

# Min-Max Optimization Is Strictly Easier Than Variational Inequalities

Henry Shugart  
UPenn  
hshugart@upenn.edu

Jason M. Altschuler  
UPenn  
alts@upenn.edu

## Abstract

Classically, a mainstream approach for solving a convex-concave min-max problem is to instead solve the variational inequality problem arising from its first-order optimality conditions. Is it possible to solve min-max problems faster by bypassing this reduction? This paper initiates this investigation. We show that the answer is yes in the textbook setting of unconstrained quadratic objectives: the optimal convergence rate for first-order algorithms is strictly better for min-max problems than for the corresponding variational inequalities. The key reason that min-max algorithms can be faster is that they can exploit the asymmetry of the min and max variables—a property that is lost in the reduction to variational inequalities. Central to our analyses are sharp characterizations of optimal convergence rates in terms of extremal polynomials which we compute using Green’s functions and conformal mappings.

## 1 Introduction

This paper shows a fundamental gap between two well-studied classes of problems. The first is *min-max problems* with convex-concave objectives  $f$ : find a saddle-point solution  $z^* = (x^*, y^*)$  to

$$\min_x \max_y f(x, y). \quad (1.1)$$

The second is *variational inequality (VI) problems* with monotone operators  $F$ : find  $z^*$  satisfying

$$\langle F(z^*), z - z^* \rangle \geq 0, \quad \forall z. \quad (1.2)$$

Classically, these two problems are intimately connected because the former problem (1.1) can be cast as an instance of the latter problem (1.2) by considering first-order optimality conditions, concatenating the variables  $z = (x, y)$ , and defining  $F(z) = (\nabla_x f(z), -\nabla_y f(z))$  which is guaranteed to be a monotone operator whenever  $f$  is convex-concave [24].

Today, this classical connection is central to much of modern algorithm design for min-max problems: simply appeal to standard VI algorithms. This reduction is popular for good reasons: it enables leveraging powerful existing algorithms, it is typically quite effective in both theory and practice, and it is flexible to different problem settings. See for example the textbooks [24, 25].

The ubiquity of this reduction necessitates a fundamental (and remarkably unstudied) question: is it possible to solve min-max problems faster by bypassing this reduction? In other words, does solving the more general problem (1.2) inherently lead to worse algorithmic guarantees than solving the more specific problem (1.1)?

## 1.1 Contribution

This paper initiates this investigation. We show that the answer is yes in the classical setting of unconstrained quadratic objectives. This uncovers a fundamental gap between the algorithmic complexity of convex-concave min-max problems (1.1) and the corresponding VI problems (1.2).

Specifically, we prove that the optimal convergence rate obtained by first-order algorithms is strictly better for the former than for the latter. We characterize this gap for unconstrained, smooth, and (possibly strongly) convex-concave quadratic  $f$  and their corresponding monotone operators  $F$ . In these settings, we can express the optimal convergence rate in terms of an extremal polynomial problem of the form  $\min_p \max_{\lambda \in S} |p(\lambda)|$  where  $p$  ranges over polynomials whose degree is bounded in terms of the number of iterations that the algorithm is run, and  $S$  is a “spectral range” (i.e., the set of all possible eigenvalues for an associated linear operator). Importantly,  $S$  is an interval for the min-max problem but is a half-disc in  $\mathbb{C}$  for the VI problem. In particular, modulo rotation (which is irrelevant for the extremal polynomial problem), the spectral range  $S$  in the min-max problem is a strict subset of the spectral range  $S$  in the VI problem. This makes the resulting value  $\min_p \max_{\lambda \in S} |p(\lambda)|$  smaller—and therefore the convergence rate faster—for min-max problems. This gap is precisely quantified by a certain measure of the relative size of the spectral ranges  $S$ , namely via the ratio of (certain quantities of) the Green’s functions for the sets  $S$ . Combining these ideas, we establish that the optimal convergence rate is faster for min-max problems than VI problems by a factor of  $3\sqrt{3}/4 \approx 1.3$  for the strongly-convex-strongly-concave setting and  $3\sqrt{3}/2 \approx 2.6$  for the convex-concave setting. Note that in order to prove a separation, we establish a lower bound on the convergence of symmetric algorithms which is slower (by the aforementioned factors) than an upper bound we establish for the convergence of asymmetric algorithms. All our upper and lower bounds are order-optimal. See [Tables 1 and 2](#) for a summary.

This modest but fundamental gap uncovers a missed algorithmic opportunity for solving min-max problems. In particular, our result shows that in order to obtain optimal convergence rates, one must directly design algorithms for min-max problems rather than rely on the classical reduction to VI. This is true even if the min-max problem has identical<sup>1</sup> structural assumptions in  $x$  and  $y$ . A key distinction in such “direct” algorithms is that they exploit the intrinsic *asymmetry* of the  $x$  and  $y$  variables arising in min-max optimization (as opposed to the VI approach which concatenates the variables  $z = (x, y)$  at the outset and then treats them symmetrically). This asymmetry is a key aspect of a few min-max algorithms, such as gradient-descent-ascent with slingshot stepsizes [27] or alternating stepsizes [10, 12, 33]. Indeed, a primary motivation of this paper is that the convergence rate of slingshot stepsizes for quadratic min-max problems was better than all existing algorithms for the corresponding VI problems, including even algorithms that use momentum, extragredients, optimism, etc [27]. The results of this paper show that this gap is fundamental: no symmetric algorithm can converge as fast.

**Outlook.** This paper focuses on demonstrating this phenomenon in its most foundational form. The algorithmic opportunity we uncover opens the door to several new directions for future work, such as showcasing larger gaps in more general settings (e.g., general convex-concave  $f$  that are not necessarily quadratic), exploiting these gaps algorithmically (e.g., as done with our slingshot stepsizes in [27] for the quadratic setting), and investigating how this gap changes for different algorithm classes (e.g., algorithms that use higher-order information). We believe that this new

---

<sup>1</sup>An orthogonal line of work has developed fast asymmetric algorithms for min-max settings in which the optimization problems for  $x$  and  $y$  have asymmetric structural assumptions, such as differing smoothness or strong convexity parameters [3, 6, 9, 11, 31]. The thesis of this paper is complementary: even if the structural assumptions are symmetric in  $x$  and  $y$ , asymmetric algorithms enable faster convergence.

line of inquiry will lead to a finer-grained complexity of these fundamental problems as well as lead to better algorithms that exploit the missed algorithmic opportunity we uncover.

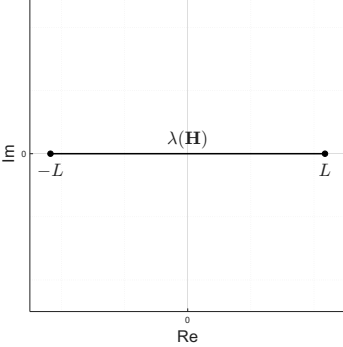
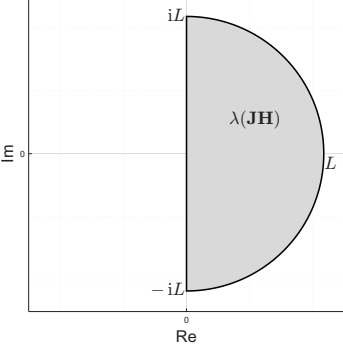
|                  | Convex-Concave Min-Max                                                            | Monotone VI                                                                        |
|------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| Convergence rate | $\frac{\ \nabla f(z_T)\ }{\ z_0 - z^*\ } \leq \frac{L}{T}$                        | $\frac{\ \nabla f(z_T)\ }{\ z_0 - z^*\ } \geq \frac{3\sqrt{3}}{2} \frac{L}{T}$     |
| Spectral range   |  |  |
| Green's function | $\frac{\partial}{\partial \mathbf{n}} g(0) = 1$                                   | $\frac{1}{2} \frac{\partial}{\partial \mathbf{n}} g(0) = \frac{2}{3\sqrt{3}}$      |

TABLE 1: **Top:** We establish a fundamental gap between the fastest possible convergence rate for convex-concave quadratic min-max optimization (**left**) and the corresponding monotone variational inequalities (**right**). Here  $T$  denotes the number of iterations and  $L$  denotes the smoothness. For simplicity we omit lower-order terms  $o(1/T)$ ; see [Theorems 3.2](#) and [4.4](#) for full details. **Middle:** the underlying geometric reason for this algorithmic gap is that the relevant spectral shape is a strictly smaller subset of  $\mathbb{C}$  for min-max problems than for VI problems. **Bottom:** This gap is precisely quantified by (the derivative of) Green's function for these spectral shapes. The additional factor of  $1/2$  appears because the spectral range has positive Lebesgue measure (see [Lemma 3.3](#)).

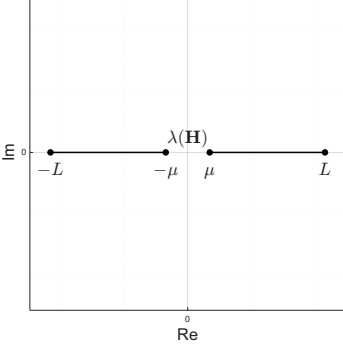
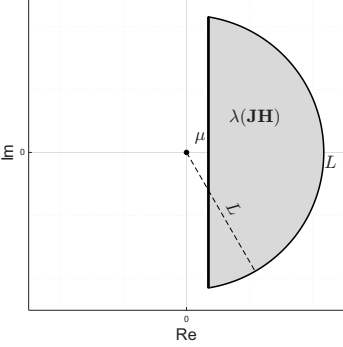
|                  | SCSC Min-Max                                                                        | Strongly-Monotone VI                                                                              |
|------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| Convergence rate | $\frac{\ z_T - z^*\ }{\ z_0 - z^*\ } \leq \exp\left(-\frac{T}{\kappa}\right)$       | $\frac{\ z_T - z^*\ }{\ z_0 - z^*\ } \geq \exp\left(-\frac{4}{3\sqrt{3}} \frac{T}{\kappa}\right)$ |
| Spectral range   |  |               |
| Green's function | $g(0) \approx \frac{1}{\kappa}$                                                     | $g(0) \approx \frac{4}{3\sqrt{3}} \frac{1}{\kappa}$                                               |

TABLE 2: Analog of [Table 1](#) for **strongly-convex-strongly-concave** settings. Here  $\mu$  denotes the strong convexity,  $L$  denotes the smoothness, and their ratio  $\kappa = L/\mu$  denotes the condition number. For simplicity we omit lower-order terms  $o_{T,\kappa}(1)$ ; see [Theorems 3.1](#) and [4.2](#) for full details.

