The Annals of Statistics
2020, Vol. 48, No. 6, 3619–3642
https://doi.org/10.1214/20-AOS1945

ANALYSIS OF A TWO-LAYER NEURAL NETWORK VIA DISPLACEMENT CONVEXITY

BY ADEL JAVANMARD[1], MARCO MONDELLI[2] AND ANDREA MONTANARI[3]

[1]*Data Sciences and Operations Department, Marshall School of Business, University of Southern California, ajavanma@usc.edu*

[2]*Institute of Science and Technology (IST) Austria, marco.mondelli@ist.ac.at*

[3]*Department of Electrical Engineering and Department of Statistics, Stanford University, montanar@stanford.edu*

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 nonconvex 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 nonconvex. 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 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 $\sigma(\boldsymbol{x}; \boldsymbol{w}) =$

Received January 2019; revised December 2019.
MSC2020 subject classifications. Primary 62F10, 62J02; secondary 62H12.
Key words and phrases. Neural networks, stochastic gradient descent, Wasserstein gradient flow, function regression, convergence rate, displacement convexity.

$(1+\exp(-\langle \boldsymbol{w}, \boldsymbol{x}\rangle))^{-1}$ (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

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

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

- Two-layer neural networks [2, 35];
- Sparse deconvolution [9, 18];
- Kernel ridge regression and related random feature methods [15, 34];
- Boosting [8, 21, 38].

Despite the impressive practical success of these methods, the risk function $R_N(\boldsymbol{w})$ is highly nonconvex 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 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 learned. While this is a fruitful idea, it gives up the degrees of freedom afforded by the $\boldsymbol{w}_i$'s.

Boosting overcomes nonconvexity by fitting the components $\boldsymbol{w}_1, \dots, \boldsymbol{w}_N$ 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, for example, $\hat{f}(x) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$. 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 [3]. 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 $K^{\delta}(\boldsymbol{x}) = \delta^{-d} K(\boldsymbol{x}/\delta)$, $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a first-order kernel with compact support, and $\boldsymbol{w}_i \in \Omega^{\delta}$ for $i \leq N$. 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, that is, $d = D$.) We refer to Section 5 for a formal statement of our assumptions. From equation (1.2), we have

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

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,

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

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 [32]. 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, that is, 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 nonuniform 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 equation (1.1). We deliberately set the weights to a fixed value because the risk function is convex in $\boldsymbol{a} = (a_i)_{i \leq N}$, 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 nonconvex regardless of the concavity properties of f. Our proof unveils a novel mechanism by which global convergence takes place. Convergence results for nonconvex 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 [10–12]. 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

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

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 $(\rho_t^\delta)_{t\ge 0}$ 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 [14, 29, 36, 39]. 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 proven in [29] and in [14] by showing that, for all $t \ge 0$ ρ_t^δ has a density that is either smooth, or strictly positive everywhere. However, these convergence results are nonquantitative, 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 for small nonzero δ. In fact, up to an additive constant, the risk is equal to

$$-2\nu_0 \int K^\delta * f(\boldsymbol{x})\rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \nu_0 \int K^\delta * K^\delta(\boldsymbol{x}-\boldsymbol{x}')\rho(\boldsymbol{x})\rho(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}',$$

which is not displacement convex unless $K^\delta * K^\delta - 2K^\delta * f$ is convex (see Lemma H.1), and this cannot hold in our setting (see Lemma H.2). 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 this $\delta \to 0$ approximation rigorous. 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 $R(\rho)$.

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

[2] An argument indicating convergence in a time polynomial in d was put forward in [47], but for a different type of continuous flow.

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 [10–12]. Namely, if f is α-strongly concave, then $R(\rho_t) \le R(\rho_0)e^{-2\alpha t}$. By using the approximation results outlined above, we obtain global convergence for SGD. With high probability,

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

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, for example, [4, 44, 48]. 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, for example, [24, 40]. However, global convergence was only proved in the extremely overparametrized regime in which the neural network essentially behaves as kernel ridge regression [19].

Classical theory of neural networks was largely devoted to the two-layer case [2], although the focus was on representation and approximation questions [5, 16], 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 [6], and was exploited again under the label of "convex neural networks" in [7].

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 [14, 28–30, 36, 39]. In particular, it was proven in [29] 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, that is, projecting at each step the weights onto the set Ω^δ. We generalize the analysis of [29], obtaining convergence to a PDE with Neumann (reflecting) boundary conditions. As in [29], we build on ideas that were first developed in the context of interacting particle systems [17, 41].

The Wasserstein gradient flow approach was used in [14, 29] to establish global convergence results. However, these results fall short of our objectives for several reasons:

- The global convergence result of [14] rely on certain homogeneity properties of the neurons that are lacking here. We could obtain homogeneity by adding coefficients to equation (1.4), that is, 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 [14, 29] 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, [14, 28, 29] 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, [29] 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 [14, 29] 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 toward dimension-independent convergence rate, in a more restricted setting than [14, 28, 29].

