NEURAL NETWORKS AS INTERACTING PARTICLE SYSTEMS: ASYMPTOTIC CONVEXITY OF THE LOSS LANDSCAPE AND UNIVERSAL SCALING OF THE APPROXIMATION ERROR

GRANT M. ROTSKOFF AND ERIC VANDEN-EIJNDEN

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, potentially of great use in computational and applied mathematics. That said, there are few rigorous results about the representation error and trainability of neural networks. Here we characterize both the error and the scaling of the error with the size of the network by reinterpreting the standard optimization algorithm used in machine learning applications, stochastic gradient descent, 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 parameters is large, the empirical distribution of the particles descends on a convex landscape towards a minimizer at a rate independent of n. We establish a Law of Large Numbers and a Central Limit Theorem for the empirical distribution, which together show that the approximation error of the network universally scales as $O(n^{-1})$. Remarkably, these properties do not depend on the dimensionality of the domain of the function that we seek to represent. Our analysis also quantifies the scale and nature of the noise introduced by stochastic gradient descent 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 network 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 $d = 25$.

CONTENTS

We thank Weinan E for discussions about the approximation error of neural networks and Sylvia Serfaty for her insights about interacting particle systems. We are also grateful to Andrea Montanari and Matthieu Wyart for their helpful comments.

1. Motivation and main results

While classification problems continue to be an active area research, extraordinary progress has been made on both speech and image recognition, problems that appeared intractable only a decade ago [19]. 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 represent high dimensional functions with accuracy.

These successes open exciting possibilities in applied and computational mathematics that are only beginning to be explored [5–7, 12, 16, 25, 30]. Any numerical calculation that uses a given function begins with a finite-dimensional representation 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, their effective dimensionality is much higher than the number of their parameters, which may explain their better capabilities observed in practice to approximate functions even when d is large. Characterizing this observation with analysis is non-trivial 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 neural networks are the "Universal Approximation Theorems" that specify the conditions under which a neural network can represent a target function with arbitrary accuracy [4, 10, 23]. These results do not, however, indicate how the network parameters should be determined to achieve maximal accuracy when their number is fixed [8]. Additionally, these theorems do not provide guidance on how the error scales with the number of parameters. Several recent papers have focused on the analysis of the shape and properties of the objective or "loss" function landscape [3, 9, 24]. These studies have mainly focused on the fine features of this landscape, trying to understand how non-convex 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 a more formal analysis. Similar to what was recently proposed in [22,29], 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 is thought of as the evolution of the particles in this interaction potential. We also consider the empirical distribution of the n interacting particles / parameters and analyze the properties of this distribution when the number n is large using standard limit theorems [17, 18, 26, 27]. This viewpoint allows us to bypass many of the difficulties that arise with approaches that attempt to study the dynamics of the individual particles. For example, we rederive the Universal Approximation Theorem as a corollary to the Law of Large Numbers (LLN) for the empirical distribution of the particles. We also establish that the loss landscape is asymptotically convex for large n in the space of the empirical distribution of the particles, and assert that convergence towards equilibrium of this distribution occurs on a time scale that is independent of n to leading order—similar results were obtained in [22, 29]. Finally, we prove a Central Limit Theorem for the empirical distribution, and thereby conclude that the approximation error of the function representation by a neural network is universal and scales as $O(n^{-1})$ as $n \to \infty$ in any d. These results are established first in situation where gradient descent (GD) is used to perform the training of 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 shed light on the nature of the noise introduced in SGD, and indicates how the time step and he batch size should be scaled to achieve the optimal error. Let us briefly elaborate on these statements next, starting with a precise formulation of the problem.

1.1. **Problem set-up.** Given a function $f : \Omega \to \mathbb{R}$ defined on $\Omega \subseteq \mathbb{R}^d$, consider its approximation by

$$f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i \varphi(\boldsymbol{x}, \boldsymbol{y}_i) \tag{1.1}$$

where $n \in \mathbb{N}$, $(c_i, \boldsymbol{y}_i) \in \mathbb{R} \times D$ with $D \subset \mathbb{R}^N$ are parameters to be learned for $i = 1, \ldots, n$, and $\varphi : \Omega \times D \to \mathbb{R}$ is some kernel—for simplicity we assume throughout this paper that D is compact. Many models used in machine learning can be cast in the form (1.1):

- **Radial basis function networks.** In this case $D \subset \Omega$ and $\varphi(\boldsymbol{x}, \boldsymbol{y}) \equiv \phi(\boldsymbol{x} - \boldsymbol{y})$ 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-layer neural networks.** In this case, $D \subset \mathbb{R}^{d+1}$ and $\varphi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x}, \boldsymbol{a}, b)$ with $\boldsymbol{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$, 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 $h(z) = 1/(1 + e^{-z})$.
- **Multi-layer neural networks.** These are essentially iterated versions of single-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
$$h^{(1)}_j(\boldsymbol{v}) = h(v_j), \qquad \boldsymbol{v} = (v_1, \ldots, v_m) \in \mathbb{R}^m, \quad j = 1, \ldots, m$$
then set
$$f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i h\left(\boldsymbol{a}^{(0)}_i \cdot \boldsymbol{h}^{(1)}\left(A^{(1)}_i \boldsymbol{x} + \boldsymbol{b}^{(1)}_i\right) + b^{(0)}_i\right)$$
where $\boldsymbol{a}^{(0)}_i \in \mathbb{R}^m$, $b^{(0)}_i \in \mathbb{R}$, $A^{(1)}_i \in \mathbb{R}^{m\times d}$, $\boldsymbol{b}^{(1)}_i \in \mathbb{R}^m$, $i = 1, \ldots, n$. Therefore here we have $\boldsymbol{y} = (\boldsymbol{a}^{(0)}, b^{(0)}, A^{(1)}, \boldsymbol{b}^{(1)}) \in D \subset \mathbb{R}^{m+1+m\times d+m}$ (where with a slight abuse of notation we view the matrix $A^{(1)}$ has a vector in $\mathbb{R}^{m\times d}$). Three-layer networks, etc. can be constructed

similarly. Note that our results apply to deep neural networks when the final layer grows large, not the total number of parameters.

In view of the growing range of applications of these methods, it is natural to ask:

(1) How good can the approximation (1.1) be if we optimize $\{(c_i, \boldsymbol{y}_i)\}_{i=1}^n$?
(2) Can we guarantee the convergence of the commonly used optimization algorithms?

To answer the first question, we need to introduce a distance, or loss function, between f and f_n. A natural candidate often used in practice is

$$\ell(f, f_n) = \tfrac{1}{2}\int_\Omega \left|f(\boldsymbol{x}) - f_n(\boldsymbol{x})\right|^2 d\mu(\boldsymbol{x}) \tag{1.2}$$

where μ is some measure on Ω with a positive density and such that $\mu(\Omega) < \infty$ (for example the Lebesgue measure, $d\mu(\boldsymbol{x}) = d\boldsymbol{x}$, if Ω is compact). This also offers a way to address the second question, since we can view $\ell(f, f_n)$ as an objective function for $\{(c_i, \boldsymbol{y}_i)\}_{i=1}^n$:

$$\ell(f, f_n) = C_f - \frac{1}{n}\sum_{i=1}^n c_i F(\boldsymbol{y}_i) + \frac{1}{2n^2}\sum_{i,j=1}^n c_i c_j K(\boldsymbol{y}_i, \boldsymbol{y}_j) \tag{1.3}$$

where $C_f = \frac{1}{2}\int_\Omega \left|f(\boldsymbol{x})\right|^2 d\mu(\boldsymbol{x})$ and we defined

$$F(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{x}), \qquad K(\boldsymbol{y}, \boldsymbol{z}) = \int_\Omega \varphi(\boldsymbol{x}, \boldsymbol{y})\varphi(\boldsymbol{x}, \boldsymbol{z}) d\mu(\boldsymbol{x}) \equiv K(\boldsymbol{z}, \boldsymbol{y}). \tag{1.4}$$

Trying to minimize (1.3) over $\{(c_i, \boldsymbol{y}_i)\}_{i=1}^n$ leads to difficulties, however, since this is potentially (and presumably) a non-convex optimization problem, which typically have local minimizers. As a result determining the distance (1.2) at the minimum (and its scaling with n, say) is also nontrivial. To bypass these difficulties we study the problem in terms of its empirical distribution.

1.2. **Functional formulation.** Assume that the set $\{(c_i, \boldsymbol{y}_i)\}_{i=1}^n$ is such that

$$f_n \to \tilde{f} = \int_D \varphi(\cdot, \boldsymbol{y}) G(\boldsymbol{y}) d\boldsymbol{y} \qquad \text{as} \quad n \to \infty \tag{1.5}$$

for some signed density G on D. If $\varphi(\boldsymbol{x}, \cdot)$ is smooth, a sufficient condition is that the empirical measure

$$d\gamma_n(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n c_i \delta_{\boldsymbol{y}_i}(d\boldsymbol{y}) \equiv \frac{1}{n}\sum_{i=1}^n c_i \delta(\boldsymbol{y} - \boldsymbol{y}_i) d\boldsymbol{y} \tag{1.6}$$

converges weakly to $G(\boldsymbol{y})d\boldsymbol{y}$ as $n \to \infty$. Clearly, this can be achieved, e.g. by drawing the $\boldsymbol{y}_i$'s independently from some probability density function $\bar{\rho}(\boldsymbol{y})$ such that $\bar{\rho}(\boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in D$ and $\int_D \bar{\rho}(\boldsymbol{y}) d\boldsymbol{y} = 1$, and setting $c_i = G(\boldsymbol{y}_j)/\bar{\rho}(\boldsymbol{y}_i)$: we then have $d\gamma_n(\boldsymbol{y}) \rightharpoonup G(\boldsymbol{y})d\boldsymbol{y}$ as $n \to \infty$ by the Law of Large Numbers, with an error scaling as $O(n^{-1/2})$ by the Central Limit Theorem. We show this scaling can be improved upon. In the large n limit, (1.3) converges to an objective function for G:

$$\ell(f, \tilde{f}) = C_f - \int_D F(\boldsymbol{y}) G(\boldsymbol{y}) d\boldsymbol{y} + \tfrac{1}{2}\int_{D\times D} K(\boldsymbol{y}, \boldsymbol{z}) G(\boldsymbol{y}) G(\boldsymbol{z}) d\boldsymbol{y} d\boldsymbol{z} \tag{1.7}$$

Unlike (1.3), this objective function is quadratic in G. This means that minimizing (1.7) over G rather than (1.3) over $\{(c_i, \boldsymbol{y}_i)\}_{i=1}^n$ is conceptually simpler than a direct minimization.

1.3. **Universal Representation Theorem.** To formalize these ideas, it is useful to view φ as a map from $L^2(D)$ to $L^2(\Omega, \mu)$ and introduce also its adjoint $\varphi^\dagger : L^2(\Omega, \mu) \to L^2(D)$. These operators are defined for any suitable $G \in L^2(D)$ and $g \in L^2(\Omega, \mu)$ respectively as

$$\varphi G = \int_D \varphi(\cdot, \boldsymbol{y}) G(\boldsymbol{y}) d\boldsymbol{y}, \qquad \varphi^\dagger g = \int_\Omega g(\boldsymbol{x}) \varphi(\boldsymbol{x}, \cdot) d\mu(\boldsymbol{x}) \tag{1.8}$$

We can then write the loss function in (1.2) evaluated on $f_n = \varphi G$ as

$$\ell(f, \varphi G) = \tfrac{1}{2}\|f - \varphi G\|^2_{L^2(\Omega,\mu)} \tag{1.9}$$

and the question becomes whether there exists a minimizer G^* of this objective function with $\varphi G^* = f$. Note that any such minimizer is also a solution to the Euler-Lagrange equation for both (1.9) and (1.7):

$$F = KG \qquad \text{explicitly:} \quad F(\boldsymbol{y}) = \int_D K(\boldsymbol{y}, \boldsymbol{y}')G(\boldsymbol{y}')d\boldsymbol{y}' \tag{1.10}$$

where F and K are given explicitly in (1.4). We can also express these operators as $F = \varphi^\dagger f$ and $K = \varphi^\dagger \varphi$: viewed as an operator mapping $L^2(D)$ into itself, K is symmetric and nonnegative.

This question of existence of solutions to (1.15) is within the realm of Fredholm theory of integral equations of the first kind [13]. For neural networks, it is natural to make:

Assumption 1.1. *The operator φ is bounded, i.e. $\exists c > 0 \ : \ \|\varphi G\|_{L^2(\Omega,\mu)} \le c\|G\|_{L^2(D)}$ $\forall G \in L^2(D)$, and square integrable:*

$$\int_D \int_\Omega |\varphi(\boldsymbol{x}, \boldsymbol{y})|^2 d\mu(\boldsymbol{x}) d\boldsymbol{y} < \infty \tag{1.11}$$

Indeed this assumption holds for the kernels used in machine learning, and it guarantees that both φ and $\varphi^\dagger$ are compact operators whose domains are $L^2(D)$ and $L^2(\Omega,\mu)$, respectively. Under Assumption 1.1, it is well known that (1.9) admits at least one minimizer (and (1.15) at least one solution) if and only if $f \in \operatorname{ran} \varphi \cup \operatorname{ran} \varphi^\dagger$ (where 'ran' denotes the range of the operator), which is a dense subspace of $L^2(\Omega,\mu)$. This solution/minimizer may not be unique, since we can add to it any element of null φ, which is not necessarily trivial. In the present context, this is no real issue, however, since we only care that φG gives f for any solution G of (1.15). This is why the kernels used in machine learning are designed so that they satisfy:

Assumption 1.2 (Discriminating Kernel)**.** *The null space of the adjoint $\varphi^\dagger : L^2(\Omega,\mu) \to L^2(D)$ is trivial, i.e.*

$$\varphi^\dagger g = 0 \quad \text{a.e. in } D \quad \Rightarrow \quad g = 0 \quad \text{a.e. in } \Omega \tag{1.12}$$

We can summarize the discussion above into:

Theorem 1.3 (Universal Representation Theorem)**.** *Under Assumptions 1.1 and 1.2, given any $f \in L^2(\Omega,\mu)$ and $\epsilon > 0$, there exists $f^* \in \operatorname{ran} \varphi \cup \operatorname{ran} \varphi^\dagger$ which is such that*

$$\|f - f^*\|_{L^2(\Omega,\mu)} \le \epsilon \tag{1.13}$$

and admits the representation

$$\varphi G^* = f^* \quad \text{a.e. in } \Omega \tag{1.14}$$

where G^* solves

$$F^* = KG^* \qquad \text{with } F^* = \varphi^\dagger f^* \tag{1.15}$$

The function f^* can also be realized as

$$f^* = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n c_j \varphi(\cdot, \boldsymbol{y}_j) \tag{1.16}$$

for some choice of $\{\boldsymbol{y}_i, c_i\}_{i\in\mathbb{N}}$.

Proof. By Assumption 1.1, (1.15) admits at least one solution G^* if $F^* = \varphi^\dagger f^*$ and $f^* \in \operatorname{ran} \varphi \cup \operatorname{ran} \varphi^\dagger$. Since this subspace is dense in $L^2(\Omega,\mu)$, (1.13) can be satisfied. By writing (1.15) as

$$0 = \varphi^\dagger (f^* - \varphi G^*) \tag{1.17}$$

we see that (1.14) follows by Assumption 1.2. To show that there is a choice of $\{\boldsymbol{y}_i, c_i\}_{i\in\mathbb{N}}$ such that (1.16) holds pick, for example, the $\boldsymbol{y}_i$ independently from some probability density function $\bar\rho(\boldsymbol{y})$ such that $\bar\rho(\boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in D$ and $\int_D \bar\rho(\boldsymbol{y}) d\boldsymbol{y} = 1$, and set $c_i = G^*(\boldsymbol{y}_i)/\bar\rho(\boldsymbol{y}_i)$. (1.16) then follows by the Law of Large Numbers. □

Note that the Universal Representation Theorem only gives f^*, not f. This suffices in our case since we cannot manipulate (1.15) directly, but rather need to go back to the representation (1.6). Using f^* introduces errors due to the finiteness of n which, even if they decrease as $n \to \infty$, remain larger than those induced by replacing f by f^* if we pick ϵ small enough. In other words, we can bound $\ell(f, f_n) \le \ell(f, f^*) + \ell(f^*, f_n)$ and make $\ell(f, f^*)$ smaller than $\ell(f^*, f_n)$. For this reason, in what follows we do not distinguish f from f^*. What remains to be shown, however, is that the Universal Representation Theorem can be approximately realized in practice via dynamic training of the parameters in (1.1) (recall that we have no direct access to G^*). In addition we seek to assess quantitatively the rate of convergence in time toward this approximation and the error due to the finiteness of n. We achieve this aim by treating the parameters $\{\boldsymbol{y}_i, c_i\}_{i=1}^n$ as a set of interacting particles.

Remark 1.1. Depending on the situation, it may be useful to consider other objective functions than (1.2). For example, in the context of elliptic problems, one is led to minimize

$$\int_\Omega \left(\tfrac{1}{2}\langle \nabla f, a(\boldsymbol{x})\nabla f\rangle - b(\boldsymbol{x}) f\right) d\boldsymbol{x} \tag{1.18}$$

where $a : \Omega \to \mathbb{R}^d \times \mathbb{R}^d$ is some positive-definite tensor and $b : \Omega \to \mathbb{R}$, and we impose e.g. $f = 0$ on $\partial\Omega$. This problem lends itself naturally to the present framework upon straightforward modifications, such as redefinition of the norm. For simplicity, in the present paper we only address the minimization of (1.2).

1.4. **Parameters as particles with loss function as interaction potential.** It is useful to view the set $\{\boldsymbol{y}_i, c_i\}_{i=1}^n$ as particles (with $\boldsymbol{y}_i$ viewed as a particle position and the weight c_i as its charge), and use (1.3) as an interaction potential between them. In this interpretation, we can perform the training by making these parameters evolve by gradient descent (GD) in this potential, or stochastic gradient descent (SGD)—the method of choice used in machine learning to train neural networks. If we denote these time-dependent parameters as $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ with $t \ge 0$, we study the way

$$f_n(t, \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n C_i(t)\varphi(\boldsymbol{x}, \boldsymbol{Y}_i(t)) \tag{1.19}$$

evolves in time, as well as the behavior of this function for large n. In particular, we establish a Law of Large Numbers (LLN) for $f_n(t)$ as well as a Central Limit Theorem (CLT), and thereby assess the error and trainability of neural networks representations.

1.5. **Discrete training set and stochastic gradient descent.** For most choices of the kernel φ commonly encountered in machine learning it is not possible to calculate (1.2) and (1.4) exactly. Rather, we approximate these integrals using a training set, i.e. a set of points $\{\boldsymbol{x}_p\}_{p=1}^P$ distributed according to μ, possibly independent, and over which f is known. We then approximate $\ell(f, f_n)$ by

$$\ell_P(f, f_n) = \frac{1}{P}\sum_{p=1}^P |f(\boldsymbol{x}_p) - f_n(\boldsymbol{x}_p)|^2 \tag{1.20}$$

and F and/or K by

$$F_P(\boldsymbol{y}) = \frac{1}{P}\sum_{i=1}^P f(\boldsymbol{x}_p)\varphi(\boldsymbol{x}_p, \boldsymbol{y}), \qquad K_P(\boldsymbol{y}, \boldsymbol{z}) = \frac{1}{P}\sum_{p=1}^P \varphi(\boldsymbol{x}_p, \boldsymbol{y})\varphi(\boldsymbol{x}_p, \boldsymbol{z}). \tag{1.21}$$

These approximations are precisely what SGD relies upon to calculate the gradient of the interaction potential / loss function, and we assess the errors they introduce.

