Global convergence of neuron birth-death dynamics

Grant M. Rotskoff [1] **Samy Jelassi** [2 3] **Joan Bruna** [1 2] **Eric Vanden-Eijnden** [1]

Abstract

Neural networks with a large number of units admit a mean-field description, which has recently served as a theoretical explanation for the favorable training properties of "overparameterized" models. In this regime, gradient descent obeys a deterministic partial differential equation (PDE) that converges to a globally optimal solution for networks with a single hidden layer under appropriate assumptions. In this work, we propose a non-local mass transport dynamics that leads to a modified PDE with the same minimizer. We implement this non-local dynamics as a stochastic neuronal birth-death process and we prove that it accelerates the rate of convergence in the mean-field limit. We subsequently realize this PDE with two classes of numerical schemes that converge to the mean-field equation, each of which can easily be implemented for neural networks with finite numbers of units. We illustrate our algorithms with two models to provide intuition for the mechanism through which convergence is accelerated.

1. Introduction

As a consequence of the universal approximation theorems, sufficiently wide single layer neural networks are expressive enough to accurately represent a broad class of functions (Cybenko, 1989; Barron, 1993; Park & Sandberg, 1991). The existence of a neural network function arbitrarily close to a given target function, however, is not a guarantee that any particular optimization procedure can identify the optimal parameters. Recently, using mathematical tools from optimal transport theory and interacting particle systems, it was shown that gradient descent (Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2018; Sirignano & Spiliopoulos, 2018; Chizat & Bach, 2018b) and stochastic gradient descent converge asymptotically to the target function in the large data limit.

This analysis relies on taking a "mean-field" limit in which the number of units n tends to infinity. In this setting, gradient descent optimization dynamics is described by a partial differential equation (PDE), corresponding to a Wasserstein gradient flow on a convex energy functional. This PDE provides a powerful conceptual framework for analyzing the properties of neural networks evolving under gradient descent dynamics. In addition, the analysis of this Wasserstein gradient flow motivates the interesting possibility of altering the dynamics to accelerate convergence.

In this work, we propose a dynamical scheme involving a birth/death process over the units of the neural network. It can be defined on systems of interacting (e.g., neural network optimization) or non-interacting particles, and in the mean-field limit it amounts to an *unbalanced* transport (Chizat et al., 2018) in which mass can be locally 'teletransported' with finite cost. We prove that the resulting modified transport equation converges to the global minimum of the loss in both interacting and non-interacting regimes (under appropriate assumptions), and we provide an explicit rate of convergence in the latter case for the mean-field limit. Interestingly—and unlike the gradient flow—the *only* fixed point of the dynamics is the global minimum of the loss function. We study the fluctuations of finite particle dynamics around this mean-field convergent solution, showing that they are of the same order throughout the dynamics and therefore providing algorithmic guarantees directly applicable to finite single-layer neural network optimization. Finally, we derive algorithms that converge to the birth-death PDEs and verify numerically that these schemes accelerate convergence even for finite numbers of parameters.

Summarily, we describe:

Global convergence and monotonicity of the energy with birth-death dynamics — We propose in Section 3 a modification of the original gradient flow that can be interpreted as a birth-death process with the ability to perform non-local mass transport in the equation governing the parameter distribution. We prove that the scheme we introduce guarantee global convergence and increase the rate of contraction of the energy compared to gradient descent

[1]Courant Institute, New York University, New York, USA [2]Center for Data Science, New York University, New York, USA [3]Princeton University, Princeton, New Jersey, USA. Correspondence to: Grant M. Rotskoff <rotskoff@cims.nyu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

and stochastic gradient descent for fixed μ. We also derive asymptotic rates of convergence (Section 4).

Analysis of fluctuations and self-quenching — The birth-death dynamics introduces additional fluctuations that are not present in gradient descent dynamics. In Section 5 we calculate these fluctuations using tools from the theory of measure-valued Markov processes. We show that these fluctuations, for n sufficiently large, are of order $O(n^{-1/2})$ and "self-quenching" in the sense that they diminish in magnitude as the quality as the optimization dynamics approaches the optimum.

Algorithms for realizing the birth-death schemes — In Section 6 we detail numerical schemes (and provide implementations in `PyTorch`) of the birth-death schemes described below. In the particular case of neural networks, the computational cost of implementing our procedure is minimal because no additional gradient computations are required. We demonstrate the efficacy of these algorithms on simple, illustrative examples in Section 7.

2. Related Work

Non-local update rules appear in various areas of machine learning and optimization. Derivative-free optimization (Rios & Sahinidis, 2013) offers a general framework for optimizing complex non-convex functions using non-local search heuristics. Some notable examples include Particle Swarm Optimization (Kennedy, 2011) and Evolutionary Strategies, such as the Covariance Matrix Adaptation method (Hansen, 2006). These approaches have found some renewed interest in the optimization of neural networks in the context of Reinforcement Learning (Salimans et al., 2017; Such et al., 2017) and hyperparameter optimization (Jaderberg et al., 2017).

Our setup of non-interacting potentials is closely related to the so-called Estimation of Distribution Algorithms (Baluja & Caruana, 1995; Larrañaga & Lozano, 2001), which define update rules for a probability distribution over a search space by querying the values of a given function to be optimized. In particular, Information Geometric Optimization Algorithms (Ollivier et al., 2017) study the dynamics of parametric densities using ordinary differential equations, focusing on invariance properties. In contrast, our focus in on the combination of transport (gradient-based) and birth/death dynamics.

Dropout (Srivastava et al., 2014) is a regularization technique popularized by the AlexNet CNN (Krizhevsky et al., 2012) reminiscent of a birth/death process, but we note that its mechanism is very different: rather than killing a neuron and replacing it by a new one with some rate, Dropout momentarily masks neurons, which become active again at the same position; in other words, Dropout implements a purely local transport scheme, as opposed to our non-local dynamics.

