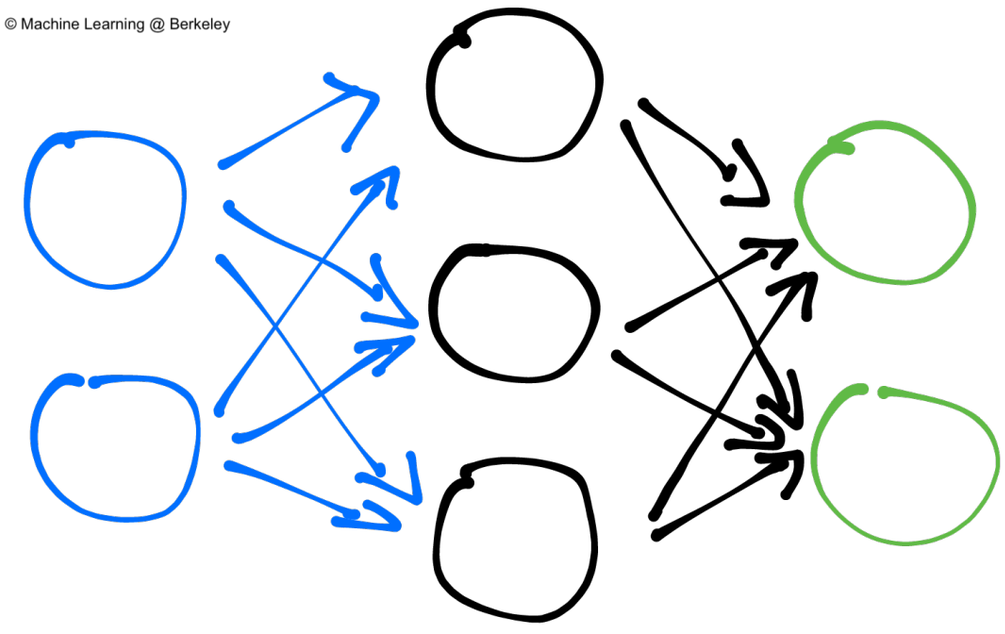
Neural Networks as Interacting Particle Systems



Eric Vanden-Eijnden
Courant Institute

Joint work with Grant Rotskoff

Ref: arXiv:1805.00915

Advances in Computational Statistical Physics, CIRM 2018

The unreasonable effectiveness of machine learning

- (Deep) neural networks have led to extraordinary progress in speech and image recognition, language processing and translation, object detection, etc.

- Problems assumed to be intractable a decade ago are now routine. Are image / speech recognition inherently low dimensional?

- *Alternatively, can neural networks accurately represent high-dimensional data / functions?*

$$f : \{0,1\}^{N^2} \to \{0,1,\ldots,9,\mathrm{NaN}\}$$

- Standard representations, like Galerkin truncations or finite element decompositions, are linear and cannot be scaled to high dimensional problem — *curse of dimensionality.*

- In contrast, neural networks are nonlinear in their adjusting parameters — their effective dimensionality is much higher than the number of these parameters.

Basic roadmap of neural network training

- Approximate a target function $f : \Omega \to \mathbb{R}$ defined on $\Omega \subseteq \mathbb{R}^d$ by a neural network representation, e.g. a *single-layer network with sigmoid nonlinearity*

$$f_n(\boldsymbol{x}) = \sum_{i=1}^{n} c_i h(\boldsymbol{a}_i \cdot \boldsymbol{x} + b_i), \qquad h(z) = 1/(1 + e^{-z})$$

non-linear approximation

- Measure the approximation error via the *loss function*

$$\ell(f, f_n) = \tfrac{1}{2} \int_\Omega |f(\boldsymbol{x}) - f_n(\boldsymbol{x})|^2 \, d\mu(\boldsymbol{x}) = \tfrac{1}{2} \mathbb{E}_{\text{data}} \, |f - f_n|^2$$

- In practice, estimate $\ell(f, f_n)$ via the *empirical loss function*

$$\ell_P(f, f_n) = \frac{1}{P} \sum_{p=1}^{P} |f(\boldsymbol{x}_p) - f_n(\boldsymbol{x}_p)|^2, \qquad \{\boldsymbol{x}_p\}_{p=1}^{P} = \textit{iid drawn from } \mu.$$

- Train the network via *stochastic gradient descent (SGD)* to minimize the loss over the parameters

non-convex optimization problem

$$\boldsymbol{z}(t + \Delta t) = \boldsymbol{z}(t) - \Delta t \nabla_{\boldsymbol{z}} L_P(\boldsymbol{z}(t)),$$

