Technical detail. Well caught! The situation regarding [17] is even worse than Reviewer 3 highlighted: in infinite dimensional spaces, one cannot simply exchange trace and expectation by assuming linearity. The good news, however, is that we can prove $tr(T_1) < \infty$ under the mild assumptions in Hypotheses 2 and 3, rescuing the theorems in both [17] and our work. We will include this proof and discussion in the document, and alert the authors of [17] to this issue.

In Hypotheses 2 and 3, we assume that instrument space $\mathcal{Z}$ is separable, and that RKHS $\mathcal{H}_{\mathcal{Z}}$ has continuous, bounded kernel $k_{\mathcal{Z}}$ with feature map $\phi(z)$. By Proposition 3, $\mathcal{H}_{\mathcal{Z}}$ is separable, i.e. it has countable orthonormal basis $\{e_i\}_{i=1}^{\infty}$. Consider the space $\mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{Z}})$ of Hilbert-Schmidt operators $A : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{Z}}$ with inner product $\langle A, B\rangle_{\mathcal{L}_2} = \sum_{i=1}^{\infty} \langle Ae_i, Be_i\rangle_{\mathcal{H}_{\mathcal{Z}}}$. Recall tensor product notation: for $a, b, c \in \mathcal{H}_{\mathcal{Z}}$, $[a \otimes b]c = \langle b, c\rangle_{\mathcal{H}_{\mathcal{Z}}} a$. By Parseval's identity, we have two helpful results: $\|a \otimes b\|^2_{\mathcal{L}_2} = \|a\|^2_{\mathcal{H}_{\mathcal{Z}}} \|b\|^2_{\mathcal{H}_{\mathcal{Z}}}$ so $a \otimes b \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{Z}})$ [G, eq. 3.6]; and if $C \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{Z}})$ then $\langle C, a \otimes b\rangle_{\mathcal{L}_2} = \langle a, Cb\rangle_{\mathcal{H}_{\mathcal{Z}}}$ [G, eq. 3.7].

First, we verify the existence of covariance operator $T_1 \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{Z}})$ satisfying $\langle T_1, A\rangle_{\mathcal{L}_2} = \mathbb{E}\langle \phi(Z) \otimes \phi(Z), A\rangle_{\mathcal{L}_2}$. By Riesz representation theorem, T_1 exists if the RHS is a bounded linear operator. Linearity follows by definition. Boundedness of $k_{\mathcal{Z}}$ in Hypothesis 3 implies $\mathbb{E}[k_{\mathcal{Z}}(Z, Z)] < \infty$ and hence

$$|\mathbb{E}\langle \phi(Z) \otimes \phi(Z), A\rangle_{\mathcal{L}_2}| \leq \mathbb{E}|\langle \phi(Z) \otimes \phi(Z), A\rangle_{\mathcal{L}_2}| \leq \|A\|_{\mathcal{L}_2} \mathbb{E}\|\phi(Z) \otimes \phi(Z)\|_{\mathcal{L}_2} = \|A\|_{\mathcal{L}_2} \mathbb{E}[k_{\mathcal{Z}}(Z, Z)] < \infty$$

Second, we verify T_1 is indeed a covariance operator with $tr(T_1) < \infty$.

$$\langle f, T_1 g\rangle_{\mathcal{H}_{\mathcal{Z}}} = \langle T_1, f \otimes g\rangle_{\mathcal{L}_2} = \mathbb{E}\langle \phi(Z) \otimes \phi(Z), f \otimes g\rangle_{\mathcal{L}_2} = \mathbb{E}\langle f, \phi(Z)\rangle_{\mathcal{H}_{\mathcal{Z}}} \langle g, \phi(Z)\rangle_{\mathcal{H}_{\mathcal{Z}}} = \mathbb{E}[f(Z)g(Z)]$$

$$tr(T_1) = \sum_{i=1}^{\infty} \langle e_i, T_1 e_i\rangle_{\mathcal{H}_{\mathcal{Z}}} = \sum_{i=1}^{\infty} \mathbb{E}\langle e_i, \phi(Z)\rangle^2_{\mathcal{H}_{\mathcal{Z}}} = \mathbb{E} \sum_{i=1}^{\infty} \langle e_i, \phi(Z)\rangle^2_{\mathcal{H}_{\mathcal{Z}}} = \mathbb{E}\|\phi(Z)\|^2_{\mathcal{H}_{\mathcal{Z}}} = \mathbb{E}[k_{\mathcal{Z}}(Z, Z)] < \infty$$

where the second line uses definition of trace, the penultimate expression in the first line, monotone convergence theorem [43, Theorem A.3.5] with upper bound $\|\phi(z)\|^2$, Parseval's identity, and boundedness of $k_{\mathcal{Z}}$.