Finally, closest to our motivation is (Wei et al., 2018), who, building on the recent body of works that leverage optimal transport techniques to study optimization in the large parameter limit (Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018b; Mei et al., 2018; Sirignano & Spiliopoulos, 2018), proposed a modification of the dynamics that replaced traditional stochastic noise by a resampling of a fraction of neurons from a base, fixed measure. Our model has significant differences to this scheme, namely we show that the dynamics preserves the same global minimizers and accelerates the rate of convergence. Finally, our interpretation of the modified dynamics in terms of a generalized gradient flow is related to the unbalanced optimal transport setups of (Kondratyev et al., 2016; Liero et al., 2018; Chizat et al., 2018). Our analysis of the resulting dynamics in terms of proximal operators was also studied in (Gallouët & Monsaingeon, 2017) in the context of unbalanced transport.

3. Mean-field PDE and Birth-death Dynamics

3.1. Mean-Field Limit and Liouville dynamics

Gradient descent propagates the parameters locally in proportion to the gradient of the objective function. In some cases, an optimization algorithm can benefit from nonlocal dynamics, for example, by allowing new units to appear at favorable values and existing units to be removed if they diminish the quality of the representation. In order to exploit a nonlocal dynamical scheme, it is useful to interpret the units as a system of n particles, $\boldsymbol{\theta}_i \in D$, a k-dimensional differentiable manifold, which for $i = 1, \dots, n$ evolve on a landscape determined by the objective function $\ell(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)$. Here we will focus on situations where the objective function may involve interactions between pairs of units:

$$\ell(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) = \sum_{i=1}^{n} F(\boldsymbol{\theta}_i) + \frac{1}{2n} \sum_{i,j=1}^{n} K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \quad (1)$$

where $F : D \to \mathbb{R}$ is a single particle energy function and $K : D \times D \to \mathbb{R}$ is a symmetric semi-positive definite interaction kernel. Interestingly, optimizing neural networks with the mean-squared loss function fits precisely this framework (Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2018; Chizat & Bach, 2018b). Consider a supervised learning problem using a neural network with nonlinearity φ. If we write the neural network as

$$f_n(\boldsymbol{x}; \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{x}, \boldsymbol{\theta}_i) \quad (2)$$

and expand the loss function,

$$\ell(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) = \tfrac{1}{2} \mathbb{E}_{y,\boldsymbol{x}} \left| y - f_n(\boldsymbol{x}; \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) \right|^2, \quad (3)$$

we see that, up to an irrelevant constant depending only on the data distribution, we arrive at (1) with

$$F(\boldsymbol{\theta}) = -\mathbb{E}_{y,\boldsymbol{x}}\big[y\varphi(\boldsymbol{x},\boldsymbol{\theta})\big], \tag{4}$$

and,

$$K(\boldsymbol{\theta},\boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{x}}\big[\varphi(\boldsymbol{x},\boldsymbol{\theta})\varphi(\boldsymbol{x},\boldsymbol{\theta}')\big]. \tag{5}$$

We also consider *non-interacting* objective functions in which $K = 0$ in (1). Optimization problems that fit this framework include resource allocation tasks in which, e.g., weak performers are eliminated, Evolution Strategies, and Information Geometric Optimization (Ollivier et al., 2017).

In the case of gradient descent dynamics, the evolution of the particles $\boldsymbol{\theta}_i$ is governed for $i = 1, \dots, n$ by

$$\dot{\boldsymbol{\theta}}_i = -\nabla_{\boldsymbol{\theta}_i}\ell(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n). \tag{6}$$

To analyze the dynamics of this particle system, we consider the "mean-field" limit $n \to \infty$. As the number of particles becomes large, the empirical distribution of particles

$$\mu_t^{(n)}(d\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\boldsymbol{\theta}_j(t)}(d\boldsymbol{\theta}) \tag{7}$$

leads to a deterministic partial differential equation at first order (Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2018; Chizat & Bach, 2018b; Sirignano & Spiliopoulos, 2018),

$$\partial_t\mu_t = \nabla\cdot(\mu_t\nabla V), \tag{8}$$

where μ_t is the weak limit of μ_t^n and μ_0 is some distribution from which the initial particle positions $\boldsymbol{\theta}_i(0)$ are drawn independently. The potential $V : D \to \mathbb{R}$ is specified by the objective function ℓ as

$$V(\boldsymbol{\theta}, [\mu]) = F(\boldsymbol{\theta}) + \int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu(d\boldsymbol{\theta}'). \tag{9}$$

and (8) should be interpreted in the weak sense, i.e., we require $\forall\phi \in C_c^\infty(D)$

$$\partial_t\int_D \phi(\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}) = -\int_D \nabla\phi(\boldsymbol{\theta})\cdot\nabla V(\boldsymbol{\theta}, [\mu_t])\mu_t(d\boldsymbol{\theta}), \tag{10}$$

where $C_c^\infty(D)$ denotes the space of smooth functions with compact support on D.

Because V is the gradient with respect to μ of an energy functional $\mathcal{E}[\mu]$,

$$\mathcal{E}[\mu] = \int_D F(\boldsymbol{\theta})\mu(d\boldsymbol{\theta}) + \tfrac{1}{2}\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu(d\boldsymbol{\theta})\mu(d\boldsymbol{\theta}'), \tag{11}$$

the nonlinear Liouville equation (8) is the Wasserstein gradient flow with respect to the energy functional $\mathcal{E}[\mu]$. Local minima of V (where $\nabla V = 0$) are clearly fixed points of this gradient flow, but these fixed points may not always be minimizers of the energy when $\operatorname{supp}\mu \subset D$. When the initial distribution of units has full support, neural networks evolving with gradient descent avoid these spurious fixed points under appropriate assumptions about their nonlinearity (Chizat & Bach, 2018b; Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2018).