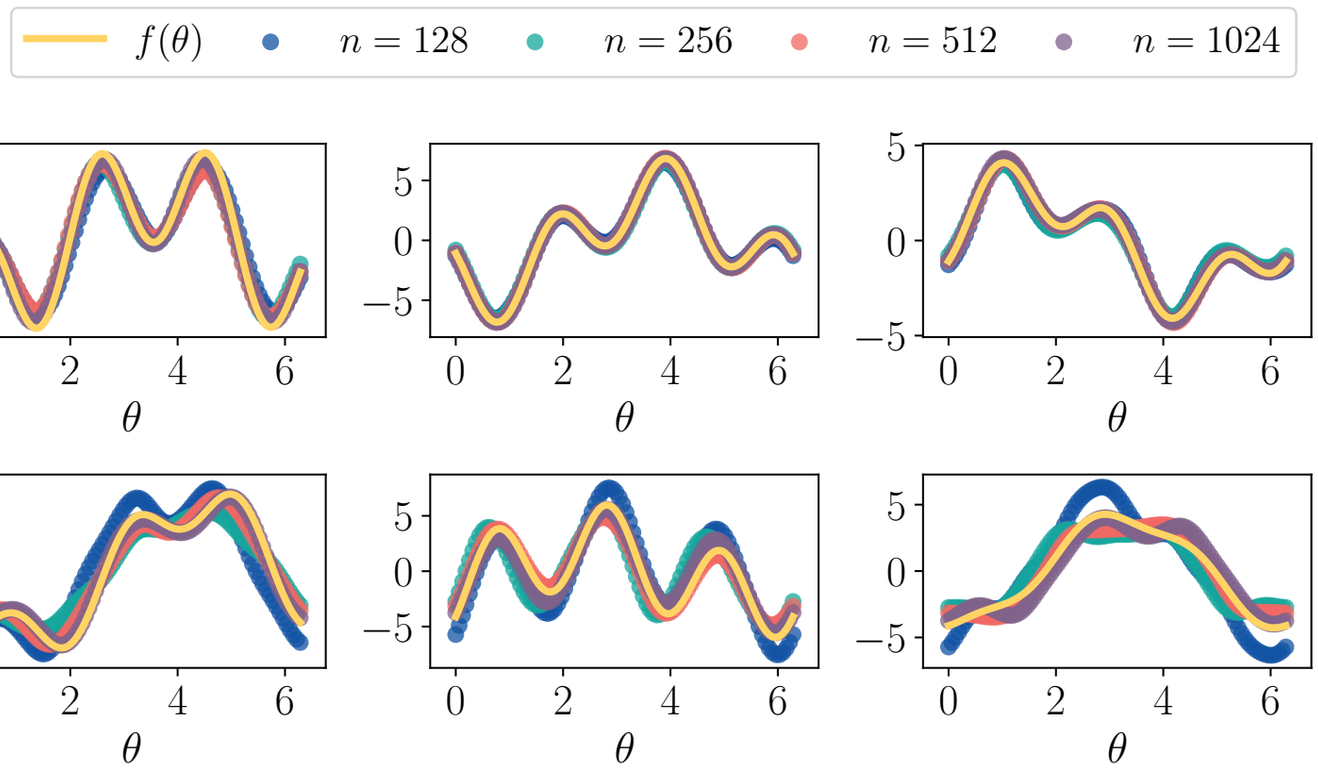
where $\boldsymbol{z}$ denotes the network parameters (e.g. $\boldsymbol{z} = (\boldsymbol{a}_1, b_1, c_1, \ldots, \boldsymbol{a}_n, b_n, c_n)$) and $L_P(\boldsymbol{z}) = \ell_P(f, f_n)$ the empirical loss viewed as function of $\boldsymbol{z}$.

A test case: Spherical 3-spin model

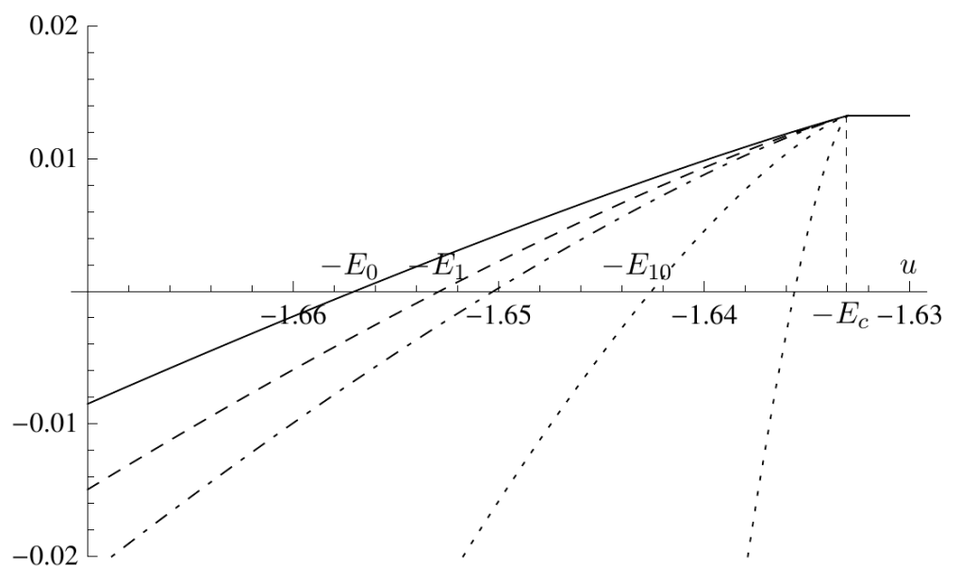
$$f(\boldsymbol{x}) = \frac{1}{d} \sum_{p,q,r=1}^{d} a_{p,q,r} x_p x_q x_r, \qquad \boldsymbol{x} \in S^{d-1}\left(\sqrt{d}\right) \subset \mathbb{R}^d$$

