Wonders of high-dimensions:
the maths and physics of ML

Bruno Loureiro
Département d'Informatique
École Normale Supérieure & CNRS

brloureiro@gmail.com

ACDL 2023, 10-14.06.2023

Part III

Two layer neural networks in the rich regime

SGD dynamics of two-layer NNs

Recall from last lecture...

Assuming that $a_{0,i} = O(1)$ and introducing a scaling:

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

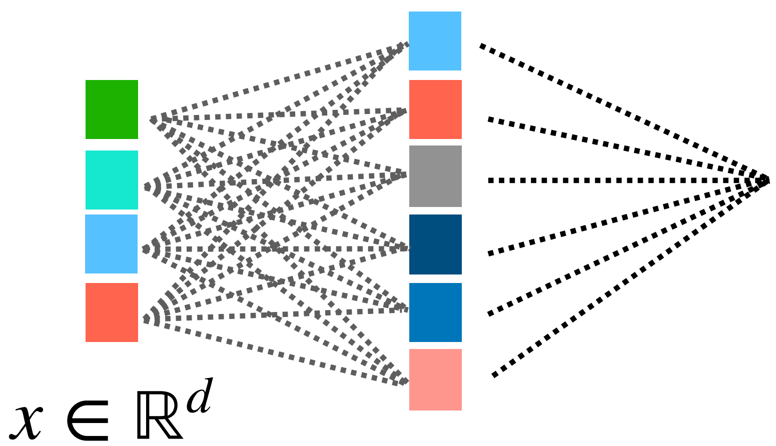
It can be shown that for $p \gg 1$: [Chizat, Oyallon & Bach '19]

$$\mathbb{E}[\kappa(\Theta_0)] \lesssim \frac{1}{\sqrt{p}} + \frac{1}{p\alpha(p)}$$

Which means $f(x;\Theta) \approx \bar{f}_{\text{lin}}(x;\Theta_0)$ if $p\alpha(p) \to \infty$ as $p \to \infty$

a.k.a. "lazy" regime

Teacher-student setup



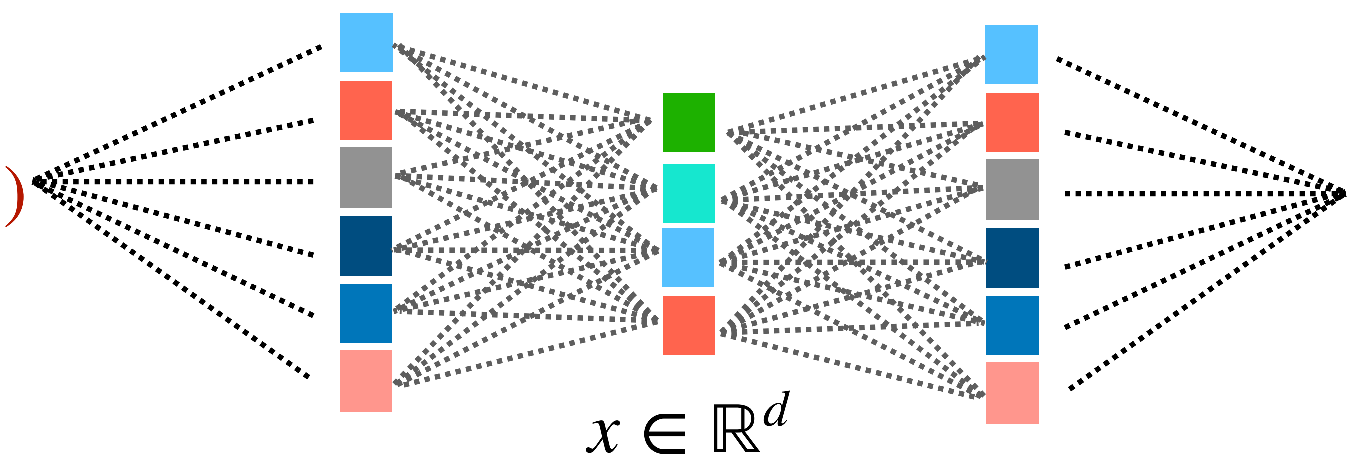
$f(x;\Theta) = a^\top \sigma(Wx)$

Hypothesis: $$f(x;\Theta) = \frac{1}{p}\sum_{i=1}^{p} a_i \sigma(w_i^\top x)$$

Teacher-student setup

Teacher

$f(x;\Theta_\star) = a_\star^\top \sigma_\star(W_\star x))$



Student

$f(x;\Theta) = a^\top \sigma(Wx)$

Hypothesis: $f(x;\Theta) = \frac{1}{p}\sum_{i=1}^{p} a_i \sigma(w_i^\top x)$

Data: $(x^\nu, y^\nu)_{\nu\in[n]} \in \mathbb{R}^d \times \mathcal{Y}$ generated as:

$$y^\nu = \frac{1}{k}\sum_{r=1}^{k} a_{\star,r}\sigma(w_\star^\top x^\nu) + \sqrt{\Delta} z^\nu$$

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

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

Algorithm: SGD

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

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

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

mini-batch

Algorithm: SGD

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

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

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

mini-batch

Gradient descent (GD)

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

Algorithm: SGD

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

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

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

mini-batch

Gradient descent (GD)

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

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

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

Algorithm: SGD

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

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

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

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.

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

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

mini-batch

Gradient descent (GD)	One-pass SGD
$b_k = [n], \quad \forall k$	b_k independent
$\gamma \to 0^+$ at fixed d, p:	$\gamma \to 0^+$ at fixed d, p:
$\dot{\Theta}(t) = -\nabla_\Theta \hat{\mathcal{R}}_n\left(\Theta(t)\right)$	$\dot{\Theta}(t) = -\nabla_\Theta \mathcal{R}\left(\Theta(t)\right)$

Algorithm: SGD

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

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

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

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:

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

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?

Summary of setting

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

Architecture:

$$f(x;\Theta) = \frac{1}{p}\sum_{i=1}^{p} a_i \sigma(w_i \cdot x)$$

Data model:

$$y^\nu = \frac{1}{k}\sum_{r=1}^{k} a_r^\star \sigma(w_r^\star \cdot x^\nu) + \sqrt{\Delta} z^\nu$$

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

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

Algorithm:

$$\Theta^{\nu+1} = \Theta^\nu - \gamma_\nu \nabla_{\Theta^\nu} (y^\nu - f(x^\nu;\Theta^\nu))^2$$

Sufficient statistics

Goal: track population error exactly throughout the dynamics

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

Sufficient statistics

Goal: track population error exactly throughout the dynamics

$$\mathcal{R}(\Theta^\nu) = \frac{1}{2}\mathbb{E}_{(\lambda_\star{}^\nu, \lambda^\nu)\sim\mathcal{N}(0,\Omega^\nu)}\left[\left(\frac{1}{k}\sum_{r=1}^{k} a_{\star,r}\sigma(\lambda^\nu_{\star,r}) - \frac{1}{p}\sum_{i=1}^{p} a_i^\nu \sigma(\lambda_i^\nu)\right)^2\right] + \frac{\Delta}{2}$$

Where:

$$\Omega^\nu = \frac{1}{d}\begin{pmatrix} W_\star W_\star^\top & W_\star W^{\nu\top} \\ W^\nu W_\star^\top & W^\nu W^{\nu\top} \end{pmatrix} = \begin{pmatrix} P & M^\nu \\ M^{\nu\top} & Q^\nu \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)}$$

Sufficient statistics

Goal: track population error exactly throughout the dynamics

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

Where:

$$\Omega^\nu = \frac{1}{d}\begin{pmatrix} W_\star W_\star^\top & W_\star W^{\nu\top} \\ W^\nu W_\star^\top & W^\nu W^{\nu\top} \end{pmatrix} = \begin{pmatrix} P & M^\nu \\ M^{\nu\top} & Q^\nu \end{pmatrix} \in \mathbb{R}^{(k+p)\times(k+p)}$$

Key idea:

One-pass SGD $\longrightarrow$ $\Omega^{\nu+1} = \psi(\Omega^\nu)$

Tracking overlaps

Starting point: one-pass SGD

$$\Theta^{\nu+1} = \Theta^{\nu} - \frac{\gamma_\nu}{2} \nabla_{\Theta^\nu} (y^\nu - f(x^\nu; \Theta^\nu))^2$$

Tracking overlaps

Starting point: one-pass SGD

$$\Theta^{\nu+1} = \Theta^{\nu} - \frac{\gamma_\nu}{2} \nabla_{\Theta^\nu} (y^\nu - f(x^\nu; \Theta^\nu))^2$$
$$= \Theta^{\nu} + \gamma_\nu (y^\nu - f(x^\nu; \Theta^\nu)) \nabla_{\Theta^\nu} f(x^\nu; \Theta^\nu)$$

Tracking overlaps

Starting point: one-pass SGD

$$\Theta^{\nu+1} = \Theta^{\nu} - \frac{\gamma_\nu}{2} \nabla_{\Theta^\nu} (y^\nu - f(x^\nu; \Theta^\nu))^2$$

$$= \Theta^{\nu} + \gamma_\nu (y^\nu - f(x^\nu; \Theta^\nu)) \nabla_{\Theta^\nu} f(x^\nu; \Theta^\nu)$$

$$= \Theta^{\nu} + \gamma_\nu \mathscr{E}^\nu \nabla_{\Theta^\nu} f(x^\nu; \Theta^\nu)$$

Tracking overlaps

Starting point: one-pass SGD

$$\Theta^{\nu+1} = \Theta^{\nu} - \frac{\gamma_\nu}{2} \nabla_{\Theta^\nu} (y^\nu - f(x^\nu; \Theta^\nu))^2$$
$$= \Theta^\nu + \gamma_\nu (y^\nu - f(x^\nu; \Theta^\nu)) \nabla_{\Theta^\nu} f(x^\nu; \Theta^\nu)$$
$$= \Theta^\nu + \gamma_\nu \mathscr{E}^\nu \nabla_{\Theta^\nu} f(x^\nu; \Theta^\nu)$$

But:

$$\nabla_{a_i} f(x; \Theta) = \frac{1}{p} \sigma(w_i^\top x)$$

$$\nabla_{w_i} f(x; \Theta) = \frac{1}{p} a_i \sigma'(w_i^\top x) x$$

Tracking overlaps

Starting point: one-pass SGD

$$a_i^{\nu+1} = a_i^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}^\nu \sigma({w_i^\nu}^\top x^\nu)$$

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}^\nu a_i \sigma'({w_i^\nu}^\top x^\nu) x^\nu$$

Tracking overlaps

Starting point: one-pass SGD

$$a_i^{\nu+1} = a_i^{\nu} + \frac{\gamma_\nu}{p} \mathscr{E}^{\nu} \sigma({w_i^{\nu}}^\top x^{\nu})$$

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p} \mathscr{E}^{\nu} a_i \sigma'({w_i^{\nu}}^\top x^{\nu}) x^{\nu}$$

Goal: Go from this to equation for Ω^{ν}

Tracking overlaps

Starting point: one-pass SGD

$$a_i^{\nu+1} = a_i^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}^{\nu} \sigma(w_i^{\nu\top} x^{\nu})$$

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}^{\nu} a_i \sigma'(w_i^{\nu\top} x^{\nu}) x^{\nu}$$

