We thank all the reviewers for their constructive comments. We explain the intuition behind DIAG (Algorithm 1) for strongly-convex-concave minimax problems first, which we will add in the final revision.

**Conceptual DIAG:** The intuition behind Algorithm 1 stems from a "conceptual" version of DIAG (also specified in Algorithm 1, Step 4), which is inspired from the conceptual version of Mirror-Prox (MP) (cf. Section 2.2):

1. \( x_{k+1} = \arg\min_{x} g(x, y_{k+1}), \) and \( y_{k+1} = P_{Y}(w_{k} + \frac{1}{\eta} \nabla_{y} g(x_{k+1}, w_{k})) \)
2. \( z_{k+1} = P_{Y}(w_{k} + \eta \nabla_{y} g(x_{k+1}, w_{k})) \)

The main idea is to apply an MP-like update for \( x \) on \( g(\cdot, y_{k+1}) \) and an AGD step for \( y \) on \( g(x_{k+1}, \cdot) \). In the final estimate, we use \( \bar{x}_{k} = (2/(K+1)) \sum_{i=1}^{K} (i \cdot x) \), because MP-like updates give ergodic guarantees, but use \( y_{k} \), because AGD has final iterate guarantees. The MP-like update is crucial in this algorithm so as to inherit the well-known fast convergence rate of AGD for smooth-convex optimization.

**Implementable DIAG:** The above step (b) requires \( g(\cdot, y_{k+1}) \) and \( g(x_{k+1}, \cdot) \) which are not a priori available at the \( k \)-th step. But we can implement this step up to \( \varepsilon_{\text{step}} \) error (step 4, Algorithm 1), using Imp-STEP subroutine (Algorithm 1). Just like the fact that conceptual MP can be realized in \( \log(1/\varepsilon) \) steps (in fact, just two steps suffice), Imp-STEP converges in \( R = \log(\frac{2KL}{\varepsilon_{\text{mp}}}) = O(\log(\frac{1}{\varepsilon_{\text{step}}})) \) steps, because the following mapping is a contraction for small enough stepsize 1/\beta:

\[
y_{k+1} = P_{Y}(w_{k} + (1/\beta) \nabla_{y} g(x_{k+1}, y), w_{k}),
\]

where \( x_{k+1} = \min_{x} g(x, y) \). This follows from (i) the \( L \)-smoothness of \( g \), and (ii) the Lipschitzness of \( x_{k+1} \) in \( y \) (due to strong convexity of \( g(\cdot, y) \)). Further, again by \( \sigma \)-strong-convexity of \( g(\cdot, y) \), \( x_{k+1} = \min_{x} g(x, y) \) could be approximately found in \( O(\sqrt{\frac{1}{L}} \log(\frac{1}{\varepsilon_{\text{step}}})) \) steps. Thus the overall speed of Imp-STEP is \( O(\sqrt{\frac{1}{L}} \log(\frac{1}{\varepsilon_{\text{step}}})) \) steps.

**Response to reviewer 1:** We agree with and will include, the reviewer’s comment, that the non-smoothness of \( f(x) = \max_{y} g(x, y) \), more precisely the non-Lipschitzness of the maximizer of \( g(x, \cdot) \) is the reason why naive AGD is sub-optimal. We will devote more space to explaining the DIAG algorithm and discussing more related works.

1. We will clarify that steps (5) & (6) is the Euclidean version of Mirror-Prox and discuss the extra-gradient method.
2. Criteria in [26] is weaker in the following sense. Consider \( g(x, y) = (x^2 - y^2)/2 \) (\( f(x) = x^2/2 \) and \( h(y) = -y^2/2 \)) with domain \( \mathbb{R} \times [0, 1] \). To reach \( (\hat{x}, \hat{y}) \) s.t. \( \hat{x} = \hat{y} \leq \varepsilon \), DIAG requires \( O(\varepsilon^{-3}) \) steps since \( \nabla f(\hat{x}) = \nabla h(\hat{y}) = \varepsilon \). However, [26] requires \( O((\varepsilon^2)^{-4.5}) = O(\varepsilon^{-7}) \) steps since \( \hat{y}(\hat{x}, \hat{y}) = \max_{y \in [0, 1]} \langle \nabla_{y} g(\hat{x}, y), y - \hat{y} \rangle = \langle -\varepsilon, -\varepsilon \rangle = \varepsilon^2 \). We will add a precise justification (which was omitted due to the lack of space) in the next revision.

3. We refer the reviewer to the above explanation of DIAG algorithm.

4. **Bilinear** coupling: a) we focus on non-linear coupling and in general, bilinear results do not apply to our setting, b) when we specialize our result to standard bilinear coupling setting, our results match the optimal \( 1/K^2 \) rates. Further assumptions like unbounded domain and full-rank coupling matrix give linear convergence rates [R1] (will be cited), but this follows directly from the fact that the Fenchel dual of a smooth function is strongly convex (Theorem 6 of [12]).

5. We will include citations to similar saddle point problems and algorithms, including [R4] and [R5]. However, we again note that none of the suggested (or other) references obtain results similar to ours in the setting that we consider.

**Response to reviewer 3:** We will include numerical experiments; as a preliminary experiment we consider the following min-max problem (P3): \( \min_{x \in \mathbb{R}^d} \max_{y \in [0, 1]} f(x) = \max_{x \in \mathbb{R}^d} f(x) \) with random quadratic functions (hence weakly-convex). In the figure right, we plot the norm of gradient of Moreau envelope \( \| \nabla f_{\text{mp}}(x) \| \) against the number of first-order gradient oracle calls in log-log scale. We see that, Prox-FDIA has a faster convergence rate than subgradient method. We will also include other practical use-cases such as robust learning, multi-task learning, and adversarial training.

**Response to reviewer 4:** We will incorporate all suggestions by the reviewer and clarify all ambiguous/missing explanations in the final version. We discuss important ones below.

- **Chen et al.:** their result only handles bilinear case (also see response to R1, point 4) and gets a rate of \( O(1/\varepsilon) \), but can handle prox-function friendly non-smoothness w.r.t. \( y \). In contrast, we can handle non-linear coupling between \( x \) and \( y \) for bilinear case (with strong convexity w.r.t. \( x \) and smoothness w.r.t. \( y \)) can obtain \( O(1/\sqrt{\varepsilon}) \) rate.
- **-** We assume \( X = \mathbb{R}^p \) since we use [Theorem 6, 12] in the proof, which requires the domain to be the full vector space.
- **-** The sub-routine Imp-STEP has a typo: In Step 10, \( x_{\varepsilon} \) should be \( x_{\varepsilon} \). That is, given \( y_{\varepsilon} \), we compute \( x_{\varepsilon} \) such that \( g(\hat{x}_{\varepsilon_{\varepsilon}}, y_{\varepsilon}) \leq \min_{x} g(x, y_{\varepsilon}) + \varepsilon_{\text{agd}} \) and then Step 11 updates: \( y_{\varepsilon+1} = P_{Y}(w_{\varepsilon} + \frac{1}{\eta} \nabla_{y} g(x_{\varepsilon}, w_{\varepsilon})) \). This gives the new \( (x, y_{\varepsilon+1}) \) pair, and the process is repeated. We refer the reviewer to the explanation of DIAG algorithm at the top.
- **-** In line 196: We meant that \( \min_{x} \max_{y} g(x, y) - \max_{y} \min_{x} g(x, y) \) (which we call the minimum primal dual gap) is unknown for non-convex functions. We will make the statement precise.
- **-** In line 203: We are citing the result of [8], which uses the same convergence criteria as our paper.