Dimension-free limits of SGD for two-layers neural networks

(and some applications)

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

Based on
arXiv: 2202.00293,
2302.05882, 2305.18502

Statistical Physics and Machine Learning back together again
08.08.2023

Dimension-free limits of SGD for two-layers neural networks

(and some applications)

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

Based on
arXiv: 2202.00293,
2302.05882, 2305.18502

Statistical Physics and Machine Learning back together again
08.08.2023

In collaboration with

Luca Arnaboldi
(EPFL)

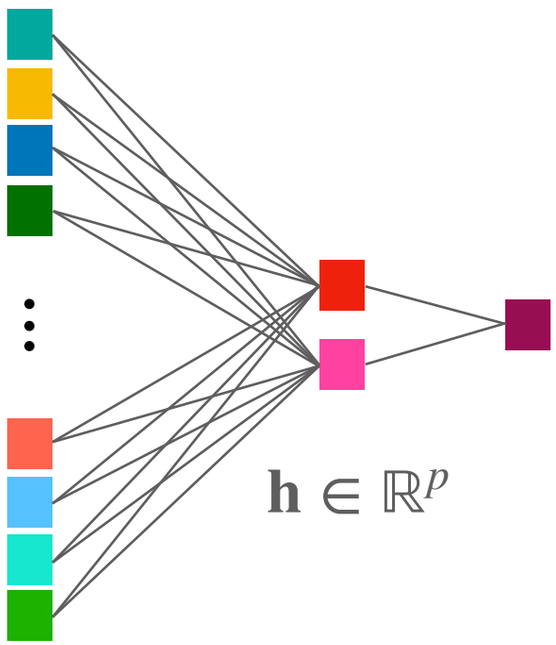
Rodrigo Veiga
(EPFL)

Ludovic Stéphan
(EPFL)

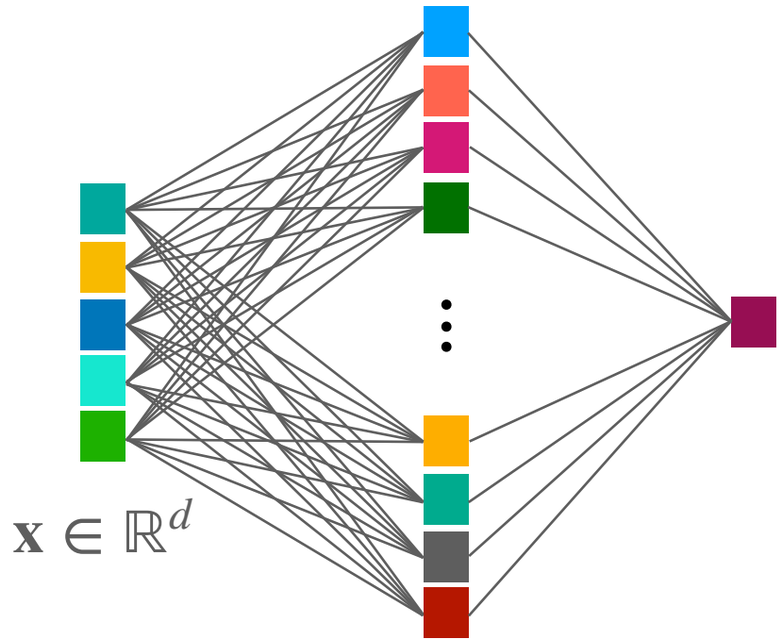
Lenka Zdeborová
(EPFL)

Florent Krzakala
(EPFL)

Narrow networks



Wide networks



$\frac{p}{d}$

$p \ll d$

???

$p \gg d$

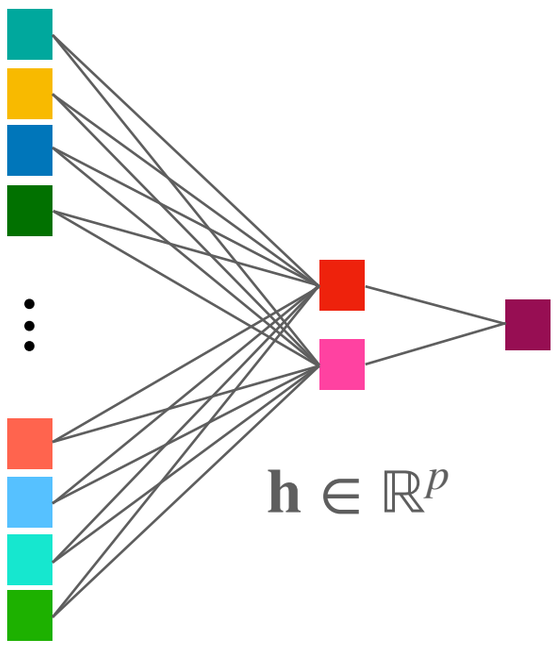
(Saad & Solla)

[Saad & Solla '95;
Goldt et al. '19;
Ben Arous, Gheissari,
Jagannath '21, '22]

(Mean-field limit)

[Mei, Montanari, Nguyen 18';
Chizat, Bach 18';
Rotskoff, **Vanden-Eijnden** 18';
Sirignano, Spiliopoulos 18']

Narrow networks

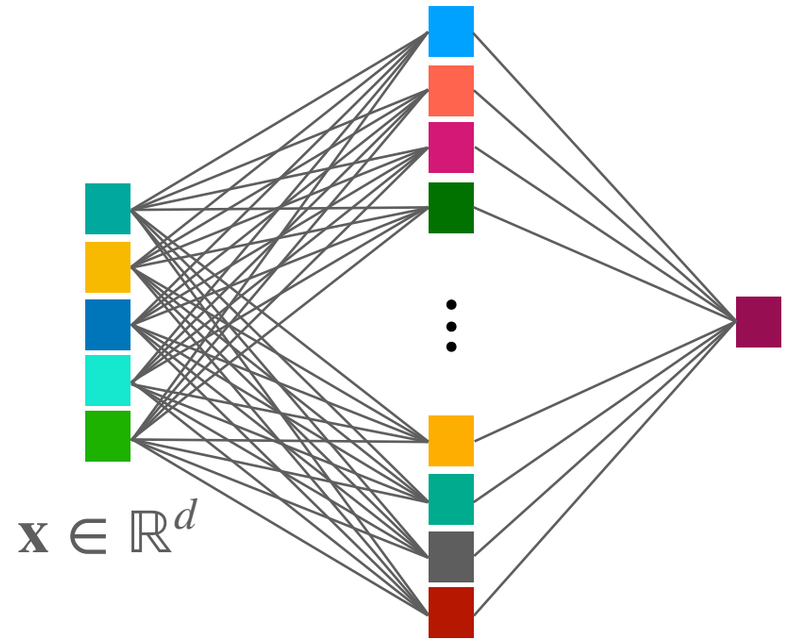


$p \ll d$

(Saad & Solla)

[Saad & Solla '95;
Goldt et al. '19;
Ben Arous, Gheissari,
Jagannath '21, '22]

Wide networks



$p \gg d$

(Mean-field limit)

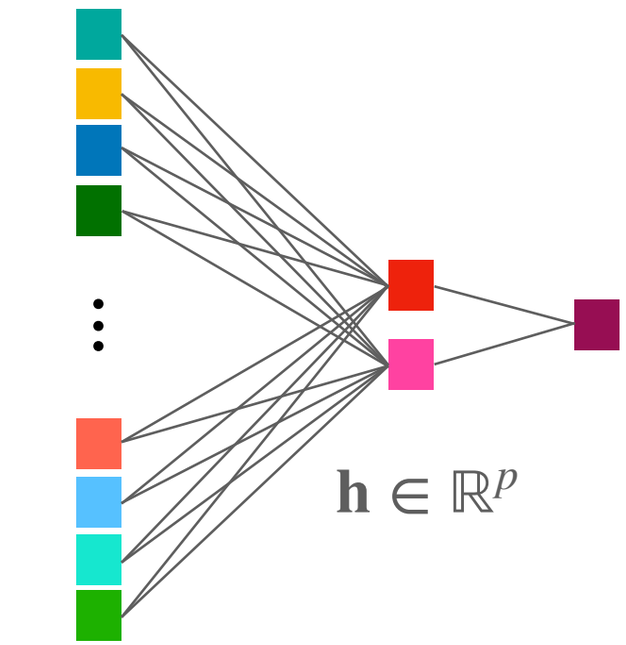
[Mei, Montanari, Nguyen 18';
Chizat, Bach 18';
Rotskoff, **Vanden-Eijnden** 18';
Sirignano, Spiliopoulos 18']

$\frac{p}{d}$

???

[Mei, **Misiakiewicz**, Montanari '19;
Abbe, Adsera, **Misiakiewicz** '22, '23;
Hajjar, Chizat '22;
Berthier, Montanari, Zhou '23]

Narrow networks

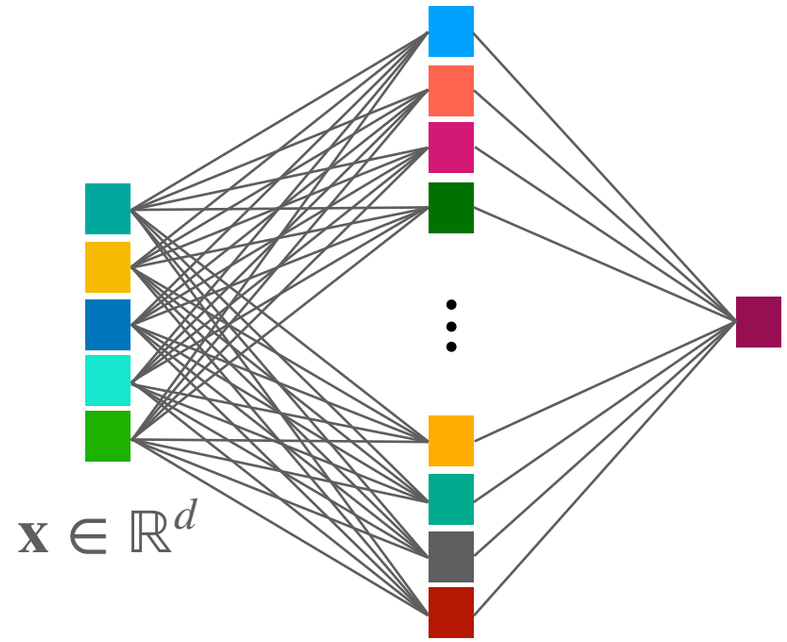


$p \ll d$

(Saad & Solla)

[Saad & Solla '95;
Goldt et al. '19;
Ben Arous, Gheissari,
Jagannath '21, '22]

Wide networks



$p \gg d$

(Mean-field limit)

[Mei, Montanari, Nguyen 18';
Chizat, Bach 18';
Rotskoff, **Vanden-Eijnden** 18';
Sirignano, Spiliopoulos 18']

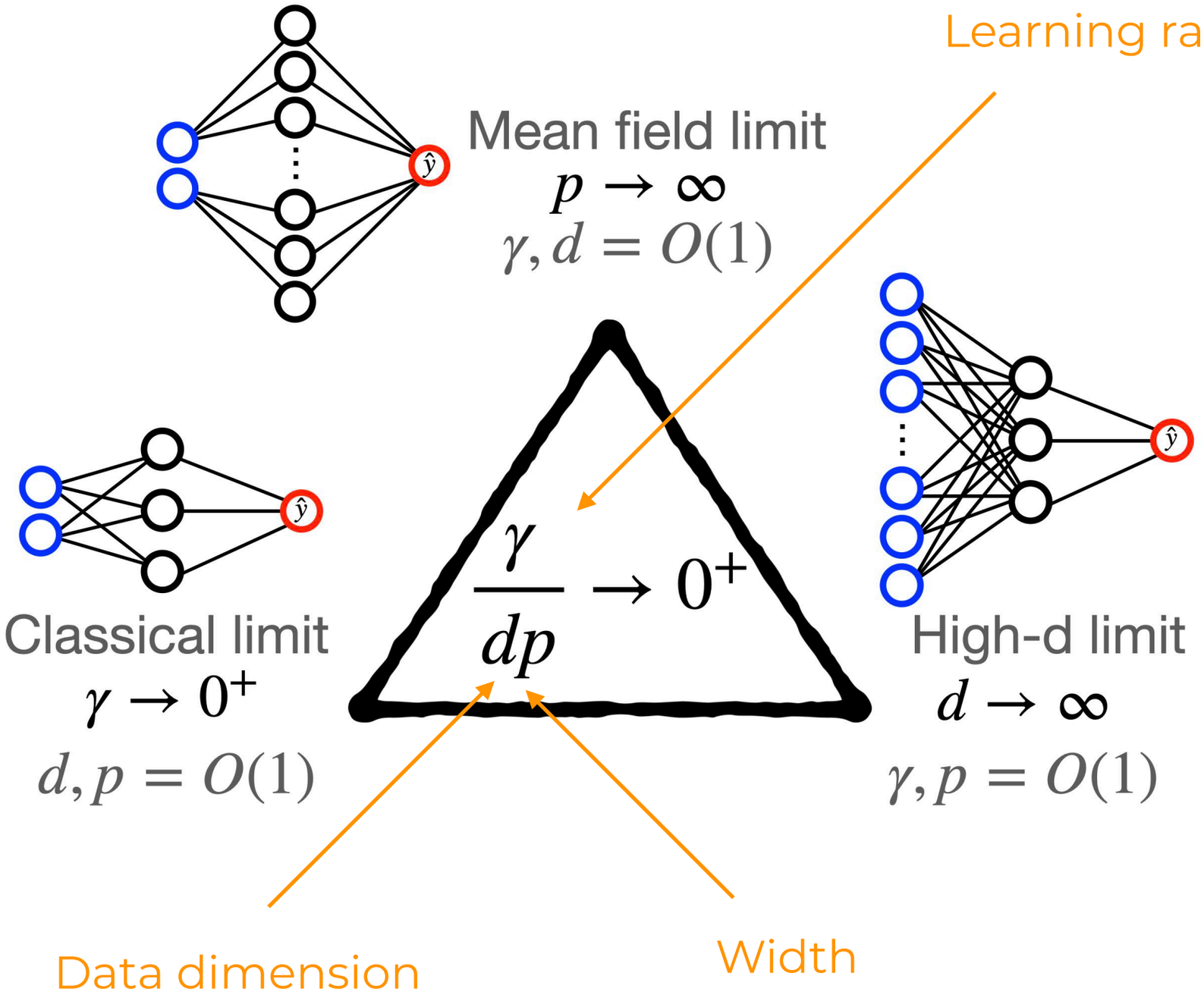
$\frac{p}{d}$

???

This work

[**Veiga et al. '22**;
Arnaboldi et al. '23]

SPOILER ALERT



Empirical risk minimisation

Let $(x^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^d \times \mathbb{R}$ ind. sampled from ρ .

Empirical risk minimisation

Let $(x^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^d \times \mathbb{R}$ ind. sampled from ρ .

Want: Function $\hat{y} = f_\Theta(x)$ that "generalises" well, i.e.

minimise $$\mathscr{R}(\Theta) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\rho}\left[\left(y - f_\Theta(x)\right)^2\right]$$

Population Risk

Empirical risk minimisation

Let $(x^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^d \times \mathbb{R}$ ind. sampled from ρ .

Want: Function $\hat{y} = f_\Theta(x)$ that "generalises" well, i.e.

minimise $$\mathscr{R}(\Theta) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\rho}\left[\left(y - f_\Theta(x)\right)^2\right]$$

Population Risk

☹ Problem: in practice, does't know ρ. So instead...

