# RENORMALIZED ENTROPY FROM SYMMETRIES: SIMPLICITY BIAS AND NEURAL HEAT DEATH

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

## ABSTRACT

Neural networks have parameter symmetries: parameter transformations that do not affect the represented function. How do these symmetries affect the functions that neural networks ultimately learn? We study the implicit bias arising from these symmetries by calculating the equilibrium distribution over represented functions under Langevin dynamics with Gaussian regularization. To isolate the effect of symmetries on this distribution, we study the limit in which the strength of the regularization is taken to zero. In the low temperature limit, we recover known effective regularizations. At positive temperature, a renormalization procedure allows us to derive a novel entropy which predicts symmetry-induced preference for certain functions. For linear networks, we obtain explicit entropy terms that favor singular and low-rank functions. For networks with homogeneous activations, we find entropy terms that favor inactive neurons, which can lead to neural "heat death" in which equilibrium mass concentrates on networks with vanishing output. We interpret these results through the neural tangent kernel, as a simplicity bias selecting for functions that vary slowly under variations in the parameters.

## 1 INTRODUCTION

Neural networks generally exhibit *parameter symmetry*, meaning there are directions in parameter space which do not affect the input-output map represented by the network. Examples include rescaling in homogeneous neural networks, and changes of bases for matrix products in linear networks and attention, among others (Saxe et al., 2014; Kunin et al., 2021; Zhao et al., 2025). How does training (which occurs in parameter space) affect what functions are actually learned?

In general, we can describe the situation as follows. Let $\theta \in \Theta \subseteq \mathbb{R}^N$ be the neural network parameters and $f = \pi(\theta)$ be the represented function. Assume $\theta$ is optimized to minimize the loss $L(\theta)$. Since the loss should only depend on $f$, we write $L(\theta) = \ell(f)$. Optimization in parameter space induces non-trivial function space dynamics, for example for gradient flow (with rate $\eta > 0$)

$$\dot{\theta} = -\eta \nabla L(\theta) \qquad \Longrightarrow \qquad \dot{f} = -\eta D\pi_\theta D\pi_\theta^\top \nabla \ell(f).$$

Here, $D\pi_\theta$ denotes the Jacobian of $\pi$ at $\theta$, and $D\pi_\theta D\pi_\theta^\top$ is the neural tangent kernel (Jacot et al., 2018). The difficulty in analyzing implicit bias in $f$ dynamically is that the NTK evolves over time, and in general depends on $\theta$ in complicated ways. In certain cases this dependence can be simplified, as characterized by Marcotte et al. (2025), or in various limits using sophisticated techniques to find order parameters (Bordelon & Pehlevan, 2022).

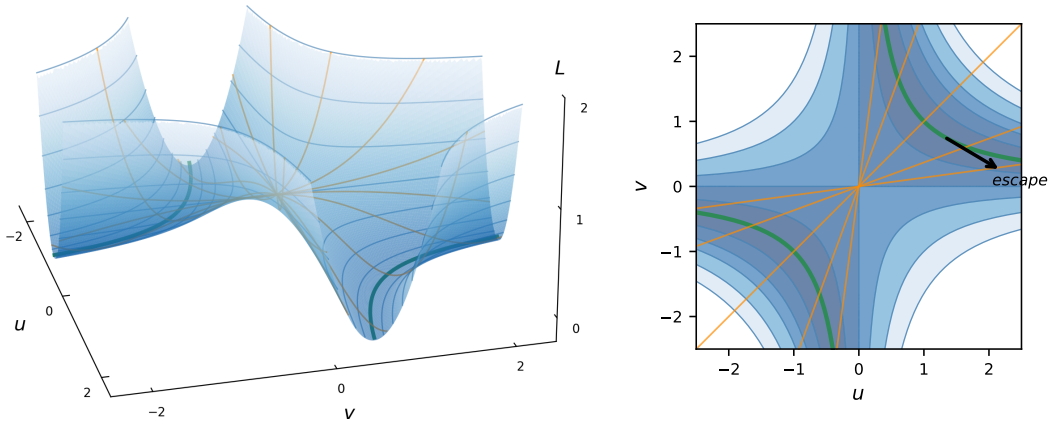
We consider a different approach, looking at the steady state distribution $p(\theta)$ of parameters subjected to noisy training dynamics. This allows us to abstract away the time course of the dynamics and study the steady state distribution they induce over functions,

$$\text{steady state} = p(\theta) \qquad \Longrightarrow \qquad \Pr(f \in B) = \int_{\pi^{-1}(B)} p(\theta)\, d\theta.$$

Neural network parameter steady state distributions exist in practice, though for SGD they are generally non-equilibrium (having probability currents) (Mandt et al., 2017; Chaudhari & Soatto, 2018; Yaida, 2019; Zhang et al., 2022), and may be complicated to write down analytically. A principled but tractable alternative is provided by Langevin dynamics, which can be thought of as gradient



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




(a) Surface view (b) Contour view

Figure 1: The loss landscape of $f(x) = uvx$ with $L(u, v) = (uv - 1)^2/2$, plotted with the coordinates $w = uv$ (blue) and $a$ (orange, defined in Section 3). The loss and $w$ have scaling symmetry $u \mapsto cu$ and $v \mapsto v/c$. Under Langevin dynamics, $\theta$ escapes to infinity along $uv = 1$ (green).

flow perturbed by white noise. This is a classical model for studying Bayesian inference in neural networks (Welling & Teh, 2011; Mandt et al., 2017) as well as the behavior of biological neural systems (Kappel et al., 2015; Aitchison et al., 2021), and allows us to make concrete predictions.

A central obstacle in our proposal is that the equilibrium distribution only exists when the loss landscape $L(\theta)$ prevents $\theta$ from diffusing "to infinity." Common neural network symmetries include scaling, which is "infinite" (technically, *non-compact*): you can always scale a number to make it bigger. Since scaling is a symmetry of the minimum-loss set, $\theta$ can escape along this set, meaning no stationary distribution exists (Figure 1). To circumvent this we consider squared-norm regularization (a choice we argue is well suited to Langevin, see Section 3) with strength $\lambda > 0$,

$$L(\theta) + \lambda R(\theta), \qquad R(\theta) = \frac{1}{2}\|\theta\|^2.$$

The regularization "pulls up" the loss landscape further from the origin, eliminating the possibility of escape and ensuring a stationary distribution exists. Such distributions are studied as Bayesian priors over functions (Li & Sompolinsky, 2021; Izmailov et al., 2021; Zavatone-Veth & Pehlevan, 2021; Noci et al., 2021; Kobialka et al., 2026). Regularization, however, necessarily breaks the non-compact symmetries whose effect we wish to study. We therefore study the limit in which $\lambda$ is taken to zero. If the temperature in Langevin dynamics (defined below) is taken to zero first, we recover predictions from the representation cost literature (Srebro et al., 2004; Gunasekar et al., 2017; Dai et al., 2021), corresponding to selecting $f$ which minimizes $R$ subject to minimizing $\ell$. If temperature is instead fixed to be positive, using a renormalization procedure, we obtain a novel entropy-driven selection principle.

Our contributions in this paper are:

- a conceptually simple but powerful framework for analyzing parameter symmetry, recovering known implicit biases in one limit, and a novel renormalized entropy in another;
- theoretical derivations of the renormalized entropy for diagonal, linear, and homogeneous networks, including the dominance of the zero-function for ReLU networks trained on radially coercive losses, which we call *heat death*;
- a heuristic connection between the predictions of renormalized entropy and the long-time limit of unregularized Langevin dynamics, supported by experimental agreement between theory and simulation;
- an interpretation of the entropy as selecting simple functions, with low NTK volume.

To the best of our knowledge, our work contains the first derivation of entropic implicit bias from the renormalized volume growth of non-compact symmetry fibers.

## 2 THEORETICAL FRAMEWORK

**Langevin dynamics** Following the introduction, let $\theta \in \Theta \subseteq \mathbb{R}^N$ be neural network parameters, $f = \pi(\theta)$ and $L(\theta) = \ell(f)$. For a potential $U(\theta)$ and temperature $T > 0$, Langevin dynamics is

$$d\theta_t = -\nabla U(\theta_t)\, dt + \sqrt{2T} dB_t$$

where $B_t$ is $N$-dimensional Brownian motion. Under confinement and normalizability conditions such as, (1) $\langle \theta, \nabla U(\theta) \rangle \to \infty$ as $\|\theta\| \to \infty$, and (2) $\int \exp(-U(\theta)/T)\, d\theta$ is finite, there exists the Gibbs stationary distribution $p(\theta)\, d\theta \propto \exp(-U(\theta)/T)\, d\theta$. This is the limiting distribution of $\theta_t$; that is if $p_t(\theta)$ is the probability density of $\theta_t$, then $p(\theta) = \lim_{t\to\infty} p_t(\theta)$. For losses $L(\theta)$ such as those shown in Figure 1, the normalization integral is not finite, so we let $U(\theta) = L(\theta) + \lambda R(\theta)$ and write the stationary distribution

$$p_{\lambda,T}(\theta)\, d\theta = \frac{e^{-\frac{1}{T}(L(\theta)+\lambda R(\theta))}}{\int e^{-\frac{1}{T}(L(\tilde{\theta})+\lambda R(\tilde{\theta}))}\, d\tilde{\theta}}\, d\theta.$$

The set $\pi^{-1}(f) = \{\theta \in \Theta : f = \pi(\theta)\}$ of parameters implementing the same function is called the *fiber over* $f$. Since $L(\theta) = \ell(f)$, integrating over the fiber gives the induced distribution $f$

$$\Pr_{\lambda,T}(df) \propto e^{-\ell(f)/T} \nu_{\lambda/T}(df), \qquad \nu_\alpha(B) = \int_{\pi^{-1}(B)} e^{-\alpha R(\theta)}\, d\theta.$$

The probability distribution over $f$ is the Gibbs distribution for loss $\ell$, tilted by a factor $\nu_{\lambda/T}$ depending on the parameter symmetries encoded in $\pi$ — counting representations of $f$ exponentially down-weighted by how much they cost in regularization.[1] Studying this factor $\nu_{\lambda/T}$ for $\lambda \approx 0$ reveals the implicit bias of the parameterization. However, for parameterizations with fibers extending to infinity (such as in Figure 1) $\nu_{0/T} = \infty$, meaning we must perform a limiting analysis.

**Symmetry coordinates** On a *regular region* (or *stratum*) of parameter space, where the Jacobian $D\pi$ has constant rank, we can locally define *symmetry coordinates* $(f, a)$ (formally, a *chart*), where $a$ parametrizes (a local part of) the fiber $\pi^{-1}(f)$ (see Figure 2). Writing the Jacobian of the change of variables as $d\theta = |J(f,a)|\, df\, da$, on such a region, we define the *fiber partition function*

$$Z_\alpha(f) = \int e^{-\alpha R(f,a)} |J(f,a)|\, da,$$

where with an abuse of notation $R(f,a) = R(\theta)$. Then $\nu_\alpha(df) = Z_\alpha(f)\, df$, and we can write a density over functions,

$$p_{\lambda,T}(f) \propto e^{-\ell(f)/T} Z_{\lambda/T}(f).$$

**Zero-regularization limit** We consider two regimes in which to take $\lambda \to 0$. The simpler case is to take $T \to 0$ first. For the fiber free energy $F_\alpha(f) = -\alpha^{-1} \log Z_\alpha(f)$, Laplace's principle gives

$$\lim_{\alpha\to\infty} F_\alpha(f) = \inf_a R(f,a) =: R_{\text{eff}}(f), \qquad p_{\lambda,T}(f) \propto e^{-\frac{1}{T}(\ell(f)+\lambda F_{\lambda/T}(f))}.$$

We see that taking $T \to 0$ concentrates the distribution on the minimizers of $\ell(f)+\lambda R_{\text{eff}}(f)$. Taking $\lambda \to 0$ next selects the minimizers of $\ell(f)$ of smallest effective regularization $R_{\text{eff}}(f)$.

