We thank the reviewers for their comments, and are very glad to see that all reviewers appreciate the novel and interesting insights that are clearly conveyed by our presentation and simplified derivations (which we believe to carry pedagogical values). They also raise some great points and criticisms, which we respond to below.

[All] More Experiments All reviewers want more experiments, though R1 "understands that the focus of the paper is rather theoretical" and R2 thinks "this is not a major weakness as there are still enough results to demonstrate the ideas presented in the paper". While we have many interesting theoretical results (some of which were deferred to the appendix due to page limit, e.g., App. B, E, the nontrivial proof of Prop.8, etc.) and accompanied them with proof-of-concept experiments, it is hard to deny that more experiments would always help! Here are what we can do:
- We can show how CI changes with $\mathcal{W}$, as suggested by R1.
- We can run the experiments on more domains (probably in the appendix, as we want the main text to focus on theory).
- The minimax objective is difficult to optimize in practice, which is a common and somewhat unsolved issue in this line of work. We have tried many methods to stablize training and documented the working tricks in F.2. In fact, we have also conducted additional experiments on a tabular environment to understand how the behavior of our optimization procedure deviates from an "ideal one" (which we can calculate only in simple environments). We will add these results to the appendix. We find them more insightful than simply running the same experiments on more domains, especially given that the community's understanding in the optimization aspect of these algorithms is still rather immature.

[R2] Abuse of "CI" Agreed. We are considering changing "confidence interval" to "value interval" to avoid this issue.

[R2] Sampling Errors We fully agree that addressing this issue is of high practical significance. In fact, we **do** handle it in the experiments (see below). While we can add related discussions to make the paper conceptually more complete, we also genuinely think that handling sampling errors in a way that is ***simultaneously*** theoretically sound and practically useful is highly nontrivial (we have thought about it for quite a while) and may require an entire follow-up paper (or more) and possibly some breakthroughs in theory.
- [Theory] We can handle sampling errors by adding e.g., Rademacher complexity-based generalization error bounds (similar to Thm 9 of Uehara et al) to the interval, which is straightforward from textbook. It makes the paper's theory more complete, but as we all know, even the tightest generalization bounds are often too loose to be practically useful.
- [Experiments] **We do handle sampling errors in the experiments via bootstrapping** (Fig 3), which is often the most practical option when strict CIs tend to be loose. Even so, there are still many open questions: for example, the difficulty of optimization—which is serious in this setting—could affect the CI's validity. In light of this, maybe we should keep the outer player fixed (e.g., w in $\inf_w \sup_q$), and only bootstrap the inner optimization for a more conservative interval? How does this affect the theoretical properties of bootstrapping? After all, confidence intervals for such minimax statistics (over neural nets!) are not something that classical statistics literature usually handle, which means a lot need to be done in theory. Of course, maybe this is just our ignorance and pointers are always welcomed!
- While we leave sampling error to future work, we expanded the paper's scope by discussing policy optimization under insufficient data (Section 5). This is a highly important yet understudied problem (L228), and the results we derived for OPE were nicely applied to this problem, producing novel insights that are useful for future research and discussions.

[R1] Describe MWL/MQL We have 2 "nutshell" paragraphs in L114 and 166, which compactly summarizes the core ideas of MWL/MQL and foreshadows our derivations. If you think there are other aspects of MWL/MQL that the readers need to know to understand our work, please let us know and we are happy to integrate them into the paper.

[R1] AlgaeDICE The policy evaluation component of Fenchel AlgaeDICE (their Eq.16, with $\alpha = 0$ and without "$\max_\pi$") is precisely our UB_q ($= \text{LB}_w$ with convex classes): their ν is our q, their ζ is our w, and there are some superficial differences due to e.g., normalization conventions; we will explain in detail in the revision.

We don't think MUB-PO can be recovered by negative α. Your guess is probably based on the first line of their Eq.15, $\max_\pi \mathbb{E}_{d^\pi}[r] - \alpha D_f(d^\pi \| d^D)$, so negative α should encourage exploration. However, this expression is only equal to the next line when *unrestricted* function spaces for ζ and ν are used (which they implicitly assumed throughout their derivations), which does not hold in general. Moreover, the α term can be viewed as regularization (see our App.E), but MUB-PO achieves exploration even *without* regularization. The key of MUB-PO is $\inf_w \sup_q$, which differs from $\sup_w \inf_q$ in MLB-PO/AlgaeDICE in a fundamental way. In fact, MUB-PO enjoys an exploration guarantee (Prop.10), but it's unclear if AlgaeDICE with negative α enjoys the same type of guarantees under similar assumptions.

[R1] Validity Ratio Our bad. It's the relative frequency that the groundtruth is contained in the predicted interval.

[R1] Fig 1 First, inclusion of groundtruth is only guaranteed when one of the classes is realizable, data is unlimited, and optimization is exact. Since all conditions only hold approximately in practice, slight exclusion is expected. Now about larger CI with larger $\mathcal{Q}$: when $\mathcal{W}$ is fixed, increasing $\mathcal{Q}$ reduces $\inf_q \sup_w$ and increases $\sup_q \inf_w$, that is, the red bar in Fig 1 should monotonically decrease, and the blue bar increase, which is roughly the case. Therefore, once the CI reversed (Q-net size ≥ 10), increasing $\mathcal{Q}$ would only make the interval larger. The interval would only collapse to a point and stay at 0 length with realizable $\mathcal{W}$ among other idealized assumptions.