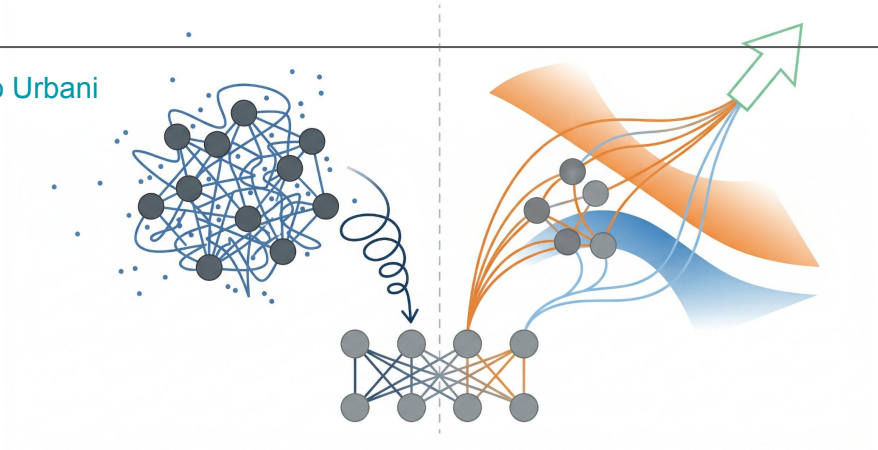
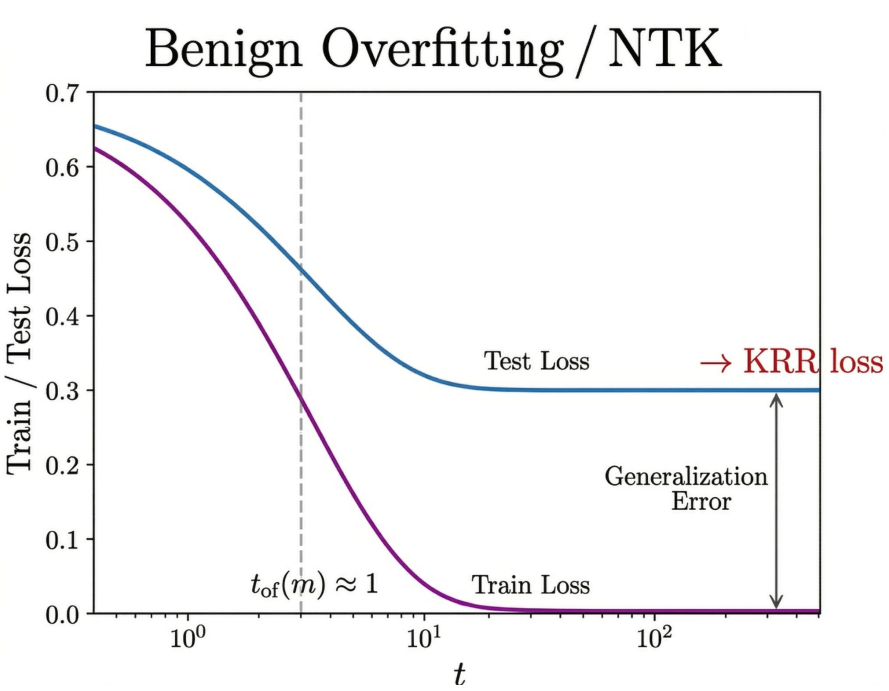
Dynamical Decoupling of Generalization and Overfitting in Large Two-Layer Networks

