A. Algorithm specifications

For the sake of clarity, we precisely specify all the algorithms discussed in this work.

Simultaneous gradient descent for smooth minimax optimization is defined as

$$\begin{aligned}\mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha \nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}^k, \mathbf{y}^k)\\ \mathbf{y}^{k+1} &= \mathbf{y}^k + \alpha \nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}^k, \mathbf{y}^k).\end{aligned}$$

The notation becomes more concise with the joint variable notation $\mathbf{z}^k = (\mathbf{x}^k, \mathbf{y}^k)$ and the saddle operator (2), where the sign change in $\mathbf{y}$-gradient is already included:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha \mathbf{G}(\mathbf{z}^k).$$

Alternating gradient descent-ascent is defined as

$$\begin{aligned}\mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha \nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}^k, \mathbf{y}^k)\\ \mathbf{y}^{k+1} &= \mathbf{y}^k + \alpha \nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}^{k+1}, \mathbf{y}^k).\end{aligned}$$

Note that we update the $\mathbf{x}$ variable first and then use it to update the $\mathbf{y}$-iterate.

The *extragradient (EG) algorithm* is defined as

$$\begin{aligned}\mathbf{z}^{k+1/2} &= \mathbf{z}^k - \alpha \mathbf{G}(\mathbf{z}^k),\\ \mathbf{z}^{k+1} &= \mathbf{z}^k - \alpha \mathbf{G}(\mathbf{z}^{k+1/2}).\end{aligned}$$

Popov's algorithm, or *optimistic descent*, is defined as

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha \mathbf{G}(\mathbf{z}^k) - \alpha \left(\mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k-1})\right).$$

Simultaneous gradient descent with anchoring (SimGD-A) (Ryu et al., 2019) is defined as

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{1-p}{(k+1)^p}\mathbf{G}(\mathbf{z}^k) + \frac{(1-p)\gamma}{k+1}(\mathbf{z}^0 - \mathbf{z}^k),$$

where $p \in (1/2, 1)$ and $\gamma > 0$. It has been proved in Ryu et al. (2019) that SimGD-A converges at $\mathcal{O}(1/k^{2-2p})$ rate. In this paper, we always used $\gamma = 1$ and $p = \frac{1}{2} + 10^{-2}$.

B. Omitted proofs of Section 2

The following identities follow directly from the definition of EAG iterates:

$$\mathbf{z}^k - \mathbf{z}^{k+1} = \beta_k(\mathbf{z}^k - \mathbf{z}^0) + \alpha_k \, \mathbf{G}(\mathbf{z}^{k+1/2}) \tag{26}$$

$$\mathbf{z}^{k+1/2} - \mathbf{z}^{k+1} = \alpha_k \left(\mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^k)\right) \tag{27}$$

$$\mathbf{z}^0 - \mathbf{z}^{k+1} = (1-\beta_k)(\mathbf{z}^0 - \mathbf{z}^k) + \alpha_k \, \mathbf{G}(\mathbf{z}^{k+1/2}). \tag{28}$$

B.1. Proof of Lemma 2

Recall that $\mathbf{G}$ is a monotone operator, so that

$$0 \le \left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1})\right\rangle.$$

Therefore,

$$\begin{aligned}
&V_k - V_{k+1} \\
&\geq V_k - V_{k+1} - \frac{B_k}{\beta_k}\left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1})\right\rangle \\
&= A_k \left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 + B_k \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0\right\rangle \\
&\quad - A_{k+1}\left\|\mathbf{G}(\mathbf{z}^{k+1})\right\|^2 - B_{k+1}\left\langle \mathbf{G}(\mathbf{z}^{k+1}), \mathbf{z}^{k+1} - \mathbf{z}^0\right\rangle - \frac{B_k}{\beta_k}\left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1})\right\rangle \\
&\overset{\text{(a)}}{=} A_k \left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 + B_k \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0\right\rangle \\
&\quad - A_{k+1}\left\|\mathbf{G}(\mathbf{z}^{k+1})\right\|^2 + B_{k+1}\left\langle \mathbf{G}(\mathbf{z}^{k+1}), (1-\beta_k)(\mathbf{z}^0 - \mathbf{z}^k) + \alpha_k\,\mathbf{G}(\mathbf{z}^{k+1/2})\right\rangle \\
&\quad - B_k\left\langle \mathbf{z}^k - \mathbf{z}^0, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1})\right\rangle - \frac{\alpha_k B_k}{\beta_k}\left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1})\right\rangle \\
&\overset{\text{(b)}}{=} A_k \left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 - A_{k+1}\left\|\mathbf{G}(\mathbf{z}^{k+1})\right\|^2 + \alpha_k B_{k+1}\left\langle \mathbf{G}(\mathbf{z}^{k+1}), \mathbf{G}(\mathbf{z}^{k+1/2})\right\rangle \\
&\quad - \frac{\alpha_k B_k}{\beta_k}\left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1})\right\rangle,
\end{aligned} \tag{29}$$

where (a) follows from (26) and (28), and (b) results from cancellation and collection of terms using (7). Next, we have

$$\begin{aligned}
0 &\leq R^2\left\|\mathbf{z}^{k+1/2} - \mathbf{z}^{k+1}\right\|^2 - \left\|\mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1})\right\|^2 \\
&= \alpha_k^2 R^2\left\|\mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1/2})\right\|^2 - \left\|\mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1})\right\|^2
\end{aligned} \tag{30}$$

from R-Lipschitzness of $\mathbf{G}$ and (27). Now multiplying the factor $\frac{A_k}{\alpha_k^2 R^2}$ to (30) and subtracting from (29) gives

$$\begin{aligned}
&V_k - V_{k+1} \\
&\geq A_k \left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 - A_{k+1}\left\|\mathbf{G}(\mathbf{z}^{k+1})\right\|^2 + \alpha_k B_{k+1}\left\langle \mathbf{G}(\mathbf{z}^{k+1}), \mathbf{G}(\mathbf{z}^{k+1/2})\right\rangle \\
&\quad - \frac{\alpha_k B_k}{\beta_k}\left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1})\right\rangle \\
&\quad - A_k\left\|\mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1/2})\right\|^2 + \frac{A_k}{\alpha_k^2 R^2}\left\|\mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1})\right\|^2 \\
&= \frac{A_k(1-\alpha_k^2 R^2)}{\alpha_k^2 R^2}\left\|\mathbf{G}(\mathbf{z}^{k+1/2})\right\|^2 + \left(\frac{A_k}{\alpha_k^2 R^2} - A_{k+1}\right)\left\|\mathbf{G}(\mathbf{z}^{k+1})\right\|^2 \\
&\quad + \left(2A_k - \frac{\alpha_k B_k}{\beta_k}\right)\left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{G}(\mathbf{z}^{k+1/2})\right\rangle \\
&\quad + \left(\alpha_k B_{k+1} + \frac{\alpha_k B_k}{\beta_k} - \frac{2A_k}{\alpha_k^2 R^2}\right)\left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^{k+1})\right\rangle.
\end{aligned} \tag{31}$$

Observe that the $\left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{G}(\mathbf{z}^{k+1/2})\right\rangle$ term vanishes because of (6), and that

$$\alpha_k B_{k+1} + \frac{\alpha_k B_k}{\beta_k} = \alpha_k\left(\frac{B_k}{1-\beta_k} + \frac{B_k}{\beta_k}\right) = \frac{\alpha_k B_k}{\beta_k(1-\beta_k)} = \frac{2A_k}{1-\beta_k}.$$

Furthermore, by (8), we have

$$A_{k+1} = \alpha_{k+1}\frac{B_{k+1}}{2\beta_{k+1}} = \frac{\alpha_k\beta_{k+1}(1-\alpha_k^2R^2-\beta_k^2)}{(1-\alpha_k^2R^2)\beta_k(1-\beta_k)}\,\frac{B_k}{2\beta_{k+1}(1-\beta_k)} = \frac{A_k(1-\alpha_k^2R^2-\beta_k^2)}{(1-\alpha_k^2R^2)(1-\beta_k)^2}.$$

Plugging these identities into (31) and simplifying, we get

$$\begin{aligned}
&V_k - V_{k+1} \\
&\geq \frac{A_k(1-\alpha_k^2R^2)}{\alpha_k^2R^2}\left\|\mathbf{G}(\mathbf{z}^{k+1/2})\right\|^2 + \frac{A_k(1-\alpha_k^2R^2-\beta_k)^2}{\alpha_k^2R^2(1-\alpha_k^2R^2)(1-\beta_k)^2}\left\|\mathbf{G}(\mathbf{z}^{k+1})\right\|^2 \\
&\quad - \frac{2A_k(1-\alpha_k^2R^2-\beta_k)}{\alpha_k^2R^2(1-\beta_k)}\left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^{k+1})\right\rangle \\
&\geq 0,
\end{aligned}$$

where the last inequality is an application of Young's inequality.

B.2. Proof of Lemma 1

We may assume $R = 1$ without loss of generality because we can recover the general case by replacing α_k with $\alpha_k R$. Rewrite (5) as

$$\alpha_k - \alpha_{k+1} = \frac{\alpha_k^3}{(k+1)(k+3)(1-\alpha_k^2)}. \tag{32}$$

Suppose that we have already established $0 < \alpha_N < \rho$ for some $N \geq 0$ and $\rho \in (0,1)$, where ρ satisfies

$$\gamma := \frac{1}{2}\left(\frac{1}{N+1} + \frac{1}{N+2}\right)\frac{\rho^2}{1-\rho^2} < 1. \tag{33}$$

Note that (33) holds true for all $N \geq 0$ if $\rho < \frac{3}{4}$. Now we will show that given (33),

$$\alpha_N > \alpha_{N+1} > \cdots > \alpha_{N+k} > (1-\gamma)\alpha_N \quad \text{for all } k \geq 0,$$

so that $\alpha_k \downarrow \alpha$ for some $\alpha \geq (1-\gamma)\alpha_N$. It suffices to prove that $(1-\gamma)\alpha_N < \alpha_{N+k} < \rho$ for all $k \geq 0$, because it is clear from (32) that $\{\alpha_k\}_{k\geq 0}$ is decreasing.

We use induction on k to prove that $\alpha_{N+k} \in ((1-\gamma)\alpha_N, \rho)$. The case $k = 0$ is trivial. Now suppose that $(1-\gamma)\alpha_N < \alpha_{N+j} < \rho$ holds true for all $j = 0, \ldots, k$. Then by (32), for each $0 \leq j \leq k$ we have

$$\begin{aligned}
0 < \alpha_{N+j} - \alpha_{N+j+1} &= \frac{1}{(N+j+1)(N+j+3)}\frac{\alpha_{N+j}^3}{1-\alpha_{N+j}^2} \\
&< \frac{1}{(N+j+1)(N+j+3)}\frac{\rho^2\alpha_N}{1-\rho^2}.
\end{aligned}$$

Summing up the inequalities for $j = 0, \ldots, k$, we obtain

$$\begin{aligned}
0 < \alpha_N - \alpha_{N+k+1} &< \sum_{j=0}^{k}\frac{1}{(N+j+1)(N+j+3)}\frac{\rho^2\alpha_N}{1-\rho^2} \\
&< \frac{\rho^2\alpha_N}{1-\rho^2}\sum_{j=0}^{\infty}\frac{1}{(N+j+1)(N+j+3)} \\
&= \frac{\rho^2\alpha_N}{1-\rho^2}\frac{1}{2}\left(\frac{1}{N+1} + \frac{1}{N+2}\right) = \gamma\alpha_N,
\end{aligned}$$

which gives $(1-\gamma)\alpha_N < \alpha_{N+k+1} < \alpha_N < \rho$, completing the induction.

In particular, when $\alpha_0 = 0.618$, direct calculation gives $0.437 > \alpha_N > 0.4366$ when $N = 1000$. With $\rho = 0.437$ and $N = 1000$, we have $\gamma = \frac{1}{2}\left(\frac{1}{N+1} + \frac{1}{N+2}\right)\frac{\rho^2}{1-\rho^2} < 2.5 \times 10^{-4}$, which gives $\alpha \geq (1-\gamma)\alpha_N \approx 0.4365$.