3.2. Birth-Death augmented Dynamics

Here we consider a more general dynamical scheme that involves nonlocal transport of particle mass. As we shall see in Section 4, this dynamics avoids spurious fixed points and local minima, and converges asymptotically to the global minimum. Consider the following modification of the Wasserstein gradient flow above:

$$\partial_t\mu_t = \nabla\cdot(\mu_t\nabla V) - \alpha V\mu_t \qquad (\alpha > 0). \tag{12}$$

The additional term $-\alpha V\mu_t$ is a birth/death term that modifies the mass of μ. If V is positive, this mass will decrease, corresponding to the removal or "death" of units. If V is negative, this mass will increase, which can be implemented as duplication or "cloning" of units. For a finite number of units, this dynamics could lead to changes in the architecture of the network. In many applications it is preferable to fix the total population, achieved by simply adding a conservation term to the dynamics,

$$\partial_t\mu_t = \nabla\cdot(\mu_t\nabla V) - \alpha V\mu_t + \alpha\bar{V}\mu_t, \tag{13}$$

where $\bar{V} \equiv \int_D V d\mu_t$. This equation (like (12)) should in general be interpreted in the weak sense. Here we will focus on solutions of (13) for the initial condition $\mu_0 \in \mathcal{M}(D)$, the space of probability measures on D, that satisfy

$$\int_D \phi(\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}) = C^{-1}(t)\int_D \phi(\boldsymbol{\theta}_t)e^{-\alpha\int_0^t V(\boldsymbol{\theta}_s,[\mu_s])ds}\mu_0(d\boldsymbol{\theta}) \tag{14}$$

where $\phi : D \to \mathbb{R}$ is any bounded differentiable function with bounded gradient, $C(t)$ is given by

$$C(t) = e^{-\alpha\int_0^t \bar{V}[\mu_s]ds} \equiv \int_D e^{-\alpha\int_0^t V(\boldsymbol{\theta}_s,[\mu_s])ds}\mu_0(d\boldsymbol{\theta}), \tag{15}$$

and $\boldsymbol{\theta}_t$ satisfies $\dot{\boldsymbol{\theta}}_t = -\nabla V(\boldsymbol{\theta}_t, [\mu_t])$ with $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$. Formula (14) can be formally established by solving (13) by the method of characteristics (Appendix C). In the non-interacting case, since $V(\boldsymbol{\theta}, [\mu_t]) = F(\boldsymbol{\theta})$, (14) is explicit and well-posed under appropriate assumptions on F (see Assumption C.1 below). In the interacting case, (14) is implicit since the right hand side depends on μ_t. Following Chizat & Bach (Chizat & Bach, 2018b), we know that under appropriate assumptions on F and K (see Assumption 4.2 below), solutions to (14) exist for all $t > 0$ for appropriate initial μ_0 that are compactly supported in D. Here we will

assume global existence of solutions to this equation for μ_0 such that $\operatorname{supp}\mu_0 = D$ with D open: if μ_0 decays sufficiently fast at infinity, this assumption is supported by the alternative derivation of (12) based on a proximal gradient formulation given in Appendix B.

Note that solutions of (12) that satisfy (14) are probability measures since they are positive by definition and we can set $\phi = 1$ in (14) to deduce that $\mu_t(D) = 1$. We can also show that the birth-death terms improve the rate of energy decay, as stated in the following proposition:

Proposition 3.1. *Let μ_t be a solution of (13) for the initial condition $\mu_0 \in \mathcal{M}(D)$ that satisfies (14) for all $t \geq 0$. Then, $\mu_t(D) = 1$ for all $t \geq 0$, and $E(t) = \mathcal{E}[\mu(t)]$ satisfies*

$$\begin{aligned}\dot{E}(t) = &-\int_D |\nabla V(\boldsymbol{\theta}, [\mu_t])|^2 \mu_t(d\boldsymbol{\theta}) \\ &- \alpha \int_D \left(V(\boldsymbol{\theta}, [\mu_t]) - \bar{V}[\mu_t]\right)^2 \mu_t(d\boldsymbol{\theta}) \leq 0.\end{aligned} \tag{16}$$

Proof: (16) can be formally obtained by testing (13) against $V(\boldsymbol{\theta}, [\mu_t])$ and using the chain rule to deduce that $d\mathcal{E}[\mu_t]/dt = \int_D V(\boldsymbol{\theta}, [\mu_t])\partial_t \mu_t(d\boldsymbol{\theta})$. To complete the proof, we need to show that this testing is legitimate and the terms at the right hand side of (16) are well-defined; this is done in Appendix E by differentiating $C(t)$. □

The birth-death term thus contributes to increase the rate of decay of the energy at all times. A natural question is whether such improved energy decay can lead to global convergence of the dynamics to the global minimum of the energy. As it turns out, the answer is yes: the fixed points of the birth-death PDEs (12) and (13) are the global minimizers of the energy $\mathcal{E}[\mu]$, as we prove in Section 4. How to implement a particle dynamics consistent with (13) is discussed in Sections 5 and 6.

We also note that there are several ways in which we can modify (13) to certain advantages: this is discussed in Appendix A.

4. Convergence of Transport Dynamics with Birth-death

Here, we compare the solutions of the original PDE (8) with those of the PDE (13) with birth-death. We restrict ourselves to situations where F and K in (11) are such that $\mathcal{E}[\mu]$ is bounded from below. Our main technical contributions are results about convergence towards global energy minimizer as well as convergence rates as the dynamics approaches these minimizers. We consider in this section the interacting case and describe the easier non-interacting case in Appendix C.