$$a_{p,q,r} \sim N(0,1)$$

Number of critical points exponential in d
(Ben Arous, etc.)



$$x_i(\theta) = \sqrt{d}\cos(\theta), \qquad x_j(\theta) = \sqrt{d}\sin(\theta), \qquad x_k(\theta) = 0 \quad \forall k \neq i, j.$$



$d = 25$ with $n = 1024$,

Two grid points per dimension
$= 2^{25} = 33,554,432$ grid points

Representation power of neural networks

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

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

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

    1. Can the network be trained (i.e. how should we get the parameters)?

    2. Do the typical machine learning algorithms converge?

    3. How does the error scale with the network size?

Neural networks as particle systems

- Using $f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i \varphi(\boldsymbol{x}, \boldsymbol{y}_i)$, the loss function $\ell(f, f_n) = \frac{1}{2}\int_\Omega |f(\boldsymbol{x}) - f_n(\boldsymbol{x})|^2 \, d\mu(\boldsymbol{x})$ can be expanded as

$$\ell(f, f_n) = C_f - \frac{1}{n}\sum_{i=1}^{n} c_i F(\boldsymbol{y}_i) + \frac{1}{2n^2}\sum_{i,j=1}^{n} c_i c_j K(\boldsymbol{y}_i, \boldsymbol{y}_j), \qquad C_f = \frac{1}{2}\int_\Omega |f(\boldsymbol{x})|^2 \, d\mu(\boldsymbol{x})$$

where

$$F(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{x}), \qquad K(\boldsymbol{y}, \boldsymbol{z}) = \int_\Omega \varphi(\boldsymbol{x}, \boldsymbol{y})\varphi(\boldsymbol{x}, \boldsymbol{z}) d\mu(\boldsymbol{x}) \equiv K(\boldsymbol{z}, \boldsymbol{y}).$$

- Minimization of the loss over $\{c_i, \boldsymbol{y}_i\}_{i=1}^n$ is a complex, presumably non-convex optimization problem.

Neural networks as particle systems

- Using $f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i \varphi(\boldsymbol{x}, \boldsymbol{y}_i)$, the loss function $\ell(f, f_n) = \frac{1}{2}\int_\Omega |f(\boldsymbol{x}) - f_n(\boldsymbol{x})|^2 \, d\mu(\boldsymbol{x})$ can be expanded as

$$\ell(f, f_n) = C_f - \frac{1}{n}\sum_{i=1}^{n} c_i F(\boldsymbol{y}_i) + \frac{1}{2n^2}\sum_{i,j=1}^{n} c_i c_j K(\boldsymbol{y}_i, \boldsymbol{y}_j), \qquad C_f = \frac{1}{2}\int_\Omega |f(\boldsymbol{x})|^2 \, d\mu(\boldsymbol{x})$$

where

interaction potential

$$F(\boldsymbol{y}) = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{x}), \qquad K(\boldsymbol{y}, \boldsymbol{z}) = \int_\Omega \varphi(\boldsymbol{x}, \boldsymbol{y})\varphi(\boldsymbol{x}, \boldsymbol{z}) d\mu(\boldsymbol{x}) \equiv K(\boldsymbol{z}, \boldsymbol{y}).$$

one-body interaction — two-body interaction

- Minimization of the loss over $\{c_i, \boldsymbol{y}_i\}_{i=1}^n$ is a complex, presumably non-convex optimization problem.

