A Constructive Orbit-Based Theory of Learnability

Yunied Puig puigdedios@gmail.com
Department of Mathematical Sciences, Lafayette College
Easton, PA 18042, USA

Editor:

Abstract

We propose a structural theory of learnability from an orbit-based operator-theoretic perspective in which learnability is organized by dataset-induced orbit geometry and sparse constructive realizability rather than by unrestricted parametrization alone. We develop this perspective in finite dimensions, while showing that its core constructive mechanisms already persist in infinite-dimensional Banach settings.

Keywords: operator-theoretic machine learning, orbit-based learning, weighted backward shifts, constructive realizability, sparse approximation

1 Introduction

Deep learning excels across vision, language, and time series, but much of its practical success still relies on stacked nonlinearities, heavy overparameterization, and architectural heuristics such as attention, which often obscure mechanism and demand large labeled datasets and careful tuning. In this work we ask a different question: can learning be organized around simple orbit dynamics, explicit geometry, and constructive realization rather than around increasingly opaque parametrization? We introduce an orbit-based framework in which scaffold geometry, realizability, and prediction are treated as three linked structural bottlenecks rather than as disconnected stages of a pipeline. The predictive instantiation of this framework, which we call a Weighted Backward Shift Neural Network (WBSNN), has a general-purpose design: the same orbit-based mechanism can be deployed across varied and demanding settings without requiring the pipeline to be tailored to each dataset. Moreover, although the backbone is linear, the model's predictions become nonlinear through their dependence on orbit evolution and data-dependent coefficient selection, yielding both interpretability and expressive capacity. Empirically, WBSNN often recovers the orbit scaffold from just 1% of the training data while matching or surpassing strong baselines, offering a data-efficient and structurally lightweight alternative to heavily overparameterized deep networks.

The philosophy is the following. Instead of taking the model as the primary object and asking it to discover structure implicitly, we first extract from the data a structured orbit scaffold encoding part of the dataset's intrinsic geometry. On that scaffold we study whether a target operator, or in the predictive setting a target dataset map, can be constructively realized with explicit and analyzable error. Prediction is then built on top of this constructive stage: rather than learning outputs freely, the predictor learns a coefficient law that tracks and extends the constructive rule found on the training set. This leads to a framework organized around three explicit bottlenecks: a *scaffold bottleneck*, measuring how well the selected orbit scaffold captures the relevant geometry of the data; a *realizability bottleneck*, measuring how well the projected target can be constructively synthesized on that scaffold; and a *predictive coefficient bottleneck*, measuring how well the learned predictor reproduces and generalizes the constructive coefficient rule.

The theoretical contribution of the paper has two tightly connected parts. On the realizability side, we prove exact interpolation theorems showing that orbit-generated realizations exist under explicit structural conditions; we derive a two-term decomposition separating projection and

parametrization errors; we develop an algorithm that searches for an optimal orbit-constrained scaffold; and we introduce CCDS, a constructive deviation-controlled synthesis mechanism that yields explicit certificates and recovers classical greedy pursuit as a boundary case. On the predictive side, we show that the same structural logic carries over to learning: the WBSNN predictor inherits the scaffold and realizability bottlenecks and adds only one new one, namely the learned approximation of the effective CCDS coefficient map. This yields in-sample structural prediction bounds and an out-of-sample transfer principle under geometric proximity assumptions between training and test points. In summary, the paper should be read as a single theory of learnability: Section 2 develops its constructive realizability backbone, and Section 3 shows that prediction is obtained by extending the same scaffold-realization logic rather than by introducing a separate model class.

Why should the machine-learning community care about such a framework? Because it shifts the focus of learnability away from opaque feature mixing and unrestricted parametrization, and toward a more explicit structural question: what geometry the dataset induces, what targets can be constructively realized on that geometry, and how prediction can be built by extending that constructive rule. In this sense, the framework is interpretable end-to-end: the scaffold is data-induced and geometric, the realization step is explicit and constructive, and prediction is constrained by orbit coefficients rather than delegated to a black-box parametrization. The constructive theory also yields two properties of direct practical value. First, CCDS supplies explicit sparsity certificates, and the experiments confirm that accurate prediction is supported by highly sparse effective realizations. Second, the theory shows that exact block interpolation, while canonical, is not required for robust prediction. The framework therefore justifies the cheaper, more robust under noisy local blocks and better-conditioned relaxed regularized interpolant as the practical default, with predictive performance governed by scaffold geometry invariance rather than machine-precision interpolation.

More fundamentally, the framework suggests that datasets should not be viewed merely as collections of samples to be fit, but as objects carrying intrinsic geometric structure that directly shapes what can be learned from them.

Our construction is rooted in the theory of linear dynamics, where the weighted backward shift operator plays a foundational role. In infinite-dimensional separable sequence Banach spaces, this operator acts on sequences $x = (x_0, x_1, x_2, \dots)$ by shifting coordinates backward and scaling by a bounded weight sequence $w = (w_i)_{i\geq 0}$:

$$B_w x = (w_1 x_1, w_2 x_2, w_3 x_3, \dots). \tag{1}$$

The orbit of a point x under B_w is the sequence $\mathrm{Orb}(x, B_w) = \{B_w^m x : m \geq 0\}$, that is, the collection of all iterates generated by repeatedly applying the shift. This simple mechanism gives rise to rich dynamical behavior, including dense orbits and hypercyclicity, and is therefore a natural source of structured representations. To adapt this dynamics to finite-dimensional learning problems, we introduce a wrapped finite-dimensional surrogate that cyclically reintroduces lost coordinates while preserving the shift-based orbit structure. This weighted backward shift mechanism serves as the backbone of the entire framework: it generates the orbit dictionary used to build scaffolds, supports the exact interpolation theory underlying realizability, and provides the structured representation class on which WBSNN performs prediction.

In this sense, the paper shifts the center of gravity of learning theory from parametrization alone to the interaction between dataset geometry, orbit structure, and constructive realization.

The framework's backbone. At the core of our framework lies a structured, shift-based architecture $W^{(L)}$ (introduced and defined below), which cyclically shifts and scales input features across layers. Unlike traditional feedforward networks, its modulo-d design preserves all input dimensions throughout depth, enabling persistent feature recombination. This recurrence serves as a finite-dimensional surrogate of orbit dynamics from infinite-dimensional operator theory, where iterated linear actions generate structured, dense trajectories. With only d parameters, $W^{(L)}$ builds global representations from local shifts, providing an interpretable and expressive backbone for both interpolation and generalization.

Definition 1 *We introduce the* wrapped *finite-dimensional iterate, denoted $W^{(m)}$, used as a technical surrogate for backward-shift dynamics in infinite dimension: rather than shifting mass out of the ambient space, indices are taken modulo d, so that orbit iterates remain in $\mathbb{R}^d$ while still exhibiting shift-like propagation. Concretely, for $x \in \mathbb{R}^d$ and weights $w = (w_0, \ldots, w_{d-1}) \in \mathbb{R}^d$, we set*

$$\left(W^{(m)}x\right)_{i-1} := \left(\prod_{k=0}^{m-1} w_{(i+k) \bmod d}\right) x_{(i+m-1) \bmod d}, \tag{2}$$

for $i = 1, \ldots, d$ and $m \geq 1$, with the empty product interpreted as 1 for $m = 0$.

This wrapping is not a standard finite-dimensional dynamical system; rather, it is a technical device that adapts shift-based orbit constructions, naturally defined in infinite-dimensional spaces, to a fixed finite dimension.

Note. For intuition, the reader may view $W^{(L)}x$ as a shift-and-scale transformation of x, beginning at the L-th entry and preserving the cyclic order of the components modulo d.

Example 1 *Let $d = 4$, $L = 2$, and weights (w_0, w_1, w_2, w_3). Then for an input vector $x = (x_0, x_1, x_2, x_3) \in \mathbb{R}^4$, then*

$$W^{(2)}x = \begin{bmatrix} 0 & 0 & w_1w_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_2w_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_3w_0 & 0 \\ 0 & 0 & 0 & 0 & 0 & w_0w_1 \end{bmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} w_1w_2x_2 \\ w_2w_3x_3 \\ w_3w_0x_0 \\ w_0w_1x_1 \end{pmatrix}.$$

Remark 2 *The modulo-d wrap-around prevents the iterates $W^{(L)}$ from vanishing as in the classical finite-dimensional weighted backward shift. Hence, unlike the classical case, iteration does not destroy coordinates, but instead propagates and recombines information across all positions. Thus the wrapped shifts provide a structured linear backbone in which information can circulate across all coordinates under iteration, with low parameter complexity.*

Example 2 (Weight regimes ($d = 3, L = 2$)) *Let $w = (c, c, c)$, then $W^{(2)}x = \left(c^2x_2,\ c^2x_0,\ c^2x_1\right)$. Thus, $c = 0.5$ (vanishing) $\Rightarrow$ decay by 0.5^2; $c = 1$ (neutral) $\Rightarrow$ pure permutation (no scaling); $c = 1.5$ (exploding) $\Rightarrow$ amplification by 1.5^2. In general, the product $\prod_{k=0}^{d-1} w_k$ governs energy growth/decay, providing a simple knob for stability and expressivity.*

The following lemma reveals a key structural property of W that limits the distinctiveness of large powers.

Lemma 3 *Given W with d predictors, for every $X \in \mathbb{R}^d$ and every $L \geq 1$, we have that $W^{(L)}X = \lambda W^{(L \bmod d)}X$, where $\lambda = \left(\prod_{k=0}^{d-1} w_k\right)^m$ with $L = md + (L \mod d)$.*

Proof For $1 \leq i \leq d$, (2) gives $(W^{(L)}X)_{i-1} = \left(\prod_{k=0}^{L-1} w_{(i+k) \bmod d}\right) X_{(i+L-1) \bmod d}$. Similarly,

$$(W^{(L \bmod d)}X)_{i-1} = \left(\prod_{k=0}^{(L \bmod d)-1} w_{(i+k) \bmod d}\right) X_{(i+(L \bmod d)-1) \bmod d}.$$

Since $(i + L - 1) \bmod d = (i + (L \bmod d) - 1) \bmod d$, the indices match, so $X_{(i+L-1) \bmod d} = X_{(i+(L \bmod d)-1) \bmod d}$. For $L \geq 1$, there exists $m \geq 0$ such that $L = md + (L \bmod d)$, where $0 \leq L \bmod d < d$. The product in $W^{(L)}$ splits as:

$$\prod_{k=0}^{L-1} w_{(i+k) \bmod d} = \left(\prod_{n=0}^{m-1} \prod_{k=0}^{d-1} w_{(i+nd+k) \bmod d}\right) \prod_{k=0}^{(L \bmod d)-1} w_{(i+md+k) \bmod d}. \tag{3}$$

Indeed, the first md terms of $\prod_{k=0}^{L-1} w_{(i+k) \bmod d}$ are m full cycles (of length d) with indices $i+nd+k$ for n from 0 to $m-1$ and k from 0 to $d-1$. So, for each n, we get

$$w_{(i+nd) \bmod d} \cdot w_{(i+nd+1) \bmod d} \cdot \ldots \cdot w_{(i+nd+(d-1)) \bmod d}$$

which covers one full cycle of indices. The remaining $L \bmod d$ terms (from $k = md$ to $k = L-1$) are

$$w_{(i+md) \bmod d} \cdot w_{(i+md+1) \bmod d} \cdot \ldots \cdot w_{(i+md+(L \bmod d)-1) \bmod d}$$

which is exactly $\prod_{k=0}^{(L \bmod d)-1} w_{(i+md+k) \bmod d}$.

For each n, the inner product $\prod_{k=0}^{d-1} w_{(i+nd+k) \bmod d} = \prod_{k=0}^{d-1} w_{(i+k) \bmod d} = \prod_{k=0}^{d-1} w_k$, since the indices $(i+nd+k) \bmod d$ cycle through $i, i+1, \ldots, i+d-1$. Thus, the first factor in parentheses of the right-hand side of (3) is $\left(\prod_{k=0}^{d-1} w_k\right)^m$. Hence, $(W^{(L)}X)_{i-1} = \left(\prod_{k=0}^{d-1} w_k\right)^m (W^{(L \bmod d)}X)_{i-1}$, for all $i = 1, \ldots, d$. Therefore, the scalar $\lambda = \left(\prod_{k=0}^{d-1} w_k\right)^m$ satisfies the lemma. ■

2 Orbit-Based Constructive Realization of Linear Operators

This section develops the constructive backbone of the paper. The central question is whether a target operator can be realized, with explicit and analyzable error, on a data-induced orbit scaffold extracted from the dataset itself. The objects introduced here—scaffold, projected target, projection error, parametrization error, and CCDS realization—will remain the governing objects of the predictive theory in Section 3.

From the viewpoint adopted here, the scaffold is not an auxiliary representation layer but the primary geometric object through which learnability is organized. More precisely, the data first determines an orbit scaffold encoding part of its intrinsic geometry. The target is then projected onto that scaffold, and the resulting projected target is constructively synthesized there. This immediately separates two structural bottlenecks: a geometric bottleneck, measuring how well the scaffold captures the target's action on the data, and a constructive bottleneck, measuring how well the projected target can be realized once the scaffold has been fixed. The remainder of this section develops these two bottlenecks in order: scaffold selection first, then constructive realization on the selected scaffold. Given any linear operator $A : \mathbb{R}^d \to \mathbb{R}^C$, a finite input–target dataset

$$\Xi = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^C,$$

a fixed dictionary $\mathcal{D} \subset \mathbb{R}^C$, and a tolerance $\varepsilon > 0$, we ask whether there exists a map $\hat{y} : \mathbb{R}^d \to \mathbb{R}^C$ with

$$\{\hat{y}(x) : (x, y) \in \Xi\} \subset \operatorname{span}(\mathcal{D})$$

such that

$$\|A(\cdot) - \hat{y}(\cdot)\|_\Xi < \varepsilon,$$

where the empirical mean seminorm on Ξ is defined by $\|F\|_\Xi := \frac{1}{n} \sum_{i=1}^n \|F(x_i)\|_2$.

This manuscript initiates a principled study of this realizability problem in the case where the dictionary $\mathcal{D}$ is not arbitrary, but is instead a structured, data-induced orbit scaffold encoding the intrinsic geometry of Ξ, as formalized below.

2.1 Exact interpolation Theorem.

The following is the foundational layer of our theory of learnability. Thanks to Hahn-Banach Theorem, it shows that for both finite or infinite dimensional Banach spaces, under a technical independence condition, any finite dataset admits an exact realization within a class of transformed weighted shifts. We now present and prove an explicit construction of such a realization.

Convention for orbit iterates. Let X be a Banach space and let $T \in \mathcal{L}(X, X)$. We use T^L to denote the admissible orbit iterate at depth L. In the finite-dimensional regime $X = \mathbb{R}^d$, T^L denotes the wrapped weighted-shift iterate $W^{(L)}$ from Definition 1, and the admissible depth set is $\mathcal{I}_T = \{0, \dots, d-1\}$. In the infinite-dimensional sequence-space regime, $T = B_w$ is the classical weighted backward shift defined in (1), $T^L = B_w^L$, and the admissible depth set is $\mathcal{I}_T = \mathbb{N}_0$.

Theorem 4 (Exact interpolation along independent orbits) *Let X and Y be real Banach spaces, finite- or infinite-dimensional. Let $T : X \to X$ be a bounded linear operator, with admissible depth set $\mathcal{I}_T$ as specified above, and let $\{(x_i, y_i)\}_{i=1}^{n_0} \subset X \times Y$ be a finite input–target set. Suppose that for each $i = 1, \dots, n_0$ there exists an integer $L_i \in \mathcal{I}_T$ such that*

$$z_i := T^{L_i} x_i \notin \operatorname{span}\{z_1, \dots, z_{i-1}\}, \qquad i \geq 2. \tag{4}$$

Then there exists a bounded linear operator $J : X \to Y$ such that $JT^{L_i} x_i = y_i,\ i = 1, \dots, n_0$.

Proof We construct J recursively. Let $J_0 := 0 \in \mathcal{L}(X, Y)$. For $i = 1$, pick any $L_1 \in \mathcal{I}_T$, set $z_1 := T^{L_1} x_1$ and find a linear functional $f_1 \in X^*$ with $f_1(z_1) = 1$. Define $J_1 := J_0 + \big(y_1 - J_0 z_1\big) \otimes f_1$, where $(y \otimes f)(v) := f(v)y$. Equivalently, for all $v \in X$ we have $J_1 v := J_0 v + (y_1 - J_0 z_1) \otimes f_1(v) = 0 + (y_1 - 0) \otimes f_1(v) = f_1(v) y_1$. Thus, $J_1 \in \mathcal{L}(X, Y)$ and $J_1 z_1 = 1 \cdot y_1 = y_1$.

Set $z_i := T^{L_i} x_i$, for $i \geq 2$ satisfying (4). Suppose that for some $i \geq 2$, the operator $J_{i-1} \in \mathcal{L}(X, Y)$ has already been constructed so that

$$J_{i-1} z_j = y_j, \qquad j = 1, \dots, i-1.$$

By (4), we have $z_i \notin \operatorname{span}\{z_1, \dots, z_{i-1}\}$. Let $M_i := \operatorname{span}\{z_1, \dots, z_{i-1}, z_i\}$. Define a linear functional $g_i : M_i \to \mathbb{R}$ by $g_i(z_j) = 0, \quad j = 1, \dots, i-1,\ g_i(z_i) = 1$, and extending linearly on M_i. This is well-defined because $z_i \notin \operatorname{span}\{z_1, \dots, z_{i-1}\}$. Since M_i is finite-dimensional, g_i is continuous. By the Hahn–Banach Theorem (Conway, 1990, Corollary 3.15, p.112), there exists a bounded linear functional $f_i \in X^*$ extending g_i, hence $f_i(z_j) = 0, \quad j < i,\ f_i(z_i) = 1$. Define

$$J_i v := J_{i-1} v + f_i(v)\big(y_i - J_{i-1} z_i\big), \qquad v \in X.$$

Equivalently, $J_i = J_{i-1} + \big(y_i - J_{i-1} z_i\big) \otimes f_i$. Since J_{i-1} and f_i are bounded, $J_i \in \mathcal{L}(X, Y)$.

Now, $J_i z_i = J_{i-1} z_i + f_i(z_i)\big(y_i - J_{i-1} z_i\big) = J_{i-1} z_i + \big(y_i - J_{i-1} z_i\big) = y_i$. For $j < i$, we have $f_i(z_j) = 0$, so $J_i z_j = J_{i-1} z_j + f_i(z_j)\big(y_i - J_{i-1} z_i\big) = J_{i-1} z_j = y_j$. Thus $J_i z_j = y_j$ for every $j \leq i$. By induction, after n_0 steps, for $1 \leq i \leq n_0$: $J_i z_j = y_j$ for $j \leq i$. Finally, by setting $J := J_{n_0} \in \mathcal{L}(X, Y)$ we have $J z_i = J T^{L_i} x_i = y_i,\ i = 1, \dots, n_0$. ∎

Remark 5 (Operator-agnostic interpolation) *The exact interpolation argument is not specific to weighted shifts. It only uses the existence of orbit representatives $z_i := T^{L_i} x_i$ satisfying the recursive linear-independence condition (4). Thus, for interpolation alone, T may be any linear operator. Weighted shifts enter the framework because they provide the structured orbit geometry used later for scaffold construction, projection control, and constructive realization.*

Remark 6 (Invertible interpolation when $X = Y$) *In the special case $X = Y$, the interpolation map may be chosen as an invertible perturbation of the identity. Indeed, if the recursive construction is initialized with $J_0 = I_X$, then*

$$J = I_X + \sum_{i=1}^{N} \big(y_i - J_{i-1} T^{L_i} x_i\big) \otimes f_i.$$

Hence, if $\sum_{i=1}^{N} \|y_i - J_{i-1} T^{L_i} x_i\| \, \|f_i\| < 1$, then $\|J - I_X\| < 1$, so J is an isomorphism by the Neumann lemma (Conway, 1990, Lemma 2.1, p 192). Thus, when $X = Y$, exact interpolation can be realized by an isomorphism $J = I_X + F$, where F has finite rank.

A finite-dimensional quantitative sharpening of exact interpolation Theorem 4 is an existential realizability result. Its importance is structural: it shows that, under the orbit-independence condition (4), exact interpolation is possible, in finite/infinite dimensional Banach spaces. In the finite-dimensional Euclidean setting, however, one can sharpen this statement quantitatively by selecting a canonical exact interpolant and controlling its norm explicitly. Fix one block $D = \{(x_1, y_1), \dots, (x_s, y_s)\} \subset \Xi_{\text{sub}}$, and choose depths $L_1, \dots, L_s$ such that $z_i := W^{(L_i)} x_i \in \mathbb{R}^d,\ i = 1, \dots, s$, are linearly independent. Write

$$Z := \begin{bmatrix} z_1 & \cdots & z_s \end{bmatrix} \in \mathbb{R}^{d \times s}, \qquad Y := \begin{bmatrix} y_1 & \cdots & y_s \end{bmatrix} \in \mathbb{R}^{C \times s}. \tag{5}$$

Since the columns of Z are linearly independent, Z has full column rank, so $Z^\dagger = (Z^\top Z)^{-1} Z^\top$, and $\sigma_{\min}(Z) > 0$.

Theorem 7 (Minimal-norm exact interpolant and conditioning bound) *Define $J^{\min} := YZ^\dagger \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^C)$. Then:*

1. *$J^{\min} z_i = y_i$ for every $i = 1, \dots, s$.*
2. *Among all linear maps $J : \mathbb{R}^d \to \mathbb{R}^C$ satisfying $Jz_i = y_i,\ i = 1, \dots, s$, the operator $J^{\min}$ has minimal operator norm.*
3. *One has the quantitative bound $\|J^{\min}\| \le \frac{\|Y\|}{\sigma_{\min}(Z)} \le \frac{\|Y\|_F}{\sigma_{\min}(Z)}$.*

Proof Because Z has full column rank, $Z^\dagger Z = I_s$. Hence, for each canonical basis vector $e_i \in \mathbb{R}^s$,

$$J^{\min} z_i = YZ^\dagger z_i = YZ^\dagger Z e_i = Ye_i = y_i.$$

This proves exact interpolation. Let $E := \operatorname{span}\{z_1, \dots, z_s\} \subset \mathbb{R}^d$. Since the columns of Z are linearly independent, the map $Z : \mathbb{R}^s \to E,\ a \mapsto Za$, is a vector space isomorphism. Hence every $u \in E$ can be written uniquely as $u = Za$, and $J^{\min} u = YZ^\dagger Za = Ya$. Thus the restriction of $J^{\min}$ to E is exactly the unique linear map $J_E : E \to \mathbb{R}^C$ such that $J_E z_i = y_i$ for all i. Now let $\widetilde{J} : \mathbb{R}^d \to \mathbb{R}^C$ be any other linear map satisfying $\widetilde{J} z_i = y_i,\ i = 1, \dots, s$. If $u = Za \in E$, then

$$\widetilde{J} u = \widetilde{J} Za = \sum_{i=1}^{s} a_i \widetilde{J} z_i = \sum_{i=1}^{s} a_i y_i = Ya = J^{\min} u.$$

So $\widetilde{J}$ and $J^{\min}$ agree on E. Therefore

$$\|\widetilde{J}\| \ge \sup_{\substack{u \in E \\ \|u\| = 1}} \|\widetilde{J} u\| = \sup_{\substack{u \in E \\ \|u\| = 1}} \|J^{\min} u\|.$$

On the other hand, if $v \in E^\perp$, then $Z^\top v = 0$, hence $Z^\dagger v = (Z^\top Z)^{-1} Z^\top v = 0$, so $J^{\min} v = YZ^\dagger v = 0$. Thus $J^{\min}$ vanishes on $E^\perp$, and therefore

$$\|J^{\min}\| = \sup_{\substack{u \in \mathbb{R}^d \\ \|u\| = 1}} \|J^{\min} u\| = \sup_{\substack{u \in E \\ \|u\| = 1}} \|J^{\min} u\|.$$

Combining the last two displays yields $\|\widetilde{J}\| \ge \|J^{\min}\|$. Hence $J^{\min}$ has minimal operator norm among all exact interpolants. For the quantitative estimate, $\|J^{\min}\| = \|YZ^\dagger\| \le \|Y\|\,\|Z^\dagger\|$. Since Z has full column rank, $\|Z^\dagger\| = \frac{1}{\sigma_{\min}(Z)}$. Thus $\|J^{\min}\| \le \frac{\|Y\|}{\sigma_{\min}(Z)}$. Finally, $\|Y\| \le \|Y\|_F$, which gives $\|J^{\min}\| \le \frac{\|Y\|_F}{\sigma_{\min}(Z)}$. ∎

From exact interpolation to local stability. Theorem 7 shows that exact interpolation is stable only when the local orbit matrix $Z = [W^{(L_i)}x_i]_i$ is well conditioned: the factor $\sigma_{\min}(Z)^{-1}$ measures the amplification caused by solving the blockwise interpolation problem.

Local-chart control. The next result shows that, on each block, good conditioning of the input matrix X and the orbit matrix Z, together with small orbit displacement $\left(\sum_i \|x_i - W^{(L_i)}x_i\|^2\right)^{1/2}$, forces the exact interpolant J_k to stay close to an observable local chart. Thus the operators J_k act as implicit atlas regularizers: they regularize the scaffold by localizing interpolation on well-conditioned blocks whose orbit representatives remain geometrically faithful to the original data.

