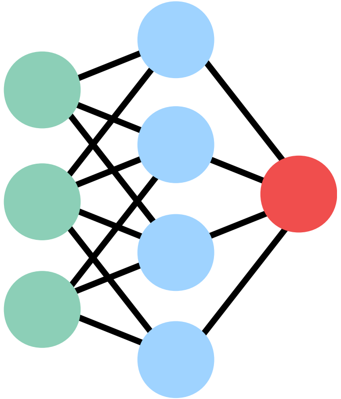
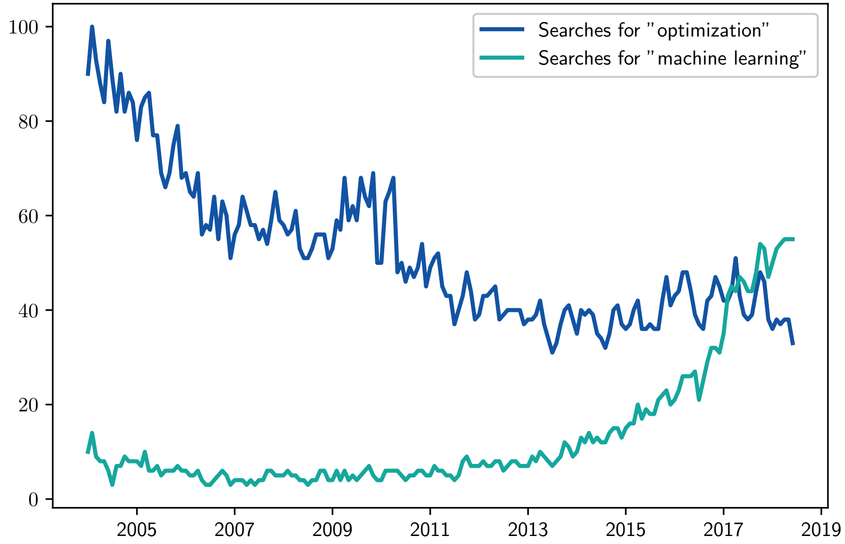
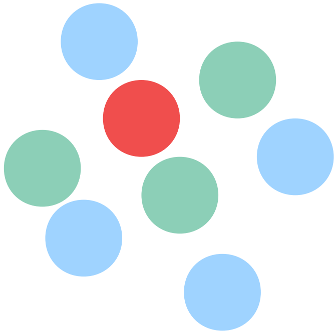
Neural networks as interacting particle systems







Cargèse
August 2018

Grant M. Rotskoff
Courant Institute of Mathematical Sciences
New York University

Joint work with Eric Vanden-Eijnden
arXiv:1805.00915

In principle, *any* function can be represented

Universal Approximation Theorems (Barron, Cybenko, Park, others)

- Says that neural network representations are dense in the space of square-integrable target functions. *There's a neural network arbitrarily close to any such function.*

- The theorems **do not** answer: How do we construct the representation?

1. How should one get to the desired parameters?

2. Do the typical machine learning algorithms converge?

3. Are there guarantees on the error for finite n?

Formalizing neural net optimization

Write the representation as $f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i \varphi(\boldsymbol{x}, \boldsymbol{y}_i)$

- **Radial basis function networks.** $D \subset \Omega$ and $\varphi(\boldsymbol{x}, \boldsymbol{y}) \equiv \phi(\boldsymbol{x} - \boldsymbol{y})$ where ϕ is a radial function

  $$\phi(\boldsymbol{x}) = \exp\left(-\tfrac{1}{2}\kappa|\boldsymbol{x}|^2\right)$$

  where $\kappa > 0$ is a fixed constant.

- **Single-layer neural networks.** $D \subset \mathbb{R}^{d+1}$ and $\varphi(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x}, \boldsymbol{a}, b)$ with $\boldsymbol{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$, and

  $$\varphi(\boldsymbol{x}, \boldsymbol{a}, b) = h(\boldsymbol{a} \cdot \boldsymbol{x} + b)$$

  where $h : \mathbb{R} \to \mathbb{R}$ is e.g. a sigmoid function $h(z) = 1/(1 + e^{-z})$.