B.3. Proof of Theorem 1

As in the proof of Theorem 2, assume without loss of generality that $R = 1$. The strategy of the proof is basically the same as in Theorem 2; we construct a nonincreasing Lyapunov function by combining the same set of inequalities, but with different (more intricate) coefficients. For $k \geq 0$, let

$$V_k = A_k \left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 + B_k \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \right\rangle.$$

As in Lemma 2, we will use $B_k = \frac{1}{1-\beta_k} = k+1$, and $a_k \geq 0$ will be specified later. Because we have the fixed step-size α, the identities (26), (27), and (28) become

$$\begin{aligned}
\mathbf{z}^{k+1/2} - \mathbf{z}^{k+1} &= \alpha \left(\mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^k) \right) \\
\mathbf{z}^k - \mathbf{z}^{k+1} &= \frac{1}{k+2}(\mathbf{z}^k - \mathbf{z}^0) + \alpha\, \mathbf{G}(\mathbf{z}^{k+1/2}) \\
\mathbf{z}^{k+1} - \mathbf{z}^0 &= \frac{k+1}{k+2}(\mathbf{z}^k - \mathbf{z}^0) - \alpha\, \mathbf{G}(\mathbf{z}^{k+1/2}).
\end{aligned}$$

Now, subtracting the same inequalities from monotonicity and Lipschitzness from $V_k - V_{k+1}$ as in Lemma 2, each with coefficients $(k+1)(k+2)$ and $\tau_k \geq 0$ (to be specified later), we obtain

$$\begin{aligned}
&V_k - V_{k+1} \\
&\geq V_k - V_{k+1} - (k+1)(k+2) \left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\
&\quad - \tau_k \left(\left\| \mathbf{z}^{k+1/2} - \mathbf{z}^{k+1} \right\|^2 - \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \right) \\
&= (A_k - \alpha^2 \tau_k) \left\| \mathbf{G}(\mathbf{z}^k) \right\|^2 + \tau_k (1-\alpha^2) \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) \right\|^2 + (\tau_k - A_{k+1}) \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \\
&\quad + \left(2\alpha^2 \tau_k - \alpha(k+1)(k+2) \right) \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{G}(\mathbf{z}^{k+1/2}) \right\rangle + \left(\alpha(k+2)^2 - 2\tau_k \right) \left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\
&= \operatorname{Tr}\left(\mathbf{M}_k \mathbf{S}_k \mathbf{M}_k^\intercal \right),
\end{aligned}$$

where we define $\mathbf{M}_k := \begin{bmatrix} \mathbf{G}(\mathbf{z}^k) & \mathbf{G}(\mathbf{z}^{k+1/2}) & \mathbf{G}(\mathbf{z}^{k+1}) \end{bmatrix}$ and

$$\mathbf{S}_k := \begin{bmatrix} A_k - \alpha^2 \tau_k & \alpha^2 \tau_k - \frac{\alpha}{2}(k+1)(k+2) & 0 \\ \alpha^2 \tau_k - \frac{\alpha}{2}(k+1)(k+2) & \tau_k(1-\alpha^2) & \frac{\alpha}{2}(k+2)^2 - \tau_k \\ 0 & \frac{\alpha}{2}(k+2)^2 - \tau_k & \tau_k - A_{k+1} \end{bmatrix}. \tag{34}$$

If $\mathbf{S}_k \succeq \mathbf{O}$, then $\operatorname{Tr}\left(\mathbf{M}_k \mathbf{S}_k \mathbf{M}_k^\intercal\right) = \operatorname{Tr}\left(\mathbf{S}_k \mathbf{M}_k^\intercal \mathbf{M}_k\right) \geq 0$ because the positive semidefinite cone is self-dual with respect to the matrix inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{Tr}(\mathbf{A}^\intercal \mathbf{B})$. Because $B_k = k+1$ grows linearly, provided that the sequence $\{A_k\}$ grows quadratically, we can derive $\mathcal{O}(1/k^2)$ convergence by using similar line of arguments as in the proof of Theorem 2. This reduction of the proof into a search of appropriate parameters (i.e., τ_k) that meet semidefiniteness constraints ($\mathbf{S}_k \succeq \mathbf{O}$ in our case) while allowing for desired rate of growth in Lyapunov function coefficients (A_k in our case) was inspired by works of Taylor et al. (2017) and Taylor & Bach (2019). In the following, we demonstrate that careful choices of A_0 and τ_k make A_k asymptotically close to $\frac{\alpha(k+1)(k+2)}{2}$, so quadratic growth is guaranteed. We begin with the following lemma, which will be used throughout the proof.

Lemma 5. *Let $k \in \mathbb{N}_{\geq 0}$ and $\alpha \in \left(0, \frac{1}{2}\right]$ be fixed, and define*

$$\ell_k := \frac{\alpha(k+2)(k+1+k\alpha)}{2(1+\alpha)}, \quad u_k := \frac{\alpha(k+2)(k+1-k\alpha)}{2(1-\alpha)}.$$

Then,

$$\begin{aligned}
u_k > \frac{\alpha(k+1)(k+2)}{2} > \ell_k & \qquad (35)\\
\geq \frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)} & \qquad (36)\\
\geq \frac{\alpha(k+1)^2 - \alpha^3 k(k+2)}{2(1-\alpha^2)} & \qquad (37)\\
\geq \max\left\{\frac{\alpha(k+1)(k+1-\alpha(k+2))}{2(1-\alpha)}, \frac{\alpha^2(k+1)(k+2)}{1+\alpha}\right\} & \qquad (38)\\
\geq \frac{\alpha^2(k+1)(k+2)+\alpha^3(k+2)^2}{2(1+\alpha)}. & \qquad (39)
\end{aligned}$$

We shall prove Lemma 5 after the proof of the main theorem and for now, focus on why we need such results. Observe that all the quantities within the lines (35) through (37) are asymptotically close to $\frac{\alpha k^2}{2}$. We show that $A_k \in I_k := [\ell_k, u_k]$ for all $k \geq 0$, which implies the quadratic growth. The quantities in Lemma 5 are used for choosing the right τ_k and for showing the positive semidefiniteness of $\mathbf{S}_k$.

Subdivide the interval I_k into two parts:

$$I_k^- = \left[\ell_k, \frac{\alpha(k+1)(k+2)}{2}\right], \quad I_k^+ = \left[\frac{\alpha(k+1)(k+2)}{2}, u_k\right].$$

We divide cases: $A_k \in I_k^-$ and $A_k \in I_k^+$. However, the latter case is in fact not needed unless we wish to extend the proof for α beyond $\frac{0.1265}{R}$. If that is not the case, we recommend the readers to refer to Case 1 only. Nevertheless, we exhibit analysis of both cases because Case 2 might provide useful data for enlarging or even completely determining the range of convergent step-sizes for EAG-C.

Case 1. Suppose that $A_k \in I_k^-$. In this case, we choose

$$\tau_k = \frac{(k+2)^2\left(2(1-\alpha)A_k - \alpha(k+1)(k+1-\alpha(k+2))\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)}. \tag{40}$$

The denominator and numerator of (40) are both positive because $u_k > A_k > \frac{\alpha(k+1)(k+1-\alpha(k+2))}{2(1-\alpha)}$ (see (38)). Thus, $\tau_k > 0$. Next, define A_{k+1} as

$$\begin{aligned}
A_{k+1} &= \frac{\alpha(k+2)^2\left(4(1-\alpha)A_k - \alpha(k+1-\alpha(k+2))^2\right)}{4(1-\alpha)\left((1-\alpha)A_k + \alpha^2(k+1)(k+2)\right)}\\
&= \frac{\alpha(k+2)^2}{1-\alpha}\left(1 - \frac{\alpha(k+1+\alpha(k+2))^2}{4((1-\alpha)A_k + \alpha^2(k+1)(k+2))}\right).
\end{aligned} \tag{41}$$

Then (34) can be rewritten as

$$\mathbf{S}_k = \begin{bmatrix} s_{11} & s_{12} & 0 \\ s_{12} & s_{22} & s_{23} \\ 0 & s_{23} & s_{33} \end{bmatrix},$$

where

$$s_{11} = \frac{(\alpha(k+1)(k+2) - 2A_k)\left(2(1-\alpha)A_k + \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{42}$$

$$s_{12} = -\frac{\alpha(1-\alpha)(k+2)(k+1+\alpha(k+2))(\alpha(k+1)(k+2) - 2A_k)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{43}$$

$$s_{22} = \frac{(1-\alpha^2)(k+2)^2\left(2(1-\alpha)A_k - \alpha(k+1)(k+1-\alpha(k+2))\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{44}$$

$$s_{23} = -\frac{(k+2)^2\left(2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3 k(k+2)\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{45}$$

$$s_{33} = \frac{(k+2)^2\left(2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3 k(k+2)\right)\left(2(1-\alpha)A_k + \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2\right)}{4(1-\alpha)\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)\left((1-\alpha)A_k + \alpha^2(k+1)(k+2)\right)}. \tag{46}$$

The expressions seem ridiculously complicated, but there are a number of repeating terms. Let

$$\begin{aligned} E_1 &= \alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k \\ E_2 &= \alpha(k+1)(k+2) - 2A_k. \end{aligned}$$

Because $A_k \le \frac{\alpha(k+1)(k+2)}{2} < u_k$ (see (35)), we have $E_1 > 0, E_2 \ge 0$. (Note that $E_2 = 0$ only in the boundary case $A_k = \sup I_k^-$.) Next, put

$$E_3 = 2(1-\alpha)A_k - \alpha(k+1)(k+1-\alpha(k+2)),$$

which is a factor that appears within the definition of τ_k (40); we have already seen that $E_3 > 0$. Further, let

$$\begin{aligned} E_4 &= 2(1-\alpha)A_k + \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2 \\ E_5 &= (1-\alpha)A_k + \alpha^2(k+1)(k+2) \\ E_6 &= 2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3 k(k+2) \\ E_7 &= k+1+\alpha(k+2). \end{aligned}$$

It is obvious that $E_5, E_7 > 0$, and $E_6 > 0$ follows directly from (37). To see that $E_4 > 0$, observe that $k+1-\alpha(k+2) = (k+2)\left(\frac{k+1}{k+2} - \alpha\right) \ge (k+2)\left(\frac{1}{2} - \alpha\right) \ge 0$, provided that $\alpha \le \frac{1}{2}$. This implies

$$E_4 = 2(1-\alpha)A_k + \alpha^2(k+2)\left(k+1-(k+2)\alpha\right) > 0.$$

Now we can rewrite (42) through (46) as

$$\begin{aligned} s_{11} &= \frac{E_2E_4}{2E_1} \\ s_{12} &= -\frac{\alpha(1-\alpha)(k+2)E_2E_7}{2E_1} \\ s_{22} &= \frac{(1-\alpha^2)(k+2)^2E_3}{2E_1} \\ s_{23} &= -\frac{(k+2)^2E_6}{2E_1} \\ s_{33} &= \frac{(k+2)^2E_4E_6}{4(1-\alpha)E_1E_5}. \end{aligned}$$

This immediately shows that the diagonal entries s_{ii} are nonnegative for $i = 1, 2, 3$. By brute-force calculation, it is not difficult to verify the identity

$$(1+\alpha)E_3E_4 = \alpha^2(1-\alpha)E_2E_7^2 + 2E_5E_6.$$

Using this, we see that $\mathbf{v} := \begin{bmatrix} \frac{\alpha(k+2)E_7}{2E_5} & \frac{E_4}{2(1-\alpha)E_5} & 1 \end{bmatrix}^\intercal$ satisfies $\mathbf{S}_k\mathbf{v} = 0$, and this implies $\det \mathbf{S}_k = 0$. The cofactor-expansion of $\det \mathbf{S}_k$ along the first row gives

$$0 = \det \mathbf{S}_k = s_{11}\begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix} - s_{12}\begin{vmatrix} s_{12} & s_{23} \\ 0 & s_{33} \end{vmatrix} \iff \begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix} = \frac{s_{12}^2 s_{33}}{s_{11}} > 0$$

when $s_{11} > 0$, and via continuity argument we can argue that $\begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix} \geq 0$ even in the boundary case $s_{11} = 0$. Similarly one can show that $\begin{vmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{vmatrix} \geq 0$. Therefore, we have shown that all diagonal submatrices of $\mathbf{S}_k$ (including the trivial case $\begin{vmatrix} s_{11} & 0 \\ 0 & s_{33} \end{vmatrix} = s_{11}s_{33} \geq 0$) have nonnegative determinants, that is, $\mathbf{S}_k \succeq \mathbf{O}$.

Finally, (41) shows that A_{k+1} is increasing with respect to A_k. We see that

$$A_{k+1}\Big|_{A_k=\frac{\alpha(k+1)(k+2)}{2}} = \frac{\alpha(k+2)((k+1)(k+3)-\alpha^2(k+2)^2)}{2(1-\alpha^2)(k+1)} < \frac{\alpha(k+2)(k+3)}{2} \tag{47}$$

and

$$A_{k+1}|_{A_k=\ell_k} - \ell_{k+1} = \frac{\alpha^2\left((1-3\alpha-\alpha^2-\alpha^3)k+1-8\alpha+\alpha^2-2\alpha^3\right)}{2(1-\alpha^2)\left((1+\alpha)^2k+1+\alpha+2\alpha^2\right)},$$

and the last expression is nonnegative because of the assumption (4), which we restate here for the case $R = 1$ for convenience: $1-3\alpha-\alpha^2-\alpha^3 \geq 0$ and $1-8\alpha+\alpha^2-2\alpha^3 \geq 0$. This proves that $A_{k+1} \in I_{k+1}^- \subset I_{k+1}$, as desired.

