Trainability and Accuracy of Artificial Neural Networks: An Interacting Particle System Approach

G. M. ROTSKOFF
Stanford University

E. VANDEN-EIJNDEN
Courant Institute

Abstract

Neural networks, a central tool in machine learning, have demonstrated remarkable, high fidelity performance on image recognition and classification tasks. These successes evince an ability to accurately represent high-dimensional functions, but rigorous results about the approximation error of neural networks after training are few. Here we establish conditions for global convergence of the standard optimization algorithm used in machine learning applications, stochastic gradient descent (SGD), and quantifying the scaling of its error with the size of the network. This is done by reinterpreting SGD as the evolution of a particle system with interactions governed by a potential related to the objective or "loss" function used to train the network. We show that, when the number n of units is large, the empirical distribution of the particles descends on a convex landscape towards the global minimum at a rate independent of n, with a resulting approximation error that universally scales as $O(n^{-1})$. These properties are established in the form of a law of large numbers and a central limit theorem for the empirical distribution. Our analysis also quantifies the scale and nature of the noise introduced by SGD and provides guidelines for the step size and batch size to use when training a neural network. We illustrate our findings on examples in which we train neural networks to learn the energy function of the continuous 3-spin model on the sphere. The approximation error scales as our analysis predicts in as high a dimension as $d = 25$. © 2022 Courant Institute of Mathematics and Wiley Periodicals LLC.

Contents

1 Introduction

While both speech recognition and image classification remain active areas of research, extraordinary progress has been made on both problems—ones that appeared intractable only a decade ago [24]. By harvesting the power of neural networks while simultaneously benefiting from advances in computational hardware, complex tasks such as automatic language translation are now routinely performed by computers with a high degree of reliability. The underlying explanation for these significant advances seems to be related to the expressive power of neural networks, and their ability to accurately represent high-dimensional functions.

These successes open exciting possibilities in applied and computational mathematics that are only beginning to be explored [7–9, 17, 21, 30, 36]. Any numerical calculation that uses a given function begins with a finite-dimensional approximation of that function. Because standard approximations, e.g., Galerkin truncations or finite element decompositions, suffer from the curse of dimensionality; it is nearly impossible to scale such methods to large dimensions d. Fundamentally, these representations are linear combinations of basis functions. The issue arises because the dimensionality of the representation is equal to that of the truncation. Neural networks, on the other hand, are highly nonlinear in their adjusting parameters. As a result, the effective dimensionality of a neural network is much higher than its total number of parameters, which may explain the impressive function approximation capabilities observed in practice, even when d is large. Characterizing this observation with analysis is nontrivial though, precisely because the representation of a function by a neural network is nonlinear in its parameters. This renders many of the standard tools of numerical analysis useless, since they are in large part based on linear algebra.

The significant achievements of machine learning have inspired many efforts to provide theoretical justification to a vast and growing body of empirical knowledge. At the core of our understanding of the approximation properties of neural networks are the well-known "universal approximation theorems" that specify the conditions under which a neural network can represent a target function with arbitrary accuracy [6, 14, 28]. Despite the power of these results, they do not indicate how the network parameters should be optimized to achieve maximal accuracy in practice [11]. In particular, these theorems do not provide general guidance on how the error of the network scales with its size at the end of training. Several recent papers have focused on the analysis of the shape and properties of the objective or "loss" function landscape [5, 13, 29]. These studies have mainly focused on the fine features of this landscape, trying to understand how nonconvex it is and making analogies with glassy landscapes. Additionally, some analysis has been performed in cases where the number of parameters vastly exceeds the amount of

training data, a setting that guarantees convexity and dramatically simplifies the landscape. Further studies have examined the dynamics of the parameters on the loss landscape to understand the properties of optimization procedures based on stochastic gradient descent.

In this paper, we adopt a different perspective which enables powerful tools for analysis. Similar to what was recently proposed in [12, 27, 33], we view the parameters in the network as particles and the loss function as a potential that dictates the interaction between them. Correspondingly, training the network can be interpreted as the evolution of the particles in this interaction potential. Using the interchangeability of the n interacting particles/parameters in the neural representation, we focus on their empirical distribution and analyze its properties when n is large using standard limit theorems [22, 23, 31, 32]. This viewpoint allows us to bypass many of the difficulties that arise with approaches that attempt to study the dynamics of the individual particles. In particular:

(1) We derive an evolution equation for the empirical distribution of the particles, and show that it evolves by gradient descent in the 2-Wasserstein metric on a convex energy landscape. This observation allows us to assert that convergence towards equilibrium of the empirical distribution occurs on a time scale that is independent of n to leading order—similar results were obtained in [12, 27, 33]. The results are obtained in the form of law of large numbers (LLN) for the empirical distribution of the parameters. As a consequence, we rederive the universal approximation theorem and establish that it can be realized dynamically.

(2) We quantify the fluctuations of the empirical distribution at finite n above its limit. We show that these fluctuations are of order $O(n^{-1/2})$ and controlled at all $t < \infty$. In addition, we establish conditions under which these fluctuations heal and become $O(n^{-1})$ as $t \to \infty$. These results rely on a central limit theorem (CLT) and indicate that the neural network approximation error is universal and scales as $O(n^{-1})$ as $n \to \infty$ in any d.

These results are established first in situations where gradient descent (GD) on the loss function is used to optimize or "train" the parameters in the network, and then shown to also apply in the context of stochastic gradient descent (SGD). In the latter case, our analysis sheds light on the nature of the noise introduced in SGD, and indicates how the time step and the batch size should be scaled to achieve the optimal error. We briefly elaborate on these statements below, first precisely formulating the problem.

1.1 Problem setup

Given a function $f : \Omega \to \mathbb{R}$ defined on the closed manifold $\Omega \subseteq \mathbb{R}^d$, consider its approximation by a neural network of the form

$$(1.1) \qquad f^{(n)}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} c_i \widehat{\varphi}(\boldsymbol{x}, z_i)$$

where $n \in \mathbb{N}$, $(c_i, z_i) \in D \equiv \mathbb{R} \times \widehat{D}$ are parameters to be learned for $i = 1, \dots, n$, and $\varphi : \Omega \times D \to \mathbb{R}$ is some function—we assume throughout this paper that $\widehat{D}$ is a closed manifold in $\mathbb{R}^N$. The function $\widehat{\varphi}$ is usually referred to as the "nonlinearity" or "unit" and n as the width of the network. To simplify notation, we use $\boldsymbol{\theta} = (c, z) \in D$ and $\varphi(\boldsymbol{x}, \boldsymbol{\theta}) = c\widehat{\varphi}(\boldsymbol{x}, z)$ in terms of which (1.1) reads

$$(1.2) \qquad f^{(n)}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{x}, \boldsymbol{\theta}_i).$$

Many models used in machine learning can be cast in the form (1.1)–(1.2):

- **Radial basis function networks.** In this case $\widehat{D} \equiv \Omega$ and $\widehat{\varphi}(\boldsymbol{x}, z) \equiv \phi(\boldsymbol{x} - z)$ where ϕ is some kernel, for example, that of a radial function such as
  $$\phi(\boldsymbol{x}) = \exp\left(-\tfrac{1}{2}\kappa |\boldsymbol{x}|^2\right)$$
  where $\kappa > 0$ is a fixed constant.

- **Single hidden layer neural networks.** In this case, $\widehat{D} \subset \mathbb{S}^d$ and $\widehat{\varphi}(\boldsymbol{x}, z) = \widehat{\varphi}(\boldsymbol{x}, \boldsymbol{a}, b)$ with, e.g., $\boldsymbol{a} \in \mathbb{S}^{d-1}$, $b \in [-1, 1]$, and
  $$\varphi(\boldsymbol{x}, \boldsymbol{a}, b) = h(\boldsymbol{a} \cdot \boldsymbol{x} + b)$$
  where $h : \mathbb{R} \to \mathbb{R}$ is, e.g., a sigmoid function $h(z) = 1/(1 + e^{-z})$ or a rectified linear unit (ReLU) $h(z) = \max(z, 0)$.

- **Iterated neural networks.** These are structurally similar to single hidden layer neural networks. For example, to construct a two-layer network we take h as above and for $m \in \mathbb{N}$, $m \le d$ define $\boldsymbol{h}^{(1)} : \mathbb{R}^m \to \mathbb{R}^m$ such that
  $$h_j^{(1)}(\boldsymbol{v}) = h(v_j), \quad \boldsymbol{v} = (v_1, \dots, v_m) \in \mathbb{R}^m, \quad j = 1, \dots, m$$
  then set
  $$f^{(n)}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} c_i h\left(\boldsymbol{a}_i^{(0)} \cdot \boldsymbol{h}^{(1)}\left(A_i^{(1)}\boldsymbol{x} + \boldsymbol{b}_i^{(1)}\right) + b_i^{(0)}\right)$$
  where $\boldsymbol{a}_i^{(0)} \in \mathbb{R}^m$, $b_i^{(0)} \in \mathbb{R}$, $A_i^{(1)} \in \mathbb{R}^{m \times d}$, $\boldsymbol{b}_i^{(1)} \in \mathbb{R}^m$, $i = 1, \dots, n$. Therefore here we have $z = (\boldsymbol{a}^{(0)}, b^{(0)}, A^{(1)}, \boldsymbol{b}^{(1)}) \in \widehat{D} \subset \mathbb{R}^{m+1+m \times d + m}$ (where with a slight abuse of notation we view the matrix $A^{(1)}$ as having a vector in $\mathbb{R}^{m \times d}$). Three-layer networks can be constructed similarly. Note that our results apply to deep neural networks when their final layer grows large, with fixed depth.

To measure the discrepancy between the target function f and its neural network approximation $f^{(n)}$, we need to introduce a distance, or loss function, between f and $f^{(n)}$. A natural candidate often used in practice is

$$(1.3) \qquad \mathcal{L}[f^{(n)}] = \frac{1}{2} \int_{\Omega} |f(x) - f^{(n)}(x)|^2 \nu(dx) = \frac{1}{2} \mathbb{E}_{\nu} |f - f^{(n)}|^2$$

where ν, the data distribution, is some positive measure on Ω such that $\nu(\Omega) < \infty$ (for example, the Hausdorff measure on Ω, which we will denote by dx). We can view $\mathcal{L}[f^{(n)}]$ as an objective function for $\{\theta_i\}_{i=1}^n$:

$$(1.4) \qquad \mathcal{L}[f^{(n)}] = C_f - \frac{1}{n} \sum_{i=1}^n F(\theta_i) + \frac{1}{2n^2} \sum_{i,j=1}^n K(\theta_i, \theta_j)$$

where $C_f = \frac{1}{2} \mathbb{E}_{\nu} |f|^2$, and we defined

$$(1.5) \qquad F(\theta) = c\widehat{F}(z), \quad K(\theta, \theta') = cc'\widehat{K}(z, z')$$

with

$$(1.6) \qquad \begin{aligned} \widehat{F}(z) &= \mathbb{E}_{\nu}[f\widehat{\varphi}(\cdot, z)], \\ \widehat{K}(z, z') &= \mathbb{E}_{\nu}[\widehat{\varphi}(\cdot, z)\widehat{\varphi}(\cdot, z')] \equiv \widehat{K}(z', z). \end{aligned}$$

Trying to minimize (1.4) over $\{\theta_i\}_{i=1}^n$ leads to difficulties, however, since this is potentially (and presumably) a nonconvex optimization problem, which has local minimizers. In particular, if we perform training by making $\{\theta_i\}_{i=1}^n$ evolve via gradient descent (GD) over the loss, i.e., if we use

$$(1.7) \qquad \dot{\theta}_i = \nabla F(\theta_i) - \frac{1}{n} \sum_{j=1}^n \nabla K(\theta_i, \theta_j),$$

there is no guarantee a priori that these parameters will reach the global minimum of the loss or even a local minimum with a value for the loss that is close to that of the global minimum. As a result, determining the value of (1.3) after training (and its scaling with n, say) is nontrivial. It is therefore natural to ask:

> *How accurate is the approximation* (1.1)–(1.2) *if we optimize* $\{\theta_i\}_{i=1}^n$ *by applying the algorithms commonly used in machine learning?*

This is the main question we investigate in the present paper.

1.2 Main results and organization

We will consider the evolution in time of the representation

$$(1.8) \qquad f_t^{(n)} = \frac{1}{n} \sum_{i=1}^n \varphi(\cdot, \theta_i(t))$$

and study the behavior of this function for large n and large t. To this end we use tools from interacting particle systems, and also build on known results about the

nature of the loss function (1.4) in the limit as $n \to \infty$—these results are recalled in Section 2.

In Section 3 we consider the situation where $\theta_i(t)$ are the solution of the GD flow (1.7)—as explained below, this is somewhat of an idealized situation since we typically must work with the empirical loss rather than the exact one, but it is more easily amenable to analysis. By looking at the evolution of the empirical distribution of $\{\theta_i(t)\}_{i=1}^n$ rather than that of the parameters themselves, we establish a law of large numbers (LLN) for $f_t^{(n)}$, namely that $\lim_{n\to\infty} f_t^{(n)} = f_t$, where f_t evolves as

$$(1.9) \qquad \partial_t f_t(x) = -\int_\Omega M_t(x, x') \left(f_t(x') - f(x') \right) \nu(dx'),$$

where f is the target function and $M_t(x, x')$ a positive semidefinite kernel whose form is explicit—see Proposition 3.3. The evolution equation (1.9) can be interpreted as GD for f_t over the loss in some metric inherited from the 2-Wasserstein metric, and in Proposition 3.5 we show the flow converges to the target function, i.e.,

$$(1.10) \qquad \lim_{t\to\infty} f_t = \lim_{t\to\infty} \lim_{n\to\infty} f_t^{(n)} = f.$$

We also establish that the limit in n and t commute. Regarding the scaling of the fluctuations above f_t when n is finite, in Proposition 3.7 we establish a central limit theorem (CLT) that asserts that these fluctuations are of size $O(n^{-1/2})$, i.e., $n^{1/2}(f_t^{(n)} - f_t)$ has a limit in law as $n \to \infty$. In addition, in Proposition 3.8 we show that these fluctuations are controlled at all times, and in Proposition 3.9 that under certain conditions they heal as $t \to \infty$ in the sense that

$$(1.11) \qquad f_{a_n}^{(n)} = f + O(n^{-1}) \quad \text{as } n \to \infty \text{ with } a_n/\log n \to \infty.$$

In Section 4 we analyze the typical situation in which it is not possible to calculate (1.3) or (1.6) exactly. Rather, we must approximate these expectations using a "training set"; i.e., a set of points $\{x_p\}_{p=1}^P$ distributed according to ν on which f is known, so that instead of $\mathcal{L}[f^{(n)}]$ we must use

$$(1.12) \qquad \mathcal{L}_P[f^{(n)}] = \frac{1}{P} \sum_{p=1}^P \left| f(x_p) - f^{(n)}(x_p) \right|^2$$

and instead of $\hat{F}$ and $\hat{K}$

$$\hat{F}_P(z) = \frac{1}{P} \sum_{p=1}^P f(x_p) \hat{\varphi}(x_p, z),$$

$$(1.13)$$

$$\hat{K}_P(z, z') = \frac{1}{P} \sum_{p=1}^P \hat{\varphi}(x_p, z) \hat{\varphi}(x_p, z').$$

If in (1.7) we replace $F(\theta)$ and $K(\theta, \theta')$ by their empirical estimates over a subset of the training points $F_P(\theta) = c\widehat{F}_P(z)$ and $K_P(\theta, \theta') = cc'\widehat{K}_P(z, z')$, we arrive at what is referred to as stochastic gradient descent (SGD)—the method of choice to train neural networks. We focus on situations in which we can redraw the training set as often as we need, namely, at every step during the learning process, an algorithm called online learning. In this case, in the limit as the optimization time step Δt used in SGD tends to 0, SGD becomes asymptotically equivalent to an SDE whose drift terms coincide with those of GD but with multiplicative noise terms added. In this context, we establish that (1.9) and (1.10) also hold if we choose the size P of the batch used in (1.13) at every SGD step such that as $P = O(n^{2\alpha})$ with $\alpha > 0$. Regarding the scaling of the fluctuations, if we set $\alpha \in (0, 1)$, we lose accuracy and (1.11) is replaced by

$$(1.14) \qquad f_{a_n}^{(n)} = f + O(n^{-\alpha}) \quad \text{as } n \to \infty \text{ with } a_n / \log n \to \infty.$$

However if $\alpha \geq 1$, we get (1.11) back (meaning also that there is no advantage in taking α bigger than 1). These results are stated in Propositions 4.3 and 4.4.

In Section 5 we illustrate these results, using a spherical p-spin model with $p = 3$ as test function to represent with a neural network. We show that the network accurately approximates this function in up to $d = 25$ dimensions, with a scaling of the error consistent with the results established in Sections 3 and 4. These results are obtained using both a radial basis function network, and a single hidden layer network using sigmoid functions.

Concluding remarks are made in Section 6, and in the Appendix we establish a finite-temperature variant (Langevin dynamics) of (1.11), which applies when additive noise terms are added in the GD equations (1.7). This result reads

$$(1.15) \qquad \lim_{T \to -\infty} f_t^{(n)} = f + n^{-1} \widetilde{f}_t + o(n^{-1}) \quad \text{with } \widetilde{f}_t = \beta^{-1} \epsilon^* + \beta^{-1/2} \widetilde{\epsilon}_t$$

where T is the time at which we initiate the training. Here $\beta > 0$ is a parameter playing the role of inverse temperature, $\epsilon^* : \Omega \to \mathbb{R}$ is some given (nonrandom) function and $\widetilde{\epsilon}_t : \Omega \to \mathbb{R}$ is a Gaussian process with mean zero and covariance $\mathbb{E}[\widetilde{\epsilon}_t(x)\widetilde{\epsilon}_t(x')] \propto \delta(x - x')$. Note that (1.15) gives (1.11) back after quenching (i.e., by sending $\beta \to \infty$). The result in (1.15) is stated in Proposition A.5

As we have emphasized, our approach has strong ties with the statistical mechanics of systems of large numbers of interacting particles. Our main aim here is to introduce a framework showing how results and concepts developed in this context are useful to address questions in machine learning. Conversely, we seek to illustrate that ML provides new mathematical questions about an interesting class of particle systems. With this in mind, we adopt a presentation style that relies on formal asymptotic arguments to derive our results, though we are confident that providing rigorous proofs to our propositions is achievable. To a certain extent, this program was already started in [12, 27, 33].

2 Functional Formulation of the Learning Problem

As discussed in Bach [4], it is useful to give conditions under which (1.1) has a limit as $n \to \infty$, for two main reasons: First it shows which functions can be represented as in (1.1) if we allow the number of units n to grow to infinity, and second, while the loss function (1.3) may be nonconvex for $\{\boldsymbol{\theta}_i\}_{i=1}^n$, the limiting functional for the parameter distribution is convex.

