Phys. Rev. E

Dynamics of the Adaptive Natural Gradient Descent Method for Soft Committee Machines

Masato Inoue,* Hyeyoung Park,† and Masato Okada‡
Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Saitama 351-0198, Japan
(Received 13 November 2003; Published 28 May 2004)

Adaptive natural gradient descent (ANGD) realizes natural gradient descent (NGD) without needing to know the input distribution of learning data and reduces the calculation cost from a cubic order to a square order. However, no performance analysis of ANGD has been done. We have developed a statistical-mechanical theory of the simplified version of ANGDPresent address: ft commPresent address: Present address: Present address: Present address: Present address: Present address: ittee machines in on-line learning; this method provides deterministic learning dynamics expressed through a few order parameters, even though ANGD intrinsically holds a large approximated Fisher information matrix. Numerical results obtained using this theory were consistent with those of a simulation, with respect not only to the learning curve but also to the learning failure. Utilizing this method, we numerically evaluated ANGD efficiency and found that ANGD generally performs as well as NGD. We also revealed the key condition affecting the learning plateau in ANGD.

PACS numbers: 02.50.-r, 05.20.-y, 07.05.Mh

I. INTRODUCTION

Feed-forward multilayer perceptrons are known to have difficulty determining their parameters using a set of training data. This is because of the non-linearity of their activation functions, which prevents the use of analytical estimation methods; e.g., maximum likelihood estimation. An alternative approach is to use the stochastic gradient descent, which introduces an error function for a given learning sample in a supervised learning framework and adjusts the network parameters step by step to reduce the error.

Steepest gradient descent (SGD), equivalent to backpropagation, is a simple and useful gradient descent method, but it suffers from a learning plateau, which is a long learning period with poor error reduction. This learning plateau is caused by the permutation symmetry (i.e., an exchange of two hidden units in the same layer has no effect on the network output) because it creates a saddle structure in the generalization error function [1, 2]. Moreover, the plateau period is prolonged when the weight vectors of hidden units in the teacher network are correlated [3].

In contrast, previous works have shown that natural gradient descent (NGD) [4, 5] has almost optimal learning performance (Fisher efficiency). NGD does not have any plateau if the learning rate is set low enough [6]. NGD is also unaffected by the correlation between the teacher weight vectors [3]. From a general view, one of the greatest advantages of NGD could be its independence of the parameterization of a given network model. Another advantage may be that the premultiplier of the gradient of the error – the inverse of the Fisher information matrix – is not dependent on learning data or the error function and is necessarily positive definite. There are similar methods that use the inverse of the Hessian matrix as the premultiplier. However, these methods may be unstable, because their premultiplier intrinsically depends on learning data and is not necessarily positive definite [7].

Adaptive natural gradient descent (ANGD) [8, 9] is an attractive form of NGD with respect to both the calculation cost and the input distribution of training data. ANGD reduces the calculation cost from a cubic order of the number of network parameters to a square order, and does not need to know the input distribution. Moreover, ANGD retains some of the advantages of NGD; e.g., its approximated inverse of the Fisher information matrix is necessarily positive definite. However, no previous research has quantified the learning performance of ANGD with respect to its practical applicability.

In this paper, we evaluate the learning efficiency of a simplified version of ANGD for soft committee machines. (Soft committee machines consist of simplified two-layer perceptrons.) ANGD is intrinsically elaborated for on-line learning. On-line learning [7, 10] facilitates analysis because it uses each learning sample only once, so the network state is independent of each learning sample. We employed statistical-mechanical techniques which extract order parameters and make the stochastic learning dynamics converge towards deterministic at the large limit of the input dimension N [1, 2].

II. MODEL

We define teacher and student network models, stochastic gradient learning rules of the student parameters (SGD, NGD,

*Present address: Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226-502, Japan; Also at "Intelligent Cooperation and Control", PRESTO, JST and Department of Otolaryngology, Head and Neck Surgery, Graduate School of Medicine, Kyoto University; Electronic address: minoue@brain.riken.jp
†Also at Department of Computer Science, College of Natural Science, Kyungpook National University, Daegu, 702-701, Korea; Electronic address: hypark@knu.ac.kr
‡Also at "Intelligent Cooperation and Control", PRESTO, JST; Electronic address: okada@brain.riken.jp

and ANGD), and an adaptive estimation of the inverse of the Fisher information matrix for ANGD. We use a soft committee machine with M hidden units as a teacher network and one with K hidden units as a student network:

$$f_{\boldsymbol{B}}(\boldsymbol{\xi}) \equiv \sum_{k=1}^{M} g(\boldsymbol{B}_k^T \boldsymbol{\xi}), \tag{1}$$

$$f_{\boldsymbol{J}}(\boldsymbol{\xi}) \equiv \sum_{k=1}^{K} g(\boldsymbol{J}_k^T \boldsymbol{\xi}), \tag{2}$$

where $\bullet^T$ denotes the transpose, $\boldsymbol{B}_k \in \mathbb{R}^N$ and $\boldsymbol{J}_k \in \mathbb{R}^N$ are column vectors that represent the kth weight vectors for N-dimensional input $\boldsymbol{\xi} \in \mathbb{R}^N$, and g denotes the activation function. We define the probability density function for each network for NGD and ANGD. The function for the teacher network is defined with input $\boldsymbol{\xi}$ and output ζ as

$$p_{\boldsymbol{B}}(\boldsymbol{\xi}, \zeta) \equiv p(\boldsymbol{\xi}) \delta(\zeta - f_{\boldsymbol{B}}(\boldsymbol{\xi})), \tag{3}$$

where $p(\boldsymbol{\xi})$ is the input distribution and δ is the Dirac delta function. The probability density function for the student network is defined with input $\boldsymbol{\xi}$ and output ζ' using normal distribution $\mathcal{N}(f_{\boldsymbol{J}}(\boldsymbol{\xi}), 1)$:

$$p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta') \equiv p(\boldsymbol{\xi}) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\{\zeta' - f_{\boldsymbol{J}}(\boldsymbol{\xi})\}^2\right). \tag{4}$$

The student distribution is modified incrementally by adjusting its parameter vector $\boldsymbol{J} \equiv [\boldsymbol{J}_1^T, \boldsymbol{J}_2^T, .., \boldsymbol{J}_K^T]^T \in \mathbb{R}^{KN}$ to approximate the teacher distribution.

The error of student output from teacher output is defined as

$$\epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}) \equiv \frac{1}{2}\{\check{\epsilon}_{\boldsymbol{J}}(\boldsymbol{\xi})\}^2, \tag{5}$$

where

$$\check{\epsilon}_{\boldsymbol{J}}(\boldsymbol{\xi}) \equiv f_{\boldsymbol{J}}(\boldsymbol{\xi}) - f_{\boldsymbol{B}}(\boldsymbol{\xi}). \tag{6}$$

The generalization error is also defined as the expected error:

$$\epsilon_g(\boldsymbol{J}) \equiv \langle \epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}}, \tag{7}$$

where $\langle \bullet \rangle_\circ$ denotes the expectation of $\bullet$ with respect to a random variable $\circ$.

The parameter update rule in gradient descent can be written, in general, as $\boldsymbol{J} := \boldsymbol{J} + \Delta \boldsymbol{J}$, where

$$\Delta \boldsymbol{J} = -\frac{\eta}{N} \boldsymbol{M} \nabla_{\boldsymbol{J}} \epsilon_{\boldsymbol{J}}(\boldsymbol{\xi}) = -\frac{\eta}{N} \check{\epsilon}_{\boldsymbol{J}}(\boldsymbol{\xi}) \boldsymbol{M} \nabla f, \tag{8}$$

while $\boldsymbol{M} \in \mathbb{R}^{NK \times NK}$, $\frac{\eta}{N} > 0$ is a scaled update rate, ∇ denotes the gradient, and $\nabla f \in \mathbb{R}^{NK}$ denotes $\nabla_{\boldsymbol{J}} f_{\boldsymbol{J}}(\boldsymbol{\xi})$. We can implement SGD by setting $\boldsymbol{M} = \boldsymbol{I}$ (unit matrix), NGD by setting $\boldsymbol{M} = \boldsymbol{G}^{-1}$ [4], and ANGD by setting $\boldsymbol{M} = \hat{\boldsymbol{G}}^{-1}$ [8, 9]. The $\boldsymbol{G}$ denotes the Fisher information matrix of parameter vector $\boldsymbol{J}$ defined as

$$\begin{aligned} \boldsymbol{G} &\equiv \left\langle [\nabla_{\boldsymbol{J}} \ln p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta')][\nabla_{\boldsymbol{J}} \ln p_{\boldsymbol{J}}(\boldsymbol{\xi}, \zeta')]^T \right\rangle_{\{\boldsymbol{\xi}, \zeta'\}} \\ &= \left\langle [\nabla f][\nabla f]^T \right\rangle_{\boldsymbol{\xi}}, \end{aligned} \tag{9}$$

while $\hat{\boldsymbol{G}}$ is an adaptively approximated matrix of $\boldsymbol{G}$ obtained by ANGD. The update rule of $\hat{\boldsymbol{G}}$ is given as $\hat{\boldsymbol{G}} := \hat{\boldsymbol{G}} + \Delta \hat{\boldsymbol{G}}$, where

$$\Delta \hat{\boldsymbol{G}} \equiv \frac{\rho}{N} \left[-\hat{\boldsymbol{G}} + [\nabla f][\nabla f]^T \right], \tag{10}$$

while $0 < \frac{\rho}{N} < 1$ is an update rate. ANGD does not use the input distribution $p(\boldsymbol{\xi})$, but shares each input sample $\boldsymbol{\xi}$ with the update rule of $\boldsymbol{J}$, and approximates $\boldsymbol{G}$ step by step. Realistically, rather than Eq. (10), ANGD adopts an exactly equivalent rule using the Sherman-Morrison formula: $\hat{\boldsymbol{G}}^{-1} := \hat{\boldsymbol{G}}^{-1} + \Delta \hat{\boldsymbol{G}}^{-1}$, where

$$\Delta \hat{\boldsymbol{G}}^{-1} = \frac{\frac{\rho}{N}}{1 - \frac{\rho}{N}} \left[\hat{\boldsymbol{G}}^{-1} - \frac{\hat{\boldsymbol{G}}^{-1}[\nabla f][\nabla f]^T \hat{\boldsymbol{G}}^{-1}}{1 + \frac{\rho}{N}\{[\nabla f]^T \hat{\boldsymbol{G}}^{-1}[\nabla f] - 1\}} \right]. \tag{11}$$

Equation (11) offers the great advantage that we can omit the expensive matrix inversion ($O(\{NK\}^3)$) and achieve lower calculation cost $O(\{NK\}^2)$. Here, $\hat{\boldsymbol{G}}^{-1}$ is always positive definite and symmetric if the initial value is positive definite and symmetric. For an initial value, we choose $\boldsymbol{I}$ for simplicity. ANGD is always applicable if SGD is applicable, because ∇f is required even in the case of SGD, whereas NGD cannot always be applied when the input distribution is unknown.

When $\rho \ll 1$, Eq. (11) can be reduced to a simple form:

$$\Delta \hat{\boldsymbol{G}}^{-1} = \frac{\rho}{N} \left[\hat{\boldsymbol{G}}^{-1} - \hat{\boldsymbol{G}}^{-1}[\nabla f][\nabla f]^T \hat{\boldsymbol{G}}^{-1} \right]. \tag{12}$$

This approximation was introduced by Amari et al [8]. In this paper, we investigate this simplified version of ANGD in detail under the assumption of small ρ. We also elucidate what happens when this assumption is violated.

III. THEORETICAL RESULTS

In this section, we show the order parameter expression of the system dynamics in ANGD, where we use both the usual and newly introduced order parameters. With respect to SGD and NGD, the usual order parameters are sufficient to explain the system state because the system has rotation invariance under the assumption of Gaussian input ($\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$); i.e., the system is equivalent to one with rotated weight vectors $\boldsymbol{J}_i, \boldsymbol{B}_i$. The usual order parameters can also describe the Fisher information matrix $\boldsymbol{G}$ and $\boldsymbol{G}^{-1}$. We need new order parameters for ANGD, though, to describe the approximated inverse of the Fisher information matrix $\hat{\boldsymbol{G}}^{-1}$. To make the present paper self-contained, we first briefly summarize the derivation of the usual order parameter equations of the soft committee machines for SGD [1, 2] and NGD [5, 6]. Then, we explain the new order parameter expression for ANGD.

A. Generalization error

The usual order parameters are the inner products among all weight vectors:

$$\begin{aligned} &\boldsymbol{Q} \equiv [Q_{ij}]_{i,j=1,..,K}, && Q_{ij} \equiv \boldsymbol{J}_i^T \boldsymbol{J}_j, \\ &\boldsymbol{R} \equiv [R_{ij}]_{i=1,..,K,j=1,..,M}, && R_{ij} \equiv \boldsymbol{J}_i^T \boldsymbol{B}_j, \\ &\boldsymbol{T} \equiv [T_{ij}]_{i,j=1,..,M}, && T_{ij} \equiv \boldsymbol{B}_i^T \boldsymbol{B}_j. \end{aligned} \tag{13}$$

Here, $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$ means the inner products matrix for the student weight vectors, while $\boldsymbol{R} \in \mathbb{R}^{K\times M}$ means the matrix containing the inner products between the student and teacher weight vectors. The $\boldsymbol{Q}$ and $\boldsymbol{R}$ are updated according to the updating of $\boldsymbol{J}$. Here, $\boldsymbol{T} \in \mathbb{R}^{M\times M}$ means the inner products among the teacher weight vectors; this is fixed. The square length of each input, and the inner products between the input and the weight vectors are temporarily used to describe the micro dynamics:

$$\begin{aligned} \chi &\equiv \boldsymbol{\xi}^T\boldsymbol{\xi}, \\ \boldsymbol{x} \equiv [x_i]_{i=1,..,K},\ x_i &\equiv \boldsymbol{J}_i^T\boldsymbol{\xi}, \\ \boldsymbol{y} \equiv [y_i]_{i=1,..,M},\ y_i &\equiv \boldsymbol{B}_i^T\boldsymbol{\xi}. \end{aligned} \tag{14}$$

Here, $\chi \in \mathbb{R}$ stochastically converges to N ($\chi \xrightarrow{P} N$) at the large limit of N. Also, $\boldsymbol{x} \in \mathbb{R}^K$ and $\boldsymbol{y} \in \mathbb{R}^M$ are random vectors dependent on input $\boldsymbol{\xi}$. The distribution of $\boldsymbol{z} \equiv [\boldsymbol{x}^T, \boldsymbol{y}^T]^T \in \mathbb{R}^{K+M}$ is determined using the order parameters: $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$ where

$$\boldsymbol{C} \equiv \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{R}^T & \boldsymbol{T} \end{bmatrix} \in \mathbb{R}^{\{K+M\}\times\{K+M\}}. \tag{15}$$

We can then substitute the order parameters and these random variables for all N-dimensional vectors $\boldsymbol{B}_i$, $\boldsymbol{J}_i$, and $\boldsymbol{\xi}$. The number of order parameters is sufficiently small because it does not depend on N, and this facilitates system analysis. For example, the generalization error can be expressed as

$$\epsilon_g(\boldsymbol{C}) = \frac{1}{2}\int \mathrm{d}\boldsymbol{z} p(\boldsymbol{z}) \left\{ \sum_{k=1}^{K+M} c_k g(z_k) \right\}^2, \tag{16}$$

where c_k is 1 if $k \le K$ or -1 if $k > K$. Here and hereafter, we assume $g(x) = \mathrm{erf}(x/\sqrt{2})$, where $\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{d}t\, e^{-t^2}$ is the conventional error function. The generalization error can then be rewritten as an analytical function [1],

$$\epsilon_g(\boldsymbol{C}) = \frac{1}{\pi}\sum_{i,j=1}^{K+M} c_i c_j \arcsin\frac{C_{ij}}{\sqrt{\{C_{ii}+1\}\{C_{jj}+1\}}}. \tag{17}$$