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 nonasymptotic convergence rates, they point clearly in that direction.

3. Model and assumptions.

3.1. *Notation.* We will use lowercase boldface for vectors, for example, $\boldsymbol{x}$, $\boldsymbol{y}$, ..., uppercase for random variables, for example, X, Y, ..., and uppercase boldface for random vectors, for example, $\boldsymbol{X}$, $\boldsymbol{Y}$, ... 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 pth 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, for example, a, b, c, we will write $C(a, b, c)$. When they are absolute numerical constants, we will emphasize this by writing C_*. (In particular, a constant C_* is independent of the dimension d and of the domain Ω.)

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

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

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

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 f has nonnegative range, this is equivalent to assuming that f is a density. 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 distributions 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 nonnegative symmetric first-order kernel with compact support. Formally, we assume that

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

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

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

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

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, $\Omega \supseteq \mathsf{B}(\mathbf{0}; r)$ implies $\lambda_\delta > 0$ for all $\delta < r/c_0$. Finally, $\lambda_{\delta=0} = 1$, whence $\Omega^{\delta=0} = \Omega$. In our running example, $\Omega^\delta = \mathsf{B}(\mathbf{0}; r - c_0\delta)$ 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

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

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:

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

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

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

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 $\rho_{\text{init}} \in \mathscr{P}_2(\Omega)$, that is, $\rho_{\text{init}}^\delta(S) = \rho_{\text{init}}(S/\lambda_\delta)$. 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 equation (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

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

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," that is, 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 equation (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 equation (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

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

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. equation (1.7).

The corresponding Wasserstein gradient flow is also known as the *viscous porous medium equation* [45] and it is given by

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

with initial and boundary conditions

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

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 equation (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 [26]. For the reader's convenience, we recall its definition here, referring to [1, 37, 46] 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, that is, 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 for any two distributions $\rho_0, \rho_1 \in \mathscr{P}_2(\Omega)$,

$$(1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \mathcal{F}(\rho_t) \geq \frac{1}{2}\lambda t(1-t) W_2(\rho_0, \rho_1)^2 \quad \forall t \in [0, 1]. \tag{3.15}$$

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 [26] 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., [13, 22]), 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 of f being concave or not. Indeed, our numerical

simulations in Section 4.4 indicate convergence to a near-global optimum even for nonconcave functions f.

From a 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. We therefore conjecture that a more general analysis can relax the concavity assumption and show that as $N \to \infty$ and $\delta \to 0$, SGD converges to the global minimizer in a more general setting. We defer this investigation to future work.

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 nonnegative, 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 function g_0 that is twice differentiable on the closure of Ω, the function $f_0(\boldsymbol{x}) = g_0(\boldsymbol{x}) - c_*\|\boldsymbol{x}\|_2^2$ is strongly concave for c_* large enough.

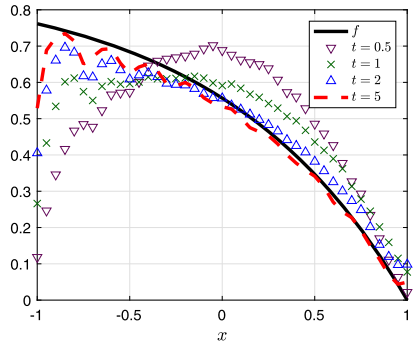
4.1. *A one-dimensional concave function.* We set $\Omega = [-1, 1]$ and $f(x) = (1 - e^{x-1})/(1 - e^{-2})$ (we choose the normalization so that $\int_{-1}^{1} f(x)\,\mathrm{d}x = 1$). Note that f is uniformly concave in $[-1, 1]$. We set the kernel K as follows:

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

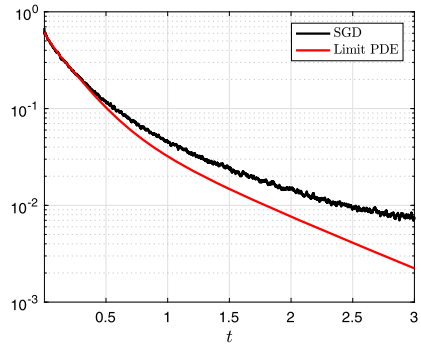
where C_d is a normalization constant ensuring that $\int_{-1}^{1} K(x)\,\mathrm{d}x = 1$. The initialization ρ_{init} is a truncated Gaussian: $\rho_{\text{init}}(x) = c \cdot \exp(-x^2/(2\sigma^2))\mathbf{1}_{[-1,1]}(x)$, with $\sigma = 1/3$.

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.