We also consider fixing $T > 0$ and taking $\lambda \to 0$ directly. If a fiber $\pi^{-1}(f)$ had finite volume $\Omega(f) = \int |J(f,a)|\, da$, then $Z_\alpha(f)$ would converge to said volume, and we would have an entropic correction, $p_{0,T}(f) \propto \exp(-\ell(f)/T + \log \Omega(f))$. In general, however, $Z_\alpha(f)$ diverges as $\alpha \to 0$. But suppose that on each regular region this divergence occurs at a well-defined rate, meaning

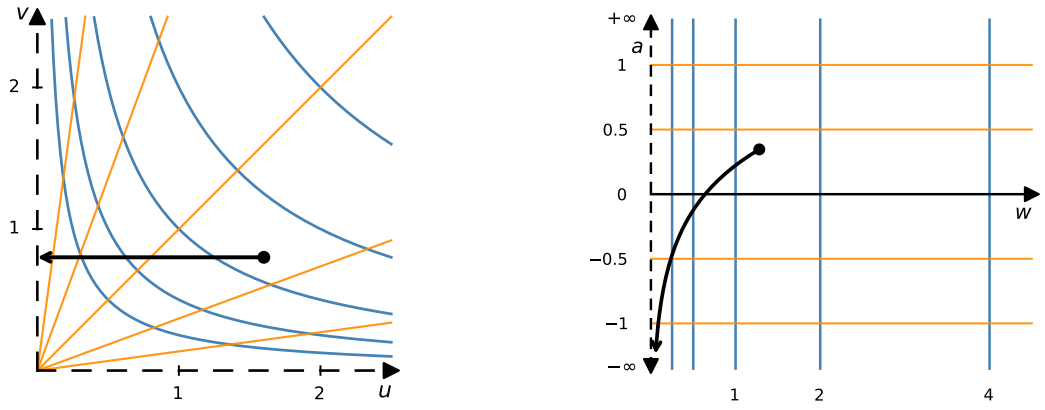
$$\Omega(f) = \lim_{\alpha\to 0} \frac{Z_\alpha(f)}{C_\alpha} \quad \text{exists for some } C_\alpha.$$

Then assuming the limit can be passed through the normalization integral over $f$, and no other stratum contributes, we get

$$\lim_{\lambda\to 0} p_{\lambda,T}(f) \propto e^{-\ell(f)/T + \log \Omega(f)}.$$

In this case, we call $\log \Omega(f)$ the *renormalized entropy*. Is is the relative volume of $\pi^{-1}(f)$ after removing a common divergence. Note that $\Omega(f)$ can depend on the choice of regularization. We argue below that squared-norm is natural here because it matches the Gaussian diffusion.

[1] Having $R(\theta) = \|\theta\|^2/2$ as the regularization, $\exp(-\alpha R(\theta))$ is proportional to a Gaussian prior of precision (inverse variance) $\alpha$. Thus $\nu_\alpha(df)$ is proportional to the distribution of $\pi(\alpha^{-1/2}\xi)$ for $\xi \sim \mathcal{N}(0, I_N)$.




(a) $u > 0, v > 0$ quadrant

(b) Corresponding symmetry coordinates

Figure 2: In the positive $(u, v)$ quadrant (excluding the axes), the diagonal network $f(x) = uvx$ can be given symmetry coordinates $w = uv$ (blue) and $a = \frac{1}{2}\log(u/v)$ (orange). These are ill-defined at the boundary, as shown by a path (black) approaching the $v$-axis diverging to $(0, -\infty)$ in the $(w, a)$ plane. The finite $a$-axis corresponds to approaching $(u, v) = (0, 0)$ along different rays.

## 3 Diagonal networks

As a simple example of the calculation, we begin with the simplest *diagonal network* Berthier (2023); Pesme & Flammarion (2023); Berthier (2026) $f(x) = uvx$ for a pair of scalar parameters $\theta = (u, v)$ and input $x$. We identify $f$ with the product $w = uv$, noting it has the scaling symmetry $u \mapsto cu, v \mapsto v/c$ for any $c \neq 0$ (see Figure 1 for an example loss landscape). On a fixed branch of the hyperbolic fiber, a convenient choice of symmetry coordinates on the set where $u \neq 0$ and $v \neq 0$ (which is regular) is $a = \frac{1}{2}\log|u/v|$, so that

$$u = e^a\sqrt{|w|}, \qquad v = e^{-a}\text{sign}(w)\sqrt{|w|}$$

as depicted in Figure 2. The Jacobian and regularization then take the convenient forms

$$d\theta = du\, dv = dw\, da, \qquad R(w, a) = \frac{u^2 + v^2}{2} = |w|\cosh(2a).$$

Ignoring the factor of two from summing over branches, the fiber partition function in this case is

$$Z_\alpha(w) = \int_{-\infty}^{\infty} e^{-\alpha|w|\cosh(2a)}\, da = K_0(\alpha|w|),$$

where $K_0$ is the modified Bessel function of the second kind. While its asymptotic expansions around $\alpha = \infty$ ($T \to 0$) and $\alpha = 0$ ($\lambda \to 0$) are classical (Abramowitz & Stegun, 1964), our framework gives simpler paths to eliminating the regularization strength $\lambda$ in $\alpha = \lambda/T$. Namely, taking $T \to 0$ first gives effective regularization

$$R_{\text{eff}}(w) = \inf_{a\in\mathbb{R}} |w|\cosh(2a) = |w|.$$

Suppose instead we fix $T > 0$ and take $\lambda \to 0$. For this and our subsequent calculations of the scaling of fiber partition functions, the idea is to determine how the regularization cuts off the non-compact symmetry directions. Here, the integrand $\exp(-\alpha|w|\cosh(2a))$ is close to one when $\alpha|w|\cosh(2a) \lesssim 1$, and decays rapidly outside this region. Since $\cosh(2a)$ grows exponentially with $|a|$, the region contributing to $Z_\alpha$ scales like $|a| \lesssim \frac{1}{2}\log\frac{1}{\alpha} + O_w(1)$. This suggests reparameterizing the fiber coordinate $a = \frac{1}{2}\log\left(\frac{1}{\alpha}\right)s$, such that the relevant region has finite size, Then

$$Z_\alpha(w) = \frac{\log(1/\alpha)}{2}\int_{-\infty}^{\infty} e^{-\frac{|w|}{2}(\alpha^{1-s}+\alpha^{1+s})}\, ds.$$

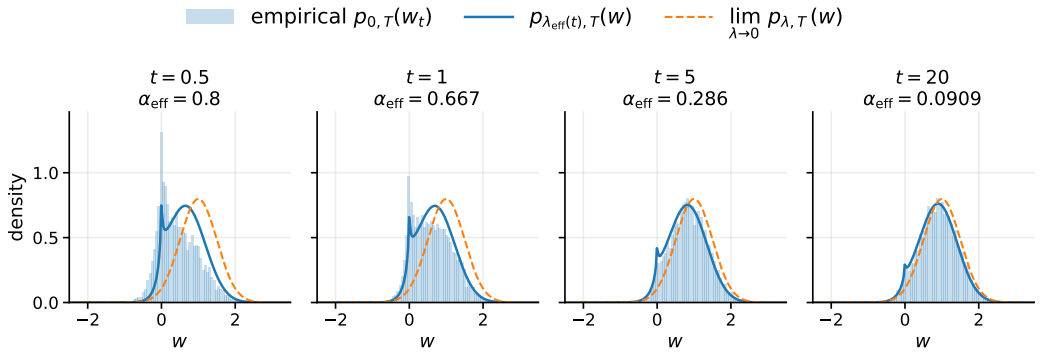



Figure 3: Langevin densities for $w = uv$ with $\ell(w) = (uv-1)^2/2$ and $T = 0.25$. Finite-time $\lambda = 0$ densities are well predicted by $p_{\lambda_{\text{eff}},T}(w)$ (blue), and approach the $\lambda = 0$ law (orange, dashed).

As $\alpha \to 0$, the $\alpha$ terms vanish for $|s| < 1$, making the integrand close to one, and at least one diverges for $|s| > 1$, making the integrand close to zero. Thus the expanding relevant region of the fiber in $a$-coordinates becomes $s \in [-1, 1]$. Letting $C_\alpha = \log(1/\alpha)$, we obtain that

$$\Omega(w) = \lim_{\alpha\to 0} \frac{Z_\alpha(w)}{C_\alpha} = \frac{1}{2}\int_{-1}^{1} ds = 1.$$

(The limit is justified in Appendix C.1.) For this simple example, the entropic term $\log \Omega(w) = 0$ is trivial, and the limiting distribution has no symmetry-induced bias: $\lim_{\lambda\to 0} p_{\lambda,T}(w) \propto \exp(-\ell(w)/T)$. It is worth pausing to make two observations. First, while we derive a limiting distribution for $w$ as $\lambda \to 0$, one cannot do the same in the original $(u, v)$ space, as probability escapes along the non-compact loss valley (see Appendix Figure 6). Second, the entropic correction happens to be trivial in this case, but there is no a priori reason for it to be so. In the models we examine later, the renormalized entropy retains a non-trivial dependence on the represented function.

**Connection to unregularized Langevin** The Bessel function prediction is exact for positive $\lambda$ and $T$, matching simulated Langevin statistics (see Appendix Figure 7). We additionally test whether the renormalized entropy is predictive of unregularized dynamics. In particular, it is not clear that we can interchange limits, so that the limiting distribution $\lim_{t\to\infty} p_{t,0,T}(w)$ (if it exists) corresponds to our limit

$$\lim_{\lambda\to 0} p_{\lambda,T}(w) = \lim_{\lambda\to 0}\lim_{t\to\infty} p_{t,\lambda,T}(w).$$

Surprisingly, we that these limits do agree empirically. We give a heuristic explanation below, which also justifies the choice of squared-norm regularization before taking the $\lambda \to 0$ limit.

Consider unregularized Langevin dynamics. Neglecting the drift $\nabla L(\theta_t)$, Itô's lemma gives $d\mathbb{E}(\theta_t)_i^2/dt \approx 2T$ in each coordinate $i$, and so $\mathbb{E}\|(\theta_t)_i\|^2 \approx s_0^2 + 2Tt$ for some initial scale $s_0^2 \geq 0$. Since in the regularized case $\nu_\alpha$ is the pushforward of a Gaussian prior on $\theta$ of precision $\alpha$, this motivates defining an effective precision in the unregularized case, $\alpha_{\text{eff}}(t) = (s_0^2 + 2Tt)^{-1}$. We find that the density under unregularized Langevin dynamics $p_{0,T}(w_t)$ is well described by $p_{\lambda_{\text{eff}}(t),T}(w)$ with $\lambda_{\text{eff}}(t) = \alpha_{\text{eff}}(t)T$ (Figure 3), the approximation becoming better as $t$ grows. But $\lambda_{\text{eff}}(t) \to 0$ as $t \to \infty$, suggesting the $\lambda \to 0$ limit indeed is predictive of unregularized Langevin,

$$\lim_{\lambda\to 0} p_{\lambda,T}(w) = \lim_{t\to\infty} p_{\lambda_{\text{eff}}(t),T}(w) \approx \lim_{t\to\infty} p_{t,0,T}(w).$$

**Diagonal networks of arbitrary width and depth** The story for general diagonal networks is much the same. A width $n$ and depth $D$ network is $f(x) = w^\top x$ where $w_i = \prod_{d=1}^{D} u_{di}$ with parameters $u_{di}$ for $d \in [D]$ and $i \in [n]$. The regularization $R = \frac{1}{2}\sum_{di} u_{di}^2$ decouples the coordinates $i$. The $T \to 0$ effective regularization recovers the known representation cost $R_{\text{eff}}(w) = \frac{D}{2}\sum_i |w_i|^{2/D}$ (Gunasekar et al., 2018; Dai et al., 2021), whereas the $\lambda \to 0$ limit gives $\Omega(w) = 1$ after renormalizing by $C_\alpha \sim (\log 1/\alpha)^{(D-1)n}$.

