# Learning Dynamics in Linear VAE: Posterior Collapse Threshold, Superfluous Latent Space Pitfalls, and Speedup with KL Annealing

**Yuma Ichikawa**
University of Tokyo, Meguro, Tokyo, Japan
Fujitsu Limited, Kawasaki, Kanagawa, Japan
ichikawa-yuma1@g.ecc.u-tokyo.ac.jp
ichikawa.yuma@fujitsu.com

**Koji Hukushima**
University of Tokyo, Meguro, Tokyo, Japan
k-hukushima@g.ecc.u-tokyo.ac.jp

## Abstract

Variational autoencoders (VAEs) face a notorious problem wherein the variational posterior often aligns closely with the prior, a phenomenon known as posterior collapse, which hinders the quality of representation learning. To mitigate this problem, an adjustable hyperparameter $\beta$ and a strategy for annealing this parameter, called KL annealing, are proposed. This study presents a theoretical analysis of the learning dynamics in a minimal VAE. It is rigorously proved that the dynamics converge to a deterministic process within the limit of large input dimensions, thereby enabling a detailed dynamical analysis of the generalization error. Furthermore, the analysis shows that the VAE initially learns entangled representations and gradually acquires disentangled representations. A fixed-point analysis of the deterministic process reveals that when $\beta$ exceeds a certain threshold, posterior collapse becomes inevitable regardless of the learning period. Additionally, the superfluous latent variables for the data-generative factors lead to overfitting of the background noise; this adversely affects both generalization and learning convergence. The analysis further unveiled that appropriately tuned KL annealing can accelerate convergence.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

## 1 INTRODUCTION

Deep latent variable models are generative models that convert latent variables generated from a prior distribution into samples that closely resemble data through a neural network. Variational autoencoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014), one of the deep latent variable models, have been applied in various fields such as image generation (Child, 2020; Vahdat and Kautz, 2020), text generation (Bowman et al., 2015), music generation (Roberts et al., 2018), clustering (Jiang et al., 2016), dimensionality reduction (Akkari et al., 2022), data augmentation (Norouzi et al., 2020), and anomaly detection (An and Cho, 2015; Park et al., 2022). The objective function of the VAE can be decomposed into the reconstruction error (*distortion*) and KL divergence term (*rate*), which have different roles and a trade-off relationship. In practice, VAEs are generally trained with the $\beta$-VAE objective (Higgins et al., 2016), which balances the reconstruction error and KL divergence term by introducing a weight parameter $\beta$.

In addition to data generation tasks, $\beta$-VAEs are state-of-the-art models for representation learning. In particular, $\beta$-VAEs have gained attention owing to their capability for obtaining representations in which a single latent variable is sensitive to changes in a single generative factor and is relatively invariant to changes in other factors (Higgins et al., 2016). This property of representations is called "disentanglement". For example, a disentangled representation of 3D objects is sensitive to a single independent data-generative factor, such as object identity, position, scale, and color. In $\beta$-VAE, the degree of disentanglement can be controlled by tuning the weight $\beta$. However, this $\beta$-tuning causes a notorious problem in which the variational posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ tends to align with the prior $p(\boldsymbol{z})$ during learning, thereby hindering the quality of representation learning. This phenomenon is commonly referred to as "posterior collapse".

Although several studies have theoretically analyzed the relationship between tuning, disentanglement, and posterior collapse, the understanding remains limited. In particular, the learning dynamics of -VAEs have not been fully explored thus far. On the other hand, several attempts have been made to mitigate the posterior collapse (Yang et al., 2017; Dieng et al., 2019; Zhao et al., 2017; Kim et al., 2018). Among these, the simplest strategy is monotonic KL annealing, in which the weight is scheduled to gradually increase during training (Bowman et al., 2015). Although this heuristic method is recognized for its e ectiveness, it cannot be guaranteed owing to its limited theoretical understanding.

This study theoretically analyzes a minimal model known as a linear VAE (Lucas et al., 2019), which captures the essence of -VAEs. Our results elucidate the formation process of disentangled features, the relationship between and the posterior collapse, and the e ect of super uous latent variables on the generative factors. In addition, we reveal the in uence of KL annealing on the generalization performance.

Contributions. This study develops a theory of learning dynamics for VAEs. Speci cally, this study rigorously proved that the one-pass gradient descent dynamics (SGD) converges to a deterministic process characterized by ordinary di erential equations (ODEs) within the limit of large input dimensions, thereby providing the asymptotically exact dynamics of the generalization error. Consequently, the relationships between the generalization error, the posterior collapse, the disentanglement, and are revealed in two scenarios: the \model-matched case" wherein the number of generative factors in the training data matches the dimension of the latent space, and the \model-mismatched case" wherein the latent dimension exceeds the number of the generative factors. The main contributions of this study are as follows.

- An asymptotic exact analysis of the learning dynamics by the one-pass SGD is derived. The results demonstrate that the learning dynamics converge to a deterministic process characterized by ODEs within the limit of large input dimensions.
- The stability analysis of the xed points of the limiting ODEs indicates that when exceeds a certain threshold, posterior collapse is inevitable regardless of the learning time.
- Theoretical analysis of the well-known replica method in statistical mechanics and the dynamics of SGD are shown to have a complementary relationship. Speci cally, a steady state of the SGD dynamics coincides exactly with the global optimum derived by the replica method, indicating the reachability to the global optimum using SGD.
- The numerical integration of the ODEs uncovers a phenomenon, wherein the VAE initially learns entangled representations and gradually acquires those that are disentangled. The stability of xed points indicates that disentangled representations can be achieved for any .
- The analysis of the model-mismatched case demonstrates that the super uous latent variable over ts the background noise with a small , degrading generalization. The stability of the xed points reveals that while an optimal generalization is achieved for the same value in both the model-matched and model-mismatched cases, the convergence time for the model-mismatched case is signi cantly longer.
- Appropriately tuned KL annealing accelerates the convergence of learning. Additionally, the stability analysis provides a speci c annealing rate beyond which the convergence decelerates.

### 1.1 Preliminaries

Here, we summarize the notations. The expression $k \quad k_F$ denotes the Frobenius norm. $I_N \ 2 \ R^{N \ N}$ denotes an $N \quad N$ identity matrix, whereas $0_N$ denotes the vector $(0; : : : ; 0)^> \ 2 \ R^N$. $D_{KL}[\,k\,]$ denotes the Kullback{Leibler (KL) divergence.

## 2 BACKGROUND

### 2.1 Variational Autoencoders

The VAE (Kingma and Welling, 2013) is a latent generative model. Let $D = f x \ g_{=1}^{P}$ with $x \ 2 \ R^D$ be the training data, and $p_D(x)$ indicate the empirical distribution of the training dataset. In practical applications, VAEs are typically trained using the -VAE objective (Higgins et al., 2016) de ned by

$$E_{p_D} \ E_q \ [ \ \log p \ (x j z)] + \ D_{KL}[q \ (z j x) k p(z)]$$
$$= E_{p_D}[l( \ ; \ ; x; \ )]; \quad (1)$$

where $p(z)$ is a prior for the latent variables, and the parameter 0 is introduced to control the trade-o between the rst and second terms in Eq. (1). Distributions $p \ (x j z)$ characterized by parameters and $q \ (z j x)$ by are commonly referred to as the decoder and encoder, respectively. Subsequently, VAEs optimize both the encoder parameters and decoder parameters by minimizing the objective of Eq. (1).

Note that when $\beta = 0$, the objective becomes a deterministic autoencoder that focuses more on minimizing the first term, which is referred to as the reconstruction error.

### 2.2 Posterior Collapse and KL Annealing

A notorious problem in VAE optimization is that the variational posterior $q(z|x)$ frequently aligns closely with the prior $p(z)$, a phenomenon which is known as posterior collapse, hindering the quality of representation learning. Several attempts have been made to mitigate this problem (Yang et al., 2017; Dieng et al., 2019; Zhao et al., 2017; Kim et al., 2018), among which a simple remedy called monotonic KL annealing has been proposed in (Bowman et al., 2015), where $\beta = 0$ is set at the beginning of the training and gradually increases until $\beta = 1$ is reached. In practice, $\beta$ is defined as follows:

$$\beta^{t+1} \leftarrow \beta^{t} + \epsilon \tag{2}$$

where $t$ denotes the index of each step of the parameter updates using an optimization algorithm, and $\epsilon \in \mathbb{R}$ represents the annealing rate. Monotonic annealing has become a standard method for training VAEs, particularly in numerous natural language processing applications. Although this heuristic is simple and often effective, it is not theoretically guaranteed. Additionally, cyclical KL Annealing (Fu et al., 2019) was utilized, which repeatedly applies monotonic KL annealing in a cyclical manner.

## 3 SETTING

**Generative Model for Dataset.** We derive our theoretical results for dataset $\mathcal{D} = \{x^{\mu}\}_{\mu=1}^{P}$ drawn from spiked covariance model (SCM) (Wishart, 1928; Potters and Bouchaud, 2020), which has been widely studied in statistics to analyze the performance of unsupervised learning methods such as principle component analysis (PCA) (Ipsen and Hansen, 2019; Biehl and Mietzner, 1993; Hoyle and Rattray, 2004), sparse PCA (Lesieur et al., 2015), and deterministic autoencoders (Refinetti and Goldt, 2022). Specifically, the dataset are sampled according to

$$x = \sqrt{\frac{\gamma}{N}} W c + \sqrt{\Delta} n \ ; \tag{3}$$

where $W \in \mathbb{R}^{N \times M}$ is a deterministic unknown feature matrix with $M$ features, $c \in \mathbb{R}^{M}$ is a random vector drawn from a standard normal distribution $\mathcal{N}(0_M; I_M)$, $n$ is a background noise vector whose components are i.i.d. from the standard normal distribution $\mathcal{N}(0_N; I_N)$, and $\Delta \in \mathbb{R}$ and $\gamma \in \mathbb{R}$ are the scalar parameters that control the strength of the noise and signal, respectively. Despite $W$ not being orthogonal, $Wc$ can be rewritten as $(WR)(R^{-1}c)$, where $R$ is a matrix that orthogonalizes and normalizes the columns of $W$. This can be considered as an equivalent system in which the new feature vector is $R^{-1}c$. Therefore, without the loss of generality, we assume that $(W)^{\top} W = I_M$.

Figure 1: The architectures of spiked covariance model (SCM) and linear variational autoencoder (VAE).

**Spectral of Covariance Matrix of the Dataset.** The spectrum of the empirical covariance matrix of $\mathcal{D}$ is characterized by $W$ and $c$. When $c = 0$, the dataset are Gaussian vectors, whose empirical covariance matrix with $P = O(N)$ samples has a Marchenko-Pastur distribution characterized by the noise strength $\Delta$ (Marchenko and Pastur, 1967). In contrast, by sampling the $c \sim p(c)$, the covariance matrix has $M$ eigenvalues, i.e., "spike", with the columns of $W$ as the corresponding eigenvectors. The remaining $N - M$ eigenvalues, i.e., "bulk", of the empirical covariance matrix still follow the Marchenko-Pastur distribution. This spectral is similar to that of the empirical covariance matrix of real datasets such as CIFAR10 and MNIST as explained in Refinetti and Goldt (2022). Moreover, the validity of the assumption of SCM (i.e., Gaussian model) as a realistic data distribution has recently been supported by "Gaussian Equivalence", which indicates that the learning dynamics with real data, irrespective of the machine learning models, closely agree with those with the Gaussian model with the empirical covariance matrix of the data (Liao and Couillet, 2018; Mei and Montanari, 2022; Hu and Lu, 2022; Goldt et al., 2022).

**Linear VAE Model** The linear VAE model (Dai et al., 2018; Lucas et al., 2019; Sicks et al., 2021) con-

sists of a linear decoder and encoder given by

$$p_W(x|z) = \mathcal{N}\left(x; \frac{1}{\sqrt{N}} W z, I_N\right); \quad (4)$$

$$q_{V,D}(z|x) = \mathcal{N}\left(z; \frac{1}{\sqrt{N}} V^\top x, D\right); \quad (5)$$

$$p(z) = \mathcal{N}(z; 0_N, I_N); \quad (6)$$

where the diagonal covariance matrix $D \in \mathbb{R}^{M \times M}$ indicates the learning parameters, and $W \in \mathbb{R}^{N \times M}$ and $V \in \mathbb{R}^{N \times M}$ also indicate the learning parameters. We assume a fixed identity covariance matrix $I_N$ because it is often used in practice. The architectural diagram is shown in Fig. 1.

Training Algorithm. The VAE is trained to learn the generative model using the following optimization problem:

$$(W(D), V(D), D(D)) = \operatorname*{argmin}_{W,V,D} R(W, V, D, D, \gamma, \lambda); \quad (7)$$

where

$$R(W, V, D, D, \gamma, \lambda) = \sum_{\mu=1}^{P} l(W, V, D, x^\mu, \gamma) + \frac{\lambda}{2}\|W\|_F^2 + \frac{\lambda}{2}\|V\|_F^2. \quad (8)$$

Here, $l(W, V, D, x^\mu, \gamma)$ is defined by Eq. (1), and the last two terms regulate the magnitudes of the parameters $W$ and $V$ which is called weight decay, where $\lambda > 0$ is a regularization parameter. Many practitioners often include a weight decay term in VAE training (Kingma and Welling, 2013; Louizos et al., 2017). This study broadens the theory to cover such situations. The following theoretical results are also applicable to scenarios without weight decay by setting $\lambda = 0$.

