

1 **Reviewer 1:** Thanks for your encouraging comments. We would like to note that even when $L_x = L_{xy} = L_y$, our
2 work still improves the result of Lin et al.'s by reducing the dependence on $\ln(1/\epsilon)$.

3 **Reviewer 3:** Thank you for the detailed comments and the pointers to several related papers.

- 4 1. For a quadratic function $\frac{1}{2}\mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$, querying the gradient oracle returns $\mathbf{A} \mathbf{x} - \mathbf{b}$, i.e. the matrix-vector product
5 (plus a constant). In other words, one can implement a matrix-vector product oracle via the gradient oracle, and vice
6 versa. Thus counting the number of matrix-vector products is indeed equivalent to counting the gradient complexity.
- 7 2. The lower bound in [28] is for the convex-concave setting, while the lower bound in [37] is for the strongly
8 convex-strongly concave setting.
- 9 3. Motivation behind the class of studied problems: the class of strongly convex-strongly concave minimax problems is
10 a fundamental class of minimax problems and has been studied extensively in the literature. Moreover, efficient
11 algorithms for this class can be translated to efficient algorithms for more general strongly convex-concave and
12 convex concave problems (via the reduction developed in [19]).
- 13 4. Thank you for bringing our attention to the paper by Ostrovskii et al., whose algorithm also relies on solving proximal
14 problems in a multi-loop manner. However, there are important differences between our algorithm and theirs. The
15 outer loop of their algorithm updates both \mathbf{x} and \mathbf{y} by solving a proximal point problem (without momentum), while
16 the two inner loops use accelerated gradient method to solve the regularized problem. In comparison, The two outer
17 loops of our algorithm use accelerated proximal method (with momentum) on \mathbf{x} and \mathbf{y} respectively to reduce the
18 problem to solving a regularized problem, and the innermost loop solves this using the alternating best response
19 scheme. The analysis and the results of both papers are also quite different.
- 20 5. The explicit form of the subpolynomial factor in Corollary 3 can be found on line 309 of the SM.
- 21 6. The full description of AGD is deferred to page 3 of the SM due to lack of space. Our final bounds hold for any L_x
22 and L_y . We only assume $L_x = L_y$ for ease of presentation. It is without loss of generality as the more general case
23 $L_x \neq L_y$ can be handled by scaling (see Line 70-73).
- 24 7. **Relation to Prior Work:** Thank you for bringing our attention to the literature on variational inequalities. We will
25 cite and discuss them accordingly in our revision.

26 **Reviewer 4:** Thank you for your detailed review comments.

27 **About correctness of the proof:**

- 28 1) Yes, the minimax theorem holds for non-compact sets in the strongly convex-strongly concave case; see e.g. Hartung,
29 An extension of Sion's minimax theorem with an application to a method for constrained games.
- 30 2) We can prove Theorem 1 from line 53 by multiplying the upper bound of T with the number of gradient evaluations
31 per t , which is given in the algorithm ($2\sqrt{\kappa_x} \ln(24\kappa_x) + 2\sqrt{\kappa_y} \ln(24\kappa_y)$).
- 32 3) $\tilde{\mathbf{x}}_t$ is not used in the proof. We will remove both definitions in our revision as they are redundant.
- 33 4) We meant to cite an earlier version of Lin et al.'s paper (2002.02417v1), where the sequence $\{\Lambda_t\}$ appears in the
34 proof of Lemma B.1 (page 24) as $\{\Lambda_t(\mathbf{x}^*)\}$.
- 35 5) This is a typo; it should be $t = 1$ instead of $t = 0$.

36 We will clarify the above points and fix the typos in the next version.

37 **About comparison to related work:** Thank you for the pointers and we will cite them accordingly in our revision.

- 38 1) We agree that previous work on monotone variational inequalities are very relevant.
- 39 2) Thanks for pointing it out. Indeed, this paper provides the same lower bound as in [37], although it only stated it for
40 a weaker class of algorithms. However, we have verified that this lower bound in fact holds for a larger class of
41 algorithms, and is directly comparable to our upper bounds. We will mention this explicitly in the next version.
- 42 3) These algorithms have suboptimal dependency on the condition number: The bound for GDA and Hamiltonian Gra-
43 dient Descent is $O\left(\left(\frac{L}{m_x} + \frac{L}{m_y}\right)^2 \ln\left(\frac{1}{\epsilon}\right)\right)$, while the bound for ExtraGradient and OGD is $O\left(\left(\frac{L}{m_x} + \frac{L}{m_y}\right) \ln\left(\frac{1}{\epsilon}\right)\right)$.
44 Suppose that $m_x = 1$, $m_y = A^2$, $L = A^4$ ($A \gg 1$), then the two upper bounds become $O(A^8 \ln(1/\epsilon))$ and
45 $O(A^4 \ln(1/\epsilon))$ respectively, while the lower bound is $\Omega(A^3 \ln(1/\epsilon))$.
- 46 4) First, as we understand it, Azizian et al. only provide local convergence guarantees, while we can show global
47 convergence. Second, the spectral shape framework does not fully capture the properties of a minimax optimization
48 problem. E.g., the spectral shape does not reflect both m_x and m_y , only $\min\{m_x, m_y\}$. Consequently, a minimax
49 optimization algorithm that is optimal for its spectral shape could still be suboptimal if m_x and m_y are very different.

50 **About using CG to solve the quadratic problem:** One can indeed solve $\mathbf{J} \mathbf{z} = \mathbf{b}$ by solving the squared equation
51 $\mathbf{J}^T \mathbf{J} \mathbf{z} = \mathbf{J}^T \mathbf{b}$. However the condition number of $\mathbf{J}^T \mathbf{J}$ can be very large, so the resulting complexity bound, which is
52 $\tilde{O}\left(\frac{L}{\min\{m_x, m_y\}}\right)$, has suboptimal dependence on condition numbers and is much worse than our result.