Generalization error bounds for deep models

Sho Sonoda

September 26, 2026

Chapter 1

Introduction

This blueprint formalizes the paper *Why and When Deep is Better than Shallow: Implementation-Agnostic State-Transition Model of Deep Learning*. Each hidden layer is a self-map of a metric state space $\mathcal{X}$, the depth-k hidden class $B(k, F)$ is the word ball of the semigroup generated by a hidden-layer class F, and the depth-k hypothesis class is $\mathcal{H}_k = H \circ B(k, F)$ for an output-layer class H. The blueprint is generated from the `@[blueprint]` annotations in the Lean sources; the informal plan is kept in `PLAN.md`.

The general-purpose tools developed for this formalization (results that belong in Mathlib, and tools for lean-rademacher: contraction for arbitrary index sets, one-sided uniform-deviation bounds, Dudley's entropy integral for sub-Gaussian processes, Hilbert-space Hoeffding bounds, Gaussian comparison and the Bernoulli–Sudakov minoration) now live in lean-rademacher (`FoML.ToMathlib`, `FoML.ToFoML`) and are used here as a dependency; they are not part of this blueprint.

The manuscript's appendix was reorganized on 2026-09-26 into Appendices A–M (see `SUMMARY.md` for the section-by-section correspondence); the chapter and section titles below note, in parentheses, which manuscript appendix each block of Lean modules now corresponds to. Labels (`thm:...`, `prop:...`, ...) are unchanged by the reorganization.

Chapter 2

Setting

2.1 Word balls and the semigroup generated by the hidden layers

Definition 1. Let $F \subseteq \mathcal{X}^{\mathcal{X}}$ be a hidden-layer class. The depth-k hidden class is the word ball $B(k, F) = \{f_m \circ \cdots \circ f_1 : 0 \le m \le k,\ f_i \in F\}$, with $B(0, F) = \{\mathrm{id}\}$. It is defined recursively by $B(k + 1, F) = B(k, F) \cup F \circ B(k, F)$.

Definition 2. The words of length exactly m: $F^m = \{f_m \circ \cdots \circ f_1 : f_i \in F\}$, with $F^0 = \{\mathrm{id}\}$ and $F^{m+1} = F \circ F^m$.

Definition 3. The semigroup (monoid) generated by F is $\langle F\rangle = \bigcup_{k\ge 0} B(k, F)$.

Theorem 4. $B(0, F) = \{\mathrm{id}\}$.

Proof. □

Theorem 5. $B(k + 1, F) = B(k, F) \cup \{g \circ f : g \in F,\ f \in B(k, F)\}$.

Proof. □

Theorem 6. $\mathrm{id} \in B(k, F)$ *for every* k.

Proof. Induction on k: $\mathrm{id} \in B(0, F)$ and $B(k, F) \subseteq B(k + 1, F)$. □

Theorem 7. *The word balls are increasing:* $B(k, F) \subseteq B(k + 1, F)$*, hence* $B(k, F) \subseteq B(l, F)$ *for* $k \le l$*.*

Proof. □

Theorem 8. *If* $g \in F$ *and* $f \in B(k, F)$ *then* $g \circ f \in B(k + 1, F)$*.*

Proof. □

Theorem 9. $B(k, F) \subseteq \langle F\rangle$ *for every* k*.*

Proof. □

Theorem 10. *If* $F \subseteq G$ *then* $B(k, F) \subseteq B(k, G)$*.*

Proof. Induction on k, using that $\circ$-images are monotone in both arguments. $\square$

Theorem 11. *If $f \in B(k, F)$ and $g \in B(l, F)$ then $g \circ f \in B(l + k, F)$.*

Proof. Induction on l (the length of the word g). If $g = \mathrm{id}$ this is $f \in B(k, F)$; if $g = a \circ b$ with $a \in F$, $b \in B(l, F)$ then $g \circ f = a \circ (b \circ f)$ with $b \circ f \in B(l + k, F)$ by induction. $\square$

Theorem 12. *$B(k+1, F) = B(k, F) \cup F^{k+1}$: the word ball of radius $k+1$ is the word ball of radius k together with the words of length exactly $k+1$.*

Proof. Induction on k: $F \circ B(k+1, F) = F \circ B(k, F) \cup F \circ F^{k+1}$ and $F \circ B(k, F) \subseteq B(k+1, F)$. $\square$

Theorem 13. *$|F^m| \le |F|^m$ (as extended natural numbers).*

Proof. Induction on k: F^{k+1} is the image of $F \times F^k$ under composition. $\square$

Theorem 14. *$|B(k, F)| \le \sum_{m=0}^{k} |F|^m$ (as extended natural numbers).*

Proof. Induction on k using $B(k+1, F) = B(k, F) \cup F^{k+1}$ and $|F^{k+1}| \le |F|^{k+1}$. $\square$

Theorem 15. *If F is finite with $|F| = r \ge 2$ then $|B(k, F)| \le \sum_{m=0}^{k} r^m \le r^{k+1}$.*

Proof. Combine the previous bound with the geometric sum estimate $\sum_{m \le k} r^m = (r^{k+1}-1)/(r-1) < r^{k+1}$ for $r \ge 2$. $\square$

2.2 Uniform and empirical metrics

Definition 16. The uniform distance on $\mathcal{X}^{\mathcal{X}}$ is $d_\infty(f, g) = \sup_{x \in \mathcal{X}} d(f(x), g(x)) \in [0, \infty]$.

Definition 17. The real-valued uniform distance $d_\infty(f, g)$, with the convention that it is 0 when $d_\infty(f, g) = \infty$.

Theorem 18. *$d(f(x), g(x)) \le d_\infty(f, g)$ for every x.*

Proof. $\square$

Theorem 19. *Right composition is 1-Lipschitz for d_∞: $d_\infty(a \circ f, b \circ f) \le d_\infty(a, b)$.*

Proof. $\square$

Theorem 20. *Left composition with a K-Lipschitz map f is K-Lipschitz for d_∞: $d_\infty(f \circ a, f \circ b) \le K\, d_\infty(a, b)$.*

Proof. Pointwise, $d(f(a(x)), f(b(x))) \le K d(a(x), b(x)) \le K d_\infty(a, b)$. $\square$

Definition 21. The pseudo-emetric space $(\mathcal{X}^{\mathcal{X}}, d_\infty)$ of self-maps equipped with the uniform distance (a type synonym of $\mathcal{X}^{\mathcal{X}}$; in Lean it is Mathlib's 'X → X', the function space with the uniform structure).

Definition 22. d_∞ is a pseudo-emetric on $\mathcal{X}^{\mathcal{X}}$ (it may take the value ∞). This is Mathlib's instance on 'X → X'.

Definition 23. The (identity) map $\mathcal{X}^{\mathcal{X}} \to (\mathcal{X}^{\mathcal{X}}, d_\infty)$ viewing a self-map as a point of the pseudo-emetric space.

Theorem 24. *Viewing f in $(\mathcal{X}^{\mathcal{X}}, d_\infty)$ does not change its values.*

Proof. $\square$

Theorem 25. *On $(\mathcal{X}^{\mathcal{X}}, d_\infty)$ the extended distance is d_∞.*

Proof. $\square$

Theorem 26. *d_∞ is the distance of $(\mathcal{X}^{\mathcal{X}}, d_\infty)$: $\operatorname{edist}(f, g) = d_\infty(f, g)$ for $f, g \in \mathcal{X}^{\mathcal{X}}$.*

Proof. $\square$

Theorem 27. *$d(f(x), g(x)) \le d_\infty(f, g)$ for $f, g \in (\mathcal{X}^{\mathcal{X}}, d_\infty)$.*

Proof. $\square$

Definition 28. For a sample $S = (x_1, \dots, x_n)$ the empirical distance is $d_S(f, g) = \left(\frac{1}{n} \sum_{i=1}^{n} d(f(x_i), g(x_i))^2\right)^{1/2}$.

Definition 29. For real-valued u, v the empirical L^2 distance is $\|u - v\|_S = \left(\frac{1}{n} \sum_{i=1}^{n} |u(x_i) - v(x_i)|^2\right)^{1/2}$.

Definition 30. The empirical sup norm $\|u\|_{S,\infty} = \max_{i \le n} |u(x_i)|$ (equal to 0 when $n = 0$).

Definition 31. The empirical diameter of a class F is $\operatorname{diam}_S(F) = \sup_{f,g \in F} d_S(f, g)$.

Theorem 32. *$d_S(f, f) = 0$.*

Proof. $\square$

Theorem 33. *$d_S(f, g) = d_S(g, f)$.*

Proof. $\square$

Theorem 34. *$d_S(f, g) \ge 0$.*

Proof. $\square$

Theorem 35. *Minkowski's inequality for the Euclidean norm on $\mathbb{R}^n$: $\left(\sum_i (u_i + v_i)^2\right)^{1/2} \le \left(\sum_i u_i^2\right)^{1/2} + \left(\sum_i v_i^2\right)^{1/2}$.*

Proof. This is the triangle inequality in the Euclidean space $\ell^2(\{1, \dots, n\})$. $\square$

Theorem 36. *The empirical distance satisfies the triangle inequality: $d_S(f, h) \le d_S(f, g) + d_S(g, h)$. Hence d_S is a pseudometric on $\mathcal{X}^{\mathcal{X}}$.*

Proof. Pointwise $d(f(x_i), h(x_i)) \le d(f(x_i), g(x_i)) + d(g(x_i), h(x_i))$, then apply Minkowski's inequality to the vectors of pointwise distances. $\square$

Definition 37. The pseudometric space $(\mathcal{X}^{\mathcal{X}}, d_S)$ of self-maps equipped with the empirical distance of the sample S (a type synonym of $\mathcal{X}^{\mathcal{X}}$).

Definition 38. d_S is a pseudometric on $\mathcal{X}^{\mathcal{X}}$.

Theorem 39. *On $(\mathcal{X}^{\mathcal{X}}, d_S)$ the distance is d_S.*

Proof. $\square$

Theorem 40. *If $d_\infty(f,g) < \infty$ then $d(f(x), g(x)) \le d_\infty(f,g)$ as real numbers.*

Proof. □

Theorem 41. *The empirical distance is dominated by the uniform distance: $d_S(f,g) \le d_\infty(f,g)$ for every sample S.*

Proof. If $d_\infty(f,g) = \infty$ there is nothing to prove. Otherwise each term satisfies $d(f(x_i), g(x_i))^2 \le d_\infty(f,g)^2$, so the average is at most $d_\infty(f,g)^2$ and we take square roots. □

2.3 Covering and packing numbers

Definition 42. The metric entropy of A at scale ε is $\log N(A,\varepsilon)$, where $N(A,\varepsilon)$ is the (internal) covering number of A by closed ε-balls (with the convention $\log\infty = \log 0 = 0$ in Lean).

2.4 Hypothesis classes and implementation error

Definition 43. For $H \subseteq \mathbb{R}^{\mathcal{X}}$ and $F \subseteq \mathcal{X}^{\mathcal{X}}$, $H \circ F = \{h \circ f : h \in H,\ f \in F\}$.

Definition 44. The depth-k hypothesis class is $\mathcal{H}_k = H \circ B(k, F)$.

Definition 45. The implementation error of an implementation map ι on a class $\mathcal{H}$ is $\varepsilon_{\text{imp}} = \sup_{f\in\mathcal{H}} \|f - \iota f\|_\infty$.

Theorem 46. *If $F \subseteq F'$ then $H \circ F \subseteq H \circ F'$.*

Proof. □

Theorem 47. *$H \circ \{\mathrm{id}\} = H$.*

Proof. □

Theorem 48. *$\mathcal{H}_0 = H \circ \{\mathrm{id}\} = H$.*

Proof. □

Theorem 49. *The hypothesis classes are increasing in depth: $\mathcal{H}_k \subseteq \mathcal{H}_{k+1}$.*

Proof. □

2.5 Loss, risk, and empirical minimizers

Definition 50. A loss is a function $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$ which is β_ℓ-Lipschitz in its first argument: $|\ell(a,y) - \ell(a',y)| \le \beta_\ell |a - a'|$.

Definition 51. The empirical risk on a sample $D = ((x_1, y_1), \dots, (x_n, y_n))$ is $\hat{L}[f] = \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)$.

Definition 52. The (population) risk under a distribution P on $\mathcal{X} \times \mathcal{Y}$ is $L[f] = \mathbb{E}_{(X,Y)\sim P}\, \ell(f(X), Y)$.

Definition 53. f is an η-empirical minimizer over $\mathcal{H}$ if $f \in \mathcal{H}$ and $\hat{L}[f] \le \inf_{g\in\mathcal{H}} \hat{L}[g] + \eta$.

Definition 54. The model error (approximation term) of $\mathcal{H}$ relative to a target class $\mathcal{C}$ is $\varepsilon_{\text{model}} = \inf_{f\in\mathcal{H}} L[f] - \inf_{c\in\mathcal{C}} L[c]$.

2.6 Rademacher complexity

Definition 55. The empirical Rademacher complexity of $G \subseteq \mathbb{R}^{\mathcal{X}}$ on the sample $S = (x_1, \dots, x_n)$ is $\hat{\mathfrak{R}}_S(G) = \mathbb{E}_\sigma \sup_{g \in G} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i) = 2^{-n} \sum_{\sigma \in \{\pm 1\}^n} \sup_{g \in G} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i)$ (no absolute value). In Lean this is FoML's ‘empiricalRademacherComplexity_without_abs‘ for the class indexed by G itself.

Definition 56. The (population) Rademacher complexity is $\mathfrak{R}_n(G) = \mathbb{E}_{S \sim P^{\otimes n}} \hat{\mathfrak{R}}_S(G)$. In Lean we take FoML's ‘rademacherComplexity‘, i.e. the expectation over $S \sim P^{\otimes n}$ of the absolute version $\mathbb{E}_\sigma \sup_{g \in G} \left| \frac{1}{n} \sum_i \sigma_i g(x_i) \right| \geq \hat{\mathfrak{R}}_S(G)$; this is the quantity for which FoML's deviation bounds are stated, and it dominates the paper's $\mathfrak{R}_n(G)$.

Theorem 57. *If $G_1 \subseteq G_2$, $G_1 \neq \emptyset$ and the Rademacher averages $\{\frac{1}{n} \sum_i \sigma_i g(x_i) : g \in G_2\}$ are bounded above for every sign pattern σ, then $\hat{\mathfrak{R}}_S(G_1) \leq \hat{\mathfrak{R}}_S(G_2)$.*

Proof. Compare the suprema termwise for each sign pattern σ: every element of G_1 is an element of G_2. □

Theorem 58. *If $|g(x_i)| \leq C$ for all $g \in G$ and $i \leq n$, then the one-sided empirical Rademacher complexity is dominated by the absolute one: $\hat{\mathfrak{R}}_S(G) \leq \mathbb{E}_\sigma \sup_{g \in G} \left| \frac{1}{n} \sum_i \sigma_i g(x_i) \right|$.*

Proof. □

2.7 The entropy integral

Definition 59. The entropy integral of A up to scale D is $\mathsf{V}(D, A) = \int_0^D \sqrt{\log N(A, \varepsilon)}\, d\varepsilon$. In the paper, $\mathsf{V}_k(S) = \int_0^{D_k(S)} \sqrt{\log N(B(k, F), d_S, \varepsilon)}\, d\varepsilon$ with $D_k(S) = \operatorname{diam}_S B(k, F)$ is $\mathsf{V}(\operatorname{diam}_S B(k, F), B(k, F))$ for the empirical pseudometric d_S.

2.8 Assumptions and auxiliary definitions of the main theorems

Definition 60. A class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is (sup-norm) separable if it has a countable subset $\mathcal{D} \subseteq \mathcal{H}$ which is dense for the uniform norm: for every $f \in \mathcal{H}$ and $\varepsilon > 0$ there is $g \in \mathcal{D}$ with $\sup_x |f(x) - g(x)| \leq \varepsilon$.

Definition 61. For a sample $S = (x_1, \dots, x_n)$, an output-layer class H and a hidden map f, the hidden-indexed process is $Z_f(\sigma) = \sup_{h \in H} \frac{1}{n} \sum_{i=1}^n \sigma_i h(f(x_i))$, $\sigma \in \{\pm 1\}^n$.

Definition 62. **Sub-Gaussian output-layer increments.** Let $\mathfrak{F}$ be a hidden-layer class and A_H, L constants. For all $f, g \in \mathfrak{F}$ and $t > 0$,

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

where $\mathbb{P}_\sigma$ is the uniform distribution on $\{\pm 1\}^n$; and when $d_S(f, g) = 0$ the requirement is $Z_f = Z_g$ (for every σ), which is the limiting interpretation of the display (in Lean the display with $d_S(f, g) = 0$ reads ... $\leq 2\exp(0)$, hence the separate conjunct).

Definition 63. For a Hilbert space $\mathscr{H}$ and a feature map $\Phi : \mathcal{X} \to \mathscr{H}$, the norm-bounded linear output-layer class is $H_\Phi = \{x \mapsto \langle w, \Phi(x)\rangle : \|w\| \le 1\}$.

Definition 64. Output-layer realization of hidden geometry. For a sample S, an output-layer class H, a hidden class B_k and constants κ, R_{out}: there is, for each $g \in B_k$, an output layer $h_g \in H$ such that the map $\Psi_k(g) = h_g \circ g$ satisfies $\|\Psi_k(g) - \Psi_k(g')\|_S \ge \kappa\, d_S(g, g')$ for $g, g' \in B_k$ and $\|\Psi_k(g)\|_{S,\infty} \le R_{\text{out}}$ for $g \in B_k$.

Definition 65. The pseudo-emetric space $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_\infty)$ of real-valued functions with the uniform distance $\|u - v\|_\infty = \sup_x |u(x) - v(x)| \in [0, \infty]$ (a type synonym of $\mathbb{R}^{\mathcal{X}}$; in Lean Mathlib's 'X → ').

Definition 66. $\|\cdot\|_\infty$ is a pseudo-emetric on $\mathbb{R}^{\mathcal{X}}$.

Theorem 67. *On $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_\infty)$ the extended distance is $\sup_x \operatorname{edist}(u(x), v(x))$.*

Proof. □

Definition 68. Uniform output-layer regularity. The output-layer class $H \subseteq \mathbb{R}^{\mathcal{X}}$ has finite uniform covering numbers $N(H, \|\cdot\|_\infty, u) < \infty$ for all $u > 0$, and there are constants B_H, L_H with $\|h\|_\infty \le B_H$ and $|h(x) - h(x')| \le L_H d(x, x')$ for all $h \in H$.

Definition 69. Uniform transition covering. The hidden-layer class $F \subseteq \mathcal{X}^{\mathcal{X}}$ has finite uniform covering numbers $N(F, d_\infty, v) < \infty$ for all $v > 0$.

Definition 70. The (identity) map $\mathbb{R}^{\mathcal{X}} \to (\mathbb{R}^{\mathcal{X}}, \|\cdot\|_\infty)$ viewing a real-valued function as a point of the pseudo-emetric space.

Theorem 71. $\operatorname{edist}(u, v) = \sup_x \operatorname{edist}(u(x), v(x))$ *in* $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_\infty)$.

Proof. □

Definition 72. The root entropy of the output-layer class at scale u: $\mathcal{E}_H(u) = \sqrt{\log N^{\text{ext}}(H, \|\cdot\|_\infty, u)}$ (with $u \mapsto \max(u, 0)$ and $\log \infty = \log 0 = 0$ in Lean).

Definition 73. The root entropy of the hidden class at scale v: $\mathcal{E}_F(v) = \sqrt{\log N^{\text{ext}}(F, d_\infty, v)}$.

Definition 74. The pseudometric space $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_S)$ of real-valued functions equipped with the empirical L^2 distance $\|u - v\|_S = \left(\frac{1}{n}\sum_i |u(x_i) - v(x_i)|^2\right)^{1/2}$ of the sample S (a type synonym of $\mathbb{R}^{\mathcal{X}}$).

Definition 75. $\|\cdot\|_S$ is a pseudometric on $\mathbb{R}^{\mathcal{X}}$. In Lean this is FoML's 'empiricalPMet S' (whose distance is FoML's 'empiricalDist S').

Definition 76. The (identity) map $\mathbb{R}^{\mathcal{X}} \to (\mathbb{R}^{\mathcal{X}}, \|\cdot\|_S)$ viewing a real-valued function as a point of the pseudometric space.

Theorem 77. *On $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_S)$ the distance is FoML's 'empiricalDist S'.*

Proof. □

Theorem 78. *On $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_S)$ the distance is $\|u - v\|_S$.*

Proof. □

Definition 79. The set of points reached by the sample through the hidden class: $\mathcal{R}_S(F) = \{f(x_i) : f \in F,\ 1 \le i \le n\} \subseteq \mathcal{X}$.

Theorem 80. *$f(x_i) \in \mathcal{R}_S(F)$ for $f \in F$.*

Proof. ☐

Definition 81. The uniform pushed-forward covering number of the output-layer class: $N_{S,F}(H, u) = \sup_{f \in F} N^{\text{ext}}(H, \|\cdot\|_{f \circ S}, u)$, the largest external covering number of H in the empirical metric of a pushed-forward sample $f \circ S = (f(x_1), \dots, f(x_n))$, $f \in F$. (Any uniform bound $N^{\text{ext}}(H, \|\cdot\|_{f \circ S}, u) \le \bar{N}(u)$ for all $f \in F$ dominates it.)

Definition 82. **Sample output-layer regularity.** For a sample S and a hidden class F, the output-layer class $H \subseteq \mathbb{R}^{\mathcal{X}}$ has finite uniform pushed-forward covering numbers $N_{S,F}(H, u) < \infty$ for all $u > 0$, and there are constants B_H, L_H such that every $h \in H$ satisfies $|h(x)| \le B_H$ and $|h(x) - h(x')| \le L_H d(x, x')$ for all $x, x' \in \mathcal{R}_S(F)$ (boundedness and the Lipschitz estimate are only required on the points reached by the sample; compare `ass:ent-readout`, where both hold on all of $\mathcal{X}$ and H is covered in the sup norm).

Definition 83. **Sample transition covering.** The hidden-layer class $F \subseteq \mathcal{X}^{\mathcal{X}}$ has finite covering numbers in the empirical metric: $N^{\text{ext}}(F, d_S, v) < \infty$ for all $v > 0$.

Definition 84. The root entropy of the output-layer class at scale u on the sample: $\mathcal{E}_{H,S}(u) = \sqrt{\log N_{S,F}(H, u)}$ (with $u \mapsto \max(u, 0)$ and $\log \infty = \log 0 = 0$ in Lean).

Definition 85. The root entropy of the hidden class at scale v on the sample: $\mathcal{E}_{F,S}(v) = \sqrt{\log N^{\text{ext}}(F, d_S, v)}$.

Definition 86. The identity map $(\mathcal{X}^{\mathcal{X}}, d_\infty) \to (\mathcal{X}^{\mathcal{X}}, d_S)$.

Definition 87. The depth-dependent part of the estimation term is

$$\operatorname{var}(k, n) := \frac{12 A_H L}{\sqrt{n}} \mathsf{V}_k(S), \qquad \mathsf{V}_k(S) = \int_0^{D_k(S)} \sqrt{\log N(B(k, F), d_S, \varepsilon)}\, d\varepsilon,$$

with $D_k(S) = \operatorname{diam}_S B(k, F)$.

Chapter 3

Implementation-agnostic generalization bounds

3.1 Implementation-free bias–variance decomposition (proof: Appendix C, thm:bv, thm:bv-general)

Theorem 88. *If ℓ is β_ℓ-Lipschitz in its first argument (with $\beta_\ell \geq 0$), f, g are measurable and $\sup_x |f(x) - g(x)| \leq c$, then $|L[f] - L[g]| \leq \beta_\ell c$.*

Proof. Both loss integrands are bounded and measurable, hence integrable; the difference of the integrals is the integral of the pointwise difference, which is bounded by $\beta_\ell |f(x) - g(x)| \leq \beta_\ell c$. ☐

Theorem 89. *If ℓ is β_ℓ-Lipschitz in its first argument (with $\beta_\ell \geq 0$) and $\sup_x |f(x) - g(x)| \leq c$ with $c \geq 0$, then $|\hat{L}[f] - \hat{L}[g]| \leq \beta_\ell c$ for every sample.*

Proof. Termwise bound $|\ell(f(x_i), y_i) - \ell(g(x_i), y_i)| \leq \beta_\ell c$ and average. ☐

Theorem 90. ***Deterministic excess-risk bound.*** *Fix a sample $\mathcal{D}$ and suppose $|L[f] - \hat{L}[f]| \leq \Delta_0$ for every $f \in \mathcal{H}$. If $d_T(\iota f, f) \leq \varepsilon_{\text{imp}}$ and $|L[\iota f] - L[f]| \leq \beta_L\, d_T(\iota f, f)$ for all $f \in \mathcal{H}$ ($\beta_L \geq 0$), then every η-empirical minimizer $\hat{f} \in \mathcal{H}$ satisfies, with $\hat{h} = \iota(\hat{f})$, $L[\hat{h}] - \inf_{\mathcal{C}} L \leq \beta_L \varepsilon_{\text{imp}} + \varepsilon_{\text{model}} + \eta + 2\Delta_0$.*

Proof. $L[\hat{h}] \leq L[\hat{f}] + \beta_L \varepsilon_{\text{imp}}$ and, for every $f \in \mathcal{H}$, $L[\hat{f}] \leq \hat{L}[\hat{f}] + \Delta_0 \leq \hat{L}[f] + \eta + \Delta_0 \leq L[f] + \eta + 2\Delta_0$; take the infimum over f. ☐

Theorem 91. ***Deterministic gap bound.*** *Fix a sample $\mathcal{D}$ and suppose $|L[f] - \hat{L}[f]| \leq \Delta_0$ for every $f \in \mathcal{H}$. Under the implementation hypotheses ($d_T(\iota f, f) \leq \varepsilon_{\text{imp}}$, $|L[\iota f] - L[f]| \leq \beta_L d_T(\iota f, f)$, $|\hat{L}[\iota f] - \hat{L}[f]| \leq \beta_{\hat{L}} d_T(\iota f, f)$), every $f \in \mathcal{H}$ satisfies $L[\iota f] - \hat{L}[\iota f] \leq (\beta_L + \beta_{\hat{L}})\varepsilon_{\text{imp}} + \Delta_0$.*

Proof. ☐

Theorem 92. *The one-sided empirical Rademacher complexity of the class $\{(x, y) \mapsto f(x) : f \in \mathcal{H}\}$ on the labelled sample $\mathcal{D}$ equals $\hat{\mathfrak{R}}_S(\mathcal{H})$ on the inputs $S = (x_1, \ldots, x_n)$.*

Proof. ☐

Theorem 93. ***One-sided Rademacher complexity of the loss class.*** *Let $\mathcal{H} \neq \emptyset$ be pointwise bounded and $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be β_ℓ-Lipschitz in its first argument ($\beta_\ell \geq 0$). Then on every sample $\mathcal{D} = ((x_i, y_i))_i$ with $S = (x_i)_i$,*

$$\widehat{\mathfrak{R}}_{\mathcal{D}}(\ell \circ \mathcal{H}) \leq \beta_\ell \widehat{\mathfrak{R}}_S(\mathcal{H}) \quad \text{and} \quad \widehat{\mathfrak{R}}_{\mathcal{D}}(-\ell \circ \mathcal{H}) \leq \beta_\ell \widehat{\mathfrak{R}}_S(\mathcal{H}),$$

where $\widehat{\mathfrak{R}}$ is the one-sided (no absolute value) empirical Rademacher complexity. Proof: the one-sided contraction `lem:contraction-without-abs` applied to $\psi(z, u) = \pm\ell(u, y)$, which is β_ℓ-Lipschitz in u (no vanishing condition at $u = 0$ is needed).

Proof. ☐

Theorem 94. ***Uniform deviation for a sup-separable Lipschitz-loss class.*** *Let $n \geq 1$, $\mathcal{H}$ be a sup-norm separable, pointwise bounded class of measurable functions, $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$ measurable and β_ℓ-Lipschitz in its first argument ($b > 0$, $\beta_\ell \geq 0$), $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ over $\mathcal{D} \sim P^{\otimes n}$,*

$$\sup_{f \in \mathcal{H}} |L[f] - \hat{L}[f]| \leq 2\beta_\ell \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3b\sqrt{\tfrac{2\log(4/\delta)}{n}}.$$

Proof: apply the two-sided bound `lem:two-sided-tail-empirical` (one-sided symmetrization for $\ell \circ \mathcal{H}$ and $-\ell \circ \mathcal{H}$) with $C(\mathcal{D}) = \beta_\ell \widehat{\mathfrak{R}}_S(\mathcal{H})$ from `lem:bv-loss-rademacher` and $\varepsilon = b\sqrt{2\log(4/\delta)/n}$, so that $4\exp(-n\varepsilon^2/(2b^2)) = \delta$; the topology on $\mathcal{H}$ is the uniform-convergence topology, which is separable and first countable by `lem:sup-dense-separableSpace`, and evaluations are continuous.

Proof. ☐

Theorem 95. ***General bias–variance decomposition (excess risk).*** *Let $\mathcal{H}$ be a sup-norm separable, pointwise bounded class of measurable functions $\mathcal{X} \to \mathbb{R}$, $\mathcal{C}$ a nonempty class of measurable benchmark functions, and $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$ measurable and β_ℓ-Lipschitz in its first argument ($b > 0$, $\beta_\ell \geq 0$). Let ι be an implementation map, d_T a nonnegative function (pseudo-metric) and assume, for constants $\beta_L, \beta_{\hat{L}} \geq 0$ and $\varepsilon_{\text{imp}} \geq 0$ (any upper bound for $\sup_{f \in \mathcal{H}} d_T(\iota f, f)$), that for every $f \in \mathcal{H}$: $d_T(\iota f, f) \leq \varepsilon_{\text{imp}}$, $|L[\iota f] - L[f]| \leq \beta_L d_T(\iota f, f)$ and $|\hat{L}[\iota f] - \hat{L}[f]| \leq \beta_{\hat{L}} d_T(\iota f, f)$ for every sample. Let $n \geq 1$, $\eta \geq 0$ and $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ over $\mathcal{D} \sim P^{\otimes n}$, every η-empirical minimizer $\hat{f} \in \mathcal{H}$ satisfies, with $\hat{h} := \iota(\hat{f})$,*

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

The Rademacher constant $4\beta_\ell$ is the paper's; the deviation constant is explicit ($Cb\sqrt{\log(1/\delta)/n}$ in the paper, with unspecified universal C, becomes $6b\sqrt{2\log(4/\delta)/n}$, from `lem:bv-uniform-deviation-onesided`). The pointwise boundedness of $\mathcal{H}$ makes $\widehat{\mathfrak{R}}_S(\mathcal{H})$ a genuine supremum.

Proof. On the complement of the bad event of `lem:bv-uniform-deviation-onesided`, apply `lem:bv-deterministic` with $\Delta_0 = 2\beta_\ell \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3b\sqrt{2\log(4/\delta)/n}$. ☐

Theorem 96. ***General bias–variance decomposition (generalization gap).*** *Under the hypotheses of `thm:bv-general`, with probability at least $1 - \delta$ over $\mathcal{D} \sim P^{\otimes n}$,* every *$f \in \mathcal{H}$ (in particular every η-empirical minimizer $\hat{f}$) satisfies, with $h := \iota(f)$,*

$$L[h] - \hat{L}[h] \leq (\beta_L + \beta_{\hat{L}})\varepsilon_{\text{imp}} + 2\beta_\ell \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3b\sqrt{\tfrac{2\log(4/\delta)}{n}}.$$

(The paper states this for the η-empirical minimizer only, with $2\beta_\ell \hat{\mathfrak{R}}_S(\mathcal{H}) + Cb\sqrt{\log(1/\delta)/n}$; its proof gives it for every $f \in \mathcal{H}$, which is what we state. The deviation constant is explicit, see `thm:bv-general`.)

Proof. □

Theorem 97. *If $\varepsilon_{\mathrm{imp}} = \sup_{f\in\mathcal{H}} \|f - \iota f\|_\infty \le \varepsilon$ with $\varepsilon \ge 0$, then $|\iota f(x) - f(x)| \le \varepsilon$ for all $f \in \mathcal{H}$, $x \in \mathcal{X}$.*

Proof. □

Theorem 98. ***Implementation-free bias–variance decomposition.*** *Fix $k \ge 0$, $\eta \ge 0$, $n \ge 1$. Assume $\mathcal{H}_k = H \circ B(k, F)$ consists of measurable functions, is sup-norm separable and pointwise bounded, $\mathcal{C}$ is a nonempty class of measurable benchmarks, the loss $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$ is measurable and β_ℓ-Lipschitz in its first argument ($b > 0$, $\beta_\ell \ge 0$), the implementation map ι produces measurable functions, and $\varepsilon_{\mathrm{imp}}(k) = \sup_{f\in\mathcal{H}_k} \|f - \iota f\|_\infty \le \varepsilon_{\mathrm{imp}}$ for a real $\varepsilon_{\mathrm{imp}} \ge 0$. Then for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $\mathcal{D} \sim P^{\otimes n}$, every η-empirical minimizer $\hat{f} \in \mathcal{H}_k$ satisfies, with $\hat{h} := \iota(\hat{f})$,*

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

This is `thm:bv-general` with $d_T(f, g) = \|f-g\|_\infty$ and $\beta_L = \beta_{\hat{L}} = \beta_\ell$ (the loss is β_ℓ-Lipschitz); the Rademacher constant $4\beta_\ell$ is the paper's and the deviation constant is explicit (as in `thm:bv-general`); $\mathcal{H}_k$ is assumed pointwise bounded.

Proof. Apply the general theorem with $d_T(f, g) = \sup_x |f(x) - g(x)|$; the Lipschitz transfer lemmas give $\beta_L = \beta_{\hat{L}} = \beta_\ell$. □

Theorem 99. ***Implementation-free generalization gap.*** *Under the hypotheses of `thm:bv`, with probability at least $1 - \delta$ over $\mathcal{D} \sim P^{\otimes n}$, every $f \in \mathcal{H}_k$ (in particular every η-empirical minimizer) satisfies, with $h := \iota(f)$,*

$$L[h] - \hat{L}[h] \le 2\beta_\ell \varepsilon_{\mathrm{imp}} + 2\beta_\ell \hat{\mathfrak{R}}_S(\mathcal{H}_k) + 3b\sqrt{\tfrac{2\log(4/\delta)}{n}}.$$

(The Rademacher constant $2\beta_\ell$ is the paper's; the deviation constant is explicit, as in `thm:bv-general-gap`.)

Proof. □

3.2 Hidden–output decomposition (proof: Appendix D, thm:hidden-decomp / thm:mixed-sg, prop:hilbert-sg, prop:finite-lipschitz-sg)

Theorem 100. *Let $F \ne \emptyset$ and fix a sign pattern σ. If for every $f \in F$ the Rademacher averages $\{\frac{1}{n}\sum_i \sigma_i h(f(x_i)) : h \in H\}$ are bounded above, and $f \mapsto Z_f(\sigma)$ is bounded above on F, then $\sup_{u\in H\circ F} \frac{1}{n}\sum_i \sigma_i u(x_i) = \sup_{f\in F} Z_f(\sigma)$.*

Proof. □

Theorem 101. *For a single hidden map f_0 with bounded Rademacher averages, $\hat{\mathfrak{R}}_S(H \circ \{f_0\}) = \mathbb{E}_\sigma Z_{f_0}(\sigma)$.*

Proof. □

Theorem 102. ***Hidden–output decomposition under a sub-Gaussian increment condition.*** *Let $n \geq 1$, let $F \subseteq \mathcal{X}^{\mathcal{X}}$ be totally bounded in d_S, let $f_0 \in F$ be an arbitrary anchor, and suppose Assumption* `ass:sg-increment-main` *holds on $\mathfrak{F} \supseteq F$ with constants $A_H, L > 0$. (We also assume that the suprema defining $Z_f(\sigma)$, $f \in F$, are finite, i.e. the Rademacher averages over H are bounded above; this is implicit in the paper.) Then*

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

where $\hat{\mathfrak{R}}_S(H \circ \{f_0\}) = \mathbb{E}_\sigma Z_{f_0}(\sigma)$ (`lem:emp-rademacher-comp-singleton`*). This generalizes the paper, which assumes $\mathrm{id} \in F$ and anchors at $f_0 = \mathrm{id}$, giving $\hat{\mathfrak{R}}_S(H)$ on the right-hand side (*`thm:hidden-decomp-id`*); the anchor is arbitrary because the Dudley bound* `thm:dudley-subgaussian-finite-space` *is anchored at any point. We assume in addition that the entropy integrand $\varepsilon \mapsto \sqrt{\log N(F, d_S, \varepsilon)}$ is integrable on $[0, \operatorname{diam}_S(F)]$; the paper's bound is trivially true with right-hand side $+\infty$ otherwise, whereas Lean's Bochner integral of a non-integrable function is 0.*

Proof. □

Theorem 103. ***Hidden–output decomposition, anchored at the identity (the paper's form).*** *Under the hypotheses of* `thm:hidden-decomp` *with $\mathrm{id} \in F$,*

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

Proof: `thm:hidden-decomp` *with $f_0 = \mathrm{id}$ and $H \circ \{\mathrm{id}\} = H$.*

Proof. □

Theorem 104. *If Assumption* `ass:sg-increment-main` *holds on the semigroup $\langle F_0 \rangle$ with constants (A_H, L), then for every depth $k \geq 0$ with $B(k, F_0)$ totally bounded in d_S,*

$$\hat{\mathfrak{R}}_S(\mathcal{H}_k) \leq \hat{\mathfrak{R}}_S(H) + \frac{12 A_H L}{\sqrt{n}} \int_0^{\operatorname{diam}_S(B(k,F_0))} \sqrt{\log N(B(k, F_0), d_S, \varepsilon)}\, d\varepsilon,$$

with the same constants A_H, L (assuming, as in `thm:hidden-decomp`*, that the entropy integrand of $B(k, F_0)$ is integrable on $[0, \operatorname{diam}_S(B(k, F_0))]$).*

Proof. $\mathrm{id} \in B(k, F_0) \subseteq \langle F_0 \rangle$ and $\mathcal{H}_k = H \circ B(k, F_0)$. □

Theorem 105. *If $h(f(x_i)) = h(g(x_i))$ for all $h \in H$ and $i \leq n$, then $Z_f(\sigma) = Z_g(\sigma)$ for every σ.*

Proof. □

Theorem 106. *If $d_S(f, g) = 0$ then $d(f(x_i), g(x_i)) = 0$ for every $i \leq n$.*

Proof. □

Theorem 107. $n\, d_S(f, g)^2 = \sum_{i=1}^{n} d(f(x_i), g(x_i))^2$.