Physical interpretation as a particle system

The loss function is $\ell(f, f_n) = \frac{1}{2} \int_\Omega |f(\boldsymbol{x}) - f_n(\boldsymbol{x})|^2 \, d\mu(\boldsymbol{x})$

Energy function

Which we can expand as $\ell(f, f_n) = C_f - \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} c_i F(\boldsymbol{y}_i) + \dfrac{1}{2n^2} \displaystyle\sum_{i,j=1}^{n} c_i c_j K(\boldsymbol{y}_i, \boldsymbol{y}_j)$

charges

particles

Using $f_n(\boldsymbol{x}) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} c_i \varphi(\boldsymbol{x}, \boldsymbol{y}_i)$, we define,

Single body potential

$$F(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x}) \varphi(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{x})$$

Interaction potential

$$K(\boldsymbol{y}, \boldsymbol{z}) = \int_\Omega \varphi(\boldsymbol{x}, \boldsymbol{y}) \varphi(\boldsymbol{x}, \boldsymbol{z}) d\mu(\boldsymbol{x})$$

The nonequilibrium dynamics of optimization

The neural net representation

$$f_n(t,\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} C_i(t)\varphi(\boldsymbol{x},\boldsymbol{Y}_i(t))$$

Evolves according to gradient descent on a complex landscape,

$$\begin{cases} \dot{\boldsymbol{Y}}_i = C_i \nabla F(\boldsymbol{Y}_i) - \frac{1}{n}\sum_{j=1}^{n} C_i C_j \nabla K(\boldsymbol{Y}_i,\boldsymbol{Y}_j), \\ \dot{C}_i = F(\boldsymbol{Y}_i) - \frac{1}{n}\sum_{j=1}^{n} C_j K(\boldsymbol{Y}_i,\boldsymbol{Y}_j) \end{cases}$$

We can (and will) extend this to Langevin dynamics, stochastic gradient descent

Interpreting the limit: McKean-Vlasov Equation

Work with the particle density:

$$\rho_n(t,\boldsymbol{y},c) = \frac{1}{n}\sum_{i=1}^{n}\delta(c - C_i(t))\delta(\boldsymbol{y} - \boldsymbol{Y}_i(t))$$

So that the representation satisfies

$$f_n(t,\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} C_i(t)\varphi(\boldsymbol{x},\boldsymbol{Y}_i(t)) = \int_{D\times R} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_n(t,\boldsymbol{y},c)d\boldsymbol{y}dc$$

$$\begin{aligned}\partial_t\rho_n = \nabla\cdot&\left(-c\nabla F\rho_n + \int_{D\times\mathbb{R}} cc'\nabla K(\boldsymbol{y},\boldsymbol{y}')\rho_n'\rho_n d\boldsymbol{y}'dc'\right)\\ &+\partial_c\left(-F\rho_n + \int_{D\times\mathbb{R}} c'K(\boldsymbol{y},\boldsymbol{y}')\rho_n'\rho_n d\boldsymbol{y}'dc'\right)\end{aligned}$$

Asymptotic convexity (but that's not the whole story)

$$\partial_t \rho_0 = \nabla \cdot \left(\rho_0 \nabla \frac{\delta \mathcal{E}_0}{\delta \rho_0} \right) + \partial_c \left(\rho_0 \partial_c \frac{\delta \mathcal{E}_0}{\delta \rho_0} \right)$$

$$\begin{aligned} \mathcal{E}_0[\rho_0] &= C_f - \int_{D\times\mathbb{R}} cF\rho_0 d\boldsymbol{y} dc + \tfrac{1}{2} \int_{(D\times\mathbb{R})^2} cc' K(\boldsymbol{y}, \boldsymbol{y}') \rho_0 \rho_0' d\boldsymbol{y} dc d\boldsymbol{y}' dc' \\ &= \tfrac{1}{2} \int_\Omega \left(f(\boldsymbol{x}) - \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x}, \boldsymbol{y}) \rho_0 d\boldsymbol{y} dc \right)^2 d\mu(\boldsymbol{x}) \geq 0 \end{aligned}$$

- Fixed points of this equation, written as above, can occur if the density singular
- Not all fixed points are energy minimizers
- Continuous dynamics in the charges changes this story

Law of Large Numbers (Gradient Descent)

