Analysis of a Two-Layer Neural Network via Displacement Convexity

Adel Javanmard[*] Marco Mondelli[†] and Andrea Montanari[‡]

August 20, 2019

Abstract

Fitting a function by using linear combinations of a large number N of 'simple' components is one of the most fruitful ideas in statistical learning. This idea lies at the core of a variety of methods, from two-layer neural networks to kernel regression, to boosting. In general, the resulting risk minimization problem is non-convex and is solved by gradient descent or its variants. Unfortunately, little is known about global convergence properties of these approaches.

Here we consider the problem of learning a concave function f on a compact convex domain $\Omega \subset \mathbb{R}^d$, using linear combinations of 'bump-like' components (neurons). The parameters to be fitted are the centers of N bumps, and the resulting empirical risk minimization problem is highly non-convex. We prove that, in the limit in which the number of neurons diverges, the evolution of gradient descent converges to a Wasserstein gradient flow in the space of probability distributions over Ω. Further, when the bump width δ tends to 0, this gradient flow has a limit which is a viscous porous medium equation. Remarkably, the cost function optimized by this gradient flow exhibits a special property known as *displacement convexity*, which implies exponential convergence rates for $N \to \infty$, $\delta \to 0$.

Surprisingly, this asymptotic theory appears to capture well the behavior for moderate values of δ, N. Explaining this phenomenon, and understanding the dependence on δ, N in a quantitative manner remains an outstanding challenge.

1 Introduction

In supervised learning, we are given data $\{(y_j, \boldsymbol{x}_j)\}_{j \le n}$ which are often assumed to be independent and identically distributed from a common law $\mathbb{P}$ on $\mathbb{R} \times \mathbb{R}^d$ (here $\boldsymbol{x}_j \in \mathbb{R}^d$ is a feature vector, and $y_j \in \mathbb{R}$ is a label or response variable). We would like to find a function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ to predict the labels at new points $\boldsymbol{x} \in \mathbb{R}^d$. Throughout this paper, we will quantify the quality of our prediction by square loss, hence we are interested in minimizing $R(\hat{f}) = \mathbb{E}\{(y - \hat{f}(\boldsymbol{x}))^2\}$.

One of the most fruitful ideas in this context is to use functions that are linear combinations of

[*]Data Science and Operations Department, Marshall School of Business, University of Southern California
[†]Department of Electrical Engineering, Stanford University
[‡]Department of Electrical Engineering and Department of Statistics, Stanford University

simple components:

$$\hat{f}(\boldsymbol{x};\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma(\boldsymbol{x};\boldsymbol{w}_i)\,. \tag{1.1}$$

Here $\sigma : \mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$ is a component function (a 'neuron' or 'unit' in the neural network parlance), and $\boldsymbol{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_N) \in \mathbb{R}^{D\times N}$, $\boldsymbol{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$ are parameters to be learnt from data. Standard choices for the activation function are $\sigma(\boldsymbol{x};\boldsymbol{w}) = (1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x}\rangle))^{-1}$ (sigmoid) or $\sigma(\boldsymbol{x};\boldsymbol{w}) = \max(\langle \boldsymbol{w}, \boldsymbol{x}\rangle; 0)$ (ReLU). In this paper we will instead study a class of activation that depends on the difference $\boldsymbol{x} - \boldsymbol{w}$. The objective is to minimize the population (prediction) risk

$$R_N(\boldsymbol{a},\boldsymbol{w}) = \mathbb{E}\Big\{\Big[y - \frac{1}{N}\sum_{i=1}^{N} a_i \sigma(\boldsymbol{x};\boldsymbol{w}_i)\Big]^2\Big\}\,. \tag{1.2}$$

Special instantiations of this idea include (we provide only pointers to the immense literature on each topic):

- Two-layer neural networks [Ros62, AB09];
- Sparse deconvolution [Don92, CFG14];
- Kernel ridge regression and related random feature methods [CST00, RR08];
- Boosting [Sch03, Fri01, BY03].

Despite the impressive practical success of these methods, the risk function $R_N(\boldsymbol{w})$ is highly non-convex and little is known about global convergence of algorithms that try to minimize it (we refer to Section 2 for further discussion of the related literature).

Notable exceptions to the last statement are provided by random features and by boosting algorithms. In random feature methods, the parameters $\boldsymbol{w}_i$ are not optimized over (they are drawn i.i.d. from some common distribution), and the resulting risk function becomes convex in the weights $(a_1, \dots, a_N)$ to be learnt. While this is a fruitful idea, it gives up the degrees of freedom afforded by the $\boldsymbol{w}_i$'s.

Boosting overcomes non-convexity by fitting the components $\boldsymbol{w}_1$, ..., $\boldsymbol{w}_N$ one at the time, sequentially. The underlying assumption is that the problem of minimizing $R_N(\boldsymbol{w})$ with respect to one of the hidden units $\boldsymbol{w}_i$ is tractable. However, this is generally not the case when the parameters $\boldsymbol{w}_i$ belong to a high-dimensional space.

The risk function (1.2) crystalizes a central conundrum in statistical learning. In a number of applications (especially at low noise), it is rarely the case that low prediction error can be achieved through a function that is linear in the raw covariates, e.g. $\hat{f}(x) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$. In a classical setting, the statistician would craft nonlinear features out of the covariates on the basis of expert knowledge. For the model of Eq. (1.1), this amounts to constructing vectors $\boldsymbol{w}_1, \dots, \boldsymbol{w}_N$. Statistical methods would then be confined to the convex task of fitting the coefficients $a_1, \dots, a_N$. This step is well understood from a statistical and computational perspective.

Modern machine learning approaches (boosting, neural networks, etc.) hold the promise of automatizing feature extraction, hence producing superior performances in a wide variety of applications. Unfortunately, we are still far from understanding in which cases optimizing over the

$\boldsymbol{w}_i$'s yields a significant improvement over –say– choosing them randomly. This central challenge intertwines statistical and computational aspects. It is not hard to see that varying the weights $\boldsymbol{w}_i$'s produces a significantly larger function class [Bac17]. The relevant question is what part of this class can be accessed using gradient descent or other practical algorithms.

The main objective of this paper is to introduce a nonparametric regression model in which these questions can be addressed rigorously. The model is interesting for at least two reasons: (*i*) From a theoretical point of view, global convergence can be proved in the limit of a large neurons. The proof relies on a mathematical mechanism that has not been explored in the statistics or machine learning literature before. (*ii*) From a practical point of view, the model is nontrivial enough to illustrate the potential advantage of fitting the features $\boldsymbol{w}_i$ (we demonstrate this numerically in Section 4.)

Let $\Omega \subset \mathbb{R}^d$ be a compact convex set with $\mathscr{C}^2$ boundary. We assume $\{(y_j, \boldsymbol{x}_j)\}_{j\ge 1}$ to be i.i.d. where $\boldsymbol{x}_j \sim \mathsf{Unif}(\Omega)$ and

$$\mathbb{E}(y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j)\,, \tag{1.3}$$

with $f : \Omega \to \mathbb{R}$ a smooth function. We try to fit these data using a combination of bumps, namely

$$\hat{f}(\boldsymbol{x}; \boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} K^{\delta}(\boldsymbol{x} - \boldsymbol{w}_i)\,, \tag{1.4}$$

where $K^{\delta}(\boldsymbol{x}) = \delta^{-d}K(\boldsymbol{x}/\delta)$, $K : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ is a first order kernel with compact support, and $\boldsymbol{w}_i \in \Omega^{\delta}$ for $i \le N$. Here Ω^{δ} is a slightly smaller compact set, with $\Omega^{\delta} \to \Omega$ as $\delta \to 0$. (Note that in our setting the hidden units $\boldsymbol{w}_i$ and input data $\boldsymbol{x}_j$ have same dimensions, i.e., $d = D$.) We refer to Section 5 for a formal statement of our assumptions. From Eq. (1.2), we have

$$R_N(\boldsymbol{w}) = R_{\#} + \mathbb{E}\big\{\big[f(\boldsymbol{x}) - \frac{1}{N}\sum_{i=1}^{N} K^{\delta}(\boldsymbol{x} - \boldsymbol{w}_i)\big]^2\big\}\,,$$

where $R_{\#} = \mathbb{E}[(y - f(\boldsymbol{x}))^2]$ and we use the fact that $\mathbb{E}[y - f(\boldsymbol{x})|\boldsymbol{x}] = 0$. Since the constant $R_{\#}$ does not depend on parameters $\boldsymbol{w}$, it does not matter in optimizing $R_N(\boldsymbol{w})$ over $\boldsymbol{w}$ and henceforth we write, with a slight abuse of notation,

$$R_N(\boldsymbol{w}) = \mathbb{E}\big\{\big[f(\boldsymbol{x}) - \frac{1}{N}\sum_{i=1}^{N} K^{\delta}(\boldsymbol{x} - \boldsymbol{w}_i)\big]^2\big\}\,.$$

The model (1.4) is general enough to include a broad class of radial-basis function (RBF) networks which are known to be universal function approximators [PS91]. To the best of our knowledge, there is no result on the global convergence of stochastic gradient descent for learning RBF networks, and this paper establishes the first result of this type.

It is important to emphasize a few differences with respect to standard RBF networks. First of all, we do not require the kernel $K(\boldsymbol{x})$ to be radial, i.e. to depend uniquely on the norm $|\boldsymbol{x}|$. Second, we require K to have compact support. This is mainly a technical requirement that simplifies some arguments: we expect our results to be generalizable to kernels that decay rapidly enough. Finally, and most crucially, the form (1.4) does not include non-uniform weights for the N components.

A more standard formulation would posit $\hat{f}(\boldsymbol{x};\boldsymbol{w}) = \sum_{i=1}^N a_i K^\delta(\boldsymbol{x}-\boldsymbol{w}_i)$ and learn the weights a_i from data, see Eq. (1.1). We deliberately set the weights to a fixed value because the risk function is convex in $\boldsymbol{a} = (a_i)_{i\le N}$, and hence fitting $\boldsymbol{a}$'s to global optimality is 'easy.' Indeed, universal approximation could be achieved by keeping the centers $\boldsymbol{w}_i$ fixed (and sufficiently dense in Ω) and only adjusting $\boldsymbol{a}$. As discussed above, our focus is on the role of the $\boldsymbol{w}_i$'s.

Our main result is a proof that, for sufficiently large N and small δ, gradient descent algorithms converge to weights $\boldsymbol{w}$ with nearly optimum prediction error, provided f is strongly concave. Let us emphasize that the resulting population risk $R_N(\boldsymbol{w})$ is non-convex regardless of the concavity properties of f. Our proof unveils a novel mechanism by which global convergence takes place. Convergence results for non-convex empirical risk minimization are generally proved by carefully ruling out local minima in the cost function (see Section 2 for pointers to this literature). Instead we prove that, as $N\to\infty$, $\delta\to 0$, the gradient descent dynamics converges to a gradient flow in Wasserstein space, and that the corresponding cost function is 'displacement convex.' Breakthrough results in optimal transport theory guarantee dimension-free convergence rates for this limiting dynamics [CJM+01, CMV03, CMV06]. In particular, we expect the cost function $R_N(\boldsymbol{w})$ to have many local minima, which are however completely neglected by the gradient descent dynamics.

More specifically, our first step is to show that – for large N – the evolution of the weights $\boldsymbol{w}_1,\dots,\boldsymbol{w}_N$ under gradient descent can be replaced by the evolution of a probability distribution[1] $\rho^\delta\in\mathscr{P}_2(\Omega)$, which approximates their empirical distribution. Namely, if $(\boldsymbol{w}_1^k,\dots,\boldsymbol{w}_N^k)$ denote the weights after k iterations with step size ε, and $\hat{\rho}_k^{(N)} = \sum_{i=1}^N \delta_{\boldsymbol{w}_i^k}/N$ is their empirical distribution, then we have

$$\lim_{N\to\infty,\varepsilon\to 0} \hat{\rho}^{(N)}_{t/\varepsilon} = \rho^\delta_t\,, \tag{1.5}$$

where the limit holds in the sense of weak convergence or in W_1 distance (the two are equivalent since Ω is compact). The limit evolution $(\rho^\delta_t)_{t\ge 0}$ satisfies a partial differential equation (PDE) that can also be described as the Wasserstein W_2 gradient flow (i.e. gradient flow in $\mathscr{P}_2(\Omega)$), for the following effective risk

$$R^\delta(\rho) = \nu_0\int_\Omega \big[f(\boldsymbol{x}) - K^\delta * \rho(\boldsymbol{x})\big]^2 \mathrm{d}\boldsymbol{x}\,, \tag{1.6}$$

where $\nu_0 = 1/|\Omega|$ and $|\Omega|$ denotes the volume of the set Ω. Here $*$ denotes the usual convolution. Let us emphasize that the convergence to Wasserstein gradient flow holds regardless of the concavity of f.

The use of W_2 gradient flows to analyze two-layer neural networks was recently developed in several papers [MMN18, RVE18, CB18, SS18]. However, we cannot rely on earlier results because of the specific boundary conditions in our problem. We constrain the $\boldsymbol{w}_i\in\Omega^\delta$ by running projected stochastic gradient descent (SGD): at each step $\boldsymbol{w}_i$ moves in the direction of a stochastic gradient of $R_N(\boldsymbol{w})$ and then projected back to Ω^δ. This results in a PDE with Neumann boundary condition on Ω^δ, which is not covered by previous theory. We establish a quantitative version of the limit (1.5) via propagation-of-chaos techniques.

Even if the cost (1.6) is quadratic and convex in ρ, its W_2 gradient flow can have multiple fixed points, and hence global convergence cannot be guaranteed. Global convergence results were

[1] Throughout, $\mathscr{P}_2(\mathcal{X})$ denotes the space of probability distributions on $\mathcal{X}$, endowed with Wasserstein metric W_2.

proven in [MMN18] and in [CB18] by showing that, for all $t \geq 0$ ρ_t^δ has a density that is either smooth, or strictly positive everywhere. However, these convergence results are non-quantitative, and do not provide convergence rates[2].

Indeed, the mathematical property that controls global convergence of W_2 gradient flow is not ordinary convexity but *displacement convexity*. Roughly speaking, displacement convexity is convexity along geodesics of the W_2 metric, see Section 3.5. The risk function (1.6) is not displacement convex. Indeed, its quadratic term reads $\nu_0 \int K_\delta * K_\delta(\boldsymbol{x}-\boldsymbol{x}')\rho(\boldsymbol{x})\rho(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'$ which is not displacement convex unless $K_\delta * K_\delta$ is convex (see Lemma H.1), which cannot be in our setting. However, for small δ, we can formally approximate $K^\delta * \rho \approx \rho$, and hence hope to replace the risk function (1.6) with a simpler one

$$R(\rho) = \nu_0 \int_\Omega \big[f(\boldsymbol{x}) - \rho(\boldsymbol{x})\big]^2 \mathrm{d}\boldsymbol{x}\,. \tag{1.7}$$

Most of our technical work is devoted to making rigorous this $\delta \to 0$ approximation. Namely, we prove that, as $\delta \to 0$, $\rho_t^\delta \Rightarrow \rho_t$ where ρ_t follows the W_2 gradient flow for the risk $R(\rho)$.

Remarkably, the risk function $R(\rho)$ is strongly displacement convex (provided f is strongly concave). A long line of work in PDE and optimal transport theory establishes dimension-free convergence rates for its W_2 gradient flow [CJM+01, CMV03, CMV06]. Namely, if f is α-strongly concave, then $R(\rho_t) \leq R(\rho_0)\, e^{-2\alpha t}$. By using the approximation results outlined above, we obtain global convergence for SGD. With high probability,

$$R_N(\boldsymbol{w}^k) \leq R_N(\boldsymbol{w}^0)\, e^{-2\alpha k\varepsilon} + \mathsf{err}(N, d, \varepsilon, \delta)\,, \tag{1.8}$$

where the error term err vanishes as $N \to \infty$, $\varepsilon, \delta \to 0$ in a suitable order.

This result implies that SGD converges exponentially fast to a near-global optimum with a rate that is controlled by the convexity parameter α.

Our bounds are not sharp enough to provide quantitative control on the error term $\mathsf{err}(N, d, \varepsilon, \delta)$, especially in high dimension. Nevertheless, the convergence rate predicted by our asymptotic theory is in excellent agreement with numerical simulations, cf. Section 4. Explaining this surprising quantitative agreement is an outstanding challenge.

2 Related literature

The present work ties in several lines of research, some of which were already mentioned in the introduction. A substantial amount of work has been devoted to analyzing two-layer neural networks and developing algorithms with convergence guarantees, see e.g. [ZSJ+17, Tia17, BJW18]. However these approaches are typically based on tensor factorization or similar initialization steps that are not used in practice, and do not scale well (although polynomially) in high dimension.

The landscape of empirical risk minimization was also studied in a number of papers, see e.g. [LY17, SJL18]. However, global convergence was only proved in the extremely overparametrized regime in which the neural network essentially behaves as kernel ridge regression [DZPS18].

[2] An argument indicating convergence in a time polynomial in d was put forward in [WLLM18], but for a different type of continuous flow.

Classical theory of neural networks was largely devoted to the two-layer case [AB09], although the focus was on representation and approximation questions [Cyb89, Bar93], as well as on generalization error. It was already clear in that context that a two-layer network is conveniently characterized by the empirical distribution of the hidden neurons, and that it is useful to relax this from a distribution with N atoms, to a general probability measure. This representation plays an important role, for instance, in [Bar98], and was exploited again under the label of 'convex neural networks' in [BRV$^+$06].

Over the last year, several groups independently revisited this connection, with the objective of understanding the landscape structure of two-layer networks, and the dynamics of gradient descent methods [NS17, MMN18, RVE18, SS18, CB18, MMM19]. In particular, it was proven in [MMN18] that, under certain smoothness condition on the underlying data distribution, the gradient descent evolution is well approximated by a Wasserstein gradient flow, provided that the number of neurons exceeds the data dimensions. As mentioned above, the algorithm treated here differs from the ones analyzed in earlier work, because the weights $\boldsymbol{w}_i$ are constrained to lie in the convex set Ω^δ. We enforce this constraint by using projected SGD, i.e. projecting at each step the weights onto the set Ω^δ. We generalize the analysis of [MMN18], obtaining convergence to a PDE with Neumann (reflecting) boundary conditions. As in [MMN18], we build on ideas that were first developed in the context of interacting particle systems [Dob79, Szn91].

The Wasserstein gradient flow approach was used in [MMN18, CB18] to establish global convergence results. However, these results fall short of our objectives for several reasons:

- The global convergence result of [CB18] rely on certain homogeneity properties of the neurons that are lacking here. We could obtain homogeneity by adding coefficients to Eq. (1.4), i.e. considering $\hat{f}(\boldsymbol{x};\boldsymbol{w}) = \sum_{i=1}^N a_i K^\delta(\boldsymbol{x}-\boldsymbol{w}_i)$ and minimizing the risk with respect to the coefficients a_i. As mentioned above, we refrain from introducing coefficients not to oversimplify the problem: when $N\to\infty$, it is sufficient to fit the coefficients a_i to achieve vanishing risk. Fitting the a_i's is a least squares problem.

- Most importantly, the techniques [MMN18, CB18] do not establish any convergence rates. This is not surprising, as those results hold under weak assumptions on the data distribution and the activation function. In particular, [MMN18, CB18, MMM19] cover general risk functions of the form (1.2) under certain smoothness and boundedness conditions on σ and on the functions $V(\boldsymbol{w}) = -\mathbb{E}\{f(\boldsymbol{x})\sigma(\boldsymbol{x};\boldsymbol{w})\}$, $U(\boldsymbol{w}_1,\boldsymbol{w}_2) = \mathbb{E}\{\sigma(\boldsymbol{x};\boldsymbol{w}_1)\sigma(\boldsymbol{x};\boldsymbol{w}_2)\}$. In such a general setting [MMN18] provides examples in which the Wasserstein gradient flow has multiple fixed points, which are singular with respect to the Lebesgue measure. Global convergence is established in [MMN18, CB18] by proving that PDE solution ρ_t has a strictly positive density. However, it is difficult to imagine this condition to hold in a quantitative dimension-independent manner.

  In contrast, our results are a first step towards dimension-independent convergence rate, in a more restricted setting than [MMN18, CB18, MMM19].

In summary, our results do not subsume earlier work, that assumes a more general setting, but rather establish stronger results in narrower context. Indeed, we believe that specific structural conditions must be imposed on the data distribution and activation function for the Wasserstein gradient flow approach to yield quantitative convergence rates. This paper presents one specific

set of assumptions. Although our results are not strong enough to establish non-asymptotic convergence rates, they point clearly in that direction.

3 Model and assumptions

3.1 Notations

We will use lowercase boldface for vectors, e.g. $\boldsymbol{x}, \boldsymbol{y}, \dots$, uppercase for random variables, e.g. $X, Y, \dots$, and uppercase boldface for random vectors, e.g. $\boldsymbol{X}, \boldsymbol{Y}, \dots$. The scalar product of two vectors is denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{d} x_i y_i$, and the ℓ_2 norm of a vector is denoted by $|\boldsymbol{x}|$. The Euclidean ball in $\mathbb{R}^d$ with center $\boldsymbol{x}$ and radius r is denoted by $\mathsf{B}(\boldsymbol{x}; r)$. Given a set $\Omega \subseteq \mathbb{R}^d$, we denote by $|\Omega|$ its volume.

We will refer to several function spaces in what follows. The most common is the space of p-th integrable functions $\mathscr{L}^p(\mathcal{X})$ on a measure space $(\mathcal{X}, \mathcal{F}, \mu)$. Given a function $f : \mathcal{X} \to \mathbb{R}$, we denote by $\|f\|_{\mathscr{L}^p(\mathcal{X})}$ its $\mathscr{L}^p$ norm, namely $\|f\|^p_{\mathscr{L}^p(\mathcal{X})} = \int_{\mathcal{X}} |f(x)|^p \, \mu(\mathrm{d}x)$. For $S \subseteq \mathbb{R}^m$, $\mathscr{C}^k(S)$ denotes the space of continuous functions $f : S \to \mathbb{R}$ with continuous derivatives up to order k. In particular, $\mathscr{C}(S)$ denotes the space of continuous real-valued functions defined on S. In addition, for $T \in \mathbb{R}_+$ and a metric space $\mathcal{M}$ (with distance $d_{\mathcal{M}}$), $\mathscr{C}([0,T], \mathcal{M})$ denotes the set of continuous functions $f : [0,T] \to \mathcal{M}$, endowed with the distance between two functions $f, g \in \mathscr{C}([0,T], \mathcal{M})$ defined as $d_{\mathscr{C}([0,T],\mathcal{M})}(f,g) \equiv \sup_{t \in [0,T]} d_{\mathcal{M}}(f(t), g(t))$. For a function $f : S \to \mathbb{R}$, we let $\|f\|_{\mathrm{Lip}} \equiv \sup_{\boldsymbol{x} \neq \boldsymbol{y} \in S} |f(\boldsymbol{x}) - f(\boldsymbol{y})| / |\boldsymbol{x} - \boldsymbol{y}|$ be the Lipschitz constant of the function f. Finally, as mentioned above, $\mathscr{P}_2(\mathcal{X})$ denotes the space of probability distributions on $\mathcal{X}$, endowed with the Wasserstein metric W_2

Throughout the paper, we use C to denote finite constants, which can vary from point to point. When these constants can depend on some of the problem parameters, e.g. a, b, c, we will write $C(a, b, c)$. When they are absolute numerical constants, we will emphasize this by writing C_*.

3.2 Data

As mentioned above, we are given data $(y_j, \boldsymbol{x}_j) \sim_{\text{i.i.d.}} \mathbb{P}$ where $\boldsymbol{x}_j \sim \mathsf{Unif}(\Omega)$, with $\Omega \subset \mathbb{R}^d$ a compact convex set, and $y_j = f(\boldsymbol{x}_j) + \varepsilon_j$, with $f : \Omega \to \mathbb{R}_{\geq 0}$. We assume the ε_j to be i.i.d. σ^2-subgaussian random variables with $\mathbb{E}(\varepsilon_j | \boldsymbol{x}_j) = 0$. We assume the function f to be concave and smooth.

Our formal assumptions on the set Ω and the function f are as follows:

(A1) $\Omega \supseteq \mathsf{B}(\mathbf{0}; r)$, with $r > 0$, is a compact convex set with $\mathscr{C}^2$ boundary.

(A2) $f : \Omega \to \mathbb{R}_{\geq 0}$ uniformly concave, i.e., there exists $\alpha > 0$ such that

$$\langle \boldsymbol{y}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{y} \rangle \leq -\alpha |\boldsymbol{y}|^2, \qquad \forall \boldsymbol{x} \in \Omega, \ \boldsymbol{y} \in \mathbb{R}^d, \tag{3.1}$$

where $\nabla^2 f$ denotes the Hessian of f.

(A3) $f \in \mathscr{C}^\infty(\Omega)$, with $\|f\|_{\mathscr{L}^\infty(\Omega)}, \|\nabla f\|_{\mathscr{L}^\infty(\Omega)} \leq C_*$ for an absolute constant C_*.

Without loss of generality, we can also assume that $\int_\Omega f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 1$. As a running example, we will use $\Omega = \mathsf{B}(\mathbf{0}; r)$, where we remind r is defined in Assumption (A1).

Remark 3.1. The assumption $\boldsymbol{x}_j \sim \mathsf{Unif}(\Omega)$ is quite strong but simplifies our analysis. We believe our approach can be generalized to a broader family of probability distribution for the covariates $\boldsymbol{x}_j$, but defer these generalizations to future work.

3.3 Neural network and SGD

Let $K \in \mathscr{C}^2(\mathbb{R}^d)$ be a non-negative symmetric first order kernel with compact support. Formally, we assume that

$$\text{(A4)} \qquad \int K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1\,, \quad K(\boldsymbol{x}) \ge 0, \quad \int K(\boldsymbol{x})\,\boldsymbol{x}\,\mathrm{d}\boldsymbol{x} = 0, \tag{3.2}$$

$$K(-\boldsymbol{x}) = K(\boldsymbol{x})\,, \qquad \mathrm{supp}(K) \subseteq \mathsf{B}(\boldsymbol{0}, c_0)\,. \tag{3.3}$$

The assumptions of symmetry and compact support are not crucial, but simplify some of the technical details later. We will further assume $\|\nabla K\|_{\mathscr{L}^\infty(\mathbb{R}^d)}$, $\|\nabla^2 K\|_{\mathscr{L}^\infty(\mathbb{R}^d)}$ and c_0 to be independent of the ambient dimension d. Notice that this requirement follows from the differentiability and compact support assumptions if $K(\boldsymbol{x}) = \kappa(\|\boldsymbol{x}\|_2)$ is a radial function.

For $\delta > 0$, let $K^\delta(\boldsymbol{x}) = \delta^{-d} K(\boldsymbol{x}/\delta)$. We try to fit the function (1.4) with parameters $\boldsymbol{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_N)$. These parameters are constrained to $\boldsymbol{w}_i \in \Omega^\delta$ which is a suitable scaling of Ω, as defined in the following. Given $\delta < r/c_0$, with r defined in (A1), define

$$\Omega^\delta = \lambda_\delta\,\Omega,$$

where

$$\lambda_\delta = \sup\big\{\lambda \ge 0 :\ \lambda\Omega \oplus \mathsf{B}(\boldsymbol{0}, c_0\,\delta) \subseteq \Omega\big\}\,. \tag{3.4}$$

For two sets $A, B \subseteq \mathbb{R}^d$, their Minkowski sum is defined as $A \oplus B = \{\boldsymbol{x} + \boldsymbol{y} : \boldsymbol{x} \in A, \boldsymbol{y} \in B\}$. Note that $\lambda_\delta \in [0,1]$ for all δ. Furthermore, $\Omega \supseteq \mathsf{B}(\boldsymbol{0}; r)$ implies $\lambda_\delta > 0$ for all $\delta < r/c_0$. Finally, $\lambda_{\delta=0} = 1$, whence $\Omega^{\delta=0} = \Omega$. In our running example, $\Omega^\delta = \mathsf{B}(\boldsymbol{0}; r - c_0\delta)$ is a ball of slightly smaller radius. Clearly, since Ω is convex, Ω^δ is convex as well.

We use stochastic gradient descent to minimize the population risk (1.2). At each step, we use a new data point $(y_k, \boldsymbol{x}_k)$, thus the sample size is equal to the number of iterations of the algorithm. Assuming for simplicity constant step size $\varepsilon > 0$, we update the parameters by

$$\boldsymbol{w}_i^{k+1} = \mathsf{P}\left\{\boldsymbol{w}_i^k - \varepsilon \nabla K^\delta(\boldsymbol{x}_{k+1} - \boldsymbol{w}_i^k)\,\Big(y_{k+1} - \hat{f}(\boldsymbol{x}_{k+1}; \boldsymbol{w}^k)\Big) + \sqrt{2\varepsilon\tau}\,\boldsymbol{g}_i^{k+1}\right\}. \tag{3.5}$$

Here $\boldsymbol{g}_i^{k+1} \sim \mathsf{N}(0, \boldsymbol{I}_d)$ is Gaussian noise which we take to be i.i.d. across time and neuron indices, k and i, and P is the orthogonal projector onto Ω^δ:

$$\mathsf{P}(\boldsymbol{z}) = \arg\min\big\{|\boldsymbol{z} - \boldsymbol{x}| :\quad \boldsymbol{x} \in \Omega^\delta\big\}\,. \tag{3.6}$$

The noise term $\sqrt{2\varepsilon\tau}\,\boldsymbol{g}_i^{k+1}$ is added mainly for technical reasons. Namely, it allows us to control the smoothness of the solutions of the resulting PDE. In simulations we do not find it useful, and we believe that a more careful analysis would be able to establish smoothness without the noise term.

Again, in our running example, we have

$$\mathsf{P}(\boldsymbol{z}) = \begin{cases} \boldsymbol{z} & \text{if } |\boldsymbol{z}| \le r - c_0\delta, \\ (r - c_0\delta)\boldsymbol{z}/|\boldsymbol{z}| & \text{if } |\boldsymbol{z}| > r - c_0\delta. \end{cases} \tag{3.7}$$

We initialize SGD with $(\boldsymbol{w}_i^0)_{i\le N} \sim_{\text{i.i.d.}} \rho_{\text{init}}^\delta \in \mathscr{P}_2(\Omega^\delta)$, where $\rho_{\text{init}}^\delta$ is a scaling of a fixed distribution $\rho_{\text{init}} \in \mathscr{P}_2(\Omega)$, i.e. $\rho_{\text{init}}^\delta(S) = \rho_{\text{init}}(S/\lambda_\delta)$. We assume that the initialization is smooth:

(A5) $\rho_{\text{init}} \in \mathscr{C}^\infty(\Omega^\delta)$.

3.4 PDE Model, $\delta > 0$

In the $N \to \infty$ limit the population risk is approximated by the effective risk $R^\delta : \mathscr{P}_2(\Omega^\delta) \to \mathbb{R}$ defined in Eq. (1.6). We emphasize that ρ is a probability distribution supported on Ω^δ. Note that

$$\inf_\rho R^\delta(\rho) \le R^\delta(f) = \nu_0 \int_\Omega \big[f(\boldsymbol{x}) - K^\delta * f(\boldsymbol{x})\big]^2 \mathrm{d}\boldsymbol{x}\,. \tag{3.8}$$

In particular $\lim_{\delta\to 0} \inf_{\rho\in\mathscr{P}_2(\Omega)} R^\delta(\rho) = 0$.

Our first main result is that the dynamics of SGD is well approximated by the following PDE (see Section 5.1 for a formal statement):

$$\begin{aligned} &\partial_t \rho_t(\boldsymbol{w}) = \nabla\cdot\big(\rho_t(\boldsymbol{w})\nabla\Psi(\boldsymbol{w};\rho_t)\big) + \tau\Delta\rho_t(\boldsymbol{w})\,, \\ &\Psi(\boldsymbol{w};\rho) \equiv -\nu_0\, K^\delta * f(\boldsymbol{w}) + \nu_0\, K^\delta * K^\delta * \rho(\boldsymbol{w})\,, \end{aligned} \tag{3.9}$$

with initial and boundary conditions

$$\begin{aligned} \rho_0 &= \rho_{\text{init}}^\delta, \\ \langle \boldsymbol{n}(\boldsymbol{w}), \rho_t(\boldsymbol{w})\nabla\Psi(\boldsymbol{w};\rho_t) + \tau\nabla\rho_t(\boldsymbol{w})\rangle &= 0 \qquad \forall \boldsymbol{w}\in\partial\Omega^\delta\,, \end{aligned} \tag{3.10}$$

where $\boldsymbol{n}(\boldsymbol{x})$ denotes the inward normal vector to $\partial\Omega^\delta$ at $\boldsymbol{x}$.

A rigorous definition of solutions of this PDE, along with some of their properties, is given in Appendix B. In Appendix C, we discuss the connection between the PDE (3.9) and the so-called "nonlinear dynamics", i.e. a stochastic differential equation that captures the trajectories of the weights $\boldsymbol{w}_i^k$. Using this connection, we prove existence and uniqueness of weak solutions of Eq. (3.9). In the proofs, we will often assume $\nu_0 = 1$, which amounts to a rescaling of time t.

For $\tau = 0$, the evolution defined by Eq. (3.9) corresponds to the gradient flow in Wasserstein metric for the risk function $R^\delta(\rho)$. For $\tau > 0$, it is the gradient flow for the free energy functional $F^\delta(\rho)$ defined below

$$F^\delta(\rho) = \frac{1}{2}R^\delta(\rho) - \tau\, S(\rho)\,, \quad S(\rho) = -\int \rho(\boldsymbol{w})\log\rho(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\,. \tag{3.11}$$

3.5 Limit PDE, $\delta = 0$

As mentioned above, in the limit $\delta \to 0$ the risk function $R^\delta(\rho)$ is well approximated by $R : \mathscr{L}^2(\Omega) \to \mathbb{R}$, where $R(\rho) = \nu_0\|f - \rho\|^2_{\mathscr{L}^2(\Omega)}$, cf. Eq. (1.7).

The corresponding Wasserstein gradient flow is also known as *viscous porous medium equation* [Váz07] and it is given by

$$\partial_t \rho_t(\boldsymbol{w}) = -\nu_0 \nabla \cdot \big(\rho_t(\boldsymbol{w}) \nabla f(\boldsymbol{w})\big) + \frac{\nu_0}{2} \Delta(\rho_t^2(\boldsymbol{w})) + \tau \Delta \rho_t(\boldsymbol{w})\,, \tag{3.12}$$

with initial and boundary conditions

$$\begin{aligned} \rho_0 &= \rho_{\text{init}}\,, \\ \langle \boldsymbol{n}(\boldsymbol{w}), \nu_0 \rho_t(\boldsymbol{w})\, \nabla(f(\boldsymbol{w}) - \rho_t(\boldsymbol{w})) - \tau \nabla \rho_t(\boldsymbol{w}) \rangle &= 0 \qquad \forall \boldsymbol{w} \in \partial\Omega\,. \end{aligned} \tag{3.13}$$

In Appendix A, we give the definition of a weak solution for the PDE (3.12) with initial and boundary conditions (3.13). We also prove that the weak solution of the PDE (3.12) is unique, under a mild integrability condition. Again, in proofs we will assume without loss of generality $\nu_0 = 1$.

As in the $\delta > 0$ case, the evolution defined by Eq. (3.12) is the gradient flow for the free energy $F(\rho) = (1/2)R(\rho) - \tau S(\rho)$. Our analysis uses a key property of the risk function $R(\rho) = \nu_0 \|f - \rho\|^2_{\mathscr{L}^2(\Omega)}$ (and the free energy): displacement convexity [McC97]. For the reader's convenience, we recall its definition here, referring to [AGS08, Vil08, San15] for further background. Given two probability measures $\rho_0, \rho_1 \in \mathscr{P}_2(\Omega)$, their W_2 distance is defined by

$$W_2(\rho_0, \rho_1)^2 = \inf_{\gamma \in \Gamma(\rho_0, \rho_1)} \int \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \, \gamma(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\,, \tag{3.14}$$

where the infimum is taken over the set $\Gamma(\rho_0, \rho_1)$ of couplings of ρ_0, ρ_1 (i.e. probability measures on $\Omega \times \Omega$ whose first marginal coincides with ρ_0, and second with ρ_1). The infimum is achieved by weak compactness of $\mathscr{P}_2(\Omega)$.

The metric space $(\mathscr{P}_2(\Omega), W_2)$ is a 'length space,' and in particular it is possible to construct geodesics, i.e. paths of minimum length connecting any two probability measures ρ_0, ρ_1. Geodesics have a simple description. Let γ_* be the coupling achieving the infimum in the definition of $W_2(\rho_0, \rho_1)$. Letting $(\boldsymbol{X}_0, \boldsymbol{X}_1) \sim \gamma_*$, we define ρ_t to be the distribution of $\boldsymbol{X}_t = (1-t)\boldsymbol{X}_0 + t\boldsymbol{X}_1$. The curve $t \mapsto \rho_t$, indexed by $t \in [0,1]$ turns out to be the geodesic between ρ_0 and ρ_1 in $(\mathscr{P}_2(\Omega), W_2)$.

