Mean-field Neural Networks

Grant M. Rotskoff
Deep Learning Theory Summer School SJTU

July 16, 2020

Stanford University

rotskoff@stanford.edu
https://statmech.stanford.edu

Outline

The expressive power of neural networks
- Functional formulation

Optimizing neural networks
- Interlude: Wasserstein Gradient Flows

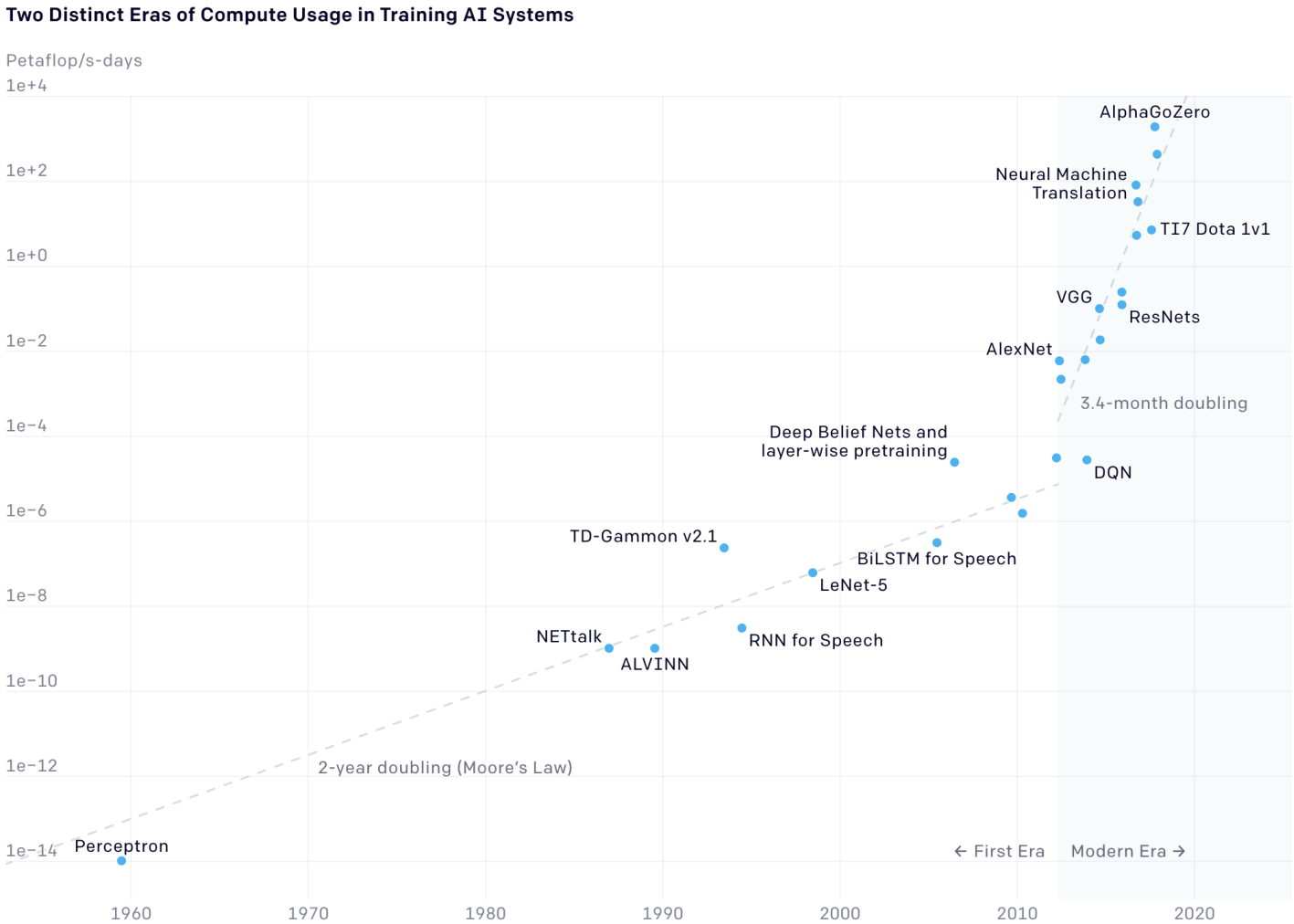
Trainability and convergence results
- Law of large numbers
- Central limit theorem

Nonlocal algorithms
- Guarantees of global convergence
- Convergence rate results

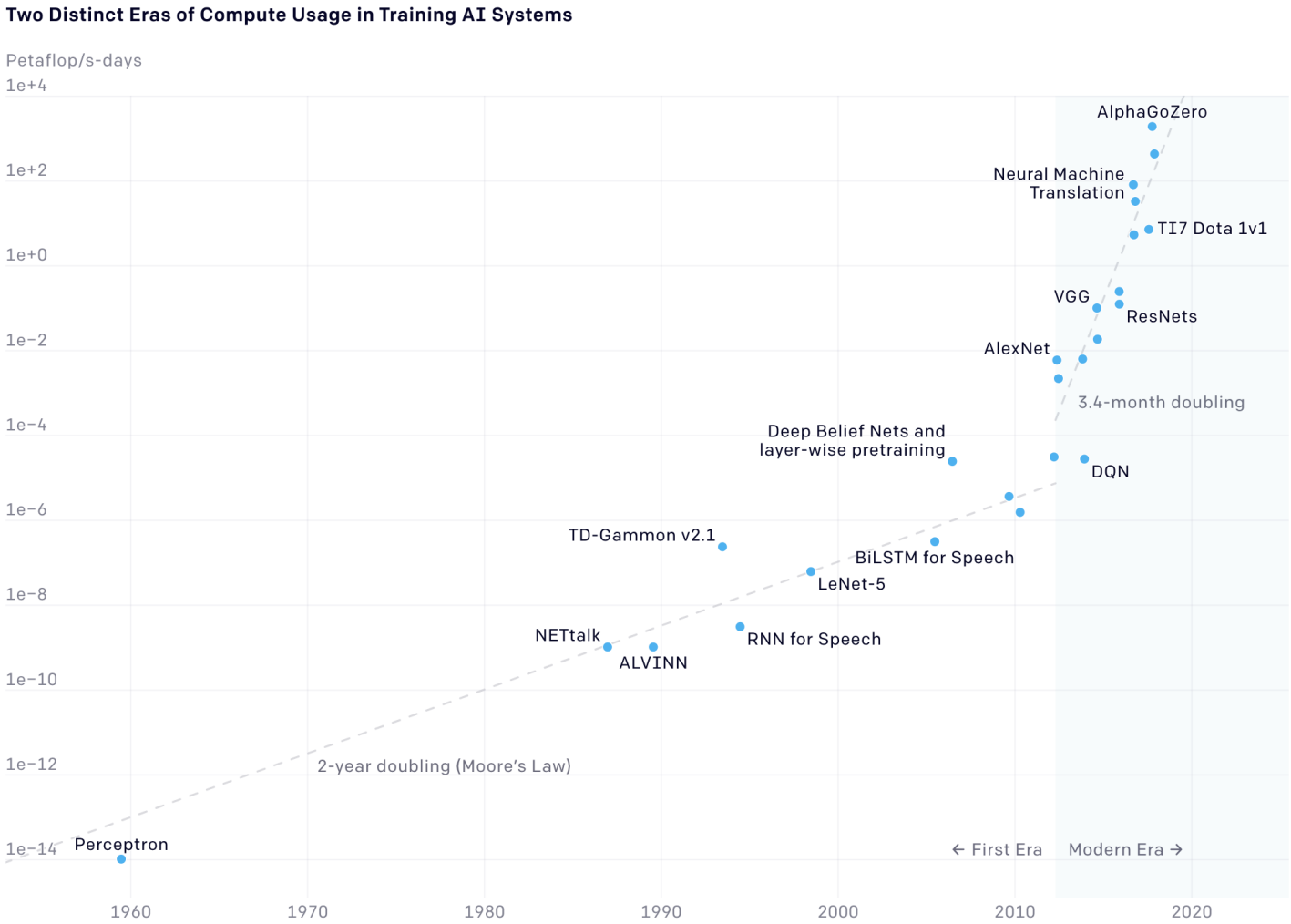
Open problems

The expressive power of neural networks

What is the origin of deep learning's empirical success?