Proof. □

Theorem 108. *If φ is Lipschitz into a metric space and $d_S(f,g) = 0$, then $\varphi(f(x_i)) = \varphi(g(x_i))$ for every $i \le n$.*

Proof. ☐

Theorem 109. *If φ is L-Lipschitz then $\sum_{i=1}^{n} d(\varphi(f(x_i)), \varphi(g(x_i)))^2 \le L^2\, n\, d_S(f,g)^2$.*

Proof. ☐

Theorem 110. *A counting fraction on $\{\pm 1\}^n$ is at most 1: $\#\{\sigma : p(\sigma)\}/2^n \le 1$.*

Proof. ☐

Theorem 111. $\frac{1}{n}\sum_k \sigma_k \langle w, u_k\rangle = \langle w, \frac{1}{n}\sum_k \sigma_k u_k\rangle$.

Proof. ☐

Theorem 112. ***Hilbert output layers satisfy the increment condition, given the vector Hoeffding inequality.*** *Let $\mathcal{H}$ be a real inner product space and $\Phi : \mathcal{X} \to \mathcal{H}$ be L-Lipschitz. Assume the Rademacher tail bound* `lem:rademacher-hilbert-tail` *holds in $\mathcal{H}$ for samples of size n: $\mathbb{P}_\sigma(\|\sum_i \sigma_i v_i\| > t) \le 2\exp(-t^2/(2\sum_i \|v_i\|^2))$. Then Assumption* `ass:sg-increment-main` *holds for $H = H_\Phi$ with $A_H = 1$ and the same L, for every hidden-layer class $\mathfrak{F}$. Proof: $|Z_f - Z_g| \le \sup_{\|w\|\le 1} |\langle w, \frac{1}{n}\sum_i \sigma_i v_i\rangle| = \|\frac{1}{n}\sum_i \sigma_i v_i\|$ with $v_i = \Phi(f(x_i)) - \Phi(g(x_i))$, and $\sum_i \|v_i\|^2 \le L^2 n d_S(f,g)^2$.*

Proof. ☐

Theorem 113. ***Hilbert output layers satisfy the increment condition.*** *Let $\mathcal{H}$ be a real Hilbert space and $\Phi : \mathcal{X} \to \mathcal{H}$ be L-Lipschitz. Then Assumption* `ass:sg-increment-main` *holds for $H = H_\Phi$ with $A_H = 1$ (and the same L), for every hidden-layer class $\mathfrak{F}$.*

Proof. Combine `prop:hilbert-sg-of-tail` with the vector Hoeffding inequality `lem:rademacher-hilbert-tail` (the only unproved ingredient). ☐

Theorem 114. ***One-dimensional Hilbert output layers.*** *Let $\Phi : \mathcal{X} \to \mathbb{R}$ be L-Lipschitz and $H_\Phi = \{x \mapsto w\,\Phi(x) : |w| \le 1\}$. Then Assumption* `ass:sg-increment-main` *holds for H_Φ with $A_H = 1$ and the same L (this is* `prop:hilbert-sg` *for $\mathcal{H} = \mathbb{R}$, fully proved from the real Hoeffding inequality).*

Proof. ☐

Theorem 115. ***Finite Lipschitz scalar output layers.*** *Let $H = \{h_1, \dots, h_m\}$ be a finite class of real-valued functions on $\mathcal{X}$, each L-Lipschitz. Then Assumption* `ass:sg-increment-main` *holds with $A_H = (1 + \log m/\log 2)^{1/2}$, for every hidden-layer class $\mathfrak{F}$.*

Proof. $|Z_f - Z_g| \le \max_j |S_j|$ with $S_j = \frac{1}{n}\sum_i \sigma_i(h_j(f(x_i)) - h_j(g(x_i)))$; real Hoeffding for each S_j (`lem:rademacher-real-tail`), a union bound over the m functions, and `lem:union-bound-absorb` to absorb m into A_H. ☐

3.3 A conditional Sudakov-type converse (proof: Appendix E, thm:sudakov-type, cor:matching, cor:sudakov-rates)

Theorem 116. *If $G \neq \emptyset$ and for every sign pattern σ the Rademacher averages $\{\frac{1}{n}\sum_i \sigma_i g(x_i) : g \in G\}$ are bounded above, then $\hat{\mathfrak{R}}_S(G) \geq 0$.*

Proof. Apply `lem:rademacher-avg-sup-nonneg` to the family $(g(x_i))_i$, $g \in G$. ☐

Theorem 117. *If $|g(x_i)| \leq C$ for all $g \in G$ and $i \leq n$, then for every sign pattern σ the Rademacher averages $\{\frac{1}{n}\sum_i \sigma_i g(x_i) : g \in G\}$ are bounded above (by $|C|$).*

Proof. ☐

Theorem 118. ***Conditional Sudakov-type lower bound.*** *There is a universal constant $c > 0$ such that the following holds. Let $(\mathcal{X}, d)$ be a pseudometric space, $S = (x_1, \dots, x_n)$ a sample with $n \geq 1$, H an output-layer class, $B \subseteq \mathcal{X}^{\mathcal{X}}$ an arbitrary hidden class, and suppose Assumption `ass:readout-realization-main` holds for B with constants $\kappa, R_{\text{out}} > 0$. Then for every $\varepsilon > 0$,*

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

equivalently the bound with $\sup_{\varepsilon>0}$ on the right. (In Lean $\log M$ is read as 0 when $M = \infty$, which only weakens the inequality.) This generalizes the paper, which states the bound for the word ball $B = B(k, F)$ (`thm:sudakov-type-wordball`); the word-ball structure is not used in the proof, and $B = \emptyset$ is allowed (both sides are then 0). In Lean we add the hypothesis that, for every sign pattern, the Rademacher averages over $H \circ B$ are bounded above (equivalently, that the supremum defining $\hat{\mathfrak{R}}_S(H \circ B)$ is finite); without it the Lean statement is false because an unbounded supremum evaluates to 0. The proof is complete modulo the Bernoulli–Sudakov minoration `thm:bernoulli-sudakov`. Proof: take a maximal 2ε-packing $g_1, \dots, g_M$ of B for d_S ($M = M(B, d_S, 2\varepsilon)$; if $M = \infty$ the right-hand side is 0 in Lean and the claim is `lem:emp-rademacher-nonneg`); the transported vectors $u_j = (h_{g_j}(g_j(x_i)))_i$ are $2\kappa\varepsilon$-separated in $\|\cdot\|_S$ and bounded by R_{out}, so `thm:bernoulli-sudakov` gives $\mathbb{E}_\sigma \max_j \frac{1}{n}\sum_i \sigma_i u_{j,i} \geq c\min\{2\kappa\varepsilon\sqrt{\log M/n}, 4\kappa^2\varepsilon^2/R_{\text{out}}\}$, and the left-hand side is at most $\hat{\mathfrak{R}}_S(H \circ B)$ since $\{h_{g_j} \circ g_j\} \subseteq H \circ B$.

Proof. ☐

Theorem 119. ***Conditional Sudakov-type lower bound for word balls (the paper's form).*** *There is a universal constant $c > 0$ such that, for F a hidden-layer class, $k \geq 0$ and Assumption `ass:readout-realization-main` for $B_k = B(k, F)$ with constants $\kappa, R_{\text{out}} > 0$ (and the boundedness hypothesis of `thm:sudakov-type`), for every $\varepsilon > 0$,*

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

Proof: `thm:sudakov-type` with $B = B(k, F)$, since $\mathcal{H}_k = H \circ B(k, F)$.

Proof. ☐

Theorem 120. ***Conditional Sudakov-type lower bound for a bounded hypothesis class.*** *Under the hypotheses of `thm:sudakov-type`, the boundedness hypothesis holds as soon as $\mathcal{H}_k$ is uniformly bounded on the sample, $|g(x_i)| \leq C$ for all $g \in \mathcal{H}_k$ and $i \leq n$; hence the same lower bound.*

Proof. ☐

Theorem 121. *(Substitution step.) There is a universal constant $c > 0$ such that under Assumption* `ass:readout-realization-main` *for B_k and the boundedness hypothesis of* `thm:sudakov-type`*, for every $\varepsilon_0 > 0$ and every $q \geq 0$ with $q \leq \log M(B_k, d_S, 2\varepsilon_0)$, $\widehat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c \min\{\kappa\varepsilon_0\sqrt{q/n},\ \kappa^2\varepsilon_0^2/R_{\text{out}}\}$.*

Proof. Monotonicity of $q \mapsto \kappa\varepsilon_0\sqrt{q/n}$ in `thm:sudakov-type-wordball`. ☐

Theorem 122. ***Rates under fixed-scale packing lower bounds (exponential).*** *There is a universal constant $c > 0$ such that under Assumption* `ass:readout-realization-main` *for B_k: if $M(B_k, d_S, 2\varepsilon_0) \geq e^{\alpha k}$ for some $\varepsilon_0, \alpha > 0$, then*

$$\widehat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c \min\Big\{\kappa\varepsilon_0\sqrt{\alpha k/n},\ \kappa^2\varepsilon_0^2/R_{\text{out}}\Big\}.$$

Proof. $\alpha k = \log e^{\alpha k} \leq \log M$. ☐

Theorem 123. ***Rates under fixed-scale packing lower bounds (polynomial).*** *There is a universal constant $c > 0$ such that under Assumption* `ass:readout-realization-main` *for B_k: if $k \geq 1$ and $M(B_k, d_S, 2\varepsilon_0) \geq k^\beta$ for some $\varepsilon_0, \beta > 0$, then*

$$\widehat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c \min\Big\{\kappa\varepsilon_0\sqrt{\beta \log k/n},\ \kappa^2\varepsilon_0^2/R_{\text{out}}\Big\}.$$

Proof. $\beta \log k = \log k^\beta \leq \log M$. ☐

Theorem 124. *For $\kappa, \varepsilon_0, R_{\text{out}} > 0$, $n \geq 1$ and $q \geq 0$, the first branch of the minimum is the smaller one as soon as $n \geq R_{\text{out}}^2 q/(\kappa^2\varepsilon_0^2)$: $\kappa\varepsilon_0\sqrt{q/n} \leq \kappa^2\varepsilon_0^2/R_{\text{out}}$.*

Proof. ☐

Theorem 125. ***Matching depth dependence (i).*** *There is a universal constant $c > 0$ such that under Assumption* `ass:readout-realization-main` *for B_k (constants κ, R_{out}), the boundedness hypothesis of* `thm:sudakov-type` *and $\varepsilon_0 > 0$: if $M(B_k, d_S, 2\varepsilon_0) \geq e^{\alpha k}$ for some $\alpha > 0$, then $\widehat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c\,\kappa\varepsilon_0\sqrt{\alpha k/n}$ whenever $n \geq R_{\text{out}}^2 \alpha k/(\kappa^2\varepsilon_0^2)$.*

Proof. Substitute $\varepsilon = \varepsilon_0$; the threshold on n makes the first branch of the minimum the smaller one. ☐

Theorem 126. ***Matching depth dependence (ii).*** *There is a universal constant $c > 0$ such that under Assumption* `ass:readout-realization-main` *for B_k (constants κ, R_{out}), the boundedness hypothesis of* `thm:sudakov-type`*, $k \geq 1$ and $\varepsilon_0 > 0$: if $M(B_k, d_S, 2\varepsilon_0) \geq k^\beta$ for some $\beta > 0$, then $\widehat{\mathfrak{R}}_S(\mathcal{H}_k) \geq c\,\kappa\varepsilon_0\sqrt{\beta \log k/n}$ whenever $n \geq R_{\text{out}}^2 \beta \log k/(\kappa^2\varepsilon_0^2)$.*

Proof. ☐

3.4 Output-layer realization of hidden geometry (Appendix E, prop:global_scalar_observable, prop:linear-interpolation, cor:rkhs-readout)

Definition 127. For a real Hilbert space $\mathcal{H}$, a feature map $\Phi : \mathcal{X} \to \mathcal{H}$ and a radius $R_H \geq 0$, the norm-bounded linear output-layer class is

$$H_{R_H}(\Phi) = \{h_w(x) = \langle w, \Phi(x)\rangle_{\mathcal{H}} : \|w\|_{\mathcal{H}} \leq R_H\}.$$

Definition 128. The reachable sample set of a hidden class B_k on the sample $S = (x_1, \dots, x_n)$ is $U_{k,S} = \{f(x_i) : f \in B_k,\ i \in [n]\}$.

Theorem 129. ***(Global scalar observable.)*** *Fix a sample $S = (x_i)_{i=1}^n$, a hidden class B_k and $U_{k,S} = \{f(x_i) : f \in B_k, i \in [n]\}$. Let $\Phi : \mathcal{X} \to \mathcal{H}$ and assume there are $u \in \mathcal{H}$ with $\|u\| \leq R_H$ and constants $\kappa, R_\Phi > 0$ such that*

$$|\langle u, \Phi(x) - \Phi(y)\rangle| \geq \kappa\, d(x, y) \quad (x, y \in U_{k,S}), \qquad \sup_{x \in \mathcal{X}} \|\Phi(x)\| \leq R_\Phi.$$

Then the readout-realization assumption holds for $H_{R_H}(\Phi)$ on B_k with constants κ and $R_{\text{out}} = R_H R_\Phi$, with the single choice $h_f = h_u = \langle u, \Phi(\cdot)\rangle$ for all $f \in B_k$: $\|h_u \circ f - h_u \circ g\|_S \geq \kappa\, d_S(f, g)$ and $\|h_u \circ f\|_{S,\infty} \leq R_H R_\Phi$ for $f, g \in B_k$.

Proof. Termwise, $\kappa^2 d(f(x_i), g(x_i))^2 \leq |\langle u, \Phi(f(x_i)) - \Phi(g(x_i))\rangle|^2$ by co-Lipschitzness on $U_{k,S}$; average and take square roots. The bound $|\langle u, \Phi(f(x))\rangle| \leq \|u\|\|\Phi(f(x))\| \leq R_H R_\Phi$ is Cauchy–Schwarz. ☐

Theorem 130. ***(Map-dependent finite-set interpolation criterion for linear heads.)*** *Fix $f_1, \dots, f_M \in B_k$ and a sample $S = (x_i)_{i=1}^n$, and let $\Phi : \mathcal{X} \to \mathcal{H}$. Assume that for each j the evaluation operator $T_j : w \mapsto (\langle w, \Phi(f_j(x_i))\rangle)_{i \in [n]}$ admits a right inverse of norm $\leq \Lambda_j$: for every $c \in \mathbb{R}^n$ there is $w \in \mathcal{H}$ with $\|w\| \leq \Lambda_j \|c\|_2$ and $\langle w, \Phi(f_j(x_i))\rangle = c_i$ for all i. Then for every family of code vectors $u^{(1)}, \dots, u^{(M)} \in \mathbb{R}^n$ with $\Lambda_j \|u^{(j)}\|_2 \leq R_H$ there are output layers $h_j \in H_{R_H}(\Phi)$ with $h_j(f_j(x_i)) = u_i^{(j)}$. Consequently, if $\left(\frac{1}{n}\sum_i |u_i^{(j)} - u_i^{(\ell)}|^2\right)^{1/2} \geq \rho$ for $j \neq \ell$ then $\|h_j \circ f_j - h_\ell \circ f_\ell\|_S \geq \rho$ for $j \neq \ell$, and if $\max_{j,i} |u_i^{(j)}| \leq R_{\text{out}}$ (with $R_{\text{out}} \geq 0$) then $\|h_j \circ f_j\|_{S,\infty} \leq R_{\text{out}}$.*

Proof. Set $w_j := R_j u^{(j)}$ and $h_j := \langle w_j, \Phi(\cdot)\rangle$; then $h_j(f_j(x_i)) = u_i^{(j)}$ and $\|w_j\| \leq \Lambda_j \|u^{(j)}\|_2 \leq R_H$. The separation and boundedness claims follow by evaluating on the sample. ☐

Theorem 131. *If $|u_i| \leq R_{\text{out}}$ for all $i \in [n]$ then $\|u\|_2 \leq \sqrt{n}\, R_{\text{out}}$.*

Proof. ☐

Theorem 132. *(Bounded codes.) In ‘prop:linear-interpolation‘, for R_{out}-bounded codes it suffices to take $R_H \geq \sqrt{n}\, R_{\text{out}} \max_j \Lambda_j$ (with $\Lambda_j \geq 0$), which is independent of the number M of maps.*

Proof. ‘lem:code-norm-le‘ gives $\Lambda_j \|u^{(j)}\|_2 \leq \Lambda_j \sqrt{n} R_{\text{out}} \leq R_H$; apply ‘prop:linear-interpolation‘. ☐

Theorem 133. *(Readout realization from interpolation.) Under the hypotheses of ‘prop:linear-interpolation‘, if moreover $\kappa\, d_S(f_j, f_\ell) \leq \rho$ for all $j \neq \ell$, then the readout-realization assumption holds for $H_{R_H}(\Phi)$ on $B_k = \{f_1, \dots, f_M\}$ with constants κ and R_{out}.*

Proof. Take the h_j of ‘prop:linear-interpolation‘ and set $h_{f_j} := h_j$ (choosing an index j for each element of the range). For $f_j \neq f_\ell$ one has $j \neq \ell$, so $\kappa\, d_S(f_j, f_\ell) \leq \rho \leq \|h_j \circ f_j - h_\ell \circ f_\ell\|_S$; for $f_j = f_\ell$ the left-hand side vanishes. □

Theorem 134. *(Right inverse from a well-conditioned Gram matrix.) Let $\varphi_1, \dots, \varphi_n \in \mathcal{H}$ and suppose the Gram matrix $G = (\langle \varphi_i, \varphi_{i'} \rangle)_{i,i'}$ satisfies $c^\top G c \geq \lambda_{\min} \|c\|_2^2$ for all $c \in \mathbb{R}^n$, with $\lambda_{\min} > 0$. Then for every $c \in \mathbb{R}^n$ there is $w \in \mathcal{H}$ with $\langle w, \varphi_i \rangle = c_i$ for all i and $\|w\| \leq \lambda_{\min}^{-1/2} \|c\|_2$.*

Proof. G is injective (if $Ga = 0$ then $\lambda_{\min}\|a\|^2 \leq a^\top G a = 0$), hence surjective: pick a with $Ga = c$ and set $w = \sum_i a_i \varphi_i$. Then $\langle w, \varphi_i \rangle = (Ga)_i = c_i$ and $p := \|w\|^2 = a^\top G a = a^\top c$. From $\lambda_{\min}\|a\|^2 \leq p$ and $p^2 \leq \|a\|^2\|c\|^2$ (Cauchy–Schwarz) we get $\lambda_{\min} p \leq \|c\|^2$, i.e. $\|w\| \leq \lambda_{\min}^{-1/2}\|c\|_2$. □

Theorem 135. ***(RKHS / kernel output layer.)*** *Let $\Phi : \mathcal{X} \to \mathcal{H}$ be a feature map (e.g. the canonical feature map of a positive definite kernel K, so that $\langle \Phi(a), \Phi(a') \rangle = K(a, a')$) and fix $f_1, \dots, f_M$ and a sample S. Suppose that for each j the Gram matrix $G_j = (\langle \Phi(f_j(x_i)), \Phi(f_j(x_{i'})) \rangle)_{i,i'}$ satisfies $c^\top G_j c \geq \lambda_{\min,j} \|c\|_2^2$ with $\lambda_{\min,j} > 0$. Then ‘prop:linear-interpolation‘ applies with $\Lambda_j = \lambda_{\min,j}^{-1/2}$: for all codes $u^{(j)}$ with $\lambda_{\min,j}^{-1/2}\|u^{(j)}\|_2 \leq R_H$, ρ-separated and R_{out}-bounded, there are $h_j \in H_{R_H}(\Phi)$ with $h_j(f_j(x_i)) = u_i^{(j)}$, $\|h_j \circ f_j - h_\ell \circ f_\ell\|_S \geq \rho$ ($j \neq \ell$) and $\|h_j \circ f_j\|_{S,\infty} \leq R_{\text{out}}$.*

Proof. ‘lem:gram-right-inverse‘ applied to $\varphi_i = \Phi(f_j(x_i))$ gives the right-inverse hypothesis of ‘prop:linear-interpolation‘ with $\Lambda_j = \lambda_{\min,j}^{-1/2}$. □

Theorem 136. ***(Finite-dimensional feature map.)*** *Let $\mathcal{H} = \mathbb{R}^m$ and $\Phi : \mathcal{X} \to \mathbb{R}^m$. If for each j the feature matrix $\Phi_j = (\Phi(f_j(x_i)))_{i \in [n]} \in \mathbb{R}^{m \times n}$ has full column rank, quantified as $\sigma_{\min}(\Phi_j)^2 = \lambda_{\min}(\Phi_j^\top \Phi_j) \geq \lambda_{\min,j} > 0$, i.e. $c^\top \Phi_j^\top \Phi_j c \geq \lambda_{\min,j}\|c\|_2^2$, then ‘prop:linear-interpolation‘ applies with $\Lambda_j = \lambda_{\min,j}^{-1/2} = \sigma_{\min}(\Phi_j)^{-1}$. (This is ‘cor:rkhs-readout‘ for $\mathcal{H} = \mathbb{R}^m$, since $\Phi_j^\top \Phi_j$ is the Gram matrix.)*

Proof. Specialization of ‘cor:rkhs-readout‘ to $\mathcal{H} = \mathbb{R}^m$. □

3.5 Tools not printed in the manuscript: a deterministic entropy decomposition (formerly App. E of the ICLR draft; kept in the library, see comparator/README.md)

Theorem 137. ***Composition of covers.*** *Let every $h \in H$ be L_H-Lipschitz, let C_H be an r_1-cover of H for $\|\cdot\|_\infty$ and C_F an r_2-cover of F for d_∞ (centres anywhere). Then $\{h_a \circ f_b : h_a \in C_H, f_b \in C_F\}$ is an $(r_1 + L_H r_2)$-cover of $H \circ F$ for $\|\cdot\|_\infty$: $|h(f(x)) - h_a(f_b(x))| \leq |h(f(x)) - h(f_b(x))| + |h(f_b(x)) - h_a(f_b(x))| \leq L_H d_\infty(f, f_b) + \|h - h_a\|_\infty$.*

Proof. □

Theorem 138. *For L_H-Lipschitz H and radii $r_1, r_2 \geq 0$, $N^{\text{ext}}(H \circ F, \|\cdot\|_\infty, r_1 + L_H r_2) \leq N^{\text{ext}}(H, \|\cdot\|_\infty, r_1) \cdot N^{\text{ext}}(F, d_\infty, r_2)$.*

Proof. □

Theorem 139. ***Composition covering lemma.*** *If every $h \in H$ is L_H-Lipschitz then for every $\varepsilon \geq 0$,*

$$N^{\text{ext}}(H \circ F, \|\cdot\|_\infty, \varepsilon) \leq N^{\text{ext}}(H, \|\cdot\|_\infty, \varepsilon/2) \cdot N^{\text{ext}}(F, d_\infty, \varepsilon/(2L_H)).$$

(For $L_H = 0$ the second radius is 0 by the convention $x/0 = 0$, and the inequality still holds.)

Proof. □

Theorem 140. *$\mathcal{E}_H(u) \geq 0$.*

Proof. □

Theorem 141. *$\mathcal{E}_F(v) \geq 0$.*

Proof. □

Theorem 142. *If $c \leq a\,b$ in $\mathbb{N} \cup \{\infty\}$ with $a, b < \infty$, then $\log c \leq \log a + \log b$ (with $\log 0 = 0$; all three logarithms are ≥ 0).*

Proof. □

Theorem 143. ***Entropy of the composition class.*** *Let every $h \in H$ be L_H-Lipschitz with $L_H > 0$, and let $\varepsilon \in \mathbb{R}$ be such that $N^{\text{ext}}(H, \|\cdot\|_\infty, \varepsilon/2)$ and $N^{\text{ext}}(F, d_\infty, \varepsilon/(2L_H))$ are finite. Then*

$$\sqrt{\log N^{\text{ext}}(H \circ F, \|\cdot\|_\infty, \varepsilon)} \leq \mathcal{E}_H(\varepsilon/2) + \mathcal{E}_F(\varepsilon/(2L_H)).$$

Proof. □

Theorem 144. *FoML's empirical distance is dominated by the uniform distance: $\|u - v\|_S \leq \|u - v\|_\infty$ whenever the latter is finite ($n \geq 1$).*

Proof. □

Theorem 145. *If $G \subseteq \mathbb{R}^{\mathcal{X}}$ is nonempty with finite uniform covering numbers $N^{\text{ext}}(G, \|\cdot\|_\infty, r) < \infty$ for all $r > 0$, then G is totally bounded for the empirical pseudometric $\|\cdot\|_S$ (FoML's 'EmpiricalFunctionSpace'), $n \geq 1$.*

Proof. A finite internal closed $(\varepsilon/2)$-cover for $\|\cdot\|_\infty$ (which exists since $N(G, 2r) \leq N^{\text{ext}}(G, r) < \infty$) is a finite open ε-cover for $\|\cdot\|_S$. □

Theorem 146. ***FoML covering numbers versus uniform external covering numbers.*** *Let $G \subseteq \mathbb{R}^{\mathcal{X}}$ be nonempty and totally bounded for $\|\cdot\|_S$, and let $0 \leq 2r < x$. Then FoML's open-ball covering number of G for $\|\cdot\|_S$ satisfies $N_S^{\text{open}}(G, x) \leq N(G, \|\cdot\|_\infty, 2r) \leq N^{\text{ext}}(G, \|\cdot\|_\infty, r)$ (internal closed $2r$-balls for the sup norm with centres in G are contained in open x-balls for $\|\cdot\|_S$).*

Proof. □

Theorem 147. *If $H \neq \emptyset$ has finite uniform covering numbers at every positive scale, then $u \mapsto \mathcal{E}_H(u)$ is antitone on $(0, \infty)$.*

Proof. □

Theorem 148. *If $F \neq \emptyset$ has finite d_∞ covering numbers at every positive scale, then $v \mapsto \mathcal{E}_F(v)$ is antitone on $(0, \infty)$.*

Proof. □

Theorem 149. $\hat{\mathfrak{R}}_S(\emptyset) = 0$ *(the supremum over the empty class is* 0 *by convention).*

Proof. □

Theorem 150. ***Almost-everywhere comparison of the entropy integrands.*** *Let $H \neq \emptyset$, $F \neq \emptyset$ satisfy Assumptions `ass:ent-readout`, `ass:ent-transition` with $L_H > 0$, $n \geq 1$, and let $N_S(x)$ be FoML's open-ball covering number of $H \circ F$ for $\|\cdot\|_S$. Then for every $\varepsilon > 0$ and Lebesgue-a.e. $x > \varepsilon$,*

$$\sqrt{\log N_S(x)} \leq \mathcal{E}_H(x/4) + \mathcal{E}_F(x/(4L_H)).$$

Proof: for every $y < x$, $N_S(x) \leq N^{\text{ext}}(H \circ F, \|\cdot\|_\infty, y/2) \leq N^{\text{ext}}(H, y/4) N^{\text{ext}}(F, y/(4L_H))$; the right-hand side $\psi(y) = \mathcal{E}_H(y/4) + \mathcal{E}_F(y/(4L_H))$ is antitone in y, hence continuous outside a countable set, and letting $y \uparrow x$ at a continuity point gives the claim.

Proof. □

Theorem 151. ***Deterministic entropy decomposition (FoML form).*** *Under Assumptions `ass:ent-readout` and `ass:ent-transition` with $L_H > 0$, for every sample S of size $n \geq 1$ and every $0 < \varepsilon < B_H/2$,*

$$\hat{\mathfrak{R}}_S(H \circ F) \leq 4\varepsilon + \frac{12}{\sqrt{n}} \int_\varepsilon^{B_H/2} \Big\{ \mathcal{E}_H(x/4) + \mathcal{E}_F(x/(4L_H)) \Big\}\, dx,$$

where $\mathcal{E}_H(u) = \sqrt{\log N^{\text{ext}}(H, \|\cdot\|_\infty, u)}$ and $\mathcal{E}_F(v) = \sqrt{\log N^{\text{ext}}(F, d_\infty, v)}$ (external covering numbers by closed balls). Proof: FoML's Dudley integral for the class $H \circ F$ with the empirical pseudometric $\|\cdot\|_S$ (`dudley_entropy_integral'`), whose internal open-ball covering number at scale x is at most $N^{\text{ext}}(H \circ F, \|\cdot\|_\infty, y/2)$ for every $y < x$ (`lem:foml-covering-le-external-unifFun`, losing a factor 2 in the radius when passing from external to internal covers), followed by the composition covering lemma `lem:ent-composition-entropy` at scale $y/2$ and the limit $y \uparrow x$ almost everywhere (`lem:entropy-integrand-ae`). Compared with the paper's statement (quoted in `thm:rad.decomp.ent.ent`), the scales $x/2$, $x/(2L_H)$ are replaced by $x/4$, $x/(4L_H)$ (the paper implicitly uses covers with centres in the class, while the assumptions here are stated with external covering numbers), the integration range is $[\varepsilon, B_H/2]$ instead of $[0, 2B_H]$, and there is the additive term 4ε from FoML's Dudley bound.

Proof. □

Theorem 152. ***Deterministic entropy decomposition.*** *The paper states: under Assumptions `ass:ent-readout` and `ass:ent-transition`, for every sample $S = (x_1, \dots, x_n)$ with $n \geq 1$,*

$$\hat{\mathfrak{R}}_S(H \circ F) \leq \frac{12}{\sqrt{n}} \int_0^{2B_H} \Big\{ \sqrt{\log N(H, \|\cdot\|_\infty, \varepsilon/2)} + \sqrt{\log N(F, d_\infty, \varepsilon/(2L_H))} \Big\}\, d\varepsilon.$$

The formalized statement is: under Assumptions `ass:ent-readout` and `ass:ent-transition` with $L_H > 0$ and $B_H > 0$, if the entropy integrand $x \mapsto \mathcal{E}_H(x/4) + \mathcal{E}_F(x/(4L_H))$ is integrable on $[0, B_H/2]$, then for every sample S of size $n \geq 1$,

$$\hat{\mathfrak{R}}_S(H \circ F) \leq \frac{12}{\sqrt{n}} \int_0^{B_H/2} \Big\{ \mathcal{E}_H(x/4) + \mathcal{E}_F(x/(4L_H)) \Big\}\, dx,$$

where $\mathcal{E}_H(u) = \sqrt{\log N^{\mathrm{ext}}(H, \|\cdot\|_\infty, u)}$ and $\mathcal{E}_F(v) = \sqrt{\log N^{\mathrm{ext}}(F, d_\infty, v)}$ are the root entropies for the external covering numbers by closed balls of the assumptions.

Deviations from the paper. *(i) We assume in addition that the entropy integrand is integrable on $[0, B_H/2]$; the paper's bound is trivially true with right-hand side $+\infty$ otherwise (e.g. Lipschitz classes on a two-dimensional domain), whereas Lean's Bochner integral of a non-integrable function is 0, so the statement without this hypothesis is false as transcribed. (ii) FoML's Dudley integral (`dudley_entropy_integral'`) uses internal open-ball covers, and converting the external closed-ball covering numbers of the assumptions costs a factor 2 in the radius (`lem:foml-covering-le-external-unifFun`), so the scales are $x/4$ and $x/(4L_H)$ on $[0, B_H/2]$ instead of $\varepsilon/2$, $\varepsilon/(2L_H)$ on $[0, 2B_H]$. Up to these constant rescalings the statement is the paper's.*

Proof: let $\varepsilon \to 0$ in `thm:rad.decomp.ent.ent-foml`, using the integrability hypothesis to compare the integrals over $[\varepsilon, B_H/2]$ and $[0, B_H/2]$.

Proof. □

3.6 Tools not printed in the manuscript: the deterministic entropy decomposition on the sample (formerly App. E of the ICLR draft; kept in the library)

Theorem 153. *$\|u \circ f - v \circ f\|_S = \|u - v\|_{f\circ S}$: the empirical norm of a composition is the empirical norm on the pushed-forward sample $f \circ S = (f(x_1), \dots, f(x_n))$.*

Proof. □

Theorem 154. *A set C is an ε-cover of A for $\|\cdot\|_S$ iff every $u \in A$ has some $c \in C$ with $\|u - c\|_S \le \varepsilon$.*

Proof. □

Theorem 155. *A set C is an ε-cover of A for d_S iff every $f \in A$ has some $c \in C$ with $d_S(f, c) \le \varepsilon$.*

Proof. □

Theorem 156. *$N^{\mathrm{ext}}(H, \|\cdot\|_{f\circ S}, u) \le N_{S,F}(H, u)$ for $f \in F$.*

Proof. □

Theorem 157. *$u \mapsto N_{S,F}(H, u)$ is antitone.*

Proof. □

Theorem 158. *$N_{S,F}(H, u) > 0$ when $H, F \neq \emptyset$.*

Proof. □

Theorem 159. *If $|h(f(x_i)) - h(g(x_i))| \le c\, d(f(x_i), g(x_i))$ for every i ($c \ge 0$), then $\|h \circ f - h \circ g\|_S \le c\, d_S(f, g)$.*

Proof. □

Theorem 160. ***Composition of covers on the sample.*** *Let every $h \in H$ be L_H-Lipschitz on $\mathcal{R}_S(F)$, let $C_F \subseteq F$ be an r_2-cover of F for d_S with centres in F, and for every $f_b \in C_F$ let $C_H(f_b)$ be an r_1-cover of H for $\|\cdot\|_{f_b \circ S}$ (centres anywhere). Then $\{h_a \circ f_b : f_b \in C_F,\ h_a \in C_H(f_b)\}$ is an $(r_1 + L_H r_2)$-cover of $H \circ F$ for $\|\cdot\|_S$: $\|h \circ f - h_a \circ f_b\|_S \le \|h \circ f - h \circ f_b\|_S + \|h \circ f_b - h_a \circ f_b\|_S \le L_H d_S(f, f_b) + \|h - h_a\|_{f_b \circ S}$, where the first estimate uses the Lipschitz property of h at the reachable points $f(x_i), f_b(x_i)$ (this is where $f_b \in F$ is needed).*

Proof. □

Theorem 161. *$H \circ F = \emptyset$ iff $H = \emptyset$ or $F = \emptyset$.*

Proof. □

Theorem 162. *For H L_H-Lipschitz on $\mathcal{R}_S(F)$ and radii $r_1, r_2 \ge 0$, $N^{\mathrm{ext}}(H \circ F, \|\cdot\|_S, r_1 + L_H r_2) \le N_{S,F}(H, r_1) \cdot N(F, d_S, r_2)$, where $N(F, d_S, \cdot)$ is the internal covering number (centres in F). Proof: choose a minimal internal cover C_F of F and, for each $f_b \in C_F$, a minimal external cover of H for $\|\cdot\|_{f_b \circ S}$; apply `lem:comp-cover-sample` and count.*

Proof. □

Theorem 163. ***Composition covering lemma on the sample.*** *If every $h \in H$ is L_H-Lipschitz on $\mathcal{R}_S(F)$ then for every $\varepsilon \ge 0$,*

$$N^{\mathrm{ext}}(H \circ F, \|\cdot\|_S, \varepsilon) \le N_{S,F}(H, \varepsilon/2) \cdot N^{\mathrm{ext}}(F, d_S, \varepsilon/(4L_H)).$$

(For $L_H = 0$ the second radius is 0 by the convention $x/0 = 0$.) Compared with the sup-norm lemma `lem:ent-composition-covering`, the radius of F is $\varepsilon/(4L_H)$ instead of $\varepsilon/(2L_H)$: the cover of F must have its centres in F, and $N(F, d_S, 2r) \le N^{\mathrm{ext}}(F, d_S, r)$.

Proof. □

Theorem 164. *$\mathcal{E}_{H,S}(u) \ge 0$.*

Proof. □

Theorem 165. *$\mathcal{E}_{F,S}(v) \ge 0$.*

Proof. □

Theorem 166. ***Entropy of the composition class on the sample.*** *Let every $h \in H$ be L_H-Lipschitz on $\mathcal{R}_S(F)$ with $L_H > 0$, and let $\varepsilon \in \mathbb{R}$ be such that $N_{S,F}(H, \varepsilon/2)$ and $N^{\mathrm{ext}}(F, d_S, \varepsilon/(4L_H))$ are finite. Then*

$$\sqrt{\log N^{\mathrm{ext}}(H \circ F, \|\cdot\|_S, \varepsilon)} \le \mathcal{E}_{H,S}(\varepsilon/2) + \mathcal{E}_{F,S}(\varepsilon/(4L_H)).$$

Proof. □

Theorem 167. *If $H, F \ne \emptyset$ and $N_{S,F}(H, u) < \infty$ for every $u > 0$, then $u \mapsto \mathcal{E}_{H,S}(u)$ is antitone on $(0, \infty)$.*

Proof. □

Theorem 168. *If $F \ne \emptyset$ has finite d_S covering numbers at every positive scale, then $v \mapsto \mathcal{E}_{F,S}(v)$ is antitone on $(0, \infty)$.*

Proof. □

Theorem 169. *FoML's 'EmpiricalFunctionSpace' of a class G (metric $\|\cdot\|_S$) embeds isometrically into $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_S)$ by evaluating the index.*

Proof. □

Theorem 170. *If $G \subseteq \mathbb{R}^{\mathcal{X}}$ has finite empirical covering numbers $N^{\mathrm{ext}}(G, \|\cdot\|_S, r) < \infty$ for all $r > 0$, then G is totally bounded for $\|\cdot\|_S$ (FoML's 'EmpiricalFunctionSpace').*

Proof. □

Theorem 171. ***FoML covering numbers versus empirical external covering numbers.*** *Let $G \subseteq \mathbb{R}^{\mathcal{X}}$ be nonempty and totally bounded for $\|\cdot\|_S$, and let $0 \le 2r < x$. Then FoML's open-ball covering number of G for $\|\cdot\|_S$ satisfies $N_S^{\mathrm{open}}(G, x) \le N(G, \|\cdot\|_S, 2r) \le N^{\mathrm{ext}}(G, \|\cdot\|_S, r)$ (the two metrics coincide; the factor 2 is the price of the internal cover).*

Proof. □

Theorem 172. ***Almost-everywhere comparison of the entropy integrands on the sample.*** *Let $H, F \neq \emptyset$ satisfy `ass:ent-readout-sample` and `ass:ent-transition-sample` with $L_H > 0$, and let $N_S(x)$ be FoML's open-ball covering number of $H \circ F$ for $\|\cdot\|_S$. Then for every $\varepsilon > 0$ and Lebesgue-a.e. $x > \varepsilon$,*

$$\sqrt{\log N_S(x)} \le \mathcal{E}_{H,S}(x/4) + \mathcal{E}_{F,S}(x/(8L_H)).$$

Proof: for every $y < x$, $N_S(x) \le N^{\mathrm{ext}}(H \circ F, \|\cdot\|_S, y/2) \le N_{S,F}(H, y/4)\, N^{\mathrm{ext}}(F, d_S, y/(8L_H))$; the right-hand side is antitone in y, hence continuous outside a countable set, and we let $y \uparrow x$ at a continuity point.