Under gradient descent dynamics, global convergence can be established with appropriate assumptions on the initialization and architecture of the neural network. (Mei et al., 2018) establishes global convergence and provides a rate for neural networks with bounded activation functions evolving under stochastic gradient descent. Similar results were obtained in (Chizat & Bach, 2018b; Rotskoff & Vanden-Eijnden, 2018), in which it is proven that gradient descent converges to the globally optimal solution for neural networks with particular homogeneity conditions on the activation functions and regularizers. Closely related to the present work, (Wei et al., 2018) provides a convergence rate for a "perturbed" gradient flow in which uniform noise is added to the PDE (8). It should be emphasized that, unlike our formulation, the addition of uniform noise changes the fixed point of the PDE and convergence to only an approximate global solution can be obtained in that setting.

Let us now consider the interacting case, when V is given by (9) with $K \neq 0$. We make

Assumption 4.1. *The set D is a k-dimensional differentiable manifold which is either closed (i.e. compact, with no boundaries), or open (i.e. with no closed subset), or the Cartesian product of a closed and an open manifold.*

Assumption 4.2. *The kernel K is symmetric, positive semi-definite, and twice differentiable in its arguments, $K \in C^2(D \times D)$; $F \in C^2(D)$; and F and K are such that the energy is bounded from below, i.e. $\exists m \in \mathbb{R}$ such that $\forall \mu \in \mathcal{M}(D) : \mathcal{E}[\mu] \geq m$.*

This technical assumption typically holds for neural networks. Assumption 4.2 guarantees that the quadratic energy $\mathcal{E}[\mu]$ in (11) has a (unique) minimum value. While we cannot guarantee in general that this minimum is reached only by minimizers, below we will work under the assumption that minimizers exist. These are solutions in $\mathcal{M}(D)$ of following Euler-Lagrange equations:

$$\begin{cases} V(\boldsymbol{\theta}, [\mu_*]) = \bar{V}[\mu_*] & \forall \boldsymbol{\theta} \in \operatorname{supp}\mu_* \\ V(\boldsymbol{\theta}, [\mu_*]) \geq \bar{V}[\mu_*] & \forall \boldsymbol{\theta} \in D. \end{cases} \tag{17}$$

where $\bar{V}[\mu] \equiv \int_D V(\boldsymbol{\theta}, [\mu])\mu(d\boldsymbol{\theta})$. These equations are well-known (Serfaty, 2015): we recall their derivation in Appendix D.

Minimizers of the energy should not be confused with fixed points of the dynamics. In particular, a well-known issue with the PDE (8) is that it potentially has many more fixed points than $\mathcal{E}[\mu]$ has minimizers: Indeed, rather than (17), these fixed points only need to satisfy

$$\nabla V(\boldsymbol{\theta}, [\mu]) = 0 \qquad \forall \boldsymbol{\theta} \in \operatorname{supp}\mu. \tag{18}$$

It is therefore remarkable that, if we pick an initial condition μ_0 for the birth-death PDE (13) that has full support, the solution to this equation converges to a global minimizer of $\mathcal{E}[\mu]$:

Theorem 4.3 (Global Convergence to Global Minimizers: Interacting Case)**.** *Let μ_t denote the solution of (13) that satisfies (14) for the initial condition μ_0 with $\operatorname{supp}\mu_0 = D$. If $\mu_t \rightharpoonup \mu_*$ as $t \to \infty$ for some probability measure $\mu_* \in \mathcal{M}(D)$, then under Assumptions 4.1 and 4.2 μ_* is a global minimizer of $\mathcal{E}[\mu]$.*

This theorem is proven in Appendix E. Note that the theorem holds under the assumption that μ_t converges to a fixed point μ_*, which we cannot guarantee *a priori* but should be true for a wide class of F and K and initial conditions μ_0 satisfying properties like $\mathcal{E}[\mu_0], \infty$—for more details on these conditions see the proof in Appendix E. One aspect of this proof is based on the evolution equation (16) for $\mathcal{E}[\mu_t]$. Since $d\mathcal{E}[\mu_t]/dt \leq 0$ and since $\mathcal{E}[\mu_t]$ is bounded from below by Assumption 4.2, by the bounded convergence theorem, the evolution must stop eventually. By assumption, this involves μ_t converging weakly towards some μ_*. This happens when both integrals in (16) are zero, i.e. μ_* must satisfy the first equation in (17) as well as (18). What remains to be shown is that μ_* must also satisfy the second equation in (17), which we check in Appendix E.

Regarding the rate of convergence, we have the following result:

Theorem 4.4 (Asymptotic Convergence Rate: Interacting Case)**.** *Under the same conditions as in Theorem 4.3, $\exists C > 0$ and $t_C > 0$ such that $E(t) = \mathcal{E}[\mu_t] - \mathcal{E}[\mu_*] \geq 0$ satisfies*

$$E(t) \leq Ct^{-1} \quad \text{if} \quad t \geq t_C \tag{19}$$

The proof of this theorem is given in Appendix F where we show that

$$\lim_{t\to\infty} tE(t) \leq C \in (0, \infty]. \tag{20}$$

5. From Mean-field to Particle Dynamics with Birth-Death

In practice the number of units n is finite, so we must verify that we can implement dynamics at finite particle numbers that is consistent with the PDEs with birth-death terms introduced in Sec. 3 in the mean-field limit $n \to \infty$. We must also ensure that the fluctuations arising from the discrete particles do not pose a problem for the optimization dynamics. In this section, we carry out this program in the context of the PDE (13). Analogous calculations can be performed in the case of (2). These results rely on the theory of measure-valued Markov processes (Dawson, 2006), and are detailed in Appendix G.

