Can Stationary Distributions of Scale-Invariant Neural Networks Be Described by the Thermodynamics of an Ideal Gas?

Ildus Sadrtdinov[1], **Ekaterina Lobacheva**[2,3], **Ivan Klimov**[1], **Mikhail Burtsev**[4], **Mikhail I. Katsnelson**[1,5] and **Dmitry Vetrov**[1]
[1]Constructor University
[2]Mila – Quebec AI Institute
[3]Université de Montréal
[4]London Institute for Mathematical Sciences
[5]Institute for Molecules and Materials, Radboud University
isadrtdinov@constructor.university

Abstract

Understanding the training dynamics of deep neural networks remains a major open problem, with physics-inspired approaches offering promising insights. Building on this perspective, we develop a thermodynamic framework to describe the stationary distributions of stochastic gradient descent (SGD) with weight decay for scale-invariant neural networks, a setting that both reflects practical architectures with normalization layers and permits theoretical analysis. We establish analogies between training hyperparameters (e.g., learning rate, weight decay) and thermodynamic variables such as temperature, pressure, and volume. Starting with a simplified isotropic noise model, we uncover a close correspondence between SGD dynamics and ideal gas behavior, validated through theory and simulation. Extending to training of neural networks, we show that key predictions of the framework, including the behavior of stationary entropy, align closely with experimental observations. This framework provides a principled foundation for interpreting training dynamics and may guide future work on hyperparameter tuning and the design of learning rate schedulers.

1 Introduction

The optimization dynamics of deep neural networks remain poorly understood despite their empirical success. A promising approach is to draw inspiration from physics, which studies systems with many degrees of freedom and complex interactions. In particular, thermodynamics has motivated numerous connections to neural networks [Jastrzębski *et al.*, 2017; Chen *et al.*, 2024; Zhang, 2024; Kozyrev, 2025]. In this work, we extend this line of research and propose a framework that relates neural network optimization to thermodynamic systems. Specifically, we interpret stochastic gradient noise as thermal fluctuations and introduce macroscopic thermodynamic variables, like temperature, pressure, and volume, to describe stationary distributions of stochastic gradient descent (SGD). We also connect these variables to training hyperparameters such as learning rate and weight decay.

We focus on *scale-invariant* neural networks, which resemble common architectures with normalization layers (e.g., BatchNorm [Ioffe and Szegedy, 2015], LayerNorm [Ba *et al.*, 2016]), while remaining theoretically tractable. Theories not accounting for scale invariance would fail to explain phenomena observed in practical training. Unlike prior works that focused on energy, entropy, and temperature, we show that SGD with weight decay on scale-invariant networks naturally introduces pressure and volume, establishing a direct analogy with the ideal gas law. Our **main contributions** are:

1. We study scale-invariant neural networks under three training protocols: (1) fixed parameter norm, (2) fixed effective learning rate (ELR), and (3) fixed learning rate (LR). For each setting, we derive the corresponding stochastic differential equation (SDE) that governs the dynamics and characterize their stationary distributions.
2. We introduce a thermodynamic framework for analyzing stationary distributions of SGD. Under a simplified *isotropic noise model*, we establish the thermodynamic analogy rigorously and validate it empirically, revealing a close correspondence with the ideal gas model.
3. We discuss how the ideal gas model generalizes to actual training of scale-invariant neural networks and show that our framework accurately predicts how the entropy of the stationary distribution varies with learning rate and weight decay.

Overall, our framework reveals a deep link between optimization and thermodynamics, providing a physics-based foundation for interpreting neural network training. This insight can inform future work on training hyperparameters, including learning rate scheduling[1].

2 Background

In this section, we outline the main concepts from thermodynamics and the optimization of scale-invariant neural net-

[1]Full paper is available at https://arxiv.org/abs/2511.07308

works that form the basis of our analogy. For convenience, a complete list of notations is provided in Appendix A. A broader overview of related work is given in Appendix B.

Thermodynamics. Below we summarize the key principles of thermodynamics relevant to our discussion. A more detailed introduction is given in Appendix C.

- **T1** A *thermodynamic system* is described by *state variables*, including *internal energy* U (the total energy of all particles in the system), *entropy* S (a measure of disorder), *temperature* T, *pressure* p, and *volume* V.
- **T2** The *First Law of Thermodynamics* expresses energy conservation: $\mathrm{d}U = \delta Q - p\,\mathrm{d}V$, where δQ is the infinitesimal heat supplied to the system. The *Second Law of Thermodynamics* requires $\mathrm{d}S \geq \delta Q/T$, implying that systems evolve toward *equilibrium*, where an appropriate *thermodynamic potential* is minimized (**T4**, **T5**). A *reversible process* satisfies $\mathrm{d}S = \delta Q/T$.
- **T3** The *Gibbs distribution* connects thermodynamics with statistical mechanics. At equilibrium, the probability of microstate i with energy E_i is $p_i \propto \exp(-E_i/T)$, with the temperature T controlling the distribution's spread. Internal energy and entropy can be expressed as $U = \mathbb{E}_{p_i}[E_i]$ and $S = \mathbb{E}_{p_i}[-\log p_i]$.
- **T4** At fixed T and V, the *Helmholtz energy* $F = U - TS$ is minimized at equilibrium. The *Maxwell relation* links derivatives of state variables and helps to determine changes in entropy, which is otherwise difficult to measure. In this case, it is given by $\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial p}{\partial T}\right)_V$ [2].
- **T5** At fixed T and p, the *Gibbs energy* $G = U - TS + pV$ is minimized at equilibrium, with Maxwell relation $-\left(\frac{\partial S}{\partial p}\right)_T = \left(\frac{\partial V}{\partial T}\right)_p$.
- **T6** An *ideal gas* is a simplified model neglecting intermolecular interactions. At equilibrium, its state variables satisfy the *ideal gas law* $pV = RT$, where R is the gas constant.
- **T7** The *isochoric* and *isobaric heat capacities* are $C_V = \left(\frac{\delta Q}{\mathrm{d}T}\right)_V$ and $C_p = \left(\frac{\delta Q}{\mathrm{d}T}\right)_p$. For an ideal gas, both are constants with $C_p - C_V = R$. In an *adiabatic process*, defined by $\delta Q = 0$, the ideal gas follows $pV^\gamma = \text{const}$ with $\gamma = C_p/C_V$.

Stationary behavior of SGD. In classical optimization theory, SGD differs from full-batch gradient descent: instead of converging to a single stationary point of the loss function, SGD converges to a *stationary distribution* centered around it. We denote this distribution by $\rho_{\boldsymbol{w}}(\boldsymbol{w})$, where $\boldsymbol{w} \in \mathbb{R}^d$ are the model parameters. This behavior is driven by the finite learning rate and stochastic gradient estimates, both of which inject noise into the dynamics, analogous to thermal fluctuations and naturally motivating a thermodynamic perspective.

