# WHY AND WHEN DEEP IS BETTER THAN SHALLOW: IMPLEMENTATION-AGNOSTIC STATE-TRANSITION MODEL OF DEEP LEARNING

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

## ABSTRACT

We ask when adding hidden layers improves generalization, in a model that keeps the layers fixed and varies only their number. A hidden layer is a self-map of a state space, a depth-$k$ network composes at most $k$ hidden layers with an output layer, and depth is compared within the nested family $\mathcal{H}_0 \subset \mathcal{H}_1 \subset \cdots$ built from one class $F$ of hidden layers; this compares a deep network with shallower networks built from the same layers, not with wider ones. Our message is that the statistical cost of depth is the metric entropy of the set $B(k, F)$ of compositions. The estimation error is bounded by an entropy integral over $B(k, F)$ with constants that do not depend on the depth, and the bound is matched from below when the output layer can see the hidden states. For Lipschitz layers on a bounded state space this entropy grows at most polynomially in $k$, and it stays bounded, or grows only like $\log k$, under contraction, equicontinuity, or nilpotent structure. Balanced against the approximation error, this gives depths $k^*(n)$ that grow with the sample size, and it separates the models whose estimation error is independent of the depth from those whose estimation error grows with it. Deep ReLU networks, unrolled solvers, and chain-of-thought computation are worked out; for the last two the number of steps is derived rather than assumed. All statements are machine-checked in Lean 4.

## 1 INTRODUCTION

Deeper models often generalize better, and learning theory does not make this automatic: every added layer enlarges the hypothesis class, and naive complexity bounds grow exponentially in the depth $k$. This paper asks a precise version of the question. Fix a class $F$ of hidden layers and a class $H$ of output layers, let $\mathcal{H}_k$ be the networks with an output layer from $H$ and at most $k$ hidden layers from $F$, and consider the nested family $\mathcal{H}_0 \subset \mathcal{H}_1 \subset \cdots$. What is the statistical cost of the depth $k$ in this family, and when is the best depth $k^*(n)$ larger than one and growing with the sample size $n$? This compares a deep network with shallower networks built from the same layers. It is the design question for models whose layers are shared by construction, such as unrolled solvers, iterative denoisers, samplers of diffusion models, neural operators, and step-by-step reasoning, where the depth is the number of iterations or steps. It is not a comparison of deep and wide networks at equal parameter count; width enters through the layer class $F$ and is held fixed when depth varies.

**The model.** We describe a network as a state-transition model (Figure 1). The *state space* $\mathcal{X}$ is a metric space in which the input and all hidden representations live, for instance a bounded subset

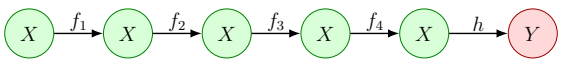


Figure 1: A network of depth $k = 4$ in the state-transition model. The input and every hidden representation are points of one state space $\mathcal{X}$; each hidden layer is a map $f_i : \mathcal{X} \to \mathcal{X}$; the output layer $h : \mathcal{X} \to \mathbb{R}$ reads the answer from the last state. Depth is the number of hidden maps composed before the output layer.



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.

of $\mathbb{R}^m$. A *hidden layer* is a continuous map $f : \mathcal{X} \to \mathcal{X}$, and $F$ is the set of admissible hidden layers. The *depth-$k$ hidden class* $B(k, F)$ is the set of all compositions $f_m \circ \cdots \circ f_1$ of at most $k$ maps from $F$, including the identity. An *output layer* is a map $h : \mathcal{X} \to \mathbb{R}$ that reads the answer from the last state, and $H$ is the set of admissible output layers. The *hypothesis class* at depth $k$ is $\mathcal{H}_k := H \circ B(k, F) = \{h \circ f : h \in H,\ f \in B(k, F)\}$. The *target class* $\mathcal{C}$ is the set of functions the learner is compared with; it need not be contained in any $\mathcal{H}_k$. Only the metric of $\mathcal{X}$ and the composition of maps in $F$ are used, so the same objects describe fully connected networks, unrolled solvers, and symbolic computation.

**Message and evidence.** The statistical cost of depth is the metric entropy of $B(k, F)$, the logarithm of the number of maps in $B(k, F)$ that can be told apart at a given resolution. The estimation error is independent of the depth when composition creates no new distinguishable maps, and it grows with the depth when it does. Three results support this message.

1. (Section 3.) Under three standing assumptions, the depth-dependent part of the estimation error is bounded by an entropy integral over $B(k, F)$, with a constant that depends only on the output-layer class and not on the depth, and when the output layer can tell the hidden states apart on the sample the same entropy appears in a lower bound (Theorem 1, Corollary 1).
2. (Section 4.) The entropy of $B(k, F)$ has a universal upper bound, the layerwise covering bound: for Lipschitz parametric layers on a bounded state space it grows at most polynomially in $k$, and the exponentially growing Lipschitz constant of a composition enters only through its logarithm (Theorem 2). Five mechanisms decide where a layer class sits below this bound: contraction, equicontinuity, and nilpotent control make the entropy constant or logarithmic in $k$, separation makes it linear, and memory-preserving expansion attains the bound (Theorem 3).
3. (Sections 5 and 6.) Balancing the resulting variance profiles against the approximation error gives depths $k^*(n)$ that grow with $n$ (Corollary 2). If the estimation bound is independent of the depth, the model may be run to convergence; if it grows polynomially, the depth has to be balanced against the bias. Three examples carry this out: for deep ReLU networks the Lipschitz constant of a layer selects the mechanism, and for unrolled solvers and chain-of-thought computation both sides of the balance are derived within the framework, so the step counts are consequences rather than inputs (Examples 1 and 2; unrolled solvers in Section M).

**Background and position.** Two lines of work motivate the model. The first is the theory of depth in deep ReLU networks, summarized below, which established sharp rates for one architecture but not a notion of depth that survives a change of architecture. The second is the family of models in which depth is an iteration count: unrolled optimization and fixed-point solvers, deep equilibrium models, neural ODEs, the samplers of diffusion and flow-matching models, Picard-type neural operators, and chain-of-thought reasoning in language models. For these models the layer is fixed by design and the question "how many steps" is asked at deployment time, so a theory that keeps the layer class fixed and varies only the depth speaks to them directly; Section B reviews both lines and places the known ReLU results in the coordinates of this paper, and Section J works out the teacher–student and neural-operator examples that connect them.

**Relation to the theory of deep ReLU networks.** Most rigorous results on depth are proved for one implementation, the fully connected ReLU network: depth separation in approximation (Eldan & Shamir, 2016; Telgarsky, 2016; Yarotsky, 2017; Siegel, 2023), VC-dimension, norm-based, and compression-based estimation bounds (Bartlett et al., 2019; 2017; Golowich et al., 2018; Arora et al., 2018; Suzuki et al., 2020), and minimax rates on smoothness classes and hierarchical compositions (Schmidt-Hieber, 2020; Suzuki, 2019). These results are sharp, but their depth dependence is read off an explicit parameterization, and a change in the layer structure, the state space, or the parameter sharing requires a new proof. Our framework isolates depth in $B(k, F)$, separately from the layer class $F$; the known ReLU rates then appear as one case of the analysis (the polynomial profile), and the question of what makes depth cheap is answered by the geometry of composition rather than by the parameter count. A detailed comparison with the literature is given in Section B.

**Roadmap.** Section 2 fixes the learning problem. Section 3 reduces the cost of depth to the entropy of $B(k, F)$; the precise statements and proofs are in Sections C to E. Section 4 classifies the growth of this entropy; the formal conditions, examples, and proofs are in Sections F to H. Section 5 derives the balancing depths; the four regimes and their calculations are in Section I. Section 6 explains how to apply the framework to a concrete model. The worked examples are in Section J (teacher–student models, neural operators, and the implementation layer), Section L (chain-of-thought computation), Section M (unrolled solvers and diffusion samplers), and Section K (deep ReLU networks, whose Lipschitz constant selects the growth mechanism). Every statement of the paper is formalized and machine-checked in Lean 4; the anonymized source and a blueprint linking each statement to its formal counterpart are provided as supplementary material.

## 2 SETTING

**Learning problem.** We observe an i.i.d. sample $\mathcal{D} = ((X_1, Y_1), \dots, (X_n, Y_n))$ from a distribution $P$ on $\mathcal{X} \times \mathbb{R}$, and write $S = (X_1, \dots, X_n)$ for the inputs. The loss $\ell : \mathbb{R} \times \mathbb{R} \to [0, b]$ is bounded and Lipschitz in the prediction (Assumption 1 below; clipped squared loss, hinge, ramp, or logistic loss on a bounded range; cross-entropy needs clipping, Remark 2). The risk and the empirical risk are $L[f] := \mathbb{E}\,\ell(f(X), Y)$ and $\hat{L}[f] := n^{-1}\sum_i \ell(f(X_i), Y_i)$. The learner picks an $\eta$-approximate empirical risk minimizer $\hat{f} \in \mathcal{H}_k$, that is, $\hat{L}[\hat{f}] \le \inf_{f\in\mathcal{H}_k} \hat{L}[f] + \eta$, and outputs an implementation $\hat{h} := \Phi_{\mathrm{embed}}(\hat{f})$, where the *implementation map* $\Phi_{\mathrm{embed}}$ sends each abstract predictor to a concrete one, for instance a finite-width network. Two errors are attached to the depth $k$:

$$\varepsilon_{\mathrm{imp}}(k) := \sup_{f\in\mathcal{H}_k} \|f - \Phi_{\mathrm{embed}}(f)\|_\infty, \qquad \varepsilon_{\mathrm{model}}(k) := \inf_{f\in\mathcal{H}_k} L[f] - \inf_{c\in\mathcal{C}} L[c],$$

the *implementation error* and the *approximation error*. The latter is a gap between risks; a uniform approximation rate transfers to it through the Lipschitz loss (Remark 1),

$$\varepsilon_{\mathrm{model}}(k) \le \beta_\ell \sup_{c\in\mathcal{C}} \inf_{f\in\mathcal{H}_k} \|f - c\|_\infty. \tag{1}$$

**Distances, covering numbers, and complexity.** Two distances between hidden maps $f, g : \mathcal{X} \to \mathcal{X}$ are used: the uniform distance $d_\infty(f, g) := \sup_x d(f(x), g(x))$ and the empirical distance $d_S(f, g) := (n^{-1}\sum_i d(f(X_i), g(X_i))^2)^{1/2}$, which only looks at the sample and satisfies $d_S \le d_\infty$. For a set $A$ of maps and a resolution $\varepsilon > 0$, the *covering number* $N(A, d, \varepsilon)$ is the smallest number of balls of radius $\varepsilon$ (in the distance $d$) needed to cover $A$; it counts the maps in $A$ that can be told apart at resolution $\varepsilon$, and $\log N(A, d, \varepsilon)$ is the *metric entropy*. For a class $G$ of real-valued functions, the empirical Rademacher complexity

$$\hat{\mathfrak{R}}_S(G) := \mathbb{E}_\sigma\Big[\sup_{g\in G} \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\Big],$$

with independent random signs $\sigma_i$, measures how well $G$ can fit random signs on the sample; it is the standard measure of statistical complexity in uniform-convergence bounds. Finer conventions (external covering numbers, the semigroup view of $B(k, F)$) are recorded in Section A.

**Standing assumptions.** Three assumptions are used throughout; each is stated once here with the place where it enters.

**Assumption 1** (Bounded Lipschitz loss)**.** The loss $\ell : \mathbb{R} \times \mathbb{R} \to [0, b]$ is measurable and $\beta_\ell$-Lipschitz in its first argument, with $b$ and $\beta_\ell$ fixed with the loss, so that neither depends on the depth. It enters the bias–variance decomposition, Theorem 1(i).

**Assumption 2** (Output layer does not amplify hidden distances)**.** For a hidden map $f$ let $Z_f := \sup_{h\in H} n^{-1}\sum_i \sigma_i h(f(X_i))$ be the Rademacher average of the output layers applied after $f$. There is a constant $L_H$, depending only on $H$ and the metric of $\mathcal{X}$, such that for all compositions $f, g$ of maps from $F$, of any depth, and all $t > 0$, $\mathbb{P}_\sigma(|Z_f - Z_g| > t) \le 2\exp\big(-nt^2/(2L_H^2 d_S(f, g)^2)\big)$. It holds with $L_H$ equal to the Lipschitz constant of the feature map for norm-bounded linear output layers on any feature map, and for finite classes of Lipschitz output layers, including bounded scalarizations of a softmax head (Section D). Its point is that $L_H$ is fixed with the output layer and is the same at every depth: the hidden layers stay inside $B(k, F)$ and never enter $L_H$. It enters the hidden–output decomposition, Theorem 1(ii).

**Assumption 3** (Bounded state space, Lipschitz layers)**.** The state space has finite diameter $D := \operatorname{diam}(\mathcal{X})$, and every map in $F$ is $\Lambda$-Lipschitz for some $\Lambda \geq 0$. Boundedness is provided in practice by bounded activations, normalization layers, or clipping. It enters the growth analysis of Section 4 and the profiles of Section 5.

## 3 THE COST OF DEPTH IS THE ENTROPY OF $B(k, F)$

This section proves the first claim: the statistical cost of depth is the metric entropy of $B(k, F)$. The argument has three steps, collected in one theorem. The generalization error splits into an implementation, an approximation, and a statistical term; the statistical term splits into an output-layer part, which does not depend on the depth, and a hidden-layer part, which is an entropy integral over $B(k, F)$; and a converse shows that this entropy is paid whenever the output layer can see the hidden states.

**Theorem 1** (The statistical cost of depth)**.** *Fix $k \geq 0$ and a sample $S$.*

*(i) (Bias–variance decomposition.) Under Assumption 1, if $\mathcal{H}_k$ is separable in the uniform norm and pointwise bounded, then for every $\delta \in (0, 1)$, with probability at least $1 - \delta$, the $\eta$-approximate empirical risk minimizer satisfies*

$$L[\hat{h}] - \inf_{c \in \mathcal{C}} L[c] \leq \beta_\ell \, \varepsilon_{\mathit{imp}}(k) + \varepsilon_{\mathit{model}}(k) + \eta + 4\beta_\ell \, \hat{\mathfrak{R}}_S(\mathcal{H}_k) + 6b\sqrt{\tfrac{2\log(4/\delta)}{n}}. \quad (2)$$

*(ii) (Hidden–output decomposition.) Under Assumption 2,*

$$\hat{\mathfrak{R}}_S(\mathcal{H}_k) \leq \hat{\mathfrak{R}}_S(H) + \frac{12 L_H}{\sqrt{n}} \, \mathsf{V}_k(S), \qquad \mathsf{V}_k(S) := \int_0^{\operatorname{diam}_S(B(k,F))} \sqrt{\log N\big(B(k, F), d_S, \varepsilon\big)} \, \mathrm{d}\varepsilon. \quad (3)$$

*(iii) (Converse under output-visible separation.) If for each map $g$ in $B(k, F)$ some output layer $h_g \in H$ realizes the hidden geometry on the sample, $\|h_g \circ g - h_{g'} \circ g'\|_S \geq \kappa \, d_S(g, g')$ and $\|h_g \circ g\|_{S,\infty} \leq R_{\mathrm{out}}$ for all $g, g'$, then, with $M$ the packing number and $c$ a universal constant,*

$$\hat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c \sup_{\varepsilon > 0} \min\Big\{ \kappa\varepsilon \sqrt{\tfrac{\log M(B(k,F), d_S, 2\varepsilon)}{n}}, \ \tfrac{\kappa^2 \varepsilon^2}{R_{\mathrm{out}}} \Big\}. \quad (4)$$

*The precise statements, with all regularity conditions, are Theorem 4 in Section C, the restatement of (ii) in Section D, and Theorem 7 in Section E.*

**Reading the theorem.** Part (i) says that the statistical cost of depth is entirely carried by $\hat{\mathfrak{R}}_S(\mathcal{H}_k)$: the implementation error is controlled separately (Proposition 6 in Section J) and the approximation error is a property of the target class. Part (ii) removes the output layer from this cost: the first term is the complexity of the output layer alone, and the second is the *entropy integral* of $B(k, F)$, which adds up, over all resolutions $\varepsilon$ up to the diameter of $B(k, F)$ on the sample, the square root of the number of maps in $B(k, F)$ that are distinguishable at resolution $\varepsilon$. It is Dudley's chaining bound (Vershynin, 2018, Chapter 8) applied to the process $f \mapsto Z_f$, whose increments are controlled by Assumption 2; the constant is $12L_H$. Part (iii) is a Sudakov-type minoration (Vershynin, 2018, Chapter 7): the packing number $M(B(k, F), d_S, 2\varepsilon)$ is the largest number of maps in $B(k, F)$ at mutual distance at least $2\varepsilon$ on the sample, and it is comparable to the covering number, so (ii) and (iii) bound the hidden term by the same quantity from both sides. The converse concerns the complexity of the class, not the risk of every algorithm; it is informative exactly when the output layer sees the hidden states, and when hidden representations collapse the entropy of $B(k, F)$ is not paid and the upper bound is correspondingly loose (Section E.4).

**Corollary 1** (Matching depth dependence)**.** *Let the hypotheses of Theorem 1(iii) hold for every $k$ with $\kappa, R_{\mathrm{out}}$ independent of $k$, and fix a resolution $\varepsilon_0 > 0$. If $B(k, F)$ contains $e^{\Theta(k)}$ maps that are $2\varepsilon_0$-separated on the sample, then $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \gtrsim \sqrt{k/n}$ for $n \gtrsim k$; if it contains $k^{\Theta(1)}$ such maps, then $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \gtrsim \sqrt{\log k / n}$ for $n \gtrsim \log k$. When the covering numbers obey the matching upper bounds of Section 4, the hidden term therefore has the same depth dependence, $\sqrt{k}$ or $\sqrt{\log k}$, in the upper and the lower bound (Corollary 3 in Section E).*

The separation in (iii) has to be visible on the sample, not only in the uniform distance: Section L gives a family whose maps are separated in $d_\infty$ but not in $d_S$, for every sample. After this section, everything the depth does to the estimation error is contained in two quantities, the covering numbers of $B(k, F)$ at every resolution and its diameter; the next section determines how they grow with $k$.

# 4 How the entropy of $B(k, F)$ grows with the depth

This section proves the second claim. By Theorem 1(ii), the depth-dependent part of the estimation bound is $\mathsf{var}(k, n) := 12L_H \mathsf{V}_k(S)/\sqrt{n}$, and its growth with $k$ is the growth of the entropy integral $\mathsf{V}_k(S)$; we call the function $k \mapsto \mathsf{V}_k(S)$, up to constants, the *profile* of the variance bound. Since $d_S \le d_\infty$, it suffices to bound the covering numbers of $B(k, F)$ in the uniform distance together with its diameter. Two results do this: a bound that holds for every Lipschitz layer class (Theorem 2), and a list of the structures that lower the growth below it or attain it (Theorem 3). Throughout, Assumption 3 is in force.

## 4.1 A layerwise covering bound: depth costs at most polynomially

**Theorem 2** (Layerwise covering bound)**.** *Let every map in $F$ be $\Lambda$-Lipschitz, $\Lambda \ge 0$. Then for every $k \ge 0$ and $\varepsilon > 0$,*

$$N\big(B(k, F), d_\infty, \varepsilon\big) \le 1 + k\, N\Big(F, d_\infty, \frac{\varepsilon}{1 + \Lambda + \cdots + \Lambda^{k-1}}\Big)^k. \tag{5}$$

*Consequently, if the state space is bounded and $F$ is parametric, that is, $N(F, d_\infty, \varepsilon) < \infty$ with $\log N(F, d_\infty, \varepsilon) \le p \log(1/\varepsilon) + O(1)$, then $\mathsf{V}_k(S) = O(\sqrt{kp \log k})$ when $\Lambda \le 1$ and $\mathsf{V}_k(S) = O(k\sqrt{p \log \Lambda})$ when $\Lambda > 1$; if instead $\log N(F, d_\infty, \varepsilon) \le c\, \varepsilon^{-q}$ with $q < 2$, then $\mathsf{V}_k(S) = O(k^{(1+q)/2})$ for $\Lambda \le 1$ and $O(k^{1/2} \Lambda^{qk/2})$ for $\Lambda > 1$ (explicit constants in Corollary 9, Section G).*

The proof is the telescoping argument that also controls the implementation error: replacing each of $k$ layers by a nearby one moves the composition by at most $\varepsilon$. Three consequences are worth stating. First, for parametric layers on a bounded state space, depth costs at most polynomially, and the exponentially growing Lipschitz constant $\Lambda^k$ of a composition enters the covering numbers only through its logarithm; for norm-bounded ReLU networks this is the order of the VC-dimension bounds of Bartlett et al. (2019) up to a factor $\sqrt{k}$ and logarithms, obtained here without any piecewise-linear structure and hence for every layer class. Second, the bound is attained: a construction whose layers write a symbol, shift the memory, and expand it (mechanism (v) below) reaches the $k^2$ entropy of the parametric case and the exponential entropy of the nonparametric case (Section G). Third, $\Lambda^k$ does appear on an unbounded state space, where the reachable states lie in a ball of radius $\Lambda^k$ times that of the inputs (Lemma 8), so that $D_k(S)$ can grow like $\Lambda^k$; this is the product of layerwise norms of norm-based bounds (Neyshabur et al., 2015; Bartlett et al., 2017), it enters only through the diameter, and bounded activations, normalization layers (Ioffe & Szegedy, 2015; Ba et al., 2016), or a projection onto a bounded set remove it.

## 4.2 Structures below the layerwise bound, and mechanisms that attain it

Depth becomes cheaper than the layerwise bound when composition does not create many well-separated maps. Whether it does is not decided by the compactness of $\mathcal{X}$ or by the algebra of $F$ alone, but by how composition interacts with the metric. The following theorem lists the mechanisms; the formal conditions, named P1, P1', P2, E1, E2, and E3, are stated and proved with examples in Section F, and the theorem summarizes them.

**Theorem 3** (Growth mechanisms)**.** *Let $N_k(\varepsilon) := N(B(k, F), d_\infty, \varepsilon)$.*

- *(i) (Saturation by equicontinuity, P1; Section F.1.) If $\mathcal{X}$ is compact and the compositions of maps from $F$ are equicontinuous, in particular if the maps in $F$ are non-expanding, then $\sup_k N_k(\varepsilon) < \infty$ for every $\varepsilon > 0$.*
- *(ii) (Saturation by contraction, P1'; Section F.1.) If the maps in $F$ are uniformly contractive, with a nonempty invariant set $A$ and a bounded absorbing set, then for every $\varepsilon > 0$ and every*

$k$, $N_k(\varepsilon) \le N(A, \varepsilon/2) + \sum_{\ell < m(\varepsilon)} N(F^\ell, d_\infty, \varepsilon)$ *with* $m(\varepsilon) = O(\log(1/\varepsilon))$*, a bound that does not depend on* $k$*.*

*(iii) (Polynomial growth by nilpotent control, P2; Section F.2.) If the maps in* $B(k, F)$ *are images, under a Lipschitz action, of the ball of radius* $O(k)$ *in a group of polynomial growth of degree* $D$*, such as translations, shears, or upper-triangular matrices, then* $N_k(\varepsilon) \le C(1 + k/\varepsilon)^D$ *and* $D_k(S) = O(k)$*, with* $D_k(S) \le \operatorname{diam}(\mathcal{X})$ *on a compact state space.*

*(iv) (Exponential growth by separation, E1 and E2; Section F.3.) If* $F = \{f_1, \dots, f_r\}$ *and distinct compositions of the same length are separated by a fixed amount at a common point (E1), or* $F$ *satisfies the ping-pong coding condition, expanding on disjoint chambers and resetting elsewhere, with marker–anchor separation* $\alpha$ *(E2), then* $N_k(\varepsilon) \ge r^k$ *for all* $k$ *and all sufficiently small* $\varepsilon$*.*

*(v) (Attainment of the layerwise bound by memory-preserving expansion, E3; Section F.4.) There is a class* $F$ *of* $\lambda$*-Lipschitz maps whose layers write a symbol, shift the stored symbols, and expand them, for which* $\log N_k(\varepsilon) \asymp k^2$ *when the written symbols range over a finite-dimensional set, and* $\log N_k(\varepsilon) \asymp \lambda^{pk}$ *when they range over a set of entropy* $\delta^{-p}$*.*

The mechanisms explain one another. Contraction and equicontinuity saturate because Ascoli's theorem makes the set of all compositions totally bounded: after a few layers no new distinguishable maps appear (Section F.1). Nilpotent control saturates nothing but organizes the compositions by a group in which the number of elements at distance $k$ grows polynomially; these layers are realized exactly by affine ReLU layers, so the mechanism is not restricted to abstract examples. Freeness of the compositions is not enough for a lower bound, because contracting maps may collapse many compositions to nearly the same map; what is needed is a probe point (Lemma 7) at which distinct compositions stay apart, and expanding maps with resets provide one probe per composition. Expansion alone still does not create growth beyond $r^k$, because later layers can erase what earlier layers distinguished; growth beyond $\sqrt{k}$ in the profile requires a memory, and with a memory the entropy accumulates across layers up to the layerwise bound. Symbolic shifts with guarded steps are the basic example of (iv) (Example 6), and branching symbolic computation is of this kind in $d_\infty$, although its empirical entropy is much smaller (Section L).

Proposition 5 in Section H converts each mechanism, with its diameter bound, into the profile of $\mathsf{V}_k(S)$: constant under (i) and (ii), $\sqrt{\log k}$ under (iii) and for branching computation on any sample, $\sqrt{k}$ under (iv) and for the layerwise bound of non-expanding parametric layers, $k$ for the layerwise bound of expanding parametric layers, which (v) attains, and exponential for expanding nonparametric layers (listed by mechanism in Section H). All these bounds are stated in $d_\infty$ and transferred to $d_S$ through $d_S \le d_\infty$; the transfer is one-sided, and Section L shows a family whose entropy grows like $k$ in $d_\infty$ but only like $\log k$ in $d_S$, for every sample. These five profiles are the variance side of the balance in the next section.

## 5 Balancing the bias against the estimation error

This section proves the third claim: the profile of the variance bound, balanced against a law for the approximation error, gives a depth $k^*(n)$ that grows with the sample size, and the kind of profile decides whether the depth has to be balanced at all. Write $\mathsf{bias}(k) := \varepsilon_{\text{model}}(k)$ for the approximation error, which decreases with the depth and depends on the target class, and $\mathsf{gen}(k, n) \lesssim \mathsf{bias}(k) + \mathsf{var}(k, n)$ for the excess-risk bound of Theorem 1(i) with the depth-independent terms omitted. Since $\mathsf{bias}(k)$ is nonincreasing and $\mathsf{var}(k, n)$ is nondecreasing in $k$, the bound is minimized, up to a factor 2, at the depth $k^* = k^*(n)$ where the two terms are of the same order (Lemma 10). Two laws for each term cover the mechanisms of Section 4 and the approximation results of the examples. The bias decays exponentially, $\mathsf{bias}(k) = e^{-\Theta(\alpha k)}$, when the target is itself an iteration that converges geometrically, as for contractive iterations, or polynomially, $\mathsf{bias}(k) = \Theta(k^{-\beta})$, as for discretization errors and smoothness classes. The variance profile is logarithmic, $\mathsf{var}(k, n) = O(\sqrt{\log k / n})$, which also covers the saturated case, or polynomial, $\mathsf{var}(k, n) = O(\sqrt{k^\gamma / n})$ with $\gamma = 1$ for the $\sqrt{k}$ profile and $\gamma = 2$ for the $k$ profile. We name the combinations by two letters, the bias first: EL is exponential bias with logarithmic variance, PP is polynomial bias with polynomial variance (Figure 2); the mixed cases EP and PL are treated in Section I.

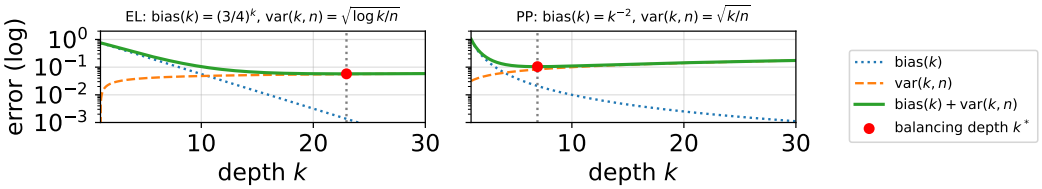


Figure 2: Balancing the bias (dotted) and the variance bound (dashed) in the depth $k$; the solid curve is their sum and the red dot its minimizer, the balancing depth $k^*$; $n = 1000$, same axes in both panels. Left, EL: bias $(3/4)^k$, as for a contractive iteration with contraction factor $3/4$, and the logarithmic variance profile $\sqrt{\log k/n}$; $k^*$ is a threshold, and deeper models are as good in the bound. Right, PP: bias $k^{-2}$, as for an order-2 discretization or a Jackson rate with $s/d = 2$, and the polynomial profile $\sqrt{k/n}$; depth is a budget, and the bound increases beyond $k^*$. The other two combinations are derived in Section I.

**Corollary 2** (Profiles and balancing depths)**.** *Under Assumptions 1 to 3, the variance bound* $\mathsf{var}(k, n)$ *of Theorem 1 has, up to constants, the profile* $n^{-1/2}$ *under the saturation mechanisms of Theorem 3(i)–(ii),* $\sqrt{\log k/n}$ *under nilpotent control (iii),* $\sqrt{k/n}$ *under the layerwise bound of Theorem 2 for non-expanding parametric layers and under the separation mechanisms (iv), and* $k/\sqrt{n}$ *under the layerwise bound for expanding parametric layers, which (v) attains (Proposition 5 in Section H). Balanced against the two bias laws, the balancing depth and the resulting rate are* $k^* = \frac{\log n}{2\alpha}(1+o(1))$ *and* $\mathsf{gen}(k^*, n) \asymp \sqrt{\log\log n/n}$ *in the EL case, and* $k^* \asymp n^{1/(2\beta+\gamma)}$ *and* $\mathsf{gen}(k^*, n) \asymp n^{-\beta/(2\beta+\gamma)}$ *in the PP case, with the mixed cases EP and PL and the explicit constants in Section I.*

In both cases the balancing depth grows with the sample size: larger datasets support deeper networks. An exponentially decaying bias asks for a logarithmic depth, a polynomially decaying bias for a polynomial depth.

**Depth-independent and depth-dependent estimation error.** The two panels of Figure 2 differ in kind, not only in rate. When the entropy integral saturates, $\mathsf{var}(k, n)$ is bounded by a quantity that does not depend on $k$, so the bound is nonincreasing in the depth: $k^*$ is a threshold rather than an optimum, every depth beyond it is as good, and the model may be run to convergence. This is the case of contractive iterations, fixed-point solvers, iterative denoisers, and append-only symbolic computation (Sections L and M). When the profile is logarithmic, the estimation bound grows so slowly that $k^*$ is set by the bias. When the profile is polynomial, which is the case of the layerwise bound and hence of generic parametric layers, the estimation bound grows with the depth and has to be balanced against the bias, and deeper is better only as far as the bias improves. Width is held fixed throughout; it enters the one-layer covering numbers, with $p$ of the order of the number of parameters per layer, the output-layer class, the bias, and the implementation error (Remark 3). The three results are now in place; the next section applies them to three architectures, and in two of them both sides of the balance are derived rather than assumed.

## 6 Using the Framework

### 6.1 A recipe for a given architecture

To apply the results to an architecture one has to (1) choose the state space, the layer class $F$, and the output-layer class $H$, and check Assumption 2, which holds for fixed linear or finite output layers; (2) bound the covering numbers and the diameter of $B(k, F)$, by the layerwise bound or by one of the mechanisms of Theorem 3, which gives the variance profile; (3) bound the approximation error through (1); (4) control the implementation error of the concrete network (Proposition 6 in Section J); and (5) balance the two terms. Step (2) decides the cost of depth, and it depends on the geometry of the layers, not on their parameter count. We carry the recipe out for fully connected ReLU networks, where the layer class is given and the framework reads off the cost of depth, and for chain-of-thought computation, where the layers are shared by design, the depth is a number of steps,

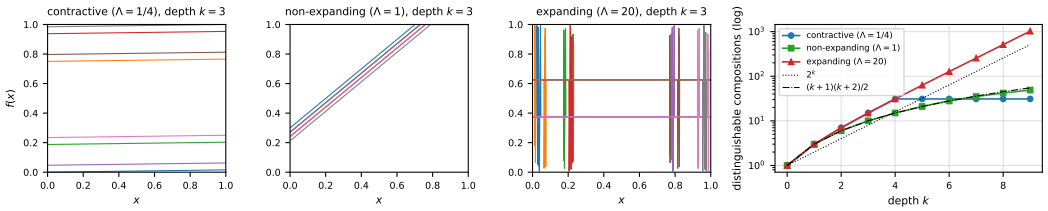


Figure 3: Three classes of two ReLU maps of $[0, 1]$ with Lipschitz constants $1/4$, 1, and 20. Left three panels: all compositions of depth 3. Right: distinguishable compositions of depth at most $k$ at resolution $10^{-2}$ (blue circles contractive, green squares non-expanding, red triangles expanding), with the references $(k + 1)(k + 2)/2$ (dash-dotted) and $2^k$ (dotted). See Sections 6.2 and K.4.

and both the profile and the bias are derived, so that the number of steps a sample of size $n$ justifies is a consequence rather than an input. Unrolled solvers and diffusion samplers are of the second kind and are worked out in Section M: contractive iterations are in the EL case with $k^* \asymp \log n$, and fixed-horizon integrators in the PL case with $k^* \asymp n^{1/(2p)}$. Each example comes with a computed figure that shows the same quantity, the number of compositions in $B(k, F)$ that are distinguishable at a fixed resolution, as a function of the depth; Section J lists what is derived and what is assumed.