Empirical risk minimisation

Let $(x^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^d \times \mathbb{R}$ ind. sampled from ρ .

Want: Function $\hat{y} = f_\Theta(x)$ that "generalises" well, i.e.

minimise $$\mathscr{R}(\Theta) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\rho}\left[\left(y - f_\Theta(x)\right)^2\right]$$

Population Risk

Problem: in practice, does't know ρ. So instead...

minimise $$\hat{\mathscr{R}}_n(\Theta) = \frac{1}{2n}\sum_{\nu\in[n]}\left(y^\nu - f_\Theta(x^\nu)\right)^2$$

Empirical Risk

Algorithm: SGD

Algorithm: Let $b_k \subset [n]$ be mini-batch.

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \hat{\mathscr{R}}_{b_k}\left(\Theta^k\right)$$

$$\hat{\mathscr{R}}_b(\Theta) = \frac{1}{2|b|} \sum_{\nu \in b} \left(y^\nu - f_\Theta(x^\nu)\right)^2$$

mini-batch

Algorithm: SGD

Algorithm: Let $b_k \subset [n]$ be mini-batch.

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \hat{\mathscr{R}}_{b_k}\left(\Theta^k\right)$$

$$\hat{\mathscr{R}}_b(\Theta) = \frac{1}{2|b|} \sum_{\nu \in b} \left(y^\nu - f_\Theta(x^\nu)\right)^2$$

mini-batch

Gradient descent (GD)

$$b_k = [n], \quad \forall k$$

Algorithm: SGD

Algorithm: Let $b_k \subset [n]$ be mini-batch.

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \hat{\mathscr{R}}_{b_k}\left(\Theta^k\right)$$

$$\hat{\mathscr{R}}_b(\Theta) = \frac{1}{2|b|} \sum_{\nu \in b} \left(y^\nu - f_\Theta(x^\nu)\right)^2$$

mini-batch

Gradient descent (GD)

$$b_k = [n], \quad \forall k$$

$\gamma \to 0^+$ at fixed d, p:

$$\dot{\Theta}(t) = -\nabla_\Theta \hat{\mathscr{R}}_n\left(\Theta(t)\right)$$

Algorithm: SGD

Algorithm: Let $b_k \subset [n]$ be mini-batch.

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \hat{\mathscr{R}}_{b_k}\left(\Theta^k\right)$$

$$\hat{\mathscr{R}}_b(\Theta) = \frac{1}{2|b|} \sum_{\nu \in b} \left(y^\nu - f_\Theta(x^\nu)\right)^2$$

mini-batch

Gradient descent (GD)

$$b_k = [n], \quad \forall k$$

$\gamma \to 0^+$ at fixed d, p:

$$\dot{\Theta}(t) = -\nabla_\Theta \hat{\mathscr{R}}_n\left(\Theta(t)\right)$$

One-pass SGD

b_k independent

Algorithm: SGD

Algorithm: Let $b_k \subset [n]$ be mini-batch.

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \hat{\mathscr{R}}_{b_k}\left(\Theta^k\right)$$

$$\hat{\mathscr{R}}_b(\Theta) = \frac{1}{2|b|} \sum_{\nu \in b} \left(y^\nu - f_\Theta(x^\nu)\right)^2$$

mini-batch

Gradient descent (GD)

$$b_k = [n], \quad \forall k$$

$\gamma \to 0^+$ at fixed d, p:

$$\dot{\Theta}(t) = -\nabla_\Theta \hat{\mathscr{R}}_n\left(\Theta(t)\right)$$

One-pass SGD

b_k independent

$\gamma \to 0^+$ at fixed d, p:

$$\dot{\Theta}(t) = -\nabla_\Theta \mathscr{R}\left(\Theta(t)\right)$$

[Robbins & Monro 51']

Algorithm: SGD

Algorithm: Let $b_k \subset [n]$ be mini-batch.

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \hat{\mathscr{R}}_{b_k}\left(\Theta^k\right)$$

$$\hat{\mathscr{R}}_b(\Theta) = \frac{1}{2|b|} \sum_{\nu \in b} \left(y^\nu - f_\Theta(x^\nu)\right)^2$$

mini-batch

Gradient descent (GD)

$$b_k = [n], \quad \forall k$$

$\gamma \to 0^+$ at fixed d, p:

$$\dot{\Theta}(t) = -\nabla_\Theta \hat{\mathscr{R}}_n\left(\Theta(t)\right)$$

One-pass SGD

b_k independent

$\gamma \to 0^+$ at fixed d, p:

$$\dot{\Theta}(t) = -\nabla_\Theta \mathscr{R}\left(\Theta(t)\right)$$

[Robbins & Monro 51']

Another look at SGD

Rewrite SGD:

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \mathscr{R}\left(\Theta^k\right) + \gamma_k \varepsilon^k$$

GD on population

Effective Noise

Where:

$$\varepsilon^k = \nabla_{\Theta^k}\left[\mathscr{R}\left(\Theta^k\right) - \hat{\mathscr{R}}_{B_k}\left(\Theta^k\right)\right]$$

Another look at SGD

Rewrite SGD:

$$\Theta^{k+1} = \Theta^k - \gamma_k \nabla_{\Theta^k} \mathscr{R}\left(\Theta^k\right) + \gamma_k \varepsilon^k$$

GD on population

Effective Noise

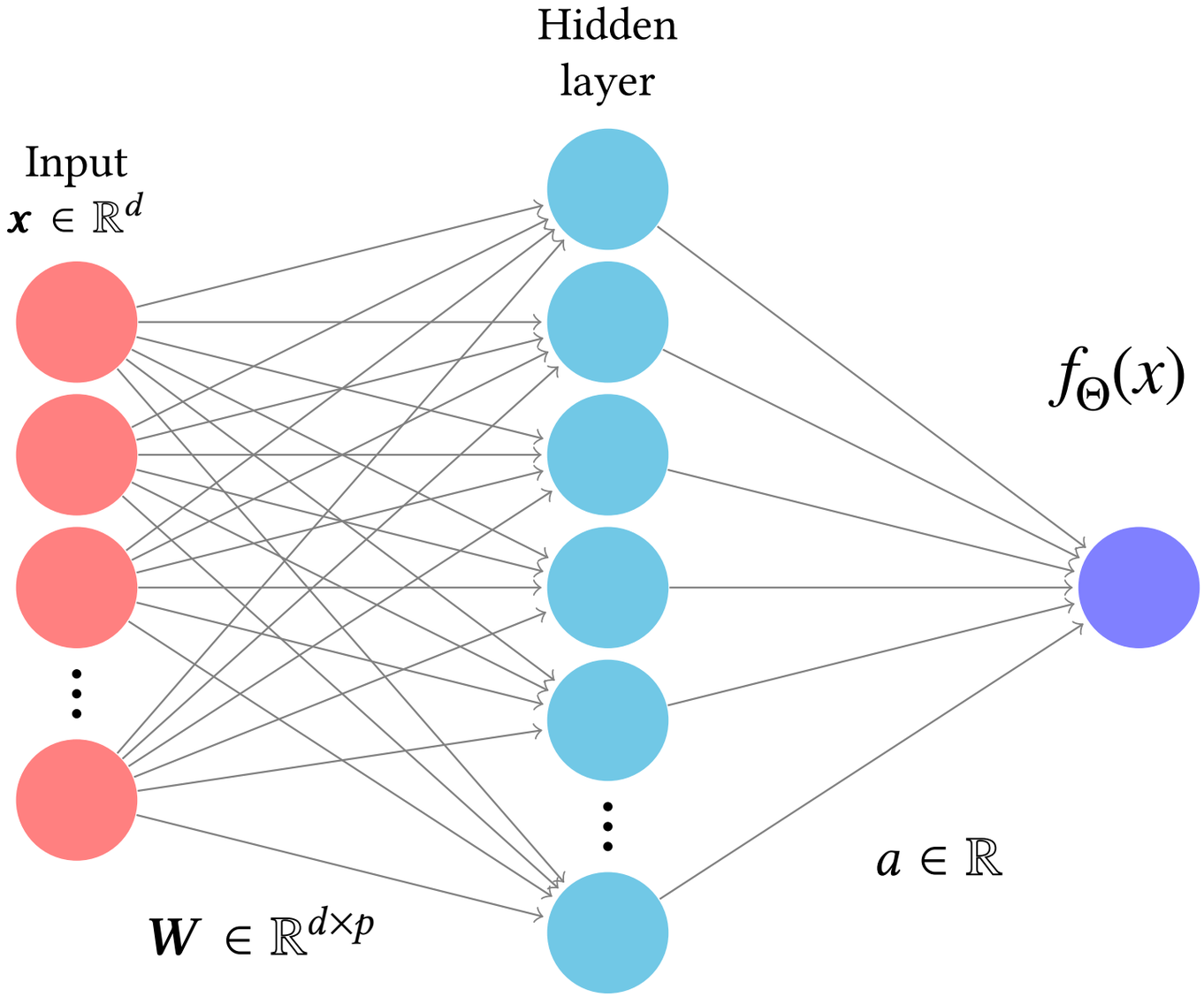
Where:

$$\varepsilon^k = \nabla_{\Theta^k}\left[\mathscr{R}\left(\Theta^k\right) - \hat{\mathscr{R}}_{B_k}\left(\Theta^k\right)\right]$$

🤔 Question: How to characterise this?

Two-layer neural networks

Let $(\mathbf{x}^\nu, y^\nu) \in \mathbb{R}^d \times \mathbb{R}$ denote $\nu = 1, \cdots, n$ i.i.d. samples from p

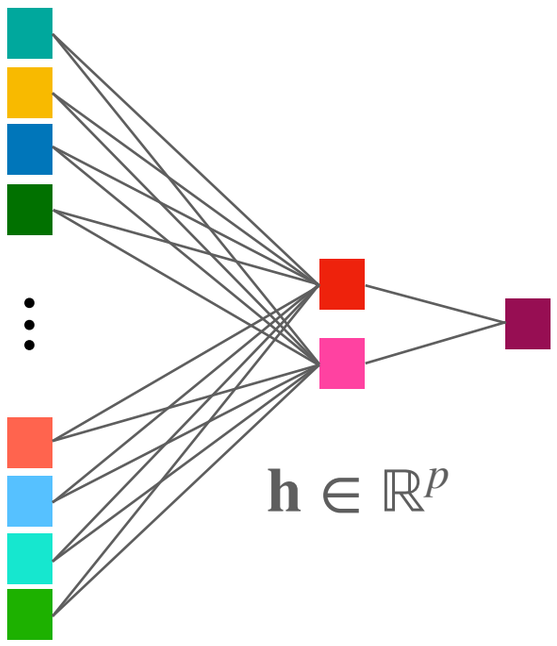


$$f_\Theta(x) = \sum_{i=1}^{p} a_i \sigma(w_i^\top x)$$

$$\Theta \equiv (\mathbf{a}, \mathbf{W}) \in \mathbb{R}^p \times \mathbb{R}^{p \times d}$$

What is known?

Narrow networks

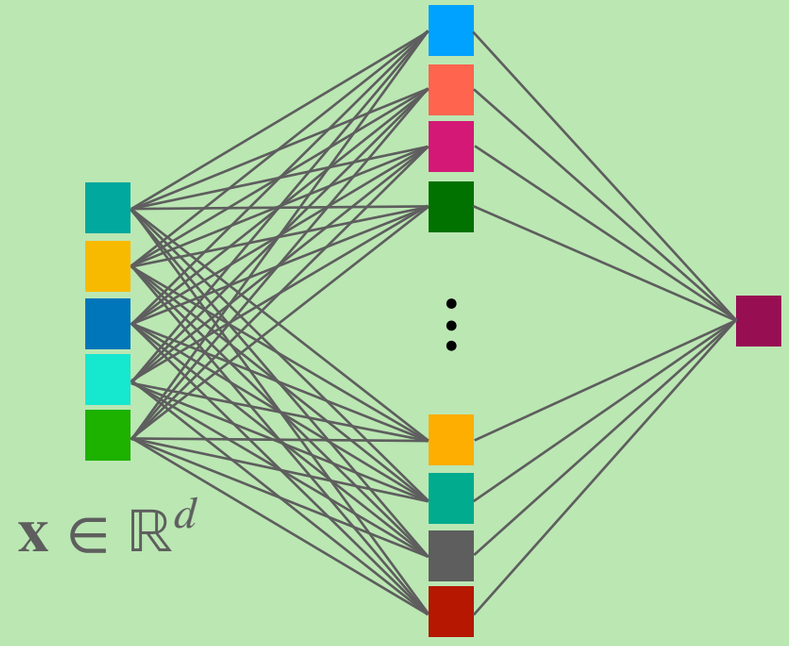


$p \ll d$

(Saad & Solla)

[Saad & Solla '95;
Goldt et al. '19;
Ben Arous, Gheissari,
Jagannath '21, '22]

Wide networks



$p \gg d$

(Mean-field limit)

[Mei, Montanari, Nguyen 18';
Chizat, Bach 18';
Rotskoff, **Vanden-Eijnden** 18';
Sirignano, Spiliopoulos 18']

$\frac{p}{d}$

Mean-field limit

On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport

Lénaïc Chizat
INRIA, ENS, PSL Research University
Paris, France
lenaic.chizat@inria.fr

Francis Bach
INRIA, ENS, PSL Research University
Paris, France
francis.bach@inria.fr

TRAINABILITY AND ACCURACY OF NEURAL NETWORKS: AN INTERACTING PARTICLE SYSTEM APPROACH

GRANT M. ROTSKOFF AND ERIC VANDEN-EIJNDEN

Mean field analysis of neural networks: A central limit theorem

Justin Sirignano[a,*], Konstantinos Spiliopoulos[b,1]

A mean field view of the landscape of two-layer neural networks

Song Mei[a], Andrea Montanari[b,c,1], and Phan-Minh Nguyen[b]

Mean-field limit

Idea: Define empirical density of weights:

$$\rho_p^{\nu}(\Theta) = \frac{1}{p}\sum_{i=1}^{p}\delta(\theta - \theta_i^{\nu}) \qquad\qquad \theta_i = (a_i, w_i) \in \mathbb{R}^{d+1}$$

Mean-field limit

Idea: Define empirical density of weights:

$$\rho_p^{\nu}(\Theta) = \frac{1}{p}\sum_{i=1}^{p}\delta(\theta - \theta_i^{\nu}) \qquad \theta_i = (a_i, w_i) \in \mathbb{R}^{d+1}$$

The risk is linear in $\hat{\rho}_p$!

$$\mathscr{R}(\Theta) = \mathbb{E}\left(y - \int \hat{\rho}_p(\mathrm{d}a, \mathrm{d}w) a\sigma(w \cdot x)\right)^2$$

Mean-field limit

Idea: Define empirical density of weights:

$$\rho_p^{\nu}(\Theta) = \frac{1}{p}\sum_{i=1}^{p}\delta(\theta - \theta_i^{\nu}) \qquad\qquad \theta_i = (a_i, w_i) \in \mathbb{R}^{d+1}$$

The risk is linear in $\hat{\rho}_p$!

$$\mathscr{R}(\Theta) = \mathbb{E}\left(y - \int \hat{\rho}_p(\mathrm{d}a, \mathrm{d}w) a\sigma(w \cdot x)\right)^2$$

Show that, at fixed d and $\gamma_k \ll 1/d$:

One-pass SGD $\xrightarrow{p \to \infty}$ $\partial_t \rho_t = \gamma \nabla_\theta\left(\rho_t \nabla_\theta \mathscr{C}(\theta; \rho_t)\right)$

"Mean-field" limit

[Mei, Montanari, Nguyen 18'; Chizat, Bach 18'; Rotskoff, Vanden-Eijnden 18'; Sirignano, Spiliopoulos 18']

Global convergence

From [Chizat, Bach 21', arXiv: 2110.08084]

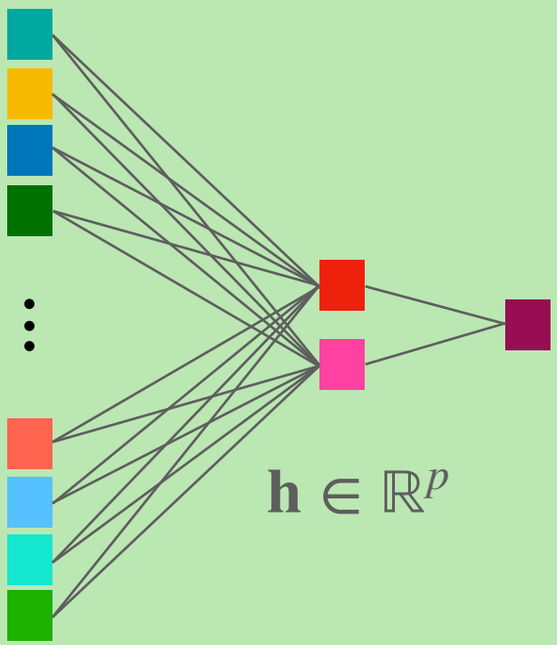
Theorem 2 (Informal) *If the support of the initial distribution includes all directions in* $\mathbb{R}^{d+1}$, *and if the function* Ψ *is positively 2-homogeneous then if the Wasserstein gradient flow weakly converges to a distribution, it can only be to a global optimum of* F.