The dynamics of the particles $\{\boldsymbol{\theta}_i(t)\}_{i=1}^n$ is specified by a Markov process defined as follows: the birth-death part of the evolution is realized by equipping each particle $\boldsymbol{\theta}_i$ with an independent exponential clock with (signed) rate

$$\begin{aligned}\tilde{V}(\boldsymbol{\theta}_i) = F(\boldsymbol{\theta}_i) + \frac{1}{n}\sum_{j=1}^n K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \\ - \frac{1}{n}\sum_{j=1}^n \left(F(\boldsymbol{\theta}_j) + \frac{1}{n}\sum_{k=1}^n K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_k) \right)\end{aligned} \tag{21}$$

such that:

1. If $\tilde{V}(\boldsymbol{\theta}_i(t)) > 0$, the particle $\boldsymbol{\theta}_i$ is duplicated with instantaneous rate $\alpha\tilde{V}(\boldsymbol{\theta}_i(t))$, and a particle $\boldsymbol{\theta}_j$ chosen at random in the stack is killed to preserve the population size.
2. If $\tilde{V}(\boldsymbol{\theta}_i(t)) < 0$, the particle $\boldsymbol{\theta}_i$ is killed with instantaneous rate $\alpha|\tilde{V}(\boldsymbol{\theta}_i(t))|$, and a particle $\boldsymbol{\theta}_j$ chosen at random in the stack is duplicated to preserve the population size.

Between these birth events the particles evolve by the GD flow (6).

Due to the interchangeability of the particles, the evolution of their empirical distribution $\mu_t^{(n)}$ defined in (7) is also Markovian: it is referred to in the probability literature as a *measured-valued Markov process* (Dawson, 2006). We can write down the generator of this process, which specifies the evolution of the expectation of functionals of $\mu_t^{(n)}$, and analyze its behavior as $n \to \infty$. These calculations are performed in Appendix G, and they lead to:

Proposition 5.1 (Law of Large Numbers)**.** *Let the empirical distribution of the initial position of the particles be $\mu_0^{(n)} = n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{\theta}_i(0)}$ and assume that $\mu_0^{(n)} \rightharpoonup \mu_0$ as $n \to \infty$. Then, for all for $t \in [0, \infty)$, $\mu_t^{(n)} = n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{\theta}_i(t)} \rightharpoonup \mu_t$ in law as $n \to \infty$, where μ_t satisfies (13) with the initial condition $\mu_{t=0} = \mu_0$.*

This statement verifies that, to leading order, the large particle limit recovers the mean-field PDE (13).

While the limit gives rise to the birth-death term of the PDE as expected, we can also quantify the scale and asymptotic behavior of the higher order fluctuations at finite n. This computation ensures that finite n fluctuations do not overcome the convergence expected from the mean-field analysis. To do so, we we introduce the discrepancy distribution defined by the difference, scaled by $\sqrt{n}$, between the empirical distribution and its mean-field limit

$$\omega_t^{(n)} \equiv \sqrt{n}\left(\mu_t^{(n)} - \mu_t\right) \tag{22}$$

where $\mu_t^{(n)}$ is the empirical distribution defined in (7) and μ_t is limit satisfying (1). We can then analyze the generator of the joint process $(\mu_t, \omega_t^{(n)})$ and deduce the following proposition:

Proposition 5.2 (Central Limit Theorem)**.** *In the limit as $n \to \infty$, we have*

$$\omega_t^{(n)} \rightharpoonup \omega_t \qquad \text{in law} \tag{23}$$

where ω_t is Gaussian random distribution with zero mean and whose covariance satisfies a linear equation with a source term proportional to $\alpha|\tilde{V}(\boldsymbol{\theta}, [\mu_t])|\mu_t$, see (117) in Appendix G.

The key consequence of this proposition is that it specifies the scale of the fluctuations of $\mu_t^{(n)}$ above its mean field limit μ_t. First it shows that these fluctuations are on a scale $O(\sqrt{\alpha/n})$. This is why α should be kept $O(1)$ relative to n. While it may appear that increasing α accelerates the rate of convergence at mean-field level, the fluctuations would grow and the $n \to \infty$ and $\alpha \to \infty$ limit do not commute. Second, the relation between the scale of the noise and the magnitude of $|\tilde{V}|\mu_t$ has an important consequence for the convergence of the dynamics: because $|\tilde{V}|\mu_t \to 0$ as $t \to \infty$, the fluctuations are "self-quenching" in the sense that their amplitude diminishes and eventually vanishes as $\mu_t \to \mu_*$. In particular, for both the interacting and non-interacting cases, the only stable fixed point of the equation for the covariance of ω_t is zero.

6. Algorithms

Numerical schemes that converge to the PDEs presented in Sec. 3 are both straightforward to design and easy to implement. In absence of the GD part of the dynamics, we could use Kinetic Monte Carlo (also called the Gillespie algorithm) to simulate birth-death without time-discretization error. However, in the large parameter regime, this would be computationally expensive: every particle has its own exponential clock, and the time between successive birth-death events scales like $1/n$. Because we must time-discretize the GD flow, we carry out the birth-death dynamics using the same time-discretization.

Denote by $\{\boldsymbol{\theta}_i\}_{i=1}^n$ the current configuration of n particles in the interacting potential ℓ in (1). To update the state of these particles, we first consider the effect of the GD flow alone, using a time-discretized approximation of this flow with step of size $\Delta t > 0$. With the forward Euler scheme, this amounts to updating the particle positions as

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \nabla F(\boldsymbol{\theta}_i)\Delta t - \frac{1}{n}\sum_{j=1}^n \nabla K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\Delta t \tag{24}$$

While this type of update is standard in machine learning, more accurate integration schemes could be used.