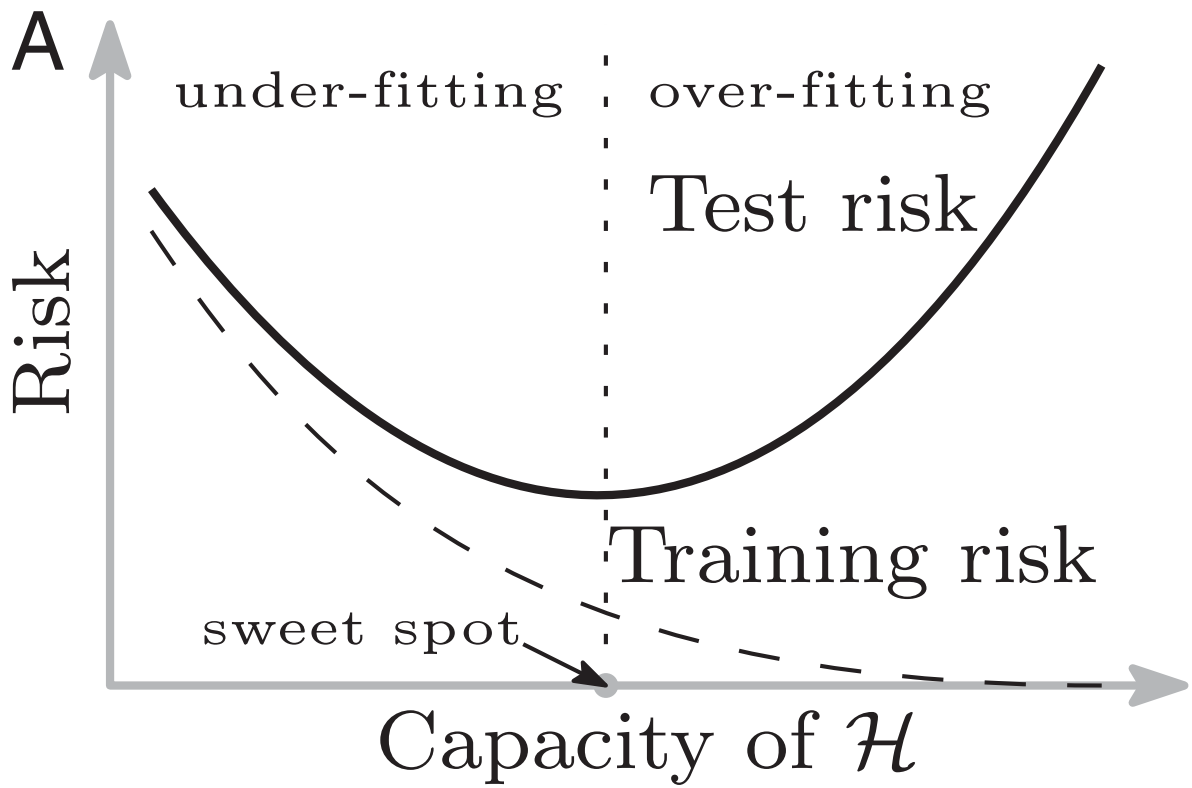
What is the origin of deep learning's empirical success?



curse of dimensionality?

[AH18]

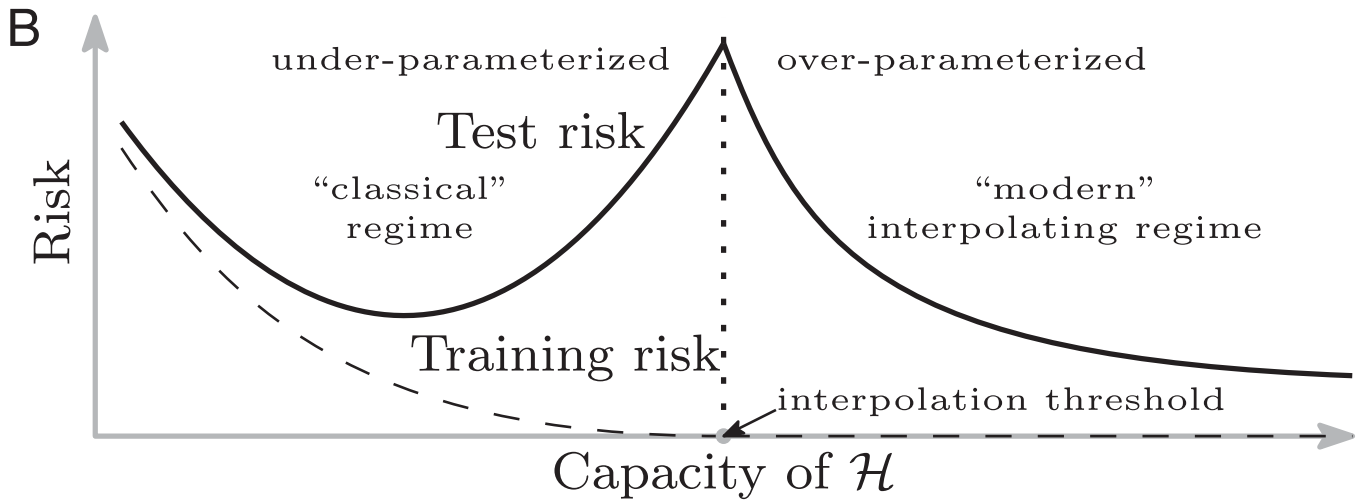
The right number of parameters is determined by the "right" model.



[BHMM19]

Evidence for "double descent"

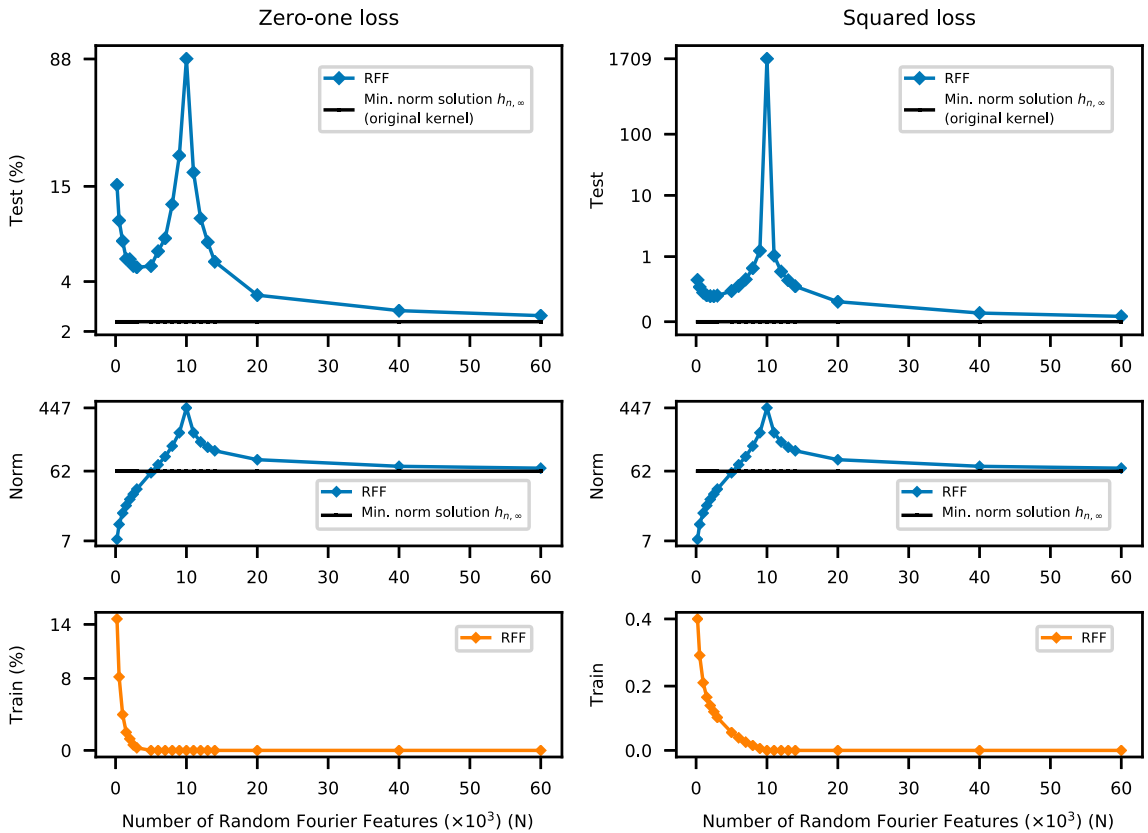
Emerging theoretical picture challenges this paradigm



[BHMM19]

Evidence for "double descent"

Neural networks are far from the *interpolation* regime



Overparameterization is *not* harmful empirically. [BHMM19]

Focus on the simplest model, for analytical purposes: *single hidden layer neural network*

$$f^{(n)}(\boldsymbol{x};\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x},\boldsymbol{\theta}_i); \quad \mathbb{R}^k \to \mathbb{R} \tag{1}$$

