

High-Dimensional Scaling Limits of Online Stochastic Gradient Descent in Single-Index Models

by

Parsa Rangriz

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2025

© Parsa Rangriz 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor: Aukosh Jagannath
Canada Research Chair in Mathematical
Foundations of Data Science,
Associate Professor,
Department of Statistics and Actuarial Science,
Department of Applied Mathematics,
Cheriton School of Computer Science,
University of Waterloo

Internal Member: Gregory Rice
Associate Chair of Graduate Studies,
Associate Professor,
Department of Statistics and Actuarial Science,
University of Waterloo

Internal Member: Shoja'eddin Chenouri
Professor,
Department of Statistics and Actuarial Science,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We analyze the scaling limits of stochastic gradient descent (SGD) with a constant step size in the high-dimensional regime in single-index models. Specifically, we prove limit theorems for the trajectories of finite-dimensional summary statistics of SGD as the dimension tends to infinity. These scaling limits enable the analysis of both ballistic dynamics, described by a system of ordinary differential equations (ODEs), and diffusive dynamics, captured by a system of stochastic differential equations (SDEs). Additionally, we analyze a critical step-size scaling regime where, below this threshold, the effective ballistic dynamics align with the gradient flow of the population loss. In contrast, a new diffusive correction term appears at the threshold due to fluctuations around the fixed points. Furthermore, we discuss nearly sharp thresholds for the number of samples required for consistent estimation, which depend solely on an intrinsic property of the activation function known as the information exponent. Our main contribution is demonstrating that if a single-index model has an information exponent greater than two, the deterministic scaling limit, corresponding to the ballistic phase, or so-called dynamical mean-field theory in statistical physics, fails to achieve consistent estimation in high-dimensional inference problems. This shows the necessity of diffusive correction terms to accurately describe the dynamics of online SGD in single-index models via SDEs such as an Ornstein-Uhlenbeck process.

Acknowledgements

I am deeply thankful to Prof. Aukosh Jagannath for his outstanding guidance, constant support, and inspiration throughout my time at the University of Waterloo. His supervision has greatly influenced my thesis and my entire academic journey.

I also want to thank Prof. Shoja'eddin Chenouri and Prof. Gregory Rice for their caring guidance and support during my master's degree and for their valuable roles on my thesis committee.

Finally, I would like to thank Kuroush Allameh, especially, for his valuable discussions and collaboration over the past two years. I also extend my acknowledgement to my friends Kasra Fallah, Daneshvar Amrollahi, Navid Akbari, Onur Kılınç, Atahan Bayır, Sina Kalantarzadeh, Amirmohammad Ghasemi, Amirhossein Boreiri, Mojtaba Rahmani, and many others for their friendship, helpful advice, and continuous support during my studies.

Dedication

To my dear parents, Shirin and Hassan, whose endless love, encouragement, and support have been my strength and motivation throughout these studies.

Table of Contents

Examining Committee	ii
Author’s Declaration	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Abbreviations	ix
1 Introduction	1
1.1 Overview	1
1.2 Contribution	3
1.3 Organization	4
2 Background	5
2.1 Dynamical mean field theory	5
2.1.1 Gradient flow	5
2.1.2 Distributional dynamics	6
2.1.3 Deterministic effective dynamics	10
2.2 Sample complexity	15
2.2.1 Search and descent phases	15
2.2.2 Information exponent	17
2.2.3 Single-index models	19
2.3 Effective dynamics	20

2.3.1	Summary statistics	20
2.3.2	Localizability and asymptotic closability	21
2.3.3	Diffusive scaling limit	22
2.3.4	Sub-critical and critical scaling regimes	23
2.3.5	Ballistic and diffusive phases	24
2.3.6	Rescaled effective dynamics	24
3	Main Results	26
3.1	Effective dynamics of single-index models	26
3.2	Proofs	28
3.2.1	Proof of Theorem 3.1 and Corollary 3.2	28
3.2.2	Proof of Theorem 3.3 and Corollary 3.4	31
3.2.3	Proof of Proposition 2.10	33
4	Discussion	35
	References	38
	APPENDICES	43
A	Sub-gaussian and sub-exponential	44
B	Distributional dynamics	47
C	Effective dynamics	48

List of Abbreviations

CLT Central Limit Theorem [1](#)

DMFT Dynamical Mean Field Theory [2](#)

i.i.d. independent and identically distributed [5](#)

LLN Law of Large Numbers [3](#)

ODE Ordinary Differential Equation [3](#)

OU Ornstein-Uhlenbeck [3](#)

PDE Partial Differential Equation [3](#)

SDE Stochastic Differential Equation [2](#)

SGD Stochastic Gradient Descent [1](#)

Chapter 1

Introduction

1.1 Overview

[Stochastic Gradient Descent \(SGD\)](#) is one of the most important algorithms for machine learning and related optimization problems. Since its introduction by Robbins and Monro (1951) [45], researchers have worked on SGD to understand how it iterates and behaves when applied to various learning problems. In the fixed-dimensional regime, which is now called the classical theory, many results have been discovered regarding the asymptotic theory of SGD, leading to well-established path-wise limit theorems, such as *functional Central Limit Theorem (CLT)* and *large deviation principles* by McLeish (1976) and others [8, 9, 27, 33, 35, 39]. Indeed, these results help explain how the scaling limits of SGD behave when the step size (learning rate) is sufficiently small and how its trajectory relates to the *gradient flow problem*. However, in high dimensions, where the dimensions of data and parameters grow significantly, the classical approach does not fully capture the complexities, as one may not be able to study the high-dimensional trajectory of SGD.

An essential question regarding the sample complexity of SGD is the number of data points the algorithm needs to learn successfully. Recent research regarding the sample complexity indicates that the evolution of SGD can be divided into two distinct phases: the *search phase* and the *descent phase*. In the search phase, the algorithm navigates a complex and frequently non-convex landscape characterized by substantial fluctuations due to updates that rely on stochastic gradients. In contrast, the descent phase occurs once SGD has effectively located a specific landscape region. Then, the algorithm evolves within an effective trust region, where it behaves deterministically, and the updates become more stable and converge to a local minimum. In fact, in this regime, the approximation to the true gradient of the *population loss* is more concrete than in the search phase, leading to a more predictable trajectory.

In the fixed-dimensional setting, one may ignore the analysis of the search phase by bypassing an initial *burn-in* period and assuming that SGD begins directly in the descent phase. This simplification is sometimes used to suppose convexity or quasi-convexity restrictions in the study of SGD, a topic that has been widely studied in the literature

[26, 31, 41, 42, 12, 13]. However, it is less clear whether the burn-in period can be ignored in high dimensions, as the relationship between SGD’s convergence duration and the dimension is not well-studied. Additionally, in many interesting problems, random initializations are often concentrated around an uninformative critical point of the population loss, which makes the study of SGD in high dimensions tricky. Nevertheless, many numerical experiments suggest that the algorithm can perform efficiently in these settings. This implies that SGD may be able to recover from a random initialization within a computationally feasible timescale, even in cases where the behavior from a random start differs from what it is when starting in a trust region.

In recent years, there has been a significant effort to interpret these challenges better and create general frameworks for analyzing the convergence of SGD for non-convex landscapes in high-dimensional settings. While many results still require restrictions such as convexity, some only need uniform control over the derivatives of the empirical loss, i.e., possibly a uniform bounded Lipschitzness. These methods often focus on Monte Carlo approximations or [Stochastic Differential Equation \(SDE\)](#) and then analyze how the corresponding processes converge to their invariant measures [43, 56, 15, 36].

Ben Arous, Gheissari, and Jagannath (2021) [6] classified the high-dimensional challenges of the sample complexity of SGD based on a geometric quantity known as the *information exponent*, which captures the non-linearity of the population loss near the trust region, i.e., maximal entropy. The information exponent is the degree of the first nonzero term in the Taylor expansion of the population loss around the *correlation* between the iterates of SGD and the hidden variable. It has been shown that thresholds for sample complexity for online spherical SGD to exit the search phase depend on whether the information exponent is 1, 2, or at least 3, as they can be understood as linear, quasi-linear, or polynomial sample complexities in the dimension, respectively.

Similar gradient flows of *spin glasses* have also been studied in statistical physics, often referred to as [Dynamical Mean Field Theory \(DMFT\)](#) approximation. By analyzing these flows, one may understand the dynamics of complex systems, such as neural networks that store and retrieve patterns, while establishing connections between optimization, energy landscape, and learning processes. This type of dynamics was studied in the context of the *Sherrington-Kirkpatrick model* [51], the *spherical p -spin glass models* [21, 22], and the *Hopfield model* [44] where can be modified by adding a noise term, leading to *Langevin dynamics*. These models introduce stochastic fluctuations into the evolution equation, capturing the effects of *thermal noises* in the system, and the zero-noise case is well-studied as the *zero-temperature* framework [14, 22, 21, 34, 38]. This well-established framework for analyzing gradient flows and Langevin flows in the context of spin glasses is central to the study of *out-of-equilibrium statistical mechanics* [22, 50, 51]. However, studying a concrete mathematical background of these problems remains a long-standing challenge, but recent significant progress has been made through probability theory [14, 5, 4].

Neural networks have also been frequently used as computational tools in statistical mechanics, machine learning, and various other fields. Numerous researchers have been studying *shallow networks* in the asymptotic regime of large network size and a large number of iterations, using the CLT for the empirical distribution of such networks and

the [Law of Large Numbers \(LLN\)](#) which means the empirical loss fluctuates around its mean-field limit (population loss) [48, 52, 46, 32, 52, 16, 25, 20]. The weak convergence analysis for interacting particle systems is the crucial idea of mean-field analysis of neural networks, which suggests a line of research on the [Partial Differential Equation \(PDE\)](#) limits of SGD that study the concentration of the empirical measure. [2, 17, 40, 46, 48].

On the other hand, since tracking the full trajectory of SGD in high dimensions is impractical, we instead focus on the evolution of *summary statistics*, which are referred to as *macroscopic variables* in the statistical physics literature which is introduced by Saad and Sola (1995) [30, 29, 10, 47]. These macroscopic variables allow us to characterize the overall behavior of SGD by studying the joint distribution of *order parameters*. In training, SGD generally suggests two distinct phases: the *ballistic* phase, where the summary statistics change macroscopically, and the *diffusive* phase, where the summary statistics fluctuate on microscopic scales.

In the initial ballistic phases, the *effective dynamics* are well-approximated by an [Ordinary Differential Equation \(ODE\)](#), meaning that the summary statistics follow a deterministic scaling limit similar to the gradient flow of the population loss. This scaling limit holds when the learning rate is sufficiently small with respect to the dimension [30, 29, 10, 47, 54, 14, 19, 30]. However, an additional correction term appears in the dynamics in a critical scaling regime of the learning rate. Near the corresponding fixed points, Ben Arous, Gheissari, and Jagannath (2024) [7] showed that the dynamics transition into a diffusive regime, where macroscopic fluctuation dominates, such that the effective dynamics are studied by SDEs. The difference between ballistic and diffusive phases provides a refined understanding of how SGD updates in high-dimensional landscapes.

1.2 Contribution

This thesis focuses on the high-dimensional asymptotic theory of SGD, particularly in the context of supervised *single-layer neural networks*, or the so-called *single-index models*. We demonstrate that with a random initialization, DMFT does not adequately represent the diffusive dynamics of SGD when it fluctuates microscopically close to the fixed point of the associated ODE, despite its success in assessing the ballistic phase of training. Specifically, we show that DMFT is severely limited in handling the recovery problem since it does not consider the corrective terms that emerge at the high-dimensional scale.

The main contribution of this thesis, which is discussed in detail in Chapter 3, is the derivation of the effective dynamics of SGD for training supervised single-layer neural networks. We establish that in high-dimensional settings, SGD follows a trajectory different from that predicted by DMFT. Our analysis reveals that the effective dynamics require additional correction terms, which emerge naturally through a diffusion process. These correction terms include diffusive components that capture small fluctuations in the learning curve. Furthermore, we show that near the fixed points of the associated SDEs, the learning dynamics behaves like an [Ornstein-Uhlenbeck \(OU\)](#) process.

1.3 Organization

In Chapter 2, we begin with a brief overview of the DMFT, which is based on the statistical physics of disordered systems and developed using heuristics. Then there is an introduction to the distributional dynamics of shallow networks, using PDEs, and an examination of deterministic scaling limits through ODEs and a regularized SDE for SGD, known as homogeneous SGD, in Section 2.1. Section 2.2 will provide information on the geometric aspects of recovery by analyzing the sample complexity of spherical SGD and its associated information exponent. As the primary mathematical framework for the main results of this paper, Section 2.3 introduces the generalized formalism of effective dynamics, which serves as the basis for the rest of the thesis.

Chapter 3 discusses the thesis's primary contributions. The effective dynamics of online SGD for single-index models in the macroscopic and microscopic regimes are shown in Section 3.1, which starts the section. Section 3.2 then presents the proofs. In conclusion, Chapter 4 provides a more comprehensive view of the research findings, providing an overview of the key findings and discussing related research and possible future research topics.

Chapter 2

Background

2.1 Dynamical mean field theory

2.1.1 Gradient flow

Inference and the statistical physics of disordered systems are closely related in the *teacher-student scenario*. The concept is as follows: a *teacher* generates data using some hidden ground truth information and a probabilistic model, and then passes on that data to a *student*. The student must determine the original ground truth by relying only on the data and model provided. A similar setup is found in neural networks [28], where the teacher follows a specific rule to generate outputs, and the student attempts to learn this rule to generate the same outputs.

In this study, we use the teacher-student network (in other literature, it is called a supervised single-layer neural network, or single-index models), in which a neural network is trained using a training dataset. This is how it works: First, the teacher generates a set of hidden variables using an already established prior distribution, $X^* \in \mathbb{R}^N$. Based on a statistical model, the teacher then generates a set of **independent and identically distributed (i.i.d.)** data points, $(Y_k)_k$, using these hidden values. Parameterized by values in $\mathcal{Y}_N \subseteq \mathbb{R}^N$, this model belongs to a family of probability distributions and follows a likelihood function. The number of data points, $k \in \{1, \dots, N\}$. Finally, the teacher passes on the dataset $(Y_k)_k$ to the student, along with some information about the distribution. The student's goal is to use this information to estimate the hidden variables X^* .

Consider a supervised learning where a parameterized set of iterates $(X_k)_k$ belongs to $\mathcal{X}_N \subseteq \mathbb{R}^N$, training set $(Y_k)_k$ taking values in $\mathcal{Y}_N \subseteq \mathbb{R}^N$, and with a loss function $L_N : \mathcal{X}_N \times \mathcal{Y}_N \rightarrow \mathbb{R}$ via online SGD with the learning rate, δ_k , which is given by

$$X_{k+1} = X_k - \delta_k \nabla L(X_k; Y_{k+1}) \quad (2.1.1)$$

with a random initialization $X_0 \sim \mu_N \in \mathcal{M}_1(\mathbb{R}^N)$ where $\mathcal{M}_1(\mathbb{R})$ is the probability measures on the real line $\mathbb{R}$. Our interest is to understand this evolution, (X_k) , in the high-dimensional regime where $N \rightarrow \infty$.

Understanding how SGD evolves in non-convex loss landscapes is an important challenge in many topics, including machine learning, high-dimensional statistics, statistical physics, and applied mathematics. One approach to evaluate the structure of solutions and near-solutions in such problems is to study the following gradient flow in $\mathbb{R}^N$,

$$\frac{dX_t}{dt} = -\nabla L_N(X_t) \quad (2.1.2)$$

When we discretize the gradient flow, we obtain the gradient descent algorithm. The discretization process enables this algorithm to solve different optimization challenges. It is the foundation for SGD and is extensively useful in machine learning and computational areas. Specifically, the iterative updates follow the rule

$$X_{k+1} = X_k - \delta_k \nabla L_N(X_k) \quad (2.1.3)$$

The discrete updates closely follow gradient flow when the step size δ_k is sufficiently small, such as $\delta_k = c_k/N$, approximating the optimization dynamics effectively.

2.1.2 Distributional dynamics

To implement our approach, we consider the following model for supervised learning using one-layer neural networks. In a single-index model, the empirical estimation of the latent variable is given by

$$\hat{y}_N = \frac{1}{N} \sum_{k=1}^N f(\langle a_k, X_k \rangle) \quad (2.1.4)$$

For the sake of simplicity, we use the quadratic form of the loss functions. To this end, we consider as data the pairs $Y_k = (y_k, a_k)$ and our goal is then to optimize a quadratic loss L_N of the form

$$L_N(x, Y) = L(x, (y, a)) = (y - \hat{y}_N)^2 \quad (2.1.5)$$

Note that, typically, in supervised learning, one minimizes the empirical loss over the entire dataset; however, by using online SGD, the optimization process directly targets the population loss, rather than the empirical loss. This means that, instead of considering the average error across all data points in the dataset, online SGD iteratively updates the model by processing one data point at a time, using stochastic estimates of the gradient to minimize the expected loss over the underlying distribution of the data. This approach is particularly useful when the dataset is large or when data is arriving sequentially, as it allows for more efficient and scalable learning, i.e., the LLN and the functional CLT can be used to control the fluctuations of the empirical loss around its mean, population loss [48, 52].

The expected value of the squared difference between the true output and the predicted output gives the population loss for a quadratic loss function:

$$\Phi_N(x) = \mathbb{E}[L_N(x, Y)] = \mathbb{E}[(y - \hat{y}_N)^2] \quad (2.1.6)$$

Given a sample (a_k, y_k) , the weights are updated sequentially by online SGD,

$$X_{k+1} = X_k - \frac{2c_k}{N}(y - \hat{y}_N)\nabla_{X_k}f(\langle a_k, X_k \rangle) \quad (2.1.7)$$

A useful starting point is to express the population loss $\Phi_N(x)$ as:

$$\Phi_N(X) = \Phi_{\#} + \frac{2}{N} \sum_{k=1}^N V(X_k) + \frac{1}{N^2} \sum_{k,l=1}^N U(X_k, X_l) \quad (2.1.8)$$

where we defined the potentials

$$V(x) = -\mathbb{E}[yf(\langle a, x \rangle)], \quad \text{and} \quad U(x, x') = \mathbb{E}[f(\langle a, x \rangle)f(\langle a, x' \rangle)] \quad (2.1.9)$$

In particular, U is a symmetric positive semidefinite kernel and the constant $\Phi_{\#} = \mathbb{E}[y^2]$ is the loss of the trivial estimator.