Theorem 8 (Blockwise exact interpolant as an observable local chart) *Let the block $D = \{(x_i, y_i)\}_{i=1}^s \subset \mathbb{R}^d \times \mathbb{R}^C$, and set $U_D := \operatorname{span}\{x_1, \dots, x_s\}$. Choose depths $L_1, \dots, L_s$ and define $z_i := W^{(L_i)}x_i,\ i = 1, \dots, s$, assuming that $z_1, \dots, z_s$ are linearly independent. Write*

$$Z := \begin{bmatrix} z_1 & \cdots & z_s \end{bmatrix} \in \mathbb{R}^{d\times s}, \qquad Y := \begin{bmatrix} y_1 & \cdots & y_s \end{bmatrix} \in \mathbb{R}^{C\times s}.$$

Define the exact block interpolant $J_D := YZ^\dagger$. Let $E := \operatorname{span}\{z_1, \dots, z_s\}$, and let Π_E denote the orthogonal projection onto E. If

$$\sum_{i=1}^s c_i x_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^s c_i y_i = 0, \tag{6}$$

then there exists $A_D \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^C)$ satisfying $A_D x_i = y_i,\ i = 1, \dots, s$. Moreover,

$$\|J_D - A_D\Pi_E\| \le \frac{\|Y\|_F}{\sigma^+_{\min}(X)\,\sigma_{\min}(Z)} \left(\sum_{i=1}^s \|x_i - W^{(L_i)}x_i\|^2\right)^{1/2}, \tag{7}$$

where $\sigma^+_{\min}(X)$ denotes the smallest nonzero singular value of X.

Proof Let $X := \begin{bmatrix} x_1 & \cdots & x_s \end{bmatrix}$. The compatibility condition (6) implies that $\ker X \subset \ker Y$. Since $X^\dagger X$ is the orthogonal projection of $\mathbb{R}^s$ onto $(\ker X)^\perp$, and since Y vanishes on $\ker X$, we have $Y = YX^\dagger X$. indeed, every $c \in \mathbb{R}^s$ decomposes uniquely as $c = c_0 + c_1$ with $c_0 \in \ker X$ and $c_1 \in (\ker X)^\perp$, then $YX^\dagger Xc = Yc_1 = Y(c_0 + c_1) = Yc$. Therefore the operator $A_D := YX^\dagger : \mathbb{R}^d \to \mathbb{R}^C$ satisfies $YX^\dagger x_i = YX^\dagger Xe_i = Ye_i = y_i$. Moreover,

$$\|A_D\| = \|YX^\dagger\| \le \|Y\|\,\|X^\dagger\| = \frac{\|Y\|}{\sigma^+_{\min}(X)} \le \frac{\|Y\|_F}{\sigma^+_{\min}(X)}. \tag{8}$$

Now define the observable defect matrix

$$\Delta_D^{\mathrm{obs}} := Y - A_D Z = \begin{bmatrix} y_1 - A_D z_1 & \cdots & y_s - A_D z_s \end{bmatrix}.$$

Since Z has full column rank, $ZZ^\dagger = \Pi_E$. Therefore

$$J_D - A_D\Pi_E = YZ^\dagger - A_D ZZ^\dagger = (Y - A_D Z)Z^\dagger = \Delta_D^{\mathrm{obs}} Z^\dagger.$$

Taking operator norms gives

$$\|J_D - A_D\Pi_E\| \le \|\Delta_D^{\mathrm{obs}}\|\,\|Z^\dagger\| = \frac{\|\Delta_D^{\mathrm{obs}}\|}{\sigma_{\min}(Z)} \le \frac{\|\Delta_D^{\mathrm{obs}}\|_F}{\sigma_{\min}(Z)}. \tag{9}$$

Finally, $y_i - A_D z_i = A_D x_i - A_D z_i = A_D(x_i - z_i)$. Thus $\|\Delta_D^{\mathrm{obs}}\|_F^2 = \sum_{i=1}^s \|A_D(x_i - z_i)\|^2 \le \|A_D\|^2 \sum_{i=1}^s \|x_i - z_i\|^2$. Using this estimate together with (9) and (8) yields (7). ∎

Remark 9 *We will return to the local-chart defect $\|J_k - A_{D_k}\Pi_{E_k}\|$ later in Theorem 32, where it is linked to the most important learning bottleneck of our framework: the projection error of the orbit scaffold. Thus block geometry is one mechanism by which local interpolation becomes global learnability.*

The two learning bottlenecks. The main goal of this section is operator realizability, which is mediated by orbit dictionaries, and the first structural step is the construction of a data-induced orbit scaffold associated with the dataset Ξ.

Select a small anchor subset $\Xi_{\text{sub}} \subseteq \Xi$ and a partition $\{D_k\}_{k=1}^{K}$ of Ξ_{sub}. The construction of the scaffold proceeds as follows: we learn a wrapped weighted backward shift operator W, then we apply the exact interpolation Theorem 4/Theorem 7 blockwise to the sets D_k, producing linear operators J_k that interpolate exactly on each block.

More precisely, for each $1 \le k \le K$, let $D_k = \{(x_{k,1}, y_{k,1}), \dots, (x_{k,|D_k|}, y_{k,|D_k|})\} \subseteq \Xi_{\text{sub}}$, for every training pair $(x_{k,j}, y_{k,j}) \in D_k$, assume one can choose an integer $L_{k,j} \in \{0, \dots, d-1\}$ so that condition (4) holds on D_k; then Theorem 4/Theorem 7 yields a linear operator J_k such that $y_{k,j} = J_k W^{(L_{k,j})} x_{k,j}$. Thus wrapped orbit iterates generated by W are mapped by the corresponding operators J_k exactly onto the target outputs on the anchor subset. This leads to the orbit-based scaffold associated with Ξ_{sub}:

$$U(\Xi_{\text{sub}}) := \operatorname{span}\Big\{ J_k W^{(m)} x_{k,j} : (x_{k,j}, y_{k,j}) \in D_k,\ 1 \le k \le K,\ 0 \le m \le d-1 \Big\} \subseteq \mathbb{R}^C. \tag{10}$$

Hence $U(\Xi_{\text{sub}})$ is a finite-rank, data-induced subspace of the output Hilbert space encoding the orbit geometry of the dataset. The following decomposition is an immediate consequence of the triangle inequality and is therefore purely structural. It identifies the two fundamental bottlenecks of our learning framework.

Theorem 10 (Two–term error decomposition for operator realization on an orbit scaffold) *Let $\Xi = \{(x_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^d \times \mathbb{R}^C$ be a finite dataset, $\Xi_{\text{sub}} \subseteq \Xi$, and let $A : \mathbb{R}^d \to \mathbb{R}^C$ be a linear operator. Let $U(\Xi_{\text{sub}})$ be the orbit scaffold relative to Ξ_{sub} and $\Pi_{U(\Xi_{\text{sub}})} : \mathbb{R}^C \to \mathbb{R}^C$ the associated orthogonal projector and $\hat{y} : \mathbb{R}^d \to \mathbb{R}^C$. Then the following decomposition holds:*

$$\|A - \hat{y}\|_{\Xi} \ \le\ \underbrace{\|A - \Pi_{U_{\Xi_{\text{sub}}}} A\|_{\Xi}}_{\varepsilon_{\text{proj}}(A, U_{\Xi_{\text{sub}}};\, \Xi)} \ +\ \underbrace{\|\Pi_{U_{\Xi_{\text{sub}}}} A - \hat{y}\|_{\Xi}}_{\varepsilon_{\text{param}}(A, U_{\Xi_{\text{sub}}}, \hat{y};\, \Xi)}. \tag{11}$$

In practice, the projection term $\varepsilon_{\text{proj}}$ is typically the dominant bottleneck; the experiments in Section 4 confirm that once a high-quality scaffold is found, the parametrization term collapses reliably with our constructive realizer mechanism CCDS introduced and developed below.

The meaning of the decomposition (11) is not limited to an upper bound for a particular realization algorithm. The dataset Ξ induces a scaffold $U(\Xi_{\text{sub}})$ (a geometric "DNA" extracted from Ξ) that is a data-only object, independent of the operator A. The projection term $\varepsilon_{\text{proj}}(A, U(\Xi_{\text{sub}});\ \Xi)$ depends only on A and on this dataset-induced scaffold, and measures how much of the action of A on Ξ lies in the orbit geometry intrinsic to Ξ_{sub}. Equivalently, it quantifies how much of the structure inherent to the operator A is carried by the dataset Ξ_{sub}. Small values indicate that the outputs Ax are largely contained in the subspaces encoded by $U(\Xi_{\text{sub}})$, while large values signal directions of A that are transversal to the dataset's orbit structure. The parametrization term $\varepsilon_{\text{param}}(A, U(\Xi_{\text{sub}}), \hat{y};\ \Xi)$ instead measures whether, once this geometric compatibility holds, the realizer $\hat{y}$ constrained to $U(\Xi_{\text{sub}})$ can constructively synthesize the projected target. The decomposition therefore separates operator complexity, operator–dataset alignment, and algorithmic realizability, and can be read simultaneously as an approximation bound, a realizability statement, and a geometric diagnostic.

Theorem 10 identifies scaffold selection and realization as the two fundamental learning bottlenecks in orbit-based models. These two components will be the focus of the remainder of this section.

We begin addressing both bottlenecks within this paper. On the geometric side, we cast scaffold selection as an orbit-constrained optimization problem and develop a certifiable deviation-controlled algorithm for selecting the scaffold. On the realization side, we introduce a certifiable deviation-controlled algorithm for constructing a realizer on a fixed scaffold. When both the projection and

parametrization errors collapse, the target linear operator can be constructively realized on the data subspace with arbitrarily small error. In particular, if the target operator is the data-dependent linear operator induced by attention on the input sequence (i.e., for a fixed token sequence $X = (x_1, \ldots, x_n) \in \mathbb{R}^{d\times n}$, single-head attention determines a data-dependent linear operator $A(X)$ acting on token/value representations, and the corresponding output is obtained by applying the map $\nu \mapsto A(X)\nu$), then attention requires no separate treatment: it is simply a concrete instance of the general operator-realization problem studied here. Attention therefore appears as an immediate special case of a broader operator-realization principle. Taken together, these observations suggest that attention may be viewed less as a primitive architectural ingredient and more as a problem of realizing input-dependent linear operators, a task that may be approached, for instance, through our orbit-based framework without relying on softmax algebra or probabilistic arguments.

2.2 Collapsing the projection error: the scaffold selection problem.

We first show that the existence of optimal scaffolds under which projection error collapses, is a mathematically well-posed problem. Denote by $\mathcal{U}$ the set of all scaffolds relative to Ξ.

Theorem 11 (Existence of the best orbit-constrained scaffold) *Let $\Xi = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^C$, and let $A : \mathbb{R}^d \to \mathbb{R}^C$ be linear, then there exists $U_\star \in \mathcal{U}$ such that $\|A - \Pi_{U_\star} A\|_\Xi = \min_{U\in\mathcal{U}} \|A - \Pi_U A\|_\Xi$.*

Proof First, let us compute an upper bound for the dimension of any scaffold U relative to Ξ. Pick any subdataset $\Xi_{\text{sub}} \subset \Xi$ and assume it is partitioned as $\Xi_{\text{sub}} = D_0 \sqcup \cdots \sqcup D_{K-1}$. Then, each sample $x_i \in \Xi_{\text{sub}}$ belongs to exactly one block D_k. Hence, in the dictionary construction, x_i contributes only the d atoms $J_k W^{(m)} x_i, \qquad 0 \le m \le d-1$, associated with its own interpolant J_k. Therefore the total number of generating atoms is at most $d\,|\Xi_{\text{sub}}|$, and since the scaffold is a subspace of $\mathbb{R}^C$, we obtain $\dim U \le \min\big(C,\ d\,|\Xi_{\text{sub}}|\big) \le \min\big(C,\ d\,n\big)$. If $\mathcal{U}_k$ denotes the set of all k-dimensional scaffolds relative to Ξ. Hence,

$$\|A - \Pi_{U_\star} A\|_\Xi = \min_{1\le k\le \min(C,dn)} \ \min_{U\in\mathcal{U}_k} \|A - \Pi_U A\|_\Xi.$$

In order to conclude our proof, it will be enough to prove the existence of a best orbit-constrained scaffold of dimension k, for all $1 \le k \le \min(C, dn)$. Since Ξ is finite, there are only finitely many admissible choices of subsets and partitions on which the exact interpolation theorem can be applied. Each such admissible choice produces one scaffold. Therefore, for each fixed k, the family of all k-dimensional scaffolds induced by Ξ is finite, hence compact. On the other hand, the map $U \mapsto \Pi_U$ is continuous (Conway, 1990). Define $f(U) = \|A - \Pi_U A\|_\Xi$. Since A is fixed and $U \mapsto \Pi_U$ is continuous in operator norm, f is continuous on $\mathcal{U}_k$. Because $\mathcal{U}_k$ is compact and f is continuous, the extreme value theorem implies that f attains its minimum on $\mathcal{U}_k$. Hence, there exists $U_\star^{(k)} \in \mathcal{U}_k$ such that $f(U_\star^{(k)}) = \min_{U\in\mathcal{U}_k} f(U) = \min_{U\in\mathcal{U}_k} \|A - \Pi_U A\|_\Xi$. ■

From existence to computation: a local geometric control theory for scaffold expansion. Theorem 11 guarantees the existence of an optimal orbit-constrained scaffold, but it is nonconstructive. The missing issue is algorithmic: given the admissible orbit geometry induced by the dataset, how can one enlarge the scaffold in a controlled way so as to decrease the projection term. This is not a classical unconstrained low-rank approximation problem, because the admissible subspaces are not arbitrary subspaces of $\mathbb{R}^C$; they must arise from the orbit mechanism described above. We therefore isolate the projection term and develop a constructive *scaffold-level* deviation-control theory. The object being controlled is the projection residual associated with nested admissible scaffold expansions.

The resulting theory has four parts. First, the projection error is rewritten as a Frobenius residual of a snapshot matrix. Second, for one-dimensional scaffold expansions, we prove an exact energy

identity describing the drop in residual energy. Third, we characterize all one-step feasible drops and show how to realize any prescribed feasible drop explicitly. Fourth, we obtain a constructive theorem showing that any stepwise feasible projection-error profile can be enforced exactly. The greedy spectral step then appears as a distinguished special case.

Mean projection error versus RMS projection residual. Before developing the scaffold-selection mechanism, we clarify a minor but important change of empirical geometry. The projection term appearing in the two-term decomposition is measured in the empirical mean-norm seminorm, while the scaffold-selection algorithm below works with an empirical RMS projection residual, equivalently the Frobenius residual of the target snapshot matrix. This choice is algorithmic rather than conceptual: the RMS residual has a Hilbertian squared-energy structure, which allows exact one-step energy identities, spectral feasibility intervals, and greedy singular-vector updates. The two quantities are equivalent on every fixed finite dataset. More importantly, the RMS projection residual upper-bounds the empirical mean-norm projection error. Therefore, any certified collapse of the scaffold-selection residual implies collapse of the projection bottleneck appearing in the two-term decomposition.

Let $\Xi = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^C$, let $A : \mathbb{R}^d \to \mathbb{R}^C$ be linear, and let $U \subset \mathbb{R}^C$ be a scaffold with orthogonal projector Π_U. We have defined the empirical mean-norm projection error by $\varepsilon_{\mathrm{proj}}(A, U; \Xi) := \frac{1}{N} \sum_{i=1}^N \|(I - \Pi_U)Ax_i\|_2$. Now we define the *empirical RMS projection residual* by

$$\varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi) := \left(\frac{1}{N} \sum_{i=1}^N \|(I - \Pi_U)Ax_i\|_2^2 \right)^{1/2}.$$

Equivalently, we have the following.

Proposition 12 (Snapshot representation of the projection error) *Define the snapshot matrix $Y := \big[Ax_1, \ldots, Ax_N\big] \in \mathbb{R}^{C \times N}$. Then, for every orthogonal projector Π_U,*

$$\varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi)^2 = \frac{1}{N} \|(I - \Pi_U)Y\|_F^2.$$

Proof The i-th column of $(I - \Pi_U)Y$ is precisely $(I - \Pi_U)Ax_i$. Therefore, by the definition of the Frobenius norm, $\|(I - \Pi_U)Y\|_F^2 = \sum_{i=1}^N \|(I - \Pi_U)Ax_i\|_2^2$. Dividing by N, we obtain

$$\frac{1}{N} \|(I - \Pi_U)Y\|_F^2 = \frac{1}{N} \sum_{i=1}^N \|(I - \Pi_U)Ax_i\|_2^2 = \varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi)^2.$$

■

We now record the precise relationship between the mean and RMS projection errors.

Lemma 13 (Equivalence of mean and RMS projection errors) *For every scaffold $U \subset \mathbb{R}^C$, one has*

$$\varepsilon_{\mathrm{proj}}(A, U; \Xi) \le \varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi) \le \sqrt{N}\, \varepsilon_{\mathrm{proj}}(A, U; \Xi).$$

Consequently, on every fixed finite dataset Ξ, the empirical mean-norm projection error and the empirical RMS projection residual are equivalent. In particular,

$$\varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi) \to 0 \quad \Longrightarrow \quad \varepsilon_{\mathrm{proj}}(A, U; \Xi) \to 0.$$

Proof Set $r_i := (I - \Pi_U)Ax_i \in \mathbb{R}^C$, $a_i := \|r_i\|_2 \ge 0$. Then $\varepsilon_{\mathrm{proj}}(A, U; \Xi) = \frac{1}{N} \sum_{i=1}^N a_i$, whereas $\varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi) = \left(\frac{1}{N} \sum_{i=1}^N a_i^2 \right)^{1/2}$. We first prove the upper bound of the mean projection error by the

RMS residual. By the Cauchy–Schwarz inequality applied to the vectors $(a_1, \dots, a_N)$ and $(1, \dots, 1)$ in $\mathbb{R}^N$, we have

$$\sum_{i=1}^{N} a_i = \langle (a_1, \dots, a_N), (1, \dots, 1) \rangle \le \left(\sum_{i=1}^{N} a_i^2 \right)^{1/2} \left(\sum_{i=1}^{N} 1^2 \right)^{1/2}.$$

Since $\left(\sum_{i=1}^{N} 1^2\right)^{1/2} = \sqrt{N}$, we obtain $\sum_{i=1}^{N} a_i \le \sqrt{N} \left(\sum_{i=1}^{N} a_i^2\right)^{1/2}$. Dividing by N, this becomes $\frac{1}{N}\sum_{i=1}^{N} a_i \le \frac{1}{N}\sqrt{N} \left(\sum_{i=1}^{N} a_i^2\right)^{1/2} = \left(\frac{1}{N}\sum_{i=1}^{N} a_i^2\right)^{1/2}$. Therefore, $\varepsilon_{\mathrm{proj}}(A, U; \Xi) \le \varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi)$.

Conversely, since $a_i \ge 0$ for every i, we have $\sum_{i=1}^{N} a_i^2 \le \left(\sum_{i=1}^{N} a_i\right)^2$. Indeed, $\left(\sum_{i=1}^{N} a_i\right)^2 = \sum_{i=1}^{N} a_i^2 + 2\sum_{1 \le i < j \le N} a_i a_j \ge \sum_{i=1}^{N} a_i^2$. Taking square roots gives $\left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \le \sum_{i=1}^{N} a_i$. Dividing by $\sqrt{N}$, we obtain $\left(\frac{1}{N}\sum_{i=1}^{N} a_i^2\right)^{1/2} \le \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_i = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} a_i\right)$. Therefore, $\varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi) \le \sqrt{N}\, \varepsilon_{\mathrm{proj}}(A, U; \Xi)$. Combining the two inequalities yields the claim. ■

Fix an admissible scaffold space $V \subset \mathbb{R}^C$ arising from the orbit construction, and let $U \subseteq V$ be an admissible scaffold.

One-step scaffold expansion. We now work in the ambient Hilbert space $H = \mathbb{R}^C$ with its standard inner product. Let $U_0 \subset U_1 \subset \cdots \subset U_T \subset V$ be a nested chain of admissible subspaces satisfying $\dim(U_t/U_{t-1}) = 1,\ t = 1, \dots, T$. Equivalently, for each t there exists a unit vector $u_t \in V \cap U_{t-1}^{\perp}$ such that $U_t = U_{t-1} \oplus \mathrm{span}\{u_t\}$. For each t, define the residual matrix and its energy by

$$R_t := (I - \Pi_{U_t})Y, \qquad E_t := \|R_t\|_F^2 = N\varepsilon_{\mathrm{proj}}^{\mathrm{rms}}(A, U; \Xi)^2.$$

where last equality holds by Proposition 12.

Lemma 14 (One-step energy identity) *Let $U_t = U_{t-1} \oplus \mathrm{span}\{u\},\ u \perp U_{t-1}$, and $\|u\| = 1$. Then $E_t = E_{t-1} - \|u^\top Y\|_2^2 = E_{t-1} - \|u^\top R_{t-1}\|_2^2$.*

Proof Since $u \perp U_{t-1}$ and $\|u\| = 1$, the orthogonal projector onto U_t is $\Pi_{U_t} = \Pi_{U_{t-1}} + uu^\top$. Therefore

$$R_t = (I - \Pi_{U_t})Y = (I - \Pi_{U_{t-1}} - uu^\top)Y = R_{t-1} - u(u^\top Y).$$

Taking Frobenius norms and using $\|A - B\|_F^2 = \|A\|_F^2 - 2\langle A, B\rangle_F + \|B\|_F^2$, we get

$$\|R_t\|_F^2 = \|R_{t-1}\|_F^2 - 2\langle R_{t-1}, u(u^\top Y)\rangle_F + \|u(u^\top Y)\|_F^2.$$

We compute the two remaining terms separately.

First, using the Frobenius inner product $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$,

$$\begin{aligned} \langle R_{t-1}, u(u^\top Y)\rangle_F &= \mathrm{tr}\left(R_{t-1}^\top u(u^\top Y)\right) = \mathrm{tr}\left(u^\top Y(R_{t-1}^\top u)\right) \\ &= u^\top Y R_{t-1}^\top u \qquad \text{since } u^\top Y(R_{t-1}^\top u) \text{ is a scalar} \\ &= u^\top R_{t-1} R_{t-1}^\top u \qquad \text{because } u^\top R_{t-1} = u^\top Y \\ &= \|u^\top R_{t-1}\|_2^2. \end{aligned}$$

The identity $u^\top R_{t-1} = u^\top Y$ follows from $R_{t-1} = (I - \Pi_{U_{t-1}})Y$ and the fact that $u \perp U_{t-1}$, which implies $u^\top \Pi_{U_{t-1}} = 0$. Second, since $\|u\| = 1$,

$$\|u(u^\top Y)\|_F^2 = \|u\|_2^2\, \|u^\top Y\|_2^2 = \|u^\top Y\|_2^2.$$

But again $u^\top R_{t-1} = u^\top Y$, so $\|u(u^\top Y)\|_F^2 = \|u^\top R_{t-1}\|_2^2$. Substituting into the Frobenius expansion yields

$$E_t = E_{t-1} - 2\|u^\top R_{t-1}\|_2^2 + \|u^\top R_{t-1}\|_2^2 = E_{t-1} - \|u^\top R_{t-1}\|_2^2.$$

Since $u^\top R_{t-1} = u^\top Y$, the first identity follows as well. ■

Feasible one-step drops. Fix an admissible scaffold $U \subseteq V$, and write

$$S := V \cap U^\perp, \qquad R := (I - \Pi_U)Y, \qquad E := \|R\|_F^2.$$

Thus S is the admissible complement in which a new one-dimensional scaffold increment must lie. For $u \in S$ with $\|u\| = 1$, define the one-step capture functional $q(u) := \|u^\top R\|_2^2$. This is the amount of residual energy removed by adjoining the direction u.

Lemma 15 (Feasible one-step residual energies) *Let $M := \Pi_S R R^\top \Pi_S$. Then M is a symmetric positive semidefinite operator with range contained in S. Define*

$$\lambda_{\min} := \min_{\substack{u \in S \\ \|u\|=1}} u^\top M u, \qquad \lambda_{\max} := \max_{\substack{u \in S \\ \|u\|=1}} u^\top M u.$$

Then the following hold:

1. *$q(u) = u^\top M u$ for every $u \in S$ with $\|u\| = 1$.*
2. *The set of achievable capture energies $\{q(u) : u \in S,\ \|u\| = 1\}$ is exactly the interval $[\lambda_{\min}, \lambda_{\max}]$.*
3. *Consequently, the set of achievable next residual energies is exactly $\big[E - \lambda_{\max},\, E - \lambda_{\min}\big]$.*

Proof We first note that M is symmetric because both Π_S and $RR^\top$ are symmetric: $M^\top = (\Pi_S R R^\top \Pi_S)^\top = \Pi_S R R^\top \Pi_S = M$. Moreover, for every $x \in \mathbb{R}^C$, $x^\top M x = x^\top \Pi_S R R^\top \Pi_S x = \|x^\top \Pi_S R\|_2^2 \geq 0$, so M is positive semidefinite.

Proof of (1). Let $u \in S$ with $\|u\| = 1$. Since $\Pi_S u = u$ and $u^\top \Pi_S = (\Pi_S u)^\top = u^\top$, we have

$$u^\top M u = u^\top \Pi_S R R^\top \Pi_S u = u^\top R R^\top u.$$

On the other hand,

$$u^\top R R^\top u = (u^\top R)(u^\top R)^\top = \|u^\top R\|_2^2 = q(u).$$

Thus $q(u) = u^\top M u$.

Proof of (2). Because $u \mapsto u^\top M u$ is continuous on the unit sphere of S, the extreme value theorem shows that the minimum and maximum are attained. Hence

$$\{q(u) : u \in S,\ \|u\| = 1\} \subseteq [\lambda_{\min}, \lambda_{\max}],$$

and both endpoints belong to the image. It remains to show that every intermediate value is also attained. If $\dim S = 1$, then the unit sphere of S is $\{u, -u\}$ for some unit vector u, and since the quadratic form is even, $q(-u) = q(u)$, the image is a singleton. In this case $\lambda_{\min} = \lambda_{\max}$, so the image is precisely the degenerate interval $[\lambda_{\min}, \lambda_{\max}]$.