Goal: Go from this to equation for Ω^{ν}

$$\begin{aligned}\mathcal{E}^{\nu} &= (y^{\nu} - f(x^{\nu}; \Theta^{\nu})) \\ &= \left(\frac{1}{k} \sum_{r=1}^{k} a_{\star,r} \sigma_{\star}(w_{\star,r}^{\top} x^{\nu}) - \sqrt{\Delta} z^{\nu} - \frac{1}{p} \sum_{i=1}^{p} a_i^{\nu} \sigma(w_i^{\nu\top} x^{\nu}) \right)\end{aligned}$$

Tracking overlaps

Starting point: one-pass SGD

$$a_i^{\nu+1} = a_i^\nu + \frac{\gamma_\nu}{p} \mathcal{E}^\nu \sigma(\lambda_i^\nu)$$

$$w_i^{\nu+1} = w_i^\nu + \frac{\gamma_\nu}{p} \mathcal{E}^\nu a_i \sigma'(\lambda_i^\nu) x^\nu$$

Goal: Go from this to equation for Ω^ν

$$\begin{aligned}\mathcal{E}^\nu &= (y^\nu - f(x^\nu;\Theta^\nu)) \\ &= \left(\frac{1}{k} \sum_{r=1}^{k} a_{\star,r} \sigma_\star(\lambda_{\star,r}) - \sqrt{\Delta} z^\nu - \frac{1}{p} \sum_{i=1}^{p} a_i^\nu \sigma(\lambda_i^\nu) \right)\end{aligned}$$

Tracking overlaps

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p}\mathscr{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu}$$

Equation for $M^{\nu} = \frac{1}{d} W_{\star} W^{\nu\top}$:

Tracking overlaps

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p}\mathcal{E}^\nu a_i \sigma'(\lambda_i^\nu) x^\nu$$

Equation for $M^\nu = \frac{1}{d} W_\star W^{\nu\top}$:

$$w_{\star,r}^\top w_i^{\nu+1} = w_{\star,r}^\top w_i^{\nu} + \frac{\gamma_\nu}{p}\mathcal{E}^\nu a_i \sigma'(\lambda_i^\nu) w_{\star,r}^\top x^\nu$$

Tracking overlaps

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p}\mathcal{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu}$$

Equation for $M^{\nu} = \frac{1}{d} W_\star W^{\nu\top}$:

$$\frac{1}{d} w_{\star,r}^{\top} w_i^{\nu+1} = \frac{1}{d} w_{\star,r}^{\top} w_i^{\nu} + \frac{\gamma_\nu}{dp}\mathcal{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) w_{\star,r}^{\top} x^{\nu}$$

Tracking overlaps

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p} \mathcal{E}^\nu a_i \sigma'(\lambda_i^\nu) x^\nu$$

Equation for $M^\nu = \frac{1}{d} W_\star W^{\nu \top}$:

$$M_{ri}^{\nu+1} = M_{ri}^{\nu} + \frac{\gamma_\nu}{dp} \mathcal{E}^\nu a_i \sigma'(\lambda_i^\nu) \lambda_{\star,r}^\nu$$

Tracking overlaps

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p}\mathcal{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu}$$

Equation for $M^{\nu} = \frac{1}{d} W_\star W^{\nu \top}$:

$$M_{ri}^{\nu+1} = M_{ri}^{\nu} + \frac{\gamma_\nu}{dp}\mathcal{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) \lambda_{\star,r}^{\nu}$$

Equation for $Q^{\nu} = \frac{1}{d} W^{\nu} W^{\nu \top}$:

$$w_j^{\nu+1\,\top} w_i^{\nu+1} = w_j^{\nu+1\,\top} \left(w_i^{\nu} + \frac{\gamma_\nu}{p}\mathcal{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu} \right)$$

Tracking overlaps

$$w_i^{\nu+1} = w_i^{\nu} + \frac{\gamma_\nu}{p}\mathscr{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu}$$

Equation for $M^{\nu} = \frac{1}{d} W_{\star} W^{\nu\top}$:

$$M_{ri}^{\nu+1} = M_{ri}^{\nu} + \frac{\gamma_\nu}{dp}\mathscr{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) \lambda_{\star,r}^{\nu}$$

Equation for $Q^{\nu} = \frac{1}{d} W^{\nu} W^{\nu\top}$:

$$w_j^{\nu+1\top} w_i^{\nu+1} = w_j^{\nu+1\top}\left(w_i^{\nu} + \frac{\gamma_\nu}{p}\mathscr{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu}\right)$$

$$= \left(w_i^{\nu} + \frac{\gamma_\nu}{p}\mathscr{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu}\right)\left(w_i^{\nu} + \frac{\gamma_\nu}{p}\mathscr{E}^{\nu} a_i \sigma'(\lambda_i^{\nu}) x^{\nu}\right)$$

Tracking overlaps

$$w_i^{\nu+1} = w_i^\nu + \frac{\gamma_\nu}{p} \mathcal{E}^\nu a_i \sigma'(\lambda_i^\nu) x^\nu$$

Equation for $M^\nu = \frac{1}{d} W_\star W^{\nu \top}$:

$$M_{ri}^{\nu+1} = M_{ri}^\nu + \frac{\gamma_\nu}{dp} \mathcal{E}^\nu a_i \sigma'(\lambda_i^\nu) \lambda_{\star,r}^\nu$$

Equation for $Q^\nu = \frac{1}{d} W^\nu W^{\nu \top}$:

$$Q_{ji}^{\nu+1} = Q_{ji}^\nu + \frac{\gamma_\nu}{dp} \left(\mathcal{E}^\nu \sigma'(\lambda_j^\nu) \lambda_i^\nu + \mathcal{E}^\nu \sigma'(\lambda_i^\nu) \lambda_j^\nu \right)$$
$$+ \frac{\gamma_\nu^2}{dp^2} (\mathcal{E}^\nu)^2 \sigma'(\lambda_i^\nu) \sigma'(\lambda_j^\nu) \, || x^\nu ||_2^2$$

Tracking overlaps: summary

$$w_i^{\nu+1} = w_i^\nu + \frac{\gamma_\nu}{p} \mathscr{E}^\nu a_i \sigma'(\lambda_i^\nu) x^\nu$$

Stochastic process in $\mathbb{R}^{p \times d}$

$$M_{ri}^{\nu+1} - M_{ri}^\nu = \frac{\gamma_\nu}{dp} \Psi_M^{(\mathrm{GF})}(M, Q)$$

$$Q_{ji}^{\nu+1} - Q_{ji}^\nu = \frac{\gamma_\nu}{dp} \Psi_Q^{(\mathrm{GF})}(M, Q) + \frac{\gamma_\nu^2}{dp^2} \Psi_Q^{(\mathrm{var})}(M, Q)$$

Stochastic process in $\mathbb{R}^{p(p+k)}$

Concentration result

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

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

Concentration result

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

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

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

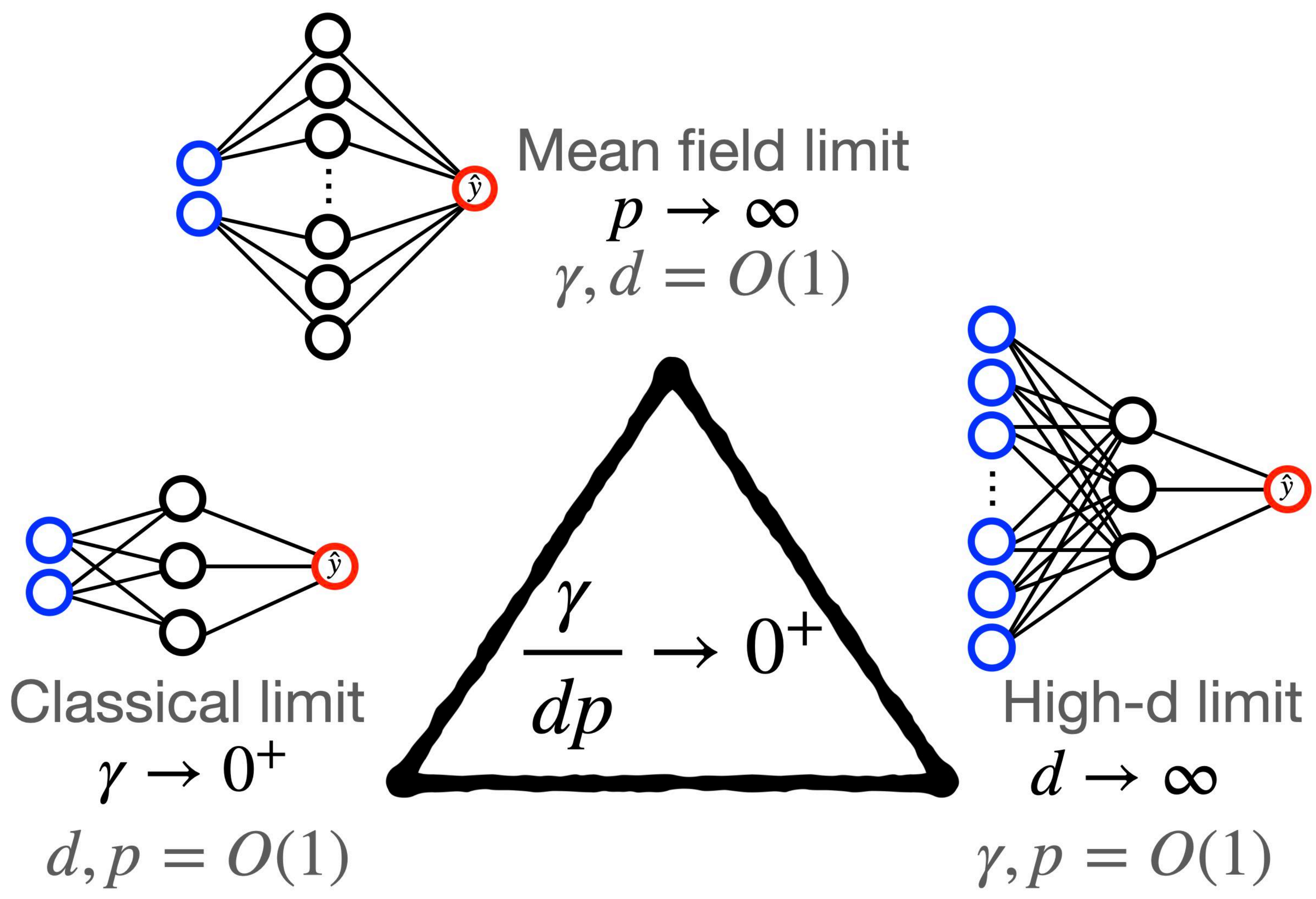
Then $\forall 0 \leq \nu \leq \left\lfloor \frac{n}{\delta t} \right\rfloor$:

