Precise Dynamics of Diagonal Linear Networks: A Unifying Analysis by Dynamical Mean-Field Theory

Sota Nishiyama
The University of Tokyo
RIKEN AIP

Masaaki Imaizumi
The University of Tokyo
RIKEN AIP
Kyoto University

Abstract

Diagonal linear networks (DLNs) are a tractable model that captures several nontrivial behaviors in neural network training, such as initialization-dependent solutions and incremental learning. These phenomena are typically studied in isolation, leaving the overall dynamics insufficiently understood. In this work, we present a unified analysis of various phenomena in the gradient flow dynamics of DLNs. Using Dynamical Mean-Field Theory (DMFT), we derive a low-dimensional effective process that captures the asymptotic gradient flow dynamics in high dimensions. Analyzing this effective process yields new insights into DLN dynamics, including loss convergence rates and their trade-off with generalization, and systematically reproduces many of the previously observed phenomena. These findings deepen our understanding of DLNs and demonstrate the effectiveness of the DMFT approach in analyzing high-dimensional learning dynamics of neural networks.

1 INTRODUCTION

The training dynamics of neural networks have attracted significant attention in deep learning theory. It has been suggested that the dynamics induced by training algorithms strongly influence the generalization performance of neural networks. This effect is captured in the idea of *implicit bias* (Neyshabur et al., 2015), in which the algorithm selects a certain solution among many induced by nonconvexity of the loss and overparameterization of networks. Accordingly, many recent works have studied the interplay between models and optimizers, aiming to characterize the resulting implicit biases (Neyshabur, 2017; Soudry et al., 2018; Arora et al., 2019; Bartlett et al., 2021). Moreover, understanding the convergence speed and timescales of the training dynamics contributes to efficient training of high-performance models in practice, especially in the context of modern large-scale neural networks in which the training is stopped at a compute-optimal point (Kaplan et al., 2020).

For a refined understanding of the dynamics, *diagonal linear networks* (DLNs) have emerged as a tractable theoretical model that captures several nontrivial behaviors of learning algorithms, making them a valuable tool for studying neural network dynamics. Recent studies have uncovered various phenomena, such as the dependence of solutions on algorithmic parameters (Woodworth et al., 2020; Nacson et al., 2022; Pesme et al., 2021; Even et al., 2023) and incremental learning dynamics (Berthier, 2023; Pesme and Flammarion, 2023).

One of the challenges in the study of DLNs is that relationships among these phenomena remain unclear. This is because existing analyses often rely on case-specific techniques. In addition, there are aspects of the dynamics that have not yet been investigated, such as convergence speed to long-term behaviors, and these unexplored elements hinder a comprehensive understanding of the overall dynamics. Specifically, we raise the following questions:

1. Which dynamical regimes and timescales arise under different initializations, and how does performance evolve in each?
2. What solution do trained DLNs converge to, and at what rate?

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Contributions. In this work, we develop a unified framework to describe DLN dynamics and conduct a comprehensive analysis of diverse phenomena. Specifically, by leveraging Dynamical Mean-Field Theory (DMFT), which provides a precise characterization in high-dimensional limits, we derive a system of equations that characterizes the gradient flow training dynamics of DLNs in sparse regression. By analyzing the derived equations, we elucidate the long-time behavior and timescale structure of the learning process, thereby deriving insights into the dynamics and implicit bias of DLNs.

Our main contributions are summarized as follows.

- We identify distinct dynamical regimes which depend on training time and initialization scales. For large initialization, we observe a sharp transition from memorizing solutions (fit all data but generalize poorly) to generalizing solutions; for small initialization, an early *search* plateau and *incremental learning* that follows it.
- We characterize the fixed point of the gradient flow and its dependence on initialization, showing that *a smaller initialization leads to better generalization*, close to minimum ℓ_1 norm solutions. This provides an alternative derivation of the result from Woodworth et al. (2020).
- We derive convergence rates of losses in time and show that *a smaller initialization leads to slower convergence*.
- Together with the fixed point and convergence rate result, we establish *a trade-off between optimization speed and generalization performance*.

Overall, our findings deepen the theoretical understanding of DLNs and highlight the utility of DMFT as a powerful tool for probing the dynamics of high-dimensional, nonlinear learning systems.

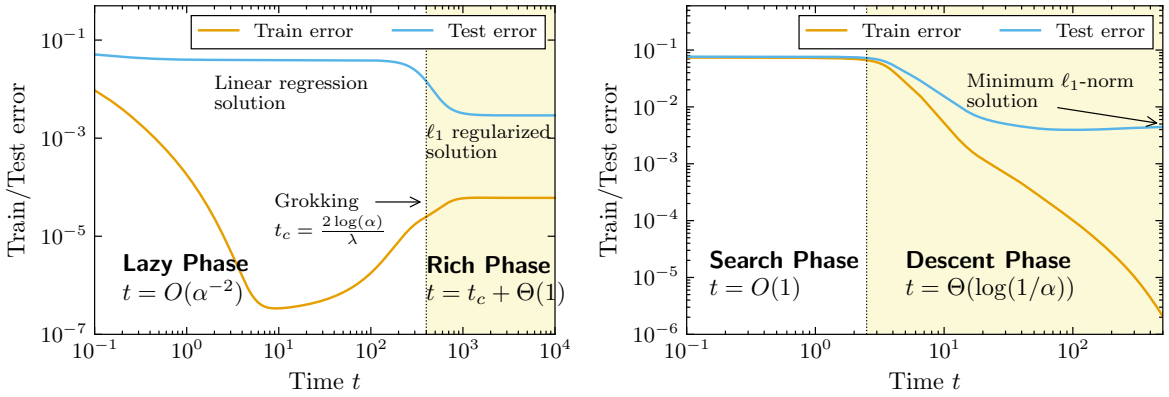
1.1 Related Work

Diagonal Linear Networks. DLNs were studied in Gunasekar et al. (2018) as a simple model that captures the rich implicit bias of neural networks. Vaskevicius et al. (2019) showed that DLNs trained with gradient descent and small initialization can implicitly perform sparse recovery. Woodworth et al. (2020) studied the implicit bias of gradient flow training for DLNs and uncovered a transition between the *kernel regime* (large initialization) and the *rich regime* (small initialization), showing that smaller initialization leads to a sparser, richer bias. Moroshko et al. (2020) studied similar phenomena in classification settings. DLNs have since become a testbed to gain insight into the implicit bias of various optimization algorithms and their hyperparameter choices, including the relative scale of layers (Azulay et al., 2021), gradient noise in stochastic gradient descent (SGD) (HaoChen et al., 2021; Pesme et al., 2021; Even et al., 2023), step size (Nacson et al., 2022), early stopping time (Li et al., 2021), and other optimizers (Papazov et al., 2024; Clara et al., 2025).

Beyond implicit bias, several works investigated the training dynamics of DLNs. Berthier (2023) and Pesme et al. (2021) identified an *incremental learning* or *saddle-to-saddle dynamics* in DLNs with small initialization, where the parameter coordinates are sequentially activated to learn the true solution.

Dynamical Mean-Field Theory. Dynamical mean-field theory (DMFT) is a technique to reduce high-dimensional random dynamics into a low-dimensional effective process characterized by a system of integro-differential equations. Originally developed in statistical physics to analyze the Langevin dynamics of spin glasses (Sompolinsky and Zippelius, 1981, 1982; Crisanti et al., 1993; Cugliandolo and Kurchan, 1993), DMFT has been applied to a wide range of problems involving many degrees of freedom with random interactions; see Cugliandolo (2024) for a recent survey. Over the last decade, DMFT has been applied to several high-dimensional optimization and estimation problems (Agoritsas et al., 2018; Sarao Mannelli et al., 2020; Mignacco et al., 2020; Bordelon and Pehlevan, 2022; Montanari and Urbani, 2025), with rigorous derivations established in certain settings (Celentano et al., 2021; Gerbelot et al., 2024; Fan et al., 2025).

The most closely related to ours is that of Montanari and Urbani (2025), who applied DMFT to wide two-layer networks and uncovered a timescale separation for generalization and overfitting. Our work differs from theirs in several aspects. Regarding the model, while two-layer DLNs analyzed in this work can be interpreted as a special case of a general two-layer neural network, they consider a narrow (compared to the input dimension) two-layer neural network with fully-connected first layer, while the DLN we consider has a diagonal first layer with width equal to the input dimension. We also consider a weight decay term not considered in their work. Regarding the analytical focus, we analyze timescale structures of gradient flow training in Section 4 in a similar spirit to Montanari and Urbani (2025); however, the result is qualitatively different. In addition, we go beyond the timescale analysis and analyze long-time behaviors of the dynamics in Section 5.



(a) Large initialization ($\alpha \gg 1$). (b) Small initialization ($\alpha \ll 1$).

Figure 1: Schematic illustrations of the timescale structures of gradient flow dynamics in DLNs.

2 PRELIMINARIES

2.1 Notation

For vectors $\boldsymbol{x} = (x_1, ..., x_d)^\mathsf{T}, \boldsymbol{y} = (y_1, ..., y_d)^\mathsf{T} \in \mathbb{R}^d$, $\boldsymbol{x} \odot \boldsymbol{y}$ denotes entry-wise multiplication, i.e., $\boldsymbol{x} \odot \boldsymbol{y} = (x_1 y_1, \ldots, x_d y_d)^\mathsf{T} \in \mathbb{R}^d$. For $\boldsymbol{x} \in \mathbb{R}^d$ and $L \in \mathbb{N}$, $\boldsymbol{x}^L$ denotes entry-wise power. $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ denotes the $d \times d$ identity matrix. $\mathbf{1}_d \in \mathbb{R}^d$ denotes the all-ones vector $\mathbf{1}_d = (1, \ldots, 1)^\mathsf{T}$. $\mathsf{GP}(0, Q)$ denotes a centered Gaussian process with covariance kernel Q. We denote by ST the soft thresholding function $\mathrm{ST}(x; \tau) \coloneqq \mathrm{sign}(x) \max\{|x| - \tau, 0\}$.

2.2 Setup

Data Model. We consider n i.i.d. samples $(\boldsymbol{x}_\mu, y_\mu) \in \mathbb{R}^d \times \mathbb{R}$ indexed by $\mu = 1, \ldots, n$. The input vectors $\boldsymbol{x}_\mu$ are sampled independently from the isotropic Gaussian distribution $\mathsf{N}(0, \mathbf{I}_d/d)$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be a data matrix whose μ-th row is $\boldsymbol{x}_\mu$. The labels y_μ follow a linear model $y_\mu = \boldsymbol{w}^{*\mathsf{T}} \boldsymbol{x}_\mu + \xi_\mu$, where $\xi_\mu \sim \mathsf{N}(0, \sigma^2)$ is some independent noise with mean 0 and variance σ^2, and $\boldsymbol{w}^* \in \mathbb{R}^d$ is a target vector. The empirical distribution of the entries of $\boldsymbol{w}^*$ converges to P_* as $d \to \infty$. We define a label vector $\boldsymbol{y} \coloneqq (y_1, \ldots, y_n)^\mathsf{T} \in \mathbb{R}^n$ and a scale term $\rho^2 \coloneqq \|\boldsymbol{w}^*\|_2^2/d$.

Diagonal Linear Network. We consider a two-layer diagonal linear network:

$$f(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{w}^\mathsf{T} \boldsymbol{x}\,, \quad \boldsymbol{w} = \frac{1}{2}(\boldsymbol{u}^2 - \boldsymbol{v}^2)\,, \tag{1}$$

for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $\boldsymbol{u}^2, \boldsymbol{v}^2$ denote entry-wise squares. This can be considered as a two-layer linear neural network with diagonal first layers $\mathrm{diag}(\boldsymbol{u}), \mathrm{diag}(\boldsymbol{v})$ and second layers $\boldsymbol{u}, \boldsymbol{v}$. Although it represents only linear functions in $\boldsymbol{x}$, the nonlinear reparameterization of $\boldsymbol{w}$ induces nontrivial dynamics.

Training Algorithm. We train the DLN by minimizing the following regularized quadratic loss:

$$L(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{2d}(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2), \tag{2}$$

where $\lambda \geq 0$ is a regularization parameter. We consider full-batch gradient flow (continuous-time gradient descent) for time $t \geq 0$ to minimize the loss:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{u}(t), \boldsymbol{v}(t)) = -\frac{d}{2}(\nabla_{\boldsymbol{u}}, \nabla_{\boldsymbol{v}}) L(\boldsymbol{u}(t), \boldsymbol{v}(t))\,, \tag{3}$$

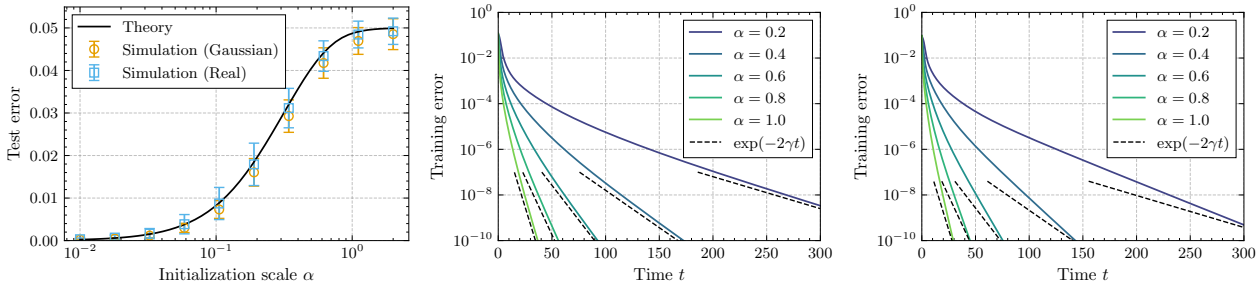
with initial values $\boldsymbol{u}(0) = \boldsymbol{v}(0) = \alpha \mathbf{1}_d$ for $\alpha > 0$. We denote the loss at time t by $L(t) \coloneqq L(\boldsymbol{u}(t), \boldsymbol{v}(t))$.

Proportional Asymptotics. We analyze the proportional asymptotic regime where $n, d \to \infty$ with $n/d \to \delta \in (0, \infty)$. Analyses in this regime have yielded significant insights into high-dimensional learning systems through exact predictions of asymptotic performance via tools from statistical physics and random matrix theory (Zdeborová and Krzakala, 2016; Mei and Montanari, 2022).

2.3 Overview of Main Findings

First, we analyze timescale structures of the dynamics to show that they exhibit *qualitatively different behaviors depending on the initialization scale* α, as depicted in Figure 1. We discuss the details in Section 4.

Large Initialization ($\alpha \gg 1$). DLNs initially behave as approximately linear models (*lazy regime*). In this phase, the loss rapidly decreases, but the model generalizes poorly. When the model is regularized ($\lambda > 0$), it then transitions to a sparse, generalizing solution (*rich regime*) around the time $t \approx 2\log(\alpha)/\lambda$ (*grokking*).



(a) Test errors at fixed points.

(b) Convergence rates of training errors for Gaussian data.

(c) Convergence rates of training errors for real data.

Figure 2: Long-time behaviors of DLNs for $\lambda = 0$ and $\delta = 0.5$. (a): Smaller initialization α leads to better generalization at the fixed point (Result 1 Case (iii)). Simulations are run 10 times on independent data, and the error bars indicate one standard deviation. (b), (c): Smaller initialization α leads to slower convergence (Result 2), thus showing a trade-off with the generalization performance. The slopes of the dashed lines represent theoretical predictions for the convergence rate. Experimental details are discussed in Section 7.

Small Initialization ($\alpha \ll 1$). We observe an early plateau with negligible change in the loss (*search phase*). The loss then decreases on a timescale of $\Theta(\log(1/\alpha))$ (*descent phase*) via *incremental learning*, in which the model learns target coordinates one by one.

Second, we analyze long-time behaviors of the dynamics to identify their fixed points (long-time limit) and convergence rates (speed of convergence in time to the fixed point). Here, we focus on the unregularized ($\lambda = 0$) and overparameterized ($\delta < 1$) case, as it exhibits the most distinctive behaviors, as illustrated in Figure 2. Details are discussed in Section 5.

Smaller Initialization Improves Generalization. We show that the fixed point of the gradient flow matches the solution of a minimum norm interpolation problem, i.e., minimizes a certain norm of $\boldsymbol{w}$ while fitting all data. The norm depends on the initialization scale α, with a smaller α enforcing a stronger bias towards sparse solutions. This implies that *smaller initialization leads to better generalization* in sparse regression settings (see Figure 2a).

Trade-off Between Generalization and Convergence. We show that the loss converges exponentially as $L(t) \sim \mathrm{e}^{-2\gamma t}$, with the exponent γ monotonically increasing with initialization α. Thus, *smaller initialization leads to slower convergence* (see Figure 2b) and, combined with the fixed point characterization, this reveals a *trade-off between the generalization performance and the convergence speed.*

3 DMFT ANALYSIS

We apply Dynamical Mean-Field Theory (DMFT) to the gradient flow (3) with randomness coming from samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$. DMFT is a technique from statistical physics that provides a low-dimensional effective description by averaging out microscopic fluctuations, thereby capturing the macroscopic behavior of high-dimensional systems in a tractable manner. In particular, the DMFT equation consists of stochastic process that characterize the high-dimensional dynamics of the model parameters and deterministic functions called *correlation* and *response* functions which encode the evolution of the macroscopic properties of the system.

In our setting, the DMFT formalism for the gradient flow (3) yields the following system, involving correlation and response functions $C_w, C_f, R_w, R_f \colon \mathbb{R}^2_{\geq 0} \to \mathbb{R}$ and stochastic processes $w, g \colon \mathbb{R}_{\geq 0} \to \mathbb{R}$:

$$C_w(t,t') = \mathbb{E}[(w(t) - w^*)(w(t') - w^*)]\,, \tag{4a}$$

$$R_w(t,t') = -\,\mathbb{E}\left[\frac{\partial w(t)}{\partial z(t')}\right], \tag{4b}$$

$$\begin{aligned} C_f(t,t') = C_w(t,t') + \sigma^2 \\ - \int_0^{t'} R_f(t',s)(C_w(t,s) + \sigma^2)\,\mathrm{d}s \\ - \int_0^{t} R_w(t,s) C_f(t',s)\,\mathrm{d}s\,, \end{aligned} \tag{4c}$$

$$R_f(t,t') = R_w(t,t') - \int_{t'}^{t} R_w(t,s) R_f(s,t')\,\mathrm{d}s\,, \tag{4d}$$

$$\begin{aligned} g(t) = \frac{z(t)}{\delta} + w(t) - w^* \\ - \int_0^t R_f(t,s)(w(s) - w^*)\,\mathrm{d}s\,, \end{aligned} \tag{4e}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = -\sqrt{w(t)^2 + \alpha^4 \mathrm{e}^{-2\lambda t}}g(t) - \lambda w(t)\,, \qquad (4f)$$

where $z \sim \mathsf{GP}(0, \delta C_f)$ and $w^* \sim P_*$.

As $d \to \infty$, the empirical distribution of entries of $\boldsymbol{w}(t)$ converges to the law of the process $w(t)$. In this way, the high-dimensional dynamics (3) is reduced to a scalar-valued stochastic process (4f), which is more tractable and amenable to theoretical analysis.

Macroscopic quantities, such as training and test errors, can be asymptotically computed from the solution (C_w, C_f, R_w, R_f) of the DMFT equation (4):

$$E_{\text{train}}(t) \coloneqq \frac{1}{n}\sum_{\mu=1}^{n}(y_\mu - \boldsymbol{w}(t)^\mathsf{T}\boldsymbol{x}_\mu)^2 \to C_f(t,t)\,, \quad (5)$$

$$E_{\text{test}}(t) \coloneqq \mathbb{E}_{\boldsymbol{x},y}[(y - \boldsymbol{w}(t)^\mathsf{T}\boldsymbol{x})^2] \to C_w(t,t) + \sigma^2\,. \quad (6)$$

We provide a heuristic derivation of the DMFT equation (4) based on statistical physics in Appendix A and present a rigorous justification in Section 6. It is also validated against numerical simulations in Section 7.

4 LEARNING TIMESCALES

In this section, we analyze the timescale structure of the DMFT equation (4) for the gradient flow (3). We identify qualitatively distinct behaviors unfolding across different timescales, depending on the initialization scale α, as depicted in Figures 1 and 3.

To illustrate our technique, we analyze the simplified case of the infinite-data limit ($\delta \to \infty$), where the DMFT equation (4) reduces to the following scalar ordinary differential equation (ODE).

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = -\sqrt{w(t)^2 + \alpha^4 \mathrm{e}^{-2\lambda t}}(w(t) - w^*) - \lambda w(t)\,. \qquad (7)$$

The full analysis for general δ appears in Appendix B.

4.1 Technique: Singular Perturbation Theory

We analyze the dynamics (7) in $\alpha \to \infty$ and $\alpha \to 0$ limits using *singular perturbation theory* (Bender and Orszag, 1999). It is a useful technique in the study of dynamical systems, which allows us to separate the behaviors of dynamical systems into different timescales. Following Montanari and Urbani (2025), we proceed heuristically: posit an ansatz on a given timescale and check consistency with the DMFT equation. We also validate against numerical simulations. Although widely used, its rigorous treatment is challenging and is beyond the scope of this paper.

4.2 Large Initialization Limit $\alpha \to \infty$

In this case, the dynamics exhibit two distinct dynamical regimes: *lazy phase* for $t = O(\alpha^{-2})$ and *rich phase* for $t = 2\log(\alpha)/\lambda + \Theta(1)$. In each phase, we first analyze the dynamics in the $\delta \to \infty$ limit, and then discuss the behavior for general δ. Note that the discussion for general δ is based on the analysis in Appendix B.

Lazy Phase: $t = O(\alpha^{-2})$. In this timescale, the factor $\sqrt{w(t)^2 + \alpha^4 \mathrm{e}^{-2\lambda t}}$ in Equation (7) can be approximated by α^2 (This is because $w(t)^2 \ll \alpha^4$ as $w(0) = 0$ and $\mathrm{e}^{-2\lambda t} \approx 1$ as $\lambda t \ll 1$ for small t). Thus, we have the following approximate ODE (with random $w^* \sim P_*$):

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = -\alpha^2(w(t) - w^*)\,, \qquad (8)$$

with an explicit solution $w(t) = w^*(1 - \mathrm{e}^{-\alpha^2 t})$, showing that $w(t)$ converges exponentially to the target w^*.

Thus, DLNs essentially behave as unregularized linear models on this timescale. This phenomenon corresponds to the *lazy training* in which models with large weights behave as linearized models around their initializations (Jacot et al., 2018; Chizat et al., 2019). In Appendix B.1, we show for general δ that DLNs behave as linear models and converge to linear regression solutions in time $O(\alpha^{-2})$.

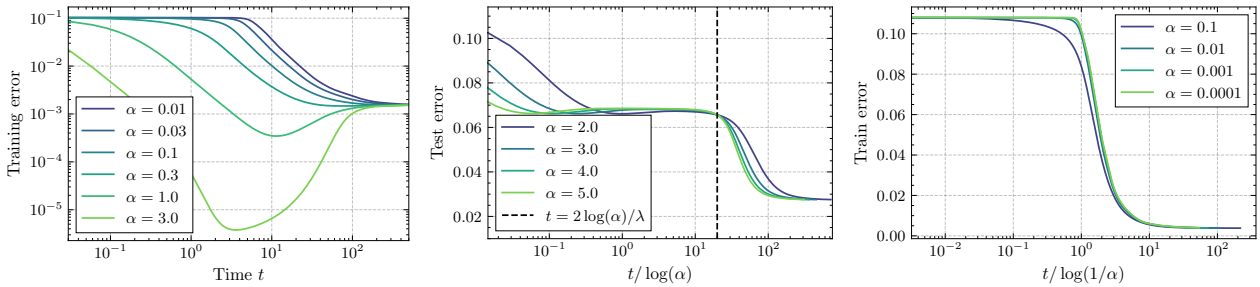
Rich Phase: $t = 2\log(\alpha)/\lambda + \Theta(1)$. When $\lambda > 0$, as t grows, the $\mathrm{e}^{-2\lambda t}$ factor in Equation (7) becomes small and eventually breaks the approximation $\sqrt{w(t)^2 + \alpha^4 \mathrm{e}^{-2\lambda t}} \approx \alpha^2$. This occurs when the two terms inside the square root become of the same order, which occurs at time $t \approx t_c \coloneqq 2\log(\alpha)/\lambda$.

After the approximation breaks down, the dynamics transitions to the next dynamical regime governed by a different equation. Shifting the time as $\tau = t - t_c$ and dropping the $\alpha^2 \mathrm{e}^{-2\lambda t} = \mathrm{e}^{-2\lambda \tau}$ term for $\tau \gg 1$, the dynamics (7) is expressed as

$$\frac{\mathrm{d}}{\mathrm{d}\tau}w(\tau) \approx -|w(\tau)|(w(\tau) - w^*) - \lambda w(\tau)\,. \qquad (9)$$

This is a logistic equation that can be solved explicitly. Setting $\Delta \coloneqq |w^*| - \lambda$, the solution is given by $w(\tau) = \mathrm{sign}(w^*)\Delta/(1 + C\mathrm{e}^{-\Delta\tau})$, where C is a constant. The fixed point as $\tau \to \infty$ can be expressed using the soft-thresholding function as $w(\infty) = \max\{\Delta, 0\} = \mathrm{ST}(w^*; \lambda)$, reminiscent of ℓ_1 regularization. Furthermore, the convergence rate is $|w(\tau) - w(\infty)| \sim \mathrm{e}^{-|\Delta|\tau}$, showing slower convergence for paths with target $|w^*|$ closer to the threshold λ.

In summary, the dynamics transitions at time $t \approx 2\log(\alpha)/\lambda$ to the second dynamical regime, the *rich*



(a) Training error dynamics. (b) Grokking timescales for large α. (c) Descent timescales for small α.

Figure 3: (a): Training error dynamics for various initialization scales α. Plots are simulations of DLNs with $d = 200$. We observe qualitatively different dynamics depicted in Figure 1. The monotonicity of the training error changes at around $\alpha \approx 0.3$. (b): Test error dynamics for large α. Once the time is rescaled by $\log(\alpha)$, the transition times to the second dynamical regime collapse, showing that it is the correct scaling for the transition time. (c): Training error dynamics for small α. Once the time is rescaled by $\log(1/\alpha)$, descent phases start and proceed on the same timescales.

phase, which exhibits nonlinear dynamics and sparsity bias in contrast to the lazy phase. The transition timescale is validated against simulations in Figure 3b.

Connection to grokking: The transition to the rich phase is sharp in the sense that the dynamics after the transition converge in a time of $\Theta(1)$, faster than the transition time of $\Theta(\log \alpha)$. This sharp transition is related to *grokking* (Power et al., 2022), a phenomenon where a model quickly transitions to a generalizing solution long after it interpolates data with poor generalization. Indeed, in the overparameterized case ($\delta < 1$), DLNs first interpolate the data with bad generalization in the lazy phase (due to the ℓ_2 implicit bias induced by the linear dynamics (Bartlett et al., 2021), which favors dense solutions) and then transition to a sparse, generalizing solution in the rich phase. This sudden transition is caused by different implicit biases in the lazy and rich phases, an explanation for grokking given by Lyu et al. (2024) and Kumar et al. (2024).

4.3 Small Initialization Limit $\alpha \to 0$

As with the large initialization case, the dynamics exhibit two dynamical regimes: *search phase* for $t = O(1)$ and *descent phase* for $t = \Theta(\log(1/\alpha))$. These names are adopted from Arous et al. (2021), which established analogous two-stage dynamics for online SGD learning in high-dimensional inference.

Search Phase: $t = O(1)$**.** We introduce the rescaled parameter $W(t) := w(t)/\alpha^2$. Assuming that $|w(t)| \ll |w^*|$, the dynamics (7) is approximated as

$$\frac{\mathrm{d}}{\mathrm{d}t} W(t) = w^* \sqrt{W(t)^2 + \mathrm{e}^{-2\lambda t}} - \lambda W(t)\,, \qquad (10)$$

with an explicit solution

$$W(t) = \frac{\mathrm{sign}(w^*)}{2}(1 - \mathrm{e}^{-2|w^*|t})\mathrm{e}^{(|w^*|-\lambda)t}\,. \qquad (11)$$

The behavior of this solution depends on the relative scales of w^* and λ. Again, let $\Delta := |w^*| - \lambda$. For large t, the solution behaves as $|W(t)| \approx (1/2)\mathrm{e}^{\Delta t}$. When $\Delta < 0$, $W(t)$ converges to zero; when $\Delta > 0$, $|W(t)|$ grows exponentially.

Since changes in $w(t)$ are small (of $O(\alpha^2)$), the loss does not change appreciably, and hence we observe a plateau at the beginning of training.

In this dynamical regime, the algorithm searches and identifies entries of $\boldsymbol{w}$ to be activated and suppresses others. Specifically, entries with $|w^*|$ smaller than the threshold λ are suppressed, and those above the threshold grow. Similar dynamics hold for general δ, but with a different definition of Δ; see Appendix B.2.