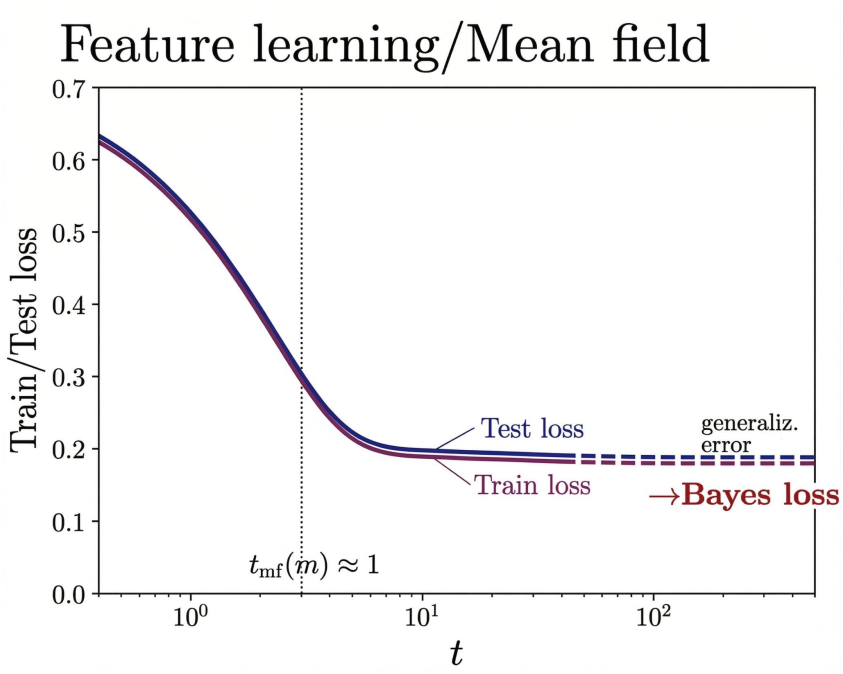
Andrea Montanari, Pierfrancesco Urbani



Mahdi Zamani



No feature learning



No overfitting?

Setup

Model $$f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W}) = \frac{1}{m} \sum_{i=1}^{m} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \qquad \|\boldsymbol{w}_i\|^2 = 1 \qquad \boldsymbol{\theta} = (\boldsymbol{a}, \boldsymbol{W})$$

Data Generation $$y_i = \varphi(\boldsymbol{U}^{\mathsf{T}} \boldsymbol{x}_i) + \varepsilon_i \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2) \qquad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$$

Empirical Risk $$\widehat{\mathscr{R}}_n(\boldsymbol{\theta}) := \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta})\big)^2$$

Gradient Flow (GF) $$\dot{\boldsymbol{\theta}}(t) = -\frac{n}{d} \boldsymbol{P}_{\boldsymbol{\theta}} \nabla \widehat{\mathscr{R}}_n(\boldsymbol{\theta}(t))$$

Overparameterization ratio $$\alpha = \frac{n}{md}$$

Pre-Techniques

Gaussian Approximation

$$\widehat{\mathscr{R}}_n^g(\boldsymbol{a}, \boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(f_i^g(\boldsymbol{a}, \boldsymbol{W}) - \varphi_i^g - \varepsilon_i \right)^2$$

Covariance Structure

$$\mathbb{E}\{ f(\boldsymbol{x}; \boldsymbol{a}_1, \boldsymbol{W}_1)\, f(\boldsymbol{x}; \boldsymbol{a}_2, \boldsymbol{W}_2) \} = \frac{1}{m^2} \sum_{i,j=1}^{m} a_{1,i} a_{2,j} h\left(\langle \boldsymbol{w}_{1,i}, \boldsymbol{w}_{2,j} \rangle \right)$$

$$h(q) = \mathbb{E}\{\sigma(G_1)\sigma(G_2)\} \qquad \mathbb{E}\{G_1 G_2\} = q$$

$$\mathbb{E}\{ f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W})\, \varphi(\boldsymbol{U}^\mathsf{T} \boldsymbol{x}) \} = \frac{1}{m} \sum_{i=1}^{m} a_i \widehat{\varphi}(\boldsymbol{U}^\mathsf{T} \boldsymbol{w}_i)$$

$$\widehat{\varphi}(\boldsymbol{v}) := \mathbb{E}\Big\{ \sigma(\langle \boldsymbol{v}, \boldsymbol{G} \rangle + \sqrt{1 - \|\boldsymbol{v}\|^2} G_0) \varphi(\boldsymbol{G}) \Big\} \qquad \boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_k) \text{ independent of } G_0 \sim \mathsf{N}(0,1)$$

DMFT

$$\dot{\boldsymbol{a}}(t) = -\frac{n}{d}\nabla_{\boldsymbol{a}}\widehat{\mathscr{R}}_n(\boldsymbol{a}(t), \boldsymbol{W}(t))\,,$$
$$\dot{\boldsymbol{w}}_i(t) = -\frac{n}{d}\nabla_{\boldsymbol{w}_i}\widehat{\mathscr{R}}_n(\boldsymbol{a}(t), \boldsymbol{W}(t)) - \nu_i(t)\boldsymbol{w}_i(t) \quad \forall i = 1, \dots, m\,.$$

