Math and Data

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU CENTER FOR DATA SCIENCE

JOAN BRUNA

ON LEARNING WITH LARGE NEURAL NETWORKS

joint work with Grant Rotskoff, Samy Jelassi, Zhengdao Chen and Eric Vanden-Eijnden




- 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: $f(x; \Theta)$, $\Theta \in \mathcal{D}$. $\quad \mathcal{F} := \{f(\cdot, \Theta); \Theta \in \mathcal{D}\}$.

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: $f(x;\Theta)$, $\Theta \in \mathcal{D}$. $\quad \mathcal{F} := \{f(\cdot,\Theta); \Theta \in \mathcal{D}\}$.
- 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:

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

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

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

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

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

- "Fundamental Theorem of ML": [Bottou & Bousquet]

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

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

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

- "Fundamental Theorem of ML": [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 Supervised ML:
  - Approximation: Functional Approximation that is not cursed by input dimensionality.
  - Generalisation: Statistical Error handled with uniform concentration bounds.
  - Optimization: How to solve (*) efficiently in the high-dimensional regime?

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

- "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].
  - $\mathcal{F} = \mathcal{H}^{s,p} : \text{ Sobolev spaces}$. Minimax rate of approximation is cursed unless $s \geq d/2$: only very smooth functions are allowed.

- "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].
  - $\mathcal{F} = \mathcal{H}^{s,p} : \text{ Sobolev spaces}$. 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?

- "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 <u>deep</u> neural networks?
- … with <u>deep convolutional</u> neural networks?

THIS WORK

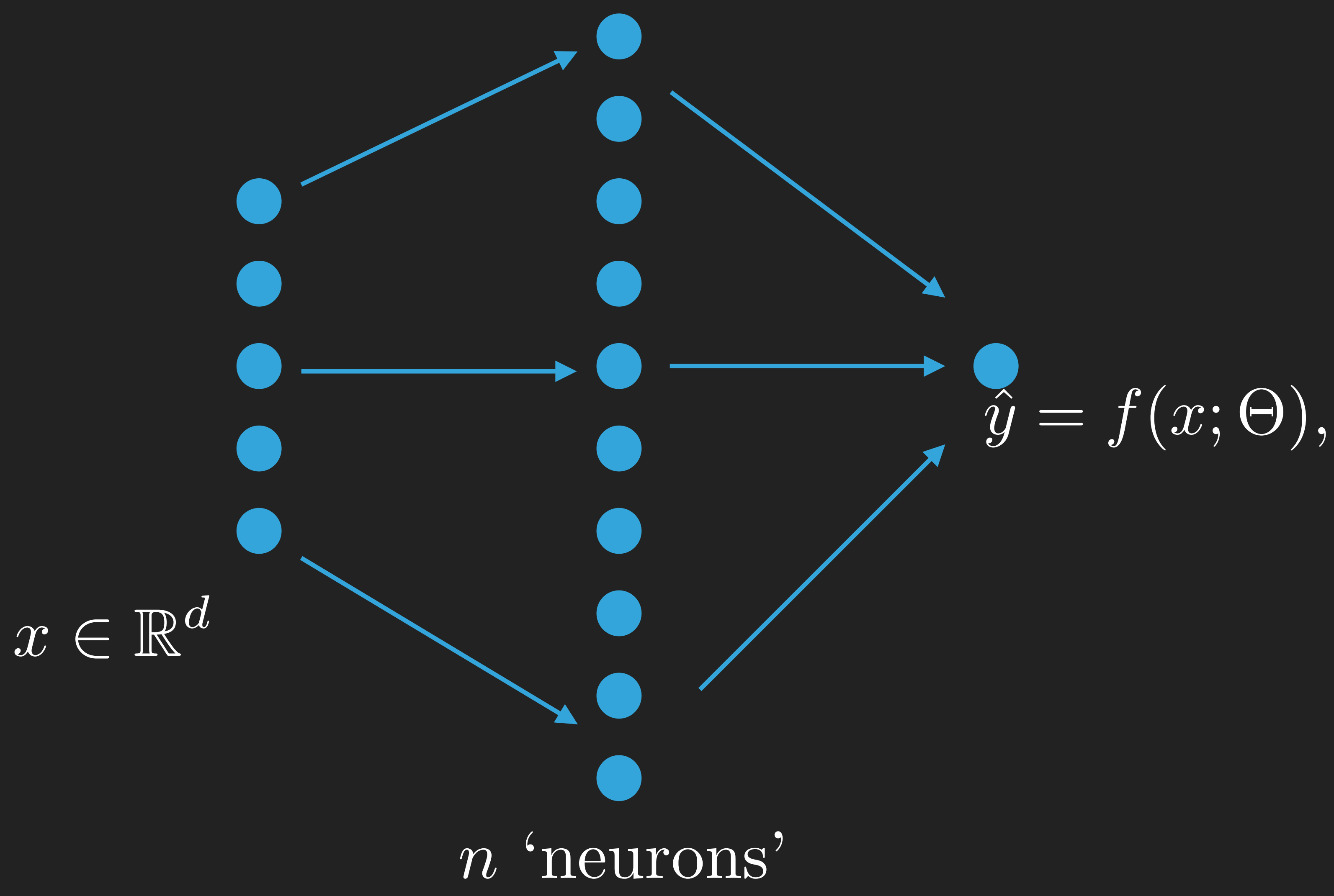
- Focus on simplest neural network family: single hidden-layer.
  - With appropriate scaling, overparametrised regime admits mean-field limit, in which dynamics become tractable and described by a PDE.
  - Scaling is consistent with variation norm spaces, which avoid curse of dimensionality for “sums of simple functions”.

