# A mean field view of the landscape of two-layer neural networks

**Song Mei**[a], **Andrea Montanari**[b,c,1], **and Phan-Minh Nguyen**[b]

[a]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305; [b]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; and [c]Department of Statistics, Stanford University, Stanford, CA 94305

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved June 21, 2018 (received for review April 16, 2018)

**Multilayer neural networks are among the most powerful models in machine learning, yet the fundamental reasons for this success defy mathematical understanding. Learning a neural network requires optimizing a nonconvex high-dimensional objective (risk function), a problem that is usually attacked using stochastic gradient descent (SGD). Does SGD converge to a global optimum of the risk or only to a local optimum? In the former case, does this happen because local minima are absent or because SGD somehow avoids them? In the latter, why do local minima reached by SGD have good generalization properties? In this paper, we consider a simple case, namely two-layer neural networks, and prove that—in a suitable scaling limit—SGD dynamics is captured by a certain nonlinear partial differential equation (PDE) that we call distributional dynamics (DD). We then consider several specific examples and show how DD can be used to prove convergence of SGD to networks with nearly ideal generalization error. This description allows for "averaging out" some of the complexities of the landscape of neural networks and can be used to prove a general convergence result for noisy SGD.**

neural networks | stochastic gradient descent | gradient flow | Wasserstein space | partial differential equations

Multilayer neural networks are one of the oldest approaches to statistical machine learning, dating back at least to the 1960s (1). Over the last 10 years, under the impulse of increasing computer power and larger data availability, they have emerged as a powerful tool for a wide variety of learning tasks (2, 3).

In this paper, we focus on the classical setting of supervised learning, whereby we are given data points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, indexed by $i \in \mathbb{N}$, which are assumed to be independent and identically distributed from an unknown distribution $\mathbb{P}$ on $\mathbb{R}^d \times \mathbb{R}$. Here $\boldsymbol{x}_i \in \mathbb{R}^d$ is a feature vector (e.g., a set of descriptors of an image), and $y_i \in \mathbb{R}$ is a label (e.g., labeling the object in the image). Our objective is to model the dependence of the label $y_i$ on the feature vector $\boldsymbol{x}_i$ to assign labels to previously unlabeled examples. In a two-layer neural network, this dependence is modeled as

$$\hat{y}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i). \qquad [1]$$

Here, $N$ is the number of hidden units (neurons), $\sigma_* : \mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$ is an activation function, and $\boldsymbol{\theta}_i \in \mathbb{R}^D$ are parameters, which we collectively denote by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$. The factor $(1/N)$ is introduced for convenience and can be eliminated by redefining the activation. Often $\boldsymbol{\theta}_i = (a_i, b_i, \boldsymbol{w}_i)$ and

$$\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i) = a_i\, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle + b_i), \qquad [2]$$

for some $\sigma : \mathbb{R} \to \mathbb{R}$. Ideally, the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i \le N}$ should be chosen as to minimize the risk (generalization error) $R_N(\boldsymbol{\theta}) = \mathbb{E}\{\ell(y, \hat{y}(\boldsymbol{x}; \boldsymbol{\theta}))\}$, where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a certain loss function. For the sake of simplicity, we will focus on the square loss $\ell(y, \hat{y}) = (y - \hat{y})^2$, but more general choices can be treated along the same lines.

In practice, the parameters of neural networks are learned by stochastic gradient descent (SGD) (4) or its variants. In the present case, this amounts to the iteration

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + 2 s_k \left( y_k - \hat{y}(\boldsymbol{x}_k; \boldsymbol{\theta}^k) \right) \nabla_{\boldsymbol{\theta}_i} \sigma_*(\boldsymbol{x}_k; \boldsymbol{\theta}_i^k). \qquad [3]$$

Here $\boldsymbol{\theta}^k = (\boldsymbol{\theta}_i^k)_{i \le N}$ denotes the parameters after $k$ iterations, $s_k$ is a step size, and $(\boldsymbol{x}_k, y_k)$ is the $k$th example. Throughout the paper, we make the following "One-Pass Assumption": Training examples are never revisited. Equivalently, $\{(\boldsymbol{x}_k, y_k)\}_{k \ge 1}$ are independent and identically distributed. $(\boldsymbol{x}_k, y_k) \sim \mathbb{P}$.

In large-scale applications, this is not far from truth: The data are so large that each example is visited at most a few times (5). Further, theoretical guarantees suggest that there is limited advantage to be gained from multiple passes (6). For recent work deriving scaling limits under such an assumption (in different problems), see ref. 7.

Understanding the optimization landscape of two-layer neural networks is largely an open problem even when we have access to an infinite number of examples—that is, to the population risk $R_N(\boldsymbol{\theta})$. Several studies have focused on special choices of the activation function $\sigma_*$ and of the data distribution $\mathbb{P}$, proving that the population risk has no bad local minima (8–10). This type of analysis requires delicate calculations that are somewhat sensitive to the specific choice of the model. Another line of work proposes new algorithms with theoretical guarantees (11–16), which use initializations based on tensor factorization.

In this paper, we prove that—in a suitable scaling limit—the SGD dynamics admits an asymptotic description in terms of a certain nonlinear partial differential equation (PDE). This PDE has a remarkable mathematical structure, in that it corresponds to a gradient flow in the metric space $(\mathscr{P}(\mathbb{R}^D), W_2)$: the space of probability measures on $\mathbb{R}^D$, endowed with the

## Significance

**Multilayer neural networks have proven extremely successful in a variety of tasks, from image classification to robotics. However, the reasons for this practical success and its precise domain of applicability are unknown. Learning a neural network from data requires solving a complex optimization problem with millions of variables. This is done by stochastic gradient descent (SGD) algorithms. We study the case of two-layer networks and derive a compact description of the SGD dynamics in terms of a limiting partial differential equation. Among other consequences, this shows that SGD dynamics does not become more complex when the network size increases.**

Author contributions: S.M., A.M., and P.-M.N. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: montanari@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1806579115/-/DCSupplemental.

Published online July 27, 2018.

www.pnas.org/cgi/doi/10.1073/pnas.1806579115

Wasserstein metric. This gradient flow minimizes an asymptotic version of the population risk, which is defined for $\rho \in \mathscr{P}(\mathbb{R}^D)$ and will be denoted by $R(\rho)$. This description simplifies the analysis of the landscape of two-layer neural networks, for instance by exploiting underlying symmetries. We illustrate this by obtaining results on several concrete examples as well as a general convergence result for "noisy SGD." In the next section, we provide an informal outline, focusing on basic intuitions rather than on formal results. We then present the consequences of these ideas on a few specific examples and subsequently state our general results.

## An Informal Overview

A good starting point is to rewrite the population risk $R_N(\boldsymbol{\theta}) = \mathbb{E}\{[y - \hat{y}(\boldsymbol{x};\boldsymbol{\theta})]^2\}$ as

$$R_N(\boldsymbol{\theta}) = R_{\#} + \frac{2}{N}\sum_{i=1}^{N} V(\boldsymbol{\theta}_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j), \quad \mathbf{[4]}$$

where we defined the potentials $V(\boldsymbol{\theta}) = -\mathbb{E}\{y\,\sigma_*(\boldsymbol{x};\boldsymbol{\theta})\}$, $U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}\{\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_1)\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_2)\}$. In particular, $U(\,\cdot\,,\,\cdot\,)$ is a symmetric positive semidefinite kernel. The constant $R_{\#} = \mathbb{E}\{y^2\}$ is the risk of the trivial predictor $\hat{y} = 0$.