Displacement convexity is convexity along geodesics. Namely, a function $\mathcal{F} : \mathscr{P}_2(\Omega) \to \mathbb{R}$ is λ-strongly displacement convex if

$$(1-t)\mathcal{F}(\rho_0) + t\,\mathcal{F}(\rho_1) - \mathcal{F}(\rho_t) \geq \frac{1}{2}\lambda\, t(1-t) W_2(\rho_0, \rho_1)^2\,. \tag{3.15}$$

A useful observation is that displacement convexity implies that all local minima of $\mathcal{F}$ are global minimizer. Indeed, by (3.15) it is straightforward to see that $\mathcal{F}$ has at most one global minimizer ρ^*. Also, for every other point ρ, the geodesic between ρ and ρ_* is a strictly decreasing path for the function $\mathcal{F}$. Now, suppose that $\bar{\rho} \neq \rho_*$ is a local minimum. Then, there exists a neighborhood U around $\bar{\rho}$ such that, for any $\rho \in U$, $\mathcal{F}(\rho) \geq \mathcal{F}(\bar{\rho})$. However, the strictly decreasing path between $\bar{\rho}$ and ρ_* passes through the neighborhood U, which leads to a contradiction and so $\rho = \rho_*$

It follows from [McC97] that the risk function $R(\rho)$ and the free energy $F(\rho)$ are strongly displacement convex.

Remark 3.2. The concavity assumption on the regression function f (Assumption (A2)) defines a nonparametric class under which global convergence can be established, with convergence rates

uniquely determined by the curvature α (in the limit $N \to \infty$, $\delta \to 0$). Nonparametric estimation of concave functions has attracted considerable attention over recent years, see e.g. [HD13, CS16], and is –by itself– an interesting domain of applicability.

However, our projected SGD algorithm is potentially applicable to any data set, and will return a meaningful estimate $\hat{f}$ regardless whether f is concave or not. Indeed, in the next section we present numerical simulations indicating convergence to a near-global optimum even for non-concave functions f.

From mathematical point of view, Assumption (A2) is only used to show the convergence of the solution of the viscous porous medium equation (limit PDE, $\delta = 0$) to the unique global minimizer of the free energy $F(\rho) = (1/2)R(\rho) - \tau S(\rho)$, as formally stated in Theorem F.8. Concavity is not needed for the other results in the paper, namely approximating the SGD trajectory with the solution of the PDE ($\delta > 0$), see Theorem 5.1, and the convergence of the solution of the PDE ($\delta > 0$) to the solution of the viscous porous medium equation, see Theorem 5.2. It is therefore foreseeable a more general analysis that relaxes the concavity assumption.

4 Numerical illustrations

In this section we provide some simple numerical illustrations of our setting, and compare numerical results with the predictions of the Wasserstein gradient flow theory.

It is easy to construct examples of strongly concave functions, satisfying our assumptions. One can start from any strongly concave continuous function f_0 on a compact convex set Ω, add a constant to make it non-negative, and multiply it by a constant to normalize its integral. The resulting function $f(\boldsymbol{x}) = (c_1 + f_0(\boldsymbol{x}))/c_2$ satisfies our conditions. Notable examples of concave functions are given by log-moment generating functions $f_0(\boldsymbol{x}) = -\log \mathbb{E}_{\boldsymbol{Z}} \exp\{\langle \boldsymbol{x}, \boldsymbol{Z}\rangle\}$, where the random variable $\boldsymbol{Z}$ satisfies mild assumptions (e.g., it is bounded and its distribution is not supported on a proper subspace of $\mathbb{R}^d$). In general, given any twice differentiable function g_0, the function $f_0(\boldsymbol{x}) = g_0(\boldsymbol{x}) - c_*\|\boldsymbol{x}\|_2^2$ is strongly concave for c_* large enough.

4.1 A one-dimensional concave function

We set $\Omega = [-1, 1]$ and $f(x) = (1 - e^{x-1})/(1 - e^{-2})$ (we choose the normalization so that $\int_{-1}^{1} f(x)\mathrm{d}x = 1$). Note that f is uniformly concave in $[-1, 1]$. We set the kernel K as follows:

$$K(x) = C_d \kappa(|x|)\,, \quad \kappa(t) = \begin{cases} 1 - t^2 - 2t^3 + 2t^4 & \text{for } t \le c_0 = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where C_d is a normalization constant ensuring that $\int_{-1}^{1} K(x)\,\mathrm{d}x = 1$. The initialization ρ_{init} is a truncated Gaussian: $\rho_{\text{init}}(x) = c \cdot \exp(-x^2/(2\sigma^2))\,\mathbf{1}_{[-1,1]}(x)$, with $\sigma = 1/3$.

We find empirically that standard stochastic gradient descent (SGD) without the projection P onto Ω^δ works well in this example, and consider this algorithm for simplicity in our first illustrations. We pick $N = 200$, $\tau = 0$ (noiseless SGD), and constant step size $\varepsilon = 10^{-6}$. In Figure 1, left column, we plot the true function $f(\,\cdot\,)$ together with the neural network estimate $\hat{f}(\,\cdot\,; \boldsymbol{w}^k)$ at several points in time t (time is related to the number of iterations k via $t = k\varepsilon$). Different

plots correspond to different values of δ with $\delta \in \{1/5, 1/10, 1/20\}$. We observe that the network estimates $\hat{f}(\,\cdot\,; \boldsymbol{w}^k)$ seem to converge to a limit curve which is an approximation of the true function f. As expected, the quality of the approximation improves as δ gets smaller.

In the right column, we report the evolution of the population risk (1.2) normalized by $\|f\|^2_{\mathscr{L}^2(\Omega)}$. For comparison, we plot the evolution of the risk (1.7) as predicted by the limit PDE (3.12) with $\tau = 0$. We solve the PDE (3.12) numerically using a finite difference scheme that enforces the conservation law $\int \rho(x,t)\mathrm{d}x = 1$, see, e.g., [Tho13]. In the finite difference scheme, we choose time step and spatial step $\Delta t = 10^{-5}$ and $\Delta x = 10^{-2}$, respectively. The curve obtained by this numerical solution appears to capture well the evolution of SGD towards optimality. The main difference is that, while the PDE (3.12) corresponds to $\delta = 0$, and hence evolves towards a global optimum at zero risk, SGD converges to a non-zero risk value, which can be interpreted as the approximation error, decreasing with δ.

In Figure 2, we illustrate the numerical solution of the PDE (3.12) by plotting *(i)* the regression function f together with the PDE solution ρ_t (which coincides with the prediction $\hat{f}$ at $\delta = 0$) at several times t, and *(ii)* the PDE prediction for the risk $R(\rho_t)$ (1.7) normalized with respect to $\|f\|^2_{\mathscr{L}^2(\Omega)}$ (this plot aggregates data from Figs. 1.(b), (d), (f)). We also compare the risk (1.7) to the population risk $R_N(\boldsymbol{w}^k)$ achieved by SGD for different values of δ. Note that, as δ becomes smaller, the risk converges to the predicted curve. The risk of the limit PDE (3.12) converges to 0 exponentially fast in t, as predicted by the strong displacement convexity of $R(\rho)$.

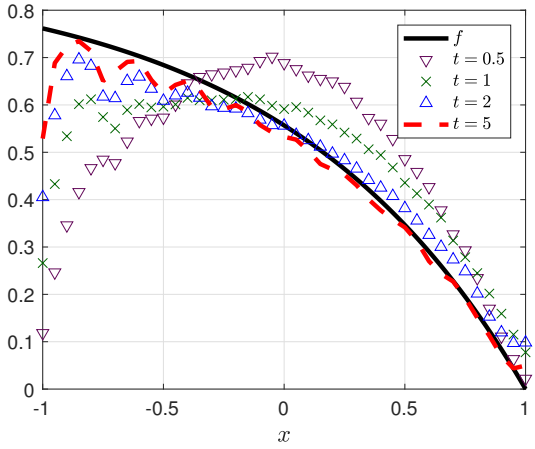
In Figure 3, we consider the SGD algorithm with projection P, see (3.5). We pick $N = 200$, $\tau = 0$, $\varepsilon = 10^{-6}$ and $\delta = 1/20$. On the left, we illustrate the evolution of the value of 40 weights chosen at random; and on the right, we plot the histogram of their empirical distribution at $t = 5$. Note that this histogram matches well the regression function f plotted in black.

4.2 A two-dimensional concave example

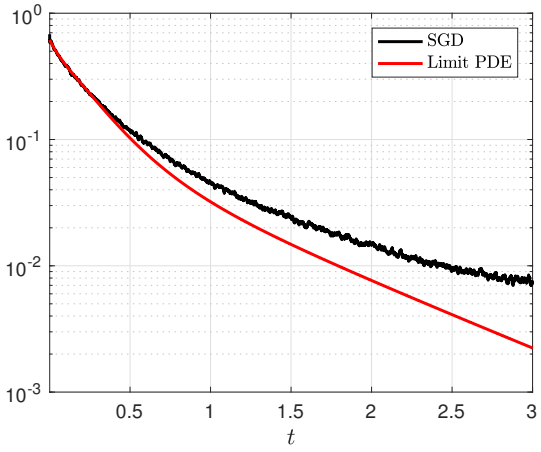
Next, we consider a two-dimensional example. We set $\Omega = [-1,1]^2$ and

$$f(\boldsymbol{x}) = \frac{c_1 - \log(e^{\langle \boldsymbol{q}_1, \boldsymbol{x}\rangle} + e^{\langle \boldsymbol{q}_2, \boldsymbol{x}\rangle})}{c_2}, \tag{4.2}$$

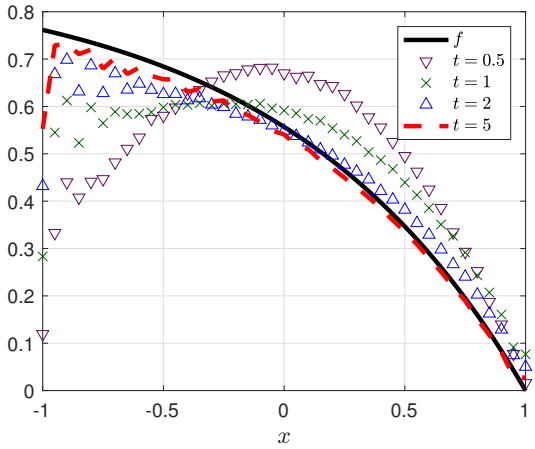
with $\boldsymbol{q}_1 = (2.5127, -2.4490)$, $\boldsymbol{q}_2 = (0.0596, 1.9908)$ and where c_1 and c_2 are chosen so that f is non-negative and $\int_\Omega f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$. The kernel K is given by $K(\boldsymbol{x}) = C_d\kappa(|\boldsymbol{x}|)$, where κ is defined in (4.1) and C_d is a normalization constant ensuring that $\int_{\mathsf{B}(\mathbf{0};1)} K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$. Again, the initialization ρ_{init} is a truncated Gaussian: $\rho_{\text{init}}(\boldsymbol{x}) = c \cdot \exp(-|\boldsymbol{x}|^2/(2\sigma^2))\,\mathbf{1}_{[-1,1]^2}(\boldsymbol{x})$, with $\sigma = 1/3$. We compare the normalized risk of SGD with no projection P ($N = 2000$, $\tau = 0$ and $\varepsilon = 10^{-6}$) for $\delta \in \{1/3, 1/5, 1/10\}$ with that of the limit PDE (3.12). Figure 4 shows that, already at $\delta = 1/10$, the risk of SGD converges to the predicted curve and the risk of the limit PDE (3.12) tends to 0 exponentially fast in t.



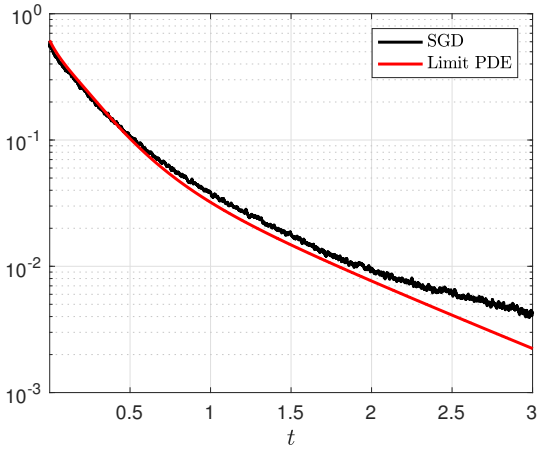
(a) Function f and SGD estimates, $\delta = 1/5$.



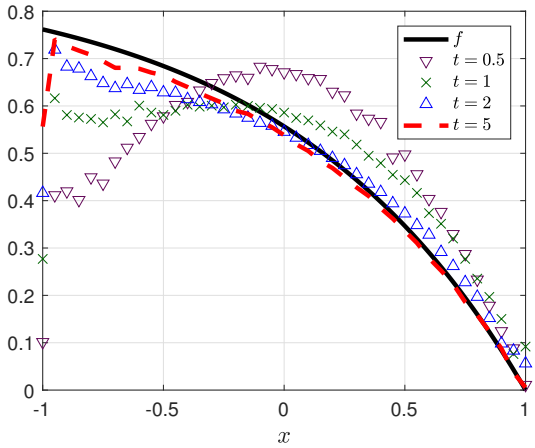
(b) Normalized risk, $\delta = 1/5$.



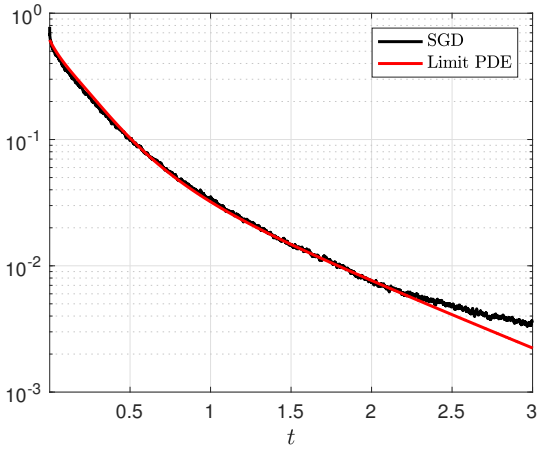
(c) Function f and SGD estimates, $\delta = 1/10$.



(d) Normalized risk, $\delta = 1/10$.

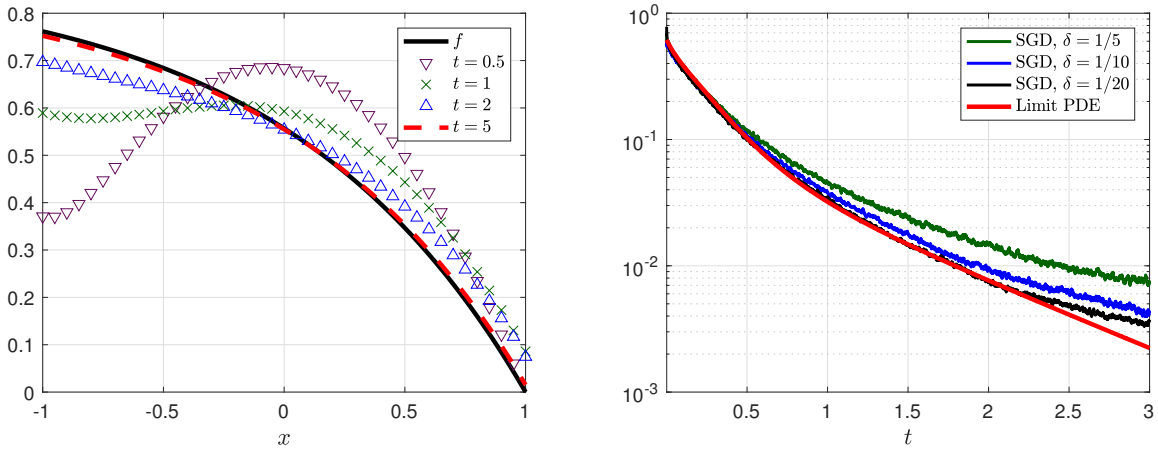


(e) Function f and SGD estimates, $\delta = 1/20$.



(f) Normalized risk, $\delta = 1/20$.

Figure 1: Dynamics of SGD update (3.5) at different times t and for different values of δ.



(a) Function f and PDE solution.

(b) Normalized risk.

Figure 2: Dynamics of limit PDE (3.12) at different times t.

4.3 Comparing feature learning to random features

As discussed in the introduction, it is useful to consider the more general model

$$\hat{f}(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a}) = \sum_{i=1}^{N} a_i K^{\delta}(\boldsymbol{x}-\boldsymbol{w}_i)\,, \tag{4.3}$$

with parameters $\boldsymbol{a} = (a_1,\dots,a_N)$ as well as $\boldsymbol{w} = (\boldsymbol{w}_1,\dots,\boldsymbol{w}_N)$. This setting allows to compare two different approaches:

(*i*) *Random feature regression*: the weights $\boldsymbol{w}$ are chosen independently of the labels y_i (we allow for dependence on the covariates $\boldsymbol{x}_i$).

(*ii*) *Feature learning*: the weights $\boldsymbol{w}$ depend on the data $(y_i,\boldsymbol{x}_i)$.

In order to compare these two approaches, we assume to be given i.i.d. data $\{(y_i,\boldsymbol{x}_i)\}_{i\le n}$, with $\boldsymbol{x}_i \sim \mathsf{Unif}(\Omega)$, $y_i = f(\boldsymbol{x}_i)$ and determine the parameters $\boldsymbol{a}$ by the same method, ridge regression. More explicitly, define the matrix $\boldsymbol{Z}\in\mathbb{R}^{n\times N}$ as $(\boldsymbol{Z})_{i,j} = K^{\delta}(\boldsymbol{x}_i-\boldsymbol{w}_j)$. Then, we estimate $\boldsymbol{a}$ via

$$\hat{\boldsymbol{a}} = (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\lambda\boldsymbol{I})^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}, \tag{4.4}$$

where λ is chosen via cross-validation on a hold-out set, comprising 10% of the samples.

In Figure 5, we compare the performance of three different ways to construct the weights $\boldsymbol{w}$: *'random $\boldsymbol{w}$,'* we choose the weights $\boldsymbol{w}_i$ independently and uniformly at random in Ω (blue triangles pointing down); *'$\boldsymbol{w}$ = data points,'* we choose the weights $\boldsymbol{w}_i$ uniformly at random among the data points (green circles); *'optimized $\boldsymbol{w}$,'* we use the output of the projected SGD algorithm of the previous sections (red triangles pointing up). The first two can be regarded as 'random features' approaches, while the latter is a 'feature learning' method.

For the optimized $\boldsymbol{w}$, we use exactly the same algorithm in as in (3.5) (without coefficients $\boldsymbol{a}$ in the SGD update), with the only difference that each SGD step is carried out with respect

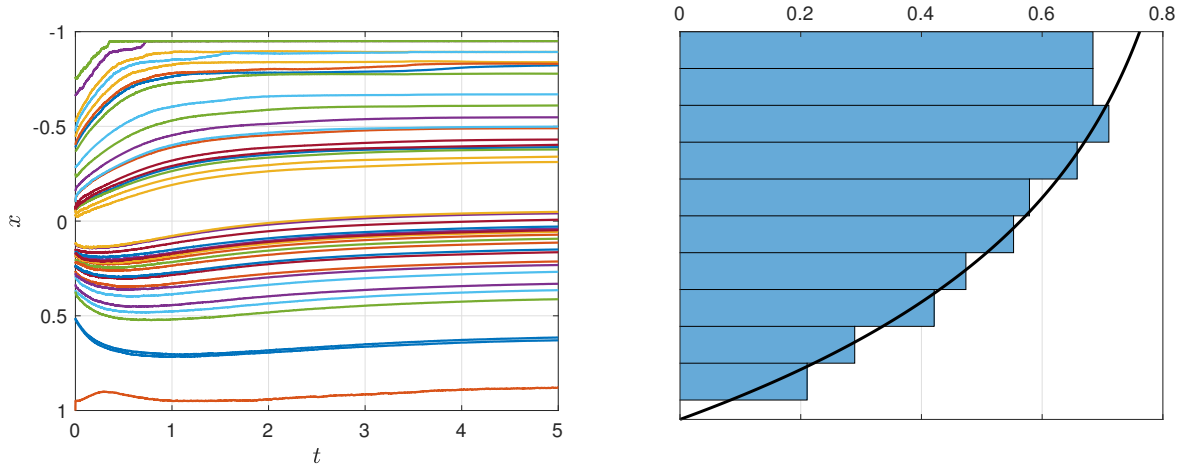


Figure 3: Evolution of the value of 40 weights chosen at random and histogram of their empirical distribution at time $t = 5$.

to an independent sample from the empirical data, with replacement. SGD is stopped after $k_{\max}$ iteration, and the coefficient $\hat{\boldsymbol{a}}$ are computed according to (4.3). Notice that this procedure is probably suboptimal, and it would be better to optimize $\boldsymbol{a}$ and $\boldsymbol{w}$ jointly: we choose this simpler two-stage procedure to have a more direct application of the algorithm analyzed in the paper, and a comparison with the random feature methods. We set $\tau = 0$ (noiseless SGD), and constant step size $\varepsilon = 5 \cdot 10^{-4}$. The number of iterations $k_{\max} \in \{5 \cdot 10^3, 15 \cdot 10^3, 5 \cdot 10^4, 15 \cdot 10^4, 5 \cdot 10^5, 15 \cdot 10^5\}$ is chosen via cross-validation, by using the same hold-out set employed to optimize λ.

We set $\Omega = [-1, 1]^4$ and define $y_j = f(\boldsymbol{x}_j)$, where $f(\boldsymbol{x})$ takes the form (4.2) with $\boldsymbol{q}_1 = (-0.3832, 0.3074, -0.3198, 0.4792)$ and $\boldsymbol{q}_2 = (0.3502, -0.1471, 0.1685, 0.0546)$. Again, c_1 and c_2 are chosen so that f is non-negative and $\int_\Omega f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$; the kernel K is given by $K(\boldsymbol{x}) = C_d \kappa(|\boldsymbol{x}|)$, where κ is defined in Eq. (4.1) and C_d ensures that $\int_{\mathsf{B}(\mathbf{0};1)} K(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$.

After estimating $\boldsymbol{w}_i$ and a_i by either methods, we generate a test set of $10,000$ samples and use it to estimate the generalization error. We perform 20 independent trials of the experiment, and we plot the average risk normalized by $\|f\|^2_{\mathscr{L}^2(\Omega)}$ together with the error bar at 1 standard deviation. In Figure 5-(a), we fix the number of neurons $N = 200$ and we plot the normalized risk as a function of the number of data points n. In Figure 5-(b), we fix the number of samples n to 2000 and we plot the normalized risk as a function of the number of neurons N. The data set used for cross-validation has size $\max(n/10, 40)$. Note that feature learning leads to improved performance in both settings. The improvement becomes more pronounced with the sample size n, presumably because a better set of weights $\boldsymbol{w}_i$ can be learnt. On the other hand, when the number of neurons N becomes very large, random $\boldsymbol{w}_i$'s are already covering Ω densely enough, and there is no significant advantage in feature learning.

4.4 A non-concave one-dimensional example

We set $\Omega = [-1, 1]$ and $f(x) = (x + \sin(5x - \pi/2) - c_1)/c_2$, where c_1 and c_2 are chosen so that f is non-negative and $\int_\Omega f(x)\, \mathrm{d}x = 1$. Note that the target function f is bimodal, thus it is not concave. We perform the same numerical experiment described in Section 4.1. In Figure 6, left column, we

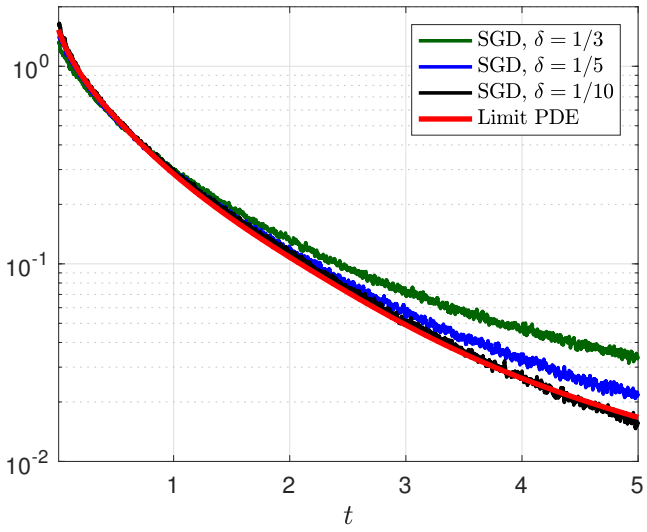
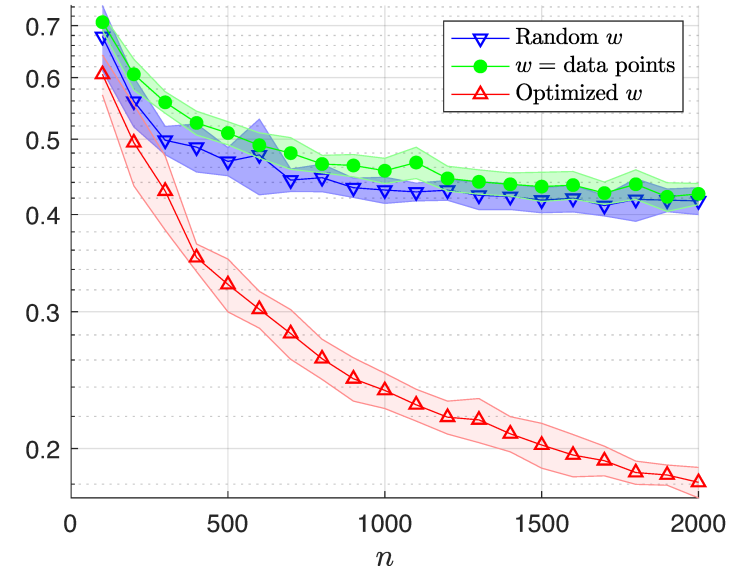


Figure 4: Normalized risk of SGD for different values of δ compared with that of the limit PDE for a two-dimensional example.

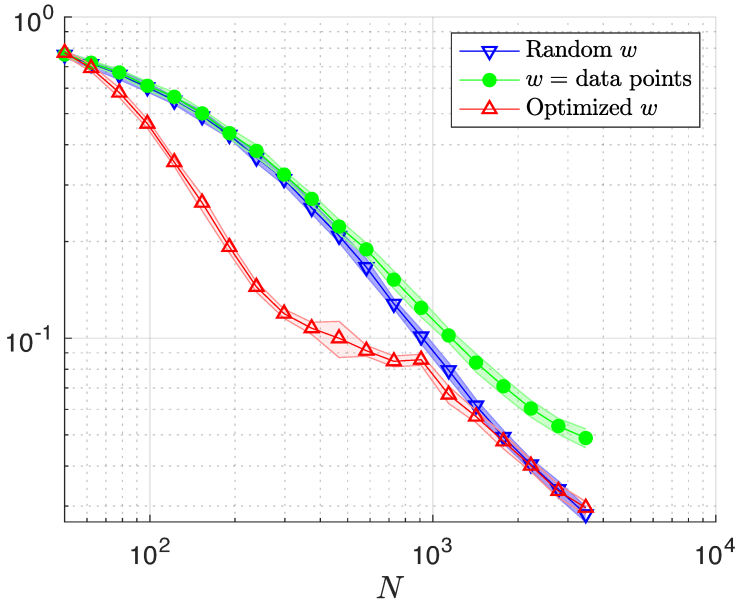
plot the true function $f(\cdot)$ together with the neural network estimate $\hat{f}(\cdot\,; \boldsymbol{w}^k)$ at several points in time t, where different plots correspond to different values of $\delta \in \{1/5, 1/10, 1/20\}$. In the right column, we report the evolution of the population risk (1.2) normalized by $\|f\|^2_{\mathscr{L}^2(\Omega)}$. In Figure 7, we plot *(i)* the regression function f together with the PDE solution ρ_t at several times t, and *(ii)* the PDE prediction for the risk $R(\rho_t)$ (1.7) (normalized with respect to $\|f\|^2_{\mathscr{L}^2(\Omega)}$) compared with the population risk $R_N(\boldsymbol{w}^k)$ achieved by SGD for different values of δ. Even if the target function is not concave, the results are similar to those presented in the concave case: *(i)* the network estimates $\hat{f}(\cdot\,; \boldsymbol{w}^k)$ seem to converge to a limit curve which is an approximation of the true function f, *(ii)* the quality of the approximation improves as δ gets smaller, and *(iii)* the risk of the limit PDE (3.12) converges to 0 exponentially fast in t.

4.5 Failure for small N

We repeat the same experiment described in Section 4.1 for a smaller number of neurons $N = 20$. As can be seen in Figures 8 and 9, the quality of the approximation becomes worse as δ gets smaller. This is expected because with small number of activations, reducing their bandwidth δ leads to a worse performance as they are all zero on a large part of the space. Put differently, the number of neurons is too small to guarantee convergence of SGD to the predictions of the Wasserstein gradient flow theory.

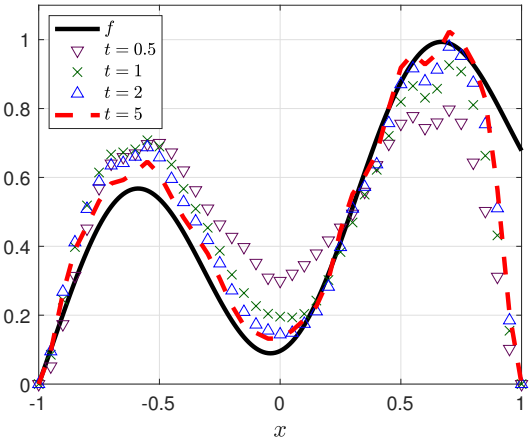


(a) $N = 200$, n varies on the x-axis.

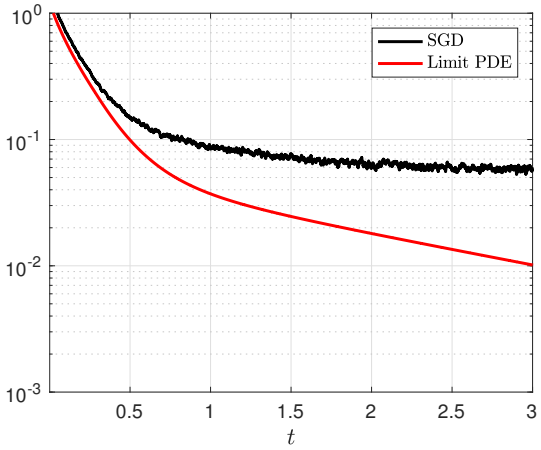


(b) $n = 2000$, N varies on the x-axis.

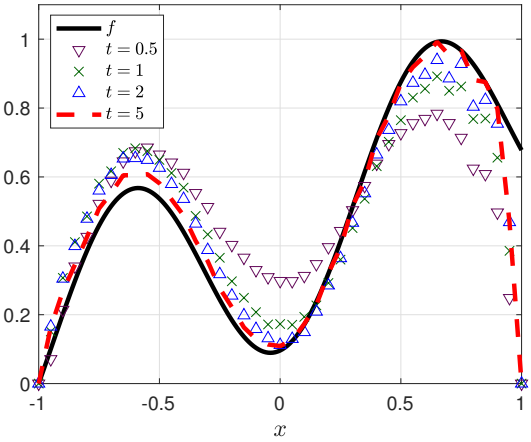
Figure 5: Generalization error achieved by fitting $\boldsymbol{a}$ from the data for three different choices of the weights $\boldsymbol{w}$: in red, the $\boldsymbol{w}_i$'s are optimized before-hand via SGD, as suggested in this paper; in blue, the $\boldsymbol{w}_i$'s are uniform in Ω; and in green, the $\boldsymbol{w}_i$'s are equal to random data points.



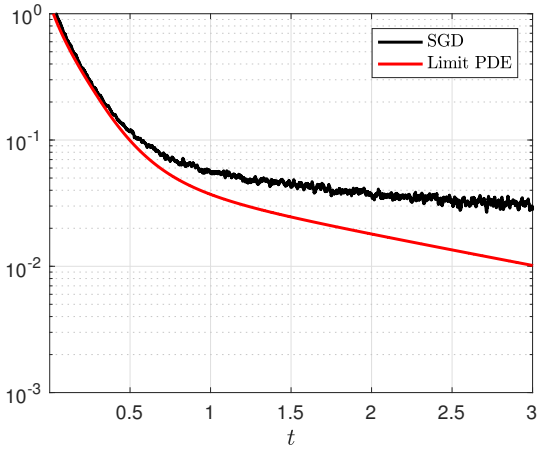
(a) Function f and SGD estimates, $\delta = 1/5$.



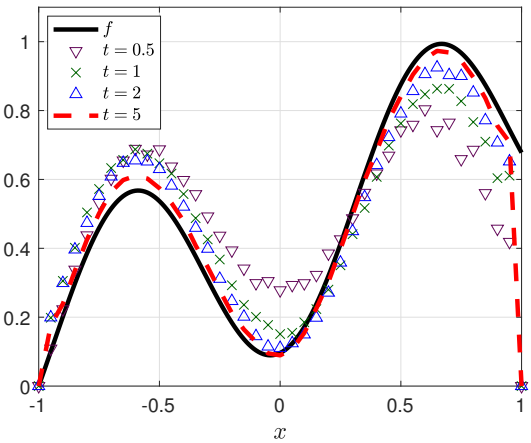
(b) Normalized risk, $\delta = 1/5$.



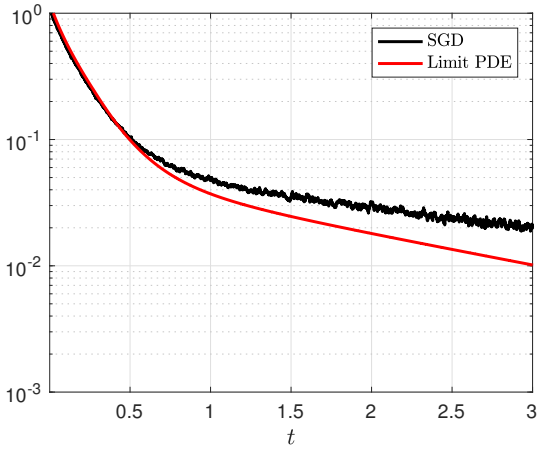
(c) Function f and SGD estimates, $\delta = 1/10$.



(d) Normalized risk, $\delta = 1/10$.

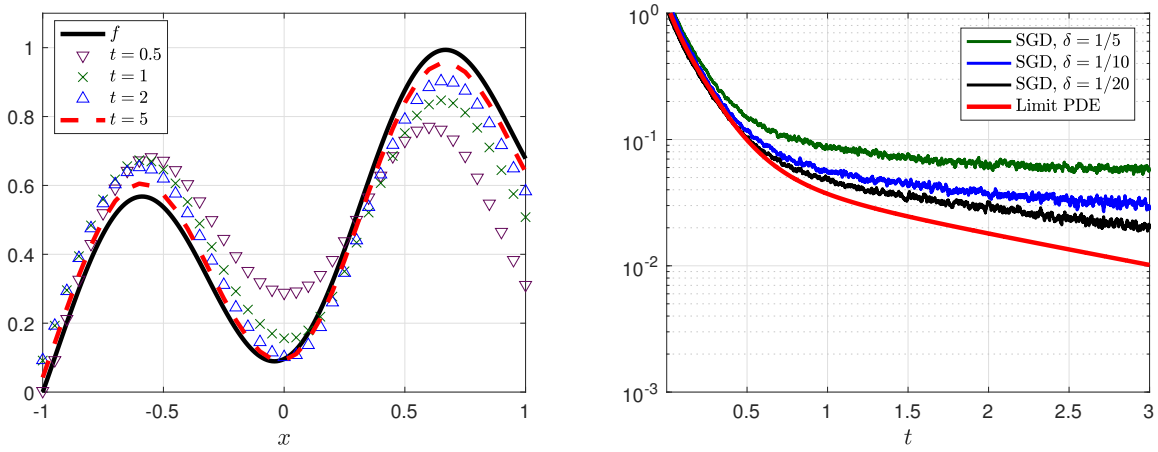


(e) Function f and SGD estimates, $\delta = 1/20$.



(f) Normalized risk, $\delta = 1/20$.

Figure 6: Dynamics of SGD update (3.5) at different times t and for different values of δ for a non-concave target function f.



(a) Function f and PDE solution.

(b) Normalized risk.

Figure 7: Dynamics of limit PDE (3.12) at different times t for a non-concave target function f.

5 Main results

5.1 Convergence of SGD to the PDE (3.9) at $\delta > 0$ fixed

We now state our result concerning the convergence of the SGD dynamics (3.5) to the PDE (3.9). Note that this result does not require concavity of f. Its proof is presented in Appendix D.