- Parameters: $\{\boldsymbol{\theta}_i\} \in D^{\otimes n}$.
- Nonlinear unit: φ e.g., ReLU
- Proper scaling is important! (cf. Jacot's lecture)

Neural net with n finite,

$$f^{(n)}(\boldsymbol{x};\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x},\boldsymbol{\theta}_i) \tag{2}$$

Replace discrete parameters with a discrete distribution

$$\mu^{(n)}(d\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\delta_{\boldsymbol{\theta}_i}(d\boldsymbol{\theta}) \tag{3}$$

Exact rewriting:

$$f^{(n)}(\boldsymbol{x}) = \int_D \varphi(\boldsymbol{x},\boldsymbol{\theta})\mu^{(n)}(d\boldsymbol{\theta}) \tag{4}$$

Homogeneous units:

$$\varphi(\boldsymbol{x}, \boldsymbol{\theta}) = c\hat{\varphi}(\boldsymbol{x}, z)$$

$$\boldsymbol{\theta} = (c, z) \in D \equiv \mathbb{R} \times \hat{D}$$

$$f^{(n)}(\boldsymbol{x}) = \int_{\hat{D}} \hat{\varphi}(\boldsymbol{x}, z)\gamma^{(n)}(z) \tag{5}$$

$\gamma^{(n)}$ is a signed Radon measure.

$$\mathcal{F}_1(\hat{\varphi}) = \left\{ g \;\middle|\; g = \int_{\hat{D}} \hat{\varphi}(\cdot, z)\gamma(dz), \quad \gamma \text{ s.t. } |\gamma|_{TV} \leq \infty \right\} \tag{6}$$

[Bac17]

This is the random kernel limit, closely related to the Neural Tangent Kernel.

- "Features" are sampled randomly
- Linear coefficients are optimized (can be done via regression)

$$\mathcal{F}_2(\varphi) = \left\{ g \;\middle|\; g = \int_{\hat{D}} \varphi(\cdot, \boldsymbol{\theta}) \mu(d\boldsymbol{\theta}), \quad d\mu(\boldsymbol{\theta}) = \rho(\boldsymbol{\theta}) d\tau(\boldsymbol{\theta}) \right\} \quad (7)$$

ρ square-integrable

$\operatorname{supp} \tau = \hat{D}$

This is the random kernel limit, closely related to the Neural Tangent Kernel.

- "Features" are sampled randomly
- Linear coefficients are optimized (can be done via regression)

$$\mathcal{F}_2(\varphi) = \left\{ g \;\middle|\; g = \int_{\hat{D}} \varphi(\cdot, \boldsymbol{\theta}) \mu(d\boldsymbol{\theta}), \quad d\mu(\boldsymbol{\theta}) = \rho(\boldsymbol{\theta}) d\tau(\boldsymbol{\theta}) \right\} \tag{7}$$

ρ square-integrable

$\operatorname{supp} \tau = \hat{D}$

NB: $\mathcal{F}_2$ is a RKHS with $k(\boldsymbol{x}, \boldsymbol{y}) = \int_D \varphi(\boldsymbol{x}, \boldsymbol{\theta}) \varphi(\boldsymbol{y}, \boldsymbol{\theta}) d\tau$

[Bac17, CB18a]

Question: How expressive are neural networks?

Approximation problem: given function $f : \mathbb{R}^k \to \mathbb{R}$. Goal: learn f

Question: How expressive are neural networks?

Approximation problem: given function $f : \mathbb{R}^k \to \mathbb{R}$. Goal: learn f

Theorem ([Bar93, Cyb89])

Let φ be a continuous, non-polynomial function. Assume $f \in L_2(\Omega, \nu)$ and $\epsilon > 0$. Then there exists a signed Radon measure $\gamma_* \in \mathcal{M}(D)$ such that

$$f_* = \int \hat{\varphi}(\boldsymbol{x}, \boldsymbol{z}) \gamma_*(\boldsymbol{z}) \tag{8}$$

and

$$\|f - f^*\|_{2,\nu} \leq \epsilon \tag{9}$$

What these theorems *do not* say

$$f_* = \int \hat{\varphi}(\boldsymbol{x}, \boldsymbol{z}) \gamma_*(\boldsymbol{z}) \tag{10}$$

- Can we train to find f_*?
- Does training converge (in any limit)?
- How does the error scale with n? (Barron discusses in special case)

Questions?

Optimizing neural networks

Data

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{P} \overset{iid}{\sim} \nu$$

Learning on the mean-squared loss:

$$\ell(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = \tfrac{1}{2}\mathbb{E}_{y,\boldsymbol{x}} \left| y - f_n(\boldsymbol{x}; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) \right|^2, \tag{11}$$

expand the quadratic:

$$F(\boldsymbol{\theta}) = -\mathbb{E}_{y,\boldsymbol{x}} \left[y\varphi(\boldsymbol{x}, \boldsymbol{\theta}) \right], \tag{12}$$

and,

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathbb{E}_{\boldsymbol{x}} \left[\varphi(\boldsymbol{x}, \boldsymbol{\theta}) \varphi(\boldsymbol{x}, \boldsymbol{\theta}') \right]. \tag{13}$$

Another framework

Alternative formulation: *interacting particle system*

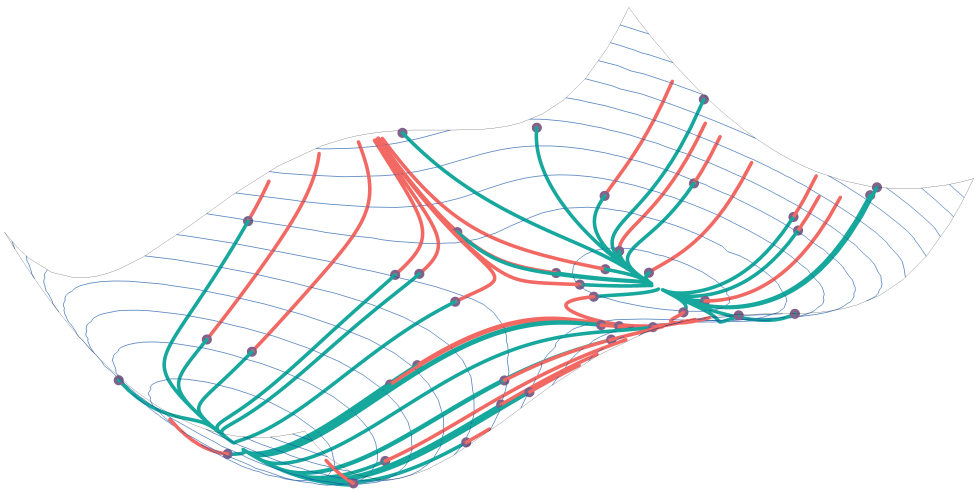
$$\ell(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = \sum_{i=1}^{n} F(\boldsymbol{\theta}_i) + \frac{1}{2n} \sum_{i,j=1}^{n} K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \tag{14}$$

$F : D \to \mathbb{R}$—single particle energy function
$K : D \times D \to \mathbb{R}$—symmetric semi-positive definite interaction kernel

parameters $\leftrightarrow$ interacting particles

Discrete time gradient updates

$$\Theta^{(k+1)} = \Theta^{(k)} - \tau \nabla F(\Theta^{(k)}) \tag{15}$$



Gradient flows $\tau \to 0$

An implicit formulation, sometimes used for numerical schemes:

$$\Theta^{(k+1)} = \text{argmin}_{\Theta} \, F(\Theta) + \frac{1}{2\tau} \|\Theta - \Theta^{(k)}\|_2^2 \tag{16}$$

The Euclidean metric is the "proximity" metric.

$$\Theta^{(k+1)} = \Theta^{(k)} - \tau \nabla F(\Theta^{(k+1)}) \tag{17}$$

An implicit formulation, sometimes used for numerical schemes:

$$\Theta^{(k+1)} = \text{argmin}_{\Theta} \, F(\Theta) + \frac{1}{2\tau} \|\Theta - \Theta^{(k)}\|_2^2 \tag{16}$$

The Euclidean metric is the "proximity" metric.

$$\Theta^{(k+1)} = \Theta^{(k)} - \tau \nabla F(\Theta^{(k+1)}) \tag{17}$$

Closely related to "mirror descent" algorithms.