We consider a standard training algorithm using the stochastic gradient descent to solve the optimization problem defined in Eq. (7). To simplify the theoretical analysis, we assume a one-pass setting, where each data sample $x^\mu$ is used only once. At $t$ steps, the model parameters $W^t$, $V^t$ and $D^t$ are updated using a new sample $x^t$ according to the following:

$$W^{t+1} = W^t - \eta_W \nabla_{W^t} r(W^t, V^t, D^t, \gamma, \lambda, x^t); \quad (9)$$

$$V^{t+1} = V^t - \eta_V \nabla_{V^t} r(W^t, V^t, D^t, \gamma, \lambda, x^t); \quad (10)$$

$$D^{t+1} = D^t - \eta_D \nabla_{D^t} r(W^t, V^t, D^t, \gamma, \lambda, x^t)/N; \quad (11)$$

where $r$ represents the loss for a given sample defined as follows:

$$r(W^t, V^t, D^t, \gamma, \lambda, x^t) = l(W^t, V^t, D^t, x^t, \gamma) + \frac{\lambda}{2N}\|W^t\|_F^2 + \frac{\lambda}{2N}\|V^t\|_F^2.$$

Parameters $\eta_W$, $\eta_V$ and $\eta_D$ in the expressions above are the learning rates. The SGD algorithm characterizes a Markov process $X^t = [W^t, V^t, D^t]$ with an updated rule. Hereafter, $X^t$ is referred to as the microscopic state. Note that the analysis presented in this study can be naturally extended to the mini-batch SGD where the mini-batch size remains a finite number, that is, $O(N^0)$.

Generalization Metric. The VAE can generate a sample $x \sim p_W(x)$ through the following procedure. First, a latent variable $z \sim p(z)$ is generated, followed by a sample $x \sim p_W(x|z)$. Thus, we evaluate the generalization performance for the data distribution $p(x)$ defined as Eq. 3 using the following metric:

$$\mathbb{E}_c D_{KL}[p(x|c)\|p_W(x|c)] = \frac{1}{N}\mathbb{E}_c\left[\left\|\sqrt{\Sigma}\left(W - W^*\right)c\right\|^2\right];$$

where $\mathbb{E}_c[\cdot]$ is the average over $p(c) = \mathcal{N}(0_M, I_M)$; thus, we define the generalization error $\epsilon_g$ as

$$\epsilon_g(W; W^*) = \frac{1}{N}\mathbb{E}_c\left[\left\|\sqrt{\Sigma}\left(W - W^*\right)c\right\|^2\right]. \quad (12)$$

The generalization error, $\epsilon_g$, measures the extent of the signal recovery from the training data.

## 4 MACROSCOPIC DYNAMICS OF VAE

From a statistical physics perspective, $\epsilon_g(W; W^*)$ can be expressed as a function of the following set of macroscopic variables, called *order parameters*. Based on this idea, we attempt to express the dynamics of $\epsilon_g(W^t; W^*)$ by explicitly using the time evolution of the order parameters.

**Definition 4.1.** For $X^t = [W^t, V^t, D^t]$, the macroscopic variables are defined as follows:

$$m^t = \frac{1}{N}(W^t)^\top W^*; \quad d^t = \frac{1}{N}(V^t)^\top W^*;$$
$$Q^t = \frac{1}{N}(W^t)^\top W^t; \quad E^t = \frac{1}{N}(V^t)^\top V^t;$$
$$R^t = \frac{1}{N}(W^t)^\top V^t.$$

Subsequently, to compactly represent the macroscopic variables, the macroscopic state $M^t$ of the Markov chain in $X^t$ is defined as follows:

$$M^t = (m^t, d^t, Q^t, E^t, R^t, V^t, D^t) \in \mathbb{R}^{M(2M+5M)}.$$

Intuitively, the overlaps $m_{ij}^t$ and $d_{ij}^t$ measure the similarity to the $j$-th representation of the true model, i.e., the $j$-th column of $W^*$; the overlaps $Q_{ij}^t$, $E_{ij}^t$, and $R_{ij}^t$ measure the similarities between the decoder

weights, specically the $i$-th and $j$-th columns of $W^t$, the encoder weights, i.e., the $i$-th and $j$-th columns of $V^t$, and between the decoder and encoder weights, i.e., the $i$-th column of $W^t$ and the $j$-th column of $V^t$, respectively. The o-diagonal elements of $E^t$ represent the independence of the encoded representations. Thus, if the o-diagonal elements of $E^t$ are zero, a disentangled representation is obtained; otherwise, an entangled representation is obtained.

We investigate the dynamics of the training algorithm expressed by Eq. (9)-(11) for the macroscopic variables. Our rst contribution is to provide rigorous theoretical results under the following assumptions:

(A.1) The sequences $c^t$ and $n^t$ for $t = 1; : : :;$ are i.i.d. random variables, and $c^t$ is drawn from the standard normal distribution $N(0_M; I_M)$.

(A.2) The sequence $n^t$ is drawn from the standard normal distribution $N(0_N; I_N)$, and $f n^t g$ is independent of $f c^t g$.

(A.3) The initial macroscopic state $M^0$ satises $EkM^0 \quad M^0 k_F \quad C=\sqrt{N}$, where $M^0$ is a deterministic matrix and $C$ is a constant independent of $N$.

(A.4) For $i = 1; 2; : : :; N$, the initial microscopic state $X^0 = [W^0; V^0; D^0]$ satises $E[\sum_{m=1}^{M} f(W_{im}^0)^4 + (V_{im}^0)^4 + (D_m^0)^4 g + \sum_{m=1}^{M} (W_{im})^4] \quad C$, where $C$ is a constant independent of $N$ and $D^0$ s $0_M \quad M$.

Assumptions (A.1) and (A.2) for $c^t$ and $n^t$ can be relaxed to non-Gaussian cases if all moments are bounded; however, we use the Gaussian assumption to simplify the proof. Assumption (A.3) ensures that the initial macroscopic states converge to deterministic values as the input dimension $N$ approaches innity. Assumption (A.4) requires that the elements in the feature matrix $W$ and initial microscopic state $X^0$ are $O(1)$. The following theorem proves that the stochastic process of the macroscopic states converges to a deterministic process in the $N \to 1$ limit characterized by ODEs.

Theorem 4.2. For all $T > 0$, it holds under assumptions (A.1)-(A.4) that

$$\max_{0 \quad NT} EkM^t \quad M \quad (t=N)k_F \quad \frac{C(T)}{\sqrt{N}}; \tag{13}$$

where $C(T)$ is a constant that depends on $T$ but not on $N$, and $M(t)$ is a unique solution of the ODE

$$\frac{dM(t)}{dt} = F(M(t)); \tag{14}$$

with the initial condition $M(0) = M^0$ and $F : R^{M(2M+5M)} \to R^{M(2M+5M)}$ is uniformly Lipschitz continuous in $M(t)$. A specic expression is not demonstrated owing to its length; however, the entire function is provided in Supplementary Materials B.

The convergence theory of stochastic processes and a coupling trick (Wang et al., 2018) can prove the theorem. To prove this, decompose $M^t$ into the following:

$$M^{t+1} \quad M^t = E_t M^{t+1} \quad M^t + M^{t+1} \quad E_t M^{t+1}$$

where $E_t$ denotes the conditional expectation given the state of the Markov chain $X^t$. Thus, it is sucient to show that the following two conditions hold for all $t \quad NT$:

$$EkE_t M^{t+1} \quad M^t \quad F(M^t)=N k_F \quad C(T)N^{2=3}$$
$$EkM^{t+1} \quad E_t M^{t+1} k_F^2 \quad C(T)N^{2}$$

The rst condition ensures that the leading order of the average increment is captured by the ODEs in the Theorem 4.2. The second condition guarantees that the stochastic part can be ignored in the large $N$ limit. Further details regarding the derivation of these two conditions and the proof of the Theorem 4.2 can be found in Supplementary Materials C.

This theorem indicates that the macroscopic stochastic process $M^t$ converges to the deterministic process $M(t)$ at a convergence rate of $O(1=\sqrt{N})$. Furthermore, the generalization error $"_g$ can be expressed as a function of the macroscopic state, which allows us to investigate the dynamics from the ODEs in Eq. (14). In the following section, we present the results obtained by using Eq. (14).

## 5 RESULTS

We investigate the learning dynamics of VAE with a high-dimensional data limit using Eq. (14). Specically, we focus on the following representative cases: (i) the model-matched setting ($M = M = 1$) where the number of generative factors in the generative model, i.e., the number of columns in $W$, is equal to the latent space dimension; and (ii) the model-mismatched setting ($M = 2$ and $M = 1$), where the latent space dimension is larger than the number of the generative factors. In addition, numerical experiments are conducted to verify the consistency of our theory and to compare the results obtained by training the VAE. The source code is available at https://github.com/Yuma-Ichikawa/LearningDynamicsVAE .

### 5.1 Dynamics of Generalization Error

The dependence of learning dynamics is discussed by observing the time evolution of the generalization.

Figure 2: (Left) Generalization error, (middle) order parameters m and Q, and (right) order parameter $E_{12}$ as a function of time t for varying values with fixed parameters $= 0$; $_W = _V = _D = 0.01$, and $= = 1$ for both model-matched and model-mismatched cases. Each point on the plots represents the averages of five different numerical simulations with $N = 500$, and the error bars represent the standard deviations of the results.

The results are summarized as follows:

Peak and Long Plateau in $"_g$. Fig. 2 demonstrates the time dependence of the generalization error $"_g$ for various values along with the numerical experimental results with finite data dimension. For a smaller , the generalization error $"_g$ peaks in the early stages of learning, which tends to smoothly disappear as increases. Furthermore, for a larger , a long plateau appears in the range of t, and the length of this plateau increases as increases. When the value of exceeds 2, the decrease in the generalization error $"_g$ appears to completely disappear. We will discuss whether this decrease exists in the infinite time in the following section, based on the fixed points of the ODEs.

Overfitting with a Small . As shown in Fig. 2, the generalization error $"_g$ decreases followed by an increase near $t$ 1200 for a small , where the difference between order parameters $m_{11}(t)$ and $Q_{11}(t)$ is minimal. After passing this point, m(t) saturates to a certain value, and Q(t) continues to increase. This behavior indicates that while the recovery of the feature vector becomes saturated, the VAE starts to overfit the background noise. This suggests that the early stopping method, which stops the SGD update when the generalization error begins to increase, is effective for small .

Formation Process of Disentanglement. As discussed in Sec. 4, the off-diagonal terms of the order parameter E can be used to measure the disentanglement of the obtained representation. When these off-diagonal terms $E_{ij}$ are zero, the corresponding representations $z_i; z_j$ $q_{V;D}(zjx)$ are disentangled. Conversely, when $E_{ij} \neq 0$, the corresponding representations are entangled. The right panel of Fig. 2 shows the time dependence of the off-diagonal term, meaning the formation process of a disentangled representation. The representation is entangled, i.e., $E_{12} \neq 0$, in the early stages of learning, and a peak then appears at some time t. Subsequently, the representations gradually become disentangled as time progresses; that is, $E_{12} = 0$. The stability of the fixed points determines whether the disentanglement representations are obtained for any in the limit $t \to 1$ .

Figure 3: Asymptotic generalization error as a function of time t for the varying learning rate with fixed parameters $= 0$, $= = 1$ for both model-matched (solid line) and model-mismatched cases (dashed line).

### 5.2 Steady State of Generalization Error

Considering the analysis of the dynamics in the previous section, it remains unclear whether it is possible to escape from the plateau and reduce the generalization error $"_g$ for any given , or to obtain disentangled features in the long-time limit. In this section, we discuss these issues using a local stability analysis of the ODEs in Eq. (14). To further reduce the degrees of

freedom of the ODEs, we assume that the regularization parameter $= 0$ and a common learning rate $= {}_W = {}_V = {}_D$. In the subsequent analysis, if the Jacobian matrix of the ODEs has only negative eigenvalues, the xed point is called locally stable, and if the Jacobian matrix has both zero eigenvalues and negative eigenvalues, the xed point is called marginally stable.

Stability of Model-Matched Case. We investigate the local stability of the xed points of the ODEs in the model-matched case to clarify the conditions under which the VAE encounters a posterior collapse.

Theorem 5.1. For a small learning rate limit and $= 0$, the xed points of ODEs in the model-matched case with $M = M = 1$ have the following properties.

- For $<$ $+$, the following xed point is locally stable:

$$m = \sqrt{\quad + \quad}; \tag{15}$$

$$"_g = \frac{}{+ \quad}(2 - \sqrt{\quad + \quad}); \tag{16}$$

- For $=$ $+$, the xed point, $m = 0; "_g =$ , is marginally stable.
- For $>$ $+$, the xed point, $m = 0; "_g =$ , is locally stable.

Theorem 5.1 elucidates that once exceeds the threshold $=$ $+$, the generalization error can not escape from the plateau, despite $t$ increasing, which indicates that the posterior collapse cannot be avoided.

Furthermore, the limiting value of the generalization error $"_g$ coincides with that obtained from the analysis of the global optimum of Eq. (8) (Ichikawa and Hukushima, 2022); namely, following Remark holds.

Remark 5.2. The limiting value of the generalization error in Eq. (16) exactly equals the generalization error derived in the in nite data size limit by the analysis of the global optimum using the replica method (Ichikawa and Hukushima, 2023).

This result implies that it is possible to reach a global optimum solution using SGD with a small learning rate limit. To our best knowledge, the exact correspondence between the global optima obtained using the replica method and the steady state of the one-pass SGD and the reachability to the global optima has not yet been explored in the statistical physics community.

Stability of Model Mismatched-Case. We also clarify the condition under which the VAE encounters a posterior collapse in the model-mismatched case and obtains disentangled representations.

Theorem 5.3. For a small learning rate limit and $= 0$, the xed points of ODEs in the model mismatched case with $M = 2$ and $M = 1$ have the following properties.