Let $f_n(t) = f_n(t, \boldsymbol{x})$ with $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ evolve according to gradient descent with initial condition drawn from some $\mathbb{P}_{\text{in}}$. Then

$$\lim_{n\to\infty} f_n(t) = f_0(t) \qquad \mathbb{P}_{\text{in}}\text{-almost surely} \tag{1}$$

where $f_0(t)$ solves the differential equation below and satisfies

$$\lim_{t\to\infty} f_0(t) = f \quad \text{a.e. in} \quad \Omega \tag{2}$$

See also: Mei, Montanari, Pham and Sirignano, Spilliopolous

Evolution equation for the representation is inherited from evolution of density,

$$\partial_t f_0(t, \boldsymbol{x}) = \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x}, \boldsymbol{y})\partial_t \rho_0(t, \boldsymbol{y}, c) d\boldsymbol{y} dc$$

Error term (and error scaling)

Let $f_n(t) = f_n(t, \boldsymbol{x})$ and with $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ evolve according to gradient descent with initial condition drawn from $\mathbb{P}_{\text{in}}$. Then for any $\bar{\xi} < 1$ and any $a_n > 0$ such that $a_n / \log n \to \infty$ as $n \to \infty$, we have

$$\lim_{n\to\infty} n^{\bar{\xi}} \left(f_n(a_n) - f_0(a_n) \right) = 0 \qquad \text{almost surely}$$

where satisfies $f_0(t) \to f$ as $t \to \infty$.

Initial scaling from CLT | Asymptotic scaling after optimization

$$\xi = 1/2 \longrightarrow \xi = 1$$

The story with stochastic gradient descent

Stochastic samples of the potential / interaction term:

$$F_P(t,\boldsymbol{y}) = \frac{1}{P}\sum_{p=1}^{P} f(\boldsymbol{X}_p(t))\varphi(\boldsymbol{X}_p(t),\boldsymbol{y}), \qquad K_P(t,\boldsymbol{y},\boldsymbol{y}') = \frac{1}{P}\sum_{p=1}^{P}\varphi(\boldsymbol{X}_p(t),\boldsymbol{y})\varphi(\boldsymbol{X}_p(t),\boldsymbol{y}')$$

Leading to an SDE,

$$d\boldsymbol{Z} =_{\boldsymbol{z}} \ell(f, f_n(\boldsymbol{z}))dt + \sqrt{\theta}d\boldsymbol{B}$$

Where the quadratic variation of the noise is

$$\mathbb{E}\left(\nabla_{\boldsymbol{z}}(L_P(\boldsymbol{z}) - n\ell(f, f_n(\boldsymbol{z})))\right) \otimes \left(\nabla_{\boldsymbol{z}}(L_P(\boldsymbol{z}') - n\ell(f, f_n(\boldsymbol{z}')))\right) = \frac{1}{P}R(\boldsymbol{z})$$

$$\theta = \Delta t/P = \text{“timestep/batch size”}$$

Dean's equation for correlated noise terms

$$
\begin{aligned}
\partial_t \rho_n = \nabla \cdot \left(-c \nabla F \rho_n + \int_{D\times\mathbb{R}} c c' \nabla K(\boldsymbol{y}, \boldsymbol{y}') \rho_n' \rho_n d\boldsymbol{y}' dc' \right) \\
+ \partial_c \left(-F \rho_n + \int_{D\times\mathbb{R}} c' K(\boldsymbol{y}, \boldsymbol{y}') \rho_n' \rho_n d\boldsymbol{y}' dc' \right) \\
+ \tfrac{1}{2} \theta \nabla\nabla : \left(\rho_n c^2 A_2([f_n(t) - f], \boldsymbol{y}, \boldsymbol{y}) \right) + \tfrac{1}{2} \theta \partial_c^2 \left(\rho_n A_0([f_n(t) - f], \boldsymbol{y}, \boldsymbol{y}) \right) \\
+ \theta \partial_c \nabla \cdot \left(\rho_n c A_1([f_n(t) - f], \boldsymbol{y}, \boldsymbol{y}) \right) \\
+ \sqrt{\theta}\, \dot{\eta}_n(t, \boldsymbol{y}, c)
\end{aligned}
$$

