NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES
NYU CENTER FOR DATA SCIENCE
FLATIRON INSTITUTE

JOAN BRUNA

MATHEMATICAL ASPECTS OF NEURAL NETWORK APPROXIMATION AND LEARNING

joint work with


Zhengdao Chen


Aaron Zweig


Samy Jelassi


Grant Rotskoff

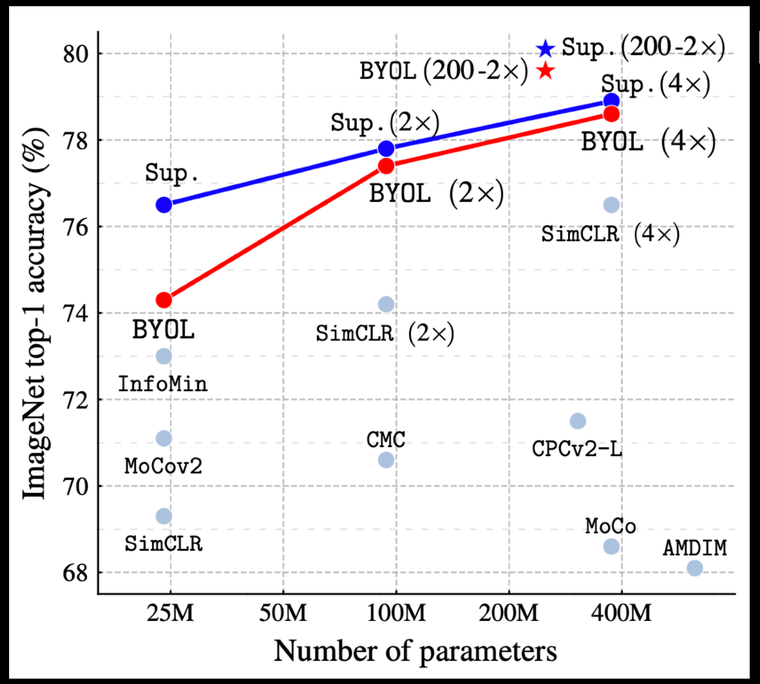

E.Vanden-Eijnden


Luca Venturi

DEEP LEARNING TODAY: EXPERIMENTAL REVOLUTION

[v.d. Oord et al.'19]

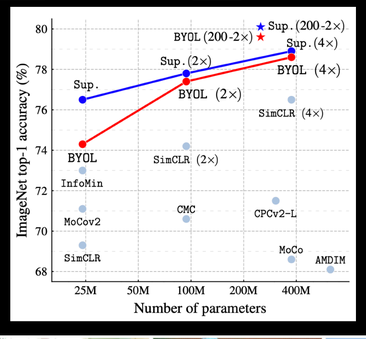
Gatys et al'14



[Grill et al'20]

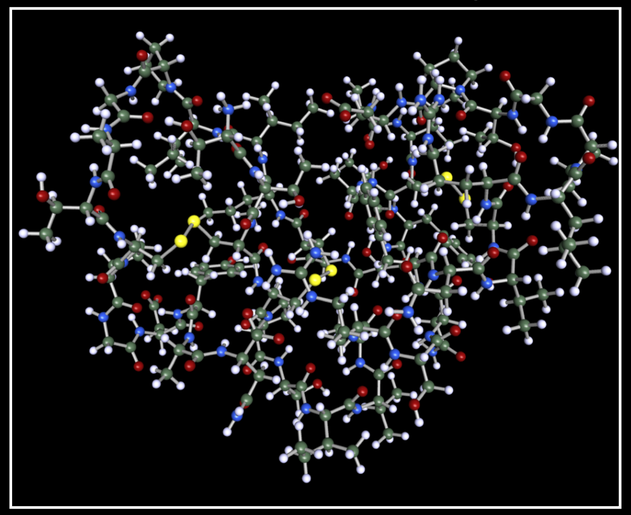
[He et al.'17]

DEEP LEARNING TODAY: EXPERIMENTAL REVOLUTION



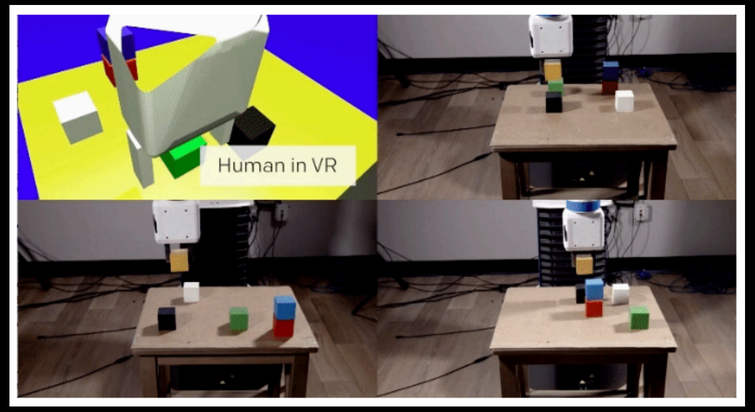
Computational Biology

Quantum Chemistry

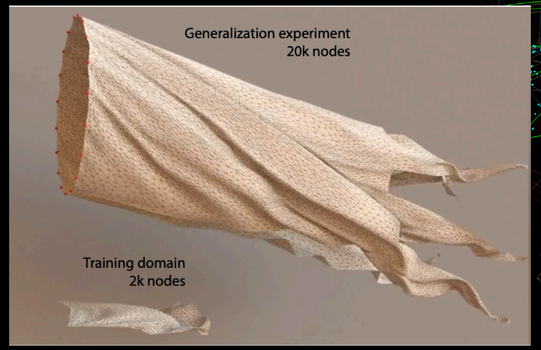


Games

Robotics

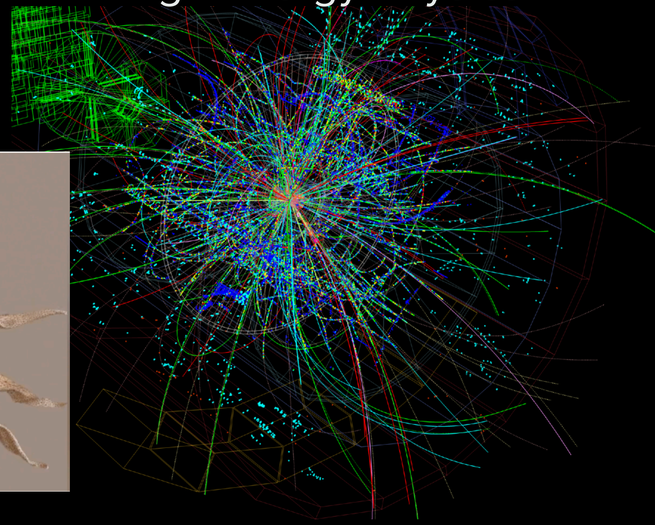


Scientific Computing

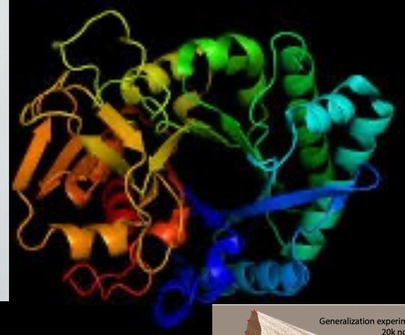


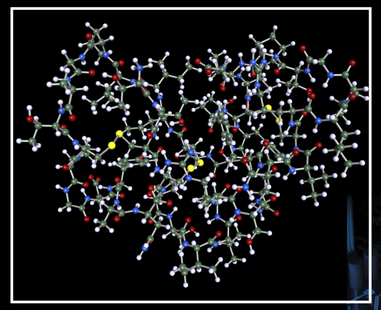
[Pfaff et al]

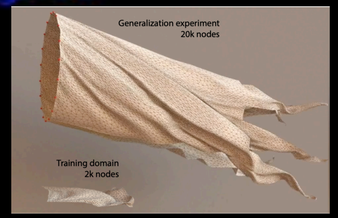
High Energy Physics

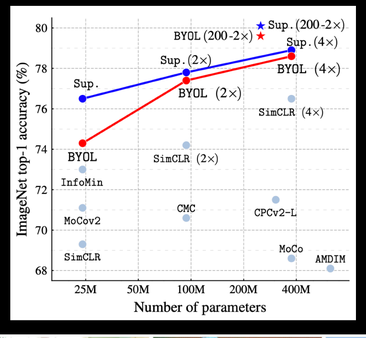


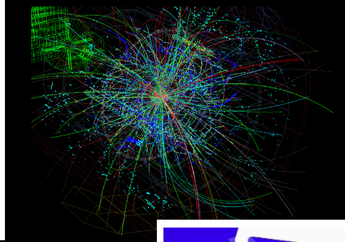
DEEP LEARNING TODAY: EXPERIMENTAL REVOLUTION

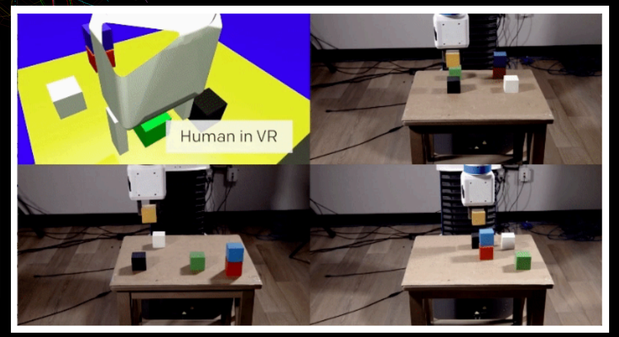










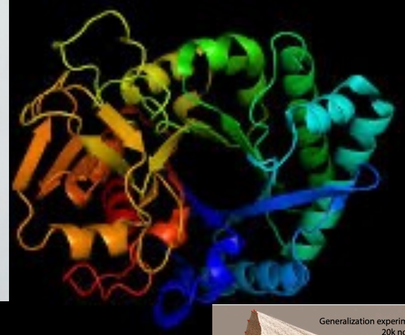


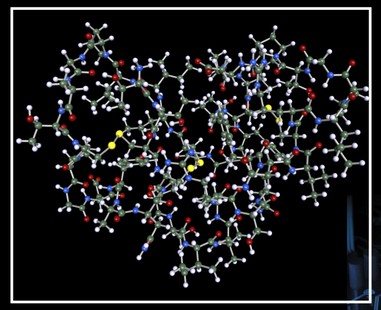
- Phenomenal ability to extract information from ***high-dimensional*** observations $x \in \Omega$
- In essence: ***non-linear***, compositional ***representation learning.***

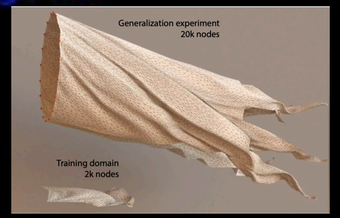
$$f(x) = \theta^\top \Phi(x) \qquad \Phi : \Omega \to \mathbb{R}^K$$

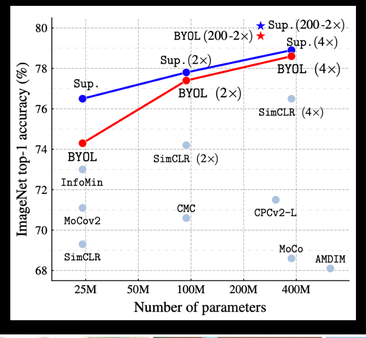
  - Φ and θ are both ***learnt*** from data.
- Learning algorithm = gradient descent.
- Structure in Φ based on physics (e.g symmetries, multiscale)

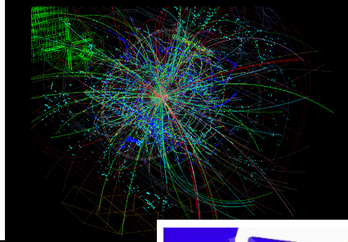
DEEP LEARNING TODAY: EXPERIMENTAL REVOLUTION

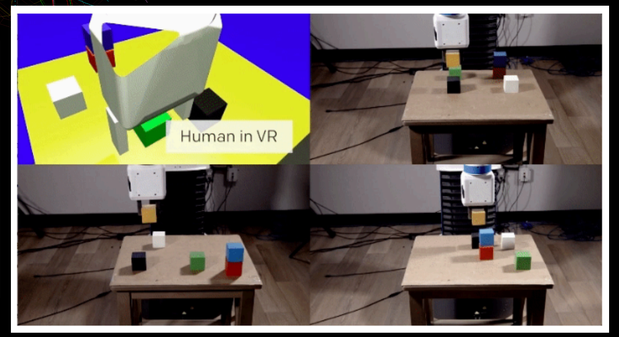












- Phenomenal ability to extract information from ***high-dimensional*** observations $x \in \Omega$
- In essence: ***non-linear***, compositional ***representation learning.***

$$f(x) = \theta^\top \Phi(x) \qquad \Phi : \Omega \to \mathbb{R}^K$$

  - Φ and θ are both ***learnt*** from data.
- Learning algorithm = gradient descent.
- Structure in Φ based on physics (e.g symmetries, multiscale)

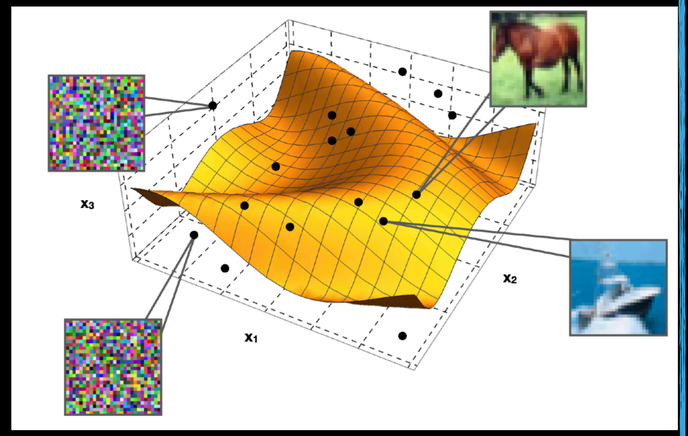
How, when, and why can DL systems approximate high-dimensional functions from data?

SUPERVISED LEARNING BASIC SETUP

DATA

$$x_i \sim \nu : \text{ data distribution in } \Omega.$$
$$y_i = f^*(x_i) \text{ for some } f^* \in L^2(\mathbb{R}^m, \mathrm{d}\nu).$$

$$(x_i, y_i) \in \Omega \times \mathbb{R}$$



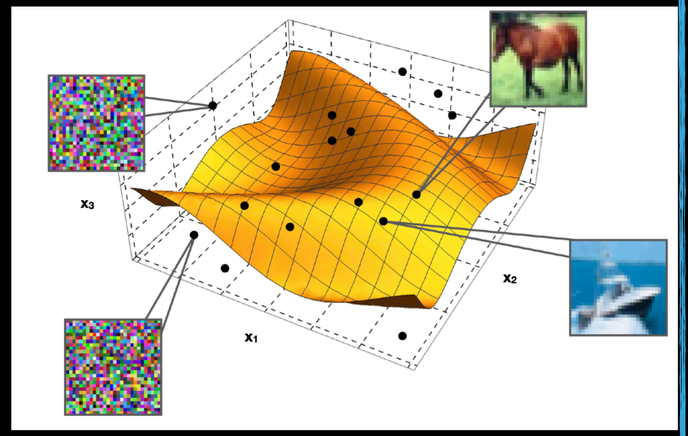
[Goldt, Zdeborova, Krzakala et al]

SUPERVISED LEARNING BASIC SETUP

DATA

$x_i \sim \nu$: data distribution in Ω.
$y_i = f^*(x_i)$ for some $f^* \in L^2(\mathbb{R}^m, \mathrm{d}\nu)$.

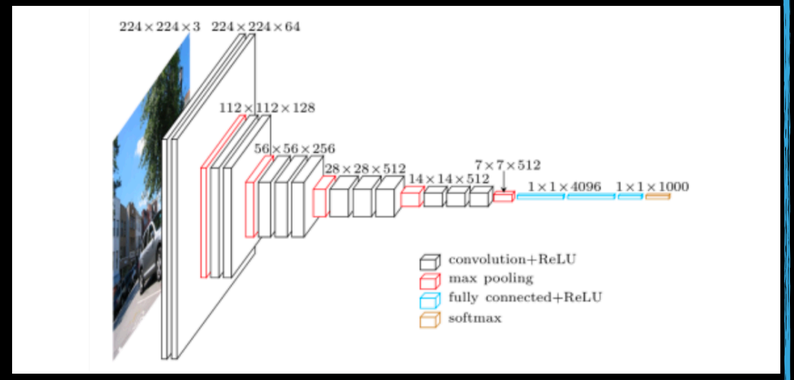
$(x_i, y_i) \in \Omega \times \mathbb{R}$



[Goldt, Zdeborova, Krzakala et al]

MODEL

$\mathcal{F} \subset \{f : \Omega \to \mathbb{R}\}$ e.g. $f(x; \Theta)$, $\Theta \in \mathcal{D}$.
Normed space: $\gamma(f)$, $f \in \mathcal{F}$.



