
Demystifying Spectral Feature Learning for Instrumental Variable Regression

Dimitri Meunier*
Gatsby Unit, UCL

Antoine Moulin
Universitat Pompeu Fabra

Jakub Wornbard
Gatsby Unit, UCL

Vladimir R. Kostic
Istituto Italiano di Tecnologia & University of Novi Sad

Arthur Gretton
Gatsby Unit, UCL

Abstract

We address the problem of causal effect estimation in the presence of hidden confounders, using nonparametric instrumental variable (IV) regression. A leading strategy employs *spectral features* - that is, learned features spanning the top eigensubspaces of the operator linking treatments to instruments. We derive a generalization error bound for a two-stage least squares estimator based on spectral features, and gain insights into the method’s performance and failure modes. We show that performance depends on two key factors, leading to a clear taxonomy of outcomes. In a *good* scenario, the approach is optimal. This occurs with strong *spectral alignment*, meaning the structural function is well-represented by the top eigenfunctions of the conditional operator, coupled with this operator’s slow eigenvalue decay, indicating a strong instrument. Performance degrades in a *bad* scenario: spectral alignment remains strong, but rapid eigenvalue decay (indicating a weaker instrument) demands significantly more samples for effective feature learning. Finally, in the *ugly* scenario, weak spectral alignment causes the method to fail, regardless of the eigenvalues’ characteristics. Our synthetic experiments empirically validate this taxonomy. We further introduce a practical procedure to estimate these spectral properties from data, allowing practitioners to diagnose which regime a given problem falls into. We apply this method to the dSprites dataset, demonstrating its utility.

1 Introduction

We study the nonparametric instrumental variable (NPIV) model [1, 32]

$$Y = h_0(X) + U, \quad \mathbb{E}[U \mid Z] = 0, \quad Z \not\perp X, \quad (1)$$

where X is the treatment, Y the outcome, Z the instrument, U an unobserved confounder, and $h_0 \in L_2(X)$ is the structural function to be learned, where we denoted $L_2(X)$ the L_2 space corresponding to the distribution \mathbb{P}_X . We assume no common confounder of Z and Y . To illustrate this setting, consider a simplified example inspired by [20]. Suppose we want to determine how ticket price (X) affects the number of flight tickets sold (Y). A potential confounder (U) could be a major event (e.g., a large conference or festival) occurring at the destination. This event would likely increase demand for flights, and airline pricing algorithms would result in more expensive seats being sold; the confounder U influences both X and Y . A naive regression of sales on price might then misleadingly suggest that higher prices lead to higher sales. This occurs because the analysis fails to account for the event, which independently drives up both demand (sales) and prices. To obtain a less biased estimate of the price effect, we could use an instrument (Z), such as the cost of jet fuel. The

*direct correspondence to dimitri.meunier.21@ucl.ac.uk

cost of fuel is a plausible instrument because: (i) it likely affects ticket prices (X) as airlines adjust fares based on operational costs, and (ii) it is unlikely to be directly correlated with whether a major event (U) is happening at the destination, thus satisfying $\mathbb{E}[U | Z] = 0$. By using fuel cost as an instrument, we can isolate the variation in ticket prices that is independent of the unobserved demand shock caused by the event, allowing for a more accurate estimation of the structural function h_0 .

Central to the NPIV model is the **conditional expectation operator** $\mathcal{T} : L_2(X) \rightarrow L_2(Z)$, defined by $\mathcal{T}h(Z) = \mathbb{E}[h(X) | Z]$. Estimating h_0 amounts to inverting this operator from the following integral equation

$$\mathcal{T}h_0 = r_0, \quad \text{where } r_0(Z) \doteq \mathbb{E}[Y | Z]. \quad (2)$$

In practice, the operator \mathcal{T} is unknown and must be estimated from data. Moreover, even if \mathcal{T} were known, its inverse is typically unbounded, making recovery of h_0 an ill-posed inverse problem that requires regularization.

A classical approach to solving NPIV is nonlinear two-stage least squares (2SLS), which projects X and Z into suitable feature spaces and performs regression in these representations. Let φ and ψ denote feature maps for X and Z , respectively. In **Stage 1**, one estimates the conditional expectation $\mathbb{E}[\varphi(X) | Z]$ by regressing $\varphi(X)$ on $\psi(Z)$, yielding an approximation of the form $\mathbb{E}[\varphi(X) | Z] \approx A\psi(Z)$, for some linear map A . This step isolates the component of $\varphi(X)$ that is predictable from Z , filtering out variation due to the unobserved confounder. In **Stage 2**, the outcome Y is regressed on the predicted features $\mathbb{E}[\varphi(X) | Z]$, yielding an estimator of the structural function h_0 , which is modeled as a linear combination $h_0(X) \approx \beta^\top \varphi(X)$, for some coefficient vector β . Given that the feature maps φ and ψ can be fixed (e.g., polynomials, splines, or radial basis functions from a kernel) or learned from data (e.g., using neural networks), a natural question arises: **What specific information must these features encode to enable efficient estimation?**

The algorithm recently proposed by [36], which learns features by minimizing a spectral contrastive loss, suggests that the answer lies in the spectral structure of \mathcal{T} . In this paper, we rigorously investigate this perspective. We show that the performance of this approach hinges on two key factors: the **spectral alignment** of the structural function h_0 with the operator \mathcal{T} , and the rate of its **singular value decay**. Namely, we identify three distinct regimes governing the effectiveness of spectral methods for causal inference:²

1. The **good** regime, where the structural function aligns strongly with the leading eigenspaces of the conditional expectation operator \mathcal{T} , and the singular values of \mathcal{T} decay slowly;
2. The **bad** regime, where spectral alignment holds, but rapid singular value decay leads to instability and degraded estimation quality;
3. The **ugly** regime, where the structural function is misaligned with the top eigenspaces of \mathcal{T} , rendering spectral methods ineffective regardless of decay rate.

Overview of contributions. We investigate the effectiveness of spectral features in NPIV estimation, grounding their empirical success in a principled theoretical analysis. Specifically:

- By leveraging an upper bound on the generalization error for the 2SLS estimator and specializing it to spectral features, we characterize the conditions under which each regime arises.
- We clarify the theoretical foundation of the state-of-the-art causal estimation algorithm of [36] by showing that it implements a 2SLS estimator with learned features that explicitly target the leading eigenspaces of the conditional expectation operator \mathcal{T} . Although [36] suggested a heuristic link to the Singular Value Decomposition (SVD) of \mathcal{T} , we make this connection explicit by proving that their contrastive objective corresponds to a Hilbert-Schmidt approximation of \mathcal{T} .
- We empirically validate the “good–bad–ugly” taxonomy through synthetic experiments designed to isolate the effects of spectral alignment and singular value decay on generalization error. We further introduce a practical, data-driven procedure to estimate these key spectral properties (spectral alignment and singular value decay), allowing practitioners to diagnose their problem’s regime. We demonstrate this procedure on the dSprites dataset [30].

²The “good–bad–ugly” terminology is borrowed from [24], who used it to describe performance regimes in Koopman operator learning. We adopt the same naming convention in a different context: spectral feature learning for causal inference. The phrase itself, of course, originates from the classic [27].

Related work. Instrumental variable regression is often implemented as a two-stage procedure, where the key design choice is the feature representation. Fixed features include polynomials, splines, or kernel methods [3, 6, 7, 31, 35], while learned features are either derived from the conditional expectation operator \mathcal{T} [36] or jointly optimized with the 2SLS objective [23, 33, 40]. Some approaches replace first-stage regression with conditional density estimation [11, 18, 20, 28, 34].

An alternative line of work reformulates IV regression as a saddle-point optimization problem to avoid the bias introduced by nested conditional expectation estimation [2, 13, 29, 38, 43]. These methods frame estimation as a min-max game between a candidate solution h and a dual witness function enforcing the IV moment condition. This approach sidesteps explicit conditional expectation estimation but relies on selecting tractable, fixed function classes to ensure the well-posedness of the optimization problem.

Features can be learned in various ways. One line of work learns them jointly with the regression objective, as in [16, 23, 33, 40], who propose end-to-end frameworks that simultaneously optimize over the feature representations and second-stage parameters. In contrast, [36, 38] propose to learn features that reflect the eigenstructure of \mathcal{T} , independently of the outcome Y , using a spectral contrastive loss. This may appear suboptimal, since incorporating Y could, in principle, yield more informative representations, however it avoids the optimization difficulties inherent to nonconvex, joint objectives involving all three variables (X, Y, Z) . This decoupled spectral approach was shown in certain experimental settings to outperform end-to-end alternatives, as demonstrated in [36]. Our work places this empirical finding in context, since it likely arises due to the experiments in that work satisfying the “good” regime in the present work. We provide further discussion in Appendix D.

Structure of the paper. Section 2 introduces the notation and preliminaries. Section 3 reviews the Sieve 2SLS estimator and presents generalization bounds highlighting the role of spectral features. Section 4 shows how to learn such features via a contrastive loss and connects this to the method of [36]. Section 5 presents a practical procedure to estimate the spectral properties from data, enabling an empirical diagnosis of our taxonomy. Section 6 provides synthetic experiments, including an application to the dSprites dataset, that validate our theoretical taxonomy. Proofs are deferred to the Appendix.

2 Preliminaries

2.1 Background

Probability Spaces and Function Spaces. Y is defined on \mathbb{R} , while X and Z take values in measurable spaces \mathcal{X} and \mathcal{Z} , respectively, endowed with their σ -fields $\mathcal{F}_{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{Z}}$. We let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space with expectation operator \mathbb{E} . For $R, S \in \{X, Z, (X, Z)\}$, let π_R denote the push-forward of \mathbb{P} under R , and let $\pi_R \otimes \pi_S$ denote the product measure on the corresponding product σ -field, defined by $(\pi_R \otimes \pi_S)(A \times B) = \pi_R(A)\pi_S(B)$ for measurable sets A, B . For $R \in \{X, Z\}$ and corresponding domain $\mathcal{R} \in \{\mathcal{X}, \mathcal{Z}\}$, we abbreviate $L_2(R) \equiv L_2(\mathcal{R}, \pi_R)$ as the space of measurable functions $f : \mathcal{R} \rightarrow \mathbb{R}$ such that $\int |f|^2 d\pi_R < \infty$, defined up to π_R -almost everywhere equivalence.

Operators on Hilbert Spaces. Let H be a Hilbert space. For a bounded linear operator A acting on H , we denote its operator norm by $\|A\|_{\text{op}}$, its Hilbert–Schmidt norm by $\|A\|_{\text{HS}}$, its Moore–Penrose inverse by A^\dagger , and its adjoint by A^* . For finite-dimensional operators, the Hilbert–Schmidt norm coincides with the Frobenius norm. We denote the range and null space of A by $\mathcal{R}(A)$ and $\mathcal{N}(A)$, respectively. Given a closed subspace $M \subseteq H$, we write M^\perp for its orthogonal complement, \overline{M} for its closure, and Π_M for the orthogonal projection onto M . The orthogonal projection onto M^\perp is denoted by $(\Pi_M)^\perp \doteq I_H - \Pi_M = \Pi_{M^\perp}$. For functions $f, h \in L_2(X)$, $g \in L_2(Z)$, the rank-one operator $g \otimes f$ is defined as $(g \otimes f)(h) = \langle h, f \rangle_{L_2(X)} g$. This generalizes the standard outer product. For vectors $x \in \mathbb{R}^d$, we write $\|x\|_{\ell_2}$ for the Euclidean norm.

Data Splitting and Empirical Expectations. Given $n, m \geq 0$, we consider two independent datasets: an unlabeled dataset, $\tilde{\mathcal{D}}_m = \{(\tilde{z}_i, \tilde{x}_i)\}_{i=1}^m$, used to learn features for X and Z , and a labeled dataset, $\mathcal{D}_n = \{(z_i, x_i, y_i)\}_{i=1}^n$, used to estimate the structural function. For $R \in \{X, Z, Y, (X, Z), (X, Y), (Z, Y), (X, Z, Y)\}$ and a measurable function f , we define the empirical expectation over $\tilde{\mathcal{D}}_m$ as $\hat{\mathbb{E}}_m[f(R)] \doteq \frac{1}{m} \sum_{i=1}^m f(\tilde{r}_i)$, and similarly for $\hat{\mathbb{E}}_n$ on \mathcal{D}_n .

Feature Maps and Projection Operators. Let $d \geq 1$, and let $\varphi : \mathcal{X} \rightarrow \mathbb{R}^d$ be a feature map with components $\varphi_i \in L_2(X)$ for $i = 1, \dots, d$. We define $\mathcal{H}_{\varphi,d} \doteq \text{span}\{\varphi_1, \dots, \varphi_d\}$, and let $\Pi_{\varphi,d}$ be the L_2 -orthogonal projection onto this space. Analogous definitions apply for feature maps $\psi : \mathcal{Z} \rightarrow \mathbb{R}^d$, yielding $\mathcal{H}_{\psi,d}$ and $\Pi_{\psi,d}$.

2.2 On the Conditional Expectation Operator: NPIV as an Ill-posed Inverse Problem

We begin by defining the conditional expectation operator. Let

$$\mathcal{T} : L_2(X) \rightarrow L_2(Z), \quad h \mapsto \mathbb{E}[h(X) \mid Z].$$

As highlighted by Eq. (2), this operator plays a central role in the NPIV model. It encodes the conditional distribution of X given Z and provides a convenient lens through which to analyze how efficiently one can estimate h_0 . The NPIV equation admits a solution if and only if $r_0 \in \mathcal{R}(\mathcal{T})$, and is identifiable if and only if \mathcal{T} is injective. We therefore adopt the following assumption.

Assumption 1. \mathcal{T} is injective and there exists a solution to the NPIV problem, i.e., $\mathcal{T}^{-1}(\{r_0\}) \neq \emptyset$.

The existence part is essential: without it, there is no structural function h_0 satisfying the model in Eq. (1). When \mathcal{T} is not injective, we lose identifiability (i.e., unicity of the solution). In such scenarios, it is possible to establish consistency to a minimum-norm solution of the inverse problem Eq. (2), namely $h_* \doteq \mathcal{T}^\dagger r_0$ [15].³ We assume injectivity henceforth to simplify the discussion.

Eq. (2) is an inverse problem in which both the operator \mathcal{T} and the right-hand side r_0 are unknown. This inverse problem is typically **ill-posed**, as \mathcal{T}^{-1} is generally not continuous.⁴ Hence, even if \mathcal{T} were known, an approximation $\hat{r} \approx r_0$ would not ensure that $\mathcal{T}^{-1}\hat{r}$ is close to $h_0 = \mathcal{T}^{-1}r_0$. In such settings, regularization is required to obtain stable approximate inverses. Two common regularization approaches are **Tikhonov regularization**, where $\mathcal{T}^{-1} \approx (\mathcal{T}^* \mathcal{T} + \lambda I)^{-1} \mathcal{T}^*$, $\lambda > 0$, and **spectral cut-off**, which truncates the Singular Value Decomposition (SVD) [14].

When \mathcal{T} is compact (and therefore it admits a countable sequence of singular values), the **degree of ill-posedness** of the inverse problem is characterized by the decay rate of these singular values. Since the singular values of \mathcal{T}^{-1} are the reciprocals of those of \mathcal{T} , faster decay implies worse ill-posedness. The NPIV problem can be viewed as particularly challenging, as in addition to the ill-posedness of the operator \mathcal{T} , one must also estimate it from data. We now introduce a sufficient condition for compactness of \mathcal{T} .

Assumption 2. The joint distribution π_{XZ} is dominated by the product measure $\pi_X \otimes \pi_Z$, and its Radon–Nikodym derivative, or “density”, $p \doteq \frac{d\pi_{XZ}}{d\pi_X \otimes \pi_Z}$ belongs to $L_2(\mathcal{X} \times \mathcal{Z}, \pi_X \otimes \pi_Z)$.

Assumption 2 is standard in the NPIV literature (Assumption A.1 in [11]). It may fail when X and Z share variables or are deterministically related. For example, if $Z = (W, V)$ and $X = V$, then the joint distribution is not dominated by the product measure.

Proposition 1. Under Assumption 2, \mathcal{T} is a Hilbert–Schmidt operator and thus compact.

This result is well known in the inverse problems literature; we provide a proof in Appendix C.1.

Compact Decomposition and Eigensubspaces. As \mathcal{T} is a conditional expectation operator, it is non-expansive, i.e., $\|\mathcal{T}\|_{\text{op}} \leq 1$. Under Assumption 2, \mathcal{T} is compact, and therefore admits a SVD

$$\mathcal{T} = \sum_{i \geq 1} \sigma_i u_i \otimes v_i, \quad (3)$$

where $\{u_i\}$ and $\{v_i\}$ are orthonormal basis of $L_2(Z)$ and $L_2(X)$ respectively (left and right singular functions, respectively), and σ_i are the nonnegative singular values in nonincreasing order. The leading singular triplet is known explicitly: $\sigma_1 = 1$, $u_1 = \mathbb{1}_Z$ and $v_1 = \mathbb{1}_X$ corresponding to the constant functions. Under the assumption that \mathcal{T} is injective (Assumption 1) $\sigma_i > 0$ for all $i \geq 1$. Finally, since \mathcal{T} is Hilbert–Schmidt, $\sum_{i \geq 1} \sigma_i^2 < \infty$. For any $d \geq 1$, define the leading eigensubspaces of \mathcal{T} as

$$\begin{aligned} \mathcal{U}_d &\doteq \text{span}\{u_1, \dots, u_d\}, & \Pi_{Z,d} &\doteq \text{orthogonal projection onto } \mathcal{U}_d, \\ \mathcal{V}_d &\doteq \text{span}\{v_1, \dots, v_d\}, & \Pi_{X,d} &\doteq \text{orthogonal projection onto } \mathcal{V}_d. \end{aligned}$$

³This corresponds to the unique element in the pre-image of r_0 under \mathcal{T} that minimizes the L_2 -norm and is equivalently characterized as the unique element of the set $\mathcal{T}^{-1}(\{r_0\}) \cap \mathcal{N}(\mathcal{T})^\perp$ [2].

