OUTCOME-AWARE SPECTRAL FEATURE LEARNING FOR INSTRUMENTAL VARIABLE REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We address the problem of causal effect estimation in the presence of hidden confounders using nonparametric instrumental variable (IV) regression. An established approach is to use estimators based on learned ***spectral features***, that is, features spanning the top singular subspaces of the operator linking treatments to instruments. While powerful, such features are agnostic to the outcome variable. Consequently, the method can fail when the true causal function is poorly represented by these dominant singular functions. To mitigate, we introduce ***Augmented Spectral Feature Learning***, a framework that makes the feature learning process ***outcome-aware***. Our method learns features by minimizing a novel contrastive loss derived from an ***augmented*** operator that incorporates information from the outcome. By learning these task-specific features, our approach remains effective even under spectral misalignment. We provide a theoretical analysis of this framework and validate our approach on challenging benchmarks.

1 INTRODUCTION

We study the nonparametric instrumental variable (NPIV) model, a cornerstone of causal inference in the presence of unobserved confounding (Newey & Powell, 2003), which assumes a relationship

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

where the confounder U is conditionally mean zero with respect to the instrument Z. The goal is to recover the causal effect from treatment X to outcome Y by estimating the structural function h_0 from i.i.d. samples of (Y, X, Z). An example is the economic problem of estimating the effect of education on earnings (Card, 1993). X represents years of schooling and Y an individual's wage. A direct regression is likely biased because unobserved factors like innate ability or family background (U) can influence both educational attainment and earning potential. To disentangle this effect, one could use an individual's proximity to a college as an instrument (Z), as living closer may increase years of schooling (X) but is unlikely to be directly correlated with innate ability (U).

The NPIV model can be reformulated as a linear inverse problem (Darolles et al., 2011). Taking the conditional expectation with respect to Z on both sides of Eq. (1) yields the integral equation

$$\mathcal{T}h_0 = r_0, \qquad r_0 \doteq \mathbb{E}[Y \mid Z], \tag{2}$$

where $\mathcal{T}\colon L_2(X) \to L_2(Z)$ is a bounded linear operator that maps every function $h \in L_2(X)$ to its conditional expectation $\mathbb{E}[h(X)|Z] \in L_2(Z)$. Here both the function r_0 and the operator $\mathcal{T}$ are unknown, and we only have access to the set of i.i.d. observations. Throughout this work, we assume that there exists a solution to the NPIV problem, that is r_0 is in the range of $\mathcal{T}$.

State-of-the-art techniques for NPIV estimation rely on learning adaptive features and integrating them into classic algorithms like two-stage least squares (2SLS) (Xu et al., 2021; Petrulionyte et al., 2024) or primal-dual strategies (Dikkala et al., 2020; Liao et al., 2020; Bennett et al., 2023). One successful technique is SpecIV (Sun et al., 2025), which learns neural net features by approximating a low-rank decomposition of the operator $\mathcal{T}$. Meunier et al. (2025) showed SpecIV to be optimal when the structural function h_0 is well-aligned with the top singular functions of $\mathcal{T}$, but degrades otherwise. The issue is that the feature learning process is agnostic to the outcome Y; it only captures

the main aspects of the relationship between the treatment X and the instrument Z. If h_0 lies outside of this dominant subspace, the resulting features are uninformative for the final task, leading to failure. In this paper, we address this limitation by proposing a framework for outcome-aware feature learning in NPIV estimation. We introduce a new spectral objective that incorporates information from the outcome Y, guiding the feature learning to identify components of X that are predictable from Z and predictive of Y. This is equivalent to learning a low-rank decomposition of a perturbed version of $\mathcal{T}$. Our approach ensures good performance even in cases of spectral misalignment, where target-agnostic methods fail. While our work focuses on IV regression for causal effect estimation, our method may also be useful in settings where learning spectral decompositions of conditional operators is relevant. For instance, in off-policy evaluation in reinforcement learning, where value function estimation can be framed as an IV problem (Hu et al., 2024), or when learning evolution operators in fields like molecular dynamics and climate science (Turri et al., 2025).

Contributions. Our contributions are as follows:

1. We identify a fundamental limitation of existing spectral methods for NPIV: the learned features are outcome-agnostic, degrading performance in cases of spectral misalignment. To address this, we propose ***Augmented Spectral Feature Learning*** and introduce an augmented operator, $\mathcal{T}_\delta$, which incorporates information on the outcome into the feature learning problem. This leads to a new, principled contrastive loss function for learning task-specific spectral features.
2. We provide a comprehensive theoretical analysis of our method. This includes a full generalization error bound for the resulting 2SLS estimator, characterizing the settings where the augmented approach remains robust to the spectral misalignment issues of previous methods.
3. We validate our theory on challenging synthetic and semi-synthetic examples, including a new and more challenging version of a dSprites IV benchmark (Xu et al., 2021). The results demonstrate the practical benefits of outcome-aware feature learning in the challenging regimes where standard SpecIV fails. In addition, we include an Off-Policy Evaluation (OPE) experiment in the context of reinforcement learning (Chen et al., 2022), showing that our approach remains robust and competitive in challenging, dynamically changing environments.

Paper Organization. The remainder of the paper is structured as follows. Sec. 2 introduces the notation, reviews the 2SLS estimator, and the SpecIV method (Sun et al., 2025). In Sec. 3, we introduce our outcome-aware framework. Sec. 4 presents our main theoretical results. We situate our contribution within the broader literature in Sec. 5. In Sec. 6 we present numerical experiments that validate our theory and demonstrate the effectiveness of our approach. All proofs are deferred to the appendix.

2 Preliminaries

Function Spaces. Y is defined on $\mathbb{R}$, while X and Z take values in measurable spaces $\mathcal{X}$ and $\mathcal{Z}$, respectively. For $R \in \{X, Z\}$, $L_2(R)$ is the space of square-integrable functions ($\mathbb{E}[f(R)^2] < \infty$).

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

Data Splitting and Empirical Expectations. We consider two independent datasets: $\tilde{\mathcal{D}}_m = \{(\tilde{z}_i, \tilde{x}_i, \tilde{y}_i)\}_{i=1}^m$, to learn features for X and Z, and $\mathcal{D}_n = \{(z_i, x_i, y_i)\}_{i=1}^n$, to estimate the structural function. $\widehat{\mathbb{E}}_m$ and $\widehat{\mathbb{E}}_n$ denote empirical expectations with respect to $\tilde{\mathcal{D}}_m$ and $\mathcal{D}_n$, respectively.

Feature Maps, Covariance Operators and Projections. Let $d \in \mathbb{N}^*$, $\varphi^{(d)} \colon \mathcal{X} \to \mathbb{R}^d$ be a feature map with linearly independent components $\varphi_1^{(d)}, \dots, \varphi_d^{(d)} \in L_2(X)$, and $\Phi^{(d)} \doteq [\varphi_1^{(d)}, \dots \varphi_d^{(d)}]$ be the operator defined as $\Phi^{(d)} \colon \mathbb{R}^d \to L_2(X), \alpha \mapsto \sum_{i=1}^d \alpha_i \varphi_i^{(d)}$. Its adjoint is given by $\Phi^{(d)\star} \colon h \mapsto$

$(\langle h, \varphi_i^{(d)} \rangle_{L_2(X)})_{i=1}^d$. Let $C_{\varphi^{(d)}} \doteq \Phi^{(d)\star}\Phi^{(d)} = \mathbb{E}[\varphi^{(d)}(X)\varphi^{(d)}(X)^\intercal]$ be the (uncentered) covariance operator, $\widehat{C}_{\varphi^{(d)}} \doteq \widehat{\mathbb{E}}_n[\varphi^{(d)}(X)\varphi^{(d)}(X)^\intercal]$ its empirical counterpart, and $\Pi_{\varphi^{(d)}} = \Phi^{(d)}(C_{\varphi^{(d)}})^{-1}\Phi^{(d)\star}$ be the corresponding orthogonal projection operator. Analogous definitions apply for a feature map $\psi^{(d)} \colon \mathcal{Z} \to \mathbb{R}^d$, yielding $\Psi^{(d)}, C_{\psi^{(d)}}, \widehat{C}_{\psi^{(d)}}$ and $\Pi_{\psi^{(d)}}$. The (uncentered) cross-covariance operator between $\varphi^{(d)}$ and $\psi^{(d)}$ is defined as $C_{\psi^{(d)},\varphi^{(d)}} \doteq \Psi^{(d)\star}\Phi^{(d)} = \mathbb{E}[\psi^{(d)}(Z)\varphi^{(d)}(X)^\intercal] = C_{\varphi^{(d)},\psi^{(d)}}^\intercal$. Denote $\widehat{C}_{\psi^{(d)},\varphi^{(d)}}$ its empirical counterpart. We drop the superscript d when it is clear from context.

2SLS in feature space. Given feature maps $\varphi^{(d)} \colon \mathcal{X} \to \mathbb{R}^d$ and $\psi^{(p)} \colon \mathcal{Z} \to \mathbb{R}^p$, a prominent estimator for NPIV is the 2SLS estimator (see, *e.g.*, Blundell et al., 2007), given by

$$\widehat{h}_{\text{2SLS}}(x) = \varphi^{(d)}(x)^\intercal \widehat{\beta}_{\text{2SLS}}, \quad \text{with} \quad \widehat{\beta}_{\text{2SLS}} = \left\{ \widehat{C}_{\varphi^{(d)},\psi^{(p)}} \widehat{C}_{\psi^{(p)}}^{-1} \widehat{C}_{\psi^{(p)},\varphi^{(d)}} \right\}^{-1} \widehat{C}_{\varphi^{(d)},\psi^{(p)}} \widehat{C}_{\psi^{(p)}}^{-1} \widehat{\mathbb{E}}_n[Y\psi^{(p)}(Z)].$$

When $d{=}p$ and the cross-covariance matrix is invertible, it simplifies to $\widehat{\beta}_{\text{2SLS}} = \widehat{C}_{\psi^{(d)},\varphi^{(d)}}^{-1} \widehat{\mathbb{E}}_n[Y\psi^{(d)}(Z)]$.

2SLS with spectral features. Features plugged into the 2SLS estimator can be either fixed or learned adaptively from data (see Section 5 for a discussion on related work). In the SpecIV approach (Sun et al., 2025), features are learned to approximate the top eigenstructure of $\mathcal{T}$. Throughout the paper, we make the following mild assumption (Darolles et al., 2011; Meunier et al., 2025).

Assumption 1. *$\mathcal{T} \colon L_2(X) \to L_2(Z)$ is a compact operator.*

This allows us to write a countable singular value decomposition (SVD) for $\mathcal{T}$,

$$\mathcal{T} = \textstyle\sum_{i \geq 1} \lambda_i u_i \otimes v_i, \quad u_i \in L_2(Z), \quad v_i \in L_2(X), \quad \lambda_1 \geq \lambda_2 \geq \cdots > 0,$$

where u_i's and v_i's are orthonormal basis for $\overline{\mathcal{R}(\mathcal{T})} \subseteq L_2(Z)$ and $\mathcal{N}(\mathcal{T})^\perp \subseteq L_2(X)$, respectively. The operator $\mathcal{T}^{(d)} \doteq \sum_{i=1}^d \lambda_i u_i \otimes v_i$ is the best (in terms of operator or Hilbert–Schmidt norm) rank-d approximation to $\mathcal{T}$. To avoid ambiguity, we assume that $\lambda_d > \lambda_{d+1}$. We do not assume that $\mathcal{T}$ is injective, since we can always target the minimum-norm solution (Florens et al., 2011)

$$\bar{h}_0 \doteq \mathcal{T}^\dagger r_0 = \textstyle\sum_{i \geq 1} \frac{1}{\lambda_i} \langle r_0, u_i \rangle_{L_2(Z)} v_i.$$

When $\mathcal{T}$ is injective, $h_0 = \bar{h}_0$. In what follows, we denote $\bar{h}_0$ as h_0 and do not distinguish between the structural function and the minimal solution to the NPIV problem.

Given feature maps $\varphi_\theta^{(d)} \colon \mathcal{X} \to \mathbb{R}^d$ and $\psi_\theta^{(d)} \colon \mathcal{Z} \to \mathbb{R}^d$, parametrized by neural networks, Sun et al. (2025) proposed to learn the features by minimizing the empirical counterpart to the following loss

$$\mathcal{L}_0^{(d)}(\theta) \doteq \mathbb{E}_X \mathbb{E}_Z[(\varphi_\theta^{(d)}(X)^\intercal \psi_\theta^{(d)}(Z))^2] - 2\mathbb{E}[\varphi_\theta^{(d)}(X)^\intercal \psi_\theta^{(d)}(Z)], \tag{3}$$

where the first expectation is over the product of the marginals of X and Z, and the second is over the joint distribution. It is shown by Meunier et al. (2025, Theorem 2) that $\mathcal{L}_0^{(d)} \geq -\|\mathcal{T}\|_{\text{HS}}^2$, and that the minimum is achieved if and only if $\Psi_\theta^{(d)}\Phi_\theta^{(d)\star} = \mathcal{T}_d$. Therefore, by minimizing the empirical counterpart to Eq. (3), one learns features that approximate the best low-rank approximation to $\mathcal{T}$.

3 Outcome-Aware Spectral Feature Learning

As discussed in Section 1, standard SpecIV learns features that are agnostic to the outcome Y. This can lead to poor performance when the structural function h_0 is not well-aligned with the top singular functions of $\mathcal{T}$ (Meunier et al., 2025). To mitigate this, we augment the SpecIV loss with a regularization term that incorporates information from the outcome Y by projecting it onto the orthonormal basis of the Z-features. The resulting loss is defined as

$$\mathcal{L}_\delta^{(d)}(\theta) = \mathcal{L}_0^{(d)}(\theta) - \delta^2 \mathbb{E}[Y\psi_\theta^{(d)}(Z)]^\intercal C_{\psi_\theta^{(d)}}^{-1} \mathbb{E}[Y\psi_\theta^{(d)}(Z)]. \tag{4}$$

We propose to learn features by minimizing the empirical counterpart of Eq. (4) over the training set $\tilde{\mathcal{D}}_m$. The regularization term, controlled by the hyperparameter δ, encourages the learned instrument features to be predictive of Y. As backpropagation through the inverse covariance matrix can be numerically unstable, we instead minimize the following equivalent loss jointly over θ and $\omega \in \mathbb{R}^d$:

$$\mathcal{L}_\delta^{(d)}(\theta, \omega) = \mathcal{L}_0^{(d)}(\theta) - 2\delta \mathbb{E}[Y\psi_\theta^{(d)}(Z)]^\intercal \omega + \omega^\intercal C_{\psi_\theta^{(d)}} \omega. \tag{5}$$

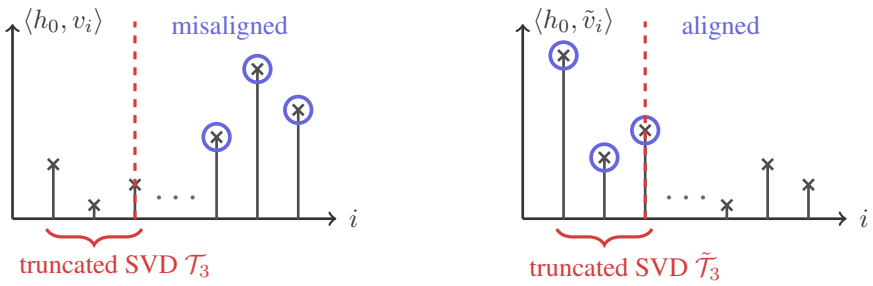


Figure 1: In case of severe misalignment of h_0 and $\mathcal{T}$ (left), an ideal solution would aim to find another operator $\tilde{\mathcal{T}}$ whose top singular functions $\tilde{v}_i$ capture the signal in h_0 (right).

For any fixed θ, this loss is convex in ω and its minimum is attained at $\omega_\theta^{(d)} = \delta C_{\psi_\theta^{(d)}}^{-1} \mathbb{E}[Y \psi_\theta^{(d)}(Z)]$. Substituting this solution back into Eq. (5) recovers the profile loss in Eq. (4).

Operator learning perspective. We now show this modification is equivalent to learning a low-rank approximation of an augmented version of the operator $\mathcal{T}$. For $\delta \in \mathbb{R}$, we define the operator

$$\mathcal{T}_\delta : L_2(X) \times \mathbb{R} \to L_2(Z), (h, a) \mapsto \mathcal{T}h + a \cdot \delta \cdot r_0.$$

$\mathcal{T}_\delta$ augments $\mathcal{T}$ with an additional "column" aligned with r_0, as captured by the compact notation $\mathcal{T}_\delta = [\mathcal{T} \mid \delta r_0] = \mathcal{T}[I_{L_2(X)} \mid \delta h_0]$. As $\mathcal{T}$ is compact, $\mathcal{T}_\delta$ is also compact and admits an SVD

$$\mathcal{T}_\delta = \textstyle\sum_{i \geq 1} \sigma_{*,i} \psi_{*,i} \otimes (\varphi_{*,i}, \omega_{*,i}), \quad \psi_{*,i} \in L_2(Z), \quad (\varphi_{*,i}, \omega_{*,i}) \in L_2(X) \times \mathbb{R}, \tag{6}$$

with $\sigma_{*,1} \geq \sigma_{*,2} \geq \cdots > 0$. The following proposition formalizes the connection: minimizing the augmented loss $\mathcal{L}_\delta^{(d)}$ is equivalent to finding the best rank-d approximation of the augmented operator $\mathcal{T}_\delta$. We denote this best approximation by $\mathcal{T}_\delta^{(d)} = \sum_{i=1}^{d} \sigma_{*,i} \psi_{*,i} \otimes (\varphi_{*,i}, \omega_{*,i})$.

Proposition 1. *Given $\delta \in \mathbb{R}$, for all parameters θ and ω it holds that $\mathcal{L}_\delta^{(d)}(\theta, \omega) \geq -\|\mathcal{T}_\delta\|_{\text{HS}}^2$. The lower bound is achieved if and only if the learned operator $\Psi_\theta^{(d)}[\Phi_\theta^{(d)\star} \mid \omega]$ is equal to $\mathcal{T}_\delta^{(d)}$.*

Intuitively, augmenting the operator with outcome information amplifies the components of h_0 that would otherwise lie in the low singular value region of $\mathcal{T}$, thereby improving their alignment with the top spectral features of the augmented operator, as illustrated in Figure 1. Section 4 formalizes this effect by bounding the distance between the learned subspaces and the signal subspaces of $\mathcal{T}$. One can also encourage the learned features to retain predictive information about additional aspects of the outcome by extending the augmentation to multiple functions of Y, such as higher conditional moments $\mathbb{E}[Y^k \mid Z]$. We discuss this higher rank extension in Appendix E. The learning objective and optimality characterization via truncated SVD remain unchanged. A complete theoretical analysis of general rank K perturbations requires further development of the perturbation framework and is left for future work.

4 Analysis

We now present the statistical guarantees for our estimator of h_0. Our first result is a non-asymptotic, high-probability error bound for the 2SLS estimator that is agnostic to the choice of representation $(\varphi_\theta^{(d)}, \psi_\theta^{(d)})$. This improves upon prior results, such as Chen & Christensen (2018), which typically provide guarantees in expectation. To this end, we introduce three standard assumptions. The first requires the whitened features to be uniformly bounded.