- Parameters = particles; loss function = interaction potential

Non-equilibrium dynamics of training / optimization

- Gradient descent (GD) dynamics over loss function: $f_n(t,\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} C_i(t)\varphi(\boldsymbol{x}, \boldsymbol{Y}_i(t))$ with

$$\frac{d\boldsymbol{Y}_i}{dt} = C_i \nabla F(\boldsymbol{Y}_i) - \frac{1}{n}\sum_{j=1}^{n} C_i C_j \nabla K(\boldsymbol{Y}_i, \boldsymbol{Y}_j),$$

$$\frac{dC_i}{dt} = F(\boldsymbol{Y}_i) - \frac{1}{n}\sum_{j=1}^{n} C_j K(\boldsymbol{Y}_i, \boldsymbol{Y}_j)$$

(to be generalized to SGD)

- Empirical distribution:

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

- Dynamics = McKean-Vlasov equation

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

Asymptotic convexity in the mean field limit

- McKean-Vlasov equation = gradient descent flow in Wasserstein metric (Otto, Villani, Serfaty, etc.)
- Mean-field limit: $\rho_n(t) \rightharpoonup \rho_0(t)$ as $n \to \infty$, where $\rho_0(t)$ descents on quadratic energy:

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

Propagation of chaos

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

- Ruggedness of the microscopic landscape viewed by particles / parameters disappears at the level of their empirical distribution.

Asymptotic convexity

Not the full story: particular shape of interaction potential from network architecture matters to avoid stationary points

Similar results appeared recently in:

Mei, Montanari, & Nguyen arXiv:1804.06561;
Sirigano & Spiliopoulos arXiv:1805.01053

Law of Large Numbers (LLN)

- $\rho_n(t) \rightharpoonup \rho_0(t)$ as $n \to \infty$, where $f_0(t,\boldsymbol{x}) = \int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_0(t,\boldsymbol{y},c)d\boldsymbol{y}dc$ satisfies gradient flow over loss function:

$$\begin{aligned}\partial_t f_0(t,\boldsymbol{x}) &= -\int_\Omega M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')\left(f_0(t,\boldsymbol{x}') - f(\boldsymbol{x}')\right)d\mu(\boldsymbol{x}') \\ &= -\int_\Omega M([\rho_0(t)],\boldsymbol{x},\boldsymbol{x}')\, D_{f_0(t,\boldsymbol{x}')}\ell(f,f_0(t))d\mu(\boldsymbol{x}')\end{aligned} \tag{1}$$

where $D_{f(\boldsymbol{x})}$ denotes the gradient with respect to $f(\boldsymbol{x})$ in the $L^2(\Omega,\mu)$-norm and

$$M([\rho],\boldsymbol{x},\boldsymbol{x}') = \int_{D\times\mathbb{R}} \left(c^2\nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x},\boldsymbol{y})\nabla_{\boldsymbol{y}}\varphi(\boldsymbol{x}',\boldsymbol{y}) + \varphi(\boldsymbol{x},\boldsymbol{y})\varphi(\boldsymbol{x}',\boldsymbol{y})\right)\rho(\boldsymbol{y},c)d\boldsymbol{y}dc.$$

Prop 1 (LLN) *Let* $f_n(t) = \frac{1}{n}\sum_{i=1}^n C_i(t)\varphi(\cdot,\boldsymbol{Y}_i(t))$. *Then*

$$\lim_{n\to\infty} f_n(t) = f_0(t) \qquad \textit{almost surely}$$

where $f_0(t)$ *solves* (1) *and satisfies*

$$\lim_{t\to\infty} f_0(t) = f \quad \textit{a.e. in } \Omega \qquad \textit{if} \quad \lim_{t\to\infty}\int_{\mathbb{R}} \rho_0(\cdot,c)dc > 0 \quad \textit{a.e. in } D$$

- Dynamical variant of Universal Representation Theorem which also indicates how to realize it in practice.

- Condition $\int_{\mathbb{R}} \rho_0(\cdot,c)dc > 0$ a.e.D generic under small perturbations.

Error scaling

parameters drawn independently at t=0

Prop 2 (CLT) *Let* $f_n(t) = \frac{1}{n}\sum_{i=1}^{n} C_i(t)\varphi(\cdot, \boldsymbol{Y}_i(t))$ *with well-prepared initial conditions.*

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*

$$\lim_{n\to\infty} n^{\bar{\xi}} \left(f_n(a_n) - f\right) = 0 \quad \textit{almost surely} \qquad \textit{if} \quad \lim_{t\to\infty} \int_{\mathbb{R}} \rho_0(\cdot, c)dc > 0 \quad \textit{a.e. in } D$$

Healing of errors by training:

- ξ = ½ initially — CLT scaling of *iid* initial conditions;
- ξ = 1 after optimization.