⁴In particular, \mathcal{T}^{-1} is discontinuous if \mathcal{T} is compact and not finite rank.

In addition, define $\mathcal{T}_d \doteq \sum_{i=1}^d \sigma_i u_i \otimes v_i$ the rank- d truncation of \mathcal{T} . To avoid ambiguity in the definition of $\mathcal{U}_d, \mathcal{V}_d$ and \mathcal{T}_d , we assume that $\sigma_d > \sigma_{d+1}$. We refer to Appendix A for more details on spectral theory.

3 Solving NPIV with 2SLS

To understand how and when spectral features should be used, we first recall results from the Sieve 2SLS literature. This gives us a clear picture of what terms are to be controlled in order to achieve good generalization, which we will use to demonstrate the effectiveness of spectral features and reveal our good-bad-ugly taxonomy.

3.1 Sieve 2SLS estimator

We start with the following characterization of the structural function:

$$h_0 = \arg \min_{h \in L_2(X)} \mathbb{E} \left[(Y - \mathcal{T}h(Z))^2 \right],$$

where the minimizer, h_0 , is unique under Assumption 1. A popular strategy to estimate h_0 is to consider a hypothesis space $\mathcal{H}_{\mathcal{X}} \subset L_2(X)$ and a procedure to estimate the action of \mathcal{T} : $\hat{\mathcal{T}}h \approx \mathcal{T}h$ for all $h \in \mathcal{H}_{\mathcal{X}}$, in order to obtain

$$\hat{h} \in \arg \min_{h \in \mathcal{H}_{\mathcal{X}}} \hat{\mathbb{E}}_n \left[(Y - \hat{\mathcal{T}}h(Z))^2 \right], \quad (4)$$

possibly subject to regularization. In the algorithm 2SLS, we consider features on both \mathcal{X} and \mathcal{Z} , and both steps (estimating $h \mapsto \mathcal{T}h$ and solving Eq. (4)) are carried out via linear regression in feature space. The two sets of features may be infinite-dimensional—*e.g.*, using Reproducing Kernel Hilbert Spaces (RKHS)—or finite-dimensional—*e.g.*, using splines, wavelets, or neural network features. When the feature dimension grows with sample size, we obtain a *sieve*. Since our focus is on neural network-based features, we present the finite-dimensional version, noting that the ideas naturally extend to infinite-dimensional cases with proper regularization [35, 31]. We refer to Appendix D for a description of other 2SLS methods.

Consider a feature map $\varphi : \mathcal{X} \rightarrow \mathbb{R}^d$ such that $\mathcal{H}_{\mathcal{X}} \doteq \{x \mapsto \theta^\top \varphi(x), \theta \in \mathbb{R}^d\}$ is included in $L_2(X)$. Then for any $h = \theta^\top \varphi(\cdot) \in \mathcal{H}_{\mathcal{X}}$, $\mathcal{T}h(Z) = \theta^\top \mathbb{E}[\varphi(X) | Z]$. Therefore, estimating $F_\star(Z) \doteq \mathbb{E}[\varphi(X) | Z]$ allows us to estimate $\mathcal{T}h$ for any $h \in \mathcal{H}_{\mathcal{X}}$. Let $\psi : \mathcal{Z} \rightarrow \mathbb{R}^d$ be another feature map with $\mathcal{H}_{\mathcal{Z}} \doteq \{z \mapsto \alpha^\top \psi(z), \alpha \in \mathbb{R}^d\} \subset L_2(Z)$. We estimate F_\star with vector-valued linear regression in feature space by approximating F_\star with $A\psi(\cdot)$, where $A \in \mathbb{R}^{d \times d}$, solving

$$\hat{A} \doteq \hat{\mathbb{E}}_n[\varphi(X)\psi(Z)^\top] \hat{\mathbb{E}}_n[\psi(Z)\psi(Z)^\top]^\dagger \in \arg \min_{A \in \mathbb{R}^{d \times d}} \hat{\mathbb{E}}_n[\|\varphi(X) - A\psi(Z)\|_{\ell_2}^2].$$

The pseudo-inverse yields the *minimum-norm solution* in the case of non-uniqueness. Regularized versions, such as Tikhonov-regularized least squares (also known as ridge regression), can be obtained by penalizing with $\eta \|A\|_{\text{HS}}^2$ for $\eta > 0$. Computing \hat{A} is referred to as the **first stage** of the 2SLS algorithm. For the **second stage**, for any $h = \theta^\top \varphi(\cdot) \in \mathcal{H}_{\mathcal{X}}$, we define $\hat{\mathcal{T}}h(Z) \doteq \theta^\top \hat{F}(Z)$ with $\hat{F}(Z) = \hat{A}\psi(Z)$. Plugging this into Eq. (4) yields $\hat{h} = \hat{\theta}^\top \varphi(\cdot)$ with

$$\hat{\theta} \doteq \hat{\mathbb{E}}_n[\hat{F}(Z)\hat{F}(Z)^\top]^\dagger \hat{\mathbb{E}}_n[\hat{F}(Z)Y] \in \arg \min_{\theta \in \mathbb{R}^d} \hat{\mathbb{E}}_n[(Y - \theta^\top \hat{F}(Z))^2]. \quad (5)$$

Similarly to the first stage, we can also consider a regularized version of the second stage, where we penalize with $\lambda \|\theta\|_{\ell_2}^2$ for $\lambda > 0$. \hat{h} is the sieve NPIV estimator [3, 7, 6].

2SLS versus saddle-point estimator. Based on features φ and ψ , [43, 36] proposed an estimator in a saddle-point form: $\hat{h}_{bis} = \hat{\theta}_{bis}^\top \varphi(\cdot)$, with

$$\hat{\theta}_{bis} = \arg \min_{\theta \in \mathbb{R}^d} \max_{\alpha \in \mathbb{R}^d} \hat{\mathbb{E}}_n[\alpha^\top \psi(Z) (Y - \theta^\top \varphi(X)) - 0.5 \cdot (\alpha^\top \psi(Z))^2] + \lambda \|\theta\|_{\ell_2}^2.$$

Albeit looking different, solving this min-max problem in closed-form leads to $\hat{\theta}_{bis} = \hat{\theta}$ (when we introduce regularization in Eq. (5)), and thus $\hat{h}_{bis} = \hat{h}$. The saddle-point form could be useful for deriving convergence rates, see [2] for example, as it allows us to use the theory of saddle-point problems. However, it is not necessary for the implementation of the algorithm.

3.2 Convergence rates for Sieve 2SLS

In order to present convergence rates for Sieve 2SLS, we introduce the sieve measure of ill-posedness.

Definition 1. Given $d \geq 1$ and $\mathcal{H}_{\varphi,d} = \text{span}\{\varphi_1, \dots, \varphi_d\}$, the sieve measure of ill-posedness is defined as

$$\tau_{\varphi,d} \doteq \sup_{h \in \mathcal{H}_{\varphi,d}, h \neq 0} \frac{\|h\|_{L_2(X)}}{\|\mathcal{T}h\|_{L_2(Z)}} = \left(\inf_{h \in \mathcal{H}_{\varphi,d}, \|h\|_{L_2}=1} \|\mathcal{T}h\|_{L_2(Z)} \right)^{-1}.$$

This quantity can be interpreted as the operator norm of \mathcal{T}^{-1} restricted to $\mathcal{T}(\mathcal{H}_{\varphi,d})$; that is, $\tau_{\varphi,d} = \sup_{g \in \mathcal{T}(\mathcal{H}_{\varphi,d}), g \neq 0} \|\mathcal{T}^{-1}g\|_{L_2(X)} / \|g\|_{L_2(Z)}$. While the full operator \mathcal{T}^{-1} is typically unbounded, the restricted norm $\tau_{\varphi,d}$ remains finite. It captures the degree of ill-posedness specific to the chosen hypothesis space $\mathcal{H}_{\varphi,d}$. As d increases and the hypothesis class becomes richer, $\tau_{\varphi,d}$ increases. Equipped with this quantity, we now introduce stability conditions that will allow us to establish generalization bounds for Sieve 2SLS.

Assumption 3. (i) $\sup_{h \in \mathcal{H}_{\varphi,d}, \|h\|_{L_2}=1} \|(\Pi_{\psi,d})_{\perp} \mathcal{T}h\|_{L_2(Z)} = o_d(\tau_{\varphi,d}^{-1})$; (ii) there is a constant $C > 0$ such that $\|\mathcal{T}(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(Z)} \leq C \times \tau_{\varphi,d}^{-1} \times \|(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(X)}$.

Assumption 3-(i) is a condition on the approximation properties of the features used for the instrument space. It says the image of $\mathcal{H}_{\varphi,d}$ through \mathcal{T} lies almost entirely inside $\mathcal{H}_{\psi,d}$, i.e., the projection onto the orthogonal complement vanishes faster than the inverse ill-posedness. Assumption 3-(ii) is a stability condition sometimes referred as **link condition** that is standard in the NPIV literature (see Assumption 6 in [3], Assumption 5.2(ii) in [7] and Assumption 3 in [23]). It says that \mathcal{T} becomes increasingly contractive on the component of h_0 lying outside the sieve space as the sieve dimension increases. The next assumption ensures the conditional variance of the noise is uniformly bounded.

Assumption 4. There exists $\bar{\sigma} > 0$ such that $\mathbb{E}[U^2 \mid Z] \leq \bar{\sigma}^2 < \infty$ almost everywhere.

The following result is an upper bound on the generalization error from [6].

Theorem 1 (Theorem B.1. [6]). Let \hat{h} be the 2SLS estimator from Eq. (5), using features $\varphi : \mathcal{X} \rightarrow \mathbb{R}^d$ and $\psi : \mathcal{Z} \rightarrow \mathbb{R}^d$. Suppose Assumptions 1, 3 and 4 hold and $\tau_{\varphi,d} \zeta_{\varphi,\psi,d} \sqrt{(\log d)/n} = o(1)$, where $\zeta_{\varphi,\psi,d} \doteq \max\{\sup_x \|\mathbb{E}[\varphi(X)\varphi(X)^{\top}]^{-1/2}\varphi(x)\|_{\ell_2}, \sup_z \|\mathbb{E}[\psi(Z)\psi(Z)^{\top}]^{-1/2}\psi(z)\|_{\ell_2}\}$. Then:

$$\|\hat{h} - h_0\|_{L_2(X)} = O_p\left(\|(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(X)} + \tau_{\varphi,d} \sqrt{d/n}\right).$$

The first term captures the **approximation error**, while the second term corresponds to the **estimation error** arising from finite samples. When standard bases such as cosine, spline, or wavelet functions are used, the approximation error is well-understood. If the sieve ill-posedness $\tau_{\varphi,d}$ grows at a polynomial or exponential rate in d , one can choose $d = d(n)$ to balance the two terms optimally [3]. Under such growth conditions and smoothness assumptions on h_0 (e.g., h_0 belongs to a Sobolev ball), the resulting convergence rate is minimax optimal [8].

3.3 Convergence rates for Sieve 2SLS with spectral features

A drawback of using non-adaptive features (such as cosine or spline bases) is that verifying Assumption 3 or characterizing the growth of $\tau_{\varphi,d}$ is generally difficult. In contrast, we now consider features constructed from the SVD of \mathcal{T} , for which both conditions can be exactly characterized. Moreover, while estimators based on universal bases can achieve minimax optimality under smoothness assumptions on \mathcal{T} and h_0 , they are often outperformed in practice by estimators using data-dependent features. This highlights that minimax rates, though important, do not capture the full picture of empirical performance. Recall from Eq. (3) the singular value decomposition of \mathcal{T} , and the definition of the top- d right eigenspace \mathcal{V}_d . The following result shows that this subspace achieves the minimal possible measure of ill-posedness among all d -dimensional subspaces of $L_2(X)$.

Proposition 2 (Lemma 1 [3]). Let Assumption 2 hold. For any $d \geq 1$, the smallest possible value of $\tau_{\varphi,d}$ is $\tau_{\varphi,d} = \sigma_d^{-1}$, achieved when $\mathcal{H}_{\varphi,d} = \mathcal{V}_d$.

Definition 2 (Spectral features). We say that the features φ and ψ are **spectral features** if

$$\mathcal{H}_{\varphi,d} = \text{span}\{v_1, \dots, v_d\} = \mathcal{V}_d \quad \text{and} \quad \mathcal{H}_{\psi,d} = \text{span}\{u_1, \dots, u_d\} = \mathcal{U}_d.$$

By Proposition 2, spectral features minimize the sieve measure of ill-posedness, and therefore the estimation error term. This leads to the following corollary.

Corollary 1 (Sieve 2SLS with spectral features). *Let \hat{h} be the 2SLS estimator from Eq. (5) using spectral features. Let Assumptions 1, 2 and 4 hold and $\sigma_d^{-1} \zeta_{\varphi, \psi, d} \sqrt{(\log d)/n} = o(1)$. Then:*

$$\|\hat{h} - h_0\|_{L^2(X)} = O_p \left(\|(\Pi_{\mathcal{X}, d})_{\perp} h_0\|_{L^2(X)} + \sqrt{\frac{d}{n\sigma_d^2}} \right).$$

We postpone the proof to Appendix C.2.

The good, the bad, and the ugly. The bound in Corollary 1 decomposes the generalization error into two terms: *spectral alignment*, encoded in the approximation error $\|(\Pi_{\mathcal{X}, d})_{\perp} h_0\|$, which measures how well h_0 lies in the top- d singular space; and *ill-posedness*, captured by the inverse singular value σ_d^{-1} . In the *good* regime, h_0 lies mostly in \mathcal{U}_d and σ_d decays slowly, so both terms are small. In the *bad* regime, alignment is favorable but fast spectral decay inflates the estimation error. In the *ugly* regime, h_0 is misaligned with the top eigenspaces, so the approximation error dominates and estimation fails regardless of σ_d or n . In Appendix B, we show that the bound in Corollary 1 is tight under a source condition and a singular value decay condition.

4 Sieve 2SLS with learned spectral features

In the previous section, we showed that spectral features minimize the sieve measure of ill-posedness and lead to fast generalization rates when the structural function h_0 aligns well with the top eigenspaces of the conditional expectation operator \mathcal{T} . This motivates learning such features directly from data, especially in the case where the conditional distribution of X given Z is informative (*i.e.*, the instrument is strong) and the alignment with the spectrum of \mathcal{T} is favorable. We now demonstrate that the contrastive learning approach of [36] can be equivalently seen as explicitly targeting the leading eigenspaces of \mathcal{T} . Consider the following objective:

$$\mathcal{L}_d(\varphi, \psi) \doteq \|\mathcal{T}_d(\varphi, \psi) - \mathcal{T}\|_{\text{HS}}^2, \quad \mathcal{T}_d(\varphi, \psi) \doteq \sum_{i=1}^d \psi_i \otimes \varphi_i, \quad (6)$$

where $\varphi_i \in L_2(X)$, $\psi_i \in L_2(Z)$, $i = 1, \dots, d$. As any rank- d operator can be decomposed as $\mathcal{T}_d(\varphi, \psi)$, we have the following version of the Eckart–Young–Mirsky Theorem for the Hilbert–Schmidt norm:

Theorem 2 (Eckart–Young–Mirsky Theorem). *Let Assumption 2 hold and let $d \geq 1$ be such that $\sigma_d > \sigma_{d+1}$. For all feature maps $\varphi : \mathcal{X} \rightarrow \mathbb{R}^d$, $\psi : \mathcal{Z} \rightarrow \mathbb{R}^d$, we have*

$$\mathcal{L}_d(\varphi, \psi) \geq \|\mathcal{T} - \mathcal{T}_d\|_{\text{HS}}^2 = \sum_{i>d} \sigma_i^2,$$

where equality holds if and only if $\mathcal{T}_d(\varphi, \psi) = \mathcal{T}_d$, in which case $\mathcal{H}_{\varphi, d} = \mathcal{V}_d$, $\mathcal{H}_{\psi, d} = \mathcal{U}_d$. \mathcal{T}_d is the rank- d truncation of \mathcal{T} introduced in Section 2.