Scale-invariant networks. Modern architectures often include normalization layers that make preceding parameters *scale-invariant*, i.e., the output of the normalization layer is independent of the parameter norm. We focus on fully scale-invariant networks, where the whole output of the network does not depend on the parameter norm (i.e., only on the unit direction vector of the weights). The dynamics of the model on the unit sphere (i.e., the loss evolution) is governed by the *effective learning rate* (ELR), defined as $\eta_{\text{eff}} = \eta/\|\boldsymbol{w}\|^2$, where η is the usual learning rate (LR) used in SGD iterates.

3 Stationary Distributions of SDE

In this section, we start from a stochastic differential equation (SDE) introduced in prior work [Li *et al.*, 2020; Wang and Wang, 2022] that captures the training dynamics of scale-invariant neural networks. We adapt this SDE to three training protocols of increasing complexity and derive their stationary distributions under an *isotropic noise model*, introduced below. Although these stationary distributions have been studied previously, we present them in a unified notation and systematize the results across the three protocols.

We begin by introducing several assumptions that enable a theoretical analysis of SGD. First, we approximate the stochastic gradient noise by a Gaussian random vector, a standard assumption in prior work [Jastrzębski *et al.*, 2017; Mandt *et al.*, 2017] naturally leading to an SDE formulation.

Assumption 1 (Gaussian noise)**.** *Let $L(\boldsymbol{w})$ and $L_{\mathcal{B}}(\boldsymbol{w})$ denote the (full-batch) training loss and the loss on a mini-batch $\mathcal{B}$, respectively. Then,*

$$\nabla L_{\mathcal{B}}(\boldsymbol{w}) \approx \nabla L(\boldsymbol{w}) + \left(\boldsymbol{\Sigma}_{\boldsymbol{w}}\right)^{1/2}\boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ is a spatially dependent covariance matrix of stochastic gradients[3]*.*

Second, we focus on scale-invariant networks, which induce scale invariance of the loss function.

Assumption 2 (Scale invariance)**.** *For any mini-batch $\mathcal{B}$ (including the full dataset), the loss satisfies $L_{\mathcal{B}}(\alpha\boldsymbol{w}) = L_{\mathcal{B}}(\boldsymbol{w})$ for all $\alpha > 0$.*

Let $r = \|\boldsymbol{w}\|_2$ and $\overline{\boldsymbol{w}} = \boldsymbol{w}/r$. Then $L(\boldsymbol{w}) = L(\overline{\boldsymbol{w}})$. As shown in [Li *et al.*, 2020] (Lemmas 3.1 and B.1), the full-batch gradient and the covariance matrix of a scale-invariant loss satisfy

$$\nabla L(\boldsymbol{w}) = \nabla L(\overline{\boldsymbol{w}})/r \qquad \boldsymbol{w}^T \nabla L(\boldsymbol{w}) = 0 \tag{2}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}}/r^2 \qquad \left(\boldsymbol{\Sigma}_{\boldsymbol{w}}\right)^{1/2}\boldsymbol{w} = 0 \tag{3}$$

Next, we assume that the discrete-time dynamics of SGD can be approximated by a continuous-time SDE.

Assumption 3 (SDE approximation)**.** *For sufficiently small (effective) learning rate and weight decay coefficient, the SGD dynamics in the weight space are well approximated by the corresponding continuous SDE.*

Finally, we introduce a simplified *isotropic noise model*, which yields explicit stationary distributions of the resulting SDE. A related assumption has been considered in [Jastrzębski *et al.*, 2017; Wang and Wang, 2022].

[2] $\left(\frac{\partial f}{\partial x}\right)_y$ denotes the partial derivative of f w.r.t. x under fixed y.

[3] We treat $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ as implicitly dependent on the batch size and therefore do not model the batch size explicitly.

Assumption 4 (Isotropic noise model)**.** *The covariance matrix at a unit vector $\overline{\boldsymbol{w}}$ takes a form $\boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}} = \boldsymbol{P}_{\overline{\boldsymbol{w}}}\boldsymbol{\Sigma}\boldsymbol{P}_{\overline{\boldsymbol{w}}}$, where $\boldsymbol{P}_{\overline{\boldsymbol{w}}} = \boldsymbol{I}_d - \overline{\boldsymbol{w}}\,\overline{\boldsymbol{w}}^T$ projects onto the tangent subspace orthogonal to $\overline{\boldsymbol{w}}$, ensuring the constraints of Eq. 3, and $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_d$ for some $\sigma > 0$. Hence, $\boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}} = \sigma^2\boldsymbol{P}_{\overline{\boldsymbol{w}}}$ and $\boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}}^{1/2} = \sigma\boldsymbol{P}_{\overline{\boldsymbol{w}}}$.*

With these assumptions in place, we analyze three training protocols and present their corresponding discrete dynamics, SDEs, and stationary distributions. In the main text, we report only the expressions characterizing the stationary distributions; detailed derivations are deferred to Appendix D.

SGD on a Fixed Sphere

A natural way to handle scale-invariant parameters is to optimize them directly on their intrinsic domain, a sphere of fixed radius. This can be implemented by projecting the weights onto the sphere after each update [Kodryan *et al.*, 2022; Loshchilov *et al.*, 2025] or by using Riemannian optimization methods [Cho and Lee, 2017]. Here, we consider projected SGD on the sphere $\mathbb{S}^{d-1}(r)$ with discrete dynamics

$$\boldsymbol{w}_{k+1} = \mathrm{proj}_{\mathbb{S}^{d-1}(r)}\Big(\boldsymbol{w}_k - \eta\nabla L_{\mathcal{B}_k}(\boldsymbol{w}_k)\Big), \tag{4}$$

where η is the learning rate and $\mathcal{B}_k$ is the mini-batch sampled at iteration k. Taking the continuous-time limit of Eq. 4 yields the SDE governing the direction vector $\overline{\boldsymbol{W}}_t$:

$$\mathrm{d}\overline{\boldsymbol{W}}_t = -\left[\eta_{\mathrm{eff}}\nabla L(\overline{\boldsymbol{W}}_t) + \tfrac{\eta_{\mathrm{eff}}^2}{2}\,\mathrm{Tr}\,(\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t})\overline{\boldsymbol{W}}_t\right]\mathrm{d}t + \eta_{\mathrm{eff}}(\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t})^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t, \tag{5}$$

where $\boldsymbol{B}_t$ denotes standard Brownian motion in $\mathbb{R}^d$ and $\eta_{\mathrm{eff}} = \eta/r^2$ is the effective learning rate (ELR). While [Wang and Wang, 2022] analyze the Riemannian version of this SDE, we remain with the Euclidean formulation.

