Vandermonde asymptotics and scale-aware singular learning

Dmitry Vaintrob

May 2026

Abstract

We estimate the Bayesian reduced free energy $F_H(n)$ of a width-H two-layer cosine network learning the zero function, jointly in the width H and the sample size/inverse temperature n. For $n \geq H$ we show, with absolute constants, that $F_H(n) = \Theta(m_n \log n)$ where $m_n = \min(\log n/\log\log n, \sqrt{H})$. Equivalently: below the crossover $\log n \sim \sqrt{H}\log H$ the free energy is $\Theta((\log n)^2/\log\log n)$, independent of H to leading order; above it, $F_H(n) = \Theta(\sqrt{H}\log n)$, the scale of the fixed-H real log-canonical threshold. The limits $n \to \infty$ at fixed H and $H \to \infty$ at polynomial n therefore do not commute: at polynomial sample size the leading term is set by the finite Hermite band resolved at scale n, not by the full analytic germ. The proof integrates out the readout weights and controls the resulting determinant through an exact positive expansion: the cosine kernel is $\exp(-(w_i^2+w_j^2)/2)\cosh(w_i w_j)$, and its Cauchy–Binet minor expansion is a positive series in squared Schur polynomials with inverse-factorial weights. All bounds are uniform in (H, n) and the argument is self-contained. An appendix gives a tropical/Schur variational formula for the fixed-H RLCT of parity-pure analytic activations, including the cosine case and the odd/tanh case of Aoyagi–Watanabe, and a comparison squeeze for mixed-parity supports.

1 Overview and results

Singular learning theory (SLT) studies the large-sample asymptotics of singular statistical models. In Watanabe's theory [19], if the architecture is fixed and the inverse temperature/sample size n tends to infinity, the reduced free energy has the form

$$\bar{F}(n) = \lambda \log n - (m-1)\log\log n + O(1),$$

where λ is the real log-canonical threshold (RLCT) of the zero-loss equations and m is its multiplicity. The RLCT is thus the endpoint of the fixed-architecture limit. The error terms in this expansion depend on the architecture, so the statement by itself says nothing about the joint behavior in (H, n) when the width H is large.

This paper computes that joint behavior in a deliberately simple model. We work with Gaussian input $x \sim \mathcal{N}(0,1)$, no biases, bounded hidden weights $|w_i| \le 1$, a Gaussian prior on the readout weights, and network

$$f_\theta(x) = \sum_{i=1}^{H} a_i \cos(w_i x).$$

The population loss is $K(\theta) = \|f_\theta\|^2_{L^2(\mathcal{N}(0,1))}$, and the zero network realizes the minimum $K^* = 0$. We write $F_H(n)$ for the reduced population free energy at inverse temperature n; Appendix B records the standard reduction from empirical Bayesian free energy to this population quantity.

Theorem 1.1. *There are absolute constants $0 < c \le C$ such that for all $H \ge 1$ and all $n \ge H$,*

$$c\, m_n \log n \;\le\; F_H(n) \;\le\; C\, m_n \log n, \qquad m_n := \min\left(\frac{\log n}{\log\log n}, \sqrt{H}\right).$$

Equivalently, up to absolute constants,

$$F_H(n) = \begin{cases} \Theta\left(\dfrac{(\log n)^2}{\log\log n}\right), & n \le (2\sqrt{H})!, \\ \Theta\left(\sqrt{H}\log n\right), & n > (2\sqrt{H})!. \end{cases}$$

Here

$$(2\sqrt{H})! \approx \exp\Big(\sqrt{H}(\log H - O(1))\Big),$$

with the factorial interpreted as $\Gamma(2\sqrt{H}+1)$ when $2\sqrt{H}$ is not an integer; replacing it by $\lfloor 2\sqrt{H}\rfloor!$ changes none of the Θ estimates.

Remark 1.2 (equivalence of the two displays)**.** *The minimum in m_n switches branches when $\log n/\log\log n \asymp \sqrt{H}$, i.e. when $\log n \asymp \sqrt{H}\log H$, which is the scale of $\log\big((2\sqrt{H})!\big)$. Near this scale $\log\log n \asymp \log H$, so*

$$\frac{(\log n)^2}{\log\log n} \asymp \sqrt{H}\log n,$$

and the two regime formulas agree within absolute constants. In particular the joint scaling law is continuous across the crossover, and each regime formula may be read off from the unified form.

The unified form has the reading

$$F_H(n) \asymp (\#\text{constrained Hermite modes}) \cdot \log n,$$

the count being the effective rank of the ridged design matrix on the region of weight space that dominates the integral. Below the crossover this is the Hermite cutoff $L_n \asymp \log n/\log\log n$, the index at which the coefficients $1/\sqrt{(2k)!}$ pass

the ridge scale $n^{-1/2}$, independently of H: heuristically, the partition function is there dominated by the generic-w fiber, and F_H grows in $\log n$ by the successive freezing of Hermite modes as their coefficients cross the ridge scale. Past the crossover, a degenerate region of clustered small weights dominates instead—the Vandermonde collision geometry makes it affordable in prior volume to leave all but $O(\sqrt{H})$ modes unconstrained—and the count saturates at $\sqrt{H}$. (The proof proceeds by two-sided bounds and does not formalize this domination picture.)

Theorem 1.1 is proved by uniform determinant estimates; the proof does not use Watanabe's expansion or the appendix. For fixed H, the post-crossover coefficient matches the RLCT scale $\lambda_H = \Theta(\sqrt{H})$; Appendix A computes the fixed-H constant by a tropical/Schur variational formula in the parity-pure case, including the cosine activation used here and the odd/tanh case of Aoyagi–Watanabe [3].

For the finite-temperature learning coefficient $\widehat{\lambda}_H(n) := \partial F_H(n)/\partial \log n$ (Section 2), the same Hermite tail bound gives the pointwise estimate

$$\widehat{\lambda}_H(n) \;\le\; C\,\frac{\log n}{\log\log n}, \qquad n \ge H,$$

uniformly in w and hence in the posterior (Proposition 6.3). Two-sided pointwise control of $\widehat{\lambda}_H$ would require localizing the posterior distribution of the hidden weights, which we do not pursue here; note that two-sided bounds on F_H constrain only averages of $\widehat{\lambda}_H$ over multiplicative windows in n.

Remark 1.3 (small n)**.** *The theorem is stated for $n \ge H$, the regime needed for the interpolation. Appendix C records the standard random-kernel heuristic for $n \le H$: under finite-width kernel concentration assumptions, the same Hermite cutoff gives*

$$F_H(n) = \Theta\!\left(\frac{(\log(nH+5))^2}{\log\log(nH+5)}\right).$$

We keep this as a supporting estimate rather than a second theorem, since making the finite-width kernel comparison uniform would require concentration bookkeeping not central to the Vandermonde argument.

1.1 Proof outline

Integrating out the readout weights (Section 3) reduces the partition function to

$$Z_H(\beta) = \int_{[-1,1]^H} \det\bigl(I + \beta Q(w)\bigr)^{-1/2} d\mu_w, \qquad Q_{ij}(w) = \langle \cos(w_i x), \cos(w_j x)\rangle,$$

with $\beta = n$. In the Hermite basis, $Q = AA^T$ for an $H \times \infty$ design matrix A whose kth column carries the spectral weight $1/\sqrt{(2k)!}$ and whose polynomial part is a Vandermonde matrix in the squared weights $t_i = w_i^2$.

The structural input (Section 4) is an exact positive expansion. The kernel has the closed form $Q_{ij} = e^{-(w_i^2+w_j^2)/2}\cosh(w_i w_j)$, and Cauchy–Binet turns

every principal $\ell \times \ell$ minor of Q into a series over partitions with at most ℓ parts,

$$\det(Q_I) = e^{-\sum_{i\in I} w_i^2} \operatorname{Vand}(t_I)^2 \sum_{\lambda} \frac{s_\lambda(t_I)^2}{\prod_{j=1}^{\ell} \big(2(\lambda_{\ell+1-j} + j - 1)\big)!},$$

in which every term is a squared minor. The ground state $\lambda = \emptyset$ is the discriminant term $\operatorname{discr}_I^2 / \prod_{k<\ell}(2k)!$, and the excited states are dominated by the same factorial decay (Theorem 5.1). Consequently the determinant is equivalent, up to lower-order errors, to the best weighted discriminant minor: the competition is between the statistical gain $\ell \log \beta$, the Hermite spectral penalty $\sum_{k<\ell} \log(2k)!$, and the clustering penalty $-2 \log \operatorname{discr}_\ell(w)$.

The theorem then follows from three uniform estimates (Section 6). The bound $F_H \lesssim (\log n)^2 / \log\log n$ comes from the worst-case Hermite tail, uniformly in w. The bound $F_H \lesssim \sqrt{H} \log n$ comes from restricting the integral to the radial region $|w_i| \le e^{-\log n/(4\sqrt{H})}$, on which only $O(\sqrt{H})$ modes lie above the ridge scale, at a matching volume cost. The lower bound $F_H \gtrsim m_n \log n$ comes from a single dichotomy at the parameters $m \asymp m_n$, $\delta = e^{-\log n/(8m)}$: either some m of the squared weights are pairwise δ-separated, in which case the ground-state minor forces the determinant up to $e^{\Omega(m \log n)}$; or all squared weights cluster into fewer than m intervals of length 2δ, an event of prior volume $e^{-\Omega(m \log n)}$. The same parameter choice covers both regimes, which is what makes the estimate uniform across the crossover.

2 Bayesian learning and SLT

We summarize the two definitions from Watanabe's theory used in this paper, for neural networks with MSE loss; Appendix B records the reduction from finite-data Bayesian statistics to the population quantities below.