## 4 LINEAR NETWORKS

We encounter our first non-trivial entropic correction in linear networks. Let us begin by considering $f(x) = UV^\top x$ for $x \in \mathbb{R}^n$, and $U, V$ square matrices. We now identify $f$ with $W = UV^\top$, which has the symmetry $U \mapsto UC$ and $V \mapsto VC^{-\top}$ for $C \in \mathrm{GL}(n)$ (an invertible matrix). We work on the regular stratum where $W$ is full-rank, and consider its singular value decomposition $W = PSQ^\top$. Then any $U$ and $V$ may be written as $U = PS^{1/2}C$ and $V = QS^{1/2}C^{-\top}$ for some choice of $C$. The change of measure and regularization are then

$$dU\, dV = |C|^{-n}\, dW\, dC, \qquad R = \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2) = \frac{1}{2}\operatorname{Tr}\left(S\left(CC^\top + (CC^\top)^{-1}\right)\right),$$

where we write $|C| = |\det C|$. The fiber partition function has a closed form in terms of a matrix Bessel function. Of interest to us, however, are the two limiting regimes described previously.

Taking $T \to 0$, and noting $CC^\top$ is positive definite, gives effective nuclear norm regularization

$$R_{\text{eff}}(W) = \frac{1}{2}\inf_{X \in \mathrm{PD}(n)} \operatorname{Tr}\left(S\left(X + X^{-1}\right)\right) = \operatorname{Tr}(S) = \|W\|_*$$

the second-to-last equality being a lower bound which is tight when $X$ is the identity. This corresponds to the matrix factorization results of Srebro et al. (2004); Dai et al. (2021). For the $\lambda \to 0$ limit, we examine the fiber partition function

$$Z_\alpha(W) = \int e^{-\frac{\alpha}{2}\operatorname{Tr}\left(S\left(CC^\top + (CC^\top)^{-1}\right)\right)} d\mu(C) \propto \int_{X \succ 0} e^{-\frac{\alpha}{2}\operatorname{Tr}\left(S\left(X + X^{-1}\right)\right)} |X|^{-(n+1)/2}\, dX.$$

A generalization of the argument from the Section 3 (Appendix C.2), centered around the integrand being non-negligible where $X$ has eigenvalues $\alpha \lesssim \lambda_i(X) \lesssim \frac{1}{\alpha}$, gives that $Z_\alpha(W)$ scales as

$$\alpha^{-kr}\left(\log\frac{1}{\alpha^2}\right)^{\mathbf{1}\{n \text{ is odd}\}} \int_{\mathrm{Gr}(k,n)} \left|Q^\top S^2 Q\right|^{-r/2}\, d\nu_{k,n}(Q).$$

Here, $k = \lfloor n/2 \rfloor$, and $r = n - k = \lceil n/2 \rceil$, which maximizes $\alpha^{-kr}$ over the choices of $k$ for $\alpha \to 0$. The measure $\nu_{k,n}$ is the invariant probability measure on the Grassmanian, the space of $k$ dimensional subspaces of $\mathbb{R}^n$, and $Q$ is any orthonormal basis of the chosen subspace. The renormalized entropy is thus

$$\log \Omega(W) = \log \int_{\mathrm{Gr}(k,n)} \mathrm{Vol}_k(W; Q)^{-r}\, d\nu_{k,n}(Q),$$

where $\mathrm{Vol}_k(W; Q) = |Q^\top W^\top W Q|^{1/2}$ measures how much $W$ expands $k$-dimensional volume in the subspace spanned by $Q$. (Note we can replace $S^2$ by $W^\top W$ in the integrand by basis-invariance). The entropic bias thus has an intuitive interpretation: those $W$ are preferred which contract some $k$-dimensional subspace. Equivalently, by the variational characterization of the singular values $s_1 \le \cdots \le s_n$, we have $\min_Q \mathrm{Vol}_k(W; Q) = s_1 \cdots s_k$, so the bias is towards matrices whose products of the $k$ smallest singular values are close to zero.

**Comparison to unregularized Langevin** As with diagonal networks, we can ask whether the entropic correction above is predictive of unregularized Langevin dynamics. We find this relation to hold empirically in this case too (Figure 4). The rank bias one can read off from the Grassmanian integral is apparent in the skew of $s_1(W)$ towards zero. Surprisingly, even in this more complex matrix model, the $\lambda_{\text{eff}}(t)$ heuristic appears to closely match the true distribution $p_{0,T}(W_t)$.

## 5 HOMOGENEOUS NEURAL NETWORKS

We turn to neural networks $f(x) = W_D\sigma(W_{D-1}\cdots\sigma(W_1 x))$ with $W_d \in \mathbb{R}^{n_d \times n_{d-1}}$ and one-homogeneous activation, $\sigma(cz) = c\sigma(z)$ for $c > 0$. This includes ReLU and leaky ReLU. (For homogeneity of different orders see Appendix C.5.) Such networks have neuron-wise scaling symmetry $W_d \mapsto P_d W_d P_{d-1}^{-1}$, where $P_d$ are diagonal matrices with $P_0$ and $P_D$ identity matrices. We introduce coordinates by recursively normalizing each layer's incoming weights: $W_d = P_d \bar{W}_d P_{d-1}^{-1}$

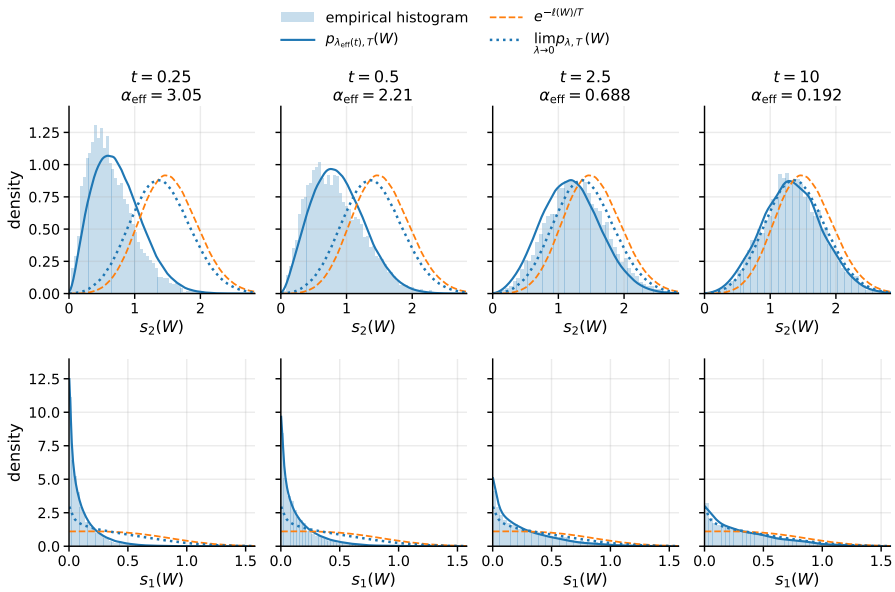



Figure 4: Langevin densities for the singular values of $W = UV^\top$ with $\ell(W) = \|W - W_*\|_F^2/2$ for $W_* = \text{diag}(1.25, 0.55)$, with $T = 0.25$. Empirical $\lambda = 0$ densities track $\alpha_{\text{eff}}(t)$ (blue), and approach $\lim_{\lambda\to 0} p_{\lambda,T}$ (blue, dotted), as opposed to the naïve distribution (orange, dashed).

where $\|(\bar{W}_d)_{i\cdot}\| = 1$ for all $d < D$ and $i \in [n_d]$, and $P_d = \text{diag}(\rho_d)$. The change of measure and regularization have simple forms in terms of edges $e = (u, v)$ between nodes:

$$dW = d\omega(\bar{W}_{1:D-1})\, d\bar{W}_D \prod_{\text{hidden } v} \rho_v^{\delta_v - 1}\, d\rho_v \qquad R(\bar{W}, \rho) = \frac{1}{2} \sum_{e=(u,v)} \bar{W}_e^2 \left(\frac{\rho_v}{\rho_u}\right)^2$$

where $d\omega$ is the product of spherical surface measures on the rows of $\bar{W}_d$ for $d < D$, and $\delta_v$ is the in-degree of $v$ minus its out-degree. The normalized weights $\bar{W}$ are reduced parameters eliminating the node scaling symmetry, but may still contain "hidden symmetries" (Rolnick & Kording, 2020; Grigsby et al., 2023), whose general characterization is open (Grillo & Montúfar, 2026). On regular regions where scaling is the only local symmetry, $f$ locally determines $\bar{W}$ and $(\bar{W}, P)$ are the symmetry coordinates of Section 2. Otherwise, integrating over $P$ gives implicit bias just over $\bar{W}$.

With that in mind, the $T \to 0$ limit gives, with an abuse of notation

$$R_{\text{eff}}(f) = \inf_{\bar{W}: f_{\bar{W}} = f} R_{\text{eff}}(\bar{W}), \qquad R_{\text{eff}}(\bar{W}) = \inf_{\rho > 0} R(\bar{W}, \rho).$$

Differentiating $R(\bar{W}, \rho)$ with respect to $\log \rho_v$ for any node $v$, we find the infimum in $R_{\text{eff}}(\bar{W})$ is attained when $\sum_{(u,v)} W_{uv}^2 = \sum_{(v,z)} W_{vz}^2$, at *balance* (Stock et al., 2019). Minimizing over any remaining hidden symmetries in $R_{\text{eff}}(f)$ recovers a version of the cost of Neyshabur et al. (2015).

For $\lambda \to 0$ we consider a tractable but illustrative case: all layers having the same width $n_d = n$. Then $\delta_v = 0$ for all nodes $v$. Following the diagonal network case and letting $\rho_v = \alpha^{-s_v/2}$,

$$Z_\alpha(\bar{W}) = \left(\frac{1}{2} \log \frac{1}{\alpha}\right)^{(D-1)n} \int \exp\left(-\frac{1}{2} \sum_{e=(u,v)} \bar{W}_e^2 \alpha^{1-(s_v - s_u)}\right) \prod_v ds_v.$$

As $\alpha \to 0$ the integrand approaches one when $s_v - s_u < 1$ for all $e = (u, v)$, and vanishes otherwise. Exchanging limits (as justified in Appendix C.3), we get $Z_\alpha(\bar{W}) C_{D,n} \sim \log(1/\alpha)^{(D-1)n}$. We make two observations. First, there is no entropic selection among choices of $\bar{W}$ on this stratum, but as we exhibit in the next section, this does not preclude entropic bias in the function $f$ itself. (Different layer shapes do give non-trivial entropy for $\bar{W}$ encouraging feature sparsity. See Appendix C.4.)

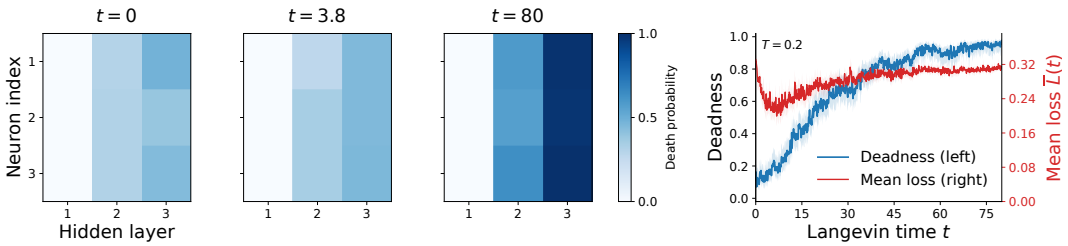



Figure 5: For a three hidden layer ReLU network trained using unregularized Langevin dynamics to match $f_\star(x) = \mathrm{ReLU}(x) + \frac{1}{2}\mathrm{ReLU}(-x)$ with inputs $x = -1, +1$ and squared loss, the probabilities of dead neurons (with activation zero across all inputs) increase over time (heatmaps). Correspondingly, the probability that the function outputs zero (deadness) goes to one (right plot).

Second, as noted above, integrating over $P$ may not integrate over all of $\pi^{-1}(f)$. Despite this, we can determine the complete implicit bias in $f$ in the case of ReLU networks without characterizing the hidden symmetries. A direct calculation shows the zero function dominates the entropy.

