We thank all reviewers for their careful reading, insightful comments and feedback. We apologize for typos and the lack of clarity and heavy notation in some places. We will take all comments into account and fix typos and improve clarity.

R1: Our convergence is robust even in the absence of completeness or point identifiability (PI). Having PI allows us to argue stronger convergence. Hence our results are simply stronger than assuming PI from the get-go as is typically done. Our bounds are applicable to weak instruments where PI exists in the limit but is brittle with finite n. Moreover (as noted in prior work) projected RMSE convergence suffices to estimate a nuisance parameter in a semiparametric model and making good predictions does not require parameter identification. We could assume completeness and together with a bound on ill-posedness (Apx C.4) our bounds imply identifiability. We will add a Corollary to this affect. Note that we present a general criterion (lines 62-65) for test function selection to ensure equivalence of our objective to $E[y - h(x)|z] = 0$. We also give many general function classes in our paper where this criterion yields a natural function class. η_n in line 148 appears in our bound on RMSE in line 152. It measures how good of an approximation our norm-constrained test function family is to the optimal test function family by measuring the maximum error of the best function in our norm-constrained family over all hypotheses in the space $\mathcal{H}$. We provide theoretical lower bounds on the regularization hyper-parameters which can be used to select them. In addition to perform cross-validation for hyper-parameter selection we have found that we can approximate the supremum over $\mathcal{F}$ by the supremum over the set of functions f encountered in a pre-training phase where we store the set of test functions f we evaluate against at each point in the pre-training. We can use this set later on for cross-validation to effectively simulate a supremum over $\mathcal{F}$. Lastly, we do not address the question of hypothesis testing in this work ,but primarily address estimation. Testing and estimation are two orthogonal tasks that are typically complementary but are also orthogonal and each of own interest.

R2: We apologize for the terse presentation of our experimental results in the main body which appears to have caused an impression that we are not comparing to the recent works in IV regression. We will present a clearer, more comprehensive experimental comparison in future versions. First, we would like to point that we do contain a more comprehensive presentation of our experimental results in the Apx. For non-parametric IV there is no prior Random Forest (RF) algorithm, as we outline in the RF section. We present the first RF algorithm for this setting. Prior RF algorithms for IV setup only work when one makes the assumption of linearity w.r.t. to treatment and estimates heterogeneity with respect to exogenous features (such as the IV forest of Athey and Wager). In Fig 2, for neural nets we compare with AGMM which is reported to outperform DeepIV (hence we exclude DeepIV in the table). Fig 3, deals with a sparse linear setting where the dimension of the input can be much larger than the number of available samples. Many of the prior works cited do not have an explicit focus on handling sparsity and without such a focus would not scale well in the high-dimensional setting. That being said, for the sake of completeness, we will add a comparison of their performance with our approach in the Apx. In Fig 4, we compare with DeepGMM for two reasons. Many of the other works implementations do not scale computationally to such high-dimensional instrument and treatment spaces. Primarily neural-net based approaches scale well. DeepIV is one previous approach which works when the instrument space is high-dimensional but since the DeepGMM paper reported a better performance of their estimator compared to DeepIV in this setting, and since we outperform DeepGMM in this setting we left out a comparison with DeepIV in our table. We apologize for the lack of clarity in some places. We will fix all of them in the paper. We will 1) make explicit the zero-sum aspect of our min-max formulation, 2) fix references to Theorems in Apx, and 3) clarify the dependence structure for IV regression. We do need bounds on ill-posedness for good RMSE and we provide explicit bounds for some of the function classes we consider. However these constraints are different from the regularization terms of eq (2). The regularization terms represent bounds on the norms of the function classes (and consequently their complexity) and are necessary to get convergence. We will make this more clear. U in Theorem 1 denotes a bound on the norm of functions in $\mathcal{F}_U$ which is essential for controlling the complexity of the class of test functions and thereby getting convergence in RMSE.

R3: Selecting test function family is indeed important for our approach. We provide strong theoretical guidance to do this depending on the richness of our hypothesis class (for e.g. lines 62-65). Once we have selected a class of test functions, we show how hyper-parameters for regularized estimators can be picked in many instantiations. Experiments on real-world data is indeed an important direction and we leave it for future work. We do demonstrate the robustness of our approach to partial real-world data by showing its efficacy on data comprised of MNIST images. Prior work: The work on ML estimation for hetero effects assumes that the function $h(T, X)$ is linear in the endogenous treatment T and only heterogeneous wrt to the exogenous variables X. The linearity is the main assumption that enables the results of that work and makes a significant qualitative and technical difference. The unpublished arXiv work of AGMM does not provide statistical guarantees of the resulting estimator apart from a fully non-parametric rate that grows exponentially with dimension. A crucial difference is they don't penalize the objective with the norm of the test function which is the key idea that enables our fast rates (based on critical radius of $\mathcal{F}$). Finally, AGMM only provides experimental results for neural nets, while here we provide experimental and theoretical results for many other function classes of interest.