## 2 Preliminaries

### 2.1 Problem setup

**Quadratic min-max problems.** We focus on unconstrained min-max problems of the form

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} f(x, y) \text{ where } f(x, y) = \frac{1}{2} \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix}^\top \underbrace{\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & -\mathbf{C} \end{bmatrix}}_{\mathbf{H}} \begin{bmatrix} x - x^* \\ y - y^* \end{bmatrix}. \quad (2.1)$$

For convex-concave  $f$ , stationary points coincide with saddle points; thus all the solutions  $z = (x, y)$  of (2.1) satisfy  $z - z^* \in \text{Null}(\mathbf{H})$ , where  $z^* = (x^*, y^*)$ . Indeed this is the criteria for  $\nabla f(z) = \mathbf{H}(z - z^*)$  to vanish. For simplicity we write problems in the form (2.1) which is homogeneous around a stationary point  $z^*$ . This is without loss of generality since, by translating, this captures quadratic objectives  $\frac{1}{2}z^\top \mathbf{H}z + l^\top z$  with arbitrary linear terms  $l$ , provided they admit at least one stationary point  $\nabla f(z) = 0$ . Throughout  $\mathbf{A}$  and  $\mathbf{C}$  are assumed symmetric without loss of generality, since otherwise one can replace them with their symmetrizations  $(\mathbf{A} + \mathbf{A}^\top)/2$  and  $(\mathbf{C} + \mathbf{C}^\top)/2$ .

**Reduction to a variational inequality.** The mainstream approach for solving min-max problems is to rewrite the first-order optimality conditions as a variational inequality. The variational inequality associated with (2.1) is:

$$\text{Find } z \in \mathbb{R}^{d_z} \text{ such that } \langle F(z), z' - z \rangle \geq 0, \forall z' \in \mathbb{R}^{d_z}, \text{ where } F(z) = \underbrace{\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^\top & \mathbf{C} \end{bmatrix}}_{\mathbf{JH}} (z - z^*). \quad (2.2)$$

Above  $\mathbf{J} = \text{diag}(\mathbf{I}, -\mathbf{I})$  and  $F = (\nabla_x f, -\nabla_y f)$ . We write (2.2) since this is the standard general way to define variational inequalities, although of course in the unconstrained setting, (2.2) simplifies to finding  $z \in \mathbb{R}^{d_z}$  such that  $F(z) = 0$ . Such a solution corresponds to a stationary point of  $f$  and thus a solution of the min-max problem (2.1).

**Convexity-concavity and monotonicity.** Convergence rates for solving such problems depend on the structure of the objective  $f$  and corresponding operator  $F$ . We focus on the classic setting of (strongly) convex-concave  $f$ , which corresponds to (strongly) monotone operators  $F$ . Below we recall these definitions and the correspondences.

**Definition 2.1** ((Strongly) convex-concave functions). *For  $\mu \geq 0$ , a function  $f(x, y)$  is  $\mu$ -strongly-convex-strongly-concave ( $\mu$ -SCSC for short) if  $f(\cdot, y)$  is  $\mu$ -strongly-convex for every  $y$ , and  $f(x, \cdot)$  is  $\mu$ -strongly-concave for every  $x$ .<sup>2</sup> If  $f$  satisfies this for  $\mu = 0$ ,  $f$  is convex-concave.*

For quadratic min-max problems (2.1), the condition that  $f$  be  $\mu$ -SCSC is equivalent to the condition  $\mathbf{A}, \mathbf{C} \geq \mu \mathbf{I}$ . For variational inequalities, the analogous property is (strong) monotonicity.

**Definition 2.2** ((Strongly) monotone operators). *For  $\mu \geq 0$ , an operator  $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is  $\mu$ -strongly-monotone if  $\langle F(z) - F(z'), z - z' \rangle \geq \mu \|z - z'\|^2$  for all  $z, z'$ . If  $F$  satisfies this for  $\mu = 0$ ,  $F$  is monotone.*

---

<sup>2</sup>Recall that a function  $g$  is said to be  $\mu$ -strongly-convex if  $g(x) - \frac{\mu}{2}\|x\|^2$  is convex. For twice-differentiable  $g$ , this condition is equivalent to  $\nabla^2 g \geq \mu \mathbf{I}$ . A function  $g$  is said to be  $\mu$ -strongly-concave if  $-g$  is  $\mu$ -strongly-convex.

In particular, (strong) convexity-concavity of  $f$  implies (strong) monotonicity of the saddle operator  $F(z) = (\nabla_x f(z), -\nabla_y f(z))$  in the associated variational inequality.

**Lemma 2.3** (Convexity-concavity implies monotonicity [23, Theorem 1]). *Let  $\mu \geq 0$ . If  $f(x, y)$  is  $\mu$ -strongly-convex-strongly-concave, then  $F = (\nabla_x f, -\nabla_y f)$  is  $\mu$ -strongly-monotone.*

*Proof.* The cited textbook proves this for  $\mu = 0$ . The proof extends to  $\mu > 0$  in a straightforward way:  $g(x, y) = f(x, y) - \frac{\mu}{2}\|x\|^2 + \frac{\mu}{2}\|y\|^2$  is convex-concave, hence  $G = (\nabla_x g, -\nabla_y g)$  is monotone (this is the case  $\mu = 0$ ), hence  $\langle G(z) - G(z'), z - z' \rangle \geq 0$  for any  $z, z'$ . Plugging in  $G = F - \mu z$  and re-arranging establishes  $\langle F(z) - F(z'), z - z' \rangle \geq \mu\|z - z'\|^2$ , hence  $F$  is  $\mu$ -strongly monotone.  $\square$

Aside from monotonicity, throughout we make the standard assumption that  $f$  is  $L$ -smooth, i.e.,  $\nabla f$  is  $L$ -Lipschitz. This is equivalent to the saddle-operator  $F = (\nabla_x f, -\nabla_y f)$  being  $L$ -Lipschitz. In the quadratic settings (2.1) and (2.2), this equivalently simplifies to the assumption  $\|\mathbf{H}\| \leq L$ .

## 2.2 Spectral range

The spectra of  $\mathbf{H}$  and  $\mathbf{JH}$  play a central role in the convergence rate of first-order algorithms. We start by defining shorthand  $\mathcal{H}_\mu$  and  $\mathcal{J}_\mu$  for the sets of possible matrices  $\mathbf{H}$  and  $\mathbf{JH}$ , respectively, associated with quadratic min-max problems (2.1) that are  $L$ -smooth and  $\mu$ -strongly-convex-concave:

$$\mathcal{H}_\mu = \left\{ \mathbf{H} : \mathbf{H} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & -\mathbf{C} \end{bmatrix}, \|\mathbf{H}\| \leq L, \mathbf{A} \geq \mu \mathbf{I}, \mathbf{C} \geq \mu \mathbf{I} \right\} \quad \text{and} \quad \mathcal{J}_\mu = \left\{ \mathbf{JH} : \mathbf{H} \in \mathcal{H}_\mu \right\}. \quad (2.3)$$

To avoid notational overhead, we suppress the dependence on  $L$  when writing  $\mathcal{H}_\mu$  and  $\mathcal{J}_\mu$ . We emphasize the dependence on  $\mu$  since separating the cases  $\mu > 0$  and  $\mu = 0$  lets us develop in parallel the strongly and non-strongly convex-concave settings.

**Definition 2.4** (Spectral range). *The spectral range of a set of matrices  $\mathcal{M}$  is*

$$\sigma(\mathcal{M}) = \bigcup_{\mathbf{M} \in \mathcal{M}} \sigma(\mathbf{M}).$$

where  $\sigma(\mathbf{M})$  denotes the spectrum of a matrix  $\mathbf{M}$ .

The spectral ranges of  $\mathcal{H}_\mu$  and  $\mathcal{J}_\mu$  are explicit in terms of  $\mu$  and  $L$ . For  $\mathcal{J}_\mu$  a proof can be found, e.g., in [1]. For  $\mathcal{H}_\mu$  we are not aware of a reference and therefore provide a short proof here.

**Lemma 2.5** (Spectral range of  $\mathcal{J}_\mu$ ). *For any  $0 \leq \mu \leq L < \infty$ , the spectral range of  $\mathcal{J}_\mu$  is*

$$\sigma(\mathcal{J}_\mu) = \{z : |z| \leq L, \operatorname{Re}(z) \geq \mu\}.$$

**Lemma 2.6** (Spectral range of  $\mathcal{H}_\mu$ ). *For any  $0 \leq \mu \leq L < \infty$ , the spectral range of  $\mathcal{H}_\mu$  is*

$$\sigma(\mathcal{H}_\mu) = [-L, -\mu] \cup [\mu, L].$$

*Proof.* The direction “ $\supseteq$ ” is clear by considering  $\mathbf{A} = \mathbf{C} = |\lambda|$  and  $\mathbf{B} = 0$  for any  $\lambda \in [\mu, L]$ . We prove the other direction “ $\subseteq$ ” by combining three observations. First,  $\sigma(\mathcal{H}_\mu)$  is real since the matrices  $\mathbf{H} \in \mathcal{H}_\mu$  are symmetric and thus have real eigenvalues. Second,  $\sigma(\mathcal{H}_\mu) \subseteq [-L, L]$  since the spectral radius (i.e., maximum magnitude eigenvalue) of a matrix is upper bounded by the spectral norm, and  $\|\mathbf{H}\| \leq L$  for all  $\mathbf{H} \in \mathcal{H}_\mu$ . Third,  $\sigma(\mathcal{H}_\mu) \subseteq (-\infty, -\mu] \cup [\mu, \infty)$ . To show this, it suffices to argue that  $\mathbf{H} - r\mathbf{I}$  is invertible for any  $\mathbf{H} \in \mathcal{H}_\mu$  and  $r \in (-\mu, \mu)$ . Observe that the diagonal blocks  $\mathbf{A} - r\mathbf{I} \geq (\mu - r)\mathbf{I} > 0$  and  $-(\mathbf{C} + r\mathbf{I}) \leq -(\mu + r)\mathbf{I} < 0$  are both invertible. Thus the Schur complement  $(\mathbf{A} - r\mathbf{I}) + \mathbf{B}(\mathbf{C} + r\mathbf{I})^{-1}\mathbf{B}^\top > 0$  is also invertible, and hence so is the full matrix  $\mathbf{H} - r\mathbf{I}$ , as desired.  $\square$

See [Tables 1 and 2](#) for an illustration of these spectral ranges in the cases  $\mu > 0$  and  $\mu = 0$ , respectively. Geometrically, when  $\mu > 0$ , the spectral range  $\sigma(\mathcal{H}_\mu)$  is the union of two real intervals that are symmetric around 0, whereas  $\sigma(\mathcal{J}_\mu)$  is the intersection of the complex disc of radius  $L$  with the half-plane  $\{z \in \mathbb{C} : \operatorname{Re}(z) \geq \mu\}$ . Both sets simplify when  $\mu = 0$ : then  $\sigma(\mathcal{H}_\mu)$  becomes a single interval  $[-L, L]$  and  $\sigma(\mathcal{J}_\mu)$  becomes the half disc of radius  $L$  with positive real part.