How to formulate explicit scheme? Proximal scheme: need a *metric*. The Wasserstein p-distance is

$$W_p(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\Omega\times\Omega} |\boldsymbol{x}-\boldsymbol{y}|^p d\pi(\boldsymbol{x},\boldsymbol{y}) \right)^{1/p} \tag{18}$$

where the connection π is a probability measure in the set

$$\Pi(\mu,\nu) = \left\{ \pi : \Omega\times\Omega \to [0,1] \;\middle|\; d\mu(\boldsymbol{x}) = \int_{\boldsymbol{y}\in\Omega} d\pi(\boldsymbol{x},\boldsymbol{y}), \; d\nu(\boldsymbol{x}) = \int_{\boldsymbol{x}\in\Omega} d\pi(\boldsymbol{x},\boldsymbol{y}) \right\}$$

[JKO98] established the relationship betweent the Fokker-Planck equation and Wasserstein gradient flows.

$$\partial_t \rho = \nabla \cdot \nabla(\rho \nabla V) + \beta^{-1} \Delta \rho \tag{19}$$

Is gradient flow on W_2 of "free energy" functional,

$$\mathcal{F}[\rho] = \int V d\rho + \beta^{-1} \int \log \rho d\rho \tag{20}$$

(cf. blog post)

[JKO98] established the relationship betweent the Fokker-Planck equation and Wasserstein gradient flows.

$$\partial_t \rho = \nabla \cdot \nabla(\rho \nabla V) + \beta^{-1} \Delta \rho \tag{19}$$

Is gradient flow on W_2 of "free energy" functional,

$$\mathcal{F}[\rho] = \int V d\rho + \beta^{-1} \int \log \rho d\rho \tag{20}$$

(cf. blog post)

End of interlude

Deriving a PDE for training in $\mathcal{F}_1$

$$\int_D \chi(\boldsymbol{\theta}) \partial_t \mu_t^{(n)}(d\boldsymbol{\theta}) \qquad \mu_t^{(n)} = \frac{1}{n} \sum \delta_{\boldsymbol{\theta}_i(t)}(d\boldsymbol{\theta})$$

$$= \frac{1}{n} \sum_{i=1}^{n} \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot \dot{\boldsymbol{\theta}}_i(t)$$

Deriving a PDE for training in $\mathcal{F}_1$

$$\int_D \chi(\boldsymbol{\theta}) \partial_t \mu_t^{(n)}(d\boldsymbol{\theta}) \qquad \mu_t^{(n)} = \frac{1}{n} \sum \delta_{\boldsymbol{\theta}_i(t)}(d\boldsymbol{\theta})$$

$$= \frac{1}{n} \sum_{i=1}^{n} \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot \dot{\boldsymbol{\theta}}_i(t)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot \left(\nabla F(\boldsymbol{\theta}_i(t)) + \frac{1}{n} \sum_{j=1}^{n} \nabla K(\boldsymbol{\theta}_i(t), \boldsymbol{\theta}_j(t)) \right)$$

Deriving a PDE for training in $\mathcal{F}_1$

$$\int_D \chi(\boldsymbol{\theta}) \partial_t \mu_t^{(n)}(d\boldsymbol{\theta}) \qquad \mu_t^{(n)} = \frac{1}{n} \sum \delta_{\boldsymbol{\theta}_i(t)}(d\boldsymbol{\theta})$$

$$= \frac{1}{n} \sum_{i=1}^{n} \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot \dot{\boldsymbol{\theta}}_i(t)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \nabla \chi(\boldsymbol{\theta}_i(t)) \cdot \left(\nabla F(\boldsymbol{\theta}_i(t)) + \frac{1}{n} \sum_{j=1}^{n} \nabla K(\boldsymbol{\theta}_i(t), \boldsymbol{\theta}_j(t)) \right)$$

$$= \int_D \nabla \chi(\boldsymbol{\theta}) \cdot \left(\nabla F(\boldsymbol{\theta}) + \int_D \nabla K(\boldsymbol{\theta}, \boldsymbol{\theta}') \mu_t^{(n)}(d\boldsymbol{\theta}') \right) \mu_t^{(n)}(d\boldsymbol{\theta})$$

Nonlinear Liouville Equation

$$\partial_t \mu_t = \nabla \cdot (\nabla V(\boldsymbol{\theta}, \mu_t)\mu_t) \tag{21}$$

where

$$V(\boldsymbol{\theta}, \mu) = F(\boldsymbol{\theta}) + \int K(\boldsymbol{\theta}, \boldsymbol{\theta}')\mu(d\boldsymbol{\theta}') \tag{22}$$

[MMN18, CB18b, RV18, SS18]

Questions?

Trainability and convergence results

The Wasserstein gradient flow,

$$\partial_t \mu_t = \nabla \cdot (\nabla V(\boldsymbol{\theta}, \mu_t)\mu_t) \tag{23}$$

is a flow on the *convex* energy functional

$$\begin{aligned}\mathcal{E}[\mu] &= C_f - \int_D F(\boldsymbol{\theta})\mu(d\boldsymbol{\theta}) + \tfrac{1}{2}\int_{(D)^2} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\mu(d\boldsymbol{\theta})\mu(d\boldsymbol{\theta}') \\ &= \tfrac{1}{2}\int_\Omega \left(f(\boldsymbol{x}) - \int_D \varphi(\boldsymbol{x}, \boldsymbol{\theta})\mu(d\boldsymbol{\theta}) \right)^2 \nu(d\boldsymbol{x}) \geq 0\end{aligned} \tag{24}$$

Dynamics at the level of the function representation:

$$\begin{aligned}\partial_t f_t(\boldsymbol{x}) &= \int_D \varphi(\boldsymbol{x},\boldsymbol{\theta})\partial_t \mu_t(d\boldsymbol{\theta}) \\ &= -\int_D \nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{x},\boldsymbol{\theta}) \cdot \left(\int_\Omega \nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{x}',\boldsymbol{\theta})\,(f_t(\boldsymbol{x}') - f(\boldsymbol{x}'))\,\nu(d\boldsymbol{x}')\mu_t(d\boldsymbol{\theta}) \right)\end{aligned} \tag{25}$$

Dynamics at the level of the function representation:

$$\begin{aligned}\partial_t f_t(\boldsymbol{x}) &= \int_D \varphi(\boldsymbol{x},\boldsymbol{\theta})\partial_t \mu_t(d\boldsymbol{\theta}) \\ &= -\int_D \nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{x},\boldsymbol{\theta}) \cdot \left(\int_\Omega \nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{x}',\boldsymbol{\theta}) \left(f_t(\boldsymbol{x}') - f(\boldsymbol{x}')\right) \nu(d\boldsymbol{x}')\mu_t(d\boldsymbol{\theta}) \right)\end{aligned} \tag{25}$$

Theorem (Law of Large Numbers [RV18])

*For *well-prepared initial conditions*, as $n \to \infty$, $f_t^{(n)} \to f_t$ a.s. pointwise, where f_t satisfies*

$$\partial_t f_t(\boldsymbol{x}) = -\int_\Omega M([\mu_t], \boldsymbol{x}, \boldsymbol{x}') \left(f_t(\boldsymbol{x}') - f(\boldsymbol{x}')\right) \nu(d\boldsymbol{x}'). \tag{26}$$

The kernel is positive semi-definite. Explicitly,

$$M([\mu], \boldsymbol{x}, \boldsymbol{x}') = \int_D \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{x}, \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} \varphi(\boldsymbol{x}', \boldsymbol{\theta}) \mu(d\boldsymbol{\theta})$$
$$= \int_{\mathbb{R} \times \hat{D}} \left(c^2 \nabla_z \hat{\varphi}(\boldsymbol{x}, z) \cdot \nabla_z \hat{\varphi}(\boldsymbol{x}', z) + \hat{\varphi}(\boldsymbol{x}, z) \hat{\varphi}(\boldsymbol{x}', z) \right) \mu(dc, dz). \tag{27}$$

What are the conditions for *global convergence*?

