A Solvable High-Dimensional Model of GAN

Chuang Wang[1,2]
wangchuang@ia.ac.cn

Hong Hu[2]
honghu@g.harvard.edu

Yue M. Lu[2]
yuelu@seas.harvard.edu

1. State Key Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Science, 95 Zhong Guan Cun Dong Lu, Beijing 100190, China
2. John A. Paulson School of Engineering and Applied Sciences, Harvard University
33 Oxford Street, Cambridge, MA 02138, USA

Abstract

We present a theoretical analysis of the training process for a single-layer GAN fed by high-dimensional input data. The training dynamics of the proposed model at both microscopic and macroscopic scales can be exactly analyzed in the high-dimensional limit. In particular, we prove that the macroscopic quantities measuring the quality of the training process converge to a deterministic process characterized by an ordinary differential equation (ODE), whereas the microscopic states containing all the detailed weights remain stochastic, whose dynamics can be described by a stochastic differential equation (SDE). This analysis provides a new perspective different from recent analyses in the limit of small learning rate, where the microscopic state is always considered deterministic, and the contribution of noise is ignored. From our analysis, we show that the level of the background noise is essential to the convergence of the training process: setting the noise level too strong leads to failure of feature recovery, whereas setting the noise too weak causes oscillation. Although this work focuses on a simple copy model of GAN, we believe the analysis methods and insights developed here would prove useful in the theoretical understanding of other variants of GANs with more advanced training algorithms.

1 Introduction

A generative adversarial network (GAN) [1] seeks to learn a high-dimensional probability distribution from samples. While there have been numerous advances on the application front [2–6], considerably less is known about the underlying theory and conditions that can explain or guarantee the successful trainings of GANs.

Recently, it has been a very active area of research to study either the equilibrium properties [7–9] or the training dynamics [10, 11]. Specifically, there is a line of works studying the dynamics of the gradient-based training algorithms *e.g.*, [11–16]. The basic idea is the following. The evolution of the learnable parameters in the training dynamics can be considered as a discrete-time process. With a proper time scaling, this discrete-time process converges to a deterministic continuous-time process as the learning rates tend to 0, which is characterized by an ordinary differential equation (ODE). By studying local stability of the ODE's fixed points, [12] shows that oscillation in the training algorithm is due to the eigenvalues of the Jacobian of the gradient vector field with zero real part and large imaginary part. Due to this fact, various stabilization approaches are proposed, for example adding additional regularizers [13, 14], and using two timescale [15] training. Very recently, [16] argues that those stabilization techniques may encourage the algorithms to converge non-Nash stationary points. All above works consider a small-learning-rates limit, where the limiting process

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

is always deterministic. The stochasticity and the effect of the noise is essentially ignored, which may not reflect practical situations. Thus, a new analysis paradigm to study the dynamics with the consideration of the intrinsic stochasticity is needed.

In this paper, we present a *high-dimensional* and *exactly solvable* model of GAN. Its dynamics can be precisely characterized at both macroscopic and microscopic scales, where the former is deterministic and the latter remains stochastic. Interestingly, our theoretical analysis shows that injecting additional noise can stabilize the training. Specifically, our main technical contributions are twofold:

- We present an asymptotically exact analysis of the training process of the proposed GAN model. Our analysis is carried out on both the *macroscopic* and the *microscopic* levels. The macroscopic state measures the overall performance of the training process, whereas the microscopic state contains all the detailed weights information. In the high-dimensional limit ($n \to \infty$), we show that the former converges to a deterministic process governed by an ordinary differential equation (ODE), whereas the latter stays stochastic described by a stochastic differential equation (SDE).
- We show that depending on the choice of the learning rates and the strength of noise, the training process can reach either a successful, a failed, an oscillating, or a mode-collapsing phase. By studying the stabilities of the fixed points of the limiting ODEs, we precisely characterize when each phase takes place. The analysis reveals a condition on the learning rates and the noise strength for successful training. We show that the level of the background noise is essential to the convergence of the training process: setting the noise level too strong (small signal-to-noise ratio) leads to failure of feature recovery, whereas setting the noise too weak (large signal-to-noise ratio) causes oscillation.

Our work builds upon a general analysis framework [17] for studying the scaling limits of high-dimensional exchangeable stochastic processes with applications to nonlinear regression problems. Similar techniques have also been used in the literature to study Monte Carlo methods [18], online perceptron learning [19, 20], online sparse PCA [21], subspace estimation [22], online ICA [23] and more recently, the supervised learning of two-layer neural networks [24], but to our best knowledge, this technique has not yet been used in analyzing GANs.

The rest of the paper is organized as follows. We present the proposed GAN model and the associated training algorithm in Section 2. Our main results are presented in Section 3, where we show that the macroscopic and microscopic dynamics of the training process converge to their respective limiting processes that are characterized by an ODE and SDE, respectively. In Section S-I, we analyze the stationary solutions of the limiting ODEs and precisely characterizes the long-term behaviors of the training process. We conclude in Section 5.

2 Formulations

In this section, we introduce the proposed GAN model and specify the associated training algorithm.

Model for the real data. In order to establish the theoretical analysis, we first impose a model for the probability distribution from which we draw our real data samples. We assume that the real data $\boldsymbol{y}_k \in \mathbb{R}^n$, $k = 0, 1, \ldots$ are drawn according to the following generative model:

$$\boldsymbol{y}_k = \mathcal{G}(\boldsymbol{c}_k, \boldsymbol{a}_k; \boldsymbol{U}, \eta_{\mathrm{T}}) \stackrel{\text{def}}{=} \boldsymbol{U}\boldsymbol{c}_k + \sqrt{\eta_{\mathrm{T}}}\boldsymbol{a}_k, \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times d}$ is a deterministic unknown feature matrix with d features; $\boldsymbol{c}_k \in \mathbb{R}^d$ is a random vector drawn from an unknown distribution $\mathcal{P}_{\boldsymbol{c}}$; $\boldsymbol{a}_k$ is an n-dimensional random vector acting as the background noise; and η_{T} is a parameter to control the strength of noise. Without loss of generality [1], we assume $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_d$, where $\boldsymbol{I}_d$ is the $d \times d$ identity matrix.

This generative model, referred to as the spiked covariance model [25] in the literature, is commonly used in the theoretical study of principal component analysis (PCA). We note that this model is not a trivial task for PCA even when $d = 1$ if the variance of the noise $\boldsymbol{a}_k$ is a non-zero constant. As

[1] If $\boldsymbol{U}$ is not orthogonal, we can rewrite $\boldsymbol{U}\boldsymbol{c}$ in (1) as $(\boldsymbol{U}\boldsymbol{R})(\boldsymbol{R}^{-1}\boldsymbol{c})$, where $\boldsymbol{R}$ is a matrix that orthogonalizes and normalizes the columns of $\boldsymbol{U}$. We can then study an equivalent system where the new feature vector is $\boldsymbol{R}^{-1}\boldsymbol{c}$.

proved in [25], the best estimator can not perfectly recover the signal $\boldsymbol{U}$ given an $\mathcal{O}(n)$ number of samples $\boldsymbol{y}_k$. Thus, it is of sufficient interest to investigate whether a GAN can retrieve informative results for the principal components in the same scaling limit.

The GAN model The GAN we are going to analyze is defined as follows. We assume that the generator $\mathcal{G}$ has the same linear structure as the real data model (1) given above:

$$\widetilde{\boldsymbol{y}}_k = \mathcal{G}(\widetilde{\boldsymbol{c}}_k, \widetilde{\boldsymbol{a}}_k; \boldsymbol{V}, \eta_{\text{G}}) \tag{2}$$

but the parameters are different. Here, $\widetilde{\boldsymbol{y}}_k$ denotes a fake sample produced by the generator; $\widetilde{\boldsymbol{a}}_k$ is an n-dimensional random noise vector; the random variable $\widetilde{\boldsymbol{c}}_k$ is drawn from a fixed distribution $\mathcal{P}_{\widetilde{\boldsymbol{c}}}$; η_{G} is the noise strength; and the matrix $\boldsymbol{V} \in \mathbb{R}^{n\times d}$ represents the parameters of the generator. (In an ideal case in which the generator learns the underlying true probability distribution perfectly, we have $\boldsymbol{V} = \boldsymbol{U}$.) Throughout the paper, we follow the notational convention that all the symbols that are decorated with a tilde (*e.g.*, $\widetilde{\boldsymbol{y}}_k$, $\widetilde{\boldsymbol{c}}_k$, $\widetilde{\boldsymbol{a}}_k$) denote quantities associated with the generator.

We define the discriminator $\mathcal{D}$ of our GAN model as

$$\mathcal{D}(\boldsymbol{y}; \boldsymbol{w}) \overset{\text{def}}{=} \widehat{D}(\boldsymbol{y}^\top \boldsymbol{w}).$$

Here, $\boldsymbol{y}$ is an input vector, which can be either the real data $\boldsymbol{y}_k$ from (1) or the fake one $\widetilde{\boldsymbol{y}}_k$ from (2); $\widehat{D} : \mathbb{R} \mapsto \mathbb{R}$ can be any function; and the vector $\boldsymbol{w} \in \mathbb{R}^n$ represents the parameters associated with the discriminator. Later, we will show that the generator can learn multiple features even though the discriminator only has one feature vector $\boldsymbol{w}$. Discriminators with multiple features can also be analyzed in a similar way, but in this paper we consider the single-feature discriminator for simplicity.

The training algorithm. The proposed GAN model has two set of parameters $\boldsymbol{V}$ and $\boldsymbol{w}$ to be learned from the data. The training process is formulated as the following MinMax problem

$$\min_{\boldsymbol{V}} \max_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{y}\sim\mathcal{P}(\boldsymbol{y};\boldsymbol{U})} \mathbb{E}_{\widetilde{\boldsymbol{y}}\sim\widetilde{\mathcal{P}}(\widetilde{\boldsymbol{y}},\boldsymbol{V})}\, \mathcal{L}(\boldsymbol{y}, \widetilde{\boldsymbol{y}}; \boldsymbol{w}), \tag{3}$$

where the two probability distributions $\mathcal{P}(\boldsymbol{y}; \boldsymbol{U})$ and $\widetilde{\mathcal{P}}(\widetilde{\boldsymbol{y}}; \boldsymbol{V})$ represent the distributions of the real data $\boldsymbol{y}$ and the fake data $\widetilde{\boldsymbol{y}}$ as specified by (1) and (2) respectively, and

$$\mathcal{L}(\boldsymbol{y}, \widetilde{\boldsymbol{y}}; \boldsymbol{w}) \overset{\text{def}}{=} F(\widehat{D}(\boldsymbol{y}^\top \boldsymbol{w})) - \widetilde{F}(\widehat{D}(\widetilde{\boldsymbol{y}}^\top \boldsymbol{w})) - \tfrac{\lambda}{2} H(\boldsymbol{w}^\top \boldsymbol{w}) + \tfrac{\lambda}{2} \text{tr}\big(H(\boldsymbol{V}^\top \boldsymbol{V})\big) \tag{4}$$

with $F(\cdot)$ and $\widetilde{F}(\cdot)$ being two functions that quantify the performance of the discriminator and $\lambda > 0$ being a constant. The function $H(\cdot)$ acts as a regularization term introduced to control the magnitude of the parameters $\boldsymbol{w}$ and $\boldsymbol{V}$. It can be an arbitrary real-valued function, which is applied element-wisely if the input is a matrix.

We consider a standard training algorithm that uses the vanilla stochastic gradient descent/ascent (SGDA) to seek a solution of (3). To simplify the theoretical analysis, we consider an online (*i.e.*, streaming) setting where each data sample $\boldsymbol{y}_k$ is used only once. At step k, the model parameters $\boldsymbol{w}_k$ and $\boldsymbol{V}_k$ are updated using a new real sample $\boldsymbol{y}_k$ and two fake samples $\widetilde{\boldsymbol{y}}_{2k}$ and $\widetilde{\boldsymbol{y}}_{2k+1}$, according to

$$\begin{aligned} \boldsymbol{w}_{k+1} &= \boldsymbol{w}_k + \tfrac{\tau}{n} \nabla_{\boldsymbol{w}_k} \mathcal{L}(\boldsymbol{y}_k, \widetilde{\boldsymbol{y}}_{2k}; \boldsymbol{w}_k) \\ \boldsymbol{V}_{k+1} &= \boldsymbol{V}_k - \tfrac{\widetilde{\tau}}{n} \nabla_{\boldsymbol{V}_k} \mathcal{L}\big(\boldsymbol{y}_k, \mathcal{G}(\widetilde{\boldsymbol{c}}_{2k+1}, \widetilde{\boldsymbol{a}}_{2k+1}; \boldsymbol{V}_k; \eta_{\text{G}}); \boldsymbol{w}_k\big), \end{aligned} \tag{5}$$

where $\widetilde{\boldsymbol{c}}_{2k+1}, \widetilde{\boldsymbol{a}}_{2k+1}$ are random variables that generates the fake sample $\widetilde{\boldsymbol{y}}_{2k+1}$ according to (2). The two parameters τ and $\widetilde{\tau}$ in the above expressions control the learning rates of the discriminator and the generator, respectively. In (5), we only consider a single-step update for $\boldsymbol{w}_k$. This is a special case of Algorithm 1 in [1] with the batch-size m set to 1. We note that the analysis presented in this paper can be naturally extended to the mini-batch case where m is a finite number.

Example 1. We define $F(\widehat{D}(x)) = \widetilde{F}(\widehat{D}(x)) = x^2/2$, and the regularizer function $H(\boldsymbol{A}) = \log\cosh(\boldsymbol{A} - \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix with the same dimension of $\boldsymbol{A}$, and the function $\log\cosh(\cdot)$ transforms the input matrix element-wisely. We use this specific regularizer to control the magnitude of the model parameters $\boldsymbol{V}$ and $\boldsymbol{w}$. In practice, any convex function with its minimum reached at zero would be fine. Our choice $\log\cosh(\boldsymbol{A} - \boldsymbol{I})$ here is is just a convenient special case since its derivative $H'(x) = \tanh(x)$ is smooth and bounded. Furthermore, we set the regularization parameter $\lambda \to \infty$, the original problem (3) becomes a constrained MinMax problem

$$\min_{\text{diag}(\boldsymbol{V}^\top \boldsymbol{V}) = \boldsymbol{I}_d} \max_{\|\boldsymbol{w}\|=1} \mathbb{E}_{\boldsymbol{y}\sim\mathcal{P}} \mathbb{E}_{\widetilde{\boldsymbol{y}}\sim\widetilde{\mathcal{P}}} \left[(\boldsymbol{y}^\top \boldsymbol{w})^2 - (\widetilde{\boldsymbol{y}}^\top \boldsymbol{w})^2 \right],$$

in which the diagonal operation $\mathrm{diag}(\boldsymbol{A})$ returns a matrix where the diagonal entries are the same as $\boldsymbol{A}$ and the off-diagonal entries are all zero. The condition $\mathrm{diag}(\boldsymbol{V}^\top \boldsymbol{V}) = \boldsymbol{I}_d$ ensures that each column vector of $\boldsymbol{V}$ is normalized.

3 Dynamics of the GAN

Definition 1. Let $\boldsymbol{X}_k \stackrel{\text{def}}{=} [\boldsymbol{U}, \boldsymbol{V}_k, \boldsymbol{w}_k] \in \mathbb{R}^{n\times(2d+1)}$. We call $\boldsymbol{X}_k$ the *microscopic state* of the training process at iteration step k.

The microscopic state $\boldsymbol{X}_k$ contains all the information about the training process. In fact, the sequence $\{\boldsymbol{X}_k\}_{k=0,1,2,\ldots}$ forms a Markov chain on $\mathbb{R}^{n\times(2d+1)}$. This can be easily verified from the update rule of $\boldsymbol{X}_k$ as defined in (5), in which the real data $\boldsymbol{y}_k$ and fake data $\widetilde{\boldsymbol{y}}_k$ are drawn according to (1) and (2) respectively. The Markov chain is driven by the initial state $\boldsymbol{X}_0$ and the sequence of random variables $\{(\boldsymbol{c}_k, \boldsymbol{a}_k, \widetilde{\boldsymbol{c}}_{2k}, \widetilde{\boldsymbol{a}}_{2k}, \widetilde{\boldsymbol{c}}_{2k+1}, \widetilde{\boldsymbol{a}}_{2k+1})\}_{k=0,1,2,\ldots}$.

Definition 2. Let $\boldsymbol{P}_k \stackrel{\text{def}}{=} \boldsymbol{U}^\top \boldsymbol{V}_k$, $\boldsymbol{q}_k \stackrel{\text{def}}{=} \boldsymbol{U}^\top \boldsymbol{w}_k$, $\boldsymbol{r}_k \stackrel{\text{def}}{=} \boldsymbol{V}_k^\top \boldsymbol{w}_k$, $\boldsymbol{S}_k \stackrel{\text{def}}{=} \boldsymbol{V}_k^\top \boldsymbol{V}_k$, and $z_k \stackrel{\text{def}}{=} \boldsymbol{w}_k^\top \boldsymbol{w}_k$. We call the tuple $\{\boldsymbol{P}_k, \boldsymbol{q}_k, \boldsymbol{r}_k, \boldsymbol{S}_k, z_k\}$ the *macroscopic state* of the Markov chain $\boldsymbol{X}_k$ at step k.

Those macroscopic quantities measure the cosine similarities among the feature vectors of the true model $\boldsymbol{U}$, the generator $\boldsymbol{V}_k$ and the discriminator $\boldsymbol{w}_k$. For example, the cosine of the angle between the ith true feature (*i.e.*, the ith column of $\boldsymbol{U}$) and the jth feature estimated in the generator (*i.e.*, the jth column of $\boldsymbol{V}_k$) is $[\boldsymbol{P}_k]_{i,j}/\sqrt{[\boldsymbol{S}_k]_{j,j}}$, where $[\boldsymbol{P}_k]_{i,j}$ is the inner product between the two feature vectors and $\sqrt{[\boldsymbol{S}_k]_{j,j}}$ is the norm of the jth column of $\boldsymbol{V}_k$. (The columns of $\boldsymbol{U}$ are unit vectors and need not be normalized here.) For simplicity, we introduce a compact notation for the macroscopic state:

$$\boldsymbol{M}_k \stackrel{\text{def}}{=} \boldsymbol{X}_k^\top \boldsymbol{X}_k = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{P}_k & \boldsymbol{q}_k \\ \boldsymbol{P}_k^\top & \boldsymbol{S}_k & \boldsymbol{r}_k \\ \boldsymbol{q}_k^\top & \boldsymbol{r}_k^\top & z_k \end{bmatrix}. \tag{6}$$

In what follows, we investigate the dynamics of the training algorithm (5) at both the macroscopic and the microscopic levels. At the macroscopic level, by examining the cosines of the angles, we study how closely the model parameters $\boldsymbol{V}_k$, $\boldsymbol{w}_k$ associated with the generator and discriminator can align with the ground truth feature vectors, *i.e.,* the columns of $\boldsymbol{U}$. At the microscopic level, we study how the elements in the matrix $\boldsymbol{V}_k$ and the vector $\boldsymbol{w}_k$ evolve as a stochastic process. As our analysis will reveal, the mechanisms behind the two levels are different: the macroscopic dynamics is asymptotically deterministic whereas the microscopic dynamics stays stochastic even as $n \to \infty$.

3.1 Macroscopic dynamics

We first study the asymptotic dynamics of the macroscopic state $\boldsymbol{M}_k$. Our theoretical analysis is carried out under the following assumptions.

- (A.1) The sequences of $\boldsymbol{c}_k \sim \mathcal{P}_{\boldsymbol{c}}$ and $\widetilde{\boldsymbol{c}}_k \sim \mathcal{P}_{\widetilde{\boldsymbol{c}}}$ for $k = 0, 1, \ldots$ are i.i.d. random variables with bounded moments of all orders, and $\{\boldsymbol{c}_k\}$ is independent of $\{\widetilde{\boldsymbol{c}}_k\}$.
- (A.2) The sequences $\{\boldsymbol{a}_k\}$ and $\{\widetilde{\boldsymbol{a}}_k\}$ for $k = 0, 1, \ldots$ are both independent Gaussian vectors with zero mean and the covariance matrix $\boldsymbol{I}_n$. Moreover, $\{\boldsymbol{a}_k\}$, $\{\widetilde{\boldsymbol{a}}_k\}$ are independent of $\{\boldsymbol{c}_k\}$ and $\{\widetilde{\boldsymbol{c}}_k\}$.
- (A.3) The first-order derivative of $H(\cdot)$ and the derivatives up to fourth order of the functions $F(\widehat{D}(\cdot))$ and $\widetilde{F}(\widehat{D}(\cdot))$ exist and they are also uniformly bounded.
- (A.4) Let $[\boldsymbol{U}, \boldsymbol{V}_0, \boldsymbol{w}_0]$ be the initial microscopic state. For $i = 1, 2, \ldots, n$, we have $\mathbb{E}\,[\sum_{\ell=1}^d ([\boldsymbol{U}]_{i,\ell}^4 + [\boldsymbol{V}_0]_{i,\ell}^4 + [\boldsymbol{w}_0]_i^4]) \le C/n^2$, where C is a constant not depending on n.
- (A.5) The initial macroscopic state $\boldsymbol{M}_0$ satisfies $\mathbb{E}\,\|\boldsymbol{M}_0 - \boldsymbol{M}_0^*\| \le C/\sqrt{n}$, where $\boldsymbol{M}_0^*$ is a deterministic matrix and C is a constant not depending on n.