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. In this case, in the limit as the updating time step Δt used in SGD tends to zero, SGD becomes asymptotically equivalent to an SDE whose drift terms coincide with those of GD but with multiplicative noise terms added.

1.6. **Universal scaling error of neural networks.** The main results we obtain can be summarized as follows: the function (1.19) evolves by GD in Wasserstein metric on a convex landscape (quadratic at leading order in n) and is such that

$$f_n(t) = f_0(t) + O(n^{-1}) \quad \text{with} \quad f_0(t) \to f \quad \text{as} \quad t \to \infty \tag{1.22}$$

with a convergence rate in time that is independent of n to leading order as $n \to \infty$. The convergence (1.22) holds for initial conditions satisfying specific conditions outlined in Sec. 2.3. This scaling also holds generically for SGD if we choose the size P of the batch used in (1.21) such that as $P = O(n^2)$. If we set $P = O(n^{2\alpha})$ with $\alpha \in (0,1)$, we lose accuracy and (1.22) is replaced by

$$f_n(t) = f_0(t) + O(n^{-\alpha}) \quad \text{with} \quad f_0(t) \to f \quad \text{as} \quad t \to \infty \tag{1.23}$$

If we set $P = O(n^{2\alpha})$ with $\alpha > 1$, there is no gain and we get (1.22) back. These results are stated in Proposition 2.6 in the context of GD and in Proposition 3.2 in the context of SGD. In Appendix A we also establish a finite-temperature variant (Langevin dynamics) of (1.22) which applies when additive noise terms are added in the evolution equation for the particles / parameters. This result reads

$$f_n(t) = f_0(t) + n^{-1} f_1(t) + o(n^{-1}) \quad \text{with} \quad \lim_{t\to\infty} f_0(t) = f \quad \text{and} \quad \lim_{t\to\infty} f_1(t) = \beta^{-1}\epsilon^* + \beta^{-1/2}\tilde{\epsilon} \tag{1.24}$$

where $\beta > 0$ is a parameter playing the role of inverse temperature, $\epsilon^* : \Omega \to \mathbb{R}$ is some given (non-random) function and $\tilde{\epsilon} : \Omega \to \mathbb{R}$ is a white-noise process, i.e. a Gaussian random field with zero mean and covariance $\propto \delta(\boldsymbol{x} - \boldsymbol{x}')$. Note that (1.24) gives (1.22) back after quenching (i.e. by sending $\beta \to \infty$). The result in (1.24) is stated in Proposition A.3

1.7. **Style and organization.** As is apparent from the discussion above, our approach has strong ties with the statistical mechanics of systems of many interacting particles. This is an active area of research in which rigorous results have been obtained recently, primarily in the context of Coulomb or Riesz interaction potentials. These potentials lead to kernels that are not compact operators. The situation we consider here is therefore different, and simpler in some technical ways. The main additional issues we face are that (i) our kernels are degenerate, i.e. φ is not injective in general, and (ii) the weights / charges are not fixed, but rather evolve alongside the particles.

Despite these difficulties, we believe that providing rigorous proof to each of our statements can be achieved using the mathematical apparatuses developed in the context of interacting particle systems—to a certain extent, this program was already started in [22, 29]. In this paper we adopt a semi-rigorous presentation and rely on formal asymptotic arguments to derive our results. This has the advantage of making the developments easier to follow. We also exploit the specific structure of the interaction potentials arising in the context of neural networks. The structure of the class of problems we study alleviates certain difficulties arising in standard interacting particle systems.

The remainder of this paper is organized as follows:

In Sec. 2 we study a model system of interacting particles undergoing gradient descent on the network loss function—this dynamics is similar to stochastic gradient descent, which is used in machine learning to train networks, but is more readily amenable to analysis. The equations for the empirical distribution of the particles are derived in Sec. 2.1. In Sec. 2.2 we study the limiting behavior of the empirical distribution and analyze the scaling of the fluctuations around this limit. The equations derived this way are then used in Sec. 2.3 to establish a Law of Large Numbers and in Sec. 2.4 a Central Limit Theorem. These results have direct implications in terms of the approximation error of the function representation by the neural network and its scaling with the number of particles, as well as the dynamics of training.

We then study stochastic gradient descent in Sec. 3, where we revisit all our results in a more practical context: In Sec. 3.1 we derive the stochastic differential equation (SDE) to which SGD is asymptotically equivalent for small time steps. The SDEs we obtain have multiplicative noise terms added to the drift terms in the GD equations studied in Sec. 2. The limiting behavior of the empirical distribution and the scaling of the fluctuations around it are then analyzed in Sec. 3.4 where we rederive a LLN and a CLT in the context of SGD, and discuss their implications in terms of approximation error and trainability.

These results are illustrated in Sec. 4, where we use a spherical p-spin model with $p = 3$ as test complex 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 Secs. 2 and 3. These results are obtained using both a radial basis function network, and a single-layer network using sigmoid functions.

Some concluding remarks are made in Sec. 5 and the situation where the optimization is done with GD at finite temperature is treated in Appendix A.

2. Interacting particles with adaptive fractional charges for training

We define an idealized set of dynamical equations for $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ that can be used to train the network by updating its parameters dynamically, and we analyze these equations as $n \to \infty$ and $t \to \infty$. Specifically, we assume that $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ satisfy the following system of ordinary differential equations (ODEs):

$$\begin{cases} \dot{\boldsymbol{Y}}_i = C_i \nabla F(\boldsymbol{Y}_i) - \dfrac{1}{n}\displaystyle\sum_{j=1}^n C_i C_j \nabla K(\boldsymbol{Y}_i, \boldsymbol{Y}_j), \\ \dot{C}_i = F(\boldsymbol{Y}_i) - \dfrac{1}{n}\displaystyle\sum_{j=1}^n C_j K(\boldsymbol{Y}_i, \boldsymbol{Y}_j) \end{cases} \tag{2.1}$$

for $i = 1, \dots, n$. As we show in Sec. 3, (2.1) share many properties with the stochastic gradient descent, though in SGD a multiplicative noise term persists in the equations. In Appendix A we also consider a finite-temperature version of these equations which have additive noise terms. The ODEs in (2.1) are the gradient descent flow on the energy:

$$E(\boldsymbol{y}_1, c_1 \dots, \boldsymbol{y}_n, c_n) = nC_f - \sum_{i=1}^n c_i F(\boldsymbol{y}_i) + \frac{1}{2n}\sum_{i,j=1}^n c_i c_j K(\boldsymbol{y}_i, \boldsymbol{y}_j) \tag{2.2}$$

This energy is simply the loss function in (1.3) rescaled by n.

We consider (2.1) with initial conditions such that every pair in $\{\boldsymbol{Y}_i(0), C_i(0)\}_{i=1}^n$ is drawn independently from some probability density function $\rho_{\text{in}}(\boldsymbol{y}, c)$ such that:

Assumption 2.1. *The density ρ_{in} is smooth in both its arguments, and such that $\rho_{in} > 0$ in $D \times \mathbb{R}$ and $\int_{\mathbb{R}} c\rho_{in}(\cdot, c)dc \in L^2(D)$.*

We denote the measure for the infinite set $\{(\boldsymbol{Y}_i(0), C_i(0))\}_{i\in\mathbb{N}}$ constructed this way by $\mathbb{P}_{\text{in}}$. Initial conditions of this type are frequently used in practice.

In order to guarantee global existence and uniqueness of the solution to (2.1) we also make

Assumption 2.2. *The kernel $\varphi(\cdot, \boldsymbol{x})$ is a continuously differentiable function of $\boldsymbol{y}$ for all $\boldsymbol{x} \in \Omega$.*

This assumption guarantees that the functions F and K are continuously differentiable in their arguments, and that the energy E is continuously differentiable and coercive, i.e., for every $C \in \mathbb{R}$ the sub-level set

$$E_C = \left\{(\boldsymbol{y}_1, c_1, \dots, \boldsymbol{y}_n, c_n) \in (D \times \mathbb{R})^n : E(\boldsymbol{y}_1, c_1, \dots, \boldsymbol{y}_n, c_n) \le C\right\} \quad \text{is bounded.} \tag{2.3}$$

2.1. **Empirical distribution and McKean-Vlasov equations.** To proceed, we consider the empirical distribution

$$\rho_n(t,\boldsymbol{y},c) = \frac{1}{n}\sum_{i=1}^{n} \delta(c - C_i(t))\delta(\boldsymbol{y} - \boldsymbol{Y}_i(t)) \tag{2.4}$$

in terms of which we can express (1.19) as

$$f_n(t,\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} C_i(t)\varphi(\boldsymbol{x},\boldsymbol{Y}_i(t)) = \int_{D\times R} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_n(t,\boldsymbol{y},c)d\boldsymbol{y}dc \tag{2.5}$$

The empirical distribution (2.4) is useful to work with because it satisfies the McKean-Vlasov equation [21]

$$\begin{aligned}\partial_t\rho_n = \nabla\cdot&\left(-c\nabla F\rho_n + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\rho_n'\rho_n d\boldsymbol{y}'dc'\right)\\ &+\partial_c\left(-F\rho_n + \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_n'\rho_n d\boldsymbol{y}'dc'\right)\end{aligned} \tag{2.6}$$

where we used the shorthands $\rho_n = \rho_n(t,\boldsymbol{y},c)$ and $\rho_n' = \rho_n(t,\boldsymbol{y}',c')$.

When there is noise, (2.6) is often referred to as Dean's equation [11]. It should be viewed as a formal identity which is useful to analyze the properties of ρ_n as $n\to\infty$, as we do in Secs. 2.2, 2.3 and 2.4.

Derivation of Dean's equation (2.6). Let us take the time derivative of (2.4). By the chain rule:

$$\begin{aligned}\partial_t\rho_n(t,\boldsymbol{y},c) = &-\frac{1}{n}\sum_{i=1}^{n}\delta(c-C_i)\nabla\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\cdot\dot{\boldsymbol{Y}}_i\\ &-\frac{1}{n}\sum_{i=1}^{n}\partial_c\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\,\dot{C}_i\end{aligned} \tag{2.7}$$

Pulling the derivatives in front of the sums and using (2.1) we can write this equation as

$$\begin{aligned}\partial_t\rho_n(t,\boldsymbol{y},c) = &-\nabla\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i(t))\left(c\nabla F(\boldsymbol{y}) - \frac{1}{n}\sum_{j=1}^{n} cC_j\nabla K(\boldsymbol{y},\boldsymbol{Y}_j)\right)\right)\\ &-\partial_c\left(\frac{1}{n}\sum_{i=1}^{n}\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\left(F(\boldsymbol{y}) - \frac{1}{n}\sum_{j=1}^{n} C_jK(\boldsymbol{y},\boldsymbol{Y}_j)\right)\right)\end{aligned} \tag{2.8}$$

where we used the Dirac delta to replace $\boldsymbol{Y}_i$ by $\boldsymbol{y}$ and C_i by c. We can now use the definition of ρ_n to replace $\frac{1}{n}\sum_{i=1}^{n}\delta(c-C_i(t))\delta(\boldsymbol{y}-\boldsymbol{Y}_i(t))$ by $\rho_n(t,\boldsymbol{y},c)$. In addition, if we use

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n} cC_j\nabla K(\boldsymbol{y},\boldsymbol{Y}_j) &= \frac{1}{n}\sum_{j=1}^{n}\int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\delta(c'-C_j)\delta(\boldsymbol{y}'-\boldsymbol{Y}_j)d\boldsymbol{y}'dc'\\ &= \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\rho_n(t,c',\boldsymbol{y}')d\boldsymbol{y}'dc'\end{aligned} \tag{2.9}$$

and similarly for $\frac{1}{n}\sum_{j=1}^{n} C_jK(\boldsymbol{y},\boldsymbol{Y}_j)$, we see that we can write the right hand side of (2.8) precisely as that in (2.6).

2.2. **Limit behavior and fluctuations scaling.** Let us now use Dean's equation (2.6) to derive equations for the limit of ρ_n as $n\to\infty$ and for the fluctuations around this limit. All limits should be understood in the weak (or distributional) sense because in the end we care about $f_n(t)$, not $\rho_n(t)$ itself, and $f_n(t)$ is obtained by testing $\rho_n(t)$ against $c\varphi(\cdot,\boldsymbol{y})$, as in (2.5).

2.2.1. *Zeroth order term—mean field limit.* If we formally take the limit as $n \to \infty$ of (2.6), we deduce that $\rho_n(t) \rightharpoonup \rho_0(t)$, where $\rho_0(t)$ satisfies

$$\begin{aligned}\partial_t \rho_0 = \nabla \cdot \left(-c\nabla F \rho_0 + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\rho_0'\rho_0 d\boldsymbol{y}'dc' \right) \\ + \partial_c \left(-F\rho_0 + \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_0'\rho_0 d\boldsymbol{y}'dc' \right).\end{aligned} \tag{2.10}$$

Even though (2.10) is formally identical to (2.6), it has a different meaning: We can look for a solution of it for the smooth initial condition $\rho_0(0) = \rho_{\text{in}}$ since we know that $\mathbb{P}_{\text{in}}$-almost surely as $n \to \infty$, $\rho_n(0,\boldsymbol{y},c) \rightharpoonup \rho_{\text{in}}(\boldsymbol{y},c)$ by the Law of Large Numbers. However, to justify that $\rho_0(t)$, the solution of (2.10) for this initial condition, is indeed the weak limit of the empirical distribution $\rho_n(t)$ at later times, we have to control the fluctuations of $\rho_n(t)$ around $\rho_0(t)$. We do this next and postpone the analysis of (2.10) until Sec. 2.3.

Notice that (2.10) can be written as

$$\partial_t \rho_0 = \nabla \cdot \left(\rho_0 \nabla \frac{\delta \mathscr{E}_0}{\delta \rho_0} \right) + \partial_c \left(\rho_0 \partial_c \frac{\delta \mathscr{E}_0}{\delta \rho_0} \right) \tag{2.11}$$

where $\mathscr{E}_0[\rho_0]$ is given by:

$$\begin{aligned}\mathscr{E}_0[\rho_0] &= C_f - \int_{D\times\mathbb{R}} cF\rho_0 d\boldsymbol{y}dc + \tfrac{1}{2}\int_{(D\times\mathbb{R})^2} cc'K(\boldsymbol{y},\boldsymbol{y}')\rho_0\rho_0' d\boldsymbol{y}dcd\boldsymbol{y}'dc' \\ &= \tfrac{1}{2}\int_\Omega \left(f(\boldsymbol{x}) - \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_0 d\boldsymbol{y}dc \right)^2 d\mu(\boldsymbol{x}) \geq 0\end{aligned} \tag{2.12}$$

In (2.11) we can also write

$$\frac{\delta \mathscr{E}_0}{\delta \rho_0} = cF_0(t) - cF \tag{2.13}$$

where we defined

$$F_0(t,\boldsymbol{y}) = \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_0(t,\boldsymbol{y}',c')d\boldsymbol{y}'dc' \tag{2.14}$$

The energy in (2.12) is the continuous limit of (2.2) scaled by n^{-1}, and (2.11) is the gradient decent flow on this energy in the Wasserstein metric [15,28].

2.2.2. *Fluctuations around the mean.* Let us now consider the fluctuations of $\rho_n(t)$ around $\rho_0(t)$. The scale of these fluctuations change with time and to account for this effect, we define $\tilde{\rho}_\xi(t)$ via:

$$\rho_n = \rho_0 + n^{-\xi(t)}\tilde{\rho}_{\xi(t)}, \tag{2.15}$$

where the exponent $\xi(t)$ depends on t as specified below. Explicitly, (2.15) means:

$$\tilde{\rho}_{\xi(t)}(t,\boldsymbol{y},c) = n^{\xi(t)-1}\sum_{i=1}^n \big(\delta(\boldsymbol{y}-\boldsymbol{Y}_i(t))\delta(c-C_i(t)) - \rho_0(t,\boldsymbol{y},c)\big) \tag{2.16}$$

As we show next, the function $\xi(t)$ must initially be $\xi(0) = 1/2$, the scale set by the Central Limit Theorem applied to the initial parameter selection. However, subsequent annealing of the error will allow us to change $\xi(t)$ so that $\lim_{t\to\infty}\xi(t) = 1$.

To see that choosing $\xi(0) = 1/2$ sets the right scale to look at the fluctuations around the initial conditions, notice that if we pick a test function $\chi : D \times \mathbb{R} \to \mathbb{R}$ the CLT tells us that under $\mathbb{P}_{\text{in}}$

$$\int_{D\times\mathbb{R}} \chi(\boldsymbol{y},c)\tilde{\rho}_{\xi(0)}(0,\boldsymbol{y},c)d\boldsymbol{y}dc = n^{-1/2}\sum_{i=1}^n \tilde{\chi}(\boldsymbol{Y}_i(0),C_i(0)) \to N(0,C_\chi) \quad \text{in law as } n\to\infty \tag{2.17}$$

where $N(0,C_\chi)$ denotes the Gaussian random variable with mean zero and variance C_χ, and we defined

$$\tilde{\chi}(\boldsymbol{y},c) = \chi(\boldsymbol{y},c) - \int_{D\times\mathbb{R}} \chi(\boldsymbol{y},c)\rho_{\text{in}}(\boldsymbol{y},c)d\boldsymbol{y}dc, \qquad C_\chi = \int_{D\times\mathbb{R}} |\tilde{\chi}(\boldsymbol{y},c)|^2 \rho_{\text{in}}(\boldsymbol{y},c)d\boldsymbol{y}dc, \tag{2.18}$$

We can write (2.17) distributionally as

$$\tilde\rho_{\xi(0)}(0,\boldsymbol{y},c) \rightharpoonup N\big(0,\rho_{\text{in}}(\boldsymbol{y},c)\delta(\boldsymbol{y}-\boldsymbol{y}')\delta(c-c')\big) \quad \text{in law as } \ n\to\infty \tag{2.19}$$

To see what happens at later times, we derive an equation for $\tilde\rho_{\xi(t)}$ by subtracting (2.10) from (2.6) and using (2.15)

$$\begin{aligned}\partial_t\tilde\rho_{\xi(t)} = \nabla\cdot\Big(&-c\nabla F\tilde\rho_{\xi(t)} + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\Big(\tilde\rho'_{\xi(t)}\rho_0 + \rho'_0\tilde\rho_{\xi(t)} + n^{-\xi(t)}\tilde\rho'_{\xi(t)}\tilde\rho_{\xi(t)}\Big)\,d\boldsymbol{y}'dc'\Big)\\ &+\partial_c\Big(-F\tilde\rho_{\xi(t)} + \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\Big(\tilde\rho'_{\xi(t)}\rho_0 + \rho'_0\tilde\rho_{\xi(t)} + n^{-\xi(t)}\tilde\rho'_{\xi(t)}\tilde\rho_{\xi}\Big)\,d\boldsymbol{y}'dc'\Big)\\ &+\dot\xi(t)\log n\,\tilde\rho_{\xi(t)}\end{aligned} \tag{2.20}$$

