Noise Balance and Stationary Distribution of Stochastic Gradient Descent

Anonymous Author(s)
Affiliation
Address
email

Abstract

How the stochastic gradient descent (SGD) navigates the loss landscape of a neural network remains poorly understood. This work shows that the minibatch noise of SGD regularizes the solution towards a noise-balanced solution whenever the loss function contains a rescaling symmetry. We prove that when the rescaling symmetry exists, the SGD dynamics is limited to only a low-dimensional subspace and prefers a special set of solutions in an infinitely large degenerate manifold, which offers a partial explanation of the effectiveness of SGD in training neural networks. We then apply this result to derive the stationary distribution of stochastic gradient flow for a diagonal linear network with arbitrary depth and width, which is the first analytical expression of the stationary distribution of SGD in a high-dimensional non-quadratic potential. The stationary distribution exhibits complicated nonlinear phenomena such as phase transitions, loss of ergodicity, memory effects, and fluctuation inversion. These phenomena are shown to exist uniquely in deep networks, highlighting a fundamental difference between deep and shallow models. Lastly, we discuss the implication of the proposed theory for the practical problem of variational Bayesian inference.

1 Introduction

In natural and social sciences, one of the most important objects of study of a stochastic system is its stationary distribution, which is often found to offer fundamental insights into understanding a given stochastic process [36, 29]. Arguably, a great deal of insights into SGD can be obtained if we have an analytical understanding of its stationary distribution, which remains unknown until today. The stochastic gradient descent (SGD) algorithm is defined as $\Delta\theta_t = -\frac{\eta}{S}\sum_{x\in B}\nabla_\theta \ell(\theta, x)$, where θ is the model parameter and $\ell(\theta, x)$ is a per-sample loss whose expectation over x gives the training loss: $L(\theta) = \mathbb{E}_x[\ell(\theta, x)]$. B is a randomly sampled minibatch of data points, each independently sampled from the training set, and S is the minibatch size. Two aspects of the algorithm make it difficult to understand this algorithm: (1) its dynamics is discrete in time, and (2) the randomness is highly nonlinear and parameter-dependent. This work relies on the continuous-time approximation and deals with the second aspect.

The main contributions are

1. the derivation of the "law of balance," which shows that SGD converges to a special subset of noised-balanced solutions when the rescaling symmetry is present;
2. the first-of-its-kind solution of the stationary distribution of an analytical model trained by SGD;
3. discovery of novel phenomena such as phase transitions, loss of ergodicity, memory effects, and fluctuation inversion, all implied by our theory.

Organization. The next section discusses the closely related works. In Section 3, we prove the law of balance, the first main result of this work, and discuss its implications for common neural

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

networks. In Section 4, we apply the law of balance to derive the stationary distribution of SGD for a highly nontrivial loss landscape. The last section concludes this work. All proofs and derivations are given in Appendix A.

2 Related Works

Solution of the Fokker Planck (FP) Equation. The FP equation is a high-dimensional partial differential equation whose solution (and its existence) is an open problem in mathematics and many fields of sciences and only known for a few celebrated special cases [28]. Our solution is the first of its kind in a deep-learning setting. **Stationary distribution of SGD**. One of the earliest works that computes the stationary distribution of SGD is the Lemma 20 of Ref. [3], which assumes that the noise has a constant covariance and shows that if the loss function is quadratic, then the stationary distribution is Gaussian. Similarly, using a saddle point expansion and assuming that the noise is parameter-independent, a series of recent works showed that the stationary distribution of SGD is exponential in the model parameters close to a local minimum: $p(\theta) \propto \exp[-a\theta^T H\theta]$, for some constant a and matrix H [21, 41, 19]. Assuming that the noise covariance only depends on the loss function value $L(\theta)$, Refs. [24] and [39] showed that the stationary distribution is power-law-like and proportional to $L(\theta)^{-c_0}$ for some constant c_0. A primary feature of these previous results is that stationary distribution does not exhibit any memory effect and also preserves ergodicity. Until now, no analytical solution to the stationary distribution of SGD is known, making it impossible to judge how good the previous approximate results are. Our result is the first to derive an exact solution to the stationary distribution of SGD without any approximation. We will see that in contrast to the approximate solutions in the previous results, the actual distribution of SGD has both a memory effect and features the loss of ergodicity.

Symmetry and SGD dynamics. Also related to our work is the study of how symmetry affects the learning dynamics of SGD. A major prior work is [17], which studies the dynamics of SGD when there is scale invariance, conjecturing that SGD reaches a fast equilibrium state at the early stage of training. Our result is different as we study a different type of symmetry, the rescaling symmetry.

3 Noise Balance

We consider the continuous-time limit of SGD [15, 16, 18, 32, 8, 11]:

$$d\theta = -\nabla_\theta L dt + \sqrt{TC(\theta)}dW_t, \tag{1}$$

where $C(\theta) = \mathbb{E}[\nabla\ell(\theta)\nabla^T\ell(\theta)]$ is the gradient covariance, dW_t is a stochastic process satisfying $dW_t \sim N(0, Idt)$ and $\mathbb{E}[dW_t dW_{t'}^T] = \delta(t-t')I$, and $T = \eta/S$. Apparently, T gives the average noise level in the dynamics. Previous works have suggested that the ratio T is a main factor determining the behavior of SGD, and using different T often leads to different generalization performance [31, 19, 44].

3.1 Rescaling Symmetry and Law of Balance

Due to standard architecture designs, a type of invariance – the rescaling symmetry – often appears in the loss function and it is preserved for all sampling of minibatches. The per-sample loss ℓ is said to have the rescaling symmetry for all x if $\ell(u, w, x) = \ell(\lambda u, w/\lambda, x)$ for a scalar $\lambda \in \mathbb{R}_+$. This type of symmetry appears in many scenarios in deep learning. For example, it appears in any neural network with the ReLU activation. It also appears in the self-attention of transformers, often in the form of key and query matrices [37]. When this symmetry exists between u and w, one can prove the following result, which we refer to as the law of balance.

Theorem 3.1. *Let u, w, and v be parameters of arbitrary dimensions. Let $\ell(u, w, v, x)$ satisfy $\ell(u, w, v, x) = \ell(\lambda u, w/\lambda, v, x)$ for arbitrary x and any $\lambda \in \mathbb{R}_+$. Then,*

$$\frac{d}{dt}(||u||^2 - ||w||^2) = -T(u^T C_1 u - w^T C_2 w), \tag{2}$$

where $C_1 = \mathbb{E}[A^T A] - \mathbb{E}[A^T]\mathbb{E}[A]$, $C_2 = \mathbb{E}[AA^T] - \mathbb{E}[A]\mathbb{E}[A^T]$ and $A_{ki} = \partial\tilde{\ell}/\partial(u_i w_k)$ with $\tilde{\ell}(u_i w_k, v, x) \equiv \ell(u_i, w_k, v, x)$.[1]

[1]This result also holds using the modified loss (See Appendix A.3).

Here, v stands for the parameters that are irrelevant to the symmetry, and C_1 and C_2 are positive semi-definite by definition. The theorem still applies if the model has parameters other than u and w. The theorem can be applied recursively when multiple rescaling symmetries exist. See Figure 1 for an illustration the the dynamics and how it differs from other types of GD.

While the matrices C_1 and C_2 may not always be full-rank, we emphasize that in common deep-learning settings with rescaling symmetry, the law of balance is almost always well-defined and applicable. In Appendix A.4, we prove that under very general settings, for all *active* hidden neurons of a two-layer ReLU net, C_1 and C_2 are always full-rank. Equation (2) is the law of balance, and it implies two different types of balance. The first type of balance is the balance of gradient noise. The proof of the theorem shows that the stationary point of the law in (2) is equivalent to

$$\mathrm{Tr}_w[C(w)] = \mathrm{Tr}_u[C(u)], \tag{3}$$

where $C(w)$ and $C(u)$ are the gradient covariance of w and u, respectively. Therefore, SGD prefers a solution where the gradient noise between the two layers is balanced. Also, this implies that the balance conditions of the law is only dependent on the diagonal terms of the Fisher information (if we regard the loss as a log probability), which is often well-behaved. As a last caveat, we emphasize that the fact that the noise will balance does not imply that either trace will converge or stay close to a fixed value – it is also possible for both terms to oscillate while their difference is close to zero.

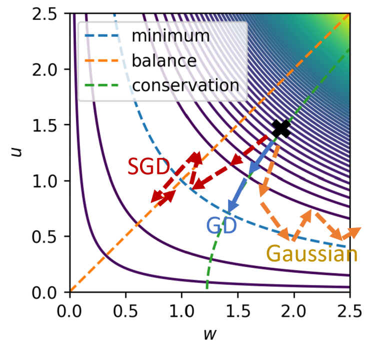

Figure 1: Dynamics of GD and SGD and GD with injected Gaussian noise for the simple problem $\ell(u, w) = (uwx - y)^2$. Due to the rescaling symmetry between u and w, GD follows a conservation law: $u^2(t) - w^2(t) = u^2(0) - w^2(0)$, SGD converges to the balanced solution $u^2 = w^2$, while GD with injected noise diverges due to simple diffusion in the degenerate directions.

The second type is the norm ratio balance between layers, though the norm ratio may not necessarily be finite. Equation (2) implies that in the degenerate direction of the rescaling symmetry, a single and unique point is favored by SGD. Let $u = \lambda u^*$ and $w = \lambda^{-1} w^*$ for arbitrary u^* and w^*, then, the stationary point of the law is reached at $\lambda^4 = \frac{(w^*)^T C_2 w^*}{(u^*)^T C_1 u^*}$. The quantity λ can be called the "balancedness" of the norm, and the law states that when a rescaling symmetry exists, a special balancedness is preferred by the SGD algorithm. When C_1 or C_2 vanishes, λ or λ^{-1} diverges, and so does SGD. Therefore, having a nonvanishing noise actually implies that SGD training will be more stable. For common problems, C_1 and C_2 are positive definite and, thus, if we know the spectrum of C_1 and C_2 at the end of training, we can estimate a rough norm ratio at convergence:

$$-T(\lambda_{1M}||u||^2 - \lambda_{2m}||w||^2) \le \frac{d}{dt}(||u||^2 - ||w||^2) \le -T(\lambda_{1m}||u||^2 - \lambda_{2M}||w||^2),$$

where $\lambda_{1m(2m)}$ and $\lambda_{1M(2M)}$ represent the minimal and maximal eigenvalue of the matrix $C_{1(2)}$, respectively. Thefore, the value of $||u||^2/||w||^2$ is restricted by (See Section A.5)

$$\frac{\lambda_{2m}}{\lambda_{1M}} \le \frac{||u||^2}{||w||^2} \le \frac{\lambda_{2M}}{\lambda_{1m}}. \tag{4}$$

Thus, a remaining question is whether the quantities $u^T C_1 u$ and $w^T C_2 w$ are generally well-defined and nonvanishing or not. The following proposition shows that for a generic two-layer ReLU net, $u^T C_1 u$ and $w^T C_2 w$ are almost everywhere strictly positive. We define a two-layer ReLU net as

$$f(x) = \sum_i^d u_i \mathrm{ReLU}(w_i^T x + b_i), \tag{5}$$

where $u_i \in \mathbb{R}^{d_u}$, $w_i \in \mathbb{R}^{d_w}$ and b_i is a scalar with i being the index of the hidden neuron. For each i, the model has the rescaling symmetry: $u_i \to \lambda u_i$, $(w_i, b_i) \to (\lambda^{-1} w_i, \lambda^{-1} b_i)$. We thus apply the law of balance to each neuron separately. The per-sample loss function is

$$\ell(\theta, x) = \|f(x) - y(x, \epsilon)\|^2. \tag{6}$$

Here, x has a full-rank covariance Σ_x, and $y = g(x) + \epsilon$ for some function g and ϵ is a zero-mean random vector independent of x and have the full-rank covariance Σ_ϵ. The following theorem shows that for this network, C_1 and C_2 are full rank unless the neuron is "dead".