2.1 Universal approximation theorem

Consider the space $\mathcal{F}_1$ of all functions that can be represented as

$$(2.1) \qquad f = \int_{\widehat{D}} \widehat{\varphi}(\cdot, z) \gamma(dz),$$

where γ is some (signed) Radon measure on D with finite total variation (L^1-norm), $|\gamma|_{\text{TV}} = \int_{\widehat{D}} |\gamma(dz)| < \infty$: we will denote the space of these Radon measures by $\mathcal{M}(\widehat{D})$ and that of probability measures by $\mathcal{M}_+(\widehat{D})$. The space $\mathcal{F}_1$ is important in our context, since any $f \in \mathcal{F}_1$ can be realized as

$$(2.2) \qquad f = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n c_i \widehat{\varphi}(\cdot, z_j)$$

by drawing $\{c_i, z_i\}_{i \in \mathbb{N}}$ as follows. Start from the Jordan decomposition for γ [10],

$$(2.3) \qquad \gamma = \gamma_+ - \gamma_-,$$

where γ_+ and γ_- are positive measures with $\text{supp}\, \gamma_+ \cup \text{supp}\, \gamma_- = \text{supp}\, \gamma$ and $\text{supp}\, \gamma_+ \cap \text{supp}\, \gamma_- = \varnothing$. Using this decomposition, we can draw z_i's independently from $(\gamma_+ + \gamma_-)/|\gamma|_{\text{TV}} \in \mathcal{M}_+(\widehat{D})$, where $|\gamma|_{\text{TV}} = \int_{\widehat{D}}(\gamma_+(dz) + \gamma_-(dz)) < \infty$, and set $c_i = +|\gamma|_{\text{TV}}$ if $z_i \in \text{supp}\, \gamma_+$ and $c_i = -|\gamma|_{\text{TV}}$ if $z_i \in \text{supp}\, \gamma_-$. By the law of large numbers we then have

$$(2.4) \qquad \gamma_n = \frac{1}{n} \sum_{i=1}^n c_i \delta_{z_i} \rightharpoonup \gamma \quad \text{as } n \to \infty.$$

which implies (2.2). We can also use the central limit theorem to get an approximation error at finite n, a calculation we carry out in Section 3.4.

Since the space $\mathcal{F}_1$ depends on the choice of unit $\widehat{\varphi}$, to characterize it we make the following assumptions:

ASSUMPTION 2.1. *Both the input space Ω and the feature space $\widehat{D}$ are closed (i.e., compact with no boundaries) smooth Riemannian manifolds. The unit is continuously differentiable in z, i.e., $\forall x \in \Omega$, $\widehat{\varphi}(x, \cdot) \in C^1(\widehat{D})$.*

ASSUMPTION 2.2 (Discriminating unit). *The unit satisfies*

$$(2.5) \qquad \int_{\Omega} g(x)\widehat{\varphi}(x, \cdot)\nu(dx) = 0 \quad \text{a.e. in } \widehat{D} \Rightarrow g = 0 \text{ a.e. in } \Omega.$$

While ReLU is not strictly differentiable at $x = 0$, the uniqueness of the gradient flow can be ensured through the use of subgradients [12], The differentiability of $\widehat{\varphi}$ in z is required to guarantee uniqueness of the GD flow.

THEOREM 2.3 (Universal approximation theorem [6, 14, 28]). *Under Assumptions 2.1 and 2.2, $\mathcal{F}_1$ is a dense subspace of $L^2(\Omega, \nu)$; i.e., given any $f \in L^2(\Omega, \nu)$ and $\epsilon > 0$, there exists $\gamma^* \in \mathcal{M}(\widehat{D})$ such that $|\gamma^*|_{TV} < \infty$ and*

$$(2.6) \qquad f^* = \int_{\widehat{D}} \widehat{\varphi}(\cdot, z)\gamma^*(dz) \in \mathcal{F}_1$$

satisfies

$$(2.7) \qquad \|f - f^*\|_{L^2(\Omega, \nu)} \leq \epsilon.$$

A similar theorem was originally stated in [14]. Since its proof is elementary let us reproduce it here:

PROOF. The space $\mathcal{F}_1$ is a linear subspace of $L^2(\Omega, \nu)$ since, if

$$f = \int_{\widehat{D}} \widehat{\varphi}(\cdot, z)\gamma(dz) \in \mathcal{F}_1,$$

$$\|f\|_{L^2(\Omega, \nu)}^2 = \int_{\Omega} \left(\int_{\widehat{D}} \widehat{\varphi}(x, z)\gamma(dz) \right)^2 \nu(dx)$$

$$= \int_{\widehat{D} \times \widehat{D}} \widehat{K}(z, z')\gamma(dz)\gamma(dz')$$

$$\leq \|\widehat{K}\|_{\infty} |\gamma|_{TV}^2 < \infty$$

where we used $\|\widehat{K}\|_{\infty} = \sup_{(z,z') \in \widehat{D} \times \widehat{D}} |\widehat{K}(z, z')| < \infty$, which follows from Assumption 2.1. To show that $\mathcal{F}_1$ is dense in $L^2(\Omega, \nu)$, we proceed by contradiction. Assuming that $\mathcal{F}_1$ is not dense, by the Hahn-Banach theorem, there exists a nonzero linear functional $L : L^2(\Omega, \nu) \to \mathbb{R}$ such that $Lf = 0$ for all $f \in \mathcal{F}_1$. By the Riesz representation theorem, the action of L on f can be represented as the inner product in $L^2(\Omega, \nu)$ between f and some $g \in L^2(\Omega, \nu)$; i.e., there must exist $g \neq 0$ such that for all $f = \int_{\widehat{D}} \widehat{\varphi}(\cdot, z)\gamma(dz) \in \mathcal{F}_1$ (i.e., all $\gamma \in \mathcal{M}(\widehat{D})$ with finite variation)

$$0 = \int_{\Omega} g(x) \left(\int_{\widehat{D}} \widehat{\varphi}(x, z)\gamma(dz) \right) \nu(dx)$$

$$= \int_{\widehat{D}} \left(\int_{\Omega} g(x)\widehat{\varphi}(x, z)\nu(dx) \right) \gamma(dz),$$

This requires that

$$0 = \int_{\Omega} g(x)\widehat{\varphi}(x, \cdot)\nu(dx) \quad \text{a.e. in } \widehat{D}.$$

which, by Assumption 2.2, implies that $g = 0$ a.e. in Ω, a contradiction. $\qquad \square$

From now on, we will make the following:

ASSUMPTION 2.4. *The target function is representable by the network, i.e.,* $f \in \mathcal{F}_1$.

This means that we can take $f = f^*$ in Theorem 2.3.

2.2 Convexification at distributional level

Another advantage of taking the $n \to \infty$ limit of (1.1) is that it turns (1.4) into a quadratic objective function for γ:

$$(2.8) \quad \begin{aligned} &\mathcal{L}[\int_D \widehat{\varphi}(\cdot, z)\gamma(dz)] \\ &= C_f - \int_{\widehat{D}} \widehat{F}(z)\gamma(dz) + \frac{1}{2} \int_{\widehat{D} \times \widehat{D}} \widehat{K}(z, z')\gamma(dz)\gamma(dz') \end{aligned}$$

This means that minimizing (2.8) over γ rather than (1.4) over $\{\theta_i\}_{i=1}^n$ is conceptually simpler. In particular, any minimizer γ^* of (2.8) solves the linear Euler-Lagrange equation:

$$(2.9) \qquad \forall z \in \widehat{D} : \widehat{F}(z) = \int_{\widehat{D}} \widehat{K}(z, z')\gamma^*(dz'),$$

and the loss evaluated on any γ^* has value zero. Indeed, using the definitions of $\widehat{F}$ and $\widehat{K}$ in (1.6), (2.9) can be written as

$$(2.10) \qquad \int_{\Omega} \widehat{\varphi}(x, z) \left(f(x) - \int_{\widehat{D}} \widehat{\varphi}(x, z')\gamma^*(dz') \right), \nu(dx) = 0$$

which, by Assumptions 2.2 and 2.4, has a solution such that $f = \int_{\widehat{D}} \widehat{\varphi}(\cdot, z)\gamma^*(dz)$, and, as a result, the loss evaluated on $\int_{\widehat{D}} \widehat{\varphi}(\cdot, z)\gamma^*(dz)$ is zero.

Of course, the results above are not necessarily an assurance of convergence in practice. Indeed, we do not know how to pick $\gamma \in \mathcal{M}(\widehat{D})$ to represent an $f \in \mathcal{F}_1$, nor can we manipulate these Radon measures explicitly: rather we will have to learn finite n approximations of the form $\gamma^{(n)} = n^{-1} \sum_{i=1}^n c_i \delta_{z_i}$ by adjusting the parameters $\{\theta_i\}_{i=1}^n = \{c_i, z_i\}_{i=1}^n$ dynamically. Furthermore, even though the energy can be expressed in terms of $\gamma^{(n)}$, as we will see the dynamics can only be closed at the level of the empirical distribution

$$(2.11) \quad \mu^{(n)}(dc, dz) = \frac{1}{n} \sum_{i=1}^n \delta_{c_i}(dc)\delta_{z_i}(dz) \equiv \frac{1}{n} \sum_{i=1}^n \delta_{\theta_i}(d\theta) = \mu^{(n)}(d\theta)$$

with $\gamma^{(n)}$ given by

$$(2.12) \qquad \gamma^{(n)} = \int_{\mathbb{R}} c\mu^{(n)}(dc, \cdot).$$

Viewed as a functional of $\mu \in \mathcal{M}_+(D)$ such that $\int_{\mathbb{R}} c\mu(dc, \cdot) = \gamma \in \mathcal{M}(\widehat{D})$, the loss function (2.8) becomes

$$(2.13) \quad \begin{aligned} \mathcal{E}[\mu] &= C_f - \int_D F(\boldsymbol{\theta})\mu(d\boldsymbol{\theta}) + \frac{1}{2}\int_{D \times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\mu(d\boldsymbol{\theta})\mu(d\boldsymbol{\theta}') \\ &= \frac{1}{2}\mathbb{E}_\nu\left(f - \int_D \varphi(\cdot, \boldsymbol{\theta})\mu(d\boldsymbol{\theta})\right)^2 \geq 0 \end{aligned}$$

3 Training by Gradient Descent on the Exact Loss

Here we assume that we train the network by evolving dynamically the parameters $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$ according to the GD flow (1.7), which we recall is given by the coupled ODEs,

$$(3.1) \qquad \dot{\boldsymbol{\theta}}_i = \nabla F(\boldsymbol{\theta}_i) - \frac{1}{n}\sum_{j=1}^n \nabla K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j),$$

for $i = 1, \ldots, n$. As we show in Section 4, (3.1) shares many properties with the stochastic gradient descent (SGD) used in applications, though in SGD a multiplicative noise term persists in the equations. The ODEs in (3.1) are the GD flow on the energy:

$$(3.2) \qquad E(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = nC_f - \sum_{i=1}^n F(\boldsymbol{\theta}_i) + \frac{1}{2n}\sum_{i,j=1}^n K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j).$$

This energy is simply the loss function in (1.4) rescaled by n.

We consider (3.1) with initial conditions such that every $\boldsymbol{\theta}_i(0)$ for $i = 1, \ldots, n$ is drawn independently from some probability distribution μ_{in} satisfying the following:

ASSUMPTION 3.1. *The distribution μ_{in} is such that:* (i) *its support contains a smooth manifold that separates the regions in $D = \mathbb{R} \times \widehat{D}$ where $c > c_0$ and $c < -c_0$ for some large enough $c_0 > 0$;* (ii) *$\gamma_{in} = \int_{\mathbb{R}} c\mu_{in}(dc, \cdot)$ has finite total variation, $|\gamma_{in}|_{TV} < \infty$; and* (iii) *$\forall b \in \mathbb{R} : \int_{\mathbb{R} \times \widehat{D}} e^{bc}\mu_{in}(dc, dz) < \infty$.*

Note that property (i) guarantees that $\widehat{\mu}_{\text{in}} = \int_{\mathbb{R}} \mu_{\text{in}}(dc, \cdot)$ has full support in $\widehat{D}$, supp $\widehat{\mu}_{\text{in}} = \widehat{D}$—we show below that this property is preserved by the dynamics. Distributions μ_{in} that satisfy Assumption 3.1 include, e.g.,

$$\delta_0(dc)\,\widehat{\mu}_{\text{in}}(dz) \quad \text{and} \quad (2\pi)^{-1/2}e^{-\frac{1}{2}c^2}\,dc\,\widehat{\mu}_{\text{in}}(dz),$$

if supp $\widehat{\mu}_{\text{in}} = \widehat{D}$ in both. We denote the measure for the infinite set $\{\boldsymbol{\theta}_i(0)\}_{i \in \mathbb{N}}$ constructed this way by $\mathbb{P}_{\text{in}}$. Initial conditions of this type are used in practice.

3.1 Empirical distribution and nonlinear Liouville equation

To proceed, we consider the empirical distribution

$$(3.3) \qquad \mu_t^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i(t)}$$

in terms of which we can express (1.8) as

$$(3.4) \qquad f_t^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \varphi(\cdot, \boldsymbol{\theta}_i(t)) = \int_{D \times R} \varphi(\cdot, \boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta}).$$

The empirical distribution (3.3) is useful to work with because it satisfies a nonlinear Liouville-type equation

$$(3.5) \qquad \partial_t \mu_t^{(n)} = \nabla \cdot \left(\nabla V(\boldsymbol{\theta}, [\mu_t^{(n)}]) \mu_t^{(n)} \right),$$

where we defined

$$(3.6) \qquad V(\boldsymbol{\theta}, [\mu]) = -F(\boldsymbol{\theta}) + \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}') \mu(d\boldsymbol{\theta}').$$

Throughout, we will interpret (3.5) in the standard weak sense, as in (3.8) below. When there is a Laplacian term in (3.5) this equation is called the McKean-Vlasov equation [15, 18, 26, 34]; with an additional noise term added it is often referred to as Dean's equation [16]. To prove asymptotic trainability results, we analyze the properties of the solution to this equation as $n \to \infty$ and $t \to \infty$.

Derivation of (3.5).0 Let $\chi : D \to \mathbb{R}$ be a test function, and consider

$$(3.7) \qquad \int_D \chi(\boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \chi(\boldsymbol{\theta}_i(t)).$$

Taking the time derivative of this equation and using (3.1), we deduce

$$(3.8) \qquad \begin{aligned}
&\int_D \chi(\boldsymbol{\theta}) \partial_t \mu_t^{(n)}(d\boldsymbol{\theta}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot \dot{\boldsymbol{\theta}}_i(t) \\
&= \frac{1}{n} \sum_{i=1}^{n} \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot \left(\nabla F(\boldsymbol{\theta}_i(t)) - \frac{1}{n} \sum_{j=1}^{n} \nabla K(\boldsymbol{\theta}_i(t), \boldsymbol{\theta}_j(t)) \right) \\
&= \int_D \nabla \chi(\boldsymbol{\theta}) \cdot \left(\nabla F(\boldsymbol{\theta}) - \int_D \nabla K(\boldsymbol{\theta}, \boldsymbol{\theta}') \mu_t^{(n)}(d\boldsymbol{\theta}') \right) \mu_t^{(n)}(d\boldsymbol{\theta}).
\end{aligned}$$

This is the weak form of (3.5).

3.2 Law of large numbers (LLN)—mean field limit

Since we know that, $\mathbb{P}_{\text{in}}$-almost surely as $n \to \infty$, $\mu_0^{(n)} \rightharpoonup \mu_{\text{in}}$ by the law of large numbers, we can take the limit as $n \to \infty$ of (3.5) to formally deduce:

PROPOSITION 3.2. *Let $\mu_t^{(n)}$ be given by (3.3) with $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$ the solution of (3.1) with initial condition drawn from $\mathbb{P}_{\text{in}}$. Then, as $n \to \infty$, $\mu_t^{(n)} \rightharpoonup \mu_t$ a.s. where μ_t satisfies*

$$(3.9) \qquad \partial_t \mu_t = \nabla \cdot (\nabla V(\boldsymbol{\theta}, [\mu_t]) \mu_t) \quad \mu_0 = \mu_{in},$$

interpreted in the weak sense.

Note that (3.9) is the same as (3.5) but with a different initial condition. Note also that (3.9) is the GD flow in the Wasserstein metric [1,35]. Indeed, this equation can be written as the $\tau \to 0$ limit of the Jordan-Kinderlehrer-Otto (JKO) proximal scheme [20]

$$(3.10) \qquad \mu_{t+\tau} \in \operatorname{argmin}\left(\mathcal{E}[\mu] + \frac{1}{2}\tau^{-1} W_2^2(\mu, \mu_t)\right), \quad \mu_0 = \mu_{\text{in}},$$

where $W_2(\mu, \mu_t)$ is the 2-Wasserstein distance between μ and μ_t and $\mathcal{E}[\mu]$ is defined in (2.13). Finally, note that the weak solutions of (3.9) satisfy: for any test function $\chi : D \to \mathbb{R}$,

$$(3.11) \qquad \int_D \chi(\boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta}) = \int_D \chi(\boldsymbol{\Theta}_t(\boldsymbol{\theta})) \mu_{\text{in}}(d\boldsymbol{\theta}).$$

where $\boldsymbol{\Theta}_t(\boldsymbol{\theta})$ solves is given in terms of characteristics

$$(3.12) \qquad \dot{\boldsymbol{\Theta}}_t(\boldsymbol{\theta}) = -\nabla V(\boldsymbol{\Theta}_t(\boldsymbol{\theta}), [\mu_t]), \qquad \boldsymbol{\Theta}_0(\boldsymbol{\theta}) = \boldsymbol{\theta}.$$

Of course, (3.11) is not explicit since (3.12) depends on μ_t, but this representation formula is useful in the sequel. In particular, notice that it implies that: (i) $\mu_t \in \mathcal{M}_+(D)$ for all $t < \infty$ since the velocity field in (3.12) is globally Lipschitz on $\mathbb{R} \times \hat{D}$ by Assumption 2.1 and hence the solutions to this equation exist for all $t < \infty$; and (ii) $\operatorname{supp} \hat{\mu}_t = \hat{D}$ with $\hat{\mu}_t = \int_{\mathbb{R}} \mu_t(dc, \cdot)$ by Assumption 3.1, and $\operatorname{supp} \mu_t = D$ if $\operatorname{supp} \mu_{\text{in}} = D$.

