# SCALING LAWS FOR QUANTIZED NEURAL COMPUTATION: GEOMETRY, SYNTHESIS, AND EXECUTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Classical empirical neural scaling laws commonly treat scaling exponents as fitted constants, limiting mechanistic prediction before costly resource sweeps are executed. We develop a first-principles theory of quantized sequential neural computation that derives joint scaling surfaces directly from independently measurable structural profiles. First, we prove that marginal depth and library-size scaling rates cannot identify their joint law because depth dynamically reuses a finite operation library as a decoder. We resolve this bottleneck by introducing the task-quotiented shared-library reachable width. For a canonical spectral residual class, we solve the finite-resource minimax problem exactly, obtaining $\Delta_a(D, M) = \frac{1}{2D}\sum_{i=1}^{2D} \max\{a_i, a_{M+1}\}$, where $D$ is depth, $M$ is library cardinality, and $(a_i)$ is the task-visible operation spectrum. This formula shows that depth resolves a clipped Cesàro profile rather than a primitive power law. Monotone regular variation yields a sharp summability transition in which nonsummable spectra retain spectrum-dependent rates, summable spectra enter a universal $\Theta(D^{-1})$ synthesis regime, and critical spectra are governed by an integrated slowly varying profile. Routed nonlinear residual adapters realize these phases, while task norm and loss geometry transform the observed exponent. Furthermore, execution semantics supplies a second structural law for arithmetic transport. Under fixed-horizon residual refinement, exact Green-operator identities show that full-state write-back has $\Theta(D)$ arithmetic amplification via backward-propagator mass, whereas first-order error feedback remains $\Theta(1)$ via adjoint semivariation. On a frozen pretrained ResNet-18 representation, a task-visible spectrum estimated from 5,000 calibration examples predicts all 49 registered depth–library errors on untouched test data with 1.02% median and 2.49% worst relative error without fitting a held-out exponent. Holding the ideal route fixed while altering only arithmetic semantics shifts the measured depth slope from 0.90 under write-back to 0.00 under error feedback, isolating a bit-width law of $2^{-1.01b}$. Consequently, neural scaling exponents emerge not as primitive constants, but as identifiable, predictable summaries of operation geometry, finite synthesis, task geometry, and execution-error transport.

## 1 INTRODUCTION

A scaling law is most informative when its exponent can be predicted before the expensive resource sweep is run. Classical neural scaling summarizes loss through

$$L(R) = L_\infty + cR^{-\alpha} + o(R^{-\alpha}), \tag{1}$$

and has become a central instrument for allocating parameters, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Yet $\alpha$ is commonly treated as a fitted constant. A growing theoretical literature instead relates scaling to data dimension, kernel and feature spectra, skill frequencies, language statistics, and task complexity (Sharma & Kaplan, 2022; Bahri et al., 2024; Bordelon et al., 2024; Michaud et al., 2023; Cagnetta et al., 2026; Barkeshli et al., 2026). Precision-aware studies likewise fit how bit width changes loss, effective capacity, or optimal training allocation (Kumar et al., 2025; Xu et al., 2024; Frantar et al., 2025; Chen et al., 2025; Sun et al., 2025; Dremov et al., 2026).

**Here is the question:** *Can the scaling surface of a quantized neural computation be predicted from independently measured structure, without fitting the final exponent?*



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

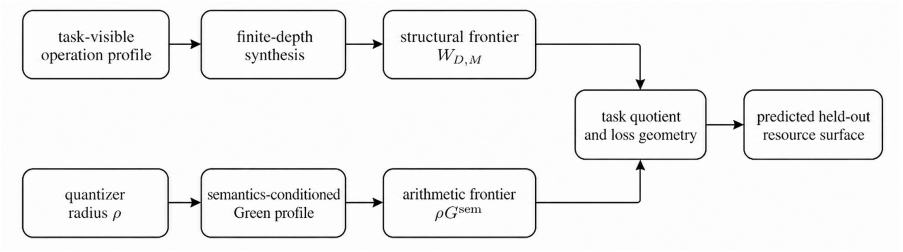



Figure 1: **A structural account of scaling.** The operation profile determines what a finite library leaves unresolved; depth determines how that profile can be synthesized. Arithmetic error is transported according to execution semantics. Task and loss geometry convert the structural and arithmetic frontiers into the observed resource surface.

The question is different from static tensor quantization. We consider a residual computation of depth $D$ that repeatedly selects from one operation library of cardinality $M$ and executes at arithmetic radius $\rho$. The library is an alphabet and depth forms words over it. Reuse may convexify the alphabet, shape or cancel error, generate finer effective addresses, or amplify perturbations. Consequently, the law cannot in general be reconstructed from separate depth-only and precision-only curves. The fixed-library resource theory of Soltanalian (2026) identifies the relaxed infinite-depth limit and separates write-back from error-feedback execution. The present work determines the joint law when library resolution, synthesis depth, and arithmetic precision vary together, and shows how to measure the structural objects that select its exponent.

**Contributions.** The paper follows the causal chain in Figure 1.

**C1. The joint resource object.** We prove that marginal exponents do not determine a joint law and introduce the task-quotiented shared-library reachable width. A sharp operational sandwich states precisely when an additive or max-type law follows.

**C2. An exact finite-resource frontier.** For a weighted operation simplex, we solve the full $(D, M)$ minimax problem. The answer is a clipped Cesàro profile with an exact library saturation threshold. Regular variation converts the profile into spectrum-controlled, critical, or universal $D^{-1}$ phases.

**C3. A neural realization and task-dependent exponent.** Routed nonlinear residual adapters realize the geometry. In $L^q$ task geometry the critical spectrum moves to $p = 1/q$, and local loss curvature further transforms the measured exponent.

**C4. Execution changes the law.** We derive exact Green operators: write-back depends on backward-propagator mass, whereas first-order error feedback depends on adjoint semivariation. Combining them with the spectral frontier gives exact robust joint laws and different useful-depth scales for identical ideal dynamics.

**C5. Structural prediction on a pretrained representation.** The structural map is stable and supports finite-horizon prediction. In a frozen ResNet-18 experiment, the calibration spectrum reproduces on untouched data and predicts all 49 registered depth–library errors; a semantics-only intervention then exposes the predicted growing-versus-bounded depth laws.

Quantized networks do not follow one power law. A fitted exponent is an asymptotic summary of identifiable profiles: operation geometry, synthesis, task geometry, and error transport. The supplement fixes the complete resource accounting and notation (Section A), supplies every proof and counterexample (Sections B to H), documents the registered experiments (Section I), and records symbolic and finite-dimensional verification (Sections J and K).

## 2 THE JOINT RESOURCE OBJECT

Let $Y$ be a behavior space, $\mathcal{O} : Y \to Z$ the task observation, and

$$d_{\mathsf{T}}(y, y') = \rho(\mathcal{O}y, \mathcal{O}y') \tag{2}$$

the induced task pseudometric. For nonempty sets, write $\mathrm{exc}(A, B) = \sup_{a\in A} \inf_{b\in B} d_{\mathsf{T}}(a, b)$. Let $\mathcal{A}_\infty(\mathcal{U})$ be the ideal relaxed class generated by the full operation family $\mathcal{U}$, and $\mathcal{A}_D(Q)$ the depth-$D$ pure class generated by one library $Q$ reused at every layer.

**Definition 1** (Shared-library reachable width)**.**

$$W^{\rightarrow}_{D,M} = \inf_{Q\subset\mathcal{U}:\,|Q|\le M} \mathrm{exc}(\mathcal{A}_\infty(\mathcal{U}), \mathcal{A}_D(Q)) \,. \tag{3}$$

*The library is chosen before the target is revealed; the route may then adapt to that target.*

Task quotienting is exact: the width equals the directed excess between the observed classes $\mathcal{OA}_\infty$ and $\mathcal{OA}_D$ (Section B.5).

**Theorem 2** (Marginal laws are nonidentifying)**.** *There are staged residual/register computations with the same marginal asymptotic rates $a_D \asymp D^{-1}$ and $c_M \asymp M^{-1}$ but*

$$W^{(1)}_{D,M} \asymp \max\{D^{-1}, M^{-1}\}, \tag{4}$$

$$W^{(2)}_{D,M} \asymp \max\{D^{-1}, M^{-1}, (DM)^{-1/3}\}. \tag{5}$$

Thus along $D = M = n$ the joint rates are $n^{-1}$ and $n^{-2/3}$. The construction in Section B changes only how the same reusable phase and counters are observed. Word-generated classes can be more extreme, with resolution exponential in $D$. Hence:

$$\boxed{\text{No function of marginal exponents alone identifies the joint law.}} \tag{6}$$

For a fixed $Q$, let $h(Q) = \mathrm{exc}(\mathcal{A}_\infty(\mathcal{U}), \mathcal{A}_\infty(Q))$ be its relaxed library defect, $\pi^{\rightarrow}_D(Q) = \mathrm{exc}(\mathcal{A}_\infty(Q), \mathcal{A}_D(Q))$ its synthesis cost, and $\bar\pi_D(Q)$ the reverse excess.

**Theorem 3** (Operational sandwich)**.**

$$\max\{\pi^{\rightarrow}_D(Q), [h(Q) - \bar\pi_D(Q)]_+\} \le \mathrm{exc}(\mathcal{A}_\infty(\mathcal{U}), \mathcal{A}_D(Q)) \le h(Q) + \pi^{\rightarrow}_D(Q). \tag{7}$$

*If $\mathcal{A}_D(Q) \subseteq \mathcal{A}_\infty(Q)$, then*

$$\frac{1}{2} \inf_{|Q|\le M} \{h(Q) + \pi^{\rightarrow}_D(Q)\} \le W^{\rightarrow}_{D,M} \le \inf_{|Q|\le M} \{h(Q) + \pi^{\rightarrow}_D(Q)\}. \tag{8}$$

Thus an additive scaling ansatz is a consequence of a *common* library sequence that controls both terms uniformly, not a consequence of separate marginal fits. We next solve this coupling exactly in a spectral class.

## 3 An exact law from the operation spectrum

Let $a_1 \ge a_2 \ge \cdots \downarrow 0$, $v_i = a_i e_i \in \ell_1$, and

$$K_a = \overline{\mathrm{conv}}\{0, v_1, v_2, \ldots\} = \{(a_i\lambda_i) : \lambda_i \ge 0,\ \sum_i \lambda_i \le 1\}. \tag{9}$$

A library $I$ exposes at most $M$ modes. A depth-$D$ word realizes the denominator-$D$ grid

$$R_{D,I} = \left\{ D^{-1} \sum_{i\in I} n_i v_i : n_i \in \mathbb{N}_0,\ \sum_i n_i \le D \right\}, \tag{10}$$

and

$$\Delta_a(D, M) = \inf_{|I|\le M} \sup_{x\in K_a} \mathrm{dist}_{\ell_1}(x, R_{D,I}). \tag{11}$$

**Theorem 4** (Exact clipped-Cesàro frontier)**.** *For every $D, M \ge 1$,*

$$\Delta_a(D, M) = a_{M+1} + \frac{1}{2D} \sum_{i=1}^{\min\{M,2D\}} (a_i - a_{M+1}) \tag{12}$$

$$= \frac{1}{2D} \sum_{i=1}^{2D} \max\{a_i, a_{M+1}\}. \tag{13}$$

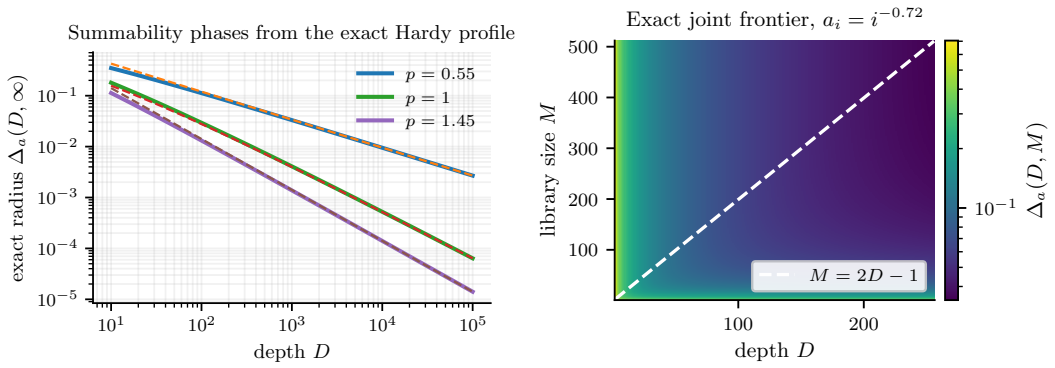



Figure 2: **The spectrum determines both the asymptotic phase and the finite-resource crossover.** Left: the exact Hardy profile for nonsummable, critical, and summable spectra; guides are theorem asymptotes, not fitted lines. Right: the exact $(D, M)$ surface and its saturation boundary $M = 2D-1$.

*Equivalently, for* $S_n = \sum_{i\le n} a_i$,

$$\Delta_a(D,M) = \begin{cases} [S_M + (2D-M)a_{M+1}]/(2D), & M < 2D, \\ S_{2D}/(2D), & M \ge 2D. \end{cases} \tag{14}$$

*An initial segment is optimal; for strictly decreasing spectra,* $2D-1$ *is the smallest universal saturation library.*

The theorem separates omitted mass from finite-depth rounding. Its proof (Section C) uses the largest missing amplitude, a half-grid witness, and a capped weighted rounding problem; the resulting ordered weighted minimax value is adjacent to approximate Carathéodory bounds (Barman, 2015; Mirrokni et al., 2017; Tinarrage, 2026).

With unlimited library,

$$\Delta_a(D,\infty) = \frac{1}{2D}\sum_{i=1}^{2D} a_i, \tag{15}$$

so depth resolves a Hardy/Cesàro profile. Let

$$a_n = n^{-p}L(n), \tag{16}$$

with slowly varying $L$ (Bingham et al., 1987).

**Corollary 5** (Summability phases)**.**

$$0 < p < 1: \quad \Delta_a(D,\infty) \sim \frac{2^{-p}}{1-p} D^{-p} L(D), \tag{17}$$

$$p > 1: \quad \Delta_a(D,\infty) \sim \frac{\sum_i a_i}{2D}. \tag{18}$$

*At* $p = 1$,

$$\Delta_a(D,\infty) = \frac{H_L(2D)}{2D}, \qquad H_L(n) = \sum_{i\le n} \frac{L(i)}{i}. \tag{19}$$

*Thus the critical correction can be* $\log D$, $\log\log D$, *another slowly varying profile, or a finite limit. The three possibilities are separated numerically in Figure 5.*

The structural transition is summability: nonsummable spectra retain a spectrum-dependent exponent, while finite total mass yields universal first-order synthesis.

Figure 2 shows both the asymptotic phases and the exact saturation boundary. For the pure power $a_i \asymp i^{-p}$, the joint rate is

$$\Delta_a(D,M) \asymp \begin{cases} \max\{M^{-p}, D^{-p}\}, & p < 1, \\ \max\{M^{-1}, (\log D)/D\}, & p = 1, \\ \max\{M^{-p}, D^{-1}\}, & p > 1. \end{cases} \tag{20}$$

## 4 NEURAL REALIZATION AND TASK GEOMETRY

The spectral class is realized by a routed residual adapter. Let $\{\varphi_i\}_{i=1}^{J} \subset L^q(P)$ be normalized features, set $a_i = 0$ for $i > J$, and let a depth-$D$ route choose $r_k \in \{0, 1, \dots, \min(M, J)\}$. The additive network writes

$$f_D(x) = \frac{1}{D} \sum_{k=1}^{D} a_{r_k} \varphi_{r_k}(x) e_{r_k}, \qquad a_0 = 0. \tag{21}$$

The output-coordinate separation gives the exact isometry

$$\left\| \sum_i a_i \theta_i \varphi_i e_i \right\|_{L^q(P;\ell_q)} = \left( \sum_i a_i^q |\theta_i|^q \right)^{1/q}. \tag{22}$$

This is a shared library of rank-one residual modes, not a static parameter grid.

**Theorem 6** (Routed residual spectral law)**.** *Let $A_{q,n} = (\sum_{i \le n} a_i^q)^{1/q}$. For the worst-case task-visible radius over the relaxed simplex,*

$$\max\left\{ a_{M+1}, \frac{A_{q,\min\{D,J\}}}{2D} \right\} \le \Delta^{\mathrm{wr}}_{q,J}(D, M) \le a_{M+1} + \frac{A_{q,\min\{D,M,J\}}}{D}. \tag{23}$$

*For $a_i \asymp i^{-p}$ and nonbinding $J$, define*

$$\Psi_{p,q}(D) := D^{-1} \left( \sum_{i \le D} a_i^q \right)^{1/q} \asymp \begin{cases} D^{-(p+1-1/q)}, & p < 1/q, \\ (\log D)^{1/q}/D, & p = 1/q, \\ D^{-1}, & p > 1/q. \end{cases} \tag{24}$$

*In this power-law regime,* (23) *gives $\Delta^{\mathrm{wr}}_{q,J}(D, M) \asymp \max\{a_{M+1}, \Psi_{p,q}(D)\}$, with $a_{M+1} = 0$ for $M \ge J$.*

The critical spectrum therefore depends on the task geometry: $p_{\mathrm{crit}} = 1/q$. The architecture and the resulting task-dependent phase diagram are visualized in Figure 6; full proofs and finite-width saturation are in Section E.

The same phase survives genuine state dependence. Initialize $h_{0,i} = 0$ and observe $f_D(x) = \sum_i h_{D,i} \varphi_i(x) e_i$.

**Theorem 7** (State-dependent neural lifting)**.** *For the coordinate-separable recursion*

$$h_{k+1,i} = h_{k,i} + D^{-1} \mathbf{1}\{r_k = i\} a_i (1 + h_{k,i}), \tag{25}$$

*the ideal relaxed flow is $h_i = e^{a_i \lambda_i} - 1$. If $a_1 \le 1$ and $c_0 = 2 - \sqrt{e}$, then*

$$\max\left\{ a_{M+1}, c_0 \frac{A_{q,\min\{D,J\}}}{D} \right\} \le \Delta^{\mathrm{nl}}_{q,J}(D, M) \le e^{a_1} \left[ a_{M+1} + \left(1 + \frac{a_1}{2}\right) \frac{A_{q,\min\{D,M,J\}}}{D} \right]. \tag{26}$$

The finite-depth nonlinear experiment in Figure 7 verifies the state-dependent phase before the real pretrained-representation study of Section 7. Paired ReLU, exact (erf) GELU, or SiLU units implement each signed scalar mode exactly because $\sigma(t) - \sigma(-t) = t$ for these activations. The architecture can therefore be nonlinear while retaining an exactly analyzable task-visible spectrum.

