Department of Economics

Athens University of Economics and Business

WORKING PAPER no. 12-2026

Learning Dimension-Reduced State-Adaptive Model Averaging in Regime-Dependent Environments

Stelios Arvanitis and Thanasis Stengos

September 2026

The Working Papers in this series circulate mainly for early presentation and discussion, as well as for the information of the Academic Community and all interested in our current research activity.

The authors assume full responsibility for the accuracy of their paper as well as for the opinions expressed therein.

Learning Dimension-Reduced State-Adaptive Model Averaging in Regime-Dependent Environments

Stelios Arvanitis* Thanasis Stengos†

September 14, 2026

Abstract

We develop a statistical framework for learning low-dimensional state-adaptive predictive model averaging under model uncertainty. Rather than assigning fixed global weights to competing models, the method learns a simplex-valued allocation map that describes how their local predictive credibility changes across observable states. Sparse state learning discerns the coordinates organizing this geometry, and neural-network sieves approximate the resulting nonparametric allocation surface. We establish uniform concentration, sparse-oracle inequalities, set consistency for possibly nonidentified pseudo-true geometries, and finite-iteration guarantees for forward selection. In a cross-country growth data illustration, the method produces economically interpretable adaptive geometry while retaining competitive predictive performance; its out-of-period advantage over strong fixed-weight averaging is more modest. In a similar framework, Monte Carlo evidence shows that accurate adaptive prediction can coexist with uncertainty about the particular support or weight representation, consistent with a set-valued inferential target.

Keywords: Adaptive model averaging; Model uncertainty; Statistical learning; Neural sieves; Sparse state learning; Predictive geometry; Oracle inequalities; Forecast combination.

1 Introduction

Modern statistical learning increasingly operates under two simultaneous forms of uncertainty. The first concerns the predictive mechanism: several models may provide plausible but

*Athens University of Economics and Business, Greece. Email: stelios@aueb.gr
†University of Guelph, Canada. Email: tstengos@uoguelph.ca

incomplete descriptions of the data-generating process, while none is globally correct. The second concerns the environment in which prediction is performed. Relationships that are informative in one region of the covariate space may be less informative elsewhere because of structural heterogeneity, changing regimes, or varying local data geometry. These considerations arise in economic forecasting, finance, personalized medicine, climate science, and ensemble learning (White, 1994; Hastie et al., 2009; Steel, 2020).

Model averaging is a principal response to the first form of uncertainty. Rather than selecting one specification, frequentist and Bayesian model averaging, stacking, forecast combinations, and ensemble methods combine candidate predictions to improve robustness under misspecification (Claeskens and Hjort, 2003; Hansen, 2007; Hansen and Racine, 2012; Steel, 2020; Yang, 2001). Most such procedures nevertheless estimate one global weight vector and therefore assume that the relative credibility of the candidate mechanisms is constant throughout the sample.

That restriction can be consequential in heterogeneous environments. Economic relationships, biological responses, and financial dynamics can change across observable states, so a model that is useful in one region may be less useful in another. Regime-switching and threshold models, mixture-of-experts architectures, adaptive ensembles, and neural gates all recognize this possibility (Tong, 1990; Hansen, 2000; Jacobs et al., 1991; Jordan and Jacobs, 1994). They commonly treat the gate as a computational device, however, rather than making the state-dependent allocation of predictive credibility the statistical object of interest.

This paper places that adaptive predictive geometry at the center of the analysis. Instead of estimating fixed model-averaging weights, we estimate a simplex-valued function assigning local credibility to competing mechanisms throughout the state space. The resulting surface records how the models cooperate and substitute for one another across heterogeneous environments. Fixed averaging is the constant-function special case; model selection and hard regime assignment arise as boundary cases.

The relationship between fixed and state-adaptive averaging parallels the relationship

between parametric and nonparametric regression. A fixed weight vector imposes a low-dimensional global restriction, analogous to a parametric regression function. A state-dependent gate replaces it with an unknown smooth allocation surface, while retaining the candidate models as structured predictive components. The resulting procedure is therefore semi-(non)-parametric in spirit: the (potentially semi-) candidate mechanisms are fixed, but their relative credibility is learned nonparametrically. Network width plays the role of a smoothing or sieve-complexity parameter, trading approximation bias against estimation variance. This analogy also explains why a richer adaptive geometry need not deliver a large OOS MSE gain in every sample: greater flexibility can refine the state-based description while adding estimation variability.

A second contribution concerns representation of the state space. The variables governing predictive heterogeneity may be numerous, correlated, and partially redundant. Our objective is not to estimate a sparse predictive regression, but to discover a low-dimensional subset of states sufficient to organize the adaptive allocation of credibility. The candidate predictive models remain fixed; sparsity is imposed only on the arguments of the weighting function. The estimand is consequently a sparse representation of predictive geometry rather than a sparse predictive equation.

This distinction also clarifies the role of predictive performance. The purpose of adaptive weighting is not merely to minimize MSE relative to every fixed benchmark. It is to refine state-based analysis by revealing where competing mechanisms are locally credible, subject to the requirement that the resulting flexibility does not sacrifice predictive ability. Forecast performance therefore validates and disciplines the learned geometry. A useful adaptive representation may improve MSE, but it may also be substantively valuable when it preserves the performance of a strong fixed-weight predictor while exposing economically meaningful heterogeneity that fixed weights conceal.

We separate this statistical target from its computational implementation. The population problem is defined over continuous simplex-valued allocation maps, and neural networks are

introduced as finite-dimensional sieve approximations. This separation permits shallow and deep networks or other computational variants to be analyzed through their architecture-specific complexity and approximation properties (Hornik, 1991; Yarotsky, 2017; Czarnecki et al., 2017; Belkin et al., 2006). For the implemented shallow ReLU–softmax gate, the theory makes the network-width bias–variance tradeoff explicit.

The theoretical analysis establishes four complementary results. First, finite-net concentration together with the entropy of the implemented gate yields uniform concentration over neural gates and sparse state subsets. Second, oracle inequalities show that the estimator approaches best sparse state-dependent convex approximations generated by the candidate library. Third, the limiting object is formulated as a generally set-valued pseudo-true geometry, reflecting potential observational equivalence among correlated states or locally substitutable candidate mechanisms, and we establish outer and Hausdorff set consistency. Fourth, forward selection is analyzed under weak submodularity, including finite-iteration, neural- optimization, marginal-gain, and empirical-curvature errors.

The set-valued formulation has direct inferential implications. Bootstrap or subsampling draws may move among supports, greedy paths, or neural gates with nearly identical population risk. Weight-function bands can therefore be wide even when the associated adaptive predictors and their MSEs are stable. Such variation need not be evidence against the predictive geometry; it may instead distinguish uncertainty about its coordinate or weight representation from uncertainty about its predictive image. Because sampling, partial identification, numerical optimization, and weak local curvature can all contribute, the empirical analysis treats this mechanism as a qualified interpretation rather than attributing band width to any one source without further diagnostics.

The illustrative empirical application combines competing threshold models of cross-country growth, an environment with substantial model uncertainty and heterogeneous development mechanisms (Levine and Renelt, 1992; Fernandez et al., 2001; Durlauf et al., 2008; Kourtellos et al., 2010; Amini and Parmeter, 2012; Magnus et al., 2010). The nonlinear

gate learns an economically interpretable low-dimensional allocation surface and improves the in-sample and clustered-bootstrap MSE ranking. Under stricter leave-one-period-out evaluation, it retains the best average MSE but is very close to fixed least-squares averaging. We interpret this combination as evidence that the method refines the state-based description without losing predictive ability, rather than as a claim of uniformly large forecasting gains. Monte Carlo experiments further show that predictive accuracy can remain strong under library misspecification and imperfect support or weight identification.

The remainder of the paper is organized as follows. Section 2 introduces predictive geometry, Section 3 develops sparse neural estimation, and Section 4 establishes the statistical theory. Sections 5 and 6 present the empirical application and Monte Carlo evidence, respectively. Section 7 examines whether switching among approximately equivalent representations helps explain the wide bootstrap confidence strips reported in the preceding sections. The Supplement contains proofs, primitive conditions, computational details, and additional inference and robustness results.

2 State-Adaptive Predictive Geometry

Economic relationships may vary across observable states. A predictive mechanism that performs well in one macroeconomic environment may perform poorly in another, even when no candidate model is globally correct (White, 1994; Hastie et al., 2009). We therefore consider a framework in which the relative predictive credibility of a collection of competing models varies across the state space.

Let $Z = (Y, X, S)$ denote a generic observation, where $Y \in \mathbb{R}$ is the response, $X \in \mathbb{R}^p$ contains predictive covariates, and $S \in \mathcal{S} \subseteq \mathbb{R}^d$ is an observable state vector. Throughout the paper, statistical targets are defined with respect to the joint distribution $\mathbb{P}$ of Z, independently of the sampling scheme used to construct the empirical measure. The distinction between X and S is core: variables in X enter the candidate predictive laws, whereas variables in S

determine how their relative predictive credibility changes across economic environments; the two sets of variables need not be disjoint. Notice that for any integrable measurable function h, we write $\mathbb{P}[h] := \int h \, d\mathbb{P} = \mathbb{E}[h(Z)]$, $\mathbb{P}_n[h] := \int h \, d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$. Thus $\mathbb{P}$ and $\mathbb{E}$ are used interchangeably when denoting population integration, while $\mathbb{P}_n$ denotes integration with respect to the empirical law.

Let $\mathcal{M} = \{f_1, \ldots, f_M\}$ denote a library of candidate predictive mechanisms. We consider predictors of the form

$$f_\Gamma(X, S) = \sum_{m=1}^{M} \Gamma_m(S) f_m(X), \qquad \Gamma(S) \in \Delta^M, \tag{1}$$

where $\Delta^M = \{a \in [0,1]^M : \sum_{m=1}^{M} a_m = 1\}$. We refer to the map $\Gamma : \mathcal{S} \to \Delta^M$, where Δ^M is the $M-1$ dimensional simplex, as the predictive geometry. Its component $\Gamma_m(s)$ is interpretable as the local predictive credibility assigned to mechanism m at state s.

This formulation provides a continuous relaxation of conventional regime models. Fixed model averaging is obtained when $\Gamma(S)$ is constant (White, 1994; Hastie et al., 2009). Local model selection is approximated when $\Gamma(S)$ lies near a vertex of the simplex, while interior values correspond to local mixtures of predictive mechanisms. Threshold and switching models can similarly be viewed as special cases in which the allocation map changes sharply across particular regions of the state space (Tong, 1990; Hansen, 2000). The present framework instead allows predictive dominance to vary continuously and nonlinearly with the economic state.

Suppose that

$$Y = f_0(X, S) + u, \; \mathbb{P}[u \mid X, S] = 0. \tag{2}$$

Under squared-error loss–(van der Vaart, 2000; Giné and Nickl, 2021), the population risk associated with a predictive geometry is $R(\Gamma) = \mathbb{P}\left[(Y - f_\Gamma(X,S))^2\right]$, which–due to (2)–satisfies

$$R(\Gamma) = \mathbb{P}[u^2] + \mathbb{P}\left[(f_0(X,S) - f_\Gamma(X,S))^2\right]. \tag{3}$$

Hence estimation of the predictive geometry can be interpreted as an $L^2(\mathbb{P})$-projection of the unknown conditional mean onto the state-adaptive convex hull generated by the candidate library. No candidate model is required to be correctly specified. The target is instead the locally best allocation of predictive credibility among the available mechanisms.

This interpretation is particularly transparent when $f_m(X) = X^\top \beta^{(m)}$. In that case,

$$f_\Gamma(X, S) = X^\top \widetilde{\beta}_\Gamma(S), \qquad \widetilde{\beta}_\Gamma(S) = \sum_{m=1}^{M} \Gamma_m(S)\beta^{(m)}, \tag{4}$$

so that adaptive model averaging induces a state-dependent coefficient surface taking values in the convex hull of the candidate coefficient vectors. Thus nonlinear variation in the predictive geometry generates state-dependent economic relationships even when the individual candidate specifications have fixed coefficients.

The weight representation need not be unique. If candidate predictions are locally redundant, distinct geometries Γ and $\widetilde{\Gamma}$ may satisfy

$$\sum_{m=1}^{M} \Gamma_m(S) f_m(X) = \sum_{m=1}^{M} \widetilde{\Gamma}_m(S) f_m(X) \quad \mathbb{P}\text{-almost surely.} \tag{5}$$

Such geometries are observationally equivalent for prediction–(Molchanov, 2006; Rockafellar and Wets, 2009). Without additional affine-independence restrictions on the candidate library, the identified statistical object is therefore naturally an equivalence class of weight representations, or, more directly, the induced predictive surface. This observation motivates the set-valued pseudo-true targets developed in Section 4.

The same issue arises in the choice of state representation. Let $S_A = P_A S$ denote the projection of S onto coordinates $A \subseteq \{1, \ldots, d\}$. We say that a predictive geometry is s-sparse if $\Gamma(S) = \Gamma_A(S_A)$ for some $|A| \leq s$. The corresponding sparse population target is

$$(\Gamma_s^\star, A_s^\star) \in \underset{\substack{A \subseteq \{1,\ldots,d\} \\ |A| \leq s}}{\arg\min} \inf_{\Gamma_A \in \mathcal{G}_A} \mathbb{E}\left[\left(f_0(X, S) - \sum_{m=1}^{M} \Gamma_{A,m}(S_A) f_m(X)\right)^2\right], \tag{6}$$

where $\mathcal{G}_A$ denotes the admissible class of simplex-valued maps on $\mathcal{S}_A$.

The selected coordinates should not be interpreted as regressors chosen for direct inclusion in the predictive equation. They identify a low-dimensional representation of the state dependence of predictive credibility. Moreover, correlated or informationally overlapping state variables may generate equivalent or nearly equivalent allocation maps. Exact recovery of a unique state subset is therefore generally stronger than necessary. The relevant statistical objective is recovery of the low-dimensional predictive geometry, potentially through any of several equivalent state representations.

The framework separates three distinct objects: the candidate library determines the predictive mechanisms available to the researcher, the predictive geometry determines how their local credibility varies across the state space, and the sampling scheme determines how these population objects are estimated from data. The first two are statistical objects defined under the population distribution $\mathbb{P}$, whereas the third affects only the asymptotic analysis developed in Section 4. The next section develops a neural-sieve approximation to this geometry together with a sparse learning procedure for discovering its relevant state coordinates. Additional formal connections with measurable regime partitions, coefficient-surface representations, and observational equivalence are provided in the Supplement.

3 Sparse Neural Learning of Predictive Geometry

We now formulate estimation of the predictive geometry as a statistical learning problem. We first define the population target over an abstract class of allocation maps and then introduce neural networks as finite-dimensional sieve approximations. This distinction separates the statistical estimand from the computational architecture used to estimate it.

For a state subset A, let $\mathcal{W}_A = \left\{w : \mathcal{S}_A \to \Delta^M : w \text{ continuous}\right\}$. For $w \in \mathcal{W}_A$, define

$$f_{w,A}(X, S) = \sum_{m=1}^{M} w_m(S_A) f_m(X). \tag{7}$$

Suppose that an observed sample induces an empirical measure $\mathbb{P}_n$ approximating the population distribution $\mathbb{P}$. The corresponding empirical and population risks are

$$R_n(w, A) = \mathbb{P}_n\left[(Y - f_{w,A}(X, S))^2\right], \qquad R(w, A) = \mathbb{P}\left[(Y - f_{w,A}(X, S))^2\right]. \tag{8}$$

The statistical target for a fixed state representation is the minimizer of $R(w, A)$ over $\mathcal{W}_A$ (van der Vaart, 2000; Giné and Nickl, 2021; Hastie et al., 2009). The statistical target for a fixed state representation is the minimizer of $R(w, A)$ over $\mathcal{W}_A$–(van der Vaart, 2000; Giné and Nickl, 2021; Hastie et al., 2009).

Due to the complexity of this infinite dimensional optimization problem, we approximate $\mathcal{W}_A$ by neural-network sieves–(Hornik, 1991; Yarotsky, 2017; Mao and Zhou, 2023). Let $g_\theta : \mathcal{S}_A \to \mathbb{R}^M$ be a feedforward neural network and define

$$w_m(S_A; \theta) = \frac{\exp(g_m(S_A; \theta))}{\sum_{j=1}^{M} \exp(g_j(S_A; \theta))}. \tag{9}$$

The softmax transformation ensures that the estimated weights remain in Δ^M for every state. Let $\mathcal{W}_{A,n} = \{w(\cdot; \theta) : \theta \in \Theta_{A,n}\}$ denote the resulting sieve. As its complexity increases with n, the sieve can approximate increasingly rich continuous allocation maps. The uniform approximation properties required for this argument are established in Section 4.

The estimator additionally restricts the allocation map to depend on at most s state coordinates. With $\mathcal{A}_s = \{A \subseteq \{1, \ldots, d\} : |A| \leq s\}$, the ideal empirical sparse-sieve problem is

$$(\widehat{w}_s, \widehat{A}_s) \in \arg\min_{A \in \mathcal{A}_s} \inf_{w \in \mathcal{W}_{A,n}} R_n(w, A). \tag{10}$$

This formulation separates functional learning from state dimensionality reduction. This differs fundamentally from sparse regression, where sparsity is imposed on the predictive equation itself (Tibshirani, 1996; Hastie et al., 2009). Conditional on A, the neural sieve learns how predictive credibility varies with the selected state coordinates; optimization over

A determines which coordinates are needed to organize that variation.

Exact minimization of (10) is combinatorial. We employ forward selection–see for example (Das and Kempe, 2011; Elenberg et al., 2018; Efron et al., 2004). Starting from $A_0 = \varnothing$, at iteration k we choose

$$j_{k+1} \in \arg\min_{j \notin A_k} \inf_{w \in \mathcal{W}_{A_k \cup \{j\}, n}} R_n(w, A_k \cup \{j\}) \tag{11}$$

and set $A_{k+1} = A_k \cup \{j_{k+1}\}$. The procedure continues until the sparsity budget is reached or a prespecified stopping criterion is satisfied. Section 4 provides approximation guarantees under weak submodularity and relates the empirical greedy solution to the corresponding population oracle accumulation set.

In computation, the empirical criterion may be augmented by a regularization functional,

$$Q_n(w, A) = R_n(w, A) + P_n(w, A). \tag{12}$$

The implementations considered in the paper include ridge regularization of the network parameters and, as alternative specifications, entropy and smoothness regularization of the allocation map. More generally, P_n may control the Sobolev regularity of the learned geometry through a curvature penalty such as

$$P_n^C(w, A) = \lambda_{C,n} \sum_{m=1}^{M} |w_m|^2_{H^2(\mathcal{S}_A)}\,, \tag{13}$$

where the H^2 penalizes second-order weak derivatives, providing a classical notion of functional smoothness (Czarnecki et al., 2017). Graph- or manifold-based penalties provide alternative notions of smoothness when the observed state distribution is concentrated near a lower-dimensional structure–(Belkin et al., 2006).

These regularizers alter the finite-sample computational approximation but not the population target when their contribution vanishes at the rates imposed in Section 4. The

theoretical results therefore cover regularized empirical estimation provided the regularization error is asymptotically negligible relative to the relevant stochastic and sieve-approximation errors.

For the baseline implementation, we use a simple single-hidden-layer feedforward network with sixteen hidden units, ReLU activation, standardized state variables, Adam optimization, early stopping, and ridge regularization. Sparse state learning is implemented by forward selection. Linear softmax and entropy-regularized gates are included as benchmarks. The Monte Carlo experiments additionally examine the behavior of sparse and unrestricted neural gates under controlled misspecification.

The architecture is modest: the objective is to isolate the statistical contribution of state-adaptive weighting and sparse geometry learning rather than improvements attributable to increasingly complex neural architectures. Deeper architectures and alternative forms of functional regularization, including Sobolev and graph- or manifold-based penalties, can be accommodated within the same statistical framework under the corresponding complexity, approximation, and vanishing-regularization conditions (Hornik, 1991; Yarotsky, 2017; Czarnecki et al., 2017; Belkin et al., 2006). Further details on architectural definitions, regularization criteria, algorithmic details, and additional computational variants are reported in the Supplement.

4 Statistical Theory

This section establishes the statistical guarantees for sparse state-adaptive predictive geometry. The results are organized around four objects. First, we derive uniform concentration over both neural-gating classes and sparse state subsets. Second, we show that the regularized sparse estimator approaches the ideal sparse adaptive $L^2(\mathbb{P})$-oracle. Third, we establish convergence of the estimated predictive geometry to the possibly set-valued pseudo-true oracle. Finally, we analyze the additional approximation error introduced by greedy forward

selection. The results are thus relevant to familiar themes from statistical learning, empirical-process theory, sieve estimation and greedy approximation to the setting of state-adaptive predictive geometry (van der Vaart, 2000; Giné and Nickl, 2021; Hornik, 1991; Claeskens and Hjort, 2003; Elenberg et al., 2018).

4.1 Setup and Regularity Conditions

The estimator minimizes the potentially regularized criterion

$$Q_n(w, A) = R_n(w, A) + \mathcal{P}_n(w, A). \tag{14}$$

The population target remains defined by the unregularized risk $R(w, A)$. The observations are allowed to be temporally dependent but are assumed to share the common marginal law $\mathbb{P}$. Dependence enters through the finite-net concentration condition below. For each $A \in \mathcal{A}_s$, define $\mathcal{L}_{A,n} = \left\{ \ell_{w,A}(Z) = \big(Y - f_{w,A}(X, S)\big)^2 : w \in \mathcal{W}_{A,n} \right\}$.

Assumption 4.1 (Finite-net empirical-process concentration)**.** *For every $A \in \mathcal{A}_s$, let $\mathcal{W}^{\varepsilon}_{A,n}$ be a deterministic finite ε-net of $\mathcal{W}_{A,n}$ under d_∞, and define the induced sparse loss net $\mathcal{L}^{\varepsilon}_{s,n} = \bigcup_{A \in \mathcal{A}_s} \left\{ \ell_{w,A} : w \in \mathcal{W}^{\varepsilon}_{A,n} \right\}$. There exist constants $C_0, C_1, c > 0$ and a deterministic dependence sequence $b_n \geq 1$ such that, for every $0 < \varepsilon < 1$ and $1 \leq x \leq n$,*

$$\mathbb{P}\left[\max_{\ell \in \mathcal{L}^{\varepsilon}_{s,n}} |(\mathbb{P}_n - \mathbb{P})\ell| > C_0 \left(\sqrt{\frac{\log |\mathcal{L}^{\varepsilon}_{s,n}| + x}{n}} + \frac{b_n(\log |\mathcal{L}^{\varepsilon}_{s,n}| + x)}{n} \right) \right] \leq C_1 e^{-cx}. \tag{15}$$

Assumption 4.1 is imposed on finite loss nets. Uniform concentration results over the complete neural loss class are derived from (15), the neural-gate entropy bound, and the loss oscillation envelope.

Assumption 4.2 (Conditional sub-Gaussian primitives)**.** *The state space $\mathcal{S}$ is compact.*

There exist finite constants $C_Y, C_F, \sigma_Y, \sigma_F > 0$ *such that*

$$|\mathbb{E}[Y \mid X, S]| \le C_Y, \qquad \left\|\mathbb{E}[\widehat{\boldsymbol{f}}(X) \mid S]\right\|_2 \le C_F, \quad \mathbb{P}\text{-}a.s., \tag{16}$$

where $\widehat{\boldsymbol{f}}(X) = \big(\widehat{f}_1(X), \ldots, \widehat{f}_M(X)\big)'$. *Moreover, for every* $t \in \mathbb{R}$,

$$\mathbb{E}\left[\exp t\big(Y - \mathbb{E}[Y \mid X, S]\big)\big)\big|X, S\right] \le \exp\left(\frac{\sigma_Y^2 t^2}{2}\right), \tag{17}$$

and

$$\sup_{\|a\|_2 \le 1} \mathbb{E}\left[\exp\left(ta'\big(\widehat{\boldsymbol{f}}(X) - \mathbb{E}[\widehat{\boldsymbol{f}}(X) \mid S]\big)\right)\Big|S\right] \le \exp\left(\frac{\sigma_F^2 t^2}{2}\right). \tag{18}$$

Assumption 4.2 implies uniform sub-exponential control of the squared losses and their gate increments. In particular, $\|\ell_{w,A} - \ell_{v,A}\|_{\psi_1} \le C_L d_\infty(w, v)$ uniformly over admissible A, w, v, for some finite constant C_L. The conditional formulation is important because the gate is allowed to depend on S. The assumption also supplies the uniform sub-exponential loss envelope and gate-increment control needed for finite-net concentration, but it does not itself restrict dependence. Combined with a sampling condition yielding a scalar Bernstein inequality uniformly over the induced finite loss nets—such as independence, exponentially decaying strong mixing under the tail restrictions stated in the Supplement, conditionally centered exchangeability, or the martingale Bernstein condition—it implies Assumption 4.1; see Proposition 3.4 in the Supplement.

For the implemented shallow architecture, let

$$g_\theta(s_A) = W_2\sigma(W_1 s_A + b_1) + b_2, \qquad \sigma(u) = \max\{u, 0\}, \tag{19}$$

$$w(s_A; \theta) = \mathrm{softmax}\{g_\theta(s_A)\}. \tag{20}$$

The network has H_n hidden units, $W_1 \in \mathbb{R}^{H_n \times |A|}$, $b_1 \in \mathbb{R}^{H_n}$, $W_2 \in \mathbb{R}^{M \times H_n}$, and $b_2 \in \mathbb{R}^M$. Its

parameter count is therefore

$$K_{A,n} = H_n(|A|+1)+M(H_n+1) = H_n(|A|+M+1)+M, \qquad K_{s,n} = H_n(s+M+1)+M. \tag{21}$$

Assumption 4.3 (Implemented neural-sieve complexity)**.** *After rescaling,* $\mathcal{S}_A \subset [-R_S, R_S]^{|A|}$ *for a finite constant* R_S*. All entries of* W_1, b_1, W_2, b_2 *are bounded in absolute value by* $B_n \geq 1$*, uniformly over* $A \in \mathcal{A}_s$*. Define* $\mathfrak{L}_{A,n} = B_n\left(1 + H_n B_n(|A|+1)(R_S+1)\right)$*,* $\mathfrak{L}_{s,n} = \max_{A\in\mathcal{A}_s} \mathfrak{L}_{A,n}$*. For every* $0 < \varepsilon < 1$*,*

$$\log N\left(\varepsilon, \mathcal{W}_{A,n}, d_\infty\right) \leq K_{A,n} \log\left(\frac{C\mathfrak{L}_{A,n}}{\varepsilon}\right). \tag{22}$$

The entropy bound in (22) concerns only the implemented neural-gate class. The conditional moment conditions transfer this gate cover to the induced loss class, that is $\log N\left(\varepsilon, \mathcal{L}_{A,n}, \|\cdot\|_{\psi_1}\right) \leq K_{A,n} \log\left(\frac{CC_L\mathfrak{L}_{A,n}}{\varepsilon}\right)$. Since $|\mathcal{A}_s| \leq \left(\frac{ed}{s}\right)^s$, define

$$V_{s,n} = CK_{s,n} \log\left(C\mathfrak{L}_{s,n}n\right) + s \log\left(\frac{ed}{s}\right) \tag{23}$$

and

$$r_n = \sqrt{\frac{V_{s,n}}{n}} + \frac{b_n V_{s,n}}{n}. \tag{24}$$

Assume $V_{s,n} = o(n)$, $\frac{b_n V_{s,n}}{n} = o(1)$. When $b_n = O(1)$, the square-root component of r_n is first order. The Supplement proves that Assumptions 4.1–4.3 imply the convergence rate $\sup_{A\in\mathcal{A}_s} \sup_{w\in\mathcal{W}_{A,n}} |(\mathbb{P}_n - \mathbb{P})\ell_{w,A}| = O_p(r_n)$.

Entropy conditions of this type are standard in sieve estimation and empirical-process theory (van der Vaart and Wellner, 1996; Giné and Nickl, 2021). The explicit forms of $K_{A,n}$, $\mathfrak{L}_{A,n}$, and $V_{s,n}$ specialize these conditions to the implemented one-hidden-layer ReLU–softmax gate.

Assumption 4.4 (Vanishing regularization)**.** *There exists a deterministic sequence* $\pi_n \to 0$

such that

$$\sup_{A\in\mathcal{A}_s}\sup_{w\in\mathcal{W}_{A,n}}|\mathcal{P}_n(w,A)|\le\pi_n. \tag{25}$$

Thus $\pi_n=o(1)$ preserves the unregularized population target, $\pi_n=O(r_n)$ preserves the first-order convergence rate, and $\pi_n=o(r_n)$ makes regularization asymptotically negligible relative to the leading statistical error.