The dynamics of $f_t = \lim_{n \to \infty} f_t^{(n)}$ We now discuss the implications of the limiting PDE for the evolution of

$$(3.13) \qquad \lim_{n \to \infty} f_t^{(n)} = \lim_{n \to \infty} \int_D \varphi(\cdot, \boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta}) = \int_D \varphi(\cdot, \boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta}) \equiv f_t$$

To begin, note that from (3.6) we can express $V(\boldsymbol{\theta}, [\mu_t])$ as

$$(3.14) \qquad V(\boldsymbol{\theta}, [\mu_t]) = \int_\Omega (f_t(\boldsymbol{x}) - f(\boldsymbol{x})) \, \varphi(\boldsymbol{x}, \boldsymbol{\theta}) \nu(d\boldsymbol{x}).$$

As a result (3.9) can be written as

$$(3.15) \qquad \partial_t \mu_t = \nabla \cdot \left(\int_\Omega \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{x}, \boldsymbol{\theta}) \, (f_t(\boldsymbol{x}) - f(\boldsymbol{x})) \, \nu(d\boldsymbol{x}) \mu_t\right)$$

and we deduce, using (3.13),

$$
\begin{aligned}
(3.16) \qquad \partial_t f_t &= \int_D \varphi(\cdot, \boldsymbol{\theta}) \partial_t \mu_t(d\boldsymbol{\theta}) \\
&= - \int_D \nabla_{\boldsymbol{\theta}} \varphi(\cdot, \boldsymbol{\theta}) \cdot \left(\int_\Omega \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{x}', \boldsymbol{\theta}) \left(f_t(\boldsymbol{x}') - f(\boldsymbol{x}') \right) \nu(d\boldsymbol{x}') \mu_t(d\boldsymbol{\theta}) \right).
\end{aligned}
$$

Interchanging the order of integration gives:

PROPOSITION 3.3 (LLN). *Let $f_t^{(n)}$ be given by (3.4) with $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$ solution of (3.1) with initial condition drawn from $\mathbb{P}_{\mathrm{in}}$. Then, as $n \to \infty$, $f_t^{(n)} \to f_t$ a.s. pointwise, where f_t satisfies*

$$
(3.17) \qquad \partial_t f_t(\boldsymbol{x}) = - \int_\Omega M([\mu_t], \boldsymbol{x}, \boldsymbol{x}')(f_t(\boldsymbol{x}') - f(\boldsymbol{x}'))\nu(d\boldsymbol{x}')
$$

where we defined the kernel

$$
\begin{aligned}
(3.18) \qquad & M([\mu], \boldsymbol{x}, \boldsymbol{x}') \\
&= \int_D \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{x}, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{x}', \boldsymbol{\theta}) \mu(d\boldsymbol{\theta}) \\
&= \int_{\mathbb{R} \times \widehat{D}} \left(c^2 \nabla_z \widehat{\varphi}(\boldsymbol{x}, z) \cdot \nabla_z \widehat{\varphi}(\boldsymbol{x}', z) + \widehat{\varphi}(\boldsymbol{x}, z)\widehat{\varphi}(\boldsymbol{x}', z) \right) \mu(dc, dz).
\end{aligned}
$$

The kernel (3.18) is symmetric in $\boldsymbol{x}$ and $\boldsymbol{x}'$ for any $\mu \in \mathcal{M}(D)$ and positive semi-definite if $\mu \in \mathcal{M}_+(D)$ since, given any $r \in L^2(\Omega, \nu)$, we then have

$$
\begin{aligned}
(3.19) \qquad & \int_{\Omega^2} r(\boldsymbol{x})r(\boldsymbol{x}')M([\mu], \boldsymbol{x}, \boldsymbol{x}')\nu(d\boldsymbol{x})\nu(d\boldsymbol{x}') \\
&= \int_{\mathbb{R} \times \widehat{D}} \left(c^2 |\nabla_z R(z)|^2 + |R(z)|^2 \right) \mu(dc, dz) \geq 0
\end{aligned}
$$

where

$$
(3.20) \qquad R(z) = \int_\Omega r(\boldsymbol{x})\widehat{\varphi}(\boldsymbol{x}, z)\nu(d\boldsymbol{x}).
$$

Equation (3.17) also confirms that f_t evolves on a quadratic landscape, namely the loss function (1.3) itself: Indeed, this equation can be written as

$$
(3.21) \qquad \partial_t f_t(\boldsymbol{x}) = - \int_\Omega M([\mu_t], \boldsymbol{x}, \boldsymbol{x}') \, D_{f_t(\boldsymbol{x}')} \mathcal{L}[f_t] \nu(d\boldsymbol{x}'),
$$

where $D_{f(\boldsymbol{x})}$ denotes the gradient with respect to $f(\boldsymbol{x})$ in the $L^2(\Omega, \nu)$-norm, i.e., given a functional $\mathcal{F}[f]$,

$$
\begin{aligned}
(3.22) \qquad \forall h : \Omega \to \mathbb{R} : \lim_{z \to 0} \frac{d}{dz}\mathcal{F}[f + zh] &= \langle h, D_f \mathcal{F}[f] \rangle_{L^2(\Omega, \nu)} \\
&= \int_\Omega h(\boldsymbol{x}) D_{f(\boldsymbol{x})}\mathcal{F}[f]\nu(d\boldsymbol{x}).
\end{aligned}
$$

That is, $D_{f(\boldsymbol{x})}$ reduces to $\delta/\delta f(\boldsymbol{x})$ if $\nu(d\boldsymbol{x}) = d\boldsymbol{x}$.

3.3 Long-time behavior—global convergence

Let us now analyze the long-time solutions of (3.9) for the weak limit μ_t of $\mu_t^{(n)}$ and (3.17) for the limit f_t of $f_t^{(n)}$. As is well-known, (3.9) has more stationary points than $\mathcal{E}[\mu]$ has minimizers. Since (3.9) is the Wasserstein GD flow on $\mathcal{E}[\mu]$, a direct calculation shows that $E_t = \mathcal{E}[\mu_t]$ satisfies

$$(3.23) \qquad \frac{dE_t}{dt} = -\int_D |\nabla V(\boldsymbol{\theta}, [\mu_t])|^2 \mu_t(d\boldsymbol{\theta}).$$

This equation implies that the stationary points μ^s of (3.9) are the solutions of

$$(3.24) \qquad \nabla V(\boldsymbol{\theta}, [\mu^s]) = 0 \quad \text{for } \boldsymbol{\theta} \in \operatorname{supp} \mu^s.$$

This should be contrasted with the minimizers of $\mathcal{E}[\mu]$, which satisfy:

$$(3.25) \qquad \begin{cases} V(\boldsymbol{\theta}, [\mu^*]) \geq \bar{V}[\mu^*] & \text{for } \boldsymbol{\theta} \in D, \\ V(\boldsymbol{\theta}, [\mu^*]) = \bar{V}[\mu^*] & \text{for } \boldsymbol{\theta} \in \operatorname{supp} \mu^*, \end{cases}$$

where $\bar{V}[\mu] = \int_D V(\boldsymbol{\theta}, [\mu])\mu(d\boldsymbol{\theta})$. In general, we cannot guarantee that the solutions to (3.24) also solve (3.25). However, due to the specific form of the unit, $\varphi(\boldsymbol{x}, \boldsymbol{\theta}) = c\widehat{\varphi}(\boldsymbol{x}, z)$, the rate of decay of the energy (3.23) reads

$$(3.26) \qquad \begin{aligned} \frac{dE}{dt} &= -\int_{\mathbb{R} \times \widehat{D}} \left(c^2 |\nabla \widehat{V}(z, [\mu_t])|^2 + |\widehat{V}(z, [\mu_t])|^2\right) \mu_t(dc, dz) \\ &= -\int_{\mathbb{R} \times \widehat{D}} c^2 |\nabla \widehat{V}(z, [\mu_t])|^2 \mu_t(dc, dz) - \int_{\widehat{D}} |\widehat{V}(z, [\mu_t])|^2 \widehat{\mu}_t(dz) \end{aligned}$$

where $\widehat{\mu}_t = \int_{\mathbb{R}} \mu_t(dc, \cdot)$ and

$$(3.27) \qquad \widehat{V}(z, [\mu]) = -\widehat{F}(z) + \int_{\mathbb{R} \times \widehat{D}} c' \widehat{K}(z, z') \mu(dc', dz')$$

(3.26) implies that the stationary points μ^s of (3.9) satisfy

$$(3.28) \qquad \widehat{V}(z, [\mu^s]) = 0 \quad \text{for } z \in \operatorname{supp} \widehat{\mu}^s = \int_{\mathbb{R}} \mu^s(dc, \cdot).$$

As a result, $V(\boldsymbol{\theta}, [\mu^s]) = c\widehat{V}(z, [\mu^s]) = 0$ for $\boldsymbol{\theta} = (c, z) \in \operatorname{supp} \mu^s$, and this shows that the second equation in (3.25) is automatically satisfied, noting that $\bar{V} = 0$ for a global minimizer.

To show that first equation of (3.28) also holds, we establish that $\widehat{V}(z, [\mu^s]) = 0$ everywhere in $\widehat{D}$. We proceed by contradiction: Suppose that μ_t converges to some μ^s such that $\widehat{V}(z, [\mu^s]) \neq 0$ for $z = \widehat{D}_s^c$ where $\widehat{D}_s^c$ is the complement in $\widehat{D}$ of $\widehat{D}_s = \operatorname{supp} \widehat{\mu}^s$—the relevant case is when $\widehat{D}_s^c$ has nonzero Hausdorff measure in $\widehat{D}$. Looking at the characteristic equations (3.12) written in terms of $\boldsymbol{\Theta}_t = (C_t, \boldsymbol{Z}_t)$ as

$$(3.29) \qquad \begin{cases} \dot{C}_t(c, z) = -\widehat{V}(\boldsymbol{Z}_t(c, z), [\mu_t]), & C_0(c, z) = c, \\ \dot{\boldsymbol{Z}}_t(c, z) = -C_t(c, z)\nabla \widehat{V}(\boldsymbol{Z}_t(c, z), [\mu_t]), & \boldsymbol{Z}_0(c, z) = z. \end{cases}$$

Since we know that supp $\widehat{\mu}_t = \widehat{D}$ at all positive time $t < \infty$, in order that $\mu_t \to \mu^s$ as $t \to \infty$, all the mass must be expelled from $\widehat{D}_s^c$. That is, all the solutions to (3.29) must leave this domain, or at least accumulate at its boundary, while at the same time we must have $\lim_{t\to\infty} \widehat{V}(z, [\mu_t]) \neq 0$. To show that this scenario is impossible, note that (using the fact that $\widehat{D}$ is compact)

$$(3.30) \qquad \forall \delta > 0, \ \exists t_c > 0 : \sup_z |\widehat{V}(z, [\mu_t]) - \widehat{V}(z, [\mu_s])| \leq \delta \quad \text{if } t \geq t_c.$$

This means that, for $t \geq t_c$, to leading order in δ, (3.29) reads

$$(3.31) \qquad \begin{cases} \dot{C}_t(c, z) = -\widehat{V}(\boldsymbol{Z}_t(c, z), [\mu^s]), & C_0(c, z) = c, \\ \dot{\boldsymbol{Z}}_t(c, z) = -C_t(c, z)\nabla\widehat{V}(\boldsymbol{Z}_t(c, z), [\mu^s]), & \boldsymbol{Z}_0(c, z) = z \end{cases}$$

which is the GD flow on

$$(3.32) \qquad\qquad\qquad\qquad c\widehat{V}(z, [\mu^s]).$$

Suppose that $\widehat{V}(z, [\mu^s]) > 0$ somewhere in $\widehat{D}_s^c$—the case when $\widehat{V}(z, [\mu^s]) < 0$ somewhere in $\widehat{D}_s^c$ can be treated similarly. Since $\widehat{D}_s^c$ is compact, $\widehat{V}(z, [\mu^s])$ must then have a maximum in $\widehat{D}_s^c$, i.e., using the differentiability of the unit, there exists $z_1 \in \widehat{D}_s^c$ with $z_1 \notin \partial\widehat{D}_s^c$ and such that $\nabla\widehat{V}(z_1, [\mu^s]) = 0$, $\widehat{V}(z_1, [\mu^s]) = \widehat{V}_1 > 0$, and $\widehat{V}(z_1, [\mu^s]) > \widehat{V}(z, [\mu^s])$ for $z \in \widehat{D}_s^c$. Consider the solutions to (3.31) for initial (c, z) such that $\boldsymbol{Z}_t(c, z)$ is very close to z_1 at $t = t_c$—these solutions must exist since supp $\mu_t = \widehat{D}$ for all $t < \infty$. If among these solutions there are some such that $C_t(c, z)$ is negative at time $t = t_c$ (which is always the case if supp $\mu_{\text{in}} = D$ since supp $\mu_t = D$ for all $t < \infty$ in that case), then by (3.31) $C_t(c, z)$ becomes more negative and $\boldsymbol{Z}_t(c, z)$ gets closer to z_1 for $t > t_c$. If all $C_{t+\tau}(c, z)$ are positive at time $t = t_c(\delta)$, then the corresponding $\boldsymbol{Z}_t(c, z)$ go away from z_1 for as long as their $C_t(c, z)$ remains positive; however, eventually some $C_t(c, z)$ become negative (since $C_{t+\tau}(c, z_1) = C_t(c, z_1) - \tau\widehat{V}(z_1, [\mu^s]) = C_t(c, z_1) - \tau\widehat{V}_1$ under (3.31)), at which point we go back to the first case and $\boldsymbol{Z}_t(c, z)$ gets closer to z_1. Either way, we can always find solutions with $\boldsymbol{Z}_t(c, z)$ sufficiently close to z_1 at time $t_c(\delta)$ that will eventually converge to z_1 rather than exiting $\widehat{D}_s^c$, a contradiction with our assumption that all solutions must either exit this domain or accumulate at its boundary. This argument is based on (3.31) rather than the original (3.29), but by setting δ small enough (and t_c large enough) we can make the terms left over in (3.31) arbitrarily small so that they do not affect the result.

This concludes the justification that the stationary points μ^s of (3.9) are such that $\widehat{V}(z, [\mu^s]) = 0$ everywhere in $\widehat{D}$; i.e., they are minimizers of $\mathcal{E}[\mu]$, which from (3.28) implies

$$(3.33) \qquad \forall z \in \widehat{D} : 0 = \int_\Omega \widehat{\varphi}(\boldsymbol{x}, z) \left(f(\boldsymbol{x}) - \int_{\widehat{D}} \widehat{\varphi}(\boldsymbol{x}, z')\gamma^s(dz') \right) \nu(d\boldsymbol{x})$$

where $\gamma^s = \int_{\mathbb{R}} c\mu^s(dc, \cdot)$. As a result, by Assumptions 2.2 and 2.4,

$$(3.34) \qquad f = \int_{\widehat{D}} \widehat{\varphi}(\cdot, z)\gamma^s(dz),$$

In other other words, we have established:

PROPOSITION 3.4 (Global convergence). *Let μ_t be the solution to (3.9) for the initial condition $\mu_0 = \mu_{in}$ that satisfies Assumption 3.1. If $\mu_t \to \mu^* \in \mathcal{M}_+(D)$ as $t \to \infty$, then under Assumptions 2.2 and 2.4 μ^* is a minimizer of $\mathcal{E}[\mu]$ and we have*

$$(3.35) \qquad \lim_{t \to \infty} \int_D \varphi(\cdot, \boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}) = \int_D \varphi(\cdot, \boldsymbol{\theta})\mu^*(d\boldsymbol{\theta}) = f.$$

Note that we assume that μ_t converges to some probability measure to establish this proposition. This is because we cannot exclude a priori that the dynamics eventually loses mass at infinity, e.g., if some of the solutions of the characteristic equation (3.12) eventually diverge as $t \to \infty$. We do not expect this scenario to occur for most initial conditions, and one way to preclude it altogether is to add regularizing terms in the loss function.

The argument that leads to (3.35) would be simple if it were the case that $\widehat{\mu}^* = \int_{\mathbb{R}} \mu^*(dc, \cdot)$ has full support in $\widehat{D}$. Indeed, this would imply that the kernel (3.18) evaluated on μ_t is positive definite not only for all $t \geq 0$ but also in the limit as $t \to \infty$, and hence the only fixed point of (3.17) is f. It is reasonable to assume that $\operatorname{supp} \widehat{\mu}^* = \widehat{D}$ because: (i) $\operatorname{supp} \widehat{\mu}_t = \widehat{D}$ for all $t < \infty$ as mentioned before and (ii) there is no energetic incentive to shrink the support, even when $t \to \infty$. To see why, note that if μ^* is an energy minimizer such that $\operatorname{supp} \widehat{\mu}^* \neq \widehat{D}$, then a direct calculation shows that for any $\alpha \in (0, 1)$ and any $\widehat{\mu} \in \mathcal{M}_+(\widehat{D})$ with $\operatorname{supp} \widehat{\mu} = \widehat{D}$,

$$(3.36) \qquad \mu^{**}(dc, dz) = (1-\alpha)^2 \mu^*((1-\alpha)dc, dz) + \alpha\delta_0(dc)\widehat{\mu}(dz)$$

is also a energy minimizer in $\mathcal{M}_+(D)$ such that $\widehat{\mu}^{**} = \int_{\mathbb{R}} \mu^{**}(dc, \cdot)$ has support $\widehat{D}$.

In the Appendix, we analyze the behavior of μ_t on a longer timescale and show that, with noise and certain regularizing terms added in (3.1), μ_t reaches a unique fixed point $\mu^* \in \mathcal{M}_+(D)$ such that $\int_D \log(d\mu^*/d\mu^0)d\mu^* < \infty$, where μ^0 is some prior measure used for regularization.

We can summarize the results of Sections 3.2 and 3.3 into:

PROPOSITION 3.5 (LLN and global convergence). *Let $f_t^{(n)}$ be given by (3.4) with $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$ the solution of (3.1) with initial condition drawn from $\mathbb{P}_{\text{in}}$. Then under the conditions of Proposition 3.4 we have*

$$(3.37) \qquad \lim_{n \to \infty} f_t^{(n)} = f_t \quad \text{pointwise, } \mathbb{P}_{\text{in}}\text{-almost surely}$$

where f_t solves (3.17) *and satisfies*

$$(3.38) \qquad \lim_{t \to \infty} f_t = f \quad a.e.\ in\ \Omega.$$

The convergence in (3.38) is equivalent to the statement in Proposition 3.4. Notice that, since the evolution of f_t occurs via (3.17), which is independent of n, for any $\delta > 0$ we can find t_c independent of n such that for $t > t_c$, $\mathbb{E}_\nu |f_t - f|^2 < \delta$. Since for any $\delta > 0$ and $t > 0$ we can also find n_c such that for $n > n_c$,

$$\mathbb{E}_\nu \big| f_t^{(n)} - f_t \big|^2 < \delta,$$

we can interchange the limits in n and t in Theorem 3.5; i.e., we also have

$$(3.39) \qquad \lim_{n \to \infty} \lim_{t \to \infty} f_t^{(n)} = f.$$

3.4 Central limit theorem (CLT)

Let us now consider the fluctuations of $\mu_t^{(n)}$ around its limit μ_t. To this end, we define $\omega_t^{(n)}$ via

$$(3.40) \qquad \omega_t^{(n)} = n^{1/2} \big(\mu_t^{(n)} - \mu_t \big),$$

Explicitly, (3.40) means

$$(3.41) \qquad \omega_t^{(n)} = n^{-1/2} \sum_{i=1}^{n} (\delta_{\boldsymbol{\theta}_i(t)} - \mu_t).$$

The scaling factor $n^{1/2}$ is set by the fluctuations in the initial conditions: if we pick a test function $\chi : D \to \mathbb{R}$, the CLT tells us that under $\mathbb{P}_{\text{in}}$

$$(3.42) \quad \int_D \chi(\boldsymbol{\theta}) \omega_0^{(n)}(d\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^{n} \widetilde{\chi}(\boldsymbol{\theta}_i(0)) \to N(0, C_\chi) \quad \text{in law as } n \to \infty$$

where $\widetilde{\chi}(\boldsymbol{\theta}) = \chi(\boldsymbol{\theta}) - \int_D \chi(\boldsymbol{\theta}) \mu_{\text{in}}(d\boldsymbol{\theta})$ and $N(0, C_\chi)$ denotes the Gaussian random variable with mean zero and variance

$$(3.43) \qquad C_\chi = \int_D |\widetilde{\chi}(\boldsymbol{\theta})|^2 \mu_{\text{in}}(d\boldsymbol{\theta}),$$