Proposition 1 (Fixed sphere)**.** *Under Assumptions 1-4, the stationary distribution on the unit sphere $\rho^*_{\overline{\boldsymbol{w}}}$ of Eq. 5 is*

$$\rho^*_{\overline{\boldsymbol{w}}}(\overline{\boldsymbol{w}}) \propto \exp\Big(-\frac{L(\overline{\boldsymbol{w}})}{\tau_{\mathrm{eff}}}\Big),\ \text{with}\ \tau_{\mathrm{eff}} = \frac{\eta_{\mathrm{eff}}\sigma^2}{2} \tag{6}$$

SGD With Fixed ELR

We next allow the radius to evolve and explicitly include weight decay into the dynamics. We consider training with a fixed ELR, which makes the loss dynamics independent of the parameter norm. This is achieved by scaling the learning rate with the squared norm of the parameters, $\eta = \eta_{\mathrm{eff}}\|\boldsymbol{w}_k\|^2$. Although less common, such ELR control has been used in practice [Lyle *et al.*, 2024; Ali Mehmeti-Göpel and Wand, 2024], for example to enhance plasticity in reinforcement learning. The discrete update rule is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta_{\mathrm{eff}}\|\boldsymbol{w}_k\|^2\Big(\nabla L_{\mathcal{B}_k}(\boldsymbol{w}_k) + \lambda\boldsymbol{w}_k\Big), \tag{7}$$

where λ is the weight decay (WD) coefficient. The corresponding SDE is

$$\mathrm{d}\boldsymbol{W}_t = -\eta_{\mathrm{eff}}\|\boldsymbol{W}_t\|^2\Big(\nabla L(\boldsymbol{W}_t) + \lambda\boldsymbol{W}_t\Big)\mathrm{d}t + \eta_{\mathrm{eff}}\|\boldsymbol{W}_t\|^2(\boldsymbol{\Sigma}_{\boldsymbol{W}_t})^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t \tag{8}$$

Applying Ito's formula separates the dynamics of the direction vector $\overline{\boldsymbol{W}}_t$ and the radius r_t:

$$\mathrm{d}\overline{\boldsymbol{W}}_t = -\left[\eta_{\mathrm{eff}}\nabla L(\overline{\boldsymbol{W}}_t) + \tfrac{\eta_{\mathrm{eff}}^2}{2}\,\mathrm{Tr}\,\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t}\cdot\overline{\boldsymbol{W}}_t\right]\mathrm{d}t + \eta_{\mathrm{eff}}(\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t})^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t \tag{9}$$

$$\tfrac{\mathrm{d}r_t}{\mathrm{d}t} = -\eta_{\mathrm{eff}}\lambda r_t^3 + \tfrac{\eta_{\mathrm{eff}}^2}{2}r_t\,\mathrm{Tr}\,\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t} \tag{10}$$

Importantly, the radius follows an ordinary differential equation and therefore evolves deterministically. Under the isotropic noise model, $\mathrm{Tr}\,\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t} = \sigma^2(d-1)$ is constant. Consequently, the system converges to a stationary distribution for $\overline{\boldsymbol{w}}$ on the unit sphere and a fixed stationary radius r^*.

Proposition 2 (Fixed ELR)**.** *Under Assumptions 1-4, the stationary distribution $\rho^*_{\overline{\boldsymbol{w}}}$ on the unit sphere of Eq. 9 and stationary radius r^* of Eq. 10 are*

$$\rho^*_{\overline{\boldsymbol{w}}}(\overline{\boldsymbol{w}}) \propto \exp\Big(-\frac{L(\overline{\boldsymbol{w}})}{\tau_{\mathrm{eff}}}\Big),\ \text{with}\ \tau_{\mathrm{eff}} = \frac{\eta_{\mathrm{eff}}\sigma^2}{2} \tag{11}$$

$$r^* = \sqrt{\frac{\tau_{\mathrm{eff}}(d-1)}{\lambda}} = \sqrt{\frac{\eta_{\mathrm{eff}}\sigma^2(d-1)}{2\lambda}} \tag{12}$$

SGD With Fixed LR

Finally, we consider standard training with a fixed LR. In this case, the dynamics on the unit sphere depends explicitly on the parameter norm. The SGD update is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta\Big(\nabla L_{\mathcal{B}_k}(\boldsymbol{w}_k) + \lambda\boldsymbol{w}_k\Big), \tag{13}$$

which corresponds to the SDE

$$\mathrm{d}\boldsymbol{W}_t = -\eta\Big(\nabla L(\boldsymbol{W}_t) + \lambda\boldsymbol{W}_t\Big)\mathrm{d}t + \eta(\boldsymbol{\Sigma}_{\boldsymbol{W}_t})^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t \tag{14}$$

The induced direction $\overline{\boldsymbol{W}}_t$ and radius r_t dynamics are

$$\mathrm{d}\overline{\boldsymbol{W}}_t = -\left[\tfrac{\eta}{r_t^2}\nabla L(\overline{\boldsymbol{W}}_t) + \tfrac{\eta^2}{2r_t^4}\,\mathrm{Tr}\,\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t}\cdot\overline{\boldsymbol{W}}_t\right]\mathrm{d}t + \tfrac{\eta}{r_t^2}(\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t})^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t \tag{15}$$

$$\tfrac{\mathrm{d}r_t}{\mathrm{d}t} = -\eta\lambda r_t + \tfrac{\eta^2}{2r_t^3}\,\mathrm{Tr}\,\boldsymbol{\Sigma}_{\overline{\boldsymbol{W}}_t} \tag{16}$$

As in the fixed-ELR case, the radius evolves deterministically and converges to a fixed stationary value r^*.

Proposition 3 (Fixed LR)**.** *Under Assumptions 1-4, the stationary distribution $\rho^*_{\overline{\boldsymbol{w}}}$ of Eq. 15 and stationary radius r^* of Eq. 16 are*

$$\rho^*_{\overline{\boldsymbol{w}}}(\overline{\boldsymbol{w}}) \propto \exp\Big(-\frac{L(\overline{\boldsymbol{w}})}{\tau/(r^*)^2}\Big),\ \text{with}\ \tau = \frac{\eta\sigma^2}{2} \tag{17}$$

$$r^* = \sqrt[4]{\frac{\tau(d-1)}{\lambda}} = \sqrt[4]{\frac{\eta\sigma^2(d-1)}{2\lambda}} \tag{18}$$

4 Thermodynamic Framework

In this section, we introduce optimization quantities that correspond to thermodynamic state variables in **T1**. Our framework is based on the stationary distribution of the weights, $\rho^*_{\boldsymbol{w}}$, which we relate to the thermodynamic Gibbs distribution (**T3**). This correspondence naturally identifies thermodynamic microstates i with weight vectors $\boldsymbol{w}$.