In order to take the limit as $n\to\infty$ of this equation, we need to consider carefully the behavior of the the factors in (2.20) that contain n explicitly, that is, $n^{-\xi(t)}\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}$ and $\dot\xi(t)\log n\,\tilde\rho_{\xi(t)}$. Consider the latter first. If we set

$$\dot\xi(t)\log n = o(1) \tag{2.21}$$

the last term at the right hand side of (2.20) is higher order. Note that (2.21) means that we can vary $\xi(t)$, but only slowly. For the factor $n^{-\xi(t)}\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}$, a direct calculation shows that, for any $p\in\mathbb{N}$ and $\xi\in\mathbb{R}$,

$$\mathbb{E}_{\text{in}}\left(n^{-\xi}\int_{(D\times\mathbb{R})^2}\chi(\boldsymbol{y},c)\chi(\boldsymbol{y}',c')\tilde\rho_\xi\tilde\rho'_\xi\,d\boldsymbol{y}dcd\boldsymbol{y}'dc'\right)^p = O\left(n^{(\xi-1)p}\right) \tag{2.22}$$

where $\mathbb{E}_{\text{in}}$ denotes expectation with respect to $\mathbb{P}_{\text{in}}$—for example, if $p=1$, this expectation is $n^{(\xi-1)}C_\chi$ where C_χ is given in (2.18). Equation (2.22) implies that $\mathbb{P}_{\text{in}}$-almost surely as $n\to\infty$, $n^{-\xi}\tilde\rho_n(0)\tilde\rho'_n(0)\rightharpoonup 0$ at $t=0$ if $\xi<1$. We can now argue that this statement also holds at times $t>0$ if (2.21) holds. To this end we write (2.20) compactly as

$$\partial_t\tilde\rho_{\xi(t)} = L\tilde\rho_{\xi(t)} + R_{\xi(t)} + \dot\xi(t)\log n\,\tilde\rho_{\xi(t)} \tag{2.23}$$

where $L\tilde\rho_n$ contains the terms at the right hand side of (2.20) that are linear in $\tilde\rho_{\xi(t)}$, and $R_{\xi(t)}$ contains the terms involving $n^{-\xi(t)}\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}$. In order to control the $R_{\xi(t)}$ term, we can write an equation for $n^{-\xi(t)}\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}$: this equation is of the form (2.23) with an additional linear term involving L' (same as L but acting on $(\boldsymbol{y}',c')$), the source term $R_{\xi(t)}$ replaced by one involving $n^{-2\xi(t)}\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}\tilde\rho''_{\xi(t)}$, and the last term in (2.23) replaced by $2\dot\xi(t)\log n\,\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}$: a calculation similar to the one that gives (2.22) indicates that at $t=0$ this source term is higher order than the rest and goes to zero $\mathbb{P}_{\text{in}}$-almost surely as $n\to\infty$. The same is true for $2\dot\xi(t)\log n\,\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}$ if (2.21) holds. We can then derive equations for $n^{-2\chi(t)}\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}\tilde\rho''_{\xi(t)}$ and so on, and each time reach the same conclusion: they involve a linear part made of operators L, L', etc. and a remainder that is higher order. The specific form of the operator L appearing in (2.23) is given by its action on any function $g: D\times\mathbb{R}\to\mathbb{R}$

$$\begin{aligned}Lg(\boldsymbol{y},c) = \nabla\cdot\left(\rho_0(t,\boldsymbol{y},c)\nabla\frac{\delta\mathscr{E}_0[g]}{\delta g}\right) + \partial_c\cdot\left(\rho_0(t,\boldsymbol{y},c)\partial_c\frac{\delta\mathscr{E}_0[g]}{\delta g}\right)\\ +\nabla\cdot\big(\nabla\big(F_0(t,\boldsymbol{y})-F(\boldsymbol{y})\big)cg\big) + \partial_c\big(\big(F_0(t,\boldsymbol{y})-F(\boldsymbol{y})\big)g\big)\end{aligned} \tag{2.24}$$

where $\mathscr{E}_0$ is defined in (2.13) and $F_0(t)$ in (2.14). The gradient part in this operator implies that it is dissipative, which in turns means that in (2.23) the linear term $L\tilde\rho_{\xi(t)}$ damps the effect of the source $R_{\xi(t)}$ and of $\dot\xi(t)\log n\,\tilde\rho_{\xi(t)}$. Since this source term itself satisfies an equation where the linear term damps the effect the source, and so on, we can formally conclude that each of the terms $n^{-\xi(t)}\tilde\rho_{\xi(t)}\rho'_{\xi(t)}$, $n^{-2\xi(t)}\tilde\rho_{\xi(t)}\tilde\rho'_{\xi(t)}\tilde\rho''_{\xi(t)}$, etc. remains at all times $t>0$ on the scale set by $\mathbb{P}_{\text{in}}$.

This argument implies that, as long as (2.21) holds and $\xi(t)<1$ at all times, $\tilde\rho_{\xi(t)}$ has a limit as $n\to\infty$. If we take this limit at any fixed time, then (2.21) implies that $\xi(t)=\xi(0)=\frac{1}{2}$ as $n\to\infty$, that is, we have remained on the original scale set by $\mathbb{P}_{\text{in}}$. On that scale, we have $\tilde\rho_{1/2}(t) \rightharpoonup \rho_{1/2}(t)$ as $n\to\infty$, where $\rho_{1/2}(t)$ solves

$$\begin{aligned}\partial_t\rho_{1/2} = \nabla\cdot&\Big(-c\nabla F\rho_{1/2} + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\big(\rho'_{1/2}\rho_0+\rho'_0\rho_{1/2}\big)\,d\boldsymbol{y}'dc'\Big)\\ &+\partial_c\Big(-F\rho_{1/2}+\int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\big(\rho'_{1/2}\rho_0+\rho'_0\rho_{1/2}\big)\,d\boldsymbol{y}'dc'\Big)\end{aligned}\tag{2.25}$$

this equation should be solved with the Gaussian initial conditions read from (2.19):

$$\rho_{1/2}(0,\boldsymbol{y},c) = N\big(0,\rho_{\text{in}}(\boldsymbol{y},c)\delta(\boldsymbol{y}-\boldsymbol{y}')\delta(c-c')\big).\tag{2.26}$$

Note that since the mean of $\rho_{1/2}$ is zero initially and (2.25) is linear, this mean remains zero for all times, and we can focus in the evolution of its covariance. Denoting this covariance by

$$\omega_{1/2}(t,\boldsymbol{y},c,\boldsymbol{y}',c') = \mathbb{E}_{\text{in}}\rho_{1/2}(t,\boldsymbol{y},c)\rho_{1/2}(t,\boldsymbol{y}',c'),\tag{2.27}$$

from (2.25) it satisfies

$$\begin{aligned}\partial_t\omega_{1/2} = \nabla\cdot&\Big(-c\nabla F\omega_{1/2}+\int_{D\times\mathbb{R}} cc''\nabla K(\boldsymbol{y},\boldsymbol{y}'')\big(\omega_{1/2}\rho_0+\omega_{1/2}\rho''_0\big)\,d\boldsymbol{y}''dc''\Big)\\ &+\partial_c\Big(-F\omega_{1/2}+\int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}'')\big(\omega_{1/2}\rho_0+\omega_{1/2}\rho''_0\big)\,d\boldsymbol{y}''dc''\Big)\\ &+\nabla'\cdot\Big(-c'\nabla'F'\omega_{1/2}+\int_{D\times\mathbb{R}} c'c''\nabla K(\boldsymbol{y},\boldsymbol{y}'')\big(\omega_{1/2}\rho'_0+\omega_{1/2}\rho''_0\big)\,d\boldsymbol{y}''dc''\Big)\\ &+\partial_{c'}\Big(-F'\omega_{1/2}+\int_{D\times\mathbb{R}} c''K(\boldsymbol{y}',\boldsymbol{y}'')\big(\omega_{1/2}\rho'_0+\omega_{1/2}\rho''_0\big)\,d\boldsymbol{y}''dc''\Big)\end{aligned}\tag{2.28}$$

where we use the shorthands $\omega_{1/2}\rho_0 = \omega_{1/2}(t,\boldsymbol{y}',c',\boldsymbol{y}'',c'')\rho_0(t,\boldsymbol{y},c)$, $\omega_{1/2}\rho'_0 = \omega_{1/2}(t,\boldsymbol{y}'',c'',\boldsymbol{y},c)\rho_0(t,\boldsymbol{y}',c')$, and $\omega_{1/2}\rho''_0 = \omega_{1/2}(t,\boldsymbol{y},c,\boldsymbol{y}',c')\rho_0(t,\boldsymbol{y}'',c'')$. The initial condition for (2.28) is

$$\omega_{1/2}(0,\boldsymbol{y},c,\boldsymbol{y}',c') = \rho_{\text{in}}(\boldsymbol{y},c)\delta(\boldsymbol{y}-\boldsymbol{y}')\delta(c-c').\tag{2.29}$$

The existence of the weak limit of $\tilde\rho_{1/2}$ even with $\xi(t)=\xi(0)=\frac12$ is enough to confirm that $\rho_n(t)\rightharpoonup\rho_0(t)$ as $n\to\infty$, where $\rho_0(t)$ solves (2.10). However, we would like to consider different values of $\xi(t)$ to get a better handle on the size of the fluctuations at long times. In practice this can be done by choosing, for example,

$$\xi(t)=\bar\xi(t/a_n)\qquad\text{with}\qquad \lim_{n\to\infty} a_n/\log n=\infty,\qquad \bar\xi(0)=\tfrac12,\qquad \bar\xi(s)<1\quad\forall s>0\tag{2.30}$$

so that $\dot\xi(t)=\bar\xi'(t/a_n)/a_n$ and satisfies (2.21). If we then set the time to be a_n+t, we can conclude that $\tilde\rho_{\xi(a_n+t)}(a_n+t)\rightharpoonup\rho_{\bar\xi}(t)$, where $\bar\xi=\bar\xi(1)$ and $\rho_{\bar\xi}(t)$ solves

$$\begin{aligned}\partial_t\rho_{\bar\xi} = \nabla\cdot&\Big(-c\nabla F\rho_{\bar\xi}+\int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\big(\rho'_{\bar\xi}\rho_0+\rho'_0\rho_{\bar\xi}\big)\,d\boldsymbol{y}'dc'\Big)\\ &+\partial_c\Big(-F\rho_{\bar\xi}+\int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\big(\rho'_{\bar\xi}\rho_0+\rho'_0\rho_{\bar\xi}\big)\,d\boldsymbol{y}'dc'\Big)\end{aligned}\tag{2.31}$$

where ρ_0 in this equation is understood as $\lim_{n\to\infty}\rho_0(a_n+t)$, assuming that this limit exists (we check in Sec. 2.3 that it does). Because of the way we have taken the limit to arrive at this equation, it is valid at times that are infinitely far in the future of the initial conditions. This changes the interpretation we need to give to (2.31): Since (2.31) is linear and homogeneous in $\rho_{\bar\xi}$, either zero is the stable fixed point of this equation, and it means that the size of the fluctuations at time a_n are bounded from above by $O(n^{-\bar\xi})$: $\tilde\rho_{\bar\xi}(a_n)\rightharpoonup 0$; or zero is an unstable fixed point of (2.31), and these fluctuations go to infinity even on the scale $O(n^{-1/2})$. In Sec. 2.4 we check that the former statement holds under some conditions. Note that in that case, it means we can take $\bar\xi$ as large as we want in $[1/2,1)$.

Summing up, we have obtained:

Proposition 2.3. *Let $\rho_n(t)$ be the empirical distribution in* (2.4) *in which $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ solve* (2.1) *with initial conditions drawn from $\mathbb{P}_{in}$. Then, as $n \to \infty$:*

$$\rho_n(t,\boldsymbol{y},c) \rightharpoonup \rho_0(t,\boldsymbol{y},c), \qquad \textit{almost surely} \tag{2.32}$$

where $\rho_0(t)$ solves (2.10)*, and*

$$n^{1/2}\left(\tilde\rho_{1/2}(t,\boldsymbol{y},c) - \rho_0(t,\boldsymbol{y},c)\right) \rightharpoonup \rho_{1/2}(t,\boldsymbol{y},c), \qquad \textit{in law} \tag{2.33}$$

where $\rho_{1/2}(t)$ is the zero-mean Gaussian process whose covariance $\omega_\xi(t)$ solves (2.28)*. In addition, if zero is the stable fixed point of* (2.31)*, then*

$$n^{\bar\xi}\left(\tilde\rho_{\bar\xi}(a_n,\boldsymbol{y},c) - \rho_0(a_n,\boldsymbol{y},c)\right) \rightharpoonup 0, \qquad \textit{almost surely} \tag{2.34}$$

for any $\bar\xi < 1$ and a_n such that $a_n/\log n \to \infty$ as $n \to \infty$.

Note that we have yet to check that the condition that leads to (2.34) can be satisfied: we do this in Sec. 2.4.

Remark 2.1. The argument that led us to conclude that in (2.20) the term $n^{-\xi}\tilde\rho_\xi\tilde\rho'_\xi$ converges to zero almost surely if $\xi < 1$ is related to the property of *propagation of molecular chaos.* This property holds if $\mathbb{P}_{\text{in}}$ being a product measure implies that its push-forward in time is also a product measure to leading order in n^{-1}, and it guarantees that $n^{-\xi}\tilde\rho_\xi\tilde\rho'_\xi = O(n^{\xi-1})$

Remark 2.2. In order that (2.16) have a limit at $t = 0$ for initial condition drawn from $\mathbb{P}_{\text{in}}$ it is required that $\xi(0) < \frac{1}{2}$. At the same time, $n^{-\xi}\tilde\rho_\xi\tilde\rho'_\xi \rightharpoonup 0$ as $n \to \infty$ only requires that $\xi < 1$. This is why we were allowed to vary $\xi(t)$ slowly with time in a way consistent with (2.30), and it shows that the fluctuations present in the initial data diminish as time progresses. We revisit the consequences of this property in Sec. 2.4.

2.3. **Law of Large Numbers (LLN).**

Let us now analyze the evolution equation (2.10) for the weak limit ρ_0 of ρ_n. We begin with a discussion about the stationary points of this equation.

2.3.1. *Stationary points of* (2.10).

This equation for $\rho_0(t)$ is the GD flow in Wasserstein metric over the energy $\mathscr{E}_0[\rho_0]$, and it has more stable fixed points than $\mathscr{E}_0[\rho_0]$ has minimizers. To see why, write (2.10) as

$$\partial_t\rho_0 = \nabla\cdot\left(\rho_0 c\nabla U(t,\boldsymbol{y})\right) + \partial_c\left(\rho_0 U(t,\boldsymbol{y})\right) \tag{2.35}$$

where we defined

$$\begin{aligned} U(t,\boldsymbol{y}) &= -F(\boldsymbol{y}) + \int_{D\times\mathbb{R}} c' K(\boldsymbol{y},\boldsymbol{y}')\rho_0(t,\boldsymbol{y}',c')d\boldsymbol{y}'dc' \\ &= \int_\Omega \left(f_0(t,\boldsymbol{x}) - f(\boldsymbol{x})\right)\varphi(\boldsymbol{x},\boldsymbol{y})d\mu(\boldsymbol{x}) \end{aligned} \tag{2.36}$$

If we introduce the characteristics equations

$$\begin{cases} \dot{\boldsymbol{Y}}(t,\boldsymbol{y},c) = -C(t,\boldsymbol{y},c)\nabla U(t,\boldsymbol{Y}(t,\boldsymbol{y},c)), & \boldsymbol{Y}(0,\boldsymbol{y},c) = \boldsymbol{y} \\ \dot{C}(t,\boldsymbol{y},c) = -U(t,\boldsymbol{Y}(t,\boldsymbol{y},c)), & C(0,\boldsymbol{y},c) = c \end{cases} \tag{2.37}$$

we can turn (2.10) for the initial condition $\rho_0(0) = \rho_{\text{in}}$ into

$$\rho_0(t,\boldsymbol{y},c) = \rho_{\text{in}}\left(\boldsymbol{Y}(-t,\boldsymbol{y},c), C(-t,\boldsymbol{y},c)\right)\exp\left(\int_0^t C(s-t,\boldsymbol{y},c)\Delta U(s,\boldsymbol{Y}(s-t,\boldsymbol{y},c))ds\right) \tag{2.38}$$

This representation formula is readily verified by taking the time derivative of $\rho_0(t,\boldsymbol{Y}(t),C(t))$ and using the property $\boldsymbol{Y}(t,\boldsymbol{Y}(s,\boldsymbol{y},c),C(s,\boldsymbol{y},c)) = \boldsymbol{Y}(t+s,\boldsymbol{y},c)$ and $C(t,\boldsymbol{Y}(s,\boldsymbol{y},c),C(s,\boldsymbol{y},c)) = C(t+s,\boldsymbol{y},c)$ as well as (2.35) and (2.37). (2.38) is not explicit, since $U(t)$ depends on $\rho_0(t)$ through (2.36), but it can be used to deduce some properties of $\rho_0(t)$. For example, this formula shows that $\rho_0(t) > 0$ for all times $t > 0$ if $\rho_{\text{in}} > 0$, though we cannot guarantee that this property holds as $t \to \infty$. Indeed, the existence of a fixed point for the second equation in (2.37), which is guaranteed by the gradient nature of the dynamics in (2.35), implies that (see Remark 2.3 for an alternative derivation of this equation and (2.41))

$$\lim_{t\to\infty} U(t,\boldsymbol{Y}(t,\boldsymbol{y},c)) = 0 \qquad \text{for almost all } (\boldsymbol{y},c)\in D\times\mathbb{R} \tag{2.39}$$

To interpret this equation, we introduce the set

$$D_0 = \Big\{\boldsymbol{y}^* = \lim_{t\to\infty}\boldsymbol{Y}(t,\boldsymbol{y},c) : (\boldsymbol{y},c)\in D\times\mathbb{R}\Big\} \tag{2.40}$$

This set can be equivalently defined as the region in D where $\lim_{t\to\infty}\int_{\mathbb{R}}\rho_0(t,\cdot,c)dc > 0$. Using the definition of $U(t,\boldsymbol{y})$ in (2.36) we see that equation (2.39) can be written as

$$-F(\boldsymbol{y}) + \int_{D_0} K(\boldsymbol{y},\boldsymbol{y}')d\gamma_0(\boldsymbol{y}') = 0 \qquad \text{for } \nu_0\text{-almost all } \boldsymbol{y}\in D_0 = \operatorname{supp}\nu_0 \tag{2.41}$$

where we defined the measures ν_0 and γ_0 as: for all test function $\chi : D\to\mathbb{R}$,

$$\begin{aligned} \lim_{t\to\infty}\int_{D\times\mathbb{R}}\chi(\boldsymbol{y})\rho_0(t,\cdot,c)d\boldsymbol{y}dc &= \int_D \chi(\boldsymbol{y})d\nu_0(\boldsymbol{y}) \\ \lim_{t\to\infty}\int_{D\times\mathbb{R}}\chi(\boldsymbol{y})c\rho_0(t,\cdot,c)d\boldsymbol{y}dc &= \int_D \chi(\boldsymbol{y})d\gamma_0(\boldsymbol{y}) \end{aligned} \tag{2.42}$$

By definition, the measure ν_0 is positive and normalized on its support, $\nu_0(D_0) = 1$, whereas γ_0 is a signed measure whose support is included in (and will generically be) D_0. (2.41) is similar to (1.15), except that it is restricted to $D_0 = \operatorname{supp}\nu_0$. This is a subset of D, $D_0 \subseteq D$, which may not contain all of D and may even be singular in it, i.e. if we denote by λ the Lebesgue measure on $D\subset\mathbb{R}^N$ we have

$$\lambda(D_0)\le\lambda(D) \qquad \text{with} \quad \lambda(D_0) = 0 \quad \text{possibly} \tag{2.43}$$

Even though D_0 may be singular in D, (2.41) well-posed since the term $\int_{D_0} K(\cdot,\boldsymbol{y}')d\gamma_0(\boldsymbol{y}')$ in this equation is a continuously differentiable function of $\boldsymbol{y}$ by Assumption 2.2. It is useful to write (2.41) as

$$0 = \int_\Omega\left(f(\boldsymbol{x}) - \int_{D_0}\varphi(\boldsymbol{x},\boldsymbol{y}')d\gamma_0(\boldsymbol{y}')d\boldsymbol{y}'\right)\varphi(\boldsymbol{x},\boldsymbol{y})d\mu(\boldsymbol{x}) \qquad \text{for } \nu_0\text{-almost all } \boldsymbol{y}\in D_0 \tag{2.44}$$