To implement the birth-death part of the dynamics, we calculate the probability of survival of the particles assuming that their position was fixed at the current values $\{\boldsymbol{\theta}_i\}_{i=1}^n$ using the empirical value $\tilde{V}(\boldsymbol{\theta}_i)$ given in (21) for the rate $V - \bar{V}$. If $\tilde{V}(\boldsymbol{\theta}_i) > 0$ the probability that particle $\boldsymbol{\theta}_i$ be killed in the time interval of size Δt is

$$1 - \exp(\tilde{V}(\boldsymbol{\theta}_i)\Delta t) \tag{25}$$

Similarly, the probability that it is duplicated in that time interval if $\tilde{V}(\boldsymbol{\theta}_i) < 0$ is

$$1 - \exp(|\tilde{V}(\boldsymbol{\theta}_i)|\Delta t) \tag{26}$$

Particles are killed and duplicated in a loop according to this rule. Since $\sum_{i=1}^n \tilde{V}(\boldsymbol{\theta}_i) = 0$ by construction, this operation preserves the number of particles on average. To enforce strict population control, we add an additional loop that guarantees the total population remains fixed after the dynamics above. The details are given in Algorithm 1.

The corresponding particle system is a discretized version, both in particle number and time, of the PDE (13) and it converges to this equation as $n \to \infty$ and $\Delta t \to 0$. The error we make at finite n is analyzed in Sec. 5; the error we make at finite Δt can be deduced from standard results about time discretization of differential equations: with the Euler scheme used above, this error scales as $O(\Delta t)$.

Algorithm 1 Parameter birth-death dynamics consistent with (13)

Δt, initial $\{\boldsymbol{\theta}_i\}_{i=1}^n$ given
$\epsilon = \epsilon_{\rm tol}$, the tolerance
while $\epsilon \geq \epsilon_{\rm tol}$ **do**
  for $i = 1 : n$ **do**
    set $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \nabla F(\boldsymbol{\theta}_i)\Delta t - \frac{1}{n}\sum_{j=1}^n \nabla K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\Delta t$
    calculate $\tilde{V}(\boldsymbol{\theta}_i) = F(\boldsymbol{\theta}_i) + n^{-1}\sum_{j=1}^n K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - n^{-1}\sum_{j=1}^n \left(F(\boldsymbol{\theta}_j) + n^{-1}\sum_{k=1}^n K(\boldsymbol{\theta}_j, \boldsymbol{\theta}_k)\right)$
    if $\tilde{V}(\boldsymbol{\theta}_i) > 0$ **then**
      kill $\boldsymbol{\theta}_i$ w/ prob $1 - \exp(-\alpha\tilde{V}(\boldsymbol{\theta}_i)\Delta t)$
    else if **if** $\tilde{V}(\boldsymbol{\theta}_i) < 0$ **then**
      duplicate $\boldsymbol{\theta}_i$ w/ prob $1 - \exp(-\alpha|\tilde{V}(\boldsymbol{\theta}_i)|\Delta t)$
    end if
  end for
  N_1: total number of particles after the loop
  if $N_1 > N$ **then**
    kill $N_1 - N$ randomly selected particles
  else if **if** $N_1 < N$ **then**
    duplicate $N - N_1$ randomly selected particles
  end if
end while

In the case of neural network parameter optimization, the birth-death algorithm does not incur any significant computational cost beyond regular stochastic gradient descent. Denoting the parameters $\boldsymbol{\theta}_i = (c_i, \boldsymbol{z}_i)$ and writing the neural network function as

$$f_n(\boldsymbol{x}; \{c_i, \boldsymbol{z}_i\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n c_i \phi(\boldsymbol{x}, \boldsymbol{z}_i), \tag{27}$$

the potential $V(\boldsymbol{\theta}_i) = F(\boldsymbol{\theta}_i) + n^{-1}\sum_{j=1}^n K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ is given by $V(\boldsymbol{\theta}_i) = c_i \hat{V}(\boldsymbol{z}_i)$ where

$$\hat{V}(\boldsymbol{z}_i) = \int_\Omega \phi(\boldsymbol{x}, \boldsymbol{z}_i)\left(f_n(\boldsymbol{x}; \{c_i, \boldsymbol{z}_i\}_{i=1}^n) - f(\boldsymbol{x})\right)\nu(d\boldsymbol{x}) \tag{28}$$

Note that $\hat{V}$ is the gradient of the loss with respect to the linear coefficient vector $\partial_{c_i} V = \hat{V}(\boldsymbol{z}_i)$. Because we do not typically have access to the exact loss function, the integrals required to compute $\hat{V}$ are estimated using a finite number of data points. Using a batch of P points in an update leads to an estimate $\hat{V}_P$ of $\hat{V}$, which is used to determine the rate of killing/duplication. In this particular case, the only change to Algorithm 1 is that the computation of $\tilde{V}$ is replaced with $c_i \hat{V}_P(\boldsymbol{z}_i) - n^{-1}\sum_{j=1}^n c_j \hat{V}_P(\boldsymbol{z}_j)$ with

$$\hat{V}_P(\boldsymbol{z}_i) = \frac{1}{P}\sum_{p=1}^P \phi(\boldsymbol{x}_p, \boldsymbol{z}_i)\left(f_n(\boldsymbol{x}_p; \{c_i, \boldsymbol{z}_i\}_{i=1}^n) - f(\boldsymbol{x}_p)\right) \tag{29}$$

where the "batch" is $\{\boldsymbol{x}_p\}_{p=1}^P$. Since this quantity is computed in the SGD update, the only additional computation is the sum of V_P over the n particles. The cost of the algorithm is $O(nP)$ at every iteration.

For neural networks of the form given in Eq. (27) a particularly simple modification of Algorithm 1 enables particle creation from a prior distribution. The algorithm proceeds through the initial birth-death loop as in Algorithm 1. At the end of the initial loop, if the total population has decreased, then additional particle are sampled with configurations $(c, \boldsymbol{z})$ distributed according to the prior distribution

$$\mu_{\mathrm{b}}(dc, d\boldsymbol{z}) = \delta_0(dc)\bar{\rho}(\boldsymbol{z})d\boldsymbol{z} \tag{30}$$

so that a reinjected particle makes no contribution to the energy. Finally, one can alternatively implement the same dynamics using the proximal interpretation of unbalanced transport. This is discussed in Appendix B.1.