Theorem 5.1. *Assume that conditions* (A1), (A3)-(A5) *hold. Consider the SGD update* (3.5) *with initialization* $(\boldsymbol{w}_i^0)_{i\le N} \sim_{\text{i.i.d.}} \rho_{\text{init}}^{\delta}$ *and constant step size* ε*. For* $t \ge 0$*, let* ρ_t *be the unique solution of the PDE* (3.9) *with initial and boundary conditions* (3.10)*, and assume* $\text{supp}(\rho_{\text{init}}^{\delta}) \subseteq \mathsf{B}(\mathbf{0}, r)$ *Then, for any fixed* $t \ge 0$*,* $\rho^{(N)}_{\lfloor t/\varepsilon \rfloor} \Rightarrow \rho_t$ *almost surely along any sequence* $(N, \varepsilon = \varepsilon_N)$ *such that* $N \to \infty$*,* $\varepsilon_N \to 0$*.*

Furthermore, for any $\delta \le 1$*,* $T \ge 1$*,* $\varepsilon \le 1$*,* $p \in \mathbb{N}$*, and for any* $g : \mathbb{R}^d \to \mathbb{R}$ *with* $\|g\|_{\text{Lip}} \le 1$*, the following happens with probability at least* $1 - z^{-2p}$*,*

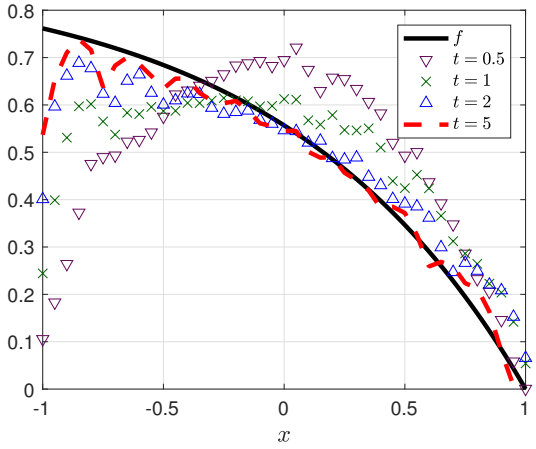
$$\begin{aligned} \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \left| \sum_{i=1}^{N} g(\boldsymbol{w}_i^k) - \int g(\boldsymbol{w})\rho_{k\varepsilon}(\mathrm{d}\boldsymbol{w}) \right| &\le z\, \mathsf{err}(N,d,\varepsilon,\delta)\; e^{C_* p \delta^{-(d+2)} T}, \\ \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\boldsymbol{w}^k) - R^{\delta}(\rho_{k\epsilon})| &\le z\, \mathsf{err}(N,d,\varepsilon,\delta)\; e^{C_* p \delta^{-(d+2)} T}, \end{aligned} \tag{5.1}$$

where

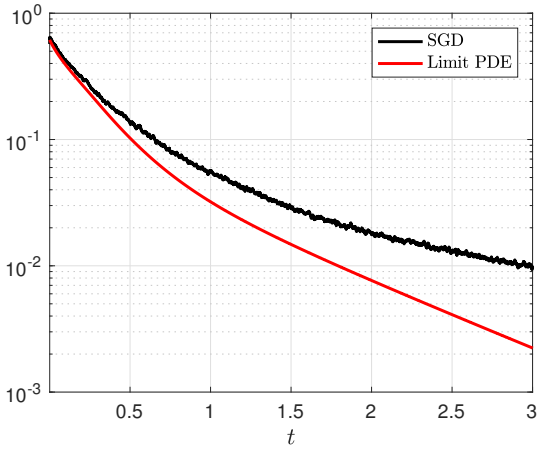
$$\mathsf{err}(N,d,\varepsilon,\delta) = \sqrt{\frac{d}{N}} \vee \left(\delta^{-2d-1} r (d^2 \varepsilon \log(1/\varepsilon))^{1/4}\right). \tag{5.2}$$

Our proof is based on the same approach developed in [MMN18]. We prove that solutions of the PDE (3.9) are in correspondence with distributions over trajectories $(\boldsymbol{X}_t)_{t\ge 0}$ in Ω satisfying the following stochastic differential equation

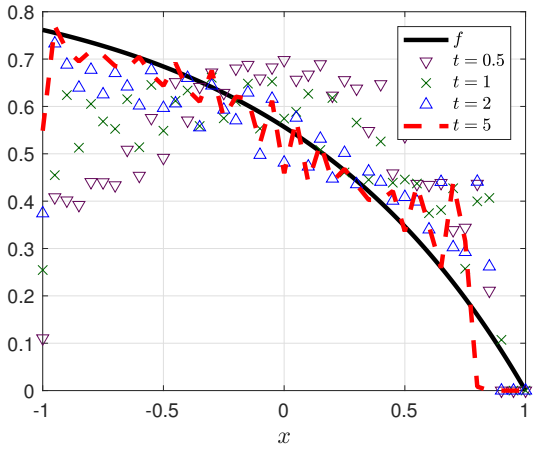
$$\mathrm{d}\boldsymbol{X}_t = -\nabla\Psi(\boldsymbol{X}_t, \rho_t)\,\mathrm{d}t + \sqrt{2\tau}\,\mathrm{d}\boldsymbol{B}_t + \mathrm{d}\boldsymbol{\Phi}_t\,, \tag{5.3}$$



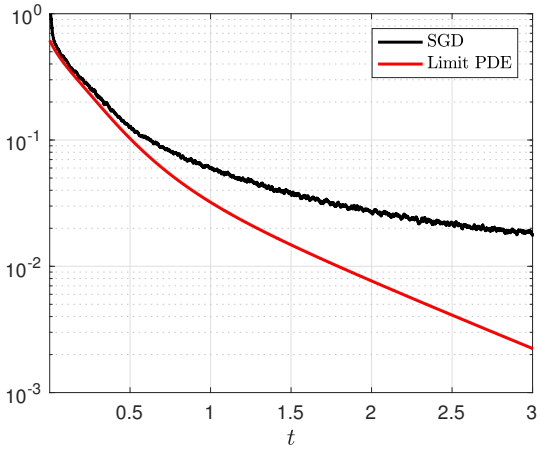
(a) Function f and SGD estimates, $\delta = 1/5$.



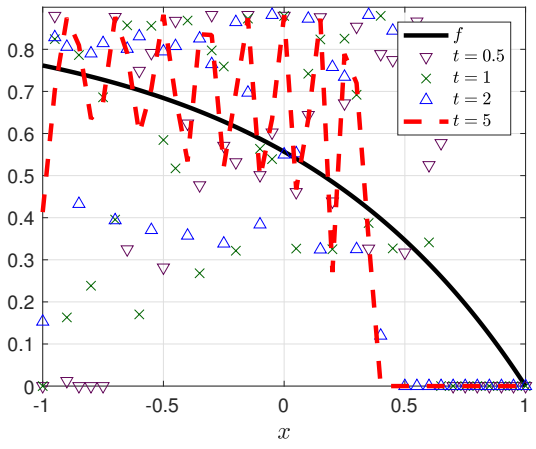
(b) Normalized risk, $\delta = 1/5$.



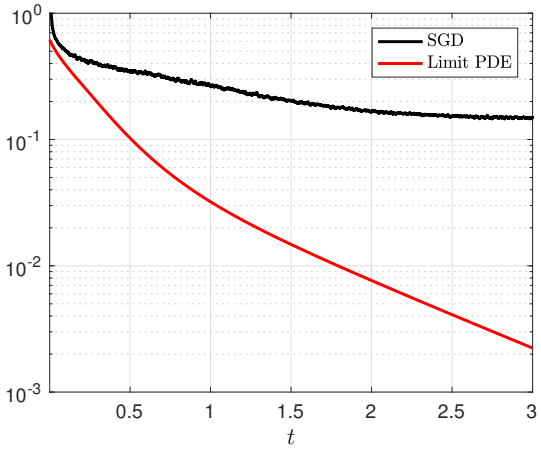
(c) Function f and SGD estimates, $\delta = 1/10$.



(d) Normalized risk, $\delta = 1/10$.



(e) Function f and SGD estimates, $\delta = 1/20$.



(f) Normalized risk, $\delta = 1/20$.

Figure 8: Dynamics of SGD update (3.5) at different times t and for different values of δ when the number of neurons is too small ($N = 20$).

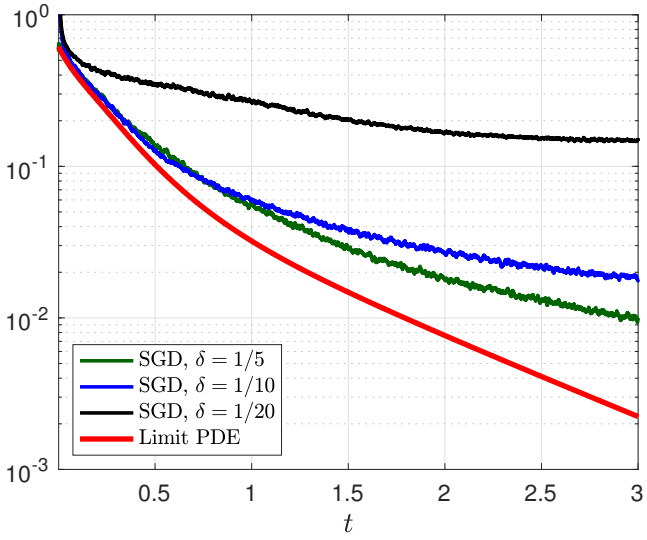


Figure 9: Normalized risk of the limit PDE (3.12) and of the SGD update (3.5) when the number of neurons is too small ($N = 20$).

where $(\boldsymbol{B}_t)_{t\geq 0}$ is a standard Brownian motion and $\mathrm{d}\boldsymbol{\Phi}_t$ is the boundary reflection (in the sense of a Skorokhod problem). The density ρ_t is determined, self consistently, via $\rho_t = \mathrm{Law}(\boldsymbol{X}_t)$. We prove existence and uniqueness of solutions to this problem, and refer to the corresponding stochastic process $(\boldsymbol{X}_t)_{t\geq 0}$ as *nonlinear dynamics*. This in turn implies existence and uniqueness of the solutions of the PDE (3.9).

We next construct a coupling between the network weights $(\boldsymbol{w}_1^k, \dots, \boldsymbol{w}_N^k) \in (\Omega^\delta)^N$, and N i.i.d. trajectories of the nonlinear dynamics $(\boldsymbol{X}_1^t, \dots, \boldsymbol{X}_N^t) \in (\Omega^\delta)^N$. Controlling the expected distance in this coupling yields Theorem 5.1.

Remark 5.1. The error term in Eq. (5.1) is completely analogous to the error in a similar theorem proved in [MMN18]. The constant δ^{-d} appearing here is obtained by bounding the Lipschitz constant of $\nabla\Psi(\boldsymbol{w};\rho)$. As already mentioned, the main technical difficulty with respect to [MMN18] is posed by the Neumann (reflecting) boundary conditions. Indeed, even if we are given a solution of the PDE (3.9), existence and uniqueness of solutions of the Skorokhod problem (5.3) is a highly non-trivial fact first established in [Tan79, LS84]. As a consequence, while the main proof idea is similar to the one in [MMN18], its implementation is significantly different.

Remark 5.2. As discussed in Appendix D, our proof applies to a more general version of the PDE (3.9) and correspondingly of the SGD dynamics (3.5), where Ψ takes the form $\Psi(\boldsymbol{w},\rho) = V(\boldsymbol{w}) + \int U(\boldsymbol{w},\boldsymbol{w}')\,\rho(\mathrm{d}\boldsymbol{w}')$, for $V : \Omega \to \mathbb{R}$, $U : \Omega\times\Omega \to \mathbb{R}$ two smooth functions. The SGD update (3.5) is generalized as in [MMN18], and Theorem 5.1 holds with the terms containing δ (i.e., δ^{-2d-1} and $\delta^{-(d+2)}$) replaced by a constant that depends uniquely on $\|\nabla V\|_{\mathscr{L}^\infty(\Omega)}$, $\|\nabla U\|_{\mathscr{L}^\infty(\Omega\times\Omega)}$, $\|\nabla^2 V\|_{\mathscr{L}^\infty(\Omega)}$, $\|\nabla^2 U\|_{\mathscr{L}^\infty(\Omega\times\Omega)}$.

5.2 Convergence to the solutions of porous medium equation

We next prove that the solution of the PDE (3.9) converges, as $\delta \to 0$, to the unique solution of the porous medium equation (3.12). As for Theorem 5.1, this result does not rely on the concavity assumption for f.

Theorem 5.2. *Assume that conditions* (A1) *and* (A3)-(A5) *hold. Denote by* ρ^δ *the unique solution of the PDE* (3.9) *with initial condition* $\rho_0^\delta = \rho_{\text{init}}$. *Then*

(a) *The porous medium equation* (3.12) *admits a weak solution* $\rho : (t, \boldsymbol{x}) \mapsto \rho_t(\boldsymbol{x})$ *with initial and boundary conditions* (3.13). *Further, this solution is unique under the additional condition* $\rho \in \mathscr{L}^4([0,T] \times \Omega)$.

(b) *For almost all* $t \in [0,T]$, *we have* $\rho_t^\delta \to \rho_t$ *in* $\mathscr{L}^2(\Omega)$ *as* $\delta \to 0$.

While this statement is very natural at a heuristic level, its proof is actually the bulk of our technical work. Similar approximation results have been proved in the past by Oelschläger, Philipowski, Figalli [Oel02, Phi07, FP08], but they do not apply directly to the present case unless $f = 0$ (also, we have to deal with different boundary conditions).

Our proof follows a classical compactness argument, generalizing the approach of [FP08]. Namely we consider the sequence of trajectories $(\rho_t^\delta)_{t\in[0,T]}$ indexed by the width δ. We prove that that this family is bounded and equicontinuous in $\mathscr{C}([0,T], \mathscr{P}_2(\Omega))$, and hence admits converging subsequences $(\rho_t^{\delta_n})_{t\in[0,T]} \to (\rho_t)_{t\in[0,T]}$. We next prove that any such converging subsequence converges in $\mathscr{L}^2(\Omega \times [0,T])$ and that the limit is a weak solution of the porous medium equation (3.12). Unfortunately, uniqueness of weak solutions of the PME (3.12) is –to the best of our knowledge– an open problem. However, we generalize methods from [Oel02] to show that any subsequential limit is actually in $\mathscr{L}^4(\Omega \times [0,T])$, and prove that the weak solution is unique under this condition. This allows us to conclude that $(\rho_t^\delta)_{t\in[0,T]}$ converges to this unique weak solution $(\rho_t)_{t\in[0,T]}$.

5.3 Global convergence of SGD

Let us now state the main result of this paper: SGD converges to a model with nearly optimal risk.

Theorem 5.3. *Assume that conditions* (A1)-(A5) *hold, and recall that* $\alpha > 0$ *is the concavity parameter of the function* f, *i.e.,* $\langle \boldsymbol{y}, \nabla^2 f(\boldsymbol{x})\boldsymbol{y}\rangle \le -\alpha|\boldsymbol{y}|^2$ *for all* $\boldsymbol{x} \in \Omega$, $\boldsymbol{y} \in \mathbb{R}^d$.

Consider the SGD update (3.5) *with initialization* $(\boldsymbol{w}_i^0)_{i\le N} \sim_{\text{i.i.d.}} \rho_{\text{init}}$ *and constant step size* ε. *Assume* $\text{supp}(\rho_{\text{init}}) \subseteq \mathsf{B}(\boldsymbol{0}; r)$. *Then, for any* $k \le T/\varepsilon$, *the following holds with probability at least* $1 - 1/z$,

$$R_N(\boldsymbol{w}^k) \le R_N(\boldsymbol{w}^0) e^{-2\alpha k\varepsilon} + 2\tau\, \Delta'(k,\varepsilon,d) + \Delta(N,\varepsilon,T,d,\delta,z)\,, \tag{5.4}$$

where

$$\Delta'(k,\varepsilon,d) = \log|\Omega| - (1 - e^{-2\alpha k\varepsilon}) S(f) - S(\rho_{\text{init}})\, e^{-2\alpha k\varepsilon}\,, \tag{5.5}$$

$$\lim_{\delta\to 0} \lim_{N\to\infty,\varepsilon\to 0} \Delta(N,\varepsilon,T,d,\delta,z) = 0\,. \tag{5.6}$$

Remark 5.3. The error term $2\tau\,\Delta'(k,\varepsilon,d)$ in Eq. (5.4) is always non-negative. In fact, $\Delta'(k,\varepsilon,d) \geq 0$ as $S(\rho) \leq \log|\Omega|$ for any $\rho \in \mathcal{P}_2(\Omega)$. Furthermore, by applying Jensen's inequality, we have that, for any $\rho \in \mathcal{P}_2(\Omega)$,

$$S(\rho) = -\int \rho(\boldsymbol{x}) \log \rho(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \geq -\log \int \rho(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} = -2\,\log \|\rho\|_{\mathscr{L}^2(\Omega)},$$

which gives the following upper bound

$$\Delta'(k,\varepsilon,d) \leq \log|\Omega| + 2\,\left|\log \|f\|_{\mathscr{L}^2(\Omega)}\right| + 2\,\left|\log \|\rho_{\mathrm{init}}\|_{\mathscr{L}^2(\Omega)}\right|.$$

Recall that τ controls the variance of the noise, which is added at each step of the SGD algorithm for technical purposes. Thus, we can take τ sufficiently small so that the term $2\tau\Delta'(k,\varepsilon,d)$ is arbitrarily small.

Remark 5.4. The proof of Theorem 5.3 provides a somewhat more explicit expression for the error term $\Delta(N,\varepsilon,T,d,\delta,z)$ in Eq. (5.4). Namely, for an arbitrary but fixed $p \in \mathbb{N}$,

$$\Delta(N,\varepsilon,T,d,\delta,z) = \Delta_1(N,\varepsilon,T,d,z) + \Delta_2(\delta,T,d)\,, \tag{5.7}$$

$$\Delta_1(N,\varepsilon,T,d,z) = \left(\sqrt{\frac{d}{N}} \vee \left(r\delta^{-2d-1}(d^2\varepsilon\log(1/\varepsilon))^{1/4}\right)\right) \cdot \exp\left\{\sqrt{2C_*\delta^{-(d+2)}\,T\log(z)}\right\} \tag{5.8}$$

$$\lim_{\delta\to 0} \Delta_2(\delta,T,d) = 0\,. \tag{5.9}$$

The term Δ_1 bounds the error due to describing the SGD dynamics using the PDE (3.9). It vanishes when $N \to \infty$, $\varepsilon \to 0$, under the stated conditions. The term Δ_2 captures the error due to approximating the PDE (3.9) with the porous medium equation (3.12). Finally, the term $e^{-2\alpha k\varepsilon}$ describes the convergence to equilibrium of the solution of the porous medium equation.

The proof of Theorem 5.3 is presented in Appendix F and relies crucially on regularity results for the PDE (3.9) which are established in Appendix E.

More specifically, the proof is based on three steps, which we spell out once more:

(*i*) We approximate the dynamics of SGD by the PDE (3.9) at $\delta > 0$ fixed. In doing so, we incur an error Δ_1 which is controlled using Theorem 5.1.

(*ii*) We approximate the solution ρ_t^δ of the PDE (3.9) at $\delta > 0$ using the solution ρ_t of the porous medium equation (3.12), as stated in Theorem 5.2.

(*iii*) We use results from [CJM+01, CMV03, CMV06] to prove that the latter solution converges exponentially fast to the global optimum, with rate $O(e^{-2\alpha t})$.

Given Theorems 5.1, 5.2, and the results of [CJM+01, CMV03, CMV06], this proof is relatively direct. We emphasize that, unlike Theorems 5.1, 5.2, the proof Theorem 5.3 relies in a crucial way on our structural assumptions, namely the concavity of f, and the structure of the bump-like activation $K_\delta(\boldsymbol{x} - \boldsymbol{w}_i)$.

Remark 5.5. If we settle for the less ambitious goal of proving global convergence without the explicit dimension-independent rate $e^{-2\alpha k\varepsilon}$, and there are no boundary conditions ($\Omega = \mathbb{R}^d$), we can achieve this goal using [MMN18, Theorem 5]. This result guarantees convergence in a number of SGD steps that potentially depends on τ (the noise injected in SGD) as well as the dimensions d, and the width δ, but does not require to assume strong concavity of f. On the other hand, numerical experiments are consistent with the conclusion that rates are independent of these parameters, cf. e.g. Fig. 1 where dependence on δ is explored.

6 Discussion

It is instructive to compare the general strategy followed in this paper (and in related work, e.g. [MMN18, MMM19]) and the results we obtain, to a more classical approach in theoretical statistics. For the sake of clarity, we will abstract away most of the details of the present problem, and focus on the most important differences.

Consider a general setting in which we want to minimize the population risk $R(\boldsymbol{w}) = \mathbb{E}_{y,\boldsymbol{x}} L(\boldsymbol{w}; y, \boldsymbol{x})$, where L is a non-convex loss function and $\boldsymbol{w} \in \mathbb{R}^D$ are parameters (in our problem $\boldsymbol{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_N)$ are the first-layer weights and $D = dN$). We are given n i.i.d. samples $\{(y_j, \boldsymbol{x}_j)\}_{j\le n}$.

A standard theoretical analysis of this problem uses empirical risk minimization. Namely, we define the empirical risk $\widehat{R}_n(\boldsymbol{w}) = \widehat{\mathbb{E}}_{y,\boldsymbol{x}} L(\boldsymbol{w}; y, \boldsymbol{x})$ (with $\widehat{\mathbb{E}}_n$ denoting the empirical average), and compute the minimizer $\hat{\boldsymbol{w}}_n \in \arg\min_{\boldsymbol{w}} \widehat{R}_n(\boldsymbol{w})$, for instance by gradient descent. Theoretical analysis proceeds –conceptually– in two steps. First, one proves that the empirical risk minimizer is a near-minimizer of the population risk. Namely

$$R(\hat{\boldsymbol{w}}_n) \le \min_{\boldsymbol{w}} R(\boldsymbol{w}) + \mathsf{err}(D, n)\,. \tag{6.1}$$

This is normally proved through a uniform convergence argument to establish a bound $\sup_{\boldsymbol{w}} |\widehat{R}_n(\boldsymbol{w}) - R(\boldsymbol{w})| \le \mathsf{err}(D, n)/2$. Here $\mathsf{err}(D, n)$ is an error term that (hopefully) vanishes as $n \to \infty$ for D fixed. Second, one proves that gradient descent (with respect to the cost function $\widehat{R}_n$) converges to a minimizer $\hat{\boldsymbol{w}}_n$. This is achieved by showing that, with high probability, the landscape $\boldsymbol{w} \mapsto \widehat{R}_n(\boldsymbol{w})$ satisfies some strong conditions that guarantee convergence of gradient descent (or other algorithms). For instance, one desirable (although not sufficient) property is that $\widehat{R}_n$ does not have local minima other than the global minima, provided that the sample size is large enough. A substantial literature applies this general scheme (with significant refinements) to a variety of non-convex problems in high-dimensional statistics, including phase retrieval, clustering, matrix completion, error-in-variables models, and so on. We refer to [MBM+18] for examples and a more detailed survey.

Unfortunately this approach runs into substantial difficulties when treating complex models such as multi-layer neural networks. We can name at least two sources of difficulties. First of all, the number of parameters D in the model is often comparable with the sample size n, and therefore uniform convergence of the empirical risk to population risk does not hold. For instance, in the present model, we could use a number of parameters $Nd \gtrsim n$: indeed, such an example is considered in Figure 5-(a), where $Nd = 800$ and $n \in \{100, \dots, 2000\}$. Of course this problem can be addressed by constraining other measures of complexity than the number of parameters [Bar98], but the common practice is not to add such regularizers in the training.

The second source of difficulties is that studying the risk landscape, and ruling out local minima is extremely difficult, even if we limit ourselves to the $n = \infty$ limit, i.e. the population risk $R(\boldsymbol{w})$. In two-layers neural networks, part of this difficulty is due to the fact that the risk (1.2) is invariant under permutations of the N neurons, and hence it has (generically) at least $N!$ global minima related by permutations, and a large number of saddle points connecting them.

The approach pursued in this paper builds on two simple remarks, which are connected to the previous difficulties:

(*i*) Uniform convergence of the empirical risk $\widehat{R}_n(\boldsymbol{w})$ to the population risk $R(\boldsymbol{w})$ is not necessary, nor it is necessary to control the random deviations of the whole landscape of the empirical risk. What is instead important is to control the landscape of the empirical risk along the trajectory of gradient descent from a given initialization.

A convenient way to implement this idea is to consider SGD in a one-pass setting in which each sample is used only once. In the limit of small step size, this converges to gradient flow with respect to $R(\boldsymbol{w})$.

(*ii*) Absence of local minima in the population landscape $R(\boldsymbol{w})$ is not necessary either. What is instead important is absence of local minima along the gradient flow trajectory for $R(\boldsymbol{w})$ or, more precisely, the fact that the gradient flow trajectory converges to a global minimum.

These remarks suggest the following proof strategy. Let $\boldsymbol{w}(t)$ denote the gradient flow trajectory from a given initialization $\boldsymbol{w}(0) = \boldsymbol{w}_0$ (namely $\dot{\boldsymbol{w}}(t) = -\nabla R(\boldsymbol{w}(t))$), and $\boldsymbol{w}^k$ be the (random) parameters produced after k SGD steps. We first prove that gradient flow converges to a global optimum, possibly with explicit convergence rate $\Delta(t)$:

$$R(\boldsymbol{w}(t)) \le \min_{\boldsymbol{w}} R(\boldsymbol{w}) + \Delta(t)\,, \tag{6.2}$$

where $\Delta(t) \to 0$ as $t \to \infty$. We then show that the SGD trajectory, after k steps, is well approximated by the gradient flow for $R(\boldsymbol{w})$ provided the step size ε is small. For instance we might prove that there exists a numerical constant c_0 such that, for any $k\varepsilon \le T$, with high probability

$$\big|R(\boldsymbol{w}^k) - R(\boldsymbol{w}(k\varepsilon))\big| \le \varepsilon^{c_0}\,\mathsf{err}(T)\,. \tag{6.3}$$

The reader might recognize that the last estimate is analogous to the one obtained in Theorem 5.1, while the estimate 6.2 is what we obtain from displacement convexity (after taking the limit $\delta \to 0$ using Theorem 5.2). Putting the two estimates together, and recalling that we can run a total of n SGD steps (in the one-pass setting), we get

$$R(\hat{\boldsymbol{w}}) \le \min_{\boldsymbol{w}} R(\boldsymbol{w}) + \Delta(n\varepsilon) + \varepsilon^{c_0}\,\mathsf{err}(n\varepsilon)\,, \tag{6.4}$$

where we set $\hat{\boldsymbol{w}} = \boldsymbol{w}^k$. The error is reminiscent of a bias-variance tradeoff: the first term is a bias due to early stopping; the second is instead the stochastic approximation error. We can now optimize n as to minimize this error. For instance, if $\Delta(t) = e^{-c_1 t}$, and $\mathsf{err}(T) = e^{c_2 T}$, we can choose $\varepsilon \propto (\log n/n)$, yielding $R(\hat{\boldsymbol{w}}) \le \min_{\boldsymbol{w}} R(\boldsymbol{w}) + C(\log n)^{c_0}/n^{c'}$ where $c' = c_0 c_1/(c_1 + c_2)$.

In summary, within the present approach, the generalization error is bounded via a tradeoff between the convergence rate of gradient flow in the population risk, and the error of approximating

the gradient flow by SGD. A side benefit of this proof strategy is that it guarantees the existence of an efficient algorithm to compute the weights $\hat{\boldsymbol{w}}$.

As mentioned, the above discussion omits several challenges that are posed by the model treated in this paper. Most notably: (1) We are trying to optimize N weight vectors $\boldsymbol{w}_1, \dots, \boldsymbol{w}_N \in \mathbb{R}^d$, but the loss only depends on the empirical distribution of these vectors $\hat{\rho}^{(N)} = N^{-1} \sum_{i=1}^N \delta_{\boldsymbol{w}_i}$. It is therefore natural to define a gradient flow in the space of probability distributions, which is nothing but the PDE (3.9). This also help addressing the challenge posed by by the fact that, as N increases, the dimension of the parameter space increases and convergence to the population behavior might fail. We are embedding all the values of N in the space $\mathscr{P}(\mathbb{R}^d)$. (2) We cannot prove a bound of the form (6.2) for the original PDE (3.9) and have to approximate this by the porous medium equation (3.12).

Because of these additional challenges, our bounds are not nearly as neat as in Eqs. (6.2), 6.3 and depend on the additional parameters d, δ: in particular, the approximation by the porous medium equation in Theorem 5.2 is non-quantitative. We therefore refrain from optimizing the tradeoff between convergence rate of gradient flow, and error in stochastic approximation, which would result in suboptimal statistical guarantees, and defer this objective to future work.

Acknowledgements

A. Javanmard was partially supported by an Outlier Research in Business (iORB) grant from the USC Marshall School of Business, a Google Faculty Research award and the NSF CAREER award DMS-1844481. M. Mondelli was supported by an Early Postdoc.Mobility fellowship from the Swiss National Science Foundation and by the Simons Institute for the Theory of Computing. A. Montanari was partially supported by grants NSF DMS-1613091, CCF-1714305, IIS-1741162 and ONR N00014-18-1-2729. This work was carried out in part while the authors were visiting the Simons Institute for the Theory of Computing.

A Uniqueness of weak solutions of limit PDE ($\delta = 0$)

In this appendix, we prove that the limit PDE obtained for $\delta \to 0$, namely the porous medium equation (3.12) has at most one solution in $\mathscr{L}^4(\Omega \times [0,T])$. Existence of such solutions will follow from the results of Appendix F, and in particular from Lemma F.4.

For the sake of clarity, we repeat the definitions of Section 3.5. Let $\Omega \subseteq \mathbb{R}^d$ be a compact convex set with $\mathscr{C}^2$ boundary. We denote by $\mathscr{P}_2(\Omega)$ the space of probability measures on Ω endowed with Wasserstein's W_2 distance. Since Ω is compact, the induced topology is equivalent to weak convergence. We consider the following PDE:

$$\partial_t \rho_t(\boldsymbol{w}) = -\nu_0 \nabla \cdot \big(\rho_t(\boldsymbol{w}) \nabla f(\boldsymbol{w})\big) + \frac{\nu_0}{2} \Delta(\rho_t^2(\boldsymbol{w})) + \tau \Delta \rho_t(\boldsymbol{w})\,, \tag{A.1}$$

with initial and boundary conditions

$$\begin{aligned} \rho_0 &= \rho_{\mathrm{init}}\,, \\ \Big\langle \boldsymbol{n}(\boldsymbol{w}), \nu_0 \rho_t(\boldsymbol{w})\, \nabla(f(\boldsymbol{w}) - \rho_t(\boldsymbol{w})) - \tau \nabla \rho_t(\boldsymbol{w}) \Big\rangle &= 0 \qquad \forall \boldsymbol{w} \in \partial\Omega\,. \end{aligned} \tag{A.2}$$

Throughout this appendix, we adopt the notation $\Phi(\rho) = \tau\rho + \nu_0\, \rho^2/2$. Let us formally define the concept of weak solutions for the PDE (A.1).

For the next statement, it is useful to recall that $\mathscr{C}^{2,1}(\Omega \times [0,T])$ denotes the class of functions $f : \Omega \times [0,T] \to \mathbb{R}$ with continuous partial derivatives $D_t f(\boldsymbol{x},t)$, $D^{\boldsymbol{\alpha}}_{\boldsymbol{x}} f(\boldsymbol{x},t)$ for all $\|\boldsymbol{\alpha}\|_2 \le 2$.

Definition A.1 (Weak solution of limit PDE)**.** *We say that $\rho \in \mathscr{C}([0,T],\, \mathscr{P}_2(\Omega))$ is a* weak solution *of the PDE (A.1), with initial and boundary conditions (A.2) if*

1. *ρ_t has density $\rho(\,\cdot\,, t)$ with respect to Lebesgue measure, and $\rho \in \mathscr{L}^2(\Omega \times [0,T])$.*
2. *For any test function $h \in \mathscr{C}^{2,1}(\Omega \times [0,T])$, satisfying $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h(\boldsymbol{x},t)\rangle = 0$ for all $\boldsymbol{x} \in \partial\Omega, t \in [0,T]$, we have*

$$\begin{aligned} &\int_\Omega h(\boldsymbol{x},T)\, \rho(\boldsymbol{x},T)\, \mathrm{d}\boldsymbol{x} - \int_\Omega h(\boldsymbol{x},0)\, \rho_{\mathrm{init}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &\quad = \int_0^T \int_\Omega \Big[\partial_t h(\boldsymbol{x},t) + \nu_0 \langle \nabla f(\boldsymbol{x}), \nabla h(\boldsymbol{x},t)\rangle\Big]\, \rho(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t \\ &\quad + \int_0^T \int_\Omega \Delta h(\boldsymbol{x},t) \Phi(\rho)(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t\,. \end{aligned} \tag{A.3}$$

We now prove a uniqueness result, under a mild integrability condition.

Lemma A.2 (Uniqueness of limit PDE)**.** *Let $\rho, \tilde\rho \in \mathscr{L}^4(\Omega \times [0,T])$ be two weak solutions of the PDE (A.1) with initial and boundary conditions (A.2), in the sense of Definition A.1. Then, $\rho = \tilde\rho$, almost everywhere.*

Proof. Note that setting $\nu_0 = 1$ corresponds to scaling time by a factor ν_0 and to substituting τ with $\tau\,\nu_0$. Since the proof holds for any $\tau > 0$, without loss of generality we can set $\nu_0 = 1$.

The proof follows ideas from [Váz07, Theorem 6.5]. We write the identity (A.3) for ρ and $\tilde{\rho}$ and subtract them to get

$$\begin{aligned}
&\int_\Omega h_T(\boldsymbol{x})\,(\rho_T(\boldsymbol{x}) - \tilde{\rho}_T(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} \qquad (A.4)\\
&\quad = \int_0^T \int_\Omega \Big[\partial_t h_t(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle\Big]\,(\rho_t(\boldsymbol{x}) - \tilde{\rho}_t(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\\
&\quad + \int_0^T \int_\Omega \Delta h_t(\boldsymbol{x})(\Phi(\rho_t(\boldsymbol{x})) - \Phi(\tilde{\rho}_t(\boldsymbol{x})))\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\,,
\end{aligned}$$

where we use the shorthand $\rho_t(\boldsymbol{x}) \equiv \rho(\boldsymbol{x}, t)$ and $h_t(\boldsymbol{x}) \equiv h(\boldsymbol{x}, t)$. Define $u_t = \rho_t - \tilde{\rho}_t$ and $\eta_t = \tau + (\rho_t + \tilde{\rho}_t)/2$. Then,

$$\begin{aligned}
\int_\Omega h_T(\boldsymbol{x})\,u_T(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int_0^T \int_\Omega \Big[\partial_t h_t(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle\Big]\,u_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \qquad (A.5)\\
&\quad + \int_0^T \int_\Omega \Delta h_t(\boldsymbol{x})\eta_t(\boldsymbol{x})u_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\,.
\end{aligned}$$

Note that $\eta_t(\boldsymbol{x}) \geq \tau$ and define the truncated function $\eta_t^M = \min(M, \eta_t)$. We next choose a smooth test function $\theta : \Omega \times [0,T] \to \mathbb{R}_{\geq 0}$, $(\boldsymbol{x}, t) \mapsto \theta_t(\boldsymbol{x})$ and consider the following backward problem:

$$\begin{cases}
\partial_t h_t(\boldsymbol{x}) + \hat{\eta}_t(\boldsymbol{x})\Delta h_t(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle + \theta_t(\boldsymbol{x}) = 0\,, & \forall \boldsymbol{x} \in \Omega, t \in [0,T],\\
\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle = 0\,, & \forall \boldsymbol{x} \in \partial\Omega, t \in [0,T],\\
h_T(\boldsymbol{x}) = 0\,, & \forall \boldsymbol{x} \in \Omega\,.
\end{cases} \qquad (A.6)$$

Here, $\hat{\eta}_t$ is a smooth approximation of η_t^M, such that $\tau \leq \hat{\eta}_t(\boldsymbol{x}) \leq M$. (We will make precise below in what sense $\hat{\eta}_t$ has to approximate η_t^M. For the moment, it can be a general smooth function satisfying the bounds $\tau \leq \hat{\eta}_t(\boldsymbol{x}) \leq M$.) Note that (A.6) is a backward parabolic problem with smooth coefficients and with Neumann boundary conditions. Hence, by classical results on quasilinear parabolic PDEs [LSU88], it admits a solution $h_t \in \mathscr{C}^{2,1}(\Omega \times [0,T])$. Rewriting (A.5) for such a test function h_t, we get

$$\int_0^T \int_\Omega \Delta h_t(\boldsymbol{x})(\eta_t(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x}))u_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t = \int_0^T \int_\Omega \theta_t(\boldsymbol{x})u_t(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\,.$$

This immediately implies that

$$\int_0^T \int_\Omega \theta_t(\boldsymbol{x})u_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \leq \int_0^T \int_\Omega |\Delta h_t(\boldsymbol{x})|\,|\eta_t(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x})||u_t(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\,. \qquad (A.7)$$

By applying Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
&\int_0^T \int_\Omega |\Delta h_t(\boldsymbol{x})|\,|\eta_t(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x})|\,|u_t(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\\
&\quad \leq \left(\int_{\Omega\times[0,T]} \hat{\eta}_t(\boldsymbol{x})\,(\Delta h_t(\boldsymbol{x}))^2\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\right)^{1/2}\\
&\qquad \left(\int_{\Omega\times[0,T]} \frac{|\eta_t(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x})|^2}{\hat{\eta}_t(\boldsymbol{x})} u_t^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\right)^{1/2}. \qquad (A.8)
\end{aligned}$$

To bound the first term on the right-hand side of (A.8), we consider a smooth positive bounded function $\mu(t)$, defined on $[0,T]$, whose properties will be discussed later. Define the shorthand $\tilde{\theta}_t(\boldsymbol{x}) \equiv \theta_t(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \nabla h_t(\boldsymbol{x}) \rangle$. We multiply the parabolic PDE (A.6) by $\mu(t)\Delta h_t(\boldsymbol{x})$ and integrate to obtain

$$\int_{\Omega\times[0,T]} \mu(t)\partial_t h_t(\boldsymbol{x})\Delta h_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_{\Omega\times[0,T]} \mu(t)\hat{\eta}_t(\boldsymbol{x})(\Delta h_t(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_{\Omega\times[0,T]} \mu(t)\tilde{\theta}_t(\boldsymbol{x})\Delta h_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t = 0. \tag{A.9}$$

We next write

$$\begin{aligned}
&\int_{\Omega\times[0,T]} \mu(t)\partial_t h_t(\boldsymbol{x})\Delta h_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\overset{(a)}{=} -\int_{\Omega\times[0,T]} \mu(t)\langle \nabla h_t(\boldsymbol{x}), \nabla(\partial_t h_t(\boldsymbol{x})\rangle \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= -\int_{\Omega\times[0,T]} \mu(t)\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\nabla h_t(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\overset{(b)}{=} \frac{1}{2}\int_{\Omega} \mu(0)|\nabla h_0(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int_{\Omega} \mu(T)|\nabla h_T(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x} \\
&\qquad + \frac{1}{2}\int_{\Omega\times[0,T]} \mu'(t)|\nabla h_t(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\overset{(c)}{\geq} \frac{1}{2}\int_{\Omega\times[0,T]} \mu'(t)|\nabla h_t(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t
\end{aligned} \tag{A.10}$$

Here (a) follows from integration by parts in the integral over Ω and using the fact that $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle = 0$ for $\boldsymbol{x} \in \partial\Omega$ and $t \in [0,T]$. Also, (b) follows from integration by parts in the integral over t. Finally (c) holds because $h_T(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \Omega$ and $\mu(0) \geq 0$.