Case 2. Suppose that $A_k \in I_k^+$. The proof would be similar to Case 1, but choices of τ_k and A_{k+1} are different. We let

$$\tau_k = \frac{(k+2)^2\left(2(1+\alpha)A_k - \alpha(k+1)(k+1+\alpha(k+2))\right)}{4(1+\alpha)A_k - 2\alpha(k+2)(k+1+k\alpha)}. \tag{48}$$

Since $A_k > \ell_k > \frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)}$, the denominator and numerator of (48) are both positive and thus $\tau_k > 0$. Next, let

$$\begin{aligned} A_{k+1} &= \frac{\alpha(k+2)^2\left(4(1+\alpha)A_k - \alpha(k+1+\alpha(k+2))^2\right)}{4(1+\alpha)\left((1+\alpha)A_k - \alpha^2(k+1)(k+2)\right)} \\ &= \frac{\alpha(k+2)^2}{1+\alpha}\left(1 - \frac{\alpha(k+1-\alpha(k+2))^2}{4((1+\alpha)A_k - \alpha^2(k+1)(k+2))}\right). \end{aligned} \tag{49}$$

Then we can check that

$$\begin{aligned} s_{11} &= \frac{(2A_k - \alpha(k+1)(k+2))(2(1+\alpha)A_k - \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2)}{4(1+\alpha)A_k - 2\alpha(k+2)(k+1+k\alpha)} \\ s_{33} &= \frac{(k+2)^2\left(2(1+\alpha)A_k - \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2\right)\left(2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3k(k+2)\right)}{4(1+\alpha)\left(2(1+\alpha)A_k - \alpha(k+2)(k+1+k\alpha)\right)\left(2(1+\alpha)A_k - \alpha^2(k+1)(k+2)\right)}, \end{aligned}$$

and so on. (Note that $2A_k - \alpha(k+1)(k+2) \geq 0$ because now we are assuming that $A_k \in I_k^+$.) We omit further details of calculations, but with the above choices of τ_k and A_{k+1} it can be shown that $\det \mathbf{S}_k = 0$ and $s_{11}, s_{33} \geq 0$, using (36) through (39). As in Case 1, this implies $\mathbf{S}_k \succeq \mathbf{O}$.

The identity (49) shows that A_{k+1} is increasing with respect to A_k. Interestingly, although (41) and (49) have distinct forms, for the boundary value $A_k = \frac{\alpha(k+1)(k+2)}{2}$, they evaluate to the same expression (47) and thus arguments from Case 1 readily show that $A_{k+1} > \ell_{k+1}$. On the other hand, we have

$$u_{k+1} - A_{k+1}|_{A_k=u_k} = \frac{\alpha^2\left(\left(1+3\alpha-\alpha^2+\alpha^3\right)k+1+8\alpha+\alpha^2+2\alpha^3\right)}{2(1-\alpha^2)\left((1-\alpha)^2k+1-\alpha+2\alpha^2\right)}$$

and the last term is positive for any $\alpha \in (0,1)$, i.e., $A_{k+1} < u_{k+1}$. This completes Case 2.

Proof of the theorem statement. Given that $A_k \in I_k^-$ implies $A_{k+1} \in I_{k+1}^-$ (which has been proved in Case 1), the rest is easy. If we take $A_0 = \ell_0 = \frac{\alpha}{1+\alpha}$, then because $\mathbf{S}_k \succeq \mathbf{O}$ for all $k \geq 0$, we see that V_k is nonincreasing:

$$\frac{\alpha}{1+\alpha}\|\mathbf{z}^0 - \mathbf{z}^\star\|^2 \geq \frac{\alpha}{1+\alpha}\left\|\mathbf{G}(\mathbf{z}^0)\right\|^2 = V_0 \geq \cdots \geq V_k = A_k\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 + (k+1)\left\langle \mathbf{z}^k - \mathbf{z}^0, \mathbf{G}(\mathbf{z}^k)\right\rangle,$$

where the first inequality follows from Lipschitzness of $\mathbf{G}$ (recall that we are assuming that $R = 1$). Also by (35) and (36),

$$A_k \geq \ell_k > \frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)} = \frac{\alpha(k+1)}{2}\frac{(1+\alpha)(k+1)+\alpha}{1+\alpha} > \frac{\alpha(k+1)^2}{2}. \tag{50}$$

Hence, we obtain

$$\begin{aligned}
\frac{\alpha}{1+\alpha}\|\mathbf{z}^0-\mathbf{z}^\star\|^2 \geq V_k &\geq \ell_k \left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 + (k+1)\left\langle \mathbf{z}^k-\mathbf{z}^0, \mathbf{G}(\mathbf{z}^k)\right\rangle \\
&\overset{\text{(a)}}{\geq} \frac{\alpha(k+1)^2}{2}\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 + (k+1)\left\langle \mathbf{z}^\star-\mathbf{z}^0, \mathbf{G}(\mathbf{z}^k)\right\rangle \\
&\overset{\text{(b)}}{\geq} \frac{\alpha(k+1)^2}{2}\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 - (k+1)\left(\frac{1}{\alpha(k+1)}\|\mathbf{z}^\star-\mathbf{z}^0\|^2 + \frac{\alpha(k+1)}{4}\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2\right),
\end{aligned}$$

where (a) follows from (50) and the monotonicity inequality $\langle \mathbf{z}^k-\mathbf{z}^\star, \mathbf{G}(\mathbf{z}^k)\rangle \geq 0$, and (b) follows from Young's inequality. Rearranging terms, we conclude that

$$\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 \leq \frac{4}{\alpha(k+1)^2}\left(\frac{\alpha}{1+\alpha}+\frac{1}{\alpha}\right)\|\mathbf{z}^0-\mathbf{z}^\star\|^2 = \frac{C\|\mathbf{z}^0-\mathbf{z}^\star\|^2}{(k+1)^2},$$

where $C = \frac{4(1+\alpha+\alpha^2)}{\alpha^2(1+\alpha)}$.

Proof of Lemma 5. Direct calculation gives

$$\begin{aligned}
u_k - \frac{\alpha(k+1)(k+2)}{2} &= \frac{\alpha^2(k+2)}{2(1-\alpha)} > 0 \\
\frac{\alpha(k+1)(k+2)}{2} - \ell_k &= \frac{\alpha^2(k+2)}{2(1+\alpha)} > 0,
\end{aligned}$$

showing (35). Next,

$$\ell_k - \frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)} = \frac{\alpha(k+1-\alpha(k+2))}{2(1+\alpha)} \geq 0$$

because $k+1-\alpha(k+2) = (k+2)(\frac{k+1}{k+2}-\alpha) \geq (k+2)(\frac{1}{2}-\alpha) \geq 0$, which shows (36). Similarly, we observe that

$$\begin{aligned}
\frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)} - \frac{\alpha(k+1)^2-\alpha^3k(k+2)}{2(1-\alpha^2)} &= \frac{\alpha^2(k+1-\alpha(k+2))}{2(1-\alpha^2)} \geq 0 \\
\frac{\alpha(k+1)^2-\alpha^3k(k+2)}{2(1-\alpha^2)} - \frac{\alpha(k+1)(k+1-\alpha(k+2))}{2(1-\alpha)} &= \frac{\alpha^2(k+1+\alpha(k+2))}{2(1-\alpha^2)} > 0 \\
\frac{\alpha(k+1)^2-\alpha^3k(k+2)}{2(1-\alpha^2)} - \frac{\alpha^2(k+1)(k+2)}{1+\alpha} &= \frac{\alpha(k+1-\alpha(k+2))^2}{2(1-\alpha^2)} \geq 0 \\
\frac{\alpha^2(k+1)(k+2)}{1+\alpha} - \frac{\alpha^2(k+1)(k+2)+\alpha^3(k+2)^2}{2(1+\alpha)} &= \frac{\alpha^2(k+2)(k+1-\alpha(k+2))}{2(1+\alpha)} \geq 0,
\end{aligned}$$

and each line corresponds to an inequality within (37), (38) and (39).

C. Omitted proofs of Section 3

In this section, we provide a self-contained discussion on the complexity lower bound results for linear operator equations from Nemirovsky (1991; 1992).

C.1. Proof of Theorem 3

The proof of Theorem 3 was essentially completed in the main body of the paper, except the argument regarding translation, (13), and the proof of Lemma 3.

We first provide the precise meaning of the translation invariance that we are to prove. Given a saddle function $\mathbf{L}$ and $\mathbf{z} \in \mathbb{R}^n \times \mathbb{R}^n$, let $\mathbf{z}^\star_{\mathbf{L}}(\mathbf{z})$ be the saddle point of $\mathbf{L}$ nearest to $\mathbf{z}$. For any $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^n$, $k \geq 0$ and $D > 0$, define

$$\mathfrak{T}\left(\mathbf{z}^0; k, D\right) := \left\{ \mathbf{z}^k \;\middle|\; \begin{array}{l} \mathbf{L}(\mathbf{x},\mathbf{y}) = \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} - \mathbf{c}\rangle,\ \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{b},\mathbf{c} \in \mathbb{R}^n,\ \left\|\mathbf{z}^\star_{\mathbf{L}}\left(\mathbf{z}^0\right) - \mathbf{z}^0\right\| \leq D, \\ \mathbf{z}^j = \mathcal{A}(\mathbf{z}^0, \dots, \mathbf{z}^{j-1}; \mathbf{L}),\ j = 1, \dots, k,\ \mathcal{A} \in \mathfrak{A}_{\text{sep}} \end{array} \right\}.$$

We will show that

$$\mathfrak{T}\left(\mathbf{z}^0; k, D\right) = \mathbf{z}^0 + \mathfrak{T}\left(0; k, D\right)$$

holds for any $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^n$.

Let $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0)$ and $\mathbf{L}(\mathbf{x},\mathbf{y}) = \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} - \mathbf{c}\rangle$ be given, and assume that $\|\mathbf{z}^\star_{\mathbf{L}}(\mathbf{z}^0) - \mathbf{z}^0\| \leq D$. Let $\mathbf{b}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}^0$ and $\mathbf{c}_0 = \mathbf{c} - \mathbf{y}^0$. Then

$$\begin{aligned} \nabla_{\mathbf{x}} \mathbf{L}_0(\mathbf{x}^0, \mathbf{y}^0) &= \mathbf{A}^\intercal(\mathbf{y}^0 - \mathbf{c}) = -\mathbf{A}^\intercal \mathbf{c}_0 \\ \nabla_{\mathbf{y}} \mathbf{L}_0(\mathbf{x}^0, \mathbf{y}^0) &= \mathbf{A}\mathbf{x}^0 - \mathbf{b} = -\mathbf{b}_0. \end{aligned}$$

Hence, (11) with $k = 1$ reads as

$$\begin{aligned} \mathbf{x}^1 - \mathbf{x}^0 &\in \operatorname{span}\{\mathbf{A}^\intercal \mathbf{c}_0\} \triangleq \mathcal{X}_1(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \\ \mathbf{y}^1 - \mathbf{y}^0 &\in \operatorname{span}\{\mathbf{b}_0\} \triangleq \mathcal{Y}_1(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0). \end{aligned}$$

This further shows that

$$\begin{aligned} \nabla_{\mathbf{x}} \mathbf{L}_0(\mathbf{x}^1, \mathbf{y}^1) &= \mathbf{A}^\intercal(\mathbf{y}^1 - \mathbf{c}) = \mathbf{A}^\intercal(\mathbf{y}^1 - \mathbf{y}^0) - \mathbf{A}^\intercal \mathbf{c}_0 \in \operatorname{span}\{\mathbf{A}^\intercal \mathbf{b}_0, \mathbf{A}^\intercal \mathbf{c}_0\} \\ \nabla_{\mathbf{y}} \mathbf{L}_0(\mathbf{x}^1, \mathbf{y}^1) &= \mathbf{A}\mathbf{x}^1 - \mathbf{b} = \mathbf{A}(\mathbf{x}^1 - \mathbf{x}^0) - \mathbf{b}_0 \in \operatorname{span}\{\mathbf{A}(\mathbf{A}^\intercal \mathbf{c}_0), \mathbf{b}_0\}, \end{aligned}$$

and (11) with $k = 2$ becomes

$$\begin{aligned} \mathbf{x}^2 - \mathbf{x}^0 &\in \operatorname{span}\{\mathbf{A}^\intercal \mathbf{c}_0, \mathbf{A}^\intercal \mathbf{b}_0\} \triangleq \mathcal{X}_2(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \\ \mathbf{y}^2 - \mathbf{y}^0 &\in \operatorname{span}\{\mathbf{b}_0, \mathbf{A}\mathbf{A}^\intercal \mathbf{c}_0\} \triangleq \mathcal{Y}_2(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0). \end{aligned}$$

As one can see, we have $\mathbf{x}^k - \mathbf{x}^0 \in \mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0)$ and $\mathbf{y}^k - \mathbf{y}^0 \in \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0)$, where we inductively define

$$\begin{aligned} \mathcal{X}_{k+1}(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\{\mathbf{A}^\intercal \mathbf{c}_0\} + \mathbf{A}^\intercal \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \\ \mathcal{Y}_{k+1}(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\{\mathbf{b}_0\} + \mathbf{A}\mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0). \end{aligned}$$

Then it is not difficult to see that for $k \geq 2$,

$$\begin{aligned} \mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\left\{\mathbf{A}^\intercal \mathbf{c}_0, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)\mathbf{c}_0, \dots, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k-1}{2} \rfloor}\mathbf{c}_0\right\} + \operatorname{span}\left\{\mathbf{A}^\intercal \mathbf{b}_0, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)\mathbf{b}_0, \dots, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k}{2} \rfloor - 1}\mathbf{b}_0\right\} \\ \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\left\{\mathbf{b}_0, (\mathbf{A}\mathbf{A}^\intercal)\mathbf{b}_0, \dots, (\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k-1}{2} \rfloor}\mathbf{b}_0\right\} + \operatorname{span}\left\{\mathbf{A}\mathbf{A}^\intercal \mathbf{c}_0, \dots, (\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k}{2} \rfloor}\mathbf{c}_0\right\}. \end{aligned}$$

Now consider $\mathbf{L}_0(\mathbf{x}, \mathbf{y}) := \langle \mathbf{A}\mathbf{x} - \mathbf{b}_0, \mathbf{y} - \mathbf{c}_0\rangle = \left\langle \mathbf{A}(\mathbf{x} + \mathbf{x}^0) - \mathbf{b}, \mathbf{y} + \mathbf{y}^0 - \mathbf{c}\right\rangle$. Because $\mathbf{z}^\star_{\mathbf{L}_0}$ is a saddle point of $\mathbf{L}_0$ if and only if $\mathbf{z}^\star_{\mathbf{L}_0} + \mathbf{z}^0$ is a saddle point of $\mathbf{L}$, we have $\mathbf{z}^\star_{\mathbf{L}_0}(0) = \mathbf{z}^\star_{\mathbf{L}}(\mathbf{z}^0) - \mathbf{z}^0$, and thus $\|\mathbf{z}^\star_{\mathbf{L}_0}(0)\| \leq D$. Therefore, if we let

$$\mathcal{S}(\mathbf{A}; D) \triangleq \left\{ (\tilde{\mathbf{b}}, \tilde{\mathbf{c}}) \in \mathbb{R}^n \times \mathbb{R}^n \;\middle|\; \left\|\mathbf{z}^\star_{\tilde{\mathbf{L}}}(0)\right\| \leq D,\ \text{where } \tilde{\mathbf{L}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{A}\mathbf{x} - \tilde{\mathbf{b}}, \mathbf{y} - \tilde{\mathbf{c}}\rangle \right\},$$

then

$$\mathfrak{T}\left(\mathbf{z}^0; k, D\right) = \bigcup_{\substack{\mathbf{A} \in \mathbb{R}^{n\times n} \\ (\mathbf{b}_0, \mathbf{c}_0) \in \mathcal{S}(\mathbf{A}; D)}} \mathbf{z}^0 + \left(\mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \times \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0)\right).$$

This proves that the translation invariance holds with $\mathfrak{T}(0; k, D) = \bigcup_{\substack{\mathbf{A} \in \mathbb{R}^{n\times n} \\ (\mathbf{b}_0, \mathbf{c}_0) \in \mathcal{S}(\mathbf{A}; D)}} \left(\mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \times \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0)\right)$ and in particular, shows (13).

