Parameter Symmetry and Noise Equilibrium of Stochastic Gradient Descent

Liu Ziyin
Massachusetts Institute of Technology,
NTT Research
ziyinl@mit.edu

Mingze Wang
Peking University
mingzewang@stu.pku.edu.cn

Hongchao Li
The University of Tokyo
lhc@cat.phys.s.u-tokyo.ac.jp

Lei Wu
Peking University
leiwu@math.pku.edu.cn

Abstract

Symmetries are abundant in the loss functions of neural networks, and understanding their impact on optimization algorithms is crucial for deep learning. We investigate the learning dynamics of Stochastic Gradient Descent (SGD) through the lens of exponential symmetries, a broad subclass of continuous symmetries in loss functions. Our analysis reveals that when gradient noise is imbalanced, SGD inherently drives model parameters toward a noise-balanced state, leading to the emergence of unique and attractive fixed points along degenerate directions. We prove that every parameter θ connects without barriers to a unique noise-balanced fixed point θ^*. This finding offers a unified perspective on how symmetry and gradient noise influence SGD. The theory provides novel insights into deep learning phenomena such as progressive sharpening/flattening and warmup, demonstrating that noise balancing is a key mechanism underlying these effects.

1 Introduction

Stochastic gradient descent (SGD) and its variants have become the cornerstone algorithms used in deep learning. In the continuous-time limit, the algorithm can be written as [17, 11, 19, 28, 7]:

$$\mathrm{d}\theta_t = -\nabla L(\theta_t)\,\mathrm{d}t + \sqrt{2\sigma^2\Sigma(\theta_t)}\,\mathrm{d}W_t, \tag{1}$$

where $\Sigma(\theta)$ is the covariance matrix of gradient noise (Section 2) with the prefactor $\sigma^2 = \eta/(2S)$ modeling the impact of a finite learning rate η and batch size S; W_t denotes the Brownian motion. When $\sigma = 0$, Eq. (1) corresponds to gradient descent (GD)[1]. However, SGD and GD can exhibit significantly different behaviors, often converging to solutions with different levels of performance [27, 31, 33, 20, 36]. Notably, even when $\sigma^2 \ll 1$, where we expect a close resemblance between SGD and GD over finite time [17], their long-time behaviors can still differ substantially [23]. These observations indicate that gradient noise can systematically bias the dynamics, and revealing its underlying mechanism is thus crucial for understanding the disparities between SGD and GD.

Contribution. In this paper, we study the how of SGD noise biases training through the lens of symmetry. Our key contributions are summarized as follows. We show that

1. when symmetry exists in the loss function, the dynamics of SGD can be precisely characterized and is different from GD along the degenerate direction;

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

[1] "Gradient descent" and "gradient flow" are used interchangeably as we work in the continuous-time limit.

2. the treatment of common symmetries, including the rescaling and scaling symmetry in deep learning, can be unified in a single theoretical framework that we call the exponential symmetry;
3. for any θ, every exponential symmetry implies the existence of a unique and attractive fixed point along the degenerate direction for SGD;
4. symmetry and balancing of noise can serve as novel mechanisms for important deep learning phenomena such as progressive sharpening/flattening and warmup.

2 Preliminaries

Setup and Notations. Let $\ell : \Omega \times \mathcal{Z} \mapsto \mathbb{R}$ denote the per-sample loss, with Ω and $\mathcal{Z}$ denoting the parameter and sample space, respectively. Here, $z \in \mathcal{Z}$ includes both the input and label and accordingly. We use $\mathbb{E}_z = \mathbb{E}$ to denote the expectation over a given training set. Therefore, $L(\theta) = \mathbb{E}_z[\ell(\theta, z)]$ is the empirical risk function. The covariance of gradient noise is given by

$$\Sigma(\theta) = \mathbb{E}_z[\nabla \ell(\theta, z)\nabla \ell(\theta, z)^\top] - \nabla L(\theta)\nabla L(\theta)^\top.$$

Additionally, we use $\Sigma_v(\theta) := \mathbb{E}_z[\nabla_v \ell(\theta, z)\nabla_v \ell(\theta, z)^\top] - \nabla_v L(\theta)\nabla_v L(\theta)^\top$ to denote the covariance of gradient noise impacting on the subset of parameters v. Denote by $(\theta_t)_{t\geq 0}$ the trajectory of SGD or GD. For any $h : \Omega \mapsto \mathbb{R}$, we write $h_t = h(\theta_t)$ and $\dot{h}(\theta_t) = \frac{\mathrm{d}}{\mathrm{d}t}h(\theta_t)$ for brevity. When the context is clear, we also use $\ell(\theta)$ to denote $\ell(\theta, z)$.

Symmetry. The per-sample loss $\ell(\cdot, \cdot)$ is said to possess the Q-symmetry if

$$\ell(\theta, z) = \ell(Q_\rho(\theta), z), \forall \rho \in \mathbb{R}, \tag{2}$$

where $(Q_\rho)_{\rho\in\mathbb{R}}$ is a set of continuous transformation parameterized by $\rho \in \mathbb{R}$. Without loss of generality, we assume $Q_0 = \mathrm{id}$. The most common symmetries exist within the model f, namely f_θ is invariant under certain transformations of θ. However, our formalism is slightly more general in the sense that it is also possible for the model to be variant while the per-sample loss remains unchanged, which appears in self-supervised learning [37], for example.

3 Continuous Symmetry Creates Noise Equilibria

Taking the derivative with respect to ρ at $\rho = 0$ in Eq. (2), we have

$$0 = \nabla_\theta \ell(\theta, z) \cdot J(\theta), \tag{3}$$

where $J(\theta) = \frac{\mathrm{d}Q_\rho(\theta)}{\mathrm{d}\rho}|_{\rho=0}$. Denote by C be the antiderivative of J, that is, $\nabla C(\theta) = J(\theta)$. Then, taking the expectation over z in (3) gives the following conservation law for GD solutions $(\theta_t)_{t\geq 0}$:

$$\dot{C}(\theta_t) = 0. \tag{4}$$

Essentially, this is a consequence of Noether's theorem [22], and C will be called a "Noether charge" in analogy to theoretical physics. The conservation law (4) implies that the GD trajectory is constrained on the manifold $\{\theta : C(\theta) = C(\theta_0)\}$. We refer to Ref. [16] for a study of this type of conservation law under the Bregman Lagrangian [14].

3.1 Noether Flow in Degenerate Directions

In this paper, we are interested in how $C(\theta_t)$ changes, if it changes at all, under SGD. By Ito's lemma, we have the following *Noether flow* (namely, the flow of the Noether charge):

$$\dot{C}(\theta_t) = \sigma^2 \mathrm{Tr}\left[\Sigma(\theta_t)\nabla^2 C(\theta_t)\right], \tag{5}$$

where $\nabla^2 C$ denotes the Hessian matrix of C. The derivation is deferred to Appendix A. By definition, $\Sigma(\theta_t)$ is always positive semidefinite (PSD). Thus, we immediately have: if $\nabla^2 C$ is PSD throughout training, $C(\theta_t)$ is a monotonically increasing function of time. Conversely, if $\nabla^2_\theta C$ is negative semidefinite (NPD), $C(\theta_t)$ is a monotonically decreasing function of time.

The existence of symmetry implies that (with suitable conditions of smoothness) any solution θ resides within a connected, loss-invariant manifold, defined as $\mathcal{M}_\theta := \{Q_\rho(\theta) : \rho \in \mathbb{R}\}$. We term directions within this manifold as "degenerate directions" since movement along them does not change the loss value. Notably, the biased flow (5) suggests that SGD noise can drive SGD to explore within this manifold along these degenerate directions since the value of $C(\theta)$ for $\theta \in \mathcal{M}_\theta$ can vary.

3.2 Exponential symmetries

Now, let us focus on a family of symmetries that is common in deep learning. Since the corresponding conserved quantities are quadratic functions of the model parameters, we will refer to this class of symmetries as *exponential symmetries*.

Definition 3.1. $(Q_\rho)_\rho$ is said to be a exponential symmetry if $J(\theta) := \frac{\mathrm{d}}{\mathrm{d}\rho}Q_\rho(\theta)|_{\rho=0} = A\theta$ for a symmetric matrix A.

This implies when $\rho \ll 1$, $Q_\rho = \mathrm{id} + \rho A + o(\rho)$. In the sequel, we also use the words "A-symmetry" and "Q-symmetry" interchangeably since all properties of Q_ρ we need can be derived from A. This definition applies to the following symmetries that are common in deep learning:

- *Rescaling symmetry*: $Q_\rho(a,b) = (a(\rho+1), b/(\rho+1))$, which appears in linear and ReLU networks [6, 35]. In this symmetry, $A = \mathrm{diag}(I_a, -I_b)$, where I is the identity matrix with dimensions matching that of a and b.
- *Scaling symmetry*: $Q_\rho\theta = (\rho+1)\theta$, which exists whenever part of the model normalized using techniques like batch normalization [12], layer normalization [2], or weight normalization [25]. In this case, $A = I$.
- *Double rotation symmetry:* This symmetry appears when parts of the model involve a matrix factorization problem, where for an arbitrary invertible matrix B $\ell = \ell(UW) = \ell(UBB^{-1}W)$.

It is possible for only a subset of parameters to have a given symmetry. Mathematically, this corresponds to the case when A is low-rank. It is also common for ℓ to have multiple exponential symmetries at once, often for different (but not necessarily disjoint) subsets of parameters. For example, a ReLU network has a different rescaling symmetry for every hidden neuron.

It is obvious that under this Q symmetry, the Noether charge has a simple quadratic form:

$$C(\theta) = \theta^\top A\theta. \tag{6}$$

Moreover, the interplay between this symmetry and weight decay can be explicitly characterized in our framework. To this end, we need the following definition.