We then define the analogues of temperature T, pressure p, and volume V so that the appropriate thermodynamic potential, Helmholtz energy F (**T4**) or Gibbs energy G (**T5**), is minimized at stationary. Concretely, we associate T with the noise level through the ELR and gradient variance, V with the radius, and p with the L2 regularization strength. This correspondence reflects an intuition that higher pressure leads to smaller volume, just as stronger weight decay leads to smaller parameter norms. These interpretations are confirmed by explicit minimization of the corresponding potentials.

Microstates and Weights

Under Assumptions 1-4, the stationary distribution $\rho^*_{\boldsymbol{w}}$ is supported on a sphere $\mathbb{S}^{d-1}(r^*)$. The radius r^* is either predefined (fixed-sphere protocol) or determined by the training hyperparameters (fixed-ELR and fixed-LR protocols). In this setting, the density on $\mathbb{S}^{d-1}(r^*)$ can be expressed in terms of a density on the unit sphere.

Lemma 1. *Let $\rho_{\boldsymbol{w}}$ be a density on the sphere $\mathbb{S}^{d-1}(r)$. Then,*

$$\rho_{\boldsymbol{w}}(\boldsymbol{w}) = \rho_{\overline{\boldsymbol{w}}}(\overline{\boldsymbol{w}})/r^{d-1}, \tag{19}$$

where $\rho_{\overline{\boldsymbol{w}}}$ is the corresponding density on the unit sphere. The associated spherical entropy $S(\rho_{\boldsymbol{w}}) = \mathbb{E}_{\rho_{\boldsymbol{w}}}[-\log \rho_{\boldsymbol{w}}(\boldsymbol{w})]$ is

$$S(\rho_{\boldsymbol{w}}) = S(\rho_{\overline{\boldsymbol{w}}}) + (d-1)\log r, \tag{20}$$

with $S(\rho_{\overline{\boldsymbol{w}}})$ denoting the spherical entropy of $\rho_{\overline{\boldsymbol{w}}}$.

As a consequence, $\rho^*_{\boldsymbol{w}}$ has the same functional form as $\rho^*_{\overline{\boldsymbol{w}}}$ and takes the Gibbs form $\rho^*_{\boldsymbol{w}} \propto \exp(-L(\boldsymbol{w})/T)$ for all three training protocols (Eqs. 6, 11, 17). This mirrors the equilibrium distribution of a thermodynamic system, $p_i \propto \exp(-E_i/T)$ (**T3**), motivating the identifications $i \leftrightarrow \boldsymbol{w}$, $E_i \leftrightarrow L(\boldsymbol{w})$, $U \leftrightarrow U(\rho_{\boldsymbol{w}}) = \mathbb{E}_{\rho_{\boldsymbol{w}}}[L(\boldsymbol{w})]$, and $S \leftrightarrow S(\rho_{\boldsymbol{w}})$.

Thermodynamic Potentials

We now define T, p, and V through the requirement that the corresponding potential is minimized at stationary.

Theorem 1 (Helmholtz energy minimization)**.** *Under Assumptions 1-4, consider SGD on a fixed sphere with radius r and ELR η_{eff}. Defining $T = \frac{\eta_{\text{eff}}\sigma^2}{2}$, the Helmholtz energy $F = U - TS$ is minimized at stationary distribution $\rho_{\overline{\boldsymbol{w}}}$ among all distributions on the unit sphere $\mathbb{S}^{d-1}$.*

Theorem 2 (Gibbs energy minimization)**.** *Under Assumptions 1-4, consider SGD with a fixed ELR η_{eff} (or a fixed LR η) and WD λ. Let $T = \frac{\eta_{\text{eff}}\sigma^2}{2}$ (or $T = \sqrt{\frac{\eta\lambda\sigma^2}{2(d-1)}}$), $p = \lambda$ and $V = \frac{r^2}{2}$. Then, the Gibbs energy $G = U - TS + pV$ is minimized at the stationary distribution $\rho_{\boldsymbol{w}}$ among all distributions supported on a sphere $\mathbb{S}^{d-1}(r)$ with deterministic radius $r > 0$.*

The proofs and a detailed discussion of the choice of p and V are provided in Appendix D. Importantly, the choice of potential follows thermodynamics: F applies when the volume (radius) is fixed, while G applies when the pressure (weight decay coefficient) is fixed. In both cases, the temperature T is set by constant hyperparameters and thus remains fixed. Therefore, our framework offers an alternative thermodynamic description of SGD stationary distributions[4].

Ideal Gas Law

Remarkably, inserting the definitions of T, p, and V into the stationary radius r^* (Eq. 12, 18) yields exactly the ideal gas law $pV = RT$ (**T6**) with $R = \frac{d-1}{2}$ both for fixed-ELR (Eq. 21) and fixed-LR (Eq. 22) training protocols. A similar expression also holds in the fixed-sphere case if we introduce "effective" weight decay as pressure (see Appendix D.4).

$$V = \frac{(r^*)^2}{2} = \frac{\eta_{\text{eff}}\sigma^2}{2} \cdot \frac{d-1}{2\lambda} = \frac{RT}{p} \tag{21}$$

$$V = \frac{(r^*)^2}{2} = \frac{1}{2}\sqrt{\frac{\eta\lambda\sigma^2}{2(d-1)}}\frac{d-1}{2\lambda} = \frac{RT}{p} \tag{22}$$

This correspondence is particularly striking given the very different microscopic assumptions. In statistical mechanics, the ideal gas law assumes non-interacting particles, so the energy of a microstate i decomposes as $E_i = \sum_j e_{ij}$, where e_{ij} is the energy of particle j. By contrast, the loss $L(\boldsymbol{w})$ typically exhibits strong parameter coupling and does not admit a decomposition of the form $L(\boldsymbol{w}) = \sum_j l_j(w_j)$. Nevertheless, under scale invariance, the stationary optimization dynamics obey the same equation of state. While such an assumption would be unnatural in physical systems, requiring interactions to depend on angular coordinates rather than distances, it is natural for neural networks and yields the ideal gas law despite strong parameter interactions.

5 Empirical Validation

To support our theoretical framework, we design experiments inspired by the thermodynamic analogy. We focus on tests that do not follow directly from the SDE formulation, providing strong evidence for the thermodynamic perspective.

(V1) Stationary radius. The SDE predicts that r^* scales as $\sqrt[4]{\eta/\lambda}$ when training with a fixed LR, and as $\sqrt{\eta_{\text{eff}}/\lambda}$ when training with a fixed ELR. Since this prediction follows directly from the SDE (without invoking thermodynamics), it primarily serves as a sanity check. At the same time, it also verifies the ideal gas law (**T6**) in our analogy.