To define the ideal sparse target, let $\mathcal{W}_A^0$ denote a compact class of admissible simplex-valued weight functions on $\mathcal{S}_A$. The Supplement constructs $\mathcal{W}_A^0$ using bounded Sobolev classes and establishes the corresponding compactness and neural-approximation properties.

Define

$$a_n=\max_{A\in\mathcal{A}_s}\sup_{w\in\mathcal{W}_A^0}\inf_{v\in\mathcal{W}_{A,n}}d_\infty(v,w), \tag{26}$$

and assume $a_n\to 0$. For the implemented shallow network, the Supplement provides the explicit bound (Mao and Zhou, 2023)

$$a_n(H_n)\le\beta_n+CR_{g,n}(\log H_n)^{s+1/2}H_n^{-\alpha_{\nu,s}},\qquad \alpha_{\nu,s}=\frac{\nu}{s}\frac{s+2}{s+4}, \tag{27}$$

under the stated smoothness, interior-surrogate, and coefficient-compatibility conditions. The Sobolev assumptions on $\mathcal{W}_A^0$ enter through this approximation bound rather than through (22). The ideal sparse approximation risk is

$$R_s^\star=\min_{A\in\mathcal{A}_s}\min_{w\in\mathcal{W}_A^0}R(w,A). \tag{28}$$

Under $\mathbb{P}\left[Y^2+\|\widehat{\boldsymbol{f}}(X)\|_1^2\right]<\infty$, the risk is Lipschitz in d_∞ over the relevant classes. Consequently, the induced population-risk approximation error is bounded by Ca_n. Thus a_n retains its original interpretation as the neural-sieve bias in the subsequent oracle inequalities.

Remark 4.5 (Balanced shallow-network rate)**.** *For fixed s and M, the parameter count satisfies $K_{s,n}\asymp H_n$. Ignoring logarithmic factors and the linear Bernstein remainder, the width-dependent estimation and approximation terms are therefore $\sqrt{\frac{H_n}{n}}$ and $H_n^{-\alpha_{\nu,s}}$. If β_n,*

regularization, and optimization errors are asymptotically negligible, $R_{g,n} = O(1)$, $\log \mathfrak{L}_{s,n} = O(\log n)$, *and* B_n *is large enough to contain the constructive approximants, balancing these terms gives*

$$H_n \asymp n^{1/(2\alpha_{\nu,s}+1)}, \qquad r_n + a_n = \widetilde{O}_p\left(n^{-\alpha_{\nu,s}/(2\alpha_{\nu,s}+1)}\right), \tag{29}$$

where $\widetilde{O}_p$ *suppresses logarithmic factors. This is the balanced rate for the stated Lipschitz risk-approximation bound; if approximation enters excess risk quadratically, the balancing exponent changes accordingly. The derivation and its feasibility conditions are given in the Supplement.*

Remark 4.6 (Deep-network substitution principle)**.** *The downstream theory is not restricted to one hidden layer. For a deep ReLU–softmax sieve, let* $K_{s,n}^{\mathrm{deep}}$ *denote its effective parameter count,* $\mathfrak{L}_{s,n}^{\mathrm{deep}}$ *its parameter-to-gate Lipschitz scale, and* a_n^{deep} *an architecture-specific uniform approximation bound. Define*

$$V_{s,n}^{\mathrm{deep}} = CK_{s,n}^{\mathrm{deep}} \log\big(C\mathfrak{L}_{s,n}^{\mathrm{deep}} n\big) + s\log(ed/s), \qquad r_n^{\mathrm{deep}} = \sqrt{\frac{V_{s,n}^{\mathrm{deep}}}{n}} + \frac{b_n V_{s,n}^{\mathrm{deep}}}{n}.$$

Provided the corresponding entropy, coefficient-control, and approximation conditions hold, all oracle, set-convergence, and forward-selection results remain valid after the substitutions

$$(r_n, a_n, K_{s,n}, \mathfrak{L}_{s,n}) \longmapsto (r_n^{\mathrm{deep}}, a_n^{\mathrm{deep}}, K_{s,n}^{\mathrm{deep}}, \mathfrak{L}_{s,n}^{\mathrm{deep}}).$$

Depth may improve approximation for a given parameter budget, especially for compositional targets, but it also changes entropy, coefficient growth, and optimization error. Consequently, no universal deep-network rate or dominance over the implemented shallow architecture is asserted without specifying the exact depth, width, sparsity, and norm constraints. The Supplement states the formal substitution result.

4.2 Uniform Concentration and Sparse-Oracle Risk

The first result derives uniform concentration over both the implemented neural-gate classes and the admissible sparse state representations.

Theorem 4.7 (Uniform sparse-state concentration)**.** *Under Assumptions 4.1-4.4*

$$\sup_{A\in\mathcal{A}_s}\sup_{w\in\mathcal{W}_{A,n}}|Q_n(w,A)-R(w,A)|=O_p(r_n+\pi_n). \tag{30}$$

The rate $r_n=\sqrt{\frac{V_{s,n}}{n}}+\frac{b_nV_{s,n}}{n}$ has three components. The term $K_{s,n}\log(C\mathfrak{L}_{s,n}n)$ in $V_{s,n}$ is the cost of learning the nonlinear neural allocation map, while $s\log(ed/s)$ is the combinatorial cost of searching over sparse state representations. The second component of r_n, $b_nV_{s,n}/n$, records the linear part of the finite-net Bernstein bound and the effect of temporal dependence. When $b_n=O(1)$ and $V_{s,n}=o(n)$, the square-root component is first order.

The proof combines the finite-net inequality in Assumption 4.1 with the neural-gate entropy bound in Assumption 4.3. The conditional sub-Gaussian conditions in Assumption 4.2 provide the integrable random envelope required to pass from the finite net to the complete neural class. It is given in the Supplement.

The neural criterion need not be minimized exactly. Let $\varepsilon_n^{\mathrm{NN}}\geq 0$ denote the achieved criterion-value tolerance and suppose that $(\widehat{w}_s,\widehat{A}_s)$ satisfies

$$Q_n(\widehat{w}_s,\widehat{A}_s)\leq\inf_{\substack{A\in\mathcal{A}_s\\ w\in\mathcal{W}_{A,n}}}Q_n(w,A)+\varepsilon_n^{\mathrm{NN}}. \tag{31}$$

The exact regularized sparse-sieve estimator is the special case $\varepsilon_n^{\mathrm{NN}}=0$. The tolerance $\varepsilon_n^{\mathrm{NN}}$ measures the achieved objective gap; it does not require a claim that a particular optimization algorithm attains a parameter-space global minimum.

Theorem 4.8 (Approximate ideal sparse-oracle inequality)**.** *Under the preceding assumptions*

and the moment condition used to obtain the risk-Lipschitz bound,

$$R(\widehat{w}_s, \widehat{A}_s) - R_s^\star \leq 2\zeta_n + C_R a_n + 2\pi_n + \varepsilon_n^{\mathrm{NN}}, \tag{32}$$

where $\zeta_n = \sup_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_{A,n}} |R_n(w, A) - R(w, A)| = O_p(r_n)$ *and* $C_R < \infty$ *is the risk-Lipschitz constant. Consequently,*

$$R(\widehat{w}_s, \widehat{A}_s) - R_s^\star = O_p\left(r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}}\right). \tag{33}$$

If $r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}} \to 0$, *then* $R(\widehat{w}_s, \widehat{A}_s) \xrightarrow{p} R_s^\star$. *If, in addition,* $a_n = O(r_n)$, $\pi_n = O(r_n)$, $\varepsilon_n^{\mathrm{NN}} = O_p(r_n)$, *then*

$$R(\widehat{w}_s, \widehat{A}_s) - R_s^\star = O_p(r_n). \tag{34}$$

The theorem is analogous in spirit to classical oracle results for model averaging and statistical learning, but the oracle here is a sparse state-dependent predictive geometry rather than a single predictive model (Claeskens and Hjort, 2003; Hansen, 2007; Hansen and Racine, 2012). Universal misspecification is allowed; the target is the best state-dependent convex approximation generated by the candidate model library. Moreover, the result does not by itself imply identification of the selected support or of the weight function.

4.3 Pseudo-True Geometry and Set Consistency

The sparse oracle need not be unique. Define the pseudo-true sparse oracle set

$$\mathcal{E}_s^\star = \left\{(A, w) : A \in \mathcal{A}_s,\ w \in \mathcal{W}_A^0,\ R(w, A) = R_s^\star\right\}. \tag{35}$$

The Supplement establishes that $\mathcal{E}_s^\star$ is nonempty and compact under the compactness conditions imposed on the ideal weight classes. Under the moment assumptions, continuity of $R(w, A)$ under the uniform metric follows from $\mathbb{P}\left[Y^2 + \|\widehat{\boldsymbol{f}}(X)\|_1^2\right] < \infty$. Set-valued

statistical targets arise naturally whenever several state representations or allocation rules are observationally equivalent (Molchanov, 2006; Rockafellar and Wets, 2009).

To compare weight functions defined on different state subsets, identify every $w_A : \mathcal{S}_A \to \Delta^M$ with its canonical extension $\widetilde{w}_A(s) = w_A(P_A s),\ s \in \mathcal{S}$, where $P_A s$ denotes the projection of s onto the coordinates in A. Consider $d_\infty(w_A, w_B) = \sup_{s\in\mathcal{S}} \|\widetilde{w}_A(s) - \widetilde{w}_B(s)\|_\infty$ and, for state-weight pairs, define the discrete-uniform metric

$$d\big((A,w),(B,v)\big) = \mathbf{1}\{A \neq B\} + d_\infty(w,v). \tag{36}$$

For a pair $e = (A, w)$ and a nonempty set $\mathcal{E}$, write $d(e, \mathcal{E}) = \inf_{e^\star\in\mathcal{E}} d(e, e^\star)$.

Assumption 4.9 (Risk separation)**.** *For every $\varepsilon > 0$,*

$$\Delta_s(\varepsilon) := \inf\left\{ R(w,A) - R_s^\star : \begin{array}{l} A \in \mathcal{A}_s,\ w \text{ belongs to an admissible ideal or neural-sieve class,} \\ d\big((A,w),\mathcal{E}_s^\star\big) \geq \varepsilon \end{array} \right\} > 0. \tag{37}$$

The infimum is understood uniformly over the neural sieves used for sufficiently large n.

Assumption 4.9 requires every admissible state-weight pair that remains uniformly separated from the complete pseudo-true oracle set to have strictly larger population risk. Let $Q_{s,n}^{\inf} = \inf_{\substack{A\in\mathcal{A}_s \\ w\in\mathcal{W}_{A,n}}} Q_n(w,A)$. For $\tau_n \geq 0$, define the regularized empirical near-minimizer set

$$\widehat{\mathcal{E}}_{s,n}(\tau_n) = \left\{ (A,w) : \begin{array}{l} A \in \mathcal{A}_s,\ w \in \mathcal{W}_{A,n}, \\ Q_n(w,A) \leq Q_{s,n}^{\inf} + \tau_n + \varepsilon_n^{\mathrm{NN}} \end{array} \right\}. \tag{38}$$

Here τ_n determines the inferential near-minimizer set, whereas $\varepsilon_n^{\mathrm{NN}}$ is the numerical objective tolerance introduced in (31). If the criterion is minimized exactly, set $\varepsilon_n^{\mathrm{NN}} = 0$.

Theorem 4.10 (Set consistency)**.** *Suppose that Assumptions 4.1, 4.2, 4.3, 4.4, and 4.9 hold. If $r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}} + \tau_n \to 0$, then $\sup_{(A,w)\in\widehat{\mathcal{E}}_{s,n}(\tau_n)} d\left((A,w),\mathcal{E}_s^\star\right) \xrightarrow{p} 0$. If, in addition,*

every element of $\mathcal{E}_s^\star$ *admits the uniform neural-sieve approximation in* (26)*,* $\tau_n \to 0$*, and* $r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}} = o(\tau_n)$*, then* $d_H\left(\widehat{\mathcal{E}}_{s,n}(\tau_n), \mathcal{E}_s^\star\right) \xrightarrow{p} 0$.

The first conclusion is an outer-consistency result: every empirical near minimizer approaches the pseudo-true oracle set. The second also establishes the reverse inclusion. Its stronger tolerance condition is needed because the empirical near-minimizer set must be wide enough to contain neural approximations to every element of $\mathcal{E}_s^\star$, rather than only an approximation to one population minimizer.

The theorem formalizes the recovery of a predictive geometry. Correlated state coordinates, locally substitutable candidate predictions, or several neural allocation rules with nearly equal risk may prevent literal support or weight-function identification while still producing near-optimal adaptive predictors. The appropriate asymptotic target is therefore the complete set $\mathcal{E}_s^\star$.

Under a local error bound, the relevant set-distance rate is governed by the statistical, sieve, regularization, neural-optimization, greedy, and curvature remainders. The associated bootstrap confidence strips may therefore reflect both sampling variation around $\mathcal{E}_s^\star$ and switching among near-optimal supports, paths, or neural basins.

4.4 Forward Selection

Exact minimization over $\mathcal{A}_s$ is combinatorial; the greedy procedure used in the implementation is now studied. Define the population profile risk $L(A) = \inf_{w \in \mathcal{W}_A^0} R(w, A)$ and the corresponding gain $G(A) = L(\varnothing) - L(A)$. Assume that G is monotone and weakly submodular: for some $\gamma_s \in (0, 1]$,

$$\sum_{j \in B} \left(G(A \cup \{j\}) - G(A)\right) \geq \gamma_s \left(G(A \cup B) - G(A)\right) \tag{39}$$

for every pair of admissible disjoint sparse sets A and B; see, for example, Das and Kempe (2011); Elenberg et al. (2018).

Let A_k^{FS} denote the population forward-selection set after $k \leq s$ iterations, and let $A_s^\star \in \arg\max_{A \in \mathcal{A}_s} G(A) = \arg\min_{A \in \mathcal{A}_s} L(A)$.

Theorem 4.11 (Population finite-iteration forward-selection guarantee)**.** *For every $k \leq s$,*

$$G(A_k^{\mathrm{FS}}) \geq \left(1 - \left(1 - \frac{\gamma_s}{s}\right)^k\right) G(A_s^\star) \geq \left(1 - e^{-k\gamma_s/s}\right) G(A_s^\star).$$

Equivalently,

$$L(A_k^{\mathrm{FS}}) - L(A_s^\star) \leq \rho_{s,k}^{\mathrm{FS}}, \qquad \rho_{s,k}^{\mathrm{FS}} := e^{-k\gamma_s/s} G(A_s^\star). \tag{40}$$

At $k = s$, this gives the familiar bound $G(A_s^{\mathrm{FS}}) \geq (1 - e^{-\gamma_s}) G(A_s^\star)$.

The finite-k formulation is relevant when forward selection is stopped before reaching the maximal sparsity level. Increasing k reduces the population truncation error $\rho_{s,k}^{\mathrm{FS}}$, but the empirical procedure must estimate more profiled marginal gains and may therefore accumulate additional statistical and numerical error.

Weak submodularity may arise from restricted curvature of the population gating problem. In particular, if the Gâteaux Hessian of the population profile objective has restricted eigenvalues bounded between $m_s > 0$ and $M_s < \infty$ along the relevant sparse perturbation directions, then $\gamma_s \geq c_\gamma \frac{m_s}{M_s}$ for an appropriate constant $c_\gamma > 0$. Poor conditioning, a small restricted minimum eigenvalue, or strongly complementary state variables therefore weaken the greedy guarantee.

For the empirical procedure, let $\widehat{A}_k^{\mathrm{FS}}$ denote the subset selected after $1 \leq k \leq s$ iterations and define $\delta_n = r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}}$. The Supplement shows that uniform restricted-Hessian convergence at rate $h_n = o(1)$ implies $\widehat{\gamma}_{s,n} = \gamma_s + O_p(h_n)$, while uniform profile accuracy implies that the empirical marginal gains approximate their population counterparts at rate

$O_p(\delta_n)$. Consequently,

$$L(\widehat{A}_k^{\mathrm{FS}}) - L(A_s^\star) \le \rho_{s,k}^{\mathrm{FS}} + O_p(k\delta_n + h_n), \tag{41}$$

$$d\left((\widehat{A}_k^{\mathrm{FS}}, \widehat{w}_{\widehat{A}_k^{\mathrm{FS}}}), \mathcal{E}_s^\star\right) = O_p\left[\left(\rho_{s,k}^{\mathrm{FS}} + k\delta_n + h_n\right)^\alpha\right], \tag{42}$$

under a local metric error bound with exponent $\alpha \in (0, 1]$. For fixed s, the factor k is absorbed into the stochastic order.

Thus forward selection approaches the exact sparse oracle only when $\rho_{s,k}^{\mathrm{FS}} \to 0$; otherwise its natural limit is the population forward-selection accumulation set. Nearly tied gains, neural optimization error, and weak empirical curvature may consequently induce switching among approximately optimal paths across bootstrap samples.

4.5 Implications

For the implementable forward-selection estimator, the preceding results give the bound

$$\rho_{s,k}^{\mathrm{FS}} + O_p\big[k\left(r_n + C_R a_n + \pi_n + \varepsilon_n^{\mathrm{NN}}\right) + h_n\big]. \tag{43}$$

The terms inside parentheses represent statistical, neural-sieve, regularization, and neural-optimization errors. Forward selection adds the population greedy gap $\rho_{s,k}^{\mathrm{FS}}$, accumulates profile errors over k iterations, and depends on restricted-curvature estimation through h_n. These components need not be independent.

Recovery is therefore expected generally to be set-valued and need not imply support consistency. Different state subsets, greedy paths, or neural gates may generate nearly equivalent predictive geometries, particularly when economic state variables or candidate predictions are locally substitutable.

5 Empirical Application

Cross-country growth provides a natural application because competing theories emphasize capital accumulation, institutions, geography, and international integration, while no single specification performs uniformly well across countries (Mankiw et al., 1992; Acemoglu et al., 2001; Gallup et al., 1999; Durlauf et al., 2008). Related work has therefore emphasized both model uncertainty and threshold heterogeneity (Fernandez et al., 2001; Sala-i Martin et al., 2004; Kourtellos et al., 2010, 2013). We ask whether the local credibility of such competing mechanisms can be learned from a low-dimensional subset of the macroeconomic state space. Seven threshold models are combined using fixed, linear-gated, and neural-gated averaging; sparse states are chosen by forward selection. Additional cross-fitted, bootstrap, and specification diagnostics are reported in the Supplement.

5.1 Candidate Predictive Mechanisms and Data

Each candidate combines common growth determinants with a different threshold variable. Let $X_t = (\mathrm{GDP0}_t, \mathrm{POP}_t, \mathrm{INV}_t, \mathrm{HC}_t)'$, where the components denote initial income, population growth, the investment share, and educational attainment. Candidate m is

$$Y_t = X_t^\top \beta_m + \delta_{m,L}(q_{mt} - \gamma_m)_- + \delta_{m,H}(q_{mt} - \gamma_m)_+ + u_t, \tag{44}$$

where $(x)_- = xI(x \le 0)$, $(x)_+ = xI(x > 0)$, and γ_m minimizes the residual sum of squares over a trimmed grid (Hansen, 2000; Seo and Linton, 2007). The seven threshold variables are inflation, life expectancy, fertility, government consumption, debt, democracy, and openness. These models are treated as complementary, possibly misspecified local approximations rather than candidate true data-generating processes.

The balanced panel contains 71 countries observed over nine non-overlapping five-year periods from 1980 to 2023, for 639 country-period observations. The data combine the Penn World Tables, World Development Indicators, IMF public-debt data, Barro–Lee educational-

attainment data, and Polity V; complete definitions and sources appear in the Supplement. State variables are standardized before neural estimation. Openness is defined as the trade balance relative to GDP, so positive values describe export-oriented economies and negative values describe external-deficit economies. This measure can distinguish countries with similar debt ratios but different external financing conditions.

Table 1: Estimated Threshold Models

Threshold Variable	Estimated Threshold	RSS
Inflation	0.1066	51.03
Life Expectancy	4.2228	49.40
Fertility	0.0282	50.76
Government Consumption	0.2388	49.78
Debt	1.0089	50.60
Democracy	4.0506	50.11
Openness	-0.2279	51.18

Life expectancy and government consumption have the smallest residual sums of squares in Table 1, but the differences are modest. The absence of a uniformly dominant candidate motivates learning state-dependent combinations rather than selecting one threshold mechanism.

5.2 Forecasting Performance

We compare the best single model, equal weights, fixed least-squares averaging, linear and nonlinear softmax gates, and an entropy-regularized neural gate.

Table 2: Empirical Model-Averaging Performance

Method	MSE
Best Single	0.0773
Naive Average	0.0776
Fixed LS Average	0.0765
Linear Gate	0.0774
Neural Gate	0.0719
Entropy Neural Gate	0.0771

The nonlinear neural gate has the smallest MSE, improving on fixed LS by 5.98% and on the best single model by 7.00%. The linear and entropy-regularized gates remain close to fixed averaging, indicating that the gain comes from a flexible nonlinear allocation rule rather than state dependence alone.

Table 3: Predictive Superiority Tests

Benchmark	Model	Mean Loss Difference	DM Statistic	One-Sided p-Value
Best Single	Neural Gate	0.0054	4.521	3.08×10^{-6}
Fixed LS Average	Neural Gate	0.0046	3.898	4.85×10^{-5}
Linear Gate	Neural Gate	0.0055	4.040	2.67×10^{-5}
Entropy Neural Gate	Neural Gate	0.0052	4.442	4.46×10^{-6}

The one-sided Diebold–Mariano tests in Table 3, based on HAC estimates of the long-run variance, reject equal predictive performance against each benchmark (Diebold and Mariano, 2002). Country-clustered bootstrap results in the Supplement likewise preserve the MSE ranking: the neural gate's bootstrap mean MSE is 0.0675, compared with 0.0761 for fixed LS, and the bootstrap interval for their relative gain is $[0.0566, 0.2176]$.

The stricter leave-one-period-out evidence is more qualified. The neural gate retains the smallest blocked MSE (0.0808, versus 0.0809 for fixed LS), but the gain is economically small; in the corresponding cross-fitted comparison, the mean loss difference is 0.0001 and the one-sided p-value is 0.481. In a standard nonparametric bias–variance interpretation, the

nonlinear gate can reduce approximation bias by adapting to heterogeneous states, while estimation of the richer gate and sparse representation adds variance. We therefore interpret the evidence as strong support for informative adaptive geometry and in-sample fit, but only a modest OOS advantage over the strongest fixed-weight benchmark.

5.3 Learning the Predictive Geometry

The gate weights describe the local credibility of each mechanism. Table 4 reports their sample averages, which summarize—but do not identify—the heterogeneous allocation surface.

Table 4: Empirical Average Model Weights

Candidate Model	Fixed LS	Linear Gate	Neural Gate	Entropy Neural Gate
Inflation	0.000	0.002	0.050	0.140
Life Expectancy	0.304	0.004	0.311	0.147
Fertility	0.134	0.004	0.027	0.141
Government Consumption	0.289	0.329	0.442	0.146
Debt	0.054	0.004	0.025	0.142
Democracy	0.219	0.655	0.124	0.144
Openness	0.000	0.002	0.021	0.140

Government consumption and life expectancy receive the largest average neural weights, while most mechanisms receive positive mass somewhere in the sample. These averages are not model-selection probabilities: the substantive object is the observation-specific allocation map.

Figure 1 displays a one-dimensional cross-section of that map against openness, holding the remaining states at their medians. The trade-balance definition makes openness an economically useful axis for distinguishing external financing environments.

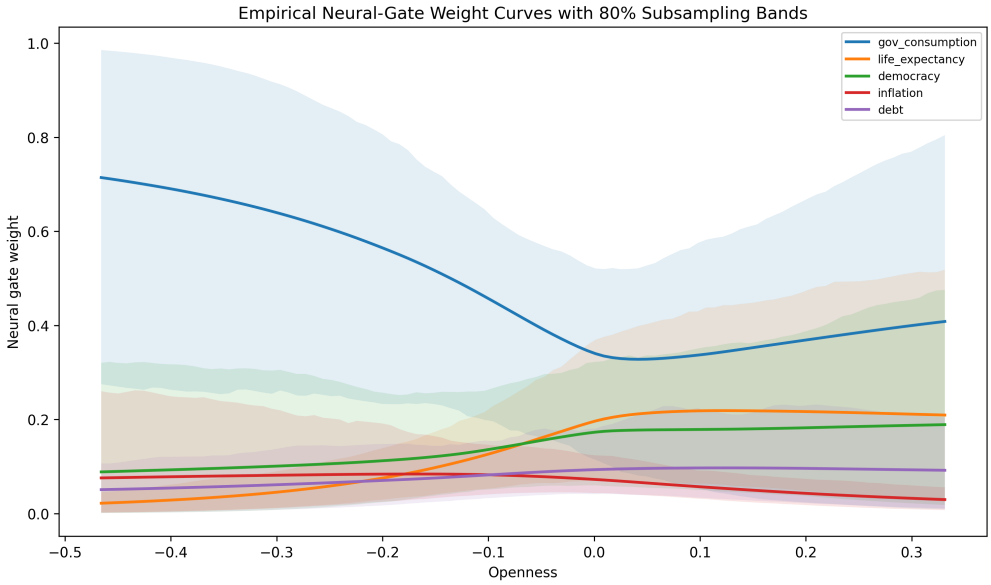

Figure 1: Empirical Neural-Gate Weight Curves with 80% Subsampling Bands

The curves vary smoothly, with government consumption and life expectancy dominant over much of the observed range and other mechanisms important more locally. Candidate weights become more similar for highly export-oriented economies, whereas external-deficit economies display greater differentiation. The bands are constructed by drawing 80% of countries without replacement, re-estimating the entire pipeline, and evaluating each fitted gate on a common grid.

Their substantial width need not represent pointwise sampling noise alone. The theory permits several supports, greedy paths, or neural gates to have nearly identical risks, while weak restricted curvature and numerical optimization can make the selected representation sensitive to resampling. The combination of comparatively stable bootstrap MSE gains with wider gate strips is therefore consistent with partial identification and switching among approximately optimal solutions–see Section 7 and the Supplement.

Overall, the empirical results support an economically interpretable low-dimensional predictive geometry, while the OOS and uncertainty results appropriately qualify claims about the magnitude of forecasting gains and point identification of its particular representation.

6 Monte Carlo Experiments

The simulations evaluate sparse state-adaptive averaging as an $L^2(\mathbb{P})$-approximation method. Experiment 1 asks whether the procedure recovers a low-dimensional adaptive geometry when the candidate library contains the relevant local mechanisms; Experiment 2 introduces an omitted nonlinear interaction. In both designs, the joint covariate-state vector is resampled from the empirical application and innovations are drawn from centered empirical residuals, preserving the observed support, dependence, and marginal asymmetry. Results use $R = 5000$ replications; full calibrations, selection frequencies, regional errors, and weight-function bands appear in the Supplement.

6.1 Designs and Comparators

Let $W_t = (X_t, S_t)$ contain the four baseline controls and the seven candidate regime coordinates used in Section 5. In each replication,

$$W_t^{\mathrm{MC}} \sim \widehat{F}_W, \qquad Y_t^{\mathrm{MC}} = f_0(X_t^{\mathrm{MC}}, S_t^{\mathrm{MC}}) + u_t^{\mathrm{MC}}.$$

In Experiment 1,

$$f_0(X_t, S_t) = \sum_{m=1}^{7} \pi_m(S_t) f_m(X_t, q_{mt}), \tag{45}$$

where f_m are the empirical threshold mechanisms and the sparse softmax weights depend principally on inflation, debt, and democracy. No single model is globally correct, but the library contains the local mechanisms.

Experiment 2 adds the unavailable interaction

$$f_8^0(X_t, S_t) = \mathrm{base}_t + \eta_1(\mathrm{debt}_t - \gamma_d)_+(\mathrm{openness}_t - \gamma_o)_+ + \eta_2(\mathrm{inflation}_t - \gamma_\pi)_+(\mathrm{debt}_t - \gamma_d)_+. \tag{46}$$

The econometrician still averages only the seven single-threshold models, so the target is the best state-dependent convex projection rather than the true conditional mean.