Descent Phase: $t = \Theta(\log(1/\alpha))$**.** Paths with $\Delta > 0$ exhibit a transition to the second dynamical regime. This occurs when $|w(t)|$ and $|w^*|$ become of the same order. Equating $|w(t)| = |\alpha^2 W(t)| \approx (\alpha^2/2)\mathrm{e}^{\Delta t}$ to $|w^*|$, we obtain $t \approx t_c := 2\log(1/\alpha)/\Delta$ as the transition time.

Setting $\tau := t - t_c$, the dynamics after the transition are given as follows.

$$\frac{\mathrm{d}}{\mathrm{d}\tau} w(\tau) = -|w(\tau)|(w(\tau) - w^*) - \lambda w(\tau)\,. \qquad (12)$$

This is the same equation as Equation (9) and thus behaves similarly.

An important difference from the rich phase for a large initialization is that the shifted time τ is defined dif-

ferently. In the large initialization case, the transition time $t_c = 2\log(\alpha)/\lambda$ is common for all paths, and the dynamics (9) proceed on the same timescale for all paths. In contrast, in the small initialization case, the transition time $t_c = 2\log(1/\alpha)/\Delta$ is different for each path. Since the dynamics (12) proceed in time $O(1)$, which is much faster than the transition timescale of $\Theta(\log(1/\alpha))$, activated paths (paths that have transitioned to the descent phase) converge quickly to their fixed points. Thus, training in this descent phase proceeds via *incremental learning*, successive activations of entries of $\boldsymbol{w}(t)$ (Berthier, 2023; Pesme and Flammarion, 2023). The timescale of $\Theta(\log(1/\alpha))$ in this regime is checked against numerical simulations in Figure 3c.

5 LONG-TIME BEHAVIOR

We analyze long-time behaviors of the DMFT equation (4) and establish a trade-off between generalization performance and optimization speed. Our results are summarized in Table 1.

Table 1: Summary of long-time behaviors under different regularization λ and the aspect ratio δ. When $\lambda = 0$ and $\delta < 1$, decreasing α results in a trade-off: the generalization performance of the fixed point improves (lower test error), but convergence to the fixed point slows down.

λ	δ	**Fixed Point**	**Convergence Rate**
> 0	Any	ℓ_1-regularized	Sub-Exponential
$= 0$	> 1	Ridgeless	Exponential (*slower* in $\alpha \searrow$)
	< 1	Minimum Norm (*better* in $\alpha \searrow$)	

5.1 Fixed Points and the Benefit of Small Initialization

We analyze the solution obtained by the gradient flow (3) through a fixed point analysis of the DMFT equation (4). In the following result, we show that the solution can be characterized as a minimizer of a certain estimation problem. Note that our result is stated as a *result* and not as a *theorem*, due to the non-rigorous derivation of the DMFT equation (4) and derivation of the fixed point.

Result 1 (Fixed point of gradient flow)**.** *Let $\boldsymbol{w}(\infty) \in \mathbb{R}^d$ be the fixed point of the gradient flow. Let $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ be the solution of a minimization problem as follows.*

Case (i): $\lambda > 0$**.** *ℓ_1-regularized linear regression:*

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{d}\|\boldsymbol{w}\|_1 . \quad (13)$$

Case (ii): $\lambda = 0$**,** $\delta > 1$**.** *Ridgeless linear regression:*

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 . \quad (14)$$

Case (iii): $\lambda = 0$**,** $\delta < 1$**.** *Minimum norm interpolation:*

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} J_\alpha(\boldsymbol{w}) \text{ subject to } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} , \quad (15)$$

with a norm $J_\alpha(\boldsymbol{w}) = \alpha^2 \sum_{i=1}^d J(w_i/\alpha^2)$ with $J(x) = x\sinh^{-1}(x) - \sqrt{1+x^2} + 1$.

As $d \to \infty$, the joint empirical distributions of the entries of $(\boldsymbol{w}(\infty), \boldsymbol{w}^*)$ and of $(\hat{\boldsymbol{w}}, \boldsymbol{w}^*)$ approach the same limiting distribution characterized by the fixed point of the DMFT equation (4) shown in Appendix C.2.

Case (i) is intuitive: ℓ_2 regularization on $(\boldsymbol{u}, \boldsymbol{v})$ translates into ℓ_1 regularization on $\boldsymbol{w} = (\boldsymbol{u}^2 - \boldsymbol{v}^2)/2$. Case (ii) is also natural, since in the underparameterized case ($\delta > 1$), there exists a unique minimizer of the loss (2) almost surely as $d \to \infty$ with the minimizer given by the ridgeless solution.

In Case (iii), there are multiple minimizers of the loss because of overparameterization ($\delta < 1$), and the implicit bias of the algorithm plays a role in selecting a solution among them. Result 1 indicates that the gradient flow selects the solution that minimizes a norm J_α dependent on the initialization α. Properties of this norm are discussed in detail in Woodworth et al. (2020). As $\alpha \to \infty$, J_α approximately behaves as the ℓ_2 norm, resulting in the same implicit bias as linear models. As $\alpha \to 0$, J_α is approximately proportional to the ℓ_1 norm, which exhibits a stronger bias toward sparse solutions. Thus, in the case of a sparse target, smaller initialization yields better final performance, as illustrated in Figure 2a.

Our result for Case (iii) is derived under a more restricted setting than Woodworth et al. (2020, Theorem 1), which holds for any dimension d and any data distributions, yet provides several advantages. First, our result allows for a precise prediction of performances in high dimensions as a solution to a system of equations. Second, our alternative derivation based on DMFT enhances our toolkit for studying implicit biases and has the potential to tackle problems that their method does not apply to.

Sketch of Derivation. Our derivation of Result 1 proceeds as follows; see Appendix C.2 for details.

1. We obtain a system of equations that the fixed point of the DMFT equation (4) satisfies to derive the limiting distribution of the entries of $\boldsymbol{w}(\infty)$.
2. We obtain a characterization of the solutions $\hat{\boldsymbol{w}}$ of the minimization problems given in Result 1 in the high-dimensional limit using *approximate message passing* (AMP) (Donoho et al., 2009; Feng et al., 2022), and show that these two characterizations match in each case.

5.2 Convergence Rates and Their Trade-off with Generalization

Next, we analyze convergence rates of the loss $L(t)$.

Result 2 (Average-case convergence rate of gradient flow)**.** *The paths $w(t)$ converge exponentially with different rates for each path.*

Regularized ($\lambda > 0$). *There are paths with arbitrarily slow rates, and the convergence of the loss $L(t)$ is subexponential.*

Unregularized ($\lambda = 0$). *The loss $L(t)$ converges exponentially as $|L(t) - L(\infty)| = \exp(-2\gamma t + o(1))$, where the exponent $\gamma > 0$ depends on α, δ, σ^2, and P_* and can be computed as a solution of a nonlinear equation (C.65). Furthermore, γ is monotonically increasing with respect to α.*

To the best of our knowledge, this provides the first theoretical characterization of the average-case convergence rate of gradient flow for DLNs in high dimensions and its monotonicity with the initialization scale. Result 2 indicates that the convergence is slower for a smaller initialization α, as shown in Figure 2b. Together with Result 1 Case (iii), it establishes a *trade-off between generalization performance and the convergence speed*. Note that the trade-off is only meaningful in overparameterized settings where multiple solutions exist. Although previous works already discuss that small initialization implies initialization near a saddle point of the loss and hence leads to slow escape from the initial saddle (Woodworth et al., 2020), our result is concerned with the long-time behavior and shows that the slow dynamics persists in the entire dynamics with a quantitative characterization of the convergence rate.

We note that a similar phenomenon is observed in a different setting by Pesme et al. (2021), who studied SGD dynamics of DLNs and found that slower training leads to sparser solutions. This suggests a general principle in DLNs: *better solutions are harder to find.*

With non-zero regularization $\lambda > 0$, the convergence is subexponential, and we do not show a monotonicity result with respect to the initialization scale α. However, for small but nonzero regularization $\lambda > 0$, the transient dynamics still resemble the unregularized case $\lambda = 0$ for a significant period of time. Since regularization affects the dynamics only after time $t \sim \lambda^{-1}$, for $t \ll \lambda^{-1}$, the dynamics behave similar to the unregularized model, and hence the smaller initialization still leads to a slower dynamics (until time $t \sim \lambda^{-1}$).

Sketch of Derivation. We derive Result 2 as follows; see Appendix C.3 for details.

1. We linearize the DMFT equation (4) around the fixed point.
2. We employ the *Laplace transform* to analyze the linearized dynamics and find singularities of the Laplace transforms to derive the convergence rate.

6 RIGOROUS THEORY

While the DMFT equation (4) is derived heuristically, we can rigorously justify it for a closely related model: *truncated diagonal linear networks*. We define truncated (two-layer) DLNs as follows:

$$f(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{w}^\mathsf{T} \boldsymbol{x}\,, \ \boldsymbol{w} = \frac{1}{2}(\eta_M(\boldsymbol{u}^2) - \eta_M(\boldsymbol{v}^2))\,, \quad (16)$$

where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $\eta_M \colon \mathbb{R} \to \mathbb{R}$ is a smooth Lipschitz function with $\eta_M(x) = x$ for $|x| \le M$ and $\eta_M(x) = 0$ for $|x| \ge M + 1$, for $M > 0$. This entry-wise truncation ensures that f_M and its gradients are Lipschitz continuous in $\boldsymbol{u}$ and $\boldsymbol{v}$, which makes the model more amenable to rigorous treatment. When the entries of $\boldsymbol{u}$ and $\boldsymbol{v}$ remain within $[-M, M]$, η_M acts as the identity; hence, for large enough M, the truncated DLN closely approximates the original model (1).

We consider gradient flow training of truncated DLNs as described in Section 2.2. To characterize its behavior, we extend the theory of Celentano et al. (2021), which rigorously establishes DMFT characterization of gradient flow for a class of models that includes generalized linear models and narrow two-layer neural networks, but not truncated DLNs. We show that the empirical distribution of the entries of $\boldsymbol{w}(t)$ for truncated DLNs is asymptotically equivalent to the distribution of the unique solution of a DMFT equation.

Theorem (Informal version of Corollary 5)**.** *Assume that the entries $\boldsymbol{X} = (x_{ij})_{i\in[n], j\in[d]}$ are independent and satisfy $\mathbb{E}\, x_{ij} = 0, \mathbb{E}\, x_{ij}^2 = 1/d, \|x_{ij}\|_{\psi_2} \le C/\sqrt{d}$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm and $C > 0$ is a constant. For any $T > 0$, there exists a unique solution $(w(t))_{t=0}^T$ of the DMFT equation (D.9), and we have*

$$\frac{1}{d}\sum_{i=1}^{d} \delta_{(w_i(t))_{t=0}^T, w_i^*} \xrightarrow{W_2} \mathsf{P}((w(t))_{t=0}^T, w^*)\,, \quad (17)$$

almost surely as $n, d \to \infty$, *where* $\xrightarrow{W_2}$ *denotes convergence in the Wasserstein-2 distance and* $\mathsf{P}((w(t))_{t=0}^{T}, w^*)$ *denotes the joint law of the process* $(w(t))_{t=0}^{T}$ *and the random variable* w^*.

For full statements and proofs, see Appendix D.

Note that the distribution of entries of $\boldsymbol{X}$ is not restricted to the Gaussian distribution, and our result is *universal* with respect to the input distribution.

7 NUMERICAL EXPERIMENTS

The code to reproduce the numerical experiments is available at https://github.com/sotanishy/dmft-dln.

Simulations with Gaussian Data. To validate our theoretical results, we have conducted numerical simulations with Gaussian data, with $\boldsymbol{x}_\mu$ sampled from the Gaussian distribution $\mathsf{N}(0, \mathbf{I}_d/d)$ with $d = 500$. We take the target distribution P_* to be a Bernoulli distribution with $\mathbb{P}\{w^* = 1\} = 0.1$. Gradient flow is discretized with step size $\eta = 0.1$. To obtain the fixed point, we run the dynamics up to $t = 1000$. Results in Figures 2 and 3 show remarkable agreement with theoretical predictions.

Simulations with Real-World Data. To check the universality of our results, we have tested on real-world data. We use a gene expression dataset (Ellrott, 2013; Fiorini, 2016) and take a random subset of $n = 100$ samples and $d = 200$ features ($\delta = 0.5$). Results for fixed points are shown in Figure 2a and show excellent agreement with the theoretical prediction. Results on convergence rates are shown in Figure 2c. Although they deviate from the theoretical prediction (see discussions in Appendix E), qualitative aspects, in particular the monotonicity of the convergence rates with respect to initialization, are well captured by our theory.

Further experiments and discussions are available in Appendix E.

8 DISCUSSION

It would be interesting to explore whether the trade-off between generalization and optimization, specifically the idea that *'better solutions are harder to find'*, which we discussed in Section 5, can be established and extended to general neural networks.

In addition, extending our DMFT analysis to other architectures (such as deep linear networks, nonlinear networks, and transformers) and algorithms (such as SGD) would be a promising future direction. DMFT can handle a wide range of complex architectures in a common formalism (Bordelon and Pehlevan, 2022; Bordelon et al., 2024b; Bordelon and Pehlevan, 2025; Montanari and Urbani, 2025), and transferring the insights developed in this work to these architectures would be a fruitful avenue. Furthermore, prior studies suggest that different optimizers induce distinct implicit biases compared to gradient flow (see Section 1.1), and a deeper theoretical understanding is of great interest.

Acknowledgements

Sota Nishiyama was supported by WINGS-FMSP at the University of Tokyo. Masaaki Imaizumi was supported by JSPS KAKENHI (24K02904), JST FOREST (JPMJFR216I), and JST BOOST (JPMJBY24A9).

References

E. Agoritsas, G. Biroli, P. Urbani, and F. Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018. doi: 10.1088/1751-8121/aaa68d.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

G. B. Arous, R. Gheissari, and A. Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

S. Azulay, E. Moroshko, M. S. Nacson, B. E. Woodworth, N. Srebro, A. Globerson, and D. Soudry. On the implicit bias of initialization shape: beyond infinitesimal mirror descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 468–477, 2021.

P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I*. Springer New York, 1999. doi: 10.1007/978-1-4757-3069-2.

R. Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023.

B. Bordelon and C. Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 32240–32256, 2022.

B. Bordelon and C. Pehlevan. Deep linear network training dynamics from random initialization: data, width, depth, and hyperparameter transfer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 4968–4997, 2025.

B. Bordelon, A. Atanasov, and C. Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 4345–4382, 2024a.

B. Bordelon, H. Chaudhry, and C. Pehlevan. Infinite limits of multi-head transformer dynamics. In *Advances in Neural Information Processing Systems*, volume 37, pages 35824–35878, 2024b. doi: 10.52202/079017-1130.

M. Celentano, C. Cheng, and A. Montanari. The high-dimensional asymptotics of first order methods with random data. arXiv:2112.07572, 2021.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

G. Clara, S. Langer, and J. Schmidt-Hieber. Training diagonal linear networks with stochastic sharpness-aware minimization. arXiv:2503.11891, 2025.

A. Crisanti, H. Horner, and H.-J. Sommers. The spherical p-spin interaction spin-glass model: the dynamics. *Zeitschrift für Physik B Condensed Matter*, 92 (2):257–271, 1993.

L. F. Cugliandolo. Recent applications of dynamical mean-field methods. *Annual Review of Condensed Matter Physics*, 15(1):177–213, 2024. doi: 10.1146/annurev-conmatphys-040721-022848.

L. F. Cugliandolo and J. Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173–176, 1993. doi: 10.1103/PhysRevLett.71.173.

D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45): 18914–18919, 2009. doi: 10.1073/pnas.0909892106.

K. Ellrott. TCGA Pancancer, 2013. Synapse. doi:10.7303/SYN300013.

M. Even, S. Pesme, S. Gunasekar, and N. Flammarion. (S)GD over diagonal linear networks: implicit bias, large stepsizes and edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pages 29406–29448, 2023.

Z. Fan, J. Ko, B. Loureiro, Y. M. Lu, and Y. Shen. Dynamical mean-field analysis of adaptive Langevin diffusions: propagation-of-chaos and convergence of the linear response. arXiv:2504.15556, 2025.

O. Y. Feng, R. Venkataramanan, C. Rush, and R. J. Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends in Machine Learning*, 15(4):335–536, 2022. doi: 10.1561/2200000092.

S. Fiorini. Gene Expression Cancer RNA-Seq, 2016. UCI Machine Learning Repository. doi:10.24432/C5R88H.

C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024. doi: 10.1137/23M1594388.

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

J. Z. HaoChen, C. Wei, J. Lee, and T. Ma. Shape matters: understanding the implicit bias of the noise covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 2315–2357, 2021.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50 (2), 2022. doi: 10.1214/21-AOS2133.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. arXiv:2001.08361, 2020.

T. Kumar, B. Bordelon, S. J. Gershman, and C. Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, 2024.

J. Li, T. Nguyen, C. Hegde, and K. W. Wong. Implicit sparse regularization: the impact of depth and early stopping. In *Advances in Neural Information Processing Systems*, volume 34, pages 28298–28309, 2021.

K. Lyu, J. Jin, Z. Li, S. S. Du, J. D. Lee, and W. Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, 2024.

P. C. Martin, E. D. Siggia, and H. A. Rose. Statistical dynamics of classical systems. *Physical Review A*, 8 (1):423–437, 1973. doi: 10.1103/PhysRevA.8.423.

S. Mei and A. Montanari. The generalization error of random features regression: precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008.

F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 9540–9550, 2020.

A. Montanari and P. Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. In *Advances in Neural Information Processing Systems*, 2025.

E. Moroshko, B. E. Woodworth, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry. Implicit bias in deep linear classification: initialization scale vs training accuracy. In *Advances in Neural Information Processing Systems*, volume 33, pages 22182–22193, 2020.

M. S. Nacson, K. Ravichandran, N. Srebro, and D. Soudry. Implicit bias of the step size in linear diagonal neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 16270–16295, 2022.

B. Neyshabur. *Implicit Regularization in Deep Learning*. PhD thesis, Toyota Technological Institute at Chicago, 2017.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: on the role of implicit regularization in deep learning. In *International Conference on Learning Representations, Workshop Track Proceedings*, 2015.

H. Papazov, S. Pesme, and N. Flammarion. Leveraging continuous time to understand momentum when training diagonal linear networks. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 3556–3564, 2024.

S. Pesme and N. Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 7475–7505, 2023.

S. Pesme, L. Pillaud-Vivien, and N. Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pages 29218–29230, 2021.

A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: generalization beyond overfitting on small algorithmic datasets. arXiv:2201.02177, 2022.

S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory*, 62(12):7464–7474, 2016. doi: 10.1109/TIT.2016.2619365.

S. Sarao Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani, and L. Zdeborová. Marvels and pitfalls of the Langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1): 011057, 2020. doi: 10.1103/PhysRevX.10.011057.

H. Sompolinsky and A. Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5): 359–362, 1981. doi: 10.1103/PhysRevLett.47.359.

H. Sompolinsky and A. Zippelius. Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Physical Review B*, 25(11):6860–6875, 1982. doi: 10.1103/PhysRevB.25.6860.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

T. Vaskevicius, V. Kanade, and P. Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

T. Wang, X. Zhong, and Z. Fan. Universality of approximate message passing algorithms and tensor networks. *The Annals of Applied Probability*, 34(4), 2024. doi: 10.1214/24-AAP2056.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3635–3673, 2020.

L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. doi: 10.1080/00018732.2016.1211393.

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. See Section 2.2.]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes. See Sections 2.2 and 6.]

(b) Complete proofs of all theoretical results. [Yes. For results in Sections 4 and 5, we provide heuristic derivations in Appendices B and C. For results in Section 6, we provide rigorous proofs in Appendix D.]

(c) Clear explanations of any assumptions. [Yes. See Sections 2.2 and 6.]

3. For all figures 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. See Section 7 and Appendix E.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific 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). [Yes. See Appendix E.]

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. [Yes (Ellrott, 2013; Fiorini, 2016).]

(b) The license information of the assets, if applicable. [Yes]

(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 identifiable 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]

Supplementary Materials

Contents

A HEURISTIC DERIVATION OF THE DMFT EQUATION

A.1 Derivation of the DMFT Equation Using Path Integrals

We heuristically derive the DMFT equation (4) using the *path integral* approach in statistical physics. Specifically, we base our derivation on the Martin–Siggia–Rose–De Dominicis–Janssen (MSRDJ) formalism (Martin et al., 1973). The derivation proceeds by expressing the dynamics in a path integral form and using the saddle-point method in the $d \to \infty$ limit to obtain self-consistent equations. Similar computations can be found, for example, in Agoritsas et al. (2018); Sarao Mannelli et al. (2020); Mignacco et al. (2020); Bordelon and Pehlevan (2022); Montanari and Urbani (2025).

The gradient flow dynamics (3) for our setup is

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(t) = -\frac{1}{2}\left(\boldsymbol{u}(t)\odot\frac{1}{\delta}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}(\boldsymbol{w}(t)-\boldsymbol{w}^*)-\boldsymbol{\xi})+\lambda\boldsymbol{u}(t)\right), \tag{A.1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{v}(t) = -\frac{1}{2}\left(-\boldsymbol{v}(t)\odot\frac{1}{\delta}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}(\boldsymbol{w}(t)-\boldsymbol{w}^*)-\boldsymbol{\xi})+\lambda\boldsymbol{v}(t)\right), \tag{A.2}$$

where $\boldsymbol{\xi} \coloneqq (\xi_1,\dots,\xi_n)^\mathsf{T} \in \mathbb{R}^n$ is a noise vector.

Defining fields $\boldsymbol{f}(t) \in \mathbb{R}^n$ and $\boldsymbol{g}(t) \in \mathbb{R}^d$ for $t \geq 0$ as

$$\boldsymbol{f}(t) \coloneqq \boldsymbol{X}(\boldsymbol{w}(t)-\boldsymbol{w}^*), \quad \boldsymbol{g}(t) \coloneqq \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}(\boldsymbol{f}(t)-\boldsymbol{\xi}), \tag{A.3}$$

the dynamics can be expressed as

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(t) = -\frac{1}{2}(\boldsymbol{u}(t)\odot\boldsymbol{g}(t)+\lambda\boldsymbol{u}(t)), \quad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{v}(t) = -\frac{1}{2}(-\boldsymbol{v}(t)\odot\boldsymbol{g}(t)+\lambda\boldsymbol{v}(t)). \tag{A.4}$$

We define a dynamical partition function Z as

$$Z \coloneqq \int \mathrm{D}\boldsymbol{f}\,\mathrm{D}\boldsymbol{g}\,\mathrm{D}\boldsymbol{u}\,\mathrm{D}\boldsymbol{v}\,\mathrm{D}\boldsymbol{w}\;\delta(\boldsymbol{f}(t)-\boldsymbol{X}(\boldsymbol{w}(t)-\boldsymbol{w}^*))\,\delta(\delta\boldsymbol{g}(t)-\boldsymbol{X}^\mathsf{T}(\boldsymbol{f}(t)-\boldsymbol{\xi})), \tag{A.5}$$

where δ is the Dirac delta function (an upright δ is used to distinguish it from δ), and the path measures $\mathrm{D}\boldsymbol{f},\mathrm{D}\boldsymbol{g},\mathrm{D}\boldsymbol{u},\mathrm{D}\boldsymbol{v},\mathrm{D}\boldsymbol{w}$ are implicitly defined with constraints (A.4) and $\boldsymbol{w}(t) = (\boldsymbol{u}(t)^2-\boldsymbol{v}(t)^2)/2$.

Using the Fourier representation of the delta function $\delta(x) = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\mathrm{d}\hat{x}/(2\pi\mathrm{i})\,\mathrm{e}^{\hat{x}x}$ (here and in the following, $\hat{\cdot}$ denotes conjugate variables), we compute the dataset-averaged partition function $\mathbb{E}\,Z$ as

$$\begin{aligned}\mathbb{E}\,Z &\propto \mathbb{E}\int \mathrm{D}\boldsymbol{f}\,\mathrm{D}\hat{\boldsymbol{f}}\,\mathrm{D}\boldsymbol{g}\,\mathrm{D}\hat{\boldsymbol{g}}\,\mathrm{D}\boldsymbol{u}\,\mathrm{D}\boldsymbol{v}\,\mathrm{D}\boldsymbol{w}\;\exp\left(\int\mathrm{d}t\,\hat{\boldsymbol{f}}(t)^\mathsf{T}(\boldsymbol{f}(t)-\boldsymbol{X}(\boldsymbol{w}(t)-\boldsymbol{w}^*))+\int\mathrm{d}t\,\hat{\boldsymbol{g}}(t)^\mathsf{T}(\delta\boldsymbol{g}(t)-\boldsymbol{X}^\mathsf{T}(\boldsymbol{f}(t)-\boldsymbol{\xi}))\right)\\ &= \mathbb{E}_{\boldsymbol{\xi}}\int \mathrm{D}\boldsymbol{f}\,\mathrm{D}\hat{\boldsymbol{f}}\,\mathrm{D}\boldsymbol{g}\,\mathrm{D}\hat{\boldsymbol{g}}\,\mathrm{D}\boldsymbol{u}\,\mathrm{D}\boldsymbol{v}\,\mathrm{D}\boldsymbol{w}\;\exp\left(\int\mathrm{d}t\,\hat{\boldsymbol{f}}(t)^\mathsf{T}\boldsymbol{f}(t)+\delta\int\mathrm{d}t\,\hat{\boldsymbol{g}}(t)^\mathsf{T}\boldsymbol{g}(t)\right)\mathbb{E}_{\boldsymbol{X}}\,\mathrm{e}^{A},\end{aligned} \tag{A.6}$$

where $\mathbb{E}_{\boldsymbol{X}},\mathbb{E}_{\boldsymbol{\xi}}$ denote expectations over $\boldsymbol{X}$, $\boldsymbol{\xi}$, respectively, and

$$A \coloneqq -\int\mathrm{d}t\,\hat{\boldsymbol{f}}(t)^\mathsf{T}\boldsymbol{X}(\boldsymbol{w}(t)-\boldsymbol{w}^*)-\int\mathrm{d}t\,(\boldsymbol{f}(t)-\boldsymbol{\xi})^\mathsf{T}\boldsymbol{X}\hat{\boldsymbol{g}}(t). \tag{A.7}$$

We calculate the expectation $\mathbb{E}_{\boldsymbol{X}}\,\mathrm{e}^{A}$ using a Gaussian integration over $\boldsymbol{X}$ as follows.

$$\begin{aligned}\log\mathbb{E}_{\boldsymbol{X}}\,\mathrm{e}^{A} &= \log\int\mathrm{d}\boldsymbol{X}\;\exp\left(-\frac{d}{2}\operatorname{tr}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})-\int\mathrm{d}t\,\hat{\boldsymbol{f}}(t)^\mathsf{T}\boldsymbol{X}(\boldsymbol{w}(t)-\boldsymbol{w}^*)-\int\mathrm{d}t\,(\boldsymbol{f}(t)-\boldsymbol{\xi})^\mathsf{T}\boldsymbol{X}\hat{\boldsymbol{g}}(t)\right)+\text{const}\\ &= \frac{1}{2d}\int\mathrm{d}t\,\mathrm{d}t'\,\hat{\boldsymbol{f}}(t)^\mathsf{T}\hat{\boldsymbol{f}}(t')(\boldsymbol{w}(t)-\boldsymbol{w}^*)^\mathsf{T}(\boldsymbol{w}(t')-\boldsymbol{w}^*)+\frac{1}{2d}\int\mathrm{d}t\,\mathrm{d}t'\,\hat{\boldsymbol{g}}(t)^\mathsf{T}\hat{\boldsymbol{g}}(t')(\boldsymbol{f}(t)-\boldsymbol{\xi})^\mathsf{T}(\boldsymbol{f}(t')-\boldsymbol{\xi})\\ &\quad+\frac{1}{d}\int\mathrm{d}t\,\mathrm{d}t'\,(\boldsymbol{f}(t)-\boldsymbol{\xi})^\mathsf{T}\hat{\boldsymbol{f}}(t')\hat{\boldsymbol{g}}(t)^\mathsf{T}(\boldsymbol{w}(t')-\boldsymbol{w}^*)+\text{const}.\end{aligned} \tag{A.8}$$

We introduce order parameters as follows.

$$C_w(t,t') \coloneqq \frac{1}{d}(\boldsymbol{w}(t)-\boldsymbol{w}^*)^\mathsf{T}(\boldsymbol{w}(t')-\boldsymbol{w}^*), \quad R_w(t,t') \coloneqq \frac{1}{d}(\boldsymbol{w}(t)-\boldsymbol{w}^*)^\mathsf{T}\hat{\boldsymbol{g}}(t'), \tag{A.9}$$

$$C_f(t,t') \coloneqq \frac{1}{\delta d}(\boldsymbol{f}(t)-\boldsymbol{\xi})^\mathsf{T}(\boldsymbol{f}(t')-\boldsymbol{\xi}), \qquad R_f(t,t') \coloneqq \frac{1}{\delta d}(\boldsymbol{f}(t)-\boldsymbol{\xi})^\mathsf{T}\hat{\boldsymbol{f}}(t'). \tag{A.10}$$

Using these order parameters, we have

$$\log\mathbb{E}_{\boldsymbol{X}}\,\mathrm{e}^{A} = \frac{1}{2}\int\mathrm{d}t\,\mathrm{d}t'\,C_w(t,t')\hat{\boldsymbol{f}}(t)^\mathsf{T}\hat{\boldsymbol{f}}(t')+\frac{\delta}{2}\int\mathrm{d}t\,\mathrm{d}t'\,C_f(t,t')\hat{\boldsymbol{g}}(t)^\mathsf{T}\hat{\boldsymbol{g}}(t')+\delta d\int\mathrm{d}t\,\mathrm{d}t'\,R_f(t,t')R_w(t',t)+\text{const}. \tag{A.11}$$

Inserting the definitions of the order parameters into Equation (A.6) using the delta function as $\delta(dC_w(t,t') - (\boldsymbol{w}(t) - \boldsymbol{w}^*)^\mathsf{T}(\boldsymbol{w}(t') - \boldsymbol{w}^*))$ and using the Fourier representation of the delta function, we get

$$\mathbb{E}\, Z \propto \int \mathrm{D}C_w\, \mathrm{D}\hat{C}_w\, \mathrm{D}C_f\, \mathrm{D}\hat{C}_f\, \mathrm{D}R_w\, \mathrm{D}\hat{R}_w\, \mathrm{D}R_f\, \mathrm{D}\hat{R}_f\, \mathrm{e}^{d\Phi}\,, \tag{A.12}$$

where the action Φ is defined as

$$\begin{aligned}\Phi \coloneqq &- \int \mathrm{d}t\, \mathrm{d}t' \left(\hat{C}_w(t,t')C_w(t,t') + \delta\hat{C}_f(t,t')C_f(t,t') + \hat{R}_w(t,t')R_w(t,t') + \delta\hat{R}_f(t,t')R_f(t,t') - \delta R_f(t,t')R_w(t',t)\right) \\ &+ \log Z_w + \delta \log Z_f\,, \end{aligned} \tag{A.13}$$

$$Z_w \coloneqq \int \mathrm{D}g\, \mathrm{D}\hat{g}\, \mathrm{D}u\, \mathrm{D}v\, \mathrm{D}w\, \mathrm{e}^{\Phi_w}\,, \tag{A.14}$$

$$Z_f \coloneqq \int \mathrm{D}f\, \mathrm{D}\hat{f}\, \mathrm{e}^{\Phi_f}\,, \tag{A.15}$$

$$\begin{aligned}\Phi_w \coloneqq &\delta \int \mathrm{d}t\, \hat{g}(t)g(t) + \frac{\delta}{2}\int \mathrm{d}t\, \mathrm{d}t'\, C_f(t,t')\hat{g}(t)\hat{g}(t') \\ &+ \int \mathrm{d}t\, \mathrm{d}t' \left(\hat{C}_w(t,t')(w(t) - w^*)(w(t') - w^*) + \hat{R}_w(t,t')(w(t) - w^*)\hat{g}(t')\right), \end{aligned} \tag{A.16}$$

$$\begin{aligned}\Phi_f \coloneqq &\int \mathrm{d}t\, \hat{f}(t)f(t) + \frac{1}{2}\int \mathrm{d}t\, \mathrm{d}t'\, C_w(t,t')\hat{f}(t)\hat{f}(t') \\ &+ \int \mathrm{d}t\, \mathrm{d}t' \left(\hat{C}_f(t,t')(f(t) - \xi)(f(t') - \xi) + \hat{R}_f(t,t')(f(t) - \xi)\hat{f}(t')\right). \end{aligned} \tag{A.17}$$

Here, the paths f, g, u, v, w are now one-dimensional, and the dimensionality of the problem has been effectively reduced from d to one.