- For $<$ , the following xed point is locally stable:

$$m = (\sqrt{\quad + \quad}; 0); (0; \sqrt{\quad + \quad}); \; E_{12} = 0$$

$$Q = \begin{matrix} + & 0 \\ 0 & \end{matrix}; \quad \begin{matrix} & 0 \\ 0 & + \end{matrix}$$

$$"_g = \frac{}{+ \quad}(2 - \sqrt{\quad + \quad}) + \quad ;$$

- For $=$ , the xed point is marginally stable:

$$m = (\sqrt{\quad}; 0); (0; \sqrt{\quad}); \; E_{12} = 0$$

$$Q = \begin{matrix} 0 & \\ 0 & 0 \end{matrix}; \quad \begin{matrix} 0 & 0 \\ 0 & \end{matrix}; \quad "_g = 0:$$

- For $<$ $<$ $+$, the xed point is locally stable:

$$m = (\sqrt{\quad + \quad}; 0); (0; \sqrt{\quad + \quad}); \; E_{12} = 0$$

$$Q = \begin{matrix} + & 0 \\ 0 & 0 \end{matrix}; \quad \begin{matrix} 0 & 0 \\ 0 & + \end{matrix}$$

$$"_g = \frac{}{+ \quad}(2 - \sqrt{\quad + \quad}):$$

- For $=$ $+$, the xed point, $m = Q = 0_{2\ 2}; E_{12} = 0; "_g =$ , is marginally stable.
- For $>$ $+$, the xed point, $m = Q = 0_{2\ 2}; E_{12} = 0; "_g =$ , is locally stable.

This theorem indicates that disentangled representations can be obtained in the small learning rate limit for any , that is, $8 ; E_{12} = 0$. The threshold for the posterior collapse is the same as that of the model-matched case. Thus, Theorem 5.1 and 5.3 suggest that can be a risky parameter since the posterior collapse is inevitable regardless of the training period. Furthermore, the extremum calculations of the generalization error in Theorems 5.1 and 5.3 demonstrate that the generalization error is minimized when $=$ , which means that the best generalization is achieved when is equal to the strength of the background noise . Note that the generalization error in the model-mismatched case at $=$ is marginally stable. However, the other xed points are unstable, indicating that the dynamics converges to the optimal xed point, but the convergence is signi cantly slow.

Another noteworthy observation is that Theorem 5.3 shows a new stable xed point; when $<$ , despite $m = (\sqrt{\quad + \quad}; 0); (0; \sqrt{\quad + \quad})$ having the same stable xed point as in the range $<$ $<$ $+$, a non-corresponding element of $Q$ becomes nite, i.e.,

when $m_{11} \neq 0$, $q_{22} \neq 0$, and when $m_{12} \neq 0$, $q_{11} \neq 0$. This suggests that when $<$ , the super uous latent variable for the data-generative factor over ts the background noise and a ects the generalization.

### 5.3 Learning Dynamics with KL Annealing

We now discuss the e ectiveness of monotonic KL annealing for the learning dynamics. A stability analysis of the xed point is conducted for the continuous tanh KL annealing, given by $(t) = \tanh( t)$, where denotes the annealing rate. This annealing satis es

$$\frac{d (t)}{dt} = (1 \quad ^2(t)); \quad (0) = 0: \qquad (17)$$

Compared to monotonic KL annealing expressed in Eq. 2, the trajectories of both tanh KL annealing and monotonic KL annealing are qualitatively similar. The learning curve with monotonic KL annealing is similar to that with tanh KL annealing; see Supplementary Materials E.1 for the detailed results. In particular, we focus on the representative model-matched case $M = M = 1$ with tanh KL annealing. The results are summarized as follows.

**Dynamical Properties of KL Annealing** The top panel of Fig. 4 demonstrates a comparison of the learning dynamics $"_g$ with constant $= 1$ and tanh KL annealing. The bottom panel of Fig. 4 shows the convergence time to the quasi-steady state $"_g + 0:001$ as a function of the annealing rate . This gure indicates the existence of an optimal annealing rate that maximizes the convergence speed to the quasi-steady state, and that an extremely slow KL annealing rate delays the convergence time. The annealing rate of the learning dynamics using tanh KL annealing, shown in Fig. 4 (Top), is selected as the optimal rate based on the bottom gure. Fig. 4 demonstrates that the convergence of the generalization error $"_g$ becomes faster with tanh KL annealing than without it. Subsequent discussions will focus on the threshold value of the annealing rate , which adversely a ects the learning dynamics.

**Steady State with KL Annealing.** Based on the stability analysis of the xed points, including the time-dependent $(t)$, the learning dynamics using tanh KL annealing exhibit the same stable xed points. Furthermore, unless excessively slow tanh KL annealing is used, the convergence speed to the steady state coincides with that without the tanh KL annealing. Formally, the following theorem holds:

**Theorem 5.4.** Even when tanh KL annealing is used, its steady state coincides with the steady state of the model-matched case at $= 1$ and $= 0$ without tanh KL annealing. Moreover, when $= 2$ and $=$ , tanh KL annealing leads to a slow convergence under the condition, $J_{max} = 2$ where

$$J_{max} = \begin{cases} \frac{1}{2}(\sqrt{5} \quad 3); & \frac{(1 \quad 2^{\sqrt{2}+\sqrt{5})}}{\sqrt{\frac{4}{4 \ (2 \quad 1) + 1}}} \quad \frac{(1+2^{\sqrt{2}+\sqrt{5})}}{4} \\ (2 \ + 1) + & \end{cases}$$

and the convergence using tanh KL annealing becomes the same as that without annealing when $> J_{max} = 2$.

The proof of this theorem can be found in Supplementary Materials D.3.

Figure 4: (Top) Time dependence of the generalization error and with both tanh KL annealing for and the constant $= 1$ under xed parameters $= 0$, $= =$ $1$, and $= 1$. (Bottom) Annealing-rate dependence of convergence time to the quasi-steady state deviating by 0:001, i.e., $"_g + 0:001$. The annealing rate of the learning dynamics with the tanh KL annealing in the top gure is used as the optimal value obtained from the bottom gure.

### 5.4 Related Work

**Deterministic Dynamical Descriptions of SGD.** Deterministic dynamical descriptions of SGD at a

high-dimensional input limit have been studied in the statistical physics community. This started with single- and two-layer neural networks with a few hidden units (Kinzel and Rujan, 1990; Kinouchi and Caticha, 1992; Copelli and Caticha, 1995; Biehl and Schwarze, 1995; Riegler and Biehl, 1995; Vicente et al., 1998), based on a heuristic derivation of ODEs describing typical learning dynamics. These results have recently been rigorously proven using the concentration phenomena in stochastic processes (Wang et al., 2018), based on which the analysis of the SGD for the two-layer neural networks was proven (Goldt et al., 2019; Veiga et al., 2022). For generative models, the SGD of generative adversarial networks has been investigated (Wang et al., 2019). However, to our best knowledge, this analysis has not been applied to the analysis of VAEs thus far.

Linear VAEs. The linear VAE is a simple model in which both the encoder and decoder are restricted to a ne transformations (Lucas et al., 2019). Although deriving analytical results for deep latent models is often intractable, a linear VAE can provide analytical results, facilitating a deeper understanding of VAEs. Furthermore, despite this simplicity, the theoretical results can su ciently explain the behavior of deeper and intricately structured VAEs (Lucas et al., 2019; Bae et al., 2022). In fact, results proven to be e ective for linear models have been applied to deeper models, leading to the new algorithms (Bae et al., 2022). In addition, several theoretical results have been obtained; Dai et al. (2018) demonstrated the connections between linear VAE, probabilistic PCA (Tipping and Bishop, 1999), and robust PCA (Candes et al., 2011; Chandrasekaran et al., 2011). Simultaneously, studies by Lucas et al. (2019) and Wang and Ziyin (2022) used linear VAEs to explore the origins of posterior collapse. However, these analyses did not address the learning dynamics indicated in our study.

## 6 CONCLUSION

This study rigorously proves that the SGD dynamics of a linear VAE converges to a deterministic process at a high-dimensional input limit. Our analysis reveals that the VAE initially learns entangled representations and then learns disentangled representations. Based on the stability analysis, we demonstrate that a posterior collapse occurs at a certain threshold of , and super uous latent spaces can over t the background noise of training data. We also demonstrate that appropriately adjusting KL annealing can accelerate the convergence of training. Although the linear VAE is a simple model, our results present a new perspective and some insights for the study of more realistic settings. This study has the following limitations. First, our analysis is based on a one-pass SGD, indicating that each data can be used only once; however, this is not the case in practical scenarios. Second, the data generation processes in the real world and VAEs are more complex than those in our data generative model and linear VAE. Thus, a more robust and minimal setup that can overcome these limitations will be developed in the future, along with a novel theoretical method.

Acknowledgements

This work was supported by JST Grant Number JPMJPF2221 and JPSJ Grant-in-Aid for Scienti c Research Number 23H01095. KH also acknowledges the partial support provided by Fujitsu Research Grand for this study. Additionally, YI was supported by the WINGS-FMSP program at the University of Tokyo.

References

Akkari, N., Casenave, F., Hachem, E., and Ryckelynck, D. (2022). A bayesian nonlinear reduced order modeling using variational autoencodersFluids, 7(10):334.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. Special lecture on IE, 2(1):1{18.

Bae, J., Zhang, M. R., Ruan, M., Wang, E., Hasegawa, S., Ba, J., and Grosse, R. (2022). Multi-rate vae: Train once, get the full rate-distortion curve. arXiv preprint arXiv:2212.03905.

Biehl, M. and Mietzner, A. (1993). Statistical mechanics of unsupervised learning. Europhysics Letters, 24(5):421.

Biehl, M. and Schwarze, H. (1995). Learning by on-line gradient descent. Journal of Physics A: Mathematical and general, 28(3):643.

Billingsley, P. (2013). Convergence of probability measures. John Wiley & Sons.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous spacearXiv preprint arXiv:1511.06349.

Candes, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? Journal of the ACM (JACM) , 58(3):1{37.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. SIAM Journal on Optimization , 21(2):572{596.

Child, R. (2020). Very deep vaes generalize autoregressive models and can outperform them on images. arXiv preprint arXiv:2011.10650.

Copelli, M. and Caticha, N. (1995). On-line learning in the committee machine. Journal of Physics A: Mathematical and General, 28(6):1615.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. (2018). Connections with robust pca and the role of emergent sparsity in variational autoencoder models. The Journal of Machine Learning Research, 19(1):1573–1614.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. (2019). Avoiding latent variable collapse with generative skip models. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 2397–2405. PMLR.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. arXiv preprint arXiv:1903.10145.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. (2019). Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. Advances in neural information processing systems, 32.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. (2022). The gaussian equivalence of generative models for learning with shallow neural networks. In Mathematical and Scientific Machine Learning, pages 426–471. PMLR.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In International conference on learning representations

Hoyle, D. C. and Rattray, M. (2004). Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. Physical Review E, 69(2):026124.

Hu, H. and Lu, Y. M. (2022). Universality laws for high-dimensional learning with random features. IEEE Transactions on Information Theory.

Ichikawa, Y. and Hukushima, K. (2022). Statistical-mechanical study of deep boltzmann machine given weight parameters after training by singular value decomposition. Journal of the Physical Society of Japan, 91(11):114001.

Ichikawa, Y. and Hukushima, K. (2023). Dataset size dependence of rate-distortion curve and threshold of posterior collapse in linear vae. arXiv preprint arXiv:2309.07663.

Ipsen, N. and Hansen, L. K. (2019). Phase transition in pca with missing data: Reduced signal-to-noise ratio, not sample size! In International Conference on Machine Learning, pages 2951–2960. PMLR.

Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H. (2016). Variational deep embedding: An unsupervised and generative approach to clustering. arXiv preprint arXiv:1611.05148.

Kim, Y., Wiseman, S., Miller, A., Sontag, D., and Rush, A. (2018). Semi-amortized variational autoencoders. In International Conference on Machine Learning, pages 2678–2687. PMLR.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kinouchi, O. and Caticha, N. (1992). Optimal generalization in perceptions. Journal of Physics A: mathematical and General, 25(23):6243.

Kinzel, W. and Rujan, P. (1990). Improving a network generalization ability by selecting examples. Europhysics Letters, 13(5):473.

Kushner, H. J. (2009). Stochastic Approximation and Recursive Algorithms and Applications (Stochastic Modelling and Applied Probability, 35). Springer New York.

Lesieur, T., Krzakala, F., and Zdeborová, L. (2015). Phase transitions in sparse pca. In 2015 IEEE International Symposium on Information Theory (ISIT), pages 1635–1639. IEEE.

Liao, Z. and Couillet, R. (2018). On the spectrum of random features maps of high dimensional data. In International Conference on Machine Learning, pages 3063–3071. PMLR.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. Advances in neural information processing systems, 30.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. (2019). Don't blame the elbo! a linear vae perspective on posterior collapse. Advances in Neural Information Processing Systems, 32.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. Matematicheskii Sbornik, 114(4):507–536.

Mei, S. and Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. Communications on Pure and Applied Mathematics, 75(4):667–766.

Norouzi, S., Fleet, D. J., and Norouzi, M. (2020). Exemplar vae: Linking generative models, nearest

neighbor retrieval, and data augmentation. Advances in Neural Information Processing Systems, 33:8753{8764.

Park, S., Adosoglou, G., and Pardalos, P. M. (2022). Interpreting rate-distortion of variational autoencoder and using model uncertainty for anomaly detection. Annals of Mathematics and Arti cial Intelligence, pages 1{18.

Potters, M. and Bouchaud, J.-P. (2020). A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists. Cambridge University Press.

Re netti, M. and Goldt, S. (2022). The dynamics of representation learning in shallow, non-linear autoencoders. InInternational Conference on Machine Learning, pages 18499{18519. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. InInternational conference on machine learning, pages 1278{1286. PMLR.

Riegler, P. and Biehl, M. (1995). On-line backpropagation in two-layered neural networks. Journal of Physics A: Mathematical and General, 28(20):L507.

Roberts, A., Engel, J., Ra el, C., Hawthorne, C., and Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In International conference on machine learning, pages 4364{4373. PMLR.

