SPARSE NEURAL SIV: LEARNING NONLINEAR CAUSAL RELATIONSHIPS WITH SYNTHETIC INSTRUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider learning nonlinear causal relationships between multiple treatments and an outcome under unobserved confounding. We leverage information in the joint treatment distribution to learn these relationships without external instruments, assuming that only a sparse, unknown subset of treatments has a causal effect on the outcome. We develop a latent-factor framework under which valid instrumental variables can be derived from the treatments themselves. These synthetic instruments (SIVs) enable identification of nonlinear causal relationships through functional sparsity and a plurality rule. For estimation, we propose Sparse Neural SIV, which incorporates treatment-level sparsity into neural instrumental variable regression to jointly select active treatments and estimate the causal relationship. We fit the network by matching outcomes to predictions averaged over treatments sampled conditional on the synthetic instruments. Under suitable regularity conditions, we establish an estimation-error bound for a global minimizer of the regularized fitting criterion and give conditions for recovering the active treatments. Simulations show lower estimation error than the competing methods in the interaction settings studied under Gaussian and mixture treatment distributions. A mouse-genomics application illustrates nonlinear fitted relationships between gene expression and body weight.

1 INTRODUCTION

Learning how multiple treatments jointly affect an outcome is central to the social and biomedical sciences (Imbens & Rubin, 2015), public health (Hernán & Robins, 2006a), and machine learning (Peters et al., 2017). When measured covariates control confounding (Pearl, 2009), regression adjustment and double machine learning (Chernozhukov et al., 2018) can estimate these relationships. Unobserved common causes can invalidate this condition.

Instrumental variables (IVs) address unmeasured confounding by providing treatment variation unrelated to the confounders and affecting the outcome only through treatment (Wright, 1928; Angrist et al., 1996). Under suitable assumptions, this variation identifies causal quantities such as average treatment effects (Wang & Tchetgen Tchetgen, 2018). More generally, dose–response curves may be nonlinear and treatments may interact.

The structural model $Y = h_0(X) + u_0$ allows such relationships, with $X \in \mathbb{R}^p$ collecting the treatments and u_0 potentially correlated with X. Nonparametric IV (NPIV) methods recover h_0 from $E(u_0 \mid Z) = 0$ under completeness of the conditional treatment law (Newey & Powell, 2003), using regularization for the resulting inverse problem (Hall & Horowitz, 2005; Darolles et al., 2011; Chen & Pouzo, 2012). Neural, kernel, and minimax methods provide flexible NPIV estimators (Singh et al., 2019; Bennett et al., 2019; Dikkala et al., 2020; Xu et al., 2021; Bennett & Kallus, 2023), but typically require valid external instruments, which can be difficult to find and justify in practice (Hernán & Robins, 2006b).

An alternative uses dependence among treatments to learn about shared latent factors, as in genomics (Lin et al., 2015). A factor model separates shared and idiosyncratic treatment variation. Wang & Blei (2019) construct substitute confounders from the shared component, although treatment dependence alone does not guarantee identification (D'Amour, 2019). Miao et al. (2023) obtain identification using auxiliary variables or sufficiently many null treatments. Tang et al. (2026) instead construct synthetic instruments (SIV) from the idiosyncratic component for sparse causal estimation. Their

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

framework covers linear effects and prespecified finite-dimensional nonlinear models; unknown response shapes and interactions motivate a nonparametric extension.

Without a prespecified functional form, each candidate set of active treatments supports an infinite-dimensional class of functions. Identification requires ruling out competing functions, including those supported on different treatment subsets, that satisfy the same synthetic-IV moment equation $E\{h_0(X) \mid Z\} = E(Y \mid Z)$. Even when the causal relationship is identified, substantially different response functions can yield nearly identical conditional outcome means. This makes the inverse problem unstable and calls for regularization in estimation.

To address these identification and estimation challenges, we develop a nonparametric SIV framework for learning nonlinear causal relationships when the active treatments are unknown. Under the factor model $X = \Lambda U + \epsilon$, projecting X onto the orthogonal complement of $\mathrm{col}(\Lambda)$ removes the shared latent-factor component. The restriction $E(u_0 \mid \epsilon) = 0$ makes the resulting synthetic instruments valid (Figure 1). A functional plurality argument identifies h_0 by comparing solutions across candidate treatment sets: sets containing the true support agree on h_0, whereas incorrect solutions receive fewer votes. We give sufficient conditions in terms of the conditional treatment law and loading geometry, and show that they hold for almost every loading matrix in a Gaussian subclass.

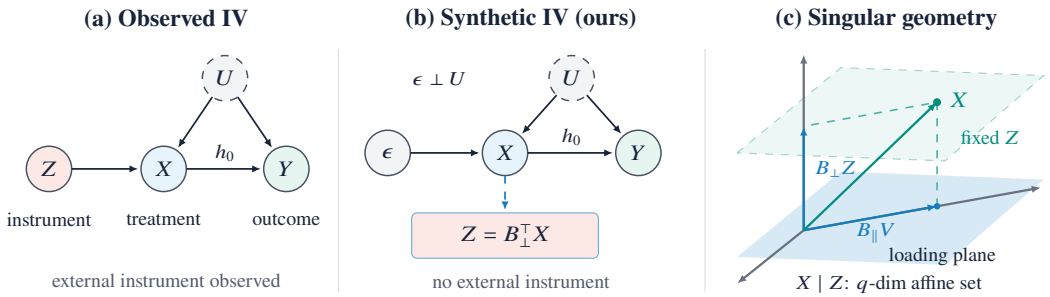


Figure 1: Conventional IV methods rely on an externally available instrument Z. Synthetic IV constructs $Z = B_\perp^\top X$ by projecting treatments onto the orthogonal complement of the factor-loading space.

For estimation, we build on neural IV regression. DeepIV integrates network predictions over a fitted conditional treatment distribution and uses independent Monte Carlo samples for unbiased gradients (Hartford et al., 2017); Regularized DeepIV adds Tikhonov regularization and statistical guarantees (Li et al., 2024). Our setting additionally requires estimating the loading space that defines the instruments, respecting the lower-dimensional support of $X \mid Z$, and selecting active treatments.

Sparse Neural SIV averages network predictions over conditional treatment draws generated in loading coordinates. A product of residuals from two independent batches removes Monte Carlo variance bias. Penalizing the squared L_2 norm of the response function regularizes the inverse problem, while limiting nonzero input-layer columns selects treatments and allows interactions among them.

Our contributions are:

(1) **Identification.** We identify nonlinear responses with unknown active treatments through functional plurality, provide primitive sufficient conditions, and establish a Gaussian genericity result.
(2) **Estimation.** We adapt regularized neural IV regression to estimated synthetic instruments through conditional sampling in loading coordinates and treatment selection via input-group sparsity.
(3) **Statistical guarantees.** For fixed treatment dimension and known sparsity level, we establish an L_2 estimation rate for a global sparse neural minimizer with unknown support, and support recovery under functional separation.

Anonymous code is available online.[1]

[1] https://anonymous.4open.science/r/neural_siv-6E95/

2 Problem Setup and Preliminaries

2.1 Nonparametric instrumental variables

We first describe the conditional-moment problem with an available instrument Z (Newey & Powell, 2003; Darolles et al., 2011). For treatments $X \in \mathbb{R}^p$ and outcome $Y \in \mathbb{R}$, consider the additively separable structural model

$$Y = h_0(X) + u_0, \qquad E(u_0 \mid Z) = 0.$$

The disturbance u_0 may be correlated with X, so h_0 need not equal $E(Y \mid X)$. The instrument supplies conditional moments that isolate the structural function h_0, which represents the causal relationships between X and Y. Figure 1(a) illustrates a standard observed-IV setting.

Define the conditional outcome regression and conditional-expectation operator by

$$r(z) = E(Y \mid Z = z), \qquad (Th)(z) = E\{h(X) \mid Z = z\}, \qquad Th_0 = r.$$

For functions of X and Z, respectively, we use $\|h\|_2^2 = E\{h(X)^2\}$ and $\|g\|_2^2 = E\{g(Z)^2\}$.

Linear versus nonparametric IV. The information needed from the first stage depends on the function class. If $h_\beta(x) = x^T\beta$, then $(Th_\beta)(z) = E(X \mid Z = z)^T\beta$, so evaluating the conditional moment requires only the finite-dimensional conditional mean $E(X \mid Z)$. For a general nonlinear function h, however, $(Th)(z) = E\{h(X) \mid Z = z\}$ cannot in general be recovered from $E(X \mid Z)$; it depends on the conditional distribution of $X \mid Z$. Thus the nonparametric problem concerns an operator T acting on an infinite-dimensional function class.

For a coordinate set S, write $T_S f(z) = E\{f(X_S) \mid Z = z\}$. Completeness requires T_S to be injective on the admissible function class: distinct functions on S must yield distinct conditional moments. Even then, inversion may be unstable, motivating regularization (Hall & Horowitz, 2005). Our setting introduces two additional challenges: constructing instruments from the treatments themselves and identifying an unknown subset of active treatments while allowing nonlinear effects and interactions. We observe only (X, Y) and use the factor model below to obtain a synthetic instrument.

2.2 Structural causal model and factor model

We observe independent draws (X_i, Y_i). The outcome model separates the structural component $h_0(X)$ from a disturbance u_0, which may depend on an unobserved confounder $U \in \mathbb{R}^q$. A factor model describes how that confounder also affects the treatments:

$$X = \Lambda U + \epsilon, \qquad Y = h_0(X) + u_0. \tag{1}$$

Here $\Lambda \in \mathbb{R}^{p\times q}$ has full column rank, and ϵ is idiosyncratic treatment noise. We assume $\epsilon \perp U$, normalize the latent factors, and impose a conditional-mean restriction on the disturbance:

$$E(U) = 0, \quad \mathrm{Cov}(U) = I_q, \quad E(u_0 \mid \epsilon) = 0.$$

All displayed variables and candidate functions are square integrable. The conditional-mean restriction implies $E(u_0) = 0$, fixing the additive normalization. Subject to this restriction, u_0 may depend arbitrarily on U: we allow $E(u_0 \mid U) \neq 0$ and correlation between u_0 and X, so unobserved confounding remains possible. Under interventions that set $X = x$ while preserving the structural disturbance u_0, mean outcome contrasts are $h_0(x) - h_0(x')$.

The factor model separates shared variation ΛU from idiosyncratic variation ϵ; classical likelihood-based estimators include Lawley (1940) and Rubin & Thayer (1982). Under the row-deletion condition on Λ in Assumption 1 (Anderson & Rubin, 1956) and diagonal positive-definite $D = \mathrm{Cov}(\epsilon)$, the loading subspace $\mathrm{col}(\Lambda)$ is identifiable. Its basis is determined only up to rotation, and the latent factors themselves need not be uniquely recovered.

3 Nonparametric Identification via Functional Plurality

3.1 Synthetic instruments from factor geometry

The factor model supplies treatment variation that is independent of the latent factors. To isolate this variation, let $B_\parallel \in \mathbb{R}^{p\times q}$ and $B_\perp \in \mathbb{R}^{p\times(p-q)}$ be orthonormal bases for the loading space and its

orthogonal complement. Define

$$V = B_{\|}^T X, \qquad Z = B_{\perp}^T X, \qquad X = B_{\|} V + B_{\perp} Z.$$

Since $B_{\perp}^T \Lambda = 0$, the projected treatments satisfy

$$Z = B_{\perp}^T X = B_{\perp}^T \epsilon \perp U.$$

Thus the projection removes the latent-factor contribution, following the synthetic-instrument construction of Tang et al. (2026). The instrument Z is a function of ϵ, so $E(u_0 \mid \epsilon) = 0$ and iterated expectation give

$$E(u_0 \mid Z) = E\{E(u_0 \mid \epsilon) \mid Z\} = 0. \tag{2}$$

Thus $Th_0 = r$ also holds for the synthetic instrument. Rotating the basis preserves the information in Z, so identifying the loading subspace suffices for this construction. Recovering h_0 from the moment equation $Th_0 = r$ requires additional restrictions on candidate functions, stated next.

3.2 Functional sparsity and plurality

A set $S \subseteq \{1, \ldots, p\}$ supports h if $h(X) = h_S(X_S)$ almost surely for some measurable h_S. We write $S(h)$ for the unique minimal support when it exists. Write $S_0 = S(h_0)$ and $s = |S_0| < p - q$. Functional sparsity limits the number of active treatments while allowing nonlinearities and interactions among them.

To compare candidate explanations of the observed moments, consider each set S of $p-q$ coordinates and solve $E\{h_S(X_S) \mid Z\} = r(Z)$. Under candidate injectivity, each solvable set has a unique solution, embedded as $h^S(x) = h_S(x_S)$. We call this solution the set's vote. Two sets agree when their embedded functions are equal almost surely, even if the sets contain different coordinates.

A nonlinear four-treatment example. Take $p = 4, q = 1$ and

$$X_j = U + \epsilon_j, \quad j = 1, \ldots, 4, \qquad Y = (X_1 + 2X_2)^2 + U + \eta,$$

with mutually independent standard normal $U, \epsilon_1, \ldots, \epsilon_4, \eta$. Here $h_0(x) = (x_1 + 2x_2)^2$ includes a treatment interaction and has support $S_0 = \{1, 2\}$. Candidates contain $p - q = 3$ treatments. To construct competing functions, let $L(x) = \sum_{j=1}^4 x_j$. Because these linear combinations induce the same conditional mean direction given Z, subtracting L or $2L$ eliminates x_1 or x_2, respectively. After squaring, constant variance corrections match their conditional moments.

Candidate set S	Embedded solution $h^S(x)$	Vote
$\{1, 2, 3\}$	$(x_1 + 2x_2)^2$	h_0
$\{1, 2, 4\}$	$(x_1 + 2x_2)^2$	h_0
$\{1, 3, 4\}$	$(-x_1 - 2x_3 - 2x_4)^2 - 20$	First incorrect function
$\{2, 3, 4\}$	$(x_2 - x_3 - x_4)^2 + 10$	Second incorrect function

The two sets containing S_0 vote for h_0, giving a $2:1:1$ pattern: plurality does not require a majority. Appendix A.4 derives the variance corrections -20 and $+10$ and proves uniqueness of each solution in $L_2(P_{X_S})$.

In general, define $\mathcal{S}_{p-q} = \{S \subseteq \{1, \ldots, p\} : |S| = p - q\}$. For $S \in \mathcal{S}_{p-q}$, let $T_S f = E\{f(X_S) \mid Z\}$, and define

$$\mathcal{S}_{\text{sol}} = \{S \in \mathcal{S}_{p-q} : r \in \text{Range}(T_S)\}.$$

Candidate sets of size $p - q$ compare solutions with potentially smaller supports. Each set containing S_0 votes for h_0, yielding $\binom{p-s}{p-q-s} \geq q + 1$ votes when $s < p - q$. Functional plurality assumes at most q votes for each incorrect function. This is similar in spirit to plurality identification with invalid instruments (Guo et al., 2018), with votes here cast for functions. With positive treatment density and injective candidate operators, this plurality condition is equivalent to sparse uniqueness (Appendix A). Sufficient conditions are given below.

$$\text{Let} \quad \mathcal{H}_{<p-q} = \{h : S(h) \text{ exists uniquely and } |S(h)| < p - q\}, \quad Q_0(h) = \|Th - r\|_2^2.$$

The following result combines factor-space identifiability and functional sparsity with candidate injectivity and plurality, as formalized in Appendix A.

Theorem 1 (Sparse synthetic-IV identification). *Under Assumptions 1–3,* $\arg\min_{h \in \mathcal{H}_{<p-q}} Q_0(h) = \{h_0\}$ *up to almost-sure equivalence.*

The theorem gives no identification guarantee for h_0 with $p - q$ or more active treatments, since it lies outside the optimization domain. Voting establishes population identification. Our estimator directly minimizes a regularized conditional-moment criterion over sparse neural networks, without enumerating candidate sets or computing votes.

When do the identification conditions hold? One sufficient route is a conditional location model $X = \Gamma Z + \eta_X$, with $\eta_X \perp Z$ and positive densities for X and Z. Identification then follows from three ingredients: the conditional-smoothing operator is injective, the relevant rows of Γ are full rank, and no incorrect support can absorb all first-stage directions along which h_0 varies. Appendix A.2 formalizes these conditions and proves that they imply candidate injectivity and plurality.

Corollary 1 (Gaussian identification). *In model* (1)*, fix a positive diagonal* D *and a bounded or polynomial-growth* h_0 *with unique minimal support* S_0*, where* $s = |S_0| < p - q$*. Suppose* $U \sim N(0, I_q)$*,* $\epsilon \sim N(0, D)$*,* $U \perp \epsilon$*,* $E(u_0 \mid \epsilon) = 0$*, and* $p \geq 2q + 1$*. Then* h_0 *is the unique solution of* $Th = r$ *in* $\mathcal{H}_{<p-q}$ *for Lebesgue-almost every loading matrix* $\Lambda \in \mathbb{R}^{p \times q}$*.*

The exceptional loadings satisfy rank or response-alignment equations; Appendix A.3 proves that they form a measure-zero set.

4 Sparse Neural Synthetic-IV Method

Sparse Neural SIV learns causal relationships by combining a fitted conditional treatment sampler with a sparse neural network. Figure 2 summarizes the sampling and estimation steps.

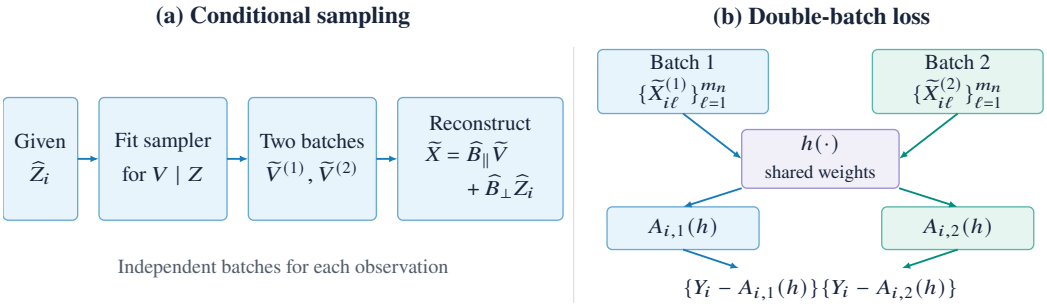


Figure 2: Conditional sampling and double-Monte-Carlo estimation. Two independent batches of reconstructed treatments pass through a shared network; multiplying their residuals removes the single-batch variance term.

4.1 Geometry-aware conditional sampling

For an estimated loading basis, define

$$\widehat{V} = \widehat{B}_\parallel^T X, \qquad \widehat{Z} = \widehat{B}_\perp^T X. \tag{3}$$

Conditioning on $\widehat{Z} = z$ fixes $p - q$ linear combinations of the treatments. The remaining variation lies in the affine set $\{\, \widehat{B}_\parallel v + \widehat{B}_\perp z : v \in \mathbb{R}^q \,\}$, so the conditional law has no density with respect to p-dimensional Lebesgue measure. We therefore fit $\widehat{P}_V(\cdot \mid z)$ in the q-dimensional loading coordinates and reconstruct treatment draws. This defines the fitted operator

$$(\widehat{T}_{\widehat{B}} h)(z) = E_{\widetilde{V} \sim \widehat{P}_V(\cdot \mid z)} \left[h(\widehat{B}_\parallel \widetilde{V} + \widehat{B}_\perp z) \right]. \tag{4}$$

Evaluating (4) requires only draws from the fitted law. The algorithm can therefore use different conditional generative models. Statistical guarantees for a given sampler require the accuracy and moment bounds in Assumption 6.

We call the estimated bases $(\widehat{B}_{\|}, \widehat{B}_{\perp})$ and the fitted sampler $\widehat{P}_V$ the *nuisance estimates*. The statistical analysis computes them from a sample $\mathcal{D}^N$ of size n_N and estimates the structural function on an independent sample $\mathcal{D}^T$ of size n_T. In the experiments, nuisance estimation and structural fitting use the same training data (Appendix G).

4.2 Double-Monte-Carlo structural criterion

The moment condition $Th_0 = r$ motivates minimizing $\|Th - r\|_2^2$. Although $r(z) = E(Y \mid Z = z)$ is unknown, squared error against the observed outcome provides an equivalent population objective:

$$E[\{Y - (Th)(Z)\}^2] = \|Th - r\|_2^2 + E\{\mathrm{Var}(Y \mid Z)\}.$$

The last term does not depend on h, so minimizing the left-hand side targets the same conditional-moment fit without a separate estimator of r. We regularize this problem with the Tikhonov criterion

$$Q_\alpha(h) = \|Th - r\|_2^2 + \alpha\|h\|_2^2. \tag{5}$$

Fix h, the fitted nuisance estimates, and an observation (X_i, Y_i). Define the fitted conditional mean and its Monte Carlo approximation by

$$\mu_i(h) = (\widehat{T}_{\widehat{B}}h)(\widehat{Z}_i), \qquad A_{i,b}(h) = \frac{1}{m_n}\sum_{\ell=1}^{m_n} h(\widetilde{X}_{i\ell}^{(b)}),$$

where each batch contains i.i.d. treatment draws reconstructed from $\widehat{P}_V(\cdot \mid \widehat{Z}_i)$, as in Algorithm 1. Then $E_{\mathrm{MC}}[A_{i,b}(h)] = \mu_i(h)$, but squaring a single-batch residual gives

$$E_{\mathrm{MC}}[\{Y_i - A_{i,1}(h)\}^2] = \{Y_i - \mu_i(h)\}^2 + \mathrm{Var}_{\mathrm{MC}}\{A_{i,1}(h)\},$$

with

$$\mathrm{Var}_{\mathrm{MC}}\{A_{i,1}(h)\} = \frac{1}{m_n}\mathrm{Var}_{\widehat{P}_V}\{h(\widetilde{X}) \mid \widehat{Z}_i\}.$$

This additional term depends on h and changes the expected objective. Using two independent batches instead yields

$$E_{\mathrm{MC}}[\{Y_i - A_{i,1}(h)\}\{Y_i - A_{i,2}(h)\}] = \{Y_i - \mu_i(h)\}^2. \tag{6}$$

Multiplying residuals from the two independent Monte Carlo batches removes the simulation-induced variance term, while first-stage estimation error remains. Algorithm 1 gives the resulting empirical objective with its Tikhonov penalty.

4.3 Sparse neural estimator and selection of k

We impose treatment-level sparsity by allowing at most k nonzero columns in the first-layer weight matrix W_1: $\sum_{j=1}^{p} \mathbf{1}\{\|(W_1)_{\cdot j}\|_2 > 0\} \le k$. Column j contains all weights from treatment X_j to the first hidden layer, so zeroing it removes every path from that treatment to the output. The retained treatments can interact through subsequent layers. Our truncated ReLU class $\mathcal{H}_{k,n} = \mathcal{G}(D_n, p, 1, N_n, M, B_n, k)$ has D_n hidden layers of common width N_n, scalar output truncated at M, and weights and biases bounded by B_n. A nonzero column need not belong to the minimal functional support; it provides only a computational input screen.

For each candidate k, the implementation first trains for 80 epochs with group shrinkage targeting $\lambda_g \sum_{j=1}^{p} \|(W_1)_{\cdot j}\|_2$. It then retains the k columns with largest Euclidean norms, zeros the others, and refits for 80 epochs with this mask fixed and the group penalty removed. Tikhonov and numerical weight penalties remain in both phases (Appendix G). Cross-validation selects k by held-out double-Monte-Carlo loss, followed by a full-data refit (Appendix G.6).

Algorithm 1 Sparse Neural SIV with CV-selected sparsity

Require: Training data $\mathcal{D}_n$; factor dimension q; candidate budgets $\mathcal{K} \subset \{1, \dots, p-q-1\}$ and network classes $\{\mathcal{H}_{k,n} : k \in \mathcal{K}\}$; $\alpha_n > 0$; m_n training draws per observation per Monte Carlo batch; F CV folds and m_{val} validation draws per observation per batch.

1: Select $\widehat{k}$ by F-fold cross-validation of the double-Monte-Carlo validation loss (Appendix G.6).
2: On all of $\mathcal{D}_n$, refit input normalization and estimate $\widehat{B}_{\|}, \widehat{B}_{\perp}$, form the pairs in (3), and fit any conditional sampler $\widehat{P}_V(\cdot \mid z)$. Set $\mathcal{D}^T = \mathcal{D}_n$, $n_T = n$ for the final structural fit.
3: **for** each $i \in \mathcal{D}^T$ and $b \in \{1, 2\}$ **do**
4: Independently draw $\widetilde{V}_{i\ell}^{(b)} \sim \widehat{P}_V(\cdot \mid \widehat{Z}_i)$, $\ell = 1, \dots, m_n$, and reconstruct

$$\widetilde{X}_{i\ell}^{(b)} = \widehat{B}_{\|} \widetilde{V}_{i\ell}^{(b)} + \widehat{B}_{\perp} \widehat{Z}_i.$$

5: **end for**
6: Set

$$A_{i,b}(h) = \frac{1}{m_n} \sum_{\ell=1}^{m_n} h(\widetilde{X}_{i\ell}^{(b)})$$

and train a new sparse network $\widehat{h} \in \mathcal{H}_{\widehat{k},n}$ to minimize $\widehat{Q}_{\alpha_n}$, where

$$\widehat{Q}_{\alpha}(h) = \frac{1}{n_T} \sum_{i \in \mathcal{D}^T} \{Y_i - A_{i,1}(h)\}\{Y_i - A_{i,2}(h)\} + \alpha \frac{1}{n_T} \sum_{i \in \mathcal{D}^T} h(X_i)^2. \tag{7}$$

4.4 Statistical guarantee

The guarantee holds for fixed p, q, k, s, while the neural sieve may grow with n_T. Appendix B states the required loading-space and sampler accuracy, moment bounds, and neural approximation conditions. Restricted regularized curvature converts objective differences into estimation error across supports; it is an additional population restriction beyond completeness and plurality.