In the $d \to \infty$ limit, we evaluate the integral (A.12) using the saddle-point method. In the following, $\mathbb{E}$ denotes an expectation over measures e^{Φ_w}/Z_w and e^{Φ_f}/Z_f. Taking derivatives of the action Φ with respect to the order parameters and setting them to zero, we get

$$\frac{\partial\Phi}{\partial\hat{C}_w(t,t')} = -C_w(t,t') + \mathbb{E}[(w(t) - w^*)(w(t') - w^*)] = 0\,, \quad \frac{\partial\Phi}{\partial\hat{R}_w(t,t')} = -R_w(t,t') + \mathbb{E}[(w(t) - w^*)\hat{g}(t')] = 0\,, \tag{A.18}$$

$$\frac{\partial\Phi}{\partial\hat{C}_f(t,t')} = -\delta C_f(t,t') + \delta\,\mathbb{E}[(f(t) - \xi)(f(t') - \xi)] = 0\,, \quad \frac{\partial\Phi}{\partial\hat{R}_f(t,t')} = -\delta R_f(t,t') + \delta\,\mathbb{E}[(f(t) - \xi)\hat{f}(t')] = 0\,, \tag{A.19}$$

and

$$\frac{\partial\Phi}{\partial C_w(t,t')} = -\hat{C}_w(t,t') + \frac{\delta}{2}\,\mathbb{E}[\hat{f}(t)\hat{f}(t')] = 0\,, \quad \frac{\partial\Phi}{\partial R_w(t,t')} = -\hat{R}_w(t,t') + \delta R_f(t',t) = 0\,, \tag{A.20}$$

$$\frac{\partial\Phi}{\partial C_f(t,t')} = -\delta\hat{C}_f(t,t') + \frac{\delta}{2}\,\mathbb{E}[\hat{g}(t)\hat{g}(t')] = 0\,, \quad \frac{\partial\Phi}{\partial R_f(t,t')} = -\delta\hat{R}_f(t,t') + \delta R_w(t',t) = 0\,. \tag{A.21}$$

We obtain

$$\hat{R}_w(t,t') = \delta R_f(t',t)\,, \quad \hat{R}_f(t,t') = R_w(t',t)\,. \tag{A.22}$$

Moreover, we can show the causality of the response functions, i.e.,

$$R_w(t',t) = R_f(t',t) = 0 \tag{A.23}$$

for $t' > t$.

To obtain effective dynamics, we rewrite the effective path measures e^{Φ_w} and e^{Φ_f}. We have

$$\int \mathrm{D}\hat{g}\, \mathrm{e}^{\Phi_w} = \int \mathrm{D}\hat{g}\, \exp\left(\delta \int \mathrm{d}t\, \hat{g}(t)g(t) + \frac{\delta}{2}\int \mathrm{d}t\, \mathrm{d}t'\, C_f(t,t')\hat{g}(t)\hat{g}(t') + \delta \int \mathrm{d}t\, \mathrm{d}t'\, R_f(t,t')\hat{g}(t)(w(t') - w^*)\right)$$

$$
\begin{aligned}
&\times \exp\left(\int \mathrm{d}t\,\mathrm{d}t'\,\hat{C}_w(t,t')(w(t)-w^*)(w(t')-w^*)\right)\\
&\propto \mathbb{E}_{z_g\sim\mathsf{GP}(0,\delta C_f)}\int \mathrm{D}\hat{g}\,\exp\left(\delta\int \mathrm{d}t\,\hat{g}(t)\left(g(t)-\frac{z_g(t)}{\delta}+\int \mathrm{d}t'\,R_f(t,t')(w(t')-w^*)\right)\right)\\
&\times \exp\left(\int \mathrm{d}t\,\mathrm{d}t'\,\hat{C}_w(t,t')(w(t)-w^*)(w(t')-w^*)\right)\\
&\propto \mathbb{E}_{z_g\sim\mathsf{GP}(0,\delta C_f)}\,\delta\left(g(t)-\frac{z_g(t)}{\delta}+\int_0^t \mathrm{d}t'\,R_f(t,t')(w(t')-w^*)\right)\\
&\times \exp\left(\int \mathrm{d}t\,\mathrm{d}t'\,\hat{C}_w(t,t')(w(t)-w^*)(w(t')-w^*)\right),
\end{aligned}
\tag{A.24}
$$

where in the second line we used the *Hubbard–Stratonovich transformation*

$$
\exp\left(\frac{1}{2}\int \mathrm{d}t\,\mathrm{d}t'A(t,t')x(t)x(t')\right)\propto\int \mathrm{D}z\,\exp\left(-\frac{1}{2}\int \mathrm{d}t\,\mathrm{d}t'A^{-1}(t,t')z(t)z(t')-\int \mathrm{d}t\,x(t)z(t)\right) \tag{A.25}
$$

$$
\propto \mathbb{E}_{z\sim\mathsf{GP}(0,A)}\exp\left(-\int \mathrm{d}t\,x(t)z(t)\right). \tag{A.26}
$$

This result indicates that the effective path $g(t)$ satisfies the following equation.

$$
g(t)=\frac{z_g(t)}{\delta}-\int_0^t \mathrm{d}t'\,R_f(t,t')(w(t')-w^*)\,. \tag{A.27}
$$

Similarly, for e^{Φ_f}, we get

$$
\begin{aligned}
\int \mathrm{D}\hat{f}\,\mathrm{e}^{\Phi_f}\propto\mathbb{E}_{z_f\sim\mathsf{GP}(0,C_w)}\,\delta\left(f(t)-z_f(t)+\int_0^t \mathrm{d}t'\,R_w(t,t')(f(t')-\xi)\right)\\
\times\exp\left(\int \mathrm{d}t\,\mathrm{d}t'\,\hat{C}_f(t,t')(f(t)-\xi)(f(t')-\xi)\right),
\end{aligned}
\tag{A.28}
$$

and thus the effective path $f(t)$ satisfies

$$
f(t)=z_f(t)-\int_0^t \mathrm{d}t'\,R_w(t,t')(f(t')-\xi)\,. \tag{A.29}
$$

Finally, we compute $R_w(t,t')$. We insert a source term into our effective action as

$$
Z_w[J]\coloneqq\int \mathrm{D}g\,\mathrm{D}\hat{g}\,\mathrm{D}u\,\mathrm{D}v\,\mathrm{D}w\,\mathrm{e}^{\Phi_w[J]}\,,\quad \Phi_w[J]\coloneqq\Phi_w+\int_0^t \mathrm{d}t\,J(t)\hat{g}(t)\,, \tag{A.30}
$$

and compute $R_w(t,t')$ as

$$
R_w(t,t')=\mathbb{E}[(w(t)-w^*)\hat{g}(t')]=\lim_{J\to0}\frac{\partial}{\partial J(t')}\,\mathbb{E}_J[w(t)-w^*]\,, \tag{A.31}
$$

where $\mathbb{E}_J$ denotes an expectation over the measure $\mathrm{e}^{\Phi_w[J]}/Z_w[J]$. Then, the effective path corresponding to this measure is

$$
g(t)=\frac{z_g(t)-J(t)}{\delta}-\int_0^t \mathrm{d}t'\,R_f(t,t')(w(t')-w^*)\,. \tag{A.32}
$$

Therefore, we have

$$
R_w(t,t')=\lim_{J\to0}\mathbb{E}_J\left[\frac{\partial(w(t)-w^*)}{\partial J(t')}\right]=-\,\mathbb{E}\left[\frac{\partial w(t)}{\partial z_g(t')}\right]. \tag{A.33}
$$

Similarly, we have

$$R_f(t,t') = \mathbb{E}[(f(t)-\xi)\hat{f}(t')] = -\mathbb{E}\left[\frac{\partial f(t)}{\partial z_f(t')}\right]. \tag{A.34}$$

$R_f(t,t')$ consists of a continuous bulk $R_f(t,t')$ for $t > t'$ and a delta spike at $t = t'$ with value -1. Separating these two contributions, Equation (A.27) is written as

$$g(t) = \frac{z_g(t)}{\delta} + w(t) - w^* - \int_0^t \mathrm{d}t'\, R_f(t,t')(w(t') - w^*)\,, \tag{A.35}$$

where we abused the notation and used $R_f(t,t')$ for only the bulk contribution.

Collecting the above expressions, we obtain the following.

$$C_w(t,t') = \mathbb{E}[(w(t)-w^*)(w(t')-w^*)]\,, \quad C_f(t,t') = \mathbb{E}[(f(t)-\xi)(f(t')-\xi)]\,, \tag{A.36a}$$

$$R_w(t,t') = -\mathbb{E}\left[\frac{\partial w(t)}{\partial z_g(t')}\right], \qquad R_f(t,t') = -\mathbb{E}\left[\frac{\partial f(t)}{\partial z_f(t')}\right], \tag{A.36b}$$

$$f(t) = z_f(t) - \int_0^t R_w(t,s)(f(s)-\xi)\,\mathrm{d}s\,, \qquad z_f \sim \mathsf{GP}(0, C_w)\,, \tag{A.36c}$$

$$g(t) = \frac{z_g(t)}{\delta} + w(t) - w^* - \int_0^t R_f(t,s)(w(s)-w^*)\,\mathrm{d}s\,, \qquad z_g \sim \mathsf{GP}(0, \delta C_f)\,, \tag{A.36d}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = -\frac{1}{2}u(t)(g(t)+\lambda)\,, \quad \frac{\mathrm{d}}{\mathrm{d}t}v(t) = -\frac{1}{2}v(t)(-g(t)+\lambda)\,, \quad w(t) = \frac{1}{2}(u(t)^2 - v(t)^2)\,. \tag{A.36e}$$

A.2 Simplifying the DMFT Equation

The stochastic process $f(t)$ can be eliminated. Differentiating both sides of Equation (A.36c) by $z_f(t')$ $(t > t')$ and averaging, we get

$$R_f(t,t') = R_w(t,t') - \int_{t'}^t R_w(t,s)R_f(s,t')\,\mathrm{d}s\,. \tag{A.37}$$

Multiplying both sides of Equation (A.36c) by $(f(t') - \xi)$ and averaging, we get

$$C_f(t,t') = \mathbb{E}[(z_f(t)-\xi)(f(t')-\xi)] - \int_0^t R_w(t,s)\,\mathbb{E}[(f(s)-\xi)(f(t')-\xi)]\,\mathrm{d}s\,. \tag{A.38}$$

By Stein's lemma (Gaussian integration-by-parts formula), we have

$$\begin{aligned}\mathbb{E}[z_f(t)(f(t')-\xi)] &= \int_0^{t'} \mathrm{Cov}(z_f(t), z_f(s))\,\mathbb{E}\left[\frac{\partial(f(t')-\xi)}{\partial z_f(s)}\right]\mathrm{d}s \\ &= C_w(t,t') - \int_0^{t'} C_w(t,s)R_f(t',s)\,\mathrm{d}s\,.\end{aligned} \tag{A.39}$$

By Equation (A.36c), we have

$$\mathbb{E}[\xi f(t)] = \sigma^2 \int_0^t R_w(t,s)\,\mathrm{d}s - \int_0^t R_w(t,s)\,\mathbb{E}[\xi f(s)]\,\mathrm{d}s\,, \tag{A.40}$$

where we used $\mathbb{E}[\xi^2] = \sigma^2$. Comparing this with Equation (A.37), we have

$$\mathbb{E}[\xi f(t)] = \sigma^2 \int_0^t R_f(t,s)\,\mathrm{d}s\,, \tag{A.41}$$

and thus

$$C_f(t,t') = C_w(t,t') + \sigma^2 - \int_0^{t'} R_f(t',s)(C_w(t,s)+\sigma^2)\,\mathrm{d}s - \int_0^{t} R_w(t,s)C_f(s,t')\,\mathrm{d}s\,. \tag{A.42}$$

We can eliminate $u(t)$ and $v(t)$ and express the dynamics in terms of $w(t)$. The product $u(t)v(t)$ obeys a solvable dynamics:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(u(t)v(t)) &= v(t)\frac{\mathrm{d}}{\mathrm{d}t}u(t) + u(t)\frac{\mathrm{d}}{\mathrm{d}t}v(t)\\ &= -\frac{1}{2}u(t)v(t)(g(t)+\lambda) - \frac{1}{2}u(t)v(t)(-g(t)+\lambda)\\ &= -\lambda u(t)v(t)\,,\end{aligned} \tag{A.43}$$

from which we get

$$u(t)v(t) = u(0)v(0)\mathrm{e}^{-\lambda t} = \alpha^2\mathrm{e}^{-\lambda t}\,. \tag{A.44}$$

The dynamics of $w(t)$ is thus

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}w(t) &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(u(t)^2 - v(t)^2)\\ &= u(t)\frac{\mathrm{d}}{\mathrm{d}t}u(t) - v(t)\frac{\mathrm{d}}{\mathrm{d}t}v(t)\\ &= -\frac{1}{2}u(t)^2(g(t)+\lambda) + \frac{1}{2}v(t)^2(-g(t)+\lambda)\\ &= -\frac{1}{2}(u(t)^2+v(t)^2)g(t) - \lambda w(t)\\ &= -\sqrt{w(t)^2+\alpha^4\mathrm{e}^{-2\lambda t}}g(t) - \lambda w(t)\,.\end{aligned} \tag{A.45}$$

In the last line, we used that

$$u(t)^2 + v(t)^2 = \sqrt{(u(t)^2-v(t)^2)^2 + 4u(t)^2v(t)^2} = 2\sqrt{w(t)^2+\alpha^4\mathrm{e}^{-2\lambda t}}\,. \tag{A.46}$$

Combining these results, we obtain the simplified DMFT equation (4) for DLNs.

B DERIVATION OF THE LEARNING TIMESCALES

In this section, we analyze timescale structures of the gradient flow dynamics in $\alpha \to \infty$ and $\alpha \to 0$ limits for general δ. We follow the approach outlined in Section 4, utilizing singular perturbation theory.

B.1 Large Initialization: $\alpha \gg 1$

We show that the dynamics for large α (simulation shown in Figure 4) consists of two phases: *lazy phase* and *rich phase*.

Lazy Phase: $t = O(\alpha^{-2})$. In this short timescale, the factor $\sqrt{w(t)^2+\alpha^4\mathrm{e}^{-2\lambda t}}$ in Equation (4) is approximated as $\alpha^2 + o(1)$. This motivates a time rescaling $\bar{t} \coloneqq \alpha^2 t$. We then make the following ansatz for the DMFT solution on this timescale.

$$C_w(\bar{t}/\alpha^2, \bar{t}'/\alpha^2) = C_w^{\mathrm{lazy}}(\bar{t},\bar{t}') + o(1)\,, \quad C_f(\bar{t}/\alpha^2,\bar{t}'/\alpha^2) = C_f^{\mathrm{lazy}}(\bar{t},\bar{t}') + o(1)\,, \tag{B.1a}$$

$$R_w(\bar{t}/\alpha^2, \bar{t}'/\alpha^2) = \alpha^2 R_w^{\mathrm{lazy}}(\bar{t},\bar{t}') + o(\alpha^2)\,, \quad R_f(\bar{t}/\alpha^2,\bar{t}'/\alpha^2) = \alpha^2 R_f^{\mathrm{lazy}}(\bar{t},\bar{t}') + o(\alpha^2)\,, \tag{B.1b}$$

$$C_w^{\mathrm{lazy}}(\bar{t},\bar{t}') = \mathbb{E}[(w^{\mathrm{lazy}}(\bar{t}) - w^*)(w^{\mathrm{lazy}}(\bar{t}') - w^*)]\,, \quad R_w^{\mathrm{lazy}}(\bar{t},\bar{t}') = -\,\mathbb{E}\left[\frac{\partial w^{\mathrm{lazy}}(\bar{t})}{\partial z^{\mathrm{lazy}}(\bar{t}')}\right], \tag{B.1c}$$

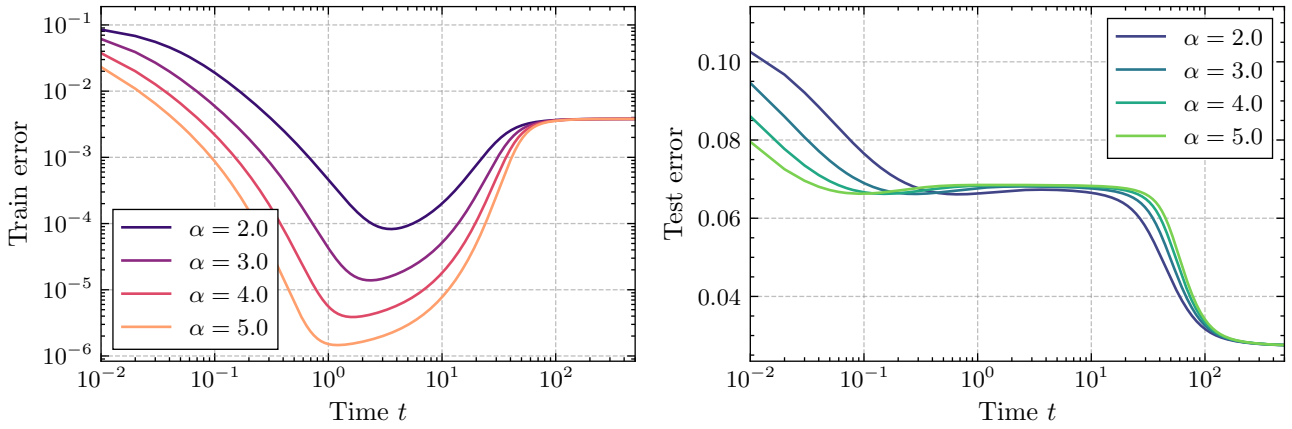


Figure 4: Training and test error dynamics for large α simulated with $d = 200$.

$$w(\bar{t}/\alpha^2) = w^{\text{lazy}}(\bar{t}) + o(1)\,, \quad z^{\text{lazy}} \sim \mathsf{GP}(0, \delta C_f^{\text{lazy}})\,, \tag{B.1d}$$

where $C_f^{\text{lazy}}, R_f^{\text{lazy}}$ are functions independent of α and w^{lazy} is a stochastic process independent of α.

Up to the leading order, the dynamics of $w(t)$ are written as follows.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\bar{t}} w(\bar{t}/\alpha^2) &= -g(\bar{t}/\alpha^2) + O(\alpha^{-2}) \\ &= -\frac{z(\bar{t}/\alpha^2)}{\delta} - (w(\bar{t}/\alpha^2) - w^*) + \int_0^{\bar{t}} R_f^{\text{lazy}}(\bar{t}, \bar{s})(w(\bar{s}/\alpha^2) - w^*)\,\mathrm{d}\bar{s} + o(1)\,.\end{aligned} \tag{B.2}$$

Thus, $w^{\text{lazy}}(\bar{t})$ satisfies a linear integro-differential equation

$$\frac{\mathrm{d}}{\mathrm{d}\bar{t}} w^{\text{lazy}}(\bar{t}) = -\frac{z^{\text{lazy}}(\bar{t})}{\delta} - (w^{\text{lazy}}(\bar{t}) - w^*) + \int_0^{\bar{t}} R_f^{\text{lazy}}(\bar{t}, \bar{s})(w^{\text{lazy}}(\bar{s}) - w^*)\,\mathrm{d}\bar{s}\,, \quad z^{\text{lazy}} \sim \mathsf{GP}(0, \delta C_f^{\text{lazy}})\,. \tag{B.3}$$

The stochastic process $w^{\text{lazy}}(\bar{t})$ can be eliminated along the same lines as in Appendix A.2 using the linearity of its dynamics to yield a closed system for correlation and response functions.

$$\frac{\partial}{\partial \bar{t}} C_w^{\text{lazy}}(\bar{t}, \bar{t}') = -C_w^{\text{lazy}}(\bar{t}, \bar{t}') + \int_0^{\bar{t}'} R_w^{\text{lazy}}(\bar{t}', \bar{s}) C_f^{\text{lazy}}(\bar{t}, \bar{s})\,\mathrm{d}\bar{s} + \int_0^{\bar{t}} R_f^{\text{lazy}}(\bar{t}, \bar{s}) C_w^{\text{lazy}}(\bar{t}', \bar{s})\,\mathrm{d}\bar{s}\,, \tag{B.4a}$$

$$C_f^{\text{lazy}}(\bar{t}, \bar{t}') = C_w^{\text{lazy}}(\bar{t}, \bar{t}') + \sigma^2 - \int_0^{\bar{t}'} R_f^{\text{lazy}}(\bar{t}', \bar{s})(C_w^{\text{lazy}}(\bar{t}, \bar{s}) + \sigma^2)\,\mathrm{d}\bar{s} - \int_0^{\bar{t}} R_w^{\text{lazy}}(\bar{t}, \bar{s}) C_f^{\text{lazy}}(\bar{t}', \bar{s})\,\mathrm{d}\bar{s}\,, \tag{B.4b}$$

$$\frac{\partial}{\partial \bar{t}} R_w^{\text{lazy}}(\bar{t}, \bar{t}') = -R_w^{\text{lazy}}(\bar{t}, \bar{t}') + \int_{\bar{t}'}^{\bar{t}} R_f^{\text{lazy}}(\bar{t}, \bar{s}) R_w^{\text{lazy}}(\bar{s}, \bar{t}')\,\mathrm{d}\bar{s}\,, \tag{B.4c}$$

$$R_f^{\text{lazy}}(\bar{t}, \bar{t}') = R_w^{\text{lazy}}(\bar{t}, \bar{t}') - \int_{\bar{t}'}^{\bar{t}} R_w^{\text{lazy}}(\bar{t}, \bar{s}) R_f^{\text{lazy}}(\bar{s}, \bar{t}')\,\mathrm{d}\bar{s}\,, \tag{B.4d}$$

with boundary conditions $C_w^{\text{lazy}}(\bar{t}, 0) = C_w^{\text{lazy}}(0, \bar{t}) = 0$ and $R_w^{\text{lazy}}(\bar{t}, \bar{t}) = 1/\delta$ for $\bar{t} \geq 0$.

These equations are equivalent to the ones for (ridgeless) linear regression derived in Fan et al. (2025) and Bordelon et al. (2024a). Thus, in this dynamical regime, DLNs behave as linear models. Equation (B.4) can be solved explicitly as follows.

$$R_w^{\text{lazy}}(\bar{t}, \bar{t}') = \frac{1}{\delta} \int \mathrm{e}^{-x(\bar{t} - \bar{t}')}\,\mathrm{d}\mu_{\text{MP}}(x)\,, \tag{B.5a}$$

$$R_f^{\text{lazy}}(\bar{t}, \bar{t}') = \frac{1}{\delta} \int x \mathrm{e}^{-x(\bar{t} - \bar{t}')}\,\mathrm{d}\mu_{\text{MP}}(x)\,, \tag{B.5b}$$

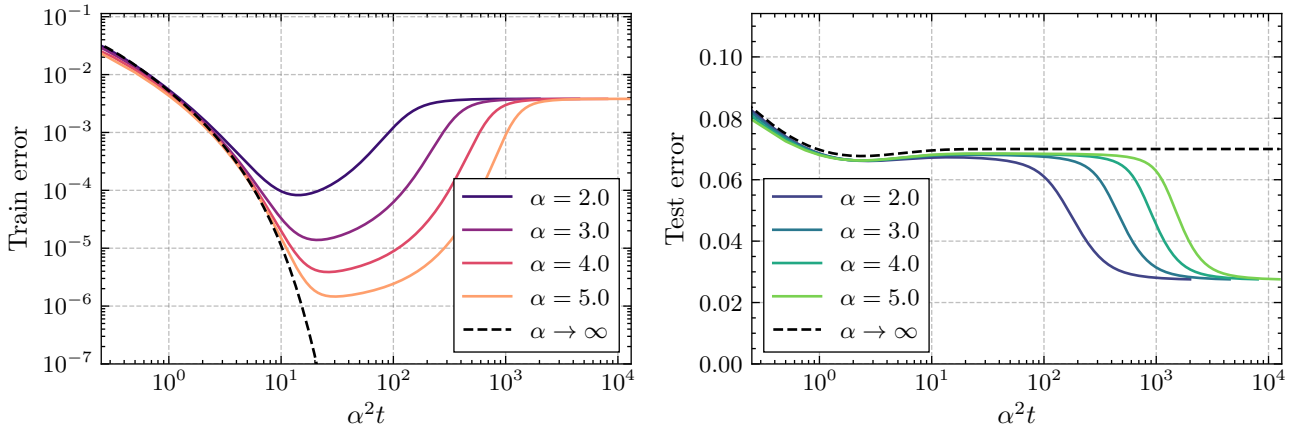


Figure 5: Training and test error dynamics for large α, with time rescaled by α^2. The initial descent of training and test errors collapses onto the limiting solution (B.7)

$$C_w^{\text{lazy}}(\bar{t}, \bar{t}') = \rho^2 \int \mathrm{e}^{-x(\bar{t}+\bar{t}')} \, \mathrm{d}\mu_{\text{MP}}(x) + \frac{\sigma^2}{\delta} \int \frac{1}{x}(1 - \mathrm{e}^{-x\bar{t}})(1 - \mathrm{e}^{-x\bar{t}'}) \, \mathrm{d}\mu_{\text{MP}}(x) \,, \tag{B.5c}$$

$$C_f^{\text{lazy}}(\bar{t}, \bar{t}') = \rho^2 \int x\mathrm{e}^{-x(\bar{t}+\bar{t}')} \, \mathrm{d}\mu_{\text{MP}}(x) + \frac{\sigma^2}{\delta} \int \mathrm{e}^{-x(\bar{t}+\bar{t}')} \, \mathrm{d}\mu_{\text{MP}}(x) + \frac{\delta - 1}{\delta}\sigma^2 \,, \tag{B.5d}$$

where μ_{MP} is the *Marchenko–Pastur law*, the limiting eigenvalue spectrum of a random matrix $\delta^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$, whose density is given explicitly as follows.

$$\mathrm{d}\mu_{\text{MP}}(x) = \frac{\delta\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi x} + (1 - \delta)\,\delta(x)\mathbf{1}_{\delta<1}, \quad \text{where } \lambda_\pm = \left(1 \pm \frac{1}{\sqrt{\delta}}\right)^2 , \tag{B.6}$$

where $\delta(x)$ is the Dirac delta function. Training and test errors are given by

$$E_{\text{train}}^{\text{lazy}}(\bar{t}) = C_f^{\text{lazy}}(\bar{t}, \bar{t}) \,, \quad E_{\text{test}}^{\text{lazy}}(\bar{t}) = C_w^{\text{lazy}}(\bar{t}, \bar{t}) + \sigma^2 \,. \tag{B.7}$$

These solutions are checked against simulations in Figure 5 and show good agreement.