B. Steepest Gradient Descent

The dynamics of the order parameters for SGD can be expressed using the order parameters themselves [1, 2]. From Eq. (8), the update rule for parameter $\boldsymbol{J}$ is

$$\Delta\boldsymbol{J}_i = -\frac{\eta}{N}\delta_i\boldsymbol{\xi}, \tag{18}$$

where

$$\boldsymbol{\delta} \equiv [\delta_i]_{i=1,..,K},\ \delta_i \equiv \check{\epsilon}_{\boldsymbol{J}}(\boldsymbol{z})g'(x_i), \tag{19}$$

while g' is the derivative of g. Thus, the update rules of the order parameters can be written as

$$\begin{aligned} \Delta R_{ij} &= [\Delta\boldsymbol{J}_i]^T\boldsymbol{B}_j = -\frac{\eta}{N}\delta_i y_j, \\ \Delta Q_{ij} &= [\boldsymbol{J}_i + \Delta\boldsymbol{J}_i]^T[\boldsymbol{J}_j + \Delta\boldsymbol{J}_j] - \boldsymbol{J}_i^T\boldsymbol{J}_j \\ &= \Delta S_{ij} + \Delta S_{ji} + \Delta\phi_{ij}, \end{aligned} \tag{20}$$

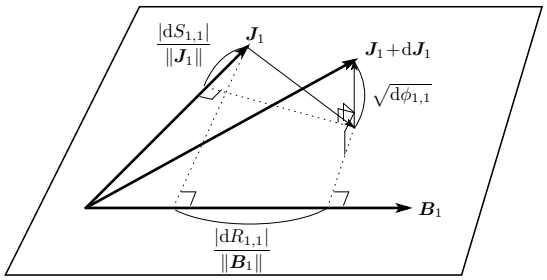

FIG. 1: Intuitive schema of the learning dynamics at the large limit of N. In the case of $K = M = 1$, $\boldsymbol{J}_1$ mainly moves in the current subspace made by $\boldsymbol{J}_1$ and $\boldsymbol{B}_1$ according to the gradient of the error, but necessarily moves out of this subspace into the null space, or the complementary subspace, of $\boldsymbol{J}_1$ and $\boldsymbol{B}_1$, because the small fluctuation of $\boldsymbol{J}_1$ in each dimension is summed up resulting in the non-zero term $\phi_{1,1}$ in Eq. (25).

where

$$\begin{aligned} \Delta S_{ij} &\equiv [\Delta\boldsymbol{J}_i]^T\boldsymbol{J}_j = -\frac{\eta}{N}\delta_i x_j, \\ \Delta\phi_{ij} &\equiv [\Delta\boldsymbol{J}_i]^T\Delta\boldsymbol{J}_j = \frac{\eta^2\chi}{N^2}\delta_i\delta_j. \end{aligned} \tag{21}$$

Here and hereafter, we use $\boldsymbol{S} \equiv [S_{ij}]_{i,j=1,..,K} \in \mathbb{R}^{K\times K}$ and $\boldsymbol{\phi} \equiv [\phi_{ij}]_{i,j=1,..,K} \in \mathbb{R}^{K\times K}$ for convenience.

Next, we introduce time α, and specify that time $1/N$ is consumed by each update. At the large limit of N, the dynamics of the order parameters become continuous and deterministic. For example, the value of R_{ij} over an infinitesimal time interval ($\mathrm{d}\alpha$) after current time α is not a random variable, although each ΔR_{ij} is a random variable.

$$\begin{aligned} R_{ij}^{(\alpha+\mathrm{d}\alpha)} &= R_{ij}^{(\alpha)} + \lim_{N\to\infty}\sum_{\mu=0}^{N\mathrm{d}\alpha-1}\Delta R_{ij}^{(\alpha+\mu/N)} \\ &= R_{ij}^{(\alpha)} - \eta\left\langle\delta_i y_j\right\rangle_{\boldsymbol{z}}\mathrm{d}\alpha. \end{aligned} \tag{22}$$

Therefore, the time derivation of R_{ij} is

$$\frac{\mathrm{d}R_{ij}^{(\alpha)}}{\mathrm{d}\alpha} = -\eta\left\langle\delta_i y_j\right\rangle_{\boldsymbol{z}}. \tag{23}$$

This expectation with respect to $\boldsymbol{z}$ can be solved analytically [1]. The dynamics of Q_{ij} can be determined similarly, and we get

$$\frac{\mathrm{d}Q_{ij}}{\mathrm{d}\alpha} = \frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha} + \frac{\mathrm{d}S_{ji}}{\mathrm{d}\alpha} + \frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha}, \tag{24}$$

where

$$\begin{aligned} \frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha} &= -\eta\left\langle\delta_i x_j\right\rangle_{\boldsymbol{z}}, \\ \frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha} &= \eta^2\left\langle\delta_i\delta_j\right\rangle_{\boldsymbol{z}}. \end{aligned} \tag{25}$$

The order parameter dynamics suggest that, at the large limit of N, $\boldsymbol{J}_i$ necessarily moves out of the current direct sum

subspace made by all the weight vectors $\boldsymbol{J}_1, .., \boldsymbol{J}_K, \boldsymbol{B}_1, .., \boldsymbol{B}_M$, although it mainly moves toward the subspace made by $\boldsymbol{B}_1, .., \boldsymbol{B}_M$ (Fig. 1). The direction of this orthogonal movement to that subspace is always chosen randomly, while the time derivative of the square distance of this movement is represented as $\frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha}$ in Eq. (24). This orthogonal movement also appears in NGD and ANGD as the $\frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha}$ term included in the dynamics of Q_{ij}, and, moreover, in the $\hat{\boldsymbol{G}}^{-1}$ dynamics for ANGD.

C. Natural Gradient Descent

The dynamics for NGD can also be expressed through the order parameters. In this subsection, we first determine the Fisher information matrix, $\boldsymbol{G}$, and its inverse, $\boldsymbol{G}^{-1}$, by using only $\boldsymbol{Q}$. Then we derive the motion equations of the order parameters.

Generally, for any $\boldsymbol{G}_{ij} \in \mathbb{R}^{K\times K}$, the (i, j) block of $\boldsymbol{G}$, can be expressed with both the student weight vectors, $\boldsymbol{J}' \equiv [\boldsymbol{J}_1, .., \boldsymbol{J}_K] \in \mathbb{R}^{N\times K}$, and all the orthonormal bases of the null space of the student weight vectors, $\boldsymbol{V} \in \mathbb{R}^{N\times\{N-K\}}$, which satisfies $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{J}'^T\boldsymbol{V} = \boldsymbol{0}$, as

$$\boldsymbol{G}_{ij} = \lambda_{ij}\boldsymbol{I} + [\boldsymbol{J}', \boldsymbol{V}]\begin{bmatrix}\boldsymbol{\Lambda}_{ij} & \boldsymbol{\Lambda}'_{ij}\\ \boldsymbol{\Lambda}''_{ij} & \boldsymbol{\Lambda}'''_{ij}\end{bmatrix}[\boldsymbol{J}', \boldsymbol{V}]^T, \tag{26}$$

where $\lambda_{ij} \in \mathbb{R}, \boldsymbol{\Lambda}_{ij} \in \mathbb{R}^{K\times K}, \boldsymbol{\Lambda}'_{ij} \in \mathbb{R}^{K\times\{N-K\}}, \boldsymbol{\Lambda}''_{ij} \in \mathbb{R}^{\{N-K\}\times K}$, and $\boldsymbol{\Lambda}'''_{ij} \in \mathbb{R}^{\{N-K\}\times\{N-K\}}$ (to be exact, λ_{ij} is verbose). We can then rewrite Eq. (26) as,

$$\boldsymbol{G}_{ij} = \lambda_{ij}\boldsymbol{I} + \boldsymbol{J}'\boldsymbol{\Lambda}_{ij}\boldsymbol{J}'^T + \boldsymbol{E}_{ij}, \tag{27}$$

$$\boldsymbol{E}_{ij} \equiv [\boldsymbol{J}', \boldsymbol{V}]\begin{bmatrix}\boldsymbol{0} & \boldsymbol{\Lambda}'_{ij}\\ \boldsymbol{\Lambda}''_{ij} & \boldsymbol{\Lambda}'''_{ij}\end{bmatrix}[\boldsymbol{J}', \boldsymbol{V}]^T, \tag{28}$$

where $\boldsymbol{E}_{ij} \in \mathbb{R}^{N\times N}$.

We next prove $\boldsymbol{E}_{ij} = \boldsymbol{0}$; in other words, λ_{ij} and $\boldsymbol{\Lambda}_{ij}$ are sufficient parameters to express $\boldsymbol{G}$. By multiplying the identity matrix, $[\boldsymbol{J}', \boldsymbol{V}]\left[[\boldsymbol{J}', \boldsymbol{V}]^T[\boldsymbol{J}', \boldsymbol{V}]\right]^{-1}[\boldsymbol{J}', \boldsymbol{V}]^T$, we can rewrite $\boldsymbol{G}_{ij}$ as

$$\begin{aligned}
\boldsymbol{G}_{ij} &= \left\langle [\nabla f_i][\nabla f_j]^T \right\rangle_{\boldsymbol{\xi}}\\
&= [\boldsymbol{J}', \boldsymbol{V}]\left[[\boldsymbol{J}', \boldsymbol{V}]^T[\boldsymbol{J}', \boldsymbol{V}]\right]^{-1}[\boldsymbol{J}', \boldsymbol{V}]^T \left\langle [\nabla f_i][\nabla f_j]^T \right\rangle_{\boldsymbol{\xi}} [\boldsymbol{J}', \boldsymbol{V}]\left[[\boldsymbol{J}', \boldsymbol{V}]^T[\boldsymbol{J}', \boldsymbol{V}]\right]^{-1}[\boldsymbol{J}', \boldsymbol{V}]^T\\
&= [\boldsymbol{J}', \boldsymbol{V}]\begin{bmatrix}\boldsymbol{Q}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\left\langle g'(x_i)g'(x_j)\begin{bmatrix}\boldsymbol{x}\boldsymbol{x}^T & \boldsymbol{x}\boldsymbol{v}^T\\ \boldsymbol{v}\boldsymbol{x}^T & \boldsymbol{v}\boldsymbol{v}^T\end{bmatrix}\right\rangle_{\boldsymbol{\xi}}\begin{bmatrix}\boldsymbol{Q}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}[\boldsymbol{J}', \boldsymbol{V}]^T\\
&= [\boldsymbol{J}', \boldsymbol{V}]\begin{bmatrix}\boldsymbol{Q}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\left\langle g'(x_i)g'(x_j)\begin{bmatrix}\boldsymbol{x}\boldsymbol{x}^T & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\right\rangle_{\boldsymbol{x}}\begin{bmatrix}\boldsymbol{Q}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}[\boldsymbol{J}', \boldsymbol{V}]^T\\
&= [\boldsymbol{J}', \boldsymbol{V}]\begin{bmatrix}\boldsymbol{Q}^{-1}\left\langle g'(x_i)g'(x_j)\boldsymbol{x}\boldsymbol{x}^T\right\rangle_{\boldsymbol{x}}\boldsymbol{Q}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \left\langle g'(x_i)g'(x_j)\right\rangle_{\boldsymbol{x}}\boldsymbol{I}\end{bmatrix}[\boldsymbol{J}', \boldsymbol{V}]^T\\
&= \left\langle g'(x_i)g'(x_j)\right\rangle_{\boldsymbol{x}}\boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{J}'\left[\boldsymbol{Q}^{-1}\left\langle g'(x_i)g'(x_j)\boldsymbol{x}\boldsymbol{x}^T\right\rangle_{\boldsymbol{x}}\boldsymbol{Q}^{-1}\right]\boldsymbol{J}'^T\\
&= \left\langle g'(x_i)g'(x_j)\right\rangle_{\boldsymbol{x}}\boldsymbol{I} + \boldsymbol{J}'\left[\boldsymbol{Q}^{-1}\left\langle g'(x_i)g'(x_j)\boldsymbol{x}\boldsymbol{x}^T\right\rangle_{\boldsymbol{x}}\boldsymbol{Q}^{-1} - \left\langle g'(x_i)g'(x_j)\right\rangle_{\boldsymbol{x}}\boldsymbol{Q}^{-1}\right]\boldsymbol{J}'^T,
\end{aligned} \tag{29}$$

where $\nabla f_i \equiv g'(x_i)\boldsymbol{\xi}$, while $\boldsymbol{v} \equiv \boldsymbol{V}^T\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. We used $\left\langle g'(x_i)g'(x_j)\boldsymbol{x}\boldsymbol{v}^T\right\rangle_{\boldsymbol{\xi}} = \left\langle g'(x_i)g'(x_j)\boldsymbol{x}\right\rangle_{\boldsymbol{x}}\left\langle \boldsymbol{v}^T\right\rangle_{\boldsymbol{v}} = \boldsymbol{0}$, and $\left\langle \boldsymbol{v}\boldsymbol{v}^T\right\rangle_{\boldsymbol{\xi}} = \boldsymbol{I}$. Thus, we can prove $\boldsymbol{E}_{ij} = \boldsymbol{0}$ by letting

$$\lambda_{ij} \equiv \left\langle g'(x_i)g'(x_j)\right\rangle_{\boldsymbol{x}} = \frac{2}{\pi}\begin{vmatrix}Q_{ii}+1 & Q_{ij}\\ Q_{ij} & Q_{jj}+1\end{vmatrix}^{-\frac{1}{2}}, \tag{30}$$

$$\begin{aligned}
\boldsymbol{\Lambda}_{ij} &\equiv \boldsymbol{Q}^{-1}\left\langle g'(x_i)g'(x_j)\boldsymbol{x}\boldsymbol{x}^T\right\rangle_{\boldsymbol{x}}\boldsymbol{Q}^{-1} - \lambda_{ij}\boldsymbol{Q}^{-1}\\
&= -\lambda_{ij}[\boldsymbol{e}_i, \boldsymbol{e}_j]\begin{bmatrix}Q_{ii}+1 & Q_{ij}\\ Q_{ij} & Q_{jj}+1\end{bmatrix}^{-1}[\boldsymbol{e}_i, \boldsymbol{e}_j]^T,
\end{aligned} \tag{31}$$

where $|\bullet|$ denotes a determinant, while $\boldsymbol{e}_i \in \mathbb{R}^K$ is a unit vector whose ith element is 1. The λ_{ij} and $\boldsymbol{\Lambda}_{ij}$ are expressed with certainty by $\boldsymbol{Q}$.