Interacting ODEs

Dynamical Partition function
+
Grassmann Variables
+
Concentration of
Path Integral around its saddle point

DMFT
(d, n→∞)
d/n→const

Average behavior of GF

$$\frac{\mathrm{d}a_i(t)}{\mathrm{d}t} = -\frac{\overline{\alpha}}{m}\int_0^t R_A(t,s)\left[\frac{1}{m}\sum_{l=1}^m a_l(s)h\left(C_{li}(s,t)\right) - \hat{\varphi}(\boldsymbol{v}_i(t))\right]\mathrm{d}s \tag{C.7}$$
$$-\frac{\overline{\alpha}}{m}\int_0^t C_A(t,s)\frac{1}{m}\sum_{l=1}^m a_l(s)h'\big(C_{li}(s,t)\big)R_{il}(t,s)\,\mathrm{d}s\,,$$

$$\frac{\mathrm{d}\boldsymbol{v}_i(t)}{\mathrm{d}t} = -\nu_i(t)\boldsymbol{v}_i(t) + \frac{\overline{\alpha}}{m}a_i(t)\nabla\hat{\varphi}(\boldsymbol{v}_i(t))\int_0^t R_A(t,s)\,\mathrm{d}s - \frac{1}{m}\sum_{j=1}^m\int_0^t M_{ij}^R(t,s)\,\boldsymbol{v}_j(s)\,\mathrm{d}s\,, \tag{C.8}$$

# of equations O(m^2)

Effectively one ODE
(particle) + noise + memory

$$\frac{\partial C_{ij}(t_a,t_b)}{\partial t_a} = -\nu_i(t_a)C_{ij}(t_a,t_b) + \frac{\overline{\alpha}}{m}a_i(t_a)\langle\nabla\hat{\varphi}(\boldsymbol{v}_i(t_a)),\boldsymbol{v}_j(t_b)\rangle\int_0^{t_a} R_A(t_a,s)\,\mathrm{d}s \tag{C.9}$$
$$-\frac{1}{m}\sum_{l=1}^m\int_0^{t_a} M_{il}^R(t_a,s)\,C_{lj}(s,t_b)\,\mathrm{d}s - \frac{1}{m}\sum_{l=1}^m\int_0^{t_b} M_{il}^C(t_a,s)\,R_{jl}(t_b,s)\mathrm{d}s\,,$$

$$\frac{\partial R_{ij}(t_a,t_b)}{\partial t_a} = -\nu_i(t_a)R_{ij}(t_a,t_b) + \delta_{ij}\delta(t_a-t_b) - \frac{1}{m}\sum_{l=1}^m\int_{t_b}^{t_a} M_{il}^R(t_a,s)\,R_{lj}(s,t_b)\,\mathrm{d}s\,. \tag{C.10}$$

SymmDMFT

Can be significantly simplified by considering a symmetric initialization

$$\boldsymbol{w}_i^n(0) \sim \mathsf{Unif}(\mathbb{S}^{d-1}) \qquad a_i^n(0) = a_0$$

This symmetric initialization is preserved by GF and we end up with O(3 + 1) integro-differential equations and can derive Test and Train error's (limiting) evolution:

$$e_{\mathrm{tr}}(t) = -\frac{1}{2} C_A(t,t)\,,$$

$$e_{\mathrm{ts}}(t) = \frac{1}{2}\left[\tau^2 + \|\varphi\|^2 - 2a(t)\hat{\varphi}(\boldsymbol{v}(t)) + \frac{1}{m}a^2(t)h(1) + \frac{m-1}{m}a^2(t)h(C_o(t,t))\right]$$

Singular Perturbation Theory

However, it's still hard to derive the limiting distribution for the evolution of weights. i.e w(t) and a(t)

So we also let m go to infinity! ¯_(ツ)_/¯

(i) Hypothesize a certain asymptotic behavior of the DMFT solution in a specific time-scale

(ii) Check consistency with the DMFT equations

(iii) Check that this behavior is observed in the numerical solution of the DMFT equations.