C.2. Complexity of solving linear operator equations and minimax polynomials

We first make some general observations. Suppose that we are given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, and an integer $k \geq 1$. Then any $\mathbf{x} \in \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b}) = \operatorname{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b}\}$ can be expressed in the form

$$\mathbf{x} = q(\mathbf{A})\mathbf{b}, \quad \text{where } q(t) = q_0 + q_1 t + \cdots + q_{k-1}t^{k-1},$$

for some $q_0, \dots, q_{k-1} \in \mathbb{R}$. Then we can write

$$\mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{b} - \mathbf{A}q(\mathbf{A})\mathbf{b} = (\mathbf{I} - \mathbf{A}q(A))\mathbf{b} = p(\mathbf{A})\mathbf{b}, \tag{51}$$

where $p(t) = 1 - tq(t)$ is a polynomial of degree at most k satisfying $p(0) = 1$. Note that conversely, given any polynomial $\tilde{p}(t)$ with degree $\leq k$ and constant term 1, one can decompose it as $\tilde{p}(t) = 1 - t\tilde{q}(t)$ and recover a polynomial $\tilde{q}$ of degree $\leq k-1$ corresponding to $\mathbf{x}$.

Now suppose further there exists $\mathbf{x}^\star \in \mathbb{R}^n$ such that $\mathbf{b} = \mathbf{A}\mathbf{x}^\star$ and $\|\mathbf{x}^\star\| \leq D$. The symmetric matrix $\mathbf{A}$ has an orthonormal eigenbasis $\mathbf{v}_1, \dots, \mathbf{v}_n$, corresponding to eigenvalues $\lambda_1, \dots, \lambda_n$, so we can write $\mathbf{x}^\star = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ for some $c_1, \dots, c_n \in \mathbb{R}$. Using (51), we obtain

$$\begin{aligned}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = \|p(\mathbf{A})\mathbf{A}\mathbf{x}^\star\|^2 = \left\|\sum_{j=1}^n c_j\mathbf{A}p(\mathbf{A})\mathbf{v}_j\right\|^2 &= \left\|\sum_{j=1}^n c_j\lambda_j p(\lambda_j)\mathbf{v}_j\right\|^2 \\ &= \sum_{j=1}^n c_j^2\lambda_j^2 p(\lambda_j)^2 \leq D^2\left(\max_{j=1,\dots,n} \lambda_j^2 p(\lambda_j)^2\right).\end{aligned} \tag{52}$$

We define the problem class by $\|\mathbf{A}\| \leq R$, which is equivalent to $\lambda_j \in [-R, R]$ for all $j = 1, \dots, n$. Therefore, we consider a method corresponding to a polynomial $q(t)$ such that $p(t) = 1 - tq(t)$ minimizes

$$\max_{\lambda\in[-R,R]} \lambda^2 p(\lambda)^2 = \left(\max_{\lambda\in[-R,R]} |\lambda p(\lambda)|\right)^2.$$

More precisely, if $p_k^\star(t) = 1 - tq_k^\star(t)$ minimizes the last quantity among all $p(t)$ such that $\deg p \leq k$ and $p(0) = 1$, and if we put $\mathbf{x}^k = q_k^\star(\mathbf{A})\mathbf{b}$, then (52) implies

$$\begin{aligned}\left\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\right\|^2 &= \sum_{j=1}^n c_j^2\lambda_j^2\left(p_k^\star(\lambda_j)\right)^2 \leq D^2 M^\star(k, R)^2 \\ M^\star(k, R) &\triangleq \min_{\substack{\deg p\leq k \\ p(0)=1}} \max_{\lambda\in[-R,R]} |\lambda p(\lambda)|,\end{aligned} \tag{53}$$

for all $\mathbf{A}$ whose spectrum belongs to $[-R, R]$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\star$ with $\|\mathbf{x}^\star\| \leq D$. As $p_k^\star$ solves (53), it is called a *minimax polynomial*.

In order to establish Lemma 3, we present a two-fold analysis in the following. First, we compute the quantity (53) by explicitly naming $p_k^\star$ for each $k \geq 1$. (This was given by Nemirovsky (1992), but without a proof.) Then, following the exposition from (Nemirovsky, 1991), we show that there exists an instance of $(\mathbf{A}, \mathbf{b})$ such that

$$\|\mathbf{A}q(\mathbf{A})\mathbf{b} - \mathbf{b}\|^2 \geq D^2 M^\star(k, R)^2$$

holds for any polynomial q of degree $\leq k-1$.

C.3. Proof of Lemma 3

The solutions to (53) are characterized using the *Chebyshev polynomials of first kind*, defined by

$$T_N(\cos\theta) = \cos(N\theta), \quad N \geq 1,$$

or equivalently by $T_N(t) = \cos(N \arccos t)$. If $N = 2d$ for some nonnegative integer d, then T_N is an even polynomial satisfying $T_N(0) = \cos(d\pi) = (-1)^d$. On the other hand, if $N = 2d + 1$, then T_N is an odd polynomial of the form

$$T_{2d+1}(t) = (-1)^d(2d + 1)t + \cdots , \tag{54}$$

which can be shown via induction using the recurrence relation $T_{N+1}(t) = 2tT_N(t) - T_{N-1}(t)$, which follows from the trigonometric identity

$$\cos((N + 1)\theta) + \cos((N - 1)\theta) = 2\cos(N\theta)\cos\theta.$$

Based on arguments from (Nemirovsky, 1992; Mason & Handscomb, 2002), we will show that given $k \geq 1$ and $m := \lfloor \frac{k}{2} \rfloor$,

$$p_k^\star(t) := \frac{(-1)^m}{2m+1}\left(\frac{R}{t}\right)T_{2m+1}\left(\frac{t}{R}\right)$$

solves (53).

The Chebyshev polynomials satisfy the *equioscillation property* which makes them so special: the extrema of T_N within $[-1, 1]$ occur at $t_j = \cos\frac{(N-j)\pi}{N}$ for $j = 0, \dots, N$, and the signs of the extremal values alternate. Indeed, we have $|T_N(t) = \cos(N \arccos t)| \leq 1$ for all $t \in [-1, 1]$, and for each $j = 0, \dots, N$,

$$T_N(t_j) = \cos\left(N\frac{(N-j)\pi}{N}\right) = \cos(N - j)\pi = (-1)^{N-j}.$$

Also, we have $T_N(t_j) = -T_N(t_{j-1})$ for each $j = 1, \dots, n$.

Given $k \geq 1$, we denote by $\mathcal{P}_k$ the collection of all polynomials p of degree $\leq k$ with $p(0) = 1$. Recall that we are to minimize

$$M(p, R) := \max_{\lambda\in[-R,R]} |\lambda\, p(\lambda)| \tag{55}$$

over $p \in \mathcal{P}_k$. If $p \in \mathcal{P}_k$ minimizes (55), then so does $p_{\mathrm{ev}}(t) := \frac{p(t)+p(-t)}{2}$, since for all $\lambda \in [-R, R]$

$$|\lambda p_{\mathrm{ev}}(\lambda)| = |\lambda| \cdot \left|\frac{p(\lambda) + p(-\lambda)}{2}\right| \leq \frac{|\lambda p(\lambda)|}{2} + \frac{|(-\lambda)p(-\lambda)|}{2} \leq \frac{M(p, R)}{2} + \frac{M(p, R)}{2} = M(p, R) \tag{56}$$

holds, which implies that $M(p_{\mathrm{ev}}, R) \leq M(p, R)$.

Observe that $p_k^\star \in \mathcal{P}_k$ due to (54). Next, note that $\lambda p_k^\star(\lambda) = \frac{(-1)^m R}{2m+1}T_{2m+1}(\frac{\lambda}{R})$ has extrema of alternating signs and same magnitude within $[-R, R]$, which occur precisely at $\lambda_j := R\cos\frac{(2m+1-j)\pi}{2m+1}$, where $j = 0, \dots, 2m + 1$. Suppose that $p_k^\star$ is not a minimizer of $M(p, R)$ over $\mathcal{P}_k$, so that there exists $p \in \mathcal{P}_k$ such that

$$|\lambda_j p(\lambda_j)| \leq M(p, R) < M(p_k^\star, R) = |\lambda_j p_k^\star(\lambda_j)| \quad (j = 0, \dots, 2m + 1). \tag{57}$$

Due to (56), by replacing p with p_{ev} if necessary, we may assume that p is even and has degree $\leq 2m$. Since $\lambda_j \neq 0$ for all $j = 0, \dots, 2m + 1$, the condition (57) reduces to $|p(\lambda_j)| < |p_k^\star(\lambda_j)|$.

As p and $p_k^\star$ are both polynomials of degree $\leq 2m$ and constant terms 1, we can write

$$p_k^\star(\lambda) - p(\lambda) = \lambda q(\lambda)$$

for some polynomial q of degree $\leq 2m - 1$. But then $|p(\lambda_j)| = |p_k^\star(\lambda_j) - \lambda_j q(\lambda_j)| < |p_k^\star(\lambda_j)|$, which implies that $p_k^\star(\lambda_j)$ and $\lambda_j q(\lambda_j)$ have same signs for $j = 0, \dots, 2m + 1$. Now, because $p_k^\star(\lambda_j)$ have alternating signs and

$$\lambda_0 < \cdots < \lambda_m < 0 < \lambda_{m+1} < \cdots < \lambda_{2m+1},$$

we see that the signs of $q(\lambda_j)$ alternate over $j = 0, \dots, m$ and over $j = m + 1, \dots, 2m + 1$, respectively. Therefore, q must have at least one zero in each open interval $(\lambda_j, \lambda_{j+1})$ for $j = 0, \dots, m - 1, m + 1, \dots, 2m$. This implies that $q(t) \equiv 0$ since $\deg q \leq 2m - 1$, while q has at least $2m$ zeros. Therefore, we arrive at $p_k^\star = p$, which is a contradiction.