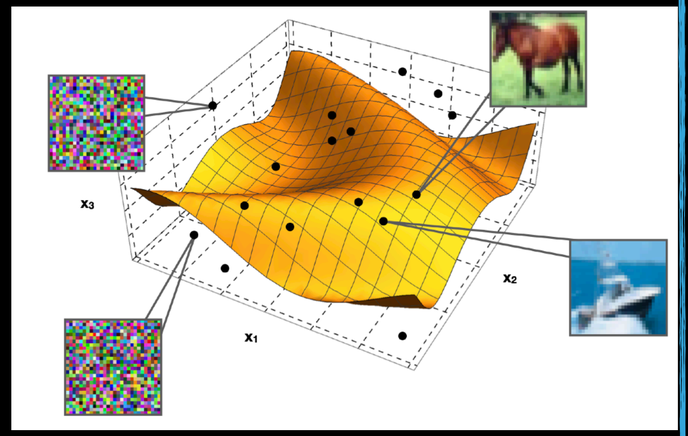
Vgg

SUPERVISED LEARNING BASIC SETUP

DATA

$x_i \sim \nu$: data distribution in Ω.
$y_i = f^*(x_i)$ for some $f^* \in L^2(\mathbb{R}^m, \mathrm{d}\nu)$.

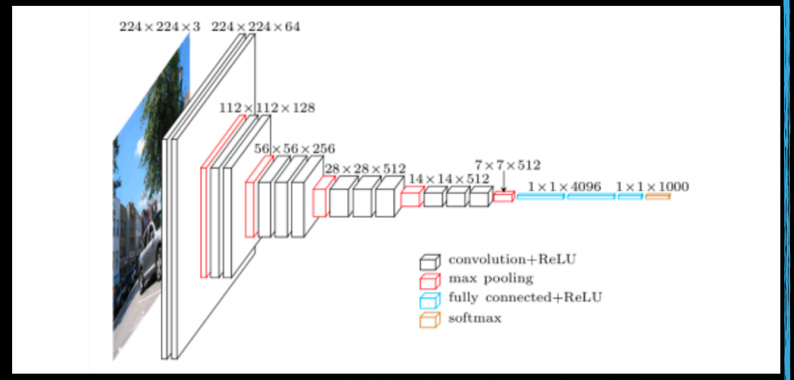
$(x_i, y_i) \in \Omega \times \mathbb{R}$



[Goldt, Zdeborova, Krzakala et al]

MODEL

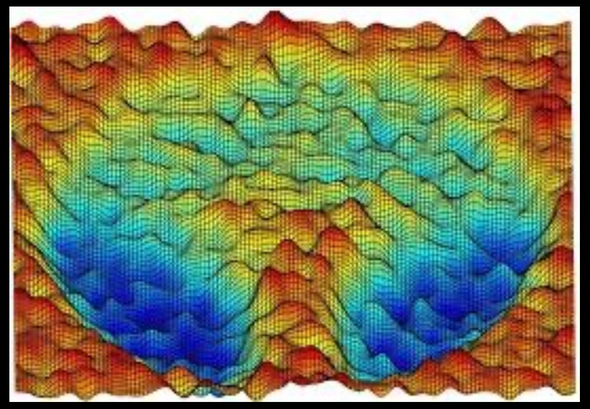
$\mathcal{F} \subset \{f : \Omega \to \mathbb{R}\}$ e.g. $f(x; \Theta)$, $\Theta \in \mathcal{D}$.
Normed space: $\gamma(f)$, $f \in \mathcal{F}$.



Vgg

ERROR METRIC

$\mathcal{R}(f)$ convex, e.g.
$\mathcal{R}(f) = \mathbb{E}_\nu |f(x) - f^*(x)|^2$



[fig credit E. Vanden-Eijnden]

Empirical Loss:

$$\widehat{\mathcal{R}}(f) = \frac{1}{L} \sum_{i=1}^{L} |f(x_i) - f^*(x_i)|^2.$$

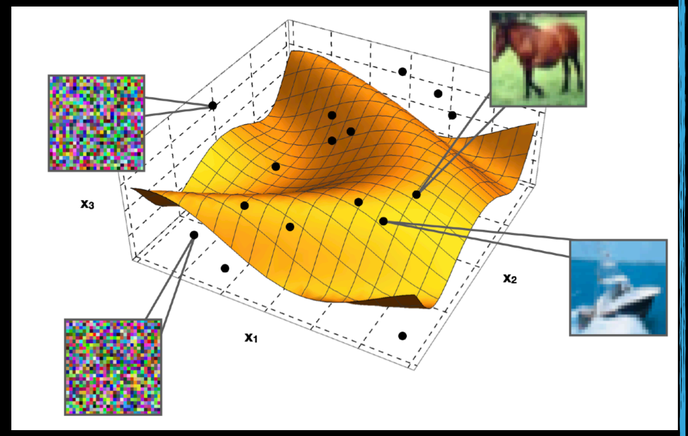
SUPERVISED LEARNING BASIC SETUP

DATA

$x_i \sim \nu$: data distribution in Ω.

$y_i = f^*(x_i)$ for some $f^* \in L^2(\mathbb{R}^m, \mathrm{d}\nu)$.

$(x_i, y_i) \in \Omega \times \mathbb{R}$

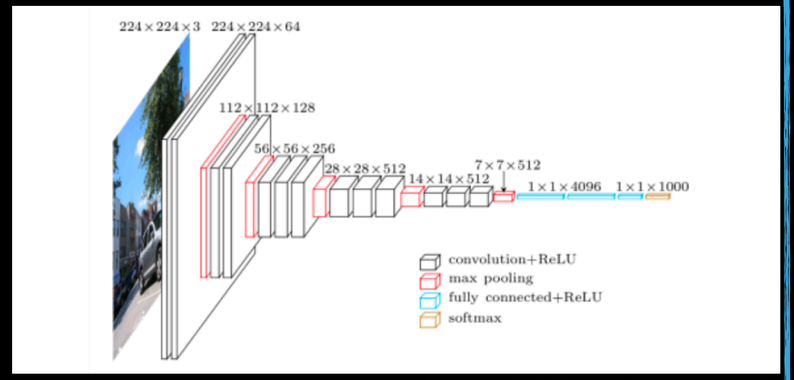


[Goldt, Zdeborova, Krzakala et al]

MODEL

$\mathcal{F} \subset \{f : \Omega \to \mathbb{R}\}$ e.g. $f(x; \Theta)$, $\Theta \in \mathcal{D}$.

Normed space: $\gamma(f)$, $f \in \mathcal{F}$.

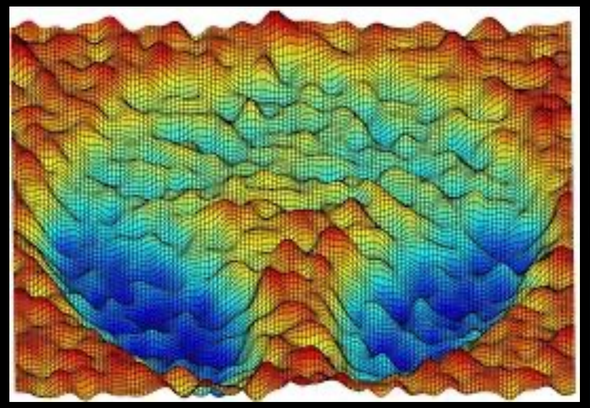


Vgg

ERROR METRIC

$\mathcal{R}(f)$ convex, e.g.

$\mathcal{R}(f) = \mathbb{E}_\nu |f(x) - f^*(x)|^2$



[fig credit E. Vanden-Eijnden]

Empirical Loss:

$$\widehat{\mathcal{R}}(f) = \frac{1}{L} \sum_{i=1}^{L} |f(x_i) - f^*(x_i)|^2.$$

ALGORITHM

Empirical Risk Minimization

Ball $\mathcal{F}_\delta = \{f \in \mathcal{F}; \gamma(f) \leq \delta\}$.

Find $\hat{f}$ such that $\widehat{R}(\hat{f}) \leq \min_{f \in \mathcal{F}_\delta} \widehat{R}(f) + \epsilon$.

(constrained)

$\hat{f}$ such that $\widehat{\mathcal{R}} \leq \min_f \left(\widehat{\mathcal{R}}(f) + \lambda \gamma(f) \right) + \epsilon$.

(penalized)

CHALLENGES OF HIGH-DIMENSIONAL LEARNING

- Basic decomposition of error:

$$\text{Find } \hat{f} \text{ such that } \widehat{R}(\hat{f}) \leq \min_{f \in \mathcal{F}_\delta} \widehat{R}(f) + \epsilon.$$

$$\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) \leq \underbrace{\inf_{f \in \mathcal{F}_\delta} \mathcal{R}(f) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)}_{\text{approx error}} + \underbrace{2 \sup_{\mathcal{F}_\delta} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)|}_{\text{statistical error}} + \underbrace{\epsilon}_{\text{optim. error}}.$$

[Bottou & Bousquet]

CHALLENGES OF HIGH-DIMENSIONAL LEARNING

- Basic decomposition of error:

$$\text{Find } \hat{f} \text{ such that } \widehat{R}(\hat{f}) \leq \min_{f \in \mathcal{F}_\delta} \widehat{R}(f) + \epsilon.$$

$$\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) \leq \underbrace{\inf_{f \in \mathcal{F}_\delta} \mathcal{R}(f) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)}_{\text{approx error}} + \underbrace{2 \sup_{\mathcal{F}_\delta} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)|}_{\text{statistical error}} + \underbrace{\epsilon}_{\text{optim. error}}.$$

[Bottou & Bousquet]

APPROXIMATION

Functional Approximation that is not cursed by input dimensionality, capturing right notion of regularity

STATISTICAL

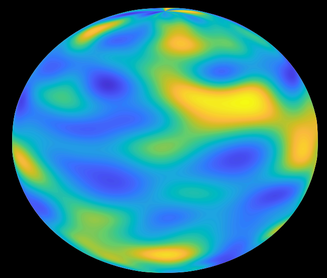
Concentration Bounds: Requires spaces $\mathcal{F}_\delta$ not to grow too quickly with dimension

COMPUTATIONAL

How to solve the ERM efficiently in the high-dimensional regime for the chosen hypothesis space?

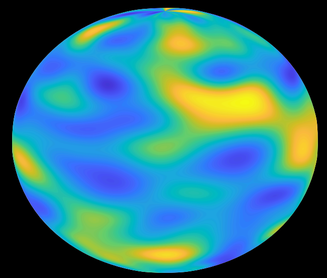
THE CURSE OF DIMENSIONALITY

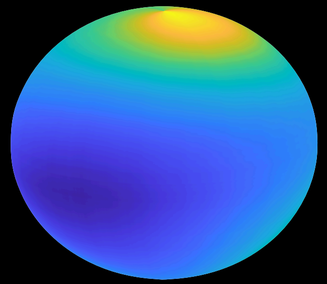
- "Classic" functional spaces do not play well with this tradeoff:
  - $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \text{ is } \mathrm{Lipschitz}\}$ is too big: the number of samples required to identify $f^* \in \mathcal{F}$ up to error ϵ is $\Omega(\epsilon^{-d})$ [von Luxburg & Bousquet].



THE CURSE OF DIMENSIONALITY

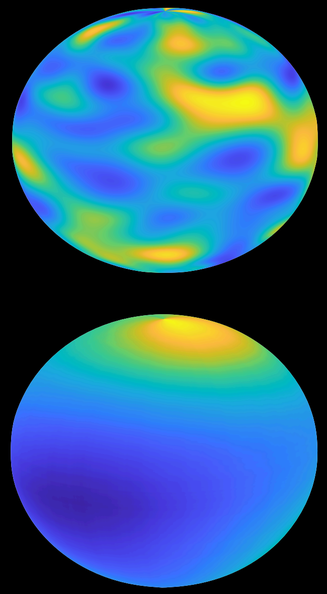
- "Classic" functional spaces do not play well with this tradeoff.
  - $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \text{ is } \mathrm{Lipschitz}\}$ is too big: the number of samples required to identify $f^* \in \mathcal{F}$ up to error ϵ is $\Omega(\epsilon^{-d})$ [von Luxburg & Bousquet].
  - $\mathcal{F} = \mathcal{H}^{s,p}$: $\mathrm{Sobolev\ spaces}$. Minimax rate of approximation is cursed unless $s \geq d/2$ [Tsybakov]: only very smooth functions are allowed!





THE CURSE OF DIMENSIONALITY

- "Classic" functional spaces do not play well with this tradeoff.
  - $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \text{ is } \mathrm{Lipschitz}\}$ is too big: the number of samples required to identify $f^* \in \mathcal{F}$ up to error ϵ is $\Omega(\epsilon^{-d})$ [von Luxburg & Bousquet].
  - $\mathcal{F} = \mathcal{H}^{s,p}$: $\mathrm{Sobolev\ spaces}$. Minimax rate of approximation is cursed unless $s \geq d/2$ [Tsybakov]: only very smooth functions are allowed!

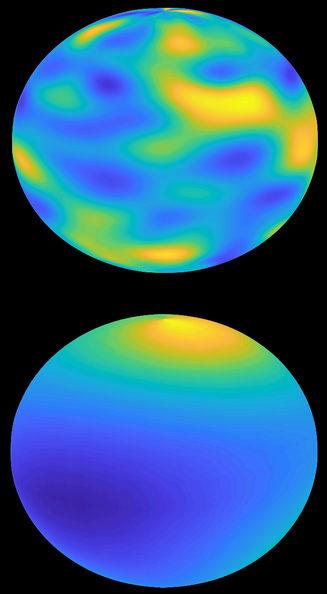


Which functions can be learnt efficiently in the high-dimensional regime?

...with ***neural networks*** and using ***gradient descent***?

THE CURSE OF DIMENSIONALITY

- "Classic" functional spaces do not play well with this tradeoff.
  - $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \text{ is } \mathrm{Lipschitz}\}$ is too big: the number of samples required to identify $f^* \in \mathcal{F}$ up to error ϵ is $\Omega(\epsilon^{-d})$ [von Luxburg & Bousquet].
  - $\mathcal{F} = \mathcal{H}^{s,p}$: $\mathrm{Sobolev\ spaces}$. Minimax rate of approximation is cursed unless $s \geq d/2$ [Tsybakov]: only very smooth functions are allowed!



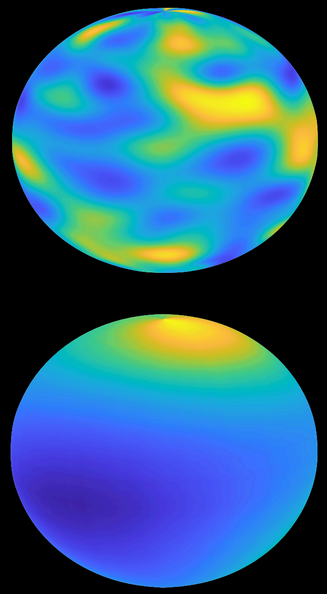
Which functions can be learnt efficiently in the high-dimensional regime?

...with ***neural networks*** and using ***gradient descent***?

...with ***deep*** neural networks?

THE CURSE OF DIMENSIONALITY

- "Classic" functional spaces do not play well with this tradeoff.
  - $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \text{ is } \mathrm{Lipschitz}\}$ is too big: the number of samples required to identify $f^* \in \mathcal{F}$ up to error ϵ is $\Omega(\epsilon^{-d})$ [von Luxburg & Bousquet].
  - $\mathcal{F} = \mathcal{H}^{s,p}$: $\mathrm{Sobolev\ spaces}$. Minimax rate of approximation is cursed unless $s \geq d/2$ [Tsybakov]: only very smooth functions are allowed!



Which functions can be learnt efficiently in the high-dimensional regime?

...with ***neural networks*** and using ***gradient descent***?

...with ***deep*** neural networks?

...with ***deep structured*** neural networks?

THIS TALK

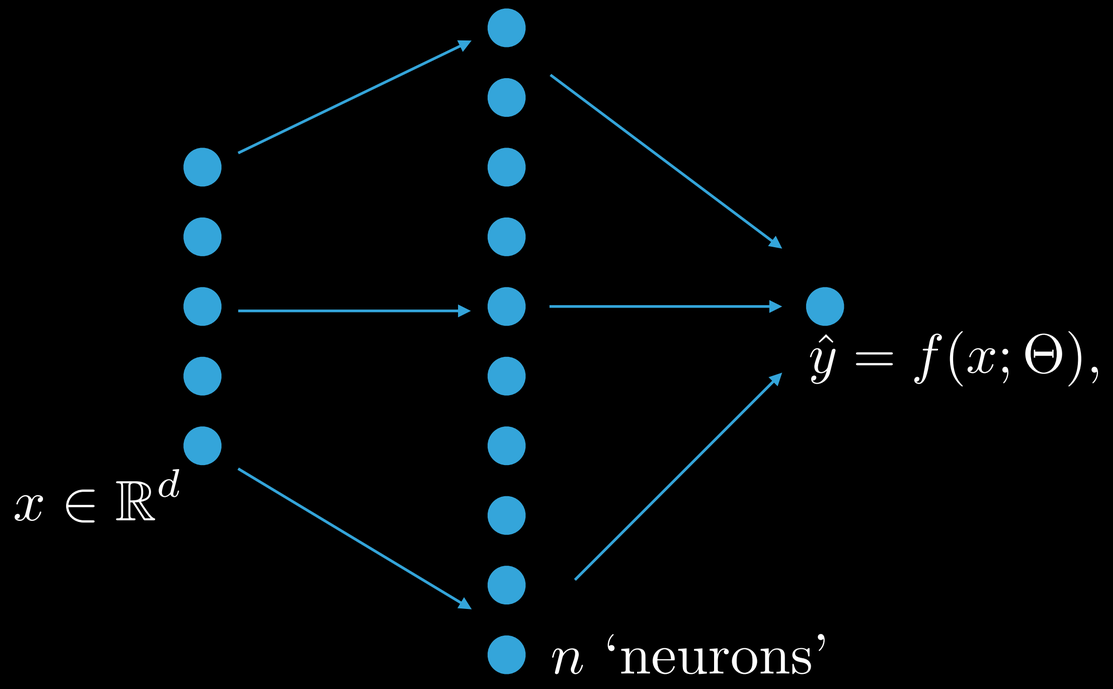
- Simplest instance of nonlinear feature learning: shallow NNs.
  - Gradient-descent Optimization analyzed as measure dynamics. Retains non-linear essence with Mean-field global convergence guarantees.
  - Towards Finite-width guarantees by CLT and fine-grained analysis of ReLU activations.

