Stochastic Gradient Descent and Anomaly of Variance-Flatness Relation in Artificial Neural Networks

Xia Xiong(熊霞)[1], Yong-Cong Chen(陈永聪)[1*], Chunxiao Shi(石春晓)[1], and Ping Ao(敖平)[2]

[1]*Shanghai Center for Quantitative Life Sciences and Physics Department, Shanghai University, Shanghai 200444, China*
[2]*Colloge of Biomedical Engineering, Sichuan University, Chengdu 610065, China*

(Received 19 May 2023; accepted manuscript online 19 July 2023)

Stochastic gradient descent (SGD), a widely used algorithm in deep-learning neural networks, has attracted continuing research interests for the theoretical principles behind its success. A recent work reported an anomaly (inverse) relation between the variance of neural weights and the landscape flatness of the loss function driven under SGD [Feng Y and Tu Y *Proc. Natl. Acad. Sci. USA* **118** e2015617118 (2021)]. To investigate this seeming violation of statistical physics principle, the properties of SGD near fixed points are analyzed with a dynamic decomposition method. Our approach recovers the true "energy" function under which the universal Boltzmann distribution holds. It differs from the cost function in general and resolves the paradox raised by the the anomaly. The study bridges the gap between the classical statistical mechanics and the emerging discipline of artificial intelligence, with potential for better algorithms to the latter.

DOI: 10.1088/0256-307X/40/8/080202

An artificial neural network (ANN) is a machine learning platform which resembles a human neural system and its learning process. The system includes at least three layers: an input layer, a hidden layer, and an output layer. When there are multiple hidden layers, it is called a deep neural network (DNN), whose efficient training has become a major topic in the emerging discipline.[1–3] Most DNNs engage routines that adjust the weights of neural connections and minimize the so-called loss function for the learning process.[4–6] Among them, stochastic gradient descent (SGD) has been a particularly successful algorithm, which appears in almost all DNN applications. However, the theoretical principle behind its tremendous success is still an on-going research topic.[7–9] Possibly, the high degree of stochastic anisotropy associated with SGD enables the optimization to escape from local minima en route to the best configuration.[10]

In search for the answer, there has been a number of ideas borrowed from the classical statistical physics.[11–14] Recently, Feng and Tu[15] regarded the loss function as resembling the "thermodynamic" energy function for a "physical" system. They derived an equivalent stochastic equation and performed principal component analysis (PCA) on the weight dynamics. The latter reduced the dimensions of the complex stochastic process in accordance with their significance: The weights were projected onto the PCA axes and further analyzed. When the loss function[16,17] was characterized by its distribution width on the PCA directions, they found a robust anomaly between the weight variance and the flatness of the landscape, in striking contrast to what would follow from the conventional Boltzmann distribution. Apparently, ANN appeared to violate a fundamental principle of physics and display what they referred to as an "inverse Einstein relation".

This work investigates such unconventional characteristic in detail. We start with a formalism of stochastic decomposition developed by Ao *et al.*[18] The method obtains for the dynamics a Lyapunov functional which is an alternative to the cost function. It is a close analogue to the energy function of a physical system under which statistical principle holds.[19] For a linear system, the deterministic driving force in the process can be strictly decomposed into two parts. One of them leads to balance of probabilistic distribution and the other to cyclic motion on the surface of constant potential function. Together they offer near a fixed point a Boltzmann-like probability distribution as well as flux-carrying stationary states without detailed balance. This methodology has been applied to a number of practical systems,[20–22] in particular, complex networks for biological studies.[23]

In our approach, the covariance matrix for the dynamics is in fact inversely proportional to the energy matrix $\boldsymbol{U}$ for the stochastic potential. Such a relationship well resolves the paradox raised by the anomaly in the variance-flatness relation (VFR). We further examine the scaling behaves of the weight variance and the flatness by explicit reconstruction of the cost function in the linear region from the fluctuation data (i.e., diffusion matrix). The study of stochastic decomposition in ANN may offer better algorithm that can address problems such as the well-known "forgetting" catastrophe[24,25] while performing multi tasks.

