define the following $n \times n$ matrices:

$$\begin{aligned}
\mathbf{M}^{(t)} &:= I - \eta \mathbf{A}^{(t)} + \mathbf{C}^{(t)} \\
\mathbf{N}^{(t)} &:= \eta(\eta \mathbf{A}^{(t)} - \mathbf{C}^{(t)}) \mathbf{B}^{(t)}.
\end{aligned}$$

Moreover, for a positive semidefinite (PSD) matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{v} \in \mathbb{R}^n$, write $\|\mathbf{v}\|_{\mathbf{S}}^2 := \mathbf{v}^\top \mathbf{S} \mathbf{v}$, so that for a matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{v} \in \mathbb{R}^n$, we have

$$\|\mathbf{v}\|_{\mathbf{M}^\top \mathbf{M}}^2 := \mathbf{v}^\top \mathbf{M}^\top \mathbf{M} \mathbf{v} = \|\mathbf{M}\mathbf{v}\|_2^2.$$

Then by (26),

$$\begin{aligned}
\|\tilde{F}^{(t+1)}\|^2 &= \|\mathbf{M}^{(t)} \cdot F(\mathbf{w}^{(t)}) + \mathbf{N}^{(t)} \cdot F(\mathbf{z}^{(t-1)})\|^2 \\
&= \|F(\mathbf{w}^{(t)}) + (\mathbf{M}^{(t)})^{-1} \mathbf{N}^{(t)} \cdot F(\mathbf{z}^{(t-1)})\|^2_{(\mathbf{M}^{(t)})^\top \mathbf{M}^{(t)}}.
\end{aligned} \tag{27}$$

Next we define $\mathbf{C}^{(T)} = \mathbf{0}$ and for $-1 \le t < T$,[12]

$$\mathbf{C}^{(t-1)} := (\mathbf{M}^{(t)})^{-1}\mathbf{N}^{(t)}. \tag{28}$$

Notice that the definition of $\mathbf{C}^{(t-1)}$ in (28) depends on $\mathbf{C}^{(t)}$, which depends on $\mathbf{C}^{(t+1)}$, and so on. By (27) and (25), it follows that

$$\begin{aligned}
\|\tilde{F}^{(t+1)}\|^2 &= \|F(\mathbf{w}^{(t)}) + \mathbf{C}^{(t-1)} \cdot F(\mathbf{z}^{(t-1)})\|^2_{(\mathbf{M}^{(t)})^\top \mathbf{M}^{(t)}} && (29)\\
&= \|\tilde{F}^{(t)}\|^2_{(\mathbf{M}^{(t)})^\top \mathbf{M}^{(t)}} \\
&= \|(I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)})\tilde{F}^{(t)}\|^2 \\
&\le \|I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)}\|^2_\sigma \|\tilde{F}^{(t)}\|^2. && (30)
\end{aligned}$$

Our goal from here on is two-fold: (1) to prove an upper bound on $\|I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)}\|_\sigma$, which will ensure, by (30), that $\|\tilde{F}^{(t+1)}\| \lesssim \|\tilde{F}^{(t)}\|$, and (2) to ensure that $\|\tilde{F}^{(t)}\|$ is an (approximate) upper bound on $\|F(\mathbf{z}^{(t)})\|$ for all t, so that in particular upper bounding $\|\tilde{F}^{(T)}\|$ suffices to upper bound $\|F(\mathbf{z}^{(T)})\|$. These tasks will be performed in the following section; we first make a few remarks on the choice of $\mathbf{C}^{(t-1)}$ in (28):

Remark 6 (Specialization to the linear case & experiments)**.** In the case that the monotone operator F is linear, i.e., $F(\mathbf{z}) = \mathbf{A}\mathbf{z}$, it is straightforward to check that the matrices $\mathbf{C}^{(t-1)}$ as defined in (28) are all equal to the matrix $\mathbf{C}$ defined in (21) and $\mathbf{A}^{(t)} = \mathbf{B}^{(t)} = \mathbf{A}$ for all t. A special case of a linear operator F is that corresponding to a two-player zero-sum matrix game, i.e., where the payoffs of the players given actions $\mathbf{x}, \mathbf{y}$, are $\pm\mathbf{x}^\top\mathbf{M}\mathbf{y}$. In experiments we conducted for random instances of such matrix games, we observe that the adaptive potential function $\tilde{F}^{(t)}$ closely tracks $F(\mathbf{z}^{(t)})$, and both are monotonically decreasing with t. It seems that any "interesting" behavior whereby $F(\mathbf{z}^{(t)})$ grows by (say) a constant factor over the course of one or more iterations, but where $\tilde{F}^{(t)}$ grows only by much less, must occur for more complicated monotone operators (if at all). We leave a detailed experimental evaluation of such possibilities to future work.

Remark 7 (Alternative choice of $\mathbf{C}^{(t)}$)**.** It is not necessary to choose $\mathbf{C}^{(t-1)}$ as in (28). Indeed, in light of the fact that it is the spectral norms $\|I - \eta\mathbf{A}^{(t-1)} + \mathbf{C}^{(t-1)}\|_\sigma$ that control the increase in $\|\tilde{F}^{(t-1)}\|$ to $\|\tilde{F}^{(t)}\|$, it is natural to try to set

$$\tilde{\mathbf{C}}^{(t-1)} = \arg\min_{\mathbf{C}\in\mathbb{R}^{n\times n}} \left[\|I - \eta\mathbf{A}^{(t-1)} + \mathbf{C}\|_\sigma \,\middle|\, \left\{\|\mathbf{C}\|_\sigma \le \frac{1}{10}\right\} \text{ and } * \right], \tag{31}$$

where

$$* = \left\{ \|F(\mathbf{w}^{(t)}) + \mathbf{C} \cdot F(\mathbf{z}^{(t-1)})\|^2_{(\mathbf{M}^{(t)})^\top \mathbf{M}^{(t)}} \ge \|F(\mathbf{w}^{(t)}) + (\mathbf{M}^{(t)})^{-1}\mathbf{N}^{(t)} \cdot F(\mathbf{z}^{(t-1)})\|^2_{(\mathbf{M}^{(t)})^\top \mathbf{M}^{(t)}} \right\}. \tag{32}$$

The reason for the constraint $*$ defined in (32) is to ensure that $\|\tilde{F}^{(t+1)}\|^2 \le \|\tilde{F}^{(t)}\|^2_{(\mathbf{M}^{(t)})^\top \mathbf{M}^{(t)}}$ (so that (29) is replaced with an inequality). The reason for the constraint $\|\mathbf{C}\|_\sigma \le 1/10$ is to ensure that $\|F(\mathbf{z}^{(T)})\| \le O\left(\|\tilde{F}^{(T)}\|\right)$. Though the asymptotic rate of $O(1/\sqrt{T})$ established by the choice of $\mathbf{C}^{(t-1)}$ in (28) is tight in light of Theorem 7, it is possible that a choice of $\mathbf{C}^{(t-1)}$ as in (31) could lead to an improvement in the absolute constant. We leave an exploration of this possibility to future work.

B.4 Proof of Theorem 5

In this section we prove Theorem 5 using the definition of $\tilde{F}^{(t)}$ in (25), where $\mathbf{C}^{(t-1)}$ is defined in (28). We begin with a few definitions: for positive semidefinie matrices $\mathbf{S}, \mathbf{T}$, write $\mathbf{S} \preceq \mathbf{T}$ if $\mathbf{T} - \mathbf{S}$ is positive semidefinite (this is known as the *Loewner ordering*). We also define

$$\mathbf{D}^{(t)} := -\eta\mathbf{C}^{(t)}\mathbf{B}^{(t)} + (I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)} - \mathbf{C}^{(t)})^2\eta\mathbf{B}^{(t)} \qquad \forall t \le T-1. \tag{33}$$

[12] The invertibility of $\mathbf{M}^{(t)}$, and thus the well-definedness of $\mathbf{C}^{(t-1)}$, is established in Lemma 15.

To understand the definition of the matrices $\mathbf{D}^{(t)}$ in (33), note that, in light of the equality

$$(I-\mathbf{X})^{-1}\mathbf{X} = \mathbf{X} + (I-\mathbf{X})^{-1}\mathbf{X}^2 \tag{34}$$

for a square matrix $\mathbf{X}$ for which $I-\mathbf{X}$ is invertible, we have, for $t \le T$,

$$\begin{aligned}
&I - \eta\mathbf{A}^{(t-1)} + \mathbf{C}^{(t-1)} \\
=&I - \eta\mathbf{A}^{(t-1)} + (I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)} - \mathbf{C}^{(t)})\eta\mathbf{B}^{(t)} \\
=&I - \eta\mathbf{A}^{(t-1)} + \eta^2\mathbf{A}^{(t)}\mathbf{B}^{(t)} + \left(-\eta\mathbf{C}^{(t)}\mathbf{B}^{(t)} + (I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)} - \mathbf{C}^{(t)})^2\eta\mathbf{B}^{(t)}\right) \\
=&I - \eta\mathbf{A}^{(t-1)} + \eta^2\mathbf{A}^{(t)}\mathbf{B}^{(t)} + \mathbf{D}^{(t)}.
\end{aligned} \tag{35}$$

Thus, to upper bound $\|I - \mathbf{A}^{(t-1)} + \mathbf{C}^{(t-1)}\|$, it will suffice to use the below lemma, which generalizes [GPDO20, Lemma 12] and can be used to give an upper bound on the spectral norm of $I - \eta\mathbf{A}^{(t-1)} + \eta^2\mathbf{A}^{(t)}\mathbf{B}^{(t)} + \mathbf{D}^{(t)}$ for each t:

Lemma 13. *Suppose $\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}, \mathbf{D} \in \mathbb{R}^{n\times n}$ are matrices and $K, L_0, L_1, L_2, \delta > 0$ so that:*

- *$\mathbf{A}_1 + \mathbf{A}_2^\top$, $\mathbf{A}_2 + \mathbf{A}_2^\top$, and $\mathbf{B} + \mathbf{B}^\top$ are PSD;*
- *$\|\mathbf{A}_1\|_\sigma, \|\mathbf{A}_2\|_\sigma, \|\mathbf{B}\|_\sigma \le L_0 \le 1/106$;*
- *$\mathbf{D} + \mathbf{D}^\top \preceq L_1 \cdot \left(\mathbf{B}^\top\mathbf{B} + \mathbf{A}_1\mathbf{A}_1^\top\right) + K\delta^2 \cdot I$.*
- *$\mathbf{D}^\top\mathbf{D} \preceq L_2 \cdot \mathbf{B}^\top\mathbf{B}$.*
- *$10L_0 + \frac{4L_2}{L_0^2} + 5L_1 \le 24/50$.*
- *For any two matrices $\mathbf{X}, \mathbf{Y} \in \{\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}\}$, $\|\mathbf{X} - \mathbf{Y}\|_\sigma \le \delta$.*

It follows that

$$\|I - \mathbf{A}_1 + \mathbf{A}_2\mathbf{B} + \mathbf{D}\|_\sigma \le \sqrt{1 + (K+400)\,\delta^2}.$$

Proof of Lemma 13. We wish to show that

$$(I - \mathbf{A}_1 + \mathbf{A}_2\mathbf{B} + \mathbf{D})^\top(I - \mathbf{A}_1 + \mathbf{A}_2\mathbf{B} + \mathbf{D}) \preceq \left(1 + (K+400)\cdot\delta^2\right) I,$$

or equivalently

$$\begin{aligned}
&(\mathbf{A}_1 + \mathbf{A}_1^\top) - (\mathbf{B}^\top\mathbf{A}_2^\top + \mathbf{A}_2\mathbf{B}) - \mathbf{A}_1^\top\mathbf{A}_1 + (\mathbf{B}^\top\mathbf{A}_2^\top\mathbf{A}_1 + \mathbf{A}_1^\top\mathbf{A}_2\mathbf{B}) - \mathbf{B}^\top\mathbf{A}_2^\top\mathbf{A}_2\mathbf{B} \\
&- (\mathbf{D}^\top + \mathbf{D}) + (\mathbf{D}^\top\mathbf{A}_1 + \mathbf{A}_1^\top\mathbf{D}) - (\mathbf{D}^\top\mathbf{A}_2\mathbf{B} + \mathbf{B}^\top\mathbf{A}_2^\top\mathbf{D}) - \mathbf{D}^\top\mathbf{D} \succeq -(K+400)\cdot\delta^2 I.
\end{aligned} \tag{36}$$

For $i \in \{1,2\}$, let us write $\mathbf{J}_i = (\mathbf{A}_i - \mathbf{A}_i^\top)/2, \mathbf{R}_i = (\mathbf{A}_i + \mathbf{A}_i^\top)/2$, and $\mathbf{K} = (\mathbf{B} - \mathbf{B}^\top)/2, \mathbf{S} = (\mathbf{B} + \mathbf{B}^\top)/2$, so that $\mathbf{R}_1, \mathbf{R}_2, \mathbf{S}$ are positive semidefinite and $\mathbf{J}_1, \mathbf{J}_2, \mathbf{K}$ are anti-symmetric.

Next we will show (in (42) below) that the sum of all terms in (36) apart from the first four are preceded by a constant (depending on L_0, L_1) times $\mathbf{B}^\top\mathbf{B}$ in the Loewner ordering. To show this we begin as follows: for any $\epsilon, \epsilon_1 > 0$, we have:

$$\text{(Lemma 18)} \qquad \mathbf{A}_1^\top\mathbf{A}_1 \preceq (1+\epsilon_1)\cdot\mathbf{B}^\top\mathbf{B} + \left(1 + \frac{1}{\epsilon_1}\right)\delta^2 I \tag{37}$$

$$\begin{aligned}
\text{(Lemma 17)} \qquad -\mathbf{B}^\top\mathbf{A}_2^\top\mathbf{A}_1 - \mathbf{A}_1^\top\mathbf{A}_2\mathbf{B} &\preceq \epsilon\cdot\mathbf{B}^\top\mathbf{B} + \frac{1}{\epsilon}\cdot\mathbf{A}_1^\top\mathbf{A}_2\mathbf{A}_2^\top\mathbf{A}_1 \\
\text{(Lemma 20)} \qquad &\preceq \epsilon\cdot\mathbf{B}^\top\mathbf{B} + \frac{L_0^2}{\epsilon}\cdot\mathbf{A}_1^\top\mathbf{A}_1 \\
\text{(Lemma 18)} \qquad &\preceq \left(\epsilon + \frac{2L_0^2}{\epsilon}\right)\cdot\mathbf{B}^\top\mathbf{B} + \frac{2L_0^2}{\epsilon}\delta^2 I
\end{aligned} \tag{38}$$

$$\text{(Lemma 20)} \qquad \mathbf{B}^\top\mathbf{A}_2^\top\mathbf{A}_2\mathbf{B} \preceq L_0^2\mathbf{B}^\top\mathbf{B}. \tag{39}$$

Note in particular that (37), (38), and (39) imply that

$$(\mathbf{A}_1 - \mathbf{A}_2\mathbf{B})^\top(\mathbf{A}_1 - \mathbf{A}_2\mathbf{B}) \preceq \left(1 + \epsilon + \epsilon_1 + \frac{2L_0^2}{\epsilon} + L_0^2\right) \cdot \mathbf{B}^\top\mathbf{B} + \left(1 + \frac{2L_0^2}{\epsilon} + \frac{1}{\epsilon_1}\right)\delta^2 I,$$

and choosing $\epsilon = L_0 \leq 1$ (whereas ϵ_1 is left as a free parameter to be specified below) gives

$$(\mathbf{A}_1 - \mathbf{A}_2\mathbf{B})^\top(\mathbf{A}_1 - \mathbf{A}_2\mathbf{B}) \preceq (1 + 4L_0 + \epsilon_1) \cdot \mathbf{B}^\top\mathbf{B} + \left(1 + 2L_0 + \frac{1}{\epsilon_1}\right) \cdot \delta^2 I. \tag{40}$$

It follows from (40) and Lemma 17 that

$$\begin{aligned}
&(\mathbf{A}_2\mathbf{B} - \mathbf{A}_1)^\top\mathbf{D} + \mathbf{D}^\top(\mathbf{A}_2\mathbf{B} - \mathbf{A}_1) \\
\preceq &\min_{\epsilon>0,\epsilon_1>0} \epsilon \cdot \left((1 + 4L_0 + \epsilon_1) \cdot \mathbf{B}^\top\mathbf{B} + \left(1 + 2L_0 + \frac{1}{\epsilon_1}\right) \cdot \delta^2 I\right) + \frac{1}{\epsilon} \cdot L_2\mathbf{B}^\top\mathbf{B} \\
\preceq &\left(2L_0^2 + \frac{L_2}{L_0^2}\right) \cdot \mathbf{B}^\top\mathbf{B} + (2L_0^2 + L_0) \cdot \delta^2 I,
\end{aligned} \tag{41}$$

where the last line results from the choice $\epsilon = L_0^2, \epsilon_1 = L_0$.