(V2) Minimizing thermodynamic potentials. Our analogy predicts that stationary distribution of SGD minimizes the appropriate thermodynamic potential. Numerical experiments cannot test minimization over the entire space of distributions $\rho_{\overline{\boldsymbol{w}}}$ and radii r^*. Instead, we check whether the observed stationary states minimize F or G at least among the distributions induced by different hyperparameter settings.

For example, in fixed LR training, consider a set of hyperparameter configurations $\mathcal{H} = \{(\eta_i, \lambda_i)\}_{i=1}^h$. For each (η_i, λ_i), we train the model and measure the corresponding internal energy U_i, entropy S_i, and volume V_i. We then verify that for each $(\eta^*, \lambda^*) \in \mathcal{H}$:

$$(\eta^*, \lambda^*) = \underset{(\eta_i,\lambda_i)\in\mathcal{H}}{\operatorname{argmin}} \Big\{U_i - T^* S_i + p^* V_i\Big\}, \tag{23}$$

where $T^* = \sqrt{\frac{\eta^*\lambda^*\sigma^2}{2(d-1)}}$ and $p^* = \lambda^*$. A similar procedure applies for the settings with a fixed ELR and on a fixed sphere.

(V3) Maxwell relations. Maxwell relations describe how entropy S varies when thermodynamic variables change (**T4**, **T5**). In our setting, training with fixed LR corresponds to fixing p and T, so the relevant Maxwell relation is $\left(\frac{\partial S}{\partial p}\right)_T =$

[4]Instead of Assumption 3, one may postulate (1) minimization of the potentials at stationary and (2) deterministic radius (while preserving Assumptions 1, 2, 4), and obtain the same stationary distributions without appealing to the SDE.

$\left(\frac{\partial V}{\partial T}\right)_p$. In Appendix D.4, we show this is equivalent to

$$\left(\frac{\partial S}{\partial \log \eta}\right)_\lambda - \left(\frac{\partial S}{\partial \log \lambda}\right)_\eta = \frac{d-1}{2} \tag{24}$$

This elegant equation quantifies, how the two hyperparameters influence the stationary entropy. Similar relations for the fixed sphere and fixed ELR protocols are also derived there.

(V4) Adiabatic process. Previously, we interpreted convergence to a stationary distribution as a non-reversible thermodynamic process. Here, we instead view stationary distributions corresponding to different hyperparameters as states of a *reversible* process. This perspective enables a thermodynamic analysis of how hyperparameters shape the stationary distribution. We focus on the adiabatic process, defined by $\delta Q = 0$ (**T7**). Although heat Q does not appear explicitly in our analogy, it can be inferred from the First Law (**T2**) as $\delta Q = \mathrm{d}U + p\mathrm{d}V$. This allows us to define the isochoric and isobaric heat capacities, C_V and C_p, as in **T7**. In Appendix D.5, we show that $C_p - C_V = \frac{d-1}{2} = R$, matching the ideal gas relation. To simulate an adiabatic process, we vary (η, λ) jointly such that $pV^\gamma = \text{const}$, where $\gamma = C_p/C_V$. For a reversible adiabatic process, $\mathrm{d}S = \delta Q/T = 0$ (**T2**), implying constant entropy. Empirically, this predicts that varying hyperparameters along an adiabatic process gives stationary distributions of different shape supported on spheres of different radii but with identical entropy, i.e., overall "disorder."

6 Isotropic Noise Model Experiments

Experimental setup. We start by validating the isotropic noise model through numerical simulation. Here, we show results for fixed LR, while other training protocols are presented in Appendix E.We consider a scale-invariant function $L(\boldsymbol{w}) = 1 + \frac{\boldsymbol{\mu}^T \boldsymbol{w}}{\|\boldsymbol{w}\|}$ with some fixed $\boldsymbol{\mu} \in \mathbb{R}^d, \|\boldsymbol{\mu}\| = 1$ and train it using noisy gradient descent:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta\Big(\nabla L(\boldsymbol{w}_k) + \frac{1}{\|\boldsymbol{w}_k\|}\boldsymbol{P}_{\overline{\boldsymbol{w}}_k}\sigma\boldsymbol{\varepsilon} + \lambda \boldsymbol{w}_k\Big), \tag{25}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. We select this function because the SDE theory predicts that, at stationarity, the angular distribution follows a Von Mises-Fisher (VMF) form: $\rho_{\overline{\boldsymbol{w}}}(\overline{\boldsymbol{w}}) \propto \exp(-\boldsymbol{\mu}^T\overline{\boldsymbol{w}}/T)$. For this distribution, the analytical values of U and S are known (see Appendix E), allowing direct comparison with simulation results. In this setting, we set $d = 3$.

Entropy estimation. While calculating mean training loss U is straightforward, estimating entropy S is a more complicated task. We rely on the nearest neighbor entropy estimator [Kozachenko and Leonenko, 1987], which requires a sample from the stationary distribution. In Appendix D.6, we present a detailed entropy estimation protocol. In the experiments, we wait until the loss and radius stabilize during the training run. Then, we maintain a queue of 1000 weight vectors sampled every 50 iterations along the training trajectory to reduce correlation between consecutive samples.

Results. In Figure 1, we present the results of numerical simulation. Subfigure 1a shows that both the average loss U and the entropy on the unit sphere $S(\rho_{\overline{\boldsymbol{w}}})$ match the analytical VMF predictions, providing a strong evidence, that this distribution is indeed stationary. Subfigures 1b and 1e validate **V1** and **V2**, respectively. To verify **V3**, we compute the partial derivatives of the total entropy S:

$$\left(\frac{\partial S}{\partial \log \eta}\right)_\lambda = \frac{d-1}{2}, \qquad \left(\frac{\partial S}{\partial \log \lambda}\right)_\eta = 0 \tag{26}$$

Subfigures 1c,d show that empirical measurements closely follow these theoretic predictions. For **V4**, we calculate the isochoric heat capacity

$$C_V = \left(\frac{\delta Q}{\mathrm{d}T}\right)_V = \left(\frac{\partial U}{\partial T}\right)_V = \frac{d-1}{2} \tag{27}$$

Combining it with $C_p - C_V = \frac{d-1}{2}$, we get $\gamma = \frac{C_p}{C_V} = 2$. In the adiabatic process, we maintain $pV^\gamma = \text{const}$, which reduces to

$$pV^\gamma \propto p^{1-\gamma}T^\gamma \propto \eta^{\frac{\gamma}{2}}\lambda^{1-\frac{\gamma}{2}}\Big|_{\gamma=2} \propto \eta \tag{28}$$

Therefore, keeping η fixed and varying λ yields an adiabatic process with $\mathrm{d}S = 0$. This is indeed observed in the simulation (Subfigures 1c,d): the entropy depends only on η and remains constant as λ changes. Note that this property is specific to the VMF distribution. For other distributions $\rho_{\overline{\boldsymbol{w}}}$, the values of C_V, C_p, and γ may differ, requiring coordinated changes of η and λ to achieve an adiabatic process.