Finally, loss geometry transforms the distance law. Suppose that, uniformly over the admissible teacher–predictor pairs in the minimax regime, resource-independent constants $c, C > 0$ and $s > 0$ satisfy

$$c\, d(f, f^\star)^s \le L(f; f^\star) - L(f^\star; f^\star) \le C\, d(f, f^\star)^s. \tag{27}$$

Then

$$c\, \Delta_d^s \le E_L \le C\, \Delta_d^s.$$

In particular, squared $L^2$ loss gives a global relation with $c = C = 1$ and $s = 2$, so the three $q = 2$ distance phases become

$$L - L_* \asymp \begin{cases} D^{-(2p+1)}, & p < 1/2, \\ (\log D)/D^2, & p = 1/2, \\ D^{-2}, & p > 1/2, \end{cases} \tag{28}$$

up to the active library floor; a local comparison requires a common neighborhood or a uniform lower separation outside it. Observed exponents therefore combine spectrum, task norm, and loss geometry.

## 5 Execution semantics changes the frontier

Ideal synthesis is only one component of executable error. Linearizing a residual computation around its intended trajectory gives

$$e_{k+1} = A_k e_k + u_k, \tag{29}$$

where $u_k$ is the local arithmetic disturbance and $P_j = A_{D-1} \cdots A_j$ propagates a disturbance injected at time $j$ to the endpoint. Let $C$ be the task observation.

**Theorem 8** (Exact semantics-conditioned Green operators)**.** *For independently bounded write-back disturbances $\|u_j\| \le 1$,*

$$G_D^{\mathrm{WB}}(C) = \sup_{\|\psi\|_* = 1} \sum_{j=1}^{D} \|P_j^* C^* \psi\|_*. \tag{30}$$

*For first-order error feedback $u_j = r_j - r_{j+1}$ with $\|r_j\| \le 1$,*

$$G_D^{\mathrm{EF}}(C) = \sup_{\|\psi\|_* = 1} \left[ \|C^*\psi\|_* + \sum_{j=1}^{D-1} \|(P_{j+1} - P_j)^* C^* \psi\|_* \right]. \tag{31}$$

*Equivalently, summation by parts yields*

$$e_D = -r_D + \sum_{j=1}^{D-1} (P_{j+1} - P_j) r_j. \tag{32}$$

Write-back is governed by the *mass* of the backward propagator path; error feedback is governed by its adjoint *semivariation*. The identities are exact induced norms, not Lipschitz upper bounds.

For a fixed physical horizon, let

$$A_{D,k} = I + H_{D,k}/D, \qquad D^{-1} \sum_k \|H_{D,k}\| \le V, \tag{33}$$

and $\max_k \|H_{D,k}\|/D \le \alpha < 1$.

**Corollary 9** (Residual refinement)**.** *For full-state observation,*

$$e^{-V/(1-\alpha)} D \le G_D^{\mathrm{WB}} \le e^V D, \qquad 1 \le G_D^{\mathrm{EF}} \le 1 + e^V V. \tag{34}$$

No contractivity, normality, commutativity, or time regularity is required. A nonlinear residual network admits the same leading operators through exact secant Jacobians on a controlled trajectory tube (Section F.4).

The Green identities combine sharply with the spectral frontier. In the canonical robust model, local errors range over a symmetric ball of radius $\rho$ after the ideal route is chosen.

**Theorem 10** (Exact spectral–execution law)**.** *For the additive spectral residual class,*

$$E_a^{\mathrm{WB}}(D, M, \rho) = \Delta_a(D, M) + D\rho, \qquad E_a^{\mathrm{EF}}(D, M, \rho) = \Delta_a(D, M) + \rho. \tag{35}$$

The equality follows from an outward-ball identity and a common supporting direction; assumptions and anisotropic variants are in Section G. With $a_i \asymp i^{-p}$ and $\rho_b \asymp 2^{-\kappa b}$,

$$E_p^{\mathrm{WB}} \asymp \max\{M^{-p}, \Psi_p(D), D2^{-\kappa b}\}, \tag{36}$$

$$E_p^{\mathrm{EF}} \asymp \max\{M^{-p}, \Psi_p(D), 2^{-\kappa b}\}, \tag{37}$$

where $\Psi_p(D) = D^{-p}$ for $p < 1$, $(\log D)/D$ for $p = 1$, and $D^{-1}$ for $p > 1$.

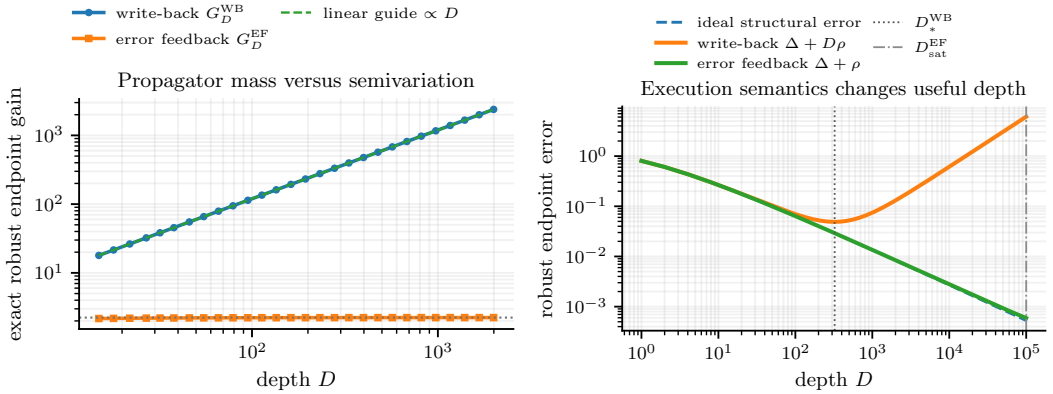



Figure 3: **Changing only execution semantics changes the resource frontier.** Left: on one time-varying nonnormal residual cocycle, write-back mass grows linearly while error-feedback semivariation stays bounded. Right: combining those measured profiles with the same ideal spectral error creates a finite write-back optimum and a feedback precision floor.

**Corollary 11** (Useful-depth change)**.** *Ignoring an already inactive library floor, write-back has $D_* \asymp 2^{\kappa b/(p+1)}$ for $p < 1$ and $D_* \asymp 2^{\kappa b/2}$ for $p > 1$. Error feedback instead reaches its arithmetic floor at $D_{\mathrm{sat}} \asymp 2^{\kappa b/p}$ and $2^{\kappa b}$, respectively. The critical $p = 1$ scales include the Lambert-W logarithmic corrections in Section G.3.*

Figure 3 visualizes this semantics-only intervention on one fixed cocycle and one fixed structural profile.

## 6 From Structural Measurement to Prediction

A mechanistic law should predict resource points whose losses were never used to fit it. The exact spectral frontier at $(D, M)$ requires only the finite prefix

$$N_*(D, M) = \min\{2D, M + 1\}. \tag{38}$$

Indeed, (14) is a positive affine combination of those coefficients, so the map from the required spectrum to $\Delta_a(D, M)$ has $\ell_\infty$ condition number one. For the scalar robust model, define the structural predictor

$$R(D, M, b) = \Delta_a(D, M) + \rho(b) G_D^{\mathrm{sem}}. \tag{39}$$

**Theorem 12** (Finite-horizon structural prediction)**.** *On a prediction region $\Omega$, suppose the required spectral coefficients are estimated within $\varepsilon_a$, the quantizer radius within $\varepsilon_\rho$, and the Green profile within $\varepsilon_G$. If $\rho \le \rho_{\max}$ and $G \le G_{\max}$, then*

$$\sup_{\Omega} |\widehat{R} - R| \le \varepsilon_a + G_{\max}\varepsilon_\rho + \rho_{\max}\varepsilon_G + \varepsilon_\rho\varepsilon_G. \tag{40}$$

*If $L - L_* = \Gamma(R)$ and $\Gamma$ is $L_\Gamma$-Lipschitz with estimation error $\varepsilon_\Gamma$, then*

$$\sup_{\Omega} |\widehat{L} - L| \le \varepsilon_\Gamma + L_\Gamma \sup_{\Omega} |\widehat{R} - R|. \tag{41}$$

No held-out loss exponent appears in the predictor. Directional support-profile versions cover anisotropic outputs. For a fixed terminal direction, one reverse-mode sweep computes all directional Green terms; maximizing over directions is efficient in low-dimensional task quotients and becomes a mixed induced-norm problem in high dimension (Section H.4).

**Theorem 13** (No assumption-free infinite extrapolation)**.** *Any finite positive monotone spectral prefix admits continuations with different regularly varying exponents. Likewise, residual cocycles can agree through depth $D_0$ and have arbitrarily different later Green growth. Hence finite structural measurements cannot certify an unbounded asymptotic law without a quantitative continuation class.*

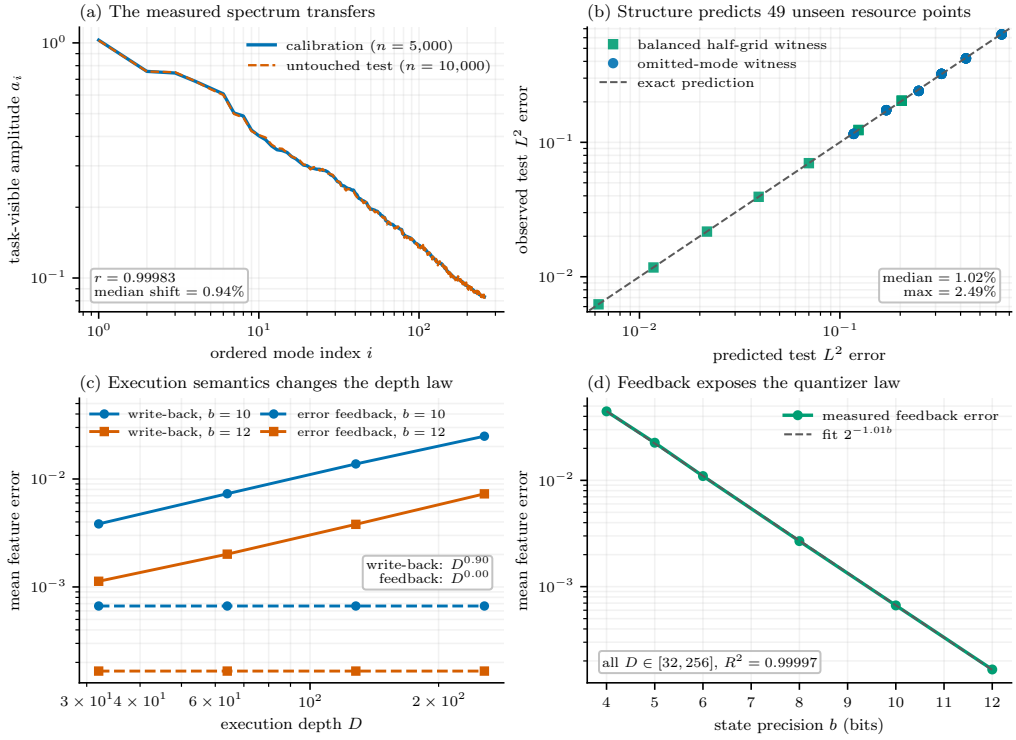



Figure 4: **Independently measured neural structure predicts the resource law.** (a) The calibration spectrum reproduces on untouched test data. (b) It predicts all 49 registered depth–library errors without fitting a held-out exponent. (c) Changing only execution semantics yields growing write-back and bounded feedback laws. (d) Feedback isolates the $2^{-1.01b}$ quantizer law.

Under exponent uncertainty $|\widehat{p} - p| \le \delta_p$, extrapolating a power tail by a factor $H$ necessarily incurs log-relative uncertainty of order $\delta_p \log H$; Figure 8 visualizes the resulting forecast horizon. The finite-prefix theorem and the no-go result together separate two questions: structural coverage certifies an unseen finite resource region, while asymptotic extrapolation requires a continuation model.

If separately trained models obey

$$E_{\text{obs}} = E_{\text{structural}} + \Delta_{\text{train}}, \tag{42}$$

the two nonnegative terms are not identifiable from observed loss alone. The present theory directly predicts compression and execution of a fixed computation, or a family whose structural profiles are measured independently. An external optimization certificate extends the conclusion to retrained families. The next section tests the fixed-computation prediction on a pretrained neural representation.

## 7 A PRETRAINED NEURAL BRIDGE: MEASURE STRUCTURE, THEN PREDICT SCALING

We test the theory on a fixed neural computation outside the analytically constructed class. An ImageNet-pretrained ResNet-18 supplies 512-dimensional CIFAR-10 features; a deterministic blur/downsample transformation gives the degraded view. On 40,000 examples we fit one ridge residual operator $A$ from centered degraded features toward clean features, together with a clean-feature linear classifier. The singular expansion $A = U\,\mathrm{diag}(s_i)V^\top$ gives routed rank-one operations

$$g_i(x) = s_i\langle x - \mu, u_i\rangle v_i, \qquad a_i = s_i\left(\mathbb{E}\langle X - \mu, u_i\rangle^2\right)^{1/2}. \tag{43}$$

The $a_i$ are task-visible $L^2$ amplitudes, not singular values alone. We order 256 modes on 5,000 calibration examples and evaluate every prediction on 10,000 untouched test examples.

**The measured profile transfers.** The calibration and test spectra have Pearson correlation $0.99983$; the median and maximum relative amplitude shifts are $0.94\%$ and $3.36\%$ (Figure 4a). A power fit over modes 32–256 is close to $i^{-0.55}$ on both populations, but the slope changes with the fitting window. Consistent with Theorem 12, the predictive object is the measured profile rather than one selected asymptotic window.

**Structure predicts the depth–library surface.** For every $D, M \in \{4, 8, 16, 32, 64, 128, 256\}$, we register an omitted-mode vertex and a balanced half-grid teacher using calibration structure only. Calibration amplitudes determine routes and predictions, while realized errors use independent test amplitudes. Across all $49$ pairs, the median and maximum relative errors are $1.02\%$ and $2.49\%$, with log-scale correlation $0.99992$ (Figure 4b); no test-grid loss fits a joint exponent. The full map and witness identities appear in Figure 9.

**Execution semantics changes the observed exponent.** We fix a 32-mode balanced teacher, its ideal route, source features, learned operator, and quantizer range, changing only the arithmetic protocol. For $D \geq 32$, write-back feature error scales as $D^{0.901}$ at 10 bits and $D^{0.899}$ at 12 bits, with $R^2 > 0.999$; feedback is depth-independent (Figure 4c). Averaged over $D \in [32, 256]$,

$$E_{\mathrm{EF}} \propto 2^{-1.011b}, \qquad R^2 = 0.99997, \tag{44}$$

matching the quantizer law without the depth factor (Figure 4d). Deterministic nearest-grid executions lie below the robust $D\rho$ and $\rho$ reference envelopes (Figure 10).

**Task geometry remains visible.** The learned correction raises test accuracy from $82.15\%$ to $84.13\%$, recovering $53\%$ of the gap to the $85.89\%$ clean representation, while the execution intervention changes classification accuracy by at most $0.049$ percentage points. Thus task loss can quotient directions visible in representation distance (Figure 11); the full protocol and verification tests are in Section I.

## 8 Related work and conclusion

**Positioning.** Empirical language-model scaling made resource curves operational for model and compute allocation (Kaplan et al., 2020; Hoffmann et al., 2022); mechanistic theories now connect exponents to data dimension, feature and kernel spectra, skill frequencies, language statistics, and task complexity (Sharma & Kaplan, 2022; Bahri et al., 2024; Bordelon et al., 2024; Michaud et al., 2023; Cagnetta et al., 2026; Barkeshli et al., 2026). Precision-aware work models bit width, compression, grouping, and QAT allocation (Kumar et al., 2025; Xu et al., 2024; Frantar et al., 2025; Panferov et al., 2025; Chen et al., 2025; Sun et al., 2025; Dremov et al., 2026). A closely related resource-theoretic treatment is Soltanalian (2026), which studies fixed-library depth–precision tradeoffs and distinguishes write-back from error-feedback execution. Our complementary mechanism is sequential: one reusable operation profile is resolved by finite-depth synthesis, observed through task geometry, and coupled to semantics-dependent error transport. The static geometry connects to metric entropy, nonlinear widths, vector quantization, quantized-network approximation, and approximate Carathéodory theory (Kolmogorov & Tikhomirov, 1961; Pinkus, 1985; Graf & Luschgy, 2000; Gersho & Gray, 1992; Ou et al., 2024; Barman, 2015; Mirrokni et al., 2017; Tinarrage, 2026); our setting is more structured through reusable libraries, equal-weight depth counts, and target-independent library selection. The execution law separates sigma–delta error shaping from error-feedback methods for compressed optimization (Daubechies & DeVore, 2003; Güntürk, 2003; Krahmer et al., 2012; Karimireddy et al., 2019). Section K gives the detailed comparison and the boundary of each exact result.

**Conclusion.** A universal exponent is not the central object of scaling laws. It is the collection of task-visible profiles that determines which resource bottleneck is active. In the solvable neural class, the clipped Cesàro spectrum gives the exact depth–library frontier; Green mass and semivariation determine how arithmetic semantics changes that frontier; and independently measured profiles predict held-out resource points without fitting their slope. This profile-first view also states its own extrapolation limit: finite structural measurements certify a finite resource horizon unless a continuation class is supplied.

## Reproducibility Statement

The supplementary package contains complete proofs, finite-width and bit-accounting details, counterexamples, raw CSV outputs, fixed configurations and hashes, one-command vector-figure generation, exhaustive small-instance regressions, Wolfram verification scripts, and the resumable neural-bridge notebook. The registered full neural run used one NVIDIA A100-SXM4-40GB and is included as an audited result archive. Every numerical value and simulation figure in the paper is regenerated from the supplied outputs and fixed analysis scripts.