Notice that $R_N(\boldsymbol{\theta})$ only depends on $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ through their empirical distribution $\hat{\rho}^{(N)} = N^{-1}\sum_{i=1}^{N}\delta_{\boldsymbol{\theta}_i}$. This suggests considering a risk function defined for $\rho \in \mathscr{P}(\mathbb{R}^D)$ [we denote by $\mathscr{P}(\Omega)$ the space of probability distributions on $\Omega$]:

$$R(\rho) = R_{\#} + 2\int V(\boldsymbol{\theta})\,\rho(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\,\rho(\mathrm{d}\boldsymbol{\theta}_1)\,\rho(\mathrm{d}\boldsymbol{\theta}_2). \quad \mathbf{[5]}$$

Formal relationships can be established between $R_N(\boldsymbol{\theta})$ and $R(\rho)$. For instance, under mild assumptions, $\inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) = \inf_{\rho} R(\rho) + O(1/N)$. We refer to the next sections for mathematical statements of this type.

Roughly speaking, $R(\rho)$ corresponds to the population risk when the number of hidden units goes to infinity, and the empirical distribution of parameters $\hat{\rho}^{(N)}$ converges to $\rho$. Since $U(\,\cdot\,,\,\cdot\,)$ is positive semidefinite, we obtain that the risk becomes convex in this limit. The fact that learning can be viewed as convex optimization in an infinite-dimensional space was indeed pointed out in the past (17, 18). Does this mean that the landscape of the population risk simplifies for large $N$ and descent algorithms will converge to a unique (or nearly unique) global optimum?

The answer to the latter question is generally negative, and a physics analogy can explain why. Think of $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ as the positions of $N$ particles in a $D$-dimensional space. When $N$ is large, the behavior of such a "gas" of particles is effectively described by a density $\rho_t(\boldsymbol{\theta})$ (with $t$ indexing time). However, not all "small" changes of this density profile can be realized in the actual physical dynamics: The dynamics conserves mass locally because particles cannot move discontinuously. For instance, if $\mathrm{supp}(\rho_t) = S_1 \cup S_2$ for two disjoint compact sets $S_1, S_2 \subseteq \mathbb{R}^D$ and all $t \in [t_1, t_2]$, then the total mass in each of these regions cannot change over time—that is, $\rho_t(S_1) = 1 - \rho_t(S_2)$ does not depend on $t \in [t_1, t_2]$.

We will prove that SGD is well approximated (in a precise quantitative sense described below) by a continuum dynamics that enforces this local mass conservation principle. Namely, assume that the step size in SGD is given by $s_k = \varepsilon\,\xi(k\varepsilon)$, for $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ a sufficiently regular function. Denoting by $\hat{\rho}^{(N)}_k = N^{-1}\sum_{i=1}^{N}\delta_{\boldsymbol{\theta}_i^k}$ the empirical distribution of parameters after $k$ SGD steps, we prove that

$$\hat{\rho}^{(N)}_{t/\varepsilon} \Rightarrow \rho_t, \quad \mathbf{[6]}$$

when $N \to \infty$, $\varepsilon \to 0$ (here $\Rightarrow$ denotes weak convergence). The asymptotic dynamics of $\rho_t$ is defined by the following PDE, which we shall refer to as distributional dynamics (DD):

$$\partial_t \rho_t = 2\xi(t)\,\nabla_{\boldsymbol{\theta}} \cdot \big(\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta};\rho_t)\big), \quad \mathbf{[7]}$$

and

$$\Psi(\boldsymbol{\theta};\rho) \equiv V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \boldsymbol{\theta}')\,\rho(\mathrm{d}\boldsymbol{\theta}'). \quad \mathbf{[8]}$$

[Here, $\nabla_{\boldsymbol{\theta}} \cdot \boldsymbol{v}(\boldsymbol{\theta})$ denotes the divergence of the vector field $\boldsymbol{v}(\boldsymbol{\theta})$.] This should be interpreted as an evolution equation in $\mathscr{P}(\mathbb{R}^D)$. While we described the convergence to this dynamics in asymptotic terms, the results in the next sections provide explicit nonasymptotic bounds. In particular, $\rho_t$ is a good approximation of $\hat{\rho}^{(N)}_k$, $k = t/\varepsilon$, as soon as $\varepsilon \ll 1/D$ and $N \gg D$.

Using these results, analyzing learning in two-layer neural networks reduces to analyzing the PDE (Eq. **7**). While this is far from being an easy task, the PDE formulation leads to several simplifications and insights. First, it factors out the invariance of the risk (Eq. **4**) (and of the SGD dynamics; Eq. **3**), with respect to permutations of the units $\{1, \dots, N\}$.

Second, it allows us to exploit symmetries in the data distribution $\mathbb{P}$. If $\mathbb{P}$ is left invariant under a group of transformations (e.g., rotations), we can look for a solution $\rho_t$ of the DD (Eq. **7**) that enjoys the same symmetry, hence reducing the dimensionality of the problem. This is impossible for the finite-$N$ dynamics (Eq. **3**), since no arrangement of the points $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N\} \subseteq \mathbb{R}^D$ is left invariant, say, under rotations. We will provide examples of this approach in the next sections.

Third, there is rich mathematical literature on the PDE (Eq. **7**) that was motivated by the study of interacting particle systems in mathematical physics. As mentioned above, a key structure exploited in this line of work is that Eq. **7** can be viewed as a gradient flow for the cost function $R(\rho)$ in the space $(\mathscr{P}(\mathbb{R}^D), W_2)$, of probability measures on $\mathbb{R}^D$ endowed with the Wasserstein metric (19–21). Roughly speaking, this means that the trajectory $t \mapsto \rho_t$ attempts to minimize the risk $R(\rho)$ while maintaining the "local mass conservation" constraint. Recall that the Wasserstein distance is defined as

$$W_2(\rho_1, \rho_2) = \left(\inf_{\gamma \in \mathcal{C}(\rho_1, \rho_2)} \int \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2\,\gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2)\right)^{1/2}, \quad \mathbf{[9]}$$

where the infimum is taken over all couplings of $\rho_1$ and $\rho_2$. Informally, the fact that $\rho_t$ is a gradient flow means that Eq. **7** is equivalent, for small $\tau$, to

$$\rho_{t+\tau} \approx \arg\min_{\rho \in \mathscr{P}(\mathbb{R}^D)} \left\{ R(\rho) + \frac{1}{2\xi(t)\tau} W_2(\rho, \rho_t)^2 \right\}. \quad \mathbf{[10]}$$

Powerful tools from the mathematical literature on gradient flows in measure spaces (20) can be exploited to study the behavior of Eq. **7**.

Most importantly, the scaling limit elucidates the dependence of the landscape of two-layer neural networks on the number of hidden units $N$.

A remarkable feature of neural networks is the observation that, while they might be dramatically overparametrized, this does not lead to performance degradation. In the case of bounded activation functions, this phenomenon was clarified in the 1990s for empirical risk minimization algorithms (see, e.g., ref. 22). The present work provides analogous insight for the SGD dynamics: Roughly speaking, our results imply that the landscape remains essentially unchanged as $N$ grows, provided $N \gg D$. In particular, assume that the PDE (Eq. **7**) converges close to an optimum in time $t_*(D)$. This might depend on $D$ but does not depend on the number of hidden units $N$ (which does not appear in the DD PDE; Eq. **7**). If $t_*(D) = O_D(1)$, we can then take $N$ arbitrarily (as long as $N \gg D$) and will achieve a population risk that is independent of $N$ (and corresponds to the optimum), using $k = t_*/\varepsilon = O(D)$ samples.