### 6.2 Fully connected ReLU networks

*Example* 1 (Deep ReLU networks with clipped states). The state space is $[0, 1]^m$, a hidden layer is one ReLU block $x \mapsto \Pi_{[0,1]^m}(W\sigma(Ax + b) + c)$ with norm-bounded weights and clipping, and the output layer is linear. If the blocks are $\Lambda$-Lipschitz with $\Lambda < 1$, the profile saturates (EL with $k^* \asymp \log n$); if they are non-expanding and commute, it is logarithmic; if they are only Lipschitz, the layerwise bound gives the polynomial profile (PP with $k^* \asymp n^{1/(2\beta+\gamma)}$), and for $\Lambda > 1$ this bound is attained by ping-pong blocks (Proposition 8 in Section K).

For layers that are only Lipschitz and norm-bounded, the layerwise bound places the network in the PP case on a bounded state space, in agreement with the VC-dimension and nonparametric analyses of deep ReLU networks (Section B); on an unbounded state space only the diameter can grow exponentially. Structure lowers the profile, and Figure 3 shows this on the smallest ReLU networks: three classes of two one-hidden-layer ReLU maps of $[0, 1]$ into itself that differ only in their Lipschitz constant, namely the two contractions $x/4$ and $(x + 3)/4$ ($\Lambda = 1/4$), two clipped translations by incommensurable positive steps, which commute ($\Lambda = 1$), and two ping-pong maps that expand a subinterval onto $[0, 1]$ and reset the rest to a constant ($\Lambda = 20$). The left panels show all eight compositions of depth 3, and the right panel counts the compositions of depth at most $k$ that are distinguishable at resolution $10^{-2}$: the contractive count saturates after four layers, the commuting count grows quadratically, and the expanding count grows like $2^k$ (maps, evaluation set, and probe points in Section K.4). The same layer type thus realizes every case of Theorem 3, and the case is decided by the Lipschitz constant and the algebra of the layers, not by the width or the parameter count.

### 6.3 Chain-of-thought computation

*Example* 2 (Programs on a scratchpad). The state is a scratchpad, an infinite sequence $x = (x_0, x_1, \dots)$ of symbols from an alphabet with $r$ active symbols and a padding symbol, with the metric $d(x, y) = \theta^{n(x,y)}$, where $n(x, y)$ is the first position at which $x$ and $y$ differ and $\theta \in (0, 1)$; a hidden layer is one reasoning step, and the output layer reads the first symbols linearly. For append-only steps the profile saturates and the bias of a long target program decays like $\theta^k$, so $k^* \asymp \log n$ (EL; Proposition 9). For branching steps the uniform entropy grows like $k$, but on every sample of size $n$ the empirical entropy grows like $\log k$, so the cost of $k$ steps is $\sqrt{\log k/n}$ (Proposition 10 in Section L).

The two step rules differ in what a step does to the prefix. An *append-only* step with symbol $a$ writes $a$ in front, $g_a(x) = (a, x_0, x_1, \dots)$, so each $g_a$ is a contraction with factor $\theta$ and the entropy

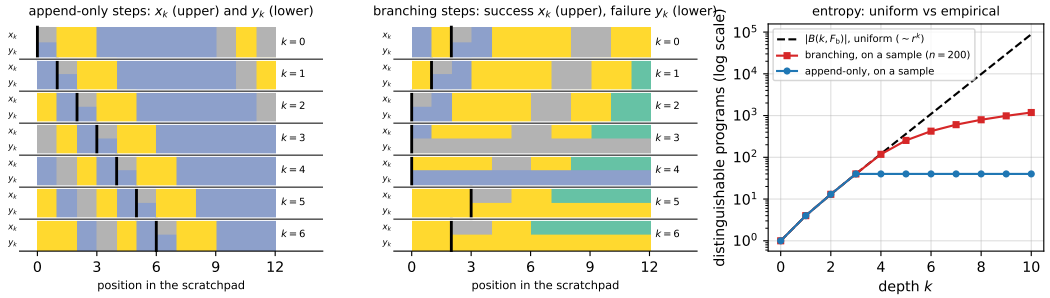


Figure 4: Chain-of-thought computation with $r = 3$ symbols. Left and middle: scratchpads $x_k$ (upper row of each pair) and $y_k$ (lower row) of two inputs after $k$ steps of one program, under append-only steps (left) and branching steps (middle); one colour per symbol, grey for padding, and a black bar at the first position where the two differ. Right: distinguishable programs of length at most $k$ on a sample of 200 inputs (red squares branching, blue circles append-only) and the number of programs (dashed). See Sections 6.3 and L.4.

saturates. A *branching* step with symbol $a$ reads the current symbol: if $x_0 = a$ it consumes it, $f_a(x) = (x_1, x_2, \dots)$, and otherwise it resets the scratchpad to the constant sequence $(a, a, \dots)$. Distinct programs of $k$ branching steps are separated in $d_\infty$, so their number grows like $r^k$; but on any sample of size $n$, programs of the same length succeed on disjoint sets of inputs, so at most $n$ of them are distinguishable from a constant map, and the empirical entropy grows only like $\log k$. Figure 4 shows both rules with $r = 3$ and $\theta = 1/2$. In the left and middle panels each pair of rows shows the scratchpads of two inputs after $k$ steps of the program $(1, 2, 3, 1, 2, 2)$, one colour per symbol, with a black bar at the first position at which they differ: under append-only steps the bar moves right by one per step, and under branching steps the lower input fails at its third step and is reset. The right panel counts, for a sample of 200 inputs, the programs of length at most $k$ that are distinguishable at resolution $\theta^3$ on the sample: for branching steps the count follows $3^k$ while that is below 200 and then grows only linearly, and for append-only steps it saturates at 40 (details in Section L.4).

## 7 Conclusion

We analyzed depth in a model in which the hidden layers are fixed and only their number varies. The statistical cost of depth is the metric entropy of the compositions $B(k, F)$: it bounds the estimation error with constants that do not depend on the depth, it is paid whenever the output layer can see the hidden states, and its growth with $k$ is decided by the geometry of composition. For Lipschitz layers on a bounded state space it grows at most polynomially, and contraction, equicontinuity, or nilpotent structure make it constant or logarithmic. Balanced against the approximation error, this gives the depth that a sample of size $n$ justifies, and it says whether that depth is a threshold beyond which further layers do not add to the estimation error or an optimum beyond which they do. The examples show that one layer type realizes every case according to its Lipschitz constant, and that for iterative models both sides of the balance follow from a single step. Deeper is better than shallower, in this family, when the target admits fast depth-driven approximation and the compositions of hidden layers do not create many distinguishable maps that the output layer and the sample can see.

### AI use statement

We used generative AI tools (large language models) for proof assistance and mathematical formulation: drafting proof steps and candidate statements that were subsequently checked, corrected, and completed by the authors, and assisting in writing the Lean 4 formalization of the results. All theorems, propositions, and lemmas in the paper and its appendices were verified by the authors by hand and, independently, are machine-checked by the Lean 4 proof assistant (the formalization is provided as anonymized supplementary material; it contains no `sorry` and its main theorems depend only on the standard axioms of Lean's kernel). We also used generative AI tools for drafting and editing parts of the text, for suggesting the structure of some sections, and for formatting refer-

ences; every reference was checked against its published source. We have not used generative AI tools for generating data, experiments, or figures. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

### Ethics statement

This is a theoretical paper and does not introduce datasets, deployed systems, or new training algorithms. Its main positive impact is conceptual: it clarifies when additional depth is statistically justified and may help researchers design or analyze architectures whose depth is used efficiently. Any general theory of deep learning can indirectly inform more capable models; such downstream uses inherit the usual application-specific concerns of machine learning systems, which are outside the scope of this paper.

### Reproducibility statement

All results are theoretical. The assumptions of each statement are given explicitly in the main text, and complete proofs are provided in the appendix: Section C (Theorem 4), Section D (Theorem 1), Section E (Theorem 7 and Corollary 3), Section F (the growth conditions and examples), Section G (Theorem 2), Section H (Proposition 5), Section I (the regime display (22)), Section J.3 (Proposition 6), and Sections L and M (the worked examples). All statements and proofs, in the form stated here, are formalized in Lean 4 and machine-checked; the anonymized Lean source, together with a human-readable blueprint that links every statement of the paper to its formal counterpart and a machine-generated certificate that the formal statements match the paper's theorems and depend only on the standard axioms, is included in the supplementary material. The figures are generated by the scripts `pics/figure-tradeoff.py` (closed-form expressions of Section I), `pics/figure-ode-example.py`, and `pics/figure-cot-example.py`, and `pics/figure-relu-example.py` (exact computations for the examples of Sections K to M), included in the supplementary material.

## References


Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger Generalization Bounds for Deep Nets via a Compression Approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 254–263. PMLR, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization, 2016.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep Equilibrium Models. In *Advances in Neural Information Processing Systems 32*, 2019.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–945, 1993.

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 6240–6249. Curran Associates, Inc., 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks. *Journal of Machine Learning Research*, 20:1–17, 2019.

Emmanuel Breuillard. Geometry of locally compact groups of polynomial growth and shape of large balls. *Group, Geometry, and Dynamics*, 8:669–732, 2014.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient Approximation of Deep ReLU Networks for Functions on Low Dimensional Manifolds. In *Advances in Neural Information Processing Systems 32*, pp. 8174–8184, 2019.

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. 6572–6583. Curran Associates, Inc., 2018.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2:303–314, 1989.

Pierre de la Harpe. *Topics in Geometric Group Theory*. University of Chicago Press, 2000.

Ronald A DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *manuscripta mathematica*, 63:469–478, 1989.

Weinan E. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5:1–11, 2017.

Ronen Eldan and Ohad Shamir. The Power of Depth for Feedforward Neural Networks. In *29th Annual Conference on Learning Theory*, volume 49, pp. 1–34, 2016.

Takashi Furuya, Koichi Taniguchi, and Satoshi Okuda. Quantitative Approximation for Neural Operators in Nonlinear Parabolic Equations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-Independent Sample Complexity of Neural Networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75, pp. 297–299. PMLR, 2018.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9:473–504, 2020.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34: 1–22, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33*, 2020.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A Formal Framework for Understanding Length Generalization in Transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Masaaki Imaizumi and Kenji Fukumizu. Deep Neural Networks Learn Non-Smooth Functions Effectively. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 869–878. PMLR, 2019.

Masaaki Imaizumi and Kenji Fukumizu. Advantage of Deep Neural Networks for Estimating Functions with Singularity on Hypersurfaces. *Journal of Machine Learning Research*, 23:1–54, 2022.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of The 32nd International Conference on Machine Learning*, volume 37, pp. 448–456. PMLR, 2015.

L K Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20:608–613, 1992.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural Operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24:1–97, 2023.

Anastasis Kratsios and Léonie Papon. Universal Approximation Theorems for Differentiable Geometric Deep Learning. *Journal of Machine Learning Research*, 23:1–73, 2022.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6:861–867, 1993.

Hongkang Li, Songtao Lu, Pin-Yu Chen, Xiaodong Cui, and Meng Wang. Training Nonlinear Transformers for Chain-of-Thought Inference: A Theoretical Generalization Analysis. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of Thought Empowers Transformers to Solve Inherently Serial Problems. In *The Twelfth International Conference on Learning Representations*, 2024.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31459–31473. Curran Associates, Inc., 2022.

Ryumei Nakada and Masaaki Imaizumi. Adaptive Approximation and Generalization of Deep Neural Network with Intrinsic Dimensionality. *Journal of Machine Learning Research*, 21:1–38, 2020.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-Based Capacity Control in Neural Networks. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pp. 1–26. JMLR W&CP, 2015.

Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148: 1567–1581, 2020.

Iosif Pinelis. Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22:1679–1706, 1994.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999.

G. Pisier. Remarques sur un r{e}sultat non publi{e} de B. Maurey. *S{e}minaire d'Analyse Fonctionnelle 1980-1981*, I:1–12, 1981.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The Little Engine That Could: Regularization by Denoising (RED). *SIAM Journal on Imaging Sciences*, 10:1804–1844, 2017.

Walter Rudin. *Functional Analysis*. McGraw-Hill Education, 2 edition, 1991.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48:1875–1897, 2020.

Hui Shen, Jingxuan Zhang, Boning Xiong, Rui Hu, Shoufa Chen, Zhongwei Wan, Xin Wang, Yu Zhang, Zixuan Gong, Guangyin Bao, Chaofan Tao, Yongfeng Huang, Ye Yuan, and Mi Zhang. Efficient Diffusion Models: A Survey. *Transactions on Machine Learning Research*, 2025. Survey Certification.

Jonathan W Siegel. Optimal Approximation Rates for Deep ReLU Neural Networks on Sobolev and Besov Spaces. *Journal of Machine Learning Research*, 24:1–52, 2023.

Jonathan W Siegel and Jinchao Xu. Sharp Bounds on the Approximation Rates, Metric Entropy, and n-Widths of Shallow Neural Networks. *Foundations of Computational Mathematics*, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.

Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20405–20422. PMLR, 2022.

Taiji Suzuki. Adaptivity of deep {ReLU} network for learning in {Besov} and mixed smooth {Besov} spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020.

Koichi Taniguchi and Sho Sonoda. Generalization Error Bounds for Picard-Type Operator Learning in Nonlinear Parabolic PDEs. In *Advances in Neural Information Processing Systems 40*, 2026.

Matus Telgarsky. Benefits of depth in neural networks. In *29th Annual Conference on Learning Theory*, pp. 1–23, 2016.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.

Yahong Yang and Juncai He. Deeper or Wider: A Perspective from Optimal Generalization Error with Sobolev Loss. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 56109–56138. PMLR, 2024.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75, pp. 639–649. PMLR, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13005–13015. Curran Associates, Inc., 2020.

## GUIDE TO THE APPENDIX

The appendix follows the order of the main text. Sections A and B collect conventions and related work. Sections C to E contain the statements and proofs of Section 3: the bias–variance decomposition, the hidden–output decomposition with the output-layer classes that satisfy Assumption 2, and the Sudakov-type converse. Sections F to H contain the growth mechanisms of Section 4 with their examples, the layerwise layerwise bound, and the variance profiles. Section I contains the four regimes of Section 5 and their balancing depths. The remaining appendices are the worked examples of Section 6, each written in the same order (setting, variance, bias, balancing, computed illustration, implementation, literature): teacher–student models and neural operators together with the implementation layer (Section J), deep ReLU networks (Section K), chain-of-thought computation (Section L), and unrolled solvers and samplers (Section M). Every statement is formalized in the Lean 4 supplement under the label shown in its LaTeX source.

# A Conventions and basic facts

This appendix fixes the conventions used in all proofs: covering and packing numbers and their elementary properties, the uniform and empirical metrics on hidden maps, the semigroup generated by a hidden-layer class, the symbolic state space used in several examples, and two remarks on the learning problem of Section 2. Nothing here is specific to a mechanism or an example; later appendices refer back to this one.

## A.1 Covering and packing numbers

For a metric space $(\mathcal{M}, \rho)$, write

$$N(A, \rho, \varepsilon) := \min \left\{ |C| \;\middle|\; C \subset \mathcal{M} \text{ s.t. } A \subset \bigcup_{y \in C} B_\rho(y, \varepsilon) \right\}$$

for the $\varepsilon$-covering number of $A \subset \mathcal{M}$, and

$$M(A, \rho, \varepsilon) := \max \{|S| \mid S \subset A, \rho(x, y) \geq \varepsilon \; \forall x \neq y \in S\}$$

for the $\varepsilon$-packing number. Thus $N$ is the external covering number by closed balls (centres anywhere in $\mathcal{M}$); it is monotone under inclusion, $A \subset A'$ implies $N(A, \rho, \varepsilon) \leq N(A', \rho, \varepsilon)$. The internal covering number $N^{\text{int}}(A, \rho, \varepsilon)$, whose centres are required to lie in $A$, satisfies $N(A, \rho, \varepsilon) \leq N^{\text{int}}(A, \rho, \varepsilon) \leq N(A, \rho, \varepsilon/2)$ but is *not* monotone under inclusion. Upper bounds below are stated for $N$; lower bounds are obtained through packing numbers, which are monotone under inclusion, and apply to both conventions.

The following facts are used throughout.

**Lemma 1** (Packing and covering numbers)**.** *For any subset $A$ of a metric space and $\varepsilon > 0$,*

$$M(A, 2\varepsilon) \leq N(A, \varepsilon) \leq M(A, \varepsilon).$$

**Lemma 2** (Lipschitz images)**.** *Let $(X, d_X), (Y, d_Y)$ be metric spaces. Suppose $\phi : X \to Y$ is $L$-Lipschitz, then for every subset $S \subset X$,*

$$N(\phi(S), d_Y, L\varepsilon) \leq N(S, d_X, \varepsilon), \quad M(\phi(S), d_Y, L\varepsilon) \leq M(S, d_X, \varepsilon).$$

**Lemma 3** (Subadditivity under unions)**.** *For any metric space $(M, d)$, subsets $F, G \subset M$, and $\varepsilon > 0$, we have*

$$N(F \cup G, \varepsilon) \leq N(F, \varepsilon) + N(G, \varepsilon).$$

*Proof.* Let $A, B$ be $\varepsilon$-coverings of $F, G$ respectively. Then, $A \cup B$ is an $\varepsilon$-covering of $F \cup G$ because for any $h \in F \cup G$ there exists $c \in A \cup B$ satisfying $d(h, c) \leq \varepsilon$. Thus, $N(F \cup G, \varepsilon) \leq |A \cup B| \leq |A| + |B| = N(F, \varepsilon) + N(G, \varepsilon)$. □

**Lemma 4** (Submultiplicativity under composition)**.** *For any bi-Lipschitz metric semigroup $(M, d)$ with $\sup_{f \in M} d(fx, fy) \leq \lambda d(x, y)$ (left Lipschitz) and $\sup_{f \in M} d(xf, yf) \leq \rho d(x, y)$ (right Lipschitz), for any subsets $F, G \subset M$, and $\varepsilon, \delta > 0$, we have*

$$N(FG, \varepsilon + \delta) \leq N(F, \varepsilon/\rho) N(G, \delta/\lambda).$$

*Proof.* Let $A, B$ be $\alpha, \beta$-coverings of $F, G$ respectively. Then, $AB$ is an $(\rho\alpha + \lambda\beta)$-covering of $FG$ because for any $fg \in FG$ there exists $ab \in AB$ satisfying $d(fg, ab) \leq d(fg, ag) + d(ag, ab) \leq \rho d(f, a) + \lambda d(g, b) = \rho\alpha + \lambda\beta$. Thus, $N(FG, \rho\alpha + \lambda\beta) \leq |AB| \leq |A||B| = N(F, \alpha)N(G, \beta)$. Letting $\alpha = \varepsilon/\rho, \beta = \delta/\lambda$ yields the assertion. □

**Lemma 5** (Right composition is 1-Lipschitz)**.** $d_\infty(a \circ f, b \circ f) \leq d_\infty(a, b)$

*Proof.* $d_\infty(a \circ f, b \circ f) = \sup_{x \in \mathcal{X}} d(a(f(x)), b(f(x))) \leq \sup_{y \in \mathcal{X}} d(a(y), b(y)) = d_\infty(a, b)$ □

**Lemma 6** (Left composition is Lipschitz)**.** $d_\infty(f \circ a, f \circ b) \leq \text{Lip}(f) d_\infty(a, b)$

*Proof.* $d_\infty(f\circ a, f\circ b) = \sup_{x\in\mathcal{X}} d(f(a(x)), f(b(x))) \le \sup_{x\in\mathcal{X}} \mathrm{Lip}(f) d(a(x), b(x)) = \mathrm{Lip}(f) d_\infty(a,b)$ □

**Lemma 7** (Packing lower bounds from finite probes)**.** *Let $P = \{x_1, \dots, x_m\} \subset \mathcal{X}$ be finite and define $d_P\big((y_j),(z_j)\big) := \max_{1\le j\le m} d(y_j, z_j)$ on $\mathcal{X}^m$, and $\mathrm{ev}_P : C(\mathcal{X},\mathcal{X}) \to \mathcal{X}^m, \mathrm{ev}_P(f) = (f(x_1), \dots, f(x_m))$. Then $\mathrm{ev}_P$ is 1-Lipschitz: $d_P\big(\mathrm{ev}_P(f), \mathrm{ev}_P(g)\big) \le d_\infty(f,g)$. If $\mathrm{ev}_P(S) \subset \mathcal{X}^m$ contains $M$ points that are pairwise $\delta$-separated (in $d_P$), then for every $\varepsilon < \delta/2$,*

$$N(S, d_\infty, \varepsilon) \ge M.$$

*Proof.* The 1-Lipschitz claim is immediate:

$$d\big(f(x_j), g(x_j)\big) \le \sup_{x\in\mathcal{X}} d\big(f(x), g(x)\big) = d_\infty(f,g).$$

If $\mathrm{ev}_P(s_1), \dots, \mathrm{ev}_P(s_M)$ are $\delta$-separated, then $d_\infty(s_i, s_j) \ge d_P(\mathrm{ev}_P(s_i), \mathrm{ev}_P(s_j)) \ge \delta$. Any ball of radius $\varepsilon < \delta/2$ in the $d_\infty$-metric can contain at most one of the $s_i$'s, so at least $M$ balls are needed. □

### A.2 Uniform and empirical metrics on hidden maps

Throughout, $(\mathcal{X}, d)$ is a metric space, not necessarily compact, and $C(\mathcal{X},\mathcal{X})$ is the set of continuous self-maps of $\mathcal{X}$.

**Uniform metric.** On $C(\mathcal{X},\mathcal{X})$ we use the uniform metric

$$d_\infty(f,g) := \sup_{x\in\mathcal{X}} d\left(f(x), g(x)\right).$$

When $\mathcal{X}$ is non-compact, this quantity may be infinite for some pairs of maps. Whenever a finite ambient metric is needed, we use one of the following two devices:

- ($d_b$) replace $d$ by its bounded version $d_b(x,y) := \min\{1, d(x,y)\}$;
- ($C_b$) fix a compact set $K_0 \subset \mathcal{X}$ and assume all maps under discussion (and all words built from them) take values in $K_0$.

These three settings—$C(\mathcal{X},\mathcal{X})$ with compact $\mathcal{X}$ and $d_\infty$, $C(\mathcal{X},\mathcal{X})$ with non-compact $\mathcal{X}$ and $d_b$, and $C_b(\mathcal{X},\mathcal{X})$ with non-compact $\mathcal{X}$ and $d_\infty$—give finite uniform metrics.

**Ascoli's theorem.** We use Ascoli's theorem in the following form (Rudin, 1991, Appendix A5): a family of continuous self-maps of a compact metric space is totally bounded in $d_\infty$ if and only if it is equicontinuous. On a non-compact space the theorem gives only compactness in the compact-open topology, which is why the bounded-metric and bounded-range devices above are needed.

**Lipschitz and co-Lipschitz constants.** For $f \in C(\mathcal{X},\mathcal{X})$ and $S \subset \mathcal{X}$ nonempty,

$$\mathrm{Lip}_S(f) := \sup_{x\ne y\in S} \frac{d(f(x), f(y))}{d(x,y)}, \qquad \mathrm{coLip}_S(f) := \inf_{x\ne y\in S} \frac{d(f(x), f(y))}{d(x,y)}.$$

We say "$f$ is uniformly expanding on $S$" if $\mathrm{coLip}_S(f) > 1$.

### A.3 Semigroups and word balls

For $F \subset C(\mathcal{X},\mathcal{X})$, write $\langle F\rangle$ for the semigroup generated by $F$ under composition. For $k \in \mathbb{N}$, let $B(k,F)$ be the set of maps obtainable as a composition of at most $k$ maps from $F$. Since $B(0,F) = \{\mathrm{id}\}$, the identity belongs to $\langle F\rangle = \bigcup_{k\ge 0} B(k,F)$, which is therefore a monoid.

**Word length.** For $h \in \langle F\rangle$, denote by $\ell_F(h)$ the least $\ell$ such that $h$ is a composition of $\ell$ maps from $F$. In all results below, when we assert that certain auxiliary maps $h$ belong to $\langle F\rangle$, we also assume that $\ell_F(h)$ is bounded by a constant independent of $k$.

**Homeomorphism and isometry groups.** $\mathrm{Homeo}(\mathcal{X}) \subset C(\mathcal{X}, \mathcal{X})$ denotes the homeomorphism group of $\mathcal{X}$, that is, the set of all *bijective* bi-continuous self-maps with the function composition as group operation. $\mathrm{Isom}(\mathcal{X}) \subset C(\mathcal{X}, \mathcal{X})$ denotes the isometry group of $(\mathcal{X}, d)$, that is, the set of all *bijective* distance-preserving self-maps. We only use these groups as sources of actions; the metric estimates below are stated explicitly in $d_\infty$.

### A.4 The full shift and its ultrametric

On the *one-sided full shift* $\Sigma_m^+ = [m]^{\mathbb{N}}$ (with $[m] = \{1, \dots, m\}$ discrete), fix $\theta \in (0, 1)$ and define the ultrametric

$$d_\theta(x, y) = \begin{cases} 0, & x = y, \\ \theta^{n(x,y)-1}, & x \neq y, \end{cases} \quad \text{where} \quad n(x, y) := \min\{i \geq 1 \mid x_i \neq y_i\}.$$

This is an ultrametric (satisfies *strong triangle inequality* $d(x, z) \leq \max\{d(x, y), d(y, z)\}$), it induces the product topology, and it makes $\Sigma_m^+$ a compact, totally disconnected, perfect (Cantor-type) space. Distances take only the values $1, \theta, \theta^2, \dots$; in particular, sequences that differ in the first symbol are at distance 1.

For a finite word $u = u_1 \cdots u_\ell$, the *(prefix) cylinder* of depth $\ell$ is given by

$$[\![u]\!] := \{w \in \Sigma_m^+ \mid w_1 = u_1, \dots, w_\ell = u_\ell\},$$

a clopen set; cylinders form a basis and the depth-$\ell$ cylinders partition $\Sigma_m^+$ into $m^\ell$ pieces. In the metric $d_\theta$, $\mathrm{diam}([\![u]\!]) = \theta^\ell$, and any two distinct depth-$\ell$ cylinders are separated by at least $\theta^{\ell-1}$. Moreover, balls coincide with cylinders: for $w \in \Sigma_m^+$ and $\ell \geq 0$, $\overline{B}_{d_\theta}(w, \theta^\ell) = [\![w_1 \cdots w_\ell]\!]$. The left shift $\sigma(w)_i = w_{i+1}$ satisfies $\sigma([\![au]\!]) = [\![u]\!]$ and $\sigma^{-1}([\![u]\!]) = \bigsqcup_{a \in [m]} [\![au]\!]$.

### A.5 Remarks on the learning problem

*Remark* 1 (Approximation term versus function approximation). $\varepsilon_{\text{model}}(k)$ is a gap between oracle risks, not the distance between a target function and its best approximant. The two notions are related through the Lipschitz loss: since $\varepsilon_{\text{model}}(k) = \sup_{c \in \mathcal{C}} \big(\inf_{f \in \mathcal{H}_k} L[f] - L[c]\big)$ and $L[f] - L[c] \leq \beta_\ell \|f - c\|_\infty$, (1) holds. Hence uniform approximation rates from approximation theory transfer to the bias term with the constant $\beta_\ell$. We call $\varepsilon_{\text{model}}(k)$ the approximation term when it is nonnegative, as is the case when $\mathcal{C}$ is a benchmark at least as good as the depth-$k$ hypothesis class in oracle risk; in complete generality it is an oracle gap relative to $\mathcal{C}$, and if a nonnegative approximation error is desired one may replace it by its positive part, at the cost of a weaker but still valid bound. The risk-based definition is used because it allows the same hypothesis class to be compared against different benchmarks and requires no norm structure on $\mathcal{C}$.

*Remark* 2 (Loss constants). For a fixed loss, $b$ and $\beta_\ell$ are fixed and do not depend on the depth; examples are the clipped squared loss, hinge and ramp losses, and the logistic loss on a bounded logit range. Cross-entropy expressed in class probabilities is unbounded and non-Lipschitz near the boundary of the simplex; expressed in logits its derivative is bounded but the loss itself is unbounded unless the logits are restricted to a bounded range. Such losses therefore enter Theorem 4 only after probability clipping or a logit range restriction. If the clipping threshold $\tau_{k,n}$ is allowed to vary with $k$ or $n$, then $b \asymp \log(1/\tau_{k,n})$ and $\beta_\ell \asymp 1/\tau_{k,n}$ must be carried through the rates in Section 5; with a fixed threshold the exponents there are unchanged.

## B Literature Overview

One motivation of this paper is to organize the generalization theory of deep learning that has been developed, for the most part, on the fully connected ReLU network. That theory contains several depth dependences, polynomial, logarithmic, and depth-independent, obtained by different techniques and under different assumptions; we review them below and then state, at the end of this appendix, how each of them appears as a case of the present framework. Section K works out the ReLU network as an example of the framework in its own right.

**Depth separation and expressivity.** A classical line of work shows that modest increases in depth can yield exponential representational advantages. Eldan & Shamir (2016) proved a three-vs-two-layer separation for a simple radial function, requiring exponential width for any depth-2

approximant, while Telgarsky (2016) established families exhibiting exponential gaps between networks of depth $O(k^3)$ and $O(k)$ with semi-algebraic gates (including ReLU) and provided constructive hard instances; subsequent work extended separations beyond radial constructions. These results clarify when *expressivity* favors depth, but do not by themselves pin down estimation behavior.

**Generalization via capacity control.** Combinatorial analyses give nearly tight VC/pseudodimension bounds for piecewise-linear networks, scaling roughly linearly in depth for fixed width/weights, providing a baseline picture of depth in classical uniform-convergence frameworks (Bartlett et al., 2019). Norm- and margin-based approaches bound estimation error through products of layer norms (path/spectral) (Neyshabur et al., 2015; Bartlett et al., 2017), sometimes yielding size-independent or gently depth-dependent bounds under additional structure (e.g., margin normalization). These strands highlight multiple possible depth dependencies—linear, polynomial, or even milder—depending on how complexity is measured.

**Rademacher/covering and size-independent bounds.** A complementary thread controls depth via data-dependent complexities (Rademacher, covering). Golowich et al. (2018; 2020) obtained bounds that (under norm constraints) improve the depth dependence and can be independent of width and depth in certain regimes; later refinements further reduced explicit depth factors. These works show how the estimation side may be decoupled from naively counted parameters and instead tied to geometric quantities of the hypothesis class.

**Compression and PAC-Bayes.** A productive viewpoint explains generalization via compressibility: if a trained network admits a succinct reparametrization, one can transfer that compression into generalization guarantees. Arora et al. (2018) formalized this link and demonstrated strong bounds in practice; follow-ups convert compression bounds to the original (non-compressed) networks and sharpen them via PAC-Bayes with subspace quantization, yielding state-of-the-art nonvacuous estimates (Suzuki et al., 2020; Lotfi et al., 2022). While depth typically enters these bounds through compressibility or margin quantities, the methodology is agnostic to architecture details.

**Nonparametric regression with deep networks.** Another large body of work analyzes approximation–estimation trade-offs of ReLU networks on smoothness classes. Schmidt-Hieber (2020) showed near-optimal rates in nonparametric regression, with depth playing an essential role; Suzuki (2019) established optimal adaptivity on (mixed) Besov spaces and improvements over linear/kernel baselines; Nakada & Imaizumi (2020) tied generalization to intrinsic (Minkowski) dimension; and Imaizumi & Fukumizu (2019; 2022) identified regimes with singularities where DNNs are minimax-superior to traditional estimators. Recent approximation results (e.g., optimal Sobolev/Besov rates) further sharpen the expressivity side (Yarotsky, 2017; 2018; Yarotsky & Zhevnerchuk, 2020; Siegel & Xu, 2022; Siegel, 2023; Yang & He, 2024). These analyses, however, are typically architecture-specific (ReLU feedforward) and hinge on smoothness assumptions.

**Iterative/hierarchical models and continuous depth.** Many modern systems are naturally modeled as *compositions* or *flows*—precisely the setting of our state-transition abstraction. Neural ODEs (Chen et al., 2018) as well as Neural Operators (Kovachki et al., 2023) treat depth as continuous time evolution; diffusion/score-based models (Shen et al., 2025) implement long iterative refinement; and chain-of-thought (Wei et al., 2022) prompting in LLMs explicitly unfolds multi-step reasoning. These families motivate studying depth-dependent generalization at the level of abstract state transitions rather than fixed architectures.

**Deep ReLU networks in the coordinates of this paper.** For Hölder, Sobolev, and Besov targets on $[0,1]^d$, the best $m$-parameter approximation has the Jackson rate $O(m^{-s/d})$ with matching lower bounds (DeVore et al., 1989); deep ReLU networks reach and, by bit-extraction, slightly surpass this rate, but the decay in the depth remains polynomial (Yarotsky, 2017; 2018; Yarotsky & Zhevnerchuk, 2020; Siegel, 2023; Yang & He, 2024). On the estimation side, VC-dimension and compression bounds grow polynomially in the depth (Bartlett et al., 2019; Arora et al., 2018; Suzuki et al., 2020; Lotfi et al., 2022). Hence deep ReLU networks on smoothness classes are in the PP case of Section 5, and the minimax analysis of Suzuki (2019) is an instance. For the hierarchical compositions of Hölder functions of Schmidt-Hieber (2020), the covering numbers of deep ReLU classes grow only logarithmically in the depth while the approximation error has an exponentially decaying compositional part, which is close to the PL case. In both cases the framework classifies known rates

rather than deriving new ones; its contribution is the quantity, the entropy of $B(k, F)$, that explains why the two cases differ.