## 2.3 First-order algorithms

**First-order algorithms.** We focus on the standard algorithmic model of first-order oracle access to  $f$ , i.e., black-box queries of the form  $f(x, y)$  and  $\nabla f(x, y)$ . We remark that our proposed algorithms use gradients  $\nabla f(x, y)$  but not function evaluations  $f(x, y)$ ; nevertheless we include function evaluations in the definition of the oracle since this is the standard setup. In [§3](#) we show that the inclusion of this information does not affect the optimal convergence rates.

**Krylov-subspace algorithms.** We focus on the standard setting of Krylov-subspace algorithms, i.e., algorithms that produce their next iterate within the span of the observed gradients (formal definition below). This is a reasonable assumption since deviating from the span of observed gradients amounts to making an uninformed guess. It is classically known from other areas of optimization (see e.g., [\[16, 17, 18\]](#)) that the Krylov-subspace assumption simplifies arguments, isolates the key conceptual ideas, and can be relaxed at the expense of more technical arguments.

**Adaptive algorithms.** All of our results hold regardless of adaptivity, i.e., whether the linear-span coefficients for producing iterates depend on observed information<sup>3</sup>. In fact our algorithmic upper bounds are achieved without adaptivity. For clarity of exposition, we first prove our lower bounds for non-adaptive algorithms in [§3](#) since these arguments are simpler; then in [§5](#) we explain the more technical argument for establishing the same results for adaptive algorithms.

**Symmetric vs asymmetric algorithms.** This paper highlights the importance of a complementary axis for distinguishing min-max algorithms: whether updates are symmetric in the minimization variable  $x$  and maximization variable  $y$ . In words, symmetric algorithms treat  $x, y$  identically in all updates. Such algorithms are the standard for variational inequality problems since, even from the outset of the problem formulation, variational inequalities do not distinguish between the blocks of  $z = (x, y)$ ; c.f., [\(2.1\)](#) versus [\(2.2\)](#). In contrast, asymmetric algorithms can update  $x$  and  $y$  differently, which enables exploiting the inherent asymmetry in the definition of min-max problems. Symmetric algorithms can be defined for both min-max problems [\(2.1\)](#) and variational inequality problems [\(2.2\)](#), whereas asymmetric algorithms are only possible to implement for the former.

**Definition 2.7** (Symmetric algorithms). *A symmetric Krylov-subspace algorithm is an iterative algorithm that produces a sequence  $\{z_t\}$  satisfying, for all  $t \geq 1$ ,*

$$z_t \in z_0 + \operatorname{span} \{F(z_0), F(z_1), \dots, F(z_{t-1})\}.$$

*In other words, there exist coefficients  $c_{t,k}$  such that*

$$z_t = z_0 + \sum_{k=0}^{t-1} c_{t,k} F(z_k). \tag{2.4}$$

---

<sup>3</sup>Formally, an algorithm is said to be non-adaptive if the coefficients  $c_{t,k}$  in [\(2.4\)](#) (or analogously  $a_{t,k}, b_{t,k}$  in [\(2.5\)](#) for asymmetric algorithms) depend on  $t, k, \mu$ , and  $L$ , but not on any information from prior iterates. Note that throughout we assume for simplicity that  $\mu$  and  $L$  are known.

**Definition 2.8** (Asymmetric algorithms). *An asymmetric Krylov-subspace algorithm is an iterative algorithm that produces a sequence  $\{z_t = (x_t, y_t)\}$  satisfying, for all  $t \geq 1$ ,*

$$\begin{aligned} x_t &\in x_0 + \text{span}\{\nabla_x f(x_0, y_0), \dots, \nabla_x f(x_{t-1}, y_{t-1})\}, \\ y_t &\in y_0 + \text{span}\{\nabla_y f(x_0, y_0), \dots, \nabla_y f(x_{t-1}, y_{t-1})\}. \end{aligned}$$

*In other words, there exist coefficients  $a_{t,k}$  and  $b_{t,k}$  such that*

$$x_t = x_0 + \sum_{k=0}^{t-1} a_{t,k} \nabla_x f(x_k, y_k) \quad \text{and} \quad y_t = y_0 + \sum_{k=0}^{t-1} b_{t,k} \nabla_y f(x_k, y_k). \quad (2.5)$$

The main result of this paper is that asymmetric algorithms lead to faster convergence rates than are possible using symmetric algorithms. Our starting point is the observation that the iterates of symmetric and asymmetric algorithms can be associated with polynomials in the matrices  $\mathbf{JH}$  and  $\mathbf{H}$ , respectively. As we detail formally below, this correspondence is 1-to-1 for the former; whereas for the latter, the class of asymmetric algorithms include these matrix polynomials as an important special case (this inclusion is sufficient for us to develop the claimed faster algorithms). Below and throughout, denote the linear space of bounded-degree polynomials with constant coefficient 1 by

$$\mathcal{P}_t = \{p : \deg(p) \leq t, p(0) = 1\}.$$

**Lemma 2.9** (Symmetric algorithms are matrix polynomials of  $\mathbf{JH}$ ). *Consider the min-max problem (2.1) or the variational inequality problem (2.2). For any symmetric Krylov-subspace algorithm, there exists a sequence of polynomials  $p_t \in \mathcal{P}_t$  such that for each  $t$ ,*

$$z_t - z^* = p_t(\mathbf{JH})(z_0 - z^*).$$

*Proof.* We prove by induction on  $t$ . The base case  $t = 0$  is trivial. Supposing true for all  $k < t$ , then  $z_t - z^* = (z_0 - z^*) + \sum_{k=0}^{t-1} c_{t,k} (F(z_k) - F(z^*)) = p_t(\mathbf{JH})(z_0 - z^*)$  where  $p_t(\lambda) = 1 + \sum_{k=0}^{t-1} c_{t,k} \lambda p_k(\lambda)$ . Here the first step uses Definition 2.7 and  $F(z^*) = 0$ ; the second step uses  $F(z) = \mathbf{JH}(z - z^*)$  and the induction hypothesis  $z_k - z^* = p_k(\mathbf{JH})(z_0 - z^*)$ . Since  $p_t \in \mathcal{P}_t$ , the claim is proved.  $\square$

**Lemma 2.10** (Asymmetric algorithms include matrix polynomials of  $\mathbf{H}$ ). *Consider the quadratic min-max problem (2.1). For any polynomial  $p_t \in \mathcal{P}_t$ , there exists an asymmetric Krylov-subspace algorithm whose  $t$ -th iterate satisfies*

$$z_t - z^* = p_t(\mathbf{H})(z_0 - z^*).$$

*Proof.* Since  $p_t \in \mathcal{P}_t$ , it can be written in factorized form  $p_t(\lambda) = \prod_{k=0}^{t-1} (1 - \lambda/r_k)$  where  $\{r_k\}$  denote its roots. Consider the asymmetric Krylov-subspace algorithm

$$x_{k+1} = x_k - \frac{1}{r_k} \nabla_x f(x_k, y_k) \quad \text{and} \quad y_{k+1} = y_k - \frac{1}{r_k} \nabla_y f(x_k, y_k) \quad (2.6)$$

for all  $k < t$ . By concatenating variables  $z = (x, y)$ , this can be written succinctly as

$$z_{k+1} - z^* = z_k - z^* - \frac{1}{r_k} \nabla f(z_k) = \left( \mathbf{I} - \frac{1}{r_k} \mathbf{H} \right) (z_k - z^*).$$

Iterating  $t$  times yields the desired identity  $z_t - z^* = \prod_{k=0}^{t-1} (\mathbf{I} - \frac{1}{r_k} \mathbf{H})(z_0 - z^*) = p_t(\mathbf{H})(z_0 - z^*)$ .  $\square$



**Remark 2.11** (Gradient-descent-ascent with negative/complex stepsizes). *The asymmetric algorithm we propose in (2.6) is gradient-descent-ascent, but with unconventional stepsizes  $\pm 1/r_k$  which are negative or complex. Such stepsizes were first explored in our previous work [27]. By showing that asymmetric algorithms can converge strictly faster than symmetric algorithms, the present paper shows that such stepsizes can lead to faster rates than arbitrary first-order symmetric algorithms—including even algorithms that use momentum, extragredients, optimism, etc.*

*Note also that the algorithmic construction of the polynomials in the proof of Lemma 2.10 is not unique. Other algorithms could be used, but the key is that any such algorithm must be asymmetric.*

## 2.4 Complex analysis and approximation theory preliminaries

Central to our analysis are sharp characterizations of the optimal convergence rates in terms of approximation-theoretic quantities which we compute using Green’s functions and conformal mappings. We briefly recall relevant background here for the convenience of the reader. For further details on these topics see for example the excellent textbooks [21, 28, 29].

Below and throughout, we use the following notational shorthands: let  $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$  denote the one-point compactification of  $\mathbb{C}$ , let  $D = \{z : |z| \leq 1\}$  denote the unit disc, and let  $\Omega = \{z \in D : \operatorname{Re}(z) \geq 0\}$  denote the unit half disc with positive real part. Note that  $\sigma(\mathcal{J}_0) = \Omega$  is the spectral range for linear monotone operators that are Lipschitz with parameter  $L = 1$  (see Lemma 2.5). Finally, we write  $\|p\|_S = \sup_{z \in S} |p(z)|$  to denote the supremum norm of a function  $p$  on a set  $S$ .

**Conformal mappings.** Recall that a conformal map is a bijective holomorphic map between open subsets of  $\hat{\mathbb{C}}$ . The Riemann mapping theorem ensures the existence of conformal maps between any non-empty, simply connected, open sets. We make particular use of the following explicit conformal map of the exterior of the half disc to the exterior of the disc; see Fig. 1 for a visualization.