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

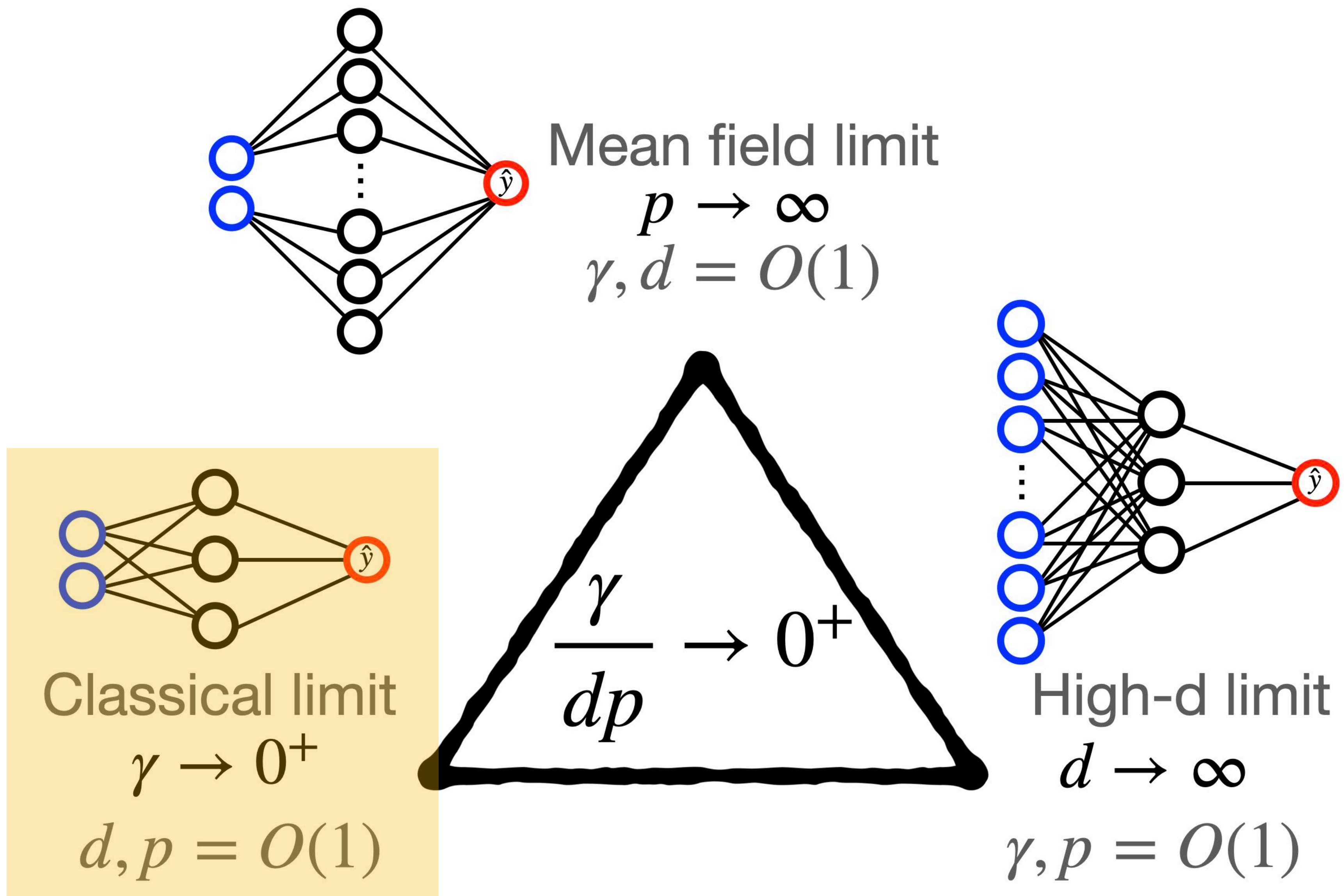
Where $\bar{\Omega}(t) = \mathbb{E}[\Omega(t)]$ is the solution of an ODE:

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

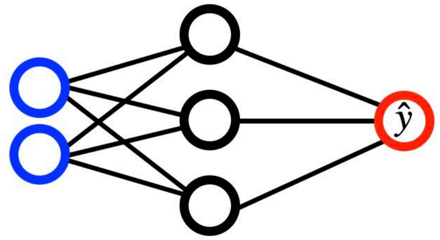
The different limiting regimes



The different limiting regimes



Classical regime



Classical limit

$\gamma \to 0^+$

$d, p = O(1)$

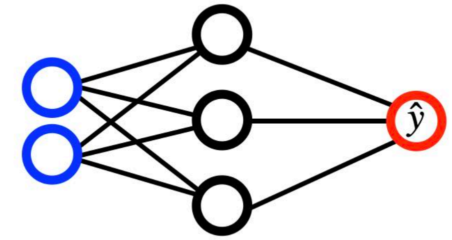
$$\Theta^{\nu+1} = \Theta^{\nu} - \frac{\gamma_\nu}{2} \nabla_{\Theta^\nu} (y^\nu - f(x^\nu; \Theta^\nu))^2$$

$$\downarrow \quad \gamma \to 0^+$$

[Robins & Monro '51]

$$\dot{\Theta}(t) = -\frac{1}{2} \nabla_\Theta \mathbb{E}_{(x,y)} \left[(y - f(x; \Theta(t)))^2 \right]$$

Classical regime



Classical limit

$\gamma \to 0^+$

$d, p = O(1)$

$$\Theta^{\nu+1} = \Theta^{\nu} - \frac{\gamma_\nu}{2} \nabla_{\Theta^\nu} (y^\nu - f(x^\nu; \Theta^\nu))^2$$

$\gamma \to 0^+$

[Robins & Monro '51]

$$\dot{\Theta}(t) = -\frac{1}{2} \nabla_\Theta \mathbb{E}_{(x,y)} \left[(y - f(x; \Theta(t)))^2 \right]$$

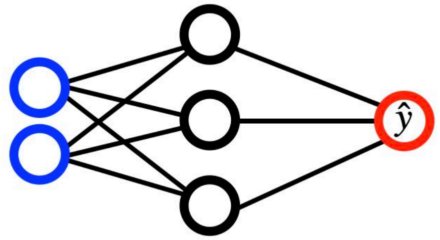
Subleading in γ

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

GD on population

Effective Noise

Classical regime



Classical limit

$\gamma \to 0^+$

$d, p = O(1)$

$$\Theta^{\nu+1} = \Theta^{\nu} - \frac{\gamma_\nu}{2} \nabla_{\Theta^\nu} (y^\nu - f(x^\nu; \Theta^\nu))^2$$

$\gamma \to 0^+$

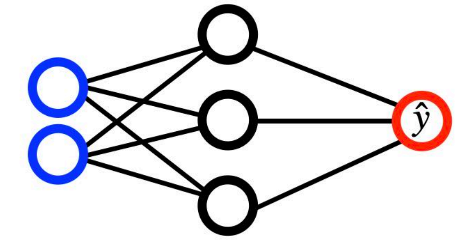
[Robins & Monro '51]

$$\dot{\Theta}(t) = -\frac{1}{2} \nabla_{\Theta} \mathbb{E}_{(x,y)} \left[(y - f(x; \Theta(t)))^2 \right]$$