We have established that

$$M^\star(k,R) = M(p_k^\star, R) = |\lambda_j p_k^\star(\lambda_j)| = \frac{R}{2m+1} = \frac{R}{2\lfloor k/2 \rfloor + 1} \quad (j = 0, \dots, 2m+1). \tag{58}$$

Furthermore, the above arguments show that the minimization of (55) over $p \in \mathcal{P}_k$ is in fact the same as the minimization of

$$\max_{j=0,\dots,2m+1} |\lambda_j p(\lambda_j)| = \max_{\lambda \in \Lambda} |\lambda p(\lambda)|, \quad \Lambda := \{\lambda_0, \lambda_1, \dots, \lambda_{2m+1}\}. \tag{59}$$

Note that the trick of replacing p by p_{ev} is still applicable to (59), but only because the set Λ is symmetric with respect to the origin. Now we can write

$$M^\star(k,R)^2 = \left(\min_{p\in\mathcal{P}_k} \max_{\lambda\in[-R,R]} |\lambda p(\lambda)| \right)^2 = \left(\min_{p\in\mathcal{P}_k} \max_{\lambda\in\Lambda} |\lambda p(\lambda)| \right)^2 = \min_{p\in\mathcal{P}_k} \max_{\lambda\in\Lambda} \lambda^2 p(\lambda)^2, \tag{60}$$

and the final problem from the line (60) is equivalent to

$$\begin{array}{ll} \underset{\nu\in\mathbb{R},\, p\in\mathcal{P}_k}{\text{minimize}} & \nu \\ \text{subject to} & \lambda_j^2 p(\lambda_j)^2 \le \nu, \quad j = 0, \dots, 2m+1. \end{array} \tag{61}$$

We can identify any $p(t) = 1 + p_1 t + \cdots + p_k t^k \in \mathcal{P}_k$ as the vector $(p_1, \dots, p_k) \in \mathbb{R}^k$. Under this identification, (61) is a second order cone program (as the constraints are convex quadratic in $p_1, \dots, p_k$), and Slater's constraint qualification is clearly satisfied. Hence $M^\star(k,R)^2$ equals the optimal value of the dual problem

$$\begin{array}{ll} \underset{\boldsymbol{\mu}\in\mathbb{R}^{2m+2}}{\text{maximize}}\ \underset{p\in\mathcal{P}_k}{\text{minimize}} & \sum_{j=0}^{2m+1} \mu_j \lambda_j^2 p(\lambda_j)^2 \\ \text{subject to} & \sum_{j=0}^{2m+1} \mu_j = 1, \\ & \boldsymbol{\mu} \ge 0. \end{array} \tag{62}$$

Let $\boldsymbol{\mu}^\star = (\mu_0^\star, \dots, \mu_{2m+1}^\star)$ be the dual optimal solution to (62). Provided that $n \ge k+2 \ge 2m+2$, we can take standard basis vectors (with 0-indexing) $\mathbf{e}_0, \dots, \mathbf{e}_{2m+1} \in \mathbb{R}^n$. Define $\mathbf{A}$ by

$$\mathbf{A}\mathbf{e}_j = \lambda_j \mathbf{e}_j \quad (j = 0, \dots, 2m+1), \quad \mathbf{A}\mathbf{v} = 0 \quad (\mathbf{v} \perp \operatorname{span}\{\mathbf{e}_0, \dots, \mathbf{e}_{2m+1}\})$$

and let

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\star, \quad \mathbf{x}^\star = D \sum_{j=0}^{2m+1} \left(\mu_j^\star\right)^{1/2} \mathbf{e}_j$$

so that $\|\mathbf{x}^\star\| = D$. For any given $\mathbf{x} = q(\mathbf{A})\mathbf{b}$ with $\deg q \le k-1$, we use (52) to rewrite $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ as

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 \left(1 - \lambda_j q(\lambda_j)\right)^2 = D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p(\lambda_j)^2,$$

where $p(t) = 1 - tq(t) \in \mathcal{P}_k$. But since $(p_k^\star, \boldsymbol{\mu}^\star)$ is the primal-dual solution pair to the problems (61) and (62), $p_k^\star$ minimizes $\sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p(\lambda_j)^2$ within $\mathcal{P}_k$. Therefore,

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p(\lambda_j)^2 \ge D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p_k^\star(\lambda_j)^2 = D^2 M^\star(k,R)^2 = \frac{R^2 D^2}{2(\lfloor k/2 \rfloor + 1)^2},$$

which establishes (14).

C.4. Proof of Lemma 4

Let $k \geq 0$ be a given (fixed) integer. Consider the polynomial $p_k^\star$ we defined in the previous section. It is an even polynomial of degree $2\lfloor \frac{k}{2} \rfloor$, and thus $p_k^\star\left(\sqrt{t}\right)$ is a polynomial in t of degree $\lfloor \frac{k}{2} \rfloor$, whose constant term is $p_k^\star(0) = 1$. Therefore, we can write $p_k^\star\left(\sqrt{t}\right) = 1 - tq_k(t)$ for some polynomial q_k. We will show that

$$\mathbf{z}^k = q_k\left(\mathbf{B}^\intercal \mathbf{B}\right) \mathbf{B}^\intercal \mathbf{v} \tag{63}$$

satisfies $\|\mathbf{B}\mathbf{z}^k - \mathbf{v}\|^2 \leq \frac{R^2 D^2}{2(\lfloor k/2 \rfloor + 1)^2}$ for any (possibly non-symmetric) $\mathbf{B} \in \mathbb{R}^{m \times m}$ and $\mathbf{v} = \mathbf{B}\mathbf{z}^\star$ satisfying $\|\mathbf{B}\| \leq R$ and $\|\mathbf{z}^\star\| \leq D$. The equation (63) defines an algorithm within the class $\mathfrak{A}_{\text{lin}}$, as q_k is of degree $\lfloor \frac{k}{2} \rfloor - 1$, so that $\mathbf{z}^k$ is determined by $2\lfloor \frac{k}{2} \rfloor - 1 \leq k - 1$ queries to the matrix multiplication oracle.

We proceed via arguments similar to derivations in C.2. First, observe that

$$\left\|\mathbf{B}\mathbf{z}^k - \mathbf{v}\right\|^2 = \left\|\mathbf{B}\mathbf{z}^k - \mathbf{B}\mathbf{z}^\star\right\|^2 = (\mathbf{z}^k - \mathbf{z}^\star)^\intercal \mathbf{B}^\intercal \mathbf{B}(\mathbf{z}^k - \mathbf{z}^\star) = (\mathbf{z}^k - \mathbf{z}^\star)^\intercal |\mathbf{B}|^2 (\mathbf{z}^k - \mathbf{z}^\star) = \left\||\mathbf{B}|\mathbf{z}^k - |\mathbf{B}|\mathbf{z}^\star\right\|^2, \tag{64}$$

where $|\mathbf{B}|$ is the matrix square root of the positive semidefinite matrix $\mathbf{B}^\intercal \mathbf{B}$. Rewriting (63) in terms of $|\mathbf{B}|$, we obtain

$$\mathbf{z}^k = q_k\left(\mathbf{B}^\intercal \mathbf{B}\right) \mathbf{B}^\intercal \mathbf{B}\mathbf{z}^\star = q_k\left(|\mathbf{B}|^2\right) |\mathbf{B}|^2 \mathbf{z}^\star.$$

Plugging the last equation into (64) gives

$$\left\||\mathbf{B}|\mathbf{z}^\star - |\mathbf{B}|\mathbf{z}^k\right\|^2 = \left\|\left(\mathbf{I} - |\mathbf{B}|^2 q_k\left(|\mathbf{B}|^2\right)\right) |\mathbf{B}|\mathbf{z}^\star\right\|^2 = \left\|p_k^\star\left(|\mathbf{B}|\right) |\mathbf{B}|\mathbf{z}^\star\right\|^2.$$

Finally, because $|\mathbf{B}|$ is a symmetric matrix whose eigenvalues are within $[0, R]$, we can apply (52) with $|\mathbf{B}|, \mathbf{z}^\star$ in places of $\mathbf{A}, \mathbf{x}^\star$, and use (58) to conclude that

$$\left\||\mathbf{B}|\mathbf{z}^\star - |\mathbf{B}|\mathbf{z}^k\right\|^2 \leq D^2 \left(\max_{\lambda \in [0,R]} \lambda^2 p_k^\star(\lambda)^2\right) \leq D^2 \left(\max_{\lambda \in [-R,R]} \lambda^2 p_k^\star(\lambda)^2\right) = \frac{R^2 D^2}{(2\lfloor k/2 \rfloor + 1)^2}.$$

C.5. Proof of Theorem 4

We first describe the general class $\mathfrak{A}$ of algorithms without the linear span assumption. An algorithm $\mathcal{A}$ within $\mathfrak{A}$ is a sequence of deterministic functions $\mathcal{A}_1, \mathcal{A}_2, \ldots$, each of which having the form

$$(\mathbf{z}^i, \overline{\mathbf{z}}^i) = \mathcal{A}_i\left(\mathbf{z}^0, \mathcal{O}(\mathbf{z}^0; \mathbf{L}), \ldots, \mathcal{O}(\mathbf{z}^{i-1}; \mathbf{L}); \mathbf{L}\right)$$

for $i \geq 1$, where $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^n \times \mathbb{R}^m$ is an initial point and $\mathcal{O} \colon (\mathbb{R}^n \times \mathbb{R}^m) \times \mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^m) \to \mathbb{R}^n \times \mathbb{R}^m$ is the gradient oracle defined as

$$\mathcal{O}((\mathbf{x}, \mathbf{y}); \mathbf{L}) = \left(\nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}, \mathbf{y})\right).$$

The sequence $\{\mathbf{z}^i\}_{i \geq 0}$ are the *inquiry points*, and $\{\overline{\mathbf{z}}^i\}_{i \geq 0}$ are the *approximate solutions* produced by $\mathcal{A}$. When $k \geq 1$ is the predefined maximum number of iterations, then we assume $\overline{\mathbf{z}}^k = \mathbf{z}^k$ without loss of generality. Similar definitions for deterministic algorithms have been considered in (Nemirovsky, 1991; Ouyang & Xu, 2021).

To clarify, given $\mathbf{L} \in \mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^m)$, an algorithm $\mathcal{A}$ uses only the previous oracle information to choose the next inquiry point and approximate solution. Therefore, if $\mathcal{O}(\mathbf{z}^i; \mathbf{L}_1) = \mathcal{O}(\mathbf{z}^i; \mathbf{L}_2)$ for all $i = 0, \ldots, k-1$, then the algorithm output $(\mathbf{z}^k, \overline{\mathbf{z}}^k)$ for the two functions will coincide, even if $\mathbf{L}_1 \neq \mathbf{L}_2$. In that sense, $\mathcal{A}$ is *deterministic*, *black-box*, and *gradient-based*.

Now we precisely restate Theorem 4.

Theorem 4. *Let $k \geq 1$ and $n \geq 3k + 2$. Let $\mathcal{A} \in \mathfrak{A}$ be a deterministic black-box gradient-based algorithm for solving convex-concave minimax problems on $\mathbb{R}^n \times \mathbb{R}^n$. Then for any initial point $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^n$, there exists $\mathbf{L} \in \mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)$ with a saddle point $\mathbf{z}^\star$, for which $\mathbf{z}^k$, the k-th iterate produced by $\mathcal{A}$, satisfies*

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\| \geq \frac{\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(2\lfloor k/2 \rfloor + 1)^2}.$$

Proof. Let $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^n \times \mathbb{R}^n$ be given. Take $\mathbf{A}$ and $\mathbf{b}$ as in Lemma 3. Denote by $\mathbf{x}^{\text{min}}$ the minimum norm solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$. Recall the construction of $\mathbf{A}$ and $\mathbf{b}$, where $\mathcal{R}(\mathbf{A}) = \operatorname{span}\{\mathbf{e}_0, \dots, \mathbf{e}_{2m+1}\} \perp \ker(\mathbf{A})$. Define

$$\mathbf{L}_0(\mathbf{x}^0, \mathbf{y}^0) = -\mathbf{b}^\intercal(\mathbf{x} - \mathbf{x}^0) + (\mathbf{x} - \mathbf{x}^0)^\intercal \mathbf{A}(\mathbf{y} - \mathbf{y}^0) - \mathbf{b}^\intercal(\mathbf{y} - \mathbf{y}^0).$$

Then $(\nabla_{\mathbf{x}} \mathbf{L}_0(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} \mathbf{L}_0(\mathbf{x}, \mathbf{y})) = \big(\mathbf{A}(\mathbf{y} - \mathbf{y}^0) - \mathbf{b}, \mathbf{A}(\mathbf{x} - \mathbf{x}^0) - \mathbf{b}\big)$, and $\mathbf{z}^0 + \big(\mathbf{x}^{\text{min}}, \mathbf{x}^{\text{min}}\big)$ is a saddle point of $\mathbf{L}_0$.

We follow the oracle-resisting proof strategy of Nemirovsky (1991), described as follows. For each $i = 1, \dots, k$, we inductively define a *rotated* biaffine function

$$\mathbf{L}_i(\mathbf{x}^0, \mathbf{y}^0) = -\mathbf{b}^\intercal(\mathbf{x} - \mathbf{x}^0) + (\mathbf{x} - \mathbf{x}^0)^\intercal \mathbf{A}_i(\mathbf{y} - \mathbf{y}^0) - \mathbf{b}^\intercal(\mathbf{y} - \mathbf{y}^0),$$

where $\mathbf{A}_i = \mathbf{U}_i \mathbf{A} \mathbf{U}_i^\intercal$ for an orthogonal matrix $\mathbf{U}_i \in \mathbb{R}^{n \times n}$. We will show that U_i can be chosen to satisfy $\mathbf{U}_i \mathbf{b} = \mathbf{b}$,

$$\mathcal{O}(\mathbf{z}^j; \mathbf{L}_i) = \mathcal{O}(\mathbf{z}^j; \mathbf{L}_{i-1}) \tag{65}$$

for $j = 0, \dots, i-1$, and

$$\mathbf{x}^j - \mathbf{x}^0, \mathbf{y}^j - \mathbf{y}^0 \in \mathcal{K}_{j-1}(\mathbf{A}_i; \mathbf{b}) \oplus \mathbf{U}_i \mathcal{N}_i = \mathbf{U}_i \mathcal{K}_{j-1}(\mathbf{A}; \mathbf{b}) \oplus \mathbf{U}_i \mathcal{N}_i \tag{66}$$

for $j = 0, \dots, i$, where $\mathcal{N}_i$ is a subspace of $\ker(\mathbf{A})$ such that $\dim(\mathcal{N}_i) \le 2i$. Note that (65) implies that the algorithm iterates $(\mathbf{z}^j, \bar{\mathbf{z}}^j)$ for $j = 1, \dots, i$ do not change when $\mathbf{L}_{i-1}$ is replaced by $\mathbf{L}_i$. Hence, this process sequentially adjusts the objective function $\mathbf{L}$ upon observing an iterate $\mathbf{z}^i$ to resist the algorithm from optimizing it efficiently. Indeed, if (66) holds with $i = j = k$, then

