Phase diagram of Stochastic Gradient Descent

in high-dimensional two-layer neural networks

Bruno Loureiro
@ CSD, DI-ENS

brloureiro@gmail.com

Based on
arXiv: 2202.00293
(NeurIPS 2022)

Seminar in Mathematics, Physics & Machine Learning, 15.12.2022

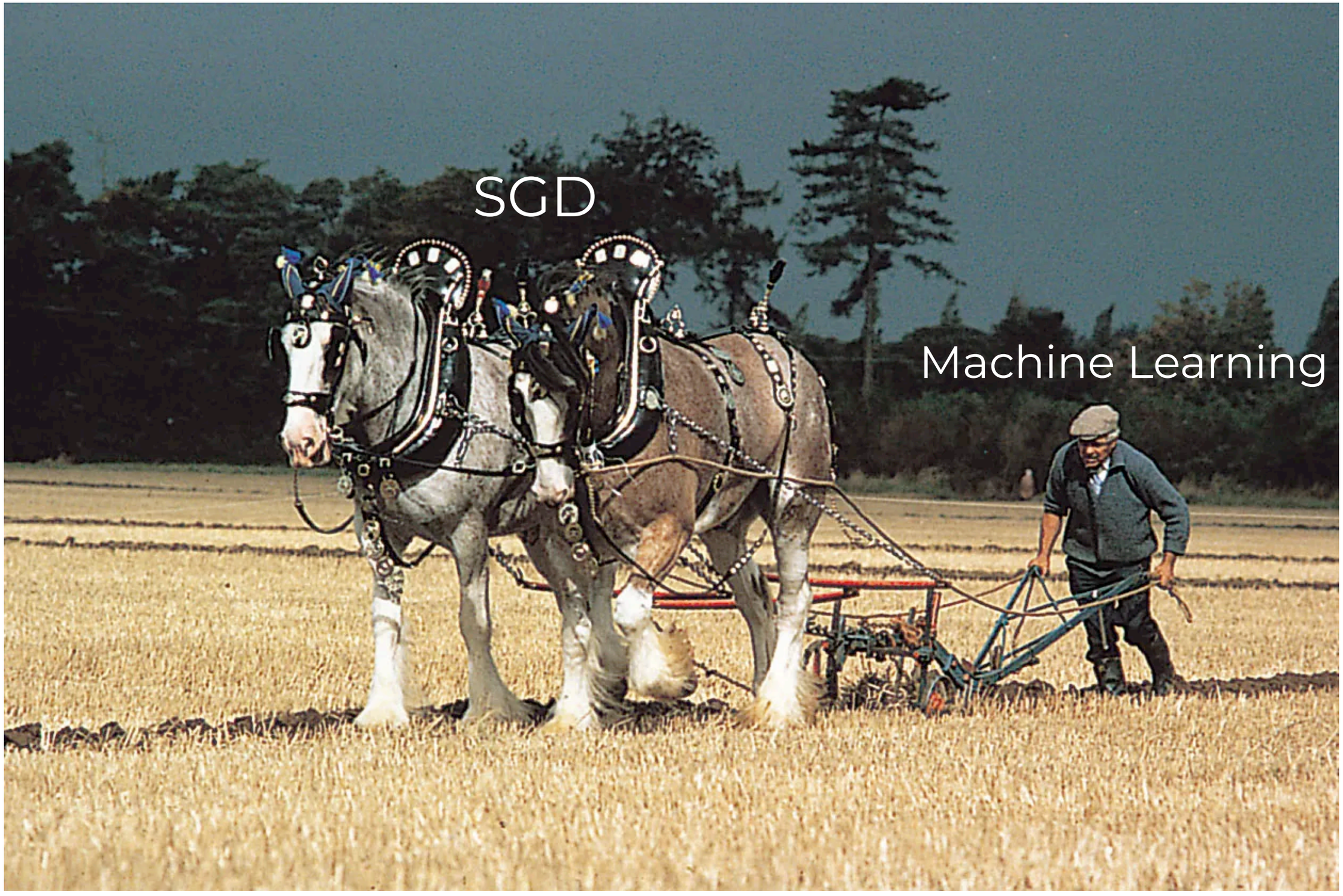
In collaboration with

Rodrigo Veiga
(EPFL)

Ludovic Stéphan
(EPFL)

Lenka Zdeborová
(EPFL)

Florent Krzakala
(EPFL)


