Tropical Boundary Complexity of Deep ReLU Binary Classifiers

Jun Li
School of Computer Science, Australian Artificial Intelligence Institute,
Faculty of Engineering and IT, University of Technology Sydney
Jun.Li@uts.edu.au

Abstract

We derive a recursion in depth for the distribution of the squared Frobenius norm of the zonotope generator matrices that govern the decision boundary of a random Gaussian deep ReLU binary classifier. For a network of depth L, let $\Gamma_L = \mathbb{E}[Z_L]$ be the expectation of the squared norm Z_L over the distribution of the network parameters. We show that for weight variances $(\sigma_l^2)_{l=1}^L$, $\Gamma_L = 8(1+1/\pi)\prod_{l=1}^L \sigma_l^2/2$, which holds at finite width and for any bias variance. Our computation of $\mathrm{Var}(Z_L)$ shows that Z_L of a single network concentrates around Γ_L when the hidden widths are large compared with the depth. Via Rice's formula, Γ_L and the variance of the network output give an estimate of how often a great circle on the input sphere crosses the decision boundary. The estimate approaches the exact crossing rate, with a relative gap of the order of the sum of the reciprocal hidden widths. Monte Carlo experiments agree with the analysis.

1 Introduction

A central theoretical question about deep neural networks is the expressivity of the function space they span, and how depth shapes it. A single perceptron computes a globally affine function, producing a flat decision boundary, and fails on simple patterns like XOR. One hidden layer changes the model qualitatively: the activation pattern of the hidden units partitions the input space into cells, on each of which the network acts as a distinct affine map, and the decision boundary becomes piecewise-linear with arbitrarily many pieces. A natural way to quantify the gain from depth is to count the linear regions a network can produce. For a piecewise-linear network with input dimension n_0 and n hidden units in total, hyperplane-arrangement counting [26] yields the upper bound $O(n^{n_0})$ [3]. Montúfar et al. [14] showed the number of linear regions can grow exponentially with depth, but a region count is a proxy for input processing and does not see the decision boundary. Alfarra et al. [2], Zhang et al. [27] characterized the boundary algebraically through tropical geometry, but the explicit zonotope description is confined to depth two. Poole et al. [19], Schoenholz et al. [23] push ensemble averages through arbitrary depth using mean-field tools, but the observables they track are pre-activation moments, not boundary geometry.

This work computes a geometric observable of the binary classification boundary for ReLU networks of arbitrary depth. The observable is determined by a ratio. The numerator is derived from the zonotope generators with which tropical geometry describes the decision boundary of 2-layer ReLU networks [2]. For a multi-layer network, we consider the squared norm of the zonotope generators and take expectation over random networks and inputs. We call the quantity the *expected generator energy*, Γ_L. The denominator is the output variance $q^{(L)}$ given by the mean-field recursion of Poole et al. [19], Schoenholz et al. [23].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

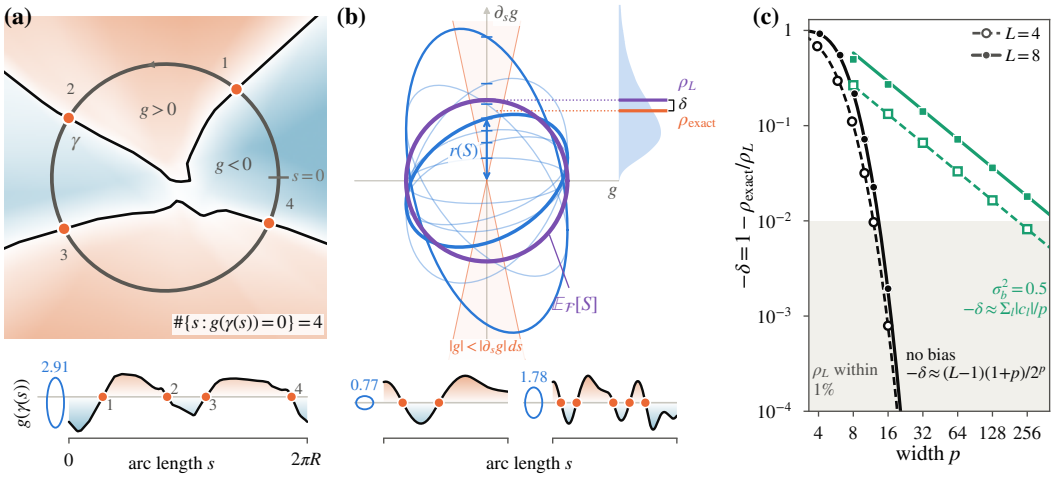


Figure 1: Complexity measured by zero-crossings. (a) A great circle γ in the input space of a random ReLU network. Colours show the log-odds $g = f_1 - f_2$ between the two classes, and the black line is the decision boundary $g = 0$. (b) The plane of g and its rate $\partial_s g$ along the arc. A large rate relative to $|g|$ makes a zero crossing nearby more likely (orange cone). Given the weights below the output layer, the pair $(g, \partial_s g)$ is Gaussian over the output-layer parameters, and each blue ellipse is its covariance for one network. $r(\cdot)$ measures the ratio $|\frac{\partial_s g}{g}|$ within a Gaussian and determines the density of zero crossings on the arc. Estimating the mean of r over the lower layers by r of the averaged ellipse (purple) has an error δ that vanishes as the network width grows (c). See Theorem 3.4 and Figures 9 and 8 in the Appendix for more details.

Specifically, the scalar Γ_L is computed as follows. For a ReLU network with L layers, the activation pattern of the lower $L-2$ layers partitions the input space into cells, and when restricted to one cell, the network acts as a 2-layer effective ReLU network applied to a *linear* function of the input. We construct the 2-layer ReLU generators as in Alfarra et al. [2] on the effective network and take expectation over the inputs and random Gaussian weights of the lower layers. We show that Γ_L satisfies a one-step recursion in depth: prepending a layer multiplies Γ_L by a factor proportional to the weight variance of the new layer.

The ratio measures the complexity of the decision boundary as follows. Take two points of equal norm $\boldsymbol{x}, \boldsymbol{x}'$ in the input space, and let $\gamma(\boldsymbol{x}, \boldsymbol{x}')$ be the great-circle arc joining the points. We show that the expected number of times γ crosses the decision boundary per unit length is linked to $\sqrt{\Gamma_L / q^{(L)}}$ by relating the ratio to Rice's formula [20, 21] (Figure 1).

Using the fourth moments of the generators in the tropical construction, we show that the variance of the generator energy can be controlled by the width-to-depth ratio, so the generator energy of an individual network concentrates around Γ_L.

The multi-layer development of the zonotope generator study of Alfarra et al. [2] in this work also connects to existing analyses of ReLU networks. Under the initialization of He et al. [8], with all weight variances $\sigma_w^2 = 2$, Γ_L stays the same for arbitrary L. The setting also keeps the gradients from exploding or vanishing as in Schoenholz et al. [23]. The concentration result in this work extends the properties from ensemble analysis to individual networks.

2 Related work

The decision boundary of a piecewise-linear network is a tropical hypersurface [27]; for depth-two networks with two outputs, Alfarra et al. [2] showed it lies in the normal fan of the convex hull of two zonotopes whose generators are given explicitly by the weights. Generalizing this explicit form to $L > 2$ is non-trivial: composing further ReLU layers turns the generators into piecewise functions of the input on the arrangement induced by

the lower layers. Other tropical-geometric studies of neural networks are surveyed and extended in Brandenburg et al. [4], Maragos et al. [12]. Our contribution is to extract an averaged generator-Frobenius quantity Γ_L from the piecewise picture and compute its mean and variance for any L.

Random feedforward networks with i.i.d. Gaussian weights behave, in the wide-network limit, layerwise as Gaussian processes [10, 13, 16], enabling layerwise recursions for the moments of pre-activations. Poole et al. [19] derived recursions for the per-neuron variance $q^{(l)}$ and the dual correlation between two inputs; Schoenholz et al. [23] read off depth scales ξ_q, ξ_c and showed the trainability boundary for tanh networks coincides with $L = 6\xi_c$. The same toolset underlies the dynamical-isometry analysis of Pennington et al. [18] and recent quantitative refinements of the wide-network Gaussian-process limit [1]. We apply expectation recursions to the second moment of the input gradient, which is proportional to Γ_L (30).

A separate line measures expressivity by the number of affine pieces of the network function: constructive lower bounds exponential in depth [14], sharper upper and lower bounds via mixed-integer programming [24], and average-case results showing that the expected number of pieces along a line is proportional to the total neuron count N, with the distance from an input to the nearest piece boundary staying of order $1/N$ throughout MNIST training [7]. Recent work refines local complexity [17] and settles the computational complexity of counting itself [25]. The boundary $\{f_1 = f_2\}$ studied here is a codimension-one set that crosses the interiors of these pieces, and Γ_L together with $q^{(L)}$ gives the estimate (19) of how often a great circle crosses it.

A third line studies wide networks through their infinite-width limit. With appropriate parametrization and $\min_l p_l \to \infty$, gradient descent becomes linear in the parameters and is governed by a deterministic neural tangent kernel [9]; subsequent work extends the analysis to training dynamics, the regime in which the linearization is exact, and finite-width corrections [5, 6, 11]. Our results are of a different kind: Γ_L is a finite-width ensemble average of a geometric quantity at initialization. Rice's formula does not need the Gaussian-process limit either: given the parameters below the output layer, g along a great circle is a Gaussian process at every width, and the estimate $\rho_L \propto \sqrt{\Gamma_L / q^{(L)}}$ approaches the exact crossing rate as the hidden widths grow.

On the relationship between init-time quantities and trained networks, Schoenholz et al. [23] predict a binary trainability outcome from the init-time depth scale alone; Hanin and Rolnick [7] report a continuous quantity (piece count and distance to nearest piece boundary, normalized by N) that stays in a fixed band throughout MNIST training. Our experiments target a continuous boundary observable defined by Rice's formula, and Monte Carlo measurements at initialization agree with the closed-form prediction.

3 Boundary complexity of deep ReLU ensembles

Random ReLU ensembles. We consider the random ensemble of depth-L fully connected ReLU binary classifiers. A network in the ensemble has $L-1$ ReLU hidden layers followed by a linear two-dimensional output layer. Writing $\boldsymbol{h}^{(0)} := \boldsymbol{x} \in \mathbb{R}^{p_0}$ for the input and $\boldsymbol{h}^{(l)}$ for the post-activation at layer l,

$$h_i^{(l)} = \max\Bigg(\sum_{j=1}^{p_{l-1}} W_{ij}^{(l)} h_j^{(l-1)} + b_i^{(l)},\ \ 0 \Bigg), \qquad l = 1, \ldots, L-1,\ \ i = 1, \ldots, p_l, \tag{1}$$

and the output layer is linear,

$$f_k(\boldsymbol{x}) \ = \ \sum_{j=1}^{p_{L-1}} W_{kj}^{(L)}\, h_j^{(L-1)} + b_k^{(L)}, \qquad k = 1, 2. \tag{2}$$

The widths are $p_0, p_1, \ldots, p_{L-1}$, with $p_L = 2$ fixed by the binary-classification setup, as in Alfarra et al. [2]. The parameters of the network follow independent centered Gaussian

distributions.

$$W_{ij}^{(l)} \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\, \frac{\sigma_l^2}{p_{l-1}}\right), \qquad b_i^{(l)} \overset{\text{iid}}{\sim} \mathcal{N}\big(0,\, \sigma_{b,l}^2\big)\,. \tag{3}$$

A network ensemble is parameterized by the per-layer variance of the parameters, $(\sigma_l,\, \sigma_{b,l})_{l=1}^{L}$. The variance scaling σ_l^2/p_{l-1} is the standard fan-in normalization.

Decision boundary in tropical geometry. We first introduce necessary notions related to the decision boundary of a ReLU binary classifier from a tropical geometry perspective. For an individual network, the decision boundary is the point set in the input space where the two output units have the same value

$$\mathcal{B} \;:=\; \{\, \boldsymbol{x} \in \mathbb{R}^{p_0} \;:\; f_1(\boldsymbol{x}) = f_2(\boldsymbol{x}) \,\}, \tag{4}$$

which is piecewise linear. For a two-layer network ($L = 2$), the linear pieces of $\mathcal{B}$ have a geometric characterization. Alfarra et al. [2], Zhang et al. [27] formulate $f_{1,2}$ in terms of tropical polynomials where the max operator corresponds to the tropical sum and the weights become tropical exponents. They show that every linear piece of the boundary is orthogonal to an edge of the convex hull of two *zonotopes*, where the *generators* of the zonotopes are the rows of two matrices $\mathbf{G}_1, \mathbf{G}_2 \in \mathbb{R}^{p_1 \times p_0}$, constructed from $\mathbf{A} := \mathbf{W}^{(1)} \in \mathbb{R}^{p_1 \times p_0}$ and $\mathbf{B} := \mathbf{W}^{(2)} \in \mathbb{R}^{2 \times p_1}$ (shorthand for the penultimate and output layer weights):

$$(\mathbf{G}_k)_{ji} = \beta_{k,j}\, A_{ji}, \tag{5}$$

$$\beta_{1,j} := (B_{1j})^+ + (B_{2j})^-,$$

$$\beta_{2,j} := (B_{2j})^+ + (B_{1j})^-, \tag{6}$$

where $a^+ := \max(a,\, 0)$ and $a^- := \max(-a,\, 0)$. It is helpful to associate each row with one neuron in the penultimate layer.

We call $\|\mathbf{G}_1\|_F^2 + \|\mathbf{G}_2\|_F^2$, the sum of squared lengths of the generators, the generator energy. Several quantities are functions of the network parameters or of the input; e.g., $\beta_{k,j}$ depends on $\mathbf{B}$. We sometimes present the dependence explicitly, e.g. as $\beta_{k,j}(\mathbf{B})$ for clarity.

Expected generator energy of random ReLU ensembles. In the random ensemble (3), the generator energy $\|\mathbf{G}_1\|_F^2 + \|\mathbf{G}_2\|_F^2$ depends on the stochastic weights through $\mathbf{G}_1, \mathbf{G}_2$ as defined by (5)–(6), and is therefore a random variable. We define the expected generator energy of the ensemble as its expectation:

$$\Gamma_2 \;:=\; \mathbb{E}_{\mathbf{A},\mathbf{B}}\Big[\, \|\mathbf{G}_1\|_F^2 + \|\mathbf{G}_2\|_F^2 \,\Big]\,. \tag{7}$$

The expectation is taken over the parameters of the network. Γ_2 depends on the ensemble parameters $\sigma_{w,1}^2, \sigma_{w,2}^2$ (not on the bias variances $\sigma_{b,1}^2, \sigma_{b,2}^2$).

Expected generator energy at depth $L > 2$. For $L > 2$, the input space is partitioned by the bottom $L-2$ hidden layers into open regions. On each of these regions, the network is a depth-2 network applied to a linear function of the input. Fix a sign pattern of the pre-activations on layers $1, \ldots, L-2$, and let $\mathcal{C} \subset \mathbb{R}^{p_0}$ be the open set of inputs realizing it (when $L = 2$, $\mathcal{C} = \mathbb{R}^{p_0}$). On $\mathcal{C}$, the activation indicators of these layers are constant; the network therefore restricts to the depth-2 network with weights $(\mathbf{A}, \mathbf{B}) := (\mathbf{W}^{(L-1)},\, \mathbf{W}^{(L)})$ acting on $\Phi_{\mathcal{C}}^{(L-2)}\, \boldsymbol{x}$, with depth-2 generators $\mathbf{G}_k(\mathbf{A}, \mathbf{B})$ as in (5)–(6), where $\Phi_{\mathcal{C}}^{(L-2)} \in \mathbb{R}^{p_{L-2} \times p_0}$ is the linear map realized by these layers on $\mathcal{C}$. The generators inherited on $\mathcal{C}$ are

$$\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)} \;:=\; \mathbf{G}_k(\mathbf{A}, \mathbf{B})\, \Phi_{\mathcal{C}}^{(L-2)}, \qquad k = 1,\, 2, \tag{8}$$

which depend on $\mathcal{C}$ through $\Phi_{\mathcal{C}}^{(L-2)}$. We define the generator energy of one network and its expectation at depth L as

$$Z_L \;:=\; \Big\|\tilde{\mathbf{G}}_{1,\mathcal{C}}^{(L)}\Big\|_F^2 + \Big\|\tilde{\mathbf{G}}_{2,\mathcal{C}}^{(L)}\Big\|_F^2\,, \tag{9}$$

$$\Gamma_L \;:=\; \mathbb{E}_{\mathcal{C},\mathbf{A},\mathbf{B}}\,[Z_L], \tag{10}$$

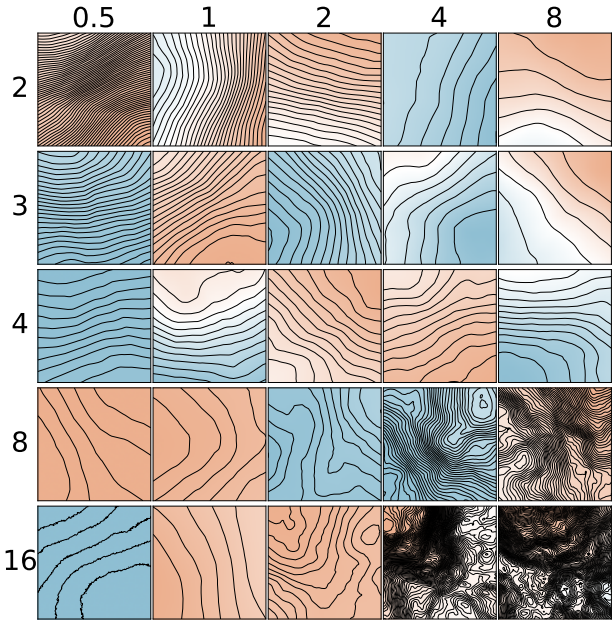


Figure 2: Level curves of $g(\boldsymbol{x}) := f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})$ on a 2D patch from a sphere $\mathbb{S}^{49}$. Rows: depth L; columns: weight variance σ_w^2. Shade is g: peach for $g > 0$, blue for $g < 0$, white at $g = 0$, with one colour scale per panel. Contour spacing fixed per column, calibrated to $L = 4$ for visual clarity.

where the expectation is over the random initialization and the cells. The cell $\mathcal{C}$ is determined by the lower-layer weights $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L-2)}$ and the input $\boldsymbol{x}$; $\mathbb{E}_{\mathcal{C}}$ marginalizes over both. The $\boldsymbol{x}$-contribution further drops out in the proof of Theorem 3.1. We therefore omit the cell subscript and write $\tilde{\mathbf{G}}_k^{(L)}$ in expressions of Γ_L. For $L = 2$ the only cell is $\mathbb{R}^{p_0}$, $\Phi_{\mathcal{C}}^{(0)}$ is the identity, and (10) reduces to (7).

$\left\|\tilde{\mathbf{G}}_k^{(L)}\right\|_F$ and expressivity Fix a cell $\mathcal{C}$, on which $\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)} = \mathbf{G}_k\, \Phi_{\mathcal{C}}^{(L-2)}$ ((8)). The depth-2 generator $\mathbf{G}_k = \mathbf{G}_k(\mathbf{A}, \mathbf{B})$ of (5) is built from $\mathbf{A} := \mathbf{W}^{(L-1)}$ and $\mathbf{B} := \mathbf{W}^{(L)}$; its p_{L-1} rows are indexed by the neurons of the last hidden layer, the j-th row encoding how neuron j contributes to $f_1 - f_2$ via its ReLU toggle. Concretely, $(\mathbf{G}_k)_{j,:} = \beta_{k,j}\, \mathbf{A}_{j,:}$, where $\beta_{k,j}$ is a scalar built from $\mathbf{B}$; right-multiplication by $\Phi_{\mathcal{C}}^{(L-2)}$ propagates this to $(\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)})_{j,:} = \beta_{k,j}\, (\mathbf{A}\, \Phi_{\mathcal{C}}^{(L-2)})_{j,:}$.

The Frobenius norm therefore decomposes into per-neuron contributions,

$$\left\|\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)}\right\|_F^2 \;=\; \sum_{j=1}^{p_{L-1}} \beta_{k,j}^2 \left\|(\mathbf{A}\, \Phi_{\mathcal{C}}^{(L-2)})_{j,:}\right\|_2^2. \tag{11}$$

The two factors of the j-th summand admit separate readings: $|\beta_{k,j}|$ is the jump that neuron j's ReLU toggle contributes to $f_1 - f_2$, and $\left\|(\mathbf{A}\, \Phi_{\mathcal{C}}^{(L-2)})_{j,:}\right\|_2$ is the rate at which its pre-activation varies as $\boldsymbol{x}$ moves within $\mathcal{C}$.

Fig. 2 illustrates the level curves of g. Each panel represents a random ReLU classifier, with depth $L \in \{2, 3, 4, 8, 16\}$ (rows) and weight variance $\sigma_w^2 \in \{0.5, 1, 2, 4, 8\}$ (columns), evaluated on the same fixed 2D patch from $\mathbb{S}^{49}$. For visual clarity, a per-column scaling is used in the figure so that at $L = 4$ the panel contains about a dozen contour lines. Reading down a column shows how σ_w^2 controls the depth scaling of the expected generator energy; reading along the $L = 4$ row shows how it controls the contour shape.

We now compute the depth dependence of Γ_L given the ensemble parameters. It has a closed form at $L = 2$ and obeys a one-step recursion for $L > 2$, as the theorem below states.

Theorem 3.1. *For the ensemble defined by* (1)–(3) *with weight-variance* $(\sigma_l)_{l=1}^{L}$, $\sigma_l > 0$,

$$\Gamma_2 \;=\; 2\,\sigma_1^2\,\sigma_2^2\,\big(1+\tfrac{1}{\pi}\big), \tag{12}$$

and, extending to networks of depth $L+1$,

$$\Gamma_{L+1} \;=\; \frac{\sigma_{New}^2}{2}\,\Gamma_L \qquad (L \geq 2). \tag{13}$$

Γ_L is the ensemble expectation defined in (10), and (13) relates its values when the network depth increases by one. As the theorem states, Γ_L is invariant to the bias variances $(\sigma_{b,l}^2)$. The quantity is defined for finite network widths. A brief derivation is provided in §4 and the detailed proof is in Appendix B [1]. Iterating (13) from (12) gives Γ_L in terms of the per-layer weight variances:

$$\Gamma_L \;=\; 8\big(1+\tfrac{1}{\pi}\big)\prod_{l=1}^{L}\frac{\sigma_l^2}{2}. \tag{14}$$

Corollary 3.2. *If* $\sigma_l \equiv \sigma_w$ *for all* l, *then*

$$\Gamma_L \;=\; 8\big(1+\tfrac{1}{\pi}\big)\left(\frac{\sigma_w^2}{2}\right)^{L}. \tag{15}$$

This follows from (14) by setting $\sigma_l \equiv \sigma_w$.

Remark. Under uniform weight variance, (13) reads $\Gamma_{L+1} = (\sigma_w^2/2)\,\Gamma_L$, so Γ_L is geometric in depth with ratio $\sigma_w^2/2$. When $\sigma_w^2 = 2$, the quantity becomes depth-invariant: $\Gamma_L = 8(1+1/\pi)$ for all $L \geq 2$. The value $\sigma_w^2 = 2$ is the variance prescribed by He et al. [8] for ReLU networks. The factor $1/2$ in the recursion is the probability that a ReLU unit is active under a symmetric pre-activation distribution.

Theorem 3.1 gives the expectation of Z_L over the network ensemble. Next, we consider the variance.

Theorem 3.3 (Variance of the generator energy)**.** *Let* $L \geq 2$, *let the widths be arbitrary, and let* $\boldsymbol{x} \neq \mathbf{0}$.

(a) $\mathrm{Var}(Z_L)$ *depends on the weight variances* $\sigma_1^2,\ldots,\sigma_L^2$ *and the widths* $p_0,\ldots,p_{L-1}$. *It depends neither on* $\boldsymbol{x}$ *nor on the bias variances* $\sigma_{b,1}^2,\ldots,\sigma_{b,L}^2$.

(b) At depth two,

$$\mathrm{Var}(Z_2) \;=\; \frac{2\,\sigma_1^4\,\sigma_2^4}{p_0\,p_1}\left[\left(5+\frac{8}{\pi}\right)(p_0+2) - 2\left(1+\frac{1}{\pi}\right)^2 p_0\right]. \tag{16}$$

(c) At depth L, $\mathrm{Var}(Z_L) = \mathbb{E}[Z_L^2] - \Gamma_L^2$, *where* Γ_L *is the closed form* (14) *and* $\mathbb{E}[Z_L^2]$ *is the combination* (85) *of three pairs of scalars* $(P^{(L)}, Q^{(L)})$. *Each pair starts from its depth-two value in* (77). *Each of the layers* $L-2, L-3, \ldots, 1$, *taken in this order, with fan-in* n', *fan-out* n *and weight variance* σ^2, *maps the pair* (P, Q) *to*

$$P' \;=\; \varsigma^2\Big(\frac{n^2+n}{4}\,P + n\,Q\Big), \qquad Q' \;=\; \varsigma^2\Big(\frac{n}{2}\,P + \frac{n^2+3n}{4}\,Q\Big), \qquad \varsigma \;:=\; \frac{\sigma^2}{n'}. \tag{17}$$

The proof is in Appendix C.

We measure the complexity of the decision boundary by its size on the input sphere per unit area of the sphere. By Crofton's formula [22], this density is proportional to the expected

[1] The new layer carries variance σ_{New}^2 rather than σ_{L+1}^2 because the derivation in §4 prepends the new layer at the input, which shifts every original σ_l to σ_{l+1}; the added layer is the new index 1, not the new index $L+1$.

number of times a great circle crosses the boundary per unit length. Γ_L is proportional to the mean squared gradient of $g := f_1 - f_2$ over random networks (30), so it measures how fast g changes along a great circle. The variance $q^{(L)}$ of each output unit measures how far g is from zero. The crossing density depends on the ratio of the two. The following theorem provides an estimation of the crossing density in terms of Γ_L with an error bound that vanishes as the hidden widths grow.

Theorem 3.4. *Let $g := f_1 - f_2$, let γ be a great circle of radius R in $\mathbb{R}^{p_0}$ parametrized by arc length s, and let $\lambda_0 := \mathrm{Var}\big(g(\gamma(s))\big)$ and $\lambda_2 := \mathrm{Var}\big(\frac{d}{ds} g(\gamma(s))\big)$ over all weights and biases. At every finite width the following hold.*

(a) Let I be an arc of γ. The expected number of crossings of the decision boundary per unit length along I is

$$\rho_{\text{exact}}(I) := \frac{1}{|I|}\, \mathbb{E}_{\mathbf{W}^{(1\dots L)}, \boldsymbol{b}^{(1\dots L)}} \left[\#\{s \in I : g(\gamma(s)) = 0\}\right] = \frac{1}{\pi}\, \mathbb{E}_{\mathcal{F}}\left[r(\mathbf{S})\right], \tag{18}$$

where $\mathcal{F} := (\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L-1)}, \boldsymbol{b}^{(1)}, \dots, \boldsymbol{b}^{(L-1)})$ denotes the weights and biases below the output layer. Given $\mathcal{F}$, $g(\gamma(s))$ is a Gaussian process. Let $\mathbf{S} := \mathrm{Cov}_{\mathbf{W}^{(L)}, \boldsymbol{b}^{(L)}}\left(\begin{bmatrix} g(\gamma(s)) \\ \frac{d}{ds} g(\gamma(s)) \end{bmatrix}\right)$ be the covariance matrix of the value and the derivative at arc parameter s. $r(\cdot)$ is a matrix function: $r(\mathbf{S}) := \sqrt{\det \mathbf{S}}/S_{11}$. Note that as a function of the random $\mathcal{F}$, $\mathbf{S}$ is a random matrix whose distribution does not depend on s or γ via rotation invariance. Hence s is dropped off from the r.h.s. of (18).

(b) We estimate ρ_{exact} by moving the expectation in (18) *inside r,*

$$\rho_L := \frac{1}{\pi} r\big(\mathbb{E}_{\mathcal{F}}[\mathbf{S}]\big) = \frac{1}{2\pi}\sqrt{\frac{1}{(1+1/\pi)p_0}}\sqrt{\frac{\Gamma_L}{q^{(L)}}}, \tag{19}$$

where $q^{(l)}$ is the per-neuron pre-activation variance at layer l [19],

$$q^{(1)} = \frac{\sigma_w^2 R^2}{p_0} + \sigma_b^2, \qquad q^{(l)} = \frac{\sigma_w^2}{2}\, q^{(l-1)} + \sigma_b^2, \quad l = 2, \dots, L. \tag{20}$$

(c) Consider the relative error of this estimate,

$$\delta \;:=\; \frac{\rho_{\text{exact}}}{\rho_L} - 1. \tag{21}$$

When all biases are zero, $\rho_L = 1/(\pi R)$ and

$$\delta \;=\; \prod_{l=1}^{L-1} \Big(1 - \frac{1+p_l}{2^{p_l}}\Big) - 1. \tag{22}$$

With Gaussian biases, to first order in $1/\min_l p_l$, $\delta = \sum_{l=1}^{L-1} c_l/p_l$ with constants $c_l \in [-\frac{5}{6}, 0]$ given in Proposition D.10.

The proofs are in Appendix D. Figure 1 illustrates the three parts of the theorem in panels (a)–(c).

4 Derivation

Theorems 3.1 and 3.3 are proved by induction on the depth L. In each step, a new layer is *prepended* as the first transform step of the input space. We consider the statistics of interest of Z_L (9). It is computed by first fixing the new layer's parameters and the input, and taking expectation over the existing network parameters with fixed feed-in via the inductive hypothesis. The result is symmetric w.r.t. the feed-in. Hence the induction is completed by taking expectation over the new weights. We sketch Theorem 3.1 as an example. We first consider the bias-free case and later show that the result is invariant with bias. Detailed proofs are provided in Appendices B and C.

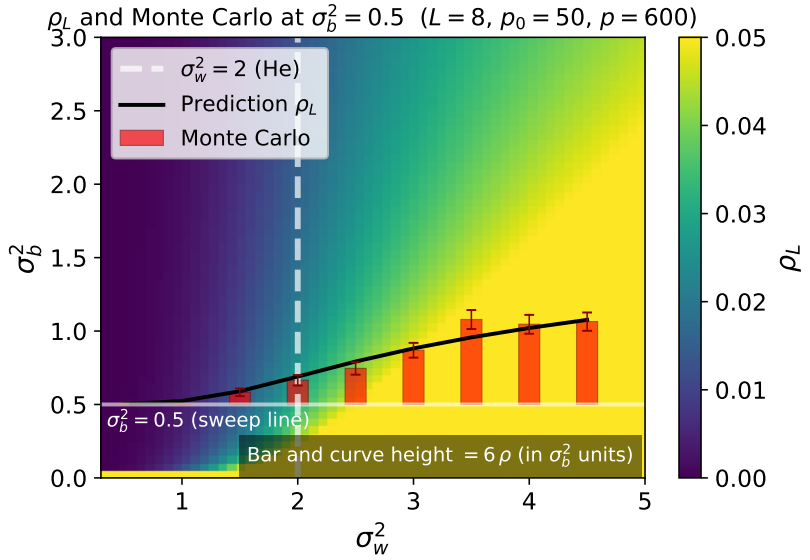


Figure 3: Phase diagram of ρ_L in (19) on the (σ_w^2, σ_b^2) plane, with Monte Carlo measurement along the slice $\sigma_b^2 = 0.5$. **Heat map:** ρ_L of predicted crossing density. **Bar and black curve overlay:** theoretical prediction and Monte Carlo measurement of the crossing density. **White dashed line:** marks $\sigma_w^2 = 2$, the initialization of He et al. [8].