THIS WORK

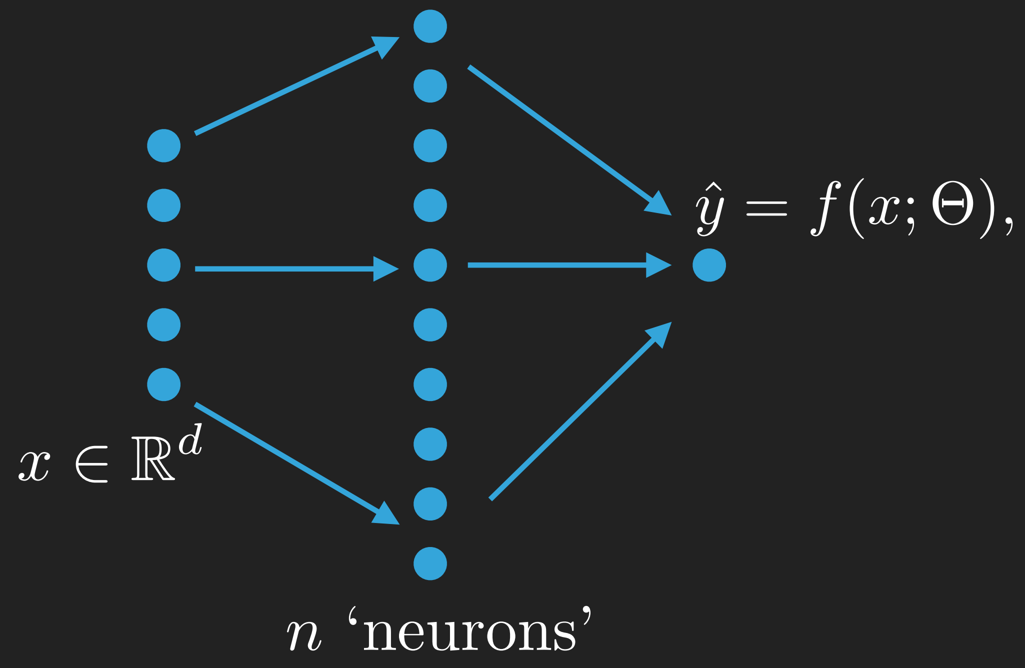
- Focus on simplest neural network family: single hidden-layer.
  - With appropriate scaling, overparametrised regime admits mean-field limit, in which dynamics become tractable and described by a PDE.
  - Scaling is consistent with variation norm spaces, which avoid curse of dimensionality for “sums of simple functions”.
- We propose non-local modification of the dynamics based on unbalanced transport using birth/death processes.
  - New dynamics with provable global convergence and generalization.
  - Although defined in the infinite limit, they admit finite-particle implementation and analysis.
  - Improved convergence with minimal algorithmic impact.