Next, we determine $\boldsymbol{G}^{-1}$ using a similar style to Eq. (27):

$$[\boldsymbol{G}^{-1}]_{ij} = \theta_{ij}\boldsymbol{I} + \boldsymbol{J}'\boldsymbol{\Theta}_{ij}\boldsymbol{J}'^T, \tag{32}$$

where $\theta_{ij} \in \mathbb{R}$ and $\boldsymbol{\Theta}_{ij} \in \mathbb{R}^{K\times K}$ correspond to λ_{ij} and $\boldsymbol{\Lambda}_{ij}$, respectively. If we temporarily adopt normalized expressions of $\boldsymbol{\Lambda}_{ij}$ and $\boldsymbol{\Theta}_{ij}$, defined as

$$\tilde{\boldsymbol{\Lambda}}_{ij} \equiv \boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Lambda}_{ij}\boldsymbol{Q}^{\frac{1}{2}} + \lambda_{ij}\boldsymbol{I}, \tag{33}$$

$$\tilde{\boldsymbol{\Theta}}_{ij} \equiv \boldsymbol{Q}^{\frac{1}{2}}\boldsymbol{\Theta}_{ij}\boldsymbol{Q}^{\frac{1}{2}} + \theta_{ij}\boldsymbol{I}, \tag{34}$$

matrix multiplication will be simplified; e.g.,

$$\boldsymbol{G}_{ij}\boldsymbol{G}_{kl} = \left[\lambda_{ij}[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{Q}^{-1}\boldsymbol{U}^T] + \boldsymbol{U}\boldsymbol{Q}^{-\frac{1}{2}}\tilde{\boldsymbol{\Lambda}}_{ij}\boldsymbol{Q}^{-\frac{1}{2}}\boldsymbol{U}^T\right]$$

$$\times\left[\lambda_{kl}[\boldsymbol{I}-\boldsymbol{U}\boldsymbol{Q}^{-1}\boldsymbol{U}^T]+\boldsymbol{U}\boldsymbol{Q}^{-\frac{1}{2}}\tilde{\boldsymbol{\Lambda}}_{kl}\boldsymbol{Q}^{-\frac{1}{2}}\boldsymbol{U}^T\right]$$
$$=\lambda_{ij}\lambda_{kl}[\boldsymbol{I}-\boldsymbol{U}\boldsymbol{Q}^{-1}\boldsymbol{U}^T]+\boldsymbol{U}\boldsymbol{Q}^{-\frac{1}{2}}\tilde{\boldsymbol{\Lambda}}_{ij}\tilde{\boldsymbol{\Lambda}}_{kl}\boldsymbol{Q}^{-\frac{1}{2}}\boldsymbol{U}^T. \tag{35}$$

Utilizing this normalization, we can easily obtain

$$\boldsymbol{\theta}=\boldsymbol{\lambda}^{-1},\ \tilde{\boldsymbol{\Theta}}=\tilde{\boldsymbol{\Lambda}}^{-1}, \tag{36}$$

where $\boldsymbol{\lambda}\equiv[\lambda_{ij}]_{i,j=1,\dots,K}$ and $\boldsymbol{\theta}\equiv[\theta_{ij}]_{i,j=1,\dots,K}$ are $\mathbb{R}^{K\times K}$ symmetric matrices, while $\boldsymbol{\Lambda}\equiv[\boldsymbol{\Lambda}_{ij}]_{i,j=1,\dots,K}$, $\tilde{\boldsymbol{\Lambda}}\equiv[\tilde{\boldsymbol{\Lambda}}_{ij}]_{i,j=1,\dots,K}$, $\boldsymbol{\Theta}\equiv[\boldsymbol{\Theta}_{ij}]_{i,j=1,\dots,K}$, and $\tilde{\boldsymbol{\Theta}}\equiv[\tilde{\boldsymbol{\Theta}}_{ij}]_{i,j=1,\dots,K}$ are $\mathbb{R}^{K^2\times K^2}$ symmetric matrices.

Finally, we obtain the order parameter dynamics for NGD. In NGD, the rule for updating $\boldsymbol{J}$ is given by

$$\Delta\boldsymbol{J}_i=-\frac{\eta}{N}\sum_{k=1}^{K}\delta_k\boldsymbol{G}_{ik}^{-1}\boldsymbol{\xi}. \tag{37}$$

Hence, we obtain the time derivatives of the order parameters in a manner similar to that for SGD:

$$\begin{aligned}
\frac{\mathrm{d}R_{ij}}{\mathrm{d}\alpha}&=-\eta\sum_{k=1}^{K}\left[\theta_{ik}\left\langle\delta_k x_j\right\rangle_{\boldsymbol{z}}+\left\langle\delta_k\boldsymbol{x}^T\right\rangle_{\boldsymbol{z}}\boldsymbol{\Theta}_{ik}^T\boldsymbol{R}_{\bullet j}\right],\\
\frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha}&=-\eta\sum_{k=1}^{K}\left[\theta_{ik}\left\langle\delta_k x_j\right\rangle_{\boldsymbol{z}}+\left\langle\delta_k\boldsymbol{x}^T\right\rangle_{\boldsymbol{z}}\boldsymbol{\Theta}_{ik}^T\boldsymbol{Q}_{\bullet j}\right],\\
\frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha}&=\eta^2\sum_{k,l=1}^{K}\theta_{ik}\left\langle\delta_k\delta_l\right\rangle_{\boldsymbol{z}}\theta_{lj},\\
\frac{\mathrm{d}Q_{ij}}{\mathrm{d}\alpha}&=\frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha}+\frac{\mathrm{d}S_{ji}}{\mathrm{d}\alpha}+\frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha},
\end{aligned} \tag{38}$$

where η and ρ are $O(1)$ with respect to N, $\rho'\equiv\frac{\rho/N}{1-\rho/N}$, and $\tau\equiv\alpha+\mu/N$, while $\nabla f_{(\bullet)}^{(\circ)}$ denotes $\nabla_{\boldsymbol{J}}f_{\boldsymbol{J}}(\boldsymbol{\xi}^{(\circ)})|_{\boldsymbol{J}=\boldsymbol{J}^{(\bullet)}}$. These two equations show that both $\boldsymbol{J}$ and $\boldsymbol{H}$ include a common random vector: $\nabla f_{(\tau)}^{(\tau)}$. Therefore, $\boldsymbol{J}$ and $\boldsymbol{H}$ become dependent on each other. To negate this dependence, we introduce a new update rule – we draw two $\boldsymbol{\xi}$ independently, one for the $\boldsymbol{J}$ update and the other for the $\boldsymbol{H}$ update in each learning step (*two $\boldsymbol{\xi}$ rule*) – so that this dependence disappears. Under this *two $\boldsymbol{\xi}$ rule*, we can fix $\boldsymbol{J}$ during $\mathrm{d}\alpha$, and reduce Eq. (40) to

$$\boldsymbol{H}^{(\alpha+\mathrm{d}\alpha)}=\{1+\rho'\}^{N\mathrm{d}\alpha}\boldsymbol{H}^{(\alpha)}$$

where $\boldsymbol{R}_{\bullet j}$ denotes the jth column of $\boldsymbol{R}$, and so on [6].

D. Dynamics of $\hat{\boldsymbol{G}}^{-1}$ for ANGD

In this subsection, we derive the dynamics of the approximated inverse of the Fisher information matrix $\hat{\boldsymbol{G}}^{-1}$ (from here on, for simplicity we write $\boldsymbol{H}$ instead of $\hat{\boldsymbol{G}}^{-1}$) in the simplified version of ANGD. Unlike NGD, there are three types of difficulty in ANGD. 1) The dependence between $\boldsymbol{J}$ and $\boldsymbol{H}$, because ANGD has two dynamics of $\boldsymbol{J}$ and $\boldsymbol{H}$ and they share each input $\boldsymbol{\xi}$. We introduce an approximation (*two $\boldsymbol{\xi}$ rule*) to negate this dependence. 2) Higher-order self-correlations of $\boldsymbol{H}$, which originate from the update rule of $\boldsymbol{H}$ itself. We negate these infinite correlations by exploiting the $\rho\ll 1$ assumption of the simplified ANGD, because n-order self-correlation is scaled by $O(\rho^n)$. 3) The high complexity of $\boldsymbol{H}$; $\boldsymbol{G}$ and $\boldsymbol{G}^{-1}$ are sufficiently characterized by the subspace of the student weight vectors, whereas $\boldsymbol{H}$ is not. This complexity will be managed in the next subsection.

First, we introduce *two $\boldsymbol{\xi}$ rule*. From Eqs. (8) and (11), the value of $\boldsymbol{J}$ and $\boldsymbol{H}$ for an infinitesimal time period ($\mathrm{d}\alpha$) after current time α are

$$\begin{aligned}
\boldsymbol{J}^{(\alpha+\mathrm{d}\alpha)}&=\boldsymbol{J}^{(\alpha)}+\sum_{\mu=0}^{N\mathrm{d}\alpha-1}\Delta\boldsymbol{J}^{(\tau)}\\
&=\boldsymbol{J}^{(\alpha)}-\frac{\eta}{N}\sum_{\mu=0}^{N\mathrm{d}\alpha-1}\check{\boldsymbol{\epsilon}}_{\boldsymbol{J}^{(\tau)}}(\boldsymbol{\xi}^{(\tau)})\boldsymbol{H}^{(\tau)}\nabla f_{(\tau)}^{(\tau)},
\end{aligned} \tag{39}$$

$$\begin{aligned}
\boldsymbol{H}^{(\alpha+\mathrm{d}\alpha)}&=\boldsymbol{H}^{(\alpha)}+\sum_{\mu=0}^{N\mathrm{d}\alpha-1}\Delta\boldsymbol{H}^{(\tau)}\\
&=\boldsymbol{H}^{(\alpha)}+\sum_{\mu=0}^{N\mathrm{d}\alpha-1}\rho'\left[\boldsymbol{H}^{(\tau)}-\boldsymbol{H}^{(\tau)}[\nabla f_{(\tau)}^{(\tau)}][\nabla f_{(\tau)}^{(\tau)}]^T\boldsymbol{H}^{(\tau)}\right]\\
&=\{1+\rho'\}^{N\mathrm{d}\alpha}\boldsymbol{H}^{(\alpha)}-\sum_{\mu=0}^{N\mathrm{d}\alpha-1}\{1+\rho'\}^{N\mathrm{d}\alpha-1-\mu}\rho'\boldsymbol{H}^{(\tau)}[\nabla f_{(\tau)}^{(\tau)}][\nabla f_{(\tau)}^{(\tau)}]^T\boldsymbol{H}^{(\tau)},
\end{aligned} \tag{40}$$

$$-\sum_{\mu=0}^{N\mathrm{d}\alpha-1}\{1+\rho'\}^{N\mathrm{d}\alpha-1-\mu}\boldsymbol{H}^{(\tau)}\boldsymbol{F}^{(\tau)}\boldsymbol{H}^{(\tau)}, \tag{41}$$

where $\boldsymbol{F}^{(\tau)}\equiv\rho'[\nabla f_{(\alpha)}^{(\tau)}][\nabla f_{(\alpha)}^{(\tau)}]^T$. This *two $\boldsymbol{\xi}$ rule* will be validated in Section IV.

Next, we negate the higher-order self-correlations of $\boldsymbol{H}$. We notice that this Eq. (41) is still difficult to solve, because it includes highly self-correlated terms with respect to the old random matrix $\boldsymbol{F}^{(\tau')}, (\alpha\le\tau'<\tau)$. For example, the most

correlated term is

$$\boldsymbol{H}^{(\alpha)}\boldsymbol{F}^{(\alpha)}\boldsymbol{H}^{(\alpha)}\boldsymbol{F}^{(\alpha+\frac{1}{N})}\boldsymbol{H}^{(\alpha)}\boldsymbol{F}^{(\alpha)}\boldsymbol{H}^{(\alpha)}\boldsymbol{F}^{(\alpha+\frac{2}{N})}... \tag{42}$$

However, we can ignore these terms including n-order self-correlation ($n \geq 2$), because they are at most $O(\rho^n)$ and small enough under the assumption of $\rho \ll 1$ for the simplified ANGD. Therefore, we sum up only the $O(\rho^0)$ and $O(\rho^1)$ terms:

$$\begin{aligned}\boldsymbol{H}^{(\alpha+\mathrm{d}\alpha)} = \{1+\rho'\}^{N\mathrm{d}\alpha}\boldsymbol{H}^{(\alpha)} + \sum_{\mu=0}^{N\mathrm{d}\alpha-1}\{1+\rho'\}^{N\mathrm{d}\alpha-1-\mu} \\ \times\Big[\{1+\rho'\}^{\mu}\boldsymbol{H}^{(\alpha)}\Big]\boldsymbol{F}^{(\tau)}\Big[\{1+\rho'\}^{\mu}\boldsymbol{H}^{(\alpha)}\Big].\end{aligned} \tag{43}$$

At the large limit of N, Eq. (43) becomes

$$\boldsymbol{H}^{(\alpha+\mathrm{d}\alpha)} = e^{\rho\mathrm{d}\alpha}\Big[\boldsymbol{H}^{(\alpha)} + \{1-e^{\rho\mathrm{d}\alpha}\}\boldsymbol{H}^{(\alpha)}\boldsymbol{G}\boldsymbol{H}^{(\alpha)}\Big], \tag{44}$$

where we use $\sum \boldsymbol{F} \rightarrow \boldsymbol{G}$ (see Appendix A for further discussion about the convergence of $\boldsymbol{F}$). Then, we obtain the dynamics of $\boldsymbol{H}$:

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{H}}{\mathrm{d}\alpha} &= \lim_{\mathrm{d}\alpha\to 0}\frac{\boldsymbol{H}^{(\alpha+\mathrm{d}\alpha)} - \boldsymbol{H}^{(\alpha)}}{\mathrm{d}\alpha} \\ &= \rho\left[\boldsymbol{H} - \boldsymbol{H}\boldsymbol{G}\boldsymbol{H}\right].\end{aligned} \tag{45}$$

We also obtain the usual order parameter dynamics. As the update equation of each $\boldsymbol{J}_i$ for ANGD is

$$\Delta\boldsymbol{J}_i = -\frac{\eta}{N}\sum_{k=1}^{K}\delta_k\boldsymbol{H}_{ik}\boldsymbol{\xi}, \tag{46}$$

we can easily get

$$\begin{aligned}\frac{\mathrm{d}R_{ij}}{\mathrm{d}\alpha} &= -\eta\sum_{k=1}^{K}\left\langle\delta_k\boldsymbol{\xi}^T\boldsymbol{H}_{ik}^T\boldsymbol{B}_j\right\rangle_{\boldsymbol{\xi}}, \\ \frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha} &= -\eta\sum_{k=1}^{K}\left\langle\delta_k\boldsymbol{\xi}^T\boldsymbol{H}_{ik}^T\boldsymbol{J}_j\right\rangle_{\boldsymbol{\xi}}, \\ \frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha} &= \frac{\eta^2}{N}\sum_{k,l=1}^{K}\left\langle\delta_k\delta_l\boldsymbol{\xi}^T\boldsymbol{H}_{ik}^T\boldsymbol{H}_{jl}\boldsymbol{\xi}\right\rangle_{\boldsymbol{\xi}}, \\ \frac{\mathrm{d}Q_{ij}}{\mathrm{d}\alpha} &= \frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha} + \frac{\mathrm{d}S_{ji}}{\mathrm{d}\alpha} + \frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha}.\end{aligned} \tag{47}$$

These order parameter dynamics still include N-dimensional vectors and $N\times N$ matrices, but we obtain the order parameter dynamics expressed by the order parameters themselves in the next subsection.

E. Order parameter representation for $\boldsymbol{H} \equiv \hat{\boldsymbol{G}}^{-1}$

In this subsection, we extract the new order parameters from $\boldsymbol{H}$. To characterize $\boldsymbol{H}$, we should consider the history of $\boldsymbol{J}$, which includes two types of movement. 1) Approach to the corresponding teacher vectors. To deal with this movement, we use not only the student weight vectors but also the teacher weight vectors to express $\boldsymbol{H}$, although $\boldsymbol{G}$ and $\boldsymbol{G}^{-1}$ are sufficiently expressed by only the current student weight vectors. 2) Escape from the subspace made by all the weight vectors. The direction of this movement is random, although its speed is deterministic as specified by the $\mathrm{d}\phi$ term. The order parameter expressions of SGD and NGD do not suffer from this randomness because they discard it by exploiting the rotation invariance of the system. In ANGD, however, this randomness upsets the order parameter expression of $\boldsymbol{H}$; it produces a hysteresis component of $\boldsymbol{H}$ or the residual fluctuating term ($\boldsymbol{E}$ as introduced below) when $\boldsymbol{H}$ is expressed through order parameters. This fluctuation term is negligible at the large limit of N, but its square is not. Here, we realistically consider the $\boldsymbol{H}$ dynamics by taking the second powers of this fluctuation term into account, and obtain an effective order parameter expression of $\boldsymbol{H}$. This theoretical result will be numerically validated in the next section.