Rich Phase: $t = 2\log(\alpha)/\lambda$**.** When $\lambda > 0$, the ansatz (B.1) breaks down when $w(t)$ and $\alpha^4\mathrm{e}^{-2\lambda t}$ are of the same order, which occurs at $t \approx t_c := 2\log\alpha/\lambda$. Introducing a new time variable as $\bar{t} = t - t_c$, the parameter $w(\bar{t})$ obeys the following new equation.

$$\frac{\mathrm{d}}{\mathrm{d}\bar{t}} w(\bar{t}) = -\sqrt{w(\bar{t})^2 + \mathrm{e}^{-2\lambda\bar{t}}}\, g(\bar{t}) - \lambda w(\bar{t}) \,. \tag{B.8}$$

A fixed point analysis in Appendix C.2 reveals that this equation converges to the ℓ_1-regularized solution in time $O(1)$. These results are checked against simulations in Figures 6 and 7.

B.2 Small Initialization: $\alpha \ll 1$

We show that the dynamics for small α (simulation shown in Figure 8) consists of two phases: *search phase* and *descent phase*.

Search Phase: $t = O(1)$**.** Let $W(t) = w(t)/\alpha^2$. Assuming that $W(t) = O(1)$, the DMFT equation is approximated up to the leading order as follows.

$$C_w(t, t') = \rho^2 + O(\alpha^2) \,, \quad C_f(t, t') = \rho^2 + \sigma^2 + O(\alpha^2) \,, \tag{B.9}$$

$$R_w(t, t') = O(\alpha^2) \,, \qquad R_f(t, t') = O(\alpha^2) \,, \tag{B.10}$$

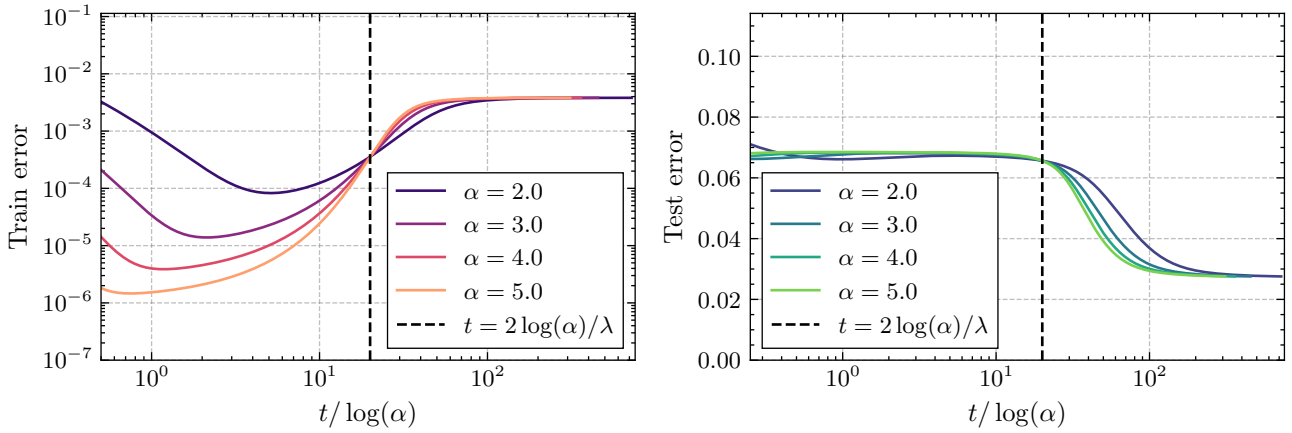


Figure 6: Training and test error dynamics for large α, with time rescaled by $\log(\alpha)$. Transition times to the rich phase collapse to the same value of $t/\log(\alpha) = 2/\lambda$.

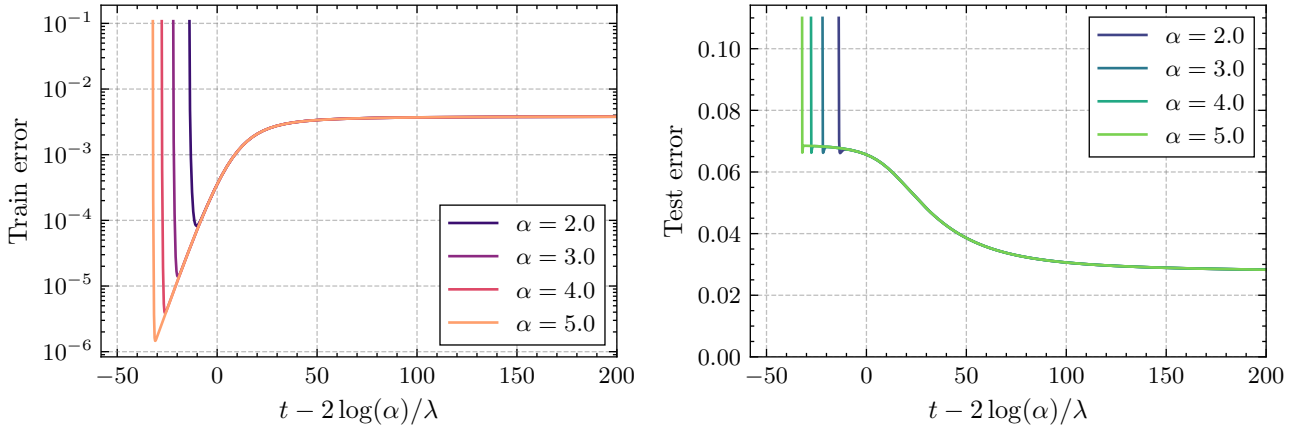


Figure 7: Training and test error dynamics for large α, with time shifted by the transition time $2\log(\alpha)/\lambda$. These curves collapse, indicating that the dynamics after the transition proceed in time $O(1)$.

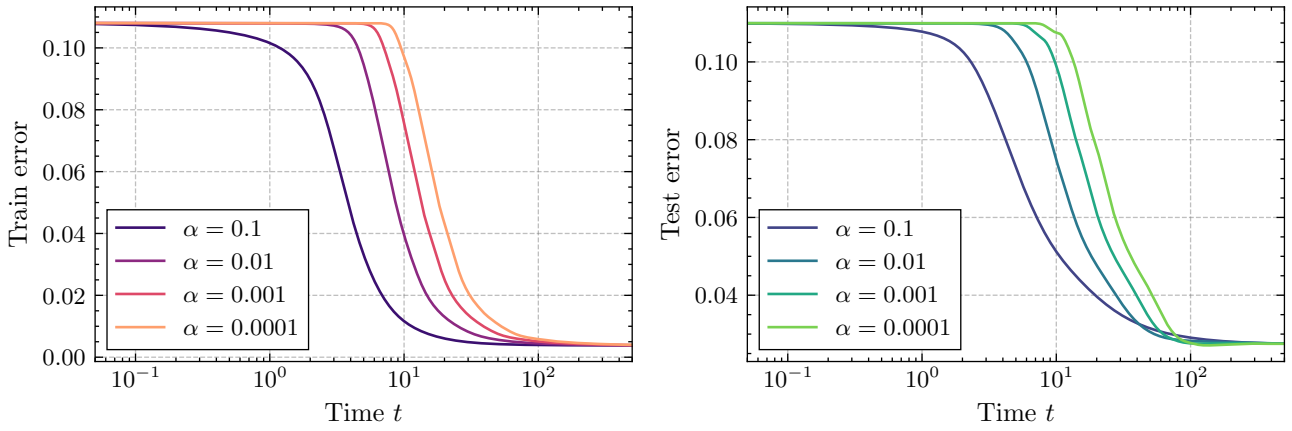


Figure 8: Training and test error dynamics for small α simulated with $d = 200$.

and the dynamics of $W(t)$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}W(t) = (w^* - z(0)/\delta)\sqrt{W(t)^2 + \mathrm{e}^{-2\lambda t}} - \lambda W(t) + o(1)\,, \quad z(0) \sim \mathsf{N}(0, \delta(\rho^2 + \sigma^2))\,. \tag{B.11}$$

This equation can be solved explicitly as

$$W(t) = \frac{\mathrm{sign}(w^* - z(0)/\delta)}{2}(1 - \mathrm{e}^{-2|w^* - z(0)/\delta|t})\mathrm{e}^{(|w^* - z(0)/\delta| - \lambda)t}\,. \tag{B.12}$$

Let $\Delta = |w^* - z(0)/\delta| - \lambda$. For large t, a sample path $W(t)$ behaves as $|W(t)| \approx (1/2)\mathrm{e}^{\Delta t}$. When $\Delta < 0$, $W(t)$ converges exponentially to zero; when $\Delta > 0$, $|W(t)|$ grows exponentially.

The noise term $z(0)/\delta$ captures a finite sample effect, which essentially acts as a noise that obscures the ground truth w^*. It vanishes as $\delta \to \infty$.

Descent Phase: $t = \Theta(\log(1/\alpha))$. As sample paths with $\Delta > 0$ grow, the assumption that $W(t) = O(1)$ breaks down. A transition to the second dynamical regime occurs when $w(t) = \alpha^2 W(t)$ becomes of $O(1)$, which happens at a timescale of $\Theta(\log(1/\alpha))$. This motivates the following rescaling of dynamical variables with $\bar{t} = t/\log(1/\alpha)$.

$$C_w(\log(1/\alpha)\bar{t}, \log(1/\alpha)\bar{t}') = C_w^{\mathrm{desc}}(\bar{t}, \bar{t}')\,, \quad C_f(\log(1/\alpha)\bar{t}, \log(1/\alpha)\bar{t}') = C_f^{\mathrm{desc}}(\bar{t}, \bar{t}')\,, \tag{B.13a}$$

$$R_w(\log(1/\alpha)\bar{t}, \log(1/\alpha)\bar{t}') = \frac{1}{\log(1/\alpha)}R_w^{\mathrm{desc}}(\bar{t}, \bar{t}')\,, \quad R_f(\log(1/\alpha)\bar{t}, \log(1/\alpha)\bar{t}') = \frac{1}{\log(1/\alpha)}R_f^{\mathrm{desc}}(\bar{t}, \bar{t}')\,, \tag{B.13b}$$

$$w(\log(1/\alpha)\bar{t}) = w^{\mathrm{desc}}(\bar{t})\,, \quad g(\log(1/\alpha)\bar{t}) = g^{\mathrm{desc}}(\bar{t})\,. \tag{B.13c}$$

The rescaled DMFT equation is

$$C_w^{\mathrm{desc}}(\bar{t}, \bar{t}') = \mathbb{E}[(w^{\mathrm{desc}}(\bar{t}) - w^*)(w^{\mathrm{desc}}(\bar{t}') - w^*)]\,, \tag{B.14a}$$

$$R_w^{\mathrm{desc}}(\bar{t}, \bar{t}') = -\log(1/\alpha)\,\mathbb{E}\left[\frac{\partial w^{\mathrm{desc}}(\bar{t})}{\partial z^{\mathrm{desc}}(\bar{t}')}\right], \tag{B.14b}$$

$$C_f^{\mathrm{desc}}(\bar{t}, \bar{t}') = C_w^{\mathrm{desc}}(\bar{t}, \bar{t}') + \sigma^2 - \int_0^{\bar{t}'} R_f^{\mathrm{desc}}(\bar{t}', \bar{s})(C_w^{\mathrm{desc}}(\bar{t}, \bar{s}) + \sigma^2)\,\mathrm{d}\bar{s} - \int_0^{\bar{t}} R_w^{\mathrm{desc}}(\bar{t}, \bar{s})C_f^{\mathrm{desc}}(\bar{t}', \bar{s})\,\mathrm{d}\bar{s}\,, \tag{B.14c}$$

$$R_f^{\mathrm{desc}}(\bar{t}, \bar{t}') = R_w^{\mathrm{desc}}(\bar{t}, \bar{t}') - \int_{\bar{t}'}^{\bar{t}} R_w^{\mathrm{desc}}(\bar{t}, \bar{s})R_f^{\mathrm{desc}}(\bar{s}, \bar{t}')\,\mathrm{d}\bar{s}\,, \tag{B.14d}$$

$$g^{\mathrm{desc}}(\bar{t}) = \frac{z^{\mathrm{desc}}(\bar{t})}{\delta} + w^{\mathrm{desc}}(\bar{t}) - w^* - \int_0^{\bar{t}} R_f^{\mathrm{desc}}(\bar{t}, \bar{s})(w^{\mathrm{desc}}(\bar{s}) - w^*)\,\mathrm{d}\bar{s}\,, \quad z^{\mathrm{desc}} \sim \mathsf{GP}(0, \delta C_f^{\mathrm{desc}})\,, \tag{B.14e}$$

$$\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}\bar{t}}w^{\mathrm{desc}}(\bar{t}) = -\sqrt{w^{\mathrm{desc}}(\bar{t})^2 + \alpha^{4+2\lambda\bar{t}}}\,g^{\mathrm{desc}}(\bar{t}) - \lambda w^{\mathrm{desc}}(\bar{t})\,. \tag{B.14f}$$

The time it takes for each path to become active (become of $\Theta(1)$) can be derived as follows. Let $W(\bar{t}) = w^{\mathrm{desc}}(\bar{t})/\alpha^2$. Assuming that $1 \ll W(\bar{t}) \ll \alpha^{-2}$, the dynamics of $W(\bar{t})$ is approximated up to the leading order as

$$\frac{1}{\log(1/\alpha)}\frac{\mathrm{d}}{\mathrm{d}\bar{t}}W(\bar{t}) \approx -|W(\bar{t})|\left(\frac{z^{\mathrm{desc}}(\bar{t})}{\delta} - w^* + \int_0^{\bar{t}} R_f^{\mathrm{desc}}(\bar{t}, \bar{s})w^*\,\mathrm{d}\bar{s}\right) - \lambda W(\bar{t})\,. \tag{B.15}$$

It can be solved as

$$W(\bar{t}) \propto \exp\left(\log(1/\alpha)\int_0^{\bar{t}}\left|\frac{z^{\mathrm{desc}}(\bar{t}')}{\delta} - w^* + \int_0^{\bar{t}'} R_f^{\mathrm{desc}}(\bar{t}', \bar{s})w^*\,\mathrm{d}\bar{s}\right|\mathrm{d}\bar{t}' - \log(1/\alpha)\lambda\bar{t}\right). \tag{B.16}$$

Thus, the time $\bar{t}_c$ at which $w(\bar{t}) = \alpha^2 W(\bar{t})$ becomes of $\Theta(1)$ is given implicitly by

$$\int_0^{\bar{t}_c}\left|\frac{z^{\mathrm{desc}}(\bar{t}')}{\delta} - w^* + \int_0^{\bar{t}'} R_f^{\mathrm{desc}}(\bar{t}', \bar{s})w^*\,\mathrm{d}\bar{s}\right|\mathrm{d}\bar{t}' - \lambda\bar{t}_c = 2\,. \tag{B.17}$$

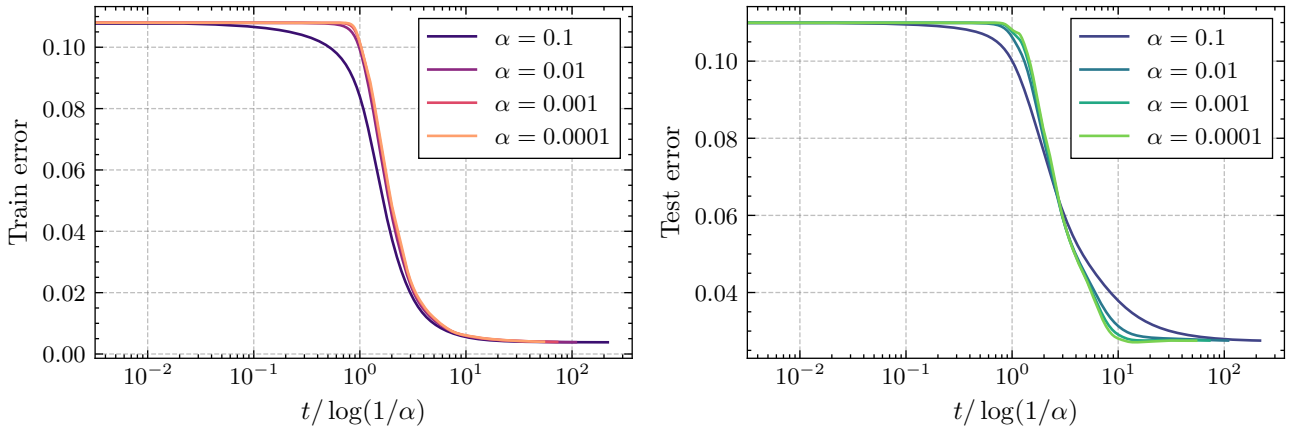


Figure 9: Training and test error dynamics for small α, with time rescaled by $\log(1/\alpha)$. Learning curves collapse, indicating that the descent phase proceeds on the timescale $\Theta(\log(1/\alpha))$.

When $\delta \to \infty$, the left-hand side of the above equation reduces to $(|w^*| - \lambda)\bar{t}_c$, and we thus have $t_c = \log(1/\alpha)\bar{t}_c = 2\log(1/\alpha)/(|w^*| - \lambda)$ for the transition time, as derived in Section 4. As already discussed in the main text, the transition times $\bar{t}_c$ are different for each path, as opposed to the large initialization ($\alpha \gg 1$) case where the transition time $t_c = 2\log(\alpha)/\lambda$ is the same for all paths. We therefore observe *incremental learning* with successive activation of paths.

After the transition, defining a new time variable $\bar{t} = t - \log(1/\alpha)\bar{t}_c$, the new dynamics is

$$\frac{\mathrm{d}}{\mathrm{d}\bar{t}} w(\bar{t}) \approx -|w(\bar{t})| g(\bar{t}) - \lambda w(\bar{t}) \,, \tag{B.18}$$

which behaves similarly to Equation (B.8) for large $\bar{t}$. $w(\bar{t})$ converges in time $O(1)$ to the ℓ_1 regularized solution for $\lambda > 0$ and minimum ℓ_1 norm interpolator for $\lambda = 0$, as described in Section 5.

The timescale is checked against numerical simulations in Figure 9, showing that $\log(1/\alpha)$ is indeed the correct time scaling.

C DERIVATION OF THE LONG-TIME BEHAVIOR

In this section, we derive Results 1 and 2 by analyzing long-time behaviors of the DMFT equation (4).

C.1 Preliminary: Laplace Transform

Throughout this section, we make extensive use of the *Laplace transform*, which is a useful technique for analyzing linear differential equations. Given a function $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, its Laplace transform $\mathcal{L}[f] = \bar{f}$ is defined as

$$\bar{f}(p) \coloneqq \int_0^\infty f(t) \mathrm{e}^{-pt} \, \mathrm{d}t \,, \tag{C.1}$$

for $p \in \mathbb{C}$ with sufficiently large real part for the integral to be convergent.

We state several of its basic properties.

- *Linearity.* For $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ and $a, b \in \mathbb{R}$, we have

$$\mathcal{L}[af + b] = a\mathcal{L}[f] + b \,. \tag{C.2}$$

- *Laplace transforms of derivatives, integrals, and convolutions.* For $f, g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, we have

$$\mathcal{L}[f'(t)](p) = p\bar{f}(p) - f(0) \,, \tag{C.3}$$

$$\mathcal{L}\left[\int_0^t f(s)\,\mathrm{d}s\right](p) = \frac{\bar{f}(p)}{p}\,, \tag{C.4}$$

$$\mathcal{L}\left[\int_0^t f(t-s)g(s)\,\mathrm{d}s\right](p) = \bar{f}(p)\bar{g}(p)\,. \tag{C.5}$$

- *The final value theorem.* For $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, we have

$$\lim_{t\to\infty} f(t) = \lim_{p\to 0} p\bar{f}(p)\,, \tag{C.6}$$

  if all singularities of $\bar{f}$ lie on the left half-plane.

- *Convergence rate.* For $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, let p_c be the singularity of $\bar{f}$ with the largest real part. Then, as $t \to \infty$, $f(t) = \exp(-(\operatorname{Re} p_c)t + o(1))$.

C.2 Fixed Point

C.2.1 Deriving the Fixed-Point Equations

First, we derive equations for the fixed point of the DMFT equation (4). The fixed point can be obtained by taking $t \to \infty$ and setting the time derivative $\mathrm{d}w(t)/\mathrm{d}t$ to zero. We make a *time-translation invariance* (TTI) ansatz to handle integrals with response functions.

Case (i): $\lambda > 0$. We assume that a long-time limit $t \to \infty$ exists with TTI response functions.

- $w(t) \to w$ and $g(t) \to g$ (random, constant for each path),
- $C_w(t,t) \to C_w$ and $C_f(t,t) \to C_f$ (constants),
- $R_w(t,t') \approx R_w(t-t')$ and $R_f(t,t') \approx R_f(t-t')$ with $R_w(t), R_f(t) \to 0$ and $R_w(t), R_f(t)$ are both integrable.

Denote the integrated responses (*susceptibilities*) as

$$\chi_w \coloneqq \int_0^\infty R_w(t)\,\mathrm{d}t\,, \quad \chi_f \coloneqq \int_0^\infty R_f(t)\,\mathrm{d}t\,. \tag{C.7}$$

Using TTI, the DMFT equation for R_f becomes

$$R_f(t) = R_w(t) - \int_0^t R_w(t-s)R_f(s)\,\mathrm{d}s\,. \tag{C.8}$$

Taking the Laplace transform of both sides, we obtain

$$\bar{R}_f(p) = \bar{R}_w(p) - \bar{R}_w(p)\bar{R}_f(p)\,, \tag{C.9}$$

from which we get

$$\bar{R}_f(p) = \frac{\bar{R}_w(p)}{1+\bar{R}_w(p)}\,. \tag{C.10}$$

Since $\bar{R}_w(0) = \chi_w$ and $\bar{R}_f(0) = \chi_f$, we get

$$\chi_f = \frac{\chi_w}{1+\chi_w}\,. \tag{C.11}$$

Similar manipulations for C_f yield

$$C_f = \frac{C_w + \sigma^2}{(1+\chi_w)^2}\,. \tag{C.12}$$

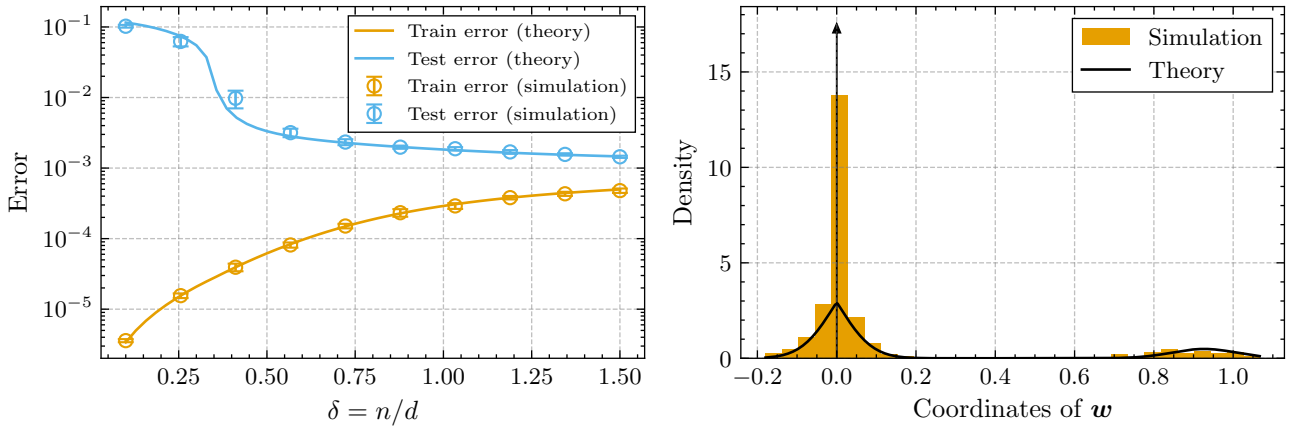


Figure 10: Fixed points for case (i) ($\lambda > 0$). (Left): Fixed point of train and test errors for different δ. Simulations are run 10 times on independent data with $d = 500$, and fixed points are obtained at $t = 1000$. Error bars indicate one standard deviation. (Right): Distribution of coordinates of $\boldsymbol{w}(\infty)$. The arrow indicates the delta spike at $w = 0$ with its height scaled by the bin size.

Next, we derive pathwise fixed points of w and g. As $t \to \infty$, the factor $\mathrm{e}^{-2\lambda t}$ vanishes, and setting $\mathrm{d}w(t)/\mathrm{d}t = 0$ yields

$$0 = -|w|g - \lambda w\,, \quad g = \frac{z}{\delta} + w - w^* - (w - w^*)\chi_f\,, \quad z \sim \mathsf{N}(0, \delta C_f)\,. \tag{C.13}$$

From the first equation, we get $w = 0$ or $g = -\lambda\,\mathrm{sign}(w)$. In the case of $g = -\lambda\,\mathrm{sign}(w)$ we get from the second equation that

$$w + (1 + \chi_w)\lambda\,\mathrm{sign}(w) = w^* - \frac{1 + \chi_w}{\delta} z\,. \tag{C.14}$$

This equation has a solution if and only if $|w| \geq (1 + \chi_w)\lambda$, otherwise we have $w = 0$. These solutions can be expressed using the soft thresholding function as

$$w = \mathrm{ST}\left(w^* - \frac{1 + \chi_w}{\delta} z; (1 + \chi_w)\lambda\right). \tag{C.15}$$

Finally, the response of w to a *constant* input z gives the integrated response χ_w.

$$\chi_w = -\,\mathbb{E}\left[\frac{\partial w}{\partial z}\right] = \frac{1 + \chi_w}{\delta}\partial_x\,\mathrm{ST}\left(w^* - \frac{1 + \chi_w}{\delta} z; (1 + \chi_w)\lambda\right). \tag{C.16}$$

Collecting these results, we obtain the following system of equations for the fixed point.

$$\boxed{\begin{gathered} C_w = \mathbb{E}[(w - w^*)^2]\,, \quad \chi_w = \frac{1 + \chi_w}{\delta}\,\mathbb{E}\left[\partial_x\,\mathrm{ST}\left(w^* - \frac{1 + \chi_w}{\delta} z; (1 + \chi_w)\lambda\right)\right], \quad C_f = \frac{C_w + \sigma^2}{(1 + \chi_w)^2}\,, \\ w = \mathrm{ST}\left(w^* - \frac{1 + \chi_w}{\delta} z; (1 + \chi_w)\lambda\right), \quad z \sim \mathsf{N}(0, \delta C_f)\,. \end{gathered}} \tag{C.17}$$

This result is validated numerically as shown in Figure 10.

We note that the susceptibility χ_w is related to the *train-test gap* because $(1 + \chi_w)^2$ is equal to the ratio between fixed points of training and test errors, $E_{\text{train}} = C_f$ and $E_{\text{test}} = C_w + \sigma^2$.