Continuous Limit of SGD. ANN optimizes its state vector, i.e., weights and biases through a training process. In general, the loss function can fairly complicated with a huge parameter space.[26–28] On a gradient descent scheme, the state vector advances a certain step from the current position along the opposite direction of the gradient. The latter is re-calculated at the new position to con-

*Corresponding author. Email: chenyongcong@shu.edu.cn
© 2023 Chinese Physical Society and IOP Publishing Ltd

tinue the optimization. When the data used for training are randomly selected mini-batches, the scheme is known as the stochastic gradient descent,[29,30] or simply SGD. Specifically, the state vector $\boldsymbol{\omega}_k$, an array of weights in question at the k^{th} stage, is now updated via

$$\boldsymbol{\omega}_{k+1} = \boldsymbol{\omega}_k - \alpha \nabla_\omega L^\mu(\boldsymbol{\omega}_k), \tag{1}$$

where $\nabla_\omega L^\mu(\boldsymbol{\omega})$ is the gradient over the loss function of the μ^{th} mini-batch (of size $B \gg 1$), and α is a small learning rate. Note that throughout this work, a boldface symbol is used for matrix or vector, while the same in normal face is for the underlying matrix element(s).

Following Feng and Tu,[15] the so-called continuous-time limit[31] of SGD can be cast into a conventional set of stochastic differential equation as

$$\dot{\boldsymbol{\omega}} = -\nabla_\omega L(\boldsymbol{\omega}) + \boldsymbol{\eta}(t). \tag{2}$$

In Eq. (2) we define $\dot{\boldsymbol{\omega}} \equiv d\boldsymbol{\omega}/dt$ with $dt \equiv \alpha$, and $L(\boldsymbol{\omega})$ is the "correct" loss function for the learning process (an average over all samples, cf. below). The SGD noise term $\boldsymbol{\eta} \equiv -\nabla_\omega(L^\mu - L)$ arises from the variation between a mini-batch and the full-batch loss functions. The noise has zero ensemble average $\langle \boldsymbol{\eta} \rangle_\mu = 0$, but non-zero variance $\langle \eta_i(t)\eta_j(t') \rangle_\mu$ for $t = t'$.

The learning set is usually assumed to be sufficiently large so that there is little correlation between the mini-batches. In an SGD process when a mini-batch is sampled with replacement, the variance reads[10]

$$\langle (\nabla L^\mu - \nabla L)(\nabla L^\mu - \nabla L) \rangle_\mu = \boldsymbol{D}(\boldsymbol{\omega})/B. \tag{3}$$

Here $\boldsymbol{D}(\boldsymbol{\omega})$ is a diffusion matrix independent of the mini-batches, which can be computed from

$$\boldsymbol{D}(\boldsymbol{\omega}) \approx \Big(\frac{1}{N_{\text{L}}} \sum_{k=1}^{N_{\text{L}}} \nabla L_k \nabla L_k \Big) - \nabla L(\boldsymbol{\omega}) \nabla L(\boldsymbol{\omega}), \tag{4}$$

where $L_k \equiv L_k(\boldsymbol{\omega})$ is the loss function for the k^{th} sample, $L(\boldsymbol{\omega}) \equiv (1/N_{\text{L}}) \sum_{k=1}^{N_{\text{L}}} L_k$, and N_{L} is the total size of the learning set. A slight variation but essentially the same for $B/N_{\text{L}} \ll 1$ can be obtained for sampling without replacement. Evidently, to cast $\langle \eta_i(t)\eta_j(t') \rangle_\mu$ into the form $2\epsilon D_{ij}\delta(t-t')$, the noise strength ϵ reads $\alpha/2B$.

Anomaly of Variance-Flatness Relation. We next summarize the main result obtained by Feng and Tu.[15] After carrying out PCA on the SGD process, the weight dynamics can be projected out for variations in the principal axes, $\boldsymbol{\omega}(t) = \langle \boldsymbol{\omega} \rangle_T + \sum_i \theta_i(t)\boldsymbol{p}_i$. Here $\langle \boldsymbol{\omega} \rangle_T \equiv \boldsymbol{\omega_0}$ is the average weight vector in a particular epoch time of length T, $\boldsymbol{p}_i$ is the i^{th} principal base vector and $\theta_i(t)$ is the projection on the direction. Let the loss function profile along $\boldsymbol{p}_i$ be $L_i(\delta\theta) \equiv L(\boldsymbol{\omega_0} + \delta\theta \boldsymbol{p}_i)$, numerous data simulations have found that L_i becomes flatter with the increase of i.