Define the dead region $\mathcal{D} = \{\theta : \pi(\theta) = 0\}$. In ReLU networks of at least two hidden layers, $\mathcal{D}$ contains an open set (a layer of negative weights zeros out later activations). Since $\pi(c\theta) = c^D\pi(\theta)$, the dead region is closed under positive scaling. Therefore, by the change of variables $\theta = \alpha^{-1/2}\xi$,

$$\nu_\alpha(\{0\}) = \int_{\mathcal{D}} e^{-\alpha\|\theta\|^2/2}\, d\theta = (2\pi)^{N/2}\alpha^{-N/2}\Pr(\xi \in \mathcal{D})$$

for $\xi \sim \mathcal{N}(0, I_N)$. The volume growth as $\alpha \to 0$ is much faster than the logarithmic growth on the regular region above. Going further, we now show the dead region dominates over all others, including those with hidden symmetry. Using the same change of variables as above in the denominator,

$$\Pr_{\lambda,T}(f = 0) = \frac{e^{-\ell(0)/T}\nu_\alpha(\{0\})}{\int e^{-\ell(f)/T}\nu_\alpha(df)} = \frac{(2\pi/\alpha)^{N/2}e^{-\ell(0)/T}\Pr(\xi \in \mathcal{D})}{(2\pi/\alpha)^{N/2}\mathbb{E}\left[e^{-\ell(\alpha^{-D/2}\pi(\xi))/T}\right]}.$$

Suppose the loss $\ell$ is radially coercive, meaning $\ell(cf) \to \infty$ as $c \to \infty$ for any non-zero $f \in \pi(\Theta)$. (This includes regression losses such as population squared error.) Then the term inside the expectation converges to $\exp(-\ell(0)/T)\mathbf{1}_{\mathcal{D}}$. Assuming $\ell$ is bounded below, the dominated convergence theorem implies the denominator approaches the numerator as $\alpha \to 0$, and so $\Pr_{\lambda,T}(f = 0) \to 1$. This means that despite potentially having a much higher loss than other functions, the zero function dominates in the no regularization limit. We call this concentration on $f = 0$ *neural heat death*. As with our previous models, our calculations are predictive of Langevin dynamics with zero regularization: Figure 5 shows the progressive heat death of an unregularized ReLU network.

## 6 Simplicity bias and the Neural Tangent Kernel

In multiple examples — linear networks, heat death, as well as non-constant width homogeneous networks (see Appendix C.4) — we see an entropic bias towards some form of sparsity. We have also seen multiple examples of models where the entropy correction in effective parameters appears trivial, such as on a regular stratum of square ReLU networks. This, however, does not mean there is no symmetry-induced implicit bias in $f$. In fact, as we show below, *even when* the effective parameters are (locally) identifiable there is a bias towards sparsity, at the level of entropy of functions $f$. The key is the change in volume captured by the NTK.

**Simplicity bias via the NTK** Suppose we are on a regular stratum on which $\mathrm{rank}(D\pi_\theta) = k$. To define a density on functions, we fix an input distribution $x \sim Q$ and take $df = d\mathcal{H}^k_Q(f)$, the induced Hausdorff measure (Appendix C.6). By the coarea formula (see e.g. Simon (2018)), the fiber partition function with respect to $df$ is

$$Z_\alpha(f) = \int_{\pi^{-1}(f)} \frac{e^{-\alpha R(\theta)}}{\sqrt{\det{}^+ K(\theta)}} d\mathcal{H}^{N-k}(\theta),$$

where $K(\theta) = D\pi_\theta D\pi_\theta^\top$ is the NTK and $\det^+$ denotes the product of non-zero eigenvalues. From this we see that functions are preferred when they vary slowly with the parameters. This is intuitive: Langevin dynamics spend more time near functions that are insensitive to parameter motion.

The same interpretation applies after symmetry coordinates have been eliminated and replaced with identifiable effective parameters $\bar{\theta}$, where $\pi(\theta) = \bar{\pi}(\bar{\theta})$, to get $\Omega(\bar{\theta})$. Then the integral above is replaced by the single point $\bar{\theta} = \bar{\pi}^{-1}(f)$. Replacing $K$ by $\bar{K}$ (defined using $\bar{\pi}$ in the obvious way), and taking $\alpha \to 0$ we get the entropic correction $\log \Omega(f) = \log \Omega(\bar{\theta}) - \frac{1}{2} \log \det^+(\bar{K}(\bar{\theta}))$.

**A simple example** Consider two-layer ReLU networks $f(x) = \sum_{i=1}^n u_i \sigma(v_i^\top x)$ where $x \in \mathbb{R}^{n_0}$. Following the normalization procedure from the previous section, we get $f(x) = \sum_{i=1}^n r_i \sigma(\varphi_i^\top x)$ where the $\varphi_i$ are unit vectors and $r_i = \|v_i\| u_i$. By our usual recipe (Appendix C.5), we find $\log \Omega(r, \varphi) = 0$ relative to $dr\, d\omega(\varphi)$. Note that if all $r_i \neq 0$ and $\varphi_i \neq \pm\varphi_j$ for $i \neq j$, then $r$ and the hyperplanes normal to $\varphi$ are determined by $f$ up to permutation (Dereich & Kassing, 2022), making these effective parameters except on a set of measure zero. Following a brief calculation of the neural tangent features (Appendix C.6), we find that outside of zero or colliding neurons

$$\log \Omega(f) = -(n_0 - 1) \sum_{i=1}^n \log |r_i| + \varphi \text{ term}.$$

From this we can directly read off a preference towards neurons of small scale $r_i$. The mechanism is exactly an NTK sparsity preference: as $r_i$ vanishes, changes in $\varphi_i$ have less effect on $f$. The effect is akin to the fact that sampling polar coordinates uniformly does not produce uniform points in space.

## 7 Discussion and related work

Our Langevin framework naturally affords and relates two limiting regimes. Taking $T \to 0$ minimizes regularization over the fiber first, giving an effective regularization. This recovers representation cost theories (Dai et al., 2021). While these have been shown not to describe gradient flow in general (Arora et al., 2019), we find them relevant to the $T \to 0$ limit of Langevin dynamics. The balance condition we recover for deep homogeneous networks occurs across gradient flow analyses, in the form of conservation laws (Saxe et al., 2014; Du et al., 2018; Kunin et al., 2021). On the other hand, works on SGD (with positive temperature) find generalizations of balance (Ziyin et al., 2024; 2025) or other dynamically attractive conditions Chen et al. (2023). This attraction within fibers highlights a difference between SGD and our other limiting regime.

Taking $\lambda \to 0$ with $T > 0$ instead reveals entropic selection based on the asymptotic growth of fibers. Aladrah et al. (2026) develop a closely related perspective on symmetry-induced bias, but handle non-compact symmetries by gauge-fixing. This essentially assumes the balanced case, consistently with their experiments, which use SGD. Similarly Menon & Yu (2025) derive an entropy on the balanced manifold of deep linear networks, whose symmetry is compact. In contrast, we study the entire non-compact fiber (though we do recover the Menon & Yu (2025) entropy as a subleading term in the $T \to 0$ expansion of $Z_\alpha$; Appendix C.7). Our results complement empirical studies of simplicity bias which suggest parameter volume as a mechanism (Valle-Pérez et al., 2019), and our explicit linear network entropy may help explain observed low-rank bias (Huh et al., 2023).

Analysis of neural network priors when $\lambda > 0$ is established (Li & Sompolinsky, 2021; Hanin & Zlokapa, 2023). Particularly relevant are Zavatone-Veth & Pehlevan (2021); Noci et al. (2021), who find exact finite-width priors for linear and ReLU networks using Meijer-G functions, which include the Bessel functions we find for diagonal networks as a special case. In particular, Zavatone-Veth & Pehlevan (2021) find an atom at zero output for a fixed input. Heat death extends this result by showing the probability of outputting zero on *all* inputs goes to one as variance increases. Related work has shown ReLU networks at initialization can die as depth increases, at finite parameter variance (Lu et al., 2020; Luther & Seung, 2020). Our method also bears some resemblance to singular learning theory (Watanabe, 2013), though the exact integrals and $\lambda \to 0$ limit are distinct.

Our work opens the door to interesting questions we do not answer here, such as what the renormalized entropy is for other architectures, and when the limit interchange between weak regularization and unregularized Langevin is formally justified. Additionally, what can be said under other kinds of noise, when a Gibbs equilibrium may not exist? We leave these questions to future work.

## REFERENCES


Milton Abramowitz and Irene A. Stegun (eds.). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, Washington, DC, 1964.

Laurence Aitchison, Jannes Jegminat, Jorge Aurelio Menendez, Jean-Pascal Pfister, Alexandre Pouget, and Peter E. Latham. Synaptic plasticity as Bayesian inference. *Nature Neuroscience*, 24:565–571, 2021. doi: 10.1038/s41593-021-00809-5. URL https://www.nature.com/articles/s41593-021-00809-5.

Nicola Aladrah, Emanuele Ballarin, Matteo Biagetti, Alessio Ansuini, Alberto d'Onofrio, and Fabio Anselmi. Understanding and inverse design of implicit bias in stochastic learning: a geometric perspective. arXiv preprint, 2026. URL https://arxiv.org/abs/2601.06597. Version 2; originally titled Implicit bias as a Gauge correction: Theory and Inverse Design.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization, 2019. URL https://arxiv.org/abs/1905.13655.

Raphaël Berthier. Incremental learning in diagonal linear networks, 2023. URL https://arxiv.org/abs/2208.14673.

Raphaël Berthier. Diagonal linear networks and the lasso regularization path, 2026. URL https://arxiv.org/abs/2509.18766.

Blake Bordelon and Cengiz Pehlevan. Self-Consistent Dynamical Field Theory of Kernel Evolution in Wide Neural Networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.09653.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.11029.

Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic Collapse: How Gradient Noise Attracts SGD Dynamics Towards Simpler Subnetworks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.04251.

Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26884–26896. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e22cb9d6bbb4c290a94e4fff4d68a831-Paper.pdf.

Steffen Dereich and Sebastian Kassing. On minimal representations of shallow ReLU networks. *Neural Networks*, 148:121–128, 2022. doi: 10.1016/j.neunet.2022.01.006. URL https://arxiv.org/abs/2108.05643.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

J. Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden Symmetries of ReLU Networks. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11734–11760, 2023. URL https://proceedings.mlr.press/v202/grigsby23a.html.

Moritz Grillo and Guido Montúfar. Most ReLU Networks Admit Identifiable Parameters. arXiv preprint, 2026. URL https://arxiv.org/abs/2605.03601.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit Regularization in Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.09280.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Boris Hanin and Alexander Zlokapa. Bayesian Interpolation with Deep Linear Networks. *Proceedings of the National Academy of Sciences*, 120(23):e2301345120, 2023. doi: 10.1073/pnas.2301345120. URL https://arxiv.org/abs/2212.14457.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The Low-Rank Simplicity Bias in Deep Networks. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bCiNWDmlY2.

Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4629–4640, 2021. URL https://proceedings.mlr.press/v139/izmailov21a.html.

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. URL https://arxiv.org/abs/1806.07572.

David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Network Plasticity as Bayesian Inference. *PLOS Computational Biology*, 11(11):e1004485, 2015. doi: 10.1371/journal.pcbi.1004485. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004485.

Julius Kobialka, Emanuel Sommer, Chris Kolb, Juntae Kwon, Daniel Dold, and David Rügamer. On the Interplay of Priors and Overparametrization in Bayesian Neural Network Posteriors. In *International Conference on Artificial Intelligence and Statistics*, 2026. URL https://arxiv.org/abs/2603.22030.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics, 2021. URL https://arxiv.org/abs/2012.04728.

Qianyi Li and Haim Sompolinsky. Statistical Mechanics of Deep Linear Neural Networks: The Back-Propagating Kernel Renormalization. *Physical Review X*, 11:031059, 2021. doi: 10.1103/PhysRevX.11.031059. URL https://arxiv.org/abs/2012.04030.

Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. Dying ReLU and Initialization: Theory and Numerical Examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020. doi: 10.4208/cicp.OA-2020-0165. URL https://arxiv.org/abs/1903.06733.

Kyle Luther and H. Sebastian Seung. Reexamining the principle of mean-variance preservation for neural network initialization. *Physical Review Research*, 2(3):033135, 2020. doi: 10.1103/PhysRevResearch.2.033135.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017. URL https://jmlr.org/papers/v18/17-214.html.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. arXiv preprint, 2025. URL https://arxiv.org/abs/2508.07370. Version 2 revised 12 March 2026.