Getting back to (A.9) and using the properties of function $\mu(t)$, we have

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega\times[0,T]} \mu'(t)|\nabla h_t(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_{\Omega\times[0,T]} \mu(t)\hat{\eta}(\boldsymbol{x})(\Delta h_t(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\leq -\int_{\Omega\times[0,T]} \mu(t)\tilde{\theta}_t(\boldsymbol{x})\Delta h_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= -\int_{\Omega\times[0,T]} \mu(t)\theta_t(\boldsymbol{x})\Delta h_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad -\int_{\Omega\times[0,T]} \mu(t)\langle \nabla f(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \Delta h_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= \int_{\Omega\times[0,T]} \mu(t)\langle \nabla\theta_t(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad -\int_{\Omega\times[0,T]} \mu(t)\langle \nabla f(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \Delta h_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\leq \int_{\Omega\times[0,T]} \mu(t)\langle \nabla\theta_t(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_{\Omega\times[0,T]} \frac{\mu(t)}{2\tau}|\nabla f(\boldsymbol{x})|^2|\nabla h_t(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&+ \int_{\Omega\times[0,T]} \mu(t)\frac{\tau}{2}(\Delta h_t(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\,,
\end{aligned} \tag{A.11}$$

The penultimate step follows from integration by parts and the constraint $\langle \nabla h_t(\boldsymbol{x}), \boldsymbol{n}(\boldsymbol{x})\rangle = 0$, for $\boldsymbol{x} \in \partial\Omega$ and $t \in [0,T]$, and the last step follows by applying Cauchy-Schwartz inequality. We continue by applying Cauchy-Schwartz inequality again to get

$$\begin{aligned}
&\int_{\Omega\times[0,T]} \mu(t)\langle \nabla\theta_t(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\leq \int_{\Omega\times[0,T]} \left[\frac{\tau\mu(t)}{2C^2}|\nabla\theta_t(\boldsymbol{x})|^2 + \frac{C^2\mu(t)}{2\tau}|\nabla h_t(\boldsymbol{x})|^2\right]\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\,,
\end{aligned} \tag{A.12}$$

where $C = \sup_{\boldsymbol{x}\in\Omega}|\nabla f(\boldsymbol{x})|$. Combining Equations (A.11) and (A.12), we get

$$\begin{aligned}
&\frac{1}{2}\int_{\Omega\times[0,T]} \left(\mu'(t) - \frac{\mu(t)}{\tau}\left(|\nabla f(\boldsymbol{x})|^2 + C^2\right)\right)|\nabla h_t(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&+ \int_{\Omega\times[0,T]} \mu(t)\left(\hat{\eta}(\boldsymbol{x}) - \frac{\tau}{2}\right)(\Delta h_t(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq \frac{\tau\mu_{\max}}{2C^2}\int_{\Omega\times[0,T]} |\nabla\theta_t(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\,,
\end{aligned} \tag{A.13}$$

where $\mu_{\max} = \sup_{t\in[0,T]}\mu(t)$. We find a smooth function $\mu(t)$ such that

1. $\mu(t) \geq \mu_{\min} > 0$, for $t \in [0,T]$,
2. $\mu'(t) - \frac{2C^2}{\tau}\mu(t) \geq 0$.

A particular choice is

$$\mu(t) = \mu_{\min} e^{\frac{2C^2}{\tau}t}.$$

We then obtain from (A.13) that

$$\int_{\Omega\times[0,T]} \hat{\eta}(\boldsymbol{x})(\Delta h_t(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq \frac{\tau\mu_{\max}}{\mu_{\min}C^2}\int_{\Omega\times[0,T]} |\nabla\theta_t(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\,. \tag{A.14}$$

Now by employing (A.14) in bound (A.8) combined with (A.7) we get

$$\begin{aligned}
\int_0^T\int_\Omega \theta_t(\boldsymbol{x})u_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \leq &\frac{1}{C}\sqrt{\frac{\tau\mu_{\max}}{\mu_{\min}}}\|\nabla\theta\|_{\mathscr{L}^2(\Omega\times[0,T])} \\
&\left(\int_{\Omega\times[0,T]} \frac{|\eta_t(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x})|^2}{\hat{\eta}_t(\boldsymbol{x})} u_t^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\right)^{1/2}.
\end{aligned} \tag{A.15}$$

Next we note that

$$\begin{aligned}
&\int_{\Omega\times[0,T]} |\eta_t(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x})|^2 u_t^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\leq 2\int_{\Omega\times[0,T]} |\eta_t^M(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x})|^2 u_t^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&+ 2\int_{\Omega\times[0,T]} ((\eta_t(\boldsymbol{x}) - M)_+)^2 u_t^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}$$

Call the first integral I_1 and denote the second one by I_2. The integrand in I_2 is pointwise bounded by

$$2\eta_t^2(\boldsymbol{x})u_t^2(\boldsymbol{x})\,\mathbb{I}(\eta_t(\boldsymbol{x}) > M) \leq 2(\Phi(\rho_t(\boldsymbol{x})) - \Phi(\tilde{\rho}_t(\boldsymbol{x})))^2\,\mathbb{I}(\eta_t(\boldsymbol{x}) > M)\,.$$

Since $\rho_t, \tilde{\rho}_t \in \mathscr{L}^4$, we have that $(\Phi(\rho_t) - \Phi(\tilde{\rho}_t))^2$ has bounded integral. Hence, we can choose M large enough such that I_2 is arbitrarily small. Moreover we can choose the smooth approximation $\hat{\eta}_t$ such that I_1 is also arbitrarily small. Putting everything together, we obtain that

$$\int_{\Omega\times[0,T]} |\eta_t(\boldsymbol{x}) - \hat{\eta}_t(\boldsymbol{x})|^2 u_t^2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \le \varepsilon\,,$$

where ε is an arbitrary small fixed constant.

In addition, since $\hat{\eta}_t(\boldsymbol{x}) \ge \tau$, invoking (A.15) we have

$$\int_0^T \int_\Omega \theta_t(\boldsymbol{x}) u_t(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \le \frac{1}{C}\sqrt{\frac{\mu_{\max}}{\mu_{\min}}} \|\nabla\theta\|_{\mathscr{L}^2(\Omega\times[0,T])} \sqrt{\varepsilon}\,.$$

Since $\frac{\mu_{\max}}{\mu_{\min}} = e^{\frac{2C^2}{\tau}T} < \infty$ and θ are independent of ε, by choosing ε arbitrarily small, we conclude that

$$\int_0^T \int_\Omega \theta_t(\boldsymbol{x}) u_t(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \le 0\,.$$

Since $\theta_t(\boldsymbol{x}) \ge 0$ was an arbitrary smooth function supported on $\Omega \times [0,T]$, this implies that $u \le 0$, almost everywhere. By repeating a similar argument, we get $u \ge 0$, almost everywhere. The result follows. □

B General results on the PDE (3.9) ($\delta > 0$)

This appendix contains some basic results on the PDE (3.9). Although these facts are standard, we collect them here for the reader's convenience.

In fact, we will consider a more general PDE, which also includes as a special case the one studied in [MMN18]. We consider a compact convex domain D, with a non-empty interior. The general PDE is parametrized by two functions $V \in \mathscr{C}^2(D)$ and $U \in \mathscr{C}^2(D \times D)$, with $U(\boldsymbol{x}_1, \boldsymbol{x}_2) = U(\boldsymbol{x}_2, \boldsymbol{x}_1)$. (Unlike in [MMN18], we consider the case of a compact domain with Neumann boundary conditions.) Given $\rho \in \mathscr{P}_2(D)$, we define

$$\Psi(\boldsymbol{x}, \rho) \equiv V(\boldsymbol{x}) + \int U(\boldsymbol{x}, \tilde{\boldsymbol{x}})\, \rho(\mathrm{d}\tilde{\boldsymbol{x}})\,, \tag{B.1}$$

and consider the PDE

$$\partial_t \rho(\boldsymbol{x}, t) = \nabla \cdot (\rho(\boldsymbol{x}, t) \nabla \Psi(\boldsymbol{x}, \rho_t)) + \tau\, \Delta \rho(\boldsymbol{x}, t)\,, \tag{B.2}$$

with initial and boundary conditions

$$\begin{aligned} \rho_0 &= \rho_{\text{init}}\,, \\ \langle \boldsymbol{n}(\boldsymbol{x}), \rho_t(\boldsymbol{x}) \nabla \Psi(\boldsymbol{x}, \rho_t) + \tau\, \nabla \rho_t(\boldsymbol{x}) \rangle &= 0\,, \qquad \forall \boldsymbol{x} \in \partial D\,. \end{aligned} \tag{B.3}$$

We will typically write $\rho_t(\,\cdot\,)$ for a solution of this equation, in order to emphasize that it is a function of t that takes values in $\mathscr{P}_2(D)$, and $\rho(\boldsymbol{x}, t)$ for the corresponding density, viewed as a function on $D \times [0,T]$. Let us formally define the concept of weak solutions for the PDE (B.2).

Note that the PDE (3.9) is a special case of this setting with $D = \Omega^\delta$, and $V(\boldsymbol{w})$ and $U(\boldsymbol{w}_1, \boldsymbol{w}_2) = U(\boldsymbol{w}_1 - \boldsymbol{w}_2)$ defined as follows:

$$\begin{aligned} V(\boldsymbol{w}) &\equiv -\nu_0\, K^\delta * f(\boldsymbol{w}) = -\nu_0 \int K^\delta(\boldsymbol{w} - \boldsymbol{x})\, f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\,, \\ U(\boldsymbol{w}) &\equiv \nu_0\, K^\delta * K^\delta(\boldsymbol{w}) = \nu_0 \int K^\delta(\boldsymbol{w} - \boldsymbol{x})\, K^\delta(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\,. \end{aligned} \tag{B.4}$$

Remark B.1. For the special choice of V and U given by (B.4) the following properties hold:

1. $V : \Omega^\delta \to \mathbb{R}$ is convex for any $\delta > 0$.
2. $\lim_{\delta \to 0} \sup_{\boldsymbol{w} \in \Omega^\delta} |V(\boldsymbol{w}) + \nu_0\, f(\boldsymbol{w})| = 0$.
3. $U(\boldsymbol{w}) = \nu_0\, \delta^{-2d} K^{(2)}(\boldsymbol{w}/\delta)$, where $K^{(2)} = K * K$.

Proof. We have $V(\boldsymbol{w}) = -\nu_0 \int K^\delta(\boldsymbol{x}) f(\boldsymbol{w} - \boldsymbol{x}) \mathrm{d}\boldsymbol{x}$. Hence,

$$\begin{aligned} V(\lambda \boldsymbol{w} + (1-\lambda)\boldsymbol{w}') &= -\nu_0 \int K^\delta(\boldsymbol{x}) f(\lambda \boldsymbol{w} + (1-\lambda)\boldsymbol{w}' - \boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= -\nu_0 \int K^\delta(\boldsymbol{x}) f(\lambda(\boldsymbol{w} - \boldsymbol{x}) + (1-\lambda)(\boldsymbol{w}' - \boldsymbol{x})) \mathrm{d}\boldsymbol{x} \\ &\le -\nu_0 \int K^\delta(\boldsymbol{x}) \left(\lambda f(\boldsymbol{w} - \boldsymbol{x}) + (1-\lambda) f(\boldsymbol{w}' - \boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \\ &= \lambda V(\boldsymbol{w}) + (1-\lambda) V(\boldsymbol{w}')\,. \end{aligned}$$

This proves that $V(\boldsymbol{w})$ is convex. The next two properties are straightforward. □

Definition B.1 (Weak solution of PDE)**.** *We say that $\rho : [0,T] \to \mathscr{P}_2(D)$ is a* weak solution *of (B.2) with initial and boundary conditions (B.3) if $\rho \in \mathscr{C}([0,T], \mathscr{P}_2(D))$ and, for any test function $h \in \mathscr{C}^{2,1}(D \times [0,T])$, satisfying $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h(\boldsymbol{x}, t)\rangle = 0$ for all $\boldsymbol{x} \in \partial D, t \in [0,T]$, we have*

$$\begin{aligned} &\int_D h(\boldsymbol{x}, T)\, \rho_T(\mathrm{d}\boldsymbol{x}) - \int_D h(\boldsymbol{x}, 0)\, \rho_{\text{init}}(\mathrm{d}\boldsymbol{x}) \\ &= \int_0^T \int_D \left[\partial_t h(\boldsymbol{x}, t) + \tau \Delta h(\boldsymbol{x}, t) - \langle \nabla \Psi(\boldsymbol{x}, \rho_t), \nabla h(\boldsymbol{x}, t)\rangle\right] \rho_t(\mathrm{d}\boldsymbol{x})\, \mathrm{d}t\,. \end{aligned} \tag{B.5}$$

We now state and prove Duhamel's principle for the PDE (B.2). Duhamel's principle follows from the fact that the right-hand side of (B.2) contains the linear diffusion term $\tau\Delta\rho$, and it will be crucial for the proofs that will follow.

Lemma B.2 (Duhamel's principle)**.** *Assume $\tau > 0$. Let $G^D(\boldsymbol{x}, \boldsymbol{y}; t)$ denote the heat kernel with Neumann boundary conditions, defined in (G.1)-(G.3). Let ρ be a weak solution of the PDE (B.2) with initial and boundary conditions (B.3). Then, for any $t > 0$, $\rho_t(\mathrm{d}\boldsymbol{x})$ has a density, denoted by $\rho(\,\cdot\,, t)$, which satisfies, for any $t > 0$,*

$$\begin{aligned} \rho(\boldsymbol{x}, t) &= \int_D G^D(\boldsymbol{x}, \boldsymbol{y}; \tau t)\, \rho_{\text{init}}(\mathrm{d}\boldsymbol{y}) \\ &\quad - \int_0^t \int_D \langle \nabla_{\boldsymbol{y}} G^D(\boldsymbol{x}, \boldsymbol{y}; \tau(t-s)), \nabla_{\boldsymbol{y}} \Psi(\boldsymbol{y}; \rho_s)\rangle\, \rho(\boldsymbol{y}, s)\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}s\,. \end{aligned} \tag{B.6}$$

Proof. By rescaling time, without loss of generality, we set $\tau = 1$. Let $\varphi \in \mathscr{C}^2(D)$, and define

$$G_\varphi(\boldsymbol{x}; t) = \int_D G^D(\boldsymbol{x}, \boldsymbol{y}; t)\, \varphi(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}\,. \tag{B.7}$$

By the properties of the heat kernel, we have:

$$\Big(\partial_t - \Delta\Big) G_\varphi(\boldsymbol{x}; t) = 0 \quad \forall t > 0\,, \tag{B.8}$$

$$\langle \boldsymbol{n}(\boldsymbol{x}), \nabla G_\varphi(\boldsymbol{x}; t)\rangle = 0 \quad \forall \boldsymbol{x} \in \partial D\,, \tag{B.9}$$

$$\lim_{t\to 0} G_\varphi(\boldsymbol{x}; t) = G_\varphi(\boldsymbol{x}; 0) = \varphi(\boldsymbol{x})\,. \tag{B.10}$$

Let ρ_t be a weak solution. We choose the test function $h(\boldsymbol{x}, s) = G_\varphi(\boldsymbol{x}; t - s)$ in (B.5) with $T = t$. Note that by (B.8), this test function satisfies the Neumann boundary condition. In addition, by (B.9) we obtain

$$\begin{aligned}\int_D \varphi(\boldsymbol{y})\, \rho_t(\mathrm{d}\boldsymbol{y}) = \int_D G_\varphi(\boldsymbol{x}; t)\, \rho_{\mathrm{init}}(\mathrm{d}\boldsymbol{x}) \\ - \int_0^t \int_D \langle \nabla\Psi(\boldsymbol{x}, \rho_s), \nabla G_\varphi(\boldsymbol{x}; t - s)\rangle\, \rho_s(\mathrm{d}\boldsymbol{x})\, \mathrm{d}s\,.\end{aligned} \tag{B.11}$$

By an application of Fubini's theorem, this implies

$$\begin{aligned}\int_D \varphi(\boldsymbol{y})\, \rho_t(\mathrm{d}\boldsymbol{y}) = \int_D \int_D G^D(\boldsymbol{x}, \boldsymbol{y}; t)\, \rho_{\mathrm{init}}(\mathrm{d}\boldsymbol{x})\, \varphi(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \\ - \int_0^t \int_D \int_D \langle \nabla\Psi(\boldsymbol{x}, \rho_s), \nabla G^D(\boldsymbol{x}, \boldsymbol{y}; t - s)\rangle\, \varphi(\boldsymbol{y})\, \rho_s(\mathrm{d}\boldsymbol{x})\mathrm{d}s\, \mathrm{d}\boldsymbol{y}\,.\end{aligned} \tag{B.12}$$

Since $\varphi \in \mathscr{C}^2(D)$ is arbitrary, we obtain that ρ_t admits a density and (B.6) follows. □

As an intermediate step towards proving existence and uniqueness, we consider a linearized problem

$$\partial_t \rho(\boldsymbol{x}, t) = \nabla \cdot (\rho(\boldsymbol{x}, t) \nabla \Psi_*(\boldsymbol{x}, t)) + \tau\, \Delta \rho(\boldsymbol{x}, t)\,, \tag{B.13}$$

with initial and boundary conditions

$$\begin{aligned}\rho_0 = \rho_{\mathrm{init}}\,, \\ \langle \boldsymbol{n}(\boldsymbol{x}), \rho_t(\boldsymbol{x}) \nabla \Psi_*(\boldsymbol{x}, t) + \tau\, \nabla \rho_t(\boldsymbol{x})\rangle = 0\,, \quad \forall \boldsymbol{x} \in \partial D\,.\end{aligned} \tag{B.14}$$

Here, $\Psi_* : D \times \mathbb{R} \to \mathbb{R}$ is independent of ρ, and weak solutions are defined as for the original problem (with Neumann boundary conditions).

Corollary B.3 (Uniqueness of linearized problem)**.** *Assume that $\tau > 0$ and also that*

$$\|\nabla \Psi_*\|_{\mathscr{L}^\infty(D\times[0,T])} \equiv \sup_{t\in[0,T]} \sup_{\boldsymbol{x}\in D} |\nabla \Psi_*(\boldsymbol{x}, t)| < \infty.$$

Then, the PDE (B.13) with initial and boundary conditions (B.14) has at most one weak solution.

Proof. Without loss of generality, we will set $\tau = 1$. Assume by contradiction that $\rho^{(1)}$, $\rho^{(2)}$ are two solutions. Fix arbitrary $0 \le t' \le t$. Then, by an application of (B.6) to $\Psi_*(\boldsymbol{x}, t)$, we have

$$\begin{aligned}
&\big|\rho^{(1)}(\boldsymbol{x}, t') - \rho^{(2)}(\boldsymbol{x}, t')\big| \le \big\|\rho^{(1)}(\,\cdot\,, 0) - \rho^{(2)}(\,\cdot\,, 0)\big\|_{\mathscr{L}^\infty(D)} \\
&+ \|\nabla\Psi_*\|_{\mathscr{L}^\infty(D\times[0,T])} \int_0^{t'}\!\!\int_D |\nabla_{\boldsymbol{y}} G^D(\boldsymbol{x}, \boldsymbol{y}; t'-s)| \, \big|\rho^{(1)}(\boldsymbol{y}, s) - \rho^{(2)}(\boldsymbol{y}, s)\big| \, \mathrm{d}\boldsymbol{y}\mathrm{d}s \\
&\le \big\|\rho^{(1)}(\,\cdot\,, 0) - \rho^{(2)}(\,\cdot\,, 0)\big\|_{\mathscr{L}^\infty(D)} \\
&\quad + C(D)\|\nabla\Psi_*\|_{\mathscr{L}^\infty(D\times[0,T])} \int_0^{t'} \frac{1}{\sqrt{t'-s}} \big\|\rho^{(1)}(\,\cdot\,, s) - \rho^{(2)}(\,\cdot\,, s)\big\|_{\mathscr{L}^\infty(D)} \mathrm{d}s \\
&\le \big\|\rho^{(1)}(\,\cdot\,, 0) - \rho^{(2)}(\,\cdot\,, 0)\big\|_{\mathscr{L}^\infty(D)} \\
&\quad + C(D)\|\nabla\Psi_*\|_{\mathscr{L}^\infty(D\times[0,T])}\sqrt{t} \, \sup_{s\le t} \big\|\rho^{(1)}(\,\cdot\,, s) - \rho^{(2)}(\,\cdot\,, s)\big\|_{\mathscr{L}^\infty(D)} ,
\end{aligned}$$

where we used the estimates of Theorem G.1. By taking supremum over $0 \le t' \le t$ form both sides, we obtain that for $t < 1/(C(D)^2\|\nabla\Psi_*\|^2_{\mathscr{L}^\infty(D\times[0,T])})$,

$$\sup_{s\le t} \big\|\rho^{(1)}(\,\cdot\,, s) - \rho^{(2)}(\,\cdot\,, s)\big\|_{\mathscr{L}^\infty(D)} \le \frac{\big\|\rho^{(1)}(\,\cdot\,, 0) - \rho^{(2)}(\,\cdot\,, 0)\big\|_{\mathscr{L}^\infty(D)}}{1 - C(D)\|\nabla\Psi_*\|_{\mathscr{L}^\infty(D\times[0,T])}\sqrt{t}} . \tag{B.15}$$

Therefore, the two solutions coincide if we fix the initial condition $\rho^{(1)}(\,\cdot\,, 0) = \rho^{(2)}(\,\cdot\,, 0) = \rho_{\mathrm{init}}$. For larger t, the claim follows by iterating the above argument. □

C Nonlinear dynamics

The 'nonlinear dynamics' plays an important role in our proof of Theorem 5.1. In this section we adopt the same general setting as in Appendix B, remembering that for our application we set $D = \Omega^\delta$ and U, V as per Eq. (B.4).

Given $\rho : [0, T] \to \mathscr{P}_2(D)$, consider the following stochastic differential equation for a process $(\boldsymbol{X}_t)_{t\in[0,T]}$, with a reflecting boundary condition (known as 'Skorokhod problem')

$$\boldsymbol{X}_0 \sim \rho_{\mathrm{init}} \tag{C.1}$$

$$\mathrm{d}\boldsymbol{X}_t = -\nabla\Psi(\boldsymbol{X}_t, \rho_t)\,\mathrm{d}t + \sqrt{2\tau}\,\mathrm{d}\boldsymbol{B}_t + \mathrm{d}\boldsymbol{\Phi}_t , \tag{C.2}$$

where $(\boldsymbol{B}_t)_{t\ge0}$ is a standard d-dimensional Brownian motion and $(\boldsymbol{\Phi}_t)_{t\ge0}$ enforces the reflecting boundary by satisfying the following constraints (recall that $\boldsymbol{n}(\boldsymbol{x})$ is the normal to ∂D at $\boldsymbol{x} \in \partial D$, directed inside):

(*i*) $(\boldsymbol{\Phi}_t)_{t\ge0}$ is adapted (and hence so is $(\boldsymbol{X}_t)_{t\ge0}$).

(*ii*) $t \mapsto \boldsymbol{\Phi}_t$ has (almost surely) bounded variation. Denoting by $\|\boldsymbol{\Phi}\|_{\mathrm{TV}}(t)$ the total variation of $\boldsymbol{\Phi}$ on the interval $[0, t]$, we define the measure μ_Φ on $[0, T]$ by $\mu_\Phi([0, t]) = \|\boldsymbol{\Phi}\|_{\mathrm{TV}}(t)$.

(*iii*) $\mu_\Phi(\{t : \boldsymbol{X}_t \in D^\circ\}) = 0$, where D° denotes the interior of D.

(*iv*) We have that, for $t \in [0,T]$,

$$\boldsymbol{\Phi}_t = \int_0^t \boldsymbol{N}_s \, \mu_\Phi(\mathrm{d}s)\,, \tag{C.3}$$

where $\boldsymbol{N}_s = \boldsymbol{n}(\boldsymbol{X}_s)$, for μ_Φ-almost every s.

Then, $(\boldsymbol{X}_t, \boldsymbol{\Phi}_t)_{t\in[0,T]}$ is said to solve the Skorokhod problem.

Lemma C.1 (Existence, uniqueness and continuity of Skorokhod problem)**.** *Fix $\rho_{\mathrm{init}} \in \mathscr{P}_2(D)$ and let $\rho : [0,T] \to \mathscr{P}_2(D)$ with $\rho_0 = \rho_{\mathrm{init}}$. Then, the Skorokhod problem (C.1), (C.2) admits a unique solution $(\boldsymbol{X}_t)_{t\geq 0}$ with continuous paths. Define $\mathscr{F}(\rho)_t \in \mathscr{P}_2(D)$, for $t \in [0,T]$, by letting $\mathscr{F}(\rho)_t = \mathrm{Law}(\boldsymbol{X}_t)$. Then, $\mathscr{F}(\rho) \in \mathscr{C}([0,T], \mathscr{P}_2(D))$.*

Proof. Let $\boldsymbol{b}(\boldsymbol{x},t) \equiv -\nabla\Psi(\boldsymbol{x},\rho_t)$ and notice that, by the smoothness of U, V, and compactness of D, this is a Lipschitz continuous function of $\boldsymbol{x}$. Hence the problem (C.1), (C.2) admits a unique solution by [Tan79, Theorem 4.1].

We are left with the task of proving that $t \mapsto \mathscr{F}(\rho)_t$ is continuous in W_2 metric. Notice that

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{b}(\boldsymbol{X}_s, s)\,\mathrm{d}s + \sqrt{2\tau}\,\boldsymbol{B}_t + \boldsymbol{\Phi}_t\,. \tag{C.4}$$

By [Tan79, Lemma 2.2], we have, for any $s \leq t$,

$$\begin{aligned}|\boldsymbol{X}_t - \boldsymbol{X}_s|^2 \leq \Big|\int_s^t \boldsymbol{b}(\boldsymbol{X}_r, r)\,\mathrm{d}r + \sqrt{2\tau}\,(\boldsymbol{B}_t - \boldsymbol{B}_s)\Big|^2 + \\ 2\int_s^t \langle \int_r^t \boldsymbol{b}(\boldsymbol{X}_u, u)\,\mathrm{d}u + \sqrt{2\tau}\,(\boldsymbol{B}_t - \boldsymbol{B}_r), \boldsymbol{N}_r\rangle \mu_\Phi(\mathrm{d}r)\,.\end{aligned}$$

Taking expectation, we get

$$\begin{aligned}\mathbb{E}\{|\boldsymbol{X}_t - \boldsymbol{X}_s|^2\} &\leq \sup_{\boldsymbol{x},t} |\boldsymbol{b}(\boldsymbol{x},t)|^2 (t-s)^2 + 2\tau(t-s) \\ &\quad + 2\sup_{\boldsymbol{x},t} |\boldsymbol{b}(\boldsymbol{x},t)| \int_s^t (t-r)\mu_\Phi(\mathrm{d}r) \\ &\leq \sup_{\boldsymbol{x},t} |\boldsymbol{b}(\boldsymbol{x},t)|^2 (t-s)^2 + 2\tau(t-s) \\ &\quad + 2\sup_{\boldsymbol{x},t} |\boldsymbol{b}(\boldsymbol{x},t)| (t-s)\mu_\Phi([0,t])\,,\end{aligned} \tag{C.5}$$

whence the continuity follows. □

Definition C.2 (Solution of nonlinear dynamics)**.** *We say that $\rho \in \mathscr{C}([0,T]; \mathscr{P}_2(D))$ is a solution of the nonlinear dynamics if $\mathscr{F}(\rho) = \rho$, namely*

$$\rho_t = \mathrm{Law}(\boldsymbol{X}_t) \quad \forall t \in [0,T]\,. \tag{C.6}$$

Lemma C.3. *Assume $\tau > 0$. If $\rho : [0,T] \to \mathscr{P}_2(D)$ is a weak solution of the PDE (B.2) with initial and boundary conditions (B.3), then it is a solution of the nonlinear dynamics. Vice versa, if $\rho : [0,T] \to \mathscr{P}_2(D)$ is a solution of the nonlinear dynamics, then it is a weak solution of PDE (B.2) with initial and boundary conditions (B.3).*

Proof. Let ρ be a weak solution of the PDE (B.2), and assume $\tau > 0$. Let $(\boldsymbol{X}_t)_{t\geq 0}$ be the unique solution of the Skorokhod problem (C.1), (C.2), cf. Lemma C.1. Let $\tilde{\rho}_t \equiv \mathrm{Law}(\boldsymbol{X}_t)$, $t \geq 0$, i.e. $\tilde{\rho} \equiv \mathscr{F}(\rho)$. For $g \in \mathscr{C}^2(D)$, satisfying $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla g(\boldsymbol{x})\rangle = 0$ for all $\boldsymbol{x} \in \partial D$, compute

$$\begin{aligned}
\int g(\boldsymbol{x})\, \tilde{\rho}_t(\mathrm{d}\boldsymbol{x}) &= \mathbb{E}\{g(\boldsymbol{X}_t)\} \\
&\overset{(a)}{=} \mathbb{E}\{g(\boldsymbol{X}_0)\} + \int_0^t \mathbb{E}\left\{-\langle \nabla\Psi(\boldsymbol{X}_s, \rho_s), \nabla g(\boldsymbol{X}_s)\rangle + \tau \Delta g(\boldsymbol{X}_s)\right\} \mathrm{d}s \\
&\qquad + \mathbb{E}\int_0^t \langle \nabla g(\boldsymbol{X}_s), \boldsymbol{N}_s\rangle\, \mu_\Phi(\mathrm{d}s) \\
&\overset{(b)}{=} \mathbb{E}\{g(\boldsymbol{X}_0)\} + \int_0^t \mathbb{E}\left\{-\langle \nabla\Psi(\boldsymbol{X}_s, \rho_s), \nabla g(\boldsymbol{X}_s)\rangle + \tau \Delta g(\boldsymbol{X}_s)\right\} \mathrm{d}s \\
&\overset{(c)}{=} \int_D g(\boldsymbol{x})\, \rho_{\mathrm{init}}(\mathrm{d}\boldsymbol{x}) \\
&\qquad + \int_0^t \int_D \left\{-\langle \nabla\Psi(\boldsymbol{x}, \rho_s), \nabla g(\boldsymbol{x})\rangle + \tau \Delta g(\boldsymbol{x})\right\}\, \tilde{\rho}_s(\mathrm{d}\boldsymbol{x})\, \mathrm{d}s\,.
\end{aligned} \tag{C.7}$$

Here (a) follows from Ito's formula for continuous semimartingales [RW94], (b) since $\boldsymbol{X}_s \in \partial D$ and $\boldsymbol{N}_s = \boldsymbol{n}(\boldsymbol{X}_s)$ for μ_Φ-almost every s, and (c) by the definition of $\tilde{\rho}$. We conclude that $\tilde{\rho}$ is a weak solution of the linearized PDE (B.13), with $\Psi_*(\boldsymbol{x}, t) = \Psi(\boldsymbol{x}, \rho_t)$. Since ρ also solves the same linearized PDE, we conclude by Lemma B.3 that $\tilde{\rho}_t = \rho_t$ for all $t \in [0, T]$, and therefore ρ is a solution of the nonlinear dynamics.

Next, assume that $\rho : [0, T] \to \mathscr{P}_2(D)$ is a solution of the nonlinear dynamics. Then by the same application of Ito's formula to the process $\boldsymbol{X}_t$, we have

$$\begin{aligned}
\int g(\boldsymbol{x})\, \rho_t(\mathrm{d}\boldsymbol{x}) &= \mathbb{E}\{g(\boldsymbol{X}_0)\} \\
&\qquad + \int_0^t \mathbb{E}\left\{-\langle \nabla\Psi(\boldsymbol{X}_s, \rho_s), \nabla g(\boldsymbol{X}_s)\rangle + \tau \Delta g(\boldsymbol{X}_s)\right\} \mathrm{d}s\,,
\end{aligned} \tag{C.8}$$

which coincides with the claim that ρ is a weak solution of the PDE (B.2). □

Theorem C.4 (Existence and uniqueness of nonlinear dynamics)**.** *For any initial condition $\rho_{\mathrm{init}} \in \mathscr{P}_2(D)$, and any $T > 0$, the nonlinear dynamics (C.6) admits a unique solution $\rho : [0, T] \to \mathscr{P}_2(D)$ with $\rho_0 = \rho_{\mathrm{init}}$. As a consequence, the PDE (B.2) with initial and boundary conditions (B.3) has a unique solution.*

Proof. Note that it is sufficient to prove the claim for $T \leq T_0$, where $T_0 > 0$ is a small enough constant, since this implies the claim for arbitrary T by breaking $[0, T]$ into intervals of size smaller than T_0.