**Lemma 2.12** (Conformal mapping of  $\hat{\mathbb{C}} \setminus \Omega$  to  $\hat{\mathbb{C}} \setminus D$  [19]). *The function*

$$\Phi_{\Omega}(\lambda) = \frac{\left(1 - \left(\frac{\lambda-i}{\lambda+i}\right)^{\frac{2}{3}}\right) - \sqrt{3}i \left(1 + \left(\frac{\lambda-i}{\lambda+i}\right)^{\frac{2}{3}}\right)}{2 \left(\left(\frac{\lambda-i}{\lambda+i}\right)^{\frac{2}{3}} - 1\right)}$$

*conformally maps the exterior  $\hat{\mathbb{C}} \setminus \Omega$  of the half disc to the exterior  $\hat{\mathbb{C}} \setminus D$  of the unit disc.*

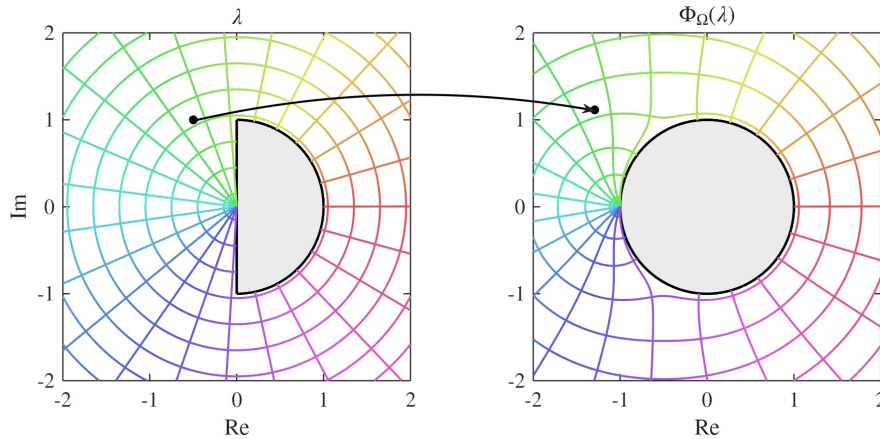


FIGURE 1: The conformal mapping  $\Phi_{\Omega}$  in Lemma 2.12 from the exterior  $\hat{\mathbb{C}} \setminus \Omega$  of the half disc to the exterior  $\hat{\mathbb{C}} \setminus D$  of the disc. In this plot, a point  $\lambda \in \hat{\mathbb{C}} \setminus \Omega$  (left) is mapped to the point  $\Phi(\lambda) \in \hat{\mathbb{C}} \setminus D$  (right) of the same color.



**Green's function.** Green's functions arise throughout PDE, complex analysis, potential theory, and more. In this paper we make use of their connections to approximation theory. Below and throughout, we consider Green's function with pole at  $\infty$  (hence we drop this qualifier as there is no ambiguity) and for sets  $S$  that are connected (hence we can use the following formula in terms of conformal mappings).

**Definition 2.13** (Green's function with pole at  $\infty$ ). *Let  $S \subset \mathbb{C}$  be a non-empty compact set such that  $\hat{\mathbb{C}} \setminus S$  is simply connected. The Green's function of  $S$  is  $g_S = \log |\Phi_S|$ , where  $\Phi_S$  is the conformal mapping from  $\hat{\mathbb{C}} \setminus S$  to  $\hat{\mathbb{C}} \setminus D$  with normalization  $\Phi_S(\infty) = \infty$  and  $\Phi'_S(\infty) > 0$ .*

Fig. 2 illustrates Green's function  $g_\Omega$  for the unit half disc  $\Omega$ .

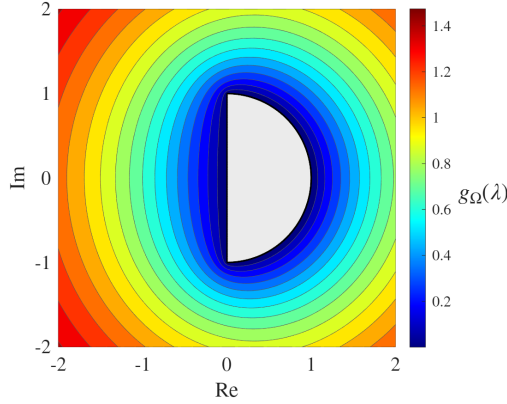


FIGURE 2: Contour plot of Green's function  $g_\Omega$  for the unit half disc  $\Omega$ .

Green's function arises as a fundamental quantity in our analysis as it captures the maximal growth rate of a polynomial outside of a set  $S$ , when constrained in sup-norm on  $S$ . This classical result is due to Bernstein and Walsh [2, 30]; see for example Lemma 3.6 of [26] for a short modern exposition, or [5, Theorem 1] for classical applications to the analysis of matrix iterations.

**Lemma 2.14** (Bernstein–Walsh Theorem). *Let  $S \subset \mathbb{C}$  be a non-empty compact set such that  $\hat{\mathbb{C}} \setminus S$  is simply connected. For any  $\lambda \in \mathbb{C} \setminus S$  and any polynomial  $p$  of degree at most  $T$ ,*

$$|p(\lambda)| \leq e^{Tg_S(\lambda)} \|p\|_S.$$

Relatedly, Green's function also captures the maximal growth rate of the *derivatives* of polynomials. Such growth bounds are often called Bernstein-type inequalities (recalled in §3.2 when we make use of them) and depend on Green's function through  $\frac{\partial}{\partial \mathbf{n}} g_S(\lambda)$  which denotes, for  $\lambda \in \partial S$ , the derivative of  $g_S$  in the direction normal to  $S$  (with orientation pointing away from  $S$ ). For ease of recall, we state here that the explicit value of this quantity for  $S = \Omega$  and  $\lambda = 0$ .

**Corollary 2.15.**  $\frac{\partial}{\partial \mathbf{n}} g_\Omega(0) = \frac{4}{3\sqrt{3}}.$

*Proof.* By Lemma 2.12 and Definition 2.13,  $\frac{\partial}{\partial \mathbf{n}} g_\Omega(0) = \frac{\partial}{\partial \mathbf{n}} \log |\Phi_\Omega(\lambda)| = \frac{|\Phi'_\Omega(0)|}{|\Phi_\Omega(0)|} = \frac{4/(3\sqrt{3})}{1}.$   $\square$

### 3 Lower bounds for symmetric algorithms

In this section we prove lower bounds on the computational complexity of *symmetric* first-order algorithms for solving (strongly) convex-concave min-max problems. This immediately implies

analogous lower bounds for solving variational inequalities for (strongly) monotone operators by the standard reduction in §2.1. For simplicity of exposition, here we restrict to non-adaptive algorithms; in §5 we show how the results extend to adaptive algorithms.

### 3.1 Strongly-convex-strongly-concave problems

Our first result is for the strongly-convex-strongly-concave (SCSC) setting. Recall that  $\kappa$ -conditioned means  $\mu$ -SCSC and  $L$ -smooth for condition number  $\kappa = L/\mu$ . This result improves on the previous state of the art lower bound of  $\exp(-2T/\kappa)$  (up to lower order terms in  $T, \kappa$ ) given in Ibrahim et al. [8, Proposition 2]. Our improved bound allows us to show a computational gap between symmetric and asymmetric algorithms for the first time—since the improvement of this lower bound from  $\exp(-2 \cdot T/\kappa)$  to roughly  $\exp(-(4\sqrt{3}/3) \cdot T/\kappa) \approx \exp(-0.77 \cdot T/\kappa)$  crosses the threshold of the  $\exp(-1 \cdot T/\kappa)$  upper bound established for asymmetric algorithms later in Theorem 4.2.

**Theorem 3.1** (Lower bound for symmetric algorithms on SCSC problems). *For any non-adaptive symmetric Krylov-subspace algorithm and any number of iterations  $T$ , there exists a  $\kappa$ -conditioned quadratic min-max problem (2.1) with solution  $z^*$  such that the convergence rate is no faster than*

$$\frac{\|z_T - z^*\|}{\|z_0 - z^*\|} \geq \left| \Phi_\Omega \left( \frac{-1}{\kappa - 1} \right) \right|^{-T} = \exp \left( - \left( \frac{4}{3\sqrt{3}} + o_\kappa(1) \right) \frac{T}{\kappa} \right).$$

*Proof.* By Lemma 2.9, the  $T$ -th iterate  $z_T$  of the algorithm can be expressed as

$$z_T - z^* = p_T(\mathbf{JH})(z_0 - z^*), \quad (3.1)$$

where  $p_T \in \mathcal{P}_T$ , i.e.,  $p_T$  is a polynomial of degree at most  $T$  satisfying the normalization constraint  $p_T(0) = 1$ . In the worst-case over  $\kappa$ -conditioned  $\mathbf{H}$  and solutions  $z^*$ , the convergence rate is no faster than

$$\max_{\mathbf{H}, z^*} \frac{\|z_T - z^*\|}{\|z_0 - z^*\|} = \max_{\mathbf{H}, z^*} \frac{\|p_T(\mathbf{JH})(z_0 - z^*)\|}{\|z_0 - z^*\|} = \max_{\mathbf{H}} \|p_T(\mathbf{JH})\| \geq \max_{\mathbf{H}} |\lambda_{\max}(p_T(\mathbf{JH}))| \geq \max_{\lambda \in \sigma(\mathcal{J}_\mu)} |p_T(\lambda)|.$$

Above, the first step is by definition of  $p_T$  in (3.1), the second step is by definition of the operator norm, the third step is because the operator norm of a matrix is bounded below by its spectral radius, and the final step is because the eigenvalues of  $p_T(\mathbf{JH})$  are simply given by  $p_T$  evaluated at the eigenvalues of  $\mathbf{JH}$  (see for example Theorem 1.1.6 of [7]).