2.1 Free energy and the learning coefficient

Let $K(\theta) = \mathbb{E}_x K(\theta, x)$ be the population loss and $K^* = \inf_\theta K$. The *reduced free energy* at inverse temperature n is

$$\bar{F}(n) := -\log \int e^{-nK(\theta)}\, d\mu(\theta) \;-\; nK^*,$$

with μ the prior. Under standard hypotheses this deterministic quantity controls, to leading order, the quenched Bayesian free energy and hence generalization at sample size n [19]; it is the continuous analogue of description length in discrete Bayesian induction [16]. The inverse temperature $\beta = n$ is the one natural for Bayesian learning, since the n-point empirical loss is the sum of n per-point losses; because n enters only as a scaling parameter, $\bar{F}(n)$ is defined for all real $n \ge 0$. When comparing basins, mechanisms, or local models with the same minimum loss, differences in $\bar{F}$ at the logarithmic scale determine the relative Bayesian weights, which is why this subleading scale matters.

The *finite-temperature learning coefficient* is

$$\widehat{\lambda}(n) := \frac{\partial \bar{F}(n)}{\partial \log n} = n\,\mathbb{E}_{\theta\sim p_n}\big[K(\theta) - K^*\big], \qquad p_n(\theta) \propto e^{-nK(\theta)}\,d\mu(\theta).$$

The second expression makes $\widehat{\lambda}(n)$ estimable by posterior sampling, and it is the quantity typically measured in empirical studies of model degeneracy [10, 15, 5], and is closely related to the widely applicable Bayesian information criterion [20].

2.2 The algebraic endpoint

For a fixed analytic model, Watanabe's theorem gives

$$\bar{F}(n) = \lambda \log n - (m-1)\log\log n + O(1), \qquad n \to \infty,$$

where λ is the RLCT of the zero-loss germ and m its multiplicity; correspondingly $\widehat{\lambda}(n) \to \lambda$. We call λ the *algebraic learning coefficient*, to distinguish it from the finite-temperature quantity $\widehat{\lambda}(n)$. The RLCT can differ from any quadratic count: the germ $(\theta_0\theta_1)^2$ has zero Hessian at the origin but nonzero RLCT, and for a two-layer network with odd analytic activation (such as tanh) learning the zero function, the Hessian at the origin is maximally degenerate while the Aoyagi–Watanabe computation gives a nontrivial intermediate RLCT [18, 3]. The algebraic singularity, not its quadratic approximation, is the correct fixed-H endpoint.

The expansion above is an asymptotic in n at fixed architecture, with architecture-dependent error terms. Theorem 1.1 quantifies the resulting non-uniformity for the cosine model: the limits $n \to \infty$ at fixed H and $H \to \infty$ at polynomial n do not commute, and the fixed-H coefficient becomes the leading term only past the width-dependent scale $\log n \asymp \sqrt{H}\log H$.

3 Model and reduction to a determinant

3.1 Setting

A parameter is $\theta = (a, w) \in \mathbb{R}^H \times \mathbb{R}^H$, and

$$f_\theta(x) = \sum_{i=1}^{H} a_i \cos(w_i x).$$

The data measure is $\mu_x = \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx$, inducing the inner product $\langle f, g\rangle := \int fg\,d\mu_x$ and norm $|f|^2 := \langle f, f\rangle$ on square-integrable functions. The population loss for the zero target is

$$K(\theta) := |f_\theta|^2.$$

The prior is

$$\mu_\theta(a,w) = \frac{e^{-|a|^2}}{\sqrt{\pi}^H}\, da\, \mu_w, \qquad \mu_w = \prod_{i=1}^{H} \tfrac{1}{2}\, \chi(w_i \in [-1,1])\, dw_i,$$

uniform on the cube $[-1,1]^H$ in w and Gaussian in a; the prior does not vary with the inverse temperature β.

Remark 3.1 (prior conventions)**.** *The normalization $e^{-a_i^2}/\sqrt{\pi}$ (rather than standard Gaussian) avoids factors of 2 in the determinant below; any fixed rescaling changes F by $\Theta(1)$. Replacing the Gaussian readout prior by $a_i \sim \mathcal{N}(0, \text{const.} \cdot H^{-\alpha})$ for any fixed α (lazy, mean-field, or anti-scaled) changes the estimates only at lower order in the regimes considered, and Appendix C records the comparison with the hard constraint $|a_i| \le 1$, which costs $\Theta(1)$ additively at the present precision.*

3.2 Integrating out the readout weights

Write $\psi_w(x) := \cos(wx)$. Then $K(a,w) = a^T Q(w)\, a$ with

$$Q_{ij}(w) = \langle \psi_{w_i}, \psi_{w_j} \rangle,$$

a positive semidefinite $H \times H$ matrix. The Gaussian a-integral gives

$$Z(w) := \int e^{-\beta K(a,w)}\, d\mu_a = \sqrt{\pi}^{-H} \int e^{-a^T (I+\beta Q(w))a}\, da = \det\big(I+\beta Q(w)\big)^{-1/2},$$

the prefactor cancelling against the Gaussian integral (as a check, $Q = 0$ integrates the prior to 1). Hence

$$Z_H(\beta) = \int_{[-1,1]^H} \det\big(I + \beta Q(w)\big)^{-1/2}\, d\mu_w, \qquad F_H(\beta) = -\log Z_H(\beta).$$

This reduction of the free energy to a determinant integral follows [3]. Writing $\sigma_0(M) \ge \sigma_1(M) \ge \cdots$ for singular values, always indexed in decreasing order and set to zero beyond the rank, the eigenvalues of the positive semidefinite Q coincide with its singular values and

$$\det(I + \beta Q) = \prod_{h=0}^{H-1} \big(1 + \beta\, \sigma_h(Q)\big).$$

3.3 Hermite basis and the design matrix

Let he_k, $k \ge 0$, be the Hermite polynomials orthonormal for the Gaussian weight,[1] so that by the generating identity $\mathbb{E}\big[\mathrm{He}_k(x) e^{iwx}\big] = (iw)^k e^{-w^2/2}$,

$$\langle \mathrm{he}_k, \psi_w \rangle = \begin{cases} \dfrac{(-1)^{k/2}}{\sqrt{k!}}\, w^k e^{-w^2/2}, & k \text{ even}, \\ 0, & k \text{ odd}, \end{cases}$$

[1] I.e. the normalized probabilist's Hermite polynomials $\mathrm{he}_k = \mathrm{He}_k/\sqrt{k!}$, satisfying $\langle \mathrm{he}_k, \mathrm{he}_\ell \rangle = \delta_{k\ell}$ for the $\mathcal{N}(0,1)$ weight.

the odd coefficients vanishing because ψ_w is even. Suppressing the odd rows and applying Parseval, $Q = AA^T$ for the *design matrix* [6]

$$A_{ik} := \langle \mathrm{he}_{2k}, \psi_{w_i} \rangle = \frac{(-1)^k}{\sqrt{(2k)!}}\, w_i^{2k}\, e^{-w_i^2/2}, \qquad i = 1, \dots, H, \quad k \geq 0.$$

A is an $H \times \infty$ matrix with square-summable rows, so AA^T is well defined, and

$$\sigma_h(Q) = \sigma_h(A)^2.$$

The Hermite basis is the energy eigenbasis of the harmonic oscillator; expanding a finite kernel in an infinite orthonormal eigenbasis is the same manipulation as in the field-theoretic treatment of wide networks [8, 9], and is a first indication that the present computation can be phrased in a statistical field theory setting.

3.4 The factorization $A = GUD$

The entries of A split into three factors:

$$A = GUD, \qquad G := \mathrm{diag}\big(e^{-w_i^2/2}\big)_{1 \leq i \leq H}, \quad U_{ik} := w_i^{2k}, \quad D := \mathrm{diag}\left(\frac{(-1)^k}{\sqrt{(2k)!}}\right)_{k \geq 0},$$

that is,

$$A = \begin{pmatrix} e^{-w_1^2/2} & & \\ & \ddots & \\ & & e^{-w_H^2/2} \end{pmatrix} \begin{pmatrix} 1 & w_1^2 & w_1^4 & \cdots \\ 1 & w_2^2 & w_2^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & w_H^2 & w_H^4 & \cdots \end{pmatrix} \begin{pmatrix} 1 & & & \\ & \frac{-1}{\sqrt{2}} & & \\ & & \frac{1}{\sqrt{24}} & \\ & & & \ddots \end{pmatrix},$$

with G carrying the w-dependent Gaussian envelope, U the polynomial (Vandermonde) structure, and D the Hermite spectral weights. The three factors play the following roles.

D: rapid spectral decay. The diagonal entries $D_k = 1/\sqrt{(2k)!}$ decay superexponentially: by Stirling,

$$-\log D_k = k\big(\log k + O(1)\big).$$

Table 1 shows the first values. The decay rate is a smoothness property of the activation: cos is entire, giving the superexponential schedule; a general analytic activation with finite complex singularities gives at least exponential decay. In the algebro-geometric limit $n \to \infty$ at fixed H, the D_k are nonzero constants and do not affect the singularity structure; at polynomial n they dominate the asymptotics, since the mode k is statistically visible only when D_k exceeds the ridge scale $n^{-1/2}$.