**How the ReLU results appear in the framework.** Read in the coordinates of this paper, the depth dependences above are instances of the mechanisms of Section 4, with the ReLU layer as the hidden-layer class (Section K). The polynomial depth dependence of VC-dimension and covering-number bounds (Bartlett et al., 2019; Schmidt-Hieber, 2020) is the layerwise bound of Theorem 2 for a parametric layer class, which reproduces its order up to a factor $\sqrt{k}$ and logarithms. The product of layer norms in norm-based bounds (Neyshabur et al., 2015; Bartlett et al., 2017) is the diameter of the reachable states on an unbounded state space (Lemma 8); on a bounded state space it disappears. The depth-independent bounds obtained under norm constraints (Golowich et al., 2018) correspond to non-expanding layers on a bounded state space, which is the saturation mechanism P1 (Proposition 8(ii)). The exponential-in-depth worst cases of uniform-convergence bounds correspond to expanding layers with separated compositions, mechanism E2, and are realized by expand-and-reset ReLU layers (Proposition 8(iii)). On the approximation side, the minimax analyses on smoothness classes and hierarchical compositions (Suzuki, 2019; Schmidt-Hieber, 2020) supply the bias laws of the PP and PL regimes. The framework does not reproduce the constants of these proofs; it reproduces their depth dependence from one quantity, the entropy of $B(k, F)$, and the same computation applies to architectures that the ReLU theory does not cover.

**This Study.** Relative to these threads, this study is *implementation-agnostic*: instead of parameterizing a specific architecture, we analyze *state-hidden-layer semigroups on metric spaces*, derive depth dependence of the *variance* via covering/Rademacher complexity of word balls, give conditions for *polynomial/logarithmic* growth, and couple them with *exponential/polynomial* bias decay to compare depth scalings across four typical regimes. This yields a unified lens for when and why depth can be statistically preferable to shallower members of the same transition model—particularly in iterative/hierarchical settings suggested above. The distinction from the architecture-specific threads is one of scope rather than of rates: the sharp depth-dependent results above are proved for one implementation, most often a fully connected ReLU network on a Sobolev, Hölder, or Besov class, and a change in the layer structure, the state space, or the parameter sharing generally requires a new proof. Here the one-layer class $F$ carries all implementation-specific quantities, including width through its covering numbers, while composition depth is isolated in the word ball $B(k, F)$; existing rates for particular networks then become examples that can be inserted into the framework (Section J).

# C THE BIAS–VARIANCE DECOMPOSITION

This appendix states the bias–variance decomposition of Section 3 precisely, proves a general version of it for an arbitrary abstract class, and records the specialization to $\mathcal{H}_k$. The decomposition is the first step of the reduction: it isolates the statistical term $\hat{\mathfrak{R}}_S(\mathcal{H}_k)$ that the rest of the paper analyzes.

## C.1 STATEMENT

**Theorem 4** (Implementation-free bias–variance decomposition)**.** *Fix $k \geq 0$ and $\eta \geq 0$. Assume that $\mathcal{H}_k \subset L^\infty(\mathcal{X})$ is separable in the uniform norm and pointwise bounded, and that the loss $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$, $b > 0$, is measurable and $\beta_\ell$-Lipschitz in its first argument. Let $\hat{f} \in \mathcal{H}_k$ be an $\eta$-empirical minimizer and $\hat{h} := \Phi_{\mathrm{embed}}(\hat{f})$. Then for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $\mathcal{D} \sim P^n$,*

$$L[\hat{h}] - \inf_{c \in \mathcal{C}} L[c] \leq \beta_\ell \varepsilon_{\mathit{imp}}(k) + \varepsilon_{\mathit{model}}(k) + \eta + 4\beta_\ell \hat{\mathfrak{R}}_S(\mathcal{H}_k) + 6b\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{6}$$

*The same bound holds with $\hat{\mathfrak{R}}_S(\mathcal{H}_k)$ replaced by $\mathfrak{R}_n(\mathcal{H}_k)$; a generalization-gap version, valid for every $f \in \mathcal{H}_k$, is given in Section C.*

### C.2 A GENERAL VERSION AND ITS PROOF

We prove a slightly more general version of Theorem 4. In this appendix, the abstract class is denoted by $\mathcal{H}_{\text{abs}}$ rather than $\mathcal{H}_k$, because the argument does not use the state-transition representation $H \circ B(k, F)$. The specialization $\mathcal{H}_{\text{abs}} = \mathcal{H}_k$ gives the theorem in the main text.

**Theorem 5** (General bias–variance decomposition)**.** *Let $\mathcal{H}_{abs}$ be a class of measurable functions $\mathcal{X} \to \mathbb{R}$ that is separable in the uniform norm and pointwise bounded ($\sup_{f\in\mathcal{H}_{abs}} |f(x)| < \infty$ for every $x$), let $\mathcal{C}$ be a nonempty class of measurable benchmark functions $\mathcal{X} \to \mathbb{R}$, and let $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$, $b > 0$, be measurable and $\beta_\ell$-Lipschitz in its first argument. Define*

$$L[f] := \mathbb{E}[\ell(f(X), Y)], \qquad \hat{L}[f] := \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i),$$

*and set*

$$L_{\mathcal{C}} := \inf_{c\in\mathcal{C}} L[c], \qquad L_{\mathcal{H}_{abs}} := \inf_{f\in\mathcal{H}_{abs}} L[f], \qquad \varepsilon_{model} := L_{\mathcal{H}_{abs}} - L_{\mathcal{C}}.$$

*Let $\Phi_{\text{embed}} : \mathcal{H}_{abs} \to \mathcal{H}_{imp}$ be a implementation map into an implementation class and let $d_T$ be a pseudo-metric on $\mathcal{H}_{abs} \cup \mathcal{H}_{imp}$. Assume that*

$$\varepsilon_{imp} := \sup_{f\in\mathcal{H}_{abs}} d_T(\Phi_{\text{embed}}(f), f) < \infty$$

*and that, for every $f \in \mathcal{H}_{abs}$,*

$$|L[\Phi_{\text{embed}}(f)] - L[f]| \le \beta_L d_T(\Phi_{\text{embed}}(f), f),$$
$$|\hat{L}[\Phi_{\text{embed}}(f)] - \hat{L}[f]| \le \beta_{\hat{L}} d_T(\Phi_{\text{embed}}(f), f).$$

*For $\eta \ge 0$, let $\hat{f} \in \mathcal{H}_{abs}$ be an $\eta$-empirical minimizer:*

$$\hat{L}[\hat{f}] \le \inf_{f\in\mathcal{H}_{abs}} \hat{L}[f] + \eta,$$

*and output $\hat{h} := \Phi_{\text{embed}}(\hat{f})$. Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $\mathcal{D} \sim P^n$,*

$$L[\hat{h}] - L_{\mathcal{C}} \le \beta_L \varepsilon_{imp} + \varepsilon_{model} + \eta + 4\beta_\ell \hat{\mathfrak{R}}_S(\mathcal{H}_{abs}) + Cb\sqrt{\frac{\log(1/\delta)}{n}}, \tag{7}$$

$$L[\hat{h}] - \hat{L}[\hat{h}] \le (\beta_L + \beta_{\hat{L}})\varepsilon_{imp} + 2\beta_\ell \hat{\mathfrak{R}}_S(\mathcal{H}_{abs}) + Cb\sqrt{\frac{\log(1/\delta)}{n}}, \tag{8}$$

*where $C$ is a universal constant; explicitly, one may take the deviation term to be $6b\sqrt{2\log(4/\delta)/n}$ in (7) and $3b\sqrt{2\log(4/\delta)/n}$ in (8). On the same event, (8) holds for $\hat{h} = \Phi_{\text{embed}}(f)$ with every $f \in \mathcal{H}_{abs}$, not only for $\eta$-empirical minimizers. The same bounds hold with $\hat{\mathfrak{R}}_S(\mathcal{H}_{abs})$ replaced by $\mathfrak{R}_n(\mathcal{H}_{abs})$.*

*Proof.* No topological structure is needed on $\mathcal{C}$. Since the loss is bounded and nonnegative, $L[c] \in [0, b]$ for every $c \in \mathcal{C}$, so $L_{\mathcal{C}} = \inf_{c\in\mathcal{C}} L[c]$ is a well-defined real number whenever $\mathcal{C}$ is nonempty.

The only probabilistic ingredient is the standard uniform deviation bound for bounded Lipschitz loss classes. Since $\mathcal{H}_{\text{abs}}$ is separable in the uniform norm and pointwise bounded, the relevant suprema are finite and measurable, and symmetrization, bounded differences, and the contraction inequality imply that, with probability at least $1 - \delta$,

$$\Delta_{\mathcal{H}_{\text{abs}}} := \sup_{f\in\mathcal{H}_{\text{abs}}} |L[f] - \hat{L}[f]| \le 2\beta_\ell \hat{\mathfrak{R}}_S(\mathcal{H}_{\text{abs}}) + Cb\sqrt{\frac{\log(1/\delta)}{n}}. \tag{9}$$

Explicitly, applying the one-sided symmetrization and bounded-differences argument to $\ell \circ \mathcal{H}_{\text{abs}}$ and to $-\ell \circ \mathcal{H}_{\text{abs}}$, each with failure probability $\delta/2$, and the contraction inequality without absolute values, gives (9) with $Cb\sqrt{\log(1/\delta)/n}$ replaced by $3b\sqrt{2\log(4/\delta)/n}$. The population Rademacher version follows from the same argument after taking expectation over $S$.

On the event (9), the implementation error is controlled directly by the assumptions:

$$L[\hat{h}] - L[\hat{f}] \le \beta_L \varepsilon_{\text{imp}}, \qquad \hat{L}[\hat{f}] - \hat{L}[\hat{h}] \le \beta_{\hat{L}} \varepsilon_{\text{imp}}.$$

For the excess risk, no compactness or exact population minimizer is needed. Using only the infimum $L_{\mathcal{H}_{\text{abs}}}$ and the $\eta$-minimality of $\hat{f}$,

$$\begin{aligned} L[\hat{f}] - L_{\mathcal{H}_{\text{abs}}} &\le L[\hat{f}] - \hat{L}[\hat{f}] + \hat{L}[\hat{f}] - \inf_{f \in \mathcal{H}_{\text{abs}}} \hat{L}[f] + \inf_{f \in \mathcal{H}_{\text{abs}}} \hat{L}[f] - L_{\mathcal{H}_{\text{abs}}} \\ &\le \Delta_{\mathcal{H}_{\text{abs}}} + \eta + \Delta_{\mathcal{H}_{\text{abs}}} = 2\Delta_{\mathcal{H}_{\text{abs}}} + \eta. \end{aligned}$$

Therefore,

$$L[\hat{h}] - L_{\mathcal{C}} \le \beta_L \varepsilon_{\text{imp}} + \big(L[\hat{f}] - L_{\mathcal{H}_{\text{abs}}}\big) + \varepsilon_{\text{model}} \le \beta_L \varepsilon_{\text{imp}} + \varepsilon_{\text{model}} + 2\Delta_{\mathcal{H}_{\text{abs}}} + \eta.$$

Substituting (9) gives (7).

The generalization gap bound does not require empirical optimality and holds for $\hat{h} = \Phi_{\text{embed}}(\hat{f})$ with an arbitrary $\hat{f} \in \mathcal{H}_{\text{abs}}$:

$$\begin{aligned} L[\hat{h}] - \hat{L}[\hat{h}] &= \big(L[\hat{h}] - L[\hat{f}]\big) + \big(L[\hat{f}] - \hat{L}[\hat{f}]\big) + \big(\hat{L}[\hat{f}] - \hat{L}[\hat{h}]\big) \\ &\le (\beta_L + \beta_{\hat{L}})\varepsilon_{\text{imp}} + \Delta_{\mathcal{H}_{\text{abs}}}. \end{aligned}$$

Again substituting (9) proves (8). □

**Remarks.** (1) Specialization to Theorem 4. Take $\mathcal{H}_{\text{abs}} = \mathcal{H}_k$, $d_T(f,g) = \|f - g\|_\infty$, $\varepsilon_{\text{imp}} = \varepsilon_{\text{imp}}(k)$, and $\varepsilon_{\text{model}} = \varepsilon_{\text{model}}(k)$. The Lipschitz condition on the loss gives $\beta_L = \beta_{\hat{L}} = \beta_\ell$. If the empirical minimizer in the main text exists, set $\eta = 0$. Without exact attainment, the same proof applies to any $\eta$-empirical minimizer and adds only the displayed $\eta$ term to the excess-risk bound.

# D THE HIDDEN–OUTPUT DECOMPOSITION

This appendix proves Theorem 1(ii), the second step of the reduction, and shows that the common output-layer classes satisfy Assumption 2 with constants that do not depend on the depth. The theorem is restated in local notation, and the proof applies Dudley's entropy integral to the anchored process $Z_f - Z_{f_0}$; the increment condition (10) below is exactly Assumption 2.

## D.1 STATEMENT AND PROOF

**Theorem 6** (Theorem 1(ii), restated)**.** *Let $(X,d)$ be a metric space, let $S = (X_1, \dots, X_n) \subset X$ be fixed, and define*

$$d_S(f,g) := \Big(\frac{1}{n}\sum_{i=1}^{n} d\big(f(X_i), g(X_i)\big)^2\Big)^{1/2} \qquad (f, g : X \to X).$$

*Let $F \subset C(X,X)$ be totally bounded in $d_S$, and fix an anchor $f_0 \in F$. Write*

$$\operatorname{diam}_S(F) := \sup_{f,g \in F} d_S(f,g).$$

*Let $H \subset \mathbb{R}^X$ be a class of real-valued functions. For each $f \in F$ define*

$$Z_f(\sigma) := \sup_{h \in H} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h\left(f(X_i)\right),$$

*where $\sigma_1, \dots, \sigma_n$ are i.i.d. Rademacher variables, and assume that each $Z_f(\sigma)$ is finite.*

*Assume that there exist constants $L > 0$ and $A_H > 0$ such that, for all $f, g \in F$ and all $t > 0$,*

$$\mathbb{P}_\sigma\left(|Z_f - Z_g| > t\right) \le 2\exp\left(-\frac{nt^2}{2A_H^2 L^2 d_S(f,g)^2}\right). \tag{10}$$

*Then*

$$\hat{\mathfrak{R}}_S(H \circ F) \le \hat{\mathfrak{R}}_S(H \circ f_0) + \frac{12 A_H L}{\sqrt{n}} \int_0^{\operatorname{diam}_S(F)} \sqrt{\log N(F, d_S, \varepsilon)}\, d\varepsilon,$$

*where $\hat{\mathfrak{R}}_S(H \circ f_0) = \mathbb{E}_\sigma Z_{f_0}$ and the right-hand side is read as $+\infty$ (so that the bound is trivial) if the entropy integrand is not integrable on $[0, \operatorname{diam}_S(F)]$. If $\mathrm{id}_X \in F$, the choice $f_0 = \mathrm{id}_X$ gives the first term $\hat{\mathfrak{R}}_S(H)$. In particular, for $F = B(k, \mathcal{F})$ one obtains the same bound, with first term $\hat{\mathfrak{R}}_S(H)$, for the depth-$k$ class $H_k := H \circ B(k, \mathcal{F})$.*

*Proof.* Anchor the process at $f_0$ and write

$$Y_f := Z_f - Z_{f_0}, \qquad f \in F.$$

Then $Y_{f_0} = 0$, and by (10),

$$\mathbb{P}_\sigma\left(|Y_f - Y_g| > t\right) = \mathbb{P}_\sigma\left(|Z_f - Z_g| > t\right) \le 2\exp\left(-\frac{nt^2}{2A_H^2 L^2 d_S(f,g)^2}\right).$$

Thus $(Y_f)_{f\in F}$ is a process with sub-Gaussian increments with respect to the metric

$$\rho(f,g) := \frac{A_H L}{\sqrt{n}} d_S(f,g).$$

By Dudley's entropy integral bound for processes with sub-Gaussian increments (this is the only probabilistic ingredient; it applies to any process on a finite probability space with sub-Gaussian increments, anchored at an arbitrary index point, with the constant 12, and the Rademacher structure of $Z_f$ enters only through (10)),

$$\mathbb{E}_\sigma\Big[\sup_{f\in F} Y_f\Big] \le 12 \int_0^{\Delta_\rho(F)} \sqrt{\log N(F,\rho,u)}\, du,$$

where

$$\Delta_\rho(F) = \sup_{f\in F} \rho(f,f_0) \le \frac{A_H L}{\sqrt{n}} \operatorname{diam}_S(F).$$

Changing variables $u = \frac{A_H L}{\sqrt{n}}\varepsilon$ and enlarging the upper limit gives

$$\mathbb{E}_\sigma\Big[\sup_{f\in F} Y_f\Big] \le \frac{12 A_H L}{\sqrt{n}} \int_0^{\operatorname{diam}_S(F)} \sqrt{\log N(F,d_S,\varepsilon)}\, d\varepsilon.$$

Finally,

$$\hat{\mathfrak{R}}_S(H\circ F) = \mathbb{E}_\sigma\Big[\sup_{f\in F} Z_f\Big] = \mathbb{E}_\sigma\Big[Z_{f_0} + \sup_{f\in F}(Z_f - Z_{f_0})\Big] = \hat{\mathfrak{R}}_S(H\circ f_0) + \mathbb{E}_\sigma\Big[\sup_{f\in F} Y_f\Big],$$

and the claim follows (if the entropy integrand is not integrable there is nothing to prove). □

### D.2 Output-layer classes satisfying Assumption 2

Assumption 2 is used only for the upper-bound decomposition: it prevents the supremum over output layers from amplifying small hidden-state perturbations beyond the $d_S(f,g)/\sqrt{n}$ scale. It is not a output-separation assumption; the latter appears only in the Sudakov-type lower bound. The condition is satisfied by standard output-layer mechanisms once the last scalarization is bounded and Lipschitz.

**Proposition 1.** *Let $\mathcal{H}$ be a real inner product space (completeness is not needed), let $\Phi : X \to \mathcal{H}$ be $L$-Lipschitz, and define*

$$H_\Phi := \Big\{x \mapsto \langle w, \Phi(x)\rangle_{\mathcal{H}} : \|w\|_{\mathcal{H}} \le 1\Big\}.$$

*Then the assumption* (10) *holds with $H = H_\Phi$ and $A_H = 1$. Consequently,*

$$\hat{\mathfrak{R}}_S(H_\Phi \circ F) \le \hat{\mathfrak{R}}_S(H_\Phi) + \frac{12L}{\sqrt{n}} \int_0^{\operatorname{diam}_S(F)} \sqrt{\log N(F,d_S,\varepsilon)}\, d\varepsilon.$$

*Proof.* For $H_\Phi$, the supremum over the unit ball of $\mathcal{H}$ can be evaluated explicitly:

$$Z_f = \sup_{\|w\|_{\mathcal{H}}\le 1} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle w, \Phi(f(X_i))\rangle_{\mathcal{H}} = \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i \Phi(f(X_i))\Big\|_{\mathcal{H}}.$$

Hence, for any $f, g \in F$,

$$|Z_f - Z_g| \le \Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i\big(\Phi(f(X_i)) - \Phi(g(X_i))\big)\Big\|_{\mathcal{H}}.$$

Set

$$v_i := \Phi(f(X_i)) - \Phi(g(X_i)) \in \mathcal{H}.$$

The Rademacher tail bound in inner product spaces (Pinelis, 1994), whose constant 2 is optimal, gives, for every $t > 0$,

$$\mathbb{P}_\sigma\left(\Big\|\frac{1}{n}\sum_{i=1}^n \sigma_i v_i\Big\|_{\mathcal{H}} > t\right) \le 2\exp\left(-\frac{n^2t^2}{2\sum_{i=1}^n \|v_i\|_{\mathcal{H}}^2}\right).$$

Because $\Phi$ is $L$-Lipschitz,

$$\sum_{i=1}^n \|v_i\|_{\mathcal{H}}^2 \le L^2 \sum_{i=1}^n d\big(f(X_i), g(X_i)\big)^2 = nL^2 d_S(f,g)^2.$$

Combining the last two displays yields

$$\mathbb{P}_\sigma\left(|Z_f - Z_g| > t\right) \le 2\exp\left(-\frac{nt^2}{2L^2 d_S(f,g)^2}\right),$$

which is exactly (10) with $A_H = 1$. The final bound is now an immediate application of Theorem 6. □

In neural-network terminology, an MLP with a linear output layer fits Proposition 1 with the hidden layers placed in the state transition class $B(k, F)$ and a fixed feature map $\Phi$ (for instance the identity on $\mathbb{R}^m$, or a fixed embedding of the terminal state) followed by a norm-bounded linear functional as the output layer. With this convention $L = \mathrm{Lip}(\Phi)$ and $A_H = 1$ do not depend on the depth, which is the depth-uniform form of Assumption 2 required in Section 3. One may alternatively absorb some lower layers into the feature map, but then $\Phi = \Phi_k$ depends on the depth and $L$ must be replaced by $\mathrm{Lip}(\Phi_k)$, which for a $k$-layer network may grow exponentially in $k$; the bound remains valid but its depth dependence is then no longer carried by the word-ball entropy alone.

**Proposition 2** (Finite Lipschitz scalar output layers)**.** *Let $H = \{h_1, \dots, h_m\}$ be a finite class of real-valued functions on $X$. Suppose every $h_j$ is $L$-Lipschitz with respect to $d$. Then the assumption* (10) *holds with*

$$A_H = \left(1 + \frac{\log m}{\log 2}\right)^{1/2}.$$

*Proof.* For $j = 1, \dots, m$, define

$$S_j(f,g) := \frac{1}{n}\sum_{i=1}^n \sigma_i\big(h_j(f(X_i)) - h_j(g(X_i))\big).$$

Since the difference of two suprema is bounded by the supremum of the differences,

$$|Z_f - Z_g| \le \max_{1\le j\le m} |S_j(f,g)|.$$

For each fixed $j$, Hoeffding's inequality and the Lipschitz property give

$$\mathbb{P}_\sigma(|S_j(f,g)| > t) \le 2\exp\left(-\frac{nt^2}{2L^2 d_S(f,g)^2}\right).$$

Writing $x := nt^2/(2L^2 d_S(f,g)^2)$, the union bound yields

$$\mathbb{P}_\sigma(|Z_f - Z_g| > t) \le \min\{1, 2me^{-x}\}.$$

With $A_H^2 = 1 + \log m/\log 2$, the right-hand side is bounded by $2\exp(-x/A_H^2)$ for all real $x$ (so no restriction on $t$, $L$, $d_S(f,g)$, or $m \ge 1$ is needed), which is (10). □

Softmax output layers can be handled by including the logits/softmax map in the state transition and taking the final scalar quantity used for prediction or loss as the output layer. Coordinate probabilities are linear output layers on the simplex, while margins and clipped ramp-type losses are bounded Lipschitz finite-coordinate scalarizations, so Proposition 2 applies. Unconstrained post-softmax output-layer classes and unbounded cross-entropy near the boundary of the simplex require additional norm, range, or clipping assumptions.

**Remarks.** (1) Theorem 1 keeps the output-layer complexity and the hidden entropy separate; a cruder alternative, which chains over the whole class $H \circ F$ with the sum of the two entropies and needs no increment condition, is not used in this paper. (2) The increment condition is used only for the upper bound; the separation condition of Section E is a different and stronger requirement on the output layer.

# E THE SUDAKOV-TYPE CONVERSE

This appendix proves the conditional lower bound of Section 3, the third step of the reduction, and describes the output-layer classes for which its separation assumption holds. The assumption asks that, for each hidden map, some output layer realizes the empirical geometry of $B(k, F)$ inside the real-valued class; under it, the packing numbers of $B(k, F)$ bound the Rademacher complexity from below, and the depth dependence of Theorem 1 is matched. The empirical Rademacher complexity is assumed finite throughout, that is, the averages $n^{-1} \sum_i \sigma_i g(X_i)$, $g \in \mathcal{H}_k$, are bounded above for every sign pattern.

## E.1 STATEMENT AND PROOF

**Assumption 4** (Output-layer realization of hidden geometry)**.** There exist constants $\kappa, R_{\mathrm{out}} > 0$ and, for each $g \in B_k$, an output layer $h_g \in H$ such that $\Psi_k(g) := h_g \circ g$ satisfies $\|\Psi_k(g) - \Psi_k(g')\|_S \geq \kappa\, d_S(g, g')$ and $\|\Psi_k(g)\|_{S,\infty} \leq R_{\mathrm{out}}$ for all $g, g' \in B_k$.

**Theorem 7** (Conditional Sudakov-type lower bound)**.** *Under Assumption 4, if $\hat{\mathfrak{R}}_S(\mathcal{H}_k) < \infty$, then for a universal constant $c > 0$*

$$\hat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c \sup_{\varepsilon > 0} \min \left\{ \kappa\varepsilon \sqrt{\frac{\log M(B_k, d_S, 2\varepsilon)}{n}}, \frac{\kappa^2 \varepsilon^2}{R_{\mathrm{out}}} \right\}, \tag{11}$$

*where the first entry is $+\infty$ if the packing number is. The word-ball structure is not used: the statement holds for any $B \subset C(\mathcal{X}, \mathcal{X})$.*

**Theorem 8** (Theorem 7, restated)**.** *Fix $k \in \mathbb{N}$, a sample $S = (X_1, \ldots, X_n) \subset X$, and write*

$$B_k := B(k, F), \qquad H_k := H \circ B_k.$$

*(The word-ball structure is not used below: the statement and proof apply verbatim to an arbitrary subset $B \subset C(X, X)$ in place of $B_k$, with $H \circ B$ in place of $H_k$.) For real-valued functions on the sample, write*

$$\|u\|_{S,\infty} := \max_{i \in [n]} |u(X_i)|.$$

*Assume that there exist constants $\kappa, R_{\mathrm{out}} > 0$ and, for each $f \in B_k$, an output layer $h_f \in H$ such that*

$$\|h_f \circ f - h_g \circ g\|_S \geq \kappa d_S(f, g) \qquad (f, g \in B_k), \tag{12}$$

$$\|h_f \circ f\|_{S,\infty} \leq R_{\mathrm{out}} \qquad (f \in B_k). \tag{13}$$

*Equivalently, the map*

$$\Psi_k : B_k \to H_k, \qquad \Psi_k(f) := h_f \circ f,$$

*is $\kappa$-co-Lipschitz from $(B_k, d_S)$ to $(H_k, \|\cdot\|_S)$, and its image is uniformly bounded on the sample. Assume also that $\hat{\mathfrak{R}}_S(H_k) < \infty$, i.e. for every sign pattern $\sigma$ the averages $n^{-1} \sum_i \sigma_i g(X_i)$, $g \in H_k$, are bounded above (for instance, $H_k$ is uniformly bounded on the sample); without this the left-hand side below is not a finite supremum.*

*Then there exists a universal constant $c > 0$ such that, for every $\varepsilon > 0$,*

$$\hat{\mathfrak{R}}_S(H_k) \geq c \sup_{\varepsilon > 0} \min \left\{ \kappa\varepsilon \sqrt{\frac{\log M(B_k, d_S, 2\varepsilon)}{n}}, \frac{\kappa^2 \varepsilon^2}{R_{\mathrm{out}}} \right\}. \tag{14}$$

*By packing-covering duality, the same conclusion holds with $M$ replaced by $N$, up to absolute constants. When $M(B_k, d_S, 2\varepsilon) = \infty$, the first entry of the minimum is read as $+\infty$, so the bound is its second entry.*

*Proof.* Fix $\varepsilon > 0$, and let $f_1, \dots, f_M \in B_k$ be a $2\varepsilon$-packing of $B_k$ with respect to $d_S$, where

$$M = M(B_k, d_S, 2\varepsilon)$$

(if this packing number is infinite, take an arbitrary finite $2\varepsilon$-packing and let its cardinality tend to infinity at the end). Set

$$g_j := \Psi_k(f_j) = h_{f_j} \circ f_j \in H_k \qquad (j = 1, \dots, M).$$

By (12),

$$\|g_j - g_\ell\|_S = \|\Psi_k(f_j) - \Psi_k(f_\ell)\|_S \geq \kappa d_S(f_j, f_\ell) \geq 2\kappa\varepsilon \qquad (j \neq \ell).$$

Hence $\{g_1, \dots, g_M\} \subset H_k$ is a $2\kappa\varepsilon$-packing in the empirical norm $\|\cdot\|_S$. Moreover, by (13),

$$\|g_j\|_{S,\infty} \leq R_{\text{out}} \qquad (j = 1, \dots, M).$$

Applying the Bernoulli–Sudakov minoration to the bounded class $\{g_1, \dots, g_M\} \subset H_k$, we obtain

$$\hat{\mathfrak{R}}_S(H_k) \geq c \min\left\{\kappa\varepsilon\sqrt{\frac{\log M}{n}}, \frac{\kappa^2\varepsilon^2}{R_{\text{out}}}\right\},$$

for a universal constant $c > 0$. Taking the supremum over $\varepsilon$ yields (14). □

**Remarks.** (1) The strengthened assumption (12)–(13) is a genuine expressivity condition on the output-layer class $H$: for each hidden map $f \in B_k$, one is allowed to choose an output layer $h_f \in H$ so that the discrete geometry of the word ball $(B_k, d_S)$ is realized inside the real-valued class $H_k$. It is a sufficient condition for the $\sqrt{k/n}$- or $\sqrt{\log k/n}$-type lower bounds in the exponential or polynomial growth regimes, respectively, up to the depth-independent output-layer term in Theorem 1(ii). (2) Only the disjointness of the hidden maps on the sample enters: the word-ball structure of $B_k$ is not used, and the statement holds for an arbitrary subset of $C(\mathcal{X}, \mathcal{X})$.

### E.2 Matching rates

**Corollary 3** (Matching depth dependence under output-visible separation)**.** *Suppose Assumption 4 holds for every $k$ with constants independent of $k$, and fix $\varepsilon_0 > 0$. If $M(B_k, d_S, 2\varepsilon_0) \geq e^{\alpha k}$, then $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c\,\kappa\varepsilon_0\sqrt{\alpha k/n}$ for $n \geq R_{\text{out}}^2\alpha k/(\kappa^2\varepsilon_0^2)$; if $k \geq 1$ and $M(B_k, d_S, 2\varepsilon_0) \geq k^\beta$, then $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c\,\kappa\varepsilon_0\sqrt{\beta \log k/n}$ for $n \geq R_{\text{out}}^2\beta \log k/(\kappa^2\varepsilon_0^2)$. Hence, when the output-layer constants and the diameter in Theorem 1 are depth-independent and the covering numbers obey the matching upper bounds of Proposition 5, the hidden complexity term has the same depth dependence, $\sqrt{k}$ or $\sqrt{\log k}$, in the upper and in the lower bound, up to constants and the depth-independent term $\hat{\mathfrak{R}}_S(H)$.*

**Corollary 4** (Rates under fixed-scale packing lower bounds)**.** *If there exist $\varepsilon_0 > 0$ and $\alpha > 0$ such that*

$$M(B_k, d_S, 2\varepsilon_0) \geq e^{\alpha k},$$

*then*

$$\hat{\mathfrak{R}}_S(H_k) \geq c \min\left\{\kappa\varepsilon_0\sqrt{\frac{\alpha k}{n}}, \frac{\kappa^2\varepsilon_0^2}{R_{\text{out}}}\right\}. \tag{15}$$

*Likewise, if $k \geq 1$ and there exist $\varepsilon_0 > 0$ and $\beta > 0$ such that*

$$M(B_k, d_S, 2\varepsilon_0) \geq k^\beta,$$

*then*

$$\hat{\mathfrak{R}}_S(H_k) \geq c \min\left\{\kappa\varepsilon_0\sqrt{\frac{\beta \log k}{n}}, \frac{\kappa^2\varepsilon_0^2}{R_{\text{out}}}\right\}. \tag{16}$$

*Proof.* (15)–(16) follow by substitution. The first branch of the minimum in (15) is the smaller one exactly when $n \geq R_{\text{out}}^2\alpha k/(\kappa^2\varepsilon_0^2)$, and in (16) exactly when $n \geq R_{\text{out}}^2\beta \log k/(\kappa^2\varepsilon_0^2)$; this gives Corollary 3. If $\kappa$, $R_{\text{out}}$ and $\varepsilon_0$ do not depend on $k$, the resulting lower bounds have the depth dependence $\sqrt{k}$ and $\sqrt{\log k}$, matching cases (iii) and (ii) of Proposition 5. □

### E.3 Output-layer classes that separate hidden states

The class

$$H_R(\Phi) := \big\{h_w(x) = \langle w, \Phi(x)\rangle_{\mathcal{H}} \ : \ \|w\|_{\mathcal{H}} \le R\big\}$$

is natural, but the mere Lipschitz continuity of $\Phi$ is in general sufficient only for the upper-bound side (Theorem 1(ii)), not for the lower-bound side. For the lower bound, one needs either a globally co-Lipschitz scalar observable (Proposition 3), or a finite-set interpolation property on the reachable sample states (Proposition 4 and its corollaries).

**Proposition 3** (Global scalar observable)**.** *Fix a sample $S = (X_i)_{i=1}^n$ and let*

$$U_{k,S} := \{f(X_i) \in X \ : \ f \in B_k,\ i \in [n]\}.$$

*Let $\Phi : X \to \mathcal{H}$ and*

$$H_{R_{\mathrm{H}}}(\Phi) := \big\{h_w(x) = \langle w, \Phi(x)\rangle_{\mathcal{H}} \ : \ \|w\|_{\mathcal{H}} \le R_{\mathrm{H}}\big\}.$$

*Assume that there exists $u \in \mathcal{H}$ with $\|u\|_{\mathcal{H}} \le R_{\mathrm{H}}$ and constants $\kappa, R_\Phi > 0$ such that*

$$|\langle u, \Phi(x) - \Phi(y)\rangle_{\mathcal{H}}| \ge \kappa\, d(x,y) \qquad (x, y \in U_{k,S}), \tag{17}$$

$$\sup_{x \in X} \|\Phi(x)\|_{\mathcal{H}} \le R_\Phi. \tag{18}$$

*Then the lower bound holds with the single choice*

$$\Psi_k(f) := h_u \circ f, \qquad h_u(x) := \langle u, \Phi(x)\rangle_{\mathcal{H}},$$

*namely*

$$\|\Psi_k(f) - \Psi_k(g)\|_S \ge \kappa\, d_S(f,g) \qquad (f, g \in B_k),$$

*and*

$$\|\Psi_k(f)\|_{S,\infty} \le R_{\mathrm{H}} R_\Phi \qquad (f \in B_k).$$

*Proof.* For $f, g \in B_k$,

$$\|\Psi_k(f) - \Psi_k(g)\|_S^2 = \frac{1}{n}\sum_{i=1}^n |\langle u, \Phi(f(X_i)) - \Phi(g(X_i))\rangle|^2 \ge \frac{1}{n}\sum_{i=1}^n \kappa^2 d(f(X_i), g(X_i))^2 = \kappa^2 d_S(f,g)^2,$$

by (17). Also,

$$|\Psi_k(f)(x)| = |\langle u, \Phi(f(x))\rangle| \le \|u\|\, \|\Phi(f(x))\| \le R_{\mathrm{H}} R_\Phi,$$

by (18). □

Proposition 3 is the cleanest way in which a linear Hilbert-space output layer satisfies the *full* lower-Lipschitz assumption. Geometrically, it says that one scalar coordinate of the feature map already observes the hidden-state geometry without collapse.