This formulation shows that if $\varphi(\cdot,\boldsymbol{y})$ is discriminating in D_0 (as it is in D by Assumption 1.2), i.e. if

$$\int_\Omega g(\boldsymbol{x})\varphi(\boldsymbol{x},\cdot)d\mu(\boldsymbol{x}) = 0 \quad \text{a.e. in } D_0 \quad \Rightarrow \quad g = 0 \quad \text{a.e. in } \Omega \tag{2.45}$$

then any solution γ_0 to (2.44) will be such that

$$\int_{D_0}\varphi(\cdot,\boldsymbol{y})d\gamma_0(\boldsymbol{y}) = f \qquad \text{a.e. in } \Omega. \tag{2.46}$$

This is a property of the stationary points of (2.10) which we can summarize into:

Proposition 2.4. *Let $\rho_0(t)$ be the solution to* (2.10) *for the initial condition $\rho_0(0) = \rho_{in}$. Then*

$$\lim_{t\to\infty}\int_{D\times\mathbb{R}} c\varphi(\cdot,\boldsymbol{y})\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc = f \tag{2.47}$$

if $\varphi(\cdot,\boldsymbol{y})$ is discriminating in $D_0 = \operatorname{supp}\nu_0$, i.e. if (2.45) *holds.*

We should stress that ν_0, γ_0, and D_0 depend on the initial $\rho_0(0) = \rho_{\text{in}}$, and we do not know *a priori* whether the condition (2.45) will hold for a given ρ_{in}. There is, however, something generic about this condition, in the sense that it can be satisfed by local (as opposed to global) transports of mass that are energetically favorable. To explain what this means, suppose that γ_0 and D_0 are such that (2.46) does not hold, so that

$$\int_\Omega\left(f(\boldsymbol{x}) - \int_{D_0}\varphi(\boldsymbol{x},\boldsymbol{y})d\gamma_0(\boldsymbol{y})\right)^2 d\mu(\boldsymbol{x}) > 0. \tag{2.48}$$

Now pick a one-parameter family of sets D_s, $s \in [0,1]$, such that $\lim_{s\to 0} D_s = D_0$, $\lambda(D_s) > 0$ for all $s \in (0,1]$, $D_s \subset D_{s'}$ for $s < s'$ with $\lim_{s'\to s} D_{s'} = D_s$, and $D_1 = D$ (or more generally, D_1 is such that (2.45) holds). Then

$$\begin{aligned}&\min_{\gamma_0}\int_\Omega\left(f(\boldsymbol{x})-\int_{D_0}\varphi(\boldsymbol{x},\boldsymbol{y})d\gamma_0(\boldsymbol{y})\right)^2 d\mu(\boldsymbol{x})\\ &\quad\ge \min_{G_0}\int_\Omega\left(f(\boldsymbol{x})-\int_{D_s}\varphi(\boldsymbol{x},\boldsymbol{y})G_0(\boldsymbol{y})d\boldsymbol{y}\right)^2 d\mu(\boldsymbol{x})\ge 0 \qquad \text{for } s\in(0,1]\end{aligned}\tag{2.49}$$

and we can make this energy decrease monotonically by increasing s, until it reaches 0 for the first D_s such that (2.45) holds. Note that the Euler-Lagrange equation for the minimization of the functional of G_0 in (2.49) is precisely (1.10) restricted to D_s—that is, we are back into the $L^2(D)$ framework discussed in Sec. 1.3 and in (2.49) we can perform the minimization of the functional of G_0 over signed densities (as opposed to measures) since $\lambda(D_s) > 0$. In addition, the minimizer G_0^* is realizable in terms of some density ρ_0^* such that $\rho_0^* > 0$ on D_s. Indeed, we can pick a probability density $\rho : D_s \times \mathbb{R} \to (0,\infty)$ such that $\rho > 0$ a.e. in $D_s \times \mathbb{R}$, $\int_{D_s\times\mathbb{R}} \rho(\boldsymbol{y},c)d\boldsymbol{y}dc = 1$ and $\int_\mathbb{R} c\rho(\boldsymbol{y},c)dc = 0$ for almost all $\boldsymbol{y} \in D_s$, and set $\rho_0^* = 0$ if $(\boldsymbol{y},c) \notin D_s \times \mathbb{R}$ and

$$\rho_0^*(\boldsymbol{y},c) = \rho(\boldsymbol{y}, c - G_0^*(\boldsymbol{y})/\bar\rho(\boldsymbol{y})) \quad \text{where} \quad \bar\rho = \int_\mathbb{R} \rho(\cdot,c)dc > 0 \quad \text{a.e. in } D_s \tag{2.50}$$

if $(\boldsymbol{y},c) \in D_s \times \mathbb{R}$. In particular, a minimizer G_0^* can be realized in terms of minimizers of $\mathscr{E}[\rho_0]$ subject to $\rho_0 = 0$ if $(\boldsymbol{y},c) \notin D_s \times \mathbb{R}$, $\rho_0 \ge 0$ if $(\boldsymbol{y},c) \in D_s \times \mathbb{R}$, and $\int_{D_s\times\mathbb{R}} \rho_0 d\boldsymbol{y}dc = 1$ since

$$\int_\Omega\left(f(\boldsymbol{x})-\int_{D_s}\varphi(\boldsymbol{x},\boldsymbol{y})G_0^*(\boldsymbol{y})d\boldsymbol{y}\right)^2 d\mu(\boldsymbol{x}) = \mathscr{E}_0[\rho_0^*] = \min \mathscr{E}_0[\rho_0] \tag{2.51}$$

where the minimization of $\mathscr{E}_0[\rho_0]$ is performed under the constraints listed.

The considerations above mean that we can continuously deform a D_0 on which (2.46) does not hold into a D_s on which

$$\int_{D_s} \varphi(\cdot,\boldsymbol{y})G_0^*(\boldsymbol{y})d\boldsymbol{y} = f \quad \text{a.e. in } \Omega \tag{2.52}$$

These deformations fatten locally the support of a stationary point of (2.10) and lead to other stationary points $\lim_{t\to\infty}\rho_0(t)$ which are absolutely continuous with respect to the Lebesgue measure on D and more favorable since they have lower energy. While these deformations are not accessible dynamically by GD, it is reasonable to assume that they will occur if we add any amount of noise in the dynamics of the particles in (2.1). That is, we expect that, if we perturb the GD flow, the dynamics will eventually reach a D_s where (2.52) is satisfied on a timescale that is inversely proportional to the size of the fluctuations (rather than exponentially large in the inverse of the fluctuation amplitude, as one would expect if escaping stationary points required global transport of mass involving crossing events over energy barriers at particle level): this point is further discussed in Appendix A, where we consider what happens when additive noise is added to the ODEs in (2.1) and show that (2.46) holds in that case. Our working assumption will be that the SGD dynamics discussed in Sec. 3 also eventually reach a D_s such that (2.46) holds.

Remark 2.3. We can also derive (2.39) and (2.41) by looking at the evolution of $\mathscr{E}_0[\rho_0(t)]$. From (2.35) and (2.36) we obtain:

$$\frac{d}{dt}\mathscr{E}_0[\rho_0(t)] = -\int_{D\times\mathbb{R}} \left(c^2|\nabla U(t)|^2 + |U(t)|^2\right)\rho_0(t)d\boldsymbol{y}dc \tag{2.53}$$

This equation shows that stationarity requires either $\lim_{t\to\infty}\int_\mathbb{R}\rho_0(t,\cdot,c)dc = 0$ or $\lim_{t\to\infty} U(t) = 0$, consistent with (2.39) and (2.41).

Remark 2.4. The features of the stationary points discussed above are quite different from those encountered in standard interacting particles systems. This is due to the fact that the interaction potential we use is also quite different; it involves fractional charges that we view as evolving alongside the particles themselves. This has the effect of making the energy for ρ_0 in (2.12) a functional of $G_0 = \int_{\mathbb{R}} \rho_0(\cdot, c)dc$,

$$\mathscr{E}_0[\rho_0] = \hat{\mathscr{E}}_0\left[\int_{\mathbb{R}}\rho_0(\cdot,c)dc\right] \quad \text{with} \quad \hat{\mathscr{E}}_0[G_0] = C_f - \int_D FG_0 d\boldsymbol{y} + \tfrac{1}{2}\int_{D^2} cc'K(\boldsymbol{y},\boldsymbol{y}')G_0G_0' d\boldsymbol{y}d\boldsymbol{y}' \tag{2.54}$$

This functional is the loss function written as in (1.7), and, if we restrict it to D_s, it is also the functional for G_0 in (2.49). Ultimately, the special features we discussed in this section are related to the fact that minimizing $\mathscr{E}[\rho_0]$ over ρ_0 with the constraints that $\rho_0 \geq 0$ and $\int_{D\times\mathbb{R}} \rho_0 d\boldsymbol{y}dc = 1$ is equivalent to the simpler problem of minimizing $\hat{\mathscr{E}}_0[G_0]$ over G_0 without any constraints.

2.3.2. *The dynamics of* $f_0(t, \boldsymbol{x})$. It is interesting to revisit the results of the previous section in terms of what they imply for the evolution of

$$f_0(t,\boldsymbol{x}) = \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc \tag{2.55}$$

By writing (2.10) as

$$\begin{aligned}\partial_t\rho_0 &= \nabla\cdot\left(c\int_\Omega \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\left(f_0(t,\boldsymbol{x})-f(\boldsymbol{x})\right)d\mu(\boldsymbol{x})\rho_0\right)\\ &\quad+\partial_c\left(\int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y})\left(f_0(t,\boldsymbol{x})-f(\boldsymbol{x})\right)d\mu(\boldsymbol{x})\rho_0\right)\end{aligned} \tag{2.56}$$

we deduce, using (2.55),

$$\begin{aligned}\partial_t f_0(t,\boldsymbol{x}) &= \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\partial_t\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc\\ &= \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\nabla\cdot\left(c\int_\Omega \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\left(f_0(t,\boldsymbol{x})-f(\boldsymbol{x})\right)d\mu(\boldsymbol{x})\rho_0\right)d\boldsymbol{y}dc\\ &\quad+\int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\partial_c\left(\int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y})\left(f_0(t,\boldsymbol{x})-f(\boldsymbol{x})\right)d\mu(\boldsymbol{x})\rho_0\right)d\boldsymbol{y}dc.\end{aligned} \tag{2.57}$$

Integrating by parts in $\boldsymbol{y}$ the first term and in c the second, and interchanging the order of integration between $(\boldsymbol{y}, c)$ and $\boldsymbol{x}$ on both these terms, this equation can be written as

$$\partial_t f_0(t,\boldsymbol{x}) = -\int_\Omega M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')\left(f_0(t,\boldsymbol{x}')-f(\boldsymbol{x}')\right)d\mu(\boldsymbol{x}') \tag{2.58}$$

where we defined the kernel

$$M([\rho],\boldsymbol{x},\boldsymbol{x}') = \int_{D\times\mathbb{R}}\left(c^2\nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x}',\boldsymbol{y})+\varphi(\boldsymbol{x},\boldsymbol{y})\varphi(\boldsymbol{x}',\boldsymbol{y})\right)\rho(\boldsymbol{y},c)d\boldsymbol{y}dc. \tag{2.59}$$

This kernel is symmetric in $\boldsymbol{x}$ and $\boldsymbol{x}'$ and positive semidefinite since, given any $r \in L^2(\Omega, \mu)$, we have

$$\int_{\Omega^2} r(\boldsymbol{x})r(\boldsymbol{x}')M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')d\mu(\boldsymbol{x})d\mu(\boldsymbol{x}') = \int_{D\times\mathbb{R}}\left(c^2|\nabla R(\boldsymbol{y})|^2+|R(\boldsymbol{y})|^2\right)\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc \geq 0 \tag{2.60}$$

where

$$R(\boldsymbol{y}) = \int_\Omega r(\boldsymbol{x})\varphi(\boldsymbol{x},\boldsymbol{y})d\mu(\boldsymbol{x}). \tag{2.61}$$

We also have

$$\int_{\Omega^2} r(\boldsymbol{x})r(\boldsymbol{x}')M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')d\mu(\boldsymbol{x})d\mu(\boldsymbol{x}') \geq \int_{D\times\mathbb{R}}|R(\boldsymbol{y})|^2\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc \geq 0 \tag{2.62}$$

and taking the limit as $t \to \infty$, we deduce

$$\lim_{t\to\infty}\int_{\Omega^2} r(\boldsymbol{x})r(\boldsymbol{x}')M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')d\mu(\boldsymbol{x})d\mu(\boldsymbol{x}') \geq \int_{D_0}|R(\boldsymbol{y})|^2 d\nu_0(\boldsymbol{y}) \geq 0 \tag{2.63}$$

where ν_0 is defined in (2.42) and D_0 is its support. As a result, (2.60) can only be zero if $R = 0$ a.e. in D_0. If D_0 is such that (2.45) holds, this implies that $r = 0$ a.e. in Ω, which shows that the only stable fixed point of (2.58) is f in that case. In other words, since $f_0(t)$ is the limit of $f_n(t)$ as $n \to \infty$, we have formally established:

Proposition 2.5 (LLN)**.** *Let $f_n(t) = f_n(t, \boldsymbol{x})$ be given by* (2.5) *with $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ solution of* (2.1) *with initial condition drawn from $\mathbb{P}_{in}$. Then*

$$\lim_{n\to\infty} f_n(t) = f_0(t) \qquad \mathbb{P}_{in}\text{-almost surely} \tag{2.64}$$

where $f_0(t)$ solves (2.58) *and satisfies*

$$\lim_{t\to\infty} f_0(t) = f \quad a.e.\ in\ \ \Omega \qquad if\ \ (2.45)\ \ holds. \tag{2.65}$$

(2.65) is equivalent to the statement in Proposition 2.4. Notice that (2.58) confirms that $f_0(t)$ evolves on a quadratic landscape, namely the loss function (1.2) itself: Indeed this equation can be written as

$$\partial_t f_0(t, \boldsymbol{x}) = -\int_\Omega M([\rho_0(t)], \boldsymbol{x}, \boldsymbol{x}')\, D_{f_0(t,\boldsymbol{x}')}\ell(f, f_0(t))\, d\mu(\boldsymbol{x}') \tag{2.66}$$

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

$$\forall h : \Omega \to \mathbb{R} \quad : \quad \lim_{z\to 0} \frac{d}{dz}\mathscr{F}[f + zh] = \langle h, D_f \mathscr{F}[f]\rangle_{L^2(\Omega,\mu)} = \int_\Omega h(\boldsymbol{x}) D_{f(\boldsymbol{x})}\mathscr{F}[f]\, d\mu(\boldsymbol{x}) \tag{2.67}$$

(That is, $D_{f(\boldsymbol{x})}$ reduces to $\delta/\delta f(\boldsymbol{x})$ if $d\mu(\boldsymbol{x}) = d\boldsymbol{x}$.) Notice also that Proposition 2.5 indicates that, if (2.45) holds, the rate in time at which $\rho_0(t)$ converges towards its fixed point and $f_0(t)$ towards f is independent of n to leading order, since n does not enter (2.10) and (2.58).

Remark 2.5. Even though Proposition 2.5 specifies fully the behavior of $\lim_{n\to\infty} f_n(t)$ as $t \to \infty$, it gives incomplete information about that of $\lim_{n\to\infty} \rho_n(t) = \rho_0(t)$: we only know that this limiting $\rho_0(t)$ is such that $\int_{D\times\mathbb{R}} c\varphi(\cdot, \boldsymbol{y})\rho_0(t) d\boldsymbol{y} dc$ converges to f as $t \to \infty$ if (2.45) holds. As already stated above, if we add noise in (2.1) it is reasonable to assume that (2.45) will automatically be satisfied. In Appendix A, we will analyze the behavior of $\rho_0(t)$ on a longer timescale and show that, with noise in (2.1), $\rho_0(t)$ reaches a unique fixed point ρ_0^* such that $\rho_0^* > 0$ in $D \times \mathbb{R}$ and $\int_{D\times\mathbb{R}} \rho_0^* \log \rho_0^* d\boldsymbol{y} dc < \infty$.

2.4. **Central Limit Theorem (CLT).** Let us now analyze (2.31) assuming that (2.45) holds. Notice first that, because $F_0(t) \to F$ as $t \to \infty$ on the timescales where (2.31) holds, this equation reduces to

$$\partial_t \rho_{\bar\xi} = \nabla \cdot \left(\int_{D\times\mathbb{R}} cc' \nabla K(\boldsymbol{y}, \boldsymbol{y}') \rho'_{\bar\xi} \rho_0 d\boldsymbol{y}' dc' \right) + \partial_c \left(\int_{D\times\mathbb{R}} c' K(\boldsymbol{y}, \boldsymbol{y}') \rho'_{\bar\xi} \rho_0 d\boldsymbol{y}' dc' \right) \tag{2.68}$$

in which $\rho_0 = \lim_{t\to\infty} \rho_0(t)$ is the fixed point reached by (2.10). Proceeding as we did to derive (2.58) we can write an equation for

$$f_{\bar\xi}(t, \boldsymbol{x}) = \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x}, \boldsymbol{y}) \rho_{\bar\xi}(t, \boldsymbol{y}, c) d\boldsymbol{y} dc \tag{2.69}$$

which is

$$\partial_t f_{\bar\xi} = -\int_\Omega M([\rho_0], \boldsymbol{x}, \boldsymbol{x}') f_{\bar\xi}(t, \boldsymbol{x}') d\mu(\boldsymbol{x}') \tag{2.70}$$

where $M([\rho_0], \boldsymbol{x}, \boldsymbol{x}')$ is the kernel defined in (2.59) evaluated on $\rho = \rho_0 = \lim_{t\to\infty} \rho_0(t)$. If (2.45) holds this kernel is positive-definite, and the only fixed point of (2.70) is zero so that

$$\lim_{t\to\infty} f_{\bar\xi}(t) = 0. \tag{2.71}$$

This also implies that $\rho_{\bar\xi} = 0$ is a stable fixed point of (2.68), which was the condition for (2.34) to hold.

Summarizing we have established:

Proposition 2.6 (CLT)**.** *Let $f_n(t) = f_n(t, \boldsymbol{x})$ be given by* (2.5) *with $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ solution of* (2.1) *with initial condition drawn from $\mathbb{P}_{in}$ and assume that* (2.45) *holds. Then for any $\bar\xi < 1$ and any $a_n > 0$ such that $a_n/\log n \to \infty$ as $n \to \infty$, we have*

$$\lim_{n\to\infty} n^{\bar\xi}\left(f_n(a_n) - f_0(a_n)\right) = 0 \qquad \textit{almost surely} \tag{2.72}$$

where $f_0(t)$ solves (2.58) *and satisfies $f_0(t) \to f$ as $t \to \infty$.*

This proposition can be stated as (1.22). It shows a remarkable self-healing property of the dynamics: the fluctuations at scale $O(n^{-1/2})$ of $f_n(t)$ around $f_0(t)$ that were present initially decrease in amplitude as time progresses, and become $O(n^{-1})$ as $t \to \infty$.

