NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU CENTER FOR DATA SCIENCE

MAD Math and Data

JOAN BRUNA

MEASURE DYNAMICS FOR NEURAL NETWORKS

joint work with

Zhengdao Chen

Jaume de Dios

Carles Domingo

Samy Jelassi

Arthur Mensch

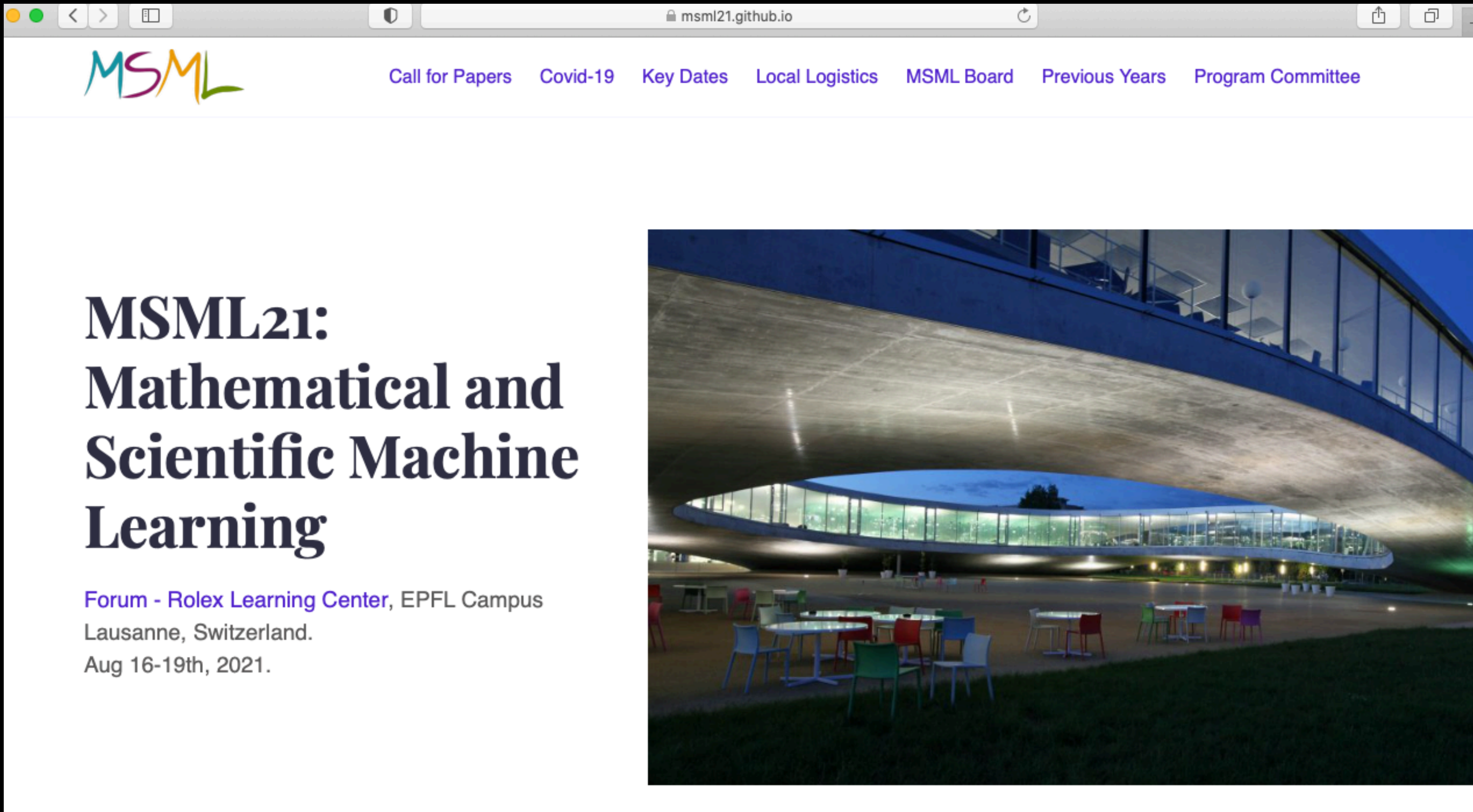
Grant Rotskoff

E.Vanden-Eijnden

Luca Venturi

Aaron Zweig

MATHEMATICAL AND SCIENTIFIC MACHINE LEARNING



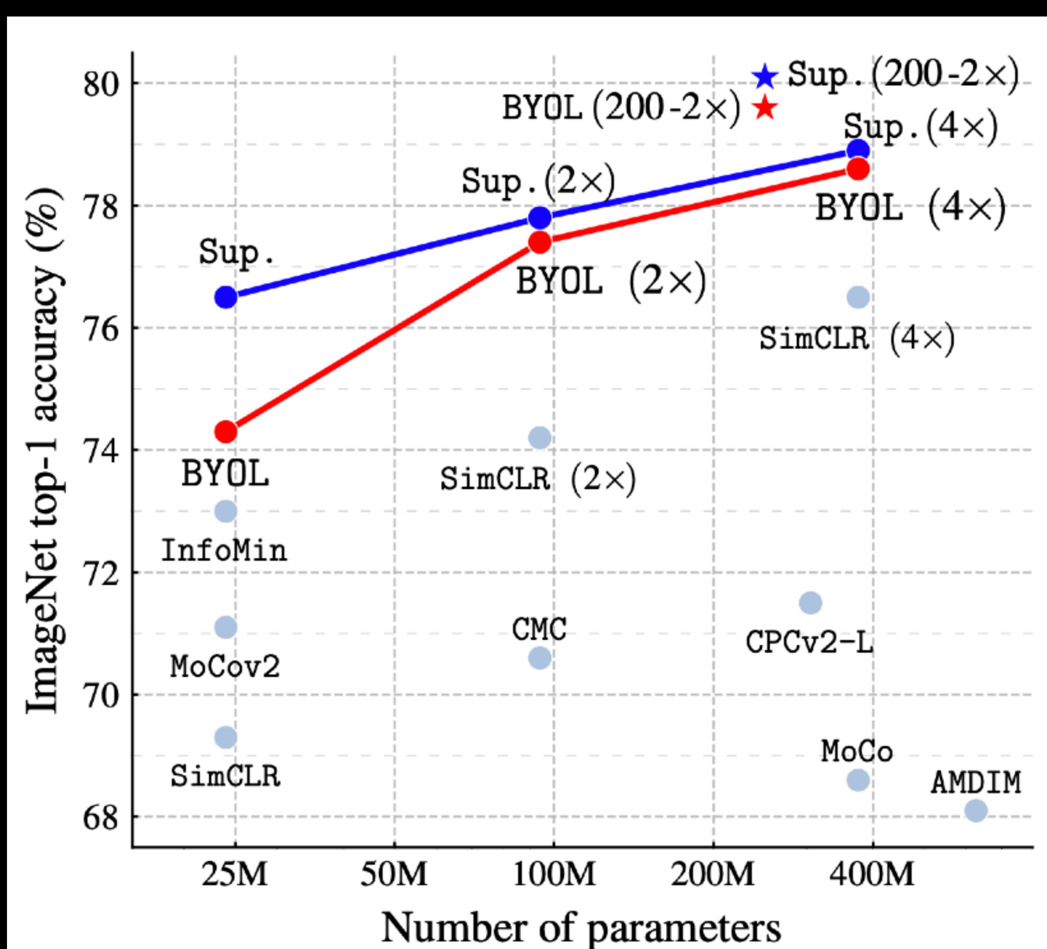
- Deadline for paper submissions: dec 4th
- General Chairs: Joan Bruna, Jan Hesthaven, Lenka Zdeborova

DEEP LEARNING TODAY: EXPERIMENTAL REVOLUTION

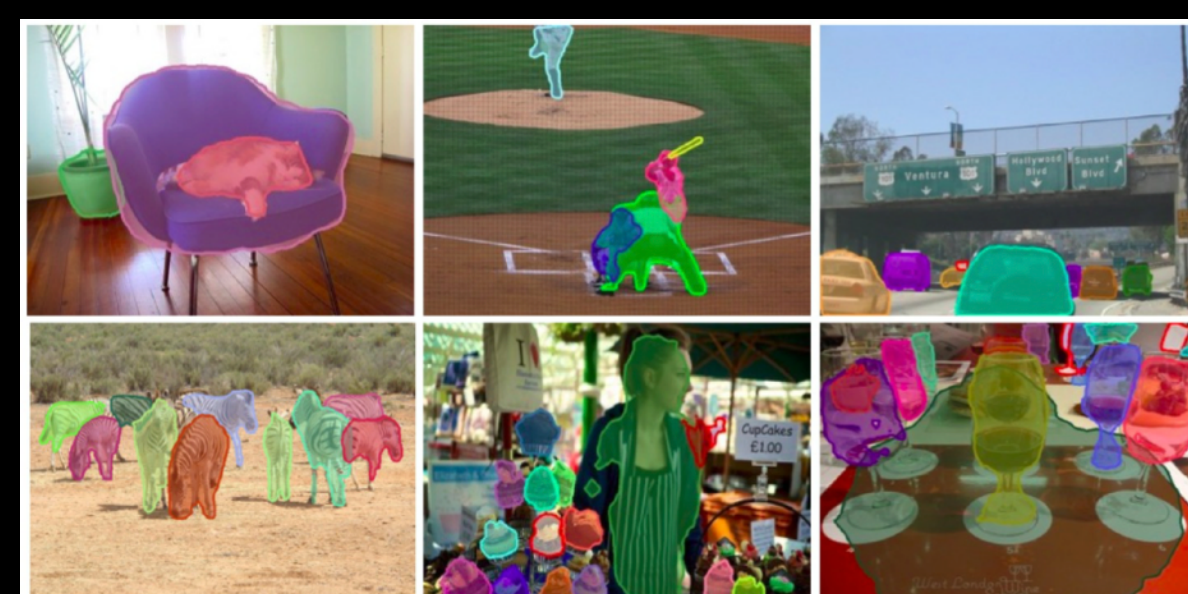
[v.d. Oord et al.'19]

Gatys et al'14

[Grill et al'20]