At $L = 2$ the only cell is $\mathbb{R}^{p_0}$ and $\tilde{\mathbf{G}}^{(2)}_{k,\mathcal{C}} = \mathbf{G}_k(\mathbf{A}, \mathbf{B})$, with

$$\mathbb{E}_{\mathbf{A},\mathbf{B}}\left[(\mathbf{G}_k)_{ji}(\mathbf{G}_k)_{ji'}\right] = c^{(2)}_{k,j}\delta_{ii'}, \quad c^{(2)}_{k,j} = \frac{\sigma_1^2\sigma_2^2(1+1/\pi)}{p_0 p_1}, \qquad \Gamma_2 = 2\sigma_1^2\sigma_2^2\left(1+\frac{1}{\pi}\right), \tag{23}$$

Note that the positive and negative parts of the elements of $\mathbf{B}$ used to define $\mathbf{G}$ (see (5)–(6)) have mean $\sigma_2/\sqrt{2\pi p_1}$.

Inductive step ($L \to L+1$). Consider a depth-$(L+1)$ network of widths $p_0, p_1, \ldots, p_{L+1} = 2$ and weight variances $(\sigma_l)_{l=1}^{L+1}$. Its layer-1 weight matrix is $\mathbf{W}_E := \mathbf{W}^{(1)} \in \mathbb{R}^{p_1 \times p_0}$ with entries $(\mathbf{W}_E)_{as} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_1^2/p_0)$. For input $\boldsymbol{x} \in \mathbb{R}^{p_0}$, write $z_a := (\mathbf{W}_E)_{a,:} \cdot \boldsymbol{x}$, $d_a := \mathbf{1}\{z_a > 0\}$, $\mathbf{D} := \mathrm{Diag}(d_1, \ldots, d_{p_1})$. Layers $2, \ldots, L+1$ form a depth-L *sub-network* (last two layers $\mathbf{A} = \mathbf{W}^{(L)}$, $\mathbf{B} = \mathbf{W}^{(L+1)}$, generators $\mathbf{G}_k(\mathbf{A}, \mathbf{B})$, variances $(\sigma_l)_{l=2}^{L+1}$) fed input $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$, with cell-aware depth-$L$ generator $\tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k,\,\mathcal{C}_{\mathrm{sub}}} \in \mathbb{R}^{p_L \times p_1}$.

Layers $1, \ldots, L-1$ of the depth-$(L+1)$ network factor through layer 1 followed by layers $2, \ldots, L-1$ (which are the first $L-2$ layers of the sub-network). Within a cell $\mathcal{C}'$ this gives $\Phi^{(L-1)}_{\mathcal{C}'} = \Phi^{(L-2)}_{\mathrm{sub},\,\mathcal{C}_{\mathrm{sub}}}\,\mathbf{D}\,\mathbf{W}_E$ and therefore

$$\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}'} \;=\; \mathbf{G}_k\,\Phi^{(L-1)}_{\mathcal{C}'} \;=\; \tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k,\,\mathcal{C}_{\mathrm{sub}}} \cdot \mathbf{D}\,\mathbf{W}_E. \tag{24}$$

(24) is the recursion for cell-aware generators: the depth-$(L+1)$ generator equals the sub-network's depth-L generator right-multiplied by the layer-1 affine map. Conditional on $(\mathbf{W}_E, \boldsymbol{x})$, the sub-network's input $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$ is fixed; the sub-network's weights are independent of $\mathbf{W}_E$, so $\tilde{\mathbf{G}}^{(L)}_{\mathrm{sub}} \perp (\mathbf{W}_E, \mathbf{D})$ conditionally.

For $j \in \{1, \ldots, p_L\}$ and $t, t' \in \{1, \ldots, p_0\}$, expand $\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt} = \sum_{a=1}^{p_1} \tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k;\,ja}\,d_a\,(\mathbf{W}_E)_{at}$ and take expectation:

$$\mathbb{E}_{\mathbf{W}_E,\Theta_{\mathrm{sub}}}\left[\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt}\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt'}\right] = \sum_{a,a'=1}^{p_1} \mathbb{E}\left[\tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},k;ja}\tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},k;ja'}d_a d_{a'}(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{a't'}\right]. \tag{25}$$

Apply the inductive hypothesis $\mathbb{E}_{\Theta_{\text{sub}}}[\tilde{\mathbf{G}}^{(L)}_{\text{sub},k;ja}\tilde{\mathbf{G}}^{(L)}_{\text{sub},k;ja'} \mid \mathbf{W}_E, \boldsymbol{x}] = c^{(L)}_{k,j}\delta_{aa'}$ (which holds by the conditional independence above, and because $c^{(L)}_{k,j}$ in (37) does not depend on the feed-in): the double sum collapses to

$$\mathbb{E}_{\mathbf{W}_E,\Theta_{\text{sub}}}\left[\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';jt}\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';jt'}\right] = c^{(L)}_{k,j}\sum_{a=1}^{p_1}\mathbb{E}_{\mathbf{W}_E}\left[d_a(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}\right]. \tag{26}$$

Flipping the sign of $(\mathbf{W}_E)_{a,:}$ maps d_a to $1-d_a$ and leaves $(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}$ unchanged, so

$$\mathbb{E}_{\mathbf{W}_E}\left[d_a(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}\right] = \frac{1}{2}\,\mathbb{E}_{\mathbf{W}_E}\left[(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}\right] = \frac{\sigma_1^2}{2p_0}\delta_{tt'}. \tag{27}$$

Substituting (27) into (26):

$$\mathbb{E}_{\mathbf{W}_E,\Theta_{\text{sub}}}\left[\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';jt}\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';jt'}\right] = c^{(L)}_{k,j}\frac{p_1\sigma_1^2}{2p_0}\delta_{tt'} = c^{(L+1)}_{k,j}\delta_{tt'}, \qquad c^{(L+1)}_{k,j} := \frac{p_1\sigma_1^2}{2p_0}c^{(L)}_{k,j}, \tag{28}$$

Summing diagonal entries: $\Gamma_{L+1} = \sum_{k,j} p_0\, c^{(L+1)}_{k,j} = (\sigma_1^2/2)\sum_{k,j} p_1\, c^{(L)}_{k,j} = (\sigma_1^2/2)\,\Gamma_L$, where the last equality uses $\Gamma_L = \sum_{k,j} p_1\, c^{(L)}_{k,j}$ (input dim of the sub-network is p_1).

Bias invariance. With biases $b_a \sim \mathcal{N}(0, \sigma^2_{b,1})$ on the prepended layer, the activation indicator becomes $d_a = \mathbf{1}\{z_a + b_a > 0\}$. Flipping the signs of $(\mathbf{W}_E)_{a,:}$ and b_a together still maps d_a to $1-d_a$ and leaves $(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}$ unchanged, so (27) holds as before, with the expectation also taken over b_a. The base case (23) is bias-free, since $\mathbf{G}_k$ depends on $\mathbf{A}$ and $\mathbf{B}$ alone. The inductive step and the closed form therefore extend to per-layer biases with variances $(\sigma^2_{b,l})_{l=1}^{L}$, $\sigma_{b,l} \geq 0$.

Induction of $\mathrm{Var}(Z_L)$. The same technique applies to the proof of Theorem 3.3, one order higher: the second moments in (28) are replaced by fourth moments over four column indices, and each prepended layer maps their coefficients (P, Q) linearly by (17).

Crossing density. Recall that the crossing density depends on the ratio of how fast g changes, $|dg/ds|$, to how far g is from zero, $|g|$. Given $\mathcal{F}$, $g \circ \gamma$ is a Gaussian process whose covariance $\mathbf{S}$ is built from the values and derivatives of the penultimate layer, and Rice's formula gives part (a). The off-diagonal entry of $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$ vanishes by reflection, so $\rho_L = \frac{1}{\pi}\sqrt{\lambda_2/\lambda_0}$, where λ_0 and λ_2 are the diagonal entries of $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$. Part (c) follows $\mathbf{S}$ layer by layer from the input upward.

For λ_0: f_1 and f_2 have independent output-layer rows in $\mathbf{W}^{(L)}$, so $\mathrm{Cov}(f_1, f_2) = 0$. At every finite width, each $f_k(\boldsymbol{x})$ has ensemble variance $q^{(L)}$ [19, 23], hence $\lambda_0 = \mathrm{Var}(f_1 - f_2) = 2q^{(L)}$.

For $\lambda_2 = \mathrm{Var}(\frac{d}{ds}g(\gamma(s)))$: the directional derivative is $\nabla g(\gamma(s)) \cdot \dot{\gamma}(s)$ with $\|\dot{\gamma}(s)\|_2 \equiv 1$. $\mathbb{E}_{\mathbf{W}^{(1\ldots L)},\boldsymbol{b}^{(1\ldots L)}}[\nabla g \nabla g^T] = (\mathbb{E}_{\mathbf{W}^{(1\ldots L)},\boldsymbol{b}^{(1\ldots L)}}\|\nabla g\|_2^2/p_0)I$, so projecting onto the unit tangent gives

$$\lambda_2 = \mathbb{E}_{\mathbf{W}^{(1\ldots L)},\boldsymbol{b}^{(1\ldots L)}}\|\nabla g(\gamma(s))\|_2^2/p_0. \tag{29}$$

At finite width g is piecewise linear: on each cell $\mathcal{C}$ the gradient is the constant $\nabla g|_{\mathcal{C}}$. The expectation of $\|\nabla g|_{\mathcal{C}}\|_2^2$ has a closed form in the weights, proportional to Γ_L:

$$\Gamma_L = 2\left(1 + \tfrac{1}{\pi}\right)\mathbb{E}_{\mathbf{W}^{(1\ldots L)},\boldsymbol{b}^{(1\ldots L)}}\|\nabla g|_{\mathcal{C}}\|_2^2. \tag{30}$$

The intuition is that both Γ_L and $\mathbb{E}_{\mathbf{W}^{(1\ldots L)},\boldsymbol{b}^{(1\ldots L)}}\|\nabla g|_{\mathcal{C}}\|_2^2$ are weighted sums of $\left\|(\mathbf{A}\Phi_{\mathcal{C}}^{(L-2)})_{j,:}\right\|_2^2$ over neurons j at layer $L-1$.

By (19), Γ_L is the geometric factor in the estimate ρ_L of the crossing density, with the bias variance σ_b^2 entering only through $q^{(L)}$ via (20) (Γ_L itself is bias-independent). Figure 3 reports Monte Carlo support for (19). Monte Carlo measurements support the closed-form prediction across the full σ_w^2 sweep.

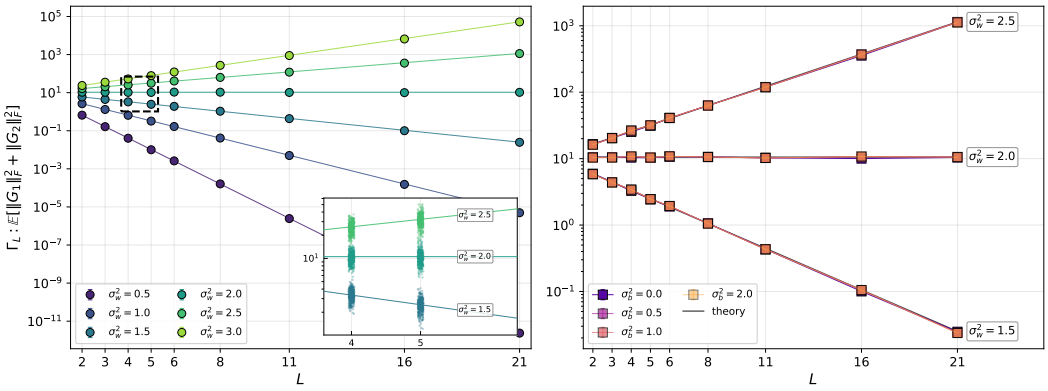


Figure 4: Numerical check of Γ_L against the closed form (15). (A) $\widehat{\Gamma}_L$ (markers) and theory (solid lines) for six values of σ_w^2, $\sigma_b^2 = 0$, with L ranging from 2 to 21; the inset enlarges $L \in \{4, 5\}$ for $\sigma_w^2 \in \{1.5, 2.0, 2.5\}$ and shows all per-input samples. (B) Same comparison at $\sigma_w^2 \in \{1.5, 2.0, 2.5\}$ with $\sigma_b^2 \in \{0, 0.5, 1.0, 2.0\}$ uniform across layers; in each σ_w^2 family the four bias curves collapse onto the single theory line.

Numerical support We checked the recursion of Theorem 3.1 and its closed form (15). For each $(L, \sigma_w^2, \sigma_b^2)$ in the grid of Figure 4 we draw 50 random networks of architecture $20 \to \underbrace{128 \to \cdots \to 128}_{L-1 \text{ hidden}} \to 2$ under (3) with $\sigma_l^2 \equiv \sigma_w^2$, $\sigma_{b,l}^2 \equiv \sigma_b^2$ throughout. For each network we sample 200 inputs uniformly from $S^{19} \subset \mathbb{R}^{20}$, form $\tilde{\mathbf{G}}_{1,\mathcal{C}(\boldsymbol{x})}^{(L)}, \tilde{\mathbf{G}}_{2,\mathcal{C}(\boldsymbol{x})}^{(L)}$ via (5)–(6) and (8), and record $\left\|\tilde{\mathbf{G}}_{1,\mathcal{C}}^{(L)}\right\|_F^2 + \left\|\tilde{\mathbf{G}}_{2,\mathcal{C}}^{(L)}\right\|_F^2$. Averaging over the resulting 50×200 samples gives the empirical $\widehat{\Gamma}_L$ plotted as markers; the prediction of (15) is the matching solid line, with no free parameter.

In Figure 4 the markers follow the theory line at every σ_w^2, and the four bias variances give the same curve.

5 Conclusion

We characterised the distribution of the squared Frobenius norm Z_L of the zonotope generator matrices of random Gaussian deep ReLU classifiers by exact, finite-width recursions in depth. The tropical description organizes the analysis by the neurons of the last hidden layer, and the crossing-density estimate inherits this organization, since given the lower layers the output covariance is a sum of one term per such neuron. Over the ensemble of random networks, the expectation $\Gamma_L = \mathbb{E}[Z_L]$ is computed from the weight variances, $\Gamma_L = 8(1 + 1/\pi) \prod_{l=1}^{L} \sigma_l^2/2$. The computation uses recursions that hold at finite width, which also provide the variance of Z_L. When the hidden widths are large compared with the depth, this variance is small relative to Γ_L^2, so Z_L of a single network concentrates around Γ_L. Through Rice's formula, Z_L of a single network and the variance $q^{(L)}$ of the network output [19] give an estimate of how often a great circle on the input sphere crosses the decision boundary. The estimate is proportional to $\sqrt{Z_L/q^{(L)}}$. As the hidden widths grow compared with the depth, the estimate approaches the ensemble crossing rate.

As to future development, our recursions start from the Gaussian distribution of the parameters at initialization. Training changes this distribution, and the next step is to follow how Z_L and the crossing rate of a network change during training. Code for numerical checks of Theorems 3.1, 3.3 and 3.4 is available at `https://github.com/junjy007/tropical-boundary-complexity`.

Acknowledgments and Disclosure of Funding

The author thanks the anonymous reviewers for their very helpful suggestions. This work received no external funding, and the author has no competing interests.

References

[1] Andrea Agazzi, Eloy Mosig García, and Dario Trevisan. Quantitative convergence of trained neural networks to Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[2] Motasem Alfarra, Adel Bibi, Hasan Hammoud, Mohamed Gaafar, and Bernard Ghanem. On the decision boundaries of neural networks: A tropical geometry perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5027–5037, 2023.

[3] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[4] Marie-Charlotte Brandenburg, Georg Loho, and Guido Montúfar. The real tropical geometry of neural networks for binary classification. *Transactions on Machine Learning Research*, 2024.

[5] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[6] Max Guillen, Philipp Misof, and Jan E. Gerken. Finite-width neural tangent kernels from Feynman diagrams. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 37700–37752. PMLR, 2026.

[7] Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2596–2604. PMLR, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[9] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[10] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

[11] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[12] Petros Maragos, Vasileios Charisopoulos, and Emmanouil Theodosis. Tropical geometry and machine learning. *Proceedings of the IEEE*, 109(5):728–755, 2021.

[13] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[14] Guido F. Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 2924–2932, 2014.

[15] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1982.

[16] Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, 1996.

[17] Niket Nikul Patel and Guido Montúfar. On the local complexity of linear regions in deep ReLU networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48335–48370. PMLR, 2025.

[18] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[19] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, volume 29, pages 3360–3368, 2016.

[20] S. O. Rice. Mathematical analysis of random noise. *Bell System Technical Journal*, 23(3): 282–332, 1944.

[21] S. O. Rice. Mathematical analysis of random noise (concluded). *Bell System Technical Journal*, 24(1):46–156, 1945.

[22] Luis A. Santaló. *Integral Geometry and Geometric Probability*, volume 1 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley, 1976.

[23] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

[24] Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4558–4566. PMLR, 2018.

[25] Moritz Stargalla, Christoph Hertrich, and Daniel Reichman. The computational complexity of counting linear regions in ReLU neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[26] Thomas Zaslavsky. Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes. *Memoirs of the American Mathematical Society*, 1(154), 1975.

[27] Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5824–5832. PMLR, 2018.

A Notation

Figure 5 shows the depth-L network of Section 3 with the symbols used in the appendices, and Table 1 lists them. Layer $l \in \{1, \dots, L\}$ applies $\mathbf{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$ and maps unit set $l-1$ to unit set l, so unit set 0 is the input, unit sets $1, \dots, L-1$ are hidden, and unit set L holds the two outputs f_1, f_2.

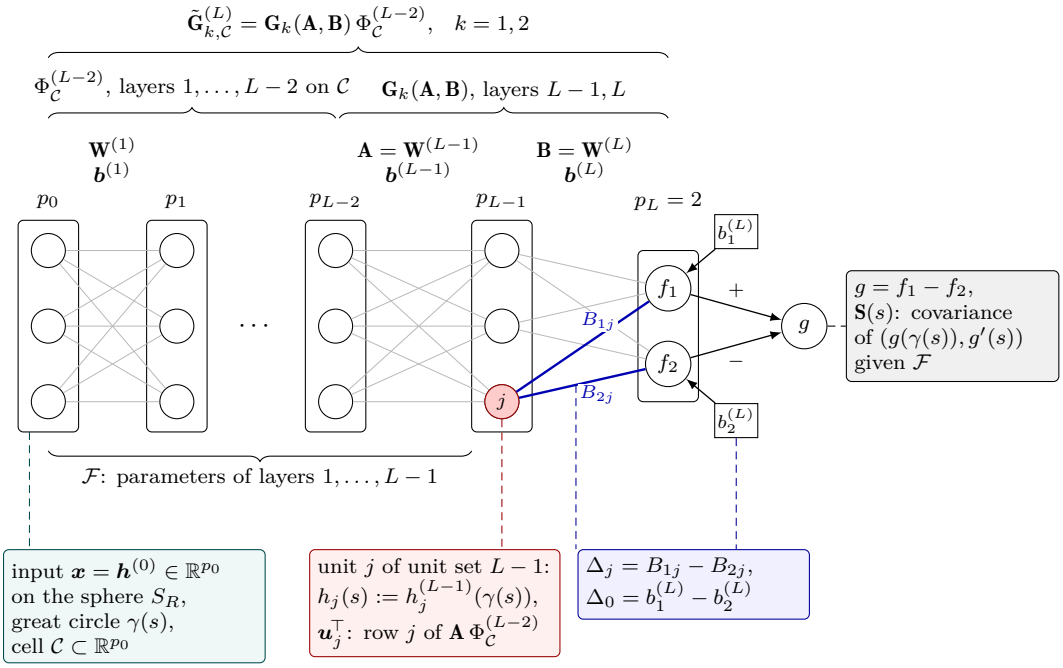


Figure 5: The depth-L network. Circles are units and each box is a unit set; unit set l holds p_l units, and p_l is written above its box. Layer l applies $\mathbf{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$, followed by ReLU for $l \le L-1$, and maps unit set $l-1$ to unit set l. On the cell $\mathcal{C}$ the activation patterns of unit sets $1, \dots, L-2$ are fixed, and layers $1, \dots, L-2$ act as $\Phi^{(L-2)}_{\mathcal{C}}$. The highlighted unit j and its two output weights carry the quantities of Appendix D.

Table 1: Main symbols used in the appendices.

Symbol	Meaning	Defined in
Network		
L	depth: layers $1,\dots,L$, unit sets $0,\dots,L$	(1)
p_l	width of unit set l; $p_L = 2$	(1)
$\boldsymbol{x} = \boldsymbol{h}^{(0)} \in \mathbb{R}^{p_0}$	input	(1)
$\boldsymbol{h}^{(l)},\ h_i^{(l)}$	output of unit set l and of its unit i	(1)
$y_i^{(l)}$	pre-activation of unit i of unit set l	Section D.3
$\mathbf{W}_E,\ \boldsymbol{z},\ z_a$	$\mathbf{W}_E = \mathbf{W}^{(1)}$ in the inductive step, $\boldsymbol{z} = \mathbf{W}_E\boldsymbol{x}$, $z_a + b_a^{(1)} = y_a^{(1)}$	(44)
$\mathbf{D},\ d_a$	activation indicators of unit set 1, $d_a = \mathbf{1}\{z_a > 0\}$	(44)
$\mathbf{W}^{(l)},\ \boldsymbol{b}^{(l)}$	weights and bias of layer l	(1), (3)
f_1, f_2	the two output units	(2)
$g := f_1 - f_2$	decision function; the decision boundary is $\{g = 0\}$	(4), Appendix D
$(\mathbf{A}, \mathbf{B})$	$(\mathbf{W}^{(L-1)}, \mathbf{W}^{(L)})$	Section 3
$u^+,\ u^-$	$\max(u, 0)$ and $\max(-u, 0)$, entrywise on matrices (the main text writes a^+, a^-)	(5)
Cells and generators		
$\mathcal{C}$	set of inputs with a fixed sign pattern on unit sets $1,\dots,L-2$	Section 3
$\Phi_{\mathcal{C}}^{(L-2)}$	linear map of layers $1,\dots,L-2$ on $\mathcal{C}$, a matrix in $\mathbb{R}^{p_{L-2}\times p_0}$	(8)
$\beta_{k,j}$	generator coefficients built from column j of $\mathbf{B}$	(5)–(6)
$\mathbf{G}_k$	depth-2 generator matrices, $(\mathbf{G}_k)_{ji} = \beta_{k,j}A_{ji}$	(5)
$\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)}$	generators on $\mathcal{C}$, $\mathbf{G}_k\,\Phi_{\mathcal{C}}^{(L-2)}$	(8)
$c_{k,j}^{(L)}$	column covariance of row j of $\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)}$	(37)
$\boldsymbol{u}_j^\top,\ \nu_j$	$\boldsymbol{u}_j^\top$: row j of $\mathbf{A}\,\Phi_{\mathcal{C}}^{(L-2)}$; ν_j: constant term of the pre-activation of unit j of unit set $L-1$ in $\mathcal{C}$	Section D.1
Ensemble quantities		
$\sigma_l^2,\ \sigma_{b,l}^2$	weight and bias variances of layer l	(3)
Γ_L	boundary complexity, expected generator energy	(10)
C_0	$\Gamma_2 = 2\sigma_1^2\sigma_2^2(1 + 1/\pi)$	(12), (43)
$q^{(l)}$	pre-activation variance of a unit of unit set l	(20), Lemma D.4
Crossing quantities		
$S_R,\ \gamma(s)$	input sphere of radius R and a great circle on it, parameterized by arc length	(105)
$h_j(s)$	$h_j^{(L-1)}(\gamma(s))$	Section D.1
$\Delta_j,\ \Delta_0$	$B_{1j} - B_{2j}$ and $b_1^{(L)} - b_2^{(L)}$	(102)
$\mathcal{F}$	parameters of layers $1,\dots,L-1$	(103)
N_ℓ	number of crossings on an arc of length ℓ	Section D.1
$\mathbf{S}(s)$	covariance of $(g(\gamma(s)), g'(s))$ given $\mathcal{F}$	(115)
$r(\mathbf{S})$	$\sqrt{\det \mathbf{S}}/S_{11}$	(116)
ρ_{exact}	expected number of crossings per unit length	(129)
ρ_L	estimate $\frac{1}{\pi} r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$	(143)
$\lambda_0,\ \lambda_2$	diagonal entries of $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$	(144)
$\kappa(p_0)$	constant in $\rho_L = \kappa(p_0)\sqrt{\Gamma_L/q^{(L)}}$	(145)
δ	relative error $\rho_{\text{exact}}/\rho_L - 1$	(156)
$\mathbf{S}^{(l)}$	Gram matrix of $(h_i^{(l)}, h_i^{(l)\prime})$ over unit set l	(158)
$\mathbf{C}^{(l)}$	covariance of $(y_i^{(l)}, y_i^{(l)\prime})$ given layers $1,\dots,l-1$	(179)
$p_{\min}$	$\min_{1\le l\le L-1} p_l$	Section D.4
$\chi_l,\ \omega_l$	share of $q^{(l)}$ from unit set $l-1$, share of $q^{(L)}$ from unit set l	(180)
$V_{11}^{(l)}, V_{22}^{(l)}, V_{\times}^{(l)}, V_{12}^{(l)}$	normalized fourth moments of $(h_i^{(l)}, h_i^{(l)\prime})$	Section D.4
c_l	first-order coefficient of δ in $1/p_l$	(183)
Z_L	generator energy of one network, $\mathbb{E}[Z_L] = \Gamma_L$	Section D.5
$\hat{\rho}_L$	estimate $\kappa(p_0)\sqrt{Z_L/q^{(L)}}$ from one network	(194)

B Proofs

We recap the setup of Theorem 3.1. The notation is collected in Section A.

Architecture and ensemble. A fully connected feedforward network with L layers has unit sets $0, 1, \ldots, L$ of widths $p_0, p_1, \ldots, p_{L-1}, p_L$, where $p_L = 2$. Layer l maps unit set $l-1$ to unit set l. Layers $1, \ldots, L-1$ apply a ReLU activation (1), so unit sets $1, \ldots, L-1$ are hidden, and layer L is linear (2). Weights and biases are drawn independently:

$$W_{ij}^{(l)} \overset{\text{iid}}{\sim} \mathcal{N}\!\left(0,\, \frac{\sigma_l^2}{p_{l-1}}\right), \qquad b_i^{(l)} \overset{\text{iid}}{\sim} \mathcal{N}\big(0,\, \sigma_{b,l}^2\big), \qquad l = 1, \ldots, L, \tag{31}$$

with $\sigma_l > 0$ and $\sigma_{b,l} \geq 0$ (cf. (3)).

Last two layers. The last two weight matrices are $\mathbf{A} = \mathbf{W}^{(L-1)} \in \mathbb{R}^{p_{L-1} \times p_{L-2}}$ and $\mathbf{B} = \mathbf{W}^{(L)} \in \mathbb{R}^{2 \times p_{L-1}}$. The entries satisfy $A_{ji} \sim \mathcal{N}(0,\, \sigma_{L-1}^2/p_{L-2})$ and $B_{kj} \sim \mathcal{N}(0,\, \sigma_L^2/p_{L-1})$, all independent.

Positive and negative parts. For $u \in \mathbb{R}$, set $u^+ := \max(u,\, 0)$ and $u^- := \max(-u,\, 0)$; then $u = u^+ - u^-$ and $|u| = u^+ + u^-$. Applying $(\cdot)^{\pm}$ entrywise to $\mathbf{B}$, define the generator coefficients

$$\beta_{1,j} := B_{1j}^+ + B_{2j}^-, \tag{32}$$

$$\beta_{2,j} := B_{2j}^+ + B_{1j}^-, \qquad j = 1, \ldots, p_{L-1}. \tag{33}$$

Each $\beta_{k,j} \geq 0$ is built from column j of $\mathbf{B}$ alone. The zonotope generator matrices are then $(\mathbf{G}_k)_{ji} := \beta_{k,j}\, A_{ji}$, $\mathbf{G}_k \in \mathbb{R}^{p_{L-1} \times p_{L-2}}$, $k = 1, 2$, as in (5)–(6).

Main theorem (restated). For a network ensemble as above,

$$\Gamma_2 \;=\; 2\,\sigma_1^2\,\sigma_2^2\!\left(1 + \frac{1}{\pi}\right), \qquad \Gamma_{L+1} \;=\; \frac{\sigma_{\text{New}}^2}{2}\,\Gamma_L \quad (L \geq 2), \tag{34}$$

where σ_{New}^2 is the variance of a layer prepended at the input (cf. the footnote to (13)); iterating gives the closed form

$$\Gamma_L \;=\; 8\!\left(1 + \frac{1}{\pi}\right)\prod_{l=1}^{L}\frac{\sigma_l^2}{2}, \tag{35}$$

where

$$\Gamma_L \;:=\; \mathbb{E}_{\mathcal{C},\mathbf{W}^{(1)},\ldots,\mathbf{W}^{(L)},\boldsymbol{b}}\!\left[\,\left\|\tilde{\mathbf{G}}_{1,\mathcal{C}}^{(L)}\right\|_F^2 + \left\|\tilde{\mathbf{G}}_{2,\mathcal{C}}^{(L)}\right\|_F^2\,\right] \tag{36}$$

is the cell-aware boundary complexity of (10). The cell-aware generators $\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)} \in \mathbb{R}^{p_{L-1} \times p_0}$ are $\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(L)}(\mathbf{A}, \mathbf{B}, \Phi_{\mathcal{C}}^{(L-2)}) := \mathbf{G}_k(\mathbf{A}, \mathbf{B})\,\Phi_{\mathcal{C}}^{(L-2)}$, with $\mathbf{G}_k$ the depth-2 generator of the last two layers (5)–(6) and $\Phi_{\mathcal{C}}^{(L-2)} \in \mathbb{R}^{p_{L-2} \times p_0}$ the linear map realized by the first $L-2$ layers on the cell $\mathcal{C}$. The expectation is jointly over all random weights, the bias $\boldsymbol{b} := (\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(L)})$, and the cell $\mathcal{C}$ (determined by the lower-layer weights and the input $\boldsymbol{x}$).

Strategy of the proof. The argument is an induction on L. The inductive invariant strengthens the closed-form identity into a column-covariance statement: for every input $\boldsymbol{x} \in \mathbb{R}^{p_0}$ and every cell $\mathcal{C}$ containing $\boldsymbol{x}$,

$$\mathbb{E}\Big[\,\tilde{\mathbf{G}}_{k,\mathcal{C};\,ji}^{(L)}\,\tilde{\mathbf{G}}_{k,\mathcal{C};\,ji'}^{(L)}\,\Big] \;=\; c_{k,j}^{(L)}\,\delta_{ii'}, \quad i, i' \in \{1, \ldots, p_0\},\; j \in \{1, \ldots, p_{L-1}\}, \tag{37}$$

for a constant $c_{k,j}^{(L)} \geq 0$. The expectation is over all random weights, biases, and cells consistent with $\boldsymbol{x}$. The base case ($L = 2$, §B.1) verifies (37) and reads off $c_{k,j}^{(2)} =$

$\sigma_1^2\sigma_2^2(1+1/\pi)/(p_0p_1)$, giving $\Gamma_2 = C_0$. The inductive step (§B.2) treats the depth-$(L+1)$ network with input width p_0 as layer 1 (weight variance σ_1^2, output width p_1) followed by a depth-L sub-network of input width p_1, and propagates (37) to depth $L+1$ with constant $c_{k,j}^{(L+1)} = (p_1\,\sigma_1^2/(2p_0))\,c_{k,j}^{(L)}$. The factor $1/p_0$ comes from the fan-in normalization σ_1^2/p_0 of the prepended layer's weights, and the factor p_1 from the sum over the inner index $a \in \{1,\dots,p_1\}$ that the inductive hypothesis collapses via $\delta_{aa'}$. Summing diagonal entries, $\Gamma_{L+1} = \sum_{k,j} p_0\,c_{k,j}^{(L+1)} = (\sigma_1^2/2)\sum_{k,j} p_1\,c_{k,j}^{(L)} = (\sigma_1^2/2)\,\Gamma_L$, the width factors cancelling. Iterating yields the closed form (§B.3). Bias invariance (§B.4) shows that the same inductive step goes through with $\mathbf{1}\{z_a > 0\}$ replaced by $\mathbf{1}\{z_a + b_a > 0\}$, with the cross-moment $\mathbb{E}[d_a(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}]$ unchanged.