First, we express $\boldsymbol{H}$ in a similar manner to $\boldsymbol{G}$ in Eq. (27) as

$$\boldsymbol{H}_{ij} = \omega_{ij}\boldsymbol{I} + [\boldsymbol{U},\boldsymbol{V}]\begin{bmatrix}\boldsymbol{\Omega}_{ij} & \boldsymbol{\Omega}'_{ij}\\ \boldsymbol{\Omega}''_{ij} & \boldsymbol{\Omega}'''_{ij}\end{bmatrix}[\boldsymbol{U},\boldsymbol{V}]^T \tag{48}$$

$$= \omega_{ij}\boldsymbol{I} + \boldsymbol{U}\boldsymbol{\Omega}_{ij}\boldsymbol{U}^T + \boldsymbol{E}_{ij}, \tag{49}$$

$$\boldsymbol{E}_{ij} \equiv [\boldsymbol{U},\boldsymbol{V}]\begin{bmatrix}\boldsymbol{0} & \boldsymbol{\Omega}'_{ij}\\ \boldsymbol{\Omega}''_{ij} & \boldsymbol{\Omega}'''_{ij}\end{bmatrix}[\boldsymbol{U},\boldsymbol{V}]^T, \tag{50}$$

and prove $\boldsymbol{E}_{ij}$ is negligible. Note that we use not only the student weight vectors, but also the teacher weight vectors; i.e., we use $\boldsymbol{U} \equiv [\boldsymbol{J}_1,..,\boldsymbol{J}_K,\boldsymbol{B}_1,..,\boldsymbol{B}_M] \in \mathbb{R}^{N\times\{K+M\}}$ instead of $\boldsymbol{J}' \equiv [\boldsymbol{J}_1,..,\boldsymbol{J}_K]$. This is because the student vectors move toward the corresponding teacher vectors, and $\boldsymbol{H}$ holds the component made from the old student vectors. Here, $\boldsymbol{V} \in \mathbb{R}^{N\times\{N-K-M\}}$ is re-defined as the orthonormal bases of the null space of $\boldsymbol{U}$ rather than $\boldsymbol{J}'$, which satisfies $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{U}^T\boldsymbol{V} = \boldsymbol{0}$. Here, $\omega_{ij} \in \mathbb{R}$ and $\boldsymbol{\Omega}_{ij} \in \mathbb{R}^{\{K+M\}\times\{K+M\}}$ are the candidates for the new order parameters. Also, $\boldsymbol{\Omega}'_{ij} \in \mathbb{R}^{\{K+M\}\times\{N-K-M\}}$, $\boldsymbol{\Omega}''_{ij} \in \mathbb{R}^{\{N-K-M\}\times\{K+M\}}$, and $\boldsymbol{\Omega}'''_{ij} \in \mathbb{R}^{\{N-K-M\}\times\{N-K-M\}}$ are large matrices. As $\boldsymbol{H} = [\boldsymbol{H}_{ij}]_{i,j=1,..,K}$ is a symmetric matrix, $\boldsymbol{\omega} \equiv [\omega_{ij}]_{i,j=1,..,K}, \boldsymbol{\Omega} \equiv [\boldsymbol{\Omega}_{ij}]_{i,j=1,..,K}$, and $\boldsymbol{\Omega}''' \equiv [\boldsymbol{\Omega}'''_{ij}]_{i,j=1,..,K}$ are also symmetric matrices, while $\boldsymbol{\Omega}' \equiv [\boldsymbol{\Omega}'_{ij}]_{i,j=1,..,K}$ and $\boldsymbol{\Omega}'' \equiv [\boldsymbol{\Omega}''_{ij}]_{i,j=1,..,K}$ are symmetric with respect to each other.

Next, we find appropriate dynamics of $\omega, \boldsymbol{\Omega}, \boldsymbol{\Omega}', \boldsymbol{\Omega}''$, and $\boldsymbol{\Omega}'''$ that satisfy the dynamics of $\boldsymbol{H}$ given by Eq. (45). For convenience, we consider an infinitesimal change of $\boldsymbol{H}_{ij}$ from Eq. (45),

$$\mathrm{d}\boldsymbol{H}_{ij} = \rho[\boldsymbol{H} - \boldsymbol{H}\boldsymbol{G}\boldsymbol{H}]_{ij}\mathrm{d}\alpha \tag{51}$$

By substituting Eq. (48) for $\boldsymbol{H}$, we easily obtain the decomposed form of $\mathrm{d}\boldsymbol{H}_{ij}$ as

$$\mathrm{d}\boldsymbol{H}_{ij} = \gamma_{ij}\boldsymbol{I} + [\boldsymbol{U},\boldsymbol{V}]\begin{bmatrix}\boldsymbol{\Gamma}_{ij} & \boldsymbol{\Gamma}'_{ij}\\ \boldsymbol{\Gamma}''_{ij} & \boldsymbol{\Gamma}'''_{ij}\end{bmatrix}[\boldsymbol{U},\boldsymbol{V}]^T, \tag{52}$$

where

$$\gamma_{ij} = \rho[\boldsymbol{\omega} - \boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij}\mathrm{d}\alpha, \tag{53}$$

$$\boldsymbol{\Gamma}_{ij} = \rho\boldsymbol{C}^{-\frac{1}{2}}\left[\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\omega}} + \tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}} - \tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\Omega}}''\right]_{ij}\boldsymbol{C}^{-\frac{1}{2}}\mathrm{d}\alpha, \tag{54}$$

$$\boldsymbol{\Gamma}'_{ij} = \rho\boldsymbol{C}^{-\frac{1}{2}}\left[\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}} - \tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\lambda}}\boldsymbol{\Omega}'''\right]_{ij}\mathrm{d}\alpha, \tag{55}$$

$$\boldsymbol{\Gamma}''_{ij} = \rho\left[\tilde{\boldsymbol{\Omega}}'' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\Omega}}'' - \boldsymbol{\Omega}'''\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\Omega}}''\right]_{ij}\boldsymbol{C}^{-\frac{1}{2}}\mathrm{d}\alpha, \tag{56}$$

$$\boldsymbol{\Gamma}'''_{ij} = \rho\left[\boldsymbol{\Omega}''' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}\boldsymbol{\Omega}''' - \boldsymbol{\Omega}'''\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}} - \boldsymbol{\Omega}'''\tilde{\boldsymbol{\lambda}}\boldsymbol{\Omega}'''\right]_{ij}\mathrm{d}\alpha, \tag{57}$$

where $\tilde{\boldsymbol{\Omega}} \equiv [\tilde{\boldsymbol{\Omega}}_{ij} \equiv \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{\Omega}_{ij}\boldsymbol{C}^{\frac{1}{2}} + \omega_{ij}\boldsymbol{I}]_{i,j=1,..,K}$, $\tilde{\boldsymbol{\Omega}}' \equiv [\tilde{\boldsymbol{\Omega}}'_{ij} \equiv \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{\Omega}_{ij}]_{i,j=1,..,K}$, and $\tilde{\boldsymbol{\Omega}}'' \equiv [\tilde{\boldsymbol{\Omega}}''_{ij} \equiv \boldsymbol{\Omega}_{ij}\boldsymbol{C}^{\frac{1}{2}}]_{i,j=1,..,K}$ are introduced to simplify multiplications, while, $\tilde{\boldsymbol{\lambda}} = [\lambda_{ij}\boldsymbol{I}]_{i,j=1,..,K}$ and $\tilde{\boldsymbol{\omega}} = [\omega_{ij}\boldsymbol{I}]_{i,j=1,..,K}$ are matrices extended to an appropriate size; e.g., the same size as that of $\boldsymbol{\Omega}$ or $\boldsymbol{\Omega}'''$. The $\tilde{\boldsymbol{\Lambda}}$ is re-defined as $\tilde{\boldsymbol{\Lambda}} \equiv [\tilde{\boldsymbol{\Lambda}}_{ij} \equiv \boldsymbol{C}^{\frac{1}{2}}\begin{bmatrix}\boldsymbol{\Lambda}_{ij} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{C}^{\frac{1}{2}} + \lambda_{ij}\boldsymbol{I}]_{i,j=1,..,K}$.

The infinitesimal change of Eq. (48) is given by

$$\mathrm{d}\boldsymbol{H}_{ij} = \{\mathrm{d}\omega_{ij}\}\boldsymbol{I} + \boldsymbol{W}\begin{bmatrix}\boldsymbol{\Omega}_{ij}+\mathrm{d}\boldsymbol{\Omega}_{ij} & \boldsymbol{\Omega}'_{ij}+\mathrm{d}\boldsymbol{\Omega}'_{ij}\\ \boldsymbol{\Omega}''_{ij}+\mathrm{d}\boldsymbol{\Omega}''_{ij} & \boldsymbol{\Omega}'''_{ij}+\mathrm{d}\boldsymbol{\Omega}'''_{ij}\end{bmatrix}\boldsymbol{W}^T - [\boldsymbol{U},\boldsymbol{V}]\begin{bmatrix}\boldsymbol{\Omega}_{ij} & \boldsymbol{\Omega}'_{ij}\\ \boldsymbol{\Omega}''_{ij} & \boldsymbol{\Omega}'''_{ij}\end{bmatrix}[\boldsymbol{U},\boldsymbol{V}]^T, \tag{58}$$

where $\boldsymbol{W} \equiv [\boldsymbol{U}+\mathrm{d}\boldsymbol{U}, \boldsymbol{V}+\mathrm{d}\boldsymbol{V}] \in \mathbb{R}^{N\times N}$. Note that the conventional total differential is not applicable because $\boldsymbol{U}$ is not differentiable. We also decompose this matrix by $\boldsymbol{W}$ rather than $[\boldsymbol{U},\boldsymbol{V}]$ because $\boldsymbol{U}$ and $\boldsymbol{V}$ are also moving,

$$\mathrm{d}\boldsymbol{H}_{ij} = l_{ij}\boldsymbol{I} + \boldsymbol{W}\begin{bmatrix}\boldsymbol{L}_{ij} & \boldsymbol{L}'_{ij}\\ \boldsymbol{L}''_{ij} & \boldsymbol{L}'''_{ij}\end{bmatrix}\boldsymbol{W}^T. \tag{59}$$

Here, $\boldsymbol{L}_{ij}, \boldsymbol{L}'_{ij}, \boldsymbol{L}''_{ij}$, and $\boldsymbol{L}'''_{ij}$ are easily determined by calculating

$$\begin{bmatrix}\boldsymbol{L}_{ij} & \boldsymbol{L}'_{ij}\\ \boldsymbol{L}''_{ij} & \boldsymbol{L}'''_{ij}\end{bmatrix} \equiv [\boldsymbol{W}^T\boldsymbol{W}]^{-1}\boldsymbol{W}^T\Big[\mathrm{d}\boldsymbol{H}_{ij} - \{\mathrm{d}\omega_{ij}\}\boldsymbol{I}\Big] \times \boldsymbol{W}[\boldsymbol{W}^T\boldsymbol{W}]^{-1}. \tag{60}$$

Note that $\boldsymbol{W}[\boldsymbol{W}^T\boldsymbol{W}]^{-1}\boldsymbol{W}^T$ is an identity matrix. By erasing small terms such as $O(\mathrm{d}^2)$, we obtain the following equation,

$$\mathrm{d}\boldsymbol{H}_{ij} = l_{ij}\boldsymbol{I} + [\boldsymbol{U},\boldsymbol{V}]\begin{bmatrix}\boldsymbol{L}_{ij} & \boldsymbol{L}'_{ij}\\ \boldsymbol{L}''_{ij} & \boldsymbol{L}'''_{ij}\end{bmatrix}[\boldsymbol{U},\boldsymbol{V}]^T, \tag{61}$$

where

$$\boldsymbol{l}_{ij} = \mathrm{d}\omega_{ij}, \tag{62}$$

$$\boldsymbol{L}_{ij} = \mathrm{d}\boldsymbol{\Omega}_{ij} + \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C} - \mathrm{d}\boldsymbol{O}]\boldsymbol{\Omega}_{ij} + \boldsymbol{\Omega}_{ij}[\mathrm{d}\boldsymbol{C} - \mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\boldsymbol{\Omega}''_{ij} - \boldsymbol{\Omega}'_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{U}]\boldsymbol{C}^{-1}, \tag{63}$$

$$\boldsymbol{L}'_{ij} = \mathrm{d}\boldsymbol{\Omega}'_{ij} + \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C} - \mathrm{d}\boldsymbol{O}]\boldsymbol{\Omega}'_{ij} - \boldsymbol{\Omega}'_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{V}] - \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\boldsymbol{\Omega}'''_{ij} - \boldsymbol{\Omega}_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}], \tag{64}$$

$$\boldsymbol{L}''_{ij} = \mathrm{d}\boldsymbol{\Omega}''_{ij} - [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}\boldsymbol{\Omega}''_{ij} + \boldsymbol{\Omega}''_{ij}[\mathrm{d}\boldsymbol{C} - \mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} - [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}_{ij} - \boldsymbol{\Omega}'''_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{U}]\boldsymbol{C}^{-1}, \tag{65}$$

$$\boldsymbol{L}'''_{ij} = \mathrm{d}\boldsymbol{\Omega}'''_{ij} - [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}\boldsymbol{\Omega}'''_{ij} - \boldsymbol{\Omega}'''_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{V}] - [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}'_{ij} - \boldsymbol{\Omega}''_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}], \tag{66}$$

where we used

$$\boldsymbol{W}^T\boldsymbol{W} = \begin{bmatrix}\boldsymbol{C}+\mathrm{d}\boldsymbol{C} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}, \tag{67}$$

$$\boldsymbol{W}^T[\boldsymbol{U},\boldsymbol{V}] = \begin{bmatrix}\boldsymbol{C}+\mathrm{d}\boldsymbol{O} & [\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\\ [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U} & \boldsymbol{I}+[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}\end{bmatrix}, \tag{68}$$

where $\boldsymbol{C} \equiv \boldsymbol{U}^T\boldsymbol{U} = \begin{bmatrix}\boldsymbol{Q} & \boldsymbol{R}\\ \boldsymbol{R}^T & \boldsymbol{T}\end{bmatrix} \in \mathbb{R}^{\{K+M\}\times\{K+M\}}$, $\mathrm{d}\boldsymbol{C} \equiv [\boldsymbol{U}+\mathrm{d}\boldsymbol{U}]^T[\boldsymbol{U}+\mathrm{d}\boldsymbol{U}] - \boldsymbol{C} = \begin{bmatrix}\mathrm{d}\boldsymbol{Q} & \mathrm{d}\boldsymbol{R}\\ \mathrm{d}\boldsymbol{R}^T & \boldsymbol{0}\end{bmatrix}$, and $\mathrm{d}\boldsymbol{O} \equiv [\mathrm{d}\boldsymbol{U}]^T\boldsymbol{U} = \begin{bmatrix}\mathrm{d}\boldsymbol{S} & \mathrm{d}\boldsymbol{R}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}$.