## AI Use Statement

Generative AI tools assisted literature discovery, theoretical and proof development, experiment and code development, symbolic and finite-dimensional verification, result interpretation, drafting, and language editing. The authors reviewed all AI-assisted content; mathematical claims are supported by the complete proofs in the appendix, and numerical claims are checked by the supplied scripts and recorded outputs. The authors take responsibility for the final content of the paper.

## References

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.09629.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121.

Maissam Barkeshli, Alberto Alfarano, and Andrey Gromov. On the origin of neural scaling laws: From random graphs to natural language. *arXiv preprint arXiv:2601.10684*, 2026. doi: 10.48550/arXiv.2601.10684.

Siddharth Barman. Approximating nash equilibria and dense bipartite subgraphs via an approximate version of carathéodory's theorem. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 361–369, 2015. doi: 10.1145/2746539.2746566.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1987. doi: 10.1017/CBO9780511721434.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382, 2024. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart. Deriving neural scaling laws from the statistics of natural language. *arXiv preprint arXiv:2602.07488*, 2026. doi: 10.48550/arXiv.2602.07488.

Mengzhao Chen, Chaoyi Zhang, Jing Liu, Yutao Zeng, Zeyue Xue, Zhiheng Liu, Yunshui Li, Jin Ma, Jie Huang, Xun Zhou, and Ping Luo. Scaling law for quantization-aware training. *arXiv preprint arXiv:2505.14302*, 2025. doi: 10.48550/arXiv.2505.14302.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6571–6583, 2018.

Ingrid Daubechies and Ronald DeVore. Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma–delta modulators of arbitrary order. *Annals of Mathematics*, 158(2):679–710, 2003. doi: 10.4007/annals.2003.158.679.

Aleksandr Dremov, David Grangier, Angelos Katharopoulos, and Awni Hannun. Compute-optimal quantization-aware training. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6e1adf459d1901edf173252b6bd40cb9-Abstract-Conference.html.

Elias Frantar, Utku Evci, Wonpyo Park, Neil Houlsby, and Dan Alistarh. Compression scaling laws: Unifying sparsity and quantization. *arXiv preprint arXiv:2502.16440*, 2025.

Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, Berlin, 2000. doi: 10.1007/BFb0103945.

C. Sinan Güntürk. One-bit sigma–delta quantization with exponential accuracy. *Communications on Pure and Applied Mathematics*, 56(11):1608–1630, 2003. doi: 10.1002/cpa.3044.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34 (1):014004, 2018. doi: 10.1088/1361-6420/aa9a90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. doi: 10.48550/arXiv.2001.08361.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261, 2019.

A. N. Kolmogorov and V. M. Tikhomirov. $\varepsilon$-entropy and $\varepsilon$-capacity of sets in functional spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961. doi: 10.1090/trans2/017/10. English translation of Uspekhi Mat. Nauk 14(2), 3–86 (1959).

Felix Krahmer, Rayan Saab, and Rachel Ward. Root-exponential accuracy for coarse quantization of finite frame expansions. *IEEE Transactions on Information Theory*, 58(2):1069–1079, 2012. doi: 10.1109/TIT.2011.2168942.

Tanishq Kumar, Zachary Ankner, Benjamin Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wg1PCg3CUP.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28699–28722, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/5b6346a05a537d4cdb2f50323452a9fe-Abstract-Conference.html.

Vahab Mirrokni, Renato Paes Leme, Adrian Vladu, and Sam Chiu wai Wong. Tight bounds for approximate carathéodory and beyond. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2440–2449, 2017.

Weigutian Ou, Philipp Schenkel, and Helmut Bölcskei. Three quantization regimes for ReLU networks. *arXiv preprint arXiv:2405.01952*, 2024. doi: 10.48550/arXiv.2405.01952.

Andrei Panferov, Alexandra Volkova, Ionut-Vlad Modoranu, Vage Egiazarian, Mher Safaryan, and Dan Alistarh. Unified scaling laws for compressed representations. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=24wDPGiDzA.

Allan Pinkus. *$n$-Widths in Approximation Theory*, volume 7 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 1985. doi: 10.1007/978-3-642-69894-1.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *Journal of Machine Learning Research*, 23(9):1–34, 2022. Preprint arXiv:2004.10802.

Mojtaba Soltanalian. When can depth replace precision? a resource theory of quantized neural computation. *arXiv preprint arXiv:2607.23390*, 2026. doi: 10.48550/arXiv.2607.23390.

Xingwu Sun, Shuaipeng Li, Ruobing Xie, Weidong Han, Kan Wu, Zhen Yang, Yixing Li, An Wang, Shuai Li, Jinbao Xue, Yu Cheng, Yangyu Tao, Zhanhui Kang, Cheng-Zhong Xu, Di Wang, and Jie Jiang. Scaling laws for floating–point quantization training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57544–57570, 2025. URL https://proceedings.mlr.press/v267/sun25j.html.

Raphaël Tinarrage. Sharp approximate carathéodory theorem and application to iterated delaunay refinement. *arXiv preprint arXiv:2606.25854*, 2026. doi: 10.48550/arXiv.2606.25854.

Zifei Xu, Alexander Y. Lan, Wanzin Yazar, Tristan Webb, Sayeh Sharify, and Xin Wang. Scaling laws for post-training quantized large language models. In *Proceedings of the 4th NeurIPS Efficient Natural Language and Speech Processing Workshop*, volume 262 of *Proceedings of Machine Learning Research*, pp. 270–285, 2024. URL https://proceedings.mlr.press/v262/xu24a.html.

Shihao Zhang, Haoyu Zhang, Ian Colbert, and Rayan Saab. Qronos: Correcting the past by shaping the future... in post-training quantization. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10011286.

# A NOTATION, RESOURCE SEMANTICS, AND THEOREM SCOPE

This appendix fixes the quantifiers and resource semantics used throughout the paper. Several attractive but false scaling statements arise from silently changing one of these choices.

## A.1 DIRECTED APPROXIMATION AND TASK QUOTIENTS

Let $(Y, d)$ be a pseudometric space. For nonempty $A, B \subset Y$,

$$\operatorname{exc}(A, B) := \sup_{a \in A} \inf_{b \in B} d(a, b) \in [0, \infty].$$

The directed excess is appropriate when every ideal target must have an implementation approximation and extra implementation behaviors are harmless. The symmetric Hausdorff distance is $d_H(A, B) = \max\{\operatorname{exc}(A, B), \operatorname{exc}(B, A)\}$ and is appropriate when spurious reachable behaviors are themselves a risk.

If a task observes $\mathcal{O} : Y \to Z$ with metric $\rho$, define $d_{\mathsf{T}}(y, y') = \rho(\mathcal{O}y, \mathcal{O}y')$. Then, directly from the definitions,

$$\operatorname{exc}_{d_{\mathsf{T}}}(A, B) = \operatorname{exc}_{\rho}(\mathcal{O}A, \mathcal{O}B). \tag{45}$$

Thus all widths may be formed after quotienting by task equivalence. We retain a pseudometric notation because the hidden trajectories may differ while their observed outputs coincide.

## A.2 LIBRARY, TARGET, ROUTE, AND ARITHMETIC QUANTIFIERS

The universal shared-library minimax game is

$$\inf_{Q:\, |Q| \le M} \sup_{f^* \in \mathcal{A}_\infty(\mathcal{U})} \inf_{f \in \mathcal{A}_D(Q)} d_{\mathsf{T}}(f^*, f). \tag{46}$$

The library $Q$ is selected once, before the target. The route or schedule may adapt to the target. A target-adaptive code instead uses $\sup_{f^*} \inf_Q \inf_f$ and can be strictly easier; randomized libraries produce a third game. The inequalities between these variants are recorded in Section B.5.

For robust arithmetic, the quantifiers are

$$\inf_{Q} \sup_{f^*} \inf_{\text{route}} \sup_{\text{admissible primitive errors}} d_{\mathsf{T}}(f^*, \widehat{f}). \tag{47}$$

The exact lower arithmetic terms in Theorem 10 use a symmetric full-ball disturbance language, which permits the adversary to align a support direction with the structural residual. Deterministic nearest rounding, stochastic rounding, or restricted vector quantizers define different primitive error sets and may have different two-sided laws.

## A.3 RESOURCE DICTIONARY

| symbol | resource or structural quantity |
|---|---|
| $D$ | number of sequential residual micro-layers / route decisions |
| $M$ | number of nonzero reusable operation modes in one universal library |
| $b_{\text{route}}$ | route-address width, $\lceil \log_2(M+1) \rceil$ including the zero mode |
| $J$ | available neural mode width; finite and explicitly charged |
| $b_{\text{state}}$ | state/arithmetic bit width under a declared quantizer |
| $\rho(b, J, \|\cdot\|)$ | actual primitive quantizer radius in the chosen state norm |
| $a_i$ | task-visible ordered operation amplitude/singular gain |
| $G_D^{\text{sem}}$ | endpoint amplification for the declared execution semantics |

The equality $M \simeq 2^{b_{\text{route}}}$ counts addressable modes, not the storage cost of their basis vectors or amplitudes. If a coordinatewise quantizer has scalar step $\delta_b$ in dimension $J$, its global $\ell_q$ radius may be $J^{1/q}\delta_b$. All arithmetic theorems use $\rho$ as the primitive and specialize to $2^{-\kappa b}$ only under an explicit quantizer model. Error feedback also requires a carry; bounded-carry conclusions assume that the carry is represented at the declared radius and does not suffer an additional unmodeled write-back.

### A.4 Relation to fixed-library reachability

Fixed-library reachability asks how a declared operation family approaches its relaxed infinite-depth class and how finite arithmetic perturbs that approximation (Soltanalian, 2026). Shared-library scaling adds a distinct optimization layer: the reusable library itself is selected under a cardinality budget before the target is revealed. The resulting quantifier order in (46) makes the depth-dependent library width $W^{\rightarrow}_{D,M}$ the relevant joint object. The exact clipped-Cesàro law, neural lifting, Green identities, and structural prediction theorems specialize this joint object in progressively more concrete settings. Section K records the assumptions under which those specializations remain valid.

### A.5 Notation and definitions

The main text and preceding appendix sections define notation locally as each object is introduced. For reference, this table collects the principal paper-specific quantities and derived objects used in the main results.

| symbol | meaning | symbol | meaning |
|---|---|---|---|
| $\mathcal{A}_\infty(\mathcal{U})$ | ideal reachable class with unlimited depth | $\mathcal{A}_D(Q)$ | depth-$D$ class generated by reusable library $Q$ |
| $\mathcal{A}_\infty(Q)$ | infinite-depth class generated by library $Q$ | $h(Q)$ | relaxed library defect |
| $\pi^{\rightarrow}_D(Q)$ | forward finite-depth synthesis cost | $\bar{\pi}_D(Q)$ | reverse finite-depth excess |
| $W_D(Q)$ | fixed-library approximation width | $W^{\rightarrow}_{D,M}$ | shared-library reachable width |
| $a_D, \bar{a}_D$ | forward/reverse full-family depth defects | $q_{D,M}$ | depth-$D$ library approximation width |
| $\mathcal{K}_a$ | canonical spectral ideal class | $\Delta_a(D,M)$ | canonical depth–library approximation frontier |
| $N^*(D,M)$ | required spectral prefix, $\min\{2D, M+1\}$ | $A_{q,n}$ | neural $L_q$ aggregation of the first $n$ amplitudes |
| $P_j$ | backward error propagator from stage $j$ to the endpoint | $G^{\mathrm{WB}}_D$, $G^{\mathrm{EF}}_D$ | write-back and error-feedback Green profiles |
| $\rho(b)$ | primitive arithmetic disturbance radius | $\epsilon_a, \epsilon_\rho, \epsilon_G$ | spectral, quantizer, and Green-profile estimation errors |
| $R(D,M,b)$ | structural/execution predictor | $\widehat{R}(D,M,b)$ | plug-in predictor estimate |

## B No-go constructions and the shared-library width

### B.1 A corrected finite-depth triangle

Let

$$K = \mathcal{A}_\infty(\mathcal{U}), \qquad B_D = \mathcal{A}_D(\mathcal{U}), \qquad C_{D,Q} = \mathcal{A}_D(Q).$$

Define

$$a_D = \operatorname{exc}(K, B_D), \qquad \bar{a}_D = \operatorname{exc}(B_D, K),$$
$$q_{D,M} = \inf_{|Q|\le M} \operatorname{exc}(B_D, C_{D,Q}), \qquad W_{D,M} = \inf_{|Q|\le M} \operatorname{exc}(K, C_{D,Q}).$$

**Proposition 14** (Corrected reachable-width sandwich)**.** *For every $D, M$ for which $q_{D,M}$ and $\bar{a}_D$ are finite,*

$$\max\{a_D, [q_{D,M} - \bar{a}_D]_+\} \le W_{D,M} \le a_D + q_{D,M}. \tag{48}$$

*If $C_{D,Q} \subseteq B_D \subseteq K$ for every $Q$, then*

$$\max\{a_D, q_{D,M}\} \le W_{D,M} \le a_D + q_{D,M} \le 2\max\{a_D, q_{D,M}\}. \tag{49}$$

*Proof.* The upper bound follows from the directed triangle inequality through $B_D$ and then taking the library infimum. Since $C_{D,Q} \subseteq B_D$, whenever this nesting holds, the supremum defining $W_{D,M}$ contains the one defining $a_D$, and the supremum defining $q_{D,M}$ is over a subset of $K$, giving the

two lower bounds. Without $B_D \subseteq K$, the directed triangle inequality $q_{D,M} \le \bar{a}_D + W_{D,M}$ yields, for finite $q_{D,M}$ and $\bar{a}_D$, only $W_{D,M} \ge [q_{D,M} - \bar{a}_D]_+$. The correction is necessary: take $K = \{0\}$, $B_D = \{0, 1\}$, and $C_{D,Q} = \{0\}$. Then $a_D = 0$, $q_{D,M} = 1$, $W_{D,M} = 0$, and $\bar{a}_D = 1$. □

### B.2 Same marginals, different joint laws

For $t \in [0, 1]$, let $\varphi_t(s) = \mathbf{1}_{[0,t]}(s)$ on $[0, 1]$. Then

$$\|\varphi_t - \varphi_{t'}\|_{L^2} = |t - t'|^{1/2}. \tag{50}$$

Use the task space

$$Y = \mathbb{R} \times \mathbb{R} \times L^2([0, 1]), \qquad d((x, y, f), (x', y', f')) = \max\{|x - x'|, |y - y'|, \|f - f'\|_2\},$$

and ideal class

$$K = \{(x, y, \varphi_t) : x, y, t \in [0, 1]\}.$$

The common operation family is $\mathcal{U} = \{0, 1\}^3 \times [0, 1]^2$. Write $q = (b_0, b_1, b_2, r, s)$. A finite $Q \subset \mathcal{U}$ is selected once and shared over all stages.

Fix $D \ge 2$ and initialize counters $N_0 = N_1 = N_2 = 0$. During stages $1, \dots, D - 1$, selecting $q_k$ performs $N_i \leftarrow N_i + b_{i,k}$. At the final stage, selecting $q_D$ writes $R \leftarrow r_D$ and $S \leftarrow s_D$. The two systems differ only in readout:

$$F_D^{\mathrm{dec}} = \left( \frac{N_0}{D - 1}, R, \varphi_{(DN_1 + N_2)/(D^2 - 1)} \right), \tag{51}$$

$$F_D^{\mathrm{ph}} = \left( \frac{N_0}{D - 1}, R, \varphi_{(N_1 + S)/D} \right). \tag{52}$$

Let $\mathcal{A}_D^j(Q)$ be the corresponding output set, and set

$$W_{D,M}^j = \inf_{|Q| \le M} \mathrm{exc}(K, \mathcal{A}_D^j(Q)).$$

Define the boundary marginals

$$a_D^j = \mathrm{exc}(K, \mathcal{A}_D^j(\mathcal{U})), \qquad c_M^j = \inf_{|Q| \le M} \mathrm{exc}\left( K, \overline{\bigcup_{D \ge 2} \mathcal{A}_D^j(Q)} \right).$$

**Theorem 15** (Strong marginal-to-joint no-go)**.** *For $D, M$ larger than fixed constants,*

$$a_D^{\mathrm{dec}} \asymp a_D^{\mathrm{ph}} \asymp D^{-1}, \qquad c_M^{\mathrm{dec}} \asymp c_M^{\mathrm{ph}} \asymp M^{-1},$$

*while*

$$W_{D,M}^{\mathrm{dec}} \asymp \max\{D^{-1}, M^{-1}\}, \tag{53}$$

$$W_{D,M}^{\mathrm{ph}} \asymp \max\{D^{-1}, M^{-1}, (DM)^{-1/3}\}. \tag{54}$$

*Proof.* Eight binary control patterns realize arbitrary counter triples in $\{0, \dots, D-1\}^3$: independently choose the stages on which each bit is one.

For the decoupled system, reserve those patterns and use the remaining atoms as a uniform $R$-grid. The $x$ mesh is $O(D^{-1})$. The integer $DN_1 + N_2$ ranges over $0, \dots, D^2 - 1$, so the raw threshold mesh is $O(D^{-2})$ and, by (50), the function error is $O(D^{-1})$. The $y$ error is $O(M^{-1})$. Conversely, the $x$ projection has at most $D$ values and the $y$ projection at most $M$ values, yielding lower bounds of order $D^{-1}$ and $M^{-1}$.

For the phase system, a final atom with metadata $(r_j, s_j)$ yields projected pairs

$$\left( r_j, \frac{n + s_j}{D} \right), \qquad 0 \le n < D.$$

Thus at most $MD$ points cover the $(y, t)$ square. Under

$$\rho((y, t), (y', t')) = \max\{|y - y'|, |t - t'|^{1/2}\},$$

a radius-$\varepsilon$ ball has Lebesgue area at most $4\varepsilon^3$. Covering the unit square gives $1 \le 4MD\varepsilon^3$, hence $\varepsilon \gtrsim (DM)^{-1/3}$. Coordinate projections again yield $D^{-1}$ and $M^{-1}$ lower bounds.