For the guarantee, replace CV selection by the fixed budget $k = s$ and fit nuisance estimates independently of $\mathcal{D}^T$. The analyzed estimator is a measurable global minimizer of (7),

$$\widehat{h} \in \operatorname*{arg\,min}_{h \in \mathcal{H}_{s,n}} \widehat{Q}_{\alpha_n}(h).$$

With bounded parameters and fixed architecture at each n, compactness and continuity ensure existence; choose any measurable minimizer. Define

$$d_n^2 = \rho_n^2 + n_N^{-1} + \delta_{F,n}^2 + \xi_{\text{MC},n}^2. \tag{8}$$

Theorem 2 (Estimation rate of Sparse Neural SIV)**.** *Suppose Assumptions 1, 2, and 4–8 hold, with known $k = s = |S_0|$ and unknown S_0. For $0 < \alpha_n \le 1$, with $\nu = \min\{\beta_{\text{iv}}, 2\}$,*

$$\|\widehat{h} - h_0\|_2^2 = O_p\left\{ \frac{d_n^2}{\alpha_n^2} + \frac{\epsilon_{H,n}^2}{\alpha_n} + \alpha_n^{\nu} \right\}. \tag{9}$$

Thus consistency follows if $\alpha_n \to 0$, $d_n/\alpha_n \to 0$, and $\epsilon_{H,n}^2/\alpha_n \to 0$.

The bound shows how regularization balances estimation error and bias. The quantity d_n^2 combines neural empirical complexity ρ_n^2, loading-space error n_N^{-1}, sampler error $\delta_{F,n}^2$, and Monte Carlo error $\xi_{\text{MC},n}^2$. Inverting the conditional operator amplifies these errors by α_n^{-2}. Neural approximation contributes $\epsilon_{H,n}^2/\alpha_n$: $\epsilon_{H,n}$ controls both approximation of h_0 and excess regularized loss on the true support. The source exponent β_{iv} measures smoothness relative to the conditional operator, giving Tikhonov bias α_n^{ν} (Assumption 7). Appendix D specifies ρ_n and $\xi_{\text{MC},n}$ in terms of network size, sample size, and Monte Carlo draws. Root-n_N loading-space estimation therefore does not imply a parametric rate for h_0. Functional-support recovery under separation is given in Appendix F.

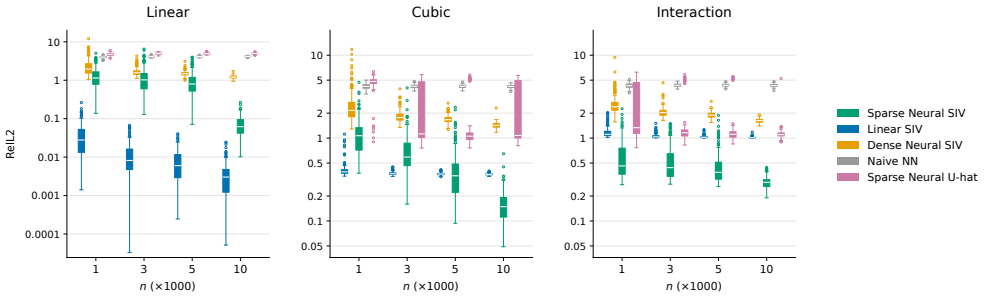


Figure 3: Relative squared structural error (RelL2; lower is better). Each boxplot summarizes 200 replications. The y-axes are shown on a logarithmic scale.

5 Experiments

5.1 Synthetic Experiments

We examine how modeling nonlinear causal relationships and selecting treatments affect structural estimation. Five methods are compared under linear, cubic, and interacting treatment effects.

Experimental setup. We fix $p = 10$, $q = 2$, $S_0 = \{1, 2\}$, and generate

$$X = \Lambda U + \epsilon, \qquad Y = h_0(X) + u_0, \qquad u_0 = 1.5U_1\epsilon_1 + 0.5\eta,$$

with independent $U \sim N(0, I_2)$, $\epsilon \sim N(0, D)$, and $\eta \sim N(0, 1)$. By independence and zero means, $E(u_0 \mid \epsilon) = 0$, and hence $E(u_0 \mid Z) = 0$. The product $U_1\epsilon_1$ provides a stress test for additive factor-score adjustment. We center and scale three structural functions:

$$f_{\text{Linear}}(x) = x_1 + x_2,$$
$$f_{\text{Cubic}}(x) = 0.3x_1^3 + 0.3x_2^3,$$
$$f_{\text{Interaction}}(x) = \sin(x_1x_2) + 0.5\cos(x_1) + 0.25x_2^2.$$

Exact calibration constants and population diagnostics are reported in Appendix G.

We compare *Sparse Neural SIV* with four alternatives. *Dense Neural SIV* uses the same architecture without the sparsity constraint; *Linear SIV* uses a sparse linear structural function (Tang et al., 2026). *Naive NN* fits outcomes by ordinary squared-error loss. *Sparse Neural $\widehat{U}$* uses additive factor-score adjustment motivated by Wang & Blei (2019). All neural methods use two width-64 hidden layers and 160 training epochs. Sparse Neural SIV selects $k \in \{1, 2, 3\}$ by CV at all four sample sizes. The other support-selecting methods use fixed $k = 2$ throughout; Appendix G gives implementation details.

For $n \in \{1000, 3000, 5000, 10000\}$, each scenario has 200 replications with 20,000 independent test observations. We evaluate relative squared L_2 error,

$$\text{RelL2}_r = \frac{\sum_t \{\widehat{h}_r(X_{rt}^{\text{test}}) - h_0(X_{rt}^{\text{test}})\}^2}{\sum_t \{h_0(X_{rt}^{\text{test}}) - \bar{h}_{0,r}^{\text{test}}\}^2},$$

where $\bar{h}_{0,r}^{\text{test}}$ is the test mean of h_0. A value of one equals the error of predicting this constant test mean. Figure 3 displays replication-level distributions.

Figure 3 highlights three patterns. Linear SIV performs best under correct linear specification. Sparse Neural SIV improves with sample size in the nonlinear settings and has the lowest mean error in every Interaction setting. At $n = 10{,}000$, its mean Interaction RelL2 is 0.292, versus 1.126 for Sparse Neural $\widehat{U}$. Correct input selection does not ensure accurate structural estimation: $\widehat{U}$ selects exactly $\{1, 2\}$ more often (99.5% versus 45.5%) despite its larger error. These percentages describe fitted input masks; the comparator fixes $k = 2$, while Sparse Neural SIV selects k by CV. Appendix G.8 reports robustness results under additive latent confounding and non-Gaussian mixture treatment noise.

5.2 Application to Mouse Obesity Genomics

We analyze the public 287-mouse, 17-gene version of the obesity data (Wang et al., 2006; Lin et al., 2015), using liver gene expression as treatments, body weight as the outcome, and $q = 1$ as in Xie et al. (2024). Prior analyses include Bayesian partial identification (Zheng et al., 2022) and linear proxy-variable selection (Xie et al., 2024). We refit Linear SIV and Sparse Neural SIV on this common dataset, sharing the loading space, Gaussian conditional model, folds, and validation draws. Both select $k \in \{1, \ldots, 6\}$ by five-fold double-Monte-Carlo validation, with normalization and nuisance fitting restricted to training folds (Appendix G.7).

CV-selected supports. Linear SIV selects $k = 1$, retaining *Gstm2*, whereas Sparse Neural SIV selects $k = 2$, retaining *Irx3* and *Dscam*. The fitted changes from the 5th to the 95th expression percentile, averaging over mice with other genes held at their observed values, are 3.57 for *Gstm2*, -1.92 for *Irx3*, and 2.60 for *Dscam*. The two fits differ in both model specification and sparsity budget, so the comparison combines these choices. The minimum CV losses, in squared body-weight units, are 33.61 for Linear SIV and 34.71 for Sparse Neural SIV, so this criterion does not favor the neural model. The neural losses at $k = 1$ and $k = 2$ are close (34.85 versus 34.71); the selected pair should not be interpreted as a precise count of causal genes.

Estimated treatment–outcome relationships. Figure 4 shows the fitted body-weight contrasts. For *Irx3*, the contrast rises slightly at low expression, then declines and levels off. For *Dscam*, it increases over most of the range, with a modest downturn at high expression. These patterns highlight the nonlinear relationships between treatment levels and the outcome.

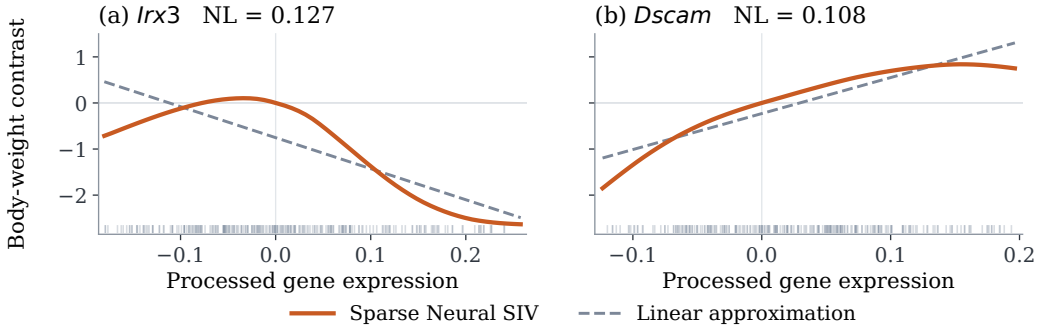


Figure 4: Sparse Neural SIV contrasts (CV-selected $k = 2$) in original body-weight units, averaging $\widehat{h}(t, X_{i,-j}) - \widehat{h}(0, X_{i,-j})$ over mice. Expression remains on the public dataset's processed scale, with zero on that scale as the reference. Dashed lines are affine fits over 61 equally spaced points between the 5th and 95th percentiles. NL measures relative departure from these lines (Appendix G.7); rugs show observations.

These fitted curves illustrate nonlinear patterns in the data, with causal interpretation subject to the identification assumptions in Section 3.

6 Conclusion

Sparse Neural SIV learns nonlinear causal relationships without external instruments. Under a latent-factor model and additional assumptions, functional sparsity, completeness, and plurality identify these relationships with unknown active treatments. Conditional sampling in loading coordinates and double Monte Carlo enable neural estimation. For a global minimizer with known sparsity level, we establish an estimation rate and support recovery under functional separation. Simulations and mouse-genomics data assess estimation and treatment selection. Guarantees for CV selection and same-sample numerical fitting remain open.

AI USE STATEMENT

Generative AI tools were used to assist with writing and polishing, literature retrieval and discovery, drafting sections of the manuscript, and developing mathematical proofs. They were also used to restructure the manuscript, migrate LaTeX and bibliography files, and check compilation and layout.

REPRODUCIBILITY STATEMENT

Algorithm 1 describes the CV-selected procedure. Section 4.3 explains the sparsity budget, and Section 4.4 defines the fixed-budget estimator analyzed in the theory. Appendices A–F provide formal conditions, proofs, and the Gaussian sampler verification. Appendix G records the simulation calibration, loading matrix, seeds, hyperparameters, implementation details, numerical tables, and limitations of the reported implementation. The anonymous code repository linked with this submission contains the main Gaussian simulations, the mixture robustness experiment, and the mouse-data analysis, including scripts for generating the reported tables and figures.

REFERENCES

Theodore W. Anderson and Yasuo Amemiya. The Asymptotic Normal Distribution of Estimators in Factor Analysis under General Conditions. *The Annals of Statistics*, 16(2):759 – 771, 1988. doi: 10.1214/aos/1176350834. URL https://doi.org/10.1214/aos/1176350834.

Theodore W. Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 5, pp. 111–150, 1956.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL https://doi.org/10.1080/01621459.1996.10476902.

Andrew Bennett and Nathan Kallus. The variational method of moments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):810–841, 2023. doi: 10.1093/jrsssb/qkad025.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

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

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3478–3486. PMLR, 2019. URL https://proceedings.mlr.press/v89/d-amour19a.html.

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. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.07201.

Zijian Guo, Hyunseung Kang, T. Tony Cai, and Dylan S. Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):793–815, 2018. doi: 10.1111/rssb.12275.

Peter Hall and Joel L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The 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 *Proceedings of the 34th International Conference on Machine Learning*, pp. 1414–1423, 2017.

Miguel A. Hernán and James M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586, 2006a. doi: 10.1136/jech.2004.029496.

Miguel A. Hernán and James M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006b. doi: 10.1097/01.ede.0000222409.00878.37.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Derrick N. Lawley. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60(1):64–82, 1940. doi: 10.1017/S037016460002006X.

Zihao Li, Hui Lan, Vasilis Syrgkanis, Mengdi Wang, and Masatoshi Uehara. Regularized DeepIV with model selection, 2024. URL https://arxiv.org/abs/2403.04236. arXiv:2403.04236.

Wei Lin, Rui Feng, and Hongzhe Li. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288, January 2015. ISSN 1537-274X. doi: 10.1080/01621459.2014.908125. URL http://dx.doi.org/10.1080/01621459.2014.908125.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Wang Miao, Wenjie Hu, Elizabeth L. Ogburn, and Xiao-Hua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, 118(543):1953–1967, 2023. doi: 10.1080/01621459.2021.2023551. URL https://doi.org/10.1080/01621459.2021.2023551.

Boris Mityagin. The Zero Set of a Real Analytic Function, 2015. URL https://arxiv.org/abs/1512.07276. arXiv:1512.07276.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

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

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. URL https://bayes.cs.ucla.edu/BOOK-2K/.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. URL https://is.mpg.de/publications/petjansch17.

Donald B. Rubin and Dorothy T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47 (1):69–76, 1982. doi: 10.1007/BF02293851.

Cyrill Scheidegger, Zijian Guo, and Peter Bühlmann. Spectral deconfounding for high-dimensional sparse additive models. *ACM/IMS Journal of Data Science*, 2(1):1–52, 2025. doi: 10.1145/3711116. Article 1.

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

Dingke Tang, Dehan Kong, and Linbo Wang. The synthetic instrument: From sparse association to sparse causation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(4): 1210–1230, 2026. doi: 10.1093/jrsssb/qkaf083.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):531–550, 2018. doi: 10.1111/rssb.12262.

Susanna Wang, Nadir Yehya, Eric E. Schadt, Hui Wang, Thomas A. Drake, and Aldons J. Lusis. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genetics*, 2(2):e15, February 2006. ISSN 1553-7404. doi: 10.1371/journal.pgen.0020015. URL http://dx.doi.org/10.1371/journal.pgen.0020015.

Yixin Wang and David M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019. doi: 10.1080/01621459.2019.1686987.

Philip G. Wright. *The Tariff on Animal and Vegetable Oils*. The Macmillan Company, New York, 1928. Appendix B, pp. 286–319.

Feng Xie, Zhengming Chen, Shanshan Luo, Wang Miao, Ruichu Cai, and Zhi Geng. Automating the selection of proxy variables of unmeasured confounders. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54430–54459. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xie24b.html.

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. URL https://openreview.net/forum?id=sy4Kg_ZQmS7.

Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Bayesian inference and partial identification in multi-treatment causal inference with unobserved confounding. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3608–3626. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/zheng22a.html.

A Formal Identification Conditions and Proofs

A.1 High-level identification conditions and voting characterization

We state the assumptions behind Theorem 1 and prove the functional plurality argument. We use the structural model from Section 2 and the instruments, supports, and candidate solutions from Section 3.

A sufficient disturbance decomposition. An additive latent disturbance provides one way to satisfy the model's conditional-mean restriction. Consider the decomposition $u_0 = g_0(U) + \eta$. If $\epsilon \perp U$, $E\{g_0(U)\} = 0$, and $E(\eta \mid X, U) = 0$, then $\epsilon = X - \Lambda U$ is measurable with respect to (X, U), and iterated expectation gives

$$E(u_0 \mid \epsilon) = E\{g_0(U)\} + E\{E(\eta \mid X, U) \mid \epsilon\} = 0.$$

The general model does not require this decomposition.

Additive separation of the structural function and disturbance places no additivity restriction on the treatments themselves. The function h_0 can be nonlinear and include treatment interactions. A related nonlinear model is studied by Scheidegger et al. (2025), who combine a factor model for the covariates with a coordinate-additive structural function and a linear confounding term. Here h_0 is nonparametric and is not restricted to a sum of univariate functions, while u_0 is constrained by $E(u_0 \mid \epsilon) = 0$. Identification for our model requires the factor-space, sparsity, and candidate-solution conditions stated below.

The admissible class attached to a coordinate set S is $L_2(P_{X_S})$, including functions that depend on a subset of S. Injectivity of T_S means that $E\{g(X_S) \mid Z\} = 0$ almost surely for $g \in L_2(P_{X_S})$ implies $g(X_S) = 0$ almost surely. All function equalities are understood almost surely.

Assumption 1 (Factor-space identifiability)**.** The idiosyncratic covariance $D = \mathrm{Cov}(\epsilon)$ is diagonal and positive definite, $p \geq 2q + 1$, and deleting any row of Λ leaves two disjoint full-rank $q \times q$ submatrices. $\mathrm{col}(\Lambda)$ is identified, and a basis is identified up to an orthogonal rotation.

This classical row-deletion condition (Anderson & Rubin, 1956) is used to identify the loading space.

Assumption 2 (Functional sparsity)**.** The structural function has a unique minimal support $S_0 = S(h_0)$, with $s = |S_0| < p - q$.

Assumption 3 (Solvable-candidate plurality)**.** Each operator T_S with $S \in \mathcal{S}_{\mathrm{sol}}$ is injective on its admissible candidate class, and no incorrect function receives more than q votes:

$$\max_{h \neq h_0} \left|\{S \in \mathcal{S}_{\mathrm{sol}} : h^S = h\}\right| \leq q.$$

Voting characterization under support regularity. Suppose Assumption 2 and $r = Th_0$ hold, X has a density positive Lebesgue-almost everywhere, $q \geq 1$, and every solvable candidate operator is injective on $L_2(P_{X_S})$. Positive density ensures that a function supported on two coordinate sets is also supported on their intersection (by Fubini's theorem), so each candidate solution has a unique minimal support. Under these conditions, the plurality bound in Assumption 3 holds if and only if h_0 is the unique solution of $Th = r$ in $\mathcal{H}_{<p-q}$. Indeed, a competing solution with support M, $|M| < p - q$, receives at least $\binom{p-|M|}{q} \geq q + 1$ votes from its supersets, since $p - |M| \geq q + 1$. Conversely, if a wrong candidate solution had support smaller than $p - q$, it would be another sparse solution. Thus every wrong candidate solution has minimal support of size $p-q$ and can be supported, hence voted for, by only its own candidate set. It receives one vote, which is at most q.

A.2 Primitive sufficient conditions for sparse identification

Put $d = p - q$. Consider the following conditional-location subclass and conditions. The location residual η_X is distinct from the outcome noise η used in the example below.

(P0) Conditional location and support regularity.
$X = \Gamma Z + \eta_X$, where $\eta_X \perp Z$ and $\Gamma \in \mathbb{R}^{p \times d}$. Both X and Z have densities positive Lebesgue-almost everywhere.

(P1) Injective conditional smoothing.
For $M \subseteq \{1, \dots, p\}$, $|M| \le d$, define $C_M f(w) = E\{f(w + \eta_{X,M})\}$. The smoothing is finite almost everywhere and injective on $L_2(P_{X_M})$: $C_M f = 0$ almost everywhere implies $f = 0$ almost everywhere. The same implication holds for every translation difference $h_0(\cdot + ta) - h_0$, $a \in \mathbb{R}^s$, $t \in \mathbb{R}$; these smoothings are assumed finite almost everywhere. Here and below, almost everywhere refers to Lebesgue measure, equivalently to the relevant marginal of X.

(P2) Full-spark first stage.
Every Γ_M with $|M| \le d$ has row rank $|M|$.

(P3) Response-design no-aliasing.
For every $M \not\supseteq S_0$ with $|M| < d$, $\Gamma_{S_0} \ker(\Gamma_M) \not\subseteq \mathcal{N}_0$, where

$$\mathcal{N}_0 = \{a \in \mathbb{R}^s : h_0(x + ta) = h_0(x) \text{ for Lebesgue-a.e. } x \text{ and every } t \in \mathbb{R}\}.$$

The subspace $\mathcal{N}_0$ records translation directions invisible to the true response. Condition (P3) excludes an incorrect support whose unobserved first-stage directions are all invisible to h_0. (P0) specifies the conditional treatment law; (P1) separately requires injectivity on the full admissible L_2 candidate classes.

Proposition A.1 (Primitive sparse identification)**.** *Under Assumptions 1 and 2, the moment restriction $r = Th_0$, and (P0)–(P3), h_0 is the unique solution of $Th = r$ in $\mathcal{H}_{<d}$. Every candidate operator T_S, $|S| = d$, is injective, so Assumption 3 follows and the conclusion of Theorem 1 holds.*

Proof. Write $G_M = C_M h_M$, so (P0) gives $T_M h_M(z) = G_M(\Gamma_M z)$. Suppose $h \in \mathcal{H}_{<d}$ solves $Th = r$ and put $M = S(h)$. For any $v \in \ker(\Gamma_M)$ and $t \in \mathbb{R}$, the equality $G_M(\Gamma_M(z + tv)) = G_M(\Gamma_M z)$ shows that $r(z + tv) = r(z)$ for almost every z, for each fixed v, t. By (P2) and $s < d$, $\operatorname{rank}(\Gamma_{S_0}) = s$, so Γ_{S_0} maps onto $\mathbb{R}^s$. Since $r(z) = C_{S_0} h_0(\Gamma_{S_0} z)$, positivity of the density of Z and this surjectivity imply that $C_{S_0} h_0(w + t\Gamma_{S_0} v) = C_{S_0} h_0(w)$ almost everywhere in w. Applying (P1) to the translation difference yields $\Gamma_{S_0} v \in \mathcal{N}_0$. Thus $\Gamma_{S_0} \ker(\Gamma_M) \subseteq \mathcal{N}_0$, and (P3) forces $M \supseteq S_0$.

Now view both h and h_0 as functions of X_M. Their smoothed difference vanishes on $\Gamma_M z$ almost everywhere in z. By (P2), Γ_M is surjective, so (P1) gives $h = h_0$ almost surely. The same argument with $|M| = d$ proves injectivity of every candidate operator. Finally, (P0) supplies the support regularity in the voting characterization above, which gives Assumption 3. □

No invariant directions. If $\mathcal{N}_0 = \{0\}$, (P2) implies (P3): for $j \in S_0 \setminus M$, the row Γ_j is outside the span of Γ_M, since $|M| + 1 \le d$. Some $v \in \ker(\Gamma_M)$ therefore has $\Gamma_j v \neq 0$. Consequently, (P0)–(P2) alone give sparse identification for such a response. This condition is not automatic for nonlinear functions: a nonconstant ridge response $h_0(x) = f(x_1 + x_2)$ is invariant along $(1, -1)$, so the general case still needs (P3).

Linear responses. For $h_0(x) = \beta_{S_0}^T x_{S_0}$, $\mathcal{N}_0 = \ker(\beta_{S_0}^T)$, and

$$\Gamma_{S_0} \ker(\Gamma_M) \subseteq \mathcal{N}_0 \quad \Longleftrightarrow \quad \beta_{S_0}^T \Gamma_{S_0} \in \operatorname{rowspan}(\Gamma_M).$$

Thus (P3) rules out an exact alignment between the effective structural direction and an incorrect support, as in linear SIV.

A.3 Gaussian idiosyncratic noise and loading genericity

Suppose $\epsilon \sim N(0, D)$, with D positive diagonal, and $U \perp \epsilon$ in the factor model (1); U need not be Gaussian. Since $Z = B_\perp^T \epsilon$, let

$$\Gamma = \operatorname{Cov}(X, Z) \operatorname{Cov}(Z)^{-1} = DB_\perp (B_\perp^T D B_\perp)^{-1}.$$

Put $G = \epsilon - \Gamma B_\perp^T \epsilon$. Then $\operatorname{Cov}(G, Z) = DB_\perp - \Gamma B_\perp^T D B_\perp = 0$. The jointly Gaussian vectors G and Z are therefore independent; $U \perp \epsilon$ also gives $U \perp (G, Z)$. Hence $X = \Gamma Z + \eta_X$ with $\eta_X = \Lambda U + G \perp Z$. The nondegenerate Gaussian ϵ makes the density of X positive everywhere, while Z is Gaussian with covariance $B_\perp^T D B_\perp \succ 0$. Thus (P0) holds without Gaussian U. Positive diagonal row scaling and invertible right multiplication preserve row dependencies; thus Γ is full spark exactly when $B_\perp$ is. For any $|S| = d$, the rows $(B_\perp)_S$ are dependent exactly when a nonzero

vector supported on S lies in $\ker(B_\perp^T) = \mathrm{col}(\Lambda)$, equivalently when the complementary $q \times q$ row minor Λ_{S^c} is singular. Hence (P2) is equivalent to nonsingularity of every such minor.