Proof. □

Theorem 173. ***Deterministic entropy decomposition on the sample (FoML form).*** *Under Assumptions `ass:ent-readout-sample` and `ass:ent-transition-sample` with $L_H > 0$, for every sample S of size $n \ge 1$ and every $0 < \varepsilon < B_H/2$,*

$$\hat{\mathfrak{R}}_S(H \circ F) \le 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{B_H/2} \Big\{ \mathcal{E}_{H,S}(x/4) + \mathcal{E}_{F,S}(x/(8L_H)) \Big\}\, dx,$$

where $\mathcal{E}_{H,S}(u) = \sqrt{\log N_{S,F}(H, u)}$ and $\mathcal{E}_{F,S}(v) = \sqrt{\log N^{\mathrm{ext}}(F, d_S, v)}$. Proof: FoML's Dudley integral for $H \circ F$ with the pseudometric $\|\cdot\|_S$ (`dudley_entropy_integral'`); its internal open-ball covering number at scale x is at most $N^{\mathrm{ext}}(H \circ F, \|\cdot\|_S, y/2)$ for every $y < x$ (`lem:foml-covering-le-external-empFun`), then the composition covering lemma on the sample `lem:ent-composition-entropy-sample` at scale $y/2$ and the limit $y \uparrow x$ almost everywhere (`lem:entropy-integrand-ae-sample`).

Proof. □

Theorem 174. ***Deterministic entropy decomposition on the sample.*** *Let S be a sample of size $n \ge 1$, and suppose `ass:ent-readout-sample` and `ass:ent-transition-sample` hold with $L_H > 0$, $B_H > 0$. If the entropy integrand $x \mapsto \mathcal{E}_{H,S}(x/4) + \mathcal{E}_{F,S}(x/(8L_H))$ is integrable on $[0, B_H/2]$, then*

$$\hat{\mathfrak{R}}_S(H \circ F) \le \frac{12}{\sqrt{n}} \int_0^{B_H/2} \Big\{ \mathcal{E}_{H,S}(x/4) + \mathcal{E}_{F,S}(x/(8L_H)) \Big\}\, dx,$$

where $\mathcal{E}_{H,S}(u) = \sqrt{\log N_{S,F}(H,u)}$ *with* $N_{S,F}(H,u) = \sup_{f\in F} N^{\mathrm{ext}}(H, \|\cdot\|_{f\circ S}, u)$*, and* $\mathcal{E}_{F,S}(v) = \sqrt{\log N^{\mathrm{ext}}(F, d_S, v)}$*.*

Comparison with the paper *(`thm:rad.decomp.ent.ent`, Appendix D). The paper assumes H uniformly L_H-Lipschitz and bounded on $\mathcal{X}$ and covered in the sup norm, and F covered in d_∞; its proof passes from $\|\cdot\|_S$ to $\|\cdot\|_\infty$ at the very first step. Here (W8 of the plan) every hypothesis is on the sample only: H is bounded and L_H-Lipschitz on the reachable points $\mathcal{R}_S(F) = \{f(x_i)\}$, F is covered in d_S, and H is covered in the empirical metrics $\|\cdot\|_{f\circ S}$ of the pushed-forward samples, uniformly in $f \in F$ (this is the honest sample analogue of the sup norm cover: in the composition estimate the output layer is evaluated at $f_b(x_i)$, not at x_i, so a cover of H for $\|\cdot\|_S$ alone would not suffice). The sup-norm assumptions imply the sample assumptions for every S (`cor:rad.decomp.ent.ent-of-sample`). The scales are $x/4$, $x/(8L_H)$ on $[0, B_H/2]$ in place of the paper's $\varepsilon/2$, $\varepsilon/(2L_H)$ on $[0, 2B_H]$: one factor 2 comes from FoML's internal open-ball covers (as in `thm:rad.decomp.ent.ent`) and, for F only, a second factor 2 from the fact that the cover of F must have its centres in F (so that the pushed-forward samples $f_b \circ S$ are admissible), see `lem:ent-composition-covering-sample`. As in the sup-norm theorem, integrability of the integrand is an additional hypothesis.*

Proof: let $\varepsilon \to 0$ in `thm:rad.decomp.ent.ent-sample-foml`.

Proof. ☐

Theorem 175. *For every sample S ($n \geq 1$) and $r \geq 0$, $N^{\mathrm{ext}}(G, \|\cdot\|_S, r) \leq N^{\mathrm{ext}}(G, \|\cdot\|_\infty, r)$: a sup-norm cover is an empirical cover, since $\|u - v\|_S \leq \|u - v\|_\infty$.*

Proof. ☐

Theorem 176. *For every sample S and $r \geq 0$, $N^{\mathrm{ext}}(F, d_S, r) \leq N^{\mathrm{ext}}(F, d_\infty, r)$, since $d_S \leq d_\infty$ (`lem:dS-le-dinf`).*

Proof. ☐

Theorem 177. *$N_{S,F}(H, r) \leq N^{\mathrm{ext}}(H, \|\cdot\|_\infty, r)$ for every sample S ($n \geq 1$): the sup-norm covering number dominates the covering numbers on all pushed-forward samples.*

Proof. ☐

Theorem 178. *Assumption `ass:ent-readout` implies `ass:ent-readout-sample` for every sample S ($n \geq 1$) and every hidden class F.*

Proof. ☐

Theorem 179. *Assumption `ass:ent-transition` implies `ass:ent-transition-sample` for every sample S.*

Proof. ☐

Theorem 180. *If $a \leq b < \infty$ in $\mathbb{N} \cup \{\infty\}$ then $\sqrt{\log a} \leq \sqrt{\log b}$ (with $\log 0 = 0$).*

Proof. ☐

Theorem 181. ***The sup-norm theorem from the sample theorem.*** *Under the sup-norm Assumptions `ass:ent-readout` and `ass:ent-transition` with $L_H > 0$, $B_H > 0$, if $x \mapsto \mathcal{E}_H(x/4) + \mathcal{E}_F(x/(8L_H))$ is integrable on $[0, B_H/2]$, then for every sample S of size $n \geq 1$,*

$$\hat{\mathfrak{R}}_S(H \circ F) \leq \frac{12}{\sqrt{n}} \int_0^{B_H/2} \Big\{ \mathcal{E}_H(x/4) + \mathcal{E}_F(x/(8L_H)) \Big\}\, dx.$$

This is `thm:rad.decomp.ent.ent` *with the scale* $x/(8L_H)$ *in place of* $x/(4L_H)$ *for* F *(the price of the internal cover of* F *in the sample version), obtained from* `thm:rad.decomp.ent.ent-sample-foml`*: the sup-norm assumptions imply the sample assumptions, and* $\mathcal{E}_{H,S} \le \mathcal{E}_H$, $\mathcal{E}_{F,S} \le \mathcal{E}_F$ *pointwise at positive scales.*

Proof. □

Chapter 4

Growth of the hidden class with depth

4.1 Definitions for the growth conditions

Definition 182. For a finite family $f = (f_1, \dots, f_r)$ of self-maps and a word $u = (i_1, \dots, i_k) \in [r]^k$, the associated map is $f_u = f_{i_k} \circ \cdots \circ f_{i_1}$ (the first letter acts first); $f_\emptyset = \mathrm{id}$. In Lean: ‘wordOf f [] = id‘ and ‘wordOf f (i :: u) = wordOf f u   f i‘.

Theorem 183. $f_\emptyset = \mathrm{id}$.

Proof. □

Theorem 184. $f_{(i,u)} = f_u \circ f_i$.

Proof. □

Theorem 185. $f_{(i,u)}(x) = f_u(f_i(x))$.

Proof. □

Definition 186. For a layerwise Lipschitz constant $\Lambda \geq 0$ and $m \geq 0$, the geometric sum $S_m(\Lambda) := \sum_{i=0}^{m-1} \Lambda^i$ (so $S_0 = 0$, $S_1 = 1$) of ‘prop:implementation‘(b); it is the amplification factor of the covering envelope ‘prop:envelope‘.

Theorem 187. $S_0(\Lambda) = 0$.

Proof. □

Theorem 188. $S_1(\Lambda) = 1$.

Proof. □

Theorem 189. *As a real number,* $S_m(\Lambda) = \sum_{i<m} \Lambda^i$.

Proof. □

Definition 190. For probes $P = (p_j)_{j \in J}$ (J finite) the evaluation map is $\mathrm{ev}_P : (\mathcal{X}^{\mathcal{X}}, d_\infty) \to (\mathcal{X}^J, d_{\max})$, $\mathrm{ev}_P(f) = (f(p_j))_{j \in J}$, where $\mathcal{X}^J$ carries the sup-metric.

Definition 191. For a monoid homomorphism $\alpha : H \to \mathcal{X}^{\mathcal{X}}$ (i.e. $\alpha(gh) = \alpha(g) \circ \alpha(h)$ and $\alpha(e) = \mathrm{id}$), the orbit map is $g \mapsto \alpha(g) \in \mathcal{X}^{\mathcal{X}}$.

Theorem 192. $\alpha(e) = \mathrm{id}$.

Proof. ☐

Theorem 193. $\alpha(gh) = \alpha(g) \circ \alpha(h)$.

Proof. ☐

Definition 194. The memory length of P1': $m(\varepsilon) := L + \lceil \log_{1/c}(2\,\mathrm{diam}(K)/\varepsilon) \rceil$ (a natural number; the ceiling of a non-positive real is 0).

Definition 195. The memory state space $X = E^{\mathbb{N}}$ (the paper's $\ell_\infty(E)$; we allow all sequences since the bounded metric below is finite anyway).

Theorem 196. *$X = E^{\mathbb{N}}$ is nonempty (it contains 0).*

Proof. ☐

Theorem 197. *If $a \leq b + c$ in $[0, \infty]$ then $\min\{1, a\} \leq \min\{1, b\} + \min\{1, c\}$.*

Proof. If $b, c \leq 1$ this is $a \leq b + c$; otherwise the right side is at least 1. ☐

Definition 198. The bounded sup metric on $X = E^{\mathbb{N}}$: $d_X(x, y) = \sup_j \min\{1, \|x_j - y_j\|_E\}$. It is a pseudo-emetric (indeed a metric with values in $[0, 1]$).

Theorem 199. $d_X(x, y) = \sup_j \min\{1, \|x_j - y_j\|_E\}$.

Proof. ☐

Theorem 200. $d_X(x, y) \leq 1$.

Proof. ☐

Definition 201. The reset map $r(x) = 0$.

Definition 202. The writer $g_u(x_0, x_1, \dots) = (u, x_0, x_1, \dots)$.

Theorem 203. $r(x)_j = 0$.

Proof. ☐

Theorem 204. $g_u(x)_0 = u$.

Proof. ☐

Theorem 205. $g_u(x)_{i+1} = x_i$.

Proof. ☐

Definition 206. The expander $A(x_0, x_1, \dots) = (\lambda x_0, \lambda x_1, \dots)$.

Theorem 207. $A(x)_j = \lambda x_j$.

Proof. ☐

Definition 208. The hidden-layer class $F = \{r, A\} \cup \{g_u : u \in G\}$.

Definition 209. For $u = (u_0, \dots, u_{k-1}) \in E^k$ the word $w_u = A \circ g_{u_{k-1}} \circ A \circ g_{u_{k-2}} \circ \cdots \circ A \circ g_{u_0} \circ r$ (so u_0 is written first and u_{k-1} last). In Lean it is defined by recursion on k: $w_{()} = r$ and $w_{(u,a)} = A \circ g_a \circ w_u$.

Theorem 210. *$w_{()} = r$.*

Proof. □

Theorem 211. *$w_u = A \circ g_{u_k} \circ w_{u|_k}$ for $u \in E^{k+1}$.*

Proof. □

Definition 212. $W_k = \{w_u : u \in G^k\}$.

4.2 Basic facts on covering and packing numbers (Appendix A, incl. lem:probes-packing)

Theorem 213. *(Sub-multiplicativity.) Let $F, G \subseteq \mathcal{X}^{\mathcal{X}}$ and suppose every $f \in F$ is λ-Lipschitz. Then for all $\varepsilon, \delta \geq 0$,*

$$N^{\text{ext}}(F \circ G, d_\infty, \varepsilon + \lambda\delta) \leq N^{\text{ext}}(F, d_\infty, \varepsilon)\, N^{\text{ext}}(G, d_\infty, \delta).$$

(This is the paper's $N(FG, \varepsilon + \delta) \leq N(F, \varepsilon/\rho)N(G, \delta/\lambda)$ with $\rho = 1$, since right composition is 1-Lipschitz for d_∞; we write $\lambda\delta$ in place of δ/λ to avoid dividing by $\lambda = 0$.)

Proof. Take an ε-cover C_F of F and a δ-cover C_G of G of minimal cardinality. For $f \in F$, $g \in G$ pick $f' \in C_F$, $g' \in C_G$ with $d_\infty(f, f') \leq \varepsilon$, $d_\infty(g, g') \leq \delta$; then $d_\infty(f \circ g, f' \circ g') \leq d_\infty(f \circ g, f \circ g') + d_\infty(f \circ g', f' \circ g') \leq \lambda\delta + \varepsilon$, so $C_F \circ C_G$ (of cardinality $\leq |C_F|\,|C_G|$) is a cover of $F \circ G$. □

Theorem 214. *The evaluation map ev_P is 1-Lipschitz from d_∞ to the sup-metric on $\mathcal{X}^J$.*

Proof. $\max_j d(f(p_j), g(p_j)) \leq \sup_x d(f(x), g(x))$. □

Theorem 215. *(Probes and packing.) Let $A \subseteq (\mathcal{X}^{\mathcal{X}}, d_\infty)$, let P be a finite family of probes and let $T \subseteq \mathrm{ev}_P(A)$ be δ-separated: $d_{\max}(y, z) \geq \delta$ for all distinct $y, z \in T$. Then for every ε with $2\varepsilon < \delta$, $|T| \leq N^{\text{ext}}(A, d_\infty, \varepsilon)$.*

Proof. T is 2ε-separated (strictly), so $|T| \leq M(\mathrm{ev}_P(A), 2\varepsilon) \leq M(A, 2\varepsilon) \leq N^{\text{ext}}(A, \varepsilon)$ by the Lipschitz embedding lemma (with $K = 1$) and the packing–covering comparison. □

Theorem 216. *Under the hypotheses of the previous lemma, also $|T| \leq N(A, d_\infty, \varepsilon)$ (internal covering number).*

Proof. □

Theorem 217. *(Indexed form of ‘lem:probes-packing‘.) Let $(g_i)_{i \in I}$ be a family of maps in $A \subseteq (\mathcal{X}^{\mathcal{X}}, d_\infty)$ and P a finite family of probes such that $d_{\max}(\mathrm{ev}_P(g_i), \mathrm{ev}_P(g_j)) \geq \delta > 0$ for all $i \neq j$. Then $|I| \leq N^{\text{ext}}(A, d_\infty, \varepsilon)$ whenever $2\varepsilon < \delta$.*

Proof. The map $i \mapsto \mathrm{ev}_P(g_i)$ is injective (distinct indices have evaluations at distance $\geq \delta > 0$), and its range is a δ-separated subset of $\mathrm{ev}_P(A)$; apply ‘lem:probes-packing‘. □

4.3 Tools not printed in the manuscript: compact Arzelà–Ascoli for self-maps (formerly App. P of the ICLR draft; kept in the library, see comparator/README.md)

Theorem 218. *If $\mathcal{X}$ is complete then $(\mathcal{X}^{\mathcal{X}}, d_\infty)$ is complete (uniform limits of Cauchy sequences). This is Mathlib's instance on 'X → X'.*

Proof. $\square$

Theorem 219. ***(Metric modulus criterion.)*** *Let $(\mathcal{X}, d)$ be a totally bounded (pseudo)metric space. If $H \subseteq \mathcal{X}^{\mathcal{X}}$ has a common modulus of continuity, namely there is $\omega : \mathbb{R} \to \mathbb{R}$ with $\omega(r) \to 0$ as $r \downarrow 0$ (for every $\varepsilon > 0$ there is $r > 0$ with $\omega(t) \le \varepsilon$ for $0 \le t \le r$) and*

$$d(f(x), f(y)) \le \omega(d(x, y)) \qquad (f \in H,\ x, y \in \mathcal{X}),$$

then H is totally bounded in d_∞.

Proof. Fix $\eta > 0$ (below the prescribed radius). Choose $\rho > 0$ with $\omega(t) \le \eta/4$ for $t \le \rho$, a finite ρ-net P and a finite $\eta/4$-net Q of $\mathcal{X}$. To $f \in H$ assign the map $P \to Q$, $p \mapsto q(f(p))$ with $d(f(p), q(f(p))) < \eta/4$. There are finitely many assignments; choose one representative $g \in H$ for each occurring assignment. If f and g have the same assignment then for every x, picking $p \in P$ with $d(x,p) < \rho$, $d(f(x), g(x)) \le d(f(x), f(p)) + d(f(p), q(f(p))) + d(q(g(p)), g(p)) + d(g(p), g(x)) \le \eta$. Thus the representatives form a finite η-net of H for d_∞. $\square$

Theorem 220. ***(Modulus extraction on compact domains.)*** *Let $(\mathcal{X}, d)$ be compact and $H \subseteq \mathcal{X}^{\mathcal{X}}$ equicontinuous. Then H admits a common monotone modulus of continuity ω with $\omega(r) \to 0$ as $r \downarrow 0$ and $d(f(x), f(y)) \le \omega(d(x, y))$ for all $f \in H$, $x, y \in \mathcal{X}$. In particular 'thm:maa' applies.*

Proof. Set $\omega(r) := \sup\{d(f(x), f(y)) : f \in H,\ d(x, y) \le r\}$ (bounded by $\operatorname{diam} \mathcal{X}$, and 0 for $r < 0$). It is monotone, and uniform equicontinuity (equicontinuity on a compact space) gives $\omega(r) \le \varepsilon$ for $r \le \delta/2$. $\square$

Theorem 221. *(Forward direction of 'thm:caa'.) If $\mathcal{X}$ is compact and $H \subseteq \mathcal{X}^{\mathcal{X}}$ is equicontinuous, then H is totally bounded in d_∞.*

Proof. Extract a common modulus ('lem:oaa') and apply the modulus criterion ('thm:maa'); a compact space is totally bounded. $\square$

Theorem 222. *(Converse direction of 'thm:caa'.) If $H \subseteq C(\mathcal{X}, \mathcal{X})$ is totally bounded in d_∞ then H is equicontinuous. (Continuity of the members of H is needed; compactness of $\mathcal{X}$ is not.)*

Proof. Fix x_0 and $\varepsilon > 0$, and take a finite $\varepsilon/3$-net $t \subseteq H$. Each $g \in t$ is continuous at x_0, so for x near x_0 we have $d(g(x_0), g(x)) < \varepsilon/3$ for all $g \in t$ simultaneously. For $f \in H$ choose $g \in t$ with $d_\infty(f, g) < \varepsilon/3$; then $d(f(x_0), f(x)) \le d(f(x_0), g(x_0)) + d(g(x_0), g(x)) + d(g(x), f(x)) < \varepsilon$. $\square$

Theorem 223. ***(Compact Arzelà–Ascoli for self-maps.)*** *Let $(\mathcal{X}, d)$ be a compact (pseudo)metric space and let $H \subseteq C(\mathcal{X}, \mathcal{X})$. Then H is totally bounded in d_∞ if and only if H is equicontinuous.*

Proof. $\square$

Theorem 224. *(Consequence of 'thm:caa'.) If $\mathcal{X}$ is compact and $H \subseteq \mathcal{X}^{\mathcal{X}}$ is equicontinuous, then the closure of H in $(\mathcal{X}^{\mathcal{X}}, d_\infty)$ is compact.*

Proof. $(\mathcal{X}^{\mathcal{X}}, d_\infty)$ is complete since $\mathcal{X}$ is compact (hence complete), and the closure of a totally bounded set in a complete space is compact. ☐

Theorem 225. ***(Saturation for compact equicontinuous semigroups.)*** *Let $(\mathcal{X}, d)$ be compact and $F \subseteq \mathcal{X}^{\mathcal{X}}$. If the generated semigroup $\langle F\rangle$ is equicontinuous, then for every $\varepsilon > 0$ and every k*

$$N^{\text{ext}}(B(k,F), d_\infty, \varepsilon) \le N^{\text{ext}}(\overline{\langle F\rangle}^{d_\infty}, d_\infty, \varepsilon) < \infty.$$

In particular the covering number of the word ball does not grow with the depth k.

Proof. $B(k,F) \subseteq \langle F\rangle \subseteq \overline{\langle F\rangle}$ and the external covering number is monotone in the set; the closure is totally bounded by 'thm:caa', so its covering number is finite. ☐

Theorem 226. *(Internal version of 'cor:aa-semigroup-saturation'.) Under the same hypotheses, for every $\varepsilon > 0$ and every k, $N(B(k,F), d_\infty, \varepsilon) \le N(\overline{\langle F\rangle}^{d_\infty}, d_\infty, \varepsilon/2) < \infty$ (internal covering numbers are not monotone in the set, whence the loss of a factor 2 in the radius).*

Proof. Mathlib's 'coveringNumber_subset_le' for the inclusion, and finiteness of the internal covering number of a totally bounded set. ☐

4.4 Saturation: equicontinuous semigroups and contraction to an invariant set (Appendix F.1, P1, P1')

Theorem 227. *$f \in B(k,F)$ if and only if $f \in F^l$ for some $l \le k$.*

Proof. Induction on k using $B(k+1,F) = B(k,F) \cup F^{k+1}$. ☐

Theorem 228. *Every word of length $n+L$ factors as $w = w_2 \circ w_1$ with $w_2 \in F^n$ and $w_1 \in F^L$.*

Proof. Induction on n: for $n = 0$ take $w_2 = \text{id}$; for $w = a \circ b$ with $a \in F$, $b \in F^{n+L}$, factor $b = b_2 \circ b_1$ and take $w_2 = a \circ b_2$. ☐

Theorem 229. *If $f(A) \subseteq A$ for every $f \in F$, then $w(a) \in A$ for every word $w \in F^m$ and every $a \in A$.*

Proof. Induction on m. ☐

Theorem 230. *If every $f \in F$ is c-Lipschitz then every word $w \in F^m$ is c^m-Lipschitz.*

Proof. Induction on m; Lipschitz constants multiply under composition. ☐

Theorem 231. *If the generators are non-expanding (1-Lipschitz) then every element of $B(k,F)$ is non-expanding.*

Proof. Induction on k; id is 1-Lipschitz and $1 \cdot 1 = 1$. ☐

Theorem 232. *If the generators are non-expanding then the whole semigroup $\langle F\rangle$ is non-expanding: $\operatorname{lip} F \le 1$ implies $\operatorname{lip}\langle F\rangle \le 1$.*

Proof. $\langle F\rangle = \bigcup_k B(k,F)$. ☐

Theorem 233. ***(P1: equicontinuous semigroup on a compact domain saturates, hypothesis 1.)*** *Assume $\mathcal{X}$ is compact and the semigroup $\langle F\rangle$ is precompact (totally bounded) in d_∞. Then, with $G := \overline{\langle F\rangle}^{d_\infty}$, for all $\varepsilon > 0$ and all k,*

$$N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon) \le N^{\mathrm{ext}}(G, d_\infty, \varepsilon) < \infty,$$

hence no dependence on k. (Compactness of $\mathcal{X}$ is not needed for this hypothesis.)

Proof. $B(k,F) \subseteq G$ and monotonicity of the external covering number; G is totally bounded as the closure of a totally bounded set, so its covering number is finite. □

Theorem 234. ***(P1, hypothesis 2a.)*** *Assume $\mathcal{X}$ is compact and the semigroup $\langle F\rangle$ is equicontinuous. Then for all $\varepsilon > 0$ and all k, $N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon) \le N^{\mathrm{ext}}(\overline{\langle F\rangle}, d_\infty, \varepsilon) < \infty$.*

Proof. □

Theorem 235. ***(P1, hypothesis 2b.)*** *Assume $\mathcal{X}$ is compact and the semigroup $\langle F\rangle$ is uniformly Lipschitz: every $g \in \langle F\rangle$ is K-Lipschitz for a common K. Then for all $\varepsilon > 0$ and all k, $N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon) \le N^{\mathrm{ext}}(\overline{\langle F\rangle}, d_\infty, \varepsilon) < \infty$.*

Proof. □

Theorem 236. ***(P1, hypothesis 2c.)*** *Assume $\mathcal{X}$ is compact and the generators are non-expanding: $\operatorname{lip} F \le 1$. Then for all $\varepsilon > 0$ and all k, $N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon) \le N^{\mathrm{ext}}(\overline{\langle F\rangle}, d_\infty, \varepsilon) < \infty$.*

Proof. □

Theorem 237. *If $0 < c < 1$, $\varepsilon > 0$ and $m(\varepsilon) \le n + L$ then $c^n \operatorname{diam}(K) \le \varepsilon/2$.*

Proof. If $\operatorname{diam}(K) = 0$ this is trivial. Otherwise $\log_{1/c}(2\operatorname{diam}(K)/\varepsilon) \le \lceil\cdot\rceil \le n$, i.e. $2\operatorname{diam}(K)/\varepsilon \le (1/c)^n$, which rearranges to the claim. □

Theorem 238. *(Long words are almost constant.) Under the hypotheses of 'cond:p1-ucont', for every word $w \in F^l$ with $l \ge m(\varepsilon)$ and all $x, y \in \mathcal{X}$,*

$$d(w(x), w(y)) \le c^{l-L} \operatorname{diam}(K) \le \varepsilon/2.$$

In particular $d_\infty(w, \mathrm{const}_{w(a_0)}) \le \varepsilon/2$ for every a_0.

Proof. Write $l = n + L$ and $w = w_2 \circ w_1$ with $w_2 \in F^n$ (c^n-Lipschitz) and $w_1 \in F^L$ (with values in K). Then $d(w(x), w(y)) \le c^n d(w_1(x), w_1(y)) \le c^n \operatorname{diam}(K) \le \varepsilon/2$. □

Theorem 239. *(Long words.) Under the hypotheses of 'cond:p1-ucont', the set of all words of length $\ge m(\varepsilon)$ satisfies*

$$N^{\mathrm{ext}}\Big(\bigcup_{l \ge m(\varepsilon)} F^l, d_\infty, \varepsilon\Big) \le N^{\mathrm{ext}}(A, d, \varepsilon/2).$$

Indeed each such word w is $\varepsilon/2$-close to the constant map $\mathrm{const}_{w(a_0)}$ with $w(a_0) \in A$, so the constant maps at the points of an $\varepsilon/2$-net of A form an ε-cover.

Proof. Take a minimal $\varepsilon/2$-cover C of A; for a long word w pick $q \in C$ with $d(w(a_0), q) \le \varepsilon/2$; then $d(w(x), q) \le d(w(x), w(a_0)) + d(w(a_0), q) \le \varepsilon$ for every x, so $\{\mathrm{const}_q : q \in C\}$ is an ε-cover. □

Theorem 240. *(Short words; subadditivity over finitely many sets.) For a finite family $(A_i)_{i\in s}$, $N^{\text{ext}}(\bigcup_{i\in s} A_i, \varepsilon) \le \sum_{i\in s} N^{\text{ext}}(A_i, \varepsilon)$. Applied to $A_l = F^l$, $l < m(\varepsilon)$, this covers the short words.*

Proof. Induction on the finite index set using 'lem:subadditivity'. □

Theorem 241. ***(P1': contraction to a compact invariant set.)*** *Assume*

1. *(uniform contraction) every $f \in F$ is c-Lipschitz with $0 < c < 1$;*
2. *(invariant set) there is a nonempty F-invariant set $A \subseteq \mathcal{X}$ ($f(A) \subseteq A$ for all $f \in F$);*
3. *(bounded absorbing set) there are $L \in \mathbb{N}$ and a bounded set $K \subseteq \mathcal{X}$ with $f(\mathcal{X}) \subseteq K$ for every word $f \in F^L$ of length L.*

Then for every $\varepsilon > 0$, with $m(\varepsilon) := L + \lceil \log_{1/c}(2\operatorname{diam}(K)/\varepsilon)\rceil$, for every $k \ge 0$,

$$N^{\text{ext}}(B(k,F), d_\infty, \varepsilon) \le N^{\text{ext}}(A, d, \varepsilon/2) + \sum_{l=0}^{m(\varepsilon)-1} N^{\text{ext}}(F^l, d_\infty, \varepsilon).$$

In particular the right-hand side is independent of k. (This generalizes the paper, which assumes A compact, K compact and $k \ge m(\varepsilon)$: the proof uses of A only that it is nonempty and invariant, of K only its boundedness, and the bound holds for every k. The right-hand side is finite whenever each $N^{\text{ext}}(F^l, d_\infty, \varepsilon)$, $l < m(\varepsilon)$, is finite, which the paper does not assume.)

Proof. $B(k,F) \subseteq \bigcup_{l\ge m(\varepsilon)} F^l \cup \bigcup_{l<m(\varepsilon)} F^l$; apply monotonicity, subadditivity, 'lem:p1ucont-long-words' and 'lem:p1ucont-short-words'. □

4.5 Polynomial growth under nilpotent control (Appendix F.2, P2)

Theorem 242. *Let $\alpha : H \to \mathcal{X}^{\mathcal{X}}$ be a homomorphism, assume the length $g \mapsto d_H(e,g)$ is subadditive, $d_H(e,gh) \le d_H(e,g) + d_H(e,h)$, and let $S \subseteq H$ satisfy $d_H(e,s) \le R_S$ for all $s \in S$ and $F \subseteq \alpha(S)$. Then for every k,*

$$B(k,F) \subseteq \alpha(\overline{B}_H(e, kR_S)).$$

Proof. Induction on k. For $k = 0$, $\mathrm{id} = \alpha(e)$ and $e \in \overline{B}_H(e,0)$. For $k+1$: an element of $B(k,F)$ lies in $\alpha(\overline{B}_H(e,kR_S)) \subseteq \alpha(\overline{B}_H(e,(k+1)R_S))$; an element $\alpha(s)\circ\alpha(h)$ with $s \in S$, $d_H(e,h) \le kR_S$ equals $\alpha(sh)$ and $d_H(e,sh) \le d_H(e,s) + d_H(e,h) \le R_S + kR_S$. □

Theorem 243. *If the orbit map is L_α-Lipschitz, $d_\infty(\alpha(g),\alpha(h)) \le L_\alpha d_H(g,h)$, then for every $B \subseteq H$ and $\delta \ge 0$,*

$$N^{\text{ext}}(\alpha(B), d_\infty, L_\alpha\delta) \le N^{\text{ext}}(B, d_H, \delta).$$

Proof. This is the Lipschitz embedding lemma 'lem:lipschitz-embedding'. □

Theorem 244. *Under the hypotheses of 'lem:p2-wordball-subset-ball' and 'lem:p2-transfer', for every k and $\delta \ge 0$,*

$$N^{\text{ext}}(B(k,F), d_\infty, L_\alpha\delta) \le N^{\text{ext}}(\overline{B}_H(e,kR_S), d_H, \delta).$$

Proof. Monotonicity of the external covering number in the set, then 'lem:p2-transfer'. □

Theorem 245. *For $\varepsilon, L_\alpha > 0$, $R_S \geq 0$, $k \geq 0$ and $D \geq 0$,*

$$\Big(1 + \frac{kR_S}{\varepsilon/L_\alpha}\Big)^D \leq \max(1, R_S L_\alpha)^D \Big(1 + \frac{k}{\varepsilon}\Big)^D.$$

Proof. With $M = \max(1, R_S L_\alpha)$ we have $1 \leq M$ and $R_S L_\alpha \leq M$, so $1 + (R_S L_\alpha)(k/\varepsilon) \leq M + M(k/\varepsilon) = M(1 + k/\varepsilon)$; raise to the power $D \geq 0$ and use $(ab)^D = a^D b^D$. ☐

Theorem 246. ***(P2: nilpotent control grows polynomially.)*** *Let (H, d_H) be a group with a pseudo-emetric, 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 its balls have polynomial entropy of degree $D \geq 0$: there is $1 \leq C_H < \infty$ such that for all $R \geq 0$ and $\delta > 0$,*

$$N^{\text{ext}}(\overline{B}_H(e, R), d_H, \delta) \leq C_H \Big(1 + \frac{R}{\delta}\Big)^D.$$

Suppose H acts on $\mathcal{X}$ through a homomorphism $\alpha : H \to \mathcal{X}^{\mathcal{X}}$, that there is a bounded set $S \subseteq H$ ($d_H(e, s) \leq R_S$ for $s \in S$) with $F \subseteq \alpha(S)$, and that the orbit map is Lipschitz in the uniform metric: $d_\infty(\alpha(g), \alpha(h)) \leq L_\alpha d_H(g, h)$ with $0 < L_\alpha < \infty$. Then with

$$C := C_H \max(1, R_S L_\alpha)^D$$

one has, for every $\varepsilon > 0$ and every $k \geq 0$,

$$N^{\text{ext}}(B(k, F), d_\infty, \varepsilon) \leq C\Big(1 + \frac{k}{\varepsilon}\Big)^D.$$

(In Lean the ball-entropy hypothesis and the conclusion are stated in $[0, \infty]$ as $N \leq \text{ofReal}(\cdots)$, which in particular asserts finiteness; the bound holds for $k = 0$ as well.)

Proof. Put $\delta := \varepsilon/L_\alpha$, so $L_\alpha \delta = \varepsilon$. By 'lem:p2-transfer-wordball', $N^{\text{ext}}(B(k, F), d_\infty, \varepsilon) \leq N^{\text{ext}}(\overline{B}_H(e, kR_S), d_H, \delta) \leq C_H(1 + kR_S/\delta)^D$, and 'lem:p2-poly-factor' bounds the last factor by $\max(1, R_S L_\alpha)^D (1 + k/\varepsilon)^D$. ☐

Theorem 247. *(Real-valued form of P2.) Under the hypotheses of 'cond:p2-nilp', for every $\varepsilon > 0$ and $k \geq 0$, $N^{\text{ext}}(B(k, F), d_\infty, \varepsilon) \leq C_H \max(1, R_S L_\alpha)^D (1 + k/\varepsilon)^D$ as real numbers (the covering number being finite).*

Proof. Take $C = C_H \max(1, R_S L_\alpha)^D$ and apply 'cond:p2-nilp'; a quantity bounded by $\text{ofReal}(b)$ is finite and has real part $\leq b$. ☐

Theorem 248. *(Diameter envelope.) Under the hypotheses of 'lem:p2-wordball-subset-ball' and with the orbit map L_α-Lipschitz, for all $f, g \in B(k, F)$,*

$$d_\infty(f, g) \leq 2L_\alpha R_S k,$$

i.e. $\text{diam}_\infty B(k, F) \leq 2L_\alpha R_S k$ *(and a fortiori $D_k(S) \leq 2L_\alpha R_S k$ for every sample S).*

Proof. Write $f = \alpha(a)$, $g = \alpha(b)$ with $d_H(e, a), d_H(e, b) \leq kR_S$ ('lem:p2-wordball-subset-ball'). Then $d_\infty(\alpha(a), \alpha(b)) \leq L_\alpha d_H(a, b) \leq L_\alpha(d_H(a, e) + d_H(e, b)) \leq 2L_\alpha R_S k$. ☐

Theorem 249. *On a bounded state space, $d(x, y) \leq D_{\mathcal{X}}$ for all $x, y \in \mathcal{X}$, every pair of self-maps satisfies $d_\infty(f, g) \leq D_{\mathcal{X}}$; in particular* $\text{diam}_\infty B(k, F) \leq \text{diam}(\mathcal{X})$ *for every k.*

Proof. $\sup_x d(f(x), g(x)) \leq D_{\mathcal{X}}$ termwise. ☐

4.6 Exponential growth: free semigroups and ping–pong coding (Appendix F.3, E1, E1', E2)

Theorem 250. *If $w \in F^m$ and $g \in F$ then $w \circ g \in F^{m+1}$: the words of length m are also closed under right composition with generators.*

Proof. Induction on m: $\mathrm{id} \circ g = g \circ \mathrm{id}$, and $(a \circ b) \circ g = a \circ (b \circ g)$. □

Theorem 251. *$f_u \in F^{|u|}$ where $F = \{f_1, \dots, f_r\}$.*

Proof. Induction on u using the previous lemma. □

Theorem 252. *$F^m \subseteq B(m, F)$.*

Proof. □

Theorem 253. *If $|u| \le k$ then $f_u \in B(k, F)$ where $F = \{f_1, \dots, f_r\}$.*

Proof. □

Theorem 254. *Let $f = (f_1, \dots, f_r)$ and let P be a finite family of probes such that the evaluations $\mathrm{ev}_P(f_u)$, $u \in [r]^k$, are pairwise at sup-distance $\ge \delta > 0$. Then $N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon) \ge r^k$ for every ε with $2\varepsilon < \delta$.*

Proof. Index the words of length k by $[r]^{[k]}$ ('List.ofFn'), which has r^k elements, and apply 'lem:probes-packing-family'. □

Theorem 255. ***(E1: free semigroup with one-point uniform separation.)*** *Let $F = \{f_1, \dots, f_r\}$, $r \ge 2$, and suppose there are a base point $x_* \in \mathcal{X}$ and $\delta > 0$ such that for every k and all distinct words $u \ne v \in [r]^k$,*

$$d(f_u(x_*), f_v(x_*)) \ge \delta.$$

Then for every k and every $\varepsilon < \delta/2$, $N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon) \ge r^k$. (The paper's hypothesis (1), that $u \mapsto f_u$ is injective on $[r]^k$, follows from the separation hypothesis since $\delta > 0$, so it is omitted; the hypothesis $r \ge 2$ is not used in the proof.)