Assumption 2 (Representation Boundedness)**.** *Denoting the covariances of the instrument and feature representations by $C_{Z,\theta} = \mathbb{E}[\psi_\theta^{(d)}(Z)\psi_\theta^{(d)}(Z)^\intercal]$ and $C_{X,\theta} = \mathbb{E}[\varphi_\theta^{(d)}(X)\varphi_\theta^{(d)}(X)^\intercal]$, respectively, there exists $\rho \geq 1$ such that representations satisfy*

$$\operatorname*{ess\,sup}_{x \sim \mathbb{P}_X} \max_{j \in [d]} \left\{ |(C_{X,\theta}^{-1/2} \varphi_\theta^{(d)}(x))_j| \right\} \bigvee \operatorname*{ess\,sup}_{z \sim \mathbb{P}_Z} \max_{j \in [d]} \left\{ |(C_{Z,\theta}^{-1/2} \psi_\theta^{(d)}(z))_j| \right\} \leq \rho.$$

Assumption 3 (Measure of ill-posedness)**.** *Denoting the cross-covariance between the instrument and feature representations by $C_{ZX,\theta} = \mathbb{E}[\psi_\theta^{(d)}(Z)\varphi_\theta^{(d)}(X)^\intercal]$, the measure of ill-posedness $c_{\varphi_\theta^{(d)}, \psi_\theta^{(d)}} := \sigma_d(C_{Z,\theta}^{-1/2} C_{ZX,\theta} C_{X,\theta}^{-1/2})$, where $\sigma_d(\cdot)$ denotes the d-th singular value, is positive.*

The measure of ill-posedness $c_{\varphi_\theta^{(d)},\psi_\theta^{(d)}}$, is equal to $\sigma_d(\Pi_{\Psi_\theta^{(d)}}\mathcal{T}\Pi_{\Phi_\theta^{(d)}})$ and captures the stability of the inverse problem when restricted to the learned feature spaces. Its positiveness implies non-singularity of $C_{ZX,\theta}$, which guarantees that there exists a vector $\beta_\theta \in \mathbb{R}^d$ such that $h_\theta(x)=\varphi_\theta^{(d)}(x)^\intercal\beta_\theta$ satisfies the instrumental moment condition: $C_{ZX,\theta}\beta_\theta=\mathbb{E}[Y\psi_\theta^{(d)}(Z)]=\mathbb{E}[r_0(Z)\psi_\theta^{(d)}(Z)]=\mathbb{E}[h_0(X)\psi_\theta^{(d)}(Z)]$. Assumption 3 is reasonable provided that some conditions on the eigenvalues of $\mathcal{T}$ are satisfied and our features capture an accurate representation of the singular spaces of $\mathcal{T}_d$ (see Proposition 8 in the appendix).

Our final assumption concerns the tail behavior of the model's error terms. For the precise definition of sub-Gaussian random variables, we refer to Definition 1 in Appendix C.2.

Assumption 4 (Sub-Gaussian distributions)**.** *The model noise $U=Y-h_0(X)$ and the function approximation error $(h_0-h_\theta)(X)$ are sub-Gaussian random variables.*

With these conditions, we can state our main result for the 2SLS estimator, which is given by $\widehat{h}_\theta(x)=\varphi_\theta^{(d)}(x)^\intercal\widehat{\mathbb{E}}_n[\psi_\theta^{(d)}(Z)\varphi_\theta^{(d)}(X)^\intercal]^{-1}\widehat{\mathbb{E}}_n[Y\,\psi_\theta^{(d)}(Z)]$.

Theorem 1. *Let Assumptions 2-4 be satisfied. Given $\tau\in(0,1)$, let $n\geq 16\,d\,\rho^2\,\log^2(4d/\tau)c^{-2}_{\varphi_\theta^{(d)},\psi_\theta^{(d)}}$. Then there exists an absolute constant $C>0$ s.t. with probability at least $1-\tau$:*

$$\|\widehat{h}_\theta - h_0\|_{L_2(X)} \leq C\Big(\|h_0-h_\theta\|_{L_2(X)} + \frac{1}{c_{\varphi_\theta^{(d)},\psi_\theta^{(d)}}}\sqrt{\frac{d}{n}}\sqrt{\sigma_U^2+\frac{\rho^2}{n}}\,\log\frac{4}{\tau}\Big),$$

where σ_U^2 is the noise variance.

This theorem provides a non-asymptotic excess-risk bound that separates the error into a deterministic approximation error ($\|h_0-h_\theta\|$) and a statistical error. The latter depends on the dimension d, sample size n, and the feature-dependent ill-posedness. It improves on existing guarantees (Theorem B.1. Chen & Christensen, 2018) by holding in high probability under a sub-Gaussian assumption.

Controlling the approximation error with learned representations. We now specialize the general bound from Theorem 1 to the features learned by our outcome-aware method. In light of Proposition 1, recalling that the leading left and right singular subspaces of $\mathcal{T}_\delta$ are the ranges of $\Psi_*^{(d)}=[\psi_{*,1}\,|\,\dots\,|\psi_{*,d}]$ and $[\,\Phi_*^{(d)}\,|\,\omega_*\,]^*$, where $\Phi_*^{(d)}=[\varphi_{*,1}\,|\,\dots\,\varphi_{*,d}]$, see Eq. (6), the feature learning stage of our approach produces $\psi_{\widehat{\theta}_m}^{(d)}$ and $\varphi_{\widehat{\theta}_m}^{(d)}$, where the parametrization $\widehat{\theta}_m$ is learned from dataset $\tilde{\mathcal{D}}_m$. To quantify their quality, as proposed in Kostic et al. (2024) and Meunier et al. (2025), we use the optimality gap

$$\mathcal{E}_d(\theta,\omega,\delta)=\|\mathcal{T}_\delta^{(d)}-\Psi_\theta^{(d)}[\Phi_\theta^{(d)*}\,|\,\omega]\|_{L^2(Z)\times\mathbb{R}\to L^2(X)}, \tag{7}$$

noting that bounding it requires architecture-specific generalization bounds for DNN training with the spectral contrastive loss, an open problem in its own right.

With this setup, our analysis hinges on relating the singular subspaces of the augmented operator $\mathcal{T}_\delta$ back to the original singular subspaces of $\mathcal{T}$. To do this, we first partition the singular components of $\mathcal{T}$ into a d-dimensional *"signal"* subspace and an infinite-dimensional *"noise"* subspace. Namely, let $\overline{N}\,\dot\cup\,\underline{N}=\mathbb{N}$ be the partition, where we take $|\overline{N}|=d$ spectral features for the signal. We define the signal components $s_d=\overline{V}_d\overline{V}_d^*h_0=\sum_{i\in\overline{N}}\overline{\alpha}_i v_i$ and noise components $q_d=\underline{V}_d\underline{V}_d^*h_0=\sum_{i\in\underline{N}}\underline{\alpha}_i v_i$, where the partition of the SVD of $\mathcal{T}$ is $\mathcal{T}=\overline{U}_d\overline{\Lambda}_d\overline{V}_d^*+\underline{U}_d\underline{\Lambda}_d\underline{V}_d^*$ with $\overline{\Lambda}_d=\mathrm{diag}(\lambda_i)_{i\in\overline{N}}$. The augmented operator $\mathcal{T}_\delta$ can be viewed as a perturbation of $\mathcal{T}$ relative to the positioning of the signal:

$$\mathcal{T}_\delta = [\,\mathcal{T}\,|\,\delta\mathcal{T}h_0\,] = \overline{U}_d\overline{\Lambda}_d\,[\,\overline{V}_d^*\,|\,\delta\overline{\alpha}\,] + \underline{U}_d\,\underline{\Lambda}_d\,[\,\underline{V}_d^*\,|\,0\,] + [\,0\,|\,\delta\,\underline{U}_d\,\underline{\Lambda}_d\,\underline{\alpha}\,].$$

This decomposition shows that if the noise $\|q_d\| = \|\underline{\alpha}\|_{\ell^2}$ is small, $\mathcal{T}_\delta$ is well-approximated by the first two terms (the noiseless part) that expose how the left and right singular subspaces of $\mathcal{T}$, relative to the partition, align with those of $\mathcal{T}_\delta$. If the singular value gap between first and the second term

$$\gamma_d(\delta)=\|[\overline{\Lambda}_d(I+\delta^2\overline{\alpha}\,\overline{\alpha}^\top)^{1/2}]^{-1}]\|^{-1} - \|\underline{\Lambda}_d\| \tag{8}$$

is positive, the dominant singular subspace of the noiseless perturbation of $\mathcal{T}$ is exactly $\mathcal{R}(\overline{U}_d)$. Therefore, by carefully applying perturbation results, we are able to control the differences in orthogonal projections $\|\Pi_{\overline{U}_d}-\Pi_{\Psi_{\widehat{\theta}_m}^{(d)}}\| \leq \|\Pi_{\overline{U}_d}-\Pi_{\Psi_*^{(d)}}\| + \|\Pi_{\Psi_*^{(d)}}-\Pi_{\Psi_{\widehat{\theta}_m}^{(d)}}\|$ to align the appropriate

left subspaces. The first term depends on the interplay between the parameter δ, the positioning, and the size of the signal, and is bounded by $\delta\|q_d\|_{L^2(X)} \,/\, \gamma_d(\delta)$. The second term is the error from learning the features of $\mathcal{T}_\delta$ from dataset $\tilde{\mathcal{D}}_m$ with our representation learning method, quantified by the optimality gap $\mathcal{E}_d(\widehat{\theta}_m, \widehat{\omega}_m, \delta)$ given in Eq. (7). A similar decomposition holds for the right subspaces. By bounding these components, we can control the total approximation error. A detailed proof is provided in Theorem 4 in Appendix B.3; in the following, we present the consequences for the "good" and "bad" scenarios from Meunier et al. (2025).

Learning in the *"good"* scenario. When the signal s_d is spanned by the top-d singular functions of $\mathcal{T}$ (i.e., $\overline{N} = \{1, \dots, d\}$), our method works efficiently even with $\delta = 0$. In this case, the spectral gap $\gamma_d(0) = \lambda_d - \lambda_{d+1}$ is positive by assumption. A direct corollary of Theorems 1, and 4 yields

$$\|h_0 - \widehat{h}_\theta\|_{L_2(X)} \lesssim \|q_d\|_{L^2(X)} + \frac{\mathcal{E}_d(\widehat{\theta}_m, \widehat{\omega}_m, 0)}{\lambda_d} + \frac{1}{\lambda_d}\sqrt{\frac{d}{n}}\log\tau^{-1} \qquad \text{w.p.a.l. } 1-\tau.$$

This result is the high probability version of Meunier et al. (2025, Theorem 3).

Learning in the *"bad"* scenario. Our outcome-aware method is designed to succeed even when the structural function h_0 is highly aligned with a "bad" singular function v_k (for large k). Meunier et al. (2025) showed that standard SpecIV would require learning a high-dimensional feature space of dimension $d \geq k$ (out of which $k-1$ are spurious). In contrast, our method can isolate the relevant signal by setting $d = 1$ with the signal concentrating mostly on space defined by v_k ($\overline{N}=\{k\}$). As shown in Appendix B.3, choosing a δ large enough guarantees that the one-dimensional spectral gap $\gamma_1(\delta)$ becomes positive. Applying Theorem 4 of Appendix B.3 in this setting shows:

$$\|h_0 - \widehat{h}_\theta\|_{L_2(X)} \lesssim \frac{1}{\lambda_k^2}\frac{\|q_1\|_{L^2(X)}}{\|s_1\|_{L^2(X)}} + \frac{\mathcal{E}_d(\widehat{\theta}_m, \widehat{\omega}_m, \delta)}{\lambda_k} + \frac{\log\tau^{-1}}{\lambda_k\sqrt{n}} \qquad \text{w.p.a.l. } 1-\tau,$$

indicating that whenever signal-to-noise ratio dominates the decay $\|s_1\|_{L^2(X)}/\|q_1\|_{L^2(X)} \gg \lambda_k^{-2}$, our method can recover the structural function with one feature, while for the standard SpecIV one would need to learn k of them, as illustrated in Figure 1.

5 Related Work

This section provides an overview of the research areas that are most relevant to our study.

2SLS methods. The classical approach to IV regression is the 2SLS method. In its nonparametric form, the first stage involves estimating the conditional expectation of the treatment given the instrument, and the second stage uses these predictions to estimate the structural function. Early influential works used sieve or series estimators, approximating unknown functions with basis functions like polynomials or splines (Newey & Powell, 2003; Hall & Horowitz, 2005; Blundell et al., 2007; Chen & Pouzo, 2012; Chen & Christensen, 2018). Other approaches include Tikhonov regularization to stabilize the inverse problem (Darolles et al., 2011) or frame the problem in reproducing kernel Hilbert spaces (Singh et al., 2019; Meunier et al., 2024). More recently, deep learning has been used to handle the nonparametric components of the 2SLS procedure. DeepIV (Hartford et al., 2017) uses a mixture density network to estimate the conditional distribution of the treatment given the instruments in the first stage, and then a second network for the structural function in the second stage. Deep Feature IV (DFIV; Xu et al., 2021) uses neural networks to learn optimal features of the instruments, which are then used as inputs for the first-stage regression.

Saddle-point methods. An alternative to 2SLS is to frame NPIV as a minimax optimization problem. These methods, often rooted in a generalized method of moments (GMM) framework, seek an equilibrium between a player that minimizes a loss function and an adversary that maximizes the violation of the moment conditions. This approach can bypass the direct estimation of conditional expectations. Different formulations exist, such as those based on the Lagrangian of a constrained least-norm problem (Bennett et al., 2023; Liao et al., 2020) or on maximizing the moment deviation directly (Lewis & Syrgkanis, 2018; Dikkala et al., 2020; Wang et al., 2022). These methods are particularly well-suited for high-dimensional settings and integration with deep learning models.

Spectral features learning. When the instrument-treatment relationship is complex, learning good features is crucial. Spectral methods use techniques like SVD to find a low-dimensional representation of the conditional expectation operator $\mathcal{T}$. Such an SVD can typically be estimated via

minimizing a contrastive loss that has been used in various contexts (Sun et al., 2025; Hu et al., 2024; Kostic et al., 2024; Turri et al., 2025). While these methods are powerful, the learned features are agnostic to the outcome; they only capture the dominant modes of the instrument-treatment relationship, which may not be enough for predicting the outcome.

Outcome-aware and adaptive methods. Our work is part of a growing literature on adaptive methods for NPIV. While spectral methods like SpecIV (Sun et al., 2025) are powerful, their outcome-agnostic nature can be a significant drawback, as we demonstrated. The features are learned based only on the instrument-treatment relationship and may not be informative for predicting the outcome. Our approach addresses this by making the feature learning process outcome-aware and ensures that the learned representations are not just predictive of the treatment, but are also relevant for the causal relationship of interest.

6 Experiments

We first illustrate the utility of our method on a synthetic example in which all parameters of $\mathcal{T}$ are controlled. To further support our claims, we benchmark outcome-aware spectral learning on two challenging benchmarks based on the dSprites dataset (Matthey et al., 2017). Finally, we include an Off-Policy Evaluation (OPE) experiment in the context of reinforcement learning, which demonstrates that the method remains robust and competitive in demanding environments. We find that a small positive δ always leads to improved performance. In most challenging cases, the resulting method outperforms standard SpecIV ($\delta = 0$) by a wide margin.

6.1 Synthetic data

Following Meunier et al. (2025), we generate a (Z, X, Y) dataset from a conditional expectation operator $\mathcal{T} = \mathbb{1}_Z \otimes \mathbb{1}_X + \sum_{i=1}^{d-1} \sigma_i u_i \otimes v_i$ with explicit σ_i, v_i, u_i. We take σ_i to decay linearly from a fixed σ_1 to $\sigma_{d-1} = c_\sigma \sigma_1$, with $c_\sigma \in [0, 1]$. u_i, v_i are random orthonormal bases of the span of $(\sin(\ell \cdot x))_{\ell=1}^{d-1}$ on $[-\pi, \pi]$. By setting the structural function $h_0 = \sum_{i=1}^{d-1} \alpha_i v_i$ with the constraint $\|\alpha\|_{\ell_2} = 1$, and having the coefficients α_i change linearly in i, we are able to control the alignment of h_0 with the spectrum of $\mathcal{T}$. We parametrise the rate at which α_i change with $c_\alpha \doteq \alpha_{d-1}/\alpha_1$.

Synthetic data results. Figure 2 shows the distribution $\|\widehat{h}_\theta - h_0\|^2$ for different values of δ, normalised so that for each c_α, the mean loss for $\delta = 0$ equals 1. The reported figures show how the losses change relative to the baseline performance of SpecIV. In cases of very poor spectral alignment, corresponding to $c_\alpha = 5.0$, increasing δ leads to improvement in the IV regression loss even when the singular values decay quickly. Moreover, when c_σ is sufficiently large, that is more singular functions can be learned easily, we observe improvement even on a very well-aligned case when $c_\alpha = 0.2$. A more extensive evaluation of how the benefit of using δ change with the rate of decay of singular values can be found in Appendix D.3.

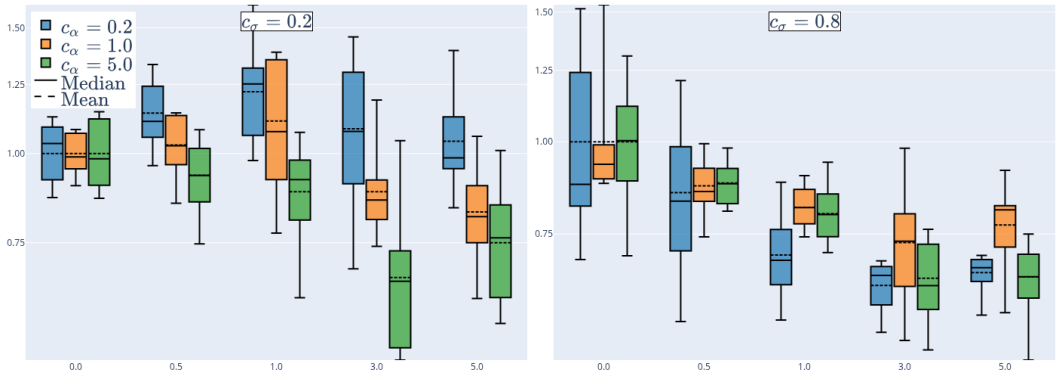


Figure 2: Distributions of relative IV regression MSEs ($\|\widehat{h}_\theta - h_0\|^2$) for the synthetic example with $\delta \in \{0, 0.5, 1.0, 3.0, 5.0\}$ and $c_\sigma \in \{0.2, 0.8\}$

6.2 dSPRITES DATA

The original dSprites dataset of Matthey et al. (2017) consists of 64×64 noisy images (sprites) of hearts, squares, and ellipses with varying position, size, and orientation. In the standard IV benchmarking setting, as described in, *e.g.*, Sun et al. (2025), only the heart images are used. The structural function takes the form $h_0(x) = (\|A \circ X\|^2 - 3000)/500$, where $A_{ij}{=}|32{-}j|/32$ measures the distance from the central vertical bar in the image and $\circ$ denotes the pointwise (Hadamard) product. The instrumental variable is defined as $Z{=}$(sprite orientation, sprite x-position, sprite scale), and the outcome is $Y{=}h(X){+}32$((sprite y-position)$-32){+}\varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.5)$. For reasons discussed below, we shall refer to this h_0 as h_{old}.

New structural function. As first argued in Meunier et al. (2025) and further discussed in Appendix D.1, the standard dSprites benchmark is an instance of the "good" case where h_0 is well-aligned with the leading singular functions of $\mathcal{T}$. Hence, we propose an alternative, more challenging structural function. We refer to the new function as h_{new}. It is based on sprite images of ellipses, as opposed to the hearts in the original case, and approximates the ellipse's orientation, which we expect to be a property that is only recovered from singular functions associated to small singular values of $\mathcal{T}$. The details of our argument and construction can be found in Appendix D.1.

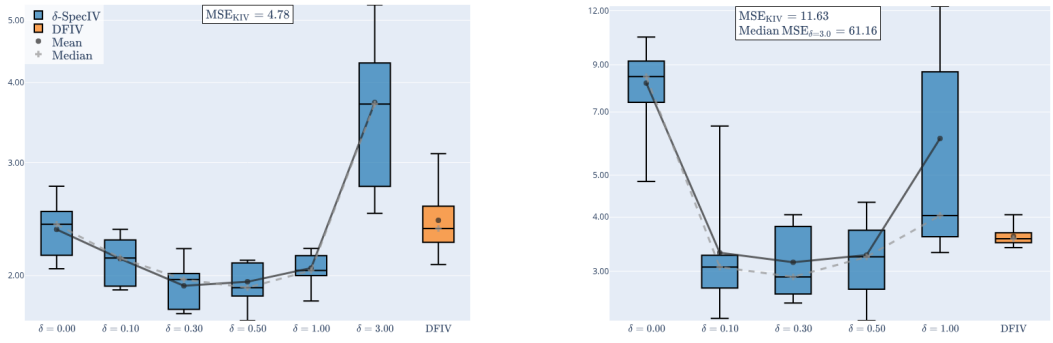


Figure 3: Distribution of $\|\widehat{h}_\theta - h_0\|^2$ on h_{old} (**left**) and h_{new} (**right**) evaluated for a range of δ values, compared to those attained by DFIV and KIV (Singh et al., 2019). Our method is evaluated on 9 independently fitted models with identical hyperparameters for each x-axis value. 16 DFIV models were fitted in both settings.