Table 1: Values of $D_k = 1/\sqrt{(2k)!}$ for $k = 0, \ldots, 20$.

k	D_k	k	D_k	k	D_k
0	1.0	7	3.4×10^{-6}	14	1.8×10^{-15}
1	7.1×10^{-1}	8	2.2×10^{-7}	15	6.1×10^{-17}
2	2.0×10^{-1}	9	1.2×10^{-8}	16	1.9×10^{-18}
3	3.7×10^{-2}	10	6.4×10^{-10}	17	5.8×10^{-20}
4	5.0×10^{-3}	11	3.0×10^{-11}	18	1.6×10^{-21}
5	5.2×10^{-4}	12	1.3×10^{-12}	19	4.4×10^{-23}
6	4.6×10^{-5}	13	5.0×10^{-14}	20	1.1×10^{-24}

Remark 3.2. *The D_k are not Hessian eigenvalues at the singular point: at $\theta = 0$ the loss K is maximally singular and the relevant Hessian data vanish identically. The D_k instead weight the singular values of the design matrix at a* general *parameter w, which is one reason an effective theory interfacing with the singularity requires the scale-dependent treatment developed below; see also Section 7.*

U: a parity-reduced Vandermonde matrix. Writing $t := w^{\odot 2} = (w_1^2, \ldots, w_H^2)$ for the coordinatewise square, $U(w) = V(t)$ where V is the (rectangular) Vandermonde matrix $V(t)_{ik} = t_i^k$. The first square minor obeys the Vandermonde identity

$$\det\big(V(t)_{:,<H}\big) = \prod_{1 \le i < j \le H} (t_j - t_i),$$

so that in w-coordinates

$$\det\big(U(w)_{:,<H}\big) = \prod_{i<j} (w_j^2 - w_i^2) = \prod_{i<j} (w_j - w_i) \prod_{i<j} (w_j + w_i).$$

The discriminant factor $\prod(w_j - w_i)$ is the generic Vandermonde vanishing; the extra factor $\prod(w_j + w_i)$ is the parity correction, and it is what distinguishes the parity-pure singular geometry analyzed in Appendix A from the mixed-parity case. General minors of U are generalized Vandermonde determinants and factor as the discriminant times a Schur polynomial in t; this Schur structure, with its positivity, is the algebraic engine of both the main estimate and the appendix.

G: exactly accounted for. Since $|w_i| \le 1$, the diagonal factor G has entries in $[e^{-1/2}, 1]$. It requires no separate treatment: the exact kernel formula of the next section splits G off multiplicatively from every minor, where it contributes the explicit factor $e^{-\sum_{i \in I} w_i^2} \in [e^{-|I|}, 1]$.

4 The exact kernel and minor expansion

The Gaussian data measure gives the kernel Q a closed form, and the closed form gives every minor of Q an exact positive expansion. This expansion is the

structural engine of the paper: the singular-value estimates of Section 5 are read off from it, with the ground state of the expansion supplying the lower bounds and a resummation of the excited states supplying the upper bounds.

Proposition 4.1 (kernel closed form)**.** *For all $u, v \in \mathbb{R}$,*

$$\langle \psi_u, \psi_v \rangle = \tfrac{1}{2}\Big(e^{-(u-v)^2/2} + e^{-(u+v)^2/2} \Big) = e^{-(u^2+v^2)/2} \cosh(uv).$$

In particular $Q = G\,\widetilde{Q}\,G$ exactly, with $\widetilde{Q}_{ij} = \cosh(w_i w_j)$ and $G = \operatorname{diag}(e^{-w_i^2/2})$.

Proof. Use $\cos a \cos b = \frac{1}{2}[\cos(a-b) + \cos(a+b)]$ and $\int \cos(tx)\, d\mu_x = e^{-t^2/2}$; the second equality is elementary. □

Remark 4.2. *The first form exhibits Q as the even symmetrization of the Gaussian RBF kernel in the* weights*: $Q_{ij} = \frac{1}{2}\big[K_{\mathrm{RBF}}(w_i, w_j) + K_{\mathrm{RBF}}(w_i, -w_j)\big]$ with $K_{\mathrm{RBF}}(u, v) = e^{-(u-v)^2/2}$. This is the real-cosine shadow of the complex-exponential model $\varphi(x) = e^{iwx}$, for which the kernel is the RBF kernel itself.*

Notation. Write $t := w^{\odot 2} = (w_1^2, \dots, w_H^2)$. For $I \subset \{1, \dots, H\}$ with elements $i_1 < \cdots < i_\ell$ set

$$\mathrm{Vand}(t_I) := \prod_{p<q} \big(t_{i_q} - t_{i_p}\big), \qquad \mathrm{discr}_I(w) := \big|\mathrm{Vand}(t_I)\big| = \prod_{i<j\in I} |w_i^2 - w_j^2|,$$

with the empty products equal to 1 when $|I| \le 1$, and

$$\mathrm{discr}_\ell(w) := \max_{|I|=\ell} \mathrm{discr}_I(w), \qquad \ell\Downarrow := \prod_{k=0}^{\ell-1} (2k)!, \qquad 0\Downarrow := 1.$$

Partitions $\lambda = (\lambda_1 \ge \cdots \ge \lambda_\ell \ge 0)$ with at most ℓ parts are in bijection with column sets $S = \{a_1 < \cdots < a_\ell\} \subset \mathbb{N}$ via

$$a_j = \lambda_{\ell+1-j} + (j-1),$$

the ground state $\lambda = \emptyset$ corresponding to $S_0 = \{0, 1, \dots, \ell-1\}$.

Proposition 4.3 (exact minor expansion)**.** *For every $I \subset \{1, \dots, H\}$ with $|I| = \ell$,*

$$\det(Q_I) \;=\; e^{-\sum_{i\in I} w_i^2}\, \mathrm{Vand}(t_I)^2\, \Sigma_\ell(t_I), \qquad \Sigma_\ell(t) := \sum_{\lambda} \frac{s_\lambda(t)^2}{\prod_{j=1}^{\ell} \big(2(\lambda_{\ell+1-j} + j - 1)\big)!}, \tag{1}$$

the sum running over all partitions λ with at most ℓ parts. Every term is nonnegative, the series converges, and the ground-state term $\lambda = \emptyset$ equals $1/\ell\Downarrow$.

Proof. Write $Q_I = A_I A_I^T$ with A_I the corresponding $\ell \times \infty$ row-submatrix of A. By Cauchy–Binet,

$$\det(Q_I) = \sum_{|S|=\ell} \det(A_{I,S})^2.$$

For $S = \{a_1 < \cdots < a_\ell\}$, factoring the row weights $e^{-w_i^2/2}$, the column signs $(-1)^{a_j}$, and the column weights $1/\sqrt{(2a_j)!}$ out of the determinant leaves the generalized Vandermonde $\det\big(t_i^{a_j}\big)_{i\in I,\, j\le\ell}$, so

$$\det(A_{I,S})^2 = e^{-\sum_{i\in I} w_i^2}\, \frac{\det\big(t_i^{a_j}\big)^2}{\prod_j (2a_j)!}.$$

The alternant formula $\det\big(t_i^{a_j}\big) = \pm \operatorname{Vand}(t_I)\, s_\lambda(t_I)$, with λ corresponding to S as above, gives (1) after summing over S. Convergence holds because the left side is a finite determinant (or directly from the estimate in Theorem 5.1 below); positivity is manifest termwise, and $S = S_0$ gives $s_\emptyset = 1$ and weight $1/\ell\Downarrow$. □

Remark 4.4 (structure of the expansion)**.** *The three factors of* (1) *are exactly the three factors of $A = GUD$: the Gaussian envelope G contributes $e^{-\sum_{i\in I} w_i^2} \in [e^{-\ell}, 1]$, the Vandermonde structure of U contributes $\operatorname{Vand}(t_I)^2\, s_\lambda(t_I)^2$, and the Hermite weights D contribute the inverse factorials. Nothing is discarded and nothing can cancel: the local integrand of the free energy is globally a positive series in squared Schur polynomials. (For the cosine model this makes the no-cancellation hypotheses of the tropical calculation in Appendix A automatic.) Summing over row sets gives the exterior-power identity*

$$\sum_{|I|=\ell} \det(Q_I) = \big\|\wedge^\ell A\big\|_F^2 = e_\ell\big(\sigma_0(A)^2, \ldots, \sigma_{H-1}(A)^2\big),$$

with e_ℓ the elementary symmetric polynomial, which is the form used below.

5 Singular value spectroscopy

We now estimate

$$\Delta_\beta(w) := \det\big(I + \beta A(w) A(w)^T\big) = \prod_{j=0}^{H-1} \big(1 + \beta\, \sigma_j(A(w))^2\big),$$

so that $Z_H(\beta) = \int_{[-1,1]^H} \Delta_\beta(w)^{-1/2} d\mu_w$ and $F_H(\beta) = -\log Z_H(\beta)$. Write

$$\Pi_\ell(M) := \prod_{j=0}^{\ell-1} \sigma_j(M), \qquad \Pi_0(M) := 1,$$

for singular products; these are controlled by minors, whereas individual singular values are awkward near collisions of the w_i.

5.1 Hermite tail and cutoff

Let $A_{\geq r}$ be the matrix keeping only the Hermite columns $k \geq r$. Since $|w_i| \leq 1$,

$$\|A_{\geq r}\|_{op}^2 \leq \|A_{\geq r}\|_F^2 \leq H \sum_{k \geq r} \frac{1}{(2k)!} \leq \frac{2H}{(2r)!},$$

and by Weyl's inequality $\sigma_r(A) \leq \|A_{\geq r}\|_{op}$, so

$$\sigma_r(A)^2 \leq \frac{2H}{(2r)!} \qquad \text{uniformly in } w. \tag{2}$$

Define the Hermite cutoff

$$L_\beta := \min\Big(H,\ \min\{r \geq 1 : (2r)! \geq 16H\beta\}\Big).$$

For $r \geq L_\beta$ we have $\beta\sigma_r(A)^2 \leq \frac{1}{8}$ (using $\sigma_r = 0$ for $r \geq H$), and summing the geometric-or-faster tail,

$$\log \Delta_\beta(w) = \sum_{j=0}^{L_\beta - 1} \log\big(1 + \beta\sigma_j(A(w))^2\big) + O(1), \tag{3}$$

uniformly in w. By Stirling, $\log(2r)! \geq r \log r$ for $r \geq 8$, so in the range $\beta \geq H$ of the main theorem,

$$L_\beta \leq C\, \frac{\log(\beta H)}{\log\log(\beta H)} \leq C'\, \frac{\log \beta}{\log\log \beta}. \tag{4}$$

5.2 Two-sided control by weighted discriminants

The following estimate is the linear-algebraic engine of the proof. Both directions are read off from the exact expansion (1).