Our analysis can accommodate some important variants of SGD, a particularly interesting one being noisy SGD:

$$\boldsymbol{\theta}_i^{k+1} = (1 - 2\lambda s_k)\boldsymbol{\theta}_i^k + 2s_k\ (y_k - \hat{y}_k)\,\nabla_{\boldsymbol{\theta}_i}\sigma_*(\boldsymbol{x}_k;\boldsymbol{\theta}_i^k) + \sqrt{2s_k/\beta}\,\boldsymbol{g}_i^k, \qquad [11]$$

where $\boldsymbol{g}_i^k \sim \mathsf{N}(0, \boldsymbol{I}_D)$ and $\hat{y}_k = \hat{y}(\boldsymbol{x}_k;\boldsymbol{\theta}^k)$. (The term $-2\lambda s_k \boldsymbol{\theta}_i^k$ corresponds to an $\ell_2$ regularization and will be useful for our analysis below.) The resulting scaling limit differs from Eq. **7** by the addition of a diffusion term:

$$\partial_t \rho_t = 2\xi(t)\,\nabla_{\boldsymbol{\theta}}\cdot(\rho_t \nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_t)) + 2\xi(t)\beta^{-1}\Delta_{\boldsymbol{\theta}}\rho_t, \qquad [12]$$

where $\Psi_\lambda(\boldsymbol{\theta};\rho) = \Psi(\boldsymbol{\theta};\rho) + (\lambda/2)\|\boldsymbol{\theta}\|_2^2$, and $\Delta_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \sum_{i=1}^d \partial_{\theta_i}^2 f(\boldsymbol{\theta})$ denotes the usual Laplacian. This can be viewed as a gradient flow for the free-energy $F_{\beta,\lambda}(\rho) = (1/2)R(\rho) + (\lambda/2)\int\|\boldsymbol{\theta}\|_2^2\rho(\mathrm{d}\boldsymbol{\theta}) - \beta^{-1}\mathrm{Ent}(\rho)$, where $\mathrm{Ent}(\rho) = -\int\rho(\boldsymbol{\theta})\log\rho(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is the entropy of $\rho$ [by definition $\mathrm{Ent}(\rho) = -\infty$ if $\rho$ is singular]. $F_{\beta,\lambda}(\rho)$ is an entropy-regularized risk, which penalizes strongly nonuniform $\rho$.

We will prove below that, for $\beta < \infty$, the evolution (Eq. **12**) generically converges to the minimizer of $F_{\beta,\lambda}(\rho)$, hence implying global convergence of noisy SGD in a number of steps independent of $N$.

## Examples

In this section, we discuss some simple applications of the general approach outlined above. Let us emphasize that these examples are not realistic. First, the data distribution $\mathbb{P}$ is extremely simple: We made this choice to be able to carry out explicit calculations. Second, the activation function $\sigma_*(\boldsymbol{x};\boldsymbol{\theta})$ is not necessarily optimal: We made this choice to illustrate some interesting phenomena.

**Centered Isotropic Gaussians.** One-neuron neural networks perform well with (nearly) linearly separable data. The simplest classification problem that requires multilayer networks is, arguably, the one of distinguishing two Gaussians with the same mean. Assume the joint law $\mathbb{P}$ of $(y, \boldsymbol{x})$ to be as follows:

with probability $1/2$: $y = +1$, $\boldsymbol{x} \sim \mathsf{N}(0, (1+\Delta)^2\boldsymbol{I}_d)$; and
with probability $1/2$: $y = -1$, $\boldsymbol{x} \sim \mathsf{N}(0, (1-\Delta)^2\boldsymbol{I}_d)$.

(This example will be generalized later.) Of course, optimal classification in this model becomes entirely trivial if we compute the feature $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$. However, it is nontrivial that an SGD-trained neural network will succeed.

We choose an activation function without offset or output weights, namely $\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i) = \sigma(\langle\boldsymbol{w}_i, \boldsymbol{x}\rangle)$. While qualitatively similar results are obtained for other choices of $\sigma$, we will use a simple piecewise linear function as a running example: $\sigma(t) = s_1$ if $t \le t_1$, $\sigma(t) = s_2$ if $t \ge t_2$, and $\sigma(t)$ interpolated linearly for $t \in (t_1, t_2)$. In simulations, we use $t_1 = 0.5$, $t_2 = 1.5$, $s_1 = -2.5$, and $s_2 = 7.5$.

We run SGD with initial weights $(\boldsymbol{w}_i^0)_{i \le N} \sim_{iid} \rho_0$, where $\rho_0$ is spherically symmetric. Fig. 1 reports the result of such an experiment. Due to the symmetry of the distribution $\mathbb{P}$, the distribution $\rho_t$ remains spherically symmetric for all $t$ and hence is completely determined by the distribution $\overline{\rho}_t$ of the norm $r = \|\boldsymbol{w}\|_2$. This distribution satisfies a one-dimensional reduced DD:

$$\partial_t \overline{\rho}_t = 2\xi(t)\,\partial_r\left(\overline{\rho}_t \partial_r \psi(r;\overline{\rho}_t)\right), \qquad [13]$$

where the form of $\psi(r;\rho)$ can be derived from $\Psi(\boldsymbol{\theta};\rho)$. This reduced PDE can be efficiently solved numerically, see *SI Appendix* for technical details. As illustrated by Fig. 1, the empirical results match closely the predictions produced by this PDE.

In Fig. 2, we compare the asymptotic risk achieved by SGD with the prediction obtained by minimizing $R(\rho)$ (cf., Eq. **5**) over spherically symmetric distributions. It turns out that, for certain values of $\Delta$, the minimum is achieved by the uniform distribution over a sphere of radius $\|\boldsymbol{w}\|_2 = r_*$, to be denoted by $\rho_{r_*}^{\mathrm{unif}}$. The value of $r_*$ is computed by minimizing

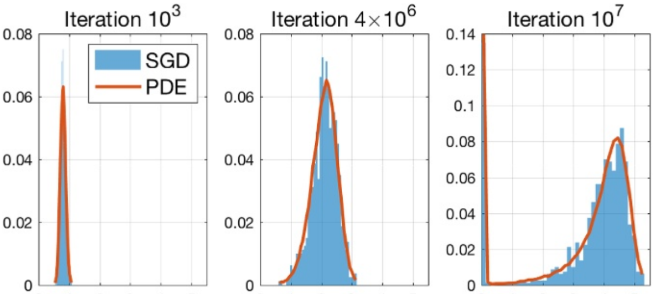

**Fig. 1.** Evolution of the radial distribution $\overline{\rho}_t$ for the isotropic Gaussian model, with $\Delta = 0.8$. Histograms are obtained from SGD experiments with $d = 40$, $N = 800$, initial weight distribution $\rho_0 = \mathsf{N}(\mathbf{0}, 0.8^2/d \cdot \boldsymbol{I}_d)$, and step size $\epsilon = 10^{-6}$ and $\xi(t) = 1$. Continuous lines correspond to a numerical solution of the DD (Eq. **13**).

$$\overline{R}_d^{(1)}(r) = 1 + 2v(r) + u_d(r, r), \qquad [14]$$

where expressions for $v(r)$, $u_d(r_1, r_2)$ can be readily derived from $V(\boldsymbol{w})$, $U(\boldsymbol{w}_1, \boldsymbol{w}_2)$ and are given in *SI Appendix*.

**Lemma 1:** *Let $r_*$ be a global minimizer of $r \mapsto R_d^{(1)}(r)$. Then $\rho_{r_*}^{\mathrm{unif}}$ is a global minimizer of $\rho \mapsto R(\rho)$ if and only if $v(r) + u_d(r, r_*) \ge v(r_*) + u_d(r_*, r_*)$ for all $r \ge 0$.*

Checking numerically, this condition yields that $\rho_{r_*}^{\mathrm{unif}}$ is a global minimizer for $\Delta$ in an interval $[\Delta_d^{\mathrm{l}}, \Delta_d^{\mathrm{h}}]$, where $\lim_{d\to\infty}\Delta_d^{\mathrm{l}} = 0$ and $\lim_{d\to\infty}\Delta_d^{\mathrm{h}} = \Delta_\infty \approx 0.47$.