We provide a few remarks on the above assumptions. In Assumption (A.1), $\mathcal{P}_{\boldsymbol{c}}$ and $\mathcal{P}_{\widetilde{\boldsymbol{c}}}$ can be different. For example, $\boldsymbol{c}$ is Gaussian, and $\widetilde{\boldsymbol{c}}$ is uniform on $[-1, 1]^d$. The assumption (A.2) can

be relaxed to non-Gaussian cases as long as all moments of $\boldsymbol{a}_k$ and $\widetilde{\boldsymbol{a}}_k$ are bounded, but we use Gaussian assumption here to simplify the proof. The assumption (A.4) requires that the elements in the parameter matrix of real data $\boldsymbol{U}$ and initial microscopic state $\boldsymbol{X}_0$ are $\mathcal{O}(1/\sqrt{n})$ numbers. Intuitively, this assumption ensures that $\boldsymbol{U}$ and $\boldsymbol{X}_0$ are generic matrices with $\mathcal{O}(1)$ Frobenius norms (*i.e.,* not the matrices that most elements are zeros and only few elements are large numbers). The assumption (A.5) ensures that the initial macroscopic states converges to a deterministic value as the system size n goes to infinity. The following theorem proves that if the initial state is convergent, then the whole training process converges to a deterministic process as $n \to \infty$, which is characterized by an ODE.

Theorem 1. *Fix $T > 0$. It holds under Assumptions (A.1)–(A.5) that*

$$\max_{0\le k\le nT} \mathbb{E}\left\|\boldsymbol{M}_k - \boldsymbol{M}\left(\tfrac{k}{n}\right)\right\| \le \tfrac{C(T)}{\sqrt{n}}, \tag{7}$$

where $C(T)$ is a constant that depends on T but not on n, and $\boldsymbol{M}(t) = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{P}_t & \boldsymbol{q}_t \\ \boldsymbol{P}_t^\top & \boldsymbol{S}_t & \boldsymbol{r}_t \\ \boldsymbol{q}_t^\top & \boldsymbol{r}_t^\top & z_t \end{bmatrix} \in \mathbb{R}^{(2d+1)\times(2d+1)}$ is a deterministic function. Moreover, $\boldsymbol{M}(t)$ is the unique solution of the following ODE:

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}_t &= \widetilde{\tau}\big(\boldsymbol{q}_t\widetilde{\boldsymbol{g}}_t^\top + \boldsymbol{P}_t\boldsymbol{L}_t\big)\\
\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}_t &= \tau\big(\boldsymbol{g}_t - \boldsymbol{P}_t\widetilde{\boldsymbol{g}}_t + \boldsymbol{q}_t h_t\big)\\
\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r}_t &= \tau\big(\boldsymbol{P}_t^T\boldsymbol{g}_t - \boldsymbol{S}_t\widetilde{\boldsymbol{g}}_t + \boldsymbol{r}_t h_t\big) + \widetilde{\tau}\big(z_t\widetilde{\boldsymbol{g}}_t + \boldsymbol{L}_t\boldsymbol{r}_t\big)\\
\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{S}_t &= \widetilde{\tau}\big(\boldsymbol{r}_t\widetilde{\boldsymbol{g}}_t^\top + \widetilde{\boldsymbol{g}}_t\boldsymbol{r}_t^\top + \boldsymbol{S}_t\boldsymbol{L}_t + \boldsymbol{L}_t\boldsymbol{S}_t\big)\\
\tfrac{\mathrm{d}}{\mathrm{d}t}z_t &= 2\tau(\boldsymbol{q}_t^\top\boldsymbol{g}_t - \boldsymbol{r}_t^\top\widetilde{\boldsymbol{g}}_t + z_t h_t) + \tau^2 b_t
\end{aligned} \tag{8}$$

with the initial condition $\boldsymbol{M}(0) = \boldsymbol{M}_0^*$, where

$$\begin{aligned}
&\boldsymbol{g}_t = \big\langle \boldsymbol{c} f(\boldsymbol{c}^\top\boldsymbol{q}_t + e\sqrt{z_t\eta_T})\big\rangle_{\boldsymbol{c},e},\ \widetilde{\boldsymbol{g}}_t = \big\langle \widetilde{\boldsymbol{c}}\widetilde{f}(\widetilde{\boldsymbol{c}}^\top\boldsymbol{r}_t + e\sqrt{z_t\eta_G})\big\rangle_{\widetilde{\boldsymbol{c}},e},\ \boldsymbol{L}_t = -\lambda \mathit{diag}(H'(\boldsymbol{S}_t))\\
&h_t = \big\langle f'(\boldsymbol{c}^\top\boldsymbol{q}_t + e\sqrt{z_t\eta_T})\big\rangle_{\boldsymbol{c},e} - \big\langle \widetilde{f}'(\widetilde{\boldsymbol{c}}^\top\boldsymbol{r}_t + e\sqrt{z_t\eta_G})\big\rangle_{\widetilde{\boldsymbol{c}},e} - \lambda H'(z_t),\\
&b_t = \eta_T\big\langle f^2(\boldsymbol{c}^\top\boldsymbol{q}_t + e\sqrt{z_t\eta_T})\big\rangle_{\boldsymbol{c},e} + \eta_G\big\langle \widetilde{f}^2(\widetilde{\boldsymbol{c}}^\top\boldsymbol{r}_t + e\sqrt{z_t\eta_G})\big\rangle_{\widetilde{\boldsymbol{c}},e}.
\end{aligned} \tag{9}$$

The two functions f, $\widetilde{f}$ stand for $f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(\widehat{D}(x))$ and $\widetilde{f}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\widetilde{F}(\widehat{D}(x))$, and f', $\widetilde{f}'$ and H' are derivatives of f, $\widetilde{f}$ and H respectively. The two constants η_T and η_G are the strength of the noise in the true data model and the generator, respectively. The brackets $\langle\cdot\rangle_{\boldsymbol{c},e}$ and $\langle\cdot\rangle_{\widetilde{\boldsymbol{c}},e}$ denote the averages over the random variables $\boldsymbol{c} \sim \mathcal{P}_{\boldsymbol{c}}$, $\widetilde{\boldsymbol{c}} \sim \mathcal{P}_{\widetilde{\boldsymbol{c}}}$, and $e \sim \mathcal{N}(0,1)$, where $\mathcal{P}_{\boldsymbol{c}}$ and $\mathcal{P}_{\widetilde{\boldsymbol{c}}}$ are the distributions involved in defining the generative model (1) and the generator (2).

This theorem implies that for each $k = \lfloor tn \rfloor$ for some $t \in [0, T]$, the macroscopic state $\boldsymbol{M}_k$ converges to a deterministic number $\boldsymbol{M}(t)$, and the convergence rate is $\mathcal{O}(1/\sqrt{n})$. The limiting ODE (8) for the macroscopic states involves $\mathcal{O}(d^2)$ variables, where d is the number of internal features often assumed to be a finite number that is much less than n. This ODE is essentially different from the ODE derived in the small-learning-rate limit [11–16], in which the number of variables is $\mathcal{O}(n)$.

The complete proof can be found in the Supplementary Materials. We briefly sketch the proof here. First, we note that $\boldsymbol{M}_k$ is a discrete-time stochastic process driven by the Markov chain $\boldsymbol{X}_k$. Then, we apply the martingale decomposition for $\boldsymbol{M}_k$ and get

$$\boldsymbol{M}_{k+1} - \boldsymbol{M}_k = \tfrac{1}{n}\phi(\boldsymbol{M}_k) + (\boldsymbol{M}_{k+1} - \mathbb{E}_k\,\boldsymbol{M}_{k+1}) + [\mathbb{E}_k\,\boldsymbol{M}_{k+1} - \boldsymbol{M}_k - \tfrac{1}{n}\phi(\boldsymbol{M}_k)],$$

where the matrix-valued function $\phi(\boldsymbol{M})$ represents the functions on the right hand sides of the ODE (8), and $\mathbb{E}_k$ denotes the conditional expectation given the state of the Markov chain $\boldsymbol{X}_k$. Finally, we show the martingale $\sum_{k'=0}^{k}(\boldsymbol{M}_{k'+1} - \mathbb{E}_{k'}\boldsymbol{M}_{k'})$ and the higher-order term $\mathbb{E}_k\,\boldsymbol{M}_{k+1} - \boldsymbol{M}_k - \frac{1}{n}\phi(\boldsymbol{M}_k)$ have no contribution when n goes to infinity.

Due to the limitation of our current proof, the constant $C(T)$ in (7) grows exponentially as T increases. This is not a problem for any finite T, but may cause some problem to study the long time behavior when $T \to \infty$. However, if we impose a sufficient large regularizer parameter λ to limit the norms of the microscopic weights $\boldsymbol{V}_k$ and $\boldsymbol{w}_k$, then the macroscopic state $\boldsymbol{M}_k$ is bounded

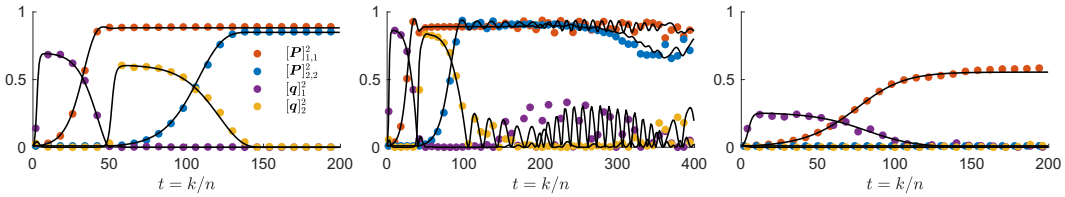

Figure 1: Macroscopic dynamics of the GAN with $d = 2$ features: $[\boldsymbol{P}_k]_{i,j}$ is the cosine of the angle between i'th column vector of the real feature matrix $\boldsymbol{U}_k$ and j'th column vector of the generator's weight matrix $\boldsymbol{V}_k$. Similarly, $[\boldsymbol{q}_k]_i$ is the cosine of angle between i'th column vector of $\boldsymbol{U}_k$ and the discriminator's weight vector $\boldsymbol{w}_k$. Colored dots are results from experiments, and the curves tracing these dots are our theoretical prediction by the ODE (8). From the left to right, the variance of background noise is $\eta_{\mathrm{T}} = \eta_{\mathrm{G}} = 2, 1, 4$ respectively, and other parameters are the same. The left figure is an example of successful training, where two features (red and blue dots) are retrieved by the generator. The center figure shows an oscillating training. It happens when noise are weak. The right figures shows a mode collapsing state, in which only the first feature are estimated by the generator.

as $[\boldsymbol{M}_k]^2_{i,j} \le [\boldsymbol{M}_k]_{i,i}[\boldsymbol{M}_k]_{j,j}$. In our experiments, $\lambda > 1$ is sufficient. In this case, the constant $C(T)$ is bounded not depending on T. In Example 1, when $\lambda \to \infty$, $[\boldsymbol{M}_k]_{i,i} = 1$, and therefore $[\boldsymbol{M}_k]^2_{i,j} \le 1$ and $C(T) \le (2d+1)^2$, where the number of features d is considered a constant not growing with n. This justifies the fixed points analysis of the ODE as discussed in Section S-I, which reflects the long-time training behavior. A better proof strategy to get rid of this dependence of T is also possible, *e.g.*, [26].

Numerical verification. We verify the theoretical prediction given by the ODE (8) via numerical simulations under the settings stated in Example 1. The results are shown in Figure 1. The number of features is $d = 2$, and $\boldsymbol{c}_k$ and $\widetilde{\boldsymbol{c}}_k$ are both Gaussian with zero mean and covariance $\mathrm{diag}([5, 3])$. The dimension is $n = 5,000$, and the learning rates of the generator and discriminator are $\widetilde{\tau} = 0.04$ and $\tau = 0.2$ respectively. After testing different noise strength $\eta_{\mathrm{T}} = \eta_{\mathrm{G}} = 2, 1, 4$, we have observed at least three nontrivial dynamical patterns: success, oscillating or mode collapsing. In all these experiments, our theoretical predictions match the actual trajectories of the macroscopic states pretty well.

Let us take a closer look at the successful case as shown in the left figure in Figure 1. The dynamics can be split into 4 stages. At the first stage, the discriminator learns the first feature of the true model. At this state, $[\boldsymbol{q}_t]_1$ quickly increases. At the second stage, the generator starts to learn the first feature and the discriminator is deceived. At this stage, $[\boldsymbol{P}_t]^2_{1,1}$ increases and $[\boldsymbol{q}_t]^2_1$ decreases. Once the discriminator completely forgets the first feature as $[\boldsymbol{q}_t]_1 \approx 0$, the third state begins. The discriminator starts to learn the second feature as $[\boldsymbol{q}_t]^2_2$ increases. Then, at the last stage, the generator learns the second feature and the discriminator is fooled again. In this region, $[\boldsymbol{P}_t]^2_{2,2}$ increases and $[\boldsymbol{q}_t]^2_2$ decreases down to 0. Eventually, the generators learns both features and the discriminator is completely fooled. It ends up at a stationary state that $\boldsymbol{q}_t = \boldsymbol{0}$ and $\boldsymbol{P}_t$ is nearly an identity matrix. Interestingly, this experiment shows that the generator learn features sequentially given a single-feature discriminator. This may be a reason why in practice, the discriminator's structure can be much simpler than the generator's.

3.2 Microscopic dynamics

In this section, we study how the elements in $\boldsymbol{X}_k = [\boldsymbol{U}, \boldsymbol{V}_k, \boldsymbol{w}_k]$ evolve during the training process. Instead of studying the trajectory of $\boldsymbol{X}_k$, we study the evolution of the *empirical measure* of the microscopic states, which is defined as

$$\mu_k(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \delta\big([\widehat{\boldsymbol{u}}^\top, \widehat{\boldsymbol{v}}^\top, \widehat{w}] - \sqrt{n}[[\boldsymbol{U}]_{i,:}, [\boldsymbol{V}_k]_{i,:}, [\boldsymbol{w}]_i]\big)$$

where $\delta(\cdot)$ is a Dirac measure on $\mathbb{R}^{2d+1}$ and $[\boldsymbol{U}]_{i,:}, [\boldsymbol{V}_k]_{i,:}$ are ith row of $\boldsymbol{U}$ and $\boldsymbol{V}_k$ respectively. The scaling factor $\sqrt{n}$ in the Dirac measures is introduced because $[\boldsymbol{U}]_{i,\ell}$, $[\boldsymbol{V}_k]_{i,\ell}$ and $[w_k,]_i$ are $\mathcal{O}(1/\sqrt{n})$ quantities.

We next embed the discrete-time measure-valued stochastic process μ_k into a continuous-time process by defining $\mu_t^{(n)} \stackrel{\text{def}}{=} \mu_k(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w})$ with $k = \lfloor nt \rfloor$ .Following the general technical approach presented in [17], we can show that under the same assumptions as Theorem 1, given $T > 0$, the sequence of measure-valued process $\{\{\mu_t^{(n)}\}_{t\in[0,T]}\}_n$ converges weakly to a deterministic process $\{\mu_t\}_{t\in[0,T]}$. In addition, μ_t is the measure of the solution to the stochastic differential equation

$$\begin{aligned} \mathrm{d}\widehat{\boldsymbol{u}}_t &= 0 \\ \mathrm{d}\widehat{\boldsymbol{v}}_t &= \widetilde{\tau}\big(\widehat{w}_t\widetilde{\boldsymbol{g}}_t + \boldsymbol{L}_t\widehat{\boldsymbol{v}}_t\big)\,\mathrm{d}t \\ \mathrm{d}\widehat{w}_t &= \tau\big(\widehat{\boldsymbol{u}}_t^\top \boldsymbol{g}_t + \widehat{\boldsymbol{v}}_t^\top\widetilde{\boldsymbol{g}}_t + \widehat{w}_t h_t\big)\,\mathrm{d}t + \tau\sqrt{b_t}\,\mathrm{d}B_t \end{aligned} \tag{10}$$

where $(\widehat{\boldsymbol{u}}_0, \widehat{\boldsymbol{v}}_0, \widehat{w}_0) \sim \mu_0$; B_t is the standard Brownian motion. The functions $\boldsymbol{g}_t$, $\widetilde{\boldsymbol{g}}_t$, $\boldsymbol{L}_t$, h_t and b_t are defined in (9), in which the macroscopic quantities $\boldsymbol{P}_t,\ \boldsymbol{S}_t,\ \boldsymbol{q}_t,\ z_t,\ \boldsymbol{r}_t$ are computed as follows

$$\boldsymbol{P}_t = \langle \mu_t, \widehat{\boldsymbol{u}}\widehat{\boldsymbol{v}}^\top\rangle, \quad \boldsymbol{S}_t = \langle \mu_t, \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top\rangle, \boldsymbol{q}_t \quad = \langle \mu_t, \widehat{\boldsymbol{u}}\widehat{w}\rangle, \quad z_t = \langle \mu_t, \widehat{w}^2\rangle, \quad \boldsymbol{r}_t = \langle \mu_t, \widehat{\boldsymbol{v}}\widehat{w}\rangle, \tag{11}$$

where $\langle \mu_t, \cdot\rangle$ denotes the expectation with respect to the measure μ_t.

The SDE (10) shows the intuitive meaning of the functions defined in (9): $\boldsymbol{g}_t$, $\widetilde{\boldsymbol{g}}_t$, $\boldsymbol{L}_t$, h_t are drift coefficients of the SDE and b_t is the diffusion coefficient of the SDE. We also note that if one follows the analysis in the small-learning-rate limit [11–16], one will get an ODE for the microscopic states. Compared to our SDE formula, the diffusion term $\tau\sqrt{b_t}dB_t$ is missing in those works, and therefore the effect of the noise can not be analyzed.

Moreover, the deterministic measure μ_t is unique solution of the following PDE (given in its weak form): for any bounded smooth test function $\varphi(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w})$,

$$\begin{aligned} &\tfrac{\mathrm{d}}{\mathrm{d}t}\big\langle \mu_t, \varphi(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w})\big\rangle = \\ &\widetilde{\tau}\big\langle \mu_t, \big(\widehat{w}\widetilde{\boldsymbol{g}}_t^\top + \widehat{\boldsymbol{v}}^\top\boldsymbol{L}_t\big)\nabla_{\widehat{\boldsymbol{v}}}\varphi\big\rangle + \tau\big\langle \mu_t, \big(\widehat{\boldsymbol{u}}^\top\boldsymbol{g}_t - \widehat{\boldsymbol{v}}^\top\widetilde{\boldsymbol{g}}_t + h_t\widehat{w}\big)\tfrac{\partial}{\partial\widehat{w}}\varphi\big\rangle + \tfrac{\tau^2}{2}b_t\big\langle \mu_t, \tfrac{\partial^2}{\partial\widehat{w}^2}\varphi\big\rangle \end{aligned} \tag{12}$$

where $\boldsymbol{q}_t$, $\boldsymbol{r}_t$, $\boldsymbol{S}_t$, and z_t are defined in (11), and the functions $\boldsymbol{g}_t$, $\widetilde{\boldsymbol{g}}_t$, b_t, h_t and $\boldsymbol{L}_t$ are defined in (9). We refer readers to [17] for a general framework for rigorously establishing the above scaling limit.

The connection between the microscopic and macroscopic dynamics can also be derived from the weak formulation of the PDE. Let φ being each element of $\widehat{\boldsymbol{u}}\widehat{\boldsymbol{v}}^\top,\ \widehat{\boldsymbol{u}}\widehat{w},\ \widehat{\boldsymbol{v}}\widehat{w},\ \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top,\ \widehat{w}^2$, and substituting those φ into the PDE (12), we can derive the ODE (8). In the setting of this paper, the macroscopic dynamics enjoys a closed ODE: We can predict the macroscopic states without solving the PDE nor SDE at microscopic scale. However, in a more general setting, e.g. when we add a regularizer other than the L2 type, the ODE itself may not be closed. In that case, one has to solve the PDE directly.