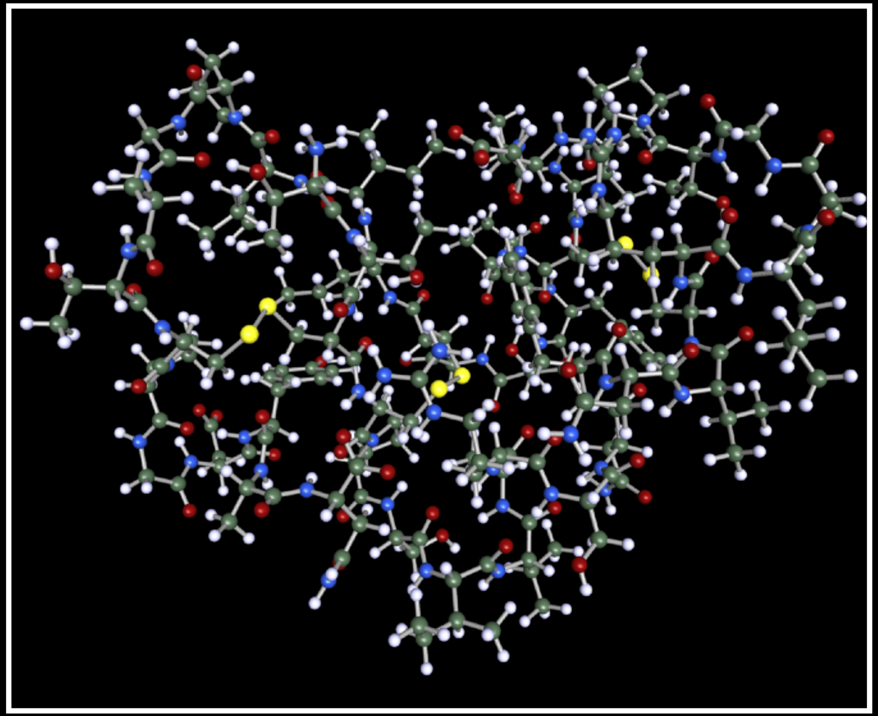
[He et al.'17]



DEEP LEARNING TODAY: EXPERIMENTAL REVOLUTION

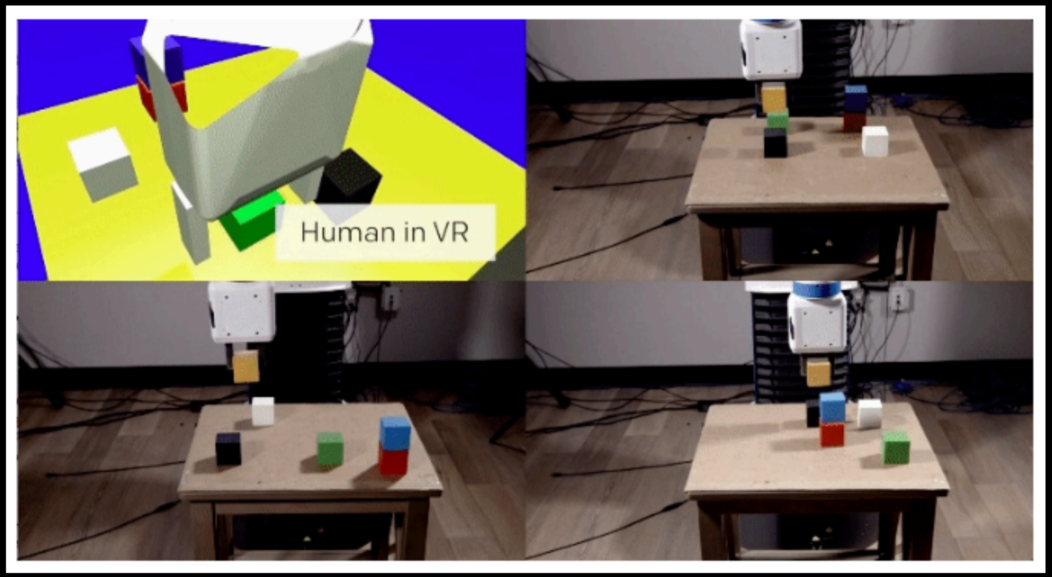
Computational Biology

Quantum Chemistry

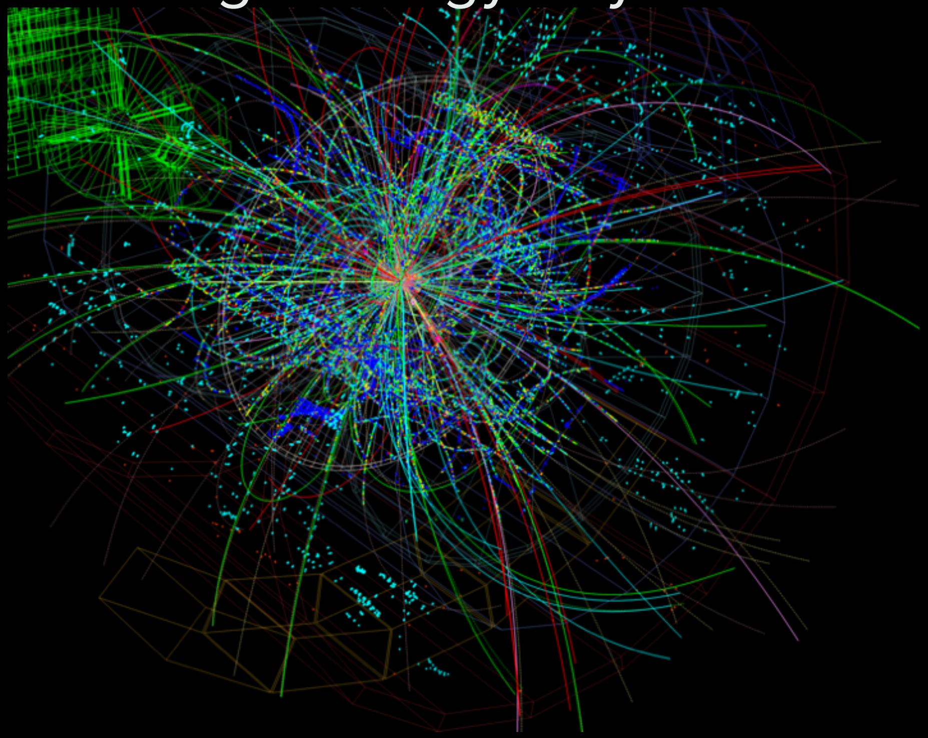


Games

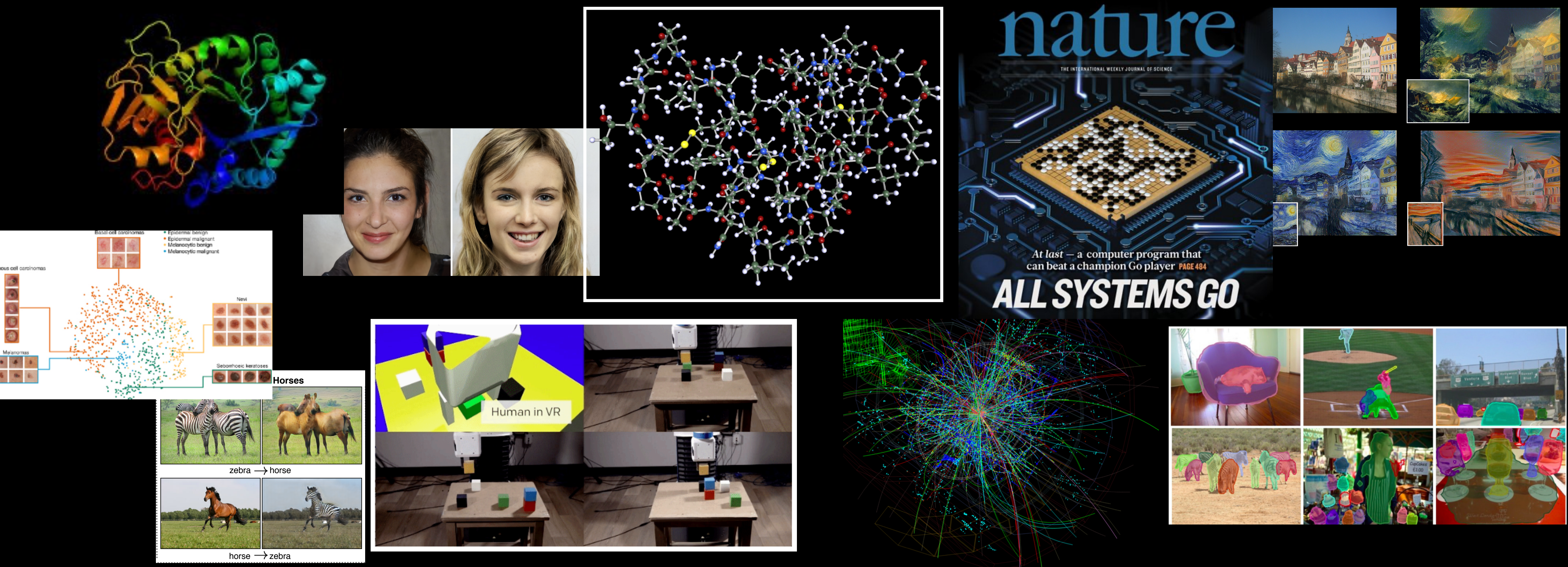
Robotics



High Energy Physics



DEEP LEARNING TODAY: EXPERIMENTAL REVOLUTION



- Phenomenal capacity to extract information from complex high-dimensional observations.
- In essence: non-linear, compositional ***feature learning.***
- "Right" balance between model-based and data-based estimation, using simple algorithmic principle (1st order optim).

SUPERVISED LEARNING SETUP

- Data : $\{(x_i, y_i)\} \sim \nu \in \mathcal{M}(\mathbb{R}^m \times \mathbb{R})$.
  - Noise-free setting: $y_i = f^*(x_i) \text{ for some } f^* \in L^2(\mathbb{R}^m, \mathrm{d}\nu)$.
- Model: $f(x; \Theta)$, $\Theta \in \mathcal{D}$. $\mathcal{F} := \{f(\cdot, \Theta); \Theta \in \mathcal{D}\}$.

SUPERVISED LEARNING SETUP

- Data : $\{(x_i, y_i)\} \sim \nu \in \mathcal{M}(\mathbb{R}^m \times \mathbb{R})$.
  - Noise-free setting: $y_i = f^*(x_i)$ for some $f^* \in L^2(\mathbb{R}^m, \mathrm{d}\nu)$.
- Model: $f(x;\Theta)$, $\Theta \in \mathcal{D}$. $\mathcal{F} := \{f(\cdot,\Theta); \Theta \in \mathcal{D}\}$.
- Loss: $\mathcal{R}(f)$ convex, e.g.

$$\mathcal{R}(f) = \int |f(x) - f^*(x)|^2 \mathrm{d}\nu(x) \ . \quad f \in \mathcal{F}.$$

- Empirical loss:

$$\widehat{\mathcal{R}}(f) = \int |f(x) - f^*(x)|^2 \mathrm{d}\hat{\nu}(x) = \frac{1}{L}\sum_{l=1}^{L} |f(x_l) - f^*(x_l)|^2 \ .$$

SUPERVISED LEARNING SETUP

- Empirical Risk Minimisation: $\mathcal{F}_\delta = \{f \in \mathcal{F}; \|f\| \leq \delta\}.$

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

SUPERVISED LEARNING SETUP

- Empirical Risk Minimisation: $\mathcal{F}_\delta = \{f \in \mathcal{F}; \|f\| \leq \delta\}$.