**Proposition 4** (Map-dependent finite-set interpolation criterion for linear heads)**.** *Fix $f_1, \dots, f_M \in B_k$, and write*

$$z_{j,i} := f_j(X_i) \qquad (j \in [M],\ i \in [n]).$$

*Let $\Phi : X \to \mathcal{H}$, and define*

$$H_{R_{\mathrm{H}}}(\Phi) := \big\{h_w(x) = \langle w, \Phi(x)\rangle_{\mathcal{H}} \ : \ \|w\|_{\mathcal{H}} \le R_{\mathrm{H}}\big\}.$$

*For each $j$, assume that the evaluation operator*

$$T_j : \mathcal{H} \to \mathbb{R}^n, \qquad T_j w := \big(\langle w, \Phi(z_{j,i})\rangle_{\mathcal{H}}\big)_{i \in [n]},$$

*admits a right inverse $R_j : \mathbb{R}^n \to \mathcal{H}$ such that*

$$T_j \circ R_j = \mathrm{Id}_{\mathbb{R}^n}, \qquad \|R_j\|_{\ell_2^n \to \mathcal{H}} \le \Lambda_j$$

*with $\Lambda_j \ge 0$ (for $n \ge 2$ this forces $z_{j,1}, \dots, z_{j,n}$ to be distinct, so no separate distinctness assumption is needed). Then, for every family of target code vectors*

$$u^{(1)}, \dots, u^{(M)} \in \mathbb{R}^n,$$

*there exist output layers $h_j \in H_{R_{\mathrm{H}}}(\Phi)$, $j \in [M]$, provided*

$$R_{\mathrm{H}} \geq \max_{j\in[M]} \Lambda_j \|u^{(j)}\|_2,$$

*such that*

$$h_j(f_j(X_i)) = u_i^{(j)} \qquad (j \in [M],\ i \in [n]).$$

*Consequently, if*

$$\Big(\frac{1}{n}\sum_{i=1}^{n} |u_i^{(j)} - u_i^{(\ell)}|^2\Big)^{1/2} \geq \rho \qquad (j \neq \ell),$$

*then*

$$\|h_j \circ f_j - h_\ell \circ f_\ell\|_S \geq \rho \qquad (j \neq \ell).$$

*If additionally $\max_{j,i} |u_i^{(j)}| \leq R_{\mathrm{out}}$, then*

$$\|h_j \circ f_j\|_{S,\infty} \leq R_{\mathrm{out}} \qquad (j \in [M]).$$

*In particular, for $R_{\mathrm{out}}$-bounded codes it is enough to take*

$$R_{\mathrm{H}} \geq \sqrt{n}\, R_{\mathrm{out}} \max_{j\in[M]} \Lambda_j,$$

*which is independent of the packing cardinality $M$.*

*Proof.* For each fixed $j$, set $w_j := R_j u^{(j)}$ and $h_j := h_{w_j}$. Then $T_j w_j = u^{(j)}$, hence

$$h_{w_j}(f_j(X_i)) = \langle w_j, \Phi(f_j(X_i))\rangle = u_i^{(j)}.$$

Moreover,

$$\|w_j\|_{\mathcal{H}} \leq \Lambda_j \|u^{(j)}\|_2,$$

so $h_{w_j} \in H_{R_{\mathrm{H}}}(\Phi)$ under the stated bound on $R_{\mathrm{H}}$. The separation and sample-boundedness claims follow by evaluating the constructed functions on the sample:

$$(h_j \circ f_j)(X_i) = u_i^{(j)}.$$

□

**Corollary 5** (Finite-dimensional feature map)**.** *Assume $\mathcal{H} = \mathbb{R}^m$ and $n \leq m$. If, for every $j$, the feature matrix*

$$\Phi_j := \big(\Phi(z_{j,i})\big)_{i\in[n]} \in \mathbb{R}^{m\times n}$$

*has full column rank (which requires the $z_{j,i}$ to be distinct), then Proposition 4 applies with*

$$\Lambda_j = \sigma_{\min}(\Phi_j)^{-1}.$$

*Hence a sufficiently wide linear last layer over the feature map $\Phi$ can realize any prescribed finite coding on each individual reachable sample set $\{z_{j,i}\}_{i\in[n]}$.*

*Proof.* Here $T_j w = \Phi_j^\top w$. Since $\Phi_j$ has full column rank, $T_j$ is surjective, and one may take the Moore–Penrose right inverse. Its operator norm is $\sigma_{\min}(\Phi_j)^{-1}$. □

**Corollary 6** (RKHS / kernel output layer)**.** *Let $K$ be a strictly positive definite kernel on $X$, with canonical feature map $\Phi : X \to \mathcal{H}_K$. Then Proposition 4 applies to*

$$H_{R_{\mathrm{H}}}(K) := \big\{h_w(x) = \langle w, \Phi(x)\rangle_{\mathcal{H}_K} \ :\ \|w\|_{\mathcal{H}_K} \leq R_{\mathrm{H}}\big\},$$

*and, provided the points $z_{j,1}, \ldots, z_{j,n}$ are distinct (so that the Gram matrix below is positive definite), one may take*

$$\Lambda_j = \lambda_{\min}(G_j)^{-1/2}, \qquad G_j := \big(K(z_{j,i}, z_{j,i'})\big)_{i,i'\in[n]}.$$

*Therefore, whenever the individual Gram matrices $G_j$ are uniformly well-conditioned, the RKHS linear output layer has the map-dependent finite-coding property needed for the packing-wise Sudakov lower bound. The output-layer norm radius scales with $\max_j \|u^{(j)}\|_2$, not with $(\sum_j \|u^{(j)}\|_2^2)^{1/2}$.*

*Proof.* Fix $j$. For $u \in \mathbb{R}^n$, let $c = G_j^{-1}u$ and set

$$w := \sum_{i=1}^{n} c_i \, \Phi(z_{j,i}).$$

Then

$$h_w(z_{j,i}) = u_i \qquad (i \in [n]),$$

and

$$\|w\|_{\mathcal{H}_K}^2 = u^\top G_j^{-1} u \le \lambda_{\min}(G_j)^{-1} \|u\|_2^2.$$

Thus $\|R_j\|_{\ell_2^n \to \mathcal{H}_K} \le \lambda_{\min}(G_j)^{-1/2}$. □

Corollary 6 gives a realistic interpretation of the lower-bound assumption: a linear output layer over a feature map $\Phi$ works if each reachable sample set is sufficiently "feature-separated" so that the corresponding evaluation map is well-conditioned. In finite dimension, this is a rank/singular-value condition on each last-layer feature matrix; in an RKHS, it is a lower bound on the smallest eigenvalue of each individual Gram matrix.

### E.4 When the converse is silent

The converse is a conditional diagnostic and is not used in the upper bounds. If the architecture collapses or over-smooths its hidden representations, Assumption 4 fails and the lower bound becomes weak or vacuous: a packing lower bound can only certify the growth of the quotient of the reachable states by the relation $x \sim y$ if $h(x) = h(y)$ for all $h \in H$, which may be much smaller than the packing in the state metric. Near collapse the same effect appears through the constants, since $\kappa$ becomes small and, for linear or RKHS output layers, the norm needed to realize a separating code grows with the conditioning of the reachable feature matrices. The upper-bound analysis still applies in these cases, possibly conservatively.

## F Growth mechanisms: conditions, proofs, and examples

This appendix states and proves the growth conditions summarized in Section 4, one subsection per mechanism: saturation (P1, P1'), polynomial growth (P2), exponential growth under separation (E1, E1', E2), and memory-preserving expansion (E3). Each subsection follows the same order: the condition and its proof, one worked example that verifies it, and closing remarks. The conventions and the basic covering-number facts are those of Section A; in this appendix $B(k, F)$ is called the word ball of radius $k$ of the semigroup generated by $F$.

### F.1 Saturation under contraction and equicontinuity (P1, P1')

The two conditions below make the covering number of $B(k, F)$ bounded in $k$ at every resolution. The first uses compactness of the state space and equicontinuity of the compositions; the second uses uniform contraction toward an invariant set and applies on unbounded state spaces. Both are used in Proposition 5(i) and in the examples of Sections J to M.

**Condition P1** (Equicontinuous semigroups on compact domains saturate)**.** *Assume $\mathcal{X}$ is compact. Suppose (at least) one of the following assumptions is satisfied:*

1. *Semigroup $\langle F \rangle$ is (pre)compact,*

2a. *Semigroup $\langle F \rangle$ is equicontinuous,*

2b. *Semigroup $\langle F \rangle$ is uniformly Lipschitz:* $\mathrm{Lip}\langle F \rangle < \infty$*, or*

2c. *Generators $F$ are non-expanding:* $\mathrm{Lip}\, F \le 1$.

*Then, closure semigroup $G := \overline{\langle F \rangle}^{d_\infty} \subset C(\mathcal{X}, \mathcal{X})$ is compact and equicontinuous, and for all $\varepsilon > 0$ and all $k$*

$$N\left(B(k, F), \varepsilon\right) \le N\left(G, \varepsilon\right) \quad (< \infty),$$

*hence no dependence on $k$. Under assumption 1 the compactness of $\mathcal{X}$ is not needed: total boundedness of $\langle F\rangle$ alone gives $N(B(k,F),\varepsilon) \le N(\langle F\rangle,\varepsilon) < \infty$ for all $k$.*

*Proof.* If $\mathcal{X}$ is compact, then $C(\mathcal{X},\mathcal{X})$ is complete in $d_\infty$. Thus the closure of any precompact subset is compact. By the compact self-map version of Ascoli's theorem (see the remark on the uniform metric above), a family in $C(\mathcal{X},\mathcal{X})$ is precompact in $d_\infty$ if and only if it is equicontinuous. Assumption 1 gives compactness of $G$ directly, and Assumption 2a gives it by Arzelà–Ascoli. Assumptions 2b and 2c imply 2a, since a uniform Lipschitz bound is a common modulus of continuity and non-expanding generators generate a non-expanding semigroup. Trivial inclusion $B(k,F) \subseteq G$ yields the bound, and compactness of $G$ gives finiteness of $N(G,\varepsilon)$. □

**Condition P1'** (Contraction to an invariant set)**.** *Assume*

1. *(uniform contraction)* $\sup_{f\in F} \mathrm{Lip}(f) \le c$ *for some* $0 < c < 1$,
2. *(invariant set) there exists a nonempty $F$-invariant set $A \subset \mathcal{X}$ (i.e., $f(A) \subset A$ for all $f \in F$), and*
3. *(bounded absorbing set) there exist $L \in \mathbb{N}$ and a bounded set $K \subset \mathcal{X}$ such that $f(\mathcal{X}) \subset K$ for all depth-$L$ maps $f \in F^L$.*

*Then for every $\varepsilon > 0$, with*

$$m(\varepsilon) := L + \left\lceil \log_{1/c}\left(\frac{2\,\mathrm{diam}(K)}{\varepsilon}\right)\right\rceil,$$

*for* every $k \ge 0$*,*

$$N\left(B(k,F), d_\infty, \varepsilon\right) \le N\left(A, d, \tfrac{\varepsilon}{2}\right) + \sum_{\ell=0}^{m(\varepsilon)-1} N\left(F^\ell, d_\infty, \varepsilon\right),$$

*an inequality in $\mathbb{N}\cup\{\infty\}$ whose right-hand side is finite as soon as $A$ is totally bounded (e.g. compact) and each $N(F^\ell, d_\infty, \varepsilon)$, $\ell < m(\varepsilon)$, is finite (for instance when $F$ is totally bounded in $d_\infty$, by sub-multiplicativity). In particular, the right-hand side is* independent of $k$*, so growth in $k$ saturates at any fixed $\varepsilon$ to the entropy of the invariant set $A$.*

*Proof.* Pick any $a_0 \in A$. For any word $w \in F^\ell$, we have $\mathrm{Lip}(w) \le c^\ell$ and $w(a_0) \in A$ (invariance). If $\ell \ge m(\varepsilon)$, then

$$\sup_{x\in\mathcal{X}} d\left(w(x), w(a_0)\right) \le c^{\ell-L}\,\mathrm{diam}(K), \quad \text{(bounded absorbing set)}$$

and the right-hand side is $\le \varepsilon/2$, i.e.

$$d_\infty\left(w, \mathrm{const}_{w(a_0)}\right) \le \varepsilon/2.$$

Hence all sufficiently deep words, i.e. $\bigcup_{\ell\ge m(\varepsilon)} F^\ell$, are covered by the $\varepsilon/2$-thickening of the set of constant maps landing in $A$: $\mathrm{const}_q : q \in A$. These constants at resolution $\varepsilon/2$ are parameterized by an $\varepsilon/2$-net of $A$, giving the bound $N(A,\varepsilon/2)$. For the finitely many shallow layers $\ell < m(\varepsilon)$, simply cover each $F^\ell$ at scale $\varepsilon$ and sum over $\ell$. These terms are independent of $k$; in the usual compact-generator setting they are finite by continuity of composition and compactness/totally boundedness of $F$. Since $B(k,F) = \bigcup_{\ell\le k} F^\ell$, the two covers together cover $B(k,F)$ for every $k \ge 0$ (for $k < m(\varepsilon)$ only the shallow part is used), which proves the bound for all $k$. Neither compactness of $A$ nor of $K$ was used, only invariance and boundedness. □

*Example* 3 (Cantor attractor (compact, free, contractive)). Let $\mathcal{X} = [0,1]$ and $F = f_0(x) = x/3, f_1(x) = (x+2)/3$. Then $\sup \mathrm{Lip}(f_i) = 1/3 = c < 1$. The attractor $A$ is the middle–third Cantor set $C$ (compact, $F$-invariant). By Condition P1' (with $L = 0$, $K = [0,1]$), for every $\varepsilon > 0$ and every $k$,

$$N\left(B(k,F),\varepsilon\right) \le N\left(C, \varepsilon/2\right) + \sum_{\ell < m(\varepsilon)} N(F^\ell,\varepsilon), \qquad m(\varepsilon) = \lceil \log_3(2/\varepsilon)\rceil,$$

independent of $k$. (Indeed, every $w \in F^\ell$ has the form $w(x) = 3^{-\ell}x + b_w$ and hence is $\varepsilon/2$-close to the constant map $x \mapsto b_w \in C$.)

**Remarks.** (1) When $\mathcal{X}$ is compact, one may take $A = K = \mathcal{X}$ and $L = 0$ in Condition P1'. (2) A single saturating map $\sigma$, such as $\tanh$, the logistic function, or clipping, provides the bounded absorbing set $K = \sigma(\mathcal{X})$ with $L = 1$. (3) Deep words are *almost constant* (their images have diameter $\leq c^{\ell-L}\operatorname{diam}(K)$),so at scale $\varepsilon$ only the landing point $w(a_0) \in A$ matters; hence the dependence on $k$ disappears once $\ell$ exceeds an $\varepsilon$-dependent "memory length" $m(\varepsilon)$.

### F.2 Polynomial growth under nilpotent control (P2)

The next condition covers layer classes that are not contractive but whose compositions are organized by a group of polynomial growth. It gives the fixed-scale bound $N(B(k,F),d_\infty,\varepsilon) \lesssim (1+k/\varepsilon)^D$ and the diameter bound $D_k(S) \lesssim k$ used in Proposition 5(ii) and (iv).

**Condition P2** (Nilpotent control gives polynomial growth)**.** *Let $(H, d_H)$ be a compactly generated group with identity $e$, and assume the length $g \mapsto d_H(e,g)$ is subadditive: $d_H(e,gh) \leq d_H(e,g) + d_H(e,h)$. Assume that its metric balls have polynomial entropy of degree $D \geq 0$: there is $1 \leq C_H < \infty$ such that for all $R \geq 0$ and $\delta > 0$,*

$$N\big(B_H(e,R), d_H, \delta\big) \leq C_H \left(1 + \frac{R}{\delta}\right)^D$$

*(in particular these covering numbers are finite). Suppose $H$ acts on $\mathcal{X}$ through a homomorphism $\alpha : H \to C(\mathcal{X},\mathcal{X})$, and suppose there is a bounded set $S \subset H$ such that $F \subset \alpha(S)$. Finally assume the orbit map is Lipschitz in the uniform metric: there is $0 < L_\alpha < \infty$ such that for all $g, h \in H$,*

$$d_\infty\big(\alpha(g), \alpha(h)\big) \leq L_\alpha d_H(g,h).$$

*Then, with $R_S := \sup_{s\in S} d_H(e,s)$ and $C := C_H \max(1, R_S L_\alpha)^D$, for every $\varepsilon > 0$ and every $k \geq 0$,*

$$N\left(B(k,F), d_\infty, \varepsilon\right) \leq C \left(1 + \frac{k}{\varepsilon}\right)^D,$$

*and moreover $\operatorname{diam}_\infty B(k,F) \leq 2L_\alpha R_S k$, hence $D_k(S) \leq 2L_\alpha R_S k$ for every sample $S$.*

*Proof.* Let $R_S := \sup_{s\in S} d_H(e,s) < \infty$. Any composition of at most $k$ maps from $F$ has a representative $\alpha(g)$ with $g = s_k \cdots s_1$ and $s_i \in S$. By subadditivity,

$$d_H(e,g) \leq \sum_{i=1}^{k} d_H(e,s_i) \leq kR_S.$$

Thus

$$B(k,F) \subset \alpha\big(B_H(e, kR_S)\big).$$

The $L_\alpha$-Lipschitz bound on $\alpha$ gives

$$N\left(B(k,F), d_\infty, \varepsilon\right) \leq N\left(B_H(e,kR_S), d_H, \varepsilon/L_\alpha\right).$$

The assumed polynomial entropy bound on $H$ gives $C_H(1 + kR_S L_\alpha/\varepsilon)^D \leq C_H \max(1, R_S L_\alpha)^D (1 + k/\varepsilon)^D$, which is the claim (for $k = 0$ as well). The diameter bound is proved in remark (2) below. □

*Example* 4 (Upper-triangular unipotent group (non-compact, nilpotent)). Let

$$H = UT_d(\mathbb{R}) = \{I_d + N : \ N \text{ is strictly upper triangular}\}$$

with a right-invariant Riemannian metric, and let $\mathcal{X} = H$. Let $\alpha(g)$ be left translation, $\alpha(g)(x) = gx$. Then

$$d_\infty(\alpha(g), \alpha(h)) = d_H(g,h),$$

because $d_H$ is right-invariant. For any compact $S \subset H$ and $F = \alpha(S)$, Condition P2 gives

$$N(B(k,F), d_\infty, \varepsilon) \lesssim (1 + k/\varepsilon)^D, \qquad D = \frac{d(d-1)(d+1)}{6}.$$

Each generator $\alpha(g)$, $x \mapsto gx$, is a linear map of the ambient matrix space $\mathbb{R}^{d\times d}$, so by Proposition 6(c) every word in $B(k,F)$ is computed exactly by a ReLU network of depth $k$ and width $2d^2$; no approximation of the transitions, and hence no exponential width, is needed to implement this example.

**Remarks.** (1) The Guivarc'h–Bass homogeneous dimension bridges nilpotency with polynomial volume and entropy growth. Connected simply connected nilpotent Lie groups with left-invariant Riemannian metrics, and finitely generated nilpotent groups with word metrics, satisfy the displayed entropy assumption with this dimension. See Breuillard (2014) and de la Harpe (2000, Chapter VII) for the growth theory of nilpotent groups used here. (2) The inclusion $B(k,F) \subset \alpha(B_H(e, kR_S))$ in the proof also controls the diameter of the word ball: for $g, h \in B_H(e, kR_S)$, $d_\infty(\alpha(g), \alpha(h)) \le L_\alpha d_H(g,h) \le 2L_\alpha R_S k$, so $\operatorname{diam}_\infty(B(k,F)) \le 2L_\alpha R_S k$ and a fortiori $D_k(S) \le 2L_\alpha R_S k$. On a compact state space the diameter is of course bounded by $\operatorname{diam}(\mathcal{X})$. These layerwise bounds are the input to Proposition 5(ii) and (iv).

### F.3 Exponential growth under separation (E1, E1', E2)

The conditions below give the lower bound $N(B(k,F), d_\infty, \varepsilon) \ge r^k$ for a class of $r$ generators at a fixed resolution. All of them are proved with Lemma 7: one exhibits, for each depth, $r^k$ compositions whose values at suitable probe points are uniformly separated. E1 and E1' use one probe point for all words; E2 uses one probe per word and applies on compact state spaces where the maps expand on part of the space and reset elsewhere. With a bounded diameter these conditions give the $\sqrt{k}$ profile of Proposition 5(iii), and Corollary 3 shows that the profile is attained when the separation is visible at the output layer.

**Condition E1** (Free semigroup with one-point uniform separation)**.** *Let $F = \{f_1, \dots, f_r\} \subset C(\mathcal{X}, \mathcal{X})$ with $r \ge 2$. Assume:*

1. *(Uniform separation at a base point) There exist $x_* \in \mathcal{X}$ and $\delta > 0$ such that for all $k$ and all distinct words $u, v \in [r]^k$, writing $f_u := f_{i_k} \circ \cdots \circ f_{i_1}$ for $u = (i_1, \dots, i_k)$, $d\big(f_u(x_*), f_v(x_*)\big) \ge \delta$.*

*Then, for every $k$ and every $\varepsilon < \delta/2$,*

$$N\big(B(k,F), d_\infty, \varepsilon\big) \ge r^k.$$

*In particular $u \mapsto f_u$ is injective on $[r]^k$ for every $k$ (freeness at each length), so no separate freeness hypothesis is needed.*

*Proof.* Apply Lemma 7 with $P = \{x_*\}$. For each fixed $k$, the $r^k$ words of length $k$ give $r^k$ maps whose images at $x_*$ are $\delta$-separated; hence for $\varepsilon < \delta/2$ at least $r^k$ balls are needed to cover $B(k,F)$. □

**Condition E1'** (Uniform same-length coding)**.** *Let $F = \{f_1, \dots, f_r\} \subset C(\mathcal{X}, \mathcal{X})$ with $r \ge 2$, and suppose (uniform same-length separation) that there exist $x_* \in \mathcal{X}$ and $c > 0$ such that, for all words $u \ne v$ of the same length,*

$$d\big(u(x_*), v(x_*)\big) \ge c.$$

*Then, for every $k$ and every $\varepsilon < c/2$,*

$$N\big(B(k,F), d_\infty, \varepsilon\big) \ge r^k.$$

*This is Condition E1 with $\delta = c$: neither isometry nor freeness nor finiteness of $d_\infty$ on $\langle F \rangle$ is needed for the lower bound. The isometric setting in which it is used below, $F \subset \mathrm{Iso}(\mathcal{X})$ with $d_\infty(g,h) < \infty$ for $g, h \in \langle F \rangle$ (as when $d$ is bi-invariant, $d(axb, ayb) = d(x,y)$), only serves to make the uniform metric finite on the semigroup.*

*Proof.* Apply Lemma 7 with $P = \{x_*\}$, exactly as for Condition E1. □

*Example* 5 (Free group with a bi-invariant metric (non-compact, free, isometric))*.* Let $\mathcal{X} = G = F_r$ $(r \ge 2)$. Let

$$S := \big\{ x a_i x^{-1}, x a_i^{-1} x^{-1} \;\big|\; x \in F_r, i = 1, \dots, r \big\}$$

and define the *conjugacy-invariant* word metric

$$d_S(g,h) := |g^{-1}h|_S,$$

the shortest length in the alphabet $S$. This metric is *bi-invariant*, hence all left translations $L_g$ are isometries and, crucially,

$$d_\infty(L_g, L_h) = \sup_x d_S(gx, hx) = \sup_x d_S\big(e, x^{-1}g^{-1}hx\big) = d_S(e, g^{-1}h) < \infty$$

(the supremum is independent of $x$).

Let $F = \{L_{a_1}, \dots, L_{a_r}\}$. The positive semigroup is free. With $x_* = e$, for words $u \neq v$ of the same length,

$$d_S\big(L_u(x_*), L_v(x_*)\big) = d_S(u, v) = |u^{-1}v|_S \geq 1,$$

so Condition E1' applies with $c = 1$: for every $\varepsilon < 1/2$,

$$N\big(B(k, F), d_\infty, \varepsilon\big) \geq r^k.$$

With the standard word metric instead, $d_\infty(L_g, L_h) = \infty$ for $g \neq h$; the lower bound would still hold, since only the probe $x_* = e$ is used, but the bi-invariant metric keeps $d_\infty$ finite.

**Ping-pong coding.** The classical *Ping-Pong* (*Table-Tennis*, or *Schottky*) lemma (e.g. de la Harpe, 2000, Chapter II.B) is a sufficient condition for freeness of generators, which itself does not immediately imply the exponential growth since the generators might be contractive. Therefore, the condition below states the uniform coding separation used in the entropy argument directly; the examples verify it through expanding branches and resets.

Here a *chamber* $U_i \subset \mathcal{X}$ is the active region for the generator $f_i$. A *coding core* $V_i \subset U_i$ is a smaller nonempty set on which $f_i$ has enough coverage to realize the next symbol. The *anchors* form a finite set $A = \{a_1, \dots, a_r\}$ of reset states; the points $a_i$ need not be pairwise distinct unless an example says so. A *marker* $q \in \mathcal{X}$ is the target state for correctly coded trajectories, and it is required to be separated from the anchor set. A *probe* $x_u \in \mathcal{X}$ is an input point chosen for a word $u$; it is constructed from the chambers and is not fixed across all words.

**Condition E2** (Ping-pong coding)**.** *Let $F = \{f_1, \dots, f_r\} \subset C(\mathcal{X}, \mathcal{X})$ with $r \geq 2$. Assume that there exist sets $V_i \subset U_i \subset \mathcal{X}$, anchors $a_i \in \mathcal{X}$, a marker $q \in \mathcal{X}$, and a constant $\alpha > 0$ such that, with $A := \{a_1, \dots, a_r\}$ and*

$$Q := \{q\} \cup \bigcup_{j=1}^{r} V_j,$$

*the following conventions and conditions hold. The marker $q$ is not an anchor, quantitatively enforced by item 4 below. No global location assumption such as $A \subset \bigcup_i U_i$ or $A \cap \bigcup_i U_i = \emptyset$ is imposed; anchors may lie inside or outside chambers. The required anchor condition is the invariance in item 3.*

1. *(Disjoint chambers)*
$$U_i \cap U_j = \emptyset \qquad (i \neq j).$$
2. *(Full coding cores) For every $i$,*
$$Q \subset f_i(V_i)$$
*(in particular each $V_i$ is nonempty, since $q \in f_i(V_i)$).*
3. *(Reset and anchor invariance) For every $i$,*
$$f_i(\mathcal{X} \setminus U_i) = \{a_i\}, \qquad f_i(A) \subset A.$$
4. *(Marker–anchor separation)*
$$\operatorname{dist}(\{q\}, A) \geq \alpha.$$

*Then, for every $k$ and every $\varepsilon < \alpha/2$,*

$$N\big(B(k, F), d_\infty, \varepsilon\big) \geq r^k.$$

*Moreover, the map $u \mapsto f_u$ is injective on words of each fixed length.*

*Proof.* Fix $k$ and a word $u = i_k \cdots i_1 \in [r]^k$. Set $z_k = q$. Since $Q \subset f_{i_s}(V_{i_s})$, we may choose recursively $z_{s-1} \in V_{i_s}$ such that

$$f_{i_s}(z_{s-1}) = z_s, \qquad s = k, k-1, \dots, 1.$$

Put $x_u := z_0$. Then $f_u(x_u) = q$. Now let $v = j_k \cdots j_1 \neq u$, and let $t$ be the first index, in the order of application, with $j_t \neq i_t$. The first $t-1$ steps agree, so the $v$-trajectory is at $z_{t-1} \in V_{i_t} \subset U_{i_t}$. Since the chambers are disjoint, $z_{t-1} \notin U_{j_t}$. Hence $f_{j_t}$ resets this trajectory to $a_{j_t} \in A$, and all later generators keep it in $A$ by anchor invariance. Therefore

$$d\big(f_u(x_u), f_v(x_u)\big) \geq \operatorname{dist}(\{q\}, A) \geq \alpha.$$

Thus, for $P_k := \{x_u \mid u \in [r]^k\}$, the $r^k$ evaluation vectors $\{\mathrm{ev}_{P_k}(f_u) \mid u \in [r]^k\}$ are pairwise $\alpha$-separated. The claim follows from Lemma 7. □

*Example* 6 (Ping-pong over the full shift (compact, expansive)). Fix an alphabet $\mathcal{A} = [r] \cup \{\bullet\}$ with $r \geq 2$ *active* symbols $[r]$ and one extra *padding* symbol $\bullet$. Let

$$\mathcal{X} = \mathcal{A}^{\mathbb{N}} = \{x = (x_0, x_1, x_2, \dots) \mid x_j \in \mathcal{A}\}$$

be the (one-sided) full shift. Equip $\mathcal{X}$ with the standard ultrametric $d_\theta$ for some fixed $\theta \in (0,1)$:

$$d_\theta(x,y) = \begin{cases} 0 & x = y \\ \theta^{\min\{n \geq 0 \mid x_n \neq y_n\}} & x \neq y \end{cases}$$

Then $(\mathcal{X}, d_\theta)$ is compact, totally disconnected, and the left shift $\sigma(x)_n = x_{n+1}$ is $L$-Lipschitz with $L = \theta^{-1} > 1$ (hence expansive).

For each $a \in \mathcal{A}$, write the clopen 1-cylinder $[\![a]\!] = \{x \in \mathcal{X} \mid x_0 = a\}$, and the *anchor* sequence $\bar{a} = (a, a, a, \dots)$.

Define $F = \{f_1, \dots, f_r\} \subset C(\mathcal{X}, \mathcal{X})$ by

$$f_a(x) = \begin{cases} \sigma(x) & x \in [\![a]\!] \\ \bar{a} & x \notin [\![a]\!] \end{cases} \qquad a \in [r].$$

Because $[\![a]\!]$ is clopen, each $f_a$ is continuous; on $[\![a]\!]$ it is $\sigma$ (Lipschitz constant $\theta^{-1} > 1$), and on $\mathcal{X} \setminus [\![a]\!]$ it is constant. Thus each $f_a$ has an expanding branch, although it is not globally expanding because of the reset branch.

Let $U_a := V_a := [\![a]\!]$ (clopen chambers and coding cores), $a \in [r]$, let $A := \{\bar{a} \mid a \in [r]\}$, and let $q := \bar{\bullet}$. Here $q \notin \bigcup_{a \in [r]} U_a$, while each anchor $\bar{a}$ lies in its corresponding chamber $U_a$. Then:

- Pairwise separation. If $a \neq b$, then for any $x \in U_a$, $y \in U_b$, $d_\theta(x,y) = \theta^0 = 1$. Hence $\operatorname{dist}(U_a, U_b) = 1$.
- Expansion/coverage on own domain. For each $a$, $f_a|_{U_a} = \sigma$ maps $U_a$ bijectively onto $\mathcal{X}$ (surjective and expanding).
- Reset off domain. On $\mathcal{X} \setminus U_a$, $f_a = \mathrm{const}_{\bar{a}}$ (exact reset; the "reset diameter" is 0).
- Anchor invariance and marker separation. For every $a \in [r]$, $f_a(A) \subset A$, and $\operatorname{dist}(\{q\}, A) = 1$.

These data yield Condition E2 with $\alpha = 1$ (the chambers are even at distance 1). Hence the separation scale can be taken as $\delta = 1$.

Let a word $u = a_k \cdots a_1 \in [r]^k$. Define the *tail-padded probe*

$$x_u := \big(a_1, a_2, \dots, a_k, \underbrace{\bullet, \bullet, \bullet, \dots}_{\text{all } \bullet}\big) \in \mathcal{X}.$$

Then, by construction,

$$f_u(x_u) = \sigma^k(x_u) = \bar{\bullet},$$

because each of the first $k$ steps sees the correct chamber and applies $\sigma$, peeling off the $k$-letter prefix and revealing the all-$\bullet$ tail.

If $v = b_k \cdots b_1 \neq u$, let $t$ be the *rightmost* index with $a_t \neq b_t$. When applying $f_v$ to $x_u$, the first $t-1$ letters match and act as shifts; at step $t$ we apply $f_{b_t}$ to a sequence whose 0-coordinate is $a_t \neq b_t$, so $f_{b_t}$ resets to the anchor $\bar{b_t}$. From then on, the state never contains $\bullet$ at the 0-coordinate (subsequent resets only use anchors $\bar{b_s}$ with $b_s \in [r]$, and $\sigma$ preserves 0-coordinate $b_s$ on the constant sequence $\bar{b_s}$). Consequently

$$f_v(x_u) \neq \bar{\bullet}.$$

Therefore

$$d_\theta\big(f_u(x_u), f_v(x_u)\big) = d_\theta\big(\bar{\bullet}, f_v(x_u)\big) = 1.$$

This shows *freeness*: distinct words $u \neq v$ define distinct maps $f_u \neq f_v$.