Fig. 2 shows good quantitative agreement between empirical results and theoretical predictions and suggests that SGD achieves a value of the risk that is close to optimum. Can we prove that this is indeed the case and that the SGD dynamics does not get stuck in local minima? It turns out that we can use our general theory (see next section) to prove that this is the case for large $d$. To state this result, we need to introduce a class of good uninformative initializations $\mathscr{P}_{\mathrm{good}} \subseteq \mathscr{P}(\mathbb{R}_{\ge 0})$ for which convergence to the optimum takes place. For $\overline{\rho} \in \mathscr{P}(\mathbb{R}_{\ge 0})$, we let $\overline{R}_d(\overline{\rho}) \equiv R(\overline{\rho} \times \mathrm{Unif}(\mathbb{S}^{d-1}))$. This risk has a well-defined limit as $d \to \infty$. We say that $\overline{\rho} \in \mathscr{P}_{\mathrm{good}}$ if $(i)$ $\overline{\rho}$ is absolutely continuous with respect to the Lebesgue measure, with bounded density, $(ii)$ $\overline{R}_\infty(\overline{\rho}) < 1$.

**Theorem 1:** *For any $\eta, \Delta, \delta > 0$ and $\overline{\rho}_0 \in \mathscr{P}_{\mathrm{good}}$, there exists $d_0 = d_0(\eta, \overline{\rho}_0, \Delta)$, $T = T(\eta, \overline{\rho}_0, \Delta)$, and $C_0 = C_0(\eta, \overline{\rho}_0, \Delta, \delta)$, such that the following holds for the problem of classifying isotropic Gaussians. For any dimension $d \ge d_0$, number of neurons $N \ge C_0 d$, consider SGD initialized with $(\boldsymbol{w}_i^0)_{i \le N} \sim_{iid} \overline{\rho}_0 \times \mathrm{Unif}(\mathbb{S}^{d-1})$ and step size $\varepsilon \in [1/N^{10}, 1/(C_0 d)]$. Then we have*

$$R_N(\boldsymbol{\theta}^k) \le \inf_{\boldsymbol{\theta} \in \mathbb{R}^{N \times d}} R(\boldsymbol{\theta}) + \eta \qquad [15]$$

*for any $k \in [T/\varepsilon, 10T/\varepsilon]$ with probability at least $1 - \delta$.*

In particular, if we set $\varepsilon = 1/(C_0 d)$, then the number of SGD steps is $k \in [(C_0 T)\,d, (10 C_0 T)\,d]$: The number of samples used by SGD does not depend on the number of hidden units $N$ and is only linear in the dimension. Unfortunately, the proof does not provide the dependence of $T$ on $\eta$, but Theorem 6 below suggests exponential local convergence.

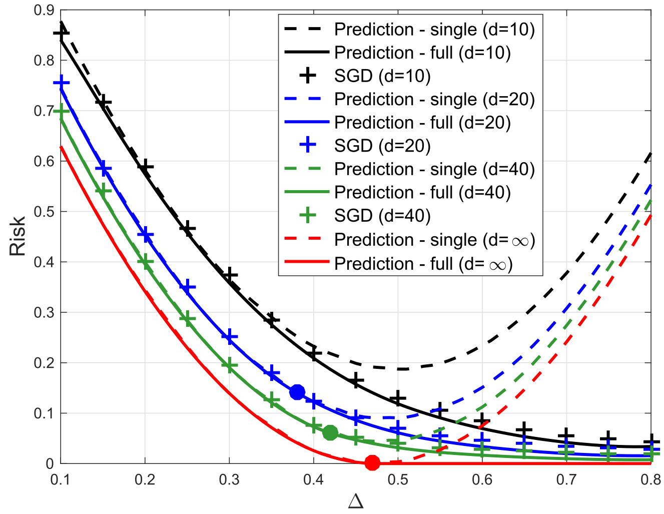

**Fig. 2.** Population risk in the problem of separating two isotropic Gaussians, as a function of the separation parameter $\Delta$. We use a two-layer network with piecewise linear activation, no offset, and output weights equal to 1. Empirical results obtained by SGD (a single run per data point) are marked "+." Continuous lines are theoretical predictions obtained by numerically minimizing $R(\rho)$ (see *SI Appendix* for details). Dashed lines are theoretical predictions from the single-delta ansatz of Eq. **14**. Notice that this ansatz is incorrect for $\Delta > \Delta_d^{\rm h}$, which is marked as a solid round dot. Here, $N = 800$.

While we stated Theorem 1 for the piecewise linear sigmoids, *SI Appendix* presents technical conditions under which it holds for a general monotone function $\sigma : \mathbb{R} \to \mathbb{R}$.

**Centered Anisotropic Gaussians.** We can generalize the previous result to a problem in which the network needs to select a subset of relevant nonlinear features out of many a priori equivalent ones. We assume the joint law of $(y, \boldsymbol{x})$ to be as follows:

with probability $1/2$: $y = +1$, $\boldsymbol{x} \sim {\sf N}(0, \boldsymbol{\Sigma}_+)$; and
with probability $1/2$: $y = -1$, $\boldsymbol{x} \sim {\sf N}(0, \boldsymbol{\Sigma}_-)$.

Given a linear subspace $\mathcal{V} \subseteq \mathbb{R}^d$ of dimension $s_0 \le d$, we assume that $\boldsymbol{\Sigma}_+$, $\boldsymbol{\Sigma}_-$ differ uniquely along $\mathcal{V}$: $\boldsymbol{\Sigma}_\pm = \boldsymbol{I}_d + (\tau_\pm^2 - 1)\boldsymbol{P}_\mathcal{V}$, where $\tau_\pm = (1 \pm \Delta)$ and $\boldsymbol{P}_\mathcal{V}$ is the orthogonal projector onto $\mathcal{V}$. In other words, the projection of $\boldsymbol{x}$ on the subspace $\mathcal{V}$ is distributed according to an isotropic Gaussian with variance $\tau_+^2$ (if $y = +1$) or $\tau_-^2$ (if $y = -1$). The projection orthogonal to $\mathcal{V}$ has instead the same variance in the two classes. A successful classifier must be able to learn the relevant subspace $\mathcal{V}$. We assume the same class of activations $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$ as for the isotropic case.

The distribution $\mathbb{P}$ is invariant under a reduced symmetry group $\mathcal{O}(s_0) \times \mathcal{O}(d - s_0)$. As a consequence, letting $r_1 = \|\boldsymbol{P}_\mathcal{V} \boldsymbol{w}\|_2$ and $r_2 \equiv \|(\boldsymbol{I}_d - \boldsymbol{P}_\mathcal{V})\boldsymbol{w}\|_2$, it is sufficient to consider distributions $\rho$ that are uniform, conditional on the values of $r_1$ and $r_2$. If we initialize $\rho_0$ to be uniform conditional on $(r_1, r_2)$, this property is preserved by the evolution (Eq. **7**). As in the isotropic case, we can use our general theory to prove convergence to a near-optimum if $d$ is large enough.