Finally, note that the results above have consequences in terms of the scaling of the loss function. We can write

$$\ell(f, f_n(a_n)) = \tfrac{1}{2}\|f - f_0(a_n)\|^2 - n^{-\bar\xi}\langle f - f_0(a_n), f_{\bar\xi}(a_n)\rangle + \tfrac{1}{2}n^{-2\bar\xi}\|f_{\bar\xi}(a_n)\|^2 + o(n^{-\bar\xi}) \tag{2.73}$$

where the norms and inner products are taken in $L^2(\Omega,\mu)$. Since $f_0(t) \to f$ and $f_{\bar\xi}(t) \to 0$ as $t \to \infty$, we deduce

Proposition 2.7. *In the same conditions as those in Proposition 2.6, the loss function satisfies*

$$\lim_{n\to\infty} n^{\bar\xi}\ell(f, f_n(a_n)) = 0 \qquad \textit{almost surely.} \tag{2.74}$$

3. Discrete training set and stochastic gradient descent (SGD)

In most applications, it is not possible to evaluate the integrals in (1.4) defining $F(\boldsymbol{y})$ and $K(\boldsymbol{y},\boldsymbol{z})$. This is especially true for $F(\boldsymbol{y})$, since we typically have limited access to $f(\boldsymbol{x})$: often, we only know its value on a discrete set of points. In these cases, unless we use radial basis function networks (as discussed in Sec. 4.1) we need to approximate the integrals in (1.4) by sum over a finite set of $\boldsymbol{x}$'s by sampling from the measure μ. Typically, this sampling is done at every step of the learning process, which introduce some noise, and the resulting scheme is referred to as stochastic gradient descent. The noise in this scheme can be used to replace the noise terms in (2.1): here we discuss in which way this modification impacts the results established before.

The SGD scheme used to train the network in applications reads

$$\begin{cases} \boldsymbol{Y}_i^P(t+\Delta t) = \boldsymbol{Y}_i^P(t) + C_i^P(t)\nabla F_P(t, \boldsymbol{Y}_i^P(t))\Delta t - \dfrac{1}{n}\displaystyle\sum_{j=1}^n C_i^P(t)C_j^P(t)\nabla K_P(t, \boldsymbol{Y}_i^P(t), \boldsymbol{Y}_j^P(t))\Delta t \\ C_i^P(t+\Delta t) = C_i(t) + F_P(t, \boldsymbol{Y}_i^P(t))\Delta t - \dfrac{1}{n}\displaystyle\sum_{j=1}^n C_j^P(t)K_P(t, \boldsymbol{Y}_i^P(t), \boldsymbol{Y}_j^P(t))\Delta t \end{cases} \tag{3.1}$$

where $\Delta t > 0$ is some time-step and we defined

$$F_P(t,\boldsymbol{y}) = \frac{1}{P}\sum_{p=1}^P f(\boldsymbol{X}_p(t))\varphi(\boldsymbol{X}_p(t),\boldsymbol{y}), \qquad K_P(t,\boldsymbol{y},\boldsymbol{y}') = \frac{1}{P}\sum_{p=1}^P \varphi(\boldsymbol{X}_p(t),\boldsymbol{y})\varphi(\boldsymbol{X}_p(t),\boldsymbol{y}') \tag{3.2}$$

in which $\{\boldsymbol{X}_p(t)\}_{p=1}^P$ are P iid variables which are redrawn from μ independently at every time step t.

3.1. **Limiting stochastic differential equation (SDE).**

To analyze the properties of (3.1), it is convenient to use compact notations and denote the set of all particles as

$$\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n) = (\boldsymbol{y}_1, c_1, \dots, \boldsymbol{y}_n, c_n) \in (D \times \mathbb{R})^n, \qquad \boldsymbol{z}_i = (\boldsymbol{y}_i, c_i) \in D \times \mathbb{R} \qquad i = 1, \dots, n \tag{3.3}$$

and use the shorthands

$$f_n(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^n c_i\varphi(\cdot, \boldsymbol{y}_i), \qquad f_n(\boldsymbol{x},\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^n c_i\varphi(\boldsymbol{x},\boldsymbol{y}_i) \tag{3.4}$$

The stochastic gradient descent in (3.1) can be written as

$$\boldsymbol{Z}(t+\Delta t) = \boldsymbol{Z}(t) - \Delta t\nabla_{\boldsymbol{z}} L_P(\boldsymbol{Z}(t)) \tag{3.5}$$

where $L_P(\boldsymbol{z})$ is the approximation of the loss function obtained with a batch of P independent samples $\{\boldsymbol{X}_p(t)\}_{p=1}^P$ drawn from μ and scaled by n:

$$L_P(\boldsymbol{z}) = \frac{n}{2P}\sum_{p=1}^{P}\left|f(\boldsymbol{X}_p) - f_n(\boldsymbol{X}_p,\boldsymbol{z})\right|^2. \tag{3.6}$$

Notice that $L_P(\boldsymbol{z})$ has expectation $n\ell(f, f_n(\boldsymbol{z}))$:

$$\mathbb{E}L_P(\boldsymbol{z}) = \frac{n}{2}\int_\Omega \left|f(\boldsymbol{x}) - f_n(\boldsymbol{x},\boldsymbol{z})\right|^2 d\mu(\boldsymbol{x}) = n\ell(f, f_n). \tag{3.7}$$

Note also that

$$\nabla_{\boldsymbol{z}} L_P(\boldsymbol{z}) = \frac{n}{P}\sum_{p=1}^{P}\left(f_n(\boldsymbol{X}_p,\boldsymbol{z}) - f(\boldsymbol{X}_p)\right)\nabla_{\boldsymbol{z}} f_n(\boldsymbol{X}_p,\boldsymbol{z}). \tag{3.8}$$

Let us introduce the covariance of this quantity:

$$\mathbb{E}\left(\nabla_{\boldsymbol{z}}(L_P(\boldsymbol{z}) - n\ell(f, f_n(\boldsymbol{z})))\right)\otimes\left(\nabla_{\boldsymbol{z}}(L_P(\boldsymbol{z}') - n\ell(f, f_n(\boldsymbol{z}'))\right) = \frac{1}{P}R(\boldsymbol{z}) \tag{3.9}$$

with

$$\begin{aligned} R(\boldsymbol{z}) = n^2 &\int_\Omega \left|f(\boldsymbol{x}) - f_n(\boldsymbol{x},\boldsymbol{z})\right|^2 \nabla_{\boldsymbol{z}} f_n(\boldsymbol{x},\boldsymbol{z}) \otimes \nabla_{\boldsymbol{z}} f_n(\boldsymbol{x},\boldsymbol{z})\, d\mu(\boldsymbol{x}) \\ &- n^2 \nabla_{\boldsymbol{z}}\ell(f, f_n(\boldsymbol{z})) \otimes \nabla_{\boldsymbol{z}}\ell(f, f_n(\boldsymbol{z})). \end{aligned} \tag{3.10}$$

It is useful in what follows to express the blocks of this tensor as

$$R_{i,j}(\boldsymbol{z}) = \begin{pmatrix} c_i c_j A_2([f - f_n], \boldsymbol{y}_i, \boldsymbol{y}_j) & c_i A_1([f - f_n], \boldsymbol{y}_i, \boldsymbol{y}_j) \\ c_j A_1([f - f_n], \boldsymbol{y}_j, \boldsymbol{y}_i) & A_0([f - f_n], \boldsymbol{y}_i, \boldsymbol{y}_j) \end{pmatrix} \tag{3.11}$$

where

$$\begin{aligned} A_0([f], \boldsymbol{y}, \boldsymbol{y}') &= \int_\Omega |f(\boldsymbol{x})|^2 \varphi(\boldsymbol{x},\boldsymbol{y})\varphi(\boldsymbol{x},\boldsymbol{y}')\, d\mu(\boldsymbol{x}) \\ &\quad - \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x},\boldsymbol{y})\, d\mu(\boldsymbol{x}) \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x},\boldsymbol{y}')\, d\mu(\boldsymbol{x}) \\ A_1([f], \boldsymbol{y}, \boldsymbol{y}') &= \int_\Omega |f(\boldsymbol{x})|^2 \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\varphi(\boldsymbol{x},\boldsymbol{y}')\, d\mu(\boldsymbol{x}) \\ &\quad - \int_\Omega f(\boldsymbol{x})\nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\, d\mu(\boldsymbol{x}) \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x},\boldsymbol{y}')\, d\mu(\boldsymbol{x}) \\ A_2([f], \boldsymbol{y}, \boldsymbol{y}') &= \int_\Omega |f(\boldsymbol{x})|^2 \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y}) \otimes \nabla_{\boldsymbol{y}'}\varphi(\boldsymbol{x},\boldsymbol{y}')\, d\mu(\boldsymbol{x}) \\ &\quad - \int_\Omega f(\boldsymbol{x})\nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\, d\mu(\boldsymbol{x}) \otimes \int_\Omega f(\boldsymbol{x})\nabla_{\boldsymbol{y}'}\varphi(\boldsymbol{x},\boldsymbol{y}')\, d\mu(\boldsymbol{x}) \end{aligned} \tag{3.12}$$

To keep track of the terms, note that $A_0 \in \mathbb{R}$, $A_1 \in \mathbb{R}^n$, and $A_2 \in \mathbb{R}^n \times \mathbb{R}^n$.

If in (3.5) we split $\nabla_{\boldsymbol{z}} L_P(\boldsymbol{Z}(t))$ into its expectation plus a zero-mean fluctuations with covariance (3.9), we can think of (3.1) as an Euler-Maruyama scheme for an SDE, except that the scaling of the noise term involves Δt rather than $\sqrt{\Delta t}$. This SDE is

$$d\boldsymbol{Z} = n\nabla_{\boldsymbol{z}}\ell(f, f_n(\boldsymbol{Z}))\,dt + \sqrt{\theta}\, d\boldsymbol{B} \tag{3.13}$$

where $\theta = \Delta t / P$ and $d\boldsymbol{B}$ is a white-noise process with quadratic variation

$$\langle d\boldsymbol{B}, d\boldsymbol{B}\rangle = R(\boldsymbol{Z})\,dt \tag{3.14}$$

In the parameter and charge variables, (3.13) reads

$$\begin{cases} d\boldsymbol{Y}_i = C_i \nabla F(\boldsymbol{Y}_i)dt - \dfrac{1}{n}\displaystyle\sum_{j=1}^n C_i C_j \nabla K(\boldsymbol{Y}_i,\boldsymbol{Y}_j)dt + \sqrt{\theta}d\boldsymbol{B}_i, \\ dC_i = F(\boldsymbol{Y}_i)dt - \dfrac{1}{n}\displaystyle\sum_{j=1}^n C_j K(\boldsymbol{Y}_i,\boldsymbol{Y}_j)dt + \sqrt{\theta}dB'_i \end{cases} \tag{3.15}$$

where $\{d\boldsymbol{B}_i, dB'_i\}_{i=1}^n$ are a white-noise processes with quadratic variation

$$\begin{aligned} \langle d\boldsymbol{B}_i, d\boldsymbol{B}_j\rangle &= C_i C_j A_2([f-f_n],\boldsymbol{Y}_i,\boldsymbol{Y}_j)dt \\ \langle d\boldsymbol{B}_i, dB'_j\rangle &= C_i A_1([f-f_n],\boldsymbol{Y}_i,\boldsymbol{Y}_j)dt \\ \langle dB'_i, dB'_j\rangle &= A_0([f-f_n],\boldsymbol{Y}_i,\boldsymbol{Y}_j)dt \end{aligned} \tag{3.16}$$

Let us state more precisely how the solution to (3.1) is close to that of (3.15). Denote

$$f_n^P(t,\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^N C_i^P(t)\varphi(\boldsymbol{x},\boldsymbol{Y}_i^P(t)), \qquad f_n(t,\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^N C_i(t)\varphi(\boldsymbol{x},\boldsymbol{Y}_i(t)), \tag{3.17}$$

where in the first sum we use the solution to (3.1) and in the second that of (3.15). Then we have [14, 20]

Proposition 3.1. *Given given any test functions $\psi:\Omega\to\mathbb{R}$ and $H:\mathbb{R}\to\mathbb{R}$, and any $T>0$, there is a constant $C>0$ such that*

$$\sup_{0\le k\Delta t\le T}\left|\mathbb{E}H\left(\int_\Omega \psi(\boldsymbol{x})f_n(k\Delta t,\boldsymbol{x})d\mu(\boldsymbol{x})\right) - \mathbb{E}H\left(\int_\Omega \psi(\boldsymbol{x})f_n^P(k\Delta t,\boldsymbol{x})d\mu(\boldsymbol{x})\right)\right| \le C\Delta t. \tag{3.18}$$

This proposition is a direct consequence of the fact that (3.1) can be viewed as the Euler-Maruyama discretization scheme for (3.15), and this scheme has weak order of accuracy 1. Note that if we let $\Delta t\to 0$, (3.15) reduces to the ODEs in (2.1) since $\theta = \Delta t/P \to 0$ in that limit. We should stress, however, that this limit is not reached in practice since the scheme (3.1) is used at small but finite Δt. As we show below θ should in fact be small—we can also adjust the size of θ at fixed Δt by changing P, i.e., by changing the batch size.

3.2. **Dean's equation for particles with correlated noise.** Proposition 3.1 indicates that we can analyze the properties of (3.15) instead of that of (3.1). Dean's equation for the empirical distribution of the process defined by (3.15) can be derived as in Sec. 2.1, except that we need to take into account the effect of the extra drift terms and the noise terms in (3.15).

Applying Itô's formula to (2.4) when $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ satisfy (3.15), we arrive at

$$\begin{aligned} d\rho_n(t,\boldsymbol{y},c) = &-\frac{1}{n}\sum_{i=1}^n \delta(c-C_i)\nabla\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\cdot d\boldsymbol{Y}_i \\ &-\frac{1}{n}\sum_{i=1}^n \partial_c\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)dC_i \\ &+\frac{\theta}{2n}\sum_{i=1}^n \delta(c-C_i)\nabla\nabla\delta(\boldsymbol{y}-\boldsymbol{Y}_i) : C_i C_i A_2([f-f_n],\boldsymbol{Y}_i,\boldsymbol{Y}_i)dt \\ &+\frac{\theta}{n}\sum_{i=1}^n \partial_c^2\delta(c-C_i)\nabla\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\cdot C_i A_1([f-f_n],\boldsymbol{Y}_i,\boldsymbol{Y}_i)dt \\ &+\frac{\theta}{2n}\sum_{i=1}^n \partial_c^2\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)A_0([f-f_n],\boldsymbol{Y}_i,\boldsymbol{Y}_i)dt. \end{aligned} \tag{3.19}$$

We use (2.4) to write $d\boldsymbol{Y}_i$ and dC_i, the drift terms that emerge can be treated as we did to derive (2.6). The noise term in (3.19) is

$$-\frac{1}{n}\sum_{i=1}^n \delta(c-C_i)\nabla\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\cdot d\boldsymbol{B}_i - \frac{1}{n}\sum_{i=1}^n \partial_c\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)dB_i \tag{3.20}$$

and it can be checked explicitly using (3.16) that its quadratic variation can be expressed as

$$\begin{aligned}
&\nabla\nabla' : \big(\rho_n(t,\boldsymbol{y},c)\rho_n(t,\boldsymbol{y}',c')cc'A_2([f_n(t)-f],\boldsymbol{y},\boldsymbol{y}')\big)\,dt\\
&+\partial_c\partial_{c'}\big(\rho_n(t,\boldsymbol{y},c)\rho_n(t,\boldsymbol{y}',c')A_0([f_n(t)-f],\boldsymbol{y},\boldsymbol{y}')\big)\,dt\\
&+\partial_c\nabla'\cdot\big(\rho_n(t,\boldsymbol{y},c)\rho_n(t,\boldsymbol{y}',c')c'A_1([f_n(t)-f],\boldsymbol{y}',\boldsymbol{y})\big)\,dt\\
&+\partial_{c'}\nabla\cdot\big(\rho_n(t,\boldsymbol{y},c)\rho_n(t,\boldsymbol{y}',c')cA_1([f_n(t)-f],\boldsymbol{y},\boldsymbol{y}')\big)\,dt.
\end{aligned}\tag{3.21}$$

With this calculation we obtain Dean's equation for the empirical distribution of the stochastic gradient descent process

$$\begin{aligned}
\partial_t\rho_n = \nabla\cdot&\left(-c\nabla F\rho_n+\int_{D\times\mathbb{R}}cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\rho_n'\rho_n d\boldsymbol{y}'dc'\right)\\
&+\partial_c\left(-F\rho_n+\int_{D\times\mathbb{R}}c'K(\boldsymbol{y},\boldsymbol{y}')\rho_n'\rho_n d\boldsymbol{y}'dc'\right)\\
&+\tfrac{1}{2}\theta\nabla\nabla:\big(\rho_n c^2A_2([f_n(t)-f],\boldsymbol{y},\boldsymbol{y})\big)+\tfrac{1}{2}\theta\partial_c^2\big(\rho_nA_0([f_n(t)-f],\boldsymbol{y},\boldsymbol{y})\big)\\
&+\theta\partial_c\nabla\cdot\big(\rho_n cA_1([f_n(t)-f],\boldsymbol{y},\boldsymbol{y})\big)\\
&+\sqrt{\theta}\,\dot{\eta}_n(t,\boldsymbol{y},c)
\end{aligned}\tag{3.22}$$

where $f_n(t)$ is given by (2.5), i.e. $f_n(t,\boldsymbol{x})=\int_{D\times\mathbb{R}}c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc$, and we defined the white-noise process $\dot{\eta}_n(t,\boldsymbol{y},c)$ with quadratic variation (3.21).

The first two terms at the right hand side of (3.22) are the same as those of (2.6). This is because these terms come from the drift terms in (3.15), which also coincide with those in (2.1). However, (3.22) also contains additional terms that were absent in (2.6). If we want to make these terms higher order so that the LLN established in Proposition 2.5 still applies, we must scale these terms with some inverse power of n. Specifically, we set

$$\theta = an^{-2\alpha}\quad\text{for some}\quad a>0\quad\text{and}\quad\alpha>0\tag{3.23}$$

This scaling can be achieved e.g. by choosing $P=O(n^{2\alpha})$, i.e. by increasing the batch size with n.

3.3. **Limit behavior and fluctuations scaling in SGD.** If we substitute $\theta=an^{-2\alpha}$ with $\alpha>0$ and take the limit as $n\to\infty$ of (3.22) , we formally conclude that $\rho_n\rightharpoonup\rho_0$, where ρ_0 solves the same deterministic equation (2.10) as before.

Turning our attention to the fluctuations of ρ_n around ρ_0, notice that there are two sources of them: some are intrinsic to the discrete nature of the particles apparent in ρ_n, and scale initially as $O(n^{-1/2})$ and eventually as $O(n^{-\bar{\xi}})$ for any $\bar{\xi}<1$, as discussed in Sec. 2.2.2. Other fluctuations come from the noise term in (3.22), and scale as $O(n^{-\alpha})$ when (3.23) holds—note that the drift terms proportional to $\theta=an^{-2\alpha}$ in (3.22) always make higher order contributions.

