Dynamical phenomena in nonlinear learning

Andrea Montanari

Stanford University

August 13, 2025

Michael Celentano

Chen Cheng

Pierfrancesco Urbani

Where we left yesterday

Two-layer neural networks ($m \equiv N$)

$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)\,, \quad \theta = ((a_i)_{i\le m}, (w_i)_{i\le m}) \in \mathbb{R}^m \times (\mathbb{S}^{d-1})^m\,.$$

- Square loss train error:

$$\widehat{\mathscr{R}}_n(\theta) := \frac{1}{2n}\sum_{i=1}^{n} \big(y_i - f(x_i;\theta)\big)^2\,.$$

- Square loss test error:

$$\mathscr{R}(\theta) := \frac{1}{2n}\mathbb{E}\big\{\big(y - f(x;\theta)\big)^2\big\}\,.$$

Two-layer neural networks $(m \equiv N)$

$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)\,, \quad \theta = ((a_i)_{i\le m}, (w_i)_{i\le m}) \in \mathbb{R}^m \times (\mathbb{S}^{d-1})^m\,.$$

- Square loss train error:

$$\widehat{\mathscr{R}}_n(\theta) := \frac{1}{2n}\sum_{i=1}^{n} \big(y_i - f(x_i;\theta)\big)^2\,.$$

- Square loss test error:

$$\mathscr{R}(\theta) := \frac{1}{2n}\mathbb{E}\big\{\big(y - f(x;\theta)\big)^2\big\}\,.$$

Single index model

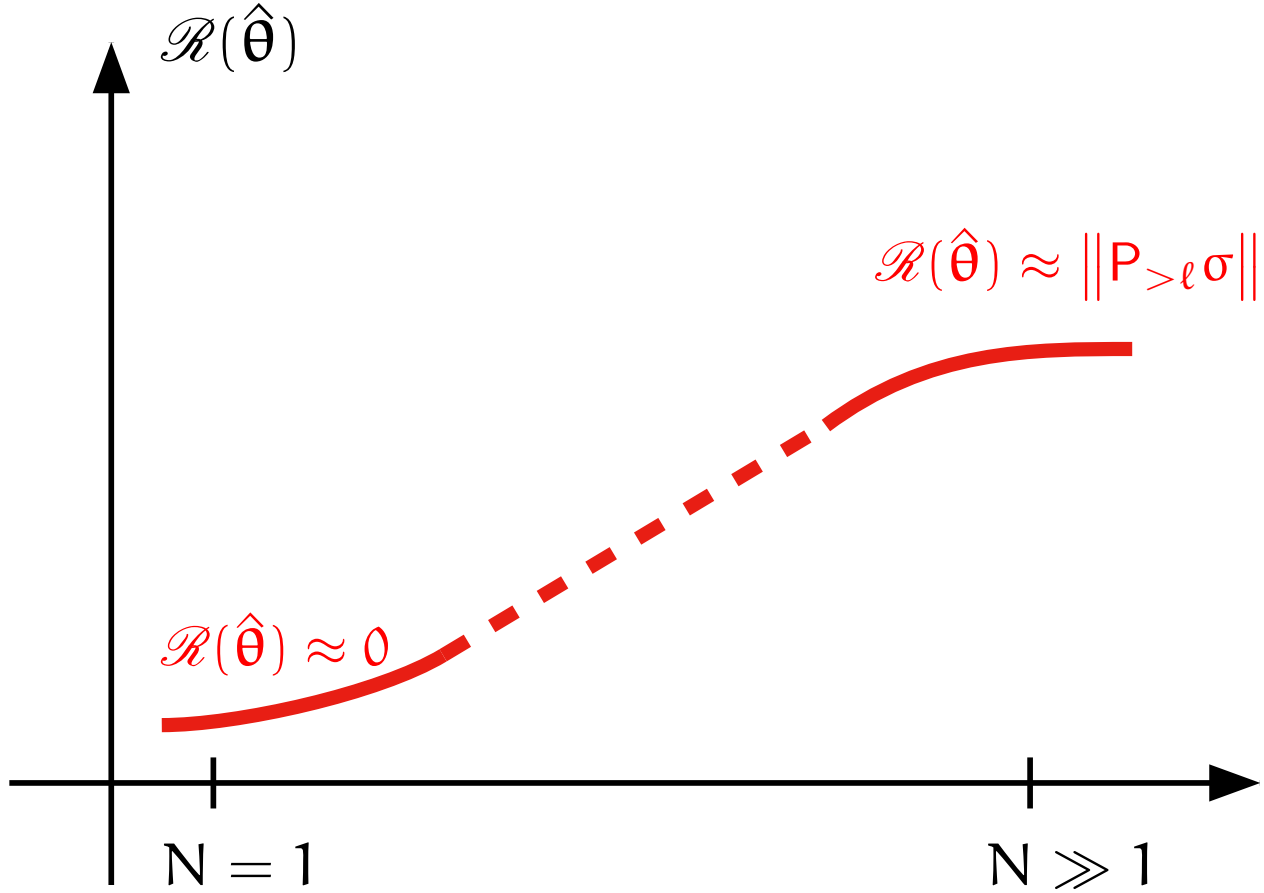
Data $\{(x_i, y_i) : i \le n\}$ iid, $\varepsilon_i \sim N(0, \tau^2)$

$$x_i \sim N(0, I_d), \quad y_i = \varphi(\langle w_*, x_i \rangle) + \varepsilon_i$$

Two theories:

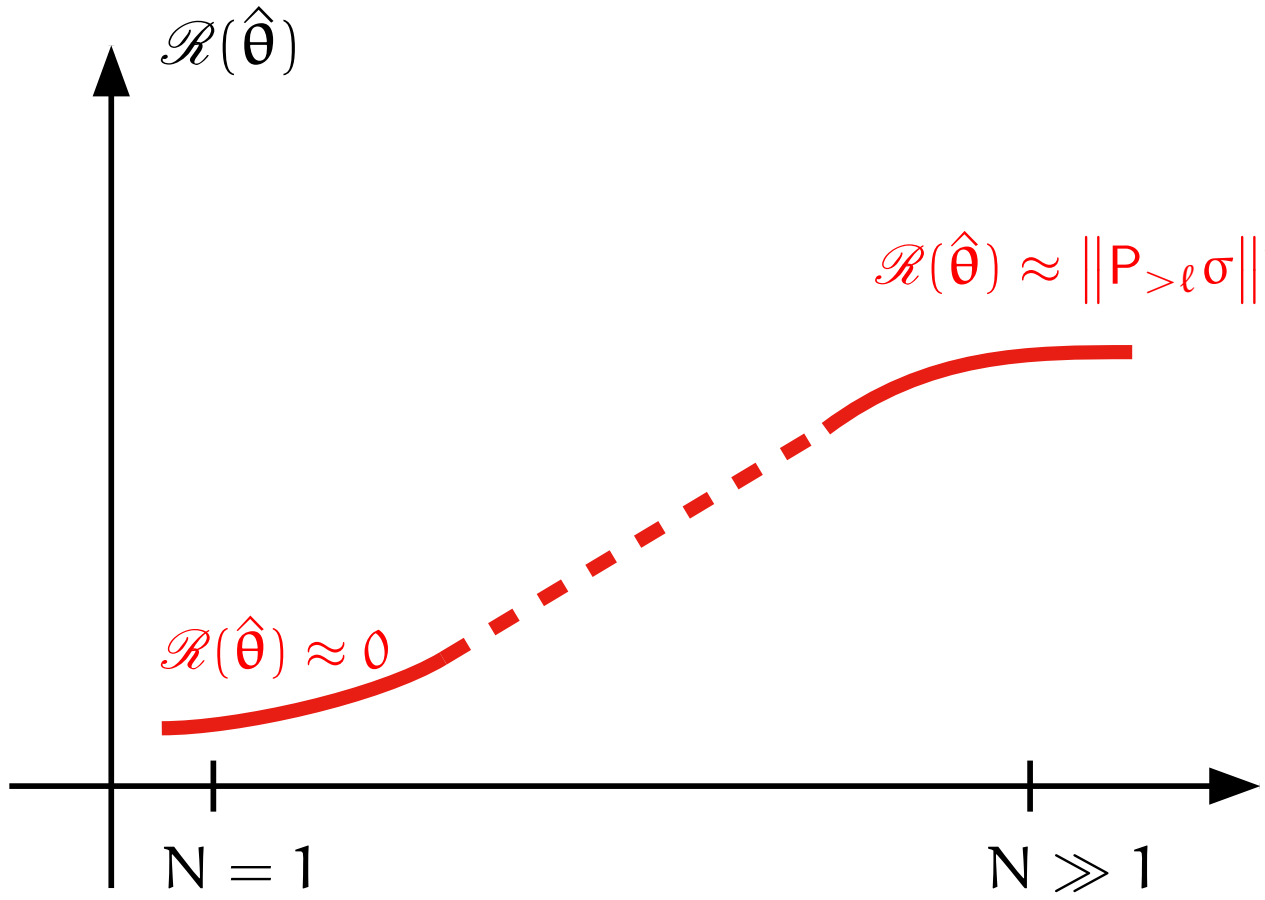
- Feature learning:
  - Weights w_i align quickly with w_*
  - No overfitting
- Neural Tangent:
  - Weights change minimally
  - Model is linear in the y_i (kernel method)
  - Overfitting/interpolation

A cartoon



$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)\,, \qquad a_i^0 = \pm\gamma\sqrt{m}$$

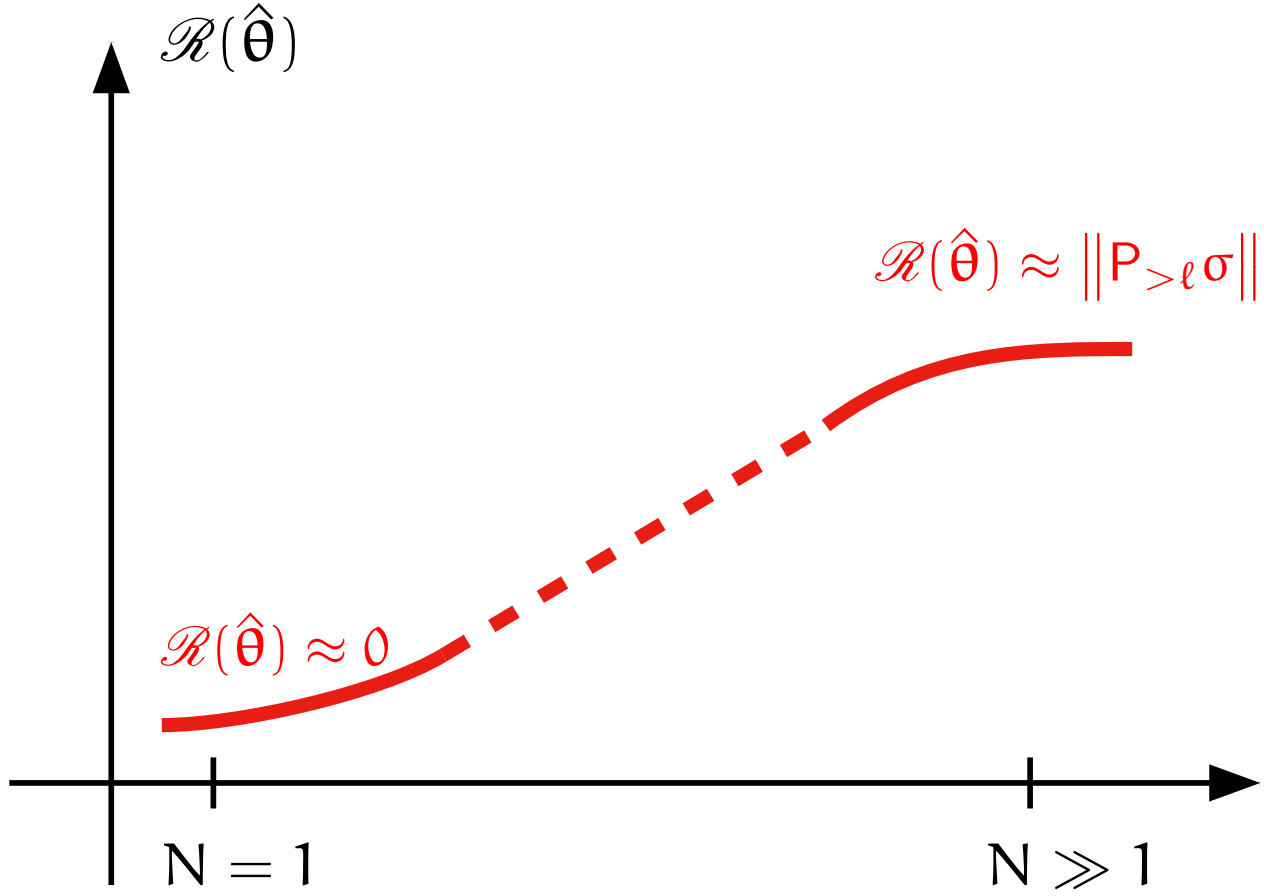
A cartoon



Are large networks necessarily bad?