To quantify the behavior, a flatness parameter can be introduced as follows:[32–34] $F_i = \theta_i^r - \theta_i^l$ ($\theta_i^l < 0$, $\theta_i^r > 0$) is taken as the difference between the nearest two points of equal value near the minimum found by SGD. Specifically $L_i(\theta_i^l) = L_i(\theta_i^r) = e \times L_0$, where e is Euler's number and L_0 is the loss function at local minimum. Now F_i is found to increase with PCA index i, whereas the SGD variance σ_i^2 decreases with i. Finally the VFR follows approximately power-law behavior as follows:

$$\sigma_i^2 \sim F_i^{-4}. \tag{5}$$

The above result is counter-intuitive due to the fact that the flatter the "energy function" is, the less diffusive the dynamics is. In statistical physics where the loss function L assumed the role of the energy function, the equilibrium probability distribution of $\boldsymbol{\theta}$ would follow Boltzmann distribution, i.e., $P(\boldsymbol{\theta}) = \exp[-L(\boldsymbol{\theta})/T]$ (with "empirical temperature" T). One would then have $\sigma_i^2 \propto F_i^2$ so that $\sigma_i^2/F_i^2 \sim$ constant, in stark contrast to Eq. (5). The latter apparently would violate a fundamental law of physics.

Covariance Matrix of the Stochastic Dynamics. As the covariance of the stochastic process plays a pivot role in the PCA,[35] we next introduce an alternative method of stochastic decomposition to analyze the dynamics. In doing so we gain the knowledge of a crucial relationship between the covariance and the "true" energy function that ought be used for the Boltzmann distribution, fulfilling a major goal of the present work.

A Review of the Stochastic Decomposition. A large class of stochastic processes in nature can be modeled by stochastic differential equations[36] in a generic form

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}) + \boldsymbol{\zeta}(t), \tag{6}$$

where $\boldsymbol{x}$ stands for a set of dynamical variables, a general state vector of an N-dimensional system. Here $f(\boldsymbol{x})$ and $\boldsymbol{\zeta}(t)$ are the deterministic and a Markovian driving force for the dynamics, respectively. In many cases, $\boldsymbol{\zeta}(t)$ can be represented by a (functional of) Gaussian white noise with $\langle \boldsymbol{\zeta}(t) \rangle = 0$ and semi-positive definite variance $\langle \boldsymbol{\zeta}(t)\boldsymbol{\zeta}^\tau(t') \rangle = 2\epsilon \boldsymbol{D}\delta(t-t')$. The superscript τ denotes the transpose of the underlying vector/matrix, $\langle \cdots \rangle$ is the average over the Gaussian distribution, and $\delta(t)$ is the Kronecker delta function for the Markovian process. Evidently $\boldsymbol{D}$ is a diffusion matrix with ϵ being the noise strength, which plays the role of temperature of the stochasticity. For the analysis below, we will focus on dynamics near a fixed point so that the first term in Eq. (6) may be approximated by $f_i(x) = -F_{ij}x_j$. Namely, it reduces to

$$\dot{\boldsymbol{x}} = -\boldsymbol{F}\,\boldsymbol{x} + \boldsymbol{\zeta}(t). \tag{7}$$

The aforementioned stochastic decomposition seeks to re-cast the dynamics into the following "canonical" form:

$$(\boldsymbol{S} + \boldsymbol{A})\dot{\boldsymbol{x}} = -\boldsymbol{U}\,\boldsymbol{x} + \boldsymbol{\xi}(t). \tag{8}$$

Among the terms, $\boldsymbol{U}$ is a symmetric potential matrix which gives rise to a stochastic potential or an energy function for the whole system,

$$u(\boldsymbol{x}) = \boldsymbol{x}^\tau \boldsymbol{U}\,\boldsymbol{x}/2. \tag{9}$$