B.1 Base case: $\Gamma_2 = C_0$ and the column-covariance identity

At $L = 2$ the network has the two layers $\mathbf{A} := \mathbf{W}^{(1)}$ and $\mathbf{B} := \mathbf{W}^{(2)}$; there are no lower layers, the cell map $\Phi_{\mathcal{C}}^{(0)}$ is the identity on $\mathbb{R}^{p_0}$, and the cell-aware generator coincides with the depth-2 generator, $\tilde{\mathbf{G}}_{k,\mathcal{C}}^{(2)} = \mathbf{G}_k(\mathbf{A},\mathbf{B}) \in \mathbb{R}^{p_1\times p_0}$, $(\mathbf{G}_k)_{ji} = \beta_{k,j}\,A_{ji}$ with $\beta_{1,j} = B_{1j}^+ + B_{2j}^-$, $\beta_{2,j} = B_{2j}^+ + B_{1j}^-$ (2, 27, recalled in (5)–(6)). $\beta_{k,j}$ depends on $\mathbf{B}$ alone, A_{ji} on $\mathbf{A}$ alone, and $\mathbf{A} \perp \mathbf{B}$; the biases $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}$ do not enter $\beta_{k,j}$ or A_{ji}.

Column-covariance identity. Fix $k \in \{1,2\}$, $j \in \{1,\dots,p_1\}$, and $i,i' \in \{1,\dots,p_0\}$. Then

$$\mathbb{E}\big[\,(\mathbf{G}_k)_{ji}\,(\mathbf{G}_k)_{ji'}\,\big] \;=\; \mathbb{E}_{\mathbf{B}}[\beta_{k,j}^2]\cdot\mathbb{E}_{\mathbf{A}}[A_{ji}\,A_{ji'}]. \tag{38}$$

The right-hand factors are computed in turn.

The entries of $\mathbf{A}$ are i.i.d. $\mathcal{N}(0,\sigma_1^2/p_0)$, so

$$\mathbb{E}_{\mathbf{A}}[A_{ji}\,A_{ji'}] \;=\; \frac{\sigma_1^2}{p_0}\,\delta_{ii'}. \tag{39}$$

For $\mathbb{E}_{\mathbf{B}}[\beta_{k,j}^2]$, set $\sigma^2 := \sigma_2^2/p_1$ and write $u := B_{1j}$, $v := B_{2j}$, both i.i.d. $\mathcal{N}(0,\sigma^2)$. The half-normal identities

$$\mathbb{E}[u^+] = \mathbb{E}[u^-] = \frac{\sigma}{\sqrt{2\pi}}, \qquad \mathbb{E}[(u^+)^2] = \mathbb{E}[(u^-)^2] = \frac{\sigma^2}{2}, \tag{40}$$

combined with $u \perp v$ give

$$\begin{aligned}\mathbb{E}[\beta_{1,j}^2] \;&=\; \mathbb{E}[(u^+)^2] + 2\,\mathbb{E}[u^+]\,\mathbb{E}[v^-] + \mathbb{E}[(v^-)^2]\\ &=\; \frac{\sigma^2}{2} \;+\; \frac{\sigma^2}{\pi} \;+\; \frac{\sigma^2}{2} \;=\; \frac{\sigma_2^2}{p_1}\left(1+\frac{1}{\pi}\right);\end{aligned} \tag{41}$$

swapping $u \leftrightarrow v$ gives the identical value for $\mathbb{E}[\beta_{2,j}^2]$. Substituting (39) and (41) into (38):

$$\mathbb{E}\big[\,(\mathbf{G}_k)_{ji}\,(\mathbf{G}_k)_{ji'}\,\big] \;=\; c_{k,j}^{(2)}\,\delta_{ii'}, \qquad c_{k,j}^{(2)} \;=\; \frac{\sigma_1^2\,\sigma_2^2(1+1/\pi)}{p_0\,p_1}. \tag{42}$$

The constant $c_{k,j}^{(2)}$ is the same for every k and every j, and is independent of i, of $\mathcal{C}$ (the only cell at $L=2$ is $\mathbb{R}^{p_0}$), and of the input $\boldsymbol{x}$. This establishes the inductive invariant (37) at $L = 2$.

From the column-covariance identity to C_0. Γ_2 is the sum of the diagonal entries of (42):

$$\begin{aligned}\Gamma_2 \;&=\; \sum_{k=1}^{2}\sum_{j=1}^{p_1}\sum_{i=1}^{p_0}\mathbb{E}\big[\,(\mathbf{G}_k)_{ji}^2\,\big] \;=\; \sum_{k=1}^{2}\sum_{j=1}^{p_1}\sum_{i=1}^{p_0} c_{k,j}^{(2)}\\ &=\; 2\,p_0\,p_1\cdot\frac{\sigma_1^2\,\sigma_2^2(1+1/\pi)}{p_0\,p_1} \;=\; 2\,\sigma_1^2\,\sigma_2^2\left(1+\frac{1}{\pi}\right) \;=\; C_0,\end{aligned} \tag{43}$$

the value asserted in Theorem 3.1. Both widths p_0 and p_1 have cancelled.

B.2 Inductive step: prepending one layer

Setup

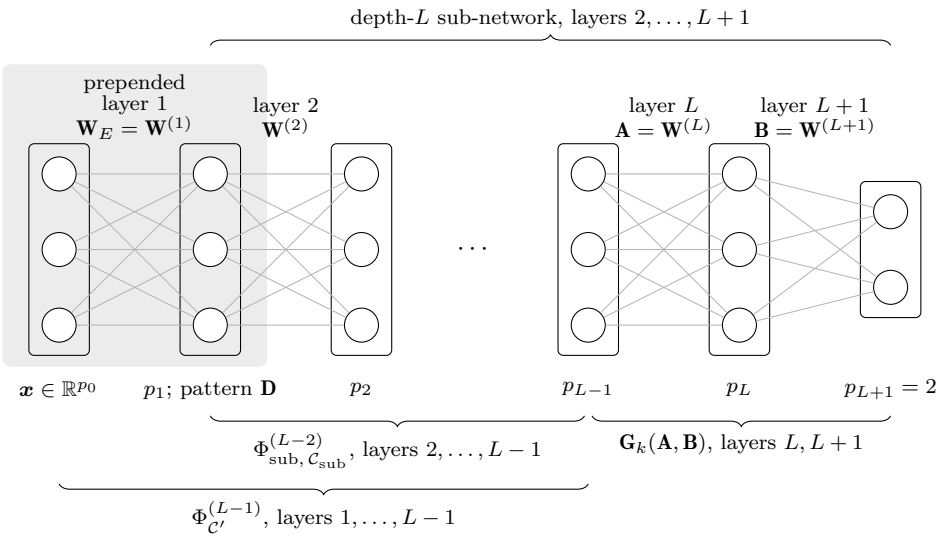


Figure 6: The depth-$(L+1)$ network of the inductive step. Circles are units and each box is a unit set; unit set l holds p_l units. Layer l denotes the l-th processing step, in which $\mathbf{W}^{(l)}$ and $\boldsymbol{b}^{(l)}$, followed by ReLU for $l \le L$, map unit set $l-1$ to unit set l. The network thus has $L+1$ layers and $L+2$ unit sets. The shaded layer 1 is the layer prepended in this step, and layers $2, \dots, L+1$ form the depth-L sub-network. On a cell $\mathcal{C}'$ the activation patterns of unit sets $1, \dots, L-1$ are fixed, whereas the pattern of unit set L is not fixed and is accounted for by the generators $\mathbf{G}_k$.

Consider a depth-$(L+1)$ network with $L+1$ layers and unit sets $0, \dots, L+1$ of widths $p_0, p_1, \dots, p_{L+1} = 2$ and weight variances $(\sigma_l)_{l=1}^{L+1}$. The network of Figure 6 is the depth-L network of Figure 5 with one layer prepended at the input, so layer l of the depth-L network becomes layer $l+1$. Write the layer 1 weight matrix as $\mathbf{W}_E := \mathbf{W}^{(1)} \in \mathbb{R}^{p_1 \times p_0}$, with entries $(\mathbf{W}_E)_{as} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_1^2/p_0)$, $a \in \{1, \dots, p_1\}$, $s \in \{1, \dots, p_0\}$. Layer 1 is the layer being prepended in this step, so its variance σ_1^2 is the σ^2_{New} of (13). For an input $\boldsymbol{x} \in \mathbb{R}^{p_0}$, the preactivation of unit set 1 is $\boldsymbol{z} := \mathbf{W}_E \boldsymbol{x} \in \mathbb{R}^{p_1}$, with components $z_a = (\mathbf{W}_E)_{a,:} \cdot \boldsymbol{x}$, so that $z_a + b_a^{(1)} = y_a^{(1)}$. Set

$$\mathbf{D} := \text{Diag}(d_1, \dots, d_{p_1}) \in \mathbb{R}^{p_1 \times p_1}, \qquad d_a := \mathbf{1}\{z_a > 0\}, \tag{44}$$

so that the output of unit set 1 within its activation cell is $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x} \in \mathbb{R}^{p_1}$.

Layers $2, \dots, L+1$ of the depth-$(L+1)$ network form a *depth-L sub-network* of widths $p_1, p_2, \dots, p_{L+1}$ and weight variances $(\sigma_l)_{l=2}^{L+1}$, fed input $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$. We index this sub-network's quantities by "sub". Its last two layers are $\mathbf{A} = \mathbf{W}^{(L)}$ $(p_L \times p_{L-1})$ and $\mathbf{B} = \mathbf{W}^{(L+1)}$ $(p_{L+1} \times p_L = 2 \times p_L)$; its depth-2 generators $\mathbf{G}_k(\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{p_L \times p_{L-1}}$ are unchanged from (5)–(6), and its cell-aware depth-L generator at sub-network cell $\mathcal{C}_{\text{sub}}$ is

$$\tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k,\,\mathcal{C}_{\text{sub}}} \;:=\; \mathbf{G}_k(\mathbf{A}, \mathbf{B})\, \Phi^{(L-2)}_{\text{sub},\,\mathcal{C}_{\text{sub}}} \;\in\; \mathbb{R}^{p_L \times p_1}, \tag{45}$$

where $\Phi^{(L-2)}_{\text{sub},\,\mathcal{C}_{\text{sub}}} \in \mathbb{R}^{p_{L-1} \times p_1}$ is the linear map of layers $2, \dots, L-1$ of the depth-$(L+1)$ network on $\mathcal{C}_{\text{sub}}$.

Decomposition of $\Phi^{(L-1)}$. The first $L-1$ layers of the depth-$(L+1)$ network are layer 1 followed by layers $2, \dots, L-1$. Within a cell $\mathcal{C}'$ of the depth-$(L+1)$ network, layer 1 acts as $\mathbf{D}\,\mathbf{W}_E$ and layers $2, \dots, L-1$ act as a linear transform $\Phi^{(L-2)}_{\text{sub},\,\mathcal{C}_{\text{sub}}}$. The $\mathcal{C}_{\text{sub}}$ is a cell in

the input space $\mathbb{R}^{p_1}$ of the depth-L sub-network, where the activation patterns of unit sets $2, \dots, L-1$ remain the same. Here $\mathcal{C}'$ is a set of inputs on which the activation patterns of unit sets $1, \dots, L-1$ are constant. The pattern of unit set 1 fixes $\mathbf{D}$ on $\mathcal{C}'$. The patterns of unit sets $2, \dots, L-1$ fix the sub-network cell that contains $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$, so this cell is the same $\mathcal{C}_{\text{sub}}$ for every $\boldsymbol{x} \in \mathcal{C}'$.

Composition gives

$$\Phi^{(L-1)}_{\mathcal{C}'} \;=\; \Phi^{(L-2)}_{\text{sub},\,\mathcal{C}_{\text{sub}}} \cdot \mathbf{D}\,\mathbf{W}_E \;\in\; \mathbb{R}^{p_{L-1} \times p_0}, \tag{46}$$

and therefore

$$\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}'} \;=\; \mathbf{G}_k\,\Phi^{(L-1)}_{\mathcal{C}'} \;=\; \tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k,\,\mathcal{C}_{\text{sub}}} \cdot \mathbf{D}\,\mathbf{W}_E \;\in\; \mathbb{R}^{p_L \times p_0}. \tag{47}$$

(47) is the recursion in cell-aware generators: the depth-$(L+1)$ generator equals the depth-L generator of the sub-network, right-multiplied by the layer-1 linear map $\mathbf{D}\,\mathbf{W}_E$.

Inductive hypothesis. Assume the column-covariance identity (37) holds at depth L for the sub-network: for every input $\boldsymbol{y} \in \mathbb{R}^{p_1}$ to the sub-network and every cell $\mathcal{C}_{\text{sub}}$ containing $\boldsymbol{y}$, and for all $j \in \{1, \dots, p_L\}$ and $a, a' \in \{1, \dots, p_1\}$,

$$\mathbb{E}\Big[\, \tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k,\,\mathcal{C}_{\text{sub}};\,ja}\, \tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k,\,\mathcal{C}_{\text{sub}};\,ja'} \,\Big] \;=\; c^{(L)}_{k,j}\,\delta_{aa'}, \tag{48}$$

where the expectation is over the sub-network's weights, biases, and the cell $\mathcal{C}_{\text{sub}}$ consistent with $\boldsymbol{y}$, and $c^{(L)}_{k,j}$ depends on (k, j) alone. The base case (42) shows (48) holds at $L = 2$.

Independence of the two stochastic groups. $\tilde{\mathbf{G}}^{(L)}_{\text{sub}}$ depends on the sub-network's weights $(\mathbf{W}^{(2)}, \dots, \mathbf{W}^{(L+1)})$, biases $(\boldsymbol{b}^{(2)}, \dots, \boldsymbol{b}^{(L+1)})$, and on $\mathcal{C}_{\text{sub}}$ (which is determined by these weights and the sub-network's input $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$). The pair $(\mathbf{W}_E, \mathbf{D})$ depends on $\mathbf{W}_E$ and $\boldsymbol{x}$ alone, with $\mathbf{W}_E$ independent of all sub-network weights and biases. Hence, conditional on the sub-network's input $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$ being fixed, $\tilde{\mathbf{G}}^{(L)}_{\text{sub}}$ and $(\mathbf{W}_E, \mathbf{D})$ are conditionally independent.

Squared norm of $\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}'}$ as a triple sum

Write $\mathbf{H} := \tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k,\,\mathcal{C}_{\text{sub}}} \in \mathbb{R}^{p_L \times p_1}$ and $\mathbf{M} := \mathbf{D}\,\mathbf{W}_E \in \mathbb{R}^{p_1 \times p_0}$, so $\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}'} = \mathbf{H}\,\mathbf{M}$. Expand the Frobenius norm:

$$\begin{aligned}
\|\mathbf{H}\,\mathbf{M}\|_F^2 \;&=\; \sum_{j=1}^{p_L}\sum_{s=1}^{p_0}(\mathbf{H}\,\mathbf{M})_{js}^2 \;=\; \sum_{j=1}^{p_L}\sum_{s=1}^{p_0}\Bigg(\sum_{a=1}^{p_1}\mathbf{H}_{ja}\,\mathbf{M}_{as}\Bigg)^2 \\
&=\; \sum_{j=1}^{p_L}\sum_{s=1}^{p_0}\sum_{a=1}^{p_1}\sum_{a'=1}^{p_1}\mathbf{H}_{ja}\,\mathbf{H}_{ja'}\,\mathbf{M}_{as}\,\mathbf{M}_{a's} \\
&=\; \sum_{j=1}^{p_L}\sum_{a=1}^{p_1}\sum_{a'=1}^{p_1}\mathbf{H}_{ja}\,\mathbf{H}_{ja'}\sum_{s=1}^{p_0}\mathbf{M}_{as}\,\mathbf{M}_{a's},
\end{aligned} \tag{49}$$

and substitute $\mathbf{M}_{as} = d_a\,(\mathbf{W}_E)_{as}$, $d_a^2 = d_a$:

$$\Big\|\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}'}\Big\|_F^2 \;=\; \sum_{j=1}^{p_L}\sum_{a=1}^{p_1}\sum_{a'=1}^{p_1}\tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k;\,ja}\,\tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k;\,ja'}\,d_a\,d_{a'}\,\langle(\mathbf{W}_E)_{a,:},\,(\mathbf{W}_E)_{a',:}\rangle, \tag{50}$$

where the cell subscripts on $\tilde{\mathbf{G}}^{(L)}_{\text{sub}}$ are suppressed for brevity.

Conditional column-covariance of $\tilde{\mathbf{G}}^{(L+1)}$

Compute $\mathbb{E}[\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt}\,\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt'}]$ for $j \in \{1, \dots, p_L\}$ and $t, t' \in \{1, \dots, p_0\}$. Expanding $\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt} = \sum_{a=1}^{p_1}\tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k;\,ja}\,d_a\,(\mathbf{W}_E)_{at}$,

$$\mathbb{E}\Big[\,\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt}\,\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt'}\,\Big] \;=\; \sum_{a=1}^{p_1}\sum_{a'=1}^{p_1}\mathbb{E}\Big[\,\tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k;\,ja}\,\tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k;\,ja'}\,d_a\,d_{a'}\,(\mathbf{W}_E)_{at}\,(\mathbf{W}_E)_{a't'}\,\Big]. \tag{51}$$

Conditional on $(\mathbf{W}_E, \boldsymbol{x})$, the sub-network's input $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$ is fixed, the factors $d_a, d_{a'}, (\mathbf{W}_E)_{at}, (\mathbf{W}_E)_{a't'}$ are deterministic, and $\tilde{\mathbf{G}}^{(L)}_{\text{sub}}$ is conditionally independent of $(\mathbf{W}_E, \mathbf{D})$ with conditional column-covariance (48):

$$\mathbb{E}\Big[\, \tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k;\,ja}\, \tilde{\mathbf{G}}^{(L)}_{\text{sub},\,k;\,ja'} \,\Big|\, \mathbf{W}_E,\, \boldsymbol{x} \Big] \;=\; c^{(L)}_{k,j}\, \delta_{aa'}. \tag{52}$$

Substituting (52) into (51) collapses the double sum over a, a' to a single sum over a:

$$\mathbb{E}\Big[\, \tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt}\, \tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt'} \Big] \;=\; c^{(L)}_{k,j} \sum_{a=1}^{p_1} \mathbb{E}[\, d_a\, (\mathbf{W}_E)_{at}\, (\mathbf{W}_E)_{at'}\,]\,. \tag{53}$$

The cross-moment $\mathbb{E}[d_a\,(\mathbf{W}_E)_{at}\,(\mathbf{W}_E)_{at'}]$

Section 4 obtains this cross-moment from the sign flip $(\mathbf{W}_E)_{a,:} \mapsto -(\mathbf{W}_E)_{a,:}$, which maps d_a to $1 - d_a$ and leaves $(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}$ unchanged. Here we verify it by direct computation.

Fix a, t, t'. Decompose $(\mathbf{W}_E)_{a,:} \in \mathbb{R}^{p_0}$ into its components parallel and orthogonal to $\hat{\boldsymbol{x}} := \boldsymbol{x}/\,\|\boldsymbol{x}\|_2$: with $z_a = (\mathbf{W}_E)_{a,:} \cdot \boldsymbol{x}$,

$$(\mathbf{W}_E)_{a,:} \;=\; \frac{z_a}{\|\boldsymbol{x}\|_2}\,\hat{\boldsymbol{x}} \;+\; \boldsymbol{w}_\perp, \qquad \boldsymbol{w}_\perp \;:=\; (\mathbf{W}_E)_{a,:} - \frac{z_a}{\|\boldsymbol{x}\|_2}\,\hat{\boldsymbol{x}}. \tag{54}$$

Coordinatewise, $(\mathbf{W}_E)_{at} = (z_a/\,\|\boldsymbol{x}\|_2)\,\hat{x}_t + (\boldsymbol{w}_\perp)_t$. The pair $((\mathbf{W}_E)_{a,:} \cdot \hat{\boldsymbol{x}}, \boldsymbol{w}_\perp)$ is jointly Gaussian, and $\mathrm{Cov}((\mathbf{W}_E)_{a,:} \cdot \hat{\boldsymbol{x}}, (\boldsymbol{w}_\perp)_t) = 0$ by direct computation, so $z_a \perp \boldsymbol{w}_\perp$. Marginal variances: $z_a \sim \mathcal{N}(0, \sigma_z^2)$ with $\sigma_z^2 := \sigma_1^2\,\|\boldsymbol{x}\|_2^2/p_0$, and the entries of $\boldsymbol{w}_\perp$ have covariance $\mathrm{Cov}((\boldsymbol{w}_\perp)_t, (\boldsymbol{w}_\perp)_{t'}) = (\sigma_1^2/p_0)\,(\delta_{tt'} - \hat{x}_t\,\hat{x}_{t'})$. Since $d_a = \mathbf{1}\{z_a > 0\}$ depends only on z_a, also $d_a \perp \boldsymbol{w}_\perp$.

Expand $(\mathbf{W}_E)_{at}\,(\mathbf{W}_E)_{at'}$:

$$(\mathbf{W}_E)_{at}\,(\mathbf{W}_E)_{at'} \;=\; \frac{z_a^2}{\|\boldsymbol{x}\|_2^2}\,\hat{x}_t\,\hat{x}_{t'} \;+\; \frac{z_a}{\|\boldsymbol{x}\|_2}\,\big(\hat{x}_t\,(\boldsymbol{w}_\perp)_{t'} + \hat{x}_{t'}\,(\boldsymbol{w}_\perp)_t\big) \;+\; (\boldsymbol{w}_\perp)_t\,(\boldsymbol{w}_\perp)_{t'}. \tag{55}$$

Multiplying by d_a and taking expectation, term by term:

(i) Symmetry of z_a gives $\mathbb{E}[d_a\,z_a^2] = \mathbb{E}[z_a^2\,\mathbf{1}\{z_a > 0\}] = \frac{1}{2}\,\mathbb{E}[z_a^2] = \sigma_z^2/2$, hence

$$\mathbb{E}\left[\, d_a\,\frac{z_a^2}{\|\boldsymbol{x}\|_2^2}\,\hat{x}_t\,\hat{x}_{t'} \right] \;=\; \hat{x}_t\,\hat{x}_{t'}\,\frac{\sigma_z^2}{2\,\|\boldsymbol{x}\|_2^2} \;=\; \hat{x}_t\,\hat{x}_{t'}\,\frac{\sigma_1^2}{2p_0}.$$

(ii) The two cross terms factor as $\mathbb{E}[d_a\,z_a]\cdot\mathbb{E}[(\boldsymbol{w}_\perp)_t]$ and $\mathbb{E}[d_a\,z_a]\cdot\mathbb{E}[(\boldsymbol{w}_\perp)_{t'}]$ by $d_a, z_a \perp \boldsymbol{w}_\perp$, and $\mathbb{E}[(\boldsymbol{w}_\perp)_t] = 0$, so both vanish.

(iii) $d_a \perp \boldsymbol{w}_\perp$ and $P(d_a = 1) = 1/2$ give

$$\mathbb{E}[\, d_a\,(\boldsymbol{w}_\perp)_t\,(\boldsymbol{w}_\perp)_{t'}\,] \;=\; \tfrac{1}{2}\,\mathbb{E}[\,(\boldsymbol{w}_\perp)_t\,(\boldsymbol{w}_\perp)_{t'}\,] \;=\; \frac{\sigma_1^2}{2p_0}\,(\delta_{tt'} - \hat{x}_t\,\hat{x}_{t'}).$$

Adding *(i)*, *(ii)*, *(iii)*:

$$\mathbb{E}[\, d_a\,(\mathbf{W}_E)_{at}\,(\mathbf{W}_E)_{at'}\,] \;=\; \frac{\sigma_1^2}{2p_0}\,\hat{x}_t\,\hat{x}_{t'} \;+\; \frac{\sigma_1^2}{2p_0}\,(\delta_{tt'} - \hat{x}_t\,\hat{x}_{t'}) \;=\; \frac{\sigma_1^2}{2p_0}\,\delta_{tt'}, \tag{56}$$

the $\hat{x}_t\,\hat{x}_{t'}$ terms cancelling. The right-hand side is independent of a and of $\boldsymbol{x}$.

Conclusion of the inductive step

Substituting (56) into (53):

$$\mathbb{E}\Big[\, \tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt}\, \tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt'} \Big] \;=\; c^{(L)}_{k,j} \sum_{a=1}^{p_1} \frac{\sigma_1^2}{2p_0}\,\delta_{tt'} \;=\; c^{(L+1)}_{k,j}\,\delta_{tt'}, \qquad c^{(L+1)}_{k,j} \;:=\; \frac{p_1\,\sigma_1^2}{2p_0}\,c^{(L)}_{k,j}. \tag{57}$$

The right-hand side has the form $c_{k,j}^{(L+1)}\,\delta_{tt'}$ with $c_{k,j}^{(L+1)}$ depending on (k,j) only, so (37) holds at depth $L+1$.

Summing the diagonal entries of (57) over $k \in \{1,2\}$, $j \in \{1,\dots,p_L\}$, and $t \in \{1,\dots,p_0\}$:

$$\Gamma_{L+1} \;=\; \sum_{k=1}^{2}\sum_{j=1}^{p_L}\sum_{t=1}^{p_0}\mathbb{E}\Big[\,\big(\tilde{\mathbf{G}}^{(L+1)}_{k,\mathcal{C}';\,jt}\big)^2\,\Big] \;=\; \sum_{k=1}^{2}\sum_{j=1}^{p_L} p_0\, c_{k,j}^{(L+1)} \;=\; \frac{\sigma_1^2}{2}\sum_{k=1}^{2}\sum_{j=1}^{p_L} p_1\, c_{k,j}^{(L)}. \tag{58}$$

The same summation applied to the sub-network at depth L (with input dim p_1, so the inner sum runs over $a \in \{1,\dots,p_1\}$) gives

$$\Gamma_{L} \;=\; \sum_{k=1}^{2}\sum_{j=1}^{p_L}\sum_{a=1}^{p_1}\mathbb{E}\Big[\,\big(\tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k;\,ja}\big)^2\,\Big] \;=\; \sum_{k=1}^{2}\sum_{j=1}^{p_L} p_1\, c_{k,j}^{(L)}. \tag{59}$$

Combining (58) and (59),

$$\Gamma_{L+1} \;=\; \frac{\sigma_1^2}{2}\,\Gamma_L, \tag{60}$$

where σ_1^2 is the variance of the prepended layer (layer 1 of the depth-$(L+1)$ network).

B.3 Iterating: the closed form

Fix a target depth $L \geq 2$ and weight variances $(\sigma_l)_{l=1}^{L}$. Apply the inductive step of §B.2 $L-2$ times, starting from the two-layer sub-network of layers $L-1, L$ and prepending one layer per iteration. At iteration $r \in \{1,\dots,L-2\}$ the existing sub-network has depth $r+1$ and consists of layers $L-r, L-r+1,\dots,L$; the prepended layer is layer $L-r-1$, with variance σ^2_{L-r-1}. The inductive step (60) (applied with the prepended layer's variance playing the role of "σ_1^2") gives

$$\Gamma_{r+2} \;=\; \frac{\sigma^2_{L-r-1}}{2}\,\Gamma_{r+1}, \qquad r = 1,\,\dots,\,L-2.$$

Starting from $\Gamma_2 = 2\,\sigma^2_{L-1}\,\sigma^2_L(1+1/\pi)$ (the base-case formula (43) applied to the top two-layer sub-network: its first layer's variance is σ^2_{L-1}, its second is σ^2_L), and chaining the $L-2$ recursions:

$$\Gamma_L \;=\; 2\,\sigma^2_{L-1}\,\sigma^2_L\Big(1+\frac{1}{\pi}\Big)\prod_{r=1}^{L-2}\frac{\sigma^2_{L-r-1}}{2} \;=\; 2\,\sigma^2_{L-1}\,\sigma^2_L\Big(1+\frac{1}{\pi}\Big)\prod_{l=1}^{L-2}\frac{\sigma^2_l}{2} \;=\; 8\Big(1+\frac{1}{\pi}\Big)\prod_{l=1}^{L}\frac{\sigma^2_l}{2}, \tag{61}$$

the second equality reindexing $l = L-r-1$ over $r \in \{1,\dots,L-2\}$, and the last equality absorbing the $\sigma^2_{L-1}\sigma^2_L$ from the base case into the product over l (using $2\cdot\sigma^2_{L-1}\sigma^2_L = 8\cdot(\sigma^2_{L-1}/2)(\sigma^2_L/2)$). This is (14) of Theorem 3.1 in the zero-bias case. The uniform-variance closed form $\Gamma_L = 8(1+1/\pi)(\sigma_w^2/2)^L$ of Corollary 3.2 is the special case $\sigma_l \equiv \sigma_w$.

B.4 Bias invariance

The argument of §B.1–§B.3 nowhere uses that the biases are zero. We now check that the inductive step (60) and the base case (43) are unchanged when each layer carries an independent Gaussian bias $\boldsymbol{b}^{(l)}$ with components $b_a^{(l)} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2_{b,l})$, $\sigma_{b,l} \geq 0$, independent of all weights and of all other biases.

Base case is bias-free by construction. The base-case generator $\mathbf{G}_k(\mathbf{A},\mathbf{B})$ is built from $\beta^{\pm}_{k,j}$ entries of $\mathbf{B} = \mathbf{W}^{(L)}$ and from A_{ji} entries of $\mathbf{A} = \mathbf{W}^{(L-1)}$ (5)–(6). Neither $\beta^{\pm}_{k,j}$ nor A_{ji} involves the biases $\boldsymbol{b}^{(L-1)}, \boldsymbol{b}^{(L)}$ of the last two layers. The column-covariance computation (42) factors through $\mathbb{E}_{\mathbf{B}}[\beta^2_{k,j}]$ and $\mathbb{E}_{\mathbf{A}}[A_{ji}A_{ji'}]$ alone, and $c^{(2)}_{k,j} = \sigma_1^2\sigma_2^2(1+1/\pi)/(p_0p_1)$, $\Gamma_2 = C_0$, are unchanged.

Inductive step under bias. The only place the inductive step §B.2 uses the activation indicator is in the cross-moment (56),

$$\mathbb{E}[\, d_a\, (\mathbf{W}_E)_{at}\, (\mathbf{W}_E)_{at'}\,] \;=\; \frac{\sigma_1^2}{2p_0}\,\delta_{tt'},$$

through the activation indicator $d_a = \mathbf{1}\{z_a > 0\}$ at the prepended layer. Replace the indicator by its biased counterpart:

$$d_a \;:=\; \mathbf{1}\{z_a + b_a > 0\}, \qquad b_a \;:=\; b_a^{(1)} \overset{\text{iid}}{\sim} \mathcal{N}(0,\, \sigma_{b,1}^2), \quad b_a \;\perp\; \mathbf{W}_E. \tag{62}$$

We verify that the cross-moment (56) is unchanged under (62).