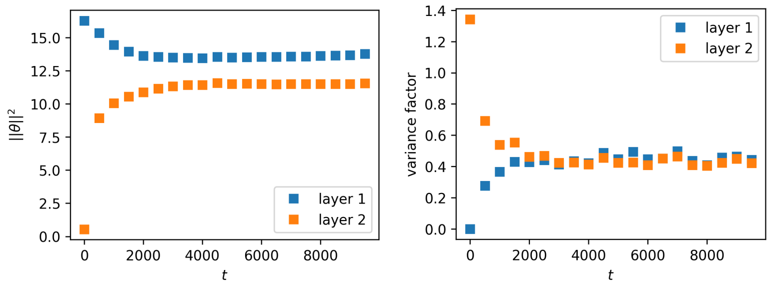

Figure 2: A two-layer ReLU network trained on a full-rank dataset. **Left**: because of the rescaling symmetry, the norms of the two layers are balanced approximately (but not exactly). **Right**: the first and second terms in Eq. (2). We see that both terms evolve towards a point where they exactly balance. In agreement with our theory, SGD training leads to an approximate norm balance and exact gradient noise balance.

Theorem 3.2. *Let the loss function be given in Eq. (6). Let $C_1^{(i)}$ and $C_2^{(i)}$ denote the corresponding noise matrices of the i-th neuron, and $p_i := \mathbb{P}(w_i^T x + b_i > 0)$. Then, $C_1^{(i)}$ and $C_2^{(i)}$ are full-rank for all i such that $p_i > 0$.*

See Figure 2. We train a two-layer ReLU network with the number of neurons: $20 \to 200 \to 20$. The dataset is a synthetic data set, where x is drawn from a normal distribution, and the labels: $y = x + \epsilon$, for an independent Gaussian noise ϵ with unit variance. While every neuron has a rescaling symmetry, we focus on the overall rescaling symmetry between the two weight matrices. The norm between the two layers reach a state of approximate balance – but not a precise balance. At the same time, the model evolves during training towards a state where $u^T C_1 u$ and $w^T C_2 w$ are balanced.

Standard analysis shows that the difference between SGD and GD is of order T^2 per unit time step, and it is thus often believed that SGD can be understood perturbatively through GD [11]. However, the law of balance implies that the difference between GD and SGD is not perturbative. As long as there is any level of noise, the difference between GD and SGD at stationarity is $O(1)$. This theorem also implies the loss of ergodicity, an important phenomenon in nonequilibrium physics [26, 34, 22, 35], because not all solutions with the same training loss will be accessed by SGD with equal probability.

3.2 1d Rescaling Symmetry

The theorem greatly simplifies when both u and w are one-dimensional.

Corollary 3.3. *If $u, w \in \mathbb{R}$, then, $\frac{d}{dt}|u^2 - w^2| = -TC_0|u^2 - w^2|$, where $C_0 = \mathrm{Var}[\frac{\partial \ell}{\partial (uw)}]$.*

Before we apply the theorem to study the stationary distributions, we stress the importance of this balance condition. This relation is closely related to Noether's theorem [23, 1, 20]. If there is no weight decay or stochasticity in training, the quantity $||u||^2 - ||w||^2$ will be a conserved quantity under gradient flow [6, 14, 33], as is evident by taking the infinite S limit. The fact that it monotonically decays to zero at a finite T may be a manifestation of some underlying fundamental mechanism. A more recent result in Ref. [38] showed that for a two-layer linear network, the norms of two layers are within a distance of order $O(\eta^{-1})$, suggesting that the norm of the two layers are balanced. Our result agrees with Ref. [38] in this case, but our result is stronger because our result is nonperturbative, only relies on the rescaling symmetry, and is independent of the loss function or architecture of the model. It is useful to note that when L_2 regularization with strength γ is present, the rate of decay changes from TC_0 to $TC_0 + \gamma$. This points to a nice interpretation that when rescaling symmetry is present, the implicit bias of SGD is equivalent to weight decay. See Figure 1 for an illustration of this point.

Example: two-layer linear network. It is instructive to illustrate the application of the law to a two-layer linear network, the simplest model that obeys the law. Let $\theta = (w, u)$ denote the set of trainable parameters; the per-sample loss is $\ell(\theta, x) = (\sum_i^d u_i w_i x - y)^2 + \gamma||\theta||^2$. Here, d is the width of the model, $\gamma||\theta||^2$ is the L_2 regularization term with strength $\gamma \geq 0$, and $\mathbb{E}_x$ denotes the averaging over the training set, which could be a continuous distribution or a discrete sum of delta distributions. It will be convenient for us also to define the shorthand: $v := \sum_i^d u_i w_i$. The distribution of v is said to be the distribution of the "model." Applying the law of balance, we obtain that

$$\frac{d}{dt}(u_i^2 - w_i^2) = -4[T(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3) + \gamma](u_i^2 - w_i^2), \tag{7}$$

where we have introduced the parameters

$$\alpha_1 := \mathrm{Var}[x^2], \quad \alpha_2 := \mathbb{E}[x^3 y] - \mathbb{E}[x^2]\mathbb{E}[xy], \quad \alpha_3 := \mathrm{Var}[xy]. \tag{8}$$

When $\alpha_1\alpha_3 - \alpha_2^2$ or $\gamma > 0$, the time evolution of $|u^2 - w^2|$ can be upper-bounded by an exponentially decreasing function in time: $|u_i^2 - w_i^2|(t) < |u_i^2 - w_i^2|(0)\exp\left(-4T(\alpha_1\alpha_3 - \alpha_2^2)t/\alpha_1 - 4\gamma t\right) \to 0$. Namely, the quantity $(u_i^2 - w_i^2)$ decays to 0 with probability 1. We thus have $u_i^2 = w_i^2$ for all $i \in \{1, \cdots, d\}$ at stationarity, in agreement with the Corollary.

4 Stationary Distribution of SGD

As an important application of the law of balance, we solve the stationary distribution of SGD for a deep diagonal linear network. While linear networks are limited in expressivity, their loss landscape and dynamics are highly nonlinear and exhibits many shared phenomenon with nonlinear neural networks [13, 30]. Let θ follow the high-dimensional Wiener process given by Eq.(1). The probability density evolves according to its Kolmogorov forward (Fokker-Planck) equation:

$$\frac{\partial}{\partial t}p(\theta,t) = -\sum_i \frac{\partial}{\partial_{\theta_i}}\left(p(\theta,t)\frac{\partial}{\partial_{\theta_i}}L(\theta)\right) + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial_{\theta_i}\partial_{\theta_j}}C_{ij}(\theta)p(\theta,t). \tag{9}$$

The solution of this partial differential equation is an open problem for almost all high-dimensional problems. This section solves it for a high-dimensional non-quadratic potential of a machine learning relevance.

4.1 Depth-0 Case

Let us first derive the stationary distribution of a one-dimensional linear regressor, which will be a basis for comparison to help us understand what is unique about having a "depth" in deep learning. The per-sample loss is $\ell(x, v) = (vx - y)^2 + \gamma v^2$. Defining

$$\beta_1 := \mathbb{E}[x^2], \quad \beta_2 := \mathbb{E}[xy], \tag{10}$$

the global minimizer of the loss can be written as: $v^* = \beta_2/\beta_1$. The gradient variance is also not trivial: $C(v) := \mathrm{Var}[\nabla_v \ell(v, x)] = 4(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)$. Note that the loss landscape L only depends on β_1 and β_2, and the gradient noise only depends on α_1, α_2 and, α_3. It is thus reasonable to call β the landscape parameters and α the noise parameters. Both β and α appear in all stationary distributions, implying that the stationary distributions of SGD are strongly data-dependent. Another relevant quantity is $\Delta := \min_v C(v) \geq 0$, which is the minimal level of noise on the landscape. It turns out that the stationary distribution is qualitatively different for $\Delta = 0$ and for $\Delta > 0$. For all the examples in this work,

$$\Delta = \mathrm{Var}[x^2]\mathrm{Var}[xy] - \mathrm{cov}(x^2, xy) = \alpha_1\alpha_3 - \alpha_2^2. \tag{11}$$

When is Δ zero? It happens when, for all samples of (x, y), $xy + c = kx^2$ for some constant k and c. We focus on the case $\Delta > 0$ in the main text, which is most likely the case for practical situations. The other cases are dealt with in Section A.

For $\Delta > 0$, the stationary distribution for linear regression is (Section A)

$$p(v) \propto (\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)^{-1-\frac{\beta_1'}{2T\alpha_1}} \exp\left[-\frac{1}{T}\frac{\alpha_2\beta_1' - \alpha_1\beta_2}{\alpha_1\sqrt{\Delta}}\arctan\left(\frac{\alpha_1 v - \alpha_2}{\sqrt{\Delta}}\right)\right], \tag{12}$$

in agreement with the previous result [24]. Two notable features exist for this distribution: (1) the power exponent for the tail of the distribution depends on the learning rate and batch size, and (2) the integral of $p(v)$ converges for an arbitrary learning rate. On the one hand, this implies that increasing the learning rate alone cannot introduce new phases of learning to a linear regression; on the other hand, it implies that the expected error is divergent as one increases the learning rate (or the feature variation), which happens at $T = \beta_1'/\alpha_1$. We will see that deeper models differ from the single-layer model in these two crucial aspects.

4.2 An Analytical Model

Now, we consider the following model with a notion of depth and width; its loss function is

$$\ell = \left[\sum_i^{d_0}\left(\prod_{k=0}^{D} u_i^{(k)}\right)x - y\right]^2, \tag{13}$$

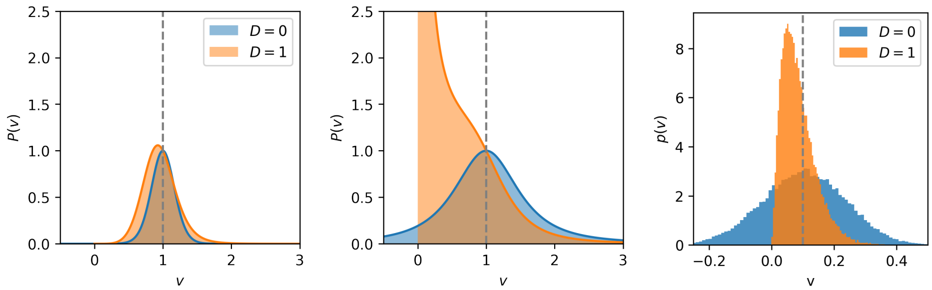

Figure 3: Stationary distributions of SGD for simple linear regression ($D = 0$), and a two-layer network ($D = 1$) across different $T = \eta/S$: $T = 0.05$ (**left**) and $T = 0.5$ (**Mid**). We see that for $D = 1$, the stationary distribution is strongly affected by the choice of the learning rate. In contrast, for $D = 0$, the stationary distribution is also centered at the global minimizer of the loss function, and the choice of the learning rate only affects the thickness of the tail. **Right**: the stationary distribution of a one-layer $\tanh$-model, $f(x) = \tanh(vx)$ ($D = 0$) and a two-layer tanh-model $f(x) = w\tanh(ux)$ ($D = 1$). For $D = 1$, we define $v := wu$. The vertical line shows the ground truth. The deeper model never learns the wrong sign of wu, whereas the shallow model can learn the wrong one.

where D can be regarded as the depth and d_0 the width. When the width $d_0 = 1$, the law of balance is sufficient to solve the model. When $d_0 > 1$, we need to eliminate additional degrees of freedom. We note that this model conceptually resembles (but not identical to) a diagonal linear network, which has been found to well approximate the dynamics of real networks [27, 25, 2, 7].

We introduce $v_i := \prod_{k=0}^{D} u_i^{(k)}$, and so $v = \sum_i v_i$, where we call v_i a "subnetwork" and v the "model." The following proposition shows that independent of d_0 and D, the dynamics of this model can be reduced to a one-dimensional form by invoking the law of balance.

Theorem 4.1. *For all $i \neq j$, one (or more) of the following conditions holds for all trajectories at stationarity: (1) $v_i = 0$, or $v_j = 0$, or $L(\theta) = 0$; (2) $\operatorname{sgn}(v_i) = \operatorname{sgn}(v_j)$. In addition, (2a) if $D = 1$, for a constant c_0, $\log|v_i| - \log|v_j| = c_0$; (2b) if $D > 1$, $|v_i|^2 - |v_j|^2 = 0$.*

This theorem contains many interesting aspects. First of all, the three situations in item 1 directly tell us the distribution of v if the initial state of of v is given by these conditions.[2] This implies a memory effect, namely, that the stationary distribution of SGD can depend on its initial state. The second aspect is the case of item 2, which we will solve below. Item 2 of the theorem implies that all the v_i of the model must be of the same sign for any network with $D \geq 1$. Namely, no subnetwork of the original network can learn an incorrect sign. This is dramatically different from the case of $D = 0$. We will discuss this point in more detail below. The third interesting aspect of the theorem is that it implies that the dynamics of SGD is qualitatively different for different depths of the model. In particular, $D = 1$ and $D > 1$ have entirely different dynamics. For $D = 1$, the ratio between every pair of v_i and v_j is a conserved quantity. In sharp contrast, for $D > 1$, the distance between different v_i is no longer conserved but decays to zero. Therefore, a new balancing condition emerges as we increase the depth. Conceptually, this qualitative distinction also corroborates the discovery in Ref. [43], where $D = 1$ models are found to be qualitatively different from models with $D > 1$.