Numerical verification. We verify the predictions given by the PDE (12) by setting $d = 1$ using a special choice of the $(n \times 1)$-dimensional target feature matrix $\boldsymbol{U}$ whose elements are all $1/\sqrt{n}$ with $n = 10,000$. We also set the initial condition $\mu_0(\widehat{v}, \widehat{w}|\widehat{u} = 1)$ to be a Gaussian distribution. (When $d = 1$, the macroscopic quantities P_t, q_t, r_t, S_t reduce to scalars, so we remove their boldface here.) In this case, the PDE (12) admits a particularly simple analytical solution: at any time t, the solution $\mu_t(\widehat{v}, \widehat{w}|\widehat{u} = 1)$ is a Gaussian distribution whose mean and covariance matrix are given by $\mathbb{E}_{\mu_t(\widehat{v},\widehat{w}|\widehat{u}=1)}\begin{bmatrix}\widehat{v}\\ \widehat{w}\end{bmatrix} = \begin{bmatrix}P_t\\ q_t\end{bmatrix}$, $\mathbb{E}_{\mu_t(\widehat{v},\widehat{w}|\widehat{u}=1)}\begin{bmatrix}\widehat{v}\\ \widehat{w}\end{bmatrix}\begin{bmatrix}\widehat{v} & \widehat{w}\end{bmatrix} = \begin{bmatrix}S_t & r_t\\ r_t & z_t\end{bmatrix}$. Figure 2 overlays the contours of the probability distribution $\mu_t(\widehat{v}, \widehat{w}|\widehat{u} = 1)$ at different times t over the point clouds of the actual experiment data $(\sqrt{n}[\boldsymbol{w}_k]_i, \sqrt{n}[\boldsymbol{V}_k]_{i,1})$. We can see that the theoretical prediction given by (12) has excellent agreement with simulation results.

4 Local Stability Analysis of the ODE for the Macroscopic States

In this section, we study how the parameters, such as the learning rates τ and $\widetilde{\tau}$, noise strength η_{G} and η_{T} affect the training algorithm. We will focus on the concrete model as described in Example 1 so that we can have analytical solutions.

In order to further reduce the degrees of freedom of the ODE (8), we let the regularization parameter $\lambda \to \infty$. In this case, the vector $\boldsymbol{w}_k$ and all columns vectors of $\boldsymbol{V}_k$ are always normalized. Thus

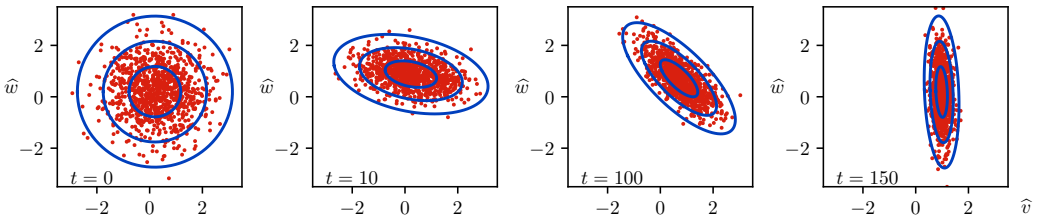

Figure 2: The evolution of the microscopic states at $t = 0$, 10, 100, and 150. For each fixed t, the red points in the corresponding figure represent the values of $(\widehat{v}, \widehat{w}) = (\sqrt{n}[\boldsymbol{V}_k]_{i,1}, \sqrt{n}[\boldsymbol{w}_k]_i)$ for $i = 1, 2, \ldots, n$, where $k = \lfloor nt \rfloor$. The blue ellipses illustrate the contours corresponding to one, two, and three standard deviations of the 2-D Gaussian distribution predicted by the PDE (12).

$z_k = 1$ and $[\boldsymbol{S}]_{i,i} = 1$. The macroscopic state is then described by $\boldsymbol{P}_k$, $\boldsymbol{q}_k$, $\boldsymbol{r}_k$ and off-diagonal terms of $\boldsymbol{S}_k$. Correspondingly, the ODE in Theorem 1 reduces to

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}_t &= \widetilde{\tau}\left(\boldsymbol{q}_t\boldsymbol{r}_t^\top\widetilde{\boldsymbol{\Lambda}} + \boldsymbol{P}_t\boldsymbol{L}_t\right) \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}_t &= \tau\big(\boldsymbol{\Lambda}\boldsymbol{q}_t - \boldsymbol{P}_t\widetilde{\boldsymbol{\Lambda}}\boldsymbol{r}_t + h_t\boldsymbol{q}_t\big) \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r}_t &= \tau\big(\boldsymbol{P}_t^T\boldsymbol{\Lambda}\boldsymbol{q}_t - \boldsymbol{S}_t\widetilde{\boldsymbol{\Lambda}}\boldsymbol{r}_t + h_t\boldsymbol{r}_t\big) + \widetilde{\tau}\big(\widetilde{\boldsymbol{\Lambda}} + \boldsymbol{L}_t\big)\boldsymbol{r}_t \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{S}_t &= \widetilde{\tau}\big(\boldsymbol{r}_t\boldsymbol{r}_t^\top\widetilde{\boldsymbol{\Lambda}} + \widetilde{\boldsymbol{\Lambda}}\boldsymbol{r}_t\boldsymbol{r}_t^\top + \boldsymbol{S}_t\boldsymbol{L}_t + \boldsymbol{L}_t\boldsymbol{S}_t\big) \end{cases} \tag{13}$$

where $\boldsymbol{\Lambda}$ and $\widetilde{\boldsymbol{\Lambda}}$ are the covariance matrices of the distributions $P_{\boldsymbol{c}}$ and $P_{\widetilde{\boldsymbol{c}}}$, respectively; and

$$h_t = (1 - \tfrac{\tau\eta_G}{2})\boldsymbol{r}_t^\top\widetilde{\boldsymbol{\Lambda}}\boldsymbol{r}_t - (1 + \tfrac{\tau\eta_T}{2})\boldsymbol{q}_t^\top\boldsymbol{\Lambda}\boldsymbol{q}_t - \tau\tfrac{\eta_G^2+\eta_T^2}{2}, \qquad \boldsymbol{L}_t = -\mathrm{diag}(\boldsymbol{r}_t\boldsymbol{r}_t^\top\widetilde{\boldsymbol{\Lambda}}), \tag{14}$$

in which η_{T} and η_{G} are the variance of noise in the true data model and generator, respectively. The derivation from the ODE (8) to (13) is presented in the Supplementary Materials.

Next, we discuss under what conditions, the GAN can reach a desirable training state by studying local stability of a particular type of fixed points of the ODE (13). The perfect estimation of the generator corresponds to $\boldsymbol{P}_t$ being an identity matrix (up to a permutation of rows and columns). A complete fail state relates to $\boldsymbol{P} = \boldsymbol{0}$. Furthermore, It is easy to verify that if $\boldsymbol{q}_t = \boldsymbol{r}_t = \boldsymbol{0}$, the ODE (13) will be stable for any $\boldsymbol{P}_t = \boldsymbol{P}$.

Claim 1. *The macroscopic states $\boldsymbol{P}_t$, $\boldsymbol{q} = \boldsymbol{r} = \boldsymbol{0}$ for all valid $\boldsymbol{P}_t$ are always the fixed points of the ODE* (13)*. Furthermore, a sufficient condition that the perfect estimation state $\boldsymbol{P}_t = \boldsymbol{I}$, $\boldsymbol{q} = \boldsymbol{r} = \boldsymbol{0}$ is locally stable and the failed state $\boldsymbol{P}_t = \boldsymbol{0}$, $\boldsymbol{q} = \boldsymbol{r} = \boldsymbol{0}$ is unstable if*

$$\tfrac{1}{2}\max_\ell\{\Lambda_\ell - \widetilde{\Lambda}_\ell + \alpha\widetilde{\Lambda}_\ell\} \le \tau\overline{\eta^2} < \min_\ell \Lambda_\ell, \tag{15}$$

where $\alpha = \frac{\widetilde{\tau}}{\tau}$, $\overline{\eta^2} = \frac{1}{2}(\eta_T^2 + \eta_G^2)$, and $\Lambda_\ell = [\boldsymbol{\Lambda}]_{\ell,\ell}$, $\widetilde{\Lambda}_\ell = [\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell}$.

The proof can be found in the Supplementary Materials. If the right inequality in (15) is violated, any feature ℓ with the signal-to-noise ratio $[\boldsymbol{\Lambda}]_{\ell,\ell} < \tau\overline{\eta^2}$ is not learned by the generator resulting *mode collapsing*. The right figure in Figure 1 demonstrates this situations, where only one of the two features is recovered. If the left inequality in (15) is violated, the training processes can be trapped in an *oscillation phase*. This phenomenon is shown in the middle figure in Figure 1. This result indicates that proper background noise can help to avoid oscillation and stabilize the training process. In fact, the trick of injecting additional noise has been used in practice to train multi-layer GANs [27]. To our best knowledge, our paper is the first theoretical study on why noise can have such a positive effect via a dynamic perspective.

In experiments, the training is not ended at the perfect recovery point due to the presence of the noise but converges at another fixed point nearby. This is because the perfect state is marginally stable, as the Jacobian matrix always has zero eigenvalues. It indicates that there are other locally stable fixed points near $\boldsymbol{P} = \boldsymbol{I}$. In fact, all points in the hyper-rectangle region satisfying $\boldsymbol{q} = \boldsymbol{r} = \boldsymbol{0}$ and $|p_\ell^*| \le |[\boldsymbol{P}]_{\ell,\ell}| \le 1$, $\forall\ \ell = 1, 2, \ldots, d$ are locally stable for some critical p_ℓ^*. In the matched case when $\Lambda_\ell = \widetilde{\Lambda}_\ell$, we have $p_\ell^* = \left[(\Lambda_\ell - \tau\overline{\eta^2})(\widetilde{\Lambda}_\ell + \tau\overline{\eta^2} - \alpha\widetilde{\Lambda}_\ell)/(\Lambda_\ell\widetilde{\Lambda}_\ell)\right]^{1/2}$, $\alpha = \frac{\widetilde{\tau}}{\tau}$ and $\overline{\eta^2} =$

$\frac{1}{2}(\eta_{\mathrm{T}}^2 + \eta_{\mathrm{G}}^2)$. Starting from a point near the origin, numerical solution of the ODE shows the training processes are ended up at the corner of this hyper-rectangle, *i.e.*, $\boldsymbol{P}^* = \mathrm{diag}(\{p_\ell^*,\ \ell = 1, 2, \ldots, d\})$. In the small-learning rate limit $\tau \to 0$ and the learning rate ratio $\alpha \to 0$, we get the perfect recovery $\boldsymbol{P}^* = \boldsymbol{I}$. The limit $\tau \to 0$, $\alpha \to 0$ was studied in the small-learning-rate analysis with the two-time scaling [15], and the result is consistent, but our analysis includes the situations with finite τ and α.

In addition, we provide a phase diagram analysis in a single-feature case $d = 1$ in the Supplementary Materials. All possible fixed points in this case are enumerated and their local stability is analyzed. This helps us understand the successful recovery condition (15), which is the intersection of the informative phases that each feature can be recovered individually.

5 Conclusion

We present a simple high-dimensional model for GAN with an exactly analyzable training process. Using the tool of scaling limits of stochastic processes, we show that the macroscopic state associated with the training process converges to a deterministic process characterized as the unique solution of an ODE, whereas the microscopic state remains stochastic described by an SDE, whose time-varying probability measure is described by a limiting PDE.

Indeed, it is a common picture in statistical physics that the macroscopic states of large systems tend to converge to deterministic values due to self-averaging. These notions, especially the mean-field dynamics, have been applied to analyzing neural networks both in shallow [19, 20] and deep models [28]. However, this mean-field regime was not considered in previous analyses of GAN. For example, a series of recent works *e.g.*, [11–16] considers a different scaling regime where the learning rate goes to zero but the system dimension n stays fixed. In that regime, the microscopic dynamics are deterministic even with the presence of the microscopic noise. In contrast, we study the regime where the learning rate is fixed but the dimension $n \to \infty$. This setting allows us to quantify the effect of training noise in the learning dynamics.

In this paper, we only consider a linear generator with a latent variable $\widetilde{\boldsymbol{c}}$ drawn from a fixed distribution $\mathcal{P}_{\widetilde{\boldsymbol{c}}}$, but our analysis can be extended to a more complex non-linear model with a learnable latent-variable distribution. Specifically, in order to compute derivatives w.r.t. $\mathcal{P}_{\widetilde{\boldsymbol{c}}}$, the latent variable $\widetilde{\boldsymbol{c}} \sim \mathcal{P}_{\widetilde{\boldsymbol{c}}}$ should be reparameterized by a deterministic function $\widetilde{\boldsymbol{c}} = f(\boldsymbol{z}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a learnable parameter and $\boldsymbol{z}$ is a random variable drawn from a simple and fixed distribution. For example, a Gaussian mixture with L equal-probability modes can be parameterized by $\widetilde{\boldsymbol{c}} = \sum_{\ell=1}^{L} (\boldsymbol{\mu}_\ell + \boldsymbol{\Sigma}_\ell \boldsymbol{\epsilon}_l)\beta_l$, where $\boldsymbol{\mu_\ell}$ and $\boldsymbol{\Sigma}_\ell$ are two learnable parameters representing the mean and covariance of the ℓth mode respectively, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$; β_ℓ is a random indicator variable where only one β_ℓ for $\ell = 1, 2, \ldots, L$ is 1 and the others are 0. In practice, $f(\boldsymbol{z}; \theta)$ is implemented by a multilayer neural network. Our analysis can be naturally extended to analyzing this model as long as the dimensions of $\widetilde{\boldsymbol{c}}$ and $\boldsymbol{\theta}$ keep finite when the data dimension n goes to infinity. More challenging situations, where the dimension of $\boldsymbol{\theta}$ is proportional to n, will be explored in future works.

Although our analysis is carried out in the asymptotic setting, numerical experiments show that our theoretical predictions can accurately capture the actual performance of the training algorithm at moderate dimensions. Our analysis also reveals several different phases of the training process that highly depend on the choice of the learning rates and noise strength. The analysis reveals a condition on the learning rates and the strength of noise to have successful training. Violating this condition results either oscillation or mode collapsing. Despite its simplicity, the proposed model of GAN provides a new perspective and some insights for the study of more realistic models and more involved training algorithms.

Acknowledgments This work was supported by the US Army Research Office under contract W911NF-16-1-0265 and by the US National Science Foundation under grants CCF-1319140, CCF-1718698, and CCF-1910410.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing System*, 2014, pp. 2672–2680.

[2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," *Proceedings of The 34th International Conference on Machine Learning*, pp. 1–32, 2017.

[3] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are gans created equal? a large-scale study," in *Advances in neural information processing systems*, 2018, pp. 698–707.

[4] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[6] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," *33rd International Conference on Machine Learning*, pp. 1060–1069, 2016.

[7] S. Arora, R. Ge, Y. Liang, and Y. Zhang, "Generalization and Equilibrium in Generative Adversarial Nets," in *International Conference on Machine Learning*, 2017, pp. 224–232.

[8] M. Arjovsky and L. Bottou, "Towards Principled Methods for Training Generative Adversarial Networks," *arXiv preprint arXiv:1701.04862*, 2017.

[9] S. Feizi, C. Suh, F. Xia, and D. Tse, "Understanding GANs: the LQG Setting," *arXiv:1710.10793*, 2017.

[10] J. Li, A. Madry, J. Peebles, and L. Schmidt, "Towards Understanding the Dynamics of Generative Adversarial Networks," *arXiv preprint arXiv:1706.09884*, 2017.

[11] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *International Conference on Machine Learning*, 2018, pp. 3478–3487.

[12] L. Mescheder, S. Nowozin, and A. Geiger, "The numerics of GANs," in *Advances in Neural Information Processing Systems*, 2017, pp. 1823–1833.

[13] V. Nagarajan and J. Z. Kolter, "Gradient descent GAN optimization is locally stable," in *Advances in Neural Information and Processing Systems*, 2017, pp. 5591–5600.

[14] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing Training of Generative Adversarial Networks through Regularization," in *Advances in Neural Information Processing Systems*, 2017, pp. 2015–2025.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6629–6640.

[16] E. V. Mazumdar, M. I. Jordan, and S. S. Sastry, "On finding local nash equilibria (and only local nash equilibria) in zero-sum games," *arXiv preprint arXiv:1901.00838*, 2019.

[17] C. Wang, J. Mattingly, and Y. M. Lu, "Scaling Limit: Exact and Tractable Analysis of Online Learning Algorithms with Applications to Regularized Regression and PCA," *arXiv preprint arXiv:1712.04332*, 2017.

[18] G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak convergence and optimal scaling of random walk Metropolis algorithms," *Annals of Applied Probability*, vol. 7, no. 1, pp. 110–120, 1997.

[19] D. Saad and S. A. Solla, "Exact Solution for On-Line Learning in Multilayer Neural Networks," *Phys. Rev. Lett.*, vol. 74, no. 21, pp. 4337–4340, 1995.

[20] M. Biehl and H. Schwarze, "Learning by on-line gradient descent," *Journal of Physics A*, vol. 28, no. 3, pp. 643–656, 1995.

[21] C. Wang and Y. M. Lu, "Online Learning for Sparse PCA in High Dimensions: Exact Dynamics and Phase Transitions," in *Information Theory Workshop (ITW), 2016 IEEE*, 2016, pp. 186–190.

[22] C. Wang, Y. C. Eldar, and Y. M. Lu, "Subspace estimation from incomplete observations: A high-dimensional analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1240–1252, Dec 2018.

[23] C. Wang and Y. M. Lu, "The Scaling Limit of High-Dimensional Online Independent Component Analysis," in *Advances in Neural Information Processing Systems*, 2017, pp. 6641–6650.

[24] S. Mei, A. Montanari, and P.-M. Nguyen, "A Mean Field View of the Landscape of Two-Layers Neural Networks," *arXiv preprint*, p. arXiv:1804.06561, 2018.

[25] I. Johnstone and A. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009.

[26] B. Jourdain, T. Lelièvre, and B. Miasojedow, "Optimal scaling for the transient phase of Metropolis Hastings algorithms: The longtime behavior," *Bernoulli*, vol. 20, no. 4, pp. 1930–1978, 2014.

[27] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," *International Conference on Learning Representations*, 2017.

[28] P.-M. Nguyen, "Mean field limit of the learning dynamics of multilayer neural networks," *arXiv preprint arXiv:1902.02880*, 2019.

[29] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.

[30] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

Supplementary Materials

These Supplementary Materials provide additional information, detailed derivations and proof of the results shown in the main text. Specifically, in Section S-I we provide a local stability analysis and draw the phase diagram in the case $d = 1$ and $d = 2$. In Section S-II, we present a heuristic derivation of the stochastic differential equation (SDE) for the microscopic states. Next, in Section S-III, we show a derivation of the ODE for the macroscopic states from the weak formulation of the PDE. We then establish the full proof of the Theorem 1 in Section S-IV. Finally, we present the local stability analysis of the ODE's fixed points in Section S-V.

Notation: Throughout the paper, we use $\boldsymbol{I}_d$ to denote the $d \times d$ identity matrix. Depending on the context, $\|\cdot\|$ denotes either the ℓ_2 norm of a vector or the spectral norm of a matrix. For any $x \in \mathbb{R}$, the floor operation $\lfloor x \rfloor$ gives the largest integer that is smaller than or equal to x. We denote $[\boldsymbol{v}]_i$ the ith element of the vector $\boldsymbol{v}$ and denote $[\boldsymbol{M}]_{i,j}$ the element at ith row and jth column of the matrix $\boldsymbol{M}$. Finally, $C(T)$ denotes a constant that depends on the terminal time T, and C denotes a general constant that does not depends on T and n. Both C and $C(T)$ can vary line to line.

S-I Phase diagram for the case $d = 1$ and $d = 2$

In what follows, we provide a thorough study of all the fixed points of the ODE (13) when the number of feature $d = 1$ and $d = 2$. In particular, three major phases are identified under different settings of the learning rates τ and $\widetilde{\tau}$ with the fixed model parameters η_{T}, η_{G}, $\boldsymbol{\Lambda}$, and $\widetilde{\boldsymbol{\Lambda}}$.

Phase diagram for $d = 1$. By analyzing the local stabilities of these fixed points as illustrated in Figure 3(a), we obtain the phase diagram as shown in Figure 3(b). For simplicity, we only present the result when $\eta_{\mathrm{T}} = \eta_{\mathrm{G}} = 1$, and $\boldsymbol{\Lambda} = \widetilde{\boldsymbol{\Lambda}}$, which is denoted by Λ used in the remaining part of this section. Detailed derivations are presented in S-V.