By (40) and (41) we have, for any $\epsilon_1 > 0$,

$$\begin{aligned}
&(\mathbf{A}_1 - \mathbf{A}_2\mathbf{B})^\top(\mathbf{A}_1 - \mathbf{A}_2\mathbf{B}) + (\mathbf{A}_2\mathbf{B} - \mathbf{A}_1)^\top\mathbf{D} + \mathbf{D}^\top(\mathbf{A}_2\mathbf{B} - \mathbf{A}_1) + (\mathbf{D}^\top + \mathbf{D}) + \mathbf{D}^\top\mathbf{D} \\
\preceq &\left(1 + 4L_0 + \epsilon_1 + 2L_0^2 + \frac{L_2}{L_0^2} + L_1 + L_2\right) \cdot \mathbf{B}^\top\mathbf{B} + L_1 \cdot \mathbf{A}_1\mathbf{A}_1^\top + \left(K + 1 + 2L_0 + \frac{1}{\epsilon_1} + 2L_0^2 + L_0\right) \cdot \delta^2 I \\
\preceq &\left(1 + 5L_0 + \epsilon_1 + \frac{2L_2}{L_0^2} + L_1\right) \cdot \mathbf{B}^\top\mathbf{B} + L_1 \cdot \mathbf{A}_1\mathbf{A}_1^\top + \left(K + 1 + 4L_0 + \frac{1}{\epsilon_1}\right) \cdot \delta^2 I.
\end{aligned} \tag{42}$$

Next, for any $\epsilon > 0$, it holds that

$$\begin{aligned}
&\mathbf{B}^\top\mathbf{A}_2^\top + \mathbf{A}_2\mathbf{B} \\
&= -(\mathbf{K}^\top\mathbf{J}_2 + \mathbf{J}_2^\top\mathbf{K}) + (\mathbf{S}\mathbf{R}_2 + \mathbf{R}_2\mathbf{S}) + (\mathbf{S}\mathbf{J}_2^\top + \mathbf{J}_2\mathbf{S}) + (\mathbf{K}^\top\mathbf{R}_2 + \mathbf{R}_2\mathbf{K}) \\
\text{(Lemma 17)} \preceq &-(\mathbf{K}^\top\mathbf{J}_2 + \mathbf{J}_2^\top\mathbf{K}) + (\mathbf{S}\mathbf{R}_2 + \mathbf{R}_2\mathbf{S}) + \frac{1}{\epsilon} \cdot \left(\mathbf{S}^2 + \mathbf{R}_2^2\right) + \epsilon \cdot \left(\mathbf{J}_2\mathbf{J}_2^\top + \mathbf{K}^\top\mathbf{K}\right) \\
\text{(Lemma 19)} \preceq &-(\mathbf{K}^\top\mathbf{J}_2 + \mathbf{J}_2^\top\mathbf{K}) + 3\mathbf{S}^2 + \frac{1}{\epsilon} \cdot \left(\mathbf{S}^2 + \mathbf{R}_2^2\right) + \epsilon \cdot \left(\mathbf{J}_2\mathbf{J}_2^\top + \mathbf{K}^\top\mathbf{K}\right) + 2\delta^2 I \\
\text{(Lemma 18)} \preceq &-(\mathbf{K}^\top\mathbf{J}_2 + \mathbf{J}_2^\top\mathbf{K}) + \left(3 + \frac{3}{\epsilon}\right)\mathbf{S}^2 + 3\epsilon \cdot \mathbf{K}^\top\mathbf{K} + \left(2 + \frac{2}{\epsilon} + 2\epsilon\right)\delta^2 I.
\end{aligned} \tag{43}$$

Next, we have for any $\epsilon > 0$,

$$\begin{aligned}
\mathbf{A}_1\mathbf{A}_1^\top =&\mathbf{R}_1\mathbf{R}_1^\top + (\mathbf{J}_1\mathbf{R}_1^\top + \mathbf{R}_1\mathbf{J}_1^\top) + \mathbf{J}_1\mathbf{J}_1^\top \\
\text{(Lemma 17)} \quad \preceq&2\mathbf{R}_1\mathbf{R}_1^\top + 2\mathbf{J}_1\mathbf{J}_1^\top \\
=&2\mathbf{R}_1\mathbf{R}_1^\top + 2\mathbf{J}_1^\top\mathbf{J}_1 \\
\text{(Lemma 18)} \quad \preceq&(2 + 2\epsilon)\,\mathbf{S}^2 + (2 + 2\epsilon)\,\mathbf{J}_2^\top\mathbf{J}_2 + \frac{4}{\epsilon} \cdot \delta^2 I.
\end{aligned} \tag{44}$$

By (43) and (44), for any $\mu, \nu \in (0,1)$ and $\epsilon > 0$ with $2\nu + 10\epsilon + \mu \cdot (2+2\epsilon) \le 1$,

$$\begin{aligned}
&\mathbf{B}^\top \mathbf{A}_2^\top + \mathbf{A}_2 \mathbf{B} + (1+\nu)\mathbf{B}^\top \mathbf{B} + \mu \mathbf{A}\mathbf{A}^\top \\
\preceq& -(\mathbf{K}^\top \mathbf{J}_2 + \mathbf{J}_2^\top \mathbf{K}) + 3\epsilon \mathbf{K}^\top \mathbf{K} + (1+\nu)\mathbf{K}^\top \mathbf{K} + \mu \cdot (2+2\epsilon)\, \mathbf{J}_2^\top \mathbf{J}_2 + \left(4 + \nu + \frac{3}{\epsilon} + \mu \cdot (2+2\epsilon)\right) \mathbf{S}^2 \\
&+ (1+\nu)(\mathbf{K}^\top \mathbf{S} + \mathbf{S}\mathbf{K}) + \left(2 + \frac{2}{\epsilon} + 2\epsilon + \frac{4\mu}{\epsilon}\right) \delta^2 I \\
\preceq& -(\mathbf{K}^\top \mathbf{J}_2 + \mathbf{J}_2^\top \mathbf{K}) + (1 + \nu + 3\epsilon + (1+\nu)\epsilon)\, \mathbf{K}^\top \mathbf{K} + \mu \cdot (2+2\epsilon)\, \mathbf{J}_2^\top \mathbf{J}_2 \\
&+ \left(4 + \nu + \frac{3}{\epsilon} + \frac{1+\nu}{\epsilon} + \mu \cdot (2+2\epsilon)\right) \mathbf{S}^2 + \left(2 + \frac{2+4\mu}{\epsilon} + 2\epsilon\right) \delta^2 I \qquad (45) \\
\preceq& -(\mathbf{K}^\top \mathbf{J}_2 + \mathbf{J}_2^\top \mathbf{K}) + \mathbf{K}^\top \mathbf{K} + (2\nu + 10\epsilon + \mu(2+2\epsilon)) \mathbf{J}_2^\top \mathbf{J}_2 \\
&+ \left(5 + \frac{5}{\epsilon} + \mu \cdot (2+2\epsilon)\right) \mathbf{S}^2 + \left(4 + \frac{2+4\mu}{\epsilon} + 2\epsilon\right) \delta^2 I \qquad (46) \\
\preceq& (\mathbf{J}_2 - \mathbf{K})^\top (\mathbf{J}_2 - \mathbf{K}) + \left(6 + \frac{5}{\epsilon}\right) \mathbf{S}^2 + \left(4 + \frac{4}{\epsilon} + 2\epsilon\right) \delta^2 I \qquad (47) \\
\preceq& \left(12 + \frac{10}{\epsilon}\right) \mathbf{R}_1^2 + \left(17 + \frac{14}{\epsilon} + 2\epsilon\right) \delta^2 I \qquad (48) \\
\preceq& \left(12 + \frac{10}{\epsilon}\right) L_0 \mathbf{R}_1 + \left(17 + \frac{14}{\epsilon} + 2\epsilon\right) \delta^2 I. \qquad (49)
\end{aligned}$$

where (45) follows from Lemma 17, (46) follows from Lemma 18 and $\nu + 5\epsilon \le 1$, (47) follows from $2\nu + 10\epsilon + \mu \cdot (2+2\epsilon) \le 1$, (48) follows from $\|\mathbf{J}_2 - \mathbf{K}\|_\sigma \le \delta$ as well as Lemma 18, and (49) follows from Lemma 20 together with $\|\mathbf{R}_1^{1/2}\|_\sigma \le \sqrt{L_0}$.

By (42) and (49), by choosing $\epsilon_1 = 1/100, \epsilon = 1/20$, $\nu = 5L_0 + \epsilon_1 + \frac{2L_2}{L_0^2} + L_1$, and $\mu = L_1$, which satisfy

$$10\epsilon + 2\nu + (2+2\epsilon)\mu = 10\epsilon + 2 \cdot \left(5L_0 + 1/100 + \frac{2L_2}{L_0^2} + L_1\right) + 3L_1 \le 1/2 + 1/50 + \left(10L_0 + \frac{4L_2}{L_0^2} + 5L_1\right) \le 1,$$

it holds that for the above choices of ϵ, ϵ_1,

$$\begin{aligned}
&(\mathbf{B}^\top \mathbf{A}_2^\top + \mathbf{A}_2 \mathbf{B}) + (\mathbf{A}_1 - \mathbf{A}_2 \mathbf{B})^\top (\mathbf{A}_1 - \mathbf{A}_2 \mathbf{B}) + (\mathbf{A}_2 \mathbf{B} - \mathbf{A}_1)^\top \mathbf{D} + \mathbf{D}^\top (\mathbf{A}_2 \mathbf{B} - \mathbf{A}_1) + (\mathbf{D}^\top + \mathbf{D}) + \mathbf{D}^\top \mathbf{D} \\
\preceq& L_0/2 \cdot \left(12 + \frac{10}{\epsilon}\right) (\mathbf{A}_1^\top + \mathbf{A}_1) + \left(K + 18 + 4L_0 + \frac{1}{\epsilon_1} + \frac{14}{\epsilon} + 2\epsilon\right) \delta^2 I \\
\preceq& 106 L_0 \cdot (\mathbf{A}_1^\top + \mathbf{A}_1) + (K + 400)\, \delta^2 I \\
\preceq& \mathbf{A}_1^\top + \mathbf{A}_1 + (K + 400) \cdot \delta^2 I,
\end{aligned}$$

establishing (36). ∎

The next several lemmas ensure that the matrices $\mathbf{D}^{(t)}$ satisfy the conditions of the matrix $\mathbf{D}$ of Lemma 13. First, Lemma 14 shows that $\|F(\mathbf{z}^{(t)})\|$ only grows by a constant factor over the course of a constant number of time steps.

Lemma 14. *Suppose that for some $t \ge 1$, we have $\max\{\|F(\mathbf{z}^{(t)})\|, \|F(\mathbf{z}^{(t-1)})\|\} \le \delta$. Then for any $s \ge 1$, we have $\|F(\mathbf{z}^{(t+s)})\| \le \delta \cdot (1 + 3\eta\ell)^s$.*

Proof. We prove the claimed bound by induction. Since F is ℓ-Lipschitz, we get

$$\|F(\mathbf{z}^{(t+s)}) - F(\mathbf{z}^{(t+s-1)})\| \le 3\eta\ell \max\{\|F(\mathbf{z}^{(t+s-1)})\|, \|F(\mathbf{z}^{(t+s-2)})\|\}$$

for each $s \ge 1$, and so if $\delta_s := \max\{\|F(\mathbf{z}^{(t+s-1)})\|, \|F(\mathbf{z}^{(t+s-2)})\|\}$, the triangle inequality gives

$$\|F(\mathbf{z}^{(t+s)})\| \le \delta_s (1 + 3\eta\ell).$$

It follows by induction that $\|F(\mathbf{z}^{(t+s)})\| \le \delta \cdot (1 + 3\eta\ell)^s$.

∎

Lemma 15 uses backwards induction (on t) to establish bounds on the matrices $\mathbf{C}^{(t)}$.

Lemma 15 (Backwards induction lemma)**.** *Suppose that there is some $L_0 > 0$ so that for all $t \le T$, we have $\max\{\eta\|\mathbf{A}^{(t)}\|_\sigma, \eta\|\mathbf{B}^{(t)}\|_\sigma\} \le L_0 \le \sqrt{1/200}$ and $\eta\ell \le 2/3$. Then:*

1. *$\|\mathbf{C}^{(t)}\|_\sigma \le 2L_0^2$ for each $t \in [T]$.*
2. *The matrices $\mathbf{C}^{(t)}$ are well-defined, i.e., $I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)}$ is invertible for each $t \in [T]$, and the spectral norm of its inverse is bounded above by $\sqrt{2}$.*
3. *$\|\eta\mathbf{A}^{(t)} - \mathbf{C}^{(t)}\|_\sigma \le 2L_0$ and $\|I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)}\|_\sigma \le 1 + 2L_0$ for each $t \in [T]$.*
4. *For all $t < T$, it holds that*

$$\begin{aligned}&(I - \eta\mathbf{A}^{(t+1)} + \mathbf{C}^{(t+1)})^{-1}(\eta\mathbf{A}^{(t+1)} - \mathbf{C}^{(t+1)})(\eta(\mathbf{A}^{(t+1)})^\top - (\mathbf{C}^{(t+1)})^\top)(I - \eta\mathbf{A}^{(t+1)} + \mathbf{C}^{(t+1)})^{-\top}\\ \preceq& 3 \cdot \left((\eta\mathbf{A}^{(t+1)})(\eta\mathbf{A}^{(t+1)})^\top + \mathbf{C}^{(t+1)}(\mathbf{C}^{(t+1)})^\top\right).\end{aligned}$$

5. *Let $\delta^{(t)} := \max\{\|F(\mathbf{z}^{(t)})\|, \|F(\mathbf{z}^{(t-1)})\|\}$ for all $t \le T$. For $t < T$, it holds that*

$$\mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top \preceq J_1 \cdot \eta\mathbf{A}^{(t)}(\eta\mathbf{A}^{(t)})^\top + J_2 \cdot (\delta^{(t)})^2 \cdot I,$$

for $J_1 = 8L_0^2$ and $J_2 = 30L_0^2\eta^2(\eta\Lambda)^2$.

Proof. The proof proceeds by backwards induction on t. The base case $t = T$ clearly holds since $\mathbf{C}^{(T)} = 0$. As for the inductive step, suppose that items 1 through 4 hold at time step t, for some $t \le T$. Then by (28) and $L_0 \le \frac{\sqrt{2}-1}{2}$,

$$\|\mathbf{C}^{(t-1)}\|_\sigma \le L_0 \cdot (L_0 + \|\mathbf{C}^{(t)}\|_\sigma) \cdot \|(I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)})^{-1}\| \le \sqrt{2}L_0 \cdot (L_0 + 2L_0^2) \le 2L_0^2,$$

establishing item 1 at time $t-1$.