To realize $\mathrm{d}\boldsymbol{H}$ of Eq. (52) through Eq. (61) under the given $\boldsymbol{U}, \boldsymbol{V}, \mathrm{d}\boldsymbol{U}$, and $\mathrm{d}\boldsymbol{V}$, $l_{ij} = \gamma_{ij}$ and $\begin{bmatrix}\boldsymbol{L}_{ij} & \boldsymbol{L}'_{ij}\\ \boldsymbol{L}''_{ij} & \boldsymbol{L}'''_{ij}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Gamma}_{ij} & \boldsymbol{\Gamma}'_{ij}\\ \boldsymbol{\Gamma}''_{ij} & \boldsymbol{\Gamma}'''_{ij}\end{bmatrix}$ should be satisfied. We then easily obtain an appropriate infinitesimal change of $\boldsymbol{\omega}, \boldsymbol{\Omega}, \boldsymbol{\Omega}', \boldsymbol{\Omega}''$, and $\boldsymbol{\Omega}'''$ as,

$$\mathrm{d}\omega_{ij} = \rho[\boldsymbol{\omega} - \boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij}\mathrm{d}\alpha, \tag{69}$$

$$\begin{aligned}\mathrm{d}\boldsymbol{\Omega}_{ij} = {} & \rho\boldsymbol{C}^{-\frac{1}{2}}\left[\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\omega} + \tilde{\omega}\tilde{\lambda}\tilde{\omega} - \tilde{\boldsymbol{\Omega}}'\tilde{\lambda}\tilde{\boldsymbol{\Omega}}''\right]_{ij}\boldsymbol{C}^{-\frac{1}{2}}\mathrm{d}\alpha \\ & -\boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]\boldsymbol{\Omega}_{ij} - \boldsymbol{\Omega}_{ij}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} + \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\boldsymbol{\Omega}''_{ij} + \boldsymbol{\Omega}'_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{U}]\boldsymbol{C}^{-1},\end{aligned} \tag{70}$$

$$\begin{aligned}\mathrm{d}\boldsymbol{\Omega}'_{ij} = {} & \rho\boldsymbol{C}^{-\frac{1}{2}}\left[\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}'\tilde{\lambda}\tilde{\omega} - \tilde{\boldsymbol{\Omega}}'\tilde{\lambda}\boldsymbol{\Omega}'''\right]_{ij}\mathrm{d}\alpha \\ & -\boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]\boldsymbol{\Omega}'_{ij} + \boldsymbol{\Omega}'_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{V}] \quad + \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\boldsymbol{\Omega}'''_{ij} + \boldsymbol{\Omega}_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}],\end{aligned} \tag{71}$$

$$\begin{aligned}\mathrm{d}\boldsymbol{\Omega}''_{ij} = {} & \rho \quad \left[\tilde{\boldsymbol{\Omega}}'' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\omega}\tilde{\lambda}\tilde{\boldsymbol{\Omega}}'' - \boldsymbol{\Omega}'''\tilde{\lambda}\tilde{\boldsymbol{\Omega}}''\right]_{ij}\boldsymbol{C}^{-\frac{1}{2}}\mathrm{d}\alpha \\ & + \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}\boldsymbol{\Omega}''_{ij} - \boldsymbol{\Omega}''_{ij}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} + \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}_{ij} + \boldsymbol{\Omega}'''_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{U}]\boldsymbol{C}^{-1},\end{aligned} \tag{72}$$

$$\begin{aligned}\mathrm{d}\boldsymbol{\Omega}'''_{ij} = {} & \rho \quad \left[\boldsymbol{\Omega}''' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\omega}\tilde{\lambda}\boldsymbol{\Omega}''' - \boldsymbol{\Omega}'''\tilde{\lambda}\tilde{\omega} - \boldsymbol{\Omega}'''\tilde{\lambda}\boldsymbol{\Omega}'''\right]_{ij}\mathrm{d}\alpha \\ & + \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}\boldsymbol{\Omega}'''_{ij} + \boldsymbol{\Omega}'''_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{V}] \quad + \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}'_{ij} + \boldsymbol{\Omega}''_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}].\end{aligned} \tag{73}$$

We notice that, with respect to N, the orders of each element of the matrices $\boldsymbol{\omega}, \boldsymbol{\Omega}, \boldsymbol{\Omega}', \boldsymbol{\Omega}''$, and $\boldsymbol{\Omega}'''$ are at most $O(1), O(1), O(\frac{1}{\sqrt{N}}), O(\frac{1}{\sqrt{N}})$, and $O(\frac{1}{N})$ (see Appendix B). This suggests that $\boldsymbol{E}$ is negligible at the large limit of N. Actually, if we explicitly assume that the escape directions of the student vectors are completely random, $\left\langle[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{U}^T]\mathrm{d}\boldsymbol{U}\right\rangle_{\boldsymbol{\xi}} = \boldsymbol{0}$ or equivalently $\left\langle\boldsymbol{V}^T\mathrm{d}\boldsymbol{U}\right\rangle_{\boldsymbol{\xi}} = \boldsymbol{0}$ or $\langle\mathrm{d}\boldsymbol{V}\rangle_{\boldsymbol{\xi}} = \boldsymbol{0}$, the time integrations of the terms including $\boldsymbol{U}^T\mathrm{d}\boldsymbol{V}, \boldsymbol{V}^T\mathrm{d}\boldsymbol{U}$, and $\boldsymbol{V}^T\mathrm{d}\boldsymbol{V}$ converge to zero, and we obtain

$$\frac{\mathrm{d}\omega_{ij}}{\mathrm{d}\alpha} = \rho[\boldsymbol{\omega} - \boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij}, \tag{74}$$

$$\frac{\mathrm{d}\boldsymbol{\Omega}_{ij}}{\mathrm{d}\alpha} = \rho\boldsymbol{C}^{-\frac{1}{2}}\left[\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\omega} + \tilde{\omega}\tilde{\lambda}\tilde{\omega} - \tilde{\boldsymbol{\Omega}}'\tilde{\lambda}\tilde{\boldsymbol{\Omega}}''\right]_{ij}\boldsymbol{C}^{-\frac{1}{2}} - \boldsymbol{C}^{-1}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha} + \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}^T\boldsymbol{\Omega}_{ij} - \boldsymbol{\Omega}_{ij}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha} + \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{C}^{-1}, \tag{75}$$

$$\frac{\mathrm{d}\boldsymbol{\Omega}'_{ij}}{\mathrm{d}\alpha} = \rho\boldsymbol{C}^{-\frac{1}{2}}\left[\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}'\tilde{\lambda}\tilde{\omega} - \tilde{\boldsymbol{\Omega}}'\tilde{\lambda}\boldsymbol{\Omega}'''\right]_{ij} - \boldsymbol{C}^{-1}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha} + \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}^T\boldsymbol{\Omega}'_{ij}, \tag{76}$$

$$\frac{\mathrm{d}\boldsymbol{\Omega}''_{ij}}{\mathrm{d}\alpha} = \rho \quad \left[\tilde{\boldsymbol{\Omega}}'' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\omega}\tilde{\lambda}\tilde{\boldsymbol{\Omega}}'' - \boldsymbol{\Omega}'''\tilde{\lambda}\tilde{\boldsymbol{\Omega}}''\right]_{ij}\boldsymbol{C}^{-\frac{1}{2}} \quad - \boldsymbol{\Omega}''_{ij}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha} + \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{C}^{-1}, \tag{77}$$

$$\frac{\mathrm{d}\boldsymbol{\Omega}'''_{ij}}{\mathrm{d}\alpha} = \rho \quad \left[\boldsymbol{\Omega}''' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\omega}\tilde{\lambda}\boldsymbol{\Omega}''' - \boldsymbol{\Omega}'''\tilde{\lambda}\tilde{\omega} - \boldsymbol{\Omega}'''\tilde{\lambda}\boldsymbol{\Omega}'''\right]_{ij}. \tag{78}$$

These equations guarantee that $\boldsymbol{\Omega}', \boldsymbol{\Omega}''$, and $\boldsymbol{\Omega}'''$ are always zero at the large limit of N, because the initial zero values of $\boldsymbol{\Omega}'_{ij}, \boldsymbol{\Omega}''_{ij}$, and $\boldsymbol{\Omega}'''_{ij}$ are preserved. Consequently, we adopt $\boldsymbol{\omega}$ and $\boldsymbol{\Omega}$ as the new order parameters, which successfully express $\boldsymbol{H}$.

Next, we consider the $\boldsymbol{HGH}$ term in Eq. (51) because a small fluctuation of $\boldsymbol{E}$ might become significant in this second-order of the $\boldsymbol{H}$ term. Specifically, we evaluate the square of the fluctuation term $\boldsymbol{E}$ (to be exact, $\boldsymbol{\omega}$ and $\boldsymbol{\Omega}$ also fluctuate, but their fluctuations do not become significant because these matrices are sufficiently small compared with N).

$$\mathrm{d}[\boldsymbol{E}^2]_{ij} = \left[[\boldsymbol{E} + \mathrm{d}\boldsymbol{E}]^2 - \boldsymbol{E}^2\right]_{ij}$$

In a similar way as above, we let

$$[\boldsymbol{E}^2]_{ij} = [\boldsymbol{U}, \boldsymbol{V}]\begin{bmatrix}\boldsymbol{\Upsilon}_{ij} & \boldsymbol{\Upsilon}'_{ij} \\ \boldsymbol{\Upsilon}''_{ij} & \boldsymbol{\Upsilon}'''_{ij}\end{bmatrix}[\boldsymbol{U}, \boldsymbol{V}]^T, \tag{79}$$

where

$$\begin{bmatrix}\boldsymbol{\Upsilon}_{ij} & \boldsymbol{\Upsilon}'_{ij} \\ \boldsymbol{\Upsilon}''_{ij} & \boldsymbol{\Upsilon}'''_{ij}\end{bmatrix} \equiv \sum_{k=1}^{K}\begin{bmatrix}\boldsymbol{\Omega}'_{ik}\boldsymbol{\Omega}''_{kj} & \boldsymbol{\Omega}'_{ik}\boldsymbol{\Omega}'''_{kj} \\ \boldsymbol{\Omega}'''_{ik}\boldsymbol{\Omega}''_{kj} & \boldsymbol{\Omega}''_{ik}\boldsymbol{C}\boldsymbol{\Omega}'_{kj} + \boldsymbol{\Omega}'''_{ik}\boldsymbol{\Omega}'''_{kj}\end{bmatrix}. \tag{80}$$

Then, an infinitesimal change of $\boldsymbol{E}^2$ is

$$= \sum_{k=1}^{K} \boldsymbol{W} \begin{bmatrix} [\boldsymbol{\Omega}'_{ik}+\mathrm{d}\boldsymbol{\Omega}'_{ik}][\boldsymbol{\Omega}''_{kj}+\mathrm{d}\boldsymbol{\Omega}''_{kj}] & [\boldsymbol{\Omega}'_{ik}+\mathrm{d}\boldsymbol{\Omega}'_{ik}][\boldsymbol{\Omega}'''_{kj}+\mathrm{d}\boldsymbol{\Omega}'''_{kj}] \\ [\boldsymbol{\Omega}'''_{ik}+\mathrm{d}\boldsymbol{\Omega}'''_{ik}][\boldsymbol{\Omega}''_{kj}+\mathrm{d}\boldsymbol{\Omega}''_{kj}] & [\boldsymbol{\Omega}''_{ik}+\mathrm{d}\boldsymbol{\Omega}''_{ik}][\boldsymbol{C}+\mathrm{d}\boldsymbol{C}][\boldsymbol{\Omega}'_{kj}+\mathrm{d}\boldsymbol{\Omega}'_{kj}]+[\boldsymbol{\Omega}'''_{ik}+\mathrm{d}\boldsymbol{\Omega}'''_{ik}][\boldsymbol{\Omega}'''_{kj}+\mathrm{d}\boldsymbol{\Omega}'''_{kj}] \end{bmatrix} \boldsymbol{W}^T$$
$$- \sum_{k=1}^{K} [\boldsymbol{U},\boldsymbol{V}] \begin{bmatrix} \boldsymbol{\Omega}'_{ik}\boldsymbol{\Omega}''_{kj} & \boldsymbol{\Omega}'_{ik}\boldsymbol{\Omega}'''_{kj} \\ \boldsymbol{\Omega}'''_{ik}\boldsymbol{\Omega}''_{kj} & \boldsymbol{\Omega}''_{ik}\boldsymbol{C}\boldsymbol{\Omega}'_{kj}+\boldsymbol{\Omega}'''_{ik}\boldsymbol{\Omega}'''_{kj} \end{bmatrix} [\boldsymbol{U},\boldsymbol{V}]^T. \tag{81}$$

Substituting Eqs. (70)-(73) for $\mathrm{d}\boldsymbol{\Omega}'$, $\mathrm{d}\boldsymbol{\Omega}''$, and $\mathrm{d}\boldsymbol{\Omega}'''$, and erasing small terms, we obtain an appropriate infinitesimal change of $\boldsymbol{\Upsilon}_{ij}$, $\boldsymbol{\Upsilon}'_{ij}$, $\boldsymbol{\Upsilon}''_{ij}$, and $\boldsymbol{\Upsilon}'''_{ij}$. Briefly, the point is that $[\mathrm{d}\boldsymbol{\Omega}'_{ik}]\mathrm{d}\boldsymbol{\Omega}''_{kj}$ is not negligible but $\boldsymbol{\Omega}_{ik}\begin{bmatrix}\mathrm{d}\boldsymbol{\phi} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{\Omega}_{kj}$, because $\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}][\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U} = \begin{bmatrix}\mathrm{d}\boldsymbol{\phi} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}$ (see Appendix B). Actually, we obtain,

$$\begin{aligned}\mathrm{d}\boldsymbol{\Upsilon}_{ij} = \rho\boldsymbol{C}^{-\frac{1}{2}} \Big[& 2\tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\Omega}}'' - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\Omega}}'' - \tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} \\ & - \tilde{\boldsymbol{\Omega}}'[\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}}+\tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}]\tilde{\boldsymbol{\Omega}}'' \Big]_{ij} \boldsymbol{C}^{-\frac{1}{2}}\mathrm{d}\alpha \\ + \sum_{k=1}^{K} \Big[& \boldsymbol{\Gamma}'_{ik}\boldsymbol{\Omega}''_{kj} - \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]\boldsymbol{\Omega}'_{ik}\boldsymbol{\Omega}''_{kj} + \boldsymbol{\Omega}'_{ik}\boldsymbol{\Gamma}''_{kj} \\ & - \boldsymbol{\Omega}'_{ik}\boldsymbol{\Omega}''_{kj}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} + \boldsymbol{\Omega}_{ik}\begin{bmatrix}\mathrm{d}\boldsymbol{\phi} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{\Omega}_{kj} \Big].\end{aligned} \tag{82}$$

We notice that each element of $\boldsymbol{\Upsilon}_{ij}$ is initially zero, and becomes $O(1)$, because $\boldsymbol{\Omega}_{ik}\begin{bmatrix}\mathrm{d}\boldsymbol{\phi} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{\Omega}_{kj}$ is $O(1)$, whereas the other terms, $\boldsymbol{\Upsilon}'_{ij}$, $\boldsymbol{\Upsilon}''_{ij}$, and $\boldsymbol{\Upsilon}'''_{ij}$, are considered to be always zero at the large limit of N.

We then let the following positive definite and symmetric matrix

$$\boldsymbol{\Upsilon} = \boldsymbol{\Omega}'\boldsymbol{\Omega}'' \in \mathbb{R}^{K\{K+M\}\times K\{K+M\}}, \tag{83}$$

be another new order parameter, where $\boldsymbol{\Upsilon} = [\boldsymbol{\Upsilon}_{ij}]_{i,j=1,..,K}$. In short, we can write

$$\boldsymbol{E}_{ij} \to \boldsymbol{0}, \tag{84}$$
$$[\boldsymbol{E}^2]_{ij} - \boldsymbol{U}\boldsymbol{\Upsilon}_{ij}\boldsymbol{U}^T \to \boldsymbol{0}, \tag{85}$$

or, equivalently,

$$\omega_{ij}\boldsymbol{I} + \boldsymbol{U}\boldsymbol{\Omega}_{ij}\boldsymbol{U}^T \to \boldsymbol{H}_{ij}, \tag{86}$$

$$\sum_{k=1}^{K} [\omega_{ik}\boldsymbol{I} + \boldsymbol{U}\boldsymbol{\Omega}_{ik}\boldsymbol{U}^T][\omega_{kj}\boldsymbol{I} + \boldsymbol{U}\boldsymbol{\Omega}_{kj}\boldsymbol{U}^T] + \boldsymbol{U}\boldsymbol{\Upsilon}_{ij}\boldsymbol{U}^T \to [\boldsymbol{H}^2]_{ij}, \tag{87}$$

at the large limit of N.