Assume now that $\dim S \geq 2$. Then the unit sphere of S is connected. Since $u \mapsto u^\top M u$ is continuous, its image is a connected subset of $\mathbb{R}$, hence an interval. Because this interval contains both $\lambda_{\min}$ and $\lambda_{\max}$, it must contain the whole closed interval between them. Therefore

$$\{q(u) : u \in S,\ \|u\| = 1\} = [\lambda_{\min}, \lambda_{\max}].$$

Proof of (3). By Lemma 14, after adjoining a unit direction $u \in S$ the new residual energy is $E^+ = E - q(u)$. Since $q(u)$ ranges exactly over $[\lambda_{\min}, \lambda_{\max}]$, the possible values of E^+ are exactly $E - [\lambda_{\min}, \lambda_{\max}] = [E - \lambda_{\max},\, E - \lambda_{\min}]$. This proves the claim. ∎

Lemma 16 (Explicit two-dimensional construction) *Assume* $\dim S \geq 2$*, and let* $v_1, v_2 \in S$ *be orthonormal eigenvectors of* M *with eigenvalues* $\lambda_1 \geq \lambda_2$*. Then for every target* $e \in [\lambda_2, \lambda_1]$*, there exists a unit vector* $u(\theta) := \cos\theta\, v_1 + \sin\theta\, v_2 \in S$ *such that* $q(u(\theta)) = e$*. More precisely, if* $\lambda_1 > \lambda_2$*, one may choose* θ *so that* $\cos^2\theta = \frac{e-\lambda_2}{\lambda_1-\lambda_2}$*. In that case, adjoining* $u(\theta)$ *enforces the one-step update* $E^+ = E - e$*.*

Proof Let $u(\theta) = \cos\theta\, v_1 + \sin\theta\, v_2$. Because $v_1, v_2 \in S$, we have $u(\theta) \in S$. Since v_1 and v_2 are orthonormal,

$$\|u(\theta)\|_2^2 = \cos^2\theta\, \|v_1\|_2^2 + \sin^2\theta\, \|v_2\|_2^2 + 2\cos\theta \sin\theta\, \langle v_1, v_2\rangle = \cos^2\theta + \sin^2\theta = 1.$$

Thus $u(\theta)$ is admissible. By Lemma 15, $q(u(\theta)) = u(\theta)^\top M u(\theta)$. Expanding and using the eigenvector relations $Mv_j = \lambda_j v_j$, we get

$$\begin{aligned} q(u(\theta)) &= (\cos\theta\, v_1 + \sin\theta\, v_2)^\top M(\cos\theta\, v_1 + \sin\theta\, v_2) \\ &= \cos^2\theta\, v_1^\top M v_1 + \sin^2\theta\, v_2^\top M v_2 + \cos\theta\sin\theta\, v_1^\top M v_2 + \cos\theta\sin\theta\, v_2^\top M v_1 \\ &= \cos^2\theta\, \lambda_1 + \sin^2\theta\, \lambda_2, \end{aligned}$$

because $v_1^\top M v_2 = \lambda_2 \langle v_1, v_2\rangle = 0$ and similarly $v_2^\top M v_1 = 0$.

If $\lambda_1 = \lambda_2$, then the interval $[\lambda_2, \lambda_1]$ is a single point and $q(u(\theta)) = \lambda_1 = \lambda_2$ for every θ, so the claim is immediate. Assume therefore that $\lambda_1 > \lambda_2$. Since

$$q(u(\theta)) = \cos^2\theta\, \lambda_1 + \sin^2\theta\, \lambda_2 = \lambda_2 + \cos^2\theta(\lambda_1 - \lambda_2),$$

the condition $q(u(\theta)) = e$ is equivalent to $\cos^2\theta = \frac{e-\lambda_2}{\lambda_1-\lambda_2}$. Because $e \in [\lambda_2, \lambda_1]$, the right-hand side belongs to $[0, 1]$, so such a θ exists. For this choice of θ, we obtain $q(u(\theta)) = e$. Finally, by Lemma 14, adjoining $u(\theta)$ changes the energy by $E^+ = E - q(u(\theta)) = E - e$. ∎

Theorem 17 (Constructive deviation-control principle for scaffold selection) *Let* $V \subset \mathbb{R}^C$ *be a closed admissible scaffold space. Let* $Y \in \mathbb{R}^{C\times N}$ *be fixed, and define*

$$R_0 := Y, \qquad U_0 := \{0\}, \qquad E_0 := \|R_0\|_F^2.$$

Fix an integer $T \geq 1$ *and a target energy profile* $E_0 \geq E_1 \geq \cdots \geq E_T \geq 0$*. For each* $t = 1, \ldots, T$*, define the admissible complement* $S_{t-1} := V \cap U_{t-1}^\perp$ *and the compressed residual covariance operator*

$$M_{t-1} := \Pi_{S_{t-1}} R_{t-1} R_{t-1}^\top \Pi_{S_{t-1}}.$$

Let

$$\lambda_{\min}^{(t-1)} := \min_{\substack{u\in S_{t-1}\\ \|u\|=1}} u^\top M_{t-1} u, \qquad \lambda_{\max}^{(t-1)} := \max_{\substack{u\in S_{t-1}\\ \|u\|=1}} u^\top M_{t-1} u.$$

Assume the feasibility inequalities

$$E_{t-1} - E_t \in \big[\lambda_{\min}^{(t-1)}, \lambda_{\max}^{(t-1)}\big], \qquad t = 1, \ldots, T.$$

Then there exists a deterministic sequence of unit vectors $u_t \in S_{t-1}$ and a nested scaffold chain $U_t := U_{t-1} \oplus \operatorname{span}\{u_t\} \subset V$ such that, with $R_t := (I - \Pi_{U_t})Y$, one has

$$\|R_t\|_F^2 = E_t, \qquad t = 1, \dots, T.$$

Moreover, whenever $\dim S_{t-1} \geq 2$, one may choose u_t explicitly by Lemma 16 so as to realize the prescribed drop $E_{t-1} - E_t$ exactly.

Proof We argue by induction on t. For $t = 0$, the conclusion holds by definition:

$$U_0 = \{0\}, \qquad R_0 = Y, \qquad \|R_0\|_F^2 = E_0.$$

Assume now that for some $t \in \{1, \dots, T\}$ we have already constructed $U_{t-1} \subset V$ and

$$R_{t-1} = (I - \Pi_{U_{t-1}})Y$$

in such a way that $\|R_{t-1}\|_F^2 = E_{t-1}$. At step t, the admissible unit directions are precisely the unit vectors in $S_{t-1} = V \cap U_{t-1}^{\perp}$. By Lemma 15, the set of achievable one-step capture energies

$$q(u) = \|u^\top R_{t-1}\|_2^2, \qquad u \in S_{t-1},\ \|u\| = 1,$$

is exactly the interval $\left[\lambda_{\min}^{(t-1)}, \lambda_{\max}^{(t-1)}\right]$. By the feasibility assumption, $E_{t-1} - E_t \in \left[\lambda_{\min}^{(t-1)}, \lambda_{\max}^{(t-1)}\right]$. Hence there exists a unit vector $u_t \in S_{t-1}$ such that $q(u_t) = E_{t-1} - E_t$. Define $U_t := U_{t-1} \oplus \operatorname{span}\{u_t\}$. Since $u_t \in V$, we have $U_t \subset V$, and since $u_t \perp U_{t-1}$, the sum is orthogonal and therefore $U_{t-1} \subset U_t$, so the chain remains nested.

Now define $R_t := (I - \Pi_{U_t})Y$. Applying Lemma 14, we obtain

$$\|R_t\|_F^2 = \|R_{t-1}\|_F^2 - q(u_t) = E_{t-1} - (E_{t-1} - E_t) = E_t.$$

This completes the induction step. Therefore, by induction, the whole sequence

$$U_0 \subset U_1 \subset \cdots \subset U_T \subset V$$

can be constructed so that $\|R_t\|_F^2 = E_t$ for every $t = 1, \dots, T$.

For the final assertion, assume $\dim S_{t-1} \geq 2$. Then, instead of choosing u_t abstractly from Lemma 15, one may choose it explicitly by Lemma 16, using two orthonormal eigenvectors of M_{t-1} and selecting the mixing angle so that the capture energy equals the prescribed drop $E_{t-1} - E_t$. ■

Corollary 18 (One-step infeasibility outside the scaffold feasibility interval) *In the setting of Lemma 15, let $e \notin [\lambda_{\min}, \lambda_{\max}]$. Then there exists no unit vector $u \in S$ such that $q(u) = e$. Equivalently, there is no one-dimensional admissible scaffold expansion $U^+ = U \oplus \operatorname{span}\{u\}$ whose residual energy satisfies $E^+ = E - e$. Consequently, any prescribed profile violating the stepwise interval constraints of Theorem 17 is unrealizable in the fixed admissible scaffold space.*

Proof By Lemma 15, $\{q(u) : u \in S,\ \|u\| = 1\} = [\lambda_{\min}, \lambda_{\max}]$. Hence if $e \notin [\lambda_{\min}, \lambda_{\max}]$, no admissible unit vector $u \in S$ can satisfy $q(u) = e$. By Lemma 14, this is equivalent to the nonexistence of a one-dimensional admissible scaffold expansion producing the residual-energy update $E^+ = E - e$. The final statement follows immediately by applying this argument at the first step where the prescribed profile violates the interval constraint. ■

Greedy maximal collapse as a distinguished special case. The constructive theorem above governs arbitrary *feasible* projection-error profiles. If one does not wish to prescribe the next drop in advance, the most natural canonical choice is instead to maximize the immediate decrease in residual energy.

Proposition 19 (Greedy one-step maximal collapse) *At step t, among all unit vectors $u \in S_{t-1}$, the choice maximizing the one-step drop $q(u) = \|u^\top R_{t-1}\|_2^2$ is any unit eigenvector of M_{t-1} associated with the largest eigenvalue $\lambda_{\max}^{(t-1)}$. The achieved drop is exactly $\lambda_{\max}^{(t-1)}$. Equivalently, such a maximizing direction is the left singular vector associated with the largest singular value of $\Pi_{S_{t-1}} R_{t-1}$.*

Proof By Lemma 15, for every unit vector $u \in S_{t-1}$, we have $q(u) = u^\top M_{t-1} u$. Therefore maximizing the one-step drop is equivalent to solving

$$\max_{\substack{u \in S_{t-1} \\ \|u\|=1}} u^\top M_{t-1} u.$$

Since M_{t-1} is symmetric, the Rayleigh–Ritz principle (Horn and Johnson, 2012) implies that this maximum is equal to the largest eigenvalue $\lambda_{\max}^{(t-1)}$, and that the maximizers are exactly the associated unit eigenvectors. Hence the largest achievable one-step drop is $\lambda_{\max}^{(t-1)}$. For the singular-vector formulation, set $B := \Pi_{S_{t-1}} R_{t-1}$. Then

$$BB^\top = (\Pi_{S_{t-1}} R_{t-1})(\Pi_{S_{t-1}} R_{t-1})^\top = \Pi_{S_{t-1}} R_{t-1} R_{t-1}^\top \Pi_{S_{t-1}} = M_{t-1},$$

because $\Pi_{S_{t-1}}$ is symmetric. The eigenvectors of $BB^\top$ are precisely the left singular vectors of B, and the corresponding eigenvalues are the squares of the singular values. Therefore the maximizing eigenvectors of M_{t-1} are exactly the left singular vectors associated with the largest singular value of $\Pi_{S_{t-1}} R_{t-1}$. ■

Transition to realizability and prediction. The scaffold-selection mechanism developed above addresses only the projection term in the two-term decomposition. Once an admissible scaffold has been selected and its projection error has been reduced in a certified way, the remaining problem is constructive synthesis on that scaffold, namely the collapse of the parametrization term. We now turn to this second bottleneck (parametrization term). In the predictive setting (developed later in Section 3), the same structural logic persists: the scaffold is first selected geometrically, the projected target is constructively realized on that scaffold, and prediction is then reduced to learning a coefficient law that extends this constructive realization over the associated orbit dictionary.

2.3 Collapsing the parametrization error: the CCDS realization mechanism.

Once a scaffold has been fixed and the projection error $\varepsilon_{\text{proj}}$ controlled, the remaining approximation error lies in the realization of the compressed operator on that scaffold. The following example shows that exact span containment does not guarantee stable or efficient collapse of the parametrization error. In particular, a purely greedy innovation rule is insufficient: it has no mechanism to enforce a prescribed decrease of the residual norm, nor to prevent oscillatory selection among nearly collinear atoms. The realizability problem is therefore not solved by classical greedy pursuit alone, even when the projection error is zero. This pathology motivates deviation-controlled solvers such as our CCDS introduced here (all details below), which explicitly regulate the residual trajectory and stabilize the coefficient growth, ensuring constructive convergence even in the presence of severe geometric degeneracies of the scaffold.

Example 3 (Canonical ill–conditioned realizability example) *Consider* $\mathbb{R}^2$ *with scaffold atoms* $v_1 = (1,0)^\top$, $v_2 = (\cos\theta, \sin\theta)^\top$, *with* $0 < \theta \ll 1$, *and target* $y = (0,1)^\top$. *Since* $\operatorname{span}\{v_1, v_2\} = \mathbb{R}^2$, *the projection error vanishes and exact realization is possible:* $\alpha_1 v_1 + \alpha_2 v_2 = y$. *The unique solution is* $\alpha_2 = 1/\sin\theta$ *and* $\alpha_1 = -\cos\theta/\sin\theta$, *so* $|\alpha_1| + |\alpha_2| \sim \theta^{-1} \to \infty$ *as* $\theta \to 0$. *Thus exact realizability requires arbitrarily large coefficients with near–perfect cancellation.*

Moreover, since the angle between v_1 *and* v_2 *is* θ, *successive orthogonal projections alternate between the two atoms and the residual admits the exact closed form* $\|r_t\| = \cos^t\theta$. *In particular, for small* θ *one has* $\cos\theta = 1 - \theta^2/2 + O(\theta^4)$, *so* $\|r_t\| = \cos^t\theta \approx \exp(-t\theta^2/2)$; *hence a constant–factor reduction of the residual requires on the order of* $1/\theta^2$ *alternating steps as* $\theta \to 0$. *The corresponding coefficients satisfy* $a_{2k+1} = \cos^{2k}\theta\sin\theta$ *and* $a_{2k+2} = -\cos^{2k+1}\theta\sin\theta$: *they alternate in sign and are successively scaled by factors of* $\cos\theta$, *building up large cancelling contributions whose sum converges to the ill–conditioned exact coefficients* $\alpha_1 = -\cos\theta/\sin\theta$ *and* $\alpha_2 = 1/\sin\theta$. *This explicit geometric–series accumulation makes precise that exact realizability is achieved only through the delicate cancellation of two large, nearly equal contributions, rendering the parametrization extremely sensitive and explaining the observed coefficient blow-up in the near-collinear regime.*

Geometrically, the atoms become almost collinear. Indeed, one has $\langle r_{2k}, v_1\rangle = 0$ *and* $\langle r_{2k}, v_2\rangle = a_{2k+1} = \cos^{2k}\theta\sin\theta$, *hence* $\langle r_{2k}, v_1\rangle - \langle r_{2k}, v_2\rangle = -a_{2k+1}$. *Likewise,* $\langle r_{2k+1}, v_2\rangle = 0$ *and* $\langle r_{2k+1}, v_1\rangle = a_{2k+2} = -\cos^{2k+1}\theta\sin\theta$, *hence* $\langle r_{2k+1}, v_1\rangle - \langle r_{2k+1}, v_2\rangle = a_{2k+2}$. *Therefore the atoms are selected in strict alternation, and the correlation gap has magnitude* $|a_{t+1}| = \cos^t\theta\sin\theta = O(\sin\theta)$, *so the selection rule becomes arbitrarily sensitive to perturbations as* $\theta \to 0$.

Example 3 shows that even with zero projection error, classical greedy pursuit is insufficient to guarantee stable collapse of parametrization error. Controlled innovation is required to prevent oscillation and coefficient blow–up, which motivates our deviation–controlled solvers introduced below.

CCDS: Constructive Controlled Deviation Synthesis. Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be a real Hilbert space with norm $\|x\| = \sqrt{\langle x, x\rangle}$. For any closed subspace $Y \subset \mathcal{H}$ let Π_Y denote the orthogonal projector and

$$\rho(x, Y) \;:=\; \inf_{y\in Y} \|x - y\| \;=\; \|x - \Pi_Y x\|$$

the distance from x to Y.

Nested scaffolds. Fix a nested chain of closed subspaces $Y_0 \subset Y_1 \subset \cdots \subset Y_T \subset \mathcal{H}$ with $\dim(Y_t/Y_{t-1}) = 1$ for $t = 1, \ldots, T$.[1] For each t pick a unit *innovation direction* $u_t \in Y_t \cap Y_{t-1}^\perp$ (hence $Y_t = Y_{t-1} \oplus \operatorname{span}\{u_t\}$).

Deviation profile. Given a target $y \in \mathcal{H}$, CCDS will construct an approximation $\hat{y} \in Y_T$ via a sequence $\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_T$ with $\hat{y}_{t-1} \in Y_{t-1}$ and residuals $r_{t-1} := y - \hat{y}_{t-1}$. The key quantity controlled at step t is the *deviation of the updated residual from the* previous *scaffold*:

$$d_t \;:=\; \rho(r_t, Y_{t-1}) \qquad (t = 1, \ldots, T), \tag{12}$$

where $r_t := y - \hat{y}_t$ after the step-t update. The idea is: at each refinement step we add a new direction, so we can prescribe (within feasibility) how far the new residual remains from the preceding scaffold.

Scalar distance equation. Fix t and abbreviate $Y := Y_{t-1}$, $u := u_t \in Y^\perp$ and $r := r_{t-1}$. Consider the scalar function $\phi(\lambda) := \rho(r - \lambda u,\, Y) = \|(I - \Pi_Y)(r - \lambda u)\| = \|a - \lambda u\|$, $a := (I - \Pi_Y)r \in Y^\perp$. This is continuous, strictly convex in λ (in squared form), and its level sets can be solved exactly.

Lemma 20 (Distance equation: explicit solutions and feasibility) *Let* $Y \subset \mathcal{H}$ *be closed, let* $u \in Y^\perp$ *with* $\|u\| = 1$, *and let* $r \in \mathcal{H}$. *Set* $a := (I - \Pi_Y)r \in Y^\perp$ *and* $\alpha := \langle a, u\rangle$. *Then for any target*

1. In applications, one either enforces $\dim(Y_t/Y_{t-1}) = 1$ by construction (e.g. one atom per step), or one works with a block version. The 1-dimensional increment case is the canonical CCDS primitive.

$d \geq 0$, the equation

$$\rho(r - \lambda u,\ Y) \ =\ d \tag{13}$$

has a real solution $\lambda \in \mathbb{R}$ if and only if

$$d \ \geq\ d_{\min} \ :=\ \inf_{\lambda \in \mathbb{R}} \rho(r - \lambda u, Y) \ =\ \sqrt{\|a\|^2 - \alpha^2}. \tag{14}$$

When feasible, (13) *has exactly two solutions*

$$\lambda_{\pm} \ =\ \alpha \pm \sqrt{d^2 - d_{\min}^2}. \tag{15}$$

Moreover, the unique minimizer is $\lambda^\star = \alpha$, attaining $\rho(r - \lambda^\star u, Y) = d_{\min}$.

Proof Since $u \in Y^\perp$ and $\|u\| = 1$,

$$\begin{aligned}\rho(r - \lambda u, Y)^2 &= \|a - \lambda u\|^2 \\ &= \|a\|^2 - 2\lambda\langle a, u\rangle + \lambda^2 \\ &= (\lambda - \alpha)^2 + (\|a\|^2 - \alpha^2).\end{aligned}$$

Thus $\rho(r - \lambda u, Y) = d$ is equivalent to $(\lambda - \alpha)^2 = d^2 - (\|a\|^2 - \alpha^2)$. Real solutions exist iff $d^2 \geq \|a\|^2 - \alpha^2$, i.e. $d \geq d_{\min}$, yielding (15). The minimizer is the vertex $\lambda^\star = \alpha$ with minimal value $d_{\min}$. ■

Constructive principle (finite-horizon CCDS). Lemma 20 yields a fully constructive, stepwise principle:

Theorem 21 (Constructive controlled deviation principle) *Fix a nested chain $Y_0 \subset \cdots \subset Y_T$ with $\dim(Y_t/Y_{t-1}) = 1$ and innovation directions $u_t \in Y_t \cap Y_{t-1}^\perp$ with $\|u_t\| = 1$. Let $y \in \mathcal{H}$ and initialize $\hat{y}_0 = 0 \in Y_0$, $r_0 = y$. Fix any target deviation profile $(d_t)_{t=1}^T$ satisfying, for each t,*

$$d_t \ \geq\ d_{t,\min} \ :=\ \sqrt{\|a_t\|^2 - \alpha_t^2}, \quad a_t := (I - \Pi_{Y_{t-1}})r_{t-1},$$

$$\alpha_t := \langle a_t, u_t\rangle. \tag{16}$$

Then, there exists a (deterministic) sequence of coefficients $(\lambda_t)_{t=1}^T$ such that, defining

$$\hat{y}_t := \hat{y}_{t-1} + \lambda_t u_t \in Y_t, \qquad r_t := y - \hat{y}_t,$$

one has for every $t = 1, \ldots, T$ the exact deviation constraints $\rho(r_t, Y_{t-1}) = d_t$. The coefficients admit the explicit formula $\lambda_t = \alpha_t \pm \sqrt{d_t^2 - d_{t,\min}^2}$, with the sign chosen by any deterministic rule.

Proof Fix t. By construction $u_t \in Y_{t-1}^\perp$, and r_{t-1} is known from previous steps. The feasibility condition (16) is exactly (14) in Lemma 20 with $Y = Y_{t-1}$, $u = u_t$, $r = r_{t-1}$ and target $d = d_t$. Hence there exists λ_t solving $\rho(r_{t-1} - \lambda_t u_t, Y_{t-1}) = d_t$. Setting $\hat{y}_t := \hat{y}_{t-1} + \lambda_t u_t$ gives $r_t = y - \hat{y}_t = r_{t-1} - \lambda_t u_t$ and thus $\rho(r_t, Y_{t-1}) = d_t$. Proceed sequentially for $t = 1, \ldots, T$. ■

CCDS algorithm. We now define CCDS as an algorithmic solver that instantiates the constructive principle above on a chain induced from a dictionary.

Dictionary and ordering. Let $\mathcal{D} \subset \mathcal{H}$ be a unit-norm dictionary. Fix a deterministic ordering rule that, given a state (Y_{t-1}, r_{t-1}), selects a new atom $q_t \in \mathcal{D}$ such that $q_t \notin Y_{t-1}$. Define

$$u_t := \frac{q_t - \Pi_{Y_{t-1}} q_t}{\|q_t - \Pi_{Y_{t-1}} q_t\|} \in Y_{t-1}^{\perp}, \quad Y_t := Y_{t-1} \oplus \operatorname{span}\{u_t\}. \tag{17}$$

Let $\eta_t := \|q_t - \Pi_{Y_{t-1}} q_t\| > 0$, and let $e_t \in \mathbb{R}^t$ denote the t-th canonical basis vector. Define recursively coefficient vectors $c^{(t)} \in \mathbb{R}^t$ by $c^{(1)} := (1)$, and, for $t \geq 2$,

$$c^{(t)} := \frac{1}{\eta_t} \left(e_t - \sum_{s=1}^{t-1} \langle q_t, u_s \rangle \, \widetilde{c}^{(s,t)} \right), \tag{18}$$

where $\widetilde{c}^{(s,t)} \in \mathbb{R}^t$ is the zero-padded extension of $c^{(s)}$, namely $\widetilde{c}^{(s,t)} := (c_1^{(s)}, \ldots, c_s^{(s)}, 0, \ldots, 0)$. Then set

$$\beta^{(t)} := (\beta^{(t-1)}, 0) + \lambda_t c^{(t)} \in \mathbb{R}^t. \tag{19}$$

Proposition 22 (Equivalent selected-atom representation of CCDS output) *For each $t = 1, \ldots, T$ one has $u_t = \sum_{j=1}^{t} c_j^{(t)} q_j$, and consequently $\widehat{y}_T = \sum_{t=1}^{T} \lambda_t u_t = \sum_{j=1}^{T} \beta_j^{(T)} q_j$.*

Proof We argue by induction on t. For $t = 1$, since $Y_0 = \{0\}$, one has $u_1 = q_1$, so $c^{(1)} = (1)$ and the claim holds. Assume now that for each $s < t$, $u_s = \sum_{j=1}^{s} c_j^{(s)} q_j$. Since $\{u_s\}_{s=1}^{t-1}$ is an orthonormal basis of Y_{t-1}, we have $\Pi_{Y_{t-1}} q_t = \sum_{s=1}^{t-1} \langle q_t, u_s \rangle u_s$. Substituting the inductive representation of each u_s,

$$q_t - \Pi_{Y_{t-1}} q_t = q_t - \sum_{s=1}^{t-1} \langle q_t, u_s \rangle \sum_{j=1}^{s} c_j^{(s)} q_j.$$

After zero-padding the vectors $c^{(s)}$ to $\mathbb{R}^t$, this becomes

$$q_t - \Pi_{Y_{t-1}} q_t = \sum_{j=1}^{t} \left(e_t - \sum_{s=1}^{t-1} \langle q_t, u_s \rangle \widetilde{c}^{(s,t)} \right)_j q_j.$$

Dividing by $\eta_t = \|q_t - \Pi_{Y_{t-1}} q_t\|$ yields $u_t = \sum_{j=1}^{t} c_j^{(t)} q_j$. This proves the first claim. Finally,

$$\widehat{y}_T = \sum_{t=1}^{T} \lambda_t u_t = \sum_{t=1}^{T} \lambda_t \sum_{j=1}^{t} c_j^{(t)} q_j = \sum_{j=1}^{T} \beta_j^{(T)} q_j,$$

by the recursive definition of $\beta^{(t)}$. This proves the result. ■

Definition 23 (CCDS algorithm with explicit atom-dictionary output) *Fix a horizon T and a deviation profile rule that provides targets $d_t \geq 0$ at each step. Given an input target $y \in H$, define $\widehat{y}_0 := 0$, $Y_0 := \{0\}$, $r_0 := y$, and initialize the dictionary-coefficient vector by $\beta^{(0)} := 0 \in \mathbb{R}^0$. For $t = 1, \ldots, T$:*

1. *Select $q_t \in \mathcal{D}$ deterministically and form u_t, Y_t by (17).*
2. *Compute $a_t := (I - \Pi_{Y_{t-1}}) r_{t-1}$, and $\alpha_t := \langle a_t, u_t \rangle$. Define $d_{t,\min} := \sqrt{\|a_t\|^2 - \alpha_t^2}$.*
3. *Choose a target deviation $d_t \geq d_{t,\min}$ (feasibility).*

4. *Set the CCDS coefficient by the explicit root* $\lambda_t := \alpha_t - \sqrt{d_t^2 - d_{t,\min}^2}$, *(or the* + *root, or any deterministic root-selection rule).*
5. *Update* $\widehat{y}_t := \widehat{y}_{t-1} + \lambda_t u_t$ *and* $r_t := y - \widehat{y}_t$.
6. Dictionary-coordinate update. *Let* $c^{(t)}$ *and* $\beta^{(t)}$ *defined as in (18) and (19) respectively.*

The intrinsic CCDS output is $\widehat{y}_T \in Y_T$. *Equivalently, the same output admits the explicit selected-atom representation*

$$\widehat{y}_T = \sum_{j=1}^{T} \beta_j^{(T)} q_j \in \operatorname{span}\{q_1, \ldots, q_T\} \subset \operatorname{span} \mathcal{D}.$$

Corollary 24 (Finite-dimensional feasibility obstruction for CCDS) *According to Lemma 20, if for some step* t *the prescribed target deviation satisfies* $d_t < d_{t,\min}$, *then there exists no scalar* $\lambda_t \in \mathbb{R}$ *such that* $\rho(r_{t-1} - \lambda_t u_t, Y_{t-1}) = d_t$. *Consequently, any deviation profile violating the stepwise feasibility thresholds is unrealizable in the fixed ambient realization space.*

The following example shows that in finite dimension, scaffold exhaustion can force u_t to be nearly orthogonal to the residual direction a, producing a strictly positive feasibility threshold $d_{\min}$. Certain deviation targets d_t are then geometrically impossible.