$\boldsymbol{S}$ is symmetric, semi-positive definite. It represents the dissipative dynamics which, in the absence of fluctuations, causes monotonic decrease of $u(\boldsymbol{x})$. The canonical noise $\boldsymbol{\xi}(t)$ is associated with $\boldsymbol{S}$ by a generalized fluctuation-dissipation theorem $\langle \boldsymbol{\xi}(t)\boldsymbol{\xi}^\tau(t') \rangle = 2\epsilon \boldsymbol{S}\delta(t-t')$. On the other hand, $\boldsymbol{A}$ is antisymmetric and conserves $u(\boldsymbol{x})$. With the canonical form it can be shown[18] that the Boltzmann-like distribution

$$\rho(\boldsymbol{x}) \propto \exp\{-u(\boldsymbol{x})/\epsilon\} \tag{10}$$

is a stationary distribution for the system. That is, the potential function so constructed bears the essence of energy in statistical physics.

Going back to Eq. (7), the two norms of noise are related by

$$\boldsymbol{\xi}(t) = (\boldsymbol{S} + \boldsymbol{A})\boldsymbol{\zeta}(t), \tag{11}$$

$$\boldsymbol{S} = (\boldsymbol{S} + \boldsymbol{A})\boldsymbol{D}(\boldsymbol{S} - \boldsymbol{A}). \tag{12}$$

Furthermore, take the inverse of $(\boldsymbol{S} + \boldsymbol{A})^{-1}$ and break it down to symmetric $\tilde{\boldsymbol{D}}$ and anti-symmetric $\boldsymbol{Q}$. We find $(\boldsymbol{S} + \boldsymbol{A})\boldsymbol{Q}(\boldsymbol{S} - \boldsymbol{A}) = -\boldsymbol{A}$, and $\tilde{\boldsymbol{D}} \equiv \boldsymbol{D}$ [as both $\tilde{\boldsymbol{D}}$ and $\boldsymbol{D}$ satisfy the same Eq. (12)].

To obtain the potential function, we rewrite the force matrix $\boldsymbol{F}$ as

$$\boldsymbol{F} = (\boldsymbol{S} + \boldsymbol{A})^{-1}\boldsymbol{U} = (\boldsymbol{D} + \boldsymbol{Q})\boldsymbol{U}. \tag{13}$$

Using the symmetry of $\boldsymbol{U}$, $\boldsymbol{D}$, and $\boldsymbol{Q}$, $\boldsymbol{U}$ and $\boldsymbol{Q}$ can be determined uniquely from $\boldsymbol{F}$ and $\boldsymbol{D}$ via

$$\boldsymbol{F}\boldsymbol{U}^{-1} + \boldsymbol{U}^{-1}\boldsymbol{F}^{\tau} = 2\boldsymbol{D}, \tag{14}$$

$$\boldsymbol{F}\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{F}^{\tau} = \boldsymbol{F}\boldsymbol{D} - \boldsymbol{D}\boldsymbol{F}^{\tau}. \tag{15}$$

The above results hold near a stable fixed point where real parts of eigenvalues of $\boldsymbol{F}$ are positive, a condition that is assumed to be satisfied in what follows. Further details can be found in, e.g., Ref. [19].

Evaluation of Covariance Matrix. To proceed further we solve Eq. (7) for the system as

$$\boldsymbol{x}(t) = e^{-\boldsymbol{F}t}\Big[\boldsymbol{x}(0) + \int_0^t e^{\boldsymbol{F}t'}\boldsymbol{\zeta}(t')dt'\Big]. \tag{16}$$

The covariance is calculated from regulated $\boldsymbol{x}$ with zero means. Namely, $\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t) - (1/t)\int_0^t \boldsymbol{x}(t')dt'$. Near a stable fixed point the homogeneous part decays exponentially. Therefore, at $t \to \infty$ the covariance matrix $\boldsymbol{\Sigma}$ can be readily obtained as

$$\boldsymbol{\Sigma} = \langle \tilde{\boldsymbol{x}}(t)\tilde{\boldsymbol{x}}^{\tau}(t)\rangle_{t\to\infty} = 2\epsilon\int_{-\infty}^{0} dt\, e^{\boldsymbol{F}t}\boldsymbol{D}e^{\boldsymbol{F}^{\tau}t}. \tag{17}$$