The parallel/orthogonal decomposition (54) is a property of $\mathbf{W}_E$ and does not involve b_a, so it still holds with the same independence $z_a \perp \boldsymbol{w}_\perp$ and the same marginal distributions. Since $b_a \perp \mathbf{W}_E$, also $b_a \perp \boldsymbol{w}_\perp$, and therefore $\boldsymbol{w}_\perp \perp (z_a, b_a)$ and $\boldsymbol{w}_\perp \perp d_a$. The expansion (55) of $(\mathbf{W}_E)_{at}\,(\mathbf{W}_E)_{at'}$ is unchanged. We re-evaluate the three terms of (55) multiplied by d_a.

(i) $\mathbb{E}[d_a\, z_a^2]$. Condition on b_a:

$$\mathbb{E}\big[\, \mathbf{1}\{z_a + b_a > 0\}\, z_a^2 \,\big|\, b_a \,\big] \;=\; \mathbb{E}\big[\, z_a^2\, \mathbf{1}\{z_a > -b_a\} \,\big|\, b_a \,\big]\,.$$

For any fixed $t \in \mathbb{R}$, by symmetry $z_a \overset{d}{=} -z_a$,

$$\begin{aligned}\mathbb{E}[z_a^2\, \mathbf{1}\{z_a > -t\}] + \mathbb{E}[z_a^2\, \mathbf{1}\{z_a > t\}] \;&=\; \mathbb{E}[z_a^2\, \mathbf{1}\{z_a > -t\}] + \mathbb{E}[(-z_a)^2\, \mathbf{1}\{-z_a > t\}] \\ &=\; \mathbb{E}[z_a^2\, \mathbf{1}\{z_a > -t\}] + \mathbb{E}[z_a^2\, \mathbf{1}\{z_a < -t\}] \;=\; \mathbb{E}[z_a^2] \;=\; \sigma_z^2.\end{aligned} \tag{63}$$

The distribution of b_a is symmetric ($b_a \overset{d}{=} -b_a$), so $\mathbb{E}_{b_a}[\mathbb{E}[z_a^2 \mathbf{1}\{z_a > -b_a\}]] = \mathbb{E}_{b_a}[\mathbb{E}[z_a^2 \mathbf{1}\{z_a > b_a\}]]$, and adding the two and applying (63) pointwise gives $2\,\mathbb{E}[d_a\, z_a^2] = \mathbb{E}[\sigma_z^2] = \sigma_z^2$, hence

$$\mathbb{E}[d_a\, z_a^2] \;=\; \sigma_z^2/2, \tag{64}$$

the same value as in the zero-bias case.

(ii) The two cross terms in (55) multiplied by d_a have the form $\hat{x}_t\, \mathbb{E}[d_a\, z_a\, (\boldsymbol{w}_\perp)_{t'}]$ and $\hat{x}_{t'}\, \mathbb{E}[d_a\, z_a\, (\boldsymbol{w}_\perp)_t]$. By $\boldsymbol{w}_\perp \perp (z_a, b_a, d_a)$,

$$\mathbb{E}[d_a\, z_a\, (\boldsymbol{w}_\perp)_{t'}] \;=\; \mathbb{E}[d_a\, z_a]\; \mathbb{E}[(\boldsymbol{w}_\perp)_{t'}] \;=\; \mathbb{E}[d_a\, z_a] \cdot 0 \;=\; 0,$$

and similarly for the second cross term, regardless of the value of $\mathbb{E}[d_a\, z_a]$ (which itself does depend on $\sigma_{b,1}$ but is irrelevant here).

(iii) For the $(\boldsymbol{w}_\perp)_t (\boldsymbol{w}_\perp)_{t'}$ piece, $\boldsymbol{w}_\perp \perp d_a$ gives $\mathbb{E}[d_a\, (\boldsymbol{w}_\perp)_t\, (\boldsymbol{w}_\perp)_{t'}] = P(d_a = 1)\, \mathbb{E}[(\boldsymbol{w}_\perp)_t\, (\boldsymbol{w}_\perp)_{t'}]$. Since $z_a + b_a \sim \mathcal{N}(0, \sigma_z^2 + \sigma_{b,1}^2)$ is symmetric about zero, $P(d_a = 1) = P(z_a + b_a > 0) = 1/2$, the same value as in the zero-bias case. The covariance $\mathbb{E}[(\boldsymbol{w}_\perp)_t (\boldsymbol{w}_\perp)_{t'}] = (\sigma_1^2/p_0)(\delta_{tt'} - \hat{x}_t \hat{x}_{t'})$ is unchanged ($\boldsymbol{w}_\perp$ is a function of $\mathbf{W}_E$ alone). Hence

$$\mathbb{E}[d_a\, (\boldsymbol{w}_\perp)_t\, (\boldsymbol{w}_\perp)_{t'}] \;=\; \frac{\sigma_1^2}{2p_0}\,(\delta_{tt'} - \hat{x}_t\, \hat{x}_{t'}), \tag{65}$$

the same value as in the zero-bias case.

Adding (i)–(iii) gives (56):

$$\mathbb{E}[\, d_a\, (\mathbf{W}_E)_{at}\, (\mathbf{W}_E)_{at'}\,] \;=\; \frac{\sigma_1^2}{2p_0}\,\hat{x}_t\, \hat{x}_{t'} \;+\; 0 \;+\; \frac{\sigma_1^2}{2p_0}\,(\delta_{tt'} - \hat{x}_t\, \hat{x}_{t'}) \;=\; \frac{\sigma_1^2}{2p_0}\,\delta_{tt'}.$$

Every other step in §B.2 from (51) onwards uses $\mathbb{E}[d_a (\mathbf{W}_E)_{at} (\mathbf{W}_E)_{at'}]$ as a black box. (57) and (60) therefore hold under the biased indicator (62).

This handles the prepended layer's bias only. The other biases are absorbed by the iteration of §B.3: at each of the $L-2$ prepend steps the layer being added plays the role of "layer 1" in this section, so its bias is treated by the argument above; the remaining biases sit inside the current sub-network and are not yet active. The biases on the top two layers (variances $\sigma_{b,L-1}^2, \sigma_{b,L}^2$) enter only the base case, which is bias-free by construction since $\beta_{k,j}$ and A_{ji} in (5)–(6) depend on $\mathbf{W}^{(L-1)}, \mathbf{W}^{(L)}$ alone. By induction the column-covariance identity (37) and the recursion $\Gamma_{L+1} = (\sigma_1^2/2)\Gamma_L$ hold for per-layer biases with variances $(\sigma_{b,l}^2)_{l=1}^{L}$, $\sigma_{b,l} \geq 0$, and iterating recovers (61). This completes the proof of Theorem 3.1. $\square$

C Variance of the generator energy

Appendix B computes the mean $\Gamma_L = \mathbb{E}[Z_L]$ of the generator energy Z_L of (9), under the ensemble (3) with $p_L = 2$. This appendix proves Theorem 3.3 of Section 3, which gives the variance of Z_L under the same ensemble, at an input $\boldsymbol{x} \neq \mathbf{0}$. Like the mean, the variance does not depend on the input or on the bias variances. We write $m := p_{L-1}$ for the width of the last hidden layer.

Theorem 3.3 (restated). Let $L \geq 2$, let the widths be arbitrary, and let $\boldsymbol{x} \neq \mathbf{0}$.

(a) $\mathrm{Var}(Z_L)$ depends on the weight variances $\sigma_1^2, \ldots, \sigma_L^2$ and the widths $p_0, \ldots, p_{L-1}$. It depends neither on $\boldsymbol{x}$ nor on the bias variances $\sigma_{b,1}^2, \ldots, \sigma_{b,L}^2$.

(b) At depth two,

$$\mathrm{Var}(Z_2) \;=\; \frac{2\,\sigma_1^4\,\sigma_2^4}{p_0\,p_1}\left[\left(5+\frac{8}{\pi}\right)(p_0+2) - 2\left(1+\frac{1}{\pi}\right)^2 p_0\right]. \tag{66}$$

(c) At depth L, $\mathrm{Var}(Z_L) = \mathbb{E}[Z_L^2] - \Gamma_L^2$, where Γ_L is the closed form (61) and $\mathbb{E}[Z_L^2]$ is the combination (85) of three pairs of scalars $(P^{(L)}, Q^{(L)})$. Each pair starts from its depth-two value in (77). Each of the layers $L-2, L-3, \ldots, 1$, taken in this order, with fan-in n', fan-out n and weight variance σ^2, maps the pair (P, Q) to

$$P' \;=\; \varsigma^2\Big(\frac{n^2+n}{4}\,P + n\,Q\Big), \qquad Q' \;=\; \varsigma^2\Big(\frac{n}{2}\,P + \frac{n^2+3n}{4}\,Q\Big), \qquad \varsigma \;:=\; \frac{\sigma^2}{n'}. \tag{67}$$

Remark C.1. Gaussianity is used only for the weights, through Wick's identity and the decomposition (54). The biases enter the proof only through the symmetry $b_a^{(l)} \overset{d}{=} -b_a^{(l)}$. So part (a) holds for any independent biases whose distributions are symmetric about zero.

C.1 The variance reduces to fourth moments of the generators

Since Z_L is the sum of the squares of the entries of the two generators, $\mathbb{E}[Z_L^2]$ is a sum of fourth moments of these entries. So Theorem 3.3 follows once these fourth moments are known. We compute them in the same way as Appendix B computes the second moments. There, the second moments of the generators of a depth-two network have the form $c\,\delta_{aa'}$ in the column indices, and prepending a layer at the input keeps this form and multiplies c by a factor (§B.2). Here, the fourth moments of the generators of a depth-two network are a combination of the three pairings of the four column indices, with two scalar coefficients P and Q. Prepending a layer keeps this form and maps (P, Q) linearly. Summing the fourth moments of the full network then gives $\mathbb{E}[Z_L^2]$.

The biases drop out of this computation for the following reason. Let $\mathbf{W}_E$ be the weight matrix of the prepended layer and d_a the activation indicator of its unit a, as in §B.2. A moment of the prepended layer depends on its bias only when it contains an odd number of weights from one row of $\mathbf{W}_E$, for example

$$\mathbb{E}[\,d_a\,(\mathbf{W}_E)_{at}\,] \;=\; \frac{\sigma_1^2}{p_0}\,x_t\,\frac{1}{\sqrt{2\pi s^2}}, \qquad s^2 \;=\; \frac{\sigma_1^2\,\|\boldsymbol{x}\|_2^2}{p_0} + \sigma_{b,1}^2, \tag{68}$$

which depends on $\sigma_{b,1}$. In the pairing form the column indices of the sub-network come in equal pairs, so every row of $\mathbf{W}_E$ contributes an even number of weights. By the sign-flip symmetry of §B.4, a moment with an even number of weights from each row does not depend on the bias. So the odd moments never enter the computation.

C.2 The fourth moments consist of three pairings

The induction of §B.2 carries the column-covariance identity (37),

$$\mathbb{E}\Big[\,\tilde{\mathbf{G}}^{(L)}_{k,\mathcal{C};\,ja}\,\tilde{\mathbf{G}}^{(L)}_{k,\mathcal{C};\,ja'}\,\Big] \;=\; c^{(L)}\,\delta_{aa'}, \tag{69}$$

for every input and every cell containing it, where the constant $c^{(L)}$ is the same for all k and j by (42). The induction of this appendix carries, in addition, the following statement about the fourth moments. At depth L, for every input, every cell containing it, all $k, k' \in \{1, 2\}$, all rows j, i and all column indices a, a', c, c',

$$\mathbb{E}\Big[\, \tilde{\mathbf{G}}^{(L)}_{k;\,ja}\, \tilde{\mathbf{G}}^{(L)}_{k;\,ja'}\, \tilde{\mathbf{G}}^{(L)}_{k';\,ic}\, \tilde{\mathbf{G}}^{(L)}_{k';\,ic'} \Big] \;=\; P^{(L)}_{kj,k'i}\, \delta_{aa'}\, \delta_{cc'} \;+\; Q^{(L)}_{kj,k'i}\, (\delta_{ac}\, \delta_{a'c'} + \delta_{ac'}\, \delta_{a'c})\,, \tag{70}$$

where the scalars $P^{(L)}_{kj,k'i}$ and $Q^{(L)}_{kj,k'i}$ do not depend on the input, the cell or any bias, and depend on the indices only through whether $j = i$ and whether $k = k'$. The three terms of (70) are the three ways of pairing the column indices a, a', c, c'. The coefficients therefore form three pairs, which we call the three tracks,

$$(P_{\text{off}}, Q_{\text{off}}) \quad (j \neq i), \qquad (P_{=}, Q_{=}) \quad (j = i,\, k = k'), \qquad (P_{\times}, Q_{\times}) \quad (j = i,\, k \neq k'). \tag{71}$$

C.3 The pairing coefficients of a two-layer network

The induction starts at $L = 2$. There the generator is $(\mathbf{G}_k)_{ja} = \beta_{k,j} A_{ja}$ with $\beta_{1,j} = B^{+}_{1j} + B^{-}_{2j}$ and $\beta_{2,j} = B^{+}_{2j} + B^{-}_{1j}$ (§B.1). It involves no input, no cell and no bias, so at $L = 2$ the independence stated in (70) needs no argument, and only the coefficients need to be computed. The factor $\beta_{k,j}$ contributes fourth moments of the output weights, and the factor A_{ja} contributes fourth moments of the first-layer weights.

Half-normal moments. For $u \sim \mathcal{N}(0, s^2)$, extending (40) to third and fourth order,

$$\mathbb{E}[u^+] = \frac{s}{\sqrt{2\pi}}, \qquad \mathbb{E}[(u^+)^2] = \frac{s^2}{2}, \qquad \mathbb{E}[(u^+)^3] = s^3\sqrt{\frac{2}{\pi}}, \qquad \mathbb{E}[(u^+)^4] = \frac{3s^4}{2}, \tag{72}$$

together with $u^+ u^- = 0$ pointwise. Set $s^2 := \sigma_2^2 / p_1$ and write $u := B_{1j}$ and $v := B_{2j}$, which are independent $\mathcal{N}(0, s^2)$. Expanding $\beta^4_{1,j} = (u^+ + v^-)^4$ binomially and using $u \perp v$,

$$\begin{aligned}
\mathbb{E}[\beta^4_{1,j}] \;&=\; \mathbb{E}[(u^+)^4] + 4\, \mathbb{E}[(u^+)^3]\, \mathbb{E}[v^-] + 6\, \mathbb{E}[(u^+)^2]\, \mathbb{E}[(v^-)^2] \\
&\qquad + 4\, \mathbb{E}[u^+]\, \mathbb{E}[(v^-)^3] + \mathbb{E}[(v^-)^4] \\
&=\; \frac{3s^4}{2} + \frac{4s^4}{\pi} + \frac{3s^4}{2} + \frac{4s^4}{\pi} + \frac{3s^4}{2} \\
&=\; s^4 \left(\frac{9}{2} + \frac{8}{\pi} \right),
\end{aligned} \tag{73}$$

using $\mathbb{E}[(u^+)^3]\, \mathbb{E}[v^-] = s^3\sqrt{2/\pi} \cdot s/\sqrt{2\pi} = s^4/\pi$. The same value holds for $\mathbb{E}[\beta^4_{2,j}]$ by the swap $u \leftrightarrow v$. For the cross product, we expand $\beta^2_{1,j}\beta^2_{2,j} = (u^+ + v^-)^2 (v^+ + u^-)^2$. Every term containing u^+u^- or v^+v^- vanishes pointwise, which leaves only $(u^+)^2(v^+)^2$ and $(v^-)^2(u^-)^2$, so

$$\mathbb{E}[\beta^2_{1,j}\, \beta^2_{2,j}] \;=\; 2 \cdot \frac{s^2}{2} \cdot \frac{s^2}{2} \;=\; \frac{s^4}{2}. \tag{74}$$

Wick contraction of A. The entries of $\mathbf{A}$ are i.i.d. $\mathcal{N}(0, \tau)$ with $\tau := \sigma_1^2 / p_0$, so the Gaussian fourth-moment (Wick) identity gives, within one row j,

$$\mathbb{E}[\, A_{ja} A_{ja'} A_{jc} A_{jc'}\,] \;=\; \tau^2\, (\delta_{aa'}\delta_{cc'} + \delta_{ac}\delta_{a'c'} + \delta_{ac'}\delta_{a'c})\,, \tag{75}$$

while for $j \neq i$ the rows factor,

$$\mathbb{E}[A_{ja} A_{ja'}]\, \mathbb{E}[A_{ic} A_{ic'}] \;=\; \tau^2 \delta_{aa'} \delta_{cc'}. \tag{76}$$

Base coefficients. The scalar $\beta_{k,j}$ depends on column j of $\mathbf{B}$ alone, the columns of $\mathbf{B}$ are independent, and $\mathbf{A} \perp \mathbf{B}$. Combining (41), (73), (74), (75) and (76),

$$\begin{aligned}
j \neq i: \quad & P^{(2)}_{\text{off}} = s^4 \Big(1 + \tfrac{1}{\pi}\Big)^2 \tau^2, && Q^{(2)}_{\text{off}} = 0; \\
j = i,\ k = k': \quad & P^{(2)}_{=} = Q^{(2)}_{=} = s^4 \Big(\tfrac{9}{2} + \tfrac{8}{\pi}\Big) \tau^2; \\
j = i,\ k \neq k': \quad & P^{(2)}_{\times} = Q^{(2)}_{\times} = \tfrac{s^4}{2}\, \tau^2.
\end{aligned} \tag{77}$$

This proves (70) at $L = 2$. For a depth-two sub-network of a deeper network, τ and s^2 are formed from the weight variances and fan-ins of its own two layers, as in §B.3.

C.4 Prepending a layer updates the coefficients linearly

Assume (70) at depth L. We prepend one layer and show that the fourth moments at depth $L+1$ have the same pairing form, with coefficients obtained from $(P^{(L)}, Q^{(L)})$ by a linear map. The setup is that of §B.2. The network has depth $L+1$, its layer 1 has weights $\mathbf{W}_E \in \mathbb{R}^{p_1 \times p_0}$ with entries of variance $\varsigma := \sigma_1^2/p_0$ and biased indicators $d_a = \mathbf{1}\{z_a + b_a > 0\}$ as in (62), and the depth-L sub-network above it receives the input $\mathbf{D}\,\mathbf{W}_E\,\boldsymbol{x}$. The recursion (47) gives $\tilde{\mathbf{G}}^{(L+1)}_{k;\,jt} = \sum_a \tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k;\,ja}\, d_a (\mathbf{W}_E)_{at}$, hence

$$\mathbb{E}\Big[\, \tilde{\mathbf{G}}^{(L+1)}_{k;\,jt}\, \tilde{\mathbf{G}}^{(L+1)}_{k;\,jt'}\, \tilde{\mathbf{G}}^{(L+1)}_{k';\,iu}\, \tilde{\mathbf{G}}^{(L+1)}_{k';\,iu'} \,\Big] = \sum_{a,a',c,c'} \mathbb{E}\Big[\, \tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k;\,ja}\, \tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k;\,ja'}\, \tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k';\,ic}\, \tilde{\mathbf{G}}^{(L)}_{\mathrm{sub},\,k';\,ic'} \times d_a d_{a'} d_c d_{c'}\, (\mathbf{W}_E)_{at}\, (\mathbf{W}_E)_{a't'}\, (\mathbf{W}_E)_{cu}\, (\mathbf{W}_E)_{c'u'} \,\Big]. \tag{78}$$

Conditional on $(\mathbf{W}_E, \boldsymbol{b}^{(1)}, \boldsymbol{x})$, the input of the sub-network is fixed and the factors $d_\bullet$ and $(\mathbf{W}_E)_\bullet$ are deterministic, while $\tilde{\mathbf{G}}^{(L)}_{\mathrm{sub}}$ is conditionally independent of them, by the argument with two independent groups of parameters in §B.2. The inductive hypothesis (70) therefore replaces the sub-network factor by the pairing form with coefficients $(P^{(L)}, Q^{(L)})$. These coefficients do not depend on the input of the sub-network, so the conditioning can be removed. The sum (78) then reduces to the three pairings of (a, a', c, c'). In each pairing every row of $\mathbf{W}_E$ carries an even number of weights, so only two one-row moments are needed.

One row, two weights. This is the cross-moment (56), whose version with bias is proved in §B.4:

$$\mathbb{E}[\, d_a\, (\mathbf{W}_E)_{at} (\mathbf{W}_E)_{at'} \,] \;=\; \frac{\varsigma}{2}\, \delta_{tt'}. \tag{79}$$

It does not depend on $\boldsymbol{x}$, because the $\hat{x}_t \hat{x}_{t'}$ contributions of the parallel and orthogonal parts cancel, and it does not depend on $\sigma_{b,1}$.

One row, four weights. The joint law of $((\mathbf{W}_E)_{a,:}, b_a)$ is invariant under the sign flip $((\mathbf{W}_E)_{a,:}, b_a) \to (-(\mathbf{W}_E)_{a,:}, -b_a)$, which maps d_a to $1 - d_a$ almost surely and leaves the product $(\mathbf{W}_E)_{at}(\mathbf{W}_E)_{at'}(\mathbf{W}_E)_{au}(\mathbf{W}_E)_{au'}$ unchanged. Hence $\mathbb{E}[d_a \cdot] = \frac{1}{2}\,\mathbb{E}[\,\cdot\,]$ for this product, and Wick's identity gives

$$\mathbb{E}[\, d_a\, (\mathbf{W}_E)_{at} (\mathbf{W}_E)_{at'} (\mathbf{W}_E)_{au} (\mathbf{W}_E)_{au'} \,] \;=\; \frac{\varsigma^2}{2}\, (\delta_{tt'}\delta_{uu'} + \delta_{tu}\delta_{t'u'} + \delta_{tu'}\delta_{t'u}), \tag{80}$$

which depends neither on $\boldsymbol{x}$ nor on $\sigma_{b,1}$. Both (79) and (80) use only the symmetry of the pair of a weight row and its bias, which is the origin of Remark C.1. An odd number of weights from one row would give the moment (68), which depends on the bias. Such a term would need the sub-network tensor to be nonzero at unpaired column indices, and (70) excludes this. This is where every $\sigma_{b,l}$ leaves the computation.

Collecting the three pairings. For the pairing $\delta_{aa'}\delta_{cc'}$, with coefficient $P^{(L)}$, the sum over $a = a'$ and $c = c'$ splits into the terms with $a \neq c$, where the two rows are independent and (79) applies twice, and the terms with $a = c$, where $d_a^2 = d_a$ and (80) applies:

$$\begin{aligned} &\sum_{a \neq c} \Big(\frac{\varsigma}{2}\Big)^2 \delta_{tt'}\delta_{uu'} + \sum_{a=c} \frac{\varsigma^2}{2}\, (\delta_{tt'}\delta_{uu'} + \delta_{tu}\delta_{t'u'} + \delta_{tu'}\delta_{t'u}) \\ &\qquad = p_1(p_1 - 1)\frac{\varsigma^2}{4}\, \delta_{tt'}\delta_{uu'} + p_1 \frac{\varsigma^2}{2}\, (\delta_{tt'}\delta_{uu'} + \delta_{tu}\delta_{t'u'} + \delta_{tu'}\delta_{t'u}). \end{aligned} \tag{81}$$

For the pairing $\delta_{ac}\delta_{a'c'}$, with coefficient $Q^{(L)}$, the same split over $a \neq a'$ and $a = a'$ gives

$$p_1(p_1 - 1)\frac{\varsigma^2}{4}\, \delta_{tu}\delta_{t'u'} + p_1 \frac{\varsigma^2}{2}\, (\delta_{tt'}\delta_{uu'} + \delta_{tu}\delta_{t'u'} + \delta_{tu'}\delta_{t'u}), \tag{82}$$

and the pairing $\delta_{ac'}\delta_{a'c}$ gives the same expression with $\delta_{tu'}\delta_{t'u}$ in the first term. We substitute these into (78) and read off the coefficients of $\delta_{tt'}\delta_{uu'}$ and of $\delta_{tu}\delta_{t'u'}$. The coefficient of $\delta_{tu'}\delta_{t'u}$ equals the latter by symmetry. This gives (70) at depth $L+1$ with

$$P^{(L+1)} = \varsigma^2\left(\frac{p_1^2+p_1}{4}\,P^{(L)} + p_1\,Q^{(L)}\right), \qquad Q^{(L+1)} = \varsigma^2\left(\frac{p_1}{2}\,P^{(L)} + \frac{p_1^2+3p_1}{4}\,Q^{(L)}\right), \tag{83}$$

the same linear map for each of the three tracks. This is (17) with $n' = p_0$ and $n = p_1$. The new coefficients do not depend on the input, the cell or any bias, which completes the induction. As in §B.3, we apply this step once for each prepended layer, with ς and p_1 replaced by σ_l^2/p_{l-1} and p_l of that layer. This carries the base values (77) of the top two-layer block down to the full depth-L network.

C.5 From the fourth moments to $\mathrm{Var}(Z_L)$

It remains to sum the fourth moments. Since $Z_L = \sum_{k,j,t}\big(\tilde{\mathbf{G}}^{(L)}_{k;\,jt}\big)^2$, the second moment of Z_L is the fourth-moment tensor evaluated at the paired column indices (t,t,u,u) and summed. By (70),

$$\mathbb{E}\Big[\big(\tilde{\mathbf{G}}^{(L)}_{k;\,jt}\big)^2\big(\tilde{\mathbf{G}}^{(L)}_{k';\,iu}\big)^2\Big] = P^{(L)}_{kj,k'i} + 2\,Q^{(L)}_{kj,k'i}\,\delta_{tu}, \qquad \sum_{t,u=1}^{p_0}(\cdots) = P^{(L)}_{kj,k'i}\,p_0^2 + 2\,Q^{(L)}_{kj,k'i}\,p_0. \tag{84}$$

Among the index pairs (k,j) and (k',i), there are $4m(m-1)$ with $j\neq i$, $2m$ with $j=i$ and $k=k'$, and $2m$ with $j=i$ and $k\neq k'$, where $m = p_{L-1}$. Summing over them,

$$\begin{aligned}\mathbb{E}[Z_L^2] \;=\; 4m(m-1)\left(P^{(L)}_{\mathrm{off}}\,p_0^2 + 2Q^{(L)}_{\mathrm{off}}\,p_0\right) + 2m\left(P^{(L)}_{=}\,p_0^2 + 2Q^{(L)}_{=}\,p_0\right)\\ + 2m\left(P^{(L)}_{\times}\,p_0^2 + 2Q^{(L)}_{\times}\,p_0\right),\end{aligned} \tag{85}$$

and $\mathrm{Var}(Z_L) = \mathbb{E}[Z_L^2] - \Gamma_L^2$ with $\Gamma_L = \mathbb{E}[Z_L]$ from (61). This proves part (c) of Theorem 3.3. No $\sigma_{b,l}$ and no component of $\boldsymbol{x}$ appears in (77), (83) or (85), which proves part (a).

Depth two. At $L=2$ we have $m = p_1$, $\tau = \sigma_1^2/p_0$ and $s^2 = \sigma_2^2/p_1$. Substituting (77) into (85) and using $\frac{9}{2} + \frac{8}{\pi} + \frac{1}{2} = 5 + \frac{8}{\pi}$,

$$\mathbb{E}[Z_2^2] = \tau^2 s^4\left[4p_1(p_1-1)\Big(1+\tfrac{1}{\pi}\Big)^2 p_0^2 + 2p_1\Big(5+\tfrac{8}{\pi}\Big)\,p_0(p_0+2)\right], \tag{86}$$

while $\Gamma_2^2 = 4p_1^2p_0^2\,\tau^2 s^4(1+1/\pi)^2$ by (43). Subtracting,

$$\mathrm{Var}(Z_2) = 2p_1p_0\,\tau^2 s^4\left[\Big(5+\tfrac{8}{\pi}\Big)(p_0+2) - 2\Big(1+\tfrac{1}{\pi}\Big)^2 p_0\right], \tag{87}$$

and inserting $\tau^2 s^4 = \sigma_1^4\sigma_2^4/(p_0^2p_1^2)$ gives (16), which is part (b).

C.6 At equal widths the relative variance vanishes as $L/p \to 0$

The recursion (83) determines how the variance grows with depth. When all hidden widths are equal, every prepended layer acts on the coefficients, normalized by the square of the mean, through the same 2×2 matrix, and the following corollary follows from the eigenvalues of this matrix. In this subsection the hidden widths are equal, $p_1 = \cdots = p_{L-1} = p$, while the input width p_0, the weight variances and the bias variances stay arbitrary. Write

$$\mathrm{CV}(Z_L) \;:=\; \frac{\sqrt{\mathrm{Var}(Z_L)}}{\Gamma_L}, \qquad k_0 \;:=\; \frac{(5+8/\pi)(p_0+2)}{2p_0\,(1+1/\pi)^2}, \qquad A \;:=\; \frac{p_0+2}{3p_0}. \tag{88}$$

The two constants k_0 and A are positive, A is at most 1, and $(p_0+2)/p_0 \geq 1$ gives $k_0 \geq (5+8/\pi)/(2(1+1/\pi)^2) > 2$.

Corollary C.2 (Concentration at equal widths)**.** *Let $L \geq 2$ and let the hidden widths be equal to p. The mean obeys $\Gamma_L = 8(1+1/\pi)\prod_{l=1}^{L}\sigma_l^2/2$ at every finite (L,p,p_0) by Theorem 3.1, so at $\sigma_l^2 \equiv 2$ it takes the value $8(1+1/\pi)$ at every depth and every width. The linear map that sends $(P^{(l)},Q^{(l)})/(c^{(l)})^2$ to its value at the next depth has the two eigenvalues $1+5/p$ and $1-1/p$, and the relative second moment satisfies*

$$\left(1-\frac{1}{p}\right)\left(1+\left(1+\frac{4}{p_0}\right)\frac{L-2}{p}\right) \;\leq\; \frac{\mathbb{E}[Z_L^2]}{\Gamma_L^2} \;\leq\; \left(1+\frac{k_0-1}{p}\right)\left(1+\frac{5}{p}\right)^{L-2} \;\leq\; k_0\, e^{5L/p}. \tag{89}$$

Consequently, along any sequence of depths and widths with p_0 fixed and $L/p \to 0$, one has $\mathrm{CV}(Z_L) \to 0$ and $Z_L/\Gamma_L \to 1$ in probability. If instead $L/p \to \theta$ with $\theta > 0$, then $\liminf \mathrm{CV}(Z_L)^2 \geq (1+4/p_0)\,\theta > 0$. At fixed p and increasing L, $\mathbb{E}[Z_L^2]/\Gamma_L^2 \geq (1-1/p)\,A\,(1+5/p)^{L-2}$, which increases without bound. Finally $\mathrm{CV}(Z_L)$ is a function of L, p and p_0.