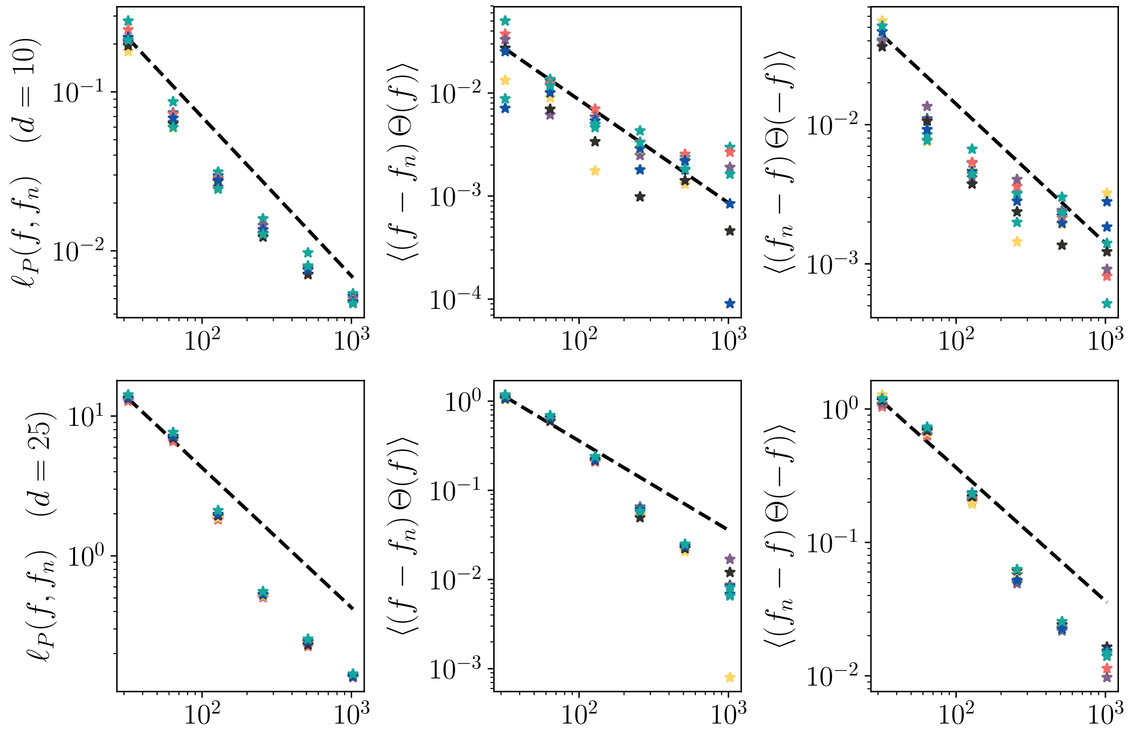
$P = n^2 \implies$

Same first order term as gradient descent

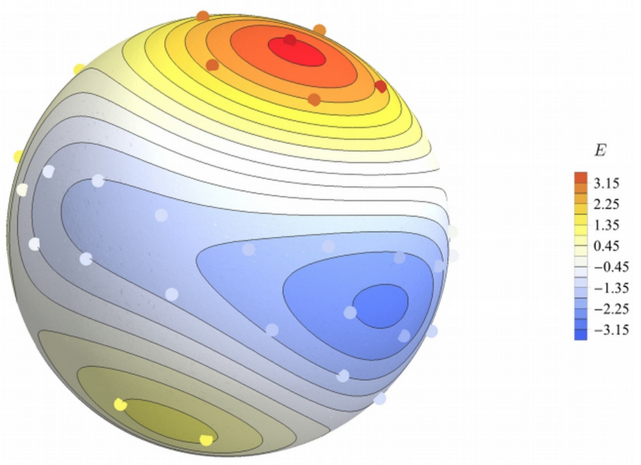
Guarantee of convergence

Recover the error scaling

Confirming the scaling



Sigmoid NN Error Scaling



RBF Centers

Conclusions

Neural networks are a potentially powerful tool for computational physics and applied mathematics. They can massively reduce the cost of representing functions in high dimensional spaces.

By interpreting the parameters as interacting particles, we can demonstrate the asymptotic convexity of the loss landscape, in the process showing that stochastic gradient descent converges to an energy minimizer, with an appropriate quench.

The error and its scaling can be identified up to a constant, which shows that errors can be controlled precisely in the limit.

Applications to free energy methods, quantum variational energy calculations, and PDEs are only beginning to be explored.