Experiment results. We compare to DFIV (Xu et al., 2021), which is currently the most competitive method to benchmark against. In the setting of h_{old} we observe an average 20% improvement from selecting a small positive δ over using standard SpecIV. The standard benchmark is well-aligned but there is still some benefit in this setting. By comparison, for h_{new}, we see that vanilla spectral learning ($\delta{=}0$) severely underperforms DFIV. This is in line with the fact that the top of the eigenspectrum of $\mathcal{T}$ is not an optimal set of features for the downstream problem. Increasing δ allows the model to tune the learned features, increasing the projection of h_{new} onto their span, and hence match or slightly exceed the performance of DFIV. Since good performance is contingent on the choice of δ, we next investigate strategies for δ selection.

6.3 OFF-POLICY EVALUATION

To further illustrate the applicability of our augmented method, we showcase its performance in Off-Policy Evaluation (OPE). OPE is a fundamental problem in reinforcement learning that can be approached via NPIV (Chen et al., 2022), but poses a significant challenge for standard SpecIV because it involves a noncompact operator that cannot be learned through the usual contrastive framework. Instead, one has to learn a simpler compact operator while iteratively updating the outcome Y, which SpecIV is agnostic to. Following the experimental pipeline of Chen et al. (2022), we showcase scenarios where our proposed modification significantly improves SpecIV and performs comparably well to state-of-the-art methods for NPIV applied to OPE (Xu et al., 2021; Chen et al., 2022; Petrulionyte et al., 2024), suggesting that our method broadens the applicability of spectral feature-based NPIV for OPE.

OPE context. The goal of OPE is to estimate the value of a *known* target policy, $\pi(a|s)$, where a denotes an action and s denotes a state. The challenge is that we only have access to a fixed, "offline" dataset of transitions, $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_i$, where r denotes a reward, and s' a next state transition. The dataset could have been collected by one or a mixture of potentially unknown policies of potentially unknown analytical form, denoted by $\pi_b(a|s)$. This problem is central to offline reinforcement learning, as it allows one to select the best-performing policy among a collection of candidate policies without having to directly interact with the environment, which can be costly or unethical, as is often the case, e.g., in healthcare (Gottesman et al., 2018). For a detailed overview of OPE, see the work of Levine et al. (2020).

OPE via NPIV. The standard objective in OPE is to estimate the Q-function Q_π of a policy π (uniformly well or in a suitable L_2 norm). Xu et al. (2021) showed that Q-function estimation can be framed as a NPIV problem (see Appendix D.7.2). However, their formulation is not amenable to spectral feature learning as the conditional expectation operator $\mathcal{T}$ they obtain is noncompact (Chen et al., 2022, Eq. (19)). To overcome this, we instead recover Q_π through a modified NPIV problem where $X = (s', a')$ and $Z = (s, a)$ are such that $a' \sim \pi(\cdot \mid s')$ and $a \sim \pi_b(\cdot \mid s)$. Q_π is then identified as the solution to $\mathcal{T}Q_\pi = \mathbb{E}[Y(Q_\pi) \mid Z]$, where $Y(Q_\pi) = -\gamma^{-1}(R - Q_\pi(X))$, where $R \mid Z$ is the random reward associated to Z, and γ is the discount factor. As the outcome $Y(Q_\pi)$ depends on Q_π, which is the very function we are estimating, we introduce an iterative procedure to estimate Q_π. We start with a random guess Q_0 (e.g. $Q_0 = 0$) and build $Y_0 = Y(Q_0)$. We then estimate Q_1 with Augmented Spec IV. We repeat this process for K steps. We defer the details of the iterative procedure and the derivation of the new NPIV problem for OPE to Appendix D.7.2. The key point is that the iterative process introduces a potential for dynamic spectral misalignment, as the target Y_k changes at every iteration. If the spectral features required to estimate Y_k are not the same as the dominant spectral features of $\mathcal{T}$, or if this direction shifts as Q_k converges, an outcome-agnostic method will fail. This is precisely the scenario where Augmented SpecIV can help.

Experiment results. We follow the experimental setting of Chen et al. (2022). In particular, we evaluate the performance of DFIV, SpecIV ($\delta = 0$), and Augmented SpecIV (AugSpecIV; $\delta > 0$) on estimating the value of policies learned by Deep Q-Networks (Mnih et al., 2015) across randomized versions of the BSuite Cartpole (Barto et al., 1983), Mountain Car (Moore, 1990), and Catch environments. The OPE datasets are the pure offline versions from Chen et al. (2022), containing approximately $n = 700\text{k}/150\text{k}/20\text{k}$ (Cartpole/Mountain Car/Catch) transition tuples (s, a, r, s') The results are shown in Appendix D.7, Figures 9 to 11. Compared with Chen et al. (2022), DFIV performed slightly worse, likely due to randomness in hyperparameter sampling. SpecIV and AugSpecIV both achieved strong performance on Catch but struggled on Mountain Car. Consistent with prior findings (Chen et al., 2022), no OPE method performed uniformly well across all tasks. In our experiments, DFIV performed poorly on Cartpole but well on Mountain Car, whereas AugSpecIV showed the opposite trend. SpecIV also underperformed on Cartpole, likely due to spectral misalignment. Since δ was automatically tuned as a hyperparameter and took the values of $1/10^{-3}/10^{-2}$ for Cartpole/Mountain Car/Catch, these results suggest that our approach can adapt to the underlying spectral alignment structure.

6.4 Selecting δ

Balancing the terms in the loss. Let θ denote the trainable parameters of the feature-learning networks. Recall the definition of the augmented loss (Eq. (4))

$$\mathcal{L}_\delta^{(d)}(\theta) = \mathcal{L}_0^{(d)}(\theta) + \mathcal{R}_\delta^{(d)}(\theta), \quad \mathcal{R}_\delta^{(d)}(\theta) \doteq -\delta^2 \mathbb{E}[Y\psi_\theta^{(d)}(Z)]^\intercal C^{-1}_{\psi_\theta^{(d)}} \mathbb{E}[Y\psi_\theta^{(d)}(Z)].$$

Given that our ability to learn the actual truncated SVD of $\mathcal{T}_\delta$ hinges on the convergence of the feature neural networks to the population-level optimum, the optimal choice of δ is in part an empirical challenge. In particular, consider what happens if δ is sufficiently large that $\mathcal{R}_\delta^{(d)}$ dominates the joint loss value. The features $\psi^{(\theta)}$ minimising $\mathcal{R}_\delta^{(d)}$ are non-unique. Any choice such that r_0 lies in their span is equally good. Therefore, the rest of the Z-features should be used to learn an approximation of the conditional expectation. However, if the gradients with respect to $\mathcal{R}_\delta$ are too large, they effectively drown out the signal needed to learn $\mathcal{T}$ and cause a significant decrease in $\mathcal{L}_0^{(d)}$.

A useful heuristic for selecting δ is to treat $\mathcal{R}_\delta^{(d)}$ as an additional regularisation on the learned features, and tune its strength so that it influences the learned features without leading the models to neglect the original $\mathcal{L}_0^{(d)}(\theta)$ term. A heuristic, which we find leads to good results, is to increase δ as

long as doing so leads to big drops in $\mathcal{R}_\delta^{(d)}(\theta)$ with minor changes in $\mathcal{L}_0^{(d)}(\theta)$. Small values of δ are always observed to lead to improved spectral alignment. As long as increasing δ remains "free" in that our ability to approximate $\mathcal{T}$, as measured by $\mathcal{L}_0^{(d)}(\theta)$, does not change by much, we continue to increase it. As seen in Figure 4, the $\mathcal{R}_\delta^{(d)}$ term eventually becomes dominant and $\mathcal{L}_0^{(d)}$ increases by a lot. Those settings, in our experiments, correspond to parameters which lead to bad results in IV regression. We also note that the performance of the method is not sensitive to minor changes to the value of δ. We see in Figures 2 and 3 that all sufficiently small values of δ lead to improvement.

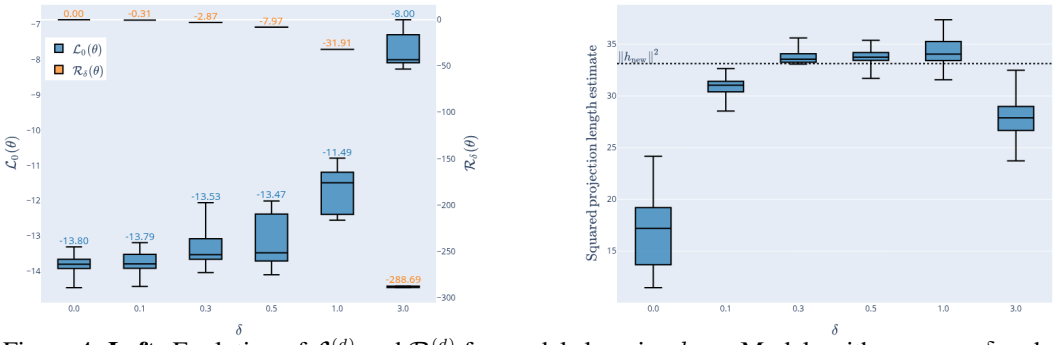


Figure 4: **Left:** Evolution of $\mathcal{L}_0^{(d)}$ and $\mathcal{R}_\delta^{(d)}$ for models learning h_{new}. Models with non-zero δ and a small $\mathcal{L}_0^{(d)}$ (close to the value attained at $\delta = 0$) demonstrate the best results. Each bar's mean value is noted above it. **Right:** Estimation of $\|\Pi_{\varphi_\star^{(d)}} h_{\text{new}}\|^2$ for a range of δ values.

Estimating alignment with h_0. By Proposition 1, our strategy targets the features $\varphi_\star^{(d)}$ from the SVD of $\mathcal{T}_\delta$ as a basis in which to learn h_0 (see Eq. (6)). The next result shows that we can estimate the length of the projection of h_0 onto the span of these features.

Proposition 2. *For $\alpha_i = \mathbb{E}[Y\psi_{\star,i}(Z)]\sigma_{\star,i}^{-1} \in \mathbb{R}^d$, it holds that* $\|\Pi_{\varphi_\star^{(d)}} h_0\|_{L_2(X)}^2 = \alpha^\top \left(I_d - \omega_\star \omega_\star^\top\right)^{-1} \alpha.$

Now, we can construct an estimator of this projection length that utilises a fitted $\widehat{\mathcal{T}}_\delta$ instead of the true operator. Having learned the operator, we can approximate its SVD using a (Z, X) dataset. We do so by decomposing the operator into features that are orthonormal with respect to the $L^2(\widehat{\mu}_X) \times \mathbb{R}$ and $L^2(\widehat{\mu}_Z)$ inner products, were $\widehat{\mu}$ denote empirical distributions based on the samples. This procedure yields $\widehat{\mathcal{T}}_\delta = \sum_{i=1}^d \widehat{\sigma}_i \widehat{\psi}_i \otimes (\widehat{\varphi}_i, \widehat{\omega}_i)$ where $(\widehat{\psi}_i)_{i=1}^d$ and $(\widehat{\varphi}_i)_{i=1}^d$ are orthonormal systems with respect to the aforementioned empirical inner products. For the plug-in estimator $\|\Pi_{\varphi_\star^{(d)}} h_0\|_{L_2(X)}^2 \approx \widehat{\alpha}^\top \left(I_d - \widehat{\omega}\widehat{\omega}^\top\right)^{-1} \widehat{\alpha}$, which shows how well our features approximate the true underlying function and can be used to select δ, in Figure 4, we observe that the estimated spectral alignment increases very rapidly with small δ and stays roughly constant for all the choices of δ that lead to good results in IV regression. The exception is $\delta = 1.0$ where our IV performance has already somewhat degraded but the estimated alignment remains high.

Minimisation of second stage loss. A natural approach to selecting an optimal IV model is to pick one which yields the smallest 2SLS error (Xu et al., 2021; Chen et al., 2022). We find that it works relatively well in experiments on dSprites and off-policy evaluation (where it is the default model selection strategy employed in standard IV-based benchmarks). However, as we discuss in Appendix D.8, the theoretical justification of this method's consistency remains elusive.

Recommended procedure. In light of these results, we believe that the first approach should be more robust. Both are informative about very poor choices of δ but the method based on balancing the loss terms provided more conclusive information about $\delta = 1.0$ being beyond the "optimal" regime. Further investigation on selection of this parameter is an important topic for future work.

7 Conclusion

We have proposed ***Augmented Spectral Feature Learning***, a new framework for outcome-aware feature learning in nonparametric instrumental variable regression. By introducing an augmented operator and a contrastive loss, our method addresses the fundamental limitation of outcome-agnostic spectral features, and achieves robustness in regimes with spectral misalignment to the targeted structural function. Our current approach relies on a rank-one augmentation; extending the framework to richer, higher-rank perturbations remains a promising direction for future research.

REFERENCES

Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. In ***IEEE Transactions on Information Theory***, volume 48-3, pp. 569–579, 2002. doi: 10.1109/TIT.2002.998035.

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. ***IEEE Transactions on Systems, Man, and Cybernetics***, SMC-13(5):834–846, 1983. doi: 10.1109/TSMC.1983.6313077.

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***, pp. 2291–2318. PMLR, 2023.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. ***Econometrica***, 75(6):1613–1669, 2007.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. ***Machine Learning***, 22(1–3):33–57, 1996. ISSN 1573-0565. doi: 10.1007/bf00114723. URL http://dx.doi.org/10.1007/BF00114723.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. ***Foundations of Computational Mathematics***, 7(3):331–368, 2007.

David Card. Using geographic variation in college proximity to estimate the return to schooling, 1993.

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.

Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. ***Econometrica***, 80(1):277–321, 2012.

Yutian Chen, Liyuan Xu, Caglar Gulcehre, Tom Le Paine, Arthur Gretton, Nando de Freitas, and Arnaud Doucet. On instrumental variable regression for deep offline policy evaluation. ***Journal of Machine Learning Research***, 23(302):1–40, 2022.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. ***Econometrica***, 79(5):1541–1565, 2011.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. ***Advances in Neural Information Processing Systems***, 33, 2020.

Jean-Pierre Florens, Jan Johannes, and Sébastien Van Bellegem. Identification and estimation by penalization in nonparametric instrumental regression. ***Econometric Theory***, 27(3):472–496, 2011.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li wei H. Lehman, Matthieu Komorowski, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings, 2018. URL https://arxiv.org/abs/1805.12298.

Peter Hall and Joel L Horowitz. Nonparametric methods for inference in the presence of instrumental variables. ***Annals of Statistics***, 33(6):2904–2929, 2005.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In ***International Conference on Machine Learning***, pp. 1414–1423. PMLR, 2017.

Yang Hu, Tianyi Chen, Na Li, Kai Wang, and Bo Dai. Primal-dual spectral representation for off-policy evaluation, 2024. URL https://arxiv.org/abs/2410.17538.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei (eds.), ***Proceedings of the 32nd International Conference on Machine Learning***, volume 37 of ***Proceedings of Machine Learning Research***, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. ***arXiv preprint arXiv:2510.04088***, 2025.

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.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020. URL https://arxiv.org/abs/2005.01643.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. ***arXiv preprint arXiv:1803.07164***, 2018.

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.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Dimitri Meunier, Zhu Li, Tim Christensen, and Arthur Gretton. Nonparametric instrumental regression via kernel methods is minimax optimal. ***arXiv preprint arXiv:2411.19653***, 2024.

Dimitri Meunier, Antoine Moulin, Jakub Wornbard, Vladimir R. Kostic, and Arthur Gretton. Demystifying spectral feature learning for instrumental variable regression, 2025. URL https://arxiv.org/abs/2506.10899.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In ***International Conference on Learning Representations***, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. ***Nature***, 518(7540):529–533, February 2015. ISSN 1476-4687. doi: 10.1038/nature14236. URL http://dx.doi.org/10.1038/nature14236.

Andrew William Moore. ***Efficient Memory-Based Learning for Robot Control***. Phd thesis, University of Cambridge, 1990.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. ***Econometrica***, 71(5):1565–1578, 2003.

Ieva Petrulionyte, Julien Mairal, and Michael Arbel. Functional bilevel optimization for machine learning. ***Advances in Neural Information Processing Systems***, 37, 2024.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. ***Advances in Neural Information Processing Systems***, 32, 2019.

Haotian Sun, Antoine Moulin, Tongzheng Ren, Arthur Gretton, and Bo Dai. Spectral representation for causal estimation with hidden confounders. In ***The 28th International Conference on Artificial Intelligence and Statistics***, pp. 2719–2727. PMLR, 2025.

Joel A Tropp. An introduction to matrix concentration inequalities. ***Foundations and Trends® in Machine Learning***, 8(1-2):1–230, 2015.

Giacomo Turri, Luigi Bonati, Kai Zhu, Massimiliano Pontil, and Pietro Novelli. Self-supervised evolution operator learning for high-dimensional dynamical systems. ***arXiv preprint arXiv:2505.18671***, 2025.

Ziyu Wang, Yuhao Zhou, and Jun Zhu. Fast instrument learning with faster rates. ***Advances in Neural Information Processing Systems***, 35:16596–16611, 2022.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. ***BIT Numerical Mathematics***, 12(1):99–111, 1972.

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***, 2021.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation, 2024. URL https://arxiv.org/abs/2106.03907.

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.

A TECHNICAL TOOLS

Notation. Throughout the appendix, for a compact operator A, $\sigma_d(A)$ denotes the d-th largest singular value of A and if A is self-adjoint, $\lambda_d(A)$ denotes the d-th largest eigenvalue of A.

Proposition 3 (Weyl's inequality)**.** *For any two compact operators A and B, the singular values are stable under perturbation: $|\sigma_i(A) - \sigma_i(B)| \leq \|A - B\|$, $i \leq \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$.*

Theorem 2 (Wedin sin-Θ Theorem)**.** *Let A and B be compact operators. Let P_A and P_B be the orthogonal projections onto the subspaces spanned by the top-d left singular vectors of A and B respectively (the same property holds with the right singular vectors). If $\gamma := \sigma_d(A) - \sigma_{d+1}(B) > 0$, then,*

$$\|P_A - P_B\| \leq \frac{\|A - B\|}{\gamma}.$$

Additionally, if $\gamma_A := \sigma_d(A) - \sigma_{d+1}(A) > 0$ and $\|A - B\| \leq \gamma_A/2$, then

$$\|P_A - P_B\| \leq 2 \cdot \frac{\|A - B\|}{\gamma_A}.$$

Proof. The first inequality is the original theorem (Wedin, 1972). We prove the second inequality. By Weyl's inequality, Proposition 3,

$$|\sigma_{d+1}(A) - \sigma_{d+1}(B)| \leq \|A - B\|.$$

Hence,

$$\gamma \geq \sigma_d(A) - \sigma_{d+1}(A) - \|A - B\| = \gamma_A - \|A - B\| \geq \gamma_A/2.$$

Therefore,

$$\|P_A - P_B\| \leq 2 \cdot \frac{\|A - B\|}{\gamma_A}.$$

□

Theorem 3 (Eckart-Young-Mirsky Theorem)**.** *Let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact operator between Hilbert spaces with Singular Value Decomposition*

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

where $\{v_i\} \subset \mathcal{H}_1$ and $\{u_i\} \subset \mathcal{H}_2$ are orthonormal sets, and $(\sigma_i)_i$ are the singular values of A, which satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. For all $d \geq 1$, $A_d \doteq \sum_{i=1}^{d} \sigma_i u_i \otimes v_i$ satisfies

$$\|A - A_d\| = \min_{B:\mathrm{rank}(B) \leq d} \|A - B\|.$$

Moreover, if $\sigma_d > \sigma_{d+1}$, A_d is the unique minimizer.

B OMITTED PROOFS

We provide the proofs omitted from the main text below.