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

- Basic decomposition of error: [Bottou & Bousquet]

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

SUPERVISED LEARNING SETUP

- Empirical Risk Minimisation: $\mathcal{F}_\delta = \{ f \in \mathcal{F}; \|f\| \leq \delta \}.$

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

- Basic decomposition of error: [Bottou & Bousquet]

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

- Main challenges in High-dimensional ML:
  - Approximation: Functional Approximation that is not cursed by input dimensionality.
  - Statistical: Statistical Error handled with uniform concentration bounds.
  - Computational: How to solve (*) efficiently in the high-dimensional regime?

THE CURSE OF DIMENSIONALITY

- “Classic” functional spaces do not play well with this tradeoff.

THE CURSE OF DIMENSIONALITY

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

THE CURSE OF DIMENSIONALITY

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

THE CURSE OF DIMENSIONALITY

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

THE CURSE OF DIMENSIONALITY

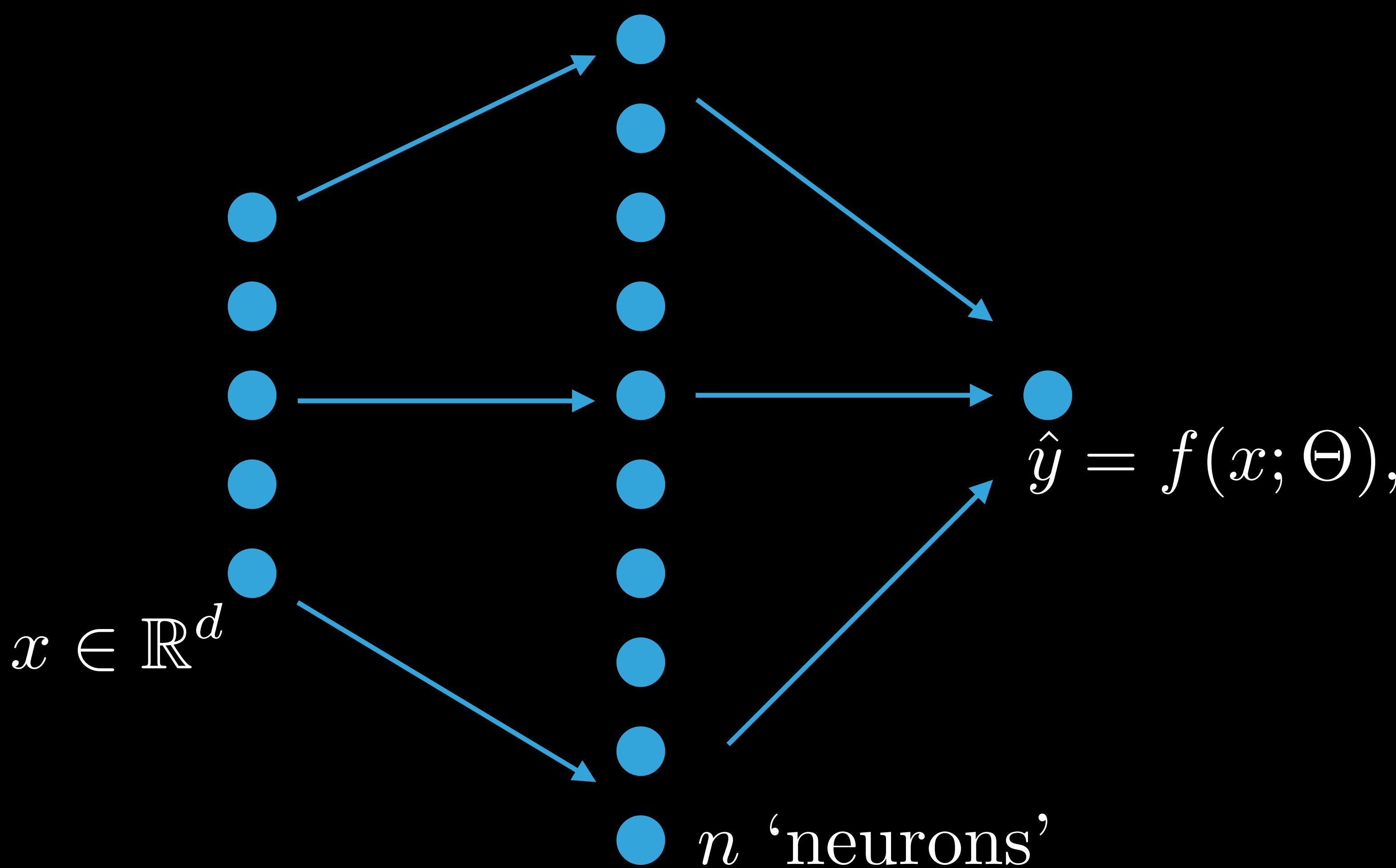
- “Classic” functional spaces do not play well with this tradeoff.
  - $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \text{ is Lipschitz}\}$ is too big: the number of samples required to identify $f^* \in \mathcal{F}$ up to error ϵ is $\Omega(\epsilon^{-m})$ [von Luxburg & Bousquet].
  - $\mathcal{F} = \mathcal{H}^{s,p}$: Sobolev spaces. Minimax rate of approximation is cursed unless $s \geq d/2$: only very smooth functions are allowed.
- Which functions can be provably learnt in the high-dimensional regime?
- … with neural networks (and using gradient descent)?
- … with deep neural networks?
- … with deep structured neural networks?

THIS TALK

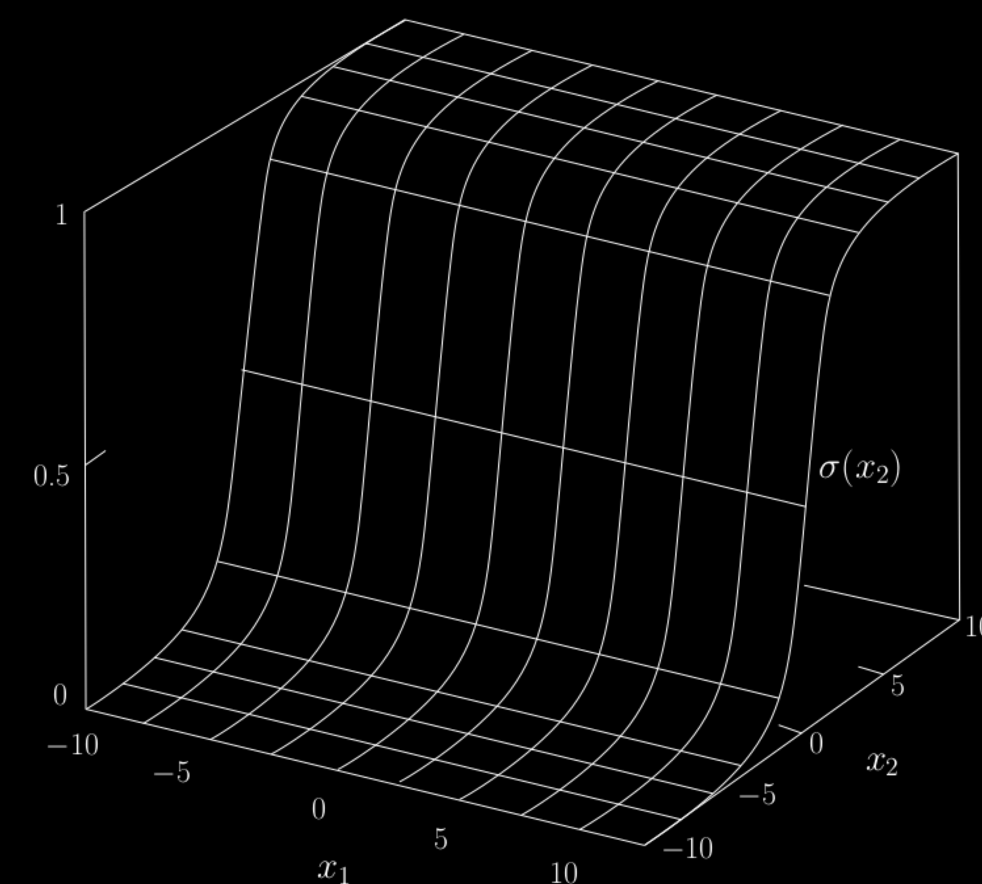
- Simplest instance of nonlinear feature learning: shallow NNs.
  - Gradient-descent Optimization analyzed as measure dynamics. Retains non-linear essence with Mean-field global convergence guarantees.
  - Towards Finite-width guarantees by CLT and fine-grained analysis of ReLU activations.
- Beyond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

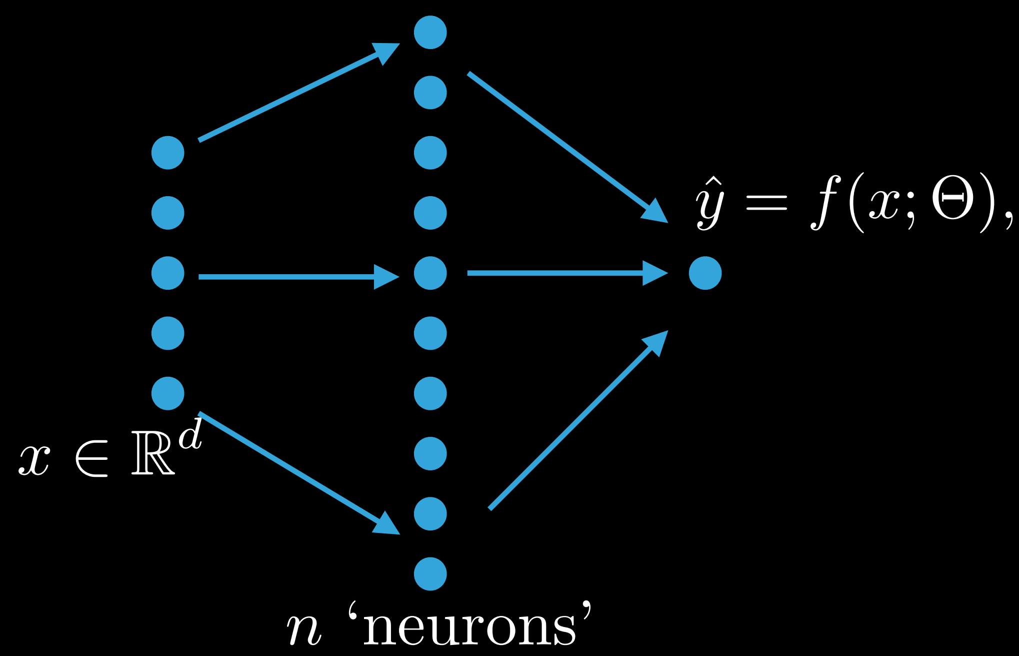


$$\tilde{\varphi}(x;\theta) = a\varphi(x;z),$$
$$\varphi(x;z) = \sigma(\langle x, w\rangle + b),$$
$$\theta = \{a, z\} \in \mathbb{R} \times \mathcal{D}.$$