7. Numerical Experiments

7.1. Mixture of Gaussians

We take as an illustrative example a mixture of Gaussians in dimension d,

$$f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m \frac{\bar{c}_i}{(2\pi\sigma_i^2)^{d/2}} e^{-|\boldsymbol{x}-\bar{\boldsymbol{z}}_i|^2/(2\sigma_i^2)}, \tag{31}$$

which we approximate as a neural network with Gaussian nonlinearities with fixed standard deviation $\sigma < \min_i \sigma_i$,

$$f_n(\boldsymbol{x}; \{c_i, \boldsymbol{z}_i\}) = \frac{1}{n}\sum_{i=1}^n \frac{c_i}{(2\pi\sigma^2)^{d/2} n} e^{-|\boldsymbol{x}-\boldsymbol{z}_i|^2/(2\sigma^2)}. \tag{32}$$

This is a useful test of our results because we can do exact gradient descent dynamics on the mean-squared loss function:

$$\ell(\{c_i, \boldsymbol{z}_i\}) = \frac{1}{2}\int_{\mathbb{R}^d} |f(\boldsymbol{x}) - f_n(\boldsymbol{x}; \{c_i, \boldsymbol{z}_i\})|^2 \, d\boldsymbol{x} \tag{33}$$

Because all the integrals are Gaussian, this loss can be computed analytically, and so can $\tilde{V}$ and its gradient.

In Fig. 6, we show convergence to the energy minimizer for a mixture of three Gaussians (details and source code are provided in the SM). The non-local mass transport dynamics dramatically accelerates convergence towards the minimizer. While gradient descent eventually converges in this setting—there is no metastability—the dynamics are particularly slow as the mass concentrates near the minimum and maxima of the target function. However, with the birth-death dynamics, this mass readily appears at those locations. The advantage of the birth-death dynamics with a reinjection distribution μ_{b} is highlighted by choosing an unfavorable initialization in which the particle mass is concentrated around $y = -2$. In this case, both GD and GD with birth-death (12) do not converge on the timescale of the dynamics. With the reinjection distribution, new mass is created near $y = 2$ and convergence is achieved.

7.2. Student-Teacher ReLU Network

In many optimization problems, it is not possible to evaluate $\tilde{V}$ exactly. Instead, typically $\tilde{V}$ is estimated as a sample mean over a batch of data. We consider a student-teacher set-up similar to (Chizat & Bach, 2018a) in which we use single hidden layer ReLU networks to approximate a network of the same type with fewer neurons. We use as the target function a ReLU network with 50-d input and 10 hidden units. We approximate the teacher with neural networks with $n = 50$ neurons (see SM). The networks are trained with stochastic gradient descent (SGD) and the mini-batch estimate of the gradient of output layer, which is computed at each step of SGD, is used to compute $\tilde{V}$, which determines the rate of birth-death. In experiments with the reinjection distribution, we use (30) with Gaussian $\bar{\rho}$.

As shown in Fig. 2, we find that the birth-death dynamics accelerates convergence to the teacher network. We emphasize that because the birth-death dynamics is stochastic at finite particle numbers, the fluctuations associated with the process could be unfavorable in some cases. In such situations, it is useful to reduce α as a function of time. On the other hand, in some cases we have observed much more dramatic accelerations from the birth-death dynamics.

8. Conclusions

The success of gradient descent requires good coverage of the parameter space so that local updates can reach the min-

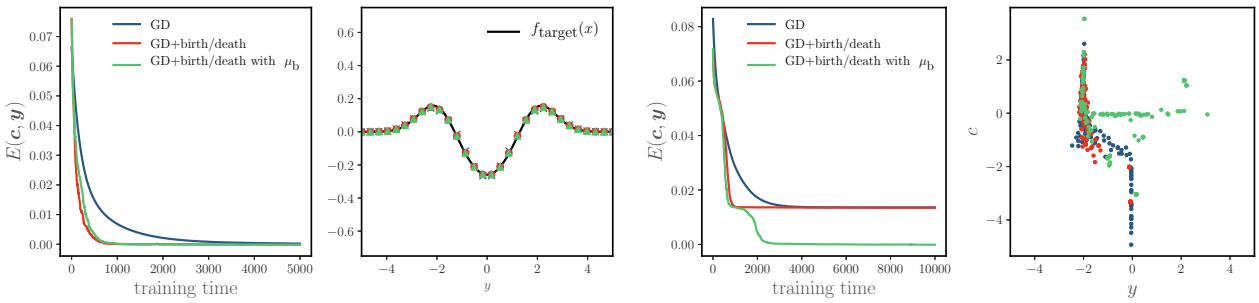

Figure 1. Top left: Convergence of the gradient descent dynamics without birth-death, with birth-death, and using a reinjection distribution. Top right: For appropriate initialization, the three dynamical schemes all converge to the target function. Bottom left: For bad initialization (narrow Gaussian distributed around y=-2), GD and GD+birth-death do not converge on this timescale. Interestingly, with the reinjection via distribution μ_b, convergence to the global minimum is rapidly achieved. Bottom right: The configuration of the particles in $\boldsymbol{\theta} = (y, c)$. Only with the reinjection distribution does mass exist near $y = 2$.

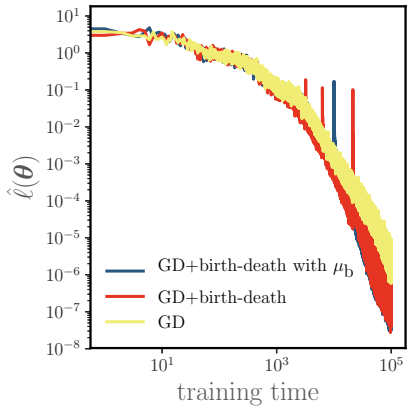

Figure 2. The batch loss as a function of training time for the student-teacher ReLU network described in Sec. 7.2. The birth-death dynamics accelerates convergence, both with and without the reinjection distribution.

ima of the loss function quickly. Our approach liberates the parameters from a purely local dynamics and allows rapid reallocation to values at which they can best reduce the approximation error. At the mean-field level, our dynamics amount to a form of unbalanced transport, that we implement at finite-particle level using a birth/death process. These new dynamics provably converge to the minimizers of the loss function for a general class of minimization problems.Remarkably, for interacting systems with we can guarantee global convergence for sufficiently regular initial conditions. We have also computed the asymptotic rate of convergence with birth-death dynamics.