Next, note that $\|\eta\mathbf{A}^{(t-1)} - \mathbf{C}^{(t-1)}\| \le L_0 + 2L_0^2 \le 2L_0$. Thus, by Equation (5.8.2) of [HJ12] and $L_0 \le \frac{1}{2} - \frac{1}{2\sqrt{2}}$, it follows that

$$\|(I - \eta\mathbf{A}^{(t-1)} + \mathbf{C}^{(t-1)})^{-1}\|_\sigma \le \frac{1}{1 - 2L_0} \le \sqrt{2},$$

which establishes item 2 at time $t-1$. It is also immediate that $\|I - \eta\mathbf{A}^{(t-1)} + \mathbf{C}^{(t-1)}\|_\sigma \le 1 + 2L_0$, establishing item 3 at time $t-1$.

Next we establish items 4 and 5 at time $t-1$. First, we have

$$\begin{aligned}&\|\mathbf{A}^{(t)} - \mathbf{A}^{(t-1)}\|_\sigma\\ \text{(Lemma 12)}\quad \le& \Lambda\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\| + \frac{\eta\Lambda}{2}\|F(\mathbf{z}^{(t)}) - F(\mathbf{z}^{(t-1)})\|\\ \le& \eta\Lambda\|F(\mathbf{z}^{(t-1)})\| + \frac{\eta\Lambda}{2}\left(2\eta\ell\|F(\mathbf{z}^{(t-1)})\| + \eta\ell\|F(\mathbf{z}^{(t-2)})\|\right)\\ \le& \delta^{(t-1)} \cdot 2\eta\Lambda,\end{aligned} \tag{50}$$

where the final inequality uses $\eta\ell \le 2/3$.

Next, by definition of $\mathbf{C}^{(t-1)}$ in (28),

$$
\begin{aligned}
&\mathbf{C}^{(t-1)}(\mathbf{C}^{(t-1)})^\top \\
=&\eta^2(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)})\mathbf{B}^{(t)}(\mathbf{B}^{(t)})^\top(\eta(\mathbf{A}^{(t)})^\top-(\mathbf{C}^{(t)})^\top)(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-\top} \\
\preceq& L_0^2(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)})(\eta(\mathbf{A}^{(t)})^\top-(\mathbf{C}^{(t)})^\top)(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-\top} && (51)\\
\preceq& \frac{L_0^2}{(1-\|\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)}\|_\sigma)^2}\cdot(\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)})(\eta(\mathbf{A}^{(t)})^\top-(\mathbf{C}^{(t)})^\top) && (52)\\
\preceq& \frac{2L_0^2}{(1-2L_0)^2}\cdot\left((\eta\mathbf{A}^{(t)})(\eta\mathbf{A}^{(t)})^\top+\mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top\right) && (53)\\
\preceq& 3L_0^2\cdot\left((\eta\mathbf{A}^{(t)})(\eta\mathbf{A}^{(t)})^\top+\mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top\right), && (54)\\
\preceq& 3L_0^2\cdot\left((\eta\mathbf{A}^{(t)})(\eta\mathbf{A}^{(t)})^\top\cdot(1+J_1)+J_2\cdot(\delta^{(t)})^2\cdot I\right) && (55)\\
\preceq& 6L_0^2(1+J_1)\cdot(\eta\mathbf{A}^{(t-1)})(\eta\mathbf{A}^{(t-1)})^\top+6L_0^2\eta^2(1+J_1)\cdot\|\mathbf{A}^{(t-1)}-\mathbf{A}^{(t)}\|_\sigma^2+3L_0^2J_2\cdot(\delta^{(t)})^2\cdot I && (56)\\
\preceq& 6L_0^2(1+J_1)\cdot(\eta\mathbf{A}^{(t-1)})(\eta\mathbf{A}^{(t-1)})^\top+24L_0^2\eta^2(1+J_1)(\eta\Lambda)^2\cdot(\delta^{(t-1)})^2+3L_0^2J_2\cdot(\delta^{(t)})^2\cdot I && (57)\\
\preceq& 6L_0^2(1+J_1)\cdot(\eta\mathbf{A}^{(t-1)})(\eta\mathbf{A}^{(t-1)})^\top+(\delta^{(t-1)})^2\cdot\left(24L_0^2\eta^2(1+J_1)(\eta\Lambda)^2+3L_0^2J_2(1+3\eta\ell)\right)\cdot I && (58)
\end{aligned}
$$

where:

- (51) follows by Lemma 20;
- (52) is by Lemma 21 with $\mathbf{X}=\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)}$;
- (53) uses Lemma 17 and item 3 at time t;
- (54) follows from $L_0\le\frac{1-\sqrt{2/3}}{2}$;
- (55) follows from the inductive hypothesis that item 5 holds at time t;
- (56) follows from Lemma 18;
- (57) follows from (50);
- (58) follows from the fact that $\delta^{(t)}\le(1+3\eta\ell)\delta^{(t-1)}$, which is a consequence of Lemma 14.

Inequalities (51) through (54) establish item 4 at time $t-1$. In order for item 5 to hold at time $t-1$, we need that

$$
\begin{aligned}
6L_0^2(1+J_1)&\le J_1 && (59)\\
24L_0^2\eta^2(1+J_1)(\eta\Lambda)^2+3L_0^2J_2(1+3\eta\ell)&\le J_2. && (60)
\end{aligned}
$$

By choosing $J_1=8L_0^2$ we satisfy (59) since $L_0<\sqrt{1/24}$. By choosing $J_2=30L_0^2\eta^2(\eta\Lambda)^2$ we satisfy (60) since

$$
24L_0^2\eta^2(1+8L_0^2)(\eta\Lambda)^2+3L_0^2\cdot J_2(1+3\eta\ell)\le 25L_0^2\eta^2(\eta\Lambda)^2+9L_0^2J_2\le J_2,
$$

where we use $L_0\le\sqrt{1/192}$ and $\eta\ell\le 2/3$. This completes the proof that item 5 holds at time $t-1$. □

Lemma 16. *Suppose that the pre-conditions of Lemma 15 (namely, those in its first sentence) hold. Then for each $t\in[T]$, we have*

$$
\mathbf{D}^{(t)}+(\mathbf{D}^{(t)})^\top\preceq 6L_0\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)}+4L_0\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top+\left(4L_0+\frac{1}{3L_0}\right)\mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top. \tag{61}
$$

and

$$
(\mathbf{D}^{(t)})^\top\mathbf{D}^{(t)}\preceq 60L_0^4\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)}. \tag{62}
$$

Proof. By Lemma 17, for any $\epsilon > 0$,

$$\begin{aligned}&-\mathbf{C}^{(t)}\eta\mathbf{B}^{(t)} - \eta(\mathbf{B}^{(t)})^\top(\mathbf{C}^{(t)})^\top \\ &\preceq \epsilon\cdot\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)} + \frac{1}{\epsilon}\cdot\mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top.\end{aligned}$$

Also, for any $\epsilon > 0$,

$$\begin{aligned}&(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)})^2\eta\mathbf{B}^{(t)} + \eta(\mathbf{B}^{(t)})^\top(\eta(\mathbf{A}^{(t)})^\top-(\mathbf{C}^{(t)})^\top)^2(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-\top} \\ &\preceq \frac{1}{\epsilon}(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)})^2(\eta(\mathbf{A}^{(t)})^\top-(\mathbf{C}^{(t)})^\top)^2(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-\top} + \epsilon\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)} \quad (63) \\ &\preceq \frac{4L_0^2}{\epsilon}(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)})(\eta(\mathbf{A}^{(t)})^\top-(\mathbf{C}^{(t)})^\top)(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-\top} + \epsilon\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)} \quad (64) \\ &\preceq \frac{12L_0^2}{\epsilon}\cdot\left(\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top + \mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top\right) + \epsilon\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)}. \quad (65)\end{aligned}$$

where (63) uses Lemma 17, (64) uses item 3 of Lemma 15 and Lemma 20, and (65) uses item 4 of Lemma 15.

Choosing $\epsilon = 3L_0$ and using the definition of $\mathbf{D}^{(t)}$ in (33), it follows from the above displays that

$$\mathbf{D}^{(t)} + (\mathbf{D}^{(t)})^\top \preceq 6L_0\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)} + 4L_0\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top + \left(4L_0 + \frac{1}{3L_0}\right)\mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top,$$

which establishes (61).

To prove (62) we first note that

$$\begin{aligned}&\left\|(I-\eta\mathbf{A}^{(t)}+\mathbf{C}^{(t)})^{-1}(\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)})^2 - \mathbf{C}^{(t)}\right\|_\sigma \\ \text{(Lemma 15, item 2)}\quad &\le \sqrt{2}\|\eta\mathbf{A}^{(t)}-\mathbf{C}^{(t)}\|_\sigma^2 + \|\mathbf{C}^{(t)}\|_\sigma \\ \text{(Lemma 15, items 1 \& 3)}\quad &\le \sqrt{2}\cdot 4L_0^2 + 2L_0^2 = (2+4\sqrt{2})L_0^2.\end{aligned}$$

By Lemma 20, it follows that

$$(\mathbf{D}^{(t)})^\top\mathbf{D}^{(t)} \preceq 60L_0^4\cdot\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)},$$

establishing (62). ∎

Finally we are ready to prove Theorem 5; for convenience we restate it here.

Theorem 5 (restated)**.** *Suppose $F:\mathbb{R}^n\to\mathbb{R}^n$ is a monotone operator that is ℓ-Lipschitz and is such that $\partial F(\cdot)$ is Λ-Lipschitz. For some $\mathbf{z}^{(-1)},\mathbf{z}^{(0)}\in\mathbb{R}^n$, suppose there is $\mathbf{z}^*\in\mathbb{R}^n$ so that $F_{\mathcal{G}}(\mathbf{z}^*)=0$ and $\|\mathbf{z}^*-\mathbf{z}^{(-1)}\|\le D, \|\mathbf{z}^*-\mathbf{z}^{(0)}\|\le D$. Then the iterates $\mathbf{z}^{(T)}$ of the OG algorithm (3) for any $\eta\le\min\left\{\frac{1}{150\ell},\frac{1}{1711D\Lambda}\right\}$ satisfy:*

$$\|F_{\mathcal{G}}(\mathbf{z}^{(T)})\| \le \frac{60D}{\eta\sqrt{T}} \quad (66)$$

Proof of Theorem 5. By Lemma 11 with $S=3$, we have that for some $t^*\in\{0,1,2,\dots,T\}$,

$$\max\{\|F(\mathbf{z}^{(t^*)})\|, \|F(\mathbf{z}^{(t^*-1)})\|, \|F(\mathbf{z}^{(t^*-2)})\|\} \le \frac{6\sqrt{3}D}{\eta\sqrt{T}\cdot\sqrt{1-10\eta^2\ell^2}} \le \frac{12D}{\eta\sqrt{T}} =: \delta_0. \quad (67)$$

Set $L_0 := \eta\ell \le 1/150$ and $\Lambda_0 := \eta\Lambda$. By Lemma 12 we have that $\|\eta\mathbf{A}^{(t)}\|_\sigma \le L_0$ and $\|\eta\mathbf{B}^{(t)}\|_\sigma \le L_0$ for all $t\le T$. Thus the preconditions of Lemma 15 hold, and in particular by item 1 of Lemma 15, it follows that

$$\begin{aligned}\|\tilde{F}^{(t^*)}\| &= \|F(\mathbf{w}^{(t^*)}) + \mathbf{C}^{(t^*-1)}\cdot F(\mathbf{z}^{(t^*-1)})\| \\ &\le \|F(\mathbf{w}^{(t^*)})\| + 2L_0^2\|F(\mathbf{z}^{(t^*-1)})\| \le \delta_0\cdot(1+L_0+2L_0^2) \le \delta_0\cdot(1+2L_0). \quad (68)\end{aligned}$$

Write $\delta := \delta_0(1 + 2L_0)$. By (30), we have that for any $t \in \{t^*, \dots, T\}$,

$$\|\tilde{F}_t\|^2 \le \prod_{t'=t^*}^{t-1} \|I - \eta \mathbf{A}^{(t')} + \mathbf{C}^{(t')}\|_\sigma^2 \cdot \delta^2. \tag{69}$$

We will prove by forwards induction (contrast with Lemma 15) that for each $t \in \{t^* - 1, \dots, T\}$, the following hold:

1. $\|\tilde{F}^{(t+1)}\| \le 2\delta$. (We will only need this item for $t^* - 1 \le t \le T - 1$.)
2. $\max\{\|F(\mathbf{z}^{(t)})\|, \|F(\mathbf{z}^{(t-1)})\|\} \le 4\delta$.
3. $\|I - \eta \mathbf{A}^{(t)} + \mathbf{C}^{(t)}\|_\sigma^2 \le 1 + 10025\Lambda_0^2\eta^2\delta^2$ if $t \ge t^*$.

The base case $t = t^* - 1$ is immediate: item 1 follows from (68), item 2 follows from (67), and item 3 states nothing for $t = t^* - 1$. We now assume that items 1 through 3 all hold for some value $t - 1 \ge t^* - 1$, and prove that they hold for t. We first establish that item 2 holds at time t, namely that $\|F(\mathbf{z}^{(t)})\| \le 4\delta$. Since item 1 holds at time $t - 1$, we get that $\|\tilde{F}_t\| \le 2\delta$, and so

$$\|F(\mathbf{w}^{(t)})\| = \|\tilde{F}_t - \mathbf{C}^{(t-1)}F(\mathbf{z}^{(t-1)})\| \le \|\tilde{F}_t\| + 2L_0^2\|F(\mathbf{z}^{(t-1)})\| \le 2\delta + 8L_0^2\delta,$$

which implies that

$$\|F(\mathbf{z}^{(t)})\| \le \|F(\mathbf{w}^{(t)})\| + \eta\ell\|F(\mathbf{z}^{(t-1)})\| \le 2\delta + 8L_0^2\delta + 4L_0\delta \le 4\delta,$$

where the last inequality holds since $8L_0^2 + 4L_0 \le 2$.

We proceed to the proof of item 1 at time t. By Lemma 12, we have that

$$\begin{aligned}
\|\mathbf{B}^{(t)} - \mathbf{B}^{(t+1)}\|_\sigma &\le \Lambda\|\mathbf{w}^{(t)} - \mathbf{w}^{(t+1)}\| + \frac{\eta\Lambda}{2}\|F(\mathbf{z}^{(t-1)}) - F(\mathbf{z}^{(t)})\| \\
&\le \eta\Lambda\|F(\mathbf{z}^{(t)})\| + \frac{\eta\Lambda}{2}(\|\eta\ell F(\mathbf{z}^{(t-2)})\| + \|2\eta\ell F(\mathbf{z}^{(t-1)})\|) \\
\text{(item 2 at times } t, t-1) \qquad &\le 4\delta\Lambda_0 \cdot (1 + 3L_0/2) \le 5\delta\Lambda_0 \qquad (70) \\
\|\mathbf{B}^{(t)} - \mathbf{A}^{(t)}\|_\sigma &\le \frac{\eta\Lambda}{2}\|F(\mathbf{z}^{(t)}) - F(\mathbf{z}^{(t-1)})\| \\
\text{(item 2 at time } t-1) \qquad &\le 6\delta\Lambda_0 L_0 \qquad (71) \\
\|\mathbf{B}^{(t+1)} - \mathbf{A}^{(t+1)}\|_\sigma &\le \frac{\eta\Lambda}{2}\|F(\mathbf{z}^{(t+1)}) - F(\mathbf{z}^{(t)})\| \\
&\le \frac{\Lambda_0}{2} \cdot (L_0\|F(\mathbf{z}^{(t)})\| + 2L_0\|F(\mathbf{z}^{(t+1)})\|) \\
\text{(item 2 at time } t \text{ \& Lemma 14}) \qquad &\le \frac{\Lambda_0}{2} \cdot (4\delta L_0 + 2L_0(1 + 3L_0) \cdot 4\delta) \\
&\le 8\Lambda_0 L_0 \delta. \qquad (72)
\end{aligned}$$

Recall that (35) gives

$$I - \eta \mathbf{A}^{(t)} + \mathbf{C}^{(t)} = I - \eta \mathbf{A}^{(t)} + \eta^2 \mathbf{A}^{(t+1)}\mathbf{B}^{(t+1)} + \mathbf{D}^{(t+1)}.$$

Now we will apply Lemma 13 with $\mathbf{A}_1 = \eta\mathbf{A}^{(t)}, \mathbf{A}_2 = \eta\mathbf{A}^{(t+1)}, \mathbf{B} = \eta\mathbf{B}^{(t+1)}, \mathbf{D} = \mathbf{D}^{(t+1)}$, $L_0 = \eta\ell$ (which is called L_0 in the present proof as well). We check that all of the preconditions of the lemma hold:

- For $\mathbf{X} \in \{\eta\mathbf{A}^{(t)}, \eta\mathbf{A}^{(t+1)}, \eta\mathbf{B}^{(t+1)}\}$, $\mathbf{X} + \mathbf{X}^\top$ is PSD by Lemma 12.
- For $\mathbf{X} \in \{\eta\mathbf{A}^{(t)}, \eta\mathbf{A}^{(t+1)}, \eta\mathbf{B}^{(t+1)}\}$, $\|\mathbf{X}\|_\sigma \le \eta\ell = L_0$ by Lemma 12, and we have $L_0 \le 1/53$.[13]

[13] As we have already noted, this observation establishes also that the preconditions of Lemmas 15 and 16 hold.