We compare the best single model, equal and fixed-LS averaging, linear and nonlinear full-state gates, entropy-regularized and sparse forward-selection neural gates, and, in Experiment 1, an oracle-state gate. Performance is measured by noisy-outcome MSE and by

$$\mathrm{MSE}_f = \frac{1}{N}\sum_{t=1}^{N}\left(f_0(X_t^{\mathrm{MC}}, S_t^{\mathrm{MC}}) - \widehat{f}_t\right)^2, \tag{47}$$

the primary measure of conditional-mean approximation.

6.2 Experiment 1: Sparse Adaptive-Mixture Geometry

Table 5: Experiment 1: Adaptive-Mixture Design

Method	MSE_Y	MSE_f	Gain vs Fixed LS	Gain vs Best Single
Best Single	1.013	0.940	17.1%	0.0%
Naive Average	2.762	2.687	-137.0%	-186.8%
Fixed LS Average	1.231	1.157	0.0%	-23.5%
Linear Full Gate	0.782	0.708	36.8%	24.2%
Neural Full Gate	0.130	0.066	94.2%	92.9%
Entropy Neural Gate	0.164	0.098	91.4%	89.5%
Sparse Forward Neural Gate	0.156	0.085	92.6%	91.0%
Oracle-State Neural Gate	0.144	0.073	93.6%	92.2%

The full neural gate has the smallest feasible approximation error ($\mathrm{MSE}_f = 0.0662$), compared with 0.9397 for the best single model and 1.1569 for fixed LS. The sparse gate attains 0.0847 and the oracle-state gate 0.0733, showing that learning a low-dimensional state representation preserves most of the nonlinear-gating gain.

Literal support recovery is less stable. Inflation is selected in 92.1% of replications, but debt and democracy are selected less frequently, while correlated non-DGP coordinates sometimes enter. The combination of low MSE_f and dispersed support selection is consistent with recovery of a near-optimal predictive representation rather than a uniquely identified structural support.

6.3 Experiment 2: Adversarial Misspecification

Table 6: Experiment 2: Adversarial Misspecification Design

Method	MSE_Y	MSE_f	Gain vs Fixed LS	Gain vs Best Single
Best Single	1.495	1.421	33.2%	0.0%
Naive Average	4.554	4.481	-106.9%	-219.8%
Fixed LS Average	2.267	2.193	0.0%	-55.1%
Linear Full Gate	2.941	2.868	-30.3%	-98.4%
Neural Full Gate	0.314	0.251	88.3%	82.7%
Entropy Neural Gate	0.344	0.280	86.9%	80.6%
Sparse Forward Neural Gate	0.376	0.306	85.8%	78.9%
Oracle-State Neural Gate	0.326	0.259	87.9%	82.2%

Approximation errors rise because the interaction in (46) is unavailable, but nonlinear adaptive averaging continues to dominate fixed aggregation. The full, oracle-state, entropy, and sparse neural gates attain respective MSE_f values of 0.2513, 0.2588, 0.2797, and 0.3056. The sparse procedure therefore retains most of the full-state gain even under library misspecification. Openness is selected more frequently, as expected because the omitted mechanism is active in high-debt/high-openness regions.

6.4 Summary of Monte Carlo Evidence

Three findings emerge. Nonlinear gating accounts for most of the gain over fixed averaging; sparse forward selection preserves most of the full-state performance; and both conclusions survive an incomplete candidate library. At the same time, dispersed state selection and wide weight bands show that predictive accuracy does not require point identification of a support or mechanism decomposition. The evidence therefore supports the paper's set-valued pseudo-true interpretation of adaptive predictive geometry–see also the following section and the Supplement.

7 Representation Uncertainty and Predictive Stability

Empirical resampling diagnostics indicate that the width of the gate-weight strips partly reflects switching among nearly tied sparse representations. Across subsamples and bootstrap replications, the procedure visits many supports and ordered forward-selection paths, with median best-versus-second-best gain gaps of approximately 2×10^{-5}. Fixing the original-sample openness support reduces the mean width of the central 80% strip by approximately 15%, whereas increasing the number of neural restarts reduces it by less than 1%. Support switching is also more strongly associated with variation in the gate than with variation in its induced prediction. The strips should therefore be interpreted as uncertainty about a potentially set-valued predictive representation, rather than as conventional confidence bands around a uniquely identified allocation surface. At the same time, the relatively modest reduction in the 90% strips shows that within-support sampling variation remains important. Detailed diagnostics and counterfactual comparisons are reported in Section 7 of the Supplement.

The Monte Carlo evidence reinforces this distinction between representation and predictive stability. Exact support recovery is infrequent, and the sparse procedure visits numerous supports and greedy paths, yet its predictive risk remains substantially below that of fixed averaging in both designs. Moreover, between-support variation accounts for a larger share of gate-weight variation than of predictive-image variation. These results support the interpretation of sparse state learning as recovery of a useful low-dimensional predictive geometry rather than necessarily of a unique state-variable representation. Consistently with this interpretation, the empirical cross-fitted MSE of the neural gate is essentially the same as that of fixed least-squares averaging, with no statistically significant predictive-superiority result. The main empirical contribution is thus the refinement of state-dependent analysis without a detectable loss of out-of-sample predictive ability, rather than a claim of uniformly improved MSE.

8 Conclusion

This paper develops a framework for learning state-adaptive predictive geometries under model uncertainty. Rather than restricting model averaging to a fixed vector of global weights, the method estimates a simplex-valued allocation surface describing how the local credibility of competing mechanisms changes across observable environments. Sparse state learning identifies a low-dimensional representation organizing this surface, while neural sieves approximate the underlying functional target. The resulting objective is not sparsification of the predictive regression, but refinement of the state-based analysis of model credibility.

The comparison between fixed and adaptive weights has the familiar form of a parametric–nonparametric tradeoff. Fixed averaging is parsimonious but cannot represent state heterogeneity; the adaptive gate reduces this restriction at the cost of estimating a richer allocation function. The theory formalizes the associated statistical, sieve, regularization, optimization, and greedy-search errors. It also treats the pseudo-true geometry as potentially set-valued: correlated states and substitutable candidate mechanisms can yield different supports or weight maps with nearly identical predictive implications. Bootstrap and subsampling variation may consequently reflect switching among approximately optimal representations as well as ordinary sampling noise and weak local curvature.

The empirical growth application illustrates the intended role of the method. The nonlinear gate uncovers an economically interpretable geometry and improves in-sample and clustered-bootstrap MSE relative to fixed averaging. Its advantage becomes much smaller under leave-one-period-out evaluation, where it performs essentially as well as the strongest fixed- weight benchmark. This is not inconsistent with the paper's objective: the principal contribution is a richer state-dependent account of changing model credibility that retains competitive predictive performance, rather than an assurance of uniformly large MSE reductions. The Monte Carlo experiments reinforce this interpretation by showing that accurate adaptive projection can coexist with uncertainty about the exact support or mechanism decomposition, including when the candidate library is incomplete.

The framework can be extended to dynamic or latent state representations, functional and high-dimensional states, distributional prediction, robust adaptive criteria, and formal inference for predictive images and set-valued geometries. These directions would further develop the central idea that predictive mechanisms can be compared not only globally, but through the low-dimensional state geometry governing where each mechanism is locally informative.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369–1401.

Amini, S. M. and Parmeter, C. F. (2012). Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics*, 27(5):870–876.

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434.

Claeskens, G. and Hjort, N. L. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Czarnecki, W. M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. (2017). Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Das, A. and Kempe, D. (2011). Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1057–1064.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20:134–144.

Durlauf, S. N., Kourtellos, A., and Tan, C. M. (2008). Are any growth theories robust? *The Economic Journal*, 118(527):329–346.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Elenberg, E. R., Khanna, R., Dimakis, A. G., and Negahban, S. (2018). Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568.

Fernandez, C., Ley, E., and Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

Gallup, J. L., Sachs, J. D., and Mellinger, A. D. (1999). Geography and economic development. *International Regional Science Review*, 22(2):179–232.

Giné, E. and Nickl, R. (2021). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2 edition.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214.

Kourtellos, A., Stengos, T., and Tan, C. M. (2010). Do institutions rule? the role of heterogeneity in the institutions versus geography debate. *Economics Bulletin*, 30(3):1710–1719.

Kourtellos, A., Stengos, T., and Tan, C. M. (2013). The effect of public debt on growth in multiple regimes. *Journal of Macroeconomics*, 38:35–43.

Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *American Economic Review*, 82(4):942–963.

Magnus, J. R., Powell, O., and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, 154(2):139–153.

Mankiw, N. G., Romer, D., and Weil, D. N. (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics*, 107(2):407–437.

Mao, T. and Zhou, D.-X. (2023). Rates of approximation by ReLU shallow neural networks. *Journal of Complexity*, 79:101784.

Molchanov, I. (2006). *Theory of Random Sets*. Springer.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational Analysis*. Springer.

Sala-i Martin, X., Doppelhofer, G., and Miller, R. I. (2004). Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *American Economic Review*, 94(4):813–835.

Seo, M. H. and Linton, O. (2007). A smoothed least squares estimator for threshold regression models. *Journal of Econometrics*, 141(2):704–735.

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, 58(3):644–719.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.

Supplement for: Learning Dimension-Reduced State-Adaptive Model Averaging in Regime-Dependent Environments

Stelios Arvanitis*, Thanasis Stengos†

September 14, 2026

1 Introduction

The supplement contains additional material for the theoretical foundations of the proposed methodology, the statistical theory underlying the estimators, the proofs of the main theorems, the descriptive statistics for the empirical application and additional robustness results as well as additional Monte Carlo designs and results, as well as several diagnostics pertaining both to the empirical application and the Monte Carlo experiments regarding set identification and bootstrap uncertainty. All the material in the supplement follow the order of the structure of the paper.

2 Additional Foundations and Computational Details

This section provides additional formal details for the predictive-geometry framework introduced in Sections 2 and 3 of the main paper. We first connect predictive geometry with conventional regime partitions and state-dependent coefficient models. We then discuss observational equivalence and identification in greater detail, before providing the full computational specification of

*Athens University of Economics and Business, Greece. Email: stelios@aueb.gr
†University of Guelph, Canada. Email: tstengos@uoguelph.ca

the neural gating architectures and regularization schemes.

2.1 Regime Dependence and Measurable State Partitions

Let $Z = (Y, X, S)$, where $Y \in \mathbb{R}$, $X \in \mathbb{R}^p$, and $S \in \mathcal{S} \subseteq \mathbb{R}^d$. A conventional regime-dependent representation begins with a measurable partition of the state space.

Definition 2.1 (Measurable state partition)**.** *A finite collection*

$$\{A_1, \ldots, A_R\} \subseteq \mathcal{B}(\mathcal{S})$$

is a measurable partition of $\mathcal{S}$ *if*

$$A_r \cap A_{r'} = \varnothing \qquad \text{for all } r \neq r',$$

and

$$\bigcup_{r=1}^{R} A_r = \mathcal{S}.$$

Definition 2.2 (Regime-dependent conditional law)**.** *A conditional expectation function is regime dependent with respect to the partition* $\{A_1, \ldots, A_R\}$ *if*

$$\mathbb{E}[Y \mid X, S] = \sum_{r=1}^{R} f_r(X) \mathbb{1}\{S \in A_r\}, \tag{1}$$

where $f_r : \mathbb{R}^p \to \mathbb{R}$ *denotes the predictive law associated with regime* r*.*

Threshold models are obtained as a special case (Tong, 1990; Hansen, 2000). Let q_t be a scalar coordinate of S_t and let γ be a threshold. The partition

$$A_1 = \{q_t \leq \gamma\}, \qquad A_2 = \{q_t > \gamma\}$$

produces a two-regime threshold law. A continuous kink specification may be written as

$$Y = X^\top \beta + \delta_L (q - \gamma)_- + \delta_H (q - \gamma)_+ + u, \tag{2}$$

where

$$(x)_- = x\mathbb{1}\{x \le 0\}, \qquad (x)_+ = x\mathbb{1}\{x > 0\}.$$

The partition representation separates the predictive mechanism operating inside a region from the state-space rule determining which region is active. Predictive geometry replaces the latter hard assignment by a simplex-valued map. In this sense, if e_r denotes the r-th vertex of the simplex, the hard regime model (1) corresponds to the allocation map $\Gamma(S) = \sum_{r=1}^{R} e_r \mathbb{1}\{S \in A_r\}$. A continuous predictive geometry replaces this vertex-valued map with $\Gamma : \mathcal{S} \to \Delta^M$, allowing intermediate states to receive nondegenerate convex combinations of the competing mechanisms.

Thus predictive geometry contains conventional threshold and switching structures as boundary cases while allowing smooth transitions between regimes.

2.2 Predictive Geometry and Convex Coefficient Surfaces

Suppose that the candidate predictive mechanisms are linear,

$$f_m(X) = X^\top \beta^{(m)}, \qquad m = 1, \dots, M.$$

For a predictive geometry Γ, the induced adaptive predictor is

$$f_\Gamma(X, S) = \sum_{m=1}^{M} \Gamma_m(S) X^\top \beta^{(m)} = X^\top \widetilde{\beta}_\Gamma(S),$$

where

$$\widetilde{\beta}_\Gamma(S) = \sum_{m=1}^{M} \Gamma_m(S) \beta^{(m)}. \tag{3}$$

The adaptive weighting problem may therefore be viewed equivalently as the estimation of a state-dependent coefficient surface constrained to lie in the convex hull of a finite collection of coefficient vectors.

Proposition 2.3 (Predictive geometries and convex coefficient surfaces)**.** *Suppose*

$$f_m(X) = X^\top \beta^{(m)}, \qquad m = 1, \dots, M.$$

Then every predictive geometry

$$\Gamma : \mathcal{S} \to \Delta^M$$

induces a coefficient surface

$$\widetilde{\beta}_\Gamma : \mathcal{S} \to \mathrm{co}\{\beta^{(1)}, \ldots, \beta^{(M)}\}.$$

Conversely, every coefficient surface $\widetilde{\beta}$ *satisfying*

$$\widetilde{\beta}(s) \in \mathrm{co}\{\beta^{(1)}, \ldots, \beta^{(M)}\}$$

for all $s \in \mathcal{S}$ *admits at least one representation*

$$\widetilde{\beta}(s) = \sum_{m=1}^{M} \Gamma_m(s)\beta^{(m)}$$

with $\Gamma(s) \in \Delta^M$. *If the coefficient vectors* $\beta^{(1)}, \ldots, \beta^{(M)}$ *are affinely independent, the representation is unique.*

Proof. For every $s \in \mathcal{S}$,

$$\Gamma_m(s) \geq 0, \qquad \sum_{m=1}^{M} \Gamma_m(s) = 1.$$

Hence

$$\widetilde{\beta}_\Gamma(s) = \sum_{m=1}^{M} \Gamma_m(s)\beta^{(m)}$$

is a convex combination of the candidate coefficient vectors and therefore belongs to their convex hull.

Conversely, if

$$\widetilde{\beta}(s) \in \mathrm{co}\{\beta^{(1)}, \ldots, \beta^{(M)}\},$$

the definition of the convex hull implies that there exists at least one vector $\Gamma(s) \in \Delta^M$ such that

$$\widetilde{\beta}(s) = \sum_{m=1}^{M} \Gamma_m(s)\beta^{(m)}.$$

If the coefficient vectors are affinely independent, the corresponding barycentric coordinates are unique, establishing uniqueness of $\Gamma(s)$. □

The proposition gives a direct economic interpretation to the adaptive weighting system. Even if every individual candidate model contains fixed coefficients, state-dependent variation in the weights generates a flexible coefficient law

$$S \mapsto \widetilde{\beta}_{\Gamma}(S).$$

Predictive geometry and state-dependent coefficients are therefore dual descriptions of the same adaptive approximation whenever the candidate library is linear.

For nonlinear candidate models, the same interpretation applies at the level of the convex hull of predictive functions. At each state s, the induced predictor belongs to $\mathrm{co}\{f_1, \ldots, f_M\}$.

2.3 Observational Equivalence and Identification

The weight representation associated with a predictive surface need not be unique.

Definition 2.4 (Observational equivalence)**.** *Two predictive geometries* Γ *and* $\widetilde{\Gamma}$ *are observationally equivalent–(Molchanov, 2006; Rockafellar and Wets, 2009), denoted* $\Gamma \sim \widetilde{\Gamma}$*, if*

$$\sum_{m=1}^{M} \Gamma_m(S) f_m(X) = \sum_{m=1}^{M} \widetilde{\Gamma}_m(S) f_m(X) \quad P\text{-almost surely.} \tag{4}$$

The equivalence class associated with Γ is $[\Gamma] = \{\widetilde{\Gamma} : \widetilde{\Gamma} \sim \Gamma\}$. All members of this class generate the same adaptive predictor and, therefore, the same population risk.

Nonuniqueness can arise when candidate models are identical, approximately collinear, or locally redundant. In such cases, the data may identify the adaptive predictor more sharply than the individual model weights.

A sufficient condition for point identification is conditional affine independence.

Assumption 2.5 (Conditional affine independence)**.** *If measurable functions*

$$a_1(S), \ldots, a_M(S)$$

satisfy

$$\sum_{m=1}^{M} a_m(S) f_m(X) = 0 \quad P\text{-almost surely}$$

and

$$\sum_{m=1}^{M} a_m(S) = 0,$$

then

$$a_m(S) = 0 \quad \mathbb{P}\text{-almost surely}$$

for every $m = 1, \ldots, M$.

Under Assumption 2.5, equality of two induced predictive surfaces implies equality of their weight maps almost surely. Without this condition, the statistically natural object is the equivalence class of predictive geometries or, equivalently, the induced predictive surface.

This distinction motivates the set-valued pseudo-true formulation developed in the main paper. The asymptotic theory deals with set convergence and does not require that a unique weight decomposition be identified.

2.4 Equivalent Sparse State Representations

A similar nonuniqueness issue arises for sparse state representations.

Let

$$A, B \subseteq \{1, \ldots, d\}$$

and define

$$S_A = P_A S, \qquad S_B = P_B S.$$

Definition 2.6 (Predictively equivalent state representations)**.** *The state subsets* A *and* B *are predictively equivalent relative to the model library if there exist allocation maps*

$$\Gamma_A : \mathcal{S}_A \to \Delta^M$$

and

$$\Gamma_B : \mathcal{S}_B \to \Delta^M$$

such that

$$\sum_{m=1}^{M} \Gamma_{A,m}(S_A) f_m(X) = \sum_{m=1}^{M} \Gamma_{B,m}(S_B) f_m(X) \quad P\text{-almost surely.} \tag{5}$$

Exact equivalence may arise, for example, when one state variable is a one-to-one transformation of another over the relevant support. Approximate equivalence may arise when state coordinates are highly correlated or when two different projections generate nearly identical partitions of the empirical state distribution.

Consequently, exact support recovery is generally stronger than recovery of the relevant predictive geometry. The latter requires only that the selected state representation generate an allocation map whose induced predictive surface is close to the pseudo-true optimum.

This distinction is especially relevant in macroeconomic applications, where inflation, public debt, institutional quality, openness, and development indicators may contain partially overlapping information about latent economic regimes.

2.5 Neural Softmax Architecture

The computational implementation approximates the infinite-dimensional allocation class by neural softmax sieves. For a state subset A, let $h^{(0)}(S_A) = S_A$. For hidden layers $\ell = 1, \ldots, L-1$, define

$$h^{(\ell)}(S_A) = \sigma\left(W^{(\ell)} h^{(\ell-1)}(S_A) + b^{(\ell)}\right). \tag{6}$$

The output logits are

$$g(S_A;\theta) = W^{(L)} h^{(L-1)}(S_A) + b^{(L)}, \tag{7}$$

where

$$g(S_A;\theta) = \left(g_1(S_A;\theta), \ldots, g_M(S_A;\theta)\right).$$

The final weights are generated by

$$w_m(S_A;\theta) = \frac{\exp\{g_m(S_A;\theta)\}}{\sum_{j=1}^{M} \exp\{g_j(S_A;\theta)\}}. \tag{8}$$

The softmax layer ensures

$$w_m(S_A;\theta) \geq 0, \qquad \sum_{m=1}^{M} w_m(S_A;\theta) = 1.$$

The corresponding prediction is

$$\widehat{f}_{\theta,A}(X,S) = \sum_{m=1}^{M} w_m(S_A;\theta)\widehat{f}_m(X). \tag{9}$$

The baseline empirical and Monte Carlo implementation uses a single hidden layer with sixteen hidden units and ReLU activation, $\sigma(x) = \max\{0, x\}$. Specifically, for $A \in \mathcal{A}_s$, the empirical architecture is

$$g_\theta(s_A) = W_2\sigma(W_1 s_A + b_1) + b_2, \qquad \sigma(u) = \max\{u, 0\}, \tag{10}$$

$$\Gamma_\theta(s_A) = \text{softmax}\{g_\theta(s_A)\}. \tag{11}$$

There are H_n hidden units, $W_1 \in \mathbb{R}^{H_n \times d_A}$, $b_1 \in \mathbb{R}^{H_n}$, $W_2 \in \mathbb{R}^{M \times H_n}$, and $b_2 \in \mathbb{R}^M$. Thus the exact parameter count is

$$K_{A,n} = H_n(d_A+1) + M(H_n+1) = H_n(d_A+M+1) + M, \qquad K_{s,n} = H_n(s+M+1) + M. \tag{12}$$

The state variables are standardized before training. Optimization uses Adam with early stopping (Kingma and Ba, 2015; Prechelt, 1998), and ridge regularization is applied to the network parameters. This is intended to isolate the gains from adaptive weighting rather than from architectural depth.

2.6 Regularized Empirical Criteria

The general regularized empirical criterion is

$$Q_n(w, A) = R_n(w, A) + \mathcal{P}_n(w, A). \tag{13}$$

The theoretical results in the main paper allow a broad class of penalties provided their uniform contribution vanishes asymptotically at the required rate.

2.6.1 Ridge Regularization

For a neural parameter vector θ, ridge regularization takes the form

$$\mathcal{P}_{R,n}(\theta) = \lambda_{R,n}\|\theta\|_2^2. \tag{14}$$

The resulting objective is $Q_{R,n}(\theta, A) = R_n(\theta, A) + \lambda_{R,n}\|\theta\|_2^2$. Ridge regularization–see for example Ch. 5 of Goodfellow et al. (2016), stabilizes optimization via the Hessian conditioning and discourages extreme parameter values.

2.6.2 Entropy Regularization

For $w = w(S_A)$, define the Shannon entropy

$$H(w) = -\sum_{m=1}^{M} w_m(S_A) \log w_m(S_A).$$

The entropy-regularized criterion used in the empirical comparisons is

$$Q_{H,n}(w, A) = R_n(w, A) - \lambda_{H,n}\frac{1}{n}\sum_{t=1}^{n} H(w). \tag{15}$$

Because larger entropy corresponds to more diffuse model weights, the negative sign rewards less concentrated allocations. Entropy regularization therefore discourages extreme local specialization–see for example Haarnoja et al. (2018). Since

$$0 \leq H(w) \leq \log M$$

for every $w \in \Delta^M$, the maximal perturbation generated by entropy regularization satisfies

$$\sup_w |\mathcal{P}_{H,n}(w)| \leq \lambda_{H,n} \log M.$$

2.6.3 Sobolev Curvature Regularization

A functional smoothness penalty may instead be imposed directly on the allocation geometry–see Czarnecki et al. (2017). Suppose $w \in W^{2,2}(\mathcal{S}_A; \mathbb{R}^M)$. Define the second-order Sobolev seminorm

$$|w|^2_{W^{2,2}(\mathcal{S}_A)} = \sum_{m=1}^{M} \sum_{|\alpha|=2} \int_{\mathcal{S}_A} |D^\alpha w_m(s)|^2 \, d\mu_A(s), \tag{16}$$

where D^α denotes a weak derivative and μ_A is a finite reference measure. The curvature-regularized objective is

$$Q_{C,n}(w, A) = R_n(w, A) + \lambda_{C,n}|w|^2_{W^{2,2}(\mathcal{S}_A)}. \tag{17}$$

This penalty discourages unnecessarily irregular variation in predictive credibility. Unlike a parameter-space penalty, it is invariant to different neural parameterizations generating the same allocation map.

The weak-derivative formulation is also appropriate for connecting functional regularization with the Sobolev approximation and compactness arguments used in the theoretical analysis.

2.6.4 Graph and Manifold Regularization

When the empirical state distribution is concentrated near a lower-dimensional structure, smoothness may alternatively be imposed through an empirical similarity graph–Belkin et al. (2006). Let $K_{ij} \geq 0$ measure similarity between states S_i and S_j. A graph penalty takes the form

$$\mathcal{P}_{G,n}(w) = \lambda_{G,n} \sum_{i,j} K_{ij} \|w(S_i) - w(S_j)\|_2^2. \tag{18}$$

Large values of K_{ij} encourage nearby or economically similar states to receive similar model allocations. The corresponding penalty can be interpreted as a discrete analogue of a Dirichlet energy on the state manifold.

2.7 Sparse State Learning and Forward Selection

For every state subset A, define the regularized profile criterion

$$\widehat{L}_n(A) = \inf_{w \in \mathcal{W}_{A,n}} Q_n(w, A). \tag{19}$$

Exact optimization over $\mathcal{A}_s = \{A \subseteq \{1, \ldots, d\} : |A| \leq s\}$ requires evaluating a combinatorial number of subsets. The implementation therefore uses forward selection (Das and Kempe, 2011; Elenberg et al., 2018; Efron et al., 2004). Starting from $A_0 = \varnothing$, the algorithm evaluates $\widehat{L}_n(A_k \cup \{j\})$ for every $j \notin A_k$. The next coordinate is

$$j_{k+1} \in \arg\min_{j \notin A_k} \widehat{L}_n(A_k \cup \{j\}), \tag{20}$$

and the state set is updated according to

$$A_{k+1} = A_k \cup \{j_{k+1}\}.$$

The algorithm terminates either when the maximum state dimension is reached or when

$$\widehat{L}_n(A_k) - \widehat{L}_n(A_{k+1}) \leq \tau_n^{FS}, \tag{21}$$

where τ_n^{FS} is the forward-selection stopping tolerance. The resulting subset is denoted $\widehat{A}_s^{FS}$. A final neural gate is then estimated conditional on this selected state representation.

2.8 Estimation Sequence

For completeness, the computational implementation used in the paper can be summarized as follows.

Step 1: Candidate-model estimation. Estimate each candidate predictive model separately and construct the prediction matrix

$$\widehat{\mathbf{f}} = \left[\widehat{f}_1(X), \ldots, \widehat{f}_M(X)\right].$$

Step 2: Benchmark aggregation. Compute equal-weight averaging, fixed least-squares averaging, and the best-single-model benchmark.

Step 3: Full-state adaptive estimation. Estimate the linear and neural softmax gates using the complete candidate state vector.

Step 4: Sparse state search. Apply forward selection. At every iteration, fit candidate neural gates corresponding to each one-variable enlargement of the current state subset and compare their validation criteria.

Step 5: Final sparse-gate estimation. Conditional on the selected state subset, estimate the final neural softmax gate and construct

$$\widehat{Y} = \sum_{m=1}^{M} \widehat{w}_m(S_{\widehat{A}_s^{FS}})\widehat{f}_m(X).$$

Step 6: Evaluation. Evaluate prediction error, approximation error when the true conditional mean is available, model-weight behavior, and state-selection stability.

The same general architecture is used in the Monte Carlo experiments and the empirical application, although the exact benchmark set differs where oracle information is available in simulation.

3 Statistical Theory: Rroofs and Additional Results

This section provides the technical details underlying the statistical theory developed in Section 4 of the main paper. The analysis proceeds in several steps. We first formalize the functional structure of the ideal weight classes and the role of regularization. We then establish the approximation properties of neural softmax sieves and connect the regularized empirical estimator to the ideal sparse adaptive $L^2(\mathbb{P})$-projection. Finally, we develop the set-valued convergence theory and provide the detailed population and empirical guarantees for forward selection.

Throughout, let

$$\mathcal{A}_s = \{A \subseteq \{1, \ldots, d\} : |A| \leq s\}$$

denote the collection of admissible sparse state representations. For $A \in \mathcal{A}_s$, write $S_A = P_A S$.