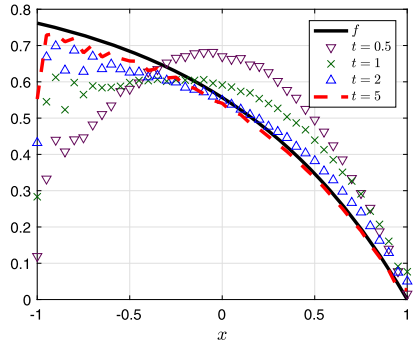
In the right column, we report the evolution of the population risk (1.2) normalized by $\|f\|^2_{\mathscr{L}^2(\Omega)}$. For comparison, we plot the evolution of the risk (1.7) as predicted by the limit PDE (3.12) with $\tau = 0$. We solve the PDE (3.12) numerically using a finite difference scheme that enforces the conservation law $\int \rho(x, t)\,\mathrm{d}x = 1$; see, for example, [43]. 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



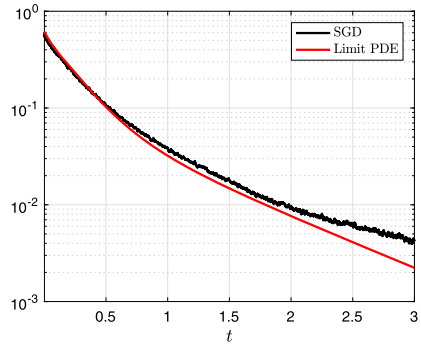
(a) Function f and SGD estimates, $\delta = 1/5$.



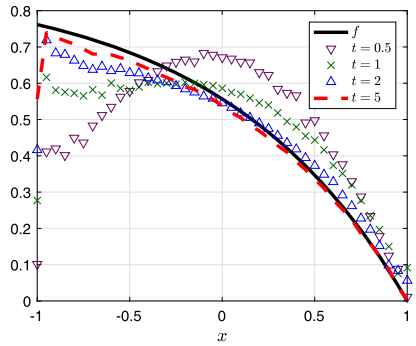
(b) Normalized risk, $\delta = 1/5$.



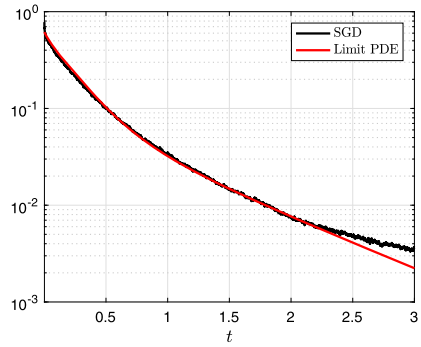
(c) Function f and SGD estimates, $\delta = 1/10$.



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



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

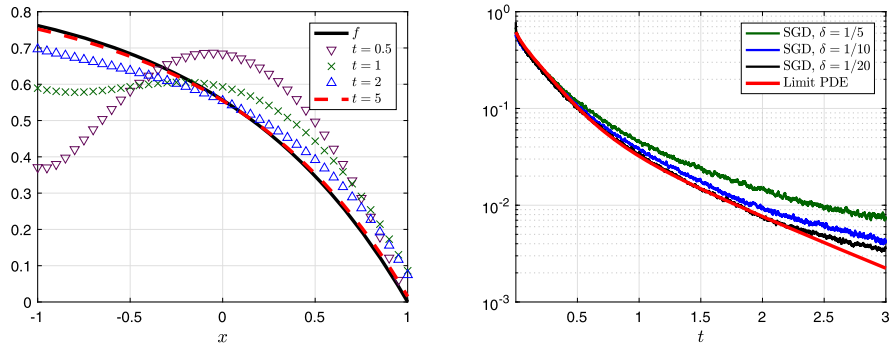


(f) Normalized risk, $\delta = 1/20$.

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

toward optimality. The main difference is that, while the PDE (3.12) corresponds to $\delta = 0$, and hence evolves toward a global optimum at zero risk, SGD converges to a nonzero 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 Figures 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



(a) Function f and PDE solution. (b) Normalized risk.

FIG. 2. *Dynamics of limit PDE* (3.12) *at different times* t.

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 nonnegative 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: $\rho_{\mathrm{init}}(\boldsymbol{x}) = c \cdot \exp(-|\boldsymbol{x}|^2/(2\sigma^2))\mathbf{1}_{[-1,1]^2}(\boldsymbol{x})$, 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.

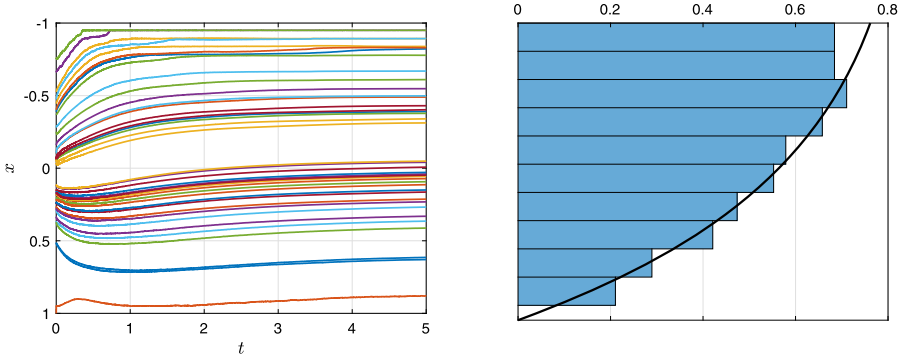


FIG. 3. *Evolution of the value of* 40 *weights chosen at random and histogram of their empirical distribution at time* $t = 5$.