With this theorem, we are ready to solve the stationary distribution. It suffices to condition on the event that v_i does not converge to zero. Let us suppose that there are d nonzero v_i that obey item 2 of Theorem 4.1 and d can be seen as an effective width of the model. We stress that the effective width $d \leq d_0$ depends on the initialization and can be arbitrary.[3] Therefore, we condition on a fixed value of d to solve for the stationary distribution of v (Appendix A):

Theorem 4.2. *Let $\delta(x)$ denote the Dirac delta function. For an arbitrary factor $z\ in[0,1]$, an invariant solution of the Fokker-Planck Equation is $p^*(v) = (1-z)\delta(v) + zp_\pm(v)$, where*

$$p_\pm(|v|) \propto \frac{1}{|v|^{3(1-1/(D+1))} g_\mp(v)} \exp\left(-\frac{1}{T}\int_0^{|v|} d|v| \frac{d^{1-2/(D+1)}(\beta_1|v| \mp \beta_2)}{(D+1)|v|^{2D/(D+1)} g_\mp(v)}\right), \tag{14}$$

where p_- is the distribution on $(-\infty, 0)$ and p_+ is that on $(0, \infty)$, and $g_\mp(v) = \alpha_1|v|^2 \mp 2\alpha_2|v| + \alpha_3$.

[2] $L \to 0$ is only possible when $\Delta = 0$ *and* $v = \beta_2/\beta_1$.

[3] One can initialize the parameters such that d takes any value between 1 and d_0. One way to achieve this is to initialize on the stationary points specified by Theorem 4.1 at the desired d.

The arbitrariness of the scalar z is due to the memory effect of SGD – if all parameters are initialized at zero, they will remain there with probability 1. This means that the stationary distribution is not unique. Since the result is symmetric in the sign of $\beta_2 = \mathbb{E}[xy]$, we assume that $\mathbb{E}[xy] > 0$ from now on.

Also, we focus on the case $\gamma = 0$ in the main text.[4] The distribution of v is

$$p_{\pm}(|v|) \propto \frac{|v|^{\pm\beta_2/2\alpha_3 T - 3/2}}{(\alpha_1|v|^2 \mp 2\alpha_2|v| + \alpha_3)^{1\pm\beta_2/4T\alpha_3}} \exp\left(-\frac{1}{2T}\frac{\alpha_3\beta_1 - \alpha_2\beta_2}{\alpha_3\sqrt{\Delta}} \arctan\frac{\alpha_1|v| \mp \alpha_2}{\sqrt{\Delta}}\right). \quad (15)$$

This measure is worth a close examination. First, the exponential term is upper and lower bounded and well-behaved in all situations. In contrast, the polynomial term becomes dominant both at infinity and close to zero. When $v < 0$, the distribution is a delta function at zero: $p(v) = \delta(v)$. To see this, note that the term $v^{-\beta_2/2\alpha_3 T - 3/2}$ integrates to give $v^{-\beta_2/2\alpha_3 T - 1/2}$ close to the origin, which is infinite. Away from the origin, the integral is finite. This signals that the only possible stationary distribution has a zero measure for $v \neq 0$. The stationary distribution is thus a delta distribution, meaning that if x and y are positively correlated, the learned subnets v_i can never be negative, independent of the initial configuration.

For $v > 0$, the distribution is nontrivial. Close to $v = 0$, the distribution is dominated by $v^{\beta_2/2\alpha_3 T - 3/2}$, which integrates to $v^{\beta_2/2\alpha_3 T - 1/2}$. It is only finite below a critical $T_c = \beta_2/\alpha_3$. This is a phase-transition-like behavior. As $T \to (\beta_2/\alpha_3)_-$, the integral diverges and tends to a delta distribution. Namely, if $T > T_c$, we have $u_i = w_i = 0$ for all i with probability 1, and no learning can happen. If $T < T_c$, the stationary distribution has a finite variance, and learning may happen. In the more general setting, where weight decay is present, this critical T shifts to $T_c = \frac{\beta_2 - \gamma}{\alpha_3}$. When $T = 0$, the phase transition occurs at $\beta_2 = \gamma$, in agreement with the threshold weight decay identified in Ref. [45]. See Figure 3 for illustrations of the distribution across different values of T. We also compare with the stationary distribution of a depth-0 model. Two characteristics of the two-layer model appear rather striking: (1) the solution becomes a delta distribution at the sparse solution $u = w = 0$ at a large learning rate; (2) the two-layer model never learns the incorrect sign (v is always non-negative). Another exotic phenomenon implied by the result is what we call the "fluctuation inversion." Naively, the variance of model parameters should increase as we increase T, which is the noise level in SGD. However, for the distribution we derived, the variance of v and u both decrease to zero as we increase T: injecting noise makes the model fluctuation vanish. We discuss more about this "fluctuation inversion" in the next section.

Also, while there is no other phase-transition behavior below T_c, there is still an interesting and practically relevant crossover behavior in the distribution of the parameters as we change the learning rate. When training a model, The most likely parameter we obtain is given by the maximum likelihood estimator of the distribution, $\hat{v} := \arg\max p(v)$. Understanding how $\hat{v}(T)$ changes as a function of T is crucial. This quantity also exhibits nontrivial crossover behaviors at critical values of T.

When $T < T_c$, a nonzero maximizer for $p(v)$ must satisfy

$$v^* = -\frac{\beta_1 - 10\alpha_2 T - \sqrt{(\beta_1 - 10\alpha_2 T)^2 + 28\alpha_1 T(\beta_2 - 3\alpha_3 T)}}{14\alpha_1 T}. \quad (16)$$

The existence of this solution is nontrivial, which we analyze in Appendix A.8. When $T \to 0$, a solution always exists and is given by $v = \beta_2/\beta_1$, which does not depend on the learning rate or noise C. Note that β_2/β_1 is also the minimum point of $L(u_i, w_i)$. This means that SGD is only a consistent estimator of the local minima in deep learning in the vanishing learning rate limit. How biased is SGD at a finite learning rate? Two limits can be computed. For a small learning rate, the leading order correction to the solution is $v = \frac{\beta_2}{\beta_1} + \left(\frac{10\alpha_2\beta_2}{\beta_1^2} - \frac{7\alpha_1\beta_2^2}{\beta_1^3} - \frac{3\alpha_3}{\beta_1}\right)T$. This implies that the common Bayesian analysis that relies on a Laplace expansion of the loss fluctuation around a local minimum is improper. The fact that the stationary distribution of SGD is very far away from the Bayesian posterior also implies that SGD is only a good Bayesian sampler at a small learning rate.

Example. It is instructive to consider an example of a structured dataset: $y = kx + \epsilon$, where $x \sim \mathcal{N}(0,1)$ and the noise ϵ obeys $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We let $\gamma = 0$ for simplicity. If $\sigma^2 > \frac{8}{21}k^2$, there always

[4] When weight decay is present, the stationary distribution is the same, except that one needs to replace β_2 with $\beta_2 - \gamma$. Other cases are also studied in detail in Appendix A and listed in Table. 1.

exists a transitional learning rate: $T^* = \frac{4k+\sqrt{42}\sigma}{4(21\sigma^2-8k^2)}$. Obviously, $T_c/3 < T^*$. One can characterize the learning of SGD by comparing T with T_c and T^*. For this simple example, SGD can be classified into roughly 5 different regimes. See Figure 4.

4.3 Power-Law Tail of Deeper Models

An interesting aspect of the depth-1 model is that its distribution is independent of the width d of the model. This is not true for a deep model, as seen from Eq. (14). The d-dependent term vanishes only if $D = 1$. Another intriguing aspect of the depth-1 distribution is that its tail is independent of any hyperparameter of the problem, dramatically different from the linear regression case. This is true for deeper models as well.

Since d only affects the non-polynomial part of the distribution, the stationary distribution scales as $p(v) \propto \frac{1}{v^{3(1-1/(D+1))}(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}$. Hence, when $v \to \infty$, the scaling behaviour is $v^{-5+3/(D+1)}$. The tail gets monotonically thinner as one increases the depth. For $D = 1$, the exponent is $7/2$; an infinite-depth network has an exponent of 5. Therefore, the tail of the model distribution only depends on the depth and is independent of the data or details of training, unlike the depth-0 model. In addition, due to the scaling $v^{5-3/(D+1)}$ for $v \to \infty$, we can see that $\mathbb{E}[v^2]$ will never diverge no matter how large the T is.

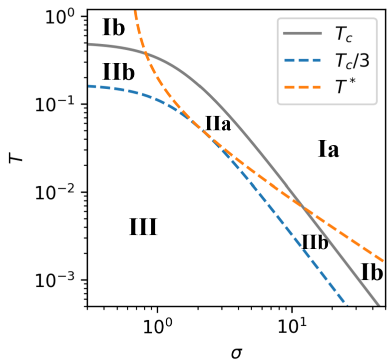

Figure 4: Regimes of learning for SGD as a function of T and the noise in the dataset σ. According to (1) whether the sparse transition has happened, (2) whether a nontrivial maximum probability estimator exists, and (3) whether the sparse solution is a maximum probability estimator, the learning of SGD can be characterized into 5 regimes. Regime **I** is where SGD converges to a sparse solution with zero variance. In regime **II**, the stationary distribution has a finite spread, but the probability of being close to the sparse solution is very high. In regime **III**, the probability density of the sparse solution is zero, and therefore the model will learn without much problem. In regime **b**, a local nontrivial probability maximum exists. The only maximum probability estimator in regime **a** is the sparse solution.

An intriguing feature of this model is that the model with at least one hidden layer will never have a divergent training loss. This directly explains the puzzling observation of the edge-of-stability phenomenon in deep learning: SGD training often gives a neural network a solution where a slight increment of the learning rate will cause discrete-time instability and divergence [40, 4]. These solutions, quite surprisingly, exhibit low training and testing loss values even when the learning rate is right at the critical learning rate of instability. This observation contradicts naive theoretical expectations. Let $\eta_{\rm sta}$ denote the largest stable learning rate. Close to a local minimum, one can expand the loss function up to the second order to show that the value of the loss function L is proportional to $\mathrm{Tr}[\Sigma]$. However, $\Sigma \propto 1/(\eta_{\rm sta} - \eta)$ should be a very large value [42, 19], and therefore L should diverge. Thus, the edge of stability phenomenon is incompatible with the naive expectation up to the second order, as pointed out by Ref. [5]. Our theory offers a direct explanation of why the divergence of loss does not happen: for deeper models, the fluctuation of model parameters decreases as the gradient noise level increases, reaching a minimal value before losing stability. Thus, SGD always has a finite loss because of the power-law tail and fluctuation inversion. See Figure 5–mid.

Infinite-D limit. As D tends to infinity, the distribution becomes

$$p(v) \propto \frac{1}{v^{3+k_1}(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)^{1-k_1/2}} \exp\left(-\frac{d}{DT}\left(\frac{\beta_2}{\alpha_3 v} + \frac{\alpha_2\alpha_3\beta_1 - 2\alpha_2^2\beta_2 + \alpha_1\alpha_3\beta_2}{\alpha_3^2\sqrt{\Delta}} \arctan(\frac{\alpha_1 v - \alpha_2}{\sqrt{\Delta}}) \right)\right),$$

where $k_1 = d(\alpha_3\beta_1 - 2\alpha_2\beta_2)/(TD\alpha_3^2)$. An interesting feature is that the architecture ratio d/D always appears simultaneously with $1/T$. This implies that for a sufficiently deep neural network, the ratio D/d also becomes proportional to the strength of the noise. Since we know that $T = \eta/S$ determines the performance of SGD, our result thus shows an extended scaling law of training: $\frac{d}{D}\frac{S}{\eta} = const$. The architecture aspect of the scaling law also agrees with an alternative analysis [9, 10], where the optimal architecture is found to have a constant ratio of d/D. See Figure 5.