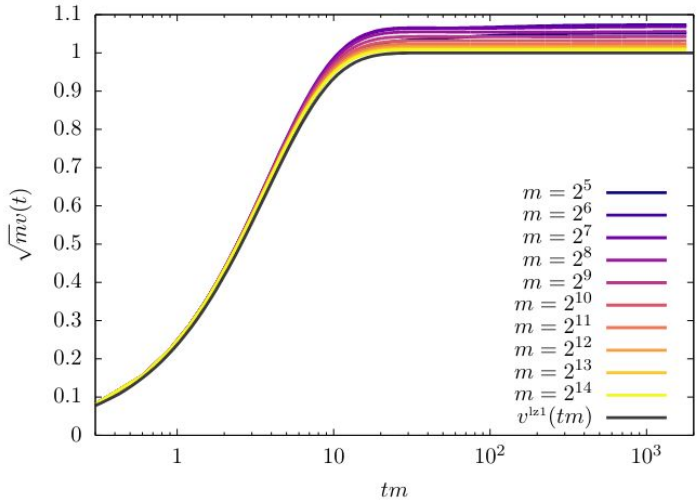
And this depends heavily on initialization of the second layer weights.

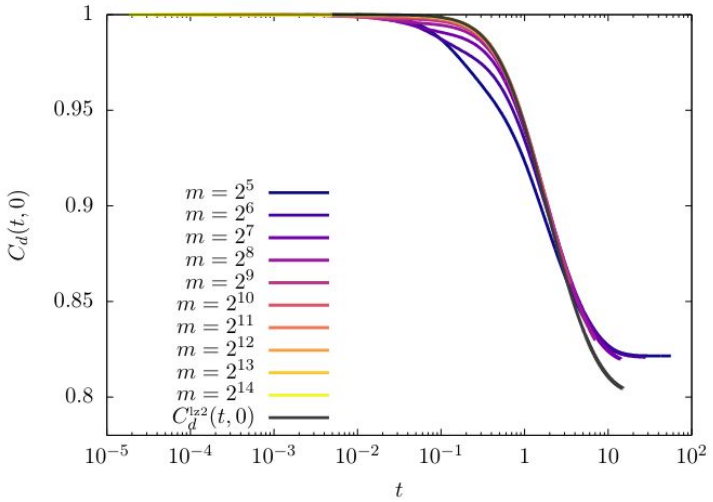
Lazy initialization

$$\gamma(t) = a(t)/\sqrt{m} \text{ (in particular, } \gamma(0) = \gamma_0)$$

First dynamical regime: t = O(1/m)

On this timescale, the SymmDMFT equations are solved, up to higher order terms, by an ansatz





$$\|\boldsymbol{w}_i(t) - \boldsymbol{w}_i(0)\| = \Theta(1/\sqrt{m})$$

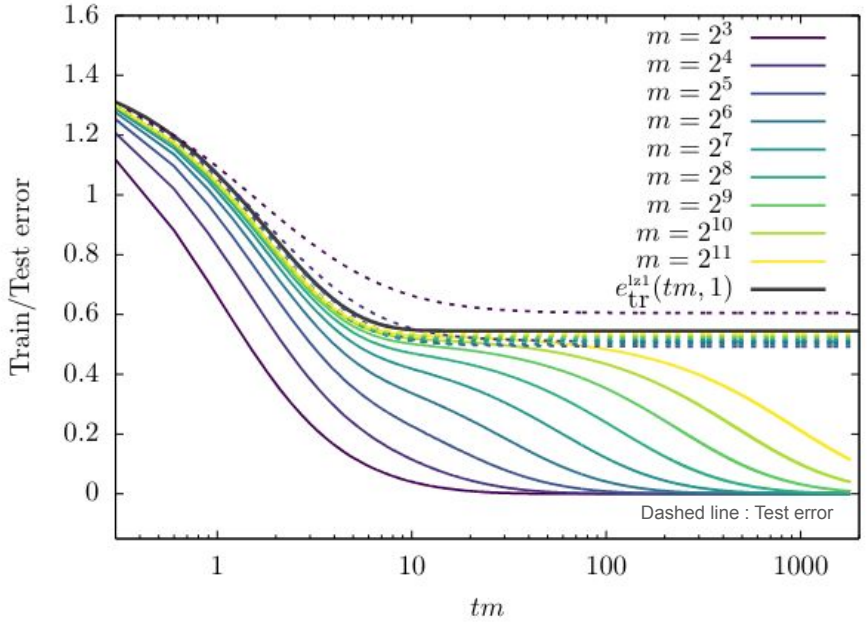
$$\|\boldsymbol{a}(t) - \boldsymbol{a}(0)\| = o_m(1)$$

$$\boldsymbol{v}_i(t) := \operatorname*{p\text{-}lim}_{n,d\to\infty} \boldsymbol{U}^\mathsf{T} \boldsymbol{w}_i^n(t)$$

$$C_{ij}(t_1,t_2) := \operatorname*{p\text{-}lim}_{n,d\to\infty} \langle \boldsymbol{w}_i^n(t_1), \boldsymbol{w}_j^n(t_2) \rangle$$