From qualitative to quantitative results? Our result states that for infinitely many particles, we can only converge to a global optimum (note that we cannot show that the flow always converges). However, it is only a qualitative result in comparison with what is known for convex optimization problems in Section 2.2:

- This is only for $m = +\infty$, and we cannot provide an estimation of the number of particles needed to approximate the mean field regime that is not exponential in t (see such results e.g. in [28]).
- We cannot provide an estimation of the performance as the function of time, that would provide an upper bound on the running time complexity.

[Mei, Montanari, Nguyen 18'; Chizat, Bach 18'; Rotskoff, Vanden-Eijnden 18'; Sirignano, Spiliopoulos 18']

Narrow networks

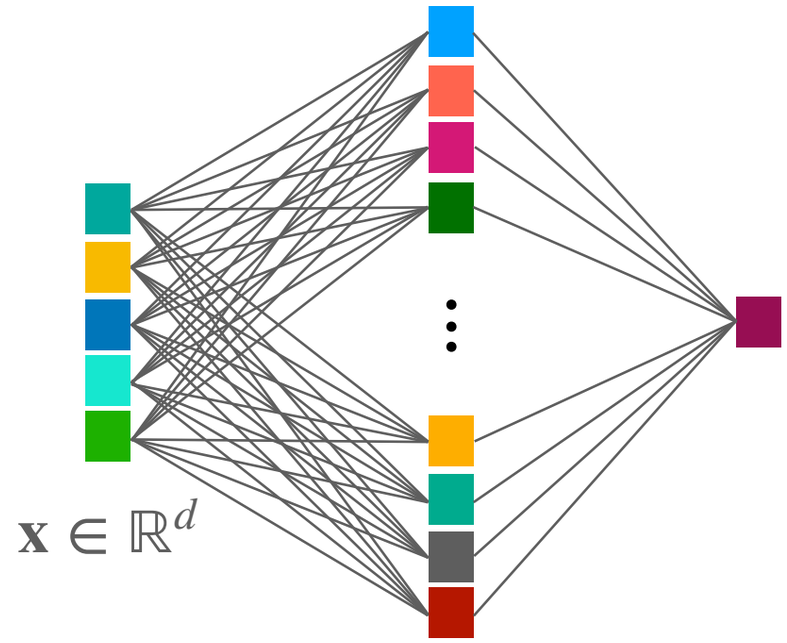


$p \ll d$

(Saad & Solla)

[Saad & Solla '95;
Goldt et al. '19;
Ben Arous, Gheissari,
Jagannath '21, '22]

Wide networks



$p \gg d$

(Mean-field limit)

[Mei, Montanari, Nguyen 18';
Chizat, Bach 18';
Rotskoff, **Vanden-Eijnden** 18';
Sirignano, Spiliopoulos 18']

$\frac{p}{d}$

Narrow networks

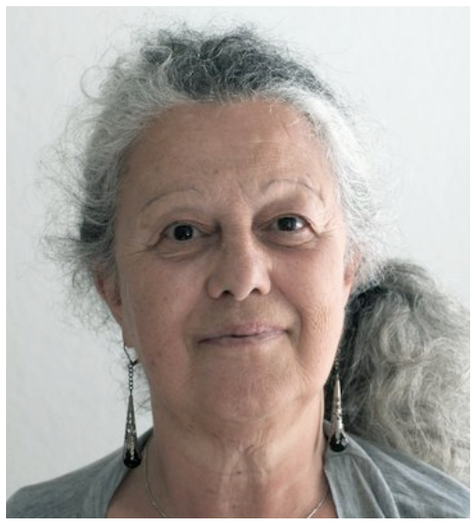
VOLUME 74, NUMBER 21 PHYSICAL REVIEW LETTERS 22 MAY 1995

Exact Solution for On-Line Learning in Multilayer Neural Networks

David Saad[1] and Sara A. Solla[2]
[1]*Department of Physics, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom*
[2]*CONNECT, The Niels Bohr Institute, Blegdamsdvej 17, Copenhagen 2100, Denmark*
(Received 14 October 1994)

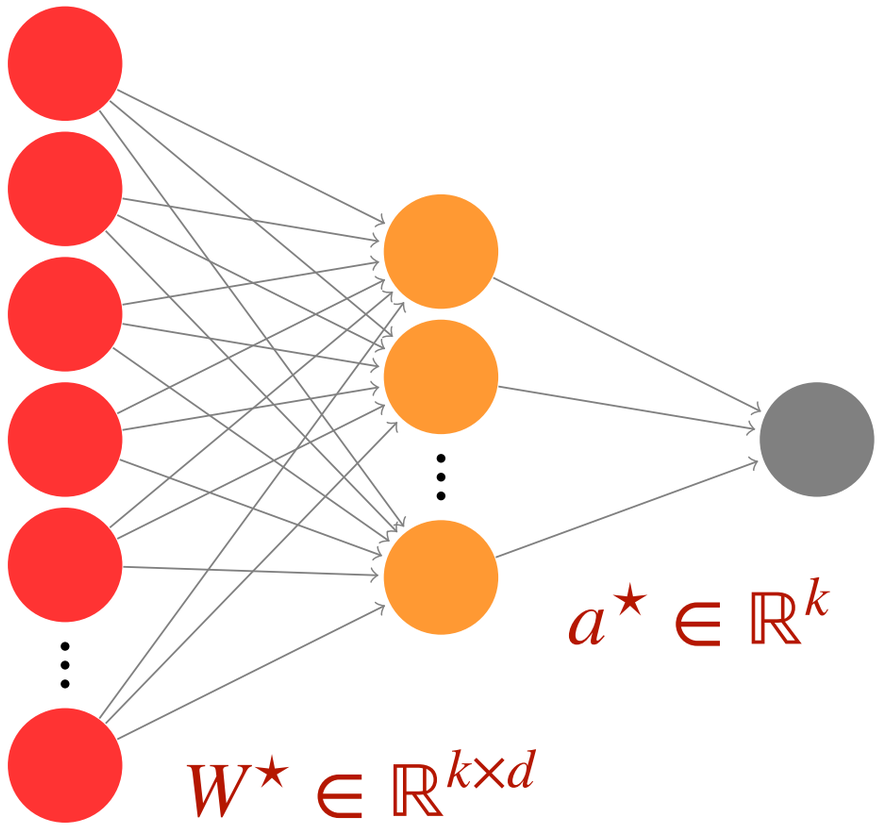
We present an analytic solution to the problem of on-line gradient-descent learning for two-layer neural networks with an arbitrary number of hidden units in both teacher and student networks.

PACS numbers: 87.10.+e, 02.50.–r, 05.20.–y



Teacher-student setting

Teacher network



$$x^\nu \sim \mathcal{N}(0, I_d) \qquad z^\nu \sim \mathcal{N}(0,1)$$

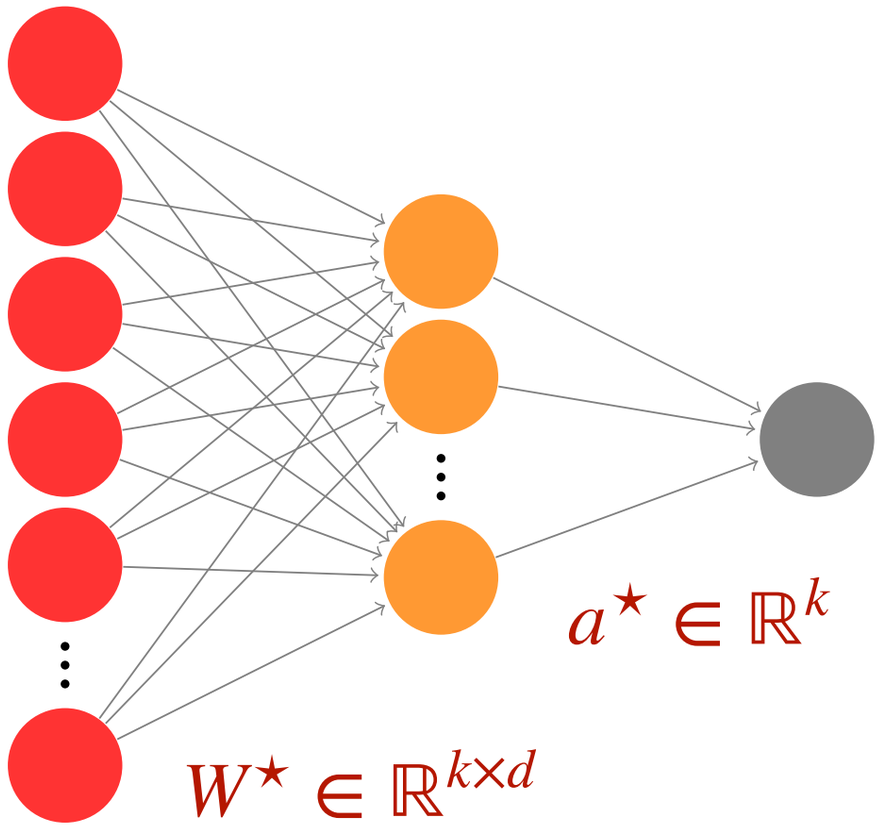
$$y^\nu = f_{W*}(x^\nu) + \sqrt{\Delta} z^\nu$$

$$f_{\Theta*}(x) = \frac{1}{k}\sum_{r=1}^{k} a_r^\star \sigma(w_r^{*\top} x)$$

Teacher-student setting

$$x^\nu \sim \mathcal{N}(0, I_d) \qquad z^\nu \sim \mathcal{N}(0,1)$$

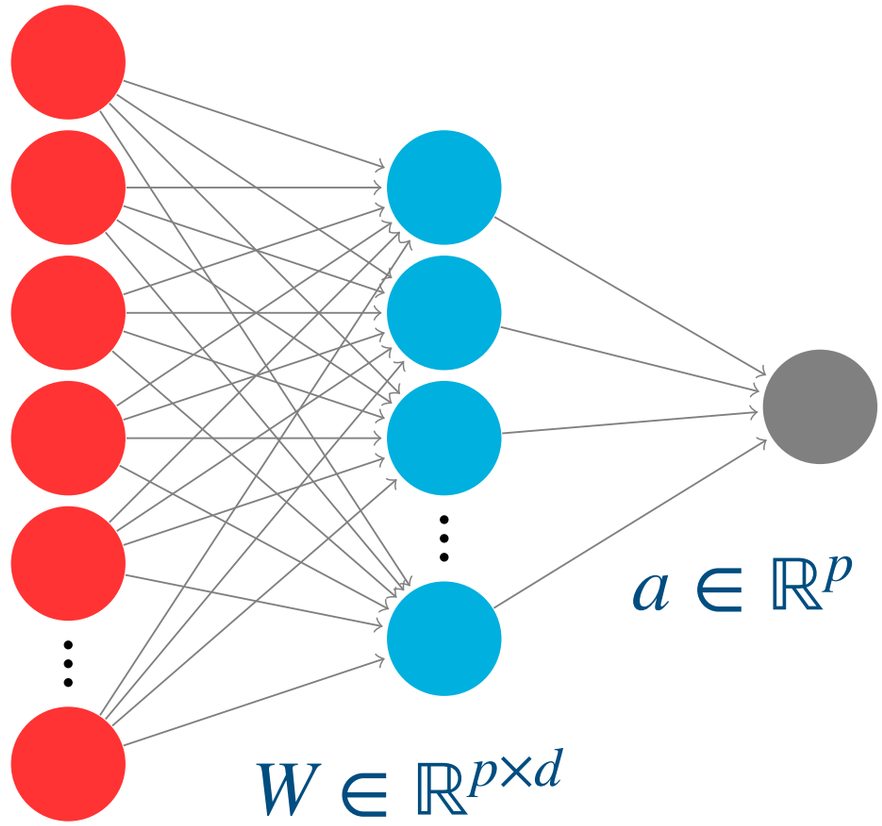
Teacher network



$$f_{\Theta*}(x) = \frac{1}{k} \sum_{r=1}^{k} a_r^\star \sigma(w_r^{*\top} x)$$

$$y^\nu = f_{\Theta*}(x^\nu) + \sqrt{\Delta} z^\nu$$

Student network



$$f_\Theta(x) = \frac{1}{p} \sum_{i=1}^{p} a_i \sigma(w_i^\top x)$$

Sufficient statistics

Goal: track population error exactly throughout the dynamics

$$\mathscr{R}(a^{\nu}, w^{\nu}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathrm{I_d})}\left[\left(\frac{1}{k}\sum_{r=1}^{k} a_r^{\star}\sigma(w_r^{*\top}x) - \frac{1}{p}\sum_{i=1}^{p} a_i^{\nu}\sigma(w_i^{\nu\top}x)\right)^2\right] + \frac{\Delta}{2}$$

Sufficient statistics

Goal: track population error exactly throughout the dynamics

$$\mathscr{R}(a^\nu, w^\nu) = \frac{1}{2}\mathbb{E}_{(\lambda^{\star\nu}, \lambda^\nu)\sim\mathcal{N}(0,\Omega^\nu)}\left[\left(\frac{1}{k}\sum_{r=1}^{k} a_r^\star \sigma(\lambda_r^{*\nu}) - \frac{1}{p}\sum_{i=1}^{p} a_i \sigma(\lambda_i^\nu)\right)^2\right] + \frac{\Delta}{2}$$

Where:

$$\Omega^\nu = \begin{pmatrix} P & M^\nu \\ M^{\nu\top} & Q^\nu \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)}$$

Sufficient statistics

Goal: track population error exactly throughout the dynamics

$$\mathscr{R}(a^\nu, w^\nu) = \frac{1}{2}\mathbb{E}_{(\lambda^{\star\nu},\lambda^\nu)\sim\mathcal{N}(0,\Omega^\nu)}\left[\left(\frac{1}{k}\sum_{r=1}^{k} a_r^\star \sigma(\lambda_r^{*\nu}) - \frac{1}{p}\sum_{i=1}^{p} a_i \sigma(\lambda_i^\nu)\right)^2\right] + \frac{\Delta}{2}$$

Where:

$$\Omega^\nu = \begin{pmatrix} P & M^\nu \\ M^{\nu\top} & Q^\nu \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)}$$

Key idea:

One-pass SGD $\longrightarrow$

$$a^{\nu+1} = \psi_1(a^\nu, \Omega^\nu)$$
$$\Omega^{\nu+1} = \psi_2(a^\nu, \Omega^\nu)$$

Sufficient statistics

Goal: track population error exactly throughout the dynamics

$$\mathscr{R}(a^\nu, w^\nu) = \frac{1}{2}\mathbb{E}_{(\lambda^{\star\nu},\lambda^\nu)\sim\mathcal{N}(0,\Omega^\nu)}\left[\left(\frac{1}{k}\sum_{r=1}^{k} a_r^\star \sigma(\lambda_r^{*\nu}) - \frac{1}{p}\sum_{i=1}^{p} a_i \sigma(\lambda_i^\nu)\right)^2\right] + \frac{\Delta}{2}$$

Where:

$$\Omega^\nu = \begin{pmatrix} P & M^\nu \\ M^{\nu\top} & Q^\nu \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)}$$

Key idea:

One-pass SGD

$$\begin{aligned} a^{\nu+1} &= \psi_1(a^\nu, \Omega^\nu) \\ \Omega^{\nu+1} &= \psi_2(a^\nu, \Omega^\nu) \end{aligned} \xrightarrow{d\to\infty} \begin{aligned} \frac{\mathrm{d}\bar{a}(t)}{\mathrm{d}t} &= \bar{\psi}_1(\bar{a}(t), \bar{\Omega}(t)) \\ \frac{\mathrm{d}\bar{\Omega}(t)}{\mathrm{d}t} &= \bar{\psi}_2(\bar{a}(t), \bar{\Omega}(t)) \end{aligned}$$

[Saad & Solla '95; Goldt et al. '19]

Can we go beyond $d \to \infty$?