Note that $\Phi_N(x)$ only depends on $(X_k)_k$ through their empirical measure

$$\hat{\rho}_N = \frac{1}{N} \sum_{k=1}^N \delta_{X_k} \quad (2.1.10)$$

This suggests considering a loss function defined for a given measure $\rho \in \mathcal{M}_1(\mathbb{R}^N)$:

$$\Phi(\rho) = \Phi_{\#} + 2 \int V(x)\rho(dx) + \int U(x_1, x_2)\rho(dx_1)\rho(dx_2) \quad (2.1.11)$$

One may establish a formal relationships between $\Phi_N(x)$ and $\Phi(\rho)$ using the concentration of the empirical measure with respect to the Wasserstein metric (see Appendix B), under mild regularization conditions such as the Lipschitz concentration, i.e., $\hat{\rho}_N \rightarrow \rho$ weakly. Specifically, we have:

$$\inf_x \Phi_N(x) = \inf_{\rho} \Phi(\rho) + O\left(\frac{1}{N}\right). \quad (2.1.12)$$

The weak convergence analysis for interacting particle systems is the crucial idea of the mean field analysis of neural networks. Thus, the dynamics conserves mass locally because particles cannot move discontinuously. Online SGD can be approximated by a continuum dynamics that enforces this *local mass conservation principle*. Denoting the empirical distribution of parameters after k updates by $\hat{\rho}_k^N = \frac{1}{N} \sum_{l=1}^N \delta_{X_k^l}$, [40] proved that $\hat{\rho}_{\lfloor tN \rfloor}^N \rightarrow \rho_t$ weakly as $N \rightarrow \infty$. Therefore, the asymptotic dynamics of ρ_t are controlled by the following PDEs, which we refer to as the *distributional dynamics*:

$$\frac{\partial \rho_t}{\partial t} = 2c_t \nabla \cdot (\rho_t \nabla \Psi(x; \rho_t)) \quad \text{where} \quad \Psi(x; \rho) = V(x) + \int U(x, x')\rho(dx') \quad (2.1.13)$$

We point out that the minimum of the asymptotic loss $\Phi(\rho)$ provides a good approximation of the minimum of the finite- N loss $\Phi_N(x)$. As N increases, the empirical distribution

of the model parameters $\hat{\rho}_N$ concentrates to the distribution ρ_t . The loss function $\Phi(\rho)$ reflects the model's true performance more accurately. This allows us to use the asymptotic loss $\Phi(\rho)$ to estimate the behavior of the finite- N system, which is useful when optimizing over the finite dataset is too complex.

Theorem 2.1 (Global vs. local minimum [40]). *Assume that either of the following conditions holds:*

1. $\inf_{\rho} \Phi(\rho)$ is achieved by a law ρ^* such that $\int U(x, x) \rho^*(dx) \leq K$;
2. There exists $\epsilon_0 > 0$ such that, for any $\rho \in \mathcal{M}_1(\mathbb{R}^N)$ such that $\Phi(\rho) \leq \inf_{\rho} \Phi(\rho) + \epsilon_0$ we have $\int U(x, x) \rho(dx) \leq K$

Then we have

$$|\inf_x \Phi_N(x) - \inf_{\rho} \Phi(\rho)| \leq \frac{K}{N} \quad (2.1.14)$$

Further, assume that $x \mapsto V(x)$ and $(x_1, x_2) \mapsto U(x_1, x_2)$ are continuous with U bounded below. A law ρ^* is a global minimum of Φ if

$$\inf_x \Psi(x; \rho^*) > -\infty \quad \text{and} \quad \text{supp}(\rho^*) \subseteq \arg \min_x \Psi(x; \rho^*) \quad (2.1.15)$$

We now turn our attention to the distributional dynamics. To demonstrate that the PDEs correctly describe the limiting behavior of the SGD dynamics as $N \rightarrow \infty$, we must first restrict our case to some assumptions. These assumptions are necessary to ensure that the dynamics of the empirical distribution of parameters, as updated by SGD, converge to the continuous distributional dynamics and that the approximations made in deriving the PDEs hold under these conditions. By making these assumptions, we can rigorously justify the connection between the discrete SGD updates and the continuous asymptotic dynamics.

Definition 2.2 (Lipschitz boundedness [40]). The following assumptions hold

1. $t \mapsto c_t$ is bounded Lipschitz:

$$\|c_t\|_{\infty}, \|c_t\|_{\text{Lip}} \leq K_1 \quad \text{with} \quad \int_0^{\infty} c(t) dt = \infty \quad (2.1.16)$$

2. The activation function $(a, x) \mapsto f(a, x)$ is bounded, with sub-Gaussian gradients (see Appendix A): $\|f\|_{\infty} \leq K_2$, $\|\nabla f(a, x)\|_{\psi_2} \leq K_2$.
3. The gradients $x \mapsto \nabla V(x)$ and $(x, x') \mapsto \nabla U(x, x')$ are bounded, Lipschitz continuous, namely

$$\|\nabla V(x)\|_2 \leq K_3, \quad \|\nabla U(x, x')\| \leq K_3 \quad (2.1.17)$$

$$\|\nabla V(x) - \nabla V(x')\|_2 \leq K_3 \|x - x'\|_2, \quad (2.1.18)$$

$$\|\nabla U(x_1, x_2) - \nabla U(x'_1, x'_2)\|_2 \leq K_3 \|(x_1, x_2) - (x'_1, x'_2)\|_2 \quad (2.1.19)$$

We also introduce the following error term, which quantifies the accuracy of our PDE model in a non-asymptotic case. This term allows us to measure the difference between the finite- N dynamics and the continuous distributional dynamics, providing a more precise understanding of how well the PDE approximation studies the behavior of the model, even for finite N ,

$$\text{err}_N(z) = \sqrt{\min\{\delta, 1/N\}} \left(\sqrt{N + \log N/\delta} + z \right) \quad (2.1.20)$$

The convergence of SGD to the PDE model is an example of a phenomenon known in probability theory as *propagation of chaos* [53].

Theorem 2.3 (Distributional dynamics [40]). *Assume that Assumption 2.2 holds. For $\rho_0 \in \mathcal{M}_1(\mathbb{R}^N)$. consider SGD with initialization $X_0 \sim \rho_0$ and step size δ . For $t \geq 0$, let ρ_t be the solution of PDE*

$$\frac{\partial \rho_t}{\partial t} = 2c_t \nabla \cdot (\rho_t \nabla \Psi(x; \rho_t)) \quad \text{where} \quad \Psi(x; \rho) = V(x) + \int U(x, x') \rho(dx') \quad (2.1.21)$$

Then, for any fixed $t \geq 0$, $\hat{\rho}_{[t/\delta]}^N \rightarrow \rho_t$ almost surely along any sequence (N, δ) such that $N \rightarrow \infty$, $\delta \rightarrow 0$, $N/\log(N/\delta) \rightarrow \infty$ and $\delta \log(N/\delta) \rightarrow 0$. Further, there exists a constant C (independent of N) such that for any 1-Lipschitz $f: \mathbb{R}^N \times \mathbb{R} \rightarrow \mathbb{R}$ and $\delta \leq 1$,

$$\sup_{k \in [0, T/\delta] \cap \mathbb{N}} \left| \frac{1}{N} \sum_{i=1}^N f(X_k^i) - \int f(x) \rho_{k\delta}(dx) \right| \leq C e^{CT} \text{err}_N(z) \quad (2.1.22)$$

$$\sup_{k \in [0, T/\delta] \cap \mathbb{N}} |\Phi_N(X_k) - \Phi(\rho_{k\delta})| \leq C e^{CT} \text{err}_N(z). \quad (2.1.23)$$

with probability $1 - e^{-z^2}$.

The model can only be applied in cases where the distributional dynamics converge quickly to a truth solution, as it shows that the error increases exponentially with the time horizon T . According to this exponential growth in error, the approximation becomes less accurate for longer time horizons, which makes it ideal for cases in which the dynamics stabilize or converge to an optimal solution in a relatively short time.

We can interpret $J(x; \rho_t) = \rho_t(x) \nabla \Psi(x; \rho_t)$ as a *current*. The fixed points of the continuum dynamics correspond to densities where the current is zero, meaning there is no net flow or change in the distribution. These fixed points represent stable solutions of the system, where the dynamics have reached equilibrium, and the current vanishes, indicating no further movement in the system's parameters. This condition is crucial for identifying the steady states of the system.

Corollary 2.4 (Steady state [40]). *Assume that V and U are differentiable with bounded gradient. If ρ_t is a solution of Theorem 2.3, then $\Phi(\rho_t)$ is non-increasing. Furthermore, law ρ is a fixed-point of the PDE of Theorem 2.3 if and only if*

$$\text{supp}(\rho) \subseteq \{x : \nabla \Psi(x; \rho) = 0\} \quad (2.1.24)$$

2.1.3 Deterministic effective dynamics

Single-index models

In a different, and older, line of research, online SGD for single-index models with a finite number of hidden units, synthetic Gaussian input data, and teacher-generated labels has been extensively studied, starting with the seminal work of Saad and Sola (1995) [47]. These studies focus on the limit of high-dimensional data and show that the stochastic process driven by gradient updates converges to a set of deterministic ODEs as the dimension grows $N \rightarrow \infty$, with the learning rate being proportional to $1/N$. This result provides a connection between the stochastic behavior of SGD and its deterministic limit, offering insights into the dynamics of neural network training in high-dimensional settings.

Suppose we are given a possibly non-linear activation function $f : \mathbb{R} \rightarrow \mathbb{R}$, a set of feature vectors $(a_k)_{k=1}^N$, and additive noisy responses of the form:

$$y_k = f(\langle a_k, X^* \rangle) + \epsilon_k \quad (2.1.25)$$

Given the teacher's data, the student's goal is to estimate X^* .

Recall the quadratic form of the single-index models, where the network output is determined by a composition of linear transformations followed by a nonlinear function. Precisely, using online SGD with a constant step-size $\delta_N = c_\delta/N$,

$$X_{k+1} = X_k - \frac{c_\delta}{N} \nabla L_N(X_k; Y_{k+1}) \quad (2.1.26)$$

where

$$L_N(x, (y, a)) = (f(\langle a, X^* \rangle) - f(\langle a, x \rangle))^2 \quad (2.1.27)$$

In this context, we define $\lambda_k = \langle a_k, X_k \rangle$ and $\lambda_k^* = \langle a_k, X^* \rangle$ as the local field vectors for the student and the teacher, respectively. These vectors represent the inner product between the input feature vectors a_k and the corresponding parameters X_k for the student, and X^* for the teacher,

$$X_{k+1} = X_k - \frac{c_\delta}{N} a_k f'(\lambda_k) \mathcal{E}_k \quad \text{where} \quad \mathcal{E}_k = f(\lambda_k) - f(\lambda_k^*) \quad (2.1.28)$$

To track the population loss along the dynamics, it is sufficient to track the joint distribution of the local fields (λ, λ^*) . Let us focus on the most studied regime, where (a_l) are i.i.d. standard Gaussian vectors in $\mathbb{R}^N$. In this setting, one can replace expectation $E_{x,y}[\cdot]$ with $\mathbb{E}_{\lambda, \lambda^*}[\cdot]$ and completely describe the dynamics using the following summary statistics, which are termed macroscopic variables in the statistical physics literature [30, 29, 10, 47]. These macroscopic variables allow us to characterize the system's overall behavior by capturing essential features of the joint distribution of the local fields, such as the mean and variance, which in turn provide insight into the dynamics of the learning process. In other words, one may define

$$R_k = \mathbb{E}[\lambda_k^2] = \frac{1}{N} \langle X_k, X_k \rangle, \quad m_k = \mathbb{E}[\lambda_k \lambda_k^*] = \frac{1}{N} \langle X_k, X^* \rangle, \quad P = \mathbb{E}[\lambda_k^{*2}] = \frac{1}{N} \langle X^*, X^* \rangle \quad (2.1.29)$$

Where the matrix elements are referred to as order parameters. The macroscopic state of the system at the learning step k is given by the overlap matrix, which captures the correlation between the student's and teacher's parameters at each step of the learning process.

To describe the dynamics of the system, one may use the overlap matrix, particularly in understanding how the student's estimates of the parameters evolve towards the true values as the learning process proceeds

$$\Omega_k = \begin{pmatrix} R_k & m_k \\ m_k^T & P \end{pmatrix} \quad (2.1.30)$$

and the population loss is completely determined by the macroscopic state, as it depends on the distribution of the local fields and the overlap matrix.

$$\Phi_N(x) = \mathbb{E}_{\lambda, \lambda^* \sim \mathcal{N}(0, \Omega)} \mathbb{E}_\epsilon [(f(\lambda) - f(\lambda^*))^2] \quad (2.1.31)$$

The training dynamics defines a discrete-time stochastic process for the evolution of the overlap matrix

$$\{\Omega_k \in \mathbb{R}^{(N+N) \times (N+N)}, k = 1, 2, \dots, N\} \quad (2.1.32)$$

with P fixed and R_k and m_k updated as

$$\begin{cases} R_{k+1} - R_k = -4c_\delta f'(\lambda_k) \mathcal{E}_k \lambda_k + 4c_\delta^2 f'^2(\lambda_k) (\mathcal{E}_k^2 + C_\epsilon) \\ m_{k+1} - m_k = -2c_\delta f'(\lambda_k) \mathcal{E}_k \lambda^* \end{cases} \quad (2.1.33)$$

The first term, referred to as the *learning* term, drives the student's learning process from the teacher. This term has been shown to correspond to the gradient flow approximation of SGD. This term pushes the population loss towards zero, helping the student converge to the true model. However, in this setting, there is an additional second term, called the *variance* term, which makes the flow approximation incomplete. This term corresponds to the fluctuations of the loss function around its expected value, i.e., $H(x) = L(x, Y) - \Phi(x)$, and will be discussed in the following sections. This variance term accounts for the noise in the learning process, reflecting the randomness in the updates and their effect on the overall dynamics.

In [54], as $N \rightarrow \infty$, Ω_k converges to the solution of the ODE, namely $\bar{\Omega}$ as follows

$$\frac{d\bar{\Omega}(t)}{dt} = \Psi(\bar{\Omega}(t)) \quad (2.1.34)$$

where $\Psi : \mathbb{R}^{(N+1) \times (N+1)} \rightarrow \mathbb{R}^{(N+1) \times (N+1)}$ is the expected value of the RHS of Eqs 2.1.32 provided that this solution stays bounded. Then the following theorem provides a deterministic scaling limit for single-index models:

Theorem 2.5 (Deterministic scaling limit [54]). *Let $\tilde{\tau}$ be the continuous time horizon and τ be a time scaling factor such that the following assumption holds:*

1. The time scaling τ satisfies for some constant c ,

$$\tau \geq c \max(\delta_N, N\delta_N^2) \quad (2.1.35)$$

2. The activation function f is L -Lipschitz,

3. The function Ψ is L' -Lipschitz.

Then, there exists a constant $C > 0$ (independent of N) such that for any $0 \leq d \leq \lfloor \tilde{\tau}/\tau \rfloor$, the following inequality holds

$$\mathbb{E}\|\Omega_k - \bar{\Omega}(d\tau)\|_\infty \leq e^{C\tilde{\tau}} \sqrt{\tau} \log N \quad (2.1.36)$$

Theorem 2.5 enables us to track the convergence rate of the discrete process to a set of ODEs. Furthermore, since the additive errors are centered, the linear contribution in $\mathcal{E}$ becomes noiseless in the high-dimensional limit. As a result, we can refer to this as the *learning term* (later we will call this *population drift*), which represents the part of the dynamics that drives the system towards the correct solution. On the other hand, the noise enters the equations through the variance computed on the quadratic form of $\mathcal{E}^2$, which we will refer to as the *noise term* (later we will call this *population corrector*).

Intuitively, this term behaves as a high-dimensional variance correction, introducing fluctuations that hinder the learning process. This noise term captures the effects of random fluctuations in the learning dynamics, which can slow down or interfere with the convergence to the optimal solution, especially in large system sizes.

Using such a convergence, one may obtain an ODE approximation for online SGD on the teacher-student network as follows:

Corollary 2.6 (ODEs for the order parameters [54]). *If one may choose $\delta_N = c_\delta/N$ for some constant c_δ , then the above discrete dynamics turn into a coupled ODEs as follows,*

$$\frac{dR_t}{dt} = -4c_\delta \mathbb{E}_{\lambda, \lambda^*} [f'(\lambda) \mathcal{E} \lambda] + 4c_\delta \mathbb{E}_{\lambda, \lambda^*} [f'^2(\lambda) (\mathcal{E}^2 + C_\epsilon)] \quad \text{and} \quad \frac{dm_t}{dt} = -2\mathbb{E}_{\lambda, \lambda^*} [f'(\lambda) \mathcal{E} \lambda^*] \quad (2.1.37)$$

Generalized formalism

In addition to the deterministic ODEs for SGD on the teacher-student network, [19] generalized the deterministic scaling limits of SGD via ODEs and also introduced an SDE with a simplified diffusion coefficient, namely homogenized SGD. This framework allows us to analyze the dynamics of SGD's summary statistics.

Our objective is to optimize the loss function L_N through online SGD, iteratively updating the parameters in response to the data,

$$X_{k+1} = X_k - \frac{c_k}{N} \nabla L_N(X_k, Y_{k+1}) \quad (2.1.38)$$

We will work with a formulation where the norms of the iterates $\{X_k\}$ stay bounded, regardless of the dimension. To ensure this, we make the following initialization assumptions. These assumptions are essential for maintaining stability in the iterates X_k and for making the learning process feasible to analyze, especially in high-dimensional cases.

Definition 2.7 (Lipschitz boundedness [19]). The following conditions hold:

1. The initialization point, X_0 and the hidden parameter X^* are bounded independent of N , i.e.,

$$\max\{\|X^*\|, \|X_0\|\} \leq C \quad (2.1.39)$$

for some $C > 0$ independent of N .