Theorem 5.1. *For every $w \in [-1,1]^H$ and $0 \leq \ell \leq H$,*

$$e^{-\ell/2}\, \frac{\mathrm{discr}_\ell(w)}{\sqrt{\ell\!\Downarrow}} \ \leq\ \Pi_\ell(A(w)) \ \leq\ e^{\ell^2/2} \sqrt{\binom{H}{\ell}}\, \frac{\mathrm{discr}_\ell(w)}{\sqrt{\ell\!\Downarrow}}.$$

Proof. Lower bound. Since $Q \succeq 0$, the eigenvalues of a principal submatrix Q_I interlace those of Q, so $\det(Q_I) \leq \Pi_\ell(A)^2$ for any $|I| = \ell$. Choose I attaining $\mathrm{discr}_\ell(w)$, keep only the ground state $\lambda = \emptyset$ in (1), and use $e^{-\sum_{i \in I} w_i^2} \geq e^{-\ell}$:

$$\Pi_\ell(A)^2 \ \geq\ \det(Q_I) \ \geq\ e^{-\ell}\, \frac{\mathrm{discr}_\ell(w)^2}{\ell\!\Downarrow}.$$

Upper bound. The product of the top ℓ eigenvalues is at most their elementary symmetric function, so by Remark 4.4,

$$\Pi_\ell(A)^2 \leq \sum_{|I| = \ell} \det(Q_I).$$

Fix I and bound the three factors of (1): $e^{-\sum w_i^2} \le 1$ and $\mathrm{Vand}(t_I)^2 \le \mathrm{discr}_\ell(w)^2$. For $\Sigma_\ell(t_I)$, since Schur polynomials have nonnegative coefficients and $t \in [0,1]^\ell$,

$$s_\lambda(t_I) \le s_\lambda(1^\ell) \le \ell^{|\lambda|},$$

the last step counting semistandard fillings entrywise. Writing $\mu_j := \lambda_{\ell+1-j}$ and using $(a+b)! \ge a!\,b!$,

$$\big(2(j-1+\mu_j)\big)! \;\ge\; (2(j-1))!\,(2\mu_j)!,$$

so, dropping the ordering constraint on $(\mu_1, \dots, \mu_\ell)$ so that the sum factors over rows,

$$\Sigma_\ell(t_I) \;\le\; \frac{1}{\ell\Downarrow} \prod_{j=1}^{\ell} \sum_{\mu \ge 0} \frac{\ell^{2\mu}}{(2\mu)!} \;=\; \frac{\cosh(\ell)^\ell}{\ell\Downarrow} \;\le\; \frac{e^{\ell^2}}{\ell\Downarrow}.$$

Summing over the $\binom{H}{\ell}$ row sets and taking square roots proves the claim. □

Remark 5.2 (sharper constant)**.** *The factor $e^{\ell^2/2}$ can be improved to $e^{C\ell}$ with C absolute: the claim is that*

$$\sup_{t \in [0,1]^\ell} \ell\Downarrow \cdot \Sigma_\ell(t) \;=\; \sum_{\lambda} s_\lambda(1^\ell)^2 \prod_{j=1}^{\ell} \frac{(2(j-1))!}{\big(2(j-1+\lambda_{\ell+1-j})\big)!} \;\le\; e^{C\ell},$$

i.e. the weighted Cauchy–Binet series is ground-state dominated up to $e^{O(\ell)}$. This follows from the hook–content evaluation

$$s_\lambda(1^\ell) = \prod_{u \in \lambda} \frac{\ell + c(u)}{h(u)} :$$

rectangular shapes $\lambda = (m^\ell)$ have $s_\lambda(1^\ell) = 1$ (the Schur polynomial is $(t_1 \cdots t_\ell)^m$) while paying the full factorial cost, and each unit increment of a row of λ multiplies $s_\lambda(1^\ell)^2$ by at most $e^{O(\ell)}$ in total while dividing the weight by a product of factorial ratios that grows faster along every excitation direction. This is the same mechanism as the main theorem in miniature: Hermite factorial decay dominates dimension growth. We do not use the refinement: the cruder constant contributes only $O(L_\beta^2)$ to (6), which is lower order in both regimes of Theorem 1.1.

5.3 The working estimate

Expanding the product in (3): each factor satisfies $\max\big(0, \log \beta\sigma_j^2\big) \le \log(1 + \beta\sigma_j^2) \le \log 2 + \max\big(0, \log \beta\sigma_j^2\big)$, and by the monotone ordering of the σ_j,

$$\sum_{j < L_\beta} \max\big(0, \log \beta\sigma_j^2\big) = \max_{0 \le \ell \le L_\beta} \sum_{j < \ell} \log \beta\sigma_j^2 = \max_{0 \le \ell \le L_\beta} \Big(\ell \log \beta + 2 \log \Pi_\ell(A(w))\Big).$$

Hence

$$\log \Delta_\beta(w) = \max_{0\le \ell \le L_\beta} \Big(\ell \log \beta + 2 \log \Pi_\ell(A(w)) \Big) + O(L_\beta). \tag{5}$$

Substituting Theorem 5.1 and $\binom{H}{\ell} \le H^\ell$,

$$\log \Delta_\beta(w) = \max_{0\le \ell \le L_\beta} \Big(\ell \log \beta - \log(\ell \Downarrow) + 2 \log \operatorname{discr}_\ell(w) \Big) + O\big(L_\beta^2 + L_\beta \log H\big). \tag{6}$$

Because Theorem 5.1 is two-sided, (6) is an equivalence, not a bound: to the stated precision, the free-energy landscape in w *is* the landscape of the weighted-discriminant objective, whose three terms are the statistical gain $\ell \log \beta$, the Hermite spectral penalty $\log(\ell \Downarrow) = \sum_{k<\ell} \log(2k)!$, and the clustering penalty $-2 \log \operatorname{discr}_\ell(w)$. The proof of Theorem 1.1 below uses only the lower half of Theorem 5.1 together with the uniform tail bound (2); the two-sided form is what justifies reading (6) as the effective description of the model; the tropical objective of Appendix A is its fixed-H, fixed-leaf analogue.

6 Proof of the main theorem

Throughout, $\beta = n \ge H \ge 1$ and $M := \log \beta$. All constants $c, C, c_0, \dots$ are absolute. We prove the estimates for $M \ge C_0$ with C_0 absolute; on the complementary range $M \le C_0$ we have $H \le e^{C_0}$, and $F_H(\beta)$ is bounded above and below by positive absolute constants there (F_H is increasing in β, positive for $\beta > 0$, and continuous on the compact remaining range of (H, β)), so the bounded range is absorbed into the constants of the theorem.

6.1 Upper bounds

Lemma 6.1. *For all $H \ge 1$ and $\beta \ge H$,*

$$\textit{(a)} \quad F_H(\beta) \le C \, \frac{M^2}{\log M}, \qquad\qquad \textit{(b)} \quad F_H(\beta) \le C \sqrt{H}\, M.$$

Consequently $F_H(\beta) \le C\, m_n M$ with $m_n = \min(M/\log M, \sqrt{H})$.

Proof. (a) By the uniform tail bound (2), for every w

$$\log \Delta_\beta(w) \le \sum_{k \ge 0} \log\Big(1 + \frac{2\beta H}{(2k)!}\Big).$$

Let $L^* := \#\{k : (2k)! \le 2\beta H\}$. Since $\log(2k)! \ge k \log k$ for $k \ge 8$, the condition $(2k)! \le 2\beta H \le 2e^{2M}$ forces $k \log k \le 3M$, hence $L^* \le CM/\log M$. Each of the first L^* terms is at most $\log 2 + \log(\beta H) \le 3M$, and the remaining terms sum to $O(1)$ by the factorial decay. Thus $\sup_w \log \Delta_\beta \le CM^2/\log M$, and

$$Z_H(\beta) \;\ge\; \exp\Big(-\tfrac{1}{2} \sup_w \log \Delta_\beta\Big)$$

gives (a).

(b) Let $p := \lceil \sqrt{H} \rceil$ and $r := e^{-M/(4p)} \in (0,1]$, and restrict the integral to the radial region $R := \{|w_i| \le r \ \forall i\}$, of prior volume $\mu_w(R) = r^H$. On R the tail estimate sharpens to

$$\sigma_k(A)^2 \le H \sum_{j \ge k} \frac{r^{4j}}{(2j)!} \le \frac{2H\, r^{4k}}{(2k)!},$$

so

$$\log \Delta_\beta(w) \le \sum_{k \ge 0} \log\Big(1 + \frac{2\beta H\, r^{4k}}{(2k)!}\Big).$$

For $k \le 3p$, each term is at most $\log(1 + 2\beta H) \le 3M$, contributing at most $(3p+1) \cdot 3M \le CpM$. For $k > 3p$, we have $\beta r^{4k} \le e^{M-3M} = e^{-2M}$, so those terms sum to at most $2H\, e^{-2M} \sum_k 1/(2k)! \le 1$. Hence $\log \Delta_\beta \le CpM + 1$ on R, and

$$Z_H(\beta) \ \ge \ \mu_w(R)\, e^{-\frac{1}{2}(CpM+1)} \ = \ \exp\Big(-\frac{HM}{4p} - \frac{CpM}{2} - \frac{1}{2}\Big) \ \ge \ \exp\big(-C'\sqrt{H}\, M\big),$$

using $H/p \le \sqrt{H}$ and $p \le 2\sqrt{H}$. This gives (b), and the consequence follows since $\min(M^2/\log M, \sqrt{H}M) = m_n M$. □

6.2 Lower bound

Lemma 6.2 (separated/clustered dichotomy)**.** *There is an absolute constant $c_0 \in (0,1]$ with the following property. For $\beta \ge H$ and $M \ge C_0$, set*

$$m := \max\Big(1, \big\lfloor c_0 \min(M/\log M, \sqrt{H})\big\rfloor\Big), \qquad \delta := e^{-M/(8m)}.$$

Call w separated *if some $I \subset \{1, \dots, H\}$ with $|I| = m$ has $|w_i^2 - w_j^2| \ge \delta$ for all distinct $i, j \in I$, and* clustered *otherwise. Then:*

(i) every separated w has $\log \Delta_\beta(w) \ge mM/2$;

(ii) $\mu_w(\{w \text{ clustered}\}) \le e^{-2mM}$.