As a result:

• if $\alpha\in(0,1)$ the fluctuations due to the noise in (3.22) eventually dominate the intrinsic ones from discreteness, and to capture them we can introduce $n^{\alpha}(\rho_n-\rho_0)$, write an equation for this quantity, and take the limit as $n\to\infty$ it. This leads to the conclusion that $n^{\alpha}(\rho_n-\rho_0)\rightharpoonup\rho_\alpha$ which satisfies (compare (2.20))

$$\begin{aligned}
\partial_t\rho_\alpha=\nabla\cdot&\left(-c\nabla F\rho_\alpha+\int_{D\times\mathbb{R}}cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\left(\rho_\alpha'\rho_0+\rho_0'\rho_\alpha\right)d\boldsymbol{y}'dc'\right)\\
&+\partial_c\left(-F\rho_\alpha+\int_{D\times\mathbb{R}}c'K(\boldsymbol{y},\boldsymbol{y}')\left(\rho_\alpha'\rho_0+\rho_0'\rho_\alpha\right)d\boldsymbol{y}'dc'\right)\\
&+\sqrt{a}\,\dot{\eta}_0(t,\boldsymbol{y},c)
\end{aligned}\tag{3.24}$$

in which $\dot\eta_0(t,\boldsymbol{y},c)$ is a white-noise process with quadratic variation

$$\begin{aligned}\langle d\eta_0(t,\boldsymbol{y},c), d\eta_0(t,\boldsymbol{y}',c')\rangle &= \nabla\nabla' : \big(\rho_0(t,\boldsymbol{y},c)\rho_0(t,\boldsymbol{y}',c')cc'A_2([f_0(t)-f],\boldsymbol{y},\boldsymbol{y}')\big)\,dt\\ &\quad+\partial_c\partial_{c'}\big(\rho_0(t,\boldsymbol{y},c)\rho_0(t,\boldsymbol{y}',c')A_0([f_0(t)-f],\boldsymbol{y},\boldsymbol{y}')\big)\,dt\\ &\quad+\partial_c\nabla'\cdot\big(\rho_0(t,\boldsymbol{y},c)\rho_0(t,\boldsymbol{y}',c')c'A_1([f_0(t)-f],\boldsymbol{y}',\boldsymbol{y})\big)\,dt\\ &\quad+\partial_{c'}\nabla\cdot\big(\rho_0(t,\boldsymbol{y},c)\rho_0(t,\boldsymbol{y}',c')cA_1([f_0(t)-f],\boldsymbol{y},\boldsymbol{y}')\big)\,dt.\end{aligned}\tag{3.25}$$

- If $\alpha \ge 1$, then the fluctuations due to the noise in (3.22) are always negligible compared to the intrinsic ones from discreteness, and we are back to the GD situation studied in Sec. 2.

3.4. **Law of Large Numbers and Central Limit Theorem for SGD.** When $\theta = an^{-2\alpha}$ with $\alpha > 0$, $\rho_n \rightharpoonup \rho_0$ as $n\to\infty$, where ρ_0 solves (2.10). This implies that $f_0(t,\boldsymbol{x}) = \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc$ satisfies (2.58) and is such that $f_0(t)\to f$ as $t\to\infty$ if (2.45) holds (and we have already argued that it should hold eventually, due to the added noise in SGD). That is, the LLN in Proposition 2.5 still holds if we use the solution of (3.29) in (2.5); in turn, this means that this proposition also holds up to discretization errors in Δt if we use the solution of (3.1) in (2.5). For further reference, notice that this also implies that the factors defined in (3.12) satisfy

$$\lim_{t\to\infty} A_k([f_0(t)-f],\boldsymbol{y},\boldsymbol{y}') = 0, \qquad k = 0,1,2.\tag{3.26}$$

We use this key property repeatedly in the sequel.

To analyze the fluctuations, let us focus on the situation where SGD differs from GD, i.e, $\alpha\in(0,1)$, and write both (3.24) as

$$\begin{aligned}\partial_t\rho_\alpha &= \nabla\cdot\left(c\int_\Omega \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\big(f_0(t,\boldsymbol{x})-f(\boldsymbol{x})\big)\,d\mu(\boldsymbol{x})\,\rho_\alpha\right)\\ &\quad+\partial_c\left(\int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y})\big(f_0(t,\boldsymbol{x})-f(\boldsymbol{x})\big)\,d\mu(\boldsymbol{x})\,\rho_\alpha\right)\\ &\quad+\nabla\cdot\left(c\int_\Omega \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})f_\alpha(t,\boldsymbol{x})d\mu(\boldsymbol{x})\,\rho_0\right)+\partial_c\left(\int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y})f_\alpha(t,\boldsymbol{x})d\mu(\boldsymbol{x})\,\rho_0\right)\\ &\quad+\sqrt{a}\,\dot\eta_0(t,\boldsymbol{y},c)\end{aligned}\tag{3.27}$$

where we defined

$$f_\alpha(t,\boldsymbol{x}) = \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_\alpha(t,\boldsymbol{y},c)d\boldsymbol{y}dc\tag{3.28}$$

This equation is structurally similar to (2.31) except that it also contains a noise term. By proceeding similarly as we did to derive (2.70) it leads to the following equation for $f_\alpha(t,\boldsymbol{x})$:

$$\begin{aligned}\partial_t f_\alpha &= -\int_\Omega M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')f_\alpha(t,\boldsymbol{x}')d\mu(\boldsymbol{x}')\\ &\quad-\int_\Omega M([\rho_\alpha(t)],\boldsymbol{x},\boldsymbol{x}')\big(f_0(t,\boldsymbol{x}')-f(\boldsymbol{x}')\big)\,d\mu(\boldsymbol{x}')+\sqrt{a}\,\dot\eta(t,\boldsymbol{x})\end{aligned}\tag{3.29}$$

where $M([\rho],\boldsymbol{x},\boldsymbol{x}')$ is given in (2.59), and the quadratic variation of $\dot\eta(t,\boldsymbol{x})$ is precisely that of

$$\int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\dot\eta_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc\tag{3.30}$$

and given by

$$\begin{aligned}\langle d\eta(t,\boldsymbol{x}), d\eta(t,\boldsymbol{x}')\rangle &= \int_\Omega N([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}',\bar{\boldsymbol{x}},\bar{\boldsymbol{x}})\,\big|f_0(t,\bar{\boldsymbol{x}})-f(\bar{\boldsymbol{x}})\big|^2\,d\mu(\bar{\boldsymbol{x}})dt\\ &\quad-\int_{\Omega^2} N([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}',\bar{\boldsymbol{x}},\bar{\boldsymbol{x}}')\big(f_0(t,\bar{\boldsymbol{x}})-f(\bar{\boldsymbol{x}})\big)\big(f_0(t,\bar{\boldsymbol{x}}')-f(\bar{\boldsymbol{x}}')\big)\,d\mu(\bar{\boldsymbol{x}})d\mu(\bar{\boldsymbol{x}}')dt\end{aligned}\tag{3.31}$$

in which

$$
\begin{aligned}
N([\rho],\boldsymbol{x},\boldsymbol{x}',\bar{\boldsymbol{x}},\bar{\boldsymbol{x}}') = \int_{(D\times\mathbb{R}^2)} & \rho(\boldsymbol{y},c)\rho(\boldsymbol{y}',c')\left(c^2\nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}}\varphi(\bar{\boldsymbol{x}},\boldsymbol{y})+\varphi(\boldsymbol{x},\boldsymbol{y})\varphi(\bar{\boldsymbol{x}},\boldsymbol{y})\right) \\
& \times\left(c'^2\nabla_{\boldsymbol{y}'}\varphi(\boldsymbol{x}',\boldsymbol{y}')\cdot\nabla_{\boldsymbol{y}'}\varphi(\bar{\boldsymbol{x}}',\boldsymbol{y}')+\varphi(\boldsymbol{x}',\boldsymbol{y}')\varphi(\bar{\boldsymbol{x}}',\boldsymbol{y}')\right)d\boldsymbol{y}dcd\boldsymbol{y}'dc'.
\end{aligned}
\tag{3.32}
$$

The SDE (3.29) has the property that it *self-quenches* as $t\to\infty$: in that limit we know that $f_0(t)\to f$ and from (3.26) we see that $\dot{\eta}(t)\to 0$ as well. Therefore, at long times (3.29) reduces to

$$
\partial_t f_\alpha = -\int_\Omega M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')f_\alpha(t,\boldsymbol{x}')d\mu(\boldsymbol{x}')
\tag{3.33}
$$

Since $M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')$ is positive-definite the only (stable) fixed point of this equation is zero and $f_\alpha(t)\to 0$ as $t\to\infty$. Of course, to guarantee that this result holds, we should be on a long timescale such that the intrinsic fluctuations discused in Sec. 2.4 have become higher order, i.e. $t=O(a_n)$ with $a_n>0$ such that $a_n/\log n\to\infty$ as $n\to\infty$.

We can summarize this result as:

Proposition 3.2 (CLT for SGD)**.** *Let $f_n(t)=f_n(t,\boldsymbol{x})$ be as in* (2.5) *with $\{\boldsymbol{Y}_i(t),C_i(t)\}_{i=1}^n$ solution to* (3.15) *with $\theta=an^{-2\alpha}$, $a>0$ $\alpha\in(0,1)$, and assume that* (2.45) *holds. Then for any $a_n>0$ such that $a_n/\log n\to\infty$ as $n\to\infty$, we have*

$$
\lim_{n\to\infty} n^\alpha\left(f_n(a_n)-f_0(a_n)\right)=0 \qquad \textit{almost surely}
\tag{3.34}
$$

where $f_0(t)$ solves (2.58) *and is such that $f_0(t)\to f$ as $t\to\infty$*

In this statement, the almost sure convergence is with respect to $\mathbb{P}_{\text{in}}$ as well as the statistics of the noise terms in (3.15). A similar statement holds if we use to the solution to (3.1) in $f_n(t)=f_n(t,\boldsymbol{x})$, but in this case discretization errors in Δt must also be accounted for. In terms of the loss function, we have

$$
\ell(f,f_n(a_n))=\tfrac{1}{2}\|f-f_0(a_n)\|^2-n^{-\alpha}\langle f-f_0(a_n),f_\alpha(a_n)\rangle+\tfrac{1}{2}n^{-2\alpha}\|f_\alpha(a_n)\|^2+o(n^{-\alpha})
\tag{3.35}
$$

and as a result we have the equivalent of Proposition 2.7 in the context of SGD

Proposition 3.3. *Under the same conditions as those in Proposition 3.2, the loss function satisfies*

$$
\lim_{n\to\infty} n^\alpha \ell(f,f_n(a_n))=0 \qquad \textit{almost surely}
\tag{3.36}
$$

4. 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

$$
f(\boldsymbol{x})=\frac{1}{d}\sum_{p,q,r=1}^{d}a_{p,q,r}x_px_qx_r, \qquad \boldsymbol{x}\in S^{d-1}(\sqrt{d})\subset\mathbb{R}^d
\tag{4.1}
$$

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 (4.1) is known to have a number of critical points that grows exponentially with the dimensionality d [1, 2, 24]. We note that previous works have sought to draw a parallel between the glassy 3-spin function and generic loss functions [9], but we are not exploring such an analogy here. Rather, we simply use the function (4.1) 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.

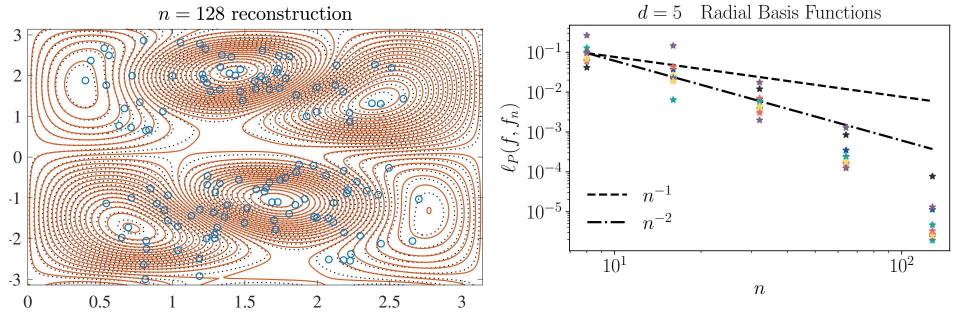

FIGURE 1. Left panel: Comparison between the level sets of the original function f in (4.1) (black dotted curves) and its approximation by the neural network in (4.4) with $n = 128$ and $d = 5$ in the slice defined by (4.7). Also shown are the projection in the slice of the particle position. Right panel: empirical loss in (4.8) 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 (4.1), the full circles show their average value.

4.1. **Learning with Gaussian kernels.** We first consider the case when $D = S^{d-1}(\sqrt{d})$ and we use

$$\varphi(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{1}{2}\alpha|\boldsymbol{x}-\boldsymbol{y}|^2} \tag{4.2}$$

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 $|\boldsymbol{x}| = |\boldsymbol{y}| = \sqrt{d}$, up to an irrelevant constant that can be absorbed in the weights c, we can also write (4.2) as

$$\varphi(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha \boldsymbol{x}\cdot\boldsymbol{y}} \tag{4.3}$$

Setting

$$f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i \varphi(\boldsymbol{x}, \boldsymbol{y}_i) = \frac{1}{n}\sum_{i=1}^{n} c_i e^{-\alpha \boldsymbol{x}\cdot\boldsymbol{y}_i}, \tag{4.4}$$

the GD flow in (2.1) can then be written explicitly as

$$\begin{cases} \dot{\boldsymbol{Y}}_i = C_i \nabla f(\boldsymbol{Y}_i) + \dfrac{\alpha}{n}\displaystyle\sum_{j=1}^{n} C_i C_j \boldsymbol{Y}_j e^{-\alpha \boldsymbol{Y}_i\cdot\boldsymbol{Y}_j} - \lambda_i \boldsymbol{Y}_i \\ \dot{C}_i = f(\boldsymbol{Y}_i) - \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} C_j e^{-\alpha \boldsymbol{Y}_i\cdot\boldsymbol{Y}_j} \end{cases} \tag{4.5}$$

where $-\lambda_i \boldsymbol{Y}_i$ is a Lagrange multiplier term added to enforce $|\boldsymbol{Y}_i| = \sqrt{d}$ for all $i = 1, \ldots, n$, $f(\boldsymbol{x})$ is given by (4.1) and $\nabla f(\boldsymbol{x})$ is given componentwise by

$$\frac{\partial f}{\partial x_p} = \frac{1}{d}\sum_{q,r=1}^{d} \left(a_{p,q,r} + a_{r,p,q} + a_{q,r,p}\right) x_q x_r, \tag{4.6}$$

As is apparent from (4.5) the advantage of using radial basis function networks is that we can use $f(\boldsymbol{x})$ and the kernel $\varphi(\boldsymbol{x}, \boldsymbol{y})$ directly, and do not need to evaluate $F(\boldsymbol{y})$ and $K(\boldsymbol{y}, \boldsymbol{y}')$ (that is, we need no batch). In other words, the cost of running (4.5) 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 $\boldsymbol{y} \in D \subset \mathbb{R}^N$. 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 Sec. 3, the cost of SGD becomes $N^2 n^{2+2\alpha}$, which is quickly becomes much worse than $(dn)^2$ as n grows.

We tested the representation (4.4) 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 (4.5) with time step

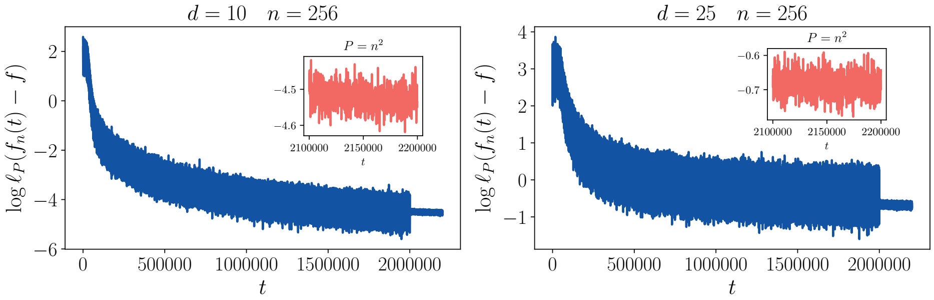

FIGURE 2. The log of the empirical loss in (4.8) as a function of training time by SGD for the sigmoid neural network in $d = 10$ (left panel) and $d = 25$ (right panel). At time $t = 2\times 10^6$, the batch size is increased to initiate a quench. The insets show the log of the empirical loss as a function of time during the final 10^5 time steps of training.

$\Delta t = 10^3$ for 2×10^5 steps: during the first 10^5 we added a bit of thermal noise to (4.5)„ which we then remove during the second half of the run. The representation (4.4) proves to be accurate even at rather low value of n: for example, the right panel of Fig. 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

$$\boldsymbol{x}(\theta) = \sqrt{d}\big(\sin(\theta)\cos(\varphi), \sin(\theta)\sin(\varphi), \cos(\theta), 0, 0\big), \qquad \theta \in [0,\pi], \quad \varphi \in [0, 2\pi). \tag{4.7}$$

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 took values that were rather 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

$$\ell_P(f, f_n(t)) = \frac{1}{2P}\sum_{p=1}^{P} \big|f(\boldsymbol{x}_p) - f_n(t, \boldsymbol{x}_p)\big|^2. \tag{4.8}$$

This empirical loss function was computed with a batch of 10^6 points $\boldsymbol{x}_p$ uniformly distributed on the sphere. The value (4.8) calculated at the end of the calculation is shown as a function of n in the right panel of Fig. 1: the empty circles show (4.8) for 4 individual realizations of the coefficient $a_{p,q,r}$ in (4.1), the full circle shows the average of (4.8) over these 4 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.

4.2. **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 parameterized by $\boldsymbol{y} = (\boldsymbol{a}, b) \in D = \mathbb{R}^{d+1}$, with $\boldsymbol{a} \in \mathbb{R}^d$, $b \in \mathbb{R}^d$, and

$$\varphi(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{a}\cdot\boldsymbol{x} + b) \tag{4.9}$$

This gives

$$f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i h(\boldsymbol{a}_i\cdot\boldsymbol{x} + b_i) \tag{4.10}$$

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 (4.10) 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 Sec. 3. Towards the end of the trajectory,

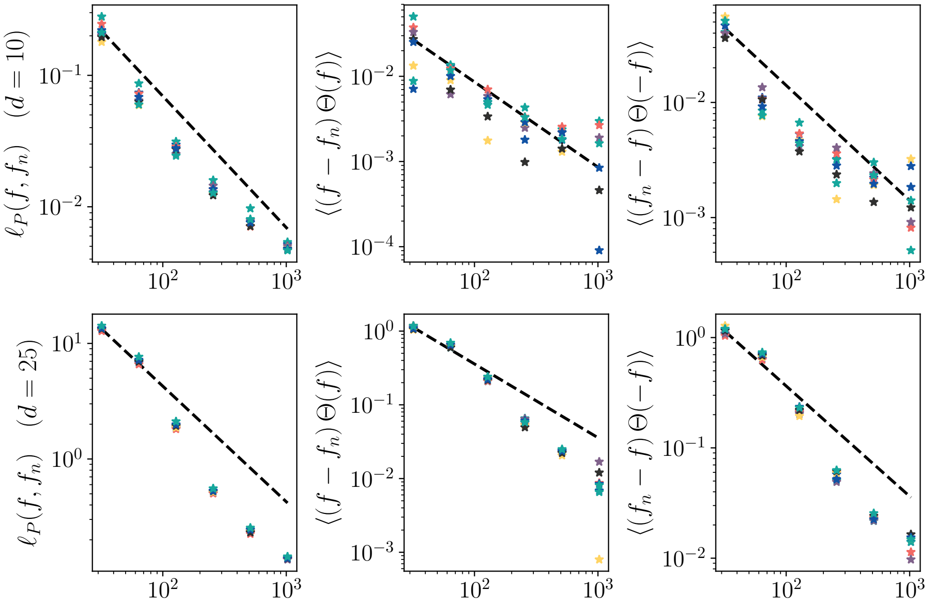

FIGURE 3. Error scaling for single layer neural network with sigmoid nonlinearities. Upper row: $d = 10$; lower row: $d = 25$. The first column shows the empirical loss in (4.8), the second column shows (4.11), and the third column shows (4.11) with $\Theta(f)$ replaced by $\Theta(-f)$. The stars show the results for 10 different realizations of the coefficients $a_{p,q,r}$ in (4.1): the dashed lines decay as n^{-1}, consistent with the predictions in (3.34) and (3.36).