- We may bound $\mathbf{D}^{(t+1)} + (\mathbf{D}^{(t+1)})^\top$ as follows:

$$\begin{aligned}
&\mathbf{D}^{(t+1)} + (\mathbf{D}^{(t+1)})^\top \\
\preceq &6L_0\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)} + 4L_0\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top + \left(4L_0 + \frac{1}{3L_0}\right)\mathbf{C}^{(t)}(\mathbf{C}^{(t)})^\top \qquad (73)\\
\preceq &6L_0\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)} + 4L_0\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top \\
&+ \left(\frac{1}{2L_0}\right)\cdot\left(8L_0^2\cdot\eta\mathbf{A}^{(t)}(\eta\mathbf{A}^{(t)})^\top + 30L_0^2\eta^4\Lambda^2(4\delta)^2\cdot I\right) \qquad (74)\\
\preceq &6L_0\eta^2(\mathbf{B}^{(t)})^\top\mathbf{B}^{(t)} + 8L_0\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top + 240L_0\eta^2\Lambda_0^2\delta^2\cdot I\\
\preceq &12L_0\eta^2(\mathbf{B}^{(t+1)})^\top\mathbf{B}^{(t+1)} + 8L_0\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top + (300\delta^2\eta^2\Lambda_0^2 + 240L_0\eta^2\Lambda_0^2\delta^2)\cdot I. \qquad (75)\\
\preceq &12L_0\eta^2(\mathbf{B}^{(t+1)})^\top\mathbf{B}^{(t+1)} + 8L_0\eta^2\mathbf{A}^{(t)}(\mathbf{A}^{(t)})^\top + 310\delta^2\eta^2\Lambda_0^2\cdot I.
\end{aligned}$$

  where (73) follows from Lemma 16, (74) follows from item 5 of Lemma 15 and item 2 of the current induction at time t, and (75) follows from Lemma 18 and (70). This shows that in our application of Lemma 13 we may take $L_1 = 12L_0$. Moreover, as we will take the parameter δ in Lemma 13 to be $5\Lambda_0\eta\delta$ (see below items), we may take $K = 14L_0$ (since $14\cdot(5\Lambda_0\eta\delta)^2 \geq 310\delta^2\eta^2\Lambda_0^2$).

- Lemma 16 gives
$$(\mathbf{D}^{(t+1)})^\top\mathbf{D}^{(t+1)} \preceq 60L_0^4\eta^2(\mathbf{B}^{(t+1)})^\top\mathbf{B}^{(t+1)},$$
  so we may take $L_2 = 60L_0^4$ in our application of Lemma 13.

- We calculate that
$$12L_0 + \frac{4L_2}{L_0^2} + 5L_1 = 12L_0 + \frac{240L_0^4}{L_0^2} + 60L_0 = 72L_0 + 240L_0^2 \leq 1/2$$
  holds as long as $L_0 \leq 1/150$.

- By (70), (71), and (72), we may take the parameter δ in Lemma 13 to be equal to $5\Lambda_0\eta\delta$ since $\max\{8\Lambda_0L_0\delta, 4\delta\Lambda_0 + 12\delta\Lambda_0L_0, 4\delta\Lambda_0 + 20\delta\Lambda_0L_0\} \leq 5\Lambda_0\delta$.

By Lemma 13, it follows that

$$\|I - \eta\mathbf{A}^{(t)} + \mathbf{C}^{(t)}\|_\sigma^2 \leq 1 + 25\Lambda_0^2\eta^2\delta^2\cdot(400 + 14L_0) \leq 1 + 10025\Lambda_0^2\eta^2\delta^2,$$

which establishes that item 3 holds at time t.

Finally we show that item 1 holds at time t. To do so, we use (69) and the fact that $\delta^2 \leq \frac{146D^2}{\eta^2T}$ to conclude that

$$\|\tilde{F}^{(t+1)}\|^2 \leq \delta^2\cdot\left(1 + 10025\Lambda_0^2\eta^2\delta^2\right)^T \leq \delta^2\cdot\left(1 + \frac{K_0\Lambda_0^2D^2}{T}\right)^T \leq 4\delta^2,$$

where $K_0 = 10025\cdot 146$ and the last inequality holds as long as $K_0\Lambda_0^2D^2 = K_0\eta^2\Lambda^2D^2 \leq 1/2$, i.e., $\eta \leq \frac{1}{\sqrt{2K_0}\cdot\Lambda D}$; in particular, it suffices to take $\eta \leq \frac{1}{1711\cdot\Lambda D}$. This verifies that item 1 holds at time t, completing the inductive step.

The conclusion of Theorem 5 is an immediate conclusion of item 2 at time T, since $4\delta \leq 5\delta_0 = \frac{60D}{\eta\sqrt{T}}$. ☐

B.5 Helpful lemmas

Lemma 17 (Young's inequality). *For square matrices $\mathbf{X}, \mathbf{Y}$, we have, for any $\epsilon > 0$,*

$$\mathbf{X}\mathbf{Y}^\top + \mathbf{Y}\mathbf{X}^\top \preceq \epsilon\mathbf{X}\mathbf{X}^\top + \frac{1}{\epsilon}\cdot\mathbf{Y}\mathbf{Y}^\top.$$

Applying the previous lemma to the cross terms in the quantity $\mathbf{X}\mathbf{X}^\top$ when using the decomposition $\mathbf{X} = \mathbf{Y} + (\mathbf{X} - \mathbf{Y})$, we obtain the following.

Lemma 18. *For square matrices $\mathbf{X}, \mathbf{Y}$, we have, for any $\epsilon > 0$,*

$$\mathbf{X}\mathbf{X}^\top \preceq (1+\epsilon)\cdot \mathbf{Y}\mathbf{Y}^\top + \left(1+\frac{1}{\epsilon}\right)\|\mathbf{X}-\mathbf{Y}\|_\sigma^2 \cdot I.$$

In particular, choosing $\epsilon = 1$ gives

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

Lemma 19 is an immediate corollary of the two lemmas above:

Lemma 19. *For square matrices $\mathbf{X}, \mathbf{Y}$, we have*

$$\mathbf{X}\mathbf{Y}^\top + \mathbf{Y}\mathbf{X}^\top \preceq 3\mathbf{Y}\mathbf{Y}^\top + 2\|\mathbf{X}-\mathbf{Y}\|_\sigma^2 \cdot I.$$

Lemma 20. *For square matrices $\mathbf{X}, \mathbf{Y}$ such that $\|\mathbf{Y}\|_\sigma \le M$, we have*

$$\mathbf{X}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{X} \preceq M^2\mathbf{X}^\top\mathbf{X}.$$

Proof. For any $\mathbf{v}$, we have

$$\|\mathbf{Y}\mathbf{X}\mathbf{v}\|^2 \le M^2\|\mathbf{X}\mathbf{v}\|^2.$$

□

Lemma 21. *For any square matrix $\mathbf{X}$ so that $\|\mathbf{X}\|_\sigma < 1$, we have*

$$(I-\mathbf{X})^{-1}\mathbf{X}\mathbf{X}^\top(I-\mathbf{X})^{-\top} \preceq \frac{1}{(1-\|\mathbf{X}\|_\sigma)^2}\cdot \mathbf{X}\mathbf{X}^\top.$$

Proof. Using the equality (34), we have that for any $\epsilon > 0$,

$$\begin{aligned}
&(I-\mathbf{X})^{-1}\mathbf{X}\mathbf{X}^\top(I-\mathbf{X})^{-\top}\\
=&(\mathbf{X}+(I-\mathbf{X})^{-1}\mathbf{X}^2)(\mathbf{X}^\top+(\mathbf{X}^\top)^2(I-\mathbf{X})^{-\top})\\
=&\mathbf{X}\mathbf{X}^\top+(I-\mathbf{X})^{-1}\mathbf{X}^2\mathbf{X}^\top+\mathbf{X}(\mathbf{X}^\top)^2(I-\mathbf{X})^{-\top}+(I-\mathbf{X})^{-1}\mathbf{X}^2(\mathbf{X}^\top)^2(I-\mathbf{X})^{-\top}\\
\text{(Lemma 17)}\quad \preceq&(1+1/\epsilon)\mathbf{X}\mathbf{X}^\top+(1+\epsilon)(I-\mathbf{X})^{-1}\mathbf{X}^2(\mathbf{X}^\top)^2(I-\mathbf{X})^{-\top}\\
\text{(Lemma 20)}\quad \preceq&(1+1/\epsilon)\mathbf{X}\mathbf{X}^\top+(1+\epsilon)\|\mathbf{X}\|_\sigma^2\cdot(I-\mathbf{X})^{-1}\mathbf{X}\mathbf{X}^\top(I-\mathbf{X})^{-\top}.
\end{aligned}$$

Rearranging gives

$$(I-\mathbf{X})^{-1}\mathbf{X}\mathbf{X}^\top(I-\mathbf{X})^{-\top} \preceq \min_{\epsilon>0:(1+\epsilon)\|\mathbf{X}\|_\sigma^2<1}\frac{(1+1/\epsilon)\mathbf{X}\mathbf{X}^\top}{1-(1+\epsilon)\|\mathbf{X}\|_\sigma^2}$$

Choosing $\epsilon = \frac{1-\|\mathbf{X}\|_\sigma}{\|\mathbf{X}\|_\sigma}$ gives the desired conclusion. □

C Proofs for Section 4

In this section we prove Theorem 7, and as byproducts of our analysis additionally prove the results mentioned at the end of Section 4.

Recall from Section 4 that $\mathcal{F}^{\text{bil}}_{n,\ell,D}$ is defined to be the set of ℓ-Lipschitz operators $F:\mathbb{R}^n\to\mathbb{R}^n$ of the form

$$F(\mathbf{z}) = \mathbf{A}\mathbf{z}+\mathbf{b} \quad \text{where} \quad \mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}, \mathbf{A} = \begin{pmatrix}\mathbf{0} & \mathbf{M}\\ -\mathbf{M}^\top & \mathbf{0}\end{pmatrix}, \mathbf{b} = \begin{pmatrix}\mathbf{b}_1\\ -\mathbf{b}_2\end{pmatrix}, \tag{76}$$

for which $\mathbf{A}$ is of full rank and $-\mathbf{A}^{-1}\mathbf{b} \in \mathcal{D}_D := \mathcal{B}_{\mathbb{R}^{n/2}}(\mathbf{0}, D)\times\mathcal{B}_{\mathbb{R}^{n/2}}(\mathbf{0}, D)$. Note that each $F\in\mathcal{F}^{\text{bil}}_{n,\ell,D}$ can be written as the min-max gradient operator $F(\mathbf{x},\mathbf{y}) = (\nabla_{\mathbf{x}}f(\mathbf{x},\mathbf{y})^\top, -\nabla_{\mathbf{y}}f(\mathbf{x},\mathbf{y})^\top)^\top$ corresponding to the function

$$f(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\mathbf{M}\mathbf{y}+\mathbf{b}_1^\top\mathbf{x}+\mathbf{b}_2^\top\mathbf{y}. \tag{77}$$

We next note that when $F\in\mathcal{F}^{\text{bil}}_{n,\ell,D}$, the p-SCLI updates of Definition 5 can be rewritten as follows:

Observation 22. *Suppose that $\mathcal{A}$ is a p-SCLI. Then there are constants $\alpha_j, \beta_j, \gamma, \delta \in \mathbb{R}$, $0 \le j \le p-1$, depending only on $\mathcal{A}$, so that for an instance F of the form $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$, and an arbitrary set of p initialization poitns $\mathbf{z}^{(0)}, \dots, \mathbf{z}^{(-p+1)} \in \mathbb{R}^n$, the iterates $\mathbf{z}^{(t)}$ of $\mathcal{A}$ satisfy*

$$\mathbf{z}^{(t)} = \sum_{j=0}^{p-1} \mathbf{C}_j(\mathbf{A})\mathbf{z}^{(t-p+j)} + \mathbf{N}(\mathbf{A})\mathbf{b}, \tag{78}$$

for $t \ge 1$ and $\mathbf{C}_j(\mathbf{A}) = \alpha_j \mathbf{A} + \beta_j I_n$ for $0 \le j \le p-1$ and $\mathbf{N}(\mathbf{A}) = \gamma \mathbf{A} + \delta I_n$.

In the case of OG with a constant step size η, for $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$, we may rewrite (15) as

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

so we have $\mathbf{C}_0(\mathbf{A}) = I_n - 2\eta\mathbf{A}, \mathbf{C}_1(\mathbf{A}) = \eta\mathbf{A}, \mathbf{N}(\mathbf{A}) = -\eta I_n$.

All lower bounds we prove in this section will apply more generally to any iterative algorithm $\mathcal{A}$ whose updates are of the form (78) when restricted to instances $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$.

The remainder of this section is organized as follows. In Section C.1, we prove Theorem 7. In Section C.2 we prove Proposition 8, which is used in the proof of Theorem 7, and Proposition 9, showing that Proposition 8 is tight in a certain sense. In Section C.3 we prove a conjecture of [ASSS15], which is similar in spirit to Proposition 8 and leads to an algorithm-independent version of Theorem 7 (with a weaker quantitative bound). Finally, in Section C.4, we discuss another byproduct of our analysis, namely a lower bound for p-SCLIs for convex function minimization.

C.1 p-SCLI lower bounds for the class $\mathcal{F}^{\text{bil}}_{n,\ell,D}$

Notation. For a square matrix $\mathbf{A}$, let $\rho(\mathbf{A})$ be its spectral radius, i.e., the maximum magnitude of an eigenvalue of $\mathbf{A}$. For matrices $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m_1}, \mathbf{A}_2 \in \mathbb{R}^{n_2 \times m_2}$, let $\mathbf{A}_1 \otimes \mathbf{A}_2 \in \mathbb{R}^{(n_1 n_2)\times(m_1 m_2)}$ be the tensor product (also known as Kronecker product) of $\mathbf{A}_1, \mathbf{A}_2$.

We will need the following standard lemma:

Lemma 23. *For a square matrix $\mathbf{C}$ and all $k \in \mathbb{N}$, we have $\|\mathbf{C}^k\|_\sigma \ge \rho(\mathbf{C})^k$.*

Next we prove Theorem 7, restated below for convenience.