For the upper bound, choose $J$ uniform $r$ centers and $L$ uniform phases, using $JL$ metadata pairs. The $y$ error is $O(J^{-1})$. The values

$$\frac{n + (b + 1/2)/L}{D} = \frac{nL + b + 1/2}{DL}$$

are midpoints of the $DL$ equal subintervals, so threshold error is $O((DL)^{-1/2})$. Therefore

$$W^{\mathrm{ph}}_{D,M} \lesssim \max\{D^{-1}, J^{-1}, (DL)^{-1/2}\}, \qquad JL \lesssim M.$$

An error $E$ requires $J \gtrsim E^{-1}$ and $L \gtrsim (DE^2)^{-1}$, hence $M \gtrsim (DE^3)^{-1}$. Conversely choose $J \asymp (DM)^{1/3}$ and $L \asymp M/J$ when $D \le M^2$, and $J = M, L = 1$ otherwise. This proves (54).

With the full operation family, the $x$ projection forces $D^{-1}$ in both systems. The decoupled threshold grid has $L^2$ error $O(D^{-1})$, while the phase system can choose $S$ to make $t$ exact; $R$ is exact. Hence $a^j_D \asymp D^{-1}$. At infinite depth, a fixed binary control sublibrary makes the counter readouts dense while $R$ still assumes at most $M$ values. A uniform $R$ grid attains and requires $M^{-1}$, proving the library marginals. □

*Proof of Theorem 2.* The two systems in Theorem 15 have the common operation family, identical $D^{-1}$ and $M^{-1}$ marginals, and the two joint laws stated in Theorem 2. The main-text theorem is therefore an immediate specialization. □

The theorem rules out any universal map from the two boundary profiles $\{a_D\}_D$ and $\{c_M\}_M$ to the joint surface. The missing information is the depth-dependent decoder geometry induced by reusing the same library.

### B.3 The operational sandwich

For a fixed $Q$, abbreviate

$$W_D(Q) = \mathrm{exc}(\mathcal{A}_\infty(\mathcal{U}), \mathcal{A}_D(Q)),$$

and define

$$\begin{aligned} h(Q) &= \mathrm{exc}(\mathcal{A}_\infty(\mathcal{U}), \mathcal{A}_\infty(Q)),\\ \pi^{\rightarrow}_D(Q) &= \mathrm{exc}(\mathcal{A}_\infty(Q), \mathcal{A}_D(Q)),\\ \bar{\pi}_D(Q) &= \mathrm{exc}(\mathcal{A}_D(Q), \mathcal{A}_\infty(Q)). \end{aligned}$$

Assume $\mathcal{A}_\infty(Q) \subseteq \mathcal{A}_\infty(\mathcal{U})$.

*Proof of Theorem 3.* The upper bound is the directed triangle inequality through $\mathcal{A}_\infty(Q)$. Because the restricted ideal class lies in the full ideal class, the supremum defining $W_D(Q)$ includes the one defining $\pi^{\rightarrow}_D(Q)$. Finally,

$$h(Q) \le W_D(Q) + \bar{\pi}_D(Q)$$

by the directed triangle inequality through $\mathcal{A}_D(Q)$. When $h(Q)$ and $\bar{\pi}_D(Q)$ are finite, this gives

$$W_D(Q) \ge [h(Q) - \bar{\pi}_D(Q)]_+.$$

If $\mathcal{A}_D(Q) \subseteq \mathcal{A}_\infty(Q)$, then $\bar{\pi}_D(Q) = 0$ and

$$W_D(Q) \ge \max\{h(Q), \pi^{\rightarrow}_D(Q)\} \ge \frac{1}{2}\big(h(Q) + \pi^{\rightarrow}_D(Q)\big).$$

Taking infima yields (8). □

### B.4 When an additive law is justified

**Proposition 16** (Uniform common-library criterion)**.** *Let*

$$\mathcal{Q}_M = \{Q \subseteq \mathcal{U} : \ |Q| \le M\},$$

*and assume*

$$\mathcal{A}_D(Q) \subseteq \mathcal{A}_\infty(Q) \subseteq \mathcal{A}_\infty(\mathcal{U}) \qquad \text{for all } Q \in \mathcal{Q}_M.$$

*Suppose there is a sequence $Q_M \in \mathcal{Q}_M$ such that, uniformly for sufficiently large $D$ and $M$,*

$$h(Q_M) \le C_h M^{-\gamma}, \qquad \pi^{\rightarrow}_D(Q_M) \le C_\pi D^{-\theta},$$

*and every $Q \in \mathcal{Q}_M$ satisfies*

$$\max\{h(Q), \pi^{\rightarrow}_D(Q)\} \ge c \max\{M^{-\gamma}, D^{-\theta}\}.$$

*Then*

$$W_{D,M} \asymp \max\{M^{-\gamma}, D^{-\theta}\} \asymp M^{-\gamma} + D^{-\theta}.$$

*Proof.* For the selected library $Q_M$, the operational sandwich gives

$$W_{D,M} \le h(Q_M) + \pi^{\rightarrow}_D(Q_M) \le C_h M^{-\gamma} + C_\pi D^{-\theta}.$$

For any $Q \in \mathcal{Q}_M$, the nesting assumption gives

$$\operatorname{exc}\big(\mathcal{A}_\infty(\mathcal{U}), \mathcal{A}_D(Q)\big) \ge \max\{h(Q), \pi^{\rightarrow}_D(Q)\}.$$

Hence

$$W_{D,M} \ge c \max\{M^{-\gamma}, D^{-\theta}\}.$$

Combining the bounds and using

$$\max\{x, y\} \le x + y \le 2 \max\{x, y\}, \qquad x, y \ge 0,$$

gives the result. □

### B.5 Task quotient and library variants

Equation (45) follows because

$$\inf_{b \in B} d_{\mathsf{T}}(a, b) = \inf_{z \in \mathcal{O}B} \rho(\mathcal{O}a, z),$$

then the supremum over $a \in A$ depends only on $\mathcal{O}A$.

For a target $f$ and library $Q$, write $r(f, Q) = d(f, \mathcal{A}_D(Q))$. Define

$$W_{\mathrm{ad}} = \sup_f \inf_Q r(f, Q), \quad W_{\mathrm{mix}} = \inf_\mu \sup_f \mathbb{E}_{Q \sim \mu} r(f, Q), \quad W_{\mathrm{det}} = \inf_Q \sup_f r(f, Q).$$

Then $W_{\mathrm{ad}} \le W_{\mathrm{mix}} \le W_{\mathrm{det}}$ by moving infima outward. Both inequalities can be strict: for $K = \{-1, +1\}$ and one-point libraries $\{-1\}$ or $\{+1\}$ under ordinary distance, the three values are $0, 1, 2$. If the random library is revealed before the adversarial target, randomization gives no minimax benefit because expectation cannot improve the best deterministic worst case.

## C Proof of the exact clipped-Cesàro theorem

We prove Theorem 4 in three steps: compress the omitted tail, solve the fixed-library rounding game exactly, and optimize the first omitted index. The argument is finite and constructive; closure of $K_a$ only permits limits of finite-support targets.

### C.1 TAIL COMPRESSION

Fix a library $I$ with selected amplitudes, sorted as $b_1 \geq \cdots \geq b_m > 0$. Let

$$c = \max\{a_j : j \notin I\}$$

be the largest omitted amplitude. If the complement is empty, set $c = 0$. For selected occupation coefficients $\lambda_i$ and total omitted occupation $\tau$, define

$$\rho_D(b; c) = \sup_{\substack{\lambda_i \geq 0,\ \tau \geq 0 \\ \sum_i \lambda_i + \tau \leq 1}} \min_{\substack{n_i \in \mathbb{N}_0 \\ \sum_i n_i \leq D}} \left[ \sum_{i=1}^{m} b_i \left| \lambda_i - \frac{n_i}{D} \right| + c\tau \right]. \tag{55}$$

**Lemma 17** (Tail compression)**.** *The covering radius of the fixed library $I$ equals $\rho_D(b; c)$.*

*Proof.* For $x \in K_a$, write the selected occupations as $\lambda_i$ and the omitted total as $\tau = \sum_{j \notin I} \lambda_j$. The omitted-coordinate error is

$$\sum_{j \notin I} a_j \lambda_j \leq c\tau,$$

which gives the upper reduction. Conversely, for any feasible $(\lambda, \tau)$, place all omitted occupation on an omitted coordinate attaining $c$; the $\ell_1$ errors add coordinatewise, giving equality in the supremum. □

### C.2 EXACT FIXED-LIBRARY ROUNDING RADIUS

**Theorem 18** (Fixed-library formula)**.** *For every $m, D \geq 1$,*

$$\boxed{\rho_D(b; c) = c + \frac{1}{2D} \sum_{i=1}^{\min\{m, 2D\}} (b_i - c)_+.} \tag{56}$$

*Proof. Lower bound.* Let

$$J = \{i \leq \min(m, 2D) : b_i > c\}, \qquad k = |J|.$$

Choose

$$\lambda_i = \frac{1}{2D} \mathbf{1}_{\{i \in J\}}, \qquad \tau = 1 - \frac{k}{2D}.$$

This is feasible. For every integer $n_i \geq 0$, $|1/(2D) - n_i/D| \geq 1/(2D)$. Hence every depth grid point incurs error at least

$$\begin{aligned} \frac{1}{2D} \sum_{i \in J} b_i + c \left(1 - \frac{k}{2D}\right) &= c + \frac{1}{2D} \sum_{i \in J} (b_i - c) \\ &= c + \frac{1}{2D} \sum_{i=1}^{\min(m, 2D)} (b_i - c)_+. \end{aligned}$$

*Upper bound.* For a feasible target set

$$t_i = D\lambda_i, \quad u = D\tau, \quad q_i = \lfloor t_i \rfloor, \quad f_i = t_i - q_i \in [0, 1), \quad R = D - \sum_i q_i.$$

Then $R \in \{0, \ldots, D\}$ and

$$\sum_i f_i + u \leq R. \tag{57}$$

If $R = 0$, every $f_i$ and $u$ vanishes. Assume $R \geq 1$. We may round coordinate $i$ upward by choosing $z_i \in \{0, 1\}$, with $\sum_i z_i \leq R$. For binary $z_i$,

$$|f_i - z_i| = f_i + z_i(1 - 2f_i).$$

Thus the residual cost multiplied by $D$ is the affine function

$$\widetilde{L}(z) = \sum_i b_i \big[f_i + z_i(1-2f_i)\big] + cu \quad \text{on} \quad \mathcal{Z}_R = \{z \in \{0,1\}^m : \sum_i z_i \le R\}. \tag{58}$$

The convex hull of $\mathcal{Z}_R$ is the integral uniform-matroid polytope

$$P_R = \{p \in [0,1]^m : \sum_i p_i \le R\}.$$

Because (58) is affine,

$$\min_{z \in \mathcal{Z}_R} \widetilde{L}(z) = \min_{p \in P_R} \widetilde{L}(p). \tag{59}$$

Extend $b_j = 0$ for $j > m$, set $h = \max\{c, b_{2R}\}$, and choose

$$p_i = \frac{(b_i - h)_+}{2b_i},$$

with $p_i = 0$ if $b_i = 0$. At most $2R$ entries are nonzero and each is at most $1/2$, hence $p \in P_R$. Direct expansion gives

$$\begin{aligned} \widetilde{L}(p) &= \frac{1}{2}\sum_i (b_i - h)_+ + \sum_i \min\{b_i, h\} f_i + cu \\ &\le \frac{1}{2}\sum_i (b_i - h)_+ + hR, \end{aligned} \tag{60}$$

where (57) and $c \le h$ were used. If $h = c$, the right-hand side is

$$Rc + \frac{1}{2}\sum_{i=1}^{\min(m,2R)} (b_i - c)_+.$$

If $h = b_{2R} > c$, it is $\frac{1}{2}\sum_{i=1}^{2R} b_i$, equal to the same expression after cancellation of $Rc$. Therefore

$$\min_z \widetilde{L}(z) \le Rc + \frac{1}{2}\sum_{i=1}^{\min(m,2R)} (b_i - c)_+. \tag{61}$$

The right side is nondecreasing in the integer $R$: increasing $R$ by one adds $c$ plus at most two nonnegative half-increments. It is therefore bounded by its value at $D$. Divide by $D$ to obtain (56). □

**Constructive rounding.** The proof is algorithmic. After flooring $D\lambda_i$, minimize the affine cost in (58) over a uniform matroid, equivalently select up to $R$ coordinates with the largest positive cost reduction $b_i(2f_i - 1)$. The threshold point $p$ is used only to prove the worst-case value; the actual minimizing binary vector is obtained by sorting these reductions.

### C.3 Optimization over universal libraries

Let $k$ be the first omitted index of a library. Then $1, \dots, k-1$ are selected, $k \le M+1$, the largest omitted amplitude is $a_k$, and selected atoms after $k$ have amplitude at most $a_k$, so they do not enter the positive-part sum.

**Lemma 19** (First-omission formula)**.** *Every library whose first omitted index is $k$ has radius*

$$F_k = a_k + \frac{1}{2D}\sum_{i=1}^{\min\{k-1,2D\}} (a_i - a_k). \tag{62}$$

*For $k < 2D$,*

$$F_{k+1} - F_k = \frac{2D-k}{2D}(a_{k+1} - a_k) \le 0, \tag{63}$$

*and for $k \ge 2D$, $F_k = S_{2D}/(2D)$.*

*Proof.* Equation (62) is Theorem 18: the selected amplitudes strictly above $a_k$ are exactly among $a_1, \dots, a_{k-1}$. Direct subtraction gives (63). If $k \geq 2D$, all $a_k$ terms cancel, leaving $S_{2D}/(2D)$. □

*Proof of Theorem 4.* The first omitted index obeys $k \leq M + 1$. By (63), pushing the first omission to the right cannot increase the radius before the plateau; after $k = 2D$ the radius is constant. Taking $k = M + 1$ yields (12). The piecewise and clipped forms follow algebraically. At $M = 2D - 1$, the first branch equals $S_{2D}/(2D)$; strict decrease shows that no smaller universal library can saturate for every decreasing spectrum. □

### C.4 Sanity checks and edge cases

If $M = 0$ were allowed, the radius would be $a_1$ because only the zero atom remains. For $D = 1$, (13) averages the two largest clipped amplitudes. Ties in $a_i$ can make non-initial libraries optimal, but the value is unchanged. For a finite spectrum of length $J$, set $a_i = 0$ for $i > J$; the formula then includes width saturation automatically.

## D Regular variation, critical corrections, and broader geometries

Let $S_n = \sum_{i \leq n} a_i$ and $\Psi(n) = S_n/n$. The exact unrestricted depth law is $\Delta_a(D, \infty) = \Psi(2D)$. This section records its asymptotic consequences.

### D.1 Karamata phases

Assume $a_n = n^{-p} L(n)$ with $L$ slowly varying and $a_n$ eventually nonincreasing.

*Proof of Theorem 5.* For $0 < p < 1$, Karamata's sum theorem gives

$$S_n \sim \frac{n^{1-p} L(n)}{1 - p}.$$

Therefore

$$\Delta_a(D, \infty) = \frac{S_{2D}}{2D} \sim \frac{(2D)^{-p} L(2D)}{1 - p} \sim \frac{2^{-p}}{1 - p} D^{-p} L(D).$$

If $p > 1$, the monotone regularly varying sequence is summable and $S_{2D} \to S_\infty$, yielding $S_\infty/(2D)$. At $p = 1$, no further universal simplification is valid:

$$S_n = H_L(n) = \sum_{i \leq n} \frac{L(i)}{i},$$

which gives (19) exactly. □

The same proof covers $p = 0$ whenever $L$ is decreasing to zero and Karamata's theorem applies: $S_n \sim nL(n)$ and the depth profile is slowly varying rather than algebraic.

### D.2 Critical examples

The following examples illustrate why the critical theorem must retain $H_L$.

1. If $L(n) \to c > 0$, then $H_L(n) \sim c \log n$ and

$$\Delta_a(D, \infty) \sim \frac{c \log(2D)}{2D}.$$

2. If $L(n) = 1/\log(e + n)$, then $H_L(n) \sim \log \log n$ and

$$\Delta_a(D, \infty) \sim \frac{\log \log(2D)}{2D}.$$

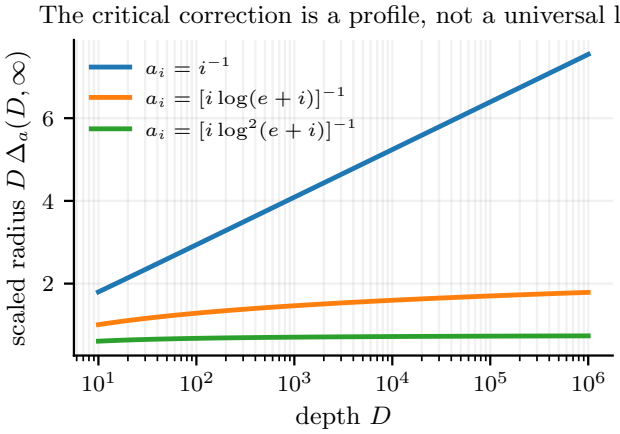



Figure 5: **The critical index does not imply a universal logarithm.** All curves have the nominal form $a_n = L(n)/n$, yet the exact depth error $H_L(2D)/(2D)$ exhibits $\log D/D$, $\log\log D/D$, and summable $D^{-1}$ behavior. The contribution is to replace a brittle critical exponent statement by the integrated slowly varying profile.

3. If $L(n) = \log(n)^\beta$ with $\beta > -1$, then
$$H_L(n) \sim \frac{\log(n)^{\beta+1}}{\beta+1}.$$
For $\beta = -1$, the correction is $\log\log n$.
4. If $L(n) = 1/\log(e+n)^{1+\eta}$ with $\eta > 0$, then $\sum_n L(n)/n < \infty$, so the nominal index $p = 1$ is already in the summable $D^{-1}$ phase.

### D.3 Finite-library crossovers

From (14), the joint law is exact at every $(D, M)$. For a regularly varying nonsummable spectrum and $M < 2D$,
$$\Delta_a(D, M) = \frac{S_M + (2D - M)a_{M+1}}{2D}.$$
If $0 < p < 1$ and $M/D \to \lambda \in (0, 2)$, Karamata's theorem gives
$$\Delta_a(D, M) \sim D^{-p} L(D) \left[\frac{\lambda^{1-p}}{2(1-p)} + \left(1 - \frac{\lambda}{2}\right)\lambda^{-p}\right]. \tag{64}$$
For $M/D \to \infty$, the unrestricted constant $2^{-p}/(1-p)$ is recovered. Thus the exact theorem yields not only exponents but ray-dependent constants throughout the joint scaling plane.