Error scaling — Central Limit Theorem (CLT)

$$d\boldsymbol{Y}_i = C_i \nabla F(\boldsymbol{Y}_i)dt - \frac{1}{n}\sum_{j=1}^{n} C_i C_j \nabla K(\boldsymbol{Y}_i, \boldsymbol{Y}_j)dt + (\beta n)^{-1}\nabla \log m(\boldsymbol{Y}_i, C_i)dt + \sqrt{2(\beta n)^{-1}}d\boldsymbol{W}_i,$$

$$dC_i = F(\boldsymbol{Y}_i)dt - \frac{1}{n}\sum_{j=1}^{n} C_j K(\boldsymbol{Y}_i, \boldsymbol{Y}_j)dt + (\beta n)^{-1}\partial_c \log m(\boldsymbol{Y}_i, C_i)dt + \sqrt{2(\beta n)^{-1}}dW_i'$$

low-temperature regime

$$\mathcal{E}_n[\rho] = \tfrac{1}{2}\int_\Omega \left(f(\boldsymbol{x}) - \tfrac{1}{2}\int_{D\times\mathbb{R}} c\varphi(\boldsymbol{x},\boldsymbol{y})\rho_0 d\boldsymbol{y}dc \right)^2 d\mu(\boldsymbol{x}) + (\beta n)^{-1}\int_{D\times\mathbb{R}} \rho \log(\rho/m) d\boldsymbol{y}dc$$

entropic correction

Prop 3 (CLT at finite temperature) *Let* $f_n(t) = \frac{1}{n}\sum_{i=1}^{n} C_i(t)\varphi(\cdot, \boldsymbol{Y}_i(t))$ *with well-prepared initial condition at* $t = T$. *Then*

$$\lim_{n\to\infty}\lim_{T\to-\infty} n\,(f_n(t) - f) = f_1(t) \qquad \textit{in law}$$

where $f_1(t)$ *is a Gaussian process such that: for any* $\chi \in L^2(\Omega, \mu)$,

$$\mathbb{E}\int_\Omega \chi(\boldsymbol{x}) f_1(t,\boldsymbol{x}) d\mu(\boldsymbol{x}) = \beta^{-1}\int_\Omega \chi(\boldsymbol{x})\epsilon^*(\boldsymbol{x}) d\mu(\boldsymbol{x})$$

$$\mathbb{E}\left(\int_\Omega \chi(\boldsymbol{x}) \left(f_1(t,\boldsymbol{x}) - \beta^{-1}\epsilon^*(\boldsymbol{x}) \right) d\mu(\boldsymbol{x}) \right)^2 = \beta^{-1}\int_\Omega |\chi(\boldsymbol{x})|^2 d\mu(\boldsymbol{x})$$

Discrete training set and stochastic gradient descent

- Loss function approximated by *empirical loss*

$$\ell_P(f, f_n) = \frac{1}{P}\sum_{p=1}^{P}|f(\boldsymbol{x}_p) - f_n(\boldsymbol{x}_p)|^2, \qquad \{\boldsymbol{x}_p\}_{p=1}^{P} = \textit{iid drawn from } \mu.$$

- Stochastic gradient descent (SGD):

$$\boldsymbol{Z}(t + \Delta t) = \boldsymbol{Z}(t) - \nabla_{\boldsymbol{z}} L_P(\boldsymbol{Z}(t))\Delta t$$

  where $\Delta t > 0$ is some time-step and $L_P(\boldsymbol{z}) = n\ell_P(f, f_n)$ is the empirical loss viewed as a function of the parameters $\boldsymbol{z} = (c_1, \boldsymbol{y}_1, \ldots, c_n, \boldsymbol{y}_n)$,

- If training set $\{\boldsymbol{x}_p\}_{p=1}^{P}$ is redrawn from μ independently at every time step t, leads to an *effective SDE*:

$$d\boldsymbol{Z} = -\nabla_z L(\boldsymbol{Z})dt + \sqrt{\theta}d\boldsymbol{B}$$

  where $L(\boldsymbol{z}) = n\ell(f, f_n(\boldsymbol{z}))$, and the quadaric variation of the noise is the covariance of the empirical loss

$$\langle d\boldsymbol{B}, d\boldsymbol{B}\rangle = \mathbb{E}\left(\nabla_z(L_{P=1}(\boldsymbol{z}) - n\ell(f, f_n(\boldsymbol{z})))\right)^{\otimes 2} dt$$

  and

$$\theta = \Delta t/P = \text{time step / batch size}$$

LLN and CLT for SGD

- Setting $\theta = \Delta t/P = an^{-2\alpha}$ with $a > 0$, $\alpha > 0$ (e.g. $P = O(n^{2\alpha})$), the empirical distribution satisfies

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

$$+ \tfrac{1}{2} a n^{-2\alpha} \nabla\nabla : \left(\rho_n c^2 A_2([f_n(t) - f], \boldsymbol{y}, \boldsymbol{y}) \right) + \tfrac{1}{2} a n^{-2\alpha} \partial_c^2 \left(\rho_n A_0([f_n(t) - f], \boldsymbol{y}, \boldsymbol{y}) \right)$$

$$+ a n^{-2\alpha} \partial_c \nabla \cdot \left(\rho_n c A_1([f_n(t) - f], \boldsymbol{y}, \boldsymbol{y}) \right)$$

$$+ \sqrt{a} n^{-\alpha} \dot{\eta}_n(t, \boldsymbol{y}, c)$$

additional terms higher order;
noise term dominates those.