Setting

One-pass SGD for two-layer neural networks in the teacher-student setting.

Architecture: $$f_\Theta(x) = \frac{1}{p}\sum_{i=1}^{p} a_i \sigma(w_i \cdot x)$$

Data model: $$y^\nu = \frac{1}{k}\sum_{r=1}^{k} a_r^\star \sigma(w_r^\star \cdot x^\nu) + \sqrt{\Delta} z^\nu \qquad x^\nu \sim \mathcal{N}(0, I_d) \qquad z^\nu \sim \mathcal{N}(0,1)$$

Algorithm: $$\Theta^{\nu+1} = \Theta^\nu - \gamma_\nu \nabla_{\Theta^\nu} (y^\nu - f_{\Theta^\nu}(x^\nu))^2$$

Focus on realisable case $p \geq k$.

Taking a closer look

Looking more closely...

$$\Omega^{\nu+1} = \Omega^{\nu} + \delta t_{\nu}\ \psi(\Omega^{\nu})$$

$$M^{\nu+1} - M^{\nu} = \frac{\gamma}{dp}\ \psi^{M}_{GF}(\Omega^{\nu})$$

$$Q^{\nu+1} - Q^{\nu} = \frac{\gamma}{dp}\ \psi^{Q}_{GF}(\Omega^{\nu}) + \frac{\gamma^2}{dp^2}\ \psi_{noise}(\Omega^{\nu})$$

Taking a closer look

Looking more closely...

$$\Omega^{\nu+1} = \Omega^{\nu} + \delta t_{\nu}\ \psi(\Omega^{\nu})$$

$$M^{\nu+1} - M^{\nu} = \frac{\gamma}{dp}\ \psi_{GF}^{M}(\Omega^{\nu})$$

$$Q^{\nu+1} - Q^{\nu} = \frac{\gamma}{dp}\ \psi_{GF}^{Q}(\Omega^{\nu}) + \frac{\gamma^2}{dp^2}\ \psi_{noise}(\Omega^{\nu})$$

Population gradient

Noise

Main theoretical result

Define step-size $\delta t = \frac{\gamma}{dp}$ and $M(t), Q(t)$ such that:

$$M(\nu\delta t) = M^{\nu} \qquad\qquad Q(\nu\delta t) = Q^{\nu}$$

Main theoretical result

Define step-size $\delta t = \frac{\gamma}{dp}$ and $\Omega(t)$ such that:

$$\Omega(\nu\delta t) = \Omega^{\nu}$$

Theorem [Veiga, Stephan, BL, Krzakala, Zdeborová '22]

Then $\forall 0 \leq \nu \leq \left\lfloor \frac{n}{\delta t} \right\rfloor$:

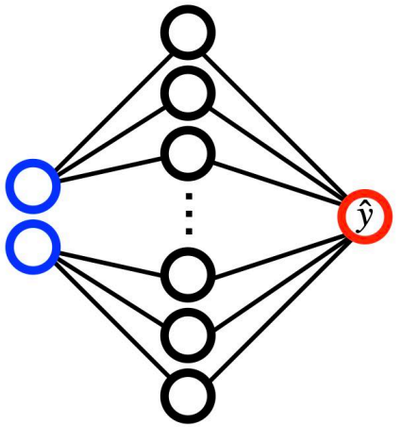
$$\mathbb{E}\,||\,\Omega^{\nu} - \bar{\Omega}(\nu\delta t)\,||_{\infty} \leq e^{C\nu\delta t}\sqrt{\frac{\gamma}{dp}}$$

Where $\bar{\Omega}(t) = \mathbb{E}[\Omega(t)]$ is the solution of an ODE:

$$\frac{\mathrm{d}\bar{\Omega}(\mathrm{t})}{\mathrm{dt}} = \mathbb{E}\left[\psi(\bar{\Omega}(t))\right]$$

Generalises [**Goldt et al**. '19], builds upon [Wang, Hu, **Lu** '18]

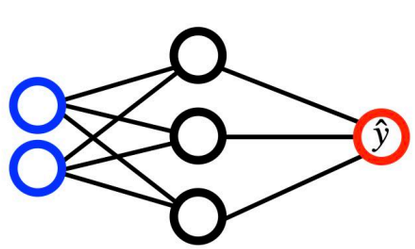
The different regimes



Mean field limit

$$p \to \infty$$

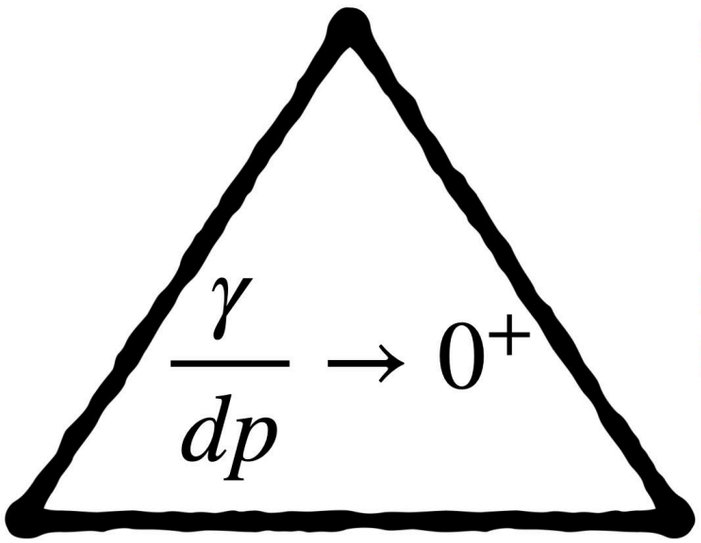
$$\gamma, d = O(1)$$

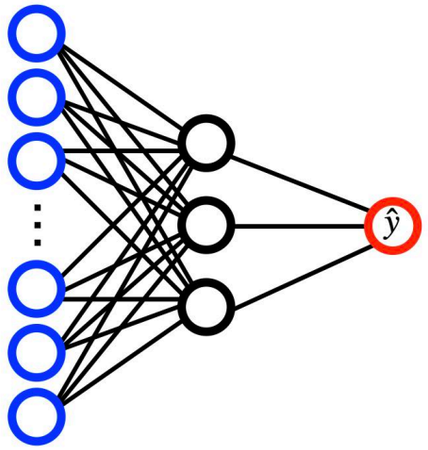


Classical limit

$$\gamma \to 0^+$$

$$d, p = O(1)$$



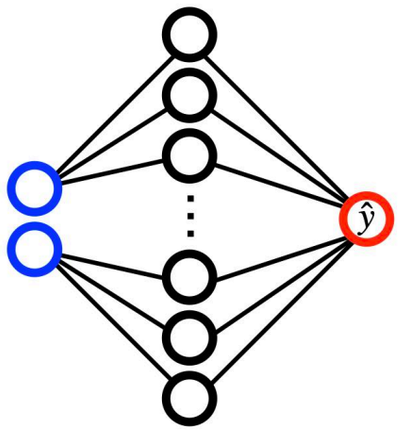


High-d limit

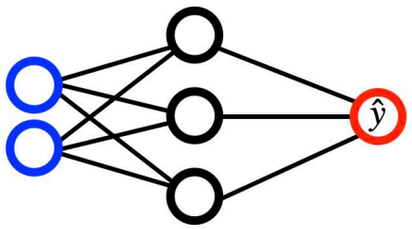
$$d \to \infty$$

$$\gamma, p = O(1)$$

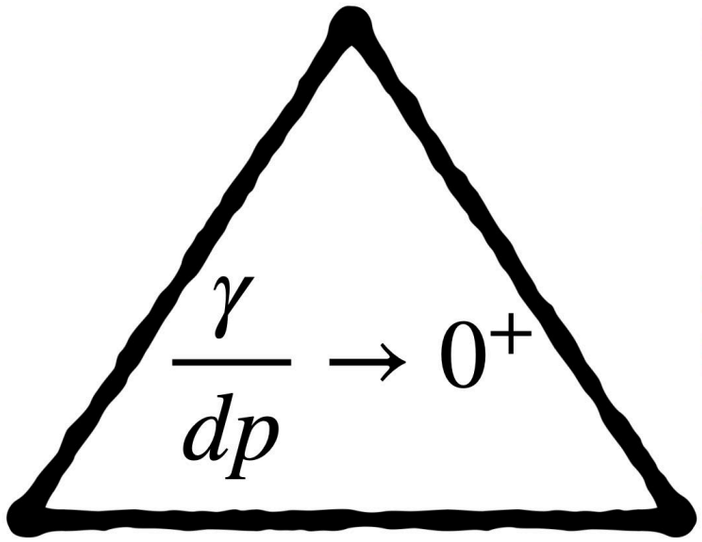
The different regimes

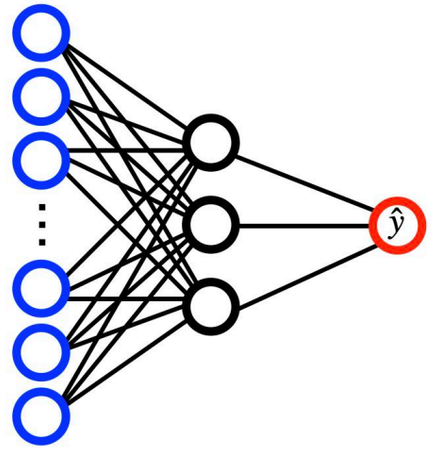


Mean field limit
$p \to \infty$
$\gamma, d = O(1)$



Classical limit
$\gamma \to 0^+$
$d, p = O(1)$



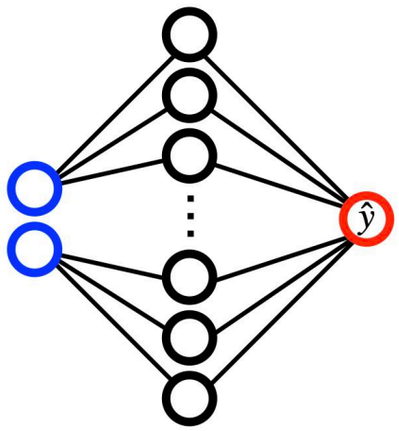


High-d limit
$d \to \infty$
$\gamma, p = O(1)$

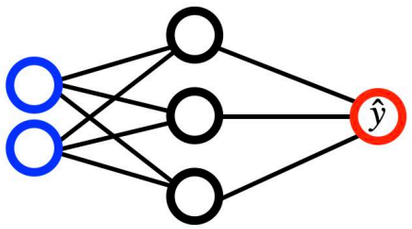
$$\dot{M}(t) = \psi^M_{\mathrm{GF}}(\bar{M}(t), \bar{Q}(t))$$

$$\dot{Q}(t) = \psi^Q_{\mathrm{GF}}(\bar{M}(t), \bar{Q}(t)) + \frac{\gamma}{p}\psi^Q_{\mathrm{noise}}(\bar{M}(t), \bar{Q}(t))$$

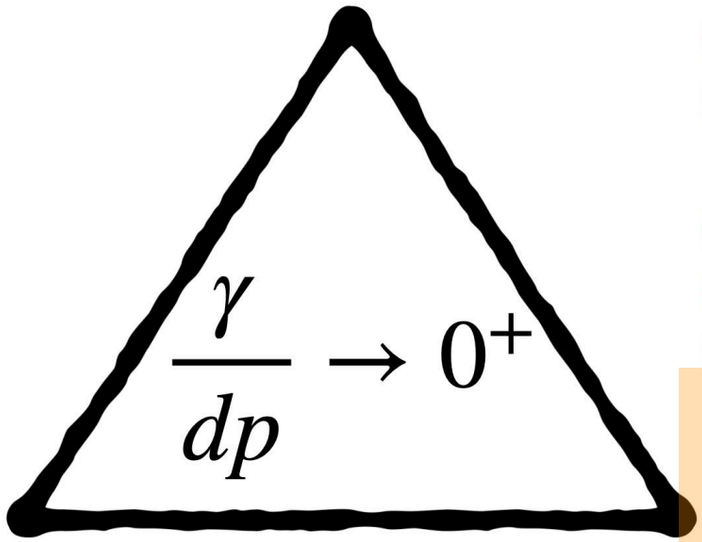
The different regimes

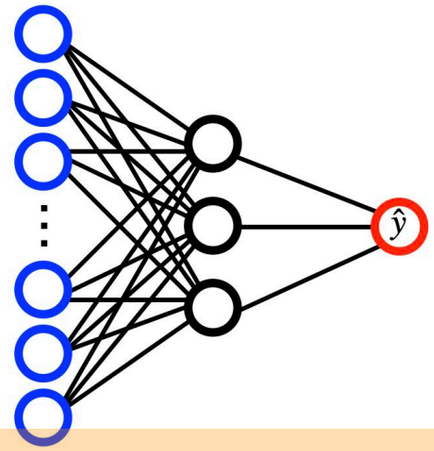


Mean field limit
$p \to \infty$
$\gamma, d = O(1)$



Classical limit
$\gamma \to 0^+$
$d, p = O(1)$



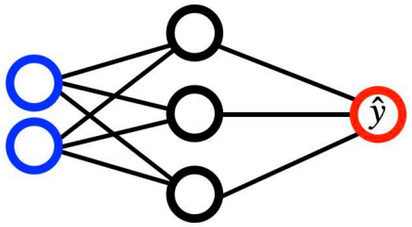


High-d limit
$d \to \infty$
$\gamma, p = O(1)$

$$\dot{M}(t) = \psi^M_{\mathrm{GF}}(\bar{M}(t), \bar{Q}(t))$$

$$\dot{Q}(t) = \psi^Q_{\mathrm{GF}}(\bar{M}(t), \bar{Q}(t)) + \frac{\gamma}{p}\psi^Q_{\mathrm{noise}}(\bar{M}(t), \bar{Q}(t))$$