Definition 3.2. For any $\gamma \in \mathbb{R}$, we say $\ell_\gamma(\theta,x) := \ell(\theta,x) + \gamma\|\theta\|^2$ has the Q symmetry as long as $\ell(\theta,x)$ has the Q symmetry.

For the SGD dynamics that minimizes $L_\gamma(\theta) = \mathbb{E}_x[\ell_\gamma(\theta,x)]$, it follows from (5) that

$$\dot{C}(\theta_t) = -4\gamma C(\theta_t) + \sigma^2 \mathrm{Tr}[\Sigma(\theta_t)A] =: G(\theta_t). \tag{7}$$

Thus, a positive γ always causes $|C(\theta_t)|$ to decay, and the influence of symmetry is determined by the spectrum of A. Denote by $A = \sum_j \mu_j n_j n_j^\top$ the eigendecomposition of A. Then,

$$\mathrm{Tr}[\Sigma(\theta_t)A] = \sum_{i:\mu_i>0} \mu_i n_i^\top \Sigma(\theta_t) n_i + \sum_{j:\mu_j<0} \mu_j n_j^\top \Sigma(\theta_t) n_j.$$

This gives a very clear interpretation of the interplay between SGD noise and the exponential symmetry: the noise along the positive directions of A causes $C(\theta_t)$ to grow, while the noise along the negative directions causes $C(\theta_t)$ to decay. In other words, the noise-induced dynamics of $C(\theta_t)$ is determined by the competition between the noise along the positive- and negative-eigenvalue directions of A.

Time Scales. The above analysis implies that the dynamics of SGD can be decomposed into two parts: the dynamics that directly reduce loss, and the dynamics along the degenerate direction of the loss, which is governed by Eq (5). These two dynamics have essentially independent time scales. The first part is independent of the σ^2 in expectation, whereas the time scale of the dynamics in the degenerate directions depends linearly on σ^2.

The first time scale t_{erm} is due to the dynamics of empirical risk minimization. The second time scale t_{equi} is the time scale for Eq. (5) to reach equilibrium, which is irrelevant to direct risk minimization. When the parameters are properly tuned, t_{erm} is of order 1, whereas t_{equi} is proportional to $\sigma^2 = \eta/(2S)$. Therefore, when σ^2 is large, the parameters will stay close to the equilibrium point early in the training, and one can expect that $\dot{C}(\theta_t)$ is approximately zero after t_{equi}. In line with Ref. [18], this can be called the fast-equilibrium phase of learning. Likewise, when $\sigma^2 \ll 1$, the approach to equilibrium will be slower than the actual time scale of risk minimization, and the dynamics in the degenerate direction only take off when the model has reached a local minimum. This can be called the slow-equilibrium phase of learning.

3.3 Noise Equilibrium and Fixed Point Theorem

It is important and practically relevant to study the stationary points of dynamics in Eq. (7). Formally, the stationary point is reached when $-\gamma C(\theta)+\eta \mathrm{Tr}[\Sigma(\theta)A] = 0$. Because we make essentially no assumption about $\ell(\theta)$ and $\Sigma(\theta)$, one might feel that it is impossible to guarantee the existence of a fixed point. Remarkably, we prove below that a fixed point exists and is unique for every connected degenerate manifold.

To start, consider the exponential maps generated by A:

$$e^{\lambda A}\theta := \lim_{\rho\to 0}(I + \rho A + o(\rho))^{\lambda/\rho}\theta,$$

which applies the symmetry transformation to θ for λ/ρ times. Then, it follows that if we apply Q_ρ transformation to θ infinitely many times and for a perturbatively small ρ,

$$\ell(\theta) = \ell\left(e^{\lambda A}\theta\right). \tag{8}$$

Thus, the exponential symmetry implies the symmetry with respect to an exponential map, a fundamental element of Lie groups [9]. Note that exponential-map symmetry is also an exponential symmetry by definition. For the exponential map, the degenerate direction is clear: for any λ, θ connects to $e^{\lambda A}\theta$ without any loss function barrier. Therefore, the degenerate direction for any exponential symmetry is unbounded. Now, we prove the following fixed point theorem, which shows that for every exponential symmetry and every θ, there is one and only one corresponding fixed point in the degenerate direction.

Theorem 3.3. *Let the per-sample loss satisfy the A exponential symmetry and $\theta_\lambda := \exp[\lambda A]\theta$. Then, for any θ and any $\gamma \geq 0$,*[2]

(1) $G(\theta_\lambda)$ (Eq. (7)) and $-C(\theta_\lambda)$ are monotonically decreasing functions of λ;
(2) there exists a $\lambda^* \in \mathbb{R}\cup\{\pm\infty\}$ such that $G(\theta_{\lambda^*}) = 0$;
(3) in addition, if $G(\theta_\lambda) \neq 0$, λ^* is unique and $G(\theta_\lambda)$ is strictly monotonic;
(4) in addition to (3), if $\Sigma(\theta)$ is differentiable, $\lambda^*(\theta)$ is a differentiable function of θ.

Part (1), together with Part (2), implies that the unique stationary point is essentially attractive. This is because $\dot{C}$ decreases with λ while C increases with it. Let $C^* = C(\theta_{\lambda^*})$. Thus, $C(\theta) - C^*$ always have the opposite sign of λ^*, while $\frac{d}{dt}C(\theta)$ will have the same sign. Conceptually, this means that C will always move to reduce its distance to C^*. Assuming that C^* is a constant in time (or close to a constant, which is often the case at the end of training), Part (1) implies that $\frac{d}{dt}(C(\theta) - C^*) \propto -\mathrm{sgn}(C(\theta) - C^*)$, signaling a convergence to $C(\theta) = C^*$. In other words, SGD will move to restore the balance if it is perturbed away from $\lambda^* = 0$. If the matrix ΣA is well-behaved, one can indeed establish the convergence to the fixed point in the relative distance even if C^* is mildly divergent due to diffusion. Because this part is strongly technical and our focus is on the fixed points, we leave the formal statement and its discussion to Appendix A.3.

Theorem 3.4. *(Informal) Let C^* follow a drifted Brownian motion and ΣA satisfy two well-behaved conditions. Then, either $C - C^* \to 0$ in L_2 or $(C - C^*)^2/(C^*)^2 \to 0$ in probability.*

Parts (2) and (3) show that a unique fixed point exists. We note that it is more common than not for the conditions of uniqueness to hold because there is generally no reason for $\mathrm{Tr}[\Sigma(\theta)A]$ or $\mathrm{Tr}[\theta\theta^\top A]$ to vanish simultaneously, except in some very restrictive subspaces. One major (perhaps the only) reason for the first trace to vanish is when the model is located at an interpolation minimum. However, interpolation minima are irrelevant for modern large-scale problems such as large language models because the amount of available text for training far exceeds the size of the largest models. Even when the interpolation minimum exists, the unique fixed point should still exist when the training is not complete. Part (4) means that the fixed points of the dynamics is well-behaved. If the parameter θ has a small fluctuation around a given location, C will also have a small fluctuation around the fixed point solution. This justifies approximating C by a constant value when θ changes slowly and with small fluctuation.

Fixed point as a Noise Equilibrium. Let θ^* be a fixed point of (7). It must satisfy

$$C(\theta^*) = \frac{\sigma^2}{4\gamma}\mathrm{Tr}\left[\Sigma(\theta^*)A\right]. \tag{9}$$

[2] While our main result is stated in terms of the continuous-time dynamics, we note that a qualitatively similar result can be proved for the discrete-time SGD. See Section E.

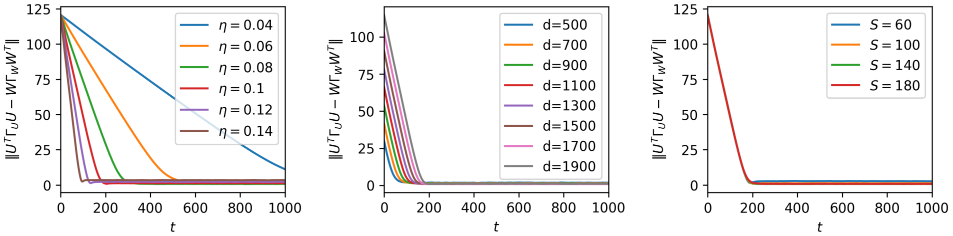

Figure 1: The convergence of matrix factorization to the noise equilibria is robust against different hyperparameter settings. The task is an autoencoding task where $y = x \in \mathbb{R}^{100}$. The distribution of x is controlled by a parameter ϕ_x: $x_{1:50} \sim \mathcal{N}(0, \phi_x)$, $x_{51:100} \sim \mathcal{N}(0, 2 - \phi_x)$. This directly controls the overall covariance of x. The output noise covariance is set to be identity. Unless it is the independent variable, η, S and d are set to be 0.1, 100 and 2000, respectively. **Left**: using different learning rates. **Mid**: different data dimension: $d_x = d_y = d$. **Right**: different batch size S.

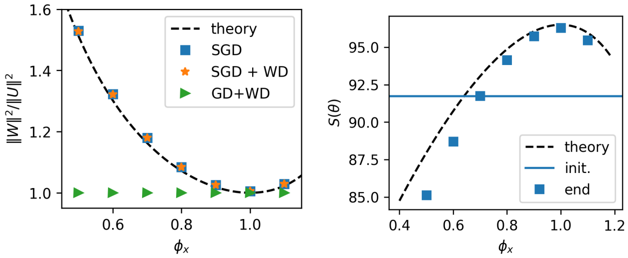

Figure 2: A two-layer linear network after training. Here, the problem setting is the same as Figure 1. The theoretical prediction is computed from Theorem 4.1. **Left** balance of the norm is only achieved when $\phi_x = 1$, namely, when the data has an isotropic covariance. We also test SGD with a small weight decay (10^{-4}), which is sufficiently small that the solution we obtained for SGD without SGD still holds approximately. In contrast, training with GD + GD always converges to a norm-balanced solution. **Right**: the sharpness of the converged model trained with SGD. We see that for some data distributions, SGD converges to a sharper solution, whereas it converges to flatter solutions for other data distributions. This flattening and sharpening effect are both due to the noise-balance effect of SGD. Here, we find that the systematic error between experiment and theory is due to the use of a finite learning rate and decreases as we decrease η.