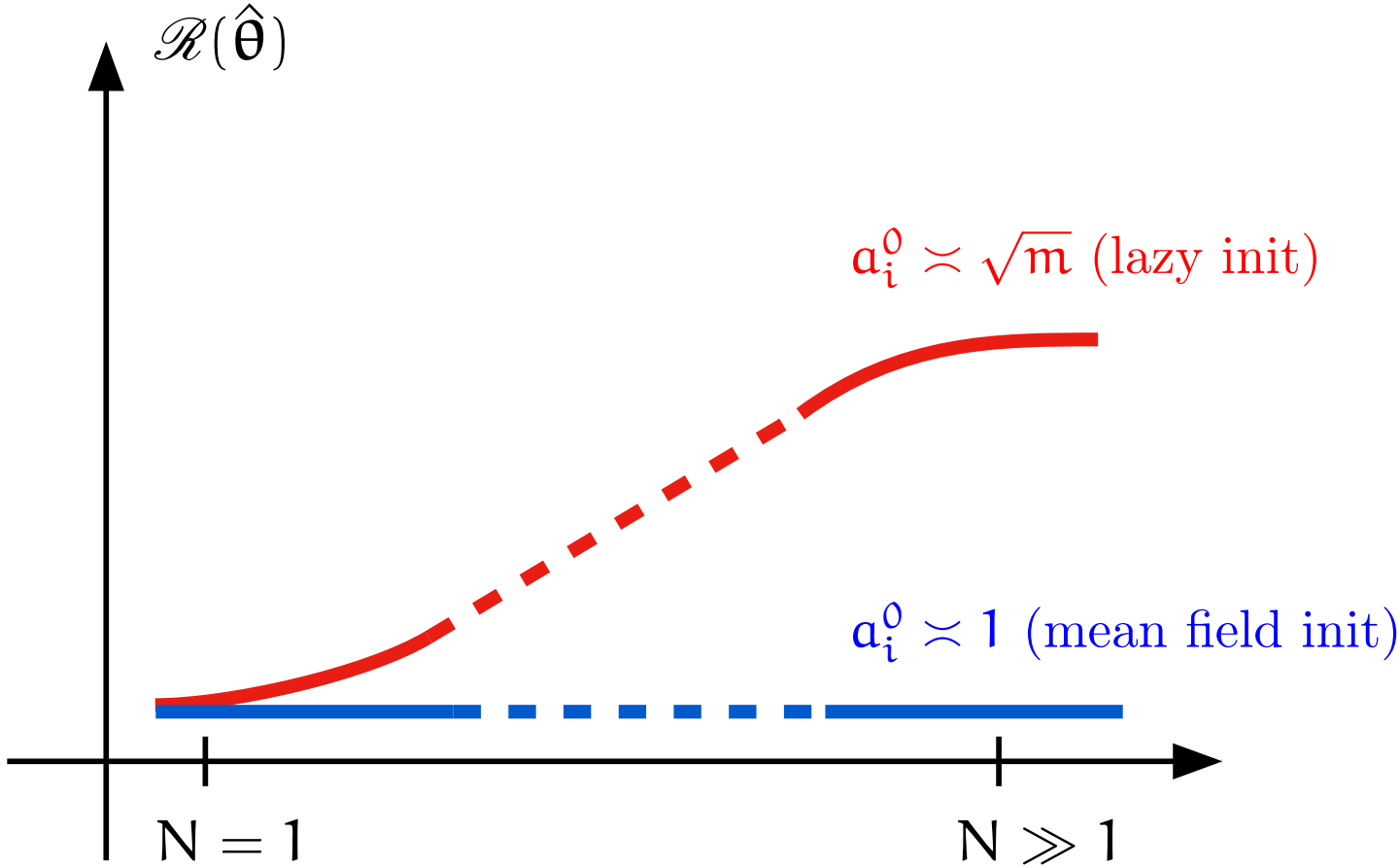
Spoiler

Spoiler



$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)\,, \qquad a_i^0 = \pm\gamma\sqrt{m}$$

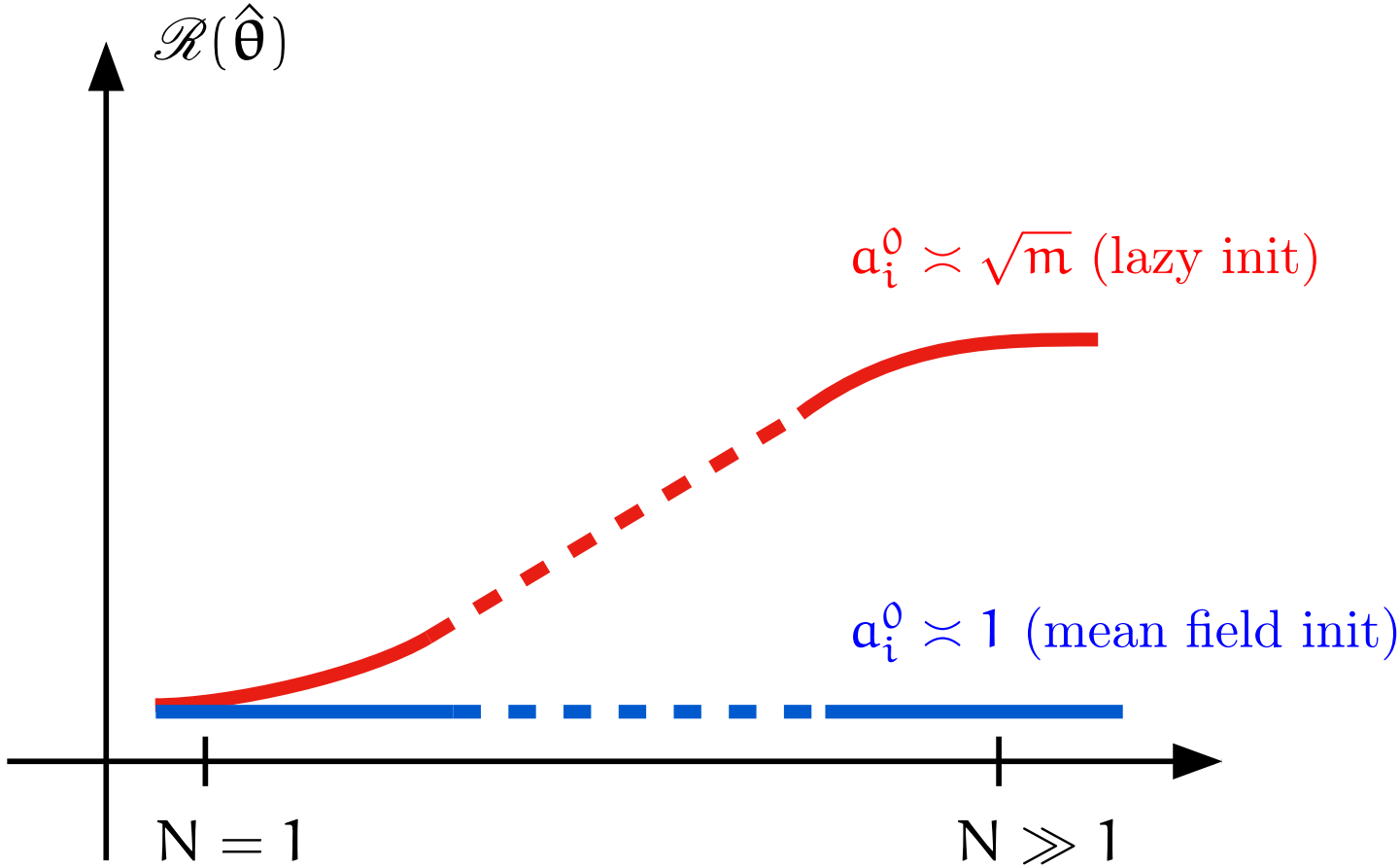
Spoiler



Good generalization: mean field initialization + early stopping

Lazy initialization is popular among practitioners.

Spoiler



Good generalization: mean field initialization + early stopping

---

Lazy initialization is popular among practitioners.

Key control parameters (large n, m, d)

t Training time,

$\alpha = \frac{n}{md}$ Overparametrization ratio,

a^0 Scale of 2nd layer weights at $t = 0$.

Questions

Q1. For which region of α, a^0 does grad flow converge $\widehat{R}_n \approx 0$?

Q2. Does the selected model provide good generalization?

Q3. When feature-learning/no-overfitting vs lazy-training/overfitting?

Q4. How does the generalization error depend on network size and number of iterations?

Approach:

- Dynamical mean field theory (DMFT): $n, d \to \infty$, $n/d \to \overline{\alpha}$.
- Take limit $m, \overline{\alpha} \to \infty$, with $\overline{\alpha}/m \to \alpha$.

Questions

Q1. For which region of α, a^0 does grad flow converge $\widehat{R}_n \approx 0$?

Q2. Does the selected model provide good generalization?

Q3. When feature-learning/no-overfitting vs lazy-training/overfitting?

Q4. How does the generalization error depend on network size and number of iterations?

Approach:

- Dynamical mean field theory (DMFT): $n, d \to \infty$, $n/d \to \overline{\alpha}$.
- Take limit $m, \overline{\alpha} \to \infty$, with $\overline{\alpha}/m \to \alpha$.

Questions

Q1. For which region of α, a^0 does grad flow converge $\widehat{R}_n \approx 0$?

Q2. Does the selected model provide good generalization?

Q3. When feature-learning/no-overfitting vs lazy-training/overfitting?

Q4. How does the generalization error depend on network size and number of iterations?

Approach:

- Dynamical mean field theory (DMFT): $n, d \to \infty$, $n/d \to \overline{\alpha}$.
- Take limit $m, \overline{\alpha} \to \infty$, with $\overline{\alpha}/m \to \alpha$.

Questions

Q1. For which region of α, a^0 does grad flow converge $\widehat{R}_n \approx 0$?

Q2. Does the selected model provide good generalization?

Q3. When feature-learning/no-overfitting vs lazy-training/overfitting?

Q4. How does the generalization error depend on network size and number of iterations?

Approach:

- Dynamical mean field theory (DMFT): $n, d \to \infty$, $n/d \to \overline{\alpha}$.
- Take limit $m, \overline{\alpha} \to \infty$, with $\overline{\alpha}/m \to \alpha$.

Questions

Q1. For which region of α, a^0 does grad flow converge $\widehat{R}_n \approx 0$?

Q2. Does the selected model provide good generalization?

Q3. When feature-learning/no-overfitting vs lazy-training/overfitting?

Q4. How does the generalization error depend on network size and number of iterations?

Approach:

- Dynamical mean field theory (DMFT): $n, d \to \infty$, $n/d \to \overline{\alpha}$.
- Take limit $m, \overline{\alpha} \to \infty$, with $\overline{\alpha}/m \to \alpha$.

Outline

Rigorous DMFT

A slightly more general setting

$\theta, \theta_* \in \mathbb{R}^{d\times m}$, regularization $t \mapsto \Lambda_t \in \mathbb{R}^{m\times m}$

$$\widehat{\mathscr{R}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} L\left(\theta^T x_i; \theta_*^T x_i, \varepsilon_i\right),$$

$$\dot{\theta}_t = -\theta_t \Lambda_t^T - \nabla \widehat{\mathscr{R}}_n(\theta_t).$$

The case of 2-layer neural nets, k–index data:

$$\theta = \left[w_1 \middle| w_2 \middle| \cdots \middle| w_m\right], \quad \theta_* = \left[w_{*,1} \middle| \cdots \middle| w_k \middle| 0 \middle| \cdots \middle| 0\right],$$

$$L(r, u, \varepsilon) = \frac{1}{2}\left(\varphi(u) + \varepsilon - \frac{1}{m}\sum_{i=1}^{m} a_i \sigma(r_i)\right)^2.$$

A slightly more general setting

$\theta, \theta_* \in \mathbb{R}^{d\times m}$, regularization $t \mapsto \Lambda_t \in \mathbb{R}^{m\times m}$

$$\widehat{\mathscr{R}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} L\big(\theta^T x_i; \theta_*^T x_i, \varepsilon_i\big)\,,$$
$$\dot{\theta}_t = -\theta_t \Lambda_t^T - \nabla \widehat{\mathscr{R}}_n(\theta_t)\,.$$

The case of 2-layer neural nets, k−index data:

$$\theta = \Big[w_1 \Big| w_2 \Big| \cdots \Big| w_m\Big], \quad \theta_* = \Big[w_{*,1} \Big| \cdots \Big| w_k \Big| 0 \Big| \cdots \Big| 0\Big]\,,$$

$$L(r, u, \varepsilon) = \frac{1}{2}\left(\varphi(u) + \varepsilon - \frac{1}{m}\sum_{i=1}^{m} a_i \sigma(r_i)\right)^2\,.$$

A slightly more general setting

$\theta, \theta_* \in \mathbb{R}^{d\times m}$, regularization $t \mapsto \Lambda_t \in \mathbb{R}^{m\times m}$

$$\widehat{\mathscr{R}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} L\big(\theta^T x_i; \theta_*^T x_i, \varepsilon_i\big)\,,$$

$$\dot\theta_t = -\theta_t \Lambda_t^T - \nabla \widehat{\mathscr{R}}_n(\theta_t)\,.$$

DMFT process: $\vartheta_t, r_t \in \mathbb{R}^m$,

$$\dot\vartheta^t = -(\Lambda^t + \Gamma^t)\vartheta^t - \int_0^t R_\ell(t,s)\vartheta^s ds - R_\ell(t,*)\vartheta^* + u^t\,,$$

$$r^t = -\frac{1}{\overline{\alpha}}\int_0^t R_\theta(t,s)\nabla_r L(r^s, w^*; \varepsilon) ds + w^t\,,$$

$$u^t \sim \mathsf{GP}(0, C_\ell/\delta)\,,\quad w^t \sim \mathsf{GP}(0, C_\theta)\,.$$

where $\Gamma_t, C_\ell(t,s), C_\theta(t,s), R_\ell(t,s), R_\ell(t,*), R_\theta(t,s) \in \mathbb{R}^{m\times m}$ are unique solution of ...