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



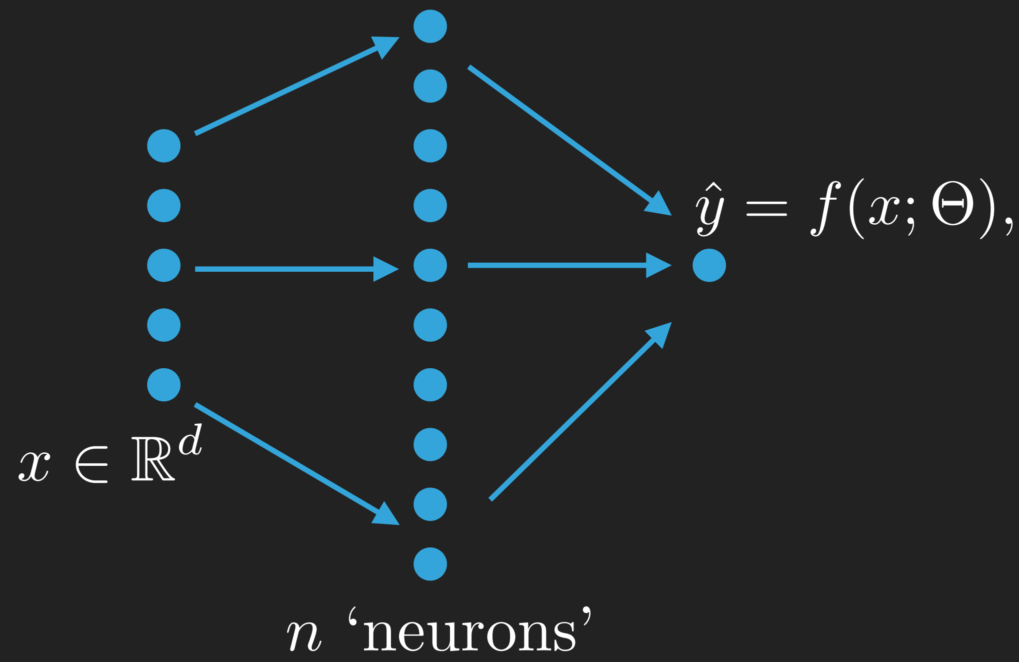
$$f(x;\Theta) = \sum_{j \le n} \tilde{\varphi}(x;\theta_j)$$

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

- 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



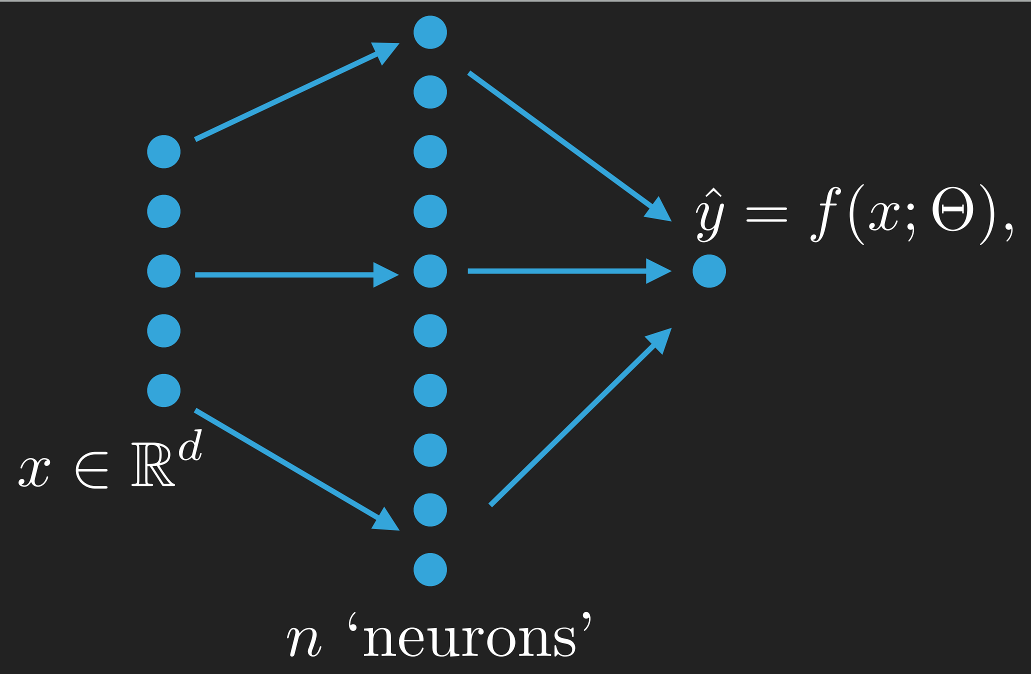
$$f(x;\Theta) = \sum_{j\leq n} \tilde{\varphi}(x;\theta_j)$$

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

- 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].