Consequently

$$Z_H(\beta) \le e^{-mM/4} + e^{-2mM} \qquad \text{and} \qquad F_H(\beta) \ge \frac{mM}{4} - \log 2 \ \ge \ c\, m_n M.$$

Proof. We first record the elementary bound

$$m \log(3m) \le 8 c_0 M : \tag{7}$$

if $m \le c_0 M/\log M$ then $\log(3m) \le 2 \log M$ (for $M \ge C_0$) and (7) follows; if instead $m \le c_0\sqrt{H}$ with $\sqrt{H} \le M/\log M$, then $\log(3m) \le 2\log\sqrt{H} \le 2\log M$ and $\sqrt{H}\log\sqrt{H} \le M$, giving (7) again. Note also $H \ge m^2/c_0^2$, since $m \le c_0\sqrt{H}$ in either branch.

(i) On the separated set, using $\Delta_\beta \ge \prod_{j<m} \beta\sigma_j^2 = \beta^m \Pi_m(A)^2$ and the lower half of Theorem 5.1 with $\mathrm{discr}_m(w) \ge \delta^{m(m-1)/2}$,

$$\log \Delta_\beta(w) \;\ge\; mM - m - \log(m\Downarrow) - m(m-1)\log(1/\delta).$$

Here $m(m-1)\log(1/\delta) \le m^2 \cdot M/(8m) = mM/8$, and

$$\log(m\Downarrow) = \sum_{k<m} \log(2k)! \le m \log(2m)! \le 2m^2 \log(2m) \le 16 c_0\, mM \le \frac{mM}{8}$$

by (7) and $c_0 \le 1/128$. Since $m \le mM/8$ for $M \ge 8$, we get $\log \Delta_\beta \ge mM(1 - \frac{1}{8} - \frac{1}{8} - \frac{1}{8}) \ge mM/2$. (For $m = 1$ the separation condition is vacuous, $\mathrm{discr}_1 = 1$, and the same display gives $\log \Delta_\beta \ge M - 1$.) Hence the separated set contributes at most $e^{-mM/4}$ to $Z_H(\beta)$.

(ii) If w is clustered, a maximal δ-separated subset of $\{t_1, \dots, t_H\} \subset [0,1]$ has fewer than m points, so every t_i lies within δ of it, hence within 2δ of the grid $\delta\mathbb{Z} \cap [0,1]$ points nearest those. Thus there is a set P of at most m grid points with all $t_i \in \bigcup_{g \in P} [g - 2\delta, g + 2\delta]$. The number of choices of P is at most

$$(1 + 1/\delta)^m \le (2/\delta)^m = \exp\bigl(m \log 2 + M/8\bigr) \le \exp\bigl(m + M/8\bigr).$$

For a fixed grid point g, the preimage $\{u \in [-1,1] : |u^2 - g| \le 2\delta\}$ has measure at most $5\sqrt{\delta}$: if $g \le 4\delta$ it is contained in $|u| \le \sqrt{6\delta}$; if $g > 4\delta$ it is two branches of length at most $2\delta/\sqrt{g - 2\delta} \le \sqrt{2\delta}$ each. So for fixed P, each coordinate w_i lies in a set of prior probability at most $3m\sqrt{\delta}$, and

$$\mu_w(\{w \text{ clustered}\}) \le \exp\bigl(m + M/8\bigr)\bigl(3m\sqrt{\delta}\bigr)^H = \exp\Bigl(m + \frac{M}{8} + H\log(3m) - \frac{HM}{16m}\Bigr).$$

By (7) and $c_0 \le 1/512$, $H \log(3m) \le HM/(64m) \le \frac{1}{2} \cdot HM/(32m)$, so the exponent is at most

$$m + \frac{M}{8} - \frac{HM}{32m} \;\le\; m + \frac{M}{8} - \frac{mM}{32c_0^2} \;\le\; -2mM,$$

using $H \ge m^2/c_0^2$ in the middle step and $c_0 \le 1/16$, $M \ge 8$ in the last. Since the integrand is at most 1, the clustered set contributes at most e^{-2mM} to $Z_H(\beta)$.

Finally, $m \ge \frac{c_0}{2} \min(M/\log M, \sqrt{H})$ once $c_0 \min(M/\log M, \sqrt{H}) \ge 2$, and the remaining range of parameters is absorbed into constants as in the preamble to this section; this gives $F_H \ge c\, m_n M$. □

Proof of Theorem 1.1. Combine Lemmas 6.1 and 6.2 with the equivalence of the two displays (Remark 1.2). □

The two regimes of the theorem correspond to which constraint binds in the choice of m: below the crossover the Hermite penalty binds ($m \log m \lesssim M$), above it the width does ($m \lesssim \sqrt{H}$). The dichotomy itself is regime-independent, which is what makes the estimate uniform across the crossover. The lemmas

also locate the mass of the integral. Below the crossover the bounds pinch on the generic fiber: by Lemma 6.2(ii) all but an e^{-2mM}-fraction of the prior is separated, and on the separated set $\log \Delta_\beta$ lies between $mM/2$ and the w-uniform bound of Lemma 6.1(a), which are of the same order in this regime; thus $Z_H(\beta)$ is carried by typical w, and the growth of F_H in $\log \beta$ is the successive freezing of Hermite modes. Above the crossover the radial region of Lemma 6.1(b) has smaller exponent than the generic fiber, and degenerate small-weight configurations carry the integral instead.

6.3 The learning coefficient

Proposition 6.3. *For all $H \geq 1$ and $n \geq H$,*

$$\widehat{\lambda}_H(n) \;\leq\; C\,\frac{\log n}{\log\log n}.$$

Proof. Differentiating under the integral (justified by smoothness in β and boundedness of the integrand),

$$\widehat{\lambda}_H(\beta) = \mathbb{E}_{\widehat{w}}\Big[\frac{1}{2}\sum_j \frac{\beta\sigma_j(A(w))^2}{1+\beta\sigma_j(A(w))^2}\Big],$$

where $\widehat{w}$ is the marginal posterior on hidden weights, $d\widehat{w} \propto \Delta_\beta(w)^{-1/2} d\mu_w$. By (2), pointwise in w,

$$\frac{1}{2}\sum_j \frac{\beta\sigma_j^2}{1+\beta\sigma_j^2} \leq \frac{1}{2}\sum_{k\geq 0} \min\Big(1, \frac{2\beta H}{(2k)!}\Big) \leq \frac{1}{2}\big(L^* + O(1)\big) \leq C\,\frac{\log\beta}{\log\log\beta},$$

with L^* as in Lemma 6.1(a). The bound is uniform in w, hence passes to the posterior expectation. □

This matches the pre-crossover order of m_n. We do not pursue two-sided pointwise control of $\widehat{\lambda}_H$: bounds on F_H constrain only averages of $\widehat{\lambda}_H$ over multiplicative windows in n, and pointwise lower bounds (or a pointwise $O(\sqrt{H})$ upper bound past the crossover) would require localizing the posterior $\widehat{w}$ beyond the determinant estimates above.

7 Discussion

Recap. For the cosine network learning 0, the wide-network and fixed-architecture limits describe different regions of a single joint (H, n) scaling law; an informal account of the phenomenon appeared in [17]. At polynomial n, the leading free energy is set by the finite Hermite band visible at scale n and is independent of H; past the crossover $\log n \asymp \sqrt{H}\log H$, the full analytic germ becomes visible and the fixed-H RLCT scale takes over. For width $H = 1024$ the crossover scale is

$$(2\sqrt{H})! = 64! \approx 1.3 \cdot 10^{89}.$$

A fixed-width RLCT can therefore be the correct $n \to \infty$ endpoint while a different, H-independent statistic is the leading term at every accessible sample size.

The corresponding positive statement: singular-geometry predictions apply at polynomial n whenever the learning problem reduces to a finite-dimensional effective sector—a polynomial network, a deep linear model, an equivariant Fourier/Hermite sector selected by the task. Absent such a reduction, analytic tails contribute width-dependent spectral scales that push the full RLCT regime beyond polynomial data.

7.1 Assumptions: locality, data, and activation

Locality. Restricting to a small fixed neighborhood in parameter space does not remove the phenomenon, because the relevant structure is multiscale: the parity-pure Vandermonde singularity contains collision patterns on a long hierarchy of scales, and a fixed local chart contains all of them unless the chart itself shrinks with H at a comparable quasi-exponential rate.

Data distribution. We use $x \sim \mathcal{N}(0,1)$ because the Hermite basis diagonalizes the calculation. Other smooth one-dimensional input distributions lead to analogous orthogonal-polynomial spectra; the constants and basis change, but the mechanism—smooth activations produce rapidly decaying spectral coefficients, and finite n resolves only the modes above the ridge scale—does not. One genuinely new effect outside our setting: with weights or data of scale $\eta \gg 1$, the Gaussian envelope G acquires singular values spread over $[e^{-\eta^2/2}, 1]$ and becomes an additional source of degeneracy, which we do not analyze here.

Activation function. The cosine is entire, so its Hermite coefficients decay superexponentially, producing the cutoff $L_n \asymp \log n / \log\log n$ and the $(\log n)^2/\log\log n$ law. Analytic activations with finite complex singularities (sigmoid-type) have exponential coefficient decay and a correspondingly different cutoff function. Non-smooth activations such as ReLU have only polynomial spectral decay in many bases; there the separation between the finite-band regime and the RLCT endpoint is much smaller, though the analytic machinery of SLT no longer applies directly.