Assumptions

- $\mathbf{X}$ has standardized sub-Gaussian independent entries
- ℓ is C^3 with bounded 2-nd, 3-rd derivatives.
- $n, d \to \infty$, $n/d \to \overline{\alpha}$
- $\theta_{0,i}$, $\theta_{*,i}$, i-th rows of θ_0, θ_*:

$$\frac{1}{d}\sum_{i=1}^{d} \delta_{\sqrt{d}\theta_{0,i},\sqrt{d}\theta_{*,i}} \overset{W_2}{\Rightarrow} \mathsf{Law}(\vartheta_0, \vartheta_*)\,, \quad \frac{1}{n}\sum_{i=1}^{n} \delta_{\varepsilon_i} \overset{W_2}{\Rightarrow} \mathsf{Law}(\varepsilon)\,.$$

DMFT characterization

Theorem (Celentano, Cheng, M, 2021)

Under the above assumptions, for any T *and any continuous bounded* $\psi : \mathbb{R}^m \times C([0,T],\mathbb{R}^m) \to \mathbb{R}$ *we have*

$$\operatorname*{p\text{-}lim}_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi\big(\sqrt{d}\theta_i^*, \sqrt{d}(\theta_i)_0^T\big) = \mathbb{E}\{\psi\big(\vartheta_{*,i}, (\vartheta_i)_0^T\big\}\,.$$

- **Informally:** The DMFT process ϑ_t characterizes the distribution of rows of θ_t.
- **Application to 2-layer nets:** $w_i(t)$: First layer weights in a 2-layer network

$$\operatorname*{p\text{-}lim}_{t,s} \langle w_i(t), w_j(s)\rangle = C_{ij}(t,s)\,,$$

where the C_{ij} solve DMFT equations.

DMFT characterization

Theorem (Celentano, Cheng, M, 2021)

Under the above assumptions, for any T *and any continuous bounded* $\psi : \mathbb{R}^m \times C([0,T],\mathbb{R}^m) \to \mathbb{R}$ *we have*

$$\operatorname*{p\text{-}lim}_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi\big(\sqrt{d}\theta_i^*, \sqrt{d}(\theta_i)_0^\top\big) = \mathbb{E}\{\psi\big(\vartheta_{*,i}, (\vartheta_i)_0^\top\big\}\,.$$

- **Informally:** The DMFT process ϑ_t characterizes the distribution of rows of θ_t.
- **Application to 2-layer nets:** $\boldsymbol{w}_i(t)$: First layer weights in a 2-layer network

$$\operatorname*{p\text{-}lim}_{t,s} \langle \boldsymbol{w}_i(t), \boldsymbol{w}_j(s)\rangle = C_{ij}(t,s)\,,$$

where the C_{ij} solve DMFT equations.

Application to 2-layer nets

$$C_{ij}^n(t,s) := \langle \boldsymbol{w}_i(t), \boldsymbol{w}_j(s) \rangle \,, \quad v_i^n(t) := \langle \boldsymbol{w}_i(t), \boldsymbol{w}_* \rangle \,, \quad a_i^n(t) \,.$$

Theorem (Celentano, Cheng, M, 2021)

As $n, d \to \infty$, $n/d \to \overline{\alpha}$, *uniformly over compacts,*

$$C_{ij}^n \xrightarrow{p} C_{ij} \,, \quad v_i^n \xrightarrow{p} v_i, \quad a_i^n \xrightarrow{p} a_i \,.$$

Further $\boldsymbol{C} = (C_{ij} : i, j \le m)$, $\boldsymbol{v} = (v_i : i \le m)$, $\boldsymbol{a} = (a_i : i \le m)$ *are deterministic and the unique solution of DMFT equations.*

Challenge: Solving the DMFT equations

$$\frac{\mathrm{d}}{\mathrm{dt}}\vartheta^t = -(\Lambda^t + \Gamma^t)\vartheta^t - \int_0^t R_\ell(t,s)\vartheta^s \mathrm{d}s - R_\ell(t,*)\vartheta^* + u^t\,, \quad u^t \sim \mathsf{GP}(0, C_\ell/\delta)\,,$$

$$r^t = -\frac{1}{\delta}\int_0^t R_\vartheta(t,s)\nabla L(r^s, w^*; z)\mathrm{d}s + w^t\,, \qquad w^t \sim \mathsf{GP}(0, C_\theta)\,,$$

$$R_\vartheta(t,s) = \mathbb{E}\left[\frac{\partial \vartheta^t}{\partial u^s}\right]\,, \qquad 0 \le s \le t < \infty\,,$$

$$R_\ell(t,s) = \mathbb{E}\left[\frac{\partial \nabla L(r^t, w^*; z)}{\partial w^s}\right]\,, \qquad 0 \le s < t < \infty\,,$$

$$R_\ell(t,*) = \mathbb{E}\left[\frac{\partial \nabla L(r^t, w^*; z)}{\partial w^*}\right]\,,$$

$$\Gamma^t = \mathbb{E}\left[\nabla^2 L(r^t, w^*; z)\right]\,,$$

$$C_\theta(t,s) = \mathbb{E}\left[\vartheta^t \vartheta^{s\,\mathsf{T}}\right]\,, \qquad 0 \le s \le t < \infty \text{ or } s = *\,,$$

$$C_\ell(t,s) = \mathbb{E}\left[\nabla L(r^t, w^*; z)\nabla L(r^s, w^*; z)^\mathsf{T}\right]\,, \qquad 0 \le s \le t < \infty\,.$$

Evaluating numerically the equations requires to compute expectation wrt the distribution of the processes ϑ, r.

Challenge: Solving the DMFT equations

$$\frac{\mathrm{d}}{\mathrm{dt}}\vartheta^t = -(\Lambda^t + \Gamma^t)\vartheta^t - \int_0^t R_\ell(t,s)\vartheta^s \mathrm{d}s - R_\ell(t,*)\vartheta^* + u^t\,, \quad u^t \sim \mathsf{GP}(0, C_\ell/\delta)\,,$$

$$r^t = -\frac{1}{\delta}\int_0^t R_\vartheta(t,s)\nabla L(r^s, w^*; z)\mathrm{d}s + w^t\,, \qquad w^t \sim \mathsf{GP}(0, C_\theta)\,,$$

$$R_\vartheta(t,s) = \mathbb{E}\left[\frac{\partial \vartheta^t}{\partial u^s}\right]\,, \qquad 0 \le s \le t < \infty\,,$$

$$R_\ell(t,s) = \mathbb{E}\left[\frac{\partial \nabla L(r^t, w^*; z)}{\partial w^s}\right]\,, \qquad 0 \le s < t < \infty\,,$$

$$R_\ell(t,*) = \mathbb{E}\left[\frac{\partial \nabla L(r^t, w^*; z)}{\partial w^*}\right]\,,$$

$$\Gamma^t = \mathbb{E}\left[\nabla^2 L(r^t, w^*; z)\right]\,,$$

$$C_\theta(t,s) = \mathbb{E}\left[\vartheta^t \vartheta^{s\,\mathsf{T}}\right]\,, \qquad 0 \le s \le t < \infty \text{ or } s = *\,,$$

$$C_\ell(t,s) = \mathbb{E}\left[\nabla L(r^t, w^*; z)\nabla L(r^s, w^*; z)^\mathsf{T}\right]\,, \qquad 0 \le s \le t < \infty\,.$$

Evaluating numerically the equations requires to compute expectation wrt the distribution of the processes ϑ, r.

Gaussian approximation

Take notice:

- Not Gaussian process regression.
- Not approximating ϑ_t, r_t by Gaussian processes.

Idea: Matching Gaussian model

$$\widehat{\mathscr{R}}_n(\theta) := \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - f(x_i;\theta)\big)^2 \quad = \frac{1}{2n}\|F(\theta)\|^2 .$$

- Replace $F(\theta)$ by Gaussian proc. $F^g(\theta)$ with matching mean and covariance.
- Study DMFT for this model.
- A piece of the covariance:

$$\mathbb{E}\{f(x;\theta_1)f(x;\theta_2)\} = \frac{1}{m^2}\sum_{i,j=1}^{m} h(w_i^\mathsf{T} w_j)\, a_i a_j .$$

Idea: Matching Gaussian model

$$\widehat{\mathscr{R}}_n(\theta) := \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - f(x_i;\theta)\big)^2 \quad = \frac{1}{2n}\|F(\theta)\|^2 .$$

- Replace $F(\theta)$ by Gaussian proc. $F^g(\theta)$ with matching mean and covariance.
- Study DMFT for this model.
- A piece of the covariance:

$$\mathbb{E}\{f(x;\theta_1)f(x;\theta_2)\} = \frac{1}{m^2}\sum_{i,j=1}^{m} h(w_i^\mathsf{T} w_j)\, a_i a_j .$$

DMFT eqs for Gaussian model are explicit

$$\frac{\mathrm{d}a}{\mathrm{d}t}(t) = \frac{\overline{\alpha}}{m}\hat{\varphi}(\boldsymbol{v}(t))\int_0^t R_A(t,s)\,\mathrm{d}s - \frac{\overline{\alpha}}{m}\int_0^t R_A(t,s)a(s)\left[\frac{1}{m}h(C_d(t,s)) + \frac{m-1}{m}h(C_o(t,s))\right]\mathrm{d}s - \frac{\overline{\alpha}}{m}\int_0^t C_A(t,s)a(s)\left[\frac{1}{m}h'(C_d(t,s))R_d(t,s) + \frac{m-1}{m}h'(C_o(t,s))R_o(t,s)\right]\mathrm{d}s\,, \tag{2.32}$$

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}(t) = -\nu(t)\boldsymbol{v}(t) + \frac{\overline{\alpha}}{m}\nabla\hat{\varphi}(\boldsymbol{v}(t))a(t)\int_0^t R_A(t,s)\,\mathrm{d}s - \frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s) + (m-1)M_R^{(o)}(t,s)\right]\boldsymbol{v}(s)\,\mathrm{d}s\,, \tag{2.33}$$

$$\partial_t C_d(t,t') = -\nu(t)C_d(t,t') + \frac{\overline{\alpha}}{m}\langle\nabla\hat{\varphi}'(\boldsymbol{v}(t)),\boldsymbol{v}(t')\rangle a(t)\int_0^t R_A(t,s)\,\mathrm{d}s - \frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s)C_d(t',s) + (m-1)M_R^{(o)}(t,s)C_o(t',s)\right]\mathrm{d}s - \frac{1}{m}\int_0^{t'} \left[M_C^{(d)}(t,s)R_d(t',s) + (m-1)M_C^{(o)}(t,s)R_o(t',s)\right]\mathrm{d}s\,, \tag{2.34}$$

$$\partial_t C_o(t,t') = -\nu(t)C_o(t,t') + \frac{\overline{\alpha}}{m}\langle\nabla\hat{\varphi}(\boldsymbol{v}(t)),\boldsymbol{v}(t')\rangle a(t)\int_0^t R_A(t,s)\,\mathrm{d}s - \frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s)C_o(t',s) + M_R^{(o)}(t,s)C_d(t',s) + (m-2)M_R^{(o)}(t,s)C_o(t',s)\right]\mathrm{d}s - \frac{1}{m}\int_0^{t'} \left[M_C^{(d)}(t,s)R_o(t',s) + M_C^{(o)}(t,s)R_d(t',s) + (m-2)M_C^{(o)}(t,s)R_o(t',s)\right]\mathrm{d}s\,, \tag{2.35}$$

$$\partial_t R_d(t,t') = -\nu(t)R_d(t,t') + \delta(t-t') - \frac{1}{m}\int_{t'}^t \left[M_R^{(d)}(t,s)R_d(s,t') + (m-1)M_R^{(o)}(t,s)R_o(s,t')\right]\mathrm{d}s\,, \tag{2.36}$$

$$\partial_t R_o(t,t') = -\nu(t)R_o(t,t') - \frac{1}{m}\int_{t'}^t \left[M_R^{(d)}(t,s)R_o(s,t') + M_R^{(o)}(t,s)R_d(s,t') + (m-2)M_R^{(o)}(t,s)R_o(s,t')\right]\mathrm{d}s\,. \tag{2.37}$$

(2) Equations for auxiliary functions. The memory kernels $M_R^{(s)}(t,s)$, $M_R^{(o)}(t,s)$ and $M_C^{(s)}(t,s)$, $M_C^{(o)}(t,s)$ are given by:

$$M_R^{(d)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)\left[R_A(t,s)h'(C_d(t,s)) + C_A(t,s)h''(C_d(t,s))R_d(t,s)\right]\,, \tag{2.38}$$

$$M_R^{(o)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)\left[R_A(t,s)h'(C_o(t,s)) + C_A(t,s)h''(C_o(t,s))R_o(t,s)\right]\,, \tag{2.39}$$

$$M_C^{(d)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)C_A(t,s)h'(C_d(t,s))\,, \tag{2.40}$$

$$M_C^{(o)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)C_A(t,s)h'(C_o(t,s))\,. \tag{2.41}$$

Further, $C_A(t,s)$, $R_A(t,s)$ are given by the same equations (2.19), where Σ_C, Σ_R are simplified as follows:

$$\Sigma_C(t,s) = \tau^2 + \|\varphi\|^2 - a(t)\hat{\varphi}(\boldsymbol{v}(t)) - a(s)\hat{\varphi}(\boldsymbol{v}(s)) + \frac{a(t)a(s)}{m}h(C_d(t,s)) + \frac{m-1}{m}a(t)a(s)h(C_o(t,s)) \tag{2.42}$$

$$\Sigma_R(t,s) = \frac{a(t)a(s)}{m}h'(C_d(t,s))R_d(t,s) + \frac{m-1}{m}a(t)a(s)h'(C_o(t,s))R_o(t,s)$$

Highly nonlinear!