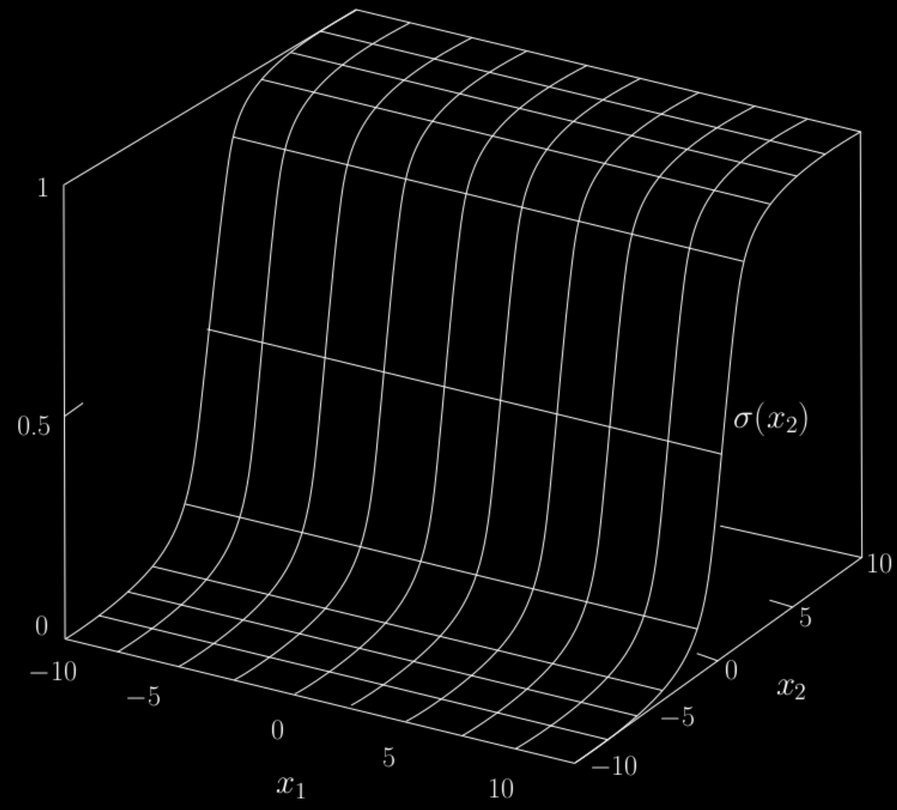
- Beyond Shallow Learning
  - Depth-Separation for ReLU networks
  - Depth-Separation and Learning for Symmetric Functions
  - [Mean-Field Dynamics on zero-sum two-player games].

SINGLE HIDDEN-LAYER NEURAL NETWORK

- $f(x;\Theta) = \sum_{j \le n} \tilde{\varphi}(x;\theta_j)$ is a sum of ridge functions:

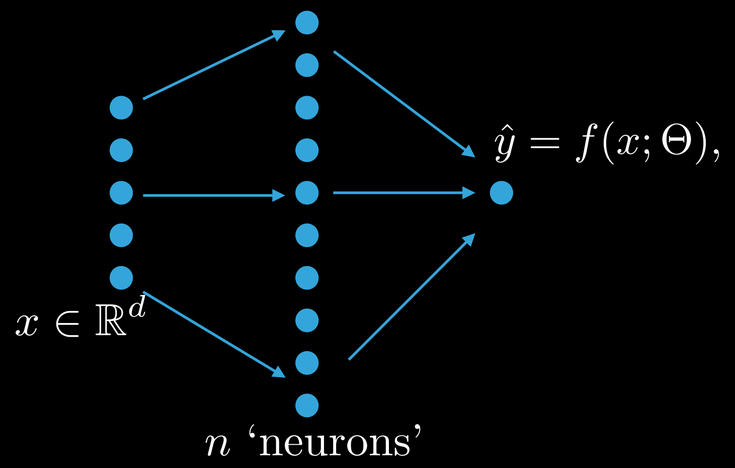
$$\tilde{\varphi}(x;\theta) = a\varphi(x;z),$$
$$\varphi(x;z) = \sigma(\langle x, w\rangle + b),$$
$$\theta = \{a, z\} \in \mathbb{R} \times \mathcal{D}.$$





- Three basic scaling quantities:
  - L datapoints, d input dimensions, n neurons.

SINGLE HIDDEN-LAYER NEURAL NETWORK



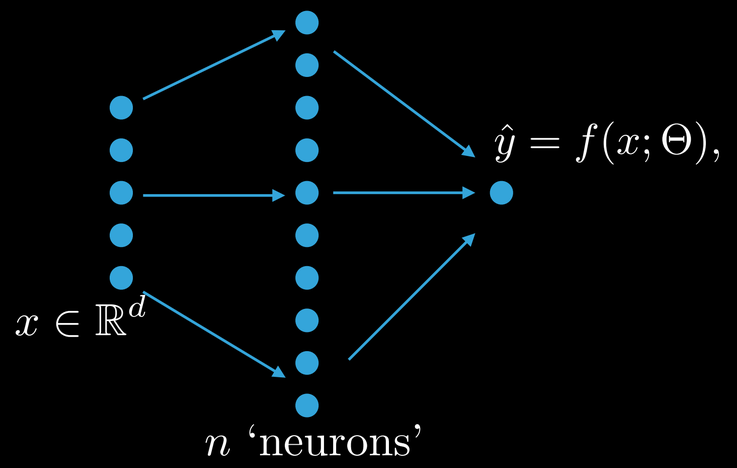
$$f(x;\Theta) = \sum_{j \le n} \tilde{\varphi}(x;\theta_j)$$

$$\tilde{\varphi}(x;\theta) = a\varphi(x;z),$$
$$\varphi(x;z) = \sigma(\langle x, w\rangle + b),$$
$$\theta = \{a, z\} \in \mathbb{R} \times \mathcal{D}.$$

- As $n \to \infty$, for appropriate base measure $\gamma \in \mathcal{M}(\mathcal{D})$, we have the integral representation

$$f(x) = \int_{\mathcal{D}} \varphi(x,z) g(z) \gamma(dz).$$

SINGLE HIDDEN-LAYER NEURAL NETWORK



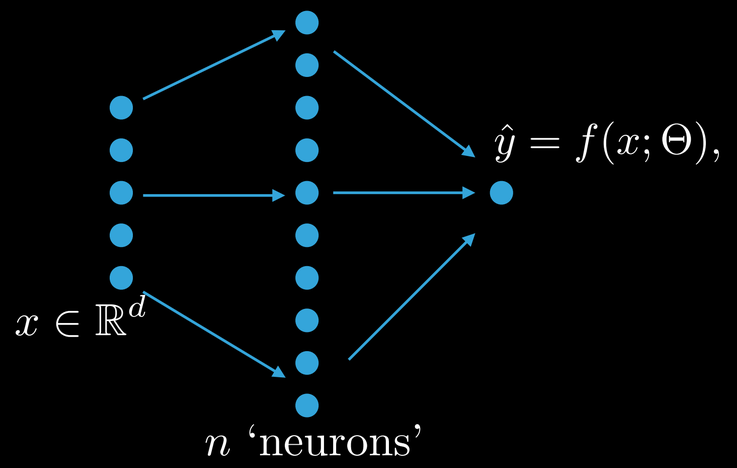
$$f(x;\Theta) = \sum_{j \le n} \tilde{\varphi}(x;\theta_j)$$

$$\tilde{\varphi}(x;\theta) = a\varphi(x;z),$$
$$\varphi(x;z) = \sigma(\langle x, w\rangle + b),$$
$$\theta = \{a, z\} \in \mathbb{R} \times \mathcal{D}.$$

- As $n \to \infty$, for appropriate base measure $\gamma \in \mathcal{M}(\mathcal{D})$, we have the integral representation
$$f(x) = \int_{\mathcal{D}} \varphi(x,z) g(z) \gamma(dz).$$

- ***Universal Approximation Theorems:*** shallow representations are dense in $\mathcal{C}(\mathbb{R}^d)$ under uniform compact convergence iff σ is not a polynomial [Barron, Bartlett, Petrushev, Lehno, Cybenko, Hornik, Pinkus].

SINGLE HIDDEN-LAYER NEURAL NETWORK



$$f(x;\Theta) = \sum_{j \le n} \tilde{\varphi}(x;\theta_j)$$

$$\tilde{\varphi}(x;\theta) = a\varphi(x;z),$$
$$\varphi(x;z) = \sigma(\langle x, w\rangle + b),$$
$$\theta = \{a, z\} \in \mathbb{R} \times \mathcal{D}.$$

- As $n \to \infty$, for appropriate base measure $\gamma \in \mathcal{M}(\mathcal{D})$, we have the integral representation
$$f(x) = \int_{\mathcal{D}} \varphi(x,z) g(z) \gamma(dz).$$

- ***Universal Approximation Theorems:*** shallow representations are dense in $\mathcal{C}(\mathbb{R}^d)$ under uniform compact convergence iff σ is not a polynomial [Barron, Bartlett, Petrushev, Lehno, Cybenko, Hornik, Pinkus].

- What are the associated functional spaces?

REPRODUCING KERNEL HILBERT SPACES

- Consider first γ_0 to be a fixed probability measure on $\mathcal{D}$.

$$\mathcal{F}_2 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,;\, f(x) = \int_{\mathcal{D}} \varphi(x,z) g(z) \mu_0(dz) \text{ and } g \in L^2(\mathcal{D}, d\mu_0) \right\}.$$

- $\mathcal{F}_2$ is a Reproducing Kernel Hilbert Space, with kernel given by

$$k(x,x') = \int \varphi(x,z)\varphi(x',z)\mu_0(dz)$$ [Bach'17a]

REPRODUCING KERNEL HILBERT SPACES

- Consider first γ_0 to be a fixed probability measure on $\mathcal{D}$.

$$\mathcal{F}_2 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,;\, f(x) = \int_{\mathcal{D}} \varphi(x,z) g(z) \mu_0(dz) \text{ and } g \in L^2(\mathcal{D}, d\mu_0) \right\}.$$

- $\mathcal{F}_2$ is a Reproducing Kernel Hilbert Space, with kernel given by

$$k(x,x') = \int \varphi(x,z)\varphi(x',z)\mu_0(dz) \qquad \text{[Bach'17a]}$$

- Learning in these RKHS is well-understood (kernel ridge regression), with efficient optimization algorithms.
  - Random feature expansions [Rahimi/Recht'08, Bach'17b] .

REPRODUCING KERNEL HILBERT SPACES

- Consider first γ_0 to be a fixed probability measure on $\mathcal{D}$.

$$\mathcal{F}_2 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,;\, f(x) = \int_{\mathcal{D}} \varphi(x,z) g(z) \mu_0(dz) \text{ and } g \in L^2(\mathcal{D}, d\mu_0) \right\}.$$

- $\mathcal{F}_2$ is a Reproducing Kernel Hilbert Space, with kernel given by

$$k(x,x') = \int \varphi(x,z)\varphi(x',z)\mu_0(dz) \qquad \text{[Bach'17a]}$$

- Learning in these RKHS is well-understood (kernel ridge regression), with efficient optimization algorithms.
  - Random feature expansions [Rahimi/Recht'08, Bach'17b] .

- However, they are cursed by dimensionality: only contain very smooth functions (derivatives of order $O(d)$ must exist).
  - Kernels arising from linearizing NNs recently studied [***NTK***, Jacot et al, Arora et al., Mei et al. Tibshirani, Belkin, Bietti & Mairal, Bietti & Bach].

VARIATION-NORM SPACES

[Bengio et al'06, Rosset et al.'07, Bach'17]

- Alternatively, we can consider

$$\mathcal{F}_1 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,;\, f(x) = \int_{\mathcal{D}} \varphi(x,z)\mu(dz); \|\mu\|_{TV} < \infty. \right\}.$$

- $\mathcal{F}_1$ is a Banach space, with norm $\|f\|_{\mathcal{F}_1} := \inf \left\{ \|\mu\|_{TV}; f = \int \varphi d\mu \right\}.$
  - Also known as ***Barron*** Spaces [Barron'90s, E et al '19].

VARIATION-NORM SPACES

[Bengio et al'06, Rosset et al.'07, Bach'17]

- Alternatively, we can consider

$$\mathcal{F}_1 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,;\, f(x) = \int_{\mathcal{D}} \varphi(x, z)\mu(dz); \|\mu\|_{TV} < \infty. \right\}.$$

- $\mathcal{F}_1$ is a Banach space, with norm $\|f\|_{\mathcal{F}_1} := \inf \left\{ \|\mu\|_{TV}; f = \int \varphi d\mu \right\}$.
  - Also known as ***Barron*** Spaces [Barron'90s, E et al '19].

- $\mathcal{F}_2 \subset \mathcal{F}_1$ (by Jensen's inequality), and $\mathcal{F}_1$ contains sums of ridge functions.
  - A single neuron $\varphi(x, z^*)$ belongs to $\mathcal{F}_1$ but not $\mathcal{F}_2$.
  - Adaptivity to low-dimensional structures via feature learning with non-cursed rates [Bach'17]
  - Simplest instance compatible with transfer learning.

VARIATION-NORM SPACES

[Bengio et al'06, Rosset et al.'07, Bach'17]

- Alternatively, we can consider

$$\mathcal{F}_1 = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,;\, f(x) = \int_{\mathcal{D}} \varphi(x, z)\mu(dz); \|\mu\|_{TV} < \infty. \right\}.$$

- $\mathcal{F}_1$ is a Banach space, with norm $\|f\|_{\mathcal{F}_1} := \inf \left\{ \|\mu\|_{TV}; f = \int \varphi d\mu \right\}$.
  - Also known as ***Barron*** Spaces [Barron'90s, E et al '19].

- $\mathcal{F}_2 \subset \mathcal{F}_1$ (by Jensen's inequality), and $\mathcal{F}_1$ contains sums of ridge functions.
  - A single neuron $\varphi(x, z^*)$ belongs to $\mathcal{F}_1$ but not $\mathcal{F}_2$.
  - Adaptivity to low-dimensional structures via feature learning with non-cursed rates [Bach'17]
  - Simplest instance compatible with transfer learning.

- How to perform optimization and approximation in these spaces?

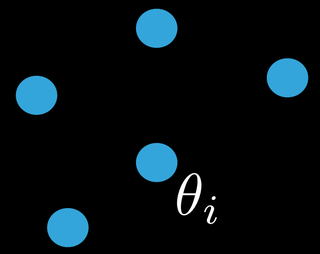
NEURAL NETWORKS AS PARTICLE INTERACTION SYSTEMS

- Regression on noise-less targets: $f^* \in L_2(\mathbb{R}^d, d\nu)$: target function.
- Single-hidden layer architecture and associated ERM:

Lifting

$$\Theta = (\theta_1 \dots \theta_n)\,, f(x;\Theta) = \frac{1}{n}\sum_{j\le n} a_j \varphi(x;z_j) = \frac{1}{n}\sum_{j\le n} \phi(x;\theta_j)\,, \theta_j = (a_j, z_j) \in \mathbb{R}\times\mathcal{D}.$$

$$\hat{\mathcal{R}}(\Theta) = \hat{\mathbb{E}}_\nu[\ell(f(x;\Theta), f^*(x))] + \lambda\mathcal{V}(\Theta)\ . \quad \mathcal{V}(\Theta) = \sum_j v(\theta_j), \text{ e.g. } v(\theta) = \|\theta\|^2$$



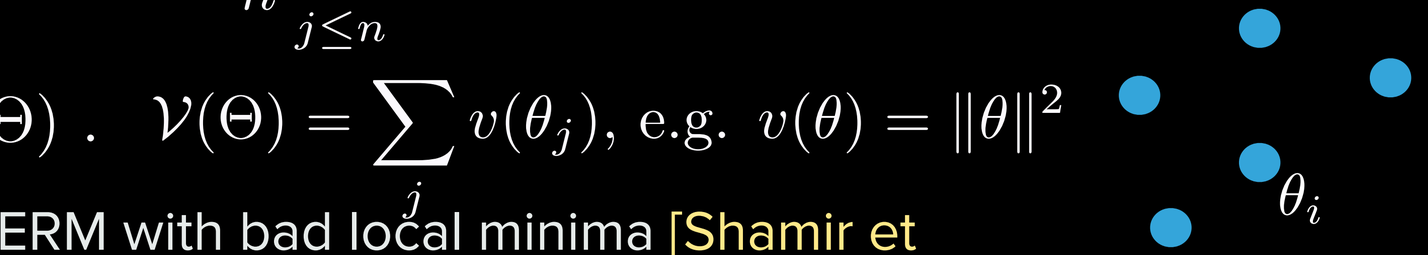
NEURAL NETWORKS AS PARTICLE INTERACTION SYSTEMS

- Regression on noise-less targets: $f^* \in L_2(\mathbb{R}^d, d\nu)$: target function.
- Single-hidden layer architecture and associated ERM:

Lifting

$$\Theta = (\theta_1 \dots \theta_n)\,, f(x;\Theta) = \frac{1}{n}\sum_{j\le n} a_j \varphi(x;z_j) = \frac{1}{n}\sum_{j\le n} \phi(x;\theta_j)\,, \theta_j = (a_j, z_j) \in \mathbb{R}\times\mathcal{D}.$$

$$\hat{\mathcal{R}}(\Theta) = \hat{\mathbb{E}}_\nu[\ell(f(x;\Theta), f^*(x))] + \lambda\mathcal{V}(\Theta)\ . \quad \mathcal{V}(\Theta) = \sum_j v(\theta_j), \text{ e.g. } v(\theta) = \|\theta\|^2$$



  - Non-linear in Θ, leads to non-convex ERM with bad local minima [Shamir et al., Venturi, Bandeira, **B,**'19]

NEURAL NETWORKS AS PARTICLE INTERACTION SYSTEMS

- Regression on noise-less targets: $f^* \in L_2(\mathbb{R}^d, d\nu)$: target function.
- Single-hidden layer architecture and associated ERM:

Lifting

$$\Theta = (\theta_1 \dots \theta_n)\,, f(x;\Theta) = \frac{1}{n}\sum_{j\le n} a_j \varphi(x;z_j) = \frac{1}{n}\sum_{j\le n} \phi(x;\theta_j)\,, \theta_j = (a_j, z_j) \in \mathbb{R}\times\mathcal{D}.$$

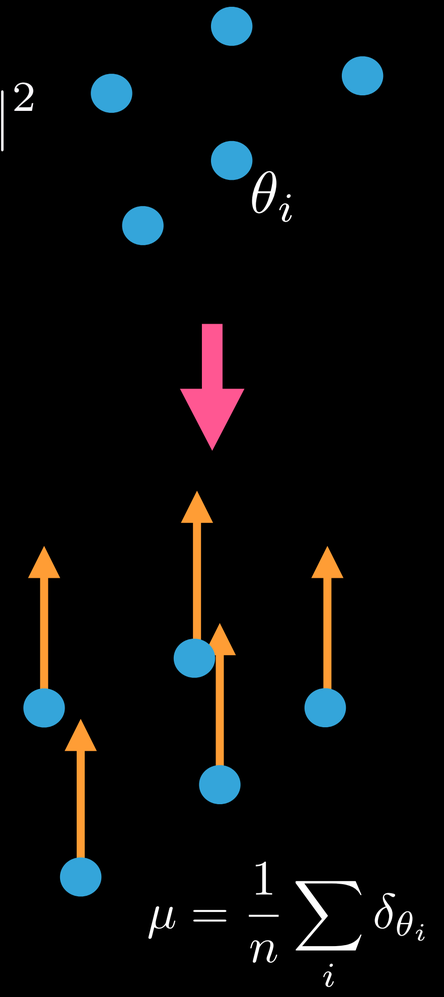
$$\hat{\mathcal{R}}(\Theta) = \hat{\mathbb{E}}_\nu[\ell(f(x;\Theta), f^*(x))] + \lambda\mathcal{V}(\Theta)\ . \quad \mathcal{V}(\Theta) = \sum_j v(\theta_j), \text{ e.g. } v(\theta) = \|\theta\|^2$$

  - Non-linear in Θ, leads to non-convex ERM with bad local minima [Shamir et al., Venturi, Bandeira, **B,**'19]

[Rosset et al., Bengio et al., Bach]

- ***Eulerian perspective:*** Rewrite the energy in terms of the empirical measure:

$$\mu^{(n)} = \frac{1}{n}\sum_{j\le n}\delta_{\theta_j}\ ,\ \text{so}\quad f(x;\Theta) = \int_{\tilde{\mathcal{D}}} \phi(x;\theta)\mu^{(n)}(d\theta)$$



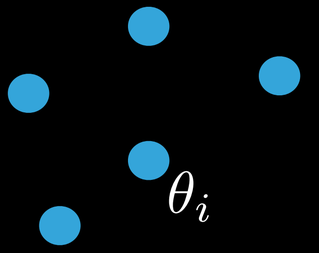
NEURAL NETWORKS AS PARTICLE INTERACTION SYSTEMS

- Regression on noise-less targets: $f^* \in L_2(\mathbb{R}^d, d\nu)$: target function.
- Single-hidden layer architecture and associated ERM:

Lifting

$$\Theta = (\theta_1 \dots \theta_n)\,, f(x;\Theta) = \frac{1}{n}\sum_{j\le n} a_j \varphi(x;z_j) = \frac{1}{n}\sum_{j\le n} \phi(x;\theta_j)\,, \theta_j = (a_j, z_j) \in \mathbb{R}\times\mathcal{D}.$$

$$\hat{\mathcal{R}}(\Theta) = \hat{\mathbb{E}}_\nu[\ell(f(x;\Theta), f^*(x))] + \lambda\mathcal{V}(\Theta)\ . \quad \mathcal{V}(\Theta) = \sum_j v(\theta_j), \text{ e.g. } v(\theta) = \|\theta\|^2$$



  - Non-linear in Θ, leads to non-convex ERM with bad local minima [Shamir et al., Venturi, Bandeira, **B,**'19]

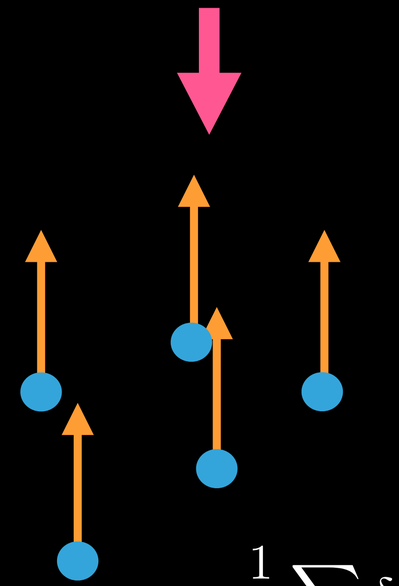
[Rosset et al., Bengio et al., Bach]

- ***Eulerian perspective:*** Rewrite the energy in terms of the empirical measure

$$\mu^{(n)} = \frac{1}{n}\sum_{j\le n}\delta_{\theta_j}\ ,\ \text{so}\quad f(x;\Theta) = \int_{\tilde{\mathcal{D}}} \phi(x;\theta)\mu^{(n)}(d\theta)$$

- ERM expressed in terms of μ is now convex:

$$\hat{\mathcal{R}}(\mu) = \hat{\mathbb{E}}_\nu\left[\ell\left(\int \phi(x,\theta)\mu(d\theta), f^*(x)\right)\right] + \lambda\int v(\theta)\mu(d\theta)\ .$$



$$\mu = \frac{1}{n}\sum_i \delta_{\theta_i}$$

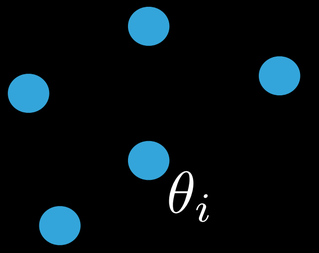
NEURAL NETWORKS AS PARTICLE INTERACTION SYSTEMS

- Regression on noise-less targets: $f^* \in L_2(\mathbb{R}^d, d\nu)$: target function.
- Single-hidden layer architecture and associated ERM:

Lifting

$$\Theta = (\theta_1 \dots \theta_n)\,, f(x;\Theta) = \frac{1}{n}\sum_{j\le n} a_j \varphi(x;z_j) = \frac{1}{n}\sum_{j\le n} \phi(x;\theta_j)\,, \theta_j = (a_j, z_j) \in \mathbb{R}\times\mathcal{D}.$$

$$\hat{\mathcal{R}}(\Theta) = \hat{\mathbb{E}}_\nu[\ell(f(x;\Theta), f^*(x))] + \lambda\mathcal{V}(\Theta)\ . \quad \mathcal{V}(\Theta) = \sum_j v(\theta_j), \text{ e.g. } v(\theta) = \|\theta\|^2$$



  - Non-linear in Θ, leads to non-convex ERM with bad local minima [Shamir et al., Venturi, Bandeira, **B,**'19]

[Rosset et al., Bengio et al., Bach]

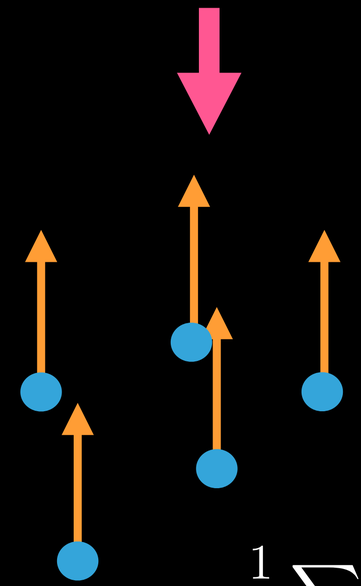
- ***Eulerian perspective:*** Rewrite the energy in terms of the empirical measure

$$\mu^{(n)} = \frac{1}{n}\sum_{j\le n}\delta_{\theta_j}\ ,\ \text{so}\quad f(x;\Theta) = \int_{\tilde{\mathcal{D}}} \phi(x;\theta)\mu^{(n)}(d\theta)$$

- ERM expressed in terms of μ is now convex:

$$\hat{\mathcal{R}}(\mu) = \hat{\mathbb{E}}_\nu\left[\ell\left(\int \phi(x,\theta)\mu(d\theta), f^*(x)\right)\right] + \lambda\int v(\theta)\mu(d\theta)\ .$$

- Training dynamics expressed in the measure domain?



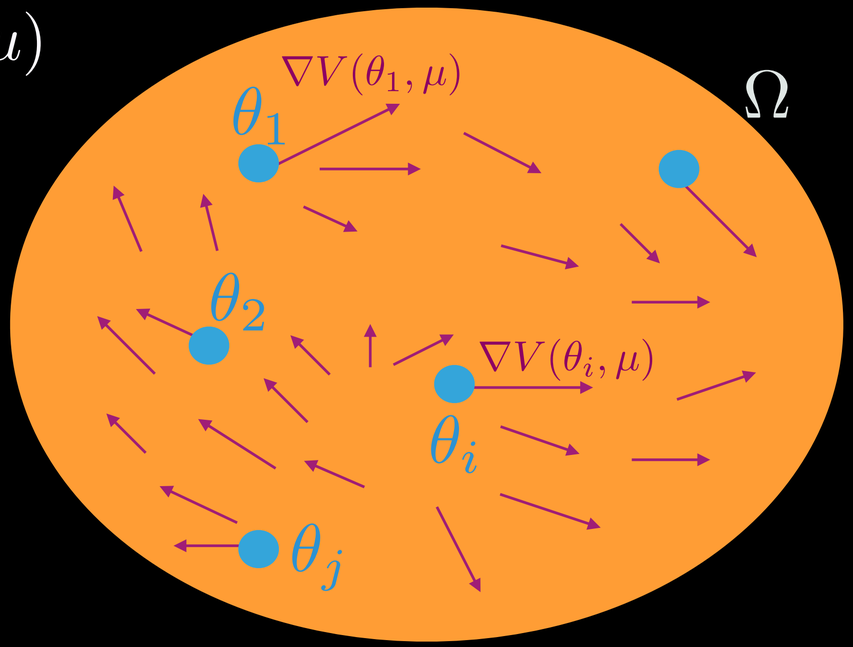
$$\mu = \frac{1}{n}\sum_i \delta_{\theta_i}$$

CONTINUITY EQUATION

[Mei, Montanari, Nguyen, PNAS'18] [Sirignano, Spiliopoulos,'18]
[Rotskoff, EVE, NeurIPS'18] [Chizat, Bach, NeurIPS'18]

- Particle gradients correspond to evaluating a scaled velocity field:

$$\frac{n}{2}\nabla_{\theta_i}\hat{\mathcal{R}}(\Theta) = \nabla V|_{\theta=\theta_i} \text{ , with } V(\cdot;\mu) = \hat{\mathcal{R}}'(\mu)$$

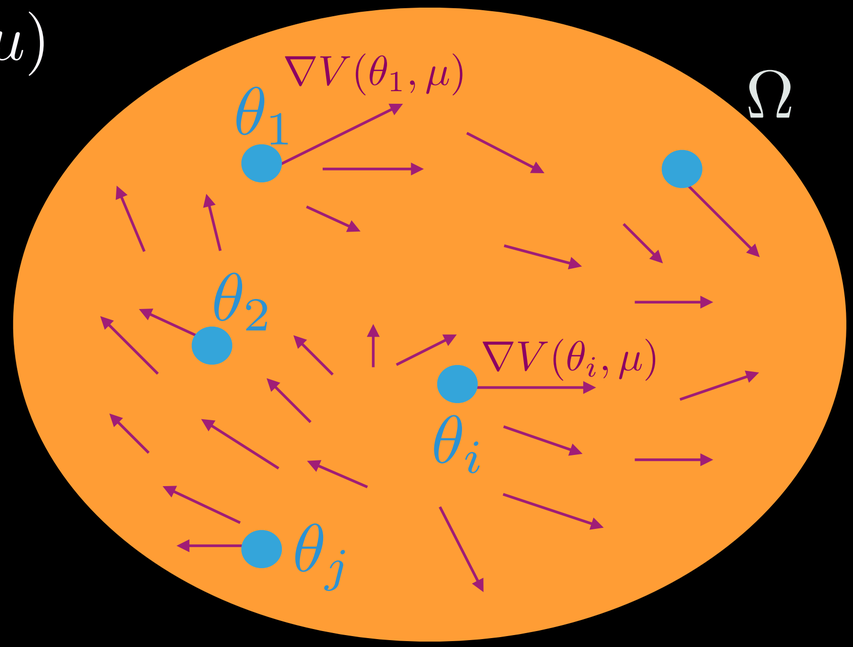


CONTINUITY EQUATION

[Mei, Montanari, Nguyen, PNAS'18] [Sirignano, Spiliopoulos,'18]
[Rotskoff, EVE, NeurIPS'18] [Chizat, Bach, NeurIPS'18]

- Particle gradients correspond to evaluating a scaled velocity field:

$$\frac{n}{2}\nabla_{\theta_i}\hat{\mathcal{R}}(\Theta) = \nabla V|_{\theta=\theta_i} \ , \text{with } V(\cdot;\mu) = \hat{\mathcal{R}}'(\mu)$$



- For general time-dependent measures μ_t, their evolution under a velocity field $V(\theta;\mu_t)$ is given by a ***continuity equation***:

$$\partial_t \mu_t = \text{div}(\mu_t \nabla V) \ , \ \mu(0) = \mu^{(0)}$$

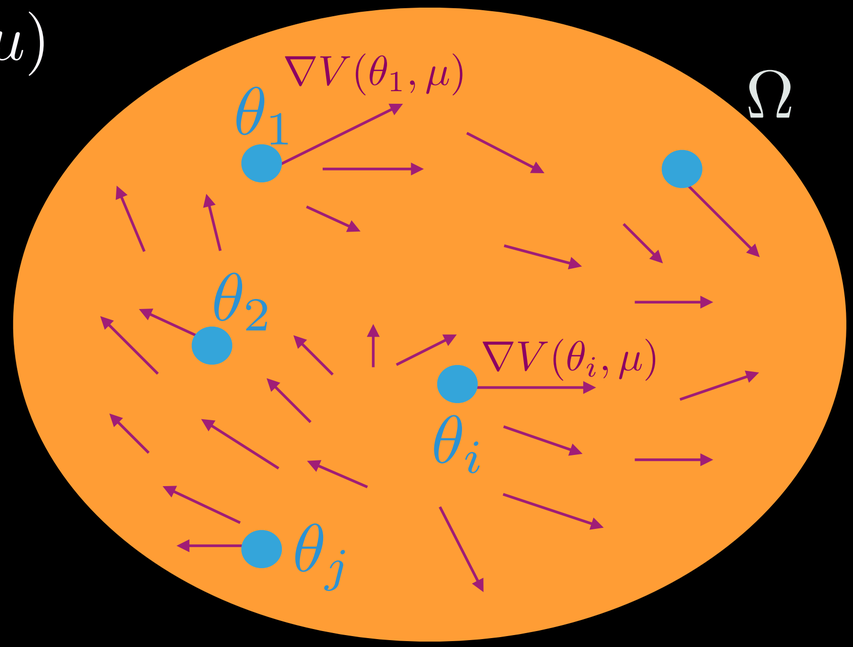
  - Gradient flow of $\hat{\mathcal{R}}$ for the Wasserstein metric W_2 in $\mathcal{P}(\Omega)$
    [Ambrosio et al.]

CONTINUITY EQUATION

[Mei, Montanari, Nguyen, PNAS'18] [Sirignano, Spiliopoulos,'18]
[Rotskoff, EVE, NeurIPS'18] [Chizat, Bach, NeurIPS'18]

- Particle gradients correspond to evaluating a scaled velocity field:

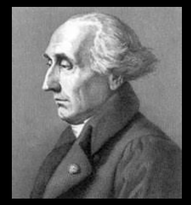
$$\frac{n}{2}\nabla_{\theta_i}\hat{\mathcal{R}}(\Theta) = \nabla V|_{\theta=\theta_i} \ , \text{with } V(\cdot;\mu) = \hat{\mathcal{R}}'(\mu)$$



- For general time-dependent measures μ_t, their evolution under a velocity field $V(\theta;\mu_t)$ is given by a ***continuity equation***:

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla V) \ , \ \ \mu(0) = \mu^{(0)}$$

  - Gradient flow of $\hat{\mathcal{R}}$ for the Wasserstein metric W_2 in $\mathcal{P}(\Omega)$
    [Ambrosio et al.]
  - ***Exact description*** of particle gradient for atomic measures.