Polynomial and linear effective models. Polynomial activations, deep linear networks, and other models whose relevant functions span a finite-dimensional polynomial sector have no analytic tail to hide beyond the statistical scale; there the fixed finite-dimensional singularity can already be the relevant one at polynomial n. This is consistent with existing SLT analyses of polynomial and linear models [7, 2, 11].

Depth. The proof is for a shallow two-layer model. Dormant-neuron and small-subnetwork mechanisms in deeper analytic networks locally reduce to sim-

ilar Hermite/Taylor expansions, but depth also introduces additional algebraic structure; in special cases such as deep linear networks the effective model is finite-dimensional from the start.

7.2 Effective finite-band singular learning

At statistical scale n, the learning problem should be analyzed after restricting to the statistically visible field modes. In the cosine model this visible sector is the Hermite band $k \lesssim L_n \asymp \log n / \log\log n$. Below the crossover the leading estimate is spectral: the visible Hermite directions do not yet interact strongly enough to produce the full width-H RLCT. Near and above the crossover, the Vandermonde collision geometry becomes visible and the fixed-H singularity emerges.

This gives a reading of SLT-style measurements in settings such as modular addition [12]: if a separate representation-theoretic, equivariant, or optimization argument shows that learning reduces to a finite Fourier/Hermite/polynomial sector, then polynomial sample sizes lie in a genuinely finite-dimensional singular-learning regime, and the relevant singularity is that of the effective sector rather than of the full analytic network. Results in the style of Abbe–Boix-Adserà–Misiakiewicz [1], where Hermite components and leap complexity are controlled in suitable regimes, are examples of the kind of input needed for such a reduction; see also the scaling arguments of [14].

It also suggests a tropical refinement of the algebraic learning coefficient: assign each Taylor or Hermite coordinate a weight according to its spectral stiffness at inverse temperature n, rather than treating all coordinates as having equal scale. In polynomial sectors the weights are trivial and ordinary SLT is recovered; in analytic models such as the cosine example, the weights produce the finite-band theory of Theorem 1.1 before the unweighted RLCT endpoint appears.

Acknowledgments

This paper was born out of a conversation with Yevgeny Liokumovich, and the author thanks Yevgeny and Chris Elliott for several useful conversations. Lauren Greenspan and Kaarel Hänni helped with drafts of this paper.

This work was supported by Princint (Principles of Intelligence) through funding from grants from the AI Security Institute and Coefficient Giving.

The author used large language model tools (Claude and ChatGPT) extensively in preparing this paper, for drafting and editing the text and in developing a number of the arguments. In particular, many of the explicit formulae for intermediate bounds (including the exact minor expansion of Proposition 4.3), the sharpened upper bound of Remark 5.2, several of the proofs in Sections 5 and 6, and the first proof of the caterpillar reduction in Appendix A were suggested by these tools. All mathematical content, references, and the final manuscript were verified by the author, who takes full responsibility for the content.

A Vandermonde/tropical RLCT for analytic activations

A.1 Dictionary with the main text

This appendix records the combinatorial calculation of the fixed-width RLCT for parity-pure analytic activations, by the same Schur-positivity mechanism as the main text. The main theorem does not depend on this appendix; conversely, the calculation here identifies the exact fixed-H constant behind the post-crossover scale $\Theta(\sqrt{H})$.

Let H be the width and $w = (w_1, \dots, w_H)$ the embedding weights. Near the zero function, the local Taylor matrix has one column per neuron and one row per Taylor monomial. For a Taylor support $T \subset \mathbb{N}$, write

$$M_T(w)_{t,i} = c_t w_i^t, \qquad t \in T, \quad i = 1, \dots, H, \tag{8}$$

where $c_t \neq 0$ are Taylor coefficients; the constants c_t never affect tropical exponents and are suppressed below. For a parity-pure support

$$\operatorname{supp}(\varphi) \subset 2T + \varepsilon, \qquad \varepsilon \in \{0, 1\}, \tag{9}$$

($\varepsilon = 0$ even, $\varepsilon = 1$ odd), we write

$$M_{T,\varepsilon}(w)_{t,i} = w_i^{\varepsilon} x_i^t, \qquad x_i = w_i^2, \quad t \in T, \tag{10}$$

so the even problem is a problem in the variables $x_i = w_i^2$, and the odd problem is the same problem with every full minor multiplied by $\prod_i w_i$.

We use the small parameter $\hbar = \beta^{-1}$. The local determinant factor is

$$\det(I + \hbar^{-1} M^* M)^{-1/2}; \tag{11}$$

changing normalization conventions (replacing $\hbar^{-1}$ by $\hbar^{-B}$, inserting factors of $1/2$ in the loss) rescales the numerical coefficients in the objective below without changing the combinatorial part.

A.2 Singular products and tropical local free energy

The only input about the local free energy used here is the exterior-power identity

$$\det(I + \hbar^{-1} M^* M) = \sum_{k=0}^{H} \hbar^{-k} \, \| \wedge^k M \|^2. \tag{12}$$

Equivalently, if $\sigma_1 \geq \cdots \geq \sigma_H \geq 0$ are the singular values and

$$\pi_k(M) = \prod_{j=1}^{k} \sigma_j(M) = \| \wedge^k M \|, \qquad \pi_0 = 1, \tag{13}$$

then the local free energy depends on the singular products π_k, not on pointwise bounds for individual singular values.

Definition A.1 (tropical singular-product exponents)**.** On a tropical stratum τ, define

$$p_k(\tau) := \mathrm{val}_\tau \| \wedge^k M\|, \qquad p_0(\tau) = 0. \tag{14}$$

Equivalently,

$$p_k(\tau) = \min_{\substack{I\subset[H],\ |I|=k \\ A\subset T,\ |A|=k}} \mathrm{val}_\tau \det(M_{A,I}), \tag{15}$$

because $\| \wedge^k M\|^2$ is a sum of squared $k \times k$ minors. Here $M_{A,I}$ denotes the submatrix with row set A and column set I.

Definition A.2 (local tropical objective)**.** The tropical local free-energy penalty of a stratum is

$$\mathcal{W}(\tau) = \max_{0\le k\le H} \left(\frac{k}{2} - p_k(\tau) \right). \tag{16}$$

If $\mathrm{Vol}(\tau)$ is the volume exponent of the stratum in w-coordinates, the tropical objective is

$$\mathcal{J}(\tau) = \mathrm{Vol}(\tau) + \mathcal{W}(\tau). \tag{17}$$

The local RLCT/free-energy exponent is obtained by minimizing (17) over the tropical leaves.

Remark A.1. *The formula* (16) *is* (12) *in tropical notation: if $p_k(\tau)$ is the exponent of π_k, the kth summand of the determinant has exponent $-k+2p_k(\tau)$, and the power $-1/2$ in* (11) *gives $k/2 - p_k(\tau)$.*

A.3 Completeness of tropical leaves

This section justifies reducing the local integral to a finite list of tropical leaves. It is the standard finite-stratification step, stated explicitly because it is where cancellation loci enter.

Lemma A.2 (finite tropical leaf decomposition)**.** *Let $g_1, \dots, g_N$ be analytic functions on a compact coordinate chart, and suppose the local integrand depends tropically only on the vector* $(\mathrm{val}(g_1), \dots, \mathrm{val}(g_N))$*. There is a finite stratification by semianalytic pieces L such that on each piece every* $\mathrm{val}(g_a)$ *is an affine function of the chosen scale parameters. Moreover, the local integral has leading tropical exponent equal to the minimum of the exponents of these pieces.*

Proof. Choose local coordinates and expand each g_a into finitely many terms up to the order relevant for the fixed compact chart and the desired tropical scale. Subdivide according to which terms have minimal valuation. On the open complement of the equal-valuation and cancellation loci, the initial form of each g_a is a nonzero unit, so its valuation is the generic affine valuation. The exceptional loci are defined by finitely many initial-form equations. Recursing on those lower-dimensional loci gives a finite tree, because at every step either the initial form is fixed and nonzero or one passes to a strictly smaller analytic locus for at least one of finitely many initial forms.

The integral over a finite disjoint union is the sum of the integrals over the pieces; in tropical notation, a finite sum is controlled by the smallest exponent. Hence the leading exponent is the minimum over the leaves. □

Remark A.3. *For Schur polynomials below, positivity prevents hidden cancellation. For general determinants or minors, cancellation loci may exist; Lemma A.2 says only that they become additional leaves. The later Schur calculation is therefore a calculation on a fixed no-cancellation leaf, or on a positive sector where cancellation is impossible. For the cosine model of the main text this discussion is global rather than leaf-by-leaf: by Proposition 4.3, the local integrand is everywhere a positive series in squared Schur polynomials.*

A.4 Schur valuation on a caterpillar

Let

$$0 \le \alpha_1 \le \alpha_2 \le \cdots \le \alpha_h, \qquad x_i \asymp \hbar^{\alpha_i}. \tag{18}$$

This is the 0-caterpillar: the variables are ordered by their rate of approach to the origin, x_1 being the largest variable (smallest exponent). The w-space volume exponent in the parity-reduced variables $x_i = w_i^2$ is

$$\mathrm{Vol}_w(\alpha) = \frac{1}{2}\sum_{i=1}^{h} \alpha_i. \tag{19}$$

Lemma A.4 (Schur valuation)**.** *Let $s_\lambda(x_1, \ldots, x_h)$ be a Schur polynomial with partition $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_h \ge 0$. On the caterpillar* (18)*,*

$$\mathrm{val}_\alpha\, s_\lambda = \sum_{i=1}^{h} \lambda_i\, \alpha_i. \tag{20}$$

That is, the largest parts of λ pair with the smallest exponents, i.e. with the largest variables.