SGD
Machine Learning

Many open questions

Bad Global Minima Exist and SGD Can Reach Them

Shengchao Liu
Quebec Artificial Intelligence Institute (Mila)
Université de Montréal
liusheng@mila.quebec

Dimitris Papailiopoulos
University of Wisconsin-Madison
dimitris@papail.io

Dimitris Achlioptas
University of Athens
optas@di.uoa.gr

Abstract

Several works have aimed to explain why overparameterized neural networks generalize well when trained by Stochastic Gradient Descent (SGD). The consensus explanation that has emerged credits the randomized nature of SGD for the bias of the training process towards low-complexity models and, thus, for implicit regularization. We take a careful look at this explanation in the context of image classification with common deep neural network architectures. We find that if we do not regularize *explicitly*, then SGD can be easily made to converge to poorly-generalizing, high-complexity models: all it takes is to first train on a random labeling on the data, before switching to properly training with the correct labels. In contrast, we find that in the presence of explicit regularization, pretraining with random labels has no detrimental effect on SGD. We believe that our results give evidence that explicit regularization plays a far more important role in the success of overparameterized neural networks than what has been understood until now. Specifically, by penalizing complicated models independently of their fit to the data, regularization affects training dynamics also far away from optima, making simple models that fit the data well discoverable by local methods, such as SGD.

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

Algorithms: GD

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

Population Risk

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

Empirical Risk

- Gradient descent (GD):

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

Algorithms: GD

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

Population Risk

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

Empirical Risk

- Gradient descent (GD):

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

Defining $t = k\delta t$ with $\delta t = \gamma_k$, at fixed n, d the limit $\gamma_k \to 0$ yields gradient flow:

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

Algorithms: SGD

- Stochastic Gradient descent (SGD): at every k, choose mini-batch $B_k \subset [n]$

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

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

mini-batch

Algorithms: SGD

- Stochastic Gradient descent (SGD): at every k, choose mini-batch $B_k \subset [n]$

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

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

mini-batch

One-pass limit: at each k, take fresh data

$$\Theta^{\nu+1} = \Theta^\nu - \gamma_\nu \nabla_{\Theta^\nu} \left[\frac{1}{2}\left(y^\nu - f_{\Theta^\nu}(x^\nu)\right)^2\right]$$

$|B_\nu| = 1$

Algorithms: SGD

- Stochastic Gradient descent (SGD): at every k, choose mini-batch $B_k \subset [n]$

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

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

mini-batch

One-pass limit: at each k, take fresh data

$$\Theta^{\nu+1} = \Theta^\nu - \gamma_\nu \nabla_{\Theta^\nu} \left[\frac{1}{2}\left(y^\nu - f_{\Theta^\nu}(x^\nu)\right)^2\right] \qquad |B_\nu| = 1$$

As before, taking the limit $\gamma_k \to 0$ at fixed n, d:

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

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

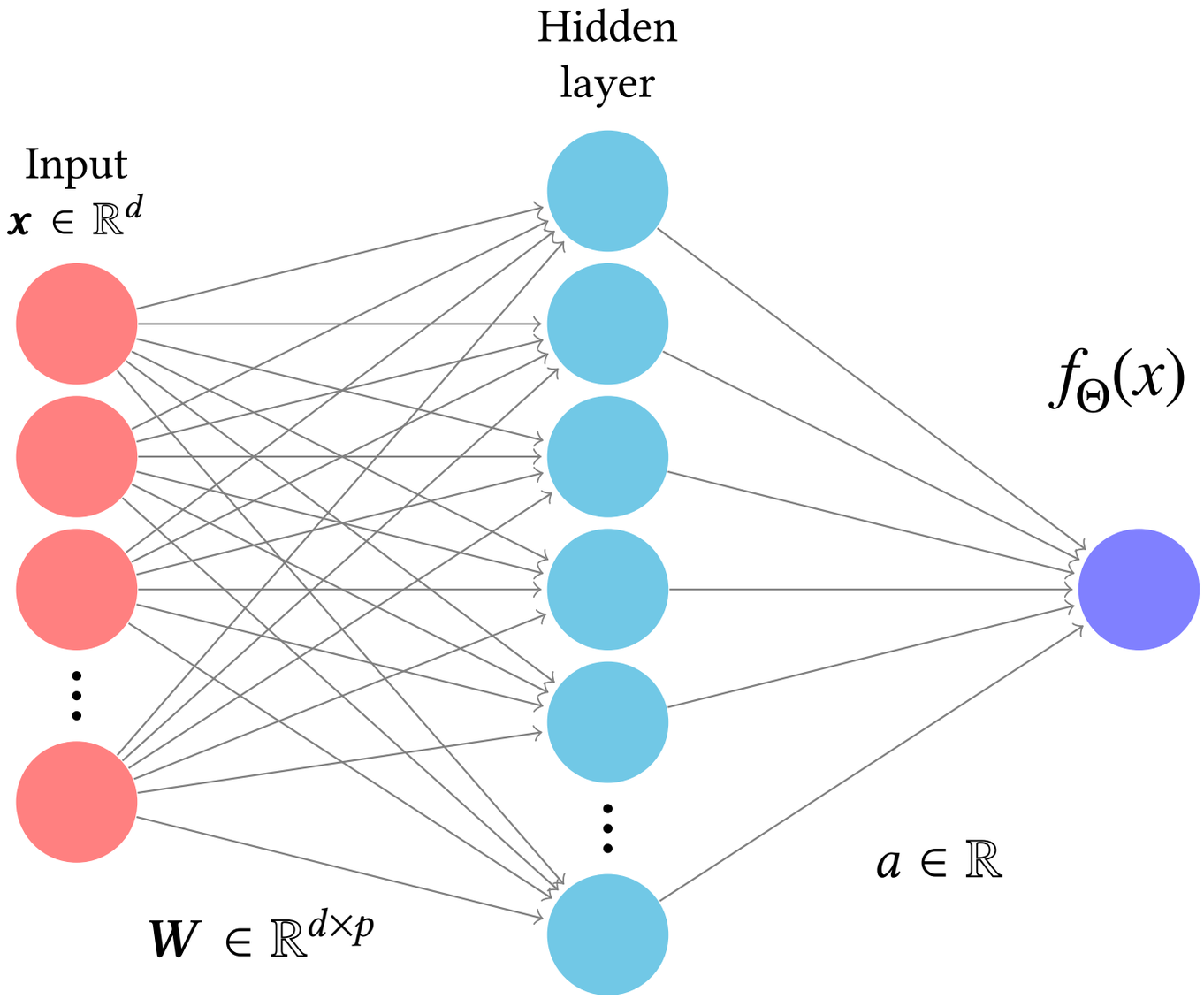
Where:

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

🤔 Question: How to characterise this?

Two-layers: a toy problem

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



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

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

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:

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

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

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

"Mean-field" limit

Where:

$$R(\theta;\rho) = V(\theta) + \int_{\mathbb{R}^{d+1}} \rho(\mathrm{d}\theta')\mathrm{U}(\theta,\theta')$$

$$V(\theta) = a\ \mathbb{E}_{(x,y)\sim\rho}\left[y\sigma(w^\mathsf{T}x)\right] \qquad U(\theta,\theta') = aa'\ \mathbb{E}_{x\sim\rho_x}\left[\sigma(w^\mathsf{T}x)\sigma(w'^\mathsf{T}x)\right]$$

[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

VOLUME 74, NUMBER 21 PHYSICAL REVIEW LETTERS 22 MAY 1995

Exact Solution for On-Line Learning in Multilayer Neural Networks

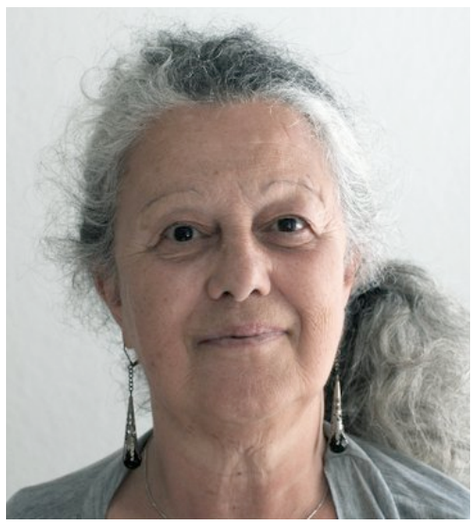
David Saad[1] and Sara A. Solla[2]

[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)