Sicks, R., Korn, R., and Schwaar, S. (2021). A generalised linear model framework for -variational autoencoders based on exponential dispersion families. The Journal of Machine Learning Research, 22(1):10539{10579.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology, 61(3):611{622.

Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. Advances in neural information processing systems, 33:19667{19679.

Veiga, R., Stephan, L., Loureiro, B., Krzakala, F., and Zdeborová, L. (2022). Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. arXiv preprint arXiv:2202.00293.

Vicente, R., Kinouchi, O., and Caticha, N. (1998). Statistical mechanics of online learning of drifting concepts: A variational approach. Machine Learning, 32:179{201.

Wang, C., Eldar, Y. C., and Lu, Y. M. (2018). Subspace estimation from incomplete observations: A high-dimensional analysis.IEEE Journal of Selected Topics in Signal Processing, 12(6):1240{1252.

Wang, C., Hu, H., and Lu, Y. (2019). A solvable high-dimensional model of gan. Advances in Neural Information Processing Systems, 32.

Wang, Z. and Ziyin, L. (2022). Posterior collapse of a linear latent variable model. Advances in Neural Information Processing Systems, 35:37537{37548.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. Biometrika, pages 32{52.

Yang, Z., Hu, Z., Salakhutdinov, R., and Berg-Kirkpatrick, T. (2017). Improved variational autoencoders for text modeling using dilated convolutions. In International conference on machine learning, pages 3881{3890. PMLR.

Zhao, T., Zhao, R., and Eskenazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. arXiv preprint arXiv:1703.10960.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with the speci cation of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] We provide comprehensive proof in the main text and Supplementary Materials.
   (c) Clear explanations of any assumptions. [Yes]
3. For all gures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear de nition of the speci c measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable] This study did not conduct performance evaluations of algorithms but only carried out numerical experiments to validate our theoretical results, thus we select Not Applicable.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identi able information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

## A OVERVIEW

This supplementary material provides additional results and detailed proofs in the main text.

## B COMPLETE FORM OF THE ORDINARY DIFFERENTIAL EQUATIONS IN THEOREM 4.2

In this section, we present the speci c function set of$F$ in Theorem 4.2 as follows:

$$\frac{dm_{ml}}{dt} = F_{m_{ml}}(M) = w\left(\sum_{n'=1}^{M} m_{n'l}\, h(d_m; d_{n'}; E_{mn'}) + m_{ml}(D_m + ) \quad h(m_l; d_m; d_{ml})\right); \tag{18}$$

$$\frac{dd_{ml}}{dt} = F_{d_{ml}}(M) = v\left(\sum_{n'=1}^{M} Q_{mn'} h(m_l; d_{n'}; d_{n'l}) + h(m_l; d_m; d_{ml}) \quad h(m_l; m_m; m_{ml}) + d_{ml}\right); \tag{19}$$

$$\begin{aligned}\frac{dQ_{mn}}{dt} = F_{Q_{mn}}(M) = {} & w\left(Q_{mn}(D_m + D_n + 2) \quad h(d_m; m_n; R_{nm}) \quad h(d_n; m_m; R_{mn})\right.\\ & \left. + \sum_{n'=1}^{M} Q_{mn'} h(d_n; d_{n'}; E_{nn'}) + \sum_{n'=1}^{M} Q_{nn'} h(d_m; d_{n'}; E_{mn'})\right) + \tfrac{2}{w} h(d_m; d_n; E_{mn});\end{aligned} \tag{20}$$

$$\begin{aligned}\frac{dE_{mn}}{dt} = F_{E_{mn}}(M) = {} & v\left(2h(d_m; d_n; E_{mn}) \quad h(m_m; d_n; R_{mn}) \quad h(m_n; d_m; R_{nm}) + 2E_{mn}\right.\\ & + \sum_{n'=1}^{M} Q_{nn'} h(d_m; d_{n'}; E_{mn'}) + \sum_{n'=1}^{M} Q_{mn'} h(d_n; d_{n'}; E_{nn'})\\ & + \tfrac{2}{v}\Big(\sum_{n';m'} Q_{mm'} Q_{nn'} h(d_{m'}; d_{n'}; E_{m'n'}) + \sum_{n'=1}^{M} Q_{mn'} h(d_{n'}; d_n; E_{nn'}) + \sum_{n'=1}^{M} Q_{nn'} h(d_{n'}; d_m; E_{mn'})\\ & + h(d_m; d_n; E_{mn}) \quad h(d_m; m_n; R_{nm}) \quad h(d_n; m_m; R_{mn})\Big)\\ & \left. + h(m_m; m_n; Q_{mn}) \sum_{n'=1}^{M} Q_{mn'} h(d_{n'}; m_n; R_{nn'}) \sum_{n'=1}^{M} Q_{nn'} h(d_{n'}; m_m; R_{mn'})\right);\end{aligned} \tag{21}$$

$$\begin{aligned}\frac{dR_{mn}}{dt} = F_{R_{mn}}(M) = {} & w\left(\sum_{n'=1}^{M} R_{n'n} h(d_{n'}; d_m; E_{mn'}) \quad h(d_m; d_n; E_{mn}) + (D_m + )R_{mn}\right.\\ & v\left(\sum_{n'=1}^{M} Q_{nn'} h(m_m; d_{n'}; R_{mn'}) + h(m_m; d_n; R_{mn}) \quad h(m_m; m_n; Q_{mn}) + R_{mn}\right)\\ & \left. + vw\left(\sum_{n'=1}^{M} Q_{nn'} h(d_m; d_{n'}; E_{mn'}) + h(d_m; d_n; E_{mn}) \quad h(d_m; m_n; R_{nm})\right);\right.\end{aligned} \tag{22}$$

$$\frac{dD_m}{dt} = F_{D_m}(M) = D \quad \frac{}{D_m} \quad (Q_{mm} + ); \tag{23}$$

where$m = (W)^{>} W/N$, and we use the shorthand expression given by

$$h(A; B; C) = \sum_{s=1}^{M} A_s B_s + C: \tag{24}$$

## C PROOF OF THEOREM 4.2

In this section, we provide a proof of Theorem 4.2 in main text from the following two Lemmas: (i) Convergence of the rst moment of the increment of the macroscopic stochastic process$M^t$, and (ii) Vanishing of the second

moment of the increment. Intuitively, these ensure that the leading order of the average increment is captured by the ODEs described in Theorem 4.2 and that the stochastic part of the increment of the macroscopic state $M^t$ vanishes as the input dimension increases.

The whole proof is divided into 4 parts. The first step is to prove the two conditions in the subsequent section. Then, it is demonstrated that these two conditions are sufficient to prove Theorem 4.2. Finally, technical Lemmas that are repeatedly used in the above proofs are summarized. The proof follows the standard scheme of the convergence of stochastic processes (Kushner, 2009; Billingsley, 2013; Wang et al., 2018).

### C.1 Convergence of First Moments of Increment to ODEs

We first review the training algorithm of SGD which characterizes a Markov process $X^t = (W^t; V^t; D^t)$. The specific update rule is given by

$$w_m^{t+1} = w_m^t - \frac{\eta}{N}\left(\sum_{n=1}^{M} w_n \left(\sqrt{\gamma}\sum_{s=1}^{M} c_s^t d_{ms}^t + \sqrt{\gamma}\,\xi_m^t\right)\left(\sqrt{\gamma}\sum_{s=1}^{M} c_s^t d_{ns}^t + \sqrt{\gamma}\,\xi_n^t\right) + (D_m^t + \Delta)w_m^t\right)$$
$$\left(\sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t w_s + \sqrt{\frac{\gamma}{N}}\,\eta^t\right)\left(\sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t d_{ms}^t + \sqrt{\gamma}\,\xi_m^t\right); \tag{25}$$

$$v_m^{t+1} = v_m^t - \frac{\eta}{N}\left(\sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t w_s + \sqrt{\frac{\gamma}{N}}\,\eta^t\right)\left(\sum_{n} Q_{mn}\left(\sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t d_{ns}^t + \sqrt{\gamma}\,\xi_n^t\right) + \sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t d_{ms}^t + \sqrt{\gamma}\,\xi_m^t\right)$$
$$\left(\sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t m_{ms} + \sqrt{\gamma}\,u_m^t\right) + v_m^t\Delta; \tag{26}$$

$$D_m^{t+1} = D_m^t - \frac{\eta}{2N}\left(Q_{mm}^t + \Delta\right) - \frac{\eta}{\sqrt{D_m^t}}; \tag{27}$$

where $w_m^t$, $w_m$ and $v_m^t$ represent $m$-th columns $W^t$, $W$ and $V^t$, respectively.

The following lemma holds for the macroscopic state $M^t$ characterized by the above updates.

Lemma C.1. Under the same assumptions as in Theorem 4.2, for all $t < NT$ the following inequality holds:

$$\mathbb{E}\left\|\mathbb{E}_t M^{t+1} - M^t - \frac{1}{N}F(M^t)\right\|_F \le \frac{C}{N^{3/2}}. \tag{28}$$

Proof. Recall that $M^t = (m^t; d^t; Q^t; E^t; R^t; V^t; D^t) \in \mathbb{R}^{M(2M+5M)}$ is composed of seven matrices. Note that defining $|A|_F = \sum_{i=1}^{N}\sum_{j=1}^{M}|a_{ij}|$ for matrix $A \in \mathbb{R}^{N\times M}$, the inequality $\|M^t\|_F \le |M^t|_F$ holds. Thus, the following inequality is sufficient to prove Eq. (28):

$$\mathbb{E}\left|\mathbb{E}_t M_{ij}^{t+1} - M_{ij}^t - \frac{1}{N}F_l(M_{ij})\right| \le \frac{C}{N^{3/2}}; \tag{29}$$

where $M_{ij}^t$ is $ij$ element of $M^t$. Subsequently, we show that the above inequality holds for each element of $M^t$.

For $m^t$, the following stronger result is obtained:

$$\mathbb{E}_t m_{ml}^{t+1} - m_{ml}^t - \frac{1}{N}F_{m_{ml}}(M^t) = 0; \tag{30}$$

where $F_{m_{ml}}(M)$ is defined in Eq. 18. This is directly proved by multiplying $(w_l)^\top/N$ from the left on both sides of Eq. 25, which yields

$$m_{ml}^{t+1} = m_{ml}^t - \frac{\eta}{N}\left(\sum_{n=1}^{M} m_{nl}^t\left(\sqrt{\gamma}\sum_{s=1}^{M} c_s^t d_{ms}^t + \sqrt{\gamma}\,\xi_m^t\right)\left(\sqrt{\gamma}\sum_{s=1}^{M} c_s^t d_{ns}^t + \sqrt{\gamma}\,\xi_n^t\right) + (D_m^t + \Delta)m_{ml}^t\right)$$
$$\left(\sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t m_{nl} + \sqrt{\gamma}\,u_l\right)\left(\sum_{s=1}^{M}\sqrt{\gamma}\,c_s^t d_{ms}^t + \sqrt{\gamma}\,\xi_m^t\right); \tag{31}$$

where $u_l = (w_l)^\top n^t/\sqrt{N}$. Note that $u_l$, $u_m^t$ and $\epsilon_m^t$ are Gaussian random variables. Then, taking the conditional expectation $E_t$ on both sides of Eq. (31), we reach Eq. 30.

Next, we can also get a stronger result for $d^t$ given by

$$E_t d_{ml}^{t+1} - d_{ml}^t - \frac{1}{N} F_{d_{ml}}(M^t) = 0; \tag{32}$$

where $F_{d_{ml}}$ is defined in Eq. (19). This is also proved by multiplying $(w_l)^\top/N$ from the left on both side of Eq. (26), which yields

$$\begin{aligned} d_{ml}^{t+1} = d_{ml}^t - \frac{V}{N}\Bigg( & \sum_{s=1}^M \sqrt{c_s^t} m_{sl} + \sqrt{u_l} \Bigg( \sum_n Q_{mn} \sum_{s=1}^M \sqrt{c_s^t} d_{ns}^t + \sqrt{\epsilon_n^t} + \sum_s \sqrt{c_s^t} d_{ms}^t + \sqrt{\epsilon_m^t} \Bigg) \\ & \sum_{s=1}^M \sqrt{c_s^t} m_{ms} + \sqrt{u_m^t} \Bigg) + d_{ml}^t. \end{aligned} \tag{33}$$

One can also take the conditional expectation $E_t$ on both sides of Eq. (33) since $u_l^t$, $u_m^t$, and $\epsilon_m^t$ are Gaussian random variables, leading to Eq. (32).