Hence, a large weight decay leads to a small $|C(\theta^*)|$, whereas a large gradient noise leads to a large $|C(\theta^*)|$. When there is no weight decay, we get a different equilibrium condition: $\mathrm{Tr}[\Sigma(\theta^*)A] = 0$, which can be finite only when A contains both positive and negative eigenvalues. This equilibrium condition is equivalent to

$$\sum_{i:\mu_i>0} \mu_i n_i^\top \Sigma(\theta^*) n_i = -\sum_{j:\mu_j<0} \mu_j n_j^\top \Sigma(\theta^*) n_j. \tag{10}$$

Namely, the overall gradient fluctuation in the two different subspaces specified by the symmetry A must balance. We will see that the main implication of this result is that the gradient noise between different layers of a deep neural network should be balanced at the end of training. Conceptually, Theorem 3.3 suggests the existence of a special type of fixed point for SGD, which the following definition formalizes.

Definition 3.5. θ is a *noise equilibrium* for a nonconstant function $C(\theta)$ if $\dot{C}(\theta) = 0$ under SGD.

4 Application: Balance, Alignment, and Stability of Matrix Factorization

As an exactly solvable example, let us consider a two-layer linear network (this can also be seen as a variant of standard matrix factorizations):

$$\ell_\gamma = \|UWx - y\|^2 + \gamma(\|U\|_F^2 + \|W\|_F^2). \tag{11}$$

where $x \in \mathbb{R}^{d_x}$ is the input data, and $y = y' + \epsilon \in \mathbb{R}^{d_y}$ is a noisy version of the label. The ground truth mapping is linear and realizable: $y' = U^* W^* x$. The second moments of the input and noise are denoted as $\Sigma_x = \mathbb{E}[xx^\top]$ and $\Sigma_\epsilon = \mathbb{E}[\epsilon\epsilon^\top]$, respectively. Note that this problem is essentially identical to a matrix factorization problem, which is not only a theoretical model of neural networks but also an important algorithm frequently in use for recommender systems [32]. The following theorem gives the fixed point of Noether flow.

Theorem 4.1. *Let $r = UWx - y$ be the prediction residual. For all symmetric B, $\dot{C}_B = 0$ if*

$$W\Gamma_W W^\top = U^\top \Gamma_U U, \tag{12}$$

where $\Gamma_W = \mathbb{E}[\|r\|^2 xx^\top] + 2\gamma I$, $\Gamma_U = \mathbb{E}[\|x\|^2 rr^\top] + 2\gamma I$.

The equilibrium condition takes a more suggestive form when the model is at the global minimum, where $U^* W^* x - y = \epsilon$. When ϵ and x are independent and $\gamma = 0$, we have:

$$W \bar{\Sigma}_x W^\top = U^\top \bar{\Sigma}_\epsilon U \tag{13}$$

Here, the bar over the matrices indicates that they have been normalized by their traces: $\bar{\Sigma} = \Sigma/\mathrm{Tr}[\Sigma]$. The matrices Γ_W and Γ_U simplifies because at the global minimum, $r_i = \epsilon_i$ and so $\mathbb{E}[\|x\|^2 rr^\top] = \mathrm{Tr}[\Sigma_x]\Sigma_\epsilon$ and $\mathbb{E}[\|r\|^2 xx^\top] = \mathrm{Tr}[\Sigma_\epsilon]\Sigma_x$. See Figure 1 for the convergence of SGD to this solution under different settings. The theory can also be leveraged to find the exact solution of a deep linear network, which we discuss in Appendix C.

Noise Driven Progressive Sharpening and Flattening. This result implies a previously unknown mechanism of progressive sharpening and flattening, where, during training, the stability of the algorithm steadily improves (during flattening) or deteriorates (during sharpening) [31, 13, 5]. To see this, we first derive a metric of sharpness for this model.

Proposition 4.2. *For the per-sample loss* (11)*, let* $S(\theta) := \mathrm{Tr}[\nabla^2 L(\theta)]$*. Then,* $S(\theta) = d_y \|W \Sigma_x^{1/2}\|_F^2 + \|U\|_F^2 \mathrm{Tr}[\Sigma_x]$.

The trace of the Hessian is a good metric of the local stability of the GD and SGD algorithm because the trace upper bounds the largest Hessian eigenvalue. The theory directly implies that SGD training can both make the model move towards sharper solutions and flatter solutions – an effect that is completely dependent on the data and initialization. This challenges the common belief that SGD seeks a flatter minimum during training. See Figure 2.

Let us analyze the simplest case of an autoencoding task, where the model is at the global minimum. Here, $\Sigma_x \propto I_{d_x}$, $\Sigma_\epsilon \propto I_{d_y}$. For a random Gaussian initialization with variance σ_W^2 and σ_U^2, the trace at initialization is, in expectation, $S_{\text{init}} = d_y d \mathrm{Tr}[\Sigma_x](\sigma_W^2 + \sigma_U^2)$. At the end of the training, the model is close to the global minimum and satisfies Proposition 4.2. Here, the rank of U and W matters and is upper bounded by $\min(d, d_x)$, and at the global minimum, U and W are full-rank (equal to $\min(d, d_x)$), and all the singular values are 1. Thus,

$$\begin{cases} S_{\text{init}} = d_x d(\sigma_U^2 + \sigma_W^2)\mathrm{Tr}[\Sigma_x], \\ S_{\text{end}} = 2\min(d, d_x)\mathrm{Tr}[\Sigma_x]. \end{cases} \tag{14}$$

The change in the sharpness during training thus depends crucially on the initialization scheme. For Xavier init, $\sigma_U^2 = (d_y + d)^{-1}$ and $\sigma_W^2 = (d + d_x)^{-1}$, and so $S_{\text{init}} \approx S_{\text{end}}$ (but S_{init} is slightly smaller). Thus, for the Xavier init., the sharpness of loss experiences a small sharpening during training. For Kaiming init., $\sigma_U^2 = 1$ and $\sigma_W^2 = d_x^{-1}$. Therefore, it always holds that $S_{\text{init}} \geq S_{\text{end}}$, and so the stability improves as the training proceeds. The only case when the Kaiming init. does not experience progressive flattening is when $d = d_x = d_y$, which agrees with the common observation that training is easier if the widths of the model are balanced [10]. In previous works, the progressive sharpening happens when the model is trained with GD [5]; our theory suggests an alternative mechanism for it.

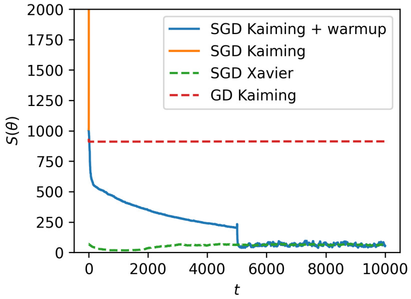

Figure 3: Dynamics of the stability condition S during the training of a rank-1 matrix factorization problem. The solid lines show the training of SGD with Kaiming init. When the learning rate ($\eta = 0.008$) is too large, SGD diverges (orange line). However, when one starts training at a small learning rate (0.001) and increases η to 0.008 after 5000 iterations, the training remains stable. This is because SGD training improves the stability condition during training, which is in agreement with the theory. In contrast, the stability condition of GD and that of SGD with a Xavier init increases only slightly. Also, note that both Xavier and Kaiming init. under SGD converges to the same stability condition because the equilibrium is unique.

A practical technique that the theory explains is using warmup to stabilize training in the early stage. This technique was first proposed in Ref. [8] for training CNNs, where it was observed that the training is divergent if we start the training at a fixed large learning rate $\eta_{\max}$. However, this divergent behavior disappears if we perform a warmup training, where the learning rate is increased gradually from a minimal value to $\eta_{\max}$. Later, the same technique is found to be crucially useful for training large language models [24]. Our theory shows that the gradient noise can drive Kaiming init. to a stabler status where a larger learning can be applied. See Figure 3.

References

[1] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[3] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

[4] Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, pages 1756–1760. PMLR, 2015.

[5] 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.

[6] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp Minima Can Generalize For Deep Nets. *ArXiv e-prints*, March 2017.

[7] 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.

[8] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[9] Brian C Hall and Brian C Hall. *Lie groups, Lie algebras, and representations*. Springer, 2013.

[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] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[13] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2019.

[14] Michael I Jordan. Dynamical, symplectic and stochastic perspectives on gradient-based optimization. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 523–549. World Scientific, 2018.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

[22] Emmy Noether. Invariante variationsprobleme. *Königlich Gesellschaft der Wissenschaften Göttingen Nachrichten Mathematik-physik Klasse*, 2:235–267, 1918.

[23] 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.

[24] Martin Popel and Ondřej Bojar. Training tips for the transformer model. *arXiv preprint arXiv:1804.00247*, 2018.

[25] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[26] 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.

[27] 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.

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

[29] Ming Chen Wang and George Eugene Uhlenbeck. On the theory of the brownian motion ii. *Reviews of modern physics*, 17(2-3):323, 1945.

[30] Zihao Wang and Liu Ziyin. Posterior collapse of a linear latent variable model. *Advances in Neural Information Processing Systems*, 35:37537–37548, 2022.

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

[32] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pages 3203–3209. Melbourne, Australia, 2017.