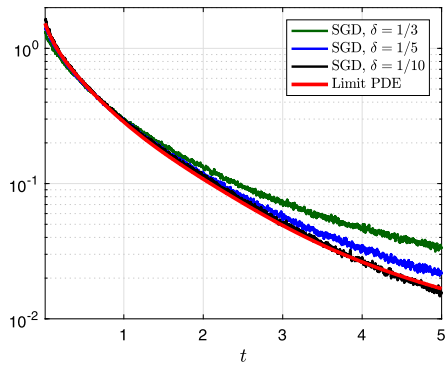


FIG. 4. *Normalized risk of SGD for different values of δ compared with that of the limit PDE for a two-dimensional example.*

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

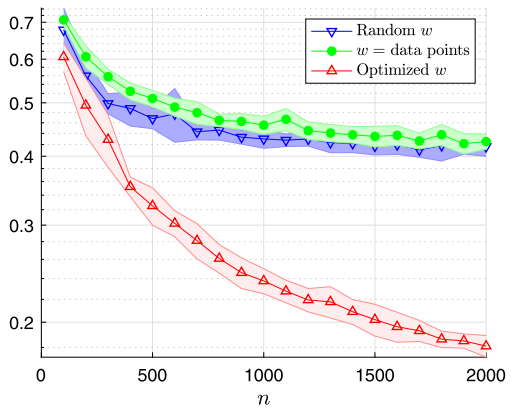
$$\hat{\boldsymbol{a}} = (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{y}, \tag{4.4}$$

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

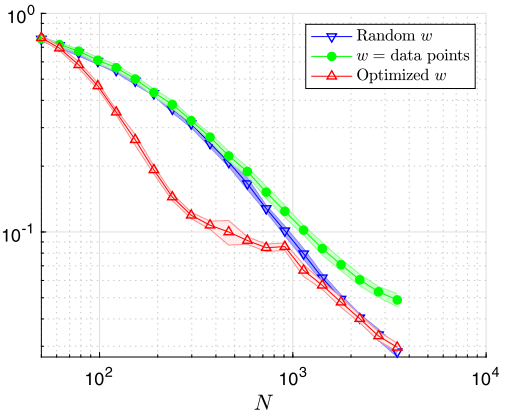
In Figure 5, we compare the performance of three different ways to construct the weights $\boldsymbol{w}$: *"random $\boldsymbol{w}$,"* we choose the weights $\boldsymbol{w}_i$ independently and uniformly at random in Ω (blue triangles pointing down); *"$\boldsymbol{w}$ = data points,"* we choose the 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 to an independent sample from the empirical data, with replacement. SGD is stopped after $k_{\max}$ iteration, and the coefficient $\hat{\boldsymbol{a}}$ is 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,



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



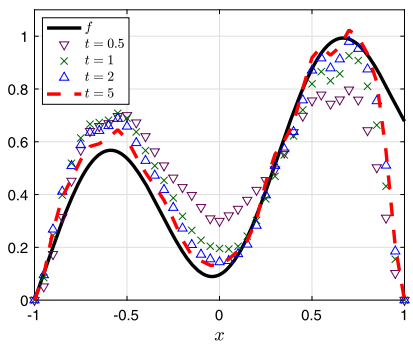
(b) $n = 2000$, N varies on the x-axis.

FIG. 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.*

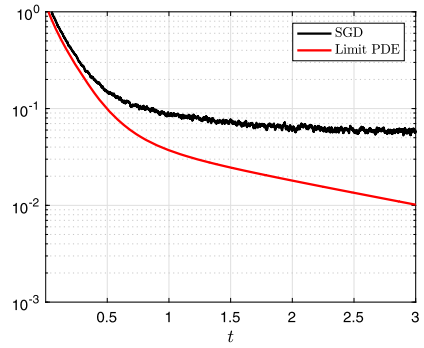
c_1 and c_2 are chosen so that f is nonnegative and $\int_\Omega f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$; the kernel K is given by $K(\boldsymbol{x}) = C_d\kappa(|\boldsymbol{x}|)$, where κ is defined in equation (4.1) and C_d ensures that $\int_{\mathsf{B}(\mathbf{0};1)} K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$.

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.