This $\boldsymbol{\Upsilon}$ is useful for estimating the matrices $\boldsymbol{\Omega}'$ and $\boldsymbol{\Omega}''$, which are required for both the $\tilde{\boldsymbol{\Omega}}'[\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}}+\tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}]\tilde{\boldsymbol{\Omega}}''$ term in Eq. (82) and the $\tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\Omega}}''$ term in Eq. (75). Intrinsically, an exact restoration of $\boldsymbol{\Omega}'$ and $\boldsymbol{\Omega}''$ using $\boldsymbol{\Upsilon}$ is impossible, because the number of elements of $\boldsymbol{\Omega}'$ and $\boldsymbol{\Omega}''$ is $O(N)$, while that of $\boldsymbol{\Upsilon}$ is $O(1)$. Hence, we substitute one of the probable candidates of $\boldsymbol{\Omega}'_{ij}$ and $\boldsymbol{\Omega}''_{ij}$ for the true $\boldsymbol{\Omega}'$ and $\boldsymbol{\Omega}''$; i.e., we use

$$\boldsymbol{\Omega}'_{ij} = \boldsymbol{\Omega}''^T_{ji} = \frac{1}{\sqrt{n}}\left[[\sqrt{\boldsymbol{\Upsilon}}]_{ij}, \cdots, [\sqrt{\boldsymbol{\Upsilon}}]_{ij} \right], \tag{88}$$

where $n \equiv \frac{N-K-M}{K+M}$ was assumed to be a natural number. This Eq. (88) certainly satisfies the necessary condition given by Eq. (83). Substituting this Eq. (88) for Eq. (82), we obtain

$$\begin{aligned}\mathrm{d}\boldsymbol{\Upsilon}_{ij} = \rho\Big[& \boldsymbol{C}^{-\frac{1}{2}} \left[2\tilde{\boldsymbol{\Upsilon}} - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Upsilon}} - \tilde{\boldsymbol{\Upsilon}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} \right]_{ij} \boldsymbol{C}^{-\frac{1}{2}} \\ & - \left[\sqrt{\boldsymbol{\Upsilon}}[\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}}+\tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}]\sqrt{\boldsymbol{\Upsilon}} \right]_{ij} \Big]\mathrm{d}\alpha \\ & - \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]\boldsymbol{\Upsilon}_{ij} - \boldsymbol{\Upsilon}_{ij}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} \\ & + \sum_{k=1}^{K} \boldsymbol{\Omega}_{ik}\begin{bmatrix}\mathrm{d}\boldsymbol{\phi} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{\Omega}_{kj},\end{aligned} \tag{89}$$

where $\tilde{\boldsymbol{\Upsilon}} \equiv [\tilde{\boldsymbol{\Upsilon}}_{ij} \equiv \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{\Upsilon}_{ij}\boldsymbol{C}^{\frac{1}{2}}]_{i,j=1,..,K}$. In a similar manner, we can substitute Eq. (88) for Eq. (75).

Finally, we obtain the new order parameter dynamics through Eqs. (74), (75), and (89) as

$$\frac{\mathrm{d}\omega_{ij}}{\mathrm{d}\alpha} = \rho[\boldsymbol{\omega}-\boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij}, \tag{90}$$

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\Omega}_{ij}}{\mathrm{d}\alpha} = & \rho\left[\boldsymbol{C}^{-\frac{1}{2}} \left[\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\omega}} + \tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}} \right]_{ij} \boldsymbol{C}^{-\frac{1}{2}} - \left[\sqrt{\boldsymbol{\Upsilon}}\tilde{\boldsymbol{\lambda}}\sqrt{\boldsymbol{\Upsilon}} \right]_{ij} \right] \\ & - \boldsymbol{C}^{-1}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha}+\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}^T\boldsymbol{\Omega}_{ij} - \boldsymbol{\Omega}_{ij}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha}+\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{C}^{-1},\end{aligned} \tag{91}$$

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\Upsilon}_{ij}}{\mathrm{d}\alpha} = & \rho\left[\boldsymbol{C}^{-\frac{1}{2}} \left[2\tilde{\boldsymbol{\Upsilon}} - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Upsilon}} - \tilde{\boldsymbol{\Upsilon}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} \right]_{ij} \boldsymbol{C}^{-\frac{1}{2}} - \left[\sqrt{\boldsymbol{\Upsilon}}[\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}}+\tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}]\sqrt{\boldsymbol{\Upsilon}} \right]_{ij} \right] \\ & - \boldsymbol{C}^{-1}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha}+\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}^T\boldsymbol{\Upsilon}_{ij} - \boldsymbol{\Upsilon}_{ij}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha}+\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{C}^{-1}\end{aligned}$$

$$+\sum_{k=1}^{K}\boldsymbol{\Omega}_{ik}\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\boldsymbol{\Omega}_{kj}. \tag{92}$$

Intuitively, in Eqs. (90)-(92), the terms including ρ correspond to the dynamics of $\boldsymbol{H}$, whereas the terms including $\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{S}}{\mathrm{d}\alpha}+\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\alpha}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ keep the component of $\boldsymbol{H}$ constant which is still expressible by $\boldsymbol{\Omega}$ or $\boldsymbol{\Upsilon}$ under the movement of $\boldsymbol{U}$. The term including $\begin{bmatrix}\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}\alpha} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ means the component of $\boldsymbol{H}$ which is no longer expressible by $\boldsymbol{\Omega}$ under the escape movement of $\boldsymbol{U}$ into the null space of $\boldsymbol{U}$.

The dynamics of the usual order parameters for ANGD given by Eq. (47) can be rewritten using the new order parameters in a manner similar to that for NGD,

$$\begin{aligned}
\frac{\mathrm{d}R_{ij}}{\mathrm{d}\alpha} &= -\eta\sum_{k=1}^{K}\left[\omega_{ik}\left\langle\delta_k\boldsymbol{x}_j\right\rangle_{\boldsymbol{z}}+\left\langle\delta_k\boldsymbol{z}^T\right\rangle_{\boldsymbol{z}}\boldsymbol{\Omega}_{ik}^T\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{T}\end{bmatrix}_{\bullet j}\right],\\
\frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha} &= -\eta\sum_{k=1}^{K}\left[\omega_{ik}\left\langle\delta_k\boldsymbol{x}_j\right\rangle_{\boldsymbol{z}}+\left\langle\delta_k\boldsymbol{z}^T\right\rangle_{\boldsymbol{z}}\boldsymbol{\Omega}_{ik}^T\begin{bmatrix}\boldsymbol{Q}\\ \boldsymbol{R}^T\end{bmatrix}_{\bullet j}\right],\\
\frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha} &= \eta^2\sum_{k,l=1}^{K}\omega_{ik}\left\langle\delta_k\delta_l\right\rangle_{\boldsymbol{z}}\omega_{lj},\\
\frac{\mathrm{d}Q_{ij}}{\mathrm{d}\alpha} &= \frac{\mathrm{d}S_{ij}}{\mathrm{d}\alpha}+\frac{\mathrm{d}S_{ji}}{\mathrm{d}\alpha}+\frac{\mathrm{d}\phi_{ij}}{\mathrm{d}\alpha}.
\end{aligned} \tag{93}$$

Note that these usual order parameter dynamics are not affected by the small fluctuation, $\boldsymbol{E}$ (see Appendix C). Equations (90)-(93) are the order parameter dynamics expressed by the order parameters themselves.

IV. NUMERICAL RESULTS

We numerically validated the theoretical results through simulation, and evaluated the performance of the simplified version of ANGD. The numerical results obtained using the theory were comparable with those of the simulation with respect to not only the learning curves but also the learning failures. We also found that the performance of ANGD is roughly comparable with that of NGD when $\frac{\eta}{\rho}$ is small. Detailed conditions of these numerical results are given in Appendix D.

First, we validated the theoretical motion equations by using a simulation with N set to 500. The learning curves (time evolution of the generalization error) of the theory are shown in Fig. 2(c), while those of the simulation are shown in Fig. 3(c). There were no significant differences between the theoretical and the simulation results. With respect to the *two $\boldsymbol{\xi}$ rule*, simulation results showed that those adopting this rule were generally comparable with those not adopting this rule, although they were slightly slower when ρ is large. (We show only the simulation results for adoption of this rule.)

We also evaluated the learning failure of the simplified version of ANGD. As this version of ANGD defined by Eq. (12) assumes $\rho \ll 1$, a large ρ could cause problems. The simulation results showed divergence of the network parameters and the system failed to learn the teacher outputs with large ρ (solid line in Fig. 4(a)). Figure 4(b) shows the borderline between this learning failure and the success areas with respect to the η and ρ conditions. Roughly, the failure area corresponded to $\rho \geq 0.05$. Our numerical solution of the theory successfully reproduced these learning failures in the simulation, which are shown as the dotted lines in Fig. 4(a) and Fig. 4(b). (We considered a learning failure to have occurred with the theory when the correlation matrix of the weight vectors, $\boldsymbol{C}$, violated its positive definiteness.)

Next, we compared the learning curves between SGD, NGD, and ANGD under various teacher weight vector correlations. (The angle between teacher weight vectors is denoted as κ.) Figure 2 shows the learning curves for (a) SGD, (b) NGD, and (c) ANGD. We can see that ANGD had almost the same performance as NGD and does not have any severe plateaus. Moreover, ANGD was not greatly affected by the teacher correlations, although SGD was.

Finally, we reveal the key condition affecting the learning plateau in ANGD. NGD is known to have a plateau when the learning rate η is too large [6]. We found that a plateau occurs in ANGD not only in the large η case, but also in the small ρ case. Figure 5(a) shows the time cost of learning under a wide range of η and ρ. This contour graph suggests that a plateau occurs when $\frac{\eta}{\rho}$ is large. Our simulation study also supported this finding (Fig. 5(b)). This phenomenon may be interpreted to mean that $\hat{\boldsymbol{G}}^{-1}$ cannot follow a change in the true $\boldsymbol{G}^{-1}$ if η is relatively large compared to ρ.

V. CONCLUSION

We have developed a new order parameter expression for a simplified version of adaptive natural gradient learning in which the learning dynamics can be expressed using only a few order parameters. We numerically validated this theory through simulation and confirmed that this theory successfully reproduces not only the learning curve, but also the learning failure. We found that the ANGD performance is generally comparable with that of NGD. We also found that we can avoid the plateau in ANGD by making the update rate of the network parameter η low enough compared to the update rate of the inverse of the Fisher information matrix ρ.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas No. 14084212 and Grant-in-Aid for Scientific Research (C) No. 14580438.

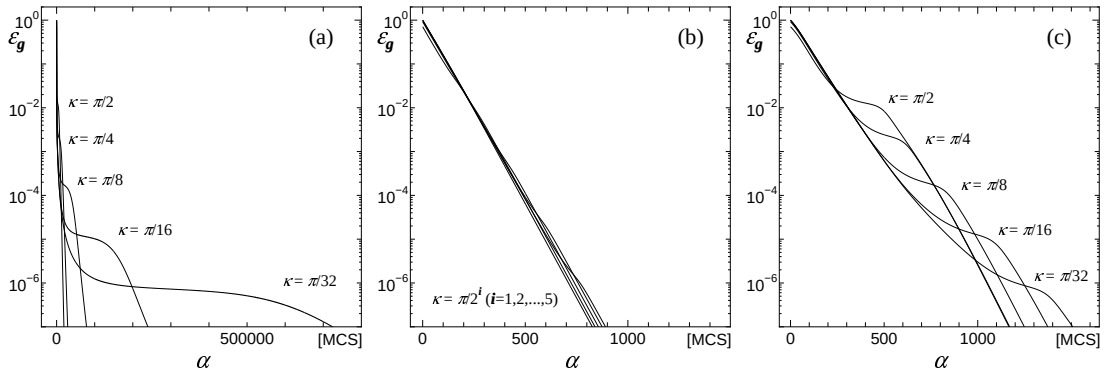

FIG. 2: Numerical results of the theory. Time evolution of the generalization error at $\eta = 0.01, \rho = 0.01,$ and $N = 500$. (a) SGD, (b) NGD, and (c) ANGD. NGD and ANGD are not greatly affected by the angle of the teacher weight vectors (κ), whereas SGD is.

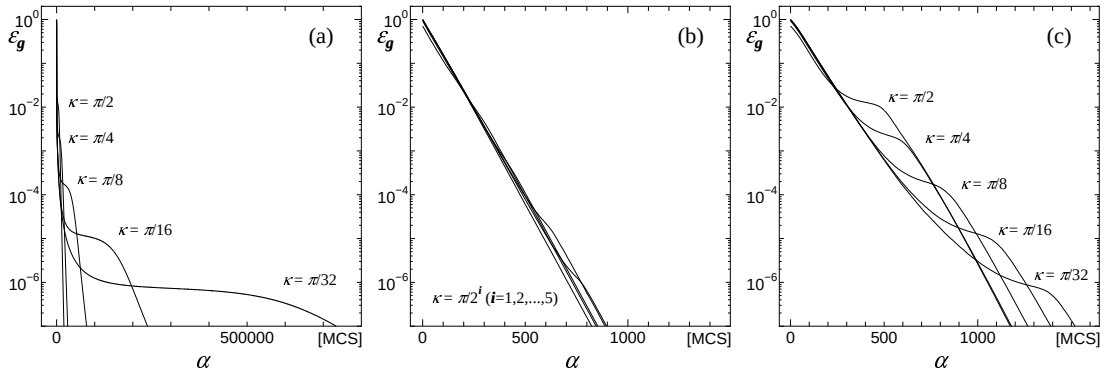

FIG. 3: Simulation results. Time evolution of the generalization error at $\eta = 0.01, \rho = 0.01,$ and $N = 500$. (a) SGD, (b) NGD, and (c) ANGD. The results obtained using the theory in Fig. 2 are comparable to the simulation results.

APPENDIX A: CONVERGENCE OF MATRICES

In this paper, we sometimes refer to matrix or vector 'convergence' in the sense of each element, although this word is usually used in the sense that the norm of the difference between the series of concern and a given matrix or vector converges to zero. As we are dealing with the large limit of the input dimension N, the average of N $\mathbb{R}^{N\times N}$-matrices or N $\mathbb{R}^N$-vectors often converges in the sense of each element, but does not in the sense of the norm. We can see one example of this phenomenon in $\boldsymbol{J}_i$ dynamics in SGD; although each element of its fluctuation is small, the norm of this fluctuation is not zero but $\Delta\phi_{ii}$ (see Eq. (21) and Fig. 1). In the following, we discuss the convergence of matrix $\boldsymbol{F}$ used in Eq. (44).