At the pure critical law $a_i \sim c/i$, balancing the library floor $M^{-1}$ with $(\log D)/D$ gives $D \asymp M \log M$ up to lower-order logarithms. For $p > 1$, balancing $M^{-p}$ with $D^{-1}$ gives $D \asymp M^p$. These are saturation scales, not universal resource exchange rates.

### D.4 Hardy and Lorentz interpretation

The map $a \mapsto \Psi$, with $\Psi(n) = n^{-1}\sum_{i \le n} a_i$, is the discrete Hardy operator. For decreasing nonnegative sequences,
$$\sup_{n \ge 1} n^p a_n < \infty, \qquad 0 < p < 1,$$
is equivalent to $a \in \ell^{1/p,\infty}$ and implies $\Psi(n) = O(n^{-p})$. Conversely, monotonicity gives $a_n \le \Psi(n)$, so a Hardy bound controls the pointwise spectrum. At $p = 1$, weak-$\ell^1$ control alone yields at most the logarithmic envelope; strong $\ell^1$ membership is the first-order phase.

This interpretation is useful when no pure power exists: the depth law can be predicted directly from the measured Hardy profile without estimating an exponent.

### D.5 BEYOND $\ell_1$

For the diagonal weighted simplex in $\ell_q$, $1 \le q < \infty$, define

$$\Delta_{q,J}(D,M) = \inf_{|I|\le M} \sup_{\lambda\in\Lambda_J} \inf_{n:\sum n_i \le D} \left( \sum_i a_i^q \left| \lambda_i - \frac{n_i}{D} \right|^q \right)^{1/q},$$

with omitted coordinates represented by $n_i = 0$. The exact $\ell_1$ equality does not transfer verbatim, because coordinate errors no longer add linearly, but Theorem 6 gives the sharp two-sided profile

$$\max\left\{ a_{M+1}, \frac{1}{2D} \left( \sum_{i\le \min(D,J)} a_i^q \right)^{1/q} \right\} \le \Delta_{q,J}(D,M) \le a_{M+1} + \frac{1}{D} \left( \sum_{i\le \min(D,M,J)} a_i^q \right)^{1/q}.$$

For a routed dictionary $U$, the same $\ell_q$ law transfers when there are resource-uniform constants

$$0 < m_U \le L_U < \infty$$

such that, on every active coefficient subspace,

$$m_U \|v\|_q \le \|Uv\|_Y \le L_U \|v\|_q.$$

A Riesz basis gives this condition for $q = 2$ when its Riesz bounds remain uniform. Unconditionality alone does not imply the required $\ell_q$ norm equivalence. For general correlated dictionaries, singular values alone are insufficient: cancellation and coherent accumulation can change the synthesis profile.

## E PROOFS FOR THE NEURAL LIFTING

### E.1 EXACT SPECTRAL-WRITE REACHABLE GEOMETRY

Let $\Lambda_J = \{\lambda_i \ge 0 : \sum_{i=1}^J \lambda_i \le 1\}$ and $\mathcal{N}_{D,I} = \{n \in \mathbb{N}_0^I : \sum_{i\in I} n_i \le D\}$.

**Proposition 20** (Complete endpoint characterization)**.** *For the spectral-write architecture* (21)*,*

$$\mathcal{K}^{\mathrm{wr}}_{J,a} = \left\{ x \mapsto \sum_{i=1}^J \lambda_i a_i \phi_i(x) e_i : \lambda \in \Lambda_J \right\},$$

$$\mathcal{R}^{\mathrm{wr}}_{D,I} = \left\{ x \mapsto \sum_{i\in I} \frac{n_i}{D} a_i \phi_i(x) e_i : n \in \mathcal{N}_{D,I} \right\},$$

*and the metric identity* (22) *holds.*

*Proof.* The operator $A_i$ writes only to $y_i$ and reads only the frozen stream $s_i$, so $A_i A_j = 0$ for all $i, j$. Iterating the residual recursion therefore gives the count formula exactly. Conversely, every admissible count vector is realized by a route word containing $n_i$ copies of each mode and zero operations for the remaining steps.

For the relaxed ODE $\dot h = \sum_i u_i(t) A_i h$, the source stream remains fixed and $\dot y_i = a_i u_i(t) \phi_i(x)$. Integration gives the occupation $\lambda_i = \int_0^1 u_i(t)\, dt$, which belongs to $\Lambda_J$. Every $\lambda \in \Lambda_J$ is realized by allocating disjoint time intervals of lengths $\lambda_i$ to the modes. Finally, Tonelli's theorem and $\|\phi_i\|_{L^q} = 1$ give

$$\mathbb{E} \sum_i |a_i(\lambda_i - \mu_i)\phi_i(X)|^q = \sum_i a_i^q |\lambda_i - \mu_i|^q.$$

□

**Lemma 21** (Capped weighted remainder)**.** *Let* $w_1 \ge \cdots \ge w_J \ge 0$*. If* $0 \le r_i \le 1/D$ *and* $\sum_i r_i \le 1$*, then*

$$\left( \sum_i w_i^q r_i^q \right)^{1/q} \le \frac{1}{D} \left( \sum_{i=1}^{\min(D,J)} w_i^q \right)^{1/q}. \tag{65}$$

*Proof.* Set $t_i = Dr_i$. Then $0 \le t_i \le 1$, $\sum_i t_i \le D$, and $t_i^q \le t_i$. Hence

$$\sum_i w_i^q t_i^q \le \sum_i w_i^q t_i.$$

The right side is a linear functional over the capped simplex and is maximized by assigning unit mass to the largest $\min(D, J)$ weights. Divide by $D^q$. □

*Proof of Theorem 6.* Fix a library $I$ with $|I| \le M$. If $M < J$, some $j \le M + 1$ is absent. The ideal teacher $a_j \phi_j e_j$ has distance at least $a_j \ge a_{M+1}$ from every predictor using $I$, proving the library lower bound.

Let $m = \min(D, J)$ and choose the balanced teacher with $\lambda_i = 1/(2D)$ for $i \le m$. It is feasible because $m/(2D) \le 1/2$. Every pure coefficient is $n_i/D$, so $|1/(2D) - n_i/D| \ge 1/(2D)$. The metric identity gives the depth lower bound $A_{q,m}/(2D)$.

For the upper bound choose the first $\min(M, J)$ modes. Given $\lambda$, set $n_i = \lfloor D\lambda_i \rfloor$ on the library. The count sum is at most $D$. The omitted tail has norm

$$\left( \sum_{i>M} a_i^q \lambda_i^q \right)^{1/q} \le a_{M+1} \|\lambda_{I^c}\|_q \le a_{M+1} \|\lambda\|_1 \le a_{M+1}.$$

The retained remainders $r_i = \lambda_i - n_i/D$ satisfy Theorem 21 with at most $\min(M, J)$ coordinates, giving $A_{q,\min(D,M,J)}/D$. The triangle inequality completes the proof. □

*Proof of the phase law* (24). For $a_i \asymp i^{-p}$, integral comparison yields

$$A_{q,n} \asymp \begin{cases} n^{1/q-p}, & pq < 1, \\ (\log(1+n))^{1/q}, & pq = 1, \\ 1, & pq > 1. \end{cases}$$

If $D \le M$, insert this into the upper and lower bounds. If $M < D$, the rounding term uses $A_{q,M}/D$. In the subcritical case,

$$\frac{A_{q,M}}{D} \asymp M^{-p} \frac{M^{1/q}}{D} \le M^{-p},$$

and the corresponding critical and supercritical terms are also bounded by the maximum of the library floor and $\Psi_{p,q}(D)$. The two lower witnesses give the reverse bound. □

### E.2 State-dependent residual architecture

For (25), set $z_i = s_i + y_i$. A pure selection of mode $i$ maps $z_i \mapsto (1 + a_i/D) z_i$ and leaves all other coordinates fixed, so after $n_i$ selections

$$y_{D,i}(x) = \phi_i(x) \left[ \left(1 + \frac{a_i}{D}\right)^{n_i} - 1 \right]. \tag{66}$$

The relaxed ODE is $\dot{z}_i = a_i u_i(t) z_i$, giving

$$y_{\lambda,i}(x) = \phi_i(x)(e^{a_i \lambda_i} - 1). \tag{67}$$

We need two elementary estimates. For $0 \le z \le 1$,

$$0 \le z - \log(1+z) \le \frac{z^2}{2}. \tag{68}$$

Also, for $0 \le a \le 1$ and every integer $n$,

$$\left| e^{a/(2D)} - \left(1 + \frac{a}{D}\right)^n \right| \ge (2 - \sqrt{e}) \frac{a}{D}. \tag{69}$$

Indeed, the closest integer candidates are $n = 0$ and $n = 1$; after multiplying by $D/a$ and using monotonicity, their smaller gap is bounded below at $a/D = 1$ by $2 - \sqrt{e}$. A detailed derivative check is in the verification script.

*Proof of Theorem 7.* For the omitted-mode teacher with $\lambda_j = 1$, every predictor using a library missing $j$ has zero in that coordinate, while the target magnitude is $e^{a_j} - 1 \geq a_j$, giving the library lower bound. For the balanced teacher $\lambda_i = 1/(2D)$ on the leading $m = \min(D, J)$ modes, (69) gives coordinate gap at least $c_0 a_i/D$, hence $c_0 A_{q,m}/D$.

For the upper bound set $N = \min(D, M, J)$, choose the first $\min(M, J)$ modes, and take $n_i = \lfloor D\lambda_i \rfloor$. Tail truncation costs at most $e^{a_1} a_{M+1}$ because $e^{a_i\lambda_i} - 1 \leq e^{a_1} a_i \lambda_i$. The difference between $e^{a_i\lambda_i}$ and $e^{a_i n_i/D}$ is at most $e^{a_1} a_i(\lambda_i - n_i/D)$, so Theorem 21 gives $e^{a_1} A_{q,N}/D$.

Finally compare exact flow at $n_i/D$ with the discrete multiplier. By (68),

$$\begin{aligned} 0 &\leq e^{a_i n_i/D} - \left(1 + \frac{a_i}{D}\right)^{n_i} \\ &\leq e^{a_1} \frac{n_i a_i^2}{2D^2}. \end{aligned}$$

The $\ell_q$ norm is at most the $\ell_1$ norm, and $\sum_i n_i a_i^2 \leq a_1^2 D$. Therefore the Euler-flow defect is at most $e^{a_1} a_1^2/(2D) \leq e^{a_1}(a_1/2) A_{q,N}/D$. Summing the three terms gives (26). □

### E.3 Singular spectrum, nonlinear branches, and decoders

Let $W_a = [\mathrm{diag}(a), \mathrm{diag}(a)] : \mathbb{R}^{2J} \to \mathbb{R}^J$. Its rows are orthogonal and row $i$ has squared norm $2a_i^2$, so

$$W_a W_a^\top = 2\,\mathrm{diag}(a_i^2), \qquad \sigma_i(W_a) = \sqrt{2} a_i.$$

The routed mode $B_i$ is the $i$th row projection written back to coordinate $i$.

If an activation satisfies $\sigma(t) - \sigma(-t) = t$, then the two-unit branch

$$F_i(s, y) = a_i e_i \big[\sigma(s_i + y_i) - \sigma(-(s_i + y_i))\big]$$

is exactly $B_i$. ReLU satisfies the identity by cases. Exact GELU is $t\Phi(t)$ and uses $\Phi(t) + \Phi(-t) = 1$; SiLU uses the analogous logistic identity. Hence no approximation error is introduced by these activations.

Let $U : \ell_q^J \to Y$ satisfy

$$m_U \|v\|_q \leq \|Uv\|_Y \leq L_U \|v\|_q$$

uniformly on the active subspace. Applying this inequality to every teacher–predictor difference before the infimum and supremum multiplies lower bounds by $m_U$ and upper bounds by $L_U$. A decoder that annihilates high modes instead changes the task-visible spectrum and can remove the lower witness.

Figure 6 summarizes the architecture and the task-dependent phase boundary.

### E.4 Loss transfer

Let $\Delta_d(D, M)$ be a minimax radius in a prediction metric $d$, and suppose

$$c\, d(f, f^*)^r \leq \mathcal{L}(f; f^*) - \mathcal{L}(f^*; f^*) \leq C\, d(f, f^*)^r$$

uniformly. Applying the inequalities before the predictor infimum, teacher supremum, and library infimum gives

$$c\Delta_d(D, M)^r \leq \mathfrak{E}_{\mathcal{L}}(D, M) \leq C\Delta_d(D, M)^r.$$

This proves (28) by setting $q = r = 2$ in (24).

### E.5 Finite width and fixed teachers

At fixed $J$, $A_{q,\min(D,J)} = A_{q,J}$ for $D \geq J$, so all synthesis rates eventually become $D^{-1}$. The subcritical phase is therefore a pre-saturation law along a growing-width sequence or in an infinite-dimensional operator model. The minimax depth witness depends on $D$, and the omitted-mode witness may depend on the selected library. A fixed teacher with rational occupations can align exactly with subsequences of denominator grids. The paper does not claim a universal fixed-teacher two-sided power law; average-case and smoothed fixed-source analogues are open.

Figure 7 verifies the state-dependent recursion on the registered depth grid.

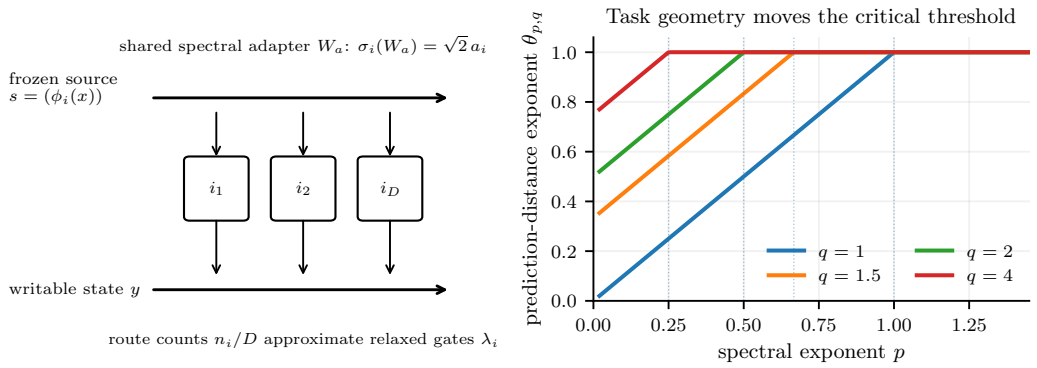



Figure 6: **Architecture and task geometry behind the neural lifting.** Left: a shared spectral adapter is reused by routed micro-layers, so route counts implement empirical occupations of its singular modes. Right: the prediction-distance exponent as a function of spectral decay for several task norms; the phase boundary moves from $p = 1$ to $p = 1/q$.

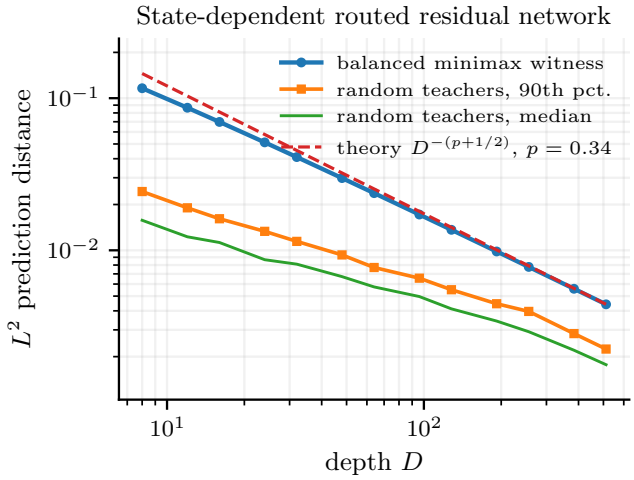



Figure 7: **Controlled simulation of the actual state-dependent routed recursion.** The balanced minimax teacher follows the theorem-predicted $L^2$ phase for $a_i = i^{-0.34}$; random relaxed teachers are typically easier. This plot validates the nonlinear Euler-flow lifting rather than only the linear weighted simplex.

## F Proofs for semantics-conditioned execution

This appendix proves the exact Green-operator identities and the fixed-horizon residual bounds used in Section 5. The distinction between a *primitive error language* and the resulting endpoint error is essential: every formula below is indexed by the admissible disturbance set and the output norm.