DMFT eqs for Gaussian model are explicit

$$\begin{aligned}\frac{\mathrm{d}a}{\mathrm{d}t}(t) =& \frac{\overline{\alpha}}{m}\hat{\varphi}(\boldsymbol{v}(t))\int_0^t R_A(t,s)\,\mathrm{d}s \\ &- \frac{\overline{\alpha}}{m}\int_0^t R_A(t,s)a(s)\left[\frac{1}{m}h(C_d(t,s)) + \frac{m-1}{m}h(C_o(t,s))\right]\mathrm{d}s \\ &- \frac{\overline{\alpha}}{m}\int_0^t C_A(t,s)a(s)\left[\frac{1}{m}h'(C_d(t,s))R_d(t,s) + \frac{m-1}{m}h'(C_o(t,s))R_o(t,s)\right]\mathrm{d}s\,,\end{aligned} \tag{2.32}$$

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}(t) =& -\nu(t)\boldsymbol{v}(t) + \frac{\overline{\alpha}}{m}\nabla\hat{\varphi}(\boldsymbol{v}(t))a(t)\int_0^t R_A(t,s)\,\mathrm{d}s \\ &- \frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s) + (m-1)M_R^{(o)}(t,s)\right]\boldsymbol{v}(s)\,\mathrm{d}s\,,\end{aligned} \tag{2.33}$$

$$\begin{aligned}\partial_t C_d(t,t') =& -\nu(t)C_d(t,t') + \frac{\overline{\alpha}}{m}\langle\nabla\hat{\varphi}'(\boldsymbol{v}(t)),\boldsymbol{v}(t')\rangle a(t)\int_0^t R_A(t,s)\,\mathrm{d}s \\ &- \frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s)C_d(t',s) + (m-1)M_R^{(o)}(t,s)C_o(t',s)\right]\mathrm{d}s \\ &- \frac{1}{m}\int_0^{t'} \left[M_C^{(d)}(t,s)R_d(t',s) + (m-1)M_C^{(o)}(t,s)R_o(t',s)\right]\mathrm{d}s\,,\end{aligned} \tag{2.34}$$

$$\begin{aligned}\partial_t C_o(t,t') =& -\nu(t)C_o(t,t') + \frac{\overline{\alpha}}{m}\langle\nabla\hat{\varphi}(\boldsymbol{v}(t)),\boldsymbol{v}(t')\rangle a(t)\int_0^t R_A(t,s)\,\mathrm{d}s \\ &- \frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s)C_o(t',s) + M_R^{(o)}(t,s)C_d(t',s) + (m-2)M_R^{(o)}(t,s)C_o(t',s)\right]\mathrm{d}s \\ &- \frac{1}{m}\int_0^{t'} \left[M_C^{(d)}(t,s)R_o(t',s) + M_C^{(o)}(t,s)R_d(t',s) + (m-2)M_C^{(o)}(t,s)R_o(t',s)\right]\mathrm{d}s\,,\end{aligned} \tag{2.35}$$

$$\begin{aligned}\partial_t R_d(t,t') =& -\nu(t)R_d(t,t') + \delta(t-t') \\ &- \frac{1}{m}\int_{t'}^t \left[M_R^{(d)}(t,s)R_d(s,t') + (m-1)M_R^{(o)}(t,s)R_o(s,t')\right]\mathrm{d}s\,,\end{aligned} \tag{2.36}$$

$$\begin{aligned}\partial_t R_o(t,t') =& -\nu(t)R_o(t,t') - \frac{1}{m}\int_{t'}^t \left[M_R^{(d)}(t,s)R_o(s,t') + M_R^{(o)}(t,s)R_d(s,t')\right. \\ &\left.+(m-2)M_R^{(o)}(t,s)R_o(s,t')\right]\mathrm{d}s\,.\end{aligned} \tag{2.37}$$

(2) Equations for auxiliary functions. The memory kernels $M_R^{(s)}(t,s)$, $M_R^{(o)}(t,s)$ and $M_C^{(s)}(t,s)$, $M_C^{(o)}(t,s)$ are given by:

$$M_R^{(d)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)\left[R_A(t,s)h'(C_d(t,s)) + C_A(t,s)h''(C_d(t,s))R_d(t,s)\right]\,, \tag{2.38}$$

$$M_R^{(o)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)\left[R_A(t,s)h'(C_o(t,s)) + C_A(t,s)h''(C_o(t,s))R_o(t,s)\right]\,, \tag{2.39}$$

$$M_C^{(d)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)C_A(t,s)h'(C_d(t,s))\,, \tag{2.40}$$

$$M_C^{(o)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)C_A(t,s)h'(C_o(t,s))\,. \tag{2.41}$$

Further, $C_A(t,s)$, $R_A(t,s)$ are given by the same equations (2.19), where Σ_C, Σ_R are simplified as follows:

$$\begin{aligned}\Sigma_C(t,s) =& \tau^2 + \|\varphi\|^2 - a(t)\hat{\varphi}(\boldsymbol{v}(t)) - a(s)\hat{\varphi}(\boldsymbol{v}(s)) + \frac{a(t)a(s)}{m}h(C_d(t,s)) \\ &+ \frac{m-1}{m}a(t)a(s)h(C_o(t,s))\end{aligned} \tag{2.42}$$

$$\Sigma_R(t,s) = \frac{a(t)a(s)}{m}h'(C_d(t,s))R_d(t,s) + \frac{m-1}{m}a(t)a(s)h'(C_o(t,s))R_o(t,s)$$

Highly nonlinear!

A related (simpler) model

Random Gaussian Equations

- Physics: Fyodorov 2019; Urbani 2023; Kamali, Urnbani, 2023
- Mathematics: M, Subag 2023, 2024

DMFT eqs for Gaussian model are explicit

$$\frac{\mathrm{d}a}{\mathrm{d}t}(t) = \frac{\overline{\alpha}}{m}\hat{\varphi}(\boldsymbol{v}(t))\int_0^t R_A(t,s)\,\mathrm{d}s \tag{2.32}$$
$$-\frac{\overline{\alpha}}{m}\int_0^t R_A(t,s)a(s)\left[\frac{1}{m}h(C_d(t,s)) + \frac{m-1}{m}h(C_o(t,s))\right]\mathrm{d}s$$
$$-\frac{\overline{\alpha}}{m}\int_0^t C_A(t,s)a(s)\left[\frac{1}{m}h'(C_d(t,s))R_d(t,s) + \frac{m-1}{m}h'(C_o(t,s))R_o(t,s)\right]\mathrm{d}s\,,$$
$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}(t) = -\nu(t)\boldsymbol{v}(t) + \frac{\overline{\alpha}}{m}\nabla\hat{\varphi}(\boldsymbol{v}(t))a(t)\int_0^t R_A(t,s)\,\mathrm{d}s \tag{2.33}$$
$$-\frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s) + (m-1)M_R^{(o)}(t,s)\right]\boldsymbol{v}(s)\,\mathrm{d}s\,,$$
$$\partial_t C_d(t,t') = -\nu(t)C_d(t,t') + \frac{\overline{\alpha}}{m}\langle\nabla\hat{\varphi}'(\boldsymbol{v}(t)),\boldsymbol{v}(t')\rangle a(t)\int_0^t R_A(t,s)\,\mathrm{d}s \tag{2.34}$$
$$-\frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s)C_d(t',s) + (m-1)M_R^{(o)}(t,s)C_o(t',s)\right]\mathrm{d}s$$
$$-\frac{1}{m}\int_0^{t'} \left[M_C^{(d)}(t,s)R_d(t',s) + (m-1)M_C^{(o)}(t,s)R_o(t',s)\right]\mathrm{d}s\,,$$
$$\partial_t C_o(t,t') = -\nu(t)C_o(t,t') + \frac{\overline{\alpha}}{m}\langle\nabla\hat{\varphi}(\boldsymbol{v}(t)),\boldsymbol{v}(t')\rangle a(t)\int_0^t R_A(t,s)\,\mathrm{d}s \tag{2.35}$$
$$-\frac{1}{m}\int_0^t \left[M_R^{(d)}(t,s)C_o(t',s) + M_R^{(o)}(t,s)C_d(t',s) + (m-2)M_R^{(o)}(t,s)C_o(t',s)\right]\mathrm{d}s$$
$$-\frac{1}{m}\int_0^{t'} \left[M_C^{(d)}(t,s)R_o(t',s) + M_C^{(o)}(t,s)R_d(t',s) + (m-2)M_C^{(o)}(t,s)R_o(t',s)\right]\mathrm{d}s\,,$$

$$\partial_t R_d(t,t') = -\nu(t)R_d(t,t') + \delta(t-t') \tag{2.36}$$
$$-\frac{1}{m}\int_{t'}^t \left[M_R^{(d)}(t,s)R_d(s,t') + (m-1)M_R^{(o)}(t,s)R_o(s,t')\right]\mathrm{d}s\,,$$
$$\partial_t R_o(t,t') = -\nu(t)R_o(t,t') - \frac{1}{m}\int_{t'}^t \left[M_R^{(d)}(t,s)R_o(s,t') + M_R^{(o)}(t,s)R_d(s,t')\right. \tag{2.37}$$
$$\left.+(m-2)M_R^{(o)}(t,s)R_o(s,t')\right]\mathrm{d}s\,.$$

(2) Equations for auxiliary functions. The memory kernels $M_R^{(s)}(t,s)$, $M_R^{(o)}(t,s)$ and $M_C^{(s)}(t,s)$, $M_C^{(o)}(t,s)$ are given by:

$$M_R^{(d)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)\left[R_A(t,s)h'(C_d(t,s)) + C_A(t,s)h''(C_d(t,s))R_d(t,s)\right]\,, \tag{2.38}$$
$$M_R^{(o)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)\left[R_A(t,s)h'(C_o(t,s)) + C_A(t,s)h''(C_o(t,s))R_o(t,s)\right]\,, \tag{2.39}$$
$$M_C^{(d)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)C_A(t,s)h'(C_d(t,s))\,, \tag{2.40}$$
$$M_C^{(o)}(t,s) = \frac{\overline{\alpha}}{m}a(t)a(s)C_A(t,s)h'(C_o(t,s))\,. \tag{2.41}$$

Further, $C_A(t,s)$, $R_A(t,s)$ are given by the same equations (2.19), where Σ_C, Σ_R are simplified as follows:

$$\Sigma_C(t,s) = \tau^2 + \|\varphi\|^2 - a(t)\hat{\varphi}(\boldsymbol{v}(t)) - a(s)\hat{\varphi}(\boldsymbol{v}(s)) + \frac{a(t)a(s)}{m}h(C_d(t,s))$$
$$+\frac{m-1}{m}a(t)a(s)h(C_o(t,s)) \tag{2.42}$$
$$\Sigma_R(t,s) = \frac{a(t)a(s)}{m}h'(C_d(t,s))R_d(t,s) + \frac{m-1}{m}a(t)a(s)h'(C_o(t,s))R_o(t,s)$$

- Symmetric initialization: $\mathfrak{a}_i(0) = \mathfrak{a}_0$, $\boldsymbol{w}_i(0) \sim \mathrm{Unif}(\mathbb{S}^{d-1})$
- Numerical solution
- Singular asymptotics: $m \to \infty$, $t \to \infty$.