Prop 7 (LLN & CLT for SGD) *Let* $f_n(t) = \frac{1}{n} \sum_{i=1}^{n} C_i(t) \varphi(\cdot, \boldsymbol{Y}_i(t))$ *with* $\{\boldsymbol{Y}_i(t), C_i(t)\}_{i=1}^n$ *solution to limiting SDE with* $\theta = an^{-2\alpha}$, $a > 0$ $\alpha \in (0, 1)$. *Then for any* $a_n > 0$ *such that* $a_n / \log n \to \infty$ *as* $n \to \infty$, *we have*

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

where $f_0(t) \to f$ *as* $t \to \infty$

- if $\alpha < 1$, training set error dominates
- if $\alpha = 1$, training set error and discretization error are in the same scale;
- if $\alpha > 1$, no gain (and convergence in time may be impeded)

3-spin model on the high-dimensional sphere

- Spherical 3-spin model: $f : S^{d-1}(\sqrt{d}) \to \mathbb{R}$ given by

$$f(\boldsymbol{x}) = \frac{1}{d} \sum_{p,q,r=1}^{d} a_{p,q,r} x_p x_q x_r, \qquad \boldsymbol{x} \in S^{d-1}\left(\sqrt{d}\right) \subset \mathbb{R}^d$$

  where the coefficients $\{a_{p,q,r}\}_{p,q,r=1}^{d}$ are independent Gaussian random variables with mean zero and variance one.

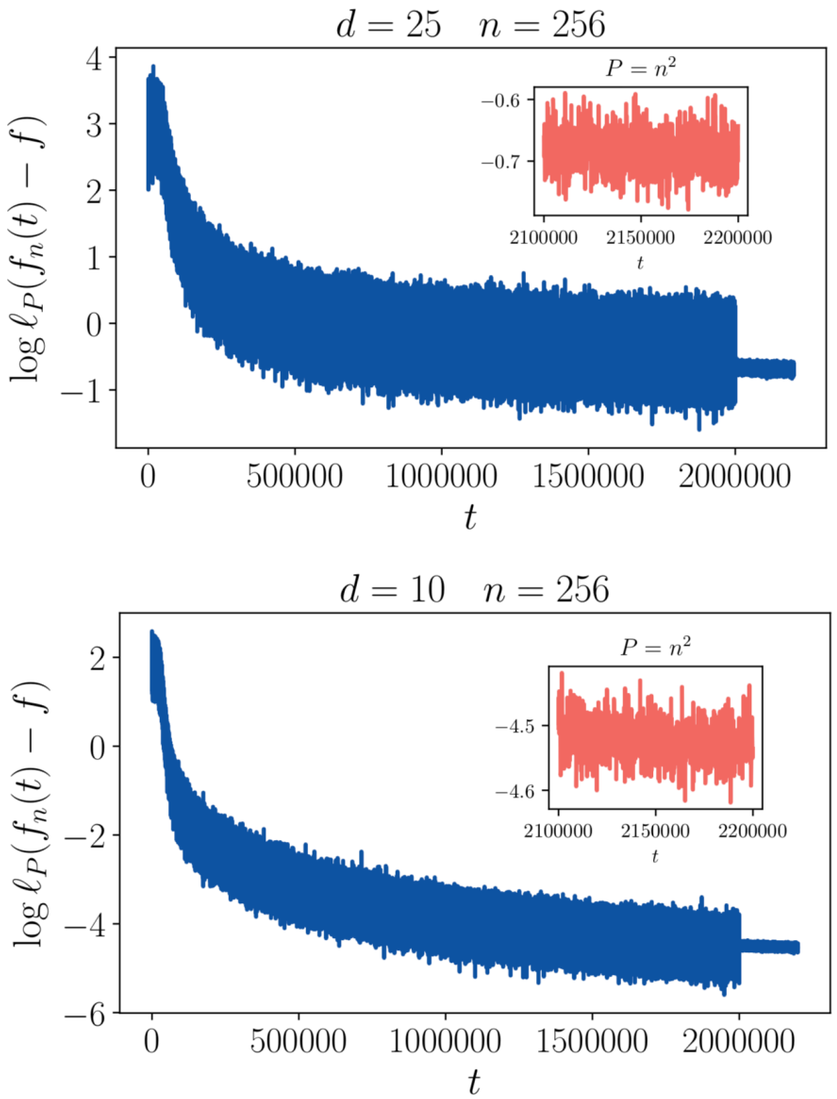
- Complex function with a number of critical points that grows exponentially with the dimensionality d.