We conclude that the fastest possible convergence rate for such algorithms is no faster than

$$\min_{p \in \mathcal{P}_T} \max_{\lambda \in \sigma(\mathcal{J}_\mu)} |p(\lambda)|. \quad (3.2)$$

Classical tools from approximation theory let us lower bound such extremal polynomial problems via an associated Green's function (Lemma 2.14). To invoke this, it is convenient to first replace  $\sigma(\mathcal{J}_\mu)$  by a slightly smaller set with an explicit Green's function, namely the half-disc  $\Omega_\mu := \{\lambda : |\lambda - \mu| \leq L - \mu, \operatorname{Re}(\lambda) \geq \mu\}$  that has center  $\mu$  and radius  $L - \mu$ . Since  $\sigma(\mathcal{J}_\mu) \supset \Omega_\mu$  by Lemma 2.5, an application of Lemma 2.14 with  $S = \Omega_\mu$  and  $\lambda = 0$  gives

$$\min_{p \in \mathcal{P}_T} \max_{\lambda \in \sigma(\mathcal{J}_\mu)} |p(\lambda)| \geq \min_{p \in \mathcal{P}_T} \max_{\lambda \in \Omega_\mu} |p(\lambda)| \geq \exp(-Tg_{\Omega_\mu}(0)).$$

This quantity involving the Green's function is explicitly computed as:

$$\exp(-Tg_{\Omega_\mu}(0)) = |\Phi_{\Omega_\mu}(0)|^{-T} = \left| \Phi_\Omega \left( \frac{-1}{\kappa - 1} \right) \right|^{-T} = \exp \left( - \left( \frac{4}{3\sqrt{3}} + o_\kappa(1) \right) \frac{T}{\kappa} \right).$$

Above, the first step is by definition of the Green's function  $g_{\Omega_\mu} = \log |\Phi_{\Omega_\mu}|$  in terms of the conformal map  $\Phi_{\Omega_\mu}$  from the exterior of  $\Omega_\mu$  to the exterior of the unit disc (Definition 2.13), the second step is by recentering and rescaling  $\Omega_\mu$  to the standard half disc  $\Omega$ , and the final step is by plugging in the explicit conformal mapping  $\Phi_\Omega$  (Lemma 2.12) and using the asymptotic expansion  $|\Phi_\Omega(-\frac{1}{\kappa-1})|^{-1} = 1 - \frac{4}{3\sqrt{3}\kappa} + O(\frac{1}{\kappa^2}) = \exp(-\frac{4}{3\sqrt{3}\kappa} + O(\frac{1}{\kappa^2}))$  for  $\kappa \rightarrow \infty$ .  $\square$

### 3.2 Convex-concave problems

We now turn to the convex-concave setting, i.e.,  $\mu = 0$ . Our main result here is the following lower bound for symmetric algorithms. This sharpens the well-known  $O(1/T)$  lower bound [32] sufficiently to establish a separation from the faster rates of asymmetric algorithms shown later.

**Theorem 3.2** (Lower bound for symmetric algorithms on convex-concave problems). *For any non-adaptive symmetric Krylov-subspace algorithm and any number of iterations  $T$ , there exists an  $L$ -smooth quadratic min-max problem (2.1) with solution  $z^*$  such that the convergence rate is no faster than*

$$\|\nabla f(z_T)\| \geq \left( \frac{3\sqrt{3}}{2} + o_T(1) \right) \frac{L\|z_0 - z^*\|}{T}.$$

Note that convergence is measured via the gradient norm  $\|\nabla f(z_T)\|$ . Although distance to optimum  $\|z_T - z^*\|$  is a meaningful metric in the SCSC setting (c.f. Theorem 3.1), it is well-known that such convergence rates are impossible in the convex-concave setting due to pathologically flat objectives.

A core ingredient in our proof of Theorem 3.2 is a Bernstein-type inequality, i.e., an inequality bounding the derivative of a polynomial by the largest value that the polynomial takes on a given set. Recall that a Jordan curve is the image of an injective continuous map of a circle, and a Jordan arc is the image of an injective continuous map of a line segment; we will apply this for the Jordan curve being the boundary of the half-disc  $\partial\Omega$  (the boundary of the relevant spectral shape for convex-concave problems, see Lemma 2.5) and the Jordan arc being a small interval of the imaginary axis around 0 (the relevant prescribed root, see the proof of Theorem 3.2 below).

**Lemma 3.3** (Bernstein inequality for polynomials with prescribed zeros on general domains). *Let  $K \subset \mathbb{C}$  be a compact set bounded by a Jordan curve. Let  $\lambda_0$  be a point on the boundary of  $K$ , and assume the boundary of  $K$  is a twice continuously differentiable Jordan arc in a neighborhood of  $\lambda_0$ . Let  $p_T$  be a polynomial of degree at most  $T$ . Further assume  $\lambda_0$  is a root of  $p_T$ . Then*

$$|p'_T(\lambda_0)| \leq (1 + o_T(1)) \cdot \frac{T}{2} \cdot \frac{\partial}{\partial \mathbf{n}} g_K(\lambda_0) \cdot \|p_T\|_K.$$

Lemma 3.3 combines two strengthenings of Bernstein's classical inequality [2]: strengthened Bernstein inequalities for polynomials with prescribed zeros [20] (note that  $\lambda_0$  is a root of  $p_T$ ) and Bernstein inequalities on general domains [14, 15] (Bernstein's original inequality was only for the disc). Theorem 1.3 of [15] establishes Lemma 3.3 but with twice as large an upper bound and without the assumption that  $\lambda_0$  is a root. Lemma 3.3 follows by the same proof—simply replace the use of Bernstein's standard inequality with the strengthening in [20] which improves the bound by a factor of 2 when  $\lambda_0$  is a root.<sup>4</sup> This twofold improvement enables our theory to tightly bound the relevant extremal polynomial (3.3), described below, as can be verified numerically.

<sup>4</sup>Details: simply carry through the twofold improvement in the three proof steps. 1) Replace the standard Bernstein inequality with the twofold improvement of [20, Corollary 1] in [14, page 452]. 2) Carry through this tightened bound on page 455 to improve [14, Theorem 1] by a factor of 2. 3) Carry through this tightened bound in equation (3) on page 193 to improve [15, Theorem 1.3] by a factor of 2.

*Proof of Theorem 3.2.* We begin by reducing to an extremal polynomial problem, similarly to our analysis of the strongly-convex-strongly-concave setting (see the proof of Theorem 3.1). As done there, let  $p_T$  denote the polynomial corresponding to the algorithm when run for  $T$  iterations. Then the worst-case convergence rate is no faster than

$$\max_{\mathbf{H}, z^*} \frac{\|\nabla f(z_T)\|}{\|z_0 - z^*\|} = \max_{\mathbf{H}, z^*} \frac{\|\mathbf{JH}p_T(\mathbf{JH})(z_0 - z^*)\|}{\|z_0 - z^*\|} = \max_{\mathbf{H}} \|\mathbf{JH}p_T(\mathbf{JH})\| \geq \max_{\lambda \in \sigma(\mathcal{J}_0)} |\lambda p_T(\lambda)|.$$

Above, the first step is because  $\|\nabla f(z_T)\| = \|\nabla f(z_T) - \nabla f(z^*)\| = \|\mathbf{H}(z_T - z^*)\| = \|\mathbf{H}p_T(\mathbf{JH})(z_0 - z^*)\| = \|\mathbf{JH}p_T(\mathbf{JH})(z_0 - z^*)\|$ . Introducing the extra factor of  $\mathbf{J}$  here does not affect the norm but ensures that the final expression in the above display depends on the matrix  $\mathbf{H}$  only through  $\mathbf{JH}$ .

We conclude that the fastest possible convergence rate for such algorithms is no faster than

$$\min_{p \in \mathcal{P}_T} \max_{\lambda \in \sigma(\mathcal{J}_0)} |\lambda p(\lambda)| = \min_{q \in \mathcal{Q}_{T+1}} \max_{\lambda \in \sigma(\mathcal{J}_0)} |q(\lambda)| \quad (3.3)$$

where we define the shorthand  $\mathcal{Q}_{T+1} = \{q : \deg(q) \leq T+1, q(0) = 0, q'(0) = 1\}$ . Note that the normalization is different than in §3.1, including constraints on both the function value and first derivative. Since 0 is a root of  $q \in \mathcal{Q}_T$ , the Bernstein inequality Lemma 3.3 gives

$$|q'(0)| \leq (1 + o_T(1)) \cdot \frac{T}{2} \cdot \frac{\partial}{\partial \mathbf{n}} g_{\sigma(\mathcal{J}_0)}(0) \cdot \max_{\lambda \in \sigma(\mathcal{J}_0)} |q(\lambda)|.$$

We now compute the quantity  $\frac{\partial}{\partial \mathbf{n}} g_{\sigma(\mathcal{J}_0)}(0)$ . Recall from Lemma 2.5 that  $\sigma(\mathcal{J}_0)$  is the half disc of radius  $L$ . This quantity is explicit for the half disc  $\Omega$  of radius 1 by Corollary 2.15. To rescale appropriately, first observe that  $g_{\sigma(\mathcal{J}_0)}(\lambda) = \log |\Phi_{\sigma(\mathcal{J}_0)}(\lambda)| = \log |\Phi_{\Omega}(\frac{\lambda}{L})| = g_{\Omega}(\frac{\lambda}{L})$  by using the definition of Green's function in terms of the conformal mapping and rescaling the conformal mapping. Hence by the chain rule and then Corollary 2.15,

$$\frac{\partial}{\partial \mathbf{n}} g_{\sigma(\mathcal{J}_0)}(0) = \frac{1}{L} \frac{\partial}{\partial \mathbf{n}} g_{\Omega}(0) = \frac{4}{3\sqrt{3}L}.$$

Combining the above three displays with the fact that  $|q'(0)| = 1$  is normalized for all  $q \in \mathcal{Q}_T$  yields

$$\min_{q \in \mathcal{Q}_T} \max_{\lambda \in \sigma(\mathcal{J}_0)} |q(\lambda)| \geq \left( \frac{3\sqrt{3}}{2} + o_T(1) \right) \frac{L\|z_0 - z^*\|}{T}.$$

□

## 4 Upper bounds for asymmetric algorithms

In this section we present *asymmetric* algorithms that break the convergence rate lower bounds established in §3 for *symmetric* algorithms. We do this for both the strongly and non-strongly convex-concave settings. In all settings, we use gradient-descent-ascent (GDA) with the slingshot stepsize schedules proposed in our previous work [27].

We begin with brief background on slingshot stepsizes. Recall that the update of GDA is

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \beta_t \nabla_y f(x_t, y_t). \end{aligned}$$

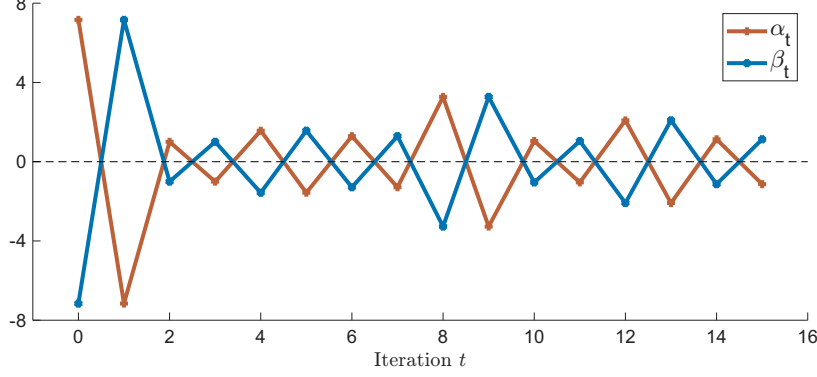


FIGURE 3: Slingshot stepsize schedule (Definition 4.1) for  $\mu = 0.1$ ,  $L = 1$ , and  $T = 16$ .