B.1 PROOF OF PROPOSITION 1

Given $d \geq 1$, $\delta \geq 0$, parameter θ and $\omega \in \mathbb{R}^d$, define the operator

$$\mathcal{T}_{\theta,\omega} = \Psi_\theta^{(d)}[\Phi_\theta^{(d)\star} \mid \omega] = \sum_{i=1}^{d} \psi_{\theta,i}^{(d)} \otimes (\varphi_{\theta,i}^{(d)}, \omega_i) : L_2(X) \times \mathbb{R} \to L_2(Z).$$

We show that

$$\mathcal{L}_\delta(\theta, \omega) = \|\mathcal{T}_\delta - \mathcal{T}_{\theta,\omega}\|_{\mathrm{HS}}^2 - \|\mathcal{T}_\delta\|_{\mathrm{HS}}^2 .$$

To simplify the notations in this proof, we drop the (d) subscript on the features. Let us define $\tilde{\Phi}_{\theta,\omega} : \mathbb{R}^d \to L_2(X) \times \mathbb{R}$ the operator whose adjoint is $\tilde{\Phi}^*_{\theta,\omega} = [\Phi^*_\theta \mid \omega] : L_2(X) \times \mathbb{R} \to \mathbb{R}^d$ such that we have $\mathcal{T}_{\theta,\omega} = \Psi_\theta \tilde{\Phi}^*_{\theta,\omega}$. $\tilde{\Phi}_{\theta,\omega}$ is such that

$$\tilde{\Phi}_{\theta,\omega}\beta = \sum_{i=1}^{d} \beta_i(\varphi_{\theta,i}, \omega_i), \qquad \tilde{\Phi}^*_{\theta,\omega}(h, a) = \Phi^*_\theta h + a \cdot \omega.$$

Therefore,

$$\tilde{\Phi}^\star_{\theta,\omega}\tilde{\Phi}_{\theta,\omega} = \mathbb{E}[\varphi_\theta(X)\varphi_\theta(X)^\intercal] + \omega\omega^\intercal = C_{\varphi_\theta} + \omega\omega^\intercal \tag{9}$$

Using the definition of the Hilbert-Schmidt norm and the cyclic property of the trace, we have

$$\|\mathcal{T}_\delta - \mathcal{T}_{\theta,\omega}\|^2_{\text{HS}} - \|\mathcal{T}_\delta\|^2_{\text{HS}} = \|\mathcal{T}_{\theta,\omega}\|^2_{\text{HS}} - 2\,\text{Tr}\left(\mathcal{T}^\star_{\theta,\omega}\mathcal{T}_\delta\right) = \text{Tr}\left(\Psi^*_\theta\Psi_\theta\tilde{\Phi}^*_{\theta,\omega}\tilde{\Phi}_{\theta,\omega}\right) - 2\,\text{Tr}\left(\Psi^*_\theta\mathcal{T}_\delta\tilde{\Phi}_{\theta,\omega}\right).$$

For the first term, exploiting Eq. (9) and the linearity of the trace, we have

$$\begin{aligned}
\text{Tr}\left(\Psi^*_\theta\Psi_\theta\tilde{\Phi}^*_{\theta,\omega}\tilde{\Phi}_{\theta,\omega}\right) &= \text{Tr}\left(C_{\psi_\theta}C_{\varphi_\theta} + C_{\psi_\theta}\omega\omega^\intercal\right) \\
&= \text{Tr}(C_{\psi_\theta}C_{\varphi_\theta}) + \omega^\intercal C_{\psi_\theta}\omega \\
&= \mathbb{E}_X\mathbb{E}_Z[(\varphi_\theta(X)^\intercal\psi_\theta(Z))^2] + \omega^\intercal C_{\psi_\theta}\omega,
\end{aligned}$$

where $\mathbb{E}_X\mathbb{E}_Z$ denotes the expectation where X and Z are treated as independent random variables drawn from their respective marginal distributions.

For the second term, recall that $\mathcal{T}_\delta = [\mathcal{T} \mid \delta r_0]$. For all $\beta \in \mathbb{R}^d$, we have

$$\begin{aligned}
[\Psi^*_\theta\mathcal{T}_\delta\tilde{\Phi}_{\theta,\omega}\beta]_j &= \sum_{i=1}^{d} \beta_i[\Psi^*_\theta\mathcal{T}_\delta(\varphi_{\theta,i}, \omega_i)]_j \\
&= \sum_{i=1}^{d} \beta_i[\Psi^*_\theta\left(\omega_i \cdot \delta \cdot r_0 + \mathcal{T}\varphi_{\theta,i}\right)]_j \\
&= \sum_{i=1}^{d} \beta_i\left(\omega_i \cdot \delta \cdot \mathbb{E}[Y\psi_{\theta,j}(Z)] + \mathbb{E}[\varphi_{\theta,i}(X)\psi_{\theta,j}(Z)]\right).
\end{aligned}$$

Therefore $\Psi^*_\theta\mathcal{T}_\delta\tilde{\Phi}_{\theta,\omega}\beta = \delta \cdot \mathbb{E}[Y\psi_\theta(Z)]\omega^\intercal + \mathbb{E}[\psi_\theta(Z)\varphi_\theta(X)^\intercal]$, and

$$\text{Tr}\left(\Psi^*_\theta\mathcal{T}_\delta\tilde{\Phi}_{\theta,\omega}\right) = \mathbb{E}[\varphi_\theta(X)^\intercal\psi_\theta(Z)] + \delta \cdot \omega^\intercal\mathbb{E}[Y\psi_\theta(Z)].$$

Putting it together, we obtain

$$\begin{aligned}
\|\mathcal{T}_\delta - \mathcal{T}_{\theta,\omega}\|^2_{\text{HS}} - \|\mathcal{T}_\delta\|^2_{\text{HS}} &= \mathbb{E}_X\mathbb{E}_Z[(\varphi_\theta(X)^\intercal\psi_\theta(Z))^2] + \omega^\intercal C_{\psi_\theta}\omega - 2\mathbb{E}[\varphi_\theta(X)^\intercal\psi_\theta(Z)] - 2\delta \cdot \omega^\intercal\mathbb{E}[Y\psi_\theta(Z)] \\
&= \mathcal{L}_0(\theta) + \omega^\intercal C_{\psi_\theta}\omega - 2\delta \cdot \omega^\intercal\mathbb{E}[Y\psi_\theta(Z)] \\
&= \mathcal{L}_\delta(\theta, \omega).
\end{aligned}$$

We conclude the proof with Theorem 3.

B.2 Proof of Proposition 2

Recall the SVD of $\mathcal{T}_\delta$ (Eq. (6))

$$\mathcal{T}_\delta = \sum_{i \geq 1} \sigma_{\star,i}\psi_{\star,i} \otimes (\varphi_{\star,i}, \omega_{\star,i}).$$

The key relationships it satisfies are $\sigma_{\star,i}\varphi_{\star,i} = \mathcal{T}^*\psi_{\star,i}$, and $\sigma_{\star,i}\omega_{\star,i} = \delta\langle r_0, \psi_{\star,i}\rangle = \delta \cdot \sigma_{\star,i}\langle h_0, \varphi_{\star,i}\rangle$. In Eq. (14), we show that $\Phi^{(d)\star}_\star\Phi^{(d)}_\star = I_d - \omega_\star\omega_\star^\intercal$. Thus

$$\Pi_{\varphi^{(d)}_\star} = \Phi^{(d)}_\star(\Phi^{(d)\star}_\star\Phi^{(d)}_\star)^{-1}\Phi^{(d)\star}_\star = \Phi^{(d)}_\star(I_d - \omega_\star\omega_\star^\intercal)\Phi^{(d)\star}_\star.$$

Next, observe that

$$\langle\varphi_{\star,i}, h_0\rangle = \sigma^{-1}_{\star,i}\langle\mathcal{T}^\star\psi_i, h_0\rangle = \sigma_i^{-1}\langle\psi_{\star,i}, \mathcal{T}h_0\rangle = \sigma^{-1}_{\star,i}\mathbb{E}[Y\psi_{\star,i}(Z)] = \alpha_i.$$

Therefore, α is such that $\alpha = \Phi^{(d)}_\star h_0$. Alignment of h_0 to the true spectral features (of $\mathcal{T}_\delta$) then satisfies:

$$\|\Pi_{\varphi^{(d)}_\star}h_0\|^2 = \langle h_0, \Pi_{\varphi^{(d)}_\star}h_0\rangle = (\Phi^{(d)\star}_\star h_0)^\intercal\left(I_d - \omega_\star\omega_\star^\intercal\right)^{-1}\Phi^{(d)\star}_\star h_0 = \alpha^\intercal\left(I_d - \omega_\star\omega_\star^\intercal\right)^{-1}\alpha.$$

B.3 Approximation Error Analysis

Recall that for $d \geq 1$, we fixed a partition $\overline{N} \dot{\cup} \underline{N} = \mathbb{N}$, such that $|\overline{N}|{=}d$, $\mathcal{T}{=}\overline{U}_d\overline{\Lambda}_d\overline{V}_d^*{+}\underline{U}_d\underline{\Lambda}_d\underline{V}_d^*$. We introduce the projection operators

$$\Pi_{\overline{U}_d} = \overline{U}_d\overline{U}_d^* \qquad \Pi_{\overline{V}_d} = \overline{V}_d\overline{V}_d^* \qquad \Pi_{\underline{U}_d} = \underline{U}_d\underline{U}_d^* = I - \Pi_{\overline{U}_d} \qquad \Pi_{\underline{V}_d} = \underline{V}_d\underline{V}_d^* = I - \Pi_{\overline{V}_d}.$$

We then decompose h_0 as

$$h_0 = \overline{h}_0 + \underline{h}_0 = \Pi_{\overline{V}_d}h_0 + \Pi_{\underline{V}_d}h_0 = \sum_{i\in\overline{N}} \overline{\alpha}_i v_i + \sum_{i\in\underline{N}} \underline{\alpha}_i v_i.$$

Let us define $\overline{\lambda}_{\min}$ as the smallest positive entries of $\overline{\Lambda}_d$.

Proposition 4. *$\overline{\lambda}_{\min}$ is such that*

$$\sup_{h\neq 0} \frac{\|\Pi_{\overline{V}_d}h\|_{L_2(X)}}{\|\mathcal{T}\,\Pi_{\overline{V}_d}h\|_{L_2(Z)}} = \overline{\lambda}_{\min}^{-1}.$$

Proof. Since the quantity is homogeneous in h, we may restrict the supremum to

$$\|\Pi_{\overline{V}_d}h\|_{L_2(X)} = 1.$$

Write $\Pi_{\overline{V}_d}h = \sum_{i\in\overline{N}} \alpha_i v_i$ so that $\sum_{i\in\overline{N}} |\alpha_i|^2 = 1$. Using $\mathcal{T}v_i = \lambda_i u_i$ for $i \in \overline{N}$ we get

$$\|\mathcal{T}\,\Pi_{\overline{V}_d}h\|_{L_2(Z)}^2 = \sum_{i\in\overline{N}} \lambda_i^2|\alpha_i|^2.$$

Hence, for such h,

$$\frac{\|\Pi_{\overline{V}_d}h\|_{L_2(X)}}{\|\mathcal{T}\,\Pi_{\overline{V}_d}h\|_{L_2(Z)}} = \frac{1}{\left(\sum_{i\in\overline{N}} \lambda_i^2|\alpha_i|^2\right)^{1/2}} \leq \frac{1}{\overline{\lambda}_{\min}},$$

Equality is attained by taking h proportional to the singular vector v_{i^*} with $\lambda_{i^*} = \overline{\lambda}_{\min}$, which conclude the proof. □

The following proposition states a general approximation error bound when the span of the learned features is close to the singular spaces of $\mathcal{T}$ associated to $\overline{N}$.

Proposition 5. *Let A, B be constants such that $\|\Pi_{\overline{U}_d} - \Pi_{\psi_\theta}\| \leq A$, $\|\Pi_{\overline{V}_d} - \Pi_{\varphi_\theta}\| \leq B$, and*

$$1 - B - \frac{A}{\overline{\lambda}_{\min}} > 0.$$

Then,

$$\|h_\theta - h_0\|_{L_2(X)} \leq \left(1 - B - \frac{A}{\overline{\lambda}_{\min}}\right)^{-1} \left(B\|h_0\|_{L_2(X)} + \|\Pi_{\underline{V}_d}h_0\|_{L_2(X)}\right).$$

Proof. Recall that under Assumption 3, $h_\theta(x) = \varphi_\theta(x)^\intercal\beta_\theta$ satisfies $C_{ZX,\theta}\beta_\theta = \mathbb{E}[r_0(Z)\psi_\theta(Z)]$. Observing that $\mathbb{E}[r_0(Z)\psi_\theta(Z)] = \Psi_\theta^* r_0 = \Psi_\theta^*\mathcal{T}h_0$ and using that $C_{ZX,\theta} = \Psi_\theta^*\mathcal{T}\Phi_\theta$, we have $C_{ZX,\theta}\beta_\theta = \Psi_\theta^*\mathcal{T}\,h_\theta$. Therefore $0 = \Psi_\theta^*\mathcal{T}(h_\theta - h_0)$, which implies

$$\Pi_{\psi_\theta}\mathcal{T}\,(h_\theta - h_0) = 0 \tag{10}$$

We then have the following chain of inequalities,

$$\begin{aligned}
\|h_\theta - h_0\|_{L_2(X)} &\leq \|\Pi_{\overline{V}_d}\,(h_\theta - h_0)\,\|_{L_2(X)} + \|\Pi_{\underline{V}_d}\,(h_\theta - h_0)\,\|_{L_2(X)} \\
&\leq \overline{\lambda}_{\min}^{-1}\|\mathcal{T}\,\Pi_{\overline{V}_d}\,(h_\theta - h_0)\,\|_{L_2(X)} + \|\Pi_{\underline{V}_d}h_\theta\|_{L_2(X)} + \|\Pi_{\underline{V}_d}h_0\|_{L_2(X)} \\
&= \overline{\lambda}_{\min}^{-1}\|\Pi_{\overline{U}_d}\mathcal{T}\,(h_\theta - h_0)\,\|_{L_2(X)} + \|\left(\Pi_{\underline{V}_d} - \Pi_{\varphi_\theta}^{\perp}\right)h_\theta\|_{L_2(X)} + \|\Pi_{\underline{V}_d}h_0\|_{L_2(X)} \\
&\leq \overline{\lambda}_{\min}^{-1}\|\Pi_{\overline{U}_d} - \Pi_{\psi_\theta}\|\|\mathcal{T}\|\|h_\theta - h_0\|_{L_2(X)} + \|\Pi_{\overline{V}_d} - \Pi_{\varphi_\theta}\|\|h_\theta\|_{L_2(X)} + \|\Pi_{\underline{V}_d}h_0\|_{L_2(X)} \\
&\leq \frac{A}{\overline{\lambda}_{\min}}\|h_\theta - h_0\|_{L_2(X)} + B\|h_\theta - h_0 + h_0\|_{L_2(X)} + \|\Pi_{\underline{V}_d}h_0\|_{L_2(X)} \\
&\leq \left(\frac{A}{\overline{\lambda}_{\min}} + B\right)\|h_\theta - h_0\|_{L_2(X)} + B\|h_0\|_{L_2(X)} + \|\Pi_{\underline{V}_d}h_0\|_{L_2(X)},
\end{aligned}$$

where in the second inequality we used Proposition 4 and the triangular inequality, in the equality we used $\mathcal{T}\,\Pi_{\overline{V}_d} = \Pi_{\overline{U}_d}\mathcal{T}$ and $\Pi^{\perp}_{\varphi_\theta} h_\theta = 0$, in the third inequality we used Eq. (10) and in the fourth inequality we used $\|\mathcal{T}\| \leq 1$. Re-arranging, we obtain

$$\left(1 - \frac{A}{\overline{\lambda}_{\min}} + B\right) \|h_\theta - h_0\|_{L_2(X)} \leq B\|h_0\|_{L_2(X)} + \|\Pi_{\underline{V}_d} h_0\|_{L_2(X)},$$

and the result follows. □

Then next step is to obtain A and B. Recall the SVD of $\mathcal{T}_\delta$ Eq. (6)

$$\mathcal{T}_\delta = \sum_{i\geq 1} \sigma_{*,i}\psi_{*,i} \otimes (\varphi_{*,i}, \omega_{*,i}), \quad \psi_{*,i} \in L_2(Z), \quad (\varphi_{*,i}, \omega_{*,i}) \in L_2(X) \times \mathbb{R},$$

We denote by $\Pi_{\psi_*^{(d)}}$ and $\Pi_{\varphi_*^{(d)}}$ and orthogonal projection onto the span of $(\psi_{*,i})_{i=1}^d$ and $(\varphi_{*,i})_{i=1}^d$ respectively. It is important to note that $\varphi_{*,i}$ is not a singular function of $\mathcal{T}_\delta$ but the first component of the singular function $(\varphi_{*,i}, \omega_{*,i})$. Therefore the family $\{\varphi_{*,i}\}_i$ is not orthonormal. From the SVD, we deduce the following relationships for all $i \geq 1$

$$\begin{cases} \sigma_{\star,i}\psi_{\star,i} = \mathcal{T}_\delta(\varphi_{\star,i}, \omega_{\star,i}), \\ \sigma_{\star,i}\varphi_{\star,i} = \mathcal{T}^*\psi_{\star,i}, \\ \sigma_{\star,i}\omega_{\star,i} = \delta\,\langle \psi_{\star,i}, r_0\rangle_{L_2(Z)}, \end{cases} \tag{11}$$

where we use the fact that $\mathcal{T}_\delta^\star r = \mathcal{T}^\star r + \delta\,\langle r, r_0\rangle_{L_2(Z)}$, $r \in L_2(Z)$.