Second dynamical regime: t = Θ(1)

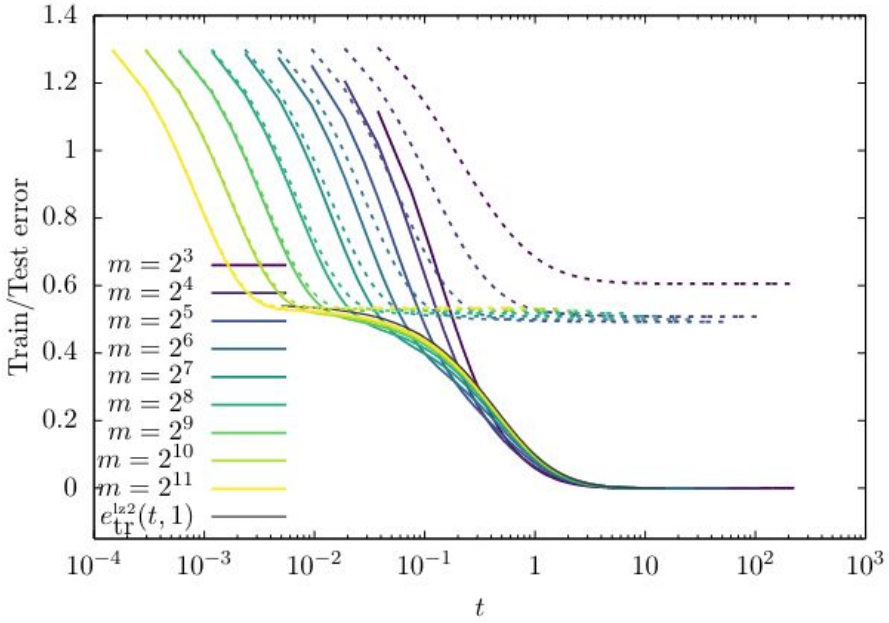
$t = \Theta(1/m)$



$train \approx test$

$train\ error\ change \approx O(1)$

$t = \Theta(1)$

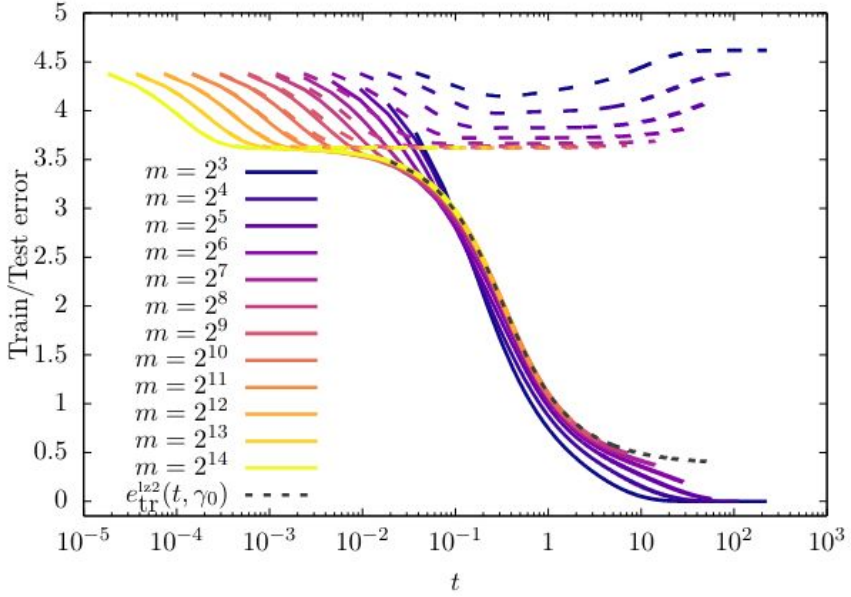


$$\|\boldsymbol{w}_i(t) - \boldsymbol{w}_i(0)\| = \Theta(1)$$

$$\|\boldsymbol{a}(t) - \boldsymbol{a}(0)\| = o_m(1)$$

No additional learning / memorization / overfitting

Third dynamical regime: t = Θ(m)