[33] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

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

[35] Liu Ziyin, Hongchao Li, and Masahito Ueda. Law of balance and stationary distribution of stochastic gradient descent. *arXiv preprint arXiv:2308.06671*, 2023.

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

[37] Liu Ziyin, Ekdeep Singh Lubana, Masahito Ueda, and Hidenori Tanaka. What shapes the loss landscape of self supervised learning? In *The Eleventh International Conference on Learning Representations*, 2023.

A Proofs

A.1 Ito's Lemma and Derivation of Eq. (5)

Let a vector X_t follow the following stochastic process:

$$\mathrm{d}X_t = \mu_t\,\mathrm{d}t + G_t\,\mathrm{d}W_t \tag{15}$$

for a matrix G_t. Then, the dynamics of any function of X_t can be written as (Ito's Lemma)

$$\mathrm{d}f(X_t) = \left(\nabla_X^\top f\mu_t + \frac{1}{2}\mathrm{Tr}[G_t^\top\nabla^2 f(X_t)G_t]\right)\mathrm{d}t + \nabla f(X_t)^\top G_t\,\mathrm{d}W_t. \tag{16}$$

Applying this result to quantity $C(\theta)$ under the SGD dynamics, we obtain that

$$dC = \left(\nabla^\top C\nabla L + \frac{\sigma^2}{2}\mathrm{Tr}[\Sigma(\theta)\nabla^2 C]\right)dt + \nabla^\top C\sqrt{\sigma^2\Sigma(\theta)}dW_t, \tag{17}$$

where we have used $\mu_t = \nabla L$, $G_t = \sqrt{\sigma^2\Sigma(\theta)}$. By Eq. (3), we have that

$$\nabla^\top C\nabla L = \mathbb{E}[\nabla^\top C\nabla\ell] = 0, \tag{18}$$

and

$$\nabla^\top C\Sigma = \mathbb{E}[\nabla^\top C\nabla\ell\nabla^\top\ell] - \mathbb{E}[\nabla^\top C\nabla\ell]\mathbb{E}[\nabla^\top\ell] = 0. \tag{19}$$

Because $\Sigma(\theta)$ and $\sqrt{\Sigma(\theta)}$ share eigenvectors, we have that

$$\nabla^\top C\sqrt{\sigma^2\Sigma(\theta)} = 0. \tag{20}$$

Therefore, we have derived:

$$dC = \frac{\sigma^2}{2}\mathrm{Tr}[\Sigma(\theta)\nabla^2 C]dt. \tag{21}$$

A.2 Proof of Theorem 3.3

We first prove a lemma that links the gradient covariance at θ to the gradient covariance at θ_λ.

Lemma A.1.

$$\mathrm{Tr}[\Sigma(\theta_\lambda)A] = \mathrm{Tr}[e^{-2\lambda A}\Sigma(\theta)A]. \tag{22}$$

Proof. By the definition of the quadrative symmetry, we have that for an arbitrary λ,

$$\ell(\theta) = \ell(e^{\lambda A}\theta). \tag{23}$$

Taking the derivative of both sides, we obtain that

$$\nabla_\theta\ell(\theta) = e^{\lambda A}\nabla_{\theta_\lambda}\ell(\theta_\lambda), \tag{24}$$

The standard result of Lie groups shows that $e^{\lambda A}$ is full-rank and symmetric, and its inverse is $e^{-\lambda A}$. Therefore, we have

$$e^{-\lambda A}\nabla_\theta\ell(\theta) = \nabla_{\theta_\lambda}\ell(\theta_\lambda). \tag{25}$$

Now, we apply this relation to the trace of interest. By definition,

$$\Sigma(\theta_\lambda) = \mathbb{E}[\nabla_{\theta_\lambda}\ell(\theta_\lambda)\nabla_{\theta_\lambda}^\top\ell(\theta_\lambda)] \tag{26}$$

$$= e^{-\lambda A}\Sigma(\theta)e^{-\lambda A}. \tag{27}$$

Because $e^{\lambda A}$ is a function of A, it commutes with A. Therefore,

$$\mathrm{Tr}[\Sigma(\theta_\lambda)A] = \mathrm{Tr}[e^{-\lambda A}\Sigma(\theta)e^{-\lambda A}A] \tag{28}$$

$$= \mathrm{Tr}[e^{-2\lambda A}\Sigma(\theta)A]. \tag{29}$$

□

Now, we are ready to prove the main theorem.

Proof. First of all, it is easy to see that $C(\theta_\lambda)$ is a monotonically increasing function of λ. By definition,

$$C(\theta_\lambda) = \theta^T e^{\lambda A} A e^{\lambda A}\theta \tag{30}$$

$$= \theta^T (Z_+ + Z_-)\theta, \tag{31}$$

where we have decomposed the matrix $e^{\lambda A} A e^{\lambda A} = Z_+ + Z_-$ into two symmetric matrices such that Z_+ only contains nonnegative eigenvalues, and Z_- only contains nonpositive eigevalues. Because $e^{\lambda A}$ commute with A, they share the eigenvectors. Using elementary Lie algebra shows that the eigenvalues of Z_+ are $a_+ e^{\lambda a_+}$ and that of Z_- are $a_- e^{\lambda a_-}$, where $a_+ \geq 0$ and $a_0 \leq 0$. This implies that $\theta^T Z_+ \theta$ and $\theta^T Z_- \theta$ are monotonically increasing functions of λ.

Now, by Lemma A.1, we have

$$\mathrm{Tr}[\Sigma(\theta_\lambda)A] = \mathrm{Tr}[e^{-2\lambda A}\Sigma(\theta)A]. \tag{32}$$

Similarly, the regularization term is

$$\gamma\theta_\lambda^\top A\theta_\lambda = \gamma \mathrm{Tr}[\theta^\top \theta A e^{2\lambda A}]. \tag{33}$$

Now, by assumption, if $G(\theta_\lambda) \neq 0$, we have either $\mathrm{Tr}[\Sigma(\theta)A] \neq 0$ or $\mathrm{Tr}[\theta\theta^\top A] \neq 0$.

If $\mathrm{Tr}[\Sigma(\theta)A] = \theta^\top A\theta = 0$, we have already proved item (2) of the theorem. Therefore, let us consider the case when either (or both) $\mathrm{Tr}[\Sigma(\theta)A] \neq 0$ or $\theta^\top A\theta \neq 0$

Without loss of generality, we assume $\gamma \geq 0$, and the case of $\gamma < 0$ follows an analogous proof. In such a case, we can write the trace in terms of the eigenvectors n_i of A:

$$-\gamma\theta_\lambda^\top A\theta_\lambda + \eta\mathrm{Tr}[\Sigma(\theta_\lambda)A] = \underbrace{\eta\sum_{\mu_i>0} e^{-2\lambda|\mu_i|}|\mu_i|\sigma_i^2 + \gamma\sum_{\mu_i<0} e^{-2\lambda|\mu_i|}|\mu_i|\tilde\theta_i^2}_{I_1(\lambda)} - \underbrace{\left(\eta\sum_{\mu_i<0} e^{2\lambda|\mu_i|}|\mu_i|\sigma_i^2 + \gamma\sum_{\mu_i>0} e^{2\lambda|\mu_i|}|\mu_i|\tilde\theta_i^2\right)}_{I_2(\lambda)}$$
$$=: I(\lambda),$$

where μ_i is the i-th eigenvalue of A, $\tilde\theta_i = (n_i^\top \theta_i)^2$, $\sigma_i^2 = n_i^\top \Sigma(\theta) n_i \geq 0$ is the norm of the projection of Σ in this direction.

By definition, I_1 is either a zero function or strictly monotonically increasing function with $I_1(-\infty) = +\infty, I_1(+\infty) = 0$ Likewise, I_2 is either a zero function or a strictly monotonically increasing function with $I_2(-\infty) = 0, I_2(+\infty) = +\infty$. By the assumption $\mathrm{Tr}(\Sigma(\theta)A) \neq 0$ or $\mathrm{Tr}(\theta\theta^\top A) \neq 0$, we have that at least one of I_1 and I_2 must be a strictly monotonic function.

- If I_1 or I_2 is zero, we can take λ to be either $+\infty$ or $-\infty$ to satisfy the condition.
- If both I_1 and I_2 are nonzero, then $I = I_1 - I_2$ is a strictly monotonically decreasing function with $I(-\infty) = +\infty$ and $I(+\infty) = -\infty$. Therefore, there must exist only a unique $\lambda^* \in \mathbb{R}$ such that $I(\lambda^*) = 0$.

For the proof of (4), we denote the multi-variable function $J(\theta;\lambda) \coloneqq G(\theta_\lambda)$. Given that $\Sigma(\theta)$ is differentiable, $\frac{\partial J}{\partial \theta}$ exists.

It is easy to see that $\frac{\partial J}{\partial \lambda}$ is continuous. Moreover, for any θ and $\lambda = \lambda^*(\theta)$,

$$-\frac{\partial J}{2\partial\lambda} = \eta\sum_{\mu_i>0} e^{-2\lambda|\mu_i|}|\mu_i|^2\sigma_i^2 + \gamma\sum_{\mu_i<0} e^{-2\lambda|\mu_i|}|\mu_i|^2\tilde\theta_i^2 + \eta\sum_{\mu_i<0} e^{2\lambda|\mu_i|}|\mu_i|^2\sigma_i^2 + \gamma\sum_{\mu_i>0} e^{2\lambda|\mu_i|}|\mu_i|^2\tilde\theta_i^2 \neq 0.$$

Consequently, according to the Implicit Function Theorem, the function $\lambda^*(\theta)$ is differentiable. Additionally, $\frac{\partial\lambda}{\partial\theta} = -\frac{\frac{\partial J}{\partial\theta}}{\frac{\partial J}{\partial\lambda}}$.