$$\begin{aligned}
\mathbf{x}^k - \mathbf{x}^0 &= \mathbf{U}_k q_{\mathbf{x}}(\mathbf{A})\mathbf{b} + \mathbf{U}_k \mathbf{v}_{\mathbf{x}}^k \\
\mathbf{y}^k - \mathbf{y}^0 &= \mathbf{U}_k q_{\mathbf{y}}(\mathbf{A})\mathbf{b} + \mathbf{U}_k \mathbf{v}_{\mathbf{y}}^k
\end{aligned}$$

for some polynomials $q_{\mathbf{x}}, q_{\mathbf{y}}$ of degree $\le k-1$ and $\mathbf{v}_{\mathbf{x}}^k, \mathbf{v}_{\mathbf{y}}^k \in \mathcal{N}_i \subseteq \ker(\mathbf{A})$. Thus

$$\nabla_{\mathbf{x}} \mathbf{L}_k(\mathbf{x}^k, \mathbf{y}^k) = \mathbf{A}_k(\mathbf{y}^k - \mathbf{y}^0) - \mathbf{b} = \mathbf{U}_k \mathbf{A} \mathbf{U}_k^\intercal \left(\mathbf{U}_k q_y(\mathbf{A})\mathbf{b} + \mathbf{U}_k \mathbf{v}_{\mathbf{y}}^k\right) - \mathbf{b} = \mathbf{U}_k \left(\mathbf{A} q_y(\mathbf{A}) - \mathbf{I}\right) \mathbf{b}$$

and similarly

$$\nabla_{\mathbf{y}} \mathbf{L}_k(\mathbf{x}^k, \mathbf{y}^k) = \mathbf{U}_k \left(\mathbf{A} q_x(\mathbf{A}) - \mathbf{I}\right) \mathbf{b},$$

showing that

$$\|\nabla \mathbf{L}_k(\mathbf{z}^k)\|^2 = \|\mathbf{U}_k \left(\mathbf{A} q_y(\mathbf{A}) - \mathbf{I}\right) \mathbf{b}\|^2 + \|\mathbf{U}_k \left(\mathbf{A} q_x(\mathbf{A}) - \mathbf{I}\right) \mathbf{b}\|^2 \ge \frac{2\|\mathbf{x}^{\text{min}}\|^2}{(2\lfloor k/2 \rfloor + 1)^2}.$$

Then the theorem statement follows from the fact that $\mathbf{z}^\star = \mathbf{z}^0 + (\mathbf{U}_k \mathbf{x}^{\text{min}}, \mathbf{U}_k \mathbf{x}^{\text{min}})$ is a saddle point of $\mathbf{L}_k$.

It remains to provide an inductive scheme for choosing $\mathbf{U}_i$. We set $\mathbf{U}_0 = \mathbf{I}$ (so that $\mathbf{A}_0 = \mathbf{A}$), $\mathcal{N}_0 = \{0\}$, and define $\mathcal{K}_{-1}(\mathbf{A}; \mathbf{b}) = \{0\}$ for convenience. Let $1 \le i \le k$, and suppose that we already have an orthogonal matrix $\mathbf{U}_{i-1}$ and $\mathcal{N}_{i-1} \subseteq \ker(\mathbf{A})$ for which $\mathbf{U}_{i-1}\mathbf{b} = \mathbf{b}$, $\dim(\mathcal{N}_{i-1}) \le 2i - 2$, and (66) holds with $i - 1$ (which is vacuously true when $i = 1$). Let

$$(\mathbf{z}^i, \bar{\mathbf{z}}^i) = \mathcal{A}_i\left(\mathbf{z}^0, \mathcal{O}(\mathbf{z}^0; \mathbf{L}_{i-1}), \dots, \mathcal{O}(\mathbf{z}^{i-1}; \mathbf{L}_{i-1})\right).$$

We want $\mathbf{U}_i$ (to be defined) to satisfy $\mathbf{s}_{\mathbf{x}}^i, \mathbf{s}_{\mathbf{y}}^i \in \mathbf{U}_i \ker(\mathbf{A})$ while $\mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}) = \mathcal{K}_{i-1}(\mathbf{A}_i; \mathbf{b})$. The latter condition is satisfied if $\mathbf{U}_i = \mathbf{Q}_i \mathbf{U}_{i-1}$ for some orthogonal matrix $\mathbf{Q}_i$ which preserves every element within

$$\mathcal{J}_{i-1} = \mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}) \oplus \mathbf{U}_{i-1} \mathcal{N}_{i-1},$$

because then it follows that $\mathbf{U}_i \mathbf{b} = \mathbf{Q}_i \mathbf{U}_{i-1} \mathbf{b} = \mathbf{Q}_i \mathbf{b} = \mathbf{b}$ and

$$\mathcal{K}_{i-1}(\mathbf{A}_i; \mathbf{b}) = \mathbf{U}_i \mathcal{K}_{i-1}(\mathbf{A}; \mathbf{b}) = \mathbf{Q}_i \mathbf{U}_{i-1} \mathcal{K}_{i-1}(\mathbf{A}; \mathbf{b}) = \mathbf{Q}_i \mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}) = \mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}).$$

Consider the decomposition

$$\begin{aligned}
\mathbf{x}^i - \mathbf{x}^0 &= \Pi_{\mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b})}(\mathbf{x}^i - \mathbf{x}^0) + \mathbf{U}_{i-1} \mathbf{r}_{\mathbf{x}}^i + \mathbf{s}_{\mathbf{x}}^i \\
\mathbf{y}^i - \mathbf{y}^0 &= \Pi_{\mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b})}(\mathbf{y}^i - \mathbf{y}^0) + \mathbf{U}_{i-1} \mathbf{r}_{\mathbf{y}}^i + \mathbf{s}_{\mathbf{y}}^i
\end{aligned}$$

where Π denotes the orthogonal projection, $\mathbf{r}^i_{\mathbf{x}}, \mathbf{r}^i_{\mathbf{y}} \in \mathcal{N}_{i-1}$ and $\mathbf{s}^i_{\mathbf{x}}, \mathbf{s}^i_{\mathbf{y}} \in \mathcal{J}^{\perp}_{i-1}$. Since $\dim\ker(\mathbf{A}) = n-2m-2 \geq n-k-2$ and $\dim\left(\mathcal{N}_{i-1}\right)^{\perp} \geq n-(2i-2) \geq n-2k+2$, we have

$$\dim\left(\ker(\mathbf{A}) \cap (\mathcal{N}_{i-1})^{\perp}\right) \geq n-3k \geq 2,$$

so there exist $\tilde{\mathbf{s}}^i_{\mathbf{x}}, \tilde{\mathbf{s}}^i_{\mathbf{y}} \in \ker(\mathbf{A}) \cap (\mathcal{N}_{i-1})^{\perp}$ such that $\|\tilde{\mathbf{s}}^i_{\mathbf{x}}\| = \|\mathbf{s}^i_{\mathbf{x}}\|$, $\|\tilde{\mathbf{s}}^i_{\mathbf{y}}\| = \|\mathbf{s}^i_{\mathbf{y}}\|$, and $\langle \tilde{\mathbf{s}}^i_{\mathbf{x}}, \tilde{\mathbf{s}}^i_{\mathbf{y}} \rangle = \langle \mathbf{s}^i_{\mathbf{x}}, \mathbf{s}^i_{\mathbf{y}} \rangle$. Also, because $\ker(\mathbf{A}) \perp \mathcal{K}_{i-1}(\mathbf{A};\mathbf{b})$,

$$\mathcal{J}_{i-1} = \mathbf{U}_{i-1}\left(\mathcal{K}_{i-1}(\mathbf{A};\mathbf{b}) + \mathcal{N}_{i-1}\right) \perp \mathbf{U}_{i-1}\left(\ker(\mathbf{A}) \cap (\mathcal{N}_{i-1})^{\perp}\right).$$

This implies that there exists an orthogonal $\mathbf{Q}_i \in \mathbb{R}^{n \times n}$ satisfying

$$\begin{aligned} \mathbf{Q}_i\big|_{\mathcal{J}_{i-1}} &= \mathrm{Id}_{\mathcal{J}_{i-1}} \\ \mathbf{Q}_i\left(\mathbf{U}_{i-1}\tilde{\mathbf{s}}^i_{\mathbf{x}}\right) &= \mathbf{s}^i_{\mathbf{x}} \\ \mathbf{Q}_i\left(\mathbf{U}_{i-1}\tilde{\mathbf{s}}^i_{\mathbf{y}}\right) &= \mathbf{s}^i_{\mathbf{y}}. \end{aligned}$$

Now let $\mathbf{v}^i_{\mathbf{x}} = \mathbf{r}^i_{\mathbf{x}} + \tilde{\mathbf{s}}^i_{\mathbf{x}} \in \ker(\mathbf{A})$, $\mathbf{v}^i_{\mathbf{y}} = \mathbf{r}^i_{\mathbf{y}} + \tilde{\mathbf{s}}^i_{\mathbf{y}} \in \ker(\mathbf{A})$, and

$$\begin{aligned} \mathbf{U}_i &\triangleq \mathbf{Q}_i\mathbf{U}_{i-1} \\ \mathcal{N}_i &\triangleq \mathcal{N}_{i-1} + \operatorname{span}\{\mathbf{v}^i_{\mathbf{x}}, \mathbf{v}^i_{\mathbf{y}}\}. \end{aligned}$$

Then clearly $\mathbf{U}_i\mathbf{b} = \mathbf{b}$, $\mathcal{N}_i \subseteq \ker(\mathbf{A})$, and $\dim\mathcal{N}_i \leq 2i$. Next, for each $j = 0, \ldots, i-1$, we have

$$\mathbf{x}^j - \mathbf{x}^0, \mathbf{y}^j - \mathbf{y}^0 \in \mathcal{K}_{j-1}(\mathbf{A}_{i-1};\mathbf{b}) \oplus \mathbf{U}_{i-1}\mathcal{N}_{i-1} \subseteq \mathcal{K}_{j-1}(\mathbf{A}_i;\mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i$$

since $\mathbf{Q}_i$ preserves $\mathcal{J}_{i-1}$ and $\mathcal{N}_{i-1} \subseteq \mathcal{N}_i$. Moreover, because $\mathbf{U}_{i-1}\mathbf{r}^i_{\mathbf{x}} = \mathbf{Q}_i\mathbf{U}_{i-1}\mathbf{r}^i_{\mathbf{x}} = \mathbf{U}_i\mathbf{r}^i_{\mathbf{x}}$ and $\mathbf{s}^i_{\mathbf{x}} = \mathbf{Q}_i\mathbf{U}_{i-1}\tilde{\mathbf{s}}^i_{\mathbf{x}} = \mathbf{U}_i\tilde{\mathbf{s}}^i_{\mathbf{x}}$,

$$\mathbf{x}^i - \mathbf{x}^0 = \Pi_{\mathcal{K}_{i-1}(\mathbf{A}_{i-1};\mathbf{b})}(\mathbf{x}^i - \mathbf{x}^0) + \mathbf{U}_i(\mathbf{r}^i_{\mathbf{x}} + \tilde{\mathbf{s}}^i_{\mathbf{x}}) \in \mathcal{K}_{i-1}(\mathbf{A}_{i-1};\mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i = \mathcal{K}_{i-1}(\mathbf{A}_i;\mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i$$

and similarly $\mathbf{y}^i - \mathbf{y}^0 \in \mathcal{K}_{i-1}(\mathbf{A}_i;\mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i$. This proves (66).

Finally, for $j = 0, \ldots, i-1$,

$$\nabla_{\mathbf{x}}\mathbf{L}_i(\mathbf{x}^j, \mathbf{y}^j) = \mathbf{A}_i(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b} = \mathbf{Q}_i\mathbf{A}_{i-1}\mathbf{Q}_i^{\intercal}(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b}.$$

But $\mathbf{Q}_i^{\intercal}(\mathbf{y}^j - \mathbf{y}^0) = \mathbf{y}^j - \mathbf{y}^0$ because $\mathbf{y}^j - \mathbf{y}^0 \in \mathcal{K}_{j-1}(\mathbf{A}_{i-1};\mathbf{b}) \oplus \mathbf{U}_{i-1}\mathcal{N}_{i-1} \subseteq \mathcal{J}_{i-1}$, and

$$\mathbf{A}_{i-1}(\mathbf{y}^j - \mathbf{y}^0) \in \mathbf{A}_{i-1}\mathcal{K}_{j-1}(\mathbf{A}_{i-1};\mathbf{b}) \oplus \mathbf{A}_{i-1}\mathbf{U}_{i-1}\mathcal{N}_{i-1} = \mathcal{K}_j(\mathbf{A}_{i-1};\mathbf{b}) \subseteq \mathcal{J}_{i-1},$$

which shows that $\nabla_{\mathbf{x}}\mathbf{L}_i(\mathbf{x}^j, \mathbf{y}^j) = \mathbf{Q}_i\mathbf{A}_{i-1}\mathbf{Q}_i^{\intercal}(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b} = \mathbf{A}_{i-1}(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b} = \nabla_{\mathbf{x}}\mathbf{L}_{i-1}(\mathbf{x}^j, \mathbf{y}^j)$. Arguing analogously for the $\mathbf{y}$-variable gives $\nabla_{\mathbf{y}}\mathbf{L}_i(\mathbf{x}^j, \mathbf{y}^j) = \nabla_{\mathbf{y}}\mathbf{L}_{i-1}(\mathbf{x}^j, \mathbf{y}^j)$, proving (65). This completes the induction step, and hence the proof. □

D. Experimental details

D.1. Exact forms of the construction from Ouyang & Xu (2021)

Following Ouyang & Xu (2021), we use

$$\mathbf{A} = \frac{1}{4}\begin{bmatrix} & & & -1 & 1 \\ & & \iddots & \iddots & \\ & -1 & 1 & & \\ -1 & 1 & & & \\ 1 & & & & \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad \mathbf{b} = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \in \mathbb{R}^n, \quad \mathbf{h} = \frac{1}{4}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^n,$$

and $\mathbf{H} = 2\mathbf{A}^{\intercal}\mathbf{A}$. Ouyang & Xu (2021) shows that $\|\mathbf{A}\| \leq \frac{1}{2}$, which implies $\|\mathbf{H}\| \leq \frac{1}{2}$. Therefore (25) is a 1-smooth saddle function.