Next, for $Q^t$, the following inequality holds:

$$E_t Q_{mn} - Q_{mn}^t - \frac{1}{N} f_{Q_{mn}}(M^t) - \frac{C(T)}{N^{\frac{3}{2}}}; \tag{34}$$

where $F_{Q_{mn}}$ is defined in Eq. (20). This is proved by evaluating $Q_{mn}^{t+1} = (w_m^{t+1})^\top w_n^{t+1}/N$ as follows:

$$\begin{aligned} Q_{mn}^{t+1} &= Q_{mn}^t - \frac{V}{N}\left( r_{w_m^t} r(X^t) \right)^\top w_n^t + (w_m^t)^\top r_{w_m^t} r(X^t) + \frac{V^2}{N} \left( r_{w_m^t} r(X^t) \right)^\top \left( r_{w_n^t} r(X^t) \right) \\ &= Q_{mn}^t - \frac{W}{N}\Bigg( \sum_{n'=1}^M Q_{mn'}^t \sqrt{c_s^t} d_{ms}^t + \sqrt{\epsilon_m^t} \sqrt{c_s^t} d_{n's}^t + \sqrt{\epsilon_{n'}^t} \Bigg) \\ &\quad + \sum_{n'=1}^M Q_{nn'}^t \sqrt{c_s^t} d_{ns}^t + \sqrt{\epsilon_n^t} \sqrt{c_s^t} d_{n's}^t + \sqrt{\epsilon_{n'}^t} \Bigg) \\ &\quad \sum_{s=1}^M \sqrt{c_s^t} m_{ns}^t + \sqrt{u_n^t} \Bigg) \Bigg( \sum_{s=1}^M \sqrt{c_s^t} d_{ms}^t + \sqrt{\epsilon_m^t} \Bigg) \Bigg( \sum_{s=1}^M \sqrt{c_s^t} m_{ms}^t + \sqrt{u_m^t} \Bigg) \Bigg( \sum_{s=1}^M \sqrt{c_s^t} d_{ns}^t + \sqrt{\epsilon_n^t} \Bigg) \\ &\quad + (D_n^t + D_m^t + 2) Q_{mn}^t + \frac{V^2}{N} \frac{k n^t k^2}{N} \sum_{s=1}^M \sqrt{c_s^t} d_{ms}^t + \sqrt{\epsilon_m^t} \Bigg) \Bigg( \sum_{s=1}^M \sqrt{c_s^t} d_{ns}^t + \sqrt{\epsilon_n^t} \Bigg) + \frac{V^2}{N^2}(M^t). \end{aligned}$$

Also, taking the conditional expectation and using $E_t|(M^t)| - \sqrt{N} C(T)$, which is proven based on Lemma C.4, we can derive Eq. 34. Then, the following inequality holds for $E^t$:

$$E_t E_{mn}^{t+1} - E_{mn}^t - \frac{1}{N} F_{E_{mn}}(M^t) - \frac{C(T)}{N^{\frac{3}{2}}}; \tag{35}$$

where $F_{E^t_{mn}}$ is defined in Eq. (20). This is proved by evaluating $E^{t+1}_{mn} = (v^{t+1}_m)^\top v^{t+1}_n / N$ as follows:

$$
\begin{aligned}
E^{t+1}_{mn} &= E^t_{mn} - \frac{\eta}{N}\Big(\big(\nabla_{v^t_m} r(X^t)\big)^\top v^t_n + (v^t_m)^\top \nabla_{v^t_n} r(X^t)\Big) + \frac{\eta^2}{N}\big(\nabla_{v^t_m} r(X^t)\big)^\top \big(\nabla_{v^t_n} r(X^t)\big);\\
&= E^t_{mn} - \frac{\eta}{N}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{ns} + \sqrt{\bar{\alpha}^t_n}\Bigg)\Bigg(\sum_{n'} Q_{mn'}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{n's} + \sqrt{\bar{\alpha}^t_{n'}}\Bigg) + \sum_{s} \sqrt{\bar{c}^t_s}\, d^t_{ms} + \sqrt{\bar{\alpha}^t_m}\Bigg)\\
&\quad - \Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, m^t_{ms} + \sqrt{\bar{u}^t_m}\Bigg)\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{ms} + \sqrt{\bar{\alpha}^t_m}\Bigg)\Bigg(\sum_{n'} Q_{nn'}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{n's} + \sqrt{\bar{\alpha}^t_{n'}}\Bigg)\Bigg)\\
&\quad + \Bigg(\sum_{s} \sqrt{\bar{c}^t_s}\, d^t_{ns} + \sqrt{\bar{\alpha}^t_n}\Bigg)\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, m^t_{ms} + \sqrt{\bar{u}^t_n}\Bigg) + 2\, E^t_{mn}\\
&\quad + \frac{\eta^2 k n^t k^2}{N^2}\Bigg(\sum_{n'} Q_{mn'}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{n's} + \sqrt{\bar{\alpha}^t_{n'}}\Bigg) + \sum_{s} \sqrt{\bar{c}^t_s}\, d^t_{ms} + \sqrt{\bar{\alpha}^t_m}\Bigg)\\
&\quad \Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, m^t_{ms} + \sqrt{\bar{u}^t_m}\Bigg)\Bigg(\sum_{n'} Q_{nn'}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{n's} + \sqrt{\bar{\alpha}^t_{n'}}\Bigg) + \sum_{s} \sqrt{\bar{c}^t_s}\, d^t_{ns} + \sqrt{\bar{\alpha}^t_n}\Bigg)\\
&\quad \Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, m^t_{ms} + \sqrt{\bar{u}^t_n}\Bigg) + \frac{\eta^2}{N^2}\,\zeta(M^t).
\end{aligned}
$$

Here, one can also take the conditional expectation and use $|\mathbb{E}_t\, \zeta(M^t)| \le \sqrt{N}\, C(T)$ that is proven based on Lemma C.4 and then reach Eq. 35.

Next, for $R_{mn}$, the following holds:

$$
\mathbb{E}_t R^t_{mn} - R^t_{mn} - \frac{1}{N} F_{R_{mn}}(M^t) - \frac{C(T)}{N^{\frac{3}{2}}}; \tag{36}
$$

where $F_{R_{mn}}$ is defined in Eq. (22). This is proved by evaluating $R^{t+1}_{mn} = (w^{t+1}_m)^\top v^{t+1}_n / N$ as follows:

$$
\begin{aligned}
R^{t+1}_{mn} &= R^t_{mn} - \frac{\eta}{N}\big(\nabla_{w^t_m} r(X^t)\big)^\top v^t_n + \frac{\eta}{N}(w^t_m)^\top \nabla_{v^t_n} r(X^t) + \frac{\eta^2}{N}\big(\nabla_{w^t_m} r(X^t)\big)^\top \big(\nabla_{v^t_n} r(X^t)\big);\\
&= R^t_{mn} - \frac{\eta}{N}\sum_{n'=1}^{N} R_{n'n}\Bigg(\sqrt{\bar{c}^t_m}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{ms} + \sqrt{\bar{\alpha}^t_m}\Bigg)\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{n's} + \sqrt{\bar{\alpha}^t_{n'}}\Bigg) + (D^t_m + \lambda) R^t_{mn}\\
&\quad - \Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{ns} + \sqrt{\bar{\alpha}^t_n}\Bigg)\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{ms} + \sqrt{\bar{\alpha}^t_m}\Bigg)\\
&\quad - \frac{\eta}{N}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, m^t_{ms} + \sqrt{\bar{u}_m}\Bigg)\Bigg(\sum_{n'} Q_{nn'}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{n's} + \sqrt{\bar{\alpha}^t_{n'}}\Bigg) + \sum_{s} \sqrt{\bar{c}^t_s}\, d^t_{ns} + \sqrt{\bar{\alpha}^t_n}\Bigg)\\
&\quad \Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, m^t_{ms} + \sqrt{\bar{u}^t_n}\Bigg) + R^t_{mn}\\
&\quad + \frac{\eta^2 k n^t k^2}{N^2}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{ms} + \sqrt{\bar{\alpha}^t_m}\Bigg)\Bigg(\sum_{n'} Q^t_{nn'}\Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, d^t_{n's} + \sqrt{\bar{\alpha}^t_{n'}}\Bigg) + \sum_{s} \sqrt{\bar{c}^t_s}\, d^t_{ns} + \sqrt{\bar{\alpha}^t_n}\Bigg)\\
&\quad \Bigg(\sum_{s=1}^{M} \sqrt{\bar{c}^t_s}\, m^t_{ms} + \sqrt{\bar{u}^t_n}\Bigg) + \frac{\eta^2}{N^2}\,\zeta(M^t);
\end{aligned}
$$

Then, one can take the conditional expectation and use $|\mathbb{E}_t\, \zeta(M^t)| \le \sqrt{N}\, C(T)$ that is proven based on Lemma C.4 and then reach Eq. 36.

Lastly, the following stronger result holds for $D^t_m$ :

$$
\mathbb{E}_t D^{t+1}_m - D^t_m - \frac{1}{N} F_{D_m}(M^t) = 0\,; \tag{37}
$$

where $F_{D\,mn}$ is de ned in Eq. (23). one can directly obtain as following

$$D_m^{t+1} = D_m^t \quad {}_D \quad (Q_{mm}^t + \quad) \quad \frac{}{D_m^t} \quad : \tag{38}$$

Then, one takes the conditional expectation and then reaches Eq. 37. Combining Eq. (30)-(37), Eq. (28) is proven, which concludes the whole proof. □

### C.2 Convergence of Second Moments of Increment

We now proceed to bound the second-order moments of the increments.

Lemma C.2. Under the same assumption as in Theorem 4.2, for all $l < NT$ the following inequality holds:

$$\mathrm{Ek}M^{t+1} \quad \mathrm{E}_t M^{t+1} k_F^2 \quad \frac{C(T)}{N^2}: \tag{39}$$

Proof. Note that

$$\begin{aligned}
\mathrm{Ek}M^{t+1} \quad \mathrm{E}_t M^{t+1} k_F^2 &= \mathrm{Ek}M^{t+1} \quad M^{\ t} \quad \mathrm{E}_t(M^{t+1} \quad M^{\ t})k_F^2; \\
&\mathrm{Ek}M^{t+1} \quad M^{\ t}k_F^2 + \mathrm{Ek}\mathrm{E}_t M^{t+1} \quad M^{\ t}k_F^2; \\
&\mathrm{Ek}M^{t+1} \quad M^{\ t}k_F^2 + \mathrm{E}\ \frac{1}{N}F(M^{\ t}) + \frac{C(T)}{N^{\frac{3}{2}}}\ ^2; \\
&\mathrm{Ek}M^{t+1} \quad M^{\ t}k_F^2 + \frac{C(T)}{N^2}:
\end{aligned}$$

Here the third line is due to Lemma C.1. Thus, it is su cient to prove that

$$\mathrm{Ek}M^{t+1} \quad M^{\ t}k_F^2 \quad \frac{C(T)}{N^2}:$$

In the following, the second moment of each element in $M^{t+1} \quad M^{\ t}$ will be bounded.

For $m^t$, the following inequality holds:

$$\begin{aligned}
\mathrm{E}(m_{ml}^{t+1} \quad m_{ml}^t)^2 = \frac{W}{N^2}\mathrm{E}\Bigg[ & \Bigg( \sum_{n=1}^{M} m_{nl}^t \quad p_- \sum_{s=1}^{M} c_s^t d_{ms}^t + p_- {}_m^t \Bigg) \Bigg( p_- \sum_{s=1}^{M} c_s^t d_{ns}^t + p_- {}_n^t \Bigg) + (D_m^t + \quad) m_{ml}^t; \\
& \sum_{s=1}^{M} p_- c_s^t m_{nl} + p_- u_l \Bigg) \Bigg( \sum_{s=1}^{M} p_- c_s^t d_{ms}^t + p_- {}_m^t \Bigg)^2 \Bigg] \\
& \frac{C}{N^2}\mathrm{E}\Bigg[ \sum_{n;h} m_{nl}^t m_{hl}^t h(d_m^t; d_n^t; E_{mn}^t) h(d_m^t; d_h^t; E_{mh}^t) + (D_m^t + \quad)^2 (m_{ml}^t)^2 + h^2(m_l; d_m; d_{ml}) \\
& + 2(D_m^t + \quad) m_{ml}^t \Bigg( \sum_n m_{nl}^t h(d_m; d_n; E_{mn}) \quad h(m_l; d_m^t; d_{ml}^t) \Bigg) \\
& 2h(d_m^t; d_n^t; E_{mn}) h(m_l; d_m^t; d_{ml}^t) \Bigg] \quad \frac{C(T)}{N^2}
\end{aligned} \tag{40}$$

Here, the last line is due to Lemma C.4.

Next, for $d^t$, one can get the following inequality in a similar way:

$$E(d_{ml}^{t+1} - d_{ml}^t)^2 = \frac{2}{N^2}E\Bigg[\Bigg(\sum_{s=1}^{M}\sqrt{c}_s^t m_{sl} + \sqrt{u}_l^t\Bigg)\Bigg(\sum_n Q_{mn}\Bigg(\sum_{s=1}^{M}\sqrt{c}_s^t d_{ns}^t + \sqrt{}_n^t\Bigg) + \sum_s \sqrt{c}_s^t d_{ms}^t + \sqrt{}_m^t\Bigg) - \sum_{s=1}^{M}\sqrt{c}_s^t m_{ms} + \sqrt{u}_m^t\Bigg) + d_{ml}^t\Bigg]^2$$
$$\frac{C}{N^2}E\Bigg[h(m_l;m_l;1)\sum_{n,h} Q_{mn}Q_{mh}h(d_n^t;d_h^t;E_{nh}^t) + {}^2h(d_m^t;d_m^t;E_{mm}^t) + h(m_m^t;m_m^t;Q_{mm}^t)$$
$$+ 2\sum_n Q_{mn}h(d_n;d_m;Q_{nm} \quad h(d_m;m_m;R_{mm})\Bigg) - \sum_n Q_{mn}^t h(d_n;m_m;R_{mn})\Bigg)$$
$$+ 2\,d_{ml}^t\Bigg(\sum_n Q_{mn}^t h(m_l;d_n^t;d_{ml}^t) + h(m_l;d_m^t;d_{ml}^t) \quad h(m_l;m_m^t;m_{ml}^t)\Bigg) + {}^2(d_{ml}^t)^2\Bigg] \quad \frac{C(T)}{N^2}. \tag{41}$$

Here, the last line is also due to Lemma C.4. Similarly, one can also prove that

$$E(Q_{mn}^{t+1} \quad Q_{mn}^t)^2 \quad \frac{C(T)}{N^2}; \tag{42}$$
$$E(E_{mn}^{t+1} \quad E_{mn}^t)^2 \quad \frac{C(T)}{N^2}; \tag{43}$$
$$E(R_{mn}^{t+1} \quad R_{mn}^t)^2 \quad \frac{C(T)}{N^2}; \tag{44}$$
$$E(D_m^{t+1} \quad D_m^t)^2 \quad \frac{C(T)}{N^2}. \tag{45}$$

Combining Eq. (40)-(45), Eq. (39) is proven, which concludes the whole proof. □

### C.3 Proof of Theorem 4.2

In this section, we nish the remaining proof of Theorem 4.2 from Lemma C.1 and C.2 by using the coupling trick.