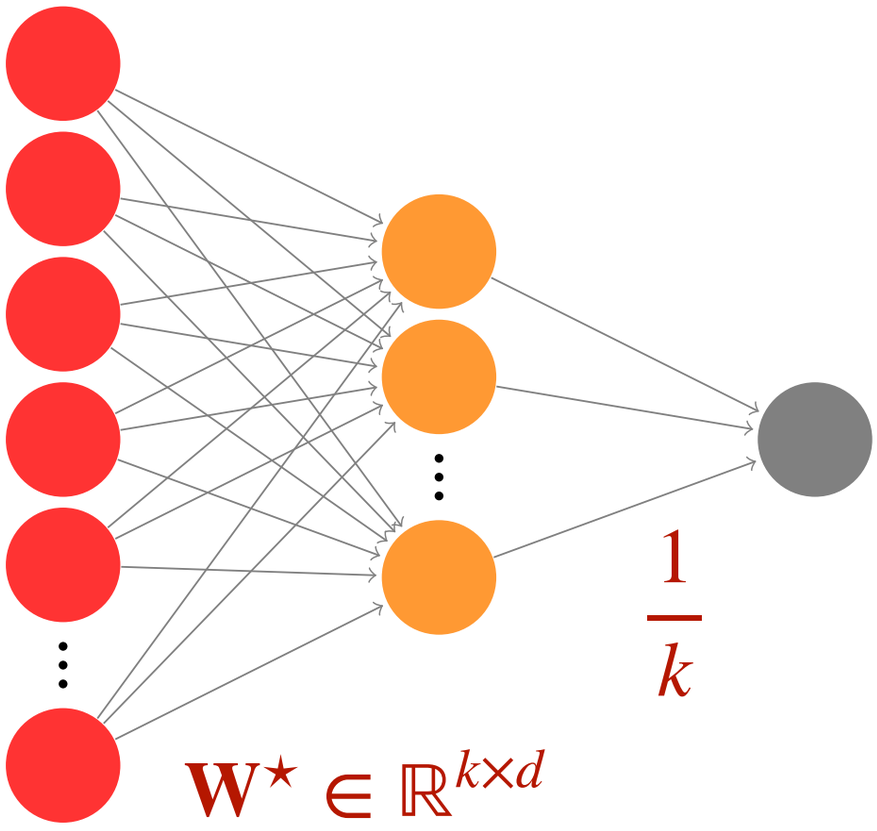
We present an analytic solution to the problem of on-line gradient-descent learning for two-layer neural networks with an arbitrary number of hidden units in both teacher and student networks.

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



Teacher-student setting

Teacher network



$\mathbf{W}^\star \in \mathbb{R}^{k\times d}$

$$\mathbf{x}^\nu \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_\mathrm{d}) \qquad \zeta^\nu \sim \mathcal{N}(0,1)$$

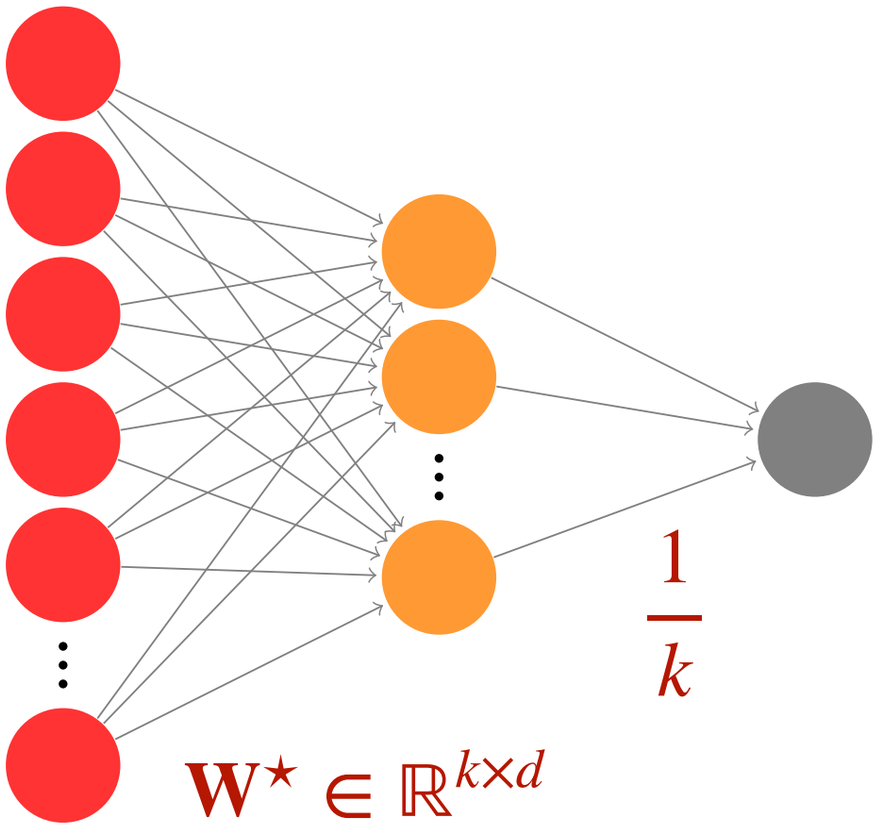
$$y^\nu = f_{\mathbf{W}*}(\mathbf{x}^\nu) + \sqrt{\Delta}\zeta^\nu$$