Now, if we T, there are three situations: (1) $d = o(D)$, (2) $d = c_0 D$ for a constant c_0, (3) $d = \Omega(D)$. If $d = o(D)$, $k_1 \to 0$ and the distribution converges to $p(v) \propto v^{-3}(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)^{-1}$, which is a delta distribution at 0. Namely, if the width is far smaller than the depth, the model will collapse to

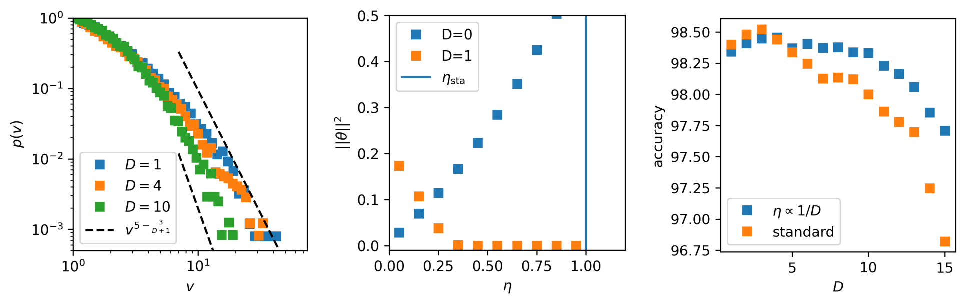

Figure 5: SGD on deep networks leads to a well-controlled distribution and training loss. **Left**: Power law of the tail of the parameter distribution of deep linear nets. The dashed lines show the upper $(-7/2)$ and lower (-5) bound of the exponent of the tail. The predicted power-law scaling agrees with the experiment, and the exponent decreases as the theory predicts. **Mid**: training loss of a tanh network. $D = 0$ is the case where only the input weight is trained, and $D = 1$ is the case where both input and output layers are trained. For $D = 0$, the model norm increases as the model loses stability. For $D = 1$, a "fluctuation inversion" effect appears. The fluctuation of the model vanishes before it loses stability. **Right**: performance of fully connected tanh nets on MNIST. Scaling the learning rate as $1/D$ keeps the model performance relatively unchanged.

zero. Therefore, we should increase the model width as we increase the depth. In the second case, d/D is a constant and can thus be absorbed into the definition of T and is the only limit where we obtain a nontrivial distribution with a finite spread. If $d = \Omega(D)$, the distribution becomes a delta distribution at the global minimum of the loss landscape, $p(v) = \delta(v - \beta_2/\beta_1)$ and achieves the global minimum.

4.4 Implication for Variational Bayesian Learning

One of the major implications of the analytical solution we found for machine learning practice is the inappropriateness of using SGD to approximate a Bayesian posterior. Because every SGD iteration can be regarded as a sampling of the model parameters. A series of recent works have argued that the stationary distribution can be used as an approximation of the Bayesian posterior for fast variational inference [21, 3], $p_{\rm Bayes}(\theta) \approx p_{\rm SGD}(\theta)$, a method that has been used for a wide variety of applications [12]. However, our result implies that such an approximation is likely to fail. Common in Bayesian deep learning, we interpret the per-sample loss as the log probability and the weight decay as a Gaussian prior over the parameters, the true model parameters have a log probability of

$$\log p_{\rm Bayes}(\theta|x) \propto \ell(\theta, x) + \gamma\|\theta\|^2. \tag{17}$$

This distribution has a nonzero measure everywhere for any differentiable loss. However, the distribution for SGD in Eq.(14) has a zero probability density almost everywhere because a 1d subspace has a zero Lebesgue measure in a high-dimensional space. This implies that the KL divergence between the two distributions (either ${\rm KL}(p_{\rm Bayes}||p_{\rm SGD})$ or ${\rm KL}(p_{\rm SGD}||p_{\rm Bayes})$) is infinite. Therefore, we can infer that in the information-theoretic sense, $p_{\rm SGD}$ cannot be used to approximate $p_{\rm Bayes}$.

5 Discussion

In this work, we first showed that SGD systematically moves towards a balanced solution when rescaling symmetry exists, a result we termed the law of balance. Applying the law of balance, we have characterized the stationary distribution of SGD analytically, which is an unanswered fundamental problem in the study of SGD. This is the first analytical expression for a globally nonconvex and beyond quadratic loss without the need for any approximation. With this solution, we have discovered many phenomena that could be relevant to deep learning that were previously unknown. We found that SGD only has probability of exploring a one-dimensional submanifold even for a very-dimensional problem, ignoring all irrelevant directions. We applied our theory to the important problem of variational inference and showed that it is, in general, not appropriate to approximate the posterior with SGD, at least when any symmetry is present in the model. If one really wants to use SGD for variational inference, special care is required to at least remove symmetries from the loss function, which could be an interesting future problem. Our theory is limited, as the model we solved is only a minimal model of reality, and it would be interesting to consider more realistic models in the future. Also, it would be interesting to extend the law of balance to a broader class of symmetries.

References

[1] John C Baez and Brendan Fong. A noether theorem for markov processes. *Journal of Mathematical Physics*, 54(1):013301, 2013.

[2] Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023.

[3] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

[4] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

[5] Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

[6] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.

[7] Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023.

[8] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1965–2058. PMLR, 15–19 Aug 2021.

[9] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in neural information processing systems*, 31, 2018.

[10] Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *Advances in Neural Information Processing Systems*, 31, 2018.

[11] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

[12] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[13] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29:586–594, 2016.

[14] Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. *arXiv preprint arXiv:2012.04728*, 2020.

[15] Jonas Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31(4):39, 2021.

[16] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[17] 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*, 33:14544–14555, 2020.

[18] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes), 2021.

[19] Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent, 2021.

[20] Agnieszka B Malinowska and Moulay Rchid Sidi Ammi. Noether's theorem for control problems on time scales. *arXiv preprint arXiv:1406.0705*, 2014.

[21] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18:1–35, 2017.

[22] John C Mauro, Prabhat K Gupta, and Roger J Loucks. Continuously broken ergodicity. *The Journal of chemical physics*, 126(18), 2007.

[23] Tetsuya Misawa. Noether's theorem in symmetric stochastic calculus of variations. *Journal of mathematical physics*, 29(10):2178–2180, 1988.

[24] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of sgd. In *International Conference on Machine Learning*, pages 15959–15975. PMLR, 2022.

[25] Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295. PMLR, 2022.

[26] Richard G Palmer. Broken ergodicity. *Advances in Physics*, 31(6):669–735, 1982.

[27] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

[28] Hannes Risken and Hannes Risken. *Fokker-planck equation*. Springer, 1996.

[29] Tomasz Rolski, Hanspeter Schmidli, Volker Schmidt, and Jozef L Teugels. *Stochastic processes for insurance and finance*. John Wiley & Sons, 2009.

[30] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[31] N. Shirish Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ArXiv e-prints*, September 2016.

[32] Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. *Stochastic Systems*, 10(2):124–151, 2020.

[33] Hidenori Tanaka and Daniel Kunin. Noether's learning dynamics: Role of symmetry breaking in neural networks, 2021.

[34] D Thirumalai and Raymond D Mountain. Activated dynamics, loss of ergodicity, and transport in supercooled liquids. *Physical Review E*, 47(1):479, 1993.

[35] Christopher J Turner, Alexios A Michailidis, Dmitry A Abanin, Maksym Serbyn, and Zlatko Papić. Weak ergodicity breaking from quantum many-body scars. *Nature Physics*, 14(7):745–749, 2018.

[36] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect, 2022.

[39] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part ii: Continuous time analysis. *Journal of Nonlinear Science*, 34(1):1–45, 2024.

[40] Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

[41] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

[42] Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. *arXiv preprint arXiv:1810.00004*, 2018.

[43] Liu Ziyin, Botao Li, and Xiangming Meng. Exact solutions of a deep linear network. In *Advances in Neural Information Processing Systems*, 2022.

[44] Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022.

[45] Liu Ziyin and Masahito Ueda. Exact phase transitions in deep learning. *arXiv preprint arXiv:2205.12510*, 2022.

A Theoretical Considerations

A.1 Background

A.1.1 Ito's Lemma

Let us consider the following stochastic differential equation (SDE) for a Wiener process $W(t)$:

$$dX_t = \mu_t dt + \sigma_t dW(t). \tag{18}$$

We are interested in the dynamics of a generic function of X_t. Let $Y_t = f(t, X_t)$; Ito's lemma states that the SDE for the new variable is

$$df(t, X_t) = \left(\frac{\partial f}{\partial t} + \mu_t \frac{\partial f}{\partial X_t} + \frac{\sigma_t^2}{2}\frac{\partial^2 f}{\partial X_t^2}\right) dt + \sigma_t \frac{\partial f}{\partial x} dW(t). \tag{19}$$

Let us take the variable $Y_t = X_t^2$ as an example. Then the SDE is

$$dY_t = \left(2\mu_t X_t + \sigma_t^2\right) dt + 2\sigma_t X_t dW(t). \tag{20}$$

Let us consider another example. Let two variables X_t and Y_t follow

$$\begin{aligned} dX_t &= \mu_t dt + \sigma_t dW(t), \\ dY_t &= \lambda_t dt + \phi_t dW(t). \end{aligned} \tag{21}$$

The SDE of $X_t Y_t$ is given by

$$d(X_t Y_t) = (\mu_t Y_t + \lambda_t X_t + \sigma_t \phi_t) dt + (\sigma_t Y_t + \phi_t X_t) dW(t). \tag{22}$$

A.1.2 Fokker Planck Equation

The general SDE of a 1d variable X is given by:

$$dX = -\mu(X) dt + B(X) dW(t). \tag{23}$$

The time evolution of the probability density $P(x, t)$ is given by the Fokker-Planck equation:

$$\frac{\partial P(X,t)}{\partial t} = -\frac{\partial}{\partial X} J(X, t), \tag{24}$$

where $J(X, t) = \mu(X)P(X, t) + \frac{1}{2}\frac{\partial}{\partial X}[B^2(X)P(X, t)]$. The stationary distribution satisfying $\partial P(X, t)/\partial t = 0$ is

$$P(X) \propto \frac{1}{B^2(X)} \exp\left[-\int dX \frac{2\mu(X)}{B^2(X)}\right] := \tilde{P}(X), \tag{25}$$

which gives a solution as a Boltzmann-type distribution if B is a constant. We will apply Eq. (25) to determine the stationary distributions in the following sections.

A.2 Proof of Theorem 3.1

Proof. We omit writing v in the argument unless necessary. By definition of the symmetry $\ell(\mathbf{u}, \mathbf{w}, x) = \ell(\lambda\mathbf{u}, \mathbf{w}/\lambda, x)$, we obtain its infinitesimal transformation $\ell(\mathbf{u}, \mathbf{w}, x) = \ell((1+\epsilon)\mathbf{u}, (1-\epsilon)\mathbf{w}/\lambda, x)$. Expanding this to first order in ϵ, we obtain

$$\sum_i u_i \frac{\partial \ell}{\partial u_i} = \sum_j w_j \frac{\partial \ell}{\partial w_j}. \tag{26}$$

The equations of motion are

$$\frac{du_i}{dt} = -\frac{\partial \ell}{\partial u_i}, \tag{27}$$

$$\frac{dw_j}{dt} = -\frac{\partial \ell}{\partial w_j}. \tag{28}$$

Using Ito's lemma, we can find the equations governing the evolutions of u_i^2 and w_j^2:

$$\frac{du_i^2}{dt} = 2u_i \frac{du_i}{dt} + \frac{(du_i)^2}{dt} = -2u_i \frac{\partial \ell}{\partial u_i} + TC_i^u,$$
$$\frac{dw_j^2}{dt} = 2w_j \frac{dw_j}{dt} + \frac{(dw_j)^2}{dt} = -2w_j \frac{\partial \ell}{\partial w_j} + TC_j^w, \tag{29}$$

where $C_i^u = \mathrm{Var}[\frac{\partial \ell}{\partial u_i}]$ and $C_j^w = \mathrm{Var}[\frac{\partial \ell}{\partial w_j}]$. With Eq. (26), we obtain

$$\frac{d}{dt}(||u||^2 - ||w||^2) = -T(\sum_j C_j^w - \sum_i C_i^u) = -T\left(\sum_j \mathrm{Var}\left[\frac{\partial \ell}{\partial w_j}\right] - \sum_i \mathrm{Var}\left[\frac{\partial \ell}{\partial u_i}\right]\right). \tag{30}$$