The integration can be further carried out [37] to get

$$\boldsymbol{F}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{F}^{\tau} = 2\epsilon\int_{-\infty}^{0} dt\,\frac{d}{dt}[e^{\boldsymbol{F}t}\boldsymbol{D}e^{\boldsymbol{F}^{\tau}t}] = 2\epsilon\boldsymbol{D}. \tag{18}$$

It follows from Eq. (14) that the covariance matrix and the potential energy matrix are the mutual inverse of each other, $\boldsymbol{\Sigma} = \epsilon\boldsymbol{U}^{-1}$.

The largest principal components of $\boldsymbol{\Sigma}$ correspond to the eigenstates of $\boldsymbol{U}$ with smallest eigenvalues.

Statistical and Scaling Properties of SGD. Returning to Eq. (2), it is evident that we can set $\boldsymbol{\omega} \equiv \boldsymbol{x}$ and map the SGD noise $\boldsymbol{\eta}(t)$ to $\boldsymbol{\zeta}(t)$ in the current context, hence establish a relationship between the loss function, characterized by the matrix $\boldsymbol{F}$ via $L(\boldsymbol{x}) = L_0 + \boldsymbol{x}^{\tau}\boldsymbol{F}\boldsymbol{x}/2$ near a fixed point, and the stochastic potential set by $\boldsymbol{U}$, cf. Eq. (9). It suggests that the mystery of anomaly VFR may be traced back to the choice of the energy function. For an ANN, if the loss function is incorrectly assumed as the energy function, the relationship between variance and flatness will not conform to the usual statistical physics. Instead, they could show the counter intuitive VFR characteristic as seen in the following.

Boltzmann Distribution of Neural Weights. The correct Boltzmann distribution, i.e., Eq. (10), can be realized when the stochastic potential is recognized as the proper energy function. Since $\boldsymbol{\Sigma} \sim \boldsymbol{U}^{-1}$, the value of variance increases as the eigenvalue of $\boldsymbol{U}$ is reduced. Hence the landscape of $u(\boldsymbol{x})$ from Eq. (9) becomes flatter, leading to larger flatness F_i in terms of the new energy function. This is a positive correlation of VFR, contrary to what appears in Eq. (5). It resolves naturally the conceptual difficulty raised by Feng and Tu. [15]

Data Simulations and Scaling Properties. The explicit relationship among $\boldsymbol{F}$, $\boldsymbol{D}$, and $\boldsymbol{\Sigma}$ in Eq. (18) allows us to recover the details of the loss function through the $\boldsymbol{F}$ matrix near a fixed point. It in turn can be used to re-examine the scaling properties reported by Feng and Tu. [15] For that purpose, additional data simulations are carried out in the Supplementary Information. The main results are given in the following.