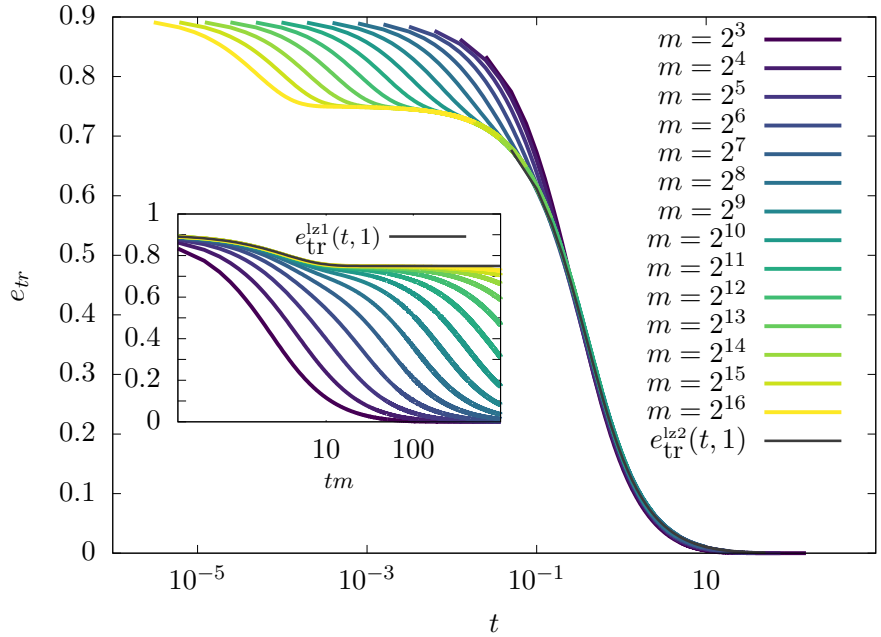
Dynamics in pure noise

Pure noise

$$(\mathbf{x}_i, y_i) \sim \mathsf{N}(0, \mathbf{I}_d) \otimes \mathsf{N}(0, \tau^2)$$

No learning, just optimization!

Fixed 2nd layer weights

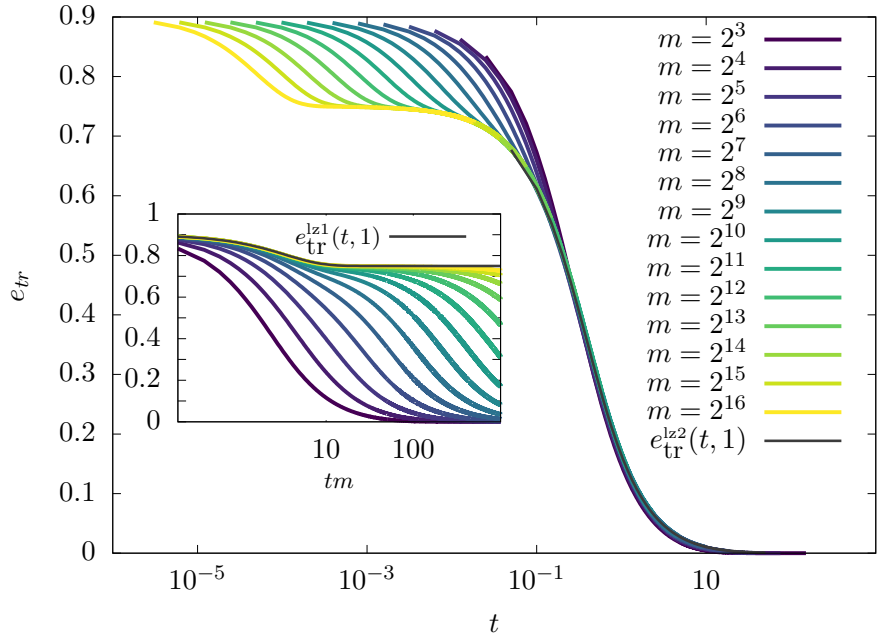


$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)\,, \qquad a_i(t) = \gamma\sqrt{m}\,, \quad \text{fixed.}$$

$$\lim_{n,d\to\infty,\, n/d=\overline{\alpha}} \widehat{\mathscr{R}}_n(\theta(t)) = e_{tr}(t;m,\overline{\alpha})\,,$$

$$\lim_{m,\alpha\to\infty,\, \overline{\alpha}/m=\alpha} e_{tr}(t;m,\overline{\alpha}) = e^{(2)}_{tr}(t;\alpha)\,.$$

Fixed 2nd layer weights

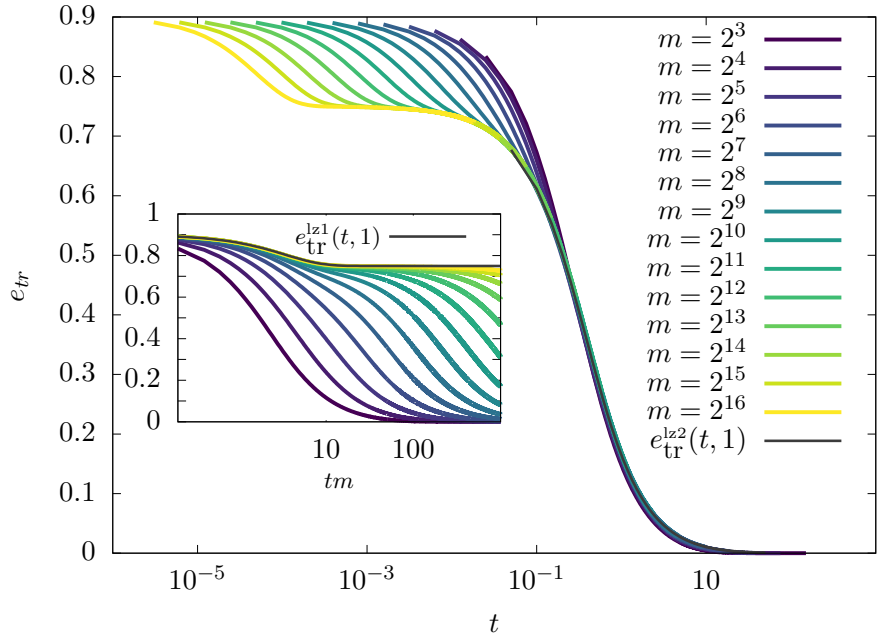


$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)\,, \qquad a_i(t) = \gamma\sqrt{m}\,, \quad \text{fixed.}$$

$$\lim_{n,d\to\infty,\, n/d=\overline{\alpha}} \widehat{\mathscr{R}}_n(\theta(t)) = e_{tr}(t;m,\overline{\alpha})\,,$$

$$\lim_{m,\alpha\to\infty,\, \overline{\alpha}/m=\alpha} e_{tr}(t;m,\overline{\alpha}) = e^{(2)}_{tr}(t;\alpha)\,.$$

Fixed 2nd layer weights

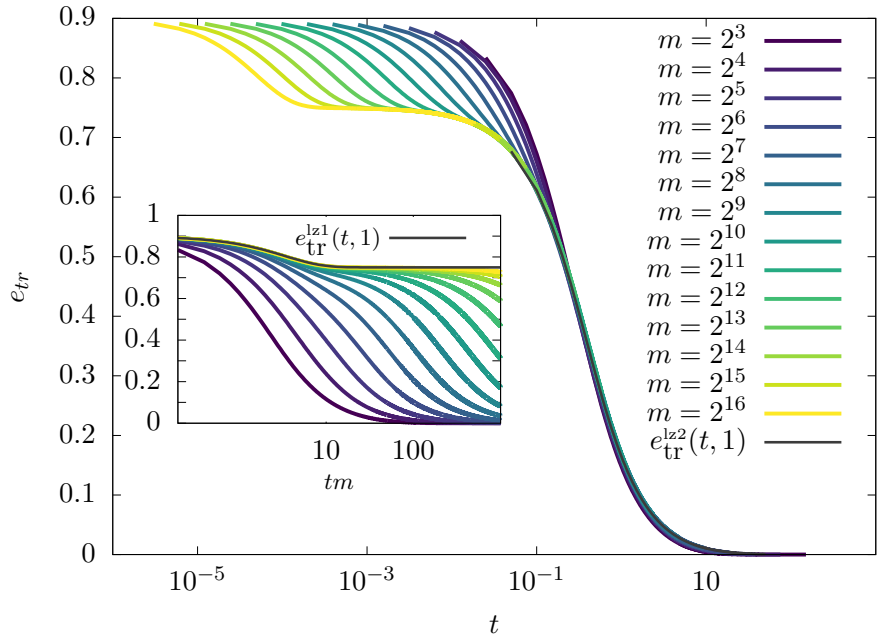


$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle)\,, \qquad a_i(t) = \gamma\sqrt{m}\,, \quad \text{fixed.}$$

$$\lim_{n,d\to\infty, n/d=\bar{\alpha}} \widehat{\mathscr{R}}_n(\theta(t)) = e_{tr}(t;m,\bar{\alpha})\,,$$

$$\lim_{m,\alpha\to\infty, \bar{\alpha}/m=\alpha} e_{tr}(t;m,\bar{\alpha}) = e^{(2)}_{tr}(t;\alpha)\,.$$

Fixed 2nd layer weights

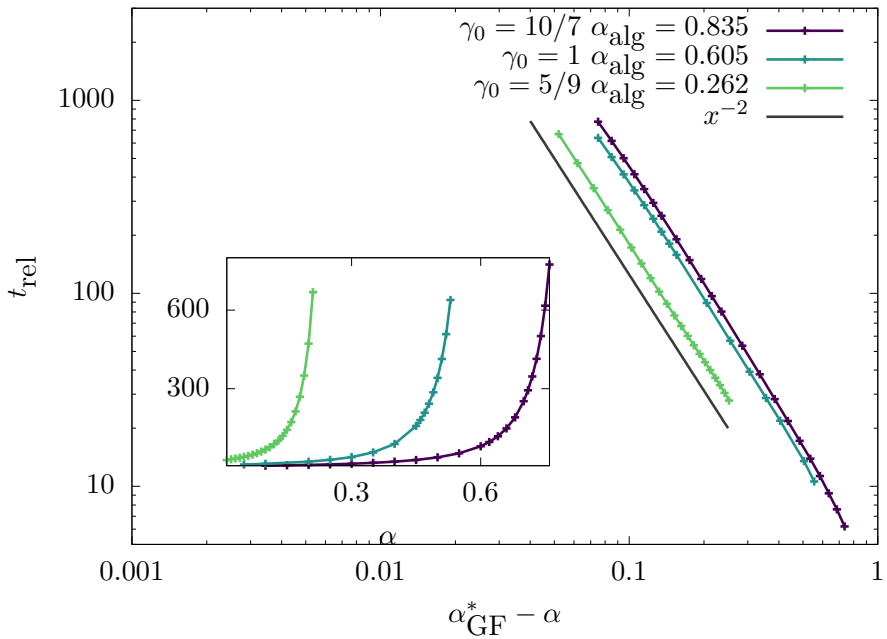


Algorithmic interpolation phase transition

$$\gamma < \gamma_{\rm GF}(\alpha) \quad \Rightarrow \quad \lim_{t\to\infty} e_{\rm tr}^{(2)}(t;\alpha) > 0\,,$$

$$\gamma > \gamma_{\rm GF}(\alpha) \quad \Rightarrow \quad \lim_{t\to\infty} e_{\rm tr}^{(2)}(t;\alpha) = 0\,.$$

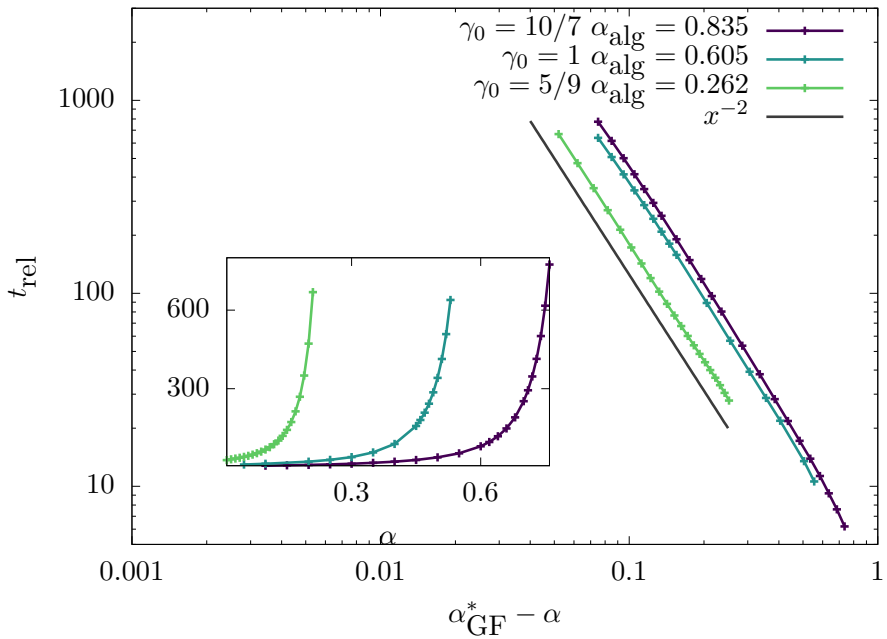
Fixed 2nd layer weights: Phase transition



$$\begin{aligned} \mathrm{t}_{\mathrm{rel}}(\alpha,\gamma) &\asymp (\gamma_{\mathrm{GF}}(\alpha) - \gamma)^{-\nu} \\ &\asymp (\alpha_{\mathrm{GF}}(\gamma) - \alpha)^{-\nu}\,. \end{aligned}$$

What happens if second-layer weights evolve?