Example 4 (Finite-dimensional infeasibility and necessity of dimension lifting) *Consider a single CCDS step with ambient Hilbert space* $H = \mathbb{R}^2$, *current scaffold subspace* $Y_{t-1} = \{0\}$, *one data point* $x_1 = e_1 = (1, 0)$, *target operator* $A = I$, *so the residual at step* $t-1$ *is* $r_{t-1}(x_1) = Ax_1 = e_1$, *and off–scaffold component* $a := (I - \Pi_{Y_{t-1}})r_{t-1}(x_1) = e_1$, *with* $\|a\| = 1$.

Assume that the current scaffold dictionary is exhausted in the sense that the only available innovation direction is $u_t = e_2 = (0, 1)$, *so that the update direction is forced to lie in the one–dimensional subspace* $\mathcal{U}_t = \operatorname{span}\{e_2\} \subset Y_{t-1}^{\perp}$. *Hence, Lemma 20 (feasibility bound) gives*

$$d_{\min} = \sqrt{\|a\|^2 - \langle a, u_t \rangle^2} = \sqrt{1 - 0} = 1,$$

so for all target deviations $d_t < 1$ *the CCDS step is infeasible in the current finite–dimensional realization space: no choice of* λ *can achieve the target deviation because the available innovation direction is orthogonal to the off–scaffold component* a.

Adding one new orthogonal dimension restores feasibility by allowing a new innovation direction with controlled alignment to a. This justifies the need for dimension lifting in order to guarantee universal deviation feasibility and, consequently, collapse of the parametrization error in CCDS.

Theorem 25 (One-step feasibility by dimension lifting) *Let* H *be a real Hilbert space,* $Y \subset H$ *a closed subspace, and* $r \in H$. *Define* $a := (I - \Pi_Y)r \in Y^{\perp}$. *Fix a target deviation* d *satisfying* $0 \le d \le \|a\|$. *Then there exists a Hilbert space* $\widetilde{H} := H \oplus \mathbb{R}$, *an isometric embedding* $\iota : H \to \widetilde{H}$, *and a unit vector* $u \in (\iota(Y))^{\perp} \subset \widetilde{H}$ *such that the CCDS distance equation* $\rho(\iota(r) - \lambda u, \iota(Y)) = d$ *admits a real solution* $\lambda \in \mathbb{R}$. *Moreover, one can choose* u *so that the minimal achievable deviation is exactly* $d_{\min}(u) = d$.

Proof Define $\widetilde{H} := H \oplus \mathbb{R}$ with inner product

$$\langle (x, s), (y, t) \rangle_{\widetilde{H}} := \langle x, y \rangle_H + st,$$

and the isometric embedding $\iota(x) = (x, 0)$. Then $\iota(Y)$ is a closed subspace and

$$(\iota(Y))^{\perp} = Y^{\perp} \oplus \mathbb{R}.$$

Let

$$\widetilde{r} := \iota(r), \qquad \widetilde{a} := (I - \Pi_{\iota(Y)})\widetilde{r} = (a, 0),$$

so that $\|\widetilde{a}\| = A$. If $A = 0$, then $d = 0$ and the claim is trivial. Assume $A > 0$ and define $e_1 := \frac{\widetilde{a}}{A}$, $e_2 := (0, 1)$. Then $e_1, e_2 \in (\iota(Y))^\perp$ are orthonormal. Choose $\theta \in [0, \pi/2]$ such that $\sin\theta = \frac{d}{A}$. Define $u := \cos\theta\, e_1 + \sin\theta\, e_2$. Then $u \in (\iota(Y))^\perp$ and $\|u\| = 1$. Moreover, $\alpha := \langle \widetilde{a}, u\rangle_{\widetilde{H}} = A\cos\theta$.

By the CCDS feasibility formula (Lemma 3.3),

$$d_{\min}(u) = \sqrt{\|\widetilde{a}\|^2 - \alpha^2} = \sqrt{A^2 - A^2\cos^2\theta} = A\sin\theta = d.$$

Hence $d \geq d_{\min}(u)$, and the equation $\rho(\widetilde{r} - \lambda u, \iota(Y)) = d$ admits a real solution λ. This completes the proof. ■

Remark 26 *To enforce the same target deviation level d simultaneously for M sample-wise residuals $\{r^{(j)}\}_{j=1}^M$ at a given CCDS step, it suffices to iterate the one–step dimension–lifting construction of Theorem 25 for each residual. This yields an isometric embedding into an ambient Hilbert space of dimension at most $\dim(H) + M$ in which the deviation constraint $\rho(r^{(j)} - \lambda u, Y) = d$ is feasible for all $j = 1, \ldots, M$.*

Resolution of Example 4 by lifting one dimension. Embed H isometrically into $\widetilde{H} = \mathbb{R}^3$ by $x \mapsto (x, 0)$, and extend the orthogonal complement with the new basis vector e_3. Choose the lifted innovation direction $\tilde{u}_t = \cos\theta\, e_1 + \sin\theta\, e_3, \qquad \sin\theta = d_t$. Then $\tilde{u}_t \in (\iota(Y_{t-1}))^\perp$, $\|\tilde{u}_t\| = 1$, and $\langle \tilde{a}, \tilde{u}_t\rangle = \cos\theta$, $d_{\min}(\tilde{u}_t) = \sqrt{1 - \cos^2\theta} = \sin\theta = d_t$. Hence the CCDS equation

$$\rho\big(\iota(r_{t-1}(x_1)) - \lambda\tilde{u}_t,\, \iota(Y_{t-1})\big) = d_t$$

admits a real solution, and the deviation d_t becomes feasible after adding a single orthogonal dimension.

Remark 27 *There is no direct analogue of the CCDS lifting theorem in the fixed-V scaffold-selection setting. In CCDS, lifting can repair infeasibility caused by a poorly aligned fixed innovation direction. In scaffold selection, however, once V is fixed, the method already optimizes over all admissible directions $S = V \cap U^\perp$. Thus there is no further local directional repair inside the same V left to exploit*

Greedy residual-alignment as the boundary case of CCDS. The CCDS mechanism constructs updates of the form $r_t = r_{t-1} - \lambda_t u_t$, where $u_t \in Y_{t-1}^\perp$ is an innovation direction and λ_t is selected by solving a one–dimensional distance equation $\rho(r_{t-1} - \lambda_t u_t, Y_{t-1}) = d_t$, for a feasible deviation target $d_t \geq d_{t,\min}$.

Consider the trivial scaffold $Y = \{0\}$. Then $\Pi_Y = 0$ and $\rho(r, Y) = \|r\|$. Innovation directions coincide with dictionary atoms: $u = q \in \mathcal{D}$. The CCDS distance equation reduces to $\|r_{t-1} - \lambda_t q_t\| = d_t$, with feasibility interval

$$d_t \in \Big[\min_\lambda \|r_{t-1} - \lambda_t q_t\|,\ \|r_{t-1}\|\Big].$$

Lemma 28 is a classical fact from Hilbert space geometry (orthogonal projection onto a one–dimensional subspace). We include a short proof for completeness and to make the paper self-contained.

Lemma 28 (Minimal–deviation update) *For fixed $r \in \mathcal{H}$ and unit $q \in \mathcal{H}$, the function $\lambda \mapsto \|r - \lambda q\|^2$ attains its unique minimum at $\lambda^\star = \langle r, q\rangle$, with minimum value $d_{t,min} = \|r\|^2 - |\langle r, q\rangle|^2$.*

Proof Let $r \in H$ and let $q \in H$ with $\|q\| = 1$. For $\lambda \in \mathbb{R}$ we expand

$$\|r - \lambda q\|^2 = \langle r - \lambda q, r - \lambda q\rangle = \|r\|^2 - 2\lambda\langle r, q\rangle + \lambda^2.$$

This quadratic polynomial in λ attains its minimum at $\lambda^\star = \langle r, q\rangle$. Evaluating at this value gives $\min_{\lambda\in\mathbb{R}} \|r - \lambda q\|^2 = \|r\|^2 - \langle r, q\rangle^2$. Hence the optimal coefficient is $\lambda^\star = \langle r, q\rangle$ and the minimal deviation is $d_{\min} = \sqrt{\|r\|^2 - \langle r, q\rangle^2}$. ■

Thus, choosing the deviation target $d_t = d_{t,\min}$ and selecting the corresponding minimizing root yields the update $r_t = r_{t-1} - \langle r_{t-1}, q_t\rangle q_t$, and the CCDS framework collapses into the Matching Pursuit algorithm. We refer to this realization regime as the *Matching Pursuit regime (MP)*. Thus, MP is defined as the specialization of CCDS to the trivial scaffold $Y = \{0\}$ (no nested subspace scaffolds) with minimal deviation targets $d_t = d_{t,\min}$, in which the general deviation equation reduces to greedy contraction and geometric residual decay.

Remark 29 *MP therefore appears as the boundary case of a deviation–equation solver on nested-subspaces orbit scaffolds.*

So we have.

Proposition 30 *MP is exactly CCDS restricted to the trivial scaffold $Y = \{0\}$ with minimal–deviation target selection.*

Remark 31 ***Why allow*** $d_t > d_{\min}$***?*** *The freedom to choose $d_t > d_{\min}$ is not merely numerical. It is what distinguishes CCDS from the extremal greedy choice $d_t = d_{\min}$, which forces the largest one-step contraction. Allowing $d_t > d_{\min}$ makes it possible to prescribe non-extremal feasible residual profiles, improves numerical stability by avoiding boundary-root behavior in the distance equation, and helps prevent repeated near-collinearity in successive innovation directions. In this sense, the flexibility $d_t > d_{\min}$ is one of the main sources of the robustness of CCDS in weak-alignment or degenerate regimes, as shown in our experiments, see Section 4.*

Relation between scaffold selection and CCDS realization mechanism. The scaffold selection mechanism is the subspace-level counterpart of the CCDS philosophy introduced later. In scaffold selection, the controlled object is the projection-error functional, and the admissible moves are nested one-dimensional scaffold expansions within the orbit geometry induced by the data. In CCDS realization, the controlled object is the parametrization residual, and the admissible moves are coefficient updates along innovation directions. Thus the two mechanisms are parallel deviation-control theories for the two main bottlenecks of the framework: scaffold geometry and constructive realization. We present scaffold selection first because the framework is logically geometric before constructive: one first selects the support on which the target can be projected, then constructs a realizer on that support.

2.4 Our realizability framework in practice.

Local-chart diagnostics before scaffold selection. Before applying scaffold selection, the next result gives a preliminary anchor-level diagnostic: local-chart defects $\|J_{k(i)} - \widetilde{A}_{D_{k(i)}}\Pi_{E_{k(i)}}\|$ measure how well the blockwise interpolants align with the scaffold geometry. Thus these defects provide a preliminary geometric signal for which blocks are well aligned before the scaffold-selection procedure is applied.

Theorem 32 (Local-chart diagnostic before scaffold selection) *Let $\Xi_{\text{sub}} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d\times\mathbb{R}^C$, and let $P = \{D_1, \dots, D_K\}$ be a partition of Ξ_{sub}. Write $D_k = \{(x_{k,1}, y_{k,1}), \dots, (x_{k,s_k}, y_{k,s_k})\}$ and set $X_k = [x_{k,1} \ \cdots \ x_{k,s_k}]$ and $Y_k = [y_{k,1} \ \cdots \ y_{k,s_k}]$. Assume that, for every k, $\sum_{j=1}^{s_k} c_j x_{k,j} = 0$*

implies $\sum_{j=1}^{s_k} c_j y_{k,j} = 0$. *Choose depths* $L_{k,j}$ *such that* $z_{k,j} := W^{(L_{k,j})} x_{k,j}$ *are linearly independent, and set* $Z_k = [z_{k,1} \ \cdots \ z_{k,s_k}]$, $E_k = \mathrm{span}\{z_{k,1}, \ldots, z_{k,s_k}\}$, *and* $J_k := Y_k Z_k^\dagger$. *Let* U_P *be the orbit scaffold generated by* $\{J_k W^{(m)} x_{k,j} : 1 \le k \le K,\ 1 \le j \le s_k,\ 0 \le m \le d-1\}$. *Then for all* $1 \le k \le K$ *and* $1 \le j \le s_k$, *we have the anchor-level scaffold diagnostic*

$$\|(I - \Pi_{U_P} y_{k,j})\| \le \frac{\|Y_k\|_F}{\sigma_{\min}^+(X_k)} \operatorname{dist}(x_{k,j}, E_k) + \delta_k \|x_{k,j}\|,$$

where $\delta_k := \|J_k - Y_k X_k^\dagger \Pi_{E_k}\|$.

Proof Fix a block D_k and set $A_k := Y_k X_k^\dagger$. The compatibility assumption gives $\ker X_k \subseteq \ker Y_k$, as in the proof of Theorem 8 we have $Y_k = Y_k X_k^\dagger X_k$. Therefore $A_k x_{k,j} = y_{k,j}$ for every j. Also,

$$\|A_k\| = \|Y_k X_k^\dagger\| \le \|Y_k\|_F \|X_k^\dagger\| = \frac{\|Y_k\|_F}{\sigma_{\min}^+(X_k)}.$$

Since $m = 0$ is allowed in the orbit scaffold, $J_k x_{k,j} \in U_P$. Hence $\operatorname{dist}(y_{k,j}, U_P) \le \|y_{k,j} - J_k x_{k,j}\|$. Using $y_{k,j} = A_k x_{k,j}$ and inserting $A_k \Pi_{E_k} x_{k,j}$, we get

$$\|y_{k,j} - J_k x_{k,j}\| \le \|A_k (I - \Pi_{E_k}) x_{k,j}\| + \|(J_k - A_k \Pi_{E_k}) x_{k,j}\|.$$

Thus

$$\|(I - \Pi_{U_P} y_{k,j})\| = \operatorname{dist}(y_{k,j}, U_P) \le \frac{\|Y_k\|_F}{\sigma_{\min}^+(X_k)} \operatorname{dist}(x_{k,j}, E_k) + \delta_k \|x_{k,j}\|.$$

■

Combining Theorem 32 and Theorem 8 yields.

Corollary 33 *We have* $\left(\sum_{k=1}^K \sum_{j=1}^{s_k} \|(I - \Pi_{U_P} y_{k,j})\| \right) / \sum_{k=1}^K s_k \le \mathcal{S}(P)$, *where*

$$\mathcal{S}(P) := \frac{\sum_{k=1}^K \sum_{j=1}^{s_k} \frac{\|Y_k\|_F}{\sigma_{\min}^+(X_k)} \left(\operatorname{dist}(x_{k,j}, E_k) + \frac{\|x_{k,j}\|}{\sigma_{\min}(Z_k)} \left(\sum_{\ell=1}^{s_k} \|x_{k,\ell} - z_{k,\ell}\|^2 \right)^{1/2} \right)}{\sum_{k=1}^K s_k}.$$

Observable guidance for the choice of block size. The diagnostic $\mathcal{S}(P)$ provides a preselection criterion for the block structure before scaffold selection is applied. It favors partitions whose local charts are well aligned with the anchor geometry. Smaller blocks often improve this diagnostic, but increase the number of blocks K and therefore the nominal dictionary size. This creates a tradeoff between local stability and scaffold complexity. The next observation shows how vanishing orbit dynamics can compensate for this increase by reducing the number of active orbit depths. Although the formal scaffold is built from the finite set of depths $0 \le m \le d-1$, it represents the full orbit family $\{J_k W^{(m)} x : m \ge 0,\ 1 \le k \le K\}$: by Lemma 3, every deeper iterate $W^{(m)}$ is a scalar multiple of one of these d residue classes. Thus the vanishing regime does not change the formal scaffold dimension; it controls how many orbit depths remain effectively active.

Vanishing orbit dynamics as implicit regularization. Assume that W has weights $w_0, \ldots, w_{d-1}$ and set $\rho := \prod_{i=0}^{d-1} |w_i|$. In the vanishing regime $\rho < 1$, deep orbit iterates become negligible. More precisely, for every $z \in \mathbb{R}^d$ and every $m \ge 1$,

$$\|W^{(m)} z\| \le \rho^{\lfloor m/d \rfloor} C_z, \qquad C_z := \max_{0 \le r \le d-1} \|W^{(r)} z\|. \tag{20}$$

Proof of (20) deferred to the Appendix A. Thus, in the vanishing regime, the orbit dynamics provide an implicit regularization mechanism: high-depth orbit contributions are exponentially damped

without adding an explicit penalty term or stochastic regularizer such as dropout. The effect is driven entirely by the geometry of the weighted shift and induces a finite effective memory horizon.

Fix $\varepsilon > 0$. For $(x, y) \in D_k$, define

$$M_{\text{eff}}(x;\varepsilon) := \min\{M \in \mathbb{N} : \|J_k W^{(m)} x\| \le \varepsilon \text{ for all } m \ge M\}.$$

If $\rho < 1$, then

$$M_{\text{eff}}(x;\varepsilon) \le d \left\lceil \frac{\log(\varepsilon/C_{k,x})}{\log \rho} \right\rceil, \qquad C_{k,x} := \max_{0 \le r \le d-1} \|J_k W^{(r)} x\|. \tag{21}$$

Proof of (21) deferred to the Appendix A. Hence the active dictionary is governed not by the nominal depth d, but by the effective horizon. We measure this pointwise by

$$r_{\text{eff}}(x;\varepsilon) := \dim \operatorname{span}\{J_k W^{(m)} x : 0 \le m < M_{\text{eff}}(x;\varepsilon)\},$$

and at dataset level by

$$r_{\text{eff}}(\Xi;\varepsilon) := \frac{1}{|\Xi|} \sum_{(x_i,y_i)\in\Xi} r_{\text{eff}}(x_i;\varepsilon).$$

Thus, in the vanishing regime, smaller effective horizon means fewer active orbit directions. This is the structural simplicity we reward together with local-chart stability.

Realizability phases. The realizability pipeline has two phases. Phase 1 addresses the scaffold bottleneck by selecting an orbit scaffold with small projection error. Phase 2 addresses the parametrization bottleneck by constructing a CCDS realizer on the selected scaffold. Let $\Xi_{\text{train}} \subseteq \Xi$ denote the training set.

Phase 1: vanishing-aware scaffold discovery. Phase 1 searches over partitions and shifts using three quantities: $\mathcal{S}(P)$, K, $\prod_{i=0}^{d-1} |w_i|$. Small $\mathcal{S}(P)$ indicates good local-chart alignment; small K controls the number of blocks; and $\prod_{i=0}^{d-1} |w_i| < 1$ enforces vanishing orbit dynamics and controls the induced effective horizons through (21). We therefore retain candidates for which $\mathcal{S}(P)$ is small, K remains moderate and the learned shift is vanishing. The deviation-controlled scaffold-selection procedure is then applied to collapse the projection error.

Phase 2: CCDS realization. Once the scaffold has been selected, Phase 2 constructs a realizer on that scaffold using CCDS:

$$f_{\text{CCDS}}(x) := \sum_{k=0}^{K-1} \sum_{m=0}^{d-1} a_{k,m}^{\text{eff}}(x)\, J_k W^{(m)} x, \qquad (x, y) \in \Xi_{\text{train}}.$$

After scaffold selection, CCDS sparsity further controls how many of the available block-depth atoms are actually used in each realization as established in Subsection 2.7 below.

2.5 Situating our operator-theoretic realizability framework in Machine Learning.

A conceptual motivation for our deviation-control framework comes from Bernstein lethargy theory in approximation theory (Bernstein, 1954; Shapiro, 1964): along nested approximation families, error decay can be arbitrarily slow, so no universal collapse rate should be expected without additional structure. Accordingly, we do not organize scaffold selection or realization around an a priori decay law. Instead, we treat the residual profile itself as an object to be prescribed and controlled. This principle appears twice in our theory: at the scaffold level, through feasibility intervals and prescribed projection-error drops; and at the realization level, through CCDS, where prescribed deviation constraints are enforced step by step. This places our framework in dialogue with two residual-based traditions: discrepancy principles in inverse problems and greedy pursuit methods in

approximation theory. In Morozov-type discrepancy principles, a target residual is typically used as a stopping or regularization criterion (Morozov, 1966; Bakushinsky and Kokurin, 2004). By contrast, our residual schedule is internalized into the construction itself: scaffold selection realizes feasible projection-error trajectories, while CCDS realizes feasible deviation profiles by solving explicit one-dimensional distance equations. Thus the residual path is not merely monitored; it is prescribed and constructively realized. Greedy methods such as Matching Pursuit, Orthogonal Matching Pursuit, and their Banach space extensions also proceed through residual reduction (Mallat and Zhang, 1993; Pati et al., 1993; Tropp and Gilbert, 2007; DeVore and Temlyakov, 1996; Temlyakov, 2011). However, their updates are typically defined by an implicit optimization, projection, or descent rule. Our framework instead recasts the update as the constructive realization of a prescribed residual profile. Classical greedy MP appears as a boundary case of the CCDS mechanism, just as greedy maximal projection collapse appears as a boundary case of the scaffold-selection mechanism.

Consequently, the novelty of our deviation-control framework is threefold: it treats the residual path as a prescribed constraint rather than an observed byproduct; it replaces implicit optimization with explicit operator-theoretic synthesis; and it provides an unified theory that recovers standard greedy algorithms as a specialized subset of a broader deviation-control class governing scaffold selection and CCDS. This yields a constructive and verifiable framework for operator realizability on data-induced scaffolds, with feasible residual trajectories built directly into the theory.

2.6 Stability under relaxed block interpolation

Exact interpolation gives the ideal certificate: for each block, the exact interpolant given by either the one-step update construction given by the proof of Exact Interpolation Theorem 4 or via pseudoinverse in Theorem 7 with $J_k = Y_k Z_k^\dagger = Y_k(Z_k^\top Z_k)^{-1} Z_k^\top$, recovers the anchor outputs exactly and anchors the scaffold on which the two bottlenecks collapse. In our experiments, however, the more useful object we found is the regularized interpolant $J_{k,\lambda} := Y_k(Z_k^\top Z_k + \lambda I)^{-1} Z_k^\top$, $\lambda > 0$, which is cheaper and better conditioned on noisy or nearly singular blocks. The next theorem is therefore crucial in our theory and serves as a bridge from the clean theoretical machinery to the implemented method: replacing the exact block maps by their relaxed regularized pseudoinverse versions preserves the scaffold bottleneck and perturbs the CCDS bottleneck only continuously. Thus the theoretical collapse mechanism survives the relaxed interpolation regime used in practice.

Theorem 34 (Stability of the scaffold/CCDS bottlenecks under relaxed interpolation) *Let $\Xi = \{(x_l, y_l)\}_{l=1}^N \subset \mathbb{R}^d \times \mathbb{R}^C$, let $A : \mathbb{R}^d \to \mathbb{R}^C$ be linear, and fix orbit data $(W, \Xi_{\mathrm{sub}}, \{D_k\}_{k=1}^K)$. For each block write Z_k for the orbit matrix and Y_k for the corresponding output matrix as defined in (5), let $J_k = Y_k Z_k^\dagger$ be the reference minimal-norm exact interpolant. Let G be the associated atom matrix and set $U := \mathrm{range}(G)$, $r := \mathrm{rank}(G)$, $\gamma := \sigma_r(G) > 0$. Assume the reference CCDS realizer f_{CCDS} on U has a finite prescribed path with positive feasibility margin and selected raw directions uniformly bounded away from zero. For every admissible relaxed block family $J'_k = Y_k R'_k$, where R'_k is a regularized pseudoinverse factor for the same block data, G' the corresponding atom matrix and $U' := \mathrm{range}(G')$, there exist $\delta_0 > 0$ such that if $\|G' - G\|_2 < \delta_0$, then $U' = U$. Consequently,*

$$\varepsilon_{\mathrm{proj}}(A, U'; \Xi) = \varepsilon_{\mathrm{proj}}(A, U; \Xi).$$

Moreover, transporting the reference CCDS path to the perturbed atoms, with the same prescribed deviations and the same root branches, gives a CCDS realizer f'_{CCDS} on U' satisfying

$$\varepsilon_{\mathrm{param}}(A, U', f'_{\mathrm{CCDS}};\ \Xi) \le \varepsilon_{\mathrm{param}}(A, U, f_{\mathrm{CCDS}};\ \Xi) + C_{\mathrm{param}} \|G' - G\|_2.$$

The same conclusion follows by taking J_k the zero-initialized one-step update constructive exact interpolant of Theorem 4. Thus, within the one-step update/minimal-norm pseudoinverse exact interpolant family constructed in Theorem 4/Theorem 7, the projection bottleneck is invariant under relaxed interpolation and the constructive CCDS bottleneck is locally Lipschitz stable.