We claim that $\mathscr{F}$ is a contraction on $\mathscr{C}([0, T], \mathcal{P}_2(D))$ endowed with the metric $d(\rho, \tilde{\rho}) \equiv \sup_{t\in[0,T]} W_2(\rho, \tilde{\rho})$. To show that this is the case, define $\boldsymbol{b}(\boldsymbol{x}, t) \equiv -\nabla\Psi(\boldsymbol{x}, \rho_t)$, $\tilde{\boldsymbol{b}}(\boldsymbol{x}, t) \equiv -\nabla\Psi(\boldsymbol{x}, \tilde{\rho}_t)$. By the smoothness of U, V and by the compactness of D, we have that $\boldsymbol{b}$ and $\tilde{\boldsymbol{b}}$ are Lipschitz continuous in $\boldsymbol{x}$, with Lipschitz constant L independent of $t, \rho, \tilde{\rho}$. Further,

$$\begin{aligned}
|\boldsymbol{b}(\boldsymbol{x}, t) - \tilde{\boldsymbol{b}}(\boldsymbol{x}, t)| &= \left| \int \nabla U(\boldsymbol{x}, \tilde{\boldsymbol{x}})\, \rho_t(\mathrm{d}\tilde{\boldsymbol{x}}) - \int \nabla U(\boldsymbol{x}, \tilde{\boldsymbol{x}})\, \tilde{\rho}_t(\mathrm{d}\tilde{\boldsymbol{x}}) \right| \\
&\leq C\, W_1(\rho_t, \tilde{\rho}_t) \leq C\, d(\rho, \tilde{\rho})\,.
\end{aligned} \tag{C.9}$$

Let $(\boldsymbol{X}_t, \boldsymbol{\Phi}_t)$ and $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{\Phi}}_t)$ are be solution of the Skorokhod problem (C.2), with drift coefficients $\boldsymbol{b}(\boldsymbol{x}, t)$, $\tilde{\boldsymbol{b}}(\boldsymbol{x}, t)$. We couple the processes $\boldsymbol{X}_t$ and $\tilde{\boldsymbol{X}}_t$ by using the same initial condition $\boldsymbol{X}_0$ and same Brownian motion $\boldsymbol{B}_t$:

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{b}(\boldsymbol{X}_s, s)\, \mathrm{d}s + \sqrt{2\tau}\, \boldsymbol{B}_t + \boldsymbol{\Phi}_t\,, \tag{C.10}$$

$$\tilde{\boldsymbol{X}}_t = \boldsymbol{X}_0 + \int_0^t \tilde{\boldsymbol{b}}(\tilde{\boldsymbol{X}}_s, s)\, \mathrm{d}s + \sqrt{2\tau}\, \boldsymbol{B}_t + \tilde{\boldsymbol{\Phi}}_t\,. \tag{C.11}$$

Define

$$\boldsymbol{D}_t \equiv \int_0^t \boldsymbol{b}(\boldsymbol{X}_s, s)\, \mathrm{d}s - \int_0^t \tilde{\boldsymbol{b}}(\tilde{\boldsymbol{X}}_s, s)\, \mathrm{d}s\,, \tag{C.12}$$

and notice that, by the above remarks,

$$|\boldsymbol{D}_t| \le L \int_0^t |\boldsymbol{X}_s - \tilde{\boldsymbol{X}}_s|\, \mathrm{d}s + C\, t\, d(\rho, \tilde{\rho})\,. \tag{C.13}$$

Further, by [Tan79, Remark 2.2], we have

$$\begin{aligned} |\boldsymbol{X}_t - \tilde{\boldsymbol{X}}_t|^2 &\le 2 \int_0^t \langle \boldsymbol{X}_s - \tilde{\boldsymbol{X}}_s, \boldsymbol{D}_s \rangle\, \mathrm{d}s \\ &\le 2 \int_0^t |\boldsymbol{X}_s - \tilde{\boldsymbol{X}}_s| \left(L \left(\int_0^s |\boldsymbol{X}_r - \tilde{\boldsymbol{X}}_r|\, \mathrm{d}r \right) + C\, s\, d(\rho, \tilde{\rho}) \right) \mathrm{d}s \\ &\le 2L \left(\int_0^t |\boldsymbol{X}_s - \tilde{\boldsymbol{X}}_s| \mathrm{d}s \right)^2 + 2C\, t\, d(\rho, \tilde{\rho}) \int_0^t |\boldsymbol{X}_s - \tilde{\boldsymbol{X}}_s| \mathrm{d}s \\ &\le 2Lt \int_0^t |\boldsymbol{X}_s - \tilde{\boldsymbol{X}}_s|^2 \mathrm{d}s + 2C\, t^{3/2}\, d(\rho, \tilde{\rho}) \left(\int_0^t |\boldsymbol{X}_s - \tilde{\boldsymbol{X}}_s|^2 \mathrm{d}s \right)^{1/2}. \end{aligned} \tag{C.14}$$

Define $\Delta(t) \equiv \mathbb{E}\{|\boldsymbol{X}_t - \tilde{\boldsymbol{X}}_t|^2\}$ and $\overline{\Delta}(t) \equiv \sup_{s \le t} \Delta(s)$. By taking the expectation of the last inequality and using Jensen's inequality, we get

$$\Delta(t) \le 2Lt \int_0^t \Delta(s) \mathrm{d}s + 2C\, t^{3/2}\, d(\rho, \tilde{\rho}) \left(\int_0^t \Delta(s) \mathrm{d}s \right)^{1/2}, \tag{C.15}$$

which immediately implies

$$\overline{\Delta}(t) \le 2Lt^2\, \overline{\Delta}(t) + 2C\, t^2\, d(\rho, \tilde{\rho})\, \overline{\Delta}(t)^{1/2}\,. \tag{C.16}$$

Hence, for $T_0 < (2L)^{-1/2}$,

$$\overline{\Delta}(t) \le \left(\frac{2CT_0^2}{1 - 2LT_0^2} \right)^2 d(\rho, \tilde{\rho})^2\,. \tag{C.17}$$

Selecting T_0 small enough, so that $(2CT_0^2)/(1 - 2LT_0^2) \le 1/2$, we obtain

$$d(\mathscr{F}(\rho), \mathscr{F}(\tilde{\rho})) \le \sqrt{\overline{\Delta}(T_0)} \le \frac{1}{2}\, d(\rho, \tilde{\rho})\,. \tag{C.18}$$

This proves that $\mathscr{F}$ is a contraction as claimed. By Lemma C.1, $\mathscr{F}$ maps $\mathscr{C}([0, T], \mathcal{P}_2(D))$ into itself. Furthermore, $\mathscr{C}([0, T], \mathcal{P}_2(D))$ is complete with respect to the metric d. As a result, there exists a unique fixed point. □

We conclude this section by stating a result about the discretization of the nonlinear dynamics. Fix a solution $(\rho_t)_{t\ge 0}$ of the PDE (B.2) with initial condition $\rho_0 = \rho_{\text{init}}$, a step size $\varepsilon > 0$ and define recursively the random variables $(\boldsymbol{X}^\varepsilon)_{k\in\mathbb{N}}$ by

$$\boldsymbol{X}_0^\varepsilon \sim \rho_{\text{init}} \tag{C.19}$$

$$\boldsymbol{X}_{k+1}^\varepsilon = \mathsf{P}\Big(\boldsymbol{X}_k^\varepsilon - \varepsilon\nabla\Psi(\boldsymbol{X}_k^\varepsilon, \rho_{k\varepsilon}) + \sqrt{2\tau\varepsilon}\,\boldsymbol{g}_k\Big)\,. \tag{C.20}$$

This can be viewed as an Euler discretization of the stochastic differential equation (C.1), (C.2), and the next theorem establishes that this is indeed a close approximation of the original process. It is just an immediate consequence of a result of Slomiński [Slo94, Slo01].

Theorem C.5 (Theorem 3.2 in [Slo01]). *Consider the nonlinear dynamics defined by Eqs. (C.1), (C.2). Assume* $\mathsf{B}(\mathbf{0}; r) \subseteq D$, *and* $\|\nabla V\|_{\mathscr{L}^\infty(D)}$, $\|\nabla U\|_{\mathscr{L}^\infty(D\times D)}$, $\|\nabla V\|_{\text{Lip}}$, $\|\nabla U\|_{\text{Lip}} \le L$. *Also assume that* $\text{supp}(\rho_{\text{init}}) \subseteq \mathsf{B}(\mathbf{0}, r)$. *Construct the Euler scheme (C.19), (C.20) on the same probability space by letting* $\boldsymbol{X}_0^\varepsilon = \boldsymbol{X}_0$ *and* $\boldsymbol{g}_k = (\boldsymbol{B}((k+1)\varepsilon) - \boldsymbol{B}(k\varepsilon))/\sqrt{\varepsilon}$. *Then, for any* $p \in \mathbb{N}$, $T \in \mathbb{R}_{\ge 0}$,

$$\mathbb{E}\Big\{\max_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \big|\boldsymbol{X}_k^\varepsilon - \boldsymbol{X}_{k\varepsilon}\big|^{2p}\Big\}^{1/(2p)} \le C_* L r\, e^{C_* pLT}(d^2\varepsilon\log(1/\varepsilon))^{1/4}\,. \tag{C.21}$$

Proof. The proof is obtained simply by chasing the constants in the proof of Theorem 3.2 (part (ii)) of [Slo01], and using the optimal constant in the Burkholder-Davis-Gundy inequality (which yields $C(p) \le (C_* p)^{2p}$ in [Slo01, Eq. (2.7)]). □

D Convergence of SGD to the PDE: Proof of Theorem 5.1

The proof is a 'propagation of chaos' argument [Szn91]. While the basic idea is similar to the one used in [MMN18], implementing it requires different estimates because of the reflecting boundary conditions. In particular, we rely on tools developed in the study of discretizations of reflecting stochastic differential equations.

We will prove a more general theorem that implies Theorem 5.1 as a special case, and also applies to the setting of [MMN18]. Namely, we consider data $\{\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)\}_{i\ge 1}$ i.i.d. with common distribution $\mathbb{P}$ on $\mathbb{R}\times\mathbb{R}^{d_0}$, and parameters $\boldsymbol{w}_i \in D \subseteq \mathbb{R}^d$. These parameters are initially sampled independently from distribution $\rho_0 \in \mathscr{P}_2(D)$, and then evolve according to

$$\boldsymbol{w}_i^{k+1} = \mathsf{P}\Big\{\boldsymbol{w}_i^k + \boldsymbol{F}_i(\boldsymbol{z}_{k+1}; \boldsymbol{w}^k)\Big\}\,, \tag{D.1}$$

$$\boldsymbol{F}_i(\boldsymbol{z}_{k+1}; \boldsymbol{w}^k) = -\varepsilon\nabla\sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k)\Big(y_{k+1} - \frac{1}{N}\sum_{i=1}^N \sigma(\boldsymbol{x}_{k+1}; \boldsymbol{w}_i^k)\Big) + \sqrt{2\tau\varepsilon}\;\boldsymbol{g}_i^{k+1}\,. \tag{D.2}$$

Here P is the projection on the closed convex domain $D \subseteq \mathbb{R}^d$ with non-empty interior. The setting of Theorem 5.1 is recovered by taking $\sigma(\boldsymbol{x}; \boldsymbol{w}) = K_\delta(\boldsymbol{x} - \boldsymbol{w})$, $D = \Omega^\delta$, $\boldsymbol{x}_k \sim \mathsf{Unif}(\Omega)$, $\mathbb{E}\{y_k|\boldsymbol{x}_k\} = f(\boldsymbol{x}_k)$.

We make the following assumptions:

(G1) $\|y\|_\infty$, $\|\sigma\|_\infty = \text{ess sup}_{\boldsymbol{w}\in D, \boldsymbol{x}} |\sigma(\boldsymbol{x}; \boldsymbol{w})| \le \sigma_\infty$, and $\nabla_{\boldsymbol{w}}\sigma(\boldsymbol{x}; \boldsymbol{w})$ is γ-subgaussian.

(G2) Letting $V(\boldsymbol{w}) = -\mathbb{E}\{y\sigma(\boldsymbol{x};\boldsymbol{w})\}$, $U(\boldsymbol{w}_1,\boldsymbol{w}_2) \equiv \mathbb{E}\{\sigma(\boldsymbol{x};\boldsymbol{w}_1)\sigma(\boldsymbol{x};\boldsymbol{w}_2)\}$, both V and U are differentiable with Lipschitz continuous derivative, namely $\|\nabla V\|_{\rm Lip}, \|\nabla U\|_{\rm Lip} \le L$. Further, we assume $\|\nabla U\|_{\mathscr{L}^\infty(D\times D)} < \infty$.

Theorem D.1. *Consider the general update (D.1) with initialization $(\boldsymbol{w}_i^0)_{i\le N} \sim_{iid} \rho_0 = \rho_{\rm init}$, under the conditions (G1), (G2) above. For $t \ge 0$, let ρ_t be the unique solution of the PDE (B.2) with initial and boundary conditions (B.3). Assume $\mathrm{supp}(\rho_{\rm init}) \subseteq \mathsf{B}(\mathbf{0}, r)$.*

Then, for $T \ge 0$ $TL \ge 1$, any $g : \mathbb{R}^d \to \mathbb{R}$ with $\|g\|_{\rm Lip} \le 1$ and for $\varepsilon \le 1$, $p \in \mathbb{N}$, the following holds with probability at least $1 - z^{-2p}$:

$$\begin{aligned}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \left|\sum_{i=1}^N g(\boldsymbol{w}_i^k) - \int g(\boldsymbol{w})\rho_{k\varepsilon}(\mathrm{d}\boldsymbol{w})\right| &\le z\,\mathsf{err}(N,d,\varepsilon)\; e^{C_* pLT},\\ \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_N(\boldsymbol{w}^k) - R^\delta(\rho_{k\epsilon})| &\le z\,\mathsf{err}(N,d,\varepsilon)\;\; e^{C_* pLT},\end{aligned} \tag{D.3}$$

where

$$\mathsf{err}(N,d,\varepsilon) = \sqrt{\frac{d}{N}} \vee \left(\sigma_\infty\gamma\sqrt{d\varepsilon}\right) \vee \left(Lr(d^2\varepsilon\log(1/\varepsilon))^{1/4}\right). \tag{D.4}$$

Theorem 5.1 follows as a special case of Theorem D.1 by considering $\sigma(\boldsymbol{x};\boldsymbol{w}) = K_\delta(\boldsymbol{x}-\boldsymbol{w})$ and letting $\sigma_\infty \le C_*\delta^{-d}$, $\gamma = C_*\delta^{-d-1}$ and $L = C_*\delta^{-2d-1}$.

Proof. Let $\mathcal{F}_k$ denote the sigma algebra generated by $(\boldsymbol{z}_j)_{j\le k}$ and denote the empirical distribution of $(\boldsymbol{w}_i^k)_{i\le N}$ by $\rho_k^{(N)} \equiv \sum_{i=1}^n \delta_{\boldsymbol{w}_i^k}$. Note that

$$\begin{aligned}\mathbb{E}\big\{\boldsymbol{F}_i(\boldsymbol{z}_{k+1};\boldsymbol{w}^k)\big|\mathcal{F}_k\big\} &= \varepsilon\boldsymbol{G}(\boldsymbol{w}_i^k;\rho_k^{(N)}),\\ \boldsymbol{G}(\boldsymbol{w};\rho) \equiv -\nabla\Psi(\boldsymbol{w};\rho) &= -\nabla V(\boldsymbol{w}) - \int \nabla U(\boldsymbol{w},\boldsymbol{w}')\,\rho(\mathrm{d}\boldsymbol{w}').\end{aligned} \tag{D.5}$$

We introduce two auxiliary processes $(\overline{\boldsymbol{w}}_i^k)_{i\le N}$ $(\hat{\boldsymbol{w}}_i^k)_{i\le N}$, with initial conditions $\overline{\boldsymbol{w}}_i^0 = \hat{\boldsymbol{w}}_i^0 = \boldsymbol{w}_i^0$, as follows:

- The trajectories $(\hat{\boldsymbol{w}}_i^k)_{k\ge0}$ are i.i.d. copies of the nonlinear dynamics introduced in Appendix C, sampled at times $t = k\varepsilon$. Namely, for any $k \in \mathbb{R}$

$$\hat{\boldsymbol{w}}_i^k = \boldsymbol{w}_i^0 + \int_0^{k\varepsilon} \boldsymbol{G}(\boldsymbol{w}_i^{s/\varepsilon};\rho_s)\,\mathrm{d}s + \sqrt{2\tau}\,\boldsymbol{B}_i(k\varepsilon) + \boldsymbol{\Phi}_i(k\varepsilon). \tag{D.6}$$

  In particular, for any k, $(\hat{\boldsymbol{w}}_i^k)_{i\le N} \sim_{iid} \rho_{k\varepsilon}$.

- The trajectories $(\overline{\boldsymbol{w}}_i^k)_{k\ge0}$ are obtained by the Euler discretization of the non-linear dynamics:

$$\overline{\boldsymbol{w}}_i^{k+1} = \mathsf{P}\Big(\overline{\boldsymbol{w}}_i^k + \varepsilon\boldsymbol{G}(\overline{\boldsymbol{w}}_i^k;\rho_{k\varepsilon}) + \sqrt{2\tau\varepsilon}\boldsymbol{g}_i^{k+1}\Big). \tag{D.7}$$

  As above, $(\rho_s)_{s\ge0}$ is the solution of the PDE (B.2). Note that, again, the $(\overline{\boldsymbol{w}}_i^k)_{i\le N}$ are i.i.d. although their distribution does not coincide with $\rho_{k\varepsilon}$.

We construct these three processes on the same space by letting $\boldsymbol{B}_i((k+1)\varepsilon) = \boldsymbol{B}_i(k\varepsilon) + \sqrt{\varepsilon}\boldsymbol{g}_i^{k+1}$, and define the distances (for $q \geq 1$)

$$\mathcal{D}_q(k) \equiv \mathbb{E}\Big\{ \max_{j\leq k} \big|\boldsymbol{w}_i^j - \overline{\boldsymbol{w}}_i^j\big|^q \Big\}^{1/q} = \mathbb{E}\left\{ \frac{1}{N}\sum_{i=1}^N \max_{j\leq k} \big|\boldsymbol{w}_i^j - \overline{\boldsymbol{w}}_i^j\big|^q \right\}^{1/q} \tag{D.8}$$

$$\geq \mathbb{E}\left\{ \max_{j\leq k} \frac{1}{N}\sum_{i=1}^N \big|\boldsymbol{w}_i^j - \overline{\boldsymbol{w}}_i^j\big|^q \right\}^{1/q},$$

$$\widehat{\mathcal{D}}_q(k) \equiv \mathbb{E}\Big\{ \max_{j\leq k} \big|\hat{\boldsymbol{w}}_i^j - \overline{\boldsymbol{w}}_i^j\big|^q \Big\}^{1/q}. \tag{D.9}$$

Theorem C.5 yields, for $p \in \mathbb{N}$,

$$\widehat{\mathcal{D}}_{2p}(k) \leq C_* L r\, e^{C_* p L(k\varepsilon)} (d^2 \varepsilon \log(1/\varepsilon))^{1/4}. \tag{D.10}$$

Note that $\boldsymbol{w}_i^k$, $\overline{\boldsymbol{w}}_i^k$ take the form

$$\boldsymbol{w}_i^k = \boldsymbol{w}_i^0 + \boldsymbol{M}_i^k + \boldsymbol{V}_i^k + \boldsymbol{\varphi}_i^k, \tag{D.11}$$

$$\overline{\boldsymbol{w}}_i^k = \boldsymbol{w}_i^0 + \overline{\boldsymbol{M}}_i^k + \overline{\boldsymbol{V}}_i^k + \overline{\boldsymbol{\varphi}}_i^k, \tag{D.12}$$

where $\boldsymbol{M}_i^k, \overline{\boldsymbol{M}}_i^k$ are martingales with respect to the filtration $\mathcal{F}_k$: $\mathbb{E}\{\boldsymbol{M}_i^k|\mathcal{F}_{k-1}\} = \boldsymbol{M}_i^{k-1}$, $\mathbb{E}\{\overline{\boldsymbol{M}}_i^k|\mathcal{F}_{k-1}\} = \overline{\boldsymbol{M}}_i^{k-1}$, and $\boldsymbol{V}_i^k$, $\overline{\boldsymbol{V}}_i^k$ are $\mathcal{F}_{k-1}$-measurable. Explicitly

$$\boldsymbol{M}_i^k = \sum_{\ell=0}^{k-1} \big(\boldsymbol{F}_i(\boldsymbol{z}_{\ell+1}; \boldsymbol{w}^\ell) - \mathbb{E}\{\boldsymbol{F}_i(\boldsymbol{z}_{\ell+1}; \boldsymbol{w}^\ell)|\mathcal{F}_\ell\}\big), \tag{D.13}$$

$$\overline{\boldsymbol{M}}_i^k = \sum_{\ell=0}^{k-1} \sqrt{2\tau\varepsilon}\, \boldsymbol{g}_i^{\ell+1}, \tag{D.14}$$

$$\boldsymbol{V}_i^k = \sum_{\ell=0}^{k-1} \mathbb{E}\{\boldsymbol{F}_i(\boldsymbol{z}_{\ell+1}; \boldsymbol{w}^\ell)|\mathcal{F}_\ell\} = \sum_{\ell=0}^{k-1} \varepsilon\, \boldsymbol{G}(\boldsymbol{w}_i^\ell; \rho_\ell^{(N)}), \tag{D.15}$$

$$\overline{\boldsymbol{V}}_i^k = \sum_{\ell=0}^{k-1} \varepsilon\, \boldsymbol{G}(\overline{\boldsymbol{w}}_i^\ell; \rho_{\ell\varepsilon}). \tag{D.16}$$

Finally, $\boldsymbol{\varphi}_i^k$, $\overline{\boldsymbol{\varphi}}_i^k$ are corrections to satisfy the constraint $\boldsymbol{w}_i^k, \overline{\boldsymbol{w}}_i^k \in D$. Indeed the above can be viewed as Skorokhod problems with unknowns $(\boldsymbol{w}_i, \boldsymbol{\varphi}_i)$ and $(\overline{\boldsymbol{w}}_i, \overline{\boldsymbol{\varphi}}_i)$.

Using [Slo94, Theorem 1] (where we can set $C_p = (C_* p)^{2p}$ which is the tight constant in the Burkholder-Davis-Gundy inequality), we get

$$\mathcal{D}_{2p}(k) \leq C_* p \Big\{ \mathbb{E}\big([\boldsymbol{M}_i - \overline{\boldsymbol{M}}_i]_k^p\big)^{1/(2p)} + \mathbb{E}\big(|\boldsymbol{V}_i - \overline{\boldsymbol{V}}_i|_k^{2p}\big)^{1/(2p)} \Big\}, \tag{D.17}$$

where $[\boldsymbol{M}]_k$ denotes the quadratic variation of the martingale $\boldsymbol{M}$, and $|\boldsymbol{V}|_k$ is the total variation

of the process $\boldsymbol{V}$. We then have

$$A_{1,p}(k) \equiv \mathbb{E}\big([\boldsymbol{M}_i - \overline{\boldsymbol{M}}_i]_k^p\big)^{1/(2p)} = \mathbb{E}\left\{\left[\sum_{\ell=0}^{k-1} |\boldsymbol{Z}_i^\ell|^2\right]^p\right\}^{1/(2p)}$$

$$\boldsymbol{Z}_i^\ell = \varepsilon \nabla\sigma(\boldsymbol{x}_{\ell+1}; \boldsymbol{w}_i^k)\Big(y_{\ell+1} - \frac{1}{N}\sum_{i=1}^N \sigma(\boldsymbol{x}_{\ell+1}; \boldsymbol{w}_i^\ell)\Big) + \varepsilon \boldsymbol{G}(\boldsymbol{w}_i^\ell; \rho_\ell^{(N)})\,, \tag{D.18}$$

Note that under the stated assumption the martingale increments $\boldsymbol{Z}_i^\ell$ are sub-Gaussian with variance proxy upper bounded by $v^2 = C_*\varepsilon^2\sigma_\infty^2\gamma^2$. Therefore, by using the moment generating function of χ_d^2 distribution, we have

$$\mathbb{E}\exp\left\{\frac{\alpha^2}{2dv^2}|\boldsymbol{Z}_i^\ell|^2\right\} \le \Big(1 - \frac{\alpha^2}{d}\Big)^{-d/2}\,. \tag{D.19}$$

Hence,

$$\mathbb{E}\exp\left\{\frac{\alpha^2}{2dv^2}\sum_{\ell=0}^{k-1}|\boldsymbol{Z}_i^\ell|^2\right\} \le \Big(1 - \frac{\alpha^2}{d}\Big)^{-dk/2}\,. \tag{D.20}$$

By using the inequality $x^p \le e^x p!$, this implies, for $\alpha \le \sqrt{d/2}$,

$$\mathbb{E}\left\{\left[\frac{\alpha^2}{2dv^2}\sum_{\ell=0}^{k-1}|\boldsymbol{Z}_i^\ell|^2\right]^p\right\} \le p!\Big(1 - \frac{\alpha^2}{d}\Big)^{-dk/2} \le p^p e^{\alpha^2 k}\,. \tag{D.21}$$

Equivalently,

$$\mathbb{E}\left\{\Big[\sum_{\ell=0}^{k-1}|\boldsymbol{Z}_i^\ell|^2\Big]^p\right\}^{1/(2p)} \le \Big(\frac{\sqrt{2d}v}{\alpha}\Big)\sqrt{p}e^{\alpha^2 k/(2p)}\,. \tag{D.22}$$

By taking $\alpha = \sqrt{p/k}$ (which is allowed provided $p \le \sqrt{kd/2}$), we obtain that

$$A_{1,p}(k) \le 10\sqrt{kd}v \le C_*\sqrt{kd}\,\varepsilon\sigma_\infty\gamma\,. \tag{D.23}$$

We next consider the total variation of the process $\boldsymbol{V}_i$ in Eq. (D.17). We have

$$\begin{aligned}
\mathbb{E}\big(|\boldsymbol{V}_i - \overline{\boldsymbol{V}}_i|_k^{2p}\big)^{1/(2p)} &= \mathbb{E}\left\{\left(\sum_{\ell=0}^{k-1}\varepsilon\Big|\boldsymbol{G}(\boldsymbol{w}_i^\ell; \rho_\ell^{(N)}) - \boldsymbol{G}(\overline{\boldsymbol{w}}_i^\ell; \rho_{\ell\varepsilon})\Big|\right)^{2p}\right\}^{1/(2p)} \\
&\le \mathbb{E}\left\{\left(\sum_{\ell=0}^{k-1}\varepsilon\Big|\boldsymbol{G}(\boldsymbol{w}_i^\ell; \rho_\ell^{(N)}) - \boldsymbol{G}(\overline{\boldsymbol{w}}_i^\ell; \rho_\ell^{(N)})\Big|\right)^{2p}\right\}^{1/(2p)} \\
&\quad + \mathbb{E}\left\{\left(\sum_{\ell=0}^{k-1}\varepsilon\Big|\boldsymbol{G}(\overline{\boldsymbol{w}}_i^\ell; \rho_\ell^{(N)}) - \boldsymbol{G}(\overline{\boldsymbol{w}}_i^\ell; \rho_{\ell\varepsilon})\Big|\right)^{2p}\right\}^{1/(2p)} \\
&\equiv A_{2,p}(k) + A_{3,p}(k)\,.
\end{aligned}$$

Using the Lipschitz property of ∇V, ∇U, we get

$$A_{2,p}(k) \le L\varepsilon\mathbb{E}\left\{\left(\sum_{\ell=0}^{k-1}\left|\boldsymbol{w}_i^\ell - \overline{\boldsymbol{w}}_i^\ell\right|\right)^{2p}\right\}^{1/(2p)}$$
$$\le L\varepsilon\sum_{\ell=0}^{k-1}\mathcal{D}_{2p}(\ell)\,. \tag{D.24}$$

For the second term, we get, by triangular inequality,

$$A_{3,p}(k) \le \varepsilon\sum_{\ell=0}^{k-1}\mathbb{E}\left\{\left|\boldsymbol{G}(\overline{\boldsymbol{w}}_i^\ell;\rho_\ell^{(N)}) - \boldsymbol{G}(\overline{\boldsymbol{w}}_i^\ell;\rho_{\ell\varepsilon})\right|^{2p}\right\}^{1/(2p)}\,. \tag{D.25}$$

We next use the expression $\boldsymbol{G}(\boldsymbol{w};\rho) = \nabla V(\boldsymbol{w}) + \int\nabla U(\boldsymbol{w},\boldsymbol{w}')\,\rho(\mathrm{d}\boldsymbol{w}')$, and the fact that $\hat{\boldsymbol{w}}_i^\ell \sim \rho_{\ell\varepsilon}$, to get

$$A_{3,p}(k) \le \varepsilon\sum_{\ell=0}^{k-1}\mathbb{E}\left\{\left|\frac{1}{N}\sum_{j=1}^N\left[\nabla U(\overline{\boldsymbol{w}}_i^\ell,\boldsymbol{w}_j^\ell) - \mathbb{E}_{\hat{\boldsymbol{w}}_j^\ell}\nabla U(\overline{\boldsymbol{w}}_i^\ell,\hat{\boldsymbol{w}}_j^\ell)\right]\right|^{2p}\right\}^{1/(2p)}$$
$$\le \varepsilon\sum_{\ell=0}^{k-1}\mathbb{E}\left\{\left|\frac{1}{N}\sum_{j=1}^N\left[\nabla U(\overline{\boldsymbol{w}}_i^\ell,\boldsymbol{w}_j^\ell) - \nabla U(\overline{\boldsymbol{w}}_i^\ell,\hat{\boldsymbol{w}}_j^\ell)\right]\right|^{2p}\right\}^{1/(2p)}$$
$$+\varepsilon\sum_{\ell=0}^{k-1}\mathbb{E}\left\{\left|\frac{1}{N}\sum_{j=1}^N\left[\nabla U(\overline{\boldsymbol{w}}_i^\ell,\hat{\boldsymbol{w}}_j^\ell) - \mathbb{E}_{\hat{\boldsymbol{w}}_j^\ell}\nabla U(\overline{\boldsymbol{w}}_i^\ell,\hat{\boldsymbol{w}}_j^\ell)\right]\right|^{2p}\right\}^{1/(2p)}$$
$$\equiv A_{3,p}^{(1)}(k) + A_{3,p}^{(2)}(k)\,.$$

Using once more the Lipschitz property of ∇U, and the symmetry of the distributions of $(\boldsymbol{w}^\ell)_{i\le N}$, $(\hat{\boldsymbol{w}}^\ell)_{i\le N}$ under permutations, we obtain

$$A_{3,p}^{(1)}(k) \le L\varepsilon\sum_{\ell=0}^{k-1}\mathbb{E}\left\{\left|\boldsymbol{w}_j^\ell - \hat{\boldsymbol{w}}_j^\ell\right|^{2p}\right\}^{1/(2p)} \tag{D.26}$$
$$\le L\varepsilon\sum_{\ell=0}^{k-1}\mathcal{D}_{2p}(\ell) + L\varepsilon\sum_{\ell=0}^{k-1}\widehat{\mathcal{D}}_{2p}(\ell)\,. \tag{D.27}$$

Finally, $|\nabla U(\overline{\boldsymbol{w}}_i^\ell,\hat{\boldsymbol{w}}_j^\ell)| \le L$ and therefore the vector

$$\boldsymbol{W} = \frac{1}{N}\sum_{j=1}^N\left[\nabla U(\overline{\boldsymbol{w}}_i^\ell,\hat{\boldsymbol{w}}_j^\ell) - \mathbb{E}_{\hat{\boldsymbol{w}}_j^\ell}\nabla U(\overline{\boldsymbol{w}}_i^\ell,\hat{\boldsymbol{w}}_j^\ell)\right]$$

is sub-Gaussian, with variance proxy upper bounded by $v^2 = L^2/N$. This implies that $\mathbb{E}\{|\boldsymbol{W}|^{2p}\}^{1/(2p)} \le C_*\sqrt{dp}\,v$, and therefore

$$A_{3,p}^{(2)}(k) \le C_*(k\varepsilon)L\sqrt{\frac{dp}{N}}\,. \tag{D.28}$$

Substituting (D.23), (D.24), (D.27), (D.28) in Eq. (D.17), we obtain

$$\mathcal{D}_{2p}(k) \le C_* L p \varepsilon \sum_{\ell=0}^{k-1} \mathcal{D}_{2p}(\ell) + C_* L p (k\varepsilon) \widehat{\mathcal{D}}_{2p}(k) + C_* p \sqrt{kd}\, \varepsilon \sigma_\infty \gamma + C_* L p (k\varepsilon) \sqrt{\frac{dp}{N}} \,. \tag{D.29}$$

Using Eq. (D.10) and Gronwall inequality, along with the fact that $k\varepsilon \le T$, this yields

$$\mathcal{D}_{2p}(T/\varepsilon) \le C_*\, e^{C_* p L T} \left[\sqrt{\frac{d}{N}} \vee \left(\sigma_\infty \gamma \sqrt{d\varepsilon}\right) \vee \left(L r (d^2 \varepsilon \log(1/\varepsilon))^{1/4} \right) \right] .$$

By using Eq. (D.10) again, we get

$$\begin{aligned} &\mathcal{D}_{2p}(T/\varepsilon) + \widehat{\mathcal{D}}_{2p}(T/\varepsilon) \\ &\le C_*\, e^{C_* p L T} \left[\sqrt{\frac{d}{N}} \vee \left(\sigma_\infty \gamma \sqrt{d\varepsilon}\right) \vee \left(L r (d^2 \varepsilon \log(1/\varepsilon))^{1/4} \right) \right] \\ &\equiv e^{C_* p L T} \mathsf{err}(N, d, \varepsilon) \,. \end{aligned} \tag{D.30}$$

By Markov inequality along with the Jensen inequality applied to the convex function x^{2p}, we have

$$\begin{aligned} \mathbb{P}\left\{ \frac{1}{N} \sum_{i=1}^N |\boldsymbol{w}_i^k - \hat{\boldsymbol{w}}_i^k| \ge \Delta \right\} &\le \frac{1}{\Delta^{2p}} \mathbb{E}\left\{ \left[\frac{1}{N} \sum_{i=1}^N |\boldsymbol{w}_i^k - \hat{\boldsymbol{w}}_i^k| \right]^{2p} \right\} \\ &\le \frac{1}{\Delta^{2p}} \mathbb{E}\left\{ \frac{1}{N} \sum_{i=1}^N |\boldsymbol{w}_i^k - \hat{\boldsymbol{w}}_i^k|^{2p} \right\} \\ &\le \frac{1}{\Delta^{2p}} \left[\mathcal{D}_{2p}(T/\varepsilon) + \widehat{\mathcal{D}}_{2p}(T/\varepsilon) \right]^{2p} \\ &\le \frac{1}{\Delta^{2p}} e^{2C_* p^2 L T} \mathsf{err}(N, d, \varepsilon)^{2p} \,, \end{aligned}$$

where in the third step we used (D.8) and (D.9). Set $\Delta = z\, e^{C_* p L T} \mathsf{err}(N, d, \varepsilon)$. Thus, we obtain

$$\frac{1}{N} \sum_{i=1}^N |\boldsymbol{w}_i^k - \hat{\boldsymbol{w}}_i^k| \le z\, e^{C_* p L t} \mathsf{err}(N, d, \varepsilon) \,, \tag{D.31}$$

with probability at least $1 - z^{-2p}$.

The bounds in Eq. (D.3) follow straightforwardly from Eq. (D.31) as in the proofs of Lemma 3.3 and 3.4 in the supplementary material of [MMN18]. □

E Regularity of the solutions of the PDE (3.9) ($\delta > 0$)

In this section we prove some standard regularity properties of the solutions of the PDE (3.9), for $\delta > 0$, and indeed for the more general PDE (B.2). First of all, we show that the weak solution of

the PDE (B.2) is in fact strong, i.e., $\rho \in \mathscr{C}^{2,1}(\Omega^\delta, [0,T])$ and the equation (B.2) holds pointwise. We will then prove upper bounds on $\nabla K^\delta * \rho$ and $\nabla U^\delta * \rho$ that are uniform in δ. These will be crucial in order to take the $\delta \to 0$ limit in the next section.

We start by proving a bound on the $\mathscr{L}^\infty$ norm of ρ. In the proofs of the two lemmas that follow, we assume without loss of generality that $\tau = 1$.

Lemma E.1 (Bound on $\mathscr{L}^\infty$ norm)**.** *Let ρ_t be a weak solution of the PDE (B.2) with initial and boundary conditions (B.3). Recall that ρ_t has a density with respect to Lebesgue measure, denoted by $\rho(\,\cdot\,,t)$. Then, there exists a constant $C(\Omega)$ such that, by letting $L = (\|\nabla V\|_{\mathscr{L}^\infty(\Omega)} \vee \|\nabla U\|_{\mathscr{L}^\infty(\Omega\times\Omega)})$, we have*

$$\|\rho(\,\cdot\,,t)\|_{\mathscr{L}^\infty(\Omega)} \le \|\rho_{\mathrm{init}}\|_{\mathscr{L}^\infty(\Omega)}\, e^{C(\Omega)L^2\, t}\,. \tag{E.1}$$

Proof. Any solution the PDE (B.2) satisfies Eq. (B.6). Given a measurable (Borel) function $\rho \in m\mathcal{B}(\Omega \times [0,T])$, denote by $\mathscr{D}(\rho) \in m\mathcal{B}(\Omega \times [0,T])$ the function given by the right-hand side of (B.2). Let $C(\Omega)$ be the constant in the statement of Theorem G.1 (part 3) and let $C_{U,V} \equiv C(\Omega)(\|\nabla V\|_{\mathscr{L}^\infty(\Omega)} + \|\nabla U\|_{\mathscr{L}^\infty(\Omega\times\Omega)})$. We then have

$$\begin{aligned}
\|\mathscr{D}(\rho)(\,\cdot\,,t)\|_{\mathscr{L}^\infty(\Omega)} &\le \|\rho_{\mathrm{init}}\|_{\mathscr{L}^\infty(\Omega)} + \big(\|\nabla V\|_{\mathscr{L}^\infty(\Omega)} + \|\nabla U\|_{\mathscr{L}^\infty(\Omega\times\Omega)}\big) \\
&\qquad \int_0^t \sup_{\boldsymbol{x}\in\Omega} \big\|\nabla G^\Omega(\boldsymbol{x},\,\cdot\,;t-s)\big\|_{\mathscr{L}^1(\Omega)} \|\rho(\,\cdot\,,s)\|_{\mathscr{L}^\infty(\Omega)}\,\mathrm{d}s \\
&\le \|\rho_{\mathrm{init}}\|_{\mathscr{L}^\infty(\Omega)} + C_{U,V}\int_0^t (t-s)^{-1/2} \|\rho(\,\cdot\,,s)\|_{\mathscr{L}^\infty(\Omega)}\,\mathrm{d}s.
\end{aligned} \tag{E.2}$$

Hence

$$\|\mathscr{D}(\rho)\|_{\mathscr{L}^\infty(\Omega\times[0,T])} \le \|\rho_{\mathrm{init}}\|_{\mathscr{L}^\infty(\Omega)} + C_{U,V}\sqrt{T}\,\|\rho\|_{\mathscr{L}^\infty(\Omega\times[0,T])}\,. \tag{E.3}$$

Proceeding analogously for two different densities $\rho, \tilde\rho$, we get

$$\|\mathscr{D}(\rho) - \mathscr{D}(\tilde\rho)\|_{\mathscr{L}^\infty(\Omega\times[0,T])} \le C_{U,V}\sqrt{T}\,\|\rho - \tilde\rho\|_{\mathscr{L}^\infty(\Omega\times[0,T])}\,. \tag{E.4}$$

Hence $\mathscr{D}$ maps $\mathscr{L}^\infty(\Omega\times[0,T])$ into itself, and is a contraction for $C_{U,V}\sqrt{T} < 1$. Therefore, it must have a unique fixed point in $\mathscr{L}^\infty$ that coincides with the unique solution of PDE (B.2). Let $T_0 = 1/(4C_{UV}^2)$. Then for that fixed point $\rho \in \mathscr{L}^\infty(\Omega\times[0,T])$ we have from Eq. (E.3)

$$\|\rho\|_{\mathscr{L}^\infty(\Omega\times[0,T_0])} \le 2\,\|\rho_{\mathrm{init}}\|_{\mathscr{L}^\infty(\Omega)}\,. \tag{E.5}$$

The desired claim follow by iterating this inequality $\lceil t/T_0 \rceil$ times. □

Lemma E.2 (Strong solutions of PDE)**.** *Let ρ_t be a weak solution of the PDE (B.2) with initial and boundary conditions (B.3), and recall that, for any $t \le T < \infty$, this has a density $\rho(\,\cdot\,,t)$, with $\rho \in \mathscr{L}^\infty(\Omega\times[0,T])$. Fix $q \in \mathbb{N}$. If $\rho_{\mathrm{init}} \in \mathscr{C}^q(\Omega)$, then $\rho \in \mathscr{C}^{q,1}(\Omega,[0,T])$.*

Proof. We prove the claim for $q = 2$. For larger values of q, the proof is similar and it only requires to iterate the argument.