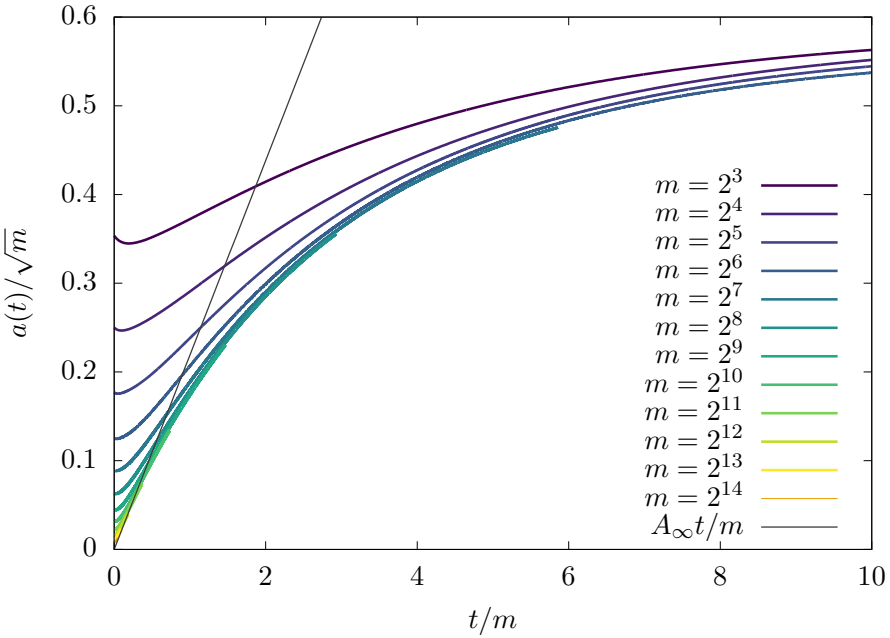
Fixed 2nd layer weights: Phase transition

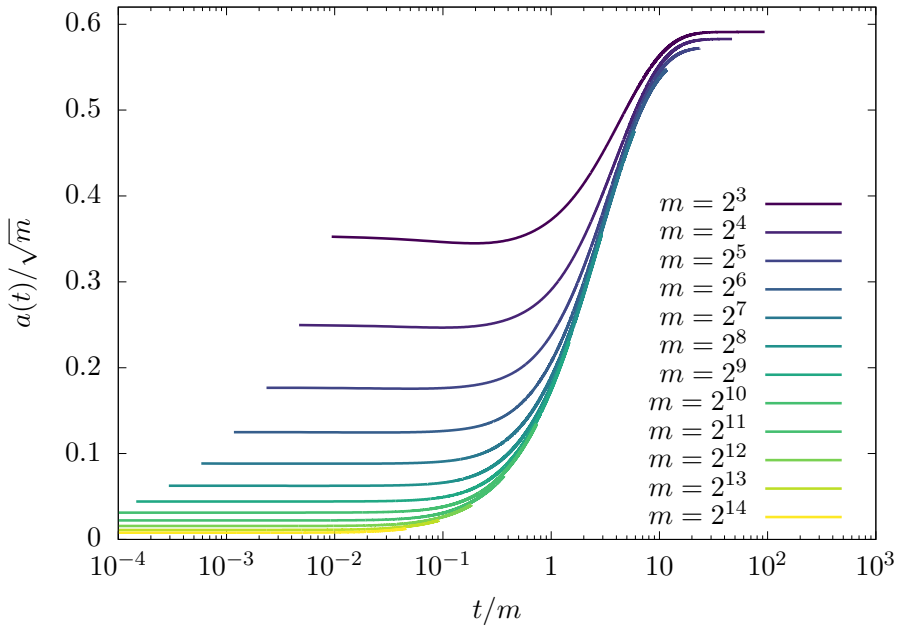


$$\mathrm{t}_{\mathrm{rel}}(\alpha,\gamma) \asymp (\gamma_{\mathrm{GF}}(\alpha) - \gamma)^{-\nu}$$
$$\asymp (\alpha_{\mathrm{GF}}(\gamma) - \alpha)^{-\nu}\,.$$

What happens if second-layer weights evolve?

Evolving 2nd layer weights: $a_i(0) = a_0$

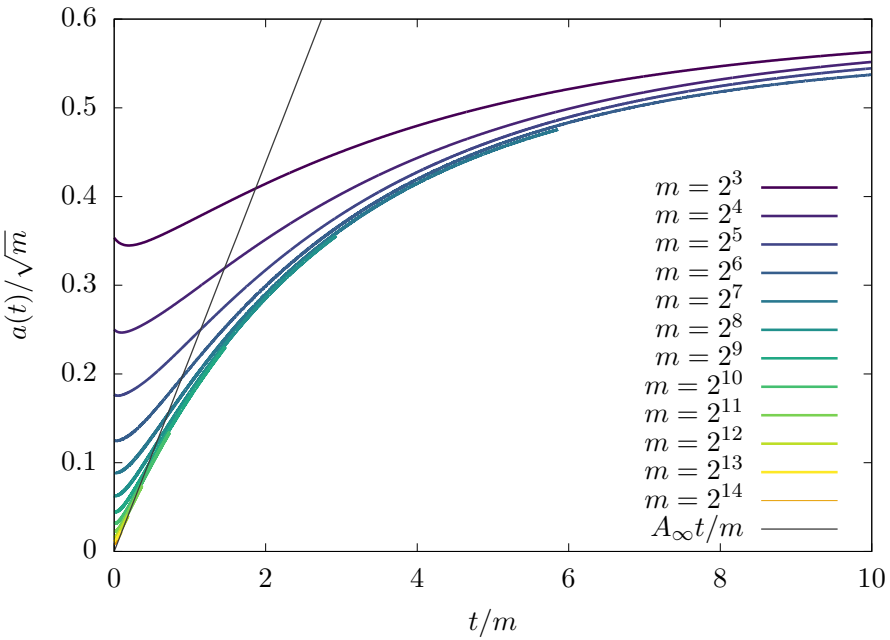


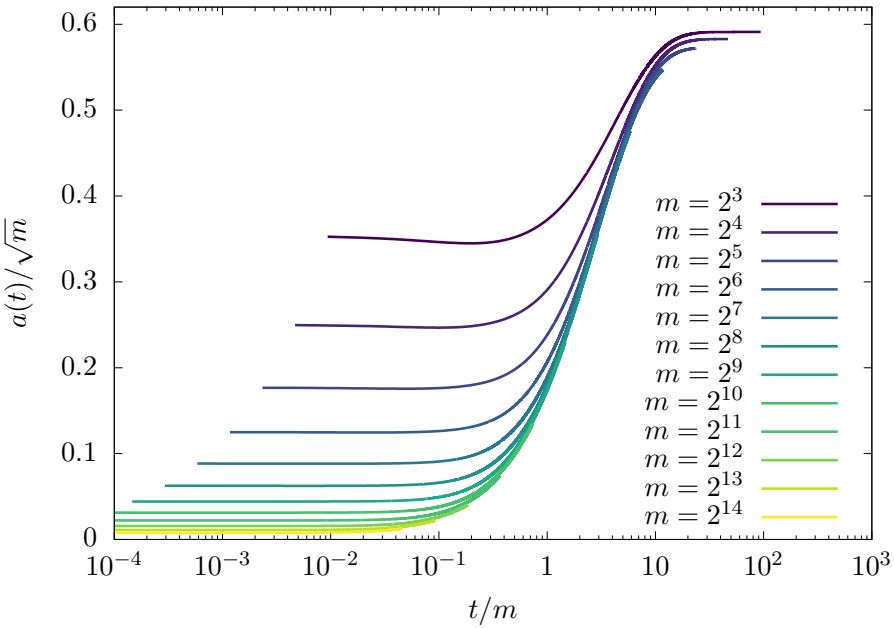


- Slowest of 3 dynamical regimes: $t/m = s = O(1)$.
- $a(t) = \sqrt{m}\gamma(t/m) + o(\sqrt{m})$
- Asymptotic behaviors:

$$s \to 0 : \quad \gamma(s) = \gamma_* s + o(s)\,,$$
$$s \to \infty : \quad \gamma(s) \to \gamma_{\rm GF}\,.$$

Evolving 2nd layer weights: $a_i(0) = a_0$



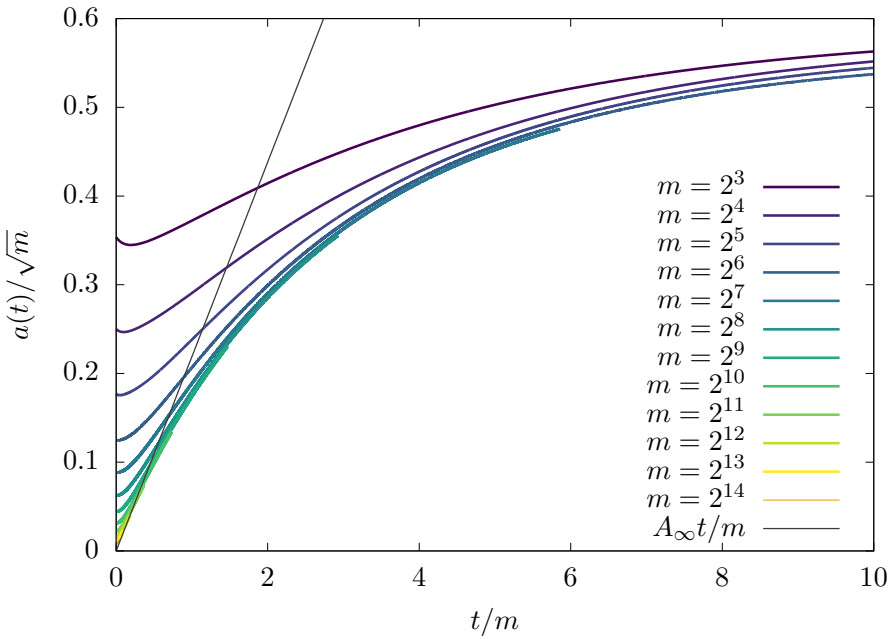


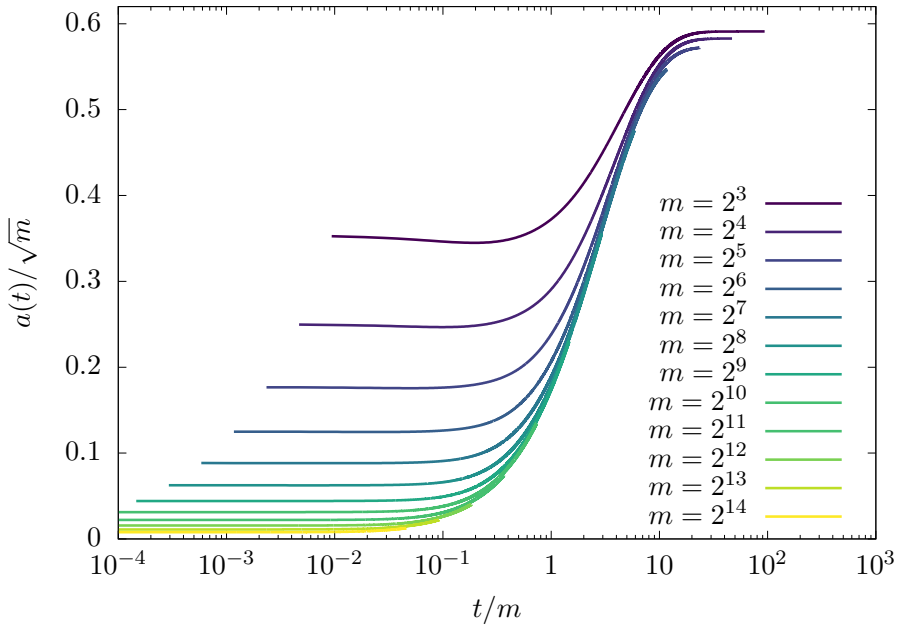
- Slowest of 3 dynamical regimes: $t/m = s = O(1)$.
- $a(t) = \sqrt{m}\gamma(t/m) + o(\sqrt{m})$
- Asymptotic behaviors:

$$s \to 0 \; : \quad \gamma(s) = \gamma_* s + o(s)\,.$$

γ_* is the threshold energy of a p-spin model!