Note that GDA is a symmetric algorithm if and only if  $\alpha_t = \beta_t$  for all  $t$ . Following [27], we use slingshot stepsize schedules of the form

$$\alpha_{2t} = -\beta_{2t} = -\alpha_{2t+1} = \beta_{2t+1} = h_t, \quad \text{for } t = 0, \dots, T/2 - 1, \quad (4.1)$$

for appropriate choices of the magnitudes  $h_t$ . See Fig. 3 for an illustration. These stepsize schedules are asymmetric, time-varying, and periodically negative. They are implementable for min-max problems, but not for general variational inequality problems due to the asymmetry in  $x$  and  $y$ ; see §1.2 of [27] for a detailed discussion. By concatenating the variables  $z = (x, y)$ , this stepsize schedule results in paired updates of the form

$$\begin{aligned} z_{2t+1} &= z_{2t} - h_t \nabla f(z_{2t}), \\ z_{2t+2} &= z_{2t+1} + h_t \nabla f(z_{2t+1}), \end{aligned}$$

which for quadratic objectives  $f(z) = \frac{1}{2}(z - z^*)^T \mathbf{H}(z - z^*)$  as in (2.1), yields the two-step update

$$z_{2t+2} - z^* = (\mathbf{I} + h_t \mathbf{H})(\mathbf{I} - h_t \mathbf{H})(z_{2t} - z^*) = (\mathbf{I} - h_t^2 \mathbf{H}^2)(z_{2t} - z^*).$$

For  $T$  even, this results in a  $T$ -step cumulative update

$$z_T - z^* = \prod_{t < T/2} (\mathbf{I} - h_t^2 \mathbf{H}^2)(z_0 - z^*) \quad (4.2)$$

which is given by a matrix polynomial of the Hessian  $\mathbf{H}$  with roots  $\pm 1/h_t$ . We choose the stepsize magnitudes  $\{h_t\}$  explicitly in terms of the roots of certain Chebyshev polynomials; the optimal choice depends on the problem setting—see §4.1 and §4.2 below for the strongly and non-strongly convex-concave settings, respectively.

#### 4.1 Strongly-convex-strongly-concave problems

For this setting, we choose the slingshot stepsize magnitudes  $\{h_t\}$  in terms of the roots of the degree- $T/2$  Chebyshev polynomial  $\mathcal{T}_{T/2}^{[\mu^2, L^2]}$  of the first kind on the interval  $[\mu^2, L^2]$ . For background on Chebyshev polynomials, see for example the textbooks [13, 22].

**Definition 4.1** (Slingshot stepsize schedules for strongly-convex-strongly-concave min-max problems). *For any even number of iterations  $T = 2N$  and any parameters  $0 < \mu \leq L < \infty$ , the slingshot stepsize schedule for strongly-convex-strongly-concave problems is*

$$\alpha_{2t} = -\beta_{2t} = -\alpha_{2t+1} = \beta_{2t+1} = h_t, \quad t \in \{0, 1, \dots, N-1\},$$

where  $\{h_t\}_{t=0}^{N-1}$  are any permutation of  $\{r_t^{-1/2}\}_{t=0}^{N-1}$ , where

$$r_t := \frac{L^2 + \mu^2}{2} + \frac{L^2 - \mu^2}{2} \cos\left(\frac{2t+1}{T}\pi\right), \quad t \in \{0, 1, \dots, N-1\},$$

are the  $N$  roots of the Chebyshev polynomial  $\mathcal{T}_N^{[\mu^2, L^2]}$ .

Two remarks. First, despite the present setting (strongly-convex-strongly-concave quadratics) being different from the setting in [27, §3.1] (bilinear objectives with non-negative singular values), the stepsizes in Definition 4.1 exactly coincide with the ones we proposed in that paper. This is because in both problem settings, the Hessian  $\mathbf{H}$  has the same spectral range. Second, note that the order of the steps in Definition 4.1 does not matter due to commutativity of the updates, at least in exact arithmetic implementations; see Appendix A of [27] for fractal-like orderings that improve numerical stability.

**Theorem 4.2** (Upper bound for strongly-convex-strongly-concave min-max problems). *Consider any even integer  $T$ , any dimensions  $d_x, d_y$ , any initialization  $z_0 = (x_0, y_0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ , and any  $\mu$ -strongly-convex-strongly-concave quadratic min-max problem that is  $L$ -smooth. Using the slingshot stepsize schedule in Definition 4.1, GDA converges to the unique saddle point  $z^*$  at rate*

$$\frac{\|z_T - z^*\|}{\|z_0 - z^*\|} \leq \frac{2(\kappa+1)^{T/2}(\kappa-1)^{T/2}}{(\kappa+1)^T + (\kappa-1)^T} = \exp\left(-\left(1 + o_{T,\kappa}(1)\right) \frac{T}{\kappa}\right). \quad (4.3)$$

*Proof.* By (4.2) and Definition 4.1, we can write the  $T$ -step update of GDA with the proposed slingshot stepsizes as the following matrix polynomial of  $\mathbf{H}^2$ :

$$z_T - z^* = \prod_{t < T/2} (\mathbf{I} - h_t^2 \mathbf{H}^2) (z_0 - z^*) = \prod_{t < T/2} (\mathbf{I} - \mathbf{H}^2 / r_t) (z_0 - z^*) = \frac{\mathcal{T}_{T/2}^{[\mu^2, L^2]}(\mathbf{H}^2)}{\mathcal{T}_{T/2}^{[\mu^2, L^2]}(0)} (z_0 - z^*).$$

The convergence rate then follows:

$$\|z_T - z^*\| \leq \frac{\|\mathcal{T}_{T/2}^{[\mu^2, L^2]}(\mathbf{H}^2)\|}{\mathcal{T}_{T/2}^{[\mu^2, L^2]}(0)} \|z_0 - z^*\| \leq \frac{2(\kappa+1)^{T/2}(\kappa-1)^{T/2}}{(\kappa+1)^T + (\kappa-1)^T} \|z_0 - z^*\|.$$

Above, the first step is by sub-multiplicativity of the operator norm. The second step uses two classical facts about Chebyshev polynomials (see e.g., Lemma 3.2 of [27]), namely the closed-form expression for  $\mathcal{T}_{T/2}^{[\mu^2, L^2]}(0) = \frac{(\kappa+1)^{T/2} + (\kappa-1)^{T/2}}{2(\kappa+1)^{T/2}(\kappa-1)^{T/2}}$  and the fact that  $\sup_{\lambda \in [\mu^2, L^2]} |\mathcal{T}_{T/2}^{[\mu^2, L^2]}(\lambda)| = 1$ , which is applicable since  $\mathbf{H}^2$  is diagonalizable with eigenvalues in  $[\mu^2, L^2]$  by Lemma 2.6.

Finally, a Taylor expansion of the rate gives the desired asymptotics

$$\frac{2(\kappa+1)^{T/2}(\kappa-1)^{T/2}}{(\kappa+1)^T + (\kappa-1)^T} = \exp\left(-\left(1 + o_{\kappa,T}(1)\right) \frac{T}{\kappa}\right).$$

□

## 4.2 Convex-concave problems

For this setting, we can directly invoke the optimal convergence rate proven in our prior work [27]. For the convenience of the reader, below we recall these optimal stepsizes and the corresponding convergence rate. These appear originally as Definition 3.5 and Theorem 3.7 in [27].

**Definition 4.3** (Slingshot stepsize schedule for convex-concave min-max problems). *For any even number of iterations  $T$  and any  $L$ -smooth, convex-concave, quadratic min-max problem, the slingshot stepsize schedule is*

$$\alpha_t = -\beta_t = h_t, \quad t \in \{0, 1, \dots, T-1\},$$

where  $\{h_t^{-1}\}_{t=0}^{T-1}$  are any permutation of

$$\rho_t := L \cos\left(\frac{2t+1}{2T+2}\pi\right), \quad t \in \{0, \dots, T\} \setminus \{T/2\},$$

which are the  $T$  non-zero roots of the Chebyshev polynomial  $\mathcal{T}_{T+1}^{[-L, L]}$ . Notice that like in [Definition 4.1](#), these roots come in positive/negative pairs because

$$\rho_t = -\rho_{T-t}.$$

**Theorem 4.4** (Upper bound for convex-concave min-max problems). *Consider any even integer  $T$ , any dimensions  $d_x, d_y$ , any initialization  $z_0 = (x_0, y_0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ , and any quadratic min-max problem that is convex-concave and  $L$ -smooth. Using the slingshot stepsize schedule in [Definition 4.3](#), GDA converges at rate*

$$\|\nabla f(z_T)\| \leq \frac{L}{T+1} \|z_0 - z^*\|, \quad (4.4)$$

where  $z^*$  is any saddle point.

This convergence rate is exactly optimal—among not just arbitrary stepsize schedules for GDA, but in fact among arbitrary first-order algorithms—as it exactly matches the lower bound in [\[32\]](#).

## 5 Adaptive algorithms

In this section we extend the lower bounds for symmetric algorithms in [§3](#) to allow for the algorithms to be *adaptive*. Conceptually, this amounts to constructing a *single* problem instance for which no (adaptive) algorithm can perform well—in contrast to the arguments in [§3](#) which construct a potentially different hard problem for each (non-adaptive) algorithm.

These results extend for both the strongly-convex-strongly-concave and convex-concave settings. For the convenience of the reader, we state both results below in their entirety; the only difference from [Theorems 3.1](#) and [3.2](#), respectively, is that the results here allow the algorithms to be adaptive.

**Theorem 5.1** (Lower bound for adaptive symmetric algorithms on SCSC problems). *There exists a  $\kappa$ -conditioned quadratic min-max problem [\(2.1\)](#) such that for any (possibly adaptive) symmetric Krylov-subspace algorithm and any number of iterations  $T$ , the convergence rate from some initialization point  $z_0$  is no faster than*

$$\|z_T - z^*\| \geq \exp\left(-\left(\frac{4}{3\sqrt{3}} + o_\kappa(1)\right)\frac{T}{\kappa}\right) \|z_0 - z^*\|.$$

**Theorem 5.2** (Lower bound for adaptive symmetric algorithms on convex-concave problems). *There exists an  $L$ -smooth quadratic min-max problem [\(2.1\)](#) such that for any (possibly adaptive) symmetric Krylov-subspace algorithm and any number of iterations  $T$ , the convergence rate from some initialization point  $z_0$  is no faster than*

$$\|\nabla f(z_T)\| \geq \left(\frac{3\sqrt{3}}{2} + o_T(1)\right) \frac{L\|z_0 - z^*\|}{T}.$$



## 5.1 Proof of Theorem 5.1

The proofs of Theorems 5.1 and 5.2 follow from nearly identical extensions of Theorems 3.1 and 3.2, respectively. Therefore for brevity we only detail the former, i.e., how to prove Theorem 5.1 from Theorem 3.1. For conceptual clarity, we first outline the argument. For shorthand, throughout this section we denote  $\sigma(\mathcal{J}_\mu)$  simply by  $\sigma$ , since here there is no possibility of confusion with  $\sigma(\mathcal{H}_\mu)$ . Recall from Lemma 2.5 that  $\sigma = \{z : |z| \leq L, \operatorname{Re}(z) \geq \mu\}$ .