To see what happens at later times, we derive an equation for $\omega_t^{(n)}$ by subtracting (3.9) from (3.5) and using (3.40)

$$(3.44) \quad \partial_t \omega_t^{(n)} = \nabla \cdot \big(\omega_t^{(n)} \nabla V(\boldsymbol{\theta}, [\mu_t]) + \big(\mu_t + n^{-1/2} \omega_t^{(n)} \big) \nabla F \big(\boldsymbol{\theta}, \big[\omega_t^{(n)} \big] \big) \big)$$

where we defined

$$(3.45) \qquad F(\boldsymbol{\theta}, [\mu]) = \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}') \mu(d\boldsymbol{\theta}').$$

If we take the limit as $n \to \infty$, the term involving $n^{-1/2} \omega_t^{(n)}$ at the right-hand side of (3.44) disappears (we quantify its rate of convergence to zero in more detail in Section 3.5) and we formally deduce the following:

PROPOSITION 3.6. *Let $\omega_t^{(n)}$ be given by* (3.41) *with* $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$ *the solution of* (3.1) *with initial conditions drawn from* $\mathbb{P}_{\mathrm{in}}$ *and* μ_t *the solution to* (3.15). *Then*

$$\omega_t^{(n)} \rightharpoonup \omega_t \quad \textit{in law as } n \to \infty \tag{3.46}$$

where ω_t satisfies

$$\partial_t \omega_t = \nabla \cdot \left(\omega_t \nabla V(\boldsymbol{\theta}, [\mu_t]) + \mu_t \nabla F(\boldsymbol{\theta}, [\omega_t]) \right) \tag{3.47}$$

to be solved in the weak sense with the Gaussian initial conditions read from (3.42):

$$\int_D \chi(\boldsymbol{\theta}) \omega_0(d\boldsymbol{\theta}) = N(0, C_\chi). \tag{3.48}$$

Note that since the mean of ω_0 is zero initially and (3.47) is linear, this mean remains zero for all times, and we can focus on the evolution of its covariance:

$$\Sigma_t(d\boldsymbol{\theta}, d\boldsymbol{\theta}') = \mathbb{E}_{\mathrm{in}}[\omega_t(d\boldsymbol{\theta})\omega_t(d\boldsymbol{\theta}')]. \tag{3.49}$$

From (3.47) it satisfies

$$\begin{aligned} \partial_t \Sigma_t &= \nabla \cdot \left(\Sigma_t \nabla V(\boldsymbol{\theta}, [\mu_t]) + \mu_t(d\boldsymbol{\theta}) \nabla G(\boldsymbol{\theta}, d\boldsymbol{\theta}', [\Sigma_t]) \right) \\ &\quad + \nabla' \cdot \left(\Sigma_t \nabla V(\boldsymbol{\theta}', [\mu_t]) + \mu_t(d\boldsymbol{\theta}') \nabla G(\boldsymbol{\theta}', d\boldsymbol{\theta}, [\Sigma_t]) \right) \end{aligned} \tag{3.50}$$

where we defined

$$G(\boldsymbol{\theta}, \cdot, [\Sigma])) = \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}'') \Sigma(d\boldsymbol{\theta}'', \cdot) \tag{3.51}$$

Equation (3.50) should be interpreted in the weak sense and solved for the initial condition

$$\Sigma_0(d\boldsymbol{\theta}, d\boldsymbol{\theta}') = \mu_{\mathrm{in}}(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}'). \tag{3.52}$$

The dynamics of $g_t = \lim_{n\to\infty} n^{1/2}(f_t^{(n)} - f_t)$. We can also test these equations against the unit, to deduce that, as $n \to \infty$,

$$\begin{aligned} g_t^{(n)} &= \int_D \varphi(\cdot, \boldsymbol{\theta}) \omega_t^{(n)}(d\boldsymbol{\theta}) = n^{1/2} \left(f_t^{(n)} - f_t \right) \\ &= n^{-1/2} \sum_{i=1}^n \left(\varphi(\cdot, \boldsymbol{\theta}_i(t)) - f_t \right) \end{aligned} \tag{3.53}$$

converges in law, $g_t^{(n)} \to g_t$, where g_t is a Gaussian process satisfying

$$\begin{aligned} \partial_t g_t &= -\int_\Omega M(\boldsymbol{x}, \boldsymbol{x}', [\omega_t]) \left(f_t(\boldsymbol{x}') - f(\boldsymbol{x}') \right) \nu(d\boldsymbol{x}') \\ &\quad - \int_\Omega M(\boldsymbol{x}, \boldsymbol{x}', [\mu_t]) g_t(\boldsymbol{x}') \nu(d\boldsymbol{x}'). \end{aligned} \tag{3.54}$$

This equation should be solved with Gaussian initial conditions with mean zero and covariance

$$
\begin{aligned}
C_0(x, x') &= \mathbb{E}_{\text{in}}[g_0(x)g_0(x')] \\
&= \int_D \varphi(x, \theta)\varphi(x', \theta)\mu_{\text{in}}(d\theta) \\
&\quad - \int_{D \times D} \varphi(x, \theta)\varphi(x', \theta')\mu_{\text{in}}(d\theta)\mu_{\text{in}}(d\theta')
\end{aligned}
$$
(3.55)

Since (3.54) is linear the mean of g_t remains zero at all times and we can again focus on the evolution of its covariance:

$$
C_t(x, x') = \mathbb{E}_{\text{in}}[g_t(x)g_t(x')]
$$
(3.56)

We obtain

$$
\begin{aligned}
\partial_t C_t &= -\int_\Omega N(x, x', x'', [\Sigma_t])\left(f_t(x'') - f(x'')\right)\nu(dx'') \\
&\quad - \int_\Omega M(x, x'', [\mu_t])C_t(x'', x)\nu(dx'') \\
&\quad - \int_\Omega M(x', x'', [\mu_t])C_t(x'', x')\nu(dx'')
\end{aligned}
$$
(3.57)

where Σ_t solves (3.50) and

$$
\begin{aligned}
&N(x, x'x'', [\Sigma]) \\
&\quad = \int_{D \times D} \nabla_\theta \varphi(x, \theta) \cdot \nabla_\theta \varphi(x'', \theta)\varphi(x', \theta')\Sigma(d\theta, d\theta') \\
&\quad + \int_{D \times D} \nabla_\theta \varphi(x', \theta) \cdot \nabla_\theta \varphi(x'', \theta)\varphi(x, \theta')\Sigma(d\theta, d\theta')
\end{aligned}
$$
(3.58)

Summarizing, we have established:

PROPOSITION 3.7 (CLT). *Let $g_t^{(n)}$ be given by (3.53) with $\{\theta_i(t)\}_{i=1}^n$ the solution of (3.1) with initial conditions drawn from $\mathbb{P}_{\text{in}}$ and μ_t the solution to (3.15). Then, as $n \to \infty$, $g_t^{(n)} \to g_t$ in law, where g_t is the zero mean Gaussian process whose covariance solves to (3.57) for the initial condition (3.55).*

3.5 Scaling of the fluctuations at long and very long times

To analyze the behavior of the fluctuations as $t \to \infty$, we revisit the results from the last section from a different perspective. Suppose that, instead of (3.41) and (3.53), we consider

$$
\bar{\omega}_t^{(n)} = n^{-1/2} \sum_{i=1}^n \left(\delta_{\Theta_i(t)} - \mu_t\right)
$$
(3.59)

and

$$(3.60) \qquad \overline{g}_t^{(n)} = n^{-1/2} \sum_{i=1}^{n} \left(\varphi(\cdot, \boldsymbol{\Theta}_i(t)) - \int_D \varphi(\cdot, \theta) \mu_t(d\theta) \right)$$

where $\boldsymbol{\Theta}_i(t)$ are independent copies of the mean-field characteristic equation (3.12). Then, direct calculations show that $\overline{\omega}_t^{(n)} \rightharpoonup \overline{\omega}_t$ and $\overline{g}_t^{(n)} \to \overline{g}_t$ in law as $n \to \infty$, where $\overline{\omega}_t$ and $\overline{g}_t$ are Gaussian processes with mean zero and covariance given explicitly by

$$(3.61) \qquad \begin{aligned} \overline{\Sigma}_t(d\boldsymbol{\theta}, d\boldsymbol{\theta}') &= \mathbb{E}_{\mathrm{in}}[\overline{\omega}_t(d\boldsymbol{\theta})\overline{\omega}_t(d\boldsymbol{\theta}')] \\ &= \mu_t(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}') - \mu_t(d\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}') \end{aligned}$$

and

$$(3.62) \qquad \begin{aligned} \overline{C}_t(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}_{\mathrm{in}}[\overline{g}_t(\boldsymbol{x})\overline{g}_t(\boldsymbol{x}')] \\ &= \int_D \varphi(\boldsymbol{x}, \boldsymbol{\theta})\varphi(\boldsymbol{x}', \boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}) - f_t(\boldsymbol{x})f_t(\boldsymbol{x}'). \end{aligned}$$

We can also easily write down evolution equations for $\overline{\omega}_t$ and $\overline{g}_t$: they read,

$$(3.63) \qquad \partial_t \overline{\omega}_t = \nabla \cdot (\overline{\omega}_t \nabla V(\boldsymbol{\theta}, [\mu_t]))$$

and

$$(3.64) \qquad \partial_t \overline{g}_t = -\int_\Omega M(\boldsymbol{x}, \boldsymbol{x}', [\overline{\omega}_t]) \left(f_t(\boldsymbol{x}') - f(\boldsymbol{x}') \right) \nu(d\boldsymbol{x}').$$

Let us focus on this last equation: it is similar to (3.54) but without the last term, $-\int_\Omega M(\boldsymbol{x}, \boldsymbol{x}', [\mu_t]) g_t(\boldsymbol{x}') \nu(d\boldsymbol{x}')$. Since the kernel M is positive semidefinite, we know that the solutions to (3.54) are controlled by those of (3.64). In particular,

$$(3.65) \qquad \mathbb{E}_{\mathrm{in}} \int_\Omega |g_t(\boldsymbol{x})|^2 \nu(d\boldsymbol{x}) = \int_\Omega C_t(\boldsymbol{x}, \boldsymbol{x}) \nu(d\boldsymbol{x}) \leq \int_\Omega \overline{C}_t(\boldsymbol{x}, \boldsymbol{x}) \nu(d\boldsymbol{x}).$$

If we assume that $\mu_t \to \mu^* \in \mathcal{M}_+(D)$ as $t \to \infty$, from (3.62) we have

$$(3.66) \qquad \lim_{t \to \infty} \int_\Omega \overline{C}_t(\boldsymbol{x}, \boldsymbol{x}) \nu(d\boldsymbol{x}) = \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta})\mu^*(d\boldsymbol{\theta}) - \int_\Omega |f(\boldsymbol{x})|^2 \nu(d\boldsymbol{x})$$

and therefore

$$(3.67) \qquad \lim_{t \to \infty} \int_\Omega C_t(\boldsymbol{x}, \boldsymbol{x}) \nu(d\boldsymbol{x}) \leq \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta})\mu^*(d\boldsymbol{\theta}) - \int_\Omega |f(\boldsymbol{x})|^2 \nu(\boldsymbol{x}).$$

Because

$$(3.68) \qquad \int_\Omega C_t(\boldsymbol{x}, \boldsymbol{x}) \nu(d\boldsymbol{x}) = \lim_{n \to \infty} n \mathbb{E}_{\mathrm{in}} \int_\Omega \left| f_t^{(n)}(\boldsymbol{x}) - f_t(\boldsymbol{x}) \right|^2 \nu(d\boldsymbol{x}),$$

the previous result gives a Monte-Carlo-type error bound on the loss. Note that this bound is only nontrivial if

$$
\int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}) \mu^*(d\boldsymbol{\theta}) = \int_{\mathbb{R} \times \widehat{D}} c^2 \widehat{K}(z, z) \mu^*(dc, dz)
$$

(3.69)

$$
\leq \|\widehat{K}\|_\infty \int_{\mathbb{R} \times \widehat{D}} c^2 \mu^*(dc, dz) < \infty.
$$

During training, we have $\int_{\mathbb{R} \times \widehat{D}} c^2 \mu_t(dc, dz) < \infty$ for all $t < \infty$, and to guarantee that this moment does not blow up as $t \to \infty$, or more generally to control its value in that limit, we may need to add a regularizing term to the loss function. If (3.69) holds, there is a situation in which we can even deduce a better bound: if $\operatorname{supp} \widehat{\mu}^* = \widehat{D}$, then $M(\boldsymbol{x}, \boldsymbol{x}, [\mu_t])$ is positive definite for all $t \geq 0$ and in the limit $t \to \infty$, indicating that the last term in (3.54) is always dissipative. In this case the argument above shows that

(3.70)
$$
\lim_{t \to \infty} \int_\Omega C_t(\boldsymbol{x}, \boldsymbol{x}) \nu(d\boldsymbol{x}) = 0.
$$

Summarizing:

PROPOSITION 3.8 (Fluctuations at long times). *Let $f_t^{(n)}$ be given with $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$ the solution of* (3.4) *with initial conditions drawn from $\mathbb{P}_{\mathrm{in}}$ and f_t the solution to* (3.17). *Then, under the conditions of Proposition* 3.4 *and assuming that* (3.69) *holds, we have*

(3.71)
$$
\lim_{t \to \infty} \lim_{n \to \infty} n \mathbb{E}_{\mathrm{in}} \int_\Omega \left| f_t^{(n)}(\boldsymbol{x}) - f_t(\boldsymbol{x}) \right|^2 \nu(d\boldsymbol{x})
$$
$$
\leq \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}) \mu^*(d\boldsymbol{\theta}) - \int_\Omega |f(\boldsymbol{x})|^2 \nu(d\boldsymbol{x}).
$$

In addition, if $\operatorname{supp} \widehat{\mu}^* = \widehat{D}$, *we have*

(3.72)
$$
\lim_{t \to \infty} \lim_{n \to \infty} n \mathbb{E}_{in} \int_\Omega |f_t^{(n)}(\boldsymbol{x}) - f_t(\boldsymbol{x})|^2 \nu(d\boldsymbol{x}) = 0.
$$

In situations where $\operatorname{supp} \widehat{\mu}^* = \widehat{D}$ and (3.72) holds, we see that the fluctuations, initially detectable on the scale $n^{-1/2}$, become higher order as time increases. To understand the scale at which the fluctuations eventually settle, consider

(3.73)
$$
\widetilde{\omega}_t^{(n)}(d\boldsymbol{\theta}) = n^{\xi(t)} \sum_{i=1}^n \left(\delta_{\boldsymbol{\theta}_i(t)}(\boldsymbol{\theta}) - \mu_t(d\boldsymbol{\theta}) \right)
$$

where $\xi(t)$ is some time-dependent exponent to be specified. By proceeding as we did to derive (3.44), we have that $\widetilde{\omega}_t^{(n)}$ satisfies

(3.74)
$$
\partial_t \widetilde{\omega}_t^{(n)} = \nabla \cdot \left(\widetilde{\omega}_t^{(n)} \nabla V(\boldsymbol{\theta}, [\mu_t]) + \mu_t \nabla F(\boldsymbol{\theta}, [\widetilde{\omega}_t^{(n)}]) \right)
$$
$$
+ n^{-\xi(t)} \nabla \cdot \left(\omega_t^{(n)} \nabla F(\boldsymbol{\theta}, [\widetilde{\omega}_t^{(n)}]) \right) + \dot{\xi}(t) \log n \, \widetilde{\omega}_t^{(n)}.
$$

In order to take the limit as $n \to \infty$ of this equation, we need to consider carefully the behavior of the factors in (3.74) that contain n explicitly, i.e., $\omega_t^{(n)} \nabla F(\boldsymbol{\theta}, [\widetilde{\omega}_t^{(n)}])$ and $\dot{\xi}(t) \log n \, \widetilde{\omega}_t^{(n)}$. Regarding the former, for any $p \in \mathbb{N}$ and $\xi \in \mathbb{R}$,

$$(3.75) \qquad \mathbb{E}_{\mathrm{in}} \left(n^{-\xi} \int_{D \times D} \chi(\boldsymbol{\theta}) \chi(\boldsymbol{\theta}') \widetilde{\omega}_0^{(n)}(d\boldsymbol{\theta}) \widetilde{\omega}_0^{(n)}(d\boldsymbol{\theta}') \right)^p = O(n^{(\xi-1)p}),$$

which can be verified by a direct calculation. For example if $p = 1$, this expectation is $n^{(\xi-1)} C_\chi$ where C_χ is given in (3.43). Equation (3.75) implies that $n^{-\xi} \widetilde{\omega}_0^{(n)}(d\boldsymbol{\theta}) \widetilde{\omega}_0^{(n)}(d\boldsymbol{\theta}) \rightharpoonup 0$ weakly in L^{2p} at $t = 0$ for any $\xi < 1$. To see whether we can bring the fluctuations to that scale, notice that if we set

$$(3.76) \qquad \dot{\xi}(t) \log n = o(1)$$

the last term on the right-hand side of (3.74) is also higher order—(3.76) means that we can vary $\xi(t)$, but only slowly. (3.76) can be achieved by choosing, e.g.,

$$(3.77) \qquad \xi(t) = \overline{\xi}(t/a_n)$$

with $\overline{\xi}(0) = \frac{1}{2}$, $\overline{\xi}'(u) > 0$, $\lim_{u \to \infty} \overline{\xi}(u) =< 1$, and a_n growing with n and such that $\lim_{n \to \infty} a_n / \log n = \infty$. With this choice, both the last two terms on the right-hand side of (3.74) are a small perturbation that vanishes as $n \to \infty$. Therefore, if we test $\widetilde{\omega}_t^{(n)}$ against the uni, and define

$$(3.78) \qquad \widetilde{g}_t^{(n)} = n^{\xi(t)} \left(f_t^{(n)} - f_t \right) = \int_D \varphi(\cdot, \boldsymbol{\theta}) \widetilde{\omega}_t^{(n)}(d\boldsymbol{\theta})$$

we know that, if $\xi(t)$ is as in (3.77) and $\operatorname{supp} \widehat{\mu}^* = \widehat{D}$, this field will be controlled and go to zero eventually. Summarizing we have established:

PROPOSITION 3.9 (Fluctuations at very long times). *Assume that the conditions of Proposition 3.8 hold and* $\operatorname{supp} \widehat{\mu}^* = \widehat{D}$. *Then*

$$(3.79) \qquad \forall \xi < 1 : \lim_{n \to \infty} n^{2\xi} \mathbb{E}_{in} \int_\Omega \left| f_{a_n}^{(n)}(\boldsymbol{x}) - f(\boldsymbol{x}) \right|^2 \nu(d\boldsymbol{x}) = 0$$

if a_n grows with n and is such that $\lim_{n \to \infty} a_n / \log n = \infty$.

This proposition can be stated as (1.11). It shows a remarkable self-healing property of the dynamics: the fluctuations at scale $O(n^{-1/2})$ of $f_t^{(n)}$ around f_t that were present initially decrease in amplitude as time progresses, and become $O(n^{-1})$ or smaller as $t \to \infty$.

4 Training by Online Stochastic Gradient Descent

In most applications, it is not possible to evaluate the expectation over the data in (1.6) defining $\widehat{F}(z)$ and $\widehat{K}(z, z')$. This is especially true for $\widehat{F}(z)$, since we typically have limited access to $f(\boldsymbol{x})$: often we can only evaluate it pointwise or only know its value on a discrete set of points. In these cases, we typically need to approximate the expectation in (1.6) by sampling a finite data set from the measure

ν by sampling a finite data set from the measure ν obtained by sampling from the measure ν.