Govind Menon and Tianmin Yu. An entropy formula for the Deep Linear Network. arXiv preprint, 2025. URL https://arxiv.org/abs/2509.09088.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-Based Capacity Control in Neural Networks. In *Proceedings of the Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, 2015. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. Precise characterization of the prior predictive distribution of deep ReLU networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/ae4503ec3da32f5e9033604744ec45ae-Abstract.html.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks, 2023. URL https://arxiv.org/abs/2304.00488.

David Rolnick and Konrad Kording. Reverse-engineering deep ReLU networks. In *Proceedings of the International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8178–8187, 2020. URL https://proceedings.mlr.press/v119/rolnick20a.html.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2014. URL https://arxiv.org/abs/1312.6120.

Leon Simon. *Introduction to Geometric Measure Theory*. Stanford University, 2018. URL https://math.stanford.edu/~lms/ntu-gmt-text.pdf. Lecture notes, 2017–2018 version; Chapter 2, Section 7, Theorem 7.3 and Remark 7.4(1), p. 76.

Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. Maximum-Margin Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 17, 2004. URL https://proceedings.neurips.cc/paper/2004/hash/e0688d13958a19e087e123148555e4b4-Abstract.html.

Pierre Stock, Benjamin Graham, Rémi Gribonval, and Hervé Jégou. Equi-normalization of Neural Networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1902.10416.

Guillermo Valle-Pérez, Camilo Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.08522.

Sumio Watanabe. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14(27):867–897, 2013. URL https://jmlr.org/papers/v14/watanabe13a.html.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011. URL https://www.stats.ox.ac.uk/~teh/research/compstats/WelTeh2011a.pdf.

Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.00004.

Jacob A. Zavatone-Veth and Cengiz Pehlevan. Exact marginal prior distributions of finite Bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3364–3375, 2021. URL https://arxiv.org/abs/2104.11734.

Jingzhao Zhang, Haochuan Li, Suvrit Sra, and Ali Jadbabaie. Neural Network Weights Do Not Converge to Stationary Points: An Invariant Measure Perspective. In *Proceedings of the International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26330–26346, 2022. URL https://proceedings.mlr.press/v162/zhang22q.html.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces, 2025. URL https://arxiv.org/abs/2506.13018.

Liu Ziyin, Mingze Wang, Hongchao Li, and Lei Wu. Parameter Symmetry and Noise Equilibrium of Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 37, pp. 93874–93906, 2024. URL https://arxiv.org/abs/2402.07193.

Liu Ziyin, Yizhou Xu, and Isaac Chuang. Neural Thermodynamics: Entropic Forces in Deep and Universal Representation Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.12387. Published at NeurIPS 2025; arXiv version 4 revised 2 February 2026.

# APPENDICES

**A AI use disclosure** **13**

**B Additional figures** **13**

**C Proofs** **14**

C.1 Diagonal network integral limit . . . . . 14

C.2 Two layer linear network entropy . . . . . 15

C.3 Square ReLU networks integral limit . . . . . 17

C.4 Square-to-scalar ReLU networks . . . . . 17

C.5 Two layer homogeneous network . . . . . 19

C.6 Functional entropy . . . . . 20

C.7 Subleading low-temperature terms and the Menon-Yu entropy . . . . . 21

**D Experiments** **23**

D.1 Diagonal networks . . . . . 23

D.2 Linear networks . . . . . 23

D.3 Neural heat death . . . . . 23

## A AI USE DISCLOSURE

Proof techniques were developed by the authors with AI assistance. AI suggestions were carefully checked and compared to existing literature, and the text and proofs themselves were written by the authors. AI was used to implement Langevin simulations, the resulting figures and descriptions of the experiments. We take responsibility for the final content of this work.

## B ADDITIONAL FIGURES

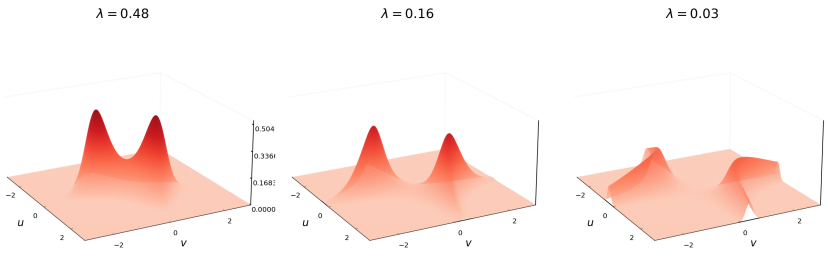



Figure 6: Stationary distributions for $w = uv$ in a two-layer diagonal network $f(x) = uvx$ trained by Langevin dynamics on the loss $L(u, v) = (uv - 1)^2/2$ (temperature $T = 0.15$). As $\lambda$ decreases, the density spreads across the loss valley, with mass escaping to infinity.

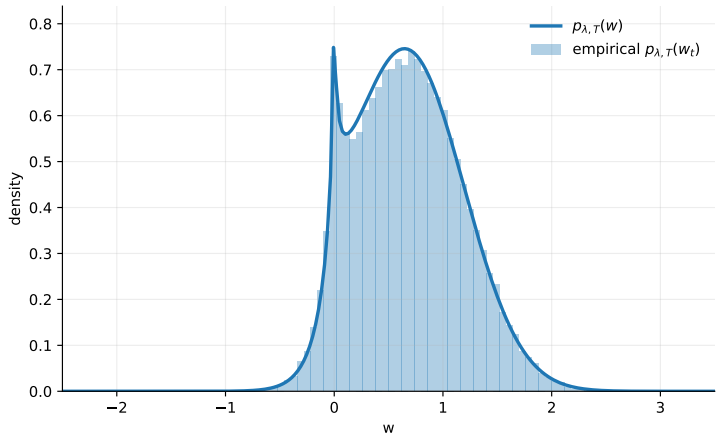



Figure 7: Analytical and empirical density of $w$ for one-layer diagonal network $f(x) = uvx$ with $L(u, v) = (uv - 1)^2/2$ trained by Langevin dynamics, with $\lambda = 0.2$ and $T = 0.25$. Samples are collected after burn-in and thinned.

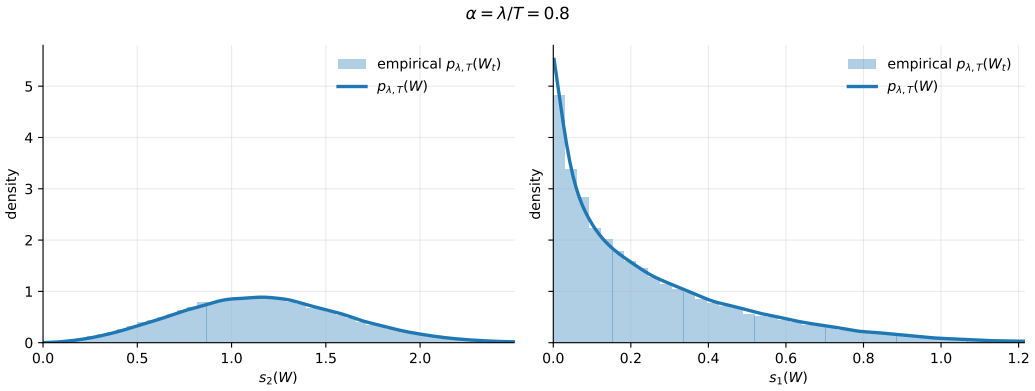



Figure 8: Analytical and empirical density of the singular values of $W$ for a one-layer square linear network $f(x) = UV^\top x$ with $\ell(W) = \|W - W_*\|_F^2/2$ for $W = UV^\top$ and $W_* = \mathrm{diag}(1.25, 0.55)$, trained by Langevin dynamics, with $\lambda = 0.2$ and $T = 0.25$. Samples are collected after burn-in and thinned.

# C Proofs

## C.1 Diagonal network integral limit

We examine the integral

$$Z_\alpha(w) = \frac{\log(1/\alpha)}{2} \int_{-\infty}^{\infty} e^{-\frac{|w|}{2}\left(\alpha^{1-s}+\alpha^{1+s}\right)}\, ds.$$

As explained in the main text, the integrand converges to the indicator $\mathbf{1}_{(-1,1)}(s)$ as $\alpha \to 0$. We apply the dominated convergence theorem to show the integral converges. Namely, for $|s| \le 1$ the integrand is bounded by one, while for $|s| > 1$, once $\alpha < e^{-1}$ the integrand is bounded by $\exp(-|w|e^{|s|-1}/2)$. The argument above generalizes readily to deep chains, and by the decoupling of coordinates, to general deep diagonal networks.

To formally obtain convergence of the distributions $p_{\lambda,T}(w)$ as $\lambda \to 0$, one needs an upper bound integrable against $e^{-\ell(w)/T}\, dw$. Such a bound holds, for example, for coercive losses such as the quadratic losses used in our examples and experiments. Namely, for a depth-$D$ chain the upper

bound

$$\frac{Z_\alpha(w)}{\log(1/\alpha)^{D-1}} \lesssim (1 + |\log|w||^{D-1})$$

is valid once $\alpha \leq e^{-1}$.

### C.2 Two layer linear network entropy

Recall that

$$Z_\alpha(S) \propto \int_{X \succ 0} e^{-\frac{\alpha}{2}\operatorname{Tr}\left(S\left(X+X^{-1}\right)\right)} |X|^{-(n+1)/2}\, dX.$$

The scaling argument proceeds by compactification, or evaluation at a boundary, of the positive definite cone. We begin by identifying which boundary dominates, i.e. how many eigenvalues of $X$ go to zero in the dominant integrand. We then calculate the limiting fiber partition function at an arbitrary boundary, and confirm that the one we identify is indeed dominant.

Let $X$ have eigenvalues $\lambda_i(X)$. The two trace terms in $Z_\alpha(S)$ define an approximate cutoff region for these eigenvalues. Since $S$ is fixed and positive definite, $\operatorname{Tr}(SX) \sim \operatorname{Tr}(X)$, so the first trace term cuts off the integral when $\alpha \operatorname{Tr}(X) \gg 1$. Similarly, $\operatorname{Tr}(SX^{-1}) \sim \operatorname{Tr}(X^{-1})$ and so the second term cuts off the integral when $\alpha \operatorname{Tr}(X^{-1}) \gg 1$. Therefore, the integral is non-negligible on the region

$$\alpha \lesssim \lambda_i(X) \lesssim \frac{1}{\alpha}.$$

Intuitively, the dominant contribution comes from having some number $k$ of eigenvalues of order $\alpha$, and $r = n - k$ of order $\alpha^{-1}$, and is of order $\alpha^{-kr}$, which in turn is maximized when $k = \lfloor n/2 \rfloor$. The argument below formalizes this intuition.

It is convenient to change variables to $Y = \alpha S^{1/2} X S^{1/2}$, so that $Z_\alpha(S) \propto I_{\alpha^2}(S^2)$ where

$$I_\epsilon(D) = \int_{Y \succ 0} e^{-\frac{1}{2}\operatorname{Tr} Y} e^{-\frac{\epsilon}{2}\operatorname{Tr}\left(DY^{-1}\right)} |Y|^{-(n+1)/2}\, dY.$$

The relevant range is $\epsilon \lesssim \lambda_i(Y) \lesssim 1$, and we study the case where $k$ eigenvalues have scale $\epsilon$ and $r$ have scale 1. Any positive definite $Y$ can be written

$$Y = \begin{pmatrix} I_r & 0 \\ B^\top & I_k \end{pmatrix} \begin{pmatrix} A_1 & 0 \\ 0 & \epsilon A_0 \end{pmatrix} \begin{pmatrix} I_r & B \\ 0 & I_k \end{pmatrix}$$

for positive definite $A_1 \in \mathbf{S}^r_{++}$ and $A_0 \in \mathbf{S}^k_{++}$, and arbitrary $B \in \mathbb{R}^{r \times k}$. Here $A_1$ and $A_0$ correspond to the order-1 and order-$\epsilon$ directions, respectively, and $B$ describes the orientation of the vanishing directions. When $\epsilon = 0$, the kernel of $Y$ is the image of

$$N_B = \begin{pmatrix} -B \\ I_k \end{pmatrix}.$$

The determinant is then $|Y| = \epsilon^k |A_1||A_0|$, and the change of measure is

$$dY = \epsilon^{k(k+1)/2} |A_1|^k\, dA_1\, dA_0\, dB.$$

Therefore,

$$|Y|^{-(n+1)/2}\, dY = \epsilon^{-kr/2} |A_1|^{k-(n+1)/2} |A_0|^{-(n+1)/2}\, dA_1\, dA_0\, dB,$$

and we see indeed that for $\epsilon = \alpha^2$, the order of the integral is $\alpha^{-kr}$.