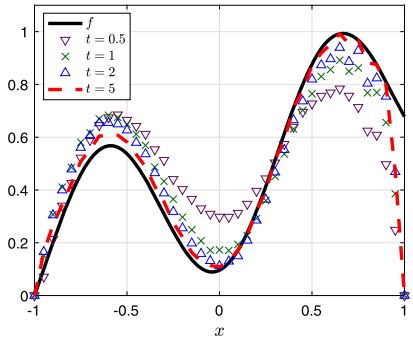
4.4. *A nonconcave one-dimensional example.* We set $\Omega = [-1, 1]$ and $f(x) = (x + \sin(5x - \pi/2) - c_1)/c_2$, where c_1 and c_2 are chosen so that f is nonnegative and $\int_\Omega f(x)\,\mathrm{d}x =$



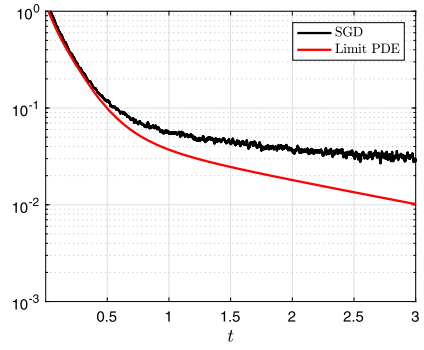
(a) Function f and SGD estimates, $\delta = 1/5$.



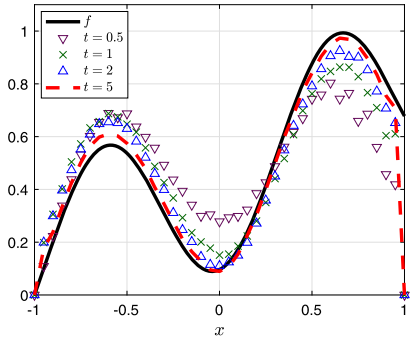
(b) Normalized risk, $\delta = 1/5$.



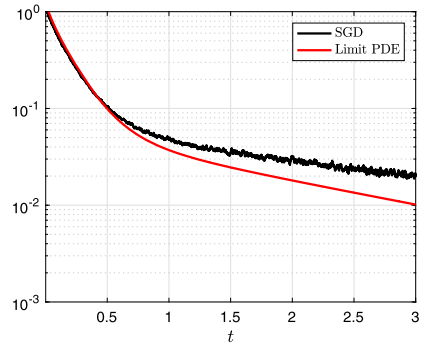
(c) Function f and SGD estimates, $\delta = 1/10$.



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



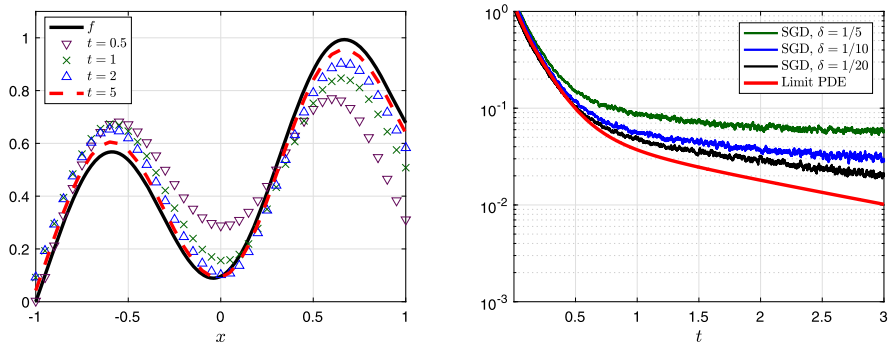
(e) Function f and SGD estimates, $\delta = 1/20$.



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

FIG. 6. *Dynamics of SGD update* (3.5) *at different times* t *and for different values of* δ *for a nonconcave target function* f.

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 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



(a) Function f and PDE solution. (b) Normalized risk.

FIG. 7. *Dynamics of limit PDE* (3.12) *at different times* t *for a nonconcave target function* f.

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.

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* $\operatorname{supp}(\rho_{\text{init}}^{\delta}) \subseteq \mathsf{B}(\mathbf{0}, r)$ *Then, for any fixed* $t \ge 0$, $\rho_{\lfloor t/\varepsilon \rfloor}^{(N)} \Rightarrow \rho_t$ *almost surely along any sequence* $(N, \varepsilon = \varepsilon_N)$ *such that* $N \to \infty$, $\varepsilon_N \to 0$.