By the triangular inequality,

$$\|\Pi_{\overline{U}_d} - \Pi_{\psi_\theta}\| \leq \|\Pi_{\overline{U}_d} - \Pi_{\psi_*^{(d)}}\| + \|\Pi_{\psi_*^{(d)}} - \Pi_{\psi_\theta}\| \qquad (Z - \text{features})$$
$$\|\Pi_{\overline{V}_d} - \Pi_{\varphi_\theta}\| \leq \|\Pi_{\overline{V}_d} - \Pi_{\varphi_*^{(d)}}\| + \|\Pi_{\varphi_*^{(d)}} - \Pi_{\varphi_\theta}\| \qquad (X - \text{features})$$

Proposition 6 (Control of the X-features)**.**

$$\|\Pi_{\varphi_*^{(d)}} - \Pi_{\overline{V}_d}\| \leq \frac{\|\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d}\|}{\overline{\lambda}_{\min}(1 - \|\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d}\|)}. \tag{12}$$

Assume in addition that $\overline{\lambda}_{\min} - \|\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d}\| > 0$*. Then,*

$$\|\Pi_{\varphi_*^{(d)}} - \Pi_{\varphi_\theta}\| \leq \frac{\|\mathcal{T}_\delta^{(d)} - \mathcal{T}_{\theta,\omega}\|}{\overline{\lambda}_{\min} - \|\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d}\|}. \tag{13}$$

Proof. We start with Eq. (13). Let $\pi : L_2(X) \times \mathbb{R} \to L_2(X)$ be the canonical projection such that $\pi(h, a) = h$. Recall that

$$\Phi_\star^{(d)}\beta = \sum_{i=1}^d \beta_i \varphi_{*,i}.$$

Let us define

$$\tilde{\Phi}_\star^{(d)}\beta = \sum_{i=1}^d \beta_i (\varphi_{*,i}, \omega_{*,i}),$$

that is such that $\pi\tilde{\Phi}_\star^{(d)} = \Phi_\star^{(d)}$. We then have the following decomposition

$$\begin{aligned} \Phi_\star^{(d)}\Sigma_\star^{(d)}\Psi_\star^{(d)\star} &= \pi(\tilde{\Phi}_\star^{(d)}\Sigma_\star^{(d)}\Psi_\star^{(d)\star}) = \pi(\mathcal{T}_\delta^{(d)})^* = \pi(\mathcal{T}_\delta^{(d)} - \mathcal{T}_{\theta,\omega})^* + \pi\mathcal{T}_{\theta,\omega}^* \\ &= \pi(\mathcal{T}_\delta^{(d)} - \mathcal{T}_{\theta,\omega})^* + \Phi_\theta\Psi_\theta^* \end{aligned}$$

We will apply Wedin Sin-Θ Theorem, Theorem 2, to $A = \Phi_\star^{(d)}\Sigma_\star^{(d)}\Psi_\star^{(d)\star}$ and $B = \Phi_\theta\Psi_\theta^*$. Note that

$$\begin{aligned} \|A - B\| &= \|\Phi_\star^{(d)}\Sigma_\star^{(d)}\Psi_\star^{(d)\star} - \Phi_\theta\Psi_\theta^*\| \\ &= \|\pi(\mathcal{T}_\delta^{(d)})^*(\Pi_{\psi_*^{(d)}} - \Pi_{\psi_\theta}) + \pi(\mathcal{T}_\delta^{(d)} - \mathcal{T}_{\theta,\omega})^*\Pi_{\psi_\theta}\| \\ &\leq \sigma_1(\mathcal{T}_\delta^{(d)})\|\Pi_{\psi_*^{(d)}} - \Pi_{\psi_\theta}\| + \|\mathcal{T}_\delta^{(d)} - \mathcal{T}_{\theta,\omega}\| \end{aligned}$$

As $(\varphi_{*,i}, \omega_{*,i})_{i=1}^d$ forms an orthonormal family, we have $\tilde{\Phi}_\star^{(d)\star}\tilde{\Phi}_\star^{(d)} = I_d$. On the other hand, similarly to Eq. (9), we have

$$I_d = \tilde{\Phi}_\star^{(d)\star}\tilde{\Phi}_\star^{(d)} = \Phi_\star^{(d)\star}\Phi_\star^{(d)} + \omega_*\omega_*^\top \tag{14}$$

Therefore,

$$\begin{aligned}
\sigma_d(A)^2 = \sigma_d(\Phi_\star^{(d)}\Sigma_\star^{(d)}\Psi_\star^{(d)\star})^2 &= \sigma_d(\Psi_\star^{(d)}\Sigma_\star^{(d)}\Phi_\star^{(d)\star}\Phi_\star^{(d)}\Sigma_\star^{(d)}\Psi_\star^{(d)\star}) \\
&= \sigma_d(\Psi_\star^{(d)}\Sigma_\star^{(d)}(I_d - \omega_\star\omega_\star^\top)\Sigma_\star^{(d)}\Psi_\star^{(d)\star}) \\
&= \sigma_d(\mathcal{T}_\delta^{(d)}(\mathcal{T}_\delta^{(d)})^* - \delta^2(\Psi_\star^{(d)}\Psi_\star^{(d)\star}r_0)\otimes(\Psi_\star^{(d)}\Psi_\star^{(d)\star}r_0)) \\
&= \sigma_d(\Pi_{\psi_*^{(d)}}(\mathcal{T}_\delta(\mathcal{T}_\delta)^* - \delta^2 r_0\otimes r_0)\Pi_{\psi_*^{(d)}}) \\
&= \sigma_d(\Pi_{\psi_*^{(d)}}\mathcal{T}\mathcal{T}^*\Pi_{\psi_*^{(d)}}) \\
&= \sigma_d(\Pi_{\psi_*^{(d)}}\mathcal{T})^2,
\end{aligned}$$

where in the third equality, we used that fact that $\omega_\star = \delta(\Sigma_\star^{(d)})^{-1}\Psi_\star^{(d)\star}r_0$ by Eq. (11) and in the fifth equality we used $\mathcal{T}_\delta(\mathcal{T}_\delta)^* = \mathcal{T}\mathcal{T}^* + \delta^2 r_0\otimes r_0$. Next, by Weyl's inequality, Proposition 3,

$$|\sigma_d(\Pi_{\psi_*^{(d)}}\mathcal{T}) - \overline{\lambda}_{\min}| = |\sigma_d(\Pi_{\psi_*^{(d)}}\mathcal{T}) - \sigma_d(\Pi_{\overline{U}_d}\mathcal{T})| \le \|(\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d})\mathcal{T}\| \le \|\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d}\|.$$

Hence $\sigma_d(A) \ge \overline{\lambda}_{\min} - \|\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d}\|$ and $\sigma_{d+1}(B) = 0$. We obtain Eq. (13) applying Theorem 2.

We now prove Eq. (12). We use

$$\Phi_\star^{(d)}\Sigma_\star^{(d)} = \mathcal{T}^*\Psi_\star^{(d)} = \overline{V}_d\,\overline{\Lambda}_d\,\overline{U}_d^*\Psi_\star^{(d)} + \underline{V}_d\,\underline{\Lambda}_d\,\underline{U}_d^*\Psi_\star^{(d)},$$

We will apply Wedin Sin-Θ Theorem, Theorem 2, to $A' = \overline{V}_d\,\overline{\Lambda}_d\,\overline{U}_d^*\Psi_\star^{(d)}$ and $B' = \Phi_\star^{(d)}\Sigma_\star^{(d)}$. Note that

$$\|A' - B'\| = \|\overline{V}_d\,\overline{\Lambda}_d\,\overline{U}_d^*\Psi_\star^{(d)} - \Phi_\star^{(d)}\Sigma_\star^{(d)}\| = \|\underline{V}_d\,\underline{\Lambda}_d\,\underline{U}_d^*\Psi_\star^{(d)}\| \le \|\Pi_{\overline{U}_d} - \Pi_{\psi_*^{(d)}}\|,$$

where for the last inequality, we use

$$\underline{U}_d^*\Psi_\star^{(d)} = \underline{U}_d^*\Pi_{\underline{U}_d}\Psi_\star^{(d)} = \underline{U}_d^*\Pi_{\overline{U}_d}^\perp\Psi_\star^{(d)} = \underline{U}_d^*[\Pi_{\overline{U}_d}^\perp - \Pi_{\psi_*^{(d)}}^\perp]\Psi_\star^{(d)} = \underline{U}_d^*[\Pi_{\overline{U}_d} - \Pi_{\psi_*^{(d)}}]\Psi_\star^{(d)}.$$

By the inequality for singular values of product of matrices, we have

$$\begin{aligned}
\sigma_d(A') = \sigma_d(\overline{V}_d\,\overline{\Lambda}_d\,\overline{U}_d^*\Psi_\star^{(d)}) &\ge \sigma_d(\overline{V}_d\,\overline{\Lambda}_d)\,\sigma_d(\overline{U}_d^*\Psi_\star^{(d)}) \\
&= \sigma_d(\overline{\Lambda}_d)\,\sigma_d(\overline{U}_d^*\Psi_\star^{(d)}) \\
&= \overline{\lambda}_{\min}\,\sigma_d(\overline{U}_d^*\Psi_\star^{(d)}) \\
&\ge \overline{\lambda}_{\min}(1 - \|\Pi_{\psi_*^{(d)}} - \Pi_{\overline{U}_d}\|).
\end{aligned}$$

We conclude with Theorem 2, using that $\sigma_{d+1}(B') = 0$. □

Proposition 7 (Control of the Z-features)**.** *Assume that $\varepsilon_d(\theta,\omega,\delta) \le \sigma_d(\mathcal{T}_\delta)/2$. Then,*

$$\|\Pi_{\psi_*^{(d)}} - \Pi_{\psi_\theta}\| \le 2\cdot\frac{\varepsilon_d(\theta,\omega,\delta)}{\sigma_d(\mathcal{T}_\delta)}$$

Assume that $\gamma_d(\delta) := \|[\overline{\Lambda}_d(I + \delta^2\overline{\alpha}\overline{\alpha}^\top)^{1/2}]^{-1}\|^{-1} - \|\underline{\Lambda}_d\| > 0$ and $\|\underline{\Lambda}_d\|\|\underline{\alpha}_0\|_{\ell_2} \le \gamma_d(\delta)/2$. Then,

$$\|\Pi_{\overline{U}_d} - \Pi_{\psi_*^{(d)}}\| \le 2\cdot\frac{\delta\|\underline{\Lambda}_d\|\|\underline{\alpha}_0\|_{\ell_2}}{\gamma_d(\delta)}.$$

Proof. The first inequality follows from Theorem 2 and the definition of $\varepsilon_d(\theta,\omega,\delta)$.

For the second inequality, using $\mathcal{T} = \overline{U}_d\,\overline{\Lambda}_d\,\overline{V}_d^* + \underline{U}_d\,\underline{\Lambda}_d\,\underline{V}_d^*$, we get the following decomposition

$$\mathcal{T}_\delta = [\mathcal{T} \mid \delta r_0] = [\mathcal{T} \mid \delta\mathcal{T}h_0] = \underbrace{[\mathcal{T} \mid \delta\overline{U}_d\,\overline{\Lambda}_d\,\overline{V}_d^*h_0]}_{\tilde{Q}} + [0 \mid \delta\underline{U}_d\,\underline{\Lambda}_d\,\underline{V}_d^*h_0]$$

Note that

$$\|\mathcal{T}_\delta - \tilde{Q}\| = \|\delta \underline{U}_d \, \underline{\Lambda}_d \, \underline{\alpha}_0\|_{L_2(Z)} \leq \delta \|\underline{\Lambda}_d\| \|\underline{\alpha}_0\|_{L_2(X)}$$

To analyze the spectrum of $\tilde{Q}$, we look at

$$\tilde{Q}\tilde{Q}^* = \mathcal{T}\mathcal{T}^* + \delta^2 (\overline{U}_d \, \overline{\Lambda}_d \, \overline{V}_d^* h_0) \otimes (\overline{U}_d \, \overline{\Lambda}_d \, \overline{V}_d^* h_0) = \left[\overline{U}_d \,|\, \underline{U}_d\right] \begin{bmatrix} \overline{M} & 0 \\ 0 & \underline{\Lambda}_d^2 \end{bmatrix} \left[\overline{U}_d \,|\, \underline{U}_d\right]^*,$$

where $\overline{M} = \overline{\Lambda}_d^2 + \delta^2 \overline{\Lambda}_d \overline{V}_d^* (h_0 \otimes h_0) \overline{V}_d \overline{\Lambda}_d$. When $\lambda_{\min}(\overline{M}) > \|\underline{\Lambda}_d^2\|$ then $\Pi_{\overline{U}_d}$ is associated to the top spectrum of $\tilde{Q}$ and we can apply Wedin Sin-Θ Theorem, Theorem 2, to $A = \tilde{Q}$ and $B = \mathcal{T}_\delta$. As

$$\overline{M} = \overline{\Lambda}_d (I + \delta^2 \overline{\alpha\alpha}^\intercal) \overline{\Lambda}_d,$$

we have

$$\lambda_{\min}(\overline{M}) = \sigma_{\min}^2 (\overline{\Lambda}_d (I + \delta^2 \overline{\alpha\alpha}^\intercal)^{1/2}) = \|[\overline{\Lambda}_d (I + \delta^2 \overline{\alpha\alpha}^\intercal)^{1/2}]^{-1}\|^{-2}.$$

Therefore $\lambda_{\min}(\overline{M}) > \|\underline{\Lambda}_d^2\|$ is equivalent to

$$\sigma_d(A) - \sigma_{d+1}(B) = \|[\overline{\Lambda}_d (I + \delta^2 \overline{\alpha\alpha}^\intercal)^{1/2}]^{-1}\|^{-1} - \|\underline{\Lambda}_d\| = \gamma_d(\delta) > 0.$$

□

Proposition 8 (Control of ill-posedness)**.** *Under the assumptions of Proposition 5, it holds that*

$$c_{\varphi_\theta, \psi_\theta} \geq \overline{\lambda}_{\min} \left(1 - B - \frac{A}{\overline{\lambda}_{\min}}\right).$$

Proof. Start by decomposing projections as

$$\Pi_{\Psi_\theta} \mathcal{T} \Pi_{\varphi_\theta} = \Pi_{\overline{U}_d} \mathcal{T} (\Pi_{\varphi_\theta} + \Pi_{\underline{U}_d}) + (\Pi_{\Psi_\theta} - \Pi_{\overline{U}_d}) \mathcal{T} \Pi_{\varphi_\theta}.$$

Now, using inequalities of sums and products of singular values we have that

$$\sigma_d(\Pi_{\Psi_\theta} \mathcal{T} \Pi_{\varphi_\theta}) \geq \sigma_d(\Pi_{\overline{U}_d} \mathcal{T}) \sigma_{\min}(\Pi_{\varphi_\theta} + \Pi_{\underline{U}_d}) - \sigma_1((\Pi_{\Psi_\theta} - \Pi_{\overline{U}_d}) \mathcal{T} \Pi_{\varphi_\theta}).$$

But, since

$$\sigma_{\min}(\Pi_{\varphi_\theta} + \Pi_{\underline{U}_d}) = \|[I - (\Pi_{\overline{U}_d} - \Pi_{\varphi_\theta})]^{-1}\|^{-1} \geq 1 - \|\Pi_{\overline{U}_d} - \Pi_{\varphi_\theta}\|,$$

whenever $\|\Pi_{\overline{U}_d} - \Pi_{\varphi_\theta}\| < 1$, the proof is completed. □

We combine the previous proposition in the following general approximation error bound.

Theorem 4. *Given any $\delta > 0$ and partition $\overline{N} \dot{\cup} \underline{N} = \mathbb{N}$, denoting $\overline{\lambda}_{\min} = \min_{i \in [\overline{N}]} \lambda_i$ and $\underline{\lambda}_{\max} = \max_{i \in [\underline{N}]} \lambda_i$, if*

$$\frac{6 \, \delta \, \underline{\lambda}_{\max}}{\overline{\lambda}_{\min} \, \gamma_d(\delta)} \|q_d\|_{L^2(X)} + \frac{2\lambda_d + \overline{\lambda}_{\min}}{\lambda_d \overline{\lambda}_{\min}} \mathcal{E}_d(\theta, \delta) \leq \kappa \leq 1/2 \tag{15}$$

then $c_{\varphi_\theta, \psi_\theta} \geq (1 - \kappa) \overline{\lambda}_{\min}$, and

$$\|h_0 - h_\theta\|_{L_2(X)} \leq \left(1 + \frac{4\delta \, \underline{\lambda}_{\max} \, \|h_0\|_{L^2(X)}}{\overline{\lambda}_{\min} \, \gamma_d(\delta)}\right) \frac{\|q_d\|_{L^2(X)}}{1 - \kappa} + \frac{(2\lambda_d + \overline{\lambda}_{\min}) \, \|h_0\|_{L^2(X)}}{\overline{\lambda}_{\min} \lambda_d} \frac{\mathcal{E}_d(\theta, \delta)}{1 - \kappa}. \tag{16}$$

Proof. First, observe that $\sigma_d(\mathcal{T}_\delta) \geq \lambda_d$ and $\sigma_1(\mathcal{T}_\delta) \leq (1 + \delta \|r_0\|_{L^2(Z)})$. Now, recalling Proposition 5, due to Proposition 7 we can take

$$A = \frac{2\delta \, \underline{\lambda}_{\max} \|q_d\|_{L^2(X)}}{\gamma_d(\delta)} + \frac{\mathcal{E}_d(\theta, \delta)}{\lambda_d}$$

But, since Eq. (15) ensures that $A / \overline{\lambda}_{\min} < 1/2$, from Proposition 6 we can set

$$B = \frac{4\delta \, \underline{\lambda}_{\max} \|q_d\|_{L^2(X)}}{\overline{\lambda}_{\min} \gamma_d(\delta)} + \mathcal{E}_d(\theta, \delta) \left(\frac{1}{\lambda_d} + \frac{2}{\overline{\lambda}_{\min}}\right)$$

and obtain that $A / \overline{\lambda}_{\min} + B \leq \kappa < 1/2$. To complete the proof we apply Proposition 8. □

Good scenario. Setting $\overline{N} = \{1, \dots, d\}$ and $\delta = 0$ we obtain the control of the approximation error for the SpecIV learning method. That is, equation 16 becomes

$$\|h_0 - h_\theta\|_{L_2(X)} \leq \frac{1}{1-\kappa}\left(\|q_d\|_{L^2(X)} + \frac{3\,\|h_0\|_{L^2(X)}}{\lambda_d}\,\mathcal{E}_d(\theta,\delta)\right), \tag{17}$$

whenever $\mathcal{E}_d(\theta,\delta) \leq \kappa\lambda_d/3$. Here we see that the representation learning error needs to scale as $\mathcal{E}_d(\theta,\delta) \asymp \lambda_d\,\|q_d\|_{L^2(X)}/\|h_0\|_{L^2(X)}$, as $d \to \infty$.

Bad scenario. Let $\overline{N} = \{k\}$ and assume that $\|s_1\|_{L^2(X)} = |\langle h_0, v_k\rangle| > \|h_0 - \alpha_k v_k\|_{L^2(X)} = \|q_1\|_{L^2(X)}$. From Eq. (8), we have that for large enough $\delta > 0$ the gap $\gamma_1(\delta) = \lambda_k\sqrt{1+\delta^2\|s_1\|^2_{L^2(x)}} - 1$ is positive. So, taking $\delta = 7(1-\lambda_k)/(\lambda_k\|s_1\|_{L^2(X)})$, we have that $\gamma_1(\delta) > \lambda_k\delta\|s_1\|_{L^2(x)} - (1-\lambda_k) = 6(1-\lambda_k)$, and, hence applying Theorem 4 of Appendix B.3), we conclude that

$$\|h_0 - h_\theta\|_{L_2(X)} \leq \frac{1}{1-\kappa}\left(\frac{5\|h_0\|_{L^2(X)} + \lambda_k^2\,\|s_1\|_{L^2(X)}}{\lambda_k^2\|s_1\|_{L^2(X)}}\|q_1\|_{L^2(X)} + \frac{2\,\|h_0\|_{L^2(X)}}{\lambda_k}\,\mathcal{E}_d(\theta,\delta)\right), \tag{18}$$

whenever $7\|q_1\|_{L^2(x)}/\|s_1\|_{L^2(x)} + 3\,\mathcal{E}_1(\theta,\omega,\delta)\,\lambda_k \leq \kappa\lambda_k^2$. Therefore, whenever $\|s_1\|_{L^2(X)} \gg \|q_1\|_{L^2(X)}$, we are able to learn the most dominant part of the structural function with just one feature.

C Statistical Analysis

We recall our estimation procedure.

Estimator: Given i.i.d. data $(Y_i, X_i, Z_i)_{i=1}^n$, we estimate h_0 via the two-stage procedure:

$$\hat{C}_{ZX,\theta} = \frac{1}{n}\sum_{i=1}^n \psi_\theta(Z_i)\varphi_\theta(X_i)^\intercal \in \mathbb{R}^{d\times d}, \quad \hat{g} = \frac{1}{n}\sum_{i=1}^n Y_i\psi_\theta(Z_i) \in \mathbb{R}^d. \tag{19}$$

$$\hat{\beta}_\theta = \hat{C}^{-1}_{ZX,\theta}\hat{g}, \quad \hat{h}_\theta(x) = \varphi_\theta(x)^\intercal\hat{\beta}_\theta. \tag{20}$$

C.1 Proof of Theorem 1

We decompose the excess risk as

$$\|\hat{h}_\theta - h_0\|_{L_2(X)} \leq \|\hat{h}_\theta - h_\theta\|_{L_2(X)} + \|h_\theta - h_0\|_{L_2(X)}.$$

Proposition 9 provides the control on the estimation error, that is, w.p.a.l. $1-\tau$

$$\|\hat{h}_\theta - h_\theta\|_{L_2(X)} \leq \frac{c}{c_{\varphi_\theta^{(d)},\psi_\theta^{(d)}}}\sqrt{\frac{d}{n}}\sqrt{\sigma^2_{\text{eff}} + \frac{\rho^2}{n}}\,\log\frac{4}{\tau}, \tag{21}$$

where $\sigma^2_{\text{eff}} := \|h_0 - h_\theta\|^2_{L^2(X)} + \sigma_U^2$ and $c > 0$ is an absolute constant.

C.2 Proof of auxiliary results

We also recall the definition of sub-Gaussian random variables.