For a simplex-valued weight function w, define

$$f_{w,A}(X,S) = \sum_{m=1}^{M} w_m(S_A)\widehat{f}_m(X).$$

The population and empirical risks are

$$R(w,A) = \mathbb{P}\left[(Y - f_{w,A}(X,S))^2\right] \tag{22}$$

and

$$R_n(w,A) = \mathbb{P}_n\left[(Y - f_{w,A}(X,S))\right]^2. \tag{23}$$

3.1 Finite-Net Concentration and Its Primitive Verification

For $A \in \mathcal{A}_s$ and $w \in \mathcal{W}_{A,n}$, write

$$\ell_{w,A}(Z) = (Y - f_{w,A}(X,S))^2, \qquad \mathcal{L}_{A,n} = \{\ell_{w,A} : w \in \mathcal{W}_{A,n}\}.$$

For any deterministic finite ε-net $\mathcal{W}^{\varepsilon}_{A,n}$ of $\mathcal{W}_{A,n}$ under d_∞, define its induced sparse loss net by

$$\mathcal{L}^{\varepsilon}_{s,n} = \bigcup_{A\in\mathcal{A}_s} \{\ell_{w,A} : w \in \mathcal{W}^{\varepsilon}_{A,n}\}. \tag{24}$$

Assumption 3.1 (Finite-net Bernstein concentration; **Assumption 4.1 in the Main Paper).** *There are constants $C_0, C_1, c > 0$ and a deterministic dependence scale $b_n \geq 1$ such that, for every $0 < \varepsilon < 1$, every induced finite loss net (24), and every $1 \leq x \leq n$,*

$$\mathbb{P}\left[\max_{\ell\in\mathcal{L}^{\varepsilon}_{s,n}} |(\mathbb{P}_n - P)\ell| > C_0\left\{\sqrt{\frac{\log|\mathcal{L}^{\varepsilon}_{s,n}| + x}{n}} + \frac{b_n\{\log|\mathcal{L}^{\varepsilon}_{s,n}| + x\}}{n}\right\}\right] \leq C_1 e^{-cx}. \tag{25}$$

Assumption 3.1 concerns only deterministic finite loss nets. It does not assume the desired supremum bound over the complete neural loss class. Theorem 3.5 derives that bound by combining (25) with the entropy of the implemented shallow architecture and a random Lipschitz envelope.

Assumption 3.2 (Conditional sub-Gaussian primitives; **Assumption 4.2 in the Main**

Paper). *For every $A \in \mathcal{A}_s$, the rescaled state domain satisfies $\mathcal{S}_A \subset [-R_S, R_S]^{d_A}$, where $d_A = |A|$ and $R_S < \infty$. There are constants $C_Y, C_f, \sigma_Y, \sigma_f < \infty$ such that*

$$|\mathbb{E}(Y \mid X, S)| \le C_Y, \qquad \|\mathbb{E}\{\widehat{\boldsymbol{f}}(X) \mid S\}\|_2 \le C_f \quad a.s.,$$

and, for every $t \in \mathbb{R}$,

$$\mathbb{E}[\exp\left(t(Y - \mathbb{E}[Y \mid X, S])\right) \mid X, S] \le \exp(\sigma_Y^2 t^2/2), \tag{26}$$

$$\sup_{\|a\|_2 \le 1} \mathbb{E}\Big[\exp\Big(ta'[\widehat{\boldsymbol{f}}(X) - \mathbb{E}\left(\widehat{\boldsymbol{f}}(X) \mid S\right)\}]\Big)\Big|S\Big] \le \exp(\sigma_f^2 t^2/2). \tag{27}$$

The implemented one-hidden-layer network is given by (43)–(44), and each entry of its input weights, hidden biases, output weights, and output biases is bounded in absolute value by $B_n \ge 1$, uniformly over $A \in \mathcal{A}_s$ and the neural sieve.

The conditional formulation is important because the gate is a function of S. Purely marginal sub-Gaussianity of each candidate prediction does not by itself control state-selected linear combinations. Equivalently, one may impose directly

$$\sup_{A,w} \|w(S_A)'\widehat{\boldsymbol{f}}(X)\|_{\psi_2} < \infty, \qquad \|\{w(S_A) - v(S_A)\}'\widehat{\boldsymbol{f}}(X)\|_{\psi_2} \le C d_\infty(w, v), \tag{28}$$

together with the corresponding sub-Gaussian condition for Y.

Lemma 3.3 (Sub-exponential losses and loss increments)**.** *Under Assumption 3.2,*

$$\sup_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_{A,n}} \|\ell_{w,A}\|_{\psi_1} < \infty \tag{29}$$

and, uniformly over A, w, v,

$$\|\ell_{w,A} - \ell_{v,A}\|_{\psi_1} \le C d_\infty(w, v). \tag{30}$$

Proof. The conditional sub-Gaussian bound for the candidate-prediction vector, its bounded conditional mean, and the simplex restriction imply

$$\sup_{A,w} \|f_{w,A}(X, S)\|_{\psi_2} < \infty.$$

The same conclusion holds for Y, so $Y - f_{w,A}$ is uniformly sub-Gaussian. Its square is therefore uniformly sub-exponential, proving (29). Moreover,

$$\ell_{w,A} - \ell_{v,A} = (f_{v,A} - f_{w,A})(2Y - f_{w,A} - f_{v,A}).$$

The first factor satisfies

$$\|f_{v,A} - f_{w,A}\|_{\psi_2} \le C d_\infty(w, v),$$

while the second is uniformly sub-Gaussian. The Orlicz product inequality (Rao and Ren, 1991) $\|UV\|_{\psi_1} \le C\|U\|_{\psi_2}\|V\|_{\psi_2}$ proves (30). □

Proposition 3.4 (Primitive verification of finite-net concentration)**.** *Suppose Assumption 3.2 holds. Then Assumption 3.1 follows under any one of the following sampling conditions, provided the stated constants are uniform over deterministic induced loss nets:*

(i) $Z_1, \dots, Z_n$ *are independent with common law* $\mathbb{P}$*;*

(ii) $(Z_i)_{i\in\mathbb{Z}}$ *is strictly stationary and strongly mixing with* $\alpha(k) \le C_\alpha \exp(-c_\alpha k^\eta)$*, and the uniform semi-exponential tail and exponent conditions of Merlevède et al. (2011a) hold for the centered induced losses;*

(iii) (Z_i) *is exchangeable and, conditionally on its directing sigma-field* $\mathcal{U}$*, the observations are independent, their conditional* ψ_1 *loss norms are uniformly bounded, and*

$$\mathbb{P}[\ell(Z_i) \mid \mathcal{U}] = \mathbb{P}[\ell] \quad \text{a.s. for every induced-net loss } \ell; \tag{31}$$

(iv) for every induced-net loss ℓ*,* $\xi_{i,\ell} = \ell(Z_i) - \mathbb{P}[\ell]$ *is a martingale difference relative to* $\mathcal{F}_i = \sigma(Z_1, \dots, Z_i)$ *and, for every integer* $k \ge 2$*,*

$$\sum_{i=1}^{n} \mathbb{E}(|\xi_{i,\ell}|^k \mid \mathcal{F}_{i-1}) \le \frac{k!}{2} n v^2 c_\ell^{k-2}, \qquad c_\ell \le C b_n, \tag{32}$$

for constants uniform over the net.

In (i) and (iii), $b_n = O(1)$*. Under (ii),* b_n *absorbs the logarithmic dependence factor in the applicable mixing Bernstein inequality.*

Proof. Fix a deterministic induced loss net of cardinality N. By Lemma 3.3, every centered net loss has a uniformly bounded ψ_1 norm. Under (i), Bernstein's inequality for independent sub-exponential variables gives

$$\mathbb{P}\Big(|(\mathbb{P}_n - \mathbb{P})[\ell]| > C\{\sqrt{y/n} + y/n\}\Big) \le 2e^{-cy}.$$

Under (ii), the Bernstein inequality of Merlevède et al. (2011b) gives the same display with the linear term replaced by $b_n y/n$. Under (iii), apply the independent inequality conditionally on $\mathcal{U}$; condition (31) supplies the centering by $\mathbb{P}[\ell]$, and integration gives the unconditional bound. Under (iv), the conditional Bernstein inequality of de la Peña (1999) gives the same form; for bounded increments, Freedman (1975) is an alternative when predictable quadratic variation is controlled.

In every case, take $y = C_2(\log N + x)$ with C_2 sufficiently large and apply a union bound over the N losses. After renaming constants, the resulting inequality is (25).

For (iii), the centering qualification is essential. A general exchangeable sequence is conditionally i.i.d. given its directing sigma-field, and $\mathbb{P}_n[\ell]$ converges to $\mathbb{P}(\ell \mid \mathcal{U})$, which need not equal $\mathbb{P}[\ell]$. □

Theorem 3.5 (Uniform sparse-state concentration)**.** *Suppose Assumption 3.1 and Lemma 3.10 hold. Let*

$$V_{s,n} = CK_{s,n}\log(C\mathfrak{L}_{s,n}n) + s\log(ed/s), \tag{33}$$

$$r_n = \sqrt{\frac{V_{s,n}}{n}} + \frac{b_n V_{s,n}}{n}, \tag{34}$$

where $K_{s,n} = \max_{A\in\mathcal{A}_s} K_{A,n}$ *and* $\mathfrak{L}_{s,n} = \max_{A\in\mathcal{A}_s} \mathfrak{L}_{A,n}$. *Assume* $V_{s,n} = o(n)$, $b_n V_{s,n}/n = o(1)$, *and*

$$\mathbb{P}[L] < \infty, \qquad L = \|\widehat{\boldsymbol{f}}(X)\|_1\{2|Y| + 2\|\widehat{\boldsymbol{f}}(X)\|_\infty\}. \tag{35}$$

Then

$$\zeta_n = \sup_{A\in\mathcal{A}_s}\sup_{w\in\mathcal{W}_{A,n}} |(\mathbb{P}_n - \mathbb{P})[\ell_{w,A}]| = O_p(r_n). \tag{36}$$

Proof. Set $\varepsilon_n = n^{-1}$ and select a deterministic ε_n-net for every gate class. By Lemma 3.10 and

$\log|\mathcal{A}_s| \le s\log(ed/s)$,

$$\log|\mathcal{L}_{s,n}^{\varepsilon_n}| \le V_{s,n}.$$

Assumption 3.1, with fixed sufficiently large x, yields

$$\max_{\ell\in\mathcal{L}_{s,n}^{\varepsilon_n}} |(\mathbb{P}_n - P)[\ell]| = O_p(r_n).$$

For arbitrary w, choose a net element w^{ε_n} with $d_\infty(w, w^{\varepsilon_n}) \le \varepsilon_n$. The squared-loss identity gives the pointwise inequality

$$|\ell_{w,A} - \ell_{w^{\varepsilon_n},A}| \le \varepsilon_n L.$$

Consequently,

$$\zeta_n \le \max_{\ell\in\mathcal{L}_{s,n}^{\varepsilon_n}} |(\mathbb{P}_n - P)[\ell]| + \varepsilon_n(\mathbb{P}_n + P)[L].$$

Because $L \ge 0$, the common marginal law and $\mathbb{P}[L] < \infty$ imply $\mathbb{P}(\mathbb{P}_n[L]) = \mathbb{P}[L]$; hence $\mathbb{P}_n[L] = O_p(1)$ by Markov's inequality, without an additional dependence assumption. The final term is $O_p(n^{-1}) = o_p(r_n)$, proving (36). □

Remark 3.6 (Heavier-tailed alternatives)**.** *Theorem 3.5 requires the finite-net inequality, the gate entropy bound, and the integrable oscillation envelope; it does not logically require sub-Gaussian primitives. Under heavier tails, downstream oracle arguments remain valid if these requirements are established by truncation, winsorization, or a robust empirical criterion. Truncation adds the uniform tail bias*

$$\sup_{A,w} \mathbb{P}[\ell_{w,A}\mathbf{1}\{\ell_{w,A} > \tau_n\}]$$

to the statistical remainder. A robust loss or robust mean requires the finite-net assumption and criterion to be restated for that estimator. Finite variance alone does not give an exponential bound uniformly over a growing net.

3.2 Regularized Empirical Learning

Generally, the regularization can obtain the following form

$$\mathcal{P}_n = \mathcal{P}_{R,n} + \mathcal{P}_{H,n} + \mathcal{P}_{C,n} + \mathcal{P}_{G,n},$$

where $\mathcal{P}_{G,n}$ may represent graph-based or manifold-based regularization over the observed state geometry. Define the uniform magnitude of the regularization perturbation by

$$\pi_n = \sup_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_{A,n}} |\mathcal{P}_n(w, A)| . \tag{37}$$

The main theory then assumes $\pi_n \to 0$. This condition guarantees that regularization does not alter the limiting unregularized population target. The stronger condition $\pi_n = O(r_n)$ ensures that regularization does not change the order of the statistical convergence rate, while the $\pi_n = o(r_n)$ makes regularization asymptotically negligible relative to the leading empirical-process fluctuation.

For ridge regularization, a sufficient condition is $\lambda_{R,n} \sup_{A \in \mathcal{A}_s} \sup_{\theta \in \Theta_{A,n}} \|\theta\|_2^2 \le \pi_n$. For entropy regularization, since $0 \le H(w) \le \log M$, it is sufficient that $\lambda_{H,n} \log M \le \pi_n$. For Sobolev curvature regularization, it is sufficient that $\lambda_{C,n} \sup_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_{A,n}} \mathcal{J}_C(w; A) \le \pi_n$.

3.3 Ideal Sparse Weight Classes

For each $A \in \mathcal{A}_s$, let $\mathcal{S}_A = P_A \mathcal{S} \subseteq \mathbb{R}^{|A|}$. Let $q \ge 1$ and $\nu > 0$ satisfy $q\nu > |A|$ for every $A \in \mathcal{A}_s$. For some constant $C_W < \infty$, define

$$\mathcal{W}_A^0 = \left\{ w = (w_1, \dots, w_M) : \begin{array}{l} w_m \in W^{\nu,q}(\mathcal{S}_A), \\ \|w_m\|_{W^{\nu,q}} \le C_W, \\ w(s_A) \in \Delta^M \quad \text{for all } s_A \in \mathcal{S}_A \end{array} \right\}. \tag{38}$$

When curvature regularization is based on second-order derivatives, we take $\nu \ge 2$. The condition $q\nu > |A|$ implies, by Sobolev embedding, that every element of $\mathcal{W}_A^0$ admits a continuous representative. Moreover, under the standard compact embedding conditions on $\mathcal{S}_A$, bounded subsets of $W^{\nu,q}(\mathcal{S}_A)$ are relatively compact in $C(\mathcal{S}_A)$.

After taking the closure of $\mathcal{W}_A^0$ in the uniform topology if necessary, the ideal class may therefore be regarded as a compact subset of $C(\mathcal{S}_A, \Delta^M)$. This functional structure serves two purposes. First, it guarantees existence of population sparse oracles. Second, it allows pointwise neural universal approximation to be strengthened into uniform approximation over the ideal class.

3.4 Existence of Pseudo-True Sparse Oracles

Define the ideal sparse approximation risk by

$$R_s^\star = \min_{A \in \mathcal{A}_s} \min_{w \in \mathcal{W}_A^0} R(w, A). \tag{39}$$

The corresponding pseudo-true sparse oracle set is

$$\mathcal{E}_s^\star = \left\{ (A, w) : A \in \mathcal{A}_s,\ w \in \mathcal{W}_A^0,\ R(w, A) = R_s^\star \right\}. \tag{40}$$

To compare gates defined on different state subsets, identify every $w_A : \mathcal{S}_A \to \Delta^M$ with its canonical extension

$$\widetilde{w}_A(s) = w_A(P_A s), \qquad s \in \mathcal{S}.$$

For the uniform metric on the state space

$$d_\infty(w_A, w_B) = \sup_{s \in \mathcal{S}} \|\widetilde{w}_A(s) - \widetilde{w}_B(s)\|_\infty \tag{41}$$

define the metric

$$d((A, w), (B, v)) = \mathbf{1}\{A \neq B\} + d_\infty(w, v). \tag{42}$$

Lemma 3.7 (Existence and compactness of ideal sparse oracles)**.** *Suppose that each $\mathcal{W}_A^0$ is nonempty and compact under d_∞ and that $\mathbb{E}[Y^2] < \infty$, $\mathbb{E}[\|\widehat{\boldsymbol{f}}(X)\|_1^2] < \infty$. Then $R(\cdot, A)$ is continuous on $\mathcal{W}_A^0$, and $\mathcal{E}_s^\star$ is nonempty and compact under (42).*

Proof. Fix A and let $d_\infty(w_j, w) \to 0$. Writing $f_j = f_{w_j, A}$ and $f = f_{w, A}$,

$$|f_j - f| \le d_\infty(w_j, w) \|\widehat{\boldsymbol{f}}(X)\|_1.$$

Moreover, because all gates are simplex valued, $|f_j| \vee |f| \le \|\widehat{\boldsymbol{f}}(X)\|_\infty$. The squared-loss identity and Cauchy–Schwarz therefore give

$$\begin{aligned} |R(w_j, A) - R(w, A)| &\le \|f_j - f\|_{L_2(P)} \|2Y - f_j - f\|_{L_2(P)} \\ &\le C, d_\infty(w_j, w) \to 0, \end{aligned}$$

where $C < \infty$ follows from the stated second moments. Thus the risk is continuous without bounded outcomes or bounded candidate predictions. It attains its minimum on each compact $\mathcal{W}_A^0$. Since $\mathcal{A}_s$ is finite, the global minimum is attained. Its argmin is a closed subset of the finite union $\bigcup_{A\in\mathcal{A}_s}\{A\}\times\mathcal{W}_A^0$, hence is compact and nonempty. □

3.5 Primitive Shallow-Sieve Conditions, Entropy, and Approximation

For $A \in \mathcal{A}_s$, put $d_A = |A|$ and let u_A denote the rescaled state vector in $[-R_S, R_S]^{d_A}$. The implemented one-hidden-layer ReLU logits and softmax gate are

$$g_{A,m}(u_A;\theta) = c_m + \sum_{h=1}^{H_n} a_{mh}\sigma(b_h' u_A + d_h), \qquad m = 1,\dots,M, \tag{43}$$

$$w_m(u_A;\theta) = \frac{\exp(g_{A,m}(u_A;\theta))}{\sum_{\ell=1}^{M}\exp(g_{A,\ell}(u_A;\theta))}, \tag{44}$$

where $\sigma(x) = \max\{x, 0\}$. Counting the shared hidden-layer parameters and the M output coordinates gives

$$K_{A,n} = H_n(d_A + 1) + M(H_n + 1). \tag{45}$$

Assumption 3.8 (Shallow-sieve regularity)**.** *For every $A \in \mathcal{A}_s$, the rescaled domain satisfies $\mathcal{S}_A \subset [-R_S, R_S]^{d_A}$, and every coefficient in (43) is bounded in absolute value by $B_n \geq 1$. For every $w \in \mathcal{W}_A^0$, there is an interior surrogate $w^{(n)}$ such that*

$$d_\infty(w^{(n)}, w) \leq \beta_n, \qquad \inf_{s_A,m} w_m^{(n)}(s_A) > 0, \tag{46}$$

whose logits $g_{A,m}^{(n)} = \log w_m^{(n)}$ belong to $W_\infty^\nu(\mathcal{S}_A)$ with norm at most $R_{g,n}$, uniformly over $A \in \mathcal{A}_s$ and $w \in \mathcal{W}_A^0$. Here $\beta_n \to 0$, while $R_{g,n}$ and B_n may vary with n. The parameter bound B_n is large enough to contain the constructive shallow approximants used below.

The two clauses of Assumption 3.8 have different roles. Compact inputs and bounded network coefficients give the finite-dimensional covering entropy used in r_n. Smoothness of the interior-surrogate logits is used only to bound the neural approximation error $a_n(H_n)$. It is not a Sobolev-entropy assumption on the full loss class. In particular, the fixed candidate predictions require tail and multiplier control, but no Sobolev covering numbers.

Remark 3.9 (Connection with the ideal Sobolev classes)**.** *The ideal classes in* (38) *use* $W^{\nu_0,q}(\mathcal{S}_A)$ *with* $q\nu_0 > d_A$*. Sobolev embedding gives Hölder regularity of every order*

$$0 < \nu < \nu_0 - d_A/q, \tag{47}$$

with the usual qualification at integer boundaries. If an interior surrogate is bounded below by $\underline{\eta}_n > 0$*, composition with* $u \mapsto \log u$ *places its logits in the corresponding Hölder class. The resulting logit radius* $R_{g,n}$ *depends on the original Sobolev radius and on* $\underline{\eta}_n^{-1}$*. This embedding supplies approximation smoothness; a Sobolev entropy term is not added to the statistical complexity of the loss class.*

Lemma 3.10 (Entropy of the implemented shallow architecture)**.** *Under Assumption 3.8, for* $0 < \varepsilon < 1$*,*

$$\log N(\varepsilon, \mathcal{W}_{A,n}, d_\infty) \le K_{A,n} \log\left(\frac{C\mathfrak{L}_{A,n}}{\varepsilon}\right), \tag{48}$$

$$\mathfrak{L}_{A,n} = B_n\{1 + H_n B_n (d_A + 1)(R_S + 1)\}. \tag{49}$$

If Assumption 3.2 also holds, then

$$\log N(\varepsilon, \mathcal{L}_{A,n}, \|\cdot\|_{\psi_1}) \le K_{A,n} \log\left(\frac{CC_L\mathfrak{L}_{A,n}}{\varepsilon}\right), \tag{50}$$

where C_L *is the loss-increment constant in Lemma 3.3. Therefore, with* $K_{s,n} = \max_A K_{A,n}$ *and* $\mathfrak{L}_{s,n} = \max_A \mathfrak{L}_{A,n}$*,*

$$\log N(\varepsilon, \mathcal{L}_{s,n}, \|\cdot\|_{\psi_1}) \le K_{s,n} \log\left(\frac{CC_L\mathfrak{L}_{s,n}}{\varepsilon}\right) + s\log(ed/s). \tag{51}$$

Proof. Let θ and θ' be two parameter vectors and put $\delta = \|\theta - \theta'\|_\infty$. The hidden preactivations differ uniformly by at most $(d_A + 1)(R_S + 1)\delta$, while the hidden activations are bounded by $CB_n(d_A + 1)(R_S + 1)$. Decomposing the logit difference into changes in the output weights, hidden activations, and output biases gives

$$\|g_\theta - g_{\theta'}\|_\infty \le C\{1 + H_n B_n (d_A + 1)(R_S + 1)\}\delta.$$

Softmax has a uniformly bounded Jacobian, so the same order controls the gate difference. Discretize the parameter cube $[-B_n, B_n]^{K_{A,n}}$ with mesh proportional to this Lipschitz scale's inverse times ε. The resulting grid has cardinality at most $(C\mathfrak{L}_{A,n}/\varepsilon)^{K_{A,n}}$, proving (48).

By Lemma 3.3,

$$\|\ell_{w,A} - \ell_{v,A}\|_{\psi_1} \le C_L d_\infty(w, v).$$

Thus an (ε/C_L)-cover of the gate class maps into an ε-cover of the induced loss class. This proves (50); neither the outcome nor the fixed candidate predictions are separately covered. Taking the union over A and using $\log|\mathcal{A}_s| \le s\log(ed/s)$ proves (51). □

For the exact fixed-depth architecture above, the applicable uniform approximation result is the shallow ReLU theorem of Mao and Zhou (2023). For $d_A \ge 1$ and $0 < \nu < d_A/2 + 2$, define

$$\alpha_{\nu,d_A} = \frac{\nu}{d_A}\frac{d_A+2}{d_A+4}, \qquad c_{d_A} = d_A + \frac{1}{2}. \tag{52}$$

The empty-support model has a constant gate represented by the output biases; no expression containing division by d_A is used when $A = \varnothing$.

Theorem 3.11 (Explicit shallow ReLU–softmax approximation)**.** *Suppose Assumption 3.8 holds and $0 < \nu < d_A/2 + 2$ for every nonempty admissible support. For every $A \in \mathcal{A}_s$ and $w \in \mathcal{W}_A^0$, there exists $v_{A,H_n} \in \mathcal{W}_{A,n}$ such that*

$$d_\infty(v_{A,H_n}, w) \le \beta_n + CR_{g,n}(\log H_n)^{c_{d_A}} H_n^{-\alpha_{\nu,d_A}}. \tag{53}$$

If

$$\mathbb{E}[Y^2 + \|\widehat{\boldsymbol{f}}(X)\|_1^2] < \infty, \tag{54}$$

then the right-hand side of (53) also bounds $|R(v_{A,H_n}, A) - R(w, A)|$ up to a constant independent of A. With the main paper's definition

$$a_n = \max_{A\in\mathcal{A}_s} \sup_{w\in\mathcal{W}_A^0} \inf_{v\in\mathcal{W}_{A,n}} d_\infty(v, w), \tag{55}$$

one may therefore use the conservative uniform bound

$$a_n(H_n) \le \beta_n + CR_{g,n}(\log H_n)^{s+1/2} H_n^{-\alpha_{\nu,s}}, \qquad \alpha_{\nu,s} = \frac{\nu}{s}\frac{s+2}{s+4}. \tag{56}$$

Proof. Fix A, w, and its interior surrogate $w^{(n)}$. Allocate $\lfloor H_n/M \rfloor$ hidden units to each logit and combine the resulting units in the multi-output network (43). Since M is fixed, this allocation changes only the approximation constant. The compatibility condition on B_n ensures that the constructed network belongs to the stated sieve. The shallow ReLU theorem (Mao and Zhou, 2023, Th. 1) gives logits $\widetilde{g}_m$ with

$$\max_{m \le M} \|\widetilde{g}_m - g_{A,m}^{(n)}\|_\infty \le CR_{g,n}(\log H_n)^{c_{d_A}} H_n^{-\alpha_{\nu,d_A}}.$$

Softmax is globally Lipschitz and $\mathrm{softmax}(g_A^{(n)}) = w^{(n)}$. The triangle inequality with (46) proves (53).

For the risk statement, put $f = f_{w,A}$ and $\widetilde{f} = f_{v_{A,H_n},A}$. By the squared-loss identity and Cauchy–Schwarz,

$$\begin{aligned} |R(v_{A,H_n}, A) - R(w, A)| &\le \|f - \widetilde{f}\|_{L_2(P)} \|2Y - f - \widetilde{f}\|_{L_2(P)} \\ &\le C d_\infty(v_{A,H_n}, w), \end{aligned}$$

because $|f - \widetilde{f}| \le d_\infty(v_{A,H_n}, w)\|\widehat{\boldsymbol{f}}(X)\|_1$ and (54) supplies the required moments. Taking the supremum over the ideal class and the maximum over supports, with the conservative dimension $d_A = s$, proves (56). □

The preceding results make the width tradeoff explicit. For fixed s and M, $K_{s,n} \asymp H_n$, so, suppressing logarithms and the linear Bernstein remainder, the width-dependent statistical and approximation terms are

$$\sqrt{H_n/n} \qquad \text{and} \qquad R_{g,n} H_n^{-\alpha_{\nu,s}}, \tag{57}$$

respectively.

Corollary 3.12 (Balanced shallow width and rate)**.** *Suppose s and M are fixed, $R_{g,n} = O(1)$, $\beta_n = o\{n^{-\alpha_{\nu,s}/(2\alpha_{\nu,s}+1)}\}$, $\log \mathfrak{L}_{s,n} = O(\log n)$, and regularization and optimization errors are*

negligible at the displayed rate. Suppose also that the chosen B_n contains the constructive approximants and that $b_n V_{s,n}/n$ is no larger than the square-root term. Ignoring logarithmic factors, balancing (57) gives

$$H_n \asymp n^{1/(2\alpha_{\nu,s}+1)}, \qquad r_n + a_n = \widetilde{O}_p\big(n^{-\alpha_{\nu,s}/(2\alpha_{\nu,s}+1)}\big). \tag{58}$$

If approximation contributes quadratically to excess risk rather than through the Lipschitz risk bound used here, both the balancing equation and the resulting exponent must be changed accordingly.

Proof. Up to logarithmic factors, equate $n^{-1/2}H_n^{1/2}$ and $H_n^{-\alpha_{\nu,s}}$. This gives $H_n^{\alpha_{\nu,s}+1/2} = n^{1/2}$, hence $H_n = n^{1/(2\alpha_{\nu,s}+1)}$. Substitution into either leading term gives the rate in (58). The remaining assumptions ensure that the omitted surrogate, dependence, parameter-feasibility, penalty, and computation terms do not dominate this balance. □

Remark 3.13 (Härdle-style bias–variance interpretation)**.** *The width H_n plays a similar role to that of a nonparametric smoothing parameter. Increasing it reduces neural approximation bias through $a_n(H_n)$ but increases estimation variance through $K_{s,n}$ and hence $r_n(H_n)$. The sparse term $s\log(ed/s)$ is the additional search cost of learning which state coordinates index the adaptive predictor. Thus the implementable out-of-sample error combines a classical sieve bias–variance tradeoff with sparse-search, regularization, and computational errors.*

3.6 The Population Neural-Sieve Oracle

Recall the population neural-sieve oracle

$$(w_{s,n}^\star, A_{s,n}^\star) \in \operatorname*{arg\,min}_{\substack{A\in\mathcal{A}_s \\ w\in\mathcal{W}_{A,n}}} R(w,A). \tag{59}$$

Existence follows from continuity of the population risk and compactness of the bounded neural parameter spaces.