Proof. We rst de ne a stochastic process $B^t$ that is coupled with the process $M^t$ as

$$B^{t+1} = B^t + \frac{1}{N}F(B^t) + M^{t+1} \quad E_t M^{t+1} \tag{46}$$

with the deterministic initial condition $B^0 = M^0$. For this stochastic process $B^t$, the following inequality holds for all $t \quad NT$:

$$EkB^t \quad M^t k_F \quad \frac{C(T)}{N^{1=2}}. \tag{47}$$

This inequality is proved as follows.

$$EkB^{t+1} \quad M^{t+1}k_F \quad EkB^t \quad M^t k_F + \frac{1}{N}EkF(B^t) \quad F(M^t)k_F + EkE_t M^{t+1} \quad M^t \quad \frac{1}{N}F(M^t)k_F :$$

From Lemma C.1 and Lemma C.6 in subsequent Sec. C.4, one can get

$$EkB^{t+1} \quad M^{t+1}k_F \quad EkB^t \quad M^t k_F + LkB^t \quad M^t k_F + C(T)N^{\frac{3}{2}}$$
$$(1 + LN^{1})kB^t \quad M^t k + CN^{\frac{3}{2}}:$$

Applying this bound iteratively, for all $t \quad NT$, one can expand as follows:

$$EkB^t \quad M^t k_F \quad e^{LT} \quad EkB^0 \quad M^0 k_F + \frac{C}{L}N^{\frac{1}{2}} \quad \frac{C(T)}{N^{\frac{1}{2}}}. \tag{48}$$

For the last inequality, we use the assumption (A.3) in the main text.

Next, we de ne a deterministic process $S^t$ as follows:

$$S^{t+1} = S^t + \frac{1}{N}F(S^t) \tag{49}$$

with the deterministic initial condition $S^0 = M^0$. Similarly, the following inequality holds for all $t \quad NT$:

$$EkB^t \quad S^t k^2 \quad \frac{C(T)}{N} \tag{50}$$

To prove this inequality, one can express as

$$EkB^{t+1} \quad S^{t+1} k_F = EkB^t \quad S^t k^2 + \frac{1}{N^2} EkF(B^t) \quad F(S^t) k^2 + \frac{2}{N} E(F(B^t) \quad F(S^t))^{>} (B^t \quad S^t) + EkM^{t+1} \quad E_t M^t k_F^2 :$$

Here, one uses the identity given by

$$E_t(M^{t+1} \quad E_t M^t)^{>} (B^t \quad S^t) = E_t(M^{t+1} \quad E_t M^t)^{>} (F(B^t) \quad F(S^t)) = 0 :$$

Then, from Lemma C.2 and Lemma C.6 in Sec. C.4 below, one can get following inequality:

$$EkB^{t+1} \quad S^{t+1} k_F^2 \quad 1 + \frac{CL}{N} \quad EkB^t \quad S^t k_F^2 + \frac{C(T)}{N^2} :$$

Applying this bound iteratively, for all $t \quad NT$, Eq. (50) is proven as follows:

$$EkB^t \quad S^t k_F^2 \quad \frac{C(T)}{N} : \tag{51}$$

Note that $S^t$ is a standard rst-order nite di erence approximation of the ODEs with the step size $1=N$. The standard Euler argument implies that

$$kS^t \quad M \quad (t)k \quad \frac{C}{N} : \tag{52}$$

Finally, combining Eq. (47), (50) and (52), Theorem 4.2 is proven as follows:

$$\begin{aligned} EkM^t \quad M \quad (t)k = & \; EkM^t \quad B^t + B^t \quad S^t + S^t \quad M \quad (t)k \\ & EkM^t \quad B^t k + EkB^t \quad S^t k + EkS^t \quad M \quad (t)k \\ & EkM^t \quad B^t k + (EkB^t \quad S^t k^2)^{\frac{1}{2}} + EkS^t \quad M \quad (t)k \\ & \frac{C(T)}{N^{\frac{1}{2}}} : \end{aligned}$$

□

### C.4 Extra Proofs

In this section, we complete the extra technical lemmas related to the proofs in the previous section.

#### C.4.1 Bound for Micoroscopic State

Lemma C.3. Under the same assumption as in Theorem 4.2, for all $t \quad NT$ and $i = 1; : : : N$, the following inequality holds:

$$E\left[ \sum_{n=1}^{M} (W_{in}^t)^4 + \sum_{n=1}^{M} (V_{in}^t)^4 + \sum_{n=1}^{M} (D_n^t)^4 \right] \quad C(T): \tag{53}$$

Proof. We rst prove $E(W_{il}^t)^4 \quad C(T)$. Note that one can expand as follows:

$$\begin{aligned} E(W_{il}^{t+1})^4 \quad E(W_{il}^t)^4 = & \; 4E[(W_{in}^t)^3 E_t(W_{in}^{t+1} \quad W_{in}^t)] + 6E[(W_{in}^t)^2 E_t(W_{in}^{t+1} \quad W_{in}^t)^2] \\ & + 4E[W_{in}^t E_t(W_{in}^{t+1} \quad W_{in}^t)^3] + E[E_t(W_{in}^{t+1} \quad W_{in}^t)^4] \end{aligned} \tag{54}$$

From Eq. 25 and the triangle inequality, the following inequality holds for $\alpha = 1, 2, 3$ and 4:

$$E_t(W_{in}^{t+1} - W_{in}^t)^\alpha \le \frac{C}{N}\Bigg|\sum_{n'=1}^{M} W_{in'}^t \sum_s d_{ms}^t d_{ns}^t + E_{mn}^t\Bigg|^\alpha + |(D_m^t + \eta)W_{in}|^\alpha + \Bigg|\sum_s W_{is} d_{ns}^t\Bigg|^\alpha + |\eta V_{in}^t|^\alpha. \quad (55)$$

Substituting Eq. (55) into Eq. (54), we have

$$E(W_{in}^{t+1})^4 \le E(W_{in}^t)^4 + \frac{C}{N}E\Bigg[(W_{in}^t)^3\Bigg|\sum_{n'=1}^{M} W_{in'}^t \sum_s d_{ms}^t d_{ns}^t + E_{mn}^t\Bigg| + |(D_m^t + \eta)(W_{in}^t)^4| + (W_{in}^t)^3\Bigg|\sum_s W_{is} d_{ns}^t\Bigg| + |(W_{in}^t)^3 V_{in}^t|\Bigg] + O(N^{-2}). \quad (56)$$

For $V_{il}^t$, one can obtain the following:

$$E(V_{il}^{t+1})^4 - E(V_{il}^t)^4 = 4E[(V_{in}^t)^3 E_t(V_{in}^{t+1} - V_{in}^t)] + 6E[(V_{in}^t)^2 E_t(V_{in}^{t+1} - V_{in}^t)^2] + 4E[V_{in}^t E_t(V_{in}^{t+1} - V_{in}^t)^3] + E[E_t(V_{in}^{t+1} - V_{in}^t)^4]. \quad (57)$$

From Eq. 26 and the triangle inequality, the following inequality holds for $\alpha = 1, 2, 3$ and 4:

$$E_t(V_{in}^{t+1} - V_{in}^t)^\alpha \le \frac{C}{N}\Bigg|\sum_s W_{is} \sum_{n'} Q_{nn'}^t \sum_s d_{n's}^t + \sum_s d_{ns}^t \sum_s m_{ns}^t + \sum_{n'} Q_{nn'}^t V_{in'}^t\Bigg|^\alpha + |\eta W_{in}^t|^\alpha + |\eta V_{in}^t|^\alpha. \quad (58)$$

Substituting Eq. (58) into Eq. (57), one can obtain the following:

$$E(V_{il}^{t+1})^4 - E(V_{il}^t)^4 \le \frac{C}{N}\Bigg[(V_{in}^t)^3\Bigg|\sum_s W_{is} \sum_{n'} Q_{nn'}^t \sum_s d_{n's}^t + \sum_s d_{ns}^t \sum_s m_{ns}^t + (V_{in}^t)^3 \sum_{n'} Q_{nn'} V_{in'}^t\Bigg| + |\eta(V_{in}^t)^3 W_{in}^t| + |\eta(V_{in}^t)^4|\Bigg] + O(N^{-2}). \quad (59)$$

Similarly, one can also get the following inequality:

$$E(D_n^{t+1})^4 - E(D_n^t)^4 \le \frac{C}{N}E\left[(D_n^t)^3 Q_{nn}^t + |\eta(D_n^t)^2|\right]. \quad (60)$$

Combining Eq. (56), Eq. (59) and Eq. (60), the following inequality holds:

$$\begin{aligned} E\left[(W_{in}^{t+1})^4 + (V_{in}^{t+1})^4 + (D_n^{t+1})^4\right] \le\ & E\left[(W_{in}^t)^4 + (V_{in}^t)^4 + (D_n^t)^4\right] \\ & \frac{C}{N}\Bigg[(W_{in}^t)^3\Bigg|\sum_{n'=1}^{M} W_{in'}^t \sum_s d_{ms}^t d_{ns}^t + E_{mn}^t\Bigg| + |\eta D_m^t (W_{in}^t)^4| \\ & + (W_{in}^t)^3\Bigg|\sum_s W_{is} d_{ns}^t\Bigg| + |(W_{in}^t)^3 V_{in}^t| + (V_{in}^t)^3\Bigg|\sum_s W_{is} \sum_{n'} Q_{nn'}^t \sum_s d_{n's}^t + \sum_s d_{ns}^t \sum_s m_{ns}^t\Bigg| \\ & + (V_{in}^t)^3 \sum_{n'} Q_{nn'}^t V_{in'}^t| + |\eta(V_{in}^t)^3 W_{in}^t| + |\eta(V_{in}^t)^4| + |\eta(D_n^t)^3 Q_{nn}^t| + |\eta(D_n^t)^2|\Bigg]. \end{aligned} \quad (61)$$

Using the above inequality iteratively, one can get

$$E\,(W_{in}^t)^4 + (V_{in}^t)^4 + (D_n^t)^4 \quad C(T)\quad (W_{in}^0)^3 \sum_{n'=1}^{M} W_{in'}^0 \sum_s d_{ms}^0 d_{ns}^0 + E_{mn}^0 \quad + jD_m^0 (W_{in}^0)^4 j$$
$$+ \quad (W_{in}^0)^3 \sum_s W_{is} d_{ns}^0 \; + j(W_{in}^0)^3 V_{in}^0 j + \quad (V_{in}^0)^3 \sum_s W_{is} \sum_{n'} Q_{nn'}^0 \sum_s d_{n's}^0 + \sum_s d_{ns}^0 \sum_s m_{ns}^0 \Bigg)$$
$$+ \quad (V_{in}^0)^3 \sum_{n'} Q_{nn'}^0 V_{in'}^0 \; + j(V_{in}^0)^3 W_{in}^0 j \; + j(V_{in}^0)^4 j + j(D_n^0)^3 Q_{nn}^0 j + j(D_n^0)^2 j \Bigg] :$$

We now reach Eq. (53) since initial microscopic states are bounded, i.e.$E[\sum_{n=1}^{M} f(W_{in}^0)^4 + (V_{in}^0)^4 + (D_n^0)^4 g + \sum_{n=1}^{M} W_{in}]$ C, because of the assumption (A.4). □

### C.4.2 Bound for Macroscopic State

Lemma C.4. Under the same assumption as in Theorem 4.2, for all NT, the following inequality holds:

$$EkQ^t k_F^2 \quad C(T);\quad EkE^t k_F^2 \quad C(T);\quad EkR^t k_F^2 \quad C(T);\quad Ekm^t k_F^2 \quad C(T);\quad Ekd^t k_F^2 \quad C(T): \tag{62}$$

Proof. It is a direct consequence of Lemma C.3. For$Q_{nn}^t$, using Hölder's inequality, one can get

$$E(Q_{nn}^t)^2 = \frac{1}{N^2} E\left[\sum_{i=1}^{n} W_{in}^t W_{in}^t\right]^2$$
$$\frac{1}{N} E \sum_{i=1}^{N} (W_{in}^t)^4 \quad C(T):$$

The last line is based on Lemma C.3.

For $Q_{mn}^t$ ; m 6 n, using Cauchy-Schwartz inequality and Hölder's inequality, one can get

$$EQ_{mn}^t = \frac{1}{N^2} E\left[\sum_{i=1}^{N} W_{im}^t W_{in}^t\right]^2$$
$$\frac{1}{N^2} E\left[\sum_{i=1}^{N} (W_{im}^t)^2\right]^2 \left[\sum_{i=1}^{N} (W_{in}^t)^2\right]^2$$
$$\frac{1}{N^2} \sqrt{E\left[\sum_{i=1}^{N} (W_{im}^t)^2\right]^2 \left[\sum_{i=1}^{N} (W_{in}^t)^2\right]^2}$$
$$\frac{1}{N} \sqrt{E\left[\sum_{i=1}^{N} (W_{im}^t)^4 \sum_{i=1}^{N} (W_{in}^t)^4\right]} \quad C(T);$$

where in reaching the last line, we use Lemma C.3. Then, we get$EkQ^t k_F \quad C(T)$. The rest bound of $EkE^t k_F^2$, $EkR^t k_F^2$, $Ekm^t k_F^2$ and $Ekd^t k_F^2$ can also be directly veri ed using the Cauchy-Schwartz inequality and Hölder's inequality and Lemma C.3. □

### C.4.3 Lipschitzness of ODEs

Lemma C.5. Under the same assumption as in Theorem 4.2, for all NT, $D^t$ 6 $0_M \quad _M$ holds.