Even in this simplest case, we find there are in total 5 types of fixed points, the locations of which are visualized in the 3-dimensional space (P, q, r) shown in Figure 3(a). Each type of the fixed points has an intuitive meaning in terms of the two-player game between $\mathcal{G}$ and $\mathcal{D}$. We list the detailed information in Table 1, in which we define a function $\beta(\tau) = \begin{cases} [1 + (\frac{\Lambda}{2} - \frac{\Lambda}{\tau})^{-1}]^{-1}, & \text{if } \tau \le \frac{2\Lambda}{\Lambda+2} \\ +\infty, & \text{otherwise} \end{cases}$.

Noninformative phase: We say that the ODE (13) is in a noninformative phase if either a type-1 or type-2 fixed point in Table 1 is stable. In this case, $P = 0$, which indicates that the generator's parameter vector $\boldsymbol{V}$ has no correlation with the true feature vector $\boldsymbol{U}$. In Figure 3(b), the region labeled as noninfo-1 is the stable region for the type-1 fixed point, and noninfo-2 is the stable region for the type-2 fixed point. The two regions have no overlap. However, we note that in noninfo-1, the type-3 fixed points can also be stable, in which case the stationary point of the ODE is determined by the initial condition.

Informative phase: We say that the ODE (13) is in an informative phase if neither type-1 nor type-2 fixed point is stable, and if at least one fixed point of type-3 and type-5 is stable. In this case, it is guaranteed that P is nonzero, indicating that the generator can achieve non-vanishing correlation with the real feature vector. In addition, the stable regions for the type-3 and type-5 fixed points are disjoint. They are shown in Figure 3(b) as info-1 and info-2, respectively. The difference between the two region is that, in info-1, q is exactly 0 indicating that the discriminator is completely fooled, whereas in info-2, q is nonzero.

Oscillating phase: We say that the ODE (13) is in an oscillating phase if none of the fixed points in Table 1 is stable. In this phase, limiting cycles emerge and the system will oscillate on these cycles indefinitely. Moreover, we found two types of limiting cycles.

To further illustrate the phase transitions, we draw ODE trajectories and phase portraits in Figure 4 corresponding to different choices of the step sizes (from left to right, $\widetilde{\tau} = 0.03, 0.2, 0.4, 0.47$).

The two figures in the first column of Figure 4 show a case in the Info-1 phase. The bottom red dot in Figure 3.(b) represents this configuration of the step sizes, where $\widetilde{\tau}/\tau$ is small. The top figure of Figure 4.(a) shows the dynamics of P_t, q_t and r_t, and the bottom figure shows the phase portrait on

Table 1: List of the fixed points of the ODE (13) when $d = 1$ and $\Lambda = \widetilde{\Lambda}$.

Type	Location	Existence	Stable Region	Intuitive Interpretation
1	$P = q = 0$, $r = 0$	always	$\tau > \Lambda^2$, $\frac{\widetilde{\tau}}{\tau} < \frac{\tau+\Lambda}{\Lambda}$	Both $\mathcal{G}$ and $\mathcal{D}$ fail, and they are uncorrelated
2	$P = q = 0$ $r = \pm r^* \neq 0$	$\frac{\widetilde{\tau}}{\tau} \geq \frac{\tau+\Lambda}{\Lambda}$ or $\frac{\widetilde{\tau}}{\tau} \leq 1 - \frac{\tau}{2}$	$\max\{2, \frac{\tau+\Lambda}{\Lambda}\} \leq \frac{\widetilde{\tau}}{\tau} \leq \beta(\tau)$	Both $\mathcal{G}$ and $\mathcal{D}$ fail, and they are correlated
3	$q = r = 0$ $\lvert P\rvert \in (0, 1]$	always	$\lvert P\rvert = 1$ is stable if $\frac{\widetilde{\tau}}{\tau} \leq \min\{\frac{2\tau}{\Lambda}, \max\{\frac{\tau^2\Lambda^{-1}}{\lvert\tau-\Lambda\rvert}, 4\}\}$	$\mathcal{G}$ wins and $\mathcal{D}$ loses
4	$P = r = 0$ $q = \pm q^* \neq 0$	always	always unstable	$\mathcal{G}$ loses and $\mathcal{D}$ wins
5	None of P, q or r is zero	not always, at most 8 fixed points	can be computed numerically	Both $\mathcal{G}$ and $\mathcal{D}$ are informative

$P - q$ plane. Top figure of Figure 4.(a) shows an interesting phenomenon that dynamics are separated into two stages. At the first stage, q_t (red dots, cosine similarity between the true feature vector and discriminator's estimation) increases drastically from 0 to some value near 1, while P_t (blue dots, cosine similarity between the true feature vector and generator's estimation) almost doesn't change. Intuitively, at this stage, the discriminator learns the true model while the generator is unchanged. In the second stage, the generator start to fool the discriminator, where $|P_t|$ increases and q_t decreases. In fact, these two-stage dynamics can be understood from the ODE (13): When τ/τ is small, the process can be decomposed into two processes in different time scales. In particular, the discriminator is associated with the faster dynamics as $\tau \gg \widetilde{\tau}$, and the generator governs the slower dynamics. Figure 1 in the main text shows that this picture is still hold for multi-feature cases in the hierarchical dynamics.

The figures in the middle two columns of Figure 4 show the two types of limiting cycles that can emerge in the oscillating phase. The middle two red dots in Figure 3.(b) represents these configurations of the step sizes. The last column of Figure 4 shows another stable phase in Info-2. In this phase, τ/τ is relatively large. The two time-scale dynamics are mixed, and another type of stable fixed points emerges.

Phase diagram for $d = 2$. Figure 5 shows the phase diagram when $d = 2$. In particular, the two red lines between Info-1 and Noninfo-1 in Figure 5 are determined by the left inequality in (15). In Info-1, both feature vectors are recovered by the generator. The dynamics of this phase are shown in Figure 1.(a) in the main text. In the Half-info phase, only the feature vector with the larger signal-to-noise ratio is recovered. The dynamics of this phase are shown in Figure 1.(c) in the main text. The blue line between Info-1 and oscillating phases shows the boundary between oscillation state and stable state.

S-II Heuristic derivations of the dynamics of the microscopic states

In this section, we derive the stochastic differential equations (10) in the main text for the microscopic states in a non-rigorous way. Specifically, we directly discard higher-order terms without any justification, in order to highlight the main ideas. In Section S-IV, we rigorously justify these steps by providing bounds on those terms.

Our starting point is the iterative algorithm (5) in the main text. Substituting the objective function $\mathcal{L}$ defined in (4) into (5), we have

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \tfrac{\tau}{n}\big[\boldsymbol{y}_k f(\boldsymbol{y}_k^\top \boldsymbol{w}_k) - \widetilde{\boldsymbol{y}}_{2k}\widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top \boldsymbol{w}_k) - \lambda \boldsymbol{w}_k H'(\boldsymbol{w}_k^\top \boldsymbol{w}_k)\big] \tag{S-1}$$

$$\boldsymbol{V}_{k+1} = \boldsymbol{V}_k + \tfrac{\widetilde{\tau}}{n}\big[\boldsymbol{w}_k \widetilde{\boldsymbol{c}}_{2k+1}^\top \widetilde{f}(\widetilde{\boldsymbol{y}}_{2k+1}^\top \boldsymbol{w}_k) - \lambda \boldsymbol{V}_k \mathrm{diag}(H'(\boldsymbol{V}_k^\top \boldsymbol{V}_k))\big], \tag{S-2}$$

where $\boldsymbol{y}_k$ and $\widetilde{\boldsymbol{y}}_k$ are true and fake samples generated according to (1) and (2) respectively. The two functions f, $\widetilde{f}$ stand for $f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(\widehat{D}(x))$ and $\widetilde{f}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\widetilde{F}(\widehat{D}(x))$. The function H' is derivative of H. If the input of $H'(\cdot)$ is a matrix, H' applies to the input matrix element-wisely. The operation $\mathrm{diag}(\boldsymbol{A})$ is a diagonal matrix of $\boldsymbol{A}$, where the off-diagonal term are set to zero.

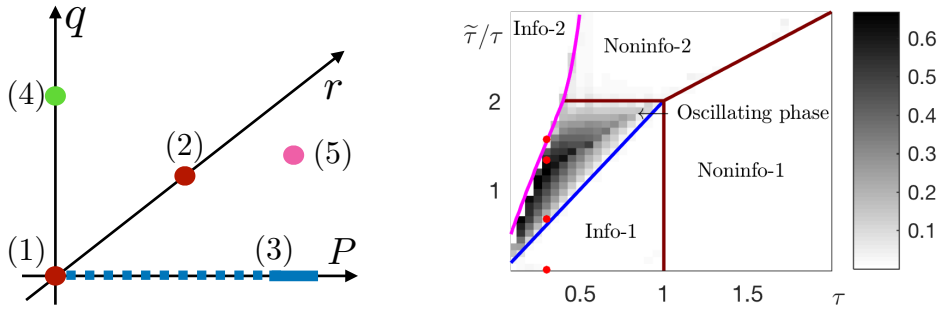

Figure 3: (a): The locations of the five types of fixed points of the ODE (13). Their properties are listed in Table 1. (b): The phase diagram for the stationary state of the ODE (13). The colored lines illustrate the theoretical prediction of the boundaries between the different phases. Simulations results for a single numerical experiment are also shown to illustrate the oscillating phase: Each grey square represents the value of $\frac{1}{200}\int_{800}^{1000}[(P_t-\langle P_t\rangle)^2+(q_t-\langle q_t\rangle)^2+(r_t-\langle r_t\rangle)^2]\,\mathrm{d}t$ where $\langle P_t\rangle=\frac{1}{200}\int_{800}^{1000}P_t\,\mathrm{d}t$, and $\langle q_t\rangle$ and $\langle r_t\rangle$ are defined similarly. Note that the above quantity measures the variation (over time) of the training process as it approaches steady states. We see that the variation is indeed nonzero in the oscillating phase (see Figure 4), whereas the variation is close to zero in all other phases.

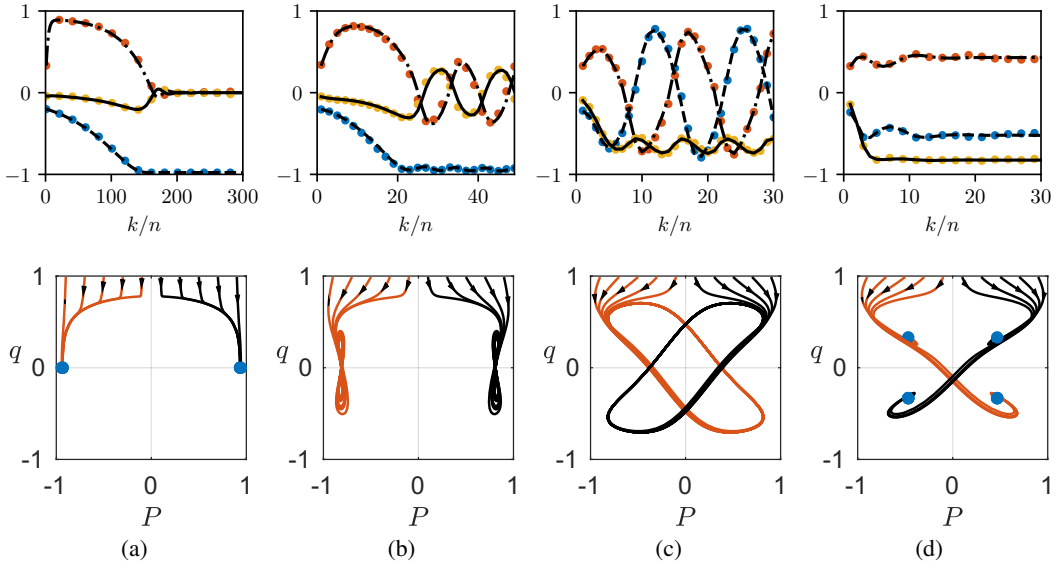

Figure 4: Macroscopic dynamics of Example 1 with $d=1$. In the first row, the red, blue and yellow dots represent P_t, q_t, and r_t respectively of the experimental results of a single trial. The black curves under the dots are theoretical predictions given by the ODE (13). We set a fix the discriminator's learning rate $\tau=0.3$ and vary the generator's learning rate $\widetilde{\tau}=0.03,\ 0.2,\ 0.4,\ 0.47$ from left to right column. These parameter settings are marked by the four red dots in the phase diagram in Figure 3. The second row is the phase portraits of the trajectories shown in the first row onto the P–q plane. Figure (a) shows a case in the phase of info-1, where a subset of type (3) fixed points are stable. Figure (b) and (c) are in the oscillating phase, and (d) is in info-2, where the fixed points of type-5 are stable. The blue dots in the figures show the stable fixed points.

We note that the elements of $\boldsymbol{w}_k$ and $\boldsymbol{V}_k$ are $\mathcal{O}(\frac{1}{\sqrt{n}})$ number as the norm of $\boldsymbol{w}_k$ and the norms of column vectors of $\boldsymbol{V}_k$ are all $\mathcal{O}(1)$ numbers. To investigate the dynamics of the microscopic state, it is convenient to rescale $\boldsymbol{w}_k$ and $\boldsymbol{V}_k$ by a factor of $\sqrt{n}$. We define $\widehat{\boldsymbol{u}}_i$ and $\widehat{\boldsymbol{v}}_{k,i}$ as the column view of

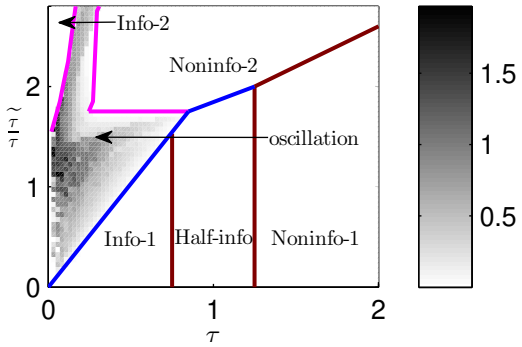

Figure 5: The phase diagram for the stationary states of the ODE (13) when $d = 2$. This phase diagram is generated by numerically computing the fixed points and eigenvalues of the Jacobian of the ODE (13).

the i'th row of the matrices $\sqrt{n}\boldsymbol{U}$ and $\sqrt{n}\boldsymbol{V}_k$ respectively, and $\widehat{w}_{k,i} \overset{\text{def}}{=} \sqrt{n}[\boldsymbol{w}_{k+1}]_i$. The update rule of $((\widehat{\boldsymbol{u}}_i, \widehat{\boldsymbol{v}}_{k,i}, \widehat{w}_{k,i})_{i=1,\dots,n})_{k=0,1,2,\dots}$ is

$$\widehat{w}_{k+1,i} - \widehat{w}_{k,i} = \frac{\tau}{n}\left[\left(\widehat{\boldsymbol{u}}_i^\top \boldsymbol{c}_k + \sqrt{n\eta_{\text{T}}}a_{k,i}\right) f_k - \left(\widehat{\boldsymbol{v}}_{k,i}^\top \widetilde{\boldsymbol{c}}_{2k} + \sqrt{n\eta_{\text{G}}}\widetilde{a}_{2k,i}\right) \widetilde{f}_{2k} - \lambda H'(z_k)\widehat{w}_{k,i}\right], \tag{S-3}$$

$$\widehat{\boldsymbol{v}}_{k+1,i} - \widehat{\boldsymbol{v}}_{k,i} = \frac{\widetilde{\tau}}{n}\left[\widehat{w}_{k,i}\widetilde{\boldsymbol{c}}_{2k+1}\widetilde{f}_{2k+1} - \lambda \text{diag}(H'(\boldsymbol{S}_k))\widehat{\boldsymbol{v}}_{k,i}\right], \tag{S-4}$$

where $a_{k,i}, \widetilde{a}_{k,i}$ are the ith elements of $\boldsymbol{a}_k$ and $\widetilde{\boldsymbol{a}}_k$ respectively, and f_k and $\widetilde{f}_k$ are shorthands for

$$f_k = f(\boldsymbol{y}_k^\top \boldsymbol{w}_k/\sqrt{n}) = f\left(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \sqrt{\tfrac{\eta_{\text{T}}}{n}}\textstyle\sum_{j=1}^n a_{k,j}\widehat{w}_{k,j}\right)$$
$$\widetilde{f}_k = \widetilde{f}(\widetilde{\boldsymbol{y}}_k^\top \boldsymbol{w}_{\lfloor k/2 \rfloor}/\sqrt{n}) = \widetilde{f}\left(\boldsymbol{r}_{\lfloor k/2 \rfloor}^\top \widetilde{\boldsymbol{c}}_k + \sqrt{\tfrac{\eta_{\text{G}}}{n}}\textstyle\sum_{j=1}^n \widetilde{a}_{k,j}\widehat{w}_{\lfloor k/2 \rfloor,j}\right),$$

respectively, and the empirical macroscopic quantities $\boldsymbol{q}_k$, $\boldsymbol{r}_k$, $\boldsymbol{z}_k$ and $\boldsymbol{S}_k$ are defined as follows

$$\begin{aligned}
\boldsymbol{q}_k &\overset{\text{def}}{=} \boldsymbol{U}^\top \boldsymbol{w}_k = \tfrac{1}{n}\textstyle\sum_{i=1}^n \widehat{\boldsymbol{u}}_i \widehat{w}_i, & \boldsymbol{r}_k &\overset{\text{def}}{=} \boldsymbol{V}_k^\top \boldsymbol{w}_k = \tfrac{1}{n}\textstyle\sum_{i=1}^n \widehat{\boldsymbol{v}}_{k,i}\widehat{w}_i, \\
z_k &\overset{\text{def}}{=} \boldsymbol{w}_k^\top \boldsymbol{w}_k = \tfrac{1}{n}\textstyle\sum_{i=1}^n \widehat{w}_{k,i}^2, & \boldsymbol{S}_k &\overset{\text{def}}{=} \boldsymbol{V}_k^\top \boldsymbol{V}_k = \tfrac{1}{n}\textstyle\sum_{i=1}^n \widehat{\boldsymbol{v}}_{k,i}\widehat{\boldsymbol{v}}_{k,i}^\top, \\
\boldsymbol{P}_k &\overset{\text{def}}{=} \boldsymbol{U}^\top \boldsymbol{V}_k = \tfrac{1}{n}\textstyle\sum_{i=1}^n \widehat{\boldsymbol{u}}_i \widehat{\boldsymbol{v}}_{k,i}^\top.
\end{aligned} \tag{S-5}$$

The matrix $\boldsymbol{P}_k$ is not used in this section, but we put it here with the other macroscopic quantities for future reference.

Now we derive (10) from (S-3) and (S-4).

First, it is trivial to get the first equation of the SDE $\mathrm{d}\widehat{\boldsymbol{u}}_t = 0$ in (10) in the main text, since $\widehat{\boldsymbol{u}}_i$ does not change over time.