Due to the rescaling symmetry, the loss function can be considered as a function of the matrix uw^T. Here we define a new loss function as $\tilde{\ell}(u_i w_j) = \ell(u_i, w_j)$. Hence, we have

$$\frac{\partial \ell}{\partial w_j} = \sum_i u_i \frac{\partial \tilde{\ell}}{\partial (u_i w_j)}, \frac{\partial \ell}{\partial u_i} = \sum_j w_j \frac{\partial \tilde{\ell}}{\partial (u_i w_j)}. \tag{31}$$

We can rewrite Eq. (30) into

$$\frac{d}{dt}(||u||^2 - ||w||^2) = -T(u^T C_1 u - w^T C_2 w),, \tag{32}$$

where

$$(C_1)_{ij} = \mathbb{E}\left[\sum_k \frac{\partial \tilde{\ell}}{\partial (u_i w_k)} \frac{\partial \tilde{\ell}}{\partial (u_j w_k)}\right] - \sum_k \mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial (u_i w_k)}\right] \mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial (u_j w_k)}\right],$$
$$\equiv \mathbb{E}[A^T A] - \mathbb{E}[A^T]\mathbb{E}[A] \tag{33}$$
$$(C_2)_{kl} = \mathbb{E}\left[\sum_i \frac{\partial \tilde{\ell}}{\partial (u_i w_k)} \frac{\partial \tilde{\ell}}{\partial (u_i w_l)}\right] - \sum_i \mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial (u_i w_k)}\right] \mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial (u_i w_l)}\right]$$
$$\equiv \mathbb{E}[AA^T] - \mathbb{E}[A]\mathbb{E}[A^T], \tag{34}$$

where

$$(A)_{ik} \equiv \frac{\partial \tilde{\ell}}{\partial (u_i w_k)}. \tag{35}$$

The proof is thus complete. □

A.3 Second-order Law of Balance

Considering the modified loss function:

$$\ell_{\text{tot}} = \ell + \frac{1}{4}T||\nabla L||^2. \tag{36}$$

In this case, the Langevin equations become

$$dw_j = -\frac{\partial \ell}{\partial w_j} dt - \frac{1}{4}T \frac{\partial ||\nabla L||^2}{\partial w_j}, \tag{37}$$
$$du_i = -- \frac{\partial \ell}{\partial u_i} dt - \frac{1}{4}T \frac{\partial ||\nabla L||^2}{\partial u_i}. \tag{38}$$

Hence, the modified SDEs of u_i^2 and w_j^2 can be rewritten as

$$\frac{du_i^2}{dt} = 2u_i \frac{du_i}{dt} + \frac{(du_i)^2}{dt} = -2u_i \frac{\partial \ell}{\partial u_i} + +TC_i^u - \frac{1}{2}Tu_i \nabla_{u_i} |\nabla L|^2, \tag{39}$$
$$\frac{dw_j^2}{dt} = 2w_j \frac{dw_j}{dt} + \frac{(dw_j)^2}{dt} = -2w_j \frac{\partial \ell}{\partial w_j} + TC_j^w - \frac{1}{2}Tw_j \nabla_{w_j} |\nabla L|^2. \tag{40}$$

In this section, we consider the effects brought by the last term in Eqs. (39) and (40). From the infinitesimal transformation of the rescaling symmetry:

$$\sum_j w_j \frac{\partial \ell}{\partial w_j} = \sum_i u_i \frac{\partial \ell}{\partial u_i}, \tag{41}$$

we take the derivative of both sides of the equation and obtain

$$\frac{\partial L}{\partial u_i} + \sum_j u_j \frac{\partial^2 L}{\partial u_i \partial u_j} = \sum_j w_j \frac{\partial^2 L}{\partial u_i \partial w_j}, \tag{42}$$

$$\sum_j u_j \frac{\partial^2 L}{\partial w_i \partial u_j} = \frac{\partial L}{\partial w_i} + \sum_j w_j \frac{\partial^2 L}{\partial w_i \partial w_j}, \tag{43}$$

where we take the expectation to ℓ at the same time. By substituting these equations into Eqs. (39) and (40), we obtain

$$\frac{d\|u\|^2}{dt} - \frac{d\|w\|\|^2}{dt} = T\sum_i (C_i^u + (\nabla_{u_i} L)^2) - T\sum_j (C_j^w + (\nabla_{w_j} L)^2). \tag{44}$$

Then following the procedure in Appendix. A.2, we can rewrite Eq. (44) as

$$\begin{aligned}\frac{d\|u\|^2}{dt} - \frac{d\|w\|^2}{dt} &= -T(u^T C_1 u + u^T D_1 u - w^T C_2 w - w^T D_2 w) \\ &= -T(u^T E_1 u - w^T E_2 w),\end{aligned} \tag{45}$$

where

$$(D_1)_{ij} = \sum_k \mathbb{E}\left[\frac{\partial \ell}{\partial (u_i w_k)}\right]\mathbb{E}\left[\frac{\partial \ell}{\partial (u_j w_k)}\right], \tag{46}$$

$$(D_2)_{kl} = \sum_i \mathbb{E}\left[\frac{\partial \ell}{\partial (u_i w_k)}\right]\mathbb{E}\left[\frac{\partial \ell}{\partial (u_i w_l)}\right], \tag{47}$$

$$(E_1)_{ij} = \mathbb{E}\left[\sum_k \frac{\partial \ell}{\partial (u_i w_k)}\frac{\partial \ell}{\partial (u_j w_k)}\right], \tag{48}$$

$$(E_2)_{kl} = \mathbb{E}\left[\sum_i \frac{\partial \ell}{\partial (u_i w_k)}\frac{\partial \ell}{\partial (u_i w_l)}\right]. \tag{49}$$

For one-dimensional parameters u, w, Eq. (45) is reduced to

$$\frac{d}{dt}(u^2 - w^2) = -\mathbb{E}\left[\left(\frac{\partial \ell}{\partial (uw)}\right)^2\right](u^2 - w^2). \tag{50}$$

Therefore, we can see this loss modification increases the speed of convergence. Now, we move to the stationary distribution of the parameter v. At the stationarity, if $u_i = -w_i$, we also have the distribution $P(v) = \delta(v)$ like before. However, when $u_i = w_i$, we have

$$\frac{dv}{dt} = -4v(\beta_1 v - \beta_2) + 4Tv(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3) - 4\beta_1^2 Tv(\beta_1 v - \beta_2)(3\beta_1 v - \beta_2) + 4v\sqrt{T(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\frac{dW}{dt}. \tag{51}$$

Hence, the stationary distribution becomes

$$P(v) \propto \frac{v^{\beta_2/2\alpha_3 T - 3/2 - \beta_2^2/2\alpha_3}}{(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)^{1+\beta_2/4T\alpha_3 + K_1}} \exp\left(-\left(\frac{1}{2T}\frac{\alpha_3\beta_1 - \alpha_2\beta_2}{\alpha_3\sqrt{\Delta}} + K_2\right)\arctan\frac{\alpha_1 v - \alpha_2}{\sqrt{\Delta}}\right), \tag{52}$$

where

$$\begin{aligned}K_1 &= \frac{3\alpha_3\beta_1^2 - \alpha_1\beta_2^2}{4\alpha_1\alpha_3}, \\ K_2 &= \frac{3\alpha_2\alpha_3\beta_1^2 - 4\alpha_1\alpha_3\beta_1\beta_2 + \alpha_1\alpha_2\beta_2^2}{2\alpha_1\alpha_3\sqrt{\Delta}}.\end{aligned} \tag{53}$$

From the expression above we can see $K_1 \ll 1 + \beta_2/4T\alpha_3$ and $K_2 \ll (\alpha_3\beta_1 - \alpha_2\beta_2)/2T\alpha_3\sqrt{\Delta}$. Hence, the effect of modification can only be seen in the term proportional to v. The phase transition point is modified as

$$T_c = \frac{\beta_2}{\alpha_3 + \beta_2^2}. \tag{54}$$

Compared with the previous result $T_c = \frac{\beta_2}{\alpha_3}$, we can see the effect of the loss modification is $\alpha_3 \to \alpha_3 + \beta_2^2$, or equivalently, $\mathrm{Var}[xy] \to \mathbb{E}[x^2y^2]$. This effect can be seen from E_1 and E_2.

A.4 Proof of Theorem 3.2

Proof. For any i, one can obtain the expressions of $C_1^{(i)}$ and $C_2^{(i)}$ from Theorem 3.1 as

$$\begin{aligned}
(C_1^{(i)})_{\alpha_1,\alpha_2} &= 4p_i\mathbb{E}_i\left[||\tilde{x}||^2(\sum_{j=1}^d u_j^{\alpha_1}v_j^T\tilde{x} - y^{\alpha_1})(\sum_{j=1}^d u_j^{\alpha_2}v_j^T\tilde{x} - y^{\alpha_2})\right] - 4p_i^2\sum_\beta \mathbb{E}_i\left[\tilde{x}^\beta(\sum_{j=1}^d u_j^{\alpha_1}v_j^T\tilde{x} - y^{\alpha_1})\right]\mathbb{E}_i\left[\tilde{x}^\beta(\sum_{j=1}^d u_j^{\alpha_2}v_j^T\tilde{x} -\right.\\
&= 4p_i\mathbb{E}_i\left[||\tilde{x}||^2 r^{\alpha_1}r^{\alpha_2}\right] - 4p_i^2\sum_\beta \mathbb{E}_i\left[\tilde{x}^\beta r^{\alpha_1}\right]\mathbb{E}_i\left[\tilde{x}^\beta r^{\alpha_2}\right],
\end{aligned} \tag{55}$$

$$\begin{aligned}
(C_2^{(i)})_{\beta_1,\beta_2} &= 4\mathbb{E}_i\left[\tilde{x}^{\beta_1}\tilde{x}^{\beta_2}||\sum_{j=1}^d u_jv_j^T\tilde{x} - y||^2\right] - 4\sum_\alpha \mathbb{E}_i\left[\tilde{x}^{\beta_1}(\sum_{j=1}^d u_j^\alpha v_j^T\tilde{x} - y^\alpha)\right]\mathbb{E}_i\left[\tilde{x}^{\beta_2}(\sum_{j=1}^d u_j^\alpha v_j^T\tilde{x} - y^\alpha)\right]\\
&= 4p_i\mathbb{E}_i\left[||r||^2\tilde{x}^{\beta_1}\tilde{x}^{\beta_2}\right] - 4p_i^2\sum_\alpha \mathbb{E}_i\left[\tilde{x}^{\beta_1}r^\alpha\right]\mathbb{E}_i\left[\tilde{x}^{\beta_2}r^\alpha\right],
\end{aligned} \tag{56}$$

where we use the notation $r^\alpha := \sum_{j=1}^d u_j^\alpha v_j^T\tilde{x} - y^\alpha, \tilde{x} := (x^T, 1)^T, v_i = (w_i^T, b_i)^T$ and $\mathbb{E}_i[O] := \mathbb{E}[O|w_i^Tx + b_i > 0]$.

We start with showing that $C_1^{(1)}$ is full-rank. Let m be an arbitrary unit vector in $\mathbb{R}^{d_u}$. We have that

$$\begin{aligned}
m^TC_1^{(i)}m &= 4p_i\mathbb{E}_i\left[||\tilde{x}||^2(m^Tr)^2\right] - 4p_i^2\sum_\beta \mathbb{E}_i\left[\tilde{x}^\beta(m^Tr)\right]\mathbb{E}_i\left[\tilde{x}^\beta(m^Tr)\right]\\
&\geq 4p_i^2\mathbb{E}_i\left[||\tilde{x}||^2(m^Tr)^2\right] - 4p_i^2\sum_\beta \mathbb{E}_i\left[\tilde{x}^\beta(m^Tr)\right]\mathbb{E}_i\left[\tilde{x}^\beta(m^Tr)\right]\\
&= 4p_i^2\sum_\beta \mathrm{Var}_i[\tilde{x}^\beta m^Tr]\\
&= 4p_i^2\sum_\beta[\mathrm{Var}_i[\tilde{x}^\beta m^T(g(x) - \sum_{j=1}^d u_jv_j^T\tilde{x})] + \mathrm{Var}_i[\tilde{x}^\beta m^T\epsilon] - 2\mathrm{Cov}_i[\tilde{x}^\beta m^T(g(x) - \sum_{j=1}^d u_jv_j^T\tilde{x}), \tilde{x}^\beta m^T\epsilon]]\\
&\geq 4p_i^2\sum_\beta \mathrm{Var}_i[\tilde{x}^\beta m^T\epsilon] > 0,
\end{aligned} \tag{57}$$

where the last inequality follows from

$$\begin{aligned}
&\mathrm{Cov}[\tilde{x}^\beta m^T(g(x) - \sum_{j=1}^d u_jv_j^T\tilde{x}), \tilde{x}^\beta m^T\epsilon]\\
=&\mathbb{E}_i[(\tilde{x}^\beta)^2m^T(g(x) - \sum_{j=1}^d u_jv_j^T\tilde{x})m^T\epsilon] - \mathbb{E}_i[\tilde{x}^\beta m^T(g(x) - \sum_{j=1}^d u_jv_j^T\tilde{x})]\mathbb{E}_i[\tilde{x}^\beta m^T\epsilon]\\
=&0.
\end{aligned} \tag{58}$$

Here we denote that $\mathrm{Var}_i[O] := \mathbb{E}_i[O^2] - \mathbb{E}_i[O]^2$ and $\mathrm{Cov}_i[O_1, O_2] := \mathbb{E}_i[O_1O_2] - \mathbb{E}_i[O_1]\mathbb{E}_i[O_2]$.