Case (ii): $\lambda = 0$, $\delta > 1$. Calculation proceeds along the same lines as Case (i). We have the same fixed point equations for χ_f and C_f:

$$\chi_f = \frac{\chi_w}{1 + \chi_w}\,, \quad C_f = \frac{C_w + \sigma^2}{(1 + \chi_w)^2}\,. \tag{C.18}$$

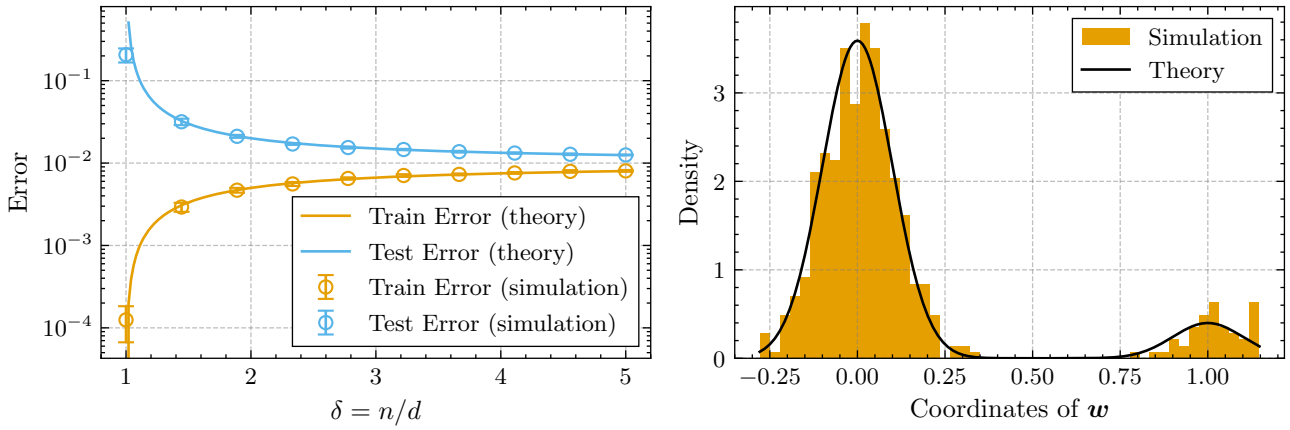


Figure 11: Fixed points for case (ii) ($\lambda = 0$, $\delta > 1$). (Left): Fixed point of train and test errors for different δ. Simulations are run 10 times on independent data with $d = 500$, and fixed points are obtained at $t = 1000$. Error bars indicate one standard deviation. (Right): Distribution of coordinates of $\boldsymbol{w}(\infty)$.

Fixed point conditions for w and g are

$$0 = -\sqrt{w^2 + \alpha^4}\, g\,, \quad g = \frac{z}{\delta} + w - w^* - (w - w^*)\chi_f\,, \quad z \sim \mathsf{N}(0, \delta C_f)\,. \tag{C.19}$$

From the first equation, we have $g = 0$. From the second equation, we get

$$w = w^* - \frac{1 + \chi_w}{\delta} z\,. \tag{C.20}$$

The susceptibility χ_w satisfies

$$\chi_w = -\,\mathbb{E}[\partial_z w] = \frac{1 + \chi_w}{\delta}\,, \quad \text{therefore } \chi_w = \frac{1}{\delta - 1}\,. \tag{C.21}$$

We can simplify the fixed point equation using the explicit form of χ_w to arrive at the following result.

$$\boxed{\begin{gathered} C_w = \frac{\sigma^2}{\delta - 1}\,, \quad \chi_w = \frac{1}{\delta - 1}\,, \quad C_f = \frac{\delta - 1}{\delta}\sigma^2\,, \\ w = w^* - \frac{z}{\delta - 1}\,, \quad z \sim \mathsf{N}(0, \delta C_f)\,. \end{gathered}} \tag{C.22}$$

This result is validated numerically as shown in Figure 11.

If $\sigma^2 > 0$, the test error $C_w + \sigma^2$ diverges as $\delta \to 1$. This is the well-known *double descent* peak (Belkin et al., 2019; Hastie et al., 2022).

Case (iii): $\lambda = 0$, $\delta < 1$. This case requires a more careful argument than the previous cases, since in this case the response function $R_w(\tau)$ does not vanish as $\tau \to \infty$. This corresponds to the fact that the minimum of the loss function is degenerate and that a perturbation to the system will permanently shift the solution.

Assuming that R_w converges to a nonzero constant, its integral χ_w diverges to infinity. By (C.18), we have $\chi_f = 0$ and $C_f = 0$. To obtain fixed points for other variables, we need to know how fast they converge. Thus, we make the following ansatz

- $w(t) \to w$ and $C_w(t, t) \to C_w$,
- $g(t) \to 0$ and integrable,
- $C_f(t, t') \to 0$ and integrable on $\mathbb{R}^2_{\geq 0}$,

- $R_w(t,t') = R_w(t-t')$ and $R_w(t) \to R_w$,
- $R_f(t,t') = R_f(t-t')$ and $1 - \int_0^t R_f(s)\,\mathrm{d}s$ is integrable (with respect to t),

and define

$$\tilde{\chi}_f \coloneqq \int_0^\infty \left(1 - \int_0^t R_f(s)\,\mathrm{d}s\right)\mathrm{d}t\,, \quad \tilde{C}_f \coloneqq \int_0^\infty \int_0^\infty C_f(t,t')\,\mathrm{d}t\,\mathrm{d}t'\,. \tag{C.23}$$

By Equation (C.10), we get

$$\frac{1-\bar{R}_f(p)}{p} = \frac{1}{p(1+\bar{R}_w(p))}\,. \tag{C.24}$$

Since $\lim_{p\to 0}(1-\bar{R}_f(p))/p = \tilde{\chi}_f$ and $\lim_{p\to 0} p\bar{R}_w(p) = R_w$, we have

$$\tilde{\chi}_f = \frac{1}{R_w}\,. \tag{C.25}$$

Similarly, for C_f, we get

$$\tilde{C}_f = \frac{C_w + \sigma^2}{R_w^2}\,. \tag{C.26}$$

Next, we derive the fixed point condition for w. The DMFT equation for $\mathrm{d}w(t)/\mathrm{d}t$ can be transformed as

$$\frac{\mathrm{d}}{\mathrm{d}t}\sinh^{-1}(w(t)/\alpha^2) = -g(t)\,. \tag{C.27}$$

Integrating both sides, we have

$$\sinh^{-1}(w/\alpha^2) = -\int_0^\infty g(t)\,\mathrm{d}t =: -\tilde{g}\,. \tag{C.28}$$

Next, we derive a fixed point condition for g. Integrating both sides of the DMFT equation for g, we get

$$\begin{aligned}\tilde{g} &= \frac{1}{\delta}\int_0^\infty z(t)\,\mathrm{d}t + (w-w^*)\int_0^\infty \left(1 - \int_0^t R_f(t-s)\,\mathrm{d}s\right)\mathrm{d}t \\ &= \frac{1}{\delta}\int_0^\infty z(t)\,\mathrm{d}t + (w-w^*)\tilde{\chi}_f\,.\end{aligned} \tag{C.29}$$

$\int_0^\infty z(t)\,\mathrm{d}t$ follows a Gaussian distribution with mean zero and variance

$$\mathbb{E}\left[\left(\int_0^\infty z(t)\,\mathrm{d}t\right)^2\right] = \int_0^\infty\int_0^\infty \mathbb{E}[z(t)z(t')]\,\mathrm{d}t\,\mathrm{d}t' = \int_0^\infty\int_0^\infty \delta C_f(t,t')\,\mathrm{d}t\,\mathrm{d}t' = \delta\tilde{C}_f\,. \tag{C.30}$$

Thus, we have

$$w + R_w \sinh^{-1}(w/\alpha^2) = w^* - \frac{R_w}{\delta}\tilde{z}\,, \quad \tilde{z} \sim \mathsf{N}(0, \delta\tilde{C}_f)\,. \tag{C.31}$$

Finally, differentiating w with respect to $\tilde{z}$ and taking the expectation gives the response R_w. Differentiating both sides of (C.31) by $\tilde{z}$,

$$\begin{aligned}\partial_{\tilde{z}} w + \frac{R_w \partial_{\tilde{z}} w}{\sqrt{w^2+\alpha^4}} &= -\frac{R_w}{\delta}\,. \\ R_w = -\,\mathbb{E}[\partial_{\tilde{z}} w] &= \frac{R_w}{\delta}\,\mathbb{E}\left[\frac{1}{1 + R_w/\sqrt{w^2+\alpha^4}}\right].\end{aligned} \tag{C.32}$$

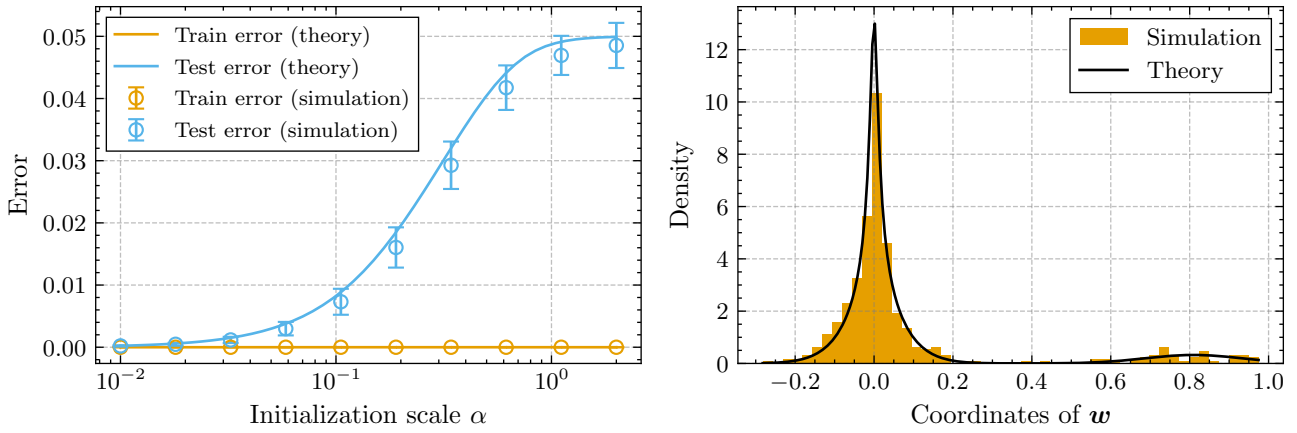


Figure 12: Fixed points for case (iii) ($\lambda = 0$, $\delta < 1$). (Left): Fixed point of train and test errors for different α. Simulations are run 10 times on independent data with $d = 500$, and fixed points are obtained at $t = 1000$. Error bars indicate one standard deviation. (Right): Distribution of coordinates of $\boldsymbol{w}(\infty)$.

Collecting these results, we have

$$\boxed{\begin{aligned} C_w = \mathbb{E}[(w - w^*)^2]\,, \quad 1 = \frac{1}{\delta}\,\mathbb{E}\left[\frac{1}{1 + R_w/\sqrt{w^2+\alpha^4}}\right], \quad \tilde{C}_f = \frac{C_w + \sigma^2}{R_w^2}\,, \\ w = f\left(w^* - \frac{R_w}{\delta}\tilde{z}; \alpha^2, R_w\right), \quad \tilde{z} \sim \mathsf{N}(0, \delta\tilde{C}_f)\,. \end{aligned}} \tag{C.33}$$

Here, $f(x; a, b)$ is the inverse (with respect to x) of a function

$$g(x; a, b) = x + b\sinh^{-1}(x/a)\,. \tag{C.34}$$

This result is validated numerically as shown in Figure 12.

C.2.2 Analyzing the Minimization Problem

We characterize the fixed point distributions (C.17), (C.22), (C.33) as solutions of minimization problems. We consider the following minimization problem.

$$\hat{\boldsymbol{w}} := \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{d}\sum_{i=1}^{d} J(w_i)\,, \tag{C.35}$$

for $\lambda > 0$ and a convex function $J\colon \mathbb{R} \to \mathbb{R}$. For $\lambda > 0$ and strictly convex J, the solution $\hat{\boldsymbol{w}}$ is unique.

We proceed as follows.

1. We characterize the empirical distribution of the entries of the minimizer $\hat{\boldsymbol{w}}$ in the high-dimensional limit $n, d \to \infty$ using *approximate message passing* (AMP). This gives a self-consistent equation for the limiting distribution.
2. We show that, with a specific choice of the norm J, the self-consistent equation becomes equivalent to the fixed point equations for the DMFT equation.

Characterizing the Empirical Distribution of the Minimizer via AMP. AMP is an iterative algorithm for solving high-dimensional statistical estimation tasks. A distinct feature of AMP is that its behavior can be rigorously tracked using a scalar recursion called *state evolution* (SE). For a tutorial on AMP, see Feng et al. (2022).

Let η be the proximal operator defined as

$$\eta(u;t) = \mathrm{prox}_{tJ}(u) \coloneqq \operatorname*{arg\,min}_{x\in\mathbb{R}} \left\{ \frac{1}{2}(x-u)^2 + tJ(x) \right\}. \tag{C.36}$$

We consider the following AMP iteration for $k = 0, 1, \ldots$.

$$\hat{\boldsymbol{r}}^k = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{w}}^k + b_k \hat{\boldsymbol{r}}^{k-1}, \quad \hat{\boldsymbol{w}}^{k+1} = \eta\left(\hat{\boldsymbol{w}}^k + \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\hat{\boldsymbol{r}}^k; t_{k+1} \right), \tag{C.37}$$

initialized with $\hat{\boldsymbol{w}}^0 = \hat{\boldsymbol{r}}^{-1} = \boldsymbol{0}$, together with its state evolution

$$\begin{aligned} \sigma_1^2 &= \frac{\sigma^2 + \mathbb{E}[(W^*)^2]}{\delta}, \quad t_1 = \lambda(1+b_0), \quad b_k = \frac{1}{\delta}\mathbb{E}[\eta'(W^* + \sigma_k G_k; t_k)], \\ \sigma_{k+1}^2 &= \frac{\sigma^2 + \mathbb{E}[(W^* - \eta(W^* + \sigma_k G_k; t_k))^2]}{\delta}, \quad t_{k+1} = \lambda + b_k t_k, \end{aligned} \tag{C.38}$$

where $W^* \sim P_*$, $G_k \sim \mathsf{N}(0,1)$ and $b_0 > 0$.

The *master theorem* (Feng et al., 2022, Theorem 4.2) states that, under some regularity conditions, we have the following for any second-order pseudo-Lipschitz function $\psi\colon \mathbb{R} \to \mathbb{R}$, almost surely as $d \to \infty$.

$$\left| \frac{1}{d}\sum_{i=1}^{d} \psi(w_i^k, w_i^*) - \mathbb{E}[\psi(\eta(W^* + \sigma_k G), W^*)] \right| \to 0, \tag{C.39}$$

where $W^* \sim P_*$ and $G \sim \mathsf{N}(0,1)$. In other words, the joint empirical distribution of the entries of $(\boldsymbol{w}^k, \boldsymbol{w}^*)$ is asymptotically equivalent to the joint distribution of $(\eta(W^* + \sigma_k G), W^*)$.

Furthermore, it can be shown that the AMP iteration (C.37) converges to the minimizer $\hat{\boldsymbol{w}}$ of the minimization problem (C.35) (Rangan et al., 2016, Theorem 1). Thus, the empirical distribution of the entries of $\hat{\boldsymbol{w}}$ can be characterized using the fixed point (b_*, σ_*, t_*) of the SE recursion:

$$b_* = \frac{1}{\delta}\mathbb{E}[\eta'(W^* + \sigma_* G; t_*)], \quad \sigma_*^2 = \frac{\sigma^2 + \mathbb{E}[(W^* - \eta(W^* + \sigma_* G; t_*))^2]}{\delta}, \quad t_* = \lambda + b_* t_*. \tag{C.40}$$

Next, we map the SE fixed point (C.40) to each of the DMFT fixed points.

Case i: $\lambda > 0$. We show that by choosing $J(w) = |w|$, the SE fixed point (C.40) becomes equivalent to the DMFT fixed point (C.17). When $J(w) = |w|$, the proximal operator η is the soft thresholding function $\eta(x;t) = \mathrm{ST}(x;t)$ and the SE fixed point corresponds to the DMFT fixed point with the following mapping.

$$b_* \to \frac{\chi_w}{1+\chi_w}, \quad \sigma_*^2 \to \frac{(1+\chi_w)^2}{\delta^2} \cdot \delta C_f = \frac{C_w + \sigma^2}{\delta}, \quad t_* = \frac{\lambda}{1-b_*} \to (1+\chi_w)\lambda. \tag{C.41}$$

Thus, the fixed point of the gradient flow for DLNs is asymptotically equivalent to the ℓ_1 regularized solution. This is natural since ℓ_2 regularization on $\boldsymbol{u}$ and $\boldsymbol{v}$ are equivalent to ℓ_1 regularization on $\boldsymbol{w} = (\boldsymbol{u}^2 - \boldsymbol{v}^2)/2$.

Case ii: $\lambda = 0$, $\delta > 1$. In this case, the penalty term vanishes since $\lambda = 0$. Then the proximal operator is the identity function $\eta(x;t) = x$ and the SE fixed point (C.40) is solved by

$$b_* = \frac{1}{\delta}, \quad \sigma_*^2 = \frac{\sigma^2}{\delta - 1}, \quad t_* = \frac{\delta}{\delta - 1}\lambda. \tag{C.42}$$

Again, with the same mapping as (C.41), we recover the DMFT fixed point (C.22).

Case iii: $\lambda = 0$, $\delta < 1$. We take $J(w) = w\sinh^{-1}(w/\alpha^2) - \sqrt{w^2 + \alpha^4} + \alpha^2$ and send $\lambda \to 0$. This corresponds to the following constrained minimization problem.

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{d}\sum_{i=1}^{d} J(w_i) \quad \text{subject to} \quad \boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}. \tag{C.43}$$

We assume that the $\lambda \to 0$ limit of the SE fixed point characterizes the solution of the above constrained minimization problem (this amounts to assuming that $\lambda \to 0$ and $d \to \infty$ limits commute).

The proximal operator is

$$\eta(u; t_*) = \arg\min_{x \in \mathbb{R}} \left\{ \frac{1}{2}(x-u)^2 + t_* J(x) \right\}, \tag{C.44}$$

and $x_* \coloneqq \eta(u; t_*)$ satisfies

$$0 = x_* - u + t_* J'(x_*) = x_* - u + t_* \sinh^{-1}(x_*/\alpha^2)\,. \tag{C.45}$$

Then $\eta'(u; t_*) = \partial_u x_*$ satisfies, by the implicit function theorem,

$$0 = \partial_u x_* - 1 + \frac{t_*}{\sqrt{x_*^2 + \alpha^4}} \partial_u x_*\,. \tag{C.46}$$

Thus, the equation for b_* is

$$1 - \frac{\lambda}{t_*} = \frac{1}{\delta}\,\mathbb{E}\left[\frac{1}{1 + t_*/\sqrt{\eta(W^* + \sigma_* G; t_*)^2 + \alpha^4}} \right]. \tag{C.47}$$

Taking $\lambda \to 0$, we recover the DMFT fixed point (C.33) with the following mapping.

$$\sigma_*^2 \to \frac{R_w^2}{\delta^2} \cdot \delta \tilde{C}_f = \frac{C_w + \sigma^2}{\delta}\,, \quad t_* \to R_w\,. \tag{C.48}$$

C.3 Convergence Rate

C.3.1 Regularized Case $\lambda > 0$

We linearize the dynamics around the fixed point. Let $w(t) = w + \Delta w(t)$ and $g(t) = g + \Delta g(t)$ where w and g are the fixed points (C.17). We make the following assumptions and approximations.

- $\Delta w(t)$ and $\Delta g(t)$ are small, and terms of second or higher order can be ignored.
- $w(t)$ converge slower than $\mathrm{e}^{-\lambda t}$ and hence the $\alpha^2 \mathrm{e}^{-2\lambda t}$ term can be ignored compared to $w(t)^2$.
- Response functions are TTI.

For paths with $w = 0$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta w(t) \approx -|\Delta w(t)|(g + \Delta g(t)) - \lambda \Delta w(t) \approx -(\lambda + \operatorname{sign}(\Delta w(t)) g)\Delta w(t) = -(\lambda - |g|)\Delta w(t)\,, \tag{C.49}$$

where in the last equality, we used $\operatorname{sign}(\Delta w(t)) = -\operatorname{sign}(g)$. From the condition $w = 0$, we have

$$|g| = \left| \frac{z}{\delta} - \frac{w^*}{1 + \chi_w} \right| \le \lambda\,, \tag{C.50}$$

and thus $\lambda - |g| \ge 0$. Therefore, paths with $w = 0$ converge to zero as $|w(t)| \sim \mathrm{e}^{-(\lambda - |g|)t}$. This rate is consistent with the assumption that $w(t)$ converges slower than $\mathrm{e}^{-\lambda t}$. This rate further implies that, the closer the observation $w^* - (1+\chi_w)z/\delta$ is to the threshold $(1+\chi_w)\lambda$, the slower the convergence. Since there are paths with $\lambda - |g|$ arbitrarily small (because of the continuous nature of the noise z), the convergence of macroscopic observables (such as training and test errors) is subexponential.

For paths with $w \neq 0$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Delta w(t) &\approx -|w + \Delta w(t)|(g + \Delta g(t)) - \lambda(w + \Delta w(t)) \\ &\approx -(\operatorname{sign}(w) g + \lambda)(w + \Delta w(t)) - |w| \Delta g(t)\end{aligned}$$

$$
\begin{aligned}
&= -|w|\Delta g(t) \\
&= -|w|\left(\frac{\Delta z(t)}{\delta} + \Delta w(t) - \int_0^t R_f(t-s)\Delta w(s)\,\mathrm{d}s\right),
\end{aligned} \tag{C.51}
$$

where in the third line we used $g = -\lambda\,\mathrm{sign}(w)$. Taking the Laplace transform, we get

$$
p\Delta\bar{w}(p) - w(0) = -|w|\left(\frac{\Delta\bar{z}(p)}{\delta} + \Delta\bar{w}(p) - \bar{R}_f(p)\Delta\bar{w}(p)\right), \tag{C.52}
$$

and we get

$$
\Delta\bar{w}(p) = \frac{w(0) - |w|\Delta\bar{z}(p)/\delta}{p + |w|(1-\bar{R}_f(p))}\,. \tag{C.53}
$$

The long-time behavior of $\Delta w(t)$ is controlled by the singularity p_c of $\Delta\bar{w}(p)$ with the largest real part. It is the point at which $p_c + |w|(1-\bar{R}_f(p_c)) = 0$. For small $|w|$, we have $p \approx 0$, and we can approximate $\bar{R}_f(p) \approx \bar{R}_f(0) = \chi_f$. Thus, the asymptotic behavior is approximately $|\Delta w(t)| \sim \exp(-|w|(1-\chi_f)t) = \exp(-\frac{|w|}{1+\chi_w}t)$. Again, when the magnitude of the observation $w^* - (1+\chi_w)z/\delta$ is closer to the threshold λ, $|w|$ is small and thus the convergence is slow. There are paths with arbitrarily small $|w|$ and hence the convergence of macroscopic observables is subexponential.

C.3.2 Unregularized Case $\lambda = 0$

The linearized dynamics around the fixed point are

$$
\frac{\mathrm{d}}{\mathrm{d}t}\Delta w(t) \approx -\sqrt{w^2+\alpha^4}\Delta g(t) = -\sqrt{w^2+\alpha^4}\left(\frac{\Delta z(t)}{\delta} + \Delta w(t) - \int_0^t R_f(t-s)\Delta w(s)\,\mathrm{d}s\right). \tag{C.54}
$$

Taking the Laplace transform, we find that the singularity of $\Delta\bar{w}(p)$ satisfies $p_c + \sqrt{w^2+\alpha^4}(1-\bar{R}_f(p_c)) = 0$, implying slower convergence for smaller w. However, unlike the case with $\lambda > 0$, we can still have $p_c \neq 0$ even for $w = 0$, which implies exponential convergence. We can explicitly determine the rate, as we discuss below.

Convergence Rates of Response Functions. We derive the convergence rates of response functions R_w and R_f. By TTI, we have

$$
\frac{\mathrm{d}}{\mathrm{d}\tau}\hat{R}_w(\tau) = \sqrt{w^2+\alpha^4}\left(-\hat{R}_w(\tau) + \int_0^\tau R_f(\tau-\sigma)\hat{R}_w(\sigma)\,\mathrm{d}\sigma\right), \tag{C.55}
$$

$$
R_f(\tau) = -\int_0^\tau R_w(\tau-\sigma)R_f(\sigma)\,\mathrm{d}\sigma + R_w(\tau)\,, \tag{C.56}
$$

where $\hat{R}_w(t,t') \coloneqq -\partial w(t)/\partial z(t')$ (without expectations) and assumed TTI for $\hat{R}_w$ as well. Taking the Laplace transform, we obtain the following equations.

$$
\bar{R}_w(p) = \frac{1+\bar{R}_w(p)}{\delta}\,\mathbb{E}\left[\frac{\sqrt{w^2+\alpha^4}}{p(1+\bar{R}_w(p)) + \sqrt{w^2+\alpha^4}}\right], \quad \bar{R}_f(p) = \frac{\bar{R}_w(p)}{1+\bar{R}_w(p)}\,. \tag{C.57}
$$

First, we consider the convergence rate of R_w. To this end, we compute the rightmost singularity p_c of $\bar{R}'_w(p) \coloneqq p\bar{R}_w(p)$. The factor p is introduced to eliminate the pole at $p = 0$, which exists when $\delta < 1$ due to the nonzero fixed point of $R_w(t)$. Note that this factor does not alter the location of other singularities, and the convergence rate is determined from the singularity with the largest negative real part.

By Equation (C.57), $\bar{R}'_w(p)$ satisfies

$$
\bar{R}'_w(p) = \frac{p+\bar{R}'_w(p)}{\delta}\,\mathbb{E}\left[\frac{\sqrt{w^2+\alpha^4}}{p+\bar{R}'_w(p) + \sqrt{w^2+\alpha^4}}\right]. \tag{C.58}
$$

The rightmost singularity p_c can be found by defining

$$h(R', p) \coloneqq R' - \frac{p + R'}{\delta} \mathbb{E}\left[\frac{\sqrt{w^2 + \alpha^4}}{p + R' + \sqrt{w^2 + \alpha^4}}\right], \tag{C.59}$$

and solving the following system of equations:

$$h(R', p_c) = R' - \frac{p_c + R'}{\delta} \mathbb{E}\left[\frac{\sqrt{w^2 + \alpha^4}}{p_c + R' + \sqrt{w^2 + \alpha^4}}\right] = 0\,, \tag{C.60}$$

$$\partial_{R'} h(R', p_c) = 1 - \frac{1}{\delta} \mathbb{E}\left[\left(\frac{\sqrt{w^2 + \alpha^4}}{p_c + R' + \sqrt{w^2 + \alpha^4}}\right)^2\right] = 0\,. \tag{C.61}$$

Let $u = R' + p_c$. By the second equation, u satisfies the following equation:

$$1 - \frac{1}{\delta} \mathbb{E}\left[\left(\frac{\sqrt{w^2 + \alpha^4}}{u + \sqrt{w^2 + \alpha^4}}\right)^2\right] = 0\,. \tag{C.62}$$

Let u_* be the solution of the above equation. By solving the equation $h(R', p_c) = 0$ for R', we have

$$R' = \frac{u_* A}{\delta}\,, \quad A \coloneqq \mathbb{E}\left[\frac{\sqrt{w^2 + \alpha^4}}{u_* + \sqrt{w^2 + \alpha^4}}\right]. \tag{C.63}$$

We thus have

$$p_c = u_* - R' = \frac{\delta - A}{\delta} u_*\,. \tag{C.64}$$

It follows that the convergence rate is $R_w(t) = \exp(-\gamma t + o(1))$ where $\gamma = -p_c$.

Next, we consider the convergence rate of R_f. By Equation (C.57), assuming that $1 + \bar{R}_w(p)$ is never zero for p with $\operatorname{Re} p \geq \operatorname{Re} p_c$, the rightmost singularity of $\bar{R}_f$ is simply that of $\bar{R}_w$, and R_f has the same convergence rate as R_w.