□

A.3 Convergence

First of all, notice an important property, which follows from Theorem 3.3: $\lambda^* = 0$ if and only if $C - C^* = 0$.

Lemma A.2. *For all* $\theta(t)$*,*

$$\frac{\dot{C}(\theta)}{\lambda^*(\theta)} \geq \begin{cases} 2\sigma^2 \mathrm{Tr}[\Sigma(\theta^*)A_+^2] & \textit{if } \lambda^* > 0; \\ 2\sigma^2 \mathrm{Tr}[\Sigma(\theta^*)A_-^2] & \textit{if } \lambda^* < 0. \end{cases} \tag{34}$$

Proof. As in the main text, let θ^* denote θ_{λ^*}, $C = C(\theta)$ and $C^* = C(\theta^*)$. Thus,

$$\frac{dC}{dt} = \sigma^2 \mathrm{Tr}[\Sigma(\theta)A] \tag{35}$$

$$= \sigma^2 \mathrm{Tr}[\Sigma(\theta^*)e^{2\lambda^* A}A], \tag{36}$$

where the second equality follows from Lemma A.1. One can decompose A as a sum of two symmetric matrices

$$A = \underbrace{Q\Sigma_+Q^\top}_{:=A_+} + \underbrace{Q\Sigma_-Q^\top}_{:=A_-}, \tag{37}$$

where Q is an orthogonal matrix, Σ_+ (Σ_-) is diagonal and contains only non-negative (non-positive) entries. Note that by the definition of λ^*, we have $\mathrm{Tr}[\Sigma(\theta^*)A] = 0$ and, thus,

$$\mathrm{Tr}[\Sigma(\theta^*)A_+] = -\mathrm{Tr}[\Sigma(\theta^*)A_-]. \tag{38}$$

Thus,

$$\mathrm{Tr}[\Sigma(\theta)A] = \mathrm{Tr}[\Sigma(\theta^*)e^{2\lambda^* A}A] \tag{39}$$

$$= \mathrm{Tr}[\Sigma(\theta^*)(e^{2\lambda^* A} - I)A] \tag{40}$$

$$= \mathrm{Tr}[\Sigma(\theta^*)(e^{2\lambda^* A_+} - I)A_+] + \mathrm{Tr}[\Sigma(\theta^*)(e^{2\lambda^* A_-} - I)A_-]. \tag{41}$$

Using the inequality $I + A \leq e^A$ (namely, that $e^A - I - A$ is PSD), we obtain a lower bound

$$\mathrm{Tr}[\Sigma(\theta)A] \geq 2\lambda^* \mathrm{Tr}[\Sigma(\theta^*)A_+^2] - \mathrm{Tr}[\Sigma(\theta^*)(I - e^{2\lambda^* A_-})A_-] \tag{42}$$

If $\lambda^* > 0$, $\mathrm{Tr}[\Sigma(\theta^*)(e^{2\lambda^* A_-} - I)A_-] < 0$,

$$\mathrm{Tr}[\Sigma(\theta)A] \geq 2\lambda^* \mathrm{Tr}[\Sigma(\theta^*)A_+^2]. \tag{43}$$

Likewise, there is an upper bound, which simplifies to the following form if $\lambda^* < 0$:

$$\mathrm{Tr}[\Sigma(\theta)A] \leq 2\lambda^* \mathrm{Tr}[\Sigma(\theta^*)A_-^2]. \tag{44}$$

This finishes the proof. □

Lemma A.3. *For any* θ*,*

$$-\frac{C - C^*}{\lambda^*} \leq \begin{cases} 2(\theta^*)^\top A_+^2 \theta^* & \textit{if } \lambda^* > 0; \\ 2(\theta^*)^\top A_-^2 \theta^* & \textit{if } \lambda^* < 0. \end{cases} \tag{45}$$

Proof. The proof is conceptually similar to the previous one. By definition, we have

$$C - C^* = (\theta^*)^\top e^{-2\lambda^* A} A \theta^* - (\theta^*)^\top A \theta^* \tag{46}$$

$$= (\theta^*)^\top A (e^{-2\lambda^* A} - I)\theta^* \tag{47}$$

$$= (\theta^*)^\top A_+ (e^{-2\lambda^* A_+} - I)\theta^* + (\theta^*)^\top A_- (e^{-2\lambda^* A_-} - I)\theta^*. \tag{48}$$

By the inequality $I + A \leq e^A$, we have an upper bound

$$C - C^* \geq -2\lambda^* (\theta^*)^\top A_+^2 \theta^* + (\theta^*)^\top A_- (e^{-2\lambda^* A_-} - I)\theta^*. \tag{49}$$

If $\lambda^* > 0$, $(\theta^*)^\top A_- (e^{-2\lambda^* A_-} - I)\theta^* \geq 0$,

$$C - C^* \geq -2\lambda^* (\theta^*)^\top A_+^2 \theta^*. \tag{50}$$

Likewise, if $\lambda^* < 0$, one can prove a lower bound:

$$C - C^* \leq -2\lambda^* (\theta^*)^\top A_-^2 \theta^*. \tag{51}$$

□

Combining the above two lemmas, one can prove the following corollary.

Corollary A.4.

$$\frac{\dot{C}}{C - C^*} \leq \begin{cases} -\sigma^2 \frac{\mathrm{Tr}[\Sigma(\theta^*)A_+^2]}{(\theta^*)^\top A_+^2 \theta^*}, & \text{if } C - C^* < 0; \\ -\sigma^2 \frac{\mathrm{Tr}[\Sigma(\theta^*)A_-^2]}{(\theta^*)^\top A_-^2 \theta^*}, & \text{if } C - C^* > 0. \end{cases} \tag{52}$$

Now, one can prove that as long as C^* is not moving too fast, C converges to C^* in mean square.

Lemma A.5. *Let the dynamics of C^* be a drifted Brownian motion: $dC^* = \mu dt + sdW$, where W is a Brownian motion with variance s^2. If there exists $c_0 > 0$ such that $\frac{\mathrm{Tr}[\Sigma(\theta^*)A_+^2]}{(\theta^*)^\top A_+^2 \theta^*} \geq c_0$ and $\frac{\mathrm{Tr}[\Sigma(\theta^*)A_-^2]}{(\theta^*)^\top A_-^2 \theta^*} > c_0$,*

$$\mathbb{E}(C - C^*)^2 \leq \frac{2\mu^2 + s^2}{2\sigma^4 c_0^2} = O(1). \tag{53}$$

Proof. By assumption,

$$\frac{\dot{C}}{C - C^*} \leq -\sigma^2 c_0. \tag{54}$$

Let us first focus on the case when $C - C^* > 0$. By the definition of C^* and Ito's lemma,

$$d(C - C^*) \leq -\sigma^2 c_0 (C - C^*)dt - \mu dt - sdW. \tag{55}$$

Let $Z = e^{\sigma^2 c_0 t}(C - C^*)$, we obtain that

$$dZ = e^{\sigma^2 c_0 t} d(C - C^*) + \sigma^2 c_0 e^{\sigma^2 c_0 t}(C - C^*)dt \tag{56}$$

$$\leq -\mu e^{\sigma^2 c_0 t} dt - s e^{\sigma^2 c_0 t} dW. \tag{57}$$

Its solution is given by

$$Z \leq -\frac{\mu e^{\sigma^2 c_0 t}}{\sigma^2 c_0} - s \int e^{\sigma^2 c_0 t} dW. \tag{58}$$

Alternatively, if $C - C^* < 0$, we let $Z = e^{\sigma^2 c_0 t}(C^* - C)$, and obtain

$$Z \leq \frac{\mu e^{\sigma^2 c_0 t}}{\sigma^2 c_0} + s \int e^{\sigma^2 c_0 t} dW. \tag{59}$$

Thus,

$$\mathbb{E}[Z^2] \leq \frac{\mu^2 e^{2\sigma^2 c_0 t}}{\sigma^4 c_0^2} + s^2 \int e^{2\sigma^2 c_0 t} dt \tag{60}$$

$$= \frac{\mu^2 e^{2\sigma^2 c_0 t}}{\sigma^4 c_0^2} + \frac{s^2 e^{2\sigma^2 c_0 t}}{2\sigma^2 c_0}. \tag{61}$$

where we have used Ito's isometry in the first line. By construction,

$$\mathbb{E}[(C - C^*)^2] \leq \frac{2\mu^2 + s^2}{2\sigma^4 c_0^2}. \tag{62}$$

The proof is complete. □

Let $\to_p$ denote convergence in probability. One can now prove the following theorem, the convergence of the relative distance to zero in probability.

Theorem A.6. *Let the assumptions be the same as Lemma. (A.5). Then, if $s = \mu = 0$, $\mathbb{E}[(C - C^*)^2] \to 0$. Otherwise,*

$$\frac{(C - C^*)^2}{(C^*)^2} \to_p 0. \tag{63}$$

Proof. By Lemma A.5 and Markov's inequality:

$$\Pr(|C(t) - C^*(t)| > t^{1/4}) \to 0. \tag{64}$$

Now, consider the distribution of $(C^*)^2$. Because C^* is a Gaussian variable with mean μt and variance $s^2 t$, we have that

$$\Pr(|C^*| > \sqrt{t}) \to 1. \tag{65}$$

Now,

$$\begin{aligned}\Pr(|C(t) - C^*(t)|/|C^*| > t^{-1/4}) &\geq \Pr(|C(t) - C^*(t)| > t^{1/4} \& |C^*| < \sqrt{t}) &(66)\\ &\geq \max\left(0, \Pr(|C(t) - C^*(t)| > t^{1/4}) + \Pr(|C^*| < \sqrt{t}) - 1\right) &(67)\\ &\to 0, &(68)\end{aligned}$$

where we have used the Frechet inequality in the second line. This finishes the proof. □

A.4 Proofs of Proposition 4.2

Proof. The loss function is

$$\ell = \|UWx - y\|^2 + \gamma(\|U\|_F^2 + \|W\|_F^2).$$

Let us adopt the following notation: $U = (\tilde{u}_1, \cdots, \tilde{u}_{d_y})^\top \in \mathbb{R}^{d_y \times d}$, $W = (\tilde{w}_1, \cdots, \tilde{w}_{d_x}) \in \mathbb{R}^{d \times d_x}$, where $\tilde{u}_i, \tilde{w}_i \in \mathbb{R}^d$. $\theta = \text{vec}(U, W) = (\tilde{u}_1^\top, \cdots, \tilde{u}_{d_y}^\top, \tilde{w}_1^\top, \cdots, \tilde{w}_{d_x}^\top)^\top \in \mathbb{R}^{(d_x + d_y)d}$.