2. We assume that $a \sim \mathcal{N}(0, K)$ and $\epsilon \sim \mathcal{N}(0, I_N)$ with covariance $K \in \mathbb{R}^{N \times N}$, which are bounded in operator norm independent of N .
3. Now we need to connect the gradients of the loss to the gradient estimators in SGD. Hence, we assume the following Lipschitzness for the population loss and the *Fisher matrix* $G_N(x) = \mathbb{E}[\nabla L_N(x, Y) \otimes \nabla L_N(x, Y)]$. There is an open set $\mathcal{U} \in \mathbb{R}^{N \times N}$ such that

$$(X_0 \otimes X^*)^T K \in \mathcal{U} \quad (2.1.40)$$

and the population loss is continuously differentiable on $\mathcal{U}$ and its derivative $\nabla \Phi_N$ is α -pseudo-Lipschitz, i.e., there is a constant $\ell_\Phi > 0$, so that for all $B, \hat{B} \in \mathcal{U}$,

$$\|\nabla \Phi_N(B) - \nabla \Phi_N(\hat{B})\| \leq \ell_\Phi \|B - \hat{B}\| (1 + \|B\|^\alpha + \|\hat{B}\|^\alpha) \quad (2.1.41)$$

Moreover, the Fisher matrix $G_N(x)$ is α -pseudo-Lipschitz.

We define a coupled family of ODEs to describe the scaling limit. These ODEs need to be sufficiently complex for studying the dynamics of the overlap matrix, which is key to the Fisher matrix G_N and the loss function L_N . We aim to give the overlap matrix a high-dimensional limit so that the ODEs can effectively represent how it changes as the dimension grows. This method allows for a more accurate understanding of the learning dynamics in the large-dimensional limit, where, as we already mentioned, the interactions between the parameters and the overlap matrix's structure are critical to the system's behavior. One may define the corresponding overlap matrices,

$$\Omega_k(X_k) = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} = \begin{pmatrix} (X_k \otimes X_k)^T K & (X_k \otimes X^*)^T K \\ K^T (X_k \otimes X^*) & (X^* \otimes X^*)^T K \end{pmatrix} \quad (2.1.42)$$

The limit variables evolve continuously over time and will be defined by an average over a family of N -dimensional limit variables. Let $\{(\lambda_i, \omega_i) : i \in 1, \dots, N\}$ represent the eigenvalues and orthonormal eigenvectors of the matrix K . We then introduce the following set of ODEs that ensures the limits of positive semidefinite matrices

$$\bar{\Omega}(t) = \begin{pmatrix} \bar{\Omega}_{11}(t) & \bar{\Omega}_{12}(t) \\ \bar{\Omega}_{12}^T(t) & \bar{\Omega}_{22}(t) \end{pmatrix} \quad \text{and} \quad \bar{\Omega}_i(t) = \begin{pmatrix} \bar{\Omega}_{11,i}(t) & \bar{\Omega}_{12,i}(t) \\ \bar{\Omega}_{12,i}^T(t) & \bar{\Omega}_{22,i}(t) \end{pmatrix} \quad (2.1.43)$$

These terms are then connected by averaging over $i \in \{1, \dots, N\}$. Additionally, at time t , we introduce a secondary average to capture the behavior of the system over the evolving time scale,

$$\bar{\Omega}(t) = \frac{1}{N} \sum_{i=1}^N \lambda_i \bar{\Omega}_i(t) \quad \text{and} \quad \omega(t) = \frac{1}{N} \sum_{i=1}^N \text{tr}(\bar{\Omega}_i(t)) \quad (2.1.44)$$

This secondary averaging process helps smooth out fluctuations and provides a more precise description of the dynamics, ensuring that the system's behavior is well-approximated as it evolves. By combining these two types of averages, we can better understand the collective properties of the system and how they evolve in the high-dimensional limit.

Now if the population loss Φ_N be defined symmetrically with respect to Ω , so that one may define

$$\nabla \Phi_N(\bar{\Omega}(t)) = \begin{pmatrix} \frac{\partial \phi_N}{\partial (X^T X)_{ij}} & \frac{\partial \phi_N}{\partial (X^T X^*)_{ij}} \\ \frac{\partial \phi_N}{\partial (X^T X^*)_{ji}} & \frac{\partial \phi_N}{\partial (X^{*T} X^*)} \end{pmatrix} = \begin{pmatrix} H_1 & H_2 \\ H_2^T & H_3 \end{pmatrix} \quad (2.1.45)$$

The following theorem provides a deterministic scaling limit that allows us to analyze a family of coupled ODEs, which can approximate the dynamics of online SGD for a well-defined class of non-convex loss functions.

Theorem 2.8 (Determinist scaling limit [19]). *Suppose Assumptions 2.7 hold. Let $\tilde{\tau}$ be the first time that either $\bar{\Omega}(t)$ or $\Omega(X_{[tN]})$ exists $\mathcal{U}$ or that $\omega(t) \geq \omega_0$. Then there is an $\epsilon > 0$ so that for any T, ω_0 , with overwhelming probability*

$$\sup_{0 \leq t \leq \min\{\tilde{\tau}, T\}} \|\bar{\Omega}(t) - \Omega(X_{[tN]})\| \leq N^{-\epsilon} \quad (2.1.46)$$

Moreover, the above discrete dynamics turn into a coupled ODEs as follows,

$$\frac{d\bar{\Omega}_{11,i}(t)}{dt} = -2\lambda_i c_t (\bar{\Omega}_{11,i}(t) H_1(t) + H_1(t) \bar{\Omega}_{11,i}(t) + \bar{\Omega}_{12,i}(t) H_2(t) + \lambda_i c_t^2 G_t) \quad (2.1.47)$$

$$\frac{d\bar{\Omega}_{12,i}(t)}{dt} = -2\lambda_i c_t (H_1(t) \bar{\Omega}_{12,i}(t) + H_2^T(t) \bar{\Omega}_{22,i}(t)) \quad (2.1.48)$$

We also note that in the case of identity covariance, the system simplifies dramatically: as all $\lambda_i = 1$, we may directly take the average on both sides.

Corollary 2.9 (Deterministic scaling limit in identity covariance [19]). *If we suppose that $K = I_N$, then $\bar{\Omega}(t)$ solves the autonomous equation*

$$\frac{d\bar{\Omega}_{11}(t)}{dt} = -2c_t (\bar{\Omega}_{11}(t) H_1(t) + H_1(t) \bar{\Omega}_{11}(t) + \bar{\Omega}_{12}(t) H_2(t)) + c_t^2 G_t \quad (2.1.49)$$

$$\frac{d\bar{\Omega}_{12}(t)}{dt} = -2c_t (H_1(t) \bar{\Omega}_{12}(t) + H_2^T(t) \bar{\Omega}_{22}(t)) \quad (2.1.50)$$

As the number of equations grows proportionately to the dimensionality of the matrix K , this system of ODEs evolves more complicated with the dimension.

We can also restate Theorem 2.8 for a special case where the covariance matrix $K = I_N$ and the loss function follows the single-index model. In this case, the overlap matrix Ω_k is given by:

$$\Omega_k = \begin{pmatrix} R_k & m_k \\ m_k^T & P \end{pmatrix} \quad \text{where} \quad R_k = \langle X_k, X_k \rangle, \quad m_k = \langle X_k, X^* \rangle, \quad P = \langle X^*, X^* \rangle \quad (2.1.51)$$

The coupled ODEs then simplify to the following proposition:

Proposition 2.10. *Suppose that the conditions of Theorem 2.8 hold. Let $(X_k)_k$ be SGD initialized from $X_0 \sim \mu_N \in \mathcal{M}_1(\mathbb{R}^N)$ with learning rate $\delta_N = 1/N$. If the quadratic loss L_N corresponds to supervised single-index models with the covariance matrix $K = I_N$, then the corresponding summary statistics $u_N = (R, m)$ converges as $N \rightarrow \infty$ to the solution of the following ODE initialized from $\lim_{N \rightarrow \infty} (u_N)_* \mu_N$*

$$\frac{dR_t}{dt} = -\langle \nabla \Phi_N, \nabla R_t \rangle + \lim_{N \rightarrow \infty} \frac{1}{N} \text{tr}(G_t) \quad \text{and} \quad \frac{dm_t}{dt} = -\langle \nabla \Phi_N, \nabla m_t \rangle \quad (2.1.52)$$

This confirms that the deterministic dynamics discussed in [19] follow the deterministic scaling limit mentioned earlier [54].

2.2 Sample complexity

2.2.1 Search and descent phases

Can one prove sample complexity bounds for SGD that are polynomial in N ? What fraction of data is used/time is spent in the search phase vs. the descent phase? How do these answers change as one varies the loss function, $L_N(x, Y)$?

The first question explores whether it's possible to establish sample complexity bounds for SGD that scale polynomially with respect to the number of samples N . This would give insights into how the algorithm's performance depends on the amount of data and whether it scales efficiently as the number of samples increases.

The second question focuses on understanding how much time or data is spent in the initial search phase of the SGD algorithm, where it explores the loss landscape, compared to the later descent phase, where the algorithm efficiently converges towards a local minimum. This question aims to quantify the different stages of the optimization process.

The latter question looks into how the answers to the previous two questions change when the form of the loss function changes. Specifically, it studies whether the behavior of SGD (e.g., the balance between search and descent phases or the sample complexity) depends on the nature of the loss function used in the optimization problem.

For the sake of simplicity, assume that our parameter space is spherical, specifically $\mathbb{S}^{N-1}$, and we will consider a spherical version of online SGD. Note that hereafter, the number of data points, $k \in \{1, \dots, M\}$ is set as $M = \alpha_N N$, where α_N is a positive constant, however, in the previous sections $\alpha_N = 1$, that is $N = M$.

Consider a supervised learning where a parameterized set of iterates $(X_k)_k$ belongs to $\mathcal{X}_N \subseteq \mathbb{R}^N$, training set $(Y_k)_k$ taking values in $\mathcal{Y}_N \subseteq \mathbb{R}^N$, and with a loss function $L_N : \mathcal{X}_N \times \mathcal{Y}_N \rightarrow \mathbb{R}$ via the spherical version of online SGD with the learning rate, δ_k , which is given by

$$\begin{cases} \tilde{X}_k = X_{k-1} - \frac{c\delta}{N} \nabla L_N(X_{k-1}; Y_k) \\ X_k = \frac{\tilde{X}_k}{\|\tilde{X}_k\|} \end{cases} \quad (2.2.1)$$

with $X_0 \sim \mu \in \mathcal{M}_1(\mathbb{S}^{N-1})$, and where ∇ denotes the spherical gradient. We take the output of this algorithm as our estimator. To ensure that the algorithm is well-defined, we assume that the loss function is almost surely differentiable with respect to the parameter for all $X \in \mathbb{S}^{N-1}$.

One of our main goals is to understand how the proportion of time SGD spends in the search phase versus the descent phase depends on the dimension. To achieve this, we need to define the search phase formally. We consider a simple setting where the population loss is a (possibly non-linear) $f : [0, \infty) \rightarrow \mathbb{R}$ of the distance $d : \mathbb{S}^{N-1} \times \mathbb{S}^{N-1} \rightarrow [0, \infty)$ between parameters. Specifically, the population loss is defined as $\Phi(x) = f(d(x, x_0))$, where x_0 is the initial parameter.

For simplicity, we focus on the case where the population loss is of the form $\Phi_N(x) = \phi(m_N(x))$, where $m_N(x)$ is a correlation term often referred to as the latitude. The set where $m_N(x) \approx 0$ is called the equator of the sphere. This terminology arises from the fact that with random initialization $X_0 \sim \mu_N = \text{Unif}(\mathbb{S}^{N-1})$, we have $m_N(X_0) \approx \frac{1}{\sqrt{N}}$. According to the Poincaré limit, it follows that $\sqrt{N}m_N(X_0) \rightarrow \mathcal{N}(0, 1)$ as $N \rightarrow \infty$.

To formalize the notion of exiting the search phase, we recall the notion of weak recovery, i.e., achieving macroscopic correlation with X^* .

Definition 2.11 (Weak Recovery). A sequence of estimators $\hat{X}^* \in \mathbb{S}^{N-1}$ is said to weakly recover (escape equator) the parameter X^* if, for some $\eta > 0$, it holds that

$$\lim_{N \rightarrow \infty} P\left(m_N(\hat{X}^*) \geq \eta\right) = 1. \quad (2.2.2)$$

To understand this scaling here, i.e., $m_N(\hat{X}^*) = o(1)$, recall the basic fact from high-dimensional probability [55] that if $\hat{X}^*$ were drawn uniformly at random, the choice is exponentially small in N . In this context, attaining weak recovery corresponds to exiting the search phase. Indeed, in this context, attaining weak recovery corresponds to the descent phase.

Definition 2.12 (Strong Recovery). A sequence of estimators $\hat{X}^* \in \mathbb{S}^{N-1}$ is said to strongly recover (consistent estimation) the parameter X^* if it holds that

$$m_N(\hat{X}^*) \rightarrow 1, \quad \text{in probability (or equivalently in } L^p\text{-norm for every } p \geq 1). \quad (2.2.3)$$

To investigate the performance of SGD in high-dimensional inference tasks, it is essential to choose an appropriate scaling for the fluctuations of the loss. These fluctuations are represented by the *sample-wise error*, defined as

$$H_N(x) = L_N(x, Y) - \Phi_N(x). \quad (2.2.4)$$

2.2.2 Information exponent

We aim for a scaling regime that avoids the issues outlined in the introduction. Consequently, we operate under an assumption that holds for the loss functions of many high-dimensional problems.

Definition 2.13 (Regularization and Fisher consistency [6]). The following assumptions hold:

1. There exists $C, \epsilon > 0$ such that the following two moment bounds hold for all N , $\sup_{x \in \mathbb{S}^{N-1}} \mathbb{E}[(\langle \nabla H_N(x), X^* \rangle)^2] \leq C$ and $\sup_{x \in \mathbb{S}^{N-1}} \mathbb{E}[\|\nabla H_N(x)\|^{4+\epsilon}] \leq CN^{\frac{4+\epsilon}{2}}$
2. ϕ is differentiable and $\phi'(t) < 0$ for all $t \in (0, 1)$.

Remark 2.14. The scaling relation above is tight when ∇H_N is an i.i.d. sub-Gaussian vector. i.e.,

- $\mathbb{E}[(\langle \nabla H_N(x), X^* \rangle)^2] \leq C$ for some $C > 0$ and all $X^* \in \mathbb{S}^{N-1}$.
- $\mathbb{E}[\|\nabla H_N(x)\|^{4+\epsilon}] \leq \sqrt{\mathbb{E}[\sum_i (\partial_i H_N(x))^2]^{\frac{4+\epsilon}{2}}} = O\left(N^{\frac{4+\epsilon}{2}}\right)$ by Jensen's inequality.

As we examine a first-order method, it is reasonable to expect that the output of gradient flow serves as a Fisher-consistent estimator for all random initial data with positive correlation, implying consistency when assessed on the population loss. Additionally, if ϕ is an even function, then $\phi'(0) = 0$.

Remark 2.15. The second part of the above assumption holds if and only if gradient flow for the population loss eventually produces a consistent estimator, when started anywhere on the upper half-sphere $\{x : m_N(x) > 0\}$.

In the following, we show that a key quantity governing the performance of online SGD is the notion, which we call the *information exponent* for a population loss.

Definition 2.16 (Information exponent). We say that a population loss Φ_N has information exponent k if $\phi \in C^{k+1}([-1, 1])$ and there exist $C, c > 0$ such that

$$\begin{cases} \frac{d^l \phi}{dm^l}(0) = 0 & 1 \leq l < k \\ \frac{d^k \phi}{dm^k}(0) \leq -c < 0 \\ \left\| \frac{d^{k+1} \phi}{dm^{k+1}}(m) \right\|_\infty \leq C \end{cases} \quad (2.2.5)$$

The first theorem establishes an upper bound on the sample complexity for consistent estimation. Specifically, we define

$$\alpha_c(N, k) = \begin{cases} 1 & k = 1 \\ \log N & k = 2 \\ N^{k-2} & k \geq 3 \end{cases} \quad (2.2.6)$$

Also, a sequence x_n is denoted as $x_n \ll y_n$ if $\frac{x_n}{y_n} \rightarrow 0$.

For clarity, we present the following result under the case of random initialization, where μ_N is the uniform measure conditioned on the upper half sphere $\{m(x) \geq 0\}$. This conditioning is made without loss of generality up to a probability of 1/2 and is implemented to mitigate symmetry issues.

Theorem 2.17 (Search phase [6]). *Suppose that Assumption 2.13 holds and the population loss has information exponent k . Let $M = \alpha_N N$ with α_N growing at most polynomial in N , namely $\alpha_N = \text{Poly}(N)$. If (α_N, δ_N) satisfy $\alpha_N^{-1} \ll \delta_N \ll \alpha_N^{-1/2}$ and*

$$\alpha_N \gg \begin{cases} \alpha_c & k = 1 \\ \alpha_c \log N & k = 2 \\ \alpha_c (\log N)^2 & k \geq 3 \end{cases} \quad (2.2.7)$$

then online SGD with the learning rate δ_N started from $X_0 \sim \mu_N$, will have

$$m_N(X_M) \rightarrow 1, \quad \text{in probability and } L^p\text{-norm for every } p \geq 1. \quad (2.2.8)$$

The second theorem provides the corresponding lower bound on the sample complexity required for exiting the search phase, depending on the information exponent.

Theorem 2.18 (Descent phase [6]). *Suppose that Assumption 2.13 holds and the population loss has information exponent k . If (α_N, δ_N) are such that*

- $(k \leq 2) : \alpha_N \ll \alpha_c$, and $\delta_N = O(1)$
- $(k \geq 3) : \alpha_N \ll \alpha_c$, and $\delta_N = O(\alpha_N^{-1/2})$,

then the online SGD with the learning rate δ_N , started from $X_0 \sim \mu_N$, will have

$$\sup_{t \leq M} |m_N(X_t)| \rightarrow 0, \quad \text{in probability and } L^p\text{-norm for every } p \geq 1. \quad (2.2.9)$$

The final argument shows that the ratio of the number of samples used in the descent phase to those used in the search phase is $O(\alpha_c(N, k)^{-1})$, which becomes negligible as $k \geq 2$. As the dimension increases, the algorithm spends an increasingly small fraction of time in the search phase compared to the descent phase. This result follows from Theorem 2.18 and is further supported by the underlying analysis, which considers the algorithm's behavior over time and the specific characteristics of the population loss function.