Classical regime



Classical limit

$$\gamma \to 0^+$$

$$d, p = O(1)$$

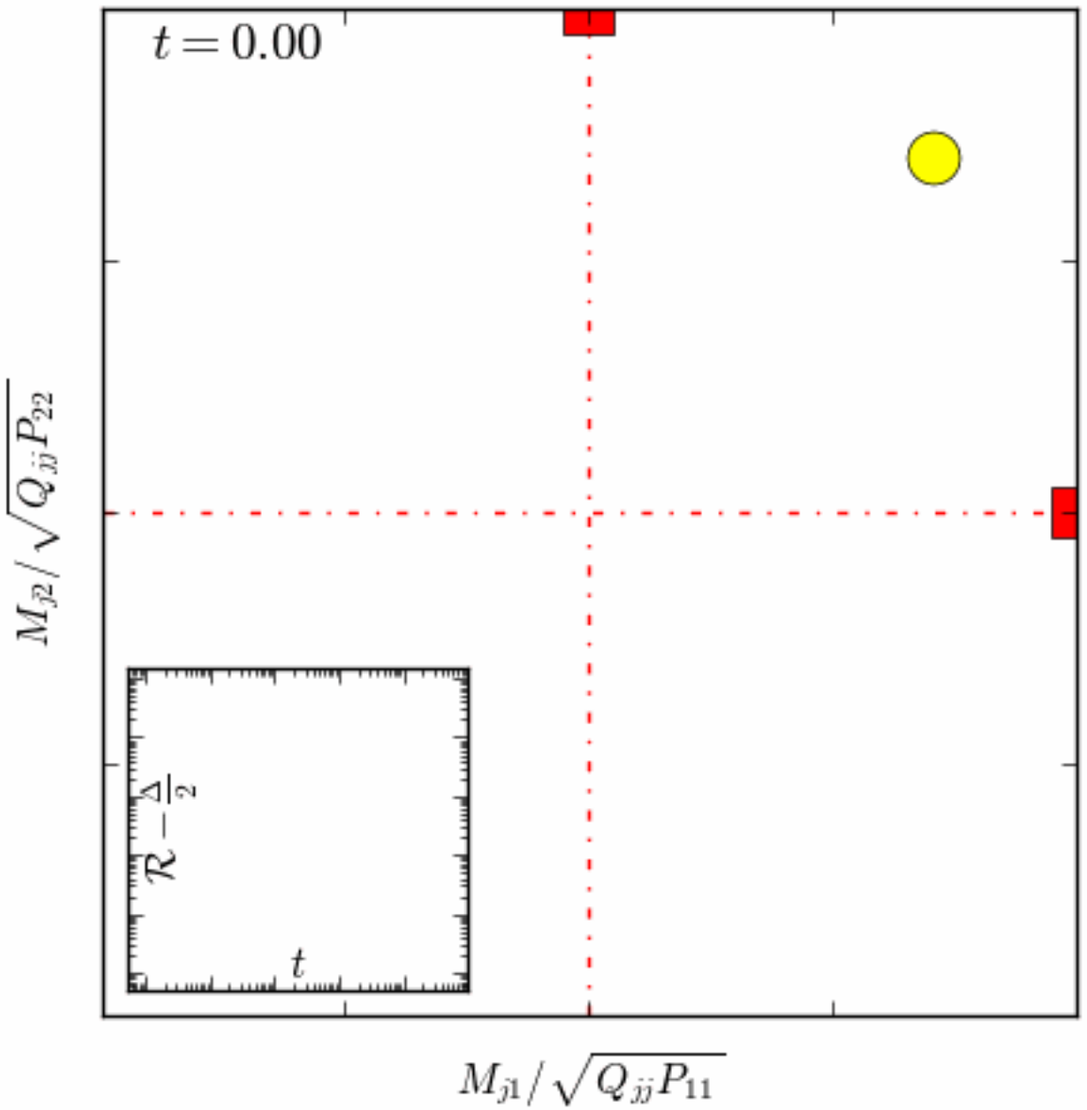
$$k = 2$$

$$p = 10$$

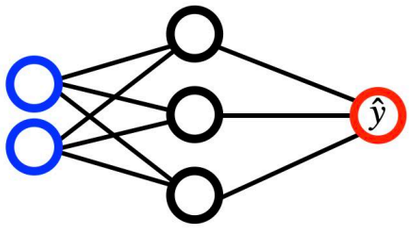
$$\Delta = 10^{-3}$$

$$\gamma = 5 \times 10^{-2}$$

$$\sigma = \sigma_\star = \mathrm{erf}$$



Classical regime



Classical limit

$$\gamma \to 0^+$$

$$d, p = O(1)$$

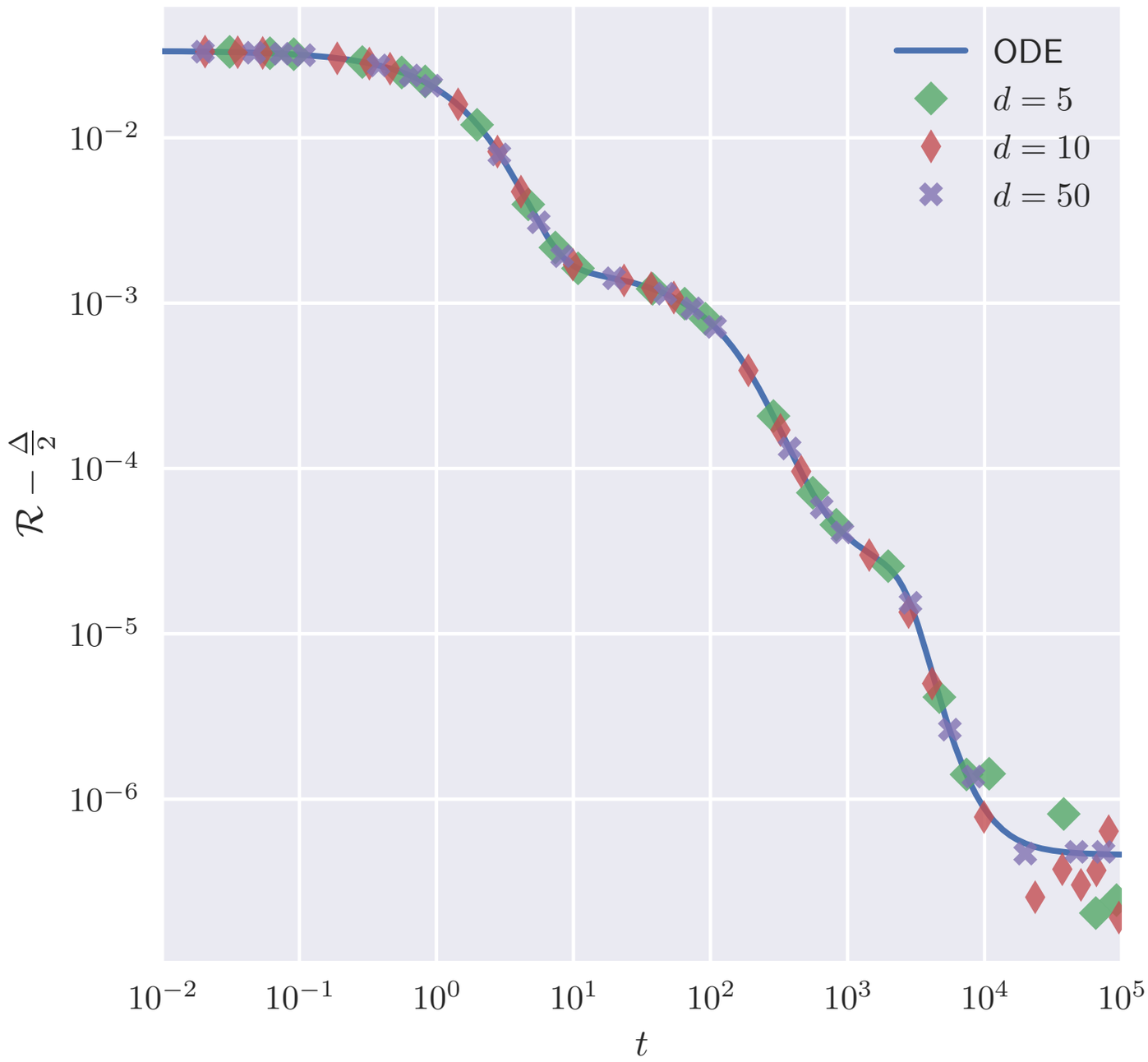
$$k = 5$$

$$p = 10$$

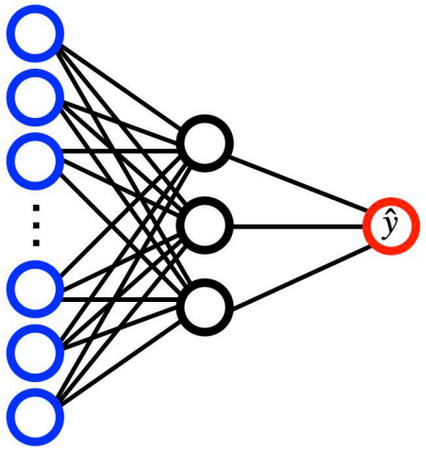
$$\Delta = 10^{-3}$$

$$\gamma = 5 \times 10^{-2}$$

$$\sigma = \sigma_\star = \mathrm{erf}$$



High-dimensional regime [Saad & Solla '95]



High-d limit

$$d \to \infty$$

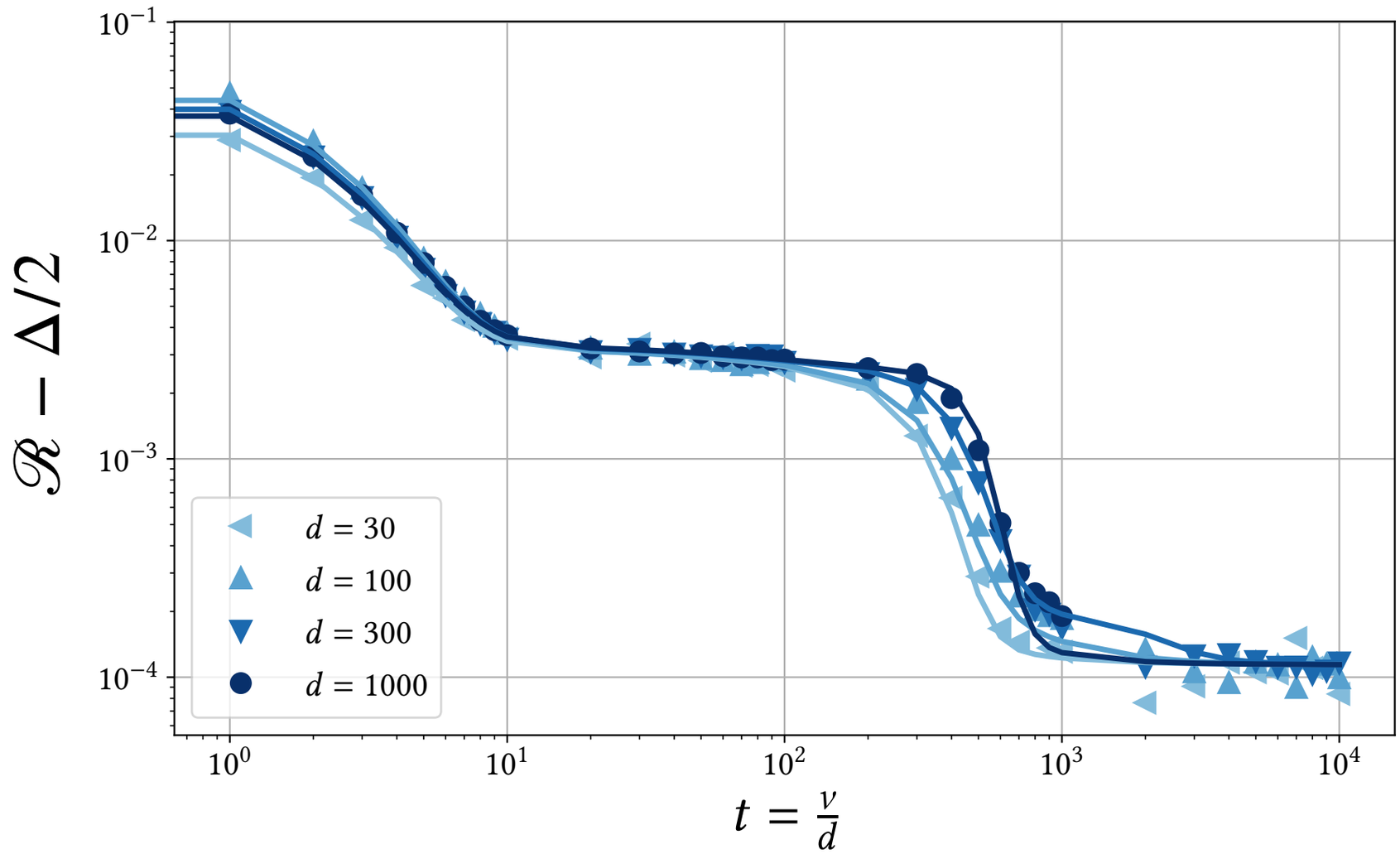
$$\gamma, p = O(1)$$

$$k = 4 \quad p = 8$$

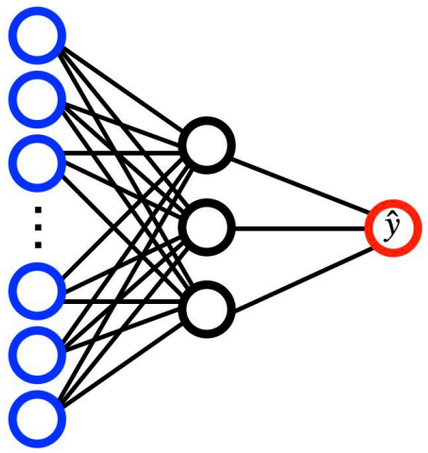
$$\Delta = 10^{-3}$$

$$\gamma = 0.1$$

$$\sigma = \sigma_\star = \mathrm{erf}$$



High-dimensional regime [Saad & Solla '95]