In order to evaluate the integral, note that

$$\begin{aligned} \operatorname{Tr} Y &= \operatorname{Tr}(A_1(I_r + BB^\top)) + \epsilon \operatorname{Tr} A_0, \\ \epsilon \operatorname{Tr} DY^{-1} &= \operatorname{Tr}(N_B^\top D N_B A_0^{-1}) + \epsilon \operatorname{Tr}(D_{11} A_1^{-1}), \end{aligned}$$

where $D_{11}$ is the upper-left $r$-by-$r$ block of $D$. Therefore, $I_\epsilon(D)$ is the integral with respect to $dB$ of the product of two integrals:

$$\int_{A_1 \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(A_1(I_r+BB^\top))-\frac{\epsilon}{2}\operatorname{Tr}(D_{11}A_1^{-1})}|A_1|^{k-(n+1)/2}\,dA_1$$

and

$$\int_{A_0 \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(N_B^\top D N_B A_0^{-1})-\frac{\epsilon}{2}\operatorname{Tr} A_0}|A_0|^{-(n+1)/2}\,dA_0.$$

Consider the integral over $A_1$, which using $n = k + r$ we write as

$$\int_{A_1 \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(A_1(I_r+BB^\top))-\frac{\epsilon}{2}\operatorname{Tr}(D_{11}A_1^{-1})}|A_1|^{\frac{k}{2}-\frac{r+1}{2}}\,dA_1.$$

The first exponential trace term keeps the integral finite as $A_1 \to \infty$. The second exponential trace term gives the cutoff $\epsilon \lesssim A_1 \lesssim 1$, and so the integral behaves like

$$\int_\epsilon^1 a^{\frac{k}{2}-\frac{r+1}{2}}\,da.$$

This is finite in the $\epsilon \to 0$ limit exactly when the exponent is greater than $-1$, that is, when $k \geq r$. When $k = r - 1$ the integral is order $\log(1/\epsilon)$, as we show below. If $k < r - 1$, then the integral diverges like a power of $\epsilon$, meaning our scaling is missing some directions which are in fact small. Similarly, the integral over $A_0$, changing variables to $H = A_0^{-1}$, is

$$\int_{H \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(N_B^\top D N_B H)-\frac{\epsilon}{2}\operatorname{Tr}(H^{-1})}|H|^{\frac{r}{2}-\frac{k+1}{2}}\,dH.$$

This is finite in the $\epsilon \to 0$ limit when $r \geq k$, and order $\log(1/\epsilon)$ when $r = k - 1$. We thus see that $k = \lfloor n/2 \rfloor$ and $r = \lceil n/2 \rceil$ is the dominant scaling.

To evaluate the integrals, recall the normalization integral for an $m$-by-$m$ Wishart matrix $H$ with $n$ degrees of freedom and scale parameter $M \in \mathbf{S}_{++}^m$,

$$\int_{H \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(M^{-1}H)}|H|^{\frac{n}{2}-\frac{m+1}{2}}\,dH = 2^{nm/2}\Gamma_m\left(\frac{n}{2}\right)|M|^{n/2},$$

where $\Gamma_m$ is the multivariate gamma function,

$$\Gamma_m(a) = \int_{H \succ 0} e^{-\operatorname{Tr} H}|H|^{a-\frac{m+1}{2}}\,dH = \pi^{m(m-1)/4}\prod_{i=1}^m \Gamma\left(a - \frac{i-1}{2}\right).$$

The above holds when $r \leq k$. Additionally,

$$\int_{H \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(PH)-\frac{\epsilon}{2}\operatorname{Tr}(QH^{-1})}|H|^{-1}\,dH \sim |P|^{-(m-1)/2}\log\left(\frac{1}{\epsilon}\right).$$

By a change of variables, it suffices to show this with $P = I$, and the scaling can be shown by changing variables to

$$H = \begin{pmatrix} h + u^\top K^{-1}u & u^\top \\ u & K \end{pmatrix}$$

for $h > 0$, $u \in \mathbb{R}^{m-1}$ and $K \in \mathbf{S}_{++}^{m-1}$. Thus, setting $k = \lfloor n/2 \rfloor$ and $r = \lceil n/2 \rceil$ we get

$$\int_{A_1 \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(A_1(I_r+BB^\top))-\frac{\epsilon}{2}\operatorname{Tr}(D_{11}A_1^{-1})}|A_1|^{\frac{k}{2}-\frac{r+1}{2}}\,dA_1 \sim \left(\log\frac{1}{\epsilon}\right)^{\mathbf{1}_{n\text{ odd}}}|I_r + BB^\top|^{-k/2}$$

and

$$\int_{A_0 \succ 0} e^{-\frac{1}{2}\operatorname{Tr}(N_B^\top D N_B A_0^{-1})-\frac{\epsilon}{2}\operatorname{Tr} A_0}|A_0|^{-(n+1)/2}\,dA_0 \sim |N_B^\top D N_B|^{-r/2}.$$

Finally, we evaluate

$$I_\epsilon(D) \sim c_n \epsilon^{-kr/2} \left(\log \frac{1}{\epsilon}\right)^{\mathbf{1}_{n \text{ odd}}} \int_{\mathbb{R}^{k\times r}} |N_B^\top D N_B|^{-r/2} |I_r + BB^\top|^{-k/2}\, dB.$$

Define the column-normalized matrix

$$Q_B = N_B (N_B^\top N_B)^{-1/2} = N_B (I_k + B^\top B)^{-1/2}.$$

Then

$$|N_B^\top D N_B| = \left|(N_B^\top N_B)^{1/2} (Q_B^\top D Q_B)(N_B^\top N_B)^{1/2}\right| = |I_k + B^\top B||Q_B^\top D Q_B|.$$

Since

$$|I_k + B^\top B| = |I_r + BB^\top|,$$

the integrand is

$$|N_B^\top D N_B|^{-r/2} |I_r + BB^\top|^{-k/2} = |Q_B^\top D Q_B|^{-r/2} |I_k + B^\top B|^{-n/2}.$$

But $|I_k + B^\top B|^{-n/2}\, dB$ is the invariant measure on $\mathrm{Gr}(k, n)$, which we write as $\nu_{k,n}$. The result is thus proved.

**Passing to the distributional limit** As in the diagonal case, formally, one must check that the convergence $Z_\alpha(W)/C_\alpha \to \Omega(W)$ holds in $L^1(e^{-\ell(W)/T}\, dW)$ to obtain convergence of $p_{\lambda,T}(W)$. For losses satisfying $\ell(W) \geq a\|W\|_F^2 - b$, we can find an integrable dominating function. In even dimensions $Z_\alpha(W)/C_\alpha \lesssim \Omega(W)$; in odd dimensions an additional factor $1 + |\log(1/s_1(W))|$ suffices (where $s_1$ is the smallest singular value). Both are Gaussian-integrable, so dominated convergence applies.

### C.3 Square ReLU networks integral limit

We examine the integral

$$\int \exp\left(-\frac{1}{2} \sum_{e=(u,v)} \bar{W}_e^2 \alpha^{1-(s_v - s_u)}\right) ds$$

for $s_v \in \mathbb{R}$. As shown in the main text, the integrand converges to the indicator on the interior of the set $\mathcal{P} = \{s : s_v - s_u \leq 1\ \forall e = (u, v\}$. This set has finite volume, since if $s_v - s_u \leq 1$ for all nodes in the network, then by induction $-(D - d) \leq s_{d,i} \leq d$ for all $d \in [D], i \in [n_d]$. To apply the dominated convergence theorem, we show the integrand is bounded by

$$\exp\left(-\frac{1}{2}\left(\min_{e=(u,v)} \bar{W}_e^2\right) e^{-1+\max_v |s_v|/D}\right)$$

(which is non-trivial on the chosen stratum). Indeed for $\max_v |s_v| > D$, some path from $v$ to the input or output has an edge $e = (u', v')$ such that $s_{v'} - s_{u'} \geq \max_v |s_v|/D$. The bound follows once $\alpha < e^{-1}$, since then $\alpha^{1-(s_{v'} - s_{u'})} \geq \exp(-1 + \max_v |s_v|/D)$. For $\max_v |s_v| \leq D$, on the other hand, we have the trivial upper bound of 1.

In this case we do not analyze whether $p_{\lambda,T}$, converges, since this point becomes moot under heat death.

### C.4 Square-to-scalar ReLU networks

We consider the ReLU network $f(x) = w_D^\top \sigma(W_{D-1} \cdots \sigma(W_1 x))$ with $W_d \in \mathbb{R}^{n\times n}$ for $d < D$ and $w_D \in \mathbb{R}^n$, with $n \geq 2$. With the normalization $W_d = P_d \bar{W}_d P_{d-1}^{-1}$ as described in the main text (with $w_D = P_{D-1}^{-1} \bar{w}_D$),

$$R(\bar{W}, \rho) = \frac{1}{2}\left(\sum_i \rho_{1,i}^2 + \sum_{d=2}^{D-1} \sum_{j,i} (\bar{W}_d)_{ji}^2 \left(\frac{\rho_{d,j}}{\rho_{d-1,i}}\right)^2 + \sum_i \left(\frac{\bar{w}_{D,i}}{\rho_{D-1,i}}\right)^2\right)$$

and

$$dW = d\omega(\bar{W}) d\bar{w}_D \left( \prod_{d=1}^{D-2} \prod_i \rho_{d,i}^{-1} \, d\rho_{d,i} \right) \left( \prod_i \rho_{D-1,i}^{n-2} \, d\rho_{D-1,i} \right).$$

Performing the change of variables $\rho_{d,i} = \alpha^{-d/2} s_{d,i}$,

$$\alpha R(\bar{W}, \rho) = \frac{1}{2} \left( \sum_i s_{1,i}^2 + \sum_{d=2}^{D-1} \sum_{j,i} (\bar{W}_d)_{ji}^2 \frac{s_{d,j}^2}{s_{d-1,i}^2} + \alpha^D \sum_i \frac{\bar{w}_{D,i}^2}{s_{D-1,i}^2} \right)$$

while the measure keeps the same form except for a scaling,

$$dW = \alpha^{-(D-1)n(n-1)/2} d\omega(\bar{W}) d\bar{w}_D \left( \prod_{d=1}^{D-2} \prod_i s_{d,i}^{-1} \, ds_{d,i} \right) \left( \prod_i s_{D-1,i}^{n-2} \, ds_{D-1,i} \right).$$

Writing $d\mu(s)$ for the $s$-dependent part of the measure above,

$$Z_\alpha(\bar{W}) = \alpha^{-(D-1)n(n-1)/2} \int e^{-\frac{1}{2}\left( \sum_i s_{1,i}^2 + \sum_{d=2}^{D-1} \sum_{j,i} (\bar{W}_d)_{ji}^2 \frac{s_{d,j}^2}{s_{d-1,i}^2} + \alpha^D \sum_i \frac{\bar{w}_{D,i}^2}{s_{D-1,i}^2} \right)} d\mu(s).$$

Only the last term in the integrand contains $\alpha$. Taking it to zero (applying dominated convergence), and provided the integral below is finite, we get $Z_\alpha(\bar{W}) \sim C_\alpha \Omega(\bar{W})$ with

$$C_\alpha = \alpha^{-(D-1)n(n-1)/2}, \qquad \Omega(\bar{W}) = \int e^{-\frac{1}{2}\left( \sum_i s_{1,i}^2 + \sum_{d=2}^{D-1} \sum_{j,i} (\bar{W}_d)_{ji}^2 \frac{s_{d,j}^2}{s_{d-1,i}^2} \right)} d\mu(s).$$

Note the scaling $C_\alpha$ is dominated by the scaling of the heat death stratum from the main text. To see the integral does indeed exists, note the integrand is upper bounded by keeping only the diagonal terms,

$$\exp\left( -\frac{1}{2} \left( \sum_i s_{1,i}^2 + \sum_{d=2}^{D-1} \sum_{j,i} (\bar{W}_d)_{ji}^2 \frac{s_{d,j}^2}{s_{d-1,i}^2} \right) \right) \leq \exp\left( -\frac{1}{2} \left( \sum_i s_{1,i}^2 + \sum_{d=2}^{D-1} \sum_i (\bar{W}_d)_{ii}^2 \frac{s_{d,i}^2}{s_{d-1,i}^2} \right) \right).$$

We obtain an upper bound with decoupled integrals

$$\Omega(\bar{W}) \leq \prod_i \int e^{-\frac{1}{2}\left( s_{1,i}^2 + \sum_{d=2}^{D-1} (\bar{W}_d)_{ii}^2 \frac{s_{d,i}^2}{s_{d-1,i}^2} \right)} d\mu_i(s_i)$$

where $d\mu_i$ is the $s_i$-dependent part of $d\mu$. Changing variables to $u_{d,i} = s_{d,i}/s_{d-1,i}$, we find the upper bound is equal to

$$\left( \prod_i \int u_{1,i}^{n-2} e^{-u_{1,i}^2/2} \, du_{1,i} \right) \prod_{d=2}^{D-1} \prod_i \int u_{d,i}^{n-2} e^{-(\bar{W}_d)_{ii}^2 u_{d,i}^2/2} \, du_{d,i}.$$

This expression is finite, being the product of Gaussian integrals.