With arbitrary U, the residual characteristic function factors as

$$\varphi_{\eta_{X,M}}(t) = \varphi_U(\Lambda_M^T t) \exp\{-t^T \operatorname{Cov}(G_M)t/2\}.$$

Even a smooth, nowhere-zero characteristic function does not by itself verify (P1) on the full $L_2(P_{X_M})$ candidate class. If U has moments of every order and φ_U has no real zeros, Fourier division applies to tempered translation differences of a polynomial-growth h_0; candidate-class injectivity still requires a separate argument. We therefore retain (P1) for non-Gaussian U.

When U is also Gaussian, the full-spark condition verifies (P1) on the candidate L_2 classes. Fix $|M| \le d$ and let $R_M = \operatorname{Cov}(Z, X_M)\operatorname{Cov}(X_M)^{-1}$. Because $\operatorname{Cov}(Z, X_M) = \operatorname{Cov}(Z)\Gamma_M^T$, R_M has full column rank. If $C_M f = 0$, then $T_M f = 0$, and Gaussian conditioning gives, for all $t \in \mathbb{R}^d$,

$$0 = E\{f(X_M)e^{\mathrm{i}t^T Z}\} = e^{-t^T \operatorname{Cov}(Z|X_M)t/2} E\{f(X_M)e^{\mathrm{i}t^T R_M X_M}\}.$$

The exponential never vanishes and R_M^T maps onto $\mathbb{R}^{|M|}$. Fourier uniqueness for the finite signed measure $f\, dP_{X_M}$ gives $f = 0$ almost surely. If h_0 and all its translates belong to $L_2(P_{X_{S_0}})$, the same argument covers the translation differences in (P1); bounded or polynomial-growth responses satisfy this requirement.

Proof of Corollary 1. Fix h_0 with unique minimal support and choose a matrix A_0 with $\ker A_0 = \mathcal{N}_0$. On the full-spark locus, (P3) fails for an incorrect M, $|M| < d$, exactly when

$$\operatorname{rank}\begin{pmatrix} \Gamma_M \\ A_0\Gamma_{S_0} \end{pmatrix} = |M|.$$

This is the vanishing of all $(|M|+1)$-row minors of the stacked matrix. Because S_0 is minimal and X has positive density, no coordinate direction e_j, $j \in S_0$, belongs to $\mathcal{N}_0$. For $j \in S_0 \setminus M$, some row of A_0 has a nonzero j-th entry. At a witness matrix, choose Γ_M independent, Γ_j outside its row span, and the other rows in $S_0 \setminus M$ to be zero. An augmented minor is then nonzero; a full-spark perturbation preserves this property.

The condition is also proper in the loading coordinates. Indeed, $\mathrm{col}(\Gamma) = \ker(\Lambda^T D^{-1})$. Extend that row of A_0 to a vector $a \in \mathbb{R}^p$ supported on S_0. Failure of (P3) gives $a - b \in \ker(\Gamma^T) = \mathrm{col}(D^{-1}\Lambda)$ for some b supported on M, and hence $a_{M^c} \in \mathrm{col}\{(D^{-1}\Lambda)_{M^c}\}$. Since $a_{M^c} \neq 0$ and $p - |M| \ge q+1$, this imposes a nontrivial $(q+1)$-minor equation in Λ. Full spark likewise fails on the zero sets of the $q \times q$ row minors of Λ. For fixed positive diagonal D, both (P2) and (P3) therefore hold almost surely when Λ has a joint density. The same holds for a real-analytic loading parameterization on a connected open domain with an absolutely continuous parameter law, provided every required full-spark minor and, for each incorrect M, at least one of its augmented minors is not identically zero (Mityagin, 2015).

Thus the loading geometry (P2)–(P3) is generic for a fixed h_0, even when $\mathcal{N}_0 \neq \{0\}$. With Gaussian U and the stated translation integrability, (P1) also holds, so Assumption 3 follows generically (with $p \ge 2q+1$). For non-Gaussian U, the same conclusion requires separate verification of (P1). Mere continuous dependence of Λ on parameters does not imply a measure-zero exceptional set.

A.4 Verification of the nonlinear four-treatment example

We verify the identification conditions for the example in Section 3.2. Let $\mathbf{1} \in \mathbb{R}^4$ be the vector of ones. Here $\Lambda = \mathbf{1}$, $D = I_4$, and $\operatorname{Cov}(X) = I_4 + \mathbf{1}\mathbf{1}^\top$. The row-deletion condition holds because every loading is nonzero, and $u_0 = U + \eta$ satisfies $E(u_0 \mid \epsilon) = 0$. Since X has a positive density on $\mathbb{R}^4$, $h_0(x) = (x_1 + 2x_2)^2$ has unique minimal support $\{1, 2\}$.

Conditional moments and variance corrections. Write $P = I_4 - \mathbf{1}\mathbf{1}^\top/4 = B_\perp B_\perp^\top$, $W = PX = B_\perp Z$, and $A = \mathbf{1}^\top X/4$. Joint Gaussianity and zero covariance give

$$X = W + A\mathbf{1}, \qquad A \perp Z, \qquad A \sim N(0, 5/4).$$

Consequently, for any $b \in \mathbb{R}^4$,

$$E\{(b^\top X)^2 \mid Z\} = (b^\top W)^2 + \frac{5}{4}(b^\top \mathbf{1})^2.$$

Set $a = (1, 2, 0, 0)^\top$. Since $\mathbf{1}^\top W = 0$, the vectors a, $a - 2\mathbf{1} = (-1, 0, -2, -2)^\top$, and $a - \mathbf{1} = (0, 1, -1, -1)^\top$ have the same inner product with W. The corresponding linear combinations of X have conditional variances $45/4$, $125/4$, and $5/4$, respectively. Thus

$$\begin{aligned} r(Z) = E(Y \mid Z) = (a^\top W)^2 + 45/4, \\ E\{(-X_1 - 2X_3 - 2X_4)^2 - 20 \mid Z\} = r(Z), \\ E\{(X_2 - X_3 - X_4)^2 + 10 \mid Z\} = r(Z). \end{aligned}$$

Together with h_0 on the two candidates containing $\{1, 2\}$, these give the four solutions displayed in the main text.

Completeness on each candidate. Fix any three-element set S, and let $J_S \in \mathbb{R}^{4\times 3}$ embed its coordinates, so $X_S = J_S^\top X$. Then

$$\mathrm{Cov}(Z, X_S) = B_\perp^\top J_S.$$

This matrix is invertible: if $B_\perp^\top J_S v = 0$, then $J_S v$ is a multiple of $\mathbf{1}$; its omitted coordinate is zero, so $v = 0$. Hence the Gaussian conditional mean matrix $M_S = \mathrm{Cov}(Z, X_S)\,\mathrm{Cov}(X_S)^{-1}$ is invertible. Write $R_S = \mathrm{Cov}(Z \mid X_S)$. For $f \in L_2(P_{X_S})$ with $E\{f(X_S) \mid Z\} = 0$, Gaussian conditioning gives, for every $t \in \mathbb{R}^3$,

$$0 = E\{f(X_S)e^{\mathrm{i}t^\top Z}\} = e^{-t^\top R_S t/2} E\{f(X_S)e^{\mathrm{i}t^\top M_S X_S}\}.$$

The exponential factor never vanishes, even though R_S may be singular. Since M_S is invertible, the Fourier transform of the finite signed measure $f\,dP_{X_S}$ vanishes everywhere. Uniqueness of the Fourier transform gives $f = 0$ almost surely. This proves injectivity on the full space $L_2(P_{X_S})$, without restricting candidates to quadratic functions.

Functional plurality. All four candidates are solvable, and completeness makes their displayed solutions unique. The two incorrect polynomials are distinct from each other and from h_0 on sets of positive probability. Each therefore receives one vote, whereas h_0 receives two. This verifies Assumption 3 with $q = 1$, and completes verification of the example's identification conditions.

A.5 Identification lemmas and proof of Theorem 1

Lemma 1 (Population synthetic-IV moment)**.** *Under model* (1) *and* $E(u_0 \mid \epsilon) = 0$*,*

$$r = Th_0, \qquad Q_0(h) = 0 \iff Th = r.$$

Proof. By construction, $B_\perp^T \Lambda = 0$, so the projection removes the latent-factor component. Substituting $X = \Lambda U + \epsilon$ gives

$$Z = B_\perp^T X = B_\perp^T \Lambda U + B_\perp^T \epsilon = B_\perp^T \epsilon.$$

Because Z is a function of ϵ, the tower property of conditional expectation applies:

$$E(u_0 \mid Z) = E\{E(u_0 \mid \epsilon) \mid Z\} = E(0 \mid Z) = 0.$$

Taking conditional expectations in $Y = h_0(X) + u_0$, we have

$$r(Z) = E(Y \mid Z) = E\{h_0(X) \mid Z\} + E(u_0 \mid Z) = (Th_0)(Z).$$

Finally, $Q_0(h) = \|Th - r\|_2^2 = E[\{(Th)(Z) - r(Z)\}^2]$ is the expectation of a nonnegative random variable; hence it vanishes if and only if $(Th)(Z) = r(Z)$ almost surely. □

Lemma 2 (Uniqueness on a solvable candidate)**.** *For every* $S \in \mathcal{S}_{\mathrm{sol}}$*, the candidate solution* h_S *is unique up to almost-sure equivalence.*

Proof. Fix $S \in \mathcal{S}_{\text{sol}}$. By the definition of $\mathcal{S}_{\text{sol}}$, the equation $T_S h_S = r$ has at least one solution in $L_2(P_{X_S})$; we show that it has only one. Let $h_{S,1}$ and $h_{S,2}$ be two solutions and put $g = h_{S,1} - h_{S,2} \in L_2(P_{X_S})$. Since conditional expectation is linear,

$$E\{g(X_S) \mid Z\} = E\{h_{S,1}(X_S) \mid Z\} - E\{h_{S,2}(X_S) \mid Z\} = r(Z) - r(Z) = 0.$$

The injectivity part of Assumption 3 now gives $g(X_S) = 0$, that is, $h_{S,1}(X_S) = h_{S,2}(X_S)$ almost surely. Thus the embedded function $h^S(x) = h_S(x_S)$, and hence the vote of S, is well defined. This lemma uses only candidate injectivity; Theorem 1 also uses the bound on incorrect votes. □

Proof of Theorem 1. We first show that every minimizer solves the moment equation. We then prove uniqueness by counting votes. Any sparse solution receives a vote from every candidate set containing its support. These votes exceed the maximum allowed for an incorrect function by Assumption 3.

By Assumption 2, h_0 has a unique minimal support of size $s < p - q$, and hence $h_0 \in \mathcal{H}_{<p-q}$. Lemma 1 gives $Q_0(h_0) = 0$. Since $Q_0 \geq 0$, the minimum of Q_0 over $\mathcal{H}_{<p-q}$ equals zero, and, by Lemma 1 again, every minimizer $h^\star$ satisfies $Th^\star = r$. It remains to show that such an $h^\star$ equals h_0.

Next, consider the votes received by an arbitrary sparse solution. Let $h^\star \in \mathcal{H}_{<p-q}$ satisfy $Th^\star = r$, and let $M = S(h^\star)$. Then $|M| \leq p - q - 1$, and $h^\star(X) = h^\star_M(X_M)$ for some function $h^\star_M$. Take any candidate set S with $|S| = p - q$ and $M \subseteq S$, and define a function of x_S that ignores the coordinates in $S \setminus M$:

$$h^\star_S(x_S) = h^\star_M(x_M).$$

Then $h^\star_S(X_S) = h^\star(X)$; hence $h^\star_S \in L_2(P_{X_S})$ and

$$(T_S h^\star_S)(Z) = E\{h^\star_S(X_S) \mid Z\} = E\{h^\star(X) \mid Z\} = (Th^\star)(Z) = r(Z).$$

Thus $S \in \mathcal{S}_{\text{sol}}$, and candidate uniqueness gives $h^S = h^\star$ by Lemma 2. Every candidate set containing M therefore votes for the same function $h^\star$.

Finally, we count these candidate sets. A candidate set containing M is obtained by choosing the remaining $p - q - |M|$ coordinates among the $p - |M|$ coordinates outside M. The number of such sets is

$$\binom{p - |M|}{p - q - |M|} = \binom{p - |M|}{q} \geq \binom{q+1}{q} = q + 1,$$

where the inequality uses $p - |M| \geq q + 1$ and the fact that $\binom{m}{q}$ is nondecreasing in m. Consequently $h^\star$ receives at least $q + 1$ votes. By Assumption 3, however, a function different from h_0 receives at most q votes. Therefore $h^\star = h_0$, and h_0 is the unique minimizer of Q_0 over $\mathcal{H}_{<p-q}$. □

B Estimation Assumptions

B.1 Scope and the class of network differences

Throughout this appendix, p, q, s are fixed and the budget is $k = s = |S_0|$. The support S_0 itself is unknown: the estimator searches over all networks with at most s nonzero first-layer columns. The width, depth, and weight bound of $\mathcal{H}_{s,n}$ may grow with n_T. The cross-validated choice of k is not analyzed. Throughout the estimation arguments, C_0 denotes a constant independent of n whose value may change between displays.

Several conditions below are required uniformly over differences of two networks,

$$\mathcal{A}_{s,n} = \{h - h' : h, h' \in \mathcal{H}_{s,n}\}.$$

We use three facts about this class. First, h and h' each depend on at most s coordinates, and hence $h - h'$ depends on at most $\min(2s, p)$. Second, the zero function is a network, which gives $\mathcal{H}_{s,n} \subseteq \mathcal{A}_{s,n}$. Third, changing the sign of the output weights shows that $-h' \in \mathcal{H}_{s,n}$, and hence the sum $h + h' = h - (-h')$ also belongs to $\mathcal{A}_{s,n}$. The target h_0 need not be a network, so $\widehat{h} - h_0$ may fall outside $\mathcal{A}_{s,n}$. The proof therefore compares $\widehat{h}$ with an approximating network h°_n, keeping the difference within the class controlled by the assumptions.

B.2 CONDITIONS FOR THE ESTIMATION THEOREM

Theorem 2 uses Assumptions 1 and 2, together with the conditions below.

Assumption 4 (True-support completeness)**.** The operator

$$T_{S_0} : L_2(P_{X_{S_0}}) \longrightarrow L_2(P_Z)$$

is injective.

In the example of Appendix A.4, true-support completeness follows by embedding any function of X_1, X_2 in a three-treatment candidate containing them.

Assumption 5 (Nuisance split and generated-instrument regularity)**.** The estimated bases $\widehat{B}_\parallel$, $\widehat{B}_\perp$, and the conditional sampler $\widehat{P}_V$ are computed from $\mathcal{D}^N$ only, and $\mathcal{D}^N$ is independent of $\mathcal{D}^T$. The estimated projector $\widehat{P}_\parallel = \widehat{B}_\parallel \widehat{B}_\parallel^T$ satisfies

$$\|\widehat{P}_\parallel - P_\parallel\|_{\mathrm{op}} = O_p(n_N^{-1/2}). \tag{10}$$

For a $p \times (p-q)$ matrix C with orthonormal columns, write $Z_C = C^T X$ and $(T_C g)(z) = E\{g(X) \mid Z_C = z\}$, so that $Z = Z_{B_\perp}$ and $T = T_{B_\perp}$. Let $\mathcal{A}_{s,n}^+ = \mathcal{A}_{s,n} \cup \{h_0\}$. There are constants $r_0 > 0$, L_T, and L_U, depending on p, q, s but not on n, such that every such C with $\|C - B_\perp\|_{\mathrm{op}} \le r_0$ satisfies

$$\sup_{0 \neq g \in \mathcal{A}_{s,n}^+} \frac{\|(T_C g)(Z_C) - (Tg)(Z)\|_2}{\|g\|_2} \le L_T \|C - B_\perp\|_{\mathrm{op}}, \tag{11}$$

$$\|E(u_0 \mid Z_C)\|_{L_2(P_{Z_C})} \le L_U \|C - B_\perp\|_{\mathrm{op}}. \tag{12}$$

The first condition requires a root-n_N loading-space estimator; see, for example, Anderson & Amemiya (1988) for factor-model estimation. Conditions (11) and (12) control distinct consequences of estimating the instrument: operator perturbation and approximate validity. Both are additional requirements beyond projector consistency. Including h_0 in (11) also controls $E(Y \mid Z_C) - r(Z)$ after evaluating both on the same observation, using (12) and $Th_0 = r$. This is needed in the proof of Proposition 1; the bound over $\mathcal{A}_{s,n}$ alone does not give it, because h_0 is in general not a network.

For the estimated coordinates, let

$$(T_{\widehat{B}} g)(z) = E\Big[g(\widehat{B}_\parallel \widehat{V} + \widehat{B}_\perp z) \mid \widehat{Z} = z, \mathcal{D}^N\Big]$$

use the true conditional law of $\widehat{V} \mid \widehat{Z} = z$, whereas $\widehat{T}_{\widehat{B}}$ in (4) uses the fitted sampler. The realized operator error $\mathcal{E}_{F,n}$ is the smallest nonnegative number such that every $g \in \mathcal{A}_{s,n}$ satisfies

$$\|(\widehat{T}_{\widehat{B}} - T_{\widehat{B}})g\|_{L_2(P_{\widehat{Z}}|\mathcal{D}^N)} \le \mathcal{E}_{F,n} \|g\|_{L_2(P_X)} + \frac{1}{n_T}. \tag{13}$$

The additive tolerance n_T^{-1} is absorbed by the error bound in Theorem 2. It accommodates tail errors that cannot be bounded uniformly relative to a network difference's potentially small L_2 norm. Appendix E derives this form for a Gaussian sampler fitted by maximum likelihood.

Assumption 6 (Conditional-sampler accuracy)**.** For a deterministic sequence $\delta_{F,n} \to 0$, $\mathcal{E}_{F,n} = O_p(\delta_{F,n})$. On an event determined by $\mathcal{D}^N$ whose probability tends to one, the observed and reconstructed treatment coordinates are sub-Gaussian, and every $g \in \mathcal{A}_{s,n}$ satisfies the second-moment bound

$$E_{\widehat{Z}}\Big[E_{\widetilde{X}|\widehat{Z},\mathcal{D}^N} g(\widetilde{X})^2\Big] \le C_V \|g\|_2^2 + \frac{1}{n_T^2}.$$

The sub-Gaussian constant and C_V do not depend on n. Given $\mathcal{D}^N$ and $\mathcal{D}^T$, the draws $\widetilde{V}_{i\ell}^{(b)}$, $i \in \mathcal{D}^T$, $b \in \{1, 2\}$, $\ell = 1, \ldots, m_n$, used in (7) are independent, and $\widetilde{V}_{i\ell}^{(b)}$ has law $\widehat{P}_V(\cdot \mid \widehat{Z}_i)$.

Assumption 6 states the statistical requirements on the conditional sampler. Using a diffusion model, GAN, flow, mixture, or another learner requires verifying these bounds for that learner. Appendix E verifies Assumption 6 with $\delta_{F,n} = n_N^{-1/2} \log n_T$ for a Gaussian conditional sampler.

The next condition controls the difficulty of recovering h_0 through the conditional operator and approximating it by a neural network. Let $\overline{\mathcal{H}}_{S_0} = L_2(P_{X_{S_0}})$, embedded in $L_2(P_X)$. Write $T_0 = T_{S_0}$, with adjoint T_0^*, and define the regularized reference on the true support by

$$h_{\alpha,0} = (T_0^* T_0 + \alpha I)^{-1} T_0^* r = \underset{h \in \overline{\mathcal{H}}_{S_0}}{\arg\min} \{\|T_0 h - r\|_2^2 + \alpha \|h\|_2^2\}.$$

Assumption 7 (Inverse problem and neural sieve)**.** For some $\beta_{\text{iv}} > 0$,

$$h_0 = (T_0^* T_0)^{\beta_{\text{iv}}/2} w_0, \qquad \|w_0\|_2 \le C_w. \tag{14}$$

The outcome and the network outputs are bounded: $|Y| \le M_Y$, and $|h| \le M$ for every $h \in \mathcal{H}_{s,n}$. There is a network $h_n^\circ \in \mathcal{H}_{s,n}$ that uses only the coordinates in S_0 and satisfies, for a constant C_H that does not depend on n,

$$\|h_n^\circ - h_0\|_2 \le \epsilon_{H,n}, \qquad Q_{\alpha_n}(h_n^\circ) - Q_{\alpha_n}(h_{\alpha_n,0}) \le C_H \epsilon_{H,n}^2. \tag{15}$$

The first bound in (15) controls approximation to h_0; the second separately controls excess regularized loss relative to $h_{\alpha_n,0}$. Both bounds are assumed for the same network.

Assumption 8 (Restricted regularized curvature)**.** Write $\nu = \min\{\beta_{\text{iv}}, 2\}$. For constants $c \in (0, 1]$ and $C_R < \infty$, independent of n, at $\alpha = \alpha_n$ every $h \in \mathcal{H}_{s,n}$ satisfies

$$Q_\alpha(h) - Q_\alpha(h_{\alpha,0}) \ge c\alpha \|h - h_{\alpha,0}\|_2^2 - C_R \alpha^{1+\nu}. \tag{16}$$

The reference $h_{\alpha,0}$ is defined on the true support, whereas the estimator searches across supports. The curvature condition connects these two domains by converting an objective bound into an L_2 error bound. It is an additional population restriction, beyond completeness and plurality. Related restrictions appear as restricted eigenvalue conditions (Bickel et al., 2009; van de Geer & Bühlmann, 2009) and restricted strong convexity (Negahban et al., 2012). The term $C_R \alpha^{1+\nu}$ allows a tolerance, as in Loh & Wainwright (2012).

B.3 Functional-support separation

Let $\mathcal{S}_s = \{S \subseteq \{1, \dots, p\} : |S| \le s\}$, and define the population functional distance

$$D_S(f) = \|f - E\{f(X) \mid X_S\}\|_2 = \inf_{v \in L_2(P_{X_S})} \|f - v\|_2.$$

The conditional expectation is the orthogonal projection onto a closed subspace of $L_2(P_X)$. Thus $D_S(f) = 0$ exactly when S supports f, regardless of the neural representation. Write

$$\kappa_0 := \min_{\substack{S \in \mathcal{S}_s \\ S \ne S_0}} D_S(h_0), \qquad D_S(h_0)^2 = E[\mathrm{Var}\{h_0(X) \mid X_S\}], \tag{17}$$

with the convention that the minimum over an empty collection is $+\infty$; if $s = 0$, support extraction is trivially the empty set. The constant κ_0 measures how well incorrect supports can approximate the true structural function. For fixed p, its positivity follows from unique minimal support by Lemma 10 in Appendix F; no separate positivity assumption is needed.

C Proof of the Main Estimation Theorem

The proof proceeds from nuisance estimation to the structural criterion. We first control errors in the estimated instruments and conditional operator, then bound differences between empirical and population criteria. Restricted curvature and the source condition convert these bounds into the estimation rate. Two conventions apply throughout.

Conditioning on the nuisance sample. The estimates $\widehat{B}_\parallel, \widehat{B}_\perp, \widehat{P}_V$ are functions of $\mathcal{D}^N$ only, and $\mathcal{D}^T$ is independent of $\mathcal{D}^N$. Conditionally on $\mathcal{D}^N$, these estimates are fixed, and the observations in $\mathcal{D}^T$ are still i.i.d. draws of (X, Y). Unless stated otherwise, expectations, L_2 norms, and probabilities are conditional on $\mathcal{D}^N$. We establish the bounds on nuisance-sample events with probability tending to one, then integrate the conditional probabilities to obtain unconditional statements.

Instruments built from another basis. For a $p \times (p-q)$ matrix C with orthonormal columns, write

$$Z_C = C^T X, \qquad (T_C g)(z) = E\{g(X) \mid Z_C = z\},$$

so that $Z = Z_{B_\perp}$ and $T = T_{B_\perp}$. If R is a $(p-q) \times (p-q)$ orthogonal matrix, then $Z_{CR} = R^T Z_C$ is a one-to-one function of Z_C. Conditioning on Z_{CR} is thus the same as conditioning on Z_C:

$$(T_{CR}\, g)(Z_{CR}) = (T_C g)(Z_C), \qquad E(u_0 \mid Z_{CR}) = E(u_0 \mid Z_C). \tag{18}$$

In particular, with $C = \widehat{B}_\perp$, rotating the estimated basis changes neither $E\{g(X) \mid \widehat{Z}, \mathcal{D}^N\}$ nor $E(u_0 \mid \widehat{Z}, \mathcal{D}^N)$.