Next, we derive the second equation in (10). Averaging over $\widetilde{\boldsymbol{c}}_{2k+1}$ and $\widetilde{\boldsymbol{a}}_{2k+1}$ on the both sides of (S-4), we get

$$\begin{aligned}
&\left\langle \widehat{\boldsymbol{v}}_{k+1,i} - \widehat{\boldsymbol{v}}_{k,i} \right\rangle_{\widetilde{\boldsymbol{c}}_{2k+1}, \widetilde{\boldsymbol{a}}_{2k+1}} \\
&= \frac{\widetilde{\tau}}{n}\left[\left\langle \widetilde{f}\left(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}} + \sqrt{\tfrac{\eta_{\text{G}}}{n}}\textstyle\sum_{j=1}^n [\widetilde{\boldsymbol{a}}]_j \widehat{w}_{k,j}\right)\widetilde{\boldsymbol{c}} \right\rangle_{\widetilde{\boldsymbol{c}}, \widetilde{\boldsymbol{a}}} \widehat{w}_{k,i} - \lambda \text{diag}(H'(\boldsymbol{S}_k))\widehat{\boldsymbol{v}}_{k,i}\right].
\end{aligned}$$

The bracket $\langle\cdot\rangle_{\widetilde{\boldsymbol{c}}, \widetilde{\boldsymbol{a}}}$ here denotes the average over $\widetilde{\boldsymbol{c}} \sim \mathcal{P}_{\widetilde{\boldsymbol{c}}}$, and standard Gaussian vector $\widetilde{\boldsymbol{a}}$, where $\widetilde{\boldsymbol{c}}$ and $\widetilde{\boldsymbol{a}}$ are the random variables generating the fake sample in the generator as described in (2). Noting that $\widetilde{\boldsymbol{a}}$ is a Gaussian vector, the term $\frac{1}{\sqrt{n}}\sum_{j=1}^n [\widetilde{\boldsymbol{a}}]_j \widehat{w}_{k,j}$ in the above equation is also a Gaussian

random variable, whose mean is zero and variance is z_k, which is defined in (S-5). Therefore, we have

$$\left\langle \widehat{\boldsymbol{v}}_{k+1,i} - \widehat{\boldsymbol{v}}_{k,i} \right\rangle_{\widetilde{\boldsymbol{c}}_{2k+1}, \widetilde{\boldsymbol{a}}_{2k+1}} = \frac{\widetilde{\tau}}{n} \Big[\widetilde{\boldsymbol{g}}_k \widehat{w}_{k,i} + \boldsymbol{L}_k \widehat{\boldsymbol{v}}_{k,i} \Big], \tag{S-6}$$

where

$$\widetilde{\boldsymbol{g}}_k = \left\langle \widetilde{f}\Big(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}} + \sqrt{z_k \eta_{\mathrm{G}}} e \Big) \widetilde{\boldsymbol{c}} \right\rangle_{\widetilde{\boldsymbol{c}}, e} \tag{S-7}$$

$$\boldsymbol{L}_k = -\lambda \mathrm{diag}(H'(\boldsymbol{S}_k)), \tag{S-8}$$

where $\langle \cdot \rangle_{\widetilde{\boldsymbol{c}}, e}$ denotes the average over $\widetilde{\boldsymbol{c}} \sim \mathcal{P}_{\widetilde{\boldsymbol{c}}}$ and $e \sim \mathcal{N}(0,1)$. In addition, from (S-4), we also know that the second moment

$$\left\langle \big(\widehat{\boldsymbol{v}}_{k+1,i} - \widehat{\boldsymbol{v}}_{k,i} \big)^2 \right\rangle_{\widetilde{\boldsymbol{c}}_{2k+1}, \widetilde{\boldsymbol{a}}_{2k+1}} = \mathcal{O}(n^{-\frac{3}{2}}). \tag{S-9}$$

The moments estimations (S-6) and (S-9) imply the second equation in (10) in the main text. Since the second moments growth smaller than $\mathcal{O}(n^{-1})$, the differential equation for $\widehat{\boldsymbol{v}}_t$ has no diffusion term.

Finally, we derive the last equation in (10) in the main text from the update rule of $\widehat{w}_k$ (S-3). We observe that both the terms inside the function f and outside of f in (S-3) depend on $a_{k,i}$. Using Taylor's expansion, we linearize the contribution of $a_{k,i}$ to the function f:

$$\begin{aligned} f_k &= f\Big(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} \textstyle\sum_{j \neq i} a_{k,j} \widehat{w}_{k,j} + \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} a_{k,i} \widehat{w}_{k,i} \Big) \\ &= f(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} \textstyle\sum_{j \neq i} a_{k,j} \widehat{w}_{k,j}) + f'(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} \textstyle\sum_{j \neq i} a_{k,j} \widehat{w}_{k,j}) \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} a_{k,i} \widehat{w}_{k,i} + \mathcal{O}(\tfrac{1}{n}). \end{aligned} \tag{S-10}$$

Similarly, we have

$$\begin{aligned} \widetilde{f}_{2k} &= \widetilde{f}(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}}_{2k} + \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \textstyle\sum_{j \neq i} \widetilde{a}_{k,j} \widehat{w}_{k,j} + \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \widetilde{a}_{2k,i} \widehat{w}_{k,i}) \\ &= \widetilde{f}(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}}_{2k} + \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \textstyle\sum_{j \neq i} \widetilde{a}_{k,j} \widehat{w}_{k,j}) + \widetilde{f}'(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}}_{2k} + \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \textstyle\sum_{j \neq i} \widetilde{a}_{k,j} \widehat{w}_{k,j}) \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \widetilde{a}_{2k,i} \widehat{w}_{k,i} + \mathcal{O}(\tfrac{1}{n}) \end{aligned} \tag{S-11}$$

Substituting (S-10) and (S-11) into (S-3), we have

$$\begin{aligned} &\frac{\widehat{w}_{k+1,i} - \widehat{w}_{k,i}}{\tau/n} \\ &= \widehat{\boldsymbol{u}}_i^\top \boldsymbol{c}_k f(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} \textstyle\sum_{j \neq i} a_{k,j} \widehat{w}_{k,j}) - \widehat{\boldsymbol{v}}_{k,i}^\top \widetilde{\boldsymbol{c}}_{2k} \widetilde{f}(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}}_{2k} + \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \textstyle\sum_{j \neq i} \widetilde{a}_{k,j} \widehat{w}_{k,j}) \\ &+ \widehat{w}_{k,i} \Big[a_{k,i}^2 f'(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} \textstyle\sum_{j \neq i} a_{k,j} \widehat{w}_{k,j}) - \widetilde{a}_{k,i}^2 \widetilde{f}'(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}}_{2k} + \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \textstyle\sum_{j \neq i} \widetilde{a}_{2k,j} \widehat{w}_{k,j}) - \lambda H'(z_k) \Big] \\ &+ \sqrt{n} \Big[a_{k,i} f(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \sqrt{\tfrac{\eta_{\mathrm{T}}}{n}} \textstyle\sum_{j \neq i} a_{k,j} \widehat{w}_{k,j}) + \widetilde{a}_{2k,i} \widetilde{f}'(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}}_{2k} + \sqrt{\tfrac{\eta_{\mathrm{G}}}{n}} \textstyle\sum_{j \neq i} \widetilde{a}_{2k,j} \widehat{w}_{k,j}) \Big] + \delta_{k,i}, \end{aligned} \tag{S-12}$$

where $\delta_{k,i}$ collects all higher-order terms whose contributions will vanish as $n \to \infty$. From this equation, we can already infer the SDE (10). Specifically, on the right hand side of (S-12), the terms in the first two lines correspond to the drift term in the SDE. Furthermore, the first term in the third line in (S-12) contributes to the SDE as a Brownian motion. More precisely, we can derive the third equation of the SDE (10) in the main text by the moments estimations. Specifically, the first-order moment is

$$\left\langle \widehat{w}_{k+1,i} - \widehat{w}_{k,i} \right\rangle_{\boldsymbol{c}_k, \boldsymbol{a}_k, \widetilde{\boldsymbol{c}}_{2k}, \widetilde{\boldsymbol{a}}_{2k}} = \frac{\tau}{n} \Big[\widehat{\boldsymbol{u}}_i^\top \boldsymbol{g}_k - \widehat{\boldsymbol{v}}_{k,i}^\top \widetilde{\boldsymbol{g}}_k + \widehat{w}_{k,i} h_k \Big] + \mathcal{O}(n^{-\frac{3}{2}}) \tag{S-13}$$

where $\widetilde{\boldsymbol{g}}_k$ is defined in (S-7), and

$$\boldsymbol{g}_k = \Big\langle \boldsymbol{c} f(\boldsymbol{q}_k^\top \boldsymbol{c} + \sqrt{z_k \eta_{\mathrm{T}}} e) \Big\rangle_{\boldsymbol{c}, e} \tag{S-14}$$

$$h_k = \eta_{\mathrm{T}} \Big\langle f'(\boldsymbol{q}_k^\top \boldsymbol{c} + \sqrt{z_k \eta_{\mathrm{T}}} e) \Big\rangle_{\boldsymbol{c}, e} - \widetilde{\eta}_{\mathrm{G}} \Big\langle \widetilde{f}'(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}} + \sqrt{z_k \eta_{\mathrm{G}}} e) \Big\rangle_{\widetilde{\boldsymbol{c}}, e} - \lambda H'(z_k). \tag{S-15}$$

The second moment is

$$\left\langle \left(\widehat{w}_{k+1,i} - \widehat{w}_{k,i}\right)^2 \right\rangle_{\boldsymbol{c}_k, \boldsymbol{a}_k, \widetilde{\boldsymbol{c}}_{2k}, \widetilde{\boldsymbol{a}}_{2k}} = \tfrac{\tau^2}{n} b_k + \mathcal{O}(n^{-\frac{3}{2}}), \tag{S-16}$$

where

$$b_k = \eta_{\mathrm{T}} \left\langle f^2(\boldsymbol{q}_k^\top \boldsymbol{c} + \sqrt{z_k \eta_{\mathrm{T}}} e) \right\rangle_{\boldsymbol{c}, e} + \eta_{\mathrm{G}} \left\langle \widetilde{f}^2(\boldsymbol{r}_k^\top \widetilde{\boldsymbol{c}} + \sqrt{z_k \eta_{\mathrm{G}}} e) \right\rangle_{\widetilde{\boldsymbol{c}}, e}. \tag{S-17}$$

From the (S-13) and (S-16), we derive the SDE for $\widehat{w}_t$ in (10) in the main text.

S-III Derive the ODE in Theorem 1 from the weak formulation of the PDE

In this section, we show how to derive the ODE (8) from the weak formulation of the PDE (12). Choosing the test function φ being each element of $\widehat{\boldsymbol{u}}\widehat{\boldsymbol{v}}^\top$, $\widehat{\boldsymbol{u}}\widehat{w}$, $\widehat{\boldsymbol{v}}\widehat{w}$, $\widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top$, $\widehat{w}^2$, and substituting those φ into the weak formulation of the PDE (12), we will get the ODE (8) as presented in Theorem 1. In what follows, we provide additional details of this derivation.

We first derive the first ODE $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}_t = \ldots$ in (8). Let $\varphi = [\widehat{\boldsymbol{u}}]_\ell [\widehat{\boldsymbol{v}}]_{\ell'}$, $\ell, \ell' = 1, 2, \ldots, d$, we have $\nabla_{\hat{\boldsymbol{v}}} \varphi = [\widehat{\boldsymbol{u}}]_\ell \boldsymbol{s}_{\ell'}$, where $\boldsymbol{s}_{\ell'}$ is the ℓ'th canonical basis (*i.e.,* all elements in $\boldsymbol{s}_{\ell'}$ are zeros, except that ℓ'th element is 1). From the PDE (12) in the main text, we have $\forall \ell, \ell' = 1, 2, \ldots, d$:

$$\left\langle \mu_t, \varphi(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w}) \right\rangle = \left\langle \mu_t, [\widehat{\boldsymbol{u}}]_\ell [\widehat{\boldsymbol{v}}]_{\ell'} \right\rangle = [\boldsymbol{P}_t]_{\ell, \ell'},$$

$$\begin{aligned} \left\langle \mu_t, (\widehat{w}\widetilde{\boldsymbol{g}}_t^\top + \widehat{\boldsymbol{v}}^\top \boldsymbol{L}_t) \nabla_{\hat{\boldsymbol{v}}} \varphi \right\rangle &= \left\langle \mu_t, ([\widehat{\boldsymbol{u}}]_\ell \widehat{w})[\widetilde{\boldsymbol{g}}_t]_{\ell'} + ([\widehat{\boldsymbol{u}}]_\ell \widehat{\boldsymbol{v}}^\top)[\boldsymbol{L}_t]_{:, \ell'} \right\rangle \\ &= [\boldsymbol{q}_t]_l [\widetilde{\boldsymbol{g}}_t]_{\ell'} + [\boldsymbol{P}_t]_{\ell, :} [\boldsymbol{L}_t]_{:, \ell'}, \end{aligned}$$

where $[\boldsymbol{P}_t]_{\ell, :}$ and $[\boldsymbol{L}_t]_{:, \ell'}$ are ℓth row of $\boldsymbol{P}_t$ and ℓ'th column of $\boldsymbol{L}$, respectively. In addition, we know that $\frac{\partial}{\partial \widehat{w}} \varphi = \frac{\partial^2}{\partial \widehat{w}^2} \varphi = 0$. Combining above results, we can recover the first ODE in (8).

Next, we derive the second ODE $\frac{d\boldsymbol{q}_t}{dt} = \ldots$ in (8). Let $\varphi = [\widehat{\boldsymbol{u}}]_\ell \widehat{w}$, $\ell = 1, 2, \ldots, d$. We have $\nabla_{\hat{\boldsymbol{v}}} \varphi = 0$, $\frac{\partial}{\partial \widehat{w}} \varphi = [\widehat{\boldsymbol{u}}]_\ell$ and $\frac{\partial^2}{\partial \widehat{w}^2} \varphi = 0$. Then $\forall \ell = 1, 2, \ldots, d$,

$$\left\langle \mu_t, \varphi(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w}) \right\rangle = \left\langle \mu_t, [\widehat{\boldsymbol{u}}]_\ell \widehat{w} \right\rangle = [\boldsymbol{q}_t]_\ell$$

and

$$\begin{aligned} \left\langle \mu_t, (\widehat{\boldsymbol{u}}^\top \boldsymbol{g}_t - \widehat{\boldsymbol{v}}^\top \widetilde{\boldsymbol{g}}_t + h_t \widehat{w}) \frac{\partial}{\partial \widehat{w}} \varphi \right\rangle &= \left\langle \mu_t, (\widehat{\boldsymbol{u}}^\top \boldsymbol{g}_t - \widehat{\boldsymbol{v}}^\top \widetilde{\boldsymbol{g}}_t + h_t \widehat{w})[\widehat{\boldsymbol{u}}]_\ell \right\rangle \\ &= [\boldsymbol{g}_t]_\ell - [\boldsymbol{P}_t]_\ell \widetilde{\boldsymbol{g}}_t + [\boldsymbol{q}_t]_\ell h_t. \end{aligned}$$

With above results, we can obtain the second ODE in (8).

Next, let's derive the ODE for $\frac{d\boldsymbol{S}_t}{dt}$. We set $\varphi = [\widehat{\boldsymbol{v}}]_\ell [\widehat{\boldsymbol{v}}]_{\ell'}$. If $\ell \neq \ell'$, we have $\nabla_{\hat{\boldsymbol{v}}} \varphi = [\widehat{\boldsymbol{v}}]_\ell \boldsymbol{s}_{\ell'} + [\widehat{\boldsymbol{v}}]_{\ell'} \boldsymbol{s}_\ell$, where $\boldsymbol{s}_{\ell'}$ is the ℓ'th canonical basis. Then

$$\left\langle \mu_t, \varphi(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w}) \right\rangle = [\boldsymbol{S}_t]_{\ell, \ell'}$$

and

$$\begin{aligned} \left\langle \mu_t, (\widehat{w}\widetilde{\boldsymbol{g}}_t^\top + \widehat{\boldsymbol{v}}^\top \boldsymbol{L}_t) \nabla_{\hat{\boldsymbol{v}}} \varphi \right\rangle &= \left\langle \mu_t, ([\widehat{\boldsymbol{v}}]_\ell \widehat{w})[\widetilde{\boldsymbol{g}}_t]_{\ell'} + ([\widehat{\boldsymbol{v}}]_\ell \widehat{\boldsymbol{v}}^\top)[\boldsymbol{L}_t]_{:, \ell'} \right\rangle \\ &\quad + \left\langle \mu_t, ([\widehat{\boldsymbol{v}}]_{\ell'} \widehat{w})[\widetilde{\boldsymbol{g}}_t]_\ell + ([\widehat{\boldsymbol{v}}]_{\ell'} \widehat{\boldsymbol{v}}^\top)[\boldsymbol{L}_t]_{:, \ell} \right\rangle \\ &= [\boldsymbol{r}_t]_\ell [\widetilde{\boldsymbol{g}}_t]_{\ell'} + [\widetilde{\boldsymbol{g}}_t]_\ell [\boldsymbol{r}_t]_{\ell'} + [\boldsymbol{S}_t]_{\ell, :} [\boldsymbol{L}_t]_{:, \ell'} + [\boldsymbol{L}_t]_{\ell, :} [\boldsymbol{S}_t]_{:, \ell'} \end{aligned}$$

If $\ell = \ell'$, we have $\nabla_{\hat{\boldsymbol{v}}} \varphi = 2[\widehat{\boldsymbol{v}}]_\ell \boldsymbol{s}_\ell$, then

$$\left\langle \mu_t, \varphi(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}, \widehat{w}) \right\rangle = [\boldsymbol{S}_t]_{\ell, \ell}$$

and

$$\left\langle \mu_t, (\widehat{w}\widetilde{\boldsymbol{g}}_t^\top + \widehat{\boldsymbol{v}}^\top \boldsymbol{L}_t) \nabla_{\hat{\boldsymbol{v}}} \varphi \right\rangle = 2([\boldsymbol{r}_t]_\ell [\widetilde{\boldsymbol{g}}_t]_\ell + [\boldsymbol{S}_t]_{\ell, :} [\boldsymbol{L}_t]_{:, \ell})$$

Plugging back the above two equations and combining the fact that $\frac{\partial}{\partial \widehat{w}} \varphi = \frac{\partial^2}{\partial \widehat{w}^2} \varphi = 0$, we recover the ODE of $\frac{d\boldsymbol{S}_t}{dt}$.

The rest two ODEs can be obtained in the similar way by letting φ to be each distinct component of $\widehat{\boldsymbol{v}}\widehat{w}$ and $\widehat{w}^2$.

S-IV Proof of Theorem 1

In this section, we prove Theorem 1 shown in the main text. In the previous section, we have already provided a derivation of the ODE in Theorem 1 from the weak formulation of the PDE for the microscopic states. In this section, we follow a different path to prove the theorem without referencing the PDE, because it is easier to establish the rigorous bound of the convergence rate. Thus, the proof itself also provides another derivation of the ODE, where the most relevant part is Lemma 5.

S-IV.1 Sketch of the proof

The proof follows the standard procedure of the convergence of stochastic processes [29, 30]. We here build the whole proof on Lemma 2 in the supplementary materials of [22]. For reader's convenient, we present that lemma below.

Lemma 1 (Lemma 2 in the supplementary materials of [22])**.** *Consider a sequence of stochastic process* $\{\boldsymbol{x}_k^{(n)}, k = 0, 1, 2, \ldots, \lfloor nT \rfloor\}_{n=1,2,\ldots}$*, with some constant* $T > 0$*. If* $\boldsymbol{x}_k^{(n)}$ *can be decomposed into three parts*

$$\boldsymbol{x}_{k+1}^{(n)} - \boldsymbol{x}_k^{(n)} = \tfrac{1}{n}\phi(\boldsymbol{x}_k^{(n)}) + \boldsymbol{\rho}_k^{(n)} + \boldsymbol{\delta}_k^{(n)} \tag{S-18}$$

such that

(C.1) The process $\sum_{k'=0}^{k} \boldsymbol{\rho}_{k'}^{(n)}$ *is a martingale, and* $\mathbb{E}\,\|\boldsymbol{\rho}_k^{(n)}\|^2 \le C(T)/n^{1+\epsilon_1}$ *for some positive* ϵ_1*;*

(C.2) $\mathbb{E}\,\|\boldsymbol{\delta}_k^{(n)}\| \le C(T)/n^{1+\epsilon_2}$ *for some positive* ϵ_2*;*

(C.3) $\phi(\boldsymbol{x})$ *is a Lipschitz function, i.e.,* $\|\phi(\boldsymbol{x}) - \phi(\widetilde{\boldsymbol{x}})\| \le C\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|$*;*

(C.4) $\mathbb{E}\,\|\boldsymbol{x}_k^{(n)}\|^2 \le C$ *for all* $k \le \lfloor nT \rfloor$*;*

(C.5) $\mathbb{E}\,\|\boldsymbol{x}_0^{(n)} - \boldsymbol{x}_0^*\| \le C/n^{\epsilon_3}$ *for some positive* ϵ_3 *and a deterministic vector* $\boldsymbol{x}_0^*$*,*

then we have

$$\|\boldsymbol{x}_k^{(n)} - \boldsymbol{x}(\tfrac{k}{n})\| \le C(T)n^{-\min\{\frac{1}{2}\epsilon_1, \epsilon_2, \epsilon_3\}},$$

where $\boldsymbol{x}(t)$ *is the solution of the ODE*

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = \phi(\boldsymbol{x}(t)), \quad \text{with } \boldsymbol{x}(0) = \boldsymbol{x}_0^*.$$

In Theorem 1, the stochastic process is the macroscopic states $\{\boldsymbol{M}_k, k = 0, 1, \ldots\}$, where $\boldsymbol{M}_k$ is a symmetric matrix consists of 5 non-trivial parts $\boldsymbol{P}_k$, $\boldsymbol{q}_k$, $\boldsymbol{r}_k$, $\boldsymbol{S}_k$, and z_k as shown in (6) in the main text. Following (S-18), we have the following decomposition for $\boldsymbol{M}_k$

$$\boldsymbol{M}_{k+1} - \boldsymbol{M}_k = \tfrac{1}{n}\phi(\boldsymbol{M}_k) + (\boldsymbol{M}_{k+1} - \mathbb{E}_k\,\boldsymbol{M}_{k+1}) + [\mathbb{E}_k\,\boldsymbol{M}_{k+1} - \boldsymbol{M}_k - \tfrac{1}{n}\phi(\boldsymbol{M}_k)], \tag{S-19}$$

in which the matrix-valued function $\phi(\boldsymbol{M})$ represents the functions on the right hand sides of the ODE (8), and $\mathbb{E}_k$ denotes the conditional expectation given the state of the Markov chain $\boldsymbol{X}_k$. Note that the stochastic process of the macroscopic state $\boldsymbol{M}_k$ is driven by the Markov chain of the microscopic state $\boldsymbol{X}_k$. Thus, $\mathbb{E}_k$ is well-defined. For future reference, we denotes $\mathbb{E}$ the unconditional expectation of all the randomness of the Markov chain $\boldsymbol{X}_k$, *i.e.*, the initial state $\boldsymbol{U}, \boldsymbol{V}_0, \boldsymbol{w}_0$ and $\{\boldsymbol{a}_k, \boldsymbol{c}_k, \widetilde{\boldsymbol{a}}_k, \widetilde{\boldsymbol{c}}_k | k = 0, 1, 2, \ldots\}$. By definition, $\sum_{k'=0}^{k}(\boldsymbol{M}_{k'+1} - \mathbb{E}_{k'}\boldsymbol{M}_{k'})$ is a Martingale.