7 Neural Network Experiments

7.1 Generalizing Isotropic Noise Model

For neural networks, the isotropic noise assumption breaks down: the covariance matrix $\boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}}$ is generally anisotropic and spatially dependent. [Chaudhari and Soatto, 2018] show that for general networks (i.e., not necessarily scale-invariant) trained with fixed LR η, the stationary distribution takes the form $\rho^*_{\boldsymbol{w}} \propto \exp(-\Phi(\boldsymbol{w})/T_0)$ with $T_0 \propto \eta$, where $\Phi(\boldsymbol{w})$ is an implicit *potential* determined by the loss $L(\boldsymbol{w})$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{w}}$, but independent of η. In this setting, Φ replaces the training loss as the effective energy. An explicit expression for Φ is known only in special cases, such as linear regression [Kunin *et al.*, 2023]. We extend this framework to scale-invariant models by restricting to the unit sphere and defining $\rho^*_{\overline{\boldsymbol{w}}} \propto \exp(-\Phi(\overline{\boldsymbol{w}})/T_0)$ with T_0 determined by training hyperparameters. This distribution arises in all three training protocols introduced in Section 3.

To extend the ideal gas analogy to neural networks, a well-defined stationary radius is required. Following [Li *et al.*, 2020; Wan *et al.*, 2021], we introduce the constant covariance trace assumption, which ensures that Eq. 10 and 16 admit a well-defined limit r^*.

Assumption 5 (Constant covariance trace)**.** *For all* $\overline{\boldsymbol{w}}$, $\operatorname{Tr}\boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}} = \text{const}$ *(while the covariance matrix* $\boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}}$ *itself might be different). We then define* $\sigma^2 = \frac{1}{d-1}\operatorname{Tr}\boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}}$.

Defining $T = \sigma^2 T_0$, minimization of $\mathbb{E}_{\rho_{\overline{\boldsymbol{w}}}}[\Phi(\overline{\boldsymbol{w}})] - T_0 S(\rho_{\overline{\boldsymbol{w}}})$, achieved at $\rho^*_{\overline{\boldsymbol{w}}}$, is equivalent to minimizing

$$U(\rho_{\overline{\boldsymbol{w}}}) - TS(\rho_{\overline{\boldsymbol{w}}}), \quad \text{with } U = \mathbb{E}_{\rho_{\overline{\boldsymbol{w}}}}[\sigma^2\Phi(\overline{\boldsymbol{w}})], \tag{29}$$

We therefore interpret this quantity as the Helmholtz energy F, minimized in the fixed-sphere case. Consequently, we define $G = U(\rho_{\overline{\boldsymbol{w}}}) - TS(\rho_{\boldsymbol{w}}) + pV(r)$, which corresponds to fixed-ELR and fixed-LR cases.

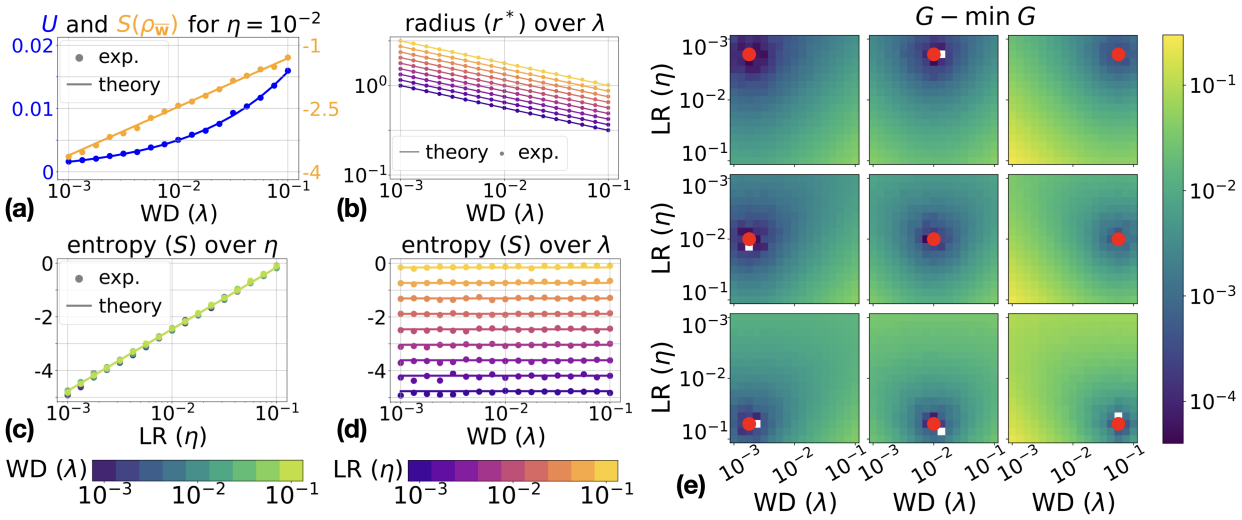

Figure 1: VMF isotropic noise model with fixed LR η and WD λ. Subfigures a–d: points are numerical measurements, solid lines are theoretical predictions: $U = \frac{d-1}{2}T$, $S(\rho_{\overline{w}}) = \frac{d-1}{2}\log(2\pi eT)$, $S = S(\rho_{\overline{w}}) + (d-1)\log r^*$, and $r^* = \sqrt{\frac{T(d-1)}{p}}$, with $T = \sqrt{\frac{\eta\lambda\sigma^2}{2(d-1)}}$ and $p = \lambda$. Subfigure e: Gibbs energy minimization (**V2**). Each subplot corresponds to a fixed pair (η^*, λ^*), denoted with red circle. The colormap shows the difference between G and its minimum across stationary distributions, with the minimizer marked by a white square. Ideally, red circles coincide with white squares; in practice, they either match or lie very close.

To summarize, the thermodynamic analogy for neural networks differs from the isotropic noise model in two points:

1. The covariance matrix $\mathbf{\Sigma}_{\overline{w}}$ is anisotropic, so the energy function is not the training loss $L(\overline{w})$, but a potential $\sigma^2\Phi(\overline{w})$ with no explicit form.
2. The formulae for temperature T and stationary radius r^* remain the same as in Section 3, but $\sigma^2 = \frac{1}{d-1}\operatorname{Tr}\mathbf{\Sigma}_{\overline{w}}$ replaces the isotropic variance.