Proof. Apply 'lem:pow-le-covering-of-probes' with the single probe x_*. □

Theorem 256. ***(E1': equal-length coding.)*** *Let $F = \{f_1, \dots, f_r\}$, $r \ge 2$, and suppose there are $x_* \in \mathcal{X}$ and $\delta > 0$ with $d(f_u(x_*), f_v(x_*)) \ge \delta$ for all distinct $u, v \in [r]^k$ and all k. Then $N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon) \ge r^k$ for all k and all $\varepsilon < \delta/2$. (This generalizes the paper, which assumes in addition that the generators are isometries, that d_∞ is finite on $\langle F \rangle$ and that $u \mapsto f_u$ is injective on $[r]^k$: none of these is needed for the lower bound, which follows from 'cond:e1-free-iso' in one line; the isometry hypothesis only serves, in the paper, to make the covering numbers meaningful.)*

Proof. □

Theorem 257. *If $A \subseteq \mathcal{X}$ satisfies $f_i(A) \subseteq A$ for all i, then $f_u(A) \subseteq A$ for every word u.*

Proof. Induction on u. □

Theorem 258. *(Probe construction for E2.) Under the hypotheses of 'cond:e2-pingpong', for every word u there is a point $x_u \in Q = \{q\} \cup \bigcup_j V_j$ with $f_u(x_u) = q$ and $f_v(x_u) \in A = \{a_1, \dots, a_r\}$ for every word $v \ne u$ of the same length.*

Proof. Induction on u, building the probe backwards. For $u = \emptyset$ take $x = q$. For $u = (i, u')$ with probe x' of u', pick $x \in V_i$ with $f_i(x) = x'$ (coding core). Then $f_u(x) = f_{u'}(x') = q$. If $v = (j, v') \neq u$: when $j = i$, $v' \neq u'$ and $f_v(x) = f_{v'}(x') \in A$ by induction; when $j \neq i$, $x \in U_i$ lies outside U_j (chambers are disjoint), so $f_j(x) = a_j \in A$ (reset) and $f_{v'}(a_j) \in A$ (invariance of A). □

Theorem 259. ***(E2: ping–pong coding.)*** *Let $F = \{f_1, \dots, f_r\}$, $r \geq 2$, and suppose there are sets $V_i \subseteq U_i \subseteq \mathcal{X}$, anchors $a_1, \dots, a_r \in \mathcal{X}$, a marker $q \in \mathcal{X}$ and a constant $\alpha > 0$ such that, with $A = \{a_1, \dots, a_r\}$ and $Q = \{q\} \cup \bigcup_j V_j$:*

1. *(disjoint chambers) $U_i \cap U_j = \emptyset$ for $i \neq j$;*
2. *(coding cores) $Q \subseteq f_i(V_i)$ for every i;*
3. *(reset) $f_i(x) = a_i$ for $x \notin U_i$, and $f_i(A) \subseteq A$;*
4. *(marker separation) $d(q, a_i) \geq \alpha$ for every i.*

Then (a) $N^{\text{ext}}(B(k, F), d_\infty, \varepsilon) \geq r^k$ for every k and every $\varepsilon < \alpha/2$, and (b) $u \mapsto f_u$ is injective on $[r]^k$ for every k. (This generalizes the paper, which assumes in (1) a uniform separation $d(U_i, U_j) \geq \Delta > 0$ and additionally $V_i \neq \emptyset$: the proof uses of (1) only the disjointness of the chambers, and $V_i \neq \emptyset$ follows from (2) since $q \in f_i(V_i)$; the hypothesis $r \geq 2$ is not used.)

Proof. Choose the probes x_u of 'lem:e2-probe'. For (a), use the r^k probes $P = (x_v)_{v \in [r]^k}$: for $u \neq w$ in $[r]^k$ the evaluations differ at the coordinate w, where $f_u(x_w) \in A$ and $f_w(x_w) = q$, so they are at distance $\geq \alpha$; conclude by 'lem:pow-le-covering-of-probes'. For (b), if $f_u = f_v$ with $u \neq v$ of the same length then $q = f_u(x_u) = f_v(x_u) \in A$, contradicting $\alpha > 0$. □

4.7 Memory-preserving expansion: super- and double-exponential growth (Appendix F.4, E3)

Theorem 260. *$w_u \in B(2k + 1, F)$ for every $u \in G^k$.*

Proof. Induction on k: $r \in F \subseteq B(1, F)$, and $A, g_{u_k} \in F$ add two letters. □

Theorem 261. *$W_k \subseteq B(2k + 1, F)$.*

Proof. □

Theorem 262. ***(E3 (i).)*** *For every $x \in X$, $w_u(x) = (\lambda u_{k-1}, \lambda^2 u_{k-2}, \dots, \lambda^k u_0, 0, 0, \dots)$: coordinate $j < k$ of $w_u(x)$ is $\lambda^{j+1} u_{k-1-j}$ and the coordinates $j \geq k$ vanish. In particular w_u is a constant map.*

Proof. Induction on k. For $k + 1$: coordinate 0 of $A(g_a(y))$ is λa with $a = u_k$, and coordinate $j + 1$ is λy_j where $y = w_{u|_k}(x)$; apply the induction hypothesis to y. □

Theorem 263. *w_u is constant: $w_u(x) = w_u(y)$ for all x, y.*

Proof. □

Theorem 264. *For $\lambda \neq 0$ the map $u \mapsto w_u$ is injective on E^k.*

Proof. Coordinate $k - 1 - i$ of $w_u(0)$ is $\lambda^{k-i} u_i$, and $\lambda^{k-i} \neq 0$. □

Theorem 265. *If $f : \mathbb{N} \to [0, \infty]$ vanishes on $j \geq k$ then $\sup_{j \in \mathbb{N}} f(j) = \max_{j<k} f(j)$.*

Proof. □

Theorem 266. ***(E3 (ii).)*** *For $u, v \in E^k$, $d_\infty(w_u, w_v) = \max_{0 \leq j < k} \min\{1, \lambda^{j+1} \|u_{k-1-j} - v_{k-1-j}\|_E\}$ (the paper's $\max_{1 \leq j \leq k} \min\{1, \lambda^{k-j+1} \|u_j - v_j\|\}$ after reindexing).*

Proof. Both maps are constant, so $d_\infty(w_u, w_v) = d_X(w_u(0), w_v(0))$; the coordinates $j \geq k$ of both vanish and coordinate $j < k$ contributes $\min\{1, \|\lambda^{j+1}(u_{k-1-j} - v_{k-1-j})\|\}$. □

Theorem 267. *The same identity for the extended distance of $(\mathcal{X}^{\mathcal{X}}, d_\infty)$.*

Proof. □

Theorem 268. ***(E3 (iii), upper bound.)*** *For every $\varepsilon \geq 0$ (and $\lambda \neq 0$),*

$$N^{\mathrm{ext}}(W_k, d_\infty, \varepsilon) \leq \prod_{j=0}^{k-1} N^{\mathrm{ext}}(G, \|\cdot\|_E, \varepsilon/\lambda^{j+1}).$$

Proof. Take minimal $\varepsilon/\lambda^{j+1}$-covers C_j of G. For $u \in G^k$ choose $c_i \in C_{k-1-i}$ with $\|u_i - c_i\| \leq \varepsilon/\lambda^{k-i}$; then by (ii) $d_\infty(w_u, w_c) \leq \max_j \lambda^{j+1} \|u_{k-1-j} - c_{k-1-j}\| \leq \varepsilon$. So $w(\prod_i C_{k-1-i})$ is an ε-cover of W_k of cardinality $\leq \prod_j |C_j|$. □

Theorem 269. *(Product packing.) Let $2\varepsilon < 1$, $\lambda \neq 0$, and let $P_j \subseteq G$ be $2\varepsilon/\lambda^{j+1}$-separated $(0 \leq j < k)$. Then $\prod_j |P_j| \leq M(W_k, d_\infty, 2\varepsilon)$.*

Proof. The set $w(\prod_i P_{k-1-i}) \subseteq W_k$ is 2ε-separated: if $u \neq v$ then $u_i \neq v_i$ for some i, and with $j = k-1-i$ the j-th term of (ii) is $\min\{1, \lambda^{j+1}\|u_i - v_i\|\} > \min\{1, 2\varepsilon\} = 2\varepsilon$. Since w is injective its cardinality is $\prod_i |P_{k-1-i}| = \prod_j |P_j|$. □

Theorem 270. *In $\mathbb{N} \cup \{\infty\}$, a finite product of suprema over nonempty index types is the supremum of the products: $\prod_{i \in s} \sup_{a \in A_i} g_i(a) = \sup_{x \in \prod_i A_i} \prod_{i \in s} g_i(x_i)$.*

Proof. Induction on s, distributing multiplication over suprema (‘ENat.mul_iSup‘, ‘ENat.iSup_mul‘) and identifying pairs (a, x) with the update $x[i := a]$. □

Theorem 271. ***(E3 (iii), lower bound.)*** *For every $0 \leq \varepsilon < 1/2$ (and $\lambda \neq 0$),*

$$\prod_{j=0}^{k-1} M(G, \|\cdot\|_E, 2\varepsilon/\lambda^{j+1}) \leq N^{\mathrm{ext}}(W_k, d_\infty, \varepsilon).$$

Proof. Each packing number is a supremum over separated subsets of G; by ‘lem:enat-prod-isup‘ the product is the supremum over families $(P_j)_j$ of $\prod_j |P_j|$, which is bounded by $M(W_k, 2\varepsilon) \leq N^{\mathrm{ext}}(W_k, \varepsilon)$ by ‘lem:e3-prod-packing-family‘ and ‘lem:packing-covering‘. □

Theorem 272. *$N^{\mathrm{ext}}(W_k, d_\infty, \varepsilon) \leq N^{\mathrm{ext}}(B(2k+1, F), d_\infty, \varepsilon)$ (monotonicity of the external covering number under $W_k \subseteq B(2k+1, F)$).*

Proof. □

Theorem 273. ***(E3: memory-preserving expansion grows super- or double-exponentially.)*** *Let E be a normed space, $G \subseteq E$, $\lambda > 1$, $X = E^{\mathbb{N}}$ with the bounded sup metric, $F = \{r, A\} \cup \{g_u : u \in G\}$ and $W_k = \{w_u : u \in G^k\} \subseteq B(2k+1, F)$. Then (i) each w_u is the constant map with value $(\lambda u_{k-1}, \dots, \lambda^k u_0, 0, \dots)$ ('cond:e3-i'), (ii) $d_\infty(w_u, w_v) = \max_{j<k} \min\{1, \lambda^{j+1}\|u_{k-1-j} - v_{k-1-j}\|\}$ ('cond:e3-ii'), and (iii) for every $0 < \varepsilon < 1/2$,*

$$\prod_{j<k} M\Big(G, \tfrac{2\varepsilon}{\lambda^{j+1}}\Big) \le N^{\mathrm{ext}}(W_k, d_\infty, \varepsilon) \le \prod_{j<k} N^{\mathrm{ext}}\Big(G, \tfrac{\varepsilon}{\lambda^{j+1}}\Big).$$

Consequently $N^{\mathrm{ext}}(B(2k+1, F), d_\infty, \varepsilon) \ge N^{\mathrm{ext}}(W_k, d_\infty, \varepsilon)$.

Proof. □

Theorem 274. *For finite nonzero $n_i \in \mathbb{N} \cup \{\infty\}$, $\log \prod_i n_i = \sum_i \log n_i$ (with the coercion to $[0, \infty]$ and 'toReal').*

Proof. □

Theorem 275. *If $G \neq \emptyset$ and all $N^{\mathrm{ext}}(G, \varepsilon/\lambda^{j+1})$ are finite, then $\log N^{\mathrm{ext}}(W_k, \varepsilon) \le \sum_{j<k} \log N^{\mathrm{ext}}(G, \varepsilon/\lambda^{j+1})$.*

Proof. Take logarithms in 'cond:e3-iii-upper'; all quantities are finite and positive since G (hence W_k) is nonempty. □

Theorem 276. *If $G \neq \emptyset$, $\varepsilon < 1/2$, all $M(G, 2\varepsilon/\lambda^{j+1})$ and $N^{\mathrm{ext}}(W_k, \varepsilon)$ are finite, then $\sum_{j<k} \log M(G, 2\varepsilon/\lambda^{j+1}) \le \log N^{\mathrm{ext}}(W_k, \varepsilon)$.*

Proof. Take logarithms in 'cond:e3-iii-lower'. □

Theorem 277. $\sum_{j<k} \log(\lambda^{j+1}/a) = \frac{k(k+1)}{2} \log \lambda - k \log a$ *for $a, \lambda > 0$.*

Proof. Induction on k using $\log(\lambda^{j+1}/a) = (j+1)\log \lambda - \log a$. □

Theorem 278. *If $0 \le a < \lambda \delta_0$ and $\lambda > 1$ then $a/\lambda^{j+1} < \delta_0$ for every j.*

Proof. □

Theorem 279. *If $c_- \log(1/\delta) \le \log M(G, \delta)$ for all small $\delta > 0$ with $c_- > 0$, then $G \neq \emptyset$.*

Proof. Otherwise $M(G, \delta) = 0$ and $\log 0 = 0 < c_- \log(1/\delta)$ for $\delta = \min\{\delta_0/2, 1/2\}$. □

Theorem 280. *(Upper bound in sum form.) Suppose $N^{\mathrm{ext}}(G, \delta) < \infty$ for all $\delta > 0$, $G \neq \emptyset$, and $\log N^{\mathrm{ext}}(G, \delta) \le \varphi(\delta)$ for $0 < \delta < \delta_0$. If $0 < \varepsilon < \lambda\delta_0$ then $\log N^{\mathrm{ext}}(W_k, \varepsilon) \le \sum_{j<k} \varphi(\varepsilon/\lambda^{j+1})$ for every k.*

Proof. All radii $\varepsilon/\lambda^{j+1} \le \varepsilon/\lambda < \delta_0$, so 'lem:e3-log-covering-le-sum' and the hypothesis apply termwise. □

Theorem 281. *If $N^{\mathrm{ext}}(G, \delta) < \infty$ for all $\delta > 0$ then $N^{\mathrm{ext}}(W_k, \varepsilon) < \infty$ for all $\varepsilon > 0$.*

Proof. □

Theorem 282. *(Lower bound in sum form.) Suppose $N^{\mathrm{ext}}(G, \delta) < \infty$ for all $\delta > 0$, $G \neq \emptyset$, and $\varphi(\delta) \le \log M(G, \delta)$ for $0 < \delta < \delta_0$. If $0 < \varepsilon < 1/2$ and $2\varepsilon < \lambda\delta_0$ then $\sum_{j<k} \varphi(2\varepsilon/\lambda^{j+1}) \le \log N^{\mathrm{ext}}(W_k, \varepsilon)$ for every k.*

Proof. All radii $2\varepsilon/\lambda^{j+1} \le 2\varepsilon/\lambda < \delta_0$; the packing numbers are finite since $M(G, 2\delta) \le N^{\text{ext}}(G, \delta)$; apply 'lem:e3-sum-log-packing-le' and the hypothesis termwise. $\square$

Theorem 283. ***(Super-exponential regime.)*** *Assume there are $c_-, c_+, \delta_0 > 0$ with $c_- \log(1/\delta) \le \log M(G, \delta)$ and $\log N^{\text{ext}}(G, \delta) \le c_+ \log(1/\delta)$ for all $0 < \delta < \delta_0$, and that $N^{\text{ext}}(G, \delta) < \infty$ for all $\delta > 0$ (automatic for compact G). Then for every fixed ε with $0 < \varepsilon < 1/2$ and $\varepsilon < \lambda\delta_0/2$ there are $C_1, C_2 > 0$ and k_0 such that $C_1 k^2 \le \log N^{\text{ext}}(W_k, d_\infty, \varepsilon) \le C_2 k^2$ for all $k \ge k_0$. (The smallness condition $\varepsilon < \lambda\delta_0/2$ makes every radius $2\varepsilon/\lambda^{j+1} \le 2\varepsilon/\lambda$ fall below δ_0; it cannot be replaced by "large k" since the $j = 0$ radius does not shrink with k.)*

Proof. By the sum forms of (iii) and the entropy hypotheses,

$$c_- \sum_{j<k} \log \frac{\lambda^{j+1}}{2\varepsilon} \le \log N^{\text{ext}}(W_k, \varepsilon) \le c_+ \sum_{j<k} \log \frac{\lambda^{j+1}}{\varepsilon},$$

and $\sum_{j<k} \log(\lambda^{j+1}/a) = \frac{k(k+1)}{2} \log \lambda + k \log(1/a)$. For $k \ge 1$ and $a \le 1$ this is between $\frac{\log \lambda}{2} k^2$ and $(\log \lambda + \log(1/a))k^2$; take $C_1 = c_- \log \lambda/2$, $C_2 = c_+(\log \lambda - \log \varepsilon)$, $k_0 = 1$. $\square$

Theorem 284. *$\sum_{j<k} (a/\lambda^{j+1})^{-p} = a^{-p} \sum_{j<k} (\lambda^p)^{j+1}$ for $a, \lambda > 0$.*

Proof. Termwise: $(a/\lambda^{j+1})^{-p} = (\lambda^{j+1})^p/a^p = a^{-p}(\lambda^p)^{j+1}$. $\square$

Theorem 285. *For $q > 1$ and $k \ge 1$: $q^k \le \sum_{j<k} q^{j+1} \le \frac{q}{q-1} q^k$.*

Proof. The lower bound is the last term; the upper bound is the geometric sum formula $q(q^k - 1)/(q - 1)$. $\square$

Theorem 286. *If $c_- \delta^{-p} \le \log M(G, \delta)$ for all small $\delta > 0$ with $c_- > 0$, then $G \ne \emptyset$.*

Proof. Otherwise $M(G, \delta) = 0$ and $\log 0 = 0 < c_- \delta^{-p}$ for $\delta = \delta_0/2$. $\square$

Theorem 287. ***(Double-exponential regime.)*** *Assume there are $p, c_-, c_+, \delta_0 > 0$ with $c_- \delta^{-p} \le \log M(G, \delta)$ and $\log N^{\text{ext}}(G, \delta) \le c_+ \delta^{-p}$ for all $0 < \delta < \delta_0$, and that $N^{\text{ext}}(G, \delta) < \infty$ for all $\delta > 0$. Then for every fixed ε with $0 < \varepsilon < 1/2$ and $\varepsilon < \lambda\delta_0/2$ there are $C_1, C_2 > 0$ and k_0 such that $C_1 \lambda^{pk} \le \log N^{\text{ext}}(W_k, d_\infty, \varepsilon) \le C_2 \lambda^{pk}$ for all $k \ge k_0$.*

Proof. By the sum forms of (iii), with $q = \lambda^p > 1$,

$$c_-(2\varepsilon)^{-p} \sum_{j<k} q^{j+1} \le \log N^{\text{ext}}(W_k, \varepsilon) \le c_+ \varepsilon^{-p} \sum_{j<k} q^{j+1},$$

and $q^k \le \sum_{j<k} q^{j+1} \le \frac{q}{q-1} q^k$ for $k \ge 1$; take $C_1 = c_-(2\varepsilon)^{-p}$, $C_2 = c_+ \varepsilon^{-p} q/(q - 1)$, $k_0 = 1$. $\square$

4.8 The layerwise covering envelope and the reachable radius (Appendix G, prop:envelope, cor:envelope-profiles, lem:reachable-radius)

Theorem 288. *$S_{m+1}(\Lambda) = 1 + \Lambda\, S_m(\Lambda)$.*

Proof. $\square$

Theorem 289. *$m \mapsto S_m(\Lambda)$ is nondecreasing.*

Proof. □

Theorem 290. *$S_m(\Lambda) \geq 1$ for $m \geq 1$.*

Proof. □

Theorem 291. *$S_m(\Lambda) > 0$ for $m \geq 1$.*

Proof. □

Theorem 292. *With $\Lambda_+ := \max\{1, \Lambda\}$, $S_k(\Lambda) \leq k\,\Lambda_+^k$ (since $\Lambda^i \leq \Lambda_+^k$ for $i < k$).*

Proof. □

Theorem 293. *If all layers of u lie in F then $f_u \in F^{|u|}$.*

Proof. □

Theorem 294. *Every $g \in F^m$ is f_u for a list u of elements of F with $|u| = m$.*

Proof. □

Theorem 295. *$B(k, \emptyset) = \{\mathrm{id}\}$.*

Proof. □

Theorem 296. *If $N < \infty$ then $1 + kN < \infty$ (in $\mathbb{N} \cup \{\infty\}$).*

Proof. □

Theorem 297. *For natural numbers k, N: $\log(1 + kN) \leq \log(k + 1) + \log N$ (with $\log 0 = 0$): $1 + kN \leq (k + 1)N$ when $N \geq 1$, and $\log 1 = 0$ when $N = 0$.*

Proof. □

Theorem 298. *If $A \leq 1 + kN$ in $\mathbb{N} \cup \{\infty\}$ with $N < \infty$, then $\log A \leq \log(k + 1) + \log N$ (as real numbers, with $\log \infty = \log 0 = 0$).*

Proof. □

Theorem 299. *(Uniform form of the telescoping estimate 'lem:impl-word-error'.) If every $f \in F$ is Λ-Lipschitz and $d_\infty(f, \tilde{f}) \leq \delta$ for $f \in F$, then $d_\infty(f_u, \tilde{f}_u) \leq \delta \sum_{i<|u|} \Lambda^i$ for every list u of layers in F.*

Proof. □

Theorem 300. *(Covering the words of length m.) If every $f \in F$ is Λ-Lipschitz then for every $\varepsilon \geq 0$ and $m \geq 0$,*

$$N^{\mathrm{ext}}(F^m, d_\infty, \varepsilon) \leq N^{\mathrm{ext}}(F, d_\infty, \varepsilon/S_m(\Lambda))^m.$$

The M^m compositions $c_{i_m} \circ \cdots \circ c_{i_1}$ of the centres of an $\varepsilon/S_m(\Lambda)$-cover $\{c_1, \ldots, c_M\}$ of F form an ε-cover of F^m by the telescoping estimate.

Proof. Take a minimal ε/S_m-cover C of F and choose for each $f \in F$ a centre $\tilde{f} \in C$ with $d_\infty(f, \tilde{f}) \leq \varepsilon/S_m$. For a word $f_u \in F^m$, $\tilde{f}_u \in C^m$ and $d_\infty(f_u, \tilde{f}_u) \leq (\varepsilon/S_m) \sum_{i<m} \Lambda^i = \varepsilon$; hence C^m is an ε-cover of F^m with $|C^m| \leq |C|^m$. □

Theorem 301. *(First inequality of 'prop:envelope'.) If every $f \in F$ is Λ-Lipschitz then for every $k \geq 0$ and $\varepsilon \geq 0$,*

$$N^{\text{ext}}(B(k,F), d_\infty, \varepsilon) \leq 1 + \sum_{m=1}^{k} N^{\text{ext}}(F, d_\infty, \varepsilon/S_m(\Lambda))^m.$$

Proof. $B(k,F) = \{\text{id}\} \cup \bigcup_{m=1}^{k} F^m$; subadditivity of the external covering number under unions and 'lem:envelope-words'. Induction on k with $B(k+1,F) = B(k,F) \cup F^{k+1}$. ☐

Theorem 302. *(Second inequality of 'prop:envelope'.) For every $k \geq 0$ and $\varepsilon \geq 0$,*

$$\sum_{m=1}^{k} N^{\text{ext}}(F, \varepsilon/S_m(\Lambda))^m \leq k\, N^{\text{ext}}(F, \varepsilon/S_k(\Lambda))^k,$$

because $S_m(\Lambda) \leq S_k(\Lambda)$ for $m \leq k$, covering numbers are nonincreasing in the radius, and $N^m \leq N^k$ for $N \geq 1$ (or $N = 0$, $m, k \geq 1$).

Proof. ☐

Theorem 303. *(Simplified envelope.) If every $f \in F$ is Λ-Lipschitz then for every $k \geq 0$ and $\varepsilon \geq 0$,*

$$N^{\text{ext}}(B(k,F), d_\infty, \varepsilon) \leq 1 + k\, N^{\text{ext}}(F, d_\infty, \varepsilon/S_k(\Lambda))^k.$$

Proof. ☐

Theorem 304. ***(Covering envelope, 'prop:envelope' restated.)*** *Let $F \subseteq \mathcal{X}^{\mathcal{X}}$ consist of Λ-Lipschitz maps, $\Lambda \geq 0$, and $S_m(\Lambda) = \sum_{i=0}^{m-1} \Lambda^i$. For every $k \geq 0$ and $\varepsilon \geq 0$,*

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

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. (In Lean $\varepsilon/S_0 = 0$ and the statement holds for $k = 0$ as well.)

Proof. ☐

Theorem 305. *If $N^{\text{ext}}(F, \varepsilon/S_k(\Lambda)) < \infty$ then $N^{\text{ext}}(B(k,F), \varepsilon) < \infty$.*

Proof. ☐

Theorem 306. *(Logarithmic form of the envelope.) If every $f \in F$ is Λ-Lipschitz, $k \geq 1$ and $N := N^{\text{ext}}(F, \varepsilon/S_k(\Lambda)) < \infty$, then $\log N^{\text{ext}}(B(k,F), \varepsilon) \leq \log(k+1) + k \log N$ (using $1 + kN^k \leq (k+1)N^k$ for $N \geq 1$; for $N = 0$ both sides are $\geq 0 = \log 1$).*

Proof. ☐

Theorem 307. *(**Parametric layers**, entropy form of 'cor:envelope-profiles'(a).) Let every $f \in F$ be Λ-Lipschitz and suppose $N^{\text{ext}}(F, d_\infty, \varepsilon) < \infty$ and $\log N^{\text{ext}}(F, d_\infty, \varepsilon) \leq p \log(C/\varepsilon)$ for all $0 < \varepsilon \leq C$, with $p \geq 0$. Then for $k \geq 1$ and $0 < \varepsilon \leq C$,*

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

Proof. Apply the logarithmic envelope at the radius $\varepsilon/S_k(\Lambda) \le \varepsilon \le C$ and the hypothesis on F there; $\log(C/(\varepsilon/S_k)) = \log(C/\varepsilon) + \log S_k$. □

Theorem 308. ***(Nonparametric layers**, entropy form of 'cor:envelope-profiles'(b).)* *Let every $f \in F$ be Λ-Lipschitz and suppose $N^{\mathrm{ext}}(F, d_\infty, \varepsilon) < \infty$ and $\log N^{\mathrm{ext}}(F, d_\infty, \varepsilon) \le c\,\varepsilon^{-q}$ for all $\varepsilon > 0$, with $c \ge 0$. Then for $k \ge 1$ and $\varepsilon > 0$,*

$$\log N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon) \le \log(k + 1) + kc\, S_k(\Lambda)^q\, \varepsilon^{-q}.$$

Proof. Apply the logarithmic envelope at the radius $\varepsilon/S_k(\Lambda)$ and the hypothesis on F there; $(\varepsilon/S_k)^{-q} = S_k^q \varepsilon^{-q}$. □

Theorem 309. *(Transfer to the empirical metric.) If every $f \in F$ is Λ-Lipschitz and $N^{\mathrm{ext}}(F, d_\infty, \varepsilon/(2S_k(\Lambda))) < \infty$, then $\log N(B(k, F), d_S, \varepsilon) \le \log N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon/2)$ for every sample S ('lem:covering-empSpace-le-external-unifMaps').*

Proof. □

Theorem 310. ***(Profiles from the envelope: parametric layers, 'cor:envelope-profiles'(a).)*** *Let every $f \in F$ be Λ-Lipschitz, let $\overline{D} > 0$, $C \ge \overline{D}$, $p \ge 0$, and suppose $N^{\mathrm{ext}}(F, d_\infty, \varepsilon) < \infty$ and $\log N^{\mathrm{ext}}(F, d_\infty, \varepsilon) \le p \log(C/\varepsilon)$ for all $0 < \varepsilon \le C$. If $D_k(S) \le \overline{D}$ then, for $k \ge 1$, with $\Lambda_+ := \max\{1, \Lambda\}$,*

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

(The paper has C in place of $2C$: the factor 2 is the price of comparing the internal covering number in d_S defining $\mathsf{V}_k(S)$ with the external one in d_∞, 'lem:covering-empSpace-le-external-unifMaps'.)

Proof. Pointwise, for $0 < \varepsilon \le \overline{D}$: $\log N(B(k, F), d_S, \varepsilon) \le \log N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon/2) \le \log(k + 1) + kp[\log(2C/\varepsilon) + \log S_k]$, with $\log(2C/\varepsilon) = \log(2C/\overline{D}) + \log(\overline{D}/\varepsilon)$ and $\log S_k \le \log k + k\log\Lambda_+$; take square roots termwise and integrate, using $\int_0^{\overline{D}} \sqrt{\log(\overline{D}/\varepsilon)}\, d\varepsilon = \frac{\sqrt{\pi}}{2}\overline{D}$. □

Theorem 311. ***(Profiles from the envelope: nonparametric layers, 'cor:envelope-profiles'(b).)*** *Let every $f \in F$ be Λ-Lipschitz, let $\overline{D} > 0$, $c \ge 0$, $0 < q < 2$, and suppose $N^{\mathrm{ext}}(F, d_\infty, \varepsilon) < \infty$ and $\log N^{\mathrm{ext}}(F, d_\infty, \varepsilon) \le c\,\varepsilon^{-q}$ for all $\varepsilon > 0$. If $D_k(S) \le \overline{D}$ then, for $k \ge 1$,*

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

(The paper has $S_k(\Lambda)^{q/2}$ in place of $(2S_k(\Lambda))^{q/2}$: the factor 2 is the price of comparing the internal covering number in d_S with the external one in d_∞.)

Proof. Pointwise, $\log N(B(k, F), d_S, \varepsilon) \le \log N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon/2) \le \log(k + 1) + kc\,(2S_k)^q\varepsilon^{-q}$, so $\sqrt{\log N} \le \sqrt{\log(k+1)} + \sqrt{kc}(2S_k)^{q/2}\varepsilon^{-q/2}$; integrate, using $\int_0^{\overline{D}} \varepsilon^{-q/2}\, d\varepsilon = \overline{D}^{\,1-q/2}/(1 - q/2)$ for $q < 2$. □

Theorem 312. ***(Profiles from the envelope, 'cor:envelope-profiles'.)*** *Let every $f \in F$ be Λ-Lipschitz, $\overline{D} > 0$, $k \ge 1$ and $D_k(S) \le \overline{D}$.*

(a) *(Parametric layers.) If* $N^{\mathrm{ext}}(F, d_\infty, \varepsilon) < \infty$ *and* $\log N^{\mathrm{ext}}(F, d_\infty, \varepsilon) \le p \log(C/\varepsilon)$ *for all* $0 < \varepsilon \le C$*, with* $p \ge 0$ *and* $C \ge \overline{D}$*, then for* $0 < \varepsilon \le \overline{D}$*,* $\log N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon) \le \log(k+1) + kp[\log(C/\varepsilon) + \log S_k(\Lambda)]$*, and* $\mathsf{V}_k(S) \le \overline{D}\big(\sqrt{\log(k+1)} + \sqrt{kp \log k} + k\sqrt{p \log \Lambda_+} + \sqrt{kp}\,\big(\sqrt{\log(2C/\overline{D})} + \sqrt{\pi}/2\big)\big)$.

(b) *(Nonparametric layers.) If* $N^{\mathrm{ext}}(F, d_\infty, \varepsilon) < \infty$ *and* $\log N^{\mathrm{ext}}(F, d_\infty, \varepsilon) \le c\,\varepsilon^{-q}$ *for all* $\varepsilon > 0$*, with* $c \ge 0$ *and* $0 < q < 2$*, then for* $\varepsilon > 0$*,* $\log N^{\mathrm{ext}}(B(k, F), d_\infty, \varepsilon) \le \log(k+1) + kc\, S_k(\Lambda)^q \varepsilon^{-q}$*, and* $\mathsf{V}_k(S) \le \overline{D}\sqrt{\log(k+1)} + \sqrt{kc}\,(2S_k(\Lambda))^{q/2}\,\overline{D}^{1-q/2}/(1 - q/2)$.

See 'cor:envelope-profiles-a' and 'cor:envelope-profiles-b' for the factors 2*.*

Proof. □

Theorem 313. *(Reachable radius, unified form.) Let every* $f \in F$ *be* Λ*-Lipschitz,* $x_0 \in \mathcal{X}$*,* $c \ge 0$ *with* $d(f(x_0), x_0) \le c$ *for all* $f \in F$*, and* $R \ge 0$*. Then every* $g \in B(k, F)$ *maps* $B(x_0, R)$ *into* $B(x_0, \Lambda_+^k R + c\, S_k(\Lambda))$*,* $\Lambda_+ := \max\{1, \Lambda\}$*: for* $d(x, x_0) \le \rho$*,* $d(f(x), x_0) \le d(f(x), f(x_0)) + d(f(x_0), x_0) \le \Lambda\rho + c$*; iterate.*

Proof. □

Theorem 314. *(Reachable radius,* $\Lambda \ge 1$*.) Under the hypotheses of 'lem:reachable-radius-core', if* $\Lambda \ge 1$ *then every* $g \in B(k, F)$ *maps* $B(x_0, R)$ *into* $B(x_0, \Lambda^k R + c\, S_k(\Lambda))$*.*

Proof. □

Theorem 315. *(Reachable radius,* $\Lambda \le 1$*.) Under the hypotheses of 'lem:reachable-radius-core', if* $\Lambda \le 1$ *then every* $g \in B(k, F)$ *maps* $B(x_0, R)$ *into* $B(x_0, R + c\, S_k(\Lambda))$*.*

Proof. □

Theorem 316. *If* $d(f(x_i), g(x_i)) \le D$ *for every sample point, with* $D \ge 0$*, then* $d_S(f, g) \le D$*.*

Proof. □

Theorem 317. *(Diameter of the reachable states.) Under the hypotheses of 'lem:reachable-radius-core', if the sample lies in* $B(x_0, R)$ *then* $D_k(S) \le 2(\Lambda_+^k R + c\, S_k(\Lambda))$*: bounded in* k *if* $\Lambda < 1$*, at most linear if* $\Lambda = 1$*, at most exponential if* $\Lambda > 1$*.*

Proof. $d(f(x_i), g(x_i)) \le d(f(x_i), x_0) + d(x_0, g(x_i)) \le 2\rho$ for $f, g \in B(k, F)$; take the supremum. □

Theorem 318. ***(Reachable radius.)*** *Let every* $f \in F$ *be* Λ*-Lipschitz,* $x_0 \in \mathcal{X}$*,* $R \ge 0$*, and suppose* $d(f(x_0), x_0) \le c$ *for all* $f \in F$ *(with* $c \ge 0$*). Then every* $g \in B(k, F)$ *maps the ball* $B(x_0, R)$ *into* $B\big(x_0, \Lambda^k R + c\, S_k(\Lambda)\big)$ *when* $\Lambda \ge 1$*, and into* $B\big(x_0, R + c\, S_k(\Lambda)\big)$ *when* $\Lambda \le 1$*. Consequently, if* $S \subset B(x_0, R)$*, then* $D_k(S) \le 2\big(\Lambda_+^k R + c\, S_k(\Lambda)\big)$ *with* $\Lambda_+ = \max\{1, \Lambda\}$*: bounded in* k *if* $\Lambda < 1$*, at most linear if* $\Lambda = 1$*, and at most exponential if* $\Lambda > 1$*.*

Proof. □

Chapter 5

From growth mechanisms to variance profiles (Appendix H)

5.1 An elementary logarithmic splitting lemma (Appendix H, lem:log-split)

Theorem 319. *For $\overline{D} > 0$, $k \geq 0$ and $0 < \varepsilon \leq \overline{D}$,* $\log(1 + \frac{k}{\varepsilon}) \leq \log(1 + \frac{k}{\overline{D}}) + \log \frac{\overline{D}}{\varepsilon}$*.*

Proof. $(1 + k/\overline{D})(\overline{D}/\varepsilon) = \overline{D}/\varepsilon + k/\varepsilon \geq 1 + k/\varepsilon$ since $\overline{D}/\varepsilon \geq 1$; take logarithms. □

5.2 Variance profiles from growth and diameter (Appendix H, prop:profiles)

Theorem 320. $\log N(A, \varepsilon) \geq 0$ *for every A and ε.*

Proof. □

Theorem 321. *The metric entropy is antitone in the scale: if $\varepsilon \leq \delta$ and $N(A, \varepsilon) < \infty$ then* $\log N(A, \delta) \leq \log N(A, \varepsilon)$*.*

Proof. □

Theorem 322. $\varepsilon \mapsto \sqrt{\log N(A, \varepsilon^+)}$ *is a measurable function on $\mathbb{R}$ (it is the composition of the antitone map $\varepsilon \mapsto N(A, \varepsilon^+)$ with measurable maps).*

Proof. $\varepsilon \mapsto N(A, \varepsilon^+) \in [0, \infty]$ is antitone, hence measurable; compose with $x \mapsto x$.toReal, log and $\sqrt{\cdot}$. □

Theorem 323. $\mathsf{V}(D, A) \geq 0$ *for $D \geq 0$.*

Proof. □

Theorem 324. *If $\sqrt{\log N(A, \varepsilon)} \leq g(\varepsilon)$ on $(0, D]$ for an interval-integrable g on $[0, D]$, then $\varepsilon \mapsto \sqrt{\log N(A, \varepsilon)}$ is interval-integrable on $[0, D]$.*

Proof. The integrand is measurable and nonnegative, and dominated by $|g|$ on $(0, D]$. □

Theorem 325. *For $0 < a \le D$ with $N(A, a) < \infty$, $\varepsilon \mapsto \sqrt{\log N(A, \varepsilon)}$ is interval-integrable on $[a, D]$ (it is antitone and bounded there).*

Proof. Antitone functions on a compact interval are interval-integrable. □

Theorem 326. $\mathsf{V}(D, A) \le \mathsf{V}(D', A)$ *for* $0 \le D \le D'$, *provided the integrand is interval-integrable on* $[0, D']$ *(the integrand is nonnegative).*

Proof. □

Theorem 327. *If* $\sqrt{\log N(A, \varepsilon)} \le g(\varepsilon)$ *on* $(0, D]$ *for an interval-integrable* g *on* $[0, D]$*, then* $\mathsf{V}(D, A) \le \int_0^D g$.