D.2. Best-iterate gradient norm bound for EG

In Figure 1, we indicated theoretical upper bounds for EG. To clarify, there is no known last-iterate convergence result for EG with respect to $\|\mathbf{G}(\cdot)\|^2$. However, it is straightforward to derive $\mathcal{O}(R^2/k)$ *best-iterate* convergence via standard summability arguments in weak convergence proofs for EG. Although there is no theoretical guarantee that $\|\mathbf{G}(\mathbf{z}^k)\|^2$ will monotonically decrease with EG, in our experiments on both examples, they did monotonically decrease (see Figures 1(a), 1(b)). Therefore, we safely used the best-iterate bounds to visualize the upper bound for EG in Figure 1. For the sake of completeness, we derive the best-iterate bound below.

Lemma 6. *Let $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be an R-smooth convex-concave saddle function with a saddle point $\mathbf{z}^\star$. Let $\mathbf{z} \in \mathbb{R}^n \times \mathbb{R}^m$ and $\alpha \in \left(0, \frac{1}{R}\right)$. Then $\mathbf{w} = \mathbf{z} - \alpha \mathbf{G}(\mathbf{z})$ and $\mathbf{z}^+ = \mathbf{z} - \alpha \mathbf{G}(\mathbf{w})$ satisfy*

$$\|\mathbf{z} - \mathbf{z}^\star\|^2 - \|\mathbf{z}^+ - \mathbf{z}^\star\|^2 \geq (1 - \alpha^2 R^2)\|\mathbf{z} - \mathbf{w}\|^2.$$

Proof.

$$\begin{aligned}
\|\mathbf{z} - \mathbf{z}^\star\|^2 - \|\mathbf{z}^+ - \mathbf{z}^\star\|^2 &= \left(\|\mathbf{z} - \mathbf{w}\|^2 + 2\langle \mathbf{z} - \mathbf{w}, \mathbf{w} - \mathbf{z}^\star\rangle + \|\mathbf{w} - \mathbf{z}^\star\|^2\right) \\
&\quad - \left(\|\mathbf{z}^+ - \mathbf{w}\|^2 + 2\langle \mathbf{z}^+ - \mathbf{w}, \mathbf{w} - \mathbf{z}^\star\rangle + \|\mathbf{w} - \mathbf{z}^\star\|^2\right) \\
&= \|\mathbf{z} - \mathbf{w}\|^2 - \|\mathbf{z}^+ - \mathbf{w}\|^2 + 2\langle \mathbf{z} - \mathbf{z}^+, \mathbf{w} - \mathbf{z}^\star\rangle \\
&\geq \|\mathbf{z} - \mathbf{w}\|^2 - \|\mathbf{z}^+ - \mathbf{w}\|^2.
\end{aligned}$$

The last inequality is just monotonicity: $\langle \mathbf{z} - \mathbf{z}^+, \mathbf{w} - \mathbf{z}^\star\rangle = \alpha\langle \mathbf{G}(\mathbf{w}), \mathbf{w} - \mathbf{z}^\star\rangle \geq 0$. Now the conclusion follows from

$$\|\mathbf{z}^+ - \mathbf{w}\|^2 = \|(\mathbf{z} - \alpha\mathbf{G}(\mathbf{w})) - (\mathbf{z} - \alpha\mathbf{G}(\mathbf{z}))\|^2 = \alpha^2\|\mathbf{G}(\mathbf{z}) - \mathbf{G}(\mathbf{w})\|^2 \leq \alpha^2 R^2 \|\mathbf{z} - \mathbf{w}\|^2,$$

where the last inequality follows from R-Lipschitzness of $\mathbf{G}$. □

Now fix an integer $k \geq 0$, and consider the EG iterations

$$\begin{aligned}
\mathbf{z}^{i+1/2} &= \mathbf{z}^i - \alpha\mathbf{G}(\mathbf{z}^i) \\
\mathbf{z}^{i+1} &= \mathbf{z}^i - \alpha\mathbf{G}(\mathbf{z}^{i+1/2})
\end{aligned}$$

for $i = 0, \dots, k$. Applying Lemma 6 with $\mathbf{z} = \mathbf{z}^i$, $\mathbf{w} = \mathbf{z}^{i+1/2}$ and $\mathbf{z}^+ = \mathbf{z}^{i+1}$, we have

$$\|\mathbf{z}^i - \mathbf{z}^\star\|^2 - \|\mathbf{z}^{i+1} - \mathbf{z}^\star\|^2 \geq (1 - \alpha^2R^2)\|\mathbf{z}^i - \mathbf{z}^{i+1/2}\|^2 = (1 - \alpha^2R^2)\alpha^2\|\mathbf{G}(\mathbf{z}^i)\|^2 \tag{67}$$

for $i = 0, \dots, k$. Summing up the inequalities (67) for all $i = 0, \dots, k$, we obtain

$$\|\mathbf{z}^0 - \mathbf{z}^\star\|^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^\star\|^2 \geq (1 - \alpha^2R^2)\alpha^2 \sum_{i=0}^{k} \|\mathbf{G}(\mathbf{z}^i)\|^2.$$

The left hand side is at most $\|\mathbf{z}^0 - \mathbf{z}^\star\|^2$, while the right hand side is lower bounded by

$$(1 - \alpha^2R^2)\alpha^2\,(k+1) \min_{i=0,\dots,k} \|\mathbf{G}(\mathbf{z}^i)\|^2.$$

Therefore we conclude that

$$\min_{i=0,\dots,k} \|\mathbf{G}(\mathbf{z}^i)\|^2 \leq \frac{C\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k+1}$$

where $C = \frac{1}{\alpha^2(1-\alpha^2R^2)}$.

D.3. ODE flows for $\mathbf{L}(x, y) = xy$

Interestingly, the continuous-time flows with $\mathbf{L}(x, y) = xy$ have exact closed-form solutions.

Note that $\mathbf{G}(x, y) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$. Therefore,

$$\mathbf{G}_\lambda(x, y) = \frac{1}{\lambda}\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & \lambda \\ -\lambda & 1 \end{bmatrix}^{-1}\right)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{\lambda}{1+\lambda^2} & \frac{1}{1+\lambda^2} \\ -\frac{1}{1+\lambda^2} & \frac{\lambda}{1+\lambda^2} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}.$$

The solution to the Moreau–Yosida regularized flow

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\frac{\lambda}{1+\lambda^2} & -\frac{1}{1+\lambda^2} \\ \frac{1}{1+\lambda^2} & -\frac{\lambda}{1+\lambda^2} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

can be obtained with the matrix exponent. The results are

$$\begin{aligned} x(t) &= \exp\left(-\frac{\lambda}{1+\lambda^2}t\right)\left(x^0 \cos\frac{t}{1+\lambda^2} - y^0 \sin\frac{t}{1+\lambda^2}\right) \\ y(t) &= \exp\left(-\frac{\lambda}{1+\lambda^2}t\right)\left(y^0 \cos\frac{t}{1+\lambda^2} + x^0 \sin\frac{t}{1+\lambda^2}\right). \end{aligned}$$

The anchored flow ODE for $\mathbf{L}(x, y) = xy$ is given by

$$\begin{aligned} \dot{x}(t) &= -y(t) + \frac{1}{t}\left(x^0 - x(t)\right) \\ \dot{y}(t) &= x(t) + \frac{1}{t}\left(y^0 - y(t)\right). \end{aligned}$$

From the first equation, we have $\frac{d}{dt}(tx(t)) = t\dot{x}(t) + x(t) = -ty(t) + x^0$, while similar manipulation of the second equation gives $\frac{d}{dt}(ty(t)) = tx(t) + y^0$. Therefore,

$$\begin{aligned} \frac{d^2}{dt^2}(tx(t)) &= -\frac{d}{dt}(ty(t)) = -tx(t) - y^0 \\ \frac{d^2}{dt^2}(ty(t)) &= \frac{d}{dt}(tx(t)) = -ty(t) + x^0, \end{aligned}$$

which gives

$$\begin{aligned} tx(t) &= c_1 \cos t - c_2 \sin t - y^0 \\ ty(t) &= c_1 \sin t + c_2 \cos t + x^0. \end{aligned}$$

Using the initial conditions to determine the coefficients c_1, c_2, we obtain

$$\begin{aligned} x(t) &= \frac{y^0 \cos t + x^0 \sin t - y^0}{t} \\ y(t) &= \frac{y^0 \sin t - x^0 \cos t + x^0}{t}. \end{aligned}$$

E. Connection to CLI lower bounds

In this section, we discuss how EAG relates to the prior work on complexity lower bounds on the class of CLI and SCLI algorithms, introduced and studied in (Arjevani et al., 2016; Arjevani & Shamir, 2016; Azizian et al., 2020; Golowich et al., 2020). Specifically, we show that EAG is not SCLI, so it can break the $\Omega(R^2/k)$ lower bound on squared gradient norm for the 1-SCLI class derived by Golowich et al. (2020). On the other hand, we show that EAG is 2-CLI in the sense of Golowich et al. (2020), and that EAG belongs to an extended class of 1-CLI algorithms.

E.1. Lower bounds for 1-SCLI and non-stationarity of EAG

We start with the notion of 1-SCLI algorithms by Golowich et al. (2020). Consider an algorithm $\mathcal{A}$ for finding saddle points of biaffine functions of the form

$$\mathbf{L}(\mathbf{x},\mathbf{y}) = \mathbf{b}^\intercal\mathbf{x} + \mathbf{x}^\intercal\mathbf{A}\mathbf{y} - \mathbf{c}^\intercal\mathbf{y},$$

where $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$. We say $\mathcal{A}$ is *1-stationary canonical linear iterative (1-SCLI)* if there exist some fixed matrix mappings $\mathbf{C},\mathbf{N} : \mathbb{R}^{2n\times 2n} \to \mathbb{R}^{2n\times 2n}$ such that

$$\mathbf{z}^{k+1} = \mathbf{C}\left(\begin{bmatrix} \mathbf{O} & \mathbf{A} \\ -\mathbf{A}^\intercal & \mathbf{O}\end{bmatrix}\right)\mathbf{z}^k + \mathbf{N}\left(\begin{bmatrix} \mathbf{O} & \mathbf{A} \\ -\mathbf{A}^\intercal & \mathbf{O}\end{bmatrix}\right)\begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix} = \mathbf{C}(\mathbf{B})\mathbf{z}^k + \mathbf{N}(\mathbf{B})\mathbf{v} \tag{68}$$

for $k \geq 0$, where

$$\mathbf{B} = \begin{bmatrix} \mathbf{O} & \mathbf{A} \\ -\mathbf{A}^\intercal & \mathbf{O}\end{bmatrix} \in \mathbb{R}^{2n\times 2n}, \quad \mathbf{v} = \begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix} \in \mathbb{R}^{2n}.$$

Following the convention of Azizian et al. (2020) and Golowich et al. (2020), we also require that $\mathbf{C},\mathbf{N}$ are matrix polynomials. The classical extragradient method (EG) is an 1-SCLI algorithm: with $\mathbf{G}(\mathbf{z}) = \mathbf{B}\mathbf{z} + \mathbf{v}$, we can express EG as

$$\begin{aligned}
\mathbf{z}^{k+1} &= \mathbf{z}^k - \alpha\mathbf{G}\left(\mathbf{z}^k - \alpha\mathbf{G}(\mathbf{z}^k)\right) \\
&= \mathbf{z}^k - \alpha\left(\mathbf{B}\left(\mathbf{z}^k - \alpha\mathbf{B}\mathbf{z}^k - \alpha\mathbf{v}\right) + \mathbf{v}\right) \\
&= \left(\mathbf{I} - \alpha\mathbf{B} + \alpha^2\mathbf{B}^2\right)\mathbf{z}^k - \alpha(\mathbf{I} - \alpha\mathbf{B})\mathbf{v},
\end{aligned}$$

which is of the 1-SCLI form.