Proof For each $1 \le k \le K$, let $D_k = \{(x_{k,1}, y_{k,1}), \dots, (x_{k,|D_k|}, y_{k,|D_k|})\} \subseteq \Xi_{\mathrm{sub}}$ and choose depths $L_{k,1}, \dots, L_{k,|D_k|}$ such that $y_{k,j} = J_k W^{(L_{k,j})} x_{k,j}$ for all $(x_{k,j}, y_{k,j}) \in D_k$. Let $\mathcal{I} = \{(k, m, j) : 1 \le k \le K,\ 0 \le m \le d-1,\ 1 \le j \le |D_k|\}$ be the fixed finite atom index set of the corresponding scaffold as defined in (10). Write the reference atoms as $g_i = J_k W^{(m)} x_{k,j}$, with $i = (k, m, j) \in \mathcal{I}$, and the perturbed atoms as $g'_i = J'_k W^{(m)} x_{k,j}$. Let $G = [g_i]_{i \in \mathcal{I}}$, $G' = [g'_i]_{i \in \mathcal{I}}$, $U = \mathrm{range}(G)$, $U' = \mathrm{range}(G')$, and $\eta := \|G' - G\|_2$. Also set $V := \sum_{k=1}^{K} \mathrm{range}(Y_k)$.

Since $J'_k = Y_k R'_k$, every perturbed atom g'_i belongs to $\mathrm{range}(Y_k)$, hence $U' \subseteq V$. On the other hand, suppose $y \in V$, then there exist scalars $(\beta_{k,j})_{1 \le k \le K, 1 \le j \le |D_k|}$ such that $y = \sum_{k=1}^{K} \sum_{j=1}^{|D_k|} y_{k,j} = \sum_{k=1}^{K} \sum_{j=1}^{|D_k|} J_k W^{(L_{k,j})} x_{k,j} \in U$. Therefore $V \subseteq U$. Conversely, every reference atom is of the form $g_i = Y_k Z_k^{\dagger} W^{(m)} x_{k,j}$, hence belongs to V. Thus $U = V$, and consequently $U' \subseteq U$.

Let $r = \mathrm{rank}(G) = \dim U$ and $\gamma = \sigma_r(G) > 0$. Choose $\delta_0 < \gamma/2$. If $\eta < \delta_0$, then Weyl's singular-value perturbation inequality (see Horn and Johnson, 2012) gives $\sigma_r(G') \ge \sigma_r(G) - \eta > \gamma/2 > 0$. Hence $\mathrm{rank}(G') \ge r$. Since $U' \subseteq U$ and $\dim U = r$, also $\mathrm{rank}(G') \le r$. Therefore $\mathrm{rank}(G') = r$, and so $U' = U$. It follows immediately that $\Pi_{U'} = \Pi_U$, and hence $\varepsilon_{\mathrm{proj}}(A, U'; \Xi) = \varepsilon_{\mathrm{proj}}(A, U; \Xi)$.

It remains to prove the parametrization bound. Since $U' = U$, the projected targets are identical: $b_l := \Pi_U A x_l = \Pi_{U'} A x_l$. Fix l and suppress the index l. Let $r_0 = b_l$, and let r_t, $t = 0, \dots, T$, be the residuals of the reference CCDS path of the CCDS realizer f_{CCDS}. At step t, the reference path uses a prescribed atom label α_t, prescribed deviation d_t, and prescribed root branch $\varepsilon_t \in \{-1, 1\}$. Let $Y_{t-1} := \mathrm{span}\{u_1, \dots, u_{t-1}\}$, $P_{t-1} := \Pi_{Y_{t-1}}$, $q_t := (I - P_{t-1}) g_{i_t}$, $u_t := q_t / \|q_t\|$, $a_t := (I - P_{t-1}) r_{t-1}$, $\alpha_t := \langle a_t, u_t \rangle$, and $d^2_{t,\min} := \|a_t\|^2 - \alpha_t^2$. The CCDS coefficient is $\lambda_t = \alpha_t + \varepsilon_t \sqrt{d_t^2 - d^2_{t,\min}}$, and the residual update is $r_t = r_{t-1} - \lambda_t u_t$.

We transport exactly this finite prescription to the perturbed atoms: the same labels i_t, the same deviations d_t, and the same root branches ε_t. Thus g_{i_t} is replaced by the corresponding g'_{i_t}, and all perturbed quantities Y'_{t-1}, P'_{t-1}, q'_t, u'_t, a'_t, α'_t, λ'_t, and r'_t are defined by the same formulas.

We prove by induction that $\|r'_t - r_t\| \le B_t \eta$ and $\|u'_t - u_t\| \le A_t \eta$ for constants depending only on the reference trajectory. The base residual estimate is $\|r'_0 - r_0\| = 0$, because $b'_l = b_l$. Assume the estimates are known up to step $t-1$. Since the directions $u_1, \dots, u_{t-1}$ and $u'_1, \dots, u'_{t-1}$ are orthonormal along the CCDS innovation construction, if $U_{t-1} := [u_1 \ \cdots \ u_{t-1}]$ and $U'_{t-1} := [u'_1 \ \cdots \ u'_{t-1}]$, then $P_{t-1} = U_{t-1} U_{t-1}^{\top}$ and $P'_{t-1} = U'_{t-1} (U'_{t-1})^{\top}$. Add and subtract $U_{t-1} (U'_{t-1})^{\top}$: $P'_{t-1} - P_{t-1} = (U'_{t-1} - U_{t-1})(U'_{t-1})^{\top} + U_{t-1}\big((U'_{t-1})^{\top} - U_{t-1}^{\top}\big)$. Taking operator norms, $\|P'_{t-1} - P_{t-1}\|_{\mathrm{op}} \le \|U'_{t-1} - U_{t-1}\|_2 \|(U'_{t-1})^{\top}\|_2 + \|U_{t-1}\|_2 \|(U'_{t-1})^{\top} - U_{t-1}^{\top}\|_2$. Since U_{t-1} and U'_{t-1} have orthonormal columns, $\|U_{t-1}\|_2 = \|U'_{t-1}\|_2 = 1$. Thus $\|P'_{t-1} - P_{t-1}\|_{\mathrm{op}} \le 2\|U'_{t-1} - U_{t-1}\|_2 \le 2\|U'_{t-1} - U_{t-1}\|_F = 2\left(\sum_{s=1}^{t-1} \|u'_s - u_s\|^2\right)^{1/2}$. Now the induction hypothesis says that for all previous steps $s < t$, $\|u'_s - u_s\| \le A_s \eta$ for some constants A_s depending only on the reference trajectory. Since there are only finitely many steps, define $A^{\max}_{t-1} := \max_{1 \le s < t} A_s$. Then $\left(\sum_{s=1}^{t-1} \|u'_s - u_s\|^2\right)^{1/2} \le \sqrt{t-1}\, A^{\max}_{t-1} \eta$. Hence

$$\|P'_{t-1} - P_{t-1}\|_{\mathrm{op}} \le 2\sqrt{t-1}\, A^{\max}_{t-1} \eta = C_{P,t} \eta. \tag{22}$$

Since corresponding atom columns differ by at most η, we have $\|g'_{i_t} - g_{i_t}\| \le \eta$. Therefore $q'_t - q_t = (I - P'_{t-1})(g'_{i_t} - g_{i_t}) + (P_{t-1} - P'_{t-1}) g_{i_t}$, and so

$$\|q'_t - q_t\| \le \eta + C_{P,t} \eta \|g_{i_t}\| \le C_{q,t} \eta. \tag{23}$$

By the nondegeneracy assumption, $\|q_t\| \ge \nu > 0$. After decreasing δ_0, we may assume $\|q'_t - q_t\| \le \nu/2$. Then by the reverse triangle inequality, $\|q'_t\| \ge \|q_t\| - \|q'_t - q_t\| \ge \nu - \nu/2 = \nu/2$. So both q and q'_t stay away from zero. Now prove the Lipschitz bound. We write $\frac{q'}{\|q'\|} - \frac{q}{\|q_t\|} = \frac{q'_t - q_t}{\|q'_t\|} + q_t \left(\frac{1}{\|q'_t\|} - \frac{1}{\|q_t\|}\right)$. Taking norms, $\left\| \frac{q'_t}{\|q'_t\|} - \frac{q_t}{\|q_t\|} \right\| \le \frac{\|q'_t - q_t\|}{\|q'_t\|} + \|q_t\| \left| \frac{1}{\|q'_t\|} - \frac{1}{\|q_t\|} \right|$. For the second term, $\left| \frac{1}{\|q'_t\|} - \frac{1}{\|q_t\|} \right| = \frac{\big|\|q_t\| - \|q'_t\|\big|}{\|q'_t\| \|q_t\|}$. By the reverse triangle inequality, $\big|\|q_t\| - \|q'_t\|\big| \le \|q_t - q'_t\|$. Therefore

$\|q_t\|\left|\frac{1}{\|q_t'\|}-\frac{1}{\|q_t\|}\right| \le \|q_t\|\frac{\|q_t-q_t'\|}{\|q_t'\|\|q_t\|} = \frac{\|q_t-q_t'\|}{\|q_t'\|}$. So $\left\|\frac{q_t'}{\|q_t'\|}-\frac{q_t}{\|q_t\|}\right\| \le \frac{2\|q_t'-q_t\|}{\|q_t'\|}$. Since $\|q_t'\| \ge \nu/2$, we get $\frac{1}{\|q_t'\|} \le \frac{2}{\nu}$. Hence $\|u_t'-u_t\| = \left\|\frac{q_t'}{\|q_t'\|}-\frac{q_t}{\|q_t\|}\right\| \le \frac{4}{\nu}\|q_t'-q_t\|$. And by (23) we have

$$\|u_t'-u_t\| \le A_t\eta. \tag{24}$$

Now compare the CCDS coefficients. We have $a_t'-a_t = (I-P_{t-1}')(r_{t-1}'-r_{t-1})+(P_{t-1}-P_{t-1}')r_{t-1}$. Since the reference residuals are bounded by (22) it follows that

$$\|a_t'-a_t\| \le C_{a,t}(\|r_{t-1}'-r_{t-1}\|+\eta). \tag{25}$$

Also $\alpha_t'-\alpha_t = \langle a_t'-a_t, u_t'\rangle + \langle a_t, u_t'-u_t\rangle$, hence

$$|\alpha_t'-\alpha_t| \le C_{\alpha,t}(\|r_{t-1}'-r_{t-1}\|+\eta). \tag{26}$$

Recall $d_{t,\min}^2 = \|a_t\|^2-\alpha_t^2$. Similarly, $(d_{t,\min}')^2 = \|a_t'\|^2-(\alpha_t')^2$. Then, $(d_{t,\min}')^2-d_{t,\min}^2 = \big(\|a_t'\|^2-\|a_t\|^2\big)-\big((\alpha_t')^2-\alpha_t^2\big)$. Taking absolute values, $\big|(d_{t,\min}')^2-d_{t,\min}^2\big| \le \big|\|a_t'\|^2-\|a_t\|^2\big| + \big|(\alpha_t')^2-\alpha_t^2\big|$. Now $\big|\|a_t'\|^2-\|a_t\|^2\big| = |(\|a_t'\|-\|a_t\|)(\|a_t'\|+\|a_t\|)|$. Using $|\|a_t'\|-\|a_t\|| \le \|a_t'-a_t\|$, we get $\big|\|a_t'\|^2-\|a_t\|^2\big| \le (\|a_t'\|+\|a_t\|)\|a_t'-a_t\|$. Similarly, $|(\alpha_t')^2-\alpha_t^2| = |\alpha_t'-\alpha_t|\,|\alpha_t'+\alpha_t|$. Along the finite trajectory, $\|a_t\|$, $\|a_t'\|$, $|\alpha_t|$, and $|\alpha_t'|$ stay bounded for sufficiently small perturbations. Therefore by (25) and (26) we have

$$|(d_{t,\min}')^2-d_{t,\min}^2| \le C_{m,t}(\|r_{t-1}'-r_{t-1}\|+\eta). \tag{27}$$

Choose the same deterministic root branch in the reference and perturbed problem, for example the minus root: $\lambda_t = \alpha_t-\sqrt{d_t^2-d_{t,\min}^2}$, $\lambda_t' = \alpha'-\sqrt{(d_t')^2-(d_{t,\min}')^2}$. Then $|\lambda_t'-\lambda_t| \le |\alpha_t'-\alpha_t| + \left|\sqrt{d_t^2-(d_{t,\min}')^2}-\sqrt{d_t^2-d_{t,\min}^2}\right|$. By the feasibility-margin assumption, $d_t^2-d_{t,\min}^2 \ge \eta_0^2 > 0$. So if the perturbation is small enough that $C_{m,t}(\|r_{t-1}'-r_{t-1}\|+\eta) \le \eta_0^2/2$, then by (27) we have $d_t^2-(d_{t,\min}')^2 \ge d_t^2-d_{t,\min}^2-|(d_t^2-(d_{t,\min}')^2)-(d_t^2-d_{t,\min}^2)| \ge \eta_0^2-\eta_0^2/2 = \eta_0^2/2$. Thus the perturbed distance equation remains strictly feasible. The map $s \mapsto \sqrt{s}$ is Lipschitz on $[\eta_0^2/2,\infty)$. Therefore the square-root Lipschitz estimate gives $|\sqrt{(d_t)^2-(d_{t,\min}')^2}-\sqrt{d_t^2-d_{t,\min}^2}| \le C_{\sqrt{}}|(d_t)^2-(d_{t,\min}')^2-(d_t^2-d_{t,\min}^2)| = C_{\sqrt{}}\big|(d_{t,\min}')^2-d_{t,\min}^2\big|$. By (26) and (27) we obtain

$$|\lambda_t'-\lambda_t| \le C_{\lambda,t}(\|r_{t-1}'-r_{t-1}\|+\eta). \tag{28}$$

Finally, $r_t'-r_t = (r_{t-1}'-r_{t-1})-(\lambda_t'u_t'-\lambda_t u_t)$, and $\lambda_t'u_t'-\lambda_t u_t = (\lambda_t'-\lambda_t)u_t'+\lambda_t(u_t'-u_t)$. Since $\|u_t'\|=1$, λ_t is bounded along the finite reference trajectory, and by (24) and (28), we get $\|r_t'-r_t\| \le L_t\|r_{t-1}'-r_{t-1}\|+D_t\eta$. The induction hypothesis then gives $\|r_t'-r_t\| \le B_t\eta$ for a suitable constant B_t. This closes the induction.

Taking the maximum of the finitely many constants over all samples and all steps, we obtain $C_{\mathrm{CCDS}} > 0$ such that $\|r_{l,T_l}'-r_{l,T_l}\| \le C_{\mathrm{CCDS}}\eta$ for every l. The transported perturbed path is an admissible CCDS realizer f_{CCDS}' on $U'=U$. Therefore $\varepsilon_{\mathrm{param}}(A,U',f_{\mathrm{CCDS}}';\,\Xi) = \frac{1}{N}\sum_{l=1}^N\|r_{l,T_l}'\| \le \frac{1}{N}\sum_{l=1}^N\|r_{l,T_l}\|+\frac{1}{N}\sum_{l=1}^N\|r_{l,T_l}'-r_{l,T_l}\| \le \varepsilon_{\mathrm{param}}(A,U,f_{\mathrm{CCDS}};\,\Xi)+C_{\mathrm{CCDS}}\eta$. Taking $C_{\mathrm{param}} := C_{\mathrm{CCDS}}$ and recalling that $\eta = \|G'-G\|_2$ gives the desired bound.

How to choose δ_0 explicitly: for every $l=1,\dots,|\Xi_{\mathrm{sub}}|$, define $\delta_{1,l,t}>0$ such that $\|G'-G\|_2 \le \delta_{1,l,t}$ implies $\|q_{l,t}'-q_{l,t}\| \le \nu/2$; and $\delta_{2,l,t}>0$ such that $C_{m,l,t}(\|r_{l,t-1}'-r_{l,t-1}\|+\delta_{2,l,t}) \le \eta_0^2/2$. For each $j=1,2$, set $\delta_{j,l} := \min_{1\le t\le T_l}\delta_{j,l,t}$ and $\delta_j := \min_{1\le l\le|\Xi_{\mathrm{sub}}|}\delta_{j,l}$. Finally, set $\delta_0 := \min(\gamma/2,\delta_1,\delta_2)$.

The same argument applies when the block interpolants are chosen as the zero-initialized one-step update constructive interpolation procedure of the Exact Interpolation Theorem 4. Indeed, everything discussed above remains intact except for the definition of J_k. The same argument we

used above to obtain $V \subseteq U$ applies, so we just have to prove $U \subseteq V$. To see this, we invite the reader to come back to the details of the proof of the Exact Interpolation Theorem 4. Fix one block $D_k = \{(x_{k,1}, y_{k,1}), \ldots, (x_{k,s}, y_{k,s})\}$, write $z_{k,i} := W^{(L_{k,i})} x_{k,i}$ as in the proof of Theorem 4, and let the recursive construction start from $\tilde{J}_{k,0} = 0$. At step i, the interpolant is defined as $\tilde{J}_{k,i} := \tilde{J}_{k,i-1} + (y_{k,i} - \tilde{J}_{k,i-1} z_{k,i}) \otimes f_{k,i}$. We claim by induction that $\mathrm{range}(\tilde{J}_{k,i}) \subseteq \mathrm{span}\{y_{k,1}, \ldots, y_{k,i}\}$. This is clear for $i = 0$. If it holds for $i-1$, then in particular, $\tilde{J}_{k,i-1} z_{k,i} \in \mathrm{span}\{y_{k,1}, \ldots, y_{k,i-1}\}$, hence $y_{k,i} - \tilde{J}_{k,i-1} z_{k,i} \in \mathrm{span}\{y_{k,1}, \ldots, y_{k,i}\}$. Consequently, the rank-one correction $(y_{k,i} - \tilde{J}_{k,i-1} z_{k,i}) \otimes f_{k,i}$ also has range contained in $\mathrm{span}\{y_{k,1}, \ldots, y_{k,i}\}$. Then $\mathrm{range}(\tilde{J}_{k,i}) \subseteq \mathrm{span}\{y_{k,1}, \ldots, y_{k,i}\}$. Thus the final block map J_k defined as $J_k := \tilde{J}_{k,s}$ satisfies $\mathrm{range}(J_k) \subseteq \mathrm{span}\{y_{k,1}, \ldots, y_{k,s}\} := \mathrm{range}(Y_k)$. Hence, for all $i = (k, m, j) \in \mathcal{I}$, $g_i = J_k W^{(m)} x_{k,j} \in range(Y_k)$. Therefore, $U = range(G) \subseteq \sum_{k=1}^{K} \mathrm{range}(Y_k) = V$. The rest of the proof is unchanged: any admissible perturbed block family $J'_k = Y_k R'_k$ with the same range property gives $U' \subseteq U$; the smallness condition $\|G' - G\|_2 < \gamma/2$ forces $\mathrm{rank}(G') = \mathrm{rank}(G)$, hence $U' = U$; the projection term is therefore invariant, and the transported CCDS path satisfies the same Lipschitz parametrization estimate.

■

2.7 CCDS sparsity.

Beyond interpretability and stability, one of the most important payoffs of the CCDS realization mechanism is sparsity. More strongly, CCDS does not merely produce a sparse realizer; it realizes the projected target using the intrinsic minimum number of orbit atoms. Let $\mathcal{H} = \mathbb{R}^C$ and fix the selected scaffold $U \subset \mathcal{H}$. Let $(x_i, y_i) \in D_k$ and set the projected target $p_i := \Pi_U y_i$. Let the active orbit dictionary for this sample be $\mathcal{A}_i = \{a_{i,j}\}_{j=1}^{M_i} = \{J_k W^{(m)} x_i\}_{k,m} \subset U$.

For a subset $S \subset \{1, \ldots, M_i\}$, define $E_{i,S} := \mathrm{span}\{a_{i,j} : j \in S\}$, and the local approximation defect $\rho_{i,S} := \mathrm{dist}(p_i, E_{i,S}) = \|p_i - \Pi_{E_{i,S}} p_i\|$. Fix a projected-parametrization tolerance $0 \le \eta \le \|p_i\|$. Now we define *the intrinsic sparsity per sample number* corresponding to (x_i, y_i) as $s_i(\eta) := \min\{|S| : \mathrm{dist}(p_i, E_{i,S}) \le \eta\}$, this is a per-sample certificate complexity We claim that CCDS produces a certificate with support at most $s_i(\eta)$. Let S_i^* be a minimizer, and set $E_i^* := \mathrm{span}\{a_{i,j} : j \in S_i^*\}$, $z_i^* := \Pi_{E_i^*} p_i$, $\rho_i^* := \|p_i - z_i^*\| \le \eta$.

Theorem 35 (Intrinsic local orbit-certificate sparsity of full CCDS) *If $z_i^* \neq 0$, then there exists a CCDS realizer $\widehat{p}_i = \sum_{j \in S_i^*} \alpha_{i,j}^* a_{i,j}$ such that $\|p_i - \widehat{p}_i\| = \eta$, and $|\operatorname{supp} \alpha_i^*| \le s_i(\eta)$.*

Proof Fix i. Write $p = p_i, \quad S = S_i^*, \quad E = E_i^*, \quad z = z_i^* = \Pi_E p, \quad \rho = \rho_i^*$. By definition, $\rho = \|p - z\| \le \eta$. Since $z \in E$, and E is spanned by exactly the atoms indexed by S, we may write $z = \sum_{j \in S} c_j a_{i,j}$. Define $u := \frac{z}{\|z\|}$. Then $u \in E$, so u is an admissible interior orbit innovation supported on the atoms in S. Start full CCDS from $Y_0 = \{0\}, \quad \widehat{p}_0 = 0, \quad r_0 = p$. The CCDS distance equation at this step is $\rho(r_0 - \lambda u, Y_0) = \eta$. Since $Y_0 = \{0\}$, this becomes $\|p - \lambda u\| = \eta$. Now compute the minimal feasible deviation along u. We have $\langle p, u \rangle = \left\langle z + (p - z), \frac{z}{\|z\|} \right\rangle$. Because $z = \Pi_E p$, the residual $p - z$ is orthogonal to E, and $u \in E$. Hence $\langle p, u \rangle = \left\langle z, \frac{z}{\|z\|} \right\rangle = \|z\|$. Therefore, $d_{\min}^2 = \|p\|^2 - |\langle p, u \rangle|^2 = \|p\|^2 - \|z\|^2$. Since $z = \Pi_E p$, $\|p\|^2 = \|z\|^2 + \|p - z\|^2$. Thus $d_{\min} = \|p - z\| = \rho \le \eta$. So the CCDS distance equation is feasible. The CCDS root formula gives $\lambda = \|z\| - \sqrt{\eta^2 - \rho^2}$. Now define $\widehat{p}_i := \lambda u$. By construction, $\|p_i - \widehat{p}_i\| = \eta$. Since $u = \frac{z}{\|z\|} = \frac{1}{\|z\|} \sum_{j \in S} c_j a_{i,j}$, we get $\widehat{p}_i = \lambda u = \sum_{j \in S} \frac{\lambda c_j}{\|z\|} a_{i,j}$. Define $\alpha_{i,j}^* := \lambda c_j / \|z\|$ for each j. Therefore, $\operatorname{supp} \alpha_i^* \subseteq S$, so $|\operatorname{supp} \alpha_i^*| \le |S| = s_i(\eta)$. This proves the theorem. ■

Definition 36 *We define the ideal coefficient map $\alpha^{(*,\eta)}$ relative to the error η as $\alpha^{(*,\eta)}(x_i) := \alpha_i^*$ for each sample (x_i, y_i); where α_i^* is the coefficient of the ideal CCDS realizer associated to (x_i, y_i) given by Theorem 35.*

3 Prediction as Extension of Constructive Orbit-Based Realizability

The philosophy before formalization. Before entering the technical development of the predictive part of the paper, we explain the guiding idea. Let $\Xi_{train} = \{(x_i, y_i) : 1 \le i \le n\}$ be the training set. In our framework, the training data does not only specify the outputs to be matched; through the constructive realizability stage, it also induces a preferred orbit-based representation of those outputs. Prediction is then built by extending this constructive structure rather than by learning the target map freely.

A natural first attempt would be to use the entire training set as the interpolation set. One could then seek a partition $\Xi_{train} = D_1 \sqcup \cdots \sqcup D_K$ such that the exact interpolation theorem applies on each block D_k, yielding a weighted shift W and interpolants J_k with $y_i = J_k W^{(L_i)} x_i$ for every training pair (x_i, y_i). This would provide an exact constructive teacher on the full training set and lead to a predictive rule of the form

$$\widetilde{f}(x) = \sum_{k=1}^{K} \sum_{m=0}^{d-1} \alpha_{k,m}(x)\, J_k W^{(m)} x.$$

But this full-training regime is not the one we ultimately want: the orbit dictionary becomes highest-dimensional, the model moves closer to memorization, and in noisy settings exact interpolation of the full training set is undesirable.