Definition 1 (Sub-Gaussian random vector). *We define the proxy variance $\overline{\sigma}_{\mathbf{x}}$ of a real-valued random variable $\mathbf{x}$, which controls the tail behavior of $\mathbf{x}$ via $\mathbb{P}(|\mathbf{x} - \mathbb{E}[\mathbf{x}]| > t) \leq 2\exp\big(-t^2/\overline{\sigma}^2_{\mathbf{x}}\big)$. A random vector $X \in \mathbb{R}^d$ will be called sub-Gaussian iff, there exists an absolute constant such that, for all $u \in \mathbb{R}^d$, $\overline{\sigma}_{\langle X,u\rangle} \leq c\|\langle X,u\rangle - \mathbb{E}\langle X,u\rangle\|_{L_2(\mathbb{P})}$.*

Proposition 9 (Concentration for Sub-Gaussian Random Variables). *Let Assumptions 2-4 be satisfied. Given $\tau \in (0,1)$, let $n \geq 16\, d\, \rho^2\, \log(4d/\tau) c^{-2}_{\varphi_\theta^{(d)}, \psi_\theta^{(d)}}$. Then there exists an absolute constant $c > 0$ such that, with probability at least $1-\tau$,*

$$\|\hat{h}_\theta - h_\theta\|_{L_2(X)} \leq \frac{c}{c_{\varphi_\theta^{(d)}, \psi_\theta^{(d)}}} \sqrt{\frac{d}{n}} \sqrt{\sigma^2_{\mathit{eff}} + \frac{\rho^2}{n}}\, \log \frac{4}{\tau}, \tag{22}$$

where $\sigma^2_{\mathit{eff}} := \|h_0 - h_\theta\|^2_{L^2(X)} + \sigma^2_U$.

Proof of Proposition 9. Let us denote

$$\hat{g} \doteq \hat{\mathbb{E}}_n[Y \psi^\theta(Z)] \qquad \text{and} \qquad g \doteq \mathbb{E}[Y \psi^\theta(Z)]. \tag{23}$$

We have

$$\|\hat{h}_\theta - h_\theta\|_{L_2(X)} = \left\| C^{1/2}_{X,\theta} \left(\hat{\beta}_\theta - \beta_\theta \right) \right\|_{\ell_2} \leq \underbrace{\left\| C^{1/2}_{X,\theta} \hat{C}^{-1}_{ZX,\theta} C^{1/2}_{Z,\theta} \right\|}_{A_1} \underbrace{\left\| C^{-1/2}_{Z,\theta} \left(\hat{g} - \hat{C}_{ZX,\theta} \beta_\theta \right) \right\|_{\ell_2}}_{A_2}. \tag{24}$$

Lemma 1 below guarantees that $\mathbb{P}(A_1 \leq 2/c_{\varphi_\theta^{(d)}, \psi_\theta^{(d)}}) \geq 1 - \tau/2$. Lemma 2 below guarantees with probability at least $1 - \tau/2$ that

$$A_2 \leq \frac{c}{\sqrt{n}} \sqrt{d\, \sigma^2_{\text{eff}} + \frac{d\, \rho^2}{n}}\, \log \frac{4}{\tau}.$$

where $\sigma^2_{\text{eff}} := \|h_0 - h_\theta\|^2_{L^2(X)} + \sigma^2_U$.

An union gives the result with an absolute constant $c > 0$ possibly different from the previous.

□

Lemma 1 (Matrix Perturbation Control). *Let Assumptions 2 and 3 be satisfied. Assume in addition that*

$$n \geq 16 \frac{d\rho^2 \log(4d/\tau)}{c^2_{\varphi_\theta^{(d)}, \psi_\theta^{(d)}}}. \tag{25}$$

Then with probability at least $1 - \tau/2$,

$$\left\| C^{1/2}_{X,\theta} \widehat{C}^{-1}_{ZX,\theta} C^{1/2}_{Z,\theta} \right\| \leq \frac{2}{c_{\varphi_\theta^{(d)}, \psi_\theta^{(d)}}}.$$

Proof of Lemma 1. Define the centered random matrix

$$\eta_i \doteq C^{-1/2}_{Z,\theta} (\psi_\theta(Z_i) \varphi_\theta(X_i)^\intercal - C_{ZX,\theta}) C^{-1/2}_{X,\theta}.$$

We have

$$C^{-1/2}_{Z,\theta} (\hat{C}_{ZX,\theta} - C_{ZX,\theta}) C^{-1/2}_{X,\theta} = \frac{1}{n} \sum_{i=1}^n \eta_i.$$

Operator norm bound: By Assumption 2, we have $\|\eta_i\| \leq d\rho^2 + 1$ almost surely. Indeed

$$\begin{aligned} \|\eta_i\| &= \|C^{-1/2}_{Z,\theta} \psi_\theta(Z_i) \varphi_\theta(X_i)^\intercal C^{-1/2}_{X,\theta} - C^{-1/2}_{Z,\theta} C_{ZX,\theta} C^{-1/2}_{X,\theta}\| \\ &\leq \|C^{-1/2}_{Z,\theta} \psi_\theta(Z_i)\|_{\ell_2} \|C^{-1/2}_{X,\theta} \varphi_\theta(X_i)\|_{\ell_2} + \|C^{-1/2}_{Z,\theta} C_{ZX,\theta} C^{-1/2}_{X,\theta}\| \\ &\leq \sqrt{d}\rho \cdot \sqrt{d}\rho + 1 \leq d\rho^2 + 1 \qquad a.s. \end{aligned}$$

Covariance bounds: For the left covariance,

$$\begin{aligned}\mathbb{E}[\eta_i\eta_i^*] &= C_{Z,\theta}^{-1/2}\mathbb{E}[(\psi_\theta(Z)\varphi_\theta(X)^\intercal - C_{ZX,\theta})C_{X,\theta}^{-1}(\psi_\theta(Z)\varphi_\theta(X)^\intercal - C_{ZX,\theta})^*]C_{Z,\theta}^{-1/2}\\ &= C_{Z,\theta}^{-1/2}\mathbb{E}[\psi_\theta(Z)\varphi_\theta(X)^\intercal C_{X,\theta}^{-1}\varphi_\theta(X)\psi_\theta(Z)^\intercal - C_{ZX,\theta}C_{X,\theta}^{-1}C_{ZX,\theta}^*]C_{Z,\theta}^{-1/2}\\ &\preceq C_{Z,\theta}^{-1/2}\mathbb{E}[\psi_\theta(Z)\varphi_\theta(X)^\intercal C_{X,\theta}^{-1}\varphi_\theta(X)\psi_\theta(Z)^\intercal]C_{Z,\theta}^{-1/2}\\ &= C_{Z,\theta}^{-1/2}\mathbb{E}[\psi_\theta(Z)\psi_\theta(Z)^\intercal\varphi_\theta(X)^\intercal C_{X,\theta}^{-1}\varphi_\theta(X)]C_{Z,\theta}^{-1/2}\\ &\preceq d\rho^2\cdot C_{Z,\theta}^{-1/2}C_{Z,\theta}C_{Z,\theta}^{-1/2} = d\rho^2 I_d.\end{aligned}$$

Similarly, $\mathbb{E}[\eta_i^*\eta_i] \preceq d\rho^2 I_d$.

Thus $\sigma_L^2 = \sigma_R^2 \le nd\rho^2$. Applying Theorem 5 gives w.p.a.l. $1-\tau/2$

$$\|C_{Z,\theta}^{-1/2}(\widehat{C}_{ZX,\theta} - C_{ZX,\theta})C_{X,\theta}^{-1/2}\| \le \sqrt{\frac{2d\rho^2\log(4d\tau^{-1})}{n}} + \frac{1+d\rho^2}{3}\frac{\log(4d\tau^{-1})}{n} =: \Delta_n(\tau). \tag{26}$$

Using Weyl's inequality (Proposition 3) and Assumption 3, we deduce that the smallest singular of $C_{Z,\theta}^{-1/2}\widehat{C}_{ZX,\theta}C_{X,\theta}^{-1/2}$ satisfies

$$\sigma_{\min}(C_{Z,\theta}^{-1/2}\widehat{C}_{ZX,\theta}C_{X,\theta}^{-1/2}) \ge c_{\varphi_\theta^{(d)},\psi_\theta^{(d)}} - \Delta_n(\tau) \ge c_{\varphi_\theta^{(d)},\psi_\theta^{(d)}}/2 > 0,$$

where the last inequality follows from sample complexity condition Eq. (25). Since $\|A\| = \sigma_{\min}^{-1}(A^{-1})$, we get the result. □

Lemma 2 (Vector Concentration with Sub-Gaussian variables)**.** *Let the assumptions of Theorem 1 be satisfied. Define*

$$\xi := (Y - h_\theta(X))C_{Z,\theta}^{-1/2}\psi_\theta(Z).$$

Then there exists an absolute constant $c>0$ such that, with probability at least $1-\tau/2$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}[\xi]\right\|_{\ell_2} \le c\sqrt{\frac{d}{n}}\sqrt{\sigma_{\mathit{eff}}^2 + \frac{\rho^2}{n}}\,\log\frac{4}{\tau},$$

where $\sigma_{\mathit{eff}}^2 := \|h_0 - h_\theta\|_{L^2(X)}^2 + \sigma_U^2$.

Proof. We need to check the moment condition of Proposition 10 with $A = \xi$. This condition is obviously satisfied for $p=2$ with $\sigma^2 = \mathbb{E}\|A\|_{\ell_2}^2$. Next for any $p \ge 3$, the Cauchy-Schwarz inequality and the equivalence of moment property give

$$\mathbb{E}\|\xi\|_{\ell_2}^m \le \left(\mathbb{E}\|\xi\|_{\ell_2}^4\right)^{1/2}\left(\mathbb{E}\|\xi\|_{\ell_2}^{2(m-2)}\right)^{1/2} \lesssim \mathbb{E}\left[\|\xi\|_{\ell_2}^2\right]\left(\mathbb{E}\|\xi\|_{\ell_2}^{2(m-2)}\right)^{1/2}.$$

For the second order moment, using Cauchy-Schwarz and the equivalence of moments for sub-Gaussian random variables:

$$\begin{aligned}\mathbb{E}[\|\xi\|_{\ell_2}^2] &= \mathbb{E}[(h_0(X) - h_\theta(X) + U)^2\|C_{Z,\theta}^{-1/2}\psi_\theta(Z)\|_{\ell_2}^2]\\ &\le \mathbb{E}^{1/2}\left[(h_0(X) - h_\theta(X) + U)^4\right]\mathbb{E}^{1/2}\left[\|C_{Z,\theta}^{-1/2}\psi_\theta(Z)\|_{\ell_2}^4\right]\\ &\le 32\,\mathbb{E}_Z\left[\|C_{Z,\theta}^{-1/2}\psi_\theta(Z)\|_{\ell_2}^2\right]\sigma_{\text{eff}}^2 = 32d\,\sigma_{\text{eff}}^2.\end{aligned} \tag{27}$$

For higher order moments, we apply Lemma 3 with $v = Y - h_\theta(X)$ and $V = C_{Z,\theta}^{-1/2}\psi_\theta(Z)$. Hence we get, for any $p \ge 3$,

$$\mathbb{E}^{1/2}[\|\xi\|_{l_2}^{2(p-2)}] \le \sqrt{2}\,\overline{\sigma}_{Y-h_\theta(X)}^{p-2}\,(\rho\sqrt{d})^{p-2}\sqrt{(p-2)!} \tag{28}$$

By Definition 1 of sub-Gaussian distributions, there exists an absolute constant $c > 0$ such that

$$\overline{\sigma}^2_{Y-h_\theta(X)} \leq 2\,c\,\big(\mathrm{Var}\big(h_0(X) - h_\theta(X)\big) + \mathrm{Var}(U)\big) \leq 2\,c\,\sigma^2_{\text{eff}}.$$

We apply Proposition 10 to get w.p.a.l. $1 - \tau/2$

$$\left\| \frac{1}{n} \sum_{i\in[n]} \xi_i - \mathbb{E}[\xi] \right\|_{l_2} \leq c \sqrt{\frac{d}{n}} \sqrt{\sigma^2_{\text{eff}} + \frac{\rho^2}{n}}\, \log \frac{4}{\tau}. \tag{29}$$

where $c > 0$ is absolute constant possibly different from the previous display.

□

Lemma 3 (Moment Bound for Sub-Gaussian Product)**.** *Let $Z = vV$ where v is a real-valued sub-Gaussian random variable with proxy variance $\overline{\sigma}_v$, and V is a d-dimensional random vector with $\|V\|_{\ell_2} \leq \rho\sqrt{d}$ almost surely. Then for any integer $p \geq 3$, we have*

$$\mathbb{E}^{1/2}[|v|^{2(p-2)} \|V\|^{2(p-2)}_{\ell_2}] \leq \sqrt{2}\,\overline{\sigma}_v^{p-2}\,(\rho\sqrt{d})^{p-2}\sqrt{(p-2)!} \tag{30}$$

Proof. Since $\|V\|_{\ell_2} \leq \rho\sqrt{d}$ almost surely, we have

$$\mathbb{E}[|v|^{2(p-2)} \|V\|^{2(p-2)}_{\ell_2}] \leq (\rho\sqrt{d})^{2(p-2)} \mathbb{E}[|v|^{2(p-2)}].$$

We will use the tail characterization of sub-Gaussian random variables and the integral representation of moments to bound $\mathbb{E}[|v|^{2(p-2)}]$.

Since v is sub-Gaussian with proxy variance $\overline{\sigma}_v$, there exists an absolute constant $c > 0$ such that for all $t \geq 0$:

$$\mathbb{P}(|v| \geq t) \leq 2\exp\left(-\frac{t^2}{\overline{\sigma}_v^2}\right). \tag{31}$$

Set $m = 2(p-2)$. Using the integral representation of moments:

$$\mathbb{E}[|v|^{2(p-2)}] = \int_0^\infty m t^{m-1} \mathbb{P}(|v| \geq t)\, dt \leq 2\int_0^\infty m\, t^{m-1} \exp\left(-\frac{t^2}{\overline{\sigma}_v^2}\right) dt. \tag{32}$$

Integration by parts gives

$$\int_0^\infty m\, t^{m-1} \exp\left(-\frac{t^2}{\overline{\sigma}_v^2}\right) dt = \frac{m}{2}\,\overline{\sigma}_v^m\,\Gamma\left(\frac{m}{2}\right)$$

where Γ is the Gamma function.

Since $m = 2(p-2)$, we have $\Gamma\left(\frac{m}{2}\right) = (p-3)!$ and consequently

$$\int_0^\infty m\, t^{m-1} \exp\left(-\frac{t^2}{\overline{\sigma}_v^2}\right) dt = (p-2)!$$

Thus we get:

$$\mathbb{E}^{1/2}[|v|^{2(p-2)} \|V\|^{2(p-2)}_{\ell_2}] \leq \sqrt{2}(\overline{\sigma}_v \rho\sqrt{d})^{(p-2)} \sqrt{(p-2)!} \tag{33}$$

□

C.3 Concentration inequalities

We present here some well-known concentration inequalities for operators that we use in our analysis.

We recall first a version of Bernstein inequality, due to Pinelis and Sakhanenko, for random variables in a separable Hilbert space, see (Caponnetto & De Vito, 2007, Proposition 2).

Proposition 10. *Let A_i, $i \in [n]$ be i.i.d copies of a random variable A in a separable Hilbert space with norm $\|\cdot\|$. If there exist constants $\Lambda > 0$ and $\sigma > 0$ such that for every $m \geq 2$, $\mathbb{E}\|A\|^m \leq \frac{1}{2} m! \Lambda^{m-2} \sigma^2$, then with probability at least $1-\delta$,*

$$\left\| \frac{1}{n} \sum_{i \in [n]} A_i - \mathbb{E}A \right\| \leq \frac{4\sqrt{2}}{\sqrt{n}} \log \frac{2}{\delta} \sqrt{\sigma^2 + \frac{\Lambda^2}{n}}. \tag{34}$$

The following result is call the noncommutative Bernstein inequality. It was first derived by Ahlswede & Winter (2002). The following version can be found in Tropp (2015).

Theorem 5 (Matrix Bernstein Inequality, Tropp (2015))**.** *Let $\{A_k\}_{k=1}^n$ be independent random matrices of size $d_1 \times d_2$ with $\mathbb{E}[A_k] = 0$. Assume that $\|A_k\| \leq R$ almost surely for all k. Define the matrix variance parameters*

$$\sigma_L^2 := \left\| \sum_{k=1}^n \mathbb{E}[A_k A_k^*] \right\|, \qquad \sigma_R^2 := \left\| \sum_{k=1}^n \mathbb{E}[A_k^* A_k] \right\|. \tag{35}$$

Then for any $t \geq 0$,

$$\mathbb{P}\left\{ \left\| \sum_{k=1}^n A_k \right\| \geq t \right\} \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{\max\{\sigma_L^2, \sigma_R^2\} + Rt/3} \right). \tag{36}$$

A more convenient and equivalent of Eq. (36) is, for any $\tau \in (0,1)$, w.p.a.l. $1-\tau$

$$\left\| \frac{1}{n} \sum_{k=1}^n A_k \right\| \leq \sqrt{\frac{2 \max\{\sigma_L^2, \sigma_R^2\} \log\left(\frac{d_1+d_2}{\tau}\right)}{n^2}} + \frac{R}{3} \frac{\log\left(\frac{d_1+d_2}{\tau}\right)}{n}. \tag{37}$$

D Experimental Details

The code needed to reproduce all figures in this work can be found at https://anonymous.4open.science/r/IV-55DE/README.md.

D.1 Structural functions and alignment in dSprites

The dSprites structural function $h_0 = h_{\text{old}}$, as used in, *e.g.*, Xu et al. (2024) is not a new idea. However, it differs from an alternative on which SpecIV(Sun et al., 2025) and DFIV(Xu et al., 2021) were originally evaluated. That one takes the form,

$$h_0(x) = (\|BX\|^2 - 5000)/1000$$

with $B \in \mathbb{R}^{10 \times 4096}$ consisting of i.i.d. $\text{Uniform}([0,1])$ entries. This function evaluates the squared norm of 10 random linear measurements of the image and returns a linear transformation of it. Since the entries of B are i.i.d., this h_0 has no clear dependence on the values of Z. To see that it should be extremely difficult to recover h_0 from the relation $\mathcal{T} h_0 = r_0$, note that moving the sprite vertically, while keeping the orientation, scale and x-position constant (i.e. not changing the instrumental variable) can lead to different values of h_0. This does not quite imply that h_0 has a non-trivial projection onto the kernel of X because the distribution of the images X is not vertical-shift-invariant. But it does suggest that the instrument Z is essentially uninformative about an important property of the image needed to recover h_0. One can reliably learn part of h_0 due its monotonicity in the sprite's scale but beyond this, any dependence on x-position and orientation is likely to be extremely irregular and not representative of real-life applications of IV regression. For these reasons we choose not to utilise this benchmark.

D.2 SPECTRAL ALIGNMENT IN DSPRITES

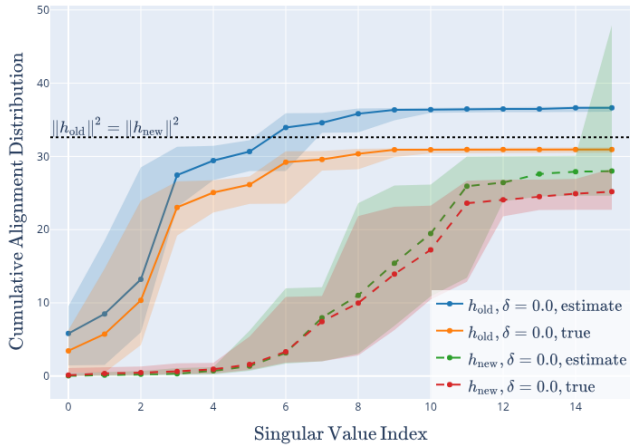


Figure 5: Distributions of cumulative alignment estimates and true values $\|\Pi_{\hat{v}^{(i)}} h_0\|^2$ for increasing i, evaluated on separately fitted models (with identical parameters) for $h_0 = h_{\text{old}}, h_{\text{new}}$ at $\delta = 0$.

Standard dSprites is well-aligned. Intuitively, most of the variability in the image should be explained by the heart's position and scale since these quantities determine where the non-zero pixels are and how many of them there are in total. By contrast, the rotation angle only explains the location of non-zero pixels on the boundary of the broad region where the sprite is located (since most of the sprite's area is preserved by rotations around its centre). Therefore, structural functions that mostly depend on scale and x-position should intuitively be easier to learn than those that are sensitive to rotation. Since A in the dSprites structural function "measures" distance of the non-zero pixels from the vertical central bar in the image, it should be mostly sensitive to x-positions and the number of non-zero points to sum over.