The proof uses the same bootstrap technique of [MMN18][Supplementary material, Lemma 6.7]. The only difference is that the Duhamel formula of Eq. (B.6) involves the Neumann heat kernel in Ω instead of the heat kernel in $\mathbb{R}^d$.

Let $S = \Omega \times [0,T]$ and, for $u : \Omega \times [0,T] \to \mathbb{R}$. For $r \in \mathbb{N}$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, let $D_t^r D_{\boldsymbol{x}}^{\boldsymbol{\alpha}} u$ be the generalized derivative of u, and define the parabolic seminorm

$$\langle\langle u \rangle\rangle^{(j)}_{\mathscr{L}^p(S)} \equiv \sum_{|\boldsymbol{\alpha}|+2r=j} \|D_t^r D_{\boldsymbol{x}}^{\boldsymbol{\alpha}} u\|_{\mathscr{L}^p(S)}. \tag{E.6}$$

The proof of [MMN18][Supplementary material, Lemma 6.7] uses the following inequality from [LSU88][Chapter IV, Section 3, Eq. (3.1)]

$$\langle\langle G *_2 u \rangle\rangle^{(2m+2)}_{\mathscr{L}^p(S)} \le C \, \langle\langle u \rangle\rangle^{(2m)}_{\mathscr{L}^p(S)}, \tag{E.7}$$

$$G *_2 u(\boldsymbol{x}, t) \equiv \int_{\mathbb{R}^d} \int_0^t G(\boldsymbol{x}, \boldsymbol{y}, \, t-s) \, u(\boldsymbol{y}, s) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}s \, . \tag{E.8}$$

Furthermore, (G.11) of Theorem G.1 yields

$$\langle\langle G^\Omega *_2 u \rangle\rangle^{(2m+2)}_{\mathscr{L}^p(S)} \le \langle\langle G *_2 u \rangle\rangle^{(2m+2)}_{\mathscr{L}^p(S)} + \langle\langle G_R^\Omega *_2 u \rangle\rangle^{(2m+2)}_{\mathscr{L}^p(S)} \, . \tag{E.9}$$

Since $G_R^\Omega \in \mathscr{C}^\infty(\Omega \times \Omega \times [0,T])$, we have that

$$\langle\langle G_R^\Omega *_2 u \rangle\rangle^{(2m+2)}_{\mathscr{L}^p(S)} \le C \|u\|_{\mathscr{L}^p(S)},$$

which immediately implies that

$$\langle\langle G^\Omega *_2 u \rangle\rangle^{(2m+2)}_{\mathscr{L}^p(S)} \le C \langle\langle u \rangle\rangle^{(2m)}_{\mathscr{L}^p(S)} + C \|u\|_{\mathscr{L}^p(S)} \, . \tag{E.10}$$

The proof of [MMN18][Supplementary material, Lemma 6.7] can be repeated verbatimly with (E.7) replaced by (E.10). □

As a consequence of the last lemma, the PDE (B.2) admits unique strong solutions $\rho \in \mathscr{C}^{2,1}(\Omega, [0,T])$ with initial condition ρ_{init} and Neumann boundary condition. We will use $\rho(t)$ as shortcut for $\rho(\,\cdot\,, t)$. The rest of this appendix is devoted to prove further regularity results for $\rho(t)$, which will be crucial in the proofs provided in Appendix F. To emphasize the dependence of ρ on δ, we will denote this solution by ρ^δ.

In what follows, we will set the initial condition $\rho^\delta(0) \equiv \rho^\delta_{\text{init}}$ at $\delta > 0$ to be defined via $\rho^\delta_{\text{init}}(\boldsymbol{w}) = \lambda_\delta^{-d} \rho_{\text{init}}(\boldsymbol{w}/\lambda_\delta)$, with λ_δ given by Eq. (3.4)

It is useful to recall the definition of free energy, which is given by

$$\begin{aligned} F^\delta(\rho) &= \frac{1}{2} R^\delta(\rho) - \tau \, S(\rho) \\ &= \frac{\nu_0}{2} \|f - K^\delta * \rho\|^2_{\mathscr{L}^2(\Omega)} + \tau \int \rho(\boldsymbol{x}) \log \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, . \end{aligned} \tag{E.11}$$

The following lemma provides an expression for the derivative of the free energy with respect to time. Such an expression immediately yields an upper bound on the $\mathscr{L}^2(\Omega)$ norm of $K^\delta * \rho^\delta(t)$ which is independent of δ.

Lemma E.3. *Let $\rho^\delta \in \mathscr{C}^{2,1}(\Omega^\delta, [0,T])$ be the solution of the PDE (B.2) with initial and boundary conditions (B.3). Then,*

$$\frac{\mathrm{d}}{\mathrm{d}t} F^\delta(\rho^\delta(t)) = - \int \left| \nabla \big(\Psi(\boldsymbol{x}; \rho^\delta(t)) + \tau \log \rho^\delta(\boldsymbol{x}, t) \big) \right|^2 \rho^\delta(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} \, . \tag{E.12}$$

Proof. By definition

$$F^\delta(\rho^\delta(t)) = \frac{\nu_0}{2}\|f\|^2_{\mathscr{L}^2(\Omega)} + \int V(\boldsymbol{w})\rho^\delta(\boldsymbol{w},t)\mathrm{d}\boldsymbol{w} + \frac{1}{2}\int U(\boldsymbol{w}_1-\boldsymbol{w}_2)\rho^\delta(\boldsymbol{w}_1,t)\mathrm{d}\boldsymbol{w}_1\rho^\delta(\boldsymbol{w}_2,t)\mathrm{d}\boldsymbol{w}_2 + \tau\int \rho^\delta(\boldsymbol{w},t)\log\rho^\delta(\boldsymbol{w},t)\mathrm{d}\boldsymbol{w}\,.$$

By differentiating $F^\delta(\rho^\delta(t))$ along the solution of (B.2), we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}F^\delta(\rho^\delta(t)) &= \int V(\boldsymbol{w})\partial_t\rho^\delta(\boldsymbol{w},t)\mathrm{d}\boldsymbol{w} \\
&\quad + \int U(\boldsymbol{w}_1-\boldsymbol{w}_2)\rho^\delta(\boldsymbol{w}_1,t)\partial_t\rho^\delta(\boldsymbol{w}_2,t)\mathrm{d}\boldsymbol{w}_1\mathrm{d}\boldsymbol{w}_2 \\
&\quad + \tau\int(1+\log\rho^\delta(\boldsymbol{w},t))\partial_t\rho^\delta(\boldsymbol{w},t)\mathrm{d}\boldsymbol{w} \\
&= \int(\Psi(\boldsymbol{w},\rho^\delta(t)) + \tau\log\rho^\delta(\boldsymbol{w},t) + \tau)\partial_t\rho^\delta(\boldsymbol{w},t)\mathrm{d}\boldsymbol{w} \\
&= \int(\Psi(\boldsymbol{w},\rho^\delta(t)) + \tau\log\rho^\delta(\boldsymbol{w},t) + \tau) \\
&\qquad \nabla\cdot\Big(\rho^\delta(\boldsymbol{w},t)\nabla\Big(\Psi(\boldsymbol{w},\rho^\delta(t)) + \tau\log\rho^\delta(\boldsymbol{w},t)\Big)\Big)\mathrm{d}\boldsymbol{w} \\
&= -\int\langle\nabla\Big(\Psi(\boldsymbol{w},\rho^\delta(t) + \tau\log\rho^\delta(\boldsymbol{w},t)\Big), \\
&\qquad \nabla(\Psi(\boldsymbol{w},\rho^\delta(t)) + \tau\log\rho^\delta(\boldsymbol{w},t))\rangle\rho^\delta(\boldsymbol{w},t)\mathrm{d}\boldsymbol{w} \\
&= -\int\left|\nabla\big(\Psi(\boldsymbol{w};\rho^\delta(t)) + \tau\log\rho^\delta(\boldsymbol{w},t)\big)\right|^2\rho^\delta(\boldsymbol{w},t)\,\mathrm{d}\boldsymbol{w} \qquad \text{(E.13)}
\end{aligned}$$

□

Corollary E.4. *Let $\rho^\delta \in \mathscr{C}^{2,1}(\Omega^\delta,[0,T])$ be the solution of the PDE (B.2) with initial and boundary conditions (B.3). Then,*

$$\nu_0\|K^\delta * \rho^\delta(t) - f\|^2_{\mathscr{L}^2(\Omega)} \le 2F^\delta(\rho^\delta(0)) + 2\tau\log|\Omega^\delta|\,, \qquad \text{(E.14)}$$

where $|\Omega^\delta|$ denotes the volume of the set Ω^δ.

Proof. By Lemma E.3 we have $F^\delta(\rho^\delta(t)) \le F^\delta(\rho^\delta(0))$. The claim follows by substituting the definition of $F^\delta(\rho^\delta)$ and using $S(\rho^\delta) \le \log|\Omega^\delta|$. □

Remark E.1. By Corollary E.4, we are able to provide a δ-free upper bound on $\nu_0\|K^\delta * \rho^\delta(t) - f\|^2_{\mathscr{L}^2(\Omega)}$. Specifically, $\Omega^\delta \subseteq \Omega$ and hence $|\Omega^\delta| \le |\Omega|$. We also have

$$F^\delta(\rho^\delta(0)) \le \nu_0\|f\|^2_{\mathscr{L}^2(\Omega)} + \nu_0\|K^\delta * \rho^\delta(0)\|_{\mathscr{L}^2(\Omega)} - \tau S(\rho^\delta(0))\,.$$

Note that

$$S(\rho^\delta(0)) = S(\rho_{\mathrm{init}}) + d\log\lambda_\delta.$$

Since $\lambda_\delta \to 1$ as $\delta \to 0$, there exists a $C_* > 0$ such that for $\delta < C_*$, $\lambda_\delta \geq 1/2$. Thus, the term $S(\rho^\delta(0))$ has a δ-free upper bound.

By Young's inequality it only remains to give a δ-free upper bound on the quantity $\|\rho^\delta(0)\|_{\mathscr{L}^2(\Omega)}$. Let us write

$$\begin{aligned}\|\rho^\delta(0)\|^2_{\mathscr{L}^2(\Omega)} &\leq \int_{\Omega^\delta} \lambda_\delta^{-2d} \rho^2_{\mathrm{init}}(\boldsymbol{w}/\lambda_\delta) \mathrm{d}\boldsymbol{w} \\ &= \int_\Omega \lambda_\delta^{-2d} \rho^2_{\mathrm{init}}(\boldsymbol{x}) \lambda_\delta^d \mathrm{d}\boldsymbol{x} = \lambda_\delta^{-d} \|\rho^2_{\mathrm{init}}\|^2_{\mathscr{L}^2(\Omega)}.\end{aligned}$$

Again, for $\delta < C_*$, $\lambda_\delta \geq 1/2$. Also, by Assumption (A5) and the fact that Ω is compact, we have $\|\rho^2_{\mathrm{init}}\|^2_{\mathscr{L}^2(\Omega)} < \infty$, which concludes the claim.

We next prove δ-free upper bound on the gradient of $\nabla K^\delta * \rho^\delta$.

Lemma E.5. *Let $\rho^\delta \in \mathscr{C}^{2,1}(\Omega^\delta, [0,T])$ be the solution of the PDE (B.2) with initial and boundary conditions (B.3). Then, the following bound holds:*

$$\begin{aligned}\int_0^T \int |\nabla U * \rho^\delta(\boldsymbol{x},t)|^2 \rho^\delta(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t + 2\tau \int_0^T \int |\nabla (K^\delta * \rho^\delta)(\boldsymbol{x},t)|^2 \mathrm{d}\boldsymbol{x}\, \mathrm{d}t \\ \leq T\|\nabla V\|^2_{\mathscr{L}^\infty(\Omega)} + 2\nu_0 \|f\|^2_{\mathscr{L}^2(\Omega)} + 4F^\delta(\rho^\delta(0)) + 4\tau \log|\Omega^\delta|\,.\end{aligned} \tag{E.15}$$

Proof. Denote by $\langle f, g\rangle = \int f(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ the standard scalar product in $\mathscr{L}^2$. Then,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle U*\rho^\delta(t), \rho^\delta(t)\rangle &= \langle U*\rho^\delta(t), \partial_t \rho^\delta(t)\rangle \\ &= \langle U*\rho^\delta(t), \nabla\cdot(\rho^\delta(t)\nabla(V + U*\rho^\delta(t))) + \tau\Delta\rho^\delta(t)\rangle \\ &= \langle U*\rho^\delta(t), \nabla\cdot(\rho^\delta(t)\nabla V)\rangle + \langle U*\rho^\delta(t), \nabla\cdot(\rho^\delta(t)\nabla(U*\rho^\delta(t))\rangle \\ &\qquad + \tau\langle U*\rho^\delta(t), \Delta\rho^\delta(t)\rangle \\ &= -\int \langle \nabla(U*\rho^\delta)(\boldsymbol{x},t), \nabla V(\boldsymbol{x})\rangle\, \rho^\delta(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x} \\ &\qquad - \int |\nabla(U*\rho^\delta)(\boldsymbol{x},t)|^2\, \rho^\delta(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x} - \tau\int |\nabla(K^\delta*\rho^\delta)(\boldsymbol{x},t)|^2 \mathrm{d}\boldsymbol{x} \\ &\leq \frac{1}{2}\int |\nabla V(\boldsymbol{x})|^2\, \rho^\delta(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x} - \frac{1}{2}\int |\nabla(U*\rho^\delta)(\boldsymbol{x},t)|^2\, \rho^\delta(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x} \\ &\qquad - \tau\int |\nabla(K^\delta*\rho^\delta)(\boldsymbol{x},t)|^2 \mathrm{d}\boldsymbol{x}\,.\end{aligned} \tag{E.16}$$

By integrating (E.16) between 0 and T, we obtain

$$\begin{aligned}\int_0^T \int |\nabla U * \rho^\delta(\boldsymbol{x},t)|^2 \rho^\delta(\boldsymbol{x},t)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t + 2\tau \int_0^T \int |\nabla (K^\delta * \rho^\delta)(\boldsymbol{x},t)|^2 \mathrm{d}\boldsymbol{x}\, \mathrm{d}t \\ \leq T\|\nabla V\|^2_\infty - \langle \rho^\delta(T), U*\rho^\delta(T)\rangle + \langle \rho^\delta(0), U*\rho^\delta(0)\rangle \\ \leq T\|\nabla V\|^2_{\mathscr{L}^\infty(\Omega)} + \nu_0 \|K^\delta*\rho^\delta(0)\|^2_{\mathscr{L}^2(\Omega)}\,.\end{aligned} \tag{E.17}$$

Hence, (E.15) follows from Corollary E.4. □

Remark E.2. Note that by virtue of Lemma E.5, we are able to get a δ-free upper bound on the left-hand side of (E.15). Indeed, by definition of ∇V as per (B.4) and using Assumption (A3), we have the δ-free bound:

$$\|\nabla V\|_{\mathscr{L}^\infty(\Omega)} \leq \nu_0 \|K^\delta\|_{\mathscr{L}^1(\Omega)} \|\nabla f\|_{\mathscr{L}^\infty(\Omega)} = \|\nabla f\|_{\mathscr{L}^\infty(\Omega)} < C_* \,. \tag{E.18}$$

In addition, by Remark E.1, $\|K^\delta * \rho^\delta(0)\|^2_{\mathscr{L}^2(\Omega)}$ has δ-free bound.

F Global convergence: Proof of Theorems 5.2 and 5.3

We start by showing that ρ^δ admits a limit in a suitable functional space as $\delta \to 0$.

Lemma F.1 (Existence of converging subsequence)**.** *Let $\rho^\delta \in \mathscr{C}^{2,1}(\Omega^\delta, [0,T])$ be the unique solution of the PDE (B.2) with initial and boundary conditions (B.3). Then, the family $(\rho^\delta)_{\delta>0}$ is relatively compact in the space $\mathscr{C}([0,T], \mathscr{P}_2(\Omega))$. In particular any sequence $(\rho^{\delta_n})_{n\geq 1}$, admits a converging subsequence.*

Proof. This follows from the Ascoli-Arzelá's theorem. Notice that $\mathscr{P}_2(\Omega)$ is compact due to the compactness of Ω. Therefore, it is sufficient to prove that the family is equicontinuous. Using the representation in terms of nonlinear dynamics (cf. Appendix C), we have

$$W_2(\rho_t, \rho_s)^2 \leq \mathbb{E}\big\{ \big| \boldsymbol{X}_t - \boldsymbol{X}_s \big|^2 \big\} \,. \tag{F.1}$$

Note that we omit for simplicity the dependence on δ. Recall that the nonlinear dynamic satisfies (for $\boldsymbol{b}(\boldsymbol{x},t) \equiv -\nabla\Psi(\boldsymbol{x},\rho_t)$)

$$\boldsymbol{X}_t = \int_0^t \boldsymbol{b}(\boldsymbol{X}_r, r)\,\mathrm{d}r + \sqrt{2\tau}\,\boldsymbol{B}_t + \boldsymbol{\Phi}_t \tag{F.2}$$

$$\equiv \boldsymbol{V}_t + \sqrt{2\tau}\,\boldsymbol{B}_t + \boldsymbol{\Phi}_t \,. \tag{F.3}$$

By [Slo01, Theorem 2.2], we have

$$\mathbb{E}\big\{ |\boldsymbol{X}_t - \boldsymbol{X}_s|^2 \big\} \leq C_* \tau \mathbb{E}\{[\boldsymbol{B}]_s^t\} + C_* \mathbb{E}\{(|\boldsymbol{V}|_s^t)^2\} \,. \tag{F.4}$$

where $[\boldsymbol{B}]_s^t$ denotes the quadratic variation of $\boldsymbol{B}$, and $|\boldsymbol{V}|_s^t$ the total variation of $\boldsymbol{V}$ between times s and t. We thus have

$$\begin{aligned} W_2(\rho_t, \rho_s)^2 \leq \mathbb{E}\big\{ |\boldsymbol{X}_t - \boldsymbol{X}_s|^2 \big\} &\leq C_* \tau (t-s) + C_* \mathbb{E}\left\{ \left(\int_s^t |\boldsymbol{b}(\boldsymbol{X}_r, r)|\,\mathrm{d}r \right)^2 \right\} \\ &\leq C_* \tau (t-s) + C_* (t-s) \int_s^t \mathbb{E}\left\{ |\boldsymbol{b}(\boldsymbol{X}_r, r)|^2 \right\} \mathrm{d}r \,. \end{aligned} \tag{F.5}$$

Hence, in order to prove uniform continuity, it is sufficient to show that, for $s, t \leq T$, $\int_s^t \mathbb{E}\big\{ |\boldsymbol{b}(\boldsymbol{X}_r, r)|^2 \big\}\,\mathrm{d}r \leq C$ where C is bounded uniformly in δ. In order to show that this is the case, notice that

$$\int_s^t \mathbb{E}\big\{ |\boldsymbol{b}(\boldsymbol{X}_r, r)|^2 \big\}\,\mathrm{d}r = \int_s^t \mathbb{E}\big\{ |\nabla V(\boldsymbol{X}_r) + \nabla U * \rho(\boldsymbol{X}_r, r)|^2 \big\}\,\mathrm{d}r \tag{F.6}$$

$$\leq 2(t-s)\|\nabla V\|^2_{\mathscr{L}^\infty(\Omega^\delta)} + 2\int_s^t \int |\nabla U * \rho(\boldsymbol{x}, r)|^2 \rho(\boldsymbol{x}, r)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}r \,, \tag{F.7}$$

and the claim follows from Lemma E.5. □

We have now proved that the sequence $(\rho^{\delta_n})_{n\geq 1}$ admits a converging subsequence, where $\delta_n \to 0$ as $n \to \infty$. Fix such a convergent subsequence and, with an abuse of notation, also denote it by $(\rho^{\delta_n})_{n\geq 1}$. Let $\rho^\infty \in \mathscr{C}([0,T], \mathscr{P}_2(\Omega))$ be its limit.

Recall that ρ^{δ_n} is supported in Ω^{δ_n}. Hence, $K^{\delta_n} * \rho^{\delta_n}$ is supported in Ω and $K^{\delta_n} * \rho^{\delta_n} \in \mathscr{P}_2(\Omega)$. We will now show that $(K^{\delta_n} * \rho^{\delta_n})_{n\geq 1}$ has the same limit as $(\rho^{\delta_n})_{n\geq 1}$ in $\mathscr{C}([0,T], \mathscr{P}_2(\Omega))$.

Lemma F.2. *The sequence $(K^{\delta_n} * \rho^{\delta_n})_{n\geq 1}$ also converges in $\mathscr{C}([0,T], \mathscr{P}_2(\Omega))$ to ρ^∞.*

Proof. By Lemma F.1, the result is implied by the following claim:

$$\lim_{n\to\infty} \sup_{0\leq t\leq T} W_2(K^{\delta_n} * \rho_t^{\delta_n}, \rho_t^{\delta_n}) = 0. \tag{F.8}$$

Note that, for bounded Ω,

$$\left(\int |\boldsymbol{x}-\boldsymbol{y}|^2 \gamma(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\right)^{1/2} \leq \mathrm{diam}(\Omega)^{1/2} \left(\int |\boldsymbol{x}-\boldsymbol{y}| \gamma(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y})\right)^{1/2},$$

for any coupling γ of the probability distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$. Hence,

$$W_2(\rho_1, \rho_2) \leq \mathrm{diam}(\Omega)^{1/2} W_1(\rho_1, \rho_2)\,. \tag{F.9}$$

As an application,

$$W_2^2(K^{\delta_n} * \rho_t^{\delta_n}, \rho_t^{\delta_n}) \leq \mathrm{diam}(\Omega)^{1/2} W_1(K^{\delta_n} * \rho_t^{\delta_n}, \rho_t^{\delta_n}). \tag{F.10}$$

Thus, it suffices to show that $\sup_{0\leq t\leq T} W_1(K^{\delta_n} * \rho_t^{\delta_n}, \rho_t^{\delta_n}) \to 0$ as $n \to \infty$.

Note that

$$W_1(K^{\delta_n} * \rho_t^{\delta_n}, \rho_t^{\delta_n}) \leq \mathbb{E}\{|(\boldsymbol{K}_{\delta_n} + \boldsymbol{X}_{\delta_n}) - \boldsymbol{X}_{\delta_n}|\} = \mathbb{E}\{|\boldsymbol{K}_{\delta_n}|\},$$

where the random variables $\boldsymbol{K}_{\delta_n}$ and $\boldsymbol{X}_{\delta_n}$ have distributions K^{δ_n} and $\rho_t^{\delta_n}$, respectively. The quantity $\mathbb{E}\{|\boldsymbol{K}_{\delta_n}|\}$ is $O(\delta)$, since K has bounded absolute first moment, which completes the proof. □

We will now prove a stronger convergence result.

Lemma F.3 (Convergence in $\mathscr{L}^2$)**.** *The measure ρ^∞ has a density, which is the limit in $\mathscr{L}^2(\Omega \times [0,T])$ of the sequence $(K^{\delta_n} * \rho^{\delta_n})_{n\geq 1}$.*

Proof. By Corollary E.4, we have that, for any $n \geq 1$, $K^{\delta_n} * \rho^{\delta_n} \in \mathscr{L}^2(\Omega \times [0,T])$. Let us show that $(K^{\delta_n} * \rho^{\delta_n})_{n\geq 1}$ is a Cauchy sequence in $\mathscr{L}^2(\Omega \times [0,T])$.

As $K^{\delta_n} * \rho_t^{\delta_n} \in \mathscr{L}^2(\Omega)$ for every $t \in [0,T]$, its Fourier transform exists and we denote it by $\widehat{K^{\delta_n} * \rho^{\delta_n}}$. Hence, by applying Parseval's theorem, we have

$$\begin{aligned}
&\limsup_{n,n'\to\infty} \|K^{\delta_n} * \rho^{\delta_n} - K^{\delta_{n'}} * \rho^{\delta_{n'}}\|^2_{\mathscr{L}^2(\Omega\times[0,T])} \\
&\quad = \limsup_{n,n'\to\infty} \int_0^T \|K^{\delta_n} * \rho_t^{\delta_n} - K^{\delta_{n'}} * \rho_t^{\delta_{n'}}\|^2_{\mathscr{L}^2(\Omega)}\,\mathrm{d}t \\
&\quad = \limsup_{n,n'\to\infty} \int_0^T \int_{\mathbb{R}^d} |\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{K^{\delta_{n'}} * \rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2\,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t.
\end{aligned} \tag{F.11}$$

Fix $\Lambda > 1$ and decompose the integral in the right-hand side of (F.11) as

$$\begin{aligned}\limsup_{n,n'\to\infty} \int_0^T \int_{|\boldsymbol{\lambda}|<\Lambda} |\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{K^{\delta_{n'}} * \rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t \\ + \limsup_{n,n'\to\infty} \int_0^T \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{K^{\delta_{n'}} * \rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t.\end{aligned} \tag{F.12}$$

Consider the first term of (F.12). By Lemma F.2, and since by Jensen's inequality $W_1(\rho_1,\rho_2) \leq W_2(\rho_1,\rho_2)$ for any two distributions ρ_1,ρ_2, we have $W_1(K^{\delta_n} * \rho_t^{\delta_n} - K^{\delta_{n'}} * \rho_t^{\delta_{n'}}) \to 0$, as $n,n' \to \infty$. Since for the complex exponential functions $\|e^{i\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle}\|_{\mathrm{Lip}} \leq |\boldsymbol{\lambda}|$, by definition of 1-Wasserstein distance, the integrand in the first term converges pointwise to 0. Furthermore, the integrand is upper bounded by an integrable function, since $|\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda})| \leq \|K^{\delta_n} * \rho_t^{\delta_n}\|_{\mathscr{L}^2(\Omega)} \leq C$ for all n and every $t \in [0,T]$. Hence, by dominated convergence, the first integral in (F.12) converges to 0.

As for the second term of (F.12), the following chain of inequalities holds:

$$\begin{aligned}&\limsup_{n,n'\to\infty} \int_0^T \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{K^{\delta_{n'}} * \rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t \\ &\quad\leq 4 \sup_{n\geq1} \int_0^T \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t \\ &\quad\leq \frac{4}{\Lambda^2} \sup_{n\geq1} \int_0^T \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\boldsymbol{\lambda}|^2 \,|\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t \\ &\quad\leq \frac{4}{\Lambda^2} \sup_{n\geq1} \int_0^T \int_{\mathbb{R}^d} |\boldsymbol{\lambda}|^2 \,|\widehat{K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t \\ &\quad= \frac{4}{\Lambda^2} \sup_{n\geq1} \int_0^T \int_{\mathbb{R}^d} |\widehat{\nabla K^{\delta_n} * \rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}\,\mathrm{d}t \\ &\quad= \frac{4}{\Lambda^2} \sup_{n\geq1} \int_0^T \int_{\Omega} |\nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t,\end{aligned} \tag{F.13}$$

where in the last equality we have applied again Parseval's theorem. By Lemma E.5, the integral in the right-hand side of (F.13) is upper bounded by a constant independent of n. Therefore, as $\Lambda \to \infty$, the second term of (F.12) converges to 0.

As a result, $(K^{\delta_n} * \rho^{\delta_n})_{n\geq1}$ is a Cauchy sequence in $\mathscr{L}^2(\Omega\times[0,T])$. Let $\tilde{\rho}^\infty \in \mathscr{L}^2(\Omega\times[0,T])$ be its limit. Furthermore, by Lemma F.2, $(K^{\delta_n} * \rho^{\delta_n})_{n\geq1}$ has limit ρ^∞ in $\mathscr{C}([0,T],\mathscr{P}_2(\Omega))$. Therefore, the measures $\rho_t^\infty(\mathrm{d}\boldsymbol{x})\mathrm{d}t$ and $\tilde{\rho}_t^\infty(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,\mathrm{d}t$ coincide. This implies that the measure ρ_t^∞ has for almost every $t \in [0,T]$ the density $\tilde{\rho}_t^\infty \in \mathscr{L}^2(\Omega\times[0,T])$, and the proof is complete. □

From now on, with an abuse of notation, we will use ρ^∞ to denote also the density which is the limit in $\mathscr{L}^2(\Omega\times[0,T])$ of the sequence $(K^{\delta_n} * \rho^{\delta_n})_{n\geq1}$.

Lemma F.4 (Convergence to a weak solution of the limit PDE)**.** *Let ρ^∞ be the limit in $\mathscr{C}([0,T],\mathscr{P}_2(\Omega))$ of the converging sequence $(\rho^{\delta_n})_{n\geq1}$. Then, ρ^∞ is a weak solution of the PDE (A.1) with initial and boundary conditions (A.2).*

Proof. By Lemma F.3, we have that $\rho^\infty \in \mathscr{L}^2(\Omega\times[0,T])$. Choose a test function $h \in \mathscr{C}^{2,1}(\Omega\times[0,T])$, satisfying $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h(\boldsymbol{x},t)\rangle = 0$ for all $\boldsymbol{x} \in \partial\Omega, t \in [0,T]$. In order to prove the claim, we need to show that (A.3) holds. Throughout the proof, we will let $\lambda_n \equiv \lambda_{\delta_n}$.

Recall that, for any $n \geq 1$, ρ^{δ_n} is a weak solution of the PDE (B.2) with initial and boundary conditions (B.3). Hence, by Definition B.1, we have that

$$\begin{aligned}\int_{\Omega^{\delta_n}} h^{\delta_n}(\boldsymbol{x},T)\,\rho_T^{\delta_n}(\mathrm{d}\boldsymbol{x}) - \int_{\Omega^{\delta_n}} h^{\delta_n}(\boldsymbol{x},0)\,\rho_0^{\delta_n}(\mathrm{d}\boldsymbol{x}) \\ = \int_0^T \int_{\Omega^{\delta_n}} \Big[\partial_t h^{\delta_n}(\boldsymbol{x},t) + \tau\Delta h^{\delta_n}(\boldsymbol{x},t) \\ - \langle \nabla\Psi(\boldsymbol{x},\rho_t^{\delta_n}), \nabla h^{\delta_n}(\boldsymbol{x},t)\rangle\Big]\,\rho_t^{\delta_n}(\mathrm{d}\boldsymbol{x})\,\mathrm{d}t\,,\end{aligned} \tag{F.14}$$

for any $h^{\delta_n} \in \mathscr{C}^{2,1}(\Omega^{\delta_n}\times[0,T])$ satisfying $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h^{\delta_n}(\boldsymbol{x},t)\rangle = 0$ for all $\boldsymbol{x}\in\partial\Omega^{\delta_n}, t\in[0,T]$. Now, we set

$$h^{\delta_n}(\boldsymbol{x},t) = h(\boldsymbol{x}/\lambda_n, t). \tag{F.15}$$

By definition of Ω_n^δ, we have that $h^{\delta_n} \in \mathscr{C}^{2,1}(\Omega^{\delta_n}\times[0,T])$ since $h\in\mathscr{C}^{2,1}(\Omega\times[0,T])$. Furthermore, $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h(\boldsymbol{x},t)\rangle = 0$ for all $\boldsymbol{x}\in\partial\Omega$, $t\in[0,T]$ immediately implies that $\langle \boldsymbol{n}(\boldsymbol{x}), \nabla h^{\delta_n}(\boldsymbol{x},t)\rangle = 0$ for all $\boldsymbol{x}\in\partial\Omega^{\delta_n}$, $t\in[0,T]$.

Recall that

$$\Psi(\boldsymbol{x},\rho_t^{\delta_n}) = V^{\delta_n}(\boldsymbol{x}) + U^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) = -\nu_0\, K^{\delta_n} * f(\boldsymbol{x}) + \nu_0\, K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}). \tag{F.16}$$

Thus, (F.14) can be rewritten as

$$\begin{aligned}&\int_{\Omega^{\delta_n}} h^{\delta_n}(\boldsymbol{x},T)\,\rho_T^{\delta_n}(\mathrm{d}\boldsymbol{x}) - \int_{\Omega^{\delta_n}} h^{\delta_n}(\boldsymbol{x},0)\,\rho_0^{\delta_n}(\mathrm{d}\boldsymbol{x}) \\ &= \int_0^T \int_{\Omega^{\delta_n}} \Big[\partial_t h^{\delta_n}(\boldsymbol{x},t) + \nu_0\,\langle \nabla K^{\delta_n} * f(\boldsymbol{x}), \nabla h^{\delta_n}(\boldsymbol{x},t)\rangle\Big]\,\rho_t^{\delta_n}(\mathrm{d}\boldsymbol{x})\,\mathrm{d}t \\ &+ \int_0^T \int_{\Omega^{\delta_n}} \Big[\tau\Delta h^{\delta_n}(\boldsymbol{x},t) - \nu_0\langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h^{\delta_n}(\boldsymbol{x},t)\rangle\Big]\,\rho_t^{\delta_n}(\mathrm{d}\boldsymbol{x})\,\mathrm{d}t\,.\end{aligned} \tag{F.17}$$

Since $(\rho^{\delta_n})_{n\geq 1}$ converges in $\mathscr{C}([0,T],\mathscr{P}_2(\Omega))$ to ρ^∞ by Lemma F.1, we have that

$$\begin{aligned}&\lim_{n\to\infty} \int_{\Omega^{\delta_n}} h^{\delta_n}(\boldsymbol{x},T)\,\rho_T^{\delta_n}(\mathrm{d}\boldsymbol{x}) = \int_\Omega h(\boldsymbol{x},T)\,\rho_T^\infty(\mathrm{d}\boldsymbol{x}), \\ &\lim_{n\to\infty} \int_0^T \int_{\Omega^{\delta_n}} \Big[\partial_t h^{\delta_n}(\boldsymbol{x},t) + \tau\Delta h^{\delta_n}(\boldsymbol{x},t)\Big]\,\rho_t^{\delta_n}(\mathrm{d}\boldsymbol{x})\,\mathrm{d}t \\ &\qquad = \int_0^T \int_\Omega \left[\partial_t h(\boldsymbol{x},t) + \tau\Delta h(\boldsymbol{x},t)\right]\,\rho_t^\infty(\mathrm{d}\boldsymbol{x})\,\mathrm{d}t.\end{aligned} \tag{F.18}$$

Furthermore, since $\rho_0^{\delta_n}(\boldsymbol{x}) = \lambda_n^{-d}\rho_{\mathrm{init}}(\boldsymbol{x}/\lambda_n)$, we have that

$$\int_{\Omega^{\delta_n}} h^{\delta_n}(\boldsymbol{x},0)\,\rho_0^{\delta_n}(\mathrm{d}\boldsymbol{x}) = \int_\Omega h(\boldsymbol{x},0)\,\rho_{\mathrm{init}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{F.19}$$

Let us use the notation $h_t(\boldsymbol{x}) = h(\boldsymbol{x},t)$ and $h_t^{\delta_n}(\boldsymbol{x}) = h^{\delta_n}(\boldsymbol{x},t)$. Again, we set $\rho_t^{\delta_n}(\boldsymbol{x}) = 0$ for $\boldsymbol{x}\notin\Omega^{\delta_n}$. We further define $\tilde{\rho}_t^{\delta_n}(\boldsymbol{x}) = \lambda_n^d\rho_t^{\delta_n}(\lambda_n\boldsymbol{x})$, which is a probability density on Ω. Since

$\rho_t^{\delta_n}(\,\cdot\,) \to \rho_t^{\infty}(\,\cdot\,)$ in $\mathscr{P}_2(\Omega)$ and $\lambda_n \to 1$, we have $\tilde{\rho}_t^{\delta_n}(\,\cdot\,) \to \rho_t^{\infty}(\,\cdot\,)$ in $\mathscr{P}_2(\Omega)$ as well. Hence

$$\begin{aligned}
&\lim_{n\to\infty} \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * f(\boldsymbol{x}), \nabla h_t^{\delta_n}(\boldsymbol{x}) \rangle \, \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\mathrm{d}t \qquad \text{(F.20)}\\
&= \lim_{n\to\infty} \lambda_n^{-1} \int_0^T \int_{\Omega} \langle \nabla K^{\delta_n} * f(\lambda_n \boldsymbol{x}), \nabla h_t(\boldsymbol{x}) \rangle \, \tilde{\rho}_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= \int_0^T \int_{\Omega} \langle \nabla f(\boldsymbol{x}), \nabla h_t^{\delta_n}(\boldsymbol{x}) \rangle \, \rho_t^{\infty}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\mathrm{d}t \,, \qquad \text{(F.21)}
\end{aligned}$$

where the last equality follows since $\lambda_n \to 1$, and $\nabla K^{\delta_n} * f(\lambda_n \boldsymbol{x}) \to \nabla f(\boldsymbol{x})$ uniformly in Ω.