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

SINGLE HIDDEN-LAYER NEURAL NETWORK



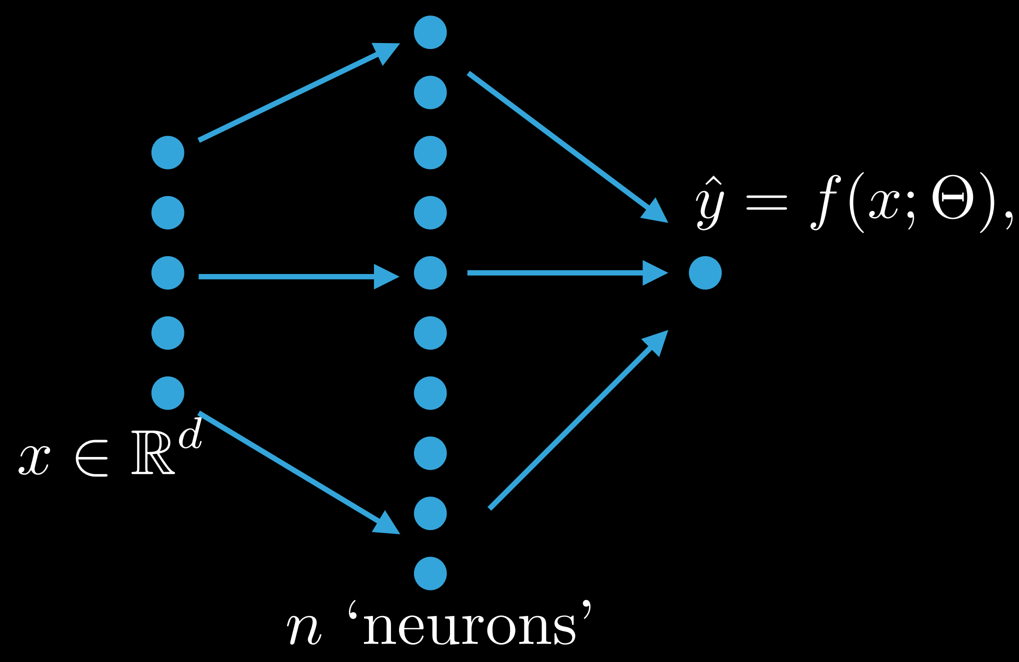
$$f(x;\Theta) = \sum_{j \leq n} \tilde{\varphi}(x; \theta_j)$$

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

- 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



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

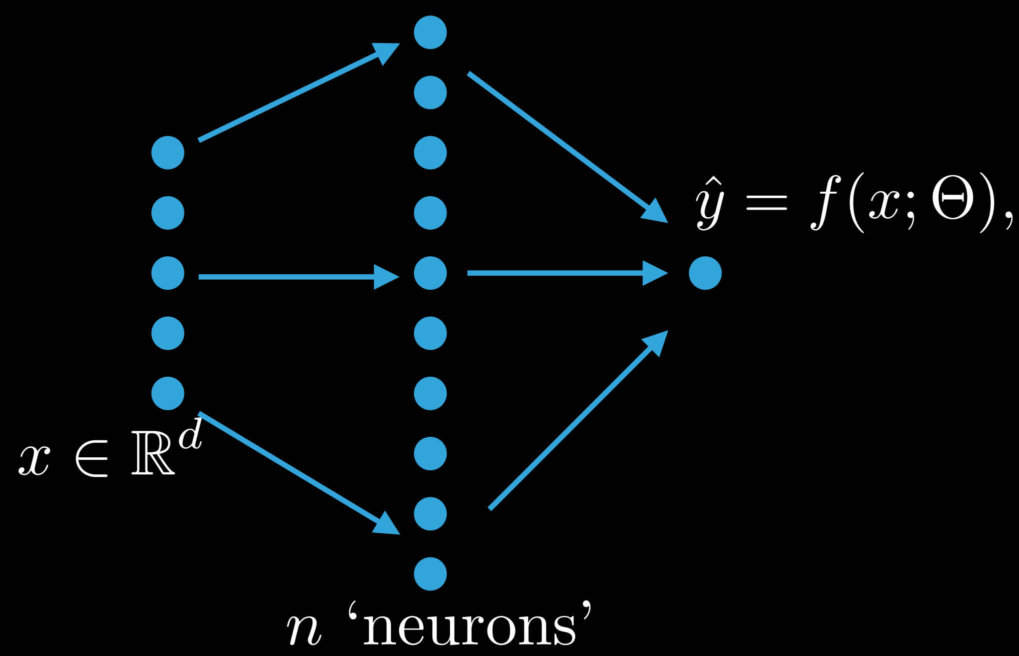
$$\tilde{\varphi}(x;\theta) = a\varphi(x;z),$$
$$\varphi(x;z) = \sigma(\langle x, w\rangle + b),$$
$$\theta = \{a, z\} \in \mathbb{R} \times \mathcal{D}.$$

- 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 Approx: 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



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

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

- 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 Approx: 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

$$k(x, x') = \int \varphi(x,z)\varphi(x',z)\mu_0(dz)$$

[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

$$k(x, x') = \int \varphi(x,z)\varphi(x',z)\mu_0(dz)$$

[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].

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.

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.
- How to perform optimization and approximation in these spaces?

NEURAL NETWORKS AS PARTICLE INTERACTION SYSTEMS

- No noise on targets: $f^* \in L_2(\mathbb{R}^d, d\nu)$: target function.
- Single-hidden layer architecture

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

NEURAL NETWORKS AS PARTICLE INTERACTION SYSTEMS

- No noise on targets: $f^* \in L_2(\mathbb{R}^d, d\nu)$: target function.
- Single-hidden layer architecture

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

- With Square loss, $\mathcal{F}_1$-penalized ERM becomes

$$\begin{aligned} \mathcal{E}(\Theta) &= \mathbb{E}_{\hat{\nu}}[|f(x;\Theta) - f^*|^2] + \lambda \mathcal{V}(\Theta) \\ &= C - \frac{2}{n} \sum_{j \leq n} F(\theta_j) + \frac{1}{n^2} \sum_{j,j'} U(\theta_j, \theta_{j'}) \end{aligned}$$

$$\mathcal{V}(\Theta) = \sum_{j \leq n} |a_j|^q \ (q \geq 1).$$

$$F(\theta) = a\mathbb{E}_{\hat{\nu}}[f^*(x)\varphi(x,\theta)] - \lambda |a|^2 \ , \ U(\theta, \theta') = aa'\mathbb{E}_{\hat{\nu}}[\varphi(x,z)\varphi(x,z')] \ .$$

- Scaling in $1/n$ contrasts with $1/\sqrt{n}$, which leads to ***lazy*** or ***NTK*** regime [Chizat et al., Jacot et al., Arora et al, etc].

EULERIAN PERSPECTIVE

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

- Taking step-size of gradient-descent to zero, we have a gradient flow in parameter space:

$$\dot{\theta}_i = -\nabla_{\theta_i}\mathcal{E}(\theta_1, \ldots, \theta_n)\,,\; i = 1 \ldots n.$$

  - Non-convex functional, generically hard [Shamir et al., Venturi et al]

EULERIAN PERSPECTIVE

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

- Taking step-size of gradient-descent to zero, we have a gradient flow in parameter space:

$$\dot{\theta}_i = -\nabla_{\theta_i} \mathcal{E}(\theta_1, \ldots, \theta_n)\,,\; i = 1 \ldots n.$$

  - Non-convex functional, generically hard [Shamir et al., Venturi et al]

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

$$\mu_n(t, \theta) = \frac{1}{n} \sum_{j \leq n} \delta_{\theta_j(t)}$$

- The regularised loss becomes

$$\mathcal{E}(\mu) = -2 \int F(\theta) \mu(d\theta) + \iint U(\theta, \theta') \mu(d\theta) \mu(d\theta')\,.$$

  - quadratic since we consider mean-squared loss.

EULERIAN PERSPECTIVE

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

- Taking step-size of gradient-descent to zero, we have a gradient flow in parameter space:

$$\dot{\theta}_i = -\nabla_{\theta_i} \mathcal{E}(\theta_1, \ldots, \theta_n)\,,\ i = 1 \ldots n.$$

  - Non-convex functional, generically hard [Shamir et al., Venturi et al]

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

$$\mu_n(t, \theta) = \frac{1}{n} \sum_{j \le n} \delta_{\theta_j(t)}$$

- The regularised loss becomes

$$\mathcal{E}(\mu) = -2 \int F(\theta) \mu(d\theta) + \iint U(\theta, \theta') \mu(d\theta) \mu(d\theta')\,.$$

  - quadratic since we consider mean-squared loss.

- Dynamics in the space of measures?

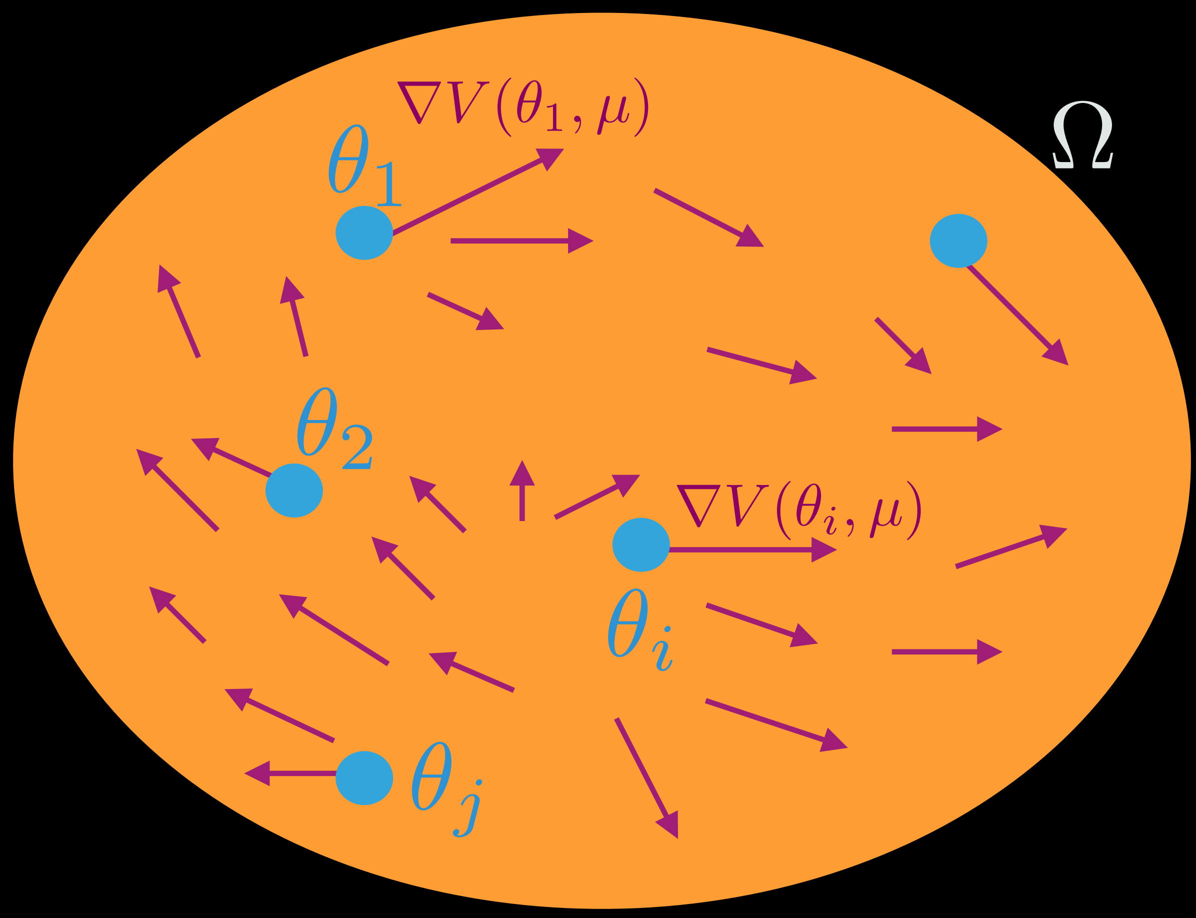
CONTINUITY EQUATION

[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:

$$\frac{n}{2}\nabla_{\theta_i}\mathcal{E}(\Theta) = \nabla V|_{\theta=\theta_i} \text{ , with}$$

$$V(\theta;\mu) = -F(\theta) + \int U(\theta,\theta')\mu(d\theta')\,.$$



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:
$$\frac{n}{2}\nabla_{\theta_i}\mathcal{E}(\Theta) = \nabla V|_{\theta=\theta_i} \text{ , with}$$
$$V(\theta;\mu) = -F(\theta) + \int U(\theta,\theta')\mu(d\theta') \,.$$
- For general time-dependent measures μ_t, their evolution under a time-varying velocity field $V(\theta;\mu_t)$ is given by a ***continuity equation***:
$$\partial_t \mu_t = \mathrm{div}(\mu_t \nabla V) \;,\; \mu(0) = \mu^{(0)} \;, \text{ with}$$
$$\forall\, \phi \in C_c^\infty(\Omega)\,,\, \partial_t\left(\int \phi \mu_t(d\theta)\right) = -\int \langle \nabla\phi, \nabla V\rangle \mu_t(d\theta)\,.$$
  - Gradient flow of $\mathcal{E}$ for the Wasserstein metric W_2 in $\mathcal{P}(\Omega)$
  - ***Exact description*** of particle gradient for atomic measures.

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

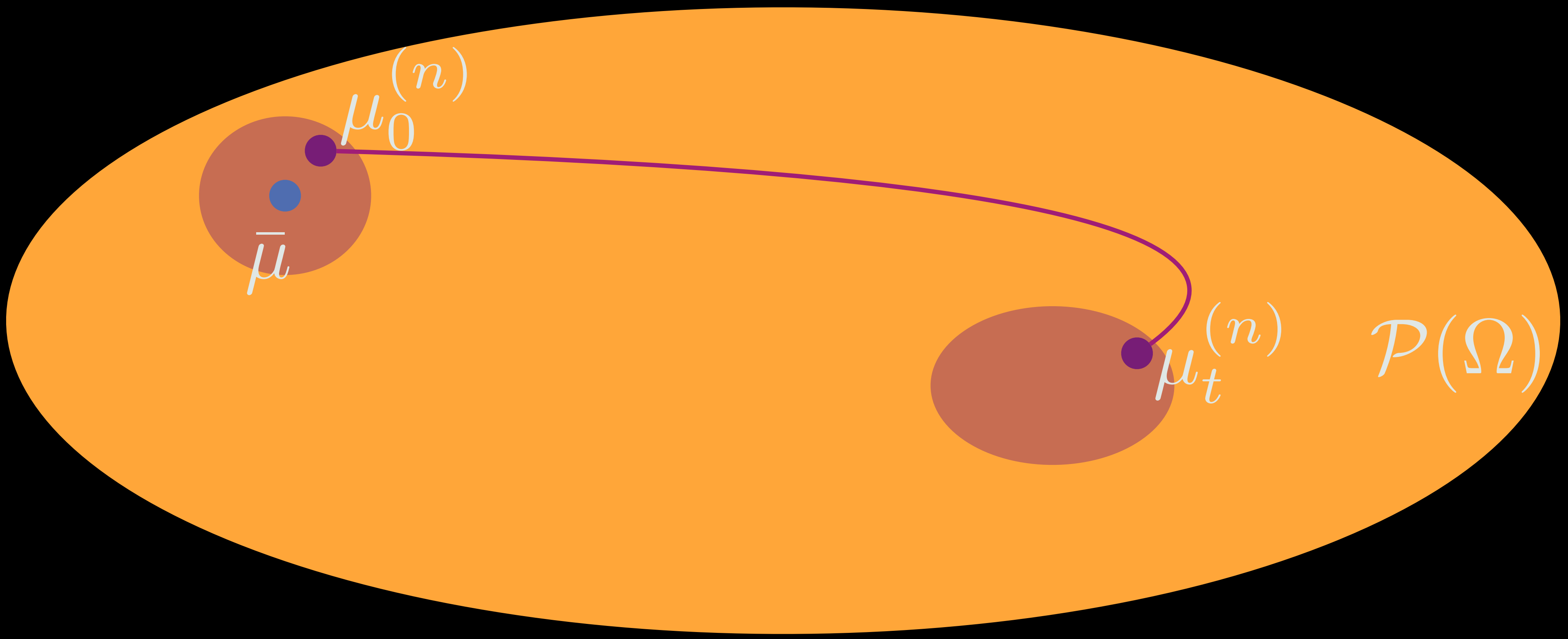
CONTINUITY EQUATION

- Particle gradients correspond to evaluating a scaled velocity field:
$$\frac{n}{2}\nabla_{\theta_i}\mathcal{E}(\Theta) = \nabla V|_{\theta=\theta_i} \text{ , with}$$
$$V(\theta;\mu) = -F(\theta) + \int U(\theta,\theta')\mu(d\theta')\,.$$
- For general time-dependent measures μ_t, their evolution under a time-varying velocity field $V(\theta;\mu_t)$ is given by a ***continuity equation***:
$$\partial_t \mu_t = \mathrm{div}(\mu_t \nabla V)\ ,\ \mu(0) = \mu^{(0)}\ , \text{ with}$$
$$\forall\, \phi \in C_c^\infty(\Omega)\,,\, \partial_t\left(\int \phi\mu_t(d\theta)\right) = -\int \langle \nabla\phi, \nabla V\rangle \mu_t(d\theta)\,.$$
  - Gradient flow of $\mathcal{E}$ for the Wasserstein metric W_2 in $\mathcal{P}(\Omega)$
  - ***Exact description*** of particle gradient for atomic measures.

LAGRANGIAN	⟷	**EULERIAN**
Non-Convexity		Convexity
Euclidean Dynamics		Non-Euclidean Dynamics

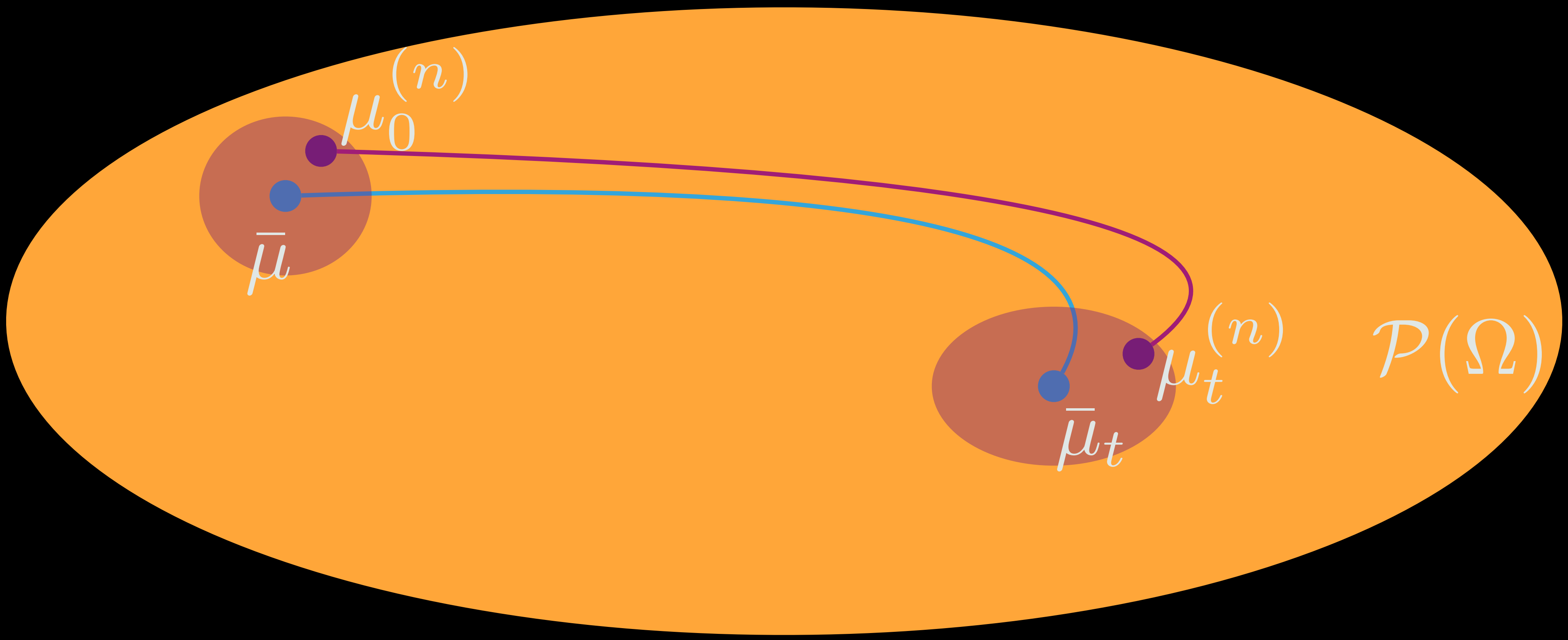
MEAN-FIELD LIMIT

- 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.