Theorem 7 (restated)**.** *Fix $\ell, D > 0$, let $\mathcal{A}$ be a p-SCLI[14], and let $\mathbf{z}^{(t)}$ denote the tth iterate of $\mathcal{A}$. Then there are constants $c_{\mathcal{A}}, T_{\mathcal{A}} > 0$ so that the following holds: For all $T \ge T_{\mathcal{A}}$, there is some $F \in \mathcal{F}^{\text{bil}}_{n,\ell,D}$ so that for some initialization $\mathbf{z}^{(0)}, \dots, \mathbf{z}^{(-p+1)} \in \mathcal{D}_D$ and some $T' \in \{T, T+1, \dots, T+p-1\}$, it holds that $\text{TGap}_F^{\mathcal{D}_{2D}}(\mathbf{z}^{(T')}) \ge \frac{c_{\mathcal{A}} \ell D^2}{\sqrt{T}}$.*

Proof of Theorem 7. Take $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$, where $\mathbf{A}, \mathbf{b}$ are of the form shown in (76), with $\mathbf{M} = \nu \cdot I$ for some $\nu \in (0, \ell]$. Notice that $\mathbf{A}$ therefore depends on the choice of ν (which will be specified later), but for simplicity of notation we do not explicitly write this dependence. The outline of the proof is to first eliminate some corner cases in which the iterates of $\mathcal{A}$ do not converge and then reduce the statement of Theorem 7 to that of Proposition 8. There are a few different ways to carry out this reduction: we follow the linear algebraic approach of [ASSS15], but an approach of a different flavor using elementary ideas from complex analysis is given in [Nev93, Section 3.7].

Since $F \in \mathcal{F}^{\text{bil}}_{n,\ell,D}$, we have that the equilibrium $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{D}_D$ satisfies $\mathcal{B}_{\mathbb{R}^{n/2}}(\mathbf{x}^*, D) \times \mathcal{B}_{\mathbb{R}^{n/2}}(\mathbf{y}^*, D) \subset \mathcal{D}_{2D}$. Then, from [GPDO20, Eq. (22)], it follows that $\text{TGap}_F^{\mathcal{D}_{2D}}(\mathbf{z}) \ge D\|F(\mathbf{z})\|$ for any $\mathbf{z} \in \mathbb{R}^n$. Therefore, to prove Theorem 7 it suffices to show the lower bound $\|F(\mathbf{z}^{(T')})\| \ge \frac{c_{\mathcal{A}} \ell D}{\sqrt{T}}$.

We consider the dynamics of the iterates of $\mathcal{A}$ for various choices of $\mathbf{z}^{(0)}, \dots, \mathbf{z}^{(-p+1)} \in \mathcal{D}_D$. To do so, we

[14] More generally, $\mathcal{A}$ may be any algorithm satisfying the conditions of Observation 22.

define the block matrices:

$$\mathbf{C}(\mathbf{A}) := \begin{pmatrix} \mathbf{0} & I_n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_n & \mathbf{0} & \cdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} & I_n \\ \mathbf{C}_0(\mathbf{A}) & \mathbf{C}_1(\mathbf{A}) & \cdots & \mathbf{C}_{p-2}(\mathbf{A}) & \mathbf{C}_{p-1}(\mathbf{A}) \end{pmatrix}, \qquad \mathbf{U} := \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ I_n \end{pmatrix} \in \mathbb{R}^{pn \times n}, \tag{79}$$

and the block vectors

$$\mathbf{w}^{(t)} := \begin{pmatrix} \mathbf{z}^{(t-p+1)} \\ \mathbf{z}^{(t-p+2)} \\ \vdots \\ \mathbf{z}^{(t)} \end{pmatrix}.$$

Then the updates of $\mathcal{A}$ as in (78) can be written in the following form, for $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$:

$$\mathbf{w}^{(t+1)} = \mathbf{C}(\mathbf{A})\mathbf{w}^{(t)} + \mathbf{U}\mathbf{N}(\mathbf{A})\mathbf{b}.$$

Hence

$$\mathbf{w}^{(t)} = \mathbf{C}(\mathbf{A})^t \cdot \mathbf{w}^{(0)} + \sum_{s=1}^{t} \mathbf{C}(\mathbf{A})^{t-s}\mathbf{U}\mathbf{N}(\mathbf{A})\mathbf{b}. \tag{80}$$

Recall that Observation 22 gives us $\mathbf{C}_j(\mathbf{A}) = \alpha_j \cdot \mathbf{A} + \beta_j \cdot I_n$, and $\mathbf{N}(\mathbf{A}) = \gamma \cdot \mathbf{A} + \delta \cdot I_n$, for some real numbers $\alpha_j, \beta_j, \gamma, \delta$ where $0 \le j \le p-1$.

We now consider several cases:

Case 1: $\mathbf{C}(\mathbf{A}) - I_{np}$ or $\mathbf{N}(\mathbf{A})$ is not invertible for some choice of $\nu \in (0, \ell]$ (which determines $\mathbf{A}$ as explained above). First suppose that $\mathbf{C}(\mathbf{A}) - I_{np}$ is not invertible. Note that the row-space of $\mathbf{C}(\mathbf{A}) - I_{np}$ contains the row-space of the following matrix:

$$\tilde{\mathbf{C}} := \begin{pmatrix} -I_n & I_n & \mathbf{0} & \cdots \\ -I_n & \mathbf{0} & I_n & \cdots \\ \vdots & \ddots & \ddots & \vdots \\ -I_n & \mathbf{0} & \cdots & I_n \\ -I_n + \mathbf{C}_0(\mathbf{A}) & \mathbf{C}_1(\mathbf{A}) & \cdots & \mathbf{C}_{p-1}(\mathbf{A}) \end{pmatrix}.$$

If $\mathbf{C}_0(\mathbf{A}) + \cdots + \mathbf{C}_{p-1}(\mathbf{A}) - I_n$ is full-rank, then the row-space of $\tilde{\mathbf{C}}$ additionally contains the row-space of $\begin{pmatrix} I_n & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n \times np}$, and thus $\tilde{\mathbf{C}}$, and so $\mathbf{C}(\mathbf{A}) - I$ would be full-rank. Thus $\mathbf{C}_0(\mathbf{A}) + \cdots + \mathbf{C}_{p-1}(\mathbf{A}) - I_n$ is not full-rank. But we can write:

$$-I_n + \sum_{j=0}^{p-1} \mathbf{C}_j(\mathbf{A}) = \left(-1 + \sum_{j=0}^{p-1} \beta_j\right) I_n + \left(\sum_{j=0}^{p-1} \alpha_j\right) \cdot \mathbf{A} = \begin{pmatrix} \left(-1 + \sum_{j=0}^{p-1} \beta_j\right) I_{n/2} & \sum_{j=0}^{p-1} \alpha_j \mathbf{M} \\ -\sum_{j=0}^{p-1} \alpha_j \mathbf{M} & \left(-1 + \sum_{j=0}^{p-1} \beta_j\right) I_{n/2} \end{pmatrix}$$

But since $\mathbf{M}$ is a nonzero multiple of the identity matrix, if the above matrix is not full-rank, it must be identically 0, i.e., $\sum_{j=0}^{p-1} \mathbf{C}_j(\mathbf{A}) = I_n$. Hence $\sum_{j=0}^{p-1} \beta_j = 1, \sum_{j=0}^{p-1} \alpha_j = 0$.

Thus, for *any* choice of the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, if we choose $\mathbf{b} = \mathbf{0}$ (so that $F(\mathbf{z}) = \mathbf{A}\mathbf{z}$), and if $\mathbf{z}^{(0)} = \cdots = \mathbf{z}^{(-p+1)} = \mathbf{z}$ for some $\mathbf{z} \in \mathbb{R}^n$, it holds that for all $t \ge 1$, the iterates $\mathbf{z}^{(t)}$ of $\mathcal{Z}$ satisfy $\mathbf{z}^{(t)} = \mathbf{z}$. We now choose $\mathbf{M} = \ell \cdot I_{n/2}$ and $\mathbf{z} = \mathbf{z}^{(0)} = D/\sqrt{n/2} \cdot \mathbf{1} \in \mathbb{R}^n$, so that $\mathbf{z}^{(0)} - \mathbf{A}^{-1}\mathbf{b} = \mathbf{z}^{(0)} \in \mathcal{D}_D$. Then for all $t \ge 0$,

$$\|F(\mathbf{z}^{(t)})\|^2 = \|F(\mathbf{z}^{(0)})\|^2 = 2\ell^2 D^2.$$

Similarly, if $\mathbf{N}(\mathbf{A})$ is not invertible for some choice of $\nu \in (0, \ell]$, then by choice of $\mathbf{A}$ we must have that $\gamma = \delta = 0$, i.e., $\mathbf{N}(\mathbf{A}) = \mathbf{0}$ for all choices of ν. Thus, choosing $\mathbf{w}^{(0)} = \mathbf{0}$ and $\mathbf{b} = D/\sqrt{n/2} \cdot \mathbf{1} \in \mathcal{D}_D$, and so for all $t \ge 0$, $\|F(\mathbf{z}^{(t)})\| = \|F(\mathbf{z}^{(0)})\| = \|\mathbf{b}\| = \sqrt{2}D$. Thus in this case we get the lower bound for $T \ge T_{\mathcal{A}} := \ell^2$.

Cases 2 & 3. In the remaining cases $\mathbf{C}(\mathbf{A}) - I_{np}$ and $\mathbf{N}(\mathbf{A})$ are invertible for all $\nu \in (0, \ell]$. Hence we can rewrite (80) as:

$$\mathbf{w}^{(t)} = \mathbf{C}(\mathbf{A})^t \cdot \mathbf{w}^{(0)} + (\mathbf{C}(\mathbf{A}) - I_{np})^{-1}(\mathbf{C}(\mathbf{A})^t - I_{np})\mathbf{U}\mathbf{N}(\mathbf{A})\mathbf{b}. \tag{81}$$

We consider three further sub-cases:

Case 2. $\rho(\mathbf{C}(\mathbf{A})) \geq 1$ for some $\nu \in (0, \ell]$. Fix such a ν (and thus $\mathbf{A}$). Since $\mathbf{C}(\mathbf{A})$ is invertible, we must in fact have $\rho(\mathbf{C}(\mathbf{A})) > 1$; write $\rho_0 := \rho(\mathbf{C}(\mathbf{A}))$. Again we choose $\mathbf{b} = \mathbf{0}$, so that $\mathbf{w}^{(t)} = \mathbf{C}(\mathbf{A})^t \cdot \mathbf{w}^{(0)}$, and so $(I_p \otimes \mathbf{A})\mathbf{w}^{(t)} = \mathbf{C}(\mathbf{A})^t \cdot (I_p \otimes \mathbf{A})\mathbf{w}^{(0)}$. By Lemma 23 we have that $\|\mathbf{C}(\mathbf{A})^t\|_\sigma \geq \rho_0^t$. Let $\tilde{\mathbf{w}}^{(0)} := ((\tilde{\mathbf{z}}^{(-p+1)})^\top, \ldots, (\tilde{\mathbf{z}}^{(0)})^\top)^\top$ be a singular vector of $\mathbf{C}(\mathbf{A})^t$ corresponding to a singular value which is at least ρ_0^t. By appropriately scaling $\tilde{\mathbf{w}}^{(0)}$, we may ensure that $\tilde{\mathbf{z}}^{(-p+1)}, \ldots, \tilde{\mathbf{z}}^{(0)} \in \mathcal{D}_D$ and $\|\tilde{\mathbf{w}}^{(0)}\| \geq D$. Moreover, we have that $\|(I_p \otimes \mathbf{A})\mathbf{w}^{(t)}\| = \nu\|\mathbf{w}^{(t)}\| \geq \nu\rho_0^t D$. This quantity can be made arbitrarily large by taking t to be arbitrarily large (as $\rho_0 > 1$), and thus in this case $\|F(\mathbf{z}^{(t)})\| = \|\mathbf{A}\mathbf{z}^{(t)}\|$ fails to converge to 0 since $\|(I_p \otimes \mathbf{A})\mathbf{w}^{(t)}\| \to \infty$ as $t \to \infty$.

Case 3. $\rho(\mathbf{C}(\mathbf{A})) < 1$; in this case we have

$$\lim_{t\to\infty} \mathbf{U}^\top \mathbf{w}^{(t)} = -\mathbf{U}^\top(\mathbf{C}(\mathbf{A}) - I_{np})^{-1}\mathbf{U}\mathbf{N}(\mathbf{A})\mathbf{b}.$$

Note that $\mathbf{U}^\top(\mathbf{C}(\mathbf{A}) - I_{np})^{-1}\mathbf{U}$ is the lower $n \times n$-submatrix of the matrix $(\mathbf{C}(\mathbf{A}) - I_{np})^{-1}$, and therefore it must be the inverse of the Schur complement of the upper $(p-1)n \times (p-1)n$-submatrix of $\mathbf{C}(\mathbf{A}) - I_{np}$. Thus $\mathbf{U}^\top(\mathbf{C}(\mathbf{A}) - I_{np})^{-1}\mathbf{U}$ is invertible, and since $\mathbf{N}(\mathbf{A})$ is as well, we may define $\mathbf{B}(\mathbf{A}) := -\left(\mathbf{U}^\top(\mathbf{C}(\mathbf{A}) - I_{np})^{-1}\mathbf{U}\mathbf{N}(\mathbf{A})\right)^{-1}$. Hence $\mathbf{U}^\top(\mathbf{C}(\mathbf{A}) - I_{np})^{-1}\mathbf{U} = -\mathbf{B}(\mathbf{A})^{-1}\mathbf{N}(\mathbf{A})^{-1}$. As shown in [ASSS15, Eqs. (68) – (70)], this implies that $\sum_{j=0}^{p-1} \mathbf{C}_j(\mathbf{A}) = I_n + \mathbf{N}(\mathbf{A})\mathbf{B}(\mathbf{A})$, which can be written as:

$$\left(\sum_{j=0}^{p-1} \alpha_j\right)\mathbf{A} + \left(\sum_{j=0}^{p-1} \beta_j\right) I_n = I + (\gamma\mathbf{A} + \delta I_n) \cdot \mathbf{B}(\mathbf{A}). \tag{82}$$

Let $\mathbf{1}_p \in \mathbb{R}^p$ be the p-vector of ones. The fact that $\mathbf{N}(\mathbf{A})\mathbf{B}(\mathbf{A}) = \sum_{j=0}^{p-1} \mathbf{C}_j(\mathbf{A}) - I_n$ and definition of $\mathbf{U}$ gives

$$\mathbf{U}\mathbf{N}(\mathbf{A})\mathbf{B}(\mathbf{A}) = (\mathbf{C}(\mathbf{A}) - I_{pn})\begin{pmatrix} I_n \\ \vdots \\ I_n \end{pmatrix} = (\mathbf{C}(\mathbf{A}) - I_{pn})(\mathbf{1}_p \otimes I_n) \quad \Rightarrow \quad (\mathbf{C}(\mathbf{A}) - I_{pn})^{-1}\mathbf{U}\mathbf{N}(\mathbf{A})\mathbf{B}(\mathbf{A}) = \mathbf{1}_p \otimes I_n.$$