$$\dot{M}_{ri}(t) = \mathbb{E}[\Psi_M^{(\mathrm{GF})}(\bar{M}, \bar{Q})]$$

$$\dot{\bar{Q}}_{ji}(t) = \mathbb{E}[\Psi_Q^{(\mathrm{GF})}(\bar{M}, \bar{Q})]$$

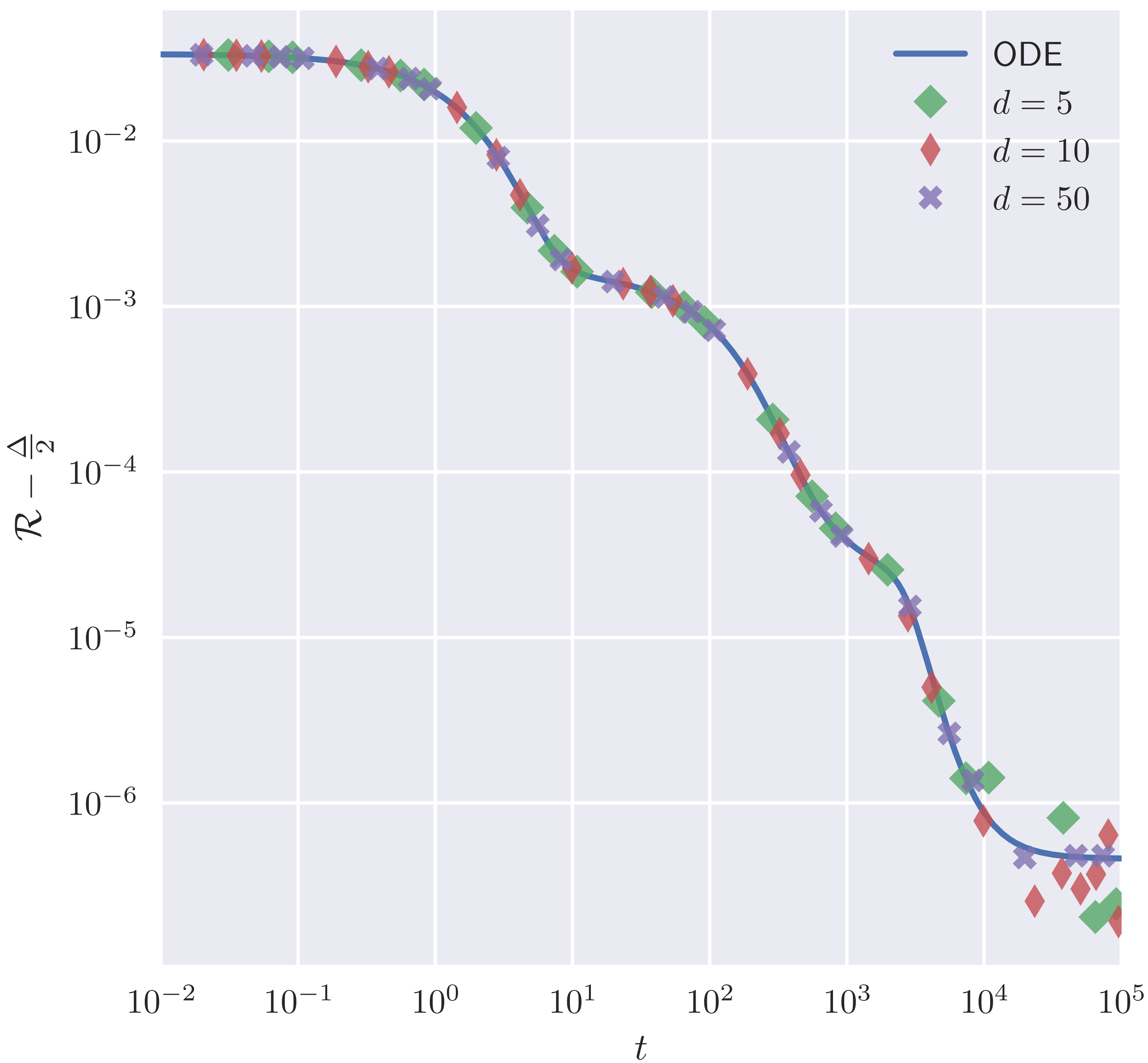
Classical regime



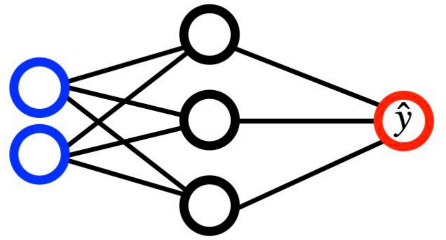
Classical limit

$\gamma \to 0^+$

$d, p = O(1)$



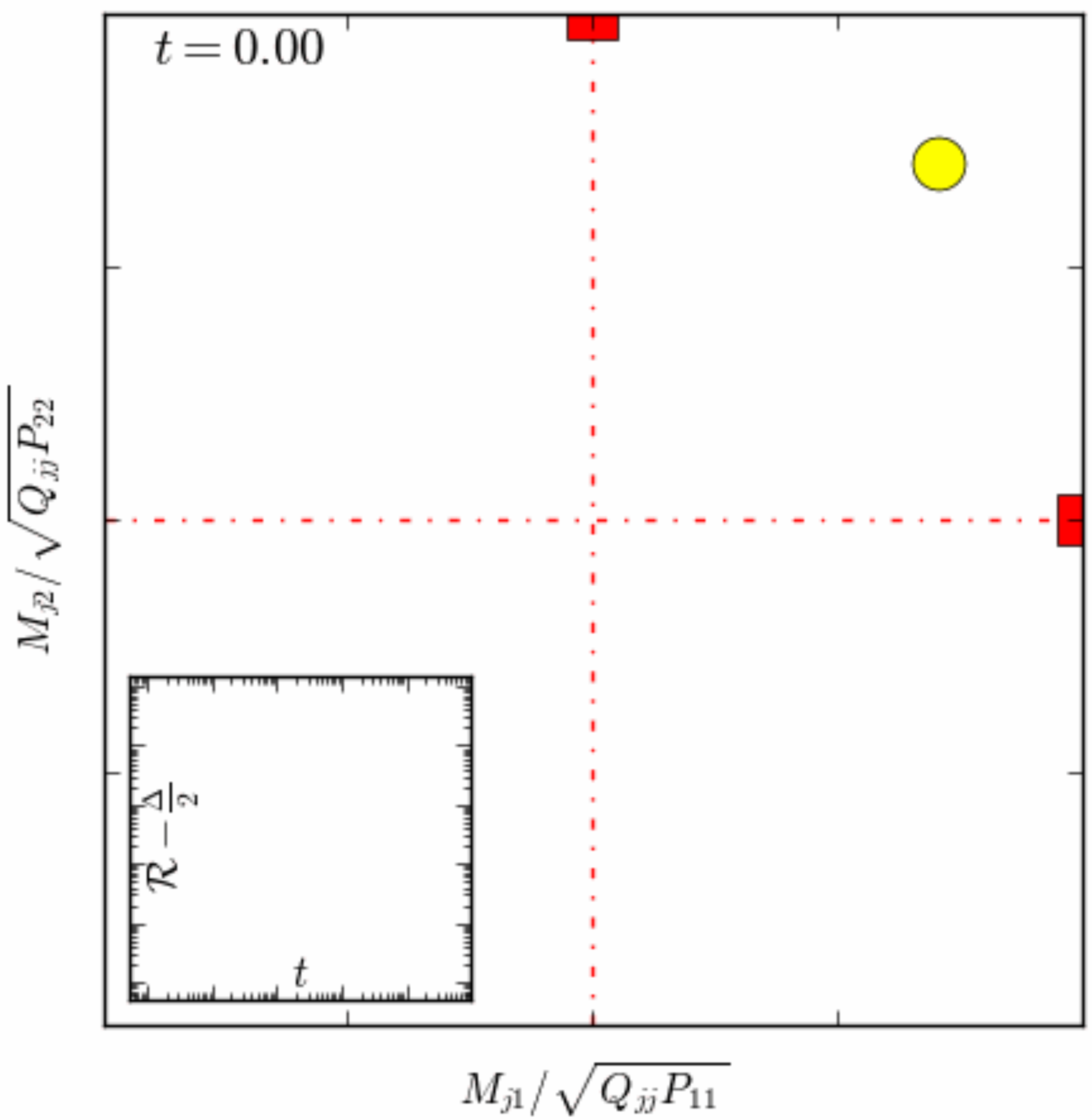
Classical regime



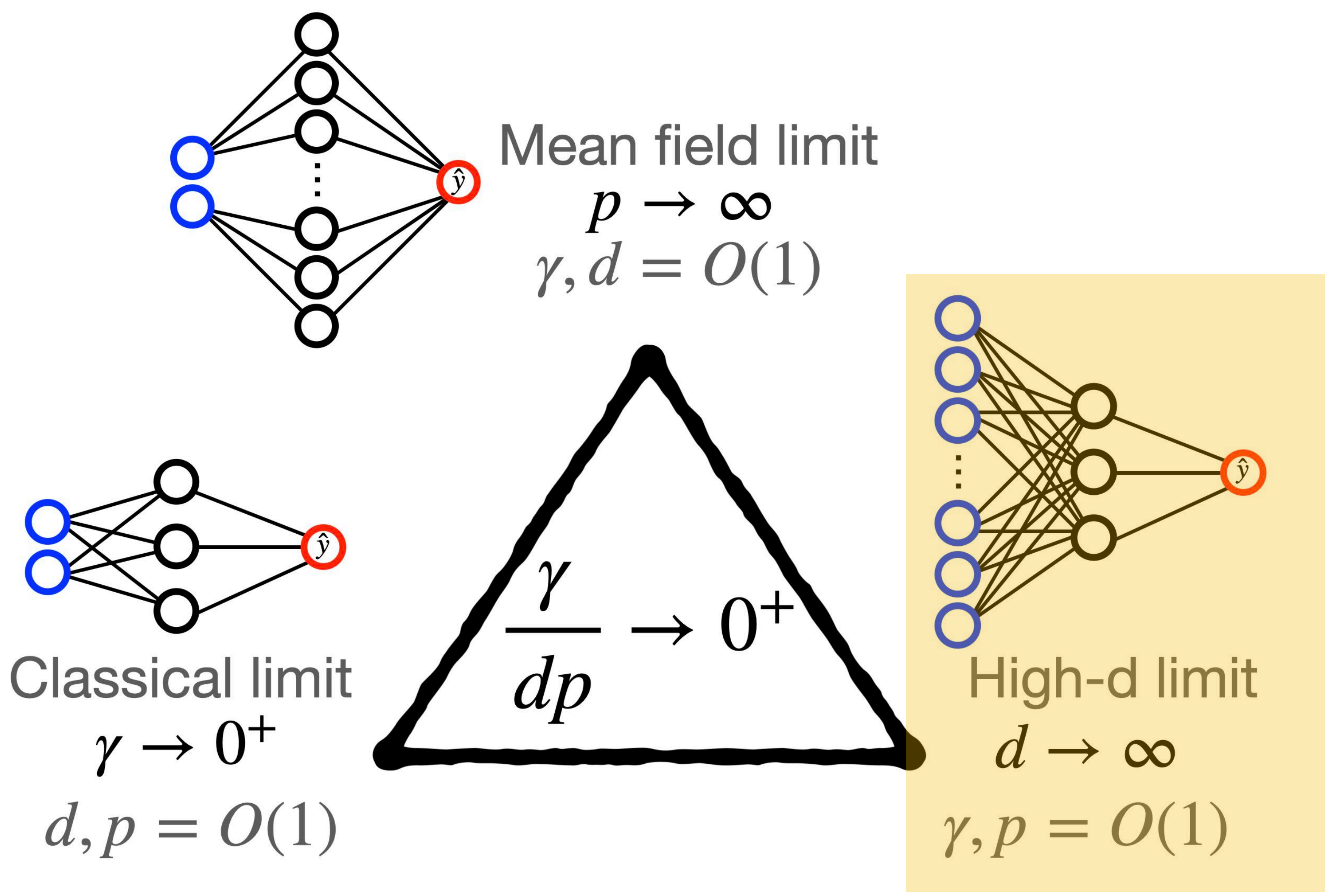
Classical limit

$\gamma \to 0^+$

$d, p = O(1)$

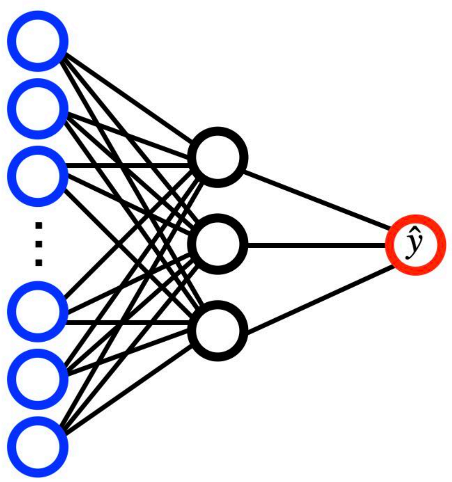


The different limiting regimes



High-dimensional regime

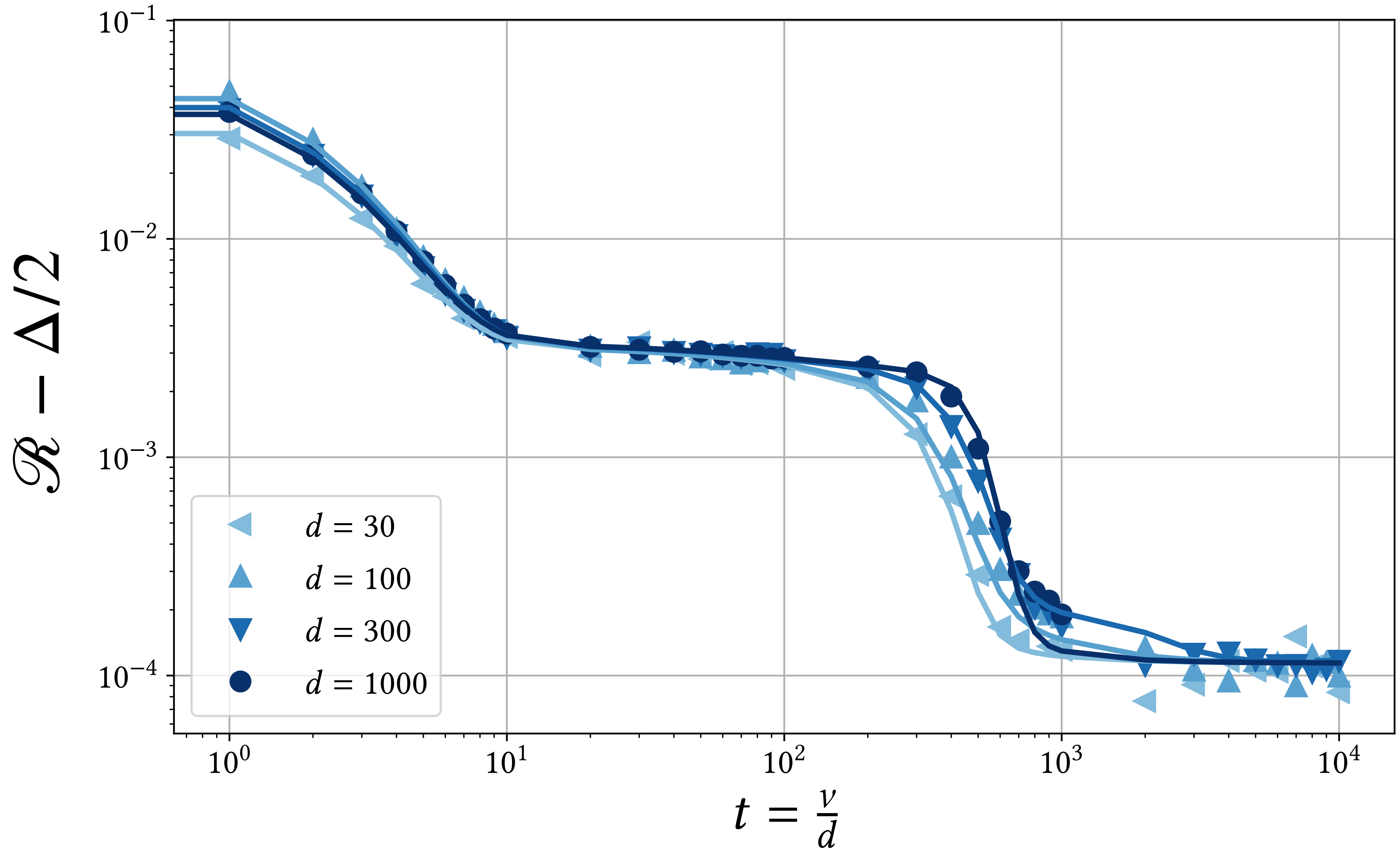
[Saad & Solla '95]