The computed regularized sparse estimator need not minimize the neural criterion exactly. We assume that

$$Q_n(\widehat{w}_s, \widehat{A}_s) \le \inf_{\substack{A\in\mathcal{A}_s \\ w\in\mathcal{W}_{A,n}}} Q_n(w,A) + \varepsilon_n^{\mathrm{NN}}, \tag{60}$$

where $\varepsilon_n^{\mathrm{NN}} \geq 0$ is the achieved criterion-value tolerance. Exact minimization corresponds to $\varepsilon_n^{\mathrm{NN}} = 0$. Define

$$\zeta_n = \sup_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_{A,n}} |R_n(w, A) - R(w, A)|. \tag{61}$$

Proof of ***Theorem 4.7 in the Main Paper****.* For each $A \in \boldsymbol{\mathcal{A}}_s$, choose a deterministic ε_n-net $\mathcal{W}_{A,n}^{\varepsilon_n}$ under d_∞, where $\varepsilon_n = n^{-1}$, and let

$$\mathcal{L}_{s,n}^{\varepsilon_n} = \bigcup_{A \in \mathcal{A}_s} \left\{ \ell_{w,A} : w \in \mathcal{W}_{A,n}^{\varepsilon_n} \right\}.$$

By the neural-gate entropy bound,

$$\log N(\varepsilon_n, \mathcal{W}_{A,n}, d_\infty) \leq K_{A,n} \log \left(C \mathfrak{L}_{A,n} n \right).$$

Since

$$|\mathcal{A}_s| \leq \left(\frac{ed}{s} \right)^s,$$

we obtain

$$\log |\mathcal{L}_{s,n}^{\varepsilon_n}| \leq C K_{s,n} \log \left(C \mathfrak{L}_{s,n} n \right) + s \log(ed/s) = V_{s,n}.$$

Applying Assumption 4.1 of the main paper to this induced finite loss net gives

$$\max_{\ell \in \mathcal{L}_{s,n}^{\varepsilon_n}} |(\mathbb{P}_n - \mathbb{P})\ell| = O_p \left(\sqrt{\frac{V_{s,n}}{n}} + \frac{b_n V_{s,n}}{n} \right) = O_p(r_n). \tag{62}$$

It remains to pass from the finite net to the complete neural loss class. For arbitrary $w \in \mathcal{W}_{A,n}$, choose $w^{\varepsilon_n} \in \mathcal{W}_{A,n}^{\varepsilon_n}$ satisfying

$$d_\infty(w, w^{\varepsilon_n}) \leq \varepsilon_n.$$

The squared-loss identity gives

$$\begin{aligned} |\ell_{w,A} - \ell_{w^{\varepsilon_n},A}| &= |f_{w,A} - f_{w^{\varepsilon_n},A}| \, |2Y - f_{w,A} - f_{w^{\varepsilon_n},A}| \\ &\leq \varepsilon_n L(Z), \end{aligned}$$

where

$$L(Z) = \|\widehat{\boldsymbol{f}}(X)\|_1 \left\{2|Y| + 2\|\widehat{\boldsymbol{f}}(X)\|_\infty\right\}. \tag{63}$$

The conditional sub-Gaussian conditions imply $\mathbb{P}[L] < \infty$. Since all observations have common marginal law $\mathbb{P}$,

$$\mathbb{E}[\mathbb{P}_n[L]] = \mathbb{P}[L] < \infty,$$

and Markov's inequality gives

$$\mathbb{P}_n[L] = O_p(1).$$

Therefore,

$$\begin{aligned}\zeta_n &\le \max_{\ell \in \mathcal{L}_{s,n}^{\varepsilon_n}} |(\mathbb{P}_n - \mathbb{P})[\ell]| + \varepsilon_n(\mathbb{P}_n + \mathbb{P})[L] \\ &= O_p(r_n) + O_p(n^{-1}) = O_p(r_n).\end{aligned}$$

This proves

$$\sup_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_{A,n}} |R_n(w, A) - R(w, A)| = O_p(r_n).$$

Finally,

$$|Q_n(w, A) - R(w, A)| \le |R_n(w, A) - R(w, A)| + |\mathcal{P}_n(w, A)|.$$

Taking suprema and using Assumption 4.4—not Assumption 4.3—gives

$$\sup_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_{A,n}} |Q_n(w, A) - R(w, A)| \le \zeta_n + \pi_n = O_p(r_n + \pi_n).$$

□

Theorem 3.14 (Approximate regularized neural-sieve oracle inequality)**.** *Under Assumptions 4.1–4.4 in the Main Paper,*

$$R(\widehat{w}_s, \widehat{A}_s) \le R(w^\star_{s,n}, A^\star_{s,n}) + 2\zeta_n + 2\pi_n + \varepsilon_n^{\mathrm{NN}}. \tag{64}$$

Consequently,

$$R(\widehat{w}_s, \widehat{A}_s) \le R(w^\star_{s,n}, A^\star_{s,n}) + O_p\left(r_n + \pi_n + \varepsilon_n^{\mathrm{NN}}\right). \tag{65}$$

Proof. Approximate empirical optimality gives

$$Q_n(\widehat{w}_s, \widehat{A}_s) \le Q_n(w^\star_{s,n}, A^\star_{s,n}) + \varepsilon_n^{\mathrm{NN}}.$$

Moreover, uniformly over the neural sieves,

$$|Q_n(w, A) - R(w, A)| \le \zeta_n + \pi_n.$$

It follows that

$$\begin{aligned} R(\widehat{w}_s, \widehat{A}_s) &\le Q_n(\widehat{w}_s, \widehat{A}_s) + \zeta_n + \pi_n \\ &\le Q_n(w^\star_{s,n}, A^\star_{s,n}) + \zeta_n + \pi_n + \varepsilon_n^{\mathrm{NN}} \\ &\le R(w^\star_{s,n}, A^\star_{s,n}) + 2\zeta_n + 2\pi_n + \varepsilon_n^{\mathrm{NN}}. \end{aligned}$$

The result follows from $\zeta_n = O_p(r_n)$. □

Proof of ***Theorem 4.8 in the Main Paper****.* By the definition of

$$a_n = \max_{A \in \mathcal{A}_s} \sup_{w \in \mathcal{W}_A^0} \inf_{v \in \mathcal{W}_{A,n}} d_\infty(v, w)$$

and the risk-Lipschitz bound, there exists a neural-sieve comparator (A_n°, w_n°) satisfying

$$R(w_n^\circ, A_n^\circ) \le R_s^\star + C_R a_n + o(a_n).$$

Therefore,

$$R(w^\star_{s,n}, A^\star_{s,n}) \le R_s^\star + C_R a_n + o(a_n).$$

Combining this inequality with Theorem 3.14 gives

$$R(\widehat{w}_s, \widehat{A}_s) - R_s^\star \le 2\zeta_n + C_R a_n + 2\pi_n + \varepsilon_n^{\mathrm{NN}} + o(a_n). \tag{66}$$

Hence

$$R(\widehat{w}_s, \widehat{A}_s) - R_s^\star = O_p\left(r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}}\right).$$

The consistency and first-order conclusions in Theorem 4.6 follow immediately. □

3.7 Deep ReLU generalization by substitution

Let $\mathcal{W}_{A,n}(L_n, K_n, B_n)$ contain ReLU–softmax gates with depth L_n, at most K_n nonzero coefficients, and coefficient bound B_n. The following principle avoids repeating every oracle and set-convergence proof.

Theorem 3.15. *Suppose a replacement architecture satisfies:*

1. *a uniform concentration bound $\zeta_n = O_p(r_n^{\mathrm{arch}})$;*
2. *a uniform population-risk approximation bound a_n^{arch};*
3. *the same compactness, separation, and approximate-minimization conditions used above.*

Then every formal result in 3.5 remains valid after the simultaneous substitutions $r_n \leftarrow r_n^{\mathrm{arch}}$ and $a_n \leftarrow a_n^{\mathrm{arch}}$.

Proof. Let ζ_n^{arch} denote the architecture-specific empirical deviation. In the oracle proof, replace the shallow comparator by an architecture-specific comparator whose risk is at most $R_s^\star + a_n^{\mathrm{arch}}$. Approximate empirical optimality then gives

$$R(\widehat{w}_s, \widehat{A}_s) - R_s^\star \le 2\zeta_n^{\mathrm{arch}} + a_n^{\mathrm{arch}} + 2\pi_n + \varepsilon_n^{NN}.$$

Every subsequent risk-to-distance argument uses only this excess-risk bound and the same separation modulus. The profiled criterion and marginal-gain errors in forward selection are obtained by applying the same deviation and approximation bounds to each admissible support; therefore δ_n changes only by the two stated substitutions. The metric-regularity, support-margin, and image-continuity arguments contain no further architecture-specific step. Applying the substitutions in each displayed bound proves the claim. □

For Sobolev-smooth target logits, Yarotsky (2017) constructs deep ReLU networks with depth growing logarithmically in inverse accuracy and, up to logarithmic factors,

$$a_n^{\mathrm{deep}} \lesssim \beta_n + R_{g,n}\left(\frac{K_n}{\log K_n}\right)^{-\nu/s}.$$

The statistical entropy is still then the parameter entropy of the participating deep neural gate. A bound of the form $K_n L_n \log(C\mathfrak{L}_n^{\mathrm{deep}} n)$ then yields

$$r_n^{\mathrm{deep}} \asymp \sqrt{\frac{K_n L_n \log(C\mathfrak{L}_n^{\mathrm{deep}} n) + s\log(ed/s)}{n}} + \frac{b_n\{K_n L_n \log(C\mathfrak{L}_n^{\mathrm{deep}} n) + s\log(ed/s)\}}{n}.$$

Depth improves approximation efficiency only when its reduction of a_n dominates the associated statistical and computational costs.

3.8 Set Convergence of Sparse Neural Gates

Define the neural-sieve population risk

$$R_{s,n}^{\inf} = \inf_{\substack{A\in\mathcal{A}_s \\ w\in\mathcal{W}_{A,n}}} R(w,A)$$

and the possible lower-side sieve discrepancy

$$\chi_n = \left[R_s^\star - R_{s,n}^{\inf}\right]_+. \tag{67}$$

If $\mathcal{W}_{A,n} \subseteq \mathcal{W}_A^0$, then $\chi_n = 0$. More generally, $\chi_n \to 0$ is a sieve-risk compatibility condition ensuring that the neural sieves do not asymptotically attain population risk strictly below the stated ideal target.

For $\tau_n \geq 0$, define the regularized empirical near-minimizer set

$$\widehat{\mathcal{E}}_{s,n}(\tau_n) = \left\{ (A,w) : \begin{array}{l} A\in\mathcal{A}_s,\ w\in\mathcal{W}_{A,n}, \\ Q_n(w,A) \leq \inf\limits_{\substack{B\in\mathcal{A}_s \\ v\in\mathcal{W}_{B,n}}} Q_n(v,B) + \tau_n + \varepsilon_n^{\mathrm{NN}} \end{array} \right\}. \tag{68}$$

The term τ_n determines the inferential near-minimizer set, while $\varepsilon_n^{\mathrm{NN}}$ accounts for numerical criterion error.

Theorem 3.16 (Set convergence of sparse neural gates-**Theorem 4.10 in the Mail Paper).** *Suppose the main paper's finite-net concentration, conditional sub-Gaussian, shallow-sieve,*

vanishing-regularization, and risk-separation conditions hold. If

$$r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}} + \tau_n \to 0,$$

then

$$\sup_{e \in \widehat{\mathcal{E}}_{s,n}(\tau_n)} d(e, \mathcal{E}_s^\star) \xrightarrow{p} 0. \tag{69}$$

If every element of $\mathcal{E}_s^\star$ admits the uniform sieve approximation above and

$$r_n + a_n + \pi_n + \varepsilon_n^{\mathrm{NN}} + \chi_n = o(\tau_n), \qquad \tau_n \to 0, \tag{70}$$

then

$$d_H(\widehat{\mathcal{E}}_{s,n}(\tau_n), \mathcal{E}_s^\star) \xrightarrow{p} 0. \tag{71}$$

Proof. Let $\zeta_n = \sup_{A,w} |R_n(w, A) - R(w, A)| = O_p(r_n)$. For $e = (A, w) \in \widehat{\mathcal{E}}_{s,n}(\tau_n)$, comparison with a neural approximation to an ideal oracle gives

$$R(e) - R_s^\star \le 2\zeta_n + 2\pi_n + C_R a_n + a u_n + \varepsilon_n^{\mathrm{NN}} + o(a_n).$$

The right-hand side is $o_p(1)$. If (69) was false, risk separation would give some $\epsilon > 0$ and, with probability bounded away from zero, an e in the near-minimizer set satisfying $R(e) - R_s^\star \ge \Delta_s(\epsilon) > 0$, a contradiction.

For the reverse inclusion, fix $e^\star = (A^\star, w^\star)$ in the oracle set and choose $w_n^\star \in \mathcal{W}_{A^\star,n}$ uniformly such that $d_\infty(w_n^\star, w^\star) \le a_n + (1 + o(1))$ and $R(w_n^\star, A^\star) \le R_s^\star + C_R a_n + o(a_n)$. Also,

$$\inf_{A,w} Q_n(w, A) \ge R_{s,n}^{\inf} - \zeta_n - \pi_n \ge R_s^\star - \chi_n - \zeta_n - \pi_n.$$

Consequently,

$$\begin{aligned} Q_n(w_n^\star, A^\star) - \inf_{A,w} Q_n(w, A) &\le C_R a_n + 2\zeta_n + 2\pi_n + \chi_n + o(a_n) \\ &= o_p(\tau_n). \end{aligned}$$

Thus $(A^\star, w_n^\star)$ belongs to the empirical near-minimizer set with probability tending to one.

Uniformity over the compact oracle set yields the reverse Hausdorff inclusion; combining it with the outer inclusion proves (71). □

3.9 Population and Empirical Forward Selection

For each admissible support, define the ideal population profile risk and gain

$$L(A) = \inf_{w \in \mathcal{W}_A^0} R(w, A), \qquad G(A) = L(\varnothing) - L(A). \tag{72}$$

Assume that G is monotone and weakly submodular: for some $\gamma_s \in (0, 1]$,

$$\sum_{j \in B} \left(G(A \cup \{j\}) - G(A) \right) \geq \gamma_s \left(G(A \cup B) - G(A) \right) \tag{73}$$

for every admissible pair of disjoint sets A, B for which the union has the required sparse cardinality. Let

$$A_s^\star \in \arg\max_{A \in \mathcal{A}_s} G(A)$$

and let A_k^{FS} be a population greedy set after k steps, starting from $A_0^{FS} = \varnothing$.

Theorem 3.17 (Population finite-iteration guarantee-**Theorem 4.11 in the Main Paper).** *For every $k \leq s$,*

$$G(A_k^{FS}) \geq \left(1 - \left(1 - \frac{\gamma_s}{s} \right)^k \right) G(A_s^\star) \geq (1 - e^{-k\gamma_s/s}) G(A_s^\star). \tag{74}$$

Consequently, the finite-iteration population greedy gap obeys

$$L(A_k^{FS}) - L(A_s^\star) = G(A_s^\star) - G(A_k^{FS}) \tag{75}$$

satisfies

$$0 \leq L(A_k^{FS}) - L(A_s^\star) \leq \rho_{s,k}^{FS}, \qquad \rho_{s,k}^{FS} := e^{-k\gamma_s/s} G(A_s^\star). \tag{76}$$

Proof. Put $D_q = G(A_s^\star) - G(A_q^{FS})$. Applying weak submodularity at A_q^{FS} gives

$$\sum_{j \in A_s^\star \setminus A_q^{FS}} \left(G(A_q^{FS} \cup \{j\}) - G(A_q^{FS}) \right) \geq \gamma_s \left(G(A_q^{FS} \cup A_s^\star) - G(A_q^{FS}) \right).$$

Monotonicity makes the right-hand side at least $\gamma_s D_q$. There are at most s summands, so some available coordinate has gain at least $\gamma_s D_q/s$. The greedy step achieves at least this gain, and hence

$$D_{q+1} \le (1 - \gamma_s/s) D_q.$$

Because $G(\varnothing) = 0$, iteration gives $D_k \le (1-\gamma_s/s)^k G(A_s^\star)$. The inequality $1-u \le e^{-u}$ completes the proof. □

3.9.1 Restricted curvature and the weak-submodularity index

The weak-submodularity condition may be verified through restricted curvature in a common local coordinate chart for the profiled gating problem. Let $\mathcal{Q}(\vartheta)$ and $\mathcal{Q}_n(\vartheta)$ denote the population and empirical local objectives.

Assumption 3.18 (Restricted population curvature and empirical Hessian convergence)**.** *There is a neighborhood $\mathcal{N}$ of the relevant population paths and constants $0 < m_s \le M_s < \infty$ such that, for every $\vartheta \in \mathcal{N}$ and every unit direction u supported on at most $2s$ coordinates,*

$$m_s \le u' \nabla^2 \mathcal{Q}(\vartheta) u \le M_s. \tag{77}$$

Moreover,

$$\sup_{\vartheta \in \mathcal{N}} \|\nabla^2 \mathcal{Q}_n(\vartheta) - \nabla^2 \mathcal{Q}(\vartheta)\|_{\mathrm{op},2s} = O_p(h_n), \qquad h_n = o(1), \tag{78}$$

where $\|H\|_{\mathrm{op},2s}$ is the supremum of $|u'Hu|$ over such unit sparse directions. The restricted-curvature argument applies to the profiled gain, so that

$$\gamma_s = c_\gamma m_s / M_s \tag{79}$$

for a fixed normalization constant $c_\gamma > 0$.

Let $\widehat{m}_{s,n}$ and $\widehat{M}_{s,n}$ be the corresponding empirical restricted eigenvalue extrema on $\mathcal{N}$, and define

$$\widehat{\gamma}_{s,n} = c_\gamma \widehat{m}_{s,n} / \widehat{M}_{s,n}. \tag{80}$$

Proposition 3.19 (Learning the weak-submodularity index)**.** *Under Assumption 3.18,*

$$\widehat{m}_{s,n} = m_s + O_p(h_n),\ \widehat{M}_{s,n} = M_s + O_p(h_n), \tag{81}$$

$$\widehat{\gamma}_{s,n} = \gamma_s + O_p(h_n). \tag{82}$$

For fixed $k \le s$,

$$1 - e^{-k\widehat{\gamma}_{s,n}/s} = 1 - e^{-k\gamma_s/s} + O_p(h_n). \tag{83}$$

Proof. For every admissible unit direction u,

$$|u'(\nabla^2 \mathcal{Q}_n(\vartheta) - \nabla^2 \mathcal{Q}(\vartheta))u| \le \|\nabla^2 \mathcal{Q}_n(\vartheta) - \nabla^2 \mathcal{Q}(\vartheta)\|_{\mathrm{op},2s}.$$

Taking the relevant infima and suprema over u and ϑ gives the restricted Weyl bounds (Horn and Johnson, 2012, Ch. 4)

$$|\widehat{m}_{s,n} - m_s| = O_p(h_n), \qquad |\widehat{M}_{s,n} - M_s| = O_p(h_n).$$

Since $M_s > 0$, $\widehat{M}_{s,n} \ge M_s/2$ with probability tending to one. On that event,

$$\frac{\widehat{m}_{s,n}}{\widehat{M}_{s,n}} - \frac{m_s}{M_s} = \frac{\widehat{m}_{s,n} - m_s}{\widehat{M}_{s,n}} + m_s \frac{M_s - \widehat{M}_{s,n}}{M_s \widehat{M}_{s,n}} = O_p(h_n).$$

Multiplication by c_γ proves (82). The last assertion follows from the mean-value theorem because the derivative of $u \mapsto 1 - e^{-ku/s}$ is bounded by $k/s \le 1$ on $[0, \infty)$. □

Assumption 3.18 separates two questions. The positive population lower eigenvalue supplies a nondegenerate greedy guarantee. The rate h_n measures how accurately that geometry is learned. When m_s is small, modest Hessian perturbations can produce large changes in the empirical greedy path even though the profile risks remain close.

3.9.2 Profile accuracy and computational errors

Let

$$L_n^{\mathcal{Q}}(A) = \inf_{w \in \mathcal{W}_{A,n}} Q_n(w, A) \tag{84}$$

be the exact regularized neural-sieve profile, and let $\widehat{L}_n(A)$ be the value returned by the neural optimizer. The profiled neural optimization tolerance is defined by

$$0 \le \widehat{L}_n(A) - L_n^Q(A) \le \varepsilon_n^{\mathrm{NN}} \tag{85}$$

uniformly over every support evaluated by the algorithm.

Assumption 3.20 (Uniform profile accuracy)**.** *Uniformly over all supports and one-coordinate extensions evaluated by forward selection,*

$$\sup_A |\widehat{L}_n(A) - L(A)| = O_p(\delta_n), \qquad \delta_n = r_n + C_R a_n + \pi_n + \varepsilon_n^{\mathrm{NN}}. \tag{86}$$

The absolute comparison in (86) follows directly when the neural sieve is contained in the ideal class. More generally, it requires the corresponding two-sided sieve-risk compatibility condition; the one-sided approximation bound alone is insufficient for an absolute profile comparison.

Define population and computed marginal gains by

$$\Delta(j \mid A) = L(A) - L(A \cup \{j\}), \tag{87}$$

$$\widehat{\Delta}_n(j \mid A) = \widehat{L}_n(A) - \widehat{L}_n(A \cup \{j\}). \tag{88}$$

Assumption 3.20 implies

$$\sup_{A,j} |\widehat{\Delta}_n(j \mid A) - \Delta(j \mid A)| = O_p(\delta_n), \tag{89}$$

because each marginal gain is a difference of two profile values.

At empirical step q, allow the implemented search to select $\widehat{j}_q$ satisfying

$$\widehat{\Delta}_n(\widehat{j}_q \mid \widehat{A}_q^{FS}) \ge \max_{j \notin \widehat{A}_q^{FS}} \widehat{\Delta}_n(j \mid \widehat{A}_q^{FS}) - \eta_{n,q}^{FS}, \tag{90}$$

where $\eta_{n,q}^{FS} \ge 0$ permits an approximate or prematurely terminated coordinate search. Set

$$\varepsilon_{n,k}^{FS} = \sum_{q=0}^{k-1} \eta_{n,q}^{FS}. \tag{91}$$

Thus $\varepsilon_n^{\mathrm{NN}}$ measures optimization within neural profiles, whereas $\varepsilon_{n,k}^{FS}$ measures perturbation of

the outer greedy coordinate choices. Neither term is the population greedy approximation gap $\rho_{s,k}^{FS}$.

3.9.3 Empirical finite-iteration guarantees

Theorem 3.21 (Empirical finite-iteration forward selection)**.** *Let $\widehat{A}_k^{FS}$ be produced by (90) for $k \le s$. Under weak submodularity and Assumption 3.20,*

$$G(\widehat{A}_k^{FS}) \ge (1 - e^{-k\gamma_s/s})G(A_s^\star) - O_p(k\delta_n) - \varepsilon_{n,k}^{FS}, \tag{92}$$

$$L(\widehat{A}_k^{FS}) - L(A_s^\star) \le \rho_{s,k}^{FS} + O_p(k\delta_n) + \varepsilon_{n,k}^{FS}. \tag{93}$$

If the guarantee is reported using the learned factor $1 - e^{-k\widehat{\gamma}_{s,n}/s}$, then under Assumption 3.18 the right-hand sides acquire an additional $O_p(h_n)$ term.

Proof. Work on events whose probability tends to one and on which (89) is bounded by $C\delta_n$. For any available coordinate j, approximate empirical greedy choice gives

$$\begin{aligned}
\Delta(\widehat{j}_q \mid \widehat{A}_q^{FS}) &\ge \widehat{\Delta}_n(\widehat{j}_q \mid \widehat{A}_q^{FS}) - C\delta_n \\
&\ge \widehat{\Delta}_n(j \mid \widehat{A}_q^{FS}) - C\delta_n - \eta_{n,q}^{FS} \\
&\ge \Delta(j \mid \widehat{A}_q^{FS}) - 2C\delta_n - \eta_{n,q}^{FS}.
\end{aligned}$$

The population weak-submodularity argument applies at the random but admissible set $\widehat{A}_q^{FS}$, yielding

$$\max_j \Delta(j \mid \widehat{A}_q^{FS}) \ge \frac{\gamma_s}{s}\{G(A_s^\star) - G(\widehat{A}_q^{FS})\}.$$

With $D_q = G(A_s^\star) - G(\widehat{A}_q^{FS})$, it follows that

$$D_{q+1} \le (1 - \gamma_s/s)D_q + 2C\delta_n + \eta_{n,q}^{FS}.$$

Iteration gives

$$\begin{aligned} D_k &\le (1-\gamma_s/s)^k G(A_s^\star) + 2C\delta_n \sum_{q=0}^{k-1}(1-\gamma_s/s)^q \\ &\quad + \sum_{q=0}^{k-1}(1-\gamma_s/s)^{k-1-q}\eta_{n,q}^{FS} \\ &\le e^{-k\gamma_s/s} G(A_s^\star) + 2Ck\delta_n + \varepsilon_{n,k}^{FS}. \end{aligned}$$

This proves both displays. Proposition 3.19 and the mean-value theorem add $O_p(h_n)$ when the estimated curvature factor replaces the population factor. □

The complete computational decomposition is therefore

$$\underbrace{\varepsilon_n^{\mathrm{NN}}}_{\text{within-support neural optimization}} + \underbrace{\varepsilon_{n,k}^{FS}}_{\text{outer greedy-choice approximation}}, \tag{94}$$

while $\rho_{s,k}^{FS}$ is a conservative population approximation bound for using only k greedy steps and h_n measures estimation of the curvature that supports the greedy guarantee.

Assumption 3.22 (Local risk-to-set error bound)**.** *There are $C_\Pi < \infty$, $\bar{r} > 0$, and $\alpha \in (0, 1]$ such that*

$$d((A, w), \mathcal{E}_s^\star) \le C_\Pi \left[R(w, A) - R_s^\star\right]^\alpha \tag{95}$$

whenever $0 \le R(w, A) - R_s^\star \le \bar{r}$.

Corollary 3.23 (Distance rate for empirical forward selection)**.** *Suppose the fitted gate on $\widehat{A}_k^{FS}$ obeys the same neural profile tolerance as (85), and suppose Assumption 3.22 holds. Then*

$$d\Big((\widehat{A}_k^{FS}, \widehat{w}_{\widehat{A}_k^{FS}}), \mathcal{E}_s^\star\Big) = O_p\left[\left(\rho_{s,k}^{FS} + k\delta_n + \varepsilon_{n,k}^{FS} + h_n\right)^\alpha\right]. \tag{96}$$

For fixed s, the factor k may be absorbed into the constant. If $h_n = O_p(r_n)$, curvature learning does not slow the first-order rate.

Proof. Theorem 3.21, the final profiled fit, and the definition of δ_n imply

$$R(\widehat{w}_{\widehat{A}_k^{FS}}, \widehat{A}_k^{FS}) - R_s^\star = O_p\left(\rho_{s,k}^{FS} + k\delta_n + \varepsilon_{n,k}^{FS} + h_n\right).$$

When this remainder converges to zero, the local error bound applies with probability tending to one. Raising the excess-risk rate to the power α proves (96). □

Remark 3.24 (Implications for bootstrap variation)**.** *When marginal population gains are tied or nearly tied, bootstrap samples can follow different near-optimal greedy paths. Small restricted eigenvalues amplify this switching because an $O_p(h_n)$ curvature perturbation can materially change the learned greedy factor and local optimizer behavior. Thus wide bootstrap confidence strips may reflect a combination of set-valued identification, $\varepsilon_n^{\mathrm{NN}}$, $\varepsilon_{n,k}^{FS}$, and weak local curvature, even when excess predictive risk is small. The empirical evidence should therefore be described as consistent with this mechanism unless stepwise gain gaps, restart dispersion, Hessian spectra, and within-path versus between-path bootstrap variation are reported.*

3.10 Interpretation of the Set-Valued Target

The set-valued formulation is important when either the candidate predictive mechanisms or the state coordinates are redundant. If two different state subsets induce the same adaptive predictor, the population risk cannot distinguish between them. Likewise, if candidate predictions are locally linearly dependent, different simplex-valued weight maps may generate the same predictive surface.