It then follows from (81) and the fact that $\mathbf{N}(\mathbf{A}), \mathbf{C}(\mathbf{A})$ commute with $\mathbf{A}$ that

$$\begin{aligned}
&(I_p \otimes \mathbf{A})\mathbf{w}^{(t)} + (\mathbf{1}_p \otimes \mathbf{b}) \\
&= (I_p \otimes \mathbf{A})\mathbf{C}(\mathbf{A})^t\mathbf{w}^{(0)} + (I_p \otimes \mathbf{A})(\mathbf{C}(\mathbf{A})^t - I_{np})(\mathbf{C}(\mathbf{A}) - I_{pn})^{-1}\mathbf{U}\mathbf{N}(\mathbf{A})\mathbf{B}(\mathbf{A})\mathbf{B}(\mathbf{A})^{-1}\mathbf{b} + (\mathbf{1}_p \otimes \mathbf{b}) \\
&= (I_p \otimes \mathbf{A})\mathbf{C}(\mathbf{A})^t\mathbf{w}^{(0)} + (\mathbf{C}(\mathbf{A})^t - I_{np})(I_p \otimes \mathbf{A})(\mathbf{1}_p \otimes \mathbf{B}(\mathbf{A})^{-1}\mathbf{b}) + (\mathbf{1}_p \otimes \mathbf{b}) \\
&= (I_p \otimes \mathbf{A})\mathbf{C}(\mathbf{A})^t\mathbf{w}^{(0)} + \mathbf{C}(\mathbf{A})^t(\mathbf{1}_p \otimes \mathbf{A}\mathbf{B}(\mathbf{A})^{-1}\mathbf{b}) + \mathbf{1}_p \otimes (I_n - \mathbf{A}\mathbf{B}(\mathbf{A})^{-1})\mathbf{b}.
\end{aligned} \tag{83}$$

Case 3a. $\sum_{j=0}^{p-1} \beta_j \neq 1$. Taking $\nu \to 0$ (i.e., $\mathbf{A} \to \mathbf{0}$) in (82), we see that $\delta \neq 0$, and moreover $\lim_{\mathbf{A}\to\mathbf{0}} \mathbf{B}(\mathbf{A}) = \delta^{-1}(\sum_{j=0}^{p-1} \beta_j - 1)I_n \neq \mathbf{0}$. Thus, there must be some $\nu_0 \in (0, \ell]$ so that $\mathbf{B}(\mathbf{A}) \neq \mathbf{A}$, and so for this choice of $\nu = \nu_0$, by (83), for an arbitrary choice of $\mathbf{w}^{(0)}$ and for some choice of $\mathbf{b}$ not in the nullspace of $I_n - \mathbf{A}\mathbf{B}(\mathbf{A})^{-1}$ with $\|\mathbf{b}\| = \nu_0 D/\sqrt{n/2} \cdot \mathbf{1}$, the following holds: for some constants $T_0 \in \mathbb{N}$, $c_0 > 0$, for all $t \geq T_0$, we have $\|(I_p \otimes \mathbf{A})\mathbf{w}^{(t)} + (\mathbf{1}_p \otimes \mathbf{b})\| \geq c_0$. This suffices to prove the desired lower bound on $\|F(\mathbf{z}^{(t)})\|$ (in particular, the constant T_0 determines $T_{\mathcal{A}}$ in the theorem statement).

Case 3b. $\sum_{j=0}^{p-1} \beta_j = 1$. This case contains the case in which the iterates $\mathbf{z}^{(t)}$ of the p-SCLI converge to the true solution $-\mathbf{A}^{-1}\mathbf{b}$ for all $\mathbf{A}, \mathbf{b}$, and is thus the main nontrivial case (in particular, it is the case in which we use Proposition 8).

We now choose $\mathbf{b} = \mathbf{0} \in \mathbb{R}^n$, and so $(I \otimes \mathbf{A})\mathbf{w}^{(t)} = \mathbf{C}(\mathbf{A})^t \mathbf{A}\mathbf{w}^{(0)}$ (we use here that $\mathbf{C}_j(\mathbf{A})$ all commute with $\mathbf{A}$). [ASSS15, Lemma 14] gives that the characteristic polynomial of $\mathbf{C}(\mathbf{A})$ is given by

$$\chi_{\mathbf{C}(\mathbf{A})}(\lambda) = (-1)^{pn} \det\left(\lambda^p I_n - \sum_{j=0}^{p-1} \lambda^j \mathbf{C}_j(\mathbf{A})\right).$$

Recall that the assumption of linear coefficient matrices gives us that $\mathbf{C}_j(\mathbf{A}) = \alpha_j \cdot \mathbf{A} + \beta_j \cdot I_n$, where $\mathbf{A}$ is defined as in (76), depending on some matrix $\mathbf{M}$. Recall our choice of $\mathbf{M} = \nu \cdot I_{n/2}$, for some $\nu \in (0, \ell]$, to be specified below. Now define $q(\lambda) := \lambda^p - \sum_{j=0}^{p-1} \beta_j \lambda^j$ and $r(\lambda) := \sum_{j=0}^{p-1} \alpha_j \lambda^j$. Then

$$\lambda^p I_n - \sum_{j=0}^{p-1} \lambda^j \mathbf{C}_j(\mathbf{A}) = q(\lambda) \cdot I_n - r(\lambda) \cdot \mathbf{A} = \begin{pmatrix} q(\lambda) \cdot I_{n/2} & \nu r(\lambda) \cdot I_{n/2} \\ -\nu r(\lambda) \cdot I_{n/2} & q(\lambda) \cdot I_{n/2} \end{pmatrix} = \begin{pmatrix} q(\lambda) & \nu r(\lambda) \\ -\nu r(\lambda) & q(\lambda) \end{pmatrix} \otimes I_{n/2}.$$

By the formula for the determinant of a tensor product of matrices,

$$\chi_{\mathbf{C}(\mathbf{A})}(\lambda) = (-1)^{pn} \cdot (q(\lambda)^2 + \nu^2 r(\lambda)^2)^{n/2},$$

and so the spectral radius of $\mathbf{C}(\mathbf{A})$ is given by $\rho(\mathbf{C}(\mathbf{A})) = \rho(q(\lambda)^2 + \nu^2 r(\lambda)^2)$. Since $\sum_{j=0}^{p-1} \beta_j = 1$, we have that $q(1)^2 = 0$; moreover, $\lambda \mapsto q(\lambda)^2$ is a degree-$2p$ monic polynomial, while $\lambda \mapsto -r(\lambda)^2$ is a degree-$(2(p-1))$ (and thus also degree-$(2p-1)$) polynomial. Thus, by Proposition 8, we get that there are some constants $\mu_{\mathcal{A}}, C_{\mathcal{A}} > 0$ (depending on the algorithm $\mathcal{A}$) so that for any $\mu \in (0, \mu_{\mathcal{A}})$, there is some $\nu \in [\mu, \ell]$ so that $\rho(q(\lambda)^2 + \nu^2 r(\lambda)^2) \geq 1 - C_{\mathcal{A}} \cdot \mu^2/\ell^2$. Let $T_{\mathcal{A}}$ be so that $\ell/(2\sqrt{T_{\mathcal{A}}}) < \mu_{\mathcal{A}}$. Now for any $T \geq T_{\mathcal{A}}$, we may choose $\mu = \ell/(2\sqrt{T})$, and set $\nu \in [\ell/(2\sqrt{T}), \ell]$ accordingly per Proposition 8. By Lemma 23, we have that, for $T \geq T_{\mathcal{A}}$,

$$\|\mathbf{C}(\mathbf{A})^T\|_\sigma \geq \rho(\mathbf{C}(\mathbf{A}))^T \geq (1 - C_{\mathcal{A}}/(4T))^T \geq \exp(-C_{\mathcal{A}}).$$

Set $c_{\mathcal{A}} = \exp(-C_{\mathcal{A}})$. Choose $\mathbf{w}^{(0)} = ((\mathbf{z}^{(-p+1)})^\top, \ldots, (\mathbf{z}^{(0)})^\top)^\top \in \mathbb{R}^{np}$ so that it is a (right) singular vector of $\mathbf{C}(\mathbf{A})^T$ corresponding to a singular value of magnitude at least $c_{\mathcal{A}}$. By scaling $\mathbf{w}^{(0)}$ appropriately, we may ensure that $\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(-p+1)} \in \mathcal{D}_D$, and that $\|\mathbf{w}^{(0)}\| \geq D$. It follows that

$$\|(I_p \otimes \mathbf{A})\mathbf{w}^{(T)}\|^2 = \|(I_p \otimes \mathbf{A})\mathbf{C}(\mathbf{A})^T \mathbf{w}^{(0)}\|^2 \geq c_{\mathcal{A}} \nu^2 D^2 \geq \frac{c_{\mathcal{A}} \ell^2 D^2}{T}.$$

Thus, for some $T' \in \{T, T-1, \ldots, T-p+1\}$, we have that $\|F(\mathbf{z}^{(T')})\| = \|\mathbf{A}\mathbf{z}^{(T')}\| \geq \sqrt{\frac{c_{\mathcal{A}} \ell^2 D^2}{pT'}}$, which establishes the desired lower bound on iteration complexity. ☐

C.2 Proof of Propositions 8 and 9

In this section we prove Propositions 8 and 9.

Proposition 8 (restated)**.** *Suppose $q(z)$ is a degree-p monic real polynomial such that $q(1) = 0$, $r(z)$ is a polynomial of degree $p-1$, and $\ell > 0$. Then there is a constant $C_0 > 0$, depending only on $q(z), r(z)$ and ℓ, and some $\mu_0 \in (0, \ell)$, so that for any $\mu \in (0, \mu_0)$,*

$$\sup_{\nu \in [\mu, \ell]} \rho(q(z) - \nu \cdot r(z)) \geq 1 - C_0 \cdot \frac{\mu}{\ell}.$$

Proof. Let $\Delta \subset \mathbb{C}$ be the unit disk in the complex plane centered at 0 and of radius 1. Set $R(z)$ to be the rational function $R(z) := \frac{q(z)}{r(z)}$. Our goal is to find some μ_0 so that for any $\mu < \mu_0$, we have

$$[\mu, \ell] \cap \{R(z) : |z| \geq 1 - C_0 \cdot \mu/\ell\} \neq \emptyset. \tag{84}$$

We may assume $r(1) \neq 0$ (if instead $r(1) = 0$, then $q(1) - \nu \cdot r(1) = 0$ for all ν, and the proof is complete). Hence $R(1) = 0$, and R is nonconstant. Since $R(z)$ is holomorphic in a neighborhood of 1, there are neighborhoods $U \ni 1$ and $V \ni 0$, with $R(U) = V$, together with conformal mappings $a : \Delta \to U$ with

$a(0) = 1$, and $b : V \to \Delta$ with $b(0) = 0$, which extend to continuous functions on $\bar{\Delta}, \bar{V}$, respectively, so that the mapping $\tilde{R} : \Delta \to \Delta$, defined by $\tilde{R} = b \circ R \circ a$, satisfies $\tilde{R}(w) = w^k$ for some $k \geq 1$.

By Cauchy's integral formula, there is a positive constant A_0, depending only on the function $R(\cdot)$, so that for $w \in \Delta$, we have that

$$|a(w) - (1 + a'(0) \cdot w)| \leq A_0 \cdot |w|^2$$

and for $z \in V$, we have that

$$|b(z) - b'(0) \cdot z| \leq A_0 \cdot |z|^2.$$

By choosing $\mu_0 > 0$ to be sufficiently small, we may ensure that $[0, \mu_0] \subset V$. Now fix any $\mu \in (0, \mu_0)$. We consider several cases:

Case 1. $k = 1$. Let $w_0 = b(\mu)$, so that

$$|w_0| \leq |b'(0)| \cdot \mu + A_0 \cdot \mu^2 \leq A_1 \cdot \mu \tag{85}$$

for some constant $A_1 > 0$. We have that $R(a(w_0)) = \mu$ by definition of $a(z)$. Moreover,

$$|a(w_0)| \geq |1 + a'(0) \cdot w_0| - A_0 \cdot |w_0|^2 \geq 1 - |a'(0)| \cdot (|b'(0)| \cdot \mu + A_0 \cdot \mu^2) - A_0 \cdot A_1^2 \mu^2,$$

and thus as long as C_0 is chosen sufficiently large as a function of $|a'(0)|, |b'(0)|, A_0, A_1, \ell$, we have $|a(w_0)| \geq 1 - C_0 \cdot \mu/\ell$, and $R(a(w_0)) = \mu$, and thus (84) is satisfied in this case.

Case 2. $k = 2$. Again let $w_0 = b(\mu)$, so that (85) holds. Let $u_0 \in \Delta$ be a square root of w_0, i.e., $u_0^2 = (-u_0)^2 = w_0$. Then $R(a(u_0)) = R(a(-u_0)) = \mu$. It must be the case that either $a'(0) \cdot u_0$ or $-a'(0) \cdot u_0$ has a non-negative real part; suppose without loss of generality that it is $a'(0) \cdot u_0$ (if not, then replace u_0 with $-u_0$). Then

$$|a(u_0)| \geq |1 + a'(0) \cdot u_0| - A_0 \cdot |u_0|^2 \geq \sqrt{1 + |a'(0) \cdot u_0|^2} - A_0 \cdot |w_0| \geq \sqrt{1} - A_0 A_1 \mu,$$

and thus as long as C_0 is chosen sufficiently large as a function of A_0, A_1, ℓ, we have that $|a(u_0)| \geq 1 - C_0 \cdot \mu/\ell$ and $R(a(u_0)) = \mu$, and again (84) is satisfied in this case.

Case 3. $k \geq 3$. In this case we have that $|R(1 - z)| \leq O(|z|^3)$ as $z \to 0$, so there are some constants $\mu_0, C > 0$ so that for $\mu \in (0, \mu_0)$ we have that any root z of $z \mapsto q(z) - \mu \cdot r(z)$ must satisfy $|z - 1| \geq C \sqrt[3]{\mu/\ell}$. Theorem 25 (in the following section) implies that $\sup_{\nu \in [\mu, \ell]} \rho(q(z) - \nu \cdot r(z)) \geq 1 - 3\sqrt{\mu/\ell}$ for all $\mu \in [0, \ell]$. By making μ_0 smaller if necessary we may assume without loss that for any $\mu \in [0, \mu_0]$, it holds that $3\sqrt{\mu/\ell} < C \sqrt[3]{\mu/\ell}$. If it holds that $\sup_{\nu \in [\mu_0, \ell]} \rho(q(z) - \nu \cdot r(z)) \geq 1$, then the lemma is established for this case. Otherwise, there is some $\mu' \in (0, \mu_0)$ so that for some $\nu \in [\mu', \mu_0]$ we have $\rho(q(z) - \nu \cdot r(z)) \geq 1 - 3\sqrt{\mu'/\ell}$. But since $\mu' \leq \nu \leq \mu_0$ we also have

$$|\rho(q(z) - \nu \cdot r(z)) - 1| \geq C \sqrt[3]{\nu/\ell} > 3\sqrt{\nu/\ell} \geq 3\sqrt{\mu'/\ell},$$

and so it must be the case that $\rho(q(z) - \nu \cdot r(z)) \geq 1 + 3\sqrt{\mu'/\ell} \geq 1$, which establishes the lemma in this case.

We remark also that the case $k \geq 3$ can be dealt with directly, without appealing to Theorem 25: again let $w_0 = b(\mu)$, so that (85) holds. Then there exists some kth root $u_0 \in \Delta$ of w_0 so that $a'(0) \cdot u_0 = re^{i\theta}$ for some $\theta \in [-\pi/3, \pi/3]$ and $r > 0$. Then

$$|a(u_0)| \geq |1 + a'(0) \cdot u_0| - A_0 \cdot |u_0|^2 \geq \frac{1}{\sqrt{3}} |a'(0)| \cdot |u_0| + 1 - A_0 \cdot |u_0|^2 \geq 1$$

for sufficiently small u_0 (which can be made arbitrarily small by taking $\mu \downarrow 0$). □

Proposition 9 (restated). *For any constant $C_0 > 0$ and $\mu_0 \in (0, \ell)$, there is some $\mu \in (0, \mu_0)$ and polynomials $q(z), r(z)$ so that $\sup_{\nu \in [\mu, \ell]} \rho(q(z) - \nu \cdot r(z)) < 1 - C_0 \cdot \mu$. Moreover, the choice of the polynomials is given by*

$$q(z) = \ell(z - \alpha)(z - 1), \qquad r(z) = -(1 + \alpha)z + \alpha \qquad \textit{for} \qquad \alpha := \frac{\sqrt{\ell} - \sqrt{\mu}}{\sqrt{\ell} + \sqrt{\mu}}. \tag{86}$$

Proof of Proposition 9. The proof of this proposition involves similar calculations as were done in [ASSS15, Section 5.2], but we spell them out in detail for completeness.