High-d limit

$$d \to \infty$$

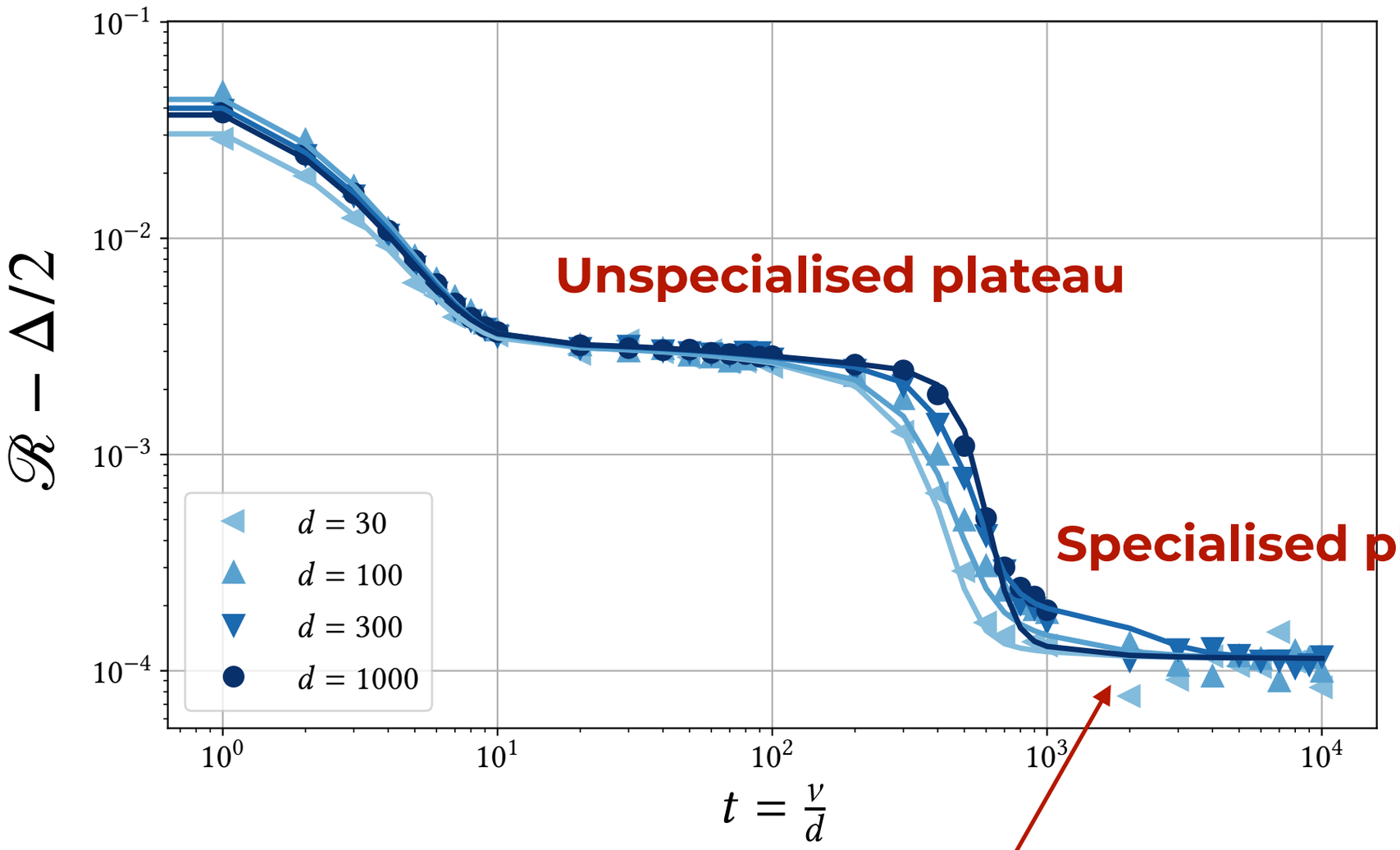
$$\gamma, p = O(1)$$

$$k = 4 \quad p = 8$$

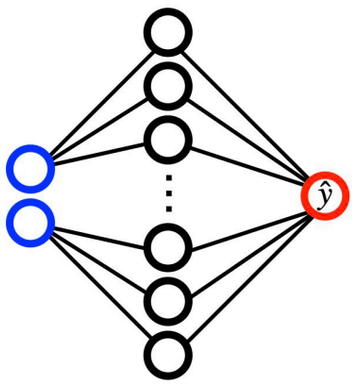
$$\Delta = 10^{-3}$$

$$\gamma = 0.1$$

$$\sigma = \sigma_\star = \mathrm{erf}$$



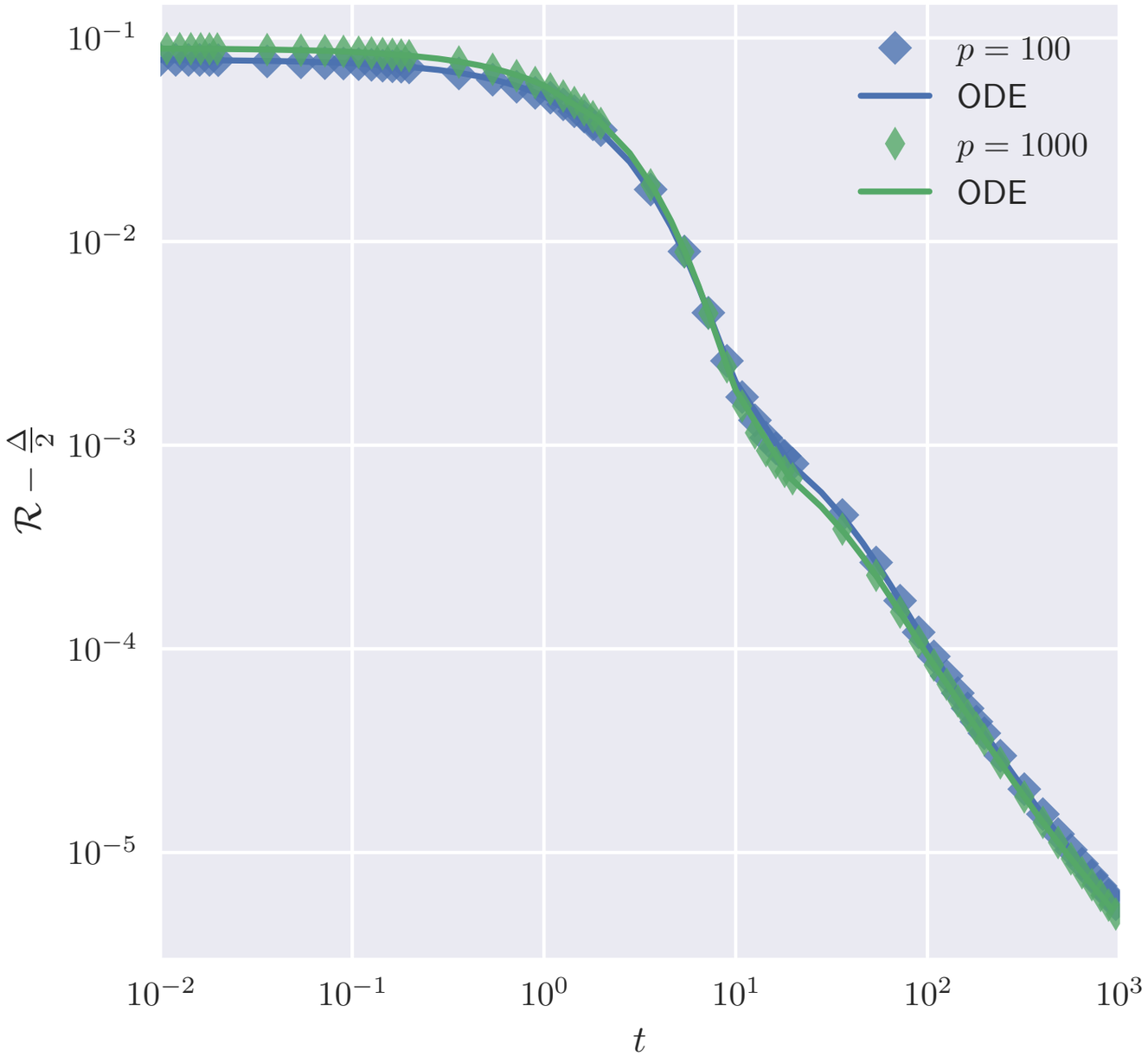
Mean-field regime



Mean field limit

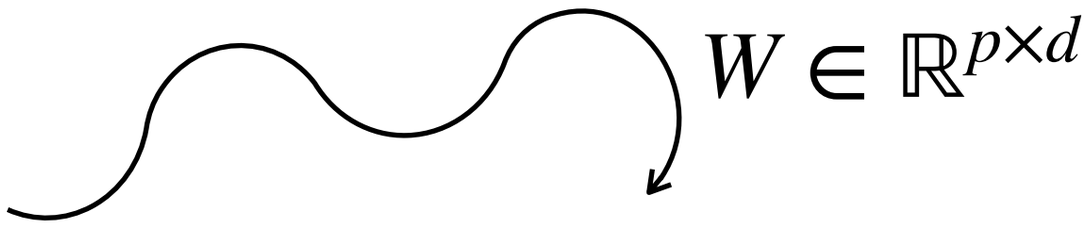
$p \to \infty$

$\gamma, d = O(1)$



Mean-field regime

🤔 But $Q \in \mathbb{R}^{p \times p}$!!!

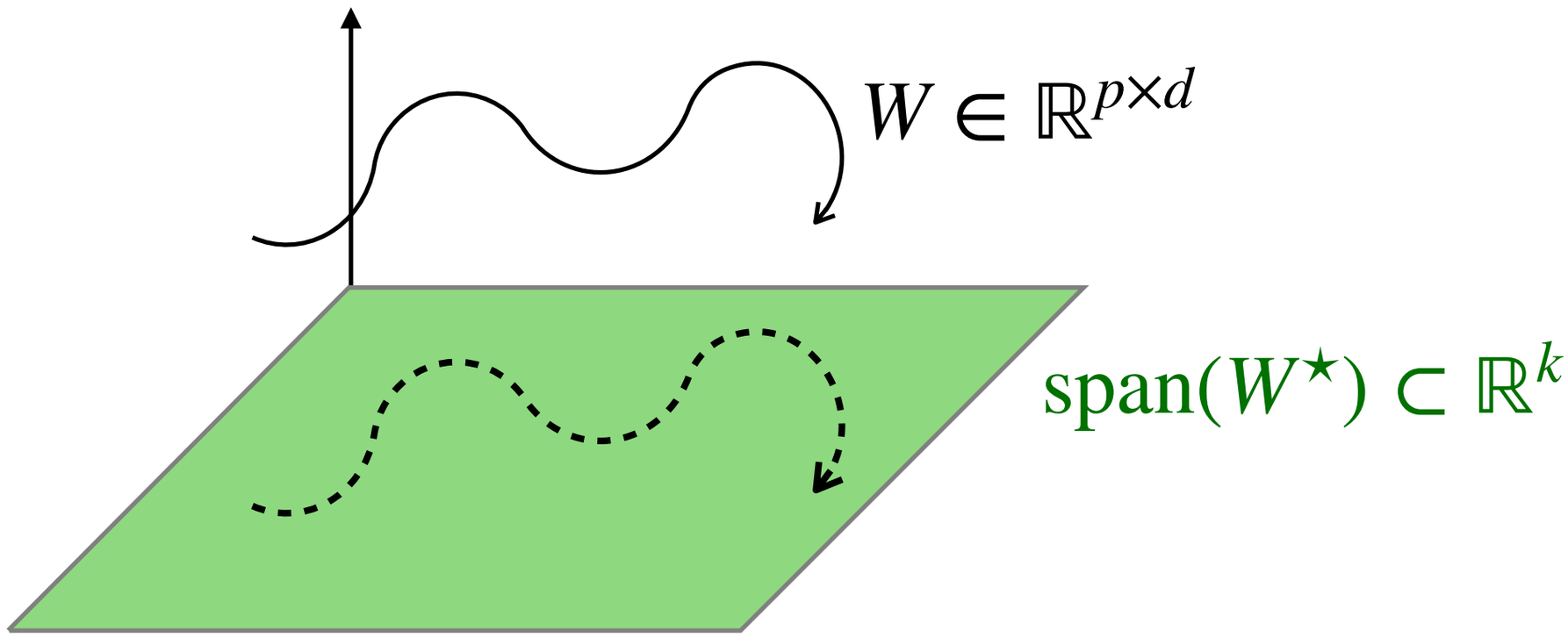


Mean-field regime

🤔 But $Q \in \mathbb{R}^{p \times p}$!!!

$$W = MP^{-1}W^\star + W^\perp$$

Teacher subspace

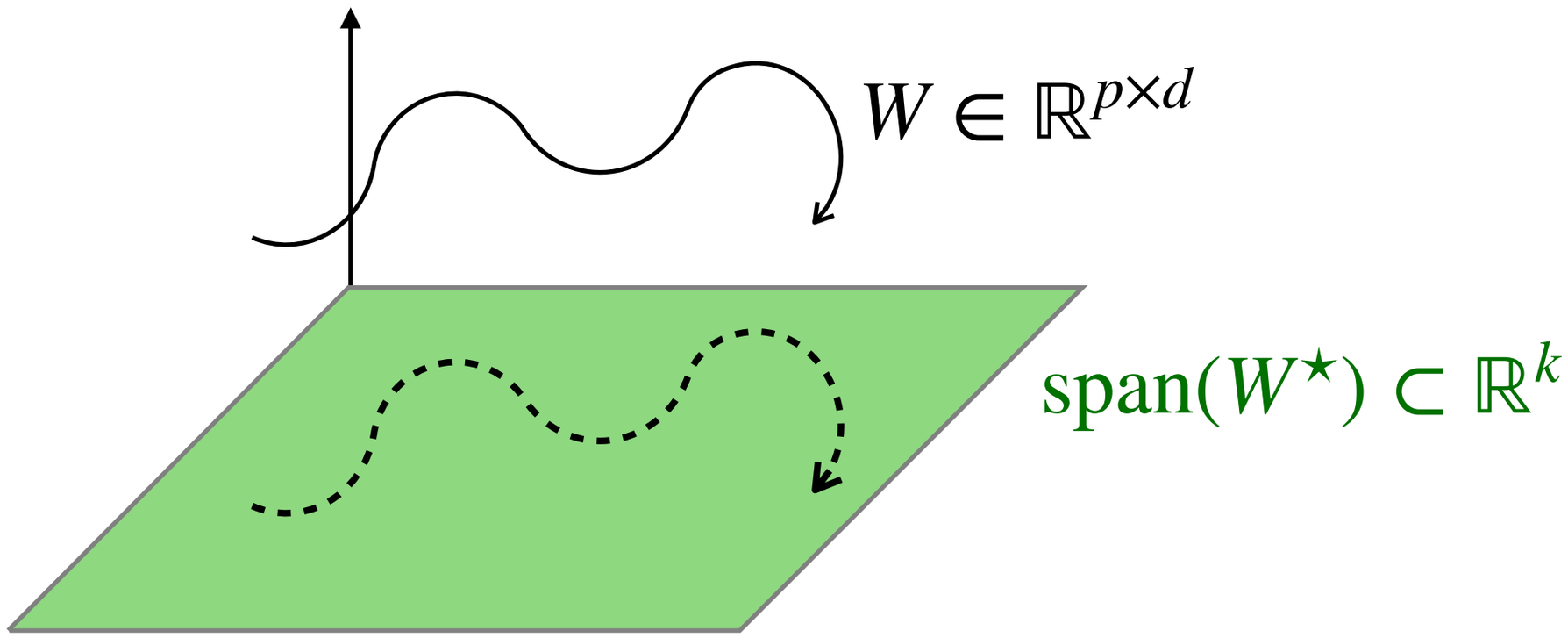


Mean-field regime

🤔 But $Q \in \mathbb{R}^{p \times p}$!!!