These theoretical results translate into significant reductions in convergence time for our illustrative examples. Importantly, the schemes we have described are straightforward to implement and come with little computational overhead. Extending this type of dynamics to deep neural network architectures could accelerate the slow dynamics at the initial layers often observed in practice. Hyperparameter selection strategies based on evolutionary algorithms (Such et al., 2017) provide another interesting potential application of our approach.

While we have characterized the basic behavior of optimization under the birth-death dynamics, many theoretical questions remain. First, we did not address generalization; understanding the role of the additional birth-death term in controlling the generalization gap is an important future question, in particular relating it to the "lazy-training" regime of (Chizat & Bach, 2018a). Next, we need to assume the existence of weak solutions via (14) with an initial measure μ_0 that has full support, though it may be possible to certify that the dynamics exist for all times if μ_0 decays sufficiently fast. In addition, more explicit calculations of global convergence rates for the interacting case and also tighter rates for the non-interacting case would be exciting additions. The proper choice of μ_b is another question worth exploring because, as highlighted in our simple example, favorable reinjection distributions can rapidly overcome slow dynamics. Finally, a mean-field perspective on deep neural networks would enable us to translate some of the guarantees here to deep architectures.

Acknowledgments

We thank Sylvia Serfaty and Yann Ollivier, as well as the anonymous reviewers, for their detailed comments. GR was supported by the James S. McDonnell Foundation. JB was partially supported by NSF grant RI-IIS 1816753, NSF CAREER CIF 1845360, the Alfred P. Sloan Fellowship, Samsung Electronics. EVE was partially supported by the Materials Research Science and Engineering Center (MRSEC) program of the NSF under award number DMR-1420073 and by NSF under award number DMS-1522767.

References

Baluja, S. and Caruana, R. Removing the genetics from the standard genetic algorithm. In *Machine Learning Proceedings 1995*, pp. 38–46. Elsevier, 1995.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993. doi: 10.1109/18.256500. URL http://ieeexplore.ieee.org/document/256500/.

Chizat, L. and Bach, F. A Note on Lazy Training in Supervised Differentiable Programming. working paper or preprint, December 2018a. URL https://hal.inria.fr/hal-01945578.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3040–3050. Curran Associates, Inc., 2018b.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090 – 3123, 2018. ISSN 0022-1236. doi: https://doi.org/10.1016/j.jfa.2018.03.008.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2(4): 303–314, December 1989. doi: 10.1007/BF02551274. URL https://link.springer.com/article/10.1007/BF02551274.

Dawson, D. Measure-valued Markov processes. In *École d'Été de Probabilités de Saint-Flour XXI—1991*, pp. 1–260. Springer Berlin Heidelberg, Berlin, Heidelberg, September 2006.

Gallouët, T. O. and Monsaingeon, L. A jko splitting scheme for kantorovich–fisher–rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

Hansen, N. The cma evolution strategy: a comparing review. In *Towards a new evolutionary computation*, pp. 75–102. Springer, 2006.

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Kennedy, J. Particle swarm optimization. In *Encyclopedia of machine learning*, pp. 760–766. Springer, 2011.

Kondratyev, S., Monsaingeon, L., and Vorotnikov, D. A new optimal transport distance on the space of finite radon measures. *Adv. Diff. Eq.*, 21(11/12):1117–1164, 11 2016.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Larrañaga, P. and Lozano, J. A. *Estimation of distribution algorithms: A new tool for evolutionary computation*, volume 2. Springer Science & Business Media, 2001.

Liero, M., Mielke, A., and Savaré, G. Optimal Entropy-Transport problems and a new Hellinger-Kantorovich distance between positive measures. *Invent. Math.*, 211 (3):969–1117, March 2018.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, August 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1806579115.

Ollivier, Y., Arnold, L., Auger, A., and Hansen, N. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, 18(18):1–65, 2017.

Park, J. and Sandberg, I. W. Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, 3(2):246–257, June 1991. doi: 10.1162/neco.1991.3.2.246. URL http://www.mitpressjournals.org/doi/10.1162/neco.1991.3.2.246.

Rios, L. M. and Sahinidis, N. V. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3): 1247–1293, 2013.

Rotskoff, G. M. and Vanden-Eijnden, E. Neural Networks as Interacting Particle Systems: Asymptotic Convexity of the Loss Landscape and Universal Scaling of the Approximation Error. *arXiv:1805.00915 [cond-mat, stat]*, May 2018. URL http://arxiv.org/abs/1805.00915. arXiv: 1805.00915.

Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Serfaty, S. *Coulomb Gases and Ginzburg–Landau Vortices*. European Mathematical Society Publishing House, Zuerich, Switzerland, March 2015. ISBN 978-3-03719-152-1. doi: 10.4171/152.

Sirignano, J. and Spiliopoulos, K. Mean Field Analysis of Neural Networks. *arXiv*, May 2018. URL http://arxiv.org/abs/1805.01053v1. arXiv: 1805.01053v1.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Such, F. P., Madhavan, V., Conti, E., Lehman, J., Stanley, K. O., and Clune, J. Deep neuroevolution: genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv preprint arXiv:1712.06567*, 2017.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. On the Margin Theory of Feedforward Neural Networks. *arXiv:1810.05369 [cs, stat]*, October 2018. arXiv: 1810.05369.