This leads to the alternative adopted in this work. Rather than building the scaffold from all of Ξ_{train}, we build it from a carefully selected subset. On that subset we again apply exact interpolation, obtaining a shift W and interpolants J_k, whose orbit atoms generate a dictionary $\mathcal{D}$ and scaffold $U = \operatorname{span} \mathcal{D}$. The subset is chosen so that the resulting scaffold captures as much as possible of the relevant output geometry of the full training set.

Once the scaffold has been selected, we do not attempt to interpolate the full target directly. Instead, we project the target onto U and construct on U a deterministic CCDS realizer. This produces an effective coefficient map on the training set. The final predictive step is then to learn, on the full training set, a coefficient map that reproduces and extends this constructive orbit-based rule from the training points to all of $\mathbb{R}^d$.

This is what makes the predictive step different from ordinary supervised fitting. The labels y_i specify what output must be produced, but the CCDS realizer specifies how that output is constructively produced through orbit atoms on the selected scaffold. The predictor is therefore not trained to learn the target map freely; it is trained to reproduce and generalize a structured constructive rule.

Our predictive philosophy is thus organized around three explicit bottlenecks: the scaffold bottleneck, the realizability bottleneck, and the predictive coefficient bottleneck. The first determines how well the selected scaffold captures the target geometry; the second measures how well the projected target can be constructively realized on that scaffold; and the third governs how well the learned predictor reproduces and extends the resulting coefficient rule beyond the training set. In this way, the realizability theory developed in the first part of the paper becomes the structural backbone of prediction rather than a separate component placed alongside it.

Formalization. One of the most striking features of our theory is that the structural decomposition developed for realizability carries over directly to prediction on the dataset. In the predictive setting the target map is simply the dataset correspondence $f(x_i) = y_i$. Theorem 37 below is an in-sample structural theorem that identifies the right bottlenecks for learning on the training data:

(i) the quality of the selected scaffold U, through the projection error $\|f - \Pi_U f\|_{\Xi}$; (ii) the quality of the CCDS realization on that scaffold through the parametrization error; and (iii) the quality of the shallow learned policy through its approximation on the effective CCDS coefficient map on the dataset.

Let $\Xi_{train} = \{(x_i, y_i)\}_{i=1}^n = D_1 \sqcup \cdots \sqcup D_K \subset \mathbb{R}^d \times \mathbb{R}^C$, $\Xi_{test} = \{(\bar{x}_j, \bar{y}_j)\}_{j=1}^m \subset \mathbb{R}^d \times \mathbb{R}^C$. Define the dataset seminorm $\|g\|_{\Xi} := (\sum_{x:(x,y)\in\Xi} \|g(x)\|_{\ell^2})/|\Xi|$ for maps g defined on $\{x : (x, y) \in \Xi\}$.

Theorem 37 (Geometric in-sample prediction bound) *Fix a scaffold $U = \operatorname{span} \mathcal{D} \subset \mathbb{R}^C$ with set of orbit atoms $\mathcal{D} = \{J_k W^{(m)} x_i : (x_i, y_i) \in D_k,\ 1 \le k \le K,\ 0 \le m \le d-1\}$. Define the target map on Ξ_{train} by $f(x_i) := y_i$. And let $f_{\mathrm{CCDS}}(x) := \sum_{k=1}^K \sum_{m=0}^{d-1} a_{k,m}^{\mathrm{eff}}(x)\, J_k W^{(m)}(x)$, with $x \in U$, be the CCDS realizer on the scaffold U. Denote $a^{\mathrm{eff}}(x) := \big(a_{k,m}^{\mathrm{eff}}(x)\big)_{k,m} \in \mathbb{R}^{Kd}$. Assume our predictor has the form*

$$\widetilde{f}(x) := \sum_{k=1}^{K} \sum_{m=0}^{d-1} \alpha_{k,m}(x)\, J_k W^{(m)}(x) \tag{29}$$

and satisfies, for each $i = 1, \ldots, n$, $\|\alpha(x_i) - a^{\mathrm{eff}}(x_i)\|_{\ell^1} \le \delta_i$. Then

$$\|f - \widetilde{f}\|_{\Xi_{train}} \le \underbrace{\|f - \Pi_U f\|_{\Xi_{train}}}_{\varepsilon_{proj}(f,U;\,\Xi_{train})} + \underbrace{\|\Pi_U f - f_{\mathrm{CCDS}}\|_{\Xi_{train}}}_{\varepsilon_{param}(f,U,f_{\mathrm{CCDS}};\,\Xi_{train})} + M\,\bar{\delta},$$

where $\bar{\delta} := \frac{1}{n}\sum_{i=1}^n \delta_i$ and $M := \max_{1\le k\le K,\ 0\le m\le d-1,\ 1\le i\le n} \|J_k W^{(m)}(x_i)\|_{\ell^2} < \infty$ by boundedness of the operators $J_k W^{(m)}$.

Proof Fix $i \in \{1, \ldots, n\}$. Then

$$(\Pi_U f)(x_i) - \widetilde{f}(x_i) = \big((\Pi_U f)(x_i) - f_{\mathrm{CCDS}}(x_i)\big) + \big(f_{\mathrm{CCDS}}(x_i) - \widetilde{f}(x_i)\big).$$

Hence, by the triangle inequality,

$$\|(\Pi_U f)(x_i) - \widetilde{f}(x_i)\|_{\ell^2} \le \|(\Pi_U f)(x_i) - f_{\mathrm{CCDS}}(x_i)\|_{\ell^2} + \|f_{\mathrm{CCDS}}(x_i) - \widetilde{f}(x_i)\|_{\ell^2}.$$

We estimate the second term. By definition,

$$f_{\mathrm{CCDS}}(x_i) - \widetilde{f}(x_i) = \sum_{k=1}^{K} \sum_{m=0}^{d-1} \big(a_{k,m}^{\mathrm{eff}}(x_i) - \alpha_{k,m}(x_i)\big)\, J_k W^{(m)} x_i.$$

Applying the triangle inequality again,

$$\|f_{\mathrm{CCDS}}(x_i) - \widetilde{f}(x_i)\|_{\ell^2} \le \sum_{k=1}^{K} \sum_{m=0}^{d-1} \big|a_{k,m}^{\mathrm{eff}}(x_i) - \alpha_{k,m}(x_i)\big|\, \|J_k W^{(m)} x_i\|_{\ell^2}.$$

Then,

$$\|f_{\mathrm{CCDS}}(x_i) - \widetilde{f}(x_i)\|_{\ell^2} \le M \sum_{k=1}^{K} \sum_{m=0}^{d-1} \big|a_{k,m}^{\mathrm{eff}}(x_i) - \alpha_{k,m}(x_i)\big|.$$

That is, $\|f_{\mathrm{CCDS}}(x_i) - \widetilde{f}(x_i)\|_{\ell^2} \le M\, \|\alpha(x_i) - a^{\mathrm{eff}}(x_i)\|_{\ell^1} \le M\, \delta_i$. Therefore,

$$\|(\Pi_U f)(x_i) - \widetilde{f}(x_i)\|_{\ell^2} \le \|(\Pi_U f)(x_i) - f_{\mathrm{CCDS}}(x_i)\|_{\ell^2} + M\, \delta_i.$$

Averaging over $i = 1, \ldots, n$, we obtain

$$\|\Pi_U f - \widetilde{f}\|_{\Xi_{train}} \le \|\Pi_U f - f_{\mathrm{CCDS}}\|_{\Xi_{train}} + M\, \bar{\delta}.$$

Finally, observe that $f - \widetilde{f} = (f - \Pi_U f) + (\Pi_U f - \widetilde{f})$. Applying the triangle inequality,

$$\|f - \widetilde{f}\|_{\Xi_{train}} \leq \|f - \Pi_U f\|_{\Xi_{train}} + \|\Pi_U f - \widetilde{f}\|_{\Xi_{train}}.$$

Substituting the bound just proved yields

$$\|f - \widetilde{f}\|_{\Xi_{train}} \leq \|f - \Pi_U f\|_{\Xi_{train}} + \|\Pi_U f - f_{\mathrm{CCDS}}\|_{\Xi_{train}} + M\,\bar{\delta}.$$

This completes the proof. ■

Theorem 37 shows that the in-sample predictive instantiation of our orbit-based framework, which we call Weighted Backward Shift Neural Networks (WBSNNs) and defined formally below in Definition 40, is naturally governed by three bottlenecks that define the three phases of the model. The first is the construction of an optimal scaffold, which determines the projection quality and therefore defines Phase 1. The second bottleneck, and hence Phase 2, is the CCDS realizability of the projected targets on that scaffold. Phase 3 is the shallow learned policy that approximates the effective CCDS coefficient map on the dataset, thereby controlling the final parametrization gap. In this sense, prediction in WBSNN inherits the same structural logic as realizability, but with an additional learned coefficient-approximation layer in the final phase.

3.1 Predictive Framework (WBSNNs).

We now formalize WBSNN Phase 3.

Phase 3: Shallow learned policy approximating f_{CCDS}. Mimicking dynamics in infinite dimensions, we define Phase 3 of WBSNN as a generalization mechanism over the orbit-induced class

$$\widetilde{f}(x) = \sum_{k=1}^{K} \sum_{m=0}^{\infty} \alpha_{k,m}(x)\, J_k W^{(m)}(x), \qquad x \in \mathbb{R}^d,$$

trained on the full dataset Ξ_{train}. Equivalently, Phase 3 learns over the family generated by the orbit geometry associated with all training samples, while producing a predictor defined for arbitrary inputs $x \in \mathbb{R}^d$. Note that this is the reason why we considered unwrapping $W^{(L)}$ modulo d; otherwise, standard backward shifts would vanish beyond depth d, preventing generalization along the infinite orbit. By Lemma 3, in finite-dimensional spaces our generalization expression collapses—up to scalar factors to the prediction rule:

$$\widetilde{f}(x) = \sum_{k=1}^{K} \sum_{m=0}^{d-1} \alpha_{k,m}(x)\, J_k W^{(m)} x, \qquad x \in \mathbb{R}^d. \tag{30}$$

In order to be able to apply the in-sample transfer Theorem 37, the Phase 3 predictor must be explicitly of the form prescribed by the prediction rule (30), and the associated coefficient map must approximate the effective CCDS coefficient map on that scaffold. Thus, Phase 3 is not trained to learn coefficients from scratch, but to approximate f_{CCDS} within the same orbit-based representation class.

The next step is to define the predictor $\widetilde{f}$ explicitly so that: (i) it has the form prescribed by the prediction rule (30); (ii) it is endowed with an explicit guiding structure that allows it to track the CCDS coefficient map effectively; and (iii) it is implemented through a hybrid shallow policy, thereby preserving interpretability.

Definition 38 (Predictive Hybrid Learned CCDS mechanism of Phase 3) *Assuming the outputs of Phases 1 and 2 have already been obtained, we now introduce the hybrid Phase 3 mechanism, which is trained on the full dataset Ξ_{train} and yields a predictor defined on all of $\mathbb{R}^d$.*

Orbit dictionary: *For each input $x \in \mathbb{R}^d$, define the orbit atoms*

$$d_{k,m}(x) := J_k W^{(m)} x \in \mathbb{R}^C, \qquad k \in \{1, \dots, K\}, \;\; m \in \{0, \dots, d-1\},$$

and their normalized versions

$$\widetilde{d}_{k,m}(x) := \frac{d_{k,m}(x)}{\|d_{k,m}(x)\| + \varepsilon_{\text{norm}}},$$

where $\varepsilon_{\text{norm}} > 0$ is fixed. We write $\mathcal{D}(x) := \{\widetilde{d}_{k,m}(x) : 1 \le k \le K,\ 0 \le m \le d-1\}$.

Learned policy variables: *Let $T \ge 1$ be the number of Phase 3 steps. For each $t = 1, \dots, T$, let $g_\theta^{(t)} : \mathbb{R}^{d+C} \to \mathbb{R}^{Kd}$, $h_\theta^{(t)} : \mathbb{R}^{d+C} \to \mathbb{R}$, be shallow networks (e.g. one-hidden-layer MLPs). Their role is only to select the* policy variables *of the CCDS step: a soft atom-mixture and a feasible deviation slack. The CCDS coefficient itself remains explicit.*

We also let $b_\theta : \mathbb{R}^d \to \mathbb{R}^C$ be a shallow initialization map, used to define the residual target for the constructive recursion.

Initialization: *Set $\widehat{y}_0(x) := 0, \quad r_0(x) := b_\theta(x), \quad Y_0(x) := \{0\}$.*

Step $t = 1, \dots, T$. Assume $\widehat{y}_{t-1}(x)$, $r_{t-1}(x)$, and $Y_{t-1}(x)$ are already defined. Define the state

$$z_t(x) := \big[x,\ r_{t-1}(x)\big] \in \mathbb{R}^{d+C}.$$

Soft policy over orbit atoms: *Let $\pi_t(x) := \operatorname{softmax}\big(g_\theta^{(t)}(z_t(x))\big) \in \Delta^{Kd-1}$, where $\Delta^{Kd-1} := \{\pi \in \mathbb{R}^{Kd} : \pi_j \ge 0,\ \sum_{j=1}^{Kd} \pi_j = 1\}$ denotes the probability simplex. And define the soft selected direction $q_t(x) := \sum_{k=0}^{K-1} \sum_{m=0}^{d-1} \pi_{t;k,m}(x)\, \widetilde{d}_{k,m}(x)$.*

Innovation direction: *Let $v_t(x) := q_t(x) - \Pi_{Y_{t-1}(x)} q_t(x)$. Assuming $\|v_t(x)\| \neq 0$, define*

$$u_t(x) := \frac{v_t(x)}{\|v_t(x)\| + \varepsilon_{\text{orth}}}, \qquad Y_t(x) := Y_{t-1}(x) \oplus \operatorname{span}\{u_t(x)\},$$

where $\varepsilon_{\text{orth}} > 0$ is fixed.

Feasible deviation target: *Set $a_t(x) := (I - \Pi_{Y_{t-1}(x)})\, r_{t-1}(x), \beta_t(x) := \langle a_t(x), u_t(x) \rangle$, and $d_{t,\min}(x) := \sqrt{\max\big(\|a_t(x)\|^2 - \beta_t(x)^2,\, 0\big)}$. Let $\Delta_t(x) := \operatorname{softplus}\big(h_\theta^{(t)}(z_t(x))\big) \ge 0$, $d_t(x) := d_{t,\min}(x) + \Delta_t(x)$. Hence $d_t(x) \ge d_{t,\min}(x)$ by construction, so feasibility is automatic.*

Explicit CCDS coefficient. *Define the step coefficient by the CCDS root*

$$\lambda_t(x) := \beta_t(x) - \sqrt{\max\big(d_t(x)^2 - d_{t,\min}(x)^2,\, 0\big)}.$$

(One may also choose the + root if desired; throughout we fix one branch consistently.)

Residual update: *Set $\widehat{y}_t(x) := \widehat{y}_{t-1}(x) + \lambda_t(x)\, u_t(x)$, $r_t(x) := b_\theta(x) - \widehat{y}_t(x)$.*

Final predictor: *The Phase 3 output is $\widetilde{f}(x) := \widehat{y}_T(x)$.*

By Proposition 22, $\widetilde{f}$ is of the form prescribed by (30), then we can apply the in-sample transfer Theorem 37 to our predictor $\widetilde{f}$ from Definition 38. So Phase 3 should be understood as a mechanism for constructing the effective coefficients in a structured and interpretable way, namely as an approximation of the CCDS coefficient map associated with the scaffold and atoms dictionary produced by Phases 1 and 2.

Coefficient-supervised Phase–3 training. To enforce both prediction accuracy and coefficient tracking, we train Phase 3 by combining the usual prediction loss with a coefficient-supervision loss against the effective CCDS coefficients. The resulting Phase 3 loss is

$$\mathcal{L}_{\text{Phase 3}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{coeff}} \frac{1}{n} \sum_{i=1}^{n} \|\alpha(x_i) - \alpha^{\text{eff}}(x_i)\|_{\ell^2}.$$

Thus, the in-sample predictive bound of Theorem 37 becomes exactly:

$$\|f-\widetilde{f}\|_{\Xi_{train}} \leq \underbrace{\|f-\Pi_U f\|_{\Xi_{train}}}_{\varepsilon_{proj}(f,U;\,\Xi_{train})} + \underbrace{\|\Pi_U f - f_{\mathrm{CCDS}}\|_{\Xi_{train}}}_{\varepsilon_{param}(f,U,f_{\mathrm{CCDS}};\,\Xi_{train})} + M\,\frac{1}{n}\sum_{i=1}^{n}\|a^{\mathrm{eff}}(x_i)-\alpha(x_i)\|_{\ell^1}, \quad (31)$$

where U is the selected scaffold, $f(x_i)=y_i$ for all $(x_i,y_i)\in\Xi_{train}$; $M=\max_{1\leq k\leq K,\ 0\leq m\leq d-1,1\leq i\leq n}\|J_k W^{(m)}(x_i)\|_{\ell^2}$; $a^{\mathrm{eff}}(x_i)=(a^{\mathrm{eff}}_{k,m}(x_i))_{k,m}\in\mathbb{R}^{Kd}$ and $\alpha(x_i)=(\alpha_{k,m}(x_i))_{k,m}\in\mathbb{R}^{Kd}$.

Hence, in-sample prediction succeeds when the scaffold selected in Phase 1 is good enough that the projected target is close to the true dataset map, the Phase 2 CCDS realizer on that scaffold is good enough to provide the model with a structured surrogate of the projected map, so that the target map is not learned in an unconstrained way, and finally the hybrid Phase 3 policy is good enough to track the CCDS coefficient map.

Remark 39 *The reader should note that:*

(i) *Noise suppression via orbit decay result via expression (20) shows that although Phase 3 in WBSNN only uses orbit powers $m\leq d-1$, these terms implicitly encode decay across the full infinite orbit $\{W^{(m)}\varepsilon\}_{m\geq 0}$ via scalar rescaling thanks to Lemma 3. The vanishing regime $\prod_{i=0}^{d-1} w_i < 1$ induces exponential suppression of long-term perturbations, acting as an intrinsic regularizer in the orbit geometry.*

(ii) *Phase 3 is deliberately defined on infinite orbits in order to pave the way for the extension of WBSNNs to infinite-dimensional spaces. This is fully natural in our framework, since both the exact interpolation Theorem 4 and the CCDS mechanism hold in infinite-dimensional Banach spaces. In this paper, however, we restrict our attention to the finite-dimensional case; a full treatment of the infinite-dimensional WBSNN framework is left for future work.*

We are now ready to formally introduce WBSNNs in the finite-dimensional setting.

Definition 40 (WBSNN on a finite-dimensional dataset) *Given a finite-dimensional training dataset Ξ_{train}, a* Weighted Backward Shift Neural Network (WBSNN) *associated to Ξ_{train} is the predictor $\widetilde{f}$ obtained from the three-phase pipeline of this paper: Phase 1 selects a scaffold U, Phase 2 constructs on that scaffold a CCDS realizer of the projected target, and Phase 3 learns a data-dependent coefficient map $\alpha_{k,m}(x)$ through the hybrid predictive CCDS mechanism from Definition 38, yielding the final prediction rule $\widetilde{f}$ via* (30).

Moreover, the predictor $\widetilde{f}$ from Definition 40 satisfies on Ξ_{train} the in-sample prediction error bound given in (31).

Remark 41 *WBSNN is a general-purpose learning paradigm that operates across modalities and tasks without requiring task-specific architectural redesign. It replaces stacked nonlinearities with structured orbit dynamics. Moreover, Theorem 37 is only an in-sample structural result: it does not by itself establish test generalization. That question is addressed later in Theorem 42. Nevertheless, the framework already suggests a structural inductive bias favorable to robust generalization: optimal scaffold budgets limit capacity, the orbit-span constraint restricts the hypothesis class, vanishing dynamics suppress noise, relaxed interpolation improves stability in noisy regimes, and the shallow Phase 3 head only reweights orbit atoms rather than learning an unrestricted predictor.*

Under the additional assumption that test points lie near the same orbit geometry captured at training, we obtain an out-of-sample prediction bound that preserves the same structural logic as the in-sample bound from Theorem 37. Thus, although the WBSNN architecture already exhibits a

structural inductive bias favorable to generalization, an explicit out-of-sample result is still needed to complete the predictive picture of the framework.

Let $U \subset \mathbb{R}^C$ be the scaffold selected from Ξ_{train}, and let $f : \mathbb{R}^d \to \mathbb{R}^C$ be the target map, so that $f(x_i) = y_i,\ f(\bar{x}_j) = \bar{y}_j$. Let $\widetilde{f}$ be the predictor built on the scaffold U.

Theorem 42 (Geometric out-of-sample prediction bound) *Assume*

1. *(Neighborhood relation between test and train) There exists $\delta \geq 0$ such that for every $\bar{x}_j \in \Xi_{test}$ there exists some $x_{i(j)} \in \Xi_{train}$ with $\|\bar{x}_j - x_{i(j)}\|_2 \leq \delta$.*
2. *(Regularity of the predictor) The map $\widetilde{f}$ is $L_{\widetilde{f}}$-Lipschitz on the same neighborhood: $\|\widetilde{f}(x) - \widetilde{f}(x')\|_2 \leq L_{\widetilde{f}}\|x - x'\|_\ell^2,\ \forall x,\ x' \in \Xi_{train} \cup \Xi_{test}$.*

Then

$$\|f - \widetilde{f}\|_{\Xi_{test}} \leq \underbrace{\|f - \Pi_U f\|_{\Xi_{test}}}_{\varepsilon_{proj}(f,U;\,\Xi_{test})} + \max_{x:(x,y)\in\Xi_{train}} \|\Pi_U f x - \widetilde{f}x\|_{\ell^2} + (1 + L_{\widetilde{f}})\delta.$$

Proof Fix $j \in \{1, \ldots, m\}$, and choose $x_{i(j)} \in \Xi_{train}$ such that $\|\bar{x}_j - x_{i(j)}\|_{\ell^2} \leq \delta$. Then

$$f(\bar{x}_j) - \widetilde{f}(\bar{x}_j) = \big(f(\bar{x}_j) - Pf(\bar{x}_j)\big) + \big(\Pi_U f(\bar{x}_j) - \Pi_U f(x_{i(j)})\big) + \big(\Pi_U f(x_{i(j)}) - \widetilde{f}(x_{i(j)})\big) + \big(\widetilde{f}(x_{i(j)}) - \widetilde{f}(\bar{x}_j)\big).$$

Applying the triangle inequality,

$$\begin{aligned}\|f(\bar{x}_j) - \widetilde{f}(\bar{x}_j)\|_2 \leq\ & \|f(\bar{x}_j) - \Pi_U f(\bar{x}_j)\|_2 + \|\Pi_U f(\bar{x}_j) - \Pi_U f(x_{i(j)})\|_2 \\ & + \|\Pi_U f(x_{i(j)}) - \widetilde{f}(x_{i(j)})\|_2 + \|\widetilde{f}(x_{i(j)}) - \widetilde{f}(\bar{x}_j)\|_2.\end{aligned}$$

We have, $\|\Pi_U f(\bar{x}_j) - \Pi_U f(x_{i(j)})\|_2 \leq \|\Pi_U\|\delta = \delta$. By Assumption 2, $\|\widetilde{f}(x_{i(j)}) - \widetilde{f}(\bar{x}_j)\|_{\ell^2} \leq L_{\widetilde{f}}\delta$. Therefore,

$$\|f(\bar{x}_j) - \widetilde{f}(\bar{x}_j)\|_{\ell^2} \leq \|f(\bar{x}_j) - \Pi_U f(\bar{x}_j)\|_{\ell^2} + \|\Pi_U f(x_{i(j)}) - \widetilde{f}(x_{i(j)})\|_{\ell^2} + (1 + L_{\widetilde{f}})\delta.$$

Averaging over $j = 1, \ldots, m$ we obtain

$$\|f - \widetilde{f}\|_{\Xi_{test}} \leq \|f - \Pi_U f\|_{\Xi_{test}} + \max_{x:(x,y)\in\Xi_{train}} \|\Pi_U f x - \widetilde{f}x\|_{\ell^2} + (1 + L_{\widetilde{f}})\delta.$$

This proves the theorem. ∎

Remark 43 *The out-of-sample version preserves the same logic as the in-sample version: (i) $\|f - \Pi_U f\|_{\Xi_{test}}$ is the scaffold bottleneck on test; (ii) $\max_{x:(x,y)\in\Xi_{train}} \|\Pi_U f(x) - \widetilde{f}(x)\|_{\ell^2}$ is the predictor bottleneck inherited from training; and (iii) $(1 + L_{\widetilde{f}})\delta$ is the new transfer term, which says that test points must lie near the training scaffold geometry. If $\delta = 0$, Theorem 42 collapses back to the same two bottlenecks: projection error on test and parametrization error of the predictor on training.*

Remark 44 (On the Lipschitzness of Phase 3 predictor) *The Lipschitz assumption on $\widetilde{f}$ in Theorem 42 is essentially already built into Definition 38. Indeed, on any bounded region $\Omega \subset X$, the orbit atoms $x \mapsto J_k W^{(m)} x$ are bounded linear maps, hence Lipschitz. The learned objects in Phase–3 are the primitive neural maps appearing in Definition 38; all other quantities $z_t,\ \pi_t,\ q_t,\ v_t,\ u_t,\ \beta_t,\ d_{t,\min},\ \Delta_t,\ d_t,\ \lambda_t,\ \widehat{y}_t,\ r_t$ are induced recursively from these learned maps, the orbit atoms, and fixed algebraic operations. Moreover, the softened normalization $u_t(x) = \frac{v_t(x)}{\|v_t(x)\| + \varepsilon_{\text{orth}}}$ removes the only singularity that would otherwise arise from orthogonalization at $v_t(x) =$*

0. *Thus, in the present formulation, the only remaining possible obstruction to Lipschitz continuity comes from the CCDS root term* $\lambda_t(x) = \beta_t(x) - \sqrt{\max\big(d_t(x)^2 - d_{t,\min}(x)^2, 0\big)}$. *Accordingly, it suffices to impose the uniform feasibility margin*

$$d_t(x)^2 - d_{t,\min}(x)^2 \geq \eta_0^2 > 0 \qquad \text{for all } x \in \Omega \text{ and all } t.$$

This holds, for instance, if the learned slack $\Delta_t(x) = d_t(x) - d_{t,\min}(x)$ *is uniformly bounded away from zero on* Ω *and* $d_{t,\min}$ *is uniformly bounded there. Under this condition, the square-root term is Lipschitz on* Ω*, hence so is* λ_t*. By finite composition of Lipschitz maps, the residual recursion and therefore the final hybrid predictor* $\widetilde{f}$ *are Lipschitz on* Ω*. Thus the additional hypothesis needed for the out-of-sample theorem is not a structural change to Phase–3, but only a uniform positive feasibility margin.*

Remark 45 (Interpolation modalities and practical considerations.) *The stability result of Theorem 34 is a realization-level statement, but it has direct consequences for prediction. The WB-SNN predictor learns the coefficient law induced by the scaffold-CCDS construction; hence stability of the scaffold and CCDS bottlenecks transfers to the predictive stage through the very definition of the coefficient map. For this reason, the experiments below use the relaxed regularized pseudoinverse interpolant throughout. We isolate the exact-versus-relaxed comparison only in the dedicated noise Ablation 2, where the noisy synthetic gauntlet at* 0%, 20%, *and* 40% *noise shows that test accuracy remains essentially unchanged across the two interpolation modalities. In summary, Theorem 4/Theorem 7 gives the canonical exact interpolation theoretical mechanism while the cheaper and more stable regularized interpolation is its numerical counterpart.*

3.2 Predictive sparsity

In our experiments, for each training sample (x_i, y_i) we compute a CCDS realizer whose coefficients we denote $\alpha^{\text{eff}}(x_i)$. This produces an effective coefficient map α^{eff} on the training set (the map used throughout Section 3). Phase 3 then learns a coefficient map α whose purpose is to track α^{eff}.