For $C_2^{(i)}$, we let the vector $\tilde{n} := (n^T, n_f)^T$ be a unit vector in $\mathbb{R}^{d_w+1}$, yielding

$$\begin{aligned}
\tilde{n}^TC_2^{(i)}\tilde{n} &= 4p_i\mathbb{E}_i\left[||r||^2(\tilde{n}^T\tilde{x})^2\right] - 4p_i^2\sum_\alpha \mathbb{E}_i\left[r^\alpha(\tilde{n}^T\tilde{x})\right]\mathbb{E}_i\left[r^\alpha(\tilde{n}^T\tilde{x})\right]\\
&\geq 4p_i^2\mathbb{E}_i\left[||r||^2(\tilde{n}^T\tilde{x})^2\right] - 4p_i^2\sum_\alpha \mathbb{E}_i\left[r^\alpha(\tilde{n}^T\tilde{x})\right]\mathbb{E}_i\left[r^\alpha(\tilde{n}^T\tilde{x})\right]\\
&= 4p_i^2\sum_\alpha \mathrm{Var}_i[r^\alpha\tilde{n}^T\tilde{x}].
\end{aligned} \tag{59}$$

Note that this quantity can be decomposed as

$$
\begin{aligned}
\sum_{\alpha} \mathrm{Var}_i[r^\alpha \tilde{n}^T \tilde{x}] &= \sum_{\alpha} \mathrm{Var}_i[(g^\alpha(x) - \sum_{j=1}^{d} u_j^\alpha v_j^T \tilde{x} + \epsilon^\alpha)(\tilde{n}^T \tilde{x})] \\
&= \sum_{\alpha} \mathrm{Var}_i[(g^\alpha(x) - \sum_{j=1}^{d} u_j^\alpha v_j^T \tilde{x})(n^T x + n_f)] + \sum_{\alpha} \mathrm{Var}_i[\epsilon^\alpha (n^T x + n_f)] \\
&\quad - 2 \sum_{\alpha} \mathrm{Cov}_i[(g^\alpha(x) - \sum_{j=1}^{d} u_j^\alpha v_j^T \tilde{x})(n^T x + n_f), \epsilon^\alpha (n^T x + n_f)]. \qquad (60)
\end{aligned}
$$

The covariance term vanishes because

$$
\begin{aligned}
&\mathrm{Cov}[(g^\alpha(x) - \sum_{j=1}^{d} u_j^\alpha v_j^T \tilde{x})(n^T x + n_f), \epsilon^\alpha (n^T x + n_f)] \\
=&\mathbb{E}_i[(g^\alpha(x) - \sum_{j=1}^{d} u_j^\alpha v_j^T \tilde{x}) \epsilon^\alpha (n^T x + n_f)^2] - \mathbb{E}_i[(g^\alpha(x) - \sum_{j=1}^{d} u_j^\alpha v_j^T \tilde{x})(n^T x + n_f)] \mathbb{E}_i[\epsilon^\alpha (n^T x + n_f)] \\
=&0. \qquad (61)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\tilde{n}^T C_2^{(i)} \tilde{n} &\geq \sum_{\alpha} \mathrm{Var}_i[(g^\alpha(x) - \sum_{j=1}^{d} u_j^\alpha v_j^T \tilde{x})(n^T x + n_f)] + \sum_{\alpha} \mathrm{Var}_i[\epsilon^\alpha (n^T x + n_f)] \\
&\geq \sum_{\alpha} \mathrm{Var}_i[\epsilon^\alpha (n^T x + n_f)] \\
&= \sum_{\alpha} \mathrm{Var}_i[\epsilon^\alpha] \mathrm{Var}_i[(n^T x + n_f)] + \sum_{\alpha} (\mathrm{Var}_i[\epsilon^\alpha] \mathbb{E}_i[(n^T x + n_f)^2] + \mathrm{Var}_i[n^T x + n_f] \mathbb{E}_i[(\epsilon^\alpha)^2]) \\
&\geq \sum_{\alpha} \mathrm{Var}_i[\epsilon^\alpha] \mathbb{E}_i[(n^T x + n_f)^2] > 0, \qquad (62)
\end{aligned}
$$

where the penultimate inequality follows from the fact that ϵ is independent of x. Hence, both the matrices $C_1^{(i)}$ and $C_2^{(i)}$ are full-rank. The proof is completed. □

A.5 Derivation of Eq. (4)

We here prove inequality (4). At stationarity, $d(\|u\|^2 - \|w\|^2)/dt = 0$, indicating

$$
\lambda_{1M} \|u\|^2 - \lambda_{2m} \|w\|^2 \geq 0, \ \lambda_{1m} \|u\|^2 - \lambda_{2M} \|w\|^2 \leq 0. \qquad (63)
$$

The first inequality in Eq. (63) gives the solution

$$
\frac{\|u\|^2}{\|w\|^2} \geq \frac{\lambda_{2m}}{\lambda_{1M}}. \qquad (64)
$$

The second inequality in Eq. (63) gives the solution

$$
\frac{\|u\|^2}{\|w\|^2} \leq \frac{\lambda_{2M}}{\lambda_{1m}}. \qquad (65)
$$

Combining these two results, we obtain

$$
\frac{\lambda_{2m}}{\lambda_{1M}} \leq \frac{\|u\|^2}{\|w\|^2} \leq \frac{\lambda_{2M}}{\lambda_{1m}}, \qquad (66)
$$

which is Eq. (4).

A.6 Proof of Theorem 4.1

Proof. This proof is based on the fact that if a certain condition is satisfied for all trajectories with probability 1, this condition is satisfied by the stationary distribution of the dynamics with probability 1.

Let us first consider the case of $D > 1$. We first show that any trajectory satisfies at least one of the following five conditions: for any i, (i) $v_i \to 0$, (ii) $L(\theta) \to 0$, or (iii) for any $k \neq l$, $(u_i^{(k)})^2 - (u_i^{(l)})^2 \to 0$.

The SDE for $u_i^{(k)}$ is

$$\frac{du_i^{(k)}}{dt} = -2\frac{v_i}{u_i^{(k)}}(\beta_1 v - \beta_2) + 2\frac{v_i}{u_i^{(k)}}\sqrt{\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\frac{dW}{dt}, \tag{67}$$

where $v_i := \prod_{k=1}^D u_i^{(k)}$, and so $v = \sum_i v_i$. There exists rescaling symmetry between $u_i^{(k)}$ and $u_i^{(l)}$ for $k \neq l$. By the law of balance, we have

$$\frac{d}{dt}[(u_i^{(k)})^2 - (u_i^{(l)})^2] = -T[(u_i^{(k)})^2 - (u_i^{(l)})^2]\text{Var}\left[\frac{\partial \ell}{\partial(u_i^{(k)}u_i^{(l)})}\right], \tag{68}$$

where

$$\text{Var}\left[\frac{\partial \ell}{\partial(u_i^{(k)}u_i^{(l)})}\right] = (\frac{v_i}{u_i^{(k)}u_i^{(l)}})^2(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3) \tag{69}$$

with $v_i/(u_i^{(k)}u_i^{(l)}) = \prod_{s\neq k,l} u_i^{(s)}$. In the long-time limit, $(u_i^{(k)})^2$ converges to $(u_i^{(l)})^2$ unless $\text{Var}\left[\frac{\partial \ell}{\partial(u_i^{(k)}u_i^{(l)})}\right] = 0$, which is equivalent to $v_i/(u_i^{(k)}u_i^{(l)}) = 0$ or $\alpha_1 v^2 - 2\alpha_2 v + \alpha_3 = 0$. These two conditions correspond to conditions (i) and (ii). The latter is because $\alpha_1 v^2 - 2\alpha_2 v + \alpha_3 = 0$ takes place if and only if $v = \alpha_2/\alpha_1$ and $\alpha_2^2 - \alpha_1\alpha_3 = 0$ together with $L(\theta) = 0$. Therefore, at stationarity, we must have conditions (i), (ii), or (iii).

Now, we prove that when (iii) holds, the condition 2-(b) in the theorem statement must hold: for $D = 1$, $(\log|v_i| - \log|v_j|) = c_0$ with $\text{sgn}(v_i) = \text{sgn}(v_j)$. When (iii) holds, there are two situations. First, if $v_i = 0$, we have $u_i^{(}k) = 0$ for all k, and v_i will stay 0 for the rest of the trajectory, which corresponds to condition (i).

If $v_i \neq 0$, we have $u_i^{(k)} \neq 0$ for all k. Therefore, the dynamics of v_i is

$$\frac{dv_i}{dt} = -2\sum_k \left(\frac{v_i}{u_i^{(k)}}\right)^2(\beta_1 v - \beta_2) + 2\sum_k \left(\frac{v_i}{u_i^{(k)}}\right)^2\sqrt{\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\frac{dW}{dt} + 4\sum_{k,l}\left(\frac{v_i^3}{(u_i^{(k)}u_i^{(l)})^2}\right)\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{70}$$

Comparing the dynamics of v_i and v_j for $i \neq j$, we obtain

$$\begin{aligned}\frac{dv_i/dt}{\sum_k (v_i/u_i^{(k)})^2} - \frac{dv_j/dt}{\sum_k (v_j/u_j^{(k)})^2} &= 4\left(\frac{\sum_{m,l} v_i^3/(u_i^{(m)}u_i^{(l)})^2}{\sum_k (v_i/u_i^{(k)})^2} - \frac{\sum_{m,l} v_j^3/(u_j^{(m)}u_j^{(l)})^2}{\sum_k (v_j/u_j^{(k)})^2}\right)\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)\\ &= 4\left(v_i\frac{\sum_{m,l} v_i^2/(u_i^{(m)}u_i^{(l)})^2}{\sum_k (v_i/u_i^{(k)})^2} - v_j\frac{\sum_{m,l} v_j^2/(u_j^{(m)}u_j^{(l)})^2}{\sum_k (v_j/u_j^{(k)})^2}\right)\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3).\end{aligned} \tag{71}$$

By condition (iii), we have $|u_i^{(0)}| = \cdots = |u_i^{(D)}|$, i.e., $(v_i/u_i^{(k)})^2 = (v_i^2)^{D/(D+1)}$ and $(v_i/u_i^{(m)}u_i^{(l)})^2 = (v_i^2)^{(D-1)/(D+1)}$.[5] Therefore, we obtain

$$\frac{dv_i/dt}{(D+1)(v_i^2)^{D/(D+1)}} - \frac{dv_j/dt}{(D+1)(v_j^2)^{D/(D+1)}} = \left(v_i\frac{D(v_i^2)^{(D-1)/(D+1)}}{2(v_i^2)^{D/(D+1)}} - v_j\frac{D(v_j^2)^{(D-1)/(D+1)}}{2(v_j^2)^{D/(D+1)}}\right)\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{72}$$

We first consider the case where v_i and v_j initially share the same sign (both positive or both negative). When $D > 1$, the left-hand side of Eq. (72) can be written as

$$\frac{1}{1-D}\frac{dv_i^{2/(D+1)-1}}{dt} + 4Dv_i^{1-2/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3) - \frac{1}{1-D}\frac{dv_j^{2/(D+1)-1}}{dt} - 4Dv_j^{1-2/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3), \tag{73}$$

[5] Here, we only consider the root on the positive real axis.

which follows from Ito's lemma:

$$\frac{dv_i^{2/(D+1)-1}}{dt} = \left(\frac{2}{D+1} - 1\right) v_i^{2/(D+1)-2}\frac{dv_i}{dt} + 2(\frac{2}{D+1} - 1)(\frac{2}{D+1} - 2)v_i^{2/(D+1)-3}\left(\sum_k (\frac{v_i}{u_i^{(k)}})^2\sqrt{\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\right)^2$$
$$= (\frac{2}{D+1} - 1)v_i^{2/(D+1)-2}\frac{dv_i}{dt} + 4D(D-1)v_i^{1-2/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{74}$$

Substitute in Eq. (72), we obtain Eq. (73).