Theorem 2.19 ([6]). Let τ_η^+ denote the first time t such that $m_N(X_t) > \eta$ (exit time), and $\tau_{1-\eta}^+$ the first time t such that $m_N(X_t) > 1 - \eta$. Assume that Assumption 2.13 holds and the population loss has information exponent $k \geq 2$.

Define $M = \alpha_N N$ with (α_N, δ_N) as in Theorem 2.17. For any $\eta > 0$, there exists a constant $C = C(k, N) > 0$ such that:

$$\tau_\eta^+ \gg \alpha_c(N, k), \quad (2.2.10)$$

$$|\tau_{1-\eta}^+ - \tau_\eta^+| \leq CN \quad \text{with probability } 1 - o(1). \quad (2.2.11)$$

Additionally, for all $\tau_{1-\eta}^+ \leq t \leq M$,

$$X_t > 1 - 2\eta \quad \text{with probability } 1 - o(1). \quad (2.2.12)$$

In another word, Theorem 2.19 says that most of the data is used in the search phase (i.e. to attain some non-trivial correlation), and that from there descent to essentially full correlation is rapid, and takes $O(N)$ samples independently of the class of the problem.

2.2.3 Single-index models

Similar to previous sections, to be concrete and implement our approach, we consider the following model for supervised learning using single-index models. Suppose that we are given some activation function $f : \mathbb{R} \rightarrow \mathbb{R}$, some feature vectors $(a_k)_{k=1}^M$, and with these, M noisy responses of the form

$$y_k = f(\langle a_k, X^* \rangle) + \epsilon_k \quad (2.2.13)$$

where $(\epsilon_k)_k$ are additive errors. Our goal is to estimate X^* given this data. Our approach is to minimize the least squares error. Consider as data the pairs $Y_k = (y_k, a_k)$ for $k = 1, \dots, M$. Using student's model, $\hat{y}(x) = f(\langle a, x \rangle)$, one may write a quadratic loss L of the form,

$$L(x, Y) = L(x, (y, a)) = (y - f(\langle a, X^* \rangle))^2 \quad (2.2.14)$$

Note that we may write the population loss as

$$\Phi(x) = \mathbb{E} [(f(\langle a, x \rangle) - f(\langle a, X^* \rangle))^2] + \mathbb{E}[\epsilon^2] \quad (2.2.15)$$

As we mentioned above, let us focus on the most studied regime, namely where $(a_k)_k$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^N$; for the (ϵ_k) we only assume they are i.i.d. mean zero with finite $4 + \delta$ -th moment for some $\delta > 0$.

Recall that the Hermite polynomials, which we denote by $(h_k(x))_{k=0}^\infty$, are the normalized orthogonal polynomials of the Gaussian distribution $\gamma(x) \propto \exp(-x^2/2)dx$. Define the k -th Hermite coefficient for an activation function $f \in L^2(\gamma)$ by

$$u_k(f) = \langle f, h_k \rangle_{L^2(\gamma)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(z) h_k(z) e^{-z^2/2} dz \quad (2.2.16)$$

As long as f' is at most polynomial growth, the population loss is differentiable, and the above exists.

Theorem 2.20 ([6]). *Suppose that the features (a_k) are i.i.d. standard Gaussian vectors, and the errors (ϵ_k) are i.i.d. mean zero, variance C_ϵ and with finite $4 + \delta$ -th moment for some $\delta > 0$. Suppose that the activation function, f , is differentiable almost everywhere and that f' has at most polynomial growth. Then*

$$\Phi(x) = \phi_f(m(x)), \quad \text{where} \quad \phi_f(m) = 2 \sum_{j=0}^{\infty} (u_j(f))^2 (1 - m^j) + C_\epsilon \quad (2.2.17)$$

and where u_j are the Hermite coefficients. Furthermore, Assumption 2.13 holds.

Determining the information exponents for various activation functions becomes straightforward with this theorem. By analyzing the Taylor expansion of the population loss around the equator, one can extract the degree of the first nonzero term, which directly gives the information exponent. The following result follows immediately: for standard activation functions such as ReLU, sigmoid, or polynomial activations, the information exponent can be explicitly computed, leading to direct implications for the sample complexity and phase transitions in the learning dynamics of SGD.

Corollary 2.21 ([6]). *Suppose the conditions of Theorem 2.20 hold. Then,*

1. *The information exponent of f is 1 if and only if $u_1(f) \neq 0$.*
2. *The information exponent of f is 2 if and only if $u_1(f) = 0$ and $u_2(f) \neq 0$.*
3. *The information exponent of f is at least 3 if and only if $u_1(f) = u_2(f) = 0$.*

2.3 Effective dynamics

Consider a supervised learning with training set $\{(X_k), (Y_k)\}_{k=1}^M$ taking values in $\mathbb{R}^N$ with a law of data $P_N \in \mathcal{M}_1(\mathbb{R}^N)$, and loss function $L_N : \mathcal{X}_N \times \mathcal{Y}_N \rightarrow \mathbb{R}$ via online SGD with a constant learning rate, δ_N , which is given by

$$X_{k+1} = X_k - \delta_N \nabla L_N(X_k; Y_{k+1}) \quad (2.3.1)$$

with a random initialization $X_0 \sim \mu_N \in \mathcal{M}_1(\mathbb{R}^N)$. Our interest is to understand this evolution, (X_k) , in the regime where $N \rightarrow \infty$.

2.3.1 Summary statistics

Since tracking the full trajectory of SGD in high dimensions is impractical, we focus on the evolution of key summary statistics of (X_k) . Specifically, we consider a finite set of functions u_N that capture essential aspects of the dynamics. These functions u_N belong to $C^1(\mathbb{R}^N; \mathbb{R}^l)$ for some fixed l , and they summarize essential properties such as the loss, the alignment with the target parameter, or other relevant macroscopic quantities. We aim to

describe how $u_N(X_k)$ evolves over time and whether a lower-dimensional deterministic or stochastic process can approximate its trajectory.

Recent research has demonstrated that, under mild regularity conditions, these summary statistics evolve to the solution of a system of SDEs as the dimension increases [7]. The nature of these effective dynamics depends heavily on factors such as the initialization scheme, the specific parameter regions in which the scaling limit is studied, and how the learning rate scales with the dimension. To establish this scaling limit, we must carefully choose an appropriate scaling for the fluctuations of the loss. These fluctuations are characterized by the sample-wise error [6, 7], which is formally defined as follows

$$H_N(x, Y) = L_N(x, Y) - \Phi_N(x) \quad (2.3.2)$$

In the following, we suppress the dependence of H on Y and instead view H as a random function of x , denoted $H(x)$. We let $V(x) = \mathbb{E}[\nabla H \otimes \nabla H]$ be the covariance matrix (Fisher matrix) for ∇H at x .

2.3.2 Localizability and asymptotic closability

To ensure the tightness of the trajectories of the summary statistics [7], we impose two key assumptions on the triplet (u_N, L_N, P_N) and the learning rate δ_N .

1. **Bound on the Learning Rate:** The learning rate must be sufficiently small, constrained by the regularity properties of the summary statistics. This ensures that the updates remain stable and prevent divergence in high dimensions.
2. **Asymptotic Closeness of Summary Statistics:** The evolution of the summary statistics must remain close asymptotically as $N \rightarrow \infty$. This guarantees that, despite stochastic fluctuations, the system's trajectory can be well-approximated by a deterministic or stochastic limit.

These conditions allow for a controlled high-dimensional analysis of the learning dynamics, ultimately leading to a well-defined scaling limit.

Definition 2.22 (Localizability [7]). A triple (u_N, L_N, P_N) is δ_N -localizable with localizing sequence $(E_K)_K$ if there is an exhaustion by compacts $(E_K)_K$ of $\mathbb{R}^l$ and constant C_K (independent of N) such that

1. $\max_i \sup_{x \in u_N^{-1}(E_k)} \|\nabla^2 u_i^N\|_{\text{op}} \leq C_K \delta_N^{-1/2}$, and $\max_i \sup_{x \in u_N^{-1}(E_k)} \|\nabla^3 u_i^N\|_{\text{op}} \leq C_K$.
2. $\sup_{x \in u_N^{-1}(E_k)} \|\nabla \Phi_N\| \leq C_K$ and $\sup_{x \in u_N^{-1}(E_k)} \mathbb{E}[\|\nabla H_N\|^8] \leq C_K \delta_N^{-4}$.
3. $\max_i \sup_{x \in u_N^{-1}(E_k)} \mathbb{E}[\langle \nabla H, \nabla u_i^N \rangle^4] \leq C_K \delta_N^{-2}$ and $\max_i \sup_{x \in u_N^{-1}(E_k)} \mathbb{E}[\langle \nabla^2 u_i^N, \nabla H \otimes \nabla H - V \rangle^2] = o(\delta_N^{-3})$

To analyze the second assumption, we introduce key differential operators that describe the evolution of the chosen summary statistics. These operators capture both the deterministic and stochastic components of the learning process.

The first-order differential operator accounts for the drift term, representing the expected change in summary statistics at each step. The second-order differential operator incorporates the variance of these changes, reflecting how randomness in the updates propagates over time.

$$\mathcal{A}_N = \sum_{i=1}^N \partial_i \Phi_N \partial_i \quad \text{and} \quad \mathcal{L}_N = \frac{1}{2} \sum_{i,j=1}^N V_{ij} \partial_i \partial_j \quad (2.3.3)$$

Alternatively, we express the differential operators as: $\mathcal{A}_N = \langle \nabla \Phi_N, \nabla \rangle$, which represents the deterministic part of the dynamics, describing how the summary statistics evolve in expectation. Moreover, $\mathcal{L}_N = \frac{1}{2} \langle V, \nabla^2 \rangle$, which captures the stochastic fluctuations in the process, reflecting the variance in the updates.

These operators help in formulating the high-dimensional scaling limit by ensuring that the trajectory of the summary statistics remains controlled and predictable as $N \rightarrow \infty$.

Definition 2.23 (Closability [7]). A family of summary statistics (u_N) are asymptotically closable for learning rate δ_N if (u_N, L_N, H_N) are δ_N -localizable with localizing sequence $(E_K)_K$, and furthermore there exists locally Lipschitz functions $\mathcal{H} : \mathbb{R}^l \rightarrow \mathbb{R}^l$ and $\Sigma : \mathbb{R}^l \rightarrow \mathbb{R}^{l \times l}$, such that

$$\sup_{x \in u_N^{-1}(E_k)} \|(-\mathcal{A}_N + \delta_N \mathcal{L}_N)u_N(x) - \mathcal{H}(u_N(x))\| \rightarrow 0 \quad (2.3.4)$$

$$\sup_{x \in u_N^{-1}(E_k)} \|\delta_N J_N V J_N^T - \Sigma(u_N(x))\| \rightarrow 0 \quad (2.3.5)$$

In this case we call $\mathcal{H}$ the effective drift, and Σ the effective volatility.

Most of the assumed bounds in the definition of localizability serve to establish tightness and ensure that higher-order terms in Taylor expansions disappear in the $N \rightarrow \infty$ limit. In principle, the second assumption is essential; without it, the same quantity would still be controlled at $O(\delta_N^{-3})$ due to other localizability conditions. However, in that case, the limiting diffusion matrix would correspond to the limit of $\delta_N J_N V J_N^T$, rather than a more refined expression capturing finer aspects of the dynamics.

2.3.3 Diffusive scaling limit

In mathematical analysis and probability theory, the push-forward measure $f_*\mu$ transforms a probability distribution using a function. If you have a random variable X that follows a distribution μ , then applying a function f to X produces a new random variable $Y = f(X)$. The distribution of Y is called the push-forward measure of μ under f , denoted as $f_*\mu$. It describes how probabilities are redistributed when the transformation f is applied.

Theorem 2.24 (Effective Dynamic [7]). *Let $(X_k^{\delta_N})_k$ be SGD initialized from $X_0 \sim \mu_N$ for $\mu_N \in \mathcal{M}_1(\mathbb{R}^N)$ with learning rate δ_N for the loss L and data distribution P_N . For a family of summary statistics $u_N = (u_i^N)_{i=1}^l$, let $(u_N(t))_t$ be the linear interpolation of $(u_N(X_{\lfloor t\delta_N^{-1} \rfloor}^{\delta_N}))_t$. Suppose that u_N are asymptotically closable with learning rate δ_N , effective drift $\mathcal{H}$, and effective volatility Σ , and that the push-forward of the initial data has $(u_N)_* \mu_N \rightarrow \nu$ weakly for some $\nu \in \mathcal{M}_1(\mathbb{R}^l)$. Then $(u_N(t))_t \rightarrow (u_t)_t$ weakly as $N \rightarrow \infty$ where u_t solves*

$$du_t = \mathcal{H}(u_t)dt + \sqrt{\Sigma(u_t)}dB_t \quad (2.3.6)$$

initialized from ν , where B_t is a standard Brownian motion in $\mathbb{R}^l$.

Proof. See Appendix C. □

2.3.4 Sub-critical and critical scaling regimes

Remark 2.25. For the sake of simplicity, suppose that not only the asymptotic closability satisfies, but each of the two terms $\mathcal{A}_N u_N$ and $\delta_N \mathcal{L}_N u_N$ individually admit $N \rightarrow \infty$ limits: namely there exists $\mathcal{F}, \mathcal{G} : \mathbb{R}^l \rightarrow \mathbb{R}^l$ such that

$$\sup_{x \in u_N^{-1}(E_K)} \|\mathcal{A}_N u_N(x) - \mathcal{F}(u_N(x))\| \rightarrow 0, \quad (2.3.7)$$

$$\sup_{x \in u_N^{-1}(E_K)} \|\delta_N \mathcal{L}_N u_N(x) - \mathcal{G}(u_N(x))\| \rightarrow 0, \quad (2.3.8)$$

in which case, evidently $\mathcal{H} = -\mathcal{F} + \mathcal{G}$, and we call $\mathcal{F}, \mathcal{G}$ and Σ the drift, the corrector, and the volatility.

Then the above SGD turns out to be as follows,

$$du_t = (-\mathcal{F}(u_t) + \mathcal{G}(u_t))dt + \sqrt{\Sigma(u_t)}dB_t \quad (2.3.9)$$

Equation $du_t = -\mathcal{F}(u_t)dt$ represents the evolution of summary statistics u_t under gradient descent on the population loss function Φ , describing a deterministic ODE. This ODE captures the dynamics of gradient flow, where $\mathcal{F}(u_t)$ corresponds to the gradient of the population loss. In the high-dimensional limit, additional factors like the population corrector $\mathcal{G}$ and the diffusion matrix Σ come into play, accounting for the stochastic nature of the system as the dimension grows. In the *sub-critical scaling regime*, where $\mathcal{G}$ and Σ are negligible, the system behaves according to the classical gradient flow model. However, in the *critical scaling regime*, where these terms are nonzero, the dynamics are influenced by additional variance and learning terms, leading to more complex behavior. In single-index models, the terms $\mathcal{F}$ and $\mathcal{G}$ can be understood as the learning and variance terms, which reflect both the deterministic learning process and the stochastic fluctuations inherent in high-dimensional SGD.

2.3.5 Ballistic and diffusive phases

In many problems, even in the critical scaling regime where the population corrector $\mathcal{H} \neq \mathcal{F}$, the diffusion matrix for the effective dynamics of the summary statistics can be zero. This situation is referred to as the ballistic limit. In this case, the dynamics of the summary statistics are purely deterministic and described by the ordinary differential equation (ODE):

$$du_t = \mathcal{H}(u_t) dt. \quad (2.3.10)$$

Here, $\mathcal{H}(u_t)$ represents the deterministic flow driving the evolution of the summary statistics. In this ballistic limit, the evolution is not influenced by noise, and the trajectory of the system follows the phase portrait of this ODE. Essentially, the system's behavior is governed entirely by the dynamics encoded in $\mathcal{H}(u_t)$, and the fluctuations typically seen in high-dimensional dynamics are absent. Thus, in these scenarios, the phase portrait of the summary statistics closely resembles that of the deterministic flow given by equation 2.3.10.

By the construction of the scaling limit, the phase portrait of the ballistic limit only describes the evolution of summary statistics at length-scales that are $o(1)$ and at iterations that are $o(1/\delta_N)$. This means that the ballistic limit captures the macroscopic, large-scale behavior of the system but does not provide detailed information about the finer, microscopic structures of the system that may emerge near the fixed points of the dynamics.

2.3.6 Rescaled effective dynamics

If one is interested in understanding the evolution of the summary statistics u_t in these microscopic, $o(1)$ neighborhoods of the fixed points of the ballistic effective dynamics, Theorem 2.24 provides a way to extend the analysis. Specifically, Theorem 2.24 allows for developing a separate diffusive limit in these regions, capturing the finer details of the evolution in these small-scale neighborhoods. In other words, while the ballistic limit describes the coarse-grained, large-scale behavior of the system, the diffusive limit can be used to analyze the system at smaller scales and near fixed points, where the dynamics might be more complex and subject to finer fluctuations.

To study the diffusive regimes near the fixed points, the recentered and rescaled summary statistics, $\tilde{u}_N(t)$, are defined as:

$$\tilde{u}_N(t) = \delta_N^{-\alpha} (u_N(t) - u_*) \quad (2.3.11)$$

Where u_* is a fixed point of the ballistic dynamics. The goal is to apply Theorem 2.24 to these rescaled statistics, ensuring that the triple $(\tilde{u}_N(t), L_N, P_N)$ is δ_N -localizable. The parameter α must be chosen appropriately, so that the drift terms for $\tilde{u}_N$ are captured correctly, especially in the microscopic regime near u_* , where the first-order drift vanishes. Additionally, the initial distribution for $\tilde{u}_N$, denoted by $(\tilde{u}_N)_* \mu_N$, must converge to some limiting measure $\tilde{\nu}$.