Fix $k \in \mathbb{N}$ and consider the finite probe set

$$P_k := \{x_u \mid u \in [r]^k\} \subset \mathcal{X},$$

of cardinality $r^k$. For any distinct $u, v$ we have just seen that at the coordinate $x_u$,

$$d_\theta\big(f_u(x_u), f_v(x_u)\big) = 1 \geq \delta.$$

Thus the $r^k$ vectors $E_{P_k}(f_u)_{|u|=k} \subset \mathcal{X}^{P_k}$ are pairwise 1-separated in the max metric. By Lemma 7, for every $\varepsilon < \frac{1}{2}$,

$$N\big(B(k, F), d_\infty, \varepsilon\big) \geq r^k.$$

**Remarks.** Only the disjointness of the chambers is used in the exact-reset proof. The interval and symbolic examples below satisfy the stronger positive separation $\operatorname{dist}(U_i, U_j) \geq \Delta > 0$, which is the robust form needed if one allows approximate resets or interpolation collars.

### F.4 Memory-preserving expansion (E3)

Expansion alone does not produce entropy growth beyond $r^k$, because later layers may erase earlier distinctions. The construction below adds a memory: each layer writes a symbol, shifts the stored symbols, and expands them. The entropy then accumulates multiplicatively across layers, and the two corollaries give $\log N \asymp k^2$ and $\log N \asymp \lambda^{pk}$; these are the two cases in which the layerwise bound of Theorem 2 is attained (Section G).

**Condition E3** (Memory-preserving expansion)**.** *Let $E$ be a normed space, let $G \subset E$ be an arbitrary subset, and fix $\lambda > 1$. Set*

$$X := \ell_\infty(E) = \Big\{x = (x_1, x_2, \dots) : \sup_{j \geq 1} \|x_j\|_E < \infty\Big\},$$

*equipped with the bounded sup metric*

$$d_X(x, y) := \sup_{j \geq 1} \min\{1, \|x_j - y_j\|_E\}.$$

*We write $d_\infty$ for the induced uniform metric on $C(X, X)$. This bounded state metric keeps the maps below at finite uniform distance while preserving the small-scale expansion estimates used in the lower bounds. Define the reset map, the expander, and the writer family by*

$$r(x) := 0, \qquad A(x_1, x_2, \dots) := (\lambda x_1, \lambda x_2, \dots),$$

*and, for each $u \in G$,*

$$g_u(x_1, x_2, \dots) := (u, x_1, x_2, \dots).$$

*Let*

$$F := \{r, A\} \cup \{g_u : u \in G\}.$$

*For each $k \geq 1$ and each $u = (u_1, \dots, u_k) \in G^k$, define the word*

$$w_u := A \circ g_{u_k} \circ A \circ g_{u_{k-1}} \circ \cdots \circ A \circ g_{u_1} \circ r.$$

*Let*

$$W_k := \{w_u : u \in G^k\} \subset B(2k+1, F).$$

*Then the following hold.*

(i) *For every* $x \in X$,

$$w_u(x) = (\lambda u_k, \lambda^2 u_{k-1}, \dots, \lambda^k u_1, 0, 0, \dots).$$

*In particular, each* $w_u$ *is a constant map.*

(ii) *For* $u = (u_1, \dots, u_k)$ *and* $v = (v_1, \dots, v_k)$,

$$d_\infty(w_u, w_v) = \max_{1 \le j \le k} \min\{1, \lambda^{k-j+1} \|u_j - v_j\|_E\}.$$

(iii) *For every* $\varepsilon > 0$,

$$N(W_k, d_\infty, \varepsilon) \le \prod_{j=1}^{k} N\left(G, \|\cdot\|_E, \frac{\varepsilon}{\lambda^{k-j+1}}\right),$$

*and for every* $0 < \varepsilon < 1/2$,

$$\prod_{j=1}^{k} M\left(G, \|\cdot\|_E, \frac{2\varepsilon}{\lambda^{k-j+1}}\right) \le N(W_k, d_\infty, \varepsilon).$$

*Consequently, by monotonicity of the (external) covering number under* $W_k \subset B(2k+1, F)$,

$$N(B(2k+1, F), d_\infty, \varepsilon) \ge N(W_k, d_\infty, \varepsilon),$$

*so the word-ball entropy inherits a multiplicative layer-by-layer growth law.*

*Proof.* We first prove (i). Since $r$ is the constant-zero map, the input $x$ is irrelevant after the first step. Starting from $0 = (0, 0, \dots)$,

$$g_{u_1}(0) = (u_1, 0, 0, \dots), \qquad A g_{u_1}(0) = (\lambda u_1, 0, 0, \dots).$$

Applying $g_{u_2}$ and then $A$ gives

$$A g_{u_2} A g_{u_1}(0) = (\lambda u_2, \lambda^2 u_1, 0, 0, \dots).$$

Iterating this computation yields

$$w_u(x) = (\lambda u_k, \lambda^2 u_{k-1}, \dots, \lambda^k u_1, 0, 0, \dots),$$

which proves (i).

Part (ii) now follows immediately from (i), because $w_u$ and $w_v$ are constant maps:

$$d_\infty(w_u, w_v) = \sup_{x \in X} d_X\big(w_u(x), w_v(x)\big) = \max_{1 \le j \le k} \min\{1, \lambda^{k-j+1} \|u_j - v_j\|_E\}.$$

For (iii), define two metrics on $G^k$:

$$\rho_k(u, v) := \max_{1 \le j \le k} \lambda^{k-j+1} \|u_j - v_j\|_E, \qquad \bar{\rho}_k(u, v) := \max_{1 \le j \le k} \min\{1, \lambda^{k-j+1} \|u_j - v_j\|_E\}.$$

By part (ii), the parameterization $u \mapsto w_u$ is an isometric embedding of $(G^k, \bar{\rho}_k)$ into $(C(X, X), d_\infty)$, so

$$N(W_k, d_\infty, \varepsilon) = N(G^k, \bar{\rho}_k, \varepsilon).$$

For the upper bound, let $C_j$ be an $\varepsilon/\lambda^{k-j+1}$-cover of $G$ in $\|\cdot\|_E$. Then the Cartesian product $C_1 \times \cdots \times C_k$ is an $\varepsilon$-cover of $(G^k, \rho_k)$, because for every $u = (u_1, \dots, u_k)$ one can choose $c_j \in C_j$ with

$$\|u_j - c_j\|_E \le \frac{\varepsilon}{\lambda^{k-j+1}}$$

for all $j$, and therefore

$$\rho_k(u, c) \le \varepsilon,$$

and hence $\bar{\rho}_k(u, c) \le \varepsilon$. Hence

$$N(W_k, d_\infty, \varepsilon) = N(G^k, \bar{\rho}_k, \varepsilon) \le \prod_{j=1}^{k} N\left(G, \|\cdot\|_E, \frac{\varepsilon}{\lambda^{k-j+1}}\right).$$

For the lower bound, let $P_j$ be a maximal $2\varepsilon/\lambda^{k-j+1}$-packing of $G$ in $\|\cdot\|_E$. Then for distinct points

$$u = (u_1, \dots, u_k),\ v = (v_1, \dots, v_k) \in P_1 \times \cdots \times P_k,$$

there exists some $j$ with $u_j \neq v_j$, and for that $j$,

$$\lambda^{k-j+1}\|u_j - v_j\|_E \ge 2\varepsilon.$$

Therefore

$$\rho_k(u, v) \ge 2\varepsilon,$$

and since $2\varepsilon < 1$, also $\bar{\rho}_k(u, v) \ge 2\varepsilon$. Thus $P_1 \times \cdots \times P_k$ is a $2\varepsilon$-packing of $(G^k, \bar{\rho}_k)$. Any $\varepsilon$-cover must contain at least as many points as any $2\varepsilon$-packing, hence

$$N(W_k, d_\infty, \varepsilon) = N(G^k, \bar{\rho}_k, \varepsilon) \ge \prod_{j=1}^{k} M\left(G, \|\cdot\|_E, \frac{2\varepsilon}{\lambda^{k-j+1}}\right).$$

Finally, since $W_k \subset B(2k+1, F)$ and external covering numbers are monotone under inclusion,

$$N(B(2k+1, F), d_\infty, \varepsilon) \ge N(W_k, d_\infty, \varepsilon).$$

This completes the proof. □

**Corollary 7** (Super-exponential growth)**.** *Assume that $N(G, \|\cdot\|_E, \delta) < \infty$ for all $\delta > 0$ (automatic for compact $G$) and that there exist constants $c_-, c_+, \delta_0 > 0$ such that, for all $0 < \delta < \delta_0$,*

$$c_- \log(1/\delta) \le \log M(G, \|\cdot\|_E, \delta) \le \log N(G, \|\cdot\|_E, \delta) \le c_+ \log(1/\delta).$$

*Then for every fixed $\varepsilon$ with $0 < \varepsilon < 1/2$ and $\varepsilon < \lambda\delta_0/2$, and for all $k \ge 1$,*

$$C_1 k^2 \le \log N(W_k, d_\infty, \varepsilon) \le C_2 k^2, \qquad C_1 := \frac{c_- \log \lambda}{2}, \quad C_2 := c_+\big(\log \lambda - \log \varepsilon\big).$$

*Consequently,*

$$\log N(B(2k+1, F), d_\infty, \varepsilon) \ge C_1 k^2,$$

*so the word-ball covering number is at least* $\exp(\Omega(k^2))$.

*Proof.* By Condition E3,

$$\log N(W_k, d_\infty, \varepsilon) \ge \sum_{j=1}^{k} \log M\left(G, \|\cdot\|_E, \frac{2\varepsilon}{\lambda^{k-j+1}}\right),$$

and

$$\log N(W_k, d_\infty, \varepsilon) \le \sum_{j=1}^{k} \log N\left(G, \|\cdot\|_E, \frac{\varepsilon}{\lambda^{k-j+1}}\right).$$

Since $0 < \varepsilon < \lambda\delta_0/2$ and $\lambda > 1$, every radius $2\varepsilon/\lambda^{k-j+1} \le 2\varepsilon/\lambda$ and $\varepsilon/\lambda^{k-j+1}$ lies below $\delta_0$; note that the $j = k$ radius does not shrink with $k$, so the smallness of $\varepsilon$ cannot be traded for largeness of $k$. Moreover $G \neq \emptyset$ by the packing lower bound. Hence the assumed entropy bounds apply to every factor, and with

$$\sum_{j=1}^{k} \log\left(\frac{\lambda^{k-j+1}}{a}\right) = (\log \lambda)\frac{k(k+1)}{2} + k \log(1/a) \qquad (a > 0)$$

we obtain, using $2\varepsilon < 1$ for the lower bound and $k + 1 \le 2k$, $k \le k^2$ for the upper bound,

$$\frac{c_- \log \lambda}{2}\, k^2 \le \log N(W_k, d_\infty, \varepsilon) \le c_+\big(\log \lambda + \log(1/\varepsilon)\big)k^2.$$

□

**Corollary 8** (Double-exponential growth)**.** *Assume that $N(G, \|\cdot\|_E, \delta) < \infty$ for all $\delta > 0$ and that there exist constants $p > 0$, $c_-, c_+, \delta_0 > 0$ such that, for all $0 < \delta < \delta_0$,*

$$c_- \delta^{-p} \le \log M(G, \|\cdot\|_E, \delta) \le \log N(G, \|\cdot\|_E, \delta) \le c_+ \delta^{-p}.$$

*Then for every fixed $\varepsilon$ with $0 < \varepsilon < 1/2$ and $\varepsilon < \lambda\delta_0/2$, and for all $k \ge 1$,*

$$C_1 \lambda^{pk} \le \log N(W_k, d_\infty, \varepsilon) \le C_2 \lambda^{pk}, \qquad C_1 := c_-(2\varepsilon)^{-p}, \quad C_2 := c_+ \varepsilon^{-p} \frac{q}{q-1}, \quad q := \lambda^p.$$

*Consequently,*

$$\log N(B(2k+1, F), d_\infty, \varepsilon) \ge C_1 \lambda^{pk},$$

*i.e. the word-ball covering number is at least double-exponential in $k$.*

*Proof.* Again by Condition E3,

$$\log N(W_k, d_\infty, \varepsilon) \ge \sum_{j=1}^{k} \log M\left(G, \|\cdot\|_E, \frac{2\varepsilon}{\lambda^{k-j+1}}\right),$$

and

$$\log N(W_k, d_\infty, \varepsilon) \le \sum_{j=1}^{k} \log N\left(G, \|\cdot\|_E, \frac{\varepsilon}{\lambda^{k-j+1}}\right).$$

As in the proof of Corollary 7, all radii lie below $\delta_0$ and $G \neq \emptyset$, so the entropy bounds apply to every factor:

$$c_-(2\varepsilon)^{-p} \sum_{r=1}^{k} q^r \le \log N(W_k, d_\infty, \varepsilon) \le c_+ \varepsilon^{-p} \sum_{r=1}^{k} q^r, \qquad q = \lambda^p > 1.$$

Since $q^k \le \sum_{r=1}^{k} q^r = q(q^k - 1)/(q-1) \le \frac{q}{q-1} q^k$, this proves the claim. □

*Example* 7 (Hölder writers on a memory state space). Let $E = C([0,1]^d)$ with the sup norm, let $0 < \alpha \le 1$, and define

$$G := \Big\{ u \in C^\alpha([0,1]^d) : \|u\|_\infty + [u]_{C^\alpha} \le 1 \Big\}.$$

Set $X = \ell_\infty(E)$ and define $r$, $A$, and $g_u$ exactly as in Condition E3. Then there exist constants $c_-, c_+, \delta_0 > 0$ such that, for all $0 < \delta < \delta_0$,

$$c_- \delta^{-d/\alpha} \le \log M(G, \|\cdot\|_\infty, \delta) \le \log N(G, \|\cdot\|_\infty, \delta) \le c_+ \delta^{-d/\alpha}.$$

Consequently, for every fixed $0 < \varepsilon < \min\{1/2, \lambda\delta_0/2\}$,

$$\log N(W_k, d_\infty, \varepsilon) = \Theta\big(\lambda^{(d/\alpha)k}\big).$$

In particular,

$$N(B(2k+1, F), d_\infty, \varepsilon) \ge \exp\Big(\Omega\big(\lambda^{(d/\alpha)k}\big)\Big),$$

so the word-ball covering number is at least double-exponential in the depth parameter $k$.

*Proof.* The displayed entropy estimate is the standard metric entropy estimate for the unit Hölder ball in the sup norm. Apply Corollary 8 with $p = d/\alpha$. □

## G THE LAYERWISE COVERING BOUND

This appendix proves the layerwise covering bound of Section 4, derives the profiles stated after it, and records the elementary bound on the diameter of the reachable states that accompanies it. Throughout, $F \subset C(\mathcal{X}, \mathcal{X})$ consists of $\Lambda$-Lipschitz maps with $\Lambda \ge 0$, and $S_m(\Lambda) := \sum_{i=0}^{m-1} \Lambda^i$ ($S_0 = 0$, $S_1 = 1$) is the geometric sum of Proposition 6(b). Covering numbers are external (Section 2).

### G.1 Statement and proof

**Theorem 9** (Theorem 2, restated)**.** *For every $k \geq 0$ and $\varepsilon > 0$,*

$$N\big(B(k,F), d_\infty, \varepsilon\big) \leq 1 + \sum_{m=1}^{k} N\big(F, d_\infty, \varepsilon/S_m(\Lambda)\big)^m \leq 1 + k\, N\big(F, d_\infty, \varepsilon/S_k(\Lambda)\big)^k, \tag{19}$$

*where the last inequality uses that $S_m(\Lambda) \leq S_k(\Lambda)$ for $m \leq k$ (for $k = 0$ both sides equal 1). The centres of the covers of $F$ need not belong to $F$ nor be Lipschitz; only the Lipschitz constant of the maps in $F$ is used, and $F = \emptyset$ is allowed.*

*Proof.* $B(k,F) = \{\mathrm{id}\} \cup \bigcup_{m=1}^{k} F^m$, where $F^m$ is the set of compositions of exactly $m$ maps of $F$, and covering numbers are subadditive under unions. Fix $1 \leq m \leq k$ and let $\{c_1, \dots, c_M\} \subset C(\mathcal{X}, \mathcal{X})$ be an $\varepsilon/S_m(\Lambda)$-cover of $F$ in $d_\infty$ with $M = N(F, d_\infty, \varepsilon/S_m(\Lambda))$. For $f_m \circ \cdots \circ f_1 \in F^m$ choose indices with $d_\infty(f_j, c_{i_j}) \leq \varepsilon/S_m(\Lambda)$. The telescoping estimate in the proof of Proposition 6(b), applied with the abstract layers $f_j$ and the implemented layers $c_{i_j}$, gives

$$d_\infty\big(f_m \circ \cdots \circ f_1,\ c_{i_m} \circ \cdots \circ c_{i_1}\big) \leq \frac{\varepsilon}{S_m(\Lambda)} \sum_{j=1}^{m} \Lambda^{m-j} = \varepsilon;$$

that estimate uses only $\mathrm{Lip}(f_j) \leq \Lambda$. Hence the $M^m$ compositions $c_{i_m} \circ \cdots \circ c_{i_1}$ form an $\varepsilon$-cover of $F^m$, which proves the first inequality. For the second, $m \leq k$ implies $S_m(\Lambda) \leq S_k(\Lambda)$, hence $\varepsilon/S_k(\Lambda) \leq \varepsilon/S_m(\Lambda)$ and $N(F, d_\infty, \varepsilon/S_m(\Lambda))^m \leq N(F, d_\infty, \varepsilon/S_k(\Lambda))^k$ because covering numbers are nonincreasing in the radius and at least 1. □

### G.2 Profiles from the layerwise bound

**Corollary 9** (Profiles from the layerwise bound)**.** *Assume $D_k(S) \leq \overline{D}$ for all $k$ with $\overline{D} > 0$, let $k \geq 1$, and assume $N(F, d_\infty, \varepsilon) < \infty$ for all $\varepsilon > 0$. Since $\mathsf{V}_k(S)$ is formed with internal $d_S$-covering numbers while the layerwise bound bounds external $d_\infty$-covering numbers, the transfer $N(A, d_S, \varepsilon) \leq N(A, d_\infty, \varepsilon/2)$ of Section H is used, which is the origin of the factors 2 below.*

*(a) (Parametric layers.) If $\log N(F, d_\infty, \varepsilon) \leq p \log(C/\varepsilon)$ for all $0 < \varepsilon \leq C$, with $p \geq 0$ and $C \geq \overline{D}$, then for $0 < \varepsilon \leq C$*

$$\log N\big(B(k,F), d_\infty, \varepsilon\big) \leq \log(k+1) + kp\big[\log(C/\varepsilon) + \log S_k(\Lambda)\big],$$

*and, with $\Lambda_+ := \max\{1, \Lambda\}$ and $S_k(\Lambda) \leq k\Lambda_+^k$,*

$$\mathsf{V}_k(S) \leq \overline{D}\Big(\sqrt{\log(k+1)} + \sqrt{kp \log k} + k\sqrt{p \log \Lambda_+} + \sqrt{kp}\,\big(\sqrt{\log(2C/\overline{D})} + \tfrac{\sqrt{\pi}}{2}\big)\Big). \tag{20}$$

*Thus $\mathsf{V}_k(S) = O(\sqrt{kp \log(kC/\overline{D})})$ if $\Lambda \leq 1$ and $\mathsf{V}_k(S) = O(k\sqrt{p \log \Lambda})$ if $\Lambda > 1$.*

*(b) (Nonparametric layers.) If $\log N(F, d_\infty, \varepsilon) \leq c\,\varepsilon^{-q}$ for all $\varepsilon > 0$, with $c \geq 0$ and $0 < q < 2$, then $\log N(B(k,F), d_\infty, \varepsilon) \leq \log(k+1) + kc\, S_k(\Lambda)^q \varepsilon^{-q}$ for all $\varepsilon > 0$, and*

$$\mathsf{V}_k(S) \leq \overline{D}\sqrt{\log(k+1)} + \sqrt{kc}\,\big(2S_k(\Lambda)\big)^{q/2}\, \frac{\overline{D}^{\,1-q/2}}{1 - q/2}. \tag{21}$$

*Thus $\mathsf{V}_k(S) = O(k^{(1+q)/2})$ if $\Lambda \leq 1$ and $\mathsf{V}_k(S) = O(k^{1/2}\Lambda^{qk/2})$ if $\Lambda > 1$.*

*Proof.* (a) For $0 < \varepsilon \leq C$ and $k \geq 1$ we have $\varepsilon/S_k(\Lambda) \leq \varepsilon \leq C$, so the hypothesis applies at radius $\varepsilon/S_k(\Lambda)$ and (19) gives $\log N(B(k,F), \varepsilon) \leq \log(1 + kN^k) \leq \log(k+1) + k \log N$ with $N := N(F, d_\infty, \varepsilon/S_k(\Lambda))$, $\log N \leq p[\log(C/\varepsilon) + \log S_k(\Lambda)]$; here $1 + kN^k \leq (k+1)N^k$ because $N \geq 1$. Since $\Lambda^i \leq \Lambda_+^k$ for $i < k$, $S_k(\Lambda) \leq k\Lambda_+^k$, so $\log S_k(\Lambda) \leq \log k + k \log \Lambda_+$. Taking square roots with $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$\sqrt{\log N(B(k,F), \varepsilon)} \leq \sqrt{\log(k+1)} + \sqrt{kp \log k} + k\sqrt{p \log \Lambda_+} + \sqrt{kp}\sqrt{\log(C/\varepsilon)},$$

for every $0 < \varepsilon \le C$. Since $N(B(k,F), d_S, \varepsilon) \le N(B(k,F), d_\infty, \varepsilon/2)$ and $D_k(S) \le \overline{D}$, $\mathsf{V}_k(S)$ is bounded by the integral of the right-hand side at radius $\varepsilon/2$ over $(0, \overline{D}]$, where $\int_0^{\overline{D}} \sqrt{\log(2C/\varepsilon)}\,\mathrm{d}\varepsilon \le \overline{D}\big(\sqrt{\log(2C/\overline{D})} + \sqrt{\pi}/2\big)$ by Lemma 9 applied to $\log(2C/\varepsilon) = \log(2C/\overline{D}) + \log(\overline{D}/\varepsilon)$; this gives (20).

(b) Now $\log N(B(k,F),\varepsilon) \le \log(k+1) + kc\,(S_k(\Lambda)/\varepsilon)^q$ for all $\varepsilon > 0$; at radius $\varepsilon/2$ the second term is $kc(2S_k(\Lambda))^q \varepsilon^{-q}$, and $\int_0^{\overline{D}} \varepsilon^{-q/2}\,\mathrm{d}\varepsilon = \overline{D}^{\,1-q/2}/(1-q/2)$ because $q < 2$. The asymptotics follow from $S_k(\Lambda) \le k$ for $\Lambda \le 1$ and $S_k(\Lambda) \le \Lambda^k/(\Lambda - 1)$ for $\Lambda > 1$. □

### G.3 Sharpness

The memory-preserving construction of Condition E3 attains both layerwise bounds. Its layer class $F = \{r, A\} \cup \{g_u : u \in G\}$ consists of $\lambda$-Lipschitz maps for the bounded metric $d_X$ (the reset and the writers are 1-Lipschitz, and $\min\{1, \lambda a\} \le \lambda \min\{1, a\}$ for $\lambda \ge 1$), and $N(F, d_\infty, \varepsilon) \le 2 + N(G, \|\cdot\|_E, \varepsilon)$ for $\varepsilon < 1$. If $G$ has finite-dimensional entropy, $\log N(G,\delta) \asymp \log(1/\delta)$, then case (a) gives $\log N(B(k,F),\varepsilon) \lesssim k^2$, and Corollary 7 gives the matching lower bound $\gtrsim k^2$ (its words $W_k$ lie in $B(2k+1, F)$, and external covering numbers are monotone under inclusion). If $\log N(G,\delta) \asymp \delta^{-p}$, case (b) gives $\log N(B(k,F),\varepsilon) \lesssim k\lambda^{pk}$, and Corollary 8 gives $\gtrsim \lambda^{pk}$. Thus the layerwise bound is attained up to polynomial factors, and no better general bound in terms of $N(F,\cdot)$ and $\Lambda$ alone is possible.

### G.4 Diameter of the reachable states

On a bounded state space $D_k(S) \le \operatorname{diam}(\mathcal{X})$ for every $k$. On an unbounded space the diameter can grow with the depth, and the growth is again governed by $\Lambda$.

**Lemma 8** (Reachable radius)**.** *Let $x_0 \in \mathcal{X}$, $R \ge 0$, $c \ge 0$, and suppose $d(f(x_0), x_0) \le c$ for all $f \in F$. Then, with $\Lambda_+ := \max\{1, \Lambda\}$, every $g \in B(k,F)$ maps the ball $B(x_0, R)$ into $B\big(x_0, \Lambda_+^k R + c\,S_k(\Lambda)\big)$. Consequently, if $S \subset B(x_0, R)$, then $D_k(S) \le 2\big(\Lambda_+^k R + c\,S_k(\Lambda)\big)$: bounded in $k$ if $\Lambda < 1$, at most linear if $\Lambda = 1$, and at most exponential if $\Lambda > 1$.*

*Proof.* For $f \in F$ and $d(x, x_0) \le \rho$, $d(f(x), x_0) \le d(f(x), f(x_0)) + d(f(x_0), x_0) \le \Lambda\rho + c$. Iterating $m \le k$ times from $\rho_0 = R$ gives $\rho_m \le \Lambda^m R + c\sum_{i<m} \Lambda^i$, which is at most $\Lambda_+^k R + cS_k(\Lambda)$ since $\Lambda^m \le \Lambda_+^k$ and $S_m \le S_k$. The diameter bound follows from the triangle inequality. □

The exponential factor $\Lambda^k R$ is the product of the layerwise Lipschitz constants that appears in norm-based generalization bounds for fully connected networks (Neyshabur et al., 2015; Bartlett et al., 2017). In the present bound it enters only through the diameter, that is, through the integration range of the entropy integral, and only when the state space is unbounded; bounded activations, normalization layers, or a projection onto a bounded set remove it, after which Corollary 9 applies with $\overline{D} = \operatorname{diam}(\mathcal{X})$.

## H Variance profiles

The following proposition converts the fixed-scale growth mechanisms of Section 4, together with a bound on the diameter, into the depth profiles of the entropy integral listed at the end of Section 4; the list after the proposition records which case applies to each mechanism and to the layerwise bound.

### H.1 Statement

**Proposition 5** (Variance profiles from growth and diameter)**.** *Let $F \subset C(\mathcal{X}, \mathcal{X})$, write $N_k(\varepsilon) := N(B(k,F), d_\infty, \varepsilon)$, and let $\overline{D} \ge 0$ ($\overline{D} > 0$ in (ii)), $D_1 > 0$, $C_0 \ge 1$, and $D \ge 0$ be constants.*

*(i) (Saturation.) If $D_k(S) \le \overline{D}$ and $N_k(\varepsilon) \le N_\infty(\varepsilon)$ for all $k$ and $\varepsilon > 0$, with $\mathsf{V}_\infty := \int_0^{\overline{D}} \sqrt{\log N_\infty(\varepsilon)}\,\mathrm{d}\varepsilon < \infty$, then $\mathsf{V}_k(S) \le \mathsf{V}_\infty$ for all $k$.*

*(ii) (Polynomial growth, bounded diameter.) If* $D_k(S) \le \overline{D}$ *and* $N_k(\varepsilon) \le C_0(1+k/\varepsilon)^D$ *for all* $k \ge 1$ *and* $\varepsilon > 0$*, then* $\mathsf{V}_k(S) \le \overline{D}\sqrt{D}\big(\sqrt{\log(1+k/\overline{D})} + \frac{\sqrt{\pi}}{2}\big) + \overline{D}\sqrt{\log C_0} = O(\sqrt{D\log k})$.

*(iii) (Exponential growth, bounded diameter.) If* $D_k(S) \le \overline{D}$ *and* $\log N_k(\varepsilon) \le \alpha k + \psi(\varepsilon)$ *with* $\Psi := \int_0^{\overline{D}} \sqrt{\psi(\varepsilon)}\,\mathrm{d}\varepsilon < \infty$*, then* $\mathsf{V}_k(S) \le \overline{D}\sqrt{\alpha k} + \Psi = O(\sqrt{k})$*; in particular, if* $|F| = r \ge 2$*, then* $N_k(\varepsilon) \le r^{k+1}$ *and* $\mathsf{V}_k(S) \le \overline{D}\sqrt{(k+1)\log r}$.

*(iv) (Polynomial growth, linearly growing diameter.) If* $D_k(S) \le D_1 k$ *and* $N_k(\varepsilon) \le C_0(1 + k/\varepsilon)^D$*, then* $\mathsf{V}_k(S) \le D_1 k\big(\sqrt{D}(\sqrt{\log(1+1/D_1)} + \frac{\sqrt{\pi}}{2}) + \sqrt{\log C_0}\big) = O(k\sqrt{D})$.

*Consequently,* $\mathsf{var}(k,n)$ *is of order* $n^{-1/2}$*,* $\sqrt{D\log k/n}$*,* $\sqrt{k/n}$*, and* $k\sqrt{D/n}$ *in the four cases, respectively.*

### H.2 Profiles by mechanism

In each item, the first entry is the fixed-scale entropy $\log N(B(k,F), d_\infty, \varepsilon)$, the second the entropy integral $\mathsf{V}_k(S)$, the third the estimation term $\mathsf{var}(k,n)$ with its letter (L for logarithmic, P for polynomial), and the last the case of Proposition 5 or of Corollary 9 and the matching lower bound where one is available. A bounded diameter is assumed except in the last item.

- P1, P1' (compact equicontinuous dynamics; contraction to a bounded set): $O(1)$; $O(1)$; $n^{-1/2}$ (L); case (i), saturated.
- P2 (nilpotent control), compact state space: $D\log(1+k/\varepsilon)$; $O(\sqrt{D\log k})$; $\sqrt{\log k/n}$ (L); case (ii), matched by Corollary 3(ii).
- E1', E2 ($r$ separated generators): $\asymp k\log r$; $O(\sqrt{k\log r})$; $\sqrt{k/n}$ (P, $\gamma = 1$); case (iii), matched by Corollary 3(i).
- Layerwise bound, parametric layers, $\Lambda \le 1$: $kp\log(kC/\varepsilon)$; $O(\sqrt{kp\log k})$; $\sqrt{k\log k/n}$ (P, $\gamma = 1$); Corollary 9(a), the order of the VC-dimension bounds of Bartlett et al. (2019).
- Layerwise bound, parametric layers, $\Lambda > 1$: $pk(k\log\Lambda + \log(C/\varepsilon))$; $O(k\sqrt{p\log\Lambda})$; $k\sqrt{p/n}$ (P, $\gamma = 2$); Corollary 9(a), attained by E3 (Corollary 7).
- Layerwise bound, nonparametric layers with entropy $c\varepsilon^{-q}$, $\Lambda > 1$: $kc(S_k(\Lambda)/\varepsilon)^q$; $O(\sqrt{k}\,\Lambda^{qk/2})$; exponential; Corollary 9(b), attained by E3 (Corollary 8).
- P2, non-compact state space ($D_k \le 2L_\alpha R_S k$): $D\log(1+k/\varepsilon)$; $O(k\sqrt{D})$; $k\sqrt{D/n}$ (P, $\gamma = 2$); case (iv).

### H.3 Proof

We now prove the elementary estimates behind Proposition 5. Throughout, $N_k(\varepsilon) := N(B(k,F), d_\infty, \varepsilon)$, and we use that $N(B(k,F), d_S, \varepsilon) \le N_k(\varepsilon)$ because $d_S \le d_\infty$ (Section 4); this holds for external and for internal covering numbers alike, since a $d_\infty$-cover of $B(k,F)$ is a $d_S$-cover with the same centres. If the entropy integral of Theorem 1 is formed with internal covering numbers while the growth conditions of Section F bound external ones, use $N^{\mathrm{int}}(A, d_S, \varepsilon) \le N(A, d_S, \varepsilon/2) \le N(A, d_\infty, \varepsilon/2)$ instead; this replaces $C_0$ by $2^D C_0$ in cases (ii) and (iv), $\psi(\varepsilon)$ by $\psi(\varepsilon/2)$ in case (iii), and $N_\infty(\varepsilon)$ by $N_\infty(\varepsilon/2)$ in case (i), and changes nothing else.

**Lemma 9.** *For* $\overline{D} > 0$*,* $k \ge 0$*, and* $0 < \varepsilon \le \overline{D}$*,*

$$\log\Big(1 + \frac{k}{\varepsilon}\Big) \le \log\Big(1 + \frac{k}{\overline{D}}\Big) + \log\frac{\overline{D}}{\varepsilon}, \qquad \int_0^{\overline{D}} \sqrt{\log\frac{\overline{D}}{\varepsilon}}\,\mathrm{d}\varepsilon = \overline{D}\,\Gamma(3/2) = \frac{\sqrt{\pi}}{2}\,\overline{D}.$$

*Proof.* For $0 < \varepsilon \le \overline{D}$, $(1 + k/\overline{D})(\overline{D}/\varepsilon) = \overline{D}/\varepsilon + k/\varepsilon \ge 1 + k/\varepsilon$, which gives the first inequality after taking logarithms. For the integral, substitute $\varepsilon = \overline{D}t$ and then $t = e^{-s}$:

$$\int_0^{\overline{D}} \sqrt{\log(\overline{D}/\varepsilon)}\,\mathrm{d}\varepsilon = \overline{D}\int_0^1 \sqrt{\log(1/t)}\,\mathrm{d}t = \overline{D}\int_0^\infty s^{1/2}e^{-s}\,\mathrm{d}s = \overline{D}\,\Gamma(3/2).$$

□

*Proof of Proposition 5.* (i) Since the integrand and the upper limit are both dominated, $\mathsf{V}_k(S) \le \int_0^{\overline{D}} \sqrt{\log N_\infty(\varepsilon)}\,\mathrm{d}\varepsilon = \mathsf{V}_\infty$.