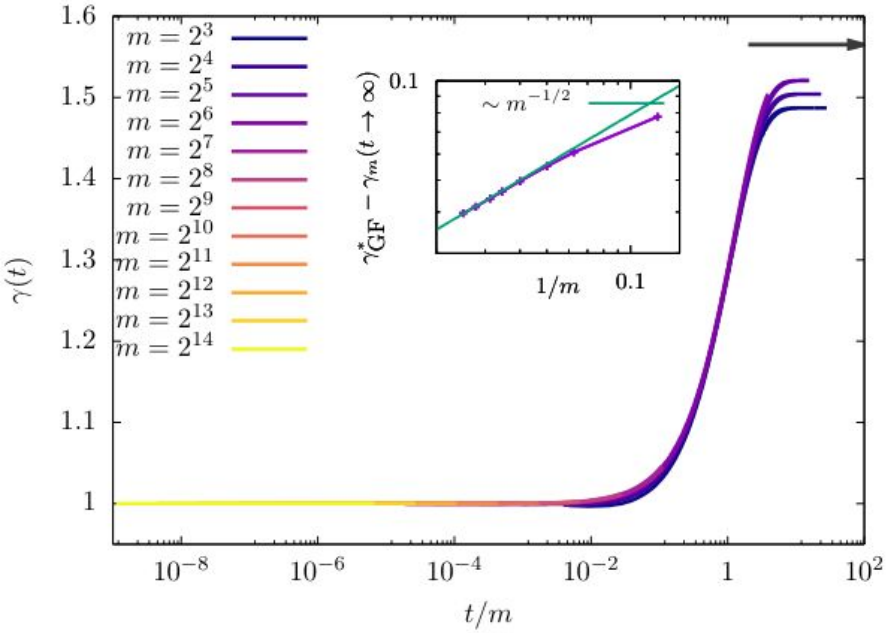


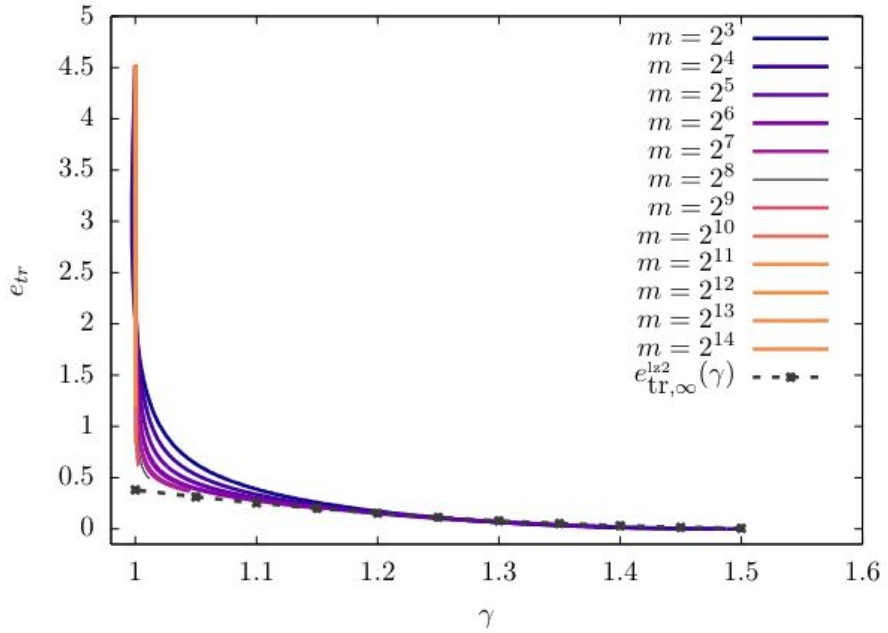
$$\|\boldsymbol{w}_i(t) - \boldsymbol{w}_i(0)\| = \Theta(1)$$

$$\|\boldsymbol{a}(t) - \boldsymbol{a}(0)\| = \Theta(1)$$

train $\downarrow 0$
test $\uparrow$

Third dynamical regime: t = Θ(m)





$$\gamma_0 < \gamma^*_{\mathrm{GF}}(\alpha, \varphi, \tau) \qquad a(t) = \gamma(t)\sqrt{m}$$