High-d limit

$d \to \infty$

$\gamma, p = O(1)$

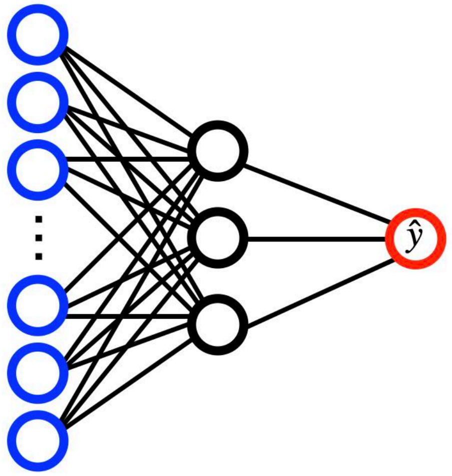


$$\dot{\bar{M}}_{ri}(t) = \mathbb{E}[\Psi_M^{(\mathrm{GF})}(\bar{M}, \bar{Q})]$$

$$\dot{\bar{Q}}_{ji}(t) = \mathbb{E}[\Psi_Q^{(\mathrm{GF})}(\bar{M}, \bar{Q})] + \frac{\gamma}{p}\mathbb{E}\left[\Psi_Q^{\mathrm{var}}(\bar{M}, \bar{Q})\right]$$

High-dimensional regime

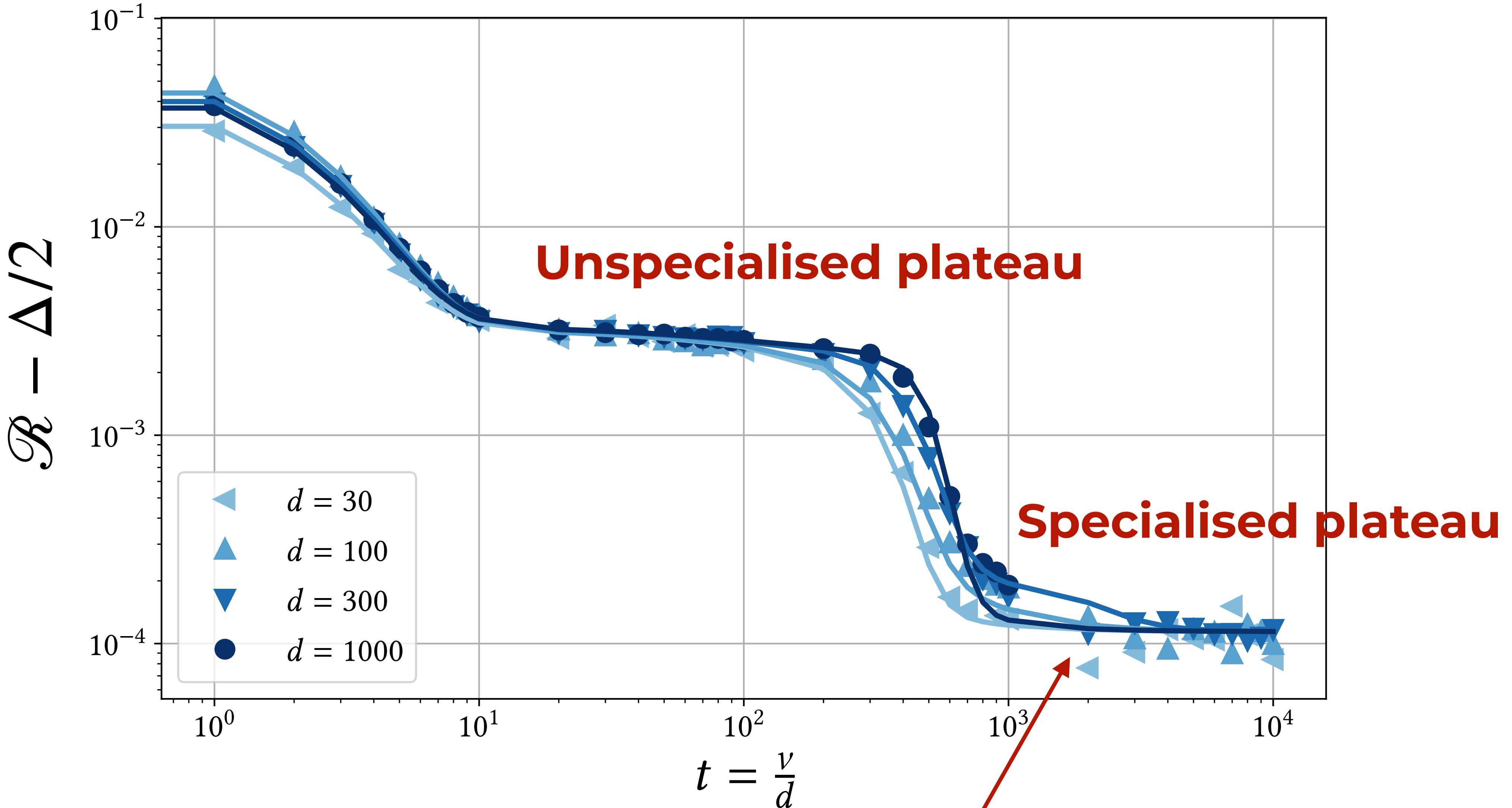
[Saad & Solla '95]