**Theorem 2:** *For any $\eta, \Delta, \delta > 0$ and $\overline{\rho}_0 \in \mathscr{P}_{\rm good}$, there exists $d_0 = d_0(\eta, \overline{\rho}_0, \Delta, \gamma)$, $T = T(\eta, \overline{\rho}_0, \Delta, \gamma)$, and $C_0 = C_0(\eta, \overline{\rho}_0, \Delta, \delta, \gamma)$, such that the following holds for the problem of classifying anisotropic Gaussians with $s_0 = \gamma d$, $\gamma \in (0, 1)$ fixed. For any dimension parameters $s_0 = \gamma d \ge d_0$, number of neurons $N \ge C_0 d$, consider SGD initialized with initialization $(\boldsymbol{w}_i^0)_{i \le N} \sim_{iid} \overline{\rho}_0 \times {\rm Unif}(\mathbb{S}^{d-1})$ and step size $\varepsilon \in [1/N^{10}, 1/(C_0 d)]$. Then, we have $R_N(\boldsymbol{\theta}^k) \le \inf_{\boldsymbol{\theta} \in \mathbb{R}^{N \times d}} R_N(\boldsymbol{\theta}) + \eta$ for any $k \in [T/\varepsilon, 10T/\varepsilon]$ with probability at least $1 - \delta$.*

Even with a reduced degree of symmetry, SGD converges to a network with nearly optimal risk, after using a number of samples $k = O(d)$, which is independent of the number of hidden units $N$.

**A Better Activation Function.** Our previous examples use activation functions $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$ without output weights or offset to simplify the analysis and illustrate some interesting phenomena. Here we consider instead a standard rectified linear unit (ReLU) activation and fit both the output weight and the offset: $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}) = a\,\sigma_{\rm ReLU}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)$, where $\sigma_{\rm ReLU}(x) = \max(x, 0)$. Hence, $\boldsymbol{\theta} = (\boldsymbol{w}, a, b) \in \mathbb{R}^{d+2}$.

We consider the same data distribution introduced in the last section (anisotropic Gaussians). Fig. 3 reports the evolution of the risk $R_N(\boldsymbol{\theta}^k)$ for three experiments with $d = 320$, $s_0 = 60$, and different values of $\Delta$. SGD is initialized by setting $a_i = 1$, $b_i = 1$, and $\boldsymbol{w}_i^0 \sim_{iid} {\sf N}(\boldsymbol{0}, 0.8^2/d \cdot \boldsymbol{I}_d)$ for $i \le N$. We observe that SGD converges to a network with very small risk, but this convergence has a nontrivial structure and presents long flat regions.

The empirical results are well captured by our predictions based on the continuum limit. In this case, we obtain a reduced PDE for the joint distribution of the four quantities $\boldsymbol{r} = (a, b, r_1 = \|\boldsymbol{P}_\mathcal{V} \boldsymbol{w}\|_2, r_2 = \|\boldsymbol{P}_\mathcal{V}^\perp \boldsymbol{w}\|_2)$, denoted by $\overline{\rho}_t$. The reduced PDE is analogous to Eq. **13** albeit in 4 dimensions rather than 1 dimension. In Fig. 3, we consider the evolution of the risk, alongside three properties of the distribution $\overline{\rho}_t$—the means of the output weight $a$, of the offset $b$, and of $r_1$.

**Predicting Failure.** SGD does not always converge to a near global optimum. Our analysis allows us to construct examples in which SGD fails. For instance, Fig. 4 reports results for the isotropic Gaussians problem. We violate the assumptions of Theorem 1 by using nonmonotone activation function. Namely, we use $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$, where $\sigma(t) = -2.5$ for $t \le 0$, $\sigma(t) = 7.5$ for $t \ge 1.5$, and $\sigma(t)$ linearly interpolates from $(0, -2.5)$ to $(0.5, -4)$, and from $(0.5, -4)$ to $(1.5, 7.5)$.

Depending on the initialization, SGD converges to two different limits, one with a small risk and the second with high risk. Again, this behavior is well tracked by solving a one-dimensional PDE for the distribution $\overline{\rho}_t$ of $r = \|\boldsymbol{w}\|_2$.

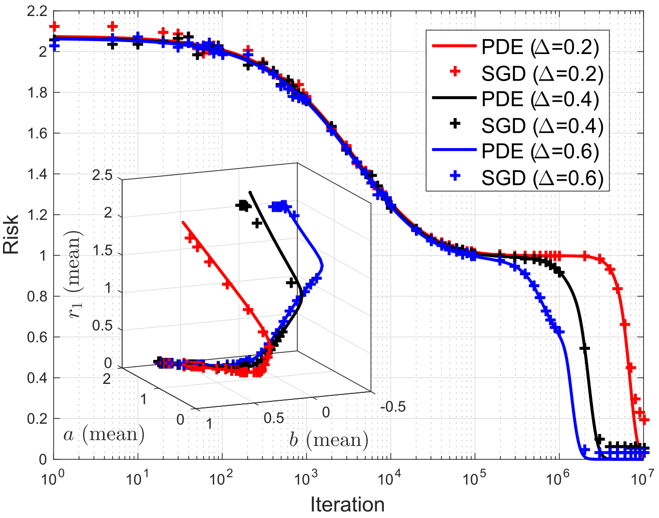

**Fig. 3.** Evolution of the population risk for the variable selection problem using a two-layer neural network with ReLU activations. Here $d = 320$, $s_0 = 60$, and $N = 800$, and we used $\xi(t) = t^{-1/4}$ and $\varepsilon = 2 \times 10^{-4}$ to set the step size. Numerical simulations using SGD (one run per data point) are marked +, and curves are solutions of the reduced PDE with $d = \infty$. (*Inset*) Evolution of three parameters of the reduced distribution $\overline{\rho}_t$ (average output weights $a$, average offsets $b$, and average $\ell_2$ norm in the relevant subspace $r_1$) for the same setting.

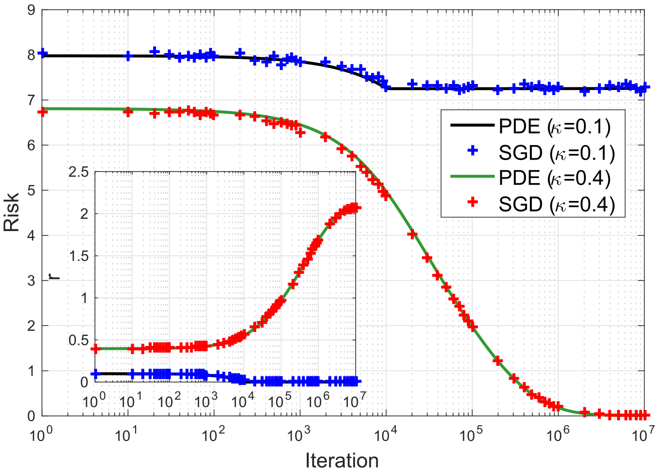

**Fig. 4.** Separating two isotropic Gaussians, with a nonmonotone activation function (see *Predicting Failure* for details). Here $N = 800$, $d = 320$, and $\Delta = 0.5$. The main frame presents the evolution of the population risk along the SGD trajectory, starting from two different initializations of $(\boldsymbol{w}_i^0)_{i \le N} \sim_{iid} \mathsf{N}(\mathbf{0}, \kappa^2/d \cdot \mathbf{I}_d)$ for either $\kappa = 0.1$ or $\kappa = 0.4$. In *Inset*, we plot the evolution of the average of $\|\boldsymbol{w}\|_2$ for the same conditions. Symbols are empirical results. Continuous lines are predictions obtained with the reduced PDE (Eq. **13**).

## General Results

In this section, we return to the general supervised learning problem described in the Introduction and describe our general results. Proofs are deferred to *SI Appendix*.

First, we note that the minimum of the asymptotic risk $R(\rho)$ of Eq. **5** provides a good approximation of the minimum of the finite-$N$ risk $R_N(\boldsymbol{\theta})$.