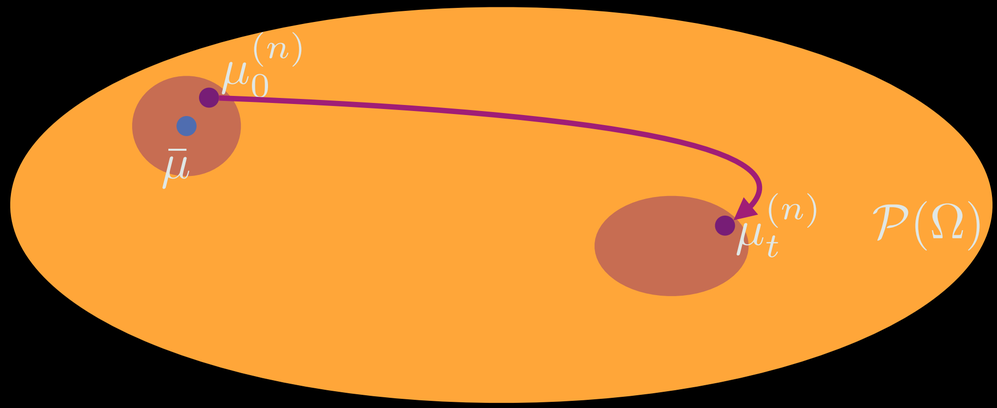
LAGRANGIAN
Non-Convexity
Euclidean Dynamics
Finite-dimensional

EULERIAN
Convexity
Non-Euclidean Dynamics
Infinite-dimensional

MEAN-FIELD LIMIT

[Mei, Montanari, Nguyen, PNAS'18] [Sirignano, Spiliopoulos,'18]
[Rotskoff, EVE, NeurIPS'18] [Chizat, Bach, NeurIPS'18]

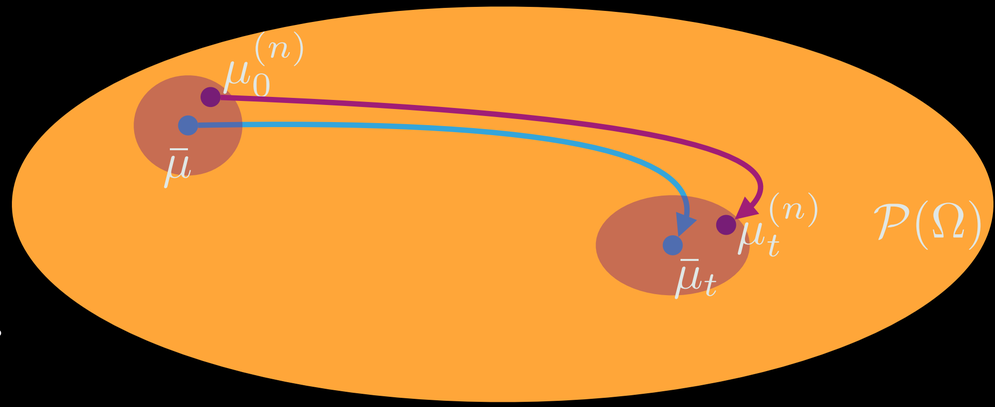
- Consider the evolution of the particle system as n grows.
- $\mu_t^{(n)}$: state of the system after time t, with $\theta_i(0) \sim \bar{\mu}$ iid.



MEAN-FIELD LIMIT

[Mei, Montanari, Nguyen, PNAS'18] [Sirignano, Spiliopoulos,'18]
[Rotskoff, EVE, NeurIPS'18] [Chizat, Bach, NeurIPS'18]

- Consider the evolution of the particle system as n grows.
- $\mu_t^{(n)}$: state of the system after time t, with $\theta_i(0) \sim \bar{\mu}$ iid.



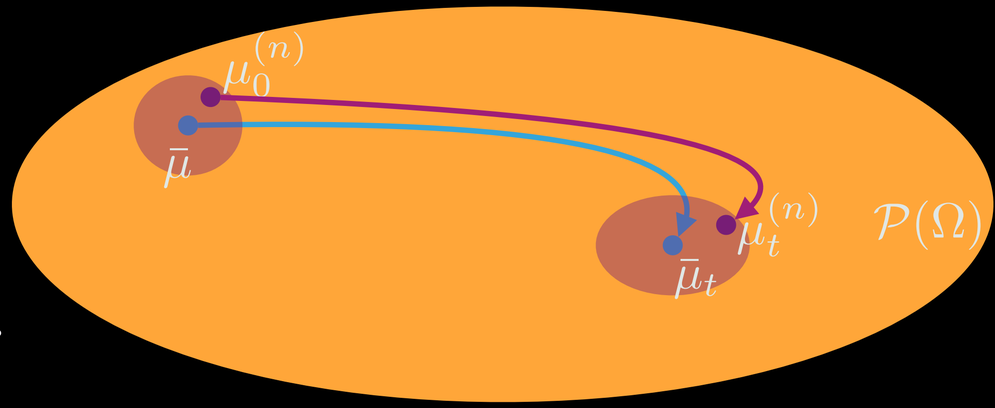
Theorem: [R,EVE'18], [CB'18], [MMN'18],[SS'18]
For any fixed $T > 0$, $\mu_T^{(n)}$ converges weakly to μ_T as $n \to \infty$, which solves $\partial_t \mu_t = \mathrm{div}(\nabla V \mu_t)$ with $\mu_0 = \bar{\mu}$.

- Dynamics and sampling commute in the limit (when it exists).

MEAN-FIELD LIMIT

[Mei, Montanari, Nguyen, PNAS'18] [Sirignano, Spiliopoulos,'18]
[Rotskoff, EVE, NeurIPS'18] [Chizat, Bach, NeurIPS'18]

- Consider the evolution of the particle system as n grows.
- $\mu_t^{(n)}$: state of the system after time t, with $\theta_i(0) \sim \bar{\mu}$ iid.



Theorem: [R,EVE'18], [CB'18], [MMN'18],[SS'18]
For any fixed $T > 0$, $\mu_T^{(n)}$ converges weakly to μ_T as $n \to \infty$, which solves $\partial_t \mu_t = \mathrm{div}(\nabla V \mu_t)$ with $\mu_0 = \bar{\mu}$.

- Dynamics and sampling commute in the limit (when it exists).

- Key convergence questions:

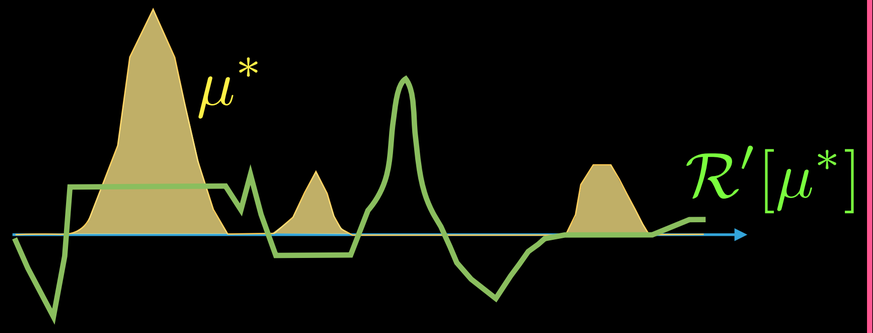
In t : Convergence properties of this PDE towards minimisers of $\mathcal{R}$?

In n: Beyond Mean-Field limit: What is the scale of the fluctuations?

GLOBAL CONVERGENCE IN THE MEAN-FIELD LIMIT

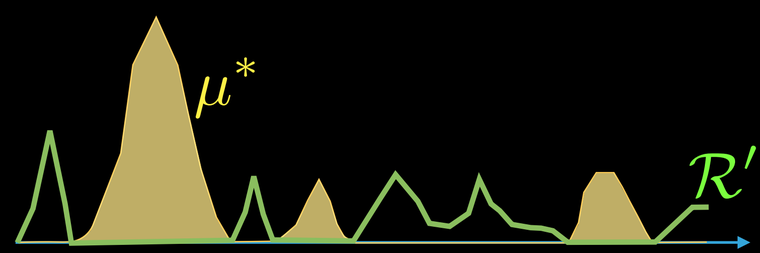
- Stationary points of Wasserstein gradient flow of $\mathcal{R}$ $\neq$global minimisers of $\mathcal{R}$ (lack of displacement convexity).

STATIONARY POINTS



$$\nabla \mathcal{R}'[\mu^*](\theta) = 0 \text{ for } \theta \in \text{supp}(\mu^*)$$

GLOBAL MINIMISERS



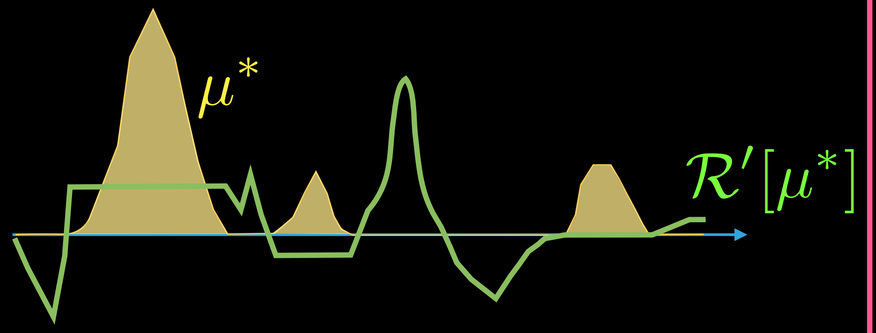
$$\mathcal{R}'[\mu^*](\theta) \geq 0 \text{ and}$$
$$\mathcal{R}'[\mu^*](\theta) = 0 \text{ for } \theta \in \text{supp}(\mu^*)$$

- Major obstacle: do not lose mass “too soon”.

GLOBAL CONVERGENCE IN THE MEAN-FIELD LIMIT

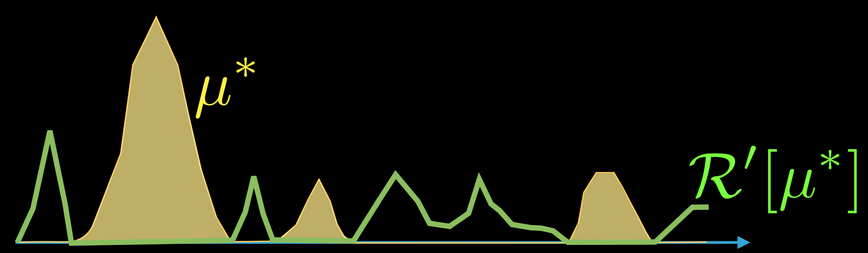
- Stationary points of Wasserstein gradient flow of $\mathcal{R}$ $\neq$global minimisers of $\mathcal{R}$ (lack of displacement convexity).

STATIONARY POINTS



$$\nabla \mathcal{R}'[\mu^*](\theta) = 0 \text{ for } \theta \in \text{supp}(\mu^*)$$

GLOBAL MINIMISERS



$$\mathcal{R}'[\mu^*](\theta) \geq 0 \text{ and}$$
$$\mathcal{R}'[\mu^*](\theta) = 0 \text{ for } \theta \in \text{supp}(\mu^*)$$

  - Major obstacle: do not lose mass “too soon”.

- Existing results:
  - Additive noise in the dynamics [Mei et al.] enforces smoothness (and full support) through an elliptic Euler-Lagrange problem: $\partial_t \mu_t = \text{div}(\nabla V \mu_t) + \beta^{-1} \Delta \mu_t$
  - Full initial support of μ_0 and homogeneity of the neuron $\phi(\lambda\theta, x) = \lambda^p \phi(\theta, x)$ [Chizat & Bach]. Captures qualitative behavior of deterministic gradient flow.
  - General unbalanced transport through extra lifting stage [Jelassi, Rotskoff, B, EVE]

TOWARDS FINITE-WIDTH GUARANTEES

[NeurIPS'20]

- Previous global convergence results hold in the mean-field $n \to \infty$ limit.
- What is the typical size of fluctuations?

TOWARDS FINITE-WIDTH GUARANTEES

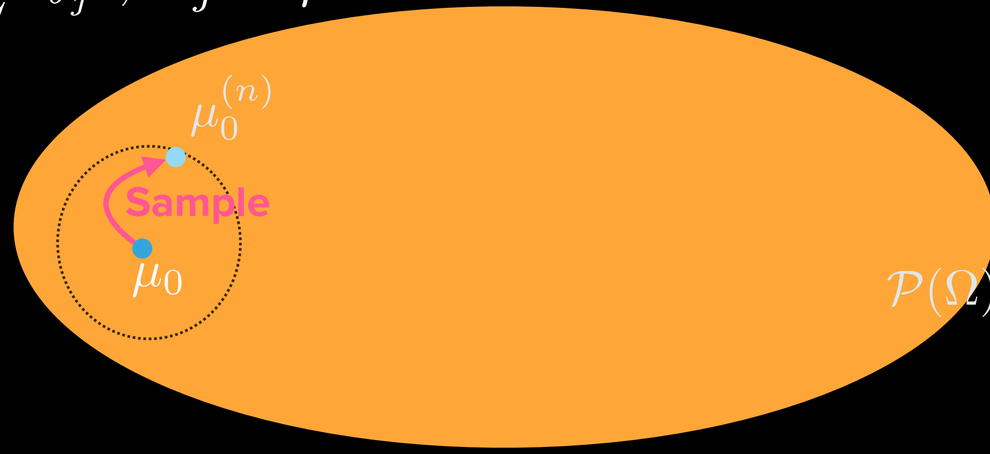
[NeurIPS'20]

- Previous global convergence results hold in the mean-field $n \to \infty$ limit.
- What is the typical size of fluctuations?
- Variation-Norm Spaces with Tykonov reg. are efficiently "samplable" :

Recall $\mathcal{F} = \left\{ f(x) = \int_{\tilde{\mathcal{D}}} \phi(x,\theta)\mu(d\theta);\ \int_{\tilde{\mathcal{D}}} \|\theta\|^2 \mu(d\theta) < \infty \right\}$ (ϕ Lipschitz)

- Monte-Carlo estimate of $f(x) = \mathbb{E}_\mu[\phi(x,\theta)]$:

$$f^{(n)}(x) = \frac{1}{n}\sum_{j \le n} \phi(x,\theta_j) = \mathbb{E}_{\mu^{(n)}}[\phi(x,\theta)]\ ,\ \mu^{(n)} = \frac{1}{n}\sum_j \delta_{\theta_j}\ ,\ \theta_j \sim \mu$$



TOWARDS FINITE-WIDTH GUARANTEES

[NeurIPS'20]

- Previous global convergence results hold in the mean-field $n \to \infty$ limit.
- What is the typical size of fluctuations?
- Variation-Norm Spaces with Tykonov reg. are efficiently "samplable" :

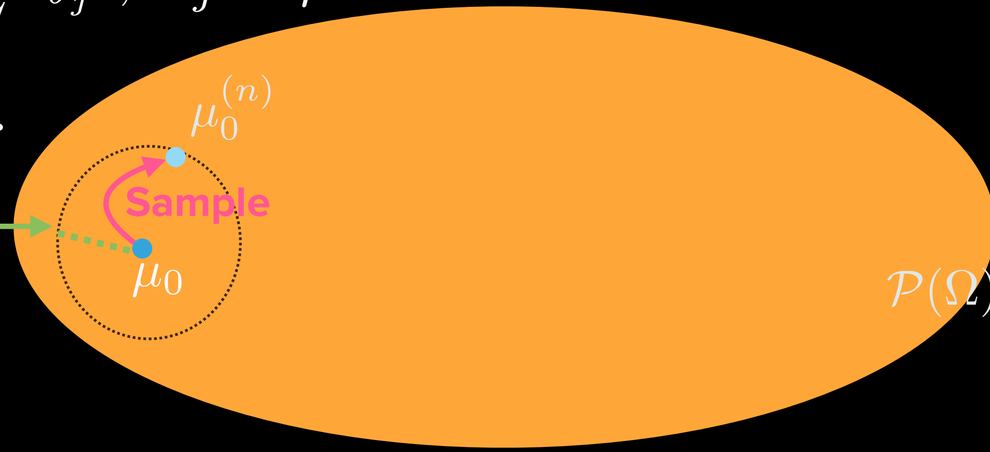
Recall $\mathcal{F} = \left\{ f(x) = \int_{\tilde{\mathcal{D}}} \phi(x,\theta)\mu(d\theta);\ \int_{\tilde{\mathcal{D}}} \|\theta\|^2 \mu(d\theta) < \infty \right\}$ (ϕ Lipschitz)

- Monte-Carlo estimate of $f(x) = \mathbb{E}_\mu[\phi(x,\theta)]$:

$$f^{(n)}(x) = \frac{1}{n}\sum_{j\le n} \phi(x,\theta_j) = \mathbb{E}_{\mu^{(n)}}[\phi(x,\theta)]\ ,\ \mu^{(n)} = \frac{1}{n}\sum_j \delta_{\theta_j}\ ,\ \theta_j \sim \mu$$

- Statistical Rate: $\mathcal{K}(\mu) := \mathbb{E}_\mu \|\phi(\cdot,\theta)\|_\nu^2 - \|\mathbb{E}_\mu \phi(\cdot,\theta)\|_\nu^2$.

$$\mathbb{E}_\theta \|f - f^{(n)}\|_\nu^2 = \frac{\mathcal{K}(\mu)}{n} \le \frac{C\|f\|_{\mathcal{F}}^2}{n}\ .$$



TOWARDS FINITE-WIDTH GUARANTEES

[NeurIPS'20]

- Previous global convergence results hold in the mean-field $n \to \infty$ limit.
- What is the typical size of fluctuations?
- Variation-Norm Spaces with Tykonov reg. are efficiently "samplable" :