Due to

$$\nabla_{\tilde{u}_i} \ell = Wx(\tilde{u}_i^\top Wx - y_i) + 2\gamma \tilde{u}_i, \quad \forall i \in [d_y];$$

$$\nabla_{\tilde{w}_j} \ell = \sum_{i=1}^{d_y} \tilde{u}_i x_j \left(\tilde{u}_i^\top Wx - y_i\right) + 2\gamma \tilde{w}_j, \quad \forall j \in [d_x];$$

the diagonal blocks of the Hessian $\nabla_{\theta\theta}^2 \ell$ have the following form:

$$\nabla^2_{\tilde{u}_i, \tilde{u}_i} \ell = Wxx^\top W^\top + 2\gamma I, \quad \forall i \in [d_y];$$

$$\nabla^2_{\tilde{w}_j, \tilde{w}_j} \ell = x_j^2 \sum_{i=1}^{d_y} \tilde{u}_i \tilde{u}_i^\top + 2\gamma I, \quad \forall j \in [d_x].$$

The trace of the Hessian is a good metric of the local stability of the GD and SGD algorithm because the trace upper bounds the largest Hessian eigenvalue. For this loss function, the trace of the Hessian of the empirical risk is

$$\begin{aligned}&S(U, W) := \text{Tr}[\nabla^2_{\theta\theta} \ell - 2\gamma I]\\ =&\sum_{i=1}^{d_y} \text{Tr}[Wxx^\top W^\top] + \sum_{j=1}^{d_x} \text{Tr}[x_j^2 \sum_{i=1}^{d_y} \tilde{u}_i \tilde{u}_i^\top]\\ =&d_y \text{Tr}[W\Sigma_x W^\top] + \|U\|_F^2 \text{Tr}[\Sigma_x] = d_y \|W\Sigma_x^{1/2}\|_F^2 + \|U\|_F^2 \text{Tr}[\Sigma_x],\end{aligned}$$

where $\Sigma_x = xx^\top$. □

B Proof of Theorem 4.1

Proof. First, we split U and W like $U = (u_1, \cdots, u_d) \in \mathbb{R}^{d_y \times d}$ and $W = (w_1^\top, \cdots, w_d^\top)^\top \in \mathbb{R}^{d \times d_x}$. The quantity under consideration is $C_B = \text{Tr}[UBU^\top] - \text{Tr}[W^\top BW]$ for an arbitrary symmetric matrix B. What will be relevant to us is the type of B that is indexed by two indices k and l such that

$$\begin{cases} B_{ij}^{(k,l)} = B_{ji}^{(k,l)} = 1 & \text{if } i = k \text{ and } j = l \text{ or } i = l \text{ and } j = k;\\ B_{ij}^{(k,l)} = 0 & \text{otherwise.}\end{cases} \tag{69}$$

Specifically, for $k, l \in [d]$, we select $B^{(k,l)}_{i,j} = \delta_{i,k}\delta_{j,l} + \delta_{i,l}\delta_{j,k}$ in C_B. With this choice, for an arbitrary pair of k and l,

$$C_{B^{(k,l)}} = u_k^\top u_l - w_k^\top w_l.$$

and

$$W^\top B^{(k,l)} W = w_k w_l^\top + w_l w_k^\top, \tag{70}$$

$$U B^{(k,l)} U^\top = u_k u_l^\top + u_l u_k^\top. \tag{71}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\sum_{i=1}^{d_y} \mathrm{Tr}\left[\Sigma(\tilde{u}_i) B^{(k)}\right] &= \mathbb{E}\sum_{i=1}^{d_y} (\tilde{u}_i^\top W x - y_i)^2 \mathrm{Tr}\left[W x x^\top W^\top B^{(k)}\right] && (72)\\
&= \mathbb{E}\left[\|r\|^2 \mathrm{Tr}[W x x^\top W^\top B^{(k,l)}]\right] && (73)\\
&= \mathrm{Tr}\left[\mathbb{E}[\|r\|^2 x x^\top] W^\top B^{(k)} W\right] && (74)\\
&= \mathrm{Tr}[\Sigma'_W (w_k w_l^\top + w_l w_k^\top)] && (75)\\
&= 2 w_k^\top \Sigma'_w w_l && (76)
\end{aligned}$$

where we have defined $r_i = \tilde{u}_i^\top W x - y_i$ and $\Sigma'_W = \mathbb{E}[\|r\|^2 x x^\top]$.

Likewise, we have that

$$\begin{aligned}
\mathbb{E}\sum_{j=1}^{d_x} \mathrm{Tr}\left[\Sigma(\tilde{w}_j) B^{(k)}\right] &= \mathbb{E}\sum_{j=1}^{d_x} x_j^2 \mathrm{Tr}\left[\left(\sum_{i=1}^{d_y} \tilde{u}_i (\tilde{u}_i^\top W x - y_i)\right)\left(\sum_{i=1}^{d_y} (\tilde{u}_i^\top W x - y_i)\tilde{u}_i^\top\right) B^{(k)}\right]\\
&= \mathrm{Tr}\left[\mathbb{E}[\|x\|^2 U^\top r r^\top U B^{(k,l)}]\right]\\
&= \mathrm{Tr}[\Sigma'_U U B^{(k,l)} U^\top]\\
&= 2 u_k^\top \Sigma'_u u_l.
\end{aligned}$$

where we have defined $\Sigma'_u = \mathbb{E}[\|x\|^2 r r^\top]$. Therefore, we have found that for arbitrary pair of k and l

$$\dot{C}_{B^{(k,l)}} = -2\gamma(u_k^\top u_l - w_k^\top w_l) + 2\eta(w_k^\top \Sigma'_w w_l - u_k^\top \Sigma'_u u_l). \tag{77}$$

The fixed point of this dynamics is:

$$w_k^\top \Sigma_w w_l = u_k^\top \Sigma_u u_l. \tag{78}$$

where $\Sigma_w = \eta \Sigma'_w + \gamma I$ and $\Sigma_U = \eta \Sigma'_u + \gamma I$. Because this holds for arbitrary k and l, the equation can be written in a matrix form:

$$W \Sigma_w W^\top = U^\top \Sigma_u U. \tag{79}$$

Let $V = UW$. To show that a solution exists for an arbitrary V. Let $W' = W\sqrt{\Sigma_w}$ and $U' = \sqrt{\Sigma_u} U$, which implies that

$$U' W' = \sqrt{\Sigma_u} V \sqrt{\Sigma_w} := V', \tag{80}$$

and

$$W' (W')^\top = (U')^\top U'. \tag{81}$$

Namely, U' and $(W')^\top$ must have the same right singular vectors and singular values. This gives us the following solution. Let $V' = LSR$ be the singular value decomposition of V', where L and R are orthogonal matrices an S is a positive diagonal matrix. Then, for an arbitrary orthogonal matrix F, the following choice of U' and W' satisfies the two conditions:

$$\begin{cases} U' = L\sqrt{S}F; \\ W' = F^\top \sqrt{S} R. \end{cases} \tag{82}$$

This finishes the proof. □

C Noise-Balanced Solution of Deep Linear Networks

Here, we apply our result to derive the exact solution of an arbitrarily deep and wide deep linear network, which has been under extensive study due to its connection in loss landscape to deep neural networks [3, 4, 15, 21, 34, 30]. Deep linear networks have also been a major model for understanding the implicit bias of GD [1]. The per-sample loss for a deep linear network can be written as:

$$\ell(\theta) = \|W_D ... W_1 x - y\|^2, \tag{83}$$

where W_i is an arbitrary dimensional matrix for all i. The global minimum is realizable: $y = Vx + \epsilon$, for i.i.d. noises ϵ. Because there is a double rotation symmetry between every two neighboring matrices, the Noether charge can be defined with respect to every such pair of matrices. Let B_i be a symmetric matrix; we define the charges to be $C_{B_i} = W_i^T B_i W_i$. The noise equilibrium solution is given by the following theorem.

Theorem C.1. *Let $W_D ... W_1 = V$. Let $V' = \sqrt{\Sigma_\epsilon} V \sqrt{\Sigma_x}$ such that $V' = LS'R$ is its SVD and $d = \mathrm{rank}(V')$. Then, for all i and all B_i, a noise equilibrium for C_{B_i} at the global minimum is*

$$\sqrt{\Sigma_\epsilon} W_D = L\Sigma_D U_{D-1}^\top,\ W_i = U_i \Sigma_i U_{i-1}^\top,\ W_1 \sqrt{\Sigma_x} = U_1 \Sigma_1 R, \tag{84}$$

for $i = 2, \cdots, D-1$. U_i are arbitrary matrices satisfying $U_i^T U_i = I_{d\times d}$, and Σ_i are diagonal matrices such that

$$\Sigma_1 = \Sigma_D = \left(\frac{d}{\mathrm{Tr}S'}\right)^{(D-2)/2D} \sqrt{S'},\ \Sigma_i = \left(\frac{\mathrm{Tr}S'}{d}\right)^{1/D} I_{d\times d}. \tag{85}$$

This solution has quite a few striking features. Surprisingly, the norms of all intermediate layers are balanced:

$$\mathrm{Tr}[\Sigma_1^2] = \mathrm{Tr}[\Sigma_i^2] = (\mathrm{Tr}S')^{2/D} d^{1-2/D}. \tag{86}$$

All intermediate layers are thus rescaled orthogonal matrices aligned with the neighboring matrices and the only two matrices that process information are the first and the last layer. This explains a puzzling experimental result first observed in Ref. [26], where the authors showed that the neural networks find similar solutions when the model is initialized with the standard init., where there is no alignment at the start, and with the aligned init. Thus, the balance and alignment between different layers in the neural networks can be attributed to the rescaling symmetry between each pair of matrices.