Proof. Consider the ODE in Eq. 23:

$$\frac{dD_m(t)}{dt} = \quad _D \quad \frac{}{D_m(t)} \quad (Q_{mm}(t) + \quad) \quad ;$$

where $_D$; 0 and $8t$ $NT; Q_{mm}(t)$ 0 by de nition. We show the behavior of the solution $D_m(t)$ based on its initial condition. For $D_m(0) > 0$, the term $_D( =D_m(t)$ $(Q_{mm}(t) + ))$ is positive as $D_m(t)$ approaches zero and negative as $D_m(t)$ grows to positive in nity. Consequently, if $D_m(t)$ attempts to approach zero, $dD_m(t)=dt$ becomes positive, indicating that $D_m(t)$ increase, and thus does not cross zero. Similarly, if $D_m(t)$ becomes very large, $dD_m(t)=dt$ becomes negative, causing $D_m(t)$ to decrease but remain positive. Therefore, given the initial condition $D_m(0) > 0$, $D_m(t)$ remains positive for all $t > 0$. Similarly, we can show that, given the initial condition $D_m(0) < 0$, $D_m(t)$ remains negative for all $t > 0$. □

Lemma C.6. Under the same assumption as Theorem 4.2, $F(M)$ is a Lipschitz function.

Proof. It su ces to verify each component of gradient $r F(M)$ is bounded. Eq. (18)-(22) are linear functions with respect to $M$ and then following inequality holds for $8M$ :

$$kr_M F_{m_{ml}}(M)k \quad L_{m_{ml}}(M); \quad kr_M F_{d_{ml}}(M)k \quad L_{d_{ml}}(M); \quad kr_M F_{Q_{mn}}(M)k \quad L_{Q_{mn}}(M);$$
$$kr_M F_{E_{mn}}(M)k \quad L_{E_{mn}}(M); \quad kr_M F_{R_{mn}}(M)k \quad L_{R_{mn}}(M);$$

where $L_{m_{ml}}(M)$, $L_{d_{ml}}(M)$, $L_{Q_{mn}}(M)$, $L_{E_{mn}}(M)$ and $L_{R_{mn}}(M)$ are constants depending on $M$. We can show the constants are bounded based on Lemma C.3. Thus, the functions satisfy the Lipschitz condition. For $F_{D_m}(M)$, gradient norm is given by

$$kr_M F_{D_m}(M)k = \sqrt{1 + \frac{2}{D_m^4}}: \tag{63}$$

The left-hand side is also bounded since Lemma C.5 indicates that for all $m = 1; : : : ; M$, $D_m(t)$ 6 0 for any $t > 0$. Thus, $F_{D_m}(M)$ also satisfy the Lipschitz condition. □

# D LOCAL STABILITY ANALYSIS OF FIXED POINTS OF ODES

In this section, we provide additional details on the local stability analysis of the ODEs. In what follows, we will omit straightforward calculations related to the eigenvalue computations.

## D.1 Stability Analysis of Model-Matched Case

For the model-matched case, the macroscopic state is described by 6 variables. For the sake of simplicity, we only consider the case $= 0$ and small learning limit $= _W = _V = _D$. The xed points are given by the condition $dM =dt = 0$. From Eq. (18)-(23), the xed point equations given by

$$\begin{cases}
F_{m_{11}}(M) = \quad d_{11}( + ) \quad m_{11}( d_{11}^2 + E_{11} + D_{11}) = 0 \\
F_{d_{11}}(M) = \quad ( + )(m_{11} \quad (Q_{11} + )d_{11}) = 0 \\
F_{Q_{11}}(M) = 2 \quad ( m_{11}d_{11} + R_{11}) \quad Q_{11}( d_{11}^2 + E_{11} + D_1) = 0 \\
F_{E_{11}}(M) = 2 \quad ( m_{11}d_{11} + R_{11}) \quad (Q_{11} + )( d_{11}^2 + E_{11}) = 0 \\
F_{R_{11}}(M) = \quad (1 \quad R_{11})( d_{11}^2 + E_{11}) \quad D_1R_{11} + ( m_{11}^2 + Q_{11}) \quad (Q_{11} + )( m_{11}d_{11} + R_{11}) = 0 \\
F_{D_1}(M) = \quad \frac{}{D_1} \quad (Q_{11} + ) = 0;
\end{cases} \tag{64}$$

where $M = ( m_{11}; d_{11}; Q_{11}; E_{11}; R_{11}; D_1)$ are the stationary macroscopic state. The local stability of a xed point is identi ed by whether the Jacobian matrix

$$J(M) = \begin{pmatrix}
\frac{@F_{m_{11}}}{@m_{11}} & \frac{@F_{m_{11}}}{@d_{11}} & \frac{@F_{m_{11}}}{@Q_{11}} & \frac{@F_{m_{11}}}{@E_{11}} & \frac{@F_{m_{11}}}{@R_{11}} & \frac{@F_{m_{11}}}{@D_1} \\
\frac{@F_{d_{11}}}{@m_{11}} & \frac{@F_{d_{11}}}{@d_{11}} & \frac{@F_{d_{11}}}{@Q_{11}} & \frac{@F_{d_{11}}}{@E_{11}} & \frac{@F_{d_{11}}}{@R_{11}} & \frac{@F_{d_{11}}}{@D_1} \\
\frac{@F_{Q_{11}}}{@m_{11}} & \frac{@F_{Q_{11}}}{@d_{11}} & \frac{@F_{Q_{11}}}{@Q_{11}} & \frac{@F_{Q_{11}}}{@E_{11}} & \frac{@F_{Q_{11}}}{@R_{11}} & \frac{@F_{Q_{11}}}{@D_1} \\
\frac{@F_{E_{11}}}{@m_{11}} & \frac{@F_{E_{11}}}{@d_{11}} & \frac{@F_{E_{11}}}{@Q_{11}} & \frac{@F_{E_{11}}}{@E_{11}} & \frac{@F_{E_{11}}}{@R_{11}} & \frac{@F_{E_{11}}}{@D_1} \\
\frac{@F_{R_{11}}}{@m_{11}} & \frac{@F_{R_{11}}}{@d_{11}} & \frac{@F_{R_{11}}}{@Q_{11}} & \frac{@F_{R_{11}}}{@E_{11}} & \frac{@F_{R_{11}}}{@R_{11}} & \frac{@F_{R_{11}}}{@D_1} \\
\frac{@F_{D_1}}{@m_{11}} & \frac{@F_{D_1}}{@d_{11}} & \frac{@F_{D_1}}{@Q_{11}} & \frac{@F_{D_1}}{@E_{11}} & \frac{@F_{D_1}}{@R_{11}} & \frac{@F_{D_1}}{@D_1}
\end{pmatrix} \tag{65}$$

has eigenvalue with non-negative real part or not. Solving Eq. 64 and computing the eigenvalues of the Jacobian, one easily finds that fixed points other than two cases have positive eigenvalues for any , and , indicating that they are unstable fixed points. Subsequently, we focus on the two cases. In the following, the shorthand expression $P = \ +\ $ is employed.

Type (1): Posterior Collapsed Fixed Point. It is easy to verify that

$$m_{11} = d_{11} = Q_{11} = E_{11} = R_{11} = 0; \quad D_1 = 1 \tag{66}$$

is a solution of Eq. (64). This fixed point indicates that the VAE encounters a posterior collapse. From a straightforward eigenvalue computation, the six eigenvalues can be expressed as follows:

$$\begin{aligned}
{}_1 &= \ = \ -\frac{}{2}; \quad {}_2 = \ = \ (1+\ ) \\
{}_3 &= \ = \ 1+\ +\sqrt{(1+\ )^2+4\ (\ \ )}\ ; \quad {}_4 = \ = \ 1+\ \ \sqrt{(1+\ )^2+4\ (\ \ )} \\
{}_5 &= \ = \ \frac{1}{2}\ 1+P+\sqrt{(1+P)^2+4P(P\ \ )}\ ; \quad {}_6 = \ = \ \frac{1}{2}\ 1+P\ \ \sqrt{(1+P)^2+4P(P\ \ )}
\end{aligned}$$

Here, ${}_4$ is positive when $<$ , ${}_6$ is when $<P$ is positive and the others are negative for any , and . Thus, type (1) fixed point is stable if $P <$ . Moreover, all other fixed points are unstable when $P <$ , which indicates that a threshold of the posterior collapse is $= P$.

Type (2): Learnable Fixed Point. The fixed points equation Eq. (64) have following solution:

$$m_{11} = \ \sqrt{P\ \ }; \ d_{11} = \ \frac{\sqrt{P}}{P}; \ Q_{11} = P\ \ ; \ E_{11} = \frac{P}{P^2}; \ R_{11} = \frac{P}{P}; \ D_1 = \frac{}{P} \tag{67}$$

The Jacobian of this fixed point possesses six eigenvalues. The three eigenvalues of them can be expressed as follows:

$$\begin{aligned}
{}_1 &= \ = \ (1+P) \\
{}_2 &= \ = \ 1+P+\sqrt{(1+P)^2\ \ 4} \\
{}_3 &= \ = \ 1+P\ \ \sqrt{(1+P)^2\ \ 4}
\end{aligned}$$

These three eigenvalues are negative for any, , and . The other three eigenvalues can be expressed as the solutions to the following equation:

$$-\ \ ^3+P^2\ \ P^2+2(1+P^2)\ \ -\ \ ^2+2P^4\ \ P^2(1+P^2)\ \ 8\ ^3+2(1+4P)\ ^2\ \ 2P\ \ -+8P^8(P\ \ )\ ^3=0:$$

One of the solutions to this equation is positive when $P >$ . Furthermore, by substituting $= P$, the equation can be expressed as

$$-\ \ -+P^4\ \ -+2P^3(1+P^2)\ \ =0;$$

indicating that $= 0$ when $= P$. Thus, type (2) fixed point is stable when $+$ .

D.2 Stability Analysis of Model-Mismatched Case

For the model-mismatched case, the macroscopic state is described by 16 variables. For the sake of simplicity, we also consider the case $= 0$ and small learning limit $= {}_W = {}_V = {}_D$. The specific fixed-point equations and their Jacobians can be derived from Eq. (18)-(23), just as in the model-matched case. However, they are not displayed here due to their length. Similarly, all fixed points other than three cases are unstable fixed points as in the model-matched case, as the eigenvalues of their Jacobians take positive values for any , and . Subsequently, we focus on the three types in detail.

**Type (1): Posterior Collapsed Fixed Point.** It is easy to verify that the following state is a solution of the ODEs:

$$m = d = \mathbf{0}_2, \quad Q = E = R = \mathbf{0}_{2\ 2}, \quad D = \mathbf{1}_2$$

The eigenvalues of the Jacobian can be expressed as follows:

$$\begin{aligned}
&\frac{\lambda_1}{\tau} = \frac{\lambda_2}{\tau} = -\frac{\beta}{2}, \quad \frac{\lambda_3}{\tau} = \frac{\lambda_4}{\tau} = \frac{\lambda_5}{\tau} = \frac{\lambda_6}{\tau} = -(1+\beta\eta) \\
&\frac{\lambda_7}{\tau} = \frac{\lambda_8}{\tau} = \frac{\lambda_9}{\tau} = -\ 1+\beta\eta+\sqrt{(1+\beta\eta)^2+4\eta(\eta-\beta)} \\
&\frac{\lambda_{10}}{\tau} = \frac{\lambda_{11}}{\tau} = \frac{\lambda_{12}}{\tau} = -\ 1+\beta\eta-\sqrt{(1+\beta\eta)^2+4\eta(\eta-\beta)} \\
&\frac{\lambda_{13}}{\tau} = \frac{\lambda_{14}}{\tau} = -\frac{1}{2}\ 1+\beta P+\sqrt{(1+\beta P)^2+4P(P-\beta)} \\
&\frac{\lambda_{15}}{\tau} = \frac{\lambda_{16}}{\tau} = -\frac{1}{2}\ 1+\beta P-\sqrt{(1+\beta P)^2+4P(P-\beta)}
\end{aligned}$$

These eigenvalue are positive when $\rho+\eta<\beta$, zero when $\rho+\eta=\beta$ and negative when $\rho+\eta>\beta$ as in the model-matched case. Thus this fixed solution is stable when $\rho+\eta\le\beta$.