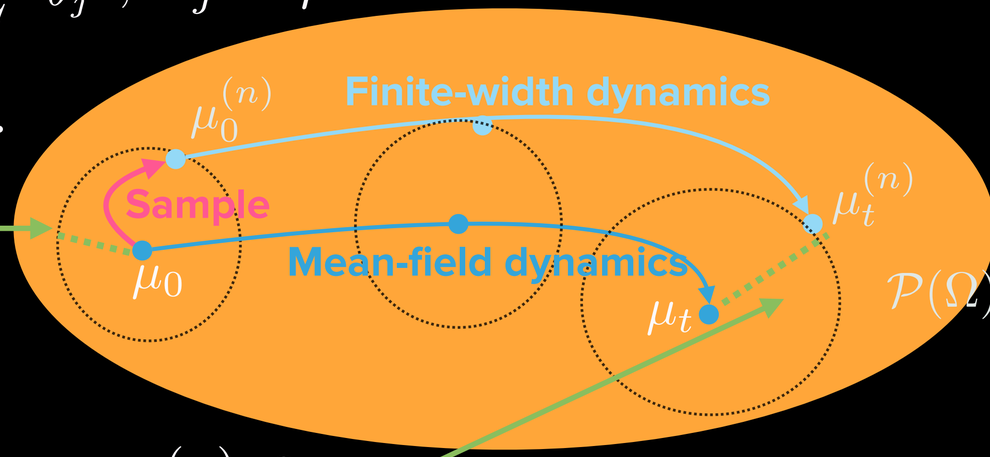
Recall $\mathcal{F} = \left\{ f(x) = \int_{\tilde{\mathcal{D}}} \phi(x,\theta)\mu(d\theta);\ \int_{\tilde{\mathcal{D}}} \|\theta\|^2 \mu(d\theta) < \infty \right\}$ (ϕ Lipschitz)

- Monte-Carlo estimate of $f(x) = \mathbb{E}_\mu[\phi(x,\theta)]$:

$$f^{(n)}(x) = \frac{1}{n}\sum_{j\le n} \phi(x,\theta_j) = \mathbb{E}_{\mu^{(n)}}[\phi(x,\theta)]\ ,\ \mu^{(n)} = \frac{1}{n}\sum_j \delta_{\theta_j}\ ,\ \theta_j \sim \mu$$

- Statistical Rate: $\mathcal{K}(\mu) := \mathbb{E}_\mu \|\phi(\cdot,\theta)\|_\nu^2 - \|\mathbb{E}_\mu \phi(\cdot,\theta)\|_\nu^2$.

$$\mathbb{E}_\theta \|f - f^{(n)}\|_\nu^2 = \frac{\mathcal{K}(\mu)}{n} \le \frac{C\|f\|_{\mathcal{F}}^2}{n}\ .$$



- How does this sampling error evolve with training?

$$f_t^{(n)} = \frac{1}{n}\sum_j \phi(x,\theta_j(t)) = \mathbb{E}_{\mu_t}[\phi(x,\theta)]$$

$$\mathbb{E}_\theta \|f_t - f_t^{(n)}\|_\nu^2$$

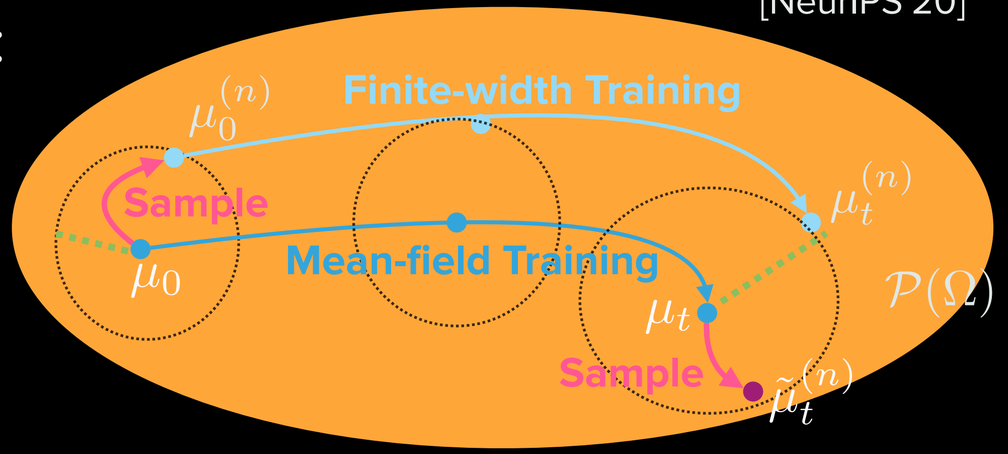
DYNAMIC CLT FOR SHALLOW NEURAL NETWORKS

[NeurIPS'20]

- A natural baseline is the MC error associated with μ_t :

$$\tilde{\mu}_t^{(n)} = \frac{1}{n}\sum_{j\le n}\delta_{\theta_j}\ ,\ \theta_j \sim \mu_t \quad \tilde{f}_t^{(n)}(x) = \int \phi(x,\theta)\tilde{\mu}_t^{(n)}(d\theta)$$

$$\mathbb{E}_\theta \| \tilde{f}_t^{(n)} - f_t \|_\nu^2 = \frac{\mathcal{K}(\mu_t)}{n}\ .$$



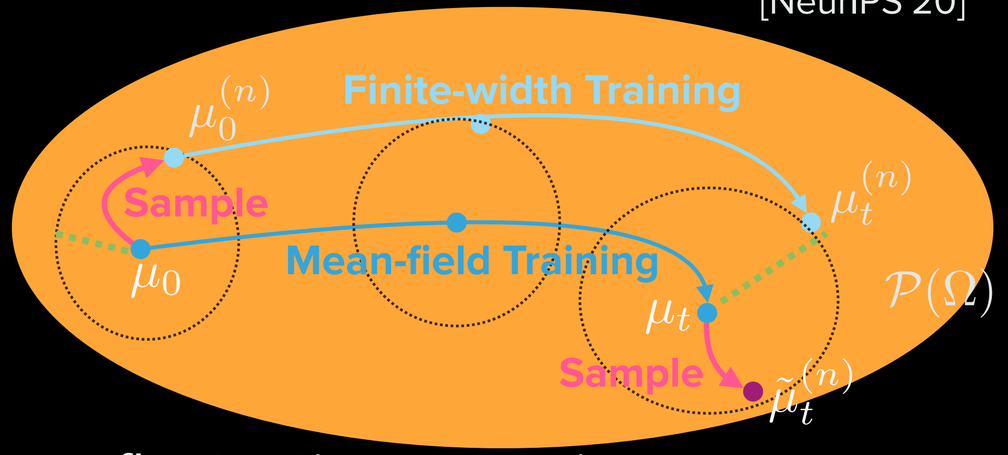
DYNAMIC CLT FOR SHALLOW NEURAL NETWORKS

[NeurIPS'20]

- A natural baseline is the MC error associated with μ_t:

$$\tilde{\mu}_t^{(n)} = \frac{1}{n}\sum_{j\le n}\delta_{\theta_j}\ ,\ \theta_j \sim \mu_t \quad \tilde{f}_t^{(n)}(x) = \int \phi(x,\theta)\tilde{\mu}_t^{(n)}(d\theta)$$

$$\mathbb{E}_\theta \| \tilde{f}_t^{(n)} - f_t \|_\nu^2 = \frac{\mathcal{K}(\mu_t)}{n}\ .$$



- Under appropriate mean-field convergence assumptions, fluctuations remain asymptotically bounded by this Monte-Carlo rate:

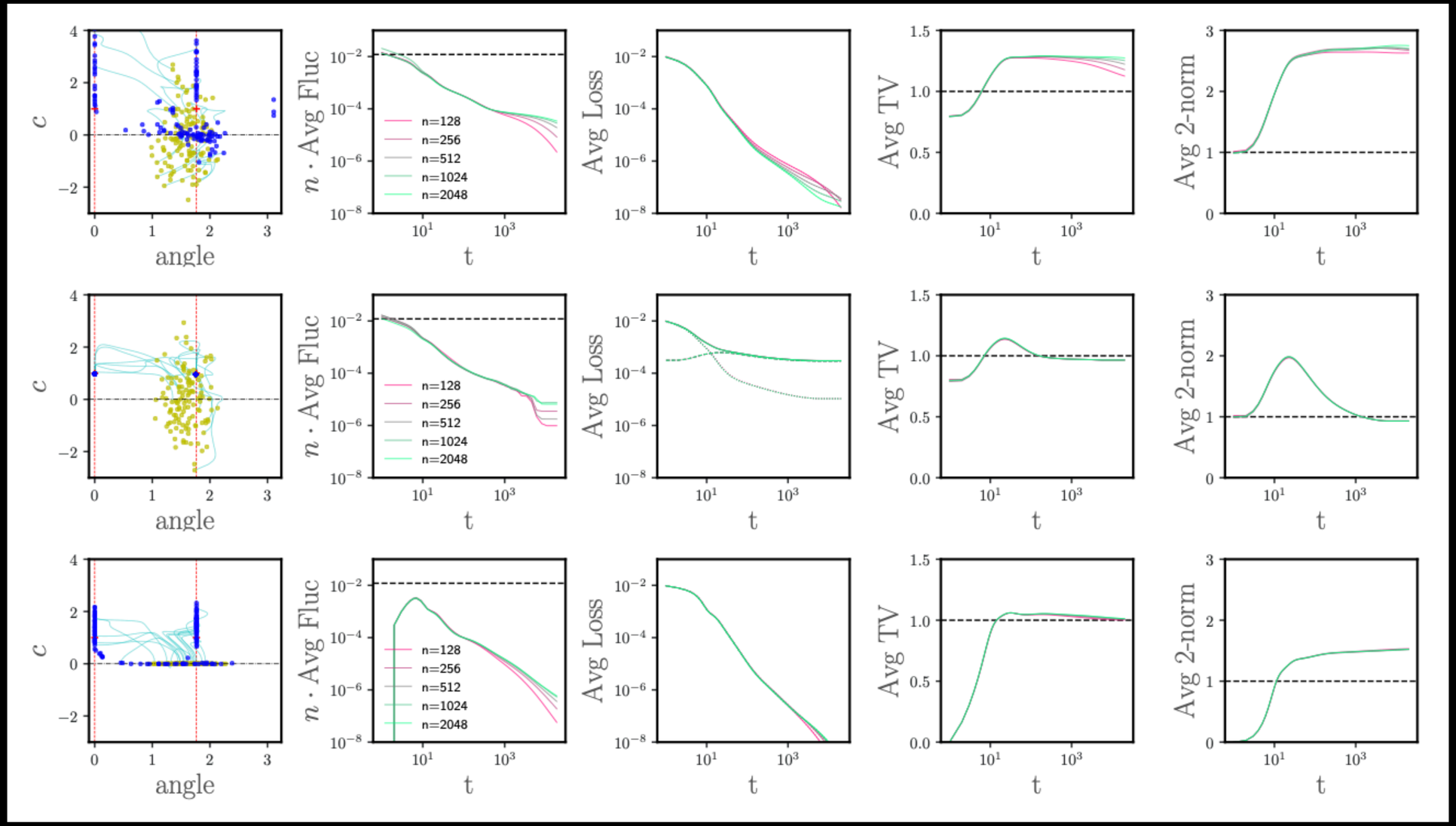
Theorem [CRBV'20]: Under appropriate smoothness and mean-field convergence assumptions, it holds

$$\lim_{t\to\infty}\lim_{n\to\infty} n\mathbb{E}_\theta \| f_t^{(n)} - f_t \|_\nu^2 \le \mathcal{K}(\mu_\infty)\ .$$

- Extends finite horizon CLT bounds from [Braun & Hepp,'70s] (also [Spilopoulos'19, De Bortoli et al.'20]). [Chizat'19] establishes zero fluctuations on sparse well-conditioned recovery problems.
- In the unregularised interpolating regime, fluctuations vanish at the MC scale.
- Proof idea: Baseline fluctuations dominate dynamic fluctuations through a Volterra Kernel.

NUMERICAL EXPERIMENTS: TEACHER-STUDENT SETUP

[NeurIPS'20]



- We verify scale of fluctuations at or below MC.

CURRENT: TOWARDS QUANTITATIVE GUARANTEES

Jon Niles-Weed

- Previous CLT results are still qualitative (limit of infinitely wide network).
  - Quantitative guarantees polynomial in the problem parameters?

CURRENT: TOWARDS QUANTITATIVE GUARANTEES

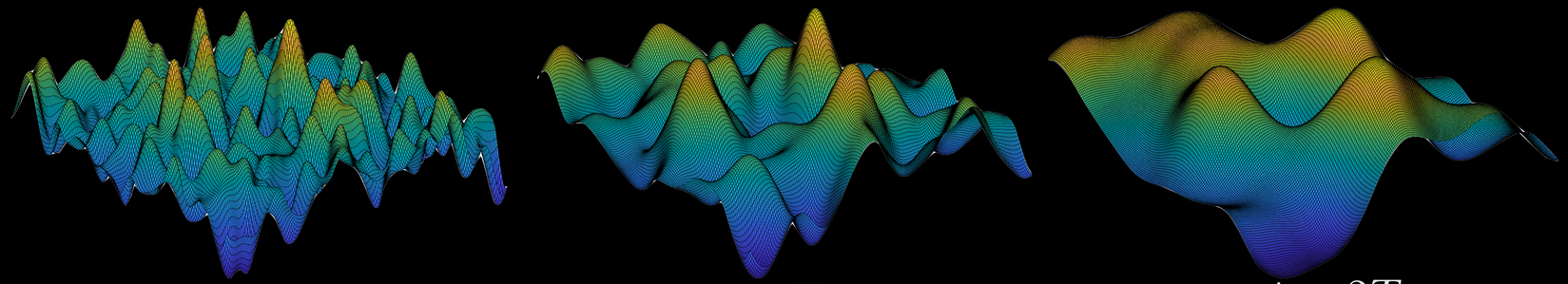
Jon Niles-Weed

- Previous CLT results are still qualitative (limit of infinitely wide network).
  - Quantitative guarantees polynomial in the problem parameters?
- Super-Polynomial Statistical-Query lower bounds for learning certain planted shallow models [Goel et al. ‘20, Diakonikolas et al.’20, Vempala’18, Kearns’94]
  - Targets of the form $f^*(x) = \sum_{k=1}^{K} \epsilon_k \rho(x^\top \theta_k)$ for adversarially chosen $\{\epsilon_k, \theta_k\}$, K diverging with d.

CURRENT: TOWARDS QUANTITATIVE GUARANTEES

Jon Niles-Weed

- Previous CLT results are still qualitative (limit of infinitely wide network).
  - Quantitative guarantees polynomial in the problem parameters?
- Super-Polynomial Statistical-Query lower bounds for learning certain planted shallow models [Goel et al. '20, Diakonikolas et al.'20, Vempala'18, Kearns'94]
  - Targets of the form $f^*(x) = \sum_{k=1}^{K} \epsilon_k \rho(x^\top \theta_k)$ for adversarially chosen $\{\epsilon_k, \theta_k\}$, K diverging with d.
- Generic planted model studied by lifting optimization landscape $\mathcal{R}(\Theta)$ using a diffusion semigroup (e.g heat kernel): $\widetilde{\mathcal{R}}(\Theta, t)$ solves $\partial_t \widetilde{\mathcal{R}}(\Theta, t) = \Delta \widetilde{\mathcal{R}}(\Theta, t)$, $\widetilde{\mathcal{R}}(\Theta, 0) = \mathcal{R}(\Theta)$.