In summary, the convergence rates γ of R_w and R_f can be obtained by solving the following system.

$$\boxed{\mathbb{E}\left[\left(\frac{\sqrt{w^2 + \alpha^4}}{u + \sqrt{w^2 + \alpha^4}}\right)^2\right] = \delta\,, \quad A = \mathbb{E}\left[\frac{\sqrt{w^2 + \alpha^4}}{u + \sqrt{w^2 + \alpha^4}}\right], \quad \gamma = \frac{A - \delta}{\delta} u\,.} \tag{C.65}$$

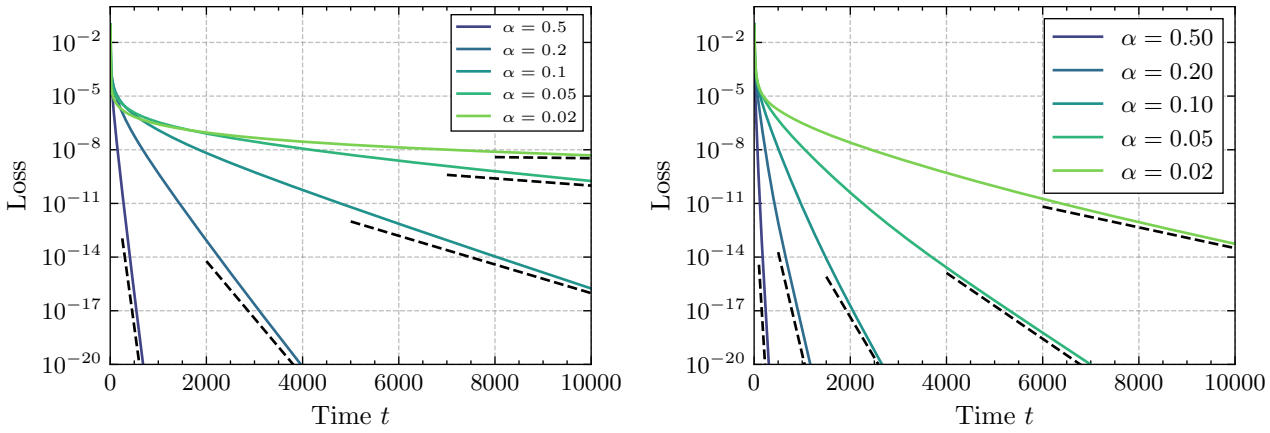
Convergence Rates of Correlation Functions. Next, we derive the rates for correlation functions C_w and C_f. Since these are bivariate functions, we use the bivariate Laplace transform, defined for $f\colon \mathbb{R}^2_{\geq 0} \to \mathbb{R}$ and $p, q \in \mathbb{C}$ with sufficiently large real parts as

$$\bar{f}(p, q) = \int_0^\infty \int_0^\infty f(t, s) \mathrm{e}^{-pt - qs} \,\mathrm{d}t \,\mathrm{d}s\,. \tag{C.66}$$

Taking the Laplace transform of the equations for C_w and C_f, we obtain

$$\bar{C}_w(p, q) = \mathbb{E}[(\bar{w}(p) - w^*/p)(\bar{w}(q) - w^*/q)]\,, \quad \bar{C}_f(p, q) = \frac{\bar{C}_w(p, q) + \sigma^2/(pq)}{(1 + \bar{R}_w(p))(1 + \bar{R}_w(q))}\,. \tag{C.67}$$

By the second equation, $pq\bar{C}_f(p, q)$ has singularities at $p = -\gamma$ and $q = -\gamma$. This implies that $C_f(t, s)$ behaves as $|C_f(t, s) - C_f(\infty, \infty)| = \exp(-\gamma(t + s) + o(1))$ and thus $|L(t) - L(\infty)| = |C_f(t, t) - C_f(\infty, \infty)| = \exp(-2\gamma t + o(1))$. This result is validated against numerical simulations in Figure 13.



(a) Convergence of the loss for $\delta = 2$.

(b) Convergence of the loss for $\delta = 0.5$.

Figure 13: Convergence of the loss for $\delta = 2 > 1$ and $\delta = 0.5 < 1$. As the initialization scale α decreases, the convergence rate γ becomes slower. Simulations are run for $d = 500$ and show good agreement with the theoretical rate.

Nonnegativity and Monotonicity of the Convergence Rate. We prove basic properties of the convergence rate γ.

Proposition C.1. *Let γ be the solution of Equation (C.65). When $\delta \neq 1$, we have $\gamma > 0$ and $\mathrm{d}\gamma/\mathrm{d}\alpha > 0$.*

Proof. Let $x(\alpha) = \sqrt{w^2 + \alpha^4}$, $y(u, \alpha) = x/(u + x)$ and define A, B as follows.

$$A(u, \alpha) = \mathbb{E}[y]\,, \quad B(u, \alpha) = \mathbb{E}[y^2]\,. \tag{C.68}$$

Since $u = u(\alpha)$ satisfies $B(u, \alpha) = \delta$, we have

$$\delta\gamma = u(A - \delta) = u(A - B) = u\,\mathbb{E}[y(1 - y)] = u^2\,\mathbb{E}\left[\frac{x}{(u + x)^2}\right] \tag{C.69}$$

and thus

$$\gamma = \frac{u^2}{\delta} C\,, \quad C(u, \alpha) \coloneqq \mathbb{E}\left[\frac{x}{(u + x)^2}\right]. \tag{C.70}$$

Since $x > 0$ and $u \neq 0$ when $\delta \neq 1$, it follows that $\gamma > 0$.

Using $\partial_\alpha x = 2\alpha^3/x$, we have

$$\partial_u B = -2\,\mathbb{E}\left[\frac{x^3}{(u + x)^3}\right], \quad \partial_\alpha B = 4\alpha^3 u\,\mathbb{E}\left[\frac{1}{(u + x)^3}\right]. \tag{C.71}$$

By the implicit function theorem, we get

$$\frac{\mathrm{d}u}{\mathrm{d}\alpha} = -\frac{\partial_\alpha B}{\partial_u B} = \frac{2\alpha^3 u\,\mathbb{E}[(u + x)^{-3}]}{\mathbb{E}[x^2(u + x)^{-3}]}\,. \tag{C.72}$$

For C, we have

$$\partial_u C = -2\,\mathbb{E}\left[\frac{x}{(u + x)^3}\right], \quad \partial_\alpha C = 2\alpha^3\,\mathbb{E}\left[\frac{1}{x(u + x)^2} - \frac{2}{(u + x)^3}\right]. \tag{C.73}$$

Thus, we have

$$\frac{\mathrm{d}\gamma}{\mathrm{d}\alpha} = \frac{1}{\delta}\left(2u\frac{\mathrm{d}u}{\mathrm{d}\alpha} C + u^2\left(\partial_u C \frac{\mathrm{d}u}{\mathrm{d}\alpha} + \partial_\alpha C\right)\right)$$

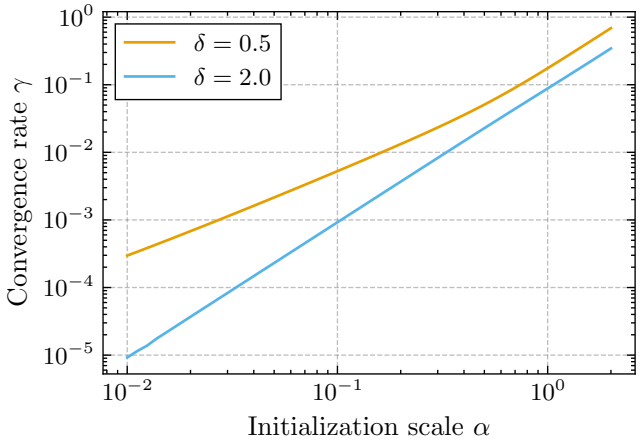


Figure 14: Theoretical convergence rates γ for $\delta = 2 > 1$ and $\delta = 0.5 < 1$. γ is monotonically increasing with respect to the initialization scale α.

$$
\begin{aligned}
&= \frac{1}{\delta}\left(\frac{\mathrm{d}u}{\mathrm{d}\alpha}(2uC + u^2\partial_u C) + u^2\partial_\alpha C\right) \\
&= \frac{1}{\delta}\left(\frac{2\alpha^3 u\,\mathbb{E}[(u+x)^{-3}]}{\mathbb{E}[x^2(u+x)^{-3}]} \cdot 2u\,\mathbb{E}\left[\frac{x^2}{(u+x)^3}\right] + 2\alpha^3 u^2\,\mathbb{E}\left[\frac{1}{x(u+x)^2} - \frac{2}{(u+x)^3}\right]\right) \\
&= \frac{2\alpha^3 u^2}{\delta}\,\mathbb{E}\left[\frac{1}{x(u+x)^2}\right] \\
&> 0\,. \qquad \text{(C.74)}
\end{aligned}
$$

In the last line, we used that $u \neq 0$ for $\delta \neq 1$.

□

These properties are illustrated in Figure 14.

Limiting Behaviors of the Convergence Rate. As $\alpha \to \infty$, the equation for u becomes, up to the leading order,

$$1 - \frac{1}{\delta}\frac{\alpha^4}{(u+\alpha^2)^2} \approx 0\,, \qquad \text{(C.75)}$$

which leads to the solution $u_* \approx \alpha^2(\delta^{-1/2} - 1)$, $A \approx \delta^{1/2}$ and $\gamma = \alpha^2(1 - \delta^{-1/2})^2$. Notice that $(1 - \delta^{-1/2})^2$ is the lower end of the support of the Marchenko–Pastur law (B.6), which is the asymptotic minimum eigenvalue of the sample covariance matrix $\delta^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$. This is consistent with the fact that the dynamics as $\alpha \to \infty$ is approximately linear (as described in Section 4), and that the convergence rate of a linear dynamics is governed by the minimum eigenvalue of the coefficients, in this case the sample covariance matrix.

As $\alpha \to 0$, we have $\gamma \to 0$, which implies subexponential decay.

D DETAILS OF THE RIGOROUS THEORY

In this section, we develop a rigorous theory for truncated DLNs. We introduce a general class of flows that includes gradient flows on truncated DLNs, characterize its high-dimensional limit using DMFT, and finally specialize to truncated DLNs.

D.1 General Setup

General Flow. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{z} \in \mathbb{R}^d$. Let $k \in \mathbb{N}$ and $\ell\colon \mathbb{R}^k \times \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}^k$, $w\colon \mathbb{R}^k \times \mathbb{R}_{\geq 0} \to \mathbb{R}^k$, $p\colon \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}_{\geq 0} \to \mathbb{R}^k$ be Lipschitz functions.

Consider the following flow for $\boldsymbol{\theta} \in \mathbb{R}^{d\times k}$, denoted as $\mathfrak{F} \coloneqq \mathfrak{F}(\boldsymbol{\theta}(0), \boldsymbol{z}, \delta, \ell, w, p)$.

$$\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = p_t(\boldsymbol{g}(t), \boldsymbol{\theta}(t))\,, \quad \boldsymbol{g}(t) = \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_t(\boldsymbol{f}(t); \boldsymbol{z})\,, \quad \boldsymbol{f}(t) = \boldsymbol{X}w_t(\boldsymbol{\theta}(t))\,. \tag{D.1}$$

Here, the functions ℓ_t, p_t, w_t are applied row-wise.

This flow is a generalization of the one defined in Celentano et al. (2021). Our generalized flow allows row-wise reparameterization of the parameter $\boldsymbol{\theta}$ through the function w_t and more general post-processing of the gradient $\boldsymbol{g}$ through the function p_t.

DMFT Equation. Given random variables $\theta(0) \in \mathbb{R}^k$ and $z \in \mathbb{R}$, we consider the following DMFT equation $\mathfrak{S} \coloneqq \mathfrak{S}(\theta(0), z, \delta, \ell, w, p)$ corresponding to the flow $\mathfrak{F}$, for unknown deterministic functions $\Gamma\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, $R_w, R_\ell, C_w, C_\ell\colon \mathbb{R}^2_{\geq 0} \to \mathbb{R}$ and stochastic processes $\theta, f, g\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$.

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = p_t(g(t), \theta(t))\,, \tag{D.2a}$$

$$g(t) = \frac{u_g(t)}{\delta} + \Gamma(t)w_t(\theta(t)) + \int_0^t R_\ell(t,s)w_s(\theta(s))\,\mathrm{d}s\,, \qquad u_g \sim \mathsf{GP}(0, \delta C_\ell)\,, \tag{D.2b}$$

$$f(t) = u_f(t) + \int_0^t R_w(t,s)\ell_s(f(s); z)\,\mathrm{d}s\,, \qquad u_f \sim \mathsf{GP}(0, C_w)\,, \tag{D.2c}$$

$$R_w(t,s) = \mathbb{E}\left[\frac{\partial w_t(\theta(t))}{\partial u_g(s)}\right], \qquad 0 \leq s \leq t\,, \tag{D.2d}$$

$$R_\ell(t,s) = \mathbb{E}\left[\frac{\partial \ell_t(f(t); z)}{\partial u_f(s)}\right], \qquad 0 \leq s \leq t\,, \tag{D.2e}$$

$$\Gamma(t) = \mathbb{E}[\nabla_f \ell_t(f(t); z)]\,, \tag{D.2f}$$

$$C_w(t,s) = \mathbb{E}[w_t(\theta(t))w_s(\theta(s))^\mathsf{T}]\,, \tag{D.2g}$$

$$C_\ell(t,s) = \mathbb{E}[\ell_t(f(t), z)\ell_s(f(s), z)^\mathsf{T}]\,. \tag{D.2h}$$

We set $R_w(t,s) = R_\ell(t,s) = 0$ for $t < s$. The quantities $\partial w_t(\theta(t))/\partial u_g(s)$ and $\partial \ell_t(f(t); z)/\partial u_f(s)$ are stochastic processes defined as follows.

$$\frac{\partial w_t(\theta(t))}{\partial u_g(s)} = \nabla_\theta w_t(\theta(t))\frac{\partial \theta(t)}{\partial u_g(s)}\,, \tag{D.3a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \theta(t)}{\partial u_g(s)} = \nabla_g p_t(g(t), \theta(t))\frac{\partial g(t)}{\partial u_g(s)} + \nabla_\theta p_t(g(t), \theta(t))\frac{\partial \theta(t)}{\partial u_g(s)}\,, \tag{D.3b}$$

$$\frac{\partial g(t)}{\partial u_g(s)} = \Gamma(t)\frac{\partial w_t(\theta(t))}{\partial u_g(s)} + \int_s^t R_\ell(t,s')\frac{\partial w_{s'}(\theta(s'))}{\partial u_g(s)}\,\mathrm{d}s'\,, \tag{D.3c}$$

$$\frac{\partial \ell_t(f(t); z)}{\partial u_f(s)} = \nabla_f \ell_t(f(t); z)\frac{\partial f(t)}{\partial u_f(s)}\,, \tag{D.3d}$$

$$\frac{\partial f(t)}{\partial u_f(s)} = R_w(t,s)\nabla_f \ell_s(f(s); z) + \int_s^t R_w(t,s')\frac{\partial \ell_{s'}(f(s'); z)}{\partial u_f(s)}\,\mathrm{d}s'\,, \tag{D.3e}$$

with the initial condition given by

$$\frac{\partial \theta(t)}{\partial u_g(t)} = \frac{1}{\delta}\nabla_g p_t(g(t), \theta(t))\,. \tag{D.4}$$

D.2 Truncated DLNs

Setup. Let $L \geq 2$ be an integer. We consider L-layer truncated diagonal linear networks defined as follows.

$$f_{L,M}(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{w}^\mathsf{T}\boldsymbol{x}, \quad \boldsymbol{w} = \frac{1}{L}(\eta_M(\boldsymbol{u}^L) - \eta_M(\boldsymbol{v}^L)), \quad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d\,, \tag{D.5}$$

where $\eta_M \colon \mathbb{R} \to \mathbb{R}$ is a truncation operator satisfying $\eta_M(x) = x$ for $|x| \le M$ and $\eta_M(x) = 0$ for $|x| \ge M+1$, for $M > 0$. Such a function can be explicitly constructed by using the smooth step function

$$\eta(t) = \begin{cases} 0 & (t \le 0)\,, \\ \dfrac{\mathrm{e}^{-1/t}}{\mathrm{e}^{-1/t} + \mathrm{e}^{-1/(1-t)}} & (0 < t < 1)\,, \\ 1 & (t \ge 1)\,, \end{cases} \tag{D.6}$$

and setting $\eta_M(x) = (1 - \eta(|x| - M))x$.

We consider a regression task with truncated DLNs. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{w}^* \in \mathbb{R}^d$, and $\boldsymbol{\xi} \in \mathbb{R}^n$. We generate labels $\boldsymbol{y} \in \mathbb{R}^n$ as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\xi}$. We consider the following loss:

$$L(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2n} \sum_{\mu=1}^{n} (y_\mu - f_{L,M}(\boldsymbol{x}_\mu; \boldsymbol{u}, \boldsymbol{v}))^2 + \frac{\lambda}{2d}(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2)\,, \tag{D.7}$$

where $\lambda \ge 0$ is a regularization parameter. The gradient flow for the loss L is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(t) &= -\frac{d}{2}\nabla_{\boldsymbol{u}} L(\boldsymbol{u}(t), \boldsymbol{v}(t)) \\ &= -\frac{1}{2}\left(\lambda \boldsymbol{u}(t) + \boldsymbol{u}(t)^{L-1} \odot \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{w}(t) - \boldsymbol{y}) \odot \eta_M'(\boldsymbol{u}(t)^L)\right), \qquad \text{(D.8a)} \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{v}(t) &= -\frac{d}{2}\nabla_{\boldsymbol{v}} L(\boldsymbol{u}(t), \boldsymbol{v}(t)) \\ &= -\frac{1}{2}\left(\lambda \boldsymbol{v}(t) - \boldsymbol{v}(t)^{L-1} \odot \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{w}(t) - \boldsymbol{y}) \odot \eta_M'(\boldsymbol{v}(t)^L)\right), \qquad \text{(D.8b)} \end{aligned}$$

for given initial values $\boldsymbol{u}(0)$ and $\boldsymbol{v}(0)$. If the entries of $\boldsymbol{u}(t)$ and $\boldsymbol{v}(t)$ stay inside $[-M, M]$, the truncation can be ignored.

DMFT Equation. Given random variables $u(0), v(0), w^*, \xi \in \mathbb{R}$, we consider the following DMFT equation corresponding to the gradient flow (D.8).

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u(t) &= p_u(g(t), u(t), v(t))\,, && \text{(D.9a)} \\ \frac{\mathrm{d}}{\mathrm{d}t}v(t) &= p_v(g(t), u(t), v(t))\,, && \text{(D.9b)} \\ w(t) &= \frac{1}{L}(\eta_M(u(t)^L) - \eta_M(v(t)^L))\,, && \text{(D.9c)} \\ g(t) &= \frac{z_g(t)}{\delta} + w(t) - w^* - \int_0^t R_f(t,s)(w(s) - w^*)\,\mathrm{d}s\,, \qquad z_g \sim \mathsf{GP}(0, \delta C_f)\,, && \text{(D.9d)} \\ f(t) &= z_f(t) - \int_0^t R_w(t,s)(f(s) - \xi)\,\mathrm{d}s\,, \qquad z_f \sim \mathsf{GP}(0, C_w)\,, && \text{(D.9e)} \\ R_w(t,s) &= -\mathbb{E}\left[\frac{\partial w(t)}{\partial z_g(s)}\right], && \text{(D.9f)} \\ R_f(t,s) &= -\mathbb{E}\left[\frac{\partial f(t)}{\partial z_f(s)}\right], && \text{(D.9g)} \\ C_w(t,s) &= \mathbb{E}[(w(t) - w^*)(w(s) - w^*)]\,, && \text{(D.9h)} \\ C_f(t,s) &= \mathbb{E}[(f(t) - \xi)(f(s) - \xi)]\,, && \text{(D.9i)} \end{aligned}$$

where

$$p_u(g,u,v) \coloneqq -\frac{1}{2}(\lambda u + u^{L-1} g \eta_M'(u^L))\,, \quad p_v(g,u,v) \coloneqq -\frac{1}{2}(-\lambda v + v^{L-1} g \eta_M'(v^L))\,, \tag{D.10}$$

and with auxiliary processes

$$\frac{\partial w(t)}{\partial z_g(s)} = u(t)^{L-1}\eta_M'(u(t)^L)\frac{\partial u(t)}{\partial z_g(s)} - v(t)^{L-1}\eta_M'(v(t))^{L-1}\frac{\partial v(t)}{\partial z_g(s)}\,, \tag{D.11a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial u(t)}{\partial z_g(s)} = \partial_g p_u(g(t),u(t),v(t))\frac{\partial g(t)}{\partial z_g(s)} + \partial_u p_u(g(t),u(t),v(t))\frac{\partial u(t)}{\partial z_g(s)} + \partial_v p_u(g(t),u(t),v(t)))\frac{\partial v(t)}{\partial z_g(s)}\,, \tag{D.11b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial u(t)}{\partial z_g(s)} = \partial_g p_v(g(t),u(t),v(t))\frac{\partial g(t)}{\partial z_g(s)} + \partial_u p_v(g(t),u(t),v(t))\frac{\partial u(t)}{\partial z_g(s)} + \partial_v p_v(g(t),u(t),v(t)))\frac{\partial v(t)}{\partial z_g(s)}\,, \tag{D.11c}$$

$$\frac{\partial g(t)}{\partial z_g(s)} = \frac{\partial w(t)}{\partial z_g(s)} - \int_0^t R_f(t,s')\frac{\partial w(s')}{\partial z_g(s)}\,\mathrm{d}s'\,, \tag{D.11d}$$

$$\frac{\partial f(t)}{\partial z_f(s)} = -\int_0^t R_w(t,s')\frac{\partial f(s')}{\partial z_f(s)}\,\mathrm{d}s' + R_w(t,s)\,, \tag{D.11e}$$

with initial conditions given by

$$\frac{\partial u(t)}{\partial z_g(t)} = \frac{1}{\delta}\partial_g p_u(g(t),u(t),v(t))\,, \quad \frac{\partial v(t)}{\partial z_g(t)} = \frac{1}{\delta}\partial_g p_v(g(t),u(t),v(t))\,. \tag{D.12}$$

If we ignore the truncation η_M, this equation for $L=2$ matches the equation (A.36) obtained heuristically. After further simplification along the lines of Appendix A.2, we obtain the DMFT equation (4) presented in the main text.

D.3 Statements of the Results

Assumption D.1.

- The entries $\boldsymbol{X} = (x_{ij})_{i\in[n],j\in[d]}$ are independent and satisfy $\mathbb{E}\,x_{ij} = 0, \mathbb{E}\,x_{ij}^2 = 1/d, \|x_{ij}\|_{\psi_2} \le C/\sqrt{d}$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm.
- $n, d \to \infty,\ n/d \to \delta \in (0,\infty)$.
- $\boldsymbol{z} \in \mathbb{R}^d, \boldsymbol{\theta}(0) \in \mathbb{R}^{d\times k}$ is independent of $\boldsymbol{X}$, and the empirical distributions $\hat{\mu}_{\theta(0)} \coloneqq d^{-1}\sum_{i=1}^d \delta_{\theta_i(0)}, \hat{\mu}_z \coloneqq n^{-1}\sum_{i=1}^n \delta_{z_i}$ converge to $\mathsf{P}(\theta(0)), \mathsf{P}(z)$, respectively, in p-Wasserstein distance for all $p \ge 1$, almost surely as $n, d \to \infty$.

Assumption D.2. The functions $\ell_t(f;z), w_t(\theta), p_t(g;\theta)$ and their Jacobians $\mathrm{D}\ell, \mathrm{D}w, \mathrm{D}p$ are Lipschitz continuous in $t \in \mathbb{R}_{\ge 0}$ and $\theta, f, g \in \mathbb{R}^k$, i.e., there exists a universal constant $M > 0$ such that for all $t_1, t_2 \in [0,T]$ and all $\theta_1, \theta_2, f_1, f_2, g_1, g_2 \in \mathbb{R}^k$,

$$\|\ell_{t_1}(f_1;z) - \ell_{t_2}(f_2;z)\|_2 \le M(\|f_1 - f_2\|_2 + |t_1 - t_2|)\,, \tag{D.13}$$

$$\|w_{t_1}(\theta_1) - w_{t_2}(\theta_2)\|_2 \le M(\|\theta_1 - \theta_2\|_2 + |t_1 - t_2|)\,, \tag{D.14}$$

$$\|p_{t_1}(g_1;\theta_1) - p_{t_2}(g_2;\theta_2)\|_2 \le M(\|g_1 - g_2\|_2 + \|\theta_1 - \theta_2\|_2 + |t_1 - t_2|)\,, \tag{D.15}$$

$$\|\mathrm{D}\ell_{t_1}(f_1;z) - \mathrm{D}\ell_{t_2}(f_2;z)\|_2 \le M(\|f_1 - f_2\|_2 + |t_1 - t_2|)\,, \tag{D.16}$$

$$\|\mathrm{D}w_{t_1}(\theta_1) - \mathrm{D}w_{t_2}(\theta_2)\|_2 \le M(\|\theta_1 - \theta_2\|_2 + |t_1 - t_2|)\,, \tag{D.17}$$

$$\|\mathrm{D}p_{t_1}(g_1;\theta_1) - \mathrm{D}p_{t_2}(g_2;\theta_2)\|_2 \le M(\|g_1 - g_2\|_2 + \|\theta_1 - \theta_2\|_2 + |t_1 - t_2|)\,. \tag{D.18}$$

The following theorem establishes the existence and uniqueness of the solution of the DMFT equation $\mathfrak{S}$. We give a proof in Appendix D.4.

Theorem 3. *Under Assumption D.2, for any $T > 0$ there exists a tuple $(\theta, g, f, R_w, R_\ell, \Gamma, C_w, C_\ell)$ that solves the DMFT system $\mathfrak{S}$. The solution is unique among all such tuples with (C_w, R_w) bounded in all compact sets in $\mathbb{R}^2_{\ge 0}$. Further, there exist functions $\Phi_{R_w}, \Phi_{R_\ell}, \Phi_{C_w}, \Phi_{C_\ell} \colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ that satisfy*

$$\|R_w(t,s)\| \le \Phi_{R_w}(t-s)\,,\ \|R_\ell(t,s)\| \le \Phi_{R_\ell}(t-s)\,,\ \|C_w(t,t)\| \le \Phi_{C_w}(t)\,,\ \|C_\ell(t,t)\| \le \Phi_{C_\ell}(t)\,, \tag{D.19}$$

and $\|\Gamma(t)\| \le M$*, for all* $t, s \ge 0$*. Further, there exists* $\lambda > 0$ *such that*

$$\lim_{t\to\infty} \mathrm{e}^{-\lambda t} \max\{\Phi_{R_w}(t), \Phi_{R_\ell}(t), \Phi_{C_w}(t), \Phi_{C_\ell}(t)\} = 0\,. \tag{D.20}$$

Finally, the stochastic processes $(\theta(t), g(t), f(t))_{t\in[0,T]}$ *have continuous sample paths.*

The following theorem characterizes the empirical distribution of the flow variable $\boldsymbol{\theta}(t)$ in $\mathfrak{F}$ as the solution of the DMFT system $\mathfrak{S}$. We give a proof in Appendix D.5.

Theorem 4. *Under Assumptions D.1 and D.2, for* $T > 0$*, let* $(\theta(t), g(t))_{t=0}^T$ *be the unique stochastic processes that solve the DMFT equation* $\mathfrak{S}$ *in Theorem 3. Then, we have*

$$\frac{1}{d}\sum_{i=1}^{d} \delta_{(\theta_i(t))_{t=0}^T} \xrightarrow{W_2} \mathsf{P}((\theta(t))_{t=0}^T)\,, \quad \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i,(f_i(t))_{t=0}^T} \xrightarrow{W_2} \mathsf{P}(z, (f(t))_{t=0}^T)\,, \tag{D.21}$$

almost surely as $n, d \to \infty$*. Here,* P *denotes the law of the given random variables.*

As a corollary of Theorem 4, we obtain a DMFT characterization of the gradient flow for truncated DLNs. We give a proof in Appendix D.6.

Corollary 5. *Under Assumption D.1 (with* $\boldsymbol{z} = \boldsymbol{\xi}$ *and* $\boldsymbol{\theta}(0) = (\boldsymbol{u}(0), \boldsymbol{v}(0), \boldsymbol{w}^*)$*), for any* $T > 0$*, there exists a unique solution to the DMFT equation (D.9) in the sense of Theorem 3, and we have*

$$\frac{1}{d}\sum_{i=1}^{d} \delta_{(u_i(t),v_i(t))_{t=0}^T} \xrightarrow{W_2} \mathsf{P}((u(t), v(t))_{t=0}^T)\,, \quad \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i,(f_i(t))_{t=0}^T} \xrightarrow{W_2} \mathsf{P}(\xi, (f(t))_{t=0}^T)\,, \tag{D.22}$$

almost surely as $n, d \to \infty$*.*

D.4 Proof of Theorem 3

We follow the approach of the proof of Celentano et al. (2021, Theorem 1). The proof proceeds as follows.

I. Define the functions $\Phi_{R_w}, \Phi_{R_\ell}, \Phi_{C_w}, \Phi_{C_\ell}$.

II. Define metric spaces $\mathcal{S}$ and $\bar{\mathcal{S}}$ for the triple (C_ℓ, R_ℓ, Γ) and the pair (C_w, R_w), respectively, and show that the stochastic processes θ, f, g are determined uniquely in these spaces.

III. Define a mapping $\mathcal{T}\colon \mathcal{S} \to \mathcal{S}$. We show that the solution (C_ℓ, R_ℓ, Γ) of the DMFT system is a fixed point of $\mathcal{T}$. We then show that the map $\mathcal{T}$ is a contraction. By Banach's fixed point theorem, it follows that $\mathcal{T}$ has a unique fixed point, establishing the existence and uniqueness of the solution.

I. Definition of the Functions Φ. Since the following quantities are bounded by assumptions, we take $M > 0$ large enough such that

$$\max\left\{1 + \mathbb{E}\|\theta(0)\|_2^2,\ \sup_{t\ge0}\mathbb{E}\|\ell_t(0;z)\|_2^2,\ \sup_{t\ge0}\|w_t(0)\|_2,\ \sup_{t\ge0}\|p_t(0,0)\|_2\right\} \le M\,. \tag{D.23}$$

Consider the following system of integro-differential equations.