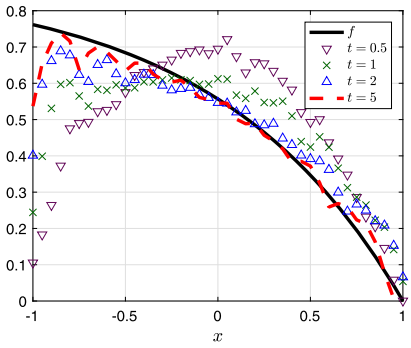
Furthermore, for any $\delta \le 1$, $T \ge 1$, $\varepsilon \le 1$, $p \in \mathbb{N}$, *and for any* $g : \mathbb{R}^d \to \mathbb{R}$ *with* $\|g\|_{\text{Lip}} \le 1$, *the following happens with probability at least* $1 - z^{-2p}$:

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

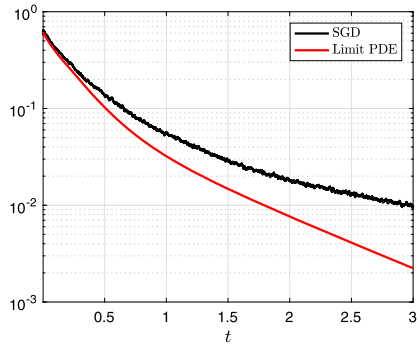
$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \left| R_N(\boldsymbol{w}^k) - R^{\delta}(\rho_{k\epsilon}) \right| \le z\,\mathsf{err}(N, d, \varepsilon, \delta) e^{C_* p \delta^{-(d+2)} T}, \tag{5.1}$$

where

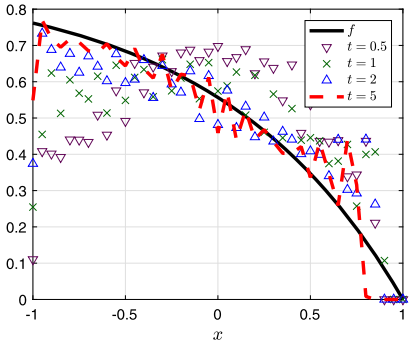
$$\mathsf{err}(N, d, \varepsilon, \delta) = \sqrt{\frac{d}{N}} \vee \left(\delta^{-2d-1} r \left(d^2 \varepsilon \log(1/\varepsilon)\right)^{1/4}\right). \tag{5.2}$$



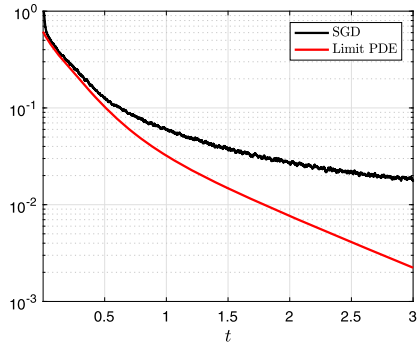
(a) Function f and SGD estimates, $\delta = 1/5$.



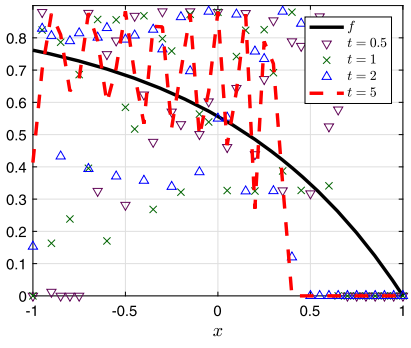
(b) Normalized risk, $\delta = 1/5$.



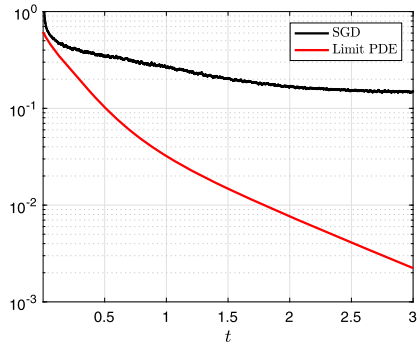
(c) Function f and SGD estimates, $\delta = 1/10$.



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



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



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

FIG. 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* ($N = 20$).

Our proof is based on the same approach developed in [29]. We prove that solutions of the PDE (3.9) are in correspondence with distributions over trajectories $(\boldsymbol{X}_t)_{t\geq 0}$ in Ω satisfying the following stochastic differential equation:

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

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).

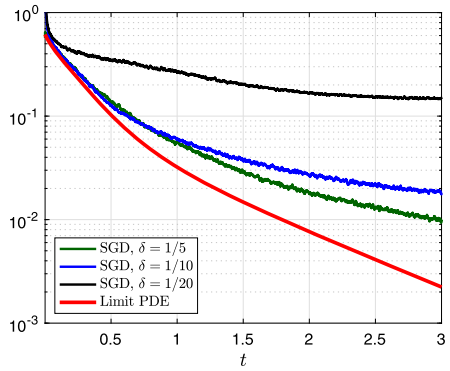


FIG. 9. *Normalized risk of the limit PDE* (3.12) *and of the SGD update* (3.5) *when the number of neurons is too small* ($N = 20$).