Proof. The Schur polynomial has the positive monomial expansion

$$s_\lambda = \sum_{\mu} K_{\lambda\mu} m_\mu, \qquad K_{\lambda\mu} \in \mathbb{N}, \tag{21}$$

so there is no cancellation and the valuation is the minimum of the monomial valuations appearing. The exponent vectors μ appearing in s_λ lie in the weight polytope of highest weight λ, the convex hull of the permutations of λ. The linear functional $\mu \mapsto \sum_i \mu_i \alpha_i$ with $\alpha_1 \le \cdots \le \alpha_h$ is minimized over this permutohedron at a vertex, and by the rearrangement inequality the minimizing vertex is the *decreasing* arrangement, $\mu = (\lambda_1, \ldots, \lambda_h)$: the largest parts sit on the smallest α_i. Thus $\min_\mu \sum_i \mu_i \alpha_i = \sum_i \lambda_i \alpha_i$. □

Lemma A.5 (generalized Vandermonde valuation)**.** *Let* $A = \{a_1 < \cdots < a_h\} \subset \mathbb{N}$*. Then*

$$D_A(x) = \det(x_i^{a_j})_{1 \le i,j \le h} \tag{22}$$

factors as

$$D_A(x) = \pm \operatorname{Vand}(x)\, s_\lambda(x), \qquad \lambda_i = a_{h+1-i} - (h-i), \tag{23}$$

and on the caterpillar (18)*,*

$$\operatorname{val}_\alpha D_A = \sum_{i=1}^{h} a_{h+1-i}\, \alpha_i. \tag{24}$$

The valuation is cancellation-free: it is attained by the Schur factor via Lemma A.4.

Proof. The factorization (23) is the alternant formula for Schur polynomials. The Vandermonde valuation is

$$\operatorname{val}_\alpha \operatorname{Vand}(x) = \sum_{1 \le i < j \le h} \min(\alpha_i, \alpha_j) = \sum_{i=1}^{h} (h-i)\, \alpha_i, \tag{25}$$

and by Lemma A.4,

$$\operatorname{val}_\alpha s_\lambda = \sum_{i=1}^{h} \lambda_i \alpha_i = \sum_{i=1}^{h} \big(a_{h+1-i} - (h-i)\big)\alpha_i.$$

Adding the two displays gives (24). Both factors are cancellation-free on the caterpillar (the Vandermonde by its product form, the Schur factor by positivity), so the valuation of the product is the sum of the valuations. □

Remark A.6 (indexing convention)**.** *The mnemonic is: larger exponents go on larger variables. Since* $\alpha_1 \le \cdots \le \alpha_h$*, the largest variable is* x_1*, and the dominant monomial of* D_A *has exponent* a_h *on* x_1*, exponent* a_{h-1} *on* x_2*, and so on—which is* (24)*.*

A.5 Parity-pure caterpillar formula

Fix a parity-pure support $2T+\varepsilon$, $\varepsilon \in \{0,1\}$, with $T = \{t_0 < t_1 < t_2 < \cdots\} \subset \mathbb{N}$.

Proposition A.7 (parity-pure caterpillar exponents)**.** *On the* 0*-caterpillar* $0 \le \alpha_1 \le \cdots \le \alpha_H$ *for* $x_i = w_i^2$*, the tropical singular-product exponents are*

$$p_k^{(\varepsilon)}(\alpha) = \sum_{i=1}^{k} \Big(t_{k-i} + \frac{\varepsilon}{2}\Big)\, \alpha_i, \qquad 0 \le k \le H, \tag{26}$$

with the empty sum for $k = 0$ *equal to* 0*. Consequently the caterpillar objective is*

$$\boxed{\Lambda^{\mathrm{cat}}_{H,\varepsilon}(T) = \inf_{0 \le \alpha_1 \le \cdots \le \alpha_H} \left\{ \frac{1}{2} \sum_{i=1}^{H} \alpha_i + \max_{0 \le k \le H} \left(\frac{k}{2} - \sum_{i=1}^{k} \Big(t_{k-i} + \frac{\varepsilon}{2}\Big)\, \alpha_i \right) \right\}.} \tag{27}$$

Proof. A $k \times k$ minor with row exponents $A = \{a_1 < \cdots < a_k\} \subset T$ and column set $I = \{i_1 < \cdots < i_k\}$ has valuation

$$\frac{\varepsilon}{2} \sum_{j=1}^{k} \alpha_{i_j} + \sum_{j=1}^{k} a_{k+1-j}\, \alpha_{i_j}$$

by Lemma A.5 (the odd case contributing $\varepsilon/2$ per chosen column through the factor $\prod w_{i_j} = \prod x_{i_j}^{1/2}$). All coefficients are nonnegative, so the minimizing row set is the first k exponents of T and the minimizing column set is $I = \{1, \ldots, k\}$, giving (26) via $a_{k+1-j} = t_{k-j}$. Substituting into (17) with the volume (19) gives (27). □

Remark A.8 (where the caterpillar reduction enters)**.** *Proposition A.7 is a complete calculation once one has reduced to* 0*-caterpillar leaves. The independent combinatorial step is that nonzero collision branches are dominated by moving them to the distinguished point* 0*. In a tree model where a nonzero collision of a cluster C at scale δ costs $(|C|-1)\delta$ in w-volume, while a zero-radius condition costs $|C|\delta/2$, this is the pair-counting inequality*

$$\binom{|C|}{2}\delta \;\geq\; (|C|-1)\,\delta,$$

enabled by the additional factors $w_i + w_j$ in $w_i^2 - w_j^2$; Schur positivity ensures the correction factors cannot cancel against it. This is the reduction used to pass from the full tropical leaf decomposition to the finite objective (27) *in the parity-pure case.*

Theorem A.9 (parity-pure RLCT formula)**.** *For a parity-pure analytic activation with Taylor support $2T+\varepsilon$, $\varepsilon \in \{0,1\}$, the local fixed-width RLCT exponent for learning the zero function is given by the caterpillar variational problem* (27)*. In particular the cosine and odd/tanh cases are obtained by taking respectively $(\varepsilon, T) = (0, \mathbb{N})$ and $(\varepsilon, T) = (1, \mathbb{N})$.*

Proof sketch. Lemma A.2 reduces the local integral to finitely many tropical leaves. For parity-pure support one passes to $x_i = w_i^2$, the odd case contributing the extra factor w_i in each column. On a 0-caterpillar leaf, Proposition A.7 gives the displayed objective. The remaining step is the domination of nonzero collision branches by the distinguished point 0 as in Remark A.8: each factor $w_i^2 - w_j^2 = (w_i - w_j)(w_i + w_j)$ supplies the additional vanishing needed for the pair-counting comparison, while Schur positivity prevents cancellation. Thus the minimizing leaves are represented by the 0-caterpillar objective. □

A.6 Exact formulas for cos and tanh

The two standard parity-pure examples have $T = \mathbb{N}$ after reducing by parity.

A.6.1 Cosine

From

$$\cos z = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} z^{2n} \tag{28}$$

we have $\varepsilon = 0$ and $T = \mathbb{N}$, so $t_j = j$ and Proposition A.7 gives

$$p_k^{\cos}(\alpha) = \sum_{i=1}^{k} (k-i)\,\alpha_i. \tag{29}$$

Hence the cosine caterpillar exponent is

$$\boxed{\Lambda_H^{\cos,\mathrm{cat}} = \inf_{0\le\alpha_1\le\cdots\le\alpha_H} \left\{ \frac{1}{2}\sum_{i=1}^{H}\alpha_i + \max_{0\le k\le H}\left(\frac{k}{2} - \sum_{i=1}^{k}(k-i)\alpha_i\right)\right\}.} \tag{30}$$

The radial subfamily $\alpha_1 = \cdots = \alpha_H = a$ gives the explicit upper bound

$$\Lambda_H^{\cos,\mathrm{cat}} \le \inf_{a\ge 0}\left\{ \frac{Ha}{2} + \max_{0\le k\le H}\left(\frac{k}{2} - \frac{a\,k(k-1)}{2}\right)\right\}, \tag{31}$$

a finite maximum of affine functions of a, convenient for comparison with known exact RLCT formulas.

A.6.2 Hyperbolic tangent

From

$$\tanh z = \sum_{n=1}^{\infty} \frac{2^{2n}(2^{2n}-1)B_{2n}}{(2n)!}\, z^{2n-1} = z - \frac{z^3}{3} + \frac{2z^5}{15} - \cdots \tag{32}$$

all odd powers occur, so $\varepsilon = 1$ and $T = \mathbb{N}$. Therefore

$$p_k^{\tanh}(\alpha) = \sum_{i=1}^{k}\left(k - i + \frac{1}{2}\right)\alpha_i, \tag{33}$$

and

$$\boxed{\Lambda_H^{\tanh,\mathrm{cat}} = \inf_{0\le\alpha_1\le\cdots\le\alpha_H} \left\{ \frac{1}{2}\sum_{i=1}^{H}\alpha_i + \max_{0\le k\le H}\left(\frac{k}{2} - \sum_{i=1}^{k}\left(k-i+\frac{1}{2}\right)\alpha_i\right)\right\}.} \tag{34}$$

The radial subfamily gives

$$\Lambda_H^{\tanh,\mathrm{cat}} \le \inf_{a\ge 0}\left\{ \frac{Ha}{2} + \max_{0\le k\le H}\left(\frac{k}{2} - \frac{a\,k^2}{2}\right)\right\}. \tag{35}$$

In the continuous relaxation of (35), the maximum over k is approximately $1/(8a)$, giving the scale $a \sim (2\sqrt{H})^{-1}$ and the size $\Lambda_H \sim \frac{1}{2}\sqrt{H}$; the exact discrete expression is the finite minimization above. Up to the normalization conventions of the local determinant, this is the Aoyagi–Watanabe odd/tanh formula [3].