Proof. Monotonicity of the integral; the integrand is integrable by domination. □

Theorem 328. *Combined comparison: if* $0 \le D \le \overline{D}$ *and* $\sqrt{\log N(A, \varepsilon)} \le g(\varepsilon)$ *on* $(0, \overline{D}]$ *for an interval-integrable* g *on* $[0, \overline{D}]$*, then* $\mathsf{V}(D, A) \le \int_0^{\overline{D}} g$.

Proof. $\mathsf{V}(D, A) \le \mathsf{V}(\overline{D}, A) \le \int_0^{\overline{D}} g$. □

Theorem 329. *(Saturation.) Let* A_k *be sets with diameters* $0 \le D_k \le \overline{D}$ *and* $N(A_k, \varepsilon) \le N_\infty(\varepsilon)$ *for all* k *and* ε*, with* $N_\infty(\varepsilon) < \infty$ *for* $\varepsilon > 0$ *and* $\mathsf{V}_\infty := \int_0^{\overline{D}} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ *(the integrand is interval-integrable). Then* $\mathsf{V}(D_k, A_k) \le \mathsf{V}_\infty$ *for all* k.

Proof. Both the integrand and the upper limit are dominated. □

Theorem 330. *If* $N(A, \varepsilon) \le C_0(1 + k/\varepsilon)^D$ *for* $\varepsilon > 0$*, with* $C_0 \ge 1$*,* $D \ge 0$*, then for* $0 < \varepsilon \le \overline{D}$*,* $\sqrt{\log N(A, \varepsilon)} \le \sqrt{\log C_0} + \sqrt{D}\sqrt{\log(1 + k/\overline{D})} + \sqrt{D}\sqrt{\log(\overline{D}/\varepsilon)}$.

Proof. $\log N \le \log C_0 + D \log(1 + k/\varepsilon) \le \log C_0 + D \log(1 + k/\overline{D}) + D \log(\overline{D}/\varepsilon)$ by the log-splitting inequality, then $\sqrt{a + b + c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$. □

Theorem 331. *(Polynomial growth, bounded diameter; single-set form.) If* $0 \le D \le \overline{D}$*,* $\overline{D} > 0$*,* $C_0 \ge 1$*,* $D \ge 0$ *(the exponent) and* $N(A, \varepsilon) \le C_0(1 + k/\varepsilon)^D$ *for all* $\varepsilon > 0$*, then* $\mathsf{V}(D, A) \le \overline{D}\sqrt{D}(\sqrt{\log(1 + k/\overline{D})} + \frac{\sqrt{\pi}}{2}) + \overline{D}\sqrt{\log C_0}$.

Proof. Integrate the pointwise bound over $(0, \overline{D}]$ and use $\int_0^{\overline{D}} \sqrt{\log(\overline{D}/\varepsilon)}\, d\varepsilon = \frac{\sqrt{\pi}}{2}\overline{D}$. □

Theorem 332. *(Polynomial growth, bounded diameter.) If* $0 \le D_k \le \overline{D}$*,* $\overline{D} > 0$*, and* $N(A_k, \varepsilon) \le C_0(1 + k/\varepsilon)^D$ *for all* $k \ge 1$ *and* $\varepsilon > 0$ *(with* $C_0 \ge 1$*,* $D \ge 0$*), then for* $k \ge 1$

$$\mathsf{V}(D_k, A_k) \le \overline{D}\sqrt{D}\Big(\sqrt{\log(1 + k/\overline{D})} + \tfrac{\sqrt{\pi}}{2}\Big) + \overline{D}\sqrt{\log C_0} = O(\sqrt{D \log k}).$$

Proof. □

Theorem 333. *(Polynomial growth, linearly growing diameter.) If* $0 \le D_k \le D_1 k$ *with* $D_1 > 0$*, and* $N(A_k, \varepsilon) \le C_0(1 + k/\varepsilon)^D$ *for all* $k \ge 1$ *and* $\varepsilon > 0$ *(with* $C_0 \ge 1$*,* $D \ge 0$*), then for* $k \ge 1$

$$\mathsf{V}(D_k, A_k) \le D_1 k\Big(\sqrt{D}(\sqrt{\log(1 + 1/D_1)} + \tfrac{\sqrt{\pi}}{2}) + \sqrt{\log C_0}\Big) = O(k\sqrt{D}).$$

Proof. Apply the bounded-diameter estimate with $\overline{D} = D_1 k$ and simplify $k/(D_1 k) = 1/D_1$. $\square$

Theorem 334. *(Exponential growth, bounded diameter.) If $0 \le D_k \le \overline{D}$ and $\log N(A_k, \varepsilon) \le \alpha k + \psi(\varepsilon)$ for all k and $\varepsilon > 0$, with $\Psi := \int_0^{\overline{D}} \sqrt{\psi(\varepsilon)}\, d\varepsilon < \infty$ (i.e. $\sqrt{\psi}$ is interval-integrable on $[0, \overline{D}]$), then $\mathsf{V}(D_k, A_k) \le \overline{D}\sqrt{\alpha k} + \Psi = O(\sqrt{k})$.*

Proof. $\sqrt{\log N_k(\varepsilon)} \le \sqrt{\alpha k} + \sqrt{\psi(\varepsilon)}$; integrate. $\square$

Theorem 335. *(Finite classes.) If $|A_k| \le r^{k+1}$ and $0 \le D_k \le \overline{D}$, then $N(A_k, \varepsilon) \le r^{k+1}$ at every scale and $\mathsf{V}(D_k, A_k) \le \overline{D}\sqrt{(k+1)\log r}$.*

Proof. $N(A_k, \varepsilon) \le |A_k| \le r^{k+1}$, so $\log N(A_k, \varepsilon) \le (k+1)\log r$; integrate the constant bound. $\square$

5.3 The variance term and the table of profiles (Appendix H)

Theorem 336. *Since $d_S \le d_\infty$, the identity map $(\mathcal{X}^{\mathcal{X}}, d_\infty) \to (\mathcal{X}^{\mathcal{X}}, d_S)$ is 1-Lipschitz.*

Proof. This is ‘lem:dS-le-dinf‘. $\square$

Theorem 337. *For every $A \subseteq \mathcal{X}^{\mathcal{X}}$ and $\varepsilon \ge 0$, $N^{\text{ext}}(A, d_S, \varepsilon) \le N^{\text{ext}}(A, d_\infty, \varepsilon)$: covering numbers in the empirical metric are dominated by those in the uniform metric, so all hypotheses may be verified in d_∞.*

Proof. ‘lem:lipschitz-embedding‘ with $K = 1$ and $\varphi = \text{id}$. $\square$

Theorem 338. $N(A, d_S, \varepsilon) \le N(A, d_\infty, \varepsilon)$ *(internal covering numbers).*

Proof. ‘lem:lipschitz-embedding-internal‘ with $K = 1$ and $\varphi = \text{id}$. $\square$

Theorem 339. $M(A, d_S, \varepsilon) \le M(A, d_\infty, \varepsilon)$ *(packing numbers).*

Proof. ‘lem:lipschitz-embedding-packing‘ with $K = 1$ and $\varphi = \text{id}$. $\square$

Theorem 340. *The bridge used by all profiles: for every $A \subseteq \mathcal{X}^{\mathcal{X}}$ and $\varepsilon \ge 0$, $N(A, d_S, \varepsilon) \le N^{\text{ext}}(A, d_\infty, \varepsilon/2)$ (internal covering number on the left, external on the right; the factor 2 is the price of comparing internal with external covers).*

Proof. $N(A, d_S, \varepsilon) = N(A, d_S, 2 \cdot \varepsilon/2) \le N^{\text{ext}}(A, d_S, \varepsilon/2) \le N^{\text{ext}}(A, d_\infty, \varepsilon/2)$. $\square$

Theorem 341. *If $N^{\text{ext}}(A, d_\infty, \varepsilon/2) \le b$ with $b \ge 0$, then $N(A, d_S, \varepsilon) \le b$ as real numbers (the covering number being finite).*

Proof. $\square$

Theorem 342. $\operatorname{diam}_S(A) \ge 0$.

Proof. $\square$

Theorem 343. *(Diameter envelopes transfer from d_∞ to d_S.) If $d_\infty(f, g) \le D$ for all $f, g \in A$, with $D \ge 0$, then $\operatorname{diam}_S(A) \le D$ for every sample S.*

Proof. $d_S(f, g) \le d_\infty(f, g) \le D$ for all $f, g \in A$; take the supremum (which is $0 \le D$ if $A = \emptyset$). $\square$

Theorem 344. *If $d_\infty(f,g) \le \overline{D} < \infty$ for all $f, g \in A$ then* $\mathrm{diam}_S(A) \le \overline{D}$.

Proof. □

Theorem 345. *On a bounded state space, $d(x,y) \le D_{\mathcal{X}}$ for all x, y ($D_{\mathcal{X}} \ge 0$), every class satisfies* $\mathrm{diam}_S(A) \le D_{\mathcal{X}}$*; in particular $D_k(S) \le \mathrm{diam}(\mathcal{X})$ for all k.*

Proof. □

Theorem 346. *On a compact state space, $D_k(S) \le \mathrm{diam}(\mathcal{X})$ for every class A and every sample S.*

Proof. □

Theorem 347. *(Diameter envelope under P2, non-compact case.) Under the hypotheses of 'lem:p2-diameter', $D_k(S) = \mathrm{diam}_S B(k,F) \le 2L_\alpha R_S k$ for every sample S.*

Proof. 'lem:p2-diameter' gives $d_\infty(f,g) \le 2L_\alpha R_S k$ on $B(k,F)$; transfer to d_S. □

Theorem 348. *(Restatement of 'thm:hidden-decomp-depth'.) Under the hypotheses of 'thm:hidden-decomp-depth' (including the integrability of the entropy integrand of $B(k,F_0)$ on $[0, D_k(S)]$), $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \le \hat{\mathfrak{R}}_S(H) + \mathrm{var}(k,n)$.*

Proof. □

Theorem 349. *(Plugging a profile into the decomposition.) Under the hypotheses of 'thm:hidden-decomp-depth' (including the integrability of the entropy integrand), if $\mathsf{V}_k(S) \le B$ then $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \le \hat{\mathfrak{R}}_S(H) + \frac{12A_H L}{\sqrt{n}} B$. The statements about $\mathrm{var}(k,n)$ in 'prop:profiles' follow by multiplying the profile bounds with $12A_H L/\sqrt{n}$.*

Proof. 'lem:hidden-decomp-var' and $12A_H L/\sqrt{n} \ge 0$. □

Theorem 350. ***(P1 ⇒ saturation; general form.)*** *Let the state metric be bounded, $d(x,y) \le D_{\mathcal{X}}$ ($D_{\mathcal{X}} \ge 0$), and let the semigroup $\langle F \rangle$ be totally bounded in d_∞ ('cond:p1'). Put $N_\infty(\varepsilon) := N^{\mathrm{ext}}(\overline{\langle F \rangle}, d_\infty, \varepsilon/2)$ (finite for $\varepsilon > 0$; the factor $\frac{1}{2}$ comes from comparing internal with external covers) and assume $\mathsf{V}_\infty := \int_0^{D_{\mathcal{X}}} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ (interval-integrability of the majorant). Then for every k and every sample S, $\mathsf{V}_k(S) \le \mathsf{V}_\infty$: profile (i), $\mathrm{var}(k,n) = O(n^{-1/2})$.*

Proof. 'prop:profiles-i' with $A_k = B(k,F)$, $D_k = D_k(S) \le D_{\mathcal{X}}$ ('lem:empDiam-le-of-bounded') and the majorant N_∞: $N(B(k,F), d_S, \varepsilon) \le N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon/2) \le N^{\mathrm{ext}}(\overline{\langle F \rangle}, d_\infty, \varepsilon/2)$ ('lem:covering-empSpace-le-external-unifMaps' and monotonicity), finite by 'cond:p1'. □

Theorem 351. ***(P1 ⇒ saturation; table row "P1 (compact, equicontinuous)".)*** *Let $\mathcal{X}$ be compact and the semigroup $\langle F \rangle$ equicontinuous. With $N_\infty(\varepsilon) := N^{\mathrm{ext}}(\overline{\langle F \rangle}, d_\infty, \varepsilon/2)$ and $\mathsf{V}_\infty := \int_0^{\mathrm{diam}(\mathcal{X})} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ (assumed interval-integrable), $D_k(S) \le \mathrm{diam}(\mathcal{X})$ and $\mathsf{V}_k(S) \le \mathsf{V}_\infty$ for all k: profile (i), $\mathsf{V}_k = O(1)$, $\mathrm{var}(k,n) = O(n^{-1/2})$.*

Proof. □

Theorem 352. ***(P1 with non-expanding generators.)*** *Let $\mathcal{X}$ be compact and every $f \in F$ be* 1*-Lipschitz. Then the conclusion of 'cor:profile-p1' holds.*

Proof. □

Theorem 353. *(P2 in the empirical metric.) Under the hypotheses of 'cond:p2-nilp', for every k and $\varepsilon > 0$,*

$$N(B(k,F), d_S, \varepsilon) \leq C_0\Big(1 + \frac{k}{\varepsilon}\Big)^D, \qquad C_0 := 2^D\, C_H \max(1, R_S L_\alpha)^D,$$

as real numbers. (The factor 2^D comes from $N(\cdot, d_S, \varepsilon) \leq N^{\mathrm{ext}}(\cdot, d_\infty, \varepsilon/2)$ and $1 + 2k/\varepsilon \leq 2(1 + k/\varepsilon)$.)

Proof. Apply 'cond:p2-nilp' at scale $\varepsilon/2$ and 'lem:covering-empSpace-toReal-le'. □

Theorem 354. $C_0 = 2^D C_H \max(1, R_S L_\alpha)^D \geq 1$.

Proof. □

Theorem 355. ***(P2 on a bounded state space; table row "P2, compact $\mathcal{X}$".)*** *Under the hypotheses of 'cond:p2-nilp', if moreover $d(x,y) \leq D_{\mathcal{X}}$ for all x, y ($D_{\mathcal{X}} > 0$), then $D_k(S) \leq D_{\mathcal{X}}$ and for every $k \geq 1$*

$$\mathsf{V}_k(S) \leq D_{\mathcal{X}}\sqrt{D}\Big(\sqrt{\log(1 + k/D_{\mathcal{X}})} + \tfrac{\sqrt{\pi}}{2}\Big) + D_{\mathcal{X}}\sqrt{\log C_0} = O(\sqrt{D \log k}),$$

with $C_0 = 2^D C_H \max(1, R_S L_\alpha)^D$: profile (ii), $\mathrm{var}(k,n) = O(\sqrt{D \log k/n})$.

Proof. □

Theorem 356. ***(P2 on a general state space; table row "P2, non-compact $\mathcal{X}$".)*** *Under the hypotheses of 'cond:p2-nilp' with $R_S > 0$, $D_k(S) \leq 2L_\alpha R_S k$ and for every $k \geq 1$*

$$\mathsf{V}_k(S) \leq 2L_\alpha R_S k\Big(\sqrt{D}(\sqrt{\log(1 + 1/(2L_\alpha R_S))} + \tfrac{\sqrt{\pi}}{2}) + \sqrt{\log C_0}\Big) = O(k\sqrt{D}),$$

with $C_0 = 2^D C_H \max(1, R_S L_\alpha)^D$: profile (iv), $\mathrm{var}(k,n) = O(k\sqrt{D/n})$.

Proof. □

Theorem 357. ***(Finite hidden-layer classes; table row "E1', E2 with $|F| = r$".)*** *If F is finite with $|F| = r \geq 2$ and the state metric is bounded, $d(x,y) \leq D_{\mathcal{X}}$ ($D_{\mathcal{X}} \geq 0$), then $|B(k,F)| \leq r^{k+1}$, hence $N(B(k,F), d_S, \varepsilon) \leq r^{k+1}$ for every $\varepsilon > 0$, and*

$$\mathsf{V}_k(S) \leq D_{\mathcal{X}}\sqrt{(k+1)\log r} = O(\sqrt{k \log r})$$

for every k: profile (iii), $\mathrm{var}(k,n) = O(\sqrt{k/n})$.

Proof. □

Theorem 358. *If $0 \leq D \leq \overline{D}$ and $\sqrt{\log N(A,\varepsilon)} \leq g(\varepsilon)$ on $(0, \overline{D}]$ for an interval-integrable g on $[0, \overline{D}]$, then $\varepsilon \mapsto \sqrt{\log N(A,\varepsilon)}$ is interval-integrable on $[0, D]$.*

Proof. □

Theorem 359. ***(Integrability under P1.)*** *Under the hypotheses of 'cor:profile-p1-totallyBounded' (in particular the interval-integrability of the majorant $\sqrt{\log N_\infty}$ on $[0, D_{\mathcal{X}}]$), the entropy integrand $\varepsilon \mapsto \sqrt{\log N(B(k,F), d_S, \varepsilon)}$ is interval-integrable on $[0, D_k(S)]$ for every k: the hypothesis of 'thm:hidden-decomp-depth' is automatic.*

Proof. Domination by the majorant N_∞, as in 'cor:profile-p1-totallyBounded'. $\square$

Theorem 360. ***(Integrability under P2.)*** *Under the hypotheses of 'cond:p2-nilp', if $D_k(S) \le \overline{D}$ with $\overline{D} > 0$, then the entropy integrand of $B(k, F)$ is interval-integrable on $[0, D_k(S)]$ (it is dominated by $\sqrt{\log C_0} + \sqrt{D}\sqrt{\log(1 + k/\overline{D})} + \sqrt{D}\sqrt{\log(\overline{D}/\varepsilon)}$, 'lem:sqrt-metric-entropy-le-of-poly').*

Proof. The polynomial majorant is interval-integrable on $[0, \overline{D}]$ ('intervalIntegrable_sqrt_log_div'). $\square$

Theorem 361. ***(Integrability for finite hidden-layer classes.)*** *If F is finite with $|F| = r \ge 2$ and $d(x, y) \le D_{\mathcal{X}}$, then the entropy integrand of $B(k, F)$ is bounded by the constant $\sqrt{(k+1)\log r}$, hence interval-integrable on $[0, D_k(S)]$.*

Proof. $N(B(k, F), d_S, \varepsilon) \le |B(k, F)| \le r^{k+1}$ ('lem:wordball-card'). $\square$

Theorem 362. ***(Estimation term under P1.)*** *Let $\mathcal{X}$ be compact, the semigroup $\langle F \rangle$ equicontinuous, and V_∞ as in 'cor:profile-p1'. Under the hypotheses of 'thm:hidden-decomp-depth', $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \le \hat{\mathfrak{R}}_S(H) + \frac{12 A_H L}{\sqrt{n}} \mathsf{V}_\infty$ for every k: $\mathrm{var}(k, n) = O(n^{-1/2})$ uniformly in the depth.*

Proof. $\square$

Theorem 363. ***(Estimation term under P2, bounded state space.)*** *Under the hypotheses of 'cor:profile-p2-compact' and 'thm:hidden-decomp-depth', for $k \ge 1$, $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \le \hat{\mathfrak{R}}_S(H) + \frac{12 A_H L}{\sqrt{n}} \Big(D_{\mathcal{X}} \sqrt{D} (\sqrt{\log(1 + k/D_{\mathcal{X}})} + \frac{\sqrt{\pi}}{2}) + D_{\mathcal{X}} \sqrt{\log C_0} \Big)$: $\mathrm{var}(k, n) = O(\sqrt{D \log k / n})$.*

Proof. $\square$

Theorem 364. ***(Estimation term under P2, general state space.)*** *Under the hypotheses of 'cor:profile-p2-noncompact' and 'thm:hidden-decomp-depth', for $k \ge 1$, $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \le \hat{\mathfrak{R}}_S(H) + \frac{12 A_H L}{\sqrt{n}} 2 L_\alpha R_S k \Big(\sqrt{D} (\sqrt{\log(1 + 1/(2 L_\alpha R_S))} + \frac{\sqrt{\pi}}{2}) + \sqrt{\log C_0} \Big)$: $\mathrm{var}(k, n) = O(k\sqrt{D/n})$.*

Proof. $\square$

Theorem 365. ***(Estimation term for finite hidden-layer classes.)*** *Under the hypotheses of 'cor:profile-finite' and 'thm:hidden-decomp-depth', $\hat{\mathfrak{R}}_S(\mathcal{H}_k) \le \hat{\mathfrak{R}}_S(H) + \frac{12 A_H L}{\sqrt{n}} D_{\mathcal{X}} \sqrt{(k+1)\log r}$ for every k: $\mathrm{var}(k, n) = O(\sqrt{k/n})$.*

Proof. $\square$

Chapter 6

Depth bias–variance trade-offs (Appendix I)

6.1 Bias laws, variance profiles, balancing depths (definitions)

Definition 366. Exponentially decaying approximation error: $\mathsf{bias}(k) = e^{-\alpha k}$, $\alpha > 0$.

Definition 367. Polynomially decaying approximation error: $\mathsf{bias}(k) = k^{-\beta}$, $\beta > 0$.

Definition 368. Root-logarithmic estimation profile: $\mathsf{var}(k, n) = \sqrt{\log k/n}$ (**??**(i)–(ii)).

Definition 369. Root-polynomial estimation profile: $\mathsf{var}(k, n) = \sqrt{k^{\gamma}/n}$, $\gamma > 0$ (**??**(iii)–(iv)).

Definition 370. The depth-dependent part of the excess-risk bound of Theorem 98: $\mathsf{gen}(k, n) = \mathsf{bias}(k) + \mathsf{var}(k, n)$.

Definition 371. PP balancing depth $k^* = n^{1/(2\beta+\gamma)}$.

Definition 372. EP balancing depth (leading terms): $k^* = \frac{1}{2\alpha}(\log n - \gamma \log\log n)$.

Definition 373. EP balanced rate $n^{-1/2}(\log n)^{\gamma/2}$.

Definition 374. EL balancing depth (leading terms): $k^* = \frac{1}{2\alpha}(\log n - \log\log\log n)$.

Definition 375. EL balanced rate $\sqrt{\log\log n/n}$.

Definition 376. PL balancing depth $k^* = \left(2\beta n/\log(2\beta n)\right)^{1/(2\beta)}$, the leading term of $\exp(W(2\beta n)/(2\beta))$ (Lambert W, $W(x) \sim \log x$).

Definition 377. PL balanced rate $\sqrt{\log n/n}$.

6.2 Balancing depths and balanced rates (Appendix I, the four regimes EL, EP, PL, PP; lem:balancing, eq:tradeoff-table)

Theorem 378. *For $\alpha > 0$, $k \mapsto e^{-\alpha k}$ is nonincreasing.*

Proof. exp is monotone and $k \mapsto -\alpha k$ is nonincreasing. $\square$

Theorem 379. *For $\beta > 0$, $k \mapsto k^{-\beta}$ is nonincreasing on $(0, \infty)$.*

Proof. Power with a nonpositive exponent is antitone on the positive reals. $\square$

Theorem 380. *For $n \geq 0$, $k \mapsto \sqrt{\log k/n}$ is nondecreasing on $(0, \infty)$.*

Proof. log is monotone on $(0, \infty)$; divide by $n \geq 0$ and take square roots. $\square$

Theorem 381. *For $\gamma \geq 0$ and $n \geq 0$, $k \mapsto \sqrt{k^\gamma/n}$ is nondecreasing on $(0, \infty)$.*

Proof. $k \mapsto k^\gamma$ is monotone for $\gamma \geq 0$; divide by n and take roots. $\square$

Theorem 382. *If $\mathsf{bias} \geq 0$ and $\mathsf{var}(\cdot, n) \geq 0$ on a set S of depths, then $\mathsf{gen}(k, n) \geq \max\{\mathsf{bias}(k), \mathsf{var}(k, n)\}$ for every $k \in S$.*

Proof. Each summand is bounded by the sum because the other one is nonnegative. $\square$

Theorem 383. ***Balancing principle.*** *Let $S \subseteq \mathbb{R}$ be a set of depths on which $\mathsf{bias} \geq 0$ is nonincreasing and $\mathsf{var}(\cdot, n) \geq 0$ is nondecreasing. If $k_0 \in S$ 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 $k \in S$, while $\mathsf{gen}(k_0, n) = 2\mathsf{bias}(k_0)$; hence k_0 minimizes the bound over S up to the factor 2.*

Proof. Case split on $k \leq k_0$ or $k \geq k_0$: in the first case $\mathsf{bias}(k) \geq \mathsf{bias}(k_0)$ and $\mathsf{var}(k, n) \geq 0$; in the second $\mathsf{var}(k, n) \geq \mathsf{var}(k_0, n) = \mathsf{bias}(k_0)$ and $\mathsf{bias}(k) \geq 0$. The identity at k_0 is the balance equation. $\square$

Theorem 384. *For $\beta, \gamma > 0$ and $n > 0$, $\mathsf{bias}(k^*) = {k^*}^{-\beta} = n^{-\beta/(2\beta+\gamma)}$.*

Proof. $(n^{1/(2\beta+\gamma)})^{-\beta} = n^{-\beta/(2\beta+\gamma)}$ by the power rule. $\square$

Theorem 385. *For $\beta, \gamma > 0$ and $n > 0$, $\mathsf{var}(k^*, n) = \sqrt{{k^*}^\gamma/n} = n^{-\beta/(2\beta+\gamma)}$.*

Proof. ${k^*}^\gamma/n = n^{\gamma/(2\beta+\gamma)-1} = n^{-2\beta/(2\beta+\gamma)}$, and the square root halves the exponent. $\square$

Theorem 386. ***PP regime*** *(**??**). Let $\beta, \gamma > 0$ and $n > 0$. At the balancing depth $k^* = n^{1/(2\beta+\gamma)}$, $\mathsf{gen}(k^*, n) = {k^*}^{-\beta} + \sqrt{{k^*}^\gamma/n} = 2\,n^{-\beta/(2\beta+\gamma)}$, and for every depth $k > 0$, $\mathsf{gen}(k, n) = k^{-\beta} + \sqrt{k^\gamma/n} \geq n^{-\beta/(2\beta+\gamma)}$. Thus k^* is optimal up to the factor 2 and $\mathsf{gen}(k^*, n) \asymp n^{-\beta/(2\beta+\gamma)}$.*

Proof. Both terms equal $n^{-\beta/(2\beta+\gamma)}$ at k^* (Theorem 384, Theorem 385), so the balancing principle Theorem 383 on $S = (0, \infty)$ gives both claims. $\square$

Theorem 387. *For every $c > 0$, $\log x \leq c\,x$ for all sufficiently large x.*

Proof. $\log x = o(x)$ (Mathlib: 'Real.isLittleO_log_id_atTop'). $\square$

Theorem 388. *For every $c > 0$, $\log \log n \leq c \log n$ for all sufficiently large n.*

Proof. Compose Theorem 387 with $\log n \to \infty$. $\square$

Theorem 389. $\log x < x$ *for $x > 0$.*

Proof. $\log x \leq x - 1 < x$. $\square$

Theorem 390. *For $n > 0$ and $L \geq 0$, $n^{-1/2} L^{\gamma/2} = \sqrt{L^\gamma/n}$.*

Proof. $\sqrt{L^\gamma/n} = \sqrt{L^\gamma}/\sqrt{n} = L^{\gamma/2}n^{-1/2}$. ☐

Theorem 391. *For $c \geq 0$, $\sqrt{c\,x/n} = \sqrt{c}\,\sqrt{x/n}$.*

Proof. Multiplicativity of the square root. ☐

Theorem 392. *$n^{-1/2}(\log n)^{\gamma/2} \geq 0$ for $n \geq 1$.*

Proof. Product of two nonnegative powers. ☐

Theorem 393. *For $\alpha > 0$ and $n > 1$, $e^{-\alpha k^*} = n^{-1/2}(\log n)^{\gamma/2}$ exactly.*

Proof. $-\alpha k^* = -\frac{1}{2}\log n + \frac{\gamma}{2}\log\log n$; exponentiate. ☐

Theorem 394. *For $\alpha, \gamma > 0$ and $\log n \geq 1$, $k^* \leq \log n/(2\alpha)$.*

Proof. $\gamma \log\log n \geq 0$ when $\log n \geq 1$. ☐

Theorem 395. *For $\alpha, \gamma > 0$, $n > 1$, $\log n \geq 1$ and $\gamma \log\log n \leq \log n$: $\sqrt{{k^*}^\gamma/n} \leq \sqrt{(2\alpha)^{-\gamma}}\, n^{-1/2}(\log n)^{\gamma/2}$.*

Proof. $0 \leq k^* \leq \log n/(2\alpha)$, so ${k^*}^\gamma \leq (2\alpha)^{-\gamma}(\log n)^\gamma$; divide by n and take roots. ☐

Theorem 396. *For $\alpha, \gamma > 0$, eventually in n: $n^{-1/2}(\log n)^{\gamma/2} \leq \mathsf{gen}(k^*, n) \leq (1+\sqrt{(2\alpha)^{-\gamma}})\, n^{-1/2}(\log n)^{\gamma/2}$.*

Proof. The bias term equals the rate exactly (Theorem 393) and the variance term is at most $\sqrt{(2\alpha)^{-\gamma}}$ times the rate (Theorem 395); the side conditions hold for large n by Theorem 388. ☐

Theorem 397. *If $0 \leq g$ eventually, $c > 0$, and $g \leq f \leq c\,g$ eventually, then $f = \Theta(g)$.*

Proof. Both O-directions, with constants c and 1. ☐

Theorem 398. ***EP regime** (??). Let $\alpha, \gamma > 0$ and $k^*(n) = \frac{1}{2\alpha}(\log n - \gamma \log\log n)$. Then $\mathsf{gen}(k^*, n) = e^{-\alpha k^*} + \sqrt{{k^*}^\gamma/n} \asymp n^{-1/2}(\log n)^{\gamma/2}$ as $n \to \infty$ (more precisely, between 1 and $1 + (2\alpha)^{-\gamma/2}$ times the rate).*

Proof. Apply Theorem 397 to the two-sided bound Theorem 396. ☐

Theorem 399. *EP regime, upper bound: $\mathsf{gen}(k^*, n) = O(n^{-1/2}(\log n)^{\gamma/2})$.*

Proof. The O-half of Theorem 398. ☐

Theorem 400. *For $\alpha > 0$ and $n > e$, $e^{-\alpha k^*} = \sqrt{\log\log n/n}$ exactly.*

Proof. $-\alpha k^* = -\frac{1}{2}\log n + \frac{1}{2}\log\log\log n$; exponentiate and rewrite $n^{-1/2}(\log\log n)^{1/2} = \sqrt{\log\log n/n}$. ☐

Theorem 401. *For $\alpha > 0$ and $\log n \geq e$: $0 < k^* \leq \log n/(2\alpha)$ and $\log k^* \leq \log\log n + |\log(2\alpha)|$.*

Proof. $\log\log\log n \geq 0$ gives the upper bound; $\log\log\log n < \log\log n < \log n$ gives positivity; then $\log k^* \leq \log(\log n/(2\alpha)) = \log\log n - \log(2\alpha)$. ☐

Theorem 402. *For $\alpha > 0$ and $\log n \geq e$: $\sqrt{\log k^*/n} \leq \sqrt{1 + |\log(2\alpha)|}\,\sqrt{\log\log n/n}$.*

Proof. $\log k^* \leq \log\log n + |\log 2\alpha| \leq (1 + |\log 2\alpha|)\log\log n$ because $\log\log n \geq 1$ (Theorem 401). ☐

Theorem 403. *For $\alpha > 0$, eventually in n: $\sqrt{\log\log n/n} \leq \mathsf{gen}(k^*, n) \leq (1+\sqrt{1 + |\log(2\alpha)|})\sqrt{\log\log n/n}$.*

Proof. Bias equals the rate (Theorem 400); variance is at most $\sqrt{1 + |\log 2\alpha|}$ times the rate (Theorem 402). ☐

Theorem 404. ***EL regime*** *(??). Let $\alpha > 0$ and $k^*(n) = \frac{1}{2\alpha}(\log n - \log\log\log n)$. Then* $\mathsf{gen}(k^*, n) = e^{-\alpha k^*} + \sqrt{\log k^*/n} \asymp \sqrt{\log\log n/n}$ *as $n \to \infty$.*

Proof. Apply Theorem 397 to Theorem 403. ☐

Theorem 405. *EL regime, upper bound:* $\mathsf{gen}(k^*, n) = O(\sqrt{\log\log n/n})$.

Proof. The O-half of Theorem 404. ☐

Theorem 406. *For $\beta > 0$, $n > 0$ and $L = \log(2\beta n) > 0$: ${k^*}^{-\beta} = \sqrt{L/(2\beta n)}$.*

Proof. ${k^*}^{-\beta} = (2\beta n/L)^{-1/2} = (L/(2\beta n))^{1/2}$. ☐

Theorem 407. *For $\beta > 0$, $n > 0$ and $L = \log(2\beta n) > 0$: $\log k^* = (L - \log L)/(2\beta)$.*

Proof. log of a power and of a quotient. ☐

Theorem 408. *The PL balancing equation $k^{2\beta} \log k \asymp n$ holds exactly up to the factor $1 - \log L/L$: for $\beta > 0$, $n > 0$ and $L = \log(2\beta n) > 0$, ${k^*}^{2\beta} \log k^* = n\,(1 - \log L/L)$.*

Proof. ${k^*}^{2\beta} = 2\beta n/L$ and $\log k^* = (L - \log L)/(2\beta)$ (Theorem 407); multiply out. ☐

Theorem 409. *${k^*}^{2\beta} \log k^*/n \to 1$ as $n \to \infty$, i.e. ${k^*}^{2\beta} \log k^* \sim n$.*

Proof. By Theorem 408 the ratio is $1 - \log L/L$ with $L = \log(2\beta n) \to \infty$, and $\log L/L \to 0$. ☐

Theorem 410. *For $\beta > 0$, $n > 0$ and $L = \log(2\beta n) \geq 1$: $\sqrt{\log k^*/n} \leq \sqrt{L/(2\beta n)}$.*

Proof. $\log k^* = (L - \log L)/(2\beta) \leq L/(2\beta)$ since $\log L \geq 0$. ☐

Theorem 411. *For $\beta > 0$ and $\log n \geq \max\{1, 2|\log 2\beta|\}$: $\frac{1}{2}\log n \leq \log(2\beta n) \leq (1+|\log 2\beta|)\log n$.*

Proof. $\log(2\beta n) = \log 2\beta + \log n$ and $-|c| \leq c \leq |c| \leq |c| \log n$. ☐

Theorem 412. *For $\beta > 0$, eventually in n: $\sqrt{1/(4\beta)}\sqrt{\log n/n} \leq \mathsf{gen}(k^*, n) \leq 2\sqrt{(1 + |\log 2\beta|)/(2\beta)}\,\sqrt{\log n/n}$.*

Proof. Both terms are at most $\sqrt{L/(2\beta n)}$ with $L = \log(2\beta n)$ (Theorem 406, Theorem 410), the bias term equals it, and $\frac{1}{2}\log n \leq L \leq (1 + |\log 2\beta|)\log n$ (Theorem 411). ☐

Theorem 413. ***PL regime*** *(??). Let $\beta > 0$ and $k^*(n) = (2\beta n/\log(2\beta n))^{1/(2\beta)}$. Then* $\mathsf{gen}(k^*, n) = {k^*}^{-\beta} + \sqrt{\log k^*/n} \asymp \sqrt{\log n/n}$ *as $n \to \infty$.*

Proof. Rescale the lower bound of Theorem 412 by $\sqrt{1/(4\beta)}^{-1}$ and apply Theorem 397 to the function $\sqrt{1/(4\beta)}\,\sqrt{\log n/n}$, which is $\Theta(\sqrt{\log n/n})$. ☐

Theorem 414. *PL regime, upper bound:* $\mathsf{gen}(k^*, n) = O(\sqrt{\log n/n})$.

Proof. The O-half of Theorem 413. ☐

Theorem 415. *For $n > 1$, $\sqrt{\log n/n} = n^{-1/2}(\log n)^{1/2}$.*

Proof. Theorem 390 with $\gamma = 1$. ☐

Theorem 416. ***Ordering note****, first half. If $\gamma \le 1$ then the EP rate is no worse than the PL rate: $n^{-1/2}(\log n)^{\gamma/2} = O(\sqrt{\log n/n})$.*

Proof. For $\log n \ge 1$, $(\log n)^{\gamma/2} \le (\log n)^{1/2}$. $\square$

Theorem 417. ***Ordering note****, second half. If $\gamma \ge 1$ then the PL rate is no worse than the EP rate: $\sqrt{\log n/n} = O(n^{-1/2}(\log n)^{\gamma/2})$.*

Proof. For $\log n \ge 1$, $(\log n)^{1/2} \le (\log n)^{\gamma/2}$. $\square$

Theorem 418. ***Ordering note*** *(**??**). The EP and PL balanced rates are not uniformly ordered: EP $\lesssim$ PL when $\gamma \le 1$, and PL $\lesssim$ EP when $\gamma \ge 1$; for $\gamma = 1$ they coincide up to constants.*

Proof. Theorem 416 and Theorem 417. $\square$

Chapter 7

Worked examples (Appendices J–M)

7.1 Shared glue for the worked examples (regimes)

Theorem 419. *$L[f] \geq 0$ since the loss is nonnegative.*

Proof. □

Theorem 420. ***(Approximation transfer, 'eq:approx-transfer'.)*** *Let ℓ be β_ℓ-Lipschitz in its first argument, $\mathcal{H}$ and $\mathcal{C} \neq \emptyset$ classes of measurable functions, and $B \in \mathbb{R}$. If every $c \in \mathcal{C}$ is uniformly approximable from $\mathcal{H}$ within $B + \varepsilon$ for every $\varepsilon > 0$ (in particular if $\sup_{c \in \mathcal{C}} \inf_{f \in \mathcal{H}} \|f - c\|_\infty \leq B$), then*

$$\varepsilon_{\text{model}} = \inf_{\mathcal{H}} L - \inf_{\mathcal{C}} L \leq \beta_\ell \, B.$$

Proof. For $c \in \mathcal{C}$ and $\varepsilon > 0$ pick $f \in \mathcal{H}$ with $\|f - c\|_\infty \leq B + \varepsilon$; then $\inf_{\mathcal{H}} L \leq L[f] \leq L[c] + \beta_\ell(B + \varepsilon)$ by 'lem:risk-lipschitz'. Let $\varepsilon \to 0$ and take the infimum over c. □

Theorem 421. *For $h = \langle w, \Phi(\cdot) \rangle \in H_R(\Phi)$, $|h(x)| \leq R \, \|\Phi(x)\|$.*

Proof. □

Theorem 422. *If Φ is L_Φ-Lipschitz and $R \geq 0$ then every $h \in H_R(\Phi)$ is RL_Φ-Lipschitz.*

Proof. □

Theorem 423. *For $R > 0$, $H_R(\Phi) = H_1(R\,\Phi)$: the radius can be absorbed into the feature map.*

Proof. □

Theorem 424. *('prop:hilbert-sg' for the radius-R class.) If Φ is L_Φ-Lipschitz and $R > 0$, then $H_R(\Phi)$ satisfies the sub-Gaussian increment condition 'ass:sg-increment-main' with $A_H = 1$ and $L = RL_\Phi$, for every hidden-layer class $\mathfrak{F}$.*

Proof. $H_R(\Phi) = H_1(R\Phi)$ and $R\Phi$ is RL_Φ-Lipschitz. □

Theorem 425. *If $|g(x)| \leq C$ for all $g \in G$ and x ($C \geq 0$), then the Rademacher averages $\{\frac{1}{n}\sum_i \sigma_i g(x_i) : g \in G\}$ are bounded above by C, for every sample and sign pattern.*

Proof. $\square$

Theorem 426. ***(Rademacher complexity of the linear readout class.)*** *If $\|\Phi(x)\| \le M$ for all x $(R, M \ge 0)$, then for every sample S of size n,*

$$\hat{\mathfrak{R}}_S(H_R(\Phi)) \le \frac{R\,M}{\sqrt{n}}.$$

(Proof: $\hat{\mathfrak{R}}_S(H_R(\Phi)) \le R\,\mathbb{E}_\sigma\|n^{-1}\sum_i \sigma_i \Phi(x_i)\| \le R\sqrt{\sum_i \|\Phi(x_i)\|^2/n} \le RM/\sqrt{n}$, via FoML's `hilbertPredictor_empiricalRademacherComplexity_le`.)