### F.1 A support-function identity

Let $U_j$ and $Y$ be finite-dimensional normed spaces, let $B_j : U_j \to Y$ be linear, and let $K_j \subset U_j$ be nonempty compact sets. Their support functions are

$$h_{K_j}(\eta) = \sup_{z \in K_j} \mathrm{Re}\langle \eta, z \rangle.$$

**Lemma 22** (Independent uncertainty cells)**.** *For arbitrary compact $K_j$,*

$$\sup_{z_j \in K_j} \left\| \sum_{j=1}^{m} B_j z_j \right\|_Y = \sup_{\|\psi\|_{Y^*}=1} \sum_{j=1}^{m} h_{K_j}(B_j^* \psi). \tag{70}$$

*If $K_j = \rho_j B_{U_j}$, the right side is $\sup_{\|\psi\|_*=1} \sum_j \rho_j \|B_j^*\psi\|_*$. More generally, for an arbitrary compact joint uncertainty set $K \subset \bigoplus_j U_j$ and $Bz = \sum_j B_j z_j$,*

$$\sup_{z\in K} \|Bz\|_Y = \sup_{\|\psi\|_{Y^*}=1} h_K(B^*\psi). \tag{71}$$

*Proof.* Norm duality and commutation of suprema give

$$\begin{aligned}\sup_{z_j\in K_j} \left\|\sum_j B_j z_j\right\|_Y &= \sup_{z_j\in K_j} \sup_{\|\psi\|_*=1} \operatorname{Re}\left\langle \psi, \sum_j B_j z_j \right\rangle \\ &= \sup_{\|\psi\|_*=1} \sum_j \sup_{z_j\in K_j} \operatorname{Re}\langle B_j^*\psi, z_j\rangle,\end{aligned}$$

which is (70). The joint-set identity is the same argument without factorizing the inner supremum. □

### F.2 Exact write-back and error-feedback gains

Consider

$$e_{k+1} = A_k e_k + u_k, \qquad e_0 = 0, \qquad k = 0, \dots, D-1, \tag{72}$$

with readout $Ce_D$. Define

$$P_j = A_{D-1}\cdots A_j \quad (0 \le j \le D-1), \qquad P_D = I.$$

Then

$$Ce_D = \sum_{j=1}^{D} CP_j u_{j-1}. \tag{73}$$

*Proof of Theorem 8: write-back.* Under $\|u_k\| \le \rho$, apply Theorem 22 to $B_j = CP_j$ and divide by $\rho$. This gives (30) exactly. The familiar estimate

$$\max_j \|CP_j\| \le G_D^{\mathrm{WB}}(C) \le \sum_j \|CP_j\|$$

is generally strict on the right because all blocks must be tested against one common terminal covector. □

For first-order error feedback, let

$$u_k = r_k - r_{k+1}, \qquad r_0 = 0, \qquad \|r_j\| \le \rho. \tag{74}$$

Reindexing (73) yields the exact Abel identity

$$\begin{aligned} e_D &= \sum_{k=0}^{D-1} P_{k+1}(r_k - r_{k+1}) \\ &= -r_D + \sum_{j=1}^{D-1} (P_{j+1} - P_j) r_j. \end{aligned} \tag{75}$$

*Proof of Theorem 8: error feedback.* Apply Theorem 22 to the coefficient blocks $C(P_{j+1} - P_j)$, $1 \le j \le D-1$, and $-C$ for $r_D$. This gives (31). For full-state observation, the terminal contribution equals one for every dual unit covector, hence

$$G_D^{\mathrm{EF}} = 1 + \sup_{\|\phi\|_*=1} \sum_{j=1}^{D-1} \|(P_{j+1} - P_j)^*\phi\|_*.$$

Thus uniform robust feedback gain is equivalent to uniform adjoint semivariation. For a general readout $C$, the exact formula is the joint supremum in (31); it need not equal $\|C\|$ plus a separately maximized variation term. □

**Other risk functionals.** The linear operator is the same, but a different primitive norm gives a different scaling law. With a total $\ell_1$ disturbance budget, the gain is $\max_j \|CP_j\|$. In Hilbert spaces, an $\ell_2$ energy budget gives

$$(G^{\mathrm{WB}}_{D,2})^2 = \left\| \sum_{j=1}^{D} CP_j P_j^* C^* \right\|, \tag{76}$$

$$(G^{\mathrm{EF}}_{D,2})^2 = \left\| CC^* + \sum_{j=1}^{D-1} C(P_{j+1} - P_j)(P_{j+1} - P_j)^* C^* \right\|. \tag{77}$$

These are endpoint controllability Gramians. They explain why martingale-like stochastic rounding often exhibits a square-root rather than linear accumulation law, while biased or correlated errors may not.

### F.3 Fixed-horizon residual refinement

Assume

$$A_{D,k} = I + \frac{H_{D,k}}{D}, \qquad V_D = \frac{1}{D} \sum_{k=0}^{D-1} \|H_{D,k}\| \le V, \qquad \alpha_D = \max_k \frac{\|H_{D,k}\|}{D} \le \alpha < 1. \tag{78}$$

Write $P_{D,j} = A_{D,D-1} \cdots A_{D,j}$.

**Lemma 23** (Product, inverse, and increment bounds)**.** *Under* (78)*,*

$$\|P_{D,j}\| \le e^V, \tag{79}$$

$$\|P_{D,j}^{-1}\| \le e^{V/(1-\alpha)}, \tag{80}$$

$$P_{D,j+1} - P_{D,j} = -\frac{1}{D} P_{D,j+1} H_{D,j}. \tag{81}$$

*Proof.* Submultiplicativity and $1 + x \le e^x$ give

$$\|P_{D,j}\| \le \prod_{k=j}^{D-1} \left(1 + \frac{\|H_{D,k}\|}{D}\right) \le \exp\left(\frac{1}{D} \sum_{k=j}^{D-1} \|H_{D,k}\|\right) \le e^V.$$

Because $\|H_{D,k}/D\| \le \alpha < 1$, the Neumann series gives

$$\left\| \left(I + \frac{H_{D,k}}{D}\right)^{-1} \right\| \le \frac{1}{1 - \|H_{D,k}\|/D} \le \exp\left(\frac{\|H_{D,k}\|/D}{1-\alpha}\right).$$

Multiplication yields (80). Finally, $P_{D,j} = P_{D,j+1}(I + H_{D,j}/D)$, proving (81). □

*Proof of Theorem 9.* The write-back upper bound follows from (30) and (79). For the lower bound, for every dual unit covector $\phi$,

$$\|P_{D,j}^* \phi\|_* \ge \|P_{D,j}^{-*}\|^{-1} \|\phi\|_* \ge e^{-V/(1-\alpha)}.$$

Summing gives the stated linear lower bound.

For error feedback, (81) and (79) imply

$$\begin{aligned} \sum_{j=1}^{D-1} \|(P_{D,j+1} - P_{D,j})^* \phi\|_* &\le \frac{e^V}{D} \sum_{j=1}^{D-1} \|H_{D,j}\| \, \|\phi\|_* \\ &\le e^V V. \end{aligned}$$

The terminal carry gives the lower bound $G_D^{\mathrm{EF}} \ge 1$. □

The constants remain uniform for nonnormal and noncommuting generators. What is essential is that the integrated generator budget $V$ remain uniform in every resource being scaled. If $V = V(W)$ grows with width, the factors $e^{V(W)}$ are genuine width–precision cross terms and must be displayed.

### F.4 NONLINEAR RESIDUAL NETWORKS

Consider a nominal residual trajectory

$$x_{k+1} = x_k + \frac{1}{D} F_{D,k}(x_k)$$

and a perturbed execution $\widetilde{x}_k$. If $F_{D,k}$ is continuously differentiable on the segment joining $x_k$ and $\widetilde{x}_k$, the error satisfies the *exact* secant-Jacobian recurrence

$$e_{k+1} = \left( I + \frac{1}{D} \bar{H}_{D,k} \right) e_k + u_k, \qquad \bar{H}_{D,k} = \int_0^1 JF_{D,k}(x_k + te_k)\, dt. \tag{82}$$

Thus the linear theorem is not merely a first-order heuristic: it is exact on any invariant tube on which $D^{-1} \sum_k \|\bar{H}_{D,k}\| \le V$ and $\max_k \|\bar{H}_{D,k}\|/D \le \alpha < 1$.

**Proposition 24** (Nonlinear execution tube)**.** *Under those tube conditions, robust write-back and error-feedback endpoint errors obey*

$$\|e_D^{\mathrm{WB}}\| \le e^V D\rho, \qquad \|e_D^{\mathrm{EF}}\| \le (1 + e^V V)\rho. \tag{83}$$

*Proof.* Equation (82) is an exact linear recurrence along the pair of trajectories. Its matrices satisfy the hypotheses of Theorem 9 by the assumed tube bounds. Substituting the primitive radius $\rho$ into the write-back and error-feedback gain estimates in (34) gives (83) directly. □

**Two failure modes worth separating.** Bounded $\sup_j \|P_j\|$ does not imply bounded semivariation: for the scalar path $P_j = (-1)^j$, every propagator has norm one but $G_D^{\mathrm{EF}} = \Theta(D)$. Conversely, an exact high-precision carry is part of the semantics. If the carry itself is rounded with direct error radius $\rho_{\mathrm{carry}}$, an additional write-back term of order $\rho_{\mathrm{carry}} G_D^{\mathrm{WB}}$ generally reappears.

## G JOINT STRUCTURAL–EXECUTION FRONTIERS

This appendix proves the exact robust composition law, records the useful-depth calculations, and separates global task-norm precision from per-coordinate bit guarantees.

### G.1 ROBUST COVERING BY ENDPOINT UNCERTAINTY

**Lemma 25** (Outward-ball identity)**.** *For every normed space, $w$, and $r \ge 0$,*

$$\sup_{\|u\| \le r} \|w - u\| = \|w\| + r. \tag{84}$$

*Proof.* The triangle inequality gives the upper bound. If $w \neq 0$, choose $u = -rw/\|w\|$; if $w = 0$, choose any vector of norm $r$. □

Let $K$ be a target class, $R_I$ the nominal reachable set under library $I$, and

$$\Delta = \inf_I \sup_{x \in K} \inf_{y \in R_I} \|x - y\|.$$

Suppose the endpoint uncertainty set $\mathcal{U}$ is independent of $I, x, y$ and satisfies $r_- B \subseteq \mathcal{U} \subseteq r_+ B$.

**Proposition 26** (Inner–outer robust transfer)**.** *The executed minimax risk satisfies*

$$\Delta + r_- \le E \le \Delta + r_+. \tag{85}$$

*In particular, if $\mathcal{U} = rB$, then $E = \Delta + r$.*

*Proof.* For the upper bound, $\sup_{u \in \mathcal{U}} \|x - y - u\| \le \|x - y\| + r_+$. For the lower bound, restrict the adversary to $r_- B$ and use Theorem 25 before taking the infimum and the remaining minimax operations. □

In the identity-transport spectral accumulator, independent write-back errors satisfy

$$\left\{\sum_{k=0}^{D-1} d_k : \|d_k\|_1 \le \rho\right\} = D\rho B_1,$$

because any $u \in D\rho B_1$ is represented by $d_k = u/D$. Under the robust carry model,

$$\{-r_D : r_0 = 0,\ \|r_j\|_1 \le \rho\} = \rho B_1.$$

*Proof of Theorem 10.* The nominal minimax radius is $\Delta_a(D, M)$ by Theorem 4. Under independent write-back, the endpoint uncertainty is exactly $D\rho B_1$: the inclusion from left to right follows by the triangle inequality, and every $u \in D\rho B_1$ is represented by $d_k = u/D$. Under first-order feedback, telescoping gives endpoint uncertainty exactly $\rho B_1$, because the endpoint defect is $-r_D$ and every $r_D \in \rho B_1$ is admissible. Applying the outward-ball identity in Theorem 26 yields $\Delta_a + D\rho$ and $\Delta_a + \rho$, respectively. □

For a general cocycle, the endpoint arithmetic set need not be a ball. If $h_{\mathrm{str}}(\psi)$ is the worst structural support profile after optimal synthesis and $h_{\mathrm{arith}}(\psi)$ the arithmetic support profile, then

$$\max\left\{\sup_{\|\psi\|_*=1} h_{\mathrm{str}}(\psi),\ \sup_{\|\psi\|_*=1} h_{\mathrm{arith}}(\psi)\right\} \le E \le \sup_{\|\psi\|_*=1}\{h_{\mathrm{str}}(\psi) + h_{\mathrm{arith}}(\psi)\}. \tag{86}$$

For the anisotropic case, the support-function sum in (86) gives a valid fixed-policy upper envelope for any fixed nominal synthesis policy. Exact minimax addition requires a common nominal minimax policy that remains jointly minimax after arithmetic uncertainty is composed; full centered balls and the stated transverse direct-sum constructions with an $\ell_1$ output norm satisfy this condition. Directional alignment alone is not sufficient. Equation (86) is therefore the appropriate general anisotropic statement. Equation (86), rather than an unconditional scalar sum, is the correct anisotropic statement.

### G.2 Power-law and critical useful-depth calculations

Assume the library floor is nonbinding and the structural depth profile is $cD^{-\theta}$. For write-back, minimize

$$f_{\mathrm{WB}}(D) = cD^{-\theta} + \rho D.$$

The continuous optimizer and optimized value are

$$D_*^{\mathrm{WB}} = \left(\frac{\theta c}{\rho}\right)^{1/(\theta+1)}, \qquad f_{\mathrm{WB}}(D_*^{\mathrm{WB}}) = (1+\theta)\theta^{-\theta/(\theta+1)} c^{1/(\theta+1)} \rho^{\theta/(\theta+1)}. \tag{87}$$

For stable error feedback, $f_{\mathrm{EF}}(D) = cD^{-\theta} + \rho$ is decreasing, so a tolerance factor $\eta > 0$ defines the saturation depth

$$D_{\mathrm{sat}}^{\mathrm{EF}}(\eta) = \left(\frac{c}{\eta\rho}\right)^{1/\theta}. \tag{88}$$

These formulas give the first and third rows of Theorem 11 with $\theta = p$ and $\theta = 1$.

### G.3 Critical logarithms and Lambert's $W$

For the pure critical profile $c \log D/D$, the write-back stationary equation is

$$\rho D^2 = c(\log D - 1).$$

Its large solution is

$$\boxed{(D_*^{\mathrm{WB}})^2 = -\frac{c}{2\rho} W_{-1}\left(-\frac{2e^2\rho}{c}\right).} \tag{89}$$

Indeed, setting $t = D^2$ reduces the equation to $[-2\rho t/c]e^{-2\rho t/c} = -2e^2\rho/c$. As $\rho \downarrow 0$,

$$D_*^{\mathrm{WB}} = \sqrt{\frac{c}{2\rho}\left[\log\frac{c}{2e^2\rho} + \log\log\frac{c}{2e^2\rho} + o(1)\right]}. \tag{90}$$

For error feedback, the saturation equation $c \log D/D = \eta\rho$ gives

$$\boxed{D_{\mathrm{sat}}^{\mathrm{EF}}(\eta) = -\frac{c}{\eta\rho} W_{-1}\Big(-\frac{\eta\rho}{c}\Big),} \tag{91}$$

which is $(c/(\eta\rho))[\log(c/(\eta\rho)) + \log\log(c/(\eta\rho)) + o(1)]$. The symbolic substitutions and asymptotic balances are independently checked in Section J.

*Proof of Theorem 11.* For $0 < p < 1$, insert $\theta = p$ into (87) and (88); with $\rho \asymp 2^{-\kappa b}$ this gives $D_*^{\mathrm{WB}} \asymp 2^{\kappa b/(p+1)}$ and $D_{\mathrm{sat}}^{\mathrm{EF}} \asymp 2^{\kappa b/p}$. For $p > 1$, the structural profile is first order, so $\theta = 1$ gives $2^{\kappa b/2}$ and $2^{\kappa b}$. At $p = 1$, (89) and (91) give the stated logarithmic corrections. A fixed nonbinding library floor leaves the preceding balances unchanged. When active, the minimum of the arithmetic and library saturation scales gives an order-level stopping scale; the exact write-back optimum minimizes $\Delta_a(D, M) + D\rho$. □

### G.4 Library saturation and a three-resource phase diagram

For the exact spectral law, the library becomes nonbinding when $a_{M+1}$ is below the synthesis term. If $a_i \asymp i^{-p}$, the saturation scale is

$$D_{\mathrm{lib}}(M) \asymp \begin{cases} M, & 0 < p < 1, \\ M \log M, & p = 1, \\ M^p, & p > 1. \end{cases} \tag{92}$$

Together with (87)–(91), this partitions the $(D, M, \rho)$ space into library-limited, synthesis-limited, and execution-limited regions. The exact finite-resource formula is preferable to this phase summary near the crossover surfaces.

### G.5 Bit accounting and non-global radii

The primitive arithmetic variable in the theorems is the *global radius in the task norm*, not an unspecified number of bits. If a quantizer satisfies

$$c_\rho 2^{-\kappa b} \le \rho_{\mathsf{Q}}(b) \le C_\rho 2^{-\kappa b}, \tag{93}$$

then Theorem 10 immediately yields the bit laws in (36)–(37). But the interpretation of $b$ must be declared:

- $b_{\mathrm{lib}} = \lceil \log_2(M+1) \rceil$ is an address cost, not the storage precision of the operations;
- $b_{\mathrm{state}}$ may be a total vector-codebook rate or a per-coordinate rate;
- a per-coordinate error $\delta$ in $N$ active coordinates has global $\ell_q$ radius at most $N^{1/q}\delta$;
- an exact error-feedback carry is a resource; rounding the carry generally reintroduces an undifferenced term of size $\rho_{\mathrm{carry}} G_D^{\mathrm{WB}}$.

Thus the dimension-free headline uses a global task-norm radius. Under a coordinatewise full-cube model with $N(D, M)$ active coordinates and per-coordinate radius $\delta$, the global $\ell_q$ radius is $N(D, M)^{1/q}\delta$ for $1 \le q < \infty$ and $\delta$ for $q = \infty$. Thus the feedback execution term is $N(D, M)^{1/q}\delta$, specializing to $N(D, M)\delta$ in the canonical $\ell_1$ model.

### G.6 Alternative execution models

The robust law is a certified worst-case statement. It is not interchangeable with deterministic or stochastic rounding:

1. Coordinatewise nearest write-back admits a deterministic dead-zone witness of order $\min\{1, D\rho\}$, but exact grid-aligned trajectories can have zero error.
2. Independent centered defects with variance $\asymp \rho^2$ yield an RMS term $\asymp \sqrt{D}\rho$ under identity transport. Conditional bias or correlation can restore linear growth or produce cancellation.

3. Saturation and wraparound require separate range bits and no-overload conditions; they are not represented by an unbounded additive ball.

These are different semantics and risk functionals, hence different scaling laws. The paper's claim is precisely that execution semantics is part of the scaling problem.

# H STRUCTURAL IDENTIFIABILITY AND CERTIFIED PREDICTION

The scaling surface is predictive only if its structural inputs can be measured and if the forward map is stable. This appendix proves both claims for the canonical model and states the unavoidable extrapolation boundary.

## H.1 EXACT FINITE-PREFIX STABILITY

The exact frontier can be written

$$\Delta_a(D, M) = \Big(1 - \frac{m}{2D}\Big)\, a_{M+1} + \frac{1}{2D} \sum_{i=1}^{m} a_i, \qquad m = \min\{M, 2D\}. \tag{94}$$

All coefficients are nonnegative and sum to one. Consequently:

**Proposition 27** (Condition-one spectral prediction)**.** *Let* $N_*(D, M) = \min\{2D, M + 1\}$. *If* $\max_{i \le N_*(D,M)} |\widehat{a}_i - a_i| \le \varepsilon_a$*, then*

$$|\Delta_{\widehat{a}}(D, M) - \Delta_a(D, M)| \le \varepsilon_a. \tag{95}$$

*The constant one is optimal. Hence an entire finite resource region is predictable from the largest required prefix over that region; no asymptotic exponent need be estimated.*

*Proof.* Subtract (94) for $a$ and $\widehat{a}$ and use that the absolute weights sum to one. Equality is attained by a common additive perturbation of every coefficient. □

This improves a generic tail-index pipeline: the exact finite-resource functional is better conditioned than an intermediate estimate of $p$, especially near the summability threshold.