Evolving 2nd layer weights: $a_i(0) = a_0$



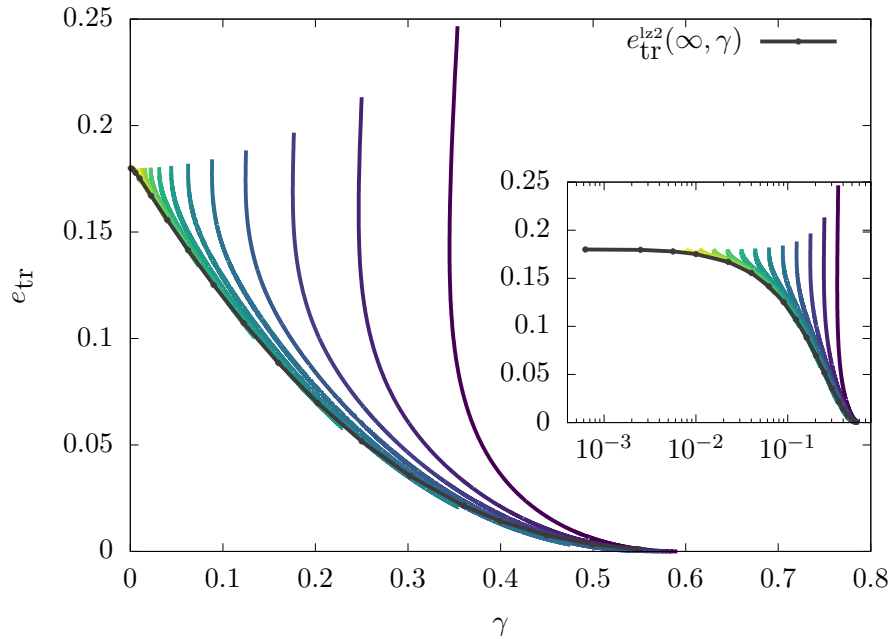


- Slowest of 3 dynamical regimes: $t/m = s = O(1)$.
- $a(t) = \sqrt{m}\gamma(t/m) + o(\sqrt{m})$
- Asymptotic behaviors:

$$s \to \infty \ : \quad \gamma(s) \to \gamma_{\mathrm{GF}} \, .$$

Complexity $\gamma(t)$ grows slowly until interpolation threshold

Adiabatic evolution of $\gamma(t) = a(t)/\sqrt{m}$



- **Black:** Minimum empirical risk achieved at γ fixed.
- **Colored:** $a(t)$ evolving

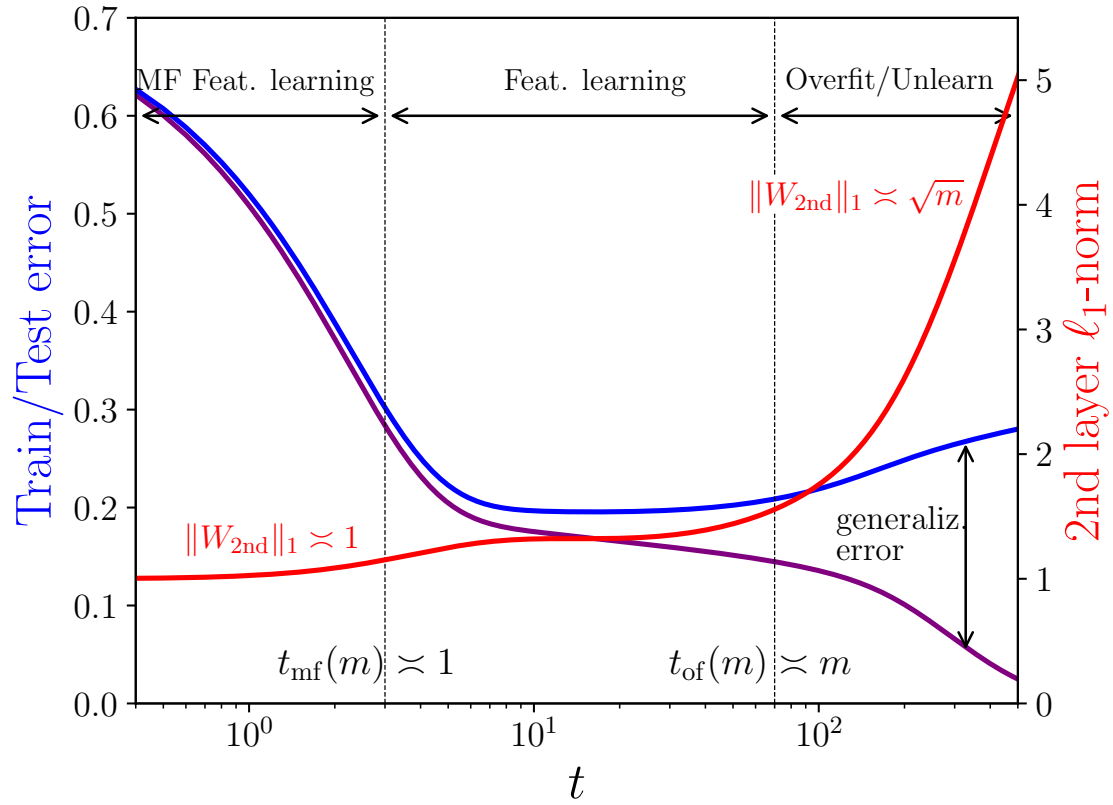
Dynamics under single-index model

Single-index model

Data $\{(\mathbf{x}_i, y_i) : i \leq n\}$ iid, $\varepsilon_i \sim \mathsf{N}(0, \tau^2)$

$$\mathbf{x}_i \sim \mathsf{N}(0, \mathbf{I}_d)\,, \quad y_i = \varphi(\langle \boldsymbol{w}_*, \mathbf{x}_i \rangle) + \varepsilon_i$$

Dynamical decoupling: Learning → Overfitting



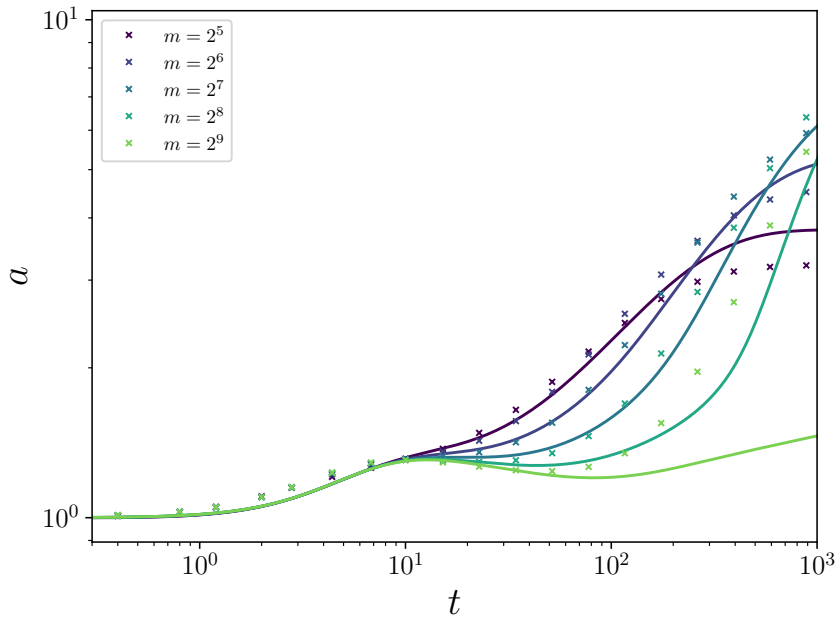
t $\asymp$ 1:

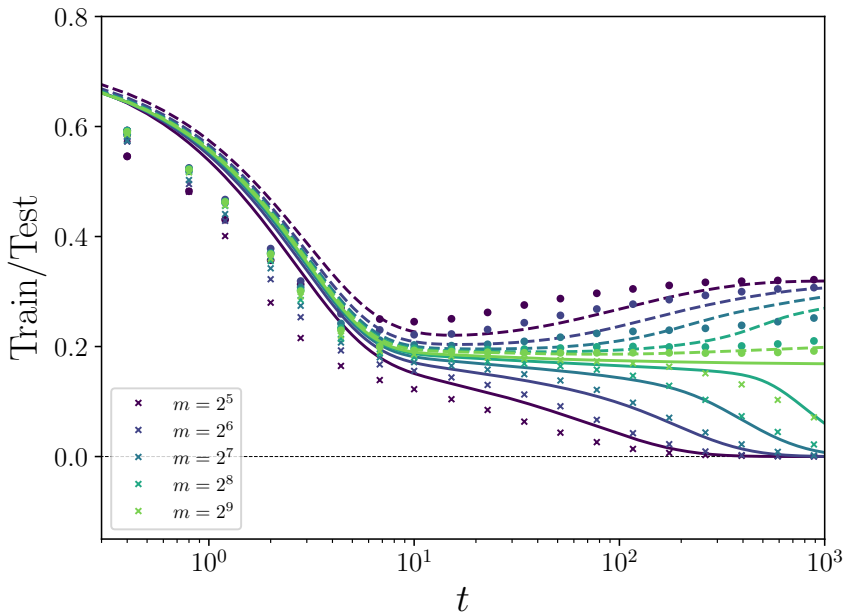
- Test error $\approx$ Train error
- Learns latent direction $\boldsymbol{w}_*$
- Mean field asymptotically correct (Mei, M, Nguyen, 2018; Chizat, Bach, 2018; Rotskoff, Vanden Eijnden, 2018)

t $\asymp$ m:

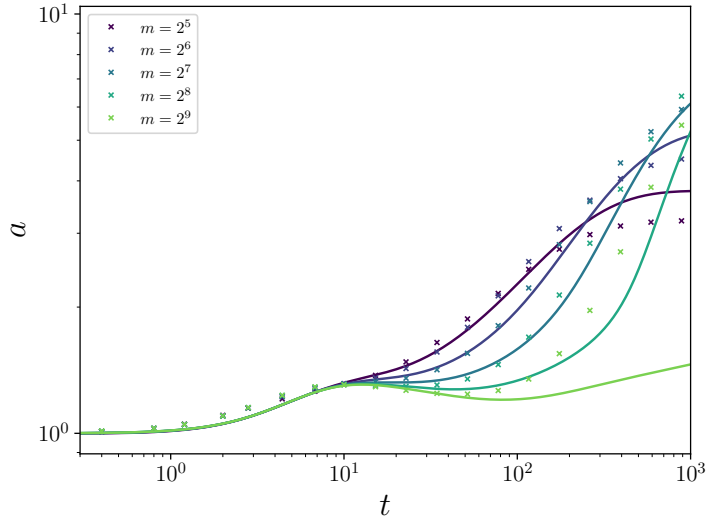
- Test error $\gg$ Train error; Train error $\to 0$
- *Unlearns* latent direction $\boldsymbol{w}_*$
- Neural tangent kernel learning

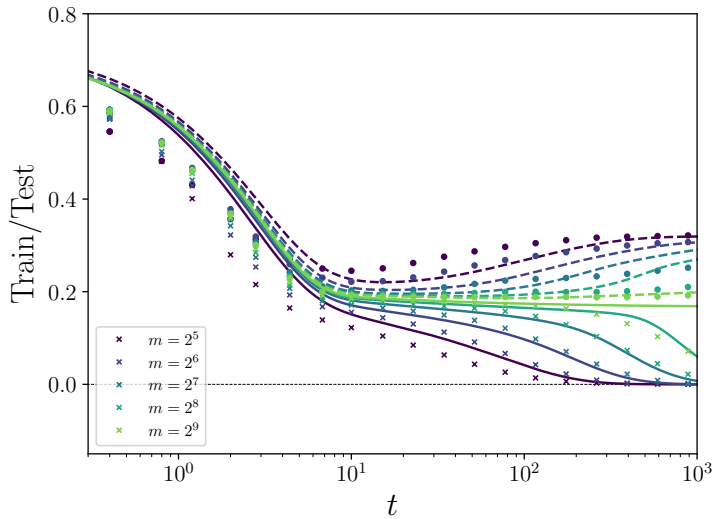
Comparing with SGD experiments





Overfitting mechanism: Adiabatic increase in complexity





- Bartlett 1996:

$$\left|\widehat{\mathscr{R}}_n(\theta) - \mathscr{R}(\theta)\right| \lesssim \underbrace{\frac{1}{m}\|a\|_1}_{\text{Rademacher comlexity}} \cdot \frac{1}{\sqrt{\alpha m}},.$$

-

$$t = \Theta(1) \quad \Rightarrow \quad \frac{1}{m}\|a(t)\|_1 = a(t) = O(1) \quad \Rightarrow \quad \text{No overfitting}$$

Lower bounding the overfitting timescale

Theorem (M, Urbani, 2025)

Assume $\|\sigma\|_{\mathrm{Lip}}, \|\sigma\|_\infty \le L$, $|\varphi(0)|, \|\varphi\|_{\mathrm{Lip}} \le L$, $\|\boldsymbol{a}(0)\|_\infty \le a_0$, $n \ge d \vee m$. *Then, with prob* $\ge 1 - 2\exp(-cn)$ *(*$\hat{t} = t_\alpha$*)*

$$\|\boldsymbol{a}(\hat{t})\|_\infty \le a_0 + a_1 \hat{t}, \quad a_1 := C_0 L(\tau + a_0 L),$$

$$\mathscr{R}(\boldsymbol{a}(\hat{t}), \boldsymbol{W}(\hat{t})) - \widehat{\mathscr{R}}_n(\boldsymbol{a}(\hat{t}), \boldsymbol{W}(\hat{t})) \le C_1 L^3 (a_0 + a_1 \hat{t})^2 \cdot \sqrt{\frac{d}{n}}.$$