- Easy to see energy is decreasing:

$$\frac{dE}{dt} = -\int_D |\nabla V(\boldsymbol{\theta}, [\mu_t])|^2 \mu_t(d\boldsymbol{\theta}) \tag{28}$$

Long time convergence

- Easy to see energy is decreasing:

$$\frac{dE}{dt} = -\int_D |\nabla V(\boldsymbol{\theta}, [\mu_t])|^2 \mu_t(d\boldsymbol{\theta}) \tag{28}$$

- Sufficient condition for fixed points $\nabla V(\boldsymbol{\theta}, \mu) = 0$

- Easy to see energy is decreasing:

$$\frac{dE}{dt} = -\int_D |\nabla V(\boldsymbol{\theta}, [\mu_t])|^2 \mu_t(d\boldsymbol{\theta}) \tag{28}$$

- Sufficient condition for fixed points $\nabla V(\boldsymbol{\theta}, \mu) = 0$
- Hard to ensure that stationary points are global minimizers:

$$\begin{cases} V(\boldsymbol{\theta}, [\mu^*]) \geq \bar{V}[\mu^*] & \text{for} \quad \boldsymbol{\theta} \in D \\ V(\boldsymbol{\theta}, [\mu^*]) = \bar{V}[\mu^*] & \text{for} \quad \boldsymbol{\theta} \in \operatorname{supp} \mu^* \end{cases} \tag{29}$$

$$\bar{V}[\mu] = \int_D V(\boldsymbol{\theta}, \mu) \mu(d\boldsymbol{\theta})$$

Homogeneous nonlinearities suffice!

Proposition (Global convergence)

If $\mu_t \to \mu^* \in \mathcal{M}_+(D)$ as $t \to \infty$, then for $f \in \mathcal{F}_1(\varphi)$, μ^* is a minimizer of $\mathcal{E}[\mu]$ and we have

$$\lim_{t\to\infty} \int_D \varphi(\cdot, \boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta}) = \int_D \varphi(\cdot, \boldsymbol{\theta}) \mu^*(d\boldsymbol{\theta}) = f. \tag{30}$$

[CB18b, RV18]

Homogeneous nonlinearities suffice!

Proposition (Global convergence)

If $\mu_t \to \mu^* \in \mathcal{M}_+(D)$ *as* $t \to \infty$, *then for* $f \in \mathcal{F}_1(\varphi)$, μ^* *is a minimizer of* $\mathcal{E}[\mu]$ *and we have*

$$\lim_{t\to\infty} \int_D \varphi(\cdot, \boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta}) = \int_D \varphi(\cdot, \boldsymbol{\theta}) \mu^*(d\boldsymbol{\theta}) = f. \tag{30}$$

[CB18b, RV18]

Pithily,

$$\lim_{n\to\infty} \lim_{t\to\infty} f_t^{(n)} = f.$$

- Discrepancy between parameter measures:

$$\omega_t^{(n)} = n^{1/2}\left(\mu_t^{(n)} - \mu_t\right),$$

- Deviation in the function representation:

$$\begin{aligned} g_t^{(n)} &= n^{1/2}\left(f_t^{(n)} - f_t\right) = \int_D \varphi(\cdot,\boldsymbol{\theta})\omega_t^{(n)}(d\boldsymbol{\theta}) \\ &= n^{-1/2}\sum_{i=1}^{n}\left(\varphi(\cdot,\boldsymbol{\theta}_i(t)) - \int_D \varphi(\cdot,\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta})\right) \end{aligned}$$

$$\begin{aligned} \partial_t g_t = &-\int_\Omega M(\boldsymbol{x},\boldsymbol{x}',\omega_t)\left(f_t(\boldsymbol{x}') - f(\boldsymbol{x}')\right)\nu(d\boldsymbol{x}') \\ &-\int_\Omega M(\boldsymbol{x},\boldsymbol{x}',\mu_t)g_t(\boldsymbol{x}')\nu(d\boldsymbol{x}') \end{aligned} \tag{31}$$

Proposition (CLT)

As $n \to \infty$, $g_t^{(n)} \to g_t$ in law, where g_t is the zero mean Gaussian process with covariance given by an explicit, closed equation of f (cf. [RV18]).

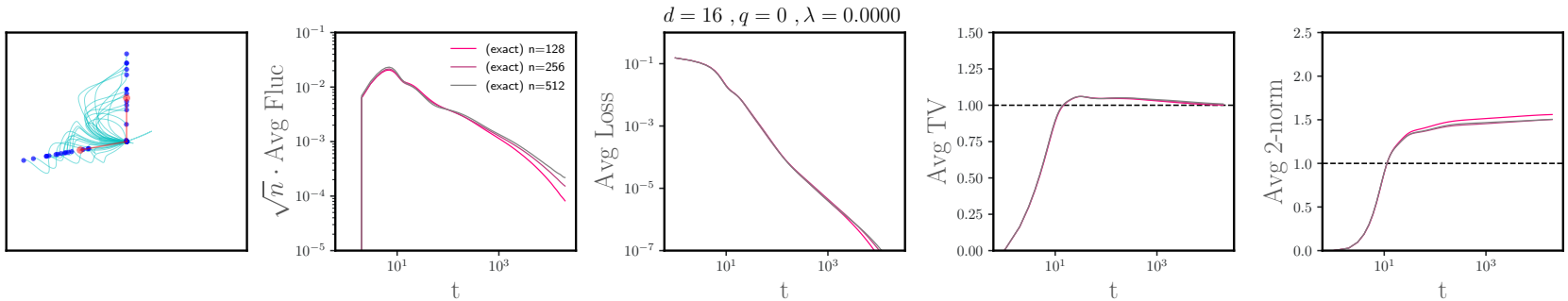
Some context:

1. Expect Monte Carlo type error (scales $n^{-1/2}$)
2. No explicit dependence on the dimension
3. Bounds on long time behavior?

Long time behavior of the fluctuations

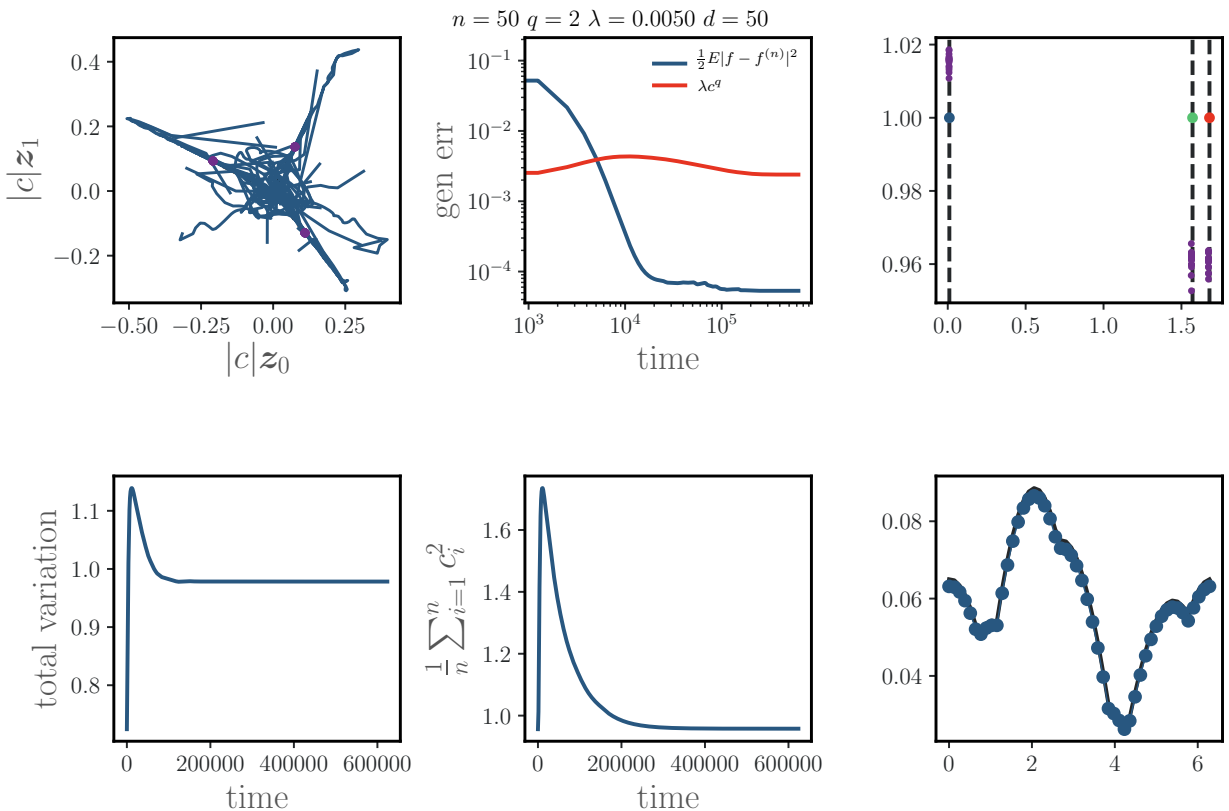
$$\lim_{t\to\infty}\lim_{n\to\infty} n\mathbb{E}_{\text{in}}\int_{\Omega}|f_t^{(n)}(\boldsymbol{x})-f_t(\boldsymbol{x})|^2\nu(d\boldsymbol{x}) \leq \int_D K(\boldsymbol{\theta},\boldsymbol{\theta})\mu^*(d\boldsymbol{\theta}) - \int_{\Omega}|f(\boldsymbol{x})|^2\nu(\boldsymbol{x}) \tag{32}$$

Goes to zero for functions in $\mathcal{F}_1$!