Lemma 3 (Projector and aligned-basis perturbation)**.** *Under Assumption 5, we have* $\|\widehat{P}_\perp - P_\perp\|_{\mathrm{op}} = O_p(n_N^{-1/2})$*, and orthogonal rotations* $R_\parallel, R_\perp$ *can be chosen so that*

$$\|\widehat{B}_\parallel R_\parallel - B_\parallel\|_{\mathrm{op}} + \|\widehat{B}_\perp R_\perp - B_\perp\|_{\mathrm{op}} = O_p(n_N^{-1/2}). \tag{19}$$

Proof. Since $P_\perp = I - P_\parallel$ and $\widehat{P}_\perp = I - \widehat{P}_\parallel$, we have $\widehat{P}_\perp - P_\perp = -(\widehat{P}_\parallel - P_\parallel)$. The two projector errors have the same operator norm, and Assumption 5 bounds it by $O_p(n_N^{-1/2})$. We prove a general fact and apply it twice. Let B and $\widehat{B}$ be $p \times d$ matrices with orthonormal columns, and let $P = BB^T$ and $\widehat{P} = \widehat{B}\widehat{B}^T$. We claim that there is a $d \times d$ orthogonal matrix R with

$$\|\widehat{B}R - B\|_{\mathrm{op}} \le \sqrt{2}\, \|\widehat{P} - P\|_{\mathrm{op}}. \tag{20}$$

Take a singular value decomposition $\widehat{B}^T B = U\Sigma V^T$, where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$. Because $\|\widehat{B}^T B\|_{\mathrm{op}} \le 1$, every σ_j lies in $[0,1]$. Put $R = UV^T$. Then $R^T \widehat{B}^T B = VU^T U\Sigma V^T = V\Sigma V^T$, which is symmetric, hence $B^T \widehat{B} R = V\Sigma V^T$ as well. Expanding the product and using $\widehat{B}^T\widehat{B} = B^T B = I$,

$$(\widehat{B}R - B)^T(\widehat{B}R - B) = 2I - R^T\widehat{B}^T B - B^T\widehat{B}R = 2V(I - \Sigma)V^T.$$

The largest eigenvalue of the right side is $2(1-\sigma_{\min})$, where $\sigma_{\min} = \min_j \sigma_j$. Hence $\|\widehat{B}R - B\|_{\mathrm{op}}^2 = 2(1 - \sigma_{\min})$. Next we relate $\sigma_{\min}$ to the projector error. Since $I - \widehat{P}$ is an orthogonal projector,

$$\{(I - \widehat{P})B\}^T\{(I - \widehat{P})B\} = B^T(I - \widehat{P})B = I - (\widehat{B}^T B)^T(\widehat{B}^T B) = V(I - \Sigma^2)V^T,$$

and hence $\|(I - \widehat{P})B\|_{\mathrm{op}}^2 = 1 - \sigma_{\min}^2$. Moreover $PB = B$, hence $(I - \widehat{P})B = (P - \widehat{P})B$ and

$$1 - \sigma_{\min}^2 = \|(P - \widehat{P})B\|_{\mathrm{op}}^2 \le \|\widehat{P} - P\|_{\mathrm{op}}^2.$$

Finally $0 \le \sigma_{\min} \le 1$ gives $1 - \sigma_{\min} \le (1 - \sigma_{\min})(1 + \sigma_{\min}) = 1 - \sigma_{\min}^2$. Combining the three displays proves (20). Apply (20) to $(B_\parallel, \widehat{B}_\parallel)$ with $d = q$ and to $(B_\perp, \widehat{B}_\perp)$ with $d = p - q$. Together with the first claim and Assumption 5, this proves (19). □

Lemma 4 (Approximate validity of the synthetic instrument)**.** *Under Assumption 5,*

$$\|E(u_0 \mid \widehat{Z}, \mathcal{D}^N)\|_{L_2(P_{\widehat{Z}}|\mathcal{D}^N)} = O_p(n_N^{-1/2}).$$

Proof. Condition on $\mathcal{D}^N$, and let $C = \widehat{B}_\perp R_\perp$ with $R_\perp$ from Lemma 3. Then C is a fixed matrix with orthonormal columns. A new observation (X, u_0) is independent of $\mathcal{D}^N$; conditioning on $\mathcal{D}^N$ therefore only fixes C, and (18) gives $E(u_0 \mid \widehat{Z}, \mathcal{D}^N) = E(u_0 \mid Z_C)$. By Lemma 3, C lies in the neighborhood of $B_\perp$ required by Assumption 5 with probability tending to one. On that event, the approximate-validity bound in Assumption 5 and Lemma 3 yield

$$\|E(u_0 \mid Z_C)\|_{L_2(P_{Z_C})} \le L_U \|C - B_\perp\|_{\mathrm{op}} = O_p(n_N^{-1/2}).$$

□

The next lemma bounds, uniformly over differences of two networks, the error of replacing the true operator T by the fitted operator $\widehat{T}_{\widehat{B}}$.

Lemma 5. *Under Assumptions 5 and 6, every $g \in \mathcal{A}_{s,n}$ satisfies*

$$\|\widehat{T}_{\widehat{B}} g(\widehat{Z}) - Tg(Z)\|_2 \le \big(\mathcal{E}_{F,n} + L_T \|\widehat{B}_\perp R_\perp - B_\perp\|_{\mathrm{op}}\big)\|g\|_2 + \frac{1}{n_T}, \tag{21}$$

where $R_\perp$ is the rotation of Lemma 3, and $\mathcal{E}_{F,n} + L_T\|\widehat{B}_\perp R_\perp - B_\perp\|_{\mathrm{op}} = O_p(\delta_{F,n} + n_N^{-1/2})$.

Proof. We evaluate both operators on the same observation X, which places their difference in a common L_2 space. For fixed $g \in \mathcal{A}_{s,n}$, add and subtract $T_{\widehat{B}} g(\widehat{Z})$, the conditional expectation of $g(X)$ given $\widehat{Z}$ under the true law, and use the triangle inequality:

$$\|\widehat{T}_{\widehat{B}} g(\widehat{Z}) - Tg(Z)\|_2 \le \|(\widehat{T}_{\widehat{B}} - T_{\widehat{B}})g(\widehat{Z})\|_2 + \|T_{\widehat{B}} g(\widehat{Z}) - Tg(Z)\|_2.$$

By the definition of $\mathcal{E}_{F,n}$ in (13), the first term on the right is at most $\mathcal{E}_{F,n}\|g\|_2 + n_T^{-1}$, and by Assumption 6, $\mathcal{E}_{F,n} = O_p(\delta_{F,n})$. For the second term, let $C = \widehat{B}_\perp R_\perp$. By (18), $T_{\widehat{B}} g(\widehat{Z}) = (T_C g)(Z_C)$. The operator-perturbation bound in Assumption 5 and Lemma 3 give

$$\|(T_C g)(Z_C) - (Tg)(Z)\|_2 \le L_T \|C - B_\perp\|_{\mathrm{op}} \|g\|_2 = O_p(n_N^{-1/2})\,\|g\|_2.$$

Adding the two bounds proves the lemma. □

THE THREE SOURCES OF ERROR IN THE CRITERION

For a function f of (X, Y), write

$$P_{n_T} f = \frac{1}{n_T} \sum_{i \in \mathcal{D}^T} f(X_i, Y_i), \qquad Pf = E\{f(X,Y) \mid \mathcal{D}^N\}.$$

Given $\mathcal{D}^N$, define the loss with its regularization term by

$$\bar{\ell}_h(X,Y) = \{Y - \widehat{T}_{\widehat{B}} h(\widehat{Z})\}^2 + \alpha_n h(X)^2, \qquad \widehat{Z} = \widehat{B}_\perp^T X,$$

which is a fixed function of (X, Y) after conditioning on the nuisance sample. This loss uses exact conditional expectations under the fitted sampler and therefore contains no Monte Carlo error. For every $h \in \mathcal{H}_{s,n}$,

$$G_n(h) = \widehat{Q}_{\alpha_n}(h) - Q_{\alpha_n}(h) = \underbrace{\widehat{Q}_{\alpha_n}(h) - P_{n_T}\bar{\ell}_h}_{\text{Monte Carlo error}} + \underbrace{(P_{n_T} - P)\bar{\ell}_h}_{\text{sampling error}} + \underbrace{P\bar{\ell}_h - Q_{\alpha_n}(h)}_{\substack{\text{nuisance contribution} \\ +\text{constant}}}. \tag{22}$$

The last term contains an h-independent outcome-variance constant, which cancels in all criterion differences below. Lemmas 8 and 9 in Appendix D control the first two terms: uniformly over $h, h' \in \mathcal{H}_{s,n}$,

$$|(P_{n_T} - P)(\bar{\ell}_h - \bar{\ell}_{h'})| = O_p\{\rho_n \|h - h'\|_2 + \rho_n^2\},$$

$$\Big|\{\widehat{Q}_{\alpha_n}(h) - P_{n_T}\bar{\ell}_h\} - \{\widehat{Q}_{\alpha_n}(h') - P_{n_T}\bar{\ell}_{h'}\}\Big| = O_p\{\xi_{\mathrm{MC},n}\|h - h'\|_2 + \xi_{\mathrm{MC},n}^2 + \rho_n \xi_{\mathrm{MC},n}\}.$$

The third term is handled in the proof of the next proposition.

Proposition 1 (Localized criterion increment)**.** *Under Assumptions 5–7, with d_n defined in* (8)*,*

$$\sup_{h,h' \in \mathcal{H}_{s,n}} \frac{|G_n(h) - G_n(h')|}{\|h - h'\|_2 + d_n} = O_p(d_n), \qquad G_n = \widehat{Q}_{\alpha_n} - Q_{\alpha_n}.$$

Proof. Fix $h, h' \in \mathcal{H}_{s,n}$. By (22), $G_n(h) - G_n(h')$ is the sum of three differences. Lemmas 8 and 9 bound sampling and Monte Carlo errors. It remains to control the nuisance contribution,

$$\Delta_n(h, h') = \{P\bar{\ell}_h - Q_{\alpha_n}(h)\} - \{P\bar{\ell}_{h'} - Q_{\alpha_n}(h')\}.$$

Define $\widehat{r}(\widehat{Z}) = E(Y \mid \widehat{Z}, \mathcal{D}^N)$, and recall that $r(Z) = E(Y \mid Z)$. Here $\widehat{r}$ is the true conditional outcome mean for the estimated instrument, introduced only for the proof; it is not a fitted outcome

regression. Since $\widehat{T}_{\widehat{B}}h(\widehat{Z})$ is a function of $\widehat{Z}$, and $Y - \widehat{r}(\widehat{Z})$ has conditional mean zero given $\widehat{Z}$, the cross term vanishes and

$$P\bar{\ell}_h = E\{Y - \widehat{r}(\widehat{Z})\}^2 + \|\widehat{T}_{\widehat{B}}h - \widehat{r}\|_2^2 + \alpha_n\|h\|_2^2.$$

The first term does not depend on h. The last term also appears in $Q_{\alpha_n}(h) = \|Th - r\|_2^2 + \alpha_n\|h\|_2^2$ and cancels. Using $\|a\|_2^2 - \|a'\|_2^2 = \langle a - a', a + a'\rangle$ for both squared norms, together with the linearity of $\widehat{T}_{\widehat{B}}$ and T,

$$\Delta_n(h, h') = \langle \widehat{T}_{\widehat{B}}(h - h'),\, \widehat{T}_{\widehat{B}}(h + h') - 2\widehat{r}\rangle - \langle T(h - h'),\, T(h + h') - 2r\rangle.$$

Here every function is evaluated at the same observation, at $\widehat{Z}$ or at Z, and $\langle\cdot,\cdot\rangle$ is the expectation of the product. Adding and subtracting $\langle T(h - h'),\, \widehat{T}_{\widehat{B}}(h + h') - 2\widehat{r}\rangle$,

$$\Delta_n(h, h') = \langle[\widehat{T}_{\widehat{B}} - T](h - h'),\, \widehat{T}_{\widehat{B}}(h + h') - 2\widehat{r}\rangle + \langle T(h - h'),\, [\widehat{T}_{\widehat{B}} - T](h + h') - 2(\widehat{r} - r)\rangle. \quad (23)$$

Let $\eta_n = \delta_{F,n} + n_N^{-1/2}$. Both $h - h'$ and $h + h'$ belong to $\mathcal{A}_{s,n}$: the difference by the definition of that class, and the sum because the output layer of the network class is symmetric in sign, which gives $-h' \in \mathcal{H}_{s,n}$ and $h + h' = h - (-h')$. Lemma 5 therefore applies to both and gives $\|\widehat{T}_{\widehat{B}}(h - h') - T(h - h')\|_2 \le O_p(\eta_n)\|h - h'\|_2 + n_T^{-1}$ and $\|\widehat{T}_{\widehat{B}}(h + h') - T(h + h')\|_2 \le O_p(\eta_n)\|h + h'\|_2 + n_T^{-1}$, where the $O_p(\eta_n)$ terms do not depend on h, h'. By Assumption 7 every network is bounded in absolute value by M; hence $\|h + h'\|_2 \le 2M$ and the second bound is $O_p(\eta_n) + n_T^{-1}$. The same bound on the networks gives $|\widehat{T}_{\widehat{B}}(h + h')| \le 2M$, while $|Y| \le M_Y$ from the same assumption gives $|\widehat{r}| \le M_Y$, because $\widehat{r}$ is a conditional expectation of Y; hence $\|\widehat{T}_{\widehat{B}}(h + h') - 2\widehat{r}\|_2 \le 2M + 2M_Y$. Conditional expectation does not increase the L_2 norm, which gives $\|T(h - h')\|_2 \le \|h - h'\|_2$.

It remains to bound $\widehat{r} - r$. Substituting $Y = h_0(X) + u_0$ and using $r = Th_0$ from Lemma 1,

$$\widehat{r}(\widehat{Z}) - r(Z) = E(u_0 \mid \widehat{Z}, \mathcal{D}^N) + \{(T_{\widehat{B}}h_0)(\widehat{Z}) - (Th_0)(Z)\}.$$

The first term is $O_p(n_N^{-1/2})$ by Lemma 4. The second is bounded exactly as the second term in the proof of Lemma 5, which is possible because the operator-perturbation bound in Assumption 5 is assumed for $g = h_0$ as well. Hence $\|\widehat{r} - r\|_2 = O_p(n_N^{-1/2})$.

Applying Cauchy–Schwarz to (23) and substituting these bounds gives

$$\begin{aligned} |\Delta_n(h, h')| &\le \{O_p(\eta_n)\|h - h'\|_2 + n_T^{-1}\}(2M + 2M_Y) \\ &\quad + \|h - h'\|_2\{O_p(\eta_n) + n_T^{-1} + O_p(n_N^{-1/2})\} \\ &= O_p(\eta_n)\,\|h - h'\|_2 + \frac{4M + 2M_Y}{n_T}, \end{aligned} \quad (24)$$

where the additive term uses $\|h - h'\|_2 \le 2M$. The O_p terms come from Lemmas 4 and 5 and do not depend on h or h'; hence the bound is uniform over $\mathcal{H}_{s,n}$.

By (22), the two displays preceding the proposition, and (24), we have

$$|G_n(h) - G_n(h')| \le O_p(1)\big\{(\rho_n + \xi_{\mathrm{MC},n} + \eta_n)\|h - h'\|_2 + \rho_n^2 + \xi_{\mathrm{MC},n}^2 + \rho_n\xi_{\mathrm{MC},n} + n_T^{-1}\big\}.$$

By the Cauchy–Schwarz inequality, we have $\rho_n + \xi_{\mathrm{MC},n} + \eta_n = \rho_n + \xi_{\mathrm{MC},n} + \delta_{F,n} + n_N^{-1/2} \le 2d_n$. Also $\rho_n\xi_{\mathrm{MC},n} \le (\rho_n^2 + \xi_{\mathrm{MC},n}^2)/2$. By the definition of ρ_n in Appendix D, $n_T^{-1} \le \rho_n^2$. Therefore $|G_n(h) - G_n(h')| \le O_p(1)\, d_n(\|h - h'\|_2 + d_n)$ uniformly over $h, h' \in \mathcal{H}_{s,n}$, which concludes the proof. □

Proof of Theorem 2. We first bound the bias introduced by regularization on the true support. Let $A = T_0^*T_0$ and write $\beta = \beta_{\mathrm{iv}}$ for short. Since $r = T_0h_0$, we have $h_{\alpha,0} = (A + \alpha I)^{-1}Ah_0$, and the source condition in Assumption 7 gives

$$h_{\alpha,0} - h_0 = -\alpha(A + \alpha I)^{-1}A^{\beta_{\mathrm{iv}}/2}w_0.$$

Conditional expectation is an L_2 contraction, which gives $\|A\| = \|T_0\|^2 \le 1$. Let E_A denote the spectral projection measure of A. As $d\|E_A(\cdot)w_0\|_2^2$ has total mass $\|w_0\|_2^2 \le C_w^2$, the spectral theorem gives

$$\|h_{\alpha,0} - h_0\|_2^2 = \int_0^1 g_\alpha \, d\|E_A w_0\|_2^2 \le C_w^2 \sup_{0\le\lambda\le 1} g_\alpha(\lambda), \qquad g_\alpha(\lambda) = \frac{\alpha^2\lambda^\beta}{(\lambda+\alpha)^2}, \tag{25}$$

and it remains to show $\sup_{[0,1]} g_\alpha \le \alpha^\nu$ with $\nu = \min\{\beta, 2\}$.

For $\beta \le 2$ the substitution $\lambda = \alpha t$ extracts the α-scale, $g_\alpha(\alpha t) = \alpha^\beta \frac{t^\beta}{(1+t)^2}$, for $t \ge 0$, and $t^\beta \le (1+t)^2$ throughout: the numerator is at most one for $t \le 1$, while $t^\beta \le t^2 < (1+t)^2$ for $t > 1$. For $\beta > 2$, we have $g_\alpha(\lambda) = \alpha^2\lambda^{\beta-2}\frac{\lambda^2}{(\lambda+\alpha)^2} \le \alpha^2\lambda^{\beta-2} \le \alpha^2$, by $\lambda \le \lambda + \alpha$ and $\lambda \le 1$. Hence $\|h_{\alpha,0} - h_0\|_2^2 \le C_w^2\alpha^\nu$.

Next we compare $\widehat{h}$ with a network close to h_0. Let $h_n^\circ \in \mathcal{H}_{s,n}$ be the network from Assumption 7. The estimator $\widehat{h}$ minimizes $\widehat{Q}_{\alpha_n}$ over $\mathcal{H}_{s,n}$, and h_n° belongs to that class, hence $\widehat{Q}_{\alpha_n}(\widehat{h}) \le \widehat{Q}_{\alpha_n}(h_n^\circ)$. Write $\widehat{Q}_{\alpha_n} = Q_{\alpha_n} + G_n$ on both sides, move the two G_n terms to the right, and subtract $Q_{\alpha_n}(h_{\alpha_n,0})$ from both sides:

$$Q_{\alpha_n}(\widehat{h}) - Q_{\alpha_n}(h_{\alpha_n,0}) \le \{G_n(h_n^\circ) - G_n(\widehat{h})\} + \{Q_{\alpha_n}(h_n^\circ) - Q_{\alpha_n}(h_{\alpha_n,0})\}.$$

Restricted curvature bounds the left side in terms of estimation error. Applying Assumption 8 to $\widehat{h}$ gives

$$Q_{\alpha_n}(\widehat{h}) - Q_{\alpha_n}(h_{\alpha_n,0}) \ge c\alpha_n\|\widehat{h} - h_{\alpha_n,0}\|_2^2 - C_R\alpha_n^{1+\nu}.$$

For the first term on the right, since both h_n° and $\widehat{h}$ belong to $\mathcal{H}_{s,n}$, Proposition 1 applies to this pair; together with the triangle inequality $\|\widehat{h} - h_n^\circ\|_2 \le \|\widehat{h} - h_{\alpha_n,0}\|_2 + \|h_n^\circ - h_{\alpha_n,0}\|_2$ it gives

$$|G_n(h_n^\circ) - G_n(\widehat{h})| \le C_0 d_n\big(\|\widehat{h} - h_{\alpha_n,0}\|_2 + \|h_n^\circ - h_{\alpha_n,0}\|_2 + d_n\big)$$

on an event whose probability can be made arbitrarily close to one by taking the constant C_0 large. The second term on the right is at most $C_H\epsilon_{H,n}^2$ by the second inequality in Assumption 7. Combining these three bounds,

$$c\alpha_n\|\widehat{h} - h_{\alpha_n,0}\|_2^2 \le C_0 d_n\|\widehat{h} - h_{\alpha_n,0}\|_2 + C_0 d_n\|h_n^\circ - h_{\alpha_n,0}\|_2 + C_0 d_n^2 + C_H\epsilon_{H,n}^2 + C_R\alpha_n^{1+\nu}. \tag{26}$$

By the triangle inequality, the first inequality in Assumption 7, and the bias bound proved above, we have

$$\|h_n^\circ - h_{\alpha_n,0}\|_2 \le \|h_n^\circ - h_0\|_2 + \|h_0 - h_{\alpha_n,0}\|_2 \le \epsilon_{H,n} + C_w\alpha_n^{\nu/2},$$

and by the Cauchy–Schwarz inequality, $\|h_n^\circ - h_{\alpha_n,0}\|_2^2 \le 2\epsilon_{H,n}^2 + 2C_w^2\alpha_n^\nu$.

For numbers $a, b \ge 0$ and $\tau > 0$, we have $ab \le \tau a^2/2 + b^2/(2\tau)$. Applying this to the first two terms on the right of (26), with $\tau = c\alpha_n/2$ and $\tau = \alpha_n$ respectively,

$$C_0 d_n\|\widehat{h} - h_{\alpha_n,0}\|_2 \le \frac{c\alpha_n}{4}\|\widehat{h} - h_{\alpha_n,0}\|_2^2 + \frac{C_0^2}{c}\frac{d_n^2}{\alpha_n},$$

$$C_0 d_n\|h_n^\circ - h_{\alpha_n,0}\|_2 \le \frac{\alpha_n}{2}\|h_n^\circ - h_{\alpha_n,0}\|_2^2 + \frac{C_0^2}{2}\frac{d_n^2}{\alpha_n}.$$

Insert these bounds into (26), and move $c\alpha_n\|\widehat{h} - h_{\alpha_n,0}\|_2^2/4$ to the left:

$$\frac{3c}{4}\alpha_n\|\widehat{h} - h_{\alpha_n,0}\|_2^2 \le C_0\Big\{\frac{d_n^2}{\alpha_n} + d_n^2 + (1+\alpha_n)\epsilon_{H,n}^2 + \alpha_n^{1+\nu}\Big\}.$$

Dividing by α_n and using $\alpha_n \le 1$, which gives $d_n^2/\alpha_n \le d_n^2/\alpha_n^2$ and $(1+\alpha_n)\epsilon_{H,n}^2/\alpha_n \le 2\epsilon_{H,n}^2/\alpha_n$,

$$\|\widehat{h} - h_{\alpha_n,0}\|_2^2 \le C_0\Big\{\frac{d_n^2}{\alpha_n^2} + \frac{\epsilon_{H,n}^2}{\alpha_n} + \alpha_n^\nu\Big\}.$$

Finally, by the triangle inequality and the bias bound again,

$$\|\widehat{h} - h_0\|_2^2 \le 2\|\widehat{h} - h_{\alpha_n,0}\|_2^2 + 2\|h_{\alpha_n,0} - h_0\|_2^2 \le 2\|\widehat{h} - h_{\alpha_n,0}\|_2^2 + 2C_w^2\alpha_n^\nu,$$

which has the order claimed in Theorem 2. □

Remark 1 (Same-sample nuisance estimation)**.** When nuisance estimation and structural fitting share observations, conditioning on the nuisance estimates no longer leaves the structural sample i.i.d. The proof above therefore does not cover same-sample fitting. Such a result would require joint uniform bounds over nuisance estimates and networks, or a leave-one-out argument. Sample splitting supplies the independence used here; fixed-fold cross-fitting supplies it within each held-out fold.

D Neural Sieve and Double-Monte-Carlo Complexity

This appendix controls sampling and Monte Carlo errors uniformly over the neural class. Both arguments approximate the class by a finite set of networks, apply Bernstein's inequality on this set, and combine the bounds. Tail control then extends the result to unbounded treatment distributions. We conclude with rates for constant-depth ReLU networks.

D.1 The network class and a covering bound

A network in $\mathcal{H}_{k,n} = \mathcal{G}(D_n, p, 1, N_n, M, B_n, k)$ is determined by weight matrices and bias vectors

$$W_1 \in \mathbb{R}^{N_n \times p}, \quad W_2, \dots, W_{D_n} \in \mathbb{R}^{N_n \times N_n}, \quad W_{D_n+1} \in \mathbb{R}^{1 \times N_n},$$
$$b_1, \dots, b_{D_n} \in \mathbb{R}^{N_n}, \quad b_{D_n+1} \in \mathbb{R},$$

all with entries in $[-B_n, B_n]$, where W_1 has at most k nonzero columns. With the ReLU function $\sigma(t) = \max(t, 0)$ applied coordinatewise, put $z_0 = x$ and $z_l = \sigma(W_l z_{l-1} + b_l)$ for $l = 1, \dots, D_n$. The network is

$$h(x) = \max\{-M, \min(M, W_{D_n+1} z_{D_n} + b_{D_n+1})\}.$$

In the main theorem $k = s$, and every choice of the s nonzero columns is allowed. We assume $B_n \geq 1$ and $N_n \geq k$; both can be arranged by enlarging the class. For $K \geq 1$ write $\|h\|_{\infty,K} = \sup\{|h(x)| : \|x\|_\infty \leq K\}$, and let $\mathcal{N}(\varepsilon, \mathcal{H}, \|\cdot\|_{\infty,K})$ be the covering number, i.e., the smallest number of functions $g_1, \dots, g_J \in \mathcal{H}$ such that every $h \in \mathcal{H}$ satisfies $\|h - g_j\|_{\infty,K} \leq \varepsilon$ for some j.