High-d limit

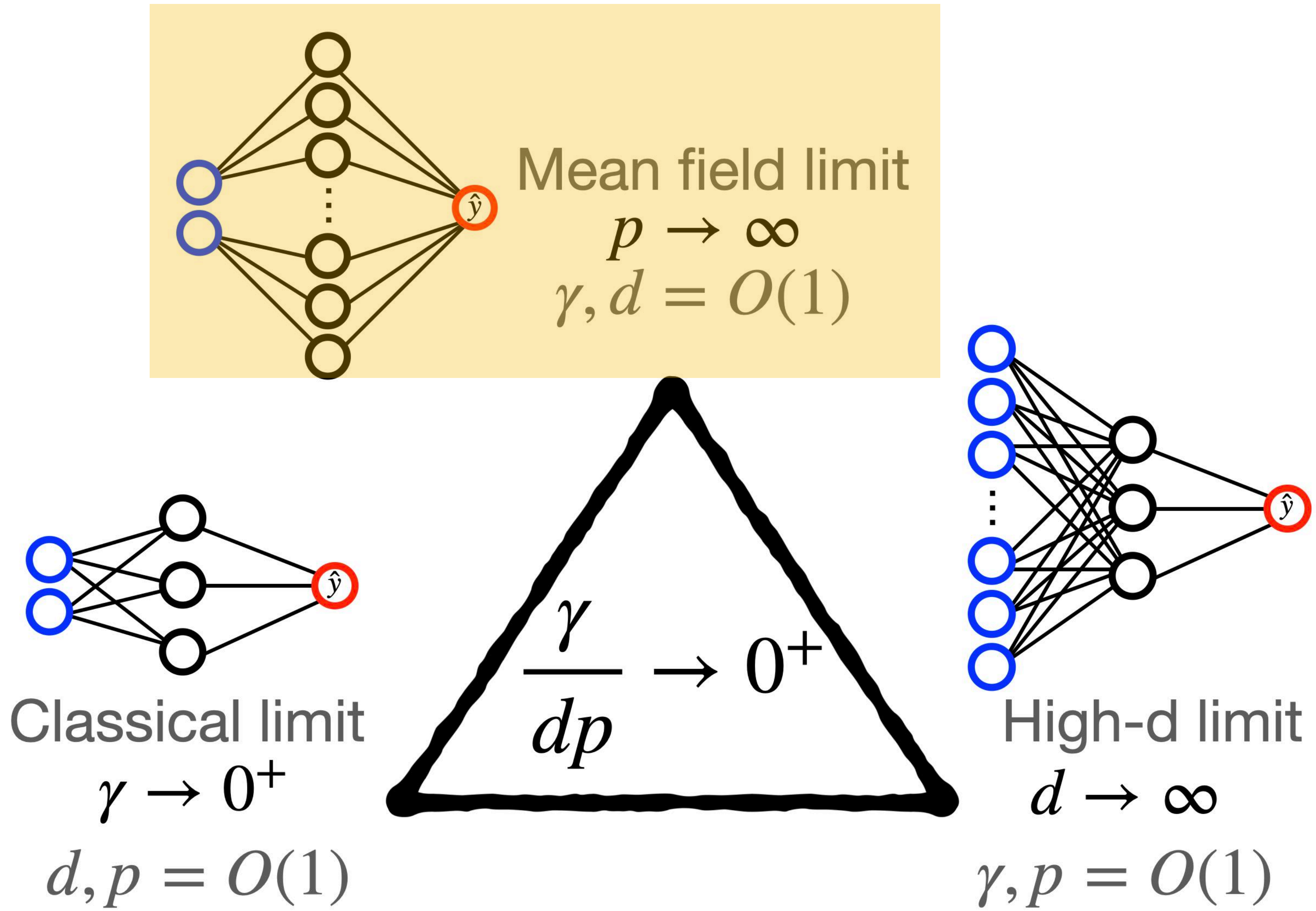
$d \to \infty$

$\gamma, p = O(1)$

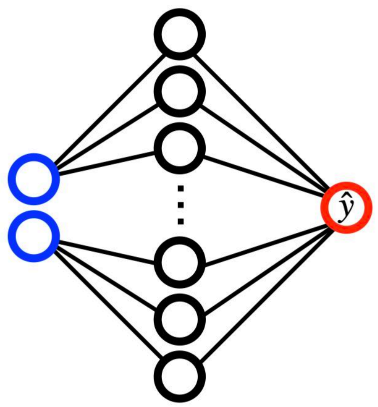


$\mathcal{R}_\infty - \Delta/2 \propto \gamma\Delta$

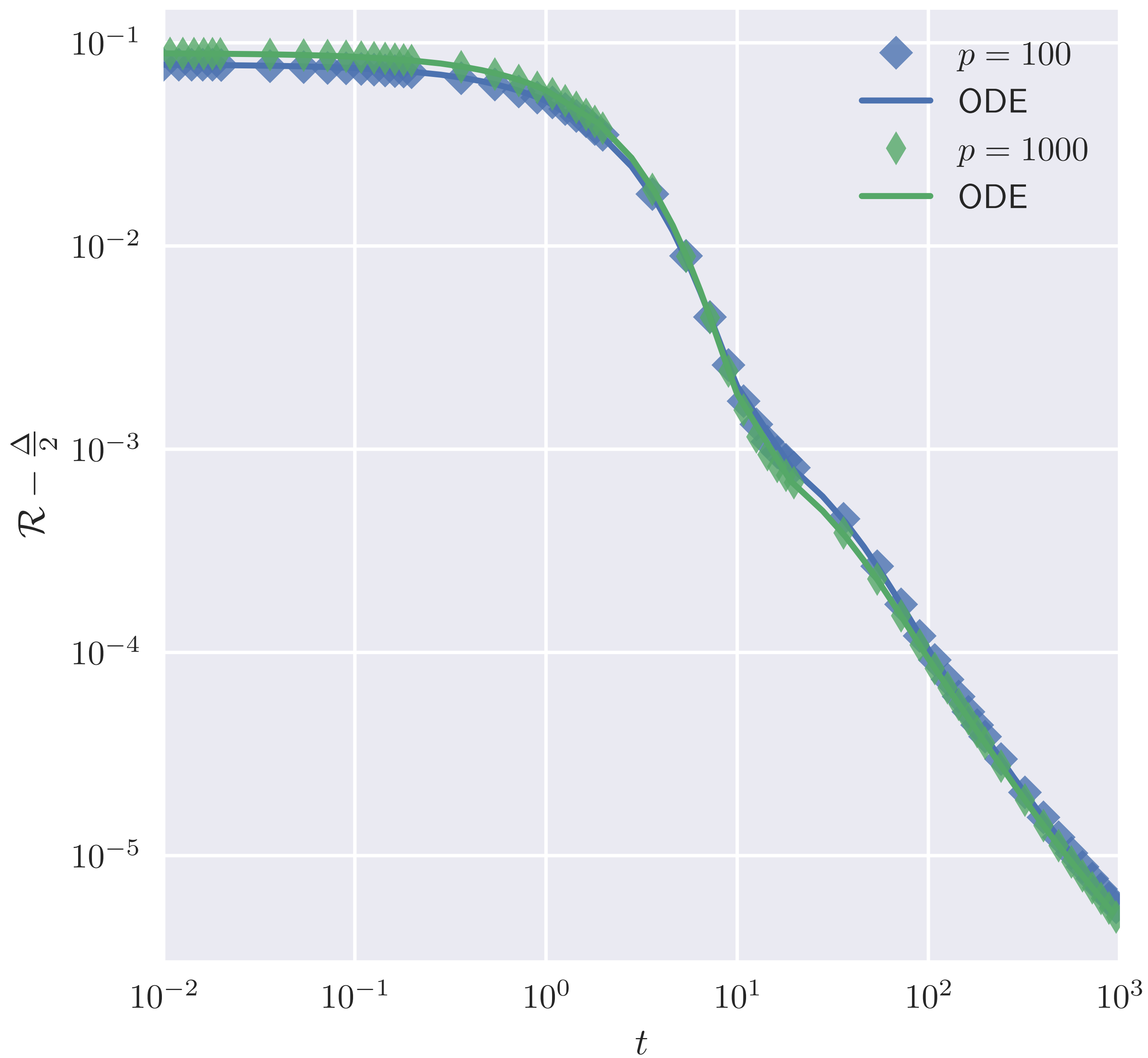
The different limiting regimes



Mean-field regime



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



$$\dot{M}_{ri}(t) = \mathbb{E}[\Psi_M^{(GF)}(\bar{M}, \bar{Q})] \qquad \dot{Q}_{ji}(t) = \mathbb{E}[\Psi_Q^{(GF)}(\bar{M}, \bar{Q})]$$

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:

$$\rho_p^{\nu}(a, w) = \frac{1}{p} \sum_{i=1}^{p} \delta(a_i - a_i^{\nu}) \delta(w_i - w_i^{\nu})$$

Mean-field limit

Idea: Define empirical density of weights:

$$\rho_p^{\nu}(a, w) = \frac{1}{p} \sum_{i=1}^{p} \delta(a_i - a_i^{\nu}) \delta(w_i - w_i^{\nu})$$

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

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

Mean-field limit

Idea: Define empirical density of weights:

$$\rho_p^\nu(a, w) = \frac{1}{p} \sum_{i=1}^{p} \delta(a_i - a_i^\nu)\delta(w_i - w_i^\nu)$$

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 $\xrightarrow{p \to \infty}$ $$\partial_t \rho_t = \gamma \nabla_\theta \left(\rho_t \nabla_\theta \ell(\theta; \rho_t)\right)$$

"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]

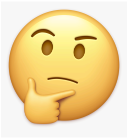
Theorem 2 (Informal) *If the support of the initial distribution includes all directions in $\mathbb{R}^{d+1}$, and if the function Ψ is positively 2-homogeneous then if the Wasserstein gradient flow weakly converges to a distribution, it can only be to a global optimum of F.*

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']

Mean-field regime



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

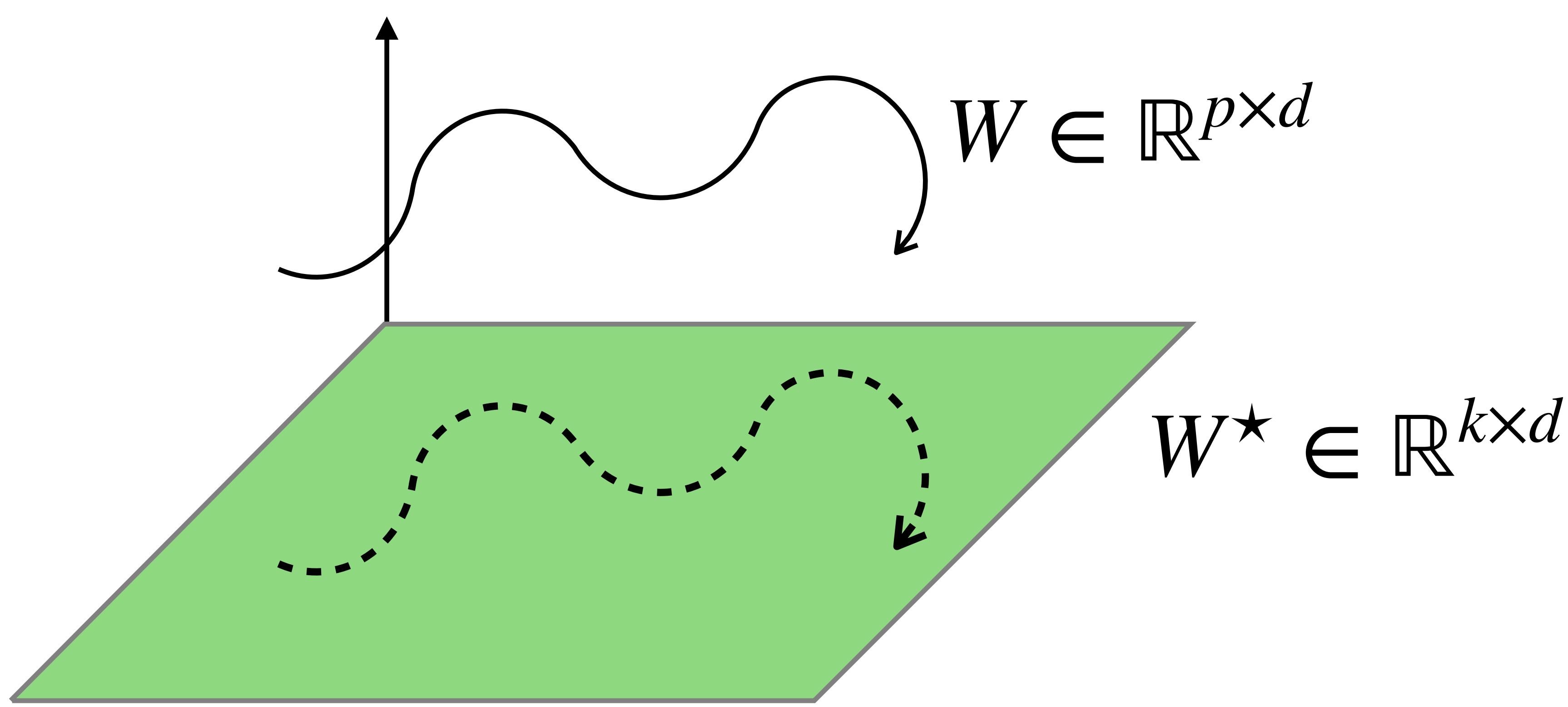
$W \in \mathbb{R}^{p \times d}$

Mean-field regime

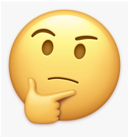
But $Q \in \mathbb{R}^{p \times p}$!!!

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

Teacher subspace



Mean-field regime

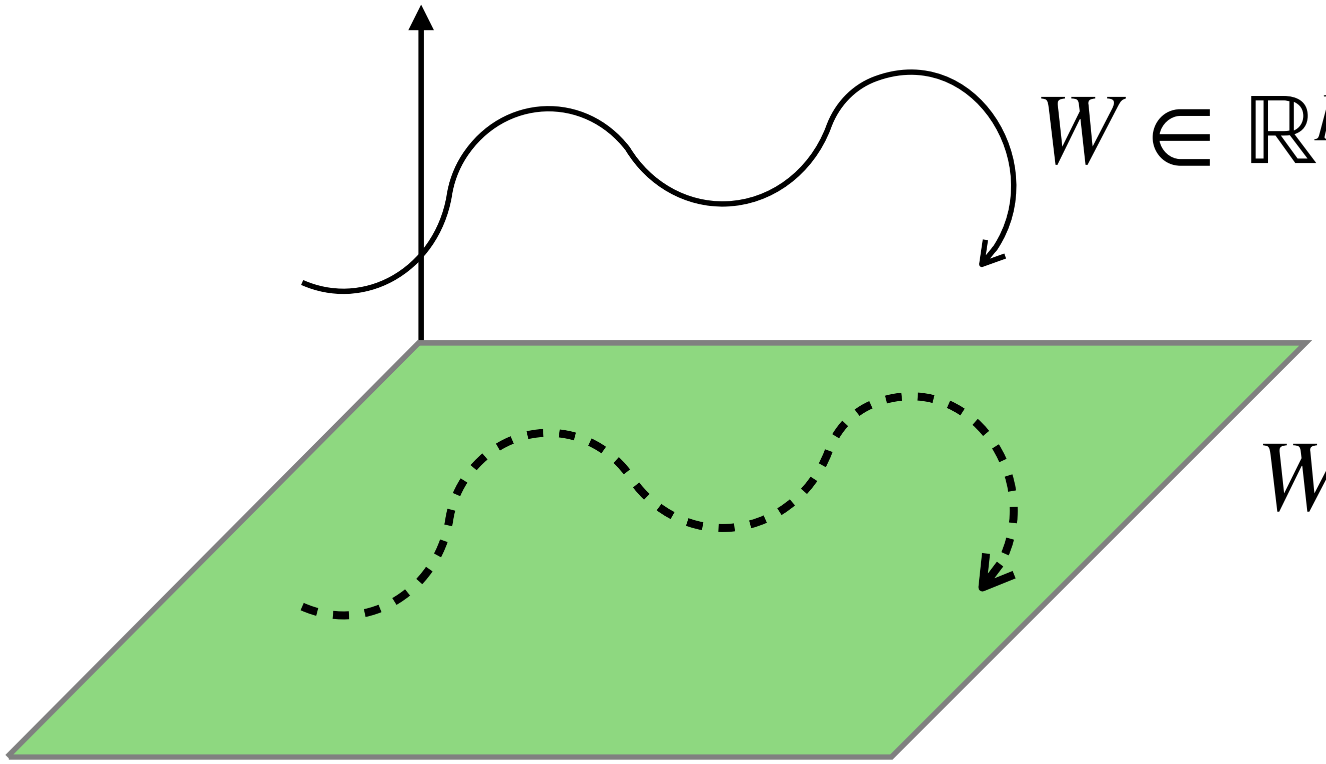
 But $Q \in \mathbb{R}^{p\times p}$!!!