A.7 A non-parity-pure squeeze

For activations whose Taylor support is not purely even or odd, we record a bound that is weaker than the parity-pure equality but robust, using the same Schur-positivity mechanism. Let $T \subset \mathbb{N}$ be a general Taylor support, and define the ordinary one-sided caterpillar functional in the variable $u = |w|$ by

$$\Lambda_m^{\mathrm{ord}}(T) := \inf_{0 \le r_1 \le \cdots \le r_m} \left\{ \sum_{i=1}^{m} r_i + \max_{0 \le k \le m} \left(\frac{k}{2} - \sum_{i=1}^{k} t_{k-i}\, r_i \right) \right\}, \tag{36}$$

where $T = \{t_0 < t_1 < \cdots\}$; the volume is $\sum r_i$ because $u_i = |w_i|$ rather than $x_i = w_i^2$.

Proposition A.10 (sign-majority lower bound)**.** *Assume that on each sign orthant the coefficients of the relevant generalized Vandermonde minors are nonzero. Let $h = \lceil H/2 \rceil$. On every sign orthant in $\mathbb{R}^H$ there is a subset $J \subset [H]$ of size h on which all signs $\mathrm{sgn}(w_i)$ agree, and the h-column minor restricted to J has the tropical valuation of a positive generalized Vandermonde in $u_i = |w_i|$. Consequently, on each orthant the exterior-power norm has the lower bound coming from the h-width ordinary caterpillar problem, and*

$$\Lambda_H^{\mathrm{nonparity}} \gtrsim \Lambda_{\lceil H/2 \rceil}^{\mathrm{ord}}(T), \tag{37}$$

with equality of constants if the same normalization of $\hbar$ and volume variables is used.

Proof. Fix a sign orthant. Among H signs, at least $h = \lceil H/2 \rceil$ share the modal sign; call this subset J. On J, write $w_i = s u_i$ with one fixed $s \in \{\pm 1\}$ and $u_i \ge 0$. For row exponents $A = \{a_1 < \cdots < a_k\} \subset T$,

$$\det(w_i^{a_j})_{i \in I,\, a_j \in A} = \Big(\prod_{j=1}^{k} s^{a_j} \Big) \det(u_i^{a_j})_{i \in I,\, a_j \in A},$$

with nonzero sign prefactor, so the valuation is that of a positive generalized Vandermonde in u, given by Lemma A.5 and Schur positivity as the ordinary formula in (36). Since $\| \wedge^k M \|^2$ is a sum of squared minors, any one minor lower-bounds the exterior-power norm and hence upper-bounds its valuation; substituting into (16) and minimizing over strata gives the stated bound. The finite decomposition into sign orthants changes the integral only by a finite sum, hence not the leading exponent. □

Proposition A.11 (full-width upper comparison)**.** *For a general support T,*

$$\Lambda_H^{\mathrm{nonparity}} \lesssim \Lambda_H^{\mathrm{ord}}(T). \tag{38}$$

Proof. Restrict the integral to the all-positive orthant and to the 0-caterpillar leaves in the variables $u_i = w_i \ge 0$. On this orthant all generalized Vandermonde minors are positive-sector determinants and Lemma A.5 applies; the

caterpillar family has positive measure at the tropical scale under consideration, so it lower-bounds the local integral, which upper-bounds the free-energy exponent. Possible additional cancellation loci outside this family can only add singular mass, so they cannot invalidate the comparison. □

Corollary A.12 (coarse squeeze)**.** *Under the nondegeneracy assumptions above, a non-parity-pure width-H model satisfies, up to the normalization constants separating u and $x = w^2$ variables,*

$$\Lambda^{\mathrm{ord}}_{\lceil H/2 \rceil}(T) \;\lesssim\; \Lambda^{\mathrm{nonparity}}_{H} \;\lesssim\; \Lambda^{\mathrm{ord}}_{H}(T). \tag{39}$$

Corollary A.13 (coarse analytic scale)**.** *Suppose the Taylor support contains an infinite arithmetic progression of step q. In the parity-pure case, and in the mixed-parity case under the nondegeneracy assumptions of Proposition A.10, the fixed-width zero-target RLCT has scale*

$$\Lambda_H = \Theta_q(\sqrt{H}),$$

with constants depending on the progression step and normalization conventions but not on H.

Proof sketch. Restricting to the arithmetic progression reduces the ordinary Vandermonde functional to exponents $a + qj$, rescaling the radial/caterpillar optimization by constants depending only on q. The radial test path gives the $O_q(\sqrt{H})$ upper bound; the singular-product/volume objective gives the matching lower bound. For mixed parity, Corollary A.12 compares the width-H problem to ordinary problems at widths $\lceil H/2 \rceil$ and H, both of the same $\sqrt{H}$ scale. □

Remark A.14 (status of the squeeze)**.** *The sign-majority argument is a genuine bound on exterior-power norms: on each orthant at least $\lceil H/2 \rceil$ columns share a sign, and on those columns the generalized Vandermonde has the positive Schur valuation. It is not an equality theorem for non-parity-pure activations: mixed signs may create additional cancellations or additional good minors, so* (39) *should be used as a squeeze rather than a formula.*

B Bayesian learning and population free energy

Let $K_D(\theta) = \sum_{i=1}^{n} K(\theta, x_i)$ be the empirical loss and $K(\theta) = \mathbb{E}_x K(\theta, x)$ the population loss. The population partition function at inverse temperature n is

$$Z_{pop}(n) = \int e^{-nK(\theta)}\, d\mu(\theta), \qquad F_{pop}(n) = -\log Z_{pop}(n).$$

Under the boundedness, analyticity, and concentration hypotheses of Watanabe's Bayesian learning theory [19], the quenched empirical free energy differs from the population free energy by lower-order terms after subtracting the minimum-loss contribution; in fixed analytic charts the standard form of this

error is logarithmic or iterated-logarithmic in n, depending on the concentration statement and normalization. The present paper therefore works with F_{pop} as the deterministic object controlling the leading scales. The main estimates, $(\log n)^2/\log\log n$ and $\sqrt{H}\log n$, are much larger than these empirical–population correction terms in the regimes where the comparison is invoked.

The finite-temperature learning coefficient used in empirical SLT is the logarithmic derivative

$$\widehat{\lambda}(n) = \frac{\partial F_{pop}(n)}{\partial \log n} = n\,\mathbb{E}_{\theta\sim p_n}\left[K(\theta) - K^*\right],$$

where $p_n(\theta) \propto e^{-nK(\theta)}d\mu(\theta)$ and $K^* = \inf_\theta K(\theta)$; for the zero-target problem $K^* = 0$.

C Priors, random kernels, and small data

The body of the paper uses the Gaussian readout prior with density $e^{-a_i^2}/\sqrt{\pi}$ because it makes the a-integral exactly Gaussian. Replacing this prior by the hard constraint $|a_i| \le 1$ changes the free energy only by lower-order terms at the precision considered here: on $|a_i| \le 1$ the Gaussian density is bounded above and below by absolute constants, while the Gaussian tail outside $[-1,1]^H$ is suppressed once the quadratic loss term is present. When this comparison is applied after the Hermite cutoff, only $O(L_n)$ readout directions are active, so the resulting free-energy error is lower order relative to the main scales.

For $n \le H$, one expects the width-H random feature kernel on the sampled data points to be close to its infinite-width limit, provided the usual random-feature concentration estimates [13, 4] hold uniformly at the required scale. Under that additional concentration input, the same Hermite cutoff calculation suggests

$$F_H(n) = \Theta\left(\frac{(\log(nH+5))^2}{\log\log(nH+5)}\right).$$

We use this only as a supporting small-data estimate; the main theorem is the deterministic population calculation for $n \ge H$.

References

[1] Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. *ArXiv*, abs/2302.11055, 2023.

[2] Miki Aoyagi. Singular leaning coefficients and efficiency in learning theory. *ArXiv*, abs/2501.12747, 2025.

[3] Miki Aoyagi and Sumio Watanabe. Resolution of singularities and the generalization error with Bayesian estimation for layered neural network. *IEICE Transactions*, J88-D-II(10):2112–2124, 2005. In Japanese.

[4] Haim Avron, Mikhail Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. *ArXiv*, abs/1804.09893, 2018.

[5] Garrett Baker, George Wang, Jesse Hoogland, and Daniel Murfet. Structural inference: Interpreting small language models with susceptibilities. *ArXiv*, abs/2504.18274, 2025.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[7] Ben Cullen, Sergio Estan-Ruiz, Riya Danait, and Jiayi Li. A basin-selection perspective on grokking via singular learning theory, 2026. Preprint.

[8] James Halverson, Anindita Maiti, and Keegan Stoner. Neural networks and quantum field theory. *Machine Learning: Science and Technology*, 2, 2020.

[9] Jessica N. Howard, Ro Jefferson, Anindita Maiti, and Zohar Ringel. Wilsonian renormalization of neural network Gaussian processes. *Machine Learning: Science and Technology*, 6, 2024.

[10] Edmund Lau, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[11] Simon Pepin Lehalleur and Richárd Rimányi. Geometry of fibers of the multiplication map of deep linear neural networks, 2024. Preprint.

[12] Nina Panickssery and Dmitry Vaintrob. Investigating the learning coefficient of modular addition: Hackathon project. `https://www.lesswrong.com/posts/4v3hMuKfsGatLXPgt/investigating-the-learning-coefficient-of-modular-addition`, October 2023. LessWrong post.

[13] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, 2007.

[14] Noa Rubin, Orit Davidovich, and Zohar Ringel. Mitigating the curse of detail: Scaling arguments for feature learning and sample complexity. *ArXiv*, abs/2512.04165, 2025.

[15] Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *International Conference on Learning Representations*, 2023.

[16] Ray J. Solomonoff. A formal theory of inductive inference. Part I. *Information and Control*, 7(1):1–22, 1964.

[17] Dmitry Vaintrob. Learning zero and what SLT gets wrong about it. https://www.lesswrong.com/posts/5hKgJy8rcqnM9ntp2/learning-zero-and-what-slt-gets-wrong-about-it, 2024. LessWrong.

[18] Sumio Watanabe. Learning efficiency of redundant neural networks in Bayesian estimation. *IEEE Transactions on Neural Networks*, 12(6):1475–1486, 2001.

[19] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2009.

[20] Sumio Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897, 2013.