Proof. $\square$

Theorem 427. *If F is finite then every word ball $B(k, F)$ is finite.*

Proof. $\square$

Theorem 428. *If Φ is continuous and every $f \in \mathfrak{F}$ is continuous, then every element of $H_R(\Phi) \circ \mathfrak{F}$ is (Borel) measurable.*

Proof. $\square$

Theorem 429. *If $\|\Phi\| \le M$ and $R \ge 0$ then $|g(x)| \le RM$ for every $g \in H_R(\Phi) \circ \mathfrak{F}$ and every x.*

Proof. $\square$

Definition 430. A hidden-layer class $\mathfrak{F}$ has a countable uniformly dense subset if there is a countable $D \subseteq \mathfrak{F}$ such that every $f \in \mathfrak{F}$ is within uniform distance ε of some $g \in D$, for every $\varepsilon > 0$.

Theorem 431. *A finite hidden-layer class has a countable uniformly dense subset (itself).*

Proof. $\square$

Theorem 432. *A hidden-layer class which is totally bounded in d_∞ has a countable uniformly dense subset (it is separable in the pseudo-metrizable space $(\mathcal{X}^{\mathcal{X}}, d_\infty)$).*

Proof. $\square$

Theorem 433. ***(Sup-norm separability of $H_R(\Phi) \circ \mathfrak{F}$.)*** *Let $\mathcal{H}$ be a separable real inner product space, $\Phi : \mathcal{X} \to \mathcal{H}$ be L_Φ-Lipschitz with $\|\Phi\| \le M$, $R \ge 0$, and let $\mathfrak{F}$ have a countable uniformly dense subset. Then $H_R(\Phi) \circ \mathfrak{F}$ is sup-norm separable: the countable set $\{\langle w, \Phi(g(\cdot))\rangle : w \in W,\ g \in D\}$, with W a countable dense subset of the ball of radius R, is uniformly dense, since $|\langle w, \Phi(fx)\rangle - \langle w', \Phi(gx)\rangle| \le \|w - w'\|M + RL_\Phi\, d(fx, gx)$.*

Proof. $\square$

Theorem 434. *A class which is totally bounded in d_∞ is totally bounded in d_S for every sample S (the identity is 1-Lipschitz, 'lem:to-emp-space-lipschitz').*

Proof. $\square$

Theorem 435. *Every element of the uniform closure of a class of measurable functions is measurable (a uniform limit of measurable functions is a pointwise limit).*

Proof. □

Definition 436. The estimation-plus-deviation term of ‘thm:bv‘ (with the constants of ‘thm:bv-general‘), as a function of an upper bound B for $\widehat{\mathfrak{R}}_S(\mathcal{H})$:

$$\mathrm{dev}_{\ell,n,\delta}(B) := 4\beta_\ell B + 6\, b\sqrt{\tfrac{2\log(4/\delta)}{n}}.$$

(All regime propositions below are stated in terms of this quantity, so that the constants of ‘thm:bv‘ enter in one place only.)

Theorem 437. *$B \mapsto \mathrm{dev}_{\ell,n,\delta}(B)$ is nondecreasing when $\beta_\ell \geq 0$.*

Proof. □

Theorem 438. ***(‘thm:bv‘ with $\iota = \mathrm{id}$ and explicit bounds.)*** *Let $\mathcal{H}$ be a sup-norm separable, pointwise bounded class of measurable functions, $\mathcal{C} \neq \emptyset$ a class of measurable benchmarks, ℓ measurable, bounded by $b > 0$ and β_ℓ-Lipschitz, $n \geq 1$, $\eta \geq 0$, $\delta \in (0,1)$. Suppose $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq B$ for every sample S of size n and $\varepsilon_{\mathrm{model}} \leq \mathrm{bias}$. Then with probability at least $1-\delta$ over $\mathcal{D} \sim P^{\otimes n}$, every η-empirical minimizer $\hat{f} \in \mathcal{H}$ satisfies*

$$L[\hat{f}] - \inf_{\mathcal{C}} L \leq \mathrm{bias} + \eta + \mathrm{dev}_{\ell,n,\delta}(B) = \mathrm{bias} + \eta + 4\beta_\ell B + 6\, b\sqrt{\tfrac{2\log(4/\delta)}{n}}.$$

Proof. ‘thm:bv-general‘ with $\iota = \mathrm{id}$, $d_T = 0$, $\varepsilon_{\mathrm{imp}} = 0$, then monotonicity of the bad event. □

Theorem 439. ***(Truncation, Sec. ‘sec:examples-regime‘.)*** *Let every $h \in H$ be L_H-Lipschitz, every $f \in F$ be c-Lipschitz, and $d(x,y) \leq D_{\mathcal{X}}$ on $\mathcal{X}$. Then every $g = h \circ w \in H \circ \langle F \rangle$ is within uniform distance $L_H c^k D_{\mathcal{X}}$ of $\mathcal{H}_k = H \circ B(k,F)$: if $w = w_2 \circ w_1$ with $|w_2| = k$ then $|h(w_2(w_1 x)) - h(w_2 x)| \leq L_H c^k d(w_1 x, x)$.*

Proof. □

Theorem 440. ***(Truncation for the uniform closure.)*** *Under the hypotheses of ‘lem:truncation-bias-exact‘, every g in the uniform closure $\overline{H \circ \langle F \rangle}^{d_\infty}$ satisfies, for every $\varepsilon > 0$, $\inf_{f \in \mathcal{H}_k} \|f - g\|_\infty \leq L_H c^k D_{\mathcal{X}} + \varepsilon$.*

Proof. □

Definition 441. The depth-independent entropy integral of ‘cor:profile-p1‘: $\mathsf{V}_\infty(F) := \int_0^{\mathrm{diam}(\mathcal{X})} \sqrt{\log N^{\mathrm{ext}}(\overline{\langle F \rangle}, d_\infty,$

Definition 442. The paper’s condition $\mathsf{V}_\infty(F) < \infty$: the majorant $\varepsilon \mapsto \sqrt{\log N^{\mathrm{ext}}(\overline{\langle F \rangle}, d_\infty, \varepsilon/2)}$ is interval-integrable on $[0, \mathrm{diam}(\mathcal{X})]$.

Theorem 443. ***(EL balance with saturated variance.)*** *For $0 < \theta < 1$ and $n \geq 1$, the depth $k := \lceil \log n / (2\log(1/\theta)) \rceil$ satisfies $\theta^k \leq n^{-1/2}$: the bias θ^k is balanced against the saturated variance $n^{-1/2}$, and $k = \frac{\log n}{2\log(1/\theta)} + O(1)$.*

Proof. $k\log(1/\theta) \geq \frac{1}{2}\log n = \log\sqrt{n}$, so $(1/\theta)^k \geq \sqrt{n}$. □

Theorem 444. ***(PL balance with saturated variance, $p = 1$.)*** *For $n \geq 1$ the depth $k := \lceil \sqrt{n} \rceil \geq 1$ satisfies $1/k \leq n^{-1/2}$: the bias k^{-1} is balanced against the saturated variance $n^{-1/2}$, and $k \asymp n^{1/2}$.*

Proof. □

7.2 Implementation with controlled uniform error (Appendix J, prop:implementation; proof in J.3)

Definition 445. For a list of layers $u = [f_m, \dots, f_1]$ (an explicit representation of a word) the composed map is $f_u = f_m \circ \cdots \circ f_1$, with $f_{[]} = \mathrm{id}$. In Lean: 'compList [] = id' and 'compList (f :: u) = f compList u' (the head of the list acts last, matching the recursion $B(k+1, F) = B(k, F) \cup F \circ B(k, F)$).

Theorem 446. $f_{[]} = \mathrm{id}$.

Proof. □

Theorem 447. $f_{f::u} = f \circ f_u$.

Proof. □

Definition 448. Given an implementation $f \mapsto \tilde{f}$ of the layers, the implemented word of a representation $u = [f_m, \dots, f_1]$ is $\tilde{f}_u = \tilde{f}_m \circ \cdots \circ \tilde{f}_1$.

Theorem 449. $\tilde{f}_{[]} = \mathrm{id}$.

Proof. □

Theorem 450. $\tilde{f}_{f::u} = \tilde{f} \circ \tilde{f}_u$.

Proof. □

Theorem 451. *If all layers of u lie in F and $|u| \le k$ then $f_u \in B(k, F)$.*

Proof. Induction on u: $\mathrm{id} \in B(k, F)$, and $f \circ f_u \in B(k'+1, F)$ when $f \in F$ and $f_u \in B(k', F)$. □

Theorem 452. *Every $g \in B(k, F)$ has a representation $g = f_m \circ \cdots \circ f_1$ with $f_i \in F$ and $m \le k$, i.e. $g = f_u$ for a list u of elements of F with $|u| \le k$.*

Proof. Induction on k along the recursion $B(k+1, F) = B(k, F) \cup F \circ B(k, F)$. □

Theorem 453. *(Telescoping error propagation.) Let every $f \in F$ be Λ-Lipschitz and let $\tilde{f}$ satisfy $d_\infty(f, \tilde{f}) \le \delta$ for $f \in F$, where $\delta \ge 0$. Then for every representation $u = [f_m, \dots, f_1]$ of layers in F and every x,*

$$d(f_m \circ \cdots \circ f_1(x),\ \tilde{f}_m \circ \cdots \circ \tilde{f}_1(x)) \le \delta \sum_{i=0}^{m-1} \Lambda^i.$$

Proof. Induction on u. For $u = f :: u'$ write $y = f_{u'}(x)$, $\tilde{y} = \tilde{f}_{u'}(x)$; then $d(f(y), \tilde{f}(\tilde{y})) \le d(f(y), f(\tilde{y})) + d(f(\tilde{y}), \tilde{f}(\tilde{y})) \le \Lambda\, d(y, \tilde{y}) + \delta \le \Lambda\delta \sum_{i<m} \Lambda^i + \delta = \delta \sum_{i<m+1} \Lambda^i$. □

Theorem 454. *(Existence of the layerwise implementation map.) Given implementations $f \mapsto \tilde{f}$ of the transitions and $h \mapsto \tilde{h}$ of the output layers, there is a map $\iota : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ such that every $g \in \mathcal{H}_k$ has a representation $g = h \circ f_m \circ \cdots \circ f_1$ with $h \in H$, $f_i \in F$, $m \le k$ and $\iota(g) = \tilde{h} \circ \tilde{f}_m \circ \cdots \circ \tilde{f}_1$.*

Proof. For each $g \in \mathcal{H}_k$ choose (by the axiom of choice) a representation, using 'lem:exists-comp-list-of-mem-wordball'; outside $\mathcal{H}_k$ the value of ι is irrelevant. □

Theorem 455. ***(Layerwise implementation.)*** *Let $F \subseteq \mathcal{X}^{\mathcal{X}}$ with $\operatorname{lip}(f) \le \Lambda$ for all $f \in F$, and $H \subseteq \mathbb{R}^{\mathcal{X}}$ with $\operatorname{lip}(h) \le L_H$ for all $h \in H$. Suppose every $f \in F$ is assigned an implemented map $\tilde{f}$ with $d_\infty(f, \tilde{f}) \le \delta$ ($\delta \ge 0$) and every $h \in H$ an implemented $\tilde{h}$ with $\|h - \tilde{h}\|_\infty \le \delta_H$. Let ι be an implementation map that sends each $g \in \mathcal{H}_k$, for one representation $g = h \circ f_m \circ \cdots \circ f_1$ with $m \le k$, to $\tilde{h} \circ \tilde{f}_m \circ \cdots \circ \tilde{f}_1$. Then*

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

(Stated for an arbitrary ι with this property; such an ι exists by 'lem:impl-map-exists'.)

Proof. Fix $g = h \circ f_u \in \mathcal{H}_k$ and x; put $T_0 = f_u(x)$, $T_m = \tilde{f}_u(x)$. By 'lem:impl-word-error', $d(T_0, T_m) \le \delta \sum_{i<m} \Lambda^i \le \delta \sum_{i<k} \Lambda^i$ (as $\Lambda \ge 0$ and $m \le k$). Then $|h(T_0) - \tilde{h}(T_m)| \le |h(T_0) - h(T_m)| + |h(T_m) - \tilde{h}(T_m)| \le L_H\, d(T_0, T_m) + \delta_H$; take suprema over x and g. ☐

Theorem 456. *Under the hypotheses of 'prop:implementation-b' there exists an implementation map ι with $\varepsilon_{\mathrm{imp}}(k) \le \delta_H + L_H\, \delta \sum_{i<k} \Lambda^i$.*

Proof. Combine 'lem:impl-map-exists' and 'prop:implementation-b'. ☐

Theorem 457. *If $0 \le \Lambda \le 1$ then $\sum_{i<k} \Lambda^i \le k$.*

Proof. ☐

Theorem 458. *If $0 \le \Lambda < 1$ then $\sum_{i<k} \Lambda^i \le 1/(1 - \Lambda)$.*

Proof. $\sum_{i<k} \Lambda^i = (1 - \Lambda^k)/(1 - \Lambda) \le 1/(1 - \Lambda)$. ☐

Theorem 459. *(Non-expanding transitions.) If moreover $\Lambda \le 1$, then $\varepsilon_{\mathrm{imp}}(k) \le \delta_H + L_H\, k\, \delta$.*

Proof. 'prop:implementation-b' and $\sum_{i<k} \Lambda^i \le k$. ☐

Theorem 460. *(Contractive transitions.) If moreover $\Lambda < 1$, then $\varepsilon_{\mathrm{imp}}(k) \le \delta_H + L_H\, \delta/(1-\Lambda)$.*

Proof. 'prop:implementation-b' and $\sum_{i<k} \Lambda^i \le 1/(1 - \Lambda)$. ☐

Definition 461. $\operatorname{relu}(t) = \max\{0, t\}$.

Definition 462. A one-hidden-layer ReLU layer on $\mathbb{R}^d$ with hidden index set m (width $|m|$) is $x \mapsto W_2 \operatorname{relu}(W_1 x) + b$, with $W_1 \in \mathbb{R}^{m \times d}$, $W_2 \in \mathbb{R}^{d \times m}$, $b \in \mathbb{R}^d$ and relu applied coordinatewise.

Theorem 463. $t = \operatorname{relu}(t) - \operatorname{relu}(-t)$.

Proof. ☐

Theorem 464. *(Exact realization of an affine map.) For $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$, with $W_1 = [I_d; -I_d] \in \mathbb{R}^{2d \times d}$ and $W_2 = [A, -A] \in \mathbb{R}^{d \times 2d}$,*

$$W_2 \operatorname{relu}(W_1 x) + b = Ax + b \qquad (x \in \mathbb{R}^d),$$

i.e. $x \mapsto Ax + b$ is one ReLU layer of width $2d$.

Proof. $\operatorname{relu}(W_1 x) = (\operatorname{relu}(x), \operatorname{relu}(-x))$ and $[A, -A](v_1, v_2) = A(v_1 - v_2)$; conclude with $\operatorname{relu}(t) - \operatorname{relu}(-t) = t$. ☐

Definition 465. The parameters (W_2, W_1, b) of a ReLU layer of width $2d$ on $\mathbb{R}^d$.

Definition 466. The ReLU layer of width $2d$ with parameters (W_2, W_1, b).

Definition 467. A map $g : \mathbb{R}^d \to \mathbb{R}^d$ is a ReLU network of depth $\le k$ and width $2d$ if it is a composition of at most k ReLU layers of width $2d$.

Theorem 468. *If every layer of the representation u is affine, then f_u is a composition of $|u|$ ReLU layers of width $2d$.*

Proof. Induction on u, replacing each affine layer by the ReLU layer of 'lem:affine-eq-relu'. ☐

Theorem 469. *The identity implementation map has zero implementation error: $\varepsilon_{\text{imp}} = 0$ for $\iota = \text{id}$ on any class.*

Proof. ☐

Theorem 470. ***(Exact implementation of affine transitions.)*** *Let $\mathcal{X} = \mathbb{R}^d$ and let every $f \in F$ be affine, $f(x) = Ax + b$. Then every element of $B(k, F)$ is a ReLU network of depth $\le k$ and width $2d$, and with the identity implementation map $\iota = \text{id}$ (the abstract class $\mathcal{H}_k$ is itself the implemented class) one has $\varepsilon_{\text{imp}}(k) = 0$.*

Proof. Take a representation of $g \in B(k, F)$ ('lem:exists-comp-list-of-mem-wordball') and apply 'lem:comp-list-affine-eq-relu-net'; the second claim is 'lem:impl-error-id'. ☐

Theorem 471. ***(Net-based implementation.)*** *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be totally bounded in the uniform norm and let $\mathcal{A} \subseteq \mathbb{R}^{\mathcal{X}}$ be uniformly dense on $\mathcal{H}$ (for every $g \in \mathcal{H}$ and $\eta > 0$ there is $a \in \mathcal{A}$ with $\|g - a\|_\infty \le \eta$). Then for every $\varepsilon > 0$ there exist a finite class $\mathcal{H}^{\varepsilon}_{\text{imp}} \subseteq \mathcal{A}$ with $|\mathcal{H}^{\varepsilon}_{\text{imp}}| \le N(\mathcal{H}, \|\cdot\|_\infty, \varepsilon/2)$ and an implementation map $\iota : \mathcal{H} \to \mathcal{H}^{\varepsilon}_{\text{imp}}$ with $\sup_{g\in\mathcal{H}} \|g - \iota(g)\|_\infty \le \varepsilon$. (The paper's compact domain K and the continuity of the functions are only used to guarantee total boundedness and density, which are the hypotheses here.)*

Proof. Let $C \subseteq \mathcal{H}$ be a minimal internal $\varepsilon/2$-net (finite by total boundedness). For each $c \in C$ choose $a_c \in \mathcal{A}$ with $\|c - a_c\|_\infty \le \varepsilon/2$, set $\mathcal{H}_{\text{imp}} = \{a_c : c \in C\}$ and $\iota(g) = a_{c(g)}$ where $c(g) \in C$ satisfies $\|g - c(g)\|_\infty \le \varepsilon/2$. Then $\|g - \iota(g)\|_\infty \le \varepsilon/2 + \varepsilon/2$. ☐

7.3 Deep ReLU networks (Appendix K, lem:relu-layer-covering, prop:relu-regimes)

Theorem 472. *Packing numbers are monotone in the set: $A \subseteq B$ implies $M(A, \varepsilon) \le M(B, \varepsilon)$.*

Proof. ☐

Theorem 473. *(Volumetric bound, finite sets.) Let E be a real normed space of dimension q, $\rho \ge 0$, $\delta > 0$. If $s \subseteq \overline{B}(x, \rho)$ is a finite δ-separated set (distinct points at distance $> \delta$) then $|s| \le (1 + 2\rho/\delta)^q$: the open balls $B(y, \delta/2)$, $y \in s$, are pairwise disjoint and contained in $B(x, \rho + \delta/2)$, and Haar measure scales like r^q.*

Proof. The balls $B(y, \delta/2)$, $y \in s$, are pairwise disjoint.
Each of them lies in $B(x, \rho + \delta/2)$.
Comparing measures: $|s| \cdot (\delta/2)^q \mu(B(0, 1)) \le (\rho + \delta/2)^q \mu(B(0, 1))$. ☐

Theorem 474. *(Volumetric bound.) In a real normed space E of dimension q, for $\rho \geq 0$ and $\delta > 0$, the packing number of the closed ball satisfies $M(\overline{B}(x, \rho), \delta) \leq \lfloor (1 + 2\rho/\delta)^q \rfloor$; in particular a δ-cover of the ball of size at most $(1 + 2\rho/\delta)^q$ exists.*

Proof. Every separated subset is finite (its finite subsets have bounded cardinality) and its cardinality is bounded by ‘lem:relu-volumetric-finset‘. □

Theorem 475. *If φ is L-Lipschitz on A then $N^{\text{ext}}(\varphi(A), L\varepsilon) \leq N(A, \varepsilon)$ (internal covering number of A on the right, since the centres must lie in A where φ is controlled).*

Proof. □

Theorem 476. ***(Parametric layers, ‘cor:envelope-profiles‘(a) for the hypothesis** $\log N(F, \varepsilon) \leq p \log(1 + C/\varepsilon)$**.)** Let every $f \in F$ be Λ-Lipschitz, let $\overline{D} > 0$, $C \geq 0$, $p \geq 0$, and suppose $N^{\text{ext}}(F, d_\infty, \varepsilon) < \infty$ and $\log N^{\text{ext}}(F, d_\infty, \varepsilon) \leq p \log(1 + C/\varepsilon)$ for all $\varepsilon > 0$. If $D_k(S) \leq \overline{D}$ then, for $k \geq 1$, with $\Lambda_+ := \max\{1, \Lambda\}$,*

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

(Same proof as ‘cor:envelope-profiles-a‘, with $1 + 2CS_k/\varepsilon \leq (1 + 2C/\overline{D})\, S_k\, \overline{D}/\varepsilon$ for $\varepsilon \leq \overline{D}$.)

Proof. Pointwise, for $0 < \varepsilon \leq \overline{D}$: $\log N(B(k, F), d_S, \varepsilon) \leq \log N^{\text{ext}}(B(k, F), d_\infty, \varepsilon/2) \leq \log(k + 1) + kp \log(1 + 2CS_k/\varepsilon)$, and $1 + 2CS_k/\varepsilon \leq (1 + 2C/\overline{D})S_k\overline{D}/\varepsilon$ with $\log S_k \leq \log k + k \log \Lambda_+$; take square roots termwise and integrate, using $\int_0^{\overline{D}} \sqrt{\log(\overline{D}/\varepsilon)}\, d\varepsilon = \frac{\sqrt{\pi}}{2}\overline{D}$. □

Definition 477. The ReLU nonlinearity acting coordinatewise on $\mathbb{R}^w$ (the scalar ‘relu‘ of ‘LeanDeepgen.Examples.Implementation‘ in every coordinate): $\text{relu}(z)_i = \max\{z_i, 0\}$.

Theorem 478. $\text{relu}(z)_i = \max\{z_i, 0\}$.

Proof. □

Theorem 479. relu *is* 1*-Lipschitz for the Euclidean norm, since* $|\max\{a, 0\} - \max\{b, 0\}| \leq |a - b|$ *coordinatewise.*

Proof. □

Theorem 480. $\text{relu}(0) = 0$.

Proof. □

Theorem 481. $\|\text{relu}(z)\| \leq \|z\|$.

Proof. □

Definition 482. The parameter space of a ReLU block of width w on $\mathbb{R}^m$ ($= E$): $\vartheta = (W, b, V, c) \in (\mathbb{R}^m \to \mathbb{R}^w) \times \mathbb{R}^w \times (\mathbb{R}^w \to \mathbb{R}^m) \times \mathbb{R}^m$, with the operator norm on the linear maps and the sup (product) norm on the tuple.

Definition 483. The ReLU block with parameters $\vartheta = (W, b, V, c)$ on the state space K: $f_\vartheta(x) = \Pi_K(V\,\text{relu}(Wx + b) + c)$, a self-map of K.

Definition 484. The admissible parameters: $\|W\|_{\text{op}}, \|V\|_{\text{op}} \leq \beta_W$, $\|b\|, \|c\| \leq \beta$ and $\text{lip}(f_\vartheta) \leq \Lambda$.

Definition 485. The hidden-layer class of ReLU blocks of width w, $F_\Lambda = \{f_\vartheta : \|W\|_{\mathrm{op}}, \|V\|_{\mathrm{op}} \le \beta_W,\ \|b\|, \|c\| \le \beta,\ \mathrm{lip}(f_\vartheta) \le \Lambda\}$.

Theorem 486. *Every $f \in F_\Lambda$ is Λ-Lipschitz.*

Proof. □

Theorem 487. *$F_\Lambda \ne \emptyset$: the zero parameters give the constant map $x \mapsto \Pi_K(0)$.*

Proof. □

Theorem 488. *The admissible parameters lie in the sup-norm ball of radius* $\max\{\beta_W, \beta\}$.

Proof. □

Theorem 489. *(Parameter-Lipschitz estimate.) Let $\|x\| \le R_K$ on K, $\|V\|_{\mathrm{op}} \le \beta_W$, $\|W'\|_{\mathrm{op}} \le \beta_W$ and $\|b'\| \le \beta$. Then*

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

(Only these three parameter bounds are used, as in the paper's proof.)

Proof. For $x \in K$ write $z = \mathrm{relu}(Wx + b)$, $z' = \mathrm{relu}(W'x + b')$; then $\|z - z'\| \le R_K\|W - W'\| + \|b - b'\|$, $\|z'\| \le \beta_W R_K + \beta$ and $\|Vz + c - V'z' - c'\| \le \|V\|\|z - z'\| + \|V - V'\|\|z'\| + \|c - c'\|$; apply the 1-Lipschitz retraction. □

Definition 490. The parameter-Lipschitz constant of $\vartheta \mapsto f_\vartheta$ on F_Λ for the sup norm on parameters: $L_F := 2\beta_W R_K + \beta_W + \beta + 1$ (the sum of the four coefficients of 'lem:relu-layer-covering-a').

Definition 491. The covering constant of one ReLU block: $C_F := 2L_F \max\{\beta_W, \beta\} = 2(2\beta_W R_K + \beta_W + \beta + 1)\max\{\beta_W, \beta\}$. (The paper's $C_F = 8\sqrt{\max\{m, w\}}\max\{\beta_W, 1\}\max\{\beta_W R_K + \beta, \beta_W, 1\}$ has the same shape; the factor $\sqrt{\max\{m, w\}}$ comes from bounding the operator norm by the Frobenius norm, which the volumetric argument in the operator norm avoids.)

Theorem 492. *$L_F \ge 1$ when $R_K \ge 0$.*

Proof. □

Theorem 493. *$\vartheta \mapsto f_\vartheta$ is L_F-Lipschitz on the admissible parameter set, from* (parameters, $\|\cdot\|_{\mathrm{sup}}$) *to $(\mathcal{X}^{\mathcal{X}}, d_\infty)$: each of the four differences in 'lem:relu-layer-covering-a' is at most $\|\vartheta - \vartheta'\|_{\mathrm{sup}}$.*

Proof. □

Theorem 494. *The parameter space has dimension $p = 2mw + w + m$, $m = \dim E$.*

Proof. □

Theorem 495. *(Covering bound, cardinality form.) Let $\|x\| \le R_K$ on K with $R_K \ge 0$ and $p = 2mw + w + m$. For every $\varepsilon > 0$, $N^{\mathrm{ext}}(F_\Lambda, d_\infty, \varepsilon) \le (1 + C_F/\varepsilon)^p$ with $C_F = 2(2\beta_W R_K + \beta_W + \beta + 1)\max\{\beta_W, \beta\}$: $N^{\mathrm{ext}}(F_\Lambda, \varepsilon) \le N(P, \varepsilon/L_F) \le M(P, \varepsilon/L_F) \le M(\overline{B}(0, \max\{\beta_W, \beta\}), \varepsilon/L_F) \le (1 + 2L_F \max\{\beta_W, \beta\}/\varepsilon)^p$ by 'lem:relu-lipschitz-on-image', 'lem:packing-covering' and 'lem:relu-volumetric'.*

Proof. □

Theorem 496. *(Covering bound, entropy form.) Under the hypotheses of 'lem:relu-layer-covering-b', for every $\varepsilon > 0$, $N^{\text{ext}}(F_\Lambda, d_\infty, \varepsilon) < \infty$ and $\log N^{\text{ext}}(F_\Lambda, d_\infty, \varepsilon) \le p \log(1 + C_F/\varepsilon)$.*

Proof. □

Theorem 497. ***(Covering numbers of one ReLU block.)*** *Let $K \subseteq \mathbb{R}^m$ with $\|x\| \le R_K$ on K ($R_K \ge 0$), let Π_K be a 1-Lipschitz retraction onto K, and let F_Λ be the class of ReLU blocks of width w with $p = 2mw + w + m$ parameters. (a) For admissible ϑ, ϑ',*

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

(b) for every $\varepsilon > 0$, $N^{\text{ext}}(F_\Lambda, d_\infty, \varepsilon) < \infty$ and $\log N^{\text{ext}}(F_\Lambda, d_\infty, \varepsilon) \le p \log(1 + C_F/\varepsilon)$ with $C_F = 2(2\beta_W R_K + \beta_W + \beta + 1) \max\{\beta_W, \beta\}$ ('def:relu-cover-const'; the paper's constant has the same shape, see there).

Proof. □

Theorem 498. *If K is bounded then so is the state space K (as a metric space in its own right).*

Proof. □

Theorem 499. *A bounded set K of a finite-dimensional space has finite covering numbers: $N^{\text{ext}}(K, \varepsilon) < \infty$ for $\varepsilon > 0$ (external covering number of the state space K by points of K; K is totally bounded since its closure is compact).*

Proof. □

Theorem 500. *(Telescoping estimate for contractive layers.) If every $f \in F$ is Λ-Lipschitz with $\Lambda < 1$ then $N^{\text{ext}}(F^l, d_\infty, \varepsilon) \le N^{\text{ext}}(F, d_\infty, (1-\Lambda)\varepsilon)^l$, since $S_l(\Lambda) = \sum_{i<l} \Lambda^i \le 1/(1-\Lambda)$ in 'lem:envelope-words'.*

Proof. □

Theorem 501. *For $N \ge 1$ in $\mathbb{N} \cup \{\infty\}$, $\sum_{l<m} N^l \le m N^m$.*

Proof. □

Theorem 502. $N^{\text{ext}}(F_\Lambda, d_\infty, \varepsilon) \ge 1$ *(the class is nonempty).*

Proof. □

Theorem 503. ***('prop:relu-regimes'(i): contractive layers, $\Lambda < 1$.)*** *Let $K \ne \emptyset$ be bounded with $\|x\| \le R_K$ on K, and $0 < \Lambda < 1$. Then 'cond:p1-ucont' applies with the invariant set K itself ($L = 0$, absorbing set K), and with $m(\varepsilon) = \lceil \log(2D_K/\varepsilon)/\log(1/\Lambda) \rceil$ ($D_K = \operatorname{diam} K$), for every $\varepsilon > 0$ and* every *k,*

$$N^{\text{ext}}(B(k, F_\Lambda), d_\infty, \varepsilon) \le N^{\text{ext}}(K, \varepsilon/2) + m(\varepsilon)\, N^{\text{ext}}(F_\Lambda, d_\infty, (1-\Lambda)\varepsilon)^{m(\varepsilon)} < \infty,$$

which does not depend on k; in particular $\sup_k N^{\text{ext}}(B(k, F_\Lambda), d_\infty, \varepsilon) < \infty$. (The paper's $k \ge m(\varepsilon)$ is not needed, cf. 'cond:p1-ucont'; the right-hand side is finite by 'lem:relu-layer-covering' and the total boundedness of K.)

Proof. 'cond:p1-ucont' with $c = \Lambda$, $A = K$, $L = 0$ and the bounded absorbing set K; the short words are bounded by 'lem:relu-words-covering-contractive' and 'lem:relu-sum-pow-le'; finiteness from 'lem:relu-covering-univ-ne-top' and 'lem:relu-layer-covering-log'. □

Definition 504. The k-independent majorant of case (i) at the scale of the empirical metric: $N_\infty(\varepsilon) := N^{\mathrm{ext}}(K, \varepsilon/4) + m(\varepsilon/2)\, N^{\mathrm{ext}}(F_\Lambda, d_\infty, (1-\Lambda)\varepsilon/2)^{m(\varepsilon/2)}$ (the factor 2 from $N(A, d_S, \varepsilon) \le N^{\mathrm{ext}}(A, d_\infty, \varepsilon/2)$).

Theorem 505. *(Case (i), saturated profile: $\mathsf{V}_k(S) = O(1)$.) Under the hypotheses of 'prop:relu-regimes-i', if $\varepsilon \mapsto \sqrt{\log N_\infty(\varepsilon)}$ is interval-integrable on $[0, D_K]$ then $\mathsf{V}_k(S) \le \mathsf{V}_\infty := \int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon$ for all k and all samples S (case (i) of 'prop:profiles').*

Proof. $N(B(k,F), d_S, \varepsilon) \le N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon/2) \le N_\infty(\varepsilon)$ by 'prop:relu-regimes-i', and $D_k(S) \le D_K$. ☐

Theorem 506. ***('prop:relu-regimes'(ii): non-expanding layers,** $\Lambda \le 1$.**)** If K is compact and $\Lambda \le 1$ then 'cond:p1' (2c, non-expanding generators on a compact state space) applies: for every $\varepsilon > 0$ and every k, $N^{\mathrm{ext}}(B(k, F_\Lambda), d_\infty, \varepsilon) \le N^{\mathrm{ext}}(\overline{\langle F_\Lambda \rangle}, d_\infty, \varepsilon) < \infty$; in particular $\sup_k N^{\mathrm{ext}}(B(k, F_\Lambda), d_\infty, \varepsilon) < \infty$.*

Proof. 'cond:p1-2c' on the compact state space K. ☐

Theorem 507. *(Case (ii), saturated profile: $\mathsf{V}_k(S) = O(1)$.) Under the hypotheses of 'prop:relu-regimes-ii', with $N_\infty(\varepsilon) := N(\overline{\langle F_\Lambda \rangle}, d_\infty, \varepsilon/2) < \infty$: if $\int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ then $\mathsf{V}_k(S) \le \int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon$ for all k and all samples S (case (i) of 'prop:profiles', as in 'lem:fp-profile').*

Proof. The non-expanding semigroup is equicontinuous, hence totally bounded in d_∞ ('thm:caa'); apply 'lem:ode-profile-saturation-generic'. ☐

Theorem 508. *(Envelope profile of the ReLU class, all Λ.) Let K be bounded with $\|x\| \le R_K$ on K, $D_K \le \overline{D}$, $\overline{D} > 0$, $p = 2mw + w + m$ and C_F as in 'lem:relu-layer-covering'. Then for $k \ge 1$ and every sample S, with $\Lambda_+ = \max\{1, \Lambda\}$,*

$$\mathsf{V}_k(S) \le \overline{D}\Big(\sqrt{\log(k+1)} + \sqrt{kp\log k} + k\sqrt{p\log\Lambda_+} + \sqrt{kp}\big(\sqrt{\log(1 + 2C_F/\overline{D})} + \tfrac{\sqrt{\pi}}{2}\big)\Big)$$

('cor:envelope-profiles-a-one-add' with 'lem:relu-layer-covering').

Proof. ☐

Theorem 509. *(Case (ii), explicit envelope: $\mathsf{V}_k(S) = O(\sqrt{kp\log k})$.) For $\Lambda \le 1$, under the hypotheses of 'lem:relu-envelope-profile',*

$$\mathsf{V}_k(S) \le \overline{D}\Big(\sqrt{\log(k+1)} + \sqrt{kp\log k} + \sqrt{kp}\big(\sqrt{\log(1 + 2C_F/\overline{D})} + \tfrac{\sqrt{\pi}}{2}\big)\Big).$$

Proof. ☐

Theorem 510. ***('prop:relu-regimes'(iii), upper bound: expanding layers,** $\Lambda > 1$**:** $\mathsf{V}_k(S) = O(k\sqrt{p\log\Lambda})$.**)** Under the hypotheses of 'lem:relu-envelope-profile', for $\Lambda > 1$,*

$$\mathsf{V}_k(S) \le \overline{D}\Big(\sqrt{\log(k+1)} + \sqrt{kp\log k} + k\sqrt{p\log\Lambda} + \sqrt{kp}\big(\sqrt{\log(1 + 2C_F/\overline{D})} + \tfrac{\sqrt{\pi}}{2}\big)\Big).$$

Proof. ☐

Theorem 511. ***(Depth profiles of ReLU networks by Lipschitz constant, upper bounds.)*** *Let $K \neq \emptyset$ be bounded with $\|x\| \leq R_K$ on K ($R_K \geq 0$), $F = F_\Lambda$, $p = 2mw + w + m$ and C_F as in 'lem:relu-layer-covering'.*

(i) (Contractive, $0 < \Lambda < 1$.) For every $\varepsilon > 0$ and k, $N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon) \leq N^{\mathrm{ext}}(K, \varepsilon/2) + m(\varepsilon) N^{\mathrm{ext}}(F, d_\infty, (1-\Lambda)\varepsilon)^{m(\varepsilon)} < \infty$ with $m(\varepsilon) = \lceil \log(2D_K/\varepsilon)/\log(1/\Lambda) \rceil$, so $\sup_k N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon) < \infty$.