Limitations. Extensive use of IV estimation in applied economic research has revealed a common pitfall: weak instrumental variables. A weak instrument satisfies Hypothesis 1, but the relationship between a weak instrument Z and input X is negligible; Z is essentially irrelevant. In this case, IV estimation becomes highly erratic [B]. In [St], the authors formalize this phenomenon with local analysis. We recommend that practitioners resist the temptation to use many weak instruments, and instead use few strong instruments such as those described in the introduction.

Experiments. We provide implementation details for `KernelIV` and its competitors in Appendix 7.10.2, including kernel choice and kernel hyperparameter tuning. Theorem 4 details the performance of KIV with suboptimal n/m, parametrized by a. In Figure 9, we present a *linear* design [14] with $h(x) = 4x - 2$. We will include Figure 5 in the main text, and move linear and sigmoid designs to the appendix. In Figure 10, we provide a robustness study of `KernelIV` applied to the sigmoid design with $n + m = 1000$, varying hyperparameter values for Guassian kernel $k_{\mathcal{X}}$. For comparison, our tuning procedure selects value 0.3. We will increase figure sizes.

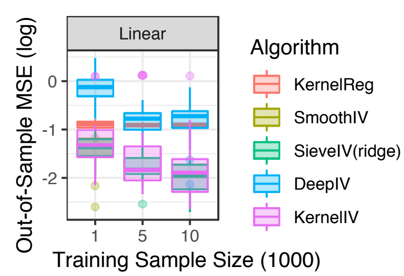

Figure 9: Linear design

Exposition. We will define e as unmeasured, confounding noise, and relate n/m to statistical efficiency earlier on. In Hypotheses 5 and 9, we will define the power of an operator in terms of its eigendecomposition. We will move the decay schedule for λ from Appendix 7.6 to Theorem 2 . We define $\Omega_{\mu(z)}$ in line 257, but we will restate this definition in Definition 2 and Hypothesis 7 for clarity. We will replace 'a.s.' with 'almost surely' in Hypothesis 8.

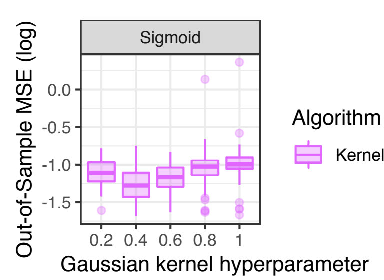

Figure 10: Robustness study

References. We agree it is important to cite early work on mean embeddings by [Sm] as summarized in [M]. We will ensure all references are cited in the main text. We cite groups of papers for the following reasons: [24, 25] introduce E and μ, which in our paper we argue are equivalent; [25, 26] and likewise [32, 33, 34] were published at the same time; [39, 40] contain different theorems that we generalize into Theorems 6 and 5 en route to Theorem 2; [46, 47] contain an original consistency argument and a stronger minimax optimality argument, respectively.

[B] J Bound, DA Jaeger, and RM Baker. Problems with IV estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *JASA*, 90(430):443–450, 1995. [G] A Gretton. RKHS in ML: Testing statistical dependence. Adv. topics in ML lecture notes, UCL Gatsby Unit, 2018. [M] K Muandet, K Fukumizu, BK Sriperumbudur, and B Schölkopf. Kernel mean embedding of distributions: A review and beyond. *FTML*, 10(1-2):1-141, 2017. [Sm] A Smola, A Gretton, L Song, and B Schölkopf. A Hilbert space embedding for distributions. In *ALT*, pages 13–31, 2007. [St] D Staiger and JH Stock. IV regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.