**Conceptual overview: hard problem instances for adaptive algorithms, via duality of the extremal polynomial problem.** Recall from §3 that our lower bounds begin with a reformulation of the optimal convergence rate in terms of an extremal polynomial problem  $\min_{p \in \mathcal{P}_T} \max_{\lambda \in \sigma} |p(\lambda)|$ , where the degree- $T$  polynomial  $p$  encodes the  $T$ -iteration evolution of the algorithm, and the eigenvalue  $\lambda$  of  $\mathbf{JH}$  encodes a worst-case problem instance for that algorithm. See (3.2). Importantly, as we considered only non-adaptive algorithms there, the polynomial  $p_T$  did not depend on the matrix  $\mathbf{JH}$ , so we could select a hard problem  $\lambda$  based on the algorithm  $p_T$  (hence the order of the min and max in  $\min_p \max_\lambda |p(\lambda)|$ ).

For adaptive algorithms, however, the polynomial  $p_T$  may depend on the matrix  $\mathbf{JH}$ , so we identify a single problem instance which is hard for all algorithms. In terms of the extremal polynomial problem, this requires a dual form where the order of the min and max are swapped. In particular, one may hope to prove a duality statement of the form

$$\min_{p \in \mathcal{P}_T} \max_{\lambda \in \sigma} |p(\lambda)|^2 = \max_{\nu \in \mathcal{M}(\sigma)} \min_{p \in \mathcal{P}_T} \mathbb{E}_{\lambda \sim \nu} |p(\lambda)|^2. \quad (5.1)$$

where  $\mathcal{M}(\sigma)$  is the set of probability measures supported on  $\sigma$ . Notice that the maximum over eigenvalues  $\lambda$  is lifted to a maximum over *probability distributions* on eigenvalues  $\nu$ ; in game-theoretic terminology this corresponds to relaxing pure strategies to mixed strategies.

At a conceptual level, our overall proof strategy amounts to proving a duality statement of the form (5.1), showing the existence of a solution  $\nu$ , and then using  $\nu$  to construct a hard problem instance for which the optimal convergence rate of any (possibly adaptive) symmetric algorithm is given by  $\min_{p \in \mathcal{P}_T} \mathbb{E}_{\lambda \sim \nu} |p(\lambda)|^2$ . This style of argument is inspired by Nemirovsky's classical lower bounds for solving symmetric linear systems using Krylov-subspace algorithms [16, 17], at least at a high level. However, implementing this proof strategy leads to additional technical considerations for min-max problems than in the simpler setting of quadratic minimization studied by Nemirovsky. Below we detail our two key steps to overcome these hurdles.

### Step 1: Nearly-optimal distributions with finite support and conjugation invariance.

The method we use to construct a hard problem from  $\nu$  is not amenable to arbitrary probability distributions  $\nu$ . We therefore show that we can impose two constraints on  $\nu$  that on one hand make our construction possible, and on the other hand only affect the final convergence rate by an arbitrarily small amount. The first condition we impose on  $\nu$  is finite support; later this will enable us to construct a hard problem instance in finite dimension. The second condition we impose on  $\nu$  is invariance under conjugation  $\nu(\lambda) = \nu(\bar{\lambda})$ ; later this will enable us to construct a hard problem whose operator  $\mathbf{JH}$  and solution  $z^*$  have all real entries.

**Lemma 5.3** (Step 1 in proof of Theorem 5.1). *For any number of iterations  $T$  and any error  $\varepsilon > 0$ , there exists a probability distribution  $\nu \in \mathcal{M}(\sigma)$  satisfying the following:*

- (i)  $\nu$  is finitely supported.

(ii)  $\nu$  is invariant under conjugation.

(iii)  $(1 - \varepsilon)^2 \min_{p \in \mathcal{P}_T} \max_{\lambda \in \sigma} |p(\lambda)|^2 \leq \min_{p \in \mathcal{P}_T} \mathbb{E}_{\lambda \sim \nu} |p(\lambda)|^2$ .

We prove this lemma in §5.2; here we focus on how we use it to prove Theorem 5.1.

**Step 2: Constructing a hard problem instance from  $\nu$ .** The second step of our argument uses  $\nu$  to construct a “hard” quadratic min-max problem, as formally stated next.

**Lemma 5.4** (Step 2 in proof of Theorem 5.1). *Suppose there exists a probability distribution  $\nu \in \mathcal{M}(\sigma)$  satisfying the following:*

(i)  $\nu$  is finitely supported.

(ii)  $\nu$  is invariant under conjugation.

(iii)  $R \leq \min_{p \in \mathcal{P}_T} \mathbb{E}_{\lambda \sim \nu} |p(\lambda)|^2$ .

Then there exists a  $\kappa$ -conditioned quadratic min-max problem (2.1) such that for any (possibly adaptive) symmetric Krylov-subspace algorithm and any number of iterations  $T$ , the convergence rate from some initialization point  $z_0$  is no faster than  $\|z_T - z^*\| \geq \sqrt{R} \|z_0 - z^*\|$ .

*Proof.* Let  $S$  denote the support of  $\nu$ ; this is finite by property (i). Define  $S^+ = \{\lambda \in S : \text{Im}(\lambda) \geq 0\}$ . We construct the  $\kappa$ -conditioned quadratic min-max problem (2.1). Define the blocks of the quadratic objective  $\mathbf{H}$  as  $\mathbf{A} = \mathbf{C} = \text{diag}(\{\text{Re}(\lambda_j) : \lambda_j \in S^+\})$  and  $\mathbf{B} = \text{diag}(\{\text{Im}(\lambda_j) : \lambda_j \in S^+\})$ . Then, modulo a permutation of rows and columns, the matrix

$$\mathbf{JH} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^\top & \mathbf{C} \end{bmatrix},$$

is block-diagonal with  $2 \times 2$  blocks of the form

$$\begin{bmatrix} \text{Re}(\lambda_j) & \text{Im}(\lambda_j) \\ -\text{Im}(\lambda_j) & \text{Re}(\lambda_j) \end{bmatrix}.$$

for every  $\lambda_j \in S^+$ . The eigenvalues of each such block are  $\lambda_j$  and  $\bar{\lambda}_j$ . (Note that for real  $\lambda_j$ , this  $2 \times 2$  block is diagonal with  $\lambda_j$  repeated twice.) Therefore the spectrum of  $\mathbf{JH}$  coincides with  $S$  because of the conjugation invariance in property (ii).

Finally we construct the initialization  $z_0$  and solution  $z^*$  so that distance to optimality on each eigenspace is  $\langle u_j, z_0 - z^* \rangle = c_j$  where  $c_j = \sqrt{\nu(\lambda_j)}/2$  if  $\lambda_j$  is real and  $c_j = \sqrt{\nu(\lambda_j)}$  otherwise, where  $u_j$  is the eigenvector associated with the eigenvalue  $\lambda_j$ . (The extra factor of 2 accounts for the double-counting of real eigenvalues described above.) This is achieved, for example, by the explicit construction  $z^* = z_0 - \sum_j c_j (e_j + e_{j+|S^+|})$ .

We now prove that this construction witnesses the desired lower bound for the convergence rate of any symmetric Krylov-subspace algorithm. Recall from §2.3 that the  $T$ -th iterate  $z_T$  of any such algorithm satisfies  $z_T - z^* = p_T(\mathbf{JH})(z_0 - z^*)$  for some polynomial  $p_T \in \mathcal{P}_T$ . Thus

$$\|z_T - z^*\|^2 = \|p_T(\mathbf{JH})(z_0 - z^*)\|^2 = \sum_j (|p_T(\lambda_j)| \cdot \langle u_j, z_0 - z^* \rangle)^2 = \mathbb{E}_{\lambda \sim \nu} |p(\lambda)|^2 \geq R.$$

Above, the first step is by definition of  $p_T$ , the second step is by block-diagonalizing, the third step is by definition of  $\nu$ , and the final step is by property (iii).  $\square$

Combining Lemma 5.3 and Lemma 5.4 immediately implies Theorem 5.1. It remains only to prove Lemma 5.3, which we do below.

## 5.2 Proof of Lemma 5.3

### 5.2.1 Helper lemmas

We begin with three helper lemmas. The first constructs the support of  $\nu$ , which we denote by  $S$ . Note that for our purposes,  $|S|$  need not be controlled so long as it is finite for every fixed  $T$  and  $\varepsilon$ .

**Lemma 5.5** (Finite mesh of  $\sigma$ ). *For every positive integer  $T$  and error  $\varepsilon > 0$ , there exists a finite subset  $S \subset \sigma$  satisfying*

$$(1 - \varepsilon)\|p\|_\sigma \leq \|p\|_S$$

for all polynomials  $p$  of degree at most  $T$ .

*Proof.* Define  $\sigma^+ = \{\lambda^+ : \exists \lambda \in \sigma, |\lambda^+ - \lambda| \leq 1\}$  to be the set of points of distance at most 1 from  $\sigma$ . Since the closure  $\text{cl}(\sigma^+ \setminus \sigma)$  is compact and  $g_\sigma$  is continuous on it,  $g_{\max} = \|g_\sigma\|_{\text{cl}(\sigma^+ \setminus \sigma)}$  is finite. Let  $S$  be any  $\delta$ -net of  $\partial\sigma$  for  $\delta = \varepsilon \exp(-Tg_{\max})$ ; that is, let  $S$  be a finite subset of  $\partial\sigma$  such that for every point in  $\partial\sigma$  there exists a point in  $S$  within distance  $\delta$ .