Mean-Field initialization

$$a(0) = 1$$

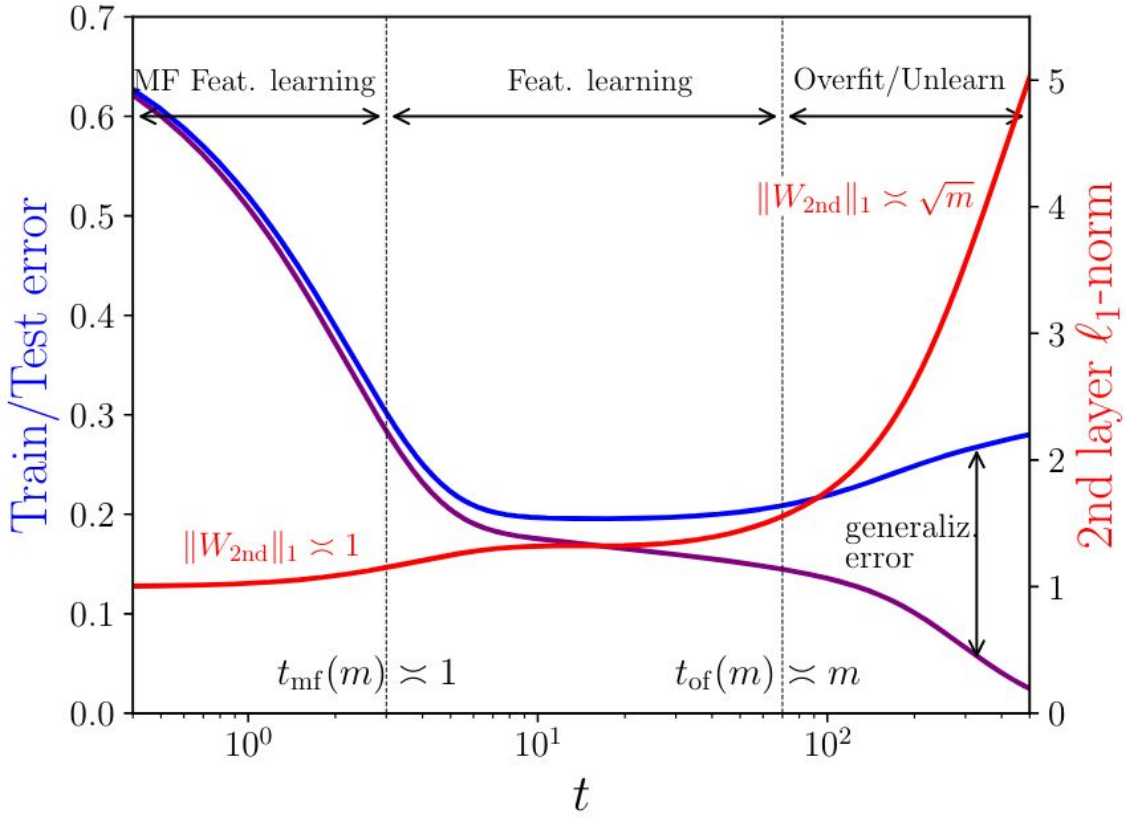
t	1st Layer Weights	2nd Layer Weights	Train Error	Test Error
Θ(1)	O(1)	O(1)	O(1) Reaches plateau	Test=Train=Bayes risk
Θ(m)	O(1)	O(√m)	Decays to 0	Increases

$$\varphi = \sigma \qquad h'(0) > 0$$

$$v^{\mathrm{mf1}}(t) \to 1$$

Fixed point

Main Result


MF Feat. learning
Feat. learning
Overfit/Unlearn
$\|W_{2\text{nd}}\|_1 \asymp \sqrt{m}$
$\|W_{2\text{nd}}\|_1 \asymp 1$
generaliz. error
$t_{\text{mf}}(m) \asymp 1$
$t_{\text{of}}(m) \asymp m$
Train/Test error
2nd layer ℓ_1-norm
t
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
5
10^0
10^1
10^2

A lot more to learn from the paper

- Had some cool connections with spin glasses
- Lower bounding the overfitting timescale for finite n,d (MuP)
- Addressing implicit bias by the second layer weights
- Mathematical Insights

Thanks

Questions