Lemma 6 (Sparse ReLU covering bound)**.** *For $K \geq 1$ and $0 < \varepsilon \leq 1$,*

$$\log \mathcal{N}(\varepsilon, \mathcal{H}_{k,n}, \|\cdot\|_{\infty,K}) \leq \log \binom{p}{k} + C_0(N_n^2 D_n + N_n D_n)\big\{D_n \log(2N_n B_n) + \log(K/\varepsilon)\big\}, \quad (27)$$

where C_0 is an absolute constant. The approximating networks can be taken in $\mathcal{H}_{k,n}$.

For the class of differences $\mathcal{A}_{s,n}$,

$$\log \mathcal{N}(2\varepsilon, \mathcal{A}_{s,n}, \|\cdot\|_{\infty,K}) \leq 2 \log \mathcal{N}(\varepsilon, \mathcal{H}_{s,n}, \|\cdot\|_{\infty,K}).$$

Proof. Write $N = N_n$, $D = D_n$, $B = B_n$ for simplicity. Fix a set S of k input coordinates and consider two networks whose first-layer weights vanish outside the columns in S. Suppose that every pair of corresponding parameters differs by at most δ. Denote the hidden layers of the two networks by z_l and z'_l, and let $\|x\|_\infty \leq K$.

First, we bound how parameter perturbations change the network output. Each coordinate of $W_l z_{l-1}$ is a sum of at most N terms (at most $k \leq N$ terms when $l = 1$), each of absolute value at most $B\|z_{l-1}\|_\infty$. Since $|\sigma(t)| \leq |t|$, we have $\|z_l\|_\infty \leq NB\|z_{l-1}\|_\infty + B$. Induction gives $\|z_l\|_\infty \leq K(2NB)^l$. Let $e_l = \|z_l - z'_l\|_\infty$; then $e_0 = 0$. Since $|\sigma(t) - \sigma(t')| \leq |t - t'|$, we have

$$\begin{aligned} e_l &\leq \|(W_l - W'_l) z_{l-1}\|_\infty + \|W'_l (z_{l-1} - z'_{l-1})\|_\infty + \delta \\ &\leq N\delta \|z_{l-1}\|_\infty + NB e_{l-1} + \delta \leq NB e_{l-1} + 2N\delta K (2NB)^{l-1}. \end{aligned}$$

Unrolling this recursion from $e_0 = 0$ gives $e_l \leq 2N\delta K\, l\, (2NB)^{l-1}$. The output layer is one more affine map, and the truncation at $\pm M$ does not increase distances. Hence, for all $\|x\|_\infty \leq K$,

$$|h(x) - h'(x)| \leq L\delta, \qquad L = 2NK(D+1)(2NB)^D. \quad (28)$$

We now construct a finite cover by discretizing the network parameters. There are $\binom{p}{k}$ choices for the set S, and a network with fewer than k nonzero columns is covered by any S containing them. For a fixed S, the number of free parameters is

$$W = (kN + N) + (D-1)(N^2 + N) + (N+1) \leq N^2 D + 3ND.$$

Replace each parameter by the nearest point of a grid in $[-B, B]$ with spacing δ; the grid has at most $2B/\delta + 1$ points, and the rounded network still belongs to $\mathcal{H}_{k,n}$. By (28), the rounded network is within $L\delta$ of the original network on $\{\|x\|_\infty \le K\}$. Choosing $\delta = \varepsilon/L$ therefore gives an ε-cover, with covering number bounded by

$$\mathcal{N}(\varepsilon, \mathcal{H}_{k,n}, \|\cdot\|_{\infty,K}) \le \binom{p}{k}\Big(\frac{2BL}{\varepsilon} + 1\Big)^W.$$

Since $2BL/\varepsilon \ge 1$, we have

$$\begin{aligned}\log\Big(\frac{2BL}{\varepsilon} + 1\Big) \le \log\frac{4BL}{\varepsilon} &= \log\{8NB(D+1)\} + D\log(2NB) + \log(K/\varepsilon) \\ &\le C_0\{D\log(2NB) + \log(K/\varepsilon)\},\end{aligned}$$

Multiplying by $W \le N^2D + 3ND$ gives the entropy bound in (27).

Finally, if $g_1, \dots, g_J$ approximate $\mathcal{H}_{s,n}$ within ε, then the J^2 differences $g_j - g_{j'}$ approximate $\mathcal{A}_{s,n}$ within 2ε, hence the second conclusion follows. □

D.2 Truncation of the inputs

The covering bound applies on a bounded cube, so we must also control the treatment tails. On the event in Assumption 6, there is a constant σ such that every coordinate ω of X and of a reconstructed draw $\widetilde{X} = \widehat{B}_\parallel \widetilde{V} + \widehat{B}_\perp \widehat{Z}$, $\widetilde{V} \sim \widehat{P}_V(\cdot \mid \widehat{Z})$, satisfies $\Pr(|\omega| > t \mid \mathcal{D}^N) \le 2\exp\{-t^2/(2\sigma^2)\}$. For $r \ge 1$ we truncate at the cube $\{\|x\|_\infty \le \sigma\sqrt{2\log(4p\, n_T^2 r^2)}\}$; for all large n its radius is at least one, and we consider only such n. Taking $t = \sigma\sqrt{2\log(4p\, n_T^2 r^2)}$ in the tail bound and a union bound over the p coordinates, X leaves the cube with probability at most $2p\exp\{-\log(4p\, n_T^2 r^2)\} = 1/(2n_T^2 r^2)$, and the same bound holds for $\widetilde{X}$ conditionally on $\mathcal{D}^N$. Define

$$\mathcal{C}_n(r) = (N_n^2 D_n^2 + N_n D_n)\log(eN_n B_n n_T r) + \log\sum_{j=0}^{s}\binom{p}{j} + \log\log(en_T r), \tag{29}$$

$$\rho_n^2 = \frac{\mathcal{C}_n(1)}{n_T}, \qquad \xi_{\mathrm{MC},n}^2 = \frac{\mathcal{C}_n(m_n)}{n_T m_n}. \tag{30}$$

Thus ρ_n^2 is the covering-complexity bound divided by the sample size, and $\xi_{\mathrm{MC},n}^2$ additionally reflects averaging over Monte Carlo draws. For constant depth, the complexity is of the order of the parameter-count bound used in Lemma 6, up to logarithmic factors. Because N_n, D_n, B_n, and $n_T r$ are at least one, the first term of $\mathcal{C}_n(r)$ is at least one, and hence $\mathcal{C}_n(r) \ge 1$; consequently $\rho_n^2 \ge 1/n_T$ and $\xi_{\mathrm{MC},n}^2 \ge 1/(n_T m_n)$. Apply Lemma 6 with $k = s$, $\varepsilon = 1/(n_T r)$, and K equal to the radius of the cube. Since this radius is at most $C_0\, n_T r$ and $D_n\log(2N_nB_n) + \log(C_0 n_T^2 r^2) \le C_0 D_n \log(eN_nB_n n_T r)$, there is a net $\mathcal{H}^{(r)} \subseteq \mathcal{H}_{s,n}$ such that every $h \in \mathcal{H}_{s,n}$ has some $g \in \mathcal{H}^{(r)}$ with

$$\log|\mathcal{H}^{(r)}| \le c_0\,\mathcal{C}_n(r), \qquad \sup_{\|x\|_\infty \le \sigma\sqrt{2\log(4p\, n_T^2 r^2)}} |h(x) - g(x)| \le \frac{1}{n_T r}, \tag{31}$$

where c_0 depends only on σ and p, and we call such a g an approximant of h. Outside the cube we only know $|h - g| \le 2M$, hence for a random vector Ξ in $\mathbb{R}^p$ with any law, this gives

$$|E\,h(\Xi) - E\,g(\Xi)| \le \frac{1}{n_T r} + 2M\Pr\Big(\|\Xi\|_\infty > \sigma\sqrt{2\log(4p\, n_T^2 r^2)}\Big). \tag{32}$$

Lemma 7 (Bernstein's inequality)**.** *Let $\zeta_1, \dots, \zeta_J$ be independent random variables with mean zero, $|\zeta_j| \le \beta_0$, and $\sum_j \mathrm{Var}(\zeta_j) \le v$. Then, for every $t > 0$,*

$$\Pr\Big(\Big|\sum_{j=1}^{J}\zeta_j\Big| > \sqrt{2vt} + \frac{\beta_0 t}{3}\Big) \le 2e^{-t}.$$

D.3 THE SAMPLING ERROR

We work in the regime $\rho_n \le 1$ and $\xi_{\mathrm{MC},n} \le 1$. The following lemma controls sampling fluctuations in differences between the losses of two networks.

Lemma 8. *Under Assumptions 5–7, suppose* $0 < \alpha_n \le 1$ *and* $\rho_n \le 1$. *Uniformly over* $h, h' \in \mathcal{H}_{s,n}$,

$$|(P_{n_T} - P)(\bar{\ell}_h - \bar{\ell}_{h'})| = O_p\{\rho_n \|h - h'\|_2 + \rho_n^2\}.$$

Proof. Conditionally on $\mathcal{D}^N$, write $n = n_T$, and put $f_{h,h'} = \bar{\ell}_h - \bar{\ell}_{h'}$. For any real numbers a, a' in $[-M, M]$ and $|Y| \le M_Y$, $|(Y-a)^2 - (Y-a')^2| = |a-a'|\,|2Y - a - a'| \le 2(M_Y + M)|a - a'|$, and $|a^2 - a'^2| \le 2M|a - a'|$. Since $\alpha_n \le 1$,

$$|f_{h,h'}(X,Y)| \le C_1\big\{|\widehat{T}_{\widehat{B}}(h-h')(\widehat{Z})| + |(h-h')(X)|\big\}, \qquad C_1 = 2(M_Y + M). \tag{33}$$

In particular, $|f_{h,h'}| \le 4MC_1$. Jensen's inequality and the second-moment condition in Assumption 6 give $\|\widehat{T}_{\widehat{B}}(h-h')\|_2^2 \le E[E\{(h-h')(\widetilde{X})^2 \mid \widehat{Z}, \mathcal{D}^N\}] \le C_V\|h - h'\|_2^2 + n^{-2}$. Squaring (33) and bounding the square of a sum of two terms by twice the sum of their squares, we obtain

$$\mathrm{Var}(f_{h,h'}) \le P f_{h,h'}^2 \le C_2(\|h - h'\|_2^2 + n^{-2}), \qquad C_2 = 2C_1^2(C_V + 1).$$

Let $\mathcal{H}^{(1)}$ be the net in (31) with $r = 1$. For each pair $g, g' \in \mathcal{H}^{(1)}$, the variables $\{[f_{g,g'}(X_i, Y_i) - P f_{g,g'}]/n,\ i \in \mathcal{D}^T\}$, are independent with mean zero, bounded by $8MC_1/n$, and have total variance at most $C_2(\|g - g'\|_2^2 + n^{-2})/n$. Apply Lemma 7 to every pair with $t = 2\log|\mathcal{H}^{(1)}| + u$, and take a union bound over the $|\mathcal{H}^{(1)}|^2$ pairs. With probability at least $1 - 2e^{-u}$, every pair satisfies

$$|(P_n - P)f_{g,g'}| \le \sqrt{2C_2 t/n}\,(\|g - g'\|_2 + n^{-1}) + \frac{8MC_1 t}{3n}, \tag{34}$$

where $\sqrt{x^2 + y^2} \le x + y$ for $x, y \ge 0$ was used. By (31), $t \le (2c_0 + u)\,\mathcal{C}_n(1)$, hence we have $t/n \le (2c_0 + u)\rho_n^2$.

Fix $h, h' \in \mathcal{H}_{s,n}$ with approximants $g, g' \in \mathcal{H}^{(1)}$, and define

$$\tau(X) = \mathbf{1}\Big\{\|X\|_\infty > \sigma\sqrt{2\log(4pn^2)}\Big\} + \Pr\Big(\|\widetilde{X}\|_\infty > \sigma\sqrt{2\log(4pn^2)} \;\Big|\; \widehat{Z}, \mathcal{D}^N\Big).$$

By (32), applied to the point mass at X and to the fitted conditional law of $\widetilde{X}$ given $\widehat{Z}$, $|h(X) - g(X)|$ and $|\widehat{T}_{\widehat{B}}h(\widehat{Z}) - \widehat{T}_{\widehat{B}}g(\widehat{Z})|$ are both at most $1/n + 2M\tau(X)$. The same holds for h', g', and the argument for (33) gives $|f_{h,h'} - f_{g,g'}| \le 4C_1\{1/n + 2M\tau(X)\}$. The sub-Gaussian tail and a union bound over the p coordinates give $P\tau \le 2 \cdot 2p\exp\{-\log(4pn^2)\} = 1/n^2$. Markov's inequality then gives $P_n\tau \le 1/n$ with probability at least $1 - 1/n$. On that event,

$$|(P_n - P)(f_{h,h'} - f_{g,g'})| \le 4C_1\{2/n + 2M(P_n\tau + P\tau)\} \le C_0/n.$$

Similarly, $\|h - g\|_2^2 \le 1/n^2 + 4M^2\Pr\{\|X\|_\infty > \sigma\sqrt{2\log(4pn^2)}\} \le C_0/n^2$, and the triangle inequality gives $\|g - g'\|_2 \le \|h - h'\|_2 + C_0/n$.

Combine (34) with the two approximation bounds just obtained, $|(P_n - P)(f_{h,h'} - f_{g,g'})| \le C_0/n$ and $\|g - g'\|_2 \le \|h - h'\|_2 + C_0/n$, and use $1/n \le \rho_n^2$ and $1/n \le \rho_n$. With probability at least $1 - 2e^{-u} - 1/n$, every pair $h, h' \in \mathcal{H}_{s,n}$ satisfies

$$|(P_n - P)f_{h,h'}| \le C_u\{\rho_n\|h - h'\|_2 + \rho_n^2\},$$

where C_u depends on u but not on n. Given $\eta > 0$, choose u with $2e^{-u} \le \eta/2$, which concludes the proof. □

D.4 THE MONTE CARLO ERROR

Lemma 9 (Double-Monte-Carlo process)**.** *Under Assumptions 5–7, suppose* $\rho_n \le 1$ *and* $\xi_{\mathrm{MC},n} \le 1$. *The two conditionally independent batches satisfy*

$$\Big|\{\widehat{Q}_{\alpha_n}(h) - P_{n_T}\bar{\ell}_h\} - \{\widehat{Q}_{\alpha_n}(h') - P_{n_T}\bar{\ell}_{h'}\}\Big| = O_p\{\xi_{\mathrm{MC},n}\|h - h'\|_2 + \xi_{\mathrm{MC},n}^2 + \rho_n\xi_{\mathrm{MC},n}\} \tag{35}$$

uniformly over $h, h' \in \mathcal{H}_{s,n}$.

Proof. The proof has four parts. We first write the Monte Carlo error as an average of conditionally centred terms, and split its increment between two networks into six sums. We then bound each sum by Bernstein's inequality. This bound involves an empirical second moment, which we next compare with $\|h - h'\|_2^2$. Finally, we combine the bounds over a finite net of networks and pass from the net to the whole class.

Given $\mathcal{D}^N$, for each value z of $\widehat{Z}$ let $\widehat{P}(\cdot \mid z)$ be the law of $\widehat{B}_\parallel \widetilde{V} + \widehat{B}_\perp z$ when $\widetilde{V} \sim \widehat{P}_V(\cdot \mid z)$, that is, the law of a reconstructed draw given $\widehat{Z} = z$. Recall the definition in (4), $(\widehat{T}_{\widehat{B}}\phi)(z) = \int \phi \, d\widehat{P}(\cdot \mid z)$ for every bounded function ϕ. By Assumption 6, given $\mathcal{D}^N \cup \mathcal{D}^T$ the draws $\widetilde{X}_{i\ell}^{(b)} = \widehat{B}_\parallel \widetilde{V}_{i\ell}^{(b)} + \widehat{B}_\perp \widehat{Z}_i$ are independent, and $\widetilde{X}_{i\ell}^{(b)}$ has law $\widehat{P}(\cdot \mid \widehat{Z}_i)$. For a bounded function ϕ, define the conditional mean, the Monte Carlo noise of batch b, and the conditional second moment by

$$\mu_i(\phi) = (\widehat{T}_{\widehat{B}}\phi)(\widehat{Z}_i), \qquad e_{i,b}(\phi) = A_{i,b}(\phi) - \mu_i(\phi), \qquad q_\phi(z) = \int \phi^2 \, d\widehat{P}(\cdot \mid z).$$

Both μ_i and $e_{i,b}$ are linear in ϕ, and Jensen's inequality gives $\mu_i(\phi)^2 \le q_\phi(\widehat{Z}_i)$.

We begin with an exact formula for the Monte Carlo error. The penalty $\alpha_n h(X_i)^2$ appears in both $\widehat{Q}_{\alpha_n}(h)$ and $P_{n_T}\bar{\ell}_h$, and therefore cancels. Writing $Y_i - A_{i,b}(h) = \{Y_i - \mu_i(h)\} - e_{i,b}(h)$ and multiplying out the two batches, we obtain

$$\widehat{Q}_{\alpha_n}(h) - P_{n_T}\bar{\ell}_h = \frac{1}{n_T}\sum_i \Delta_i(h), \quad \Delta_i(h) = -\{Y_i - \mu_i(h)\}\{e_{i,1}(h) + e_{i,2}(h)\} + e_{i,1}(h)e_{i,2}(h).$$

This expansion separates the linear and quadratic terms in the sampling noise. Now let $g, g' \in \mathcal{H}_{s,n}$ and $\delta = g - g'$. Because μ_i and $e_{i,b}$ are linear, a direct computation gives

$$\begin{aligned}\Delta_i(g) - \Delta_i(g') = & -\{Y_i - \mu_i(g)\}\{e_{i,1}(\delta) + e_{i,2}(\delta)\} + \mu_i(\delta)\{e_{i,1}(g') + e_{i,2}(g')\} \\ & + e_{i,1}(\delta)\, e_{i,2}(g) + e_{i,1}(g')\, e_{i,2}(\delta).\end{aligned}$$

Each of the six terms on the right is a coefficient c_i times the noise $e_{i,b}(\phi)$ of one batch. In the first two terms $c_i = -\{Y_i - \mu_i(g)\}$ and $\phi = \delta$; in the next two $c_i = \mu_i(\delta)$ and $\phi = g'$; in the last two c_i is the noise of the other batch, $e_{i,2}(g)$ or $e_{i,1}(g')$, and $\phi = \delta$. Thus every term contains δ, either in the noise or in the coefficient, and this is the source of the factor $\|h - h'\|_2$ in the lemma.

We next bound the average over i of one such term,

$$\frac{1}{n_T}\sum_i c_i\, e_{i,b}(\phi) = \frac{1}{n_T m_n}\sum_i \sum_{\ell=1}^{m_n} c_i\{\phi(\widetilde{X}_{i\ell}^{(b)}) - \mu_i(\phi)\}.$$

Condition on $\mathcal{D}^N \cup \mathcal{D}^T$ and, for the last two terms, also on the other batch. Because the two batches are conditionally independent, the coefficients c_i are then fixed, while the draws of batch b remain independent with law $\widehat{P}(\cdot \mid \widehat{Z}_i)$. Hence the $n_T m_n$ summands are independent with mean zero, and they are bounded by $4M(M_Y + 2M)$, since $|c_i| \le M_Y + 2M$ and $|\phi - \mu_i(\phi)| \le 4M$. Moreover, their variances are controlled by q_δ in all six cases. If $\phi = \delta$, the variance is at most $c_i^2 q_\delta(\widehat{Z}_i) \le (M_Y + 2M)^2 q_\delta(\widehat{Z}_i)$; if $\phi = g'$, then $c_i^2 = \mu_i(\delta)^2 \le q_\delta(\widehat{Z}_i)$ and the variance of $g'(\widetilde{X}_{i\ell}^{(b)})$ is at most M^2. The total variance is therefore at most $(M_Y + 2M)^2\, n_T m_n\, P_{n_T} q_\delta$, and Lemma 7 gives, with conditional probability at least $1 - 2e^{-t}$,

$$\Big|\frac{1}{n_T}\sum_i c_i\, e_{i,b}(\phi)\Big| \le (M_Y + 2M)\sqrt{\frac{2P_{n_T} q_\delta\, t}{n_T m_n}} + \frac{4M(M_Y + 2M)\, t}{3 n_T m_n}. \tag{36}$$

The bound (36) involves the empirical quantity $P_{n_T} q_\delta$, whereas the lemma is stated in terms of $\|\delta\|_2$. We therefore show that, for a fixed $u > 0$, with probability at least $1 - 2e^{-u} - 1/n_T$,

$$P_{n_T} q_\delta \le C_u(\|\delta\|_2^2 + \rho_n^2) \qquad \text{for all } \delta \in \mathcal{A}_{s,n}, \tag{37}$$

where C_u depends on u but not on n, and may change from one display to the next. The argument parallels Lemma 8: we control the mean, then a finite net, and then the whole class. The second-moment bound in Assumption 6 gives $P q_\delta \le C_V \|\delta\|_2^2 + n_T^{-2}$. On the net $\mathcal{H}^{(1)}$, given $\mathcal{D}^N$, the

variables $q_\delta(\widehat{Z}_i)$ are i.i.d., lie in $[0, 4M^2]$, and have variance at most $4M^2 P q_\delta$. By Lemma 7 with $t = 2\log|\mathcal{H}^{(1)}| + u$ and a union bound over the pairs $g, g' \in \mathcal{H}^{(1)}$, with probability at least $1 - 2e^{-u}$,

$$P_{n_T} q_\delta \le P q_\delta + \sqrt{8M^2 P q_\delta \, t/n_T} + \frac{4M^2 t}{3n_T} \le 2P q_\delta + \frac{4M^2 t}{n_T}.$$

Since $t/n_T \le (2c_0 + u)\rho_n^2$ by the proof of Lemma 8 and $n_T^{-2} \le \rho_n^2$, (37) holds on the net.

We next extend (37) to arbitrary network differences $\delta = h - h'$, where $h, h' \in \mathcal{H}_{s,n}$. Let $g, g' \in \mathcal{H}^{(1)}$ be approximants of h, h', and put $\widetilde{\delta} = g - g'$. By (31), $|\delta - \widetilde{\delta}| \le 2/n_T$ on the cube $\{\|x\|_\infty \le \sigma\sqrt{2\log(4pn_T^2)}\}$. Outside this cube, bounded outputs give $|\delta - \widetilde{\delta}| \le 4M$. Since $|\delta^2 - \widetilde{\delta}^2| \le 4M|\delta - \widetilde{\delta}|$, it follows that $|q_\delta(\widehat{Z}_i) - q_{\widetilde{\delta}}(\widehat{Z}_i)| \le 4M(2/n_T + 4M\pi_i)$, where π_i is the conditional probability, given $\widehat{Z}_i$ and $\mathcal{D}^N$, that a reconstructed draw leaves the cube. By the sub-Gaussian tail in Assumption 6 and a union bound over the p coordinates, a reconstructed draw leaves the cube with probability at most $2p\exp\{-\log(4pn_T^2)\} = 1/(2n_T^2)$, hence $n_T^{-1}\sum_i \pi_i$ has mean at most $1/(2n_T^2)$, and Markov's inequality gives $n_T^{-1}\sum_i \pi_i \le 1/n_T$ with probability at least $1 - 1/(2n_T)$. On this event, $|P_{n_T} q_\delta - P_{n_T} q_{\widetilde{\delta}}| \le C_0/n_T$. In addition, $\|\widetilde{\delta}\|_2 \le \|\delta\|_2 + C_0/n_T$ by the bound $\|h - g\|_2 \le C_0/n_T$ in the proof of Lemma 8. Because $1/n_T \le \rho_n^2$, both errors are absorbed, and (37) holds for all $\delta \in \mathcal{A}_{s,n}$.

We now bound the Monte Carlo error on a finite net. Let $\mathcal{H}^{(m_n)}$ be the net in (31) with $r = m_n$; its precision $1/(n_T m_n)$ matches the Monte Carlo scale. Apply (36) to the six terms of every pair $g, g' \in \mathcal{H}^{(m_n)}$, with $t = 2\log|\mathcal{H}^{(m_n)}| + \log 6 + u$. A union bound over these $6|\mathcal{H}^{(m_n)}|^2$ terms shows that, with probability at least $1 - 2e^{-u}$, all of them satisfy (36). Moreover, since $\mathcal{C}_n(m_n) \ge 1$, (31) gives $t \le (2c_0 + \log 6 + u)\,\mathcal{C}_n(m_n)$, that is, $t/(n_T m_n) \le C_u \xi_{\mathrm{MC},n}^2$. Substituting this and (37) into (36), and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, we find that each term is at most $C_u\{\xi_{\mathrm{MC},n}(\|\delta\|_2 + \rho_n) + \xi_{\mathrm{MC},n}^2\}$. Adding the six terms, we conclude that, with probability at least $1 - 4e^{-u} - 1/n_T$, every pair $g, g' \in \mathcal{H}^{(m_n)}$ satisfies

$$\Big|\frac{1}{n_T}\sum_i \{\Delta_i(g) - \Delta_i(g')\}\Big| \le C_u\big\{\xi_{\mathrm{MC},n}(\|g - g'\|_2 + \rho_n) + \xi_{\mathrm{MC},n}^2\big\}. \tag{38}$$

Finally, we pass from the net to the whole class. Fix $h, h' \in \mathcal{H}_{s,n}$, and let $g, g' \in \mathcal{H}^{(m_n)}$ be their approximants. We show that replacing (h, h') by (g, g') changes each side of (38) by at most $C_0/(n_T m_n)$. Since $1/(n_T m_n) \le \xi_{\mathrm{MC},n}^2$ and $\xi_{\mathrm{MC},n} \le 1$, such changes are absorbed into $C_u \xi_{\mathrm{MC},n}^2$.