S-IV.2 Check the conditions provided in Lemma 1

In this subsection, we check the condition (C.1)–(C.5) for the decomposition of (S-19). Once all conditions are proved to be satisfied, Theorem 1 will be proved.

We first note that (C.5) is the assumption (A.5) in the main text. Thus, (C.5) is satisfied. Before proving other conditions, we declare a lemma.

Lemma 2. *Under the same setting as Theorem 1, given* $T > 0$*, then*

$$\mathbb{E}\left(\sum_{\ell=1}^{d}[\boldsymbol{V}_k]_{i,\ell}^4 + [\boldsymbol{w}_k]_i^4\right) \le C(T)n^{-2}, \quad \forall i = 1, 2, \ldots, n, \text{ and } k = 0, 1, \ldots, \lfloor nT \rfloor, \tag{S-20}$$

The proof can be founded in Section S-IV.3.

Check Condition (C.4)

Lemma 3. *Under the same setting as Theorem 1, for all $k = 0, 1, \ldots, \lfloor nT \rfloor$ with a given $T > 0$, then*

$$\mathbb{E}\,\|\boldsymbol{P}_k\|^2 \le C(T), \qquad \mathbb{E}\,\|\boldsymbol{q}_k\|^2 \le C(T),$$
$$\mathbb{E}\,\|\boldsymbol{S}_k\|^2 \le C(T), \qquad \mathbb{E}\, z_k^2 \le C(T),$$
$$\mathbb{E}\,\|\boldsymbol{r}_k\|^2 \le C(T).$$

Proof. It's a direct consequence of Lemma 2. We first verify $\mathbb{E}\, z_k^2 \le C(T)$. Using Holder's inequality, we have

$$\mathbb{E}\, z_k^2 = \mathbb{E}\left(\sum_{i=1}^n w_{k,i}^2\right)^2 \le n\mathbb{E}\sum_{i=1}^n w_{k,i}^4 \le C(T)$$

For $[\boldsymbol{S}_k]_{\ell,\ell}$, $\ell = 1, \ldots, d$, similarly, we have

$$\mathbb{E}\,[\boldsymbol{S}_k]_{\ell,\ell}^2 = \mathbb{E}\left(\sum_{i=1}^n [\boldsymbol{V}_k]_{i,\ell}^2\right)^2 \le C(T).$$

and for $\mathbb{E}\,[\boldsymbol{S}_k]_{\ell,\ell'}^2$, $\ell \ne \ell'$, we have:

$$\begin{aligned}
\mathbb{E}\,[\boldsymbol{S}_k]_{\ell,\ell'}^2 &= \mathbb{E}\left(\sum_{i=1}^n [\boldsymbol{V}_k]_{i,\ell}[\boldsymbol{V}_k]_{i,\ell'}\right)^2 \\
&\le \mathbb{E}\left(\sum_{i=1}^n [\boldsymbol{V}_k]_{i,\ell}^2\right)\left(\sum_{i=1}^n [\boldsymbol{V}_k]_{i,\ell'}^2\right) \\
&\le \sqrt{\mathbb{E}\left(\sum_{i=1}^n [\boldsymbol{V}_k]_{i,\ell}^2\right)^2 \mathbb{E}\left(\sum_{i=1}^n [\boldsymbol{V}_k]_{i,\ell'}^2\right)^2} \\
&\le C(T)
\end{aligned}$$

where in reaching the third and last line, we used the Cauchy-Schwartz inequality. Now, we get $\mathbb{E}\,\|\boldsymbol{S}_k\|^2 \le C(T)$. The rest bounds of $\mathbb{E}\,\|\boldsymbol{P}_k\|^2$, $\mathbb{E}\,\|\boldsymbol{q}_k\|^2$ and $\mathbb{E}\,\|\boldsymbol{r}_k\|^2$ in Lemma 3 can also be directly verified using the Cauchy-Schwartz inequality. □

Check Condition (C.3)

Lemma 4. *If Assumption (A.3) hold, $\phi(\boldsymbol{M})$ is a Lipschitz function.*

Proof. It suffices to verify each component of gradient $\nabla\phi(\boldsymbol{M})$ is bounded. Assumption (A.3) ensures that H' is Lipschitz and the derivatives up to fourth order of the functions f, $\widetilde{f}$ exists and uniformly bounded. These conditions guarantee that the partial derivatives of $\phi(\boldsymbol{M})$ w.r.t. $\boldsymbol{P}$, $\boldsymbol{q}$, $\boldsymbol{S}$ and $\boldsymbol{r}$ are bounded. The remaining thing is to show that $\frac{\partial\phi(\boldsymbol{M})}{\partial z}$ is also bounded. Since there is a $\sqrt{z}$ term in $\phi(\boldsymbol{M})$, the boundness can be potentially broken at $z = 0$. However, we can show that it is not the case. For example, we can show that $\left\langle \boldsymbol{c} f(\boldsymbol{c}^\top \boldsymbol{q} + e\sqrt{z})\right\rangle_{\boldsymbol{c},e}$ is a Lipschitz function, because

$$\begin{aligned}
\tfrac{\partial}{\partial z}\left\langle \boldsymbol{c} f(\boldsymbol{c}^\top \boldsymbol{q} + e\sqrt{z})\right\rangle_{c,e} &= \tfrac{1}{2} z^{-\frac{1}{2}}\left\langle ecf'(cq + e\sqrt{z})\right\rangle_{c,e} \\
&= \tfrac{1}{2}\left\langle cf''(cq + e\sqrt{z})\right\rangle_{c,e}
\end{aligned}$$

is always a well-defined bounded function. In reaching the first line, we here interchanged the expectation and derivative, which is valid because of the boundness of $f(\cdot)$, and in reaching the second line, we used the Stein's lemma. Finally, other terms in (9) involving $\sqrt{z}$ can be treated in the same way. Thus, $\phi(\boldsymbol{M})$ is a Lipschitz function. □

Check Condition (C.2)

Lemma 5. *Under the same setting as Theorem 1, for all $k = 0, 1, \ldots, \lfloor nT \rfloor$ with a given $T > 0$, then*

$$\mathbb{E}\,\|\mathbb{E}_k\, \boldsymbol{M}_{k+1} - \boldsymbol{M}_k - \tfrac{1}{n}\phi(\boldsymbol{M}_k)\| \le C(T) n^{-\frac{3}{2}}.$$

Proof. The above inequality can be split into 5 parts

$$\mathbb{E}\left\|\mathbb{E}_k \boldsymbol{P}_{k+1} - \boldsymbol{P}_k - \tfrac{\widetilde{\tau}}{n}(\boldsymbol{q}_k\widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{P}_k\boldsymbol{L}_k)\right\| \le C(T)n^{-\frac{3}{2}} \tag{S-21}$$

$$\mathbb{E}\left\|\mathbb{E}_k \boldsymbol{q}_{k+1} - \boldsymbol{q}_k - \tfrac{\tau}{n}(\boldsymbol{g}_k - \boldsymbol{P}_k\widetilde{\boldsymbol{g}}_k + \boldsymbol{q}_k h_k)\right\| \le C(T)n^{-\frac{3}{2}} \tag{S-22}$$

$$\mathbb{E}\left\|\mathbb{E}_k \boldsymbol{S}_{k+1} - \boldsymbol{S}_k - \tfrac{\widetilde{\tau}}{n}\left(\boldsymbol{r}_k\widetilde{\boldsymbol{g}}_k^\top + \widetilde{\boldsymbol{g}}_k\boldsymbol{r}_k^\top + \boldsymbol{S}_k\boldsymbol{L}_k + \boldsymbol{L}_k\boldsymbol{S}_k\right)\right\| \le C(T)n^{-\frac{3}{2}} \tag{S-23}$$

$$\mathbb{E}\left\|\mathbb{E}_k z_{k+1} - z_k - \tfrac{2\tau}{n}\left(\boldsymbol{q}_k^\top\boldsymbol{g}_k - \boldsymbol{r}_k^\top\widetilde{\boldsymbol{g}}_k + z_k h_k\right) - \tfrac{\tau^2}{n}b_k\right\| \le C(T)n^{-\frac{3}{2}}, \tag{S-24}$$

$$\mathbb{E}\left\|\mathbb{E}_k \boldsymbol{r}_{k+1} - \boldsymbol{r}_k - \tfrac{\tau}{n}\left(\boldsymbol{P}_k^\top\boldsymbol{g}_k - \boldsymbol{S}_k\widetilde{\boldsymbol{g}}_k + \boldsymbol{r}_k h_k\right) - \tfrac{\widetilde{\tau}}{n}(z_k\widetilde{\boldsymbol{g}}_k + \boldsymbol{L}_k\boldsymbol{r}_k)\right\| \le C(T)n^{-\frac{3}{2}} \tag{S-25}$$

where $\widetilde{\boldsymbol{g}}_k,\ \boldsymbol{L}_k,\ \boldsymbol{g}_k,\ h_k,\ b_k$ are defined in (S-7), (S-8), (S-14), (S-15) and (S-17), respectively.

We first prove (S-21). From (S-2), we have

$$\boldsymbol{V}_{k+1} - \boldsymbol{V}_k = \tfrac{\widetilde{\tau}}{n}\big[\boldsymbol{w}_k\widetilde{\boldsymbol{c}}_{2k+1}^\top\widetilde{f}(\widetilde{\boldsymbol{c}}_{2k+1}^\top\boldsymbol{V}_k^\top\boldsymbol{w}_k + \eta_{\mathrm{G}}\widetilde{\boldsymbol{a}}_{2k+1}^\top\boldsymbol{w}_k) - \lambda\boldsymbol{V}_k\mathrm{diag}(H'(\boldsymbol{S}_k))\big]. \tag{S-26}$$

Averaging both sides of the above equation over $\widetilde{\boldsymbol{c}}_{2k+1}$ and $\widetilde{\boldsymbol{a}}_{2k+1}$, we have

$$\mathbb{E}_k\,\boldsymbol{V}_{k+1} - \boldsymbol{V}_k = \tfrac{\widetilde{\tau}}{n}\big[\boldsymbol{w}_k\widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{V}_k\boldsymbol{L}_k\big], \tag{S-27}$$

where $\widetilde{\boldsymbol{g}}_k$ and $\boldsymbol{L}_k$ are defined in (S-7) and (S-8), respectively. Multiplying $\boldsymbol{U}^\top$ from the left on the both sides of the above equation, we have

$$\mathbb{E}_k\,\boldsymbol{P}_{k+1} - \boldsymbol{P}_k = \tfrac{\widetilde{\tau}}{n}\left[\boldsymbol{q}_k\widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{P}_k\boldsymbol{L}_k\right],$$

which implies (S-21). In fact, there is no higher-order term in (S-21), and the left hand side of (S-21) is exactly zero.

Then, we prove (S-22). From (S-1), we have

$$\boldsymbol{w}_{k+1} - \boldsymbol{w}_k = \tfrac{\tau}{n}\left[\boldsymbol{y}_k f(\boldsymbol{y}_k^\top\boldsymbol{w}_k) - \widetilde{\boldsymbol{y}}_{2k}\widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top\boldsymbol{w}_k) - \lambda\boldsymbol{w}_k\mathrm{diag}(H'(z_k))\right], \tag{S-28}$$

where $\boldsymbol{y}_k = \boldsymbol{U}\boldsymbol{c}_k + \sqrt{\eta_{\mathrm{T}}}\boldsymbol{a}_k$ and $\widetilde{\boldsymbol{y}}_{2k} = \boldsymbol{V}_k\widetilde{\boldsymbol{c}}_{2k} + \sqrt{\eta_{\mathrm{G}}}\widetilde{\boldsymbol{a}}_{2k}$. Averaging both sides of the above equation over $\boldsymbol{c}_k$, $\boldsymbol{a}_k\widetilde{\boldsymbol{c}}_{2k}$ and $\widetilde{\boldsymbol{a}}_{2k}$, we have

$$\begin{aligned}\mathbb{E}_k\,\boldsymbol{w}_{k+1} - \boldsymbol{w}_k =& \tfrac{\tau}{n}\Big[\boldsymbol{U}\boldsymbol{g}_k + \Big\langle\boldsymbol{a}_k f(\boldsymbol{c}_k^\top\boldsymbol{q}_k + \sqrt{\eta_{\mathrm{T}}}\boldsymbol{a}_k^\top\boldsymbol{w}_k)\Big\rangle \\ &- \boldsymbol{V}_k\widetilde{\boldsymbol{g}}_k - \Big\langle\widetilde{\boldsymbol{a}}_{2k}\widetilde{f}(\widetilde{\boldsymbol{c}}_{2k}^\top\boldsymbol{r}_k + \sqrt{\eta_{\mathrm{G}}}\widetilde{\boldsymbol{a}}_{2k}^\top\boldsymbol{w}_k)\Big\rangle - \lambda\boldsymbol{w}_k\mathrm{diag}(H'(z_k))\Big].\end{aligned}$$

Multiplying $\boldsymbol{U}^\top$ from the left on the both sides of the above equation, we have

$$\begin{aligned}\mathbb{E}_k\,\boldsymbol{q}_{k+1} - \boldsymbol{q}_k =& \tfrac{\tau}{n}\Big[\boldsymbol{g}_k - \boldsymbol{P}_k\widetilde{\boldsymbol{g}}_k + \sqrt{\eta_{\mathrm{T}}}\Big\langle\boldsymbol{U}^\top\boldsymbol{a}_k f(\boldsymbol{c}_k^\top\boldsymbol{q}_k + \sqrt{\eta_{\mathrm{T}}}\boldsymbol{a}_k^\top\boldsymbol{w}_k)\Big\rangle_{\boldsymbol{c},\boldsymbol{a}} \\ &- \sqrt{\eta_{\mathrm{G}}}\Big\langle\boldsymbol{U}^\top\widetilde{\boldsymbol{a}}\widetilde{f}(\widetilde{\boldsymbol{c}}^\top\boldsymbol{r}_k + \sqrt{\eta_{\mathrm{G}}}\widetilde{\boldsymbol{a}}^\top\boldsymbol{w}_k)\Big\rangle_{\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{a}}} - \lambda\boldsymbol{q}_k\mathrm{diag}(H'(z_k))\Big]\end{aligned} \tag{S-29}$$

We note that $\begin{bmatrix}\boldsymbol{U}^\top\boldsymbol{a}_k \\ \boldsymbol{w}_k^\top\boldsymbol{a}_k\end{bmatrix}$ are Gaussian random vector with zero-mean and covariance matrix $\begin{bmatrix}\boldsymbol{I} & \boldsymbol{q}_k \\ \boldsymbol{q}_k^\top & z_k\end{bmatrix}$.
We can rewrite

$$\begin{aligned}\Big\langle\boldsymbol{U}^\top\boldsymbol{a}f(\boldsymbol{c}^\top\boldsymbol{q}_k + \sqrt{\eta_{\mathrm{T}}}\boldsymbol{a}^\top\boldsymbol{w}_k)\Big\rangle_{\boldsymbol{c},\boldsymbol{a}} &= z_k^{-1/2}\boldsymbol{U}^\top\boldsymbol{w}_k\Big\langle e f(\boldsymbol{c}^\top\boldsymbol{q}_k + \sqrt{z_k\eta_{\mathrm{T}}}e\Big\rangle_{\boldsymbol{c},e} \\ &= \sqrt{\eta_T}\boldsymbol{q}_k\Big\langle f'(\boldsymbol{c}^\top\boldsymbol{q}_k + \sqrt{z_k\eta_{\mathrm{T}}}e\Big\rangle_{\boldsymbol{c},e},\end{aligned} \tag{S-30}$$

where the second line is due to Stein's lemma (i.e., integral by part for Gaussian random variable.) Similarly, we have

$$\Big\langle\boldsymbol{U}^\top\widetilde{\boldsymbol{a}}\widetilde{f}(\widetilde{\boldsymbol{c}}^\top\boldsymbol{r}_k + \sqrt{\eta_{\mathrm{G}}}\widetilde{\boldsymbol{a}}^\top\boldsymbol{w}_k)\Big\rangle_{\widetilde{\boldsymbol{c}},\widetilde{\boldsymbol{a}}} = \sqrt{\eta_{\mathrm{G}}}\boldsymbol{q}_k\Big\langle\widetilde{f}'(\widetilde{\boldsymbol{c}}^\top\boldsymbol{r}_k + \sqrt{z_k\eta_{\mathrm{G}}}e\Big\rangle_{\widetilde{\boldsymbol{c}},e}. \tag{S-31}$$

Substituting (S-30) and (S-31) into (S-29), we get

$$\mathbb{E}_k\,\boldsymbol{q}_{k+1} - \boldsymbol{q}_k = \tfrac{\tau}{n}\left[\boldsymbol{g}_k - \boldsymbol{P}_k\widetilde{\boldsymbol{g}}_k + \boldsymbol{q}_k h_k\right],$$

where $\widetilde{\boldsymbol{g}}_k$, $\boldsymbol{g}_k$, and h_k are defined in (S-7), (S-14), and (S-15), respectively. Now, we proved (S-22), which again has no higher-order term.

We next prove (S-23). Note that

$$\begin{aligned}\boldsymbol{S}_{k+1} - \boldsymbol{S}_k &= (\boldsymbol{V}_k + \boldsymbol{V}_{k+1} - \boldsymbol{V}_k)^\top (\boldsymbol{V}_k + \boldsymbol{V}_{k+1} - \boldsymbol{V}_k) - \boldsymbol{S}_k \\ &= \boldsymbol{V}_k^\top (\boldsymbol{V}_{k+1} - \boldsymbol{V}_k) + (\boldsymbol{V}_{k+1} - \boldsymbol{V}_k)^\top \boldsymbol{V}_k + (\boldsymbol{V}_{k+1} - \boldsymbol{V}_k)^\top (\boldsymbol{V}_{k+1} - \boldsymbol{V}_k).\end{aligned}$$

Averaging both sides of the above equation over $\widetilde{\boldsymbol{c}}_{2k+1}$ and $\widetilde{\boldsymbol{a}}_{2k+1}$ and substituting (S-27) into above equation, we have

$$\mathbb{E}_k \boldsymbol{S}_{k+1} - \boldsymbol{S}_k = \tfrac{\widetilde{\tau}}{n}\left[\boldsymbol{r}_k \widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{S}_k \boldsymbol{L}_k + \widetilde{\boldsymbol{g}}_k \boldsymbol{r}_k^\top + \boldsymbol{L}_k \boldsymbol{S}_k\right] + \tfrac{\widetilde{\tau}^2}{n^2}\left[\boldsymbol{w}_k \widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{V}_k \boldsymbol{L}_k\right]^\top \left[\boldsymbol{w}_k \widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{V}_k \boldsymbol{L}_k\right]. \tag{S-32}$$

We know that

$$\begin{aligned}\mathbb{E}\left\|\left[\boldsymbol{w}_k \widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{V}_k \boldsymbol{L}_k\right]^\top \left[\boldsymbol{w}_k \widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{V}_k \boldsymbol{L}_k\right]\right\| &\le \mathbb{E}\left\|\boldsymbol{w}_k \widetilde{\boldsymbol{g}}_k^\top + \boldsymbol{V}_k \boldsymbol{L}_k\right\|^2 \\ &\le 2 z_k \|\widetilde{\boldsymbol{g}}_k\|^2 + 2\|\boldsymbol{S}_k\| \|\boldsymbol{L}_k\|^2 \\ &\le C\mathbb{E}\left[z_k + \|\boldsymbol{S}_k\|\right] \\ &\le C(T),\end{aligned} \tag{S-33}$$

where $\widetilde{\boldsymbol{g}}_k$, $\boldsymbol{L}_k$ are defined in (S-7) and (S-8), respectively. The third line of the above inequalities is due to the fact that $\widetilde{f}$ and H' are uniformly bounded, and in reaching the last line, we used Lemma 3. Combining (S-32) and (S-33), we reach (S-23).