- Previous works drew a parallel between the glassy 3-spin function and generic loss functions.

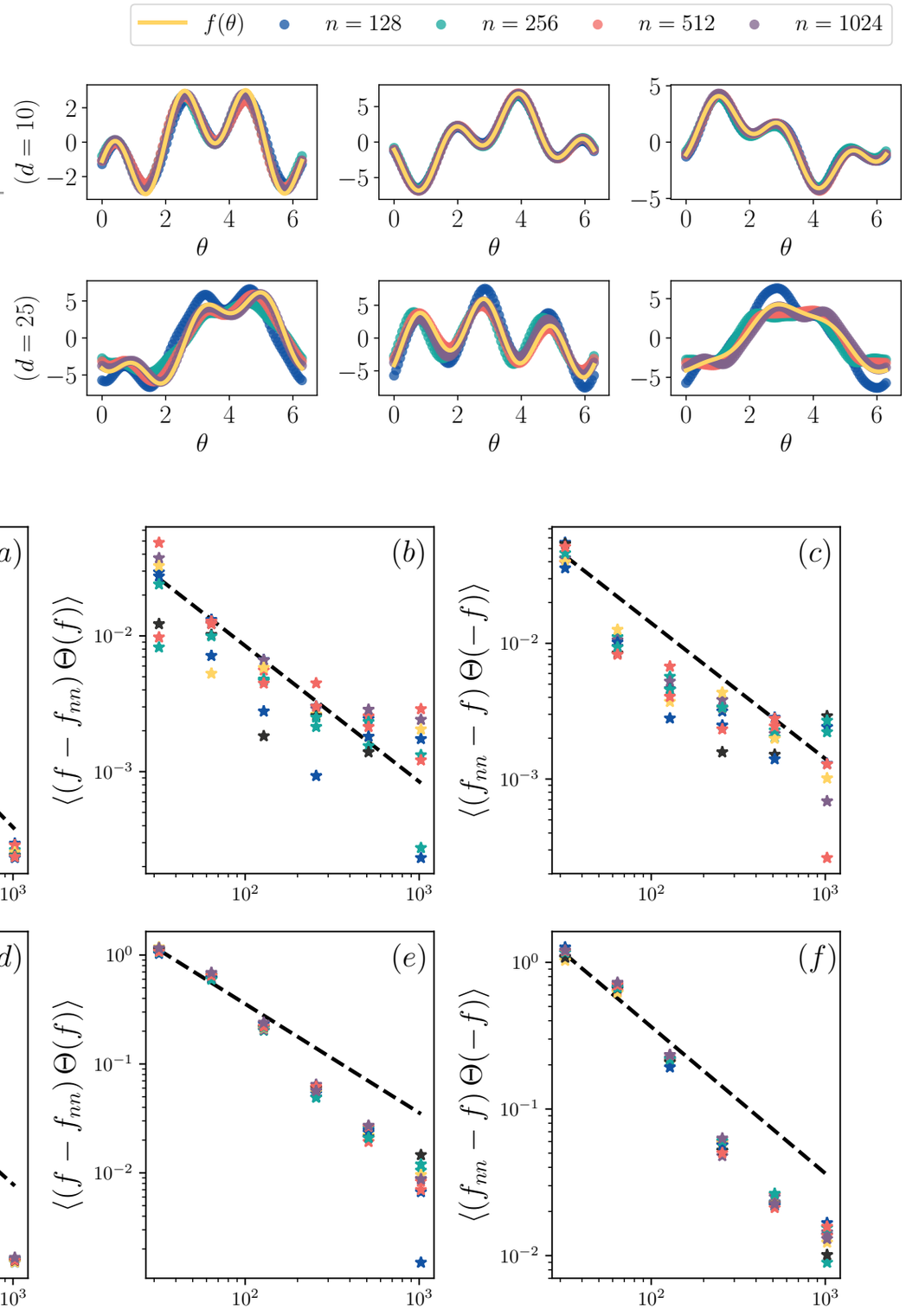
- In contrast, we use the 3-spin function as a difficult target for approximation by neural networks, that is:
  - ▷ we train networks to learn f with a particular realization of $a_{p,q,r}$, and;
  - ▷ we study the accuracy of that representation as a function of the number of particles n.

Confirming the scaling

$$f_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} c_i h(\boldsymbol{a}_i \cdot \boldsymbol{x} + b_i), \qquad h(z) = 1/(1+e^{-z}).$$



Empirical loss vs SGD training time.
At time 2E6, the batch size is increased to initiate an additional quench.