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

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

- 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 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?

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

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

- 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'16]
- Learning in these RKHS is well-understood (kernel ridge regression), with efficient optimization algorithms.
  - Efficient approximation algorithms through random features [Rahimi/Recht'08, Bach'17] .

- 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'16]

- Learning in these RKHS is well-understood (kernel ridge regression), with efficient optimization algorithms.
  - Efficient approximation algorithms through random features [Rahimi/Recht'08, Bach'17] .
- 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 [Arora et al., Mei et al. Tibshirani, Belkin]. (cf talks by S. Du, M.Belkin, J.Lee, ..)

- Alternatively, we can consider [Bach'16]

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

- $\mathcal{F}_1$ is a Banach space, with norm $\|f\|_{\mathcal{F}_1} := \inf\left\{ \|\mu\|_{TV}\,;\, f = \int \varphi d\mu \right\}$.
- $\mathcal{F}_2 \subset \mathcal{F}_1$ (by Jensen's inequality), and $\mathcal{F}_1$ contains sums of ridge functions.
- Also known as *Barron* Spaces.

- Alternatively, we can consider [Bach'16]

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

- $\mathcal{F}_1$ is a Banach space, with norm $\|f\|_{\mathcal{F}_1} := \inf \left\{ \|\mu\|_{TV}\,;\, f = \int \varphi d\mu \right\}$.

- $\mathcal{F}_2 \subset \mathcal{F}_1$ (by Jensen's inequality), and $\mathcal{F}_1$ contains sums of ridge functions.

- Also known as *Barron* Spaces.

- How to perform optimization and approximation in these spaces?

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

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

- 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, penalized objective becomes

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

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

  - Hamiltonian of a system of n interacting particles.

- 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, penalized objective becomes

$$\mathcal{E}(\Theta) = \mathbb{E}_{\hat{\nu}}[|f(x;\Theta) - f^*|^2] + \lambda \mathcal{V}(\Theta)$$

$$= C - \frac{2}{n}\sum_{j\leq n} F(\theta_j) + \frac{1}{n^2}\sum_{j,j'} U(\theta_j, \theta_{j'})$$

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

  - Hamiltonian of a system of n interacting particles.

- 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].

[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:

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

[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:

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

- *Eulerian perspective:* Rewrite the energy in terms of the empirical measure $\mu_n(t, \theta) = \frac{1}{n}\sum_{j \le n} \delta_{\theta_j(t)}$

[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:

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

- *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 loss becomes

$$\mathcal{E}(\mu) = -2\int F(\theta)\mu(d\theta) + \iint U(\theta,\theta')\mu(d\theta)\mu(d\theta')\,.$$
$$= \left\| f^* - \int_{\mathbb{R}\times\mathcal{D}} a\varphi(\cdot, z)\mu(da, dz)\right\|^2 + \lambda\int_{\mathbb{R}\times\mathcal{D}} |a|^2\mu(da, dz)$$