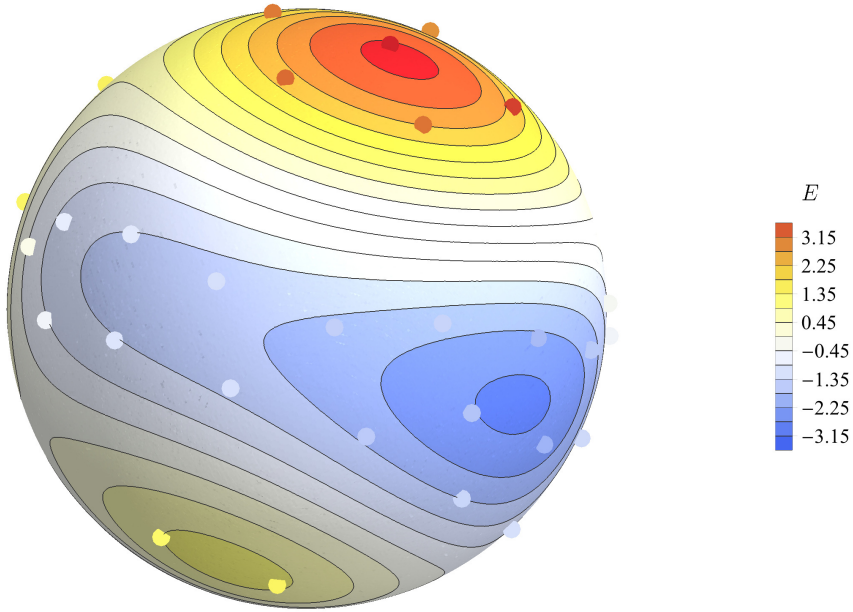
Analytically tractable population loss for uniform data on sphere.

$$f(\boldsymbol{x}) = \frac{1}{n_T} \sum_{i=1}^{n_T} c_i [\boldsymbol{x} \cdot \boldsymbol{z}_i]_+$$

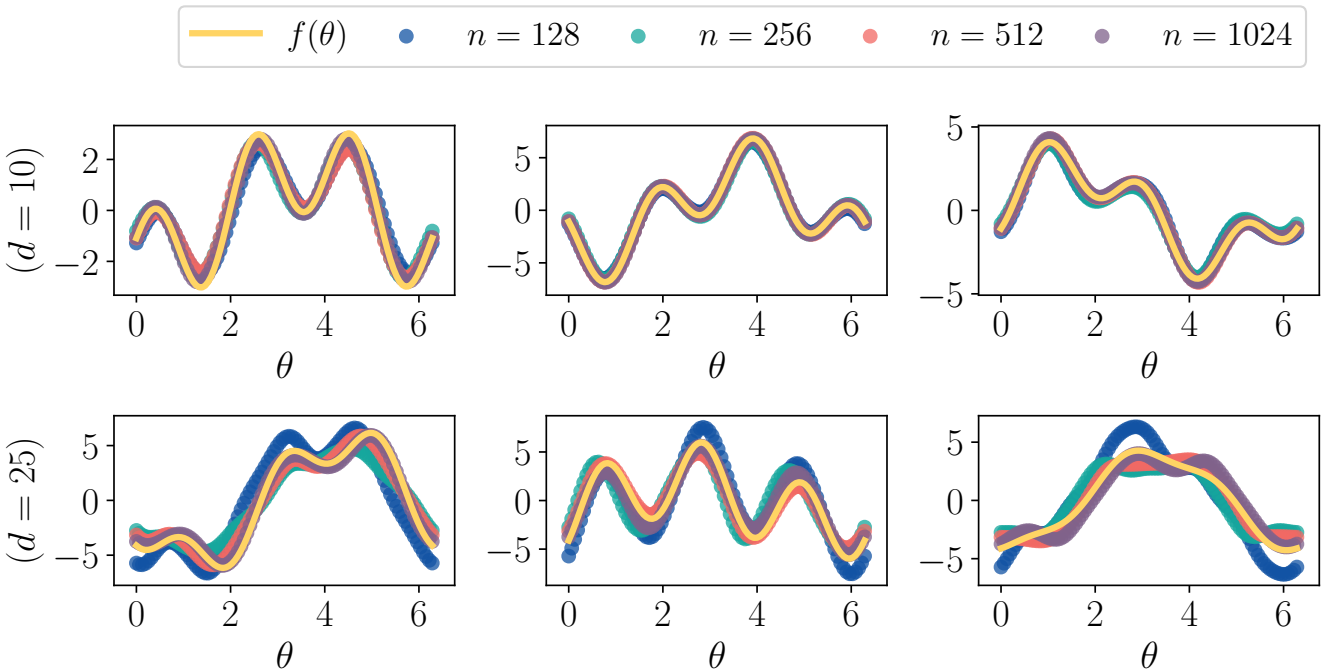


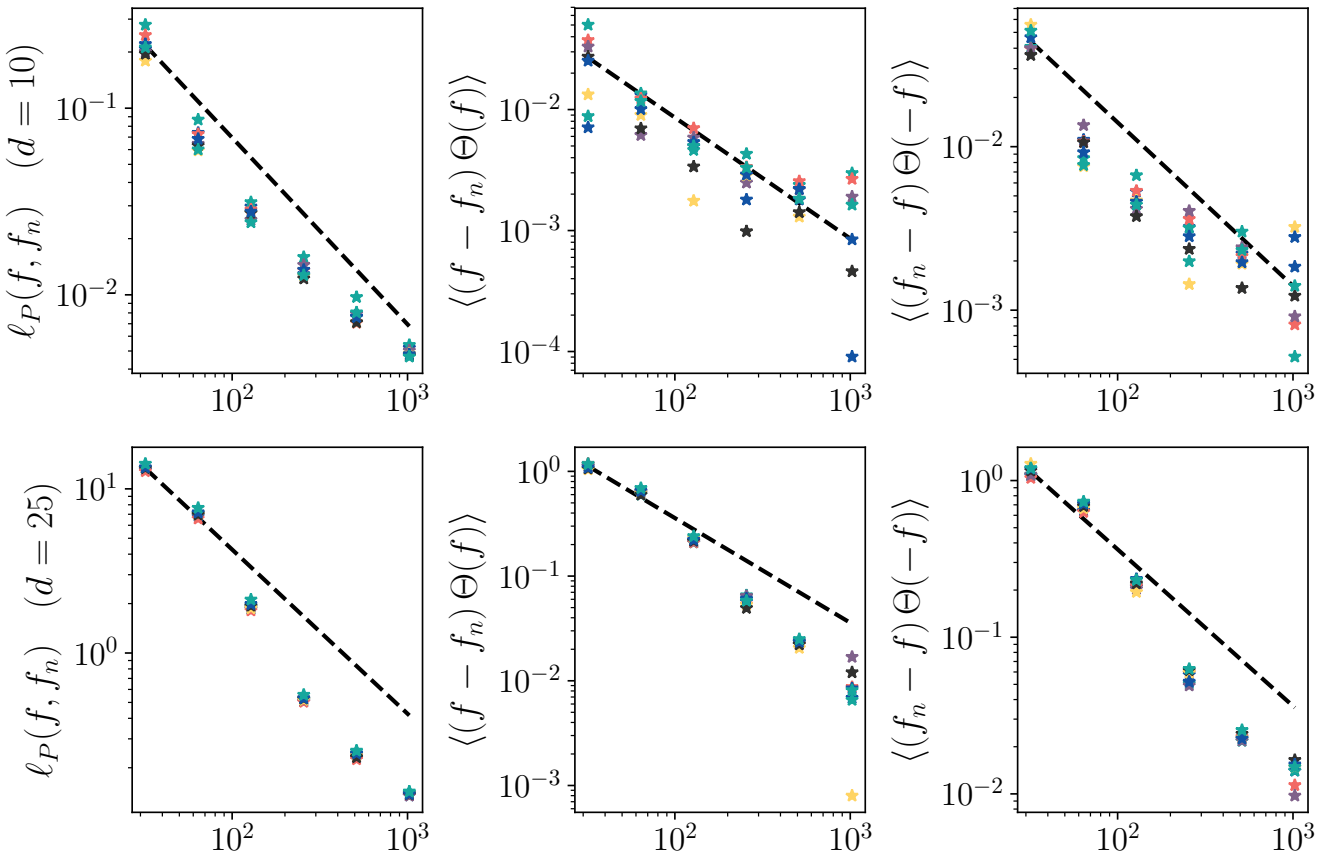
$$f : S^{d-1}(\sqrt{(d)}) \to \mathbb{R}; \quad x \mapsto d^{-1} \sum_{p,q,r}^{d} a_{p,q,r} x_p x_q x_r$$

For a_{ijk} Gaussian, we get a complicated, "rugged" function in high



p-spin models




$\ell_P(f, f_n)$ $(d = 10)$
$\langle (f - f_n)\,\Theta(f) \rangle$
$\langle (f_n - f)\,\Theta(-f) \rangle$
$\ell_P(f, f_n)$ $(d = 25)$
$\langle (f - f_n)\,\Theta(f) \rangle$
$\langle (f_n - f)\,\Theta(-f) \rangle$

Nonlocal algorithms

SGD is a fundamentally *local* algorithm. What if we could teleport mass nonlocally?