With these considerations, the rescaled effective dynamics near the fixed point u_* are given by the following stochastic differential equation:

$$d\tilde{u}_t = \tilde{\mathcal{H}}(\tilde{u}_t) + \sqrt{\tilde{\Sigma}(\tilde{u}_t)}dB_t \quad (2.3.12)$$

where $\tilde{u}_0 \sim \tilde{\nu}$, and B_t is a Brownian motion term. This equation describes the diffusive behavior of the summary statistics in the microscopic neighborhoods of the fixed point, incorporating both the drift term $\tilde{\mathcal{H}}$ and the noise term given by the diffusion matrix $\tilde{\Sigma}$.

Chapter 3

Main Results

3.1 Effective dynamics of single-index models

Consider a single-index model with training set $\{(X_k), (Y_k)\}_{k=1}^M$ taking values in $\mathbb{R}^N$ with a quadratic loss $L : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ of the form

$$L(x, a) = (f(\langle a, x \rangle) - f(\langle a, v \rangle) + \epsilon)^2 \quad (3.1.1)$$

which is given by online SGD initialized from $X_0 \sim \mu_N$ with learning rate $\delta_N = 1/N$

$$X_{k+1} = X_k - \delta_N \nabla L(X_k; Y_{k+1}) \quad (3.1.2)$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a given activation function, feature vectors is $(a_k)_k \sim \mathcal{N}(0, I_N)$ and ϵ is an additive error.

The useful summary statistics that we most frequently analyze are linear functions of the parameter space $\mathbb{R}^N$, for instance the correlation with a unite vector, i.e., $\langle X, X^* \rangle$ or some ground truth; radial statistics, such as the L^2 -norm of the parameters, or some subset of the parameters, i.e., $\|X\|^2$; and rescaled versions, usually blown up by $\sqrt{N}$ of these near their fixed points as described in Chapter 2.

In our case, the sufficient number of summary statistics we need for the single-index model is $l = 2$, and for the sake of simplicity, let $X^* \in \mathbb{S}^{N-1}$ be a fixed unit vector. Now, let us define the following modified summary statistics $u_N(x) = (u_1^N(x), u_2^N(x))$

$$u_1(x) := m(x) = \langle x, X^* \rangle, \quad u_2(x) := r_\perp^2(x) = \|x\|^2 - m^2 \quad (3.1.3)$$

where the loss distribution only depends on the pair $u_N = (m, r_\perp^2)$. Hence, the population loss is of the form

$$\Phi(x) = \mathbb{E}_a[L(x, a)] = \phi(m, r_\perp^2) \quad (3.1.4)$$

As discussed above, they are solvable and only require two summary statistics: a correlation observable and a radial term. Even with this relative simplicity, we encounter various ODE and SDE limits. Indeed, we find dynamical phase transitions corresponding

to the aforementioned threshold in these models. Our analysis focuses exclusively on the most interesting, critical step-size scaling corresponding to the proportional asymptotics regime from the random matrix theory literature.

We are now ready to present our main results.

Theorem 3.1. *Let $(X_k^{\delta_N})_k$ be SGD initialized from $X_0 \sim \mu_N$ for $\mu_N \in \mathcal{M}_1(\mathbb{R}^N)$ with learning rate δ_N for the quadratic loss L of a single-index model and the information exponent of the activation function is at least two. For the corresponding summary statistics (correlation and radius) $u_N = (u_i^N)_{i=1}^2 = (m, r_\perp^2)$, let $(u_N(t))_t$ be the linear interpolation of $(u_N(X_{\lfloor t\delta_N^{-1} \rfloor}^{\delta_N}))_t$. Then u_N are asymptotically closable with learning rate $\delta_N = 1/N$. Moreover, $u_N = (m, r_\perp^2)$ converges as $N \rightarrow \infty$ to the solution of the following ODE initialized from $\lim_{N \rightarrow \infty} (u_N)_* \mu_N$*

$$\frac{dm}{dt} = -2\mathbb{E}[a_1 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))] \quad (3.1.5)$$

$$\begin{aligned} \frac{dr_\perp^2}{dt} = & -4\mathbb{E}[a_2 r_\perp f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))] \\ & + 4\mathbb{E}[f'^2(a_1 m + a_2 r_\perp)((f(a_1 m + a_2 r_\perp) - f(a_1))^2 + C_\epsilon)] \end{aligned} \quad (3.1.6)$$

We can obtain the following result when we restrict the initialization of the algorithm to be chosen randomly from a gaussian distribution, i.e., $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{N} I_N)$, then $(u_N)_* \mu_N \rightarrow \delta_{(0, \sigma^2)}$ weakly for some fixed σ^2 .

Corollary 3.2. *Suppose that the conditions of Theorem 3.1 hold. Then $m(t) = 0$ remains constant and $r_\perp^2(t)$ converges as $N \rightarrow \infty$ to the solution of the following ODE initialized from $\lim_{N \rightarrow \infty} (u_N)_* \mu_N = \delta_{(0, 1)}$,*

$$\frac{dr_\perp^2}{dt} = 4\mathbb{E}[f'^2(a_2 r_\perp)](C_\epsilon + \|f\|^2 - r_\perp^2) + 4\mathbb{E}[f'^2(a_2 r_\perp)f^2(a_2 r_\perp)] - 4r_\perp^2 \mathbb{E}[f''(a_2 r_\perp)f(a_2 r_\perp)] \quad (3.1.7)$$

A fixed point can be challenging to find in such an effective dynamic. For example, if $f = h_3$ (Hermite polynomial with degree 3), the dynamics can diverge, and the ballistic ODE becomes obsolete. We address this issue by restricting f to be bounded. In this case, it is guaranteed that the corresponding ODE exists, and the fixed point is denoted by $r_\perp^*$.

Let us consider a rescaling regime of u_N in a microscopic neighborhood of the saddle set $m = 0$. This captures the initial phase from a random start if $\mu_N \sim \mathcal{N}(0, \frac{\sigma^2}{N} I_N)$ for some fixed $\sigma^2 > 0$, then $(u_N)_* \mu_N \rightarrow \delta_{(0, \sigma^2)}$ weakly. Now rescale and let $\tilde{u}_N = (\tilde{m}, r_\perp^2)$ where $\tilde{m} = \sqrt{N}m$. Evidently, $\bar{\nu} = \lim_{N \rightarrow \infty} (\tilde{u}_N)_* \mu_N = \mathcal{N}(0, \sigma^2) \otimes \delta_{\sigma^2}$.

Theorem 3.3. *Let $(X_k^{\delta_N})_k$ be SGD initialized from $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{N} I_N)$ for some fixed σ^2 with learning rate δ_N for the quadratic loss L of a single-index model, the information exponent of the bounded activation function is at least two, and data distribution P_N . For the corresponding summary statistics (rescaled correlation and radius) $\tilde{u}_N = (\tilde{u}_i^N)_{i=1}^2 = (\tilde{m}, r_\perp^2)$, let $(\tilde{u}_N(t))_t$ be the linear interpolation of $(\tilde{u}_N(X_{\lfloor t\delta_N^{-1} \rfloor}^{\delta_N}))_t$. Then $\tilde{u}_N$ are asymptotically closable*

with learning rate $\delta_N = 1/N$. Moreover, $\tilde{u}_N = (\tilde{m}, r_\perp^2)$ converges as $N \rightarrow \infty$ to the solution of the following SDE initialized from $\mathcal{N}(0, \sigma^2) \otimes \delta_{\sigma^2}$

$$d\tilde{m} = -2\tilde{m}\mathbb{E}[(f'^2(a_2r_\perp) + f(a_2r_\perp)f''(a_2r_\perp))]dt \quad (3.1.8)$$

$$+ 2\sqrt{\mathbb{E}[f'^2(a_2r_\perp)f^2(a_2r_\perp)] + \mathbb{E}[f^2(a_2r_\perp)](\|f'\|^2 + 2\|f'\|^2 + 2\langle f, f'' \rangle)}dB_t$$

$$\frac{dr_\perp^2}{dt} = 4\mathbb{E}[f'^2(a_2r_\perp)](C_\epsilon + \|f\|^2 - r_\perp^2) + 4\mathbb{E}[f'^2(a_2r_\perp)f^2(a_2r_\perp)] - 4r_\perp^2\mathbb{E}[f''(a_2r_\perp)f(a_2r_\perp)] \quad (3.1.9)$$

As discussed earlier, if the radius term of the rescaled summary statistics starts from the initial point $r_\perp^*$, then a mean-reverting OU process appears.

Corollary 3.4. *If $\mu_N \sim \mathcal{N}(0, \frac{r_\perp^{*2}}{N}I_N)$ where $r_\perp^{*2}$ is the fixed-point of the ODE for $r_\perp^2(t)$ in Theorem 3.3 then $\tilde{m}(t)$ converges as $N \rightarrow \infty$ to the solution of the following mean-reverting OU process,*

$$d\tilde{m} = -2\tilde{m}\mathbb{E}_{a_2}[(f'^2(a_2r_\perp^*) + f(a_2r_\perp^*)f''(a_2r_\perp^*))]dt \quad (3.1.10)$$

$$+ 2\sqrt{2\mathbb{E}_{a_2}[f'^2(a_2r_\perp^*)](\|f'\|^2 + \langle f, f'' \rangle) + r_\perp^{*2}\mathbb{E}_{a_2}[f'^2(a_2r_\perp^*) + f(a_2r_\perp^*)f''(a_2r_\perp^*)]}dB_t$$

3.2 Proofs

3.2.1 Proof of Theorem 3.1 and Corollary 3.2

Lemma 3.5. *The distribution of the loss $L_N(x, (a, y))$ depends only on $u_N = (m, r_\perp^2)$. Furthermore, the population loss has information exponent $k \geq 2$. Then u_N is δ_N -localizable for E_K being the centered balls of radius K in $\mathbb{R}^2$.*

Proof. For the sake of simplicity, let us assume $X^* = v$ where v is the first basis vector of $\mathbb{R}^N$. First note that for every x , since f is differentiable and f' is of at most polynomial growth,

$$\nabla\Phi = \partial_m\phi\nabla m + \partial_{r_\perp^2}\phi\nabla r_\perp^2 \quad (3.2.1)$$

where

$$\begin{cases} \partial_m\phi = 2\mathbb{E}_{a_1, a_2}[a_1f'(a_1m + a_2r_\perp)(f(a_1m + a_2r_\perp) - f(a_1))] \\ \partial_{r_\perp^2}\phi = \frac{1}{r_\perp}\mathbb{E}_{a_1, a_2}[a_2f'(a_1m + a_2r_\perp)(f(a_1m + a_2r_\perp) - f(a_1))] \end{cases} \quad (3.2.2)$$

One may express the derivatives for u_N as

$$\nabla m = v, \quad \nabla r_\perp^2 = 2(x - mv) \quad (3.2.3)$$

Notice that $\nabla^2 m = 0$, while $\nabla^2 r_\perp^2 = 2(I - vv^T)$, and $\nabla^l m = \nabla^l r_\perp^2 = 0$ for all $l \geq 3$. It yields that

$$\langle \nabla m, \nabla m \rangle = 1 \quad \langle \nabla m, \nabla r_\perp^2 \rangle = 0 \quad \langle \nabla r_\perp^2, \nabla r_\perp^2 \rangle = 4r_\perp^2 \quad (3.2.4)$$

Consider

$$\|\nabla\Phi\| \leq |\partial_m\phi| \|\nabla m\| + |\partial_{r_\perp^2}\phi| \|\nabla r_\perp^2\| \quad (3.2.5)$$

the bounding quantity is evidently a continuous function of $m, r_\perp^2$ and therefore as long as x is such that $(m, r_\perp^2) \in E_K$, it is bounded by some constant $C(K)$.

Now, we calculate the covariance matrix $V = \mathbb{E}[\nabla H \otimes \nabla H]$ as follows

$$H(x, a) = (f(\langle a, x \rangle) - f(a_1))^2 - \phi(x) \quad (3.2.6)$$

The derivatives of H are given by

$$\nabla H = 2af'(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1)) - \nabla\phi \quad (3.2.7)$$

Since $\|a\| = O(\sqrt{N})$ and $\|\nabla\phi\| = O(1)$, one may write

$$\|\nabla H\| = 2|f'(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))| \cdot \|a\| + O(N^{-\lambda}) \quad (3.2.8)$$

where $\lambda > 1/2$. Therefore, there exists $C_K(f) > 0$ independent of N such that

$$\mathbb{E}[\|\nabla H\|^8] \leq C_K(f)N^4 \quad (3.2.9)$$

Moreover, there exists $c_K(f) > 0$ independent of N such that

$$\mathbb{E}[\langle \nabla H, \nabla r_\perp^2 \rangle^4] \leq \mathbb{E}[\|\nabla H\|^4] \leq c_K(f)N^2 \quad (3.2.10)$$

as claimed. $\square$

Proof of Theorem 3.1. Having checked localizability for u_N , we apply Theorem 2.24. To compute $\mathcal{F}$, by the above,

$$\begin{cases} \mathcal{F}_m = 2\mathbb{E}[a_1 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))] \\ \mathcal{F}_{r_\perp^2} = 4r_\perp \mathbb{E}_{a_1, a_2}[a_2 f'(a_1 m + a_2 r_\perp)(f(a_1 m + a_2 r_\perp) - f(a_1))] \end{cases} \quad (3.2.11)$$

We next turn to calculating the corrector. Recall $V = \mathbb{E}[\nabla H \otimes \nabla H]$, we have that

$$V_{ij} = \mathbb{E}[\partial_i H \partial_j H] = \mathbb{E}[\partial_i L \partial_j L] - \partial_i \phi \partial_j \phi \quad (3.2.12)$$

where

$$\begin{aligned} \mathbb{E}[\partial_i L \partial_j L] &= 4\mathbb{E}[a_i a_j f'^2(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))^2] + 4\mathbb{E}[\epsilon^2 a_i a_j f'^2(\langle a, x \rangle)] \\ &= 4\mathbb{E}[a_i a_j] \mathbb{E}[f'^2(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))^2] + 4C_\epsilon \mathbb{E}[a_i a_j] \mathbb{E}[f'^2(\langle a, x \rangle)] \\ &\quad + 4\text{Cov}(a_i a_j, f'^2(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))^2) \end{aligned} \quad (3.2.13)$$

In particular, for $\delta = 1/N$, we have $\delta \mathcal{L}^\delta m = 0$ and

$$\delta \mathcal{L}^\delta r_\perp^2 = \frac{1}{N} \sum_{i=2}^N V_{ii} = \frac{1}{N} \sum_{i=2}^N \mathbb{E}[(\partial_i H)^2] = \frac{1}{N} \sum_{i=2}^N \mathbb{E}[(\partial_i L)^2] - \frac{1}{N} \sum_{i=2}^N (\partial_i \phi)^2 \quad (3.2.14)$$

For the first term, one may write

$$\begin{aligned} \sum_{i=2}^N \mathbb{E}[(\partial_i L)^2] &= 4(N-1) \left(\mathbb{E}[f'^2(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))^2] + C_\epsilon \mathbb{E}[f'^2(\langle a, x \rangle)] \right) \\ &\quad + \sum_{i=2}^N \text{Cov}(a_i^2, f'^2(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))^2) \end{aligned} \quad (3.2.15)$$

Using the Cauchy-Schwarz inequality,

$$\sum_{i=2}^N \text{Cov}(a_i^2, f'^2(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))^2) \leq \sum_{i=2}^N \sqrt{2(N-1) \text{Var}(f'^2(\langle a, x \rangle)(f(\langle a, x \rangle) - f(a_1))^2)} \quad (3.2.16)$$

This results in the covariance term being $O(\sqrt{N})$, which vanishes in terms of the correction. Moreover, the population term, $\sum_{i=2}^N (\partial_i \phi)^2 = 4r_\perp^2 (\partial_{r_\perp} \phi)^2 = O(1)$, can be seen to vanish in the corrector as $N \rightarrow \infty$.