(ii) (Non-expanding, $\Lambda \leq 1$.) If K is compact then for every $\varepsilon > 0$ and k, $N^{\mathrm{ext}}(B(k,F), d_\infty, \varepsilon) \leq N^{\mathrm{ext}}(\overline{\langle F \rangle}, d_\infty, \varepsilon) < \infty$; and for $D_K \leq \overline{D}$, $\overline{D} > 0$, $k \geq 1$, the envelope gives $\mathsf{V}_k(S) \leq \overline{D}(\sqrt{\log(k+1)} + \sqrt{kp \log k} + \sqrt{kp}(\sqrt{\log(1 + 2C_F/\overline{D})} + \sqrt{\pi}/2)) = O(\sqrt{kp \log k})$.

(iii) (Expanding, $\Lambda > 1$.) For $D_K \leq \overline{D}$, $\overline{D} > 0$, $k \geq 1$, $\mathsf{V}_k(S) \leq \overline{D}(\sqrt{\log(k+1)} + \sqrt{kp \log k} + k\sqrt{p \log \Lambda} + \sqrt{kp}(\sqrt{\log(1 + 2C_F/\overline{D})} + \sqrt{\pi}/2)) = O(k\sqrt{p \log \Lambda})$.

The lower bound of (iii) is 'prop:relu-regimes-iii-lower'; the saturated profiles $\mathsf{V}_k(S) = O(1)$ of (i) and (ii) are 'prop:relu-regimes-i-profile' and 'prop:relu-regimes-ii-profile'.

Proof. ☐

Definition 512. The clipping $\Pi_{[0,1]}(x) = \max\{0, \min\{1, x\}\}$, a 1-Lipschitz retraction of $\mathbb{R}$ onto $[0,1]$.

Theorem 513. $\Pi_{[0,1]}(x) = x$ *for* $x \in [0,1]$.

Proof. ☐

Theorem 514. $\Pi_{[0,1]}$ *is a* 1*-Lipschitz retraction onto* $[0,1]$ *('def:ode-projection').*

Proof. ☐

Theorem 515. *$x \mapsto ax + b$ is L-Lipschitz on $\mathbb{R}$ when $|a| \leq L$.*

Proof. ☐

Definition 516. The first expand-and-reset map of the computed illustration ('sec:relu-computed'), $\eta = 1/32$: the piecewise-linear map with breakpoints $(0, 3/8), (\eta, 0), (1/4-\eta, 1), (1/4, 3/8), (1, 3/8)$, i.e. $g_0(x) = 3/8 - 12x$ on $[0, \eta]$, $= \frac{16}{3}(x - \eta)$ on $[\eta, 1/4 - \eta]$, $= 1 - 20(x - 1/4 + \eta)$ on $[1/4 - \eta, 1/4]$ and $= 3/8$ on $[1/4, 1]$.

Definition 517. The second expand-and-reset map, with breakpoints $(0, 5/8), (3/4, 5/8), (3/4 + \eta, 0), (1 - \eta, 1), (1, 5/8)$, i.e. $g_1(x) = 5/8$ on $[0, 3/4]$, $= 5/8 - 20(x - 3/4)$ on $[3/4, 3/4 + \eta]$, $= \frac{16}{3}(x - 3/4 - \eta)$ on $[3/4 + \eta, 1 - \eta]$ and $= 1 - 12(x - 1 + \eta)$ on $[1 - \eta, 1]$.

Theorem 518. g_0 *maps* $[0,1]$ *into* $[0,1]$.

Proof. ☐

Theorem 519. g_1 *maps* $[0,1]$ *into* $[0,1]$.

Proof. ☐

Theorem 520. *On* $[0,1]$, $g_0(x) = \max\{-12x + 3/8, \min\{\frac{16}{3}x - \frac{1}{6}, \max\{3/8, -20x + 43/8\}\}\}$ *(a max–min of affine maps of slopes* ≤ 20*).*

Proof. ☐

Theorem 521. *On* $[0,1]$, $g_1(x) = \max\{\min\{5/8, -20x+125/8\}, \min\{\frac{16}{3}x-\frac{25}{6}, -12x+101/8\}\}$.

Proof. ☐

Theorem 522. g_0 *is* 20*-Lipschitz on* $[0,1]$ *(max–min of* 20*-Lipschitz affine maps).*

Proof. ☐

Theorem 523. g_1 *is* 20*-Lipschitz on* $[0,1]$.

Proof. ☐

Definition 524. g_0 as a self-map of the state space $[0,1]$.

Definition 525. g_1 as a self-map of the state space $[0,1]$.

Theorem 526. $g_0 : [0,1] \to [0,1]$ *is* 20*-Lipschitz.*

Proof. ☐

Theorem 527. $g_1 : [0,1] \to [0,1]$ *is* 20*-Lipschitz.*

Proof. ☐

Definition 528. The parameters of a one-dimensional ReLU block with w units of input weights u_i, thresholds t_i, output weights v_i and bias c: $W = (u_i)_i$, $b = (-t_i)_i$, $V = (v_i)_i$ (as a row), so that $V\,\mathrm{relu}(Wx+b)+c = \sum_i v_i\,\mathrm{relu}(u_i x - t_i) + c$.

Theorem 529. $f_\vartheta(x) = \Pi_K(\sum_i v_i \max\{u_i x - t_i, 0\} + c)$ *for the parameters 'def:relu-unit-param'.*

Proof. ☐

Definition 530. Input weights of the expand-and-reset blocks at width $4+w'$: $u = (1,1,1,1,0,\dots,0)$.

Definition 531. Thresholds of g_0: $t = (0, \eta, 1/4-\eta, 1/4, 0, \dots, 0)$.

Definition 532. Output weights of g_0 (the slope increments at the breakpoints): $v = (-12, 52/3, -76/3, 20, 0, \dots, 0)$.

Definition 533. Thresholds of g_1: $t = (3/4, 3/4+\eta, 1-\eta, 0, 0, \dots, 0)$.

Definition 534. Output weights of g_1: $v = (-20, 76/3, -52/3, 0, 0, \dots, 0)$.

Definition 535. The ReLU parameters of g_0 (bias $c = 3/8$).

Definition 536. The ReLU parameters of g_1 (bias $c = 5/8$).

Theorem 537. *On* $[0,1]$, $g_0(x) = -12\,\mathrm{relu}(x) + \frac{52}{3}\mathrm{relu}(x-\eta) - \frac{76}{3}\mathrm{relu}(x-1/4+\eta) + 20\,\mathrm{relu}(x-1/4) + 3/8$.

Proof. ☐

Theorem 538. *On* $[0,1]$, $g_1(x) = -20\,\mathrm{relu}(x-3/4) + \frac{76}{3}\mathrm{relu}(x-3/4-\eta) - \frac{52}{3}\mathrm{relu}(x-1+\eta) + 5/8$.

Proof. ☐

Theorem 539. *The unit sum of* g_0 *at width* $4+w'$ *(the padding units vanish).*

Proof. ☐

Theorem 540. *The unit sum of g_1 at width $4 + w'$.*

Proof. ☐

Theorem 541. *(g_0 **is a ReLU block of width** $4 + w'$**.**) With the clipping $\Pi_{[0,1]}$, $f_{\vartheta_0} = g_0$ on $[0, 1]$.*

Proof. ☐

Theorem 542. *(g_1 **is a ReLU block of width** $4 + w'$**.**) $f_{\vartheta_1} = g_1$ on $[0, 1]$.*

Proof. ☐

Theorem 543. *Weight norms of the representation of g_0: $\|W\|_{\mathrm{op}} = \|u\| = 2$, $\|b\| = \|t\| \le 2$, $\|V\|_{\mathrm{op}} = \|v\| \le 41$, $\|c\| = 3/8 \le 2$.*

Proof. ☐

Theorem 544. *Weight norms of the representation of g_1: $\|W\|_{\mathrm{op}} = 2$, $\|b\| \le 2$, $\|V\|_{\mathrm{op}} \le 41$, $\|c\| = 5/8 \le 2$.*

Proof. ☐

Definition 545. The two generators $f = (g_0, g_1)$.

Definition 546. The chambers $U_0 = [0, 1/4]$, $U_1 = [3/4, 1]$.

Definition 547. The coding cores $V_0 = [\eta, 1/4 - \eta]$, $V_1 = [3/4 + \eta, 1 - \eta]$.

Definition 548. The anchors $a_0 = 3/8$, $a_1 = 5/8$.

Definition 549. The marker $q = 1/2$.

Theorem 550. *$V_i \subseteq U_i$.*

Proof. ☐

Theorem 551. *(Disjoint chambers.) $U_0 \cap U_1 = \emptyset$.*

Proof. ☐

Theorem 552. *(Reset.) $g_i(x) = a_i$ for $x \notin U_i$: $g_0 = 3/8$ outside $[0, 1/4]$ and $g_1 = 5/8$ outside $[3/4, 1]$.*

Proof. ☐

Theorem 553. *The anchors lie outside both chambers.*

Proof. ☐

Theorem 554. *$g_i(A) \subseteq A$ for $A = \{a_0, a_1\}$ (each anchor is reset to a_i).*

Proof. ☐

Theorem 555. *(Coding cores.) $g_i(V_i) = [0, 1] \supseteq \{q\} \cup V_0 \cup V_1$: for $y \in [0, 1]$, $x = \eta + \frac{3}{16} y \in V_0$ has $g_0(x) = y$, and $x = 3/4 + \eta + \frac{3}{16} y \in V_1$ has $g_1(x) = y$.*

Proof. □

Theorem 556. *(Marker separation.)* $d(q, a_i) = 1/8$ *for* $i = 0, 1$.

Proof. □

Theorem 557. ***(The expand-and-reset maps satisfy 'cond:e2-pingpong'.)*** *With the chambers, cores, anchors and marker above and separation* $\alpha = 1/8$*:* $N^{\text{ext}}(B(k, \{g_0, g_1\}), d_\infty, \varepsilon) \geq 2^k$ *for all* k *and* $2\varepsilon < 1/8$*, and the words of each length are pairwise distinct.*

Proof. □

Theorem 558. *For* $\beta_W \geq 41$*,* $\beta \geq 2$*,* $\Lambda \geq 20$ *and width* $4 + w'$*,* $g_0, g_1 \in F_\Lambda$ *on* $K = [0, 1]$ *with the clipping retraction.*

Proof. □

Theorem 559. ***('prop:relu-regimes'(iii), lower bound.)*** *For* $m = 1$*,* $K = [0, 1]$ *with the clipping retraction, width* $w \geq 4$*,* $\Lambda \geq 20$*, and* $\beta_W \geq 41$*,* $\beta \geq 2$ *(at least the weight norms of the representations of the two expand-and-reset maps), the class* F_Λ *contains* g_0, g_1*, which satisfy 'cond:e2-pingpong' with two generators and separation* $1/8$*; hence*

$$N^{\text{ext}}(B(k, F_\Lambda), d_\infty, \varepsilon) \geq 2^k \quad \textit{for all } k \textit{ and all } \varepsilon < 1/16.$$

(The paper states $w \geq 5$*; the two maps have three interior breakpoints each, so width* 4 *suffices.)*

Proof. Write $w = 4 + w'$; 'lem:relu-er-pingpong' gives the bound for $\{g_0, g_1\}$, and $B(k, \{g_0, g_1\}) \subseteq B(k, F_\Lambda)$ by 'lem:relu-er-mem-class'. □

7.4 Chain-of-thought style symbolic computation (Appendix L)

Definition 560. The scratchpad space is $\mathcal{X} = \mathcal{A}^{\mathbb{N}} = \{x = (x_0, x_1, \dots) : x_j \in \mathcal{A}\}$, the space of infinite symbol sequences over the alphabet $\mathcal{A}$ (a type synonym of $\mathbb{N} \to \mathcal{A}$ carrying the parameter θ of the metric below).

Definition 561. For $x \neq y$ the first mismatch index is $n(x, y) := \min\{j \geq 0 : x_j \neq y_j\}$.

Theorem 562. $x_{n(x,y)} \neq y_{n(x,y)}$.

Proof. □

Theorem 563. $L \leq n(x, y)$ *if and only if* $x_j = y_j$ *for all* $j < L$.

Proof. □

Theorem 564. $x_j = y_j$ *for all* $j < n(x, y)$.

Proof. □

Theorem 565. $n(x, y) = n(y, x)$.

Proof. Both indices are characterized by the same prefix-agreement property. □

Definition 566. The ultrametric on $\mathcal{A}^{\mathbb{N}}$: $d_\theta(x, y) = \theta^{n(x,y)}$ for $x \neq y$ and $d_\theta(x, x) = 0$, for a fixed $\theta \in (0, 1)$.

Theorem 567. $0 < \theta$ *(as a real number).*

Proof. □

Theorem 568. $\theta < 1$ *(as a real number).*

Proof. □

Theorem 569. *For* $x \neq y$, $d_\theta(x, y) = \theta^{n(x,y)}$.

Proof. □

Theorem 570. $d_\theta(x, y) = d_\theta(y, x)$.

Proof. □

Theorem 571. *(Prefix characterization.)* $d_\theta(x, y) \leq \theta^L$ *if and only if* $x_j = y_j$ *for all* $j < L$.

Proof. If $x = y$ both sides hold. Otherwise $\theta^{n(x,y)} \leq \theta^L \iff L \leq n(x, y)$ since $0 < \theta < 1$, and $L \leq n(x, y)$ is the prefix-agreement property. □

Theorem 572. $d_\theta(x, y) \geq 0$.

Proof. □

Theorem 573. $d_\theta(x, y) \leq 1$*: the scratchpad space has diameter at most* 1.

Proof. The case $L = 0$ of the prefix characterization. □

Theorem 574. *(Ultrametric inequality.)* $d_\theta(x, z) \leq \max\{d_\theta(x, y), d_\theta(y, z)\}$.

Proof. If $x = y$ or $y = z$ this is trivial. Otherwise let $L = \min\{n(x, y), n(y, z)\}$; then x, y and y, z agree on the first L symbols, hence so do x, z, so $d_\theta(x, z) \leq \theta^L = \max\{\theta^{n(x,y)}, \theta^{n(y,z)}\}$. □

Theorem 575. $d_\theta(x, y) = 0$ *implies* $x = y$.

Proof. □

Definition 576. $(\mathcal{A}^{\mathbb{N}}, d_\theta)$ is a metric space (indeed an ultrametric space).

Theorem 577. *On* $\mathcal{A}^{\mathbb{N}}$ *the distance is* d_θ.

Proof. □

Theorem 578. *Two sequences are within distance* θ^L *if and only if they agree in the first* L *symbols:* $d_\theta(x, y) \leq \theta^L \iff \forall j < L,\ x_j = y_j$.

Proof. □

Theorem 579. $d_\theta(x, y) \leq 1$, *i.e.* $\mathrm{diam}(\mathcal{A}^{\mathbb{N}}) \leq 1$.

Proof. □

Theorem 580. $d_\theta(x, y) \leq 1$ *as an extended distance.*

Proof. □

Theorem 581. *If $x_0 \neq y_0$ then $d_\theta(x, y) = 1$.*

Proof. $x \neq y$ and $n(x, y) = 0$. □

Theorem 582. *If x and y do not agree on the first L symbols then $d_\theta(x, y) \geq \theta^{L-1}$.*

Proof. $x \neq y$ and $n(x, y) < L$, so $n(x, y) \leq L - 1$ and $\theta^{L-1} \leq \theta^{n(x,y)}$. □

Definition 583. The identity map $\mathcal{A}^{\mathbb{N}} \to (\mathcal{A}^{\mathbb{N}}, d_\theta)$ from the product of the discrete spaces.

Theorem 584. *For a finite alphabet, $(\mathcal{A}^{\mathbb{N}}, d_\theta)$ is compact. (The identity from the product of discrete spaces is continuous, since a d_θ-ball is a cylinder set.)*

Proof. Equip $\mathcal{A}$ with the discrete topology; $\mathcal{A}^{\mathbb{N}}$ with the product topology is compact (Tychonoff). The identity map to $(\mathcal{A}^{\mathbb{N}}, d_\theta)$ is continuous: given $\varepsilon > 0$ pick L with $\theta^L < \varepsilon$; the cylinder $\{y : y_j = x_j,\ j < L\}$ is a product neighbourhood of x contained in the ε-ball. The continuous image of a compact space is compact. □

Theorem 585. *Since* $\operatorname{diam}(\mathcal{A}^{\mathbb{N}}) \leq 1$ *and* $d_S \leq d_\infty$, *the empirical diameter of any class of self-maps of $\mathcal{A}^{\mathbb{N}}$ is at most 1: $D_k(S) \leq 1$.*

Proof. □

Definition 586. A write step prepends a symbol: $g_a(x) := (a, x_0, x_1, \dots)$ for $a \in \mathcal{A}$.

Definition 587. The append-only step family $F_{\mathrm{w}} := \{g_a : a \in \mathcal{A}\}$.

Theorem 588. $g_a(x)_0 = a$.

Proof. □

Theorem 589. $g_a(x)_{i+1} = x_i$.

Proof. □

Theorem 590. *$a \mapsto g_a$ is injective, so $|F_{\mathrm{w}}| = |\mathcal{A}| = m$.*

Proof. □

Theorem 591. *F_{w} is finite.*

Proof. □

Theorem 592. $|F_{\mathrm{w}}| = m = |\mathcal{A}|$.

Proof. □

Theorem 593. *Each write step is θ-Lipschitz: $d_\theta(g_a(x), g_a(y)) \leq \theta\, d_\theta(x, y)$ (in fact with equality). Hence F_{w} is a contractive hidden-layer class with* $\operatorname{lip}(g_a) = \theta < 1$.

Proof. If $x = y$ there is nothing to prove. Otherwise $g_a(x)$ and $g_a(y)$ agree on the first $n(x, y)+1$ symbols, so $d_\theta(g_a(x), g_a(y)) \leq \theta^{n(x,y)+1} = \theta\, d_\theta(x, y)$. □

Theorem 594. *Each write step is non-expanding (1-Lipschitz).*

Proof. □

Theorem 595. *(Saturation via P1.) For a finite alphabet and every $\varepsilon > 0$, $N^{\text{ext}}(B(k, F_{\text{w}}), d_\infty, \varepsilon) \le N^{\text{ext}}(\overline{\langle F_{\text{w}} \rangle}, d_\infty, \varepsilon) < \infty$ uniformly in k.*

Proof. ☐

Theorem 596. *A word of ℓ write steps determines the first ℓ symbols of its output independently of the input: for $w \in F_{\text{w}}^{\ell}$ and all x, y, $w(x)_j = w(y)_j$ for $j < \ell$.*

Proof. Induction on ℓ: for $w = g_a \circ w'$, the symbol 0 is a for both inputs and the symbols $1, \dots, \ell$ are the first $\ell - 1$ symbols of $w'(x)$, $w'(y)$. ☐

Definition 597. The resolution length $\ell(\varepsilon) := \lceil \log(1/\varepsilon)/\log(1/\theta) \rceil$ (a natural number; 0 for $\varepsilon \ge 1$).

Theorem 598. $\log(1/\theta) > 0$.

Proof. ☐

Theorem 599. *$\theta^{\ell(\varepsilon)} \le \varepsilon$ for every $\varepsilon > 0$.*

Proof. $\ell(\varepsilon) \ge \log(1/\varepsilon)/\log(1/\theta)$, so $\log(1/\varepsilon) \le \log((1/\theta)^{\ell(\varepsilon)})$ and $1/\varepsilon \le 1/\theta^{\ell(\varepsilon)}$. ☐

Theorem 600. *For $0 < \varepsilon \le 1$, $\ell(\varepsilon) \le \log(1/\varepsilon)/\log(1/\theta) + 1$.*

Proof. $\lceil a \rceil < a + 1$ for $a \ge 0$; here $a \ge 0$ since $\varepsilon \le 1$. ☐

Theorem 601. *(The short words cover the word ball.) For every $\varepsilon > 0$ and k, the word ball $B(\min\{k, \ell(\varepsilon)\}, F_{\text{w}}) \subseteq B(k, F_{\text{w}})$ is an ε-cover of $B(k, F_{\text{w}})$ in d_∞: a word g_u with $|u| > \ell$ is within $\theta^\ell \le \varepsilon$ of the word $g_{u'}$ formed by its last ℓ letters.*

Proof. Let $f \in F_{\text{w}}^{l}$ with $l \le k$. If $l \le \ell(\varepsilon)$ then f itself lies in the cover. Otherwise factor $f = w_2 \circ w_1$ with $w_2 \in F_{\text{w}}^{\ell(\varepsilon)}$ (the last $\ell(\varepsilon)$ letters); $w_2(w_1(x))$ and $w_2(x)$ agree on the first $\ell(\varepsilon)$ symbols, so $d_\infty(f, w_2) \le \theta^{\ell(\varepsilon)} \le \varepsilon$. ☐

Theorem 602. ***(Saturation for append-only steps.)*** *Let $|\mathcal{A}| = m \ge 2$. For $\varepsilon > 0$ let $\ell(\varepsilon) := \lceil \log(1/\varepsilon)/\log(1/\theta) \rceil$. Then for all $k \ge 0$,*

$$N^{\text{ext}}\big(B(k, F_{\text{w}}), d_\infty, \varepsilon\big) \le m^{\min\{k, \ell(\varepsilon)\}+1}.$$

Proof. The cover of ‘lem:cot-append-cover‘ has at most $\sum_{j \le \min\{k, \ell(\varepsilon)\}} m^j \le m^{\min\{k, \ell(\varepsilon)\}+1}$ elements (‘lem:wordball-card‘). ☐

Theorem 603. *The same bound for the internal covering number in d_∞ and in d_S: $N(B(k, F_{\text{w}}), d_S, \varepsilon) \le N(B(k, F_{\text{w}}), d_\infty, \varepsilon) \le m^{\min\{k, \ell(\varepsilon)\}+1}$ (the cover consists of elements of $B(k, F_{\text{w}})$).*

Proof. ‘lem:covering-empSpace-le-unifMaps-internal‘, then the internal cover $B(\min\{k, \ell(\varepsilon)\}, F_{\text{w}}) \subseteq B(k, F_{\text{w}})$. ☐

Theorem 604. *For $0 < \varepsilon \le 1$ and $m \ge 2$, $\sqrt{\log N(B(k, F_{\text{w}}), d_S, \varepsilon)} \le \sqrt{\log m}\Big(\frac{\sqrt{\log(1/\varepsilon)}}{\sqrt{\log(1/\theta)}} + \sqrt{2}\Big)$, using $\ell(\varepsilon) + 1 \le \log(1/\varepsilon)/\log(1/\theta) + 2$ and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$.*

Proof. $\log N \le (\ell(\varepsilon) + 1)\log m \le (\log(1/\varepsilon)/\log(1/\theta) + 2)\log m$; take square roots. ☐

Theorem 605. ***(Saturated variance profile for append-only steps.)*** *Let $|\mathcal{A}| = m \geq 2$. For every sample S and every $k \geq 0$,*

$$\mathsf{V}_k(S) = \int_0^{D_k(S)} \sqrt{\log N(B(k, F_{\mathrm{w}}), d_S, \varepsilon)}\, d\varepsilon \leq \sqrt{\log m}\Big(\frac{\sqrt{\pi}}{2\sqrt{\log(1/\theta)}} + \sqrt{2}\Big) =: \mathsf{V}_\infty.$$

(Case (i) of ‘prop:profiles‘: the entropy integral does not depend on the depth k.)

Proof. Since $D_k(S) \leq 1$, dominate the integrand on $(0, 1]$ by ‘lem:cot-append-entropy-pointwise‘ and use $\int_0^1 \sqrt{\log(1/\varepsilon)}\, d\varepsilon = \sqrt{\pi}/2$ (‘lem:log-split‘). □

Definition 606. The alphabet $\mathcal{A} = [r] \cup \{\bullet\}$ with r active symbols and one padding symbol $\bullet$.

Definition 607. The shift $\sigma(x) = (x_1, x_2, \dots)$.

Definition 608. The constant sequence $\bar{a} = (a, a, \dots)$.

Definition 609. A guarded step f_a, $a \in [r]$, reads the current symbol: if it equals a the step consumes it, $f_a(x) = \sigma(x)$; otherwise it resets to the constant error state $\bar{a}$.

Definition 610. The branching step family $F_{\mathrm{b}} := \{f_a : a \in [r]\}$.

Theorem 611. $\sigma(g_a(y)) = y$.

Proof. □

Theorem 612. $\sigma(\bar{a}) = \bar{a}$.

Proof. □

Theorem 613. *If $x_0 \neq a$ then $f_a(x) = \bar{a}$.*

Proof. □

Theorem 614. *If $x_0 = a$ then $f_a(x) = \sigma(x)$.*

Proof. □

Theorem 615. *$f_a(\bar{a}) = \bar{a}$ and $f_a(\bar{b}) = \bar{a}$ for $b \neq a$: the anchors are mapped to anchors.*

Proof. □

Theorem 616. *$a \mapsto f_a$ is injective, so $|F_{\mathrm{b}}| = r$.*

Proof. $f_a(\bar{a}) = \bar{a}$ while $f_b(\bar{a}) = \bar{b}$. □

Theorem 617. $|F_{\mathrm{b}}| = r$.

Proof. □

Theorem 618. ***(The branching family is a ping–pong family, ‘ex:e2-pingpong-subshift‘.)*** *With chambers $U_a = V_a = \{x : x_0 = a\}$, anchors $\bar{a}$, marker $\bar{\bullet}$ and $\alpha = 1$, F_{b} satisfies the hypotheses of ‘cond:e2-pingpong‘ (the chambers are disjoint; they are in fact 1-separated, which the paper's form of E2 requires). Consequently, for $r \geq 2$, $N^{\mathrm{ext}}(B(k, F_{\mathrm{b}}), d_\infty, \varepsilon) \geq r^k$ for every k and every $\varepsilon < 1/2$, and $u \mapsto f_u$ is injective on $[r]^k$.*

Proof. Verify the four ping–pong conditions: (1) sequences in different chambers differ at index 0, so the chambers are disjoint; (2) $y = f_a(g_a(y))$ with $g_a(y) \in V_a$; (3) the reset is the definition of f_a off U_a, and anchors go to anchors ('lem:cot-guarded-const'); (4) $\bar{\bullet}_0 = \bullet \neq a = \bar{a}_0$, so $d(\bar{\bullet}, \bar{a}) = 1$. ☐

Theorem 619. ***(Exponential growth for branching steps.)*** *For $r \geq 2$, every $k \geq 0$ and every $\varepsilon < 1/2$,*

$$r^k \leq N^{\text{ext}}(B(k, F_{\text{b}}), d_\infty, \varepsilon) \leq r^{k+1}.$$

The lower bound is the ping–pong bound 'cond:e2-pingpong'; the upper bound holds because F_{b} is finite, so $|B(k, F_{\text{b}})| \leq \sum_{j \leq k} r^j \leq r^{k+1}$.

Proof. ☐

Theorem 620. *For $r \geq 2$, every sample S and every k, $\mathsf{V}_k(S) \leq \sqrt{(k+1)\log r}$ (case (iii) of 'prop:profiles', with $\overline{D} = \operatorname{diam}(\mathcal{A}^{\mathbb{N}}) = 1$).*

Proof. 'prop:profiles-finite' with $|B(k, F_{\text{b}})| \leq r^{k+1}$ and $D_k(S) \leq 1$. ☐

Theorem 621. *$f_{u \cdot v} = f_v \circ f_u$ (the first letter acts first).*

Proof. ☐

Theorem 622. *Every $g \in B(k, \{f_1, \dots, f_r\})$ is f_u for a word $u \in [r]^{\leq k}$.*

Proof. ☐

Definition 623. The success cylinder of a program $u = (u_1, \dots, u_j)$: the set $[u]$ of inputs whose first j symbols are $u_1, \dots, u_j$. In Lean, membership is defined by recursion on u: $x \in [\,]$ always, and $x \in [a :: u]$ iff $x_0 = a$ and $\sigma(x) \in [u]$.

Definition 624. Membership in a cylinder is decidable (it is a finite conjunction of symbol comparisons).

Theorem 625. *$\sigma^j(x) = (x_{i+j})_{i \geq 0}$.*

Proof. ☐

Theorem 626. *On a constant state, $f_u(\bar{b}) = \bar{u}_j$ where u_j is the last letter of u (and $\bar{b}$ if u is empty): each guard either keeps the constant state or replaces it by its own.*

Proof. ☐

Theorem 627. *(Closed form, success.) If $x \in [u]$ then $f_u(x) = \sigma^{|u|}(x)$.*

Proof. ☐

Theorem 628. *(Closed form, failure.) If $x \notin [u]$ then $f_u(x) = \bar{u}_j$, the constant state of the last letter of u: once a guard fails the scratchpad is the constant state of the failed symbol, and each later guard either keeps it or replaces it by its own constant state.*

Proof. ☐

Theorem 629. ***(Closed form of the programs.)*** *For a nonempty program $u = (u_1, \dots, u_j)$ and $f_u := f_{u_j} \circ \cdots \circ f_{u_1}$,*

$$f_u(x) = \sigma^j(x) \quad \textit{if } x \in [u], \qquad f_u(x) = \bar{u}_j \quad \textit{otherwise.}$$

Proof. □

Theorem 630. *(Disjointness of the cylinders.) Programs of the same length with a common successful input coincide: if $x \in [u] \cap [v]$ and $|u| = |v|$ then $u = v$.*

Proof. □

Theorem 631. *(The cylinders of length j are pairwise disjoint.) For every sample $S = (X_1, \dots, X_n)$ and $j \geq 0$, $\sum_{|u|=j} \#\{i : X_i \in [u]\} \leq n$, i.e. $\sum_{|u|=j} \hat{P}_n([u]) \leq 1$: each X_i lies in at most one cylinder of length j.*

Proof. □

Theorem 632. ***(Programs are close to constants on the sample.)*** *For every program u (with last letter u_j, or any b if u is empty),*

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

since $f_u = \bar{u}_j$ off $[u]$ and $\operatorname{diam}(\mathcal{A}^{\mathbb{N}}) = 1$.

Proof. □

Theorem 633. *If $\#\{i : X_i \in [u]\} \leq \varepsilon^2 n$ then $d_S(f_u, \bar{u}_j) \leq \varepsilon$.*

Proof. □

Theorem 634. ***(Empirical saturation for branching steps.)*** *For every sample $S = (X_1, \dots, X_n)$, every $k \geq 0$ and every $\varepsilon > 0$ (the paper states $\varepsilon \in (0, 1]$; the bound holds for all $\varepsilon > 0$),*

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

Centres: the identity, the r constant maps $\bar{a}$, and for each $1 \leq j \leq k$ the programs u of length j with $\hat{P}_n([u]) > \varepsilon^2$ (fewer than ε^{-2} of them, at most n, at most r^j); every other program of length j is within d_S-distance ε of the constant map $\bar{u}_j$.

Proof. The heavy programs of length j.

Counting the heavy programs: at most r^j, at most n, at most $\lfloor \varepsilon^{-2} \rfloor$.

The cover. □

Theorem 635. $1 + r + \sum_{j=1}^{k} \min\{r^j, n, \lfloor \varepsilon^{-2} \rfloor\} \leq 1 + r + k \min\{n, \lfloor \varepsilon^{-2} \rfloor\}$.

Proof. □

Theorem 636. ***(Root-logarithmic empirical entropy integral for branching steps.)*** *For every sample S and every k,*

$$\mathsf{V}_k(S) \leq \sqrt{\log(k+1)} + \sqrt{\log(1+r)} + \sqrt{2\log 2} + \sqrt{\pi/2}$$

(the paper's bound without the term $\sqrt{2\log 2}$: this additive constant is the price of comparing the internal covering number in d_S, which defines $\mathsf{V}_k(S)$, with the external one at half the scale, $N(B, d_S, \varepsilon) \leq N^{\text{ext}}(B, d_S, \varepsilon/2) \leq 1 + r + 4k\varepsilon^{-2}$).

Proof. For $0 < \varepsilon \le 1$: $N(B, d_S, \varepsilon) \le N^{\mathrm{ext}}(B, d_S, \varepsilon/2) \le 1 + r + k\lfloor 4\varepsilon^{-2} \rfloor \le (1+r)(k+1)(2/\varepsilon)^2$, so $\sqrt{\log N} \le \sqrt{\log(k+1)} + \sqrt{\log(1+r)} + \sqrt{2\log 2} + \sqrt{2}\sqrt{\log(1/\varepsilon)}$; integrate over $(0, 1]$ (recall $D_k(S) \le 1$) with $\int_0^1 \sqrt{\log(1/\varepsilon)}\, d\varepsilon = \sqrt{\pi}/2$. □

Theorem 637. ***(Empirical saturation for branching steps, 'lem:cot-branch-sample'.)*** *For every sample $S = (X_1, \dots, X_n)$, every $k \ge 0$ and every $\varepsilon > 0$ (the paper states $\varepsilon \in (0, 1]$),*

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

and consequently $\mathsf{V}_k(S) \le \sqrt{\log(k+1)} + \sqrt{\log(1+r)} + \sqrt{2\log 2} + \sqrt{\pi/2}$, the root-logarithmic profile, for every sample and without any assumption on the input distribution.

Proof. □

Definition 638. For $L \ge 1$ the window feature map $\Phi_L : \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^{\mathcal{A}^L}$ is the one-hot encoding of the window $(x_0, \dots, x_{L-1})$.

Theorem 639. $\|\Phi_L(x)\| = 1$ *(in particular $\|\Phi_L\| \le 1$).*

Proof. □

Theorem 640. *If x, y agree on the first L symbols then $\Phi_L(x) = \Phi_L(y)$.*

Proof. □

Theorem 641. *Distinct one-hot vectors are at Euclidean distance $\sqrt{2}$: $\|e_u - e_v\| = \sqrt{2}$ for $u \ne v$.*

Proof. $\|e_u - e_v\|^2 = \|e_u\|^2 - 2\langle e_u, e_v \rangle + \|e_v\|^2 = 1 - 0 + 1$. □

Theorem 642. ***(Window output features.)*** *Φ_L is $\sqrt{2}\,\theta^{1-L}$-Lipschitz with respect to d_θ, and $\|\Phi_L\| \le 1$. (If $d_\theta(x, y) \le \theta^L$ the first L symbols agree and $\Phi_L(x) = \Phi_L(y)$; otherwise $d_\theta(x, y) \ge \theta^{L-1}$ and $\|\Phi_L(x) - \Phi_L(y)\| = \sqrt{2} \le \sqrt{2}\,\theta^{1-L} d_\theta(x, y)$.)*

Proof. Case analysis on whether the windows agree. □

Theorem 643. *Consequently (via 'prop:hilbert-sg') the sub-Gaussian increment condition holds for the linear readout class $H_L = \{x \mapsto \langle w, \Phi_L(x) \rangle : \|w\| \le 1\}$ and every hidden-layer class on $\mathcal{A}^{\mathbb{N}}$, with $A_H = 1$ and $L = \sqrt{2}\,\theta^{1-L}$.*

Proof. □

Definition 644. The depth-independent entropy integral of 'lem:cot-append-profile': $\mathsf{V}_\infty(m, \theta) := \sqrt{\log m}\Big(\frac{\sqrt{\pi}}{2\sqrt{\log(1/\theta)}} + \sqrt{2}\Big)$.

Theorem 645. $\mathsf{V}_\infty(m, \theta) \ge 0$.

Proof. □

Definition 646. The teacher–student target class of arbitrarily long programs read out linearly: $\mathcal{C} := \overline{H_R(\Phi) \circ \langle F_{\mathrm{w}} \rangle}^{d_\infty}$, the uniform closure of the readouts of all finite compositions of write steps.

Theorem 647. *Every element of $B(k, F_{\mathrm{w}})$ and of $\langle F_{\mathrm{w}} \rangle$ is 1-Lipschitz, hence continuous.*

Proof. □

Theorem 648. ***(Measurability.)*** *For a continuous feature map Φ (with the Borel σ-algebra on $\mathcal{A}^{\mathbb{N}}$), every element of $\mathcal{H}_k = H_R(\Phi) \circ B(k, F_{\mathrm{w}})$ is measurable.*

Proof. □

Theorem 649. *Every element of the target class $\mathcal{C}$ is measurable (a uniform limit of continuous functions).*

Proof. □

Theorem 650. *$\mathcal{C} \neq \emptyset$ (it contains the zero readout) when $R \geq 0$.*

Proof. □

Theorem 651. ***(Pointwise boundedness.)*** *If $\|\Phi\| \leq M_\Phi$ and $R \geq 0$ then $|g(x)| \leq RM_\Phi$ for every $g \in \mathcal{H}_k$.*

Proof. □

Theorem 652. ***(Sup-norm separability.)*** *For a separable feature space, a Lipschitz feature map Φ with $\|\Phi\| \leq M_\Phi$ and $R \geq 0$, the class $\mathcal{H}_k = H_R(\Phi) \circ B(k, F_{\mathrm{w}})$ is sup-norm separable ($B(k, F_{\mathrm{w}})$ is finite and the ball of radius R has a countable dense subset).*

Proof. □

Theorem 653. ***(Rademacher complexity of the readout class, 'lem:cot-output'.)*** *If $\|\Phi\| \leq M_\Phi$ then $\hat{\mathfrak{R}}_S(H_R(\Phi)) \leq RM_\Phi/\sqrt{n}$ for every sample S (for the window features, $M_\Phi = 1$).*

Proof. □

Theorem 654. ***(Estimation term for append-only steps.)*** *Let $|\mathcal{A}| = m \geq 2$, Φ be L_Φ-Lipschitz with $\|\Phi\| \leq M_\Phi$ and $R > 0$. Then for every sample S of size $n \geq 1$ and every depth k,*

$$\hat{\mathfrak{R}}_S(\mathcal{H}_k) \leq \frac{RM_\Phi}{\sqrt{n}} + \frac{12 \cdot 1 \cdot RL_\Phi}{\sqrt{n}} \mathsf{V}_\infty(m, \theta)$$

('thm:hidden-decomp-depth' with A_H $= 1$, 'lem:cot-append-profile' and 'lem:cot-readout-rademacher').

Proof. □

Theorem 655. ***(Bias for append-only steps.)*** *Let Φ be L_Φ-Lipschitz and $R \geq 0$. Against the teacher–student class $\mathcal{C} = \overline{H_R(\Phi) \circ \langle F_{\mathrm{w}} \rangle}^{d_\infty}$,*

$$\mathrm{bias}(k) = \varepsilon_{\mathrm{model}}(k) \leq \beta_\ell \, RL_\Phi \, \theta^k :$$

for a teacher $c = h_w \circ g_u \circ g_v$ with $|u| = k$, $|c(x) - h_w(g_u(x))| \leq RL_\Phi \, d_\theta(g_u(g_v x), g_u x) \leq RL_\Phi \theta^k$ since $\mathrm{lip}(g_u) = \theta^k$ and $\mathrm{diam} = 1$, and the same bound holds on the uniform closure; conclude with 'lem:approx-transfer'.