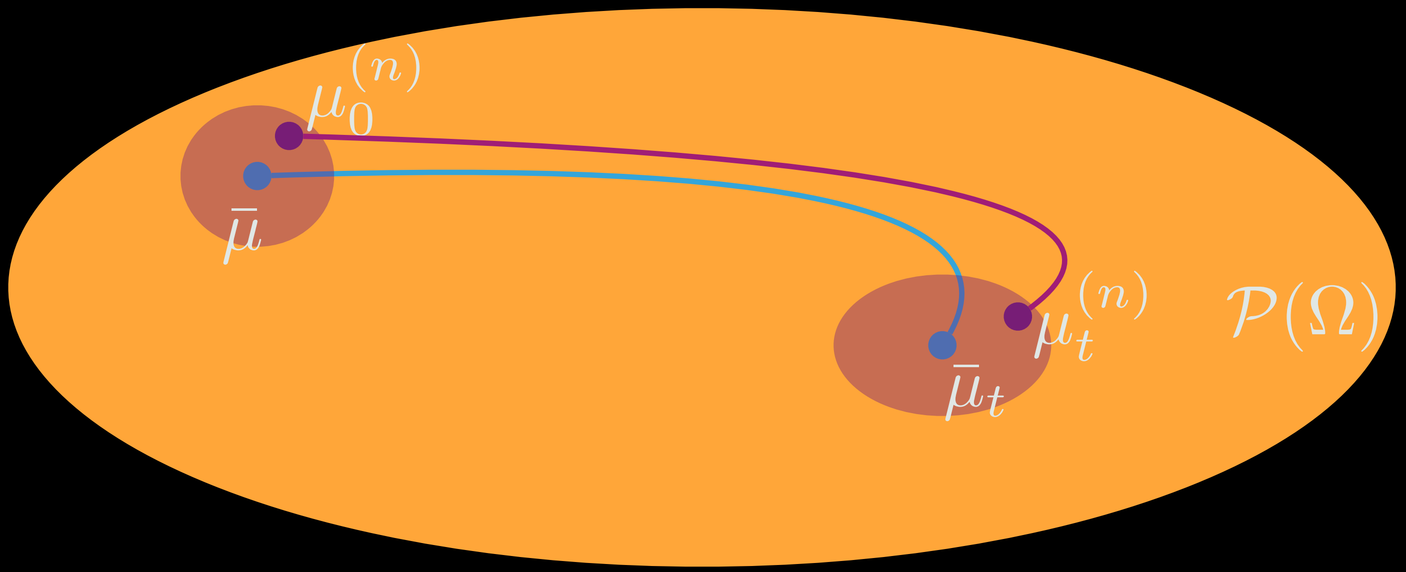
MEAN-FIELD LIMIT

- 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

- 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.

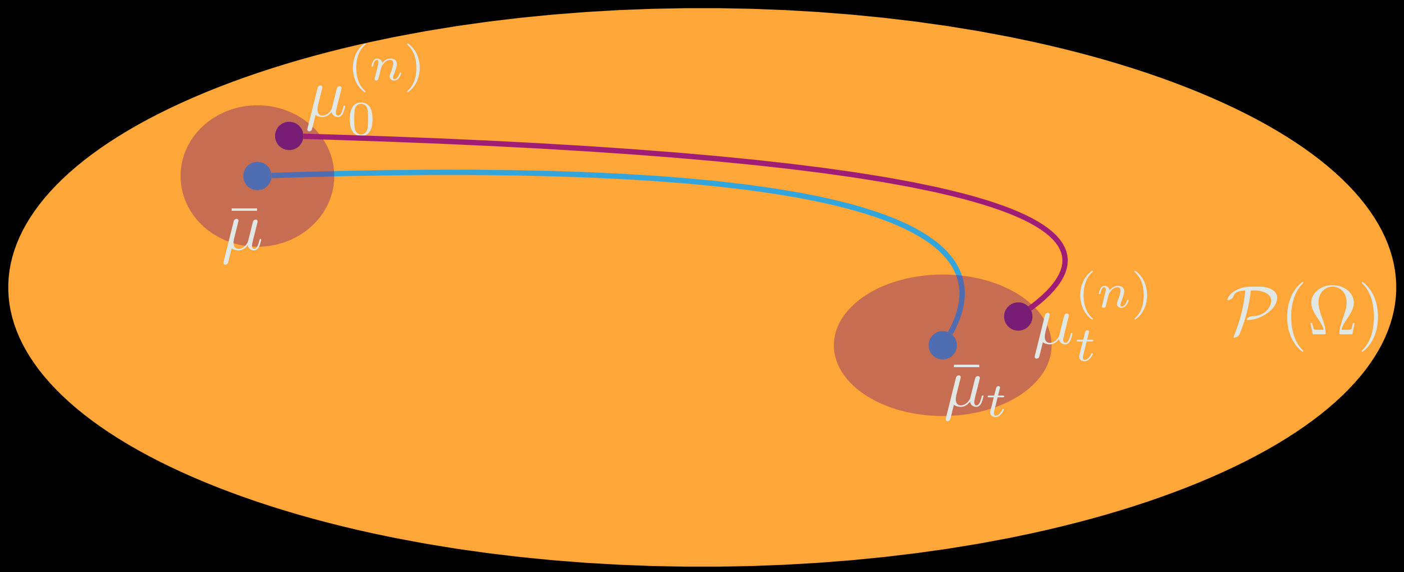


Theorem: [R,EVE,'18],[CB'18],[MMN'18],[SS'18]
For any fixed $t > 0$, $\mu_t^{(n)}$ converges weakly to μ_t as $n \to \infty$, which solves $\partial_t \mu_t = \mathrm{div}(\nabla V \mu_t)$ with $\mu_0 = \bar{\mu}$.

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

MEAN-FIELD LIMIT

- 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).
- Convergence properties of this PDE?
- LLN result. What is the scale of the fluctuations?

UNBALANCED TRANSPORT

- Inspired from [Wei et al.'18], we consider the following unbalanced modification of the dynamics:

$$\partial_t \mu_t = \mathrm{div}(\mu_t \nabla V) - \alpha V \mu_t + \alpha \bar{V} \mu_t \,, \quad \text{with}$$

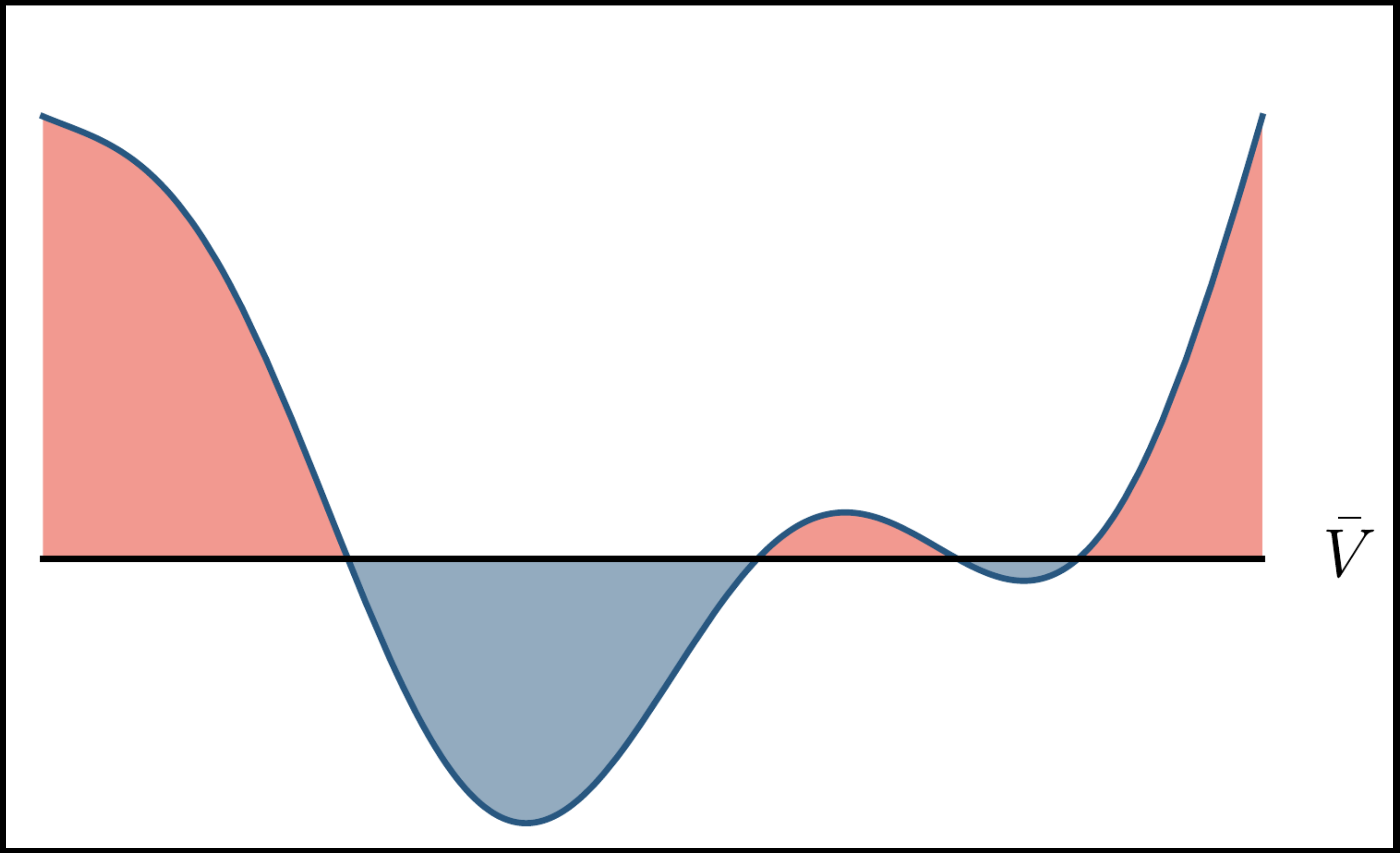
$$\alpha > 0 \,,\ \overline{V} := \int V(\theta)\mu(d\theta)\,.$$

UNBALANCED TRANSPORT

- Inspired from [Wei et al.'18], we consider the following unbalanced modification of the dynamics:

$$\partial_t \mu_t = \mathrm{div}(\mu_t \nabla V) - \alpha V \mu_t + \alpha \bar{V} \mu_t \,, \quad \text{with}$$

$$\alpha > 0 \,, \; \overline{V} := \int V(\theta)\mu(d\theta) \,.$$



UNBALANCED TRANSPORT

- Inspired from [Wei et al.'18], we consider the following unbalanced modification of the dynamics:

$$\partial_t \mu_t = \mathrm{div}(\mu_t \nabla V) - \alpha V \mu_t + \alpha \bar{V} \mu_t \ , \ \text{with}$$

$$\alpha > 0 \ , \ \overline{V} := \int V(\theta)\mu(d\theta) \, .$$

- For all μ, we verify that $\int V(\theta)\mu(d\theta) - \int \bar{V}\mu(d\theta) = 0$
  - Mass is preserved. In particular, for atomic measures, population is constant.

- Full PDE corresponds to gradient flow for the Wasserstein-Fisher-Rao metric [Kondratiev et al.], [Chizat et al.] (aka Hellinger-Kantorovich).
- Admits easy discretization using birth/death processes.
- Wasserstein-Fisher-Rao dynamics can also be used to study equilibria in zero-sum two-player games [D-E, J R, M,**B**'20].

GLOBAL CONVERGENCE

- Interaction kernel $U(\theta, \theta')$ symmetric and positive semi-definite, twice differentiable.
- $U(\theta, \theta')$ and $F(\theta)$ such that energy $\mathcal{E}[\mu]$ is bounded below.
- The only fixed points of the dynamics are global minimizers of the energy:

Theorem: [RJBV'19] Let μ_t denote the solution of the dynamics for initial condition μ_0 with full support. Then, if $\mu_t \to \mu_*$ in the weak sense, then μ_* is a global minimiser of $\mathcal{E}[\mu]$. Also, $\exists C, t_c > 0$ such that $\mathcal{E}[\mu_t] \leq \mathcal{E}[\mu_*] + Ct^{-1}$ if $t \geq t_c$.