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 equation (5.1) is completely analogous to the error in a similar theorem proved in [29]. 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 [29] 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 nontrivial fact first established in [25, 42]. As a consequence, while the main proof idea is similar to the one in [29], 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 [29], 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* $\rho_0^\delta = \rho_{\mathrm{init}}$. *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 [20, 31, 33], 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 [20]. 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 [31] 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, *that is,* $\langle \boldsymbol{y}, \nabla^2 f(\boldsymbol{x})\boldsymbol{y}\rangle \le -\alpha|\boldsymbol{y}|^2$ *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* $\operatorname{supp}(\rho_{\text{init}}) \subseteq \mathsf{B}(\mathbf{0}; r)$. *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

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

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

REMARK 5.3. The error term $2\tau\Delta'(k,\varepsilon,d)$ in equation (5.4) is always nonnegative. In fact, $\Delta'(k,\varepsilon,d) \ge 0$ as $S(\rho) \le \log|\Omega|$ for any $\rho\in\mathcal{P}_2(\Omega)$. Furthermore, by applying Jensen's inequality, we have that, for any $\rho\in\mathcal{P}_2(\Omega)$,

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

which gives the following upper bound:

$$\Delta'(k,\varepsilon,d) \le \log|\Omega| + 2\big|\log\|f\|_{\mathscr{L}^2(\Omega)}\big| + 2\big|\log\|\rho_{\text{init}}\|_{\mathscr{L}^2(\Omega)}\big|.$$

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 equation (5.4). Namely, for an arbitrary but fixed $p\in\mathbb{N}$,

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

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

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

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 [10–12] 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 [10–12], 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 [29], 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., for example, Figure 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., [28, 29]) 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 nonconvex 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 nonconvex problems in high dimensional statistics, including phase retrieval, clustering, matrix completion, error-in-variables models and so on. We refer to [27] for examples and a more detailed survey.

Unfortunately, this approach runs into substantial difficulties when treating complex models such as multilayer 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 [6], 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, that is, 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 $\Delta(t)$:

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

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

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

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 helps addressing the challenge posed 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 equations (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 nonquantitative. 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.

Acknowledgments. The first author was supported in part 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.

The second author was supported by an Early Postdoc.Mobility fellowship from the Swiss National Science Foundation and by the Simons Institute for the Theory of Computing.

The third author was supported in part 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.

SUPPLEMENTARY MATERIAL

Supplement to "Analysis of a two-layer neural network via displacement convexity" (DOI: 10.1214/20-AOS1945SUPP; .pdf). Due to space constraints, proofs of theorems and some of the technical details are provided in the Supplementary Material [23].

REFERENCES

[1] AMBROSIO, L., GIGLI, N. and SAVARÉ, G. (2008). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed. *Lectures in Mathematics ETH Zürich*. Birkhäuser, Basel. MR2401600

[2] ANTHONY, M. and BARTLETT, P. L. (2009). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press. MR1741038 https://doi.org/10.1017/CBO9780511624216

[3] BACH, F. (2017). Breaking the curse of dimensionality with convex neutral networks. *J. Mach. Learn. Res.* **18** Paper No. 19, 53. MR3634886
[4] BAKSHI, A., JAYARAM, R. and WOODRUFF, D. P. (2018). Learning two layer rectified neural networks in polynomial time. arXiv:1811.01885.
[5] BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39** 930–945. MR1237720 https://doi.org/10.1109/18.256500
[6] BARTLETT, P. L. (1998). The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Trans. Inform. Theory* **44** 525–536. MR1607706 https://doi.org/10.1109/18.661502
[7] BENGIO, Y., ROUX, N. L., VINCENT, P., DELALLEAU, O. and MARCOTTE, P. (2006). Convex neural networks. In *Advances in Neural Information Processing Systems* 123–130.
[8] BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. MR1995709 https://doi.org/10.1198/016214503000125
[9] CANDÈS, E. J. and FERNANDEZ-GRANDA, C. (2014). Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.* **67** 906–956. MR3193963 https://doi.org/10.1002/cpa.21455
[10] CARRILLO, J. A., JÜNGEL, A., MARKOWICH, P. A., TOSCANI, G. and UNTERREITER, A. (2001). Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. *Monatsh. Math.* **133** 1–82. MR1853037 https://doi.org/10.1007/s006050170032
[11] CARRILLO, J. A., MCCANN, R. J. and VILLANI, C. (2003). Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoam.* **19** 971–1018. MR2053570 https://doi.org/10.4171/RMI/376
[12] CARRILLO, J. A., MCCANN, R. J. and VILLANI, C. (2006). Contractions in the 2-Wasserstein length space and thermalization of granular media. *Arch. Ration. Mech. Anal.* **179** 217–263. MR2209130 https://doi.org/10.1007/s00205-005-0386-1
[13] CHEN, Y. and SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348 https://doi.org/10.1111/rssb.12137
[14] CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*.
[15] CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge Univ. Press.
[16] CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. MR1015670 https://doi.org/10.1007/BF02551274
[17] DOBRUSHIN, R. L. (1979). Vlasov equations. *Funct. Anal. Appl.* **13** 115–123. MR0541637
[18] DONOHO, D. L. (1992). Superresolution via sparsity constraints. *SIAM J. Math. Anal.* **23** 1309–1331. MR1177792 https://doi.org/10.1137/0523074
[19] DU, S. S., ZHAI, X., POCZOS, B. and SINGH, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. arXiv:1810.02054.
[20] FIGALLI, A. and PHILIPOWSKI, R. (2008). Convergence to the viscous porous medium equation and propagation of chaos. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 185–203. MR2421181
[21] FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451
[22] HANNAH, L. A. and DUNSON, D. B. (2013). Multivariate convex regression with adaptive partitioning. *J. Mach. Learn. Res.* **14** 3261–3294. MR3144462
[23] JAVANMARD, A., MONDELLI, M. and MONTANARI, A. (2020). Supplement to "Analysis of a two-layer neural network via displacement convexity." https://doi.org/10.1214/20-AOS1945SUPP.
[24] LI, Y. and YUAN, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems* 597–607.
[25] LIONS, P.-L. and SZNITMAN, A.-S. (1984). Stochastic differential equations with reflecting boundary conditions. *Comm. Pure Appl. Math.* **37** 511–537. MR0745330 https://doi.org/10.1002/cpa.3160370408
[26] MCCANN, R. J. (1997). A convexity principle for interacting gases. *Adv. Math.* **128** 153–179. MR1451422 https://doi.org/10.1006/aima.1997.1634
[27] MEI, S., BAI, Y. and MONTANARI, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. MR3851754 https://doi.org/10.1214/17-AOS1637
[28] MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In *Conference on Learning Theory* (*COLT*).
[29] MEI, S., MONTANARI, A. and NGUYEN, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671. MR3845070 https://doi.org/10.1073/pnas.1806579115
[30] NITANDA, A. and SUZUKI, T. (2017). Stochastic particle gradient descent for infinite ensembles. arXiv:1712.05438.