Consequently, exact recovery of a particular state subset or weight representation is generally stronger than required for predictive consistency. The pseudo-true set $\mathcal{E}_s^\star$ collects all sparse representations attaining the optimal adaptive $L^2(\mathbb{P})$-projection risk.

The forward-selection oracle set $\mathcal{E}_s^{FS,\star}$ has a related but computational interpretation. It collects the population solutions reachable through potentially nonunique greedy paths. Ties in marginal gains and approximate redundancy among state variables may therefore lead different empirical samples toward different elements of this set.

The statistical content of the 3.23 result is that this nonuniqueness does not prevent recovery of the relevant predictive geometry. The empirical procedure approaches the population greedy oracle set, while its distance from the exact sparse oracle is controlled by the intrinsic greedy approximation gap ρ_s.

Thus the appropriate asymptotic object is not necessarily a uniquely identified sparse coordinate vector. It is the collection of low-dimensional state representations that generate

optimal or near-optimal adaptive predictive geometry.

4 Descriptive statistics

Life expectancy at birth, fertility rate (total births per woman) are all from the World Bank data base (World Bank, 2026). Investment as share of GDP, Exports as share of GDP and Imports as share of GDP come from the Penn World Table 10 (Feenstra et al., 2015). The Debt variable captures public debt and comes from the IMF (Mbaye et al., 2018; International Monetary Fund, 2026), the Human Capital variable is total average school for both males and females from the Barro and Lee dataset (Barro and Lee, 2013). The variable Democracy and Executive Constraints come from from polity 5, the "Polity5 Annual Time-Series, 1946-2018" (Marshall and Gurr, 2020), where for 2020 we used the values of 2015. The variable Openness describes the trade balance of the difference between exports and imports as GDP share. It is used to capture open economies with different characteristics e.g. economies that are export oriented take positive values, while countries with external trade deficits take negative values. This heterogeneity feature with respect to the nature of the trade composition was very important during the debt crisis. In this context Japan with a substantial debt to GDP ratio, yet an export oriented economy with a substantial positive trade balance had a very different experience from a country like Greece which primarily imports most of it traded goods. Such an imbalance between the size of imports and exports constitutes an important dimension of heterogeneity in this case. In the supplement we present the descriptive statistics of the data used. The empirical application uses a cross-country macroeconomic growth dataset containing 639 observations. We have a balanced five-year raw panel covering 71 countries over nine non-overlapping five-year periods: 1980–1984,1985–1989, 1990- 1994, 1995–1999, 2000–2004, 2005–2009, 2010–2014, 2015–2019, and 2020–2023. The raw panel therefore contains 639 country-period observations. The dependent variable is annualized five-year economic growth, denoted by EG. Public debt is measured as general government debt as a percentage of GDP and is averaged within each five-year block. All variables are averaged within each five-year block. This five-year aggregation is standard in the empirical growth literature and helps smooth short-run business-cycle fluctuations and

temporary crisis episodes. [1]

Table 1 reports the descriptive statistics for the data used.

Table 1: Summary Statistics

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Initial income	8.502	1.582	5.200	7.171	8.390	9.972	11.583
Growth	0.157	0.297	-0.703	0.030	0.153	0.277	3.938
Population growth	0.016	0.011	-0.007	0.007	0.015	0.025	0.072
Inflation	0.197	1.281	-0.025	0.027	0.057	0.108	25.228
Life expectancy	4.205	0.162	3.683	4.099	4.242	4.333	4.434
Fertility rate	0.035	0.018	0.010	0.018	0.030	0.050	0.077
Investment	0.217	0.089	0.019	0.152	0.218	0.273	0.629
Government	0.171	0.065	0.042	0.125	0.163	0.207	0.536
Public debt	0.552	0.374	0.017	0.306	0.474	0.689	2.806
Openness	-0.046	0.168	-1.004	-0.098	-0.029	0.020	0.588
Education	1.853	0.568	-0.435	1.566	1.996	2.295	2.586
Democracy	5.836	3.870	0.000	2.000	7.000	10.000	10.000
Executive constraints	5.054	2.065	0.000	3.000	6.000	7.000	7.000
Religion	0.413	0.262	0.003	0.197	0.384	0.660	0.860
Language	0.389	0.299	0.002	0.105	0.335	0.687	0.898
tropical	0.403	0.444	0.000	0.000	0.117	0.946	1.000
dist_coast	0.287	0.283	0.000	0.058	0.215	0.398	1.186
LCR100km	0.403	0.370	0.000	0.094	0.281	0.774	1.000

5 Empirical Application

Theoretical work such as the seminal work of Azariadis and Drazen (1990) on threshold externalities as well as the recent work on endogenous economic take-offs (e.g., Galor and Moav (2002), Galor (2005), Chamon and Kremer (2009)) suggest that heterogeneity across countries may be characterized by threshold nonlinearities. In this context, there has been also a substantial amount of work on the effect of public debt on economic growth emanating from the financial crises of the last decade, yet the effect that different countries with similarly

[1]The data sources are from PWT 11 https://www.rug.nl/ggdc/productivity/pwt/?lang=en and the World Bank data base https://databank.worldbank.org/source/world-development-indicators

high public debt to GDP ratios (Japan and Greece for example) have been affected in a vastly different manner in their ability to borrow due to their external debt situation. In fact the openness of each economy and its ability to raise funds to finance its expenditures is an important factor that points to the heterogeneous nature of the public debt issue. It is important to recognize that the composition of trade openness also matters, since Japan for example is an export oriented economy with substantial positive trade balance, whereas a country like Greece primarily imports most of it traded goods. Such an imbalance between the size of imports and exports constitutes an important dimension of heterogeneity in this case. In our case we include the variable "openness" defined as the sum of imports (a negative number) and exports (a positive number) as a share of GDP to capture this imbalance. In this case the variable openness for economies that are export oriented take positive values, while countries with external trade deficits take negative values.

5.0.1 Inferential Procedures

Predictive superiority is evaluated using one-sided Diebold–Mariano-type procedures (Diebold and Mariano, 2002). Let:

$$L^{(a)} = (Y - \widehat{Y}^{(a)})^2$$

denote the loss associated with forecasting method a. For a benchmark method b and a competing adaptive method m, define:

$$d = L^{(b)} - L^{(m)}. \tag{97}$$

The null hypothesis is:

$$H_0 : \mathbb{P}[d] \leq 0,$$

against:

$$H_1 : \mathbb{P}[d] > 0.$$

Thus the alternative corresponds directly to superior predictive performance of the adaptive method. The test statistic is:

$$\mathrm{DM} = \frac{\sqrt{n}\,\bar{d}}{\widehat{\sigma}_{\mathrm{LR}}}, \tag{98}$$

where:

$$\bar{d} = \mathbb{P}_n[d],$$

and $\widehat{\sigma}_{\mathrm{LR}}$ is anof the variance of d.

Bootstrap inference uses clustered country-level resampling. Entire country trajectories are resampled with replacement in order to preserve within-country dependence structure. For each bootstrap replication, the full estimation pipeline is repeated, including:

- threshold estimation,
- adaptive gate training,
- forward selection,
- predictive evaluation.

The bootstrap procedure therefore captures the joint uncertainty associated with both the predictive models and the learned adaptive geometry.

Finally, sparse-state stability is evaluated through bootstrap state-selection frequencies. Let: $\widehat{A}^{(b)}$ denote the forward-selection subset obtained in bootstrap replication b. The selection frequency of state variable j is:

$$\widehat{\pi}_j = \frac{1}{B} \sum_{b=1}^{B} I(j \in \widehat{A}^{(b)}). \tag{99}$$

These frequencies quantify the stability of the learned low-dimensional regime geometry.

Below we present a number of additional robustness checks that confirm the overall superiority of the neural gate. These checks include leave-one-period-out cross-fitting, clustered bootstrap confidence intervals, diagnostics based on the dominant neural-gate model for each observation as well as various additional weight-curve visualization experiment.

5.1 Cross-Fitted Predictive Inference

We implement leave-one-period-out cross-fitting. In each fold, all observations from one five-year period are held out, the candidate library and gates are refitted on the remaining periods, and

losses are evaluated only on the held-out observations. Table 2 reports the resulting cross-fitted MSEs.

Table 2: Leave-One-Period-Out Cross-Fitted Performance

Method	Cross-Fitted MSE	Cross-Fitted RMSE
Neural Gate	0.0809	0.2844
Fixed LS Average	0.0809	0.2845
Naive Average	0.0813	0.2852
Entropy Neural Gate	0.0815	0.2855
Linear Gate	0.0823	0.2868
Best Single	0.0828	0.2877

The neural gate remains the lowest-MSE procedure under this stricter out-of-sample evaluation, but the advantage is much smaller than in the in-sample comparison. Its cross-fitted MSE is 0.0809, very close to the fixed least-squares average. Table 3 therefore reports predictive-superiority tests based on the held-out cross-fitted losses.

Table 3: Cross-Fitted Predictive-Superiority Tests

Benchmark	Model	Mean Loss Difference	DM Statistic	One-Sided p-Value
Best Single	Neural Gate	0.0019	1.129	0.129
Naive Average	Neural Gate	0.0004	0.317	0.376
Fixed LS Average	Neural Gate	0.0001	0.049	0.481
Linear Gate	Neural Gate	0.0014	1.318	0.094
Entropy Neural Gate	Neural Gate	0.0006	0.585	0.279

The cross-fitted tests are conservative. They no longer deliver small p-values against the strongest benchmarks, even though the neural gate retains the best average held-out MSE. This distinction is useful for the empirical interpretation. The in-sample and clustered-bootstrap results show that the neural gate captures meaningful nonlinear structure in the observed sample, while the cross-fitted results indicate that the out-of-period predictive advantage is modest. We therefore interpret the empirical findings as evidence of economically informative adaptive regime geometry, with predictive gains that are present but moderate under strict temporal cross-fitting.

5.2 Bootstrap Uncertainty

Table 4 reports clustered bootstrap confidence intervals. The bootstrap distribution confirms the ranking of the main procedures: the neural gate has the lowest bootstrap mean MSE, and the relative gain of the neural gate over the benchmark procedures remains positive across the bootstrap distribution.

Table 4: Bootstrap MSE and Relative-Gain Confidence Intervals

Quantity	Bootstrap Estimate	95% Interval
Best Single	0.0770	[0.0394, 0.1287]
Naive Average	0.0778	[0.0402, 0.1319]
Fixed LS Average	0.0761	[0.0391, 0.1269]
Linear Gate	0.0770	[0.0397, 0.1293]
Neural Gate	0.0675	[0.0356, 0.1058]
Entropy Neural Gate	0.0742	[0.0397, 0.1152]
Gain: Neural vs Best Single	0.1179	[0.0664, 0.2265]
Gain: Neural vs Fixed LS	0.1079	[0.0566, 0.2176]
Gain: Neural vs Linear Gate	0.1180	[0.0606, 0.2153]

The bootstrap evidence is economically meaningful because it suggests that the neural gate's gain is not driven by a small number of high-leverage observations or by a single deterministic train-test split.

5.3 Blocked Out-of-Sample Validation

To assess whether the adaptive neural gate generalizes across time, we also conduct leave-one-period-out validation. In each fold, all observations from one five-year period are held out, the model library and gates are refitted on the remaining periods, and predictions are evaluated on the held-out period. Table 5 reports the resulting validation errors.

Table 5: Blocked Out-of-Sample Validation

Block	Method	Weighted MSE	Mean Fold MSE	Weighted MAE	Bias
Time	Best Single	0.0828	0.0828	0.1776	-0.0021
Time	Naive Average	0.0813	0.0813	0.1749	-0.0031
Time	Fixed LS Average	0.0809	0.0809	0.1742	-0.0031
Time	Linear Gate	0.0823	0.0823	0.1756	-0.0033
Time	Neural Gate	0.0808	0.0808	0.1730	0.0034
Time	Entropy Neural Gate	0.0815	0.0815	0.1749	-0.0032

The leave-one-period-out results are more demanding than the in-sample comparison. The neural gate remains the best-performing procedure, although its advantage over fixed least-squares averaging is smaller. This pattern is reassuring: the nonlinear gate does not appear to achieve its gains purely by in-sample overfitting, but the out-of-sample evidence also indicates that the empirical gain is moderate once period-level validation is imposed.

5.4 Dominant-Model Regime Diagnostics

Table 6 reports diagnostics based on the dominant neural-gate model for each observation. Government consumption is the dominant local mechanism for 53.8% of observations, while life expectancy is dominant for 35.8%. Democracy is dominant for 9.7% of observations, and inflation dominates only a small high-inflation subset.

Table 6: Dominant-Model Regime Diagnostics

Dominant Model	n	Share	Avg. Weight	Growth	Inflation	Debt	Openness	RMSE
Government Consumption	344	0.538	0.708	0.190	0.093	0.609	-0.070	0.227
Life Expectancy	229	0.358	0.624	0.117	0.154	0.482	-0.017	0.199
Democracy	62	0.097	0.649	0.142	0.125	0.476	-0.020	0.554
Inflation	4	0.006	0.623	-0.097	12.633	0.736	-0.052	0.173

This diagnostic clarifies the economic content of the neural gate. The estimator does not merely produce a small forecasting improvement; it also partitions the sample into interpretable

local regimes in which different threshold mechanisms dominate. The small inflation-dominant group is especially distinctive, with much higher average inflation and negative average growth.

5.5 Gate Concentration

Table 7 summarizes the concentration of the neural-gate weights. The mean effective number of models is 2.95, well below the seven available candidate models. The maximum weight exceeds 0.50 for 79.3% of observations and exceeds 0.75 for 37.2% of observations.

Table 7: Neural-Gate Concentration Diagnostics

Diagnostic	Value
Mean entropy	1.012
Median entropy	1.051
Mean effective number of models	2.952
Share max weight > 0.50	0.793
Share max weight > 0.75	0.372
Share max weight > 0.90	0.103

These results indicate that the neural gate is not simply producing diffuse averaging. Instead, it behaves as a state-dependent local allocation rule: in most observations it concentrates predictive weight on a small number of candidate mechanisms, while retaining enough smoothness to avoid hard regime classification.

5.6 Sparse-State Selection

Table 8 reports clustered bootstrap state-selection frequencies from the forward-selection stage. Government consumption is selected most frequently, followed by debt, openness, and life expectancy. The selection frequencies should be interpreted as evidence about stable low-dimensional organizing coordinates rather than as literal structural inclusion probabilities, because several candidate state variables are correlated and can generate similar local regime partitions.

Table 8: Empirical Bootstrap State-Selection Frequencies

State Variable	Selection Frequency
Government Consumption	0.440
Debt	0.290
Openness	0.250
Life Expectancy	0.230
Fertility	0.130
Inflation	0.070
Democracy	0.010

80% band width is approximately 0.588 for the government-consumption

5.7 Conditional Gate-Parameter Inference

The country-resampling bands reported above include uncertainty from repeatedly refitting the candidate library and the neural gate. As a complementary diagnostic, we also compute narrower conditional bands that hold the fitted candidate library and neural architecture fixed and approximate uncertainty only around the final neural-gate parameters. Specifically, we use a local Laplace approximation based on the Jacobian of the fitted predictions with respect to the neural-gate parameters, draw nearby gates from the resulting Gaussian approximation, and evaluate the implied weight curves over the openness grid.

The calculation uses 5000 parameter draws and covariance scale 0.50. Because this diagnostic conditions on the selected library and architecture, the bands are not a replacement for the wider country-resampling bands. They answer a narrower question: how uncertain is the final neural allocation surface, given the estimated candidate mechanisms?

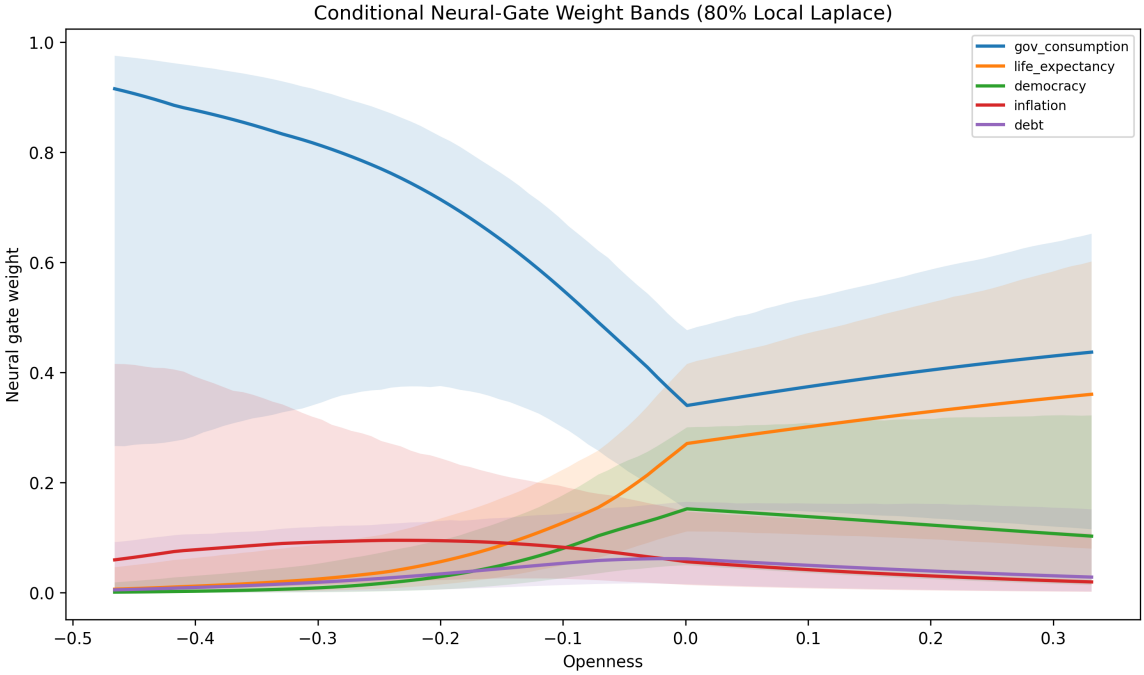

Figure 1: Conditional Neural-Gate Weight Curves with 80% Local Laplace Bands

Table 9: Average Width of Conditional Local-Laplace Neural-Gate Bands

Candidate Mechanism	80% Band	90% Band	95% Band
Gov Consumption	0.475	0.583	0.660
Life Expectancy	0.243	0.309	0.369
Democracy	0.173	0.240	0.303
Inflation	0.202	0.286	0.366
Debt	0.134	0.211	0.299

The average central 80% conditional band widths are Gov Consumption: 0.475, Life Expectancy: 0.243, Democracy: 0.173, Inflation: 0.202, Debt: 0.134. These bands are therefore much tighter than the fully refitted country-resampling bands and are useful for displaying the local shape of the learned allocation rule. The substantive interpretation should nevertheless continue to distinguish this conditional gate-parameter uncertainty from full sample-composition uncertainty.

5.8 Stabilized Neural-Gate Specification

The wide full-parameter neural-gate uncertainty bands may suggest (see also Section 7 below) that the original neural network contains weakly identified directions. We therefore estimate a small grid of more parsimonious neural gates that reduce the hidden-layer width, increase ridge regularization, and optionally add entropy regularization to avoid overly concentrated softmax weights.

Table 10: Neural-Gate Identification Tuning

Specification	Hidden Units	Ridge	Entropy	Cross-Fitted MSE	Avg. Weight Range	Avg. 80% Band Width
Moderate gate, H=4	4	1e-04	0.00	0.0802	0.196	0.186
Default neural gate	16	1e-04	0.00	0.0808	0.251	0.228
Moderate gate, H=12	12	1e-04	0.00	0.0811	0.230	0.222
Regularized gate, H=4	4	1e-02	0.00	0.0813	0.000	0.019
Regularized entropy gate, H=4	4	1e-02	0.05	0.0813	0.000	0.019
Strong entropy gate, H=4	4	1e-02	0.10	0.0813	0.000	0.019
Moderate gate, H=8	8	1e-04	0.00	0.0814	0.214	0.252
Small gate, H=8	8	1e-03	0.00	0.0816	0.000	0.062
Small gate, H=4	4	1e-03	0.00	0.0816	0.000	0.062
Mild ridge gate, H=16	16	5e-04	0.00	0.0821	0.051	0.099
Mild ridge gate, H=8	8	5e-04	0.00	0.0821	0.051	0.094
Mild ridge gate, H=4	4	5e-04	0.00	0.0823	0.033	0.091

The best tradeoff in this grid is Moderate gate, H=4. Its cross-fitted MSE is 0.0802, compared with 0.0808 for the default neural gate. Its average central 80% conditional band width is 0.186, compared with 0.228 under the default gate, while the average openness-range of the plotted weights is 0.196. This specification is therefore a compromise: it avoids the nearly flat gates produced by very strong shrinkage while still reducing conditional uncertainty relative to the most flexible network.

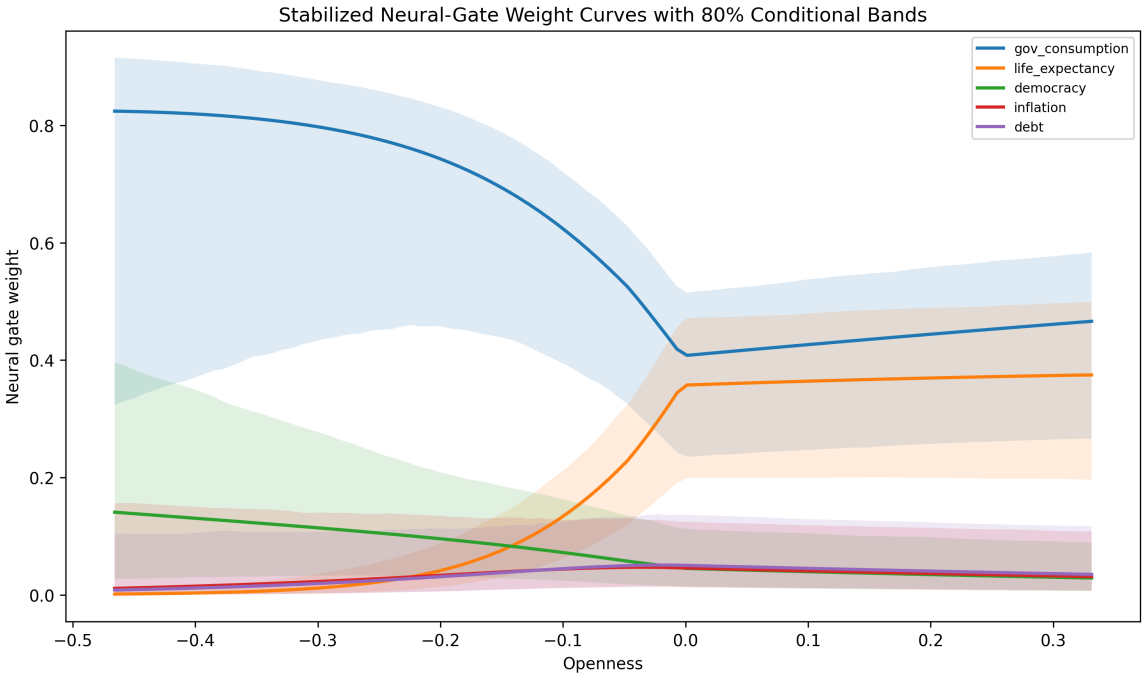

Figure 2: Stabilized Neural-Gate Weight Curves with 80% Conditional Bands

We therefore treat the selected specification as an identification-oriented robustness version of the neural gate, not as a replacement for the main flexible benchmark. The exercise shows the tradeoff clearly: very strong regularization produces tight bands by flattening the allocation surface, whereas moderate regularization preserves state dependence but leaves non-negligible uncertainty.

5.9 Curvature-Stabilized Gate Geometry

The broad country-resampling bands suggest that much of the empirical uncertainty comes from refitting the full candidate library and gate over different country compositions. To focus inference on the allocation surface itself, we estimate a curvature-stabilized neural gate. The estimator adds a penalty to the average squared second difference of the model weights along the openness grid, while leaving the candidate threshold library unchanged. The penalty is selected by leave-one-period-out cross-fitting, so the smoother gate is not chosen only for visual regularity.

Table 11: Curvature-Stabilized Gate Selection

Curvature Penalty	Cross-Fitted MSE	Weight Range	Curvature	80% Band
0	0.0805	0.308	6.95e-07	0.156
1	0.0806	0.008	8.28e-15	0.175
0.01	0.0811	0.075	8.10e-09	0.205
0.0001	0.0812	0.237	7.15e-07	0.205
0.1	0.0814	0.029	2.78e-14	0.207
0.001	0.0815	0.324	4.10e-07	0.175
0.05	0.0818	0.143	2.02e-09	0.205
0.5	0.0819	0.074	8.51e-11	0.202

The cross-fitting criterion selects the unpenalized small neural gate. Its cross-fitted MSE is 0.0805, and its average central 80% conditional band width is 0.156. Positive curvature penalties reduce the measured bending of the weight curves, but in this sample they also tend to flatten economically relevant state dependence or widen the conditional bands. Thus the curvature exercise supports using the parsimonious small neural gate as the main geometry-stabilized specification, rather than adding an additional curvature penalty.

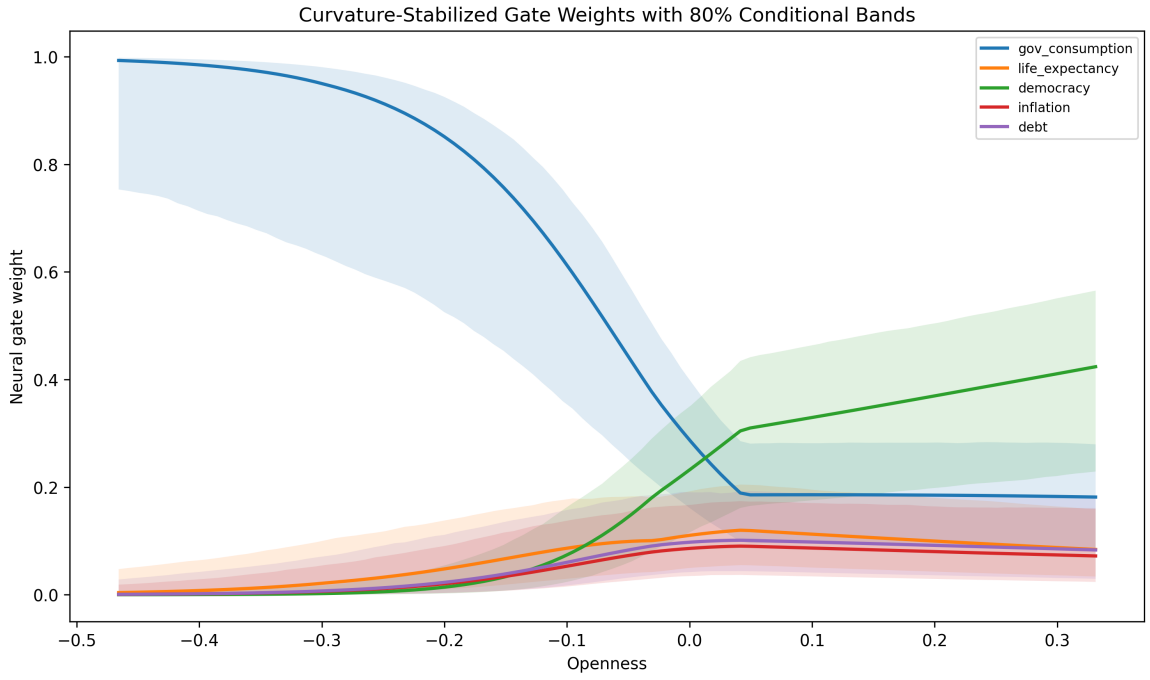

Figure 3: Curvature-Stabilized Neural-Gate Weights with 80% Conditional Bands

Table 12: Curvature-Stabilized Conditional Band Widths

Candidate Mechanism	80% Band	90% Band	95% Band
Debt	0.116	0.170	0.237
Democracy	0.165	0.210	0.255
Gov Consumption	0.278	0.369	0.455
Inflation	0.101	0.149	0.199
Life Expectancy	0.121	0.172	0.230

These bands should be interpreted as conditional uncertainty around the learned gate geometry, given the fitted candidate library and neural architecture. They are not a substitute for the wider sample-composition bands, but they are the more appropriate display when the empirical question is whether openness organizes the shape of the model-averaging rule.