If we fix this training data set, $\{x_p\}_{p=1}^P$, and denote by

$$\nu_P = P^{-1} \sum_{p=1}^P \delta_{x_p}$$

the corresponding empirical measure, then all the results in Section 3 apply at empirical level if we replace everywhere ν by ν_P. This, however, is not the question we are typically interested in, which is rather:

> *How does the test error (that is, the error obtained using the exact loss defined with the original ν) scale if we train the network on the empirical loss associated to ν_P?*

Here we will address this question in the specific setting of "online" learning algorithms, in which we can draw a training data set of batch size P at every step of the learning. This effectively assumes that we have access to infinite data, but cannot use it all at the same time, and the finite size of the batch introduces noise into the learning algorithm. The algorithm in which the gradient is estimated from a subset of training data at each step is known as stochastic gradient descent. It reads

$$(4.1)\ \ \widehat{\theta}_i(t + \Delta t) = \widehat{\theta}_i(t) + \nabla F_P(t, \widehat{\theta}_i^P(t))\Delta t - \frac{1}{n}\sum_{j=1}^n \nabla K_P(t, \widehat{\theta}_i(t), \widehat{\theta}_j(t))\Delta t$$

where $i = 1, \ldots, n$, $\Delta t > 0$ is some time step, and we defined

$$F_P(t, \theta) = \frac{1}{P}\sum_{p=1}^P f(x_p(t))\varphi(x_p(t), \theta),$$

(4.2)

$$K_P(t, \theta, \theta') = \frac{1}{P}\sum_{p=1}^P \varphi(x_p(t), \theta)\varphi(x_p(t), \theta'),$$

in which $\{x_p(t)\}_{p=1}^P$ are P i.i.d. variables that are redrawn from ν independently at every time step t. Next we analyze how the result from Section 3 must be modified when we use (4.2) rather than (3.1) to perform the training.

4.1 Limiting stochastic differential equation

To analyze the properties of (4.1), we start by noticing that the term

$$(4.3)\qquad R_i(\vec{\theta}) = \nabla F_P(t, \theta_i) - \frac{1}{n}\sum_{j=1}^n \nabla K_P(t, \theta_i, \theta_j), \quad \vec{\theta} = (\theta_1, \ldots, \theta_n),$$

is an unbiased estimator of the right-hand side of the GD equation (3.1). Indeed, conditional on $\{\boldsymbol{\theta}_i\}_{i=1}^{n}$ fixed, we have

$$(4.4) \qquad \mathbb{E}_\nu \boldsymbol{R}_i(\vec{\boldsymbol{\theta}}) = \nabla F(\boldsymbol{\theta}_i) - \frac{1}{n} \sum_{j=1}^{n} \nabla K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j).$$

This means that, if we split the right-hand side of (4.1) into its expectation plus a zero-mean fluctuation, the expression resembles an Euler-Maruyama scheme for a stochastic differential equation (SDE), except that the scaling of the noise term involves Δt rather than $\sqrt{\Delta t}$. To write this SDE explicitly, we compute the covariance of $\boldsymbol{R}(\vec{\boldsymbol{\theta}})$ conditional on $\{\boldsymbol{\theta}_i\}_{i=1}^{n}$ fixed,

$$(4.5) \qquad \mathrm{cov}_\nu \big[\boldsymbol{R}_i(\vec{\boldsymbol{\theta}}), \boldsymbol{R}_j(\vec{\boldsymbol{\theta}}') \big] = A\big([f - f^{(n)}], \boldsymbol{\theta}_i, \boldsymbol{\theta}_j' \big)$$

where $f^{(n)} = n^{-1} \sum_{i=1}^{n} \varphi(\cdot, \boldsymbol{\theta}_i)$, and we defined

$$(4.6) \qquad \begin{aligned} A([f], \boldsymbol{\theta}, \boldsymbol{\theta}') &= \mathbb{E}_\nu [|f|^2 \nabla_{\boldsymbol{\theta}} \varphi(\cdot, \boldsymbol{\theta}) \otimes \nabla_{\boldsymbol{\theta}'} \varphi(\cdot, \boldsymbol{\theta}')] \\ &\quad - \mathbb{E}_\nu [f \, \nabla_{\boldsymbol{\theta}} \varphi(\cdot, \boldsymbol{\theta})] \otimes \mathbb{E}_\nu [f \, \nabla_{\boldsymbol{\theta}'} \varphi(\cdot, \boldsymbol{\theta}')]. \end{aligned}$$

The SDE capturing the behavior of the solution to (4.1) is

$$(4.7) \qquad d\boldsymbol{\theta}_i = \nabla F(\boldsymbol{\theta}_i) dt - \frac{1}{n} \sum_{j=1}^{n} \nabla K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) dt + \sqrt{\sigma} \, d\boldsymbol{B}_i,$$

where $\sigma = \Delta t / P$ and $\{d\boldsymbol{B}_i\}_{i=1}^{n}$ is a white-noise process with quadratic variation

$$(4.8) \qquad \langle d\boldsymbol{B}_i, d\boldsymbol{B}_j \rangle = A([f - f^{(n)}], \boldsymbol{\theta}_i, \boldsymbol{\theta}_j) dt.$$

More precisely, [19, 25]:

LEMMA 4.1. *Given any test functions* $\chi : D \to \mathbb{R}$ *and any* $T > 0$, *there is a constant* $C > 0$ *such that*

$$(4.9) \qquad \sup_{0 \leq k \Delta t \leq T} \left| \frac{1}{n} \sum_{i=1}^{n} \big(\mathbb{E} \chi(\widehat{\boldsymbol{\theta}}_i(k\Delta t)) - \mathbb{E} \chi(\boldsymbol{\theta}_i(k\Delta t)) \big) \right| \leq C \Delta t.$$

where $\widehat{\boldsymbol{\theta}}_i(t)$ *and* $\boldsymbol{\theta}_i(t)$ *denote the solutions to* (4.1) *and* (4.12), *respectively.*

This lemma is a direct consequence of the fact that (4.1) can be viewed as the Euler-Maruyama discretization scheme for (4.7), and this scheme has weak order of accuracy 1. Note that if we let $\Delta t \to 0$, (4.7) reduces to the ODEs in (3.1) since $\sigma = \Delta t / P \to 0$ in that limit. We should stress, however, that this limit is not reached in practice since the scheme (4.1) is used at small but finite Δt. We analyze next what happens when we adjust the size of σ by changing Δt and/or the batch size P.

4.2 Dean's equation for particles with correlated noise

Lemma 4.1 indicates that we can analyze the properties of (4.7) instead of those of (4.1). To this end, we derive an equation for the empirical distribution $\mu_t^{(n)}$ in (3.3) when $\{\theta_i(t)\}_{i=1}^n$ satisfy the SDE (4.7); this calculation is operationally similar to the derivation of (3.5) but takes into account the extra drift term and the noise term in (4.7). By applying Itô's formula to (3.7) we deduce

$$
\begin{aligned}
d \int_D \chi(\boldsymbol{\theta})\mu_t^{(n)}(d\boldsymbol{\theta}) \\
= \frac{1}{n}\sum_{i=1}^n \nabla\chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{\theta}_i(t) \\
+ \frac{\sigma}{2n}\sum_{i=1}^n \nabla\nabla\chi(\boldsymbol{\theta}_i(t)) : A\big([f - f_t^{(n)}], \boldsymbol{\theta}_i(t), \boldsymbol{\theta}_i(t)\big)dt
\end{aligned}
$$

(4.10)

where $f_t^{(n)} = n^{-1}\sum_{i=1}^n \varphi(\cdot, \boldsymbol{\theta}_i(t)) = \int_D \varphi(\cdot, \boldsymbol{\theta})\mu_t^{(n)}(d\boldsymbol{\theta})$. Using (4.7) and the definition of $\mu_t^{(n)}$, this relation can be written as

$$
\begin{aligned}
d \int_D \chi(\boldsymbol{\theta})\mu_t^{(n)}(d\boldsymbol{\theta}) \\
= \int_D \nabla\chi(\boldsymbol{\theta}) \cdot \nabla V\big(\boldsymbol{\theta}, [\mu_t^{(n)}]\big)\mu_t^{(n)}(d\boldsymbol{\theta})dt \\
+ \frac{\sigma}{2}\int_D \nabla\nabla\chi(\boldsymbol{\theta}) : A\big([f - f_t^{(n)}], \boldsymbol{\theta}, \boldsymbol{\theta}\big)\mu_t^{(n)}(d\boldsymbol{\theta})dt \\
+ \frac{\sqrt{\sigma}}{n}\sum_{i=1}^n \nabla\chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{B}_i(t)
\end{aligned}
$$

(4.11)

The drift terms in this equation are expressed in term of $\mu_t^{(n)}$; for the noise term, notice that its quadratic variation is

$$
\begin{aligned}
\Big\langle \frac{\sqrt{\sigma}}{n}\sum_{i=1}^n \nabla\chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{B}_i(t), \frac{\sqrt{\sigma}}{n}\sum_{i=1}^n \nabla\chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{B}_i(t) \Big\rangle \\
= \sigma \int_{D \times D} \nabla\chi(\boldsymbol{\theta})\nabla\chi(\boldsymbol{\theta}') : A([f - f_t^{(n)}], \boldsymbol{\theta}, \boldsymbol{\theta}')\mu_t^{(n)}(d\boldsymbol{\theta})\mu_t^{(n)}(d\boldsymbol{\theta}')dt
\end{aligned}
$$

This means that, in law, (4.11) is equivalent to

$$
\begin{aligned}
d \int_D \chi(\boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta}) \\
= \int_D \nabla\chi(\boldsymbol{\theta}) \cdot \nabla V(\boldsymbol{\theta}, [\mu_t^{(n)}]) \mu_t^{(n)}(d\boldsymbol{\theta}) dt \\
+ \frac{\sigma}{2} \int_D \nabla\nabla\chi(\boldsymbol{\theta}) : A([f - f_t^{(n)}], \boldsymbol{\theta}, \boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta}) dt \\
+ \sqrt{\sigma} \int_D \nabla\chi(\boldsymbol{\theta}) \cdot d\boldsymbol{\eta}_t^{(n)}(d\boldsymbol{\theta})
\end{aligned}
\tag{4.12}
$$

where $d\boldsymbol{\eta}_t^{(n)}(d\boldsymbol{\theta})$ is vector-valued random measure, white in time, and with quadratic variation

$$
\langle d\boldsymbol{\eta}_t^{(n)}(d\boldsymbol{\theta}), d\boldsymbol{\eta}_t^{(n)}(d\boldsymbol{\theta}') \rangle = A([f - f_t^{(n)}], \boldsymbol{\theta}, \boldsymbol{\theta}') \mu_t^{(n)}(d\boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta}') dt
\tag{4.13}
$$

The first term on the right-hand side of (4.12) is the same as in the weak form of (3.5). This is because these terms come from the drift terms in (4.7), which also coincide with those in (3.1). However, (4.12) also contains additional terms that were absent in (3.5)—note that these terms are different from those in the standard Dean's equation, because the noise term in (4.7) is correlated between the particles instead of being independent.

4.3 LLN for SGD

If we want the result established in Proposition 3.5 to apply and also for the approximation error to vanish as $n \to \infty$, we need to make the additional terms in (4.12) compared to (3.5) higher order. We do this btthey scaling σ with some inverse power of n. Specifically, we will set

$$
\sigma = a n^{-2\alpha} \quad \text{for some } a > 0 \text{ and } \alpha > 0.
\tag{4.14}
$$

This scaling can be achieved by choosing, e.g., $P = O(n^{2\alpha})$, which amounts to increasing the batch size with n. The choice (4.14) implies that the last two terms in (4.12) disappear in the limit as $n \to \infty$. Therefore, we formally conclude that $\mu_t^{(n)} \rightharpoonup \mu_t$ as $n \to \infty$, where μ_t solves the same deterministic equation (3.9) as before. This implies that $\lim_{n\to\infty} f_t^{(n)} = f_t = \int_D \varphi(\cdot, \boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta})$ satisfies (3.17) and is such that $f_t \to f$ as $t \to \infty$. In particular, both the LLN and the global convergence result in Proposition 3.5 still hold if the assumption in this proposition is met, and we use the solution of (4.19) in (3.4). In turn, we can also conclude that this proposition holds up to discretization errors in Δt if we use the solution of (4.1) in (3.4). Importantly, the covariance associated with the estimator for the gradient, defined in (4.6), satisfies

$$
\forall (\boldsymbol{\theta}, \boldsymbol{\theta}') \in D \times D : \lim_{t\to\infty} A([f_t - f], \boldsymbol{\theta}, \boldsymbol{\theta}') = 0.
\tag{4.15}
$$

This property will be useful later.

4.4 CLT for SGD

Turning our attention to the fluctuations of $\mu_t^{(n)}$ around μ_t, notice that there are two sources of them: some are intrinsic to the discrete nature of the particles apparent in $\mu_t^{(n)}$ and scale as $O(n^{-1/2})$ for all $t < \infty$ and possibly as $O(n^{-\xi})$ for any $\xi < 1$ as $t \to \infty$, as discussed in Section 3.5. Other fluctuations come from the noise term in (4.12) and scale as $O(n^{-\alpha})$ when (4.14) holds. The Itô drift terms proportional to $\sigma = an^{-2\alpha}$ in (4.12) always make higher-order contributions.

We first consider $t < \infty$ and subsequently examine the limit $t \to \infty$ in Section 4.5. In the present case, we first observe that if $\alpha \geq \frac{1}{2}$, then for all $t < \infty$ the fluctuations due to the noise in (4.12) are negligible compared to the intrinsic ones from discreteness, and we are back to the GD situation studied in Section 3. In contrast, if $\alpha \in (0, \frac{1}{2})$, for all $t < \infty$ the fluctuations due to the noise in (4.12) dominate the intrinsic ones from discreteness, so let us focus on this case from now on. To quantify these fluctuations, we can introduce $n^\alpha(\mu_t^{(n)} - \mu_t)$, write an equation for this scaled discrepancy, and take the limit as $n \to \infty$. The derivation proceeds akin to the derivation of (3.47) and leads to the conclusion that, as $n \to \infty$, $n^\alpha(\mu_t^{(n)} - \mu_t) \rightharpoonup \omega_t^{(\alpha)}$ in law, which satisfies

$$
d \int_D \chi(\boldsymbol{\theta}) \omega_t^{(\alpha)}(d\boldsymbol{\theta}) = \int_D \nabla\chi(\boldsymbol{\theta}) \cdot \nabla V(\boldsymbol{\theta}, [\mu_t]) \omega_t^{(\alpha)}(d\boldsymbol{\theta}) dt
$$

(4.16)
$$
+ \int_D \nabla\chi(\boldsymbol{\theta}) \cdot \nabla F(\boldsymbol{\theta}, [\omega_t^{(\alpha)}]) \mu_t(d\boldsymbol{\theta}) dt
$$

$$
+ \sqrt{a} \int_D \nabla\chi(\boldsymbol{\theta}) \cdot d\boldsymbol{\eta}_t(d\boldsymbol{\theta}),
$$

in which $d\boldsymbol{\eta}_t(d\boldsymbol{\theta})$ is a vector-valued random measure, white in time, and with quadratic variation (compare (4.13))

$$
(4.17) \qquad \langle d\boldsymbol{\eta}_t(d\boldsymbol{\theta}), d\boldsymbol{\eta}_t(d\boldsymbol{\theta}') \rangle = A([f - f_t], \boldsymbol{\theta}, \boldsymbol{\theta}') \mu_t(d\boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta}') dt.
$$

Equation (4.16) should be solved with zero initial condition, since the $O(n^{-1/2})$ fluctuations arising from the initial condition are higher order compared to the scaling $O(n^{-\alpha})$ we picked to obtain (4.16). Since (4.16) is linear in $\omega_t^{(\alpha)}$ with additive noise, it indicates that $\omega_t^{(\alpha)}$, a Gaussian process with mean zero and thereby fully characterized by its covariance (we omit the equation for brevity). This also implies that

$$
(4.18) \qquad n^\alpha \left(f_t^{(n)} - f \right) \to g_t^{(\alpha)} \quad \text{in law as } n \to \infty,
$$

where $g_t^{(\alpha)}$ is a Gaussian process whose evolution equation (cf. the derivation of (3.54)) gives

$$
\begin{aligned}
(4.19) \quad dg_t^{(\alpha)} = {}& -\int_\Omega M([\omega_t^\alpha], x, x') \left(f_t(x) - f(x') \right) v(dx')dt \\
& -\int_\Omega M([\mu_t], x, x') g_t^{(\alpha)}(x') v(dx')dt + \sqrt{a}\, d\zeta_t(x),
\end{aligned}
$$

where $M([\mu], x, x')$ is given in (3.18), and the quadratic variation of $d\zeta_t$ is that of $\int_D \varphi(\cdot, \theta) d\eta_t(d\theta)$. Explicitly,

$$
\begin{aligned}
(4.20) \quad & g\langle d\zeta_t(x), d\zeta_t(x')\rangle \\
& = \int_\Omega M([\mu_t], x, x'') M([\mu_t], x', x'') \left| f_t(x'') - f(x'') \right|^2 dv(x'')dt \\
& \quad - \int_\Omega M([\mu_t], x, x'') \left(f_t(x'') - f(x'') \right) dv(x'') \\
& \qquad \times \int_\Omega M([\mu_t], x', x'') \left(f_t(x'') - f(x'') \right) dv(x'')dt.
\end{aligned}
$$

The SDE (4.19) should be solved with zero initial condition, $g_0^{(\alpha)} = 0$. Since it is linear in $g_t^{(\alpha)}$ with additive noise, it defines a Gaussian process with mean zero and is specified by its covariance

$$
(4.21) \qquad C_t^{(\alpha)}(x, x') = \mathbb{E}\left[g_t^{(\alpha)}(x) g_t^{(\alpha)}(x) \right],
$$

where $\mathbb{E}$ denotes expectation over the noise $d\zeta_t$ (that is, over the data in the batches used in SGD). With this calculation, we have established the following:

PROPOSITION 4.2 (CLT for SGD). *Consider*

$$
(4.22) \qquad g_t^{(\alpha,n)} = n^{\alpha-1} \sum_{i=1}^n \left(\varphi(\cdot, \theta_i(t)) - f_t \right) = n^\alpha \left(f_t^{(n)} - f_t \right)
$$

with $\{\theta_i(t)\}_{i=1}^n$ the solution to the SDE (4.7) with $\sigma = an^{-2\alpha}$, $\alpha \in (0, \frac{1}{2})$, and f_t the solution to (3.17). Then, as $n \to \infty$, $g_t^{(\alpha,n)}$ converges in law towards the Gaussian process $g_t^{(\alpha)}$, the solution of (4.19) for $g_0^{(\alpha)} = 0$.