While the Eckart–Young–Mirsky objective in Eq. (6) offers a clean objective to learn spectral features, it appears impractical due to the unknown nature of the operator \mathcal{T} . However, this objective admits an equivalent formulation as a spectral contrastive learning loss [37]:

$$\mathcal{L}_d(\varphi, \psi) = \mathbb{E}_X \mathbb{E}_Z \left[(\varphi(X)^\top \psi(Z))^2 \right] - 2\mathbb{E}_{X, Z} [\varphi(X)^\top \psi(Z)] + \|\mathcal{T}\|_{\text{HS}}^2, \quad (7)$$

where the last term is independent of (φ, ψ) . We provide a proof of this equivalence in Appendix C.3. The population loss can be estimated from samples $\tilde{\mathcal{D}}_m = \{(\tilde{x}_i, \tilde{z}_i)\}_{i=1}^m$ as

$$\hat{\mathcal{L}}_d(\varphi, \psi) \doteq \frac{1}{m(m-1)} \sum_{i \neq j} (\varphi(\tilde{x}_i)^\top \psi(\tilde{z}_j))^2 - \frac{2}{m} \sum_{i=1}^m \varphi(\tilde{x}_i)^\top \psi(\tilde{z}_i). \quad (8)$$

The spectral contrastive loss is the basis of the recent algorithm proposed in [36], where φ and ψ are parametrized by neural networks and optimized via stochastic gradient descent. While [36] motivate this objective heuristically, Eq. (6) shows that it directly targets the leading singular structure of \mathcal{T} . The use of objectives similar to Eq. (6) for operator learning or Eq. (7) for representation learning has a rich history that we detail in Appendix D.

Comparison to [36]. To justify the use of the contrastive loss in Eq. (8), [36] assume that the density $p(x, z)$ factorizes as

$$p(x, z) = \varphi(x)^\top \psi(z), \quad \varphi : \mathcal{X} \rightarrow \mathbb{R}^d, \quad \psi : \mathcal{Z} \rightarrow \mathbb{R}^d. \quad (9)$$

We argue that this assumption is overly strong: it holds if and only if \mathcal{T} has rank at most d , in which case the spectral features perfectly capture the structure of \mathcal{T} . This result is formalized in Proposition 5, Appendix C.3. We should instead interpret Eq. (9) as an approximation rather than a literal assumption. In practice, the learned features φ, ψ provide a rank- d approximation of \mathcal{T} , and the quality of this approximation governs the performance of the resulting Sieve 2SLS estimator.

Let $\hat{\varphi}, \hat{\psi}$ be d -dimensional features trained with loss Eq. (8) with \tilde{D}_m and define $\hat{\mathcal{T}}_d \doteq \sum_{i=1}^d \hat{\psi}_i \otimes \hat{\varphi}_i$. We make the following assumption:

Assumption 5. $\{\hat{\psi}_1, \dots, \hat{\psi}_d\}$ and $\{\hat{\varphi}_1, \dots, \hat{\varphi}_d\}$ form linearly independent families.

This assumption is only made to simplify the discussion. If $\{\hat{\psi}_1, \dots, \hat{\psi}_d\}$ or $\{\hat{\varphi}_1, \dots, \hat{\varphi}_d\}$ is not linearly independent, one can replace d in the following by the linear dimension of the family. Consider $\sigma_d > \sigma_{d+1}$ so that the unique minimizer of the population loss is given by \mathcal{T}_d . We quantify how the distance between $\hat{\mathcal{T}}_d$ and \mathcal{T}_d affects the approximation error and the sieve measure of ill-posedness $\tau_{\hat{\varphi}, d}$. We quantify how the distance between $\hat{\mathcal{T}}_d$ and \mathcal{T}_d affects the generalization error of the 2SLS estimator.

Theorem 3. Let Assumptions 1, 2, 4 and 5 hold and let $\hat{\varepsilon}_d \doteq \|\hat{\mathcal{T}}_d - \mathcal{T}_d\|_{\text{op}}$ be such that $\hat{\varepsilon}_d < (1 - 1/\sqrt{2})\sigma_d$.

- i) $\sigma_d^{-1} \leq \tau_{\hat{\varphi}, d} \leq (\sigma_d - 2\hat{\varepsilon}_d)^{-1}$;
- ii) Let \hat{h} be the 2SLS estimator from Eq. (5), using features $\hat{\varphi} : \mathcal{X} \rightarrow \mathbb{R}^d$ and $\hat{\psi} : \mathcal{Z} \rightarrow \mathbb{R}^d$. Suppose $(\sigma_d - 2\hat{\varepsilon}_d)^{-1} \zeta_{\hat{\varphi}, \hat{\psi}, d} \sqrt{(\log d)/n} = o(1)$ and $\hat{\varepsilon}_d = o_d(\sigma_d^2)$, then

$$\|\hat{h} - h_0\|_{L^2(X)} = O_p \left(\|(\Pi_{\mathcal{X}, d})^\perp h_0\|_{L_2(X)} + \frac{\sqrt{2}\hat{\varepsilon}_d \|h_0\|_{L_2(X)}}{\sigma_d} + \frac{1}{\sigma_d - 2\hat{\varepsilon}_d} \sqrt{\frac{d}{n}} \right).$$

We postpone the proof to Appendix C.3. The result quantifies how close the learned features must be to \mathcal{T}_d in order to preserve the desirable properties of spectral 2SLS. Theorem 3-i) shows that as long as the learned operator $\hat{\mathcal{T}}_d$ is a good approximation of the true rank- d truncation \mathcal{T}_d , the learned features inherit a favorable measure of ill-posedness. Theorem 3-ii) further ensures that if contrastive learning yields a near-optimal rank- d approximation of \mathcal{T} , the downstream Sieve 2SLS estimator retains strong statistical guarantees. Importantly, the spectral features generalization error $\hat{\varepsilon}_d$ encapsulates the statistical and optimization errors in learning spectral features. It depends on the number of samples used to train the contrastive loss, the expressivity and architecture of the neural networks parametrizing φ and ψ , and the complexity of the leading eigensubspaces of \mathcal{T} . As $\mathcal{L}_d(\varphi, \psi)$ converges to $\mathcal{L}_d^{\min} \doteq \|\mathcal{T} - \mathcal{T}_d\|_{\text{HS}}^2$, $\hat{\varepsilon}_d$ converges to 0 by Theorem 2. Precisely controlling $\hat{\varepsilon}_d$ lies beyond the scope of this paper and remains an important question for future work.

5 Estimating alignment and spectral decay from data

Having established the good-bad-ugly taxonomy based on spectral properties and its interaction with the structural function h_0 , we now discuss how to estimate these spectral properties from data. Namely, we present a methodology to estimate the spectral decay of the operator \mathcal{T} as well as spectral alignment with the structural function h_0 . Recall that \mathcal{T} is compact and admits an SVD of the form Eq. (3), with the key relationship that $\sigma_i v_i = \mathcal{T}^* u_i$. Therefore, for all $i \geq 1$, we have

$$\langle v_i, h_0 \rangle_{L_2(X)} = \frac{1}{\sigma_i} \langle \mathcal{T}^* u_i, h_0 \rangle_{L_2(X)} = \frac{1}{\sigma_i} \langle u_i, r_0 \rangle_{L_2(Z)} = \frac{1}{\sigma_i} \mathbb{E}[Y \cdot u_i(Z)], \quad (10)$$

where we used $\mathcal{T}h_0 = r_0$, the definition of r_0 and the tower property of the conditional expectation. This relationship shows that the alignment coefficients $\langle v_i, h_0 \rangle_{L_2(X)}$ can be estimated from data if we have estimates of the singular functions u_i and singular values σ_i . In Section 5, we provide a practical estimator for Eq. (10) using learned features and derive its estimation guarantees, which depend on the operator error $\hat{\varepsilon}_d$.

We apply this procedure in Section 6 to diagnose the spectral properties of the dSprites dataset [30], a popular benchmark in the nonparametric IV and proxy literature [40, 41, 36].

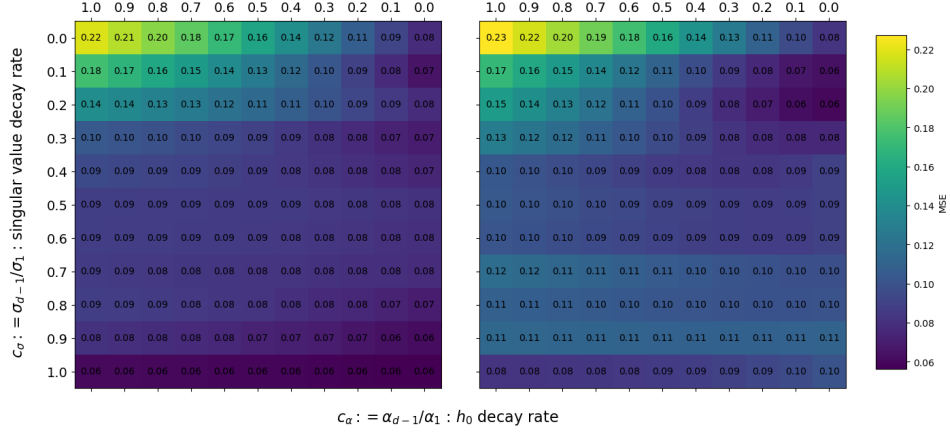


Figure 1: Dependence of the MSE of instrumental variable regression on the decay rates of the spectrum and coefficients of the structural function. IV fitting and the MSE evaluations were repeated 500 times per parameter set, rendering the standard error of these estimates negligible. **Left:** Oracle spectral features; **Right:** Learned features.

6 Experiments

6.1 Synthetic datasets

We evaluate the main theoretical insight of the paper: Sieve 2SLS with spectral features performs best when h_0 aligns with the top eigenspaces of \mathcal{T} and its singular values decay slowly. Performance degrades with faster decay and weaker alignment. We design a synthetic NPIV setting where we control the conditional expectation operator \mathcal{T} and its spectral decay and well as the alignment of h_0 with its eigenspaces. To simulate such a setting, we rely on the following procedure for generating samples from the NPIV model:

Proposition 3. *Let π_{XZ} be a probability distribution on $\mathcal{X} \times \mathcal{Z}$ with marginals π_X and π_Z . Let $\mathcal{T} : L_2(\mathcal{X}, \pi_X) \rightarrow L_2(\mathcal{Z}, \pi_Z)$ be the conditional expectation operator associated with π_{XZ} and let $h_0 \in L_2(\mathcal{X}, \pi_X)$. Sample $(X, Z) \sim \pi_{XZ}$, $V \sim \mathcal{N}(0, \sigma^2)$ and set $U = \mathcal{T}h_0(Z) - h_0(X) + V$ and $Y = h_0(X) + U$. Then (U, Z, X, Y) is a sample from the NPIV model of Eq. (1).*

Let $d = 11$, $\mathcal{X} = [0, 2\pi]$, $\mathcal{Z} = [0, 2\pi]$, $\sigma^2 = 0.1$, $f(x) = (\sin(x), \sin(2x), \dots, \sin((d-1)x))$, $g(z) = (\sin(z), \sin(2z), \dots, \sin((d-1)z))$, and let P, Q be two orthogonal $(d-1) \times (d-1)$ matrices. Define $v : \mathcal{X} \rightarrow \mathbb{R}^{d-1}$ and $u : \mathcal{Z} \rightarrow \mathbb{R}^{d-1}$ as $v(x) = Pf(x)$ and $u(z) = Qg(z)$. We set $\mathcal{T} = \mathbb{1}_{\mathcal{Z}} \otimes \mathbb{1}_{\mathcal{X}} + \sum_{i=1}^{d-1} \sigma_i u_i \otimes v_i$. (X, Z) is jointly sampled with rejection sampling. Z, X are uniform on $[0, 2\pi]$ and X given Z admits \mathcal{T} as conditional expectation operator. We then sample Y following Proposition 3. The multiplication of the trigonometric functions by the orthogonal matrices P, Q ensures that the consecutive singular functions of \mathcal{T} are comparably complex. By keeping the difficulty constant per singular function we can isolate the influence of the singular value decay on the performance of the final estimator. We set the singular values $\{\sigma_i\}_{i=1}^{d-1}$ to decay linearly, from some initial value σ_1 to a $\sigma_{d-1} = c_\sigma \sigma_1$, $c_\sigma \in [0, 1]$. To control spectral alignment, we set $h_0 = \sum_{i=1}^{d-1} \alpha_i v_i$ subject to $\|\alpha\|_{\ell_2} = 1$, and allow $\{\alpha_i\}_{i=1}^{d-1}$ to decay linearly from α_1 to $\alpha_{d-1} = c_\alpha \alpha_1$, $c_\alpha \in [0, 1]$. Each value of $c_\alpha \in [0, 1]$ controls the alignment of h_0 with the top singular functions of \mathcal{T} : larger values correspond to weaker alignment. Similarly, $c_\sigma \in [0, 1]$ controls the ill-posedness of the problem: smaller values indicate faster decay of singular values and hence more severe ill-posedness.

In Figure 1, we report the mean squared error (MSE) of two 2SLS estimators across a range of spectral decay rates (c_σ) and structural function decay rates (c_α). The left panel uses 2SLS with fixed features f and g that form the exact spectral features. This represents an oracle setting. The MSE decreases from top to bottom as the problem becomes better conditioned (i.e., as c_σ increases). Additionally, the MSE decreases from left to right, as smaller values of c_α correspond to stronger

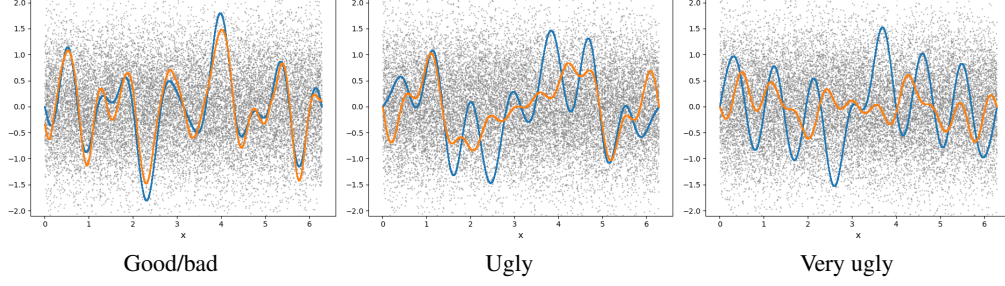


Figure 2: Comparison of the spectral methods’ performance depending on the alignment of h_0 with the singular functions of \mathcal{T} . **Blue:** True h_0 ; **Orange:** 2SLS estimate.

spectral alignment: more of h_0 is concentrated on the top singular functions of \mathcal{T} . These trends are fully consistent with the bound established in Corollary 1. In the right panel, we learn 50 features from data by minimizing the empirical contrastive loss Eq. (8) using a neural network. We then build a 2SLS estimator on top of the learned features. The same qualitative trends are observed: MSE decreases with slower singular value decay and better alignment. Moreover, when c_σ is small (top rows), learning the entire eigenspaces becomes difficult. In this regime, performance is most sensitive to c_α . In particular, if h_0 is concentrated on the top singular functions (small c_α), good recovery is still possible. In contrast, when c_σ is large (bottom rows), the singular functions are equally important and more easily learned, rendering the alignment of h_0 less critical. The mismatch between the oracle and learned settings is because we do not exactly recover the singular feature spaces; as predicted by Theorem 3. It is also worth noting that when h_0 is concentrated on high singular values, it may sometimes be beneficial for the singular values to decay quickly, by making it easier for the feature-learning model to learn the top singular functions well. This may explain the slight top-bottom increase in losses on the right end of the right panel.

The previous experiment assumes h_0 lies in the span of the singular functions of \mathcal{T} . Figure 2 illustrates three regimes. The left panel shows a well-aligned case (“good/bad”), where h_0 is entirely supported on the singular functions of \mathcal{T} , allowing for exact recovery as long as enough data is available. The middle panel represents a “partial alignment” regime, where h_0 has some overlap with the singular functions; in this case, recovery is imperfect, but the estimator can still capture some signal. Finally, the right panel illustrates the “misaligned” or “very ugly” regime, where h_0 is orthogonal to the singular functions. The instrument provides no information about the signal, and all spectral methods fail. We refer to Appendix E for further discussion of these scenarios and the experimental setup.

6.2 Dsprites dataset

The dSprites dataset, as employed in the evaluation of IV models, consists of 64×64 noisy images of hearts, with varying x, y positions, orientations and sizes. The full setting is as follows; $X \in \mathbb{R}^{64 \times 64}$ is the raw image, $Z \in \mathbb{R}^3$ consists of the x position, orientation and scale of the heart in the image, and the output is constructed as⁵

$$Y = h_0(X) + 32 \cdot (y \text{ position} - 32) + U, U \sim \mathcal{N}(0, 1),$$

where

$$h_0(X) = (\|A \circ X\|^2 - 3000)/500, A_{ij} = |32 - j|/32.$$

It was observed by Sun et al. [36] that spectral methods significantly outperform alternatives in this setting. We argue that it is because the structural function employed in this benchmark lies in the “good” regime. Utilising the alignment estimation methods outlined in Section 5, we are able to provide evidence for this claim. We observe that the structural function is spanned by the leading singular functions of the conditional expectation operator associated with this model. The details of this experiment can be found in Appendix F.

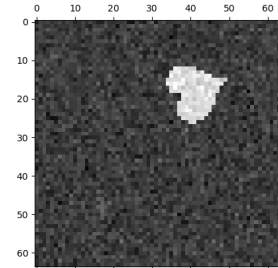


Figure 3: Example image from the dSprites dataset.

⁵The original formulation of the dSprites model for IV used $A_{i,j} \sim \mathcal{U}(0, 1)$ [40]. However, it is noted in [41] (arxiv version 18/06/2024, Appendix E.3) that it leads to identifiability issues and that the choice $A_{ij} = |32 - j|/32$ is preferred.

Acknowledgements

D.M., J.W., and A.G. are supported by the Gatsby Charitable Foundation. V.R.K. is supported from NextGenerationEU and MUR PNRR project PE0000013 CUP J53C22003010006 “Future Artificial Intelligence Research (FAIR)”. A.M. has received funding from the European Research Council (ERC), under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 950180), and travel support from ELSA (Project ID: 101070617).

