Thank you reviewers for your time and thoughtful comments! We respond to each reviewer's comments below.

R1 [Non-convexity issues.] One diagnostic on non-convexity issues is to measure the smoothness of the dose-response curve. We argue that our experiments demonstrate valid bounds are possible in many cases. We stress that many hard problems in ML involve non-convex optimization without theoretical guarantees, but these solutions are still useful.

R1 [Uncertainty intervals for the bounds.] We agree this is important, which is why we mentioned it. There are a lot of possible ways to do this including links to *(Silva & Evans. 2016)* and *(Fong, Lyddon, & Holmes. 2019)* which is why we originally thought to leave this for future work. We would be open to including an experiment if the paper is accepted.

R1 [Choice of functions for p_η.] Sorry for not being clear. We chose the distribution for p_η to be Gaussian purely as a reasonable example. Ultimately, the choice of p_η, $p_\eta(\theta|x, z)$, and f_θ is completely up to the practitioner.

R1 [Intuition on response.] Our recommendation is that if one does not have any prior knowledge about what the response function family is, then the most conservative thing is to use a flexible function family such as an MLP basis. As you point out, this may result in loose bounds as the practitioner provides limited information to constrain the model space. However, if one does have an idea of the function class of the response (e.g., through expert knowledge, prior experiments) then one should parameterize f to constrain the considered models and thus obtain tighter bounds.

R1 [Other non-linear cases.] Thanks, we completely agree with this. We have included an additional non-linear case in the supplement, but would consider including it in the main text, let us know what you think!

R1 [Choice of outcome equations.] Our criteria were (a) the simplest choices for both linear (i.e, 2SLS assumptions satisfied) and non-linear, non-additive cases (i.e., quadratic, interaction term); (b) to keep variance of Y constant for different settings; (c) obtain "reversal" of the observed vs. true effect as we vary confounding. We have run many more settings with higher-degree polynomials, trigonometric, and exponential functions, and we plan to add these!

R2 [Tradeoffs.] Thank you for this question! One way to view the trade-off is via requirements on the bounds: In some cases, bounding the true effect to be above or below zero could be a valuable insight. In other situations more details of the causal effect are required to inform decisions, requiring stronger assumptions. So the context can inform the necessary strength of assumptions. Another view on the trade-off is via domain knowledge: If one has expert knowledge that an effect is, say, linear, then one can use this to obtain tighter, more informative, bounds. If instead an expert has information about the smoothness of the causal effect, methods such as regression splines can be controlled based on this (see Gunsilius, 2019). Our aim is primarily to allow for flexible assumptions. Overall, we argue that a practitioner is much more likely to have expert knowledge about the form of the response function versus other common assumptions (e.g., additivity). In these cases they can directly tune the size of the function class to obtain bounds that are tight enough to suit the particular context at hand. Thank you for asking about this, we will add this to the text.

R2 [Include Bennet et al., 2019, Lewis & Syrgkanis 2018.] Thanks for these! We will add them to the related work.

R3 [Not in basis.] Good question. We agree there could be cases where the response is not parameterized by basis functions. In this case, a practitioner can still use our framework but with a different differentiable response parameterization. We simply chose basis functions as a flexible example that allows for efficient optimization. For example, if one has information about the smoothness of the causal effect, methods such as splines/wavelets can be controlled for this (see Gunsilius, 2019). It's all about allowing for choices as opposed to a one-size-fits-all tool.

R3 [Lack of prior knowledge.] Agreed, without any prior knowledge, the bounds may not capture the causal effect. This makes sense as IV cause-effects are unidentifiable from observations without assumptions. We argue that the classic solution: assuming additivity, is rarely justified by domain knowledge. However, it is not uncommon to have knowledge about the parametric family or the smoothness of the causal effect, which many causal-effect methods utilize.

R3 [Grid for Z, relax bound.] This is a deep question. Ultimately, we argue this must be investigated case-by-case. The more levels of Z the better if we can afford it. In practice, an analyst can choose a few treatment levels and see how the bounds on the ATEs shrink as more Zs are entered, before setting on a 'satisficing' plateau to use on other ATEs.

R3 [How constraints baked in.] Thanks. We will add more detail about this in the text.

R3 [Identifying $p(X|Z)$.] Ah let us clarify this (more in Appendix). We always identify $p(X|Z)$ using grids in "CDF space": each bin contains the same number of points. This way we don't need to assume any distribution for $p(X|Z)$.

R5 [Sensitivity w.r.t. response.] We have run many more settings where the guessed response differs from the true response: high-degree polynomials, trigonometric, and exponential functions, and we'll add these to the Appendix.

R5 [Causal effect outside bounds.] Curves may miss the bounds as these are estimated bounds, versus population ones. Higher samples sizes will control this error, but it will be there as in any learning problem.

R5 [Tests.] There is a degree of testability (e.g., linear models imply falsifiable constraints), but we would rather cover these issues in a journal version. We want to emphasize that as causal methodologists all we want is to provide tools that allow for flexible assumption-making, and put the ultimate responsibility where it should be, the domain expert.

R5 [Discretize, BP94.] Four reasons it won't work: (1) Discretizing X is a non-invertible mapping, so in general exclusion restriction is destroyed i.e. independence of Z and Y given $\{X, U\}$ stops holding when X is discretized. (2) If we discretize X and Y each into k categories, the number of response functions and constraints is $O(k^k)$. (3) "$do(X^* = x^*)$" is ill-defined for a discretization X^* of X. (4) Vacuous bounds at $k \to \infty$. Thank you for raising this.