Let us estimate the order of $\boldsymbol{F}$, specifically, the following matrix:

$$\frac{1}{N}\sum_{\mu=1}^{N}[\nabla f^{(\mu)}][\nabla f^{(\mu)}]^T, \tag{A1}$$

where $\nabla f^{(\mu)} = [g'(\boldsymbol{J}_i^T\boldsymbol{\xi}^{(\mu)})\boldsymbol{\xi}^{(\mu)}]_{i=1,..,K} \in \mathbb{R}^{NK}$ is a random vector dependent on the random input $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Here, the superscript of $\bullet^{(\circ)}$ denotes not the time but simple identification. The diagonal element of each block of this matrix, e.g., the (k, k) element of the (i, j) block, is

$$\frac{1}{N}\sum_{\mu=1}^{N} g'(\boldsymbol{J}_i^T\boldsymbol{\xi}^{(\mu)})g'(\boldsymbol{J}_j^T\boldsymbol{\xi}^{(\mu)})\{\xi_k^{(\mu)}\}^2. \tag{A2}$$

If we drop $g'(\boldsymbol{J}_i^T\boldsymbol{\xi}^{(\mu)})g'(\boldsymbol{J}_j^T\boldsymbol{\xi}^{(\mu)})$, because it is $O(1)$, we notice that the probability distribution of this element is given by an N-freedom chi-square distribution; i.e., its moment-generating function is defined as

$$\varphi(t) = \left\{1 - 2\left\{\frac{1}{N}t\right\}\right\}^{-N/2}. \tag{A3}$$

Consequently, we get the variance of this element as

$$\frac{\partial^2}{\partial t^2}\ln\varphi(t)\bigg|_{t=0} = \frac{2}{N}. \tag{A4}$$

The non-diagonal element of each block, e.g., the (k, l) element of the (i, j) block, is

$$\frac{1}{N}\sum_{\mu=1}^{N} g'(\boldsymbol{J}_i^T\boldsymbol{\xi}^{(\mu)})g'(\boldsymbol{J}_j^T\boldsymbol{\xi}^{(\mu)})\xi_k^{(\mu)}\xi_l^{(\mu)}. \tag{A5}$$

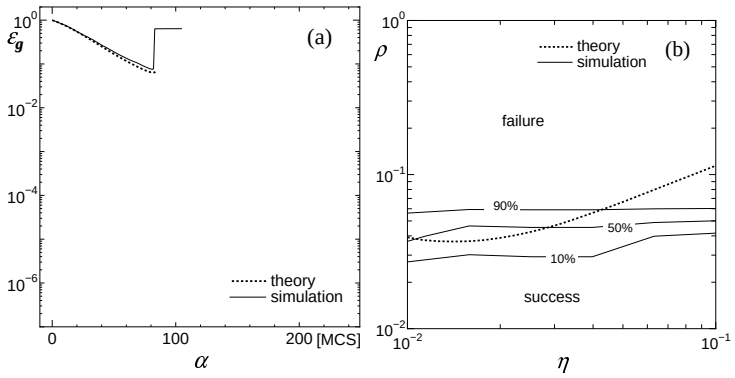

FIG. 4: Learning failure of the simplified version of ANGD with $N = 100$. (a) Time evolution of the generalization error under the conditions of $\eta = 0.02$ and $\rho = 0.1$. (b) Contour graph under various values of η and ρ. The probability of learning failure is shown with respect to the simulation results because the simulation was a stochastic process. Numerical results of the theory well predicted the learning failure in the simulation.

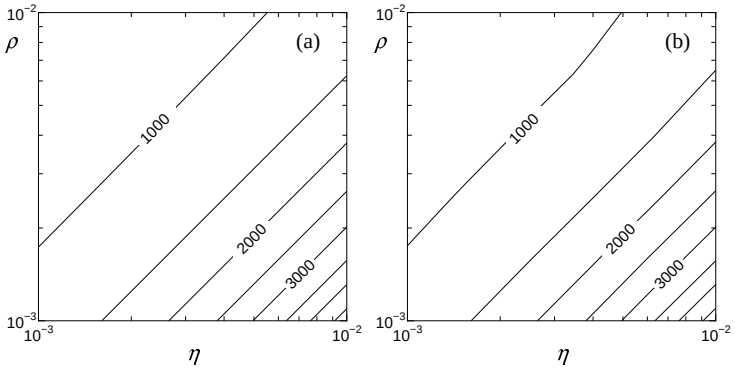

FIG. 5: Contour graphs of the time cost needed for ϵ_g to reach $\epsilon_g < 10^{-7}$ for various values of η and ρ in ANGD. Here, κ was set to $\pi/8$. (a) Results obtained using the theory, and (b) simulation results. Time α was normalized as $100\eta\alpha$. The plateau length was strongly dependent on $\frac{\eta}{\rho}$.

Similarly, we notice that this distribution is given by an average of N modified Bessel functions of the second kind. Then, we get the moment generating function,

$$\varphi(t) = \left\{1 - \left\{\frac{1}{N}t\right\}^2\right\}^{-N/2}, \tag{A6}$$

and the variance as $\frac{1}{N}$.

Therefore, each element of the matrix converges as $O(\frac{1}{\sqrt{N}})$, but the Frobenius norm diverges as $O(\sqrt{N})$ because this matrix has $NK \times NK$ elements. In other words, $\sum \boldsymbol{F} \to \boldsymbol{G}$ holds with respect to each element, but does not converge with respect to the Frobenius norm.

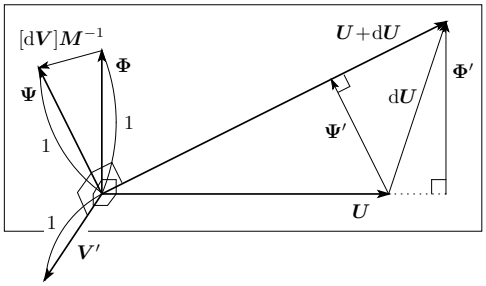

FIG. 6: Intuitive schema of the null space of $\boldsymbol{U}$ and $\boldsymbol{U}+\mathrm{d}\boldsymbol{U}$.

APPENDIX B: ORDERS OF $\boldsymbol{\omega}, \boldsymbol{\Omega}, \boldsymbol{\Omega}', \boldsymbol{\Omega}''$ AND $\boldsymbol{\Omega}'''$

We prove that, with respect to N, the orders of each element of the matrices $\boldsymbol{\omega}, \boldsymbol{\Omega}, \boldsymbol{\Omega}', \boldsymbol{\Omega}''$, and $\boldsymbol{\Omega}'''$ are at most $O(1), O(1), O(\frac{1}{\sqrt{N}}), O(\frac{1}{\sqrt{N}})$, and $O(\frac{1}{N})$. First, we evaluate the order of the $\boldsymbol{U}^T\mathrm{d}\boldsymbol{V}, \boldsymbol{V}^T\mathrm{d}\boldsymbol{U}$, and $\boldsymbol{V}^T\mathrm{d}\boldsymbol{V}$ terms, which are used in Eqs. (70)-(73). We especially pay attention to the fact that some of the inner products between infinitesimal changes of N-dimensional vectors are not $O(\mathrm{d}^2)$ but $O(\mathrm{d})$; e.g., $[\mathrm{d}\boldsymbol{U}]^T\mathrm{d}\boldsymbol{U} = \mathrm{d}\boldsymbol{\phi}' \equiv \begin{bmatrix} \mathrm{d}\boldsymbol{\phi} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{\{K+M\}\times\{K+M\}}$. We then consider Eqs. (69)-(73).

First of all, we explicitly determine $\boldsymbol{V}$, the orthonormal

bases of the null space of the weight vectors, using $\boldsymbol{U}$ and $\mathrm{d}\boldsymbol{U}$. We find that some of the orthonormal bases of the $\boldsymbol{V}$ subspace can be expressed using $\boldsymbol{U}$ and $\mathrm{d}\boldsymbol{U}$ as

$$\boldsymbol{\Phi} \equiv \boldsymbol{\Phi}' \left[\boldsymbol{\Phi}'^T \boldsymbol{\Phi}' \right]^{-\frac{1}{2}} \in \mathbb{R}^{N \times \{K+M\}}, \tag{B1}$$

where

$$\boldsymbol{\Phi}' \equiv \left[\boldsymbol{I} - \boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{U}^T \right] \mathrm{d}\boldsymbol{U} \in \mathbb{R}^{N \times \{K+M\}}. \tag{B2}$$

Here, $\boldsymbol{\Phi}'$ is the orthogonal component of the $\mathrm{d}\boldsymbol{U}$ to $\boldsymbol{U}$ subspace, while $\boldsymbol{\Phi}$ is the normalized $\boldsymbol{\Phi}'$; i.e., $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \boldsymbol{I}$ (see Fig. 6). (We assume the rank of $\mathrm{d}\boldsymbol{U}$ is K+M for simplicity, although it is actually K.) Note that $\boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{U}^T$ is a projection matrix to the $\boldsymbol{U}$ subspace. In a similar manner, we also find that some of the orthonormal bases of the $\boldsymbol{V}$+$\mathrm{d}\boldsymbol{V}$ subspace can be expressed using $\boldsymbol{U}$ and $\mathrm{d}\boldsymbol{U}$ as

$$\boldsymbol{\Psi} \equiv \boldsymbol{\Psi}' \left[\boldsymbol{\Psi}'^T \boldsymbol{\Psi}' \right]^{-\frac{1}{2}} \in \mathbb{R}^{N \times \{K+M\}}, \tag{B3}$$

where

$$\boldsymbol{\Psi}' \equiv -\left[\boldsymbol{I} - [\boldsymbol{U}+\mathrm{d}\boldsymbol{U}][\boldsymbol{C}+\mathrm{d}\boldsymbol{C}]^{-1}[\boldsymbol{U}+\mathrm{d}\boldsymbol{U}]^T \right] \boldsymbol{U} \in \mathbb{R}^{N \times \{K+M\}}. \tag{B4}$$

Here, $\boldsymbol{\Psi}'$ is the orthogonal component of $\boldsymbol{U}$ to the $\boldsymbol{U}$+$\mathrm{d}\boldsymbol{U}$ subspace, while $\boldsymbol{\Psi}$ is the normalized $\boldsymbol{\Psi}'$; i.e., $\boldsymbol{\Psi}^T\boldsymbol{\Psi} = \boldsymbol{I}$. Thus, we can think that some of the orthonormal bases of the $\boldsymbol{V}$ subspace correspond to $\boldsymbol{\Phi}$, and $\boldsymbol{\Phi}$ moves to $\boldsymbol{\Psi}$ as $\boldsymbol{V}$ moves to $\boldsymbol{V}$+$\mathrm{d}\boldsymbol{V}$. However, the column vectors of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ do not necessarily coincide with some of the column vectors of $\boldsymbol{V}$ and $\boldsymbol{V}$+$\mathrm{d}\boldsymbol{V}$, respectively; i.e., some rotation or mirror image conversion might be required. Hence, we introduce an appropriate orthonormal matrix $\boldsymbol{M} \in \mathbb{R}^{\{N-K-M\} \times \{N-K-M\}}$, and explicitly express $\boldsymbol{V}$, $\boldsymbol{V}$+$\mathrm{d}\boldsymbol{V}$, and $\mathrm{d}\boldsymbol{V}$ as

$$\boldsymbol{V} = [\boldsymbol{\Phi}, \boldsymbol{V}']\boldsymbol{M}, \tag{B5}$$
$$\boldsymbol{V}+\mathrm{d}\boldsymbol{V} = [\boldsymbol{\Psi}, \boldsymbol{V}']\boldsymbol{M}, \tag{B6}$$
$$\mathrm{d}\boldsymbol{V} = [\boldsymbol{\Psi}-\boldsymbol{\Phi}, \boldsymbol{0}]\boldsymbol{M}, \tag{B7}$$

where $\boldsymbol{V}'$ is one of the set of orthonormal bases of the null space of $\boldsymbol{U}$ and $\boldsymbol{U}$+$\mathrm{d}\boldsymbol{U}$. Note that each of

$$\begin{aligned}
&\left[\boldsymbol{U}\boldsymbol{C}^{-\frac{1}{2}}, \boldsymbol{V}\right], \\
&\left[\boldsymbol{U}\boldsymbol{C}^{-\frac{1}{2}}, \boldsymbol{\Phi}, \boldsymbol{V}'\right], \\
&\left[[\boldsymbol{U}+\mathrm{d}\boldsymbol{U}][\boldsymbol{C}+\mathrm{d}\boldsymbol{C}]^{-\frac{1}{2}}, \boldsymbol{V}+\mathrm{d}\boldsymbol{V}\right], \\
&\left[[\boldsymbol{U}+\mathrm{d}\boldsymbol{U}][\boldsymbol{C}+\mathrm{d}\boldsymbol{C}]^{-\frac{1}{2}}, \boldsymbol{\Psi}, \boldsymbol{V}'\right],
\end{aligned} \tag{B8}$$

consists of the orthonormal bases of the whole space. If we decompose $\boldsymbol{M}$ as $\begin{bmatrix}\boldsymbol{M}_0 \\ \boldsymbol{M}_1\end{bmatrix}$, we can rewrite Eq. (B5)-(B7) as

$$\boldsymbol{V} = [\boldsymbol{\Phi}, \boldsymbol{V}']\boldsymbol{M} = \boldsymbol{\Phi}\boldsymbol{M}_0 + \boldsymbol{V}'\boldsymbol{M}_1, \tag{B9}$$
$$\boldsymbol{V}+\mathrm{d}\boldsymbol{V} = [\boldsymbol{\Psi}, \boldsymbol{V}']\boldsymbol{M} = \boldsymbol{\Psi}\boldsymbol{M}_0 + \boldsymbol{V}'\boldsymbol{M}_1, \tag{B10}$$
$$\mathrm{d}\boldsymbol{V} = [\boldsymbol{\Psi}-\boldsymbol{\Phi}, \boldsymbol{0}]\boldsymbol{M} = [\boldsymbol{\Psi}-\boldsymbol{\Phi}]\boldsymbol{M}_0. \tag{B11}$$

The $\boldsymbol{V}'$ and $\boldsymbol{M}_1$ contain some arbitrariness, although $\boldsymbol{V}'\boldsymbol{M}_1$ is well-defined. However, we can avoid using them as shown below.

We can then calculate the $\boldsymbol{V}^T\mathrm{d}\boldsymbol{U}$, $[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}$, and $[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}$ terms. We find that the norms of the column vectors of $\boldsymbol{V}^T\mathrm{d}\boldsymbol{U}$,

$$\begin{aligned}
&[\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\boldsymbol{V}^T\mathrm{d}\boldsymbol{U} \\
&= [\mathrm{d}\boldsymbol{U}]^T[\boldsymbol{\Phi}, \boldsymbol{V}']\boldsymbol{M}\boldsymbol{M}^T[\boldsymbol{\Phi}, \boldsymbol{V}']^T\mathrm{d}\boldsymbol{U} \\
&= \mathrm{d}\boldsymbol{\phi}',
\end{aligned} \tag{B12}$$

are $O(1)$ with respect to N. Note that $\boldsymbol{M}\boldsymbol{M}^T = \boldsymbol{I}$ by its definition. As the orthonormal matrix $\boldsymbol{M}$ does not change a matrix norm, each element of the N-dimensional vectors $\boldsymbol{V}^T\mathrm{d}\boldsymbol{U}$ is $O(\frac{1}{\sqrt{N}})$. In a similar manner, we can calculate the norm of $[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}$,

$$\begin{aligned}
&\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}][\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U} \\
&= \boldsymbol{U}^T[\boldsymbol{\Psi}-\boldsymbol{\Phi}, \boldsymbol{0}]\boldsymbol{M}\boldsymbol{M}^T[\boldsymbol{\Psi}-\boldsymbol{\Phi}, \boldsymbol{0}]^T\boldsymbol{U} \\
&= \mathrm{d}\boldsymbol{\phi}'.
\end{aligned} \tag{B13}$$

This also means that each element of $[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}$ is $O(\frac{1}{\sqrt{N}})$. These results are used after Eq. (73). Equation (B13) is also used to derive Eq. (82). We also find that each element of $[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}$ is $O(\frac{1}{N^2})$, because

$$\begin{aligned}
&\boldsymbol{V}^T[\mathrm{d}\boldsymbol{V}] \\
&= \boldsymbol{M}^T[\boldsymbol{\Phi}, \boldsymbol{V}']^T[\boldsymbol{\Psi}-\boldsymbol{\Phi}, \boldsymbol{0}]\boldsymbol{M} \\
&= -\boldsymbol{M}^T\begin{bmatrix}\boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{M} \\
&= -\boldsymbol{M}_0^T\boldsymbol{A}\boldsymbol{M}_0,
\end{aligned} \tag{B14}$$

where each element of $\boldsymbol{A} \equiv [\mathrm{d}\boldsymbol{\phi}']^{-\frac{1}{2}}\left[[\mathrm{d}\boldsymbol{\phi}']\boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}] + [\mathrm{d}\boldsymbol{O}]\boldsymbol{C}^{-1}\mathrm{d}\boldsymbol{O}^T\right][\mathrm{d}\boldsymbol{\phi}']^{-\frac{1}{2}}$ is $O(1)$, and these elements are scattered to $N \times N$ elements by $\boldsymbol{M}$.