All bounds below use the larger cube $\{\|x\|_\infty \le \sigma\sqrt{2\log(4pn_T^2 m_n^2)}\}$. On this cube, $|h - g| \le 1/(n_T m_n)$ by (31) with $r = m_n$; outside it, we have $|h - g| \le 2M$. The sub-Gaussian tail and a union bound over the p coordinates, as in the truncation of the inputs, show that X and a reconstructed draw each leave this cube with probability at most $2p\exp\{-\log(4pn_T^2 m_n^2)\} = 1/(2n_T^2 m_n^2)$. Two events control what happens outside the cube. The first is the event Ω that all $2n_T m_n$ reconstructed draws lie in the cube; a union bound gives $\Pr(\Omega^c) \le 1/(n_T m_n)$. The second concerns the conditional probability π_i, given $\widehat{Z}_i$ and $\mathcal{D}^N$, that a reconstructed draw leaves the cube. The average $n_T^{-1}\sum_i \pi_i$ has mean at most $1/(2n_T^2 m_n^2)$, and Markov's inequality gives, with probability at least $1 - 1/(2n_T m_n)$, $\frac{1}{n_T}\sum_i \pi_i \le \frac{1}{n_T m_n}$.

We first bound the change in the left side of (38). The term $\Delta_i(h)$ depends on h only through the batch averages $A_{i,b}(h)$ and the conditional mean $\mu_i(h)$. On Ω, every draw lies in the cube, and hence the batch averages change by at most $1/(n_T m_n)$. The conditional mean also averages over draws outside the cube, and (32) accounts for them. Together,

$$|A_{i,b}(h) - A_{i,b}(g)| \le \frac{1}{n_T m_n} \quad (b = 1, 2), \qquad |\mu_i(h) - \mu_i(g)| \le \frac{1}{n_T m_n} + 2M\pi_i.$$

Moreover, $\Delta_i(h)$ is a sum of products of the factors $Y_i - \mu_i(h)$ and $e_{i,b}(h)$, each bounded by $M_Y + 2M$. Applying $|xy - x'y'| \le |x|\,|y - y'| + |y'|\,|x - x'|$ to each product, we obtain

$$|\Delta_i(h) - \Delta_i(g)| \le C_0\Big\{|\mu_i(h) - \mu_i(g)| + \sum_{b=1}^{2} |A_{i,b}(h) - A_{i,b}(g)|\Big\} \le C_0\Big(\frac{1}{n_T m_n} + \pi_i\Big).$$

Averaging over i and using the bound on $n_T^{-1}\sum_i \pi_i$, we conclude that

$$\Big|\frac{1}{n_T}\sum_i\{\Delta_i(h)-\Delta_i(h')\}-\frac{1}{n_T}\sum_i\{\Delta_i(g)-\Delta_i(g')\}\Big| \le \frac{C_0}{n_T m_n}.$$

We next bound the change in the right side of (38). Because $|h-g|$ is at most $1/(n_T m_n)$ on the cube and at most $2M$ outside it, and X leaves the cube with probability at most $1/(2n_T^2 m_n^2)$,

$$\|h-g\|_2^2 \le \frac{1}{(n_T m_n)^2}+\frac{4M^2}{2n_T^2 m_n^2} \le \frac{C_0}{(n_T m_n)^2}.$$

The same bound holds for $h'-g'$, and the triangle inequality gives $\|g-g'\|_2 \le \|h-h'\|_2 + C_0/(n_T m_n)$.

Combining the two changes with (38), we obtain

$$\begin{aligned}\Big|\frac{1}{n_T}\sum_i\{\Delta_i(h)-\Delta_i(h')\}\Big| &\le C_u\Big\{\xi_{\mathrm{MC},n}\Big(\|h-h'\|_2+\frac{C_0}{n_T m_n}+\rho_n\Big)+\xi_{\mathrm{MC},n}^2\Big\}+\frac{C_0}{n_T m_n}\\ &\le C_u\big\{\xi_{\mathrm{MC},n}\|h-h'\|_2+\rho_n\xi_{\mathrm{MC},n}+\xi_{\mathrm{MC},n}^2\big\},\end{aligned}$$

where the second inequality uses $1/(n_T m_n) \le \xi_{\mathrm{MC},n}^2$ and $\xi_{\mathrm{MC},n} \le 1$. This holds for all $h, h' \in \mathcal{H}_{s,n}$ with probability at least $1-4e^{-u}-1/n_T-3/(2n_T m_n)$, where the last term accounts for Ω^c and for the bound on $n_T^{-1}\sum_i \pi_i$. Given $\eta>0$, choose u with $4e^{-u} \le \eta/2$; for all large n, the remaining terms sum to at most $\eta/2$, which proves (35). □

D.5 Rates for constant-depth ReLU networks

We now specialize Theorem 2 to sparse ReLU networks of constant depth. Their complexity grows quadratically in width N_n, up to logarithmic factors, while approximation error decreases with width. Under the stated nuisance and approximation bounds, balancing these errors with regularization bias gives an explicit rate.

Corollary 2 (Rate for constant-depth sparse ReLU networks)**.** *Suppose that the assumptions of Theorem 2 hold, that the depth D_n is constant, and that $1 \le m_n \le n_T$. Suppose additionally that $0<\beta_{\mathrm{iv}} \le 1$, so that $\nu=\beta_{\mathrm{iv}} \in (0,1]$. Suppose also that, for some $\beta_h>0$,*

$$\log B_n \le C_0 \log n_T, \qquad n_T \le C_0 n_N, \qquad \delta_{F,n}^2 \le \frac{C_0(\log n_T)^2}{n_N}, \qquad \epsilon_{H,n}^2 \le C_0 N_n^{-4\beta_h/s}.$$

Then, for $\sqrt{\log n_T} \le N_n \le n_T$ and $0<\alpha_n \le 1$,

$$\|\widehat{h}-h_0\|_2^2 = O_p\Big\{\frac{N_n^2 \log n_T}{n_T \alpha_n^2}+\frac{N_n^{-4\beta_h/s}}{\alpha_n}+\alpha_n^{\nu}\Big\}. \tag{39}$$

In particular, the choices

$$\alpha_n=\Big(\frac{\log n_T}{n_T}\Big)^{2\beta_h/\{\nu(2\beta_h+s)+4\beta_h+s\}}, \qquad N_n=\left\lceil\Big(\frac{\alpha_n n_T}{\log n_T}\Big)^{s/(4\beta_h+2s)}\right\rceil$$

give

$$\|\widehat{h}-h_0\|_2^2 = O_p\Big\{\Big(\frac{\log n_T}{n_T}\Big)^{2\beta_h\nu/\{\nu(2\beta_h+s)+4\beta_h+s\}}\Big\}.$$

Remark 2. The displayed explicit-rate specialization is stated for $0<\beta_{\mathrm{iv}} \le 1$, where $\nu=\beta_{\mathrm{iv}}$ and the bias exponent is not saturated; Theorem 2 remains stated for general $\beta_{\mathrm{iv}}>0$, with $\nu=\min\{\beta_{\mathrm{iv}},2\}$. The balance below takes $\epsilon_{H,n} \asymp \alpha_n^{(1+\nu)/2}$, which is compatible with Assumption 7: its second inequality gives $\alpha_n\|h_n^\circ - h_{\alpha_n,0}\|_2^2 \le C_H\epsilon_{H,n}^2$, and together with $\|h_n^\circ-h_0\|_2 \le \epsilon_{H,n}$ and the Tikhonov bias of order $\alpha_n^{\nu/2}$ it requires only $\epsilon_{H,n} \gtrsim \alpha_n^{(1+\nu)/2}$. The condition on $\epsilon_{H,n}$ is the approximation rate one expects when h_0 is β_h-smooth in its s active coordinates: constant-depth ReLU networks of width N approximate such a function on a cube with uniform error of order $N^{-2\beta_h/s}$, up to logarithmic factors (Lu et al., 2021). We state it as a condition because $\epsilon_{H,n}$ is

an $L_2(P_X)$ error over an unbounded design. For the same reason, the corollary keeps the second inequality of Assumption 7 and Assumption 8 as assumptions; smoothness of h_0 alone gives neither. For the Gaussian sampler of Appendix E, $\delta_{F,n} = n_N^{-1/2} \log n_T$; the sample-size condition holds when the nuisance sample is at least a fixed positive fraction as large as the structural sample. Under these conditions, the nuisance and Monte Carlo errors do not affect the rate, because d_n^2 is of the same order as ρ_n^2. Compared with the regression rate $(\log n_T / n_T)^{2\beta_h/(2\beta_h+s)}$, the exponent is multiplied by $\nu(2\beta_h + s)/\{\nu(2\beta_h + s) + 4\beta_h + s\} < 1$. The resulting upper bound is slower, reflecting error amplification in the inverse problem under our assumptions.

Proof of Corollary 2. Under $0 < \beta_{\mathrm{iv}} \le 1$, the bias exponent in Theorem 2 is $\nu = \beta_{\mathrm{iv}}$. The proof has two parts. We first bound d_n in terms of the width and insert the bound into Theorem 2, which gives (39). We then choose N_n and α_n to balance the three terms of (39).

Consider first the complexity $\mathcal{C}_n(r)$ in (29) for $r \in \{1, m_n\}$. Its first factor satisfies $N_n^2 D_n^2 + N_n D_n \le 2 D_n^2 N_n^2$. Since $N_n \le n_T$, $\log B_n \le C_0 \log n_T$ and $r \le m_n \le n_T$, each logarithm in (29) is at most a constant multiple of $\log n_T$; moreover, $\log \sum_{j=0}^{s} \binom{p}{j} \le p \log 2$. Hence, for all large n,

$$\mathcal{C}_n(1) \le C_0 N_n^2 \log n_T, \qquad \mathcal{C}_n(m_n) \le C_0 N_n^2 \log n_T.$$

By (30), it follows that

$$\rho_n^2 \le C_0 \frac{N_n^2 \log n_T}{n_T}, \qquad \xi_{\mathrm{MC},n}^2 \le C_0 \frac{N_n^2 \log n_T}{n_T m_n} \le C_0 \frac{N_n^2 \log n_T}{n_T}.$$

The other two terms in d_n^2 are bounded by the same scale. Indeed, $n_N^{-1} \le C_0/n_T \le C_0 N_n^2 \log n_T / n_T$, and $\delta_{F,n}^2 \le C_0 (\log n_T)^2 / n_N \le C_0 (\log n_T)^2 / n_T \le C_0 N_n^2 \log n_T / n_T$ because $N_n^2 \ge \log n_T$. Therefore $d_n^2 \le C_0 \frac{N_n^2 \log n_T}{n_T}$. Inserting this bound and $\epsilon_{H,n}^2 \le C_0 N_n^{-4\beta_h/s}$ into Theorem 2 gives (39).

We next choose the width for a given α_n. In (39), the first term increases with N_n and the second decreases, and the two are equal when

$$N_n^{2+4\beta_h/s} = \frac{\alpha_n n_T}{\log n_T}, \qquad \text{that is,} \qquad N_n = \Big(\frac{\alpha_n n_T}{\log n_T}\Big)^{s/(4\beta_h+2s)}.$$

Rounding up to an integer changes each term by at most a constant factor. For this width, the second term, and hence also the first, equals

$$\frac{N_n^{-4\beta_h/s}}{\alpha_n} = \frac{1}{\alpha_n}\Big(\frac{\alpha_n n_T}{\log n_T}\Big)^{-2\beta_h/(2\beta_h+s)} = \Big(\frac{\log n_T}{n_T}\Big)^{2\beta_h/(2\beta_h+s)} \alpha_n^{-(4\beta_h+s)/(2\beta_h+s)}.$$

Finally, we choose α_n to balance this term against the bias term α_n^ν. The two are equal when

$$\alpha_n^{\nu+(4\beta_h+s)/(2\beta_h+s)} = \Big(\frac{\log n_T}{n_T}\Big)^{2\beta_h/(2\beta_h+s)},$$

which is the choice of α_n in the corollary. This α_n lies in $(0, 1]$. Moreover, its exponent is below one, and hence $\alpha_n n_T / \log n_T \to \infty$; consequently $\sqrt{\log n_T} \le N_n \le n_T$ for all large n, because N_n grows like a positive power of $n_T / \log n_T$. All three terms of (39) are then of order α_n^ν, which is the stated rate. □

E Conditional Sampler Verification

We verify the sampler conditions for a Gaussian maximum likelihood fit. The algorithm can use other conditional models, but applying Theorem 2 to them requires corresponding accuracy and moment bounds.

E.1 The Gaussian sampler in closed form

Suppose that $X \sim N(m_X, \Sigma_X)$ with Σ_X positive definite. In model (1) this holds when U and ϵ are Gaussian, with $\Sigma_X = \Lambda\Lambda^T + D$. Let $(B_\parallel, B_\perp)$ be any pair of orthonormal bases of a q-dimensional

subspace and of its orthogonal complement, and put $V = B_\parallel^T X$ and $Z = B_\perp^T X$. Note that (V, Z) is a linear function of X and is therefore jointly Gaussian. Its conditional law is

$$V \mid Z = z \sim N(Az + b, \Sigma), \tag{40}$$

where $\theta = (A, b, \Sigma)$ is the following function of the basis and of the first two moments of X:

$$\begin{aligned} A &= B_\parallel^T \Sigma_X B_\perp \, (B_\perp^T \Sigma_X B_\perp)^{-1}, \qquad b = B_\parallel^T m_X - A B_\perp^T m_X, \\ \Sigma &= B_\parallel^T \Sigma_X B_\parallel - A B_\perp^T \Sigma_X B_\parallel. \end{aligned} \tag{41}$$

We write $\theta = \Theta(B; m_X, \Sigma_X)$. The matrix Σ is a Schur complement of $[B_\parallel, B_\perp]^T \Sigma_X [B_\parallel, B_\perp]$, and hence its eigenvalues lie between the smallest and the largest eigenvalue of Σ_X.

The Gaussian maximum likelihood estimator of θ from the pairs $(\widehat{V}_i, \widehat{Z}_i)$, $i \in \mathcal{D}^N$, is the least-squares regression of $\widehat{V}_i$ on $(1, \widehat{Z}_i)$ together with the residual covariance matrix. In terms of the sample mean $\bar{X}$ and the sample covariance $\widehat{\Sigma}_X$ (with divisor n_N) of $\mathcal{D}^N$, the least-squares formulas are exactly (41) with (m_X, Σ_X) replaced by $(\bar{X}, \widehat{\Sigma}_X)$ and with the estimated basis. Thus

$$\widehat{\theta} = \Theta(\widehat{B}; \bar{X}, \widehat{\Sigma}_X), \qquad \theta_0 = \Theta(\widehat{B}; m_X, \Sigma_X), \tag{42}$$

where θ_0 is the parameter of the true conditional law of $\widehat{V}$ given $\widehat{Z}$ for a new observation, with the estimated basis held fixed. Both parameters use the estimated basis $\widehat{B}$, so their comparison isolates sampler fitting error. Lemma 5 separately accounts for loading-space estimation. Sampling from $N(\widehat{A}z + \widehat{b}, \widehat{\Sigma})$ defines $\widehat{P}_V(\cdot \mid z)$.

Let $p_\theta(v \mid z)$ be the density of $N(Az + b, \Sigma)$, and define the likelihood ratio and the score

$$R_\theta(v, z) = \frac{p_\theta(v \mid z)}{p_{\theta_0}(v \mid z)}, \qquad s_\theta(v, z) = \nabla_\theta \log p_\theta(v \mid z).$$

With $e = v - Az - b$, the components of the score with respect to b, A, and Σ are $\Sigma^{-1} e$, $\Sigma^{-1} e z^T$, and $\frac{1}{2}(\Sigma^{-1} e e^T \Sigma^{-1} - \Sigma^{-1})$; the score is therefore a polynomial of degree two in (v, z). Below, E_{θ_0} denotes expectation over a new observation X, conditionally on $\mathcal{D}^N$, with $V = \widehat{B}_\parallel^T X$ and $Z = \widehat{B}_\perp^T X$.

Throughout this appendix, $X \sim N(m_X, \Sigma_X)$ with Σ_X positive definite, and the sampler is the Gaussian maximum likelihood estimator (42).

Proposition 2. *If $\log n_T = o(n_N^{1/2})$, then*

$$\|\widehat{\theta} - \theta_0\| = O_p(n_N^{-1/2}), \qquad \mathcal{E}_{F,n} = O_p(n_N^{-1/2} \log n_T), \tag{43}$$

and the second-moment bound in Assumption 6 holds with $C_V = 2$. Hence Assumption 6 holds with $\delta_{F,n} = n_N^{-1/2} \log n_T$. The integrated squared Hellinger distance between the fitted and true conditional Gaussian distributions is $O_p(n_N^{-1})$.

Proof. Assumption 6 has three parts: the operator bound (13), the second-moment bound, and sub-Gaussian coordinates; the independence of the draws holds by construction. All three follow from one fact, namely that $\widehat{\theta}$ is within $O_p(n_N^{-1/2})$ of θ_0. We first establish the parameter rate, using the fact that $\widehat{\theta}$ and θ_0 are the same function $\Theta(\widehat{B}; \cdot, \cdot)$ evaluated at the sample and at the population moments. We then show that, on a cube of radius of order $\sqrt{\log n_T}$, the likelihood ratio $R_{\widehat{\theta}}$ is within $O_p(n_N^{-1/2} \log n_T)$ of one, because the score is a polynomial of degree two. Inside the cube this gives the relative parts of the operator and second-moment bounds; outside the cube, the sub-Gaussian tails give the additive parts. The Hellinger bound, which is not needed for Assumption 6, is proved last. Throughout, $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and the largest eigenvalue of Σ_X.

We first show that $\|\widehat{\theta} - \theta_0\| = O_p(n_N^{-1/2})$. The sample moments converge at the root-n_N rate: the entries of $\bar{X}$ and $\widehat{\Sigma}_X$ are smooth functions of averages of variables with finite variance, and Chebyshev's inequality gives

$$\|\bar{X} - m_X\| + \|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}} = O_p(n_N^{-1/2}).$$

It therefore suffices to show that $\Theta(B;\cdot,\cdot)$ is Lipschitz in the moments, with a constant that does not depend on the basis B. The uniformity in B matters, because $\widehat{B}$ is computed from the same sample $\mathcal{D}^N$ as the moments. Work on the event $\|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}} \le \lambda_{\min}/2$ and $\|\bar{X} - m_X\| \le 1$, whose probability tends to one. Put $G = B_\perp^T \Sigma_X B_\perp$ and $\widehat{G} = B_\perp^T \widehat{\Sigma}_X B_\perp$. Because $B_\perp$ has orthonormal columns, the eigenvalues of G are at least $\lambda_{\min}$, and those of $\widehat{G}$ are at least $\lambda_{\min}/2$. Using $\widehat{G}^{-1} - G^{-1} = \widehat{G}^{-1}(G - \widehat{G})G^{-1}$, we obtain

$$\widehat{A} - A_0 = B_\parallel^T(\widehat{\Sigma}_X - \Sigma_X)B_\perp \widehat{G}^{-1} + B_\parallel^T \Sigma_X B_\perp\, \widehat{G}^{-1}(G - \widehat{G})G^{-1},$$

and therefore

$$\|\widehat{A} - A_0\|_{\mathrm{op}} \le \Big(\frac{2}{\lambda_{\min}} + \frac{2\lambda_{\max}}{\lambda_{\min}^2}\Big)\|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}}.$$

The formulas for b and Σ in (41) are sums and products of A and of the moments, all of which are bounded on this event; hence the same argument gives $\|\widehat{b} - b_0\| + \|\widehat{\Sigma} - \Sigma_0\|_{\mathrm{op}} \le C_0(\|\bar{X} - m_X\| + \|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}})$. These bounds hold for every orthonormal basis, in particular for $\widehat{B}$, whichever way $\widehat{B}$ was computed from $\mathcal{D}^N$, hence $\|\widehat{\theta} - \theta_0\| = O_p(n_N^{-1/2})$. We also record that, on the same event, the eigenvalues of $\widehat{\Sigma}$ and Σ_0 lie in $[\lambda_{\min}/2, \lambda_{\max} + \lambda_{\min}/2]$, because both matrices are Schur complements of matrices with eigenvalues in this interval, and that $\widehat{A}$, $\widehat{b}$, A_0, and b_0 are bounded by a constant.

We next establish the sub-Gaussian coordinates and fix the cube. The vector X is Gaussian. A reconstructed draw $\widetilde{X} = \widehat{B}_\parallel \widetilde{V} + \widehat{B}_\perp \widehat{Z}$ is a linear function of the Gaussian vector $\widehat{Z}$ and of independent Gaussian noise with covariance $\widehat{\Sigma}$; hence it is Gaussian as well, and, on the high-probability event above, its mean and covariance are within a fixed distance of m_X and Σ_X. Consequently there is a constant σ, depending only on m_X and Σ_X, such that every coordinate ω of X and of $\widetilde{X}$ satisfies $\Pr(|\omega| > t \mid \mathcal{D}^N) \le 2\exp\{-t^2/(2\sigma^2)\}$ for all $t > 0$. This is the sub-Gaussian part of Assumption 6. Put

$$K_n = \sigma\sqrt{2\log(32pM^2n_T^2)}, \qquad \mathcal{K}_n = \{x \in \mathbb{R}^p : \|x\|_\infty \le K_n\}.$$

A union bound over the p coordinates gives $\Pr(X \notin \mathcal{K}_n) \le 2p\exp\{-K_n^2/(2\sigma^2)\} = 1/(16M^2n_T^2)$, and the same bound for $\Pr(\widetilde{X} \notin \mathcal{K}_n \mid \mathcal{D}^N)$.

We now bound the likelihood ratio on the cube. For $x \in \mathcal{K}_n$ write $v = \widehat{B}_\parallel^T x$ and $z = \widehat{B}_\perp^T x$; then $\|v\|^2 + \|z\|^2 = \|x\|_2^2 \le pK_n^2$. Let $\theta_t = \theta_0 + t(\widehat{\theta} - \theta_0)$ for $0 \le t \le 1$. On the same high-probability event, each Σ_t is a convex combination of Σ_0 and $\widehat{\Sigma}$ and hence has eigenvalues in $[\lambda_{\min}/2, \lambda_{\max} + \lambda_{\min}/2]$, and A_t, b_t are bounded. Therefore $\|v - A_t z - b_t\| \le C_0(1 + K_n)$ on $\mathcal{K}_n$, and the three components of the score give $\|s_{\theta_t}(v,z)\| \le C_0(1 + K_n^2)$ for all $x \in \mathcal{K}_n$ and $0 \le t \le 1$. Applying the mean value theorem to $t \mapsto \log p_{\theta_t}(v \mid z)$, we obtain, for all $x \in \mathcal{K}_n$,

$$|\log R_{\widehat{\theta}}(v,z)| \le \|\widehat{\theta} - \theta_0\| \sup_{0\le t\le 1} \|s_{\theta_t}(v,z)\| \le \gamma_n,$$

where $\gamma_n = C_0\|\widehat{\theta} - \theta_0\|(1 + K_n^2) = O_p(n_N^{-1/2}\log n_T)$. Since $\log n_T = o(n_N^{1/2})$, the event $\gamma_n \le 1/2$ has probability tending to one, and we work on it. Because $e^a - 1 \le 2a$ and $1 - e^{-a} \le a$ for $0 \le a \le 1/2$,

$$\sup_{x\in\mathcal{K}_n} |R_{\widehat{\theta}}(v,z) - 1| \le 2\gamma_n, \qquad \sup_{x\in\mathcal{K}_n} R_{\widehat{\theta}}(v,z) \le 2.$$

The fitted and the true conditional laws are related by a change of measure. For a bounded function ϕ on $\mathbb{R}^p$,

$$E\{\phi(\widetilde{X}) \mid \widehat{Z} = z, \mathcal{D}^N\} = E_{\theta_0}\{\phi(X)R_{\widehat{\theta}}(V,Z) \mid Z = z\}, \tag{44}$$

because both sides integrate $\phi(\widehat{B}_\parallel v + \widehat{B}_\perp z)$ over v, the left against $p_{\widehat{\theta}}(v \mid z)$ and the right against $R_{\widehat{\theta}}(v,z)p_{\theta_0}(v \mid z) = p_{\widehat{\theta}}(v \mid z)$. In particular, $\widehat{T}_{\widehat{B}}\phi(z)$ is the left side of (44), $T_{\widehat{B}}\phi(z)$ is the right side with $R_{\widehat{\theta}}$ replaced by one, and hence $(\widehat{T}_{\widehat{B}} - T_{\widehat{B}})\phi(z) = E_{\theta_0}[\phi(X)\{R_{\widehat{\theta}}(V,Z) - 1\} \mid Z = z]$.