5.10 Restricted Three-State Empirical Specification

Motivated by the desire to keep the empirical model as small and interpretable as possible, we estimate a restricted specification in which every candidate threshold regression uses the same baseline controls: initial GDP per capita, investment, population growth, and human capital. The threshold library is restricted to three economically central state variables: public debt, openness, and democracy. The model-averaging gates are also allowed to depend only on these three state variables.

Table 13: Restricted Three-State Cross-Fitted Performance

Method	Cross-Fitted MSE	Cross-Fitted RMSE
Linear Gate	0.0815	0.2856
Neural Gate	0.0816	0.2857
Fixed LS Average	0.0817	0.2858
Entropy Neural Gate	0.0817	0.2859
Naive Average	0.0817	0.2859
Best Single	0.0821	0.2866

In this restricted specification, the neural gate has cross-fitted MSE 0.0816, compared with 0.0817 for the fixed least-squares average and 0.0821 for the best single threshold model. Thus the compact three-state version remains competitive with the larger empirical system while using a much smaller and easier-to-interpret candidate library.

Table 14: Restricted Three-State Predictive-Superiority Tests

Benchmark	Loss Difference	DM Statistic	One-Sided p
Best Single	0.0005	0.770	0.221
Naive Average	0.0001	0.192	0.424
Fixed LS Average	0.0001	0.181	0.428
Linear Gate	-0.0001	-0.331	0.630
Entropy Neural Gate	0.0001	0.172	0.432

The predictive-superiority test against the fixed least-squares average has one-sided p-value 0.428. As in the larger specification, this should be interpreted as a predictive robustness diagnostic rather than as a claim of large forecast dominance.

Table 15: Restricted Three-State Average Weights

Threshold Mechanism	Fixed LS	Linear Gate	Neural Gate
Debt	0.415	0.387	0.241
Openness	0.000	0.053	0.147
Democracy	0.585	0.560	0.612

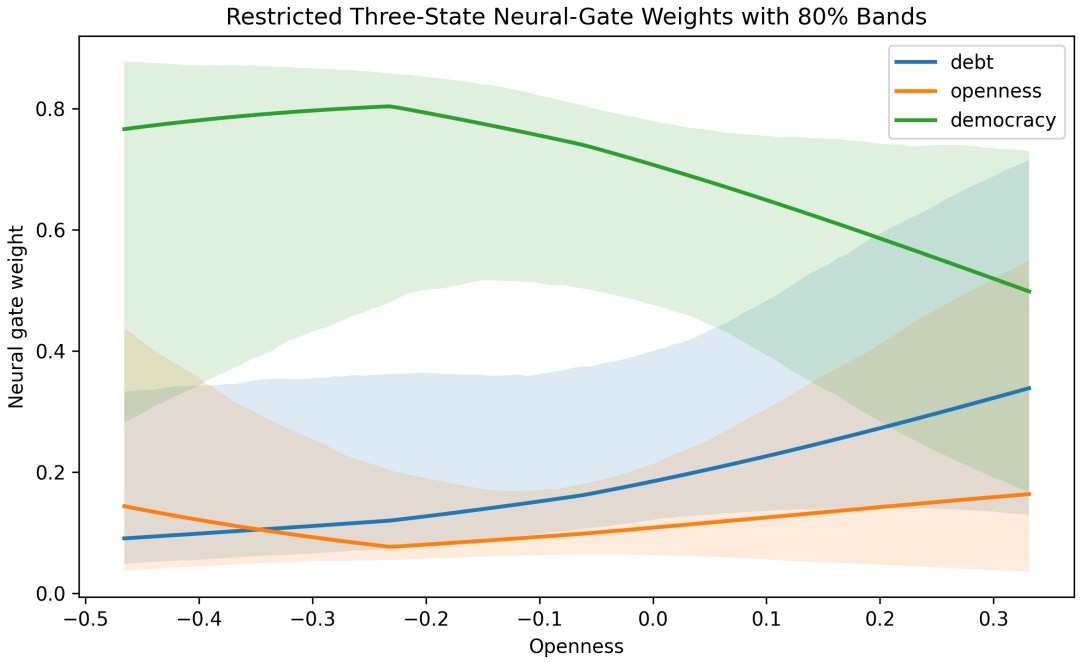

Figure 4: Restricted Three-State Neural-Gate Weights over Openness with 80% Subsampling Bands

Table 16: Restricted Three-State Average Subsampling Band Widths

Threshold Mechanism	80% Band	90% Band
Debt	0.339	0.437
Democracy	0.414	0.520
Openness	0.249	0.320

The country-level subsampling bands are computed by drawing 80% of countries without replacement and refitting the restricted three-state system. The average central 80% band widths are 0.339 for debt, 0.249 for openness, and 0.414 for democracy. These bands summarize the sensitivity of the compact allocation rule to country composition.

Overall, the restricted specification is implementable and empirically useful. It provides a tractable version of the regime-dependent averaging exercise centered on debt, openness, and an institutional state variable. If the paper emphasizes interpretability, this compact model is a natural robustness specification and may also be a candidate preferred empirical specification, depending on how much weight is placed on parsimony versus the slightly richer patterns available in the full seven-state library.

5.11 Homogeneous-Sample Robustness

The cross-fitted empirical results in the full panel suggest that the neural gate has the lowest average held-out MSE, but that its predictive advantage is modest relative to sampling uncertainty. Since a longer sample is not available, we also examine whether the results become clearer when the cross section is restricted to more homogeneous sets of countries. All restrictions are defined at the country level, so selected countries remain in the sample for all available five-year periods.

Table 17: Homogeneous-Sample Robustness: Sample Definitions

Sample	Countries	Observations	Definition
Full sample	71	639	All countries in the balanced empirical panel.
Central state countries, 5–95	51	459	Countries whose median inflation, debt, and openness are all between the cross-country 5th and 95th percentiles.
Central state countries, 10–90	38	342	Countries whose median inflation, debt, and openness are all between the cross-country 10th and 90th percentiles.
Moderate inflation/debt countries	41	369	Countries with median inflation and median debt no larger than the cross-country 75th percentiles.
High life expectancy/democracy countries	28	252	Countries with median life expectancy and median democracy no lower than the cross-country medians.

Table 18 reports the leave-one-period-out cross-fitted MSEs for the full sample and for the homogeneous country samples. The central-state restrictions remove countries with unusually high or low median inflation, debt, or openness. The moderate-inflation/debt sample removes countries with high median inflation or high median debt. The high-life-expectancy/democracy sample focuses on countries with stronger institutional and development indicators.

Table 18: Cross-Fitted Performance in Homogeneous Country Samples

Sample	Obs.	Neural MSE	Fixed-LS MSE	Linear-Gate MSE	Best-Single MSE
Full sample	639	0.0809	0.0809	0.0823	0.0828
Central state countries, 5–95	459	0.0893	0.0920	0.0927	0.1026
Central state countries, 10–90	342	0.0537	0.0540	0.0543	0.0539
Moderate inflation/debt countries	369	0.0487	0.0484	0.0483	0.0500
High life expectancy/democracy countries	252	0.0336	0.0317	0.0325	0.0312

The results are mixed in a useful way. The central-state 5–95 sample gives the strongest support for the neural gate: its cross-fitted MSE is 0.0893, compared with 0.0920 for the fixed least-squares average, 0.0927 for the linear gate, and 0.1026 for the best single model. In this sample, removing countries with the most unusual median inflation, debt, or openness therefore sharpens the adaptive-averaging result. The tighter central-state 10–90 restriction still ranks the neural gate first, with MSE 0.0537, but the margins relative to the fixed average and the best single model become very small. This suggests that some trimming helps, while very aggressive trimming also removes useful identifying variation.

The other two homogeneous samples are less favorable to the neural gate. In the moderate-inflation/debt sample, the linear gate and the fixed least-squares average have slightly lower cross-fitted MSEs than the neural gate. In the high-life-expectancy/democracy sample, the best single model and the fixed average outperform the neural gate. These results indicate that the nonlinear adaptive gains are not uniform across all relatively homogeneous subsamples.

Table 19 reports predictive-superiority tests comparing the neural gate with the main benchmarks within each restricted sample. These tests should be read as robustness diagnostics. Restricting the cross section can reduce heterogeneity, but it also reduces the number of countries and therefore does not mechanically increase precision.

Table 19: Cross-Fitted Predictive-Superiority Tests in Homogeneous Country Samples

Sample	Benchmark	Loss Diff.	DM Stat.	One-Sided p
Full sample	Fixed LS Average	0.0001	0.049	0.481
Full sample	Linear Gate	0.0014	1.318	0.094
Full sample	Best Single	0.0019	1.129	0.129
Central state countries, 5–95	Fixed LS Average	0.0028	2.804	0.003
Central state countries, 5–95	Linear Gate	0.0034	2.357	0.009
Central state countries, 5–95	Best Single	0.0133	3.684	0.000
Central state countries, 10–90	Fixed LS Average	0.0003	0.332	0.370
Central state countries, 10–90	Linear Gate	0.0006	0.692	0.245
Central state countries, 10–90	Best Single	0.0002	0.227	0.410
Moderate inflation/debt countries	Fixed LS Average	-0.0004	-0.534	0.703
Moderate inflation/debt countries	Linear Gate	-0.0004	-0.477	0.683
Moderate inflation/debt countries	Best Single	0.0013	1.089	0.138
High life expectancy/democracy countries	Fixed LS Average	-0.0019	-1.275	0.899
High life expectancy/democracy countries	Linear Gate	-0.0012	-1.366	0.914
High life expectancy/democracy countries	Best Single	-0.0024	-1.190	0.883

The predictive-superiority tests reinforce this interpretation. In the central-state 5–95 sample, the neural gate significantly improves on the fixed least-squares average, the linear gate, and the best single model, with one-sided p-values of approximately 0.003, 0.009, and 0.0001, respectively. In the tighter central-state 10–90 sample, the corresponding p-values are much larger, and in the moderate-inflation/debt and high-life-expectancy/democracy samples the neural gate is not the best cross-fitted procedure.

Overall, these checks suggest that sample homogeneity can help, but only for a particular kind of restriction. Excluding countries with the most extreme state-variable profiles makes the neural-gate evidence much stronger. By contrast, focusing only on low-inflation/low-debt or high-development and high-democracy countries does not strengthen the nonlinear gate and can favor simpler averaging rules. The empirical conclusion should therefore be stated carefully: the adaptive regime-dependent averaging result is not merely driven by the most extreme countries, but it is strongest in a moderately trimmed middle sample rather than in every homogeneous

subsample.

6 Additional Monte Carlo Design and Results

This section provides the complete specification of the Monte Carlo designs, additional state-selection diagnostics, and uncertainty analysis for the learned model-weight functions. The main paper reports only the principal design features and performance comparisons. All reported results are based on $R = 5000$ Monte Carlo replications.

6.1 Empirical-Design Sampling and Innovation Calibration

Let

$$W_t = (X_t, S_t),$$

where

$$X_t = (\mathrm{GDP0_t}, \mathrm{POP_t}, \mathrm{INV_t}, \mathrm{HC_t})$$

contains the baseline growth controls and

$$S_t = (\mathrm{inflation_t}, \mathrm{lifeexp_t}, \mathrm{fertility_t}, \mathrm{gov_t}, \mathrm{debt_t}, \mathrm{democracy_t}, \mathrm{openness_t})$$

contains the candidate regime coordinates.

Let

$$\widehat{F}_W = \frac{1}{n} \sum_{t=1}^{n} \delta_{W_t}$$

denote the empirical distribution of the observed macroeconomic sample. In each replication, simulated covariate-state observations are generated as

$$W_t^{MC} \sim \widehat{F}_W.$$

The empirical-design resampling strategy preserves the observed joint support and dependence among the predictors and state variables. In particular, it retains empirical correlations between candidate state coordinates, which is important for evaluating the distinction between exact state recovery and recovery of approximately equivalent predictive geometries.

Conditional on the resampled covariates,

$$Y_t^{MC} = f_0(X_t^{MC}, S_t^{MC}) + u_t^{MC}.$$

Let

$$\widehat{u}_t = Y_t - \widehat{f}(X_t, S_t)$$

denote residuals from the preliminary empirical growth specification. The innovations are obtained by resampling centered empirical residuals,

$$u_t^{MC} \sim \widehat{F}_{\widehat{u}}.$$

This avoids imposing Gaussianity and allows the simulated outcomes to inherit empirically relevant asymmetry and tail behavior.

6.2 Complete Specification of Experiment 1

For each candidate threshold variable q_m, define

$$(q_m - \gamma_m)_- = (q_m - \gamma_m)\mathbf{1}(q_m \leq \gamma_m), \qquad (q_m - \gamma_m)_+ = (q_m - \gamma_m)\mathbf{1}(q_m > \gamma_m).$$

The threshold values γ_m are fixed at the corresponding estimates obtained in the empirical application.

Each mechanism-specific component is

$$f_m(X_t, q_{mt}) = \mathrm{base_t} + \delta_{\mathrm{m,L}}(\mathrm{q_{mt}} - \gamma_{\mathrm{m}})_- + \delta_{\mathrm{m,H}}(\mathrm{q_{mt}} - \gamma_{\mathrm{m}})_+,$$

where

$$\mathrm{base_t} = \alpha_0 + \mathrm{X_t^\top}\alpha.$$

The kink slopes are calibrated so that the seven mechanisms exhibit different local nonlinear behavior. Inflation and debt generate comparatively strong asymmetries, whereas the demographic and openness mechanisms vary more smoothly.

The true conditional mean is

$$f_0(X_t, S_t) = \sum_{m=1}^{7} \pi_m(S_t) f_m(X_t, q_{mt}),$$

where

$$\pi_m(S_t) = \frac{\exp(r_{mt})}{\sum_{j=1}^{7} \exp(r_{jt})}.$$

The dominant allocation scores are

$$r_{1t} = 1.8\,\text{inflation}_\text{t}, \qquad \text{r}_{5\text{t}} = 1.8\,\text{debt}_\text{t}, \qquad \text{r}_{6\text{t}} = 1.8\,\text{democracy}_\text{t},$$

while the remaining mechanisms receive weaker baseline scores. Thus

$$A_0 = \{\text{inflation},\ \text{debt},\ \text{democracy}\}$$

is the principal sparse state representation generating the true predictive geometry.

The econometrician does not observe the true weights. In every replication, the seven single-threshold candidate models are re-estimated and their predictions are combined using each competing averaging procedure.

6.3 Complete Experiment 1 Results

For completeness, Table 20 reproduces the full performance comparison.

Table 20: Experiment 1: Adaptive-Mixture Design

Method	MSE_Y	MSE_f	Gain vs Fixed LS	Gain vs Best Single
Best Single	1.013	0.940	17.1%	0.0%
Naive Average	2.762	2.687	-137.0%	-186.8%
Fixed LS Average	1.231	1.157	0.0%	-23.5%
Linear Full Gate	0.782	0.708	36.8%	24.2%
Neural Full Gate	0.130	0.066	94.2%	92.9%
Entropy Neural Gate	0.164	0.098	91.4%	89.5%
Sparse Forward Neural Gate	0.156	0.085	92.6%	91.0%
Oracle-State Neural Gate	0.144	0.073	93.6%	92.2%

The full neural gate delivers the smallest approximation error among feasible estimators, while the sparse forward neural gate remains close to both the full-state and oracle-state neural procedures. The difference between MSE_Y and MSE_f reflects the irreducible innovation component present in the simulated outcome.

Table 21 reports state-selection frequencies.

Table 21: Experiment 1: State-Selection Frequencies

State Variable	Selection Frequency	True State Variable?
Inflation	0.921	Yes
Debt	0.358	Yes
Democracy	0.317	Yes
Openness	0.344	No
Government Consumption	0.330	No
Fertility	0.251	No
Life Expectancy	0.044	No

Inflation is selected in 92.1% of replications. Debt and democracy, though belonging to the true organizing set, are selected less frequently. Openness, government consumption, and fertility are also selected with non-negligible frequency.

This pattern reflects the empirical geometry inherited from the resampled macroeconomic covariates. Correlated coordinates may generate approximately equivalent partitions and hence nearly indistinguishable adaptive predictors. Selection frequencies should therefore not be interpreted as posterior probabilities that a state variable belongs to a uniquely identified structural support.

6.4 Experiment 1: Weight-Function Diagnostics

Figure 5 reports across-replication pointwise 95% bands for the full neural gate over an openness grid, with the remaining state variables fixed at their empirical medians.

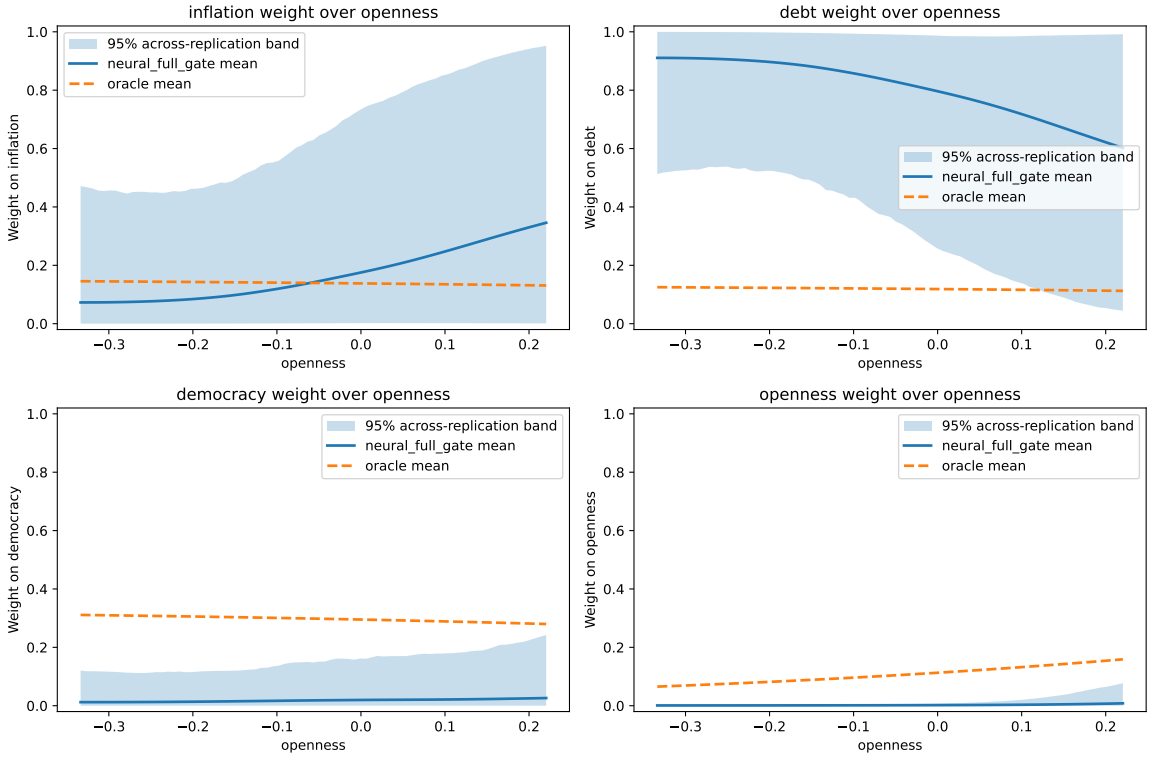

Figure 5: Experiment 1: Neural-Gate Weight Bands

Over this one-dimensional slice, the full neural gate places most average weight on the debt and inflation mechanisms. The mean weight on debt is approximately 0.807, while the mean inflation weight is approximately 0.166. The corresponding average band widths, approximately 0.662 and 0.658, indicate substantial uncertainty about the precise decomposition of the adaptive predictor across these locally substitutable mechanisms.

Figure 6 gives the corresponding result for the sparse forward neural gate.

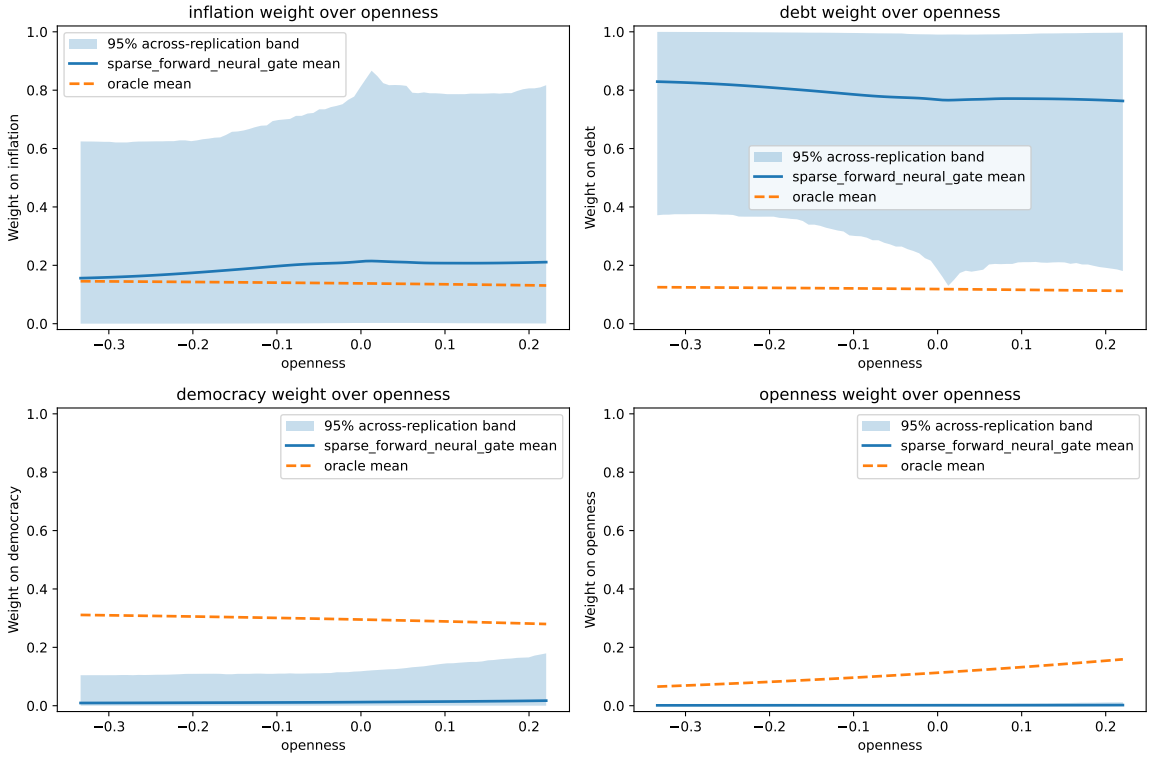

Figure 6: Experiment 1: Sparse Neural-Gate Weight Bands

The sparse gate exhibits a similar qualitative allocation, with average weights of approximately 0.788 on debt and 0.193 on inflation over the openness slice. The wide bands contrast with the very stable performance ranking in the MSE comparisons. This illustrates the distinction between prediction of the adaptive surface and identification of a unique mechanism decomposition–see also Section 7.

6.5 Complete Specification of Experiment 2

Experiment 2 augments the adaptive-mixture DGP with an additional mechanism that is omitted from the econometrician's candidate library. The omitted component is

$$f_8^0(X_t, S_t) = \text{base}_t + \eta_1(\text{debt}_t - \gamma_d)_+(\text{openness}_t - \gamma_o)_+ + \eta_2(\text{inflation}_t - \gamma_\pi)_+(\text{debt}_t - \gamma_d)_+.$$

The true conditional mean is

$$f_0(X_t, S_t) = \sum_{m=1}^{7} \pi_m^0(S_t) f_m(X_t, q_{mt}) + \pi_8^0(S_t) f_8^0(X_t, S_t),$$

while the estimated model library remains restricted to the original seven single-threshold mechanisms.

The omitted component is especially active in high-debt/high-openness and high-inflation/high-debt regions. Consequently, the true DGP cannot be represented exactly by any estimator based on the observed seven-model library. The relevant population target is therefore the best adaptive projection onto the available state-dependent convex hull.

6.6 Complete Experiment 2 Results

Table 22 reports the complete performance comparison.

Table 22: Experiment 2: Adversarial Misspecification Design

Method	MSE_Y	MSE_f	Gain vs Fixed LS	Gain vs Best Single
Best Single	1.495	1.421	33.2%	0.0%
Naive Average	4.554	4.481	-106.9%	-219.8%
Fixed LS Average	2.267	2.193	0.0%	-55.1%
Linear Full Gate	2.941	2.868	-30.3%	-98.4%
Neural Full Gate	0.314	0.251	88.3%	82.7%
Entropy Neural Gate	0.344	0.280	86.9%	80.6%
Sparse Forward Neural Gate	0.376	0.306	85.8%	78.9%
Oracle-State Neural Gate	0.326	0.259	87.9%	82.2%

The full neural gate attains mean $\text{MSE}_f = 0.2513$, followed closely by the oracle-state neural gate at 0.2588 and the entropy-regularized neural gate at 0.2797. The sparse forward neural gate achieves $\text{MSE}_f = 0.3056$. Although approximation errors are necessarily larger than in Experiment 1, nonlinear adaptive weighting continues to dominate conventional fixed averaging.

Table 23 reports the corresponding state-selection frequencies.

Table 23: Experiment 2: State-Selection Frequencies

State Variable	Selection Frequency	True State Variable?
Inflation	0.841	Yes
Openness	0.393	Yes
Democracy	0.318	Yes
Debt	0.304	Yes
Government Consumption	0.236	No
Fertility	0.188	No
Life Expectancy	0.050	No

Inflation remains the most frequently selected coordinate, while openness, democracy, and debt become particularly relevant. The greater role of openness is consistent with the omitted interaction mechanism, which becomes active in high-debt/high-openness regions. Forward selection therefore tends to choose observable coordinates that proxy for the missing regime structure.

6.7 Experiment 2: Weight-Function Diagnostics

Figure 7 reports the full neural-gate weight bands.

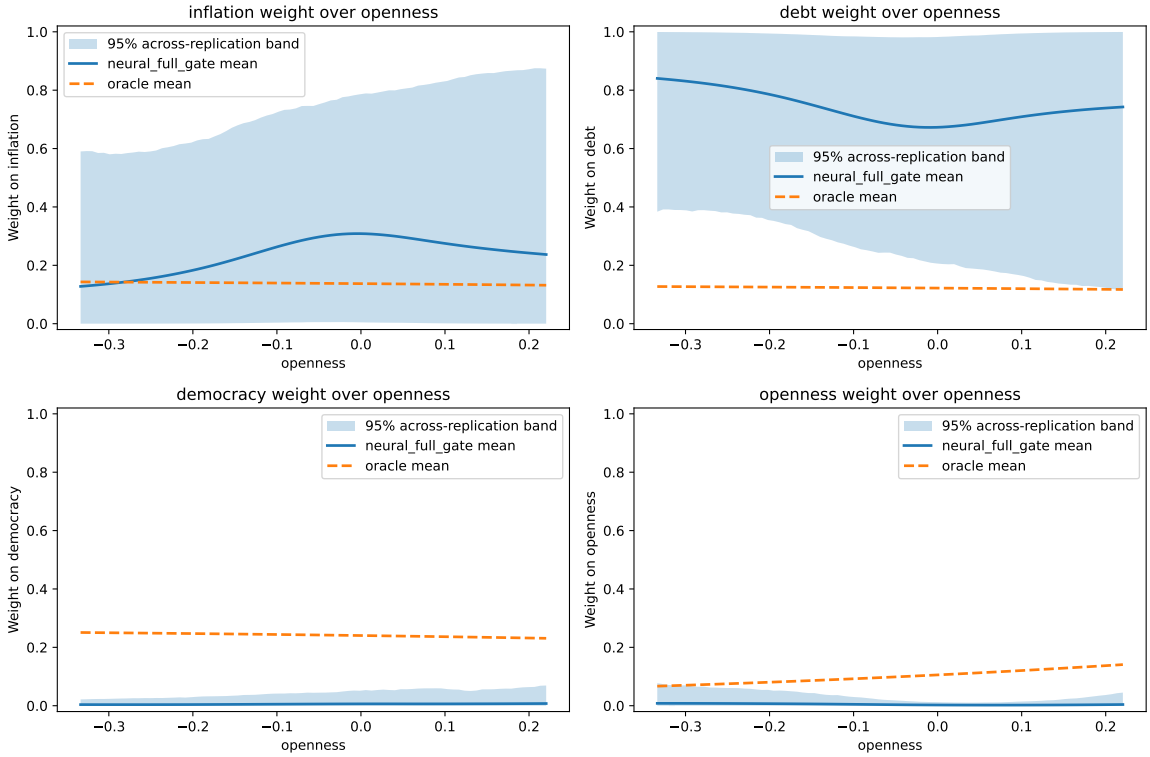

Figure 7: Experiment 2: Neural-Gate Weight Bands

Over the openness slice, the neural gate concentrates average weight on the debt and inflation candidate mechanisms, with mean weights of approximately 0.739 and 0.237, respectively. The corresponding average band widths are approximately 0.739 and 0.731.

