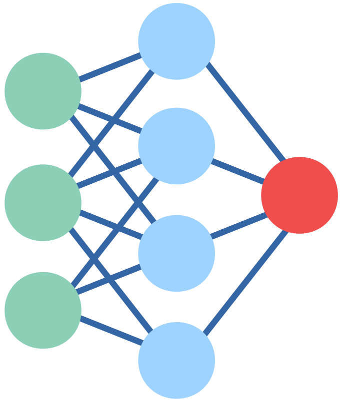
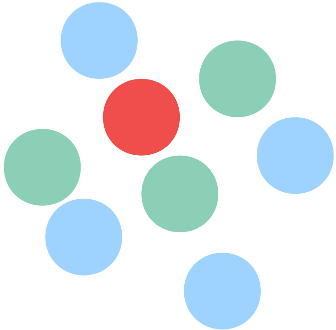
Global convergence of neuron birth-death dynamics



ICML 2019
Long Beach, CA
13 June 2019



Grant M. Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden

Center for Data Science, New York University
Courant Institute of Mathematical Sciences, New York University

Synopsis of main results

→ Modification of gradient descent (GD) with ***unbalanced optimal transport***

- Parameter *birth-death* process
- Proof of **global convergence**
- Rate of convergence scales as t^{-1}

→ Based on ***mean-field perspective*** on neural networks

- Analysis of deterministic partial differential equation (PDE)
- PDE leads to **practical, efficient algorithm**

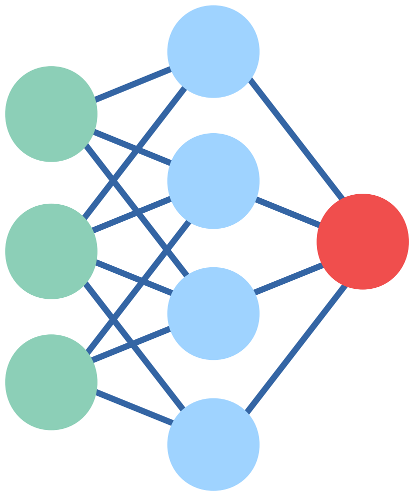
→ Experiments show ***faster convergence*** relative to GD

- Illustrative examples show effect of transport / exploration
- Easy to implement with no additional gradient computations

Single hidden layer neural network

d-dim input

$\boldsymbol{x} \in \Omega \subset \mathbb{R}^d$



n "neurons"

$$f_n(\boldsymbol{x}; \{\boldsymbol{\theta}\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n \varphi(\boldsymbol{x}, \boldsymbol{\theta}_i)$$

$$\varphi(\boldsymbol{x}, \boldsymbol{\theta}) = c\varphi(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)$$
$$\boldsymbol{\theta} = \{c, \boldsymbol{w}, b\}$$

$$\lim_{n\to\infty}$$

PDE for parameter distribution $\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V)$

Distinct from kernel learning / NTK, dynamics leads to feature selection

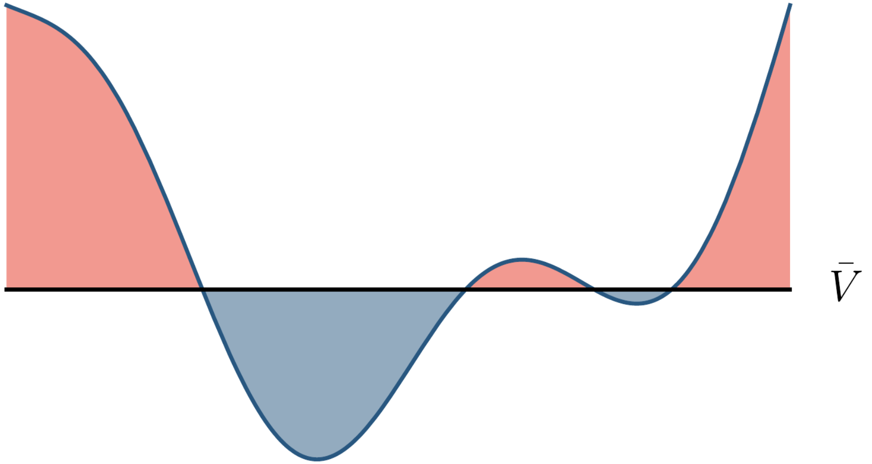
Non-local mass transport (particle birth-death)