$$W = MP^{-1}W^{\star} + W^{\perp} \longrightarrow Q \approx MPM^{\top} + D_{\sqrt{q^{\perp}}}\Xi D_{\sqrt{q^{\perp}}}$$

Teacher subspace

$$\wr$$

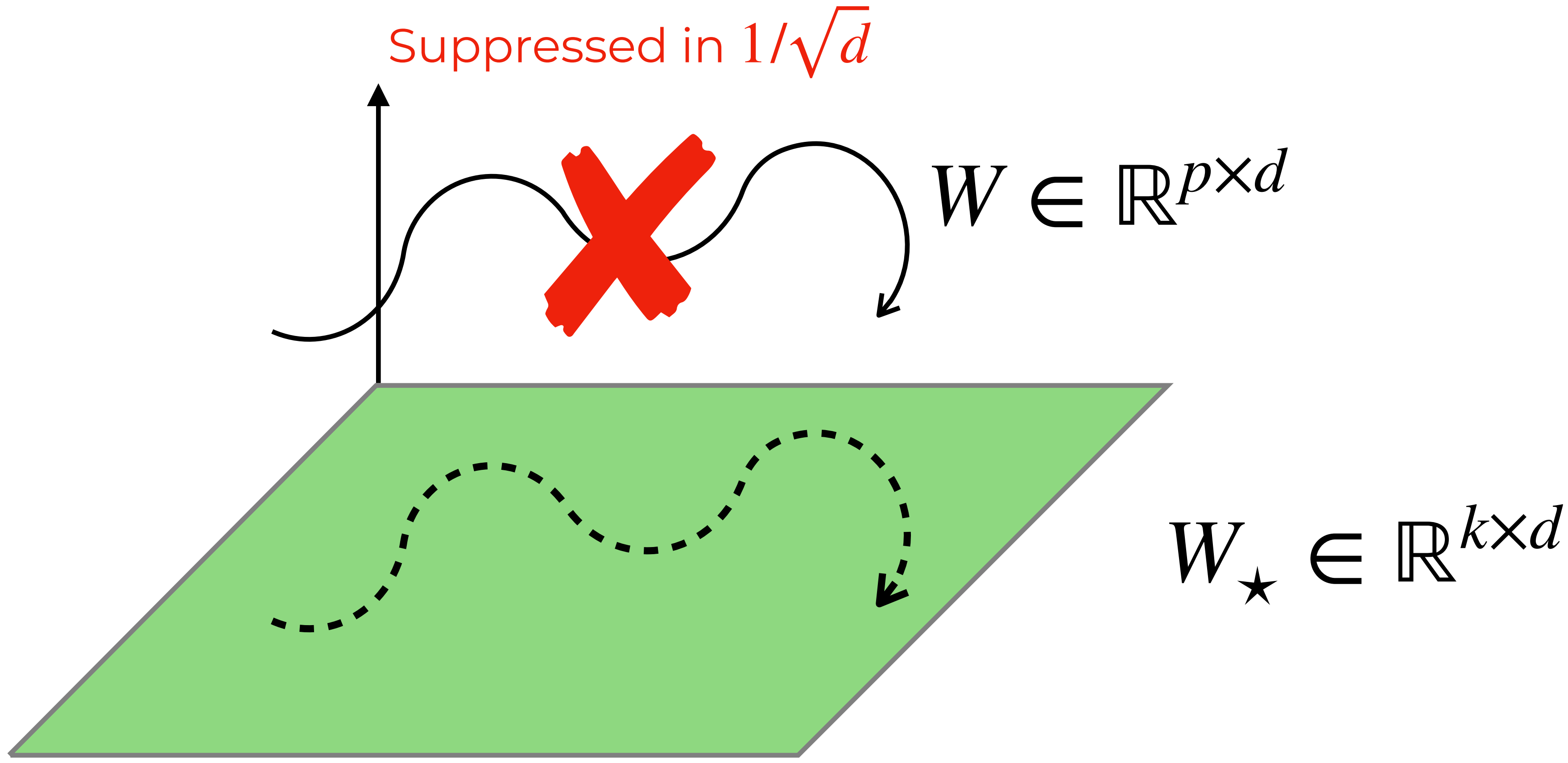
$$\mathbb{S}^{d-k-1}$$



Mean-field + high-d

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

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



Mean-field + high-d

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

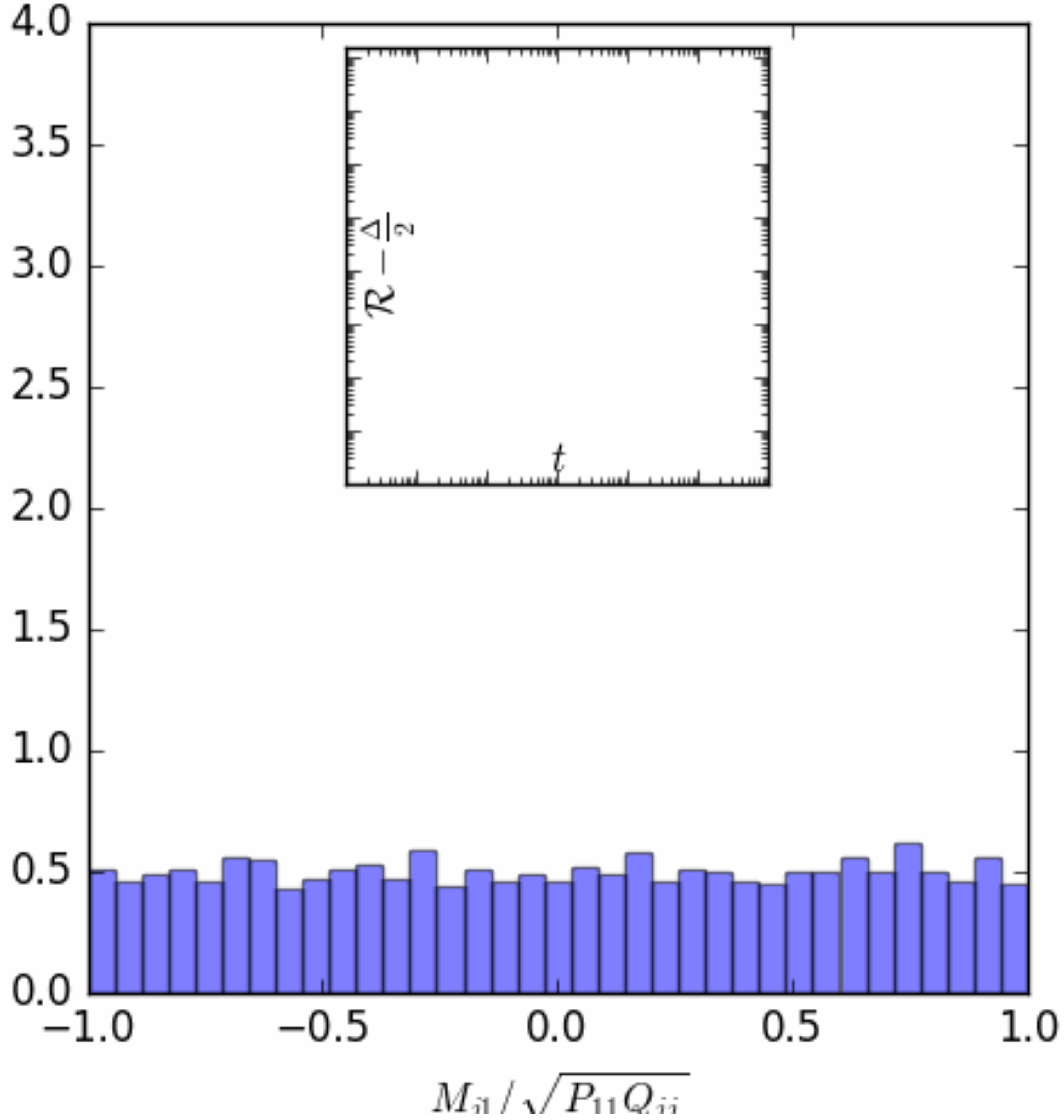
$$\mathbb{E}\,||\,Q(t) - MP^{-1}M^{\top} + \mathrm{diag}(Q^{\perp})\,||_{\infty} \leq e^{Ct}\left(p^{-1/2} + d^{-1/2}\right)$$

This implies MF-like PDE for the sufficient statistics:

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

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

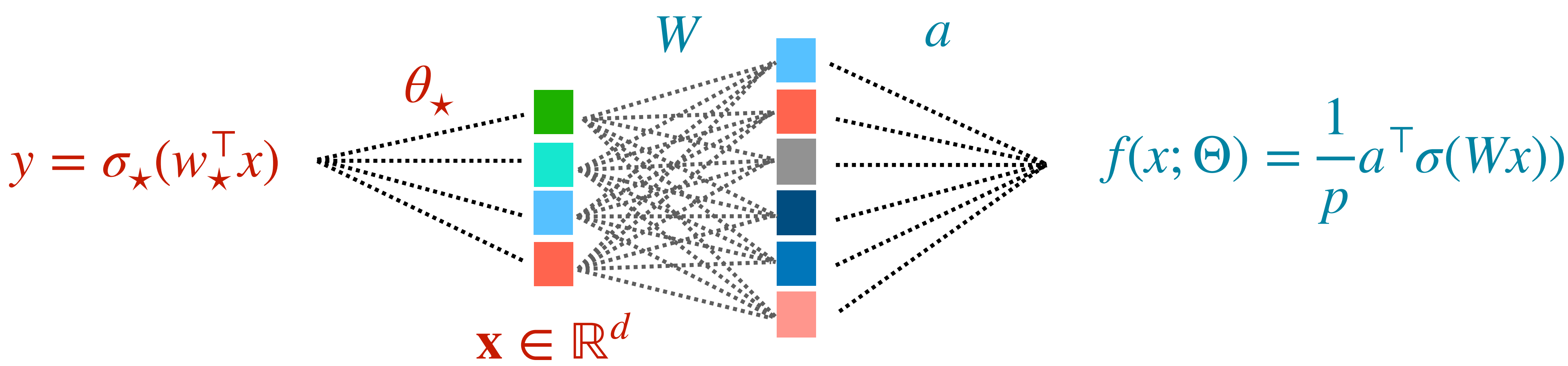
Mean-field and high-d



What can I do with that?