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. Fig. 2 shows the empirical loss in (4.8) calculated 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 (4.8) which we computed over a batch of size $\hat{P} = 10^5$ larger than P. As shown in the two right panels Fig. 3, $\ell_{\hat{P}}(f, f_n(t))$ scales as n^{-1}, consistent with the estimate in (3.36) with $\alpha = 1$. We also tested the estimate in (3.34) using

$$\frac{1}{\hat{P}} \sum_{p=1}^{\hat{P}} \Theta\big(f(\boldsymbol{x}_p)\big)\big(f(\boldsymbol{x}_p) - f_n(t, \boldsymbol{x}_p)\big), \tag{4.11}$$

and similarly with $\Theta\big(-f(\boldsymbol{x}_p)\big)$: here Θ denotes the Heaviside function. The result is shown in the four right panels in Fig. 3: (4.11) scales as n^{-1}, consistent with (3.34) 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 $\{1, \cdots, d\}$

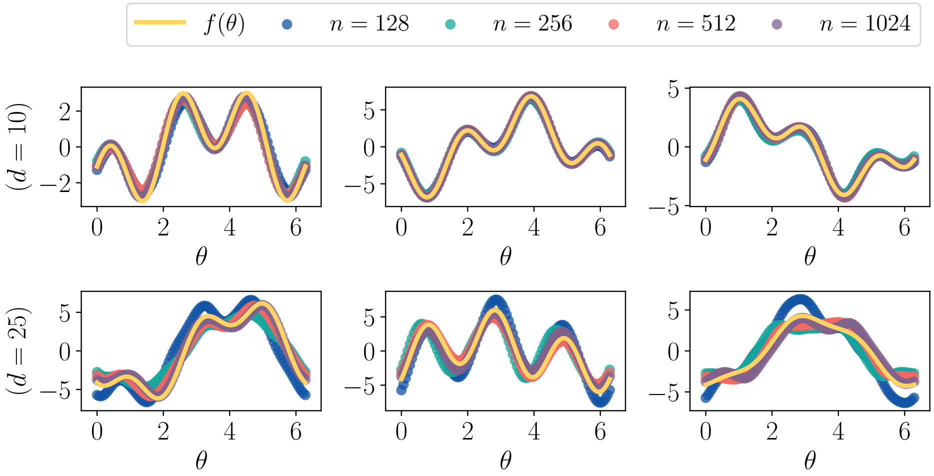

FIGURE 4. One dimensional slices through the $d = 10$ (upper row) and $d = 25$ (lower row) neural net representation f_n are shown below a yellow curve with the target function f. In $d = 10$, the function representations clearly capture the main features of the target function, with only small scale 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 setting $\boldsymbol{x} = \boldsymbol{x}(\theta) = (x_1(\theta), \dots x_d(\theta))$ with

$$x_i(\theta) = \sqrt{d}\cos(\theta), \qquad x_j(\theta) = \sqrt{d}\sin(\theta), \qquad x_k(\theta) = 0 \quad \forall k \neq i, j. \tag{4.12}$$

In Fig. 4 we plot $f(\boldsymbol{x}(\theta)$ and $f_n(\boldsymbol{x}(\theta))$ 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 $2^{25} = 33,554,432$.

5. CONCLUDING REMARKS

Viewing parameters as particles with the loss function as interaction potential enables us to leverage a powerful theoretical apparatus developed in the context of large systems of interacting particles. 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 that the Universal Approximation Theorems follows from a Law of Large Numbers for the empirical distribution of the parameters / particles. Moreover, our results enrich the more abstract derivations of the Universal Approximation Theorem with a dynamical perspective. 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 neural networks are trainable despite the non-convexity 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 particles, specifying the approximation error of neural network representation and showing that it is universal.

We derived these results first in the context of gradient descent dynamics of the particles / parameters; however, they 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.

Our numerical results confirm these predictions, as well as the capability of neural networks to represent high-dimensional function accurately with a relatively modest number of adjustable parameters. Needless to say, this feat opens 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 A. TRAINING AT FINITE (BUT SMALL) TEMPERATURE

For completeness, let us consider here the case when noise-terms are added in (2.1) and we turn these ODEs into stochastic differential equations (SDEs). Adding these terms is a way to address the non-uniqueness issues encountered in Sec. 2. To formulate these SDEs, we need a probability density $m : \mathbb{R} \to (0,\infty)$, used to regularize the dynamics. We specify its properties via:

Assumption A.1. *The density $m : D \to (0,\infty)$ is smooth and such that $m > 0$ and*

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

Next, we replace (2.1) with the SDEs

$$\begin{cases} d\boldsymbol{Y}_i = C_i \nabla F(\boldsymbol{Y}_i) dt - \dfrac{1}{n} \displaystyle\sum_{j=1}^n C_i C_j \nabla K(\boldsymbol{Y}_i, \boldsymbol{Y}_j) dt + \sqrt{2} (\beta n)^{-1/2} d\boldsymbol{W}_i, \\ dC_i = F(\boldsymbol{Y}_i) dt - \dfrac{1}{n} \displaystyle\sum_{j=1}^n C_j K(\boldsymbol{Y}_i, \boldsymbol{Y}_j) dt + (\beta n)^{-1} \partial_c \log m(C_i) dt + \sqrt{2} (\beta n)^{-1/2} dW_i' \end{cases} \tag{A.2}$$

for $i = 1, \dots, n$. Here $\boldsymbol{W}_i$ and W_i' are $2n$ independent Wiener processes, taking values in $\mathbb{R}^N$ and $\mathbb{R}$, respectively, 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.1 below. The SDEs (A.2) are overdamped Langevin equations associated with the energy:

$$E_\beta(\boldsymbol{y}_1, c_1, \dots, \boldsymbol{y}_n, c_n) = nC_f - \sum_{i=1}^n c_i F(\boldsymbol{y}_i) + \frac{1}{2n} \sum_{i,j=1}^n c_i c_j K(\boldsymbol{y}_i, \boldsymbol{y}_j) - (\beta n)^{-1} \sum_{i=1}^n \log m(c_i), \tag{A.3}$$

This energy is (2.2) plus a regularizing term (the one involving $-\log m$). This term guarantees that, for any $\beta > 0$, the following integral is finite

$$Z_n = \int_{(D\times\mathbb{R})^n} e^{-n\beta E_\beta(\boldsymbol{y}_1, c_1, \dots, \boldsymbol{y}_n, c_n)} d\boldsymbol{y}_1 dc_1 \cdots d\boldsymbol{y}_n dc_n < \infty \tag{A.4}$$

which in turns implies that

$$Z_n^{-1} \exp\left(-n\beta E_\beta(\boldsymbol{y}_1, c_1, \cdots, \boldsymbol{y}_n, c_n)\right) \tag{A.5}$$

is a normalized probability density on $(D \times \mathbb{R})^n$. As a result, the solutions to (2.1) are ergodic with respect to the equilibrium distribution with density (A.5) for any $\beta > 0$.

A.1. **Dean's equation.** As shown below, adding the terms involving β^{-1} in (A.2) modifies Dean's equation (A.6) into

$$\begin{aligned} \partial_t \rho_n = \nabla \cdot &\left(-c \nabla F \rho_n + \int_{D\times\mathbb{R}} c c' \nabla K(\boldsymbol{y}, \boldsymbol{y}') \rho_n' \rho_n d\boldsymbol{y}' dc' \right) \\ &+ (\beta n)^{-1} \Delta \rho_n + \sqrt{2} \beta^{-1/2} n^{-1} \nabla \cdot \left(\sqrt{\rho_n}\, \dot{\boldsymbol{\eta}} \right) \\ &+ \partial_c \left(-F \rho_n + \int_{D\times\mathbb{R}} c' K(\boldsymbol{y}, \boldsymbol{y}') \rho_n' \rho_n d\boldsymbol{y}' dc' - (\beta n)^{-1} \partial_c \log m\, \rho_n \right) \\ &+ (\beta n)^{-1} \partial_c^2 \rho_n + \sqrt{2} \beta^{-1/2} n^{-1} \partial_c \left(\sqrt{\rho_n}\, \dot{\eta}' \right) \end{aligned} \tag{A.6}$$

where $\dot{\boldsymbol{\eta}} = \dot{\boldsymbol{\eta}}(t, \boldsymbol{y}, c)$ and $\dot{\eta}' = \dot{\eta}'(t, \boldsymbol{y}, c)$ are independent spatio-temporal white-noises. It is difficult to give this finite-temperature Dean's equation a precise mathematical meaning: viewed as a stochastic partial differential equation (SPDE) (2.6) is ill-posed. It remains useful to analyze the properties of ρ_n as $n \to \infty$, however.

Remark A.1. 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 [22]: with this scaling, the diffusive and regularizing terms in (A.6) are replaced by $\beta^{-1}\Delta\rho_n + \beta^{-1}\partial_c^2\rho_n - \beta^{-1}\partial_c\left(\partial_c m\rho_n\right)$, and the noise terms by $\sqrt{2}(\beta n)^{-1/2}\nabla\cdot\left(\sqrt{\rho_n}\dot{\boldsymbol{\eta}}\right) + \sqrt{2}(\beta n)^{-1/2}\partial_c\left(\sqrt{\rho_n}\dot{\eta}'\right)$. This means that these diffusive and regularizing terms affect the mean field limit equation for $\rho_0(t)$, whereas the noise terms remain higher order. In particular, in that case one can prove that $\rho_n(t) \rightharpoonup \rho_0(t)$ with $\rho_0(t)$ that converges to a unique fixed point ρ_β such that $\rho_\beta > 0$ a.e.in $D\times\mathbb{R}$ but for which $\int_{D\times\mathbb{R}} c\varphi(\cdot,\boldsymbol{y})\rho_\beta(\boldsymbol{y},c)d\boldsymbol{y}dc \neq f$ (there is a correction proportional to β^{-1}). When $\alpha \in (0,1)$, the diffusive and regularizing terms in (A.6) are replaced by $\beta^{-1}n^{-\alpha}\Delta\rho_n + \beta^{-1}n^{-\alpha}\partial_c^2\rho_n - \beta^{-1}n^{-\alpha}\partial_c\left(\partial_c m\rho_n\right)$, and the noise terms by $\sqrt{2}(\beta n^{1+\alpha})^{-1/2}\nabla\cdot\left(\sqrt{\rho_n}\dot{\boldsymbol{\eta}}\right) + \sqrt{2}(\beta n^{1+\alpha})^{-1/2}\partial_c\left(\sqrt{\rho_n}\dot{\eta}'\right)$. 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.

Formal derivation of (A.6). Applying Itô's formula to (2.4) and using (A.2) gives

$$\begin{aligned} d\rho_n(t,\boldsymbol{y},c) = &-\frac{1}{n}\sum_{i=1}^{n}\delta(c-C_i(t))\nabla\delta(\boldsymbol{y}-\boldsymbol{Y}_i(t))\cdot d\boldsymbol{Y}_i(t)\\ &-\frac{1}{n}\sum_{i=1}^{n}\partial_c\delta(c-C_i(t))\delta(\boldsymbol{y}-\boldsymbol{Y}_i(t))dC_i(t)\\ &+\frac{1}{n}\sum_{i=1}^{n}\delta(c-C_i(t))\Delta\delta(\boldsymbol{y}-\boldsymbol{Y}_i(t))(\beta n)^{-1}dt\\ &+\frac{1}{n}\sum_{i=1}^{n}\partial_c^2\delta(c-C_i(t))\delta(\boldsymbol{y}-\boldsymbol{Y}_i(t))(\beta n)^{-1}dt \end{aligned} \tag{A.7}$$

The drift terms in this equation can again be treated as we did to derive (2.6), and this gives the drift terms in (A.6). The noise terms can be analyzed by looking at their quadratic variation. For example, we have

$$\begin{aligned} &\left\langle \frac{1}{n}\sum_{i=1}^{n}\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)d\boldsymbol{W}_i, \frac{1}{n}\sum_{i=1}^{n}\delta(c'-C_i(t))\delta(\boldsymbol{y}'-\boldsymbol{Y}_i(t))d\boldsymbol{W}_i \right\rangle\\ &=\frac{1}{n^2}\sum_{i=1}^{n}\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\delta(c'-C_i)\delta(\boldsymbol{y}'-\boldsymbol{Y}_i)dt\\ &=\frac{1}{n^2}\sum_{i=1}^{n}\delta(c-C_i)\delta(\boldsymbol{y}-\boldsymbol{Y}_i)\delta(c'-c)\delta(\boldsymbol{y}'-\boldsymbol{y})dt\\ &=n^{-1}\rho_n(t,c,\boldsymbol{y})\delta(c'-c)\delta(\boldsymbol{y}'-\boldsymbol{y})dt \end{aligned} \tag{A.8}$$

Proceeding similarly for the other noise term, we see that in law they formally coincide with (i.e. their quadratic variations is the same as) the noise terms in (A.6).

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.6). On long timescales, the memory of the initial conditions is lost, and we can directly pick the right scaling to analyze the fluctuations of $\rho_n(t)$ around its limit $\rho_0(t)$: as discussed in Remark A.1 and confirmed below, this scale is $O(n^{-1})$, consistent with what we reach at long times with GD as discussed in Sec. 2.

We analyze (A.6) by formal asymptotics, 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

$$\rho_n(t) = \rho_0(t,\tau) + n^{-1}\rho_1(t,\tau) + o(n^{-1}), \qquad \tau = t/n. \tag{A.9}$$

We use the rescaled time $\tau = t/n$ to look at the solution to (A.6) on $O(n)$ timescales. Not only does this fix the behavior of $\rho_0(t)$ on long timescales but also guarantees solvability of the equation for $\rho_1(t)$. Treating t and τ as independent variables, (A.9) implies that

$$\partial_t\rho_n = \partial_t\rho_0 + n^{-1}\left(\partial_t\rho_0 + \partial_\tau\rho_1\right) + o(n^{-1}) \tag{A.10}$$

Inserting (A.9) and (A.10) in (A.6) 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

$$\begin{aligned}\partial_t\rho_0 = \nabla\cdot&\left(-c\nabla F\rho_0 + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\rho_0'\rho_0 d\boldsymbol{y}'dc'\right)\\ &+\partial_c\left(-F\rho_0 + \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_0'\rho_0 d\boldsymbol{y}'dc'\right)\end{aligned} \tag{A.11}$$

and

$$\begin{aligned}\partial_\tau\rho_0 + \partial_t\rho_1 = \nabla\cdot&\left(-c\nabla F\rho_1 + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\left(\rho_0'\rho_1 + \rho_1'\rho_0\right)d\boldsymbol{y}'dc'\right)\\ &+\partial_c\left(-F\rho_1 + \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\left(\rho_0'\rho_1 + \rho_1'\rho_0\right)d\boldsymbol{y}'dc'\right)\\ &+\beta^{-1}\Delta\rho_0 - \beta^{-1}\partial_c\left(\partial_c \log m\,\rho_0\right) + \beta^{-1}\partial_c^2\rho_0\\ &+\sqrt{2}\beta^{-1/2}\nabla\cdot\left(\sqrt{\rho_0}\dot{\boldsymbol{\eta}}\right) + \sqrt{2}\beta^{-1/2}\partial_c\left(\sqrt{\rho_0}\dot\eta'\right)\end{aligned} \tag{A.12}$$

A.3. **Law of Large Numbers at finite temperature.** Since (A.11) is identical to (2.10), all the results we established in Sec. 2.3 still hold at finite temperature. In particular, Proposition 2.5 applies. As we see below, we can obtain more information about $\rho_0(t)$ by looking at the evolution of this function on longer timescales, and we will be able to deduce that $\rho_0(\tau) > 0$ a.e. on $D\times\mathbb{R}$. This guarantees that (2.45) holds, so it can be removed from the assumptions needed in Proposition 2.5.

A.4. **Asymptotic behavior of $\rho_0(t)$ on $O(n)$ timescales.** An equation governing the evolution of ρ_0 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 $\rho_0(t,\tau)\to\rho_0(\tau)$ as $t\to\infty$, where $\rho_0(\tau)$ is a stationary point of (A.11), we have

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \rho_0(t,\tau)\rho_1(t,\tau)dt = \rho_0(\tau)\bar\rho_1(\tau) \quad \text{where} \quad \bar\rho_1(\tau) =: \lim_{T\to\infty}\frac{1}{T}\int_0^T\rho_1(t,\tau)dt \tag{A.13}$$

in which we made ρ_1 depends on t and τ for consistency and we assumed that the time-average of ρ_1 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

$$\begin{aligned}\partial_\tau\rho_0 = \nabla\cdot&\left(c\nabla(F_0 - F)\rho_1 + c\nabla F_1\rho_0\right) + \beta^{-1}\Delta\rho_0\\ &+\partial_c\left((F_0-F)\rho_1 + F_1\rho_0 - \beta^{-1}\partial_c\log m\,\rho_0\right) + \beta^{-1}\partial_c^2\rho_0\end{aligned} \tag{A.14}$$

where we defined

$$F_0(\tau,\boldsymbol{y}) = \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_0(\tau,\boldsymbol{y}',c')d\boldsymbol{y}'dc', \tag{A.15}$$

and

$$F_1(\tau,\boldsymbol{y}) = \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\bar\rho_1(\tau,\boldsymbol{y}',c')d\boldsymbol{y}'dc'. \tag{A.16}$$

These two function are continuously differentiable in $\boldsymbol{y}$ by Assumption 2.2. Because of the presence of the diffusive terms $\beta^{-1}\Delta\rho_0 + \beta^{-1}\partial_c^2\rho_0$ in (A.14), we can therefore conclude that on the timescales where this equation holds we must have $\rho_0(\tau) > 0$ a.e. on $D\times\mathbb{R}$. This means that (2.45) holds and so $F_0(\tau,\boldsymbol{y}) = F(\boldsymbol{y})$ since $\int_{D\times\mathbb{R}} c\varphi(\cdot,\boldsymbol{y})\rho_0(\tau)d\boldsymbol{y}dc = f$ by Proposition 2.5. As a result (A.14) reduces to

$$\partial_\tau\rho_0 = \nabla\cdot\left(c\nabla F_1\rho_0\right) + \beta^{-1}\Delta\rho_0 + \partial_c\left(F_1\rho_0 - \beta^{-1}\partial_c\log m\,\rho_0\right) + \beta^{-1}\partial_c^2\rho_0. \tag{A.17}$$

Since (A.16) needs to be satisfied, in (A.17) we can treat the terms involving the factor F_1 as Lagrange multipliers used to enforce this constraint. Within this interpretation, it is easy to see that (A.17) can be written as

$$\partial_\tau\rho_0 = \nabla\cdot\left(\rho_0\nabla\frac{\delta\mathscr{E}_1}{\delta\rho_0}\right) + \partial_c\left(\rho_0\partial_c\frac{\delta\mathscr{E}_1}{\delta\rho_0}\right) \tag{A.18}$$

where we defined

$$\mathscr{E}_1[\rho_0] = \int_{D\times\mathbb{R}}\left(\beta^{-1}\rho_0\log(\rho_0/m) + cF_1(\tau,\boldsymbol{y})\rho_0\right)d\boldsymbol{y}dc. \tag{A.19}$$

A direct consequence of this formulation is that the stable fixed points of (A.18) are the minimizers of the energy (A.19) subject to the constraints that (A.16) holds and $\int_{D\times\mathbb{R}}\rho_0(\boldsymbol{y},c)d\boldsymbol{y}dc = 1$. These fixed points are reached on a timescale that is large compared the $O(n)$ timescale $\tau = t/n$.