**Feature sparsity bias** In terms of nodes $u, v$

$$\Omega(\bar{W}) = \int \exp\left( -\frac{1}{2} \sum_{\substack{e=(u,v) \\ v \text{ hidden}}} \bar{W}_e^2 \left( \frac{s_v}{s_u} \right)^2 \right) d\mu(s),$$

where $s_u = 1$ for $u$ a coordinate of the input $x$. Integrating over a fixed node $v$,

$$\Omega(\bar{W}) = \int (s_{-v} \text{ term}) \int \exp\left( -\frac{1}{2} \sum_{u \to v} \frac{\bar{W}_{(u,v)}^2}{s_u^2} s_v^2 - \frac{1}{2} \sum_{v \to u} \bar{W}_{(u,v)}^2 s_u^2 \frac{1}{s_v^2} \right) \frac{ds_v}{s_v} \, d\mu_{-v}(s_{-v}).$$

Suppose that the norm of the outgoing normalized weights $\bar{W}$ from $v$ is $O(\epsilon)$. To obtain a lower bound on the integral, restrict all other scales a finite box. On this region,

$$\sum_{u \to v} \frac{\bar{W}_{(u,v)}^2}{s_u^2} \sim 1 \qquad \sum_{v \to u} \bar{W}_{(u,v)}^2 s_u^2 \lesssim \epsilon^2.$$

By our usual cutoff analysis, the inner integral is lower bounded as $\log(1/\epsilon)$, and thus $\log \Omega(\bar{W}) \gtrsim \log\log(1/\epsilon)$. This is exactly a sparsity bias: the entropy increases as the feature coming from node $v$ becomes unused.

### C.5 TWO LAYER HOMOGENEOUS NETWORK

Consider a two-layer homogeneous-activation network

$$f(x) = \sum_{i=1}^{m} u_i \sigma\left(v_i^\top x\right)$$

for $u_i \in \mathbb{R}$ and $v_i \in \mathbb{R}^n$, and $\sigma : \mathbb{R} \to \mathbb{R}$ with $\sigma(cz) = c^q \sigma(z)$ for $c > 0$. This has a scaling symmetry $v_i \mapsto cv_i$, $u_i \mapsto u_i/c^q$. We may thus let

$$r_i = \rho_i^q u_i \qquad \varphi_i = \frac{v_i}{\rho_i} \qquad \rho_i = \|v_i\|,$$

so

$$f(x) = \sum_{i=1}^{m} r_i \sigma\left(\varphi_i^\top x\right).$$

Squared-norm regularization on the parameters then takes the form

$$R(u, v) = \frac{1}{2}\sum_{i=1}^{m} u_i^2 + \|v_i\|^2 = \frac{1}{2}\sum_{i=1}^{m} \frac{r_i^2}{\rho_i^{2q}} + \rho_i^2.$$

Changing measure gives

$$\prod_{i=1}^{m} du_i\, dv_i = \prod_{i=1}^{m} \rho_i^{n-q-1}\, d\rho_i\, dr_i d\omega(\varphi_i)$$

where $d\omega$ indicates the uniform measure on the unit $(n-1)$-sphere. If we treat $r$ and $\varphi$ as effective parameters, we get

$$p_{\lambda,T}(r, \varphi) \prod_{i=1}^{m} dr_i\, d\omega(\varphi_i) \propto e^{-\ell(r,\varphi)/T} Z_{\lambda/T}(r, \varphi) \prod_{i=1}^{m} dr_i\, d\omega(\varphi_i)$$

where

$$\begin{aligned} Z_\alpha(r, \varphi) &= \int_0^\infty e^{-\frac{\alpha}{2}\sum_{i=1}^m \rho_i^{-2q} r_i^2 + \rho_i^2} \prod_{i=1}^{m} \rho_i^{n-q-1}\, d\rho_i \\ &= \prod_{i=1}^{m} \int_0^\infty e^{-\frac{\alpha}{2}(\rho_i^{-2q} r_i^2 + \rho_i^2)} \rho_i^{n-q-1}\, d\rho_i. \end{aligned}$$

Taking the $T \to 0$ limit gives

$$R_{\text{eff}}(r, \varphi) = \frac{1}{2} \inf_{\rho} \sum_{i=1}^{m} \frac{r_i^2}{\rho_i^{2q}} + \rho_i^2.$$

Each term in the sum can be minimized independently, with

$$\rho_i = (\sqrt{q}|r_i|)^{\frac{1}{q+1}} \quad \Rightarrow \quad R_{\text{eff}}(r, \varphi) = \frac{1+q}{2q^{\frac{q}{q+1}}} \sum_{i=1}^{m} |r_i|^{\frac{2}{q+1}}$$

This is minimized at $r = 0$. For example, for $q = 1$ (such as for linear or ReLU activations), we get $R_{\text{eff}}(r, \varphi) = \|r\|_1$.

To obtain the $\lambda \to 0$ limit, we analyze the scaling of the integral for a single neuron,

$$I_\alpha(r) = \int_0^\infty e^{-\frac{\alpha}{2}(r^2\rho^{-2q} + \rho^2)} \rho^{n-q-1}\, d\rho,$$

where we now write $r, \rho > 0$, dropping subscripts for clarity. As $\alpha \to 0$, the exponential gives the cutoff

$$\alpha\rho^{-2q} \lesssim 1, \qquad \alpha\rho^2 \lesssim 1,$$

that is,

$$\alpha^{\frac{1}{q}} \lesssim \rho^2 \lesssim \frac{1}{\alpha}.$$

Thus, changing variables to $y = \alpha\rho^2$, we get $dy = 2\alpha\rho\, d\rho$ and

$$\begin{aligned} I_\alpha(r) &= \int_0^\infty e^{-\frac{1}{2}(\alpha^{q+1} r^2 y^{-q} + y)} \left(\frac{y}{\alpha}\right)^{\frac{n-q-1}{2}} \frac{1}{2\sqrt{\alpha y}}\, dy \\ &= \frac{1}{2}\alpha^{-\frac{n-q}{2}} \int_0^\infty e^{-\frac{1}{2}(\alpha^{q+1} r^2 y^{-q} + y)} y^{\frac{n-q-2}{2}}\, dy. \end{aligned}$$

When $n > q$, we can set $\alpha = 0$ in the integral to get

$$I_\alpha(r) \sim 2^{\frac{n-q-2}{2}} \alpha^{-\frac{n-q}{2}} \Gamma\left(\frac{n-q}{2}\right),$$

which is independent of $r$. If $n = q$, then the integral scales like $q^{-1}\log(r^{-2}\alpha^{-q-1})/2$, which gives a logarithmic correction vanishing in the renormalized $\alpha \to 0$ limit. Finally, if $n < q$, changing variables to $y = r^{2/q}\alpha^{(q+1)/q} z$ gives

$$I_\alpha(r) \sim \frac{1}{q} 2^{-\frac{n+q}{2q}} \alpha^{-\frac{q-n}{2q}} \Gamma\left(\frac{q-n}{2q}\right) |r|^{-\frac{q-n}{q}},$$

and we obtain the renormalized entropy

$$\log \Omega(r, \varphi) = -\mathbf{1}_{q>n} \frac{q-n}{q} \sum_{i=1}^m \log |r_i|.$$

We think of $q$ as usually being small (1 for ReLU or 2 for a quadratic activation), and thus we predict that there is no symmetry-induced implicit bias in neuron magnitude or direction in networks of non-trivial width.

### C.6 Functional entropy

**Hausdorff measure** Our definition of renormalized entropy over functions requires a measure $df$, ideally one which we can informally think of as uniform. We use a standard construction, by picking a distribution over inputs $x \sim Q$ and considering functions $f$ as members of $L^2(Q)$. On a regular region $S \subseteq \Theta$ on which $\text{rank}(D\pi) = k$, the image $\pi(S)$ is a $k$-dimensional manifold. For a local chart $\phi : z \in \mathbb{R}^k \mapsto f$, the volume form is

$$d\mathcal{H}^k(f) = \sqrt{\det G_Q(z)}\, dz, \qquad (G_Q)_{ij} = \langle \partial_i \phi, \partial_j \phi \rangle_{L^2(Q)}.$$

Noting $\pi = \phi \circ \iota$ where $\iota : \theta \mapsto z$ take parameters to the effective chart, and applying the coarea formula (Simon, 2018), give the partition function in the main text.

**Two layer homogeneous networks** Here we build on Appendix C.5 to analyze the entropy $\log \Omega(f)$ with respect to $df$. Recall that $f(x) = \sum_i u_i \sigma(v_i^\top x) = \sum_i r_i \sigma(\varphi_i^\top x)$ for $r_i = u_i \|v_i\|^q$ and $\varphi_i = v_i/\|v_i\|$, where $\sigma$ is $q$-homogeneous activation function. Consider the generic $n > q$ case. For the $(r, \varphi)$ parameterization above, the feature derivatives are given by

$$\frac{\partial f(x)}{\partial r_i} = \sigma(\varphi_i^\top x), \qquad D_{\varphi_i} f(x)[b_i] = r_i \langle \nabla_\varphi^S \sigma(\varphi^\top x), b_i \rangle$$

where $b_i \in T_{\varphi_i} S^{n-1}$ is any infinitesimal change in $\varphi_i$, and

$$\nabla_\varphi^S \sigma(\varphi^\top x) = \sigma'(\varphi^\top x)(I - \varphi\varphi^\top)x$$

is the spherical gradient, assuming $\sigma$ is almost everywhere differentiable. Let $\tilde{K}(\varphi)$ be the Gram matrix of the *unscaled* features,

$$\tilde{K}_{\sigma,Q}(\varphi) = \mathbb{E}_{x\sim Q}\left[\left(\sigma(\varphi_i^\top x), \sigma'(\varphi_i^\top x) B_i^\top x\right)_{i=1}^m \otimes \left(\sigma(\varphi_j^\top x), \sigma'(\varphi_j^\top x) B_j^\top x\right)_{j=1}^m\right],$$

where $B_i \in \mathbb{R}^{n\times(n-1)}$ are any orthonormal frames for the tangent spaces at $\varphi_i$, meaning $B_i^\top B_i = I_{n-1}$ and $B_i^\top \varphi_i = 0$. (The spectral structure of $\tilde{K}$ is independent of the choice of $B_i$.) The effective parameter NTK (or rather, its transpose) is given by

$$K(r,\varphi) = D(r)\tilde{K}_{\sigma,Q}(\varphi)D(r), \qquad D(r) = \mathrm{diag}(1, r_1 I_{n-1}, \ldots, 1, r_m I_{n-1}),$$

and therefore

$$\sqrt{\det K(r,\varphi)} = \sqrt{\det \tilde{K}_{\sigma,Q}(\varphi)} \prod_i |r_i|^{(n-1)},$$

(assuming no zero eigenvalues for notational simplicity). This gives the result stated in the main text.