$t = 0$ $t = T$ $t = 2T$

- Preliminary results: provably learning random planted shallow models with $K = O(d^\alpha), \alpha < 1$ using $O(d)$ Frank-Wolfe smoothed gradient descent [Jaggi, Bach, Zhong et al., Ma, Lee & Ge]

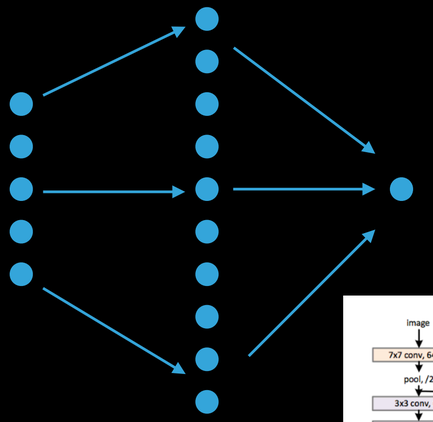
BEYOND VARIATION-NORM SPACES: DEPTH SEPARATION

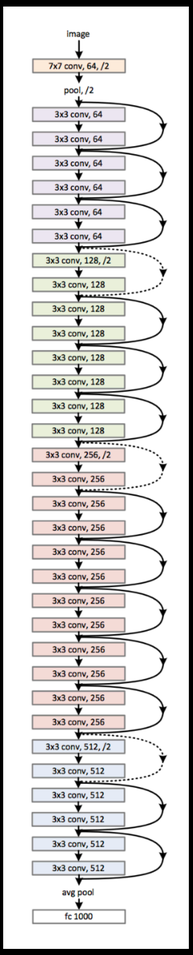
[submitted]

- Functions in $\mathcal{F}_1$ are expressed as sparse sums of ridge functions.
- We just saw they admit statistical rates of approximation using Monte-Carlo estimates: if $f \in \mathcal{F}$, there exists n-term approximation f_n such that

$$\|f - f_n\|_{L^2(\Omega)} \leq n^{-1/2} \|f\|_{\mathcal{F}_1} \sup_{z \in D} \|\varphi(., z)\|$$

  - Rates can be slightly improved in certain regimes [Bach'17]
- Explicit control using Fourier moments, e.g. $\|f\|_{\mathcal{F}_1} \lesssim \int \|\omega\|_1^2 |\hat{f}(\omega)| d\omega$ for the ReLU and compact domain Ω [Barron'93, Klusowski, Barron,'18, Ongie et al.'19]





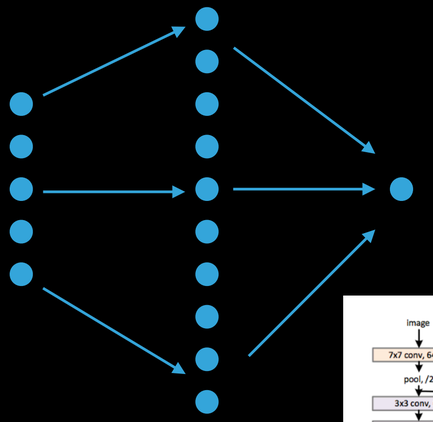
BEYOND VARIATION-NORM SPACES: DEPTH SEPARATION

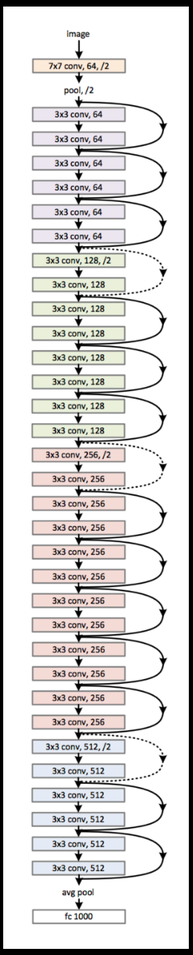
[submitted]

- Functions in $\mathcal{F}_1$ are expressed as sparse sums of ridge functions.
- We just saw they admit statistical rates of approximation using Monte-Carlo estimates: if $f \in \mathcal{F}$, there exists n-term approximation f_n such that

$$\|f - f_n\|_{L^2(\Omega)} \leq n^{-1/2} \|f\|_{\mathcal{F}_1} \sup_{z \in D} \|\varphi(., z)\|$$

  - Rates can be slightly improved in certain regimes [Bach'17]
- Explicit control using Fourier moments, e.g. $\|f\|_{\mathcal{F}_1} \lesssim \int \|\omega\|_1^2 |\hat{f}(\omega)| d\omega$ for the ReLU and compact domain Ω [Barron'93, Klusowski, Barron,'18, Ongie et al.'19]

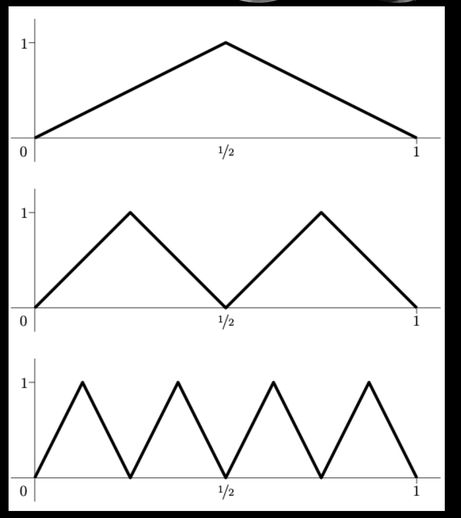




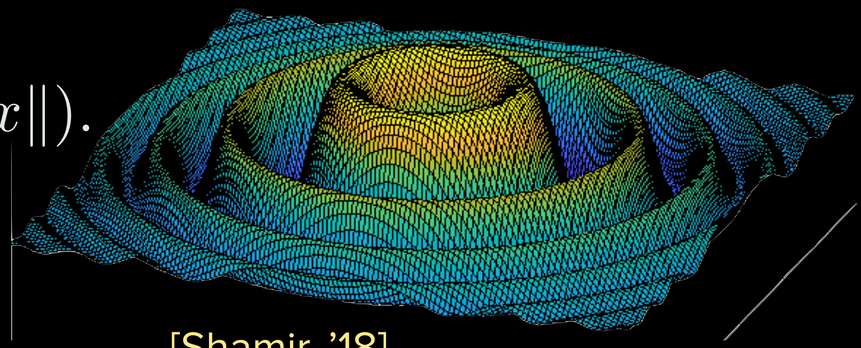
- Which function classes are not well approximated in $\mathcal{F}_1$, but are approximable/learnable by deeper architectures efficiently?
- How necessary are these (strong) Fourier decay conditions?
- Depth Separation in presence of input structure, eg images?

DEPTH SEPARATION PRIOR WORK

- Rich literature in boolean [Rossman, Hastad'68] or threshold [Hajnal'93] circuit lower bounds.
- [Martens et al'13] shows lower bounds for RBMs.
- [Telgarsky'15] Exploits combinatorial limitations of shallow networks
  - Refined periodicity analysis in [Chatziafratis et al'20].
- [Montufar et al.] bound number of linear regions of deep ReLU nets.
- [Eldan, Shamir, Safran, Daniely] construct oscillatory functions with depth-separation. Provably require $\exp(d)$ width for shallow model, but $\mathrm{poly}(d)$ for deeper neural network.
  - Constructions are inherently low-dimensional, e.g. $f(x) = g(\|x\|)$.
- Towards more general function separations?



[Telgarsky, '15]



[Shamir, '18]

LOWER BOUNDS FOR PIECE-WISE OSCILLATORY FUNCTIONS


- Towards more expressive example functions, we consider functions with piece-wise structure.

- Approximation lower bound for shallow neural networks with efficient approximation with depth-3 networks:

$$D_\mu(f,g) = \mathbb{E}_\mu |f(x) - g(x)|^2$$

Theorem [BJV'20]: Let $f^*(x) = \exp\{i\langle \omega_d, \rho(Ux+b)\rangle\}$ with $U \in \mathbb{R}^{d\times d}$, $\|\omega_d\| = \Omega(d^3)$ and $\rho(t) = \max(0,t)$. Let μ be a heavy-tailed distribution. Then

(i) f^* is not $\Omega(1)$-approximable by any shallow $\mathsf{exp}(o(d))$-wide network.

(ii) there exists a $\mathsf{poly}(d, \epsilon^{-1})$ 3-layer ReLU network f such that $D_\mu(f, f^*) \leq \epsilon$.

LOWER BOUNDS FOR PIECE-WISE OSCILLATORY FUNCTIONS

- Towards more expressive example functions, we consider functions with piece-wise structure.

- Approximation lower bound for shallow neural networks with efficient approximation with depth-3 networks:

$$D_\mu(f,g) = \mathbb{E}_\mu |f(x) - g(x)|^2$$

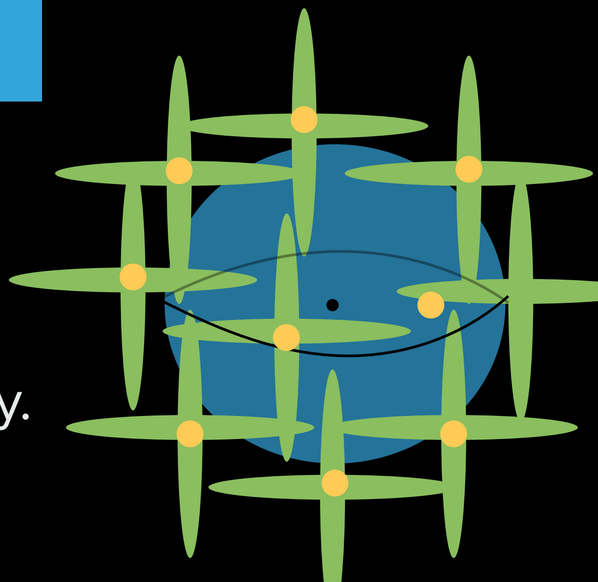
Theorem [BJV'20]: Let $f^*(x) = \exp\{i\langle \omega_d, \rho(Ux+b)\rangle\}$ with $U \in \mathbb{R}^{d\times d}$, $\|\omega_d\| = \Omega(d^3)$ and $\rho(t) = \max(0,t)$. Let μ be a heavy-tailed distribution. Then

(i) f^* is not $\Omega(1)$-approximable by any shallow $\mathsf{exp}(o(d))$-wide network.

(ii) there exists a $\mathsf{poly}(d, \epsilon^{-1})$ 3-layer ReLU network f such that $D_\mu(f, f^*) \le \epsilon$.

- Proof builds from [Eldan & Shamir'16], based on Fourier analysis, with tighter control on the data distribution.

- Heavy-tailed and increasing oscillation assumptions are (jointly) necessary.



UPPER BOUNDS FOR DEEP RELU NETWORKS

- In Compact domains, situation is more favorable towards shallow approximation:

Theorem [BJV'20]: Let $f^*(x)$ be a depth-L ReLU network with weights $\|W_l\|_\infty = \Theta(1)$ for $l \leq L$. Then $\forall \epsilon > 0$ there is a shallow ReLU network f_n such that $D_{\mathbb{S}^d,\infty}(f^*, f_n) \leq \epsilon$ of width

$$n \geq \left(\Theta(\exp L)(1+\epsilon^{-2})\mathsf{poly}(d)\right)^{\Omega(\epsilon^{-L})} .$$

- Extends previous results in [Safran, Eldan, Shamir'19] for radial functions.
- Rate is polynomial in d , but exponential in ϵ^{-1}.
- We can extend to D_μ with μ sub-Gaussian.

UPPER BOUNDS FOR DEEP RELU NETWORKS

- In Compact domains, situation is more favorable towards shallow approximation:

Theorem [BJV'20]: Let $f^*(x)$ be a depth-L ReLU network with weights $\|W_l\|_\infty = \Theta(1)$ for $l \leq L$. Then $\forall \epsilon > 0$ there is a shallow ReLU network f_n such that $D_{\mathbb{S}^d,\infty}(f^*, f_n) \leq \epsilon$ of width

$$n \geq \left(\Theta(\exp L)(1+\epsilon^{-2})\mathsf{poly}(d)\right)^{\Omega(\epsilon^{-L})} .$$

  - Extends previous results in [Safran, Eldan, Shamir'19] for radial functions.
  - Rate is polynomial in d, but exponential in ϵ^{-1}.
  - We can extend to D_μ with μ sub-Gaussian.

- Take-home: approximation gaps with depth, but no optimization guarantees (yet!)
- For ***unstructured*** inputs, depth brings unclear approximation advantages.

  Not much beyond Barron's Fourier Sparsity condition.

  Current techniques hard to extend to deeper models [Vardi & Shamir'20]

LEARNING UNDER SYMMETRY

- So far, we have considered the fully-connected setting with generic d-dimensional inputs.

LEARNING UNDER SYMMETRY

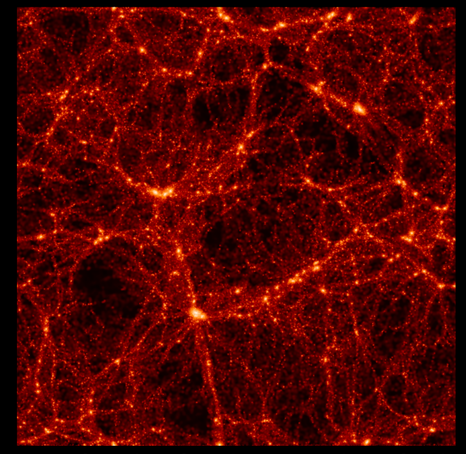
- So far, we have considered the fully-connected setting with generic d-dimensional inputs.
- Simple framework to study symmetries: permutation-invariant functions:

$$f : \{\Omega^k; k \in \mathbb{N}\} \to \mathbb{R} \text{ such that}$$

$$\Omega \subseteq \mathbb{R}^d \qquad f(x_{\pi(1)}, \ldots, x_{\pi(k)}) = f(x_1, \ldots, x_k) \, \forall \, k, x_j \in \Omega, \pi \in \mathsf{S}_k.$$

- E.g particle interaction systems, 3d point-clouds.

(Source: S. Grunewald, credit Qi et al.)



Cosmological n-body simulations
Joerg Colberg, Virgo Simulation

LEARNING UNDER SYMMETRY

- So far, we have considered the fully-connected setting with generic d-dimensional inputs.
- Simple framework to study symmetries: permutation-invariant functions:

$$f : \{\Omega^k; k \in \mathbb{N}\} \to \mathbb{R} \text{ such that}$$

$$\Omega \subseteq \mathbb{R}^d \qquad f(x_{\pi(1)}, \ldots, x_{\pi(k)}) = f(x_1, \ldots, x_k) \, \forall \, k, x_j \in \Omega, \pi \in \mathsf{S}_k.$$

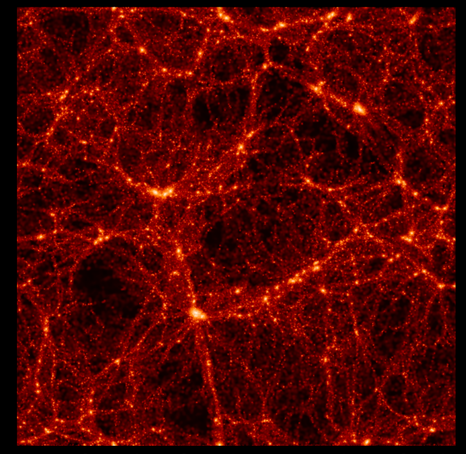
  - E.g particle interaction systems, 3d point-clouds.

(Source: S. Grunewald, credit Qi et al.)

- Input Embedding into $\mathcal{P}(\Omega)$: $(x_1, \ldots, x_k) \to \mu^{(k)} = \frac{1}{k} \sum_{j=1}^{k} \delta_{x_j}$.

[De Vie, Peyre, Cuturi]

  - Under appropriate regularity, f extended to $\overline{f} : \mathcal{P}(\Omega) \to \mathbb{R}$.
  - Input domain is not-Euclidean, infinite-dimensional.



Cosmological n-body simulations
Joerg Colberg, Virgo Simulation

- Functional neural spaces?

LEARNING UNDER SYMMETRY

- A “neuron” is now a ridge function $\varphi(\cdot, \theta) : \mathcal{P}(\Omega) \to \mathbb{R}$

$$\varphi(\mu, \theta) = a\sigma(\langle \mu, \phi \rangle),\ a \in \mathbb{R}, \boxed{\phi : \Omega \to \mathbb{R},} \langle \mu, \phi \rangle = \int_\Omega \phi(u)\mu(du).$$

  - Input “weights” ϕ are now test functions.

- Shallow invariant neural network:

$$f(\mu, \Theta) = \frac{1}{n} \sum_{i=1}^{n} a_i \varphi(\mu, \phi_i).$$

LEARNING UNDER SYMMETRY

- A “neuron” is now a ridge function $\varphi(\cdot, \theta) : \mathcal{P}(\Omega) \to \mathbb{R}$

$$\varphi(\mu, \theta) = a\sigma(\langle \mu, \phi \rangle),\ a \in \mathbb{R}, \boxed{\phi : \Omega \to \mathbb{R},} \langle \mu, \phi \rangle = \int_\Omega \phi(u)\mu(du).$$

  - Input “weights” ϕ are now test functions.

- Shallow invariant neural network:

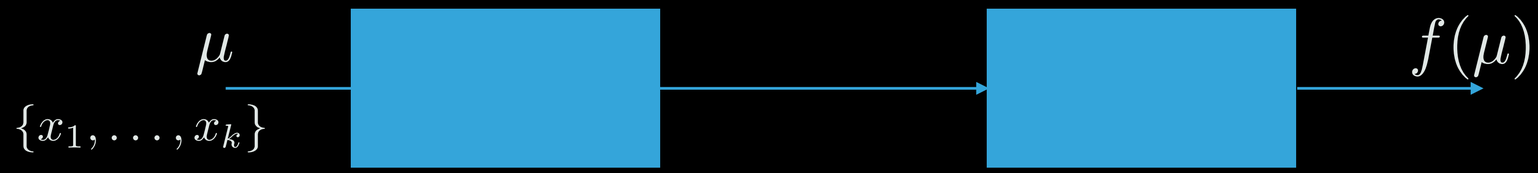
$$f(\mu, \Theta) = \frac{1}{n} \sum_{i=1}^{n} a_i \varphi(\mu, \phi_i).$$

- Integral representation:

$$f(\mu, \chi) = \int_{\mathcal{D}} \varphi(\mu, \phi)\chi(d\phi)$$

$\mathcal{D} =$ domain of test functions in Ω,
$\chi \in \mathcal{M}(\mathcal{D})$ Radon Measure over $\mathcal{D}$.

- Different over-parametrised regimes as in fully connected case?