GLOBAL CONVERGENCE

- Interaction kernel $U(\theta, \theta')$ symmetric and positive semi-definite, twice differentiable.
- $U(\theta, \theta')$ and $F(\theta)$ such that energy $\mathcal{E}[\mu]$ is bounded below.
- The only fixed points of the dynamics are global minimizers of the energy:

Theorem: [RJBV'19] Let μ_t denote the solution of the dynamics for initial condition μ_0 with full support. Then, if $\mu_t \to \mu_*$ in the weak sense, then μ_* is a global minimiser of $\mathcal{E}[\mu]$. Also, $\exists C, t_c > 0$ such that $\mathcal{E}[\mu_t] \leq \mathcal{E}[\mu_*] + Ct^{-1}$ if $t \geq t_c$.

  - We avoid the fixed points of the Liouville PDE which are not minimizers of the energy $\nabla V(\theta) = 0$ for $\theta \in \mathrm{supp}(\mu_*)$.
  - Extends results from [Chizat & Bach] beyond homogeneous models.
  - How to leverage this mean-field guarantee for finite data/units?

APPROXIMATION AND GENERALIZATION IN VARIATION-NORM

- Minimisers of $\mathcal{E}[\mu]$ can be efficiently discretized if $f^* \in \mathcal{F}_1$:

Proposition [RCBE'19]: Let $\mu^* \in \mathcal{M}_+(\mathbb{R} \times \mathcal{D})$ be a minimiser of $\mathcal{E}$. Then $\int U(\theta, \theta)\mu^*(d\theta) \leq C\|f^*\|_1^2$.

- Monte-Carlo approximation bounds $\|f_{n,t} - f_t\|_\nu^2 \leq \dfrac{C\|f^*\|_1^2}{n}$

APPROXIMATION AND GENERALIZATION IN VARIATION-NORM

- Minimisers of $\mathcal{E}[\mu]$ can be efficiently discretized if $f^* \in \mathcal{F}_1$:

Proposition [RCBE'19]: Let $\mu^* \in \mathcal{M}_+(\mathbb{R} \times \mathcal{D})$ be a minimiser of $\mathcal{E}$. Then $\int U(\theta, \theta)\mu^*(d\theta) \leq C\|f^*\|_1^2$.

  - Monte-Carlo approximation bounds $\|f_{n,t} - f_t\|_\nu^2 \leq \frac{C\|f^*\|_1^2}{n}$

- Generalisation bound: Let μ_L^* be a minimiser of the empirical (regularised) loss, and $\hat{f}_L = \int a\varphi(z)\mu_L^*(da, dz)$.

Theorem [RCBE'19]: Then
$$\mathbb{E}\|\hat{f}_L - f^*\|_\nu^2 \leq 2\|f^*\|_1 \left(\frac{R_1\|f^*\|_1 + R_2}{\sqrt{L}} + \lambda \right)$$

  - Based on Rademacher bounds for $\mathcal{F}_1$ [Bach'17]

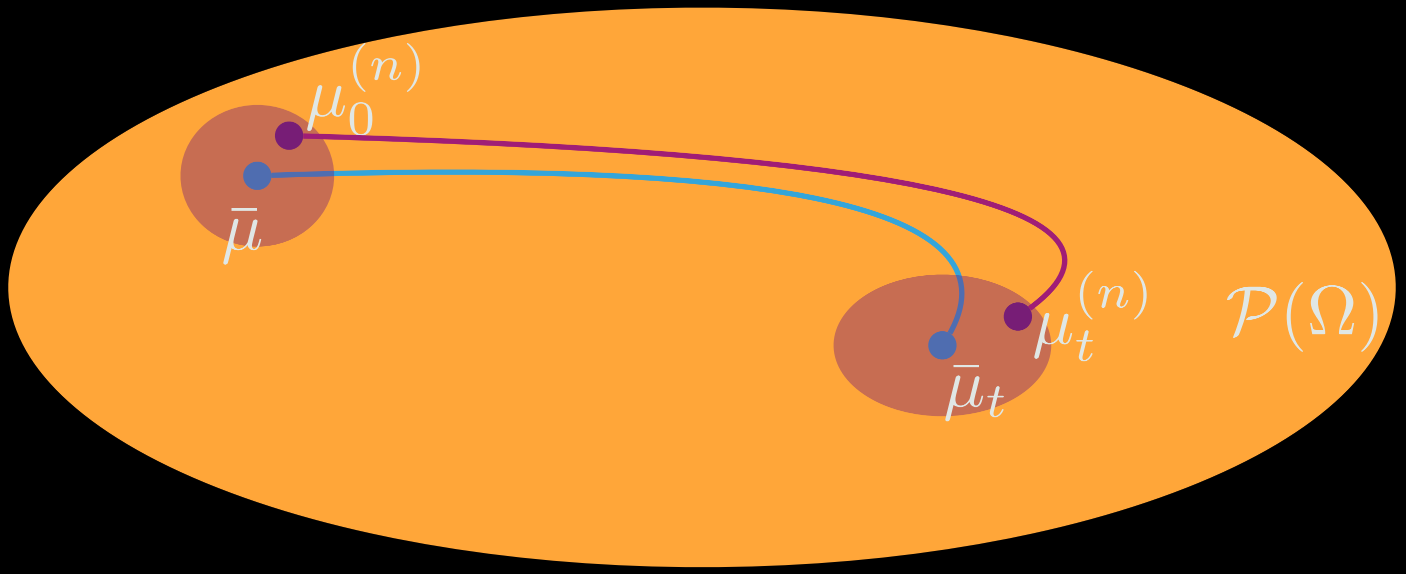
  - Terms R1,R2 only depend on activation function. Not cursed by dimensionality using e.g. ReLU.

- This suggests $\lambda \simeq L^{-1/2}, n \gtrsim \sqrt{L}$ to obtain an efficient learning algorithm in $\mathcal{F}_1$.

- However, previous Monte-Carlo bound is ***static***: if

$$f_t^{(n)} = \frac{1}{n}\sum_j a_j(t)\varphi(z_j(t)) \;, (a_j(0), z_j(0)) \sim \mu_0 \text{ iid},$$

we need to control $\|f_t^{(n)} - \int a\varphi(z)\mu_t(da, dz)\|_\nu^2$



DYNAMIC CLT FOR SHALLOW NEURAL NETWORKS

- This suggests $\lambda \simeq L^{-1/2}, n \gtrsim \sqrt{L}$ to obtain an efficient learning algorithm in $\mathcal{F}_1$.

- However, previous Monte-Carlo bound is ***static***: if

$$f_t^{(n)} = \frac{1}{n}\sum_j a_j(t)\varphi(z_j(t)) \ , (a_j(0), z_j(0)) \sim \mu_0 \text{ iid},$$

we need to control $\|f_t^{(n)} - \int a\varphi(z)\mu_t(da, dz)\|_\nu^2$

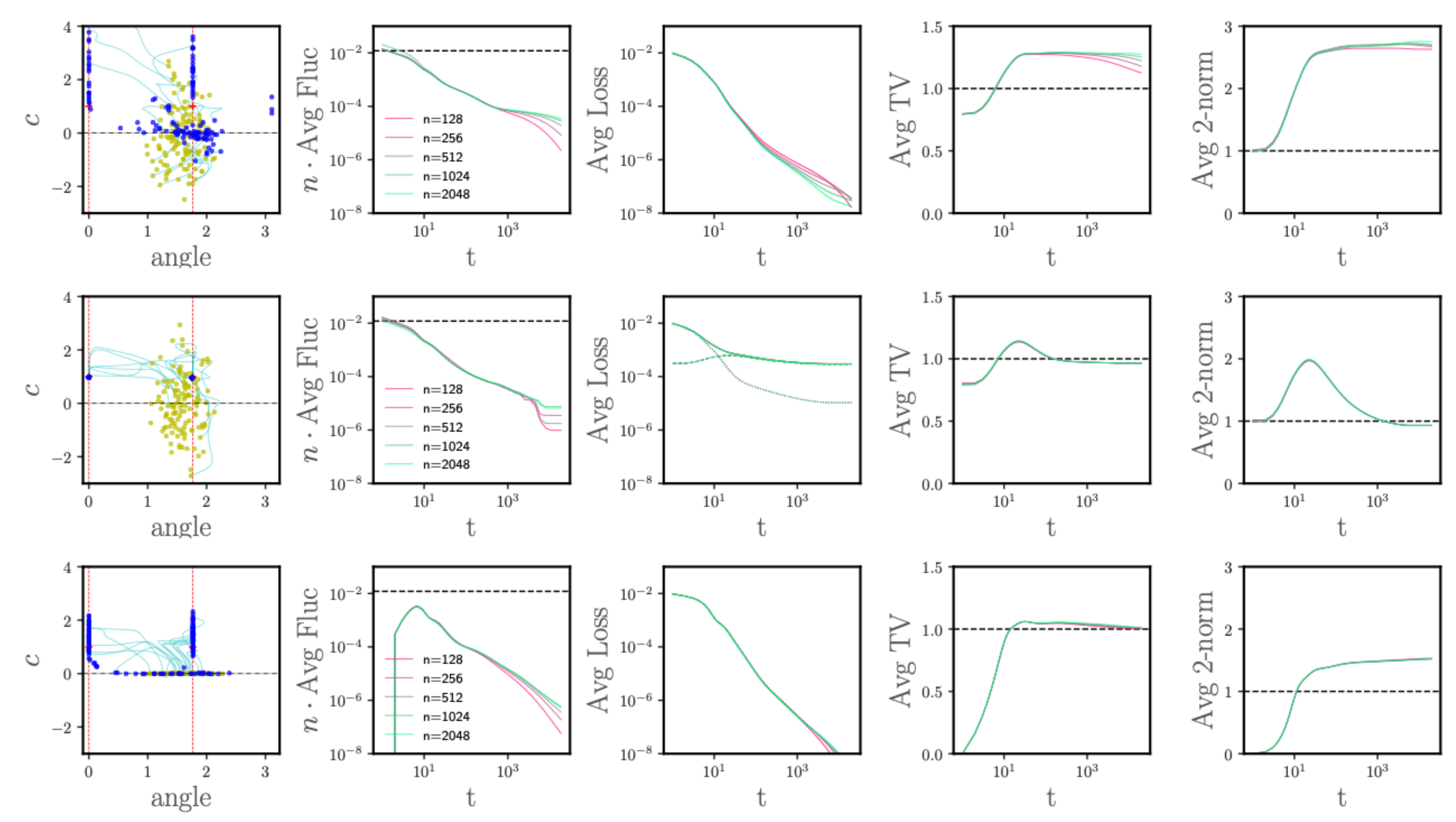
Theorem: [BCRV'19] Under Mean Field global convergence assumptions, it holds

$$\lim_{t\to\infty}\lim_{n\to\infty} n\mathbb{E}\|f_t^{(n)} - f(t)\|_\nu^2 = C < \infty$$

- Extends finite horizon CLT bounds from [Braun & Hepp,'70s] (also [Spilopoulos'19, De Bortoli et al.'20]) using Volterra systems. [Chizat'19] establishes zero fluctuations on sparse well-conditioned.