4.5 Fluctuations in SGD at long and very long times

The noise $d\zeta_t$ in (4.19) has the remarkable property that it self-quenches as $t \to \infty$ if the conditions of Proposition 3.5 are met and $f_t \to f$ as $t \to \infty$ and therefore, from (4.20):

$$
(4.23) \qquad \forall x, x' \in \Omega : \lim_{t \to \infty} \langle d\zeta_t(x), d\zeta_t(x')\rangle = 0.
$$

Since the first drift term in (4.19) also goes to zero when $f_t \to f$ and the second drift term is a damping term because $M([\mu_t], x, x')$ is positive definite for all $t < \infty$, we know that $g_t^{(\alpha)}$ will be controlled as $t \to \infty$, i.e., $C_t^{(\alpha)}$ as a limit. In

addition, if supp $\widehat{\mu}^* = \widehat{D}$ where $\widehat{\mu}^* = \int_{\mathbb{R}} \mu^*(dc, \cdot) = \lim_{t \to \infty} \int_{\mathbb{R}} \mu^*(dc, \cdot)$, then $M([\mu_t], x, x')$ is positive definite for all $t < 0$ and in the limit as $t \to \infty$, and the solution to (4.19) goes to zero. Using the the Gaussian process $g_t^{(n, \alpha)}$ defined in (4.22), we can summarize this result as follows:

PROPOSITION 4.3 (Fluctuations in SGD at long time). *Under the conditions of Proposition 3.5, if $f_t^{(n)}$ is given by (3.4) with $\{\theta_i(t)\}_{i=1}^n$ the solution of (4.7) with $\sigma = an^{-2\alpha}$, $\alpha \in (0, \frac{1}{2})$, and initial condition drawn from $\mathbb{P}_{in}$, and f_t solves (3.17), then*

$$(4.24) \quad \lim_{t \to \infty} \lim_{n \to \infty} n^{2\alpha} \mathbb{E} \int_{\Omega} \left| f_t^{(n)}(x) - f_t(x) \right|^2 \nu(dx) = \lim_{t \to \infty} C_t^{(\alpha)}(x, x') \text{ exists.}$$

In addition, if supp $\widehat{\mu}^* = \widehat{D}$, *then this limit is zero.*

If supp $\widehat{\mu}^* = \widehat{D}$ where $\widehat{\mu}^* = \int_{\mathbb{R}} \mu^*(dc, \cdot) = \lim_{t \to \infty} \int_{\mathbb{R}} \mu^*(dc, \cdot)$, then $M([\mu_t], x, x')$ is positive definite for all $t < 0$ and in the limit as $t \to \infty$. In that case, the only fixed point of (4.19) is zero. Since in this case we also know that the fluctuations from the initial conditions disappear on scale $O(n^\xi)$ for any $\xi < 0$, we can proceed as in Section 3.5 and adjust α all the way up to 1 instead of $\frac{1}{2}$. That is, we can generalize Proposition 3.9 as follows:

PROPOSITION 4.4 (Fluctuations in SGD at very long times). *Under the conditions of Proposition 4.3, if* supp $\widehat{\mu}^* = \widehat{D}$, *then for any $\alpha \in (0, 1)$,*

$$(4.25) \quad \lim_{n \to \infty} n^{2\alpha} \mathbb{E} \int_{\Omega} \left| f_{a_n}^{(n)}(x) - f(x) \right|^2 \nu(dx) = 0$$

if a_n grows with n and is such that $\lim_{n \to \infty} a_n / \log n = \infty$—here $\mathbb{E}$ denotes expectation over both the initial condition $\mathbb{P}_{in}$ and the noise in (4.7).

5 Illustrative Example: 3-Spin Model on the High-Dimensional Sphere

To test our results, we use a function known for its complex features in high dimensions: the spherical 3-spin model, $f : S^{d-1}(\sqrt{d}) \to \mathbb{R}$, given by

$$(5.1) \qquad f(x) = \frac{1}{d} \sum_{p,q,r=1}^{d} a_{p,q,r} x_p x_q x_r, \quad x \in S^{d-1}(\sqrt{d}) \subset \mathbb{R}^d,$$

where the coefficients $\{a_{p,q,r}\}_{p,q,r=1}^d$ are independent Gaussian random variables with mean zero and variance one. The function (5.1) is known to have a number of critical points that grows exponentially with the dimensionality d [2, 3, 29]. We note that previous works have sought to draw a parallel between the glassy 3-spin function and generic loss functions [13], but we are not exploring such an analogy here. Rather, we simply use the function (5.1) as a difficult target for approximation by neural networks. That is, throughout this section, we train networks to learn f

with a particular realization of $a_{p,q,r}$ and study the accuracy of that representation as a function of the number of particles n.

5.1 Learning with Gaussian kernels

We first consider the case when $D = S^{d-1}(\sqrt{d})$, and we use

$$(5.2) \qquad \varphi(x, z) = e^{-\frac{1}{2}\alpha|x-z|^2}$$

for some fixed $\alpha > 0$. In this case, the parameters are elements of the domain of the function (here the d-dimensional sphere). Note that, since $|x| = |z| = \sqrt{d}$, up to an irrelevant constant that can be absorbed in the weights c, we can also write (5.2) as

$$(5.3) \qquad \varphi(x, z) = e^{-\alpha x \cdot z}.$$

This setting allows us to simplify the problem. Using

$$(5.4) \qquad f^{(n)}(x) = \frac{1}{n}\sum_{i=1}^{n} c_i \varphi(x, z_i) = \frac{1}{n}\sum_{i=1}^{n} c_i e^{-\alpha x \cdot z_i},$$

we can use as alternative loss

$$(5.5) \qquad \mathcal{L}[f^{(n)}] = -\frac{1}{n}\sum_{i=1}^{n} c_i f(z_i) + \frac{1}{2n^2}\sum_{i,j=1}^{n} c_i c_j \varphi(z_i, z_j);$$

i.e., eliminate the need for data beside the set $\{z_i\}_{i=1}^{n}$. In terms of the empirical distribution, the loss can be represented as

$$(5.6) \quad \mathcal{L}[f^{(n)}] = -\int_{\widehat{D}} f(z)\gamma^{(n)}(dz) + \frac{1}{2}\int_{\widehat{D}\times\widehat{D}} \varphi(z, z')\gamma^{(n)}(dz)\gamma^{(n)}(dz'),$$

where $\gamma^{(n)} = \int_{\mathbb{R}} c\mu^{(n)}(dc, \cdot)$. Viewed as an integral kernel, φ is positive definite; as a result, the loss is a convex functional of $\gamma^{(n)}$ (or $\mu^{(n)}$). Hence, the results established above apply to this special case as well. The GD flow on the loss (5.5) can now be written explicitly as

$$(5.7) \qquad \begin{cases} \dot{z}_i = c_i \nabla f(z_i) + \frac{\alpha}{n}\sum_{j=1}^{n} c_i c_j z_j e^{-\alpha z_i \cdot z_j} - \lambda_i z_i, \\ \dot{c}_i = f(z_i) - \frac{1}{n}\sum_{j=1}^{n} c_j e^{-\alpha z_i \cdot z_j}, \end{cases}$$

where $-\lambda_i z_i$ is a Lagrange multiplier term added to enforce $|z_i| = \sqrt{d}$ for all $i = 1, \ldots, n$, $f(x)$ is given by (5.1), and $\nabla f(z)$ is given componentwise by

$$(5.8) \qquad \frac{\partial f}{\partial z_p} = \frac{1}{d}\sum_{q,r=1}^{d} (a_{p,q,r} + a_{r,p,q} + a_{q,r,p})z_q z_r.$$

As is apparent from (5.7), the advantage of using radial basis function networks (or, in fact, any unit $\widehat{\phi}$ which is (i) such that $\widehat{D} = \Omega$ and (ii) positive definite) is that we can use $f(x)$ and the unit $\varphi(x, z)$ directly, and do not need to evaluate $\widehat{F}(z)$ and $\widehat{K}(z, z')$ (that is, we need no batch). In other words, the cost of running (5.7)
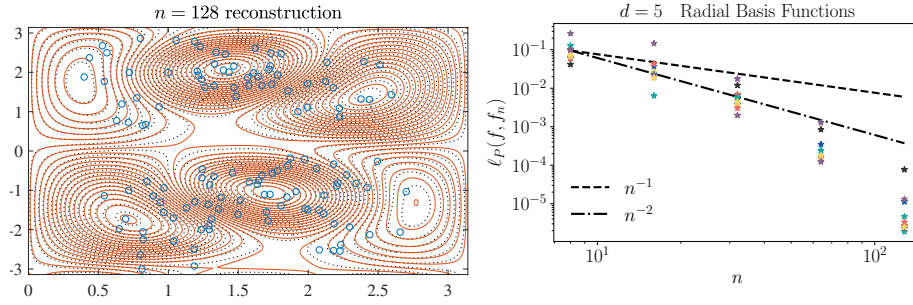
FIGURE 5.1. Left panel: Comparison between the level sets of the original function f in (5.1) (black dotted curves) and its approximation by the neural network in (5.4) with $n = 128$ and $d = 5$ in the slice defined by (5.9). Also shown is the projection in the slice of the particle position. Right panel: empirical loss in (5.10) vs. n at the end of the calculation. The stars show the empirical loss for 10 independent realizations of the coefficients $a_{p,q,r}$ in (5.1).

scales like $(dn)^2$ instead of $P(Nn)^2$ in the case of a general network optimized by SGD with a batch of size P and $z \in \hat{D} \subset \mathbb{R}^N$. If we make P scale with n, like $P = Cn^{2\alpha}$ for some $C > 0$, as we need to do to obtain the scalings discussed in Section 4, the cost of SGD becomes $N^2 n^{2+2\alpha}$, which quickly becomes much worse than $(dn)^2$ as n grows.

We tested the representation (5.4) in $d = 5$ using $n = 16, 32, 64, 128$, and 256 and setting $\alpha = 5/d = 1$. The training was done by running a time-discretized version of (5.7) with time step $\Delta t = 10^3$ for 2×10^5 steps: during the first 10^5 we added thermal noise to (5.7), which we then removed during the second half of the run. The representation (5.4) proves to be accurate even at rather low values of n: for example, the right panel of Figure 5.1 shows a contour plot of the original function f and its representation $f^{(n)}$ with $n = 128$ through a slice of the sphere defined as

$$(5.9) \qquad x(\theta) = \sqrt{d} \left(\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta), 0, 0 \right),$$

with $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. The level sets of both functions are in good agreement. Also shown on this figure is the projection on the slice of the position of the 64 particles on the sphere. In this result, the parameters c_i take values that are initially uniformly distributed by about $-40d^2 = -10^3$ and $40d^2 = 10^3$. To test the accuracy of the representation, we used the following Monte Carlo estimate of the loss function:

$$(5.10) \qquad \mathcal{L}_P\big[f_t^{(n)}\big] = \frac{1}{2P} \sum_{p=1}^{P} \big| f(x_p) - f_t^{(n)}(x_p) \big|^2.$$
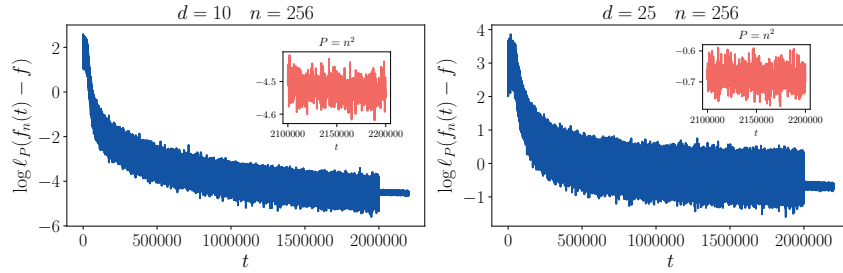
FIGURE 5.2. The log of the empirical loss in (5.10) as a function of training time by SGD for the sigmoid neural network in $d = 10$ (left panel) and $d = 25$ (right panel). At time $t = 2 \times 10^6$, the batch size is increased to initiate a quench. The insets show the log of the empirical loss as a function of time during the final 10^5 time steps of training.

This empirical loss function was computed with a batch of 10^6 points x_p uniformly distributed on the sphere. The value (5.10) calculated at the end of the calculation is shown as a function of n in the right panel of Figure 5.1: the empty circles show (5.10) for four individual realizations of the coefficient $a_{p,q,r}$ in (5.1); the full circle shows the average of (5.10) over these four realizations. The blue line scale as n^{-1}, the red one as n^{-2}: as can be seen, the empirical loss decays with n faster than n^{-1}, which is as expected.

5.2 Learning with single layer networks with sigmoid nonlinearity

To further test our predictions and also assess the learnability of high-dimensional functions, we used 3-spin models in $d = 10$ and 25 dimensions, which we approximated with a single-layer neural network with sigmoid nonlinearity parametrized by $z = (a, b) \in D = \mathbb{R}^{d+1}$, with $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, and

$$(5.11) \qquad \varphi(x, z) = h(a \cdot x + b).$$

This gives

$$(5.12) \qquad f^{(n)}(x) = \frac{1}{n} \sum_{i=1}^{n} c_i h(a_i \cdot x + b_i),$$

where $h(z) = 1/(1 + e^{-z})$. Simple networks like these, as opposed to deep neural with many parameters, provide greater assurance that we have trained sufficiently to test the scaling.

We trained the model in (5.12) using SGD with an initial batch size of $P = \lfloor n/5 \rfloor$ points uniformly sampled on the sphere for 2×10^6 time steps, resampling a new batch at every time step: this corresponds to choosing $\alpha = 1/2$ in the notation of Section 4. Towards the end of the trajectory, we initiated a partial quench by increasing the batch size to $P = \lfloor (n/5)^2 \rfloor$ (i.e., $\alpha = 1$) which we run for an additional 2×10^5 time steps. Figure 5.2 shows the empirical loss in (5.10) calculated
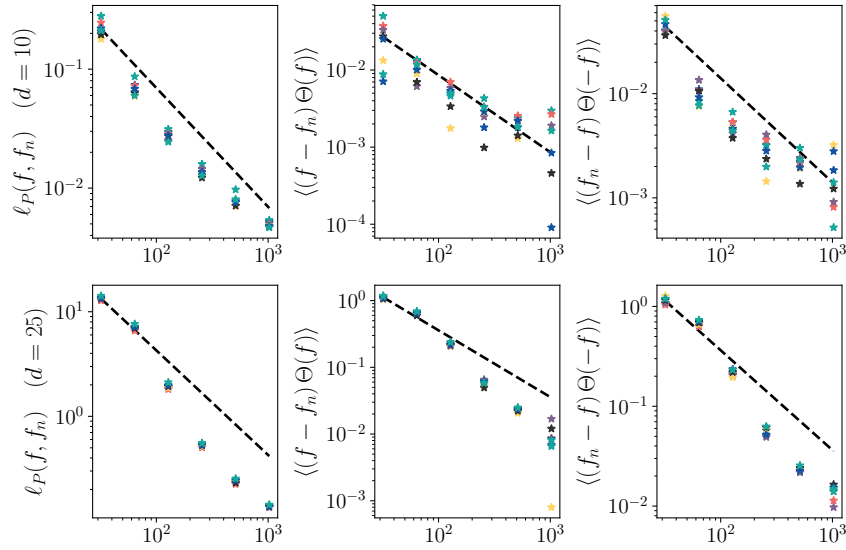
FIGURE 5.3. Error scaling for a single-layer neural network with sigmoid nonlinearities. Upper row: $d = 10$; lower row: $d = 25$. The first column shows the empirical loss in (5.10), the second column shows (5.13), and the third column shows (5.13) with $\Theta(f)$ replaced by $\Theta(-f)$. The stars show the results for 10 different realizations of the coefficients $a_{p,q,r}$ in (5.1): the dashed lines decay as n^{-1}, consistent with the predictions in (4.25) and 3.7 .

over the batch as a function of training time during the optimization with $n = 256$ particles and $d = 10$ (left panel) and $d = 25$ (right panel). Note that the lack of intermediate plateaus in the loss during training is consistent with our conclusion that the dynamics effectively descends on a quadratic energy landscape (i.e., the loss function itself) at the level of the empirical distribution of the particles. After the quench the empirical loss shows substantially smaller fluctuations as a function of time, which helps to reduce the fluctuating error. The inset shows the final 10^5 time steps in which there is negligible downward drift, indicating convergence towards stationarity at this batch size.

In these higher-dimensional examples, we tested the scaling with three different observables. First, we considered the empirical loss function in (5.10) which we computed over a batch of size $\widehat{P} = 10^5$ larger than P. As shown in the two right panels, Figure 5.3, $\mathcal{L}_{\widehat{P}}[f_t^{(n)}]$ scales as n^{-1}, as expected. We also tested the estimate in (4.25) using

$$(5.13) \qquad \frac{1}{\widehat{P}} \sum_{p=1}^{\widehat{P}} \Theta\left(f(x_p)\right)\left(f(x_p) - f_t^{(n)}(x_p)\right),$$
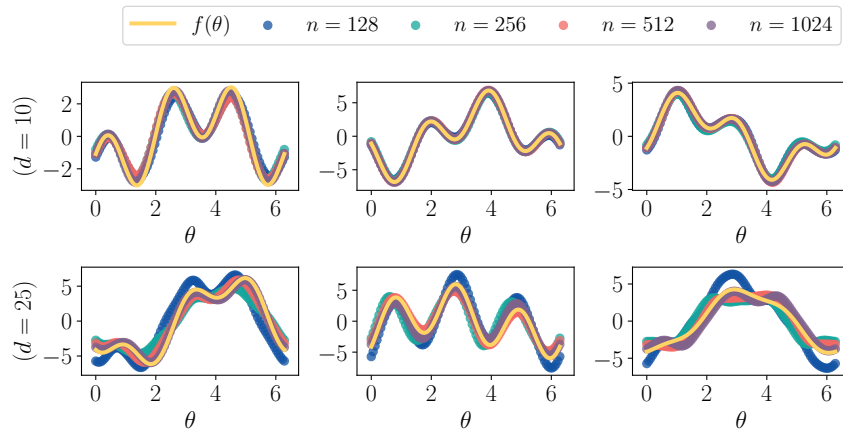
FIGURE 5.4. One-dimensional slices through the $d = 10$ (upper row) and $d = 25$ (lower row) neural net representation $f^{(n)}$ are shown below a yellow curve with the target function f. In $d = 10$, the function representations clearly capture the main features of the target function, with only small scale deviations. In $d = 25$ there is remarkably good signal when $n = 1024$ while the smaller neural network is less able to faithfully represent the target function.

and similarly with $\Theta\left(-f(x_p)\right)$: here Θ denotes the Heaviside function. The result is shown in the four right panels in Figure 5.3: (5.13) scales as n^{-1}, consistent with (4.25) and our choice of $\alpha = 1$.

To provide further confidence in the quality of the representations, we also made a visual comparison by plotting f and $f^{(n)}$ along great circles of the sphere. We do so by picking $i \neq j$ in $\{1, \ldots, d\}$ and setting $x = x(\theta) = (x_1(\theta), \ldots, x_d(\theta))$ with

$$(5.14) \quad x_i(\theta) = \sqrt{d}\cos(\theta), \quad x_j(\theta) = \sqrt{d}\sin(\theta), \quad x_k(\theta) = 0 \quad \forall k \neq i, j.$$

In Figure 5.4 we plot $f(x(\theta))$ and $f^{(n)}(x(\theta))$ along three great circles for $d = 10$ and $d = 25$. As can be seen, the agreement is quite good and confirms the quality of the final fit. A strong signal is present in $d = 25$ with $n = 1024$, a remarkable fact when considering that if we had only two grid points per dimension, the total number of points in the grid would be $2^{25} = 33,554,432$.