References

- [1] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.
- [2] Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Minimax instrumental variable regression and l_2 convergence guarantees without identification or closedness. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2291–2318. PMLR, 2023.
- [3] Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.
- [4] Vladimir Igorevich Bogachev and Maria Aparecida Soares Ruas. *Measure theory, Vol. II*. Springer, 2007.
- [5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a “siamese” time delay neural network. *Advances in Neural Information Processing Systems*, 6, 1993.
- [6] Xiaohong Chen and Timothy M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.
- [7] Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.
- [8] Xiaohong Chen and Markus Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(3):497–521, 2011.
- [9] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.
- [10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, volume 1, pages 539–546. IEEE, 2005.
- [11] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.
- [12] Zhijie Deng, Jiaxin Shi, and Jun Zhu. Neuraief: Deconstructing kernels by deep neural networks. In *International Conference on Machine Learning*, pages 4976–4992. PMLR, 2022.
- [13] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33, 2020.
- [14] Heinz Werner Engl, Martin Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 2000.
- [15] Jean-Pierre Florens, Jan Johannes, and Sébastien Van Belleghem. Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, 27(3):472–496, 2011.
- [16] Yuri R. Fonseca, Caio F. L. Peixoto, and Yuri F. Saporito. Nonparametric instrumental variable regression through stochastic approximate gradients. *Advances in Neural Information Processing Systems*, 37, 2024.

- [17] M.J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, 1984. ISBN 9780122990502. URL <https://books.google.es/books?id=LsPaAAAAAAAJ>.
- [18] Peter Hall and Joel L Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33(6):2904–2929, 2005.
- [19] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.
- [20] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.
- [21] Hsiang Hsu, Salman Salamatian, and Flavio P Calmon. Correspondence analysis using neural networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2671–2680. PMLR, 2019.
- [22] Hsiang Hsu, Salman Salamatian, and Flavio P Calmon. Generalizing correspondence analysis for applications in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9347–9362, 2021.
- [23] Juno Kim, Dimitri Meunier, Arthur Gretton, Taiji Suzuki, and Zhu Li. Optimality and adaptivity of deep neural features for instrumental variable regression. *arXiv preprint arXiv:2501.04898*, 2025.
- [24] Vladimir Kostic, Karim Lounici, Pietro Novelli, and Massimiliano Pontil. Sharp spectral rates for koopman operator learning. *Advances in Neural Information Processing Systems*, 36, 2023.
- [25] Vladimir Kostic, Grégoire Pacreau, Giacomo Turri, Pietro Novelli, Karim Lounici, and Massimiliano Pontil. Neural conditional probability for uncertainty quantification. *Advances in Neural Information Processing Systems*, 37, 2024.
- [26] Henry Oliver Lancaster. The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29(3):719–736, 1958.
- [27] Sergio Leone. The good, the bad and the ugly. *United Artists*, 1966.
- [28] Zihao Li, Hui Lan, Vasilis Syrgkanis, Mengdi Wang, and Masatoshi Uehara. Regularized deepiv with model selection. *arXiv preprint arXiv:2403.04236*, 2024.
- [29] Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, and Zhaoran Wang. Provably efficient neural estimation of structural equation models: An adversarial approach. *Advances in Neural Information Processing Systems*, 33, 2020.
- [30] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. <https://github.com/deepmind/dsprites-dataset/>, 2017.
- [31] Dimitri Meunier, Zhu Li, Tim Christensen, and Arthur Gretton. Nonparametric instrumental regression via kernel methods is minimax optimal. *arXiv preprint arXiv:2411.19653*, 2024.
- [32] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.
- [33] Ieva Petrulionyte, Julien Mairal, and Michael Arbel. Functional bilevel optimization for machine learning. *Advances in Neural Information Processing Systems*, 37, 2024.
- [34] Daqian Shao, Ashkan Soleymani, Francesco Quinzan, and Marta Kwiatkowska. Learning decision policies with instrumental variables through double machine learning. *arXiv preprint arXiv:2405.08498*, 2024.
- [35] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

- [36] Haotian Sun, Antoine Moulin, Tongzheng Ren, Arthur Gretton, and Bo Dai. Spectral representation for causal estimation with hidden confounders. *arXiv preprint arXiv:2407.10448*, 2024.
- [37] Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, and Russ R Salakhutdinov. Neural methods for point-wise dependency estimation. *Advances in Neural Information Processing Systems*, 33, 2020.
- [38] Ziyu Wang, Yucen Luo, Yueru Li, Jun Zhu, and Bernhard Schölkopf. Spectral representation learning for conditional moment models. *arXiv preprint arXiv:2210.16525*, 2022.
- [39] Ziyu Wang, Yuhao Zhou, and Jun Zhu. Fast instrument learning with faster rates. *Advances in Neural Information Processing Systems*, 35:16596–16611, 2022.
- [40] Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2020.
- [41] Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *arXiv preprint arXiv:2106.03907*, 2021.
- [42] Xiangxiang Xu, Shao-Lun Huang, Lizhong Zheng, and Gregory W Wornell. An information theoretic interpretation to deep neural networks. *Entropy*, 24(1):135, 2022.
- [43] Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1): 20220073, 2023.
- [44] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594, 2020.

Supplementary Material: Demystifying Spectral Feature Learning for Instrumental Variable Regression

A Background on Spectral Theory

For completeness, we briefly recall key notions from spectral theory used throughout the paper. Our goal is to provide intuition and the minimal mathematical tools required to interpret concepts such as compactness, singular value decomposition (SVD) of operators, and eigensubspaces.

Let H_1 and H_2 be separable Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_{H_i}$ and norms $\| \cdot \|_{H_i}$, $i = 1, 2$.

Hilbert spaces and bounded operators. A linear map $A : H_1 \rightarrow H_2$ is said to be *bounded* if there exists $C > 0$ such that $\|Ah\|_{H_2} \leq C\|h\|_{H_1}$ for all $h \in H_1$. The smallest such constant is the *operator norm*, denoted $\|A\|_{\text{op}}$.

Adjoint and self-adjoint operators. Every bounded operator $A : H_1 \rightarrow H_2$ admits an *adjoint* $A^* : H_2 \rightarrow H_1$ defined by $\langle Ah, g \rangle_{H_2} = \langle h, A^*g \rangle_{H_1}$. If $H_1 = H_2$ and $A = A^*$, the operator is *self-adjoint*.

Compact operators. A linear operator A is called *compact* if it maps bounded sets in H_1 to relatively compact sets in H_2 , that is, the image of any bounded sequence has a convergent subsequence in H_2 . All compact operators are bounded, but not all bounded operators are compact. Any compact operator is a limit (in operator norm) of finite-rank operators, so that the class of compact operators can be defined alternatively as the closure of the set of finite-rank operators in the operator norm topology. The conditional expectation operator \mathcal{T} is a canonical example of a compact operator (see Proposition 1 below).

Hilbert–Schmidt operators and norm. Let $A : H_1 \rightarrow H_2$ be a bounded linear operator. Choose any orthonormal basis $\{e_j\}_{j \geq 1}$ of H_1 . The operator A is called *Hilbert–Schmidt* if

$$\|A\|_{\text{HS}}^2 := \sum_{j \geq 1} \|Ae_j\|_{H_2}^2 < \infty.$$

The value of the sum above is independent of the chosen basis of H_1 . Hence the quantity $\langle A, B \rangle_{\text{HS}} := \sum_{j \geq 1} \langle Ae_j, Be_j \rangle_{H_2}$ defines an inner product on the space of Hilbert–Schmidt operators, and $\| \cdot \|_{\text{HS}}$ is the induced norm.

Singular Value Decomposition (SVD). When $A : H_1 \rightarrow H_2$ is compact (not necessarily self-adjoint), it admits a *singular value decomposition*:

$$A = \sum_{i \geq 1} \sigma_i u_i \otimes v_i,$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq 0$ are the singular values of A such that $\sigma_i \rightarrow 0$, and $\{v_i\} \subset H_1$, $\{u_i\} \subset H_2$ form orthonormal bases satisfying $Av_i = \sigma_i u_i$ and $A^*u_i = \sigma_i v_i$. This generalizes the matrix SVD to infinite-dimensional spaces. Important properties we use below:

- The sum above is finite if and only if A has finite rank.
- $\|A\|_{\text{op}} = \sigma_1$.
- If A is Hilbert–Schmidt then A is compact such that $\|A\|_{\text{HS}}^2 = \sum_i \sigma_i^2 < \infty$, hence $\|A\|_{\text{op}} \leq \|A\|_{\text{HS}}$. Thus Hilbert–Schmidt operators are bounded, but the converse need not hold.

B Learning Rate with Spectral Features

In this section, we formalize the learning rate that can be obtained from Corollary 1. To do so, we employ two standard assumptions from the nonparametric IV literature [8]: a source condition and a singular value decay assumption on the operator \mathcal{T} .

Assumption 6 (Singular Value Decay). *Let $p \geq 1$. There are constants $0 < c_1 \leq c_2 < \infty$ such that for all $i \geq 1$, $c_1 i^{-p} \leq \sigma_i \leq c_2 i^{-p}$.*

This assumption is standard in the nonparametric IV literature [8] and characterizes the difficulty of the inverse problem defined by the operator \mathcal{T} . A larger p indicates faster decay and thus greater ill-posedness. Alternative ill-posedness characterizations can be considered, such as exponential decay [8].

Assumption 7 (Source Condition). *Let $r > 0$. There is a constant $B < \infty$ such that*

$$\|h_0\|_r^2 := \|(\mathcal{T}^* \mathcal{T})^{-r/2} h_0\|_{L^2(X)}^2 = \sum_{i \geq 1} \frac{\langle h_0, u_i \rangle_{L^2(X)}^2}{\sigma_i^{2r}} \leq B.$$

This condition quantifies how “smooth” h_0 is relative to the spectral decomposition of \mathcal{T} , or equivalently, how quickly the alignment coefficients $\langle h_0, u_i \rangle_{L^2(X)}$ decay relative to the singular values. A larger r implies h_0 is “smoother” or better aligned with the less ill-posed directions of \mathcal{T} . Note that under the “bad” scenario where the eigen-decay is fast (p is large), a stronger alignment is necessary for $\|h_0\|_r^2$ to be finite for a large value of r . While in the “good” scenario with $p \sim 1$, $\|h_0\|_r^2 < +\infty$ is more likely to hold with a large value of r .

Under Assumption 7, we have

$$\|(\Pi_{\mathcal{X},d})_{\perp} h_0\|_{L^2(X)}^2 = \sum_{i>d} \langle h_0, u_i \rangle_{L^2(X)}^2 \leq \sigma_d^{2r} \sum_{i>d} \frac{|\langle h_0, u_i \rangle_{L^2(X)}|^2}{\sigma_i^{2r}} \leq \sigma_d^{2r} \|h_0\|_r^2.$$

Combined with Assumption 6, we obtain from Corollary 1 that:

$$\|\hat{h} - h_0\|_{L^2(X)} = \mathcal{O}_p \left(\|h_0\|_r d^{-rp} + \sqrt{\frac{d^{1+2p}}{n}} \right).$$

Balancing both terms in d as a function of n , we obtain

$$\|\hat{h} - h_0\|_{L^2(X)} = \mathcal{O}_p \left(n^{-\frac{r}{2(r+1)+p-1}} \right),$$

with

$$d(n) = n^{\frac{1}{2p(r+1)+1}}.$$

This bound is tight as a matching lower bound is obtained in Theorem 1-(i) [8]. From this tight rate, we see that both the ill-posedness (p) and the relative smoothness (r) fundamentally capture the effectiveness of 2SLS with spectral features, and their impacts are inherently intertwined.

- A low relative smoothness (small r) means h_0 has significant components aligned with directions corresponding to very small singular values, even after accounting for the singular value decay. This also directly slows the rate, effectively rendering the problem very hard or even impossible ($r \rightarrow 0_+$ makes the rate vacuous, representing our “ugly” scenario).
- A fast decay (large p) directly slows down the rate regardless of r . It means singular values drop quickly, making it fundamentally harder to resolve higher-frequency components of h_0 . When p is large we are either in the bad or in the ugly scenario. A large p also slows down the rate indirectly by making it harder to achieve a high relative smoothness r .

When working with learned features. We see from Theorem 3 that when we learn the spectral features using contrastive learning, we incur an additional penalty related to $\hat{\varepsilon}_d/\sigma_d$ in the upper bound. This error shows that a faster eigendecay makes the 2SLS estimator more sensitive to errors in the learned features. Alignment of h_0 does not enter this factor, as learning the eigensubspaces of \mathcal{T} is unrelated to h_0 . Intuitively, features corresponding to small singular values can be empirically difficult to learn, even with a very large number of (Z, X) samples.

C Proofs

In this section we recall and prove the results of the main section.

C.1 Proofs of Section 2

We refer the reader to [4] for background on measure-theoretic notions used throughout this section. We assume that \mathcal{X} and \mathcal{Z} are standard Borel spaces, *i.e.*, Borel subsets of complete separable metric spaces. This ensures the existence of regular conditional distributions and supports the disintegration of the joint law $\pi_{X,Z}$. Working directly with measures allows us to treat both continuous settings, where distributions admit densities with respect to the Lebesgue measure, and discrete or more general settings, where such densities may not exist.

To prove properties of the conditional expectation operator \mathcal{T} , it is helpful to express it in terms of a Markov kernel. We define a Markov kernel $p : \mathcal{Z} \times \mathcal{F}_{\mathcal{X}} \rightarrow [0, 1]$ such that for any measurable set $A \in \mathcal{F}_{\mathcal{X}}$,

$$p(z, A) = \mathbb{P}[X \in A \mid Z = z],$$

meaning $p(z, \cdot)$ is a regular conditional distribution of X given $Z = z$. Then for all $h \in L_2(X)$,

$$\mathcal{T}h(z) = \int_{\mathcal{X}} h(x) p(z, dx), \quad \pi_Z\text{-a.e.} \quad (11)$$

By the disintegration theorem, for all measurable sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Z}$, the joint distribution admits the disintegration

$$\pi_{X,Z}(A \times B) = \int_B \left(\int_A p(z, dx) \right) d\pi_Z(z),$$

and therefore, the joint distribution $\pi_{X,Z}$ can be decomposed as

$$d\pi_{X,Z}(x, z) = p(z, dx) d\pi_Z(z), \quad (12)$$

where $p(z, dx)$ is a Markov kernel as defined above. This decomposition holds in general, even when $\pi_{X,Z}$ is not absolutely continuous with respect to $\pi_X \otimes \pi_Z$.

Under Assumption 2, the joint distribution $\pi_{X,Z}$ is absolutely continuous with respect to the product measure $\pi_X \otimes \pi_Z$, and thus admits a density

$$p(x, z) \doteq \frac{d\pi_{X,Z}}{d(\pi_X \otimes \pi_Z)}(x, z). \quad (13)$$

In order to prove Proposition 1, we use the following lemma.

Lemma 1. *Under Assumption 2, for $\pi_X \otimes \pi_Z$ -almost every (x, z) ,*

$$p(x, z) = \frac{dp(z, \cdot)}{d\pi_X}(x).$$

Proof. From Eq. (12), we have

$$d\pi_{X,Z}(x, z) = p(z, dx) d\pi_Z(z).$$

On the other hand, by Eq. (13),

$$d\pi_{X,Z}(x, z) = p(x, z) d\pi_X(x) d\pi_Z(z).$$

Comparing both expressions and using the uniqueness of the Radon–Nikodym derivative, we conclude that for $\pi_X \otimes \pi_Z$ -almost every (x, z) ,

$$p(z, dx) = p(x, z) d\pi_X(x),$$

i.e., $p(z, \cdot) \ll \pi_X$, and

$$\frac{dp(z, \cdot)}{d\pi_X}(x) = p(x, z).$$

This concludes the proof. \square

By Lemma 1, under Assumption 2, for all $h \in L_2(X)$ and π_Z -a.e.,

$$\mathcal{T}h(z) = \int_{\mathcal{X}} h(x) p(z, dx) = \int_{\mathcal{X}} h(x) p(x, z) d\pi_X(x) = \langle h, p(\cdot, z) \rangle_{L_2(X)}. \quad (14)$$

We recall and prove Proposition 1. This is a classical result, see, *e.g.*, [26].

Proposition 1. *Under Assumption 2, \mathcal{T} is a Hilbert–Schmidt operator and thus compact.*

Proof. Take $(e_i)_{i \geq 1}$ any orthonormal basis (ONB) of $L_2(X)$. Then, using the definition of the Hilbert–Schmidt norm:

$$\begin{aligned}
\|\mathcal{T}\|_{\text{HS}}^2 &= \sum_{i \geq 1} \|\mathcal{T}e_i\|_{L_2(Z)}^2 && \text{(definition of HS norm via ONB)} \\
&= \sum_{i \geq 1} \int_Z |\mathcal{T}e_i(z)|^2 d\pi_Z(z) && \text{(expand } L_2(Z) \text{ norm)} \\
&= \sum_{i \geq 1} \int_Z \langle e_i, p(\cdot, z) \rangle_{L_2(X)}^2 d\pi_Z(z) && \text{(by Eq. (14))} \\
&= \int_Z \sum_{i \geq 1} \langle e_i, p(\cdot, z) \rangle_{L_2(X)}^2 d\pi_Z(z) && \text{(Fubini's theorem)} \\
&= \int_Z \|p(\cdot, z)\|_{L_2(X)}^2 d\pi_Z(z) && \text{(Parseval's identity)} \\
&= \int_{\mathcal{X} \times \mathcal{Z}} p(x, z)^2 d\pi_X(x) d\pi_Z(z) && \text{(expand } L_2 \text{ norm)} \\
&< +\infty && \text{(by Assumption 2, since } p \in L_2(\pi_X \otimes \pi_Z)).
\end{aligned}$$

□

Note that the proof shows in addition that $\|\mathcal{T}\|_{\text{HS}} = \|p\|_{L_2(\pi_X \otimes \pi_Z)}$.