Finally, the corrector is given by

$$\mathcal{G}_{r_\perp^2} = 4\mathbb{E} [f'^2(a_1 m + a_2 r_\perp) ((f(a_1 m + a_2 r_\perp) - f(a_1))^2 + C_\epsilon)] \quad (3.2.17)$$

Together, these yield the ODE system for $(m, r_\perp^2)$,

$$\frac{dm}{dt} = -2\mathbb{E}_{a_1, a_2} [a_1 f'(a_1 m + a_2 r_\perp) (f(a_1 m + a_2 r_\perp) - f(a_1))] \quad (3.2.18)$$

$$\begin{aligned} \frac{dr_\perp^2}{dt} &= -4\mathbb{E}_{a_1, a_2} [a_2 r_\perp f'(a_1 m + a_2 r_\perp) (f(a_1 m + a_2 r_\perp) - f(a_1))] \\ &\quad + 4\mathbb{E}_{a_1, a_2} [f'^2(a_1 m + a_2 r_\perp) ((f(a_1 m + a_2 r_\perp) - f(a_1))^2 + C_\epsilon)] \end{aligned} \quad (3.2.19)$$

Now, let us find the saddle set of m in the above system. If we fix $m = 0$, then the ODE for m is given by

$$\frac{dm}{dt} = -2\mathbb{E}[a_1 f'(a_2 r_\perp) f(a_2 r_\perp)] + 2\mathbb{E}[a_1 f(a_1) f'(a_2 r_\perp)] \quad (3.2.20)$$

Then, using Stein's lemma, one may write

$$\frac{dm}{dt} = 2\mathbb{E}[f'(a_1)]\mathbb{E}[f'(a_2 r_\perp)] \quad (3.2.21)$$

Recalling the Hermite polynomials, $(h_k(x))_{k=0}^\infty$, the activation function can be expressed in terms of Hermite polynomials as follows

$$f(x) = \sum_k u_k h_k(x) \quad \text{where,} \quad u_k = \langle f, h_k \rangle_{L^2(\gamma)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(z) h_k(z) e^{-z^2/2} dz \quad (3.2.22)$$

Hence $\mathbb{E}[f'(a_1)] = u_1$ and if the population loss has the information exponent larger than 2, then $dm/dt = 0$. $\square$

Proof of Corollary 3.2. Since $m = 0$ is the initial point of the dynamic, one may write

$$\begin{aligned} \frac{dr_{\perp}^2}{dt} = & -4\mathbb{E}_{a_1, a_2}[a_2 r_{\perp} f'(a_2 r_{\perp})(f(a_2 r_{\perp}) - f(a_1))] \\ & + 4\mathbb{E}_{a_1, a_2}[f'^2(a_2 r_{\perp})((f(a_2 r_{\perp}) - f(a_1))^2 + C_{\epsilon})] \end{aligned} \quad (3.2.23)$$

Using Stein's lemma, the first term can be expressed as follows

$$\mathbb{E}[a_2 r_{\perp} f'(a_2 r_{\perp})f(a_2 r_{\perp})] = r_{\perp}^2 \mathbb{E}[f''(a_2 r_{\perp})f(a_2 r_{\perp}) + f'^2(a_2 r_{\perp})] \quad (3.2.24)$$

And for the second term

$$\mathbb{E}[f'^2(a_2 r_{\perp})f^2(a_2 r_{\perp})] + \mathbb{E}[f'^2(a_2 r_{\perp})](\|f\|^2 + C_{\epsilon}) \quad (3.2.25)$$

Hence,

$$\begin{aligned} \frac{dr_{\perp}^2}{dt} = & 4\mathbb{E}[f'^2(a_2 r_{\perp})](C_{\epsilon} + \|f\|^2 - r_{\perp}^2) + 4\mathbb{E}[f'^2(a_2 r_{\perp})f^2(a_2 r_{\perp})] - 4r_{\perp}^2 \mathbb{E}[f''(a_2 r_{\perp})f(a_2 r_{\perp})] \end{aligned} \quad (3.2.26)$$

□

3.2.2 Proof of Theorem 3.3 and Corollary 3.4

Lemma 3.6. *The distribution of the loss $L(x, (a, y))$ depends only on $u_N = (\sqrt{N}m, r_{\perp}^2)$. Furthermore, the population loss has information exponent $k \geq 2$. Then u_N is δ_N -localizable for E_K being the centered balls of radius K in $\mathbb{R}^2$.*

Proof. For the sake of simplicity, let us assume $X^* = v$ where v is the first basis vector of $\mathbb{R}^N$. We have checked localizability in lemma 3.5, but the change from the original variables is in the J matrix, in which now $\nabla \tilde{m} = \sqrt{N} \nabla m = \sqrt{N} v$. This does not affect the first two conditions of localizability; for the third condition, notice that for some $C > 0$

$$\mathbb{E}[\langle \nabla H, \nabla \tilde{m} \rangle^4] = N^2 \mathbb{E}[\langle \nabla H, v \rangle^4] = N^2 \mathbb{E}[(\partial_1 H)^2] \leq N^2 C \quad (3.2.27)$$

and the second part of the third condition is unchanged since $\nabla^2 \tilde{m} = 0$. □

Proof of Theorem 3.3. In a neighborhood of the saddle point $m = 0$, using Taylor expansion for some $\epsilon > 1$,

$$f(a_1 m + a_2 r_{\perp}) = f(a_2 r_{\perp}) + \frac{\tilde{m}}{\sqrt{N}} a_1 f'(a_2 r_{\perp}) + \frac{\tilde{m}^2}{2N} a_1^2 f''(a_2 r_{\perp}) + O(N^{-\epsilon}) \quad (3.2.28)$$

Then the population can be expressed as follows

$$\begin{aligned} \phi(\tilde{m}, r_{\perp}^2) = & \mathbb{E}_{a_1, a_2}[(f(a_2 r_{\perp}) - f(a_1))^2] + \frac{\tilde{m}^2}{N} \mathbb{E}_{a_2}[f''(a_2 r_{\perp})f(a_2 r_{\perp})] + \frac{\tilde{m}^2}{N} \mathbb{E}_{a_2}[f'^2(a_2 r_{\perp})] + O(N^{-\epsilon}) \end{aligned} \quad (3.2.29)$$

One may write the derivatives for $\tilde{\mathbf{u}}_N$ as

$$\nabla \tilde{m} = \sqrt{N}v, \quad \nabla r_{\perp}^2 = 2(x - mv) \quad (3.2.30)$$

Similar to the standard summary statistics, $\nabla^2 \tilde{m} = 0$, while $\nabla^2 r_{\perp}^2 = 2(I - vv^T)$, and $\nabla^l r_{\perp}^2 = 0$ for all $l \geq 3$. It yields that

$$\langle \nabla \tilde{m}, \nabla \tilde{m} \rangle = N \quad \langle \nabla \tilde{m}, \nabla r_{\perp}^2 \rangle = 0 \quad \langle \nabla r_{\perp}^2, \nabla r_{\perp}^2 \rangle = 4r_{\perp}^2 \quad (3.2.31)$$

Having checked localizability for $\mathbf{u}_N$, we apply Theorem 2.24. To compute $\mathcal{F}$, by the above

$$\begin{cases} \mathcal{F}_{\tilde{m}} = 2\tilde{m}\mathbb{E}[(f'^2(a_2r_{\perp}) + f(a_2r_{\perp})f''(a_2r_{\perp}))] \\ \mathcal{F}_{r_{\perp}^2} = 4r_{\perp}\mathbb{E}_{a_2}[a_2f(a_2r_{\perp})f'(a_2r_{\perp})] \end{cases} \quad (3.2.32)$$

In particular, for $\delta = 1/N$, we have $\delta \mathcal{L}^{\delta} m = 0$. Thus, the corrector $\mathcal{G}_m = 0$. Moreover, the corrector for $r_{\perp}^2$ is given by

$$\mathcal{G}_{r_{\perp}^2} = 4\mathbb{E}_{a_2}[f'^2(a_2r_{\perp})f^2(a_2r_{\perp})] + 4\mathbb{E}_{a_2}[f'^2(a_2r_{\perp})]\mathbb{E}_{a_1}[f^2(a_1)] \quad (3.2.33)$$

Finally, we consider the volatility of the stochastic process one gets in the limit. Rescaling JVJ^T by noticing that the rescaling $J \rightarrow \tilde{J}$ multiplies its $(1, 1)$ -entry by N and its off-diagonal entries by $\sqrt{N}$, one may obtain the entries as follows. Thus,

$$J = \begin{pmatrix} \nabla m \\ \nabla r^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 2x_2 & \cdots & 2x_N \end{pmatrix} \quad \text{and} \quad JVJ^T = \begin{pmatrix} V_{11} & 2\sum_{i=2}^N x_i V_{1i} \\ 2\sum_{i=2}^N x_i V_{1i} & 4\sum_{i=2}^N x_i x_j V_{ij} \end{pmatrix} \quad (3.2.34)$$

The $(1, 2)$ -entry of the volatility is given by

$$\begin{aligned} \frac{1}{2}(JVJ^T)_{12} &= \sum_{i=2}^N x_i V_{1i} = \sum_{i=2}^N x_i \mathbb{E}[\partial_1 H \partial_i H] \\ &= 4 \sum_{i=2}^N \mathbb{E}_a[a_1 a_i x_i f'^2(\langle a, x \rangle)((f(\langle a, x \rangle) - f(a_1))^2 + C_{\epsilon})] \\ &= 4\mathbb{E}_{a_1, a_2}[a_1 a_2 r_{\perp} f'^2(a_2 r_{\perp})((f(a_2 r_{\perp}) - f(a_1))^2 + C_{\epsilon})] \end{aligned} \quad (3.2.35)$$

and since the rescaled volatility is $\delta(\tilde{J}V\tilde{J}^T)_{1,2} = \frac{1}{N}\sqrt{N}(JVJ^T)_{1,2}$, after taking limits, this term will be vanished. We could also say the same reason for the $(2, 1)$ entry.

The $(2, 2)$ -entry, one may write

$$\begin{aligned} \frac{1}{4}(JVJ^T)_{22} &= \sum_{i,j=2}^N x_i x_j V_{ij} = \sum_{i,j=2}^N x_i x_j \mathbb{E}[\partial_i H \partial_j H] \\ &= 4 \sum_{i,j=2}^N \mathbb{E}_a[a_i x_i a_j x_j f'^2(\langle a, x \rangle)((f(\langle a, x \rangle) - f(a_1))^2 + C_{\epsilon})] \\ &= 4\mathbb{E}_{a_2}[a_2^2 r_{\perp}^2 f'^2(a_2 r_{\perp})((f(a_2 r_{\perp}) - f(a_1))^2 + C_{\epsilon})] \end{aligned} \quad (3.2.36)$$

Again, the rescaled volatility is $\delta(\tilde{J}V\tilde{J}^T)_{2,2} = \frac{1}{N}(JVJ^T)_{2,2}$ vanishes when $N \rightarrow \infty$. It means the The only surviving entry of the volatility is the (1, 1)-entry that is given by

$$(JVJ^T)_{11} = V_{11} = \mathbb{E}[(\partial_1 H)^2] = 4\mathbb{E}[a_1^2 f'^2(a_2 r_\perp)((f(a_2 r_\perp) - f(a_1))^2 + C_\epsilon)] \quad (3.2.37)$$

Hence, the volatility is of the form

$$\Sigma_{11} = 4\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)f^2(a_2 r_\perp)] + 4\mathbb{E}_{a_1}[a_1^2 f^2(a_1)]\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)], \quad \Sigma_{21} = \Sigma_{12} = \Sigma_{22} = 0 \quad (3.2.38)$$

Using the integral by parts and Stein's lemma, one may write

$$\mathbb{E}_{a_1}[a_1^2 f^2(a_1)] = \mathbb{E}_{a_1}[f^2(a_1) + 2f'^2(a_1) + 2f(a_1)f''(a_1)] \quad (3.2.39)$$

Therefore,

$$\Sigma_{11} = 4\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)f^2(a_2 r_\perp)] + 4(\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)] + C_\epsilon)(\|f\|^2 + 2\|f'\|^2 + 2\langle f, f'' \rangle) \quad (3.2.40)$$

Together, these yield the SDE system for $(\tilde{m}, r_\perp^2)$,

$$d\tilde{m} = -2\tilde{m}\mathbb{E}_{a_2}[(f'^2(a_2 r_\perp) + f(a_2 r_\perp)f''(a_2 r_\perp))]dt \quad (3.2.41)$$

$$\begin{aligned} & + 2\sqrt{\mathbb{E}_{a_2}[f'^2(a_2 r_\perp)f^2(a_2 r_\perp)] + (\mathbb{E}_{a_2}[f^2(a_2 r_\perp)] + C_\epsilon)(\|f\|^2 + 2\|f'\|^2 + 2\langle f, f'' \rangle)}dB_t \\ \frac{dr_\perp^2}{dt} & = 4\mathbb{E}[f'^2(a_2 r_\perp)](C_\epsilon + \|f\|^2 - r_\perp^2) + 4\mathbb{E}[f'^2(a_2 r_\perp)f^2(a_2 r_\perp)] - 4r_\perp^2\mathbb{E}[f''(a_2 r_\perp)f(a_2 r_\perp)] \end{aligned} \quad (3.2.42)$$

□

Proof of Corollary 3.4. Now, if we replace $r_\perp$ with the fixed point of the corresponding ODE, then

$$\Sigma_{11} = 4\mathbb{E}[f'^2(a_2 r_\perp)](r_\perp^2 + \|f'\|_{L(\gamma)}^2 + \langle f, f'' \rangle_{L(\gamma)}) + 4r_\perp^2\mathbb{E}[f(a_2 r_\perp)f''(a_2 r_\perp)] \quad (3.2.43)$$

At the end, let's determine the sign of the drift. One may write

$$\mathbb{E}_{a_2}[f'^2(a_2 r_\perp) + f(a_2 r_\perp)f''(a_2 r_\perp)] = \frac{1}{r_\perp^2}\mathbb{E}[f'^2(a_2 r_\perp)f^2(a_2 r_\perp)] + \frac{1}{r_\perp^2}\mathbb{E}[f'^2(a_2 r_\perp)]\mathbb{E}[f^2(a_1)] > 0 \quad (3.2.44)$$

as $\mathbb{E}[g^2(\alpha X)] > 0$ for any real function g , $X \sim \mathcal{N}(0, 1)$, and $\alpha > 0$. □

3.2.3 Proof of Proposition 2.10

Proof of Proposition 2.10. According to Theorem 2.8, one may write the dynamics with respect to the elements of the overlap matrix Ω . For the first block, R , we have

$$\frac{d(X^T X)_{ij}}{dt} = -4 \sum_{l=1}^N (X^T X)_{il}(H_1)_{lj} - 2 \sum_{l=1}^N (X^T X^*)_{il}(H_2)_{lj} + G_{ij} \quad (3.2.45)$$

Also, for the second block, m , we have

$$\frac{d(X^T X^*)_{ij}}{dt} = -2 \sum_{l=1}^N (X^T X^*)_{il} (H_1)_{lj} - 2 \sum_{l=1}^N (X^{*T} X^*)_{il} (H_2)_{lj} \quad (3.2.46)$$

Recall the definition of the summary statistics of $u_N = (R, m)$,

$$R_k = \frac{1}{N} \langle X_k, X_k \rangle = \frac{1}{N} \sum_{i=1}^N X_{k,i}^2, \quad m_k = \frac{1}{N} \langle X_k, X^* \rangle = \frac{1}{N} \sum_{i=1}^N X_{k,i} X_i^* \quad (3.2.47)$$

If we sum the above ODEs to get the dynamics for R and m , one may obtain

$$\frac{dR_t}{dt} = -\frac{4}{N} \sum_{i,l} X_i X_l \frac{\partial \phi}{\partial (X_i X_l)} - \frac{2}{N} \sum_{i,l} X_i X_l^* \frac{\partial \phi}{\partial (X_i^* X_l)} + \frac{1}{N} \text{tr}(G_t) \quad (3.2.48)$$

and

$$\frac{dm_t}{dt} = -\frac{2}{N} \sum_{i,l} X_i X_l^* \frac{\partial \phi}{\partial (X_i X_l^*)} - \frac{2}{N} \sum_{i,l} X_i^* X_l^* \frac{\partial \phi}{\partial (X_i^* X_l^*)} \quad (3.2.49)$$

Since the population loss $\Phi_N(X)$ is of the form $\Phi_N(X_t) = \phi(R_t, m_t)$, then there is a specific rule for the derivatives, namely,

$$\frac{\partial \phi}{\partial (X_i X_j)} = \delta_{ij} \frac{\partial \phi}{\partial X_i^2} = \frac{\delta_{ij}}{2X_i} \frac{\partial \phi}{\partial X_i}, \quad \text{and} \quad \frac{\partial \phi}{\partial (X_i X_j^*)} = \delta_{ij} \frac{\partial \phi}{\partial X_i X_i^*} = \frac{\delta_{ij}}{X_i^*} \frac{\partial \phi}{\partial X_i} \quad (3.2.50)$$

Hence, the effective dynamics for $u = (R, m)$ is given by

$$\frac{dR_t}{dt} = \lim_{N \rightarrow \infty} -\frac{4}{N} \sum_i X_i \frac{\partial \phi}{\partial X_i} + \frac{1}{N} \text{tr}(G_t), \quad \text{and} \quad \frac{dm_t}{dt} = \lim_{N \rightarrow \infty} -\frac{2}{N} \sum_i X_i^* \frac{\partial \phi}{\partial X_i} \quad (3.2.51)$$

Moreover, one may reparametrize the above dynamics by the following ones,

$$\frac{dR_t}{dt} = -\langle \nabla \Phi_N, \nabla R_t \rangle + \lim_{N \rightarrow \infty} \frac{1}{2N} \text{tr}(G_t) \quad \text{and} \quad \frac{dm_t}{dt} = -\langle \nabla \Phi_N, \nabla m_t \rangle \quad (3.2.52)$$

as claimed. $\square$

Chapter 4

Discussion

We start this discussion by briefly comparing our approach to three related approaches that have been mentioned to investigate similar questions in recent years. The classical approach to such problems is the *ODE method* which dates back at least to the work of Ljung (1977) [35]. Here, one proves convergence of the trajectory with a small step-size to the solution of the population dynamics. While this approach is similar to ours, it is designed for fixed dimensions. Indeed, in the scaling regime studied here, namely that corresponding to the initial search phase, one cannot neglect the effect of the error martingales as diffusive fluctuations.

Another approach to estimation problems via gradient-based algorithms is to study a diffusion approximation to this problem. Diffusion approximations to stochastic approximation algorithms date at least back to the work of McLeish (1976) [39] in finite dimensions; more recently, they have been studied in high dimensions for online SGD, e.g., distributional dynamics [40]. There is a lot of mathematical research on distributional dynamics, inspired by the study of interacting particle systems in mathematical physics. A key idea in this research is that the corresponding PDE can be seen as a gradient flow for the population loss $\Phi(\rho)$ in the space of probability measures on $\mathbb{R}^N$ with the Wasserstein metric. This means the trajectory $t \mapsto \rho_t$ tries to minimize the population loss while keeping *local mass conservation*.

On the other hand, several statistical physics motivated approaches have recently been introduced to study such questions in online settings, e.g., deterministic scaling limits. Their results develop a scaling limit, in the regime where one first sends $N \rightarrow \infty$ then $T \rightarrow \infty$. In this setting, however, the solution to the corresponding limiting evolution equation admits a trivial zero solution if $m_0 = 0$ and a meaningful solution if $m_0 > 0$. This limiting system is then analyzed for its behavior and recovery thresholds.

Additionally, the scaling limit helps us understand how the shape of the multi-index model's loss landscape depends on the dimension N . This links the SGD dynamics of these models to a diffusion equation. The distributional dynamics approach studies the empirical distribution, while the effective dynamics focus on summary statistics that describe the relationship between the teacher and student networks.

Regarding the regularization we use in our problem, in Definition 2.22, item (1) holds automatically for linear functions. In the case of radial statistics, the Hessian forms a block identity matrix, meaning item (1) is satisfied as long as δ_N remains $O(1)$. Therefore, item (1) imposes the most substantial restriction when rescaling nonlinear statistics. For item (2), the regularity assumptions on Φ_N and L_N here are weaker than the standard uniform Lipschitz condition commonly found in the literature, such as in Theorem 2.5 and Definition 2.7 for Theorem 2.8. Finally, the scaling conditions in item (3) ensure some level of independence between H and the values of ∇u and ∇u^2 at x .