Next, we consider the orders of Eqs. (69)-(73):

$$\mathrm{d}\omega_{ij} = \rho[\boldsymbol{\omega} - \boldsymbol{\omega}\boldsymbol{\lambda}\boldsymbol{\omega}]_{ij}\mathrm{d}\alpha, \tag{B15}$$

$$\begin{aligned}
\mathrm{d}\boldsymbol{\Omega}_{ij} = {}& \rho\boldsymbol{C}^{-\frac{1}{2}}\left[\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\omega} + \tilde{\omega}\tilde{\lambda}\tilde{\omega} - \tilde{\boldsymbol{\Omega}}'\tilde{\lambda}\tilde{\boldsymbol{\Omega}}''\right]_{ij}\boldsymbol{C}^{-\frac{1}{2}}\mathrm{d}\alpha \\
&-\boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]\boldsymbol{\Omega}_{ij} - \boldsymbol{\Omega}_{ij}[\mathrm{d}\boldsymbol{C}-\mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} + \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\boldsymbol{\Omega}''_{ij} + \boldsymbol{\Omega}'_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{U}]\boldsymbol{C}^{-1},
\end{aligned} \tag{B16}$$

$$\mathrm{d}\boldsymbol{\Omega}'_{ij} = \rho \boldsymbol{C}^{-\frac{1}{2}} \left[\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}} - \tilde{\boldsymbol{\Omega}}'\tilde{\boldsymbol{\lambda}}\boldsymbol{\Omega}''' \right]_{ij} \mathrm{d}\alpha$$
$$-\boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{C} - \mathrm{d}\boldsymbol{O}]\boldsymbol{\Omega}'_{ij} + \boldsymbol{\Omega}'_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{V}] \quad + \boldsymbol{C}^{-1}[\mathrm{d}\boldsymbol{U}]^T\boldsymbol{V}\boldsymbol{\Omega}'''_{ij} + \boldsymbol{\Omega}_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}], \tag{B17}$$

$$\mathrm{d}\boldsymbol{\Omega}''_{ij} = \rho \quad \left[\tilde{\boldsymbol{\Omega}}'' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}} - \tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\Omega}}'' - \boldsymbol{\Omega}'''\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\Omega}}'' \right]_{ij} \boldsymbol{C}^{-\frac{1}{2}} \mathrm{d}\alpha$$
$$+ \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}\boldsymbol{\Omega}''_{ij} - \boldsymbol{\Omega}''_{ij}[\mathrm{d}\boldsymbol{C} - \mathrm{d}\boldsymbol{O}]^T\boldsymbol{C}^{-1} + \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}_{ij} + \boldsymbol{\Omega}'''_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{U}]\boldsymbol{C}^{-1}, \tag{B18}$$

$$\mathrm{d}\boldsymbol{\Omega}'''_{ij} = \rho \quad \left[\boldsymbol{\Omega}''' - \tilde{\boldsymbol{\Omega}}''\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Omega}}' - \tilde{\boldsymbol{\omega}}\tilde{\boldsymbol{\lambda}}\boldsymbol{\Omega}''' - \boldsymbol{\Omega}'''\tilde{\boldsymbol{\lambda}}\tilde{\boldsymbol{\omega}} - \boldsymbol{\Omega}'''\tilde{\boldsymbol{\lambda}}\boldsymbol{\Omega}''' \right]_{ij} \mathrm{d}\alpha$$
$$+ \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{V}\boldsymbol{\Omega}'''_{ij} + \boldsymbol{\Omega}'''_{ij}\boldsymbol{V}^T[\mathrm{d}\boldsymbol{V}] \quad + \quad [\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}'_{ij} + \boldsymbol{\Omega}''_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}]. \tag{B19}$$

Each element of the matrices $\boldsymbol{C}, \mathrm{d}\boldsymbol{C}$, and $\mathrm{d}\boldsymbol{O}$ are considered to be $O(1)$ with respect to N. The $\boldsymbol{\lambda}$ and $\boldsymbol{\Lambda}$ are also considered to be $O(1)$. Matrix normalization denoted with tilde (e.g., $\tilde{\boldsymbol{\Omega}}$) is considered to not change the order. Matrix size extension (e.g., $\tilde{\omega}$) is also considered to not change the order. As the initial value of $\boldsymbol{H}$ is defined as an unit matrix, we can let $\boldsymbol{\omega} = \boldsymbol{I}, \boldsymbol{\Omega} = \boldsymbol{0}, \boldsymbol{\Omega}' = \boldsymbol{0}, \boldsymbol{\Omega}'' = \boldsymbol{0}$, and $\boldsymbol{\Omega}''' = \boldsymbol{0}$ as initial values. Then, we notice the following. 1) $\boldsymbol{\omega}$ is $O(1)$ from the initial state. 2) Then, $\boldsymbol{\Omega}$ soon becomes $O(1)$, because $\mathrm{d}\boldsymbol{\Omega}_{ij}$ has a $O(1)$ term: $-\tilde{\omega} + \tilde{\omega}\tilde{\lambda}\tilde{\omega}$. 3) Then, $\boldsymbol{\Omega}'$ and $\boldsymbol{\Omega}''$ soon become $O(\frac{1}{\sqrt{N}})$, because $\mathrm{d}\boldsymbol{\Omega}'_{ij}$ and $\mathrm{d}\boldsymbol{\Omega}''_{ij}$ have a $O(\frac{1}{\sqrt{N}})$ term: $\boldsymbol{\Omega}_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}]$ and $[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}_{ij}$, respectively. 4) Then, $\boldsymbol{\Omega}'''$ soon becomes $O(\frac{1}{N})$, because $\mathrm{d}\boldsymbol{\Omega}'''_{ij}$ has a $O(\frac{1}{N})$ term: $[\mathrm{d}\boldsymbol{V}]^T\boldsymbol{U}\boldsymbol{\Omega}'_{ij} + \boldsymbol{\Omega}''_{ij}\boldsymbol{U}^T[\mathrm{d}\boldsymbol{V}]$. 5) There is no contradiction if we assume these orders are preserved.

APPENDIX C: EFFECT OF $\boldsymbol{E}$

We find that the fluctuation term $\boldsymbol{E}$ defined by Eq. (50) as

$$\boldsymbol{E}_{ij} \equiv [\boldsymbol{U}, \boldsymbol{V}] \begin{bmatrix} \boldsymbol{0} & \boldsymbol{\Omega}'_{ij} \\ \boldsymbol{\Omega}''_{ij} & \boldsymbol{\Omega}'''_{ij} \end{bmatrix} [\boldsymbol{U}, \boldsymbol{V}]^T, \tag{C1}$$

does not affect the usual order parameter dynamics under the assumption that each element of $\boldsymbol{\Omega}'_{ij}, \boldsymbol{\Omega}'_{ij}$ and $\boldsymbol{\Omega}'''_{ij}$ are $O(\frac{1}{\sqrt{N}}), O(\frac{1}{\sqrt{N}})$, and $O(\frac{1}{N})$, respectively (see Appendix B).

Here, $\boldsymbol{E}$ is negligible with respect to the dynamics of $\boldsymbol{R}$ and $\boldsymbol{S}$ in Eq. (47). This is because all the terms including $\boldsymbol{E}$ are 0 as shown

$$\left\langle \delta_k \boldsymbol{\xi}^T \boldsymbol{E}_{ik}^T \boldsymbol{B}_j \right\rangle_{\boldsymbol{\xi}} = \left\langle \delta_k \boldsymbol{\xi}^T \boldsymbol{V} \boldsymbol{\Omega}''^T_{ik} \boldsymbol{U}^T \boldsymbol{B}_j \right\rangle_{\boldsymbol{\xi}}$$
$$= \left\langle \check{\epsilon}(\boldsymbol{z}) g'(x_k) \right\rangle_{\boldsymbol{z}} \left\langle \boldsymbol{v}^T \right\rangle_{\boldsymbol{v}} \boldsymbol{\Omega}''^T_{ik} \boldsymbol{U}^T \boldsymbol{B}_j = 0,$$
$$\left\langle \delta_k \boldsymbol{\xi}^T \boldsymbol{E}_{ik}^T \boldsymbol{J}_j \right\rangle_{\boldsymbol{\xi}} = \left\langle \delta_k \boldsymbol{\xi}^T \boldsymbol{V} \boldsymbol{\Omega}''^T_{ik} \boldsymbol{U}^T \boldsymbol{J}_j \right\rangle_{\boldsymbol{\xi}}$$
$$= \left\langle \check{\epsilon}(\boldsymbol{z}) g'(x_k) \right\rangle_{\boldsymbol{z}} \left\langle \boldsymbol{v}^T \right\rangle_{\boldsymbol{v}} \boldsymbol{\Omega}''^T_{ik} \boldsymbol{U}^T \boldsymbol{J}_j = 0, \tag{C2}$$

where $\boldsymbol{v} \equiv \boldsymbol{V}^T\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is independent of $\boldsymbol{z} \equiv \boldsymbol{U}^T\boldsymbol{\xi} = [x_1, .., x_K, y_1, .., y_M]^T$, and $\langle \boldsymbol{v} \rangle_{\boldsymbol{v}} = \boldsymbol{0}$.

$\boldsymbol{E}$ is also negligible with respect to the dynamics of $\boldsymbol{\phi}$. All the terms including $\boldsymbol{E}$ can be expressed as

$$\frac{1}{N} \left\langle \delta_k \delta_l \boldsymbol{\xi}^T [\boldsymbol{U}, \boldsymbol{V}] \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}' \\ \boldsymbol{A}'' & \boldsymbol{A}''' \end{bmatrix} [\boldsymbol{U}, \boldsymbol{V}]^T \boldsymbol{\xi} \right\rangle_{\boldsymbol{\xi}}, \tag{C3}$$

where

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}' \\ \boldsymbol{A}'' & \boldsymbol{A}''' \end{bmatrix} = \boldsymbol{H}_{ik}\boldsymbol{H}_{jl} - [\omega_{ik} + \boldsymbol{U}\boldsymbol{\Omega}_{ik}\boldsymbol{U}^T][\omega_{jl} + \boldsymbol{U}\boldsymbol{\Omega}_{jl}\boldsymbol{U}^T] \tag{C4}$$

Then, we notice that $\left\langle \delta_k \delta_l \boldsymbol{\xi}^T \boldsymbol{U}\boldsymbol{A}\boldsymbol{U}\boldsymbol{\xi} \right\rangle_{\boldsymbol{\xi}}$ is $O(1)$, while $\left\langle \delta_k \delta_l \boldsymbol{\xi}^T \boldsymbol{U}\boldsymbol{A}'\boldsymbol{V}\boldsymbol{\xi} \right\rangle_{\boldsymbol{\xi}}$ and $\left\langle \delta_k \delta_l \boldsymbol{\xi}^T \boldsymbol{V}\boldsymbol{A}''\boldsymbol{U}\boldsymbol{\xi} \right\rangle_{\boldsymbol{\xi}}$ are 0. Moreover, $\left\langle \delta_k \delta_l \boldsymbol{\xi}^T \boldsymbol{V}\boldsymbol{A}'''\boldsymbol{V}\boldsymbol{\xi} \right\rangle_{\boldsymbol{\xi}}$ is at most $O(1)$ because each element of $\boldsymbol{A}'''$,

$$\boldsymbol{A}''' = \omega_{ik}\boldsymbol{\Omega}'''_{jl} + \omega_{jl}\boldsymbol{\Omega}'''_{ik} + \boldsymbol{\Omega}''_{ik}\boldsymbol{C}\boldsymbol{\Omega}'_{jl} + \boldsymbol{\Omega}'''_{ik}\boldsymbol{C}\boldsymbol{\Omega}'''_{ik}, \tag{C5}$$

is $O(\frac{1}{N})$. As all the terms in Eq. (C3) are scaled by $\frac{1}{N}$, we can ignore them. Therefore, we can completely ignore the effect of $\boldsymbol{E}$ in the usual order parameter dynamics.

APPENDIX D: DETAILED CONDITIONS FOR NUMERICAL RESULTS

For numerical results, we considered a realizable case, in which the numbers of the hidden units for both the teacher and student networks were set to two ($K = M = 2$). With respect to the order parameter dynamics, the initial conditions of the usual order parameters were set as follows. The square lengths of all teacher weight vectors T_{ii} were set to 1, while the angle between the teacher weight vectors, $\kappa \equiv \arccos \frac{T_{1,2}}{\sqrt{T_{1,1}T_{2,2}}}$, was set to a moderately correlated value, $\pi/8$, unless otherwise stated. The initial conditions with respect to the student weight vectors were determined according to the corresponding expected values of random choice $\boldsymbol{J} \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{N}\boldsymbol{I})$; i.e., $Q_{ii} = 1, Q_{ij} = 0$, and $R_{ii} = 0$. Only $R_{ij} (i \neq j)$, the inner products between the student weight vectors and non-corresponding teacher vectors, were set to a small negative value $R_{ii} – r_{ij}$ to break the permutation symmetry a little, where we adopted a 1 standard deviation (S.D.) rule; i.e.,

$$r_{ij} \equiv \sqrt{\mathrm{Var}_{\boldsymbol{J}}(R_{ii} - R_{ij})} = \sqrt{\frac{T_{ii} + T_{jj} - 2T_{ij}}{N}}, \tag{D1}$$

where $\mathrm{Var}_{\boldsymbol{J}}(\bullet)$ denotes the variance of $\bullet$ with respect to $\boldsymbol{J}$. Then, we solved the order parameter equations using the Runge-Kutta method with time interval $\Delta\alpha = 0.1$. With respect to our simulation, the initial states of $\boldsymbol{B}$ and $\boldsymbol{J}$ were determined to satisfy the conditions of the order parameters above.

[1] D. Saad and S. A. Solla: Phys. Rev. Lett. **74** 4337 (1995).
D. Saad and S. A. Solla: Phys. Rev. E **52** 4225 (1995).
[2] P. Riegler and M. Biehl: J. Phys. A **28** L507 (1995).
[3] M. Inoue, H. Park and M. Okada: J. Phys. Soc. Jpn. **72** 805 (2003).
[4] S. Amari: Neural Comput. **10** 251 (1998).
[5] H. H. Yang and S. Amari: Neural Comput. **10** 2137 (1998).
[6] M. Rattray, D. Saad and S. Amari: Phys. Rev. Lett. **81** 5461 (1998).
M. Rattray and D. Saad: Phys. Rev. E **59** 4523 (1999).
[7] Online Learning in Neural Networks, edited by D. Saad, Cambridge University Press, London (1998).
[8] S. Amari, H. Park and K. Fukumizu: Neural Comput. **12** 1399 (2000).
[9] H. Park, S. Amari and K. Fukumizu: Neural Networks **13** 755 (2000).
[10] J. Inoue, H Nishimori and Y. Kabashima: J. Phys. A **30** 3795 (1997).
[11] D. Saad and M. Rattray: Physical Review Letters **79** 2578 (1997).
M. Rattray and D. Saad: Physical Review E **58**, 6379 (1998).