Thus, under Assumption 5, the ideal gas laws still hold, and we expect the tests **V1**–**V4** to hold. However, since Φ is unknown, we can only directly verify **V1** and **V3**.

7.2 Experimental Setup

We train ResNet-18 [He *et al.*, 2016] on CIFAR-10 [Krizhevsky *et al.*, 2009a] with varying LR η and WD λ, fixed batch size $B = 128$, for $t = 10^6$ iterations. The entropy estimation queue stores 1000 vectors with new weights added every 25 iterations. All models are made fully scale-invariant following [Li and Arora, 2020]. We also use a "thin" model with the width multiplier $k = 4$ (while the default value is $k = 64$), the resulting number of trainable parameters is $d = 43692$. This choice is motivated by (1) memory constraints for storing the entropy estimation queue, and (2) avoiding *overparameterization*, which imposes a specific behavior of stochastic gradients for small LRs later in training (which advocates convergence to loss minimum, as discussed in Appendix I).In the main text, we show the results only for the fixed LR case, two other training protocols are presented in the Appendix F, supported by the results for the ConvNet architecture from [Kodryan *et al.*, 2022] and the CIFAR-100 dataset [Krizhevsky *et al.*, 2009b]. In Appendix G, we also experiment with partially scale-invariant models.

Estimating entropy in high dimensions is challenging: the nearest-neighbor estimator is unbiased only asymptotically, with a bias of order $\mathcal{O}(N^{-2/d})$ [Delattre and Fournier, 2017]. We assume the bias is roughly constant across different stationary distributions, allowing accurate reconstruction of entropy derivatives for Maxwell relations. Empirically, **V3** holds with high precision, supporting this assumption.

7.3 Results

We present the results in Figure 2. Subfigure 2a shows that σ^2 varies across stationary distributions. While theory assumes a fixed σ^2, in practice it depends on the hyperparameters, primarily the product $\eta\lambda$. We therefore define temperature as $T = \sqrt{\frac{\eta\lambda\sigma^2}{2(d-1)}}$ using the measured σ^2, with values shown in Subfigure 2b. T generally increases with η and λ, but for large λ it can be non-monotonic, e.g., at $\lambda = 10^{-1}$, T decreases for $10^{-1} \le \eta \le 10^{-0.5}$ and rises again for $\eta \ge 10^{-0.5}$, reflecting deviations in σ^2 from the trend seen for $\eta\lambda < 10^{-2}$. These results suggest that, for large $\eta\lambda$, stationary distributions explore different regions of the weight space, consistent with prior observations by [Sadrtdinov *et al.*, 2024], who analyze the loss landscape under large LRs.

Subfigure 2c shows that the stationary radius r^* closely follows theoretical predictions, confirming **V1**. Deviations appear only at large η and λ, due to *discretization error* in the SDE: the continuous model neglects the norm of the full-batch gradient, which acts like a centrifugal force and enlarges r^*. In Appendix H, we derive a correction that predicts r^* more accurately for larger values of η and λ.

Finally, to verify **V3**, we represent S as a function of $\log\eta$ and $\log\lambda$, as shown in Subfigures 2e and f. The empirical estimates are noisy, so we smooth them using polynomial regression with quadratic terms (see details in Appendix F).

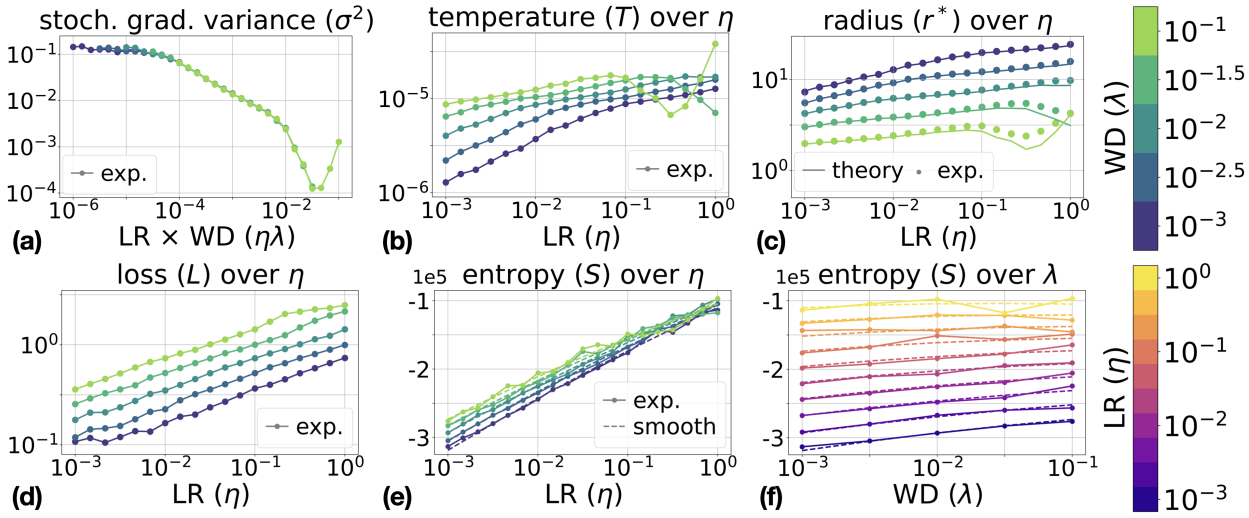

Figure 2: ResNet-18 on CIFAR-10 with fixed LR η and WD λ. Subfigures a, b, d: empirically measured σ^2, mean loss L, and temperature $T = \sqrt{\frac{\eta\lambda\sigma^2}{2(d-1)}}$, respectively. Subfigure c: stationary radius $r^* = \sqrt{\frac{T(d-1)}{p}}$ (solid lines, theory) vs. experimental values (points). Subfigures e and f: entropy S as a function of η and λ; solid lines with markers show experimental estimates, dashed lines their smoothed versions.

This approximation achieves a coefficient of determination of $R^2 \approx 0.99$. We then substitute the partial derivatives of the approximation into Eq. 24 and obtain a maximal absolute error of $\approx 3\%$ between the left- and right-hand sides, indicating that the Maxwell relation holds with high precision.

8 Discussion

Extending Ideal Gas Analogy

Answering the question posed in the title, when the assumption $\operatorname{Tr} \boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}} = \text{const}$ holds, the stationary behavior of SGD is fully consistent with the ideal gas model, as in the isotropic noise setting. In practice, however, we observe that the value of σ^2 varies with training hyperparameters. Nevertheless, our experiments show that **V1** and **V3** remain valid.