We now verify the operator bound (13). Fix $g \in \mathcal{A}_{s,n}$ and split $g = g\mathbf{1}_{\mathcal{K}_n} + g\mathbf{1}_{\mathcal{K}_n^c}$, where $\mathbf{1}_{\mathcal{K}_n}$ is the indicator of the cube. For the part inside the cube, squaring the last display, applying Jensen's

inequality to the conditional expectation, integrating over z, and using the bound on $|R_{\widehat{\theta}} - 1|$ on $\mathcal{K}_n$, we obtain

$$\|(\widehat{T}_{\widehat{B}} - T_{\widehat{B}})(g\mathbf{1}_{\mathcal{K}_n})\|_2^2 \le E_{\theta_0}\big[g(X)^2\mathbf{1}_{\mathcal{K}_n}(X)\{R_{\widehat{\theta}}(V,Z)-1\}^2\big] \le 4\gamma_n^2\|g\|_2^2.$$

For the part outside the cube, Jensen's inequality and $|g| \le 2M$ give $\|T_{\widehat{B}}(g\mathbf{1}_{\mathcal{K}_n^c})\|_2^2 \le E\{g(X)^2\mathbf{1}_{\mathcal{K}_n^c}(X)\} \le 4M^2\Pr(X \notin \mathcal{K}_n) \le 1/(4n_T^2)$, and in the same way $\|\widehat{T}_{\widehat{B}}(g\mathbf{1}_{\mathcal{K}_n^c})\|_2^2 \le 4M^2\Pr(\widetilde{X} \notin \mathcal{K}_n \mid \mathcal{D}^N) \le 1/(4n_T^2)$. Since both operators are linear, the triangle inequality gives

$$\|(\widehat{T}_{\widehat{B}} - T_{\widehat{B}})g\|_2 \le 2\gamma_n\|g\|_2 + \frac{1}{2n_T} + \frac{1}{2n_T} = 2\gamma_n\|g\|_2 + \frac{1}{n_T}.$$

As $g \in \mathcal{A}_{s,n}$ was arbitrary, the definition of $\mathcal{E}_{F,n}$ in (13) gives $\mathcal{E}_{F,n} \le 2\gamma_n = O_p(n_N^{-1/2}\log n_T)$.

The second-moment bound follows from the same split. For $g \in \mathcal{A}_{s,n}$, (44) with $\phi = g^2\mathbf{1}_{\mathcal{K}_n}$, the bound $R_{\widehat{\theta}} \le 2$ on the cube, and $|g| \le 2M$ give

$$\begin{aligned} E_{\widehat{Z}}\big[E_{\widetilde{X}|\widehat{Z},\mathcal{D}^N} g(\widetilde{X})^2\big] &= E_{\theta_0}\big[g(X)^2\mathbf{1}_{\mathcal{K}_n}(X)R_{\widehat{\theta}}(V,Z)\big] + E\big\{g(\widetilde{X})^2\mathbf{1}_{\mathcal{K}_n^c}(\widetilde{X}) \mid \mathcal{D}^N\big\} \\ &\le 2\|g\|_2^2 + 4M^2\Pr(\widetilde{X} \notin \mathcal{K}_n \mid \mathcal{D}^N) \le 2\|g\|_2^2 + \frac{1}{n_T^2}, \end{aligned}$$

which is the second-moment bound with $C_V = 2$. All parts of Assumption 6 are now verified on the intersection of the events above, with $\delta_{F,n} = n_N^{-1/2}\log n_T$.

Finally, we prove the Hellinger bound. We bound the Hellinger distance by the Kullback–Leibler divergence, which has a closed form for Gaussian distributions. For densities p_1, p_2 let $H^2(p_1,p_2) = 1 - \int\sqrt{p_1p_2}$ and $\mathrm{KL}(p_1\|p_2) = \int p_1\log(p_1/p_2)$. Jensen's inequality gives $\mathrm{KL}(p_1\|p_2) \ge -2\log\int\sqrt{p_1p_2} \ge 2H^2(p_1,p_2)$. For two Gaussian distributions on $\mathbb{R}^q$,

$$\mathrm{KL}\{N(\mu_1,\Sigma_1)\|N(\mu_2,\Sigma_2)\} = \frac{1}{2}\sum_{j=1}^{q}(\lambda_j - 1 - \log\lambda_j) + \frac{1}{2}(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2),$$

where $\lambda_1,\dots,\lambda_q$ are the eigenvalues of $\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}$. We apply this formula with $\Sigma_1 = \Sigma_0$, $\Sigma_2 = \widehat{\Sigma}$, and $\mu_1 - \mu_2 = (A_0 - \widehat{A})z + (b_0 - \widehat{b})$. The covariance part is quadratic in $\Sigma_0 - \widehat{\Sigma}$. Indeed, the function $\lambda - 1 - \log\lambda$ is at most $(\lambda-1)^2$ for $\lambda \ge 1/2$, and $\sum_j(\lambda_j-1)^2 \le \|\Sigma_2^{-1}\|_{\mathrm{op}}^2\|\Sigma_1-\Sigma_2\|_F^2$; moreover, all $\lambda_j \ge 1/2$ with probability tending to one, by parameter convergence and the eigenvalue bounds recorded above. The mean part is quadratic in $(A_0 - \widehat{A}, b_0 - \widehat{b})$. Integrating both parts over the law of $\widehat{Z}$, which has a finite second moment, and using parameter convergence, we obtain

$$\int H^2\{p_{\widehat{\theta}}(\cdot \mid z), p_{\theta_0}(\cdot \mid z)\}\, dP_{\widehat{Z}}(z) \le C_0\|\widehat{\theta} - \theta_0\|^2 = O_p(n_N^{-1}). \tag{45}$$

This Hellinger bound is a property of the Gaussian example only; the proof of $\mathcal{E}_{F,n} = O_p(n_N^{-1/2}\log n_T)$ above does not use it. □

F Functional Support Recovery

This appendix derives support recovery from L_2 consistency. The argument uses the functional distances D_S, candidate family $\mathcal{S}_s$, and separation constant κ_0 defined in Appendix B.

For any function supported on at most s coordinates define

$$\mathfrak{S}_s(f) = \text{the lexicographically first member of } \arg\min_{\substack{S\in\mathcal{S}_s \\ D_S(f)=0}} |S|. \tag{46}$$

The extractor $\mathfrak{S}_s(f)$ selects a supporting set of minimum cardinality, breaking ties lexicographically. It depends on both f and the population law of X; no data-based computation is asserted. Every function in $\mathcal{H}_{s,n}$ is supported by at least one set in $\mathcal{S}_s$, namely the set of its nonzero first-layer columns. If f has a unique minimal support, then $\mathfrak{S}_s(f)$ equals it, by the reduction below. The

uniqueness condition in Corollary 3 holds, for example, when a function supported on both S and S' is also supported on $S \cap S'$.

Both results below use the following reduction. If a set T supports a function f, remove from T a coordinate j whenever $T \setminus \{j\}$ still supports f. After finitely many removals no coordinate can be removed. Because every superset of a support is again a support, no proper subset of the resulting set supports f; that is, the resulting set is a minimal support of f contained in T. The first result shows that the separation constant is positive.

Lemma 10 (Positive separation)**.** *Under Assumption 2, $\kappa_0 > 0$.*

Proof. Since p is fixed, $\mathcal{S}_s$ is finite, and κ_0 is the minimum of the finitely many distances $D_S(h_0)$ with $S \in \mathcal{S}_s$ and $S \neq S_0$. It therefore suffices to show that each of these distances is positive. Let $S \in \mathcal{S}_s$ with $S \neq S_0$, and suppose that S supported h_0. The reduction above would give a minimal support of h_0 contained in S, and this minimal support equals S_0 by the uniqueness in Assumption 2; hence $S_0 \subseteq S$. Since $|S| \leq s = |S_0|$, this forces $S = S_0$, a contradiction. Therefore S does not support h_0. Because $L_2(P_{X_S})$ is closed, the distance $D_S(h_0)$ from h_0 to this subspace is then positive. This argument concerns subspaces of $L_2(P_X)$ only; it uses no lower bound on $\|T(h - h_0)\|_2$. □

Corollary 3 (Functional-support recovery)**.** *Under the assumptions of Theorem 2 and its consistency conditions $\alpha_n \to 0$, $d_n/\alpha_n \to 0$, and $\epsilon_{H,n}^2/\alpha_n \to 0$, the extractor in* (46) *satisfies*

$$\Pr\{\mathfrak{S}_s(\widehat{h}) = S_0\} \longrightarrow 1.$$

If every function in $\mathcal{H}_{s,n}$ has a unique minimal support, then

$$\Pr\{S(\widehat{h}) = S_0\} \longrightarrow 1.$$

For a triangular array, the data-generating law may vary with sample size. If the estimation assumptions hold uniformly and the separation constants satisfy $\kappa_n \to 0$, the same conclusions hold whenever

$$\frac{d_n^2}{\alpha_n^2} + \frac{\epsilon_{H,n}^2}{\alpha_n} + \alpha_n^{\nu} = o(\kappa_n^2).$$

Proof of Corollary 3. An estimate within L_2 distance κ_0 of h_0 cannot have an incorrect support of size at most s. We prove this deterministic statement and then apply the consistency result of Theorem 2:

$$\{\mathfrak{S}_s(\widehat{h}) \neq S_0\} \subseteq \{\|\widehat{h} - h_0\|_2 \geq \kappa_0\}. \tag{47}$$

Since $\|\widehat{h} - h_0\|_2 \to 0$ in probability and $\kappa_0 > 0$, the probability of the right side tends to zero. The second claim reduces to the first once minimal supports are unique, and the triangular-array claim follows by replacing κ_0 with κ_n.

We first prove (47). The key observation is that a set supporting $\widehat{h}$ nearly supports h_0. Indeed, if $S \in \mathcal{S}_s$ supports $\widehat{h}$, then $\widehat{h}$ belongs to $L_2(P_{X_S})$, and since $D_S(h_0)$ is the distance from h_0 to this subspace,

$$D_S(h_0) \leq \|h_0 - \widehat{h}\|_2.$$

Now suppose that $\|h_0 - \widehat{h}\|_2 < \kappa_0$. Then no $S \in \mathcal{S}_s$ with $S \neq S_0$ supports $\widehat{h}$, because such a set would satisfy $D_S(h_0) < \kappa_0$, contradicting the definition of κ_0; in particular, no proper subset of S_0 supports $\widehat{h}$. On the other hand, $\widehat{h}$ is a network with at most s nonzero first-layer columns, and hence at least one set in $\mathcal{S}_s$ supports it. That set must be S_0. The extractor (46) returns the smallest supporting set in $\mathcal{S}_s$, which is therefore S_0. Taking complements gives (47).

We next derive the first claim. By (47),

$$\Pr\{\mathfrak{S}_s(\widehat{h}) \neq S_0\} \leq \Pr\{\|\widehat{h} - h_0\|_2^2 \geq \kappa_0^2\}.$$

Under the three consistency conditions, the rate in Theorem 2 tends to zero, and hence $\|\widehat{h} - h_0\|_2^2 \to 0$ in probability. Since κ_0 is a fixed positive number by Lemma 10, the right side tends to zero.

For the second claim, suppose that every function in $\mathcal{H}_{s,n}$ has a unique minimal support. For $\widehat{h}$, the reduction above, applied to any supporting set, gives a minimal support contained in that set, and by uniqueness this minimal support is $S(\widehat{h})$. Hence $S(\widehat{h})$ is contained in every supporting set of $\widehat{h}$; in particular, $|S(\widehat{h})| \le s$, and $S(\widehat{h})$ is the smallest supporting set in $\mathcal{S}_s$. Therefore $S(\widehat{h}) = \mathfrak{S}_s(\widehat{h})$, and the second claim follows from the first. Without this uniqueness property the first claim still holds, including for laws of X with redundant coordinates.

Finally, consider a triangular array with separation constants $\kappa_n \to 0$. The inclusion (47) holds with κ_n in place of κ_0, because its proof uses only the definition of the separation constant. The first claim then requires $\Pr\{\|\widehat{h} - h_0\|_2^2 \ge \kappa_n^2\} \to 0$, which holds when the rate in Theorem 2 is $o(\kappa_n^2)$. The second claim follows from the first as before. □

G Additional Experimental and Reproducibility Details

G.1 Loading realization, calibration, and population diagnostics

The main experiment's three scenarios in Section 5.1 use $u_0 = 1.5U_1\epsilon_1 + 0.5\eta$, with the same ϵ_1 that enters X. The interaction coefficient is fixed at $\gamma = 1.5$; there is no parameter sweep. Independence and centering give

$$\begin{aligned} \mathrm{Var}(1.5U_1\epsilon_1) &= 2.25D_{11} = 5.113166713877538, \\ \mathrm{Var}(u_0) &= 2.25D_{11} + 0.25 = 5.363166713877538. \end{aligned}$$

Although disturbance variance is shared, $\mathrm{Var}(Y)$ can differ across structural functions. We chose this strong-interaction design after exploratory comparisons with weaker confounding. The figures and tables report only the selected design.

We draw an initial matrix $L \in \mathbb{R}^{10\times 2}$ with independent entries $L_{ja} \sim N(0, 1/2)$, and set

$$\Lambda = \frac{\sqrt{10}}{\|L\|_F} L, \qquad \frac{1}{10}\,\mathrm{tr}(\Lambda\Lambda^\top) = 1.$$

A single realized Λ is retained across all three structural functions, sample sizes, and Monte Carlo replications. We use the heteroskedastic idiosyncratic covariance

$$D = \mathrm{diag}\left(2.272518539501128,\ 2.272518539501128,\ \underbrace{0.681870365124718, \ldots, 0.681870365124718}_{8 \text{ entries}}\right).$$

The diagonal entries specify variances, with $\mathrm{tr}(D)/10 = 1$. The design assigns larger idiosyncratic variances to the active coordinates using S_0, which remains hidden from the fitted estimators. For the fixed loading matrix, define the operator-strength measure for h_0 by

$$\kappa^2(h_0) = \frac{\mathrm{Var}\{E[h_0(X) \mid Z]\}}{\mathrm{Var}\{h_0(X)\}}$$

which equals 0.4 for the Cubic scenario. This measure depends on h_0, so the same loading matrix and noise covariance need not give the same strength for Linear or Interaction.

The experiment uses base seed 20260827, with separate deterministic namespaces for the loading realization, calibration, population validation, training data, test data, network initialization, observation shuffling, and the two conditional Monte Carlo batches. The loading-generation scenario seed is 89896278289785659. The retained loading matrix, rounded for display, is

Table 1: Saved structural calibration constants, rounded for display, and independent population-validation moments from 200,000 observations

Scenario	μ_a	c_a	Validation mean	Validation variance
Linear	-0.001896490841	2.844830018055	0.0013992335	0.9980795040
Cubic	-0.006046551743	10.926143491903	0.0007966416	0.9944218009
Interaction	1.013933423395	1.457018458087	0.0006845912	0.9939244054

$$\Lambda \simeq \begin{pmatrix} -0.43460549 & -0.85976743 \\ 0.51829244 & -1.02452823 \\ -0.05597407 & 0.68627474 \\ 0.54912199 & -0.29432275 \\ 0.37841028 & -0.00669939 \\ 0.79551375 & 0.42340360 \\ -0.15596328 & 0.91099695 \\ 0.43404938 & 1.38625066 \\ 1.36161025 & 0.30566193 \\ 0.18819092 & -0.99441274 \end{pmatrix}.$$

To place the structural signals on a common variance scale, we generate one independent calibration sample $X_1^{\mathrm{cal}}, \ldots, X_{N_{\mathrm{cal}}}^{\mathrm{cal}}$, with $N_{\mathrm{cal}} = 10^6$, from the common treatment distribution. For each scenario a, define

$$\mu_a = \frac{1}{N_{\mathrm{cal}}} \sum_{t=1}^{N_{\mathrm{cal}}} f_a(X_t^{\mathrm{cal}}), \qquad c_a^2 = \frac{1}{N_{\mathrm{cal}}} \sum_{t=1}^{N_{\mathrm{cal}}} \{f_a(X_t^{\mathrm{cal}}) - \mu_a\}^2, \qquad h_{0,a}(x) = \frac{f_a(x) - \mu_a}{c_a}.$$

The constants are saved once and reused throughout the experiment; neither the structural functions nor the disturbances are restandardized within replications. The structural components $h_0(X)$ have approximately zero population mean and unit population variance, up to calibration Monte Carlo error. Independent population diagnostics use 200,000 observations.

Let $B_{\|}$ and $B_{\perp}$ be orthonormal bases for the loading space and its complement. At the population basis,

$$Z = B_{\perp}^{\top} X = B_{\perp}^{\top} \epsilon \perp U, \qquad E\{Y - h_0(X) \mid Z\} = 0.$$

The conditional first stage is nontrivial: for the realized design,

$$\| \operatorname{Cov}(V, Z)\|_F = \|B_{\|}^{\top} D B_{\perp}\|_F \simeq 0.8915.$$

The full-precision matrix is retained in the resolved experiment configuration. Calibration uses seed 2902207984129644426. The independent population diagnostics use seed 3518224067295180325. Table 1 reports the saved calibration constants and independently estimated validation moments.

The diagnostic variance of u_0 is 5.3801905132, close to the population value 5.3631667139 up to Monte Carlo error. No diagnostic moment is forced to its population target. Evaluation uses each replication's test variance in the RelL2 denominator.

G.2 Nuisance fitting and structural implementations

Sparse Neural SIV selects k by minimum five-fold IV validation loss over $\mathcal{K} = \{1, 2, 3\}$ at all four sample sizes, followed by a full-sample refit. Linear SIV and Sparse Neural $\widehat{U}$ use fixed $k = 2$ throughout. Appendix G.6 details the CV procedure. Every fitted estimator receives $q = 2$, but none receives the true $U, \epsilon, \Lambda, D, h_0$, or S_0. Test values of h_0 are used only for evaluation.

The three SIV methods share the estimated loading space and Gaussian conditional sampler within each replication. We estimate the loading space from X alone by Gaussian factor analysis (FA), using maximum likelihood with a diagonal positive idiosyncratic covariance, using a deterministic principal-factor initialization followed by EM. We then form

$$\widehat{V}_i = \widehat{B}_{\|}^{\top} X_i, \qquad \widehat{Z}_i = \widehat{B}_{\perp}^{\top} X_i,$$

and fit the affine Gaussian conditional model

$$\widehat{P}_V(\cdot \mid z) = N(\widehat{A}z + \widehat{b}, \widehat{\Sigma}).$$

Multivariate least squares with an intercept gives $(\widehat{A}, \widehat{b})$, and the residual covariance uses divisor n. Gaussian treatments have a Gaussian conditional law in every fixed orthogonal coordinate system. The main experiment therefore evaluates nuisance estimation within a correctly specified conditional family.

Gaussian factor-analysis EM uses at most 500 iterations, tolerance 10^{-8}, and a coordinatewise uniqueness floor equal to 10^{-8} times the corresponding sample variance. A nonconverged factor fit is recorded as a failure rather than silently accepted. Conditional Gaussian fitting uses the empirical residual covariance without a ridge penalty and rejects a non-positive-definite covariance.

Linear SIV. Motivated by the sparse linear synthetic-instrument approach of Tang, Kong, and Wang (Tang et al., 2026), we implement a fixed-k sparse linear baseline. Define the reconstructed conditional treatment means

$$\widehat{M}_i = \widehat{B}_{\|}(\widehat{A}\widehat{Z}_i + \widehat{b}) + \widehat{B}_{\perp}\widehat{Z}_i.$$

We enumerate all $\binom{10}{2} = 45$ two-coordinate subsets and solve

$$(\widehat{S}_{\mathrm{lin}}, \widehat{a}, \widehat{\beta}_{\widehat{S}_{\mathrm{lin}}}) \in \underset{\substack{S \subseteq \{1,\dots,10\},\ |S|=2 \\ a \in \mathbb{R},\ \beta_S \in \mathbb{R}^2}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - a - \widehat{M}_{i,S}^{\top} \beta_S\}^2.$$

The intercept and coefficients are unpenalized. Prediction uses the observed treatment vector: $\widehat{h}_{\mathrm{lin}}(x) = \widehat{a} + x_{\widehat{S}_{\mathrm{lin}}}^{\top} \widehat{\beta}_{\widehat{S}_{\mathrm{lin}}}$. The conditional means are evaluated analytically, so Linear SIV requires no Monte Carlo approximation. This baseline implements a fixed-budget linear version of the approach in Tang et al. (2026).

Dense Neural SIV. This method uses a fully connected ReLU network with all ten inputs, two hidden layers of width 64, and a scalar output clipped to $[-50, 50]$. Inputs are centered and scaled using training-sample moments. For each training observation, we generate two independent conditional batches of size $m = 48$,

$$\widetilde{V}_{i\ell}^{(b)} \sim \widehat{P}_V(\cdot \mid \widehat{Z}_i), \qquad \widetilde{X}_{i\ell}^{(b)} = \widehat{B}_{\|}\widetilde{V}_{i\ell}^{(b)} + \widehat{B}_{\perp}\widehat{Z}_i, \qquad b \in \{1, 2\},$$

and set

$$A_{i,b}(h) = \frac{1}{m} \sum_{\ell=1}^{m} h(\widetilde{X}_{i\ell}^{(b)}).$$

The conditional draws are generated once per fit and reused during training. The neural fitting criterion is

$$\widehat{Q}_{\alpha,\lambda_w}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - A_{i,1}(h_\theta)\}\{Y_i - A_{i,2}(h_\theta)\} + \frac{\alpha}{n} \sum_{i=1}^{n} h_\theta(X_i)^2 + \frac{\lambda_w}{2} \sum_{\ell} \|W_\ell\|_F^2,$$

where $\alpha = 0.01$, $\lambda_w = 10^{-4}$, and the last sum includes all network weight matrices but excludes biases. The first two terms implement the double-Monte-Carlo SIV criterion; the final term is an additional numerical weight penalty. The dense fit has no group penalty or input-support constraint, and its support metrics are not applicable.

Proposed Sparse Neural SIV. The proposed implementation uses the same architecture, input normalization, conditional batches, and base criterion as Dense Neural SIV, with a two-phase sparsity schedule. During the first 80 epochs, it applies group shrinkage to the first-layer columns, targeting the additional penalty

$$\lambda_g \sum_{j=1}^{10} \|(W_1)_{\cdot j}\|_2, \qquad \lambda_g = 1.$$

Training alternates Adam updates of the base criterion with a Euclidean groupwise proximal step. After 80 epochs, the algorithm retains the k columns with largest norms and zeros the remaining columns. It freezes this mask and refits for 80 further epochs, retaining the Tikhonov and weight penalties while removing the group penalty. Refitting starts from the screened weights with a fresh optimizer. The reported support is the frozen mask; ties are resolved deterministically.

Naive NN. This comparator uses all ten observed treatments to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - h_\theta(X_i)\}^2 + \frac{\lambda_w}{2}\sum_{\ell}\|W_\ell\|_F^2.$$

It has no first stage, sparse screening, Tikhonov term, or conditional Monte Carlo samples. Its network prediction is evaluated against h_0, and selected-support metrics are not applicable.

Sparse Neural $\widehat{U}$. We construct this two-network comparator to evaluate additive factor-score adjustment. It fits Gaussian factor analysis to X using the same settings as the SIV methods. The resulting score map was checked against their shared factor fit. The comparator uses these scores directly and requires no SIV conditional sampler. With training mean $\bar{X}$, form the empirical covariance with divisor n and compute scores by

$$\widehat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top, \qquad \widehat{U}_i^\top = (X_i - \bar{X})^\top \widehat{\Sigma}_X^{-1}\widehat{\Lambda}.$$

The implementation uses a linear solve, rather than explicitly forming an inverse. The training mean and score map are fixed for test observations. The second stage jointly fits $h_\theta(X) + a_\phi(\widehat{U})$ using

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - h_\theta(X_i) - a_\phi(\widehat{U}_i)\}^2 + \frac{\lambda_w}{2}\left(\sum_{\ell}\|W_{\theta,\ell}\|_F^2 + \sum_{\ell}\|W_{\phi,\ell}\|_F^2\right).$$

The h branch follows the same 80-epoch group-shrinkage and 80-epoch frozen-top-two refitting schedule as Sparse Neural SIV. The a branch uses both factor scores without a sparsity constraint. Both networks allow interactions among their inputs; the restriction is additivity between the treatment and factor-score branches. No Tikhonov or conditional-Monte-Carlo term is used. After training, define

$$\widehat{c} = \frac{1}{n}\sum_{i=1}^{n} a_\phi(\widehat{U}_i), \qquad \widehat{h}(x) = h_\theta(x) + \widehat{c}, \qquad \widehat{a}(u) = a_\phi(u) - \widehat{c}.$$

Test evaluation fixes this training value of $\widehat{c}$. RelL2 uses the estimated structural function $\widehat{h}$, not the combined outcome prediction $h_\theta + a_\phi$.