Proof. □

Theorem 656. ***(Append-only scratchpad, rigorous form.)*** *Let $|\mathcal{A}| = m \geq 2$, let $\Phi : \mathcal{A}^{\mathbb{N}} \to \mathcal{H}$ be an L_Φ-Lipschitz feature map into a separable Hilbert space with $\|\Phi\| \leq M_\Phi$ (e.g. the window features Φ_L), $R > 0$, $H = H_R(\Phi)$, $F = F_{\mathrm{w}}$, and let the target class be $\mathcal{C} = \overline{H \circ \langle F_{\mathrm{w}} \rangle}^{d_\infty}$. For a measurable loss $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$, β_ℓ-Lipschitz in its first argument, $n \geq 1$, $\eta \geq 0$ and $\delta \in (0, 1)$: with probability at least $1 - \delta$ over $\mathcal{D} \sim P^{\otimes n}$, every η-empirical minimizer $\hat{h} \in \mathcal{H}_k = H \circ B(k, F_{\mathrm{w}})$ satisfies*

$$L[\hat{h}] - \inf_{\mathcal{C}} L \leq \beta_\ell R L_\Phi\, \theta^k + \eta + \mathrm{dev}_{\ell,n,\delta}\Big(\frac{RM_\Phi}{\sqrt{n}} + \frac{12RL_\Phi}{\sqrt{n}} \mathsf{V}_\infty(m, \theta)\Big),$$

where $\mathrm{dev}_{\ell,n,\delta}(B) = 4\beta_\ell B + 6b\sqrt{2\log(4/\delta)/n}$ is the estimation-plus-deviation term of 'thm:bv' ('def:bv-dev'). This is the EL regime with $\alpha = \log(1/\theta)$ and saturated variance. (The implementation map is the identity, $\varepsilon_{\mathrm{imp}} = 0$.)

Proof. □

Theorem 657. ***(EL balancing with saturated variance.)*** *For $n \geq 1$ the depth $k^* := \lceil \log n/(2\log(1/\theta)) \rceil = \frac{\log n}{2\log(1/\theta)} + O(1)$ satisfies $\theta^{k^*} \leq n^{-1/2}$, so the bias term matches the saturated variance $n^{-1/2}$.*

Proof. □

Theorem 658. ***(Balanced value $O(n^{-1/2})$.)*** *Under the hypotheses of 'prop:cot-append', at the depth $k^* = \lceil \log n/(2\log(1/\theta)) \rceil$ the bound reads*

$$L[\hat{h}] - \inf_{\mathcal{C}} L \leq \frac{\beta_\ell R L_\Phi}{\sqrt{n}} + \eta + \mathrm{dev}_{\ell,n,\delta}\Big(\frac{RM_\Phi + 12RL_\Phi \mathsf{V}_\infty(m, \theta)}{\sqrt{n}}\Big)$$

with probability at least $1 - \delta$: every term is of order $n^{-1/2}$ (up to the $\sqrt{\log(1/\delta)}$ factor of the deviation term), so the balanced value is $O(n^{-1/2})$.

Proof. □

7.5 Unrolled fixed-point iterations and ODE solvers (Appendix M)

Theorem 659. *(Saturated profile from total boundedness.) Let $\mathcal{X}$ be compact, let $G \subseteq \mathcal{X}^{\mathcal{X}}$ be totally bounded in d_∞ and $A_k \subseteq G$ for all k. With $N_\infty(\varepsilon) := N(\overline{G}, d_\infty, \varepsilon/2) < \infty$ and $\overline{D} := \mathrm{diam}(\mathcal{X})$, if $\int_0^{\overline{D}} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ then $\mathsf{V}(\mathrm{diam}_S(A_k), A_k) \leq \int_0^{\overline{D}} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon$ for every sample S and every k (case (i) of 'prop:profiles').*

Proof. 'prop:profiles-i' with $N(A_k, d_S, \varepsilon) \leq N(A_k, d_\infty, \varepsilon) \leq N(\overline{G}, d_\infty, \varepsilon/2)$ ('lem:covering-empSpace-le-unifMaps-internal' and Mathlib's 'coveringNumber_subset_le'), which is finite since $\overline{G}$ is totally bounded, and $D_k(S) \leq \mathrm{diam}(\mathcal{X})$. □

Theorem 660. *(Integrability of the entropy integrand from total boundedness.) Under the hypotheses of 'lem:ode-profile-saturation-generic', the entropy integrand $\varepsilon \mapsto \sqrt{\log N(A_k, d_S, \varepsilon)}$ is interval-integrable on $[0, \mathrm{diam}_S(A_k)]$ (it is dominated by the integrable majorant).*

Proof. □

Definition 661. A map $\Pi_K : \mathbb{R}^d \to \mathbb{R}^d$ is a (Euclidean) projection onto K if $\Pi_K(x) \in K$ for all x, $\Pi_K(x) = x$ for $x \in K$, and Π_K is 1-Lipschitz. (For a nonempty closed convex K the nearest-point map has these properties; Mathlib provides only the existence of nearest points, so we take the projection and its properties as hypotheses.)

Definition 662. The projected step of size h along a vector field $s : K \to \mathbb{R}^d$: $T_s(x) := \Pi_K(x + h\, s(x))$, a self-map of K.

Theorem 663. *Since Π_K is 1-Lipschitz, $\|T_s(x) - T_s(y)\| \le \|(x-y) + h\,(s(x) - s(y))\|$.*

Proof. □

Theorem 664. *A step of size 0 is the identity: $T_{s,0} = \mathrm{id}$.*

Proof. □

Definition 665. A vector field $s : K \to \mathbb{R}^d$ is *μ-strongly monotone and Λ-co-coercive* (the properties of $s = \nabla\phi$ for ϕ μ-strongly concave and Λ-smooth) if for all $x, y \in K$

$$\langle x-y, s(x)-s(y)\rangle \le -\mu\|x-y\|^2, \qquad \langle x-y, s(x)-s(y)\rangle \le -\frac{\mu\Lambda}{\mu+\Lambda}\|x-y\|^2 - \frac{1}{\mu+\Lambda}\|s(x)-s(y)\|^2.$$

(The first is the gradient characterization of strong concavity; the second is the standard consequence of strong concavity and smoothness. We take both as the definition.)

Theorem 666. ***(Contraction of projected gradient steps.)*** *Let $0 < \mu \le \Lambda$, let s be μ-strongly monotone and Λ-co-coercive, and let $0 < h \le 2/(\mu+\Lambda)$. Then $T_s = \Pi_K(\cdot + hs(\cdot))$ is λ-Lipschitz on K with $\lambda := 1 - h\mu \in [0,1)$.*

Proof. Write $u = x - y$, $v = s(x) - s(y)$, $a = \|u\|$, $b = \|v\|$. Strong monotonicity and Cauchy–Schwarz give $\mu a \le b$. Then $\|u+hv\|^2 = a^2 + 2h\langle u, v\rangle + h^2b^2 \le a^2(1 - \frac{2h\mu\Lambda}{\mu+\Lambda}) + (h^2 - \frac{2h}{\mu+\Lambda})b^2$ and, since $h^2 - 2h/(\mu+\Lambda) \le 0$ and $b^2 \ge \mu^2 a^2$, this is at most $a^2(1 - \frac{2h\mu\Lambda}{\mu+\Lambda}) + (h^2 - \frac{2h}{\mu+\Lambda})\mu^2 a^2 = (1-h\mu)^2 a^2$. Finally Π_K is 1-Lipschitz. □

Theorem 667. *(Stability of one step.) If s is Λ_s-Lipschitz and $h \ge 0$ then $x \mapsto \Pi_K(x + hs(x))$ is $(1 + h\Lambda_s)$-Lipschitz.*

Proof. $\|(x-y) + h(s(x) - s(y))\| \le \|x-y\| + h\Lambda_s\|x-y\|$. □

Definition 668. The fixed-point refinement class $F_{\mathrm{fp}} := \{T_s : x \mapsto \Pi_K(x + hs(x)) : s \in \mathcal{S}\}$ for a class $\mathcal{S}$ of vector fields on K and a fixed step size h.

Theorem 669. *Every $T \in F_{\mathrm{fp}}$ is $(1-h\mu)$-Lipschitz, under the hypotheses of 'lem:fp-contraction' for every $s \in \mathcal{S}$.*

Proof. □

Theorem 670. *Every $T \in F_{\mathrm{fp}}$ is non-expanding:* $\mathrm{lip}\, F_{\mathrm{fp}} \le 1 - h\mu \le 1$.

Proof. □

Theorem 671. ***(Saturation for fixed-point refinement, 'cond:p1'.)*** *If K is compact then for every $\varepsilon > 0$ and every k, $N^{\mathrm{ext}}(B(k, F_{\mathrm{fp}}), d_\infty, \varepsilon) \le N^{\mathrm{ext}}(\overline{\langle F_{\mathrm{fp}}\rangle}, d_\infty, \varepsilon) < \infty$; in particular $\sup_k N^{\mathrm{ext}}(B(k, F_{\mathrm{fp}}), d_\infty, \varepsilon) < \infty$.*

Proof. 'cond:p1-2c' on the compact state space K with non-expanding generators. □

Theorem 672. *(Explicit entropy bound via 'cond:p1-ucont'.) Assume moreover $h\mu < 1$ and $K \neq \emptyset$, and let $\lambda = 1 - h\mu$, $m(\varepsilon) = \lceil \log_{1/\lambda}(2D_K/\varepsilon) \rceil$ with $D_K = \operatorname{diam}(K)$. Then for every $\varepsilon > 0$ and every k (generalizing the paper, which assumes $k \geq m(\varepsilon)$; see 'cond:p1-ucont'),*

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

(In Lean the absorbing set is K itself, with $L = 0$.)

Proof. 'cond:p1-ucont' with $c = 1 - h\mu \in (0, 1)$, invariant set $A = K$ and bounded absorbing set K (words of length $L = 0$). ☐

Theorem 673. *(Saturated variance profile for fixed-point refinement.) Under the hypotheses of 'lem:fp-saturation', with $N_\infty(\varepsilon) := N(\overline{\langle F_{\mathrm{fp}} \rangle}, d_\infty, \varepsilon/2) < \infty$ and $D_K = \operatorname{diam}(K)$: if $\int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ then $\mathsf{V}_k(S) \leq \mathsf{V}_\infty := \int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon$ for all k and all samples S (case (i) of 'prop:profiles').*

Proof. The non-expanding semigroup is equicontinuous, hence totally bounded in d_∞ ('thm:caa'); apply 'lem:ode-profile-saturation-generic'. ☐

Definition 674. An explicit Euler layer of step size h at time stamp τ for a time-dependent vector field $s : K \times [0, T] \to \mathbb{R}^d$: $T_{s,h,\tau}(x) := \Pi_K(x + h\, s(x, \tau))$.

Definition 675. The Euler layer class $F_T := \{T_{s,h,\tau} : s \in \mathcal{S}_T,\ h \in [0, h_0],\ \tau \in [0, T]\}$.

Theorem 676. *Each Euler layer with $h \geq 0$ and Λ_s-Lipschitz drift $s(\cdot, \tau)$ is $(1+h\Lambda_s)$-Lipschitz.*

Proof. ☐

Definition 677. A scheme of n steps with a single drift s, step sizes $h = (h_0, \dots, h_{n-1})$ and time stamps $\tau = (\tau_0, \dots, \tau_{n-1})$ is the composition $T_{s,h_{n-1},\tau_{n-1}} \circ \cdots \circ T_{s,h_0,\tau_0}$ (the empty scheme is id).

Theorem 678. *A scheme of n steps with drift $s \in \mathcal{S}_T$, step sizes in $[0, h_0]$ and time stamps in $[0, T]$ belongs to $B(n, F_T)$.*

Proof. ☐

Theorem 679. ***(Stability of schemes.)*** *If $s(\cdot, \tau)$ is Λ_s-Lipschitz for every τ and $h_i \geq 0$, then the scheme with steps $h_0, \dots, h_{n-1}$ is $\prod_i (1 + h_i\Lambda_s) \leq \exp(\Lambda_s \sum_i h_i)$-Lipschitz.*

Proof. ☐

Definition 680. The class $B_T(k)$ of schemes of at most k steps with a single drift $s \in \mathcal{S}_T$, step sizes $h_i \in [0, h_0]$, time stamps $\tau_i \in [0, T]$ and total time $\sum_i h_i \leq T$ (it contains id, the empty scheme). The consistency condition $\tau_i = \sum_{j<i} h_j$ of the paper is dropped; the bounds below hold for this larger class.

Definition 681. All schemes: $\bigcup_k B_T(k)$.

Theorem 682. *$B_T(k) \subseteq B_T(k + 1)$: the scheme classes are nested.*

Proof. ☐

Theorem 683. *$B_T(k) \subseteq B(k, F_T)$.*

Proof. ☐

Theorem 684. *Every scheme in $\bigcup_k B_T(k)$ is $e^{\Lambda_s T}$-Lipschitz on K.*

Proof. 'lem:ode-scheme-lipschitz' and $\sum_i h_i \le T$. ☐

Theorem 685. ***(Stability and saturation.)*** *Let K be compact and let every drift $s \in \mathcal{S}_T$ be Λ_s-Lipschitz in x. 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_∞ ('thm:caa'), and*

$$\sup_k N^{\text{ext}}(B_T(k), d_\infty, \varepsilon) \le N^{\text{ext}}\Big(\bigcup_k B_T(k), d_\infty, \varepsilon\Big) =: N_\infty(\varepsilon) < \infty \quad (\varepsilon > 0).$$

Proof. A uniformly Lipschitz family is (uniformly) equicontinuous; Arzelà–Ascoli ('thm:caa-of-equicontinuous') gives total boundedness in d_∞, monotonicity of the external covering number and 'lem:aa-external-covering-ne-top' the rest. ☐

Theorem 686. *(Saturated variance profile for Euler schemes.) Under the hypotheses of 'lem:ode-saturation', with $N_\infty(\varepsilon) := N(\overline{\bigcup_k B_T(k)}, d_\infty, \varepsilon/2) < \infty$ and $D_K = \operatorname{diam}(K)$: if $\int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ then $\mathsf{V}_k(S) \le \mathsf{V}_\infty := \int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon$ for all k and all samples S (case (i) of 'prop:profiles').*

Proof. 'lem:ode-profile-saturation-generic' with $G = \bigcup_k B_T(k)$. ☐

Definition 687. The identity feature map $K \hookrightarrow \mathbb{R}^d$, so that $H_R = \{x \mapsto \langle w, x\rangle : \|w\| \le R\}$ is the linear readout class $H_R(\Phi)$ with $\Phi = \mathrm{id}_K$.

Theorem 688. id_K *is* 1*-Lipschitz and continuous.*

Proof. ☐

Definition 689. The teacher–student target class of the fixed-point example: $\mathcal{C} := \overline{H_R \circ \langle F_{\text{fp}}\rangle}^{d_\infty}$, which contains the readouts $x \mapsto \langle w, x_s^*\rangle$ of the fixed points (limits of the refinement).

Theorem 690. *Under the step-size condition $0 < h \le 2/(\mu + \Lambda)$ with $0 < \mu \le \Lambda$, $0 \le 1 - h\mu$.*

Proof. ☐

Theorem 691. *Every element of $\langle F_{\text{fp}}\rangle$ is* 1*-Lipschitz, hence continuous.*

Proof. ☐

Theorem 692. *For compact K, $\langle F_{\text{fp}}\rangle$ is totally bounded in d_∞ (Arzelà–Ascoli for the non-expanding semigroup).*

Proof. ☐

Theorem 693. ***(Estimation term for fixed-point refinement.)*** *Let K be compact with $\|x\| \le M_K$ on K, $R > 0$, and assume $\mathsf{V}_\infty(F_{\text{fp}}) < \infty$ ('def:sat-integrable'). Then for every sample S of size $n \ge 1$ and every depth k,*

$$\hat{\mathfrak{R}}_S(H_R \circ B(k, F_{\text{fp}})) \le \frac{RM_K}{\sqrt{n}} + \frac{12 \cdot 1 \cdot R}{\sqrt{n}} \mathsf{V}_\infty(F_{\text{fp}})$$

('cor:var-profiles-p1' with A_H $= 1$, $L = R$, and 'lem:linear-readouts-rademacher').

Proof. ☐

Theorem 694. *Every element of $H_R \circ B(k, F_{\mathrm{fp}})$ is Borel measurable.*

Proof. ☐

Theorem 695. *Every element of the target class $\mathcal{C}$ is measurable.*

Proof. ☐

Theorem 696. *$\mathcal{C} \neq \emptyset$ for $R \geq 0$.*

Proof. ☐

Theorem 697. ***(Bias for fixed-point refinement.)*** *Let K be compact with $D_K = \mathrm{diam}(K)$ and $R \geq 0$. Against the teacher–student class $\mathcal{C} = \overline{H_R \circ \langle F_{\mathrm{fp}} \rangle}^{d_\infty}$,*

$$\mathrm{bias}(k) = \varepsilon_{\mathrm{model}}(k) \leq \beta_\ell \, R \, D_K \, (1 - h\mu)^k$$

by the truncation argument 'lem:truncation-bias' ($\mathrm{lip}(h) \leq R$, $\mathrm{lip}(u) \leq (1 - h\mu)^k$ for $u \in F_{\mathrm{fp}}^k$) and 'lem:approx-transfer'.

Proof. ☐

Theorem 698. ***(Fixed-point refinement, rigorous form.)*** *Let $K \subseteq \mathbb{R}^d$ (a separable Hilbert space) be compact with $\|x\| \leq M_K$ on K and $D_K = \mathrm{diam}(K)$, let F_{fp} be the projected gradient steps of 'lem:fp-contraction' with $0 < h \leq 2/(\mu + \Lambda)$, $H = H_R$ ($R > 0$), and let the target class be $\mathcal{C} = \overline{H_R \circ \langle F_{\mathrm{fp}} \rangle}^{d_\infty}$. Assume $\mathsf{V}_\infty(F_{\mathrm{fp}}) < \infty$. For a measurable loss $\ell : \mathbb{R} \times \mathcal{Y} \to [0, b]$, β_ℓ-Lipschitz in its first argument, $n \geq 1$, $\eta \geq 0$, $\delta \in (0, 1)$: with probability at least $1 - \delta$ over $\mathcal{D} \sim P^{\otimes n}$, every η-empirical minimizer $\hat{h} \in H_R \circ B(k, F_{\mathrm{fp}})$ satisfies*

$$L[\hat{h}] - \inf_{\mathcal{C}} L \leq \beta_\ell R D_K (1 - h\mu)^k + \eta + \mathrm{dev}_{\ell,n,\delta}\Big(\frac{R M_K}{\sqrt{n}} + \frac{12R}{\sqrt{n}} \mathsf{V}_\infty(F_{\mathrm{fp}}) \Big),$$

with $\mathrm{dev}_{\ell,n,\delta}(B) = 4\beta_\ell B + 6b\sqrt{2\log(4/\delta)/n}$ ('def:bv-dev'): the EL regime with $\alpha = \log(1/(1 - h\mu)) \geq h\mu$ and saturated variance.

Proof. ☐

Theorem 699. ***(EL balancing for fixed-point refinement.)*** *If $h\mu < 1$ and $n \geq 1$, the depth $k^* := \lceil \log n / (2 \log(1/(1 - h\mu))) \rceil = \frac{\log n}{2 \log(1/(1-h\mu))} + O(1) \leq \frac{\log n}{2h\mu} + O(1)$ satisfies $(1 - h\mu)^{k^*} \leq n^{-1/2}$.*

Proof. ☐

Theorem 700. ***(Balanced value $O(n^{-1/2})$.)*** *Under the hypotheses of 'prop:ode-fixedpoint' with $h\mu < 1$, at the depth $k^* = \lceil \log n / (2 \log(1/(1 - h\mu))) \rceil$,*

$$L[\hat{h}] - \inf_{\mathcal{C}} L \leq \frac{\beta_\ell R D_K}{\sqrt{n}} + \eta + \mathrm{dev}_{\ell,n,\delta}\Big(\frac{R M_K + 12R \, \mathsf{V}_\infty(F_{\mathrm{fp}})}{\sqrt{n}} \Big)$$

with probability at least $1 - \delta$: the balanced value is of order $n^{-1/2}$.

Proof. ☐

Definition 701. The explicit Euler iterates on the ambient space with constant step h from time 0: $y_0 = x$, $y_{i+1} = y_i + h\, s(y_i, ih)$.

Theorem 702. ***(One-step consistency of the Euler scheme.)*** *Let s be Λ_s-Lipschitz in x, Λ_τ-Lipschitz in τ and bounded by M_s, and let x solve $\dot{x}(t) = s(x(t), t)$ on $[t_0, t_0 + h]$ $(h \geq 0)$. Then*

$$\|x(t_0 + h) - x(t_0) - h\, s(x(t_0), t_0)\| \leq \frac{(\Lambda_s M_s + \Lambda_\tau)\, h^2}{2}.$$

(The function $g(t) = x(t) - x(t_0) - (t - t_0)s(x(t_0), t_0)$ has $\|g'(t)\| \leq \Lambda_s\|x(t) - x(t_0)\| + \Lambda_\tau(t - t_0) \leq (\Lambda_s M_s + \Lambda_\tau)(t - t_0)$, and the mean value inequality with the quadratic boundary $B(t) = (\Lambda_s M_s + \Lambda_\tau)(t - t_0)^2/2$ gives the claim.)

Proof. □

Theorem 703. ***(Global error of the explicit Euler scheme.)*** *Let s be Λ_s-Lipschitz in x, Λ_τ-Lipschitz in τ and bounded by M_s, let x solve $\dot{x}(t) = s(x(t), t)$ on $[0, T]$ $(T \geq 0)$, and let $y_0, \dots, y_k$ be the Euler iterates with $k \geq 1$ equal steps $h = T/k$ started at $y_0 = x(0)$. Then*

$$\|x(T) - y_k\| \leq C_E\, \frac{T^2}{k}, \qquad C_E := \frac{(\Lambda_s M_s + \Lambda_\tau)\, e^{\Lambda_s T}}{2}.$$

(Discrete Grönwall: $e_{i+1} \leq (1 + h\Lambda_s)e_i + ch^2/2$ with $c = \Lambda_s M_s + \Lambda_\tau$, hence $e_k \leq \frac{ch^2}{2} \sum_{j<k}(1 + h\Lambda_s)^j \leq \frac{ch^2}{2}\, k\, e^{\Lambda_s T}$.)

Proof. □

Definition 704. The equal-step scheme with k Euler steps of size $h = T/k$ and time stamps $\tau_i = ih$: $T_{s,h,\tau_{k-1}} \circ \cdots \circ T_{s,h,\tau_0}$.

Theorem 705. *For $s \in \mathcal{S}_T$, $T \geq 0$, $k \geq 1$ and $T/k \leq h_0$, the equal-step scheme with k steps belongs to $B_T(k)$.*

Proof. □

Theorem 706. ***(Inactive projection.)*** *Let $\tilde{s} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ extend the drift s from K, and suppose K is invariant under the Euler steps, $x + h\, s(x, \tau) \in K$ for $x \in K$. Then the projection is inactive and the scheme with k equal steps of size h coincides with the Euler iterates: $T_{s,h,(k-1)h} \circ \cdots \circ T_{s,h,0}(x_0) = y_k(x_0)$.*

Proof. □

Definition 707. The depth-independent entropy integral of 'lem:ode-saturation-profile': $\mathsf{V}_\infty := \int_0^{D_K} \sqrt{\log N(\overline{\bigcup_k B_T(k)}, d_\infty, \varepsilon/2)}\, d\varepsilon$.

Definition 708. The paper's condition $\int_0^{D_K} \sqrt{\log N_\infty(\varepsilon)}\, d\varepsilon < \infty$ for the scheme classes: the majorant is interval-integrable.

Theorem 709. ***(Estimation term for fixed-horizon schemes.)*** *Let K be compact with $\|x\| \leq M_K$ on K, every drift in $\mathcal{S}_T \neq \emptyset$ be Λ_s-Lipschitz in x, $T \geq 0$, $R > 0$, and assume the entropy condition of 'lem:ode-saturation-profile'. Then for every sample S of size $n \geq 1$ and every k,*

$$\hat{\mathfrak{R}}_S(H_R \circ B_T(k)) \leq \frac{RM_K}{\sqrt{n}} + \frac{12 \cdot 1 \cdot R}{\sqrt{n}}\, \mathsf{V}_\infty$$

('thm:hidden-decomp' on the scheme class, 'lem:ode-saturation-profile' and 'lem:linear-readouts-rademacher').

Proof. ∎

Theorem 710. *Every scheme in $\bigcup_k B_T(k)$ is Lipschitz, hence continuous.*

Proof. ∎

Theorem 711. *Every element of $H_R \circ B_T(k)$ is Borel measurable.*

Proof. ∎

Theorem 712. ***(Bias for fixed-horizon integration.)*** *Let the target class be the readouts of the endpoint map of the exact flow of $s^* \in \mathcal{S}_T$: $\mathcal{C} = \{x_0 \mapsto \langle w, \Phi_T(x_0)\rangle : \|w\| \le R\}$, where for every $x_0 \in K$ the flow $\Phi_T(x_0) = x(T)$ is given by a solution of $\dot{x} = \tilde{s}(x,t)$, $x(0) = x_0$, for an extension $\tilde{s}$ of s^* which is Λ_s-Lipschitz in x, Λ_τ-Lipschitz in τ and bounded by M_s. Assume the projection is inactive along the Euler steps of size T/k (K is invariant), $k \ge 1$ and $T/k \le h_0$. Then*

$$\mathrm{bias}(k) = \varepsilon_{\mathrm{model}}(k) \le \beta_\ell\, R\, C_E\, \frac{T^2}{k}, \qquad C_E = \frac{(\Lambda_s M_s + \Lambda_\tau)e^{\Lambda_s T}}{2},$$

by 'lem:ode-euler-error', 'lem:ode-scheme-eq-euler' and 'lem:approx-transfer'.

Proof. ∎

Theorem 713. ***(Fixed-horizon integration, rigorous form,*** $p = 1$***.)*** *Let $K \subseteq \mathbb{R}^d$ (a separable Hilbert space) be compact with $\|x\| \le M_K$ on K, let every drift in $\mathcal{S}_T$ be Λ_s-Lipschitz in x, $H = H_R$ ($R > 0$), and let the target class be the flow readouts $\mathcal{C} = \{\langle w, \Phi_T(\cdot)\rangle : \|w\| \le R\}$ of a teacher $s^* \in \mathcal{S}_T$ as in 'lem:ode-horizon-bias' (projection inactive, $k \ge 1$, $T/k \le h_0$). Assume the entropy condition $\mathsf{V}_\infty < \infty$ of 'lem:ode-saturation-profile'. For a measurable loss $\ell : \mathbb{R}\times\mathcal{Y} \to [0,b]$, β_ℓ-Lipschitz in its first argument, $n \ge 1$, $\eta \ge 0$, $\delta \in (0,1)$: with probability at least $1-\delta$ over $\mathcal{D} \sim P^{\otimes n}$, every η-empirical minimizer $\hat{h} \in H_R \circ B_T(k)$ satisfies*

$$L[\hat{h}] - \inf_{\mathcal{C}} L \le \beta_\ell R C_E\, \frac{T^2}{k} + \eta + \mathrm{dev}_{\ell,n,\delta}\Big(\frac{RM_K}{\sqrt{n}} + \frac{12R}{\sqrt{n}}\,\mathsf{V}_\infty\Big),$$

with $C_E = (\Lambda_s M_s + \Lambda_\tau)e^{\Lambda_s T}/2$ and $\mathrm{dev}_{\ell,n,\delta}(B) = 4\beta_\ell B + 6b\sqrt{2\log(4/\delta)/n}$ ('def:bv-dev'): the PL regime with $\beta = p = 1$ and saturated variance.

Proof. ∎

Theorem 714. ***(PL balancing with saturated variance,*** $p = 1$***.)*** *For $n \ge 1$ the depth $k^* := \lceil\sqrt{n}\rceil \ge 1$ satisfies $1/k^* \le n^{-1/2}$, so $k^* \asymp n^{1/2} = n^{1/(2p)}$ balances the bias k^{-1} against the saturated variance $n^{-1/2}$.*

Proof. ∎

Theorem 715. ***(Balanced value*** $O(n^{-1/2})$***.)*** *Under the hypotheses of 'prop:ode-horizon' with $k^* = \lceil\sqrt{n}\rceil$ (and $T/k^* \le h_0$, projection inactive at step size T/k^*),*

$$L[\hat{h}] - \inf_{\mathcal{C}} L \le \frac{\beta_\ell R C_E T^2}{\sqrt{n}} + \eta + \mathrm{dev}_{\ell,n,\delta}\Big(\frac{RM_K + 12R\,\mathsf{V}_\infty}{\sqrt{n}}\Big)$$

with probability at least $1-\delta$: the balanced value is of order $n^{-1/2}$.

Proof. ∎

Definition 716. The class of equal-step schemes of at most k steps, $E_T(k) := \{\Phi_{s,m} : s \in \mathcal{S}_T,\ 0 \le m \le k\}$, where $\Phi_{s,m} = T_{s,T/m,\tau_m} \circ \cdots \circ T_{s,T/m,\tau_1}$ with $\tau_i = (i-1)T/m$ is the m-step equal-step Euler scheme and $\Phi_{s,0} = \mathrm{id}$.

Definition 717. The space of drifts restricted to $K \times [0,T]$ with the sup-norm (pseudo-e)metric $\|s - s'\|_\infty := \sup_{x \in K,\ \tau \in [0,T]} \|s(x,\tau) - s'(x,\tau)\|$; in Lean it is Mathlib's $(K \times [0,T]) \to_{\mathrm{u}} E$.

Definition 718. $\|\cdot\|_\infty$ is a pseudo-emetric on the restricted drifts (Mathlib's instance on the uniform function space).

Definition 719. The restriction of a drift $s : K \times \mathbb{R} \to \mathbb{R}^d$ to $K \times [0,T]$, viewed in the sup-norm space.

Theorem 720. $\|s - s'\|_\infty = \sup_{x \in K,\ \tau \in [0,T]} \|s(x,\tau) - s'(x,\tau)\|$ *(as an extended distance).*

Proof. □

Theorem 721. $\|s(x,\tau) - s'(x,\tau)\| \le \|s - s'\|_\infty$ *for* $x \in K$ *and* $\tau \in [0,T]$.

Proof. □

Definition 722. The extension of a restricted drift $c : K \times [0,T] \to \mathbb{R}^d$ to $K \times \mathbb{R}$ by 0 outside $[0,T]$ (used to lift the centres of a cover of $\mathcal{S}_T$ to drifts).

Theorem 723. *Restricting the extension gives back the restricted drift.*

Proof. □

Theorem 724. *(Two drifts.) Since* Π_K *is* 1*-Lipschitz,* $\|T_s(x) - T_{s'}(y)\| \le \|(x-y) + h\,(s(x) - s'(y))\|$.

Proof. □

Theorem 725. ***(Stability with respect to the drift; discrete Grönwall.)*** *Let* $s(\cdot,\tau)$ *be* Λ_s*-Lipschitz for every* τ*, let* $h_i \ge 0$*, and suppose* $\|s(x,\tau_i) - s'(x,\tau_i)\| \le \delta$ *for all* $x \in K$ *and all stamps* τ_i *used by the scheme. Then for every* x*,*

$$\|y_n - y_n'\| \le \delta \Big(\sum_i h_i\Big) \exp\Big(\Lambda_s \sum_i h_i\Big),$$

where y_n, y_n' *are the schemes of* s *and* s' *with the same steps and stamps started at* x*. Indeed* $\|y_i - y_i'\| \le (1 + h_i\Lambda_s)\|y_{i-1} - y_{i-1}'\| + h_i\delta$ *and* $1 + h\Lambda_s \le e^{h\Lambda_s}$*.*

Proof. □

Theorem 726. *A scheme all of whose steps have size* 0 *is the identity.*

Proof. □

Theorem 727. *With horizon* $T = 0$ *every equal-step scheme is the identity.*

Proof. □

Theorem 728. *(Pointwise stability of equal-step schemes.) If* $s(\cdot,\tau)$ *is* Λ_s*-Lipschitz for every* τ*,* $T \ge 0$*, and* $\|s(x,\tau) - s'(x,\tau)\| \le \delta$ *for all* $x \in K$*,* $\tau \in [0,T]$*, then* $\|\Phi_{s,m}(x) - \Phi_{s',m}(x)\| \le T e^{\Lambda_s T}\delta$ *for all* $m \ge 0$ *and* x*.*

Proof. 'lem:ode-scheme-dist' with $h_i = T/m$, $\sum_i h_i = T$ and stamps $\tau_i = (i-1)T/m \in [0, T]$. $\square$

Theorem 729. ***(Stability of equal-step schemes in d_∞.)*** *If $s(\cdot, \tau)$ is Λ_s-Lipschitz for every τ and $T \geq 0$, then for every drift s' and every $m \geq 0$,*

$$d_\infty(\Phi_{s,m}, \Phi_{s',m}) \leq T e^{\Lambda_s T} \|s - s'\|_\infty, \qquad \|s - s'\|_\infty = \sup_{x \in K,\ \tau \in [0,T]} \|s(x, \tau) - s'(x, \tau)\|.$$

(Only the Lipschitz constant of s is used; s' may be any drift.)

Proof. If $\|s - s'\|_\infty = \infty$ the bound is trivial unless $T = 0$, when both schemes are the identity; otherwise apply 'lem:ode-equal-scheme-dist' with $\delta = \|s - s'\|_\infty$ and take the supremum over x. $\square$

Theorem 730. *(Covering the m-step schemes.) Under the hypotheses of 'lem:ode-equal-scheme-uniformdist' for every $s \in \mathcal{S}_T$, the image of an $\varepsilon e^{-\Lambda_s T}/T$-cover of $\mathcal{S}_T$ (in $\|\cdot\|_\infty$ on $K \times [0, T]$, centres extended by 0 outside $[0, T]$) under $s \mapsto \Phi_{s,m}$ is an ε-cover of $\{\Phi_{s,m} : s \in \mathcal{S}_T\}$; hence $N^{\text{ext}}(\{\Phi_{s,m} : s \in \mathcal{S}_T\}, d_\infty, \varepsilon) \leq N^{\text{ext}}(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon e^{-\Lambda_s T}/T)$.*

Proof. $\square$

Theorem 731. $E_T(k) \subseteq \{\text{id}\} \cup \bigcup_{m=1}^{k} \{\Phi_{s,m} : s \in \mathcal{S}_T\}$.

Proof. $\square$

Theorem 732. ***(Covering bound for equal-step schemes.)*** *If every drift $s \in \mathcal{S}_T$ is Λ_s-Lipschitz in x and $T \geq 0$, then for every $k \geq 0$ and $\varepsilon \geq 0$,*

$$N^{\text{ext}}(E_T(k), d_\infty, \varepsilon) \leq 1 + k\, N^{\text{ext}}(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon e^{-\Lambda_s T}/T)$$

(the radius is $\varepsilon/(T e^{\Lambda_s T})$, equal to 0 when $T = 0$).

Proof. Subadditivity over the union $\{\text{id}\} \cup \bigcup_{m=1}^{k} \{\Phi_{s,m}\}$ and 'lem:ode-equal-scheme-image-covering' for each m. $\square$

Theorem 733. *The equal-step scheme classes are nested: $E_T(k) \subseteq E_T(k+1)$.*

Proof. $\square$

Theorem 734. *$E_T(k) \subseteq B_T(k)$ when $0 \leq T \leq h_0$ (the equal steps T/m, $1 \leq m \leq k$, are admissible).*

Proof. $\square$

Theorem 735. ***(Explicit entropy of equal-step schemes.)*** *Let every drift $s \in \mathcal{S}_T$ be Λ_s-Lipschitz in x and $T \geq 0$. Then for all $s, s' \in \mathcal{S}_T$ and $m \geq 0$,*

$$d_\infty(\Phi_{s,m}, \Phi_{s',m}) \leq T e^{\Lambda_s T} \|s - s'\|_\infty, \qquad \|s - s'\|_\infty := \sup_{x \in K,\ \tau \in [0,T]} \|s(x, \tau) - s'(x, \tau)\|,$$

and consequently, for every $k \geq 0$ and $\varepsilon \geq 0$,

$$N^{\text{ext}}(E_T(k), d_\infty, \varepsilon) \leq 1 + k\, N^{\text{ext}}(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon e^{-\Lambda_s T}/T).$$

The entropy-integral consequence is 'lem:ode-scheme-entropy-profile'.

Proof. □

Theorem 736. ***(Entropy integral of equal-step schemes.)*** *Let K be compact, every drift $s \in \mathcal{S}_T$ be Λ_s-Lipschitz in x, $T > 0$, and suppose $N^{\mathrm{ext}}(\mathcal{S}_T, \|\cdot\|_\infty, \rho) < \infty$ for all $\rho > 0$ and that $\varepsilon \mapsto \sqrt{\log N^{\mathrm{ext}}(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon e^{-\Lambda_s T}/(2T))}$ is integrable on $[0, D_K]$, $D_K = \mathrm{diam}(K)$. Then for every sample S and every k, for the class $E_T(k)$,*

$$\mathsf{V}_k(S) \le D_K \sqrt{\log(k+1)} + \int_0^{D_K} \sqrt{\log N^{\mathrm{ext}}(\mathcal{S}_T, \|\cdot\|_\infty, \varepsilon e^{-\Lambda_s T}/(2T))}\, d\varepsilon.$$

(The paper has $\varepsilon e^{-\Lambda_s T}/T$: the factor 2 is the price of comparing the internal covering number in d_S defining $\mathsf{V}_k(S)$ with the external one in d_∞, 'lem:covering-empSpace-le-external-unifMaps'.)

Proof. Pointwise, $\log N(E_T(k), d_S, \varepsilon) \le \log N^{\mathrm{ext}}(E_T(k), d_\infty, \varepsilon/2) \le \log(1 + kN) \le \log(k+1) + \log N$ with $N = N^{\mathrm{ext}}(\mathcal{S}_T, \varepsilon e^{-\Lambda_s T}/(2T))$; take square roots termwise and integrate over $(0, D_K]$, using $D_k(S) \le D_K$. □