Moreover, following Meunier et al. (2025), we can compute the empirical-distribution-based singular value decomposition of a fitted $\hat{\mathcal{T}}_d$ for $\delta = 0$ (which we just treat as an operator $L^2(X) \to L^2(Z)$). That is, we compute features $\hat{u}_i, \hat{v}_i$ and positive real numbers $\hat{\sigma}_i$ such that,

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

The feature functions $(u_{1:d}), (v_{1:d})$ are orthonormal systems with respect to the empirical distributions of Z, X based on the samples used for the SVD estimation. If $\|\hat{\mathcal{T}}_d - \mathcal{T}_d\|$ is sufficiently small, we are guaranteed (Meunier et al., 2025) that the projections onto singular features $\hat{v}_i$ are close to the projections onto the singular features of the true truncated conditional mean operator $\mathcal{T}_d$.

We compute the individual-feature projection lengths $|\langle \hat{v}_i, h_0 \rangle|$ to conclude that most of h_{old} is supported on the leading singular functions whose singular values are close to 1. Moreover, we compare these true alignment values to their estimates that are not based on h_0 (Meunier et al., 2025). This is to ensure that spectral (mis-)alignment in the dSprites setting can also be detected in a real-life setting where the structural function is not available. Denote the projection onto the leading i singular X-features of $\hat{\mathcal{T}}_d$ with $\Pi_{\hat{v}^{(i)}}$ As seen in Figure 5, both the estimates of the squared projection length onto the leading i features of $\hat{\mathcal{T}}_d$ and the true values $\|\Pi_{\hat{v}^{(i)}} h_0\|^2$ are in agreement with one another and indicate that h_{old} lies in the top of the spectrum $\mathcal{T}$.

A more difficult structural function. Inspired by the discussion above, we would like to evaluate our methods in a more challenging setting. An intuitive method of obtaining such a benchmark is constructing a structural function which is most sensitive to the rotation of the sprite as opposed to its position or scale. In order to work with shapes whose orientation is easier to estimate from the image,

we replace hearts with ellipses in our benchmark. For an image X of an ellipse, let $\tilde{X}$ be the result of convolving it with a smoothing gaussian kernel with a small bandwidth of around 4 pixels. This is done in order to decrease our method's sensitivity to noise. Then let $x_{\max,i} \doteq \max_{j\in\{1,\dots,64\}} \tilde{X}_{ji}$ be the maximum of $\tilde{X}$ along its vertical bars, and $y_{\max,i} \doteq \max_{j\in\{1,\dots,64\}} \tilde{X}_{ij}$ be the corresponding maximum over horizontal bars. Then, one expects the norms of $x_{\max}$ and $y_{\max}$ to be roughly proportional to the length of the ellipses projection onto the horizontal and vertical axes of the image. Hence, taking

$$h_{\text{new}} \doteq a \cdot (\|x_{\max}\|_{\ell_2} / \|y_{\max}\|_{\ell_2} - b),$$

one should obtain a function which depends on the ellipse's orientation and is insensitive to its scale or position. The scalars a, b are chosen in order to match the first two moments of the original structural function h_{old}. This is done in order to not artificially change the downstream problem's difficulty by making the norm of the signal relative to noise different.

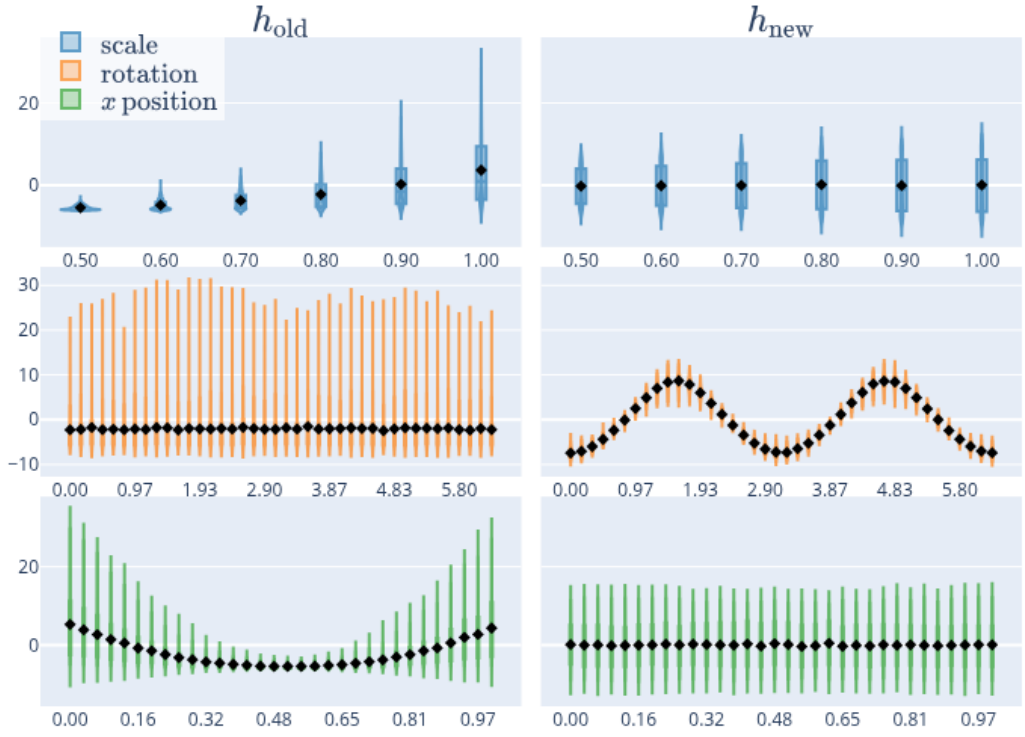


Figure 6: Comparison of how the distributions of h_{old} and h_{new} vary with each component of the instrument Z (scale, orientation or x position). The values of each component of Z in dSprites are quantised. The x-axis positions in the figure correspond to those values. For each value of a component of Z, we display the distribution of the values of $h_0 = h_{\text{old}}, h_{\text{new}}$ evaluated on images where the component takes that value. The marked values are the means of each bin.

As seen in Figure 6, while h_{old} has a clear dependence on the sprite position and orientation, h_{new} is only visibly sinusoidal in its orientation and insensitive to the latter two. Figure 5 confirms that data generated using h_{new} should be considerably more challenging for spectral-feature-based IV regression, since its projection onto the leading singular functions is close to zero for all models with $\delta = 0$ that we have fitted. It is supported further in the spectrum, and in fact $\|\Pi_{\hat{v}^{(d)}} h_{\text{new}}\|^2$ is far from $\|h_{\text{new}}\|^2$. This means that not even all of the learned features span h_{new}.

D.3 Synthetic data models

We train our models using the same architecture as (Meunier et al., 2025), which is described in Table 1. Notably, the first layer of this network utilises the $\sin(x)^2 + x$ activation proposed by Ziyin et al. (2020), which enables it to learn the oscillatory basis functions more reliably. To make

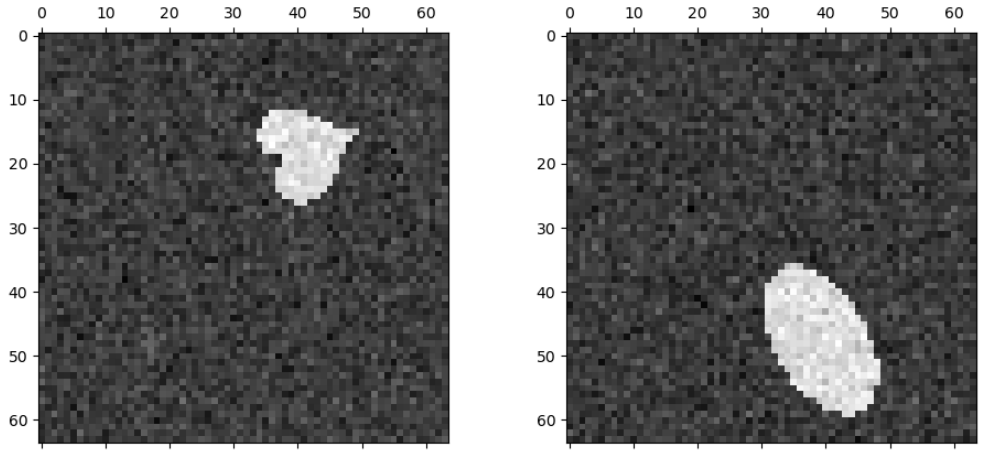


Figure 7: Examples of sprite images on which h_{old} (left) and h_{new} (right) are evaluated.

the improvements in the alignment of leading features with δ clearer, we learn 10 Z, X features on synthetic datasets with $d = 11$. The models are trained on 50000 (Z, X, Y) samples using the Adam algorithm.

Layer	Configuration
1	Input: 1
2	FC(1,50), $x \mapsto x + \sin^2 x$
3	FC(50, 50), GeLU
4	FC(50, 10)

Table 1: Z, X feature networks for synthetic data

Evaluation of the benefits of using a positive δ on a bigger range of c_σ values than in the main body of the work can be found in Figure 8

D.4 Spectral learning models for dSprites

For learning spectral features we utilise nearly the same models as Sun et al. (2025). The only difference is replacing ReLU activations with GeLU which we found to lead to slightly easier training. "BN" in Table 2 refers to Batch Normalisation (Ioffe & Szegedy, 2015). The models are trained on 25000 (Z, X, Y) samples and optimised with the Adam algorithm.

We observed that feature models utilising 32 features, as originally proposed in Sun et al. (2025), were very prone to overfitting, often before this became apparent in the observed loss, making early stopping difficult to implement. Hence, we trained the models with 16 features instead. This led to better performance across the board for h_{old}. In h_{new}, vanilla SpecIV, trained with 32 features attained a smaller loss than it did for 16. However, the performance was still significantly below that of DFIV or spectral learning with a non-zero δ. For all the "optimal" values of δ (between 0.1 and 1.0) we observed benefits from using 16 features over 32.

D.5 DFIV models

For the comparison to DFIV, which is the only viable competitor to SpecIV, we utilise the same architecture as proposed in the orignal work Xu et al. (2021). "SN" in Table 3 refers to spectral

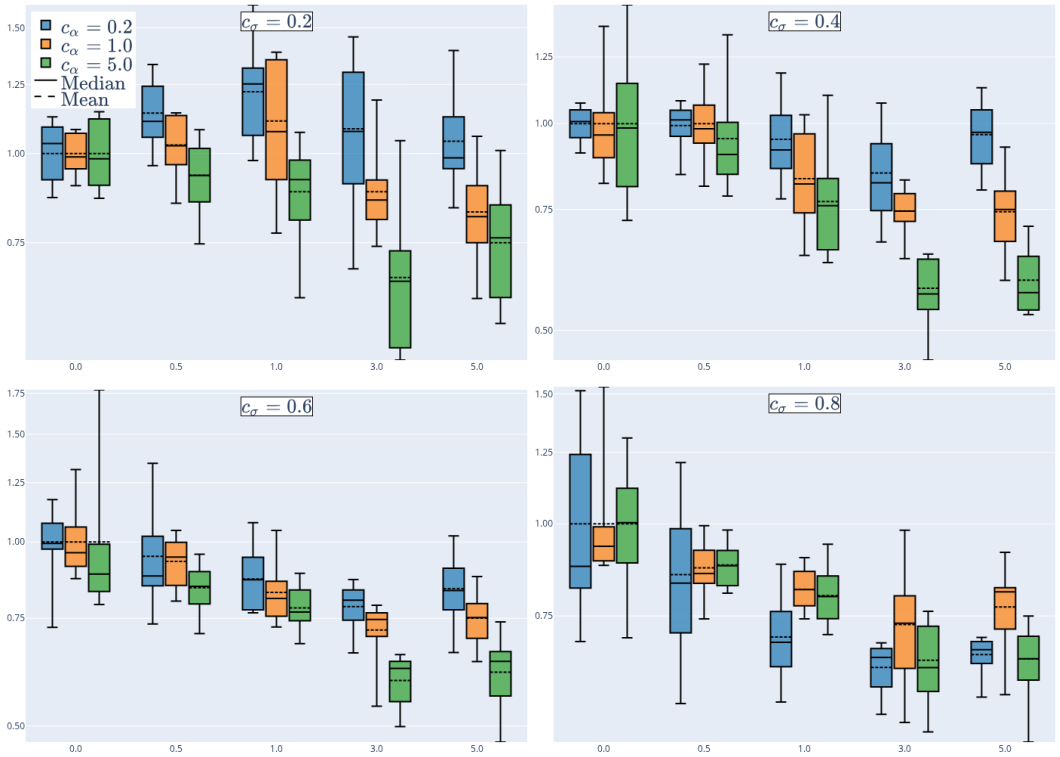


Figure 8: Distributions of relative IV regression MSEs for the synthetic example with $\delta \in \{0, 0.5, 1.0, 3.0, 5.0\}$.

Layer	Configuration
1	Input: 4096
2	FC(4096,1024), BN, GeLU
3	FC(1024,512), BN, GeLU
4	FC(512,128), BN, GeLU
5	FC(128,16)

X - features

Layer	Configuration
1	Input: 3
2	FC(3,256), BN, GeLU
3	FC(256,128), BN, GeLU
4	FC(128,128), BN, GeLU
5	FC(128,16)

Z - features

Table 2: Architectures of spectral learning networks for dSprites datasets

normalisation proposed by Miyato et al. (2018) and "BN" is Batch Normalisation. The models are trained on 25000 (Z, X, Y) samples and optimised with the Adam algorithm.

Since we decided to decrease the number of features employed by spectral learning, relative to the standard architecture used in the literature, we also investigated whether doing so leads to better performance in DFIV. The opposite was observed and hence we retain the original models.

D.6 Comparison to KIV

Since Kernel IV (KIV) proposed by Singh et al. (2019) is a very popular and easily applicable method, we also include it in our benchmarks. We evaluated it using the Gaussian kernel with a bandwidth proportional to $m\sqrt{d}$ where m is the median distance between pairs samples on which the kernel is evaluated, and d is the dimension of the samples (i.e. 3 for Z and 4096 for X).

Layer	Configuration	Layer	Configuration
1	Input: 4096	1	Input: 3
2	FC(4096,1024), SN, ReLU	2	FC(3,256), SpectralNorm, ReLU
3	FC(1024,512), SN, ReLU, BN	3	FC(256,128), SN, ReLU, BN
4	FC(512,128), SN, ReLU	4	FC(128,128), SN, ReLU, BN
5	FC(128,32), SN, BN, Tanh	5	FC(128,32), BN, ReLU
	X - features		Z - features

Table 3: Architectures of DFIV networks for dSprites datasets

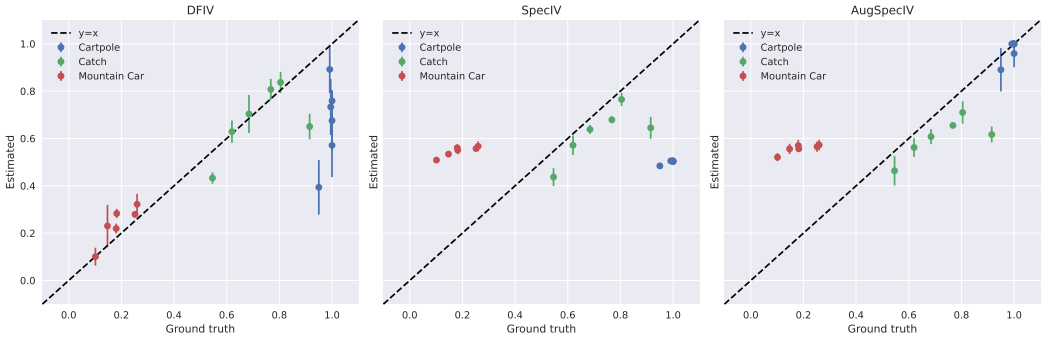


Figure 9: Scatter plot of estimated policy values vs. ground truth.

D.7 OPE Experiment Protocol and Results

Hyperparameter optimization. For each method and task, we randomly sample 100 hyperparameter combinations from the grids in Tables 4 and 5, and select the configuration achieving the lowest stage-2 mean squared error on the corresponding task with $p = 0.2$. Orthonormal regularization is applied to both spectral methods, while the δ parameter is specific to AugSpecIV, i.e., $\delta = 0$ for SpecIV and is tuned as any other hyperparameter for AugSpecIV. With the chosen hyperparameters, we run each method five times across all tasks and noise levels and report the mean and standard deviation of the estimated policy values.

Hyperparameter	Values
Training Steps	10^5
Batch Size	2048
Stage-1 reg.	$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$
Stage-2 reg.	$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$
Value reg.	$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$
Instrumental reg.	$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$
Value learning rate	$\{10^{-5}, 3{\cdot}10^{-5}, 10^{-4}, 3{\cdot}10^{-4}, 10^{-3}\}$
Instrumental learning rate	$\{10^{-5}, 3{\cdot}10^{-5}, 10^{-4}, 3{\cdot}10^{-4}, 10^{-3}\}$
X net layer sizes	$(50, 50)$
Z net layer sizes	$\{(50, 50), (100, 100), (150, 150)\}$

Table 4: DFIV hyperparameters.

D.7.1 OPE Context: Policies and Data

In OPE, the "offline" constraint is critical:

- We cannot interact with the environment. We cannot take a state s and an action $a \sim \pi$ to observe a new transition (s, a, r, s').

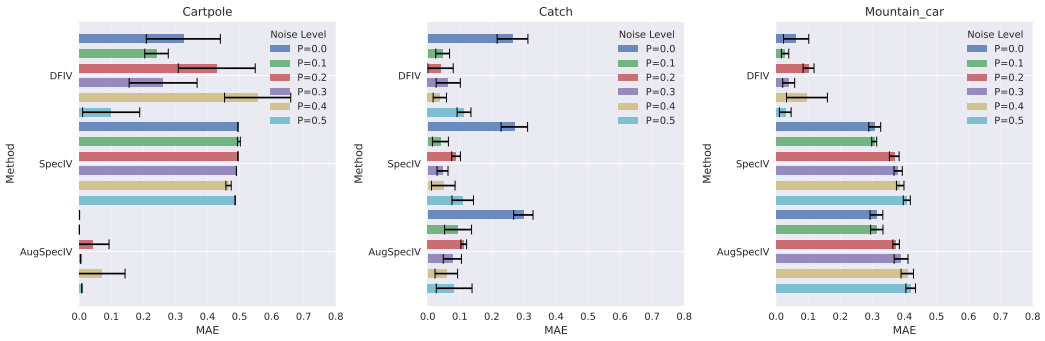


Figure 10: Box plot of mean absolute error (MAE) between policy value and ground-truth.

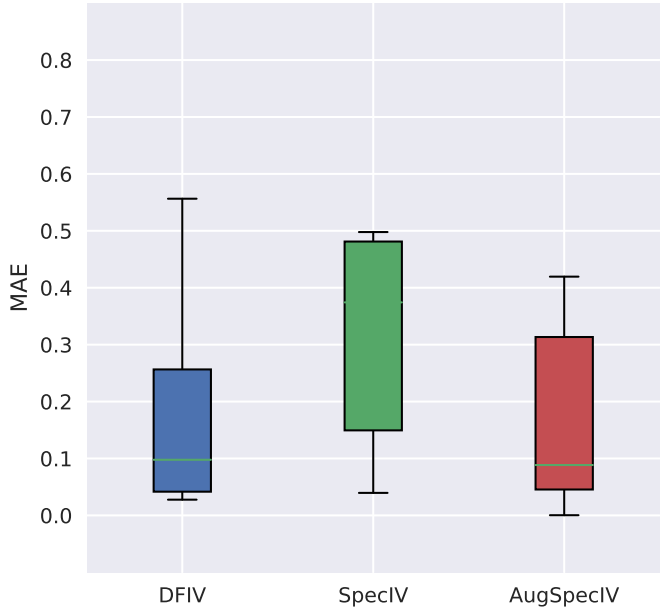


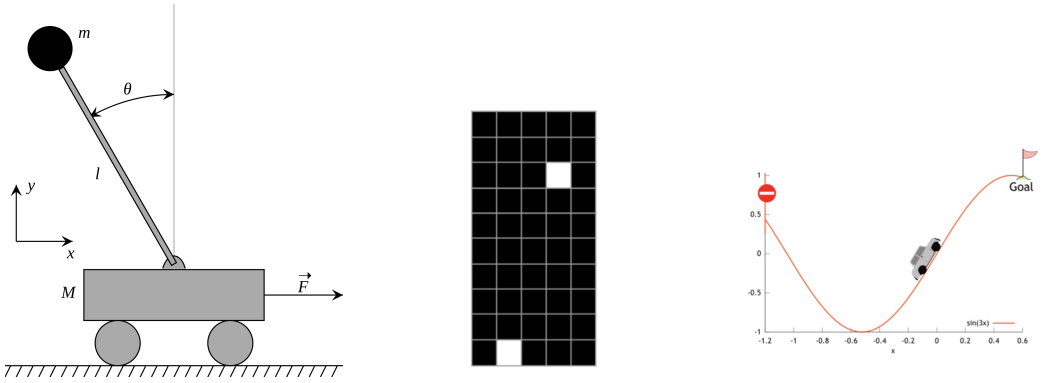
Figure 11: Distribution of MAE across NPIV methods.