Developing a more general thermodynamic description that accounts for variable $\operatorname{Tr} \boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}}$ is an interesting direction for future work. One possibility is to draw an analogy with a *real gas*, described by the state equation $V = Z(p,T)\frac{RT}{p}$, where Z is the compressibility factor capturing deviations from the ideal gas. This mirrors the relation $\frac{(r^*)^2}{2} = \operatorname{Tr} \boldsymbol{\Sigma}_{\overline{\boldsymbol{w}}} \frac{\eta_{\text{eff}}}{4\lambda}$ (from Eq. 10), suggesting that extending the analogy to real gases may be as natural for scale-invariant networks as the ideal gas correspondence established here. Another promising direction is extending the framework beyond fully scale-invariant networks and to more complex optimizers, such as SGD with momentum or Adam [Kingma and Ba, 2015], which could yield insights directly relevant to practical training.

Practical Implications

Our work adopts a thermodynamic perspective to clarify how training hyperparameters, particularly learning rate and weight decay, shape the stationary distribution of SGD. Since neural networks are rarely trained with a fixed learning rate, these insights may inform the design of more effective LR schedulers [Liu *et al.*, 2025]. A key outcome of our framework is the Maxwell relations, which suggest a possible approach to explicitly control the rate of entropy decay. Properly regulating entropy decay may help reduce training loss while avoiding premature convergence to sharp minima often associated with poor generalization [Kodryan *et al.*, 2022].

Moreover, hyperparameters determine the effective temperature, governing the trade-off between internal energy and entropy. This balance is crucial for weight averaging [Izmailov *et al.*, 2018], which improves performance by combining multiple low-loss checkpoints. While low loss is necessary for accuracy, high entropy promotes diversity of solutions and thus effective averaging, making hyperparameter selection non-trivial. Indeed, [Sadrtdinov *et al.*, 2024] show that LRs optimal for weight averaging often exceed the convergence threshold. Hence, understanding how entropy dynamics affect weight averaging is a promising direction for future work, with Maxwell relations providing a quantitative tool for both theoretical and empirical analysis.

9 Conclusion

In this work, we introduced and empirically validated a thermodynamic framework for describing the stationary distributions of SGD in scale-invariant neural networks. We defined macroscopic thermodynamic variables and related them to learning rate, weight decay, and parameter norm. We rigorously showed that, under the simplified isotropic noise model, stationary distributions follow ideal gas laws. Importantly, the key predictions of this framework also hold in experiments with neural networks. Future work may extend this analogy to settings with variable noise, non-scale-invariant architectures, or more complex optimizers, and may further support principled approaches to hyperparameter tuning for learning rate scheduling and weight averaging.

Acknowledgments

We would like to thank Alexander Shabalin for insightful conversations and personal support. Ekaterina Lobacheva was supported by IVADO and the Canada First Research Excellence Fund. The authors gratefully acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Research, Technology and Space of Germany and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner). The empirical results were also enabled by compute resources and technical support provided by Mila - Quebec AI Institute (mila.quebec).

References

[Ali Mehmeti-Göpel and Wand, 2024] Christian Ali Mehmeti-Göpel and Michael Wand. On the weight dynamics of deep normalized networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 992–1007. PMLR, 21–27 Jul 2024.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Chaudhari and Soatto, 2018] 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.

[Chen *et al.*, 2024] Xiaoli Chen, Beatrice W Soh, Zi-En Ooi, Eleonore Vissol-Gaudin, Haijun Yu, Kostya S Novoselov, Kedar Hippalgaonkar, and Qianxiao Li. Constructing custom thermodynamics using deep learning. *Nature Computational Science*, 4(1):66–85, 2024.

[Cho and Lee, 2017] Minhyung Cho and Jaehyung Lee. Riemannian approach to batch normalization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5231–5241. Curran Associates Inc., 2017.

[Delattre and Fournier, 2017] Sylvain Delattre and Nicolas Fournier. On the kozachenko–leonenko entropy estimator. *Journal of Statistical Planning and Inference*, 185:69–93, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[Izmailov *et al.*, 2018] Pavel Izmailov, Andrew Gordon Wilson, Dmitrii Podoprikhin, Dmitry Vetrov, and Timur Garipov. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885, 2018.

[Jastrzębski *et al.*, 2017] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Kodryan *et al.*, 2022] Maxim Kodryan, Ekaterina Lobacheva, Maksim Nakhodnov, and Dmitry Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. In *Advances in Neural Information Processing Systems*, 2022.

[Kozachenko and Leonenko, 1987] Leonid Fedorovich Kozachenko and Nikolai Nikolaevich Leonenko. Sample estimate of the entropy of a random vector. *Problems Inform. Transmission*, 23:95–101, 1987.

[Kozyrev, 2025] Sergei Kozyrev. How to explain grokking, 2025.

[Krizhevsky *et al.*, 2009a] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 (canadian institute for advanced research). 2009.

[Krizhevsky *et al.*, 2009b] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-100 (canadian institute for advanced research). 2009.

[Kunin *et al.*, 2023] Daniel Kunin, Javier Sagastuy-Brena, Lauren Gillespie, Eshed Margalit, Hidenori Tanaka, Surya Ganguli, and Daniel L. K. Yamins. The limiting dynamics of SGD: Modified loss, phase-space oscillations, and anomalous diffusion. *Neural Computation*, 36(1):151–174, 2023.

[Li and Arora, 2020] Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020.

[Li *et al.*, 2020] Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020.

[Liu *et al.*, 2025] Ziming Liu, Yizhou Liu, Jeff Gore, and Max Tegmark. Neural thermodynamic laws for large language model training, 2025.

[Loshchilov *et al.*, 2025] Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Lyle *et al.*, 2024] Clare Lyle, Zeyu Zheng, Khimya Khetarpal, James Martens, Hado van Hasselt, Razvan Pascanu, and Will Dabney. Normalization and effective

learning rates in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 106440–106473. Curran Associates, Inc., 2024.

[Mandt *et al.*, 2017] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *J. Mach. Learn. Res.*, 18(1):4873–4907, January 2017.

[Sadrtdinov *et al.*, 2024] Ildus Sadrtdinov, Maxim Kodryan, Eduard Pokonechny, Ekaterina Lobacheva, and Dmitry Vetrov. Where do large learning rates lead us? In *Advances in Neural Information Processing Systems*, volume 37, pages 58445–58479. Curran Associates, Inc., 2024.

[Wan *et al.*, 2021] Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6380–6391. Curran Associates, Inc., 2021.

[Wang and Wang, 2022] Yi Wang and Zhiren Wang. Three-stage evolution and fast equilibrium for SGD with non-degerate critical points. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23092–23113. PMLR, 17–23 Jul 2022.

[Zhang, 2024] Dong Zhang. On the temperature of machine learning systems, 2024.