Easy at the level of the PDE…Birth-Death dynamics:

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V) - \alpha V \mu_t \tag{33}$$

SGD is a fundamentally *local* algorithm. What if we could teleport mass nonlocally?

Easy at the level of the PDE…Birth-Death dynamics:

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V) - \alpha V \mu_t \tag{33}$$

Conserving population

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V) - \alpha (V - \bar{V}) \mu_t \tag{34}$$

[RJBV19]

We can view this as a change of metric. The PDE

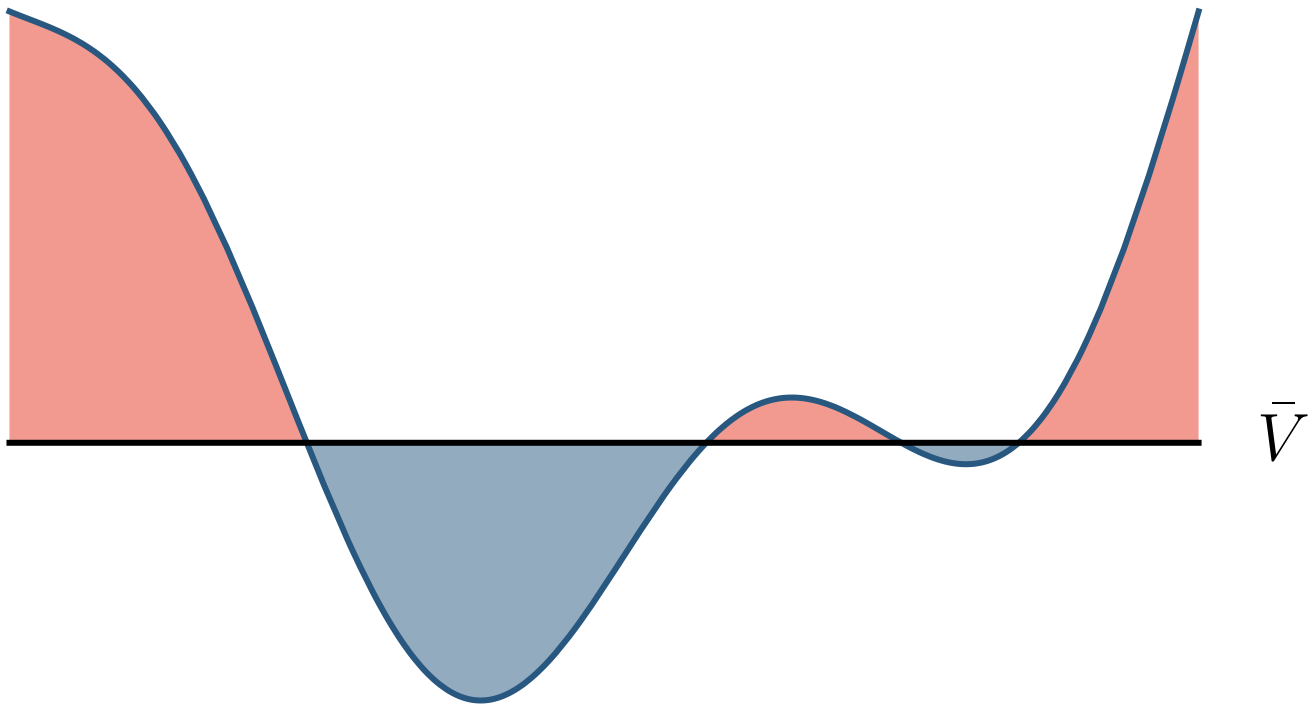
$$\partial_t \mu_t = -\alpha V \mu_t + \alpha \bar{V} \mu_t, \tag{35}$$

corresponds to the proximal update

$$\mu_{k+1} \in \operatorname{argmin} \left(\mathcal{E}[\mu] + (\alpha\tau)^{-1} D_{\mathrm{KL}}(\mu || \mu_k) \right) \tag{36}$$

Splitting scheme between W_2 step and D_{KL} step leads to WFR metric.

Implement the PDE at "particle level" by *killing* and *cloning* particles according to a Markovian dynamics.



Theorem (Global Convergence to Global Minimizers)

Let μ_t denote the solution of the birth-death PDE with initial condition μ_0 with $\operatorname{supp} \mu_0 = D$. *If $\mu_t \rightharpoonup \mu_*$ as $t \to \infty$ for some probability measure $\mu_* \in \mathcal{M}(D)$, then μ_* is a global minimizer of $\mathcal{E}[\mu]$.*

Cf. [RJBV19] for convergence rate, non-interacting case, and detailed proofs.

Proof details

Some comments about the proof:

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V) - \alpha(V - \bar{V})\mu_t \tag{37}$$

1. Any fixed point *must* satisfy $V(\boldsymbol{\theta}, \mu) = \bar{V}(\mu)$ on $\operatorname{supp} \mu$
2. Still must check that $V \geq \bar{V}$ outside $\operatorname{supp} \mu$!

Straightforward (but not simple) argument by contradiction.

Energy decay rate

$E(t) = \mathcal{E}[\mu(t)]$ satisfies

$$\dot{E}(t) = -\int_D |\nabla V(\boldsymbol{\theta}, [\mu_t])|^2 \mu_t(d\boldsymbol{\theta})$$

$$-\alpha \int_D \left(V(\boldsymbol{\theta}, [\mu_t]) - \bar{V}[\mu_t]\right)^2 \mu_t(d\boldsymbol{\theta}) \leq 0.$$

That rate of decrease is *faster* than plain GD (with a caveat!)

Theorem (Asymptotic Convergence Rate)

Same setup as for global convergence, $\exists C > 0$ *and* $t_C > 0$ *such that* $E(t) = \mathcal{E}[\mu_t] - \mathcal{E}[\mu_*] \geq 0$ *satisfies*

$$E(t) \leq Ct^{-1} \quad \textit{if} \quad t \geq t_C \tag{38}$$

Theorem (Asymptotic Convergence Rate)

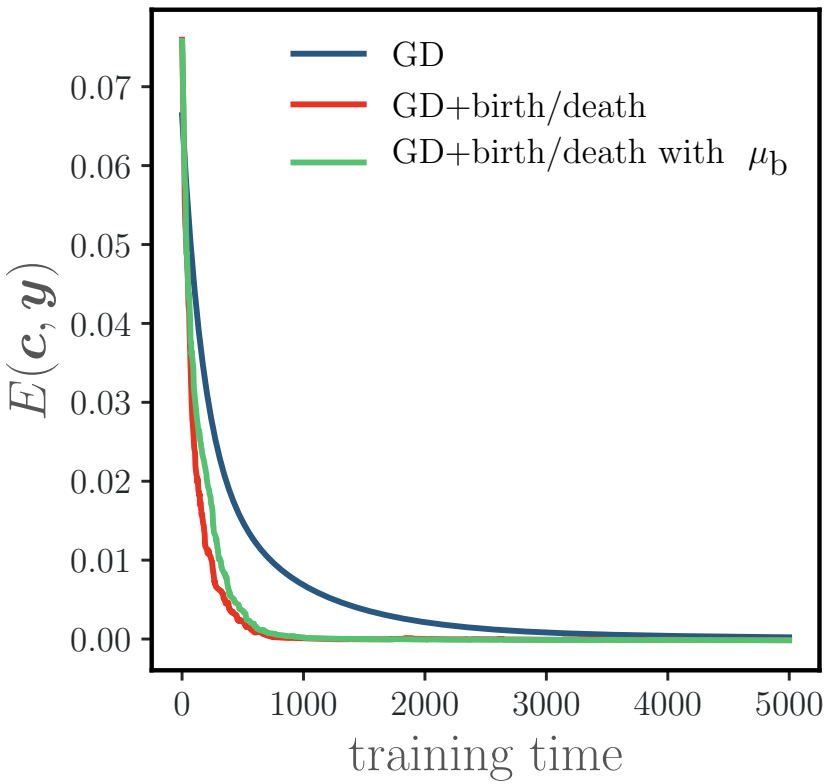
Same setup as for global convergence, $\exists C > 0$ *and* $t_C > 0$ *such that* $E(t) = \mathcal{E}[\mu_t] - \mathcal{E}[\mu_*] \geq 0$ *satisfies*