Proof. Normalized recursion. For each of the three coefficient tracks of (70) put $\hat{P}^{(l)} := P^{(l)}/(c^{(l)})^2$ and $\hat{Q}^{(l)} := Q^{(l)}/(c^{(l)})^2$, with $c^{(l)}$ the column-covariance constant of (37). Prepending a layer of fan-in n', fan-out n and weight variance σ^2 multiplies c by $n\varsigma/2$, where $\varsigma = \sigma^2/n'$, by (57), and acts on (P,Q) by (83). Dividing the second of these two statements by the square of the first,

$$\hat{P}' \;=\; \frac{\varsigma^2\left(\frac{n^2+n}{4}\,P + n\,Q\right)}{\left(\frac{n\varsigma}{2}\right)^2 c^2} \;=\; \frac{4}{n^2}\left(\frac{n^2+n}{4}\,\hat{P} + n\,\hat{Q}\right) \;=\; \Big(1+\frac{1}{n}\Big)\hat{P} + \frac{4}{n}\,\hat{Q}, \tag{90}$$

$$\hat{Q}' \;=\; \frac{\varsigma^2\left(\frac{n}{2}\,P + \frac{n^2+3n}{4}\,Q\right)}{\left(\frac{n\varsigma}{2}\right)^2 c^2} \;=\; \frac{4}{n^2}\left(\frac{n}{2}\,\hat{P} + \frac{n^2+3n}{4}\,\hat{Q}\right) \;=\; \frac{2}{n}\,\hat{P} + \Big(1+\frac{3}{n}\Big)\hat{Q}. \tag{91}$$

Neither ς nor the fan-in n' appears after the division, so the normalized pair is mapped by a matrix that depends on the fan-out. Every layer prepended here has fan-out p, so that map is

$$M_p \;=\; \mathbf{I} + \frac{1}{p}\,J, \qquad J \;:=\; \begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}. \tag{92}$$

Eigenvalues. The characteristic polynomial of J is $\lambda^2 - 4\lambda - 5 = (\lambda-5)(\lambda+1)$, so the eigenvalues of J are 5 and -1. The equation $(J - 5\,\mathbf{I})v = \mathbf{0}$ reads $-4v_1 + 4v_2 = 0$, solved by $v = (1,1)^\top$, and $(J + \mathbf{I})v = \mathbf{0}$ reads $2v_1 + 4v_2 = 0$, solved by $v = (2,-1)^\top$. Hence M_p has eigenvalues $1+5/p$ and $1-1/p$ with the same two eigenvectors. Decomposing $(\hat{P},\hat{Q})^\top = a\,(1,1)^\top + b\,(2,-1)^\top$ gives $\hat{P} = a + 2b$ and $\hat{Q} = a - b$, hence $a = (\hat{P}+2\hat{Q})/3$ and $b = (\hat{P}-\hat{Q})/3$, and for every integer $k \geq 0$

$$M_p^{\,k}\begin{pmatrix}\hat{P}\\ \hat{Q}\end{pmatrix} \;=\; a\Big(1+\frac{5}{p}\Big)^k\begin{pmatrix}1\\1\end{pmatrix} \;+\; b\Big(1-\frac{1}{p}\Big)^k\begin{pmatrix}2\\-1\end{pmatrix}. \tag{93}$$

Base values. At the top two-layer block $c^{(2)} = \tau s^2(1+1/\pi)$ by (42), so $(c^{(2)})^2 = \tau^2 s^4(1+1/\pi)^2$. Dividing the three lines of (77) by this number,

$$\big(\hat{P}^{(2)}_{\mathrm{off}},\hat{Q}^{(2)}_{\mathrm{off}}\big) = (1,0), \qquad \big(\hat{P}^{(2)}_{=},\hat{Q}^{(2)}_{=}\big) = (R_=,R_=), \qquad \big(\hat{P}^{(2)}_{\times},\hat{Q}^{(2)}_{\times}\big) = (r_\times,r_\times), \tag{94}$$

with $R_= := (9/2 + 8/\pi)/(1+1/\pi)^2$ and $r_\times := (1/2)/(1+1/\pi)^2$, whose sum is $R_= + r_\times = (5+8/\pi)/(1+1/\pi)^2 = 2k_0p_0/(p_0+2)$. The coefficients (a,b) of the previous step are $(1/3,1/3)$ for the track $j \neq i$, $(R_=,0)$ for the track $j = i$, $k = k'$, and $(r_\times,0)$ for the track $j = i$, $k \neq k'$.

The three tracks at depth L. A depth-L network is reached from the top two-layer block by prepending $L-2$ layers, each of fan-out p. Write $\mu := (1+5/p)^{L-2}$ and $\nu := (1-1/p)^{L-2}$. Then (93) with $k = L-2$ and the coefficients just listed give

$$\begin{aligned}&\big(\hat{P}^{(L)}_{\mathrm{off}},\hat{Q}^{(L)}_{\mathrm{off}}\big) = \Big(\frac{\mu+2\nu}{3},\ \frac{\mu-\nu}{3}\Big),\\ &\big(\hat{P}^{(L)}_{=},\hat{Q}^{(L)}_{=}\big) = (R_=\mu,\ R_=\mu), \quad \big(\hat{P}^{(L)}_{\times},\hat{Q}^{(L)}_{\times}\big) = (r_\times\mu,\ r_\times\mu).\end{aligned} \tag{95}$$

Normalized assembly. Summing the diagonal entries of (37) over $k \in \{1,2\}$, $j \in \{1,\dots,m\}$ and $t \in \{1,\dots,p_0\}$ gives $\Gamma_L = 2mp_0\, c^{(L)}$, hence $\Gamma_L^2 = 4m^2 p_0^2 (c^{(L)})^2$ with $m = p$. Dividing (85) by this quantity and inserting (95),

$$\begin{aligned}\frac{\mathbb{E}[Z_L^2]}{\Gamma_L^2} &= \frac{4p(p-1)\big(\hat{P}_{\text{off}}^{(L)}\, p_0^2 + 2\hat{Q}_{\text{off}}^{(L)}\, p_0\big) + 2p(R_= + r_\times)\,\mu\,\big(p_0^2 + 2p_0\big)}{4p^2 p_0^2} \\ &= \frac{(p-1)\big[(\mu+2\nu)\,p_0 + 2(\mu-\nu)\big]}{3p\,p_0} \;+\; \frac{(R_= + r_\times)\,\mu\,(p_0+2)}{2p\,p_0} \\ &= \Big(1-\frac{1}{p}\Big)\frac{\mu\,(p_0+2) + \nu\,(2p_0-2)}{3p_0} \;+\; \frac{k_0}{p}\,\mu, \end{aligned} \tag{96}$$

the last line using $(\mu+2\nu)p_0+2(\mu-\nu) = \mu(p_0+2)+\nu(2p_0-2)$ and $(R_=+r_\times)(p_0+2)/(2p_0) = k_0$.

Upper bound. Since $\nu \le 1 \le \mu$ and $2p_0 - 2 \ge 0$, the numerator of the first term of (96) is at most $\mu(p_0+2) + \mu(2p_0-2) = 3p_0\,\mu$, so

$$\frac{\mathbb{E}[Z_L^2]}{\Gamma_L^2} \;\le\; \Big(1-\frac{1}{p}\Big)\mu + \frac{k_0}{p}\,\mu \;=\; \Big(1+\frac{k_0-1}{p}\Big)\Big(1+\frac{5}{p}\Big)^{L-2}, \tag{97}$$

the second inequality of (89). As $k_0 > 2$ and $p \ge 1$, one has $1 + (k_0-1)/p \le k_0$, and $1 + u \le e^u$ gives $(1+5/p)^{L-2} \le e^{5(L-2)/p} \le e^{5L/p}$; multiplying the two yields the third inequality.

Lower bound. Dropping the nonnegative term $k_0\mu/p$ in (96) and using $A = (p_0+2)/(3p_0)$ and $1 - A = (2p_0-2)/(3p_0)$,

$$\frac{\mathbb{E}[Z_L^2]}{\Gamma_L^2} \;\ge\; \Big(1-\frac{1}{p}\Big)\big[A\,\mu + (1-A)\,\nu\big]. \tag{98}$$

Bernoulli's inequality gives $\mu \ge 1 + 5(L-2)/p$ and $\nu \ge 1 - (L-2)/p$, and $1 - A \ge 0$, so

$$A\,\mu + (1-A)\,\nu \;\ge\; A + (1-A) + \big[5A - (1-A)\big]\frac{L-2}{p} \;=\; 1 + \Big(1+\frac{4}{p_0}\Big)\frac{L-2}{p}, \tag{99}$$

where $6A - 1 = 2(p_0+2)/p_0 - 1 = 1 + 4/p_0$. Inserting this into (98) gives the first inequality of (89).

The three regimes. Since $\mathbb{E}[Z_L/\Gamma_L] = 1$, one has $\mathrm{CV}(Z_L)^2 = \mathbb{E}[Z_L^2]/\Gamma_L^2 - 1$. If $L/p \to 0$ at fixed p_0, then $p \to \infty$ because $L \ge 2$, so the upper bound in (89) tends to 1 and $\mathrm{CV}(Z_L)^2 \to 0$. Chebyshev's inequality applied to Z_L/Γ_L gives, for every $\varepsilon > 0$,

$$\Pr\!\left(\left|\frac{Z_L}{\Gamma_L} - 1\right| \ge \varepsilon\right) \;\le\; \frac{\mathrm{Var}(Z_L)}{\varepsilon^2\,\Gamma_L^2} \;=\; \frac{\mathrm{CV}(Z_L)^2}{\varepsilon^2} \;\longrightarrow\; 0, \tag{100}$$

which is convergence to 1 in probability. If $L/p \to \theta$ with $\theta > 0$, the first inequality of (89) gives $\liminf \mathrm{CV}(Z_L)^2 \ge (1 + 4/p_0)\,\theta$, which is positive. At fixed p, keeping only the term containing μ in (98) leaves $(1-1/p)A(1+5/p)^{L-2}$, which grows geometrically in L at the rate $1 + 5/p$.

Independence of the variances. The base values (94) are pure numbers and the map (92) contains no weight variance, so (96) expresses $\mathbb{E}[Z_L^2]/\Gamma_L^2$, and with it $\mathrm{CV}(Z_L)^2$, through L, p and p_0. Equivalently, Γ_L^2 and the pair $(P^{(L)}, Q^{(L)})$ contain the same factor $\prod_{l=1}^{L}\sigma_l^4$, which cancels in the ratio, and no $\sigma_{b,l}$ enters either of them by Theorem 3.3. □

C.7 Comparison with direct computation and Monte Carlo

We compare the closed form with a direct computation at depth two and with simulations at larger depths.

Depth two, computed directly. At $L = 2$ the variance can be computed without the recursion: $Z_2 = \sum_{j=1}^{p_1} S_j R_j$ is a sum of p_1 i.i.d. terms (S_j from column j of $\mathbf{B}$, R_j from row j of $\mathbf{A}$), and its variance follows from the half-normal moments alone. Dividing (16) by $\sigma_1^4\sigma_2^4/p_1$ and multiplying out the factor $2/p_0$,

$$\mathrm{Var}(Z_2) \;=\; \frac{\sigma_1^4\sigma_2^4}{p_1}\left[\Big(1+\frac{2}{p_0}\Big)\Big(10+\frac{16}{\pi}\Big) - 4\Big(1+\frac{1}{\pi}\Big)^2\right], \tag{101}$$

which agrees, term by term, with the direct computation. At $p_0 = 20$, $p_1 = 128$ and $\sigma_1^2 = \sigma_2^2 = 2$ this gives $\mathrm{Var}(Z_2) = 1.206311$, against the Monte Carlo value 1.205075 over 200000 independent networks, a relative difference of 1.0×10^{-3} with a Monte Carlo standard error of 3.2×10^{-3} on the variance. Written as $\mathrm{CV}(Z_2)^2 = c(2)/p_1$, the same numbers give $c(2) = 1.388$ and $\mathrm{CV}(Z_2) = 0.1041$.

Across depths, no free parameter. Table 2 compares the closed-form coefficient of variation $\mathrm{CV}(Z_L) = \sqrt{\mathrm{Var}(Z_L)}/\Gamma_L$, obtained from (77)–(85), with Monte Carlo estimates over independent networks (He initialization $\sigma_l^2 = 2$, hidden widths 128, $p_0 = 20$; Monte Carlo standard errors range from 0.0005 at $L = 2$ to 0.0023 at $L = 16$). Simulations with per-layer biases $\sigma_{b,l}^2 \in [0, 4]$ reproduce the same values up to Monte Carlo fluctuation and without a trend in the bias variance, as Theorem 3.3 requires.

Table 2: Coefficient of variation of the generator energy: closed form vs. Monte Carlo (hidden widths 128, $p_0 = 20$, $\sigma_l^2 = 2$).

depth L	2	4	8	16
closed form	0.1041	0.1776	0.2843	0.4659
Monte Carlo	0.1044	0.1773	0.2846	0.4674

C.8 Distribution over independent networks

One sample of Z_L is one network drawn from (3) together with one input drawn uniformly on the sphere of radius $R = 1$ in $\mathbb{R}^{p_0}$, and each network contributes one sample. The spread reported in Table 3 is therefore the spread of Z_L across independent networks. The inset of Figure 4 uses a different sampling scheme, in which 200 inputs are drawn inside each of 50 networks, so the spread in that panel also reflects the variation of Z_L across inputs at fixed weights.

The two sampling schemes are related by the law of total variance, $\mathrm{Var}(Z_L) = \mathrm{Var}_W(\mathbb{E}[Z_L \mid W]) + \mathbb{E}_W[\mathrm{Var}(Z_L \mid W)]$, where W is the network. Panel (b) of Figure 7 measures the first term as a share of the total. The share is 1 at $L = 2$, where there is a single cell, and about 0.4 to 0.5 for $L \geq 3$ at both widths $p = 64$ and $p = 128$, so about 0.5 to 0.6 of the variance of Z_L is the variance across inputs and cells inside one network. Since $\mathrm{Var}(\bar{Z}_L) \leq \mathrm{Var}(Z_L)$ for the per-network input average $\bar{Z}_L(W) := \mathbb{E}_{\boldsymbol{x}}[Z_L \mid W]$, Corollary C.2 also gives $\bar{Z}_L/\Gamma_L \to 1$ in probability when $L/p \to 0$.

Table 3 reports, for each depth and hidden width, the closed-form coefficient of variation from (77)–(85), its Monte Carlo counterpart, the ratio of the mean to the median, and five quantiles of Z_L/Γ_L, over 20000 independent networks per configuration at $p_0 = 20$, $\sigma_l^2 = 2$ and zero bias. The quantiles are normalized by the closed form $\Gamma_L = 8(1 + 1/\pi)$, and the empirical mean stays within 0.6% of that value in every configuration. The median decreases with depth while the mean does not, and at a given depth the upper quantiles depart from 1 by more than the lower ones. Panel (a) of Figure 7 plots these quantiles against depth, with $L = 3$ added.

A separate sweep gives an independent Gaussian bias of common variance σ_b^2 to every layer below the top two; these lower layers fix the cell. The sweep uses hidden width 128 and again 20000 networks per configuration. Over $\sigma_b^2 \in \{0, 0.5, 1, 2, 4\}$ the ratio of the empirical mean to Γ_L stays between 0.997 and 1.002, and the coefficient of variation shows no trend in σ_b^2. It is 0.177 ± 0.001 at $L = 4$ and lies between 0.460 and 0.467 at $L = 16$, against the closed-form values 0.1776 and 0.4659, in which no $\sigma_{b,l}^2$ appears.

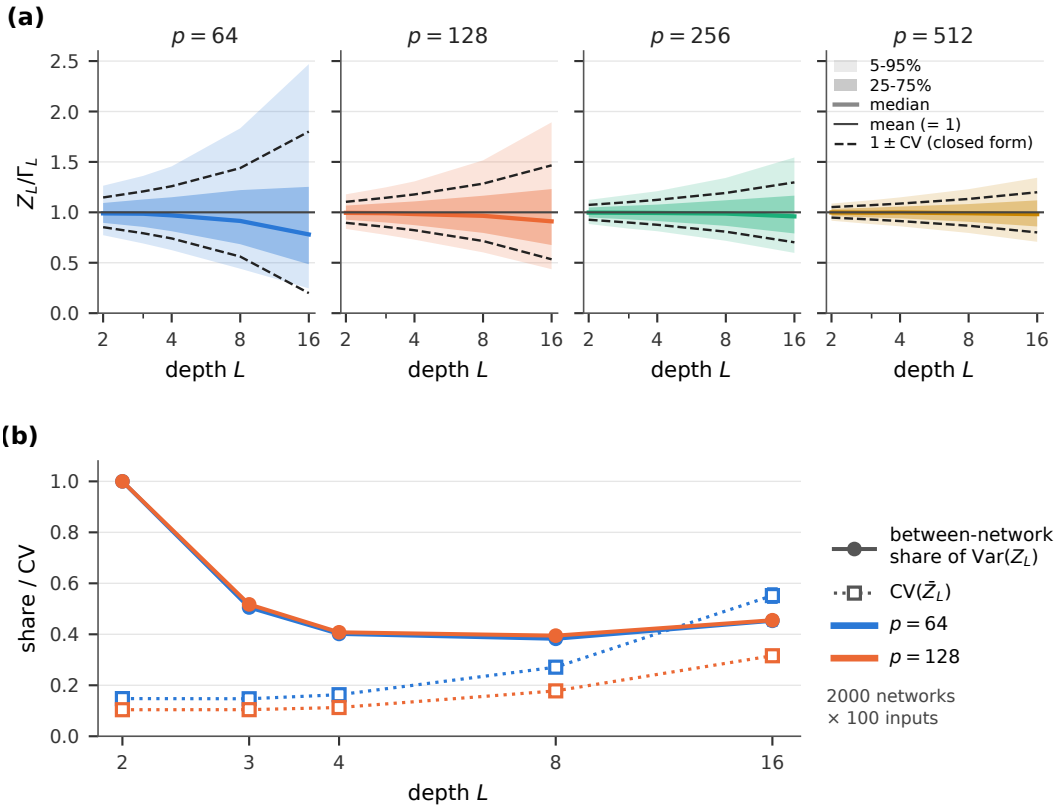


Figure 7: Distribution of Z_L/Γ_L over independent networks at $p_0 = 20$, $\sigma_l^2 = 2$, zero bias. (a) Quantiles against depth $L \in \{2, 3, 4, 8, 16\}$ (logarithmic axis), one sub-panel per hidden width $p \in \{64, 128, 256, 512\}$, over 2×10^4 networks with one input each: the shaded bands span the 5 to 95 and the 25 to 75 percent quantiles, the line inside them is the median, the solid line is the mean 1, and the dashed lines are $1 \pm \mathrm{CV}(Z_L)$ from the closed form (85). (b) Between-network share $\mathrm{Var}_W(\mathbb{E}[Z_L \mid W])/\mathrm{Var}(Z_L)$ of the variance of Z_L (filled circles) against L, from 2000 networks times 100 inputs at $p \in \{64, 128\}$; the remainder of the variance is across inputs and cells inside one network. The coefficient of variation of the per-network input average $\bar{Z}_L(W) := \mathbb{E}_{\boldsymbol{x}}[Z_L \mid W]$ is drawn on the same axis (open squares). The between-network component is the unbiased one-way analysis-of-variance estimator, and the error bars in (b) are bootstrap standard errors over networks, which are smaller than the markers.

Table 3: Distribution of Z_L/Γ_L over independent networks: 20000 networks per configuration, one input per network, $p_0 = 20$, $\sigma_l^2 = 2$, zero bias, hidden widths all equal to p. The exact coefficient of variation is (77)–(85); the standard error of the Monte Carlo one runs from 0.0003 to 0.0040 across the configurations.

L	p	CV exact	CV MC	mean/median	$q_{0.05}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.95}$
2	64	0.1473	0.1493	1.011	0.773	0.896	0.990	1.093	1.264
2	128	0.1041	0.1044	1.005	0.835	0.927	0.995	1.067	1.178
2	256	0.0736	0.0730	1.003	0.884	0.949	0.997	1.048	1.123
2	512	0.0521	0.0518	1.001	0.917	0.964	0.999	1.034	1.087
4	64	0.2589	0.2579	1.027	0.625	0.813	0.970	1.150	1.456
4	128	0.1776	0.1773	1.014	0.729	0.873	0.983	1.109	1.307
4	256	0.1236	0.1219	1.006	0.813	0.914	0.994	1.077	1.210
4	512	0.0867	0.0869	1.002	0.861	0.940	0.998	1.057	1.148
8	64	0.4389	0.4449	1.090	0.441	0.684	0.915	1.220	1.833
8	128	0.2843	0.2846	1.036	0.603	0.796	0.966	1.165	1.515
8	256	0.1920	0.1914	1.015	0.717	0.866	0.988	1.120	1.342
8	512	0.1326	0.1329	1.008	0.795	0.906	0.991	1.082	1.230
16	64	0.8011	0.7913	1.273	0.246	0.485	0.781	1.253	2.470
16	128	0.4659	0.4674	1.104	0.436	0.675	0.911	1.231	1.893
16	256	0.2971	0.2952	1.042	0.597	0.790	0.960	1.166	1.544
16	512	0.1991	0.1958	1.017	0.707	0.860	0.982	1.119	1.343

D Crossing density

This appendix computes the expected number of times a great circle on the input sphere crosses the decision boundary. The expected number of crossings per unit length is the same for every great circle, and integrating the number of crossings over all great circles gives the expected volume of the decision boundary on the sphere times a constant that depends only on p_0 (§D.6).

Throughout, the network is given by (1) and the parameters are drawn as in (3). As in (2), the output layer carries a bias $\boldsymbol{b}^{(L)}$ of variance $\sigma_{b,L}^2 \geq 0$, so that $f_1 = \sum_j B_{1j} h_j^{(L-1)} + b_1^{(L)}$ and $f_2 = \sum_j B_{2j} h_j^{(L-1)} + b_2^{(L)}$, where $\mathbf{B} = \mathbf{W}^{(L)}$. We write $g := f_1 - f_2$, so the decision boundary is $\{g = 0\}$. We also write

$$\Delta_j \;:=\; B_{1j} - B_{2j} \quad (j = 1, \ldots, p_{L-1}), \qquad \Delta_0 \;:=\; b_1^{(L)} - b_2^{(L)}, \tag{102}$$

so that $g = \Delta_0 + \sum_{j=1}^{p_{L-1}} \Delta_j h_j^{(L-1)}$. Let

$$\mathcal{F} \;:=\; \big(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L-1)}, \boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(L-1)}\big) \tag{103}$$

collect all parameters below the output layer. Given $\mathcal{F}$, each $h_j^{(L-1)}$ is a deterministic function of $\boldsymbol{x}$, the only random variables are $\mathbf{B}$ and $\boldsymbol{b}^{(L)}$, and they enter g only through $\Delta_0, \ldots, \Delta_{p_{L-1}}$.

Hence, given $\mathcal{F}$, g is a linear combination of $h_1^{(L-1)}, \ldots, h_{p_{L-1}}^{(L-1)}$ with the random coefficients Δ_j, plus the random constant Δ_0. As we move along a great circle, the value of each hidden unit and its rate of change form a point in the plane, and this point traces a small ellipse; with a bias, the ellipse is shifted along the value direction. The output bias adds the variance $2\sigma_{b,L}^2$ of Δ_0 to the variance of g, and in the plane it corresponds to the direction $(1, 0)$ with weight $2\sigma_{b,L}^2$. Given $\mathcal{F}$, the expected number of times the great circle crosses the decision boundary is therefore determined by these p_{L-1} small ellipses and the variance of the output bias (§D.2). The more spread out the directions of these points, the larger the area they span, the less the value and the rate of change of g move together, and the more crossings there are. If all the points lie on one line, the rate of change of g is a fixed multiple of its value, and g does not cross zero.

These points determine a 2×2 matrix $\mathbf{S}(s)$, the covariance matrix, given $\mathcal{F}$, of the value and the rate of change of g at the point s of the great circle (§D.1); it is therefore a function of $\mathcal{F}$. The number N_ℓ of crossings on an arc of length ℓ satisfies

$$\mathbb{E}[N_\ell \mid \mathcal{F}] = \frac{1}{\pi} \int_0^\ell r\big(\mathbf{S}(s)\big)\, ds, \qquad r(\mathbf{S}) = \frac{\sqrt{S_{11}S_{22} - S_{12}^2}}{S_{11}} \tag{104}$$

(§D.2). Taking the expectation over $\mathcal{F}$ gives the crossing density, an expectation over all parameters, $\rho_{\text{exact}} = \frac{1}{\pi}\,\mathbb{E}_{\mathcal{F}}[r(\mathbf{S})]$. The estimate (19) takes the expectation of $\mathbf{S}$ first and then evaluates r, that is, $\rho_L = \frac{1}{\pi} r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$. The matrix $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$ is diagonal, with diagonal entries $2q^{(L)}$ and a number proportional to Γ_L, so $\rho_L = \kappa(p_0)\sqrt{\Gamma_L / q^{(L)}}$ (§D.3). The relative difference $\delta = \rho_{\text{exact}}/\rho_L - 1$ comes from the fluctuation of $\mathbf{S}$ across different $\mathcal{F}$ (§D.4). Finally, when all parameters $(\mathcal{F}, \mathbf{B}, \boldsymbol{b}^{(L)})$ are fixed, the generator energy Z_L of one specific network satisfies $\mathbb{E}[Z_L] = \Gamma_L$ and gives the estimate $\hat{\rho}_L = \kappa(p_0)\sqrt{Z_L / q^{(L)}}$ for this network. When every hidden width equals p and $L/p \to 0$, $\hat{\rho}_L / \rho_L \to 1$ in probability (§D.5). Integrating the number of crossings over all great circles, ρ_L gives the expected volume of the decision boundary on the sphere (§D.6).

D.1 Circles and the conditional covariance

The inputs lie on the sphere $S_R := \{\boldsymbol{x} \in \mathbb{R}^{p_0} : \|\boldsymbol{x}\|_2 = R\}$. For a pair of orthonormal vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$, the great circle in the plane they span is parameterized by arc length as

$$\gamma(s) \;:=\; R\big(\boldsymbol{e}_1 \cos(s/R) + \boldsymbol{e}_2 \sin(s/R)\big), \qquad s \in \mathbb{R}, \tag{105}$$

so that $\|\dot\gamma\|_2 = 1$ and $\langle\gamma,\dot\gamma\rangle = 0$. A crossing is an isolated zero of the function $s \mapsto g(\gamma(s))$ at which the function changes sign. For an arc of length ℓ, we write N_ℓ for its number of crossings.

Lemma D.1. *For every orthogonal matrix $U \in O(p_0)$, the hidden features $\boldsymbol{h}^{(1)},\dots,\boldsymbol{h}^{(L-1)}$ and g, as random functions of $\boldsymbol{x}$, have the same joint distribution at $U\boldsymbol{x}$ as at $\boldsymbol{x}$.*

Proof. The input enters the network only through $\mathbf{W}^{(1)}\boldsymbol{x}$, so each feature at $U\boldsymbol{x}$ equals the same feature computed at $\boldsymbol{x}$ with $\mathbf{W}^{(1)}U$ in place of $\mathbf{W}^{(1)}$. The rows of $\mathbf{W}^{(1)}$ are i.i.d. $\mathcal{N}(0,(\sigma_1^2/p_0)\,\mathbf{I})$, this distribution is invariant under right multiplication by U, and the other parameters are independent of $\mathbf{W}^{(1)}$. □

Moving along the circle by an arc is a rotation in the plane spanned by $\boldsymbol{e}_1,\boldsymbol{e}_2$, and every great circle is the image of γ under some rotation. By Lemma D.1, the distribution of the crossings on an arc depends only on its length, and the expectation of any function of the network and its derivatives along the circle at $\gamma(s)$ does not depend on s.

We write

$$h_j(s) \;:=\; h_j^{(L-1)}(\gamma(s)), \qquad j = 1,\dots,p_{L-1}, \tag{106}$$

for the output of unit j of unit set $L-1$ at the point s of the great circle, and $h_j'(s)$ for its derivative with respect to s. The derivative h_j' exists except at the finitely many s where some unit changes its activation state. The small ellipse described at the beginning is the trajectory traced by the point $(h_j(s),h_j'(s))$ in the plane as s varies. For a cell $\mathcal{C}$ of Section 3, let $\boldsymbol{u}_j^\top$ be the j-th row of $\mathbf{W}^{(L-1)}\Phi_{\mathcal{C}}^{(L-2)}$, and let ν_j be the constant term of the pre-activation of unit j inside $\mathcal{C}$. Inside $\mathcal{C}$, the pre-activation of unit j is then the affine function $\langle\boldsymbol{u}_j,\boldsymbol{x}\rangle+\nu_j$ of $\boldsymbol{x}$, and every generator of the zonotopes generated by the rows of (8) is a nonnegative multiple of some $\boldsymbol{u}_j$. When unit j is active in this cell, $(h_j,h_j') = (\langle\boldsymbol{u}_j,\gamma(s)\rangle+\nu_j,\langle\boldsymbol{u}_j,\dot\gamma(s)\rangle)$. Substituting (105) and its derivative $\dot\gamma(s) = -\boldsymbol{e}_1\sin(s/R)+\boldsymbol{e}_2\cos(s/R)$,

$$\begin{aligned}\langle\boldsymbol{u}_j,\gamma(s)\rangle &= R\big(\langle\boldsymbol{u}_j,\boldsymbol{e}_1\rangle\cos(s/R)+\langle\boldsymbol{u}_j,\boldsymbol{e}_2\rangle\sin(s/R)\big),\\ \langle\boldsymbol{u}_j,\dot\gamma(s)\rangle &= -\langle\boldsymbol{u}_j,\boldsymbol{e}_1\rangle\sin(s/R)+\langle\boldsymbol{u}_j,\boldsymbol{e}_2\rangle\cos(s/R).\end{aligned} \tag{107}$$

Dividing the first line by R, squaring both lines and adding them, the cross terms cancel, and $\cos^2+\sin^2 = 1$ gives

$$\frac{\langle\boldsymbol{u}_j,\gamma(s)\rangle^2}{R^2}+\langle\boldsymbol{u}_j,\dot\gamma(s)\rangle^2 = \langle\boldsymbol{u}_j,\boldsymbol{e}_1\rangle^2+\langle\boldsymbol{u}_j,\boldsymbol{e}_2\rangle^2. \tag{108}$$

So the point lies on the ellipse with center $(\nu_j,0)$ and semi-axes $R\,(\langle\boldsymbol{u}_j,\boldsymbol{e}_1\rangle^2+\langle\boldsymbol{u}_j,\boldsymbol{e}_2\rangle^2)^{1/2}$ and $(\langle\boldsymbol{u}_j,\boldsymbol{e}_1\rangle^2+\langle\boldsymbol{u}_j,\boldsymbol{e}_2\rangle^2)^{1/2}$. When unit j is inactive, $(h_j,h_j') = (0,0)$. Hence, while $\gamma(s)$ stays in $\mathcal{C}$, the point (h_j,h_j') moves along the part of this ellipse with positive first coordinate, or rests at the origin. The whole trajectory is the union of the elliptic arcs belonging to the cells that γ passes through.