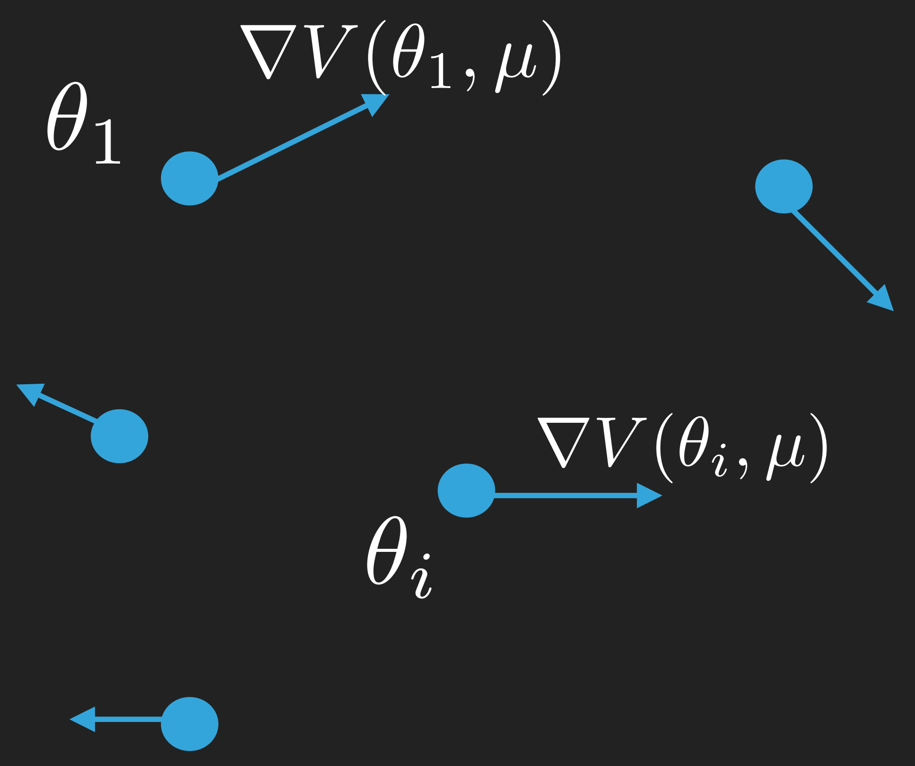
  - quadratic since we consider L2 loss
  - convex in the geometry of linear mixtures (not in general).

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

- It follows that the particle gradients correspond to evaluating a scaled velocity field:

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

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



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

- It follows that the 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*:

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

$$\forall\,\phi \in C_c^\infty(\Omega)\,,\ \partial_t\left(\int \phi\mu_t(d\theta)\right) = -\int \langle \nabla\phi, \nabla V\rangle \mu_t(d\theta)\,.$$

  - This PDE corresponds to a non-linear Liouville equation.
  - Gradient flow of $\mathcal{E}$ for the Wasserstein metric W_2 in $\mathcal{M}(\Omega)$
  - Exact description of particle gradient for atomic measures.