$$W = \underbrace{MP^{-1}W^\star}_{\text{Teacher subspace}} + W^\perp \xrightarrow[p \to \infty]{} Q \approx MPM^\top + D_{\sqrt{q^\perp}} \Xi D_{\sqrt{q^\perp}}$$

$$\Xi \sim \mathbb{S}^{d-k-1}$$

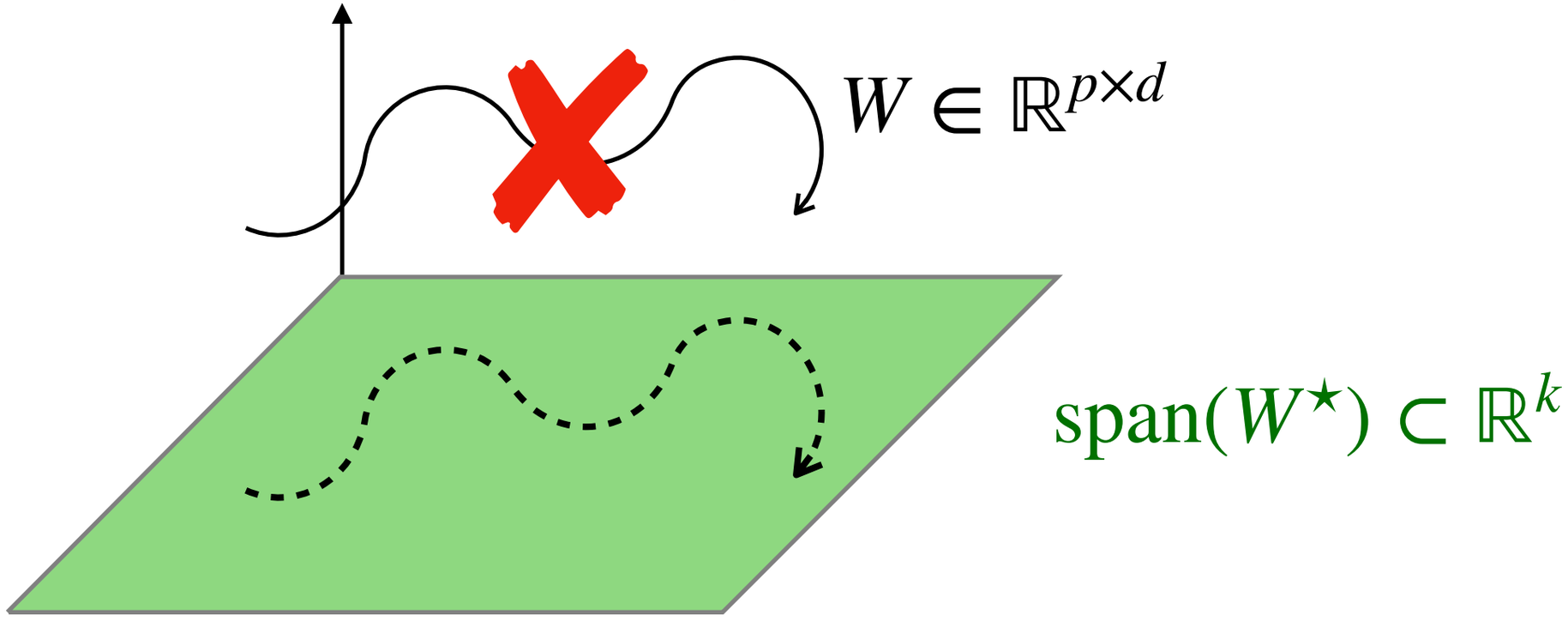


Mean-field + high-d

Theorem [Arnaboldi, Stephan, BL, Krzakala '23]

$$\mathbb{E}\,||\,Q(t) - MP^{-1}M^{\mathsf{T}} + \mathrm{diag}(Q^{\perp})\,||_{\infty} \leq e^{Ct}\left(p^{-1/2} + d^{-1/2}\right)$$

Suppressed in $1/\sqrt{d}$



Mean-field + high-d

Theorem [Arnaboldi, Stephan, BL, Krzakala '23]

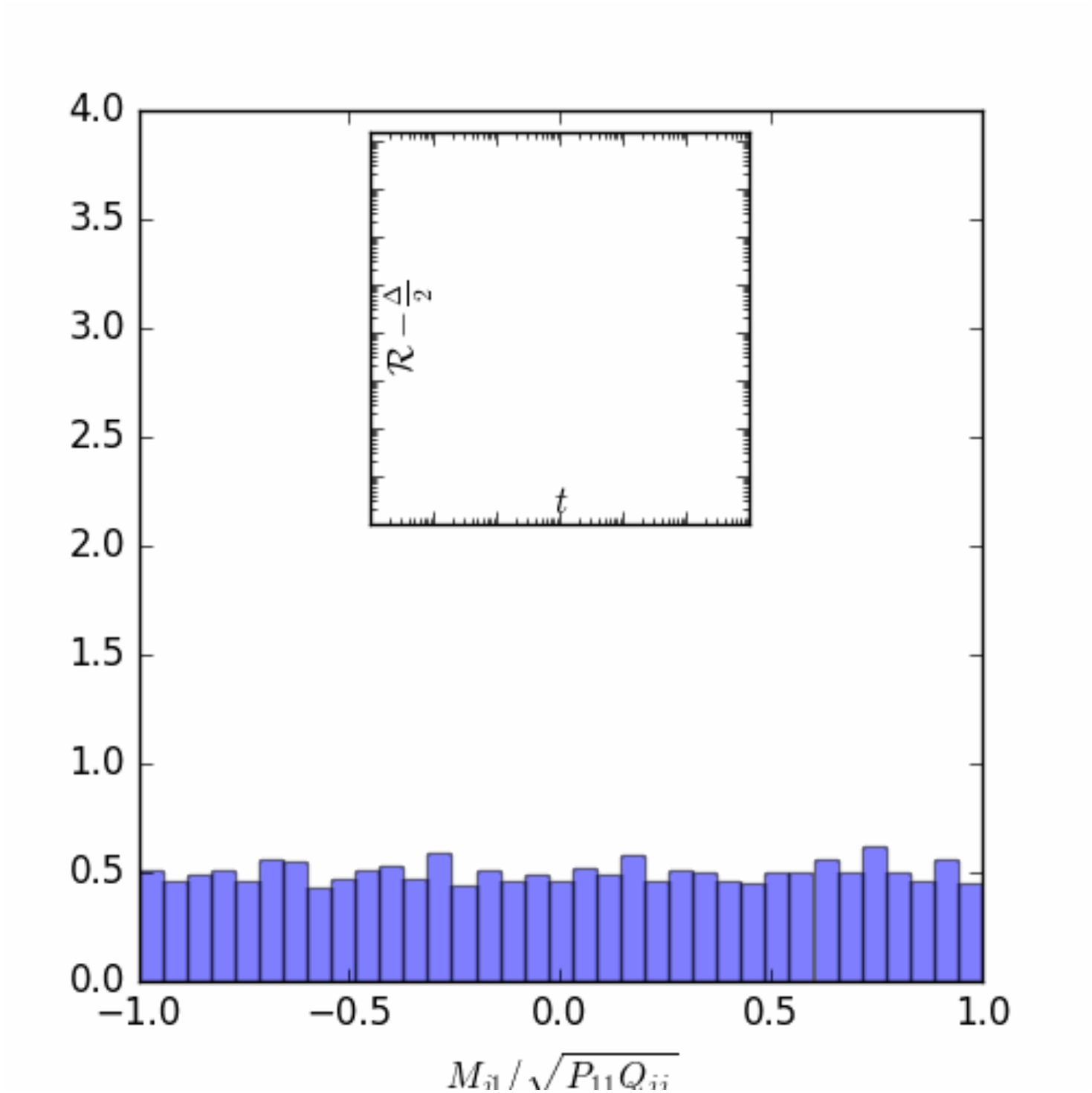
$$\mathbb{E}\,||\,Q(t) - MP^{-1}M^{\mathsf{T}} + \mathrm{diag}(Q^{\perp})\,||_{\infty} \leq e^{Ct}\left(p^{-1/2} + d^{-1/2}\right)$$

This implies MF-like PDE for the sufficient statistics:

$$\hat{\mu}_t(m,q) = \frac{1}{p}\sum_{i=1}^{p} \delta(m - m_i(t))\delta(q - Q^{\perp}_{ii}(t))$$

$$\partial_t \hat{\mu}_p(m,q) = \nabla_{(m,q)} \cdot (\hat{\mu}_t \varphi(\,\cdot\,, \hat{\mu}_t))$$

Mean-field and high-d



But what can we do with these equations?

Single-index models

"Simple case" $k = 1$ $\qquad y^\nu = \sigma_\star(w^{\star\top} x^\nu)$

- $p = 1, \sigma_\star = \sigma$: information exponent [Ben Arous, Gheissari, Jagannath '21]

$$n = O(1) \qquad \kappa = 1$$

$$n = O(d \log d) \qquad \kappa = 2$$ [Chen et al. '19; Tan, Vershynin '19]

$$n = O(d^{k-1}) \qquad \kappa \geq 2$$

See also: Joan's talk for beyond Gaussian data
Alex's on landscape smoothening

Single-index models

"Simple case" $k = 1$ $\qquad y^\nu = \sigma_\star(w^{\star\top} x^\nu)$

- $p = 1, \sigma_\star = \sigma$: information exponent [Ben Arous, Gheissari, Jagannath '21]

$$n = O(1) \qquad \kappa = 1$$

$$n = O(d \log d) \qquad \kappa = 2$$ [Chen et al. '19; Tan, Vershynin '19]

$$n = O(d^{k-1}) \qquad \kappa \geq 2$$

See also: Joan's talk for beyond Gaussian data
Alex's on landscape smoothening

- $p \to \infty$: Symmetric targets [Hajjar & Chizat '22]

Staircase functions
Leap complexity [Abbe, Adsera, Misiakiewicz '22, '23]

$\sigma_\star \neq \sigma, \kappa = 1$ [Berthier, Montanari, Zhou '23]

See also: Ludovic's and Joan's for multi-index target

Single-index models

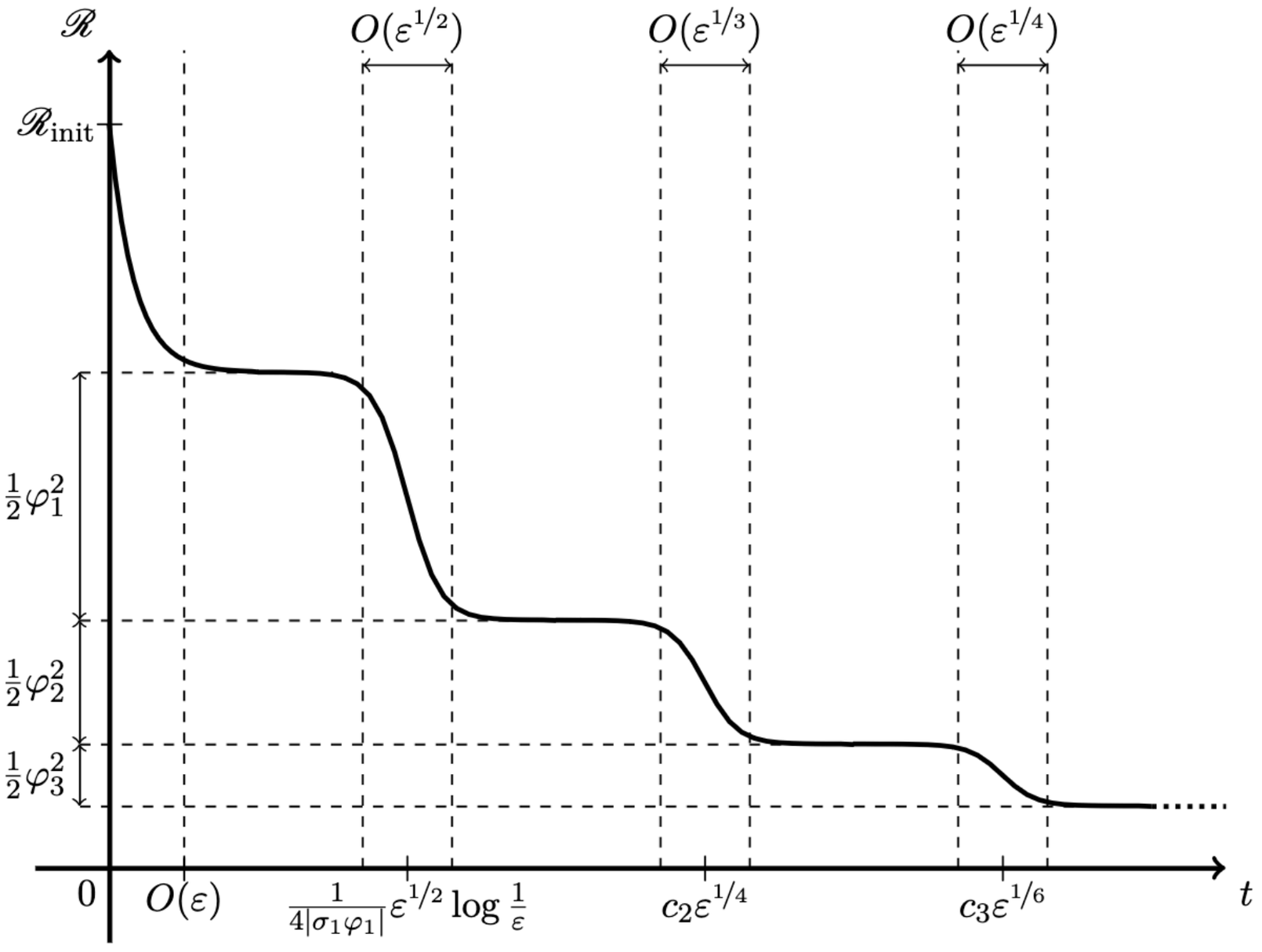
[Berthier, Montanari, Zhou '23]

$$\gamma_a = \varepsilon\gamma$$

$$y^\nu = \sigma_\star(w^{\star\top} x^\nu)$$

$$f_\Theta(x) = \int \rho(da, dw) a\sigma(w^\top x)$$

$(\sigma_\star, \sigma)$ "standard"
$(\kappa = 1)$



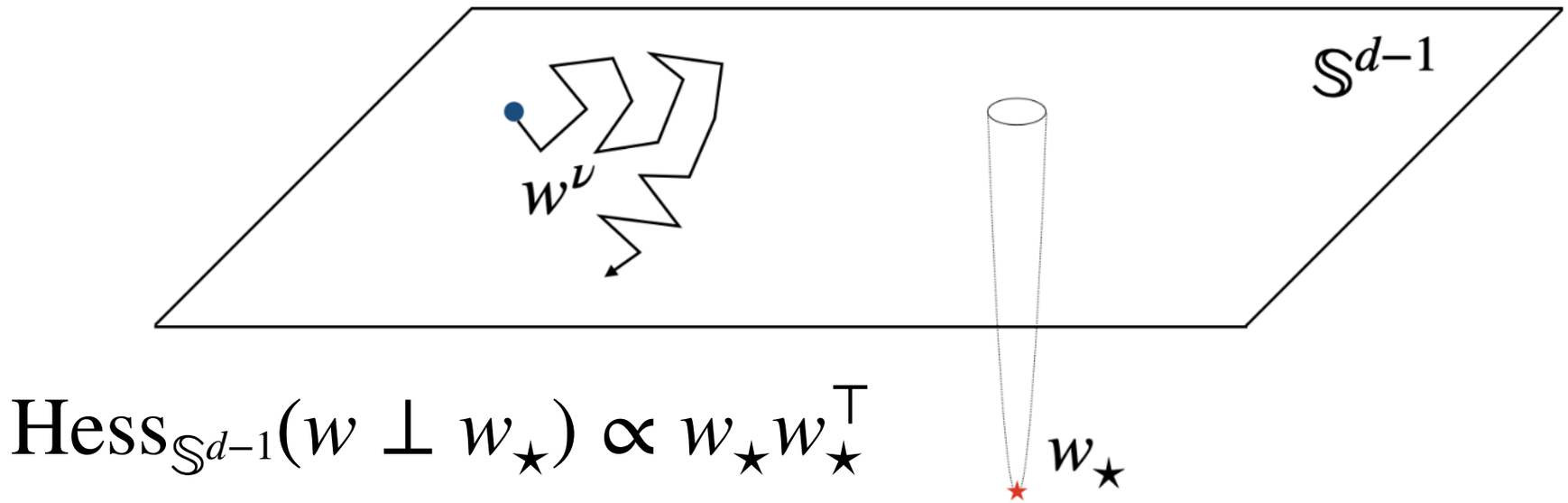
What about $\kappa > 1$?

