We would like to thank the reviewers for their useful, detailed feedback! We will update the paper with the suggested minor revisions regarding typos and presentation improvements, and respond to individual reviewers comments below.

Reviewer #1:

1. Corollary 1 doesn't actually imply the consistency proven in Theorem 1, because it only applies to a finite number of moment conditions, whereas in Theorem 1 we deal with an infinite number of moment conditions and a function space defined by a neural network architecture, which is why proving consistency there is important and necessary.
2. You're correct that training the adversarial loss function can be finicky, especially if done naively. We find that in practice when we include the $-\frac{1}{4}C$ term inspired by Lemma 1 learning is very stable in particular in the low dimensional scenarios, and other tweaks like OAdam and early stopping using our validation scheme seem to improve this further.
3. We agree that some of the prose discussing GMM+NN results is confusing vis-a-vis results in Table 1. We will alter this discussion accordingly to more clearly reflect the numbers.
4. We acknowledge your point about lack of misspecification in DeepIV's first stage in the low-dim scenarios, and so forbidden regression might not be the right explanation for its performance there. However, the DeepIV second stage network does have the capacity to fit the problems better than a polynomial. Regarding "full training": we used an existing implementation of DeepIV.
5. The network architectures and hyperparams were mistakenly omitted from the supplement and will be added. Note our code is public. For baseline methods we used the official implementations provided by the authors.
6. Thanks for catching, we apologize for the typo. Both 0.5's should be 1's: $X = Z_1 + e + \gamma$.
7. Yes, test MSE is w.r.t. the "true response" g_0 over the X population, as mentioned in 2nd paragraph of Sec 5.

Reviewer #3:

1. Good idea. We will use the notation $\|x\|_C = x^T C^{-1} x$.
2. We will add a proof in the appendix. This is just self-duality of L_2.
3. Regarding line 79, our comment is simply pointing out that standard GMM usually starts from the assumption that the moment conditions specify the problem. Indeed, the assumption that a finite number of moment conditions identify theta is very strong (too strong) when theta is complex because it easily gives us statistically efficient methods for estimating theta if true. We will clarify.
4. Yes; the limit of $\tilde{\theta}_n$ can be anything. This may seem counterintuitive, but one way of understanding this is that having the "correct" limit corresponds to using the optimally weighted norm, whereas having some other limit corresponds to some non-optimal norm, but using an non-optimal norm for GMM just means that estimator won't be statistically efficient, not that it won't be consistent.
5. All theory in AGMM applies to an *a priori* fixed finite collection of moment conditions; the only "learning [of] moment functions" in AGMM is in a heuristic jitter step added in the experiments. The significant differences between DeepGMM and AGMM, other than the drastic difference in performance, are that our method is inspired by the statistically efficient optimally weighted GMM with a corresponding regularization term which AGMM lacks, that we directly optimize the neural net critic f, and that we do learn a critic from an infinite collection rather than using a finite ensemble of critics. We will update this to make it clearer.
6. Yes; as line 190 says, what we mean is $\tilde{\theta}$ is treated as constant. The second term of U in Equation (9) has zero partial derivative in θ; so $\tilde{\theta}$ does not appear in the θ gradient. We will rephrase to make this clearer.
7. This means we use $f_i(Z) = Z_i$ for $i = 1, \ldots, 784$. We'll add this clarification.

Reviewer #4:

1. Identification assumption for neural nets: as referenced in line 161, we can easily relax identification and instead converge to *some* θ satisfying the moment conditions (will clarify this simple extension in the proof). Moreover, this immediately gives that even if there are redundancies in that two θs give the same function g (e.g., permuting hidden layers), if $g \in G$ is identified (i.e., all identified thetas give rise to the same function g), we will obtain *some* parameterization θ of the unique g. It's a good point so we will add this discussion.
2. Regarding the $-\frac{1}{4}C(f, f)$ term, we found that other regularizers/controls on f do not perform well as they induce suboptimal weighting that ignores the covariance of the moment conditions for different f, whereas our new regularizer, as you write, is motivated by optimal weighting. This a key driver of our improved performance over, e.g., AGMM.
3. The network architectures and learning hyperparams were mistakenly omitted from the supplement and will be added. Note our code is public. Re "sufficiently rich" moments for Poly2SLS: since its degree is variable, Poly2SLS can be thought of as a sieve as in Newey and Powell [23], giving universal consistency.
4. We will cite the recent paper you reference on kernelized 2SLS. It wasn't on our radar (first appearance online June 2019). We will also cite the Ravuri et al. paper re future work. Thanks.