Fix $C_0 > 0, \mu_0 \in (0, \ell)$. We will show that for some $\mu \in (0, \mu_0)$, we have that $\rho(q(z) - \nu \cdot r(z)) < 1 - C_0 \cdot \mu$ for all $\nu \in [\mu, \ell]$, for the choice of $q(z), r(z), \alpha$ in (86).

Fix any $\nu \in [\mu, \ell]$. Solving $q(z) - \nu \cdot r(z) = 0$ gives

$$z = \frac{(\alpha+1)(1-\nu/\ell) \pm \sqrt{(\alpha+1)^2(1-\nu/\ell)^2 - 4\alpha}}{2}. \tag{87}$$

Let us write $\alpha = \frac{\sqrt{\ell}-\sqrt{\mu}}{\sqrt{\ell}+\sqrt{\mu}} = 1 - 2\epsilon$ for some $\epsilon \in [\sqrt{\mu/\ell}, 2\sqrt{\mu/\ell}]$. Note that, since $\nu \geq \mu$,

$$(\alpha+1)^2(1-\nu/\ell)^2 - 4\alpha \leq (\alpha+1)^2(1-\mu/\ell)^2 - 4\alpha = 4((1-\sqrt{\mu/\ell})^2 - \alpha) < 0,$$

so the values of z in (87) have absolute value equal to $\sqrt{\alpha} \leq 1 - \epsilon \leq 1 - \sqrt{\mu/\ell}$ for any $\nu \in [\mu, \ell]$. For sufficiently small μ, we have $\sqrt{\mu/\ell} > C_0\mu$, and thus $1 - \sqrt{\mu/\ell} < 1 - C_0\mu$. □

The polynomials in (86) are closely related to Nesterov's accelerated gradient descent (AGD); we discuss this connection further in Remark 8.

C.3 Proof of a conjecture of [ASSS15]

In this section we prove the following conjecture:

Conjecture 24 ([ASSS15])**.** *Suppose $q(z)$ is a degree-p monic real polynomial such that $q(1) = 0$. Then for any polynomial $r(z)$ of degree $p-1$ and for any $0 < \mu < \ell$, there exists $\nu \in [\mu, \ell]$ so that*

$$\rho(q(z) - \nu \cdot r(z)) \geq \frac{\sqrt{\ell/\mu} - 1}{\sqrt{\ell/\mu} + 1}. \tag{88}$$

Theorem 25. *Conjecture 24 is true.*

We are not aware of any reference in the literature directly claiming to prove the statement of Conjecture 24. However, we will show two distinct proofs of Conjecture 24: the first is an indirect proof showing how Conjecture 24 may be derived indirectly as a consequence of prior works ([Nev93, AS16]), and the second is a direct proof using basic principles from complex analysis.

Before continuing, we introduce some further notation.

Notation. For a polynomial $s(z)$, write $\rho(s)$ to be the spectral radius of s, i.e., $\rho(s) = \max\{|z| : s(z) = 0\}$ is the maximum magnitude of a root of s. Let $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ denote the Riemann sphere. For $z \in \mathbb{C}, r > 0$, let $D(z, r) := \{w \in \mathbb{C} : |w - z| < r\}$ denote the (open) disk of radius r centered at z. Set $\Delta = D(0, 1)$ and $\mathbb{H} := \{z \in \mathbb{C} : \Im(z) > 0\}$ to be the upper half-plane (here $\Im(z)$ denotes the imaginary part of z). We refer the reader to [Ahl79] for further background on complex analysis.

Indirect proof of Theorem 25 using prior works. We first make the simplifying assumption that there is no $\nu \in [\mu, \ell]$ so that $q(z) - \nu \cdot r(z) = z^p$. (We remove this assumption at the end of the proof.) Let us write $q(z) = z^p - q_{p-1}z^{p-1} - \cdots - q_1 z - q_0, r(z) = r_0 + r_1 z + \cdots + r_{p-1}z^{p-1}$. We have that $q_0 + \cdots + q_{p-1} = 1$ since $q(1) = 0$. Similar to the proof of Theorem 7, define, for $\nu \in [\mu, \ell]$,

$$\mathbf{C}(\nu) := \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & 1 \\ C_0(\nu) & C_1(\nu) & \cdots & C_{p-2}(\nu) & C_{p-1}(\nu) \end{pmatrix},$$

where $C_j(\nu) = q_j + r_j\nu$ for $0 \le j \le p-1$. By our initial simplifying assumption, there is no $\nu \in [\mu, \ell]$ so that $C_0(\nu) = \cdots = C_{p-1}(\nu) = 0$. Then by [ASSS15, Lemma 14], we have that

$$\rho(\mathbf{C}(\nu)) = \rho\left(z^p - \sum_{j=0}^{p-1} C_j(\nu) z^j\right) = \rho\left(q(z) - \nu \cdot r(z)\right). \tag{89}$$

Let $\mathbf{e} := \frac{1}{\sqrt{p}}(1, 1, \ldots, 1)^\top \in \mathbb{R}^p$. Note that $\mathbf{e}^\top \mathbf{C}(\nu)^t \mathbf{e}$ is a polynomial in ν, which we write as $p_t(\nu)$, of degree at most t. It is also immediate that $p_t(0) = 1$ for all t. Moreover, p_t satisfies

$$|p_t(\nu)| \le |\mathbf{e}^\top \mathbf{C}(\nu)^t \mathbf{e}| \le \|\mathbf{c}(\nu)^t \mathbf{e}\| \le \|\mathbf{C}(\nu)^t\|_\sigma. \tag{90}$$

Next we will need the following lemma:

Lemma 26. *It holds that*

$$\sup_{\nu \in [\mu,\ell]} \rho(\mathbf{C}(\nu)) = \sup_{\nu \in [\mu,\ell]} \liminf_{t\to\infty} \|\mathbf{C}(\nu)^t\|_\sigma^{1/t} \ge \liminf_{t\to\infty} \sup_{\nu \in [\mu,\ell]} \|\mathbf{C}(\nu)^t\|_\sigma^{1/t}. \tag{91}$$

Notice that the opposite direction of the inequality in (91) holds trivially, and thus we have equality. Notice also that the first equality in (91) follows by Gelfand's formula.

Proof of Lemma 26. Note that if at least one of $C_0(\nu), \ldots, C_{p-1}(\nu)$ is nonzero, then $\mathbf{C}(\nu)^p \ne \mathbf{0}$: this is the case since there is some vector $\mathbf{v} \in \mathbb{R}^p$ so that $\langle \mathbf{v}, (C_0(\nu), \ldots, C_{p-1}(\nu))\rangle \ne 0$, and the first entry of $\mathbf{C}(\nu)^p \mathbf{v}$ is $\langle \mathbf{v}, (C_0(\nu), \ldots, C_{p-1}(\nu))\rangle$. Since $[\mu, \ell]$ is compact, it follows that the function $\nu \mapsto \frac{\|\mathbf{C}(\nu)\|^p}{\|\mathbf{C}(\nu)^p\|}$ is bounded for $\nu \in [\mu, \ell]$. Let $S := \sup_{\nu\in[\mu,\ell]} \frac{\|\mathbf{C}(\nu)\|^p}{\|\mathbf{C}(\nu)^p\|}$, $\sigma := \max\left\{1/2, \frac{\ln(p-1)}{\ln(p)}\right\}$, and $A_p = 2^p$. Then [Koz09, Theorem 1] gives that for all $\nu \in [\mu, \ell]$ and $t \ge 1$, we have

$$\begin{aligned}
\|\mathbf{C}(\nu)^t\|^{1/t} &\le \rho(\mathbf{C}(\nu)) \cdot A_p^{A_p \cdot t^{\sigma-1}} \cdot \left(\frac{\|\mathbf{C}(\nu)\|^p}{\|\mathbf{C}(\nu)^p\|}\right)^{A_p \cdot t^{\sigma-1}} \\
&\le \rho(\mathbf{C}(\nu)) \cdot A_p^{A_p \cdot t^{\sigma-1}} \cdot S^{A_p \cdot t^{\sigma-1}}.
\end{aligned}$$

Since $\sigma < 1$, it follows that

$$\liminf_{t\to\infty} \sup_{\nu\in[\mu,\ell]} \|\mathbf{C}(\nu)^t\|^{1/t} \le \liminf_{t\to\infty} \sup_{\nu\in[\mu,\ell]} \rho(\mathbf{C}(\nu)) \cdot A_p^{A_p \cdot t^{\sigma-1}} \cdot S^{A_p \cdot t^{\sigma-1}} = \sup_{\nu\in[\mu,\ell]} \rho(\mathbf{C}(\nu)).$$

□

By (90) and Lemma 26, we have

$$\begin{aligned}
\liminf_{t\to\infty} \sup_{\nu\in[\mu,\ell]} |p_t(\nu)|^{1/t} &= \liminf_{t\to\infty} \sup_{\nu\in[\mu,\ell]} |\mathbf{e}^\top \mathbf{C}(\nu)^t \mathbf{e}|^{1/t} \\
&\le \liminf_{t\to\infty} \sup_{\nu\in[\mu,\ell]} \|\mathbf{C}(\nu)^t \mathbf{e}\|^{1/t} \\
&\le \liminf_{t\to\infty} \sup_{\nu\in[\mu,\ell]} \|\mathbf{C}(\nu)^t\|_\sigma^{1/t} \\
&\le \sup_{\nu\in[\mu,\ell]} \rho(\mathbf{C}(\nu)).
\end{aligned} \tag{92}$$

(We use Lemma 26 in (92).) Let $\mathcal{S}_t$ denote the set of polynomials s_t with complex coefficients of degree at most t such that $s_t(0) = 1$. (Note in particular that the polynomials p_t defined above belong to $\mathcal{S}_t$ for each t.) It follows from Theorem 3.6.3, and Example 3.8.3 of [Nev93] that

$$\inf_{t>0} \inf_{s_t\in\mathcal{S}_t} \sup_{\nu\in[\mu,\ell]} |s_t(\nu)|^{1/t} = \frac{\sqrt{\ell/\mu} - 1}{\sqrt{\ell/\mu} + 1}. \tag{93}$$

(In more detail, the quantity on the left-hand-side of (93), which is called the *optimal reduction factor* of the region $[\mu, \ell]$ in [Nev93] and denoted by $\eta_{[\mu,\ell]}$ therein, is shown in [Nev93, Theorem 3.6.3] to be equal to $e^{-G(0)}$, where $G : \mathbb{C} - [\mu, \ell] \to \mathbb{R}$ is the Green's function for the region $\mathbb{C} - [\mu, \ell]$. Then [Nev93, Example 3.8.3] explicitly computes the Green's function and shows that $e^{-G(0)}$ is the quantity on the right-hand-side of (93)).

Combining (89), (92), and (93), we see that

$$\sup_{\nu \in [\mu,\ell]} \rho(q(z) - \nu \cdot r(z)) = \sup_{\nu \in [\mu,\ell]} \rho(\mathbf{C}(\nu)) \geq \inf_{t>0} \inf_{s_t \in \mathcal{S}_t} \sup_{\nu \in [\mu,\ell]} |s_t(\nu)|^{1/t} = \frac{\sqrt{\ell/\mu} - 1}{\sqrt{\ell/\mu} + 1}.$$

Finally, we deal with the case that for some $\nu \in [\mu, \ell]$, we have $q(z) - \nu \cdot r(z) = z^p$. Since the roots of a polynomial are continuous functions of its coefficients and a continuous function defined on a compact set is uniformly continuous, for any $\epsilon > 0$, there is some $\delta > 0$ so that for any polynomial $\tilde{r}(z) = \tilde{r}_0 + \cdots + \tilde{r}_p z^{p-1}$ with $|\tilde{r}_j - r_j| \leq \delta$ for each j, we have that $|\rho(q(z) - \nu \cdot r(z)) - \rho(q(z) - \nu \cdot \tilde{r}(z))| \leq \epsilon$ for all $\nu \in [\mu, \ell]$. Such a polynomial $\tilde{r}$ may be found so that $q(z) - \nu \cdot \tilde{r}(z) \neq z^p$ for all $\nu \in [\mu, \ell]$, and so by the proof above we have

$$\sup_{\nu \in [\mu,\ell]} \rho(q(z) - \nu \cdot r(z)) \geq \sup_{\nu \in [\mu,\ell]} \rho(q(z) - \nu \cdot \tilde{r}(z)) - \epsilon \geq \frac{\sqrt{\ell/\mu} - 1}{\sqrt{\ell/\mu} + 1} - \epsilon.$$

The desired conclusion follows by taking $\epsilon \downarrow 0$, thus completing the proof of Theorem 25.

We remark that an alternative approach to establishing (93) without appealing to the heavy machinery of Green's functions is to use [AS16, Lemma 2] directly, which shows that

$$\inf_{s_t \in \mathcal{S}_t} \sup_{\nu \in [\mu,\ell]} |s_t(\nu)| \geq \left(\frac{\sqrt{\ell/\nu} - 1}{\sqrt{\ell/\nu} + 1} \right)^t.$$

□

The approach to proving Conjecture 24 described above is unsatisfying in that it first passes a statement about polynomials (namely, Conjecture 24) to a statement about matrices (namely, about $\liminf_t \sup_{\nu \in [\mu,\ell]} \|\mathbf{C}(\nu)^t\|_\sigma^{1/t}$), relying on a nontrivial uniform version of Gelfand's formula ([Koz09]), before passing back to a statement about polynomials and using either [AS16] or [Nev93] to establish (93). It is natural to wonder whether there is a *direct* proof of Conjecture 24 which operates on the polynomials $q(z), r(z)$ directly, without bounding the optimal reduction factor in (93) and constructing the matrices $\mathbf{C}(\nu)$. We next give such a direct proof of Conjecture 24, which follows from basic facts from complex analysis.

Direct proof of Theorem 25. Fix some polynomials q, r satisfying the conditions of Conjecture 24. Choose $\delta \in \mathbb{R}$ so that $\max_{\nu \in [\mu,\ell]} \rho(q(z) - \nu \cdot r(z)) = 1 - \delta$. Notice that the maximum exists since the roots of a polynomial are a continuous function of its coefficients. Our goal is to show that $\delta \leq 1 - \frac{\sqrt{\ell/\mu}-1}{\sqrt{\ell/\mu}+1}$. Define the rational function $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ by $R(z) = \frac{q(z)}{r(z)}$. If, for some z_0 with $|z_0| > 1 - \delta$, $R(z_0) =: \nu \in [\mu, \ell]$, then we have $q(z_0) - \nu \cdot r(z_0) = 0$, and so $\rho(q(z) - \nu \cdot r(z)) \geq |z_0| > 1 - \delta$, a contradiction. Hence the restriction of R to $\hat{\mathbb{C}} - \overline{D(0, 1-\delta)}$ is in fact a holomorphic function to the Riemann surface $\hat{\mathbb{C}} - [\mu, \ell]$, i.e., $R : \hat{\mathbb{C}} - \overline{D(0, 1-\delta)} \to \hat{\mathbb{C}} - [\mu, \ell]$. (Recall that $\overline{D(0, 1-\delta)}$ denotes the closed disc of radius $1 - \delta$ centered at 0.) We next need the following standard lemma:

Lemma 27. *There is a holomorphic map $G : \hat{\mathbb{C}} - [\mu, \ell] \to \Delta$ from $\hat{\mathbb{C}} - [\mu, \ell]$ to the unit disk Δ, so that $G(0) = \frac{1-\sqrt{\ell/\mu}}{1+\sqrt{\ell/\mu}}$ and $G(\infty) = 0$.*[15]

For completeness we prove Lemma 27 below; we first complete the proof of Theorem 25 assuming Lemma 27.