## H.2 PLUG-IN ERROR FOR THE COMPLETE SURFACE

Let

$$R(D, M, b) = \Delta_a(D, M) + \rho(b) G_D,$$

and define the plug-in predictor by replacing each structural profile with an estimate.

*Proof of Theorem 12.* By Theorem 27, the spectral error is at most $\varepsilon_a$. The product perturbation identity gives

$$\begin{aligned} |\widehat{\rho}\widehat{G} - \rho G| &\le G|\widehat{\rho} - \rho| + \rho|\widehat{G} - G| + |\widehat{\rho} - \rho|\,|\widehat{G} - G| \\ &\le G_{\max}\varepsilon_\rho + \rho_{\max}\varepsilon_G + \varepsilon_\rho\varepsilon_G. \end{aligned}$$

Add the terms and take the supremum over the prediction region. Composition with a Lipschitz loss calibration gives (41). For two-sided power calibration, monotonicity maps the certified radius interval to the stated loss interval. □

A multiplicative variant is often more informative near zero. If $e^{-\eta_a}\Delta \le \widehat{\Delta} \le e^{\eta_a}\Delta$ and similarly for $\rho$ and $G$, then

$$e^{-\eta_R} R \le \widehat{R} \le e^{\eta_R} R, \qquad \eta_R = \max\{\eta_a, \eta_\rho + \eta_G\}. \tag{96}$$

If $L - L_* = cR^r$, the log-relative loss error is at most $r\eta_R$ when $c$ and $r$ are known.

### H.3 ANISOTROPIC SUPPORT-PROFILE PREDICTION

Scalar radii discard direction. Let $h_{\mathrm{str}}(\psi; D, M)$ and $h_{\mathrm{arith}}(\psi; D, b)$ be structural and arithmetic support profiles on a task-dual unit sphere $\Psi$. Define

$$R(D, M, b) = \sup_{\psi \in \Psi} \{h_{\mathrm{str}}(\psi; D, M) + h_{\mathrm{arith}}(\psi; D, b)\}.$$

When $h_{\mathrm{str}}$ is obtained from a fixed nominal synthesis policy, $R(D, M, b)$ is a declared-policy support-profile predictor. It equals the executed minimax risk only under the same common-policy composition condition required for exact minimax addition in Section G. If the two profiles are uniformly estimated to errors $\epsilon_{\mathrm{str}}$ and $\epsilon_{\mathrm{arith}}$, then

$$|\widehat{R} - R| \le \epsilon_{\mathrm{str}} + \epsilon_{\mathrm{arith}}.$$

This follows because the supremum functional is one-Lipschitz in the uniform norm. The directional formulation is the appropriate one when worst structural and arithmetic directions do not align.

### H.4 MEASURING GREEN PROFILES

For coefficient blocks $B_j : X_j \to Y$, define

$$\mathfrak{G}(B_1, \dots, B_m) = \sup_{\|\phi\|_{Y^*} \le 1} \sum_{j=1}^{m} \|B_j^* \phi\|_{X_j^*}.$$

For a fixed terminal covector $\phi$, one reverse sweep through a stored residual graph returns all $P_j^* C^* \phi$. The same sweep therefore computes both the directional write-back mass and the directional feedback semivariation in $O(D)$ accumulation time.

**Proposition 28** (Green-profile stability)**.** *If $\|\widehat{B}_j - B_j\| \le \delta_j$, then*

$$|\mathfrak{G}(\widehat{B}_1, \dots, \widehat{B}_m) - \mathfrak{G}(B_1, \dots, B_m)| \le \sum_{j=1}^{m} \delta_j. \tag{97}$$

*For error feedback, if*

$$\|\widehat{P}_j - P_j\| \le \delta_j$$

*and the readout $C$ is exact, the feedback variation-profile error is at most*

$$\|C\| \sum_{j=1}^{D-1} (\delta_j + \delta_{j+1}) \le 2\|C\| \sum_j \delta_j,$$

*apart from the exactly known boundary term. The factor $\|C\|$ may be omitted for full-state observation or when $\delta_j$ already measures the error in $CP_j$.*

In Euclidean task dimension $r$, let $F(\phi) = \sum_j \|B_j^\top \phi\|_2$ and $U = \sum_j \|B_j\|_{2\to 2}$. Then $F$ is $U$-Lipschitz on the sphere. Let

$$\phi_1, \dots, \phi_N$$

be independent and uniformly distributed on $S^{r-1}$, and let

$$\widehat{G}_N = \max_{1 \le \ell \le N} F(\phi_\ell).$$

If $\mu_r(\tau)$ denotes the normalized measure of a radius-$\tau$ spherical cap around a maximizer of $F$, then

$$\Pr\{G - \widehat{G}_N > U\tau\} \le (1 - \mu_r(\tau))^N \le \exp[-N\mu_r(\tau)], \tag{98}$$

where $\mu_r(\tau)$ is the normalized measure of the cap. Since $\mu_r(\tau) = \Theta_r(\tau^{r-1})$ for small $\tau$, robust near-exact estimation by isotropic probes is exponentially costly in the *task-quotient* dimension. Scalar or low-dimensional readouts are therefore not a cosmetic simplification; they make the robust Green quantity identifiable.

A computable Gramian bracket is

$$g_2 \le \mathfrak{G}(B_1, \dots, B_m) \le \sqrt{m}\, g_2, \qquad g_2^2 = \lambda_{\max}\left( \sum_j B_j B_j^* \right). \tag{99}$$

It is useful as a certificate, but the factor $\sqrt{m}$ can hide the distinction between bounded and growing semivariation.

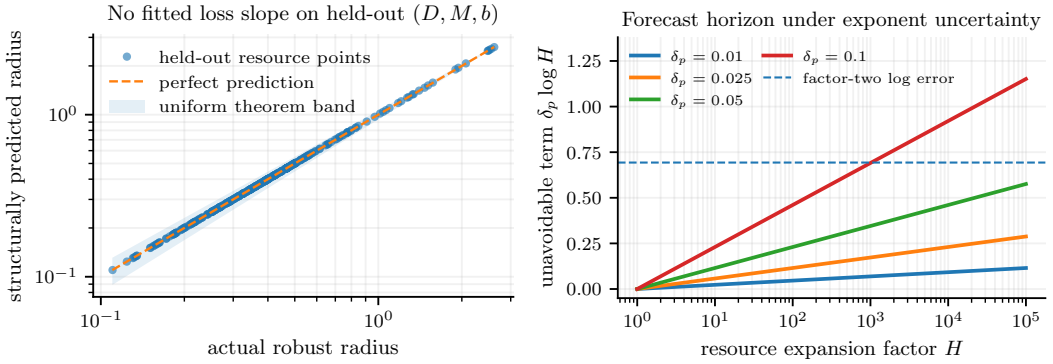



Figure 8: **Finite structural coverage and continuation are different sources of prediction.** Left: a controlled closure test plugs independently perturbed structural profiles into (39) and covers every withheld resource point with (40). Right: once prediction extends beyond the measured prefix, exponent uncertainty grows as $\delta_p \log H$. The reader should distinguish certified finite-horizon prediction from assumption-dependent tail continuation.

### H.5 No-free-lunch and finite forecast horizons

*Proof of Theorem 13: spectra.* Fix any finite positive nonincreasing prefix $a_1 \geq \cdots \geq a_K > 0$ and two distinct indices $p, q > 0$. Continue the same prefix by

$$a_n^{(p)} = a_K \left(\frac{n}{K}\right)^{-p}, \qquad a_n^{(q)} = a_K \left(\frac{n}{K}\right)^{-q}, \qquad n > K.$$

Both continuations are positive and nonincreasing, agree exactly through $K$, and are regularly varying with indices $-p$ and $-q$. Every estimator based only on the prefix receives identical data under the two models, so its worst-case index error is at least $|p - q|/2$. □

The critical boundary is correspondingly hard to resolve. If the slowly varying factor is only known to lie in $[C^{-1}, C]$, distinguishing $p = 1$ from $p = 1 + \epsilon$ uniformly can require $K > C^{1/\epsilon}$, exponential in $1/\epsilon$.

*Proof of Theorem 13: Green profiles.* Fix an observed horizon $D_0$ and consider scalar recurrences. Let both families have $A_k = 1$ for $k < D_0$. Continue the first with $A_k = 1$ for all $k \geq D_0$ and the second with $A_k = 2$ for all $k \geq D_0$. Every observation through $D_0$ is identical. At a later horizon $D$, the first family has write-back gain $D$ and feedback gain 1, whereas the second has write-back gain at least $2^{D-D_0}$ and feedback semivariation of the same exponential order. Hence a finite observed prefix cannot bound or identify unrestricted later Green growth. □

A quantitative continuation class restores a finite forecast. Suppose for $1 \leq \lambda \leq H$,

$$\left|\log \frac{f(\lambda K)}{f(K)} + p \log \lambda\right| \leq R_f(K, \lambda), \tag{100}$$

and $|\log(\widehat{f}(K)/f(K))| \leq \eta_0$, $|\widehat{p} - p| \leq \delta_p$. Then the extrapolator $\widehat{f}(\lambda K) = \widehat{f}(K)\lambda^{-\widehat{p}}$ obeys

$$\left|\log \frac{\widehat{f}(\lambda K)}{f(\lambda K)}\right| \leq \eta_0 + \delta_p \log \lambda + R_f(K, \lambda). \tag{101}$$

The $\delta_p \log H$ term is unavoidable: exponent uncertainty is amplified by logarithmic resource distance.

### H.6 Why marginals and observed loss are insufficient

Even complete one-dimensional slices do not identify an interior surface. If $F_0(D, M)$ is any admissible monotone surface and $h(D, M)$ is a nonnegative monotone interaction that vanishes on

the observed slices, then $F_0$ and $F_0 + h$ share every marginal observation but differ off-axis. This is the statistical counterpart of Theorem 2.

Likewise, if $E_{\mathrm{obs}} = E_{\mathrm{structural}} + \Delta_{\mathrm{train}}$ with both terms nonnegative, the decomposition is not identifiable from $E_{\mathrm{obs}}$ alone: any nonnegative subcurve below $E_{\mathrm{obs}}$ can be designated structural and the remainder training-related. An optimization certificate or external upper bound on $\Delta_{\mathrm{train}}$ is necessary. If $\Delta_{\mathrm{train}} = o(E_{\mathrm{structural}})$, the structural exponent survives; if the training gap decays more slowly, the observed exponent belongs to the trainer.

# I Experimental methodology and reproducibility

The numerical program has two roles. Controlled experiments serve as theorem regressions: every guide is computed from a declared spectrum, dynamics, quantizer, or exact formula. The frozen-ResNet study then asks whether the same structural measurement pipeline predicts a real neural resource surface that was not used to construct the theorem. The source package contains fixed seeds, configurations, raw CSV files, hashes, analysis code, and vector figures with LaTeX-rendered text.

## I.1 Controlled theorem regressions

**Exact finite-resource regression.** For each of 24 seeds, we draw a decreasing three-mode spectrum, append $a_4 = 0$, and choose $D \in \{1, 2, 3, 4\}$ and $M \in \{1, 2, 3\}$. The depth grid is enumerated exactly. The target simplex is evaluated on denominator 48, which contains every half-step witness $1/(2D)$. All 24 radii agree with Theorem 4 to floating-point precision; the complete record is `data/exact_grid_regression.csv`. This regression detects indexing, truncation, and rounding-code errors independently of the proof.

**Spectral phases and critical corrections.** We evaluate

$$\Delta_a(D, \infty) = \frac{1}{2D} \sum_{i=1}^{2D} a_i$$

through $D = 10^5$ for pure powers $p \in \{0.55, 1, 1.45\}$. Tail slopes over the final 25 points are $-0.54746$, $-0.91314$, and $-0.99735$, against $-0.55$, a logarithmically corrected $-1$, and $-1$. The critical-profile test extends to $D = 10^6$ and compares

$$a_i = i^{-1}, \qquad a_i = [i \log(e + i)]^{-1}, \qquad a_i = [i \log^2(e + i)]^{-1}.$$

The scaled quantity $D\Delta_a$ exhibits logarithmic growth, log–log growth, and convergence, respectively, as shown in Figure 5. This tests the integrated slowly varying profile in (19), not a preselected logarithm.

**Joint surface and nonlinear lifting.** For $a_i = i^{-0.72}$, the exact $(D, M)$ surface is evaluated over $2 \le D \le 256$ and $1 \le M \le 512$; the boundary $M = 2D - 1$ is the theorem's exact saturation line. The state-dependent network uses (25) with $J = M = 1024$, $q = 2$, $p = 0.34$, and depths from 8 to 512. The balanced-witness slope is $-0.81210$, compared with the asymptotic $-(p + 1/2) = -0.84$; sparse random teachers are easier, as the minimax theorem permits (Figure 7).

**Green profiles and controlled prediction.** The two-dimensional nonnormal cocycle in Figure 3 uses $A_{D,k} = I + H((k + 1/2)/D)/D$ with time-varying shear and rotation. Euclidean Green gains are estimated on a uniform grid of 4096 terminal directions; the reported scan maximum $\widehat{G}$ is a lower estimate of the continuous gain $G$, with $\widehat{G} \le G \le \sec(\delta)\widehat{G}$ for maximum angular half-gap $\delta$. The final-depth slopes are 0.9999878 for write-back and 0.0005705 for feedback, with $G_D^{\mathrm{EF}} < 2.2347$.

A separate closure test perturbs a declared spectrum, quantizer radius, and Green profile, withholds 352 resource triples, and predicts them only through (39). The maximum absolute error is $5.5522 \times 10^{-3}$, below the uniform bound $2.0989 \times 10^{-2}$ at every point; the closure and forecast-horizon diagnostics are shown in Figure 8.

| Component | Registered protocol |
|---|---|
| Representation | ImageNet-pretrained ResNet-18 with its final classifier removed; 512-dimensional CIFAR-10 features. Clean images are resized to $224 \times 224$. The deterministic degraded view is downsampled to $56 \times 56$, restored to $224 \times 224$, and blurred with a $5 \times 5$ Gaussian kernel of standard deviation $1.2$. |
| Fixed computation | One ridge residual operator and one clean-feature ridge classifier fitted on $40{,}000$ training examples. No model is retrained as $D$, $M$, or $b$ changes. |
| Structural calibration | $5{,}000$ held-out training examples estimate the first $256$ task-visible amplitudes and determine mode ordering, teacher registration, and quantizer range, using centering statistics $\mu_{\mathrm{c}}, \mu_{\mathrm{d}}$ learned from the $40{,}000$-example training set. The residual operator and classifier are also fitted on that training set. |
| Untouched evaluation | The complete spectrum-transfer and $7 \times 7$ depth–library grid use all $10{,}000$ CIFAR-10 test examples. The execution intervention uses the first $2{,}048$ test examples and a fixed 32-mode route. |
| Resource grids | $D, M \in \{4, 8, 16, 32, 64, 128, 256\}$; execution depths $D \in \{8, 16, 32, 64, 128, 256\}$; state precision $b \in \{4, 5, 6, 8, 10, 12\}$. |
| Hardware | NVIDIA A100-SXM4-40GB; PyTorch 2.11.0 and torchvision 0.26.0; seed 20260828. |

Table 1: **Registered neural-bridge protocol.** Training, structural calibration, untouched evaluation, and resource variation are separated. The reader should note that the experiment studies quantized execution of one fixed learned computation rather than a separately retrained model at every resource point.

### I.2 Frozen-ResNet neural bridge

#### I.2.1 One fixed residual operator and its task-visible spectrum

Let $x_{\mathrm{c}}, x_{\mathrm{d}} \in \mathbb{R}^{512}$ denote clean and degraded frozen features. Means $\mu_{\mathrm{c}}$ and $\mu_{\mathrm{d}}$ are learned from the $40{,}000$ training examples. With $z = x_{\mathrm{d}} - \mu_{\mathrm{d}}$, the residual operator is the ridge solution

$$A = \arg\min_{B} \sum_{n} \|(x_{\mathrm{c},n} - \mu_{\mathrm{c}}) - z_n - z_n B\|_2^2 + 0.05\|B\|_F^2. \tag{102}$$

A separate ridge classifier with penalty $0.01$ is fitted on clean features and held fixed thereafter. Writing $A = U \operatorname{diag}(s_i) V^\top$, the routed mode output is

$$g_i(x) = s_i \langle x - \mu_{\mathrm{d}}, u_i \rangle v_i.$$

Because the $v_i$ are orthonormal, the population $L^2$ amplitude of mode $i$ is

$$a_i = s_i \left( \mathbb{E} \langle X_{\mathrm{d}} - \mu_{\mathrm{d}}, u_i \rangle^2 \right)^{1/2}. \tag{103}$$

The calibration population orders the modes and estimates $a_i^{\mathrm{cal}}$; the test population independently estimates $a_i^{\mathrm{test}}$ without changing the order. Their Pearson and Spearman correlations are $0.999826$ and $0.998526$. Median and maximum relative amplitude shifts are $0.9397\%$ and $3.3574\%$. A pure-power diagnostic over modes 32–256 gives slopes $-0.5543$ and $-0.5510$ with $R^2 = 0.9984$ and $0.9965$ on calibration and test. The fitted slope changes across windows, so all structural predictions use the measured coefficients directly.