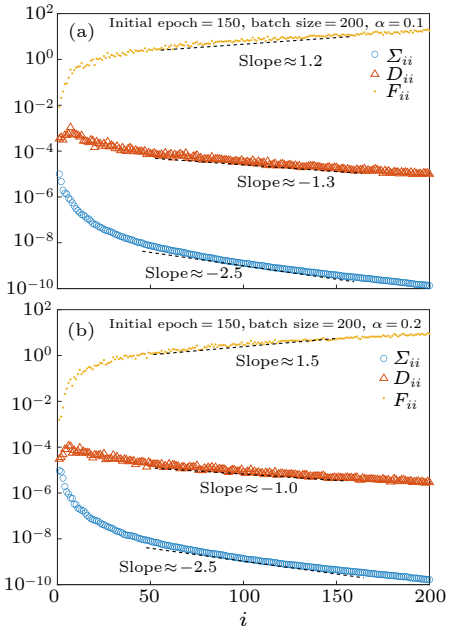

Fig. 1. The diagonal matrix elements of $\boldsymbol{\Sigma}$, $\boldsymbol{D}$, and $\boldsymbol{F}$ at different learning rates under the eigen-space of $\boldsymbol{\Sigma}$. The horizontal axis i represents the PCA index. (a) Learning rate $\alpha = 0.1$, which shows the scaling behavior as $\Sigma_{ii} \propto i^{-2.5}$, $D_{ii} \propto i^{-1.3}$, and $F_{ii} \propto i^{1.2}$. (b) Learning rate $\alpha = 0.2$, with slightly different scaling indexes as $\Sigma_{ii} \propto i^{-2.5}$, $D_{ii} \propto i^{-1.0}$, $F_{ii} \propto i^{1.5}$.

In an exemplary setup, we explore a system of fully connected neural network with two hidden layers. The weight elements between the two hidden layers are extracted to form the set of weights $\boldsymbol{\omega}$, i.e., the state vector $\boldsymbol{x}$ in the current context. The covariance matrix of the weights is calculated and diagonalized with eigenvalues ranked in descending order. The eigenvectors can be used as the convenience basis for the weight space. Furthermore, the diffusion matrix $\boldsymbol{D}$ can be sampled directly via Eq. (3) or calculated via Eq. (4). We can find that the difference is negligible. Note that $\boldsymbol{F}$ (being second derivatives on the loss function) is also symmetric, so it can be

easily obtained via Eq. (18) in the diagonal representation of $\boldsymbol{\Sigma}$, see the Supplementary Information for more details. The diagonal matrix elements of $\boldsymbol{D}$ and $\boldsymbol{F}$, D_{ii} and F_{ii} are shown in Fig. 1.

There are a few key points taken away from the results. First of all, D_{ii} of the diffusion matrix $\boldsymbol{D}$ also decreases with the PCA index i, *albeit* at a slower rate than Σ_{ii}. Secondly, it is (learning rate) α-dependent, which appears to suggest that larger α's lead to local minima of loss function that are smoother and less fluctuating under SGD. In the vicinity of a local minimum, we have $F_{ii} \sim F_i^{-2}$ by definition of flatness in the loss function. Again we take that function as the energy function for statistical distribution, and have the same anomaly that $\Sigma_{ii} \times F_{ii} \sim \Sigma_{ii}/F_i^2$ does not scale to a constant, rather it diminishes as i increases. On the other hand, we do not find the inverse relationship of Eq. (5) either. The discrepancy may be due to the flatness definition employed by Feng and Tu.[15] Their definition is more an intuitive and global one and the measure is obtained rather far away from the local minimum. Another factor might arise from the limitation on the continuous-time approximation of SGD. More remains to be explored.

Potential Algorithm Improvement. The stochastic decomposition itself can reveal rich structural features. In a related study,[22] the eigenvalues and eigenvectors of the $\boldsymbol{S}$ and $\boldsymbol{A}$ matrices [cf. Eq. (8)] are analyzed to a great extent. Under an isotropic diffusion environment, the dissipation matrix $\boldsymbol{S}$ can be expressed as the direct sum of multiple matrices, $\boldsymbol{S} = \boldsymbol{s}_1 \oplus \boldsymbol{s}_2 \oplus \cdots \boldsymbol{s}_{N/2}$, where $\boldsymbol{s}_i = s_i \mathbf{1}$ with the 2×2 unit matrix $\mathbf{1}$. Here, $s_i > 0$ is the i^{th} eigenvalue which can be sorted by $s_1 \leq s_2 \leq s_3 \leq \cdots \leq s_{N/2}$. In the same subspaces, $\boldsymbol{A}$ can also be diagonalized into a direct sum of pairs of 2×2 antisymmetric matrices. The system is often found circulating in the subspaces with small s_i's, i.e., it shows vortex-like behavior when the damping or viscosity is small.