[31] OELSCHLÄGER, K. (2002). Simulation of the solution of a viscous porous medium equation by a particle method. *SIAM J. Numer. Anal.* **40** 1716–1762. MR1950620 https://doi.org/10.1137/S0036142900363377

[32] PARK, J. and SANDBERG, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Comput.* **3** 246–257.

[33] PHILIPOWSKI, R. (2007). Interacting diffusions approximating the porous medium equation and propagation of chaos. *Stochastic Process. Appl.* **117** 526–538. MR2305385 https://doi.org/10.1016/j.spa.2006.09.003

[34] RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* 1177–1184.

[35] ROSENBLATT, F. (1962). *Principles of Neurodynamics*. Spartan Book.

[36] ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. In *Advances in Neural Information Processing Systems*.

[37] SANTAMBROGIO, F. (2015). *Optimal Transport for Applied Mathematicians*: *Calculus of Variations*, *PDEs*, *and Modeling*. *Progress in Nonlinear Differential Equations and Their Applications* **87**. Birkhäuser/Springer, Cham. MR3409718 https://doi.org/10.1007/978-3-319-20828-2

[38] SCHAPIRE, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear Estimation and Classification* (*Berkeley*, *CA*, 2001). *Lect. Notes Stat.* **171** 149–171. Springer, New York. MR2005788 https://doi.org/10.1007/978-0-387-21579-2_9

[39] SIRIGNANO, J. and SPILIOPOULOS, K. (2018). Mean field analysis of neural networks. arXiv:1805.01053.

[40] SOLTANOLKOTABI, M., JAVANMARD, A. and LEE, J. D. (2019). Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Trans. Inform. Theory* **65** 742–769. MR3904911 https://doi.org/10.1109/TIT.2018.2854560

[41] SZNITMAN, A.-S. (1991). Topics in propagation of chaos. In *École D'Été de Probabilités de Saint-Flour XIX*—1989. *Lecture Notes in Math.* **1464** 165–251. Springer, Berlin. MR1108185 https://doi.org/10.1007/BFb0085169

[42] TANAKA, H. (1979). Stochastic differential equations with reflecting boundary condition in convex regions. *Hiroshima Math. J.* **9** 163–177. MR0529332

[43] THOMAS, J. W. (2013). *Numerical Partial Differential Equations*: *Finite Difference Methods* **22**. Springer Science & Business Media. MR1367964 https://doi.org/10.1007/978-1-4899-7278-1

[44] TIAN, Y. (2017). Symmetry-breaking convergence analysis of certain two-layered neural networks with ReLU nonlinearity. In *Workshop at International Conference on Learning Representation* (*ICLR*).

[45] VÁZQUEZ, J. L. (2007). *The Porous Medium Equation*: *Mathematical Theory*. *Oxford Mathematical Monographs*. Oxford University Press, Oxford. MR2286292

[46] VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[47] WEI, C., LEE, J. D., LIU, Q. and MA, T. (2018). On the margin theory of feedforward neural networks. arXiv:1810.05369.

[48] ZHONG, K., SONG, Z., JAIN, P., BARTLETT, P. L. and DHILLON, I. S. (2017). Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning* 4140–4149.