#### I.2.2 Registered depth–library prediction

For each $(D, M)$, two candidate teachers are declared before inspecting test error. If $M < J$, the omitted-mode teacher places unit mass on mode $M + 1$. The balanced teacher places $1/(2D)$ on the first $\min(D, J)$ modes. Let

$$I_M = \{1, \ldots, \min(M, J)\}.$$

For either relaxed coefficient vector $\lambda$, define the denominator-$D$ route by

$$n^*(\lambda) \in \operatorname*{arg\,min}_{\substack{n \in \mathbb{N}_0^J,\ \sum_i n_i \le D \\ n_i = 0 \text{ for } i \notin I_M}} \left[ \sum_i (a_i^{\mathrm{cal}})^2 \left| \lambda_i - \frac{n_i}{D} \right|^2 \right]^{1/2}. \tag{104}$$

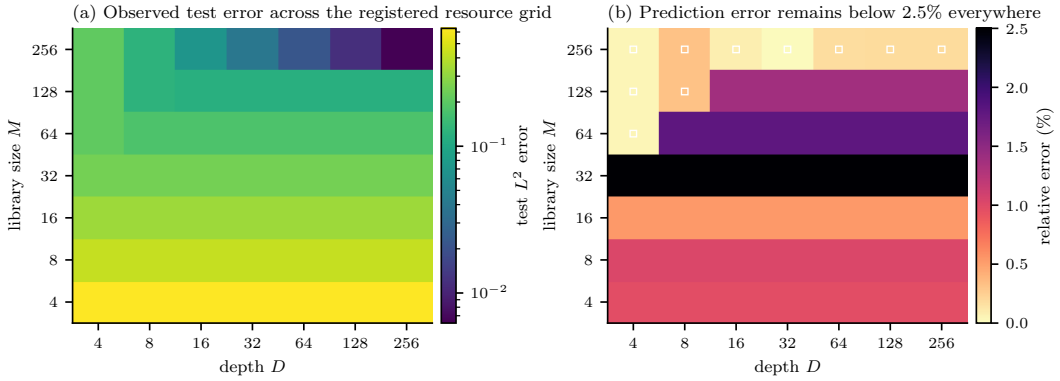



Figure 9: **The structural prediction remains accurate over the complete depth–library plane.** Left: observed test error for the registered teacher at every $(D, M)$. Right: relative prediction error from the calibration spectrum; open squares identify balanced half-grid witnesses and all other cells use omitted-mode witnesses. The reader should see both the library floor and depth saturation while noting that every cell is predicted within $2.5\%$.

When multiple count vectors attain the same minimum, choose the lexicographically smallest count vector. The same deterministic rule is used for calibration and test evaluation. The candidate with larger calibration error is registered. Its predicted test error substitutes $a_i^{\mathrm{cal}}$ in (104); its observed test error substitutes the independently measured $a_i^{\mathrm{test}}$ while keeping the teacher and route unchanged.

Across the 49 resource pairs, the median and maximum relative errors are $1.0161\%$ and $2.4899\%$, the log-scale correlation is $0.999921$, and RMSE is $3.864 \times 10^{-3}$. The balanced half-grid subset is even tighter: median $0.1237\%$, maximum $0.3214\%$. Figure 9 shows that the prediction remains accurate throughout the full two-dimensional grid rather than only along a fitted diagonal.

### I.2.3 Causal execution intervention

The execution study fixes the first 32 ordered modes and the balanced coefficient vector $\lambda_i = 1/32$. Calibration data set one symmetric coordinate range $R$: 1.2 times the $0.9999$ quantile of the absolute full-target residual. A signed $b$-bit nearest-grid quantizer uses

$$q_{\max} = 2^{b-1} - 1, \qquad \delta_b = R/q_{\max}, \qquad Q_b(x) = \delta_b \,\mathrm{round}(\mathrm{clip}(x, -R, R)/\delta_b). \tag{105}$$

Its declared global Euclidean cube radius is $\rho_b = \sqrt{512}\,\delta_b/2$. A deterministic round-robin schedule interleaves the registered counts. The ideal route adds each increment in floating point. Write-back quantizes the complete residual state after every micro-layer. Error feedback quantizes the increment plus a persistent carry and accumulates the quantized increment in high precision.

At $D = 8$, the fixed target $\lambda_i = 1/32$ has a unique nearest grid point; at $D = 16$, it lies at a half-grid tie, so the deterministic tie-breaking rule fixes the registered count vector. The reported zero execution errors at $D = 8$ and $D = 16$ are observed outcomes of these fixed routes; slope diagnostics therefore use $D \geq 32$. At $b = 10$ and $b = 12$, write-back slopes are $0.9013$ and $0.8988$, with $R^2 = 0.99961$ and $0.99928$. Feedback errors are constant across all tested $D \geq 32$ to numerical precision. Averaging across those depths gives a bit law $2^{-1.0107b}$ with $R^2 = 0.999975$. Maximum overload is $2.813 \times 10^{-4}$ of scalar quantizer calls. The observed execution errors are compared against the nominal $D\rho_b^{\mathrm{round}}$ and $\rho_b^{\mathrm{round}}$ reference envelopes (Figure 10).

### I.2.4 Downstream task relevance and task quotienting

The clean-feature classifier reaches $85.89\%$ test accuracy. Degradation lowers accuracy to $82.15\%$; applying the fixed full residual correction raises it to $84.13\%$, recovering $52.94\%$ of the clean–degraded gap. Thus the learned operator is relevant to the downstream task. Across the execution grid, however, the maximum change from the ideal-route classification accuracy is only $0.0488$ percentage

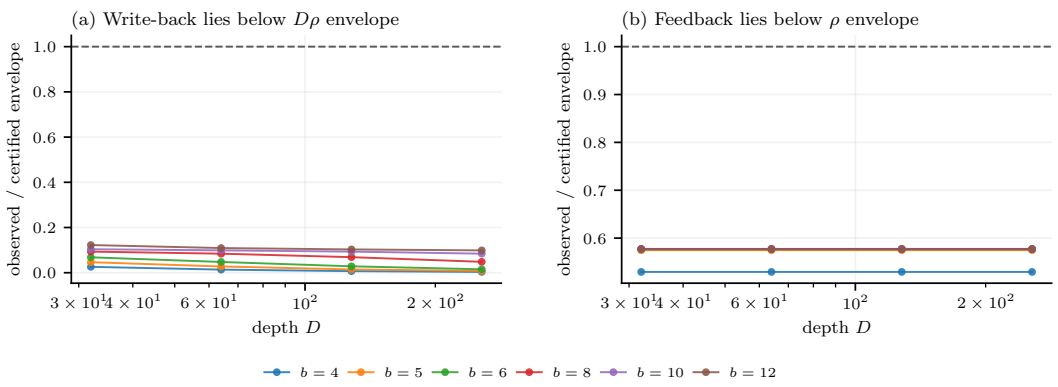



Figure 10: **Practical nearest-grid executions relative to the robust reference envelopes.** Each curve reports observed feature error divided by the write-back $D\rho_b$ or feedback $\rho_b$ reference envelope. Here the "certified envelope" label denotes the nominal robust reference envelope under the declared uncertainty model; the smaller observed ratios reflect alignment and cancellation not captured by the adversarial worst case.

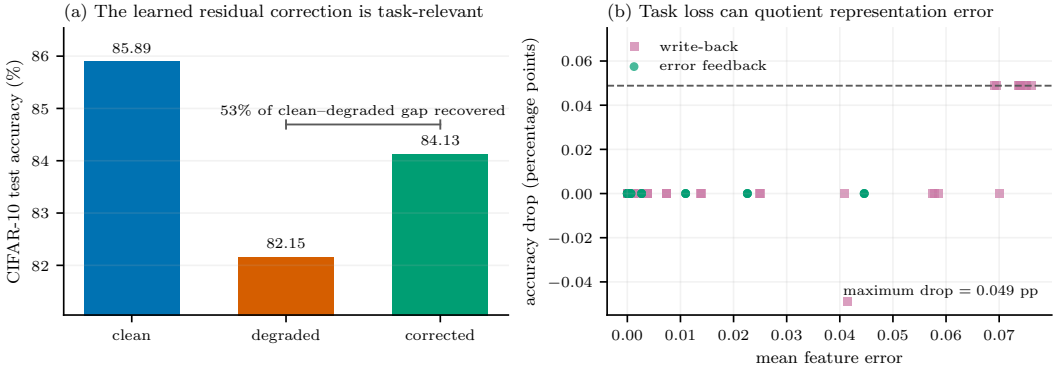



Figure 11: **Task geometry converts representation scaling into loss scaling.** Left: the learned residual computation recovers 53% of the classification gap created by deterministic image degradation. Right: the same execution perturbations can produce visible feature error while changing $0/1$ accuracy by at most $0.049$ percentage points. The reader should not equate a representation-space exponent with a classification-accuracy exponent without a loss-calibration condition.

points. Figure 11 makes the distinction explicit: execution semantics strongly changes representation error while the coarse $0/1$ loss quotients most of that variation.

### I.3 Artifacts and falsification criteria

The neural-run inputs and outputs are stored under `data/neural_bridge/`: the measured spectrum, structural grid, execution grid, fixed configuration, result manifest, summary, and audited metrics. The analysis script is `scripts/analyze_neural_bridge_results.py`; the resumable notebook is `notebooks/QNC_neural_bridge.ipynb`.

The registered numerical claims fail if any of the following occurs:

1. an exhaustive finite-grid radius differs from Theorem 4 beyond the declared target-grid resolution;
2. $D\Delta_a$ fails to separate the critical profiles of (19);
3. the balanced state-dependent neural witness has a persistent tail outside the two-sided theorem range;

4. the declared cocycle fails to exhibit linear write-back mass or bounded feedback semivariation;
5. a controlled withheld prediction lies outside (40);
6. the independently measured test spectrum fails to reproduce the calibration profile, or a registered $(D, M)$ error exceeds the reported $2.5\%$ prediction tolerance;
7. the semantics-only execution fit loses the growing-versus-bounded distinction on the registered high-bit range.

The supplied scripts exit nonzero on exact-grid, schema, finiteness, or interval-coverage failures and reproduce all quoted statistics from the archived CSV files.

## J PROOF, SYMBOLIC, AND IMPLEMENTATION VERIFICATION

Verification has three independent layers with different logical status: written proofs, symbolic checks, and executable theorem regressions.

### J.1 PROOF ARCHITECTURE AND QUANTIFIER CHECKS

Every main theorem has a complete written proof in the appendices. The proof system keeps the following distinctions explicit:

1. quantifier order: library before target, synthesis before robust arithmetic;
2. the distinction between one-sided reachable width and symmetric Hausdorff distance;
3. finite width $J$, library cardinality $M$, global arithmetic radius, and address bits as separate resources;
4. exact versus asymptotic claims and the uniformity of constants;
5. robust, deterministic, and stochastic execution as separate semantics;
6. task-visible lower witnesses, including whether a decoder annihilates them.

The weighted-simplex upper bound in Section C uses the affine vertex-cost extension

$$\widetilde{L}(p) = \sum_i b_i\{f_i + p_i(1 - 2f_i)\} + cu.$$

Because the objective is affine, the integral and convex-hull minima coincide, and the explicit fractional point gives an upper bound. The half-grid teacher supplies the lower-bound witness.

### J.2 WOLFRAM CHECKS

The supplied file `scripts/wolfram_checks.wl` verifies the algebraic and asymptotic subclaims most vulnerable to transcription error:

1. the subcritical Hardy limit
$$\lim_{n\to\infty} \frac{H_{n,p}/n}{n^{-p}/(1-p)} = 1, \qquad 0 < p < 1;$$
2. the summable constant
$$\lim_{n\to\infty} n\frac{H_{2n,p}}{2n} = \frac{\zeta(p)}{2}, \qquad p > 1;$$
3. the nonlinear half-grid constant
$$\min_{0<x\le 1} \min\left\{\frac{e^{x/2}-1}{x}, \frac{1+x-e^{x/2}}{x}\right\} = 2 - \sqrt{e};$$
4. a complete $J = 3$, $D = 2$, $M = 2$ spectral minimax instance, for which numerical global optimization and the closed formula both equal $1/2$;
5. numerical residuals of the $W_{-1}$ expressions (89) and (91) over four decades of $\rho$.

The recorded stationary-equation residuals are below $2 \times 10^{-21}$ at the working precision used for the critical checks.

### J.3 Executable theorem regressions

The Python regression suite performs:

1. exhaustive count-grid evaluation for 24 small spectral instances;
2. exact evaluation of all finite-resource figures from the theorem formulas;
3. direct simulation of the state-dependent routed residual recursion;
4. dense dual-sphere computation of Green mass and semivariation in two dimensions;
5. coverage checks for every held-out structural prediction interval.

## K Applicability, resource accounting, and theoretical frontier

The structural profiles in the main text are useful because their assumptions are explicit. This appendix states the boundary of each exact law and the mathematical questions that remain once those assumptions are relaxed.

### K.1 Canonical solvable class and broader dictionaries

The exact spectral law is proved for a shared routed basis with positive ordered modes, and the neural lifting preserves that structure. General residual networks may have coherent, noncommuting, or state-dependent dictionaries whose compositions create cancellation, algebraic closure, or word-level superresolution. For example, on $[0, 1]$, the reusable maps

$$T_0(x) = x/2, \qquad T_1(x) = (x + 1)/2, \qquad x_0 = 0,$$

generate every endpoint $x_D = m/2^D$, $m = 0, \dots, 2^D - 1$, after $D$ stages, giving $2^D$ reachable endpoints with spacing $2^{-D}$. This explicit construction shows how word-generated classes can attain exponential depth resolution. In those settings the shared-library reachable width remains well defined, but its synthesis profile need not be a clipped Cesaro average. The measured spectrum and Green profile provide falsifiable diagnostics: agreement supports the canonical mechanism; systematic residual structure identifies which assumption must be enlarged.

### K.2 Width and storage are charged

The subcritical $L^q$ neural phase requires the number of task-visible modes $J$ to remain nonbinding over the depth range. At fixed $J$, the synthesis term eventually becomes $A_{q,J}/D$, and the library floor disappears once $M \geq J$. Thus $D^{-(p+1-1/q)}$ is a joint growing-mode or neural-operator regime, not a free fixed-width asymptotic. Likewise, $\lceil \log_2(M + 1) \rceil$ counts route addresses, not the storage precision of the basis vectors or amplitudes. The primitive arithmetic resource is the norm-specific radius $\rho(b, J, \| \cdot \|)$.

### K.3 Minimax and fixed-target behavior

The sharp library witness follows the declared minimax order: one universal library is chosen before the target. A fixed target with rational occupation weights can align exactly with denominator grids on subsequences, so its pointwise error may oscillate or vanish faster than the minimax radius. Average-case, smoothed, and multiscale fixed-target laws require additional assumptions on the target distribution. The minimax theorem characterizes universal approximation capacity; it does not replace those source-specific notions.

### K.4 Execution semantics and carry resources

The exact $\Delta + D\rho$ and $\Delta + \rho$ identities use a robust task-norm uncertainty ball, exact first-order carry, and no overload. Deterministic nearest rounding can be smaller on aligned trajectories; stochastic RMS laws depend on bias and correlation; coordinatewise precision introduces a dimension factor; and quantizing the carry can restore accumulation. Each choice defines a different primitive disturbance set and therefore a different semantics-conditioned Green profile. The deterministic neural bridge reports both the observed error and the corresponding robust envelope.

### K.5 Prediction horizon and evolving computations

A finite spectral prefix exactly determines every resource point whose required indices have been measured. It cannot identify an unrestricted asymptotic continuation. Likewise, Green gains observed through depth $D_0$ do not determine an arbitrary future cocycle. Certified prediction therefore reports either a structural horizon or an explicit continuation remainder. When a new model is retrained at every resource point, the structural profiles themselves become resource-dependent and an optimization gap can control the observed curve. Predicting that family requires a model of profile evolution or an independent training certificate in addition to the fixed-computation theory developed here.

### K.6 Detailed relation to prior work

Classical neural scaling laws established smooth empirical relationships among model size, data, compute, and loss (Kaplan et al., 2020; Hoffmann et al., 2022). Mechanistic accounts derive or constrain exponents through intrinsic dimension, kernel and feature spectra, frequency-ordered skills, language correlations and conditional entropy, or controlled task families (Sharma & Kaplan, 2022; Bahri et al., 2024; Bordelon et al., 2024; Michaud et al., 2023; Cagnetta et al., 2026; Barkeshli et al., 2026). The present theory adds a distinct mechanism: finite depth repeatedly decodes a shared operation alphabet, so the joint resource law depends on synthesis of the task-visible operation profile rather than on a static capacity variable alone.

Precision-aware scaling work studies how integer or floating-point precision interacts with model size, grouping, training compute, compression, or post-training error (Kumar et al., 2025; Xu et al., 2024; Frantar et al., 2025; Panferov et al., 2025; Chen et al., 2025; Sun et al., 2025; Dremov et al., 2026). Static finite-library approximation is naturally related to metric entropy, vector quantization, rate-distortion ideas, nonlinear widths, and quantized-network approximation (Kolmogorov & Tikhomirov, 1961; Pinkus, 1985; Graf & Luschgy, 2000; Gersho & Gray, 1992; Ou et al., 2024). Equal-weight synthesis is adjacent to approximate Carathéodory theory (Barman, 2015; Mirrokni et al., 2017; Tinarrage, 2026). Our exact clipped-Cesàro formula is specific to a reusable ordered spectral dictionary and exposes a finite $(D, M)$ saturation boundary that these general-purpose bounds do not by themselves identify.

Residual and continuous-depth networks motivate the fixed-horizon residual refinement model studied in this paper (He et al., 2016; Haber & Ruthotto, 2018; Chen et al., 2018). Sigma–delta quantization uses feedback and oversampling to shape quantization error and improve reconstruction (Daubechies & DeVore, 2003; Güntürk, 2003; Krahmer et al., 2012); error-feedback methods in compressed optimization feed accumulated compression error into subsequent updates (Karimireddy et al., 2019); recent quantization methods likewise propagate or correct errors across layers (Arai & Ichikawa, 2025; Zhang et al., 2026). The Green identities here isolate the exact endpoint objects for the declared residual execution game: propagator mass under full-state write-back and adjoint semivariation under first-order feedback. Their role is not to replace those literatures, but to connect operation synthesis and arithmetic transport in one semantics-indexed scaling surface.

### K.7 Open theoretical frontier

The profile-based viewpoint isolates five concrete problems:

1. characterize shared-library reachable widths for coherent and noncommuting operation dictionaries;
2. derive average-case or smoothed fixed-target spectral laws;
3. extend first-order feedback to higher-order error shaping through higher differences of the backward propagator path;
4. estimate task-quotiented spectra and semivariation profiles in large trained residual models with dimension-dependent guarantees; and
5. couple structural scaling with independently certified optimization gaps for retrained model families.