Furthermore, we have that

$$\begin{aligned}
&\bigg| \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t^{\delta_n}(\boldsymbol{x}) \rangle \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&\qquad + \frac{1}{2} \int_0^T \int_{\Omega} \Delta h_t(\boldsymbol{x}) (\rho_t^{\infty}(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \bigg| \\
&\le \bigg| \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t^{\delta_n}(\boldsymbol{x}) \rangle \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&\qquad - \int_0^T \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x}) \rangle K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \bigg| \\
&+ \bigg| \int_0^T \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x}) \rangle K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&\qquad + \frac{1}{2} \int_0^T \int_{\Omega} \Delta h_t(\boldsymbol{x}) (K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \bigg| \\
&+ \bigg| \frac{1}{2} \int_0^T \int_{\Omega} \Delta h_t(\boldsymbol{x}) (K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&\qquad - \frac{1}{2} \int_0^T \int_{\Omega} \Delta h_t(\boldsymbol{x}) (\rho_t^{\infty}(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \bigg| .
\end{aligned} \qquad \text{(F.22)}$$

The second term in the right-hand side of (F.22) is equal to 0 by integration by parts. The third integral in the right-hand side of (F.22) is upper bounded as follows:

$$\begin{aligned}
&\bigg| \frac{1}{2} \int_0^T \int_{\Omega} \Delta h_t(\boldsymbol{x}) (K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t - \frac{1}{2} \int_0^T \int_{\Omega} \Delta h_t(\boldsymbol{x}) (\rho_t^{\infty}(\boldsymbol{x}))^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \bigg| \\
&\le C \int_0^T \int_{\Omega} \Big| (K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}))^2 - (\rho_t^{\infty}(\boldsymbol{x}))^2 \Big| \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&\le C \int_0^T \int_{\Omega} \Big| K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) - \rho_t^{\infty}(\boldsymbol{x}) \Big| \, \Big| K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) + \rho_t^{\infty}(\boldsymbol{x}) \Big| \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \qquad \text{(F.23)} \\
&\le C \left(\int_0^T \int_{\Omega} \Big| K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) - \rho_t^{\infty}(\boldsymbol{x}) \Big|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \right)^{1/2} \\
&\qquad \left(\int_0^T \int_{\Omega} \Big| K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) + \rho_t^{\infty}(\boldsymbol{x}) \Big|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \right)^{1/2} ,
\end{aligned}$$

which converges to 0, as $(K^{\delta_n} * \rho^{\delta_n})_{n\geq 1}$ converges in $\mathscr{L}^2(\Omega \times [0,T])$ to ρ^∞. The first term in the right-hand side of (F.22) is upper bounded as follows:

$$\begin{aligned}
&\Big| \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t^{\delta_n}(\boldsymbol{x})\rangle \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\
&\qquad - \int_0^T \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \Big| \\
&\leq \Big| \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t^{\delta_n}(\boldsymbol{x})\rangle \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\
&\qquad - \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \Big| \\
&+ \Big| \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\
&\qquad - \int_0^T \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \Big|.
\end{aligned} \tag{F.24}$$

The first term is upper bounded using

$$\begin{aligned}
&\Big| \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t^{\delta_n}(\boldsymbol{x})\rangle \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\
&\qquad - \int_0^T \int_{\Omega} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x})\rangle \rho_t^{\delta_n}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \Big| \\
&\leq \|\nabla h^{\delta_n} - \nabla h\|_{\mathscr{L}^\infty(\Omega\times[0,T])} \Big\| \Big| \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n} \Big| \times \rho_t^{\delta_n} \Big\|_{\mathscr{L}^1(\Omega^{\delta_n}\times[0,T])}.
\end{aligned} \tag{F.25}$$

Notice that

$$\begin{aligned}
&\Big\| \Big| \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n} \Big| \times \rho_t^{\delta_n} \Big\|_{\mathscr{L}^1(\Omega^{\delta_n}\times[0,T])} \\
&\overset{(a)}{\leq} \Big\| \Big| \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n} \Big| \times \sqrt{\rho_t^{\delta_n}} \Big\|_{\mathscr{L}^2(\Omega^{\delta_n}\times[0,T])} \Big\| \sqrt{\rho_t^{\delta_n}} \Big\|_{\mathscr{L}^2(\Omega^{\delta_n}\times[0,T])} \\
&\leq \sqrt{T} \Big\| \Big| \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n} \Big| \times \sqrt{\rho_t^{\delta_n}} \Big\|_{\mathscr{L}^2(\Omega^{\delta_n}\times[0,T])},
\end{aligned} \tag{F.26}$$

where (a) follows from an application of Cauchy-Schwartz. By Lemma E.5, we deduce that the right-hand side of (F.26) is bounded uniformly in δ_n. Thus, the first term of (F.24) converges to 0

because of Eq. (F.25). As concerns the second term of (F.24), we have that

$$\begin{aligned}
&\Big| \int_0^T \int_{\Omega^{\delta_n}} \langle \nabla K^{\delta_n} * K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x}) \rangle \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&\qquad - \int_0^T \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}), \nabla h_t(\boldsymbol{x}) \rangle K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \Big| \\
&\leq \Big| \int_0^T \int_{\Omega} \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y}), \nabla h_t(\boldsymbol{x}) \rangle K^{\delta_n}(\boldsymbol{x}-\boldsymbol{y}) \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&\qquad - \int_0^T \int_{\Omega} \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y}), \nabla h_t(\boldsymbol{y}) \rangle K^{\delta_n}(\boldsymbol{x}-\boldsymbol{y}) \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \Big| \\
&= \Big| \int_0^T \int_{\Omega} \int_{\Omega} \langle \nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y}), \nabla h_t(\boldsymbol{x}) - \nabla h_t(\boldsymbol{y}) \rangle K^{\delta_n}(\boldsymbol{x}-\boldsymbol{y}) \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \Big| \\
&\leq C \int_0^T \int_{\Omega} \int_{\Omega} |\boldsymbol{x}-\boldsymbol{y}| \, |\nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y})| K^{\delta_n}(\boldsymbol{x}-\boldsymbol{y}) \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t.
\end{aligned} \tag{F.27}$$

Recall that $\rho_t^{\delta_n}$ is supported on $\Omega^{\delta_n} \subseteq \Omega$, and Ω is bounded. In addition, since the kernel K has bounded support, the diameter of the support of K^{δ_n} is at most δ_n times a constant. Consequently, the last term in the right-hand side of (F.27) is upper bounded by

$$\begin{aligned}
&\delta_n C_1 \int_0^T \int_{\Omega} \int_{\Omega} |\nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y})| K^{\delta_n}(\boldsymbol{x}-\boldsymbol{y}) \rho_t^{\delta_n}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\
&= \delta_n C_1 \int_0^T \int_{\Omega} |\nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y})| K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}t \\
&\leq \delta_n C_1 \left(\int_0^T \int_{\Omega} |\nabla K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y})|^2 \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}t \right)^{1/2} \\
&\qquad \left(\int_0^T \int_{\Omega} (K^{\delta_n} * \rho_t^{\delta_n}(\boldsymbol{y}))^2 \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}t \right)^{1/2}.
\end{aligned} \tag{F.28}$$

By using that $K^{\delta_n} * \rho^{\delta_n} \in \mathscr{L}^2(\Omega \times [0,T])$ and the result of Lemma E.5, we have that the two last integrals are bounded uniformly in δ. As a result, the right-hand side of (F.28) converges to 0, which implies that the right-hand side of (F.22) also converges to 0. By putting this fact together with (F.18) and (F.21), the desired result follows. □

We have now proved that $(\rho^{\delta_n})_{n\geq 1}$ converges to *a* weak solution of the limit PDE (A.1). In order to prove the uniqueness of the weak solutions of the limit PDE, we next prove a bound on $\|\rho_t^{\delta_n}\|_{\mathscr{L}^4(\Omega)}$, which along with Lemma A.2 proves the uniqueness claim.

Lemma F.5 (Uniform bound in $\mathscr{L}^4$)**.** *Assume that $\rho_{\text{init}}, f \in \mathscr{C}^\infty(\Omega)$ and consider the sequence $(\rho^{\delta_n})_{n\geq 1}$. Then,*

$$\sup_{n\geq 1, t\in[0,T_0]} \|\rho_t^{\delta_n}\|_{\mathscr{L}^4(\Omega)} \leq C(\Omega)(1+T)\,, \tag{F.29}$$

where

$$T_0 = \frac{1}{C(\Omega)(1+T)}, \tag{F.30}$$

for some bounded constant $0 < C(\Omega) < \infty$.

Proof. For simplicity, we indicate the norms $\mathscr{L}^p(\Omega)$ by $\|\cdot\|_p$. For a function $g \in \mathscr{C}^m(\Omega)$, we let $\nabla^{\otimes m} g$ be the vector with coordinates $\partial^m g/(\partial_{i_1} \dots \partial_{i_m})$, with $1 \le i_1, i_2, \dots, i_m \le d$. The proof strategy to prove this lemma is to first bound $\|\nabla^{\otimes m} \rho_t^{\delta_n}\|_2$, for some $m \ge d/4$, and then apply the Gagliardo-Nirenberg interpolation inequality (cf. Lemma H.3) to bound $\|\rho_t^{\delta_n}\|_4$. Throughout this proof, we will use C, C_k and so on to denote constants that can depend on the domain Ω, but do not depend on t or δ.

Before proceeding, we need to establish some notations and definitions.

For a function g and an integer $k \ge 0$, we denote its Sobolev norms by

$$\|g\|_{(k)} = \left(\sum_{m=0}^{k} \|\nabla^{\otimes m} g\|_2^2 \right)^{1/2} . \tag{F.31}$$

We will use the following relations on Sobolev norms (see [Oel01, Equation (1.14)]):

$$\|g\|_{(k)} \le \begin{cases} \bar{C}_k (\|g\|_2^2 + \|(-\Delta)^{k/2} g\|_2^2)^{1/2} & \text{if } k \text{ is even}, \\ \bar{C}_k (\|g\|_2^2 + \|\nabla(-\Delta)^{(k-1)/2} g\|_2^2)^{1/2} & \text{if } k \text{ is odd}. \end{cases} \tag{F.32}$$

Instead of bounding $\|\nabla^{\otimes m} \rho_t^{\delta_n}\|_2$, we will bound the dominating quantity $\|\rho_t^{\delta_n}\|_{(m)}$. To this end, we follow a similar strategy as in [Oel01]. Namely, we derive descriptions of the evolution of $\|(-\Delta)^m \rho_t^{\delta_n}\|_2$ and $\|(-\Delta)^m (\rho_t^{\delta_n} * K^{\delta_n} - f)\|_2$. More precisely, we derive a recursive equation (on m) for the evolution of a suitably chosen linear combination of these two quantities.

Since ρ^{δ_n} is a solution of the PDE (B.2), we have

$$\partial_t (-\Delta)^m \rho_t^{\delta_n}(\boldsymbol{x}) = -\tau (-\Delta)^{m+1} \rho_t^{\delta_n}(\boldsymbol{x}) + (-\Delta)^m \nabla \cdot \left(\rho_t^{\delta_n}(\boldsymbol{x}) \nabla (V + U * \rho_t^{\delta_n}) \right) \tag{F.33}$$

Following along the same lines as in derivation of [Oel01, Equation (3.12)], we obtain

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} \|(-\Delta)^m \rho_t^{\delta_n}\|_2^2 \\ &\quad \le (C C_m - 2\tau) \|\nabla (-\Delta)^m \rho_t^{\delta_n}\|_2^2 \\ &\quad + \frac{2}{C} \|\rho_t^{\delta_n}\|_\infty \left\langle \nabla (-\Delta)^m (V + U * \rho_t^{\delta_n}), \rho_t^{\delta_n} \nabla (-\Delta)^m (V + U * \rho_t^{\delta_n}) \right\rangle \\ &\quad + \frac{\tilde{C}_m}{C} \sum_{q=1}^{m} \left(\left\| V + U * \rho_t^{\delta_n} \right\|_{(2m+1-q)}^2 \|\rho_t^{\delta_n}\|_{(q+1+d/2)}^2 \right. \\ &\quad \left. + \left\| V + U * \rho_t^{\delta_n} \right\|_{(q+1+d/2)}^2 \|\rho_t^{\delta_n}\|_{(2m+1-q)}^2 \right), \end{aligned} \tag{F.34}$$

where C_m and $\tilde{C}_m$ are positive constants that depend on m and $C > 0$ is a constant which can be chosen arbitrarily.

We set $m = \lceil 1 + d/2 \rceil$ for which we can upper bound the right-hand side of (F.34) as

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left\|(-\Delta)^m \rho_t^{\delta_n}\right\|_2^2 \\
&\quad \le (CC_m - 2\tau)\left\|\nabla(-\Delta)^m \rho_t^{\delta_n}\right\|_2^2 \\
&\quad + \frac{2}{C}\|\rho_t^{\delta_n}\|_\infty \left\langle \nabla(-\Delta)^m (V + U * \rho_t), \rho_t^{\delta_n} \nabla(-\Delta)^m (V + U * \rho_t^{\delta_n}) \right\rangle \\
&\quad + \frac{2m\tilde{C}_m}{C}\left\|V + U * \rho_t^{\delta_n}\right\|_{(2m)}^2 \|\rho_t^{\delta_n}\|_{(2m)}^2 .
\end{aligned} \tag{F.35}$$

We next move to the next quantity. Write

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left\|(-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n} - f)\right\|_2^2 \\
&\le 2\left\langle (-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n} - f), \partial_t (-\Delta)^m \rho_t^{\delta_n} * K^{\delta_n} \right\rangle \\
&= 2\left\langle (-\Delta)^m (V^{\delta_n} + U^{\delta_n} * \rho_t^{\delta_n}), \partial_t (-\Delta)^m \rho_t^{\delta_n} \right\rangle \\
&= -2\tau\left\langle (-\Delta)^m (V^{\delta_n} + U^{\delta_n} * \rho_t^{\delta_n}), (-\Delta)^{m+1} \rho_t^{\delta_n} \right\rangle \\
&+ 2\left\langle (-\Delta)^m (V^{\delta_n} + U^{\delta_n} * \rho_t^{\delta_n}), (-\Delta)^m \nabla \cdot \left(\rho_t^{\delta_n}(\boldsymbol{x}) \nabla (V^{\delta_n} + U^{\delta_n} * \rho_t^{\delta_n}) \right) \right\rangle ,
\end{aligned} \tag{F.36}$$

where the last step follows from (F.33). Note that the first term on the right-hand side can be bounded as

$$\begin{aligned}
&- 2\tau\left\langle (-\Delta)^m (V^{\delta_n} + U^{\delta_n} * \rho_t^{\delta_n}), (-\Delta)^{m+1} \rho_t^{\delta_n} \right\rangle \\
&= -2\tau\left\langle (-\Delta)^{m+1} (K^{\delta_n} * f), (-\Delta)^m \rho_t^{\delta_n} \right\rangle - 2\tau\left\|\nabla(-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n})\right\|_2^2 \\
&\le 2\tau \|(-\Delta)^{m+1} f\|_2 \|(-\Delta)^m \rho_t^{\delta_n}\|_2 - 2\tau\left\|\nabla(-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n})\right\|_2^2 ,
\end{aligned} \tag{F.37}$$

where the last step follows from Young's convolution inequality and the fact that $\|K^{\delta_n}\|_1 = 1$.

The second term in (F.36) can be bounded following the same lines as in derivation of [Oel01, Equations (3.3) and (3.16)], which along with (F.37) gives

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left\|(-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n} - f)\right\|_2^2 \\
&\quad \le CC_m \|\nabla(-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n} - f)\|_2^2 + 2\tau \|(-\Delta)^{m+1} f\|_2 \|(-\Delta)^m \rho_t^{\delta_n}\|_2 \\
&\qquad\qquad - 2\tau\left\|\nabla(-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n})\right\|_2^2 \\
&\quad - 2\left\langle \nabla(-\Delta)^m (V + U * \rho_t^{\delta_n}), \rho_t^{\delta_n} \nabla(-\Delta)^m (V + U * \rho_t^{\delta_n}) \right\rangle \\
&\quad + \frac{\tilde{C}_m}{C}\sum_{q=1}^{m} \left(\left\|V + U * \rho_t^{\delta_n}\right\|_{(2m+1-q)}^2 \|\rho_t^{\delta_n}\|_{(q+1+d/2)}^2 \right. \\
&\quad \left. + \left\|V + U * \rho_t^{\delta_n}\right\|_{(q+1+d/2)}^2 \|\rho_t^{\delta_n}\|_{(2m+1-q)}^2 \right) .
\end{aligned} \tag{F.38}$$

Since $f \in \mathscr{C}^\infty(\Omega)$, there exists constant $M > 0$, such that $\|(-\Delta)^{m+1} f\|_2 \le M$, $\|\nabla(-\Delta)^m f\|_2 \le M$. Using the particular choice of m, we can upper bound the right-hand side of (F.38) as

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left\|(-\Delta)^m(K^{\delta_n} * \rho_t^{\delta_n} - f)\right\|_2^2 \\
&\quad \le (2CC_m - 2\tau)\|\nabla(-\Delta)^m(K^{\delta_n} * \rho_t^{\delta_n})\|_2^2 + 2\tau M\|(-\Delta)^m \rho_t^{\delta_n}\|_2 + 2M^2 CC_m \\
&\quad - 2\Big\langle \nabla(-\Delta)^m(V + U * \rho_t^{\delta_n}), \rho_t^{\delta_n} \nabla(-\Delta)^m(V + U * \rho_t^{\delta_n})\Big\rangle \\
&\quad + \frac{2m\tilde{C}_m}{C}\left\|V + U * \rho_t^{\delta_n}\right\|_{(2m)}^2 \|\rho_t^{\delta_n}\|_{(2m)}^2 . \qquad \text{(F.39)}
\end{aligned}$$

Define $C_1 \equiv 2\|\rho_{\mathrm{init}}\|_{(2m)}$ and let

$$T_n \equiv \inf\left\{t \in [0,T] : \|\rho_t^{\delta_n}\|_{(2m)} > C_1\right\} \wedge T ,$$

for $n \ge 1$. Clearly, $T_n > 0$ by choice of C_1. In addition, by applying Sobolev's inequality (see e.g. [Oel01, Equation (1.12)]), we have

$$\|\rho_t^{\delta_n}\|_\infty \le C_2\|\rho_t^{\delta_n}\|_{(2m)} \le C_1 C_2 , \quad \text{for } t \in [0, T_n],\ n \ge 1 .$$

where $C_2 > 0$ is a constant depending on d. We let $C_* \equiv C_1 C_2 / C$. Recall that the constant $C > 0$ in (F.35) and (F.39) was arbitrary. We choose it in a way that $C < \tau/(2C_m)$. We then consider the evolution of the following linear combination of the two quantities we analyzed above. Note that by Equations (F.35) and (F.39), we have for $t \in [0, T_n]$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left(\left\|(-\Delta)^m \rho_t^{\delta_n}\right\|_2^2 + C_*\left\|(-\Delta)^m(K^{\delta_n} * \rho_t^{\delta_n} - f)\right\|_2^2\right) \\
&\quad \le -\frac{3\tau}{2}\left\|\nabla(-\Delta)^m \rho_t^{\delta_n}\right\|_2^2 - C_*\tau\|\nabla(-\Delta)^m(K^{\delta_n} * \rho_t^{\delta_n})\|_2^2 \\
&\qquad + C_*\left(2\tau M\|(-\Delta)^m \rho_t^{\delta_n}\|_2 + \tau M^2\right) \\
&\qquad + \frac{2m\tilde{C}_m}{C}(1 + C_*)\left\|V + U * \rho_t^{\delta_n}\right\|_{(2m)}^2 \|\rho_t^{\delta_n}\|_{(2m)}^2 \\
&\quad \le C_*\left(2\tau M\|(-\Delta)^m \rho_t^{\delta_n}\|_2 + \tau M^2\right) \\
&\qquad + \frac{2m\tilde{C}_m}{C}(1 + C_*)\left(\|\rho_t^{\delta_n}\|_{(2m)}^2 + \left\|V + U * \rho_t^{\delta_n}\right\|_{(2m)}^2\right)^2 \\
&\quad \overset{(a)}{\le} \tau M^2 C_* + C_3\left(\|\rho_t^{\delta_n}\|_{(2m)}^2 + C_*\left\|V + U * \rho_t^{\delta_n}\right\|_{(2m)}^2\right)^2 \\
&\quad \overset{(b)}{\le} \tau M^2 C_* + C_3\left(\|\rho_t^{\delta_n}\|_{(2m)}^2 + C_*\left\| - f + K^{\delta_n} * \rho_t^{\delta_n}\right\|_{(2m)}^2\right)^2 , \qquad \text{(F.40)}
\end{aligned}$$

where in (a) we use the fact that $\|(-\Delta)^m \rho_t^{\delta_n}\|_2 \le \|\rho_t^{\delta_n}\|_{(2m)}$, which follows immediately from (F.31); (b) follows from the fact that for any function $g \in \mathscr{L}^2(\Omega)$, $\|g * K^{\delta_n}\|_2 \le \|K^{\delta_n}\|_1 \|g\|_2 = \|g\|_2$, by Young's inequality for convolution.

Another observation that will be used later is that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\|\rho_t^{\delta_n}\|_2^2 + C_*\|K^{\delta_n} * \rho_t^{\delta_n} - f\|_2^2\right) \le 0 . \qquad \text{(F.41)}$$

This claim follows by repeating the same argument we had to derive (F.40), for $m = 0$. In this case, we have analogous equations to (F.35) and (F.39), where only the first two terms appear.

Next note that by (F.32), we have for $t \in [0, T_n]$,

$$\begin{aligned}
&\|\rho_t^{\delta_n}\|_{(2m)}^2 + C_*\|K^{\delta_n} * \rho_t^{\delta_n} - f\|_{(2m)}^2 \\
&\leq \bar{C}_m \Bigg(\|\rho_t^{\delta_n}\|_2^2 + \|(-\Delta)^m \rho_t^{\delta_n}\|_2^2 \\
&\qquad + C_* \left(\|K^{\delta_n} * \rho_t^{\delta_n} - f\|_2^2 + \left\|(-\Delta)^m (K^{\delta_n} * \rho_t^{\delta_n} - f)\right\|_2^2 \right) \Bigg) \\
&\leq \bar{C}_m \left(\|\rho_{\text{init}}\|_{(2m)}^2 + C_*\|K^{\delta_n} * \rho_{\text{init}} - f\|_{(2m)}^2 \right) \\
&+ \bar{C}_m \tau M^2 C_* t + \bar{C}_m C_3 \int_0^t \left(\|\rho_s^{\delta_n}\|_{(2m)}^2 + C_* \left\|K^{\delta_n} * \rho_s^{\delta_n} - f\right\|_{(2m)}^2 \right)^2 \mathrm{d}s \,,
\end{aligned} \tag{F.42}$$

where the last step is a result of (F.41) and (F.40). Let us stress that $\bar{C}_m$, C_*, C_3 are constants that are independent of n.

We further note that

$$\begin{aligned}
\|\rho_{\text{init}}\|_{(2m)}^2 + C_*\|K^{\delta_n} * \rho_{\text{init}} - f\|_{(2m)}^2 &\leq \|\rho_{\text{init}}\|_{(2m)}^2 (1 + 2C_*) + 2C_*\|f\|_{(2m)}^2 \\
&\leq (1 + 2C_*)\frac{C_1^2}{4} + 2C_*\|f\|_{(2m)}^2 \,,
\end{aligned} \tag{F.43}$$

for $n \geq 1$. Here, the first step is a result of triangle inequality and the Young's inequality for convolution along with the fact that $\|K^{\delta_n}\|_1 = 1$. The second step follows from definition of C_1. Since $f \in \mathscr{C}^\infty(\Omega)$, $\|f\|_{(2m)}^2$ is uniformly bounded over Ω. We denote the right-hand side of (F.43) by the constant C_4. Using bound (F.43) into (F.42) results in

$$\begin{aligned}
&\|\rho_t^{\delta_n}\|_{(2m)}^2 + C_*\|K^{\delta_n} * \rho_t^{\delta_n} - f\|_{(2m)}^2 \\
&\leq \bar{C}_m C_4 + \bar{C}_m \tau M^2 C_* t \\
&\qquad + \bar{C}_m C_3 \int_0^t \left(\|\rho_s^{\delta_n}\|_{(2m)}^2 + C_* \left\|K^{\delta_n} * \rho_s^{\delta_n} - f\right\|_{(2m)}^2 \right)^2 \mathrm{d}s \,,
\end{aligned} \tag{F.44}$$

for $t \in [0, T_n]$. By employing a generalization of Gronwall's inequality (cf. Lemma H.2 and Remark H.1) we get

$$\|\rho_t^{\delta_n}\|_{(2m)}^2 + C_*\|K^{\delta_n} * \rho_t^{\delta_n} - f\|_{(2m)}^2 \leq \frac{\bar{C}_m C_4 + \bar{C}_m \tau M^2 C_* T_n}{1 - (\bar{C}_m C_4 + \bar{C}_m \tau M^2 C_* T_n)\bar{C}_m C_3 t} \,. \tag{F.45}$$

Therefore, for $t \in [0, T_0]$, with

$$T_0 = \frac{1}{2\bar{C}_m C_4 + 2\bar{C}_m \tau M^2 C_* T}, \tag{F.46}$$

we have that

$$\|\rho_t^{\delta_n}\|_{(2m)}^2 + C_*\|K^{\delta_n} * \rho_t^{\delta_n} - f\|_{(2m)}^2 \leq C_5 + C_6 T \,, \tag{F.47}$$

with $C_5 \equiv \bar{C}_m C_4$ and $C_6 \equiv \bar{C}_m \tau M^2 C_*$. Note that C_5, C_6 and T_0 are independent of n, but depend on d. Let $m_0 = \lceil d/4 \rceil$. Then, by the choice of $m = 1 + \lceil d/2 \rceil$ we have $\|\nabla^{\otimes m_0} \rho_t^{\delta_n}\|_2 \leq \|\rho_t^{\delta_n}\|_{(2m)}$, and hence as a result of (F.47), we obtain

$$\sup_{n \geq 1, t \in [0,T_0]} \|\nabla^{\otimes m_0} \rho_t^{\delta_n}\|_2 \leq C_5 + C_6 T\,, \tag{F.48}$$

Finally, by applying Gagliardo-Nirenberg interpolation inequality (cf. Lemma H.3) we get

$$\sup_{n \geq 1, t \in [0,T_0]} \|\rho_t^{\delta_n}\|_4 \leq C_7 + C_8 T\,,$$

for some constant $C_7, C_8 > 0$, which completes the proof. □

Lemma F.6 (Convergence to the unique weak solution of limit PDE)**.** *Let ρ^∞ be the limit in $\mathscr{C}([0,T], \mathscr{P}_2(\Omega))$ of the converging sequence $(\rho^{\delta_n})_{n \geq 1}$. Then, ρ^∞ is the unique weak solution of the PDE (A.1) in $\mathscr{L}^4(\Omega \times [0,T])$ with initial and boundary conditions (A.2).*

Proof. From Lemma F.3, we have that the sequence $(K^{\delta_n} * \rho^{\delta_n})_{n \geq 1}$ converges in $\mathscr{L}^2(\Omega \times [0,T])$ to ρ^∞. Furthermore, by Lemma F.5, $\|\rho_t^\delta\|_{\mathscr{L}^4(\Omega)} \leq C(1+T)$ for any $t \in [0, T_0]$, where C is a universal constant. By using Young's convolution inequality, we also deduce that $\|K^{\delta_n} * \rho_t^\delta\|_{\mathscr{L}^4(\Omega)} \leq C(1+T)$ for any $t \in [0, T_0]$.

Note that $\mathscr{L}^4(\Omega)$ is a reflexive Banach space. Thus, by applying the Banach-Alaoglu theorem, every bounded sequence in $\mathscr{L}^4(\Omega)$ has a weakly convergent subsequence. This means that there exist a subsequence $K^{\delta_{n_k}} * \rho^{\delta_{n_k}}$ and a function $\tilde{\rho} \in \mathscr{L}^4(\Omega)$ such that, for any $g \in \mathscr{L}^{4/3}(\Omega)$, we have

$$\int_\Omega (K^{\delta_{n_k}} * \rho^{\delta_{n_k}}(\boldsymbol{x}) - \tilde{\rho}(\boldsymbol{x})) g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \to 0. \tag{F.49}$$

Now, since Ω is bounded, $K^{\delta_{n_k}} * \rho^{\delta_{n_k}}$ and $\tilde{\rho}$ are also in $\mathscr{L}^2(\Omega)$ (as they are in $\mathscr{L}^4(\Omega)$). Thus, $K^{\delta_{n_k}} * \rho^{\delta_{n_k}} - \tilde{\rho}$ is in $\mathscr{L}^2(\Omega)$, hence it is also in $\mathscr{L}^{4/3}(\Omega)$. As a result, we can pick $g = K^{\delta_{n_k}} * \rho^{\delta_{n_k}} - \tilde{\rho}$ and obtain

$$\int_\Omega |K^{\delta_{n_k}} * \rho^{\delta_{n_k}}(\boldsymbol{x}) - \tilde{\rho}(\boldsymbol{x})|^2\, \mathrm{d}\boldsymbol{x} \to 0. \tag{F.50}$$

Therefore, $\tilde{\rho}$ is the limit in $\mathscr{L}^2(\Omega)$ of the sequence $K^{\delta_{n_k}} * \rho^{\delta_{n_k}}$. By uniqueness of the limit, we conclude that $\tilde{\rho} = \rho^\infty$. As a result, $\rho^\infty \in \mathscr{L}^4(\Omega)$ for any $t \in [0, T_0]$, which implies that $\rho^\infty \in \mathscr{L}^4(\Omega \times [0, T_0])$. Thus, by Lemma F.4 and Lemma A.2, ρ^∞ is the unique weak solution of the PDE (A.1) for $t \in [0, T_0]$. Note that T_0 is decreasing with T. Thus, we can repeat the same argument with $T - T_0$ instead of T and obtain that ρ^∞ is the unique weak solution of the PDE (A.1) for $t \in [T_0, 2\,T_0]$. By iterating this procedure T/T_0 times, the result follows. □

At this point, we state and prove a lemma showing that the sequence $(\rho_t^{\delta_n})_{n \geq 1}$ converges in $\mathscr{L}^2(\Omega)$ to ρ_t^∞.

Lemma F.7. *For almost all $t \in [0,T]$, the measure ρ_t^∞ is the limit in $\mathscr{L}^2(\Omega)$ of the sequence $(\rho_t^{\delta_n})_{n \geq 1}$.*

Proof. The proof is similar to that of Lemma F.3. Suppose that $t \in [0, T_0]$, where T_0 is defined in the statement of Lemma F.5. Note that, for any $n \geq 1$, $\rho^{\delta_n} \in \mathscr{L}^2(\Omega)$. Let us show that $(\rho^{\delta_n})_{n\geq 1}$ is a Cauchy sequence in $\mathscr{L}^2(\Omega)$.

As $\rho_t^{\delta_n} \in \mathscr{L}^2(\Omega)$ for every $t \in [0, T_0]$, its Fourier transform exists and we denote it by $\widehat{\rho^{\delta_n}}$. Hence, by applying Parseval's theorem, we have

$$\begin{aligned}\limsup_{n,n'\to\infty} \|\rho^{\delta_n} - \rho^{\delta_{n'}}\|^2_{\mathscr{L}^2(\Omega)} &= \limsup_{n,n'\to\infty} \|\rho_t^{\delta_n} - \rho_t^{\delta_{n'}}\|^2_{\mathscr{L}^2(\Omega)} \\ &= \limsup_{n,n'\to\infty} \int_{\mathbb{R}^d} |\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{\rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}.\end{aligned} \tag{F.51}$$

Fix $\Lambda > 1$ and decompose the integral in the right-hand side of (F.51) as

$$\limsup_{n,n'\to\infty} \int_{|\boldsymbol{\lambda}|<\Lambda} |\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{\rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda} + \limsup_{n,n'\to\infty} \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{\rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda}. \tag{F.52}$$

Consider the first term of (F.52). By Lemma F.1, and since by Jensen's inequality $W_1(\rho_1, \rho_2) \leq W_2(\rho_1, \rho_2)$ for any two distributions ρ_1, ρ_2, we have $W_1(\rho_t^{\delta_n} - \rho_t^{\delta_{n'}}) \to 0$, as $n, n' \to \infty$. Since for the complex exponential functions $\|e^{i\langle\boldsymbol{\lambda},\boldsymbol{x}\rangle}\|_{\mathrm{Lip}} \leq |\boldsymbol{\lambda}|$, by definition of 1-Wasserstein distance, the integrand in the first term converges pointwise to 0. Furthermore, the integrand is upper bounded by an integrable function, since $|\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda})| \leq \|\rho_t^{\delta_n}\|_{\mathscr{L}^2(\Omega)} \leq C$ for all n and every $t \in [0, T_0]$. Hence, by dominated convergence, the first integral in (F.52) converges to 0.

As for the second term of (F.52), the following chain of inequalities holds:

$$\begin{aligned}&\limsup_{n,n'\to\infty} \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda}) - \widehat{\rho_t^{\delta_{n'}}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda} \\ &\qquad \leq 4 \sup_{n\geq 1} \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda} \\ &\qquad \leq \frac{4}{\Lambda^2} \sup_{n\geq 1} \int_{|\boldsymbol{\lambda}|\geq\Lambda} |\boldsymbol{\lambda}|^2 \, |\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda} \\ &\qquad \leq \frac{4}{\Lambda^2} \sup_{n\geq 1} \int_{\mathbb{R}^d} |\boldsymbol{\lambda}|^2 \, |\widehat{\rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda} \\ &\qquad = \frac{4}{\Lambda^2} \sup_{n\geq 1} \int_{\mathbb{R}^d} |\widehat{\nabla\rho_t^{\delta_n}}(\boldsymbol{\lambda})|^2 \,\mathrm{d}\boldsymbol{\lambda} \\ &\qquad = \frac{4}{\Lambda^2} \sup_{n\geq 1} \int_{\Omega} |\nabla\rho_t^{\delta_n}(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x},\end{aligned} \tag{F.53}$$

where in the last equality we have applied again Parseval's theorem. In the proof of Lemma F.5, we provide an upper bound, which does not depend on n, on the Sobolev norm of $\rho_t^{\delta_n}$ (see (F.47)). Thus, as $\Lambda \to \infty$, the second term of (F.52) converges to 0.

By iterating the argument T/T_0 times, we obtain that $(\rho_t^{\delta_n})_{n\geq 1}$ is a Cauchy sequence in $\mathscr{L}^2(\Omega)$ for $t \in [0, T]$. Let $\tilde{\rho}_t^{\infty} \in \mathscr{L}^2(\Omega)$ be its limit. Furthermore, by Lemma F.1, $(\rho^{\delta_n})_{n\geq 1}$ has limit ρ^{∞} in $\mathscr{C}([0, T], \mathscr{P}_2(\Omega))$. Therefore, the measure ρ_t^{∞} has for almost every $t \in [0, T]$ the density $\tilde{\rho}_t^{\infty} \in \mathscr{L}^2(\Omega)$, and the proof is complete. □

Theorem 5.2 follows from Lemma A.2, Lemma F.6 and Lemma F.7.

Let us define the free energy associated to the PDE (A.1) as

$$\begin{aligned} F(\rho) &= \frac{1}{2} R(\rho) - \tau S(\rho) \\ &= \frac{\nu_0}{2} \|f - \rho\|^2_{\mathscr{L}^2(\Omega)} + \tau \int \rho(\boldsymbol{x}, t) \log \rho(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} \,. \end{aligned} \tag{F.54}$$

As explained in Section 3.5, this limit free energy is displacement convex, and hence its W_2 gradient flow converges to the unique minimizer of (F.54). These facts are stated and proved formally in the theorem that follows.