Dynamics on the feature learning timescale

Mean field asymptotics: $(\mathbf{Q}_{\boldsymbol{v}} := \mathbf{I}_k - \boldsymbol{v}\boldsymbol{v}^\mathsf{T},\ \hat{\varphi}, h = \dots)$

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}}\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}) = a_i^{\mathrm{mf}}(\hat{t})\mathbf{Q}_{\boldsymbol{v}_i^{\mathrm{mf}}}\Big(\nabla\hat{\varphi}(\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m}\sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t})h'(\langle\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}), \boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\rangle)\boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\Big),$$

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}}a_i^{\mathrm{mf}}(\hat{t}) = \hat{\varphi}(\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m}\sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t})h(\langle\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}), \boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\rangle)\,.$$

- Gradient flow wrt population risk.
- High-dimensional asymptotics: $\langle \boldsymbol{w}_i(t), \boldsymbol{w}_j(t)\rangle \approx \langle \boldsymbol{v}_i(t), \boldsymbol{v}_j(t)\rangle$.
- 2-dimensional under symmetric initialization

Mei, M, Nguyen, 2018; Chizat, Bach, 2018; Rotskoff, Vanden Eijnden, 2018; Berthier, M, Zhong, 2024

Dynamics on the feature learning timescale

Mean field asymptotics: $(\mathbf{Q}_{\boldsymbol{v}} := \mathbf{I}_k - \boldsymbol{v}\boldsymbol{v}^\mathsf{T},\ \hat{\varphi}, h = \dots)$

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}}\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}) = a_i^{\mathrm{mf}}(\hat{t})\mathbf{Q}_{\boldsymbol{v}_i^{\mathrm{mf}}}\Big(\nabla\hat{\varphi}(\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m}\sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t})h'(\langle \boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}), \boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\rangle)\boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\Big),$$

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}}a_i^{\mathrm{mf}}(\hat{t}) = \hat{\varphi}(\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m}\sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t})h(\langle \boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}), \boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\rangle)\,.$$

- Gradient flow wrt population risk.
- High-dimensional asymptotics: $\langle \boldsymbol{w}_i(t), \boldsymbol{w}_j(t)\rangle \approx \langle \boldsymbol{v}_i(t), \boldsymbol{v}_j(t)\rangle$.
- 2-dimensional under symmetric initialization

Mei, M, Nguyen, 2018; Chizat, Bach, 2018; Rotskoff, Vanden Eijnden, 2018; Berthier, M, Zhong, 2024

Dynamics on the feature learning timescale

Mean field asymptotics: $(\mathbf{Q}_{\boldsymbol{v}} := \mathbf{I}_k - \boldsymbol{v}\boldsymbol{v}^\mathsf{T},\ \hat{\varphi}, h = \dots)$

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}}\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}) = a_i^{\mathrm{mf}}(\hat{t})\mathbf{Q}_{\boldsymbol{v}_i^{\mathrm{mf}}}\Big(\nabla\hat{\varphi}(\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m}\sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t})h'(\langle \boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}), \boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\rangle)\boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\Big),$$

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}}a_i^{\mathrm{mf}}(\hat{t}) = \hat{\varphi}(\boldsymbol{v}_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m}\sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t})h(\langle \boldsymbol{v}_i^{\mathrm{mf}}(\hat{t}), \boldsymbol{v}_j^{\mathrm{mf}}(\hat{t})\rangle)\,.$$

- Gradient flow wrt population risk.
- High-dimensional asymptotics: $\langle \boldsymbol{w}_i(t), \boldsymbol{w}_j(t)\rangle \approx \langle \boldsymbol{v}_i(t), \boldsymbol{v}_j(t)\rangle$.
- 2-dimensional under symmetric initialization

Mei, M, Nguyen, 2018; Chizat, Bach, 2018; Rotskoff, Vanden Eijnden, 2018; Berthier, M, Zhong, 2024

Dynamics on the feature learning timescale

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}} v_i^{\mathrm{mf}}(\hat{t}) = a_i^{\mathrm{mf}}(\hat{t}) Q_{v_i^{\mathrm{mf}}} \Big(\nabla \hat{\varphi}(v_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m} \sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t}) h'(\langle v_i^{\mathrm{mf}}(\hat{t}), v_j^{\mathrm{mf}}(\hat{t}) \rangle) v_j^{\mathrm{mf}}(\hat{t}) \Big),$$

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}} a_i^{\mathrm{mf}}(\hat{t}) = \hat{\varphi}(v_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m} \sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t}) h(\langle v_i^{\mathrm{mf}}(\hat{t}), v_j^{\mathrm{mf}}(\hat{t}) \rangle).$$

Theorem (M, Urbani, 2025)

Under same assumptions, let $T_{\mathrm{lb}} = c_0 (\log m)^{1/3} \wedge (\log n/d)^{1/3}$. *Then with prob* $\geq 1 - 2\exp(-c_1 d)$,

$$\sup_{\hat{t} \leq T_{\mathrm{lb}}} \frac{1}{m} \sum_{i=1}^{m} \Big(|a_i(\hat{t}) - a_i^{\mathrm{mf}}(\hat{t})| + \|v_i(\hat{t}) - v_i^{\mathrm{mf}}(\hat{t})\| \Big) \leq C \left(\frac{1}{m} \vee \frac{1}{d} \vee \frac{d}{n} \right)^{1/2-\delta}$$

Dynamics on the feature learning timescale

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}} v_i^{\mathrm{mf}}(\hat{t}) = a_i^{\mathrm{mf}}(\hat{t}) \mathbf{Q}_{v_i^{\mathrm{mf}}} \Big(\nabla \hat{\varphi}(v_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m} \sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t}) h'(\langle v_i^{\mathrm{mf}}(\hat{t}), v_j^{\mathrm{mf}}(\hat{t}) \rangle) v_j^{\mathrm{mf}}(\hat{t}) \Big),$$

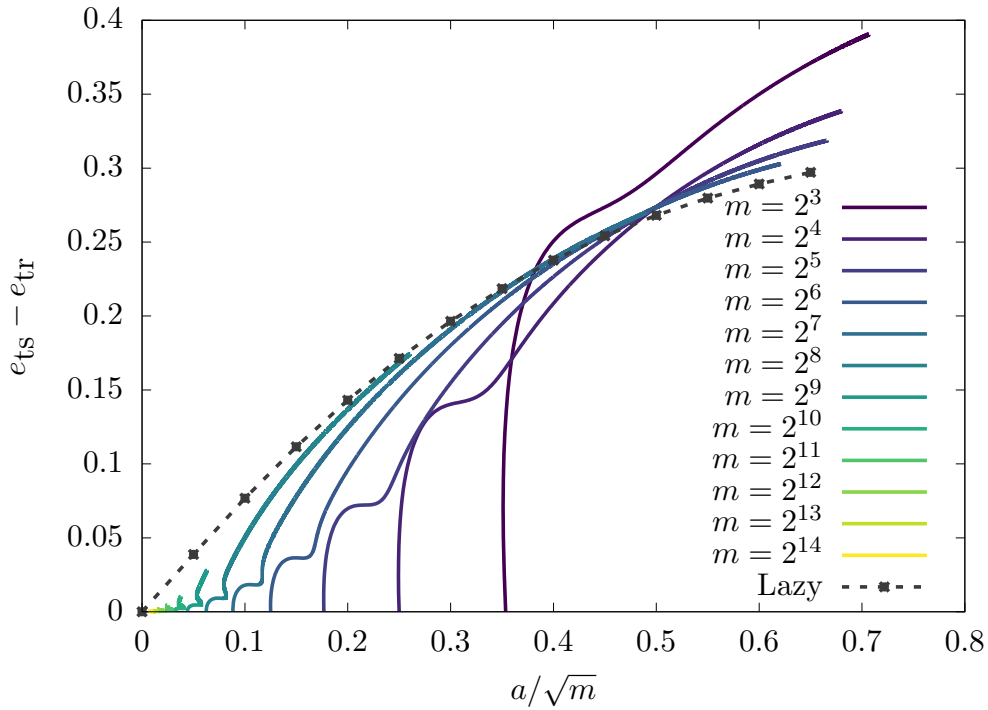
$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}} a_i^{\mathrm{mf}}(\hat{t}) = \hat{\varphi}(v_i^{\mathrm{mf}}(\hat{t})) - \frac{1}{m} \sum_{j=1}^{m} a_j^{\mathrm{mf}}(\hat{t}) h(\langle v_i^{\mathrm{mf}}(\hat{t}), v_j^{\mathrm{mf}}(\hat{t}) \rangle).$$

Theorem (M, Urbani, 2025)

Under same assumptions, let $T_{\mathrm{lb}} = c_0 (\log m)^{1/3} \wedge (\log n/d)^{1/3}$. *Then with prob* $\geq 1 - 2\exp(-c_1 d)$,

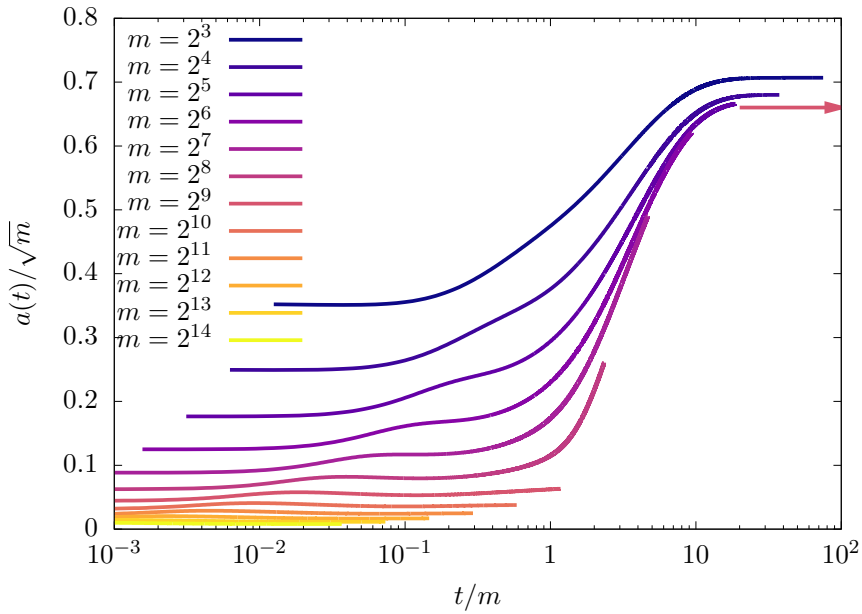
$$\sup_{\hat{t} \leq T_{\mathrm{lb}}} \frac{1}{m} \sum_{i=1}^{m} \Big(|a_i(\hat{t}) - a_i^{\mathrm{mf}}(\hat{t})| + \|v_i(\hat{t}) - v_i^{\mathrm{mf}}(\hat{t})\| \Big) \leq C \left(\frac{1}{m} \vee \frac{1}{d} \vee \frac{d}{n} \right)^{1/2-\delta}$$

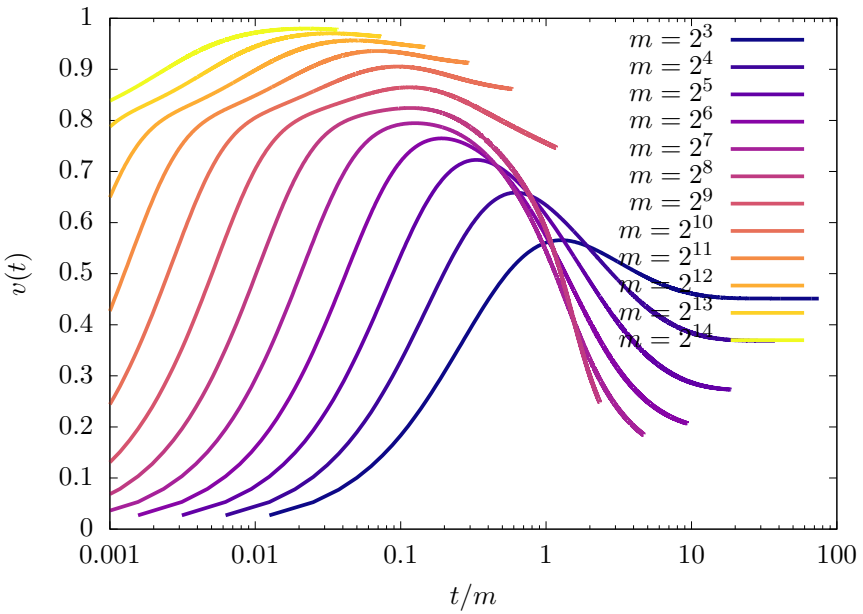
Mechanism: Adiabatic increase in complexity



- Slow timescale: Divergence of 2nd layer weights
- $\Leftrightarrow$ Growth of Gaussian/Rademacher complexity

Unlearning: $\mathbf{t = \Theta(m)}$





- Projection onto the latent direction vanishes.

Conclusion

Conclusion

- How can modern ML models generalize well?
- Requires to undertand dynamics
- ...

Thank you!

Conclusion

- How can modern ML models generalize well?
- Requires to undertand dynamics
- ...

Thank you!