C.2 Proofs of Section 3

Proposition 4. *Let $\mathcal{H}_{\varphi, d} = \text{span}\{\varphi_1, \dots, \varphi_d\}$, and let $\omega = \dim(\mathcal{H}_{\varphi, d})$. Then:*

- i) $\tau_{\varphi, d} \geq \sigma_{\omega}^{-1}$.
- ii) If $\mathcal{H}_{\varphi, d} \subseteq \mathcal{V}_\ell = \text{span}\{v_1, \dots, v_\ell\}$, then $\tau_{\varphi, d} \leq \sigma_\ell^{-1}$.
- iii) For any $d \geq 1$, the minimal value $\tau_{\varphi, d} = \sigma_d^{-1}$ is achieved when $\mathcal{H}_{\varphi, d} = \mathcal{V}_d$.

Proposition 4 is due to Lemma 1 of [3] and Proposition 2 corresponds to part (iii). We provide a full proof for completeness.

Proof. We first prove (i). Since $\dim(\mathcal{H}_{\varphi, d}) = \omega$, and the subspace $\mathcal{V}_{\omega-1}$ has dimension $\omega - 1$, there exists $\tilde{h} \in \mathcal{H}_{\varphi, d} \cap \mathcal{V}_{\omega-1}^\perp$ with $\|\tilde{h}\|_{L_2(X)} = 1$. Then,

$$\begin{aligned}
\tau_{\varphi, d}^{-2} &= \inf_{\substack{h \in \mathcal{H}_{\varphi, d} \\ \|h\|_{L_2(X)} = 1}} \|\mathcal{T}h\|_{L_2(Z)}^2 \\
&\leq \|\mathcal{T}\tilde{h}\|_{L_2(Z)}^2 \\
&\leq \sup_{\substack{h \in \mathcal{V}_{\omega-1}^\perp \\ \|h\|_{L_2(X)} = 1}} \|\mathcal{T}h\|_{L_2(Z)}^2 \\
&= \sup_{\substack{h \in \mathcal{V}_{\omega-1}^\perp \\ \|h\| = 1}} \langle \mathcal{T}^* \mathcal{T}h, h \rangle_{L_2(X)} = \sigma_\omega^2,
\end{aligned}$$

where the last line follows from the min–max theorem for compact self-adjoint operators.

We now prove (ii). If $h \in \mathcal{H}_{\varphi, d} \subseteq \mathcal{V}_\ell$ and $\|h\|_{L_2(X)} = 1$, then

$$h = \sum_{i=1}^{\ell} \langle h, v_i \rangle v_i,$$

so by orthonormality of (u_i) ,

$$\|\mathcal{T}h\|_{L_2(Z)}^2 = \left\| \sum_{i=1}^{\ell} \sigma_i \langle h, v_i \rangle u_i \right\|_{L_2(Z)}^2 = \sum_{i=1}^{\ell} \sigma_i^2 \langle h, v_i \rangle^2 \geq \sigma_{\ell}^2 \sum_{i=1}^{\ell} \langle h, v_i \rangle^2 = \sigma_{\ell}^2.$$

Taking the infimum over unit-norm $h \in \mathcal{H}_{\varphi,d}$ gives $\tau_{\varphi,d}^{-2} \geq \sigma_{\ell}^2$, i.e., $\tau_{\varphi,d} \leq \sigma_{\ell}^{-1}$.

Finally, (iii) follows by taking $\mathcal{H}_{\varphi,d} = \mathcal{V}_d$, so that $\omega = d = \iota$, and both bounds in (i) and (ii) match. \square

We recall and prove Corollary 1:

Corollary 1 (Sieve 2SLS with spectral features). *Let \hat{h} be the 2SLS estimator from Eq. (5) using spectral features. Let Assumptions 1, 2 and 4 hold and $\sigma_d^{-1} \zeta_{\varphi,\psi,d} \sqrt{(\log d)/n} = o(1)$. Then:*

$$\|\hat{h} - h_0\|_{L_2(X)} = O_p \left(\|(\Pi_{\mathcal{X},d})_{\perp} h_0\|_{L_2(X)} + \sqrt{\frac{d}{n\sigma_d^2}} \right).$$

Proof. Let us verify that Assumption 3 holds under Definition 2, i.e., when $\mathcal{H}_{\varphi,d} = \mathcal{V}_d$ and $\mathcal{H}_{\psi,d} = \mathcal{U}_d$.

(i) Since the image of \mathcal{V}_d under \mathcal{T} lies in \mathcal{U}_d , we have

$$(\Pi_{\psi,d})_{\perp} \mathcal{T}h = 0 \quad \text{for all } h \in \mathcal{H}_{\varphi,d}.$$

Hence,

$$\sup_{\substack{h \in \mathcal{H}_{\varphi,d} \\ \|h\|_{L_2} = 1}} \|(\Pi_{\psi,d})_{\perp} \mathcal{T}h\|_{L_2(Z)} = 0 = o\left(\tau_{\varphi,d}^{-1}\right),$$

and condition (i) is satisfied.

(ii) Let $h_0 \in L_2(X)$, and consider its projection onto the orthogonal complement of $\mathcal{H}_{\varphi,d}$. Since \mathcal{T} is compact with singular value decomposition $\mathcal{T} = \sum_{i=1}^{\infty} \sigma_i u_i \otimes v_i$ under Assumption 2, we have

$$\|\mathcal{T}(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(Z)}^2 = \sum_{i>d} \sigma_i^2 \langle h_0, v_i \rangle^2 \leq \sigma_d^2 \sum_{i>d} \langle h_0, v_i \rangle^2 = \sigma_d^2 \|(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(X)}^2.$$

This implies, using that $\tau_{\varphi,d} = \sigma_d^{-1}$ by Proposition 2,

$$\|\mathcal{T}(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(Z)} \leq \sigma_d \|(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(X)} = \tau_{\varphi,d}^{-1} \|(\Pi_{\varphi,d})_{\perp} h_0\|_{L_2(X)},$$

so Assumption 3-(ii) holds with constant $C = 1$.

Since both parts of Assumption 3 hold, and $\tau_{\varphi,d} = \sigma_d^{-1}$ by Proposition 2, we may apply Theorem 1, which yields the desired result. \square

C.3 Proofs of Section 4

The following result shows that Eq. (9) holds if and only if $\text{rank}(\mathcal{T}) \leq d$.

Proposition 5. *Let $\varphi : \mathcal{X} \rightarrow \mathbb{R}^d$, $\psi : \mathcal{Z} \rightarrow \mathbb{R}^d$ with components $\varphi_i \in L_2(X)$, $\psi_i \in L_2(Z)$. Then*

$$p(x, z) = \varphi(x)^{\top} \psi(z) \quad \text{if and only if} \quad \mathcal{T} = \sum_{i=1}^d \psi_i \otimes \varphi_i.$$

Proof. \Rightarrow Assume $p(x, z) = \varphi(x)^{\top} \psi(z)$. Then, by Eq. (11), for all $h \in L_2(X)$,

$$\begin{aligned} \mathcal{T}h(z) &= \int h(x) p(z, dx) = \int h(x) p(x, z) d\pi_X(x) && \text{(by Lemma 1)} \\ &= \int h(x) \varphi(x)^{\top} \psi(z) d\pi_X(x) = \psi(z)^{\top} \int h(x) \varphi(x) d\pi_X(x) \\ &= \sum_{i=1}^d \psi_i(z) \langle h, \varphi_i \rangle_{L_2(X)} = \left(\sum_{i=1}^d \psi_i \otimes \varphi_i \right) (h)(z), \end{aligned}$$

so $\mathcal{T} = \sum_{i=1}^d \psi_i \otimes \varphi_i$.

\Leftarrow Now assume $\mathcal{T} = \sum_{i=1}^d \psi_i \otimes \varphi_i$. Let $f \in L_2(\pi_X \otimes \pi_Z)$. We want to show:

$$\int f(x, z) p(x, z) d(\pi_X \otimes \pi_Z)(x, z) = \int f(x, z) \varphi(x)^\top \psi(z) d(\pi_X \otimes \pi_Z)(x, z).$$

Starting with the right-hand side:

$$\begin{aligned} \int f(x, z) \varphi(x)^\top \psi(z) d\pi_X(x) d\pi_Z(z) &= \sum_{i=1}^d \int f(x, z) \varphi_i(x) \psi_i(z) d\pi_X(x) d\pi_Z(z) \\ &= \sum_{i=1}^d \int_Z \psi_i(z) \left(\int_{\mathcal{X}} f(x, z) \varphi_i(x) d\pi_X(x) \right) d\pi_Z(z) \\ &= \int_Z \left(\sum_{i=1}^d \psi_i(z) \langle f(\cdot, z), \varphi_i \rangle \right) d\pi_Z(z) \\ &= \int_Z [\mathcal{T}f(\cdot, z)](z) d\pi_Z(z) \quad (\text{since } \mathcal{T} = \sum \psi_i \otimes \varphi_i) \\ &= \int_Z \left(\int_{\mathcal{X}} f(x, z) p(x, z) d\pi_X(x) \right) d\pi_Z(z) \quad (\text{by Eq. (11)}) \\ &= \int_{\mathcal{X} \times \mathcal{Z}} f(x, z) p(x, z) d\pi_X(x) d\pi_Z(z) \quad (\text{by Lemma 1}). \end{aligned}$$

Since both integrals agree for all $f \in L_2(\pi_X \otimes \pi_Z)$, and $p(x, z)$ is the Radon–Nikodym derivative of $\pi_{X,Z}$ with respect to $\pi_X \otimes \pi_Z$, it follows by uniqueness that

$$p(x, z) = \varphi(x)^\top \psi(z) \quad \text{for } (\pi_X \otimes \pi_Z)\text{-almost every } (x, z).$$

□

In order to prove Theorem 3 we prove intermediate results.

Proposition 6. *Assumption 5 holds if and only if $\text{rank}(\hat{\mathcal{T}}_d) = d$.*

Proof. Assume that Assumption 5 holds, i.e., $\{\hat{\varphi}_1, \dots, \hat{\varphi}_d\}$ and $\{\hat{\psi}_1, \dots, \hat{\psi}_d\}$ are linearly independent. We show that $\text{rank}(\hat{\mathcal{T}}_d) = d$.

Since $\hat{\mathcal{T}}_d h = \sum_{i=1}^d \langle h, \hat{\varphi}_i \rangle \hat{\psi}_i$, we can write $\hat{\mathcal{T}}_d = A \circ \Phi$, where:

- $\Phi : L_2(X) \rightarrow \mathbb{R}^d$, defined by $h \mapsto (\langle h, \hat{\varphi}_1 \rangle, \dots, \langle h, \hat{\varphi}_d \rangle)$,
- $A : \mathbb{R}^d \rightarrow L_2(Z)$, defined by $a \mapsto \sum_{i=1}^d a_i \hat{\psi}_i$.

Under Assumption 5, the family $\{\hat{\varphi}_i\}$ is linearly independent, so Φ is surjective. Similarly, the family $\{\hat{\psi}_i\}$ is linearly independent, so A is injective. Therefore, the composition $A \circ \Phi$ has image of dimension

$$\text{rank}(\hat{\mathcal{T}}_d) = \dim(\mathcal{R}(A \circ \Phi)) = \dim(\mathcal{R}(A)) = d,$$

where the second equality follows from the surjectivity of Φ , and the third from the injectivity of A .

Conversely, suppose $\text{rank}(\hat{\mathcal{T}}_d) = d$. Then $\hat{\mathcal{T}}_d$ has image of dimension d , which implies that the $\hat{\psi}_i$ must be linearly independent. The same reasoning applied to \mathcal{T}^* shows that $\{\hat{\varphi}_1, \dots, \hat{\varphi}_d\}$ is linearly independent.

□

Under Assumption 5, we can therefore write $\hat{\mathcal{T}}_d$ in the following SVD form:

$$\hat{\mathcal{T}}_d = \sum_{i=1}^d \hat{\sigma}_i \hat{u}_i \otimes \hat{v}_i, \tag{15}$$

where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \dots \geq \hat{\sigma}_d > 0$ are the singular values of $\hat{\mathcal{T}}$ and \hat{u}_i and \hat{v}_i are the left and right singular functions of $\hat{\mathcal{T}}$.

Proposition 7. *Under Assumption 5, $\mathcal{H}_{\hat{\varphi},d} = \text{span}\{\hat{v}_1, \dots, \hat{v}_d\}$ and $\mathcal{H}_{\hat{\psi},d} = \text{span}\{\hat{u}_1, \dots, \hat{u}_d\}$.*

Proof. We have by definition of the SVD that

$$\mathcal{N}(\hat{\mathcal{T}}_d)^\perp = \text{span}\{\hat{v}_1, \dots, \hat{v}_d\}.$$

To show that $\mathcal{H}_{\hat{\varphi},d} = \text{span}\{\hat{v}_1, \dots, \hat{v}_d\}$, it suffices to prove that

$$\mathcal{N}(\hat{\mathcal{T}}_d) = \text{span}\{\hat{\varphi}_1, \dots, \hat{\varphi}_d\}^\perp.$$

Let $h \in L_2(X)$. Then:

$$\begin{aligned} h \in \mathcal{N}(\hat{\mathcal{T}}_d) &\iff \hat{\mathcal{T}}_d h = 0 = \sum_{i=1}^d \langle h, \hat{\varphi}_i \rangle \hat{\psi}_i \\ &\iff \sum_{i=1}^d \langle h, \hat{\varphi}_i \rangle \hat{\psi}_i = 0 \\ &\iff \langle h, \hat{\varphi}_i \rangle = 0 \quad \forall i \in [d], \end{aligned}$$

where the last equivalence follows from the linear independence of $\{\hat{\psi}_i\}_{i=1}^d$.

Therefore, $\mathcal{N}(\hat{\mathcal{T}}_d) = \text{span}\{\hat{\varphi}_1, \dots, \hat{\varphi}_d\}^\perp$, which implies

$$\mathcal{H}_{\hat{\varphi},d} = \mathcal{N}(\hat{\mathcal{T}}_d)^\perp = \text{span}\{\hat{v}_1, \dots, \hat{v}_d\}.$$

The proof for $\mathcal{H}_{\hat{\psi},d} = \text{span}\{\hat{u}_1, \dots, \hat{u}_d\}$ is analogous, using the adjoint $\hat{\mathcal{T}}_d^*$. \square

We now recall and prove Theorem 3:

Theorem 3. *Let Assumptions 1, 2, 4 and 5 hold and let $\hat{\varepsilon}_d \doteq \|\hat{\mathcal{T}}_d - \mathcal{T}_d\|_{\text{op}}$ be such that $\hat{\varepsilon}_d < (1 - 1/\sqrt{2})\sigma_d$.*

- i) $\sigma_d^{-1} \leq \tau_{\hat{\varphi},d} \leq (\sigma_d - 2\hat{\varepsilon}_d)^{-1}$;
- ii) *Let \hat{h} be the 2SLS estimator from Eq. (5), using features $\hat{\varphi} : \mathcal{X} \rightarrow \mathbb{R}^d$ and $\hat{\psi} : \mathcal{Z} \rightarrow \mathbb{R}^d$. Suppose $(\sigma_d - 2\hat{\varepsilon}_d)^{-1} \zeta_{\hat{\varphi},\hat{\psi},d} \sqrt{(\log d)/n} = o(1)$ and $\hat{\varepsilon}_d = o_d(\sigma_d^2)$, then*

$$\|\hat{h} - h_0\|_{L^2(X)} = O_p \left(\|(\Pi_{\mathcal{X},d})^\perp h_0\|_{L_2(X)} + \frac{\sqrt{2}\hat{\varepsilon}_d \|h_0\|_{L_2(X)}}{\sigma_d} + \frac{1}{\sigma_d - 2\hat{\varepsilon}_d} \sqrt{\frac{d}{n}} \right).$$

Proof. We first show that $\sigma_d^{-1} \leq \tau_{\hat{\varphi},d} \leq (\sigma_d - 2\hat{\varepsilon}_d)^{-1}$. First note that if $h \in \mathcal{H}_{\hat{\varphi},d}$, then by Proposition 7, $h \in \mathcal{V}_d$. Using the SVD of $\hat{\mathcal{T}}_d$ from Eq. (15), and writing $h = \sum_{i=1}^d \langle h, \hat{v}_i \rangle \hat{v}_i$, we obtain:

$$\hat{\mathcal{T}}_d h = \sum_{i=1}^d \hat{\sigma}_i \langle h, \hat{v}_i \rangle \hat{u}_i, \quad \text{so} \quad \|\hat{\mathcal{T}}_d h\|_{L_2(Z)}^2 = \sum_{i=1}^d \hat{\sigma}_i^2 \langle h, \hat{v}_i \rangle^2 \geq \hat{\sigma}_d^2 \|h\|_{L_2(X)}^2.$$