Theorem F.8. *Assume that the initial condition $\rho^\infty(0) \in \mathscr{C}^\infty(\Omega)$. Then, the following results hold:*

1. *There exists a unique minimizer in $\mathscr{P}_2(\Omega)$, call it ρ^*, of the free energy F defined in (F.54).*
2. *For any $t \geq 0$, we have*

$$F(\rho^\infty(t)) - F(\rho^*) \leq (F(\rho^\infty(0)) - F(\rho^*)) e^{-2\alpha t}, \tag{F.55}$$

   where α is defined in (3.1).
3. *For any $n \geq 1$ and for almost any $t \geq 0$, we have*

$$F(\rho^\delta(t)) - F(\rho^*) \leq (F(\rho^\infty(0)) - F(\rho^*)) e^{-2\alpha t} + \Delta(\delta, T, d), \tag{F.56}$$

   where α is defined in (3.1) and $\Delta(\delta, T, d) \to 0$ as $\delta \to 0$.

Proof. The proof follows from the results of [CJM+01]. The technical assumptions required by [CJM+01] are satisfied by the PDE (A.1), since Ω is convex and bounded, the initial condition $\rho^\infty(0) \in \mathscr{L}^\infty(\Omega)$, and f satisfies the assumptions (A2) and (A3). Note also that the condition $\inf_\Omega V = 0$ coming from assumption (HV3) of [CJM+01] can be relaxed. In fact, adding a constant to V does not change the entropy functional in [CJM+01, Eq. (3)] (which corresponds to the free energy (F.54)) and the PDE in [CJM+01, Eq. (46)] (which corresponds to the PDE (A.1)).

The uniqueness of the minimizer ρ^* follows from [CJM+01, Lemma 6], which proves the first result. Since ρ^∞ is the unique weak solution of the PDE (A.1) with initial and boundary conditions (A.2), then it coincides with the unique, non-negative mass-preserving solution of [CJM+01, Theorem 16]. Thus, the inequality (F.55) readily follows from [CJM+01, Theorem 16].

It remains to prove inequality (F.56). By definition of free energy, we obtain

$$F(\rho^\delta(t)) - F(\rho^\infty(t)) = \frac{1}{2}(R(\rho^\delta(t)) - R(\rho^\infty(t))) - \tau(S(\rho^\delta(t)) - S(\rho^\infty(t))). \tag{F.57}$$

Recall that, by Lemma F.7, $\rho^\delta(t)$ converges to $\rho^\infty(t)$ in $\mathscr{L}^2(\Omega)$. Consequently, by using the triangle inequality, we have that the term $R(\rho^\delta(t)) - R(\rho^\infty(t))$ tends to 0 as $\delta \to 0$.

In order to complete the proof, it remains to show that $S(\rho^\delta(t)) - S(\rho^\infty(t))$ tends to 0 as $\delta \to 0$. To do so, define

$$\begin{aligned} A &= \{\boldsymbol{x} \in \Omega \ : \ |\rho^\delta(\boldsymbol{x}, t) - \rho^\infty(\boldsymbol{x}, t)| > 1/4\}, \\ B &= \{\boldsymbol{x} \in \Omega \ : \ \rho^\delta(\boldsymbol{x}, t) \in [0, 1/2], \, \rho^\infty(\boldsymbol{x}, t) \in [0, 1/2]\}, \\ C &= \{\boldsymbol{x} \in \Omega \ : \ \rho^\delta(\boldsymbol{x}, t) > 1/4, \, \rho^\infty(\boldsymbol{x}, t) > 1/4\}. \end{aligned} \tag{F.58}$$

Note that $A \cup B \cup C = \Omega$. In fact, suppose that $\boldsymbol{x} \notin B$ and $\boldsymbol{x} \notin C$. Then, one between $\rho^\delta(\boldsymbol{x},t)$ and $\rho^\infty(\boldsymbol{x},t)$ is $\in [0,1/4]$ and the other is $> 1/2$. Consequently, $|\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)| > 1/4$ and $\boldsymbol{x} \in A$. This immediately implies that

$$\begin{aligned} |S(\rho^\delta(t)) - S(\rho^\infty(t))| \leq & \left| \int_A \left(\rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \right) \mathrm{d}\boldsymbol{x} \right| \\ & + \left| \int_B \left(\rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \right) \mathrm{d}\boldsymbol{x} \right| \\ & + \left| \int_C \left(\rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \right) \mathrm{d}\boldsymbol{x} \right|. \end{aligned} \tag{F.59}$$

We will now upper bound the three integrals in the RHS of (F.59). As for the first term, note that

$$\int_\Omega |\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)|^2 \, \mathrm{d}\boldsymbol{x} \geq \int_A |\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)|^2 \, \mathrm{d}\boldsymbol{x} \geq \frac{|A|}{16}, \tag{F.60}$$

where $|A|$ denotes the volume of A. Furthermore,

$$\begin{aligned} & \left| \int_A \left(\rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \right) \mathrm{d}\boldsymbol{x} \right| \\ & \quad \leq \left| \int_A \rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) \, \mathrm{d}\boldsymbol{x} \right| + \left| \int_A \rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \, \mathrm{d}\boldsymbol{x} \right| \\ & \quad \leq |A|^{1/2} \left(\int_A \left(\rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) \right)^2 \mathrm{d}\boldsymbol{x} \right)^{1/2} \\ & \quad + |A|^{1/2} \left(\int_A \left(\rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \right)^2 \mathrm{d}\boldsymbol{x} \right)^{1/2}. \end{aligned} \tag{F.61}$$

Note that $|t \log t| \leq 1$ for $t \in [0,1]$ and $|\log t| \leq t$ for $t \geq 1$. Thus, the RHS of (F.61) is upper bounded by

$$|A|^{1/2} \left(|A| + \|\rho^\delta(t)\|_{\mathscr{L}^4(\Omega)} \right)^{1/2} + |A|^{1/2} \left(|A| + \|\rho^\infty(t)\|_{\mathscr{L}^4(\Omega)} \right)^{1/2}. \tag{F.62}$$

By Lemma F.7, for almost all $t \in [0,T]$, $\rho^\delta(t)$ converges to $\rho^\infty(t)$ in $\mathscr{L}^2(\Omega)$. Thus, by (F.60), $|A|$ tends to 0 as $\delta \to 0$. By Lemma F.6, $\rho^\infty(t) \in \mathscr{L}^4(\Omega)$ for almost all $t \in [0,T]$. Furthermore, by Lemma F.5, the quantity $\|\rho^\delta(t)\|_{\mathscr{L}^4(\Omega)}$ has a δ-free upper bound for $t \in [0,T_0]$. As a result, for almost all $t \in [0,T_0]$, the first integral in (F.59) tends to 0 as $\delta \to 0$. By iterating this argument T/T_0 times, we conclude that for almost all $t \in [0,T_0]$, the first integral in (F.59) tends to 0 as $\delta \to 0$.

In order to bound the second integral in (F.59), we write

$$\begin{aligned} & \left| \int_B \left(\rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \right) \mathrm{d}\boldsymbol{x} \right| \\ & \quad \leq \int_B \left| \rho^\delta(\boldsymbol{x},t) \log \rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t) \log \rho^\infty(\boldsymbol{x},t) \right| \mathrm{d}\boldsymbol{x} \\ & \quad \leq \int_B \left| \rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t) \right| \log \frac{1}{|\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)|} \, \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{F.63}$$

where in the last inequality we have applied [CT06, Theorem 17.3.3], since $\rho^\delta(\boldsymbol{x},t)$, $\rho^\infty(\boldsymbol{x},t) \in [0,1/2]$ by definition of B. Note that

$$|\log t| \le \max\left(2\sqrt{t}, \frac{1}{t}\right) \le 2\sqrt{t} + \frac{1}{t}.$$

Thus, the RHS of (F.63) is upper bounded by

$$\begin{aligned} &\int_B \left|\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)\right|^2 \,\mathrm{d}\boldsymbol{x} + 2\int_B \left|\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)\right|^{1/2} \,\mathrm{d}\boldsymbol{x} \\ &\quad \|\rho^\delta(t) - \rho^\infty(t)\|^2_{\mathscr{L}^2(\Omega)} + 2|\Omega|^{1/2}\|\rho^\delta(t) - \rho^\infty(t)\|^{1/2}_{\mathscr{L}^1(\Omega)}, \end{aligned} \tag{F.64}$$

where in the last step we have used Cauchy-Schwarz inequality. By Lemma F.7, for almost all $t \in [0,T]$, $\rho^\delta(t)$ converges to $\rho^\infty(t)$ in $\mathscr{L}^2(\Omega)$. As a result, the second integral in (F.59) also tends to 0 as $\delta \to 0$.

Finally, let us bound the third integral in (F.59). Define $h(x) = x\log x$. Then, for $x > 1/4$,

$$|h'(x)| \le 1 + |\log x| \le 1 + \log 4 + x. \tag{F.65}$$

Thus,

$$\begin{aligned} &\left|\int_C \left(\rho^\delta(\boldsymbol{x},t)\log\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)\log\rho^\infty(\boldsymbol{x},t)\right)\,\mathrm{d}\boldsymbol{x}\right| \\ &\quad \le \int_C \left|\rho^\delta(\boldsymbol{x},t)\log\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)\log\rho^\infty(\boldsymbol{x},t)\right|\,\mathrm{d}\boldsymbol{x} \\ &\quad \le \int_C \left|\rho^\delta(\boldsymbol{x},t) - \rho^\infty(\boldsymbol{x},t)\right| \cdot (1 + \log 4 + \rho^\delta(\boldsymbol{x},t) + \rho^\infty(\boldsymbol{x},t))\,\mathrm{d}\boldsymbol{x} \\ &\quad \le (1+\log 4)\|\rho^\delta(t) - \rho^\infty(t)\|_{\mathscr{L}^1(\Omega)} + \|(\rho^\delta(t))^2 - (\rho^\infty(t))^2\|_{\mathscr{L}^1(\Omega)} \\ &\quad \le (1+\log 4)\|\rho^\delta(t) - \rho^\infty(t)\|_{\mathscr{L}^1(\Omega)} \\ &\quad + \|\rho^\delta(t) - \rho^\infty(t)\|_{\mathscr{L}^2(\Omega)} \cdot \|\rho^\delta(t) + \rho^\infty(t)\|_{\mathscr{L}^2(\Omega)}, \end{aligned} \tag{F.66}$$

where in the last step we have used Cauchy-Schwarz inequality. By Lemma F.7, for almost all $t \in [0,T]$, $\rho^\delta(t)$ converges to $\rho^\infty(t)$ in $\mathscr{L}^2(\Omega)$. By Lemma F.6, $\rho^\infty(t) \in \mathscr{L}^2(\Omega)$ for almost all $t \in [0,T]$. Furthermore, by Lemma F.5, the quantity $\|\rho^\delta(t)\|_{\mathscr{L}^2(\Omega)}$ has a δ-free upper bound for $t \in [0,T_0]$. As a result, for almost all $t \in [0,T_0]$, the third integral in (F.59) tends to 0 as $\delta \to 0$. By iterating this argument T/T_0 times, we conclude that for almost all $t \in [0,T_0]$, the third integral in (F.59) tends to 0 as $\delta \to 0$, and the proof is complete. □

At this point, we are ready to provide the proof of Theorem 5.3.

Proof of Theorem 5.3. By substituting z with $z^{1/2p}$ in Theorem 5.1, we have that with probability at least $1 - 1/z$

$$R_N(\boldsymbol{w}^k) \le R^\delta(\rho^\delta_{k\varepsilon}) + z^{1/2p}\,\mathsf{err}(N,d,\varepsilon,\delta)\; e^{C_* p \delta^{-(d+2)} T}, \tag{F.67}$$

where $\mathsf{err}(N,d,\varepsilon,\delta)$ is defined in (5.2). The risk $R^\delta(\rho^\delta_{k\varepsilon})$ can be upper bounded as

$$\begin{aligned} R^\delta(\rho^\delta_{k\varepsilon}) &= \nu_0\,\|f - K^\delta * \rho^\delta_{k\varepsilon}\|^2_{\mathscr{L}^2(\Omega)} \\ &\le \nu_0\,\left(\|f - \rho^\delta_{k\varepsilon}\|_{\mathscr{L}^2(\Omega)} + \|K^\delta * \rho^\delta_{k\varepsilon} - \rho^\delta_{k\varepsilon}\|_{\mathscr{L}^2(\Omega)}\right)^2 \\ &= R(\rho^\delta_{k\varepsilon}) + \Delta_0(\delta,T,d), \end{aligned} \tag{F.68}$$

where $\Delta_0(\delta, T, d) \to 0$ as $\delta \to 0$, since both $K^\delta * \rho_t^\delta$ and ρ_t^δ converge in $\mathscr{L}^2(\Omega)$ to ρ_t^∞. Furthermore, by Theorem F.8,

$$\begin{aligned} R(\rho_{k\varepsilon}^\delta) &= 2\,F(\rho_{k\varepsilon}^\delta) + 2\tau S(\rho_{k\varepsilon}^\delta) \le 2\,F(\rho_{k\varepsilon}^\delta) + 2\tau \log|\Omega| \\ &\le 2\,F(\rho^*) + 2\,(F(\rho^\infty(0)) - F(\rho^*))e^{-2\alpha k\varepsilon} + 2\tau\log|\Omega| + \Delta(\delta,T,d) \\ &= 2\,F(\rho^\infty(0))e^{-2\alpha k\varepsilon} + 2\,(1-e^{-2\alpha k\varepsilon})\,F(\rho^*) + 2\tau\log|\Omega| + \Delta(\delta,T,d), \end{aligned} \tag{F.69}$$

where $\Delta(\delta, T, d) \to 0$ as $\delta \to 0$ and we recall that $|\Omega|$ denotes the volume of the set Ω.

Note that

$$F(\rho^*) \le F(f) = -\tau S(f), \tag{F.70}$$

since ρ^* is the minimizer of F. By combining (F.70) with (F.69), we deduce that

$$\begin{aligned} R(\rho_{k\varepsilon}^\delta) &\le 2\,F(\rho^\infty(0))e^{-2\alpha k\varepsilon} + 2\tau\left(\log|\Omega| - (1-e^{-2\alpha k\epsilon})S(f)\right) + \Delta(\delta,T,d) \\ &= R(\rho^\infty(0))e^{-2\alpha k\varepsilon} + 2\tau\left(\log|\Omega| - (1-e^{-2\alpha k\epsilon})S(f) - S(\rho^\infty(0))\,e^{-2\alpha k\epsilon}\right) \\ &\qquad\qquad + \Delta(\delta,T,d) \\ &\le R_N(\boldsymbol{w}^0)e^{-2\alpha k\varepsilon} + 2\tau\left(\log|\Omega| - (1-e^{-2\alpha k\epsilon})S(f) - S(\rho^\infty(0))\,e^{-2\alpha k\epsilon}\right) \\ &\qquad + z\,\mathsf{err}(N,d,\varepsilon,\delta)\,e^{C_* p\delta^{-(d+2)}\,T}\,e^{-2\alpha k\epsilon} + \Delta(\delta,T,d), \end{aligned} \tag{F.71}$$

where in the last step we use again the result of Theorem 5.1 and the fact that $R(\rho^\infty(0)) - R^\delta(\rho^\infty(0))$ tends to 0 as $\delta \to 0$.

By optimizing over p in (F.67), we will set $\Delta_1(N, \varepsilon, T, d, z)$ as in (5.8). We also let $\Delta_2(\delta, T, d) = \Delta_0(\delta, T, d) + \Delta(\delta, T, d)$. Then, the result follows by combining (F.67), (F.68) and (F.71). □

G Heat kernel in bounded domains with Neumann boundary

Given the domain $D \subseteq \mathbb{R}^d$ (compact, with $\mathscr{C}^2$ boundary ∂D), we denote by $G^D(\boldsymbol{x}, \boldsymbol{y}; t)$ the associated heat kernel, with Neumann boundary conditions. We collect here a few well known facts about this kernel (see, e.g., [Tay13, Section 6.1]).

The heat kernel can be defined as a function $G^D : D \times D \times \mathbb{R}_{>0}$ satisfying

$$\partial_t G^D(\boldsymbol{x}, \boldsymbol{y}; t) = \Delta_{\boldsymbol{y}} G^D(\boldsymbol{x}, \boldsymbol{y}; t)\,, \tag{G.1}$$

$$\langle \nabla_{\boldsymbol{y}} G^D(\boldsymbol{x}, \boldsymbol{y}; t), \boldsymbol{n}(\boldsymbol{y})\rangle = 0 \quad \forall \boldsymbol{y} \in \partial D\,, \tag{G.2}$$

$$G^D(\boldsymbol{x}, \cdot\,; t) \Rightarrow \delta_{\boldsymbol{x}}\,, \quad \text{as } t \to 0,\ \boldsymbol{x} \in D^\circ. \tag{G.3}$$

We will also denote by $G(\boldsymbol{x}, \boldsymbol{y}; t)$ the heat kernel on $\mathbb{R}^d$, namely

$$G(\boldsymbol{x}, \boldsymbol{y}; t) \equiv \frac{1}{(4\pi t)^{d/2}} \exp\left\{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{4t}\right\}. \tag{G.4}$$

The probabilistic interpretation of G^D is as follows (see, e.g., [BGL13]). Let $\mathbb{E}_{\boldsymbol{x}}$ denote expectation with respect to a Brownian motion $\boldsymbol{X}_t$, with initial condition $\boldsymbol{X}_0 = \boldsymbol{x}$, and reflected at ∂D

(see Section C for definitions of this process, following [Tan79]). Then, for any bounded continuous function $\varphi : D \to \mathbb{R}$,

$$\mathbb{E}_{\boldsymbol{x}}\{\varphi(\boldsymbol{X}_t)\} = \int G^D(\boldsymbol{x}, \boldsymbol{y}; t)\, \varphi(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \tag{G.5}$$

$$\equiv G^D_\varphi(\boldsymbol{x}; t)\,. \tag{G.6}$$

Finally, G^D can be viewed as the kernel representation of the bounded operator $e^{t\Delta/2}$ in $\mathscr{L}^2(D, \mathsf{Unif})$. We have

$$(e^{t\Delta/2} f)(\boldsymbol{y}) = \int f(\boldsymbol{x})\, G^D(\boldsymbol{x}, \boldsymbol{y}; t)\, \mathrm{d}\boldsymbol{y}\,. \tag{G.7}$$

Hence $G^D(\boldsymbol{x}, \boldsymbol{y}; t)$ can be represented in terms of the eigenfunctions ϕ_k, and eigenvalues λ_k, of $-\Delta$,

$$G^D(\boldsymbol{x}, \boldsymbol{y}; t) = \sum_{k=0}^{\infty} e^{-\lambda_k t} \phi_k(\boldsymbol{x})\, \phi_k(\boldsymbol{y})\,. \tag{G.8}$$

Here $0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \ldots$, with $\lim_{k\to\infty} \lambda_k = \infty$, and $\phi_0(\boldsymbol{x}) = \mathbf{1}_D(\boldsymbol{x})/\mathrm{Vol}(D)^{1/2}$.

Remark G.1. Since Δ is self-adjoint in $\mathscr{L}^2(D, \mathsf{Unif})$, it follows that G^D is symmetric, namely $G^D(\boldsymbol{x}, \boldsymbol{y}, t) = G^D(\boldsymbol{y}, \boldsymbol{x}; t)$, and therefore it satisfies

$$\partial_t G^D(\boldsymbol{x}, \boldsymbol{y}; t) = \Delta_{\boldsymbol{x}} G^D(\boldsymbol{x}, \boldsymbol{y}; t)\,, \tag{G.9}$$

$$\langle \nabla_{\boldsymbol{x}} G^D(\boldsymbol{x}, \boldsymbol{y}; t), \boldsymbol{n}(\boldsymbol{x})\rangle = 0 \quad \forall \boldsymbol{x} \in \partial D\,. \tag{G.10}$$

Theorem G.1. *The Neumann heat kernel satisfies the following properties:*

1. *We have that*

$$G^D(\boldsymbol{x}, \boldsymbol{y}; t) = G(\boldsymbol{x}, \boldsymbol{y}; t) + G^D_R(\boldsymbol{x}, \boldsymbol{y}; t)\,, \tag{G.11}$$

   where $G^D_R \in \mathscr{C}^\infty(D \times D \times \mathbb{R}_{\ge})$.

2. *For any* $t > 0$, $G^D(\,\cdot\,,\,\cdot\,; t) \in \mathscr{C}^\infty(D \times D)$.

3. *We have that, for a constant* $C(D)$,

$$\big\|\nabla G^D(\boldsymbol{x}, \,\cdot\,; t)\big\|_{\mathscr{L}^1(D)} \le \frac{C(D)}{\sqrt{t}}\,. \tag{G.12}$$

Proof. Substituting $G^D(\boldsymbol{x}, \boldsymbol{y}; t) = G(\boldsymbol{x}, \boldsymbol{y}; t) + G^D_R(\boldsymbol{x}, \boldsymbol{y}; t)$ into Eqs. (G.1) to (G.3) yields, for $\boldsymbol{x} \in D$,

$$\partial_t G^D_R(\boldsymbol{x}, \boldsymbol{y}; t) = \Delta_{\boldsymbol{y}} G^D_R(\boldsymbol{x}, \boldsymbol{y}; t)\,, \tag{G.13}$$

$$\langle \nabla_{\boldsymbol{y}} G^D_R(\boldsymbol{x}, \boldsymbol{y}; t), \boldsymbol{n}(\boldsymbol{y})\rangle = -\langle \nabla_{\boldsymbol{y}} G(\boldsymbol{x}, \boldsymbol{y}; t), \boldsymbol{n}(\boldsymbol{y})\rangle \quad \forall \boldsymbol{y} \in \partial D\,, \tag{G.14}$$

$$G^D_R(\boldsymbol{x}, \boldsymbol{y}; ; 0) = 0\,, \quad \boldsymbol{x}, \boldsymbol{y} \in D^\circ. \tag{G.15}$$

Thus G_R satisfies the heat equation in $D \times [0, T]$ and hence $(\boldsymbol{y}, t) \mapsto G^D_R(\boldsymbol{x}, \boldsymbol{y}; t)$ is $\mathscr{C}^\infty$ inside this domain (see, e.g., [Eva09, Chapter 2, Theorem 8], which refers to Dirichlet boundary condition, but applies equally well to the Neumann case). By symmetry, we have the claimed continuity in $(\boldsymbol{x}, \boldsymbol{y})$, thus proving point 1.

Claim 2 follows by the same decomposition.

Finally, claim 3 follows from Lemma 3.1 in [WY13]. □

H Some useful technical lemmas

Lemma H.1 (Displacement convexity of quadratic functionals)**.** *Let $U : \mathbb{R}^d \to \mathbb{R}^d$ be twice differentiable with $|U(\boldsymbol{x})| \leq C(1+|\boldsymbol{x}|^2)$, $U(\boldsymbol{x}) = U(-\boldsymbol{x})$, and define $\mathscr{U} : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$ by $\mathscr{U}(\rho) \equiv \int U(\boldsymbol{x}-\boldsymbol{x}')\,\rho(\mathrm{d}\boldsymbol{x})\,\rho(\mathrm{d}\boldsymbol{x}')$. Then $\mathscr{U}$ is displacement convex if and only if U is convex.*

Proof. Proposition 7.4 in [San15] proves that convexity of U implies displacement convexity of $\mathscr{U}$. To prove the converse implication, let $\boldsymbol{x}, \boldsymbol{\delta} \in \mathbb{R}^d$, $\boldsymbol{x} \neq \boldsymbol{0}$ and consider the two probability distributions $\rho_0 = (\delta_{\boldsymbol{0}} + \delta_{\boldsymbol{x}})/2$ and $\rho_1 = (\delta_{\boldsymbol{0}} + \delta_{\boldsymbol{x}+\boldsymbol{\delta}})/2$. For $|\boldsymbol{\delta}| < |\boldsymbol{x}|$, the geodesic path connecting these distribution is $\rho_t = (\delta_{\boldsymbol{0}} + \delta_{\boldsymbol{x}+t\boldsymbol{\delta}})/2$, $t \in [0,1]$. Substituting in the definition of $\mathscr{U}$, we get

$$\mathscr{U}(\rho_t) = \frac{1}{2}U(\boldsymbol{0}) + \frac{1}{2}U(\boldsymbol{x}+t\boldsymbol{\delta}) \tag{H.1}$$

$$= \mathscr{U}(\rho_0) + \frac{t}{2}\langle \nabla U(\boldsymbol{x}), \boldsymbol{\delta}\rangle + \frac{t^2}{4}\langle \boldsymbol{\delta}, \nabla^2 U(\boldsymbol{x})\boldsymbol{\delta}\rangle + o(t^2)\,. \tag{H.2}$$

Hence, displacement convexity implies $\langle \boldsymbol{\delta}, \nabla^2 U(\boldsymbol{x})\boldsymbol{\delta}\rangle \geq 0$. Since this holds for all $|\boldsymbol{\delta}| < |\boldsymbol{x}|$, we obtain $\nabla^2 U(\boldsymbol{x}) \succeq \boldsymbol{0}$ for all $\boldsymbol{x} \neq \boldsymbol{0}$, which in turns imply that U is convex (by a continuity argument, it is sufficient to lower bound the Hessian everywhere except at a point). □

Lemma H.2 (A Gronwall type inequality [Bih56])**.** *Let $u : [0,T] \to \mathbb{R}_+$ be a continuous function that satisfies the inequality*

$$u(t) \leq A + \int_0^t \Psi(s)\omega(u(s))\mathrm{d}s\,, \quad t \in [0,T]\,,$$

where $A \geq 0$, $\Psi : [0,T] \to \mathbb{R}_+$ is continuous and $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and monotone-increasing. Then, the following holds

$$u(t) \leq \Phi^{-1}\left(\Phi(A) + \int_0^t \Psi(s)\mathrm{d}s\right), \quad t \in [0,T]\,,$$

with $\Phi : \mathbb{R} \mapsto \mathbb{R}$ given by

$$\Phi(u) \equiv \int_{u_0}^u \frac{\mathrm{d}s}{\omega(s)}\,, \quad u \in \mathbb{R}, \quad u_0 \equiv \omega(A)\,.$$

Remark H.1. To derive Equation (F.45), we use Lemma H.2 with $\omega(u) = u^2$, $\Psi(s) = \bar{C}_m C_3$, $A = \bar{C}_m C_4 + \bar{C}_m \tau M^2 C_a s t T_n$.

Lemma H.3 (Gagliardo-Nirenberg interpolation inequality, cf. Theorem 1.5.2 of [CM12])**.** *Fix $1 \leq q, r \leq \infty$ and m a positive integer. Let $u \in \mathscr{L}^q(\Omega) \cap \mathscr{L}^r(\Omega)$ and $\nabla^{\otimes m} u \in \mathscr{L}^p(\Omega)$. For integer j, $0 \leq j \leq m$, and $\theta \in [j/m, 1]$ (with the exception $\theta \neq 1$ if $m - j - d/2$ is a non-negative integer), define p by*

$$\frac{1}{p} = \frac{j}{d} + \theta\left(\frac{1}{r} - \frac{m}{d}\right) + \frac{1-\theta}{q}\,.$$

Then $\nabla^{\otimes j} u \in \mathscr{L}^p(\Omega)$ and satisfies

$$\|\nabla^{\otimes j} u\|_p \leq C\|\nabla^{\otimes m} u\|_r^\theta \, \|u\|_q^{1-\theta} + C_1\|u\|_s\,.$$

with finite arbitrary $1 \leq s \leq \max(r,q)$ and $C > 0$ and $C_1 \geq 0$ are independent of u. The constant C is independent of Ω, while $C_1 \to 0$ as $|\Omega| \to \infty$. In particular, the choice $C_1 = 0$ is admissible if $\Omega = \mathbb{R}^d$.

References

[AB09] Martin Anthony and Peter L. Bartlett, *Neural network learning: Theoretical foundations*, Cambridge University Press, 2009. 2, 6

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008. 10

[Bac17] Francis Bach, *Breaking the curse of dimensionality with convex neural networks*, The Journal of Machine Learning Research **18** (2017), no. 1, 629–681. 3

[Bar93] Andrew R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information theory **39** (1993), no. 3, 930–945. 6

[Bar98] Peter L. Bartlett, *The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network*, IEEE Transactions on Information Theory **44** (1998), no. 2, 525–536. 6, 24

[BGL13] Dominique Bakry, Ivan Gentil, and Michel Ledoux, *Analysis and geometry of markov diffusion operators*, vol. 348, Springer Science & Business Media, 2013. 64

[Bih56] Imre Bihari, *A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations*, Acta Mathematica Hungarica **7** (1956), no. 1, 81–94. 66

[BJW18] Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff, *Learning two layer rectified neural networks in polynomial time*, arXiv:1811.01885 (2018). 5

[BRV+06] Yoshua Bengio, Nicolas L. Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte, *Convex neural networks*, Advances in Neural Information Processing Systems, 2006, pp. 123–130. 6

[BY03] Peter Bühlmann and Bin Yu, *Boosting with the L2 loss: regression and classification*, Journal of the American Statistical Association **98** (2003), no. 462, 324–339. 2

[CB18] Lenaic Chizat and Francis Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in Neural Information Processing Systems, 2018. 4, 5, 6

[CFG14] Emmanuel J. Candès and Carlos Fernandez-Granda, *Towards a mathematical theory of super-resolution*, Communications on Pure and Applied Mathematics **67** (2014), no. 6, 906–956. 2

[CJM+01] José A. Carrillo, Ansgar Jüngel, Peter A. Markowich, Giuseppe Toscani, and Andreas Unterreiter, *Entropy dissipation methods for degenerate parabolicproblems and generalized sobolev inequalities*, Monatshefte für Mathematik **133** (2001), no. 1, 1–82. 4, 5, 23, 61

[CM12] Pascal Cherrier and Albert Milani, *Linear and quasi-linear evolution equations in Hilbert spaces*, Graduate studies in mathematics, American Mathematical Soc., 2012. 66

[CMV03] José A. Carrillo, Robert J. McCann, and Cédric Villani, *Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates*, Revista Matematica Iberoamericana **19** (2003), no. 3, 971–1018. 4, 5, 23

[CMV06] ———, *Contractions in the 2-Wasserstein length space and thermalization of granular media*, Archive for Rational Mechanics and Analysis **179** (2006), no. 2, 217–263. 4, 5, 23

[CS16] Yining Chen and Richard J Samworth, *Generalized additive and index models with shape constraints*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **78** (2016), no. 4, 729–754. 11

[CST00] Nello Cristianini and John Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2000. 2

[CT06] Thomas M Cover and Joy A Thomas, *Elements of information theory*, John Wiley & Sons, 2006. 63

[Cyb89] George Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of control, signals and systems **2** (1989), no. 4, 303–314. 6

[Dob79] Roland L'vovich Dobrushin, *Vlasov equations*, Functional Analysis and Its Applications **13** (1979), no. 2, 115–123. 6

[Don92] David L. Donoho, *Superresolution via sparsity constraints*, SIAM Journal on Mathematical Analysis **23** (1992), no. 5, 1309–1331. 2

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh, *Gradient descent provably optimizes over-parameterized neural networks*, arXiv:1810.02054 (2018). 5

[Eva09] Lawrence C. Evans, *Partial differential equations*, Springer, 2009. 65

[FP08] Alessio Figalli and Robert Philipowski, *Convergence to the viscous porous medium equation and propagation of chaos*, ALEA Lat. Am. J. Probab. Math. Stat **4** (2008), 185–203. 22

[Fri01] Jerome H. Friedman, *Greedy function approximation: a gradient boosting machine*, Annals of Statistics (2001), 1189–1232. 2

[HD13] Lauren A Hannah and David B Dunson, *Multivariate convex regression with adaptive partitioning*, The Journal of Machine Learning Research **14** (2013), no. 1, 3261–3294. 11

[LS84] Pierre-Louis Lions and Alain-Sol Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Communications on Pure and Applied Mathematics **37** (1984), no. 4, 511–537. 21

[LSU88] Olga Aleksandrovna Ladyzhenskaia, Vsevolod Alekseevich Solonnikov, and Nina N. Ural'tseva, *Linear and quasi-linear equations of parabolic type*, vol. 23, American Mathematical Society, 1988. 28, 45

[LY17] Yuanzhi Li and Yang Yuan, *Convergence analysis of two-layer neural networks with relu activation*, Advances in Neural Information Processing Systems, 2017, pp. 597–607. 5

[MBM+18] Song Mei, Yu Bai, Andrea Montanari, et al., *The landscape of empirical risk for non-convex losses*, The Annals of Statistics **46** (2018), no. 6A, 2747–2774. 24

[McC97] Robert J McCann, *A convexity principle for interacting gases*, Advances in mathematics **128** (1997), no. 1, 153–179. 10

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari, *Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit*, Conference on Learning Theory (COLT), 2019. 6, 24

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences (2018). 4, 5, 6, 19, 21, 24, 31, 38, 43, 44, 45

[NS17] Atsushi Nitanda and Taiji Suzuki, *Stochastic particle gradient descent for infinite ensembles*, arXiv:1712.05438 (2017). 6

[Oel01] Karl Oelschläger, *A sequence of integro-differential equations approximating a viscous porous medium equation*, Zeitschrift für Analysis und ihre Anwendungen **20** (2001), no. 1, 55–91. 55, 56, 57

[Oel02] ———, *Simulation of the solution of a viscous porous medium equation by a particle method*, SIAM Journal on Numerical Analysis **40** (2002), no. 5, 1716–1762. 22

[Phi07] Robert Philipowski, *Interacting diffusions approximating the porous medium equation and propagation of chaos*, Stochastic Processes and their Applications **117** (2007), no. 4, 526–538. 22

[PS91] Jooyoung Park and Irwin W. Sandberg, *Universal approximation using radial-basis-function networks*, Neural computation **3** (1991), no. 2, 246–257. 3

[Ros62] Frank Rosenblatt, *Principles of neurodynamics*, Spartan Book, 1962. 2

[RR08] Ali Rahimi and Benjamin Recht, *Random features for large-scale kernel machines*, Advances in neural information processing systems, 2008, pp. 1177–1184. 2

[RVE18] Grant M. Rotskoff and Eric Vanden-Eijnden, *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, Advances in Neural Information Processing Systems, 2018. 4, 6

[RW94] L. Chris G. Rogers and David Williams, *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, vol. 2, Cambridge university press, 1994. 36

[San15] Filippo Santambrogio, *Optimal transport for applied mathematicians: Calculus of variations, PDEs, and modeling*, vol. 87, Birkhäuser, 2015. 10, 66

[Sch03] Robert E. Schapire, *The boosting approach to machine learning: An overview*, Nonlinear estimation and classification, Springer, 2003, pp. 149–171. 2

[SJL18] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee, *Theoretical insights into the optimization landscape of over-parameterized shallow neural networks*, IEEE Transactions on Information Theory (2018). 5

[Slo94] Slomiński, Leszek, *On approximation of solutions of multidimensional SDE's with reflecting boundary conditions*, Stochastic processes and their Applications **50** (1994), no. 2, 197–219. 38, 40

[Slo01] ———, *Euler's approximations of solutions of SDEs with reflecting boundary*, Stochastic processes and their applications **94** (2001), no. 2, 317–337. 38, 48

[SS18] Justin Sirignano and Konstantinos Spiliopoulos, *Mean field analysis of neural networks*, arXiv:1805.01053 (2018). 4, 6

[Szn91] Alain-Sol Sznitman, *Topics in propagation of chaos*, Ecole d'été de probabilités de Saint-Flour XIX—1989, Springer, 1991, pp. 165–251. 6, 38

[Tan79] Hiroshi Tanaka, *Stochastic differential equations with reflecting boundary condition in convex regions*, Hiroshima Mathematical Journal **9** (1979), no. 1, 163–177. 21, 35, 37, 65

[Tay13] Michael Taylor, *Partial differential equations I: Basic theory*, vol. 115, Springer Science & Business Media, 2013. 64

[Tho13] James William Thomas, *Numerical partial differential equations: finite difference methods*, vol. 22, Springer Science & Business Media, 2013. 12

[Tia17] Yuandong Tian, *Symmetry-breaking convergence analysis of certain two-layered neural networks with ReLU nonlinearity*, Workshop at International Conference on Learning Representation (ICLR), 2017. 5

[Váz07] Juan Luis Vázquez, *The porous medium equation: mathematical theory*, Oxford University Press, 2007. 10, 28

[Vil08] Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008. 10

[WLLM18] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma, *On the margin theory of feedforward neural networks*, arXiv:1810.05369 (2018). 5

[WY13] Feng-Yu Wang and Lixin Yan, *Gradient estimate on convex domains and applications*, Proceedings of the American Mathematical Society **141** (2013), no. 3, 1067–1081. 65

[ZSJ+17] Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon, *Recovery guarantees for one-hidden-layer neural networks*, International Conference on Machine Learning, 2017, pp. 4140–4149. 5