Now fix any polynomial  $p$  of degree at most  $T$ . By Cauchy's estimate<sup>5</sup> and then an extremal growth bound for polynomials via Green's function (Lemma 2.14)

$$\|p'\|_\sigma \leq \|p\|_{\sigma^+} \leq \exp(Tg_{\max})\|p\|_\sigma.$$

Now let  $\lambda^* \in \arg\max_{\lambda \in \sigma} |p(\lambda)|$ . By the maximum modulus principle,  $\lambda^* \in \partial\sigma$ . By definition of  $S$  as a  $\delta$ -net of  $\partial\sigma$ , there exists  $\lambda \in S$  for which  $|\lambda - \lambda^*| \leq \delta$ . By the above display, it follows that

$$||p(\lambda)| - |p(\lambda^*)|| \leq \delta \|p'\|_\sigma \leq \delta \exp(Tg_{\max})\|p\|_\sigma = \varepsilon \|p\|_\sigma.$$

Since  $|p(\lambda^*)| = \|p\|_\sigma$  by definition of  $\lambda^*$ , it follows that  $|p(\lambda)| \geq (1 - \varepsilon)\|p\|_\sigma$ . This proves the lemma since  $\lambda \in S$  is arbitrary.  $\square$

We also make use of the following two elementary helper lemmas about symmetry of polynomials along the real axis. Below, recall the notation that  $\mathcal{P}_T$  is the linear space of polynomials of degree at most  $T$  satisfying  $p(0) = 1$ , and let  $\mathcal{R}_T$  denote the subspace of  $\mathcal{P}_T$  containing only polynomials with real coefficients.

**Lemma 5.6** (Helper lemma 1). *Suppose  $S \subseteq \mathbb{C}$  is closed under conjugation. Then*

$$\min_{r \in \mathcal{R}_T} \|r\|_S = \min_{p \in \mathcal{P}_T} \|p\|_S.$$

*Proof.* Denote  $\tilde{p} = \overline{p(\bar{\lambda})}$  and  $r = \frac{1}{2}(p + \tilde{p})$ . Since  $S$  is closed under conjugation,  $\|p\|_S = \|\tilde{p}\|_S$ . Thus  $\|r\|_S = \frac{1}{2}\|p + \tilde{p}\|_S \leq \max\{\|p\|_S, \|\tilde{p}\|_S\} = \|p\|_S$ . Finally, to check that  $r$  has real coefficients, note that if  $p(\lambda) = \sum_t c_t \lambda^t$ , then  $\tilde{p}(\lambda) = \sum_t \bar{c}_t \lambda^t$ , hence  $r = \sum_t \text{Re}(c_t) \lambda^t$ .  $\square$

**Lemma 5.7** (Helper lemma 2). *Suppose  $\nu$  is invariant under conjugation. Then*

$$\min_{r \in \mathcal{R}_T} \mathbb{E}_{\lambda \sim \nu} |r(\lambda)|^2 = \min_{p \in \mathcal{P}_T} \mathbb{E}_{\lambda \sim \nu} |p(\lambda)|^2$$

---

<sup>5</sup>For completeness, we recall here Cauchy's estimate:  $|f^{(n)}(\lambda)| \leq \frac{n!}{r^n} \sup_{\lambda': |\lambda' - \lambda| \leq r} |f(\lambda')|$  for any holomorphic  $f$ ,  $n \in \mathbb{N}$ ,  $r > 0$ , and  $\lambda \in \mathbb{C}$ . A proof can be found, e.g., in [4, identity 2.14].

*Proof.* The direction “ $\geq$ ” is obvious since  $\mathcal{P}_T \supset \mathcal{R}_T$ . For the direction “ $\leq$ ”, observe that by linearity of expectation, it suffices to show that for any  $p \in \mathcal{P}_T$ , there exists  $r \in \mathcal{R}_T$  satisfying  $|r(\lambda)|^2 + |r(\bar{\lambda})|^2 \leq |p(\lambda)|^2 + |p(\bar{\lambda})|^2$  for all  $\lambda \in \mathbb{C}$ . To this end, define  $r$  as in the proof of [Lemma 5.6](#); i.e., let  $\tilde{p} = p(\bar{\lambda})$  and  $r = (p + \tilde{p})/2$ . Then  $r \in \mathcal{R}_T$  (as shown there) and satisfies the desired inequality since

$$|r(\lambda)|^2 + |r(\bar{\lambda})|^2 = 2|r(\lambda)|^2 = \frac{1}{2} |p(\lambda) + p(\bar{\lambda})|^2 \leq |p(\lambda)|^2 + |p(\bar{\lambda})|^2.$$

Above, the first step is because  $r(\lambda)$  and  $r(\bar{\lambda})$  are complex conjugates and thus have the same magnitude; the second step is by definition of  $r$ ; and the final step is by the elementary inequality  $|a + b|^2 \leq 2(|a|^2 + |b|^2)$  for any  $a, b \in \mathbb{C}$ .  $\square$

## 5.2.2 Combining the helper lemmas

*Proof of Lemma 5.3.* Let  $S$  be the finite set in [Lemma 5.5](#). Without loss of generality, suppose  $S$  is closed under conjugation (since otherwise we can include all conjugates). For shorthand, let  $\mathcal{M}_{\text{inv}}$  denote the subset of probability distributions in  $\mathcal{M}(S)$  that are invariant under conjugation.

Using in order: the definition of  $S$ , [Lemma 5.6](#), linearity of expectation, Sion’s minimax theorem, a symmetrization argument (replacing  $\nu(\lambda)$  by  $(\nu(\lambda) + \nu(\bar{\lambda}))/2$ ), and then [Lemma 5.7](#), we conclude

$$\begin{aligned} (1 - \varepsilon)^2 \min_{p \in \mathcal{P}_T} \max_{\lambda \in S} |p(\lambda)|^2 &\leq \min_{p \in \mathcal{P}_T} \max_{\lambda \in S} |p(\lambda)|^2 \\ &= \min_{r \in \mathcal{R}_T} \max_{\lambda \in S} |r(\lambda)|^2 \\ &= \min_{r \in \mathcal{R}_T} \max_{\nu \in \mathcal{M}(S)} \mathbb{E}_{\lambda \sim \nu} |r(\lambda)|^2 \\ &= \max_{\nu \in \mathcal{M}(S)} \min_{r \in \mathcal{R}_T} \mathbb{E}_{\lambda \sim \nu} |r(\lambda)|^2 \\ &= \max_{\nu \in \mathcal{M}_{\text{inv}}(S)} \min_{r \in \mathcal{R}_T} \mathbb{E}_{\lambda \sim \nu} |r(\lambda)|^2 \\ &= \max_{\nu \in \mathcal{M}_{\text{inv}}(S)} \min_{p \in \mathcal{P}_T} \mathbb{E}_{\lambda \sim \nu} |p(\lambda)|^2. \end{aligned}$$

Let  $\nu$  be an optimal solution to the final expression; existence is guaranteed by compactness. Then  $\nu$  satisfies all three desired properties: the first by finiteness of  $S$ , the second because  $\nu \in \mathcal{M}_{\text{inv}}(S)$ , and the third by the above display.  $\square$

**Acknowledgements.** We are grateful to Joel Tropp for helpful discussions about the literature. JMA acknowledges funding from a Sloan Research Fellowship and a Seed Grant Award from Apple.

## References

- [1] Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *International Conference on Artificial Intelligence and Statistics*, 2020.
- [2] Serge Bernstein. *Sur l’ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné*, volume 4. Hayez, imprimeur des académies royales, 1912.
- [3] Jiseok Chae, Kyuwon Kim, and Donghwan Kim. Two-timescale extragradient for finding local minimax points. In *International Conference on Learning Representations*, 2024.

- [4] John B Conway. *Functions of one complex variable II*, volume 159. Springer Science & Business Media, 2012.
- [5] Tobin A. Driscoll, Kim-Chuan Toh, and Lloyd N. Trefethen. From potential theory to matrix iterations in six steps. *SIAM Review*, 40(3):547–578, 1998.
- [6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.
- [7] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.
- [8] Adam Ibrahim, Waiss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593, 2020.
- [9] Dmitry Kovalev and Alexander Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.
- [10] Jaewook Lee, Hanseul Cho, and Chulhee Yun. Fundamental benefit of alternating updates in minimax optimization. In *International Conference on Machine Learning*, pages 26439–26514, 2024.
- [11] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, page 6083–6093, 2020.
- [12] Songtao Lu, Rahul Singh, Xiangyi Chen, Yongxin Chen, and Mingyi Hong. Alternating gradient descent ascent for nonconvex min-max problems in robust learning and GANs. In *Asilomar Conference on Signals, Systems, and Computers*, 2019.
- [13] John C Mason and David C Handscomb. *Chebyshev polynomials*. Chapman and Hall/CRC, 2002.
- [14] Béla Nagy. Asymptotic Bernstein inequality on lemniscates. *Journal of Mathematical Analysis and Applications*, 301(2):449–456, 2005. ISSN 0022-247X.
- [15] Béla Nagy and Vilmos Totik. Sharpening of Hilbert’s lemniscate theorem. *Journal d’Analyse Mathématique*, 96(1):191–223, 2005.
- [16] Arkadi S Nemirovsky. On optimality of Krylov’s information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.
- [17] Arkadi S Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.
- [18] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 1998.
- [19] Christian Pommerenke. On metric properties of complex polynomials. *Michigan Mathematical Journal*, 8(2):97–115, 1961.
- [20] Qazi Rahman and Q. G. Mohammad. Remarks on Schwarz’s lemma. *Pacific Journal of Mathematics*, 23(1):139–142, 1967.
- [21] Thomas Ransford. *Potential theory in the complex plane*. Cambridge University Press, 1995.
- [22] Theodore J Rivlin. *Chebyshev polynomials*. Courier Dover Publications, 2020.
- [23] R. T. Rockafellar. *Monotone operators associated with saddle-functions and minimax problems*, volume 18.1, page 241–250. American Mathematical Society, 1970.

- [24] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [25] Ernest K Ryu and Wotao Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.
- [26] Edward B Saff. Logarithmic potential theory with applications to approximation theory. *Preprint at arXiv:1010.3760*, 2010.
- [27] Henry Shugart and Jason M Altschuler. Negative stepsizes make gradient-descent-ascent converge. *Preprint at arXiv:2505.01423*, 2025.
- [28] Elias M Stein and Rami Shakarchi. *Complex analysis*, volume 2. Princeton University Press, 2010.
- [29] Lloyd N Trefethen. *Approximation theory and approximation practice, extended edition*. SIAM, 2019.
- [30] Joseph L Walsh. Über den grad der approximation einer analytischen funktion. 1926.
- [31] Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. *Advances in Neural Information Processing Systems*, 33:4800–4810, 2020.
- [32] Taeho Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with  $O(1/k^2)$  rate on squared gradient norm. In *International Conference on Machine Learning*, page 12098–12109, 2021.
- [33] Guodong Zhang, Yuanhao Wang, Laurent Lessard, and Roger B. Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, 2022.