We appreciate the valuable comments from the reviewers, which will lead to a largely improved final paper. Following the unanimous suggestion by three reviewers, we will provide a simulation section in the final version.

Reviewer 2. We are thrilled to find that Reviewer 2 appreciates our goal to provide a theoretical foundation for the use of NNs in econometrics models and causal inference where conditional moment equations play a key role in identifying the structural parameters. We further point to Babii and Florens [1], which includes a battery of economics models that involve conditional moment restrictions, demonstrating the potential application of our method in those models.

Comparison with Dual IV We thank R2 for noticing that we provided a detailed comparison with Dual IV in Appendix F. The main difference is that we use a variational characterization of conditional expectation while dual IV resort to maximality principle (Lemma F.1). Our derivation has a natural connection to generalized method of moments.

Reviewer 4. The notation $\mathbb{E}_{\text{init}}[\cdot]$ It is the expectation taken over the random variables Ξ_0 or $\Xi_{H,0}$, defined in Eq. (9) and (10), respectively. We will define it explicitly in the final draft.

The bound of Theorem 4.2 As for the second comment, we emphasize that it should be f that appears in (15), the solution to Eq. (1), not f^α. We will list separately several assumptions made in Theorem 4.2 so that it is easier to parse.

Reviewer 5. The need for saddle-point formulation with NNs For example, consider the nonparametric IV problem $\mathbb{E}[g(X)|Z] = \mathbb{E}[Y|Z]$ where we want to solve for g. Consider the square loss $L(g) = \mathbb{E}_Z\left[\left(\mathbb{E}[g(X) - Y|Z]\right)^2\right]$ without Tikhonov penalty. Assume g is approximated by a function parameterized with θ. Taking the gradient w.r.t. θ and assuming exchange of ∇_θ and $\mathbb{E}$, we get $\nabla_\theta L(g) = 2\mathbb{E}_Z\left[\mathbb{E}[g(X) - Y|Z] \cdot \mathbb{E}[\nabla_\theta g(X)|Z]\right]$. Assume we observe iid samples of (X, Y, Z). The product of two expectation terms implies that, to obtain an unbiased estimate of the gradient, we will need two samples of (X, Y, Z) with Z taking the same value. This is usually unlikely except for simulated environments.

We can see saddle-point formulation eliminates the need of double samples. The double-sample challenge is a consequence of the nested structure of the problem, where the conditional expectation operator appears inside the square. This point was briefly mentioned in the beginning of Sec 2.2 and will be elaborated on in the final draft. The double-sample problem is also a typical issue in reinforcement learning literature. The "double sampling" problem becomes more problematic if we use other convex losses. We use NNs as approximators due to their representation power.

Assumptions on the inverse problem under study We study an instance of inverse problem where the operator is a conditional expectation operator. We assume (i) the equation $Af = b$ admits a unique solution f, and (ii) A is compact and f is "smooth" w.r.t to A. Under (i) we are working with a well-specified and identified model. Under (ii) we have a bound on the regularization bias (Lemma D.2). A sufficient condition for compactness of A is given in Appendix E. These assumptions are common in condition moment equation problems [1]. We also emphasize that the identification problem is very important and well studied in the literature, but we do not focus on this.

Relation to literature of inverse problems We focus on a very specific type of inverse problem. We contrast our method with two lines of work in conditional moment equations (CME) estimation: (i) non-parametric approach, and (ii) recent use of NN in CME. This is discussed in the Related Work section. We will provide additional discussion on the relationship of our method to other methods for solving inverse problems.

In stochastic inverse problem where the operator must be estimated from the data, nonparametric methods proceed in a two-stage type manner. First estimate the conditional expectation operator, using either sieve estimator or kernel estimator, and then solve a Tikhonov penalized problem [5]. However, these methods have serious drawbacks: (1) their performance depends on the choice of kernel function or spline basis; and (2) they become computationally intractable in high-dimensional feature spaces or with large numbers of training examples. The disadvantages of nonparametric methods have motivated recent works incorporating NN in CME, especially in instrumental variable regression. However, these methods suffer either the need of double sampling [3,4], unknown convergence behavior [1,4], or computational burden [3]. Our method is scalable, provably efficient and does not require double sampling.

References

[1] A. Babii and J.-P. Florens. *Is completeness necessary? Estimation in nonidentified linear models*, page 5, footnote 4.

[2] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual IV: A single stage instrumental variable regression.

[3] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction.

[4] G. Lewis and V. Syrgkanis. Adversarial generalized method of moments.

[5] M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. Handbook of Econometrics.