We now compute, component by component, the covariance of the value and the rate of change of g given $\mathcal{F}$. Write $g'(s) := \frac{d}{ds}\,g(\gamma(s))$. By (102),

$$g(\gamma(s)) = \Delta_0+\Delta_1h_1(s)+\Delta_2h_2(s)+\cdots+\Delta_{p_{L-1}}h_{p_{L-1}}(s). \tag{109}$$

None of $\Delta_0,\dots,\Delta_{p_{L-1}}$ depends on s, so differentiating with respect to s gives

$$g'(s) = \Delta_0\cdot 0+\Delta_1h_1'(s)+\Delta_2h_2'(s)+\cdots+\Delta_{p_{L-1}}h_{p_{L-1}}'(s). \tag{110}$$

Given $\mathcal{F}$, $h_j(s)$ and $h_j'(s)$ are deterministic numbers, and only $\Delta_0,\dots,\Delta_{p_{L-1}}$ are random. By (3), $B_{1j},B_{2j},b_1^{(L)},b_2^{(L)}$ are mutually independent, all independent of $\mathcal{F}$, all of mean 0, and $\mathrm{Var}(B_{1j}) = \mathrm{Var}(B_{2j}) = \sigma_L^2/p_{L-1}$, $\mathrm{Var}(b_1^{(L)}) = \mathrm{Var}(b_2^{(L)}) = \sigma_{b,L}^2$. Therefore $\Delta_0,\dots,\Delta_{p_{L-1}}$ are mutually independent, independent of $\mathcal{F}$, of mean 0, $\mathbb{E}[\Delta_i\Delta_j] = 0$ for $0\le i,j\le p_{L-1}$ with $i\ne j$, and

$$\begin{aligned}\mathbb{E}[\Delta_0^2] &= \mathrm{Var}(b_1^{(L)})+\mathrm{Var}(b_2^{(L)}) = 2\sigma_{b,L}^2,\\ \mathbb{E}[\Delta_j^2] &= \mathrm{Var}(B_{1j})+\mathrm{Var}(B_{2j}) = \frac{2\sigma_L^2}{p_{L-1}} \quad (j\ge 1).\end{aligned} \tag{111}$$

Hence the conditional means of $g(\gamma(s))$ and $g'(s)$ are both 0. In what follows, the sums over i and j run from 1 to p_{L-1}. The variance expands term by term as

$$\begin{aligned}\operatorname{Var}(g(\gamma(s)) \mid \mathcal{F}) &= \mathbb{E}\left[\Big(\Delta_0 + \sum_i \Delta_i h_i\Big)\Big(\Delta_0 + \sum_j \Delta_j h_j\Big)\right] \\ &= \mathbb{E}[\Delta_0^2] + \sum_i \mathbb{E}[\Delta_0\Delta_i]\, h_i + \sum_j \mathbb{E}[\Delta_0\Delta_j]\, h_j + \sum_i \sum_j \mathbb{E}[\Delta_i\Delta_j]\, h_i h_j.\end{aligned} \tag{112}$$

Substituting $\mathbb{E}[\Delta_0\Delta_i] = 0$, the two middle terms vanish. Splitting the double sum into the parts $i = j$ and $i \neq j$ and substituting $\mathbb{E}[\Delta_i\Delta_j] = 0$ for $i \neq j$, the double sum equals $\sum_i \mathbb{E}[\Delta_i^2]\, h_i^2$. Substituting the two variances,

$$\operatorname{Var}(g(\gamma(s)) \mid \mathcal{F}) = \mathbb{E}[\Delta_0^2] + \sum_i \mathbb{E}[\Delta_i^2]\, h_i^2 = 2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}} \sum_i h_i^2. \tag{113}$$

The covariance is computed in the same way. The coefficient of Δ_0 in $g'(s)$ is 0. Each line below first expands the product, then substitutes $\mathbb{E}[\Delta_0\Delta_j] = 0$ and $\mathbb{E}[\Delta_i\Delta_j] = 0$ for $i \neq j$, and finally substitutes $\mathbb{E}[\Delta_i^2]$,

$$\begin{aligned}\operatorname{Cov}(g(\gamma(s)), g'(s) \mid \mathcal{F}) &= \mathbb{E}\left[\Big(\Delta_0 + \sum_i \Delta_i h_i\Big)\Big(\sum_j \Delta_j h_j'\Big)\right] = \sum_j \mathbb{E}[\Delta_0\Delta_j]\, h_j' \\ &\quad + \sum_i \sum_j \mathbb{E}[\Delta_i\Delta_j]\, h_i h_j' \\ &= 0 + \sum_i \mathbb{E}[\Delta_i^2]\, h_i h_i' = \frac{2\sigma_L^2}{p_{L-1}} \sum_i h_i h_i', \\ \operatorname{Var}(g'(s) \mid \mathcal{F}) &= \mathbb{E}\left[\Big(\sum_i \Delta_i h_i'\Big)\Big(\sum_j \Delta_j h_j'\Big)\right] = \sum_i \sum_j \mathbb{E}[\Delta_i\Delta_j]\, h_i' h_j' \\ &= \sum_i \mathbb{E}[\Delta_i^2]\, (h_i')^2 = \frac{2\sigma_L^2}{p_{L-1}} \sum_i (h_i')^2.\end{aligned} \tag{114}$$

Arranging the three entries in a matrix, renaming the summation index to j and splitting by j, we obtain

$$\mathbf{S}(s) \;:=\; \operatorname{Cov}\big((g(\gamma(s)), g'(s))\,\big|\,\mathcal{F}\big) \;=\; 2\sigma_{b,L}^2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{2\sigma_L^2}{p_{L-1}} \sum_{j=1}^{p_{L-1}} \begin{bmatrix} h_j^2 & h_j h_j' \\ h_j h_j' & (h_j')^2 \end{bmatrix}, \tag{115}$$

where the j-th matrix equals $(h_j, h_j')^\top (h_j, h_j')$. The matrix $\mathbf{S}(s)$ is a function of $\mathcal{F}$, with the expectation over $(\mathbf{B}, \boldsymbol{b}^{(L)})$ already taken. When $\sigma_{b,L} = 0$, the first term is zero.

For a 2×2 positive semidefinite matrix $\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}$, we write

$$r(\mathbf{S}) \;:=\; \frac{\sqrt{S_{11}S_{22} - S_{12}^2}}{S_{11}} \quad \text{if } S_{11} > 0, \qquad r(\mathbf{S}) := 0 \quad \text{if } S_{11} = 0. \tag{116}$$

For $t > 0$, the definition gives $r(t\mathbf{S}) = \sqrt{t^2 \det \mathbf{S}}/(tS_{11}) = r(\mathbf{S})$, that is, multiplying $\mathbf{S}$ by a positive number leaves this ratio unchanged.

D.2 The crossing density is an area

Given $\mathcal{F}$, $g(\gamma(s))$ and $g'(s)$ are linear combinations of the independent centered Gaussian variables $\Delta_0, \ldots, \Delta_{p_{L-1}}$, so they form a pair of centered Gaussian random variables with covariance matrix $\mathbf{S}(s)$ from (115), and the expected number of crossings is determined by $\mathbf{S}$. The crossing density is expressed through the function $r(\mathbf{S})$ of $\mathbf{S}$. The determinant $\det \mathbf{S}$ has a geometric reading, and $r(\mathbf{S})$ has a probabilistic reading.

Lemma D.2. *Let* $\mathbf{S} = \mathbf{S}(s)$ *be given by* (115)*, with* $S_{11} > 0$*. The sums over* i *and* j *below run over* $1, \dots, p_{L-1}$*.*

(a) $\det \mathbf{S}$ *is a weighted sum of two parts: the sum of the squared areas of the parallelograms spanned by the pairs* (h_i, h'_i)*,* (h_j, h'_j)*, and the sum of the squared areas of the parallelograms spanned by the bias direction* $(1, 0)$ *and each* (h_j, h'_j)*,*

$$\det \mathbf{S} = 2\sigma_{b,L}^2 \, \frac{2\sigma_L^2}{p_{L-1}} \sum_j \det \begin{bmatrix} 1 & h_j \\ 0 & h'_j \end{bmatrix}^2 + \Big(\frac{2\sigma_L^2}{p_{L-1}}\Big)^2 \sum_{i<j} \det \begin{bmatrix} h_i & h_j \\ h'_i & h'_j \end{bmatrix}^2 . \tag{117}$$

(b) Given $\mathcal{F}$*, the density of* $g(\gamma(s))$ *at* 0 *times* $\mathbb{E}\big[\, |g'(s)| \,\big|\, g(\gamma(s)) = 0, \mathcal{F}\big]$ *equals* $\frac{1}{\pi}\, r(\mathbf{S})$*.*

Proof. (a) Substituting the entries of (115) and expanding the products,

$$S_{11} S_{22} = \Big(2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}} \sum_i h_i^2\Big) \frac{2\sigma_L^2}{p_{L-1}} \sum_j (h'_j)^2 \tag{118}$$

$$= 2\sigma_{b,L}^2 \, \frac{2\sigma_L^2}{p_{L-1}} \sum_j (h'_j)^2 + \Big(\frac{2\sigma_L^2}{p_{L-1}}\Big)^2 \sum_i \sum_j h_i^2 (h'_j)^2, \tag{119}$$

$$S_{12}^2 = \Big(\frac{2\sigma_L^2}{p_{L-1}} \sum_i h_i h'_i\Big)\Big(\frac{2\sigma_L^2}{p_{L-1}} \sum_j h_j h'_j\Big) = \Big(\frac{2\sigma_L^2}{p_{L-1}}\Big)^2 \sum_i \sum_j h_i h'_i h_j h'_j. \tag{120}$$

Subtracting the two gives

$$\det \mathbf{S} = S_{11} S_{22} - S_{12}^2 \tag{121}$$

$$= 2\sigma_{b,L}^2 \, \frac{2\sigma_L^2}{p_{L-1}} \sum_j (h'_j)^2 + \Big(\frac{2\sigma_L^2}{p_{L-1}}\Big)^2 \Big(\sum_i \sum_j h_i^2 (h'_j)^2 - \sum_i \sum_j h_i h'_i h_j h'_j \Big). \tag{122}$$

Splitting each of the two double sums in the bracket into the three parts $i = j$, $i < j$ and $i > j$,

$$\sum_i \sum_j h_i^2 (h'_j)^2 - \sum_i \sum_j h_i h'_i h_j h'_j \tag{123}$$

$$= \Big(\sum_i h_i^2 (h'_i)^2 + \sum_{i<j} h_i^2 (h'_j)^2 + \sum_{i>j} h_i^2 (h'_j)^2 \Big) \tag{124}$$

$$\quad - \Big(\sum_i h_i^2 (h'_i)^2 + \sum_{i<j} h_i h'_i h_j h'_j + \sum_{i>j} h_i h'_i h_j h'_j \Big) \tag{125}$$

$$= \sum_{i<j} h_i^2 (h'_j)^2 + \sum_{i<j} h_j^2 (h'_i)^2 - \sum_{i<j} h_i h'_i h_j h'_j - \sum_{i<j} h_j h'_j h_i h'_i \tag{126}$$

$$= \sum_{i<j} \big(h_i^2 (h'_j)^2 + h_j^2 (h'_i)^2 - 2 h_i h'_j h_j h'_i \big) = \sum_{i<j} (h_i h'_j - h_j h'_i)^2, \tag{127}$$

where the second equality cancels the common term $\sum_i h_i^2 (h'_i)^2$ on both sides and renames the indices i and j in the sums over $i > j$. Finally we substitute $\det \left[\begin{smallmatrix} 1 & h_j \\ 0 & h'_j \end{smallmatrix}\right] = h'_j$ and $\det \left[\begin{smallmatrix} h_i & h_j \\ h'_i & h'_j \end{smallmatrix}\right] = h_i h'_j - h_j h'_i$.

(b) Given $\mathcal{F}$, $(g(\gamma(s)), g'(s))$ is a centered Gaussian vector with covariance $\mathbf{S}$. So the density of $g(\gamma(s))$ at 0 is $(2\pi S_{11})^{-1/2}$, and given $g(\gamma(s)) = 0$, $g'(s)$ is distributed as $\mathcal{N}(0, S_{22} - S_{12}^2/S_{11}) = \mathcal{N}(0, \det \mathbf{S}/S_{11})$, while the expected absolute value of a $\mathcal{N}(0, \det \mathbf{S}/S_{11})$ variable is $\big(2 \det \mathbf{S}/(\pi S_{11})\big)^{1/2}$. Therefore

$$(2\pi S_{11})^{-1/2} \Big(\frac{2 \det \mathbf{S}}{\pi S_{11}}\Big)^{1/2} = \frac{\sqrt{\det \mathbf{S}}}{\pi \, S_{11}} = \frac{1}{\pi}\, r(\mathbf{S}). \tag{128}$$

□

Absorbing the variances of $\Delta_0, \dots, \Delta_{p_{L-1}}$ into the coefficients, the value of g along the circle is the projection of a curve in $\mathbb{R}^{p_{L-1}+1}$ onto an isotropic random direction. The 0-th component of this curve comes from the output bias and is constant, and the other components are proportional to $h_1, \dots, h_{p_{L-1}}$. Each time the curve crosses the hyperplane orthogonal to this direction, g crosses zero once. So the faster the direction of the curve turns, the more crossings there are. The quotient rule computation in the proof gives the speed at which the direction turns. The direction turns because the points (h_j, h'_j) are not collinear with each other, or not collinear with the bias direction $(1, 0)$, and part (a) writes $\det \mathbf{S}$ as a weighted sum of the squared areas of these parallelograms. Part (b) shows that this quantity combines the two ingredients of the Rice formula, how far the process is from zero and how fast it changes.

Proposition D.3 (Crossing density)**.** *For every arc of every great circle and every $\ell > 0$, $\mathbb{E}[N_\ell] = \ell\, \rho_{\text{exact}}$, where the expectation is over all parameters, and*

$$\begin{aligned}\rho_{\text{exact}} \;&:=\; \frac{1}{\pi}\, \mathbb{E}_{\mathcal{F}}\left[r(\mathbf{S})\right] \;=\; \frac{1}{\pi}\, \mathbb{E}_{\mathcal{F}}\left[\frac{\sqrt{S_{11}S_{22} - S_{12}^2}}{S_{11}}\,;\; S_{11} > 0\right] \\ &=\; \frac{1}{\pi}\, \mathbb{E}_{\mathcal{F}}\left[\frac{\left(2\sigma_{b,L}^2\, \frac{2\sigma_L^2}{p_{L-1}} \sum_j (h'_j)^2 + \left(\frac{2\sigma_L^2}{p_{L-1}}\right)^2 \sum_{i<j} (h_i h'_j - h_j h'_i)^2\right)^{1/2}}{2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}} \sum_j h_j^2}\,;\; S_{11} > 0\right],\end{aligned} \tag{129}$$

where the entries of $\mathbf{S}$ are given by (115)*, the summation indices run over $1, \dots, p_{L-1}$, and h_j and h'_j are evaluated at any fixed s.*

Proof. Condition on $\mathcal{F}$. In this proof, index the components of $\boldsymbol{w}(s), \boldsymbol{\xi} \in \mathbb{R}^{p_{L-1}+1}$ from 0 to p_{L-1}, and set

$$w_0(s) := \sqrt{2}\, \sigma_{b,L}, \qquad w_j(s) := \sqrt{2\sigma_L^2/p_{L-1}}\; h_j(s), \tag{130}$$

$$\xi_0 := \frac{\Delta_0}{\sqrt{2}\, \sigma_{b,L}}, \qquad \xi_j := \frac{\Delta_j}{\sqrt{2\sigma_L^2/p_{L-1}}}, \tag{131}$$

where $j = 1, \dots, p_{L-1}$; when $\sigma_{b,L} = 0$, we drop the 0-th component of both. By the variances of Δ_j in §D.1, the components of $\boldsymbol{\xi}$ are i.i.d. $\mathcal{N}(0, 1)$ and independent of $\mathcal{F}$. Multiplying component by component gives

$$\langle \boldsymbol{\xi}, \boldsymbol{w}(s) \rangle = \Delta_0 + \sum_j \Delta_j h_j(s) = g(\gamma(s)), \tag{132}$$

and at every s where $\boldsymbol{w}$ is differentiable, $w'_0 = 0$ and (115) give, component by component,

$$\|\boldsymbol{w}\|_2^2 = 2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}} \sum_j h_j^2 = S_{11}, \tag{133}$$

$$\langle \boldsymbol{w}, \boldsymbol{w}' \rangle = \frac{2\sigma_L^2}{p_{L-1}} \sum_j h_j h'_j = S_{12}, \tag{134}$$

$$\|\boldsymbol{w}'\|_2^2 = \frac{2\sigma_L^2}{p_{L-1}} \sum_j (h'_j)^2 = S_{22}. \tag{135}$$

Since h'_j exists except at finitely many s, $\boldsymbol{w}$ is differentiable except at these finitely many s. At an s where $\boldsymbol{w}$ is differentiable and $S_{11}(s) > 0$, the first identity gives $\boldsymbol{w} \neq 0$. Differentiating $\|\boldsymbol{w}\|_2 = \langle \boldsymbol{w}, \boldsymbol{w} \rangle^{1/2}$ gives $\frac{d}{ds} \|\boldsymbol{w}\|_2 = \langle \boldsymbol{w}, \boldsymbol{w}' \rangle / \|\boldsymbol{w}\|_2$, so the quotient rule gives

$$\frac{d}{ds}\, \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} = \frac{\boldsymbol{w}'}{\|\boldsymbol{w}\|_2} - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}\, \frac{\langle \boldsymbol{w}, \boldsymbol{w}' \rangle}{\|\boldsymbol{w}\|_2} \tag{136}$$

$$= \frac{\boldsymbol{w}'}{\|\boldsymbol{w}\|_2} - \frac{\langle \boldsymbol{w}, \boldsymbol{w}' \rangle}{\|\boldsymbol{w}\|_2^3}\, \boldsymbol{w}. \tag{137}$$

Expanding its squared norm and substituting the three identities,

$$\Big\| \frac{d}{ds} \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \Big\|_2^2 = \frac{\|\boldsymbol{w}'\|_2^2}{\|\boldsymbol{w}\|_2^2} - 2\,\frac{\langle \boldsymbol{w}, \boldsymbol{w}' \rangle^2}{\|\boldsymbol{w}\|_2^4} + \frac{\langle \boldsymbol{w}, \boldsymbol{w}' \rangle^2\ \|\boldsymbol{w}\|_2^2}{\|\boldsymbol{w}\|_2^6} \tag{138}$$

$$= \frac{\|\boldsymbol{w}\|_2^2\, \|\boldsymbol{w}'\|_2^2 - \langle \boldsymbol{w}, \boldsymbol{w}' \rangle^2}{\|\boldsymbol{w}\|_2^4} = \frac{S_{11} S_{22} - S_{12}^2}{S_{11}^2}. \tag{139}$$

Taking the square root and comparing with (116) gives

$$\Big\| \frac{d}{ds} \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \Big\|_2 = r\big(\mathbf{S}(s)\big), \tag{140}$$

that is, $r(\mathbf{S})$ is the speed at which the direction of $\boldsymbol{w}$ moves on the unit sphere.

On an interval where $\boldsymbol{w} \neq 0$, the crossings of $g(\gamma(s))$ are the values of s at which the curve $\boldsymbol{w}/\|\boldsymbol{w}\|_2$ on the unit sphere crosses the great hypersphere $\boldsymbol{\xi}^\perp$. The direction of $\boldsymbol{\xi}$ is uniformly distributed, and the spherical Crofton formula [22] states that the expected number of intersections of a curve on the unit sphere with a uniformly random great hypersphere equals the length of the curve divided by π. The curve is piecewise analytic, so for almost every $\boldsymbol{\xi}$ it meets $\boldsymbol{\xi}^\perp$ transversally and the intersection points avoid its finitely many corners, and every intersection point is then a crossing. On an interval where $\boldsymbol{w} = 0$, the function $s \mapsto g(\gamma(s))$ is identically zero and has no crossings. An isolated zero of $\boldsymbol{w}$ can occur only when $\sigma_{b,L} = 0$, and it requires two units to change their activation state at the same s, or some pre-activation to reach zero without changing sign, so it has probability zero. The Crofton formula gives the first equality below. Identity (140) holds except at finitely many s, and these s do not affect the integral, so the second equality holds,

$$\mathbb{E}\left[N_\ell \,\middle|\, \mathcal{F}\right] \;=\; \frac{1}{\pi} \int_0^\ell \Big\| \frac{d}{ds} \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \Big\|_2 \mathbf{1}\{S_{11}(s) > 0\}\, ds \;=\; \frac{1}{\pi} \int_0^\ell r\big(\mathbf{S}(s)\big)\, ds. \tag{141}$$

We take the expectation over $\mathcal{F}$ and exchange it with the integral, which is permitted because the integrand is nonnegative. By Lemma D.1, the expectation of the integrand depends neither on s nor on the circle, which gives $\mathbb{E}[N_\ell] = \ell\, \rho_{\text{exact}}$. The second equality of (129) is (116). For the third equality, Lemma D.2(a) gives

$$S_{11} S_{22} - S_{12}^2 = 2\sigma_{b,L}^2\, \frac{2\sigma_L^2}{p_{L-1}} \sum_j (h_j')^2 + \Big(\frac{2\sigma_L^2}{p_{L-1}}\Big)^2 \sum_{i<j} (h_i h_j' - h_j h_i')^2, \tag{142}$$

and we substitute $S_{11} = 2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}} \sum_j h_j^2$. □

We count crossings through the Crofton formula because, given $\mathcal{F}$, $s \mapsto g(\gamma(s))$ is only piecewise smooth. By Lemma D.2(b), the integrand is the Kac–Rice intensity [20, 21] of the zeros of the Gaussian process $s \mapsto g(\gamma(s))$ given $\mathcal{F}$, so Proposition D.3 is the Rice formula applied conditionally on $\mathcal{F}$. The output weights do not enter the result, because the Crofton formula integrates over $(\Delta_0, \ldots, \Delta_{p_{L-1}})$, and given $\mathcal{F}$ the expected number of crossings depends only on $\mathbf{S}$.

D.3 The estimate ρ_L

The crossing density is the expectation of $r(\mathbf{S})$. Since r is a nonlinear function of the random matrix $\mathbf{S}$, and the distribution of $\mathbf{S}$ depends on the parameters of layers $1, \ldots, L-1$, this expectation has no closed form. Our approach is to move the expectation inside r and evaluate r at the expected matrix,

$$\rho_L \;:=\; \frac{1}{\pi}\, r\big(\,\mathbb{E}_{\mathcal{F}}[\mathbf{S}]\big). \tag{143}$$

This requires only the three entries of $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$. By reflection symmetry the off-diagonal entry is zero. The upper-left entry is the mean square of g, which measures how far g is from zero, and it equals the quantity $2q^{(L)}$ of (20). The lower-right entry is the mean square of

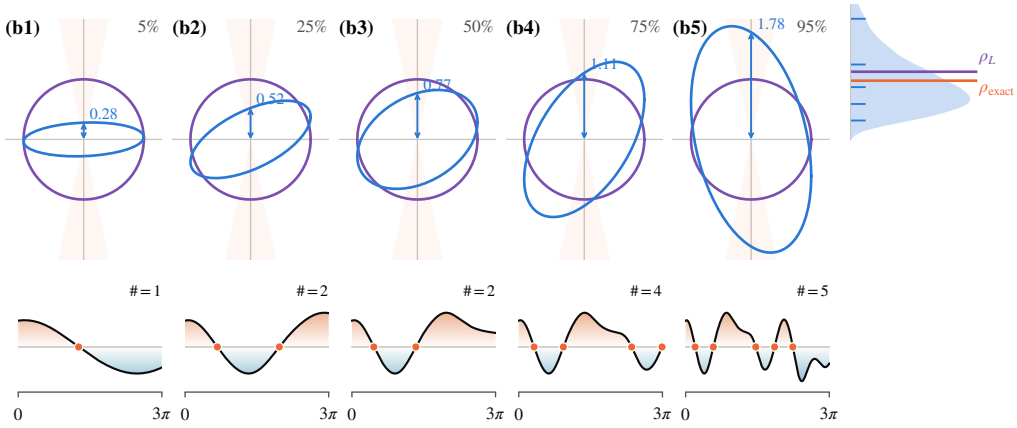


Figure 8: Five networks with $p_0 = 50$, $L = 4$, hidden widths 16, $\sigma_w^2 = 2$ and $\sigma_b^2 = 0.5$, taken at the 5%, 25%, 50%, 75% and 95% quantiles of $r(\mathbf{S})/r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$ over 2×10^5 networks. Panel (b3) is the median network, drawn thick in Figure 1(b). Top: the covariance ellipse of $(g, \partial_s g)$ of each network (blue), rescaled to half-width 1, so that the averaged ellipse $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$ is the unit circle (purple). The number is the height of the ellipse at $g = 0$, which equals $r(\mathbf{S})/r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$. A path crosses zero within the next ds exactly when $|g| < |\partial_s g|\, ds$ (orange cone), and a taller ellipse puts a larger share of its mass in the cone. Bottom: a stationary Gaussian path on an arc of length 3π whose value and derivative have variances 1 and x^2, where x is the height above, so that it has the same r as the ellipse. The five paths are rescalings $s \mapsto Z(xs)$ of one path Z, so the number of crossings (orange, count at right) grows with the height. Right: the distribution of the heights (blue), the five heights (ticks), their mean $\rho_{\text{exact}}/\rho_L$ (orange) and the height 1 of the averaged ellipse (purple). The relative gap between the orange and the purple line is δ.

the derivative of g along the circle, which measures how fast g changes. Within a cell the gradient of g is given by the zonotope generators, so this entry is the generator energy Γ_L times a constant that depends only on p_0. Hence ρ_L is proportional to $\sqrt{\Gamma_L / q^{(L)}}$, which is (19).

The estimate is accurate because $\mathbf{S}$ is a sum of p_{L-1} terms, one from each hidden unit. At large width $\mathbf{S}$ concentrates around $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$, so $\mathbb{E}_{\mathcal{F}}[r(\mathbf{S})]$ and $r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$ differ by a small amount. The relative difference between them comes from the fluctuation of $\mathbf{S}$, that is, from the relative variances of the entries of $\mathbf{S}$, and §D.4 computes this difference.

Figure 8 shows the matrices $\mathbf{S}$ of five networks and the crossings they produce. Each ellipse is rescaled so that its half-width along g is 1. After this rescaling, $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$ becomes the unit circle, and the height of an ellipse at $g = 0$ equals $r(\mathbf{S})/r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$. A Gaussian path whose value and derivative have the covariance $\mathbf{S}$ crosses zero at the rate $r(\mathbf{S})/\pi$, so a taller ellipse gives more crossings. The exact density ρ_{exact} averages the heights, whereas ρ_L is the height of the averaged ellipse.

Lemma D.4. *At every s and at every finite width,*

$$\mathbb{E}_{\mathcal{F}}\left[\mathbf{S}\right] = \begin{pmatrix} \lambda_0 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \qquad \lambda_0 = 2\, q^{(L)}, \qquad \lambda_2 = \frac{4}{p_0} \prod_{l=1}^{L} \frac{\sigma_l^2}{2} = \frac{\Gamma_L}{2(1 + 1/\pi)\, p_0}, \tag{144}$$

where $q^{(1)} = \sigma_1^2 R^2/p_0 + \sigma_{b,1}^2$ and, for $l = 2, \ldots, L$, $q^{(l)} = \frac{\sigma_l^2}{2} q^{(l-1)} + \sigma_{b,l}^2$, which is (20) *with the per-layer variances $\sigma_l^2, \sigma_{b,l}^2$.*

In Rice's notation, λ_0 and λ_2 are the zeroth and second spectral moments of $s \mapsto g(\gamma(s))$. By Lemma D.4,

$$\rho_L \;=\; \frac{1}{\pi}\sqrt{\frac{\lambda_2}{\lambda_0}} \;=\; \kappa(p_0)\sqrt{\frac{\Gamma_L}{q^{(L)}}}, \qquad \kappa(p_0) \;:=\; \frac{1}{2\pi\sqrt{(1+1/\pi)\,p_0}}, \tag{145}$$

which is a deterministic number that can be computed at every finite width.

Multiplying $\mathbf{W}^{(L-1)}$, $\boldsymbol{b}^{(L-1)}$ and $\boldsymbol{b}^{(L)}$ by $t > 0$ multiplies every h_j, every h'_j and $\sigma_{b,L}$ by t, and therefore, by (115), multiplies $\mathbf{S}$ by t^2. This multiplies Γ_L and $q^{(L)}$ by t^2 and leaves ρ_{exact} and ρ_L unchanged. When all biases are zero, multiplying the weights of any one hidden layer by t has the same effect. So Γ_L grows under such a rescaling while the crossing density stays the same.

When all biases are zero, the recursion gives $q^{(L)} = \frac{2R^2}{p_0}\prod_{l=1}^{L}\frac{\sigma_l^2}{2}$, hence $\lambda_0 = R^2\lambda_2$, and

$$\rho_L \;=\; \frac{1}{\pi R} \tag{146}$$

holds for every depth, every width and every choice of the weight variances. This is the crossing density of a hyperplane through the origin, which meets every great circle twice in a length of $2\pi R$.

For $l = 1, \dots, L-1$ and unit i of unit set l, we write its pre-activation as

$$y_i^{(l)} \;:=\; \sum_{k=1}^{p_{l-1}} W_{ik}^{(l)} h_k^{(l-1)} + b_i^{(l)}. \tag{147}$$

It is a function of $\boldsymbol{x}$, and by (1), $h_i^{(l)} = (y_i^{(l)})^+$. The proof of Lemma D.4 uses two facts.

Lemma D.5. *Let the random pair $(\boldsymbol{\theta}, \beta) \in \mathbb{R}^m \times \mathbb{R}$ satisfy $(\boldsymbol{\theta}, \beta) \overset{d}{=} (-\boldsymbol{\theta}, -\beta)$, let $\boldsymbol{v} \in \mathbb{R}^m$ be a deterministic vector, and let $\zeta := \langle \boldsymbol{\theta}, \boldsymbol{v}\rangle + \beta$. Let φ be an integrable function of $(\boldsymbol{\theta}, \beta)$ with real or real-matrix values that satisfies $\varphi(-\boldsymbol{\theta}, -\beta) = \varphi(\boldsymbol{\theta}, \beta)$. If φ vanishes on $\{\zeta = 0\}$, or if $\zeta \neq 0$ almost surely, then*

$$\mathbb{E}\left[\mathbf{1}\{\zeta > 0\}\,\varphi(\boldsymbol{\theta}, \beta)\right] = \frac{1}{2}\,\mathbb{E}\left[\varphi(\boldsymbol{\theta}, \beta)\right]. \tag{148}$$

Proof. The map $(\boldsymbol{\theta}, \beta) \mapsto (-\boldsymbol{\theta}, -\beta)$ preserves the distribution and preserves φ, and it sends ζ to $-\zeta$, hence sends $\mathbf{1}\{\zeta > 0\}$ to $\mathbf{1}\{\zeta < 0\}$. Therefore $\mathbb{E}[\mathbf{1}\{\zeta > 0\}\varphi] = \mathbb{E}[\mathbf{1}\{\zeta < 0\}\varphi]$. Adding the two gives $2\,\mathbb{E}[\mathbf{1}\{\zeta > 0\}\varphi] = \mathbb{E}[\varphi] - \mathbb{E}[\mathbf{1}\{\zeta = 0\}\varphi]$, and under either assumption the last term is zero. □

Lemma D.6. *For $\boldsymbol{x} \neq 0$ and $\Phi_{\mathcal{C}}^{(L-2)}$ at $\boldsymbol{x}$,*

$$\mathbb{E}\left[\Phi_{\mathcal{C}}^{(L-2)\top}\Phi_{\mathcal{C}}^{(L-2)}\right] \;=\; \frac{p_{L-2}}{p_0}\prod_{l=1}^{L-2}\frac{\sigma_l^2}{2}\,\mathbf{I}, \tag{149}$$

where $\Phi_{\mathcal{C}}^{(L-2)} = \mathbf{I}$ when $L = 2$.