Next, observe that for all $h \in L_2(X)$,

$$\|\mathcal{T}h\|_{L_2(Z)} \geq \|\mathcal{T}_d h\|_{L_2(Z)}.$$

By the reverse triangle inequality, the bound $\|\hat{\mathcal{T}}_d - \mathcal{T}_d\|_{\text{op}} \leq \hat{\varepsilon}_d$, and Weyl's inequality for singular values, we have, for all $h \in L_2(X)$, with $\|h\|_{L_2(X)} = 1$:

$$\begin{aligned} \|\mathcal{T}h\|_{L_2(Z)} &\geq \|\mathcal{T}_d h\|_{L_2(Z)} \\ &\geq \|\hat{\mathcal{T}}_d h\|_{L_2(Z)} - \|\hat{\mathcal{T}}_d - \mathcal{T}_d\|_{\text{op}} \cdot \|h\|_{L_2(X)} \\ &\geq (\hat{\sigma}_d - \hat{\varepsilon}_d) \|h\|_{L_2(X)} \\ &= \hat{\sigma}_d - \hat{\varepsilon}_d \\ &\geq \sigma_d - 2\hat{\varepsilon}_d. \end{aligned}$$

This shows that

$$\tau_{\hat{\varphi},d}^{-1} \geq \sigma_d - 2\hat{\varepsilon}_d \Rightarrow \tau_{\hat{\varphi},d} \leq (\sigma_d - 2\hat{\varepsilon}_d)^{-1},$$

which holds under the assumption $\hat{\varepsilon}_d < (1 - 1/\sqrt{2})\sigma_d < \sigma_d/2$. On the other hand, since $\dim(\mathcal{H}_{\hat{\varphi},d}) = d$ by Assumption 5, Proposition 4-(i) implies

$$\sigma_d \Rightarrow \tau_{\hat{\varphi},d} \geq \sigma_d^{-1}.$$

Next, we show how to bound the sieve approximation error, write:

$$\begin{aligned} \|(\Pi_{\hat{\varphi},d})^\perp h_0\|_{L_2(X)} &= \|(\Pi_{\hat{\varphi},d})^\perp ((\Pi_{\mathcal{X},d})^\perp h_0 + \Pi_{\mathcal{X},d} h_0)\|_{L_2(X)} \\ &\leq \|(\Pi_{\mathcal{X},d})^\perp h_0\|_{L_2(X)} + \|(\Pi_{\hat{\varphi},d})^\perp \Pi_{\mathcal{X},d} h_0\|_{L_2(X)} \\ &\leq \|(\Pi_{\mathcal{X},d})^\perp h_0\|_{L_2(X)} + \|(\Pi_{\hat{\varphi},d})^\perp \Pi_{\mathcal{X},d}\|_{\text{op}} \cdot \|h_0\|_{L_2(X)}. \end{aligned}$$

By Wedin's sin- Θ theorem (Theorem 2.9 and Eq. (2.26a) in [9]), and under the assumption $\hat{\varepsilon}_d < (1 - 1/\sqrt{2})\sigma_d$, we have:

$$\|(\Pi_{\hat{\varphi},d})^\perp \Pi_{\mathcal{X},d}\|_{\text{op}} \leq \frac{\sqrt{2}\hat{\varepsilon}_d}{\sigma_d}.$$

Therefore,

$$\|(\Pi_{\hat{\varphi},d})^\perp h_0\|_{L_2(X)} \leq \|(\Pi_{\mathcal{X},d})^\perp h_0\|_{L_2(X)} + \frac{\sqrt{2}\hat{\varepsilon}_d \|h_0\|_{L_2(X)}}{\sigma_d},$$

To conclude we now check Assumption 3. Let us introduce

$$s_{\hat{\psi},\hat{\varphi},d}^{-1} = \sup_{h \in \mathcal{H}_{\hat{\varphi},d}, h \neq 0} \frac{\|h\|_{L_2(X)}}{\|\Pi_{\hat{\psi},d} \mathcal{T} h\|_{L_2(Z)}} \geq \tau_{\hat{\varphi},d}.$$

The unique use of Assumption 3-i) in the proof of Theorem B.1 [6] is to prove that $s_{\hat{\psi},\hat{\varphi},d}^{-1} = O(\tau_{\hat{\varphi},d})$ as $d \rightarrow +\infty$, which we now directly prove. First, note that for any $h \in L_2(X)$,

$$\begin{aligned} \|\mathcal{T} \Pi_{\hat{\varphi},d} h\|_{L_2(Z)} &= \|\mathcal{T} (\Pi_{\hat{\varphi},d} + \Pi_{\mathcal{X},d} - \Pi_{\mathcal{X},d}) h\|_{L_2(Z)} \\ &\leq \|\mathcal{T} h\|_{L_2(Z)} + \|\Pi_{\hat{\varphi},d} - \Pi_{\mathcal{X},d}\|_{\text{op}} \|h\|_{L_2(X)} \\ &\leq \|\mathcal{T} h\|_{L_2(Z)} + \frac{\sqrt{2}\hat{\varepsilon}_d \|h\|_{L_2(X)}}{\sigma_d}, \end{aligned}$$

where we used $\|\mathcal{T}\|_{\text{op}} \leq 1$, the definition of \mathcal{T}_d and Wedin's sin- Θ theorem. On the other hand,

$$\begin{aligned} \|\Pi_{\hat{\psi},d} \mathcal{T} \Pi_{\hat{\varphi},d} h\|_{L_2(Z)} &\geq \|\Pi_{\hat{\psi},d} \mathcal{T} \Pi_{\varphi,d} h\|_{L_2(Z)} - \|\Pi_{\hat{\psi},d} \mathcal{T} (\Pi_{\hat{\varphi},d} - \Pi_{\mathcal{X},d}) h\|_{L_2(Z)} \\ &\geq \|\Pi_{\psi,d} \mathcal{T} \Pi_{\varphi,d} h\|_{L_2(Z)} - \|(\Pi_{\psi,d} - \Pi_{\hat{\psi},d}) \mathcal{T} \Pi_{\varphi,d} h\|_{L_2(Z)} - \frac{\sqrt{2}\hat{\varepsilon}_d \|h\|_{L_2(X)}}{\sigma_d} \\ &\geq \|\mathcal{T} h\|_{L_2(Z)} - \frac{2\sqrt{2}\hat{\varepsilon}_d \|h\|_{L_2(X)}}{\sigma_d}. \end{aligned}$$

We therefore obtain that,

$$s_{\hat{\psi},\hat{\varphi},d}^{-1} \leq \sup_{h \in \mathcal{H}_{\hat{\varphi},d}, h \neq 0} \frac{1}{\frac{\|\mathcal{T} \Pi_{\hat{\varphi},d} h\|_{L_2(Z)}}{\|h\|_{L_2(X)}} - \frac{3\sqrt{2}\hat{\varepsilon}_d}{\sigma_d}} \leq \tau_{\hat{\varphi},d} \times \frac{1}{1 - \frac{3\sqrt{2}\hat{\varepsilon}_d \tau_{\hat{\varphi},d}}{\sigma_d}},$$

where the inequality is valid for d large enough as long as $3\sqrt{2}\hat{\varepsilon}_d \tau_{\hat{\varphi},d} \sigma_d^{-1} = o(1)$. Re-using that $\tau_{\hat{\varphi},d} \leq (\sigma_d - 2\hat{\varepsilon}_d)^{-1}$, a sufficient condition is

$$3\sqrt{2} \frac{\hat{\varepsilon}_d}{\sigma_d} \frac{1}{\sigma_d - 2\hat{\varepsilon}_d} = o(1),$$

which is satisfied if $\hat{\varepsilon}_d = o(\sigma_d^2)$. We therefore conclude that if $\hat{\varepsilon}_d = o(\sigma_d^2)$ then $s_{\hat{\psi},\hat{\varphi},d}^{-1} = O(\tau_{\hat{\varphi},d})$. Finally, we check Assumption 3-ii). First note that by definition of $\Pi_{\hat{\varphi},d}$ and $\hat{\mathcal{T}}_d$, we have

$\hat{\mathcal{T}}_d(\Pi_{\hat{\varphi},d})_{\perp} = 0$, therefore,

$$\begin{aligned} \tau_{\hat{\varphi},d} \|\mathcal{T}(\Pi_{\hat{\varphi},d})_{\perp} h_0\| &= \tau_{\hat{\varphi},d} \left\| (\mathcal{T} - \hat{\mathcal{T}}_d)(\Pi_{\hat{\varphi},d})_{\perp} h_0 \right\| \\ &\leq \tau_{\hat{\varphi},d} \left(\|(\mathcal{T} - \mathcal{T}_d)(\Pi_{\hat{\varphi},d})_{\perp} h_0\| + \|(\mathcal{T}_d - \hat{\mathcal{T}}_d)(\Pi_{\hat{\varphi},d})_{\perp} h_0\| \right) \\ &\leq \tau_{\hat{\varphi},d} (\sigma_{d+1} + \hat{\varepsilon}_d) \|(\Pi_{\hat{\varphi},d})_{\perp} h_0\| \\ &\leq \frac{\sigma_{d+1} + \hat{\varepsilon}_d}{\sigma_d - 2\hat{\varepsilon}_d} \|(\Pi_{\hat{\varphi},d})_{\perp} h_0\|, \end{aligned}$$

We conclude using that $\hat{\varepsilon}_d = o(\sigma_d^2)$ implies $\hat{\varepsilon}_d = o(\sigma_d)$ which implies $(\sigma_{d+1} + \hat{\varepsilon}_d)(\sigma_d - 2\hat{\varepsilon}_d) = O(1)$. \square

We now prove the equivalence between Eq. (6) and Eq. (7). While the proof strategy follows that of Kostic et al. [25, Theorem 1], we include our own version for completeness, as the parametrization of the learned operator differs. Specifically, [25] directly approximate the truncated operator \mathcal{T}_d using a singular value decomposition of the form $\mathbb{1}_Z \otimes \mathbb{1}_X + \sum_{i=2}^d \hat{\sigma}_i \hat{\psi}_i \otimes \hat{\varphi}_i$, where the singular values $\hat{\sigma}_i$ are learned explicitly via a separate network. In contrast, our approach learns only the feature maps (φ_i, ψ_i) and defines the approximation $\mathcal{T}_d(\hat{\varphi}, \hat{\psi}) = \sum_{i=1}^d \hat{\psi}_i \otimes \hat{\varphi}_i$. As such, we provide a short self-contained derivation adapted to our setting for clarity. Recall that the Eckart–Young–Mirsky formulation of the objective is defined as

$$\mathcal{L}_d(\varphi, \psi) = \|\mathcal{T}_d(\varphi, \psi) - \mathcal{T}\|_{\text{HS}}^2 = \|\mathcal{T}_d(\varphi, \psi)\|_{\text{HS}}^2 - 2\langle \mathcal{T}_d(\varphi, \psi), \mathcal{T} \rangle_{\text{HS}} + \|\mathcal{T}_d(\varphi, \psi)\|_{\text{HS}}^2.$$

Proposition 8. *It holds that,*

$$\|\mathcal{T}_d(\varphi, \psi)\|_{\text{HS}}^2 - 2\langle \mathcal{T}_d(\varphi, \psi), \mathcal{T} \rangle_{\text{HS}} = \mathbb{E}_X \mathbb{E}_Z \left[(\varphi(X)^\top \psi(Z))^2 \right] - 2\mathbb{E}_{X,Z} [\varphi(X)^\top \psi(Z)].$$

Proof. Let us introduce

$$\Phi : \mathbb{R}^d \rightarrow L_2(X), \quad c \mapsto \sum_{i=1}^d c_i \varphi_i, \quad \Psi : \mathbb{R}^d \rightarrow L_2(Z), \quad c \mapsto \sum_{i=1}^d c_i \psi_i,$$

such that $\mathcal{T}_d(\varphi, \psi) = \Psi \Phi^*$, $\Psi^* \Psi = \mathbb{E}[\psi(Z) \psi(Z)^\top]$ and $\Phi^* \Phi = \mathbb{E}[\varphi(X) \varphi(X)^\top]$. On one hand, we have:

$$\begin{aligned} \|\mathcal{T}_d(\varphi, \psi)\|_{\text{HS}}^2 &= \|\Psi \Phi^*\|_{\text{HS}}^2 \\ &= \text{Tr}(\Psi^* \Psi \Phi^* \Phi) \\ &= \text{Tr}(\mathbb{E}[\psi(Z) \psi(Z)^\top] \mathbb{E}[\varphi(X) \varphi(X)^\top]) \\ &= \mathbb{E}_X \mathbb{E}_Z [\text{Tr}(\psi(Z) \psi(Z)^\top \varphi(X) \varphi(X)^\top)] \\ &= \mathbb{E}_X \mathbb{E}_Z \left[(\psi(Z)^\top \varphi(X))^2 \right]. \end{aligned}$$

On the other hand, we have:

$$\begin{aligned} \langle \mathcal{T}_d(\varphi, \psi), \mathcal{T} \rangle_{\text{HS}} &= \text{Tr}(\Psi^* \mathcal{T} \Phi) \\ &= \text{Tr}(\mathbb{E}[\psi(Z) \mathbb{E}[\varphi(X) | Z]^\top]) \\ &= \mathbb{E}_{X,Z} [\text{Tr}(\psi(Z) \varphi(X)^\top)] \\ &= \mathbb{E}_{X,Z} [\psi(Z)^\top \varphi(X)], \end{aligned}$$

which conclude the proof. \square

As noted in [25], the above result does not require \mathcal{T} to be Hilbert-Schmidt, or even compact. Indeed as $\mathcal{T}_d(\varphi, \psi)$ is a finite rank operator, $\langle \mathcal{T}_d(\varphi, \psi), \mathcal{T} \rangle_{\text{HS}}$ is always well defined.

D Extended Related Work

We now discuss the various ideas that have been proposed to solve NPIV problems. Extensive benchmarking for these methods has already been conducted in prior work; refer, for instance, to [36].

As mentioned in Section 1, a central challenge in NPIV estimation is that it constitutes an ill-posed inverse problem. The literature has evolved along several methodological lines to address it. Two broad classes of estimation strategies have become particularly prominent:

- **Two-Stage Least-Squares (2SLS).** This classical approach and its various generalizations involves a sequential estimation procedure. Typically, the first stage estimates the conditional expectation of the endogenous variable (or its features) given the instrument, and the second stage uses these predictions to estimate the structural function.
- **Saddle-point optimization.** These methods, often rooted in a generalized method of moments (GMM) framework or duality principles, reformulate the NPIV estimation as finding an equilibrium in a min-max game. This often involves optimizing an objective function over a hypothesis space for the structural function and a test function space for moment conditions.

These two methodologies developed with distinct focuses. The 2SLS methods, with roots in classical econometrics [32, 11], have progressively incorporated more sophisticated nonparametric techniques. Saddle-point methods, on the other hand, have gained traction with the rise of machine learning, offering powerful tools for handling complex optimization problems arising from conditional moment restrictions [13, 29, 2].

2SLS approaches. Early and influential nonparametric extensions of 2SLS employed sieve or series estimators. These methods approximate the unknown functions using a finite linear combination of basis functions, such as polynomials, splines, or Fourier series. The number of basis functions, or the “sieve dimension”, is allowed to increase with the sample size, enabling consistent estimation of the nonparametric functions.

A seminal contribution, by [32], provided identification results and a consistent nonparametric estimator for conditional moment restrictions. Their proposed NPIV estimator is an analog of 2SLS, where the first stage involves nonparametric estimation of conditional means of basis functions of X given Z , and the second stage uses a series approximation for the structural function based on these first-stage predictions. Regularization is achieved by controlling the number of terms in the series approximation.

Building on similar principles, [18] proposed nonparametric methods based on both kernel⁶ techniques and orthogonal series for estimating regression functions with instrumental variables. For their orthogonal series estimator, they transformed the marginal distributions of X and Z to be uniform and used Fourier expansions. The estimated Fourier coefficients of the structural function were obtained via a regularized regression involving estimated coefficients from the first stage. They were among the first to derive optimal convergence rates for this class of problems, explicitly linking these rates to the “difficulty” of the ill-posed inverse problem, which is characterized by the eigenvalues of the underlying integral operator.

[11] proposed an estimation procedure based on Tikhonov regularization. This involves regularizing the inverse of the integral operator \mathcal{T} (or its empirical counterpart) to stabilize the solution. Specifically, denoting $\hat{\mathcal{T}}$ and \hat{r}_0 empirical estimates (in their case, computed with kernel density estimators), the Tikhonov regularized solution is of the form $h_\alpha = \left(\hat{\mathcal{T}}^* \hat{\mathcal{T}} + \alpha I \right)^{-1} \hat{\mathcal{T}}^* \hat{r}_0$, where $\alpha > 0$. They presented asymptotic properties of their estimator, including consistency and convergence rates, which depend on the smoothness of the structural function and the degree of ill-posedness of \mathcal{T} .

[35] introduced Kernel Instrumental Variable Regression (KIV), a direct nonparametric generalization of 2SLS. KIV models the relationships between the different variables as nonlinear functions in reproducing kernel Hilbert spaces (RKHSs). In stage 1, it learns a conditional mean embedding $\mu(z) = \mathbb{E}[\varphi(X) \mid Z = z]$ where $\varphi(X)$ represents features of X mapped into an RKHS \mathcal{H}_X . This learning is framed as a vector-valued kernel ridge regression, effectively estimating the conditional

⁶“Kernel” is meant here in the sense of density estimation, and not reproducing kernel: see [18, eq. (2.4)].

expectation operator $E : \mathcal{H}_X \rightarrow \mathcal{H}_Z$. In stage 2, a scalar-valued kernel ridge regression of the outcome Y on the estimated means $\hat{\mu}(Z)$ is performed to estimate the structural function. The authors prove the consistency of KIV in the projected norm under mild conditions and derive conditions under which KIV achieves minimax optimal convergence rates. The analysis was later improved by [31], who established convergence in L_2 norm rather than the projected norm.