In the current context, $\boldsymbol{S}$ and $\boldsymbol{A}$ can be obtained via Eq. (13) whereas $\boldsymbol{Q}$ can be obtained from Eq. (15) once $\boldsymbol{F}$ is known. The presence of $\boldsymbol{Q}$ is an indication that $\boldsymbol{D}$ and $\boldsymbol{F}$, as in the case of anisotropic $\boldsymbol{D}$, do not commute. The magnitude of $\boldsymbol{Q}$ may appear to be insignificant when compared to $\boldsymbol{F}$. However, the correct comparison should be drawn between $\boldsymbol{F}$ and $\boldsymbol{D}$ as they are the odd and even parts of one matrix in Eq. (13). After re-scaling the state variables so that $\boldsymbol{D}$ is transformed into the isotropic identity matrix $\boldsymbol{I}$, $\boldsymbol{Q}$ is enhanced by several orders of magnitude. A large $\boldsymbol{Q}$ will lead to small $\boldsymbol{S}$. Hence there will be abundance of vortices to emerge. More details concerning $\boldsymbol{Q}$ are presented in the Supplementary Information.

The characteristic may be useful for a solution on the well-known catastrophic weight-forgetting difficulty. The latter refers to the problem that when a neural network learns a new task under the premise that it has already learned an old one, it often performs poorly on the first task afterwards. This can be due to heavy shifting of neural weights in learning the new task such that the old settings are forgotten.[24] The structural vortices identified above may be used to confine the drifting of weights in a more lenient way than a brute-force pin-down of them to reach the first task. Namely, the weights are allowed to circulate in the weight space instead. Multiple tasks can be associated to vortices from different hidden layers. Any success will be an indication that our approach may indeed be adopted in ANNs to benefit future development.

Discussion. An ANN is a layer-structured network that emulates a biological neural network with a learning objective. The training process involves a forward propagation along the layers, each of them is activated nonlinearly by the input "signals" from the preceding layer. A loss function, typically in the form of cross-entropy loss for classification problems, is computed at the output layer. It is then followed by a back-propagation where the chain rule of derivative is used to compute the gradients with respect to the loss function. The structure, flows and algorithm allow simultaneous updates of the entire network parameters, making the whole routine very efficient. It is these unique features, added by the limited number of mini-batches in an epoch time, that give rise to the characteristics of low effective dimensionality or high anisotropy for the noise studied in this work.

To summarize, a dynamic decomposition is used to analyze the stochastic properties near fixed points of SGD in ANNs. Controversies from previous studies regarding some fundamental statistical principle such as the Einstein relation are resolved by the identification of stochastic potential function, i.e., the proper energy function which differs from the loss function used in ANNs when the stochasticity of SGD is concerned. In addition, the method and data simulations have revealed further characteristics and structures of SGD.

The success of stochastic dynamic decomposition in ANNs brings us insights into the theoretical understanding of deep learning systems, which may offer ideas for the subsequent development of better algorithms in artificial intelligence. For example, we are able to identify the non-vanishing transverse $\boldsymbol{Q}$ matrix for the stochastic process. The latter is found to be closely related to the vortex-like circulations in similar systems.[22] Such features may be further explored in the current field. It may be used to construct a more natural strategy targeting the catastrophic forgetting problem in DNNs.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Grant No. 16Z103060007(PA)). One of us (X.X.) thanks Dr. Pratik Chaudhari for communications and free codes from his work. Thanks to Shanghai Nanobubble Technology Co., Ltd. for supporting this work.

References

[1] LeCun Y, Bengio Y, and Hinton G 2015 *Nature* **521** 436
[2] Goodfellow I, Bengio Y, and Courville A 2016 *Deep Learning* (Cambridge: MIT Press)
[3] Aggarwal C C 2018 *Neural Networks and Deep Learning* (Berlin: Springer) p 105
[4] Le Q V, Ngiam J, Coates A, Lahiri A, Prochnow B, and Ng A Y 2011 *Proceedings of 28th International Conference on Machine Learning*, 28 June–2 July 2011, Madison, WI, USA, p 265
[5] Martens J 2010 *Proceedings of the 27th International Conference on Machine Learning*, 21–24 June 2010, Haifa, Israel, p 735