Error scaling for single layer neural network with sigmoid nonlinearities. Dashed line are $\propto n^{-1}$

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.

- Viewing the parameters as interacting particles, we can demonstrate that the loss landscape is asymptotically convex and stochastic gradient descent converges to an energy minimizer — dynamical generalization of Universal Representation Theorems.

- The approximation error can be identified up to a constant, which shows that its scaling is universal.

- This offers exciting possibilities in scientific computing (free energy methods, quantum variational energy calculations, PDE solving, etc.) that are only beginning to be explored.

"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"

"About six inches to the mile."

"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"

"Have you used it much?" I enquired.

"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.

Lewis Carroll, Sylvie and Bruno, 1889 - 1893.

Functional formulation in the limit of large *n*

- Assume that the set $\{(c_i, \boldsymbol{y}_i)\}_{i=1}^n$ is such that for some signed density G on D we have

$$f_n = \frac{1}{n}\sum_{i=1}^{n} c_i \varphi(\cdot, \boldsymbol{y}_i) \to \tilde{f} = \int_D \varphi(\cdot, \boldsymbol{y}) G(\boldsymbol{y}) d\boldsymbol{y} \qquad \text{as} \quad n \to \infty$$

- Make the loss function quadratic in G: $\ell(f, \tilde{f}) = C_f - \int_D F(\boldsymbol{y}) G(\boldsymbol{y}) d\boldsymbol{y} + \frac{1}{2}\int_{D\times D} K(\boldsymbol{y}, \boldsymbol{z}) G(\boldsymbol{y}) G(\boldsymbol{z}) d\boldsymbol{y} d\boldsymbol{z}$

Universal Representation Theorem (Barron, Cybenko, Park,...) *If the kernel φ is discriminating, then given any $f \in L^2(\Omega, \mu)$ and $\epsilon > 0$:*

$$\exists f^* \quad \textit{such that} \quad \|f - f^*\|_{L^2(\Omega,\mu)} \leq \epsilon$$

and f^* can be represented as

$$\int_D \varphi(\cdot, \boldsymbol{y}) G^*(\boldsymbol{y}) d\boldsymbol{y} = f^* \quad \textit{a.e. in } \Omega$$

where G^* solves

$$F^*(\boldsymbol{y}) = \int_D K(\boldsymbol{y}, \boldsymbol{z}) G^*(\boldsymbol{z}) d\boldsymbol{z} \qquad \textit{with} \qquad F^*(\boldsymbol{y}) = \int_\Omega f^*(\boldsymbol{x}) \varphi(\boldsymbol{x}, \boldsymbol{y}) d\mu(\boldsymbol{x})$$

The function f^* can also be realized as $f^* = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} c_j \varphi(\cdot, \boldsymbol{y}_j)$ for some choice of $\{\boldsymbol{y}_i, c_i\}_{i\in\mathbb{N}}$.

Stationary points of McKean-Vlasov equation

- The stationary points of the McKean-Vlasov equation satify

$$-F(\boldsymbol{y}) + \int_{D_0} K(\boldsymbol{y}, \boldsymbol{y}') d\gamma_0(\boldsymbol{y}') = 0 \qquad \text{for } \nu_0\text{-almost all } \boldsymbol{y} \in D_0 = \operatorname{supp} \nu_0$$

where

$$\lim_{t\to\infty} \int_{D\times\mathbb{R}} \chi(\boldsymbol{y}) \rho_0(t, \cdot, c) d\boldsymbol{y} dc = \int_D \chi(\boldsymbol{y}) d\nu_0(\boldsymbol{y}) \qquad \lim_{t\to\infty} \int_{D\times\mathbb{R}} \chi(\boldsymbol{y}) c\rho_0(t, \cdot, c) d\boldsymbol{y} dc = \int_D \chi(\boldsymbol{y}) d\gamma_0(\boldsymbol{y})$$

- D_0 may be a (singular) subset of D; however, if $\varphi(\cdot, \boldsymbol{y})$ is discriminating in D_0, i.e. if

$$\int_\Omega g(\boldsymbol{x}) \varphi(\boldsymbol{x}, \cdot) d\mu(\boldsymbol{x}) = 0 \quad \text{a.e. in } D_0 \quad \Rightarrow \quad g = 0 \quad \text{a.e. in } \Omega \tag{2}$$

then

Universal Representation Thm

$$\int_{D_0} \varphi(\cdot, \boldsymbol{y}) d\gamma_0(\boldsymbol{y}) = f \qquad \text{a.e. in } \Omega. \tag{3}$$

- Even if $\varphi(\cdot, \boldsymbol{y})$ is not discriminating in D_0, we can continously deform D_0 into a set such that (2) and hence (3) hold. As a result these equations always hold in the presence of noise (as in SGD).