Shared neural training settings. Every network has two hidden layers of width 64, ReLU activations, and a raw scalar output clipped to $[-50, 50]$. Inputs are standardized using training-sample moments; the X and factor-score inputs of Sparse Neural $\widehat{U}$ are standardized separately. All neural fits use 160 epochs, minibatches of size 256, Adam with initial learning rate 0.002, gradient-norm

clipping at 10, and $\lambda_w = 10^{-4}$. Weight penalties always have coefficient $\lambda_w/2$, exclude biases, and include both networks for Sparse Neural $\widehat{U}$. The two sparse methods use $\lambda_g = 1$ in the first 80 epochs, then freeze the selected mask and restart the optimizer and cosine schedule for 80 refitting epochs. Dense Neural SIV and Naive NN use a single 160-epoch cosine schedule. Each schedule ends at one tenth of its initial learning rate. Fixed epoch counts do not imply equal numbers of gradient updates or equal computation times. These settings are held fixed across sample sizes and scenarios without test-set tuning.

G.3 Paired sampling, evaluation, and reproducibility

The sample-size grid is $n \in \{1000, 3000, 5000, 10000\}$, with $R = 200$ replications per scenario/sample-size cell and 20,000 independent test observations per replication. At a fixed (n, r), the three scenarios share U, ϵ, η, X and the test draws; only the structural function changes. All five methods are evaluated on the same data within a cell. Training and test data are independent. Data seeds derive from base seed 20260827, separate deterministic namespaces, and (n, r). They exclude scenario, method, and worker identifiers, preserving paired comparisons across methods and scenarios. In the fixed-budget fits, network initialization, observation shuffling, and conditional-Monte-Carlo streams are paired across applicable methods using fixed rules. CV uses separate deterministic fold and stage seeds. The a_ϕ network has a separate derived initialization seed. The fixed loading matrix and calibration constants are not redrawn across replications or workers.

Within each CV training fold, FA, conditional-model fitting, input normalization, and structural fitting use only that fold's training data; held-out observations are used for scoring. After selecting k, all stages are refitted on the full training sample. Nuisance and structural fitting share observations within each fit; no nuisance cross-fitting is used. The fixed-budget fits likewise reuse the full training sample. Calibration and population-diagnostic samples are independent of training and test data. The test set is used only for evaluation, including the RelL2 criterion in Section 5.1, which is a relative squared error without taking a square root. The boxplot summarizes replication-level errors, whereas numerical tables report means.

Only Linear SIV, Sparse Neural SIV, and Sparse Neural $\widehat{U}$ select a support. For their selected sets $\widehat{S}_r$, we report exact selection, true positive rate (TPR), and false discovery proportion (FDP):

$$\widehat{P}_{\text{exact}} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\{\widehat{S}_r = S_0\}, \qquad \text{TPR}_r = \frac{|\widehat{S}_r \cap S_0|}{2}, \qquad \text{FDP}_r = \frac{|\widehat{S}_r \setminus S_0|}{|\widehat{S}_r|}.$$

Mean TPR averages TPR_r, and false discovery rate (FDR) averages the replication-level FDP_r. Because the selected size can vary under CV, TPR and FDP need not sum to one. The selected set is the winning linear candidate or the frozen sparse-network mask, retained even if a selected group subsequently becomes zero. These operational sets are distinct from minimal functional supports in Corollary 3. Dense Neural SIV and Naive NN do not select sparse supports. Their support entries are marked not applicable, and we apply no post-fit top-two screen.

The figure combines 12,000 successful final estimates on $3 \times 4 \times 200 = 2{,}400$ simulated datasets: 2,400 CV-selected Sparse Neural SIV refits and 9,600 unchanged comparator fits. The two CV runs completed 12,000 folds and 36,000 candidate fits without failures. Their training/test data hashes and seeds match the original replications. The original 120 pilot fits and the separate CV pilots (nine fits at the smaller sample sizes and three at $n = 10000$) are excluded. Each scenario/sample-size/method cell has 200 successful final estimates; failure rates count failed final records relative to the 200 expected. Replication-level metrics, selected indices, stage-specific seeds, configurations, package versions, and status fields are retained.

All fits use CPUs and float64 arithmetic, with 48 single-threaded workers. The recorded environment is Python 3.11.7, PyTorch 2.14.0+cu130 (CPU execution), NumPy 1.26.4, SciPy 1.17.1, and Matplotlib 3.8.0. These versions describe the reported runs. The supplementary package separately pins its execution-check environment (Python 3.12.14, PyTorch 2.10.0, and NumPy 2.4.6); its README notes that numerical results can vary across environments even with identical seeds. The original fixed-budget run took 11 hours, 16 minutes, and 54 seconds; the CV runs at $n \leq 5000$ and $n = 10000$ took approximately 21 hours, 56 minutes, and 28 seconds and 16 hours, 5 minutes, and 11 seconds, respectively. These separate wall-clock totals exclude the pilots and do not measure relative computation times across methods.

Table 2: Largest-sample comparison at $n = 10{,}000$, with 200 replications per scenario and method. Sparse Neural SIV uses CV-selected k; Linear SIV and Sparse Neural $\widehat{U}$ use fixed $k = 2$. Exact selection, mean TPR, and FDR are proportions; FDR is the mean replication-level FDP. A dash means that selected-support metrics are not applicable.

Scenario	Method	Mean RelL2	Exact	Mean TPR	FDR
Linear	Linear SIV	0.0037295	1.0000	1.0000	0.0000
	Dense Neural SIV	1.2158283	—	—	—
	Sparse Neural SIV	0.0715199	0.2000	1.0000	0.2667
	Naive NN	4.1644802	—	—	—
	Sparse Neural $\widehat{U}$	4.9747281	0.0000	0.5000	0.5000
Cubic	Linear SIV	0.3650967	1.0000	1.0000	0.0000
	Dense Neural SIV	1.4319467	—	—	—
	Sparse Neural SIV	0.1614824	0.7450	1.0000	0.0850
	Naive NN	4.1903292	—	—	—
	Sparse Neural $\widehat{U}$	2.4751030	0.6400	0.8200	0.1800
Interaction	Linear SIV	1.0152320	0.0750	0.3175	0.6825
	Dense Neural SIV	1.6235734	—	—	—
	Sparse Neural SIV	0.2917848	0.4550	0.7825	0.0642
	Naive NN	4.3164759	—	—	—
	Sparse Neural $\widehat{U}$	1.1264362	0.9950	0.9975	0.0025

Table 3: Exact support-recovery rates over 200 replications, shown as percentages. Only methods that actually select a support are included; columns give training sample sizes n. Sparse Neural SIV uses CV in all four columns; the other methods use fixed $k = 2$ throughout.

Scenario	Method	1,000	3,000	5,000	10,000
Linear	Linear SIV	100%	100%	100%	100%
	Sparse Neural SIV	1.5%	0.5%	1.5%	20%
	Sparse Neural $\widehat{U}$	0%	0%	0%	0%
Cubic	Linear SIV	96%	100%	100%	100%
	Sparse Neural SIV	4%	12%	23.5%	74.5%
	Sparse Neural $\widehat{U}$	3%	71%	86.5%	64%
Interaction	Linear SIV	6%	9%	6.5%	7.5%
	Sparse Neural SIV	6.5%	4.5%	9.5%	45.5%
	Sparse Neural $\widehat{U}$	65.5%	90%	95%	99.5%

G.4 Additional numerical results and scope

Table 2 gives the largest-sample comparison, and Table 3 gives exact selection rates over the full sample-size grid. Sparse Neural SIV uses CV-selected k at every sample size.

Structural error and support selection. From $n = 1{,}000$ to $n = 10{,}000$, Sparse Neural SIV's mean RelL2 decreases from 1.31353 to 0.07152 in Linear, from 1.13300 to 0.16148 in Cubic, and from 0.60993 to 0.29178 in Interaction. Linear SIV's Cubic error decreases from 0.41647 to 0.36510; Sparse Neural SIV improves on it only at the largest sample size. Sparse Neural $\widehat{U}$'s Interaction error decreases from 2.51200 to 1.12644. Sparse Neural SIV has lower mean error than the factor-score comparator in all twelve cells, while Linear SIV remains best throughout Linear and Sparse Neural SIV remains best among all five methods throughout Interaction.

Support recovery need not track these errors or vary monotonically with sample size. Among the CV fits at $n = 5000$, the frequencies of selecting $k = 1, 2, 3$ are $8, 37, 155$ in Linear, $11, 67, 122$ in Cubic, and $68, 88, 44$ in Interaction (200 replications each). Selecting the true budget and recovering the true support are distinct events. At $n = 10{,}000$, Sparse Neural SIV's exact rates are 20%, 74.5%, and 45.5%. In Interaction, Sparse Neural $\widehat{U}$ has a higher exact rate (99.5%) but a larger mean structural error (1.12644 versus 0.29178). All percentages describe this finite set of replications, not

population recovery probabilities. At $n = 10000$, the respective $k = 1, 2, 3$ counts are $0, 40, 160$, $0, 149, 51$, and $76, 102, 22$. Linear and Cubic always include both true variables; their lower exact rates reflect an extra selected variable. Interaction selects $k = 1$ in 76 replications, omitting an active variable. Compared with fixed $k = 2$, mean RelL2 increases by 0.02505, 0.03005, and 0.01230, respectively. CV minimizes held-out IV loss, which need not minimize structural RelL2 or maximize exact recovery.

Factor-score adjustment diagnostics. In Linear, Sparse Neural $\widehat{U}$ selects $\{1, 9\}$ rather than S_0 in all 800 fits. In Cubic at $n = 10{,}000$, its 128 fits with correct support have mean RelL2 1.02077; the other 72 all select $\{1, 9\}$ and have mean 5.06058. The overall mean is 2.47510 and the median is 1.07510. Both selected-support groups contribute substantially to the overall error distribution. Among paired replications where both sparse methods select S_0, the Cubic cell has 90 pairs with mean errors 0.13249 for Sparse Neural SIV and 1.01917 for Sparse Neural $\widehat{U}$. Interaction has 90 such pairs, with means 0.26194 and 1.11251. These post hoc summaries condition on the observed selected supports. They do not use oracle refits or isolate the effect of screening errors.

The following Gaussian conditional-mean calculation explains why the additive adjustment model is restrictive in this design. Let $\mu(X) = E(U \mid X)$, $S = \mathrm{Var}(U \mid X)$, and let λ_1 be the first loading row written as a column. Then

$$E(U_1\epsilon_1 \mid X) = \mu_1(X)\{X_1 - \lambda_1^\top \mu(X)\} - (S\lambda_1)_1.$$

With the true h_0 fixed, this remaining conditional mean generally cannot be represented by a function of the ideal scores $\mu(X)$ alone, because it includes $\mu_1(X)X_1$. Estimated scores $\widehat{U}$ add a further approximation. Allowing interactions inside both networks does not remove the additive restriction between branches. The observed results are consistent with the structural branch absorbing some confounding; they do not prove an asymptotic bias bound or a universal failure of factor-score adjustment.

Scope. The design uses a single fixed loading realization, support-aligned D, same-sample nuisance and structural fitting, and fixed networks, regularization, and training epochs. The main Gaussian experiment does not evaluate conditional-sampler misspecification. Appendix G.8 adds one two-component mixture design with a fitted finite-mixture sampler; broader generative models such as flows or diffusion samplers are not studied. The reported experiments do not use nuisance cross-fitting; CV is evaluated for Sparse Neural SIV at all four sample sizes in the main design. Gaussian, polynomial, and product terms produce unbounded outcomes, unlike the bounded-outcome condition in the rate theorem. Numerical training is not certified to attain the global minimizer in Theorem 2, and the frozen mask is not the theoretical functional-support extractor. The interaction disturbance remains within the additive structural decomposition $Y = h_0(X) + u_0$; verifying its population IV moment does not verify every identification or estimation assumption.

G.5 Algorithmic implementation scope

Practical neural optimization and support screening are described above; the cross-validation procedure for selecting k, including the validation score and final refit, is given in Appendix G.6. The analysis in Section 4.4 instead fixes $k = s$, uses independent nuisance and structural samples, and assumes a global empirical minimizer. It does not cover CV selection or same-sample refitting, or assert convergence of numerical training to that minimizer.

The proof uses independent samples $\mathcal{D}^N, \mathcal{D}^T$ of sizes n_N, n_T. Fixed-K nuisance cross-fitting would instead use a loading space and sampler fitted without each structural observation. The reported experiments use neither construction. Their validation folds hold out observations for scoring, while nuisance and structural fitting share the training fold. Thus tuning by CV does not provide the independence used in the theorem.

DeepGraHTP or FastDeepGraHTP offer possible numerical implementations, but the experiments use the two-phase procedure above. The theory establishes neither restricted strong convexity and smoothness for the empirical objective nor convergence of numerical training to its global minimizer.

G.6 Cross-validation for sparsity selection

We select the sparsity budget by comparing held-out conditional-moment losses over a prespecified grid $\mathcal{K} \subset \{1, \dots, p-q-1\}$.

Partition the training indices into F nearly equal folds $I_1, \dots, I_F$. For each fold f, estimate the loading space, conditional sampler, and input normalization using only I_{-f}. Share these nuisance estimates across candidates and fit each $\widehat{h}_{k,-f}$ using the complete group-shrinkage, top-k screening, and fixed-mask refitting procedure. For each held-out observation, reconstruct two independent batches of m_{val} treatment draws from the fold's fitted sampler, conditional on $\widehat{Z}_{i,f} = \widehat{B}_{\perp,-f}^T X_i$, as in Algorithm 1. Let $A_{i,b}^{(-f)}(k) = m_{\mathrm{val}}^{-1} \sum_{\ell=1}^{m_{\mathrm{val}}} \widehat{h}_{k,-f}(\widetilde{X}_{i\ell,f}^{(b)})$. Use fresh validation draws, independent of training randomness and shared across k, to compute

$$\widehat{R}_{\mathrm{CV}}(k) = \frac{1}{n} \sum_{f=1}^{F} \sum_{i \in I_f} \{Y_i - A_{i,1}^{(-f)}(k)\}\{Y_i - A_{i,2}^{(-f)}(k)\}. \tag{48}$$

We select the budget with the smallest cross-validated risk:

$$\widehat{k} \in \arg\min_{k \in \mathcal{K}} \widehat{R}_{\mathrm{CV}}(k). \tag{49}$$

We use $F = 5$ and $m_{\mathrm{val}} = 96$, holding other training hyperparameters fixed. In the simulation CV runs, $\mathcal{K} = \{1, 2, 3\}$. Within each fold and scenario, candidates share the same initialization and first 80 group-shrinkage epochs. The implementation computes this common screening path once, clones the network and shuffle state, then performs a separate 80-epoch fixed-mask refit for each k, with a fresh optimizer. No screening weights are shared across folds or scenarios. Thus each candidate retains the full 160-epoch schedule, while shared computation reduces the total to $5(80 + 3 \times 80) + 160 = 1760$ epochs per scenario, including the final refit. Finally, refit the nuisance objects and a new sparse network with budget $\widehat{k}$ on all training observations; test data are reserved for final evaluation.

G.7 Mouse-data implementation and plotted contrasts

The supplementary package's `experiments/run_mouse.py` reads the pinned public `micedata.rda` file from the third-party CopSens dataset, selecting the 17 genes listed in `configs/mouse.yaml` and outcome `y`. It retains all 287 rows without imputation, filtering, or an additional log transformation. The configuration records the dataset URL and SHA-256 checksum. Expression coordinates in the figures are those supplied by this processed public dataset.

Both methods compare $k \in \{1, \dots, 6\}$ on the same five folds, with 96 conditional draws per observation in each validation batch. Within each training fold, treatment and outcome normalization uses that fold's sample means and standard deviations, with divisor $n_{\mathrm{train}} - 1$. For each fixed k, Linear SIV enumerates all $\binom{17}{k}$ supports, fits an intercept and coefficients by least squares on the fitted conditional treatment means, and selects the finite, full-rank fit with smallest training mean squared error. Neural candidates are fitted separately, using the architecture with two hidden layers of width 64 and 80-epoch screening plus 80-epoch fixed-mask refitting schedule described above. Unlike the simulation's shared screening path, mouse CV fits each candidate separately. Validation residual products are converted to original outcome units by multiplying by the training-fold outcome variance. Fold-size-weighted averages select k, after which both methods are refitted on all 287 mice.

For plotting, let $\widehat{h}$ denote the fitted predictor restored to the supplied expression coordinates and original outcome units. For gene j, define

$$g_j(t) = \frac{1}{287} \sum_{i=1}^{287} \{\widehat{h}(t, X_{i,-j}) - \widehat{h}(0, X_{i,-j})\}.$$

The other genes retain each mouse's observed values. The reference is zero on the public expression scale, which is transformed internally to $-\bar{X}_j / s_{X_j}$ under full-sample normalization. This reference does not assume that zero represents biological absence of expression. We evaluate g_j at 61 equally spaced points $t_1, \dots, t_{61}$ from the empirical 5th to 95th expression percentile. The reported percentile

contrast is $g_j(t_{61}) - g_j(t_1)$. Let $a_j + b_j t$ be the least-squares affine fit to these 61 curve values. The displayed nonlinearity index is

$$\mathrm{NL}_j = \frac{61^{-1}\sum_{\ell=1}^{61}\{g_j(t_\ell) - a_j - b_j t_\ell\}^2}{61^{-1}\sum_{\ell=1}^{61} g_j(t_\ell)^2 + 10^{-12}}.$$

This descriptive index is zero for an exactly affine curve and measures relative departure from its affine approximation; it is not a formal test statistic. Body-weight units are preserved as supplied by the public outcome.

G.8 Robustness to additive confounding and mixture treatment noise

Robustness design. This design examines performance with additive latent confounding and non-Gaussian treatment noise. It was fixed after directed exploration and changes both features together, so their effects are not isolated. Retain $p = 10$, $q = 2$, and $S_0 = \{1, 2\}$, but set $\Lambda = 0.3\Lambda_0$, where Λ_0 is the loading matrix in Appendix G. Independently draw $U \sim N(0, I_2)$, $\eta \sim N(0, 1)$, and an equiprobable $H \in \{-1, +1\}$, with

$$X = \Lambda U + \epsilon, \qquad \epsilon \mid H = t \sim N(0, D_t), \qquad D_t = \mathrm{diag}\{d_j(1 + 0.95ts_j)\}_{j=1}^{10},$$

where $s = (-1, -1, +1, +1, +1, +1, +1, +1, +1, +1)^\top$, $d_1 = d_2 = 2.272518539501128$, and $d_3 = \cdots = d_{10} = 0.681870365124718$. Both components are centered, with $\mathrm{Cov}(\epsilon) = D$. Although the noise coordinates are uncorrelated, their shared mixture label H makes them generally dependent, yielding a non-Gaussian conditional treatment law. As in the main design, D assigns larger variances to the active coordinates. Use the same three structural shapes as Section 5.1, recalibrated using 10^6 independent mixture observations: $(\mu_a, c_a) = (-0.0015875745, 2.2040687179)$, $(-0.0049813169, 11.2187096887)$, and $(0.8793099829, 1.3042143545)$ for Linear, Cubic, and Interaction. These constants are fixed across replications. Set

$$Y = h_0(X) + 3g(U) + 0.5\eta, \qquad g(U) = \frac{U_1^2 + U_2^2 - 2}{2}.$$

This is additive latent confounding with no U–ϵ interaction. Since $\epsilon \perp (U, \eta)$ and $E[g(U)] = 0$, $E[Y - h_0(X) \mid Z] = 0$ for $Z = B_\perp^\top \epsilon$. This calculation verifies the IV moment only; completeness, plurality, and the estimation assumptions require separate justification.

Conditional mixture sampler. All SIV methods share a loading space estimated from training X by Gaussian factor analysis (FA). Sparse Neural SIV with a Gaussian-mixture-model sampler (GMM) standardizes the training $(\widehat{V}, \widehat{Z})$ coordinates and fits two full-covariance Gaussian components, estimating their weights, means, and covariances without observing H. Scikit-learn fitting uses five k-means initializations, at most 500 iterations, tolerance 10^{-4}, and covariance regularization 10^{-5}. Component-specific Gaussian conditioning and z-dependent posterior weights yield draws of $V \mid Z = z$, followed by treatment reconstruction. Linear SIV and Dense Neural SIV (Gaussian) retain Gaussian conditional models; Sparse Neural $\widehat{U}$ uses Gaussian factor scores. Neural architecture, penalties, and the 160-epoch training schedule are unchanged; sparse fits screen for 80 epochs and refit for 80, and neural SIV uses two independent batches of 48 draws. All support-selecting methods use known $k = 2$, with unknown support and no CV; the GMM component count is fixed at two.

Structural results. Table 4 reports mean relative squared L_2 errors over 200 attempted replications per cell, with 20,000 independent test observations each and paired data across methods and scenarios. FA fails within 500 iterations for $35, 6, 0, 0$ of the 200 data replications at $n = 1000, 3000, 5000, 10000$, respectively. FA-dependent methods share the successful replications. Each treatment dataset's first-stage fit is reused across all three structural scenarios, so the $165 + 194 + 200 + 200 = 759$ successful FA replications yield 759 GMM fits, all of which succeed. Naive NN succeeds throughout. At $n = 10000$, Sparse Neural SIV (GMM) has lower error than Sparse Neural $\widehat{U}$ in every paired replication for each scenario, and the lowest mean in Cubic and Interaction; Linear SIV remains best in Linear. Linear SIV's small-sample error distributions are heavy-tailed, making their means sensitive to large errors in individual replications.

Table 4: Mixture-design mean RelL2 (relative squared error; lower is better). FA-dependent columns average 165, 194, 200, and 200 successful replications at increasing n; Naive NN averages all 200.

Scenario	n	Linear SIV	Dense Neural SIV (Gaussian)	Sparse Neural SIV (GMM)	Naive NN	Sparse Neural $\widehat{U}$
Linear	1000	1.6830	2.2970	1.3393	4.6424	1.5218
	3000	0.0488	1.6503	1.1647	5.2884	1.1700
	5000	0.0065	1.2966	0.9291	5.3867	0.9095
	10000	0.0027	0.9377	0.0194	5.2918	0.8972
Cubic	1000	3.1181	2.4957	1.2171	4.9108	1.4226
	3000	0.4687	1.7002	0.9985	5.3868	1.0712
	5000	0.4370	1.3350	0.6457	5.4724	0.8993
	10000	0.4311	0.9591	0.0433	5.3354	0.9099
Interaction	1000	1.8421	2.6334	0.9501	4.9399	1.0078
	3000	1.0618	1.8646	0.7992	5.4925	1.0157
	5000	1.0221	1.4655	0.3546	5.5511	0.9266
	10000	1.0111	1.0349	0.1282	5.4148	0.9693

First-stage diagnostics and limitations. For each scenario, we compare fitted and true mixture conditional operators on its structural function h_0, using 20,000 independent diagnostic observations to estimate

$$\frac{E[\{(\widehat{T}_{\widehat{B}}h_0)(\widehat{Z}) - (T_{\widehat{B}}h_0)(\widehat{Z})\}^2]}{\mathrm{Var}\{h_0(X)\}}.$$

Both operators use the same estimated loading coordinates; the true law enters only evaluation. The code evaluates the raw structural shape rather than its centered and scaled version; the ratio is invariant to that centering and scaling. This diagnostic concerns the specified function, not a supremum over a function class. Figure 5 shows persistent Gaussian approximation error over the sample-size grid and sharply decreasing GMM error: for Interaction, Gaussian error stays near 0.08, while GMM error falls from 0.001547 to 0.000170. These results illustrate why nonlinear SIV requires an accurate conditional distribution, even when conditional first moments suffice for linear SIV. These diagnostics do not establish injectivity or singular-value bounds.

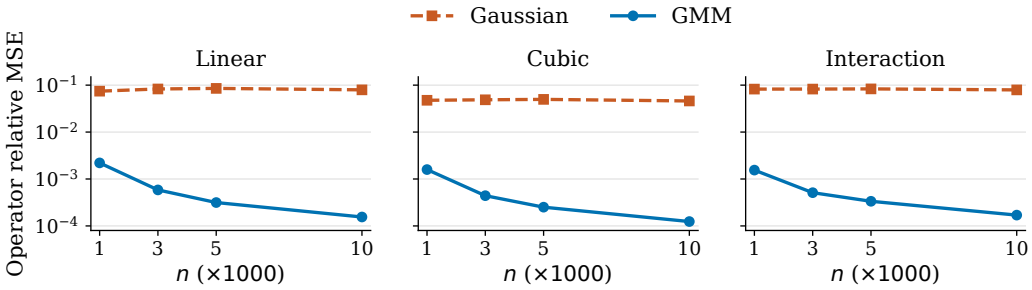


Figure 5: Mean operator relative MSE under mixture treatment noise, on a logarithmic vertical scale. Both samplers are evaluated in the same estimated loading space, over successful FA replications.

Accurate conditional sampling did not ensure reliable structural screening at every sample size. Exact support recovery at $n = 5000$ was $2.5\%/16.5\%/19.0\%$ for Linear/Cubic/Interaction, rising to $100\%/100\%/98\%$ at $n = 10000$. Better loading estimates and support separation accompanied this increase. The fixed 160-epoch schedule and batch size 256 also gave larger samples more gradient updates, preventing attribution to statistical sample size alone. The structural comparison between Sparse Neural SIV (GMM) and Dense Neural SIV (Gaussian) changes both the sampler and the sparsity constraint. Isolating the sampler's contribution would require a sparse Gaussian-SIV comparison. The separate operator diagnostic compares the conditional models directly and shows persistent Gaussian approximation error in this design.