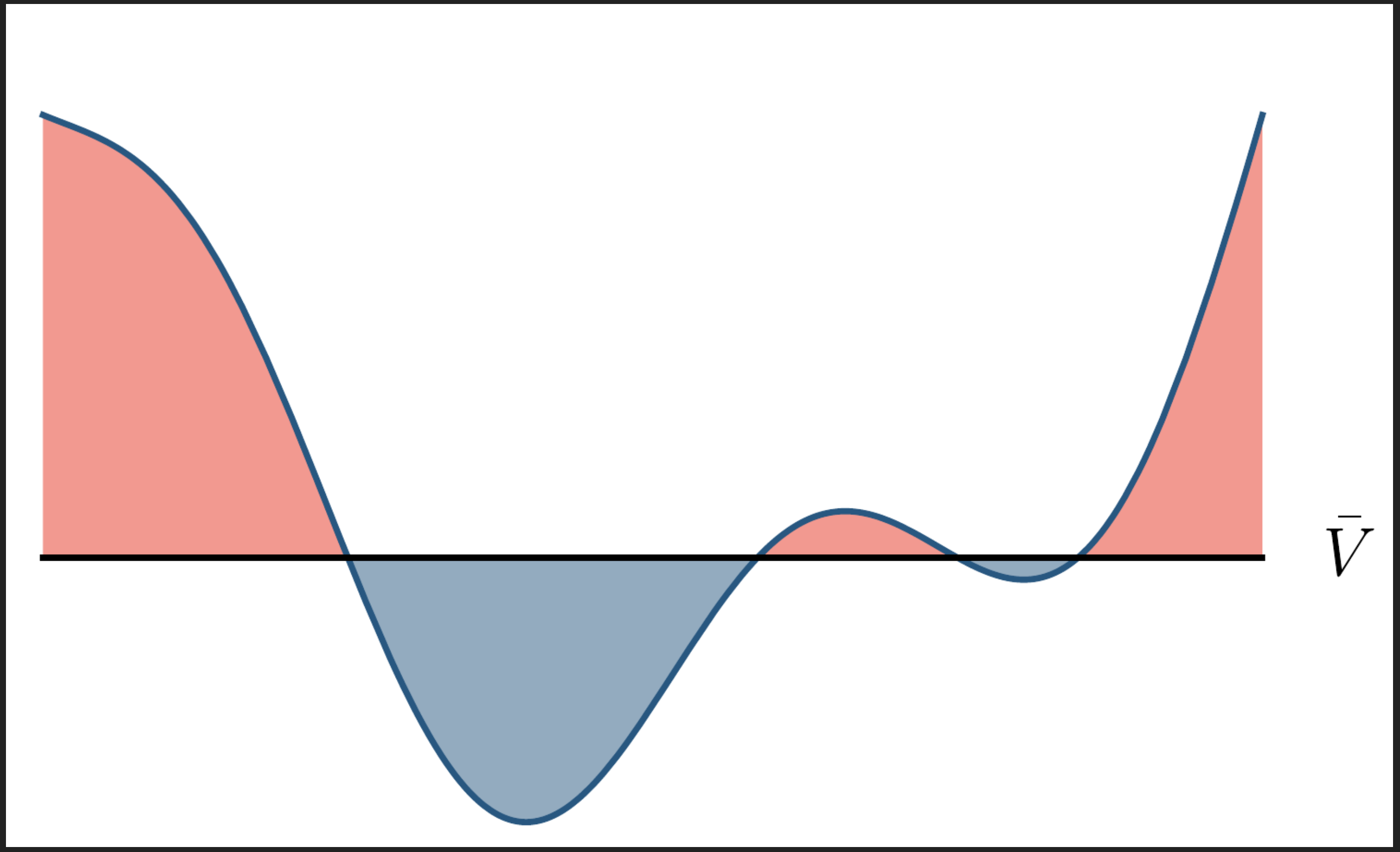
- Consider the evolution of the particle system as n grows.
- $\mu_t^{(n)}$: state of the system after time t, with $\theta_i(0) \sim \bar{\mu} \quad \text{iid.}$

- 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?
- What is the scale of the fluctuations?

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

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

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



- 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 is akin 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.

- 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 $t \geq t_c$.

- 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 $\theta \in \operatorname{supp}(\mu_*)$.
  - How to leverage this mean-field guarantee for finite data/units?

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

- 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

$$\mathbb{E}\|\hat{f}_L - f^*\|_\nu^2 \leq 2\|f^*\|_1 \left(\frac{R_1\|f^*\|_1 + R_2}{\sqrt{L}} + \lambda \right)$$

  - 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 $f_t^{(n)} = \frac{1}{n}\sum_j a_j(t)\varphi(z_j(t))$, $(a_j(0), z_j(0)) \sim \mu_0$ iid, we need to control $\|f_t^{(n)} - \int a\varphi(z)\mu_t(da, dz)\|_\nu^2$

- 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 $f_t^{(n)} = \frac{1}{n}\sum_j a_j(t)\varphi(z_j(t))\ ,(a_j(0), z_j(0)) \sim \mu_0$ iid, we need to control $\|f_t^{(n)} - \int a\varphi(z)\mu_t(da, dz)\|_\nu^2$

- Finite-horizon bounds follow from CLT results [Braun & Hepp,'70s] (also [Spilopoulos'19]).

- Related recent work: [Chizat'19] establishes global convergence for singular initializations, with convergence rates. Deterministic, but cursed by input dim.

- 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)?
- Establishing large-deviation principle for finite-particle dynamics.
- Beyond vanilla gradient descent (adagrad, etc.) ? Role of time-discretization? (cf talk by T. Ma, S. Arora).

Thanks!

References:

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

"Large Deviations for Large Neural Networks", Rotskoff, Chen, Bruna, Vanden-Eijnden (in preparation).