Let us now discuss the intuition behind the information exponent and its relation to our main result. Together, Theorems 2.17 and 2.18 show that the information exponent governs, in a sharp sense, the performance of online SGD when started from a uniform random initialization. To understand where these thresholds come from, consider the following simplification of the recovery problem given by

$$m_t = m_{t-1} - \frac{c_\delta}{N} \partial_m \phi(m_{t-1}) \|\nabla m_{t-1}\|^2 + \frac{c_\delta}{N} c n_{t-1}^{k-1} \quad (4.0.1)$$

for m_{t-1} small and some $c > 0$. This amounts to gradient descent for the population loss.

Analyzing the finite difference equation, one finds that if the initial correlation is positive but microscopic, $m_0 \asymp N^{-\gamma}$ for any $\gamma > 0$, there are three regimes with distinct behaviors:

- $k < 2$: the time for m_t to weakly recover ($m_t \geq \eta$) is linear, i.e., order $c_\delta^{-1} N$.
- $k = 2$: the time for m_t to weakly recover is quasi-linear, i.e., $c_\delta^{-1} N \log N$.
- $k > 2$: the time for m_t to weakly recover is polynomial: $c_\delta^{-1} N^{1+c_\gamma}$ for $c_\gamma = (k-2)\gamma$.

In the online setting, the number of time steps equals the number of samples used. Consequently, no matter what $\gamma \in (0, 1)$ is, there is a transition between linear sample complexity and polynomial sample complexity regimes for the gradient descent on the population loss as one varies the information exponent, through the critical $k_c = 2$.

The primary contribution of this thesis, explored in detail in Chapter 3, is analyzing the effective dynamics of SGD in training single-index models. We demonstrate that with a random initialization, SGD deviates from the trajectory predicted by DMFT in high-dimensional settings. Our calculation shows that effective dynamics need additional correction terms, which naturally arise through diffusion. These terms incorporate diffusive components that account for small fluctuations in the learning curve. When the information exponent exceeds two, we identify and classify the fixed points of this effective dynamics. We focus on the critical step-size with $c_\delta = 1$, where it follows from 3.1 that $m = 0$. Moreover, we establish that the learning dynamics resemble an OU process near the fixed points of the associated SDEs.

In Cohen (2020) [18], it was empirically observed that the best training for neural networks does not occur when step-sizes are small enough for the classical gradient flow approximation to be valid. Instead, it appears at the edge of stability where the step size is small enough for the training to remain stable. Here, the loss fluctuates for some time

before eventually converging to lower values than with a smaller step size. This critical step size scaling is defined via the sharpness, namely the largest eigenvalue of the training loss Hessian. For recent theoretical investigations of this phenomenon, see, for example, [1, 3, 24, 57].

While sharpness and edge of stability do not have direct analogues in the context of online SGD, a qualitatively similar phenomenon can be seen by taking the population loss as a summary statistic. The critical scaling of the learning rate with dimension discussed constrains the step size in terms of the top eigenvalue of the Hessian of the loss. With this scaling, the population loss fluctuates near critical regions of its ballistic flow, allowing it to escape the critical region. In contrast, with a sub-critical learning rate, the population loss stays stuck. We leave a more detailed investigation of this connection to edge-of-stability phenomena for future investigation.

References

- [1] Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability", October 2023. arXiv:2212.07469 [cs].
- [2] Dyego Araújo, Roberto I. Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks, June 2019. arXiv:1906.00193 [math].
- [3] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding Gradient Descent on the Edge of Stability in Deep Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 948–1024. PMLR, June 2022. ISSN: 2640-3498.
- [4] G. Ben Arous, A. Dembo, and A. Guionnet. Aging of spherical spin glasses. *Probability Theory and Related Fields*, 120(1):1–67, May 2001.
- [5] Gerard Ben Arous, Amir Dembo, and Alice Guionnet. Cugliandolo-Kurchan equations for dynamics of Spin-Glasses, September 2004. arXiv:math/0409273.
- [6] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.
- [7] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Communications on Pure and Applied Mathematics*, 77(3):2030–2080, 2024. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22169>.
- [8] Michel Benaïm. Dynamics of stochastic approximation algorithms. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor, editors, *Séminaire de Probabilités XXXIII*, pages 1–68, Berlin, Heidelberg, 1999. Springer.
- [9] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, Berlin, Heidelberg, 1990.
- [10] M. Biehl and H. Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643, February 1995.

- [11] Léon Bottou. On-line Learning and Stochastic Approximations. In David Saad, editor, *On-Line Learning in Neural Networks*, Publications of the Newton Institute, pages 9–42. Cambridge University Press, Cambridge, 1999.
- [12] Léon Bottou. Stochastic Learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, pages 146–168. Springer, Berlin, Heidelberg, 2004.
- [13] Léon Bottou and Yann Cun. Large Scale Online Learning. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.
- [14] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data, December 2021. arXiv:2112.07572 [math].
- [15] Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting, July 2020. arXiv:1805.01648 [stat].
- [16] Julien Chevallier. Fluctuations for mean-field interacting age-dependent Hawkes processes. *Electronic Journal of Probability*, 22(none):1–49, January 2017. Publisher: Institute of Mathematical Statistics and Bernoulli Society.
- [17] Lenaïc Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport, October 2018. arXiv:1805.09545 [math].
- [18] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. October 2020.
- [19] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the High-Dimensional Notes: An ODE for SGD learning dynamics on GLMs and multi-index models, August 2023. arXiv:2308.08977 [math].
- [20] F. Comets and Th. Eisele. Asymptotic dynamics, non-critical and critical fluctuations for a geometric long-range interacting model. *Communications in Mathematical Physics*, 118(4):531–567, December 1988.
- [21] A. Crisanti, H. Horner, and H. J. Sommers. The spherical p-spin interaction spin-glass model. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, June 1993.
- [22] 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, July 1993. Publisher: American Physical Society.
- [23] 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, July 1993.

- [24] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability. September 2022.
- [25] Donald A. Dawson. Critical dynamics and fluctuations for a mean-field model of cooperative behavior. *Journal of Statistical Physics*, 31(1):29–85, April 1983.
- [26] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3), June 2020.
- [27] Paul Dupuis and Harold J. Kushner. Stochastic Approximation and Large Deviations: Upper Bounds and w.p.1 Convergence. *SIAM Journal on Control and Optimization*, 27(5):1108–1135, September 1989. Publisher: Society for Industrial and Applied Mathematics.
- [28] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, Cambridge, 2001.
- [29] E Gardner and B Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983–1994, June 1989.
- [30] Sebastian Goldt, Madhu S. Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124010, December 2020. arXiv:1906.08632 [stat].
- [31] Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1579–1613. PMLR, June 2019. ISSN: 2640-3498.
- [32] Thomas G. Kurtz and Jie Xiong. A stochastic evolution equation arising from the fluctuations of a class of interacting particle systems. *Communications in Mathematical Sciences*, 2(3):325–358. Publisher: International Press of Boston.
- [33] Harold J. Kushner. Asymptotic behavior of stochastic approximation and large deviations. In *The 22nd IEEE Conference on Decision and Control*, pages 75–81, December 1983.
- [34] Tengyuan Liang, Subhabrata Sen, and Pragya Sur. High-dimensional Asymptotics of Langevin Dynamics in Spiked Matrix Models. *Information and Inference: A Journal of the IMA*, 12(4):2720–2752, September 2023. arXiv:2204.04476 [math].
- [35] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575, August 1977. Conference Name: IEEE Transactions on Automatic Control.

- [36] Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, October 2019. Publisher: Proceedings of the National Academy of Sciences.
- [37] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference, January 2018. arXiv:1704.04289 [stat].
- [38] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, and Lenka Zdeborová. Who is Afraid of Big Bad Minima? Analysis of Gradient-Flow in a Spiked Matrix-Tensor Model, January 2020. arXiv:1907.08226 [cs].
- [39] D. L. McLeish. Functional and random central limit theorems for the Robbins-Munro process. *Journal of Applied Probability*, 13(1):148–154, March 1976.
- [40] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, August 2018. Publisher: Proceedings of the National Academy of Sciences.
- [41] Deanna Needell and Rachel Ward. Batched Stochastic Gradient Descent with Weighted Sampling. In Gregory E. Fasshauer and Larry L. Schumaker, editors, *Approximation Theory XV: San Antonio 2016*, pages 279–306, Cham, 2017. Springer International Publishing.
- [42] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- [43] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis, June 2017. arXiv:1702.03849 [cs].
- [44] H. Rieger, M. Schreckenberg, and J. Zittartz. Glauber dynamics of the Little-Hopfield model. *Zeitschrift für Physik B Condensed Matter*, 72(4):523–533, December 1988.
- [45] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. Publisher: Institute of Mathematical Statistics.
- [46] Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and Accuracy of Neural Networks: An Interacting Particle System Approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, September 2022. arXiv:1805.00915 [stat].
- [47] David Saad and Sara A. Solla. Exact Solution for On-Line Learning in Multilayer Neural Networks. *Physical Review Letters*, 74(21):4337–4340, May 1995. Publisher: American Physical Society.

- [48] Justin Sirignano and Konstantinos Spiliopoulos. Mean Field Analysis of Neural Networks: A Central Limit Theorem, June 2019. arXiv:1808.09372 [math].
- [49] Justin Sirignano and Konstantinos Spiliopoulos. Mean Field Analysis of Neural Networks: A Law of Large Numbers, November 2019. arXiv:1805.01053 [math].
- [50] H. Sompolinsky and Annette Zippelius. Dynamic Theory of the Spin-Glass Phase. *Physical Review Letters*, 47(5):359–362, August 1981. Publisher: American Physical Society.
- [51] H. Sompolinsky and Annette Zippelius. Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Physical Review B*, 25(11):6860–6875, June 1982. Publisher: American Physical Society.
- [52] Konstantinos Spiliopoulos, Justin A. Sirignano, and Kay Giesecke. Fluctuation Analysis for the Loss From Default, February 2015. arXiv:1304.1420 [math].
- [53] Alain-Sol Sznitman. Topics in propagation of chaos. In Donald L. Burkholder, Etienne Pardoux, Alain-Sol Sznitman, and Paul-Louis Hennequin, editors, *Ecole d’Eté de Probabilités de Saint-Flour XIX — 1989*, pages 165–251, Berlin, Heidelberg, 1991. Springer.
- [54] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of Stochastic Gradient Descent in high-dimensional two-layer neural networks, June 2023. arXiv:2202.00293 [stat].
- [55] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.
- [56] Yuchen Zhang, Percy Liang, and Moses Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1980–2022. PMLR, June 2017. ISSN: 2640-3498.
- [57] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example. September 2022.

APPENDICES

Appendix A

Sub-gaussian and sub-exponential

Theorem A.1 (Sub-gaussian properties [55]). *Let X be a random variable. Then the following properties are equivalent; the parameter $K_i > 0$ apperaing in the properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*

$$P(|X| \geq t) \leq 2 \exp(-t^2/K_1^2), \quad \forall t \geq 0 \quad (\text{A.0.1})$$

2. *The moments of X satisfy*

$$\|X\|_p = (\mathbb{E}[|X|^p])^{1/p} \leq K_2 \sqrt{p}, \quad \forall p \geq 1 \quad (\text{A.0.2})$$

3. *The moment generating function of X^2 satisfies*

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(K_3^2 \lambda^2), \quad \forall |\lambda| \leq 1/K_3 \quad (\text{A.0.3})$$

4. *The moment generating function of X^2 is bounded at some point, namely*

$$\mathbb{E}[\exp(X^2/K_4^2)] \leq 2, \quad \forall \lambda \in \mathbb{R} \quad (\text{A.0.4})$$

Definition A.2 (Sub-gaussian random variables). A random variable X that satisfies one of the equivalent sub-gaussian properties is called a sub-gaussian random variable. The sub-gaussian norm of X is denoted $\|X\|_{\psi_2}$ is defined to be the smallest K_4 in Theorem A.1, i.e.,

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\} \quad (\text{A.0.5})$$

Remark A.3. Gaussian, Bernoulli, and bounded random variables are some classical exmaples of sub-gaussian random variables.

Corollary A.4 (Hoeffding inequality). *Let $X_1, \dots, X_N$ be independent centered Gaussian random variables and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then for every $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^N X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2 \sum_{i=1}^N \sigma_i^2}\right), \quad P\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{2 \max_i \sigma_i^2 \|a\|_2^2}\right) \quad (\text{A.0.6})$$

The sub-gaussian distribution class, while extensive and intuitive, excludes significant distributions with heavier tails than the Gaussian. In this section, we shift our focus to distributions featuring at least an exponential tail decay. We aim to establish an analogue of Hoeffding's inequality tailored to this class of distributions.

Theorem A.5 (Sub-exponential properties [55]). *Let X be a random variable. Then the following properties are equivalent; the parameter $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*

$$P(|X| \geq t) \leq 2 \exp(-t/K_1), \quad \forall t \geq 0 \quad (\text{A.0.7})$$

2. *The moments of X satisfy*

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq K_2 p, \quad \forall p \geq 1 \quad (\text{A.0.8})$$

3. *The MGF of X^2 satisfies*

$$\mathbb{E} \exp(\lambda |X|) \leq \exp(K_3 \lambda), \quad \forall |\lambda| \leq 1/K_3 \quad (\text{A.0.9})$$

4. *The MGF of $|X|$ is bounded at some point, namely*

$$\mathbb{E} \exp(X/K_4) \leq 2 \quad (\text{A.0.10})$$

5. *If $\mathbb{E}X = 0$, then the MGF of X satisfies*

$$\mathbb{E} \exp(\lambda X) \leq \exp(K_5^2 \lambda^2) \quad \forall |\lambda| \leq 1/K_5 \quad (\text{A.0.11})$$

Definition A.6 (Sub-exponential random variables). A random variable X that satisfies one of the equivalent sub-exponential properties is called a sub-exponential random variable. The sub-exponential norm of X , denoted $\|X\|_{\psi_1}$, is defined to be the smallest K_4 in Theorem A.5. In other words, we define

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E} \exp(|X|/t) \leq 2\} \quad (\text{A.0.12})$$

Sub-gaussian and sub-exponential distributions have a close relationship. Firstly, it's evident that any sub-Gaussian distribution is also sub-exponential. Secondly, the square of a sub-Gaussian random variable is itself sub-exponential.

Lemma A.7 (Sub-exponential is sub-Gaussian squared [55]). *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2 \quad (\text{A.0.13})$$

More generally, the product of two sub-Gaussian random variables is sub-exponential.

Lemma A.8 (The product of sub-gaussians is sub-exponential [55]). *Let X and Y be sub-Gaussian random variables. Then XY is sub-exponential.*

Proof. Without loss of generality assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. The lemma claims that if

$$\mathbb{E} \exp(X^2) \leq 2, \quad \mathbb{E} \exp(Y^2) \leq 2 \quad (\text{A.0.14})$$

then $\mathbb{E} \exp(|XY|) \leq 2$.

To prove this, one may use Young's inequality $ab \leq a^2/2 + b^2/2$. It yields,

$$\begin{aligned} \mathbb{E} \exp(|XY|) &\leq \mathbb{E} \exp\left(\frac{X^2}{2} + \frac{Y^2}{2}\right) \\ &= \mathbb{E} \left(\exp\left(\frac{X^2}{2}\right) \exp\left(\frac{Y^2}{2}\right) \right) \\ &\leq \frac{1}{2} \mathbb{E}(\exp(X^2) + \exp(Y^2)) = 2 \end{aligned} \quad (\text{A.0.15})$$

□

We are ready to state and prove a concentration inequality for sums of independent sub-exponential random variables.

Theorem A.9 (Bernstein's inequality [55]). *Let $X_1, \dots, X_N$ be independent mean-zero sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^N X_i\right|\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{\sum_{i=1}^N K_i^2}, \frac{t}{\max_i K_i}\right)\right) \quad (\text{A.0.16})$$

where $K_i = \inf\{t > 0 : \mathbb{E} \exp(|X_i|/t) \leq 2\}$.

Appendix B

Distributional dynamics

A very useful tool for the analysis of the PDE 2.3 is provided by the following non-linear dynamics. We introduce trajectories $(\bar{X}_i^t)_{1 \leq i \leq N}$ by letting $\bar{X}_i^0 = X_i^0$ to the same initialization as for SGD and, $t \geq 0$ (here P_X denotes the law of the random variable X):

$$\bar{X}_i^t = X_i^0 - 2 \int_0^t c(s) \nabla \Psi(\bar{X}_i^s, \rho_s) ds, \quad \rho_s = P_{\bar{X}_i^s} \quad (\text{B.0.1})$$

Definition B.1 (Wasserstein distance). Let $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$ be Radon probability measure. The Wasserstein distance between μ and ν is the quantity,

$$d_W(\mu, \nu) = \sup_{f \in \mathcal{L}} \left| \int f d\mu - \int f d\nu \right| \quad (\text{B.0.2})$$

where $\mathcal{L} = \{f \in C_b(\mathbb{R}) : \|f\|_\infty \leq 1 \text{ and } f \text{ is } K\text{-Lipschitz such that } K \leq 1\}$.

Lemma B.2 (Concentration of empirical measure [40]). *Assume conditions of Theorem 2.3 hold. Let $(\rho_t)_{t \geq 0}$ be the solution of the PDE 2.3. Let $(\bar{X}_i^t)_{t \geq 0}$ be the solution of non-linear dynamics. Then $t \mapsto \bar{X}_i^t$ is $K_1 K_3$ -Lipschitz continuous, and $t \mapsto \rho_t$ is $K_1 K_3$ -Lipschitz continuous in d_W Wasserstein distance. In particular, $t \mapsto \rho_t$ is continuous in the topology of weak convergence, namely*

$$d_{BL}(\rho_t, \rho_s) \leq d_W(\rho_t, \rho_s) \leq K_1 K_3 |t - s| \quad (\text{B.0.3})$$

Appendix C

Effective dynamics

Proof of Theorem 2.24 [7]. Our aim is to establish $u_N \rightarrow u_t$ weakly as random variables on $C([0, \infty])$ where u solves

$$du_t = \mathcal{H}(u_t)dt + \sqrt{\Sigma(u_t)}dB_t \quad (\text{C.0.1})$$

as mentioned in Theorem 2.24. It is equivalent to show the same on $C([0, T])$ equipped with the sup-norm for every $T > 0$.