- Fluctuations vanish at the MC scale in the interpolating, unregularised regime.

NUMERICAL EXPERIMENTS: TEACHER-STUDENT SETUP



- We verify scale of fluctuations at or below MC.

TOWARDS FINITE-WIDTH GUARANTEES

- The previous CLT results are qualitative (limit of infinitely wide networks).

TOWARDS FINITE-WIDTH GUARANTEES

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

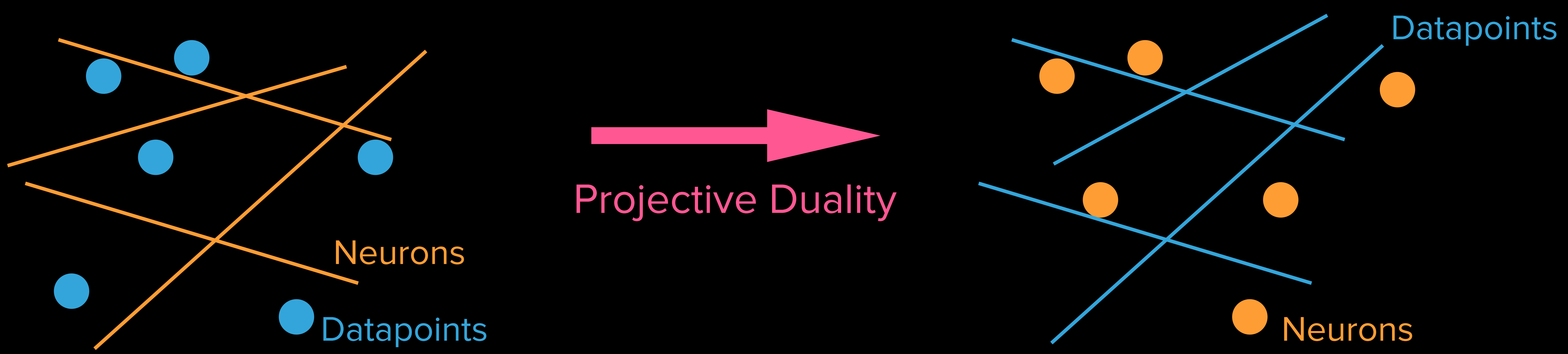
Theorem [DB'20]: The $\mathcal{F}_1$ regularised ERM using ReLU units only admits atomic minimisers, and the functional $\mathcal{E}[\mu]$ is locally strongly convex.

TOWARDS FINITE-WIDTH GUARANTEES

- The previous CLT results are 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 VARIATION-NORM SPACES: DEPTH SEPARATION


- Functions in $\mathcal{F}_1$ are expressed as sparse sums of ridge functions.
- Which function classes are not well approximated in $\mathcal{F}_1$, but are approximable/learnable by deeper architectures efficiently?

BEYOND VARIATION-NORM SPACES: DEPTH SEPARATION

- Functions in $\mathcal{F}_1$ are expressed as sparse sums of ridge functions.
- Which function classes are not well approximated in $\mathcal{F}_1$, but are approximable/learnable by deeper architectures efficiently?
- [Eldan, Shamir, Telgarski, 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 “natural” function separations?

BEYOND VARIATION-NORM SPACES: DEPTH SEPARATION


- Inhomogeneous case: Approximation lower bounds for piece-wise oscillatory functions under heavy-tailed data distributions:

Theorem [BJV'20]: Let $g(x) = \exp\{i\langle \omega_d, \rho(Ux + b)\rangle\}$ with $\|\omega_d\| = \Theta(d^3)$, and $\rho(t) = \max(0, t)$. Let μ a heavy-tailed distribution, and $\mathcal{R}_M$ the class of shallow neural networks with M hidden units. Then

$$\inf_{f \in \mathcal{R}_M} \frac{\mathbb{E}_\mu |f(x) - g(x)|^2}{\mathbb{E}_\mu |g(x)|^2} \geq 1 - M\gamma^d \text{poly}(d) \text{ with } \gamma < 1 \ .$$

  - Efficient approximation with depth-three ReLU networks.

BEYOND VARIATION-NORM SPACES: DEPTH SEPARATION


- Inhomogeneous case: Approximation lower bounds for piece-wise oscillatory functions under heavy-tailed data distributions:

Theorem [BJV'20]: Let $g(x) = \exp\{i\langle \omega_d, \rho(Ux + b)\rangle\}$ with $\|\omega_d\| = \Theta(d^3)$, and $\rho(t) = \max(0, t)$. Let μ a heavy-tailed distribution, and $\mathcal{R}_M$ the class of shallow neural networks with M hidden units. Then

$$\inf_{f \in \mathcal{R}_M} \frac{\mathbb{E}_\mu |f(x) - g(x)|^2}{\mathbb{E}_\mu |g(x)|^2} \geq 1 - M\gamma^d \text{poly}(d) \text{ with } \gamma < 1 \ .$$

  - Efficient approximation with depth-three ReLU networks.

- Homogeneous case: Approximation upper bounds for arbitrary ReLU networks on the sphere with shallow networks:

Theorem [BJV'20]: Let $g(x) = a_{D+1}\rho(A_D\rho(\ldots\rho(A_1 x)))$ be a depth-D ReLU network, with $\sup_{\|x\|=1} g(x) = 1$. Then

$$\inf_{f \in \mathcal{R}_M} \sup_{\|x\|=1} |g(x) - f(x)| \leq \epsilon \text{ if } M \geq \left(2^D C \left(1 + \epsilon^{-2}\right) d\right)^{CD(1+\epsilon^{-1})^D} .$$

  - Rate is not cursed in d (but cursed in depth D and in ϵ^{-1}).

BEYOND VARIATION-NORM SPACES: DEPTH SEPARATION

- Inhomogeneous case: Approximation lower bounds for piece-wise oscillatory functions under heavy-tailed data distributions:

Theorem [BJV'20]: Let $g(x) = \exp\{i\langle \omega_d, \rho(Ux+b)\rangle\}$ with $\|\omega_d\| = \Theta(d^3)$, and $\rho(t) = \max(0,t)$. Let μ a heavy-tailed distribution, and $\mathcal{R}_M$ the class of shallow neural networks with M hidden units. Then

$$\inf_{f \in \mathcal{R}_M} \frac{\mathbb{E}_\mu |f(x) - g(x)|^2}{\mathbb{E}_\mu |g(x)|^2} \geq 1 - M\gamma^d \mathsf{poly}(d) \text{ with } \gamma < 1 .$$

  - Efficient approximation with depth-three ReLU networks.

- Homogeneous case: Approximation upper bounds for arbitrary ReLU networks on the sphere with shallow networks:

Theorem [BJV'20]: Let $g(x) = a_{D+1}\rho(A_D\rho(\ldots\rho(A_1x)))$ be a depth-D ReLU network, with $\sup_{\|x\|=1} g(x) = 1$. Then

$$\inf_{f \in \mathcal{R}_M} \sup_{\|x\|=1} |g(x) - f(x)| \leq \epsilon \text{ if } M \geq \left(2^D C \left(1+\epsilon^{-2}\right) d\right)^{CD(1+\epsilon^{-1})^D} .$$

  - Rate is not cursed in d (but cursed in depth D and in ϵ^{-1}).

- ***Open***: close the gap between lower and upper bounds.

LEARNING UNDER SYMMETRY

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

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

- Simple framework to study symmetries: permutation-invariant functions:

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

Feature domain

$$\Omega \subseteq \mathbb{R}^d$$

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

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

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

Feature domain $\Omega \subseteq \mathbb{R}^d$

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

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

  - E.g particle interaction systems, 3d point-clouds.
- Input Embedding into $\mathcal{P}(\Omega)$: $(x_1, \ldots, x_k) \to \mu^{(k)} = \frac{1}{k} \sum_{j=1}^{k} \delta_{x_j}$.
  - Under appropriate regularity, f extended to $\overline{f} : \mathcal{P}(\Omega) \to \mathbb{R}$.
  - Input domain is not-Euclidean, infinite-dimensional.
  - Functional neural spaces?

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

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

  - Input "weights" ϕ are now test functions.

LEARNING UNDER SYMMETRY

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

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

  - 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 $\mathcal{D}$.

LEARNING UNDER SYMMETRY

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

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

  - 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:

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

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

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

- $\mathcal{S}_3 \subset \mathcal{S}_2 \subset \mathcal{S}_1$ By Jensen.

LEARNING UNDER SYMMETRY

- Hierarchy of functional spaces for learning:

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

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

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

- $\mathcal{S}_3 \subset \mathcal{S}_2 \subset \mathcal{S}_1$ By Jensen.
- Universal approximators of symmetric functions.
- Implemented with two-hidden layer neural networks using random feature kernel expansions:

	First Layer	Second Layer	Third Layer
$\mathcal{S}_1$	Trained	Trained	Trained
$\mathcal{S}_2$	Frozen	Trained	Trained
$\mathcal{S}_3$	Frozen	Frozen	Trained

LEARNING UNDER SYMMETRY

- Hierarchy of functional spaces for learning:

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

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

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

- Approximation lower bounds and generalization guarantees:

Theorem [BZ'20]: For ReLU activations, there exists f_1 with $\|f_1\|_{\mathcal{S}_1} \leq 1$ such that

$$\inf_{\|f\|_{\mathcal{S}_2} \leq \delta} |f_1 - f|_\infty \gtrsim \left| d^{-1} - \delta 2^{-d/2} \right| . \qquad \text{(depth-separation)}$$

Moreover, assuming bounded feature domain Ω, we have

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

- ***Open***: optimization guarantees.

CURRENT AND OPEN PROBLEMS

- Beyond Variation Spaces: Depth-separation
  - What is the functional space associated to deep architectures beyond feature selection? GD optimization in such space?
  - Links with dynamical systems.
- Mean-field formulation is informative in the single-hidden layer model.
  - Extension to deep architectures (ResNet). Geometric networks (CNN,GNN)?
- Polynomial finite width guarantees for typical instances?
- Beyond vanilla gradient descent (adagrad, etc.) ? Role of time-discretization?

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 for high-dimensional ReLU networks", Bruna, Jelassi, Venturi, *in prep.* 20.

"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

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:

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

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

μ_x, μ_y: players strategy distribution

$\ell(x, y)$ smooth

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

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

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

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

$\mu_x, \mu_y\text{: players strategy distribution}$ $\qquad \ell(x,y)$ 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:

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

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

μ_x, μ_y: players strategy distribution

$\ell(x,y)$ 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.