[6] Young S R, Rose D C, Karnowski T P, Lim S H, and Patton R M 2015 *Proceedings of the Workshop on Machine Learning in High-performance Computing Environments*, November 2015, New York, NY, USA, Article No. 4
[7] Advani M, Lahiri S, and Ganguli S 2013 *J. Stat. Mech.* **2013** P03014
[8] Baldassi C, Borgs C, Chayes J T, Ingrosso A, Lucibello C, Saglietti L, and Zecchina R 2016 *Proc. Natl. Acad. Sci. USA* **113** E7655
[9] Zhang C, Bengio S, Hardt M, Recht B, and Vinyals O 2021 *Commun. ACM* **64** 107
[10] Chaudhari P and Soatto S 2018 *Information Theory and Applications Workshop*, 11–16 February 2018, San Diego, CA, USA, pp 1–10
[11] Zhang Y, Saxe A M, Advani M S, and Lee A A 2018 *Mol. Phys.* **116** 3214
[12] Feng Y and Tu Y 2021 *Mach. Learn.: Sci. Technol.* **2** 043001
[13] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L, and Zdeborová L 2019 *Rev. Mod. Phys.* **91** 045002
[14] Mehta P, Bukov M, Wang C H, Day A G, Richardson C, Fisher C K, and Schwab D J 2019 *Phys. Rep.* **810** 1
[15] Feng Y and Tu Y 2021 *Proc. Natl. Acad. Sci. USA* **118** e2015617118
[16] Ghorbani B, Krishnan S, and Xiao Y 2019 *Proceedings of Machine Learning Research* **97** 2232
[17] Li H, Xu Z, Taylor G, Studer C, and Goldstein T 2018 *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, vol 31
[18] Ao P 2004 *J. Phys. A: Math. Gen.* **37** L25
[19] Kwon C, Ao P, and Thouless D J 2005 *Proc. Natl. Acad. Sci. USA* **102** 13029
[20] Chen Y C, Shi C, Kosterlitz J M, Zhu X, and Ao P 2020 *Proc. Natl. Acad. Sci. USA* **117** 23227
[21] Shi C, Chen Y C, Xiong X, and Ao P 2023 *J. Nonlinear Math. Phys.* **30** (accepted)
[22] Chen Y C, Shi C, Kosterlitz J M, Zhu X, and Ao P 2022 *Proc. Natl. Acad. Sci. USA* **119** e2211359119
[23] Yuan R, Zhu X, Wang G, Li S, and Ao P 2017 *Rep. Prog. Phys.* **80** 042701
[24] Robins A 1995 *Connection Sci.* **7** 123
[25] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu A A, Milan K, Quan J, Ramalho T, and Grabska-Barwinska A 2017 *Proc. Natl. Acad. Sci. USA* **114** 3521
[26] Bray A J and Dean D S 2007 *Phys. Rev. Lett.* **98** 150201
[27] Beer R D 2006 *Neural Comput.* **18** 3009
[28] Amari S I 1996 *Advances in Neural Information Processing Systems* vol 9
[29] Rattray M, Saad D, and Amari S I 1998 *Phys. Rev. Lett.* **81** 5461
[30] Sohl-Dickstein J, Poole B, and Ganguli S 2014 *Proceedings of the 31st International Conference on Machine Learning*, 21–26 June 2014, Beijing, China, vol 32 pp 604–612
[31] Sompolinsky H, Crisanti A, and Sommers H J 1988 *Phys. Rev. Lett.* **61** 259
[32] Hochreiter S and Schmidhuber J 1997 *Neural Comput.* **9** 1
[33] Chaudhari P, Choromanska A, Soatto S, LeCun Y, Baldassi C, Borgs C, Chayes J, Sagun L, and Zecchina R 2019 *J. Stat. Mech.* **2019** 124018
[34] Baldassi C, Pittorino F, and Zecchina R 2020 *Proc. Natl. Acad. Sci. USA* **117** 161
[35] Abdi H and Williams L J 2010 *Wiley Interdisciplinary Re41 views: Computational Statistics* **2** 433
[36] Van Kampen N G 1992 *Stochastic Processes in Physics and Chemistry* (Amsterdam: Elsevier Press)
[37] Han M, Park J, Lee T, and Han J H 2021 *Phys. Rev. E* **104** 034126