(ii) Using $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$ and Lemma 9, for $0 < \varepsilon \le \overline{D}$,

$$\sqrt{\log N_k(\varepsilon)} \le \sqrt{\log C_0} + \sqrt{D\log(1 + k/\overline{D})} + \sqrt{D\log(\overline{D}/\varepsilon)}.$$

Integrating over $(0, \overline{D}]$ and applying the identity in Lemma 9 to the last term gives the claim.

(iii) Here $\sqrt{\log N_k(\varepsilon)} \le \sqrt{\alpha k} + \sqrt{\psi(\varepsilon)}$, and integration over $(0, \overline{D}]$ gives $\mathsf{V}_k(S) \le \overline{D}\sqrt{\alpha k} + \Psi$. If $|F| = r \ge 2$, then $B(k, F)$ has at most $\sum_{m=0}^k r^m \le r^{k+1}$ elements (this uses $r \ge 2$), so $N_k(\varepsilon) \le r^{k+1}$ at every scale and one may take $\alpha = \log r$ with $\psi \equiv \log r$, or directly $\log N_k(\varepsilon) \le (k+1)\log r$, which yields $\mathsf{V}_k(S) \le \overline{D}\sqrt{(k+1)\log r}$.

(iv) Apply Lemma 9 with $\overline{D}$ replaced by $D_1 k$: for $0 < \varepsilon \le D_1 k$, $\log(1 + k/\varepsilon) \le \log(1 + 1/D_1) + \log(D_1 k/\varepsilon)$, hence

$$\begin{aligned}\mathsf{V}_k(S) &\le \int_0^{D_1 k} \Big(\sqrt{\log C_0} + \sqrt{D\log(1 + 1/D_1)} + \sqrt{D\log(D_1 k/\varepsilon)}\Big)\mathrm{d}\varepsilon \\ &= D_1 k\Big(\sqrt{\log C_0} + \sqrt{D\log(1 + 1/D_1)} + \sqrt{D}\,\tfrac{\sqrt{\pi}}{2}\Big).\end{aligned}$$

The statements about $\mathsf{var}(k, n)$ follow by multiplying with $12A_H L/\sqrt{n}$. □

### H.4 Diameter bounds by mechanism

Under Condition P1 the state space is compact, so $D_k(S) \le \mathrm{diam}_\infty(B(k, F)) \le \mathrm{diam}(\mathcal{X})$. Under Condition P1' with the bounded-metric convention $d_b = \min\{1, d\}$, $D_k(S) \le 1$; without that convention, all words of length at least $L$ take values in the bounded absorbing set $K$, so the sub-ball of such words has diameter at most $\mathrm{diam}(K)$. Under Condition P2, the proof gives $B(k, F) \subset \alpha(B_H(e, kR_S))$ and the orbit map is $L_\alpha$-Lipschitz, so $D_k(S) \le d_\infty$-diam $\alpha(B_H(e, kR_S)) \le 2L_\alpha R_S k$; on a compact state space this is improved to $\mathrm{diam}(\mathcal{X})$. Under Condition E2 the state space is compact, and the construction of Condition E3 uses a metric bounded by 1. Under Condition E1 with the standard word metric, $d_\infty$ may be infinite; the bi-invariant variant Example 5 has $D_k \le k$, and after passing to $d_b$ the diameter is at most 1.

## I The four regimes and the balancing depths

This appendix defines the four regimes of Section 5, states their balancing depths and balanced rates, and derives them.

### I.1 Two laws and four regimes

The main text names two of the four combinations of bias and variance laws; here we treat all four. The approximation error decays with depth $k$ either exponentially or polynomially,

$$\mathsf{bias}(k) = \exp(\Theta(-\alpha k)) \quad \text{or} \quad \Theta(k^{-\beta}),$$

with parameters $\alpha, \beta > 0$; the letters E and P refer to these two cases. The estimation term follows one of the two profiles produced by Proposition 5, root-logarithmic or root-polynomial growth,

$$\mathsf{var}(k, n) \lesssim \sqrt{\log(k)/n} \quad \text{or} \quad \sqrt{k^\gamma/n},$$

with parameter $\gamma > 0$; the letters L and P refer to these two cases. The root-logarithmic profile is case (ii) of Proposition 5, and it also covers the saturated case (i), for which the true dependence is constant; case (iii) is the root-polynomial profile with $\gamma = 1$, and case (iv) with $\gamma = 2$. Combining the two bias laws with the two variance profiles gives rise to the four typical regimes EL, EP, PL, and PP listed in (22): the first letter describes the bias and the second the variance. They are representative diagnostic cases rather than an exhaustive list of all possible depth dependences; for instance, the sub-exponential bias decay $\exp(-\Theta(\sqrt{k}))$ of Picard-type neural operators (Section J) combined with a saturated or root-logarithmic profile is called the SubEL regime there.

The balancing depths and balanced rates are the following, with the derivations below:

$$\begin{aligned}
&\text{EL } (\mathsf{bias} = e^{-\alpha k},\ \mathsf{var} = \sqrt{\log k/n}): && k^* = \tfrac{1}{2\alpha}(\log n - \log\log\log n), && \mathsf{gen}(k^*, n) \asymp \sqrt{\tfrac{\log\log n}{n}};\\
&\text{EP } (\mathsf{bias} = e^{-\alpha k},\ \mathsf{var} = \sqrt{k^\gamma/n}): && k^* = \tfrac{1}{2\alpha}(\log n - \gamma\log\log n), && \mathsf{gen}(k^*, n) \asymp \sqrt{\tfrac{(\log n)^\gamma}{(2\alpha)^\gamma n}};\\
&\text{PL } (\mathsf{bias} = k^{-\beta},\ \mathsf{var} = \sqrt{\log k/n}): && k^* \sim \left(\tfrac{2\beta n}{\log(2\beta n)}\right)^{\frac{1}{2\beta}}, && \mathsf{gen}(k^*, n) \asymp \sqrt{\tfrac{\log n}{2\beta n}};\\
&\text{PP } (\mathsf{bias} = k^{-\beta},\ \mathsf{var} = \sqrt{k^\gamma/n}): && k^* \asymp n^{\frac{1}{2\beta+\gamma}}, && \mathsf{gen}(k^*, n) \asymp n^{-\frac{\beta}{2\beta+\gamma}}.
\end{aligned} \tag{22}$$

Lower-order terms in $k^*$ are omitted in the EL and EP rows.

The worked examples are placed in these regimes in the summary list of Section J.

### I.2 Width and other one-layer parameters

*Remark* 3 (Width and other one-layer parameters). The display (22) fixes the hidden-layer class $F$ and the output-layer class $H$ and varies only the depth. Width, and more generally any parameter of a single layer, is not absent from the framework: if $F_w$ denotes the admissible hidden layers at width $w$, then $w$ enters the one-layer covering numbers $N(F_w, d_\infty, \varepsilon)$, hence those of $B(k, F_w)$ through Theorem 2, where $p \asymp w^2$ for a fully connected layer, and possibly the output-layer class $H_w$, the approximation term, and the implementation error (Section J.3 quantifies the last route). Writing $\mathcal{H}_{k,w} = H_w \circ B(k, F_w)$, the balancing depth $k^*(n, w)$ is obtained by minimizing the full bound over $k$ for each fixed $w$; the display does not assert that it is independent of $w$, and its dependence on $w$ need not be monotone because width can change both the bias and the variance. A joint optimization over depth and width, and a resource-matched comparison with wide shallow networks, are outside the present analysis.

### I.3 The balancing principle

Throughout, we minimize the upper bound

$$\mathsf{gen}(k, n) \lesssim \mathsf{bias}(k) + \mathsf{var}(k, n)$$

over the depth $k$, treated as a positive real number and rounded to the nearest integer at the end. Since $\mathsf{bias}(k)$ is nonincreasing and $\mathsf{var}(k, n)$ is nondecreasing in $k$, the minimizer is located where the two terms are of the same order, and the balanced value is then of the order of either term. The following lemma makes this precise.

**Lemma 10** (Balancing principle)**.** *Let* $\mathsf{bias} \geq 0$ *be nonincreasing and* $\mathsf{var}(\cdot, n) \geq 0$ *be nondecreasing on a set* $\mathcal{K}$ *of depths, and let* $\mathsf{gen}(k, n) := \mathsf{bias}(k) + \mathsf{var}(k, n)$*. If* $k_0 \in \mathcal{K}$ *balances the two terms,* $\mathsf{bias}(k_0) = \mathsf{var}(k_0, n)$*, then* $\mathsf{gen}(k, n) \geq \mathsf{bias}(k_0)$ *for every* $k \in \mathcal{K}$*, while* $\mathsf{gen}(k_0, n) = 2\mathsf{bias}(k_0)$*; hence* $k_0$ *minimizes the bound over* $\mathcal{K}$ *up to the factor* 2*.*

*Proof.* For $k \leq k_0$ we have $\mathsf{gen}(k, n) \geq \mathsf{bias}(k) \geq \mathsf{bias}(k_0)$, and for $k \geq k_0$ we have $\mathsf{gen}(k, n) \geq \mathsf{var}(k, n) \geq \mathsf{var}(k_0, n) = \mathsf{bias}(k_0)$. □

In the PP regime the balancing equation is solved exactly, so Lemma 10 applies verbatim. In the other three regimes the displayed $k^*$ balances the two terms only asymptotically; the balanced rates are nevertheless two-sided, of the form $\Theta(\cdot)$, with explicit constants. For all sufficiently large $n$, the EP rate lies between 1 and $1 + (2\alpha)^{-\gamma/2}$ times $n^{-1/2}(\log n)^{\gamma/2}$, the EL rate between 1 and $1 +$

$\sqrt{1+|\log 2\alpha|}$ times $\sqrt{\log\log n/n}$, and the PL rate between $\sqrt{1/(4\beta)}$ and $2\sqrt{(1+|\log 2\beta|)/(2\beta)}$ times $\sqrt{\log n/n}$. If the two terms do not cross, the minimizer lies at a boundary of $\mathcal{K}$; in the four regimes below they cross for all large $n$.

### I.4 Derivations by regime

#### I.4.1 EP: exponentially decaying bias, polynomially growing variance

Let $\mathsf{bias}(k) = e^{-\alpha k}$ and $\mathsf{var}(k,n) = n^{-1/2}k^{\gamma/2}$ with $\alpha, \gamma > 0$. Taking logarithms, the two terms are of the same order exactly when

$$\alpha k = \tfrac{1}{2}\log n - \tfrac{\gamma}{2}\log k + O(1).$$

As $n \to \infty$ the solution satisfies $\log k = \log\log n + O(1)$, because $\log k \ll \log n$; substituting this into the right-hand side gives

$$k^* = \frac{1}{2\alpha}\big(\log n - \gamma\log\log n + O(1)\big).$$

Evaluating either term at $k^*$ yields the balanced rate

$$\mathsf{gen}(k^*, n) \asymp n^{-1/2}(\log n)^{\gamma/2},$$

more precisely $(2\alpha)^{-\gamma/2}n^{-1/2}(\log n)^{\gamma/2}$ up to a factor bounded above and below by positive constants.

#### I.4.2 EL: exponentially decaying bias, logarithmically growing variance

Let $\mathsf{bias}(k) = e^{-\alpha k}$ and $\mathsf{var}(k,n) = \sqrt{\log k/n}$ with $\alpha > 0$. The two terms are of the same order exactly when

$$\alpha k = \tfrac{1}{2}\log n - \tfrac{1}{2}\log\log k + O(1).$$

Since $\log k = \log\log n + O(1)$ at the solution, and hence $\log\log k = \log\log\log n + O(1)$, we obtain

$$k^* = \frac{1}{2\alpha}\big(\log n - \log\log\log n + O(1)\big), \qquad \mathsf{gen}(k^*, n) \asymp \sqrt{\frac{\log\log n}{n}}.$$

#### I.4.3 PP: polynomially decaying bias, polynomially growing variance

Let $\mathsf{bias}(k) = k^{-\beta}$ and $\mathsf{var}(k,n) = n^{-1/2}k^{\gamma/2}$ with $\beta, \gamma > 0$. The two terms are equal exactly when $k^{\beta+\gamma/2} = n^{1/2}$, that is, at

$$k^* = n^{1/(2\beta+\gamma)}, \qquad \text{where} \qquad \mathsf{gen}(k^*, n) \asymp n^{-\beta/(2\beta+\gamma)}.$$

Here the balancing equation is solved exactly, so Lemma 10 shows that $k^*$ minimizes the bound up to the factor 2.

#### I.4.4 PL: polynomially decaying bias, logarithmically growing variance

Let $\mathsf{bias}(k) = k^{-\beta}$ and $\mathsf{var}(k,n) = \sqrt{\log k/n}$ with $\beta > 0$. Squaring both terms, they are of the same order exactly when $k^{2\beta}\log k \asymp n$. The equation $k^{2\beta}\log k = n$ is solved by $k = \exp(W(2\beta n)/(2\beta))$, where $W$ is the Lambert function. The Lambert function can be avoided by taking

$$k^* := \Big(\frac{2\beta n}{\log(2\beta n)}\Big)^{1/(2\beta)}, \qquad \text{for which} \qquad k^{*2\beta}\log k^* = n\Big(1 - \frac{\log L}{L}\Big), \quad L := \log(2\beta n),$$

so that $k^{*2\beta}\log k^*/n \to 1$ as $n \to \infty$. For this choice, $\mathsf{bias}(k^*) = \sqrt{L/(2\beta n)}$ and, for $L \geq 1$, $\mathsf{var}(k^*, n) = \sqrt{(L - \log L)/(2\beta n)} \leq \sqrt{L/(2\beta n)}$, so that

$$\mathsf{gen}(k^*, n) \asymp \sqrt{\frac{\log n}{2\beta n}}.$$

**Ordering of the EP and PL rates.** The EP and PL rates in (22) are not uniformly ordered without specifying $\gamma$: EP gives $n^{-1/2}(\log n)^{\gamma/2}$, whereas PL gives $n^{-1/2}(\log n)^{1/2}$, up to constants. Thus EP is no worse than PL for $\gamma \leq 1$, while PL is no worse than EP for $\gamma > 1$.

# J WORKED EXAMPLES: TEACHER–STUDENT MODELS, NEURAL OPERATORS, AND THE IMPLEMENTATION LAYER

This appendix opens the series of worked examples. Each example instantiates the framework in the same order: the setting (state space, hidden-layer class, output layer), the variance profile obtained from a growth mechanism of Section F, the bias against a target class, the balancing depth, and, where available, a computed illustration, the implementation error, and the relation to the literature. The present appendix treats the two examples that need no new construction, the contractive teacher–student model and Picard-type neural operators, and then the implementation layer of Theorem 4, which is used by all examples. The deep ReLU, chain-of-thought, and solver examples follow in Sections K to M.

**Summary of the worked examples.** For each example we record the variance profile, the bias, and the balancing depth; "proved" means derived within the framework under the stated hypotheses, "known" means an existing architecture-specific rate inserted into the framework, and "assumed" means an assumption on the target.

- Contractive teacher–student model (Section J): saturated profile (proved, P1/P1'); bias $D\lambda^k$ (proved); EL, $k^* \asymp \log n/(2\log\lambda^{-1})$.
- Fixed-point refinement, iterative denoising (Section M): saturated profile with explicit constants (proved); bias $(1-h\mu)^k$ (proved); EL, $k^* \leq \log n/(2h\mu)$.
- Fixed-horizon integration, neural ODE, diffusion sampler (Section M): $\sqrt{\log k}$ profile with explicit constants (proved); bias $C_E T^2 k^{-p}$ (proved); PL, $k^* \asymp n^{1/(2p)}$.
- Chain of thought, append-only scratchpad (Section L): saturated profile with explicit constants (proved); bias $\theta^k$ (proved); EL, $k^* \asymp \log n/(2\log\theta^{-1})$.
- Chain of thought, branching (Section L): $\sqrt{\log k}$ profile on every sample (proved), $\sqrt{k}$ in $d_\infty$; bias 0 for $k \geq k_0$ (proved) or $e^{-\alpha k}$ (assumed); $k^* = k_0$, or EL.
- Picard-type neural operator (Section J): saturated profile (proved, P1'); bias $e^{-c\sqrt{k}}$ (known); SubEL.
- Deep ReLU network (Section K): saturated profile for contractive or non-expanding layers on a bounded state space (proved), layerwise bound profile $k\sqrt{p}$ for expanding layers (proved) with the $\sqrt{k}$ lower bound realized by expand-and-reset layers (proved); bias $\Lambda^k$ for a contractive teacher (proved), of Jackson type $k^{-s/d}$ for Hölder or Besov targets (known); EL, respectively PP.

## J.1 CONTRACTIVE TEACHER–STUDENT MODELS

**Setting.** A teacher-student setting refers to the scenario where the hypothesis class (student) and the target class (teacher) share the same compositional structure. Let the depth-$k$ student class be $\mathcal{H}_k := H \circ B(k, F)$ and take the target class to be the uniform closure of the infinitely deep compositional class, $\mathcal{C} := \mathcal{H}_\infty := H \circ \overline{\bigcup_{m\geq 0} B(m,F)}^{\,d_\infty}$. Assume the input space $\mathcal{X}$ is bounded with $\operatorname{diam}(\mathcal{X}) = D < \infty$; the output layer $H$ is 1-Lipschitz; and the intermediate-layer semigroup $F \subset C(\mathcal{X},\mathcal{X})$ is *contractive*, i.e., there exists $\lambda \in (0,1)$ such that $\operatorname{Lip}(f) \leq \lambda$ for all $f \in F$. A teacher of depth less than $k$ is represented exactly by $\mathcal{H}_k$. For a finite-depth teacher of depth at least $k$, write $c = h \circ u \circ v$ with $h \in H$, $u \in F^k$, and $v \in B(\ell, F)$. Truncating the tail yields a depth-$k$ student $h \circ u \in \mathcal{H}_k$ that approximates $c$ with error

$$\|h \circ u \circ v - h \circ u\|_\infty \leq \operatorname{Lip}(h)\operatorname{Lip}(u) d_\infty(v, \mathrm{id}) \leq 1 \cdot \lambda^k \cdot D = D\lambda^k,$$

since $\mathrm{Lip}(u) \le \lambda^k$ and $d_\infty(v, \mathrm{id}) \le D$. Passing to the uniform closure gives the same bound for teachers in $\mathcal{H}_\infty$. Hence the approximation error decays exponentially:

$$\sup_{c \in \mathcal{H}_\infty} \inf_{h \in \mathcal{H}_k} \|c - h\|_\infty \le D\lambda^k.$$

Moreover, under the compact-domain P1 condition, or under the non-compact absorbing-set version P1', the contractive transition family has saturated word-ball entropy: the hidden contribution to the variance upper bound is independent of $k$ at fixed scale, and in the compact/totally bounded case the whole Dudley integral is depth-independent (case (i) of Proposition 5). Thus this example is at least as favorable as the EL proxy with root-logarithmic variance growth.

### J.2 Picard-type neural operators

**Setting and bias.** Furuya et al. (2025). Furuya et al. (2025) investigated the approximation error of Neural Operators (NOs) that learn the solution operator of nonlinear parabolic PDEs on a bounded domain $\mathcal{X} \subset \mathbb{R}^d$. By aligning a single hidden layer of the NO with one step of the Picard iteration for the PDE, a textbook iterative argument for solving differential equations, they show that the approximation error decays at a *sub-exponential* rate of the form $O(\exp(-c\sqrt{k}))$ in the network depth $k$, for a problem-dependent constant $c > 0$. Since Picard iteration is *contractive* and generated by a single operator, the resulting transition dynamics are naturally treated by P1 on compact state spaces, or by P1' when the iteration has a compact absorbing set. In either case, the transition word-ball entropy saturates rather than merely growing polynomially. The function space for approximation is a *mixed Lebesgue space* $L^r_t L^s_x$, which we can regard as the target class $\mathcal{C}$. A finite-sample generalization analysis of this Picard-type setting, including finite-observation reconstruction errors and long-time rollout, is developed by Taniguchi & Sonoda (2026).

This NO setting can be viewed as a concrete instance of the teacher-student framework discussed earlier: both the hypothesis class and the target class share a hierarchical (iterative) structure, enabling efficient depth-driven approximation. It thus exemplifies the *SubEL regime* of Section 5, sub-exponential bias decay with a saturated or root-logarithmic variance profile, in which depth is particularly advantageous. Together with the previous example, it suggests that depth is most effective when the underlying target structure, namely the target class $\mathcal{C}$, is hierarchical, as in deep compositions or differential equations.

### J.3 The implementation layer

The implementation map in Theorem 4 is the identity when the abstract class is itself the implemented network class, and then $\varepsilon_{\mathrm{imp}}(k) = 0$. The distinction becomes useful when a mathematically convenient transition model is only approximated by a finite network, for instance when $\mathcal{H}_k$ is a class of polynomials or of smooth flows and $\mathcal{H}_{\mathrm{imp}}$ is a finite-width ReLU family. The following statement gives three explicit implementation maps.

**Proposition 6** (Implementation with controlled uniform error)**.** *(a) (Net-based.) Let $\mathcal{H}_k$ be totally bounded in the uniform norm on $\mathcal{X}$, and let $\mathcal{A}$ be a class of functions that is uniformly dense on $\mathcal{H}_k$; for example, on a compact $\mathcal{X}$, one-hidden-layer networks with a non-polynomial activation. For every $\varepsilon > 0$ there is a finite implementation class $\mathcal{H}^\varepsilon_{imp} \subset \mathcal{A}$ with $|\mathcal{H}^\varepsilon_{imp}| \le N(\mathcal{H}_k, \|\cdot\|_\infty, \varepsilon/2)$ and an implementation map with $\varepsilon_{imp}(k) \le \varepsilon$.*

*(b) (Layerwise.) If every transition is $\Lambda$-Lipschitz, every output layer is $L_H$-Lipschitz, and each generator and output layer is implemented within uniform error $\delta$ and $\delta_H$, respectively ($\Lambda, \delta, \delta_H \ge 0$), then composing the implemented layers gives $\varepsilon_{imp}(k) \le \delta_H + L_H\, \delta \sum_{i=0}^{k-1} \Lambda^i$; this is $O(k\delta)$ for non-expanding transitions and $O(\delta)$ for contractive ones.*

*(c) (Exact.) If $\mathcal{X} = \mathbb{R}^d$ and the transitions are affine, as for the nilpotent matrix actions of Condition P2, then each transition is one ReLU layer of width $2d$ via $x = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x)$, every word in $B(k, F)$ is a depth-$k$ ReLU network of width $2d$, and $\varepsilon_{imp}(k) = 0$.*

Part (a) shows that the implementation class is generally a finite class that does not contain $\mathcal{H}_k$, with an error that can be made arbitrarily small at the price of a larger class; its size is controlled by the covering number of the abstract class. Part (b) shows how a per-layer resolution $\delta$, for instance a

width- or precision-dependent quantity, interacts with depth: additively under layerwise stability and geometrically without it. Part (c) answers the implementation question for the Heisenberg and upper-triangular examples of Section F: no exponential width is needed, and the polynomial growth bounds of P2 apply to a concrete ReLU network with zero implementation error. Quantitative finite-width results on manifolds and symmetric spaces (Kratsios & Papon, 2022; Chen et al., 2019; Sonoda et al., 2022; Petersen & Voigtlaender, 2020) give further examples of controlled $\varepsilon_{\text{imp}}$ under additional assumptions on the target and the region being approximated; we do not claim that P2 alone yields an efficient implementation of every nilpotent action.

**Width through the implementation error.** Part (b) makes the role of the width explicit. Suppose the abstract hidden layers are $\Lambda$-Lipschitz maps whose coordinates belong to a smoothness class on the bounded state set, and that each of them is implemented by a block of width $w$ with uniform accuracy $\delta(w)$; for example $\delta(w) \asymp w^{-s/d}$ for $C^s$ layers on a bounded subset of $\mathbb{R}^d$ (DeVore et al., 1989; Yarotsky, 2017), and $\delta(w) \asymp w^{-1/2}$ in $L^2$ for layers of bounded variation norm (Barron, 1993; Jones, 1992; Pisier, 1981). Then, with a fixed output layer,

$$\varepsilon_{\text{imp}}(k) \le \delta_H + L_H\, \delta(w)\, S_k(\Lambda), \qquad S_k(\Lambda) = \sum_{i<k} \Lambda^i,$$

and the implementation error is below the estimation term $\mathsf{var}(k,n)$ as soon as $\delta(w) \le \mathsf{var}(k,n)/(L_H S_k(\Lambda))$. For non-expanding layers and the root-polynomial profile $\mathsf{var}(k,n) \asymp \sqrt{k/n}$ this requires $w \gtrsim (L_H\sqrt{kn})^{d/s}$: the width needed to realize a depth-$k$ abstract network without changing its regime grows polynomially in $k$ and $n$, and for expanding layers geometrically in $k$. Width therefore enters the depth trade-off through two routes, the one-layer covering number $p \asymp w^2$ in Theorem 2 and the per-layer accuracy $\delta(w)$ here, and both are explicit; a joint optimization over $(k, w)$ would balance the three terms $\mathsf{bias}(k)$, $\mathsf{var}(k,n)$, and $\beta_\ell \varepsilon_{\text{imp}}(k)$ simultaneously, which we do not pursue.

**Proof of Proposition 6.** We restate Proposition 6 in slightly more general form and prove it. The point of the statement is that the implementation class in Theorem 4 is a genuine third object: it need not contain $\mathcal{H}_k$, and the uniform implementation error can be traded against the size or the resolution of the implementation.

**Proposition 7** (Proposition 6, restated)**.** *(a) (Net-based implementation.) Let $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$ be totally bounded in the uniform norm, and let $\mathcal{A} \subset \mathbb{R}^{\mathcal{X}}$ be uniformly dense on $\mathcal{H}_k$: for every $g \in \mathcal{H}_k$ and $\eta > 0$ there is $a \in \mathcal{A}$ with $\|g - a\|_\infty \le \eta$ (no compactness of the domain and no continuity is needed; on a compact $K$ any dense $\mathcal{A} \subset C(K)$ qualifies). Then for every $\varepsilon > 0$ there exist a finite class $\mathcal{H}^{\varepsilon}_{imp} \subset \mathcal{A}$ with $|\mathcal{H}^{\varepsilon}_{imp}| \le N(\mathcal{H}_k, \|\cdot\|_\infty, \varepsilon/2)$ and an implementation map $\Phi_{\text{embed}} : \mathcal{H}_k \to \mathcal{H}^{\varepsilon}_{imp}$ such that $\sup_{g\in\mathcal{H}_k} \|g - \Phi_{\text{embed}}(g)\|_\infty \le \varepsilon$.*

*(b) (Layerwise implementation.) Let $F \subset C(\mathcal{X}, \mathcal{X})$ with $\mathrm{Lip}(f) \le \Lambda$ for all $f \in F$ ($\Lambda \ge 0$), and let $H \subset C(\mathcal{X})$ with $\mathrm{Lip}(h) \le L_H$ for all $h \in H$. Suppose every $f \in F$ is assigned an implemented map $\tilde f : \mathcal{X} \to \mathcal{X}$ with $d_\infty(f, \tilde f) \le \delta$ ($\delta \ge 0$) and every $h \in H$ an implemented $\tilde h : \mathcal{X} \to \mathbb{R}$ with $\|h - \tilde h\|_\infty \le \delta_H$. For each element of $\mathcal{H}_k$ fix one representation $h \circ f_m \circ \cdots \circ f_1$ with $m \le k$ and define $\Phi_{\text{embed}}(h \circ f_m \circ \cdots \circ f_1) := \tilde h \circ \tilde f_m \circ \cdots \circ \tilde f_1$. Then*

$$\varepsilon_{imp}(k) \le \delta_H + L_H\, \delta \sum_{i=0}^{k-1} \Lambda^i,$$

*which is at most $\delta_H + L_H k\delta$ if $\Lambda \le 1$ and at most $\delta_H + L_H\delta/(1-\Lambda)$ if $\Lambda < 1$.*

*(c) (Exact implementation of affine transitions.) If $\mathcal{X} = \mathbb{R}^d$ and every $f \in F$ is affine, $f(x) = Ax + b$, then each $f$ is realized exactly by one ReLU layer of width $2d$, and every element of $B(k, F)$ by a ReLU network of depth $k$ and width $2d$. With a linear output-layer class, $\mathcal{H}_k \subset \mathcal{H}_{imp}$ for the corresponding ReLU network class and $\varepsilon_{imp}(k) = 0$.*

*Proof.* (a) Let $\{g_1, \dots, g_M\} \subset \mathcal{H}_k$ be an $\varepsilon/2$-net of $\mathcal{H}_k$ in $\|\cdot\|_\infty$ with centres in $\mathcal{H}_k$ and of minimal cardinality $M = N(\mathcal{H}_k, \|\cdot\|_\infty, \varepsilon/2)$ (internal covering number). By density on $\mathcal{H}_k$, choose $a_j \in \mathcal{A}$ with $\|a_j - g_j\|_\infty \le \varepsilon/2$, set $\mathcal{H}^{\varepsilon}_{\text{imp}} := \{a_1, \dots, a_M\}$, and let $\Phi_{\text{embed}}(g) := a_{j(g)}$

where $j(g)$ is the index of a net point with $\|g - g_{j(g)}\|_\infty \le \varepsilon/2$. Then $\|g - \Phi_{\mathrm{embed}}(g)\|_\infty \le \|g - g_{j(g)}\|_\infty + \|g_{j(g)} - a_{j(g)}\|_\infty \le \varepsilon$, and $|\mathcal{H}^{\varepsilon}_{\mathrm{imp}}| \le M$ since the $a_j$ need not be distinct.

(b) Fix a representation $h \circ f_m \circ \cdots \circ f_1$ and $x \in \mathcal{X}$. Put $y_0 := x$, $y_j := \tilde{f}_j(y_{j-1})$, and $G_j := f_m \circ \cdots \circ f_{j+1}$ for $0 \le j \le m$, with $G_m := \mathrm{id}$; then $\mathrm{Lip}(G_j) \le \Lambda^{m-j}$. Set $T_j := G_j(y_j)$, so that $T_0 = f_m \circ \cdots \circ f_1(x)$ and $T_m = \tilde{f}_m \circ \cdots \circ \tilde{f}_1(x)$. Since $T_{j-1} = G_j(f_j(y_{j-1}))$ and $T_j = G_j(\tilde{f}_j(y_{j-1}))$,

$$d(T_{j-1}, T_j) \le \Lambda^{m-j}\, d\big(f_j(y_{j-1}), \tilde{f}_j(y_{j-1})\big) \le \Lambda^{m-j}\delta.$$

By the triangle inequality, $d(T_0, T_m) \le \delta \sum_{j=1}^{m} \Lambda^{m-j} = \delta \sum_{i=0}^{m-1} \Lambda^i$. Finally,

$$|h(T_0) - \tilde{h}(T_m)| \le |h(T_0) - h(T_m)| + |h(T_m) - \tilde{h}(T_m)| \le L_H\, \delta \sum_{i=0}^{m-1} \Lambda^i + \delta_H,$$

and taking the supremum over $x$, over the representations with $m \le k$ (the sum is nondecreasing in $m$ because $\Lambda \ge 0$), and over $h \in H$ gives the claim.

(c) Since $x = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x)$ coordinatewise, one has $Ax + b = [A, -A]\,\mathrm{ReLU}\big([I_d; -I_d]\,x\big) + b$, a ReLU layer of width $2d$ followed by an affine map. Composing $k$ such blocks, and merging each output affine map with the next input affine map, gives a depth-$k$ ReLU network of width $2d$ that computes $f_k \circ \cdots \circ f_1$ exactly. A linear output layer composed with it is again such a network. □

**Remarks.** In (a), $\mathcal{A}$ may be taken to be one-hidden-layer networks with a non-polynomial activation (Cybenko, 1989; Hornik et al., 1989; Leshno et al., 1993; Pinkus, 1999), or finite expansions in a universal reproducing kernel; the required widths depend on $\varepsilon$ and on the regularity of the net points $g_j$, and quantitative rates require additional assumptions on the target and the domain (Barron, 1993; Kratsios & Papon, 2022; Chen et al., 2019; Sonoda et al., 2022; Petersen & Voigtlaender, 2020). In (b), the bound displays the interaction between per-layer resolution $\delta$ (for instance a width- or precision-dependent quantity) and depth: it is additive in $k$ for non-expanding transitions and geometric without layerwise stability. The implemented maps $\tilde{f}$ need not be continuous or Lipschitz; only the abstract transitions are used to propagate the error. In (c), the nilpotent examples of Condition P2, such as Example 4, fall under this case, so their polynomial growth bounds apply verbatim to a concrete ReLU implementation with zero implementation error; on a torus the torus itself is the state space, and the same affine layers act after reduction modulo $\mathbb{Z}^d$.

## K Worked example: deep ReLU networks

This appendix places the fully connected ReLU network, the architecture on which most of the theory of depth has been developed (Section B), in the state-transition model, and shows that the Lipschitz constant of the layers alone moves the network across the growth mechanisms of Section 4: contractive layers saturate, non-expanding layers on a bounded state space saturate as well, and expanding layers can realize exponential growth. The ReLU examples in the literature are then read as instances of the resulting regimes, and a computed illustration (Figure 3) shows the three growth behaviours for three two-layer families that differ only in their Lipschitz constant.