Another related approach, “fast IV” (FIV), was proposed in [39]. FIV studies nonlinear IV with high-dimensional instruments and proposes a two-stage pipeline that keeps the outcome model in a fixed RKHS/GP space but learns the first-stage kernel from a black-box adaptive regressor (*e.g.*, a neural network) distilled into a compact kernel basis for estimating the conditional expectation operator. The resulting estimator plugs this learned kernel into standard kernelized IV schemes, achieving rates that adapt to the dimensionality of informative instrument features. Their work focuses on learning a kernel basis for the instrument. In contrast, our method learns paired spectral features for both the instrument and the treatment to approximate the top singular subspaces of the conditional expectation operator.

Saddle-point approaches. These arise from reformulating the conditional moment restrictions that define the NPIV problem (Eq. 1), frequently conceptualized as a zero-sum game. One player selects a function h from a hypothesis space \mathcal{H} to minimize the objective. The other player, the “adversary” or “witness”, selects a test function g from a test function space \mathcal{G} to maximize the objective. The adversary’s role is to select the function that maximizes the violation of the moment condition by the current choice of h . A notable advantage of this formulation is that it can allow bypassing the direct estimation of conditional expectations. Saddle-point objectives in NPIV regression can be derived in different ways.

One can frame the NPIV problem as finding a solution h_0 that minimizes a certain criterion (*e.g.*, a norm) subject to satisfying the moment condition $\mathcal{T}h_0 = r_0$. The Lagrangian of this constrained problem then leads to a minimax objective. [2] target the least-norm solution

$$h_0 = \arg \min_{h \in L_2(X)} \frac{1}{2} \|h\|_{L_2(X)}^2 \quad \text{subject to} \quad \mathcal{T}h = r_0.$$

The corresponding Lagrangian is given by $L(f, g) = \frac{1}{2} \|h\|_{L_2(X)}^2 + \langle r_0 - \mathcal{T}h, g \rangle_{L_2(Z)}$ for $g \in L_2(Z)$. Their method achieves strong L_2 error rates under a source condition and realizability assumptions. Notably, their approach does not require the often-problematic closedness condition or uniqueness of the IV solution. [29] consider a similar approach, focusing on the setting where the function classes are formed by neural networks, but use different techniques to analyze their method (online learning and neural tangent kernel theory).

Another approach to deriving a saddle-point problem involves using Fenchel duality to transform a squared error loss involving conditional expectations. This is the approach taken by [36], which we study in this paper.

A third method is to directly use the unconditional moment formulation $\mathbb{E}[(Y - h(X))g(Z)] = 0$. [13] define their criterion as the maximum moment deviation,

$$h_0 = \arg \inf_{h \in \mathcal{H}} \sup_{g \in \mathcal{G}} \mathbb{E}[(Y - h(X))g(Z)] = \Psi(h, g),$$

and define the estimator $\hat{h} = \arg \min_h \sup_g \mathbb{E}_n[(Y - h(X))g(Z)] + \mu R_1(h) - \lambda R_2(g)$, where R_1, R_2 are regularizers. Their key theoretical result is that the statistical estimation rate, in terms of projected mean squared error $\|\mathcal{T}(\hat{h} - h_0)\|_2$, scales with the critical radii of the hypothesis space \mathcal{H} and the test function space \mathcal{G} . This holds under some closedness assumption, namely that $\mathbb{E}[h(X) - h'(X) | Z] \in \mathcal{G}$ for any $h, h' \in \mathcal{H}$. We note that the method introduced by [43] defines a risk functional in terms of the squared moment deviation and is thus related to GMM.

The role and interpretation of the “adversary” function $g \in \mathcal{G}$ and its associated objective function vary subtly across different saddle-point formulations, which in turn influences the types of assumptions required. This distinction is important: if g is intended to approximate $\mathbb{E}[Y - h(X) | Z]$, then the space \mathcal{G} must be sufficiently rich to do so, which is reflected via a closedness condition. If g is primarily a Lagrange multiplier, its existence and properties are more directly tied to conditions like a source condition. This difference helps explain why [2] can dispense with the closedness assumption while other methods may require it.

About the contrastive loss. The contrastive loss used to learn spectral features derives its name from the foundational concept of contrastive learning, which traces back to the early 1990s. Notably, [5] introduced the Siamese network architecture for signature verification, which utilized a form similar to Eq.(8) to measure the distance between input pairs. Later, [10] formalized the contrastive loss function for face verification tasks. This loss is, in turn, linked to a truncated conditional operator (see Eq. (6)). The underlying principle has been frequently redeveloped under various names, such as correspondence analysis [17], principal inertia components [21, 22] for finite alphabets, the contrastive kernel [19, 12], and pointwise dependence [37] in self-supervised representation learning. The objective in Eq. (8) was also proposed and studied by [38] and used as a local approximation to the log-loss of classification deep neural networks [42]. More recently, [25] linked this same objective to the SVD of the conditional expectation operator.

E Experimental Details

In this section, we provide additional details for the experiments presented in the main text. Let (Z, X) be generated as described in Section 6 with $n = 100000$. Figure 4 displays both the true data-generating density and the density corresponding to a set of learned spectral features. Even for $d = 11$, the resulting densities are complex. While it is feasible to conduct analogous experiments with higher values of d , we observed no qualitative changes in the outcomes. However, training and hyperparameter tuning became increasingly challenging as d grew.

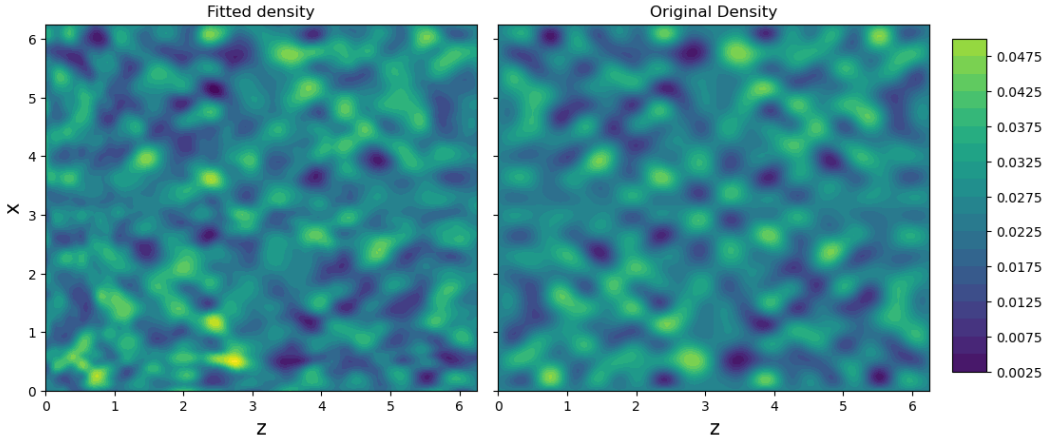


Figure 4: Comparison of a density corresponding to a set of learned spectral features (**left**) and the true data generating density (**right**).

E.1 Models Employed

The features were learned using two-hidden-layer neural networks. All models shared the same architecture, with layer widths $[1, 50, 50, 50]$: the input is one-dimensional, and the final layer outputs 50 learned features. To encourage the models to learn more oscillatory functions, the first layer used the activation $x \mapsto x + \sin^2(x)$, as introduced by [44], followed by GELU-activated hidden layers and a final linear layer.

We note that the first singular eigenfunctions of the conditional expectation operator \mathcal{T} are always the constant functions (see Section 2). Therefore, we hard-code the constant feature $\mathbb{1}_Z \otimes \mathbb{1}_X$ into the model and restrict the following learned features to be mean-zero.

In addition, we included a regularization term to penalize both feature collinearity and large feature norms. This regularizer is a sample-based approximation of the quantity defined in Equation (10) of [25]:

$$\mathbb{E} [\|\varphi(X)\varphi(X)^\top - I\|^2] + \mathbb{E} [\|\psi(Z)\psi(Z)^\top - I\|^2] + 2\mathbb{E} [\|\varphi(X)\|^2] + 2\mathbb{E} [\|\psi(Z)\|^2].$$

The spectral features were trained on 100,000 samples of (Z, X) , while the 2SLS estimator built from the learned features used a separate dataset of 10,000 samples of (Z, X, Y) .

E.2 Expanded example of the ugly scenario

As noted in the main text, the difficulty of recovering h_0 increases when it is not well supported on the singular functions of the conditional expectation operator \mathcal{T} . If the projection of h_0 onto the span of the learned singular functions is sufficiently large, the missing components may not significantly harm the quality of the estimate. We illustrate this property with a controlled experiment designed to vary the amount of support h_0 has in the singular space of \mathcal{T} .

We fix $d - 1$ orthonormal features u_i for Z and v_i for X , and define the operator:

$$\mathcal{T} = \mathbb{1}_Z \otimes \mathbb{1}_X + \sum_{i=1}^{d-1} \sigma_i u_i \otimes v_i,$$

mirroring the setup from Section 6. We vary the spectrum by setting $\sigma_{1:k} = c$ and $\sigma_{k+1:d-1} = 0$ for $k \in \{1, \dots, d-1\}$. We then define the target function:

$$h_0 = \frac{1}{\sqrt{d-1}} \sum_{i=1}^{d-1} v_i,$$

which is uniformly spread across the feature directions. As k increases, the projection of h_0 onto the singular space of \mathcal{T} increases in discrete steps from 0 to 1. This results in a corresponding qualitative improvement in the accuracy of the 2SLS estimator for h_0 . Figure 5 illustrates this behavior: for small values of k , where h_0 is largely orthogonal to the singular functions of \mathcal{T} , the estimate fails to recover h_0 . As k grows, and more of h_0 's energy lies in the span of these singular functions, the 2SLS estimate increasingly aligns with the true function.

F Estimating spectral alignment

In this section, we present the details of the methodology introduced in Section 5. We recall the central formula provided in Eq. (10).

$$\langle v_i, h_0 \rangle_{L^2(X)} = \frac{1}{\sigma_i} \mathbb{E}[Y \cdot u_i(Z)]. \quad (16)$$

To evaluate spectral alignment, we shall approximate the RHS of the above. Suppose we have learned an operator $\hat{\mathcal{T}}_d = \sum_{i=1}^d \hat{\psi}_i \otimes \hat{\phi}_i = \Psi \Phi^*$ which approximates the true conditional mean operator \mathcal{T} . Here $\Psi: \mathbb{R}^d \rightarrow L^2(Z)$ and $\Phi: \mathbb{R}^d \rightarrow L^2(X)$ are the maps sending $\alpha \mapsto \sum_{i=1}^d \alpha_i \hat{\psi}_i$ and $\alpha \mapsto \sum_{i=1}^d \alpha_i \hat{\phi}_i$. One can compute the SVD of $\hat{\mathcal{T}}_d$ in two steps. We start with the derivations assuming access to population quantities. Let C_Z , and C_X be the $d \times d$ population covariance matrices of the learned Z , and X features, respectively. For any z and x , we then introduce the whitened feature vectors $\psi'_i(z) = (C_Z^{-1/2} \hat{\psi}(z))_i$ and $\phi'_i(x) = (C_X^{-1/2} \hat{\phi}(x))_i$. The corresponding operators, $\Psi' = \Psi C_Z^{-1/2}$ and $\Phi' = \Phi C_X^{-1/2}$, are isometries. We can then write

$$\hat{\mathcal{T}}_d = \Psi' C_Z^{1/2} C_X^{1/2} (\Phi')^*.$$

$C_Z^{1/2} C_X^{1/2}$ is a $d \times d$ matrix, let its SVD be $C_Z^{1/2} C_X^{1/2} = O \Sigma_{\phi, \psi} P^*$, where $\Sigma_{\phi, \psi}$ is a positive-definite diagonal matrix with diagonal entries $\sigma_{\phi, \psi, i}$. Now we can write

$$\hat{\mathcal{T}}_d = (\Psi' O) \Sigma_{\phi, \psi} (\Phi' P)^*.$$

Finally letting $u_{\phi, \psi, i} = \Psi' O e_i$, and $v_{\phi, \psi, i} = \Phi' P e_i$, with the corresponding operators $U_{\phi, \psi}: \mathbb{R}^d \rightarrow L^2(Z)$, and $V_{\phi, \psi}: \mathbb{R}^d \rightarrow L^2(X)$, we get the SVD

$$\hat{\mathcal{T}}_d = U_{\phi, \psi} \Sigma_{\phi, \psi} V_{\phi, \psi}^*.$$

In practice, given a (Z, X) dataset, possibly the same one as was used to train $\hat{\mathcal{T}}_d$, one can perform sample-based counterparts to the above procedures to get an approximate SVD.

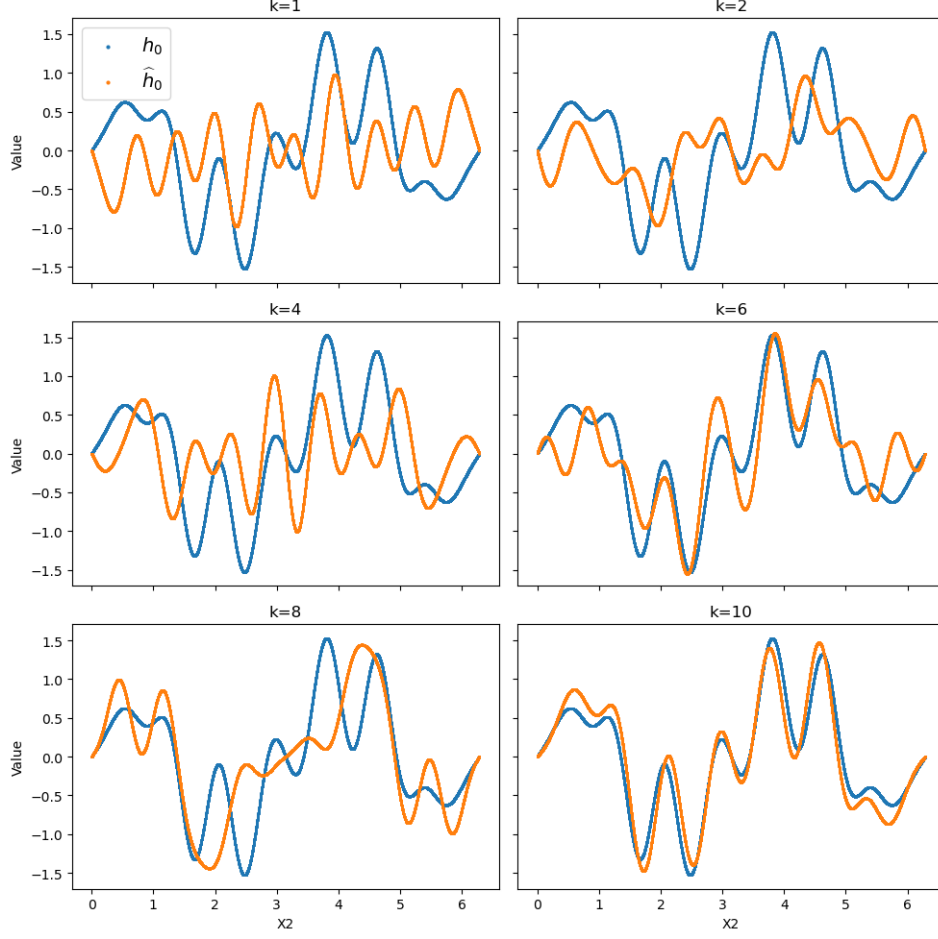


Figure 5: Qualitative improvement in the 2SLS estimate of h_0 as k increases. When $k = 1$, h_0 is orthogonal to the singular functions of \mathcal{T} ; when $k = d - 1$, it is fully contained in their span.