Proof. Let $\mathbf{D}^{(l)}$ be the diagonal matrix formed by the activation indicators $\mathbf{1}\{y_i^{(l)} > 0\}$ $(i = 1, \dots, p_l)$ of unit set l at $\boldsymbol{x}$. Let $\Phi^{(0)} := \mathbf{I}$ and, for $l \geq 1$, $\Phi^{(l)} := \mathbf{D}^{(l)}\mathbf{W}^{(l)}\Phi^{(l-1)}$, so that $\Phi^{(l)}$ is $\Phi_{\mathcal{C}}^{(l)}$ at $\boldsymbol{x}$. Fix $l \geq 1$ and condition on layers $1, \dots, l-1$; then $\Phi^{(l-1)}$ and $\boldsymbol{h}^{(l-1)}$ are deterministic. Since $\mathbf{D}^{(l)}$ is diagonal with diagonal entries 0 or 1, we have $\mathbf{D}^{(l)\top}\mathbf{D}^{(l)} = \mathbf{D}^{(l)}$, and $\mathbf{W}^{(l)\top}\mathbf{D}^{(l)}\mathbf{W}^{(l)}$ is the weighted sum of the outer products $\mathbf{W}_{i,:}^{(l)\top}\mathbf{W}_{i,:}^{(l)}$ of the rows $\mathbf{W}_{i,:}^{(l)}$ of $\mathbf{W}^{(l)}$, where the weight is the i-th diagonal entry of $\mathbf{D}^{(l)}$. Therefore

$$\Phi^{(l)\top}\Phi^{(l)} = \Phi^{(l-1)\top}\mathbf{W}^{(l)\top}\mathbf{D}^{(l)}\mathbf{W}^{(l)}\Phi^{(l-1)} = \Phi^{(l-1)\top}\Big(\sum_{i=1}^{p_l}\mathbf{1}\{y_i^{(l)} > 0\}\,\mathbf{W}_{i,:}^{(l)\top}\mathbf{W}_{i,:}^{(l)}\Big)\Phi^{(l-1)}. \tag{150}$$

When $\boldsymbol{h}^{(l-1)} \neq 0$, take, for row i, $\boldsymbol{\theta} = \mathbf{W}_{i,:}^{(l)\top}$, $\beta = b_i^{(l)}$ and $\boldsymbol{v} = \boldsymbol{h}^{(l-1)}$. Then $\zeta = y_i^{(l)}$ is a centered Gaussian variable with positive variance, and taking $\varphi(\boldsymbol{\theta}, \beta) = \boldsymbol{\theta}\boldsymbol{\theta}^\top$ in Lemma D.5 gives

$$\mathbb{E}\left[\mathbf{1}\{y_i^{(l)} > 0\}\, \mathbf{W}_{i,:}^{(l)\top} \mathbf{W}_{i,:}^{(l)}\right] = \frac{1}{2}\, \mathbb{E}\left[\mathbf{W}_{i,:}^{(l)\top} \mathbf{W}_{i,:}^{(l)}\right] = \frac{\sigma_l^2}{2p_{l-1}}\, \mathbf{I}. \tag{151}$$

Summing over i gives $\mathbb{E}[\Phi^{(l)\top}\Phi^{(l)} \mid \text{layers } 1, \ldots, l-1] = \frac{p_l \sigma_l^2}{2p_{l-1}}\, \Phi^{(l-1)\top}\Phi^{(l-1)}$. When $\boldsymbol{h}^{(l-1)} = 0$ and $l \geq 2$, no unit of unit set $l-1$ is active, so $\mathbf{D}^{(l-1)} = 0$ and $\Phi^{(l-1)} = 0$, and both sides of the same identity are zero. When $l = 1$ we have $\boldsymbol{h}^{(0)} = \boldsymbol{x} \neq 0$. Taking expectations and iterating from $l = 1$ to $l = L-2$,

$$\mathbb{E}\left[\Phi^{(L-2)\top}\Phi^{(L-2)}\right] = \prod_{l=1}^{L-2} \frac{p_l \sigma_l^2}{2p_{l-1}}\, \mathbf{I} = \frac{p_{L-2}}{p_0} \prod_{l=1}^{L-2} \frac{\sigma_l^2}{2}\, \mathbf{I}, \tag{152}$$

where the second equality cancels $p_1, \ldots, p_{L-3}$. □

S_{11}. For $l = 1, \ldots, L-1$ and unit i of unit set l, condition on layers $1, \ldots, l-1$ and take $\boldsymbol{\theta} = \mathbf{W}_{i,:}^{(l)\top}$, $\beta = b_i^{(l)}$ and $\boldsymbol{v} = \boldsymbol{h}^{(l-1)}$, so that $\zeta = y_i^{(l)}$. Taking $\varphi = \zeta^2$ in Lemma D.5, which vanishes on $\{\zeta = 0\}$, shows that the second moment of $h_i^{(l)}$ is $\frac{1}{2}\,\mathbb{E}[(y_i^{(l)})^2]$. Since the rows of $\mathbf{W}^{(l)}$ are independent of $\boldsymbol{h}^{(l-1)}$, $\mathbb{E}[(y_i^{(l)})^2] = \frac{\sigma_l^2}{p_{l-1}}\, \mathbb{E}\left\|\boldsymbol{h}^{(l-1)}\right\|_2^2 + \sigma_{b,l}^2$. This is the recursion for $q^{(l)}$, and at $l = 1$ the input satisfies $\|\boldsymbol{x}\|_2^2 = R^2$. In particular $\mathbb{E}[h_j^2] = \frac{1}{2} q^{(L-1)}$. At the output, by (115),

$$\begin{aligned} \mathbb{E}[S_{11}] &= 2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}} \sum_{j=1}^{p_{L-1}} \mathbb{E}[h_j^2] = 2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}} \cdot p_{L-1} \cdot \frac{1}{2} q^{(L-1)} \\ &= 2\Big(\frac{\sigma_L^2}{2} q^{(L-1)} + \sigma_{b,L}^2\Big) = 2q^{(L)}. \end{aligned} \tag{153}$$

S_{22}. Condition on layers $1, \ldots, L-2$, and let $\gamma(s)$ lie in a cell $\mathcal{C}$ of Section 3. With $\mathbf{A} = \mathbf{W}^{(L-1)}$, write $\Phi := \Phi_{\mathcal{C}}^{(L-2)}$. By the definitions in §D.1, $\boldsymbol{u}_j^\top$ is the j-th row of $\mathbf{A}\Phi$, and ν_j is the constant term of the pre-activation of unit j in $\mathcal{C}$. Let $d_j \in \{0, 1\}$ indicate whether unit j is active at $\gamma(s)$; then $h_j = d_j(\langle \boldsymbol{u}_j, \gamma \rangle + \nu_j)$. Since d_j and ν_j do not depend on s within the cell, differentiating in s gives $h_j' = d_j \langle \boldsymbol{u}_j, \dot{\gamma} \rangle$. From $\langle \boldsymbol{u}_j, \dot{\gamma} \rangle = \mathbf{A}_{j,:}\Phi\dot{\gamma}$ we obtain $\langle \boldsymbol{u}_j, \dot{\gamma} \rangle^2 = \dot{\gamma}^\top \Phi^\top (\mathbf{A}_{j,:}^\top \mathbf{A}_{j,:})\, \Phi\dot{\gamma}$. When $\boldsymbol{h}^{(L-2)}(\gamma(s)) \neq 0$, take $\boldsymbol{\theta} = \mathbf{A}_{j,:}^\top$, $\beta = b_j^{(L-1)}$ and $\boldsymbol{v} = \boldsymbol{h}^{(L-2)}(\gamma(s))$. Then $\zeta = y_j^{(L-1)}$ is a centered Gaussian variable with positive variance, $\mathbf{1}\{\zeta > 0\} = d_j$, and Φ is independent of $\mathbf{A}_{j,:}$. Taking $\varphi(\boldsymbol{\theta}, \beta) = \boldsymbol{\theta}\boldsymbol{\theta}^\top$ in Lemma D.5 gives $\mathbb{E}[d_j \mathbf{A}_{j,:}^\top \mathbf{A}_{j,:}] = \frac{1}{2}\,\mathbb{E}[\mathbf{A}_{j,:}^\top \mathbf{A}_{j,:}]$, and substituting this into the expression above gives $\mathbb{E}[d_j \langle \boldsymbol{u}_j, \dot{\gamma} \rangle^2] = \frac{1}{2}\,\mathbb{E}\langle \boldsymbol{u}_j, \dot{\gamma} \rangle^2$. When $\boldsymbol{h}^{(L-2)} = 0$, both sides are zero because $\Phi = 0$. The entries of $\mathbf{A}$ have variance σ_{L-1}^2 / p_{L-2} and are independent of Φ, so by Lemma D.6,

$$\mathbb{E}\left[\boldsymbol{u}_j \boldsymbol{u}_j^\top\right] \;=\; \frac{\sigma_{L-1}^2}{p_{L-2}}\, \mathbb{E}\left[\Phi^\top \Phi\right] \;=\; \frac{2}{p_0} \prod_{l=1}^{L-1} \frac{\sigma_l^2}{2}\, \mathbf{I}, \tag{154}$$

and $\|\dot{\gamma}\|_2 = 1$ gives $\mathbb{E}\langle \boldsymbol{u}_j, \dot{\gamma} \rangle^2 = \dot{\gamma}^\top\, \mathbb{E}[\boldsymbol{u}_j \boldsymbol{u}_j^\top] \dot{\gamma} = \frac{2}{p_0} \prod_{l=1}^{L-1} \frac{\sigma_l^2}{2}$. From $h_j' = d_j \langle \boldsymbol{u}_j, \dot{\gamma} \rangle$ we obtain $\mathbb{E}[(h_j')^2] = \mathbb{E}[d_j \langle \boldsymbol{u}_j, \dot{\gamma} \rangle^2] = \frac{1}{p_0} \prod_{l=1}^{L-1} \frac{\sigma_l^2}{2}$. Summing over the p_{L-1} units by (115),

$$\mathbb{E}[S_{22}] = \frac{2\sigma_L^2}{p_{L-1}} \cdot p_{L-1} \cdot \frac{1}{p_0} \prod_{l=1}^{L-1} \frac{\sigma_l^2}{2} = \frac{4}{p_0} \cdot \frac{\sigma_L^2}{2} \prod_{l=1}^{L-1} \frac{\sigma_l^2}{2} = \frac{4}{p_0} \prod_{l=1}^{L} \frac{\sigma_l^2}{2}. \tag{155}$$

Dividing $\Gamma_L = 8(1 + \frac{1}{\pi}) \prod_{l=1}^{L} \frac{\sigma_l^2}{2}$ (61) by this value gives $2(1 + \frac{1}{\pi})p_0$.

S_{12}. At $s = 0$, the reflection $\boldsymbol{e}_2 \mapsto -\boldsymbol{e}_2$ fixes $\gamma(0)$ and maps $\gamma(s)$ to $\gamma(-s)$. By Lemma D.1, $(h_1(0), \dots, h_{p_{L-1}}(0), h_1'(0), \dots, h_{p_{L-1}}'(0))$ and $(h_1(0), \dots, h_{p_{L-1}}(0), -h_1'(0), \dots, -h_{p_{L-1}}'(0))$ have the same distribution, so by (115), $\mathbb{E}[S_{12}(0)] = \frac{2\sigma_L^2}{p_{L-1}} \sum_{j=1}^{p_{L-1}} \mathbb{E}[h_j(0) h_j'(0)] = 0$. By rotation, the same conclusion holds at every s. □

D.4 Error at finite width

In §D.2 and §D.3 we obtained the expected number of crossings of the decision boundary per unit arc length along a great circle of the input space, $\rho_{\text{exact}} = \frac{1}{\pi} \mathbb{E}_{\mathcal{F}}[r(\mathbf{S})]$, and its estimate $\rho_L = \frac{1}{\pi} r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$. This expected number is taken over (i) $\mathcal{F}$ and (ii) $\{\Delta_j\}$. The expectation over $\{\Delta_j\}$ is already carried out inside $r(\mathbf{S})$, so the expectations in both formulas are over $\mathcal{F}$ only. The randomness of $\mathcal{F}$ makes $\mathbf{S}$ a random matrix. At infinite width, $\mathbf{S}$ concentrates at $\mathbb{E}_{\mathcal{F}}[\mathbf{S}]$, and then $\rho_{\text{exact}}/\rho_L \to 1$.

At finite width we study the relative error between ρ_{exact} and ρ_L,

$$\delta := \frac{\rho_{\text{exact}}}{\rho_L} - 1 = \frac{\mathbb{E}_{\mathcal{F}}[r(\mathbf{S})]}{r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])} - 1. \tag{156}$$

We show that ρ_L slightly overestimates ρ_{exact}, and that the overestimate decreases as the width grows. When all biases are zero, δ has an exact product formula in which each hidden unit set contributes one factor, and this factor tends to 1 exponentially in the width of that unit set. When the biases are nonzero, we regard δ as a function of the inverse widths $1/p_1, \dots, 1/p_{L-1}$ of the hidden unit sets and compute its first-order term in these inverses, $\sum_{l=1}^{L-1} c_l/p_l$, where every coefficient satisfies $c_l \le 0$.

To compute δ, we compare the random matrix $\mathbf{S}$ with its expectation under the distribution of $\mathcal{F}$. Since handling all of $\mathcal{F}$ at once is hard, we start from the input unit set. The matrix $\mathbf{S}^{(0)}$ at the input has no randomness at all. With each additional layer, $\mathbf{S}$ depends on one more layer of random weights. We compute how much the deviation grows when one layer is added, and then combine the layers.

For $l = 1, \dots, L-1$ and unit i of unit set l, following the notation $h_j(s)$, we write $y_i^{(l)}(s) := y_i^{(l)}(\gamma(s))$ and $h_i^{(l)}(s) := h_i^{(l)}(\gamma(s))$, so that by (1), $h_i^{(l)}(s) = (y_i^{(l)}(s))^+$. We also write $y_i^{(l)\prime}(s)$ and $h_i^{(l)\prime}(s)$ for their derivatives in s. At every s where unit i does not change its activation state,

$$h_i^{(l)} = \mathbf{1}\{y_i^{(l)} > 0\}\, y_i^{(l)}, \qquad h_i^{(l)\prime} = \mathbf{1}\{y_i^{(l)} > 0\}\, y_i^{(l)\prime}. \tag{157}$$

For $l = L-1$ this definition is the definition of $h_j(s)$. By the convention $\boldsymbol{h}^{(0)} = \boldsymbol{x}$ of (1), $h_k^{(0)}(s) = \gamma_k(s)$ is the k-th coordinate of the circle and $h_k^{(0)\prime}(s) = \dot{\gamma}_k(s)$. For $l = 0, \dots, L-1$, the Gram matrix of unit set l is

$$\mathbf{S}^{(l)} := \sum_{i=1}^{p_l} \begin{bmatrix} (h_i^{(l)})^2 & h_i^{(l)} h_i^{(l)\prime} \\ h_i^{(l)} h_i^{(l)\prime} & (h_i^{(l)\prime})^2 \end{bmatrix}. \tag{158}$$

Proposition D.7 (Layer-wise recursion for $\mathbf{S}$)**.** *Fix s.*

(a) The starting point is the deterministic matrix $\mathbf{S}^{(0)} = \operatorname{diag}(R^2, 1)$.

(b) The end point satisfies $\mathbf{S} = 2\sigma_{b,L}^2 \operatorname{diag}(1, 0) + \frac{2\sigma_L^2}{p_{L-1}} \mathbf{S}^{(L-1)}$.

(c) For $l = 1, \dots, L-1$, conditionally on layers $1, \dots, l-1$, the two-dimensional vectors $(y_i^{(l)}, y_i^{(l)\prime})$, $i = 1, \dots, p_l$, are independent and identically distributed as a centered Gaussian with covariance matrix

$$\frac{\sigma_l^2}{p_{l-1}} \mathbf{S}^{(l-1)} + \sigma_{b,l}^2 \operatorname{diag}(1, 0). \tag{159}$$

When no unit of unit set l changes its activation state at $\gamma(s)$,

$$\mathbf{S}^{(l)} = \sum_{i=1}^{p_l} \mathbf{1}\{y_i^{(l)} > 0\} \begin{bmatrix} y_i^{(l)} \\ y_i^{(l)\prime} \end{bmatrix} \begin{bmatrix} y_i^{(l)} & y_i^{(l)\prime} \end{bmatrix}. \tag{160}$$

Proof. (a) At $l = 0$, substitute $h_k^{(0)} = \gamma_k$ and $h_k^{(0)\prime} = \dot\gamma_k$, and then $\|\gamma\|_2 = R$, $\langle\gamma, \dot\gamma\rangle = 0$ and $\|\dot\gamma\|_2 = 1$, to obtain

$$\mathbf{S}^{(0)} = \begin{bmatrix} \sum_k \gamma_k^2 & \sum_k \gamma_k\dot\gamma_k \\ \sum_k \gamma_k\dot\gamma_k & \sum_k \dot\gamma_k^2 \end{bmatrix} = \begin{bmatrix} \|\gamma\|_2^2 & \langle\gamma,\dot\gamma\rangle \\ \langle\gamma,\dot\gamma\rangle & \|\dot\gamma\|_2^2 \end{bmatrix} = \begin{bmatrix} R^2 & 0 \\ 0 & 1 \end{bmatrix}. \tag{161}$$

(b) At $l = L-1$, $\mathbf{S}^{(L-1)}$ equals the sum over j in (115), so

$$\mathbf{S} \;=\; 2\sigma_{b,L}^2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{2\sigma_L^2}{p_{L-1}}\, \mathbf{S}^{(L-1)}. \tag{162}$$

(c) At $\gamma(s)$, substitute the definition of $y_i^{(l)}$, $y_i^{(l)} = \sum_k W_{ik}^{(l)} h_k^{(l-1)} + b_i^{(l)}$, where the sum over k runs from 1 to p_{l-1}. Since $W_{ik}^{(l)}$ and $b_i^{(l)}$ do not depend on s, $y_i^{(l)\prime} = \sum_k W_{ik}^{(l)} h_k^{(l-1)\prime}$. Together,

$$\begin{bmatrix} y_i^{(l)} \\ y_i^{(l)\prime} \end{bmatrix} \;=\; \sum_{k=1}^{p_{l-1}} W_{ik}^{(l)} \begin{bmatrix} h_k^{(l-1)} \\ h_k^{(l-1)\prime} \end{bmatrix} \;+\; \begin{bmatrix} b_i^{(l)} \\ 0 \end{bmatrix}. \tag{163}$$

Conditionally on layers $1, \dots, l-1$, every $(h_k^{(l-1)}, h_k^{(l-1)\prime})$ is deterministic. Different i use different rows of $\mathbf{W}^{(l)}$ and different $b_i^{(l)}$, so the pairs $(y_i^{(l)}, y_i^{(l)\prime})$ are independent and identically distributed. Each pair is a linear combination of the independent centered Gaussian variables $W_{ik}^{(l)}$ and $b_i^{(l)}$, hence a centered Gaussian vector. By $\mathrm{Var}(W_{ik}^{(l)}) = \sigma_l^2/p_{l-1}$, $\mathrm{Var}(b_i^{(l)}) = \sigma_{b,l}^2$ and their independence, its covariance matrix is

$$\sum_{k=1}^{p_{l-1}} \frac{\sigma_l^2}{p_{l-1}} \begin{bmatrix} h_k^{(l-1)} \\ h_k^{(l-1)\prime} \end{bmatrix} \begin{bmatrix} h_k^{(l-1)} & h_k^{(l-1)\prime} \end{bmatrix} + \sigma_{b,l}^2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \frac{\sigma_l^2}{p_{l-1}}\, \mathbf{S}^{(l-1)} + \sigma_{b,l}^2\, \mathrm{diag}(1,0), \tag{164}$$

where the equality substitutes the definition (158) at index $l-1$. Finally, substituting $h_i^{(l)} = \mathbf{1}\{y_i^{(l)} > 0\}\, y_i^{(l)}$ and $h_i^{(l)\prime} = \mathbf{1}\{y_i^{(l)} > 0\}\, y_i^{(l)\prime}$ into the definition (158) at index l, and using $\mathbf{1}\{\cdot\}^2 = \mathbf{1}\{\cdot\}$, gives (160). □

When $\sigma_{b,L} = 0$, part (b) shows that $\mathbf{S}$ is $2\sigma_L^2/p_{L-1}$ times $\mathbf{S}^{(L-1)}$, so by the invariance of r under multiplication by a positive number (stated after (116)), $r(\mathbf{S}) = r(\mathbf{S}^{(L-1)})$.

With zero biases, conditionally on layers $1, \dots, l-1$ and on the number of active units of unit set l, $\mathbf{S}^{(l)}$ is a Wishart matrix. A Wishart matrix has the property that its r averages to the value of r at the scale matrix. We write $\mathcal{W}_2(n, \mathbf{T})$ for the distribution of the sum of the outer products of n independent $\mathcal{N}(0, \mathbf{T})$ vectors, that is, the 2×2 Wishart distribution with n degrees of freedom and scale $\mathbf{T}$ [15].

Lemma D.8 (r of a Wishart matrix)**.** *Let $\mathbf{K} \sim \mathcal{W}_2(n, \mathbf{T})$ with $\mathbf{T} \succ 0$ and $n \geq 2$. Then*

$$\mathbb{E}\left[r(\mathbf{K})\right] \;=\; r(\mathbf{T}). \tag{165}$$

Proof. We first separate out $\mathbf{T}$. Write $\mathbf{K} = \mathbf{T}^{1/2}\mathbf{K}_0\mathbf{T}^{1/2}$ with $\mathbf{K}_0 \sim \mathcal{W}_2(n, \mathbf{I})$. Then $\det\mathbf{K} = \det\mathbf{T}\,\det\mathbf{K}_0$. Let $\boldsymbol{\iota} := (1,0)^\top$ and $\boldsymbol{u} := \mathbf{T}^{1/2}\boldsymbol{\iota} / \left\|\mathbf{T}^{1/2}\boldsymbol{\iota}\right\|_2$. It is a unit vector, and $\left\|\mathbf{T}^{1/2}\boldsymbol{\iota}\right\|_2^2 = \boldsymbol{\iota}^\top\mathbf{T}\boldsymbol{\iota} = \mathbf{T}_{11}$, so

$$\mathbf{K}_{11} = \boldsymbol{\iota}^\top\mathbf{T}^{1/2}\mathbf{K}_0\mathbf{T}^{1/2}\boldsymbol{\iota} = \mathbf{T}_{11}\, \boldsymbol{u}^\top\mathbf{K}_0\boldsymbol{u}. \tag{166}$$

Take a rotation $\mathbf{Q}$ with $\mathbf{Q}\boldsymbol{u} = \boldsymbol{\iota}$. Since $\mathbf{K}_0$ and $\mathbf{Q}\mathbf{K}_0\mathbf{Q}^\top$ have the same distribution, $(\boldsymbol{u}^\top\mathbf{K}_0\boldsymbol{u}, \det\mathbf{K}_0)$ and $((\mathbf{K}_0)_{11}, \det\mathbf{K}_0)$ have the same distribution. Substituting into (116),

$$r(\mathbf{K}) = \frac{\sqrt{\det\mathbf{T}\,\det\mathbf{K}_0}}{\mathbf{T}_{11}\,\boldsymbol{u}^\top\mathbf{K}_0\boldsymbol{u}} \;\overset{d}{=}\; \frac{\sqrt{\det\mathbf{T}}}{\mathbf{T}_{11}} \cdot \frac{\sqrt{\det\mathbf{K}_0}}{(\mathbf{K}_0)_{11}} = r(\mathbf{T})\, r(\mathbf{K}_0). \tag{167}$$

It therefore suffices to show $\mathbb{E}[r(\mathbf{K}_0)] = 1$.

Write $\mathbf{K}_0$ as the sum of the outer products of n independent $\mathcal{N}(0, \mathbf{I}_2)$ vectors. Collect the first coordinates of these n vectors into $\boldsymbol{a} \in \mathbb{R}^n$ and the second coordinates into $\boldsymbol{c} \in \mathbb{R}^n$. Then $\boldsymbol{a}$ and $\boldsymbol{c}$ are independent standard Gaussian vectors, and

$$(\mathbf{K}_0)_{11} = \|\boldsymbol{a}\|_2^2, \qquad \det \mathbf{K}_0 = \|\boldsymbol{a}\|_2^2 \, \|\boldsymbol{c}\|_2^2 - \langle \boldsymbol{a}, \boldsymbol{c} \rangle^2. \tag{168}$$

The latter is the squared area of the parallelogram spanned by $\boldsymbol{a}$ and $\boldsymbol{c}$. Let $\boldsymbol{c}_\perp$ be the component of $\boldsymbol{c}$ orthogonal to $\boldsymbol{a}$. Then this area equals $\|\boldsymbol{a}\|_2 \, \|\boldsymbol{c}_\perp\|_2$, and so

$$r(\mathbf{K}_0) = \frac{\|\boldsymbol{a}\|_2 \, \|\boldsymbol{c}_\perp\|_2}{\|\boldsymbol{a}\|_2^2} = \frac{\|\boldsymbol{c}_\perp\|_2}{\|\boldsymbol{a}\|_2}, \tag{169}$$

which is the height of the parallelogram divided by the length of its base. Given $\boldsymbol{a}$, $\boldsymbol{c}_\perp$ is a standard Gaussian vector in the $(n-1)$-dimensional orthogonal complement of $\boldsymbol{a}$, so $\|\boldsymbol{c}_\perp\|_2^2 \sim \chi^2_{n-1}$ and $\|\boldsymbol{a}\|_2^2 \sim \chi^2_n$, and the two are independent. Using $\mathbb{E}\sqrt{\chi^2_k} = \sqrt{2}\,\Gamma(\frac{k+1}{2})/\Gamma(\frac{k}{2})$ $(k \geq 1)$ and $\mathbb{E}[1/\sqrt{\chi^2_k}] = \Gamma(\frac{k-1}{2})/(\sqrt{2}\,\Gamma(\frac{k}{2}))$ $(k \geq 2)$, where Γ is the Euler Gamma function,

$$\mathbb{E}\left[r(\mathbf{K}_0)\right] = \mathbb{E}\,\|\boldsymbol{c}_\perp\|_2 \; \mathbb{E}\,\frac{1}{\|\boldsymbol{a}\|_2} = \sqrt{2}\,\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \cdot \frac{\Gamma(\frac{n-1}{2})}{\sqrt{2}\,\Gamma(\frac{n}{2})} = 1. \tag{170}$$

□

For $n = 1$, $\boldsymbol{a}$ and $\boldsymbol{c}$ are collinear, the parallelogram has zero height, and $r = 0$. The factor $1 - (1 + p_l)/2^{p_l}$ in Proposition D.9 is the probability that remains after this case (and the case of no active unit) is removed.

Proposition D.9 (Zero bias)**.** *If $\sigma_{b,1} = \cdots = \sigma_{b,L} = 0$, then*

$$\frac{\rho_{\text{exact}}}{\rho_L} \;=\; \frac{\mathbb{E}_{\mathcal{F}}[r(\mathbf{S})]}{r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])} \;=\; \prod_{l=1}^{L-1} \left(1 - \frac{1 + p_l}{2^{p_l}}\right) \tag{171}$$

for every choice of the weight variances. The l-th factor is the probability, conditionally on $\boldsymbol{h}^{(l-1)}(\gamma(s)) \neq 0$, that at least two units of unit set l are active at $\gamma(s)$.

Proof. For $l = 0, \ldots, L-1$, let

$$t_l \;:=\; \frac{\mathbb{E}_{\mathcal{F}}[r(\mathbf{S}^{(l)})]}{r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(l)}])}. \tag{172}$$

We show that $t_0 = 1$, that t_{L-1} equals the left-hand side of (171), and that each added layer multiplies t_l by the factor $1 - (1 + p_l)/2^{p_l}$.

The two ends. By Proposition D.7(a), $\mathbf{S}^{(0)} = \operatorname{diag}(R^2, 1)$ is deterministic, so $\mathbb{E}_{\mathcal{F}}[r(\mathbf{S}^{(0)})] = r(\mathbf{S}^{(0)}) = r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(0)}])$, that is, $t_0 = 1$. By Proposition D.7(b) and $\sigma_{b,L} = 0$, $\mathbf{S} = \alpha\,\mathbf{S}^{(L-1)}$, where $\alpha := 2\sigma_L^2/p_{L-1}$ is a deterministic positive number. Hence $\mathbb{E}_{\mathcal{F}}[\mathbf{S}] = \alpha\;\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(L-1)}]$, and by the invariance of r under multiplication by a positive number,

$$\frac{\mathbb{E}_{\mathcal{F}}[r(\mathbf{S})]}{r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])} = \frac{\mathbb{E}_{\mathcal{F}}[r(\alpha\,\mathbf{S}^{(L-1)})]}{r(\alpha\;\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(L-1)}])} = \frac{\mathbb{E}_{\mathcal{F}}[r(\mathbf{S}^{(L-1)})]}{r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(L-1)}])} = t_{L-1}. \tag{173}$$

The denominator. Fix $l \geq 1$ and condition on layers $1, \ldots, l-1$. By Proposition D.7(c) and zero bias, the pairs $(y_i^{(l)}, y_i^{(l)\prime})$ are independent and identically distributed as $\mathcal{N}(0, \frac{\sigma_l^2}{p_{l-1}}\mathbf{S}^{(l-1)})$. When $\boldsymbol{h}^{(l-1)} \neq 0$, take $\boldsymbol{\theta} = \mathbf{W}_{i,:}^{(l)\top}$, $\beta = 0$ and $\boldsymbol{v} = \boldsymbol{h}^{(l-1)}$ in Lemma D.5; then $\zeta = y_i^{(l)}$ has positive variance. Take φ to be $(y_i^{(l)}, y_i^{(l)\prime})^\top (y_i^{(l)}, y_i^{(l)\prime})$, which is an even function of $\boldsymbol{\theta}$. Then

$$\mathbb{E}\left[\mathbf{1}\{y_i^{(l)} > 0\} \begin{bmatrix} y_i^{(l)} \\ y_i^{(l)\prime} \end{bmatrix} \begin{bmatrix} y_i^{(l)} & y_i^{(l)\prime} \end{bmatrix} \,\middle|\, \text{layers } 1, \ldots, l-1\right] = \frac{1}{2} \cdot \frac{\sigma_l^2}{p_{l-1}}\,\mathbf{S}^{(l-1)}. \tag{174}$$

When $\boldsymbol{h}^{(l-1)} = 0$, we have $\mathbf{S}^{(l-1)} = 0$ and both sides are zero. Summing over i and using (160),

$$\mathbb{E}\left[\mathbf{S}^{(l)} \,\middle|\, \text{layers } 1, \dots, l-1\right] = \frac{p_l\, \sigma_l^2}{2\, p_{l-1}}\, \mathbf{S}^{(l-1)}. \tag{175}$$

Taking the expectation over layers $1, \dots, l-1$ gives $\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(l)}] = \frac{p_l \sigma_l^2}{2 p_{l-1}}\, \mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(l-1)}]$. Iterating from $l = 1$, every $\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(l)}]$ is a positive multiple of $\mathbf{S}^{(0)}$, so by the invariance of r under multiplication by a positive number and (116),

$$r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}^{(l)}]) = r(\mathbf{S}^{(0)}) = \frac{\sqrt{R^2 \cdot 1 - 0^2}}{R^2} = \frac{1}{R}, \qquad l = 0, \dots, L-1. \tag{176}$$

The numerator. Fix $l \geq 1$ and condition on layers $1, \dots, l-1$, with $\boldsymbol{h}^{(l-1)} \neq 0$. As in the denominator, the pairs $(y_i^{(l)}, y_i^{(l)\prime})$ are independent and identically distributed as $\mathcal{N}(0, \frac{\sigma_l^2}{p_{l-1}} \mathbf{S}^{(l-1)})$, and the distribution of each pair is invariant under $(y_i^{(l)}, y_i^{(l)\prime}) \mapsto (-y_i^{(l)}, -y_i^{(l)\prime})$. This map preserves the matrix $(y_i^{(l)}, y_i^{(l)\prime})^\top (y_i^{(l)}, y_i^{(l)\prime})$ and reverses the sign of $y_i^{(l)}$. Therefore the activation indicators $\mathbf{1}\{y_i^{(l)} > 0\}$ are independent and identically distributed as Bernoulli($\frac{1}{2}$), and they are independent of these matrices. By (160), given that n units are active, $\mathbf{S}^{(l)}$ is the sum of the outer products of n independent $\mathcal{N}(0, \frac{\sigma_l^2}{p_{l-1}} \mathbf{S}^{(l-1)})$ vectors, that is, it follows $\mathcal{W}_2(n, \frac{\sigma_l^2}{p_{l-1}} \mathbf{S}^{(l-1)})$. When $n \geq 2$ and $\mathbf{S}^{(l-1)}$ has full rank, taking $\mathbf{T} = \frac{\sigma_l^2}{p_{l-1}} \mathbf{S}^{(l-1)}$ in Lemma D.8 and using the invariance of r under multiplication by a positive number gives $\mathbb{E}[r(\mathbf{S}^{(l)}) \mid n] = r(\frac{\sigma_l^2}{p_{l-1}} \mathbf{S}^{(l-1)}) = r(\mathbf{S}^{(l-1)})$. When $n \leq 1$, or when $\mathbf{S}^{(l-1)}$ has rank one, the vectors $(h_i^{(l)}, h_i^{(l)\prime})$ of all active units lie on a common line through the origin, and $r(\mathbf{S}^{(l)}) = 0$; in the latter case $r(\mathbf{S}^{(l-1)}) = 0$ as well. When $\boldsymbol{h}^{(l-1)} = 0$, all $(y_i^{(l)}, y_i^{(l)\prime})$ are zero, and $r(\mathbf{S}^{(l)}) = 0 = r(\mathbf{S}^{(l-1)})$. In all cases

$$\mathbb{E}\left[r(\mathbf{S}^{(l)}) \,\middle|\, \text{layers } 1, \dots, l-1\right] \;=\; \left(1 - \frac{1 + p_l}{2^{p_l}}\right) r(\mathbf{S}^{(l-1)}), \tag{177}$$

because $1 - (1 + p_l) 2^{-p_l}$ is the probability that a $\mathrm{Bin}(p_l, \frac{1}{2})$ random variable is at least 2. Taking the expectation over layers $1, \dots, l-1$ gives $\mathbb{E}_{\mathcal{F}}[r(\mathbf{S}^{(l)})] = \left(1 - (1 + p_l) 2^{-p_l}\right)\, \mathbb{E}_{\mathcal{F}}[r(\mathbf{S}^{(l-1)})]$.