We continue the sparsity discussion begun in Subsection 2.7. Fix a tolerance $\eta > 0$. Theorem 35 supplies an explicit, certifiable ideal coefficient map $\alpha^{(*,\eta)}$ that realizes the projected target up to tolerance η, with the support of each $\alpha^{(*,\eta)}(x_i)$ bounded by the intrinsic per-sample sparsity number $s_i(\eta)$. For prediction, however, the algebraically exact support of this ideal certificate is not the most relevant object. A learned coefficient map α necessarily carries training, regularization, and numerical error, so exact zeros are unstable under arbitrarily small perturbations that barely change the prediction. We therefore work with a resolution-dependent notion of support obtained by thresholding the coefficients at a positive scale $\tau > 0$.

Denote by $\alpha_{i,j}^{(*,\eta)}$ the j-th entry of $\alpha^{(*,\eta)}(x_i)$.

Theorem 46 (Thresholded effective support of full CCDS) *In the notation of Subsection 2.7, for every threshold* $\tau > 0$ *one has*

$$\big|\{j : |\alpha_{i,j}^{(*,\eta)}| > \tau\}\big| \leq \min\Big(s_i(\eta), \frac{(\|z_i^*\| - \sqrt{\eta^2 - (\rho_i^*)^2})^2}{\gamma_i^2 \tau^2}\Big) \tag{32}$$

where $D_{i,S_i^*} := [a_{i,j}]_{j \in S_i^*}$ *and* $\gamma = \sigma_{\min}^+(D_{i,S_i^*})$.

Proof Fix a sample (x_i, y_i) and keep in mind all the notation from Subsection 2.7. Remember that S_i^* is the minimizer, $E_i^* := \operatorname{span}\{a_{i,j} : j \in S_i^*\}$, $z_i^* := \Pi_{E_i^*} p_i$, $\rho_i^* := \|p_i - z_i^*\| \leq \eta$. Now write $D_{i,S_i^*} := [a_{i,j}]_{j \in S_i^*}$. Then $z_i^* = D_{i,S_i^*} c$ with c chosen as the minimum-ℓ^2 norm solution. Setting $\gamma := \sigma_{\min}^+(D_{i,S_i^*})$, by the singular-value characterization of the Moore–Penrose inverse on the range of D_{i,S_i^*}, $\|D_{i,S_i^*}^\dagger\| = 1/\gamma$. Therefore $\|c\|_2 = \|D_{i,S_i^*}^\dagger z_i^*\| \leq \|D_{i,S_i^*}^\dagger\| \|z_i^*\| = \|z_i^*\|/\gamma$. Remember from

Theorem 35, our ideal CCDS coefficient $\alpha^{(*,\eta)}(x_i) := \lambda c/\|z_i^*\|$, so $\|\alpha^{(*,\eta)}(x_i)\|_2 = |\lambda|\|c\|_2/\|z_i^*\|_2 \leq |\lambda|/\gamma$. With the chosen CCDS branch $|\lambda| = \lambda = \|z_i^*\| - \sqrt{\eta^2 - (\rho_i^*)^2}$, we have $\|\alpha^{(*,\eta)}(x_i)\|_2 \leq (\|z^*\| - \sqrt{\eta^2 - (\rho^*)^2})/\gamma$.

Now define the thresholded active set $A_\tau(\alpha^{(*,\eta)}(x_i)) := \{j : |\alpha_{i,j}^{(*,\eta)}| > \tau\}$ and let $m := |A_\tau(\alpha^{(*,\eta)}(x_i))|$. Then $m\tau^2 \leq \sum_{j\in A_\tau(\alpha^{(*,\eta)}(x_i))} |\alpha_{i,j}^{(*,\eta)}|^2 \leq \sum_j |\alpha_{i,j}^{(*,\eta)}|^2 = \|\alpha^{(*,\eta)}(x_i)\|_2^2$, and therefore $m \leq \|\alpha^{(*,\eta)}(x_i)\|_2^2/\tau^2 \leq \left(\|z_i^*\| - \sqrt{\eta^2 - (\rho_i^*)^2}\right)^2/(\gamma^2\tau^2)$. Combining this with the exact-support bound from Theorem 35 proves the per-sample estimate (32). ■

Theorem 46 supplies a pruning guarantee: coefficients below τ may be removed and the next result (Proposition 47) makes explicit the payoff: the parametrization error increases by a controlled amount. The ideal CCDS realizer given by Theorem 35 is of the form $\hat{p}_i = D_{i,S_i^*}\alpha^{(*,\eta)}(x_i)$; where $D_{i,S_i^*} := [a_{i,j}]_{j\in S_i^*}$. The associated parametrization error is, $\varepsilon_{\text{param},i} = \|p_i - D_{i,S_i^*}\alpha^{(*,\eta)}(x_i)\|$. Define the thresholded vector coefficient

$$\left(\alpha^{(*,\eta,\tau)}(x_i)\right)_j = \begin{cases} \alpha_{i,j}^{(*,\eta)}, & \text{if } |\alpha_{i,j}^{(*,\eta)}| > \tau, \\ 0, & \textit{otherwise.} \end{cases}$$

Denote by $\alpha_{i,j}^{(*,\eta,\tau)}$ the j-th entry of $\alpha^{(*,\eta,\tau)}(x_i)$. The thresholded CCDS realizer has the form $\hat{p}_i^\tau := D_{i,S_i^*}\alpha^{(*,\eta,\tau)}(x_i)$. The associated parametrization error is, $\varepsilon_{\text{param},i}^\tau = \|p_i - D_{i,S_i^*}\alpha^{(*,\eta,\tau)}(x_i)\|$. Then we have the following per sample diagnostic.

Proposition 47 $\varepsilon_{\text{param},i}^\tau \leq \varepsilon_{\text{param},i} + \|D_{i,S_i^*}\|_{\text{op}}\tau\sqrt{S_i^*}$.

Proof The proof is quite immediate. Write $p_i - D_{i,S_i^*}\alpha^{(*,\eta,\tau)}(x_i) = p_i - D_{i,S_i^*}\alpha^{(*,\eta)}(x_i) + D_{i,S_i^*}(\alpha^{(*,\eta)}(x_i) - \alpha^{(*,\eta,\tau)}(x_i))$. Therefore, $\varepsilon_{\text{param},i}^\tau \leq \varepsilon_{\text{param},i} + \|D_{i,S_i^*}(\alpha^{(*,\eta)}(x_i) - \alpha^{(*,\eta,\tau)}(x_i))\|$. So pruning entries below τ may increase the parametrization error by at most $\|D_{i,S_i^*}(\alpha^{(*,\eta)}(x_i) - \alpha^{(*,\eta,\tau)}(x_i))\|$. As the support of $\alpha^{(*,\eta)}(x_i)$ has size S_i^*, and every entry of the coefficient vector $\alpha^{(*,\eta)}(x_i) - \alpha^{(*,\eta,\tau)}(x_i)$ has magnitude at most τ, then $\|\alpha^{(*,\eta)}(x_i) - \alpha^{(*,\eta,\tau)}(x_i)\|_2 \leq \tau\sqrt{S_i^*}$. Hence $\varepsilon_{\text{param},i}^\tau \leq \varepsilon_{\text{param},i} + \|D_{i,S_i^*}\|_{\text{op}}\tau\sqrt{S_i^*}$. ■

Remark 48 *With the explicit increase bound of Proposition 47, if one can already make the base parametrization error $\varepsilon_{\text{param},i}$ sufficiently small, then thresholding the CCDS coefficients at scale τ remains admissible. Thus Theorem 46 together with Proposition 47 convert ideal CCDS certificate sparsity into stable, learnable, and measurable coefficient-law sparsity at finite resolution. This motivates one of the most important diagnostics in our experiments, the effective support size defined in (33) below.*

3.3 Interpretability of the full WBSNN pipeline

A central claim of the framework is that WBSNN is interpretable end-to-end. This interpretability is phasewise and structural. In Phase 1, the model selects an explicit data-induced orbit scaffold, so the geometric support on which learning takes place is visible rather than latent. In Phase 2, the projected target is constructively realized on that scaffold through blockwise interpolation and CCDS-based synthesis, so the realizability stage is governed by explicit certificates rather than by opaque hidden representations. In Phase 3, prediction is not learned as unrestricted feature mixing, but as a data-dependent coefficient law over the orbit atoms generated by the scaffold and the interpolation operators. Thus each phase carries its own interpretable object: scaffold geometry, constructive realization, and coefficient law.

Table 1: Interpretability diagnostics for Phases 1 and 2 of main experiments for WBSNN, subsection 4.2. Lower projection residual and lower interpolation/parametrization errors indicate a scaffold that captures the target geometry well and supports accurate constructive realization. Values are mean ± standard deviation over seeds.

Dataset	$\varepsilon_{\text{proj}}$	$\varepsilon_{\text{param}}$
MNIST	$5.32\text{e}{-15} \pm 2.2\text{e}{-15}$	$1.46\text{e}{-7} \pm 6.21\text{e}{-9}$
ISOLET	$3.11\text{e}{-9} \pm 6.33\text{e}{-10}$	$3.25\text{e}{-6} \pm 2.47\text{e}{-8}$
Gas Sensor Drift	$4.21\text{e}{-11} \pm 6.03\text{e}{-11}$	$8.92\text{e}{-8} \pm 3.89\text{e}{-8}$

Table 2: Interpretability diagnostics for Phase 3 of main experiments for WBSNN, subsection 4.2. Lower coefficient-tracking error and smaller effective support indicate closer adherence to the constructive CCDS rule. Higher top-3 mass and higher neighbor support overlap indicate stronger coefficient concentration and greater local support stability. Values are mean ± standard deviation over seeds.

Dataset	E_{coeff}	Support size	Top-3 mass	Neighbor overlap
MNIST	$1.47\text{e}{-1} \pm 6.22\text{e}{-3}$	18.53 ± 0.23	$(93.13 \pm 0.5)\%$	0.663 ± 0.016
ISOLET	$3.26\text{e}{-1} \pm 3.35\text{e}{-3}$	24.92 ± 0.16	$(35.2 \pm 0.4)\%$	0.705 ± 0.015
Gas Sensor Drift	$7.05\text{e}{-2} \pm 1.47\text{e}{-2}$	9.73 ± 0.64	$(97.48 \pm 0.4)\%$	0.85 ± 0.028

To make this end-to-end interpretability measurable, we report diagnostics aligned with the three phases of WBSNN. For Phase 1/2, we use the projection and parametrization errors. For Phase 3, we report the coefficient-tracking error against the effective CCDS coefficients, the effective support size of the learned coefficient vector, the top-3 coefficient mass, and the support-overlap stability across nearby test points. Together, these diagnostics make it possible to inspect the full WBSNN pipeline rather than only its final outputs.

For Phase 3, let $\alpha(x) \in \mathbb{R}^{Kd}$ denote the learned coefficient vector and $\alpha^{\text{eff}}(x) \in \mathbb{R}^{Kd}$ the effective CCDS coefficient vector used as the constructive target in the hybrid loss. We define the *coefficient-tracking error* by

$$E_{\text{coeff}} := \frac{1}{n} \sum_{i=1}^{n} \big\| \alpha(x_i) - \alpha^{\text{eff}}(x_i) \big\|_2,$$

which measures how closely the learned predictive phase tracks the underlying CCDS rule.

To quantify sparsity, motivated by our discussion in Subsections 2.7 and 3.2, we define the *effective support size* relative to a fixed threshold $\tau > 0$ by

$$S_{\text{eff}} := \frac{1}{n} \sum_{i=1}^{n} \big| \{ j : \ |\alpha_{i,j}(x_i)| > \tau \} \big|. \tag{33}$$

Thus S_{eff} measures how many orbit atoms are meaningfully active on average in the final prediction.

To quantify coefficient concentration, let $\text{Top}_3(\alpha(x_i))$ denote the set of indices of the three largest coefficients in magnitude. We define the *top-3 coefficient mass* by

$$M_3 := \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j \in \text{Top}_3(\alpha(x_i))} |\alpha_j(x_i)|}{\sum_{j=1}^{Kd} |\alpha_j(x_i)|},$$

so that larger values indicate that most of the predictive mass is carried by only a few orbit atoms.

Finally, to measure local structural stability, fix an integer $s \geq 1$ and let

$$\mathcal{S}_s(x_i) := \text{Top}_s(\alpha(x_i))$$

be the top-s support of the coefficient vector. If x_i^{nn} denotes a nearest neighbor of x_i in the test set (or in the relevant feature space), we define the *neighbor support-overlap stability* by

$$O_{\text{nn}} := \frac{1}{n} \sum_{i=1}^{n} \frac{\big| \mathcal{S}_s(x_i) \cap \mathcal{S}_s(x_i^{\text{nn}}) \big|}{\big| \mathcal{S}_s(x_i) \cup \mathcal{S}_s(x_i^{\text{nn}}) \big|}.$$

High values of O_{nn} indicate that nearby inputs tend to activate similar scaffold-orbit supports, which is consistent with a locally stable and geometrically meaningful coefficient field.

Table 1 summarizes the geometric and realizability diagnostics of Phases 1 and 2, while Table 2 summarizes the predictive coefficient diagnostics of Phase 3. Across all representative datasets, the same qualitative pattern is observed: the selected scaffold captures a large portion of the target geometry, the Phase 2 realizer fits that projected target with low residual, and the final hybrid predictive phase remains sparse, concentrated, and close to the effective CCDS rule. This supports the intended interpretation of WBSNN as a structured orbit-based pipeline rather than as a black-box predictor.

3.4 Situating WBSNNs in Machine Learning.

Operator-theoretic ideas have entered machine learning through several distinct routes, including Koopman-inspired approaches that represent and forecast nonlinear dynamical systems through learned linear operator representations and spectral decompositions (Mezić, 2005; Lusch et al., 2018; Azencot et al., 2020; Brunton et al., 2022), operator-learning architectures such as DeepONet that are designed to learn nonlinear operators from data (Lu et al., 2021), methods such as neural integral equations that learn integral operators from data (Zappala et al., 2024), and architectural modifications aimed at efficiency or stability, such as Linformer, which approximates self-attention by a low-rank mechanism to achieve linear complexity, and NAIS-Net, which derives stable deep architectures from non-autonomous differential equations (Wang et al., 2020; Ciccone et al., 2018). WBSNNs are conceptually different from all of these. They are not built to approximate nonlinear operators in the usual neural-operator sense, nor to impose low-rank or spectral constraints on an otherwise black-box architecture. Instead, they build prediction from orbit iterates of a single orbit-generating linear operator and organize learning through three linked ingredients: scaffold selection, constructive realizability on the selected scaffold, and extension of the resulting coefficient law. Their mathematical motivation comes from linear dynamics, where iterates of a single operator generate both rich geometric and dynamical behavior, with the weighted backward shift serving as the guiding prototype (Bayart and Matheron, 2009).

4 Experiments

We instantiate the constructive theory in two experimental regimes. First, we study realizability, where the goal is to test whether a target operator can be constructively synthesized on a selected orbit scaffold, and to isolate the roles of projection quality and deviation-controlled realization. Second, we study predictability, where the same scaffold-realization mechanism is used inside WBSNN, with Phase 3 implemented by the hybrid CCDS coefficient-tracking policy as in Definition 38, rather than by an unconstrained MLP head. Throughout, the experiments are designed to test the theory's structural claims, not merely to maximize benchmark scores.

Unless otherwise stated, all features are standardized using training-set statistics only and compressed by PCA before Phase 1. This imposes a controlled low-dimensional information bottleneck in which scaffold geometry, constructive realizability, and predictive coefficient tracking can be examined consistently. In the predictive experiments, WBSNN and all matched baselines operate on the same PCA-compressed features; raw-input or domain-specific models, when reported, are identified separately as reference comparisons. Only a small *discovery fraction* of the training set is used to build the orbit scaffold. The predictive head uses the hybrid CCDS Phase 3 mechanism, so that prediction is interpreted as an approximation of the effective CCDS coefficient map rather than as unrestricted function fitting. All experiments run in a lightweight CPU-only Jupyter environment—no GPUs—so the numbers reflect true data/geometry leverage rather than horsepower. The experiments are therefore not designed around a large-scale engineering scaling narrative, but around illustrating the practical consequences of the theory in controlled finite-dimensional regimes through

scaffold quality, constructive realizability, and predictive coefficient tracking. Across all experiments, both the CCDS realizer (Phase 2) and the learned coefficient law (Phase 3) exhibit strong sparsity: the effective support size remains small while the top-3 coefficients carry most of the mass (see diagnostic tables).

Phase 1/2 protocol. Phase 1 uses the cheaper and more stable regularized pseudoinverse interpolant: this choice is fully justified by Theorem 34 and Remark 45. Phase 2 uses CCDS to realize the projected target on the selected scaffold. We report two implementations. The fast greedy MP/CCDS solver selects atoms by the matching-pursuit correlation rule, but updates coefficients through the MP/CCDS distance equation; hence it is MP only in atom selection, not in its full update rule. The full CCDS solver instead uses the nested-subspace innovation construction and is reported separately to show how prescribed deviation control affects convergence under the same selected scaffold.

4.1 Experiments on realizability

The realizability experiments isolate the two bottlenecks from Section 2: the projection bottleneck, governed by scaffold quality, and the parametrization bottleneck, governed by constructive synthesis on the selected scaffold. The first two experiments already expose the necessity of CCDS under weak or degenerate geometry. We then add one diagnostic experiment whose sole purpose is to make scaffold quality visible as an empirical object.

Experiment 1: Pathological weak-alignment stress test and deviation-control ablation. We evaluate Phase 2 realizability on an extreme synthetic weak-alignment dataset designed to destroy global dictionary coherence. Samples are generated from random label-dependent directions in independent orthogonal frames and are then re-rotated by independent random rotations; large label noise further places residual directions in near-generic position relative to any fixed orbit dictionary. This creates a small-cosine regime in which greedy one-ray alignment is unreliable and CCDS updates are expected to dominate.

We use an ill-conditioned operator family with no PCA, working in full dimension $d = 10$ (raw scaled features), 500 training samples, seven output classes, and a strict Phase 1 budget of 7% of the training set. The selected scaffold uses only 35 anchor samples, organized into 23 blocks, and reduces the projection bottleneck to numerical precision, $\varepsilon_{\text{proj}} \approx 1.35 \times 10^{-10}$. Thus the experiment no longer tests failure caused by scaffold incompleteness; instead, it isolates the ability of MP/CCDS to realize projected targets under weak alignment. The empirical alignment diagnostic confirms the intended pathology: 0% of probe samples maintain $\cos\theta \geq 0.8$ over the first 20 greedy steps, and the MP certification rate is 0%. The fast MP/CCDS solver nevertheless realizes the projected operator outputs to numerical precision, with $\varepsilon_{\text{param}} \approx 6.36 \times 10^{-8}$, using a mean of 14.11 steps. Hence the experiment shows that weak one-step alignment does not prevent constructive realization once the scaffold has captured the relevant output geometry. Full CCDS, run as a deviation-control diagnostic, reveals the expected dependence on the prescribed contraction schedule. For $\gamma_{\text{dev}} = 0$, where d_t is driven to the minimal feasible boundary, full CCDS obtains $\varepsilon_{\text{param}} \approx 2.2 \times 10^{-7}$ and a 99.2% realizability rate (i.e. 99.2% of training samples have $\varepsilon_{\text{param}} \leq 10^{-6}$). For $\gamma_{\text{dev}} = 0.5$, the imposed interior deviation profile is more conservative, increasing the mean residual to approximately 5.33×10^{-4} and reducing exact realizability to 54.4%. For $\gamma_{\text{dev}} = 0.9$, the residual contraction is more conservative within the fixed step budget: all samples reach the step limit and the full-CCDS realizability rate drops to 41.2% reaching a mean residual of 1.18×10^{-3}.

This ablation separates three effects. First, scaffold selection can collapse the geometric projection bottleneck even in an adversarial weak-alignment dataset. Second, fast MP/CCDS can still synthesize the projected target to numerical precision, despite the absence of certified pure-MP trajectories. Third, full CCDS should not be interpreted as a residual-minimizing solver in this diagnostic: as γ_{dev} increases, it deliberately realizes an interior deviation profile rather than the

smallest attainable residual. The sparse and class-structured atom usage shows that the realizations remain interpretable even in this pathological regime. We report the full ablation in Table 7 and provide representative solver traces in the complementary material.

Experiment 2: Extreme-Ray Dominance vs. Interior Feasibility in Near-Parallel Subspaces Under Extremely Low Budget Random Scaffolds. We evaluate Phase 2 realizability of MP/CCDS on a challenging synthetic dataset formed by a union of six nearly parallel 3-dimensional subspaces in $\mathbb{R}^{35}$ ($\sigma_{\text{parallel}} = 0.02$, $\sigma_{\text{noise}} = 0.05$), reduced to $d = 15$ by PCA (98.2% variance retained) and 2000 training samples. Labels are generated by linear operators C followed by a fixed readout B, yielding classification tasks with identical input geometry but diverse operator spectra (identity, orthogonal, ill-conditioned, low-rank, spiked, sparse, Toeplitz, and adversarially scaffold-misaligned operator). A deliberately weak scaffold coverage (3% random subset, no optimization) is used to decouple Phase 2 realizability from scaffold selection. Notably, the low-rank operator is the only case with a large projection error. This highlights a limitation of rank-based diagnostics often used in machine learning: low intrinsic rank does not necessarily imply geometric compatibility with the representation space on which realization is attempted. A low-dimensional operator can still be poorly captured when its active range directions are misaligned with the dataset-induced scaffold. In this sense, scaffold alignment provides a finer diagnostic than rank alone.

Despite poor projection coverage, fast MP/CCDS consistently achieves extremely small parametrization error ($\varepsilon_{\text{param}} \sim 10^{-8}$–$10^{-6}$) via very sparse representations. However, strict MP contraction is almost never certified: residuals typically lie on flat or symmetric faces of the feasible cone, where several atoms have nearly equal correlations and no single extreme ray has a margin. Full CCDS, by contrast, enforces deviation-controlled multi-atom feasibility rather than one-step contraction, achieves moderate to high realizability (4–88% of training samples have Full CCDS parametrization error below 1e-6) across spectra, constructing stable interior two-atom certificates, trading minimal residual for interior feasibility. More details on full diagnostics and interpretability metrics in Tables 3 and 8 in Appendix A.

Furthermore, the fast MP/CCDS and full CCDS traces for Sample 0 (Identity and Orthogonal operator), illustrate the sharp contrast between greedy dynamics and deviation-controlled certificates (full details in complementary materials). In fast mode, the Identity operator case converges in 14 steps with occasional CCDS fallbacks, while the Orthogonal operator case requires 50 steps and exhibits strong alternation between a few atoms (e.g., repeated switching between 327 and 194), a clear instance of boundary crawling along a flat cone face. This directly visualizes the margin-free pathology: no single atom dominates, so pure MP progresses slowly and fragily. In contrast, full CCDS always produces a clean two-step interior-feasible certificate. Together, these traces provide concrete evidence that greedy contraction is insufficient in near-parallel geometries, whereas deviation control yields short, stable, and geometrically robust realizations.

Overall, Experiment 2 exposes a core limitation of classical greedy contraction: MP requires strict single-ray dominance and fails almost universally due to margin collapse under symmetry, degeneracy, or near-parallelism. Realizability is instead governed by cone interior width and operator-induced anisotropy. This benchmark cleanly separates approximation quality from contraction geometry and shows that deviation-controlled solvers such as CCDS are essential — not optional — for achieving reliable, stable realizability under weak, unoptimized scaffolds. For a full report of interpretability diagnostics (per-sample supports, atom usage histograms, deviation paths, and stopping statistics), we refer the reader to the complementary materials.