LEARNING UNDER SYMMETRY

- Hierarchy of functional spaces for learning:

$$\mathcal{S}_1 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_1} \leq 1\}, f = \int_{\mathcal{D}} \varphi d\chi; \|\chi\|_{\mathrm{TV}} < \infty \right\}$$

$$\mathcal{S}_2 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_2} \leq 1\}, f = \int_{\mathcal{D}} \varphi d\chi; \|\chi\|_{\mathrm{TV}} < \infty \right\}$$

$$\mathcal{S}_3 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_2} \leq 1\}, f = \int_{\mathcal{D}} \varphi g(\phi) d\chi_0; \|g\|_{L^2(\mathcal{D}, d\chi_0)} < \infty \right\}$$

LEARNING UNDER SYMMETRY

- Hierarchy of functional spaces for learning:

$$\mathcal{S}_1 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_1} \le 1\}, f = \int_{\mathcal{D}} \varphi d\chi; \|\chi\|_{\mathrm{TV}} < \infty \right\}$$

$$\mathcal{S}_2 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_2} \le 1\}, f = \int_{\mathcal{D}} \varphi d\chi; \|\chi\|_{\mathrm{TV}} < \infty \right\}$$

$$\mathcal{S}_3 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_2} \le 1\}, f = \int_{\mathcal{D}} \varphi g(\phi) d\chi_0; \|g\|_{L^2(\mathcal{D}, d\chi_0)} < \infty \right\}$$

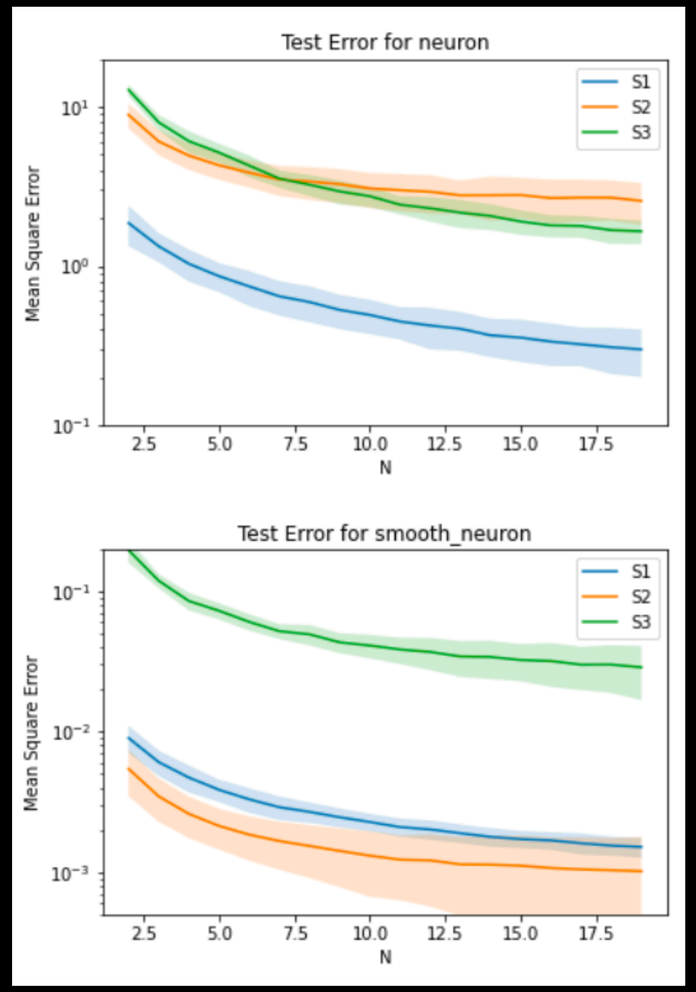
- Approximation lower bounds and generalization guarantees:

Theorem [BZ'20]: For ReLU activations, there exists f_1, f_2 with $\|f_i\|_{\mathcal{S}_i} \le 1$ such that (depth-separation)

$$\inf_{\|f\|_{\mathcal{S}_2} \le \delta} \mathrm{poly}(d) |f_1 - f|_\infty \gtrsim \left| 1 - \delta d^{-d/3} \right| , \text{ and } \inf_{\|f\|_{\mathcal{S}_3} \le \delta} \mathrm{poly}(d) \|f_2 - f\|_\infty \gtrsim \delta^{-5/d} .$$

Moreover, assuming bounded feature domain Ω, we have

$$\mathbb{E} \sup_{\|f\|_{\mathcal{S}_1} \le \delta} \left| \mathbb{E}_{\mu \sim \mathcal{D}} \ell(f^*(\mu), f(\mu)) - \frac{1}{L} \sum_{i=1}^{L} \ell(f^*(\mu_i), f(\mu_i)) \right| \lesssim \frac{\delta(1+\delta)}{\sqrt{L}} . \quad \text{(generalization bounds)}$$



LEARNING UNDER SYMMETRY

- Hierarchy of functional spaces for learning:

$$\mathcal{S}_1 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_1} \le 1\}, f = \int_{\mathcal{D}} \varphi d\chi; \|\chi\|_{\mathrm{TV}} < \infty \right\}$$

$$\mathcal{S}_2 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_2} \le 1\}, f = \int_{\mathcal{D}} \varphi d\chi; \|\chi\|_{\mathrm{TV}} < \infty \right\}$$

$$\mathcal{S}_3 = \left\{ \mathcal{D} = \{\phi; \|\phi\|_{\mathcal{F}_2} \le 1\}, f = \int_{\mathcal{D}} \varphi g(\phi) d\chi_0; \|g\|_{L^2(\mathcal{D}, d\chi_0)} < \infty \right\}$$

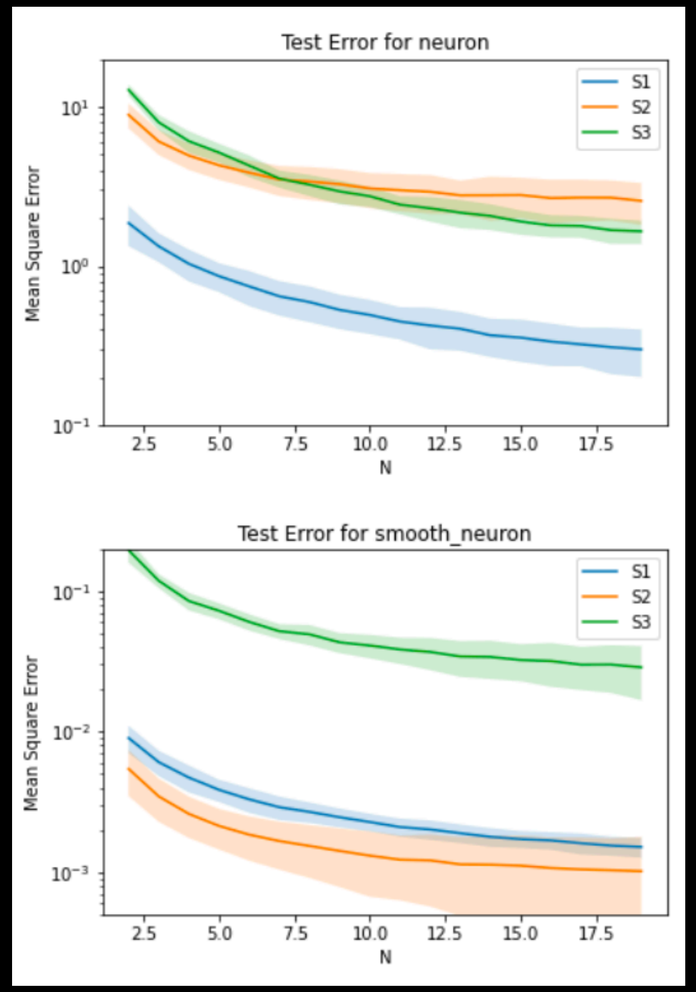
- Approximation lower bounds and generalization guarantees:

Theorem [BZ'20]: For ReLU activations, there exists f_1, f_2 with $\|f_i\|_{\mathcal{S}_i} \le 1$ such that (depth-separation)

$$\inf_{\|f\|_{\mathcal{S}_2} \le \delta} \mathsf{poly}(d) |f_1 - f|_\infty \gtrsim \left| 1 - \delta d^{-d/3} \right| , \text{ and } \inf_{\|f\|_{\mathcal{S}_3} \le \delta} \mathsf{poly}(d) \|f_2 - f\|_\infty \gtrsim \delta^{-5/d} .$$

Moreover, assuming bounded feature domain Ω, we have (generalization bounds)

$$\mathbb{E} \sup_{\|f\|_{\mathcal{S}_1} \le \delta} \left| \mathbb{E}_{\mu \sim \mathcal{D}} \ell(f^*(\mu), f(\mu)) - \frac{1}{L} \sum_{i=1}^{L} \ell(f^*(\mu_i), f(\mu_i)) \right| \lesssim \frac{\delta(1+\delta)}{\sqrt{L}} .$$



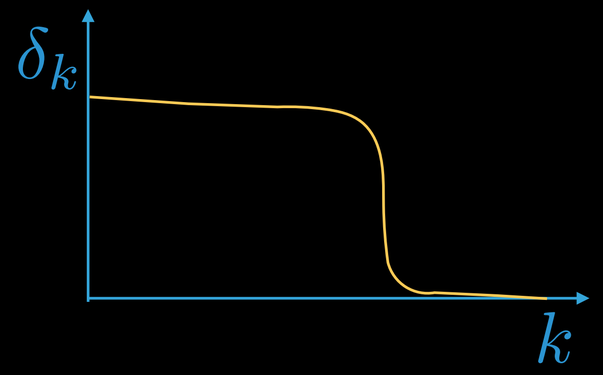
- ***Current***: extension to Fermion antisymmetry

CONCLUSIONS: APPROXIMATION VS OPTIMIZATION IN DEEP MODELS

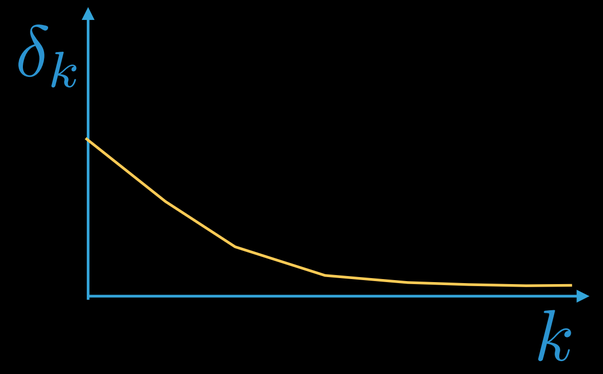
- Two necessary ingredients for learning: good approximation and efficient optimization algorithms.
- Simplified setup: $\mathcal{N}_k = \{\text{Net of depth } k \text{ and width } \text{poly}(d)\}$
  - Given target f^*, consider $\delta_k = \inf_{f \in \mathcal{N}_k} D(f, f^*)$.

CONCLUSIONS: APPROXIMATION VS OPTIMIZATION IN DEEP MODELS

- Two necessary ingredients for learning: good approximation and efficient optimization algorithms.
- Simplified setup: $\mathcal{N}_k = \{\text{Net of depth } k \text{ and width } \text{poly}(d)\}$
  - Given target f^*, consider $\delta_k = \inf_{f \in \mathcal{N}_k} D(f, f^*)$.
- ***Mild Depth Separation:*** necessary and sufficient for deep optimization?
  - Shown to be sufficient in the context of "lacunary" polynomials [Allen-Zhu & Li].
  - Shown to be necessary in the context of one-dimensional fractal distributions [Malach & Shalev-Schwartz].
- How to measure depth correlation in practice?
- ***Physical structure is key to uncover the practical benefits of depth***.
- In presence of physically-structured data (eg grids), how does depth correlation relate to Weak scale interactions (e.g. fast multipole, scattering representations)?



Strong depth separation: Shallow models provide no efficient approximation



Mild depth separation: Shallow models provide some approximation

THANKS!

References:

“Global Convergence of Neuron birth-death dynamics”, Rotskoff, Jelassi, Bruna, Vanden-Eijnden https://arxiv.org/abs/1902.01843 (ICML’19)

“A dynamical CLT for shallow Neural Networks”, Rotskoff, Chen, Bruna, Vanden-Eijnden https://arxiv.org/abs/2008.09623 (NeurIPS’20)

“Depth Separation beyond Radial Functions”, Bruna, Jelassi, Ozuch Venturi, https://arxiv.org/abs/2102.01621v2 *preprint* 2021

“On Sparsity for Overparametrised ReLU Networks”, Jaume de Dios, Bruna, https://arxiv.org/abs/2006.10225 *preprint* 2020.

“A Functional Perspective on Learning Symmetric Functions with Neural Networks”, A. Zweig, Bruna, https://arxiv.org/abs/2008.06952 *preprint* 2020.

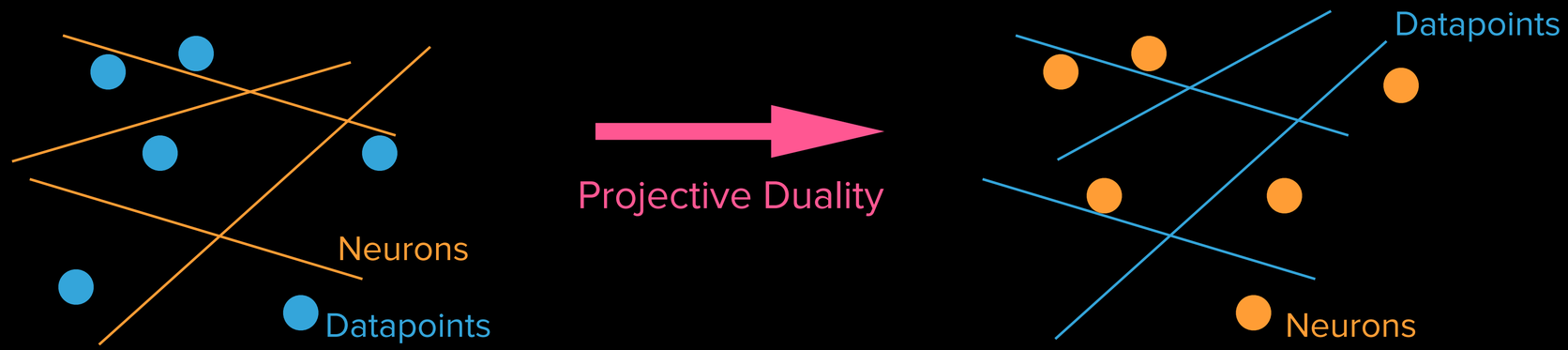
“A mean-field analysis of two-player zero-sum games”, C. Domingo-Enrich, S. Jelassi, A. Mensch, G. Rotskoff, J Bruna, https://arxiv.org/abs/2002.06277 NeurIPS’20

TOWARDS FINITE-WIDTH GUARANTEES

- The previous CLT results are still qualitative (limit of infinitely wide networks).
- For shallow ReLU networks, we can strengthen to finite-width guarantees by leveraging fine-grained ReLU structure.

> **Theorem** [DB'20]: The $\mathcal{F}_1$ regularised ERM using ReLU units only admits atomic minimisers, and the functional $\mathcal{E}[\mu]$ is locally strongly convex.

- Leveraging results from [Chizat'19] we can provide guarantees for finite width (albeit still exponential in dimension).
- ERM is reduced to a finite-dimensional linear program.



BEYOND SUPERVISED LEARNING: COMPETITIVE OPTIMIZATION

- Wasserstein-Fisher-Rao dynamics can also be used to study equilibria in games.
- Canonical setup: finding mixed strategies in two player zero-sum game:

$$\mathcal{L}[\mu_x, \mu_y] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x,y)\mu_x(dx)\mu_y(dy)\,.$$

μ_x, μ_y: players strategy distribution

$\mathcal{X}, \mathcal{Y}$: compact spaces

$\ell(x,y)$ smooth

BEYOND SUPERVISED LEARNING: COMPETITIVE OPTIMIZATION

- Wasserstein-Fisher-Rao dynamics can also be used to study equilibria in games.

- Canonical setup: finding mixed strategies in two player zero-sum game:

$$\mathcal{L}[\mu_x, \mu_y] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x,y)\mu_x(dx)\mu_y(dy)\,.$$

μ_x, μ_y: players strategy distribution

$\mathcal{X}, \mathcal{Y}$: compact spaces

$\ell(x,y)$ smooth

- (mixed) Nash Equilibria: (μ_x^*, μ_y^*) such that

$$\forall \mu_x \;,\; \mathcal{L}[\mu_x^*, \mu_y^*] \leq \mathcal{L}[\mu_x, \mu_y^*] \;,\quad \forall \mu_y \;,\; \mathcal{L}[\mu_x^*, \mu_y^*] \geq \mathcal{L}[\mu_x, \mu_y]\;.$$

  - Guaranteed to exist [Nash'50s]
  - Algorithms to find them in the high-dimensional setting?

BEYOND SUPERVISED LEARNING: COMPETITIVE OPTIMIZATION

- Wasserstein-Fisher-Rao dynamics can also be used to study equilibria in games.

- Canonical setup: finding mixed strategies in two player zero-sum game:

$$\mathcal{L}[\mu_x, \mu_y] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(x,y)\mu_x(dx)\mu_y(dy)\,.$$

μ_x, μ_y: players strategy distribution

$\mathcal{X}, \mathcal{Y}$: compact spaces

$\ell(x,y)$ smooth

- (mixed) Nash Equilibria: (μ_x^*, μ_y^*) such that

$$\forall \mu_x \;,\; \mathcal{L}[\mu_x^*, \mu_y^*] \leq \mathcal{L}[\mu_x, \mu_y^*] \;, \quad \forall \mu_y \;,\; \mathcal{L}[\mu_x^*, \mu_y^*] \geq \mathcal{L}[\mu_x, \mu_y]\;.$$

- Gradient dynamics:

$$\partial_t \mu_{x,t} = \mathrm{div}(\nabla \frac{\partial \mathcal{L}}{\partial \mu_x}) \qquad \partial_t \mu_{y,t} = -\mathrm{div}(\nabla \frac{\partial \mathcal{L}}{\partial \mu_y})$$

BEYOND SUPERVISED LEARNING: COMPETITIVE OPTIMIZATION

- Measure dynamics associated with particle gradient ascent/ descent:

$$\partial_t \mu_{x,t} = \mathrm{div}(\nabla \frac{\partial \mathcal{L}}{\partial \mu_x}) \qquad \partial_t \mu_{y,t} = -\mathrm{div}(\nabla \frac{\partial \mathcal{L}}{\partial \mu_y})$$

- We establish Global convergence to approximate Nash equilibria using WFR.

- Similar propagation-of-chaos and robustness in high-dimensions.