Recalling that $\rho_0(t)$ is the weak limit of $\rho_n(t)$ as $n\to\infty$, we can summarize these considerations into:

Proposition A.2. *If $\rho_n(t)$ be the empirical distribution defined in* (2.4) *with $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ solution to* (A.6). *Then given any $b_n > 0$ such that $b_n/n\to\infty$ as $n\to\infty$, we have*

$$\rho_n(b_n) \rightharpoonup \rho_0^* \qquad as \quad n\to\infty \tag{A.20}$$

where ρ_0^* is the minimizer of

$$\beta^{-1}\int_{D\times\mathbb{R}}\rho_0\log(\rho_0/m)d\boldsymbol{y}dc \tag{A.21}$$

subject to

$$\int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_0(\boldsymbol{y}',c')d\boldsymbol{y}'dc' = F(\boldsymbol{y}) \quad a.e.\ in\ D \quad and \quad \int_{D\times\mathbb{R}}\rho_0 d\boldsymbol{y}dc = 1. \tag{A.22}$$

If we denote by $\rho_0^*(\boldsymbol{y},c)$ the minimizer of (A.21) subject to (A.22) and by $F_1^*(\boldsymbol{y})$ the Lagrange multiplier used to satisfy the first constraint in (A.22), this Lagrange multiplier is given by

$$F_1^*(\boldsymbol{y}) = \beta^{-1}\frac{\delta}{\delta F(\boldsymbol{y})}\int_{D\times\mathbb{R}}\rho_0^*\log(\rho_0^*/m)d\boldsymbol{y}dc \tag{A.23}$$

It is easy to see that ρ_0^* is independent of β: indeed, we can drop the factor β^{-1} in front of (A.21) without affecting the minimization problem. This also means that the dependency of F_1^* in β is explicit: Indeed from (A.23)

$$F_1^*(\boldsymbol{y}) = \beta^{-1}\delta^*(\boldsymbol{y}) \tag{A.24}$$

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

$$\delta^*(\boldsymbol{y}) = \frac{\delta}{\delta F(\boldsymbol{y})}\int_{D\times\mathbb{R}}\rho_0^*\log(\rho_0^*/m)d\boldsymbol{y}dc \tag{A.25}$$

This factor is independent of β since ρ_0^* is.

It is also easy to see that the solution to the minimization problem in Proposition A.2 is such that

$$\int_{D\times\mathbb{R}}\rho_0^*\log(\rho_0^*/m)d\boldsymbol{y}dc < \infty \qquad \text{and} \qquad \rho_0^* > 0 \quad \text{a.e. in} \quad D\times\mathbb{R}. \tag{A.26}$$

Indeed, any $\rho_0 > 0$ such that $\int_{D\times\mathbb{R}} \rho_0(\boldsymbol{y},c)d\boldsymbol{y}dc = 1$ and $\int_{\mathbb{R}} c\rho_0(\boldsymbol{y},c)dc = G_0^*(\boldsymbol{y})$ where $G_0^*(\boldsymbol{y})$ solves (2.41) satisfies the constraints in (A.22). One such ρ_0 is

$$\rho_0(\boldsymbol{y},c) = m\left(c - G_0^*(\boldsymbol{y})\right), \tag{A.27}$$

for which we have

$$\begin{aligned} &\beta^{-1}\int_{D\times\mathbb{R}} \rho_0 \log(\rho_0/m) d\boldsymbol{y}dc \\ &= \beta^{-1}\int_{D\times\mathbb{R}} m(c)\left(\log m(c) - \log m\left(c + G_0^*(\boldsymbol{y})\right)\right) d\boldsymbol{y}dc \end{aligned} \tag{A.28}$$

which is finite since $m(c)$ decreases sufficiently fast as $c \to \infty$ by Assumption A.1. For example, if $m(c) = \exp(-\frac{1}{2}\lambda c^2)\sqrt{\lambda/(2\pi)}$ where $\lambda > 0$, (A.28) becomes

$$\beta^{-1}\int_{D\times\mathbb{R}} \rho_0 \log(\rho_0/m) d\boldsymbol{y}dc = \tfrac{1}{2}\lambda\beta^{-1}\int_D |G_0^*(\boldsymbol{y})|^2 d\boldsymbol{y} < \infty \tag{A.29}$$

The actual minimizer of (A.21) subject to (A.22) must do at least as well as this. To prove that $\rho_0^* > 0$, suppose by contradiction that the minimizer is such that $\rho_0^* = 0$ if $(\boldsymbol{y},c) \in B$ with $\int_B m(c)d\boldsymbol{y}dc > 0$. For $s \in [0,1]$, consider $\rho_0^s = (1-s)\rho_0^* + sm(c)/|D|$ where $|D|$ denotes the volume of D. A direct calculation shows that

$$\int_{D\times\mathbb{R}} \rho_0^s \log(\rho_0^s/m) d\boldsymbol{y}dc = \int_B \rho_0^* \log(\rho_0^*/m) d\boldsymbol{y}dc + s\log s|D|^{-1}\int_{B^c} m(c)d\boldsymbol{y}dc + O(s) \tag{A.30}$$

Since $s\log s|D|^{-1}\int_{B^c} m(c)d\boldsymbol{y}dc < 0$ for $s \in (0,1)$, (A.30) implies that for $s > 0$ small enough

$$\int_{D\times\mathbb{R}} \rho_0^s \log(\rho_0^s/m) d\boldsymbol{y}dc < \int_B \rho_0^* \log(\rho_0^*/m) d\boldsymbol{y}dc \tag{A.31}$$

which contradicts the fact that ρ_0^* was the minimizer.

Remark A.2. Compared to the case treated in [22] 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 $\rho_0(t)$ converges to a density that is positive a.e in $D\times\mathbb{R}$ as $t \to \infty$; however, the fixed point ρ_0^* we obtain satisfies $\int_{D\times\mathbb{R}} c\varphi(\cdot,\boldsymbol{y})\rho_0^* d\boldsymbol{y}dc = f$, 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 $\rho_0(0) = \rho_{\text{in}}$ is such that (2.45) 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 $\lim_{n\to\infty}\rho_n = \rho_0$ at all times, we can stop distinguishing τ from t, and focus on ρ_1. We already know that (A.15) constrain the average value of ρ_1 on long timescales, but we would also like to quantify this average value beyond (A.15) as well as the fluctuations around it. To this end, let us use (A.17) in (A.12) and look at the resulting equation on timescales where $\rho_0(t)$ has converged to ρ_0^*, the minimizer specified in Proposition A.2, so that $F_0(t)$ has also converged to F and $F_1(t)$ to $F_1^* = \beta^{-1}\delta^*$. This can be achieved by considering (A.12) with initial condition at $t = T$ and pushing back $T \to -\infty$. The resulting equation is

$$\begin{aligned} \partial_t\rho_1 = \nabla\cdot&\left(-\beta^{-1}c\nabla\delta^*\rho_0^* + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\rho_1'\rho_0^* d\boldsymbol{y}'dc'\right) \\ &+ \partial_c\left(-\beta^{-1}\delta^*\rho_0^* + \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_1'\rho_0^* d\boldsymbol{y}'dc'\right) \\ &+ \sqrt{2}\beta^{-1/2}\nabla\cdot\left(\sqrt{\rho_0^*}\dot{\boldsymbol{\eta}}\right) + \sqrt{2}\beta^{-1/2}\partial_c\left(\sqrt{\rho_0^*}\dot{\eta}'\right) \end{aligned} \tag{A.32}$$

Even though we derived it formally, the SPDE (A.42) can be given a precise meaning: since its drift is linear in ρ_1 and its noise is additive (recall that ρ_0^* is a given, non-random, function), (A.42) defines ρ_1 as a Gaussian process. This also means that

$$f_1(t,\boldsymbol{x}) = \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_1(t,\boldsymbol{y},c)d\boldsymbol{y}dc. \tag{A.33}$$

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

$$f_n(t,\boldsymbol{x}) = f_0(t,\boldsymbol{x}) + n^{-1} f_1(t,\boldsymbol{x}) + o(n^{-1}) \tag{A.34}$$

Let us derive a closed equation for $f_1(t,\boldsymbol{x})$ from (A.32). To this end, notice first that we can use

$$\int_{D\times\mathbb{R}} c' K(\boldsymbol{y},\boldsymbol{y}')\rho_1(t,\boldsymbol{y}',c')d\boldsymbol{y}'dc' = \int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y}) f_1(t,\boldsymbol{x})d\mu(\boldsymbol{x}) \tag{A.35}$$

to express the integral terms in (A.32) in terms of $f_1(t)$. Let us also define $\epsilon^*(\boldsymbol{x})$ via the equation

$$\int_D \varphi(\boldsymbol{x},\boldsymbol{y})\delta^*(\boldsymbol{y})d\boldsymbol{y} = \int_D\int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y})\varphi(\boldsymbol{x}',\boldsymbol{y})\epsilon^*(\boldsymbol{x}')d\mu(\boldsymbol{x}')d\boldsymbol{y} \tag{A.36}$$

This is the Euler-Lagrange equation for the minimizer of

$$\tfrac{1}{2}\int_D \left|\delta^*(\boldsymbol{y}) - \int_\Omega \epsilon(\boldsymbol{x})\varphi(\boldsymbol{x},\boldsymbol{y})d\mu(\boldsymbol{x})\right|^2 d\boldsymbol{y} \tag{A.37}$$

over ϵ. Therefore, (A.36) is also the equation for the least square solution of

$$\delta^*(\boldsymbol{y}) = \int_\Omega \epsilon^*(\boldsymbol{x})\varphi(\boldsymbol{x},\boldsymbol{y})d\mu(\boldsymbol{x}) \tag{A.38}$$

and we know that such a least square solution exists for a modification of $\delta^*(\boldsymbol{y})$ which is arbitrarily close to it in $L^2(D)$: any such solution for a modification of $\delta^*(\boldsymbol{y})$ 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 1.2 and it can be expressed as

$$\epsilon^*(\boldsymbol{x}) = D_{f(\boldsymbol{x})}\int_{D\times\mathbb{R}} \rho_0^* \log(\rho_0^*/m)d\boldsymbol{y}dc \tag{A.39}$$

where ρ_0^* is viewed as a functional of $f(\boldsymbol{x})$ by expressing $F(\boldsymbol{y})$ as $F(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x},\boldsymbol{y})d\mu(\boldsymbol{x})$, and $D_{f(\boldsymbol{x})}$ denotes the gradient with respect to $f(\boldsymbol{x})$ in the $L^2(\Omega,\mu)$-norm defined in (2.67). The equality (A.39) follows from (A.25) and the fact that $D_{f(\boldsymbol{x})}F(\boldsymbol{y}) = \varphi(\boldsymbol{x},\boldsymbol{y})$.

By taking the time derivative of (2.69) and using (A.32) together with (A.35) and (A.36) we derive:

$$\begin{aligned}
\partial_t f_1 &= \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\partial_t\rho_1(t,\boldsymbol{y},c)d\boldsymbol{y}dc\\
&= \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\nabla\cdot\left(\left(-\beta^{-1}c\nabla\delta^*(\boldsymbol{y}) + c\int_\Omega \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})f_1(t,\boldsymbol{x})d\mu(\boldsymbol{x})\right)\rho_0^*\right)d\boldsymbol{y}dc\\
&\quad + \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\partial_c\left(\left(-\beta^{-1}\delta^*(\boldsymbol{y}) + \int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y})f_1(t,\boldsymbol{x})d\mu(\boldsymbol{x})\right)\rho_0^*\right)d\boldsymbol{y}dc\\
&\quad + \sqrt{2}\beta^{-1/2}\int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\left(\nabla\cdot\left(\sqrt{\rho_0^*}\dot{\boldsymbol{\eta}}\right) + \partial_c\left(\sqrt{\rho_0^*}\dot{\eta}'\right)\right)d\boldsymbol{y}dc\\
&= \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\nabla\cdot\left(c\int_\Omega \nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\left(f_1(t,\boldsymbol{x}) - \beta^{-1}\epsilon^*(\boldsymbol{x})\right)d\mu(\boldsymbol{x})\rho_0^*\right)d\boldsymbol{y}dc\\
&\quad + \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\partial_c\left(\int_\Omega \varphi(\boldsymbol{x},\boldsymbol{y})\left(f_1(t,\boldsymbol{x}) - \beta^{-1}\epsilon^*(\boldsymbol{x})\right)d\mu(\boldsymbol{x})\rho_0^*\right)d\boldsymbol{y}dc\\
&\quad + \sqrt{2}\beta^{-1/2}\int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\left(\nabla\cdot\left(\sqrt{\rho_0^*}\dot{\boldsymbol{\eta}}\right) + \partial_c\left(\sqrt{\rho_0^*}\dot{\eta}'\right)\right)d\boldsymbol{y}dc
\end{aligned} \tag{A.40}$$

By performing integration by parts in $\boldsymbol{y}$ and c on the noise term in this equation, we deduce that its quadratic variation is

$$2\beta^{-1}M([\rho_0^*],\boldsymbol{x},\boldsymbol{x}')dt \tag{A.41}$$

where $M([\rho],\boldsymbol{x},\boldsymbol{x}')$ is the kernel defined in (2.59). Similarly, by integrating by parts in $\boldsymbol{y}$ the first term at the right hand side of (A.40) and in c the second, and interchanging the order of integration between $(\boldsymbol{y},c)$ and $\boldsymbol{x}$ on both these terms, we can write this equation as

$$\begin{aligned}\partial_t f_1 = &-\int_\Omega M([\rho_0^*],\boldsymbol{x},\boldsymbol{x}')\big(f_1(t,\boldsymbol{x}') - \beta^{-1}\epsilon^*(\boldsymbol{x})\big)\,d\mu(\boldsymbol{x}')\\ &+\sqrt{2}\beta^{-1/2}\int_\Omega \sigma([\rho_0^*],\boldsymbol{x},\boldsymbol{x}')\dot\eta(t,\boldsymbol{x}')d\boldsymbol{x}'\end{aligned}\tag{A.42}$$

where $\dot\eta(t,\boldsymbol{x})$ is a spatio-temporal white-noise, and $\sigma([\rho_0^*],\boldsymbol{x},\boldsymbol{x}')$ is such that

$$\int_\Omega \sigma([\rho_0^*],\boldsymbol{x},\boldsymbol{x}'')\sigma([\rho_0^*],\boldsymbol{x}',\boldsymbol{x}'')d\boldsymbol{x}'' = M([\rho_0^*],\boldsymbol{x},\boldsymbol{x}')\tag{A.43}$$

Note that this decomposition exists since $M([\rho_0^*],\boldsymbol{x},\boldsymbol{x}')$ is positive-definite. The asymptotic mean and variance of $f_1(t)$ can be readily deduced from (A.42) knowing the asymptotic behavior of $\bar f_1(t)$. We can state this as

Proposition A.3 (CLT at finite temperature)**.** *Let $f_n(t) = f_n(t,\boldsymbol{x})$ be given by* (2.5) *with $\{\boldsymbol{Y}_i(t),C_i(t)\}_{i=1}^n$ solution to* (A.6) *with initial conditions specified at $t = T$. Then*

$$\lim_{T\to-\infty}\lim_{n\to\infty} n\big(f_n(t) - f\big) = f_1(t) \qquad \textit{in law}\tag{A.44}$$

where $f_1(t)$ is the Gaussian process specified by (A.42) *and whose mean an covariance satisfy for any $\chi\in L^2(\Omega,\mu)$*

$$\begin{aligned}&\lim_{t\to\infty}\mathbb{E}\int_\Omega \chi(\boldsymbol{x})f_1(t,\boldsymbol{x})d\mu(\boldsymbol{x}) = \beta^{-1}\int_\Omega\chi(\boldsymbol{x})\epsilon^*(\boldsymbol{x})d\mu(\boldsymbol{x})\\ &\lim_{t\to\infty}\mathbb{E}\left(\int_\Omega\chi(\boldsymbol{x})\big(f_1(t,\boldsymbol{x}) - \beta^{-1}\epsilon^*(\boldsymbol{x})\big)\,d\mu(\boldsymbol{x})\right)^2 = \beta^{-1}\int_\Omega|\chi(\boldsymbol{x})|^2d\mu(\boldsymbol{x})\end{aligned}\tag{A.45}$$

where ϵ^* is given by (A.39)

Notice that if we quench the result in (A.45) (i.e. send $\beta\to\infty$), we arrive at the conclusion that $f_1(t)\to 0$ 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 2.6.

References

[1] A. Auffinger and G. Ben Arous. Complexity of random smooth functions on the high-dimensional sphere. *The Annals of Probability*, 41(6):4214–4247, Nov. 2013.

[2] A. Auffinger, G. Ben Arous, and J. Černý. Random Matrices and Complexity of Spin Glasses. *Comm. Pure Appl. Math.*, 66(2):165–201, 2012.

[3] M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. Ben Arous, C. Cammarota, Y. LeCun, M. Wyart, and G. Biroli. Comparing Dynamics: Deep Neural Networks versus Glassy Systems. *arXiv:1803.06969*, Mar. 2018.

[4] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

[5] C. Beck, W. E, and A. Jentzen. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *arXiv:1709.05963*, Sept. 2017.

[6] J. Behler and M. Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):583, Apr. 2007.

[7] J. Berg and K. Nyström. A unified deep artificial neural network approach to partial differential equations in complex geometries. *arXiv:1711.06464*, Nov. 2017.

[8] L. Bottou and Y. L. Cun. Large scale online learning. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 217–224. MIT Press, 2004.

[9] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The Loss Surfaces of Multilayer Networks. *arXiv:1412.0233*, Nov. 2014.

[10] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2(4):303–314, Dec. 1989.

[11] D. S. Dean. Langevin equation for the density of a system of interacting Langevin processes. *J. Phys. A: Math. Gen.*, 29(24):L613–L617, Jan. 1999.

[12] W. E, J. Han, and A. Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *arXiv:1706.04702*, June 2017.

[13] C. W. Groetsch. *The theory of Tikhonov regularization for Fredholm equations of the first kind*, volume 105 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1984.

[14] W. Hu, C. J. Li, L. Li, and J.-G. Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv:1705.07562*, May 2017.

[15] R. Jordan, D. Kinderlehrer, and F. Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, Jan. 1998.

[16] Y. Khoo, J. Lu, and L. Ying. Solving for high dimensional committor functions using artificial neural networks. *arXiv:1802.10275*, Feb. 2018.

[17] C. Kipnis and C. Landim. *Scaling limits of interacting particle systems*, volume 320. Springer Science & Business Media, 2013.

[18] T. Leblé and S. Serfaty. Large deviation principle for empirical fields of log and riesz gases. *Inventiones mathematicae*, 210(3):645–757, 2017.

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[20] Q. Li, C. Tai, and W. E. Dynamics of stochastic gradient algorithms. *CoRR*, abs/1511.06251, 2015.

[21] H. P. McKean. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, Dec. 1966.

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

[23] J. Park and I. W. Sandberg. Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, 3(2):246–257, June 1991.

[24] L. Sagun, V. U. Guney, G. Ben Arous, and Y. LeCun. Explorations on high dimensional landscapes. *arXiv:1412.6615*, Dec. 2014.

[25] E. Schneider, L. Dai, R. Q. Topper, C. Drechsel-Grau, and M. E. Tuckerman. Stochastic Neural Network Approach for Learning High-Dimensional Free Energy Surfaces. *Phys. Rev. Lett.*, 119(15):150601, Oct. 2017.

[26] S. Serfaty. *Coulomb gases and Ginzburg-Landau vortices* . Zurich Lectures in Advanced Mathematics. European Mathematical Society Publishing House, 2015.

[27] S. Serfaty. Systems of Points with Coulomb Interactions. *arXiv:1712.04095*, Dec. 2017.

[28] S. Serfaty. Mean Field Limit for Coulomb Flows. *arXiv.org*, Mar. 2018.

[29] J. Sirignano and K. Spiliopoulos. Mean Field Analysis of Neural Networks. *arXiv.org*, May 2018.

[30] L. Zhang, J. Han, H. Wang, R. Car, and W. E. DeePCG: constructing coarse-grained models via deep neural networks. *arXiv:1802.08549*, Feb. 2018.

COURANT INSTITUTE OF MATHEMATICAL SCIENCES, NEW YORK UNIVERSITY, 251 MERCER STREET, NEW YORK, NY 10012

E-mail address: rotskoff@cims.nyu.edu, eve2@cims.nyu.edu