**Type (2): Overfitting Fixed Point.** The fixed point equations have the following solution:

$$\begin{aligned}
m &= \pm\sqrt{P-\beta},0\ , \ d = \pm\frac{\sqrt{P-\beta}}{P},0 \\
Q &= \begin{matrix} P-\beta & 0 \\ 0 & \eta-\beta \end{matrix}\ , \ E = \begin{matrix} \frac{P-\beta}{P^2} & 0 \\ 0 & \frac{\eta-\beta}{\eta^2} \end{matrix}\ , \ R = \begin{matrix} \frac{P-\beta}{P} & 0 \\ 0 & \frac{\eta-\beta}{\eta} \end{matrix}\ , \ D = \frac{\beta}{P},\frac{\beta}{\eta}
\end{aligned}$$

and

$$\begin{aligned}
m &= 0,\pm\sqrt{P-\beta}\ , \ d = 0,\pm\frac{\sqrt{P-\beta}}{\rho+\eta} \\
Q &= \begin{matrix} \eta-\beta & 0 \\ 0 & P-\beta \end{matrix}\ , \ E = \begin{matrix} \frac{\eta-\beta}{\eta^2} & 0 \\ 0 & \frac{P-\beta}{P^2} \end{matrix}\ , \ R = \begin{matrix} \frac{\eta-\beta}{\eta} & 0 \\ 0 & \frac{P-\beta}{P} \end{matrix}\ , \ D = \frac{\beta}{\eta},\frac{\beta}{P}
\end{aligned}$$

The eigenvalues of the Jacobian can be expressed as follows:

$$\begin{aligned}
&\frac{\lambda_1}{\tau} = -2(1+\eta^2), \quad \frac{\lambda_2}{\tau} = -(1+\eta P) \\
&\frac{\lambda_3}{\tau} = -\frac{1}{2}\ 1+\eta P+2(1+\eta)+\sqrt{(1+\eta P)^2-4\eta\rho}\ , \ \frac{\lambda_4}{\tau} = -\frac{1}{2}\ 1+\eta P+2(1+\eta)-\sqrt{(1+\eta P)^2-4\eta\rho} \\
&\frac{\lambda_5}{\tau} = -\ 1+\eta P+\sqrt{(1+\eta P)^2-4\eta\rho}\ , \ \frac{\lambda_6}{\tau} = -\ 1+\eta P-\sqrt{(1+\eta P)^2-4\eta\rho} \\
&\frac{\lambda_7}{\tau} = \frac{\lambda_8}{\tau} = -\frac{1}{2}\ 1+\eta P+\sqrt{(1+\eta P)^2+8\beta\sqrt{(\beta-\eta)(\beta-P)}+\beta-P+\frac{\rho}{2}} \\
&\frac{\lambda_9}{\tau} = \frac{\lambda_{10}}{\tau} = -\frac{1}{2}\ 1+\eta P-\sqrt{(1+\eta P)^2+8\beta\sqrt{(\beta-\eta)(\beta-P)}+\beta-P+\frac{\rho}{2}}
\end{aligned}$$

Here, the real parts $\mathrm{Re}(\lambda_9)$ and $\mathrm{Re}(\lambda_{10})$ are positive when $\beta>\rho/2+\eta$ and the others are negative for any $\beta$, $\eta$ and $\rho$. Additionally, the other eigenvalues are represented as solutions to the following equations:

$$\left(\frac{\lambda}{\tau}\right)^3+(\eta^3+2\beta\eta(1+\eta^2))\left(\frac{\lambda}{\tau}\right)^2-2\beta\eta^2\ 8\beta^3+2\beta(\eta-\beta(1+4\eta))-\eta^2(1+\eta^2)\ P^4\frac{\lambda}{\tau}-8\beta^3(\beta-\eta)\eta^5P^6=0 \tag{68}$$

$$\left(\frac{\lambda}{\tau}\right)^3+\ \eta P^4+2\eta\beta P^2(1+P^2)\ \left(\frac{\lambda}{\tau}\right)^2+2\beta\eta^2P^4\ (P^2+P^4)+2\beta^2(1+4P)-2\beta(4\beta^2+P)\ \frac{\lambda}{\tau}$$
$$-8\beta^3\eta^3(\beta-P)P^8=0 \tag{69}$$

One solution of Eq. (68) is positive when $\beta > \eta$, and Eq. (68) can be expressed as follows when $\beta = \eta$:

$$\left(\frac{\lambda}{\tau}\right)^2 + \frac{\lambda}{\tau}\eta^2(2+\eta+2\eta^2)P^2 + 2\eta^5(1+\eta^2)P^4 = 0,$$

indicating that $\lambda = 0$ when $\beta = \eta$. One solution of Eq. (69) is positive when $\beta > \eta + \rho$ and Eq. (69) can be expressed as follows when $\beta = \eta$:

$$\left(\frac{\lambda}{\tau}\right)^2 + \frac{\lambda}{\tau}\eta P^3(2+P+2P^2) + 2\eta^2 P^7(1+P^2) = 0, \tag{70}$$

indicating that $\lambda = 0$ when $\beta = \eta + \rho$. Thus, type (2) is stable when $\eta \leq \beta \leq \eta + \beta$.

**Type (3): Learnable Fixed Point.** The fixed point equation has the following solution:

$$m = \pm\sqrt{P-\beta}, 0 \ , \ d = \pm\frac{\sqrt{P-\beta}}{P}, 0$$
$$Q = \begin{pmatrix} P-\beta & 0 \\ 0 & 0 \end{pmatrix} , \ E = \begin{pmatrix} \frac{P-\beta}{P^2} & 0 \\ 0 & 0 \end{pmatrix} , \ R = \begin{pmatrix} \frac{P-\beta}{P} & 0 \\ 0 & 0 \end{pmatrix} , \ D = \frac{\beta}{P}, 1$$

and

$$m = 0, \pm\sqrt{P-\beta} \ , \ d = 0, \pm\frac{\sqrt{P-\beta}}{P}$$
$$Q = \begin{pmatrix} 0 & 0 \\ 0 & P-\beta \end{pmatrix} , \ E = \begin{pmatrix} 0 & 0 \\ 0 & \frac{P-\beta}{P^2} \end{pmatrix} , \ R = \begin{pmatrix} 0 & 0 \\ 0 & \frac{P-\beta}{\eta+\rho} \end{pmatrix} , \ D = 1, \frac{\beta}{P}$$

The eigenvalue of the Jacobian can be expressed as follows:

$$\frac{\lambda_1}{\tau} = -\frac{\beta}{2}, \quad \frac{\lambda_2}{\tau} = -(1+\eta P) \quad \frac{\lambda_3}{\tau} = -(1+\beta\eta),$$
$$\frac{\lambda_4}{\tau} = -\left(1+\beta\eta+\sqrt{(1+\beta\eta)^2+4\eta(\eta-\beta)}\right) \ , \quad \frac{\lambda_5}{\tau} = -\left(1+\beta\eta-\sqrt{(1+\beta\eta)^2+4\eta(\eta-\beta)}\right) \ ,$$
$$\frac{\lambda_6}{\tau} = -\frac{1}{2}\left(1+\beta P+\sqrt{(1+\beta P)^2+4\beta(\beta-P)}\right) \ , \quad \frac{\lambda_7}{\tau} = -\frac{1}{2}\left(1+\beta P-\sqrt{(1+\beta P)^2+4\beta(\beta-P)}\right) \ ,$$
$$\frac{\lambda_8}{\tau} = -\left(1+\eta P+\sqrt{(1+\eta P)^2-4\eta\rho}\right) \ , \quad \frac{\lambda_9}{\tau} = -\left(1+\eta P-\sqrt{(1+\eta P)^2-4\eta\rho}\right) \ ,$$
$$\frac{\lambda_{10}}{\tau} = -\frac{1}{2}\left(2+\eta(\beta+P)+\left((1+\eta\beta)^2+(1+\eta P)^2+4\eta(\eta-\rho)+2\sqrt{\left((1-\beta\eta)^2+4\eta^2\right)\left((1+\eta P)^2-2\eta P\right)}\right)^{1/2}\right) \ ,$$
$$\frac{\lambda_{11}}{\tau} = -\frac{1}{2}\left(2+\eta(\beta+P)-\left((1+\eta\beta)^2+(1+\eta P)^2+4\eta(\eta-\rho)+2\sqrt{\left((1-\beta\eta)^2+4\eta^2\right)\left((1+\eta P)^2-2\eta P\right)}\right)^{1/2}\right) \ ,$$
$$\frac{\lambda_{12}}{\tau} = -\frac{1}{2}\left(2+\eta(\beta+P)+\left((1+\eta\beta)^2+(1+\eta P)^2+4\eta(\eta-\rho)-2\sqrt{\left((1-\beta\eta)^2+4\eta^2\right)\left((1+\eta P)^2-2\eta P\right)}\right)^{1/2}\right) \ ,$$
$$\frac{\lambda_{13}}{\tau} = -\frac{1}{2}\left(2+\eta(\beta+P)-\left((1+\eta\beta)^2+(1+\eta P)^2+4\eta(\eta-\rho)-2\sqrt{\left((1-\beta\eta)^2+4\eta^2\right)\left((1+\eta P)^2-2\eta P\right)}\right)^{1/2}\right) \ .$$

Here, $\lambda_7$ is positive when $\beta > \rho + \eta$, $\lambda_5$ is positive when $\beta < \eta$, $\lambda_{11}$ is positive when $\beta < \bar{\eta}$ where $\eta < \bar{\eta}$ and the others are negative for any $\beta$, $\eta$ and $\rho$. The other eigenvalues are expressed as solutions to the following equation:

$$\left(\frac{\lambda}{\tau}\right)^3 + P^2\left(P^2+2\beta(1+P^2)\right)\left(\frac{\lambda}{\tau}\right)^2 + 2\beta P^4\left(-2\beta(4\beta^2+P)+P^2(1+P^2)+2\beta^2(1+2P)\right)\frac{\lambda}{\tau} - 8\beta^3(\beta-P)P^8 = 0$$

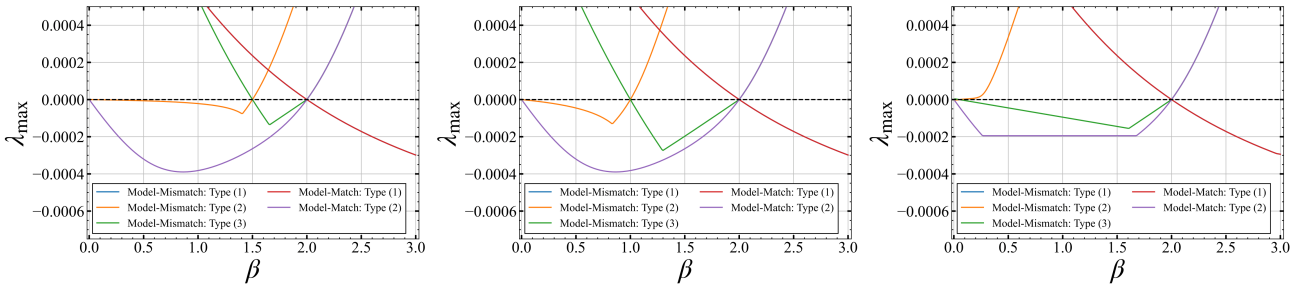


Figure 5: Max eigenvalues of Jacobian for each stable fixed points when $\rho, \eta = 0.5, 1.5$ (left), $\rho, \eta = 1.0, 1.0$ (middle) and $\rho, \eta = 1.95, 0.05$ (right) as a function of $\beta$ for both model-matched and model-mismatched cases.

eigenvalue is positive when $\beta > \rho + \eta$, $\beta = \rho + \eta$ and the equation expressed as when $\beta = P$

$$\frac{\lambda}{\tau} \quad \frac{\lambda}{\tau}^{2} + \frac{\lambda}{\tau} P^3(2 + P(1 + 2P)) + 2P^7(1 + P^2) \quad = 0$$

which indicates $\lambda = 0$. Thus, type (3) fixed point is stable when $\eta \leq \beta \leq \rho + \eta$. Fig. 5 presents all types of fixed points and their corresponding maximum eigenvalues as a function of $\beta$.

### D.3 Stability Analysis of Tanh KL Annealing

For the case of Tanh KL annealing $\beta(t) = \tanh(\gamma t)$, the fixed-point equation can be expressed as follows:

$$\begin{aligned}
&F_{m_{11}}(\mathcal{M}, \beta) = \tau \quad d_{11}(\rho + \eta) - m_{11}(\rho d_{11}^2 + \eta E_{11} + D_{11}) \quad = 0 \\
&F_{d_{11}}(\mathcal{M}, \beta) = \tau(\rho + \eta)(m_{11} - (Q_{11} + \beta)d_{11}) = 0 \\
&F_{Q_{11}}(\mathcal{M}, \beta) = 2\tau \quad (\rho m_{11} d_{11} + \eta R_{11}) - Q_{11}(\rho d_{11}^2 + \eta E_{11} + D_1) \quad = 0 \\
&F_{E_{11}}(\mathcal{M}, \beta) = 2\tau \quad (\rho m_{11} d_{11} + \eta R_{11}) - (Q_{11} + \beta)(\rho d_{11}^2 + \eta E_{11}) \quad = 0 \\
&F_{R_{11}}(\mathcal{M}, \beta) = \tau \quad (1 - R_{11})(\rho d_{11}^2 + \eta E_{11}) - D_1 R_{11} + (\rho m_{11}^2 + \eta Q_{11}) - (Q_{11} + \beta)(\rho m_{11} d_{11} + \eta R_{11}) \quad = 0 \\
&F_{D_1}(\mathcal{M}, \beta) = \tau \quad \frac{\beta}{D_1} - (Q_{11} + \beta) \quad = 0, \\
&F_{\beta}(\mathcal{M}, \beta) = \gamma(1 - \beta^2) = 0
\end{aligned}$$

This fixed-point equation has the same stable fixed points as the model-matched case; that is, type (1) posterior collapsed fixed point is stable when $\beta > \eta + \rho$ and type (2) Learnable fixed point is stable when $\beta < \eta + \rho$. Additionally, the Jacobian possesses the same eigenvalues as the model-matched case, along with a new eigenvalue of $\lambda_7 = -2\gamma$ originated from tanh KL annealing. Specifically, for the learnable fixed point, and excluding $-2\gamma$, the maximal eigenvalue can be expressed as follows when $\rho = 2 - \nu$ and $\eta = \nu$:

$$\lambda_{\max}(\nu) = \begin{cases} \frac{\tau}{2}(\sqrt{5} - 3) & \tau(1 - 2\sqrt{2} + \sqrt{5})/4 \leq \nu \leq \tau(1 + 2\sqrt{2} + \sqrt{5})/4 \\ -\tau(1 + 2\nu) + \tau \sqrt{\dfrac{}{1 - 4\nu(1 - 4\nu)}} & \text{otherwise} \end{cases} \tag{71}$$

Thus, the conditions under which tanh KL annealing slows down the convergence are expressed as

$$\gamma \leq \begin{cases} \frac{\tau}{4}(3 - \sqrt{5}), & \tau(1 - 2\sqrt{2} + \sqrt{5})/4 \leq \nu \leq \tau(1 + 2\sqrt{2} + \sqrt{5})/4 \\ \tau \quad \nu + \frac{1}{2} \quad - \tau \quad \sqrt{\nu(2\nu - 1) + \frac{1}{4}}, & \text{otherwise.} \end{cases}$$

## E ADDITIONAL RESULTS

### E.1 Linear Annealing

In this section, we demonstrate the properties of the linear annealing $\beta(t) = \gamma t$ which is used in various applications. Fig. 6 demonstrates the generalization error as a function of $t$ for both Linear and tanh KL annealing