D Proof of Theorem C.1

We first prove the following theorem, which applies to an arbitrary parameter that are not necessarily local minima of of the loss.

Theorem D.1. *Let $r = W_D \cdots W_1 x - y$, $\xi_{i+1} := W_D \cdots W_{i+2}$ and $h_i := W_{i-1} \cdots W_1$. For all layer i, the equilibrium is achieved at*

$$W_{i+1}^\top \xi_{i+1}^\top C_0^i \xi_{i+1} W_{i+1} = W_i h_i C_1^i h_i^\top W_i^\top, \tag{87}$$

where $C_0^i = \mathbb{E}[\|h_i x\|^2 r r^\top], C_1^i := \mathbb{E}[\|\xi_{i+1}^\top r\|^2 x x^\top]$. Or equivalently,

$$\xi_i^\top C_0^i \xi_i = h_{i+1} C_1^i h_{i+1}^\top. \tag{88}$$

Proof. By Proposition **??**,

$$\frac{d}{dt} C_B^i = \sigma^2 \left(\mathbb{E}\mathrm{Tr}\left[\frac{\partial \ell}{\partial W_{i+1}} B \left(\frac{\partial \ell}{\partial W_{i+1}} \right)^\top \right] - \mathbb{E}\mathrm{Tr}\left[\left(\frac{\partial \ell}{\partial W_i} \right)^\top B \frac{\partial \ell}{\partial W_i} \right] \right). \tag{89}$$

The derivatives are

$$\frac{\partial \ell}{\partial W_{i+1}} = \xi_{i+1}^\top r (W_i h_i x)^\top, \tag{90}$$

$$\frac{\partial \ell}{\partial W_i} = \xi_{i+1}^\top W_{i+1}^\top r (h_i x)^\top. \tag{91}$$

Therefore, the two terms on R.H.S of Eq. (89) are given by

$$\begin{aligned}
\mathbb{E}\mathrm{Tr}\left[\frac{\partial \ell}{\partial W_{i+1}} B \left(\frac{\partial \ell}{\partial W_{i+1}}\right)^\top\right] &= \mathbb{E}\mathrm{Tr}[\xi_{i+1}^\top r (W_i h_i x)^\top B (W_i h_i x) r^\top \xi_{i+1}], \\
&= \mathbb{E}\|\xi_{i+1}^\top r\|^2 \mathrm{Tr}[h_i x x^\top h_i^\top W_i^\top B W_i] \qquad (92) \\
\mathbb{E}\mathrm{Tr}\left[\left(\frac{\partial \ell}{\partial W_i}\right)^\top B \frac{\partial \ell}{\partial W_i}\right] &= \mathrm{Tr}[W_{i+1}^\top \xi_{i+1}^\top r (h_i x)^\top B (h_i x) r^\top \xi_{i+1} W_{i+1}] \\
&= \mathbb{E}[\|h_i x\|^2 \mathrm{Tr}[W_{i+1}^\top \xi_{i+1}^\top r r^\top \xi_{i+1} W_{i+1} B]]. \qquad (93)
\end{aligned}$$

Because the matrix B is arbitrary, we can let $B_{i,j} = \delta_{i,k}\delta_{j,l} + \delta_{i,l}\delta_{j,k}$. Then, the two terms become

$$\mathbb{E}\mathrm{Tr}\left[\frac{\partial \ell}{\partial W_{i+1}} B \left(\frac{\partial \ell}{\partial W_{i+1}}\right)^\top\right] = 2\mathbb{E}[\|\xi_{i+1}^\top r\|^2 \tilde{w}_{i,k}^\top h_i x x^\top h_i^\top \tilde{w}_{i,l}], \tag{94}$$

$$\mathbb{E}\mathrm{Tr}\left[\left(\frac{\partial \ell}{\partial W_i}\right)^\top B \frac{\partial \ell}{\partial W_i}\right] = 2\mathbb{E}[\|h_i x\|^2 \tilde{w}_{i+1,k}^\top \xi_{i+1}^\top r r^\top \xi_{i+1} \tilde{w}_{i+1,l}]. \tag{95}$$

Here, we define the vectors $W_i = (\tilde{w}_{i,1}^\top, \cdots, \tilde{w}_{i,d}^\top)^\top$ and $W_{i+1} = (\tilde{w}_{i+1,1}, \cdots, \tilde{w}_{i+1,d})$. Because Eq. (94) and (95) hold for arbitrary k, l, we have

$$\mathbb{E}\mathrm{Tr}\left[\frac{\partial \ell}{\partial W_{i+1}} B \left(\frac{\partial \ell}{\partial W_{i+1}}\right)^\top\right] = 2 W_i h_i \mathbb{E}[\|\xi_{i+1}^\top r\|^2 x x^\top] h_i^\top W_i^\top, \tag{96}$$

$$\mathbb{E}\mathrm{Tr}\left[\left(\frac{\partial \ell}{\partial W_i}\right)^\top B \frac{\partial \ell}{\partial W_i}\right] = 2 W_{i+1}^\top \xi_{i+1}^\top \mathbb{E}[\|h_i x\|^2 r r^\top] \xi_{i+1} W_{i+1}. \tag{97}$$

For Eq. (89) to be 0, we must have

$$W_i h_i \mathbb{E}[\|\xi_{i+1}^\top r\|^2 x x^\top] h_i^\top W_i^\top = W_{i+1}^\top \xi_{i+1}^\top \mathbb{E}[\|h_i x\|^2 r r^\top] \xi_{i+1} W_{i+1}, \tag{98}$$

which is Eq. (87). The proof is complete. □

We are now ready to prove Theorem C.1.

Proof. It suffices to specialize Theorem D.1 to the global minimum. At the global minimum, we can define

$$r = W_D^* \cdots W_1^* x - y = \epsilon. \tag{99}$$

Then, Eq. (87) can be written as

$$W_{i+1}^\top \frac{W_{i+2}^\top \cdots W_D^\top \Sigma_\epsilon W_D \cdots W_{i+2}}{\mathrm{Tr}[W_{i+2}^\top \cdots W_D^\top \Sigma_\epsilon W_D \cdots W_{i+2}]} W_{i+1} = W_i \frac{W_{i-1} \cdots W_1 \Sigma_x W_1^\top \cdots W_{i-1}^\top}{\mathrm{Tr}[W_{i-1} \cdots W_1 \Sigma_x W_1^\top \cdots W_{i-1}^\top]} W_i^\top. \tag{100}$$

To solve Eq. (100), we substitute W_D and W_1 with $W_1' = W_1 \sqrt{\Sigma_x}$ and $W_D' = \sqrt{\Sigma_\epsilon} W_D$, which transform Eq. (100) into

$$W_{i+1}^\top \frac{W_{i+2}^\top \cdots W_D'^\top W_D' \cdots W_{i+2}}{\mathrm{Tr}[W_{i+2}^\top \cdots W_D'^\top W_D' \cdots W_{i+2}]} W_{i+1} = W_i \frac{W_{i-1} \cdots W_1' W_1'^\top \cdots W_{i-1}^\top}{\mathrm{Tr}[W_{i-1} \cdots W_1' W_1'^\top \cdots W_{i-1}^\top]} W_i^\top. \tag{101}$$

The global minimum condition can be written as

$$W_D' W_{D-1} \cdots W_2 W_1' = \sqrt{\Sigma_\epsilon} V \sqrt{\Sigma_x} := V'. \tag{102}$$

Then, we can decompose the matrices $W_1', \cdots, W_D'$ as

$$W_D' = L \Sigma_D U_{D-1}^\top, \; W_i = U_i \Sigma_i U_{i-1}^\top (i \neq 1, D), \; W_1' = U_1 \Sigma_1 R, \tag{103}$$

where $\Sigma_D, \cdots, \Sigma_1 \in \mathbb{R}^{d \times d}$, $L \in \mathbb{R}^{d_y \times d}$, $U_i \in \mathbb{R}^{d_i \times d}$, $R \in \mathbb{R}^{d \times d_x}$ with $d := \mathrm{rank}(V')$ and arbitrary d_i. The matrices U_i satisfy $U_i^\top U_i = I_{d \times d}$. By substituting the decomposition into Eq. (101), we have

$$\frac{\Sigma_{i+1} \cdots \Sigma_D \Sigma_D \cdots \Sigma_{i+1}}{\mathrm{Tr}[\Sigma_{i+2} \cdots \Sigma_D \Sigma_D \cdots \Sigma_{i+2}]} = \frac{\Sigma_i \cdots \Sigma_1 \Sigma_1 \cdots \Sigma_i}{\mathrm{Tr}[\Sigma_{i-1} \cdots \Sigma_1 \Sigma_1 \cdots \Sigma_{i-1}]}. \tag{104}$$