Experiment 3: scaffold-quality diagnostic. This experiment is designed to empirically validate the scaffold bottleneck itself. We reuse the same synthetic generator as in Experiment 1 so that the operator family, ambient dimension, and noise model remain fixed; only the scaffold construction mechanism is varied. For each of 10 runs, we construct three candidate scaffolds with the same scaffold-size budget and the same realization-step budget: a selected orbit scaffold produced by the scaffold-selection algorithm from Section 2.2, a random orbit scaffold produced from the same

Table 3: Experiment 2: Near-parallel and symmetric degeneracy benchmark with random 3% scaffold. Full CCDS remains robust when greedy single-ray contraction becomes unstable or vacuous.

Operator	% samples MP updates	% samples CCDS updates	Fast $\varepsilon_{\text{param}}$ (mean)	Full CCDS $\varepsilon_{\text{param}}$ (mean)	Steps (fast mean)	$\varepsilon_{\text{proj}}$ (mean)
I	0.20	99.80	2.06e−06	3.43e−06	20.76	2.56e−06
ORTHO	0.10	99.90	5.09e−07	2.90e−06	20.97	2.60e−06
ILL_SVD	0.00	100	1.31e−07	1.02e−06	21.24	6.87e−07
LOWRANK	0.00	100	1.84e−07	7.26e−07	18.09	7.18e−01
SPIKED	0.00	100	2.20e−07	1.60e−06	20.41	1.44e−06
SPARSE	0.00	100	1.88e−07	1.38e−06	20.63	1.10e−06
TOEPLITZ	0.00	100	5.64e−08	6.40e−07	20.79	2.58e−07
ADV_SCAFF	0.05	99.95	6.73e−07	2.56e−06	20.93	2.40e−06

orbit-generated family but with a random admissible subset/partition choice, and a random non-orbit based Gaussian scaffold built from normalized Gaussian atoms with the same total number of atoms as the orbit dictionary. For the Gaussian baseline we impose a proper output-rank bottleneck $r \leq C - 1$, since a full-rank Gaussian scaffold spans the entire output space almost surely and makes $\varepsilon_{\text{proj}}$ degenerate. For each scaffold U, we compute projection error, parametrization error and the *explained label-space energy* defined as $E_{\text{expl}}(U) = \frac{\|\Pi_U Y\|_F^2}{\|Y\|_F^2}$. Table 9 in Appendix A reports the results. The selected orbit scaffold dominates both comparison scaffolds—the random orbit scaffold and the random non-orbit based Gaussian scaffold—not only in projection quality but also in downstream realizability. This experiment is especially important because it shows directly that orbit-based scaffolds do matter for learning: the gain is not coming merely from dictionary size or ambient dimension, but from the orbit structure itself and from selecting that structure in a data-adaptive way. The systematic gap between the selected orbit scaffold, the random orbit scaffold, and the random non-orbit based Gaussian baseline therefore gives concrete empirical evidence that learnability in our framework is driven by scaffold geometry.

4.2 Experiments on predictability

We now evaluate the predictive instantiation of the framework. The predictive section is intentionally compact but sufficient to illustrate the theory: three representative predictive runs and three focused ablations spanning five datasets. The datasets serve distinct roles: chronological distribution shift (Gas Sensor Drift), high-dimensional multiclass speech recognition under feature compression (ISOLET), compressed vision and discovery-budget saturation (MNIST), recovery of hidden manifold geometry (Swiss roll + RFF), and stability of exact versus relaxed interpolation under controlled label noise (noisy synthetic gauntlet).

Baselines. We compare against strong baselines organized by family and matched to the regime under consideration, but we deliberately keep the main-paper suite compact. Because all predictive experiments are conducted in PCA-compressed finite-dimensional settings, we restrict attention to controls that are meaningful under the same compressed-feature protocol: one strong linear baseline, one strong kernel baseline, one strong tree ensemble, one shallow neural baseline, and, when the dataset warrants it, one domain-matched baseline tailored to the underlying shift or drift structure. This preserves the family-organized comparison principle without turning the main text into an exhaustive benchmark inventory.

Main predictive results. The main predictive results are reported in Tables 1, 2 and 4.

Gas Sensor Drift. We use the batch-chronological protocol, training on earlier batches and testing strictly on later batches under sensor drift. Features are standardized on the training split and compressed to $d = 10$ principal components (95% retained variance). We use 3000 training

samples. The scaffold is discovered from only a 1% subset of the training data. This experiment therefore evaluates predictive performance under simultaneous distribution shift and an extremely limited discovery budget. Despite relying on a highly sparse orbit realization over the selected scaffold (9.7 atoms in average used per sample out of 220), WBSNN outperforms the considered baselines showing that accurate prediction can be achieved with a compact and interpretable orbit representation rather than dense parameterization.

ISOLET. We evaluate on the ISOLET spoken letter recognition dataset using PCA-compressed features $d = 25$ (71% retained variance). With the full training split of 6238 samples, the scaffold is discovered from only a 1% subset. The fast MP/CCDS realizer proves insufficient, producing a relatively high parametrization error ($\approx 5.2\times10^{-3}$) with no samples achieving pure MP certification. We therefore fall back to the full CCDS solver, which attains near machine-precision realization ($\approx 3.26\times10^{-6}$) across all samples. The resulting orbit representation is highly sparse: each prediction activates on average only 24.9 atoms out of the nominal 550-atom scaffold, with the top three coefficients carrying approximately 35.2% of the mass. Despite the extremely limited discovery budget, the WBSNN predictor trained on these coefficients reaches 91.94% test accuracy, remaining competitive with strong baselines while relying on a compact and explicitly constructed sparse orbit representation.

MNIST. We use PCA-compressed pixel features with $d = 25$ (46% retained variance), 5000 training samples, and a 1% discovery fraction for scaffold selection. This experiment evaluates whether a small orbit scaffold can support accurate prediction from a highly compressed representation. Despite relying on a sparse orbit realization learned from this limited discovery budget (18.5 atoms in average used per sample out of 525), WBSNN outperforms all classical baselines trained on the same compressed features, including SVM and Gradient Boosting, while remaining fully interpretable through its scaffold and sparse CCDS representation. Convolutional architectures such as CNNs achieve higher accuracy by operating directly on the full training set of raw images, rather than learning from a 1% scaffold over compressed features. Even under this substantially more favorable setting, the performance gap remains relatively small.

For the full set of interpretability diagnostics introduced in subsection 3.3, we refer the reader to the complementary materials.

Table 4: Main predictive results under the updated protocol: scaffold selection in Phase 1, CCDS-based realization in Phase 2, and hybrid CCDS coefficient tracking in Phase 3. For each dataset we report strong baselines organized by family and matched to the compressed-feature regime: one representative linear, kernel, tree, and shallow neural baseline, plus a domain-matched baseline when relevant. Results are mean ± standard deviation over seeds. Best result in bold.

Dataset	WBSNN 1%	LogReg	RBF-SVM	GBT	MLP	Domain-matched
MNIST	0.934 ± 0.003	0.877 ± 0.002	0.930 ± 0.001	0.919 ± 0.002	0.908 ± 0.01	CNN (raw img) $\mathbf{0.971 \pm 0.01}$
ISOLET	0.919 ± 0.006	0.915 ± 0.0	$\mathbf{0.929 \pm 0.0}$	0.891 ± 0.004	0.903 ± 0.007	Nyström RBF 0.912 ± 0.00
Gas Sensor	$\mathbf{0.71 \pm 0.03}$	0.61 ± 0.05	0.56 ± 0.02	0.64 ± 0.03	0.67 ± 0.05	LogReg+CORAL 0.61 ± 0.05

Predictive ablations

Ablation 1: recovering hidden manifold geometry through scaffold selection. Does scaffold selection recover the intrinsic manifold geometry better than random orbit or unstructured Gaussian scaffolds? This ablation isolates the contribution of scaffold geometry itself and is among the most important experiments in the paper. We use the noisy 5-class Swiss-roll with random Fourier features dataset precisely because the ground-truth manifold geometry is known while the observed representation is nonlinear, noisy and high-dimensional (30-D RFF features projected to PCA = 10) (58.4% retained variance) and 3000 training samples. All reported values are mean ± standard deviation across 10 independent runs performed with paired random seeds. This setup lets us test directly whether predictive performance requires recovering the right orbit geometry or whether it can be explained merely by dictionary size or Phase-3 capacity. Under identical Phase-1/2 budgets ($p = 1\%$) and with the same Phase-3 coefficient head, we compare three scaffolds of

comparable atom count: (i) the data-adaptive orbit scaffold returned by our selection algorithm, (ii) a random orbit scaffold drawn from the same orbit-generated family, and (iii) an unstructured Gaussian scaffold. Table 5 shows a clear, systematic ordering: the selected orbit scaffold achieves near machine-precision alignment with the target $\varepsilon_{\text{proj}} = 1.37 \times 10^{-10} \pm 3.9 \times 10^{-10}$, while a random orbit scaffold of the same size yields $\varepsilon_{\text{proj}} = 1.96 \times 10^{-2} \pm 3.4 \times 10^{-2}$ and an unstructured Gaussian scaffold yields $\varepsilon_{\text{proj}} = 4.36 \times 10^{-1} \pm 8.2 \times 10^{-2}$. Consistent with the scaffold quality metrics, test accuracy degrades sharply with increasing variance: Selected orbit: 0.8689 ± 0.0021, Random orbit: 0.8293 ± 0.0725, Gaussian: 0.7247 ± 0.1068. This ordering demonstrates that both ingredients are essential. First, orbit structure matters: random orbit scaffolds consistently outperform unstructured Gaussian scaffolds, showing that the orbit-generated family carries genuine predictive content beyond mere dictionary expansion. Second, scaffold selection matters: the data-adaptive selected orbit scaffold consistently outperforms a random orbit scaffold of identical size, proving that learnability is not driven by orbit structure in the abstract but by recovering the specific orbit geometry compatible with the dataset. In short, the ablation supplies direct empirical support for the central claim of the paper: prediction improves when the scaffold captures more of the underlying target action, and this improvement arises from structured, data-adaptive orbit geometry rather than from brute-force growth of dictionary capacity.

Ablation 2: exact vs. relaxed interpolation under controlled label noise. Theorem 34 turns exact interpolation from a fragile idealization into a practical principle: the exact interpolant is the canonical reference object, but a regularized relaxed interpolant can be used without destroying the scaffold geometry. Indeed, the theorem predicts that replacing the exact block interpolants by their regularized pseudoinverse versions should preserve the projection bottleneck and perturb the CCDS parametrization bottleneck only continuously, as long as the induced atom system remains in the local stability regime. We test this prediction on a noisy synthetic gauntlet with 15,000 samples and 50 raw features, reduced by PCA to $d = 10$ (94.1% retained variance). The dataset combines informative coordinates, correlated nuisance coordinates, and a mildly spurious feature. We add symmetric noise only to the training labels at levels $0\%, 20\%$, and 40%, while keeping test labels clean. Exact and relaxed interpolation are run with paired seeds and the same 1% Phase 1/2 discovery budget. In the relaxed arm we use a fixed ridge parameter $\lambda = 10^{-3}$, so the near-identical test accuracies are obtained despite using a genuinely regularized, non-exact local interpolant.

The outcome is striking: across all noise levels, the test accuracies of exact and relaxed interpolation remain essentially identical. Thus, machine-precision anchor interpolation is not the fragile ingredient driving prediction. What survives—and what Theorem 34 identifies—is the stability of the orbit scaffold and of the CCDS coefficient law induced on it. This justifies using the regularized relaxed interpolant as the main protocol: it keeps the predictive behavior of the exact construction while being cheaper, better conditioned, and more robust under noisy local blocks.

Ablation 3: discovery-budget curve. This ablation varies the Phase 1/2 discovery fraction while keeping all other settings fixed. Its purpose is to test whether the orbit scaffold needed for prediction can be identified from a small discovery set. Results are reported in Table 10 in Appendix A. The main pattern is that predictive performance saturates very early. On MNIST (5000 training samples, PCA = 20, 42% retained variance), moving from 1% to larger discovery budgets changes test accuracy only marginally (a few tenths of a percentage point). On Gas Sensor Drift (3000 training samples, PCA = 10, 95% retained variance), the 1% discovery budget is in fact the best. Across budgets, the projection and parametrization errors remain numerically negligible, so the scaffold and CCDS realization bottlenecks have already been solved and the limiting factor has shifted to Phase 3. Thus, the ablation supports a data-efficient scaffold-discovery principle. Once a small orbit scaffold captures the relevant label-space geometry, adding more discovery anchors mainly enlarges the dictionary and introduces additional coefficient degrees of freedom; under drift, these extra degrees of freedom can even reduce transfer stability. The relevant conclusion is therefore not "more discovery data gives better accuracy," but rather that WBSNN can recover a predictive

orbit scaffold from a very small discovery budget, after which generalization is governed by the transferability of the learned CCDS coefficient law. In this sense, small discovered scaffolds appear to act as structural regularizers, especially in drifted regimes.

Table 5: Ablation 1: Scaffold ablation on Swiss Roll + RFF. The selected orbit scaffold achieves near machine-precision alignment with the target and outperforms both random orbit and Gaussian scaffolds of the same size in test accuracy, while also exhibiting much lower variance.

Scaffold type	Test acc	$\varepsilon_{\text{proj}}$	$\varepsilon_{\text{param}}$	Effect Support
WBSNN; Selected Orbit (1%)	$\mathbf{0.8689 \pm 0.0021}$	$1.37\text{e}{-10} \pm 3.9\text{e}{-10}$	$5.32\text{e}{-8} \pm 4.2\text{e}{-8}$	5.63 ± 1.44
WBSNN; Random Orbit (1%)	0.8293 ± 0.0725	$1.96\text{e}{-2} \pm 3.4\text{e}{-2}$	$1.55\text{e}{-8} \pm 7.4\text{e}{-9}$	3.35 ± 0.85
Gaussian Scaffold (1%)	0.7247 ± 0.1068	$4.36\text{e}{-1} \pm 8.2\text{e}{-2}$	$5.34\text{e}{-3} \pm 8.5\text{e}{-3}$	1.41 ± 1.46

Table 6: Ablation 2: Exact vs relaxed interpolation under controlled training-label noise. Test labels are kept clean. Values are mean ± st dev over five seeds. Near-identical test accuracies show that the relaxed regularized interpolant preserves the predictive behavior of the exact scaffold construction, in agreement with Theorem 34.

	Test acc (0% noise)	Test acc (20% noise)	Test acc (40% noise)
WBSNN, Phase 1/2: 1 % relaxed	0.9711 ± 0.0015	0.9691 ± 0.0012	0.8925 ± 0.0077
WBSNN, Phase 1/2: 1 % exact	0.9710 ± 0.0010	$\mathbf{0.9694 \pm 0.0008}$	0.8925 ± 0.0077
Best baseline	$\mathbf{0.975 \pm 0.0}$ DRO-Log	0.958 ± 0.0 LogReg	$\mathbf{0.9327 \pm 0.0}$ LogReg
Worst baseline	0.925 ± 0.0 ExtraTrees	0.9163 ± 0.0 ExtraTrees	0.7689 ± 0.001 RdmFor

Runtime and scaffold reuse. Finally, we emphasize that the reported WBSNN runtimes are CPU-only and correspond to an unoptimized constructive prototype. They should therefore not be interpreted in the same way as mature baseline implementations, whose core routines are highly optimized. In our runs, the dominant cost is concentrated in Phase 1, the scaffold-discovery stage, whereas the CCDS realization (Phase 2) and predictive coefficient-fitting (Phase 3) stages are comparatively lightweight. This is consistent with the structure of the theory: Phase 1 is the stage where the data-induced orbit geometry is explicitly recovered, while later phases operate on the selected scaffold. Thus the expensive object is not an opaque fitted parameterization, but an interpretable orbit scaffold. Its cost is naturally amortizable: once such a scaffold has been recovered, subsequent dataset enlargements or task updates need not be viewed as requiring full geometric rediscovery. The appropriate theoretical mechanism is the residual/scaffold adaptation principle formalized by Puig (2026) in OrBA, where adaptation is organized through orbit-generated scaffold families, observable projection mismatch, and CCDS realization on the selected scaffold. Moreover, the robustness of scaffold superiority under nearby scaffold perturbations indicates that reuse is not tied to a single exceptional scaffold, but to a neighborhood of admissible orbit geometries. In this sense, Phase 1 should be understood as an offline geometry-discovery cost; optimizing it is an important implementation direction, but is separate from the structural learnability claims established here.

Conclusion.

This work establishes a unified orbit-based language for scaffold discovery, constructive realization, and prediction through a single local-to-global orbit scaffold-realization principle, laying the foundation for a broader structural theory in which seemingly distinct learning problems may be understood as instances of the same underlying mechanism.

Acknowledgments

I am especially grateful to Nicholas J. Bordelon Jr. for believing in WBSNNs even more than I did, and to Ozgur Martin for encouraging me to transition into the rich world of machine learning. This research received no external funding, and the author declares no competing interests.

Appendix A. Remaining Proofs and Tables

Proof of (20). By Lemma 3, for every $m \geq 1$, there exists a scalar $\lambda_m > 0$ such that $W^{(m)}z = \lambda_m \cdot W^{(m \bmod d)}z$. Let $m = qd + r$, with $0 \leq r \leq d-1$ and $q \in \mathbb{N}$. Then, $\lambda_m = \left(\prod_{i=0}^{d-1} w_i\right)^q = \rho^q$. Hence,

$$\|W^{(m)}z\| = \lambda_m \cdot \|W^{(r)}z\| \leq \rho^q \cdot \max_{0 \leq r < d} \|W^{(r)}z\| = \rho^{\lfloor m/d \rfloor} \cdot C,$$

where $C := \max_{0 \leq r \leq d-1} \|W^{(r)}z\|$ depends only on z and d. Since $\rho < 1$, it follows that $\|W^{(m)}z\| \to 0$ exponentially as $m \to \infty$. ■

Proof of (21). Fix a tolerance $\varepsilon > 0$, let $\rho = \prod_{i=0}^{d-1} w_i$ and the effective horizon

$$M_{\text{eff}}(x; \varepsilon) := \min\{M \geq 0 : \|J_k W^{(m)}x\| \leq \varepsilon,\ \forall m \geq M\}.$$

Let $C := \max_{0 \leq r < d-1} \|J_k W^{(r)}x\|$, by (20), a sufficient condition for $\|J_k W^{(m)}x\| \leq \varepsilon$ is $C\,\rho^{\lfloor m/d \rfloor} \leq \varepsilon$, or equivalently $\rho^{\lfloor m/d \rfloor} \leq \frac{\varepsilon}{C}$. Since $0 < \rho < 1$, the function $x \mapsto \log_\rho x$ is decreasing; thus, $\lfloor m/d \rfloor \geq \frac{\log(\varepsilon/C)}{\log \rho}$. Therefore,

$$m \ \geq\ d \left\lceil \frac{\log(\varepsilon/C)}{\log \rho} \right\rceil. \tag{34}$$

We have proved that if m satisfies (34), then $\|J_k W^{(m)}x\| \leq \varepsilon$. Thus, we arrive at the explicit upper bound $M_{\text{eff}}(x; \varepsilon) \ <\ d \left\lceil \frac{\log(\varepsilon/C)}{\log \rho} \right\rceil$. This completes the proof of (21). ■

Table 7: Experiment 1: Phase 2 MP/CCDS behavior under deviation–control ablation on the weak–alignment synthetic dataset (7% scaffold budget, $d = 10$, $M = 500$, operator family ILL–SVD).

γ_{dev}	$\varepsilon_{\text{proj}}$	$\varepsilon_{\text{param}}^{\text{fast}}$	atm x samp/Kd mean (fast solv)	Top-3 mass fast solv	$\varepsilon_{\text{param}}^{\text{full CCDS}}$	samples with $\varepsilon_{\text{param}}^{\text{full CCDS}} \leq 10^{-6}$	atm x samp/Kd mean (full CCDS)	Top-3 mass full CCDS
0.0	1.35×10^{-10}	6.36×10^{-8}	9/230	96.8%	2.20×10^{-7}	99.2 %	3/230	100%
0.5	1.35×10^{-10}	6.75×10^{-8}	10/230	95.1 %	5.34×10^{-4}	54.4 %	3/230	100%
0.9	1.35×10^{-10}	3.87×10^{-3}	9/230	91.9%	1.18×10^{-3}	41.2 %	3/230	100%

Table 8: Experiment 2: Interpretability metrics for the fast/full CCDS solver.

Operator	atm x samp/Kd mean (fast solv)	Top-3 mass fast solv	Top Atoms among classes 0-5 (fast solver)	atm x samp/Kd mean (full CCDS)	Top-3 mass full CCDS
I	13/450	88.6%	323 (29–32%), 329 (20–28%), 318(20–24%)	2/450	100%
ORTHO	13/450	88.8%	323 (25–32%), 329 (23–27%), 443 (20–26%)	2/450	100%
ILL_SVD	13/450	88.9%	323 (20–27%), 329 (22–27%), 317 (19–22%)	2/450	100%
LOWRANK	13/450	96.3%	323 (19–31%), 443 (18–24%), 329 (19–25%)	2/450	100%
SPIKED	14/450	86.5%	323 (27–33%), 329 (22–26%), 443 (18–24%)	2/450	100%
SPARSE	13/450	88.6%	323 (27–30%), 329 (23–28%), 317 (20–23%)	2/450	100%
TOEPLITZ	13/450	88.2%	329 (25–30%), 323 (24–27%), 325 (18–24%)	2/450	100%
ADV_SCAFF	13/450	88.3%	323 (26–32%), 329 (20–28%), 325 (21–26%)	2/450	100%

Table 9: Experiment 3: Scaffold-quality diagnostic on the synthetic realizability benchmark. Better scaffold quality, measured through lower projection residual and higher explained label-space energy, correlates with lower parametrization error and higher realizability success. Values are mean ± standard deviation over runs.

Scaffold type	$\varepsilon_{\text{proj}}$	E_{expl}	$\varepsilon_{\text{param}}$
Selected orbit	$1.5\text{e}{-10} \pm 9.9\text{e}{-11}$	1.00 ± 0.0	$7.39\text{e}{-8} \pm 4.8\text{e}{-9}$
Random orbit	$2.15\text{e}{-1} \pm 1.1\text{e}{-1}$	0.85 ± 0.10	$1.4\text{e}{-8} \pm 2.9\text{e}{-9}$
Gaussian	$2.1\text{e}{-1} \pm 1.8\text{e}{-1}$	0.80 ± 0.28	$1.0\text{e}{-3} \pm 1.0\text{e}{-3}$

Table 10: Ablation 3: Discovery-budget ablation. Gains are strongly front-loaded, indicating early recovery of the relevant orbit geometry and saturation of scaffold quality.

Dataset	Metric	1%	3%	5%	7%	10%
Gas Sensor Drift	Test Accuracy	0.6761	0.6661	0.6621	0.6413	0.6531
3000 train samples	$\varepsilon_{\text{proj}}$	1.4e−10	5e−15	2.4e−15	7.3e−11	8e−11
$d = 10$	$\varepsilon_{\text{param}}$	1.1e−7	1.4e−7	9.12e−8	8.8e−8	4.7e−8
	Nominal atoms	220	280	340	400	490
	Effect Support	8.88	10.29	10.42	9.95	9.04
	Top-3 mass (%)	97.1	97.9	98	97	98
MNIST	Test Accuracy	0.9298	0.9321	0.9326	0.9308	0.9305
5000 train samples	$\varepsilon_{\text{proj}}$	4.3e−15	7.1e−10	5.8e−15	6.1e−15	9.6e−15
$d = 20$	$\varepsilon_{\text{param}}$	1.4e−7	1.4e−7	9.1e−8	1.5e−7	1.7e−7
	Nominal atoms	440	540	640	740	880
	Effect Support	19.02	17.95	18.79	18.57	18.58
	Top-3 mass (%)	90.8	90.4	91.8	92.4	92.5

References

O. Azencot, N. B. Erichson, V. Lin, and M. W. Mahoney. Forecasting sequential data using consistent Koopman autoencoders. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 475–485. PMLR, 2020.

A. B. Bakushinsky and M. Y. Kokurin. *Iterative Methods for Approximate Solution of Inverse Problems*, volume 577 of *Mathematics and Its Applications*. Springer, 2004.

F. Bayart and E. Matheron. *Dynamics of Linear Operators*, volume 179 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2009.

S. N. Bernstein. On the inverse problem of the theory of the best approximation of continuous functions. In *Collected Works*, volume II, pages 292–294. Academy of Sciences of the USSR, 1954. (Original work from 1938.).

S. L. Brunton, M. Budisic, E. Kaiser, and J. N. Kutz. Modern Koopman Theory for Dynamical Systems. *SIAM Review*, 64(2):229–340, 2022.

M. Ciccone, M. Gallieri, J. Masci, C. Osendorfer, and F. Gomez. NAIS-Net: Stable Deep Networks from Non-Autonomous Differential Equations. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 2018.

J. B. Conway. *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 1990.

R. A. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Adv Comput Math*, 5: 173–187, 1996.

R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge/United Kingdom, 2nd edition, 2012.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat Mach Intell*, 3:218–229, 2021.

B. Lusch, J. N. Kutz, and S. L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(4950), 2018.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, Dec 1993.

I. Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41:309–325, 2005.

V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7(2):414–417, 1966.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 40–44. IEEE, 1993.

Y. Puig. An interpretable theory of structured operator adaptation beyond low rank, 2026. Preprint.

H. S. Shapiro. Some negative theorems of approximation theory. *Michigan Mathematical Journal*, 11(3):211–217, 1964.

V. Temlyakov. *Greedy Approximation*, volume 20 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2011.

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, Dec 2007.

S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-Attention with Linear Complexity. arXiv preprint arXiv:2006.04768, 2020.

E. Zappala, A. O. Fonseca, J. Caro, and A. H. Moberly. Learning integral operators via neural integral equations. *Nat Mach Intell*, 6:1046–1062, 2024.