Now, we consider the right-hand side of Eq. (72), which is given by

$$2Dv_i^{1-2/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3) - 2Dv_j^{1-2/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{75}$$

Combining Eq. (73) and Eq. (75), we obtain

$$\frac{1}{1-D}\frac{dv_i^{2/(D+1)-1}}{dt} - \frac{1}{1-D}\frac{dv_j^{2/(D+1)-1}}{dt} = -2D(v_i^{1-2/(D+1)} - v_j^{1-2/(D+1)})\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{76}$$

By defining $z_i = v_i^{2/(D+1)-1}$, we can further simplify the dynamics:

$$\frac{d(z_i - z_j)}{dt} = 2D(D-1)\left(\frac{1}{z_i} - \frac{1}{z_j}\right)\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)$$
$$= -2D(D-1)\frac{z_i - z_j}{z_i z_j}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{77}$$

Hence,

$$z_i(t) - z_j(t) = \exp\left[-\int dt \frac{2D(D-1)}{z_i z_j}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)\right]. \tag{78}$$

Therefore, if v_i and v_j initially have the same sign, they will decay to the same value in the long-time limit $t \to \infty$, which gives condition 2-(b). When v_i and v_j initially have different signs, we can write Eq. (72) as

$$\frac{d|v_i|/dt}{(D+1)(|v_i|^2)^{D/(D+1)}} + \frac{d|v_j|/dt}{(D+1)(|v_j|^2)^{D/(D+1)}} = \left(|v_i|\frac{D(|v_i|^2)^{(D-1)/(D+1)}}{2(|v_i|^2)^{D/(D+1)}} + |v_j|\frac{D(|v_j|^2)^{(D-1)/(D+1)}}{2(|v_j|^2)^{D/(D+1)}}\right)$$
$$\times \eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{79}$$

Hence, when $D > 1$, we simplify the equation with a similar procedure as

$$\frac{1}{1-D}\frac{d|v_i|^{2/(D+1)-1}}{dt} + \frac{1}{1-D}\frac{d|v_j|^{2/(D+1)-1}}{dt} = -2D(|v_i|^{1-2/(D+1)} + |v_j|^{1-2/(D+1)})\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{80}$$

Defining $z_i = |v_i|^{2/(D+1)-1}$, we obtain

$$\frac{d(z_i + z_j)}{dt} = 2D(D-1)\left(\frac{1}{z_i} + \frac{1}{z_j}\right)\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)$$
$$= 2D(D-1)\frac{z_i + z_j}{z_i z_j}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3), \tag{81}$$

which implies

$$z_i(t) + z_j(t) = \exp\left[\int dt \frac{2D(D-1)}{z_i z_j}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)\right]. \tag{82}$$

From this equation, we reach the conclusion that if v_i and v_j have different signs initially, one of them converges to 0 in the long-time limit $t \to \infty$, corresponding to condition 1 in the theorem statement. Hence, for $D > 1$, at least one of the conditions is always satisfied at $t \to \infty$.

Now, we prove the theorem for $D = 1$, which is similar to the proof above. The law of balance gives

$$\frac{d}{dt}[(u_i^{(1)})^2 - (u_i^{(2)})^2] = -T[(u_i^{(1)})^2 - (u_i^{(2)})^2]\mathrm{Var}\left[\frac{\partial \ell}{\partial(u_i^{(1)}u_i^{(2)})}\right]. \tag{83}$$

We can see that $|u_i^{(1)}| \to |u_i^{(2)}|$ takes place unless $\mathrm{Var}\left[\frac{\partial \ell}{\partial(u_i^{(1)} u_i^{(2)})}\right] = 0$, which is equivalent to $L(\theta) = 0$. This corresponds to condition (ii). Hence, if condition (ii) is violated, we need to prove condition (iii). In this sense, $|u_i^{(1)}| \to |u_i^{(2)}|$ occurs and Eq. (72) can be rewritten as

$$\frac{dv_i/dt}{|v_i|} - \frac{dv_j/dt}{|v_j|} = (\mathrm{sign}(v_i) - \mathrm{sign}(v_j))\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{84}$$

When v_i and v_j are both positive, we have

$$\frac{dv_i/dt}{v_i} - \frac{dv_j/dt}{v_j} = 0. \tag{85}$$

With Ito's lemma, we have

$$\frac{d\log(v_i)}{dt} = \frac{dv_i}{v_i dt} - 2\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{86}$$

Therefore, Eq. (85) can be simplified to

$$\frac{d(\log(v_i) - \log(v_j))}{dt} = 0, \tag{87}$$

which indicates that all v_i with the same sign will decay at the same rate. This differs from the case of $D > 2$ where all v_i decay to the same value. Similarly, we can prove the case where v_i and v_j are both negative.

Now, we consider the case where v_i is positive while v_j is negative and rewrite Eq. (84) as

$$\frac{dv_i/dt}{v_i} + \frac{d(|v_j|)/dt}{|v_j|} = 2\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{88}$$

Furthermore, we can derive the dynamics of v_j with Ito's lemma:

$$\frac{d\log(|v_j|)}{dt} = \frac{dv_i}{v_i dt} - 2\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{89}$$

Therefore, Eq. (88) takes the form of

$$\frac{d(\log(v_i) + \log(|v_j|))}{dt} = -2\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{90}$$

In the long-time limit, we can see $\log(v_i|v_j|)$ decays to $-\infty$, indicating that either v_i or v_j will decay to 0. This corresponds to condition 1 in the theorem statement. Combining Eq. (87) and Eq. (90), we conclude that all v_i have the same sign as $t \to \infty$, which indicates condition 2-(a) if conditions in item 1 are all violated. The proof is thus complete. □

A.7 Proof of Theorem 4.2

Proof. Following Eq. (70), we substitute $u_i^{(k)}$ with $v_i^{1/D}$ for arbitrary k and obtain

$$\begin{aligned}\frac{dv_i}{dt} =& -2(D+1)|v_i|^{2D/(D+1)}(\beta_1 v - \beta_2) + 2(D+1)|v_i|^{2D/(D+1)}\sqrt{\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\frac{dW}{dt}\\ &+ 2(D+1)Dv_i^3|v_i|^{-4/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3).\end{aligned} \tag{91}$$

With Eq. (78), we can see that for arbitrary i and j, v_i will converge to v_j in the long-time limit. In this case, we have $v = dv_i$ for each i. Then, the SDE for v can be written as

$$\begin{aligned}\frac{dv}{dt} =& -2(D+1)d^{2/(D+1)-1}|v|^{2D/(D+1)}(\beta_1 v - \beta_2) + 2(D+1)d^{2/(D+1)-1}|v|^{2D/(D+1)}\sqrt{\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\frac{dW}{dt}\\ &+ 2(D+1)Dd^{4/(D+1)-2}v^3|v|^{-4/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3).\end{aligned} \tag{92}$$

If $v > 0$, Eq. (92) becomes

$$\begin{aligned}\frac{dv}{dt} =& -2(D+1)d^{2/(D+1)-1}v^{2D/(D+1)}(\beta_1 v - \beta_2) + 2(D+1)d^{2/(D+1)-1}v^{2D/(D+1)}\sqrt{\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\frac{dW}{dt}\\ &+ 2(D+1)Dd^{4/(D+1)-2}v^{3-4/(D+1)}\eta(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3).\end{aligned} \tag{93}$$

Therefore, the stationary distribution of a general deep diagonal network is given by

$$p(v) \propto \frac{1}{v^{3(1-1/(D+1))}(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)} \exp\left(-\frac{1}{T}\int dv \frac{d^{1-2/(D+1)}(\beta_1 v - \beta_2)}{(D+1)v^{2D/(D+1)}(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)}\right). \tag{94}$$

If $v < 0$, Eq. (92) becomes

$$\begin{aligned}\frac{d|v|}{dt} =& -2(D+1)d^{2/(D+1)-1}|v|^{2D/(D+1)}(\beta_1|v| + \beta_2) - 2(D+1)d^{2/(D+1)-1}|v|^{2D/(D+1)}\sqrt{\eta(\alpha_1|v|^2 + 2\alpha_2|v| + \alpha_3)}\frac{dW}{dt}\\ &+ 2(D+1)Dd^{4/(D+1)-2}|v|^{3-4/(D+1)}\eta(\alpha_1|v|^2 + 2\alpha_2|v| + \alpha_3).\end{aligned} \tag{95}$$

The stationary distribution of $|v|$ is given by

$$p(|v|) \propto \frac{1}{|v|^{3(1-1/(D+1))}(\alpha_1 |v|^2 + 2\alpha_2 |v| + \alpha_3)} \exp\left(-\frac{1}{T}\int d|v| \frac{d^{1-2/(D+1)}(\beta_1 |v| + \beta_2)}{(D+1)|v|^{2D/(D+1)}(\alpha_1 |v|^2 + 2\alpha_2 |v| + \alpha_3)}\right). \tag{96}$$

Thus, we have obtained

$$p_\pm(|v|) \propto \frac{1}{|v|^{3(1-1/(D+1))}(\alpha_1 |v|^2 \mp 2\alpha_2 |v| + \alpha_3)} \exp\left(-\frac{1}{T}\int d|v| \frac{d^{1-2/(D+1)}(\beta_1 |v| \mp \beta_2)}{(D+1)|v|^{2D/(D+1)}(\alpha_1 |v|^2 \mp 2\alpha_2 |v| + \alpha_3)}\right). \tag{97}$$

Especially when $D = 1$, the distribution function can be simplified as

$$p_\pm(|v|) \propto \frac{|v|^{\pm\beta_2/2\alpha_3 T - 3/2}}{(\alpha_1|v|^2 \mp 2\alpha_2|v| + \alpha_3)^{1\pm\beta_2/4T\alpha_3}} \exp\left(-\frac{1}{2T}\frac{\alpha_3\beta_1 - \alpha_2\beta_2}{\alpha_3\sqrt{\Delta}} \arctan\frac{\alpha_1|v| \mp \alpha_2}{\sqrt{\Delta}}\right), \tag{98}$$

where we have used the integral

$$\int dv \frac{\beta_1 v \mp \beta_2}{\alpha_1 v^2 - 2\alpha_2 v + \alpha_3} = \frac{\alpha_3\beta_1 - \alpha_2\beta_2}{\alpha_3\sqrt{\Delta}} \arctan\frac{\alpha_1|v| \mp \alpha_2}{\sqrt{\Delta}} \pm \frac{\beta_2}{\alpha_3}\log(v) \pm \frac{\beta_2}{2\alpha_3}\log(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3). \tag{99}$$

Furthermore, we can also see that $p(v) = \delta(v)$ is also the stationary distribution of the Fokker-Planck equation of Eq. (93). Hence, the general stationary distribution of v can be expressed as

$$p^*(v) = (1-z)\delta(v) + zp_\pm(v). \tag{100}$$

The proof is complete. □

A.8 Analysis of the maximum probability point

To investigate the existence of the maximum point given in Eq. (16), we treat T as a variable and study whether $(\beta_1 - 10\alpha_2 T)^2 + 28\alpha_1 T(\beta_2 - 3\alpha_3 T) := A$ in the square root is always positive or not. When $T < \frac{\beta_2}{3\alpha_3} = T_c/3$, A is positive for arbitrary data. When $T > \frac{\beta_2}{3\alpha_3}$, we divide the discussion into several cases. First, when $\alpha_1\alpha_3 > \frac{25}{21}\alpha_2^2$, there always exists a root for the expression A. Hence, we find that

$$T = \frac{-5\alpha_2\beta_1 + 7\alpha_1\beta_2 + \sqrt{7}\sqrt{3\alpha_1\alpha_3\beta_1^2 - 10\alpha_1\alpha_2\beta_1\beta_2 + 7\alpha_1^2\beta_2^2}}{2(21\alpha_1\alpha_3 - 25\alpha_2^2)} := T^* \tag{101}$$

is a critical point. When $T_c/3 < T < T^*$, there exists a solution to the maximum condition. When $T > T^*$, there is no solution to the maximum condition.