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_{R_\theta}(t) = 2M^3\Phi_{R_\theta}(t) + M^2\int_0^t \Phi_{R_\ell}(t-s')\Phi_{R_\theta}(s'-s)\,\mathrm{d}s'\,, \tag{D.24}$$

$$\Phi_{R_\ell}(t) = M\cdot\left\{\Phi_{R_\theta}(t) + \int_0^t \Phi_{R_\theta}(t-s)\Phi_{R_\ell}(s)\,\mathrm{d}s\right\}, \tag{D.25}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\Phi_{C_\theta}(t)} = \sqrt{3}\cdot\sqrt{8M^6\Phi_{C_\theta}(t) + \frac{M^2k}{\delta}\Phi_{C_\ell}(t) + 2M^4\int_0^t (t-s+1)^2\Phi_{R_\ell}(t-s)^2\Phi_{C_\theta}(t)\,\mathrm{d}s}\,, \tag{D.26}$$

$$\Phi_{C_\ell}(t) = 3\cdot\left\{M + kM^2\Phi_{C_\theta}(t) + M^2\int_0^t (t-s+1)^2\Phi_{R_\theta}(t-s)^2\Phi_{C_\ell}(s)\,\mathrm{d}s\right\}. \tag{D.27}$$

By Celentano et al. (2021, Lemma 5.1), there exists a nondecreasing solution to the above system, given $\Phi_{R_\theta}(0) > 0, \Phi_{C_\theta}(0) > 0$. In addition, there exists $\lambda > 0$ such that the following holds.

$$\lim_{t\to\infty} \mathrm{e}^{-\lambda t} \max\{\Phi_{R_\theta}(t), \Phi_{R_\ell}(t), \Phi_{C_\theta}(t), \Phi_{C_\ell}(t)\} = 0\,. \tag{D.28}$$

Using this solution, we define

$$\Phi_{R_w}(t) \coloneqq M\Phi_{R_\theta}(t)\,, \quad \Phi_{C_w}(t) \coloneqq 2M^2\Phi_{C_\theta}(t)\,. \tag{D.29}$$

II. Definition of the Spaces $\mathcal{S}, \bar{\mathcal{S}}$. We define the following function spaces.

Definition D.3 (Function triplet spaces $\mathcal{S}$ and $\mathcal{S}_{\mathrm{cont}}$)**.** We define the space $\mathcal{S} \coloneqq \mathcal{S}(\Phi_{R_w}, \Phi_{R_\ell}, \Phi_{C_w}, \Phi_{C_\ell}, M_{\mathcal{S}}, T)$ of triples (C_ℓ, R_ℓ, Γ) that satisfy the following.

- $C_\ell(t,s)$ is a covariance kernel and satisfies $\|C_\ell(t,t)\| \le \Phi_{C_\ell}(t)$ for $t \in [0,T]$ and

$$C_\ell(0,0) = \mathbb{E}[\ell_0(f(0);z)\ell_0(f(0);z)^{\mathsf{T}}]\,, \quad f(0) \sim \mathsf{N}(0, \mathbb{E}[\theta(0)\theta(0)^{\mathsf{T}}])\,. \tag{D.30}$$

  Further, $C_\ell(t,s)$ is continuous for $s,t \in [0,T] \setminus P$ where P is a finite set, and for any $s \le t$ such that C_ℓ is continuous in $[s,t]^2$,

$$\|C_\ell(t,t) - 2C_\ell(t,s) + C_\ell(s,s)\| \le M_{\mathcal{S}}(t-s)^2\,. \tag{D.31}$$

- $R_\ell(t,s)$ is measurable, $R_\ell(t,s) = 0$ for $t \le s$, and $\|R_\ell(t,s)\| \le \Phi_{R_\ell}(t-s)$ for $0 \le s \le t \le T$.
- $\Gamma(t)$ is measurable, $\|\Gamma(t)\| \le M$ for $t \in [0,T]$, and

$$\Gamma(0) = \mathbb{E}[\nabla_f \ell_0(f(0);z)]\,, \quad f(0) \sim \mathsf{N}(0, \mathbb{E}[\theta(0)\theta(0)^{\mathsf{T}}])\,. \tag{D.32}$$

We define the space $\mathcal{S}_{\mathrm{cont}} \subset \mathcal{S}$ of all (C_ℓ, R_ℓ, Γ) such that C_ℓ is continuous (i.e., $P = \emptyset$) and for all $s, s' \le t$,

$$\|C_\ell(t,s) - C_\ell(t,s')\| \le \sqrt{\Phi_{C_\ell}(T)M_{\mathcal{S}}} \cdot |s - s'|\,. \tag{D.33}$$

Definition D.4 (Function pair spaces $\bar{\mathcal{S}}$ and $\bar{\mathcal{S}}_{\mathrm{cont}}$)**.** We define the space $\bar{\mathcal{S}} \coloneqq \bar{\mathcal{S}}(\Phi_{R_w}, \Phi_{R_\ell}, \Phi_{C_w}, \Phi_{C_\ell}, M_{\bar{\mathcal{S}}}, T)$ of pairs (C_w, R_w) that satisfy the following.

- $C_w(t,s)$ is a covariance kernel and satisfies $\|C_w(t,t)\| \le \Phi_{C_w}(t)$ for $t \in [0,T]$ and

$$C_w(0,0) = \mathbb{E}[w_0(\theta(0))w_0(\theta(0))^{\mathsf{T}}]\,. \tag{D.34}$$

  Further, $C_w(t,s)$ is continuous for all $s,t \in [0,T] \setminus P$ where P is a finite set, and for any $s \le t$ such that $C_w(t,s)$ is continuous in $[s,t]^2$,

$$\|C_w(t,t) - 2C_w(t,s) + C_w(s,s)\| \le M_{\bar{\mathcal{S}}}(t-s)^2\,. \tag{D.35}$$

- $R_w(t,s)$ is measurable, $R_w(t,s) = 0$ for $t < s$, and $\|R_w(t,s)\| \le \Phi_{R_w}(t-s)$ for $0 \le s \le t \le T$.

We define the space $\bar{\mathcal{S}}_{\mathrm{cont}} \subset \bar{\mathcal{S}}$ of all (C_w, R_w) such that C_w is continuous (i.e., $P = \emptyset$) and for all $s, s' \le t$,

$$\|C_w(t,s) - C_w(t,s')\| \le \sqrt{\Phi_{C_w}(T)M_{\bar{\mathcal{S}}}} \cdot |s - s'|\,. \tag{D.36}$$

The constants $M_{\mathcal{S}}, M_{\bar{\mathcal{S}}}$ are chosen in the proof of Lemma D.5.

For given $(C_\ell, R_\ell, \Gamma) \in \mathcal{S}$ and $(C_w, R_w) \in \bar{\mathcal{S}}$, stochastic processes $\theta(t), f(t), \partial w_t(\theta(t))/\partial u_g(s), \partial \ell_t(f(t);z)/\partial u_f(s)$ are uniquely determined by Equations (D.2) and (D.3) (Celentano et al., 2021, Lemma 5.4).

III. Definition of the Map $\mathcal{T}$. We define $\mathcal{T} = \mathcal{T}_{\bar{\mathcal{S}}\to\mathcal{S}} \circ \mathcal{T}_{\mathcal{S}\to\bar{\mathcal{S}}}$, where $\mathcal{T}_{\mathcal{S}\to\bar{\mathcal{S}}}\colon \mathcal{S} \to \bar{\mathcal{S}}, \mathcal{T}_{\bar{\mathcal{S}}\to\mathcal{S}}\colon \bar{\mathcal{S}} \to \mathcal{S}$ are defined in the following.

We define $\mathcal{T}_{\mathcal{S}\to\bar{\mathcal{S}}}\colon (C_\ell, R_\ell, \Gamma) \mapsto (\bar{C}_w, \bar{R}_w)$ as follows. For a given $(C_\ell, R_\ell, \Gamma) \in \bar{\mathcal{S}}$, take the unique stochastic processes $\theta(t), \partial w_t(\theta(t))/\partial u_g(s)$ satisfying Equations (D.2a), (D.2b) and (D.3a) to (D.3c), and define

$$\bar{C}_w(t,s) = \mathbb{E}[w_t(\theta(t))w_s(\theta(s))^\mathsf{T}], \quad \bar{R}_w(t,s) = \mathbb{E}\left[\frac{\partial w_t(\theta(t))}{\partial u_g(s)}\right]. \tag{D.37}$$

Similarly, we define $\mathcal{T}_{\bar{\mathcal{S}}\to\mathcal{S}}\colon (\bar{C}_w, \bar{R}_w) \mapsto (\bar{C}_\ell, \bar{R}_\ell, \bar{\Gamma})$ as follows. For a given $(\bar{C}_w, \bar{R}_w) \in \bar{\mathcal{S}}$, take the unique stochastic processes $f(t), \partial \ell_t(f(t); z)/\partial u_f(s)$ satisfying the equations Equations (D.2c), (D.3d) and (D.3e), and define

$$\bar{C}_\ell(t,s) = \mathbb{E}[\ell_t(f(t);z)\ell_s(f(s);z)^\mathsf{T}], \quad \bar{R}_\ell(t,s) = \mathbb{E}\left[\frac{\partial \ell_t(f(t);z)}{\partial u_f(s)}\right], \quad \bar{\Gamma}(t) = \mathbb{E}[\nabla_f \ell_t(f(t);z)]. \tag{D.38}$$

In the following lemma, we show that these mappings indeed map into $\bar{\mathcal{S}}$ and $\mathcal{S}$, respectively.

Lemma D.5. *In addition to the assumptions for Theorem 3, assume $\Phi_{C_\theta}(0) > M$ and $\Phi_{R_\theta}(0) > M/\delta$. Then, $\mathcal{T}_{\mathcal{S}\to\bar{\mathcal{S}}}$ maps $\mathcal{S}$ to $\bar{\mathcal{S}}_{\text{cont}} \subset \bar{\mathcal{S}}$, and $\mathcal{T}_{\bar{\mathcal{S}}\to\mathcal{S}}$ maps $\bar{\mathcal{S}}$ to $\mathcal{S}_{\text{cont}} \subset \mathcal{S}$.*

Once this lemma is established, it remains to show that $\mathcal{T}$ is a contraction. The rest of the proof is essentially identical to Celentano et al. (2021, Section 5.4) and is omitted.

Proof of Lemma D.5. The proof proceeds along the same lines as the proof of Celentano et al. (2021, Lemma 5.5). We need to modify the norm evaluations to account for additional processing with functions w_t and p_t.

Note that the map $\mathcal{T}_{\bar{\mathcal{S}}\to\mathcal{S}}$ defined above is identical to the one defined in Celentano et al. (2021, Section 5.4), up to rescaling. Thus, it follows immediately that $\mathcal{T}_{\bar{\mathcal{S}}\to\mathcal{S}}$ maps $\bar{\mathcal{S}}$ into $\mathcal{S}_{\text{cont}}$

It remains to show that $\mathcal{T}_{\mathcal{S}\to\bar{\mathcal{S}}}$ maps $\mathcal{S}$ into $\bar{\mathcal{S}}_{\text{cont}}$. By the Lipschitz continuity of w and p, we have

$$\|w_t(\theta(t))\| \le \|w_t(0)\|_2 + M\|\theta(t)\|_2 \le M(1 + \|\theta(t)\|_2), \tag{D.39}$$
$$\|p_t(g(t), \theta(t))\| \le \|p_t(0,0)\|_2 + M(\|g(t)\|_2 + \|\theta(t)\|_2) \le M(1 + \|g(t)\|_2 + \|\theta(t)\|_2). \tag{D.40}$$

We first show that $\|\bar{C}_w(t,t)\| \le \Phi_{C_w}(t)$. By $\mathfrak{S}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\theta(t)\|_2 \le \|p_t(g(t),\theta(t))\| \le M(1 + \|g(t)\|_2 + \|\theta(t)\|_2), \tag{D.41}$$

$$\begin{aligned}\|g(t)\|_2 &\le \frac{1}{\delta}\|u_g(t)\|_2 + \|\Gamma(t)\|\|w_t(\theta(t))\|_2 + \int_0^t \|R_\ell(t,s)\|\|w_s(\theta(s))\|_2 \,\mathrm{d}s \\ &\le \frac{1}{\delta}\|u_g(t)\|_2 + M^2(1 + \|\theta(t)\|_2) + M\int_0^t \Phi_{R_\ell}(t-s)(1 + \|\theta(s)\|_2)\,\mathrm{d}s.\end{aligned} \tag{D.42}$$

Combining, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\theta(t)\|_2 \le 2M^3(1 + \|\theta(t)\|_2) + \frac{M}{\delta}\|u_g(t)\|_2 + M^2\int_0^t \Phi_{R_\ell}(t-s)(1 + \|\theta(s)\|_2)\,\mathrm{d}s. \tag{D.43}$$

Furthermore, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{1 + \mathbb{E}\|\theta(t)\|_2^2} &= \frac{\mathbb{E}\left[\|\theta(t)\|_2 \frac{\mathrm{d}}{\mathrm{d}t}\|\theta(t)\|_2\right]}{\sqrt{1 + \mathbb{E}\|\theta(t)\|_2^2}} \overset{\text{(i)}}{\le} \sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}t}\|\theta(t)\|_2\right)^2\right]} \\ &\le \sqrt{\mathbb{E}\left[\left(2M^3(1 + \|\theta(t)\|_2) + \frac{M}{\delta}\|u_g(t)\|_2 + M^2\int_0^t \Phi_{R_\ell}(t-s)(1 + \|\theta(s)\|_2)\,\mathrm{d}s\right)^2\right]} \\ &\overset{\text{(ii)}}{\le} \sqrt{\mathbb{E}\left[\left(4M^6(1 + \|\theta(t)\|_2)^2 + \frac{M^2}{\delta^2}\|u_g(t)\|_2^2 + M^4\int_0^t (t-s+1)^2\Phi_{R_\ell}(t-s)^2(1 + \|\theta(s)\|_2)^2\,\mathrm{d}s\right)\right]}\end{aligned}$$

$$\cdot \sqrt{\left(1+1+\int_0^t (t-s+1)^{-2}\,\mathrm{d}s\right)}$$

$$\overset{\text{(iii)}}{\leq} \sqrt{3}\cdot\sqrt{8M^6(1+\mathbb{E}\|\theta(t)\|_2^2)+\frac{M^2k}{\delta}\Phi_{C_\ell}(t)+2M^4\int_0^t (t-s+1)^2\Phi_{R_\ell}(t-s)^2(1+\mathbb{E}\|\theta(s)\|_2^2)\,\mathrm{d}s}\,, \tag{D.44}$$

where in (i) and (ii) we used the Cauchy-Schwarz inequality, and in (iii) we used the following inequality.

$$\mathbb{E}\|u_g(t)\|_2^2 = \mathrm{tr}(\mathbb{E}[u_g(t)u_g(t)^\mathsf{T}]) \leq k\|\mathbb{E}[u_g(t)u_g(t)^\mathsf{T}]\| = k\delta\|C_\ell(t,t)\| \leq k\delta\Phi_{C_\ell}(t)\,. \tag{D.45}$$

By $\Phi_{C_\theta}(0) > M > 1+\mathbb{E}\|\theta(0)\|_2^2$ and Equation (D.26), $1+\mathbb{E}\|\theta(t)\|_2^2 < \Phi_{C_\theta}(t)$ holds for $t \geq 0$. Thus, we have

$$\|\bar{C}_w(t,t)\| = \|\mathbb{E}[w_t(\theta(t))w_t(\theta(t))^\mathsf{T}]\| \leq \mathbb{E}\|w_t(\theta(t))\|_2^2 \leq 2M^2(1+\mathbb{E}\|\theta(t)\|_2^2) < 2M^2\Phi_{C_\theta}(t) = \Phi_{C_w}(t)\,. \tag{D.46}$$

Next, we show that $\|\bar{R}_w(t,s)\| \leq \Phi_{R_w}(t-s)$. By $\mathfrak{S}$, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| &\leq \|\nabla_g p_t(g(t),\theta(t))\|\left\|\frac{\partial g(t)}{\partial u_g(s)}\right\| + \|\nabla_\theta p_t(g(t),\theta(t))\|\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| \\
&\leq M\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| + M\left\|\frac{\partial g(t)}{\partial u_g(s)}\right\|, \qquad \text{(D.47)} \\
\left\|\frac{\partial g(t)}{\partial u_g(s)}\right\| &\leq \|\Gamma(t)\|\|\nabla_\theta w_t(\theta(t))\|\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| + \int_0^t \|R_\ell(t,u)\|\|\nabla_\theta w_{s'}(\theta(s'))\|\left\|\frac{\partial\theta(s')}{\partial u_g(s)}\right\|\mathrm{d}s' \\
&\leq M^2\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| + M\int_0^t \Phi_{R_\ell}(t-s')\left\|\frac{\partial\theta(s')}{\partial u_g(s)}\right\|\mathrm{d}s'\,. \qquad \text{(D.48)}
\end{aligned}$$

Thus, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| \leq 2M^3\,\mathbb{E}\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| + M^2\int_0^t \Phi_{R_\ell}(t-s')\,\mathbb{E}\left\|\frac{\partial\theta(s')}{\partial u_g(s)}\right\|\mathrm{d}s'\,. \tag{D.49}$$

By $\left\|\frac{\partial\theta(t)}{\partial u_g(t)}\right\| \leq M/\delta < \Phi_{R_\theta}(0)$ and Equation (D.24), $\mathbb{E}\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| < \Phi_{R_\theta}(t-s)$ holds for $t \geq s$. Thus, we have

$$\begin{aligned}
\|\bar{R}_w(t,s)\| = \left\|\mathbb{E}\left[\frac{\partial w_t(\theta(t))}{\partial u_g(s)}\right]\right\| \leq \mathbb{E}\left[\left\|\frac{\partial w_t(\theta(t))}{\partial u_g(s)}\right\|\right] &\leq \mathbb{E}\left[\|\nabla_\theta w_t(\theta(t))\|\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\|\right] \\
&\leq M\,\mathbb{E}\left\|\frac{\partial\theta(t)}{\partial u_g(s)}\right\| < M\Phi_{R_\theta}(t-s) = \Phi_{R_w}(t-s)\,. \qquad \text{(D.50)}
\end{aligned}$$

Next, we show that $\|\bar{C}_w(t,t) - 2\bar{C}_w(t,s) + \bar{C}_w(s,s)\| \leq M_{\bar{\mathcal{S}}}(t-s)^2$. We have

$$\|\bar{C}_w(t,t) - 2\bar{C}_w(t,s) + \bar{C}_w(s,s)\| \leq \mathbb{E}\|w_t(\theta(t)) - w_s(\theta(s))\|_2^2 \leq 2M((t-s)^2 + \mathbb{E}\|\theta(t)-\theta(s)\|_2^2)\,, \tag{D.51}$$

and

$$\begin{aligned}
\mathbb{E}\|\theta(t)-\theta(s)\|_2^2 = \mathbb{E}\left\|\int_s^t \frac{\mathrm{d}}{\mathrm{d}t'}\theta(t')\,\mathrm{d}t'\right\|_2^2 &\leq (t-s)^2\sup_{0\leq t\leq T}\mathbb{E}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\theta(t)\right\|_2^2 \\
&\leq (t-s)^2\sup_{0\leq t\leq T} 3\left(8M^6\Phi_{C_\theta}(t) + \frac{M^2k}{\delta}\Phi_{C_\ell}(t) + 2M^4\int_0^t (t-s+1)^2\Phi_{R_\ell}(t-s)^2\Phi_{C_\theta}(s)\,\mathrm{d}s\right) \\
&\leq (t-s)^2\cdot 3\left(8M^6\Phi_{C_\theta}(T) + \frac{M^2k}{\delta}\Phi_{C_\ell}(T) + 2M^4T(T+1)^2\Phi_{R_\ell}(T)^2\Phi_{C_\theta}(T)\right) \\
&=: A(t-s)^2\,. \qquad \text{(D.52)}
\end{aligned}$$

Setting $M_{\bar{\mathcal{S}}} \coloneqq 2M(1+A)$, we obtain $\|\bar{C}_w(t,t) - 2\bar{C}_w(t,s) + \bar{C}_w(s,s)\| \leq M_{\bar{\mathcal{S}}}(t-s)^2$.

Finally, by the Cauchy-Schwarz inequality, we have

$$\|\bar{C}_w(t,s) - \bar{C}_w(t,s')\| \leq \sqrt{\mathbb{E}\|w_t(\theta(t))\|_2^2\cdot\mathbb{E}\|w_s(\theta(s)) - w_{s'}(\theta(s'))\|_2^2} \leq \sqrt{\Phi_{C_w(T)}M_{\bar{\mathcal{S}}}}\cdot|s-s'|\,. \tag{D.53}$$

Thus, we have $(\bar{C}_w, \bar{R}_w) \in \bar{\mathcal{S}}_{\mathrm{cont}}$.

□

D.5 Proof of Theorem 4

Again, we follow the approach of the proof of Celentano et al. (2021, Theorem 2). The proof proceeds as follows.

I. Discretize the flow $\mathfrak{F}$ with step size $\eta > 0$.

II. Map the discretized flow $\mathfrak{F}^\eta$ to an approximate message passing (AMP) iteration. Show that the state evolution of the AMP iteration is equivalent to the discretized version of the DMFT equation, $\mathfrak{S}^\eta$.

III. Show that as $\eta \to 0$, the solution of the discretized DMFT equation $\mathfrak{S}^\eta$ converges to the unique solution of the DMFT equation $\mathfrak{S}$.

First, we define the discretized flow $\mathfrak{F}^\eta$ as follows. For $i = 0, 1, \dots,$ set $t_i := i\eta$ and define

$$\frac{\boldsymbol{\theta}^\eta(t_{i+1}) - \boldsymbol{\theta}^\eta(t_i)}{\eta} = p_{t_i}(\boldsymbol{g}^\eta(t_i), \boldsymbol{\theta}^\eta(t_i))\,, \quad \boldsymbol{g}^\eta(t_i) = \frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_{t_i}(\boldsymbol{f}^\eta(t_i); \boldsymbol{z})\,, \quad \boldsymbol{f}^\eta(t_i) = \boldsymbol{X} w_{t_i}(\boldsymbol{\theta}^\eta(t_i))\,, \tag{D.54}$$

with $\boldsymbol{\theta}^\eta(0) = \boldsymbol{\theta}(0)$. We extend this flow to continuous time by piecewise linear interpolation. We denote by

Next, we define the discretized DMFT equation $\mathfrak{S}^\eta$ as follows.

$$\frac{\theta^\eta(t_{i+1}) - \theta^\eta(t_i)}{\eta} = p_{t_i}(g^\eta(t_i), \theta^\eta(t_i))\,, \tag{D.55a}$$

$$g^\eta(t_i) = \frac{u_g^\eta(t_i)}{\delta} + \Gamma^\eta(t_i) w_{t_i}(\theta^\eta(t_i)) + \eta \sum_{j=0}^{i-1} R_\ell^\eta(t_i, t_j) w_{t_j}(\theta^\eta(t_j))\,, \qquad u_g \sim \mathsf{N}(0, \delta C_\ell^\eta)\,, \tag{D.55b}$$

$$f^\eta(t_i) = u_f^\eta(t_i) + \eta \sum_{j=0}^{i-1} R_w^\eta(t_i, t_j) \ell_{t_j}(f^\eta(t_j); z)\,, \qquad u_f \sim \mathsf{N}(0, C_w^\eta)\,, \tag{D.55c}$$

$$R_w^\eta(t_i, t_j) = \eta^{-1}\,\mathbb{E}\left[\frac{\partial w_{t_i}(\theta^\eta(t_i))}{\partial u_g^\eta(t_j)}\right], \qquad 0 \le j < i\,, \tag{D.55d}$$

$$R_\ell^\eta(t_i, t_j) = \eta^{-1}\,\mathbb{E}\left[\frac{\partial \ell_{t_i}(f^\eta(t_i); z)}{\partial u_f^\eta(t_j)}\right], \qquad 0 \le j < i\,, \tag{D.55e}$$

$$\Gamma^\eta(t_i) = \mathbb{E}[\nabla_f \ell_{t_i}(f^\eta(t_i); z)]\,, \tag{D.55f}$$

$$C_w^\eta(t_i, t_j) = \mathbb{E}[w_{t_i}(\theta^\eta(t_i)) w_{t_j}(\theta^\eta(t_j))^\mathsf{T}]\,, \tag{D.55g}$$

$$C_\ell^\eta(t_i, t_j) = \mathbb{E}[\ell_{t_i}(f^\eta(t_i); z) \ell_{t_j}(f^\eta(t_j); z)^\mathsf{T}]\,. \tag{D.55h}$$

We set $R_w^\eta(t_i, t_j) = R_\ell^\eta(t_i, t_j) = 0$ for $i \le j$. The quantities $\partial w_{t_i}(\theta^\eta(t_i))/\partial u_g^\eta(t_j)$ and $\partial \ell_{t_i}(f^\eta(t_i); z)/\partial u_f^\eta(t_j)$ are stochastic processes defined as follows.

$$\frac{\partial w_{t_i}(\theta^\eta(t_i))}{\partial u_g^\eta(t_j)} = \nabla_\theta w_{t_i}(\theta^\eta(t_i)) \frac{\partial \theta^\eta(t_i)}{\partial u_g^\eta(t_j)}\,, \tag{D.56a}$$

$$\frac{1}{\eta}\left(\frac{\partial \theta^\eta(t_{i+1})}{\partial u_g^\eta(t_j)} - \frac{\partial \theta^\eta(t_i)}{\partial u_g^\eta(t_j)}\right) = \nabla_g p_{t_i}(g^\eta(t_i), \theta^\eta(t_i)) \frac{\partial g^\eta(t_i)}{\partial u_g^\eta(t_j)} + \nabla_\theta p_{t_i}(g^\eta(t_i), \theta^\eta(t_i)) \frac{\partial \theta^\eta(t_i)}{\partial u_g^\eta(t_j)}\,, \tag{D.56b}$$

$$\frac{\partial g^\eta(t_i)}{\partial u_g^\eta(t_j)} = \Gamma^\eta(t_i) \frac{\partial w_{t_i}(\theta^\eta(t_i))}{\partial u_g^\eta(t_j)} + \eta \sum_{j'=j+1}^{i-1} R_\ell^\eta(t_i, t_{j'}) \frac{\partial w_{t_{j'}}(\theta^\eta(t_{j'}))}{\partial u_g^\eta(t_j)}\,, \tag{D.56c}$$

$$\frac{\partial \ell_{t_i}(f^\eta(t_i); z)}{\partial u_f^\eta(t_j)} = \nabla_f \ell_{t_i}(f^\eta(t_i); z) \frac{\partial f^\eta(t_i)}{\partial u_f^\eta(t_j)}\,, \tag{D.56d}$$

$$\frac{\partial f^\eta(t_i)}{\partial u_f^\eta(t_j)} = \eta R_w^\eta(t_i, t_j) \nabla_f \ell_j(f_j; z) + \eta \sum_{j'=j+1}^{i-1} R_w^\eta(t_i, t_{j'}) \frac{\partial \ell_{t_{j'}}(f^\eta(t_{j'}); z)}{\partial u_f^\eta(t_j)}\,, \tag{D.56e}$$

with the initial value

$$\frac{\partial \theta^\eta(t_{i+1})}{\partial u_g^\eta(t_i)} = \frac{\eta}{\delta} \nabla_g p_{t_i}(g^\eta(t_i), \theta^\eta(t_i))\,. \tag{D.57}$$

Similarly to $\mathfrak{F}^\eta$, we extend this to continuous time by piecewise linear interpolation.

The three parts of the proof correspond to the following three lemmas.