[15] In fact, G is a conformal mapping, though we will not need this.

Notice that the mapping $z \mapsto \frac{1}{z}$ maps $D\left(0, \frac{1}{1-\delta}\right)$ to $\hat{\mathbb{C}} - \overline{D(0, 1-\delta)}$. Thus we may define $\tilde{R} : D\left(0, \frac{1}{1-\delta}\right) \to \hat{\mathbb{C}} - [\mu, \ell]$ by $\tilde{R}(z) = R\left(\frac{1}{z}\right)$, which is holomorphic since R is. Now define the function $H : \Delta \to \Delta$ by

$$H(z) = G\left(\tilde{R}\left(\frac{1}{1-\delta} \cdot z\right)\right),$$

which is well-defined since $\frac{1}{1-\delta} \cdot z \in D\left(0, \frac{1}{1-\delta}\right)$ for $z \in \Delta$. Since H is a composition of the holomorphic functions $z \mapsto \frac{1}{1-\delta} \cdot z$, $\tilde{R}$, and G, H is itself holomorphic. Note that

$$H(0) = G(\tilde{R}(0)) = G(R(\infty)) = G(\infty) = 0 \tag{94}$$

$$H(1-\delta) = G(\tilde{R}(1)) = G(R(1)) = G(0) = \frac{1 - \sqrt{\ell/\mu}}{1 + \sqrt{\ell/\mu}}. \tag{95}$$

where to derive (94) we used that $R(\infty) = \infty$ since $q(z)$ is monic of degree p and $r(z)$ is of degree $p-1$, and to derive (95) we used that $R(1) = 0$ since $q(1) = 0$ by assumption.

Next we recall the Schwarz lemma from elementary complex analysis:

Lemma 28 (Schwarz)**.** *A holomorphic function $f : \Delta \to \Delta$ with $f(0) = 0$ satisfies $|f(z)| \le |z|$ for all $z \in \Delta$.*

Since $H : \Delta \to \Delta$ is holomorphic, satisfies $H(0) = 0$ (by (94)), (95) together with Lemma 28 gives us that

$$|H(1-\delta)| = \frac{\sqrt{\ell/\mu} - 1}{\sqrt{\ell/\mu} + 1} \le 1 - \delta.$$

In particular, $\delta \le 1 - \frac{\sqrt{\ell/\mu}-1}{\sqrt{\ell/\mu}+1}$, which completes the proof.

□

Now we prove Lemma 27 for completeness.

Proof of Lemma 27. We will take

$$G(w) := \frac{\sqrt{w-\mu} - i\sqrt{\ell - w}}{\sqrt{w-\mu} + i\sqrt{\ell - w}},$$

where the choice of the branch of the square root will be explained below. In particular, G is obtained as the composition of maps $G = G_5 \circ G_4 \circ G_3 \circ G_2 \circ G_1$, where $G_1, \ldots, G_5$ are defined by:

$$\begin{aligned} G_1 : \hat{\mathbb{C}} - [\mu, \ell] \to \hat{\mathbb{C}} - [0, 1], &\qquad w \mapsto \frac{\ell - w}{\ell - \mu} \\ G_2 : \hat{\mathbb{C}} - [0, 1] \to \hat{\mathbb{C}} - [1, \infty], &\qquad w \mapsto 1/w \\ G_3 : \hat{\mathbb{C}} - [1, \infty] \to \hat{\mathbb{C}} - [0, \infty], &\qquad w \mapsto w - 1 \\ G_4 : \hat{\mathbb{C}} - [0, \infty] \to \mathbb{H}, &\qquad w \mapsto \sqrt{w} \\ G_5 : \mathbb{H} \to \Delta, &\qquad w \mapsto \frac{w - i}{w + i}, \end{aligned} \tag{96}$$

where the choice of the branch of the square root in (96) is given by $G_4(re^{i\theta}) = \sqrt{r}e^{i\theta/2}$ for $r > 0, \theta \in (0, 2\pi)$. It is clear that each of $G_1, \ldots, G_5$ are holomorphic functions between their respective Riemann surfaces, and thus $G : \hat{\mathbb{C}} - [\mu, \ell] \to \Delta$ is holomorphic.

To verify the values of $G(0), G(\infty)$, note that $G_3(G_2(G_1(0))) = -\mu/\ell$ and $G_3(G_2(G_1(\infty)) = -1$. By the choice of the branch of the square root defining G_4, we have that $G_4(G_3(G_2(G_1(0)))) = i\sqrt{\mu/\ell}$ and $G_4(G_3(G_2(G_1(\infty))) = i$. It follows that $G(\infty) = 0$ and $G(0) = \frac{1-\sqrt{\ell/\mu}}{\sqrt{\ell/\mu}+1}$. □

Theorem 25 leads to an algorithm-independent version of Theorem 7. We need the following definition: a p-SCLI in the form (78) with $\mathbf{C}_j(\mathbf{A}) = \alpha_j \mathbf{A} + \beta_j I_n$ is called *consistent* ([ASSS15]) if $\sum_{j=0}^{p-1} \beta_j = 1$. It is known that if the iterates of $\mathcal{A}$ converge for all $\mathbf{b} \in \mathbb{R}^n$, then $\mathcal{A}$ is consistent; hence consistent p-SCLIs represent all "useful" ones.

Proposition 29. *Let $\mathcal{A}$ be a consistent p-SCLI and let $\mathbf{z}^{(t)}$ denote the tth iterate of $\mathcal{A}$. Then for all $T \in \mathbb{N}$, there is some $F \in \mathcal{F}^{\text{bil}}_{n,\ell,D}$ so that for some initialization $\mathbf{z}^{(0)}, \dots, \mathbf{z}^{(-p+1)} \in \mathcal{D}_D$ and some $T' \in \{T, T-1, \dots, T-p+1\}$, it holds that $\text{TGap}_F^{\mathcal{D}_{2D}}(\mathbf{z}^{(T')}) \geq \frac{\ell D^2}{\sqrt{20pT}}$.*

Proof. The proof of Proposition 29 mirrors nearly exactly the proof of Theorem 7, except we need only consider Case 3b by consistency. Moreover, the only difference to Case 3b is the following: instead of applying Proposition 8, we apply Theorem 25 (i.e., Conjecture 24) with $\mu = \ell/2T$. Then, we may choose $\mu \in [\ell/2T, \ell]$ accordingly per the statement of Conjecture 24 to conclude that

$$\rho(\mathbf{C}(\mathbf{A}))^T \geq \left(\frac{2T-1}{2T+1}\right)^T \geq 1/5.$$

Thus it follows in the same way as in the proof of Theorem 7 that for some $T' \in \{T, T-1, \dots, T-p+1\}$ we have that $\|F(\mathbf{z}^{(T')})\| \geq \sqrt{\frac{\nu^2 D^2}{5pT}} \geq \sqrt{\frac{\ell^2 D^2}{20pT}}$. ∎

The conclusion of Proposition 29 is known even for non-stationary p-CLIs and without the superfluous $1/\sqrt{p}$ factor (e.g., it follows from Proposition 5 in [ASM+20]), but our proof is new since it involves Theorem 25, which does not seem to have been previously known in the literature. We are hopeful that Theorem 25 may have further consequences for proving lower bounds for optimization algorithms, such as in the stochastic setting.

C.4 Byproduct: Lower bound for convex function minimization

In this section we prove an (algorithm-dependent) lower bound of $\Omega(1/T)$ on the rate of convergence for p-SCLIs for convex function minimization. This statement was claimed to be proven by [AS16, Corollary 1], but in fact their results only give a linear lower bound for the strongly convex case (and not the sublinear bound of $\Omega(1/T)$ we obtain here): in particular, Corollary 1 of [AS16] is a corollary of Theorem 2 of [AS16], which should be adjusted to state that the error after T iterations cannot be upper bounded by $O\left((1-(\mu/L)^\alpha)^T\right)$, for any $\alpha < 1$.[16] This weaker version of [AS16, Theorem 2] does not imply [AS16, Corollary 1].

In this section, we show that Proposition 8 can be used to correct the above issue in [AS16]. We first introduce the function class of "hard" functions, analogously to $\mathcal{F}^{\text{bil}}_{n,\ell,D}$. Let $\mathcal{F}^{\text{quad}}_{n,\ell,D}$ be the class of ℓ-smooth[17] functions $f : \mathbb{R}^n \to \mathbb{R}$ of the form

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{S}\mathbf{x} + \mathbf{b}^\top \mathbf{x},$$

for which $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a positive definite matrix and $\mathbf{x}^* := -\mathbf{S}^{-1}\mathbf{b}$ has norm $\|\mathbf{x}^*\| \leq D$. We prove the following lower bound for p-SCLI algorithms using functions from $\mathcal{F}^{\text{quad}}_{n,\ell,D}$

Proposition 30. *Let $\mathcal{A}$ be a p-SCLI, and let $\mathbf{x}^{(t)}$ denote the tth iterate of $\mathcal{A}$. Then there are constants $c_{\mathcal{A}}, T_{\mathcal{A}} > 0$ so that the following holds: for all $T \geq T_{\mathcal{A}}$, there is some $f \in \mathcal{F}^{\text{quad}}_{n,\ell,D}$ so that for some initialization $\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(-p+1)} \in \mathcal{B}(\mathbf{0}, D)$ and some $T' \in \{T, T+1, \dots, T+p-1\}$, it holds that $f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \geq \frac{c_{\mathcal{A}} \ell D^2}{T}$.*

Proof of Proposition 30. Note that for any $\mathbf{x} \in \mathbb{R}^n$, we have that

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2}\mathbf{x}^\top \mathbf{S}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + \frac{1}{2}\mathbf{b}^\top \mathbf{S}^{-1}\mathbf{b} = \frac{1}{2}(\mathbf{S}\mathbf{x} + \mathbf{b})^\top \mathbf{S}^{-1}(\mathbf{S}\mathbf{x} + \mathbf{b}).$$

[16] In particular, this modified version can be established by only using functions for which the condition number L/μ is a constant. In more detail, one runs into the following issue when using the machinery of [AS16] to attempt to prove that the iteration complexity of a p-SCLI cannot be $O(\kappa^\alpha \ln(1/\epsilon))$ for any $\alpha < 1$: at the end of the proof of [AS16, Theorem 2], Lemma 4 of [AS16] is used to conclude the existence of some $\eta \in (L/2, L)$ satisfying a certain inequality. However, L/η represents the condition number κ of the problem, and so choosing $\eta \in (L/2, L)$ forces the condition number κ of the function to be a constant.

[17] Recall that f is ℓ-smooth iff its gradient is ℓ-Lipschitz.

Define, for each $t \geq 0$,

$$\mathbf{w}^{(t)} := \begin{pmatrix} \mathbf{x}^{(t-p+1)} \\ \mathbf{x}^{(t-p+2)} \\ \vdots \\ \mathbf{x}^{(t)} \end{pmatrix}.$$

We will choose $\mathbf{S} = \nu \cdot I_n$, for some $\nu \in (0, \ell]$ to be chosen later. Thus $f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2\nu}\|\mathbf{S}\mathbf{x} + \mathbf{b}\|^2$. Next we proceed exactly as in the proof of Theorem 7, with $\mathbf{S}$ taking the role of $\mathbf{A}$ there. In particular, we define $\mathbf{C}(\mathbf{S})$ exactly as in (79), where $\mathbf{C}_j(\mathbf{S}) = \alpha_j \cdot \mathbf{A} + \beta_j \cdot I_n, \mathbf{N}(\mathbf{S}) = \gamma \cdot \mathbf{S} + \delta \cdot I_n$, where $\alpha_j, \beta_j, \gamma, \delta \in \mathbb{R}$ are the constants associated with the p-SCLI $\mathcal{A}$. Cases 1, 2, and 3a of the proof (namely, the ones in which the algorithm does not converge) proceed in exactly the same way and we omit the details.

To deal with Case 3b (i.e., the case that $\sum_{j=0}^{p-1} \beta_j = 1$), we choose $\mathbf{b} = \mathbf{0} \in \mathbb{R}^n$, and (83) gives us that $(I_p \otimes \mathbf{S})\mathbf{w}^{(t)} = \mathbf{C}(\mathbf{S})^t \mathbf{S}\mathbf{w}^{(0)}$. Moreover, it follows from [ASSS15, Lemma 14] that

$$\rho(\mathbf{C}(\mathbf{S})) = \rho(q(z) - \nu \cdot r(z)).$$

By Proposition 8, there are some constants $\mu_{\mathcal{A}}, C_{\mathcal{A}} > 0$ so that for any $\mu \in (0, \mu_{\mathcal{A}})$, there is some $\nu \in [\mu, \ell]$ so that $\rho(q(z) - \nu \cdot r(z)) \geq 1 - C_{\mathcal{A}} \cdot \mu/\ell$. Letting $T_{\mathcal{A}}$ be so that $\ell/(4T_{\mathcal{A}}) < \mu_{\mathcal{A}}$, as long as $T \geq T_{\mathcal{A}}$, we may choose $\mu = \ell/(4T)$, and set $\nu \in [\ell/(4T), \ell]$ accordingly per Proposition 8. By Lemma 23, we have that for $T \geq T_{\mathcal{A}}$,

$$\|\mathbf{C}(\mathbf{S})^T\|_\sigma \geq \rho(\mathbf{C}(\mathbf{S}))^T \geq (1 - C_{\mathcal{A}}/(4T))^T \geq \exp(-C_{\mathcal{A}}).$$

Set $c_{\mathcal{A}} = \exp(-C_{\mathcal{A}})$. Choose $\mathbf{w}^{(0)} = ((\mathbf{x}^{(-p)})^\top, \ldots, (\mathbf{x}^{(0)})^\top)^\top \in \mathbb{R}^{np}$ so that it is a right singular vector of $\mathbf{C}(\mathbf{S})^T$ corresponding to a singular value of magnitude at least $c_{\mathcal{A}}$. By scaling $\mathbf{w}^{(0)}$ appropriately, we may ensure that $\|\mathbf{x}^{(-p+1)}\|, \ldots, \|\mathbf{x}^{(0)}\| \leq D$, and that $\|\mathbf{w}^{(0)}\| \geq D$. It follows that

$$\sum_{j=0}^{p-1} \left(f(\mathbf{x}^{(T-j)}) - f(\mathbf{x}^*) \right) = \frac{1}{2\nu}\|(I_p \otimes \mathbf{S})\mathbf{w}^{(T)}\|^2 = \frac{\nu}{2}\|\mathbf{C}(\mathbf{S})^t \mathbf{w}^{(0)}\|^2 \geq \frac{\nu D^2 c_{\mathcal{A}}}{2} \geq \frac{\ell D^2 c_{\mathcal{A}}}{8T}.$$

By replacing T with $T + p - 1$ and decreasing $c_{\mathcal{A}}$, the conclusion of Proposition 30 follows. □

Remark 8. As in Theorem 7, the lower bound in Proposition 30 involves an algorithm-dependent constant $c_{\mathcal{A}}$ due to the reliance on Proposition 8. We remark that the iterates $\mathbf{x}^{(t)}$ of gradient descent satisfy $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq O(\ell D^2/T)$ for any ℓ-smooth convex function f, so Proposition 8 is tight up to the algorithm-dependent constant $c_{\mathcal{A}}$. Nesterov's AGD improves the rate of gradient descent to $O(\ell D^2/T^2)$, but is non-stationary (i.e., requires a changing step size). The polynomials in Proposition 9 (i.e., (9)) showing the necessity of an algorithm-dependent constant in Proposition 8 correspond under the reduction outlined in the proof of Proposition 30 to running Nesterov's AGD with a fixed learning rate. We do not know if such an algorithm (for an appropriate choice of the arbitrary but fixed learning rate) can lead to an arbitrarily large constant factor speedup over the rate $O(\ell D^2/T)$ of gradient descent. We believe this is an interesting direction for future work.