The second case is $\alpha_2^2 < \alpha_1\alpha_3 < \frac{25}{21}\alpha_2^2$. In this case, we need to further compare the value between $5\alpha_2\beta_1$ and $7\alpha_1\beta_2$. If $5\alpha_2\beta_1 < 7\alpha_1\beta_2$, we have $A > 0$, which indicates that the maximum point exists. If $5\alpha_2\beta_1 > 7\alpha_1\beta_2$, we need to further check the value of minimum of A, which takes the form of

$$\min_T A(T) = \frac{(25\alpha_2^2 - 21\alpha_1\alpha_3)\beta_1^2 - (7\alpha_1\beta_2 - 5\alpha_2\beta_1)^2}{25\alpha_2^2 - 21\alpha_1\alpha_3}. \tag{102}$$

If $\frac{7\alpha_1}{5\alpha_2} < \frac{\beta_1}{\beta_2} < \frac{5\alpha_2 + \sqrt{25\alpha_2^2 - 21\alpha_1\alpha_3}}{3\alpha_3}$, the minimum of A is always positive and the maximum exists. However, if $\frac{\beta_1}{\beta_2} \geq \frac{5\alpha_2 + \sqrt{25\alpha_2^2 - 21\alpha_1\alpha_3}}{3\alpha_3}$, there is always a critical learning rate T^*. If

	without weight decay	with weight decay
single layer	$(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)^{-1-\frac{\beta_1}{2T\alpha_1}}$	$\alpha_1 (v-k)^{-2-\frac{(\beta_1+\gamma)}{T\alpha_1}}$
non-interpolation	$\frac{v^{\beta_2/2\alpha_3 T - 3/2}}{(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)^{1+\beta_2/4T\alpha_3}}$	$\frac{v^{S(\beta_2-\gamma)/2\alpha_3\lambda - 3/2}}{(\alpha_1 v^2 - 2\alpha_2 v + \alpha_3)^{1+(\beta_2-\gamma)/4T\alpha_3}}$
interpolation $y = kx$	$\frac{v^{-3/2+\beta_1/2T\alpha_1 k}}{(v-k)^{2+\beta_1/2T\alpha_1 k}}$	$\frac{v^{-3/2+\frac{1}{2T\alpha_1 k}(\beta_1 - \frac{\gamma}{k})}}{(v-k)^{2+\frac{1}{2T\alpha_1 k}(\beta_1-\frac{\gamma}{k})}} \exp(-\frac{\beta\gamma}{2T\alpha_1}\frac{1}{k(k-v)})$

Table 1: Summary of distributions $p(v)$ in a depth-1 neural network. Here, we show the distribution in the nontrivial subspace when the data x and y are positively correlated. The $\Theta(1)$ factors are neglected for concision.

$\frac{\beta_1}{\beta_2} = \frac{5\alpha_2 + \sqrt{25\alpha_2^2 - 21\alpha_1\alpha_3}}{3\alpha_3}$, there is only one critical learning rate as $T_c = \frac{5\alpha_2\beta_1 - 7\alpha_1\beta_2}{2(25\alpha_2^2 - 21\alpha_1\alpha_3)}$. When $T_c/3 < T < T^*$, there is a solution to the maximum condition, while there is no solution when $T > T^*$. If $\frac{\beta_1}{\beta_2} > \frac{5\alpha_2 + \sqrt{25\alpha_2^2 - 21\alpha_1\alpha_3}}{3\alpha_3}$, there are two critical points:

$$T_{1,2} = \frac{-5\alpha_2\beta_1 + 7\alpha_1\beta_2 \mp \sqrt{7}\sqrt{3\alpha_1\alpha_3\beta_1^2 - 10\alpha_1\alpha_2\beta_1\beta_2 + 7\alpha_1^2\beta_2^2}}{2(21\alpha_1\alpha_3 - 25\alpha_2^2)}. \tag{103}$$

For $T < T_1$ and $T > T_2$, there exists a solution to the maximum condition. For $T_1 < T < T_2$, there is no solution to the maximum condition. The last case is $\alpha_2^2 = \alpha_1\alpha_3 < \frac{25}{21}\alpha_2^2$. In this sense, the expression of A is simplified as $\beta_1^2 + 28\alpha_1\beta_2 T - 20\alpha_2\beta_1 T$. Hence, when $\frac{\beta_1}{\beta_2} < \frac{7\alpha_1}{5\alpha_2}$, there is no critical learning rate and the maximum always exists. Nevertheless, when $\frac{\beta_1}{\beta_2} > \frac{7\alpha_1}{5\alpha_2}$, there is always a critical learning rate as $T^* = \frac{\beta_1^2}{20\alpha_2\beta_1 - 28\alpha_1\beta_2}$. When $T < T^*$, there is a solution to the maximum condition, while there is no solution when $T > T^*$.

A.9 Other Cases for $D = 1$

The other cases are worth studying. For the interpolation case where the data is linear ($y = kx$ for some k), the stationary distribution is different and simpler. There exists a nontrivial fixed point for $\sum_i (u_i^2 - w_i^2)$: $\sum_j u_j w_j = \frac{\alpha_2}{\alpha_1}$, which is the global minimizer of L and also has a vanishing noise. It is helpful to note the following relationships for the data distribution when it is linear:

$$\begin{cases} \alpha_1 = \mathrm{Var}[x^2], \\ \alpha_2 = k\mathrm{Var}[x^2] = k\alpha_1, \\ \alpha_3 = k^2\alpha_1, \\ \beta_1 = \mathbb{E}[x^2], \\ \beta_2 = k\mathbb{E}[x^2] = k\beta_1. \end{cases} \tag{104}$$

Since the analysis of the Fokker-Planck equation is the same, we directly begin with the distribution function in Eq. (15) for $u_i = -w_i$ which is given by $P(|v|) \propto \delta(|v|)$. Namely, the only possible weights are $u_i = w_i = 0$, the same as the non-interpolation case. This is because the corresponding stationary distribution is

$$\begin{aligned} P(|v|) &\propto \frac{1}{|v|^2(|v|+k)^2} \exp\left(-\frac{1}{2T}\int d|v| \frac{\beta_1(|v|+k) + \alpha_1 \frac{1}{T}(|v|+k)^2}{\alpha_1 |v|(|v|+k)^2}\right) \\ &\propto |v|^{-\frac{3}{2} - \frac{\beta_1}{2T\alpha_1 k}} (|v|+k)^{-2+\frac{\beta_1}{2T\alpha_1 k}}. \end{aligned} \tag{105}$$

The integral of Eq. (105) with respect to $|v|$ diverges at the origin due to the factor $|v|^{\frac{3}{2}+\frac{\beta_1}{2T\alpha_1 k}}$.

For the case $u_i = w_i$, the stationary distribution is given from Eq. (15) as

$$\begin{aligned} P(v) &\propto \frac{1}{v^2(v-k)^2} \exp\left(-\frac{1}{2T}\int dv \frac{\beta_1(v-k) + \alpha_1 T(v-k)^2}{\alpha_1 v(v-k)^2}\right) \\ &\propto v^{-\frac{3}{2}+\frac{\beta_1}{2T\alpha_1 k}} (v-k)^{-2-\frac{\beta_1}{2T\alpha_1 k}}. \end{aligned} \tag{106}$$

Now, we consider the case of $\gamma \neq 0$. In the non-interpolation regime, when $u_i = -w_i$, the stationary distribution is still $p(v) = \delta(v)$. For the case of $u_i = w_i$, the stationary distribution is the same as in Eq. (15) after replacing β with $\beta_2' = \beta_2 - \gamma$. It still has a phase transition. The weight decay has the effect of shifting β_2 by $-\gamma$. In the interpolation regime, the stationary distribution is still $p(v) = \delta(v)$ when $u_i = -w_i$. However, when $u_i = w_i$, the phase transition still exists since the stationary distribution is

$$p(v) \propto \frac{v^{-\frac{3}{2}+\theta_2}}{(v-k)^{2+\theta_2}} \exp\left(-\frac{\beta_1 \gamma}{2T\alpha_1} \frac{1}{k(k-v)}\right), \tag{107}$$

where $\theta_2 = \frac{1}{2T\alpha_1 k}(\beta_1 - \frac{\gamma}{k})$. The phase transition point is $\theta_2 = 1/2$, which is the same as the non-interpolation one.

The last situation is rather special, which happens when $\Delta = 0$ but $y \neq kx$: $y = kx - c/x$ for some $c \neq 0$. In this case, the parameters α and β are the same as those given in Eq. (104) except for β_2:

$$\beta_2 = k\mathbb{E}[x^2] - kc = k\beta_1 - kc. \tag{108}$$

The corresponding stationary distribution is

$$P(|v|) \propto \frac{|v|^{-\frac{3}{2}-\phi_2}}{(|v|+k)^{2-\phi_2}} \exp\left(\frac{c}{2T\alpha_1} \frac{1}{k(k+|v|)}\right), \tag{109}$$

where $\phi_2 = \frac{1}{2T\alpha_1 k}(\beta_1 - c)$. Here, we see that the behavior of stationary distribution $P(|v|)$ is influenced by the sign of c. When $c < 0$, the integral of $P(|v|)$ diverges due to the factor $|v|^{-\frac{3}{2}-\phi_2} < |v|^{-3/2}$ and Eq. (109) becomes $\delta(|v|)$ again. However, when $c > 0$, the integral of $|v|$ may not diverge. The critical point is $\frac{3}{2} + \phi_2 = 1$ or equivalently: $c = \beta_1 + T\alpha_1 k$. This is because when $c < 0$, the data points are all distributed above the line $y = kx$. Hence, $u_i = -w_i$ can only give a trivial solution. However, if $c > 0$, there is the possibility to learn the negative slope k. When $0 < c < \beta_1 + T\alpha_1 k$, the integral of $P(|v|)$ still diverges and the distribution is equivalent to $\delta(|v|)$. Now, we consider the case of $u_i = w_i$. The stationary distribution is

$$P(|v|) \propto \frac{|v|^{-\frac{3}{2}+\phi_2}}{(|v|-k)^{2+\phi_2}} \exp\left(-\frac{c}{2T\alpha_1} \frac{1}{k-|v|}\right). \tag{110}$$

It also contains a critical point: $-\frac{3}{2} + \phi_2 = -1$, or equivalently, $c = \beta_1 - \alpha_1 kT$. There are two cases. When $c < 0$, the probability density only has support for $|v| > k$ since the gradient always pulls the parameter $|v|$ to the region $|v| > k$. Hence, the divergence at $|v| = 0$ is of no effect. When $c > 0$, the probability density has support on $0 < |v| < k$ for the same reason. Therefore, if $\beta_1 > \alpha_1 kT$, there exists a critical point $c = \beta_1 - \alpha_1 kT$. When $c > \beta_1 - \alpha_1 kT$, the distribution function $P(|v|)$ becomes $\delta(|v|)$. When $c < \beta_1 - \alpha_1 kT$, the integral of the distribution function is finite for $0 < |v| < k$, indicating the learning of the neural network. If $\beta_1 \leq \alpha_1 kT$, there will be no criticality and $P(|v|)$ is always equivalent to $\delta(|v|)$. The effect of having weight decay can be similarly analyzed, and the result can be systematically obtained if we replace β_1 with $\beta_1 + \gamma/k$ for the case $u_i = -w_i$ or replacing β_1 with $\beta_1 - \gamma/k$ for the case $u_i = w_i$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We believe that the abstract and introduction reflect the contributions and scope of the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have discussed the limitations of our work in the Discussion session at lines 369-371.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

Answer: [Yes]

Justification: We believe that the assumptions are clarified and complete proofs are provided for the theoretical parts.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: We believe that all of the experimental results are reproducable in our work.

Guidelines:

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

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

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

Answer: [No]

Justification: The code or data of the experiments are simple and easy to reproduce following the description in the main text.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: We have specified the training and test details in the captions of the experiments in Figs. 2,4, and 5.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [No]

   Justification: Here the dynamics is deterministic and there is no need to consider the error bars here.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments Compute Resources**

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

   Answer: [No]

   Justification: The experiments can be simply conducted on personal computers.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We have confirmed that the research is conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: Our work is a fundamental research on the dynamics of SGD and hence it does not have direct positive or negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [No]

    Justification: We believe there is no risks for misuse for the data and models.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer:[NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [No]

    Justification: Nothing introduced.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: We believe that neither the crowdsourcing nor the research with human subjects is included in our work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: Our work does not contain crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.