**Proposition 1:** *Assume that either one of the following conditions hold*: $(a)$ $\inf_\rho R(\rho)$ *is achieved by a distribution* $\rho_*$ *such that* $\int U(\boldsymbol{\theta}, \boldsymbol{\theta})\, \rho_*(\mathrm{d}\boldsymbol{\theta}) \le K$; $(b)$ *There exists* $\varepsilon_0 > 0$ *such that, for any* $\rho \in \mathscr{P}(\mathbb{R}^D)$ *such that* $R(\rho) \le \inf_\rho R(\rho) + \varepsilon_0$, *we have* $\int U(\boldsymbol{\theta}, \boldsymbol{\theta})\, \rho(\mathrm{d}\boldsymbol{\theta}) \le K$. *Then*

$$\left| \inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) - \inf_{\rho} R(\rho) \right| \le K/N. \qquad \textbf{[16]}$$

*Further, assume that* $\boldsymbol{\theta} \mapsto V(\boldsymbol{\theta})$ *and* $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \mapsto U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ *are continuous, with* $U$ *bounded below. A probability measure* $\rho_*$ *is a global minimum of* $R$ *if* $\inf_{\boldsymbol{\theta} \in \mathbb{R}^D} \Psi(\boldsymbol{\theta}; \rho_*) > -\infty$ *and*

$$\mathrm{supp}(\rho_*) \subseteq \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^D} \Psi(\boldsymbol{\theta}; \rho_*). \qquad \textbf{[17]}$$

We next consider the DDs (Eqs. **7** and **12**). These should be interpreted to hold in a weak sense (cf. *SI Appendix*). To establish that these PDEs indeed describe the limit of the SGD dynamics, we make the following assumptions:

A1. $t \mapsto \xi(t)$ is bounded Lipschitz: $\|\xi\|_\infty, \|\xi\|_{\mathrm{Lip}} \le K_1$, with $\int_0^\infty \xi(t) \mathrm{d}t = \infty$.

A2. The activation function $(\boldsymbol{x}, \boldsymbol{\theta}) \mapsto \sigma_*(\boldsymbol{x}; \boldsymbol{\theta})$ is bounded, with sub-Gaussian gradient: $\|\sigma_*\|_\infty \le K_2$, $\|\nabla_{\boldsymbol{\theta}} \sigma_*(\boldsymbol{X}; \boldsymbol{\theta})\|_{\psi_2} \le K_2$. Labels are bounded $|y_k| \le K_2$.

A3. The gradients $\boldsymbol{\theta} \mapsto \nabla V(\boldsymbol{\theta})$, $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \mapsto \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are bounded, Lipschitz continuous [namely $\|\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta})\|_2$, $\|\nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\|_2 \le K_3$, $\|\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}')\|_2 \le K_3 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, $\|\nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)\|_2 \le K_3 \|(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)\|_2$].

We also introduce the following error term that quantifies in a nonasymptotic sense the accuracy of our PDE model:

$$\mathsf{err}_{N,D}(z) \equiv \sqrt{1/N \vee \varepsilon} \cdot \left[ \sqrt{D + \log(N/\varepsilon)} + z \right]. \qquad \textbf{[18]}$$

The convergence of the SGD process to the PDE model is an example of a phenomenon that is known in probability theory as propagation of chaos (23).

**Theorem 3:** *Assume that conditions* **A1**, **A2**, **A3** *hold. For* $\rho_0 \in \mathscr{P}(\mathbb{R}^D)$, *consider SGD with initialization* $(\boldsymbol{\theta}_i^0)_{i \le N} \sim_{iid} \rho_0$ *and step size* $s_k = \varepsilon \xi(k\varepsilon)$. *For* $t \ge 0$, *let* $\rho_t$ *be the solution of PDE* (Eq. **7**). *Then, for any fixed* $k$, $\hat{\rho}_k^{(N)} \Rightarrow \rho_{k\varepsilon}$ *almost surely along any sequence* $(N, \varepsilon = \varepsilon_N)$ *such that* $N/\log(1/\varepsilon_N) \to \infty$, $\varepsilon_N \to 0$. *Further, there exists a constant* $C$ *(depending uniquely on the parameters* $K_i$ *of conditions* **A1–A3**) *such that, for any* $f : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$, *with* $\|f\|_\infty, \|f\|_{\mathrm{Lip}} \le 1$, $\varepsilon \le 1$,

$$\sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \left| \frac{1}{N} \sum_{i=1}^N f(\boldsymbol{\theta}_i^k) - \int f(\boldsymbol{\theta}) \rho_{k\varepsilon}(\mathrm{d}\boldsymbol{\theta}) \right| \le C e^{CT} \mathsf{err}_{N,D}(z),$$

$$\sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \left| R_N(\boldsymbol{\theta}^k) - R(\rho_{k\varepsilon}) \right| \le C e^{CT} \mathsf{err}_{N,D}(z), \qquad \textbf{[19]}$$

*with probability* $1 - e^{-z^2}$. *The same statements hold for noisy SGD* (Eq. **11**), *provided* Eq. **7** *is replaced by* Eq. **12**, *and if* $\beta \ge 1$, $\lambda \le 1$, *and* $\rho_0$ *is* $K_0$ *sub-Gaussian for some* $K_0 > 0$.

Notice that dependence of the error terms in $N$ and $D$ is rather benign. On the other hand, the error grows exponentially with the time horizon $T$, which limits its applicability to cases in which the DD converges rapidly to a good solution. We do not expect this behavior to be improvable within the general setting of 0.3, which a priori includes cases in which the dynamics is unstable.

We can regard $\boldsymbol{J}(\boldsymbol{\theta}; \rho_t) = \rho_t(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)$ as a current. The fixed points of the continuum dynamics are densities that correspond to zero current, as stated below.

**Proposition 2:** *Assume* $V(\cdot)$, $U(\cdot, \cdot)$ *to be differentiable with bounded gradient. If* $\rho_t$ *is a solution of the PDE* (Eq. **7**), *then* $R(\rho_t)$ *is nonincreasing. Further, probability distribution* $\rho$ *is a fixed point of the PDE* (Eq. **7**) *if and only if*

$$\mathrm{supp}(\rho) \subseteq \{ \boldsymbol{\theta} : \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho) = \mathbf{0} \}. \qquad \textbf{[20]}$$

Note that global optimizers of $R(\rho)$, defined by condition (Eq. **17**), are fixed points, but the set of fixed points is, in general, larger than the set of optimizers. Our next proposition provides an analogous characterization of the fixed points of diffusion DD (Eq. **12**) (see ref. 21 for related results).

**Proposition 3:** *Assume that conditions* **A1–A3** *hold and that* $\rho_0$ *is absolutely continuous with respect to the Lebesgue measure, with* $F_{\beta,\lambda}(\rho_0) < \infty$. *If* $(\rho_t)_{t \ge 0}$ *is a solution of the diffusion PDE* (Eq. **12**), *then* $\rho_t$ *is absolutely continuous. Further, there is at most one fixed point* $\rho_* = \rho_*^{\beta,\lambda}$ *of* Eq. **12** *satisfying* $F_{\beta,\lambda}(\rho_*) < \infty$. *This fixed point is absolutely continuous and its density satisfies*

$$\rho_*(\boldsymbol{\theta}) = \frac{1}{Z(\beta)} \exp\{-\beta \Psi_\lambda(\boldsymbol{\theta}; \rho_*)\}. \qquad \textbf{[21]}$$

In the next sections, we state our results about convergence of the DD to its fixed point. In the case of noisy SGD [and for the diffusion PDE (12)], a general convergence result can be established (although at the cost of an additional regularization). For noiseless SGD (and the continuity equation; Eq. **12**), we do not have such a general result. However, we obtain a stability condition for a fixed point containing one point mass, which is useful to characterize possible limiting points (and is used in treating the examples in the previous section).