From [Berthier, Montanari & Zhou '23]

Consider simple case: $k = 1$ and $p \to \infty$

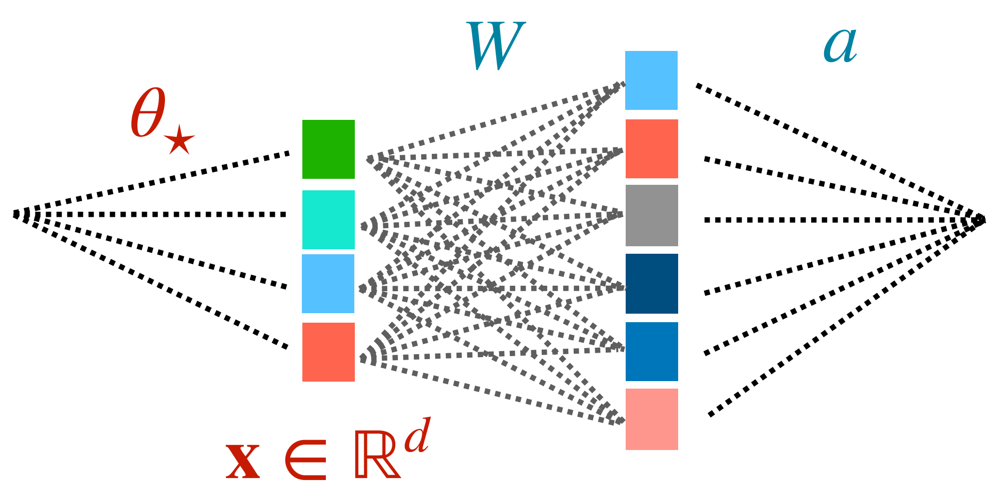


What can I do with that?

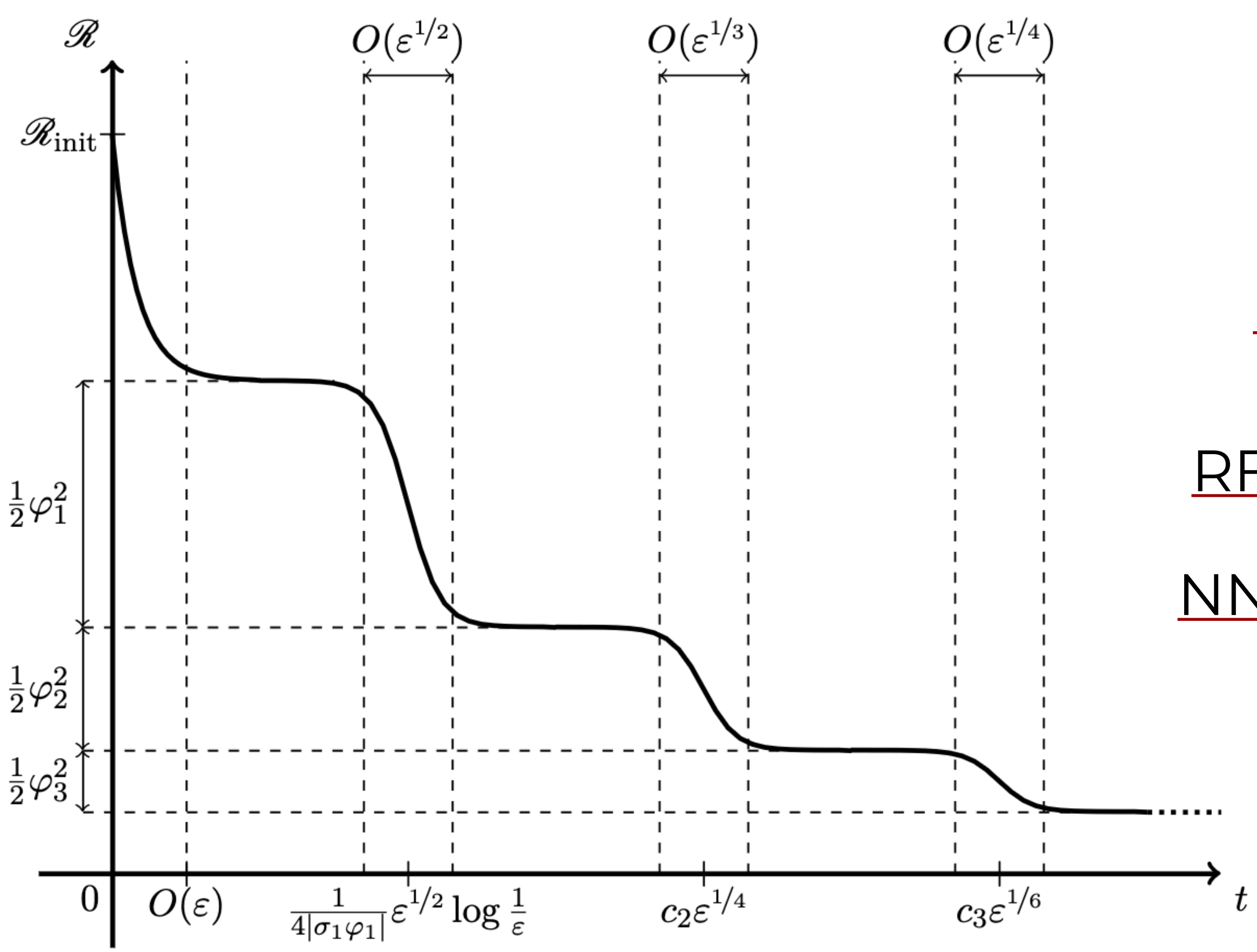
From [Berthier, Montanari & Zhou '23]

Consider simple case: $k = 1$ and $p \to \infty$

$y = \sigma_\star(w_\star^\top x)$



$f(x;\Theta) = \frac{1}{p} a^\top \sigma(Wx))$



Recall: For $n \propto d$ and $x \sim \mathcal{N}(0, I_d)$

RF / Kernels: can only learn linear part

NNs: can learn non-linear components if $(\sigma_\star, \sigma)$ "standard" (full Hermite expansion).

Follow-ups and challenges

- Characterisation of the stochastic dynamics close to fixed points as a coloured diffusion [Ben Arous, Gheissari, Jagannath NeurIPS '22]

Follow-ups and challenges

- Characterisation of the stochastic dynamics close to fixed points as a coloured diffusion [Ben Arous, Gheissari, Jagannath NeurIPS '22]

- Uniform control over the variance and convergence rates

Follow-ups and challenges

- Characterisation of the stochastic dynamics close to fixed points as a coloured diffusion — [Ben Arous, Gheissari, Jagannath NeurIPS ’22]
- Uniform control over the variance and convergence rates
- Phenomenology of the dynamics: Functions of increasing complexity? Distributions of increasing complexity? — [Abbe, Adsera, Misiakiewicz, COLT ’22; Berthier, Montanari ’23] [Refinetti, Ingrosso, Goldt, arXiv ’22]

Follow-ups and challenges

- Characterisation of the stochastic dynamics close to fixed points as a coloured diffusion

  [Ben Arous, Gheissari, Jagannath NeurIPS ’22]

- Uniform control over the variance and convergence rates

- Phenomenology of the dynamics: Functions of increasing complexity? Distributions of increasing complexity?

  [Abbe, Adsera, Misiakiewicz, COLT ’22; Berthier, Montanari ’23]

  [Refinetti, Ingrosso, Goldt, arXiv ’22]

- Role of initialisation

Lecture III: Summary

✓ Deterministic analysis of two-layer neural nets in the “rich” regime

Lecture III: Summary

- ✓ Deterministic analysis of two-layer neural nets in the “rich” regime
- ✓ Phenomenology in the classical and high-dimensional regime

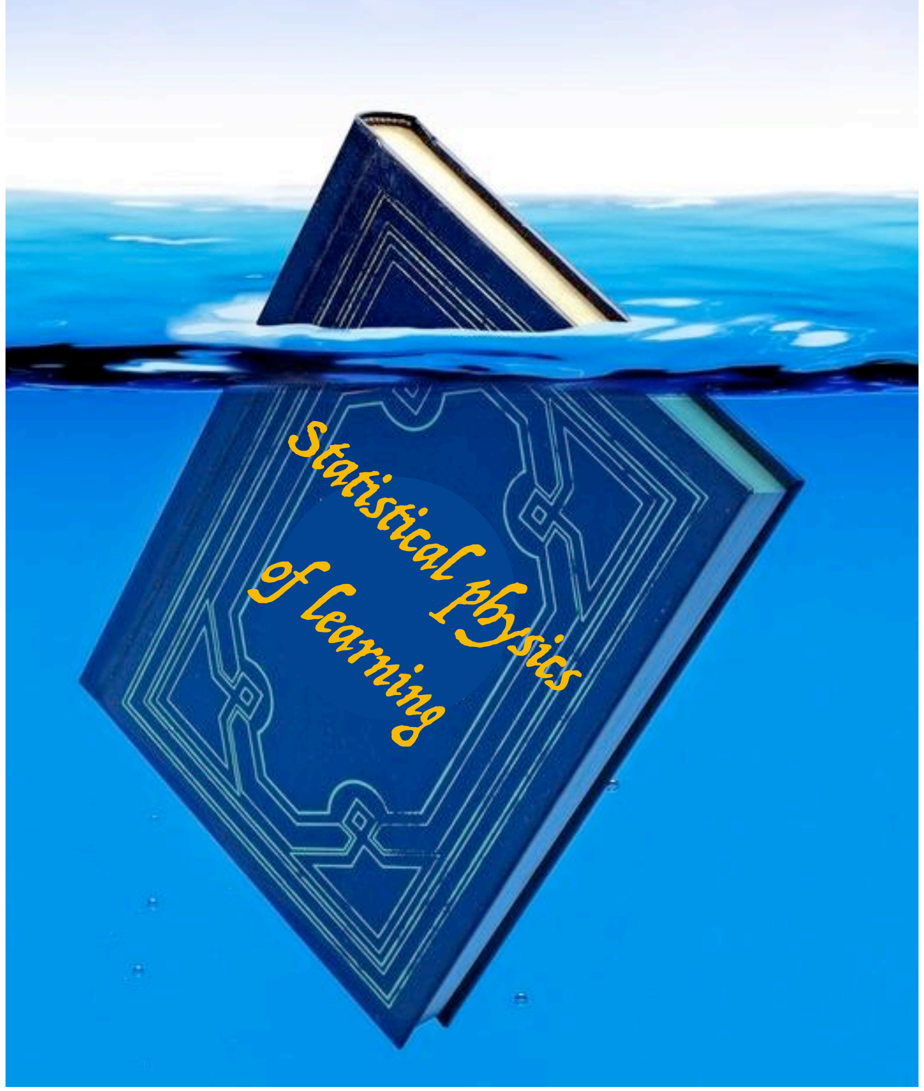
Lecture III: Summary

- Deterministic analysis of two-layer neural nets in the "rich" regime
- Phenomenology in the classical and high-dimensional regime
- Interplay between effective SGD noise and overparametrisation

Lecture III: Summary

- ✓ Deterministic analysis of two-layer neural nets in the “rich” regime
- ✓ Phenomenology in the classical and high-dimensional regime
- ✓ Interplay between effective SGD noise and overparametrisation
- ✓ Dimension free limits in the mean-field regime

But this is only the tip of an iceberg...



brloureiro@gmail.com