A 1-SCLI algorithm $\mathcal{A}$ is *consistent* with respect to an invertible matrix $\mathbf{B}$ if for any $\mathbf{v} \in \mathbb{R}^{2n}$, iterates $\{\mathbf{z}^k\}_{k\geq 0}$ produced by $\mathcal{A}$ satisfy

$$\mathbf{z}^k \to \mathbf{z}^\star = -\mathbf{B}^{-1}\mathbf{v}.$$

If $\mathcal{A}$ is consistent with respect to $\mathbf{B}$, then for any $\mathbf{w} = \mathbf{B}^{-1}\mathbf{v} \in \mathbb{R}^{2n}$, we have

$$\begin{aligned}
-\mathbf{w} = -\mathbf{B}^{-1}\mathbf{v} &= \lim_{k\to\infty} \mathbf{z}^{k+1} \\
&= \lim_{k\to\infty} \mathbf{C}(\mathbf{B})\mathbf{z}^k + \mathbf{N}(\mathbf{B})\mathbf{v} \\
&= \mathbf{C}(\mathbf{B})(-\mathbf{B}^{-1}\mathbf{v}) + \mathbf{N}(\mathbf{B})\mathbf{v} \\
&= \left(-\mathbf{C}(\mathbf{B}) + \mathbf{N}(\mathbf{B})\mathbf{B}\right)\mathbf{w}.
\end{aligned}$$

As this holds for all $\mathbf{w} \in \mathbb{R}^{2n}$, we have the following result.

Lemma 7 (Arjevani et al. (2016))**.** *If a 1-SCLI algorithm $\mathcal{A}$ described by* (68) *is consistent with respect to $\mathbf{B}$, then*

$$\mathbf{I} + \mathbf{N}(\mathbf{B})\mathbf{B} = \mathbf{C}(\mathbf{B}). \tag{69}$$

Indeed, the 1-SCLI formulation of EG satisfies (69).

For the class of consistent 1-SCLI algorithms, Golowich et al. (2020) established $\Omega(1/k)$ a complexity lower bound on squared gradient norm.

Theorem 5 (Golowich et al. (2020))**.** *Let $k \geq 0$ and $n \geq 1$. Then for any consistent 1-SCLI algorithm of the form* (68) *with* $\deg \mathbf{N} = d_{\mathbf{N}}$*, there exist a biaffine function $\mathbf{L}(\mathbf{x},\mathbf{y}) = \mathbf{b}^\intercal\mathbf{x} + \mathbf{x}^\intercal\mathbf{A}\mathbf{y} - \mathbf{c}^\intercal\mathbf{y}$ on $\mathbb{R}^n \times \mathbb{R}^n$ with invertible $\mathbf{A}$, for which*

$$\|\nabla\mathbf{L}(\mathbf{z}^k)\|^2 \geq \frac{R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{20(d_{\mathbf{N}}+1)^2 k} = \Omega\left(\frac{R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k}\right),$$

where $\mathbf{z}^\star$ is the unique saddle point of $\mathbf{L}$.

To clarify, $\deg \mathbf{N}$ refers to the degree of the matrix polynomial defining $\mathbf{N}$. 1-SCLI algorithms with $d_{\mathbf{C}} = \deg \mathbf{C} = 1$ forms a subclass of $\mathfrak{A}_{\text{sim}}$ and $\mathfrak{A}_{\text{sep}}$. (Even if $d_{\mathbf{C}} > 1$, one can still view 1-SCLI algorithms as instances of $\mathfrak{A}_{\text{sim}}$ or $\mathfrak{A}_{\text{sep}}$ by introducing $d_{\mathbf{C}} - 1$ dummy iterates for each 1-SCLI iteration.) However, EAG is an algorithm that belongs to $\mathfrak{A}_{\text{sim}}$ but is not 1-SCLI; if it was, a contradiction would occur, as $\|\nabla\mathbf{L}(\mathbf{z}^k)\|^2 \leq \mathcal{O}(1/k^2)$ for EAG. In fact, it is intuitively clear that EAG is not 1-SCLI; the S in 1-**S**CLI stands for stationary, but EAG has anchoring coefficients $\frac{1}{k+2}$ that vary over iterations.

E.2. Understanding EAG as a CLI algorithm

In this section, we show that EAG algorithms are (non-stationary) 2-CLI, and that we can expand the definition of 1-CLI algorithms to accommodate EAG.

First, we state the definition of m-CLI algorithms introduced by Arjevani & Shamir (2016) adapted to the case of biaffine saddle functions. For $m \geq 1$, an m-CLI algorithm $\mathcal{A}$ takes m initial points $\mathbf{z}_1^0, \dots, \mathbf{z}_m^0$ and at each iteration $k \geq 0$, outputs

$$\mathbf{z}_i^{k+1} = \sum_{j=1}^{m} \mathbf{C}_{ij}^{(k)}(\mathbf{B})\, \mathbf{z}_j^k + \mathbf{N}_i^{(k)}(\mathbf{B})\, \mathbf{v} \tag{70}$$

for $i = 1, \dots, m$, where $\mathbf{C}_{ij}^{(k)}, \mathbf{N}_i^{(k)} : \mathbb{R}^{2n \times 2n} \to \mathbb{R}^{2n \times 2n}$ for $i, j = 1, \dots, m$ are matrix polynomials that depend on k but not on $\{\mathbf{z}_1^k, \dots, \mathbf{z}_m^k\}_{k \geq 0}$. In the case where $\mathbf{C}_{ij}^{(k)} \equiv \mathbf{C}_{ij}$ and $\mathbf{N}_i^{(k)} \equiv \mathbf{N}_i$ for all $i, j = 1, \dots, m$ and $k \geq 0$, we say $\mathcal{A}$ is stationary. Indeed, when $m = 1$, this definition of stationary 1-CLI coincides with that of 1-SCLI given in Section E.1. Also note that the definition (70) includes algorithms that obtain $\mathbf{z}^{k+1}$ with m previous iterates $\mathbf{z}^k, \mathbf{z}^{k-1}, \dots, \mathbf{z}^{k-m+1}$, by letting $\mathbf{z}_i^k = \mathbf{z}^{k+1-i}$ for $i = 1, \dots, m$.

Performance measure	Algorithm class	Lower bound	Best known rate	Order-optimality
Duality gap (Last iterate)	1-SCLI	$\Omega\left(\frac{R}{\sqrt{k}}\right)$ (Golowich et al., 2020)	$\mathcal{O}\left(\frac{R}{\sqrt{k}}\right)$ (Golowich et al., 2020)*	Established*
	1-CLI	$\Omega\left(\frac{R}{k}\right)$ (Nemirovsky (1992), Nemirovski (2004))	$\mathcal{O}\left(\frac{R}{\sqrt{k}}\right)$ (Golowich et al., 2020)*	Unknown
	m-CLI ($m \geq 2$)	$\Omega\left(\frac{R}{k}\right)$ (Nemirovsky (1992), Nemirovski (2004))	$\mathcal{O}\left(\frac{R}{k}\right)$ (Nemirovski (2004), Golowich et al. (2020))	Established
Squared gradient norm (Last iterate)	1-SCLI	$\Omega\left(\frac{R^2}{k}\right)$ (Golowich et al., 2020)	$\mathcal{O}\left(\frac{R^2}{k}\right)$ (Golowich et al., 2020)*	Established*
	1-CLI	$\Omega\left(\frac{R^2}{k^2}\right)$ (Nemirovsky, 1992)	$\mathcal{O}\left(\frac{R^2}{k}\right)$ (Golowich et al., 2020)*	Unknown
	Translated 1-CLI	$\Omega\left(\frac{R^2}{k^2}\right)$ (Nemirovsky, 1992)	$\mathcal{O}\left(\frac{R^2}{k^2}\right)$ (This paper)	Established
	m-CLI ($m \geq 2$)	$\Omega\left(\frac{R^2}{k^2}\right)$ (Nemirovsky, 1992)	$\mathcal{O}\left(\frac{R^2}{k^2}\right)$ (This paper)	Established

Table 1. Lower bounds and best known rates for CLI algorithm classes (* means that the result holds with the additional assumption that the derivative of $\mathbf{G}$ is Lipschitz continuous).

Golowich et al. (2020) showed that the averaged EG iterates, which have rate $\mathcal{O}(1/k)$ on duality gap, can be written in 2-CLI form; hence, the $\Omega(1/\sqrt{k})$ 1-SCLI lower bound on duality gap therein cannot be generalized to m-CLI algorithms for $m \geq 2$. They then posed the open problem of whether the $\Omega(1/\sqrt{k})$ 1-SCLI lower bound on duality gap can be generalized to 1-CLI algorithms. Below, we provide a similar discussion on rates on squared gradient norm.

It is straightforward to see that EAG is 2-CLI; define $\mathbf{z}_2^{k+1} = \mathbf{z}_2^k = \cdots = \mathbf{z}_2^0 = \mathbf{z}^0 = \mathbf{z}_1^0$ for all $k \geq 0$, and

$$\begin{aligned}\mathbf{z}_1^{k+1} &= \mathbf{z}_1^k - \alpha_k \mathbf{G}\left(\mathbf{z}_1^k - \alpha_k \mathbf{G}(\mathbf{z}_1^k) + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}_1^k)\right) + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}_1^k) \\ &= \left(\frac{k+1}{k+2}\mathbf{I} - \frac{k+1}{k+2}\alpha_k \mathbf{B} + \alpha_k^2 \mathbf{B}^2\right)\mathbf{z}_1^k + \frac{1}{k+2}(\mathbf{I} - \alpha_k \mathbf{B})\mathbf{z}_2^k - \alpha_k(\mathbf{I} - \alpha_k \mathbf{B})\mathbf{v}. \end{aligned} \tag{71}$$

For EAG-C, one can alternatively eliminate the dependency on $\mathbf{z}^0$ to define $\mathbf{z}^{k+1}$ in terms of $\mathbf{z}^k$, $\mathbf{z}^{k-1}$, and $\mathbf{v}$; respectively multiply $(k+2)$ and $(k+1)$ to the following identities

$$\mathbf{z}^{k+1} = \left(\frac{k+1}{k+2}\mathbf{I} - \frac{k+1}{k+2}\alpha\mathbf{B} + \alpha^2\mathbf{B}^2\right)\mathbf{z}^k + \frac{1}{k+2}(\mathbf{I}-\alpha\mathbf{B})\mathbf{z}^0 - \alpha(\mathbf{I}-\alpha\mathbf{B})\mathbf{v}$$

$$\mathbf{z}^{k} = \left(\frac{k}{k+1}\mathbf{I} - \frac{k}{k+1}\alpha\mathbf{B} + \alpha^2\mathbf{B}^2\right)\mathbf{z}^{k-1} + \frac{1}{k+1}(\mathbf{I}-\alpha\mathbf{B})\mathbf{z}^0 - \alpha(\mathbf{I}-\alpha\mathbf{B})\mathbf{v}$$

and subtract to eliminate $\mathbf{z}^0$. Since EAG has $\mathcal{O}(1/k^2)$ rate, this reformulation shows that the $\Theta(1/k)$ 1-SCLI lower bound on the squared gradient norm cannot be generalized to 2-CLI algorithms.

Furthermore, EAG also provides a partial resolution, in the negative, of the open problem of whether the $\Theta(1/k)$ 1-SCLI lower bound on the squared gradient norm can be generalized to 1-CLI algorithms. Observe that if we translate the given problem to set $\mathbf{z}^0 = 0$, keeping the sequence $\mathbf{z}_2^k$ is no longer necessary, and (71) reduces to 1-CLI form. Such translation is not allowed in the definition (70), but it is reasonable to consider an expanded class of algorithms that are 1-CLI up to translation. Precisely, define an algorithm $\mathcal{A}$ to be *translated 1-CLI* if it takes the form

$$\mathbf{z}^{k+1} = \mathbf{C}^{(k)}(\mathbf{B})(\mathbf{z}^k) + \mathbf{N}^{(k)}(\mathbf{B})(\mathbf{v})$$

when $\mathbf{z}^0 = 0$, and is *translation invariant* in the sense that

$$\mathbf{z}^k = \mathcal{A}(\mathbf{z}^0, \mathbf{z}^1, \dots, \mathbf{z}^{k-1}; \mathbf{L}) = \mathbf{z}^0 + \mathcal{A}(0, \mathbf{z}^1 - \mathbf{z}^0, \dots, \mathbf{z}^{k-1} - \mathbf{z}^0; \mathbf{L}_{\mathbf{z}^0})$$

when $\mathbf{z}^0 \neq 0$, where $\mathbf{L}_{\mathbf{z}^0}(\mathbf{x}, \mathbf{y}) = \mathbf{L}(\mathbf{x} + \mathbf{x}^0, \mathbf{y} + \mathbf{y}^0)$. That is, the iterates of $\mathcal{A}$ are generated equivalently by starting with $\mathbf{z}^0 = 0$ and applying $\mathcal{A}$ to the translated objective $\mathbf{L}_{\mathbf{z}^0}$. The concept of translated 1-CLI can be viewed as a generalization of consistent 1-SCLI algorithms; observe that we can rewrite (68) as

$$\mathbf{z}^{k+1} - \mathbf{z}^0 = \mathbf{C}(\mathbf{B})(\mathbf{z}^k - \mathbf{z}^0) + \mathbf{N}(\mathbf{B})(\mathbf{B}\mathbf{z}^0 + \mathbf{v}) - (\mathbf{I} + \mathbf{N}(\mathbf{B})\mathbf{B} - \mathbf{C}(\mathbf{B}))\mathbf{z}^0,$$

which shows that a 1-SCLI algorithm is translation invariant if and only if it satisfies the consistency formula (69). Since EAG has $\mathcal{O}(1/k^2)$ rate and is a translated 1-CLI algorithm, our results prove that the $\Theta(1/k)$ 1-SCLI lower bound on the squared gradient norm can be generalized to translated 1-CLI algorithms.