**Convergence: Noisy SGD.** Remarkably, the diffusion PDE (Eq. **12**) generically admits a unique fixed point, which is the global minimum of $F_{\beta,\lambda}(\rho)$, and the evolution (Eq. **12**) converges to it, if initialized so that $F_{\beta,\lambda}(\rho_0) < \infty$. This statement requires some qualifications. First, we introduce sufficient regularity assumptions to guarantee the existence of sufficiently smooth solutions of Eq. **12**:

A4. $V \in C^4(\mathbb{R}^D)$, $U \in C^4(\mathbb{R}^D \times \mathbb{R}^D)$, $\nabla^k_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is uniformly bounded for $0 \le k \le 4$.

Next notice that the righthand side of the fixed point equation (Eq. **21**) is not necessarily normalizable [for instance, it is not when $V(\,\cdot\,)$, $U(\,\cdot\,,\,\cdot\,)$ are bounded]. To ensure the existence of a fixed point, we need $\lambda > 0$.

**Theorem 4:** *Assume that conditions* **A1–A4** *hold, and* $1/K_0 \le \lambda \le K_0$ *for some* $K_0 > 0$. *Then* $F_{\beta,\lambda}(\rho)$ *has a unique minimizer, denoted by* $\rho_*^{\beta,\lambda}$, *which satisfies*

$$R(\rho_*^{\beta,\lambda}) \le \inf_{\boldsymbol{\theta} \in \mathbb{R}^{N \times D}} R_N(\boldsymbol{\theta}) + C\,D/\beta, \qquad \textbf{[22]}$$

*where* $C$ *is a constant depending on* $K_0, K_1, K_2, K_3$. *Further, letting* $\rho_t$ *be a solution of the diffusion PDE* (Eq. **12**) *with initialization satisfying* $F_{\beta,\lambda}(\rho_0) < \infty$, *we have, as* $t \to \infty$,

$$\rho_t \Rightarrow \rho_*^{\beta,\lambda}. \qquad \textbf{[23]}$$

The proof of this theorem is based on the following formula that describes the free-energy decrease along the trajectories of the DD (Eq. **12**):

$$\frac{\mathrm{d}F_{\beta,\lambda}(\rho_t)}{\mathrm{d}t} = -2\xi(t) \int_{\mathbb{R}^D} \|\nabla_{\boldsymbol{\theta}} \left(\Psi_\lambda(\boldsymbol{\theta}; \rho_t) + 1/\beta \cdot \log \rho_t(\boldsymbol{\theta})\right)\|_2^2 \rho_t(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \qquad \textbf{[24]}$$

(A key technical hurdle is of course proving that this expression makes sense, which we do by showing the existence of strong solutions.) It follows that the righthand side must vanish as $t \to \infty$, from which we prove that (eventually taking subsequences) $\rho_t \Rightarrow \rho_*$ where $\rho_*$ must satisfy $\beta\Psi_\lambda(\boldsymbol{\theta}; \rho_*) + \log \rho_*(\boldsymbol{\theta}) = \text{const}$. This in turns mean $\rho_*$ is a solution of the fixed point condition 21 and is in fact a global minimum of $F_{\beta,\lambda}$ by convexity.

This result can be used in conjunction with Theorem 3, to analyze the regularized noisy SGD algorithm (Eq. **11**).

**Theorem 5:** *Assume that conditions* **A1–A4** *hold. Let* $\rho_0 \in \mathscr{P}(\mathbb{R}^D)$ *be absolutely continuous with* $F_{\beta,\lambda}(\rho_0) < \infty$ *and* $K_0$ *sub-Gaussian. Consider regularized noisy SGD (cf.* Eq. **11**) *at inverse temperature* $\beta < \infty$, *regularization* $1/K_0 \le \lambda \le K_0$ *with initialization* $(\boldsymbol{\theta}_i^0)_{i \le N} \sim_{iid} \rho_0$. *Then, for any* $\eta > 0$, *there exists* $K = K(\eta, \{K_i\})$, *and setting* $\beta \ge KD$, *there exists* $T = T(\eta, V, U, \{K_i\}, D, \beta) < \infty$ *and* $C_0 = C_0(\eta, \{K_i\}, \delta)$ *(independent of the dimension* $D$ *and temperature* $\beta$) *such that the following happens for* $N, (1/\varepsilon) \ge C_0 e^{C_0 T} D$, $\varepsilon \ge 1/N^{10}$: *For any* $k \in [T/\varepsilon, 10T/\varepsilon]$, *we have, with probability* $1 - \delta$,

$$R_N(\boldsymbol{\theta}^k) \le \inf_{\rho \in \mathscr{P}(\mathbb{R}^D)} R_\lambda(\rho) + \eta. \qquad \textbf{[25]}$$

Let us emphasize that the convergence time $T$ in the last theorem can depend on the dimension $D$ and on the data distribution $\mathbb{P}$ but is independent of the number of hidden units $N$. As illustrated by the examples in the previous section, understanding the dependence of $T$ on $D$ requires further analysis, but examining the proof of this theorem suggests $T = e^{O(D)}$ quite generally [examples in which $T = O(1)$ or $T = e^{\Theta(D)}$ can be constructed]. We expect that our techniques could be pushed to investigate the dependence of $T$ on $\eta$ (see *SI Appendix, Discussion*). In highly structured cases, the dimension $D$ can be of constant order and be much smaller than $d$.

**Convergence: Noiseless SGD.** The next theorems provide necessary and sufficient conditions for distributions containing a single point mass to be a stable fixed point of the evolution. This result is useful to characterize the large time asymptotics of the dynamics (Eq. **7**). Here, we write $\nabla_1 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for the gradient of $U$ with respect to its first argument and $\nabla^2_{1,1} U$ for the corresponding Hessian. Further, for a probability distribution $\rho_*$, we define

$$\boldsymbol{H}_0(\rho_*) = \nabla^2 V(\boldsymbol{\theta}_*) + \int \nabla^2_{1,1} U(\boldsymbol{\theta}_*, \boldsymbol{\theta})\, \rho_*(\mathrm{d}\boldsymbol{\theta}). \qquad \textbf{[26]}$$

Note that $\boldsymbol{H}_0(\rho_*)$ is nothing but the Hessian of $\boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta}; \rho_*)$ at $\boldsymbol{\theta}_*$.

**Theorem 6:** *Assume* $V, U$ *to be twice differentiable with bounded gradient and bounded continuous Hessian. Let* $\boldsymbol{\theta}_* \in \mathbb{R}^D$ *be given. Then* $\rho_* = \delta_{\boldsymbol{\theta}_*}$ *is a fixed point of the evolution* (Eq. **7**) *if and only if* $\nabla V(\boldsymbol{\theta}_*) + \nabla_1 U(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) = \mathbf{0}$.

*Define* $\boldsymbol{H}_0(\delta_{\boldsymbol{\theta}_*}) \in \mathbb{R}^{D \times D}$ *as per* Eq. **26**. *If* $\lambda_{\min}(\boldsymbol{H}_0(\delta_{\boldsymbol{\theta}_*})) > 0$, *then there exists* $r_0 > 0$ *such that, if* $\mathrm{supp}(\rho_{t_0}) \subseteq \mathsf{B}(\boldsymbol{\theta}_*; r_0) \equiv \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \le r_0\}$, *then* $\rho_t \Rightarrow \rho_*$ *as* $t \to \infty$. *In fact, convergence is exponentially fast, namely* $\int \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \rho_t(\mathrm{d}\boldsymbol{\theta}) \le e^{-\lambda(t - t_0)}$ *for some* $\lambda > 0$.