Since these diagonal matrices commute with each other, we can see $\Sigma_i = cI_{d\times d}$. Then we move on to fix the parameter c. By taking $i = 1$ and $i = D-1$ in Eq. (104), we obtain

$$\frac{\Sigma_2^2 \cdots \Sigma_D^2}{\mathrm{Tr}[\Sigma_3^2 \dots \Sigma_D^2]} = c^2 \frac{\Sigma_D^2}{\mathrm{Tr}[\Sigma_D^2]} = \frac{\Sigma_1^2}{d}, \tag{105}$$

$$\frac{\Sigma_D^2}{d} = \frac{\Sigma_1^2 \cdots \Sigma_{D-1}^2}{\mathrm{Tr}[\Sigma_1^2 \dots \Sigma_{D-1}^2]} = c^2 \frac{\Sigma_1^2}{\mathrm{Tr}[\Sigma_1^2]}, \tag{106}$$

where d represents the dimension of the learning space. By taking trace to both sides of Eqs. (105) and (106), we can see $\mathrm{Tr}[\Sigma_1^2] = \mathrm{Tr}[\Sigma_D^2]$ and hence, $\Sigma_1 = \Sigma_D$. The parameter c is given by

$$c = \sqrt{\frac{\mathrm{Tr}[\Sigma_1^2]}{d}}. \tag{107}$$

With the SVD decomposition $V' = LS'R$, we have

$$\Sigma_1^2 c^{D-2} = S'. \tag{108}$$

Therefore, the solutions for c and Σ_1 are

$$c = \left(\frac{\mathrm{Tr}S'}{d}\right)^{1/D}, \ \Sigma_1 = \frac{\sqrt{S'}}{c^{(D-2)/2}} = \left(\frac{d}{\mathrm{Tr}S'}\right)^{(D-2)/2D} \sqrt{S'}. \tag{109}$$

The scaling of the diagonal matrices are shown as

$$\mathrm{Tr}[\Sigma_1^2] = d^{1-2/D}(\mathrm{Tr}S')^{2/D}, \ \mathrm{Tr}[\Sigma_i^2] = (\mathrm{Tr}S')^{2/D} d^{1-2/D} = \mathrm{Tr}[\Sigma_1^2]. \tag{110}$$

The proof is complete. □

E Discrete-Time SGD

In fact, our results hold in a similar form for discrete-time SGD. Let us focus on the exponential symmetries with the symmetric matrix A.

The following equation holds with probability 1:

$$0 = \nabla_\theta \ell(\theta, z) \cdot A\theta. \tag{111}$$

For discrete-time SGD, it is notationally simpler and without loss of generality to regard $\ell(\theta)$ as the minibatch-averaged loss, which is the notation we adopt here. This is because if a symmetry holds for every per-sample loss, then it must also hold for every empirical average of these per-sample losses.

The dynamics of SGD gives

$$\Delta\theta_t = -\eta \nabla_\theta \ell(\theta_t, z). \tag{112}$$

This means that

$$\Delta\theta_t \cdot J(\theta) = 0. \tag{113}$$

Therefore, we have that

$$\Delta C_t = 2\Delta\theta_t^\top A\theta_t + \Delta\theta_t^\top A \Delta\theta_t \tag{114}$$

$$= \Delta\theta_t^\top A \Delta\theta_t. \tag{115}$$

Therefore,

$$\Delta C_t = \eta^2 \mathrm{Tr}[\tilde{\Sigma}_d(\theta) A], \tag{116}$$

where $\tilde{\Sigma}_d(\theta) = \theta_t \theta_t^\top$ is by definition PSD. Already, note the similarity between Eq. (116) and its continuous-time version. The qualitative discussions carry over: if A is PSD, C_t increases monotonically.

Now, while the first-order terms in η also vanish in the r.h.s, the problem is that the r.h.s. becomes stochastic because $\Sigma_d(\theta_t)$ is different for every time step. However, one can still analyze the expected flow and show that the expected flow (over the sampling of minibatches) is zero at a unique point in a way similar to the continuous-time limit of the problem. Therefore, we define

$$G_d(\theta_t) = \mathbb{E}_z[\Delta C_t], \tag{117}$$

$$\Sigma_d(\theta_t) = \mathbb{E}_z[\tilde{\Sigma}_d]. \tag{118}$$

We can now prove the following theorem.

Theorem E.1. *(Discrete-time fixed point theorem of SGD.) Let the per-sample loss satisfy the A exponential symmetry and $\theta_\lambda := \exp[\lambda A]\theta$. Then, for any θ and any $\gamma \in \mathbb{R}$,*

(1) $G_d(\theta_\lambda)$ is a monotonically decreasing function of λ;
(2) there exists a $\lambda^* \in \mathbb{R} \cup \{\pm\infty\}$ such that $G_d(\theta_\lambda) = 0$;
(3) in addition, if $\mathrm{Tr}[\Sigma_d(\theta)A] \neq 0$ or $\mathrm{Tr}[\theta\theta^\top A] \neq 0$, λ^* is unique and $G_d(\theta_\lambda)$ is strictly monotonic;
(4) in addition to (3), if $\Sigma_d(\theta)$ is differentiable, $\lambda^*(\theta)$ is a differentiable function of θ.

Proof. Similarly, let us establish the relationship between $\nabla\ell(\theta)$ and $\nabla\ell(\exp(\lambda A))$. By the definition of the exponential symmetry, we have that for an arbitrary λ,

$$\ell(\theta) = \ell(e^{\lambda A}\theta). \tag{119}$$

Taking the derivative of both sides, we obtain that

$$\nabla_\theta \ell(\theta) = e^{\lambda A}\nabla_{\theta_\lambda}\ell(\theta_\lambda), \tag{120}$$

The standard result of Lie groups shows that $e^{\lambda A}$ is full-rank and symmetric, and its inverse is $e^{-\lambda A}$. Therefore, we have

$$e^{-\lambda A}\nabla_\theta \ell(\theta) = \nabla_{\theta_\lambda}\ell(\theta_\lambda). \tag{121}$$

Now, we apply this relation to the trace of interest. By definition,

$$\Sigma_d(\theta_\lambda) = \mathbb{E}[\nabla_{\theta_\lambda}\ell(\theta_\lambda)\nabla^\top_{\theta_\lambda}\ell(\theta_\lambda)] \tag{122}$$
$$= e^{-\lambda A}\Sigma_d(\theta)e^{-\lambda A}. \tag{123}$$

Because $e^{\lambda A}$ is a function of A, it commutes with A. Therefore,

$$\mathrm{Tr}[\Sigma_d(\theta_\lambda)A] = \mathrm{Tr}[e^{-\lambda A}\Sigma_d(\theta)e^{-\lambda A}A] \tag{124}$$
$$= \mathrm{Tr}[e^{-2\lambda A}\Sigma_d(\theta)A]. \tag{125}$$

Similarly, the regularization term is

$$\gamma\theta_\lambda^\top A\theta_\lambda = \gamma\mathrm{Tr}[\theta^\top\theta A e^{2\lambda A}] \tag{126}$$

Now, if $\mathrm{Tr}[\Sigma_d(\theta)A] = \theta^\top A\theta = 0$, we have already proved item (2) of the theorem. Therefore, let us consider the case when either (or both) $\mathrm{Tr}[\Sigma_d(\theta)A] \neq 0$ or $\theta^\top A\theta \neq 0$

Without loss of generality, we assume $\gamma \geq 0$, and the case of $\gamma < 0$ follows an analogous proof. In such a case, we can write the trace in terms of the eigenvectors n_i of A:

$$-\gamma\theta_\lambda^\top A\theta_\lambda + \eta\mathrm{Tr}[\Sigma_d(\theta_\lambda)A] = \underbrace{\eta\sum_{\mu_i>0} e^{-2\lambda|\mu_i|}|\mu_i|\sigma_i^2 + \gamma\sum_{\mu_i<0} e^{-2\lambda|\mu_i|}|\mu_i|\tilde{\theta}_i^2}_{I_1(\lambda)} - \underbrace{\left(\eta\sum_{\mu_i<0} e^{2\lambda|\mu_i|}|\mu_i|\sigma_i^2 + \gamma\sum_{\mu_i>0} e^{2\lambda|\mu_i|}|\mu_i|\tilde{\theta}_i^2\right)}_{I_2(\lambda)}$$
$$=: I(\lambda),$$

where μ_i is the i-th eigenvalue of A, $\tilde{\theta}_i = (n_i^\top\theta_i)^2$, $\sigma_i^2 = n_i^\top\Sigma_d(\theta)n_i \geq 0$ is the norm of the projection of Σ_d in this direction.

By definition, I_1 is either a zero function or strictly monotonically increasing function with $I_1(-\infty) = +\infty, I_1(+\infty) = 0$ Likewise, I_2 is either a zero function or a strictly monotonically increasing function with $I_2(-\infty) = 0, I_2(+\infty) = +\infty$. By the assumption $\mathrm{Tr}(\Sigma_d(\theta)A) \neq 0$ or $\mathrm{Tr}(\theta\theta^\top A) \neq 0$, we have that at least one of I_1 and I_2 must be a strictly monotonic function.

- If I_1 or I_2 is zero, we can take λ to be either $+\infty$ or $-\infty$ to satisfy the condition.
- If both I_1 and I_2 are nonzero, then $I = I_1 - I_2$ is a strictly monotonically decreasing function with $I(-\infty) = +\infty$ and $I(+\infty) = -\infty$. Therefore, there must exist only a unique $\lambda^* \in \mathbb{R}$ such that $I(\lambda^*) = 0$.

The proof of (4) follows from the Implicit Function Theorem, as in the continuous-time case. □

The final question is this: what does it mean for θ to reach a point where $G_d(\theta) = 0$? An educated guess can be made: the fluctuation in C does not vanish, but the flow takes C towards this vanishing flow point – something like a Brownian motion trapped in a local potential well [29]. However, it is difficult to say more without specific knowledge of the systems.