The other two inequalities (S-24) and (S-25) can be proved in a similar way. We omit the details here. □

Check Condition (C.1)

Lemma 6. *Under the same setting as Theorem 1, for all $k = 0, 1, \ldots, \lfloor nT \rfloor$ with a given $T > 0$, then*

$$\mathbb{E}\left\|\boldsymbol{M}_{k+1} - \mathbb{E}_k \boldsymbol{M}_{k+1}\right\|^2 \le C(T) n^{-2}.$$

Proof. Note that $\mathbb{E}\left\|\boldsymbol{M}_{k+1} - \mathbb{E}_k \boldsymbol{M}_{k+1}\right\|^2 = \mathbb{E}\left\|\boldsymbol{M}_{k+1} - \boldsymbol{M}_k - \mathbb{E}_k\left(\boldsymbol{M}_{k+1} - \boldsymbol{M}_k\right)\right\|^2 \le \mathbb{E}\left\|\boldsymbol{M}_{k+1} - \boldsymbol{M}_k\right\|^2$. It is sufficient to prove

$$\mathbb{E}\left\|\boldsymbol{M}_{k+1} - \boldsymbol{M}_k\right\|^2 \le C(T) n^{-2}. \tag{S-34}$$

In what follows, we are going to bound the second-order moment of each element in $\boldsymbol{M}_{k+1} - \boldsymbol{M}_k$. In particular, we bound the 5 blocks $\boldsymbol{P}_k$, $\boldsymbol{S}_k$, $\boldsymbol{q}_k$, z_k and $\boldsymbol{r}_k$ of $\boldsymbol{M}_k$ separately.

We first bound $\mathbb{E}\left\|\boldsymbol{P}_{k+1} - \boldsymbol{P}_k\right\|^2$. Multiplying $\boldsymbol{U}^\top$ from left on both sides of (S-26), we have

$$\boldsymbol{P}_{k+1} - \boldsymbol{P}_k = \tfrac{\widetilde{\tau}}{n}\left[\boldsymbol{q}_k \widetilde{\boldsymbol{c}}_{2k+1}^\top \widetilde{f}(\widetilde{\boldsymbol{c}}_{2k+1}^\top \boldsymbol{V}_k^\top \boldsymbol{w}_k + \eta_{\mathrm{G}} \widetilde{\boldsymbol{a}}_{2k+1}^\top \boldsymbol{w}_k) - \lambda \boldsymbol{P}_k \mathrm{diag}(H'(\boldsymbol{V}_k^\top \boldsymbol{V}_k))\right]$$

We then get

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{P}_{k+1} - \boldsymbol{P}_k\right\|^2 &\le C n^{-2} \mathbb{E}\left[\|\boldsymbol{q}_k\|^2 \mathbb{E}_k \|\widetilde{\boldsymbol{c}}_{2k+1}\|^2 + \|\boldsymbol{P}_k\|^2\right] \\ &\le C n^{-2} \mathbb{E}\left[1 + \|\boldsymbol{q}_k\|^2 + \|\boldsymbol{P}_k\|^2\right] \\ &\le C(T) n^{-2}.\end{aligned} \tag{S-35}$$

Here the last line is due to Lemma 3.

We next bound $\mathbb{E}\left\|\boldsymbol{q}_{k+1} - \boldsymbol{q}_k\right\|^2$ in the same way. Specifically, multiplying $\boldsymbol{U}^\top$ from the left on both sides of (S-28), we get

$$\boldsymbol{q}_{k+1} - \boldsymbol{q}_k = \tfrac{\tau}{n}\left[\boldsymbol{U}^\top \boldsymbol{y}_k f(\boldsymbol{y}_k^\top \boldsymbol{w}_k) - \boldsymbol{U}^\top \widetilde{\boldsymbol{y}}_{2k} \widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top \boldsymbol{w}_k) - \lambda \boldsymbol{q}_k \mathrm{diag}(H'(\boldsymbol{w}_k^\top \boldsymbol{w}_k))\right].$$

We then have

$$\begin{aligned}
&\mathbb{E}\,\|\boldsymbol{q}_{k+1}-\boldsymbol{q}_k\|^2\\
&\le \tfrac{\tau^2}{n^2}\mathbb{E}\left[\|\boldsymbol{c}_k\|^2 f_k^2+\|\boldsymbol{U}^\top\boldsymbol{a}_k\|^2 f_k^2+\|\boldsymbol{P}_k\|^2\|\widetilde{\boldsymbol{c}}_{2k}\|^2\widetilde{f}_{2k}^2+\|\boldsymbol{U}^\top\widetilde{\boldsymbol{a}}_{2k}\|^2\widetilde{f}_{2k}^2+\|\boldsymbol{q}_k\|^2 h_k^2\right]\\
&\le Cn^{-2}\big[1+\sqrt{\mathbb{E}\,\|\boldsymbol{U}^\top\boldsymbol{a}_k\|^4}\sqrt{\mathbb{E}\,f_k^4}+\sqrt{\mathbb{E}\,\|\boldsymbol{U}^\top\widetilde{\boldsymbol{a}}_{2k}\|^4}\sqrt{\mathbb{E}\,\widetilde{f}_{2k}^4}+\mathbb{E}\,z_k^2+\mathbb{E}\,\|\boldsymbol{S}_k\|^2\big]\\
&\le Cn^{-2}[1+\mathbb{E}\,z_k^2+\mathbb{E}\,\|\boldsymbol{S}_k\|^2]\\
&\le C(T)n^{-2},
\end{aligned}\tag{S-36}$$

where f_k and $\widetilde{f}_{2k}$ are shorthands for $f(\boldsymbol{y}_k^\top\boldsymbol{w}_k)$ and $\widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top\boldsymbol{w}_k)$ respectively. In reaching the last line, we used Lemma 3 again.

Similarly, we can also prove that

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{S}_{k+1}-\boldsymbol{S}_k\|^2&\le C(T)n^{-2}\\
\mathbb{E}\,(z_{k+1}-z_k)^2&\le C(T)n^{-2}\\
\mathbb{E}\,\|\boldsymbol{r}_{k+1}-\boldsymbol{r}_k\|^2&\le C(T)n^{-2}.
\end{aligned}\tag{S-37}$$

Combining (S-35), (S-36) and (S-37), we can prove (S-34), which concludes the whole proof. □

S-IV.3 Proof of Lemma 2

Before proving Lemma 2, we first present and prove the following lemma. Let $\boldsymbol{u}_i$ and $\boldsymbol{v}_{k,i}$ denote the ith row vectors of $\boldsymbol{U}$ and $\boldsymbol{V}_k$ in column view, respectively, and let $w_{k,i}$ be the ith element of the vector $\boldsymbol{w}_k$.

Lemma 7. *Under the same setting as Theorem 1, for all $k=0,1,\ldots,\lfloor nT\rfloor$ with a given $T>0$, then*

$$\|\mathbb{E}_k\,\boldsymbol{v}_{k+1,i}-\boldsymbol{v}_{k,i}\|\le Cn^{-1}\left(\|\boldsymbol{v}_{k,i}\|+\left|w_{k,i}\right|\right)\tag{S-38}$$

$$\left|\mathbb{E}_k\,w_{k,i}-w_{k,i}\right|\le Cn^{-1}\left(\|\boldsymbol{u}_i\|+\|\boldsymbol{v}_{k,i}\|+\left|w_{k,i}\right|\right).\tag{S-39}$$

In the proof of this lemma and Lemma 2, we omit the two constants η_{T} and η_{G} for simplicity.

Proof. From (S-2) and knowing that the function $\widetilde{f}$ and H' are uniformly bounded, we can immediately prove (S-38).

Next, we are going to prove (S-39). From (S-1), we know

$$\begin{aligned}
&\left|\mathbb{E}_k\,w_{k+1,i}-w_{k,i}\right|\\
&\le\tfrac{\tau}{n}\Bigg(\left|\boldsymbol{u}_i^\top\left\langle\boldsymbol{c}_k f(\boldsymbol{y}_k^\top\boldsymbol{w}_k)\right\rangle_{\boldsymbol{c}_k,\boldsymbol{a}_k}\right|+\left|\left\langle a_{k,i}f(\boldsymbol{y}_k^\top\boldsymbol{w}_k)\right\rangle_{\boldsymbol{c}_k,\boldsymbol{a}_k}\right|\\
&+\left|\boldsymbol{v}_{k,i}^\top\left\langle\widetilde{\boldsymbol{c}}_{2k}\widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top\boldsymbol{w}_k)\right\rangle_{\widetilde{\boldsymbol{c}}_{2k},\widetilde{\boldsymbol{a}}_{2k}}\right|+\left|\left\langle\widetilde{a}_{2k,i}\widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top\boldsymbol{w}_k)\right\rangle_{\widetilde{\boldsymbol{c}}_{2k},\widetilde{\boldsymbol{a}}_{2k}}\right|+\lambda\left|w_{k,i}H'(\boldsymbol{w}_k^\top\boldsymbol{w}_k)\right|\Bigg)\\
&\le Cn^{-1}\Bigg(\|\boldsymbol{u}_i\|+\|\boldsymbol{v}_{k,i}\|+\left|w_{k,i}\right|+\left|\left\langle a_{k,i}f(\boldsymbol{y}_k^\top\boldsymbol{w}_k)\right\rangle_{\boldsymbol{c}_k,\boldsymbol{a}_k}\right|+\left|\left\langle\widetilde{a}_{2k,i}\widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top\boldsymbol{w}_k)\right\rangle_{\widetilde{\boldsymbol{c}}_{2k},\widetilde{\boldsymbol{a}}_{2k}}\right|\Bigg),
\end{aligned}\tag{S-40}$$

where the last is due to the fact that H', f and $\widetilde{f}$ are uniformly bounded. Using Taylor's expansion up-to zero-order

$$\begin{aligned}
f(\boldsymbol{y}_k^\top\boldsymbol{w}_k)&=f(\boldsymbol{q}_k^\top\boldsymbol{c}_k+\textstyle\sum_{j\neq i}w_{k,j}a_{k,j}+w_{k,j}a_{k,j})\\
&=f(\boldsymbol{q}_k^\top\boldsymbol{c}_k+\textstyle\sum_{j\neq i}w_{k,j}a_{k,j})+f'(\boldsymbol{q}_k^\top\boldsymbol{c}_k+\textstyle\sum_{j\neq i}w_{k,j}a_{k,j}+\chi_{k,i})w_{k,j}a_{k,j},
\end{aligned}$$

with $\chi_{k,i}$ being some number such that $\left|\chi_{k,i}\right| \leq \left|w_{k,i}a_{k,i}\right|$, we have

$$\begin{aligned}
&\left|\left\langle a_{k,i} f(\boldsymbol{y}_k^\top \boldsymbol{w}_k)\right\rangle_{\boldsymbol{c}_k,\boldsymbol{a}_k}\right| \\
&\leq \left|\left\langle f(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \textstyle\sum_{j\neq i} w_{k,j}a_{k,j})a_{k,i}\right\rangle_{\boldsymbol{c}_k,\boldsymbol{a}_k}\right| + \left|\left\langle f'(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \textstyle\sum_{j\neq i} w_{k,j}a_{k,j} + \chi_{k,i})w_{k,j}a_{k,j}^2\right\rangle_{\boldsymbol{c}_k,\boldsymbol{a}_k}\right| \\
&= \left|\left\langle f'(\boldsymbol{q}_k^\top \boldsymbol{c}_k + \textstyle\sum_{j\neq i} w_{k,j}a_{k,j} + \chi_{k,i})w_{k,i}a_{k,i}^2\right\rangle_{\boldsymbol{c}_k,\boldsymbol{a}_k}\right| \\
&\leq C\left|w_{k,i}\right|. \qquad \text{(S-41)}
\end{aligned}$$

The second line is due to the fact $a_{k,i}$ is zero-mean, and in reaching the last line, we used the boundness of f'. Similarly, we can get

$$\left|\left\langle \widetilde{a}_{2k,i}\widetilde{f}(\widetilde{\boldsymbol{y}}_{2k}^\top \boldsymbol{w}_k)\right\rangle_{\widetilde{\boldsymbol{c}}_{2k},\widetilde{\boldsymbol{a}}_{2k}}\right| \leq C\left|w_{k,i}\right|. \qquad \text{(S-42)}$$

Substituting (S-41) and (S-42) into (S-40), we prove (S-40). □

Now we are in the position to prove Lemma 2.

Proof of Lemma 2. Because of the exchangeability, $\mathbb{E}\, w_{k,i}^4 = \mathbb{E}\, w_{k,j}^4$, and $\mathbb{E}\,[\boldsymbol{V}_k]_{i,\ell}^4 = \mathbb{E}\,[\boldsymbol{V}_k]_{j,\ell}^4$ for all $i, j = 1, 2, \ldots, n$ and $\ell = 1, 2, \ldots, d$. Thus, we only need to prove (S-20) for any specific i.

We first prove $\mathbb{E}\, w_{k,i}^4 \leq C(T)n^{-2}$. We know that

$$\begin{aligned}
\mathbb{E}\, w_{k+1,i}^4 - \mathbb{E}\, w_{k,i}^4 = 4\mathbb{E}\left[w_{k,i}^3 \mathbb{E}_k\left(w_{k+1,i} - w_{k,i}\right)\right] + 6\mathbb{E}\left[w_{k,i}^2 \mathbb{E}_k\left(w_{k+1,i} - w_{k,i}\right)^2\right] \qquad \text{(S-43)} \\
+ 4\mathbb{E}\left[w_{k,i}\mathbb{E}_k\left(w_{k+1,i} - w_{k,i}\right)^3\right] + \mathbb{E}\mathbb{E}_k\left(w_{k+1,i} - w_{k,i}\right)^4.
\end{aligned}$$

From (S-1) and knowing that h, f and $\widetilde{f}$ are uniformly bounded, we have

$$\mathbb{E}_k\left(w_{k+1,i} - w_{k,i}\right)^\gamma \leq \frac{C}{n^\gamma}\left(1 + \|\boldsymbol{u}_i\|^\gamma + \|\boldsymbol{v}_{k,i}\|^\gamma + \left|w_{k,i}\right|^\gamma\right) \quad \text{for } \gamma = 2, 3, 4. \qquad \text{(S-44)}$$

Substituting (S-39) and (S-44) into (S-43) and using the Young's inequality, we have

$$\begin{aligned}
\mathbb{E}\, w_{k+1,i}^4 - \mathbb{E}\, w_{k,i}^4 &\leq \tfrac{C}{n}\left(n^{-2} + \mathbb{E}\,\|\boldsymbol{u}_i\|^4 + \mathbb{E}\,\|\boldsymbol{v}_{k,i}\|^4 + \mathbb{E}\, w_{k,i}^4\right). \\
&\leq \tfrac{C}{n}\mathbb{E}\left(n^{-2} + \textstyle\sum_{\ell=1}^d [\boldsymbol{V}_k]_{i,\ell}^4 + w_{k,i}^4\right), \qquad \text{(S-45)}
\end{aligned}$$

where the last line is due to Assumption A.4), which implies $\sum_\ell [\boldsymbol{U}]_{i,\ell}^4 \leq C$. Similarly, we can prove

$$\textstyle\sum_{\ell=1}^d \mathbb{E}\left([\boldsymbol{V}_{k+1}]_{i,\ell}^4 - [\boldsymbol{V}_k]_{i,\ell}^4\right) \leq \tfrac{C}{n}\mathbb{E}\left(n^{-2} + \sum_{\ell=1}^d [\boldsymbol{V}_k]_{i,\ell}^4 + w_{k,i}^4\right). \qquad \text{(S-46)}$$

Combining (S-45) and (S-46), we have

$$\mathbb{E}\,(w_{k+1,i}^4 + \textstyle\sum_{\ell=1}^d [\boldsymbol{V}_{k+1}]_{i,\ell}^4) - \mathbb{E}\left(w_{k,i}^4 + \sum_{\ell=1}^d [\boldsymbol{V}_k]_{i,\ell}^4\right) \leq \tfrac{C}{n}\left[n^{-2} + \mathbb{E}\left(w_{k,i}^4 + \textstyle\sum_{\ell=1}^d [\boldsymbol{V}_k]_{i,\ell}^4\right)\right].$$

Using the above inequality iteratively, we have

$$\mathbb{E}\left(w_{k,i}^4 + \textstyle\sum_{\ell=1}^d [\boldsymbol{V}_k]_{i,\ell}^4\right) \leq \left(n^{-2} + w_{0,i}^4 + \textstyle\sum_{\ell=1}^d [\boldsymbol{V}_0]_{i,\ell}^4\right) e^{\frac{k}{n}C}.$$

Since $\mathbb{E}\left(w_{0,i}^4 + \sum_{\ell=1}^d [\boldsymbol{V}_0]_{i,\ell}^4\right)$ are bounded in Assumption A.4), we now reach (S-20). □

S-V Local stability analysis of the fixed points of the ODE

In this section, we provide additional details on the local stability analysis of the ODE for Example 1. We first its simplified ODE (13) in the main text. Then, we provide the derivation of the local stability analysis when $d = 1$, where the main results are summarized in Section S-I. Finally, we establish the proof of Claim 1 in the main text.

S-V.1 Derive the reduced ODE for Example 1 when $\lambda \to \infty$

In Example 1, $f(x) = \widetilde{f}(x) = x$. Plugging back to (9), we obtain that

$$\begin{aligned} \boldsymbol{g}_t &= \boldsymbol{\Lambda} \boldsymbol{q}_t \\ \widetilde{\boldsymbol{g}}_t &= \widetilde{\boldsymbol{\Lambda}} \boldsymbol{r}_t \\ b_t &= \eta_{\mathrm{T}}(\boldsymbol{q}_t^\top \boldsymbol{\Lambda} \boldsymbol{q}_t + \eta_{\mathrm{T}} z_t) + \eta_{\mathrm{G}}(\boldsymbol{r}_t^\top \widetilde{\boldsymbol{\Lambda}} \boldsymbol{r}_t + \eta_{\mathrm{G}} z_t). \end{aligned} \tag{S-47}$$

Correspondingly, ODE in (8) becomes:

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}_t &= \widetilde{\tau}\big(\boldsymbol{q}_t \widetilde{\boldsymbol{r}}_t^\top \widetilde{\boldsymbol{\Lambda}} + \boldsymbol{P}_t \boldsymbol{L}_t\big) \\ \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}_t &= \tau\big(\boldsymbol{\Lambda}\boldsymbol{q}_t - \boldsymbol{P}_t \widetilde{\boldsymbol{\Lambda}} \boldsymbol{r}_t + \boldsymbol{q}_t h_t\big) \\ \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r}_t &= \tau\big(\boldsymbol{P}_t^T \boldsymbol{\Lambda}\boldsymbol{q}_t - \boldsymbol{S}_t \widetilde{\boldsymbol{\Lambda}} \boldsymbol{r}_t + \boldsymbol{r}_t h_t\big) + \widetilde{\tau}\big(\widetilde{\boldsymbol{\Lambda}}\boldsymbol{r}_t + \boldsymbol{L}_t \boldsymbol{r}_t\big) \\ \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{S}_t &= \widetilde{\tau}\big(\boldsymbol{r}_t \boldsymbol{r}_t^\top \widetilde{\boldsymbol{\Lambda}}^\top + \widetilde{\boldsymbol{\Lambda}} \boldsymbol{r}_t \boldsymbol{r}_t^\top + \boldsymbol{S}_t \boldsymbol{L}_t + \boldsymbol{L}_t \boldsymbol{S}_t\big) \\ \tfrac{\mathrm{d}}{\mathrm{d}t} z_t &= 2\tau(\boldsymbol{q}_t^\top \boldsymbol{\Lambda} \boldsymbol{q}_t - \boldsymbol{r}_t^\top \widetilde{\boldsymbol{\Lambda}} \boldsymbol{r}_t + z_t h_t) \\ &\quad + \tau^2[\eta_{\mathrm{T}}(\boldsymbol{q}_t^\top \boldsymbol{\Lambda} \boldsymbol{q}_t + z_t \eta_{\mathrm{T}}) + \eta_{\mathrm{G}}(\boldsymbol{r}_t^\top \widetilde{\boldsymbol{\Lambda}} \boldsymbol{r}_t + z_t \eta_{\mathrm{G}})] \end{aligned} \tag{S-48}$$

The first four equations are exactly (13). From last two equations of (S-48), by setting $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{diag}\{\boldsymbol{S}_t\} = \boldsymbol{0}$, $\frac{\mathrm{d}}{\mathrm{d}t} z_t = 0$, $\mathrm{diag}(\boldsymbol{S}_t) = \boldsymbol{I}$ and $z_t = 1$, we can get (14).