- We can compute with π. Since π is a known policy (e.g., a function in our code), we can sample actions $a' \sim \pi(a'|s')$ for any state s' that we observe in our dataset.

D.7.2 NPIV Formulation of OPE and the Compactness Problem

In OPE, to estimate the expected return of a policy π, a common and effective approach consists of first estimating the Q-function Q_π, and then averaging over the initial state distribution. While early methods like Least Squares Temporal Difference (LSTD; Bradtke & Barto, 1996) pioneered this direction using linear function approximation, modern approaches increasingly leverage NPIV regression to handle general function approximation.

Related work on OPE. We briefly discuss existing methods for OPE and refer to Jiang & Xie (2025) and references therein for additional details. Early approaches such as LSTD were restricted



(a) CartPole (b) Catch (c) Mountain Car

Figure 12: Depiction of the considered environments.

Hyperparameter	**Values**
Training Steps	10^4
Batch Size	2048
Stage-1 reg.	$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$
Stage-2 reg.	$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$
Orthonormal reg.	$\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}\}$
Learning rate	$\{10^{-5}, 3{\cdot}10^{-5}, 10^{-4}, 3{\cdot}10^{-4}, 10^{-3}\}$
X net layer sizes	$(50, 50)$
Z net layer sizes	$\{(50, 50), (100, 100), (150, 150)\}$
δ	$\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$

Table 5: SpecIV and AugSpecIV hyperparameters. The neural architectures are the same as those of DFIV described in (Chen et al., 2022), except for the usage of GELU activations.

to linear function approximation. For general function approximation, the dominant approach has historically been Fitted-Q Evaluation (FQE), which iteratively solves a regression problem to minimize the Bellman residual. Recent advances have highlighted that such an approach requires strong assumptions to succeed, namely: having access to a sufficiently expressive function class (*e.g.*, that is realizable and Bellman complete) and using data with good coverage. NPIV formulations offer an alternative perspective to these regression-based methods. Rather than minimizing a squared error directly, NPIV approaches Hu et al. (2024) frame the Bellman equation as a conditional moment restriction as shown below.

Q function and NPIV. Let $(\mathcal{S}, \mathcal{A}, P, R, \mu_0)$ be a Markov Decision Process (MDP) with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P(s' \mid s, a)$, reward distribution $P_{rew}(r \mid s, a)$, and initial distribution μ_0. We denote by R the random reward variable such that its distribution given any action-pair (s, a) is $P_{rew}(\cdot \mid s, a)$.

Given a target policy π and discount factor $\gamma \in (0, 1)$, its value is defined as:

$$\rho(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] = \mathbb{E}_{S_0 \sim \mu_0, A_0 \sim \pi(S_0)}[Q_\pi(S_0, A_0)], \tag{38}$$

where $R_t \sim P_{rew}(\cdot \mid S_t, A_t)$, (S_t, A_t) follows π and P, and Q_π is the state-action value function (i.e. Q-function), given by

$$Q_\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s, A_0 = a\right]. \tag{39}$$

The Q-function, Q_π, is the unique fixed point of the Bellman equation:

$$Q_\pi(s, a) = \mathbb{E}_{R \sim P_{rew}(\cdot|s,a)}[R] + \gamma \mathbb{E}_{S \sim P(\cdot|s,a), A' \sim \pi(\cdot|S')}[Q_\pi(S', A')] \tag{40}$$

We first show that we can rewrite this Bellman Equation as an NPIV problem of the form Eq. (2). A given policy π induces an occupancy measure μ_π on the state-action space $\mathcal{S} \times \mathcal{A}$. We then take $X = (S, A, S', A')$ equipped with the measure such that $(S, A) \sim \mu_\pi, S' \sim P(\cdot \mid S, A), A' \sim \pi(S')$, $Z = (S, A)$ equipped with the measure such that $(S, A) \sim \mu_\pi$ and finally $Y = R$. We take $\mathcal{T}$ as defined in the main body so that Eq. (40) can be re-written as

$$\mathcal{T} h_\pi(S, A) = \mathbb{E}[R|S, A] \qquad \text{i.e.} \qquad \mathcal{T} h_\pi(Z) = \mathbb{E}[Y|Z],$$

where $h_\pi(X) = h_\pi(S, A, S', A') = Q_\pi(S, A) - \gamma Q_\pi(S', A')$. Note that solving NPIV with respect to $\mathcal{T}$ allows us to retrieve Q_π on the support of μ_π. However instead of using μ_π for which we don't have samples, we can use the logging policy. Recall the logging policy π_b and denote by μ_b the induced state-action distribution. We introduce $\mathcal{T}_b$ as the standard conditional expectation operator but we now equip X with $(S, A) \sim \mu_b, S' \sim P(\cdot \mid S, A), A' \sim \pi(S')$ and Z with $(S, A) \sim \mu_b$. The following equation

$$\mathcal{T}_b h_\pi(S, A) = \mathbb{E}[R|S, A] \qquad \text{i.e.} \qquad \mathcal{T}_b h_\pi(Z) = \mathbb{E}[Y|Z],$$

allows to identify $h_\pi(X) = h_\pi(S, A, S', A') = Q_\pi(S, A) - \gamma Q_\pi(S', A')$ on the support μ_b. Crucially, this identification relies on a coverage assumption (Jiang & Xie, 2025) that states that the support of the distribution μ_b covers the support of the target policy. Consider $\mathcal{Q}$ a function class to approximate Q_π. We can retrieve Q_π by solving for an empirical version of the following loss

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{(S,A) \sim \mu_b, R \sim P_{\text{new}}(\cdot|S,A)}[(Y - \mathcal{T}_b(h_q)(S, A))^2] \qquad h_q(s, a, s', a') = q(s, a) - \gamma q(s', a').$$

We can then plug our estimate into Eq. (38) to get an estimate of $\rho(\pi)$. Different methods have been employed with this loss where methods differ by their choice of $\mathcal{Q}$ and their approximation of $\mathcal{T}_b$, see Chen et al. (2022).

However, this formulation is not amenable to spectral feature methods since $\mathcal{T}_b$ is non-compact as X contains Z, and thus cannot be learned using the usual contrastive loss. To overcome this, we instead recover Q_π through a modified NPIV problem. This time consider $X = (S', A')$ equipped

with the measure such that given (s, a), $S' \sim P(\cdot \mid s, a)$, $A' \sim \pi(S')$, $Z = (S, A)$ equipped with the measure such that $(S, A) \sim \mu_b$ and finally $Y(Q_\pi) = -\gamma^{-1}(R - Q_\pi(S, A))$. We take $\tilde{\mathcal{T}}$ as defined in the main body so that Eq. (40) be re-written as

$$\mathbb{E}[Y(Q_\pi)|Z] = \tilde{\mathcal{T}}(Q_\pi)(Z) \tag{41}$$

$\tilde{\mathcal{T}}$ is now compact, but this modified formulation introduces a new challenge: the target outcome $Y(Q_\pi)$ depends on Q_π, which is the very function we are trying to estimate.

This necessitates an iterative procedure for $k = 0, 1, \ldots, K-1$, similar to Fitted Q-Evaluation (FQE):

1. Start with an initial guess, Q_0 (e.g., $Q_0 = 0$).
2. At iteration $k+1$, we solve the NPIV problem defined by Q_k:
   - We construct the target outcome Y_k using our previous estimate Q_k and the observed reward $r(S, A)$ from the data:
   $$Y_k(S, A) = -\frac{1}{\gamma}(r(S, A) - Q_k(S, A))$$
   - We solve the spectral NPIV problem $\mathbb{E}[Y_k \mid Z] = \tilde{\mathcal{T}} Q_{k+1}$ to find the new estimate Q_{k+1}.
3. Repeat until convergence.

This iterative process introduces a potential for dynamic spectral misalignment. The target Y_k changes at every iteration k. If the spectral features required to estimate Y_k (the "outcome-aware" direction) are not the same as the dominant spectral features of $\tilde{\mathcal{T}}$, or if this direction shifts as Q_k converges, an outcome-agnostic method will fail. This is precisely the scenario where our **Augmented Spectral Feature Learning** could help.

D.8 Validity of hyperparameter tuning by cross-validation on the 2SLS loss

The population level optimal estimator of $\mathcal{T}$ based on a fixed set of $\psi^{(d)}, \varphi^{(d)}$ features is $\Pi_{\psi^{(d)}} \mathcal{T} \Pi_{\varphi^{(d)}}$ where $\Pi_{\varphi^{(d)}}, \Pi_{\psi^{(d)}}$ denote the orthogonal projections onto the spans of the learned X, Z features. By minimizing the contrastive loss we can ensure that our estimator is indeed close to having this structure, with whatever features are obtained in the process.

Given $\widehat{\mathcal{T}}$ an estimation of $\mathcal{T}$, consider the so-called stage-2 loss, $\ell(h) \doteq \mathbb{E}[(\widehat{\mathcal{T}} h(Z) - Y)^2$ over h spanned by the learned $X-$features. For simplicity assume it is performed with this population-level optimal estimate of the conditional mean $\widehat{\mathcal{T}} = \Pi_{\varphi^{(d)}} \mathcal{T} \Pi_{\psi^{(d)}}$, and let the resulting estimate of the structural function be $\widehat{h}$. We have,

$$\ell(\widehat{h}) = \mathbb{E}\left[(\mathcal{T} h_0(Z) - \Pi_{\psi^{(d)}} \mathcal{T} \Pi_{\varphi^{(d)}} \widehat{h}(Z))^2\right] + \mathbb{E}[(\mathcal{T} h_0(Z) - Y)^2].$$

The second term is a model-independent constant so we can disregard it. If we wanted to evaluate the quality of our model based on this loss, then we should be able to decompose it into a monotone function of $\|h_0 - \widehat{h}\|$ or (to make the task easier) of $\|\mathcal{T}(h_0 - \widehat{h})\|$ and a term that will not vary across different learned feature sets (or is sufficiently small to be negligible). However, there seems to be no clear way of achieving this. Consider the most natural approach below.

Note $\Pi_{\psi^{(d)}} \mathcal{T} \Pi_{\varphi^{(d)}} \widehat{h} = \Pi_{\psi^{(d)}} \mathcal{T} \widehat{h}$ since $\widehat{h}$ is spanned by the $X-$features. Let $\Pi^{\perp}_{\psi^{(d)}}$ be the projection onto the orthogonal complement of the span of the ψ features. Then

$$\mathbb{E}\left[(\mathcal{T} h_0 - \Pi_{\psi^{(d)}} \mathcal{T} \widehat{h})^2\right] = \mathbb{E}\left[(\mathcal{T}(h_0 - \widehat{h}))^2\right] - \mathbb{E}\left[(\Pi^{\perp}_{\psi^{(d)}} \mathcal{T} \widehat{h})^2\right] - 2\mathbb{E}\left[\mathcal{T}(h_0 - \widehat{h}) \cdot \Pi^{\perp}_{\psi^{(d)}} \mathcal{T} \widehat{h}\right].$$

If we could argue that the latter two terms are negligible, then stage 2 error should reflect $\mathbb{E}[(\mathcal{T}(h_0 - \widehat{h}))^2]$ and we would indeed be done. This condition would be satisfied if one could for instance argue that $\|\Pi^{\perp}_{\psi^{(d)}} \mathcal{T} \Pi_{\varphi^{(d)}}\|$ is always small. But there is no clear reason why this should hold. We note that this indicates that minimisation of the 2SLS loss is a generally unprincipled methodology, not only for tuning δ in our setting, but for IV model selection more broadly.

Practical performance. Regardless of the aforementioned obstacles, we find that selecting the model, and, in particular, tuning δ based on $\ell(h)$, the stage-2 loss, yields good results. On the dSprites benchmarks these are consistent with the values selected by the procedure based on maximizing the estimated projection length of h_0 onto the learned $X-$features, as shown in Figure 13.

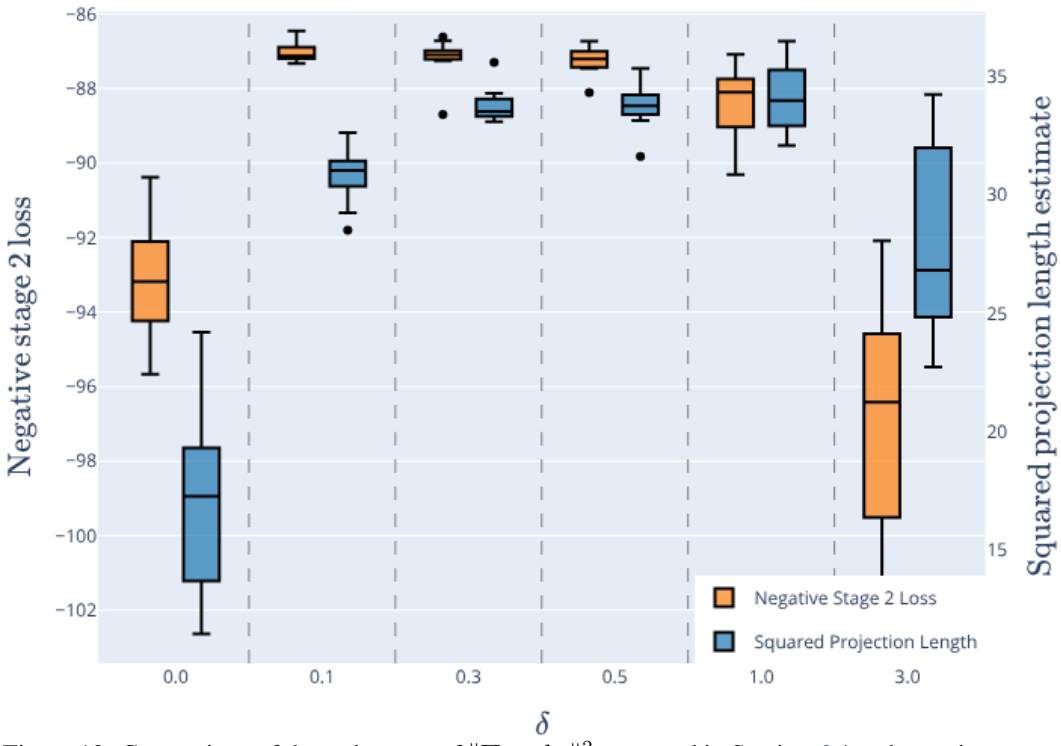


Figure 13: Comparison of the estimators of $\|\Pi_{\varphi^{(d)}_*} h_0\|^2$ proposed in Section 6.4 and negative stage-2 errors, for a range of δ values on the dSprites benchmark with $h_0 = h_{\text{new}}$. Despite the lack of theoretical backing for the latter, both methods attain their maxima near the same values of δ.

E HIGHER RANK PERTURBATIONS

The rank one augmentation we propose extends directly to higher rank perturbations. This setting includes vector–valued outcomes and also the case where one wishes to encourage the learned Z features to retain predictive information about multiple functions of Y.

Let $f_1, \ldots, f_K$ be functions of Y that we would like to be well approximated from Z. For $\underline{\delta} \in \mathbb{R}^K$ we define:

$$\mathcal{L}^{(d)}_{\underline{\delta}}(\theta) \doteq \mathcal{L}_0(\theta) - \sum_{k=1}^{K} \underline{\delta}_k^2 \left\| \Pi_{\psi^{(d)}_\theta} \mathbb{E}[f_k(Y) \mid Z] \right\|^2 .$$

As in the rank one case, to avoid differentiating through matrix inverses, this can be rewritten with auxiliary variables $\omega_k \in \mathbb{R}^d$ as:

$$\mathcal{L}_{\underline{\delta}}^{(d)}(\theta, \omega_1, \ldots, \omega_k) \doteq \mathcal{L}_0(\theta) + \sum_{k=1}^{K} \omega_k^\top C_{\psi_\theta^{(d)}} \omega_k - 2\,\underline{\delta}_k \mathbb{E}[f_k(Y)\psi_\theta^{(d)}(Z)]^\top \omega_k.$$

The optimal features correspond to the truncated SVD of the following rank K perturbed operator:

$$\mathcal{T}_{\underline{\delta}} : L^2(X) \times \mathbb{R}^K \to L^2(Z), \qquad (h, a) \mapsto \mathcal{T}h + \sum_{k=1}^{K} a_k\,\underline{\delta}_k\,f_k.$$

A complete theoretical analysis of the rank K setting is beyond the present scope and we regard it as an important direction for future work. We report preliminary results for $K = 2$ using $f_k = \mathbb{E}[Y^k \mid Z]$ in Figure 14. In the dSprites setting with $h_0 = h_{\text{new}}$, including f_2 alonde yields similar improvements as f_1 alone. Using both functions gives slightly lower error, but the differences are small and do not allow strong conclusions at this stage.

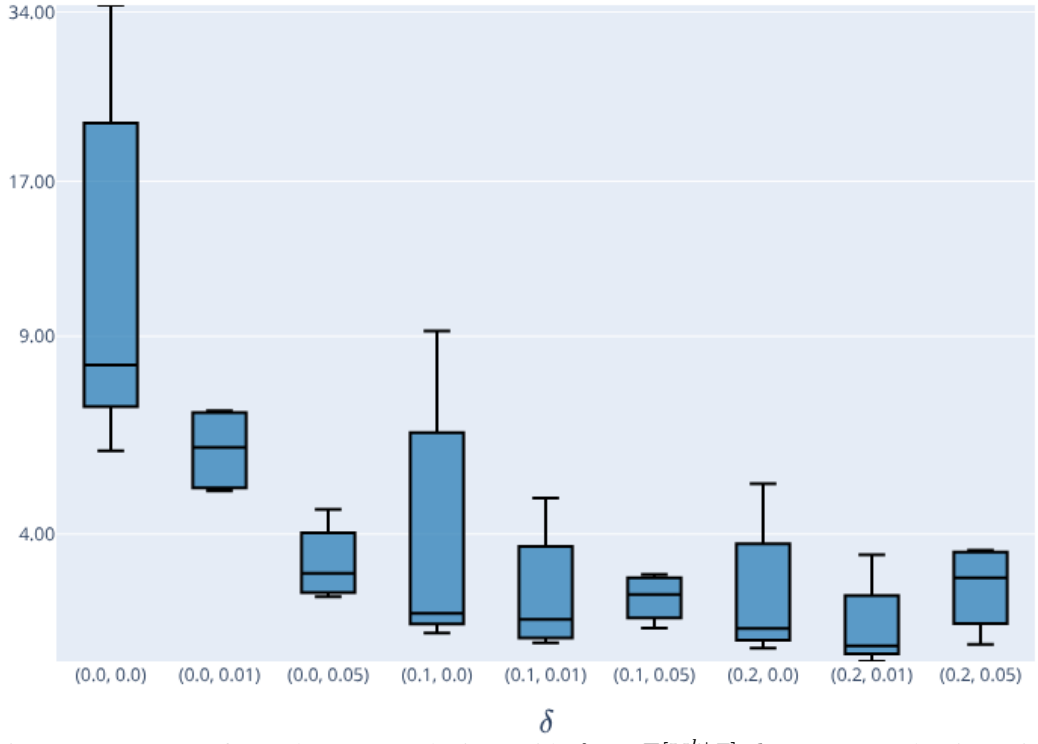


Figure 14: IV MSE for rank two perturbations with $f_k = \mathbb{E}[Y^k|Z]$, $k = 1, 2$, on dSprites with $h_0 = h_{\text{new}}$. For example, $\delta = (0.2, 0.0)$ reduces to the rank one case.