Combining. By the result for the denominator, the denominator of every t_l is $1/R$, so by the result for the numerator, for $l = 1, \dots, L-1$,

$$t_l = R\, \mathbb{E}_{\mathcal{F}}[r(\mathbf{S}^{(l)})] = \left(1 - \frac{1 + p_l}{2^{p_l}}\right) R\, \mathbb{E}_{\mathcal{F}}[r(\mathbf{S}^{(l-1)})] = \left(1 - \frac{1 + p_l}{2^{p_l}}\right) t_{l-1}. \tag{178}$$

Iterating from $t_0 = 1$ to $l = L - 1$ gives $t_{L-1} = \prod_{l=1}^{L-1} (1 - (1 + p_l) 2^{-p_l})$, and the result for the two ends then gives (171). □

With zero biases, the error comes only from the event that in some unit set at most one unit is active. In that event the parallelogram in Lemma D.8 degenerates to a segment, $r = 0$, and there is no crossing at that point. The probability of such an event decreases exponentially in the width, so δ has no term of order $1/p_{\min}$, where $p_{\min} := \min_{1 \leq l \leq L-1} p_l$. At $L = 4$, $p_l \equiv 8$ gives $\delta = -10.2\%$, and $p_l \equiv 16$ gives $\delta = -0.078\%$.

Let $\mathbf{C}^{(l)}$ denote the covariance of the vector $(y_i^{(l)}, y_i^{(l)\prime})$ of unit set l, conditioning on layers $1, \dots, l-1$. By Proposition D.7(c),

$$\mathbf{C}^{(l)} = \frac{\sigma_l^2}{p_{l-1}}\, \mathbf{S}^{(l-1)} + \sigma_{b,l}^2\, \mathrm{diag}(1, 0). \tag{179}$$

In the zero-bias argument, Lemma D.8 reduces a quantity of unit set l to the same quantity of unit set $l - 1$. Conditioning on layers $1, \dots, l-1$ and on $n \geq 2$ active units, the lemma

gives $\mathbb{E}[r(\mathbf{S}^{(l)}) \mid \cdots] = r(\mathbf{C}^{(l)})$. With zero bias, $\mathbf{C}^{(l)} = \frac{\sigma_l^2}{p_{l-1}}\mathbf{S}^{(l-1)}$, and r is invariant under multiplication by a positive number, so $r(\mathbf{C}^{(l)}) = r(\mathbf{S}^{(l-1)})$. Each application of the lemma therefore lowers the index of the argument of r by one, until it reaches the deterministic matrix $\mathbf{S}^{(0)}$.

With bias, the bias term in (179) is added only to the upper-left entry, and the output bias adds $2\sigma_{b,L}^2$ to the upper-left entry of $\mathbf{S}$ in the same way. With zero bias, the upper-left and lower-right entries of $\mathbf{S}$ come from the same random weights and are scaled together, and r is invariant under a common scale factor. A bias adds a constant without fluctuation to the upper-left entry, which reduces the relative fluctuation of that entry, while the relative fluctuation of the lower-right entry stays the same. The two entries then fluctuate by different amounts, $r(\mathbf{C}^{(l)})$ fluctuates as a consequence, and its expectation over layers $1, \ldots, l-1$ is again the expectation of a nonlinear function of a random matrix.

We treat this expectation by an expansion. At large width, $\mathbf{S}$ has small relative fluctuations around its expectation. Expanding r to second order around the expectation, δ is determined by the variances and covariances of the relative fluctuations of the entries of $\mathbf{S}$. These quantities can be computed layer by layer. Each layer produces new fluctuations. Of the fluctuation coming from the unit sets below, the part in the lower-right entry is kept in full, and the part in the upper-left entry is reduced by the bias, by a factor set by the share of $q^{(l)}$ that comes from unit set $l-1$. We first define the quantities that describe these shares and fluctuations. Let

$$\chi_l \;:=\; \frac{\sigma_l^2\, q^{(l-1)}}{2\, q^{(l)}} \quad (2 \le l \le L), \tag{180}$$

$$\omega_l \;:=\; \chi_{l+1}\chi_{l+2}\cdots\chi_L \;=\; \frac{q^{(l)}}{q^{(L)}}\,\frac{\sigma_{l+1}^2}{2}\,\frac{\sigma_{l+2}^2}{2}\cdots\frac{\sigma_L^2}{2} \quad (1 \le l \le L-1),$$

where the second equality substitutes the definitions of $\chi_{l+1}, \ldots, \chi_L$, and the factors $q^{(l+1)}, \ldots, q^{(L-1)}$ in the numerator cancel the same factors in the denominator. Thus χ_l is the fraction of $q^{(l)}$ that comes from unit set $l-1$ rather than from the bias of layer l, and ω_l is the fraction of $q^{(L)}$ that comes from unit set l. For $l = 0, \ldots, L-1$, let

$$V_{11}^{(l)} := \frac{\mathrm{Var}(\mathbf{S}_{11}^{(l)})}{(\mathbb{E}\,\mathbf{S}_{11}^{(l)})^2}, \qquad V_{22}^{(l)} := \frac{\mathrm{Var}(\mathbf{S}_{22}^{(l)})}{(\mathbb{E}\,\mathbf{S}_{22}^{(l)})^2}, \tag{181}$$

$$V_{\times}^{(l)} := \frac{\mathrm{Cov}(\mathbf{S}_{11}^{(l)}, \mathbf{S}_{22}^{(l)})}{\mathbb{E}\,\mathbf{S}_{11}^{(l)}\;\mathbb{E}\,\mathbf{S}_{22}^{(l)}}, \qquad V_{12}^{(l)} := \frac{\mathbb{E}[(\mathbf{S}_{12}^{(l)})^2]}{\mathbb{E}\,\mathbf{S}_{11}^{(l)}\;\mathbb{E}\,\mathbf{S}_{22}^{(l)}}. \tag{182}$$

These are normalized fourth moments of $(h_i^{(l)}, h_i^{(l)\prime})$.

Proposition D.10 (Nonzero bias, first order)**.** *For Gaussian biases drawn as in* (3)*, to first order in* $1/p_{\min}$*,*

$$\delta \;=\; \sum_{l=1}^{L-1} \frac{c_l}{p_l}, \qquad c_l \;=\; \tfrac{5}{8}\,(3\omega_l + 1)(\omega_l - 1) \;\in\; \big[-\tfrac{5}{6},\, 0\big]. \tag{183}$$

Proof. By (162), the three entries of $\mathbf{S}$ are

$$S_{11} = 2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}}\mathbf{S}_{11}^{(L-1)}, \qquad S_{22} = \frac{2\sigma_L^2}{p_{L-1}}\mathbf{S}_{22}^{(L-1)}, \qquad S_{12} = \frac{2\sigma_L^2}{p_{L-1}}\mathbf{S}_{12}^{(L-1)}. \tag{184}$$

Let

$$\epsilon_1 := \frac{\mathbf{S}_{11}^{(L-1)}}{\mathbb{E}\,\mathbf{S}_{11}^{(L-1)}} - 1, \qquad \epsilon_2 := \frac{\mathbf{S}_{22}^{(L-1)}}{\mathbb{E}\,\mathbf{S}_{22}^{(L-1)}} - 1, \qquad \epsilon_{12} := \frac{\mathbf{S}_{12}^{(L-1)}}{(\mathbb{E}\,\mathbf{S}_{11}^{(L-1)}\;\mathbb{E}\,\mathbf{S}_{22}^{(L-1)})^{1/2}}. \tag{185}$$

All three are of order $p_{\min}^{-1/2}$, because, conditioning on layers $1, \ldots, L-2$, every entry of $\mathbf{S}^{(L-1)}$ is a sum of p_{L-1} independent terms. By $\mathbb{E}[h_j^2] = \frac{1}{2}q^{(L-1)}$ from the proof of Lemma D.4,

$\mathbb{E}\,\mathbf{S}_{11}^{(L-1)} = \sum_{j=1}^{p_{L-1}} \mathbb{E}[h_j^2] = \frac{1}{2} p_{L-1} q^{(L-1)}$, so $\frac{2\sigma_L^2}{p_{L-1}}\,\mathbb{E}\,\mathbf{S}_{11}^{(L-1)} = \sigma_L^2 q^{(L-1)}$. Substituting $\mathbb{E}\,S_{11} = 2q^{(L)}$ and (180) gives

$$\frac{\frac{2\sigma_L^2}{p_{L-1}}\,\mathbb{E}\,\mathbf{S}_{11}^{(L-1)}}{\mathbb{E}\,S_{11}} = \frac{\sigma_L^2 q^{(L-1)}}{2q^{(L)}} = \chi_L, \tag{186}$$

that is, χ_L is the share of $\mathbb{E}\,S_{11}$ contributed by the hidden units. Substituting $\mathbf{S}_{11}^{(L-1)} = \mathbb{E}\,\mathbf{S}_{11}^{(L-1)}(1+\epsilon_1)$ and $\mathbb{E}\,S_{11} = 2\sigma_{b,L}^2 + \frac{2\sigma_L^2}{p_{L-1}}\,\mathbb{E}\,\mathbf{S}_{11}^{(L-1)}$,

$$S_{11} = \mathbb{E}\,S_{11} + \frac{2\sigma_L^2}{p_{L-1}}\,\mathbb{E}\,\mathbf{S}_{11}^{(L-1)}\,\epsilon_1 = \mathbb{E}\,S_{11}(1+\chi_L\epsilon_1). \tag{187}$$

In S_{22}, the factor $2\sigma_L^2/p_{L-1}$ appears in both S_{22} and $\mathbb{E}\,S_{22}$ and cancels, so $S_{22}/\,\mathbb{E}\,S_{22} = \mathbf{S}_{22}^{(L-1)}/\,\mathbb{E}\,\mathbf{S}_{22}^{(L-1)} = 1+\epsilon_2$. For S_{12}, substituting the three expressions above,

$$\frac{S_{12}^2}{S_{11}S_{22}} = \frac{\left(\frac{2\sigma_L^2}{p_{L-1}}\right)^2 (\mathbf{S}_{12}^{(L-1)})^2}{\mathbb{E}\,S_{11}(1+\chi_L\epsilon_1)\,\frac{2\sigma_L^2}{p_{L-1}}\,\mathbb{E}\,\mathbf{S}_{22}^{(L-1)}(1+\epsilon_2)} \tag{188}$$

$$= \frac{\frac{2\sigma_L^2}{p_{L-1}}\,\mathbb{E}\,\mathbf{S}_{11}^{(L-1)}}{\mathbb{E}\,S_{11}}\,\frac{\epsilon_{12}^2}{(1+\chi_L\epsilon_1)(1+\epsilon_2)} = \chi_L\epsilon_{12}^2\big(1+O(p_{\min}^{-1/2})\big). \tag{189}$$

By Lemma D.4, $\mathbb{E}\,S_{12} = 0$, so $r(\mathbb{E}\,\mathbf{S}) = (\mathbb{E}\,S_{22}/\,\mathbb{E}\,S_{11})^{1/2}$, and (116) gives $r(\mathbf{S}) = (S_{22}/S_{11})^{1/2}(1 - S_{12}^2/(S_{11}S_{22}))^{1/2}$. Dividing the two and substituting the results above,

$$\frac{r(\mathbf{S})}{r(\mathbb{E}\,\mathbf{S})} \;=\; (1+\epsilon_2)^{1/2}\,(1+\chi_L\epsilon_1)^{-1/2}\,\big(1-\chi_L\epsilon_{12}^2(1+O(p_{\min}^{-1/2}))\big)^{1/2}. \tag{190}$$

To second order, the three factors are $1+\frac{1}{2}\epsilon_2 - \frac{1}{8}\epsilon_2^2$, $1-\frac{\chi_L}{2}\epsilon_1 + \frac{3\chi_L^2}{8}\epsilon_1^2$, and $1-\frac{\chi_L}{2}\epsilon_{12}^2$, each with error $O(p_{\min}^{-3/2})$. Multiplying them and keeping the second-order terms,

$$\frac{r(\mathbf{S})}{r(\mathbb{E}\,\mathbf{S})} = 1 + \frac{1}{2}\epsilon_2 - \frac{\chi_L}{2}\epsilon_1 - \frac{1}{8}\epsilon_2^2 + \frac{3\chi_L^2}{8}\epsilon_1^2 - \frac{\chi_L}{4}\epsilon_1\epsilon_2 - \frac{\chi_L}{2}\epsilon_{12}^2 + O(p_{\min}^{-3/2}). \tag{191}$$

By the definitions of $\epsilon_1, \epsilon_2, \epsilon_{12}$, we have $\mathbb{E}\,\epsilon_1 = \mathbb{E}\,\epsilon_2 = 0$, $\mathbb{E}\,\epsilon_1^2 = V_{11}^{(L-1)}$, $\mathbb{E}\,\epsilon_2^2 = V_{22}^{(L-1)}$, $\mathbb{E}[\epsilon_1\epsilon_2] = V_{\times}^{(L-1)}$, and $\mathbb{E}\,\epsilon_{12}^2 = V_{12}^{(L-1)}$. Taking the expectation of the expansion above and using $\delta = \mathbb{E}[r(\mathbf{S})]/r(\mathbb{E}\,\mathbf{S}) - 1$ from (156), we obtain

$$\delta \;=\; \frac{3\chi_L^2}{8}V_{11}^{(L-1)} - \frac{1}{8}V_{22}^{(L-1)} - \frac{\chi_L}{4}V_{\times}^{(L-1)} - \frac{\chi_L}{2}V_{12}^{(L-1)}. \tag{192}$$

The quantities $V^{(l)}$ satisfy a recursion across layers. Conditioning on layers $1,\dots,l-1$, by (163), the pre-activation pair $(y_i^{(l)}, y_i^{(l)\prime})$ of unit i of unit set l is a centered Gaussian vector whose two coordinates are correlated only at relative order $p_{\min}^{-1/2}$, and this changes the moments below by $O(p_{\min}^{-2})$. The contributions of this unit to $\mathbf{S}_{11}^{(l)}$, $\mathbf{S}_{22}^{(l)}$, and $\mathbf{S}_{12}^{(l)}$ are $(h_i^{(l)})^2 = ((y_i^{(l)})^+)^2$, $(h_i^{(l)\prime})^2 = (y_i^{(l)\prime})^2\mathbf{1}\{y_i^{(l)} > 0\}$, and $h_i^{(l)}h_i^{(l)\prime} = (y_i^{(l)})^+ y_i^{(l)\prime}$. We normalize the first two by their conditional means $\frac{1}{2}\,\mathbb{E}(y_i^{(l)})^2$ and $\frac{1}{2}\,\mathbb{E}(y_i^{(l)\prime})^2$, and the third by the geometric mean of these two means. The half-normal moment $\mathbb{E}[((y_i^{(l)})^+)^4] = \frac{3}{2}(\mathbb{E}(y_i^{(l)})^2)^2$ then gives relative variance 5 for each of the first two, relative covariance 1 between them, and relative second moment 2 for the third. Summing over the p_l independent units adds these constants divided by p_l. The variance of the conditional means adds the values from the previous unit set. Because the bias enters only the first coordinate, the previous value is multiplied by χ_l^2 when it passes to $V_{11}^{(l)}$, by χ_l when it passes to $V_{\times}^{(l)}$ and $V_{12}^{(l)}$, and by 1 when it passes to $V_{22}^{(l)}$. Starting from zero at $l=0$, where the circle is deterministic,

$$\begin{aligned} V_{11}^{(l)} &= \frac{5}{p_l} + \chi_l^2 V_{11}^{(l-1)}, & V_{22}^{(l)} &= \frac{5}{p_l} + V_{22}^{(l-1)}, \\ V_{\times}^{(l)} &= \frac{1}{p_l} + \chi_l V_{\times}^{(l-1)}, & V_{12}^{(l)} &= \frac{2}{p_l} + \chi_l V_{12}^{(l-1)}, \end{aligned} \tag{193}$$

to first order in $1/p_{\min}$. At $l = 1$, χ_1 multiplies $V^{(0)} = 0$, so its value does not enter. Substituting the solution into (192), unit set l contributes c_l/p_l with $c_l = \frac{15}{8}\omega_l^2 - \frac{5}{4}\omega_l - \frac{5}{8} = \frac{5}{8}(3\omega_l + 1)(\omega_l - 1)$. On $[0, 1]$, this quadratic equals 0 at $\omega_l = 1$, equals $-\frac{5}{8}$ at $\omega_l = 0$, and attains its minimum $-\frac{5}{6}$ at $\omega_l = \frac{1}{3}$. □

With zero bias, the part of the random fluctuation that increases crossings and the part that decreases them cancel exactly. A bias adds to each small ellipse a shift that does not scale with the unit sets below, and the output bias adds the fixed term $2\sigma_{b,L}^2 \operatorname{diag}(1, 0)$ to $\mathbf{S}$. Both reduce the part of the fluctuation that increases crossings, so the increase is smaller than the decrease, and the net result is fewer crossings.

The coefficient c_l is zero exactly when $\omega_l = 1$, that is, when layers $l + 1, \ldots, L$ have no bias. To first order, finite width only lowers the crossing density, and the relative decrease is at most $\frac{5}{6}\sum_l 1/p_l$. For example, take $L = 8$, $p_0 = 50$, $p_l \equiv 128$, $R = 1$, and $\sigma_l^2 = 2$, $\sigma_{b,l}^2 = 0.5$ in every layer, including the output layer. Then $\omega_l = q^{(l)}/q^{(L)}$, $\sum_l c_l = -4.64$, and $\delta \approx -4.64/128 = -3.6\%$.

Figure 1(c) compares these results with Monte Carlo estimates for $p_0 = 50$, $\sigma_l^2 = 2$, $R = 1$ and all hidden widths equal to p, at $L = 4$ (dashed, open markers) and $L = 8$ (solid, filled markers). Without biases (black), the estimates from up to 5×10^8 networks per point follow the exact value (171), which decays exponentially in p. With $\sigma_{b,l}^2 = 0.5$ in every layer (green), the estimates from 8×10^4 to 1.2×10^6 networks per point follow the first-order value $\sum_l c_l/p$ and decay as $1/p$, with corrections of order $1/p^2$ visible at $p \le 16$ for $L = 8$. In every case $\delta \le 0$, and at fixed width $|\delta|$ grows roughly linearly in the number $L - 1$ of hidden unit sets. The shaded band marks $|\delta| < 1\%$.

D.5 A single network

Replacing Γ_L in ρ_L by the generator energy Z_L of one network gives the estimate $\hat{\rho}_L$ computed from that network. The ratio of the two is $\sqrt{Z_L/\Gamma_L}$, so the fluctuation of $\hat{\rho}_L$ is determined by the relative variance of Z_L, which is computed in Appendix C. The definitions and details follow.

Let $\gamma(s)$ lie in the cell $\mathcal{C}$ of Section 3, and let Z_L be the generator energy (9), that is, the generator energy of one network at the input $\gamma(s)$, so $\Gamma_L = \mathbb{E}[Z_L]$. Let

$$\hat{\rho}_L \;:=\; \kappa(p_0)\sqrt{\frac{Z_L}{q^{(L)}}} \tag{194}$$

be the estimate computed from the generators of that network. By Corollary C.2, if all widths of unit sets $1, \ldots, L-1$ equal p and p_0 is fixed, then as $L/p \to 0$, $\hat{\rho}_L/\rho_L = \sqrt{Z_L/\Gamma_L} \to 1$ in probability. In this regime, the numerator of ρ_L equals the generator energy of the given network with high probability, and the denominator $q^{(L)}$ is deterministic. The difference between ρ_L and the crossing density over the population of networks is the δ of §D.4.

The relative variance of S_{22} grows with depth in the same way as the generator energy. Conditioning on layers $1, \ldots, l-1$, the contribution of unit i of unit set l to $\mathbf{S}_{22}^{(l)}$ is $(h_i^{(l)\prime})^2 = (y_i^{(l)\prime})^2 \mathbf{1}\{y_i^{(l)} > 0\}$, where $y_i^{(l)\prime} = \langle \boldsymbol{\theta}, \boldsymbol{h}^{(l-1)\prime} \rangle$ and $\boldsymbol{\theta} = \mathbf{W}_{i,:}^{(l)\top}$. Taking $\varphi = \langle \boldsymbol{\theta}, \boldsymbol{h}^{(l-1)\prime} \rangle^2$ and $\varphi = \langle \boldsymbol{\theta}, \boldsymbol{h}^{(l-1)\prime} \rangle^4$ in Lemma D.5 shows that, for any correlation between $y_i^{(l)}$ and $y_i^{(l)\prime}$, the first and second moments of this contribution are $\frac{1}{2}\,\mathbb{E}(y_i^{(l)\prime})^2$ and $\frac{3}{2}(\mathbb{E}(y_i^{(l)\prime})^2)^2$. Hence

$$1 + V_{22}^{(L-1)} \;=\; \prod_{l=1}^{L-1} \Big(1 + \frac{5}{p_l}\Big) \tag{195}$$

holds exactly. The factor $(p_l^2 + 5p_l)/4$ behind each factor is the multiple by which the moment recursion of Appendix C acts along its direction $P = Q$, and $1 + 5/p$ is the per-layer growth factor in Corollary C.2. In both cases, the constant 5 is the value of $\operatorname{Var}((h_i^{(l)})^2)/(\mathbb{E}(h_i^{(l)})^2)^2$ when $y_i^{(l)}$ is a centered Gaussian.

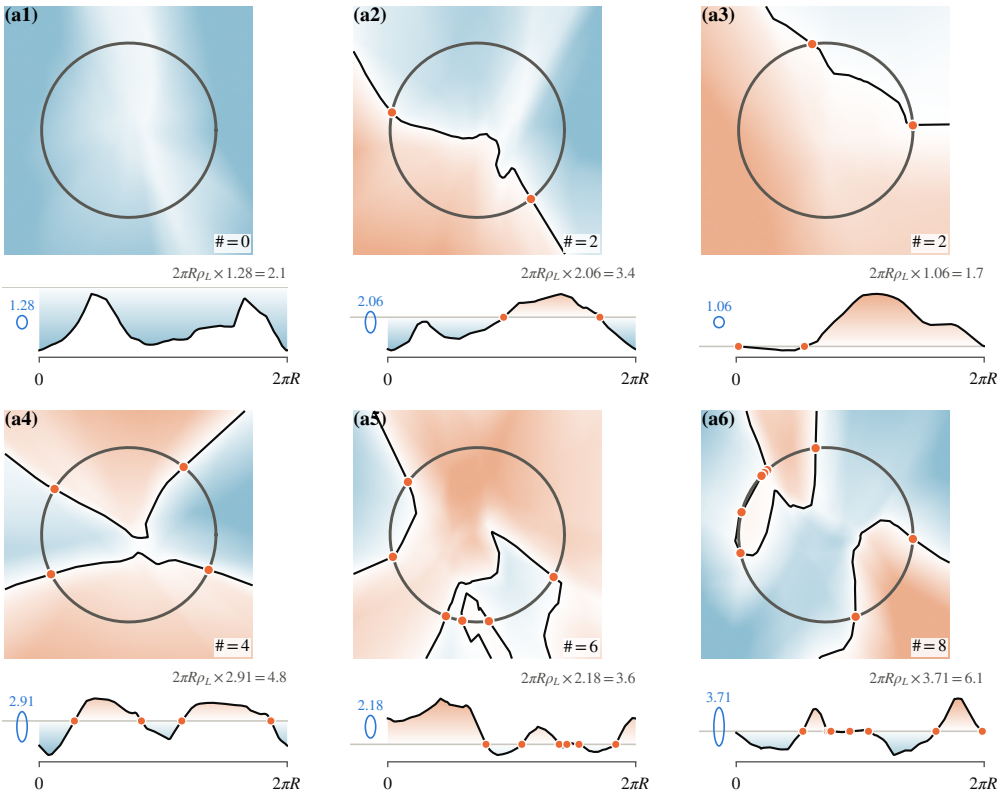


Figure 9: Six networks with $p_0 = 2$, $L = 4$, hidden widths 16, $\sigma_w^2 = 2$ and $\sigma_b^2 = 0.5$, on a circle γ of radius $R = 2$. Panel (a4) is the network of Figure 1(a). The other panels are the first networks with 0, 2, 2, 6 and 8 crossings in a scan of 2000 random seeds. Top: the regions $g > 0$ (peach) and $g < 0$ (blue), white at $g = 0$, with the colour saturated at the maximum of $|g|$ on γ. The boundary $\{g = 0\}$ (black) cuts the circle (grey) at the crossings (orange), and # is their number. Inside a linear cell g is affine, so the white band along the boundary is narrow where g changes fast. Bottom: $g(\gamma(s))$ along the circle, starting at $(R, 0)$ and running counterclockwise, shaded in the same way. The blue glyph is the ellipse of the second moments of $(g, \partial_s g)$ along this circle, drawn at a common scale, and the number beside it is its height x. The line above the strip is the prediction $2\pi R\rho_L\, x$ of the number of crossings.

Figure 9 shows six networks of the configuration of Figure 1(a). For one network and one circle, let $\mathbf{C}$ be the mean of $(g, \partial_s g)^\top (g, \partial_s g)$ over the circle, and let $x := r(\mathbf{C})/r(\mathbb{E}_{\mathcal{F}}[\mathbf{S}])$. This is the height of the ellipse of $\mathbf{C}$ in the units of Figure 8, and $2\pi R\rho_L\, x$ is the number of crossings that Rice's formula gives for a path with these second moments. The prediction increases with the observed count, which ranges from 0 to 8, and the two differ by at most about two crossings, because one circle is a short sample of the path.

D.6 Crossing density and boundary volume

This subsection converts the crossing density into the size of the decision boundary itself. Let $\sigma_{b,L} > 0$. With ρ_L in place of ρ_{exact}, the expected volume of the decision boundary on the sphere S_R, as a fraction of the area of the sphere, is

$$\pi\, \frac{|\mathbb{S}^{p_0-2}|}{|\mathbb{S}^{p_0-1}|}\, \kappa(p_0) \sqrt{\frac{\Gamma_L}{q^{(L)}}}, \tag{196}$$

with relative error δ. Thus the quantity $\Gamma_L/q^{(L)}$ in (19) determines how large the decision boundary on the sphere is. The derivation integrates the number of crossings over all great

circles. Every great circle has the same expected crossing density, so the integral contributes a constant that depends on p_0.

Proposition D.11 (Boundary volume)**.** *Let $\sigma_{b,L} > 0$, let $\mathcal{H}^{p_0-1}$ and $\mathcal{H}^{p_0-2}$ denote the (p_0-1)-dimensional and (p_0-2)-dimensional Hausdorff measures, and let $|\mathbb{S}^{p_0-1}|$ and $|\mathbb{S}^{p_0-2}|$ be the areas of the unit spheres in $\mathbb{R}^{p_0}$ and $\mathbb{R}^{p_0-1}$. Then*

$$\frac{\mathbb{E}\left[\mathcal{H}^{p_0-2}(\{g=0\}\cap S_R)\right]}{\mathcal{H}^{p_0-1}(S_R)} \;=\; \pi\,\frac{|\mathbb{S}^{p_0-2}|}{|\mathbb{S}^{p_0-1}|}\,\rho_{\text{exact}}. \tag{197}$$

For $p_0 = 2$ the constant equals 1.

Proof. Given $\mathcal{F}$, there are finitely many linear regions, and g is affine on each region. An affine function vanishes on an open set only if it is identically zero. On each region, the constant term of g equals Δ_0 plus a term independent of it, and Δ_0 has a density, so with probability one g is not identically zero on any region. Since g is piecewise affine with finitely many pieces, $M := \{g=0\}\cap S_R$ is a finite union of (p_0-2)-dimensional spherical patches, and almost every great circle meets it transversally, with the intersection points being the crossings. Let μ be the rotation-invariant probability measure on the set of great circles, and let Λ denote a great circle. The spherical Crofton formula on S_R [22], with its constant fixed by a great (p_0-2)-dimensional sphere (which almost every great circle meets twice), reads

$$\int \#(M\cap\Lambda)\,d\mu(\Lambda) \;=\; \frac{2\,\mathcal{H}^{p_0-2}(M)}{R^{p_0-2}\,|\mathbb{S}^{p_0-2}|}. \tag{198}$$

Each great circle has length $2\pi R$, and by Proposition D.3 its expected number of crossings is $2\pi R\,\rho_{\text{exact}}$. Taking the expectation and exchanging it with the integral over μ, which is permitted because the integrand is nonnegative, gives $\mathbb{E}\,\mathcal{H}^{p_0-2}(M) = \pi R^{p_0-1}|\mathbb{S}^{p_0-2}|\,\rho_{\text{exact}}$, and $\mathcal{H}^{p_0-1}(S_R) = R^{p_0-1}|\mathbb{S}^{p_0-1}|$. □

The assumption $\sigma_{b,L} > 0$ is needed. When all biases are zero, with positive probability every unit of unit set $L-1$ is inactive on some open subset of the sphere. Then g is zero there, $\mathcal{H}^{p_0-2}(\{g=0\}\cap S_R)$ is infinite with positive probability, and the left side of (197) is infinite, while by Proposition D.9, ρ_{exact} is finite.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have been careful to ensure that the claims made in the abstract and introduction accurately reflect the contributions and scope of our paper. We have proofread the abstract and introduction multiple times to remove overclaims and inaccurate statements. The current version of the abstract and introduction reflects the contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations has been discussed, and the constrains are clearly stated in the technical derivations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The

authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We either provide a rigid proof or tried our best to list necessary assumptions for each claim.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all necessary details for reproducing the experimental results. The code will be made available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code will be made available. No private data is used in this paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiment settings will be released with the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: When applicable, experiments are conducted with multiple random seeds to ensure the statistical significance; error bars are reported accordingly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The experiments are of small size, runnable on a consumer laptop.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the principles outlined in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work performed has no direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve data or models that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper uses commonly available assets in the public domain, we acknowledge all relevant sources according to academic norms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used to assist with the mathematical derivations, the writing, and the implementation of the experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.