- Mixture of Gaussians:

$$f^*(x) = \frac{1}{S}\sum_{s\leq S}\frac{c_s}{(2\pi\sigma_s^2)^{d/2}}e^{-\|x-z_s\|^2/(2\sigma_s^2)}.$$

- Gaussian activation function:

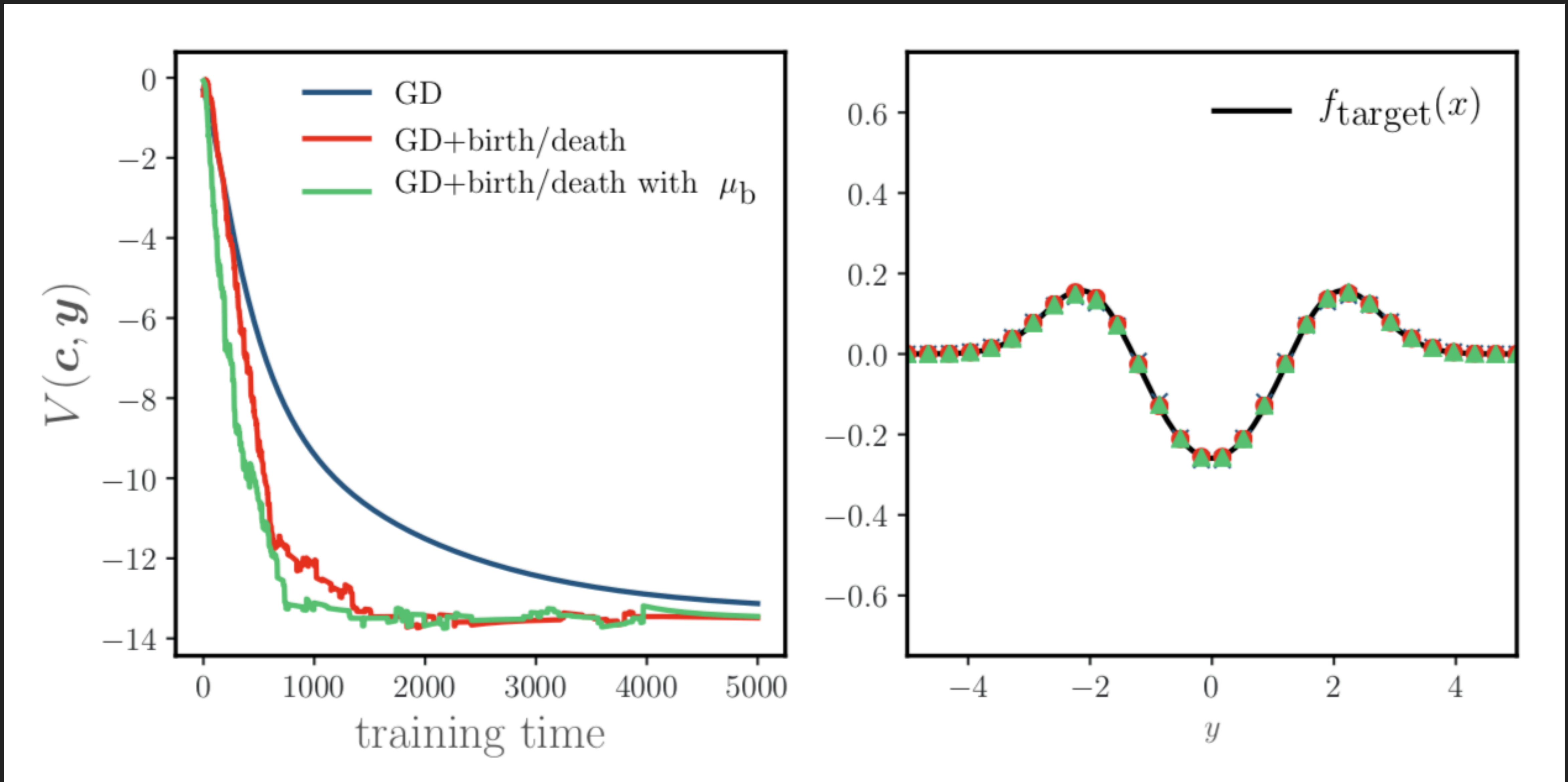
$$\varphi(x;\theta) = \frac{c}{(2\pi\sigma^2)^{d/2}}e^{-\|x-z\|^2/(2\sigma^2)}, \ \theta = (c, z).$$