S-V.2 A complete study of all fixed points when $d = 1$

We next provide the local stability analysis of the fixed points of the ODE (13). When $d = 1$ and $\lambda \to \infty$, the macroscopic state is described by only 3 scalars, P_t, q_t and r_t. The result is summarized in Table 1. For the sake of simplicity, we only consider the case $\Lambda = \widetilde{\Lambda}$, and set $\eta_{\mathrm{T}} = \eta_{\mathrm{G}} = 1$, but all analysis can be extended to general cases.

The fixed points are given by the condition $\frac{d}{dt}P_t = \frac{d}{dt}q_t = \frac{d}{dt}r_t = 0$. From (13), we get

$$\begin{cases} \widetilde{\tau}\Lambda r\,(q - rP) = 0 \\ \tau\big[\Lambda - \tau - \Lambda\left(1 + \frac{\tau}{2}\right) q^2\big] q - \tau\Lambda\big[P + \left(\frac{\tau}{2} - 1\right) rq\big] r = 0 \\ \tau\Lambda P q + \big[\Lambda(\widetilde{\tau} - \tau) - \tau^2\big] r + \Lambda\big(\tau - \widetilde{\tau} - \frac{\tau^2\}}{2}\big) r^3 - \tau\Lambda\big(1 + \frac{\tau}{2}\big) r q^2 = 0, \end{cases} \tag{S-49}$$

where P, q, r are the stationary macroscopic state. The local stability of a fixed point is identified by whether the Jacobian matrix

$$J(P, q, r) \overset{\text{def}}{=} \begin{bmatrix} \frac{\partial}{\partial P} g_1 & \frac{\partial}{\partial q} g_1 & \frac{\partial}{\partial r} g_1 \\ \frac{\partial}{\partial P} g_3 & \frac{\partial}{\partial q} g_3 & \frac{\partial}{\partial r} g_3 \\ \frac{\partial}{\partial P} g_5 & \frac{\partial}{\partial q} g_5 & \frac{\partial}{\partial r} g_5 \end{bmatrix}$$

has eigenvalue with non-negative real part or not, where $g_1 = \widetilde{\tau}\Lambda r\,(q - rP)$, $g_2 = \tau\left[\Lambda - \tau - \Lambda\left(1 + \frac{\tau}{2}\right) q^2\right] q - \tau\Lambda\left[P + \left(\frac{\tau}{2} - 1\right) rq\right] r$ and $g_5 = \tau\Lambda P q + \big[\Lambda(\widetilde{\tau} - \tau) - \frac{(\eta_{\mathrm{T}} + \eta_{\mathrm{G}})\tau^2}{2}\big] r + \Lambda\left(\tau - \widetilde{\tau} - \frac{\tau^2 \eta_{\mathrm{G}}}{2}\right) r^3 - \tau\Lambda\left(1 + \frac{\tau\eta_{\mathrm{T}}}{2}\right) r q^2$.

Type (1) fixed point at $P = q = r = 0$

It is easy to verify that $q = r = 0$ and any $P \in [-1, 1]$ is a solution of (S-49), but we first consider $P = 0$.

The Jacobian at $P = q = r = 0$ is

$$J(0, 0, 0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \tau(\Lambda - \tau) & 0 \\ 0 & 0 & \Lambda\,(\widetilde{\tau} - \tau) - \tau^2 \end{bmatrix}.$$

Thus, type (1) fixed point is stable if and only if

$$\tau \geq \Lambda \quad \text{and} \quad \tfrac{\widetilde{\tau}}{\tau} \leq \tfrac{\tau + \Lambda}{\Lambda}.$$

Type (2) fixed points at $P = q = 0$, $r = \pm r^* \neq 0$

We first analyze when such fixed point exists and then study its local stability.

If $P = q = 0$, the first two equations in (S-49) trivially hold. The third equation becomes

$$\tau[\Lambda(r^2 - 1) - \tfrac{\tau}{2}(\Lambda r^2 + 2)] - \widetilde{\tau}\Lambda(r^2 - 1) = 0.$$

The solution is

$$r^2 = \frac{\tau - \widetilde{\tau} + \tau^2/\Lambda}{\tau - \widetilde{\tau} - \tau^2/2}. \tag{S-50}$$

Since only the positive solution corresponds a fixed one. Thus, type (2) fixed point exists if

$$\tfrac{\widetilde{\tau}}{\tau} \leq 1 - \tfrac{\tau}{2} \tag{S-51}$$

$$\text{or} \quad \tfrac{\widetilde{\tau}}{\tau} \geq \tfrac{\tau+\Lambda}{\Lambda}. \tag{S-52}$$

Next, we investigate the local stability of this fixed point. The Jacobian at $\tilde{q} = q = 0$ for a given r is

$$J(0,0,r) = \begin{bmatrix} -\widetilde{\tau}\Lambda r^2 & \widetilde{\tau}\Lambda r & 0 \\ -\tau\Lambda r & \tau(\Lambda - \tau) - \Lambda\tau(\tfrac{\tau}{2} - 1)r^2 & 0 \\ 0 & 0 & 3r^2\Lambda(\tau - \frac{\tau^2}{2} - \widetilde{\tau}) - \tau^2 + \Lambda(\widetilde{\tau} - \tau) \end{bmatrix} \tag{S-53}$$

Plugging (S-50) into $[J(0,0,r)]_{3,3}$ of (S-53), then $[J(0,0,r)]_{3,3} \leq 0$ implies

$$\tfrac{\widetilde{\tau}}{\tau} \geq \tfrac{\tau}{\Lambda} + 1.$$

It indicates that the stationary points at the region (S-51) are always unstable. Thus, we only need to consider the second region specified by (S-52).

For the upper-left 2×2 sub-matrix of (S-53), the eigenvalues are non-positive if and only if

$$-\widetilde{\tau}\Lambda r^2 + \tau(\Lambda - \tau) - \Lambda\tau(\tfrac{\tau}{2} - 1)r^2 \leq 0 \tag{S-54}$$

$$\tau + \Lambda(\tfrac{\tau}{2} - 1)r^2 + \Lambda - \Lambda \geq 0. \tag{S-55}$$

Plugging (S-50) into (S-54), we can get

$$\tfrac{\widetilde{\tau}}{\tau} \geq 2. \tag{S-56}$$

Plugging (S-50) into (S-55) and combining (S-52), we can get

$$[\tau + \Lambda(\tfrac{\tau}{2} - 1)]\tilde{\tau} \geq \tau\Lambda(\tfrac{\tau}{2} - 1).$$

Solving this inequality implies that

$$\frac{\tilde{\tau}}{\tau} \leq \frac{(\frac{\tau}{2} - 1)\Lambda}{(\frac{\tau}{2} - 1)\Lambda + \tau}, \text{ when } \tau < \frac{2\Lambda}{\Lambda + 2} \tag{S-57}$$

and

$$\frac{\tilde{\tau}}{\tau} \geq \frac{(\frac{\tau}{2} - 1)\Lambda}{(\frac{\tau}{2} - 1)\Lambda + \tau}, \text{ when } \tau > \frac{2\Lambda}{\Lambda + 2}. \tag{S-58}$$

Note that (S-58) is included by (S-56), as $\frac{(\frac{\tau}{2}-1)\Lambda}{(\frac{\tau}{2}-1)\Lambda+\tau} \leq 2$ when $\tau > \frac{2\Lambda}{\Lambda+2}$.

Then, combining (S-52), (S-56), and (S-57) we obtain the stability region for $\tilde{q} = q = 0$,

$$\frac{\tilde{\tau}}{\tau} \geq 1 + \frac{\tau}{\Lambda}, \frac{\tilde{\tau}}{\tau} \geq 2, \text{ and } \frac{\tilde{\tau}}{\tau} \leq \beta(\tau),$$

where $\beta(\tau)$ is defined as

$$\beta(\tau) \stackrel{\text{def}}{=} \begin{cases} \frac{(\frac{\tau}{2}-1)\Lambda}{(\frac{\tau}{2}-1)\Lambda+\tau} & \text{if } \tau \leq \frac{2\Lambda}{\Lambda+2} \\ +\infty & \text{otherwise.} \end{cases}$$

Type (3) fixed points at $q = r = 0$ and $|P| \in (0, 1]$

As mentioned, we can check that $q = r = 0$ and any $P \in [-1, 1]$ is a solution of (S-49). We next investigate the stable region for the fixed point $P = \pm 1$ and $q = r = 0$, which represents the perfect recovery state. For general P, we can analyze its fixed point similarly.

The Jacobian at $q = r = 0$ for any given P is

$$J(1, 0, 0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \tau(\Lambda - \tau) & -\tau\Lambda \\ 0 & \tau\Lambda & \Lambda\left(\widetilde{\tau} - \tau\right) - \tau^2 \end{bmatrix}.$$

In this case, $J(1, 0, 0)$ always has an eigenvalue 0 and to calculate the rest two eigenvalues, we only need to analyze the bottom-right 2×2 sub-matrix of $J(\tilde{q})$. The characteristic polynomial of this sub-matrix is $f(\lambda) = \lambda^2 - (a + d)\lambda + ad - bc$,where $a = \tau(\Lambda - \tau)$, $b = -\tau\Lambda$, $c = \tau\Lambda$, and $d = \Lambda\left(\widetilde{\tau} - \tau\right) - \tau^2$. The roots of $f(\lambda) = 0$ both have non-positive real part if and only if $a + d \leq 0$, $ad - bc \geq 0$, which implies

$$\tfrac{\widetilde{\tau}}{\tau} \leq \tfrac{2\tau}{\Lambda} \quad \text{and} \quad \tfrac{\widetilde{\tau}}{\tau}(\tau - \Lambda) \leq \tfrac{\tau^2}{\Lambda}. \tag{S-59}$$

Noting that when $\tau < \Lambda$, the second inequality always hold, and when $\tau > \Lambda$, $\frac{\tau^2}{\Lambda(\tau - \Lambda)} \geq 4$, we can combine the two inequalities in (S-59) into compact form

$$\tfrac{\widetilde{\tau}}{\tau} \leq \min\{\tfrac{2\tau}{\Lambda}, \max\{\tfrac{\tau^2}{\Lambda|\tau - \Lambda|}, 4\}\}.$$

The stable regions of the fixed points for $q = r = 0$ and $|P| < 1$ can be derived in a similar way, which turns out to be a subset of the stable region for $P = \pm 1$.

Type (4) fixed point at $P = r = 0$ and $q \neq 0$.

From (S-49), we know when at fixed point, $\tilde{q} = r = 0$, then $q^2 = \frac{\Lambda - \tau}{\Lambda(1 + \tau/2)}$, so τ must satisfy $\tau \leq \Lambda$. The corresponding Jacobian is:

$$J(0, 0, q) = \begin{bmatrix} 0 & 0 & \widetilde{\tau}\Lambda q \\ 0 & \tau(\Lambda - \tau) - 3\tau\Lambda q^2(1 + \frac{\tau}{2}) & 0 \\ \tau\Lambda q & 0 & (\widetilde{\tau} - \tau)\Lambda - \tau^2 - \tau\Lambda q^2(1 + \frac{\tau}{2}) \end{bmatrix}.$$

After plugging in $q^2 = \frac{\Lambda - \tau}{\Lambda(1 + \tau/2)}$, we can obtain that the characteristic function $\det(\lambda \boldsymbol{I} - J(0, 0, q))$ is equal to:

$$\det(\lambda \boldsymbol{I} - J(0, 0, q)) = [\lambda + 2\tau(\Lambda - \tau)][\lambda(\lambda + (2\tau - \widetilde{\tau})\Lambda) - \tau\widetilde{\tau}\Lambda^2 q^2]$$

Clearly, $\det(\lambda \boldsymbol{I} - J(0, 0, q)) = 0$ has a non-negative root, so $J(0, 0, q)$ always has a non-negative eigenvalue. This means type (4) fixed points are always unstable.

Type (5) fixed points at $P, q, r \neq 0$

The fixed points equation (S-49) can also have solutions that none of P, q and r is zero. In what follows, we derive the analytical expression of this type of solutions. It turns out that there can be maximum 8 solutions, which are symmetric by flipping the signs. We are unable to derive the analytical expression for their stable region, but it can be computed numerically.

If $P, q, r \neq 0$, (S-49) yields

$$r = \tfrac{q}{P} \tag{S-60}$$

$$\Lambda - \tau - \Lambda(1 + \tfrac{\tau}{2})q^2 - \Lambda[\tfrac{P}{q} + (\tfrac{\tau}{2} - 1)r]r = 0 \tag{S-61}$$

$$\tau\Lambda\widetilde{P}q + r\left[\Lambda(\widetilde{\tau} - \tau) - \tau^2\right] + r^3\Lambda\left(\tau - \widetilde{\tau} - \tfrac{\tau^2}{2}\right) - rq^2\tau\Lambda\left(1 + \tfrac{\tau}{2}\right) = 0. \tag{S-62}$$

Plugging (S-60) into (S-61), we can get

$$q^{-2} = -\tfrac{1}{\tau}[\Lambda(\tfrac{\tau}{2} - 1)P^{-2} + \Lambda(1 + \tfrac{\tau}{2})]. \tag{S-63}$$

Then combining (S-60) (S-63) and (S-62), we can obtain the following equations:

$$AP^{-4} + BP^{-2} + C = 0 \tag{S-64}$$

where $A = \Lambda(\widetilde{\tau} - \tau)(\frac{1}{2} - \frac{1}{\tau}) + \widetilde{\tau}$, $B = \Lambda[\frac{\widetilde{\tau}}{\tau}(1 + \frac{\tau}{2}) - 2]$, $C = \Lambda(1 + \frac{\tau}{2})$. We can find that (S-64) is an equation of P^{-2} with at most two roots. Combining (S-63), we know there are at most 2 solutions for the pair $(q^{-2},\ P^{-2})$ and hence there are at most 8 solutions for $(q,\ P, r)$, where $r = P/q$.

S-V.3 Proof of Claim 1

Proof of Claim 1. We first compute the Jacobian $\partial\{\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{P}_t, \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{q}_t, \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r}_t, \}/\partial\{\boldsymbol{P}_t, \boldsymbol{q}_t, \boldsymbol{r}_t\}$ of the ODE (13) when $\boldsymbol{q}_t = \boldsymbol{r}_t = \boldsymbol{0}$. In the Jacobian, the $d \times d$ matrix $\boldsymbol{P}_t$ is considered as a d^2 vector. In fact, all elements in the Jacobian matrix related to $\boldsymbol{P}_t$ are 0. Specifically, the Jacobian for any $\boldsymbol{P}$ and $\boldsymbol{q}_t = \boldsymbol{r}_t = \boldsymbol{0}$ is

$$\boldsymbol{J}(\boldsymbol{P}) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \tau(\boldsymbol{\Lambda} - \tau\overline{\eta^2}\boldsymbol{I}_d) & -\tau\boldsymbol{P}\widetilde{\boldsymbol{\Lambda}} \\ \boldsymbol{0} & \tau\boldsymbol{P}^\top\boldsymbol{\Lambda} & \widetilde{\boldsymbol{\Lambda}}(\widetilde{\tau} - \tau) - \tau^2\overline{\eta^2} \end{bmatrix}, \tag{S-65}$$

where $\overline{\eta^2} = (\eta_{\mathrm{T}}^2 + \eta_{\mathrm{G}}^2)/2$.

When $\boldsymbol{P}$ is diagonal, under a suitable column-row permutation, the $\boldsymbol{J}(\boldsymbol{P})$ in (S-65) becomes a block diagonal matrix, where each non-zero block is a 2×2 matrix

$$\begin{bmatrix} \tau([\boldsymbol{\Lambda}]_{\ell,\ell} - \tau\overline{\eta^2}) & -\tau[\boldsymbol{P}]_{\ell,\ell}[\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell} \\ \tau[\boldsymbol{P}]_{\ell,\ell}[\boldsymbol{\Lambda}]_{\ell,\ell} & [\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell}(\widetilde{\tau} - \tau) - \tau^2\overline{\eta^2} \end{bmatrix} \tag{S-66}$$

for $\ell = 1, 2, \ldots, d$. Intuitively, the above matrix is the Jacobian matrix of $\partial\{\frac{\mathrm{d}}{\mathrm{d}t}[\boldsymbol{q}_t]_\ell,\ \frac{\mathrm{d}}{\mathrm{d}t}[\boldsymbol{r}_t]_\ell\}/\partial\{[\boldsymbol{q}_t]_\ell,\ [\boldsymbol{r}_t]_\ell\}$, and the Jacobian $\partial\{\frac{\mathrm{d}}{\mathrm{d}t}[\boldsymbol{q}_t]_\ell,\ \frac{\mathrm{d}}{\mathrm{d}t}[\boldsymbol{r}_t]_\ell\}/\partial\{[\boldsymbol{q}_t]_{\ell'},\ [\boldsymbol{r}_t]_{\ell'}\}$ is zero for $\ell \neq \ell'$.

Now the problem reduces into investigate eigenvalues of n 2-by-2 matrices. For any given $\ell = 1, 2, \ldots, n$, we have studied this problem in Section S-V.2 (type (1) and type (3) fixed points).

Specifically, the perfect recovery point $\boldsymbol{P} = \boldsymbol{I}$, $\boldsymbol{q} = \boldsymbol{r} = \boldsymbol{0}$ is stable if and only if $\lambda_{\max}(\boldsymbol{J}(\boldsymbol{P})) \leq 0$, where $\boldsymbol{J}(\boldsymbol{P})$ is defined in (S-65). Similar to the analysis of the type (3) fixed points in Section S-V.2, the condition that both eigenvalues of the matrix in (S-66) is non-positive implies

$$\tfrac{1}{2}([\boldsymbol{\Lambda}]_{\ell,\ell} - [\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell} + \alpha[\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell}) \leq \tau\overline{\eta^2} \tag{S-67}$$

$$\text{and} \quad \alpha(\tau\overline{\eta^2} - [\boldsymbol{\Lambda}]_{\ell,\ell}) \leq \tfrac{\tau\overline{\eta^2}}{[\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell}}(\tau\overline{\eta^2} - [\boldsymbol{\Lambda}]_{\ell,\ell} + [\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell}), \tag{S-68}$$

for all $\ell = 1, 2, \ldots, n$. The inequality (S-67) is the first inequality of (15) in Claim 1 in the main text.

Next, we investigate the condition when the trivial fixed point of the origin $\boldsymbol{P} = \boldsymbol{0}$ and $\boldsymbol{q} = \boldsymbol{r} = \boldsymbol{0}$ is unstable. Put $\boldsymbol{P} = \boldsymbol{0}$ into (S-66), we get a diagonal matrix

$$\begin{bmatrix} \tau([\boldsymbol{\Lambda}]_{\ell,\ell} - \tau\overline{\eta^2}) & 0 \\ 0 & [\widetilde{\boldsymbol{\Lambda}}]_{\ell,\ell}(\widetilde{\tau} - \tau) - \tau^2\overline{\eta^2} \end{bmatrix}.$$

When any eigenvalue of the above matrices for $\ell = 1, 2, \ldots, n$ is positive, this trivial fixed point will be unstable. A sufficient condition is the first eigenvalues of all matrices are positive:

$$\tau\overline{\eta^2} < [\boldsymbol{\Lambda}]_{\ell,\ell} \text{ for all } \ell = 1, 2, \ldots, n. \tag{S-69}$$

The above inequality is the second inequality of (15) in the main text. In addition, (S-69) implies (S-68) hold as the left hand side of (S-68) is negative. Now, we prove that (15) is a sufficient condition that the perfect fixed point is stable and the trivial fixed point is unstable.

□

We further note that (15) is not a necessary condition. There may be a region that (S-69) does not hold, but the origin is still unstable, and the perfect recovery point is stable. Such region is hard to characterize analytically, and numerically, we found the training algorithms always converge to other bad fixed points (e.g. mode collapsing state, or a state that $\boldsymbol{P}$ and $\boldsymbol{q}$ are still zero, but $\boldsymbol{r}$ is non-zero. The situation of the latter is similar to the noninfo-2 phase in the $d = 1$ case, which converges to the type (2) fixed point). Further study on those bad fixed points will be established in future works under a more general model.