As we stated therein, for ReLU networks, when all $r_i \neq 0$ and $\varphi_i \neq \pm\varphi_j$ for all $i \neq j$, the unordered collection of $(r_i, H_i)$ for hyperplanes $H_i = \{x : \varphi_i^\top x = 0\}$ is determined by $f$. This is because $\nabla f$ changes when crossing a hyperplane $H_i$, with magnitude identifying $|r_i|$ and convexity determining the sign. This allows us to more precisely analyze the angular term elided in the main text. First,

$$\Omega(f) = \prod_i |r_i|^{-(n-1)} \sum_{\varphi:\bar{\pi}(r,\varphi)=f} \frac{1}{\sqrt{\det \tilde{K}_{\sigma,Q}(\varphi)}}.$$

Since there are finitely many choices of orientations of $\varphi_i$ matching the collection of $H_i$,

$$\log \sum_{\varphi:\bar{\pi}(r,\varphi)=f} \frac{1}{\sqrt{\det \tilde{K}_{\sigma,Q}(\varphi)}} = -\frac{1}{2}\log \Delta(f) + O(1)$$

where we define the angular term

$$\Delta(f) = \min_{\varphi:\bar{\pi}(r,\varphi)=f} \det \tilde{K}_{\sigma,Q}(\varphi).$$

Therefore

$$\log \Omega(f) = (n-1)\sum_i \log \frac{1}{|r_i|} - \frac{1}{2}\log \Delta(f) + O(1).$$

If additionally $Q$ is rotationally invariant, then $\det \tilde{K}(\varphi)$ depends only on the collection of angles $\arccos(\varphi_i^\top \varphi_j)$. In particular on can show a single vanishing angle of order $\epsilon_{ij}$ contributes $(n+2)/2\ \log(1/\epsilon_{ij})$ to the entropy.

### C.7 Subleading low-temperature terms and the Menon-Yu entropy

**Subleading terms** Consider the square $n$-dimensional linear network $f(x) = UV^\top$ with parameters $\theta = (U, V)$, and let $W = UV^\top$ with singular value decomposition $W = PSQ^\top$ where $S = \mathrm{diag}(s_1, \ldots, s_n)$. Recall that

$$Z_\alpha(W) \propto \int_{X\succ 0} e^{-\frac{\alpha}{2}\mathrm{Tr}(S(X+X^{-1}))}|X|^{-(n+1)/2}\, dX.$$

We begin by finding the $T \to 0$ (that is, $\alpha \to \infty$) expansion by Laplace's method, including subleading terms. As usual, the leading term is found by noting the minimizer of the exponent, $X = I$. Now we also consider a symmetric perturbation $A = A^\top$ (the tangent space to the positive definite matrices at $I$ being $\mathrm{Sym}(n)$)

$$X = I + \alpha^{-1/2}A \qquad \Longrightarrow \qquad X^{-1} = I - \alpha^{-1/2}A + \alpha^{-1}A^2 + O(\alpha^{-3/2}).$$

The exponent thus has the expansion

$$\frac{\alpha}{2}\mathrm{Tr}(S(X+X^{-1})) = \alpha\,\mathrm{Tr}(S) + \frac{1}{2}\mathrm{Tr}(SA^2) + O(\alpha^{-1/2}).$$

While the integrating measure transforms as

$$|X|^{-(n+1)/2} = 1 + O(\alpha^{-1/2}), \qquad dX = \alpha^{-n(n+1)/4}\, dA.$$

Applying Laplace's method, as $\alpha \to \infty$,

$$Z_\alpha(W) = \alpha^{-n(n+1)/4} e^{-\alpha\|W\|_*} \int_{\mathrm{Sym}(A)} e^{-\frac{1}{2}\operatorname{Tr}(SA^2)}\, dA\, (1 + o(1)).$$

But since $A$ is symmetric

$$\operatorname{Tr}(SA^2) = \sum_i s_i A_{ii}^2 + \sum_{i<j} (s_i + s_j) A_{ij}^2,$$

and the Gaussian integral factorizes

$$\int_{\mathrm{Sym}(A)} e^{-\frac{1}{2}\operatorname{Tr}(SA^2)}\, dA = \left(\prod_i \int e^{-s_i h^2/2}\, dh\right) \left(\prod_{i<j} \int e^{-(s_i+s_j)h^2/2}\, dh\right).$$

Up to a constant factor, this is $|W|^{-1/2} \prod_{i<j} (s_i + s_j)^{-1/2}$, and so

$$Z_\alpha(W) \sim \alpha^{-n(n+1)/4} \exp\left(-\alpha\|W\|_* - \frac{1}{2}\log|W| - \frac{1}{2}\sum_{i<j}\log(s_i + s_j)\right)(1 + o(1)).$$

The subleading term, describing fluctuations in the fiber, is thus $-\frac{1}{2}\log|W| - \frac{1}{2}\sum_{i<j}\log(s_i + s_j)$.

**Relationship to Menon and Yu's entropy** Menon & Yu (2025) focus on the *balanced manifold*

$$\mathcal{M} = \left\{(U, V) : U^\top U = V^\top V\right\}.$$

The ambient space of $\theta$ is Euclidean with Frobenius inner product, and the balanced manifold inherits the corresponding surface-volume measure $d\mathrm{vol}_{\mathcal{M}}$. Menon and Yu then consider the piece of the fiber $\pi^{-1}(W)$ containing balanced factorizations

$$\mathcal{O}_W = \pi^{-1}(W) \cap \mathcal{M}.$$

Theorem 4 of Menon & Yu (2025) characterizes the entropy

$$S_{MY}(W) := \log \mathrm{vol}(\mathcal{O}_W) = \frac{1}{2}\sum_{i<j}\log(s_i + s_j) + C$$

where the volume of $\mathcal{O}_W$ is also induced by the ambient Euclidean metric. This appears related to, but is not quite, the subleading term we found above. This issue is a change of measure. Our measure over effective parameters is $Z_\alpha(W)\, dW$. The natural measure for Menon and Yu is instead their $d\mathrm{vol}_g(W)$, induced by their metric $g$. This is the metric corresponding to the NTK in our introduction, which turns parameter space gradient flow into a metric gradient flow in $W$. From their equation (3.13),

$$g_W(A, A) = \sum_i \frac{A_{ii}^2}{2s_i} + \sum_{i \neq j} \frac{A_{ij}^2}{s_i + s_j},$$

where we write $A$ in the $p_i q_j^\top$ basis: $A_{ij} = p_i^\top A q_j$. As this basis is orthonormal with respect to the Frobenius inner product inducing $dW$,

$$d\mathrm{vol}_g(W) = \left(\prod_i \frac{1}{2s_i}\right)^{1/2} \left(\prod_{i \neq j} \frac{1}{s_i + s_j}\right)^{1/2} dW \propto |W|^{-1/2} \prod_{i<j} (s_i + s_j)^{-1}\, dW.$$

Combining terms and defining the appropriate constant $C_\alpha$ we find that

$$Z_\alpha(W)\, dW \sim C_\alpha \exp\left(-\alpha\|W\|_* + S_{MY}(W)\right)\, d\mathrm{vol}_g(W).$$

**Interpretation as balanced parameter volume** Following our main text, we define the measure in effective parameter space counting balanced representatives,

$$\nu_{\mathcal{M}}(B) = \int_{\pi^{-1}(B)\cap\mathcal{M}} d\mathrm{vol}_{\mathcal{M}}(\theta).$$

(No regularization is needed here, since $\mathcal{O}_W$ is compact.) Menon and Yu's Theorem 8 states that the restriction $\pi|_{\mathcal{M}}$ is a Riemannian submersion when the image is given their metric $g$. By the coarea formula, it follows that

$$\nu_{\mathcal{M}}(dW) = \mathrm{vol}(\mathcal{O}_W) d\mathrm{vol}_g(W) = e^{S_{MY}(W)} d\mathrm{vol}_g(W).$$

What our argument above then shows is that

$$Z_\alpha(W)\, dW \sim C_\alpha \exp(-\alpha\|W\|_*)\, \nu_{\mathcal{M}}(dW).$$

That is, at low temperature, fluctuations in $W$ are governed exactly by movement in the balanced manifold.

# D Experiments

## D.1 Diagonal networks

Our figures are generated using depth two, with $\ell(w) = (w-1)^2/2$, temperature $T = 0.25$, Euler–Maruyama step size $10^{-3}$, and standard Gaussian initialization for $(u, v)$. The unregularized experiment runs 6,000 independent chains and records one product per chain at $t = 0.5, 1, 5, 20$, with effective precision $\alpha_{\text{eff}}(t) = (1 + 2Tt)^{-1}$. The finite-regularization run uses 256 chains, 8,000 steps, a 2,000-step burn-in and thinning interval 20. All predicted densities at depth two use the closed form $K_0(\alpha|w|)$.

Our code allows for deeper network simulations. These display the same agreement between prediction and experiment. Increasing depth, however, increases the curvature of the loss landscape, leading to instability in the Euler–Maruyama algorithm as time increases.

## D.2 Linear networks

Our figures are generated using $2 \times 2$ matrices $U, V$ initialized with independent $\mathcal{N}(0, 0.45)$ entries, with target $W_* = diag(1.25, 0.55)$, loss $\ell(W) = \|W - W_*\|_F^2/2$, temperature $T = 0.25$, and Euler-Maruyama step size $5 \times 10^{-4}$. The unregularized experiment uses 6,000 chains and snapshots at $t = 0.25, 0.5, 2.5, 10$, and $\alpha_{\text{eff}}(t) = (0.45^2 + 2Tt)^{-1}$. The finite-regularization run uses 192 chains, 10,000 steps, burn-in 3,000, and thinning 25. Predicted $W$ densities are plotted by drawing $2 \times 10^6$ independent $U, V$ pairs with precision $\alpha$, and importance-weighting $W = UV^\top$ by $e^{-\ell(W)/T}$. The $\alpha \to 0$ limit is plotted by sampling $W = W_* + \sqrt{T}\xi$ for standard normal $\xi$, and weighting it by the known closed form in the $2 \times 2$ case,

$$\Omega(W) = \frac{2}{\pi s_2} K\left(1 - \frac{s_1^2}{s_2^2}\right)$$

where $K$ is the complete elliptic integral. We clip $s_2$ at $10^{-12}$ for numerical stability. Sampled densities plotted as lines use Gaussian smoothing with Silverman bandwidth.

As with diagonal networks, experiments in larger dimensions behave in accordance with the $2 \times 2$ case, up to numerical instability.

## D.3 Neural heat death

We simulate a fully connected, bias-free ReLU network with scalar input, three hidden layers of width three, and a linear scalar readout. All weights are initialized independently from $\mathrm{Uniform}(-\sqrt{3}/2, \sqrt{3}/2)$, with standard deviation $0.5$. The training inputs are $x = -1, +1$, with respective targets $0.5, 1$, and the loss is

$$L(\theta) = \frac{1}{4}\left[\big(f_\theta(-1) - 0.5\big)^2 + \big(f_\theta(1) - 1\big)^2\right].$$

Starting directly from these random initializations, we evolve 128 independent unregularized Langevin trajectories at constant temperature $T = 0.2$ until $t = 80$, with snapshots at $t = 0, 3.8, 80$. We use adaptive Euler–Maruyama integration with maximum step size $10^{-3}$, curvature-safety parameter $0.0125$, and drift-increment parameter $0.0025$.

A neuron is dead when its activation is identically zero over the input domain $\mathcal{X}$; the network is dead when $f_\theta(x) = 0$ for every $x \in \mathcal{X}$. Heatmaps show each neuron's empirical death probability across the 128 trajectories. The time-dependent curves show the fraction of networks that are dead and the ensemble mean loss. Initially dead networks are retained, and death is not treated as absorbing. Shaded regions are nominal pointwise 95% sampling intervals: Wilson intervals for network deadness and Student-$t$ intervals for mean loss.