- "Overparametrised" model: $n > S$

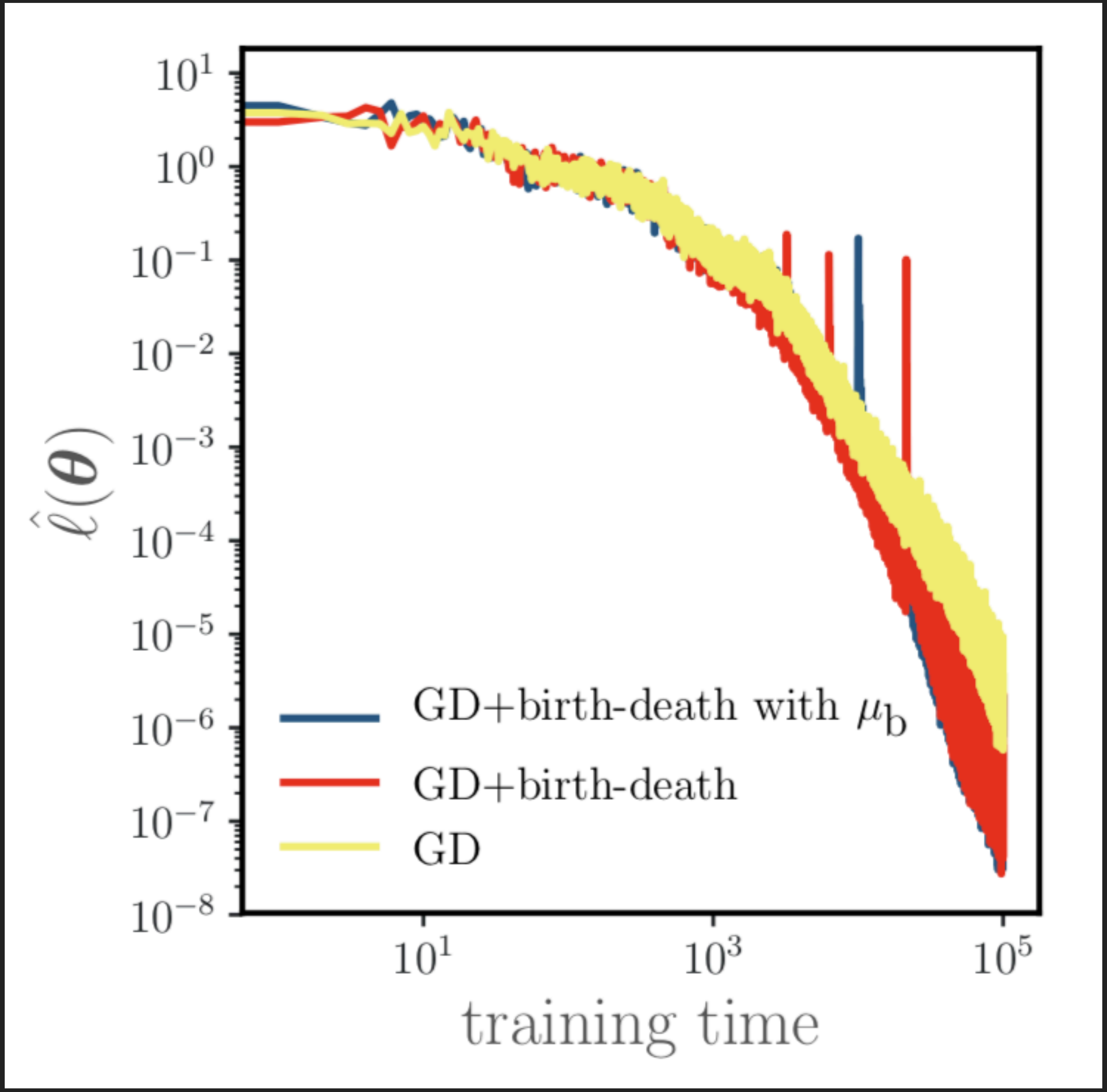
$$f(x;\Theta) = \frac{1}{n}\sum_{i\leq n}\varphi(x;\theta_i).$$

- Mixture of Gaussians:

$$f^*(x) = \frac{1}{S}\sum_{s\leq S}\frac{c_s}{(2\pi\sigma_s^2)^{d/2}}e^{-\|x-z_s\|^2/(2\sigma_s^2)}.$$



- Teacher-Student single hidden layer neural network using ReLU activations



10 planted neurons

$n = 50$

$d = 50$