Rationale for the alignment approximation method. We are interested in evaluating the length of the projection of h_0 onto the leading learned singular functions of $\hat{\mathcal{T}}_d$, that is of $\langle v_{\phi,\psi,i}, h_0 \rangle$. Following Eq. (10), we have

$$\begin{aligned} \langle v_{\phi,\psi,i}, h_0 \rangle &= \sigma_{\phi,\psi,i}^{-1} \langle \hat{\mathcal{T}}_d^* u_{\phi,\psi,i}, h_0 \rangle = \sigma_{\phi,\psi,i}^{-1} \langle u_{\phi,\psi,i}, \hat{\mathcal{T}}_d h_0 \rangle \\ &= \sigma_{\phi,\psi,i}^{-1} \langle u_{\phi,\psi,i}, \mathcal{T} h_0 \rangle + \sigma_{\phi,\psi,i}^{-1} \langle u_{\phi,\psi,i}, (\hat{\mathcal{T}}_d - \mathcal{T}) h_0 \rangle. \end{aligned}$$

The first term on the RHS above can be equivalently written as $\sigma_{\phi,\psi,i}^{-1} \langle u_{\phi,\psi,i}, \mathcal{T} h_0 \rangle = \sigma_{\phi,\psi,i}^{-1} \mathbb{E}[u_{\phi,\psi,i}(Z)Y]$, for which it is easy to compute a sample-based approximation. This approximation should be close to the true value of $\langle v_{\phi,\psi,i}, h_0 \rangle$ provided that the second RHS term is small. Letting \mathcal{T}_d be the rank- d SVD truncation of \mathcal{T} we further telescope the second term on the RHS to get

$$\sigma_{\phi,\psi,i}^{-1} \left(\langle u_{\phi,\psi,i}, (\hat{\mathcal{T}}_d - \mathcal{T}_d) h_0 \rangle + \langle u_{\phi,\psi,i}, (\mathcal{T}_d - \mathcal{T}) h_0 \rangle \right). \quad (17)$$

Since $\hat{\mathcal{T}}_d$ should converge to \mathcal{T}_d for a sufficiently flexible feature-learning model class, it is sensible to bound the norm of the first term above with $\sigma_{\phi,\psi,i}^{-1} \|\hat{\mathcal{T}}_d - \mathcal{T}_d\| \|h_0\|$. For any $i \leq d$, this upper bound should converge to 0 with the size of the feature-learning sample size. To upper-bound the second term of Eq. (17), note that by Wedin Sin- Θ Theorem, we have

$$\|\Pi_{u_{1:d}, \phi, \psi} - \Pi_{u_{1:d}}\| \leq \frac{\|\mathcal{T}_d - \hat{\mathcal{T}}_d\|}{\sigma_d}.$$

Thus

$$|\langle u_{\phi,\psi,i}, (\mathcal{T}_d - \mathcal{T})h_0 \rangle| \leq \|\Pi_{u_{\phi,\psi,1:d}}(\mathcal{T}_d - \mathcal{T})h_0\| \leq \underbrace{\|\Pi_{u_{1:d}}(\mathcal{T}_d - \mathcal{T})h_0\|}_{=0} + \frac{\|\mathcal{T}_d - \hat{\mathcal{T}}_d\|}{\sigma_d} \cdot \|(\mathcal{T}_d - \mathcal{T})h_0\|.$$

By construction, $\text{Span}(u_{1:d})$ is orthogonal to the image of $\mathcal{T}_d - \mathcal{T}$. Hence the first term on the RHS vanishes, leaving us with

$$\left| \langle v_{\phi,\psi,i}, h_0 \rangle - \sigma_{\phi,\psi,i}^{-1} \mathbb{E}[Y u_{\phi,\psi,i}] \right| \leq \sigma_{\phi,\psi,i}^{-1} \|\hat{\mathcal{T}}_d - \mathcal{T}_d\| \|h_0\| + \sigma_d^{-1} \sigma_{\phi,\psi,i}^{-1} \|\mathcal{T}_d - \hat{\mathcal{T}}_d\| \|(\mathcal{T}_d - \mathcal{T})h_0\|. \quad (18)$$

The essential takeaway from the bound above is that as long as the feature-learning sample size can increase to ensure $\|\hat{\mathcal{T}}_d - \mathcal{T}_d\|$ is small, approximating the alignment with $\sigma_{\phi,\psi,i}^{-1} \mathbb{E}[Y u_{\phi,\psi,i}]$ is consistent. Moreover, the bound becomes tighter when,

1. $\sigma_{\phi,\psi,i}$ is big (close to 1), which corresponds to a slow spectral decay or looking at the top singular function.
2. h_0 lies in the top of the spectrum of \mathcal{T} so that $\|(\mathcal{T}_d - \mathcal{T})h_0\|$ is small.

We can additionally compare how close is $\sigma_{\phi,\psi,i}^{-1} \mathbb{E}[Y u_{\phi,\psi,i}]$ to the alignment with the true eigenfunction of \mathcal{T} : $\langle v_i, h_0 \rangle$, so that we can evaluate if we are in the “good” or “bad” scenario. For $i \leq d$, denote $\alpha_i = \langle v_{\phi,\psi,i}, h_0 \rangle$ and $\hat{\alpha}_i = \sigma_{\phi,\psi,i}^{-1} \mathbb{E}[Y u_{\phi,\psi,i}(Z)]$. The previous bound shows that

$$|\alpha_i - \hat{\alpha}_i| \leq \text{Err}_i, \quad \text{Err}_i = \sigma_{\phi,\psi,i}^{-1} \|\hat{\mathcal{T}}_d - \mathcal{T}_d\| \|h_0\| + \sigma_d^{-1} \sigma_{\phi,\psi,i}^{-1} \|\mathcal{T}_d - \hat{\mathcal{T}}_d\| \|(\mathcal{T}_d - \mathcal{T})h_0\|.$$

Then

$$\begin{aligned} \left| \|\Pi_{v_{\phi,\psi,1:d}} h_0\|^2 - \sum_{i=1}^d \sigma_{\phi,\psi,i}^{-2} |\mathbb{E}[Y u_{\phi,\psi,i}(Z)]|^2 \right| &= \left| \sum_{i=1}^d (\alpha_i^2 - \hat{\alpha}_i^2) \right| = \left| \sum_{i=1}^d (\alpha_i - \hat{\alpha}_i)(\alpha_i + \hat{\alpha}_i) \right| \\ &\leq \sum_{i=1}^d |\alpha_i - \hat{\alpha}_i| \cdot |\alpha_i + \hat{\alpha}_i| \end{aligned}$$

Next,

$$|\alpha_i + \hat{\alpha}_i| = |2\alpha_i - (\alpha_i - \hat{\alpha}_i)| \leq 2|\alpha_i| + |\alpha_i - \hat{\alpha}_i| \leq 2\|h_0\| + \text{Err}_i,$$

and

$$\begin{aligned} \left| \|\Pi_{v_{\phi,\psi,1:d}} h_0\|^2 - \sum_{i=1}^d \sigma_{\phi,\psi,i}^{-2} |\mathbb{E}[Y u_{\phi,\psi,i}(Z)]|^2 \right| &\leq \sum_{i=1}^d \text{Err}_i \cdot (2\|h_0\| + \text{Err}_i) \\ &= 2\|h_0\| \sum_{i=1}^d \text{Err}_i + \sum_{i=1}^d (\text{Err}_i)^2 \end{aligned}$$

Finally, to assess how well we get the alignment to the eigenfunctions of \mathcal{T} observe that

$$\left| \|\Pi_{v_{\phi,\psi,1:d}} h_0\| - \|\Pi_{v_{1:d}} h_0\| \right| \leq \|(\Pi_{v_{\phi,\psi,1:d}} - \Pi_{v_{1:d}}) h_0\| \leq \frac{\|\mathcal{T}_d - \hat{\mathcal{T}}_d\|}{\sigma_d} \|h_0\|.$$

Practical considerations. In practice, we want to perform sample-based approximations of the procedure described above, and while the top singular functions of \mathcal{T} and the corresponding singular values can be reliably estimated, learning the bottom of the spectrum is more unreliable. Given that we divide $\mathbb{E}[Y u_{\phi,\psi,i}(Z)]$ by $\sigma_{i,\phi,\psi}$, for singular values close to 0, a small error in estimating them is inflated by the inversion. Hence, we resort to a heuristic that allows us to decide which features and singular functions are learned reliably. Let $\hat{\mathcal{T}}_d = \hat{U}_{\phi,\psi} \hat{\Sigma}_{\phi,\psi} \hat{V}_{\phi,\psi}$ be a finite-sample approximation of the SVD of $\hat{\mathcal{T}}_d^*$. That is, we perform the SVD computation procedures of the preceding paragraph but relying on sample feature covariance matrices \hat{C}_X and \hat{C}_Z . After fitting the SVD once, we recompute the covariance of $\hat{u}_{\phi,\psi,i}$ and $\hat{v}_{\phi,\psi,i}$ on resampled $(Z^{(b)}, X^{(b)})$ data with $b = 1, \dots, B$ and extract its diagonal terms, which we refer to as $\hat{\sigma}_{\phi,\psi,1:d}^{(b)}$. Letting the original singular value estimates be

$\hat{\sigma}_{\phi,\psi,i} \doteq \hat{\sigma}_{\phi,\psi,i}^{(0)}$, we can use the sets $(\hat{\sigma}_{\phi,\psi,i}^{(b)})_{b=0}^B$ to evaluate the reliability of a given singular value estimate. If the variance of the estimates in a given collection of $\hat{\sigma}_{\phi,\psi,k}^{(b)}$ is large relative to their mean, we treat the value as unreliable and discard all the singular values $\sigma_{\phi,\psi,i \geq k}$ from further analysis.

For the remaining singular values, $\hat{\sigma}_{\phi,\psi,1}, \dots, \hat{\sigma}_{\phi,\psi,k}$, which we treat as “reliably learned” we utilise the quantiles of bootstrap samples to construct pseudo-confidence-sets $[\hat{\sigma}_{\phi,\psi,i}^{\min}, \hat{\sigma}_{\phi,\psi,i}^{\max}]$ for $\sigma_{\phi,\psi,i}$. We use these as proxies for upper and lower bounds on $\sigma_{\phi,\psi,i}$. These provide us with matching pseudo-bounds on our estimates of $\langle v_{\phi,\psi,i}, h_0 \rangle$. That is, the “confidence set” for $\langle v_i, h_0 \rangle$ is the interval

$$\left[(\hat{\sigma}_{\phi,\psi,i}^{\max})^{-1} \widehat{\mathbb{E}}[Y \hat{u}_{\phi,\psi,i}(Z)], (\hat{\sigma}_{\phi,\psi,i}^{\min})^{-1} \widehat{\mathbb{E}}[Y \hat{u}_{\phi,\psi,i}(Z)] \right],$$

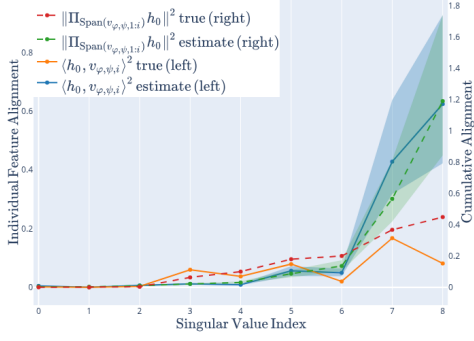
together with the central estimate $(\hat{\sigma}_{\phi,\psi,i}^{(0)})^{-1} \widehat{\mathbb{E}}[Y \hat{u}_{\phi,\psi,i}(Z)]$.

F.1 Spectral alignment in synthetic examples

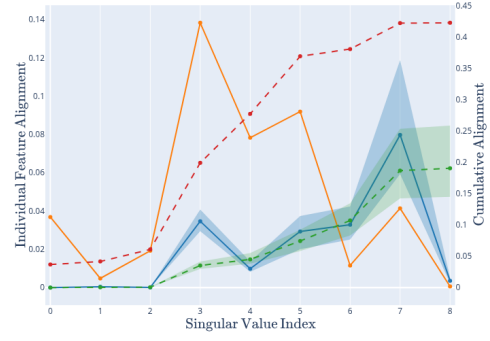
Note that whenever the structural function is available, as is the case in our one-dimensional synthetic example and in the dSprites experiment, one can evaluate the correctness of the approximation proposed above by comparing it to $\langle h_0, v_{\phi,\psi,i} \rangle$. In Figure 6, we observe that the method is reliable if the decay of singular values is sufficiently slow for feature-learning to pick up on the singular function pairs reliably. For $c_\sigma = 0.8$ the approximations are very good while for $c_\sigma = 0.2$ they become entirely unreliable. A more thorough theoretical analysis of these estimates is a topic warranting further investigation but these observations are generally in line with Eq. (18). In particular, small values of $\sigma_{\phi,\psi,i}$ inflate the approximation error.

F.2 dSprites is in the good regime

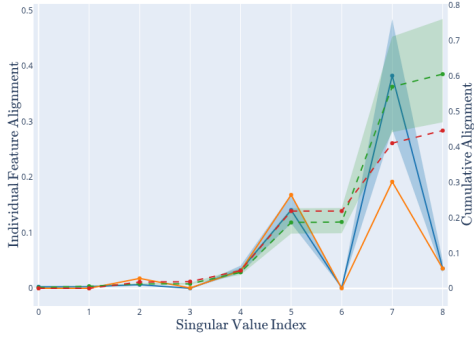
We fitted the spectral-feature models for dSprites using the same architectures as proposed by [36]. Analogously to the fully synthetic experiment, we are able to evaluate the spectrum of $\widehat{\mathcal{T}}$ and directly measure where in it, h_0 is supported. As seen in Figure 7, the top of the learned spectrum is very flat, and hence, in line with what we observed in the previous example, the alignment estimates are reliable. Our experiments confirm that h_0 is usually spanned by the leading 32 spectral features and 85% of its squared norm spanned by the leading 10 functions. Hence, dSprites is indeed an example of the good case.



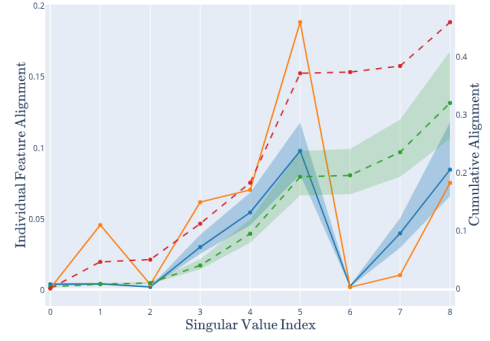
(a) $c_\sigma = 0.2, c_\alpha = 0.2$



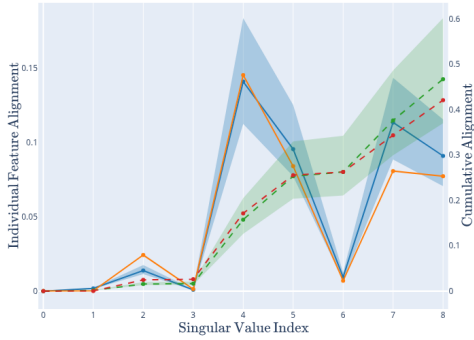
(b) $c_\sigma = 0.2, c_\alpha = 5.0$



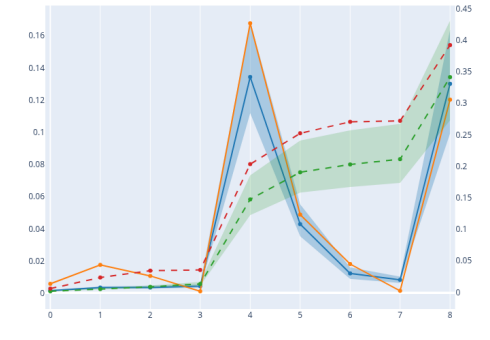
(c) $c_\sigma = 0.6, c_\alpha = 0.2$



(d) $c_\sigma = 0.6, c_\alpha = 5.0$

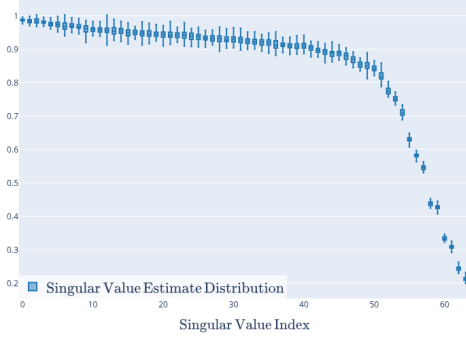


(e) $c_\sigma = 0.8, c_\alpha = 0.2$

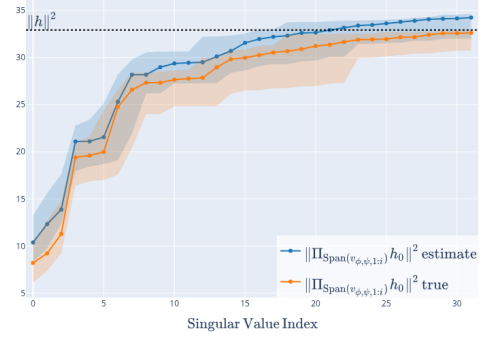


(f) $c_\sigma = 0.8, c_\alpha = 5.0$

Figure 6: Evaluation of the reliability of spectral alignment estimation depending on the real spectral alignment measured in terms of c_α and the rate of singular value decay c_σ .



(a) Distributions of resampled values of $\hat{\sigma}_{\phi, \psi, i}^{(b)}$ evaluated for a single $\hat{\mathcal{T}}_d$ model.



(b) Estimates and true values of the projection length of h_0 onto the leading i features.

Figure 7: Evaluation of dSprites spectral alignment. Left: the confidence intervals correspond to bootstrapped refits of the feature covariance for an individual model. Right: the confidence intervals are obtained from 9 independently trained models with identical hyperparameters but different random seeds.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We specifically point out the settings and detailed contributions of our work in both abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In describing our assumptions as well as discussing our theoretical results, we specifically list our limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We specifically list all assumptions in the main text and refer to each assumption individually for every theoretical result we obtain. All proofs are reported in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a complete description of the data generation process, experimental setup, and hyperparameters used in all simulations, ensuring full reproducibility of our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our code in a zip file.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We provide a complete description of the data generation process, experimental setup, and hyperparameters used in all simulations, ensuring full reproducibility of our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: Our experimental results are based on multiple realizations, standard deviation are reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Our experiments are lightweight and can be reproduced on a standard laptop without specialized hardware.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed the Code of Ethics—this paper conforms with the Code.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our paper mainly focuses on investigating the learning efficiency of a two stage least squares estimator with spectral features. We therefore believe our paper does not have any negative societal impacts. We explain how our theory can improve our understanding of various causal learning algorithms in the introduction which could be potential positive societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.