The increased uncertainty relative to the already-wide Experiment 1 bands is consistent with mechanism misspecification. Since the omitted vulnerability component is unavailable to the econometrician, the adaptive gate must approximate it through different combinations of the observed threshold mechanisms. Several such combinations can produce similar fitted conditional means.

Figure 8 reports the sparse-gate counterpart.

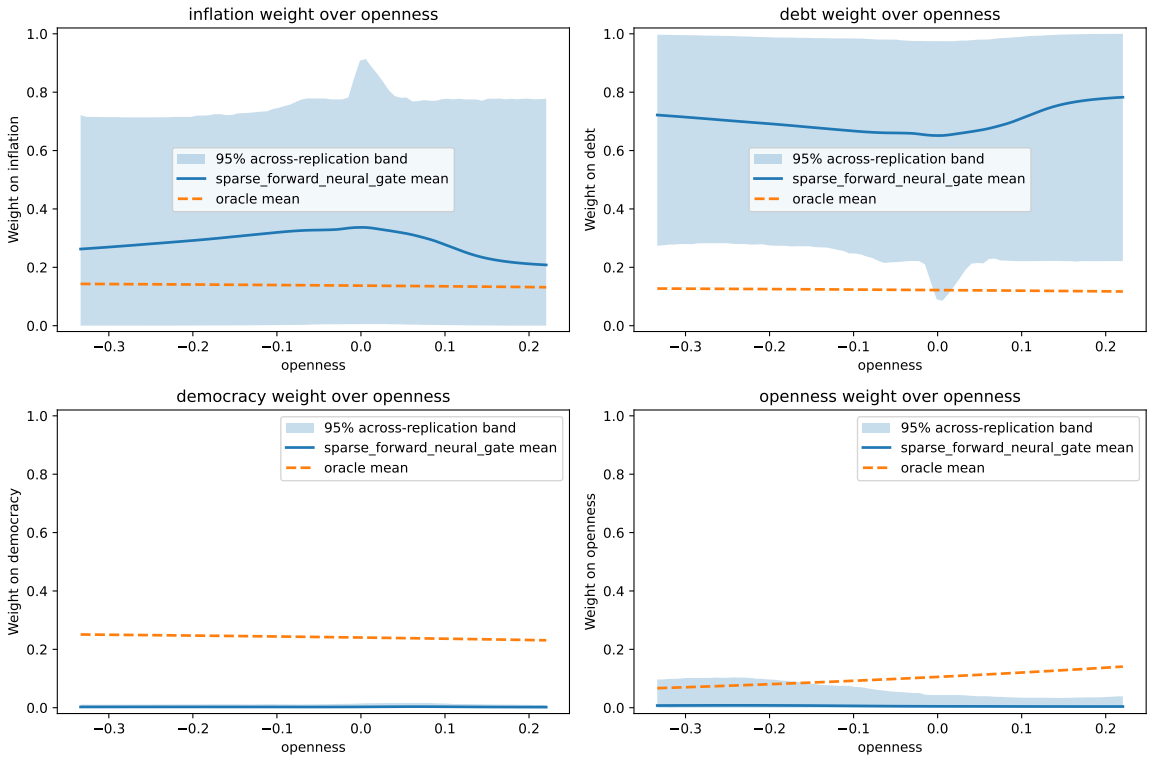

Figure 8: Experiment 2: Sparse Neural-Gate Weight Bands

The sparse gate assigns average weights of approximately 0.699 to debt and 0.288 to inflation over the same slice. Again, the bands are wide despite the stable predictive ranking of the adaptive estimators.

The omitted oracle component itself is considerably more stable over this one-dimensional slice, with mean weight approximately 0.077. The contrast illustrates how an incomplete model library induces uncertainty in the decomposition across available mechanisms even when the resulting adaptive projection remains predictively effective.

6.8 Monte Carlo Interpretation

Taken together, the additional diagnostics sharpen the interpretation of the main Monte Carlo evidence. The nonlinear gates recover highly accurate adaptive approximations in Experiment 1, while sparse state selection retains most of the full-state gains. Under adversarial misspecification, approximation error increases but nonlinear adaptive weighting remains substantially more effective than fixed aggregation.

At the same time, neither exact support recovery nor precise point identification of every

model weight is necessary for strong approximation performance. Correlated state variables can generate approximately equivalent regime representations, and locally substitutable candidate models can generate different weight decompositions with similar predictive implications. The Monte Carlo evidence therefore supports the set-valued pseudo-true interpretation underlying the theoretical analysis: the procedure learns a low-risk predictive geometry, while the coordinate representation and precise mechanism decomposition may remain only partially identified.

7 Set-Valued Identification and Bootstrap-Uncertainty Diagnostics

The set-consistency results in Section 3.8 allow the sparse population target to contain several observationally equivalent or nearly equivalent state–gate representations. This possibility has a direct implication for resampling inference. A bootstrap or subsampling replication may converge to a different element of the pseudo-true set, or to a different approximately optimal forward-selection path, even when the induced adaptive predictions are similar. Pointwise confidence strips for the gate can consequently be wide without implying equally large uncertainty about the predictive image or out-of-sample loss. This subsection reports diagnostics designed to distinguish that mechanism from ordinary sampling variation and imperfect neural optimization.

The diagnostics are descriptive finite-sample evidence. In particular, the local-information quantities reported below are not direct estimators of the abstract restricted-Hessian convergence rate used in Appendix D. Likewise, support switching across resamples is not by itself evidence of population non-identification. The set-valued interpretation is supported only by the joint behavior of representation, prediction, optimization, and curvature diagnostics.

7.0.1 Diagnostic design

For replication b, let $\widehat{\pi}^{(b)} = (\widehat{j}_1^{(b)}, \ldots, \widehat{j}_{K_b}^{(b)})$ denote the ordered forward-selection path, let $\widehat{A}^{(b)} = \{\widehat{j}_1^{(b)}, \ldots, \widehat{j}_{K_b}^{(b)}\}$ denote its unordered support, and let $\widehat{w}^{(b)}$ denote the subsequently fitted neural gate. At greedy step k, write $\widehat{\Delta}_{(1),k}^{(b)}$ and $\widehat{\Delta}_{(2),k}^{(b)}$ for the largest and second-largest empirical

marginal gains. We record the near-tie gap

$$\widehat{g}_k^{(b)} = \widehat{\Delta}_{(1),k}^{(b)} - \widehat{\Delta}_{(2),k}^{(b)}. \tag{100}$$

Small values of $\widehat{g}_k^{(b)}$ indicate that modest empirical perturbations can change the selected path. We therefore report support and ordered-path frequencies together with the minimum and average near-tie gaps.

All fitted gates are evaluated on a common reference design. For two replications b and b', we calculate the gate distances

$$d_{\mathcal{W},\infty}(b,b') = \sup_{s\in\mathcal{S}_0} \left\| \widehat{w}^{(b)}(s) - \widehat{w}^{(b')}(s) \right\|_\infty, \tag{101}$$

$$d_{\mathcal{W},2}(b,b') = \left\{ \frac{1}{|\mathcal{S}_0|} \sum_{s\in\mathcal{S}_0} \left\| \widehat{w}^{(b)}(s) - \widehat{w}^{(b')}(s) \right\|_2^2 \right\}^{1/2}. \tag{102}$$

Let $\widehat{f}^{(b)}$ denote the adaptive prediction evaluated on the same reference observations. Prediction stability is summarized by

$$d_f(b,b') = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{f}_i^{(b)} - \widehat{f}_i^{(b')} \right)^2 \right\}^{1/2}, \tag{103}$$

and by the Hausdorff distance between the corresponding finite predictive images. The Jaccard distance between selected supports and an indicator for equality of the ordered paths complete the pairwise diagnostics. Large gate or support distances accompanied by small $d_f(b,b')$, small predictive-image distance, and limited loss deterioration are the pattern predicted by representation-level set switching.

For each model m and grid point s, we also decompose resampling variation by the law of total variance. With C_b denoting either the selected support or ordered path,

$$\mathrm{Var}_b\{\widehat{w}_m^{(b)}(s)\} = \mathbb{E}_b\big[\mathrm{Var}_b\{\widehat{w}_m^{(b)}(s) \mid C_b\}\big] + \mathrm{Var}_b\big[\mathbb{E}_b\{\widehat{w}_m^{(b)}(s) \mid C_b\}\big]. \tag{104}$$

The second term measures the share of gate uncertainty associated with switching between supports or paths; it need not measure causal variation generated by the selection procedure. In

computing support frequencies and support-conditioned variances, we canonicalize the selected variables as an unordered set. This is essential because different ordered greedy paths can terminate at the same support.

7.0.2 Fixed-support and optimization counterfactuals

To separate support switching from neural optimization, the country-level subsampling exercise is repeated under three designs:

1. the support is reselected within every subsample and the sparse neural gate is refitted;
2. the support selected in the original sample is held fixed and the neural gate is estimated from one initialization; and
3. the same original-sample support is held fixed and the gate is selected from multiple independent neural initializations according to the attained criterion value.

The first-to-second comparison removes support and path switching while retaining within-support sampling and optimization variation. The second-to-third comparison isolates the effect of improved neural optimization. A marked reduction in band width after fixing the support, followed by a comparatively small change from additional restarts, is consistent with selection among nearly equivalent representations rather than optimization being the dominant source of uncertainty. Conversely, a large single-to-multiple-restart reduction would attribute an important portion of the original width to neural optimization.

For every fitted neural gate, we record the retained objective, gradient norm, best epoch, stopping reason, and dispersion of objective values and fitted gates across restarts. We additionally compute the smallest and largest eigenvalues and condition number of a local Gauss–Newton parameter-information matrix. This matrix is a finite-dimensional diagnostic for local flatness. Weak local information can amplify sampling and optimization perturbations, but it should not be identified with the restricted population curvature condition in Appendix D.

7.0.3 Empirical evidence

Table 24 reports the empirical support and path diagnostics. Table 25 decomposes gate variation into its within- and between-representation components. Table 26 compares the free-selection and

fixed-support confidence-strip widths. Every reported empirical diagnostic uses 500 successful replications.

Table 24: Empirical forward-selection and near-tie diagnostics

Resampling design	Supports	Paths	Empty	Median gap	10% gap
Country subsampling	38	58	21.4%	1.87×10^{-5}	2.19×10^{-6}
Cluster bootstrap	40	67	16.4%	2.03×10^{-5}	2.52×10^{-6}

Notes: Supports are treated as unordered sets, while paths preserve selection order. "Empty" is the frequency with which the stopping rule selects no state variable. The gain gap is the difference between the largest and second-largest profiled marginal gains; the final two columns summarize the minimum gap within a replication. The mean minimum gap is 4.78×10^{-5} under subsampling and 5.74×10^{-5} under the bootstrap.

Table 25: Within- and between-representation decomposition of empirical gate variation

Gate	Conditioning variable	Clusters	Mean share	Median share
Full-state neural gate	Unordered support	38	8.4%	7.8%
Full-state neural gate	Ordered path	58	12.4%	12.3%
Sparse neural gate	Unordered support	38	30.4%	31.7%
Sparse neural gate	Ordered path	58	34.5%	35.3%

Notes: Shares are computed pointwise and then summarized over the seven candidate-model weights and the 100-point openness grid. The support rows were recomputed after sorting each support; the generated path rows retain selection order. Because some clusters are small, these are descriptive variance allocations rather than bias-corrected variance-component estimates.

Table 26: Fixed-support and neural-optimization counterfactuals

Design	Mean 80% width	Reduction	Mean 90% width	Reduction
Reselected support, multiple starts	0.231	—	0.294	—
Fixed openness support, one start	0.198	14.3%	0.281	4.6%
Fixed openness support, multiple starts	0.196	15.0%	0.279	5.2%

Notes: Widths are averaged over the seven candidate-model weights and the 100-point openness grid. Reductions are relative to the reselected sparse gate. Moving from one to multiple starts on the fixed support reduces mean width by only 0.8% for the 80% strip and 0.6% for the 90% strip.

The results support a qualified, rather than exclusive, set-identification interpretation. Support and path switching are extensive and the near-tie gaps are small relative to the stopping threshold 10^{-4}. For the sparse gate, unordered-support switching accounts descriptively for about 30% of pointwise gate variance, rising to about 35% when path order is retained. Holding the original-sample openness support fixed narrows the central 80% strip by about 14–15%, whereas additional neural restarts change its width by less than 1%. Hence neural optimization is not the principal explanation for the average strip width. The smaller reduction in the 90%

strip indicates that support switching does not account for the tail behavior of every gate coordinate; within-support sampling variation remains quantitatively important.

The bootstrap comparison reinforces this qualification. Across all 124,750 replication pairs, the mean gate L_2 distance is 0.236 and the mean prediction distance is 0.074. The corresponding means are 0.186 and 0.064 for pairs with the same unordered support, compared with 0.240 and 0.074 for pairs with different supports. Thus changing support is associated with substantially greater movement in the gate than in its induced prediction. Moreover, the Spearman correlation between support Jaccard distance and prediction distance is only 0.017. At the same time, the nonzero prediction distances and the substantial within-support gate dispersion rule out an interpretation in which all band width is due to exact observational equivalence.

Optimization diagnostics point in the same direction. For the sparse gate, the median maximum distance across restarts is 0.011 under subsampling, while the median maximum prediction distance is 3.25×10^{-4} and the median criterion gap is 8.94×10^{-8}. The full-state gate shows much larger between-restart gate movement (median maximum distance 0.942) but a median prediction distance of only 0.015. Nearly equal criterion values can therefore correspond to visibly different allocations, particularly in the less restricted gate.

Finally, cross-fitted predictive performance is essentially preserved. The neural gate has MSE 0.08088, compared with 0.08095 for fixed least-squares averaging, but the one-sided predictive-superiority test against the fixed benchmark has $p = 0.476$. The empirical value of the adaptive gate therefore lies primarily in refining the state-dependent representation without a detectable loss of predictive ability, rather than in establishing a large out-of-sample MSE improvement.

7.0.4 Monte Carlo validation

The Monte Carlo experiments apply the same diagnostics to designs in which the data-generating predictive mechanism is known. Replication-level weight and prediction curves are evaluated on a common state grid. We report support and path frequencies, the decomposition in (104), neural convergence diagnostics, and the relation between representation distances and prediction error. This exercise evaluates whether the proposed empirical signatures correctly distinguish unstable representation from unstable prediction under controlled designs.

The simulations reproduce the distinction between representation and prediction stability.

Table 27: Monte Carlo diagnostics for representation and prediction stability

Design	Supports	Paths	Weight share	Image share	MSE(f_0)	Gain vs. fixed
Experiment 1	43	65	25.2%	16.5%	0.080	93.0%
Experiment 2	44	81	25.9%	12.4%	0.300	86.2%

Notes: "Weight share" is the mean, over seven weights and 81 grid points, of the variance share between canonical unordered supports. "Image share" is the corresponding share for the scalar predictive image. The last column is the mean reduction in MSE(f_0) relative to fixed least-squares averaging. Each design uses 500 replications.

In Experiment 1, the sparse procedure visits 43 unordered supports and 65 ordered paths; in the more adversarial Experiment 2, it visits 44 supports and 81 paths. Between-support variation accounts for approximately one quarter of weight variation, but only 16.5% and 12.4% of predictive-image variation. Exact recovery of the designated state support is 8.6% in Experiment 1 and zero in Experiment 2, while mean recovery shares are 54.1% and 46.0%. Nevertheless, the sparse gate reduces MSE(f_0) relative to fixed averaging by 93.0% and 86.2%, respectively. It also remains close to the unrestricted neural gate: the latter has mean MSE(f_0) of 0.064 in Experiment 1 and 0.252 in Experiment 2, compared with 0.080 and 0.300 for the sparse gate. Thus failure to recover a unique representation can coexist with large predictive gains, including under mechanism misspecification.

The Monte Carlo optimization diagnostics are less benign than in the empirical subsamples. For the sparse gate, the median maximum prediction distance across restarts is 0.056 in Experiment 1 and 0.098 in Experiment 2. The corresponding median criterion dispersions remain small (0.0011 and 0.0033), but local condition numbers are large, especially in Experiment 2. These results are consistent with finite-iteration optimization and weak local curvature adding variation on top of path switching in difficult designs. Because the smallest reported local-information eigenvalue is mechanically bounded by the ridge term, we treat the condition number only as a comparative flatness diagnostic, not as an estimate of the theoretical h_n sequence.

7.0.5 Implications for bootstrap inference

The reported strips remain valid descriptive summaries of resampling variation, but the diagnostics reject a uniquely identified-gate interpretation. They are consistent with a mixture of three effects: switching between approximately equivalent supports and paths, sampling variation within a fixed support, and—particularly in the adversarial Monte Carlo design—weak-curvature

optimization variation. The fixed-support and restart counterfactuals indicate that the first effect is material but not dominant for every coordinate or tail probability. We therefore report gate uncertainty together with support/path frequencies and predictive-image stability, and interpret the bands as uncertainty about the learned representation rather than conventional confidence bands around a unique allocation surface.

References

Azariadis, C. and Drazen, A. (1990). Threshold externalities in economic development. *The Quarterly Journal of Economics*, 105(2):501–526.

Barro, R. J. and Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics*, 104:184–198.

Belkin, M., Niyogi, P., and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434.

Chamon, M. and Kremer, M. (2009). Economic transformation, population growth and the long-run world income distribution. *Journal of International Economics*, 79(1):20–30.

Czarnecki, W. M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. (2017). Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Das, A. and Kempe, D. (2011). Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1057–1064.

de la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20:134–144.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Elenberg, E. R., Khanna, R., Dimakis, A. G., and Negahban, S. (2018). Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568.

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The next generation of the penn world table. *American Economic Review*, 105(10):3150–3182.

Freedman, D. A. (1975). On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118.

Galor, O. (2005). From stagnation to growth: Unified growth theory. In Aghion, P. and Durlauf, S. N., editors, *Handbook of Economic Growth*, volume 1A, chapter 4, pages 171–293. Elsevier BV.

Galor, O. and Moav, O. (2002). Natural selection and the origin of economic growth. *The Quarterly Journal of Economics*, 117(4):1133–1191.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pages 1861–1870. PMLR.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition.

International Monetary Fund (2026). Global debt database (gdd) [dataset]. Accessed: August 25, 2026.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Mao, T. and Zhou, D.-X. (2023). Rates of approximation by ReLU shallow neural networks. *Journal of Complexity*, 79:101784.

Marshall, M. G. and Gurr, T. R. (2020). Polity5: Political regime characteristics and transitions, 1800–2018. annual time-series, 1946–2018 [dataset]. Accessed: August 25, 2026.

Mbaye, S., Moreno Badia, M., and Chae, K. (2018). Global debt database: Methodology and sources. IMF Working Paper WP/18/111, International Monetary Fund.

Merlevède, F., Peligrad, M., and Rio, E. (2011a). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3):435–474.

Merlevède, F., Peligrad, M., and Rio, E. (2011b). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3):435–474.

Molchanov, I. (2006). *Theory of Random Sets*. Springer.

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767.

Rao, M. M. and Ren, Z. D. (1991). *Theory of Orlicz Spaces*, volume 146 of *Pure and Applied Mathematics*. Marcel Dekker, Inc., New York.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational Analysis*. Springer.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press.

World Bank (2026). World development indicators. Accessed: August 25, 2026.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ SCHOOL OF ECONOMIC SCIENCES
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΕΠΙΣΤΗΜΗΣ DEPARTMENT OF ECONOMICS

Department of Economics
Athens University of Economics and Business

List of Recent Working Papers

2024

01-24 Market Timing & Predictive Complexity, Stelios Arvanitis, Foteini Kyriazi, Dimitrios Thomakos
02-24 Multi-Objective Frequentistic Model Averaging with an Application to Economic Growth, Stelios Arvanitis, Mehmet Pinar, Thanasis Stengos, Nikolas Topaloglou
03-24 State dependent fiscal multipliers in a Small Open Economy, Xiaoshan Chen, Jilei Huang, Petros Varthalitis
04-24 Public debt consolidation: Aggregate and distributional implications in a small open economy of the Euro Area, Eleftherios-Theodoros Roumpanis
05-24 Intangible investment during the Global Financial Crisis in the EU, Vassiliki Dimakopoulou, Stelios Sakkas and Petros Varthalitis
06-24 Time will tell! Towards the construction of instantaneous indicators of different agent-types, Iordanis Kalaitzoglou, Stelios Arvanitis
07-24 Norm Constrained Empirical Portfolio Optimization with Stochastic Dominance: Robust Optimization Non-Asymptotics, Stelios Arvanitis
08-24 Asymptotics of a QLR-type test for optimal predictive ability, Stelios Arvanitis
09-24 Strongly Equitable General Equilibrium Allocations, Angelos Angelopoulos
10-24 The Greek macroeconomy: A note on the current situation and future outlook, Apostolis Philippopoulos
11-24 Evolution of Greek Tax System, A Survey of Legislated Tax Changes from 1974 to 2018, Panagiotis Asimakopoulos
12-24 Macroeconomic Impact of Tax Changes, The case of Greece from 1974 to 2018, Panagiotis Asimakopoulos
13-24 `Pareto, Edgeworth, Walras, Shapley' Equivalence in a Small Economy, Angelos Angelopoulos
14-24 Stimulating long-term growth and welfare in the U.S, James Malley and Apostolis Philippopoulos
15-24 Distributionally Conservative Stochastic Dominance via Subsampling, Stelios Arvanitis
16-24 Before and After the Political Transition of 1974. Institutions, Politics, and the Economy of Post-War Greece, George Alogoskoufis
17-24 Gaussian Stochastic Volatility, Misspecified Volatility Filters and Indirect Inference Estimation, Stelios Arvanitis, Antonis Demos
18-24 Endogenous Realtor Intermediation and Housing Market Liquidity, Miroslav Gabrovski, Ioannis Kospentaris, Victor Ortego-Marti
19-24 Universal Choice Spaces and Expected Utility: A Banach-type Functorial Fixed Point, Stelios Arvanitis, Pantelis Argyropoulos, Spyros Vassilakis

2025

01-25 Democracy, redistribution, and economic growth: Some evidence from post-1974 Greece, George C. Bitros

02-25 Assessing Downside Public Debt Risks in an Environment of Negative Interest Rates Growth Differentials, Yiannis Dendramis, Georgios Dimitrakopoulos and Elias Tzavalis

03-25 Wicksellian General Equilibrium, Angelos Angelopoulos

04-25 Institutions and Ethics, Angelos Angelopoulos

05-25 General Equilibrium in a Capitalist Economy, Angelos Angelopoulos

06-25 Public investment multipliers revisited: The role of production complementarities, Vasiliki Dimakopoulou, George Economides, and Apostolis Philippopoulos

07-25 The heterogeneous causal effects of the EU's Cohesion Fund, Angelos Alexopoulos, Ilias Kostarakos, Christos Mylonakis, and Petros Varthalitis

08-25 Gaussian Invariant Markov Chain Monte Carlo, Michalis K. Titsias, Angelos Alexopoulos , Siran Liu, and Petros Dellaportas

09-25 Financial Intermediation, Investment, and Monetary Policy, Ioannis Kospentaris and Florian Madison

10-25 Temperature as a premium factor. Insurance incorporating climate change. A comparative case study between Europe and North America, Nomikou-Lazarou Irene

11-25 How High will our Buildings Become? Angelos Angelopoulos

12-25 Emergency Room General Equilibrium, Angelos Angelopoulos

13-25 Updating Density Estimates using Conditional Information Projection: Stock Index Returns and Stochastic Dominance Relations, Stelios Arvanitis, Richard McGee, and Thierry Post

14-25 Measuring Textual Cohesion via Graph Curvature, Weighted Dominance, and Embedding-Based Coherence, Stelios Arvanitis

15-25 Universal Preference Spaces for Expected Utility via a Banach-type Functorial Fixed Point, Stelios Arvanitis, Pantelis Argyropoulos, and Spyros Vassilakis

16-25 Planet Earth's General Equilibrium, Angelos Angelopoulos

17-25 Electable and Stable Insiders' Coalition Governments, Tryphon Kollintzas and Lambros Pechlivanos

18-25 Long Memory from Cheeger Bottlenecks and Long Cycles in Network Dynamics, Stelios Arvanitis

19-25 Robust Multiobjective Model Averaging in Predictive Regressions, Stelios Arvanitis, Mehmet Pinar, Thanasis Stengos, and Nikolas Topaloglou

20-25 On robust Bayesian causal inference, Angelos Alexopoulos and Nikolaos Demiris

21-25 Leverage Effect, Volatility Feedback and the Influence of Jumps, Orestis Agapitos, Ioannis Papantonis, Leonidas Rompolis, and Elias Tzavalis

22-25 Leadership: The other deficit of post-1974 Greece, George C. Bitros

2026

01-26 The Tell-Tale Clock! Speed and Agent Composition in High Frequency Trading, Stelios Arvanitis

02-26 Sentiment Utopia Index: Rhetorical Structure and Corpus-Relative Financial Sentiment, Stelios Arvanitis

03-26 Draghi's report a year later: Beneficial but no remedy for EU challenges, George C. Bitros, Anastasios G. Malliaris, and Mark S. Rzepczynski

04-26 **Tax Structure Reform and Macroeconomic Strategy in Greece: Econometric Evidence and Policy Implications, Panagiotis Asimakopoulos**

05-26 **A Perspective on International Monetary Leadership. From Silver and Gold to the Dollar Standard, George Alogoskoufis**

06-26 **Behavioral Balanced Scorecard Framework for Tax Administrations: Integrating Tax Morale, Institutional Trust, and Compliance Behavior, Panagiotis Asimakopoulos**

07-26 **Education and governance: The Russell–Hayek Conundrum through the lenses of ancient Sparta and the "Sparta of the East", George C. Bitros**

08-26 **Overhauling Democracy: Layered State Authority, Corporate-Style Accountability, and the Free Market Economy, George C. Bitros and Steve Bakalis**

09-26 **ECONOMICS 200. The Economics of transition to a Steady State Economy, Theodore P. Lianos**

10-26 **Empirical Likelihood Tests for Pairwise Stochastic Dominance, Stelios Arvanitis, Richard McGee, and Thierry Post**

11-26 **Production Function Estimation under Input Distortions: Identification Using Heteroscedasticity-Based Instruments, Dimitris K. Christopoulos and Elias Tzavalis**

12-26 **Learning Dimension-Reduced State-Adaptive Model Averaging in Regime-Dependent Environments, Stelios Arvanitis and Thanasis Stengos**

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
SCHOOL OF ECONOMIC SCIENCES

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF ECONOMICS

Department of Economics
Athens University of Economics and Business

The Department is the oldest Department of Economics in Greece with a pioneering role in organising postgraduate studies in Economics since 1978. Its priority has always been to bring together highly qualified academics and top quality students. Faculty members specialize in a wide range of topics in economics, with teaching and research experience in world-class universities and publications in top academic journals.

The Department constantly strives to maintain its high level of research and teaching standards. It covers a wide range of economic studies in micro-and macroeconomic analysis, banking and finance, public and monetary economics, international and rural economics, labour economics, industrial organization and strategy, economics of the environment and natural resources, economic history and relevant quantitative tools of mathematics, statistics and econometrics.

Its undergraduate program attracts high quality students who, after successful completion of their studies, have excellent prospects for employment in the private and public sector, including areas such as business, banking, finance and advisory services. Also, graduates of the program have solid foundations in economics and related tools and are regularly admitted to top graduate programs internationally. Three specializations are offered:1. Economic Theory and Policy, 2. Business Economics and Finance and 3. International and European Economics. The postgraduate programs of the Department (M.Sc and Ph.D) are highly regarded and attract a large number of quality candidates every year.

For more information:

https://www.dept.aueb.gr/en/econ/