Lemma D.6. *Under the assumptions of Theorem 4, for any $\tau_1, \dots, \tau_m \in [0, T]$, we have, almost surely,*

$$\lim_{\eta \to 0} \limsup_{n,d \to \infty} W_2 \left(\frac{1}{d} \sum_{i=1}^{d} \delta_{\theta_i(\tau_1), \dots, \theta_i(\tau_m)}, \frac{1}{d} \sum_{i=1}^{d} \delta_{\theta_i^\eta(\tau_1), \dots, \theta_i^\eta(\tau_m)} \right) = 0 \,, \tag{D.58}$$

$$\lim_{\eta \to 0} \limsup_{n,d \to \infty} W_2 \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{f_i(\tau_1), \dots, f_i(\tau_m), z_i}, \frac{1}{n} \sum_{i=1}^{n} \delta_{f_i^\eta(\tau_1), \dots, f_i^\eta(\tau_m), z_i} \right) = 0 \,. \tag{D.59}$$

Lemma D.7. *Under the assumptions of Theorem 4, let $(\theta^\eta(t), f^\eta(t))_{t=0}^T$ be the unique solution of the discretized DMFT equation $\mathfrak{S}^\eta$. For any $\tau_1, \dots, \tau_m \in [0, T]$, we have, almost surely,*

$$\limsup_{n,d \to \infty} W_2 \left(\frac{1}{d} \sum_{i=1}^{d} \delta_{\theta_i^\eta(\tau_1), \dots, \theta_i^\eta(\tau_m)}, \mathsf{P}(\theta^\eta(\tau_1), \dots, \theta^\eta(\tau_m)) \right) = 0 \,, \tag{D.60}$$

$$\limsup_{n,d \to \infty} W_2 \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{f_i^\eta(\tau_1), \dots, f_i^\eta(\tau_m), z_i}, \mathsf{P}(f^\eta(\tau_1), \dots, f^\eta(\tau_m), z) \right) = 0 \,. \tag{D.61}$$

Lemma D.8. *Under the assumptions of Theorem 4, let $(\theta^\eta(t), f^\eta(t))_{t=0}^T$ and $(\theta(t), f(t))_{t=0}^T$ be the unique solution of $\mathfrak{S}^\eta$ and $\mathfrak{S}$, respectively. For any $\tau_1, \dots, \tau_m \in [0, T]$, we have,*

$$\lim_{\eta \to 0} W_2(\mathsf{P}(\theta^\eta(\tau_1), \dots, \theta^\eta(\tau_m)), \mathsf{P}(\theta(\tau_1), \dots, \theta(\tau_m))) = 0 \,, \tag{D.62}$$

$$\lim_{\eta \to 0} W_2(\mathsf{P}(f^\eta(\tau_1), \dots, f^\eta(\tau_m), z), \mathsf{P}(f(\tau_1), \dots, f(\tau_m), z)) = 0 \,. \tag{D.63}$$

Proofs of Lemmas D.6 and D.8 are essentially identical to those of Celentano et al. (2021, Lemmas 6.1 and 6.3) and thus omitted.

Here, we prove Lemma D.7. We prove a slightly stronger convergence result (almost sure 2-Wasserstein convergence) than Celentano et al. (2021, Lemma 6.2) (weak convergence in probability) by utilizing a recent universality result (Wang et al., 2024).

Proof of Lemma D.7. In the following, we omit the superscript η for simplicity, as we only consider the discretized systems $\mathfrak{F}^\eta$ and $\mathfrak{S}^\eta$.

Reduction to AMP. We consider the following AMP iteration. For a sequence of Lipschitz functions $F_i \colon \mathbb{R}^{k(i+1)+1} \to \mathbb{R}^m$, $G_i \colon \mathbb{R}^{k(i+1)} \to \mathbb{R}^k$ $(i = 0, 1, \dots)$, generate a sequence of matrices $\boldsymbol{a}_{i+1} \in \mathbb{R}^{d \times k}, \boldsymbol{b}_i \in \mathbb{R}^{n \times k}$ $(i \geq 0)$ as follows.

$$\boldsymbol{a}_{i+1} = \boldsymbol{X}^\mathsf{T} F_i(\boldsymbol{b}_0, \dots, \boldsymbol{b}_i; \boldsymbol{z}) - \delta \sum_{j=0}^{i} G_j(\boldsymbol{a}_1, \dots, \boldsymbol{a}_j; \boldsymbol{\theta}(0)) \xi_{i,j} \,, \tag{D.64}$$

$$\boldsymbol{b}_i = \boldsymbol{X} G_i(\boldsymbol{a}_1, \dots, \boldsymbol{a}_i; \boldsymbol{\theta}(0)) - \sum_{j=0}^{i-1} F_j(\boldsymbol{b}_0, \dots, \boldsymbol{b}_j; \boldsymbol{z}) \zeta_{i,j} \,, \tag{D.65}$$

with the initial value $G_0(\boldsymbol{\theta}(0)) = \boldsymbol{\theta}(0)$. Here, F_i, G_i are applied row-wise. The matrices $\{\xi_{i,j}\}_{0 \leq j \leq i}, \{\zeta_{i,j}\}_{0 \leq j \leq i-1} \subset \mathbb{R}^{k \times k}$ are defined by

$$\zeta_{i,j} = \mathbb{E}\left[\frac{\partial}{\partial \bar{u}_g(j+1)} G_i(\bar{u}_g(1), \dots, \bar{u}_g(i); \theta(0)) \right], \qquad 0 \leq j < i \,, \tag{D.66}$$

$$\xi_{i,j} = \mathbb{E}\left[\frac{\partial}{\partial \bar{u}_f(j)} F_i(\bar{u}_f(0), \dots, \bar{u}_f(i); z) \right], \qquad 0 \leq j \leq i \,, \tag{D.67}$$

where $(\bar{u}_g(i+1), \bar{u}_f(i))_{i\geq 0}$ is a sequence of centered Gaussian vectors in $\mathbb{R}^k$ with covariance

$$\mathbb{E}[\bar{u}_f(i)\bar{u}_f(j)^\mathsf{T}] = \mathbb{E}[G_i(\bar{u}_g(1),\dots,\bar{u}_g(i);\theta(0))G_j(\bar{u}_g(1),\dots,\bar{u}_g(j);\theta(0))^\mathsf{T}]\,, \qquad 0\leq j\leq i\,, \tag{D.68}$$

$$\mathbb{E}[\bar{u}_g(i+1)\bar{u}_g(j+1)^\mathsf{T}] = \delta\,\mathbb{E}[F_i(\bar{u}_f(0),\dots,\bar{u}_f(i);z)F_j(\bar{u}_f(0),\dots,\bar{u}_f(j);z)^\mathsf{T}]\,, \qquad 0\leq j\leq i\,. \tag{D.69}$$

This AMP iteration can be related to $\boldsymbol{\theta}(t_i)$ and $\boldsymbol{f}(t_i)$ by taking F_i and G_i as follows.

$$G_i(\boldsymbol{a}_1,\dots,\boldsymbol{a}_i;\boldsymbol{\theta}(0)) = w_{t_i}(\boldsymbol{\theta}(t_i)) \tag{D.70}$$

$$F_i(\boldsymbol{b}_0,\dots,\boldsymbol{b}_i;\boldsymbol{z}) = \ell_{t_i}(\boldsymbol{f}(t_i);\boldsymbol{z})\,. \tag{D.71}$$

We can show by induction that the right-hand sides of the above equations are Lipschitz functions of $\boldsymbol{a}_i, \boldsymbol{\theta}(0)$ and $\boldsymbol{b}_i, \boldsymbol{z}$, respectively, as below.

$$\begin{aligned}
\boldsymbol{\theta}(t_i) &= \boldsymbol{\theta}(t_{i-1}) + \eta p_{t_{i-1}}\left(\frac{1}{\delta}\boldsymbol{X}^\mathsf{T}\ell_{t_{i-1}}(\boldsymbol{f}(t_{i-1});\boldsymbol{z}), \boldsymbol{\theta}(t_{i-1})\right)\\
&= \boldsymbol{\theta}(t_{i-1}) + \eta p_{t_{i-1}}\left(\frac{1}{\delta}\boldsymbol{X}^\mathsf{T}F_{i-1}(\boldsymbol{b}_0,\dots,\boldsymbol{b}_{i-1};\boldsymbol{z}), \boldsymbol{\theta}(t_{i-1})\right)\\
&= \boldsymbol{\theta}(t_{i-1}) + \eta p_{t_{i-1}}\left(\frac{\boldsymbol{a}_i}{\delta} + \sum_{j=0}^{i-1} G_j(\boldsymbol{a}_1,\dots,\boldsymbol{a}_j;\boldsymbol{\theta}(0))\xi_{i-1,j}, \boldsymbol{\theta}(t_{i-1})\right),
\end{aligned} \tag{D.72}$$

$$\begin{aligned}
\boldsymbol{f}(t_i) &= \boldsymbol{X}w_{t_i}(\boldsymbol{\theta}(t_i))\\
&= \boldsymbol{X}G_i(\boldsymbol{a}_1,\dots,\boldsymbol{a}_i;\boldsymbol{\theta}(0))\\
&= \boldsymbol{b}_i + \sum_{j=0}^{i-1} F_j(\boldsymbol{b}_0,\dots,\boldsymbol{b}_j;\boldsymbol{z})\zeta_{i,j}\,.
\end{aligned} \tag{D.73}$$

By Wang et al. (2024, Theorem 2.21), for all second-order pseudo-Lipschitz functions $\psi, \tilde{\psi}\colon \mathbb{R}^{k(K+1)} \to \mathbb{R}$, we have, almost surely,

$$\lim_{n,d\to\infty} \frac{1}{d}\sum_{j=1}^{d} \psi((\boldsymbol{a}_1)_j,\dots,(\boldsymbol{a}_K)_j;\theta_j(0)) = \mathbb{E}[\psi(\bar{u}_g(1),\dots,\bar{u}_g(K);\theta(0))]\,, \tag{D.74}$$

$$\lim_{n,d\to\infty} \frac{1}{n}\sum_{j=1}^{n} \tilde{\psi}((\boldsymbol{b}_0)_j,\dots,(\boldsymbol{b}_K)_j;z_j) = \mathbb{E}[\tilde{\psi}(\bar{u}_f(0),\dots,\bar{u}_f(K);z)]\,. \tag{D.75}$$

Since $\boldsymbol{\theta}(i)$ is a Lipschitz function of $\boldsymbol{a}_1,\dots,\boldsymbol{a}_i,\boldsymbol{\theta}(0)$, we can take some Lipschitz function h_θ such that $\boldsymbol{\theta}(i) = h_\theta(\boldsymbol{a}_1,\dots,\boldsymbol{a}_i;\boldsymbol{\theta}(0))$. Then, we define $\bar{\theta}(i) \coloneqq h_\theta(\bar{u}_g(1),\dots,\bar{u}_g(i);\theta(0))$. Similarly, we define $\bar{f}(i) = h_f(\bar{u}_f(0),\dots,\bar{u}_f(i);z)$ where h_f is such that $\boldsymbol{f}(i) = h_f(\boldsymbol{b}_0,\dots,\boldsymbol{b}_i;\boldsymbol{z})$. Since a composition of Lipschitz functions is again Lipschitz, we have

$$\lim_{n,d\to\infty} \frac{1}{d}\sum_{j=1}^{d} \psi(\theta_j(1),\dots,\theta_j(K);\theta(0)_j) = \mathbb{E}[\psi(\bar{\theta}(1),\dots,\bar{\theta}(K);\theta(0))]\,, \tag{D.76}$$

$$\lim_{n,d\to\infty} \frac{1}{n}\sum_{j=1}^{n} \tilde{\psi}(f_j(0),\dots,f_j(K);z_j) = \mathbb{E}[\tilde{\psi}(\bar{f}(0),\dots,\bar{f}(K);z)]\,. \tag{D.77}$$

It then follows that

$$\lim_{n,d\to\infty} W_2\left(\frac{1}{d}\sum_{j=1}^{d}\delta_{\theta(0)_j,\dots,\theta_j(K)}, \mathsf{P}(\bar{\theta}(0),\dots,\bar{\theta}(K))\right) = 0\,, \tag{D.78}$$

$$\lim_{n,d\to\infty} W_2\left(\frac{1}{n}\sum_{j=1}^{n}\delta_{f_j(0),\dots,f_j(K),z_j}, \mathsf{P}(\bar{f}(0),\dots,\bar{f}(K),z)\right) = 0\,. \tag{D.79}$$

Mapping the State Evolution to DMFT. We map the state evolution variables $\bar{\theta}(i), \bar{f}(i)$ to the DMFT variables $\theta(t_i), f(t_i)$ for $\mathfrak{S}^\eta$. We show by induction on r that the unique solution of $\mathfrak{S}^\eta$ is given as follows.

$$(\theta(t_1), \dots, \theta(t_r)) \overset{\mathrm{d}}{=} (\bar{\theta}(1), \dots, \bar{\theta}(r)) \,, \tag{D.80}$$

$$(f(t_0), \dots, f(t_r)) \overset{\mathrm{d}}{=} (\bar{f}(0), \dots, \bar{f}(r)) \,, \tag{D.81}$$

$$R_w(t_i, t_j) = \zeta_{i,j}/\eta \,, \qquad 0 \le j < i \le r \,, \tag{D.82}$$

$$R_\ell(t_i, t_j) = \xi_{i,j}/\eta \,, \qquad 0 \le j < i \le r \,, \tag{D.83}$$

$$\Gamma(t_i) = \xi_{i,i} \,, \qquad 0 \le i \le r \,. \tag{D.84}$$

For $r = 0$, we have $\theta(0) \overset{\mathrm{d}}{=} \bar{\theta}(0)$, and

$$f(t_0) = u_f(t_0) \overset{\mathrm{d}}{=} \bar{u}_f(0) = \bar{f}(0) \overset{\mathrm{d}}{=} \mathsf{N}(0, \mathbb{E}[\theta(0)\theta(0)^\mathsf{T}]) \,, \tag{D.85}$$

$$\Gamma(t_0) = \mathbb{E}[\nabla_f \ell_{t_0}(f(t_0); z)] = \mathbb{E}[\nabla_f \ell_{t_0}(\bar{f}(0); z)] = \mathbb{E}\left[\frac{\partial}{\partial \bar{u}_f(0)} F_0(\bar{u}_f(0); z)\right] = \xi_{0,0} \,. \tag{D.86}$$

Next, we assume that the hypothesis holds for r, and show that equations (D.80)–(D.84) hold for $r+1$.

First, we check (D.80) and (D.82). For $0 \le j \le i \le r$, we have

$$\begin{aligned}
\mathbb{E}[\bar{u}_g(i+1)\bar{u}_g(j+1)^\mathsf{T}] &= \delta\, \mathbb{E}[F_i(\bar{u}_f(0), \dots, \bar{u}_f(i); z) F_j(\bar{u}_f(1), \dots, \bar{u}_f(j); z)^\mathsf{T}] \\
&= \delta\, \mathbb{E}[\ell_{t_i}(\bar{f}(i); z) \ell_{t_j}(\bar{f}(j); z)] \\
&= \delta\, \mathbb{E}[\ell_{t_i}(f(t_i); z) \ell_{t_j}(f(t_j); z)] \\
&= \delta C_\ell(t_i, t_j) \,,
\end{aligned} \tag{D.87}$$

and therefore $(u_g(t_0), \dots, u_g(t_r)) \overset{\mathrm{d}}{=} (\bar{u}_g(1), \dots, \bar{u}_g(r+1))$ holds.

By $\mathfrak{S}^\eta$,

$$\theta(t_{r+1}) = \theta(r) + \eta p_{t_r}(g(t_r), \theta(t_r)) \,, \tag{D.88}$$

$$\begin{aligned}
g(t_r) &= \frac{u_g(t_r)}{\delta} + \Gamma(t_r) w_{t_r}(\theta(t_r)) + \eta \sum_{j=0}^{r-1} R_\ell(t_r, t_j) w_{t_j}(\theta(t_j)) \\
&= \frac{u_g(t_r)}{\delta} + \xi_{r,r} w_{t_r}(\theta(t_r)) + \sum_{j=0}^{r-1} \xi_{r,j} w_{t_j}(\theta(t_j)) \\
&= \frac{u_g(t_r)}{\delta} + \sum_{j=0}^{r} \xi_{r,j} w_{t_j}(\theta(j)) \,.
\end{aligned} \tag{D.89}$$

By the state evolution,

$$\bar{\theta}(r+1) = \bar{\theta}(r) + \eta p_{t_r}(\bar{g}(r), \bar{\theta}(r)) \,, \tag{D.90}$$

$$\begin{aligned}
\bar{g}(r) &\coloneqq \frac{\bar{u}_g(r+1)}{\delta} + \sum_{j=0}^{r} G_j(\bar{u}_g(1), \dots, \bar{u}_g(j); \theta(0)) \xi_{r,j} \\
&= \frac{\bar{u}_g(r+1)}{\delta} + \sum_{j=0}^{r} \xi_{r,j} w_{t_j}(\bar{\theta}(j)) \,.
\end{aligned} \tag{D.91}$$

Comparing these two equations, we get $\theta(t_{r+1}) \overset{\mathrm{d}}{=} \bar{\theta}(r+1)$. A similar argument shows that, for $0 \le j \le r$, $\frac{\partial}{\partial u_g(t_j)} w_{t_{r+1}}(\theta(t_{r+1})) \overset{\mathrm{d}}{=} \frac{\partial}{\partial \bar{u}_g(j+1)} w_{t_{r+1}}(\bar{\theta}(r+1))$. Thus, it follows that $R_w(t_{r+1}, t_j) = \zeta_{r+1,j}/\eta$.

Next, we check (D.81), (D.83), and (D.84). For $0 \le j \le i \le r+1$, we have

$$\mathbb{E}[\bar{u}_f(i)\bar{u}_f(j)^\mathsf{T}] = \mathbb{E}[G_i(\bar{u}_g(1), \dots, \bar{u}_g(i); \theta(0)) G_j(\bar{u}_g(1), \dots, \bar{u}_g(j); \theta(0))^\mathsf{T}]$$

$$
\begin{aligned}
&= \mathbb{E}[w_{t_i}(\bar{\theta}(i))w_{t_j}(\bar{\theta}(j))] \\
&= \mathbb{E}[w_{t_i}(\theta(t_i))w_{t_j}(\theta(t_j))] \\
&= C_w(t_i, t_j)
\end{aligned} \tag{D.92}
$$

and therefore $(u_f(t_0), \dots, u_f(t_{r+1})) \overset{\mathrm{d}}{=} (\bar{u}_f(0), \dots, \bar{u}_f(r+1))$ holds.

By $\mathfrak{S}^\eta$,

$$
\begin{aligned}
f(t_{r+1}) &= u_f(t_{r+1}) + \eta \sum_{j=0}^{r} R_w(t_{r+1}, t_j)\ell_{t_j}(f(t_j); z) \\
&= u_f(t_{r+1}) + \sum_{j=0}^{r} \zeta_{r+1,j}\ell_{t_j}(f(t_j); z)\,,
\end{aligned} \tag{D.93}
$$

By the state evolution,

$$
\begin{aligned}
\bar{f}(r+1) &= \bar{u}_f(r+1) + \sum_{j=0}^{r} F_j(\bar{u}_f(0), \dots, \bar{u}_f(j); z)\zeta_{r+1,j} \\
&= \bar{u}_f(r+1) + \sum_{j=0}^{r} \zeta_{r+1,j}\ell_{t_j}(\bar{f}(j); z)\,.
\end{aligned} \tag{D.94}
$$

Comparing these two equations, we get $f(t_{r+1}) \overset{\mathrm{d}}{=} \bar{f}(r+1)$. A similar argument shows that, for $0 \le j \le r+1$, $\frac{\partial}{\partial u_f(t_j)}\ell_{t_{r+1}}(f(t_{r+1}); z) \overset{\mathrm{d}}{=} \frac{\partial}{\partial \bar{u}_f(j)}\ell_{t_{r+1}}(\bar{f}(r+1); z)$. Thus, it follows that $R_\ell(t_{r+1}, t_j) = \xi_{r+1,j}/\eta, \Gamma(t_{r+1}) = \xi_{r+1,r+1}$.

□

D.6 Proof of Corollary 5

The gradient flow for regression using truncated DLNs can be mapped to our general flow $\mathfrak{F}$ by setting $k = 3$ and taking $\boldsymbol{\theta}(t), w_t, p_t, \ell_t$ as follows.

$$
\boldsymbol{\theta}(t) = (\boldsymbol{u}(t), \boldsymbol{v}(t), \boldsymbol{w}^*)\,, \tag{D.95}
$$

$$
w_t(u, v, w^*) = \left(\frac{1}{L}(\eta_M(u^L) - \eta_M(v^L)), w^*, 0\right), \tag{D.96}
$$

$$
p_t(g, \_, \_; u, v, w^*) = (p_u(g, u, v), p_v(g, u, v), 0)\,, \tag{D.97}
$$

$$
\ell_t(f, f^*, \_; \xi) = (f - f^* - \xi, 0, 0)\,, \tag{D.98}
$$

where the underscore _ indicates unused entries.

It is easy to check Assumptions D.1 and D.2 for our specific choices of functions. Applying Theorems 3 and 4, we obtain the following DMFT equation.

$$
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} u(t) \\ v(t) \\ w^* \end{pmatrix} = \begin{pmatrix} p_u(g(t), u(t), v(t)) \\ p_v(g(t), u(t), v(t)) \\ 0 \end{pmatrix}, \tag{D.99}
$$

$$
\begin{pmatrix} g(t) \\ \_ \end{pmatrix} = \frac{1}{\delta}\begin{pmatrix} z_g(t) \\ z_{g*}(t) \end{pmatrix} + \Gamma(t)\begin{pmatrix} w(t) \\ w^* \end{pmatrix} + \int_0^t R_\ell(t, s)\begin{pmatrix} w(s) \\ w^* \end{pmatrix}\mathrm{d}s\,, \qquad \begin{pmatrix} z_g \\ z_{g*} \end{pmatrix} \sim \mathsf{GP}(0, \delta C_\ell)\,, \tag{D.100}
$$

$$
\begin{pmatrix} f(t) \\ f_* \end{pmatrix} = \begin{pmatrix} z_f(t) \\ z_{f*}(t) \end{pmatrix} + \int_0^t R_w(t, s)\begin{pmatrix} f(t) - f_* - \xi \\ 0 \end{pmatrix}\mathrm{d}s\,, \qquad \begin{pmatrix} z_f \\ z_{f*} \end{pmatrix} \sim \mathsf{GP}(0, C_w)\,, \tag{D.101}
$$

$$
w(t) \coloneqq \frac{1}{L}(\eta_M(u(t)^L) - \eta_M(v(t)^L))\,, \tag{D.102}
$$

where

$$R_w(t,s) = \begin{pmatrix} R_w(t,s) & 0 \\ 0 & 0 \end{pmatrix}, \quad R_\ell(t,s) = \begin{pmatrix} R_f(t,s) & R_{f*}(t,s) \\ 0 & 0 \end{pmatrix}, \quad \Gamma(t) = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}, \\ C_w(t,s) = \begin{pmatrix} C_w(t,s) & m(t) \\ m(s) & \rho^2 \end{pmatrix}, \; C_\ell(t,s) = \begin{pmatrix} C_f(t,s) & 0 \\ 0 & 0 \end{pmatrix}, \tag{D.103}$$

$$R_w(t,s) = \mathbb{E}\left[\frac{\partial w(t)}{\partial z_g(s)}\right], \quad R_f(t,s) = \mathbb{E}\left[\frac{\partial (f(t)-f^*-\xi)}{\partial z_f(s)}\right], \quad R_{f*}(t,s) = \mathbb{E}\left[\frac{\partial (f(t)-f^*-\xi)}{\partial z_{f*}(s)}\right], \\ C_w(t,s) = \mathbb{E}[w(t)w(s)], \quad m(t) = \mathbb{E}[w(t)w^*], \quad C_f(t,s) = \mathbb{E}[(f(t)-f^*-\xi)(f(s)-f^*-\xi)], \tag{D.104}$$

where $\rho^2 \coloneqq \mathbb{E}[w^{*2}]$ and we omitted unused rows and columns.

We have $R_{f*}(t,s) = -R_f(t,s)$. Resetting $f(t) - f^*$, $z_f(t) - z_{f*}(t)$ as $f(t)$, $z_f(t)$, respectively, and unfolding the expressions, and flipping the signs of R_w and R_f, we have the following.

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = p_u(g(t), u(t), v(t)), \tag{D.105a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}v(t) = p_v(g(t), u(t), v(t)), \tag{D.105b}$$

$$g(t) = \frac{z_g(t)}{\delta} + w(t) - w^* - \int_0^t R_f(t,s)(w(s)-w^*)\,\mathrm{d}s, \qquad z_g \sim \mathsf{GP}(0, \delta C_f), \quad w^* \sim P_*, \tag{D.105c}$$

$$f(t) = z_f(t) - \int_0^t R_w(t,s)(f(s)-\xi)\,\mathrm{d}s, \qquad z_f \sim \mathsf{GP}(0, C_w), \quad \xi \sim P_\xi, \tag{D.105d}$$

$$R_w(t,s) = -\mathbb{E}\left[\frac{\partial w(t)}{\partial z_g(s)}\right], \tag{D.105e}$$

$$R_f(t,s) = -\mathbb{E}\left[\frac{\partial f(t)}{\partial z_f(s)}\right], \tag{D.105f}$$

$$C_w(t,s) = \mathbb{E}[(w(t)-w^*)(w(s)-w^*)], \tag{D.105g}$$

$$C_f(t,s) = \mathbb{E}[(f(t)-\xi)(f(s)-\xi)]. \tag{D.105h}$$

This concludes the proof.

E DETAILS OF NUMERICAL EXPERIMENTS

All experiments were conducted using Python on a single Apple MacBook Pro with an 8-core CPU and 16 GB of RAM.

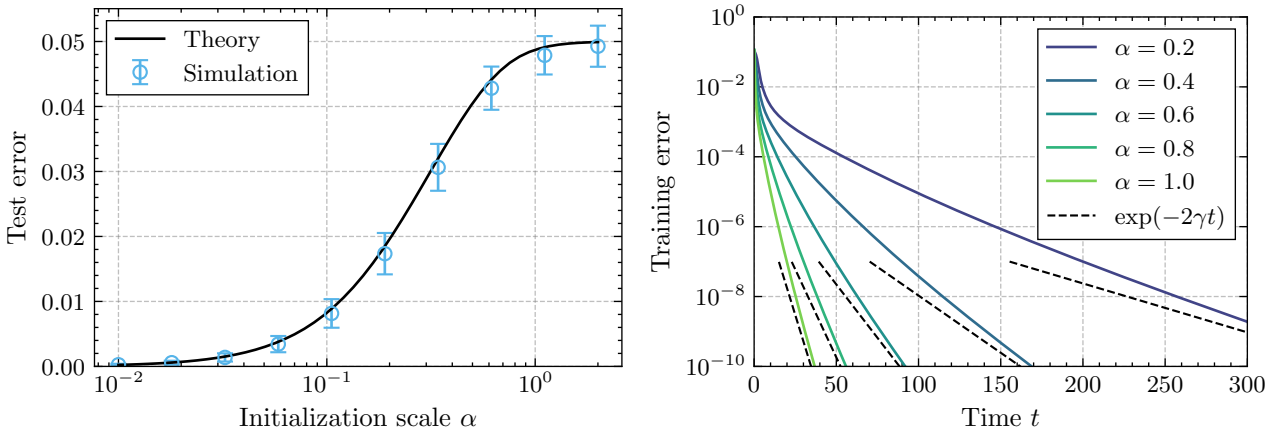
E.1 Experiments with Non-Gaussian Data

To illustrate the universality of our theoretical result, we conducted experiments with binary features $x_{\mu i} \sim \mathsf{Unif}(\{\pm 1/\sqrt{d}\})$. Other experimental setups are the same as the experiments on Gaussian data described in Section 7. Fixed points and convergence rates are shown in Figure 15, showing good agreement with the theoretical prediction.

E.2 Experiments with Real-World Data

We use a gene expression dataset (Fiorini, 2016) (a subset of Ellrott (2013)) consisting of 801 instances and 20531 features. We whiten the entire dataset using principal component analysis and randomly sample subsets of samples and features to obtain $n = 100$ times $d = 200$ data matrix. Columns are normalized to have mean zero and variance $1/d$. The empirical distribution of whitened features is shown in Figure 16.

The results on the convergence rate quantitatively deviate from the theoretical prediction as illustrated in Figure 2c. This is most likely due to a finite sample size effect. We have whitened the dataset using $n = 801$ samples, instead of ideal $n = \infty$ samples. Since the matrix is whitened using finite samples, the spectrum of the empirical



(a) Fixed points.

(b) Convergence rates of the training error.

Figure 15: Fixed points and convergence rates of the loss on binary data with $\delta = 0.5$. Theoretical predictions are shown in black and agree well with the numerical simulation, thus highlighting the universality of our result with respect to the data distribution.

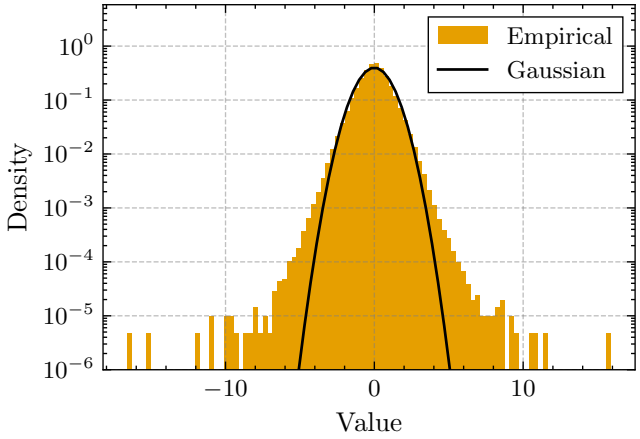


Figure 16: The empirical distribution of the whitened gene expression data. It has heavier tails than the standard Gaussian distribution shown in black.

covariance matrix for $n = 100$ samples concentrates closer to one, and the minimum eigenvalue of the empirical covariance matrix becomes larger. Since the convergence rate is closely related to the minimum eigenvalue of the sample covariance, as suggested by the $\alpha \to \infty$ analysis of the convergence rate in Appendix C.3, the finite sample size effect slightly accelerates the dynamics, resulting in the deviation from the theoretical prediction in Figure 2c.