6 Concluding Remarks

Viewing parameters as particles with the loss function as interaction potential enables us to leverage a powerful theoretical apparatus developed to analyze problems from statistical physics. Using these ideas, we can analyze the approximation quality and the trainability of neural network representations of high-dimensional functions. Several insights emerge from our analysis based on this viewpoint: First,

these tools show the dynamical realizability of the universal approximation theorems, a direct consequence of the law of large numbers for the empirical distribution of the parameters. Specifically, we conclude that the empirical distribution effectively descends on the quadratic loss function landscape when the number n of parameters in the network is large. This confirms the empirical observation that wide neural networks are trainable despite the nonconvexity of the loss function viewed from the individual particles perspective (as opposed to that of their empirical distribution). Secondly, we have derived a central limit theorem for the empirical distribution of the parameters, specifying the approximation error of the neural network representation and showing that it is universal.

We derived these results first in the context of gradient descent dynamics; however, our conclusions also apply to stochastic gradient descent. The analysis indicates how the parameters in SGD should be chosen, in particular how the batch size should be scaled with n given the time step used in the scheme, which can be done towards the end of training.

These results were derived for a quadratic loss, $\mathcal{L}[f^{(n)}] = \frac{1}{2}\mathbb{E}_\nu|f - f^{(n)}|^2$. However, they do generalize to other losses as long as they are convex in $f^{(n)}$.

We also worked in the limit of an infinite amount of training data, an idealized setting that does not address the error incurred from a finite data set. For a neural network trained on a dataset of P points, $\{x_p\}_{p=1}^P$, we can decompose the "generalization" error into components that involve the approximation error and the error from the finiteness of the data,

$$(6.1) \qquad \mathbb{E}_\nu|f - f_P^{(n)}|^2 \leq \mathbb{E}_\nu|f - f_P|^2 + \mathbb{E}_\nu|f_P - f_P^{(n)}|^2,$$

where f_P and $f_P^{(n)}$ are the approximations of f we can get if we train the network on the empirical loss build on $\{x_p\}_{p=1}^P$ with finitely ($n < \infty$) or infinitely ($n \to \infty$) many units, respectively. Our results give direct insight on the second term on the right-hand side of (6.1). We leave assessments of the first term for future work.

Our numerical results not only confirm our predictions, but also emphasize the capability of neural networks to represent high-dimensional function accurately with a relatively modest number of adjustable parameters. Needless to say, the computational achievements of neural networks open the door to developments in scientific computing that we are only beginning to grasp. Such applications may benefit from better understanding how the specific architecture of the neural networks affects the approximation error and trainability, not in the general terms of their scaling with n that we analyzed here, but in the details of the constant involved.

Appendix: Training at Finite (But Small) Temperature

For completeness, let us consider here the case when noise terms are added in (3.1) and the ODEs become stochastic differential equations (SDEs). Additive noise addresses the nonuniqueness issues encountered in Section 3. To formulate

the resulting SDEs, we need a distribution $\mu_0 \in \mathcal{M}_+(D)$ used to regularize the dynamics. We specify its properties via the following:

ASSUMPTION A.1. *The distribution μ_0* (i) *has a density ρ_0 that is continuously differentiable, $\rho_0 \in C^1(D)$;* (ii) *is such that* $\mathrm{supp}(\mu_0) = D$; *and* (iii) *satisfies*

$$(A.1) \qquad \forall b \in \mathbb{R} : \int_{\mathbb{R}} e^{bc} \mu_0(dc, \cdot) < \infty \quad and \quad \int_{\mathbb{R}} c \mu_0(dc, \cdot) = 0.$$

We then replace (3.1) with the SDEs

$$(A.2) \qquad \begin{aligned} d\boldsymbol{\theta}_i &= \nabla F(\boldsymbol{\theta}_i)dt - \frac{1}{n}\sum_{j=1}^{n} \nabla K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)dt \\ &\quad + (\beta n)^{-1}\nabla \log \rho_0(\boldsymbol{\theta}_i)dt + \sqrt{2}(\beta n)^{-1/2} d\boldsymbol{W}_i, \end{aligned}$$

for $i = 1, \ldots, n$. Here $\boldsymbol{W}_i$ are n independent Wiener processes, taking values in D, and $\beta > 0$ is a parameter playing the role of inverse temperature and controlling the amplitude of a noise added to the dynamics. Note the specific scale on which the regularizing and the noise terms act in (A.2): they are higher-order perturbations. We comment on the choice of this scaling in Remark A.2 below. The SDEs (A.2) are overdamped Langevin equations associated with the energy:

$$(A.3) \qquad \begin{aligned} E_\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) &= nC_f - \sum_{i=1}^{n} F(\boldsymbol{\theta}_i) + \frac{1}{2n}\sum_{i,j=1}^{n} K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \\ &\quad - (\beta n)^{-1}\sum_{i=1}^{n} \log \rho(\boldsymbol{\theta}_i), \end{aligned}$$

This energy is (3.2) plus a regularizing term (the one involving $-\log \rho_0$). Under Assumption A.1 this term guarantees that, for any $\beta > 0$, the following integral is finite:

$$(A.4) \qquad Z_n = \int_{D^n} e^{-n\beta E_\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)} d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_n < \infty,$$

which in turn implies that

$$(A.5) \qquad Z_n^{-1} \exp\left(-n\beta E_\beta(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)\right)$$

is a normalized probability density on D^n. As a result, the solutions to (3.1) are ergodic with respect to the equilibrium distribution with density (A.5) for any $\beta > 0$.

A.1 Dean's equation

Let $\chi : D \to \mathbb{R}$ be a test function. Applying Itô's formula to

$$n^{-1}\sum_{i=1}^{n}\chi(\boldsymbol{\theta}_i(t)) = \int_D \chi(\boldsymbol{\theta})\mu_t^{(n)}(d\boldsymbol{\theta})$$

and using (A.2) gives

$$
d \int_D \chi(\boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta})
$$

$$
= \frac{1}{n} \sum_{i=1}^n \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{\theta}_i(t) + \beta^{-1} \sum_{i=1}^n \Delta \chi(\boldsymbol{\theta}_i(t)) dt
$$

$$
= \int_D \nabla \chi(\boldsymbol{\theta}) \cdot \left(\nabla F(\boldsymbol{\theta}) \mu_t^{(n)}(\boldsymbol{\theta}) - \int_D \nabla K(\boldsymbol{\theta}, \boldsymbol{\theta}') \mu_t^{(n)}(d\boldsymbol{\theta}') \mu_t^{(n)}(d\boldsymbol{\theta}) \right) dt
$$

$$
\text{(A.6)} \qquad + (\beta n)^{-1} \int_D \nabla \chi(\boldsymbol{\theta}) \cdot \left(\nabla \log \rho_0(\boldsymbol{\theta}) \, \mu_t^{(n)}(d\boldsymbol{\theta}) \right) dt
$$

$$
+ (\beta n)^{-1} \int_D \Delta \chi(\boldsymbol{\theta}) \mu_t^{(n)}(d\boldsymbol{\theta}) dt
$$

$$
+ \sqrt{2}(\beta n^3)^{-1/2} \sum_{i=1}^n \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{W}_i(t)
$$

The drift terms in this equation are closed in terms of $\mu_t^{(n)}$; the noise term has a quadratic variation given by

$$
\left\langle (\beta n^3)^{-1/2} \sum_{i=1}^n \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{W}_i(t), (\beta n^3)^{-1/2} \sum_{i=1}^n \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot d\boldsymbol{W}_i(t) \right\rangle
$$

$$
\text{(A.7)} \qquad = \beta^{-1} n^{-3} \sum_{i=1}^n |\nabla \chi(\boldsymbol{\theta}_i(t))|^2 \, dt
$$

$$
= \beta^{-1} n^{-2} \int_D |\nabla \chi(\boldsymbol{\theta})|^2 \mu_t^{(n)}(d\boldsymbol{\theta}) dt.
$$

As a result, (A.6) is sometimes written formally as the stochastic partial differential equation (SPDE)

$$
\partial_t \mu_t^{(n)} = \nabla \cdot \left(\nabla F \mu_t^{(n)} + \int_D \nabla K(\boldsymbol{\theta}, \boldsymbol{\theta}') \mu^{(n)}(d\boldsymbol{\theta}') \mu_t^{(n)} \right)
$$

$$
\text{(A.8)} \qquad - (\beta n)^{-1} \nabla \cdot \left(\nabla \log \rho_0 \mu_t^{(n)} \right)
$$

$$
+ (\beta n)^{-1} \Delta \mu_t^{(n)} + \sqrt{2} \beta^{-1/2} n^{-1} \nabla \cdot \left(\sqrt{\mu_t^{(n)}} \, \dot{\boldsymbol{\eta}}_t \right)
$$

where $\dot{\boldsymbol{\eta}}_t = \dot{\boldsymbol{\eta}}_t(\boldsymbol{\theta})$ is a spatiotemporal white noise so that the quadratic variation of the noise term in (A.8) is formally given by (A.7). This equation is referred to as Dean's equation. It is difficult to give (3.5) a precise meaning because it is not clear how to interpret the noise term. It remains useful to analyze the properties of $\mu_t^{(n)}$ as $n \to \infty$, however, which is what we will do next.

Remark A.2. We could also consider situations where in (A.2) n^{-1} is replaced by $n^{-\alpha}$ with $\alpha \in [0, 1)$. The case $\alpha = 0$ is treated in [27]: with this scaling, the

diffusive and regularizing terms in (A.8) are replaced by

$$\beta^{-1}\Delta\mu_t^{(n)} - \beta^{-1}\nabla \cdot \left(\nabla\log\rho_0\mu_t^{(n)}\right),$$

and the noise terms by

$$\sqrt{2}(\beta n)^{-1/2}\nabla \cdot \left(\sqrt{\mu_t^{(n)}}\,\dot{\eta}_t\right).$$

This means that these diffusive and regularizing terms affect the mean field limit equation for μ_t, whereas the noise terms remain higher order. In particular, in that case one can prove that $\mu_t^{(n)} \rightharpoonup \mu_t$ with μ_t that converges to a unique fixed point μ_β such that $\mu_\beta > 0$ a.e. in D but for which $\int_D \varphi(\cdot,\theta)\mu_\beta(d\theta) \neq f$ (there is a correction proportional to β^{-1}). When $\alpha \in (0,1)$, the diffusive and regularizing terms in (A.8) are replaced by

$$\beta^{-1}n^{-\alpha}\Delta\mu_t^{(n)} - \beta^{-1}n^{-\alpha}\nabla \cdot \left(\nabla\log\rho_0\,\mu_t^{(n)}\right),$$

and the noise terms by

$$\sqrt{2}(\beta n^{1+\alpha})^{-1/2}\,\nabla \cdot \left(\sqrt{\mu^{(n)}}\,\dot{\eta}_t\right).$$

This means that none of these terms affect the mean field limit equation, but at next order, $O(n^{-\alpha})$, the diffusive and regularizing terms dominate whereas the noise terms remain higher order. In the case when $\alpha = 1$, on which we focus here, the diffusive, regularizing, and noise terms are perturbations on the $O(n^{-1})$ same scale, the same scale as the errors introduced by discretization effects (finite n) also present in GD.

A.2 Multiple-scale expansion

The advantage of adding noise terms in (A.2) is that it guarantees ergodicity of the solution to these SDEs with respect to the equilibrium distribution with density (A.5). Correspondingly, we focus on analyzing the long-time ergodicity properties of the empirical distribution satisfying (A.8). On long timescales, the memory of the initial conditions is lost, and we can directly pick the right scaling to analyze the fluctuations of $\mu_t^{(n)}$ around its limit μ_t: as discussed in Remark A.2 and confirmed below, this scale is $O(n^{-1})$, consistent with what we reach at long times with GD as discussed in Section 3.

We analyze (A.8) by formal asymptotic, using a two-timescale expansion. Consistent with the expected $O(n^{-1})$ scaling of the fluctuations, we look for a solution of this equation of the form

(A.9) $$\mu_t^{(n)} = \mu_{t,\tau} + n^{-1}\omega_{t,\tau} + o(n^{-1}), \quad \tau = t/n.$$

We use the rescaled time $\tau = t/n$ to look at the solution to (A.8) on $O(n)$ timescales. Not only does this fix the behavior of $\mu_{t,\tau}$ on long timescales but also

guarantees solvability of the equation for $\omega_{t,\tau}$. Treating t and τ as independent variables, (A.9) implies that

$$(\text{A.10}) \qquad \partial_t \mu_t^{(n)} = \partial_t \mu_{t,\tau} + n^{-1}(\partial_\tau \mu_{t,\tau} + \partial_t \omega_{t,\tau})$$

Inserting (A.9) and (A.10) in (A.8) and collecting terms of the same order in n^{-1}, we arrive at the following two equations at order $O(1)$ and $O(n^{-1})$, respectively,

$$(\text{A.11}) \qquad \partial_t \mu_{t,\tau} = \nabla \cdot \big(\nabla V(\boldsymbol{\theta}, [\mu_{t,\tau}]) \mu_{t,\tau}\big)$$

and

$$
\begin{aligned}
(\text{A.12}) \qquad \partial_\tau \mu_{t,\tau} + \partial_t \omega_{t,\tau} &= \nabla \cdot \big(\nabla V(\boldsymbol{\theta}, [\mu_{t,\tau}]) \omega_{t,\tau} + \nabla F(\boldsymbol{\theta}, [\omega_{t,\tau}]) \mu_{t,\tau}\big) \\
&\quad + \beta^{-1} \Delta \mu_{t,\tau} - \beta^{-1} \nabla \cdot \big(\nabla \log \rho_0\, \mu_{t,\tau}\big) \\
&\quad + \sqrt{2}\beta^{-1/2} \nabla \cdot \big(\sqrt{\mu_{t,\tau}}\dot{\boldsymbol{\eta}}_t\big)
\end{aligned}
$$

A.3 Law of large numbers at finite temperature

Since (A.11) is identical to (3.9), the results we established in Section 3.3 still hold at finite temperature. In particular, Proposition 3.5 applies. As we see below, we can obtain more information about μ_t by looking at the evolution of this function on longer timescales, and we will be able to deduce that $\operatorname{supp}\mu_{t,\tau} = D$. This guarantees that (2.10) holds, so it can be removed from the assumptions needed in Proposition 3.5.

A.4 Global convergence on $O(n)$ timescales

An equation governing the evolution of $\mu_{t,\tau}$ on the rescaled time $\tau = t/n$ can be derived by time averaging (A.12) over t. This equation guarantees the solvability of (A.12). Since $\mu_{t,\tau} \to \mu_\tau$ as $t \to \infty$, where μ_τ is a stationary point of (A.11), we have

$$(\text{A.13}) \qquad \lim_{T \to \infty} \frac{1}{T} \int_0^T \mu_{t,\tau} \omega_{t,\tau}\, dt = \mu_\tau \bar{\omega}_\tau$$

where

$$(\text{A.14}) \qquad \bar{\omega}_\tau =: \lim_{T \to \infty} \frac{1}{T} \int_0^T \omega_{t,\tau}\, dt$$

in which we assume that the time average of $\omega_{t,\tau}$ exists (which we check a posteriori). Using (A.13) and the fact that the white-noise terms time-average to zero almost surely, we deduce that the time average of (A.12) is

$$
\begin{aligned}
(\text{A.15}) \qquad \partial_\tau \mu_\tau &= \nabla \cdot (\nabla V(\boldsymbol{\theta}, [\mu_\tau])\bar{\omega}_\tau + \nabla F(\boldsymbol{\theta}, [\bar{\omega}_\tau])\mu_\tau) \\
&\quad + \beta^{-1} \Delta \mu_\tau - \beta^{-1} \nabla \cdot (\nabla \log \rho_0 \mu_\tau).
\end{aligned}
$$

Because of the presence of the diffusive term $\beta^{-1}\Delta\mu_\tau$ in (A.15), we can therefore conclude that on the timescales where this equation holds we must have $\mu_\tau > 0$

a.e. on D. This means that (2.10) holds and so $V(\boldsymbol{\theta}, [\mu_\tau]) = 0$ since

$$\int_D \varphi(\cdot, \boldsymbol{\theta})\mu_\tau(d\theta) = f$$

by Proposition 3.5. As a result, (A.15) reduces to

$$(\text{A.16}) \quad \partial_\tau \mu_\tau = \nabla \cdot (\nabla F(\boldsymbol{\theta}, [\bar{\omega}_\tau])\mu_\tau) + \beta^{-1}\Delta\mu_\tau - \beta^{-1}\nabla \cdot (\nabla \log \rho_0 \mu_\tau).$$

Since $V(\boldsymbol{\theta}, [\mu_\tau]) = 0$ needs to be satisfied, in (A.16) we can treat the term involving the factor $F(\boldsymbol{\theta}, [\bar{\omega}_\tau])$ as a Lagrange multiplier used to enforce this constraint. It is also easy to see that (A.16) is the Wasserstein GD flow on

$$(\text{A.17}) \qquad \int_D \left(\beta^{-1}\log(d\mu/d\mu_0) + F(\boldsymbol{\theta}, [\bar{\omega}_\tau])\right)\mu(d\boldsymbol{\theta}).$$

Since this energy is strictly convex, a direct consequence of these observations is that the stable fixed points of (A.16) are the minimizers of the energy (A.17) subject to the constraints that $V(\boldsymbol{\theta}, [\mu_\tau]) = 0$ and $\mu_\tau \in \mathcal{M}_+(D)$. These fixed points are reached on a timescale that is large compared to the $O(n)$ timescale $\tau = t/n$.

Recalling that $\mu_{t,\tau}$ is the weak limit of $\mu_t^{(n)}$ as $n \to \infty$ and $V(\boldsymbol{\theta}, [\mu]) = -F(\boldsymbol{\theta}) + \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}')\mu(d\boldsymbol{\theta}')$, we can summarize these considerations into:

PROPOSITION A.3. *If $\mu_t^{(n)}$ is the empirical distribution that is defined in* (3.3) *with* $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$, *the solution to* (A.8). *Then given any $b_n > 0$ such that $b_n/n \to \infty$ as $n \to \infty$, we have*

$$(\text{A.18}) \qquad\qquad \mu_{b_n}^{(n)} \rightharpoonup \mu^* \quad \text{as } n \to \infty$$

where μ^* is the minimizer in $\mathcal{M}_+(D)$ of

$$(\text{A.19}) \qquad\qquad \beta^{-1}\int_D \log(d\mu/d\mu_0)\mu(d\boldsymbol{\theta})$$

subject to

$$(\text{A.20}) \qquad\qquad F(\boldsymbol{\theta}) = \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}')\mu^*(d\boldsymbol{\theta}') \quad \text{a.e. in } D.$$

It is easy to see that the solution to the minimization problem in Proposition A.3 is such that

$$(\text{A.21}) \qquad\qquad \int_D \log(d\mu^*/d\mu_0)\mu^*(d\boldsymbol{\theta}) < \infty \quad \text{and} \quad \text{supp}\,\mu^* = D.$$

The first condition says that the minimizer exists, which is clear since we can find test distributions $\mu \in \mathcal{M}_+(D)$ such that (i) $\int_{\mathbb{R}} c\mu(dc, \cdot) = \gamma^*$ where γ^* solves (2.6) (i.e., such that μ satisfies the constraint in (A.20)), and (ii) μ has finite entropy with respect to μ_0. One such μ is

$$(\text{A.22}) \quad \mu(dc, dz) = |\gamma^*|_{\text{TV}}^{-1}\left(\delta_{|\gamma^*|_{\text{TV}}}(dc)\gamma_+^*(dz) + \delta_{-|\gamma^*|_{\text{TV}}}(dc)\gamma_-^*(dz)\right).$$

To prove that $\text{supp}\,\mu^* = D$, suppose by contradiction that the minimizer is such that $\mu^* = 0$ if $\boldsymbol{\theta} \in B$ with $\int_B \mu_0(d\boldsymbol{\theta}) > 0$. For $s \in [0, 1]$, consider $\mu^s = (1 - s)\mu^* + s\mu_0$. A direct calculation shows that

$$\text{(A.23)} \qquad \int_D \log(d\mu^s/d\mu_0)\mu^s(d\boldsymbol{\theta}) = \int_B \log(d\mu^*/d\mu_0)\mu^*(d\boldsymbol{\theta})$$
$$+ s \log s \int_{B^c} \mu_0(d\boldsymbol{\theta}) + O(s).$$

Since $s \log s \int_{B^c} \mu_0(d\boldsymbol{\theta}) < 0$ for $s \in (0, 1)$, (A.23) implies that for $s > 0$ small enough

$$\text{(A.24)} \qquad \int_D \log(d\mu^s/d\mu_0)\mu^s(d\boldsymbol{\theta}) < \int_B \log(d\mu^*/d\mu_0)\mu^*(d\boldsymbol{\theta}),$$

a contradiction with our assumption that μ^* is the minimizer.