### K.1 Setting

Let $K \subset \mathbb{R}^m$ be a nonempty bounded set of diameter $D_K$, for instance $K = [-1, 1]^m$, let $R_K \ge 0$ be a bound with $\|x\| \le R_K$ for all $x \in K$, and let $\Pi_K$ be a 1-Lipschitz retraction of $\mathbb{R}^m$ onto $K$, such as the coordinatewise clipping for a box. A hidden layer is a ReLU block of width $w$ that maps the state space to itself,

$$f_\vartheta(x) := \Pi_K\big(V\,\mathrm{ReLU}(Wx + b) + c\big), \qquad \vartheta = (W, b, V, c) \in \mathbb{R}^{w\times m} \times \mathbb{R}^w \times \mathbb{R}^{m\times w} \times \mathbb{R}^m,$$

where ReLU acts coordinatewise, and the hidden-layer class is

$$F_\Lambda := \{f_\vartheta :\ \|W\|_{\mathrm{op}}, \|V\|_{\mathrm{op}} \le \beta_W,\ \|b\|, \|c\| \le \beta,\ \mathrm{Lip}(f_\vartheta) \le \Lambda\},$$

with $p := 2mw + w + m$ parameters, where $\beta_W, \beta, \Lambda \geq 0$. The retraction keeps the state space bounded, which bounded activations or normalization layers achieve in practice, and it is the identity whenever the block output stays in $K$. The output layer is linear, $H = \{x \mapsto \langle u, x\rangle : \|u\| \leq R\}$, so Assumption 2 holds with depth-independent constants (Proposition 1), and a depth-$k$ network $h \circ f_{\vartheta_k} \circ \cdots \circ f_{\vartheta_1}$ is a fully connected ReLU network of depth $2k$ with clipped states of dimension $m$ after every second layer; the Lipschitz constraint restricts the class to blocks of Lipschitz constant at most $\Lambda$ (the product $\|V\|_{\mathrm{op}}\|W\|_{\mathrm{op}}$ is always an upper bound for it).

**Lemma 11** (Covering numbers of one ReLU block)**.** *For $f_\vartheta, f_{\vartheta'} \in F_\Lambda$,*

$$d_\infty(f_\vartheta, f_{\vartheta'}) \leq \beta_W R_K \|W - W'\|_{\mathrm{op}} + \beta_W \|b - b'\| + (\beta_W R_K + \beta)\|V - V'\|_{\mathrm{op}} + \|c - c'\|,$$

*and hence, for every $\varepsilon > 0$, $\log N(F_\Lambda, d_\infty, \varepsilon) \leq p \log(1 + C_F/\varepsilon)$ with $C_F := 2\,(2\beta_W R_K + \beta_W + \beta + 1)\max\{\beta_W, \beta\}$; in particular $\log N(F_\Lambda, d_\infty, \varepsilon) \leq p \log(1/\varepsilon) + O(p)$ for $\varepsilon \leq C_F$.*

*Proof.* For $x \in K$, write $z = \mathrm{ReLU}(Wx + b)$ and $z' = \mathrm{ReLU}(W'x + b')$; then $\|z - z'\| \leq R_K\|W - W'\|_{\mathrm{op}} + \|b - b'\|$ and $\|z'\| \leq \beta_W R_K + \beta$. Hence $\|Vz + c - V'z' - c'\| \leq \|V\|_{\mathrm{op}}\|z - z'\| + \|V - V'\|_{\mathrm{op}}\|z'\| + \|c - c'\|$, which gives the displayed bound after applying the 1-Lipschitz retraction. The parameter tuple lives in a real normed space of dimension $p$ (operator norms on $W, V$, Euclidean norms on $b, c$, the maximum over the four blocks), in which the parameter set is contained in the ball of radius $\max\{\beta_W, \beta\}$, and the displayed estimate shows that $\vartheta \mapsto f_\vartheta$ is $(2\beta_W R_K + \beta_W + \beta + 1)$-Lipschitz from this norm to $d_\infty$. A ball of radius $\rho$ in a real normed space of dimension $p$ has a $\delta$-separated subset of at most $(1 + 2\rho/\delta)^p$ points (compare the volumes of disjoint balls of radius $\delta/2$), hence a $\delta$-cover of that size; the image of such a cover under the Lipschitz map is an $\varepsilon$-cover of $F_\Lambda$ when $\delta = \varepsilon/(2\beta_W R_K + \beta_W + \beta + 1)$, which gives the stated $C_F$. □

### K.2 Variance: three growth regimes from the Lipschitz constant

**Proposition 8** (Depth profiles of ReLU networks by Lipschitz constant)**.** *Let $F = F_\Lambda$ on the bounded set $K$, and let $\mathsf{V}_k(S)$ be the entropy integral of Theorem 1.*

- *(i) (Contractive, $0 < \Lambda < 1$.) Condition P1' applies with the invariant set $K$ itself, so $\sup_k N(B(k, F_\Lambda), d_\infty, \varepsilon) < \infty$ for every $\varepsilon > 0$, and $\mathsf{V}_k(S) = O(1)$. Explicitly, with $m(\varepsilon) := \lceil \log(2D_K/\varepsilon)/\log(1/\Lambda) \rceil$, $\log N(B(k, F_\Lambda), d_\infty, \varepsilon) \leq \log\big(N(K, \varepsilon/2) + m(\varepsilon) N(F_\Lambda, d_\infty, (1-\Lambda)\varepsilon)^{m(\varepsilon)}\big)$, which does not depend on $k$.*
- *(ii) (Non-expanding, $\Lambda \leq 1$.) If $K$ is compact, Condition P1 applies (non-expanding generators on a compact state space), so again $\sup_k N(B(k, F_\Lambda), d_\infty, \varepsilon) < \infty$ and $\mathsf{V}_k(S) = O(1)$; the explicit layerwise bound of Theorem 2 gives the cruder but quantitative bound $\mathsf{V}_k(S) \leq D_K\big(\sqrt{\log(k+1)} + \sqrt{kp \log k} + \sqrt{kp}\,(\sqrt{\log(1 + 2C_F/D_K)} + \sqrt{\pi}/2)\big) = O(\sqrt{kp \log k})$ for $k \geq 1$, which needs only boundedness of $K$.*
- *(iii) (Expanding, $\Lambda > 1$.) Theorem 2 gives $\mathsf{V}_k(S) \leq D_K\big(\sqrt{\log(k+1)} + \sqrt{kp \log k} + k\sqrt{p \log \Lambda} + \sqrt{kp}\,(\sqrt{\log(1 + 2C_F/D_K)} + \sqrt{\pi}/2)\big) = O(k\sqrt{p \log \Lambda})$, and this order cannot be improved in general: for $m = 1$, $K = [0, 1]$ with the clipping retraction, width $w \geq 4$, $\beta_W \geq 41$, $\beta \geq 2$, and every $\Lambda \geq 20$, the class $F_\Lambda$ contains the two piecewise-linear expand-and-reset maps of Section K.4, which satisfy Condition E2 with two generators, so $N(B(k, F_\Lambda), d_\infty, \varepsilon) \geq 2^k$ for all $\varepsilon < 1/16$ and $\mathsf{V}_k(S)$ grows at least like $\sqrt{k}$ on a sample that contains the probe points (Lemma 7).*

*Proof.* (i) Every map in $F_\Lambda$ is $\Lambda$-Lipschitz with $\Lambda < 1$ and maps $K$ into $K$, so $K$ is a nonempty invariant set and a bounded absorbing set with $L = 0$; Condition P1' gives the bound with $N(F_\Lambda^\ell, d_\infty, \varepsilon) \leq N(F_\Lambda, d_\infty, (1-\Lambda)\varepsilon)^\ell$ by the telescoping estimate of Theorem 2, and the right-hand side is finite by Lemma 11. (ii) The maps are non-expanding self-maps of the compact set $K$, which is condition 2c of Condition P1. The layerwise bound statements in (ii) and (iii) are Corollary 9(a) with Lemma 11; since the lemma bounds the entropy by $p \log(1 + C_F/\varepsilon)$ rather than by $p \log(C_F/\varepsilon)$, the proof of the corollary is repeated with this hypothesis, which changes $\log(2C/\overline{D})$ into $\log(1 + 2C_F/D_K)$ and nothing else. For the lower bound, a continuous piecewise-linear map $[0, 1] \to [0, 1]$ with $q$ interior breakpoints is a ReLU block $x \mapsto V\,\mathrm{ReLU}(Wx + b) + c$ of width

$q + 1$, and the two maps of Section K.4 have Lipschitz constant 20, are ReLU blocks of width 4 with weight norms at most 41 and bias norms at most 2, and satisfy Condition E2 with separation $1/8$, as verified there; the entropy lower bound is Condition E2, and the transfer to the sample is Lemma 7. □

The three cases are the three rows of the balance in Section 5 on the variance side. Contractive and non-expanding ReLU layers on a bounded state space have a saturated profile, so their estimation bound does not grow with the depth; expanding layers have a polynomial profile at best, between $\sqrt{k}$ and $k$, so their estimation bound grows with the depth and has to be balanced against the bias. The dependence on the width is through $p = 2mw + w + m$ in the explicit bounds and through the one-layer covering numbers in the saturated ones.

### K.3 Bias and balancing

The bias depends on the target class, and the ReLU literature provides three cases. For a teacher–student target, $\mathcal{C} = H \circ \overline{\langle F_\Lambda \rangle}$ with $\Lambda < 1$, the truncation argument of Section J gives $\mathsf{bias}(k) \leq \beta_\ell R D_K \Lambda^k$, and with case (i) the network is in the EL regime: a contractive ReLU network may be run to convergence. For Hölder, Sobolev, or Besov targets, the approximation error of depth-$k$ ReLU networks decays polynomially in the number of parameters and hence in the depth (Yarotsky, 2017; Siegel, 2023; Suzuki, 2019), so the network is in the PP regime with the profile of case (ii) or (iii): this is the setting of the VC-dimension and nonparametric analyses reviewed in Section B, whose polynomial depth dependence is the layerwise bound of Theorem 2. For the hierarchical compositions of Schmidt-Hieber (2020), the approximation error has an exponentially decaying compositional part and the covering-number bounds grow logarithmically in the depth, close to the PL regime. In each case the framework does not change the known rate; it identifies which of the three mechanisms of Proposition 8 the corresponding proof uses, and the Lipschitz constant of the layers as the quantity that decides between them.

### K.4 Computed illustration (Figure 3)

Figure 3 is produced by the script `pics/figure-relu-example.py` (NumPy and Matplotlib; no randomness). The state space is $K = [0, 1]$ with the clipping retraction, and each family consists of two one-hidden-layer ReLU maps $[0, 1] \to [0, 1]$; the three families differ only in their Lipschitz constant.

- Contractive ($\Lambda = 1/4$): $f_0(x) = x/4$ and $f_1(x) = (x + 3)/4$, the affine maps of a Cantor-type attractor, as in Example 3 with ratio $1/4$.
- Non-expanding ($\Lambda = 1$): $f_0(x) = \min\{1, \mathrm{ReLU}(x + \frac{1}{10})\}$ and $f_1(x) = \min\{1, \mathrm{ReLU}(x + \frac{1}{10\sqrt{2}})\}$, translations by two incommensurable positive steps followed by clipping; each is a ReLU network of width 2. Since both steps are positive, the compositions commute exactly, $f_1 \circ f_0 = f_0 \circ f_1 = \min\{1, \mathrm{ReLU}(x + \frac{1}{10} + \frac{1}{10\sqrt{2}})\}$, and a composition of $n_0$ steps of the first kind and $n_1$ of the second is the clipped translation by $n_0/10 + n_1/(10\sqrt{2})$: this is the abelian mechanism P2 with two generators, and the sums stay below 1 for $n_0 + n_1 \leq 9$.
- Expanding ($\Lambda = 20$): the piecewise-linear maps with breakpoints $(0, \frac{3}{8}), (\eta, 0), (\frac{1}{4} - \eta, 1), (\frac{1}{4}, \frac{3}{8}), (1, \frac{3}{8})$ and $(0, \frac{5}{8}), (\frac{3}{4}, \frac{5}{8}), (\frac{3}{4} + \eta, 0), (1 - \eta, 1), (1, \frac{5}{8})$, $\eta = 1/32$. Each is affine from its coding core $V_0 = [\eta, \frac{1}{4} - \eta]$, respectively $V_1 = [\frac{3}{4} + \eta, 1 - \eta]$, onto $[0, 1]$, and constant equal to its anchor $3/8$, respectively $5/8$, outside its chamber $U_0 = [0, \frac{1}{4}]$, respectively $U_1 = [\frac{3}{4}, 1]$. This is Condition E2 with the marker $q = 1/2$: the chambers are disjoint, $f_i(V_i) = [0, 1]$, every point outside $U_i$ is reset to the anchor $a_i$, the anchors lie outside both chambers, and $\mathrm{dist}(q, \{a_0, a_1\}) = 1/8$. Each map is a ReLU block of width 4 (Proposition 8(iii)).

The left three panels show the graphs of all eight compositions of depth $k = 3$ of each family: contractive compositions are almost constant, non-expanding compositions are clipped translations, and expanding compositions are separated by their behaviour on nested cores of width $(3/16)^k$.

The right panel counts the compositions of depth at most $k$ that are distinguishable at resolution $\varepsilon = 10^{-2}$, that is, the size of a greedy $\varepsilon$-net of $B(k, F)$ in the sup distance over a finite evaluation set, for $k \leq 9$. For the contractive and the non-expanding family the evaluation set is the uniform grid of $4001$ points in $[0, 1]$. For the expanding family the grid alone cannot resolve the coding cores beyond depth $5$ (their width is $(3/16)^k$), so the evaluation set is the grid together with the probe points $x_w$ of every word $w$ of length at most $9$, computed as $x_w = g_{i_1} \circ \cdots \circ g_{i_k}(q)$ with $g_i$ the inverse of $f_i$ on $V_i$, at which $f_w(x_w) = q$ while every other word of the same length maps $x_w$ to an anchor. On this set the count equals $2^{k+1} - 1$, the number of words, as guaranteed by Condition E2; on the grid alone it would follow $2^{k+1} - 1$ up to depth $5$ and then saturate, which is the empirical-entropy effect of Section L.4. The three families realize the three growth cases of Proposition 8 and of Section 4:

- the contractive count is $1, 3, 7, 15, 31$ for $k \leq 4$ and stays at $31$ afterwards, since compositions of depth $\geq 5$ are within $4^{-5} < \varepsilon$ of constant maps; the level is the covering number of the attractor at resolution $\varepsilon$, a property of the attractor and not of the depth (case (i), P1');
- the non-expanding count is $1, 3, 6, 10, 15, 21, 28, 35, 42, 49$: the triangular numbers $(k + 1)(k + 2)/2$ up to $k = 6$, that is, one map per pair $(n_0, n_1)$ with $n_0 + n_1 \leq k$, and slightly fewer afterwards because some sums fall within $\varepsilon$ of each other; this is the polynomial fixed-scale growth of the abelian mechanism P2 (case (ii), which by P1 saturates once the sums exceed $1$);
- the expanding count grows like $2^k$ in the uniform distance (case (iii), E2).

### K.5 Implementation and literature

The hidden-layer class $F_\Lambda$ is itself a class of ReLU blocks, so the implementation map is the identity and $\varepsilon_{\text{imp}}(k) = 0$; the implementation error becomes relevant only when a ReLU block approximates a layer of another type, which is the situation of Proposition 6(b). The depth dependences established for ReLU networks in the literature, and their place among the mechanisms of this paper, are reviewed in Section B.

## L Worked example: chain-of-thought style symbolic computation

This appendix applies the framework to a stylized model of step-by-step symbolic computation, in which each hidden layer is one reasoning step acting on a scratchpad and the depth $k$ is the number of steps. The purpose is to show that the statistical cost of the number of steps can be read off from the metric growth of the compositions of steps, separately from any particular implementation (for example a recurrently applied transformer block). Two step families are analyzed: an append-only scratchpad, which is contractive and lands in the EL regime, and a branching computation that reads the current symbol, which is expansive, has exponentially growing uniform entropy, and yet has root-logarithmic empirical entropy on every sample, because a sample of size $n$ can exercise only finitely many branches. Throughout, "chain of thought" refers to this abstract model and not to a specific language model; the relation to the literature on chain-of-thought reasoning is discussed at the end.

### L.1 Setting

**Scratchpad space.** Let $\mathcal{A}$ be a finite alphabet with $|\mathcal{A}| = m \geq 2$ and let $\mathcal{X} = \mathcal{A}^{\mathbb{N}} = \{x = (x_0, x_1, \dots) : x_j \in \mathcal{A}\}$ be the space of infinite symbol sequences, equipped with the ultrametric

$$d_\theta(x, y) = \theta^{n(x,y)} \quad (x \neq y), \qquad n(x, y) := \min\{j \geq 0 : x_j \neq y_j\},$$

for a fixed $\theta \in (0, 1)$, and $d_\theta(x, x) = 0$. Then $(\mathcal{X}, d_\theta)$ is compact with $\operatorname{diam}(\mathcal{X}) = 1$, and two sequences are within distance $\theta^L$ if and only if they agree in the first $L$ symbols (see Section F for the standard facts used below). A finite scratchpad of length $N$ can be embedded by padding with a fixed symbol.

**Output layers.** The answer is read from the scratchpad through an output layer. We use two linear output-layer classes on explicit feature maps, so that Assumption 2 holds with depth-independent constants by Proposition 1.

- (Window output layer.) For $L \geq 1$, let $\Phi_L : \mathcal{X} \to \mathbb{R}^{\mathcal{A}^L}$ be the one-hot encoding of the window $(x_0, \dots, x_{L-1})$, and $H_L := \{x \mapsto \langle w, \Phi_L(x)\rangle : \|w\|_2 \leq R\}$.
- (Geometrically weighted output layer.) Let $\Phi_\infty : \mathcal{X} \to \ell^2(\mathbb{N}; \mathbb{R}^{\mathcal{A}})$, $\Phi_\infty(x) := (\theta^j e_{x_j})_{j\geq 0}$, where $e_a$ is the one-hot vector of $a \in \mathcal{A}$, and $H_\infty := \{x \mapsto \langle w, \Phi_\infty(x)\rangle : \|w\|_{\ell^2} \leq R\}$.

**Lemma 12.** *$\Phi_L$ is $\sqrt{2}\,\theta^{1-L}$-Lipschitz and $\Phi_\infty$ is $\sqrt{2/(1-\theta^2)}$-Lipschitz with respect to $d_\theta$, and $\|\Phi_L\| \leq 1$, $\|\Phi_\infty\| \leq (1-\theta^2)^{-1/2}$. Consequently, Assumption 2 holds for every hidden-layer class on $\mathcal{X}$, with $A_H = 1$ and $L = \sqrt{2}\,\theta^{1-L}$ for $H_L$, and with $A_H = 1$ and $L = \sqrt{2/(1-\theta^2)}$ for $H_\infty$; moreover $\hat{\mathfrak{R}}_S(H_L) \leq R/\sqrt{n}$ and $\hat{\mathfrak{R}}_S(H_\infty) \leq R/\sqrt{n(1-\theta^2)}$.*

*Proof.* If $d_\theta(x,y) \leq \theta^L$, the first $L$ symbols agree and $\Phi_L(x) = \Phi_L(y)$. Otherwise $d_\theta(x,y) \geq \theta^{L-1}$ and $\|\Phi_L(x) - \Phi_L(y)\| \leq \sqrt{2} \leq \sqrt{2}\,\theta^{1-L} d_\theta(x,y)$. For $\Phi_\infty$, with $n = n(x,y)$, $\|\Phi_\infty(x) - \Phi_\infty(y)\|^2 = \sum_{j\geq n} \theta^{2j}\|e_{x_j} - e_{y_j}\|^2 \leq 2\theta^{2n}/(1-\theta^2)$, so $\|\Phi_\infty(x) - \Phi_\infty(y)\| \leq \sqrt{2/(1-\theta^2)}\, d_\theta(x,y)$. The norm bounds are immediate. The increment condition follows from Proposition 1, and the Rademacher bounds from $\hat{\mathfrak{R}}_S(H) = R\,\mathbb{E}_\sigma\|n^{-1}\sum_i \sigma_i \Phi(X_i)\| \leq R \sup\|\Phi\|/\sqrt{n}$. □

**Loss and learning problem.** Labels are real-valued, the loss is bounded and $\beta_\ell$-Lipschitz as in Section 2, and the learner minimizes the empirical risk over $\mathcal{H}_k = H \circ B(k, F)$, that is, over output layers composed with programs of at most $k$ steps.

### L.2 Append-only steps: saturation and the EL regime

**Steps.** A write step prepends a symbol: $g_a(x) := (a, x_0, x_1, \dots)$ for $a \in \mathcal{A}$, and $F_{\mathrm{w}} := \{g_a : a \in \mathcal{A}\}$. For a word $u = (u_1, \dots, u_j) \in \mathcal{A}^j$ write $g_u := g_{u_j} \circ \cdots \circ g_{u_1}$, so that $g_u(x) = (u_j, u_{j-1}, \dots, u_1, x_0, x_1, \dots)$: the scratchpad records the steps in reverse order and the input is pushed back. Each $g_a$ satisfies $d_\theta(g_a(x), g_a(y)) = \theta\, d_\theta(x,y)$, so $F_{\mathrm{w}}$ is a contractive hidden-layer class with $\mathrm{Lip}(g_a) = \theta < 1$, and Condition P1 applies.

**Lemma 13** (Saturation for append-only steps)**.** *For $\varepsilon > 0$ let $\ell(\varepsilon) := \max\{0, \lceil \log(1/\varepsilon)/\log(1/\theta)\rceil\}$. Then for all $k \geq 0$ and all $\varepsilon > 0$,*

$$N\big(B(k, F_{\mathrm{w}}), d_\infty, \varepsilon\big) \leq m^{\min\{k, \ell(\varepsilon)\}+1}, \qquad \mathsf{V}_k(S) \leq \sqrt{\log m}\Big(\frac{\sqrt{\pi}}{2\sqrt{\log(1/\theta)}} + \sqrt{2}\Big) =: \mathsf{V}_\infty.$$

*In particular $\mathsf{var}(k,n) \leq 12 A_H L\, \mathsf{V}_\infty/\sqrt{n}$ for every $k$ (case (i) of Proposition 5).*

*Proof.* If $|u| \geq \ell$ and $u'$ consists of the last $\ell$ letters of $u$, then $g_u(x)$ and $g_{u'}(x)$ agree in the first $\ell$ symbols for every $x$, so $d_\infty(g_u, g_{u'}) \leq \theta^\ell$. With $\ell = \ell(\varepsilon)$ we have $\theta^\ell \leq \varepsilon$, hence the compositions of length at most $\min\{k, \ell(\varepsilon)\}$ form an $\varepsilon$-cover of $B(k, F_{\mathrm{w}})$, and there are at most $\sum_{j \leq \min\{k,\ell(\varepsilon)\}} m^j \leq m^{\min\{k,\ell(\varepsilon)\}+1}$ of them. Since $\mathrm{diam}(\mathcal{X}) = 1$ and $d_S \leq d_\infty$,

$$\mathsf{V}_k(S) \leq \int_0^1 \sqrt{(\ell(\varepsilon)+1)\log m}\,\mathrm{d}\varepsilon \leq \sqrt{\log m}\int_0^1 \Big(\sqrt{\tfrac{\log(1/\varepsilon)}{\log(1/\theta)}} + \sqrt{2}\Big)\mathrm{d}\varepsilon,$$

using $\ell(\varepsilon) + 1 \leq \log(1/\varepsilon)/\log(1/\theta) + 2$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. The identity $\int_0^1 \sqrt{\log(1/\varepsilon)}\,\mathrm{d}\varepsilon = \sqrt{\pi}/2$ (Lemma 9) gives the claim. □

**Bias.** Take the target class to be the teacher–student class of arbitrarily long programs with the geometrically weighted output layer, $\mathcal{C} := H_\infty \circ \overline{\langle F_{\mathrm{w}}\rangle}^{\,d_\infty}$. For a teacher $c = h_w \circ g_u \circ g_v$ with $|u| = k$ and any $v$, Lemma 12 and $\mathrm{Lip}(g_u) = \theta^k$ give

$$|c(x) - h_w(g_u(x))| \leq R\sqrt{\tfrac{2}{1-\theta^2}}\; d_\theta\big(g_u(g_v(x)), g_u(x)\big) \leq R\sqrt{\tfrac{2}{1-\theta^2}}\;\theta^k,$$

and the same bound holds on the uniform closure. By (1), $\mathsf{bias}(k) = \varepsilon_{\text{model}}(k) \leq \beta_\ell R\sqrt{2/(1-\theta^2)}\,\theta^k$: steps performed long ago contribute geometrically less to the answer.

**Proposition 9** (Append-only scratchpad)**.** *In the setting above, with $H = H_\infty$ and $F = F_{\mathrm{w}}$, the excess-risk bound of Theorem 4 is, up to the depth-independent terms,*

$$\mathsf{gen}(k,n) \lesssim \beta_\ell R\sqrt{\tfrac{2}{1-\theta^2}}\,\theta^k + \frac{12\sqrt{2/(1-\theta^2)}\,\mathsf{V}_\infty}{\sqrt{n}},$$

*which is the EL regime with $\alpha = \log(1/\theta)$ and saturated variance. The bound is minimized at $k^* = \frac{\log n}{2\log(1/\theta)} + O(1)$, with balanced value of order $n^{-1/2}$.*

*Proof.* Combine Theorems 1 and 4, Lemma 12, and Lemma 13; the balancing step is that of the EL row of (22) with a constant variance profile, which removes the $\log\log$ factors. □

With the window output layer $H_L$ instead, the bias vanishes exactly for $k \geq L$, because the answer depends only on the last $L$ writes; then $k^* = L$ and additional steps cost nothing in the bound. In both cases the number of useful steps is small, of order $\log n$ or a constant, and further steps do not increase the estimation bound: append-only scratchpads are forgetful at the output layer, and this forgetting is exactly the contraction that saturates the entropy.

### L.3 Branching computation: exponential uniform entropy, logarithmic empirical entropy

**Steps.** Let $\mathcal{A} = [r] \cup \{\bullet\}$ with $r \geq 2$ active symbols and one padding symbol. A guarded step $f_a$, $a \in [r]$, reads the current symbol: if it equals $a$ the step consumes it and exposes the rest of the scratchpad, $f_a(x) = \sigma(x) = (x_1, x_2, \dots)$; otherwise the step fails and resets the scratchpad to the constant error state $\bar{a} = (a, a, \dots)$. Let $F_{\mathrm{b}} := \{f_a : a \in [r]\}$. This is the ping–pong family of Example 6, which satisfies Condition E2 with separation 1: on its chamber each $f_a$ is the shift, with Lipschitz constant $\theta^{-1} > 1$, so $F_{\mathrm{b}}$ is expansive rather than contractive. A program $u = (u_1, \dots, u_j)$ succeeds on exactly those inputs whose first $j$ symbols are $u_1, \dots, u_j$, and then outputs the remainder of the scratchpad. Explicitly, writing $[u]$ for the cylinder set of inputs starting with $u$ and $f_u := f_{u_j} \circ \cdots \circ f_{u_1}$ for $j \geq 1$ ($f_{()} = \mathrm{id}$),

$$f_u(x) = \sigma^j(x) \ \text{ if } x \in [u], \qquad f_u(x) = \bar{u}_j \ \text{ otherwise,} \tag{23}$$

since once a guard fails the scratchpad is the constant state of the failed symbol, and each later guard either keeps it (if its symbol agrees) or replaces it by its own constant state, so the final state is $\bar{u}_j$.

**Lemma 14** (Exponential growth for branching steps)**.** *For every $k \geq 0$ and $\varepsilon \in (0, 1/2)$,*

$$r^k \leq N\big(B(k, F_{\mathrm{b}}), d_\infty, \varepsilon\big) \leq r^{k+1}, \qquad \mathsf{V}_k(S) \leq \sqrt{(k+1)\log r},$$

*and hence $\mathsf{var}(k,n) \leq 12 A_H L\sqrt{(k+1)\log r/n}$ (case (iii) of Proposition 5).*

*Proof.* The lower bound is Example 6. The upper bound holds because $F_{\mathrm{b}}$ is finite, so $B(k, F_{\mathrm{b}})$ has at most $\sum_{j \leq k} r^j \leq r^{k+1}$ elements (using $r \geq 2$), and $\mathrm{diam}(\mathcal{X}) = 1$ (Proposition 5(iii)). □

**Empirical entropy: what a sample can see.** The exponential growth in Lemma 14 is certified in the uniform metric, and Example 6 shows that it is real. The entropy integral of Theorem 1, however, is taken in the empirical metric $d_S$, and for branching computation the two metrics differ by an exponential factor, for every sample. The reason is (23): programs of the same length have pairwise disjoint success cylinders, so a sample of size $n$ meets at most $n$ of them, and every program whose cylinder contains no sample point is indistinguishable on the sample from a constant map.

**Lemma 15** (Empirical saturation for branching steps)**.** *For every sample $S = (X_1, \dots, X_n)$, every $k \geq 0$, and every $\varepsilon > 0$,*

$$N\big(B(k, F_{\mathrm{b}}), d_S, \varepsilon\big) \leq 1 + r + \sum_{j=1}^{k} \min\big\{r^j,\ n,\ \lfloor \varepsilon^{-2} \rfloor\big\} \leq 1 + r + k\min\{n, \varepsilon^{-2}\},$$

*and consequently*

$$\mathsf{V}_k(S) \le \sqrt{\log(k+1)} + \sqrt{\log(1+r)} + \sqrt{2\log 2} + \sqrt{\pi/2},$$
$$\mathsf{var}(k,n) \le \frac{12A_H L}{\sqrt{n}}\Big(\sqrt{\log(k+1)} + \sqrt{\log(1+r)} + \sqrt{2\log 2} + \sqrt{\pi/2}\Big),$$

*the root-logarithmic profile (case (ii) of Proposition 5), for every sample and without any assumption on the input distribution. (The constant $\sqrt{2\log 2}$ comes from the internal/external transfer $N(B, d_S, \varepsilon) \le N^{\mathrm{ext}}(B, d_S, \varepsilon/2)$, since $\mathsf{V}_k(S)$ is formed with internal covering numbers.)*

*Proof.* Let $\hat{P}_n([u]) := n^{-1}\#\{i : X_i \in [u]\}$ be the empirical mass of the success cylinder of $u$. By (23) and $\mathrm{diam}(\mathcal{X}) = 1$,

$$d_S(f_u, \bar{u}_j)^2 = \frac{1}{n}\sum_{i:X_i\in[u]} d_\theta\big(\sigma^j(X_i), \bar{u}_j\big)^2 \le \hat{P}_n([u]).$$

For fixed $j$ the cylinders $[u]$, $|u| = j$, are pairwise disjoint, hence $\sum_{|u|=j}\hat{P}_n([u]) \le 1$; therefore fewer than $\varepsilon^{-2}$ programs of length $j$ have $\hat{P}_n([u]) > \varepsilon^2$, and also at most $n$ programs of length $j$ have $\hat{P}_n([u]) > 0$. Every other program of length $j$ is within $d_S$-distance $\varepsilon$ of one of the $r$ constant maps $\bar{a}$, $a \in [r]$. Taking as centres the identity, the $r$ constant maps, and, for each $1 \le j \le k$, the at most $\min\{r^j, n, \lfloor\varepsilon^{-2}\rfloor\}$ programs with $\hat{P}_n([u]) > \varepsilon^2$, gives an external $\varepsilon$-cover of $B(k, F_{\mathrm{b}})$ of the stated cardinality. For the entropy integral, the internal covering number at radius $\varepsilon$ is at most the external one at radius $\varepsilon/2$, which is at most $1 + r + 4k\varepsilon^{-2} \le (1+r)(k+1)\,4\varepsilon^{-2}$ for $\varepsilon \le 1$, so $\sqrt{\log N} \le \sqrt{\log(1+r)} + \sqrt{\log(k+1)} + \sqrt{2\log 2} + \sqrt{2\log(1/\varepsilon)}$; integrate over $(0,1]$ (recall $D_k(S) \le \mathrm{diam}(\mathcal{X}) = 1$) and use $\int_0^1 \sqrt{\log(1/\varepsilon)}\,\mathrm{d}\varepsilon = \sqrt{\pi}/2$ (Lemma 9). □

The lemma does not contradict the lower bound of Lemma 14: the $r^k$ programs of length $k$ are 1-separated in $d_\infty$, but all except at most $n$ of them agree with a constant map on the sample. The output-layer separation condition of Corollary 3 (Assumption 4 with depth-independent $\kappa$) therefore fails for this family, and the $\sqrt{k/n}$ upper bound obtained from the uniform entropy is not sharp for any fixed sample size. This is the general phenomenon behind Theorem 1: the empirical entropy of $B(k, F)$, not its uniform entropy, is the statistical cost of depth, and the transfer $d_S \le d_\infty$ used in Section 4 is a one-sided device that may lose an exponential factor.

**Bias and balancing.** Two target models are natural.

- (Exact program.) If the target is $c = h \circ f_{u^*}$ for one program $u^*$ of length $k_0$ and $h \in H$, then $c \in \mathcal{H}_k$ for all $k \ge k_0$ and $\mathsf{bias}(k) = 0$ there. The bound of Theorem 4 is then minimized at $k = k_0$, with value of order $A_H L\sqrt{\log k_0/n} + R/\sqrt{n}$ by Lemma 15, and additional steps beyond $k_0$ cost only $\sqrt{\log k/n}$.
- (Geometric relevance.) If longer programs matter geometrically less, in the sense that $\mathsf{bias}(k) \le C_{\mathrm{b}}e^{-\alpha k}$ for some $C_{\mathrm{b}}, \alpha > 0$, then the bound is in the EL regime: by the EL row of (22), $k^* = \frac{1}{2\alpha}\log n + O(\log\log\log n)$ and the balanced value is of order $\sqrt{\log\log n/n}$. In contrast to the append-only case, this decay is an assumption on the target rather than a consequence of the step geometry, because the guarded steps do not contract.

**Proposition 10** (Branching computation)**.** *In the setting above, with $H = H_L$ or $H_\infty$ and $F = F_{\mathrm{b}}$, the excess-risk bound of Theorem 4 is, up to the depth-independent terms,*

$$\mathsf{gen}(k,n) \lesssim \mathsf{bias}(k) + \frac{12A_H L}{\sqrt{n}}\Big(\sqrt{\log(k+1)} + \sqrt{\log(1+r)} + \sqrt{2\log 2} + \sqrt{\pi/2}\Big)$$

*for every sample. With an exact program of length $k_0$ the bound is minimized at $k^* = k_0$; under geometric relevance it is in the EL regime with $k^* \asymp \log n$.*

### L.4 Computed illustration (Figure 4)

Figure 4 is produced by the script `pics/figure-cot-example.py` (NumPy and Matplotlib, random seed 0). The alphabet is $\mathcal{A} = \{1, 2, 3\} \cup \{\bullet\}$, so $r = 3$ active symbols and one padding symbol (drawn in grey); the ultrametric has $\theta = 1/2$; scratchpads are displayed through their first 12 symbols.