Phase retrieval

$$y^\nu = (w^{\star\top} x^\nu)^2 \qquad f_\Theta(x) = \frac{1}{p}\sum_{i=1}^{p} (w_i^\top x^\nu)^2 \qquad w^\star, w_i \sim \mathbb{S}^{d-1}$$

Phase retrieval

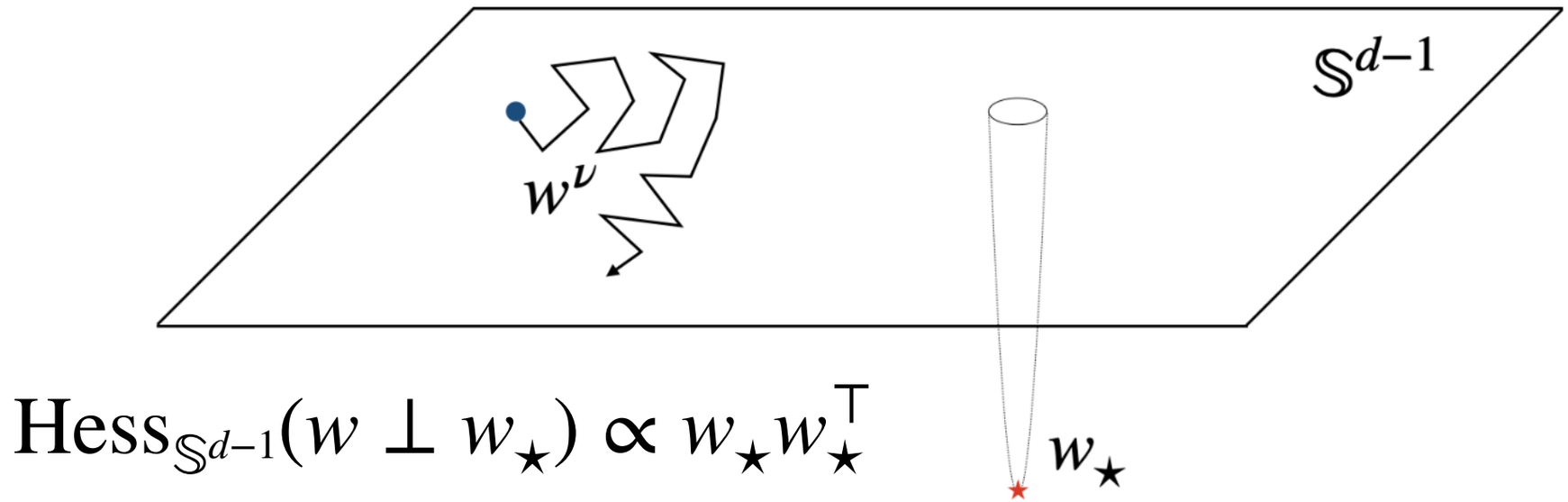
$$y^\nu = (w^{\star\top} x^\nu)^2 \qquad f_\Theta(x) = \frac{1}{p}\sum_{i=1}^{p} (w_i^\top x^\nu)^2 \qquad w^\star, w_i \sim \mathbb{S}^{d-1}$$



$$\mathrm{Hess}_{\mathbb{S}^{d-1}}(w \perp w_\star) \propto w_\star w_\star^\top$$

Phase retrieval

$$y^\nu = (w^{\star\top} x^\nu)^2 \qquad f_\Theta(x) = \frac{1}{p}\sum_{i=1}^{p} (w_i^\top x^\nu)^2 \qquad w^\star, w_i \sim \mathbb{S}^{d-1}$$



$$\mathrm{Hess}_{\mathbb{S}^{d-1}}(w \perp w_\star) \propto w_\star w_\star^\top$$

Questions:

- Know $n = O(d \log d)$, but constants? [Chen et al. '19; Tan, Vershynin '19]
- Does $p \to \infty$ helps? [Mannelli, Vanden-Eijnden, Zdeborová '20]
- Does SGD noise matters? [Ben Arous, Gheissari, Jagannath '22]

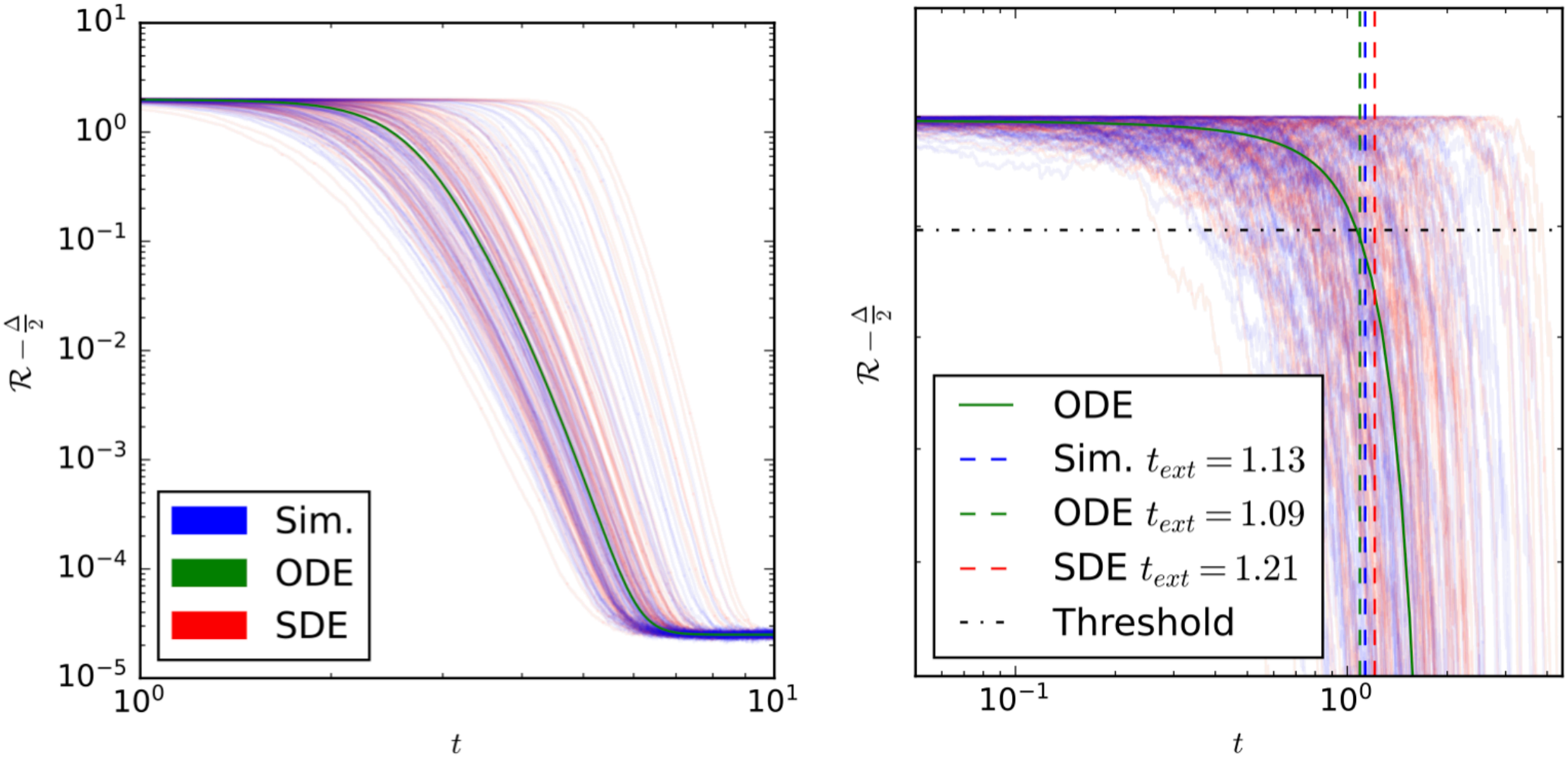
Does noise matter?

Random initial condition: $w(t=0) \sim \mathrm{Uni}(\mathbb{S}^{d-1})$ $\qquad d = 3000$

ODE: $\dot{m}(t) = \bar{\psi}(m(t))$

SDE: $\mathrm{d}m_t = \bar{\psi}(m_t)\mathrm{d}t + \Sigma_t^{1/2}\mathrm{d}B_t$

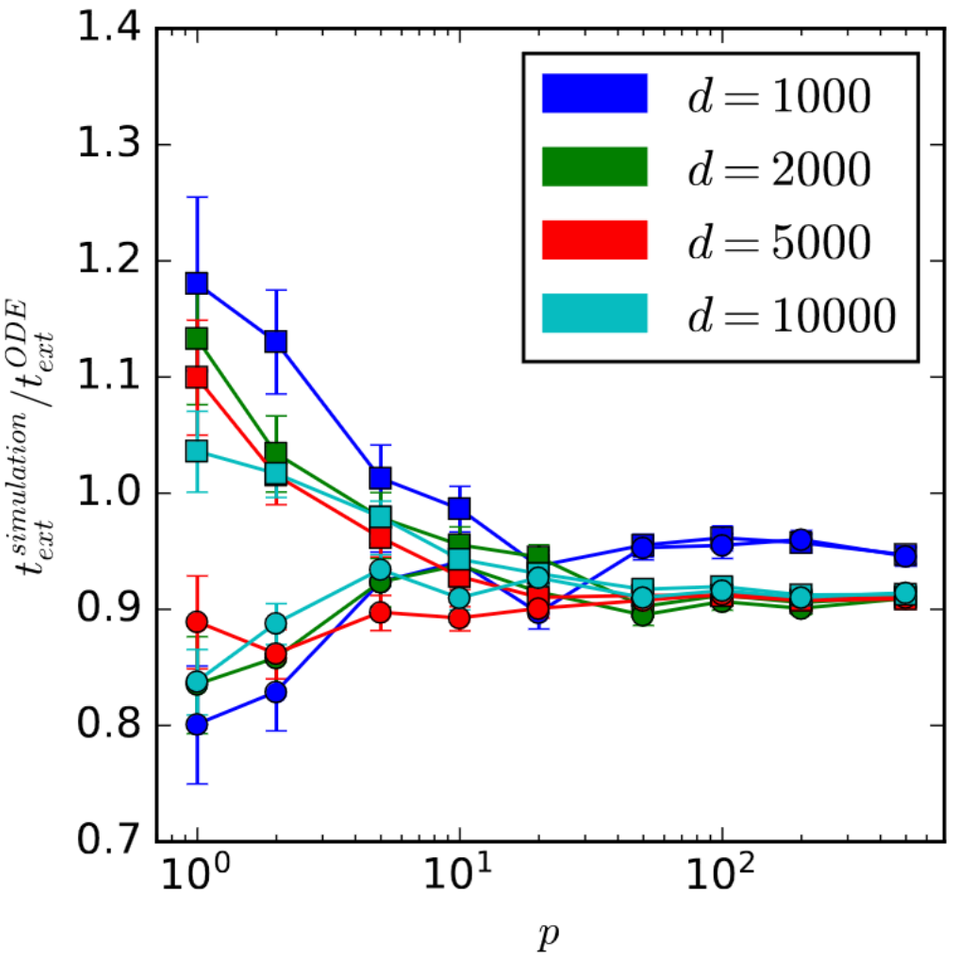
[Ben Arous, Gheissari, Jagannath '22]



[Arnaboldi, Krzakala, BL, Stephan 2023]

Does width matter?

Random initial condition: $w(t=0) \sim \mathrm{Uni}(\mathbb{S}^{d-1})$



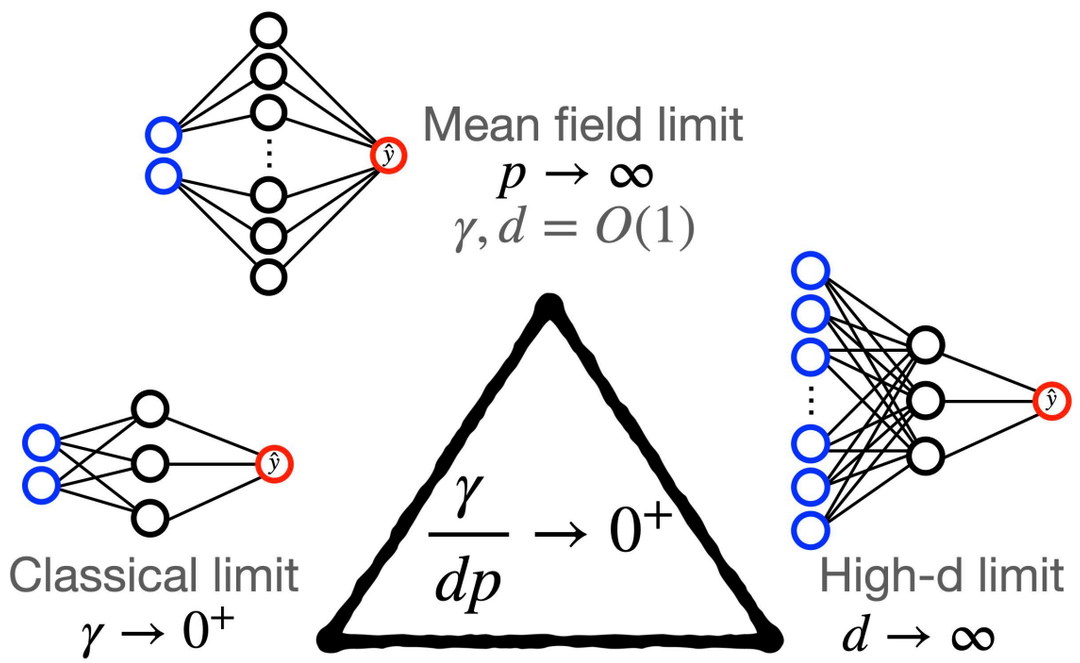
$$t_{\text{ext}}^{(\text{anl})} = \frac{\log\left[\frac{T(p+1)d + (p+1)(1-T)}{2p}\right]}{8\left[1 - \frac{\gamma}{p}\left(1 + \frac{1}{p} + \frac{4}{p^2} + \frac{\Delta}{2}\right)\right]}$$

$$t_{\text{ext}}^{(\text{qnc})} = \mathbb{E}_{\mu_0,\tau_0\sim\mathscr{P}_p^d}\left[\frac{\log\left[\frac{Tp(p+1)d + (2\mu_0 p - \tau_0)(1-T)}{2\mu_0 p}\right]}{8\left[1 - \frac{\gamma}{p}\left(1 + \frac{1}{p} + \frac{4}{p^2} + \frac{\Delta}{2}\right)\right]}\right]$$

$$\mathscr{P}_p^d \equiv \left(d\sum_{j=1}^{p}(u_j\cdot v)^2, 2d\sum_{j=1}^{p}\sum_{l=j+1}^{p}(u_j\cdot u_l)^2\right) \qquad v, u_j \sim \mathbb{S}^{d-1}$$

[Arnaboldi, Krzakala, BL, Stephan 2023]

Take aways



- Low-dimensional reductions of SGD dynamics for 2-layer NN
- Statistical Physics and mean-field limit back together (again)
- SGD noise and overparametrisation not always help you escape mediocrity