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

Parameters are killed / cloned

Total population is fixed

$$\bar{V} \equiv \int_D V d\mu_t$$



Theorem [R,J,B,V-E]: global convergence

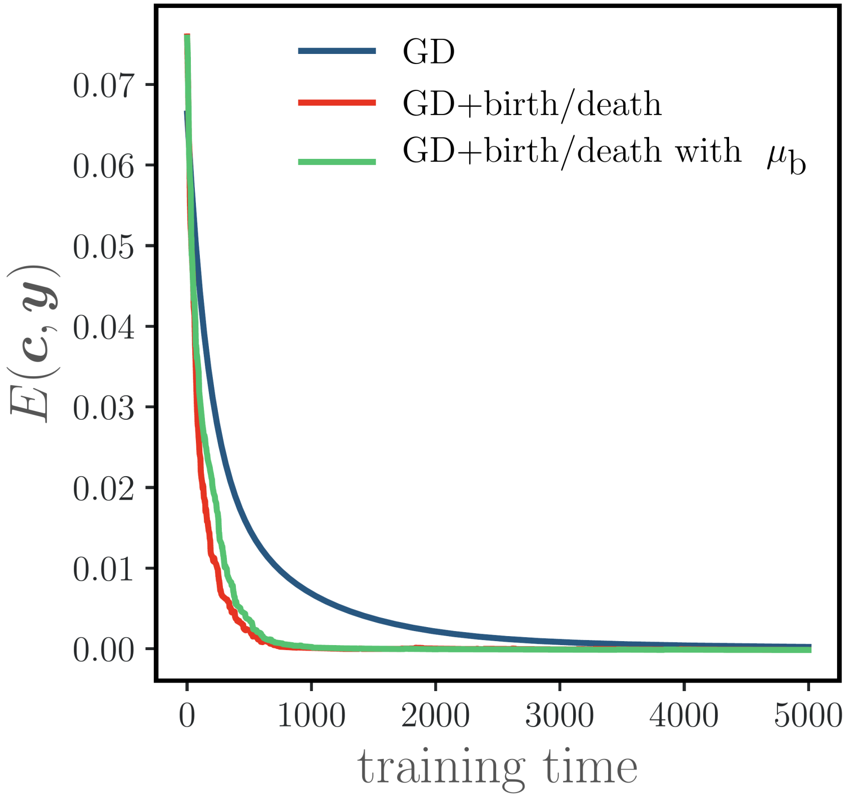
Let μ_t denote the solution of the birth-death PDE for the initial condition μ_0 with $\operatorname{supp}\mu_0 = D$ and with a density ρ_0. Then, under appropriate coerciveness assumptions for the potential,

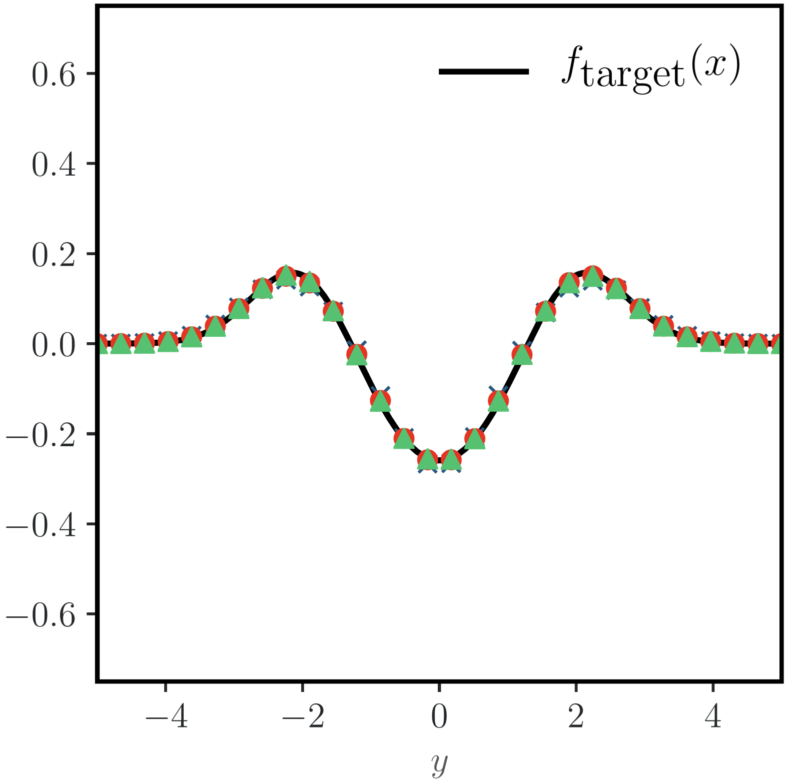
$$\lim_{t\to\infty} \mu_t = \mu_*$$

where μ_* is a global minimizer of $\mathcal{E}[\mu]$

Compare with [Chizat & Bach, 2018] without any restriction on homogeneity for the units

Dramatic effect of birth/death dynamics





Come see our poster!

Thu Jun 13th 06:30 -- 09:00 PM @ Pacific Ballroom #93

Grant Rotskoff
Samy Jelassi
Joan Bruna
Eric Vanden-Eijnden

arXiv:1902.01843