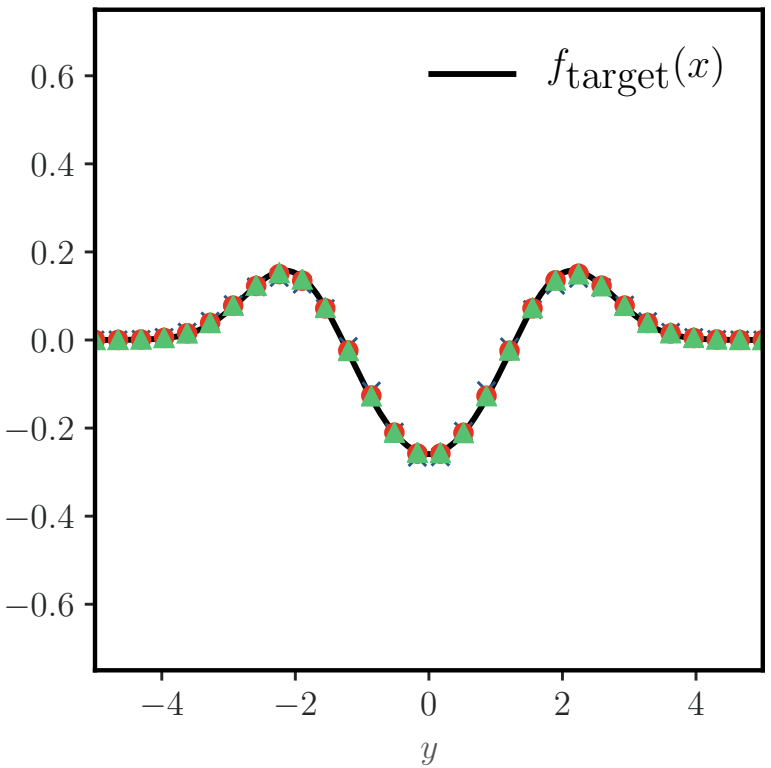
$$E(t) \leq Ct^{-1} \quad \textit{if} \quad t \geq t_C \tag{38}$$

NB: C is not explicit, and could scale poorly.

Simple example: mixtures of gaussians

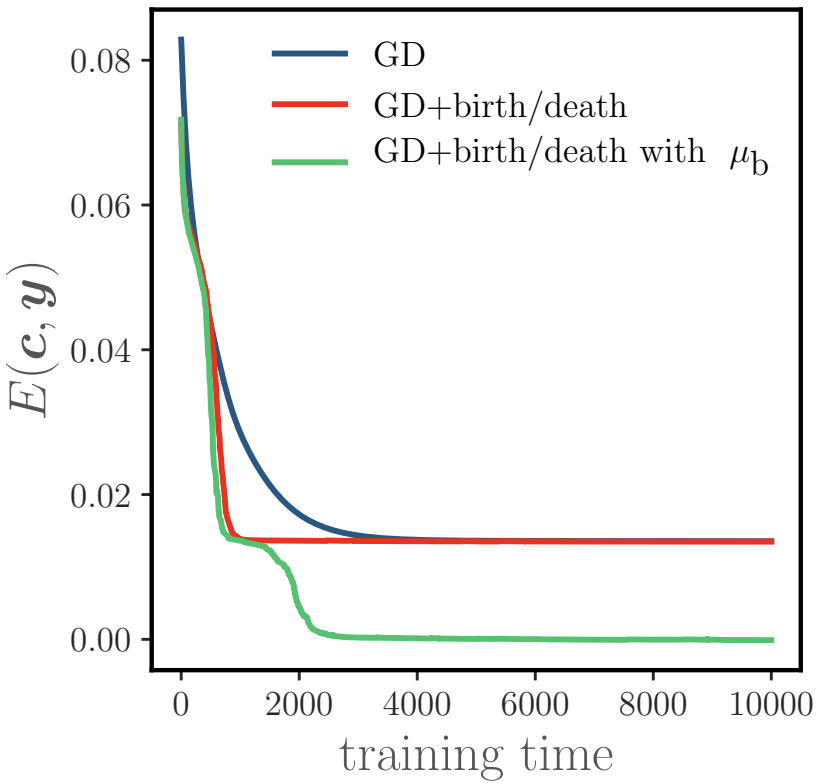
Empirically, convergence properties are good!

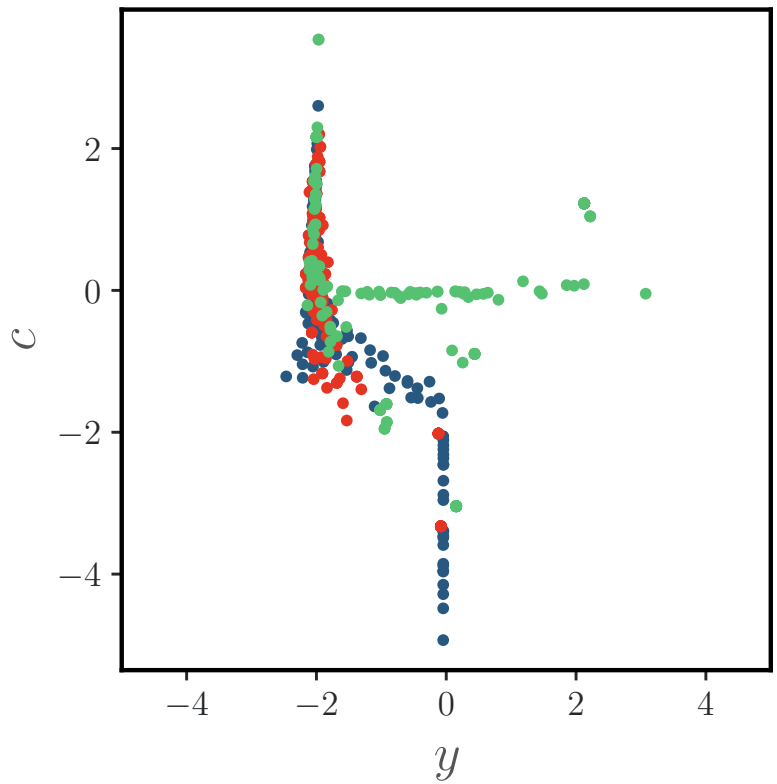




Expanding the support with a "birth" measure

Can ensure full support by reinjecting particles according to some measure with full support.





Open problems

Some things to ponder

1. Do *generic* deep architectures admit a mean-field limit?
2. How does depth change the approximation error?
3. ResNets have a clear mean-field limit—what is the function class they form?
4. Are there other metrics that aid convergence?
5. Dynamical strategies for avoiding stationary points?

Thanks to:

- Eric Vanden-Eijnden (Courant)
- Joan Bruna (NYU Center for Data Science)
- Samy Jelassi (Princeton)
- Zhengdao Chen (Courant)
- James S. McDonnell Foundation
- Stanford University, Terman Fellowship

References I

[AH18] Dario Amodei and Danny Hernandez. AI and Compute, May 2018.

[Bac17] Francis Bach. Breaking the Curse of Dimensionality with Convex Neural Networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[Bar93] A R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, August 2019.

[CB18a] Lenaic Chizat and Francis Bach. A Note on Lazy Training in Supervised Differentiable Programming. page 19, 2018.

[CB18b] Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. *arXiv:1805.09545 [cs, math, stat]*, May 2018.

[Cyb89] G Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2(4):303–314, December 1989.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A Mean Field View of the Landscape of Two-Layers Neural Networks. *arXiv*, April 2018.

[RJBV19] Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. *arXiv:1902.01843 [cs, stat]*, February 2019.

[RV18] Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and Accuracy of Neural Networks: An Interacting Particle System Approach. *arXiv:1805.00915 [cond-mat, stat]*, 2018.

[SS18] Justin Sirignano and Konstantinos Spiliopoulos. Mean Field Analysis of Neural Networks. *arXiv*, May 2018.