Let τ_K^N denote the exit time for the interpolated process $u_N(t)$ from E_K^N . Define its pre-image $E_{K,N}^* := u_N^{-1}(E_K^N)$ and let $L_{K,N}^\infty = L^\infty(E_{K,N}^*)$. For a function f , we use the shorthand f_ℓ to denote $f(X_\ell)$. By Taylor's theorem, we have that for any C^3 function f and any $\ell \leq \tau_K^N/\delta$,

$$\begin{aligned} f_\ell &= f(X_{\ell-1} - \delta \nabla \Phi_{\ell-1} - \delta \nabla H_{\ell-1}^\ell) \\ &= f_{\ell-1} + \delta [A_\ell^f - A_{\ell-1}^f] + \delta [M_\ell^f - M_{\ell-1}^f] + O\left(\delta^3 \|\nabla^3 f\|_{L_{K,N}^\infty} \cdot \|\nabla L\|_{L_{K,N}^\infty}^3\right), \end{aligned} \quad (\text{C.0.2})$$

where A_ℓ^f and M_ℓ^f are defined by their increments as follows:

$$A_\ell^f - A_{\ell-1}^f = (-\mathcal{A}_N + \delta \mathcal{L}_N) f_{\ell-1} + \frac{1}{2} \langle \nabla \Phi \otimes \nabla \Phi, \nabla^2 f \rangle_{\ell-1} \quad (\text{C.0.3})$$

$$M_\ell^f - M_{\ell-1}^f = -\langle \nabla H^\ell, \nabla f \rangle_{\ell-1} + \delta (\mathcal{E}_\ell^f - \mathcal{E}_{\ell-1}^f) \quad (\text{C.0.4})$$

$$\mathcal{E}_\ell^f - \mathcal{E}_{\ell-1}^f = \nabla^2 f(\nabla \Phi, \nabla H^\ell)_{\ell-1} + \frac{1}{2} \langle \nabla^2 f, \nabla H^\ell \otimes \nabla H^\ell - V \rangle_{\ell-1} \quad (\text{C.0.5})$$

for $\mathcal{A}_N = \langle \nabla \Phi, \nabla \rangle$, $\mathcal{L}_N = \frac{1}{2} \sum_{i,j} V_{ij} \partial_i \partial_j$ and $V = \mathbb{E}[\nabla H \otimes \nabla H]$ as in Theorem 2.24. Observe that A_ℓ^f is previsible (with respect to the filtration generated by $(Y_1, \dots, Y_{\ell-1})$), and M_ℓ^f is a martingale. We bound these for $f = u_j$ among $u_N = (u_1, \dots, u_k)$.

Recalling Definition 3.5, since u_N are δ_N -localizable, the error term in Equation C.0.2 has

$$\delta^3 \sup_{x \in E_{K,N}^*} \mathbb{E} [\|\nabla^3 u_j\| \cdot \|\nabla L\|^3] \lesssim \delta^3 \|\nabla^3 u_j\|_{L_{K,N}^\infty} \left(\|\nabla \Phi\|_{L_{K,n}^\infty}^3 + \sup_{E_{K,n}^*} \mathbb{E} \|\nabla H\|^3 \right) \lesssim_K \delta^{3/2} \quad (\text{C.0.6})$$

Since δ_N goes to 0 as $N \rightarrow \infty$, we may thus write $u_j(X_\ell)$ as

$$u_j(X_\ell) = u_j(0) + \delta \sum_{\ell' \leq \ell} (A_{\ell'}^{u_j} - A_{\ell'-1}^{u_j}) + \delta \sum_{\ell' \leq \ell} (M_{\ell'}^{u_j} - M_{\ell'-1}^{u_j}) + o(1) \quad (\text{C.0.7})$$

where the last term is $o(1)$ in L^1 uniformly for $\ell \leq \tau_K/\delta$. Now let us define for $s \in [0, T]$,

$$a'_j(s) = A_{[s/\delta]}^{u_j} - A_{[s/\delta]-1}^{u_j} \quad \text{and} \quad b'_j(s) = M_{[s/\delta]}^{u_j} - M_{[s/\delta]-1}^{u_j} \quad (\text{C.0.8})$$

If we let

$$a_j(s) = \int_0^s a'_j(s') ds' = a_j(\delta[s/\delta]) + (s - \delta[s/\delta]) (A_{[s/\delta]}^{u_j} - A_{[s/\delta]-1}^{u_j}) \quad (\text{C.0.9})$$

and similarly $b_j(s) = \int_0^s b'_j(s') ds'$, then recalling that $u_N(s)$ is the linear interpolation of $(u_j([s/\delta]))_j$, we may write

$$u_N(s) = u_N(0) + a_N(s) + b_N(s) + o(1) \quad (\text{C.0.10})$$

where $a_N(s) = (a_j(s))_j$ and $b_N(s) = (b_j(s))_j$.

We now prove that the sequence $(u_N(s \wedge \tau_K^N))$ is tight in $C([0, T])$ with limit points which are $(1/4)$ -Hölder for each K . To this end, let us define

$$v_N(s) = u_N(0) + a_N(s) + b_N(s) \quad (\text{C.0.11})$$

As the $o(1)$ error above is uniform in t , we have that

$$\sup_{0 \leq s \leq \tau_K^N} \|u_N(s) - v_N(s)\| \rightarrow 0, \quad \text{in } L^1 \quad (\text{C.0.12})$$

Thus it suffices to show the claimed tightness and Hölder properties of limit points for v_N instead of u_N . We aim to show that for all $0 \leq s, t \leq T$,

$$\mathbb{E} \|v_N(s \wedge \tau_K) - v_N(t \wedge \tau_K)\|^4 \Sigma_{K,T}(t-s)^2 \quad (\text{C.0.13})$$

from which we will get that the sequence $v_n(s \wedge \tau_K)$ is uniformly $1/4$ -Hölder by Kolmogorov's continuity theorem. Evidently, for all s, t we have that

$$\|v_N(s) - v_N(t)\| \leq \|a_N(s) - a_N(t)\| + \|b_N(s) - b_N(t)\| \quad (\text{C.0.14})$$

We control these terms in turn. We will do this coordinate wise and, for readability, fix some $j \leq k$ and let $u = u_j, a = a_j, b = b_j$ etc.

For the previsible term, we have

$$\begin{aligned} \mathbb{E}|a(s \wedge \tau_K) - a(t \wedge \tau_K)|^4 & \quad (C.0.15) \\ & \lesssim \mathbb{E} \left| \delta \sum_{\ell} ((-\mathcal{A}_N + \delta \mathcal{L}_N) u)_{\ell} \right|^4 + \mathbb{E} \left| \delta^2 \sum_{\ell} \langle \nabla \Phi \otimes \nabla \Phi, \nabla^2 u \rangle_{\ell} \right|^4 \end{aligned}$$

where these sums are over steps t ranging from $[s/\delta] \wedge \tau_K/\delta$ to $[t/\delta] \wedge \tau_K/\delta$. Let $\mathcal{H} = (h_j)_{j \leq k}$ be as in 2.3.4. Then the first term in C.0.15 satisfies

$$\begin{aligned} \mathbb{E} \left| \delta \sum_{\ell} ((-\mathcal{A}_N + \delta \mathcal{L}_N) u)_{\ell} \right|^4 & \lesssim_K \mathbb{E} \left| \delta \sum_{\ell} h_j(u_n)_{\ell} \right|^4 + o((t-s)^4) \quad (C.0.16) \\ & \leq (t-s)^4 \left(\|h_j\|_{L^\infty(E_K^n)}^4 + o(1) \right) \lesssim_K (t-s)^4 \end{aligned}$$

by continuity of h_j . For the second term in Eq C.0.15,

$$\mathbb{E} \left| \delta^2 \sum_{\ell} \langle \nabla \Phi \otimes \nabla \Phi, \nabla^2 u \rangle_{\ell} \right|^4 \leq \delta^8 \left(|(t-s)/\delta| \sup_{x \in E_{K,N}^*} \|\nabla \Phi(x)\|^2 \sup_{x \in E_{K,N}^*} \|\nabla^2 u(x)\|_{\text{op}} \right)^4 \quad (C.0.17)$$

which is $\Sigma_K \delta^2 (t-s)^4$ by items (1)-(2) δ_N -localizability. (Applying this bound for $s=0, t=T$, the last term in a is vanishing in the limit for each K whenever $\delta_N = o(1)$.) Combining the above bounds yields

$$\mathbb{E}|a(s \wedge \tau_K) - a(t \wedge \tau_K)|^4 \lesssim_K (t-s)^4 \quad (C.0.18)$$

For the martingale term, notice that by Burkholder's inequality,

$$\mathbb{E}|b(s \wedge \tau_K) - b(t \wedge \tau_K)|^4 = \mathbb{E} \left[\left(\delta \sum_{\ell} (M_{\ell}^u - M_{\ell-1}^u) \right)^4 \right] \lesssim \mathbb{E} \left[\left(\delta^2 \sum_{\ell} (M_{\ell}^u - M_{\ell-1}^u)^2 \right)^2 \right] \quad (C.0.19)$$

where the sum again runs over steps t ranging from $[s/\delta] \wedge \tau_K$ to $[t/\delta] \wedge \tau_K$. Repeatedly using the inequality $(x+y+z)^2 \lesssim x^2 + y^2 + z^2$, it suffices to bound the above quantity for each of the three terms defining the martingale difference $M_{\ell}^u - M_{\ell-1}^u$ respectively.

For the first term in that martingale difference, observe that

$$\begin{aligned} \left[\left(\delta^2 \sum_{\ell} \langle \nabla H^{\ell}, \nabla u \rangle_{\ell-1}^2 \right)^2 \right] & = \delta^4 \sum_{\ell, \ell'} \mathbb{E} \left[\langle \nabla H^{\ell}, \nabla u \rangle_{\ell-1}^2 \langle \nabla H^{\ell'}, \nabla u \rangle_{\ell'-1}^2 \right] \quad (C.0.20) \\ & \leq \left(\delta \sum_{\ell} \left(\delta^2 \mathbb{E} \langle \nabla H^{\ell}, \nabla u \rangle_{\ell-1}^4 \right)^{1/2} \right)^2 \lesssim_K (t-s)^2 \end{aligned}$$

where in the second line we used Cauchy-Schwarz and in the last we used item (3) of δ_N -localizability.

For the second term in the martingale difference,

$$\begin{aligned} & \mathbb{E} \left[\left(\delta^4 \sum_{\ell} \left(\nabla^2 u (\nabla \Phi, \nabla H^{\ell})_{\ell-1} \right)^2 \right)^2 \right] \\ & \leq \delta^6 (t-s)^2 \sup_{x \in E_{K,n}^*} \|\nabla^2 u(x)\|_{\text{op}}^4 \cdot \|\nabla \Phi(x)\|^4 \cdot \mathbb{E} \|\nabla H(x)\|^4 \lesssim_K \delta^2 (t-s)^2, \end{aligned} \quad (\text{C.0.21})$$

by items (1)-(2) of δ_N -localizability. Finally, by the same reasoning, for the third term,

$$\begin{aligned} & \mathbb{E} \left[\left(\delta^4 \sum_{\ell} \langle \nabla^2 u, \nabla H^{\ell} \otimes \nabla H^{\ell} - V \rangle_{\ell-1}^2 \right)^2 \right] \\ & \lesssim \delta^6 (t-s)^2 \sup_{x \in E_{K,N}^+} \|\nabla^2 u(x)\|_{\text{op}}^4 \cdot \mathbb{E} [\|\nabla H(x)\|^8] \lesssim_K (t-s)^2. \end{aligned} \quad (\text{C.0.22})$$

All of the above terms are $O((t-s)^2)$ since $0 \leq s, t \leq T$. Thus we have the claimed [C.0.13](#), and by Kolmogorov's continuity theorem, $(v_N(s \wedge \tau_K))_s$, are uniformly 1/4-Holder and thus the sequence is tight with 1/4-Holder limit points. Notice furthermore that if we look at $(v_N(t \wedge \tau_K) - a_N(t \wedge \tau_K))_t$, this sequence is also tight and the limits points are continuous martingales. Let us examine their limiting quadratic variations.

Let $v_N^K(t) = v_N(t \wedge \tau_K)$ and define $a_N^K(t)$ and $b_N^K(t)$ analogously. Furthermore, let $v^K(t)$, $a^K(t)$ and $b^K(t)$ be their respective limits which we have shown to exist and be 1/4-Holder.

We will compute the limiting quadratic variation for $b^K(t)$. For ease of notation, let $\Delta M_{\ell}^{u_i} = M_{\ell}^{u_i} - M_{\ell-1}^{u_i}$ and $\Delta \mathcal{E}_{\ell}^{u_i} = \mathcal{E}_{\ell}^{u_i} - \mathcal{E}_{\ell-1}^{u_i}$. Notice first that for $1 \leq i, j \leq k$,

$$b_{N,i}^K(t) b_{N,j}^K(t) - \int_0^t \delta \mathbb{E} \left[\Delta M_{[s/\delta] \wedge \tau_K}^{u_i} \Delta M_{[s/\delta] \wedge \tau_K}^{u_j} \right] ds \quad (\text{C.0.23})$$

is a martingale. We therefore need to consider the limit as $N \rightarrow \infty$ of the integral above. Write

$$\begin{aligned} \mathbb{E} [\Delta M_{\ell}^{u_i} \Delta M_{\ell}^{u_j}] &= \langle \nabla u_i, V \nabla u_j \rangle + \delta \mathbb{E} [\langle \nabla H^{\ell}, \nabla u_i \rangle_{\ell-1} \Delta \mathcal{E}_{\ell}^{u_j}] \\ &+ \delta \mathbb{E} [\Delta \mathcal{E}_{\ell}^{u_i} \langle \nabla H^{\ell}, \nabla u_j \rangle_{\ell-1}] + \delta^2 \mathbb{E} [\Delta \mathcal{E}_{\ell}^{u_i} \Delta \mathcal{E}_{\ell}^{u_j}] \end{aligned} \quad (\text{C.0.24})$$

Consider the integrals of δ times each of these four terms separately. For the first term,

$$\sup_{t \leq T} \left| \int_0^t \delta \langle \nabla u_i, V \nabla u_j \rangle_{[s/\delta] \wedge \tau_K} - \Sigma_{ij} (v_N^K(s)) ds \right| \leq T \sup_{x \in E_{K,N}^*} |\delta \langle \nabla u_i, V \nabla u_j \rangle(x) - \Sigma_{ij}(u_N(x))| \quad (\text{C.0.25})$$

goes to zero as $N \rightarrow \infty$.

We now reason that the integrals of δ times the other three terms in C.0.24 all go to zero as $N \rightarrow \infty$. The second and third are identical: by Cauchy-Schwarz,

$$\sup_{x \in E_{K,N}^*} |\delta^2 \mathbb{E} [\langle \nabla H, \nabla u_i \rangle \Delta \varepsilon_\ell^{u_j}]| \leq \delta^2 \mathbb{E} [\langle \nabla H, \nabla u_i \rangle^2]^{1/2} \mathbb{E} [(\Delta \varepsilon_\ell^{u_i})^2]^{1/2} \quad (\text{C.0.26})$$

The first expectation contributes $\delta^{-1/2}$ by the first part of item (3) of localizability. Also,

$$\mathbb{E} [(\Delta \varepsilon_\ell^{u_i})^2]^{1/2} \lesssim \mathbb{E} [\langle \nabla^2 u_i, \nabla \Phi \otimes \nabla H \rangle^2]^{1/2} + \mathbb{E} [\langle \nabla^2 u_i, \nabla H \otimes \nabla H - V \rangle^2]^{1/2} \quad (\text{C.0.27})$$

The first of these terms is at most δ^{-1} as argued in C.0.21. The second is $o(\delta^{-3/2})$ by the second part of item (3) in the definition of localizability. As such, we are able to conclude that

$$\sup_{t \leq T} \left| \int_0^t \delta^2 \mathbb{E} [\langle \nabla H, \nabla u_i \rangle_{[s/\delta] \wedge \tau_K} \Delta \varepsilon_{[s/\delta] \wedge \tau_K}^{u_j}] ds \right| \rightarrow 0 \quad (\text{C.0.28})$$

as $N \rightarrow \infty$. The integral of δ times the fourth term in C.0.24 is handled similarly using Cauchy-Schwarz and the bound of $o(\delta^{-3/2})$ on C.0.27.

Altogether, we end up with

$$\lim_{N \rightarrow \infty} \sup_{i,j \leq k} \sup_{t \leq T} \left| \int_0^t \delta \mathbb{E} [\Delta M_{[s/\delta] \wedge \tau_K}^{u_i} \Delta M_{[s/\delta] \wedge \tau_K}^{u_j}] ds - \int_0^t \Sigma_{ij} (v_N^K(s)) ds \right| = 0 \quad (\text{C.0.29})$$

Thus, if we consider the continuous martingales given by $b^K(t)$, its angle bracket is, by definition, given by

$$\langle b^K \rangle_t = \int_0^t \Sigma (v^K(s)) ds \quad (\text{C.0.30})$$

By Ito's formula for continuous martingales, we have that $f(v_t) - \int_0^t Lf(v_s) ds$ is a martingale for all $f \in C_0^\infty(\mathbb{R}^k)$, where

$$L = \frac{1}{2} \sum_{i,j=1}^k \Sigma_{ij} \partial_i \partial_j - \sum_{i=1}^k h_i \partial_i \quad (\text{C.0.31})$$

Since, by assumption, $h, \sqrt{\Sigma}$ are locally Lipschitz-and thus Lipschitz on E_K -this property uniquely characterizes the solutions to the SDE of Theorem 2.24. Thus v_K converges to the solution of Theorem 2.24 stopped at τ_K . By a standard localization argument, every limit point $v(t)$ of $v_N(t)$ solves the SDE of Theorem 2.24 (using here that E_K is an exhaustion by compact sets of $\mathbb{R}^k$). $\square$