**Left panel (append-only steps).** The input $x$ is a uniformly random sequence of active symbols, and $y$ equals $x$ except in its first symbol. Both are run through the program $u = (1, 2, 3, 1, 2, 2)$ of Section L.2: after $k$ steps the scratchpad is $(u_k, \dots, u_1, x_0, x_1, \dots)$. Each pair of rows shows the two scratchpads after $k = 0, \dots, 6$ steps; the black bar marks $n(g_u(x), g_u(y))$, the first position at which they differ, which equals $k$. Hence $d_\theta(g_u(x), g_u(y)) = \theta^k$, the contraction behind Lemma 13, and the output layers of Section L.1 see the two states as $\theta^k$-close.

**Middle panel (branching steps).** The upper input starts with the six symbols of $u$ followed by random symbols, so every guarded step $f_{u_j}$ succeeds and consumes its symbol: after $k$ steps the scratchpad is $\sigma^k(x)$ padded with $\bullet$ at the end. The lower input equals the upper one except in its third symbol, so the third step fails and resets the scratchpad to the constant state $\bar{u}_3$, which the remaining steps turn into $\bar{u}_k$ as in (23). The black bar marks the first difference between the two rows; it stays at position 0 once the lower input has failed, so the two states are at distance 1, which is the separation used in Lemma 14.

**Right panel (distinguishable programs on a sample).** A sample of $n = 200$ uniformly random sequences of active symbols of length 12 is drawn. For each $k \le 10$, all programs of length at most $k$ are enumerated ($\sum_{j \le k} 3^j$ programs, the dashed curve), and two programs are identified when their outputs agree in the first $\ell = 3$ symbols at every sample point, that is, when their sample restrictions are within $\theta^3$ in the metric $\max_i d_\theta(f(X_i), g(X_i))$. Since the metric is an ultrametric, the classes of this relation are the balls of radius $\theta^3$, so the number of classes is the covering number of $B(k, F)$ at resolution $\theta^3$ in this sample metric, which dominates the covering number in $d_S$. The count is computed exactly by encoding the first three output symbols at every sample point as one integer and counting distinct rows. For branching steps (red squares) the count follows the uniform count $3^k$ while $3^k \le n$ and then grows linearly, by about $n$ per unit of depth, as predicted by Lemma 15: programs whose success cylinder contains no sample point restrict to one of the $r$ constant maps. For append-only steps (blue circles) the count saturates at $\sum_{j \le 3} 3^j = 40$, because programs longer than $\ell = 3$ agree with a shorter one in the first three symbols (Lemma 13). For $k = 10$ the three counts are 88573, 1184, and 40.

### L.5 Implementation

Each step is a fixed map, so a recurrently applied block that implements the step family realizes $B(k, F)$ with the same block at every depth; the implementation error of an approximate block is controlled by Proposition 6(b). For append-only steps the propagation constant is $\Lambda = \theta < 1$, so a per-step error $\delta$ yields $\varepsilon_{\text{imp}}(k) \le \delta_H + L_H \delta/(1 - \theta)$, uniformly in $k$. For branching steps the propagation constant on the chambers is $\Lambda = \theta^{-1} > 1$, and the layerwise bound grows like $\delta\theta^{-k}$; an exact implementation of the guard is required for long programs. Thus the two step families differ numerically, forgetting is stable and branching is not, while statistically both are benign on any finite sample: the price of branching is the root-logarithmic profile of Lemma 15 rather than the saturated one, together with the need for an exact implementation of the guard. In summary, both step families act on the same scratchpad space with fixed linear output layers, so the output-layer constants are depth-independent, and only the growth of $B(k, F)$ differs. For append-only steps, $B(k, F_{\text{w}})$ saturates in $d_\infty$ and in $d_S$ (Condition P1), $\mathsf{var}(k, n) \asymp n^{-1/2}$, the bias is at most $\theta^k$ (proved), and the regime is EL with $k^* \asymp \log n$. For branching steps, $r^k \le N(B(k, F_{\text{b}}), d_\infty, \varepsilon) \le r^{k+1}$ (Condition E2) but $N(B(k, F_{\text{b}}), d_S, \varepsilon) \le 1 + r + k \min\{n, \varepsilon^{-2}\}$ on every sample (Lemma 15), so $\mathsf{var}(k, n) \asymp \sqrt{\log k / n}$; the bias is 0 for $k \ge k_0$ (proved) or $e^{-\alpha k}$ (assumed), and the balancing depth is $k_0$, or $\asymp \log n$ in the EL regime.

### L.6 Relation to the literature

Chain-of-thought prompting (Wei et al., 2022) lets a language model produce intermediate steps before the answer. Expressivity analyses show that allowing a number of intermediate steps enlarges the class of problems a transformer can solve, in particular inherently serial ones (Li et al., 2024), and formal frameworks for length generalization (Huang et al., 2025) and generalization analyses for training transformers with chain of thought (Li et al., 2025) address specific architectures. The present example is complementary and does not concern a particular architecture: it isolates the statistical cost of the number of steps as the growth of the compositions of step maps, and shows that this cost is saturated for forgetful scratchpads and root-logarithmic in the number of steps for branching computation, independently of how the steps are implemented; the exponential uniform entropy of branching computation is not paid on any finite sample.

## M Worked example: unrolled iterative solvers and samplers

Many deep models are unrolled iterations: fixed-point solvers, iterative denoisers, and numerical integrators of ordinary differential equations (ODEs), including the samplers of score-based diffusion and flow-matching models. In each case a hidden layer is one iteration and the depth is the number of iterations. This appendix instantiates the framework for two such families. The first, fixed-point refinement, is contractive and lands in the EL regime with saturated variance. The second, fixed-horizon integration with a step size that decreases with the depth, is not contractive but stable, and lands in the PL regime; its variance saturates by an Arzelà–Ascoli argument. The example is summarized as follows.

*Example* 8 (Unrolled iterations). The state space is a compact convex $K \subset \mathbb{R}^d$, a hidden layer is one projected step $x \mapsto \Pi_K(x + h\, s(x))$ with a vector field $s$ from a class $\mathcal{S}$, and the output layer is linear. For strongly monotone fields the profile saturates and the bias decays geometrically, so $k^* \asymp \log n$ and further iterations do not increase the estimation bound (EL; Proposition 11). For Euler steps of a stable field over a fixed horizon the profile is $\sqrt{\log k/n}$ and the bias is the discretization error $k^{-p}$, so $k^* \asymp n^{1/(2p)}$ (PL; Proposition 12).

### M.1 Setting

Throughout, $K \subset \mathbb{R}^d$ is a compact convex set with diameter $D_K$, $\Pi_K$ is the Euclidean projection onto $K$, which maps $\mathbb{R}^d$ into $K$, is the identity on $K$, and is 1-Lipschitz, and the output-layer class is $H := \{x \mapsto \langle w, x\rangle : \|w\| \le R\}$ on $K$, for which Assumption 2 holds with $A_H = 1$ and $L = 1$ (Proposition 1) and $\hat{\mathfrak{R}}_S(H) \le R \sup_K \|x\|/\sqrt{n}$. Any fixed Lipschitz feature map may be inserted before the linear layer without changing the depth dependence.

### M.2 Fixed-point refinement: iterative denoising

**Layers.** Let $\mathcal{S}$ be a class of vector fields $s : K \to \mathbb{R}^d$ of the form $s = \nabla\phi$ with $\phi$ $\mu$-strongly concave and $\Lambda_s$-smooth on $K$, $0 < \mu \le \Lambda_s$ (for instance the score $\nabla \log p$ of a strongly log-concave density, or the residual $D(x) - x$ of a denoiser $D$ satisfying the corresponding monotonicity conditions). Fix a step size $0 < h \le 2/(\mu + \Lambda_s)$ and let

$$F_{\mathrm{fp}} := \big\{T_s : x \mapsto \Pi_K\big(x + h\, s(x)\big) \;:\; s \in \mathcal{S}\big\}.$$

**Lemma 16.** *Every $T_s \in F_{\mathrm{fp}}$ is $\lambda$-Lipschitz on $K$ with $\lambda := 1 - h\mu < 1$. Consequently Condition P1 applies: $\sup_k N(B(k, F_{\mathrm{fp}}), d_\infty, \varepsilon) < \infty$ for every $\varepsilon > 0$, $D_k(S) \le D_K$, and $\mathsf{V}_k(S) \le \mathsf{V}_\infty < \infty$ for all $k$ whenever $\int_0^{D_K} \sqrt{\log N(\overline{\langle F_{\mathrm{fp}}\rangle}, d_\infty, \varepsilon)}\, \mathrm{d}\varepsilon < \infty$.*

*Proof.* Write $\Delta := x - y$ and $\Delta_s := s(x) - s(y)$. Since $-\phi$ is $\mu$-strongly convex and $\Lambda_s$-smooth, $s = \nabla\phi$ is $\mu$-strongly monotone, $\langle \Delta_s, \Delta\rangle \le -\mu\|\Delta\|^2$, and co-coercive (Nesterov, 2004, Theorem 2.1.12):

$$\langle \Delta_s, \Delta\rangle \le -\frac{\mu\Lambda_s}{\mu + \Lambda_s}\|\Delta\|^2 - \frac{1}{\mu + \Lambda_s}\|\Delta_s\|^2.$$

Strong monotonicity and Cauchy–Schwarz give $\|\Delta_s\| \geq \mu\|\Delta\|$. Hence, for $0 < h \leq 2/(\mu + \Lambda_s)$, so that $h^2 - 2h/(\mu + \Lambda_s) \leq 0$,

$$\begin{aligned}\|\Delta + h\Delta_s\|^2 &\leq \Big(1 - \frac{2h\mu\Lambda_s}{\mu + \Lambda_s}\Big)\|\Delta\|^2 + \Big(h^2 - \frac{2h}{\mu + \Lambda_s}\Big)\|\Delta_s\|^2 \\ &\leq \Big(1 - \frac{2h\mu\Lambda_s}{\mu + \Lambda_s} + \mu^2 h^2 - \frac{2h\mu^2}{\mu + \Lambda_s}\Big)\|\Delta\|^2 = (1 - h\mu)^2\|\Delta\|^2,\end{aligned}$$

and $\Pi_K$ is 1-Lipschitz; moreover $0 \leq 1 - h\mu < 1$ because $h\mu \leq 2\mu/(\mu + \Lambda_s) \leq 1$. The remaining claims are Condition P1 and Proposition 5(i). □

**Explicit entropy bound.** The saturation can be quantified through Condition P1' with $A = K$ and $m(\varepsilon) = \lceil \log(2D_K/\varepsilon)/\log(1/\lambda)\rceil$:

$$N\big(B(k, F_{\mathrm{fp}}), d_\infty, \varepsilon\big) \leq N(K, \varepsilon/2) + \sum_{j < m(\varepsilon)} N\big(F_{\mathrm{fp}}^j, d_\infty, \varepsilon\big),$$
$$N\big(F_{\mathrm{fp}}^j, d_\infty, \varepsilon\big) \leq N\big(F_{\mathrm{fp}}, d_\infty, (1-\lambda)\varepsilon\big)^j,$$

where the last inequality follows from the telescoping estimate of Proposition 7(b): replacing each of $j$ layers by an $(1-\lambda)\varepsilon$-close layer changes the composition by at most $(1-\lambda)\varepsilon \sum_{i<j} \lambda^i \leq \varepsilon$. Hence $\log N_\infty(\varepsilon) \lesssim d \log(D_K/\varepsilon) + m(\varepsilon) \log N(F_{\mathrm{fp}}, d_\infty, (1-\lambda)\varepsilon)$, and if the one-layer class is parametric with $\log N(F_{\mathrm{fp}}, d_\infty, \varepsilon) \leq p \log(C_F/\varepsilon)$, then $\log N_\infty(\varepsilon) \lesssim (d+p)\log^2(C/\varepsilon)$, whose square root is integrable at 0. Thus $\mathsf{V}_\infty \lesssim D_K\sqrt{d+p}$ up to logarithmic factors, independently of $k$.

**Bias.** Take the target class to be the teacher–student class $\mathcal{C} := H \circ \overline{\langle F_{\mathrm{fp}} \rangle}^{\,d_\infty}$, which contains in particular the maps $x \mapsto \langle w, x_s^* \rangle$ sending each input to the fixed point $x_s^*$ of $T_s$, the limit of the refinement. By the truncation argument of Section J, with $\mathrm{Lip}(h) \leq R$ and $\mathrm{Lip}(u) \leq \lambda^k$ for $u \in F_{\mathrm{fp}}^k$,

$$\sup_{c \in \mathcal{C}} \inf_{f \in \mathcal{H}_k} \|c - f\|_\infty \leq R\, D_K\, \lambda^k, \qquad \text{hence} \qquad \mathsf{bias}(k) \leq \beta_\ell R D_K (1 - h\mu)^k$$

by (1).

**Proposition 11** (Fixed-point refinement)**.** *In the setting above, the excess-risk bound of Theorem 4 is, up to depth-independent terms,* $\mathsf{gen}(k, n) \lesssim \beta_\ell R D_K (1 - h\mu)^k + 12\,\mathsf{V}_\infty/\sqrt{n}$*, the EL regime with* $\alpha = \log(1/(1-h\mu)) \geq h\mu$ *and saturated variance. It is minimized at* $k^* = \frac{\log n}{2\log(1/(1-h\mu))} + O(1) \leq \frac{\log n}{2h\mu} + O(1)$*, with balanced value of order* $n^{-1/2}$*.*

The number of refinement iterations should therefore grow logarithmically with the sample size, at a rate set by the contraction modulus $h\mu$, and additional iterations do not increase the estimation bound. This is the abstract content of the observation that unrolled denoisers and equilibrium models can be run to convergence without a statistical penalty (Bai et al., 2019; Romano et al., 2017); Taniguchi & Sonoda (2026) develop the same mechanism for Picard iterations of parabolic equations.

### M.3 Fixed-horizon integration: neural ODEs and diffusion samplers

**Layers.** Let $\mathcal{S}_T$ be a class of time-dependent vector fields $s : K \times [0, T] \to \mathbb{R}^d$ with $\|s\| \leq M_s$, $\Lambda_s$-Lipschitz in $x$, and $\Lambda_\tau$-Lipschitz in $\tau$. A layer is one explicit Euler step of size $h \in [0, h_0]$ at time stamp $\tau \in [0, T]$:

$$F_T := \big\{T_{s,h,\tau} : x \mapsto \Pi_K\big(x + h\, s(x, \tau)\big) \;:\; s \in \mathcal{S}_T,\ h \in [0, h_0],\ \tau \in [0, T]\big\},$$

and $\mathrm{id} = T_{s,0,\tau} \in F_T$. A *scheme* of at most $k$ steps is a composition $T_{s,h_k,\tau_k} \circ \cdots \circ T_{s,h_1,\tau_1}$ with a single drift $s$, consistent time stamps $\tau_i = \sum_{j<i} h_j$, and total time $\sum_i h_i \leq T$; let $B_T(k) \subset B(k, F_T)$ be the set of schemes of at most $k$ steps, together with $\mathrm{id}$. The schemes form a nested family, $B_T(k) \subset B_T(k+1)$, by padding with steps of size 0, and the depth-$k$ hypothesis class is $H \circ B_T(k)$. All statements of Section 3 apply to this subclass with $F = B_T(k)$ in Theorem 1.

**Lemma 17** (Stability and saturation)**.** *Every scheme in $\bigcup_k B_T(k)$ is $e^{\Lambda_s T}$-Lipschitz on $K$. Consequently $\bigcup_k B_T(k)$ is equicontinuous on the compact set $K$, hence totally bounded in $d_\infty$ (Ascoli's theorem), and*

$$\sup_k N\big(B_T(k), d_\infty, \varepsilon\big) =: N_\infty(\varepsilon) < \infty \quad (\varepsilon > 0), \qquad D_k(S) \le D_K.$$

*If moreover $\int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\,\mathrm{d}\varepsilon < \infty$, then $\mathsf{V}_k(S) \le \mathsf{V}_\infty < \infty$ for all $k$ (case (i) of Proposition 5).*

*Proof.* Each layer is $(1 + h_i\Lambda_s)$-Lipschitz, since $\Pi_K$ is 1-Lipschitz and $s(\cdot, \tau)$ is $\Lambda_s$-Lipschitz. A scheme is therefore $\prod_i(1 + h_i\Lambda_s) \le \exp(\Lambda_s \sum_i h_i) \le e^{\Lambda_s T}$-Lipschitz. A uniformly Lipschitz family of self-maps of a compact metric space is equicontinuous, and Ascoli's theorem gives total boundedness in $d_\infty$. The rest is Proposition 5(i). □

No contraction is used: only the stability of the integrator over the fixed horizon. Nor is the consistency of the time stamps used: the larger class of all compositions of at most $k$ layers from $F_T$ with a common drift and $\sum_i h_i \le T$ is also $e^{\Lambda_s T}$-Lipschitz and saturates in the same way. The finiteness of $\int \sqrt{\log N_\infty}$ is a condition on the one-layer class, and for the schemes actually used by a solver, equal steps $h = T/m$, it is explicit.

**Lemma 18** (Explicit entropy of equal-step schemes)**.** *Assume $0 < T \le h_0$. For $s \in \mathcal{S}_T$ and $m \ge 1$ let $\Phi_{s,m} := T_{s,T/m,\tau_m} \circ \cdots \circ T_{s,T/m,\tau_1}$ with $\tau_i = (i-1)T/m$ be the $m$-step equal-step Euler scheme, $\Phi_{s,0} := \mathrm{id}$, and let $E_T(k) := \{\Phi_{s,m} : s \in \mathcal{S}_T,\ 0 \le m \le k\} \subset B_T(k)$. Then for all $s \in \mathcal{S}_T$, every drift $s'$ (not necessarily in $\mathcal{S}_T$), and $m \ge 0$,*

$$d_\infty(\Phi_{s,m}, \Phi_{s',m}) \le T e^{\Lambda_s T}\,\|s - s'\|_\infty, \qquad \|s - s'\|_\infty := \sup_{x \in K,\ \tau \in [0,T]} \|s(x,\tau) - s'(x,\tau)\|,$$

*and consequently, for every $k \ge 0$ and $\varepsilon > 0$,*

$$N\big(E_T(k), d_\infty, \varepsilon\big) \le 1 + k\, N\big(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon e^{-\Lambda_s T}/T\big).$$

*If moreover $N(\mathcal{S}_T, \|\cdot\|_\infty, \rho) < \infty$ for all $\rho > 0$ and the integrand below is integrable, then*

$$\mathsf{V}_k(S) \le D_K\sqrt{\log(k+1)} + \int_0^{D_K} \sqrt{\log N\big(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon e^{-\Lambda_s T}/(2T)\big)}\,\mathrm{d}\varepsilon$$

*for the hypothesis class $H \circ E_T(k)$ (the factor $1/2$ is the internal/external transfer of Section H). If the drift class is parametric, $\log N(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon) \le p \log(C/\varepsilon)$, the last integral is at most $D_K\sqrt{p}\,(\sqrt{\log(2CTe^{\Lambda_s T}/D_K)} + \sqrt{\pi}/2)$, and the variance profile is $O(\sqrt{\log k})$, case (ii) of Proposition 5.*

*Proof.* Fix $x \in K$ and let $y_i$, $y_i'$ be the Euler iterates of $s$ and $s'$ started at $y_0 = y_0' = x$, with $h = T/m$. Since $\Pi_K$ is 1-Lipschitz and $s(\cdot, \tau)$ is $\Lambda_s$-Lipschitz (only the Lipschitz constant of $s$ is used),

$$\|y_i - y_i'\| \le \|y_{i-1} - y_{i-1}'\| + h\|s(y_{i-1}, \tau_i) - s'(y_{i-1}', \tau_i)\| \le (1 + h\Lambda_s)\|y_{i-1} - y_{i-1}'\| + h\|s - s'\|_\infty,$$

so $\|y_m - y_m'\| \le h\|s - s'\|_\infty \sum_{i<m}(1 + h\Lambda_s)^i \le mh\, e^{\Lambda_s mh}\|s - s'\|_\infty = Te^{\Lambda_s T}\|s - s'\|_\infty$. This is the telescoping estimate of Proposition 6(b) with per-layer error $h\|s - s'\|_\infty$ and $\Lambda = 1 + h\Lambda_s$. Hence, for each $m \ge 1$, the image of an $\varepsilon e^{-\Lambda_s T}/T$-cover of $\mathcal{S}_T$ under $s \mapsto \Phi_{s,m}$ is an $\varepsilon$-cover of $\{\Phi_{s,m} : s \in \mathcal{S}_T\}$; adding the identity and summing over $1 \le m \le k$ gives the covering bound. Since $1 + kN \le (k+1)N$ for $N \ge 1$, $\sqrt{\log(1+kN)} \le \sqrt{\log(k+1)} + \sqrt{\log N}$, $N(E_T(k), d_S, \varepsilon) \le N(E_T(k), d_\infty, \varepsilon/2)$, and $D_k(S) \le D_K$, integration over $(0, D_K]$ gives the bound on $\mathsf{V}_k(S)$. The condition $T \le h_0$ ensures that the equal steps $T/m$ are admissible, so that $E_T(k) \subset B_T(k)$. The parametric case follows from Lemma 9 as in Corollary 9(a). □

The equal-step schemes are nested, $E_T(k) \subset E_T(k+1)$, so the balancing argument of Section 5 applies to $H \circ E_T(k)$ verbatim; the estimate is stated for this subclass because a solver with a prescribed step count uses exactly these schemes.

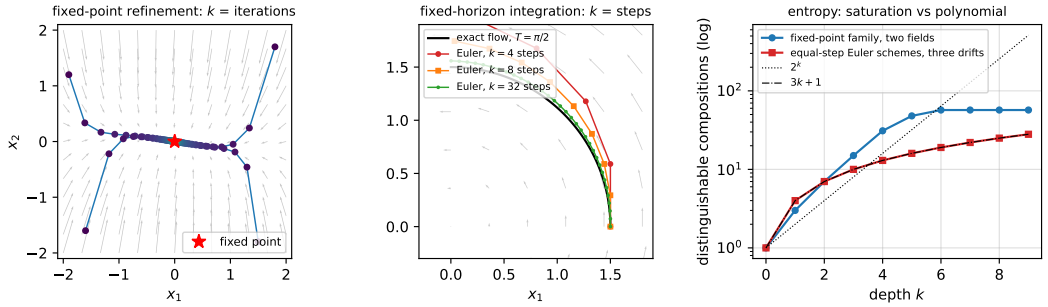


Figure 5: Unrolled solvers on $K = [-2, 2]^2$. Left: fixed-point refinement, four trajectories (dots, one per iteration) and the fixed point (star). Middle: Euler integration of a rotation field with 4, 8, and 32 steps (circles, squares, dots) and the exact flow (black). Right: distinguishable compositions of depth at most $k$ at resolution $2 \cdot 10^{-2}$ for a fixed-point family with two fields (blue circles) and for equal-step Euler schemes with three drifts (red squares), with the references $2^k$ (dotted) and $3k + 1$ (dash-dotted). See Section M.4.

**Bias.** Let the target be the endpoint map of the exact flow of some $s^* \in \mathcal{S}_T$, composed with an output layer: $\mathcal{C} := \{x \mapsto \langle w, \Phi_T^{s^*}(x)\rangle : \|w\| \le R\}$, where $\Phi_T^{s^*}(x)$ is the solution at time $T$ of $\dot{x}(t) = s^*(x(t), t)$, $x(0) = x$. Assume that the exact trajectories and the Euler trajectories started in $K$ stay in a region where $\Pi_K$ is inactive (for instance $K$ contains the $M_s T$-neighborhood of the relevant inputs). The classical global error bound for the explicit Euler scheme with $k$ equal steps $h = T/k$ gives

$$\sup_x \left\| \Phi_T^{s^*}(x) - T_{s^*,h,\tau_k} \circ \cdots \circ T_{s^*,h,\tau_1}(x) \right\| \le C_E \frac{T^2}{k}, \qquad C_E := \frac{(\Lambda_s M_s + \Lambda_\tau)\, e^{\Lambda_s T}}{2},$$

so that $\mathsf{bias}(k) \le \beta_\ell R\, C_E T^2/k$ by (1). If the layer class consists instead of one-step methods of order $p$ (for example Runge–Kutta steps), the same argument gives $\mathsf{bias}(k) \lesssim k^{-p}$.

**Proposition 12** (Fixed-horizon integration)**.** *In the setting above, the excess-risk bound of Theorem 4 is, up to depth-independent terms,* $\mathsf{gen}(k, n) \lesssim \beta_\ell R C_E T^2 k^{-p} + 12\, \mathsf{V}_\infty/\sqrt{n}$ *with* $p = 1$ *for Euler layers, the PL regime with* $\beta = p$ *and saturated variance. It is minimized at* $k^* \asymp n^{1/(2p)}$*, with balanced value of order* $n^{-1/2}$*.*

The number of integration steps should therefore grow polynomially with the sample size, while the statistical cost of the steps stays bounded. This is the depth trade-off of neural ODEs (Chen et al., 2018) and of stable residual architectures (Haber & Ruthotto, 2017; E, 2017), read in the state-transition model: depth refines the discretization and does not enlarge the effective hypothesis class beyond the equicontinuous family of stable schemes.

**Diffusion and flow-matching samplers.** The probability-flow ODE of a score-based diffusion model (Song et al., 2021; Ho et al., 2020) and the velocity field of a flow-matching model (Lipman et al., 2023) are time-dependent vector fields $s(x, \tau)$ whose learned approximants form the drift class $\mathcal{S}_T$; a sampler with $k$ steps is a scheme in $B_T(k)$, and the number of function evaluations is the depth. When the learning task is the regression of the endpoint map, for instance predicting a coordinate of the generated sample from the initial noise, or the reconstruction of clean data along the deterministic flow, Proposition 12 applies verbatim under the Lipschitz and boundedness conditions on the learned drifts. Generative quality is usually measured by a distance between distributions rather than by a regression loss, and the present analysis does not address that criterion; it addresses the statistical cost of the number of steps for the learned map.

### M.4 Computed illustration (Figure 5)

The two cases differ in one step. A strongly monotone field, as the score of a strongly log-concave density or the residual of a denoiser, makes every layer a contraction, so the entropy of $B(k, F)$ saturates by Theorem 3(ii) and the bias decays like $(1 - h\mu)^k$ because the iteration converges

geometrically to its fixed point; a denoiser or a deep equilibrium model may therefore be run to convergence without a statistical penalty. An Euler step of a stable but not contractive field over a fixed horizon, as in a neural ODE or the sampler of a diffusion model, keeps the compositions uniformly Lipschitz, so the entropy grows only logarithmically in the number of steps, while the bias is the discretization error of the scheme. Figure 5 shows both families on $K = [-2, 2]^2$: fixed-point refinement of a strongly monotone field (left, four trajectories converging to the fixed point), Euler integration of a rotation field over a fixed horizon with $4$, $8$, and $32$ steps against the exact flow (middle), and the distinguishable compositions of the two families at resolution $2 \cdot 10^{-2}$ (right): the fixed-point count follows $2^k$ until the contraction takes the compositions below the resolution and then saturates, and the count of the equal-step Euler schemes is $3k + 1$, linear in the depth (details below).

Figure 5 is produced by the script `pics/figure-ode-example.py` (NumPy and Matplotlib; no randomness). Its three panels use the following setting.

**Fixed-point refinement (left panel).** The state space is $K = [-2, 2]^2$ with the Euclidean metric and $\Pi_K$ the coordinatewise clipping to $[-2, 2]$. The vector field is

$$s(x) = -\big(x_1 + \tfrac{1}{2}\sin x_2,\ 4x_2 + \tfrac{1}{2}\sin x_1\big).$$

Its Jacobian is $-\left[\begin{smallmatrix} 1 & \frac{1}{2}\cos x_2 \\ \frac{1}{2}\cos x_1 & 4 \end{smallmatrix}\right]$, whose symmetric part has eigenvalues between $0.92$ and $4.08$, so $s$ is $\mu$-strongly monotone and $\Lambda_s$-Lipschitz with $\mu = 0.9$ and $\Lambda_s = 4.1$; the contraction estimate of Lemma 16 holds for such fields, since its proof uses only strong monotonicity and co-coercivity, which follow from these bounds on the symmetric part. The step size is $h = 1/(\mu + \Lambda_s) = 0.2$, inside the admissible range $(0, 2/(\mu + \Lambda_s)]$, so every layer is $(1 - h\mu) = 0.82$-Lipschitz. Four initial states, $(1.8, 1.7)$, $(-1.9, 1.2)$, $(1.5, -1.8)$, $(-1.6, -1.6)$, are iterated $k = 40$ times, and the fixed point $x^*$ is computed by $400$ iterations from the first initial state; the largest distance $\max_i \|x_k^{(i)} - x^*\|$ over the four trajectories, the quantity bounded by $D_K \lambda^k$ in Proposition 11, decays geometrically at a rate close to $1 - h\mu$ (printed by the script).

**Fixed-horizon integration (middle panel).** The vector field is the rotation $s(x) = (-x_2, x_1)$, which is $\Lambda_s = 1$-Lipschitz, bounded by $M_s = 2\sqrt{2}$ on $K$, and not contractive; its exact flow over time $t$ is the rotation by the angle $t$. The horizon is $T = \pi/2$, and the scheme is the explicit Euler method with $k$ equal steps $h = T/k$, the equal-step scheme $\Phi_{s,k}$ of Lemma 18; the projection is inactive along the displayed paths. The panel shows the paths from the initial state $(1.5, 0)$ for $k = 4, 8, 32$ and the exact quarter circle; the endpoint error $\max_i \|\Phi_{s,k}(x^{(i)}) - \Phi_T^s(x^{(i)})\|$ over the three initial states $(1.5, 0)$, $(1, 0.5)$, $(0.6, -1.2)$ decays like $k^{-1}$ (printed by the script), the global error of the Euler method, which is the bias of Proposition 12 with $p = 1$.

**Distinguishable compositions (right panel).** The right panel compares the entropy of the two families in the same way as Figures 3 and 4: it counts the compositions of depth at most $k$ that are distinguishable at resolution $\varepsilon = 2{\cdot}10^{-2}$, that is, the size of a greedy $\varepsilon$-net in the metric $\max_i \|f(x_i) - g(x_i)\|$ over the $21 \times 21$ grid of points $x_i$ of $K$, for $k \le 9$. The fixed-point family has two layers, $x \mapsto \Pi_K(x + h' s_j(x))$ with the strongly monotone fields $s_1(x) = -(2x_1 + \frac{1}{2}\sin x_2,\ 4x_2 + \frac{1}{2}\sin x_1)$ and $s_2(x) = -(4x_1 + \frac{1}{2}\cos x_2,\ 2x_2 + \frac{1}{2}\cos x_1)$, whose symmetric Jacobians have eigenvalues between $1.9$ and $4.1$, and the step $h' = 2/(1.9 + 4.1) = 1/3$, so that every layer is $0.37$-Lipschitz; all $2^k$ words of length $k$ are formed. The count is $1, 3, 7, 15, 31, 48, 57$ for $k \le 6$ and stays at $57$ afterwards: the words follow $2^k$ until the contraction $0.37^k$ falls below the resolution, and then saturate at the entropy of the attractor of the two fields, which is mechanism P1' (Lemma 16). The fixed-horizon family is the class $E_T(k)$ of Lemma 18 with the drift class $\{\omega\,(-x_2, x_1) : \omega \in \{0.8, 1, 1.2\}\}$, that is, the equal-step Euler schemes with $m \le k$ steps and one of three angular speeds. The count is $3k + 1$: one map per pair (drift, number of steps), all distinguishable at this resolution, which is the linear growth $1 + k\,N(\mathcal{S}_T, \cdot)$ of Lemma 18 and hence the $\sqrt{\log k}$ profile. The two curves are the two variance profiles of the examples, saturated and logarithmic; the two bias laws, geometric for the contractive iteration and $k^{-1}$ for the Euler scheme, are the values printed by the script and are the other side of the balance in Section 5.

### M.5 Implementation

For both families, Proposition 6(b) controls the implementation error of approximate layers. For fixed-point refinement, $\Lambda = 1 - h\mu < 1$ and $\varepsilon_{\text{imp}}(k) \le \delta_H + R\delta/(h\mu)$ uniformly in $k$. For fixed-horizon schemes, $\Lambda_i = 1 + h_i\Lambda_s$ and the telescoping sum is bounded by $\delta\, k\, e^{\Lambda_s T}$, so $\varepsilon_{\text{imp}}(k) \le \delta_H + R\delta k e^{\Lambda_s T}$ grows only linearly in the number of steps; with $\delta$ the per-step accuracy of the learned drift this is the usual stability of one-step methods. In both cases the implementation error does not change the regime of the depth trade-off.

### M.6 Relation to the literature

Unrolled fixed-point iterations are the model of deep equilibrium networks (Bai et al., 2019) and of regularization by denoising (Romano et al., 2017), and Picard-type iterations of the same kind underlie the neural operators analyzed by Taniguchi & Sonoda (2026). Fixed-horizon integration is the model of neural ODEs (Chen et al., 2018) and of stable residual architectures (Haber & Ruthotto, 2017; E, 2017), and the samplers of score-based diffusion and flow-matching models (Song et al., 2021; Ho et al., 2020; Lipman et al., 2023) are its instances with a learned drift. In each case the depth is the number of iterations or steps; the present appendix supplies the statistical side of the question of how many of them to use, which those works treat as a numerical or empirical choice.