**Theorem 7:** *Under the same assumptions of Theorem 6, let* $\rho_* = p_* \delta_{\boldsymbol{\theta}_*} + (1 - p_*)\tilde{\rho}_* \in \mathscr{P}(\mathbb{R}^D)$ *be a fixed point of dynamics* (Eq. **7**), *with* $p_* \in (0, 1]$ *and* $\nabla \Psi(\boldsymbol{\theta}_*; \rho_*) = \mathbf{0}$ *(which, in particular, is implied by the fixed point condition*; Eq. **20**). *Define the level sets* $\mathcal{L}(\eta) \equiv \{\boldsymbol{\theta} : \Psi(\boldsymbol{\theta}; \rho_*) \le \Psi(\boldsymbol{\theta}_*; \rho_*) - \eta\}$ *and make the following assumptions*: (B1) *The eigenvalues of* $\boldsymbol{H}_0 = \boldsymbol{H}_0(\rho_*)$ *are all different from* 0, *with* $\lambda_{\min}(\boldsymbol{H}_0) < 0$; (B2) $\tilde{\rho}_*(\mathcal{L}(\eta)) \uparrow 1$ *as* $\eta \downarrow 0$; *and* (B3) *there exists* $\eta_0 > 0$ *such that the sets* $\partial \mathcal{L}(\eta)$ *are compact for all* $\eta \in (0, \eta_0)$.

*If* $\rho_0$ *has a bounded density with respect to the Lebesgue measure, then it cannot be that* $\rho_t$ *converges weakly to* $\rho_*$ *as* $t \to \infty$.

## Discussion and Future Directions

In this paper, we developed an approach to the analysis of two-layer neural networks. Using a propagation-of-chaos argument, we proved that—if the number of hidden units satisfies $N \gg D$—SGD dynamics is well approximated by the PDE in Eq. **7**, while noisy SGD is well approximated by Eq. **12**. Both of these asymptotic descriptions correspond to Wasserstein gradient flows for certain energy (or free energy) functionals. While empirical risk minimization is known to be insensitive to overparametrization (22), the present work clarifies that the SGD behavior is also independent of the number of hidden units, as soon as this is large enough.

We illustrated our approach on several concrete examples, by proving convergence of SGD to a near-global optimum. This type of analysis provides a mechanism for avoiding the perils of nonconvexity. We do not prove that the finite-$N$ risk $R_N(\boldsymbol{\theta})$ has a unique local minimum or that all local minima are close to each other. Such claims have often been the target of earlier work but might be too strong for the case of neural networks. We prove instead that the PDE (Eq. **7**) converges to a near-global optimum, when initialized with a bounded density. This effectively gets rid of some exceptional stationary points of $R_N(\boldsymbol{\theta})$ and merges multiple finite $N$ stationary points that result into similar distributions $\rho$.

In the case of noisy SGD (Eq. **11**), we prove that it converges generically to a near-global minimum of the regularized risk, in time independent of the number of hidden units.

We emphasize that while we focused here on the case of square loss, our approach should be generalizable to other loss functions as well (cf. *SI Appendix*).

The present work opens the way to several interesting research directions. We will mention two of them: $(i)$ The PDE (Eq. **7**) corresponds to gradient flow in the Wasserstein metric for the risk $R(\rho)$ (see ref. 20). Building on this remark, tools from optimal transportation theory can be used to prove convergence. $(ii)$ Multiple finite-$N$ local minima can correspond to the same minimizer $\rho_*$ of $R(\rho)$ in the limit $N \to \infty$. Ideas from glass theory (24) might be useful to investigate this structure.

Let us finally mention that, after a first version of this paper appeared as a preprint, several other groups obtained results that are closely related to Theorem 3 (25–27).

**ACKNOWLEDGMENTS.** This work was partially supported by NSF Grants DMS-1613091, CCF-1714305, and IIS-1741162. S.M. was partially supported by an Office of Technology Licensing Stanford Graduate Fellowship. P.-M.N. was partially supported by a William R. Hewlett Stanford Graduate Fellowship.

1. Rosenblatt F (1962) *Principles of Neurodynamics* (Spartan Book, Washington, DC).
2. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, ed Vardi MY (Association for Computing Machinery, New York), pp 1097–1105.
3. Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) *Deep Learning* (MIT Press, Cambridge), Vol 1.
4. Robbins H, Monro S (1951) A stochastic approximation method. *Ann Math Stat.* 22:400–407 .
5. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, eds Lechevallier Y, Saporta G (Physica, Heidelberg), pp 177–186.
6. Shalev-Shwartz S, Ben-David S (2014) *Understanding Machine Learning: From Theory to Algorithms* (Cambridge Univ Press, New York).
7. Wang C, Mattingly J, Lu YM (2017) Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA. arXiv:1712.04332.
8. Soltanolkotabi M, Javanmard A, Lee JD (2017) Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. arXiv:1707.04926.
9. Ge R, Lee JD, Ma T (2017) Learning one-hidden-layer neural networks with landscape design. arXiv:1711.00501.
10. Brutzkus A, Globerson A (2017) Globally optimal gradient descent for a convnet with Gaussian inputs. arXiv:1702.07966.
11. Arora S, Bhaskara A, Ge R, Ma T (2014) Provable bounds for learning some deep representations. *Proceedings of International Conference on Machine Learning (ICML)*. Available at https://arxiv.org/abs/1310.6343. Accessed July 18, 2018.
12. Sedghi H, Anandkumar A (2015) Provable methods for training neural networks with sparse connectivity. *Proceedings of International Conference on Learning Representation (ICLR)*. Available at https://arxiv.org/abs/1412.2693. Accessed July 18, 2018.
13. Janzamin M, Sedghi H, Anandkumar A (2015) Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. arXiv:1506.08473.
14. Zhang Y, Lee JD, Jordan MI (2016) L1-regularized neural networks are improperly learnable in polynomial time. *Proceedings of International Conference on Machine Learning (ICML)*. Available at https://arxiv.org/abs/1510.03528. Accessed July 18, 2018.
15. Tian Y (2017) Symmetry-breaking convergence analysis of certain two-layered neural networks with ReLU nonlinearity. *International Conference on Learning Representation (ICLR)*. Available at https://openreview.net/forum?id=Hk85q85ee. Accessed July 18, 2018.
16. Zhong K, Song Z, Jain P, Bartlett PL, Dhillon IS (2017) Recovery guarantees for one-hidden-layer neural networks. arXiv:1706.03175.
17. Sun Lee W, Bartlett PL, Williamson RC (1996) Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans Inf Theor* 42:2118–2132.
18. Bengio Y, Roux NL, Vincent P, Delalleau O, Marcotte P (2006) Convex neural networks. *Advances in Neural Information Processing Systems*, eds Weiss Y, Schölkopf B, Platt JC (MIT Press, Cambridge, MA), pp 123–130.
19. Jordan R, Kinderlehrer D, Otto F (1998) The variational formulation of the Fokker–Planck equation. *SIAM J Math Anal* 29:1–17.
20. Ambrosio L, Gigli N, Savaré G (2008) *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (Birkhäuser, Basel).
21. Carrillo JA, McCann RJ, Villani C (2003) Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates. *Reva Matematica Iberoam* 19:971–1018.
22. Bartlett PL (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 Inf Theor* 44:525–536.
23. Sznitman A-S (1991) Topics in propagation of chaos. *Ecole d'été de probabilités de Saint-Flour XIX—1989*, ed Hennequin PL (Springer, Berlin), pp 165–251.
24. Mézard M, Parisi G (1999) Thermodynamics of glasses: A first principles computation. *J Phys Condens Matter* 11:A157–A165.
25. Rotskoff GM, Vanden-Eijnden E (2018) Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. arXiv:1805.00915.
26. Sirignano J, Spiliopoulos K (2018) Mean field analysis of neural networks. arXiv:1805.01053.
27. Chizat L, Bach F (2018) On the global convergence of gradient descent for over-parameterized models using optimal transport. arXiv:1805.09545.