Let us also analyze in some more detail the constrained optimization problem in Proposition A.3 since this will be useful in the next section. If we denote by μ^* the minimizer of (A.19) subject to (A.20) and by λ^* the Lagrange multiplier used to satisfy the first constraint in (A.20), this Lagrange multiplier is given by

$$\text{(A.25)} \qquad \lambda^*(\boldsymbol{\theta}) = \beta^{-1} \frac{\delta}{\delta F(\boldsymbol{\theta})} \int_D \log(d\mu^*/d\mu_0)\mu^*(d\boldsymbol{\theta}').$$

It is easy to see that μ^* is independent of β: indeed, we can drop the factor β^{-1} in front of (A.19) without affecting the minimization problem. This also means that the dependency of λ^* in β is explicit: Indeed, from (A.25)

$$\text{(A.26)} \qquad \lambda^*(\boldsymbol{\theta}) = \beta^{-1}\delta^*(\boldsymbol{\theta})$$

where $\delta^*(\boldsymbol{\theta})$ is given by

$$\text{(A.27)} \qquad \delta^*(\boldsymbol{\theta}) = \frac{\delta}{\delta F(\boldsymbol{\theta})} \int_D \log(d\mu^*/d\mu_0)\mu^*(d\boldsymbol{\theta}').$$

This factor is independent of β since μ^* is. It will be useful later to work with the function $\epsilon^*(\boldsymbol{x})$ defined via the equation

$$\text{(A.28)} \qquad \int_D \varphi(\boldsymbol{x}, \boldsymbol{\theta})\delta^*(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_D \int_\Omega \varphi(\boldsymbol{x}, \boldsymbol{\theta})\varphi(\boldsymbol{x}', \boldsymbol{\theta})\epsilon^*(\boldsymbol{x}')\nu(d\boldsymbol{x}')d\boldsymbol{\theta}$$

This is the Euler-Lagrange equation for the minimizer of

$$\text{(A.29)} \qquad \frac{1}{2} \int_D \left| \delta^*(\boldsymbol{\theta}) - \int_\Omega \epsilon(\boldsymbol{x})\varphi(\boldsymbol{x}, \boldsymbol{\theta})\nu(d\boldsymbol{x}) \right|^2 d\boldsymbol{\theta}$$

over ϵ. Therefore, (A.28) is also the equation for the least squares solution of

$$\text{(A.30)} \qquad \delta^*(\boldsymbol{\theta}) = \int_\Omega \epsilon^*(\boldsymbol{x})\varphi(\boldsymbol{x}, \boldsymbol{\theta})\nu(d\boldsymbol{x})$$

and such a least squares solution exists for a modification of $\delta^*(\boldsymbol{\theta})$ that is arbitrarily close to it in $L^2(D)$: any such solution for a modification of $\delta^*(\boldsymbol{\theta})$ that is $O(n^{-1})$

away from it is good enough for our purpose since the discrepancy can be absorbed in higher-order terms in our expansion in n^{-1}. This solution is also unique by Assumption 2.2, and it can be expressed as

$$\text{(A.31)} \qquad \epsilon^*(x) = D_{f(x)} \int_D \log(d\mu^*/d\mu_0)\mu^*(d\theta)$$

where μ^* is viewed as a functional of $f(x)$ by using

$$F(\theta) = \int_\Omega f(x)\varphi(x,\theta)\nu(dx),$$

and $D_{f(x)}$ denotes the gradient with respect to $f(x)$ in the $L^2(\Omega,\nu)$-norm defined in (3.22). The equality (A.31) follows from (A.27) and the fact that $D_{f(x)}F(\theta) = \varphi(x,\theta)$.

Remark A.4. Compared to the case treated in [27] where the noise and regularizing terms in (A.2) are scaled as β^{-1} (high temperature) rather than $(\beta n)^{-1}$ (low temperature), we see that we can also conclude that μ_t converges as $t \to \infty$ to a distribution μ^* with supp $\mu^* = D$; however, the fixed point μ^* we obtain satisfies $\int_D \varphi(\cdot,\theta)\mu^*(d\theta) = f$, whereas the one obtained at high temperature introduces a correction proportional to β^{-1} in this relation. The price we pay by working at low temperature is that convergence in time may be slower if the initial condition $\mu_0 = \mu_{\text{in}}$ is such that (2.10) is not satisfied by the GD flow without noise: specifically, this convergence should occur on timescales that are intermediate between $O(1)$ and $O(n)$.

A.5 Central limit theorem at finite temperature

Now that we have determined the behavior of $\lim_{n\to\infty}\mu_t^{(n)} = \mu_t$ at all times, we can stop distinguishing τ from t, and focus on ω_t. We already know that (A.20) constrains the average value of ω_t on long timescales, but we would also like to quantify this average value beyond what (A.20) implies, and also analyze the fluctuations around this average. To this end, let us use (A.16) in (A.12) and look at the resulting equation on timescales where μ_t has converged to μ^*, the minimizer specified in Proposition A.3, so that $V(\theta,[\mu^*]) = 0$ and λ has converged to $\lambda^* = \beta^{-1}\delta^*$. This can be achieved by considering (A.12) with initial condition at $t = T$ and pushing back $T \to -\infty$. The resulting equation is

$$\text{(A.32)} \qquad \begin{aligned} \partial_t\omega_t &= \nabla \cdot \left(-\beta^{-1}\nabla\delta^*\mu^* + \int_D \nabla K(\theta,\theta')\omega_t(d\theta')\mu^*\right) \\ &\quad + \sqrt{2}\beta^{-1/2}\nabla \cdot \left(\sqrt{\mu^*}\dot{\eta}_t\right) \end{aligned}$$

Even though we derived it formally, the SPDE (A.38) can be given a precise meaning: since its drift is linear in ω_t and its noise is additive (recall that μ^* is a given, nonrandom distribution), (A.38) defines ω_t as a Gaussian process. This also means

that

$$(A.33) \qquad g_t = \int_D \varphi(\cdot, d\boldsymbol{\theta}) \omega_t(d\boldsymbol{\theta})$$

is a Gaussian process. This is an important quantity since gives the error on f made in $f_t^{(n)}$ at order $O(n^{-1})$:

$$(A.34) \qquad f_t^{(n)} = f + n^{-1} g_t + o(n^{-1})$$

Let us derive a closed equation for g_t from (A.32). To this end, notice first that we can use

$$(A.35) \qquad \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}') \omega_t(d\boldsymbol{\theta}') = \int_\Omega \varphi(\boldsymbol{x}, \boldsymbol{\theta}) g_t(\boldsymbol{x}) \nu(d\boldsymbol{x})$$

to express the integral terms in (A.32) in terms of g_t. By taking the time derivative of (A.33) and using (A.32) together with (A.35) and (A.28) we derive

$$
\begin{aligned}
\partial_t g_t &= \int_D \varphi(\cdot, \boldsymbol{\theta}) \partial_t \omega_t(d\boldsymbol{\theta}) \\
&= \int_D \nabla_{\boldsymbol{\theta}} \varphi(\cdot, \boldsymbol{\theta}) \\
&\qquad \cdot \int_\Omega \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{x}', \boldsymbol{\theta}) \left(g_t(\boldsymbol{x}') - \beta^{-1} \epsilon^*(\boldsymbol{x}') \right) \nu(d\boldsymbol{x}') \mu^*(d\boldsymbol{\theta}) \\
&\qquad - \sqrt{2} \beta^{-1/2} \int_D \nabla_{\boldsymbol{\theta}} \varphi(\cdot, \boldsymbol{\theta}) \cdot \sqrt{\mu^*} \dot{\eta}_t \\
&= \int_\Omega M([\mu^*], \boldsymbol{x}, \boldsymbol{x}') \left(g_t(\boldsymbol{x}') - \beta^{-1} \epsilon^*(\boldsymbol{x}') \right) \nu(d\boldsymbol{x}') \\
&\qquad - \sqrt{2} \beta^{-1/2} \int_D \nabla_{\boldsymbol{\theta}} \varphi(\cdot, \boldsymbol{\theta}) \cdot \sqrt{\mu^*} \dot{\eta}_t
\end{aligned}
$$

$$(A.36)$$

where $M([\mu], \boldsymbol{x}, \boldsymbol{x}')$ is the kernel defined in (3.18). Since the quadratic variation of the noise term in this equation is

$$(A.37) \qquad 2\beta^{-1} M([\mu^*], \boldsymbol{x}, \boldsymbol{x}') dt$$

in law, it is equivalent to

$$
\begin{aligned}
\partial_t g_t &= - \int_\Omega M([\mu^*], \boldsymbol{x}, \boldsymbol{x}') \left(g_t(\boldsymbol{x}') - \beta^{-1} \epsilon^*(\boldsymbol{x}) \right) \nu(d\boldsymbol{x}') \\
&\qquad + \sqrt{2} \beta^{-1/2} \int_\Omega \sigma([\mu^*], \boldsymbol{x}, \boldsymbol{x}') \dot{\eta}_t(\boldsymbol{x}') d\boldsymbol{x}'
\end{aligned}
$$

$$(A.38)$$

where $\dot{\eta}_t(\boldsymbol{x})$ is a spatiotemporal white noise, and $\sigma([\mu^*], \boldsymbol{x}, \boldsymbol{x}')$ is such that

$$(A.39) \qquad \int_\Omega \sigma([\mu^*], \boldsymbol{x}, \boldsymbol{x}'') \sigma([\mu^*], \boldsymbol{x}', \boldsymbol{x}'') d\boldsymbol{x}'' = M([\mu^*], \boldsymbol{x}, \boldsymbol{x}').$$

Note that this decomposition exists since $\mu^* \in \mathcal{M}_+(D)$ with supp $\mu^* = D$, and hence $M([\mu^*], x, x')$ is positive definite. The asymptotic mean and variance of g_t can be readily deduced from (A.38) by noting that this Ornstein-Uhlenbeck equation is in detailed balance with respect to the Gibbs distribution associated with the energy

$$(A.40) \qquad \frac{1}{2} \int_\Omega \left| g_t(x') - \beta^{-1} \epsilon^*(x) \right|^2 \nu(dx).$$

We can state this as follows:

PROPOSITION A.5 (CLT at finite temperature). *Let $f_t^{(n)}$ be given by* (3.4) *with* $\{\theta_i(t)\}_{i=1}^n$ *a solution to* (A.8) *with initial conditions specified at $t = T$. Then*

$$(A.41) \qquad \lim_{T \to -\infty} \lim_{n \to \infty} n\left(f_t^{(n)} - f\right) = g_t \quad \text{in law}$$

where g_t is the stationary Gaussian process specified by (A.38) *and whose mean and covariance satisfy*: *for any test function $\chi : \Omega \to \mathbb{R}$*

$$(A.42) \qquad \begin{aligned} &\mathbb{E} \int_\Omega \chi(x) g_t(x) \nu(dx) = \beta^{-1} \int_\Omega \chi(x) \epsilon^*(x) \nu(dx), \\ &\mathbb{E} \left(\int_\Omega \chi(x) \left(g_t(x) - \beta^{-1} \epsilon^*(x) \right) \nu(dx) \right)^2 = \beta^{-1} \int_\Omega |\chi(x)|^2 \nu(dx), \end{aligned}$$

where ϵ^* is given by (A.31)

Notice that if we quench the result in (A.42) (i.e., send $\beta \to \infty$), we arrive at the conclusion that $g_t \to 0$ as $t \to \infty$ in that case. This is consistent with what happens at zero temperature in the limit as $\xi \to 1$; see Proposition 3.9.

Bibliography

[1] Ambrosio, L.; Gigli, N.; Savaré, G. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 2nd ed., 2008.

[2] Auffinger, A.; Ben Arous, G. Complexity of random smooth functions on then high-dimensional sphere. *Ann. Probab.* **41** (2013), no. 6, 4214–4247. doi:10.1214/13-AOP862

[3] Auffinger, A.; Ben Arous, G.; Černý, J. Random matrices and complexity of spin glasses. *Comm. Pure Appl. Math.* **66** (2013), no. 2, 165–201. doi:10.1002/cpa.21422

[4] Bach, F. Breaking the curse of dimensionality with convex neutral networks. *J. Mach. Learn. Res.* **18** (2017), 19, 53.

[5] Baity-Jesi, M.; Sagun, L.; Geiger, M.; Spigler, S.; Arous, G. B.; Cammarota, C.; LeCun, Y.; Wyart, M.; Biroli, G. Comparing dynamics: deep neural networks versus glassy systems. *J. Stat. Mech. Theory Exp.* **2019** (2019), no. 12, 124013, 15. doi:10.1088/1742-5468/ab3281

[6] Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39** (1993), no. 3, 930–945. doi:10.1109/18.256500

[7] Beck, C.; E, W.; Jentzen, A. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *J. Nonlinear Sci.* **29** (2019), no. 4, 1563–1619. doi:10.1007/s00332-018-9525-3

[8] Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98** (2007), 146401. doi:10.1103/PhysRevLett.98.146401.

[9] Berg, J.; Nyström, K. A unified deep artificial neural network approach to partial differential equations in complex geometries. *Neurocomputing* **317** (2018), 28–41. `https://www.sciencedirect.com/science/article/pii/S092523121830794X`.

[10] Billingsley, P. *Convergence of probability measures.* Second edition. Wiley Series in Probability and Statistics: Probability and Statistics. A Wiley-Interscience Publication. Wiley, New York, 1999. doi:10.1002/9780470316962

[11] Bottou, L.; Le Cun, Y. On-line learning for very large data sets. *Appl. Stoch. Models Bus. Ind.* **21** (2005), no. 2, 137–151. doi:10.1002/asmb.538

[12] Chizat, L.; Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, vol. 31, edited by S. Bengio; H. Wallach; H. Larochelle; K. Grauman; N. Cesa-Bianchi; R. Garnett. Curran Associates, 2018.

[13] Choromanska, A.; Henaff, M.; Mathieu, M.; Ben Arous, G.; LeCun, Y. The loss surfaces of multilayer networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 38, edited by G. Lebanon; S. V. N. Vishwanathan. PMLR, San Diego, California, USA, 2015 pp. 192–204. `https://proceedings.mlr.press/v38/choromanska15.html`

[14] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** (1989), no. 4, 303–314. doi:10.1007/BF02551274

[15] Dawson, D. A.; Gärtner, J. Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics* **20** (1987), no. 4, 247–308. doi:10.1080/17442508708833446

[16] Dean, D. S. Langevin equation for the density of a system of interacting Langevin processes. *J. Phys. A* **29** (1996), no. 24, L613–L617. doi:10.1088/0305-4470/29/24/001

[17] E, W.; Han, J.; Jentzen, A. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Commun. Math. Stat.* **5** (2017), no. 4, 349–380. doi:10.1007/s40304-017-0117-6

[18] Gärtner, J. On the McKean-Vlasov limit for interacting diffusions. *Math. Nachr.* **137** (1988), 197–248. doi:10.1002/mana.19881370116

[19] Hu, W.; Li, C. J.; Li, L.; Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *Ann. Math. Sci. Appl.* **4** (2019), no. 1, 3–32. doi:10.4310/AMSA.2019.v4.n1.a1

[20] Jordan, R.; Kinderlehrer, D.; Otto, F. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.* **29** (1998), no. 1, 1–17. doi:10.1137/S0036141096303359

[21] Khoo, Y.; Lu, J.; Ying, L. Solving for high-dimensional committor functions using artificial neural networks. *Res. Math. Sci.* **6** (2019), no. 1, Paper No. 1. doi:10.1007/s40687-018-0160-2

[22] Kipnis, C.; Landim, C. *Scaling limits of interacting particle systems*. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 320. Springer, Berlin, 1999. doi:10.1007/978-3-662-03752-2

[23] Leblé, T.; Serfaty, S. Large deviation principle for empirical fields of log and Riesz gases. *Invent. Math.* **210** (2017), no. 3, 645–757. doi:10.1007/s00222-017-0738-0

[24] LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **521** (2015), no. 7553, 436–444.

[25] Li, Q.; Tai, C.; E, W. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *J. Mach. Learn. Res.* **20** (2019), Paper No. 40, 47.

[26] McKean, H. P., Jr. A class of Markov processes associated with nonlinear parabolic equations. *Proc. Nat. Acad. Sci. U.S.A.* **56** (1966), 1907–1911. doi:10.1073/pnas.56.6.1907

[27] Mei, S.; Montanari, A.; Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** (2018), no. 34, E7665–E7671. doi:10.1073/pnas.1806579115

[28] Park, J.; Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural Comput.* **3** (1991), no. 2, 246–257.

[29] Sagun, L.; Guney, V. U.; Arous, G. B.; LeCun, Y. Explorations on high dimensional landscapes. 2014. Preprint, `https://arxiv.org/abs/1412.6615`.

[30] Schneider, E.; Dai, L.; Topper, R. Q.; Drechsel-Grau, C.; Tuckerman, M. E. Stochastic neural network approach for learning high-dimensional free energy surfaces. *Phys. Rev. Lett.* **119** (2017), no. 15, 150601, 5. doi:10.1103/PhysRevLett.119.150601

[31] Serfaty, S. *Coulomb gases and Ginzburg-Landau vortices*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2015. doi:10.4171/152

[32] Serfaty, S. Systems of points with Coulomb interactions. *Eur. Math. Soc. Newsl.* (2018), no. 110, 16–21.

[33] Sirignano, J.; Spiliopoulos, K. Mean field analysis of neural networks: a law of large numbers. *SIAM J. Appl. Math.* **80** (2020), no. 2, 725–752. doi:10.1137/18M1192184

[34] Sznitman, A.-S. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*, *Lecture Notes in Math.*, vol. 1464, pp. 165–251, Springer, Berlin, 1991. doi:10.1007/BFb0085169.

[35] Villani, C. *Optimal transport*. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338. Springer, Berlin, 2009. doi:10.1007/978-3-540-71050-9

[36] Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. DeePCG: Constructing coarse-grained models via deep neural networks. *J. Chem. Phys.* **149** (2018), no. 3, 034101. doi:10.1063/1.5027645

GRANT ROTSKOFF
Chemistry Department
Stanford University
165 Keck Science Building
Stanford, CA 94305
USA
E-mail: `rotskoff@stanford.edu`

ERIC VANDEN-EIJNDEN
Courant Institute
251 Mercer St., Room 617
New York, NY 10012
USA
E-mail: `eve2@cims.nyu.edu`