$$f_{\Theta*}(\mathbf{x}) = \frac{1}{k}\sum_{r=1}^{k}\sigma(\mathbf{w}_r^{*\top}\mathbf{x})$$

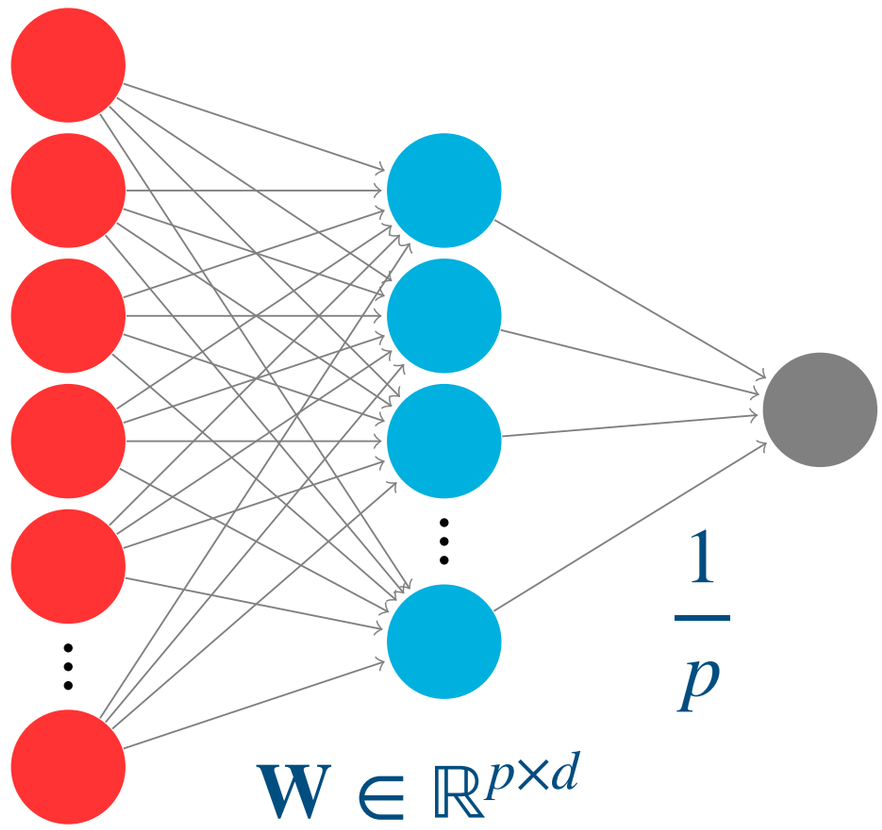
Teacher-student setting

$$\mathbf{x}^\nu \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_\mathrm{d}) \qquad \zeta^\nu \sim \mathcal{N}(0,1)$$

Teacher network



Student network



$$f_{\Theta*}(\mathbf{x}) = \frac{1}{k}\sum_{r=1}^{k} \sigma(\mathbf{w}_r^{*\top}\mathbf{x})$$

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

$$\hat{f}_\Theta(\mathbf{x}) = \frac{1}{p}\sum_{i=1}^{p} \sigma(\mathbf{w}_i^\top\mathbf{x})$$

Sufficient statistics

Goal: track population error exactly throughout the dynamics

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

Sufficient statistics

Goal: track population error exactly throughout the dynamics

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

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

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

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

Sufficient statistics

After some work...

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

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

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

Sufficient statistics

After some work...

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

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

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

Population gradient

Noise

Deterministic limit

Theorem (Saad,Solla '95; Reents, Urbanczik '98; Goldt et al '19)

Defining $t = \nu\delta t$ with $\delta t = 1/d$, for $p, k, \gamma = O(1)$,
for $d \to \infty$:

$$\dot{M}(t) = \frac{\gamma}{p}\,\bar{\psi}_M(M(t), Q(t))$$

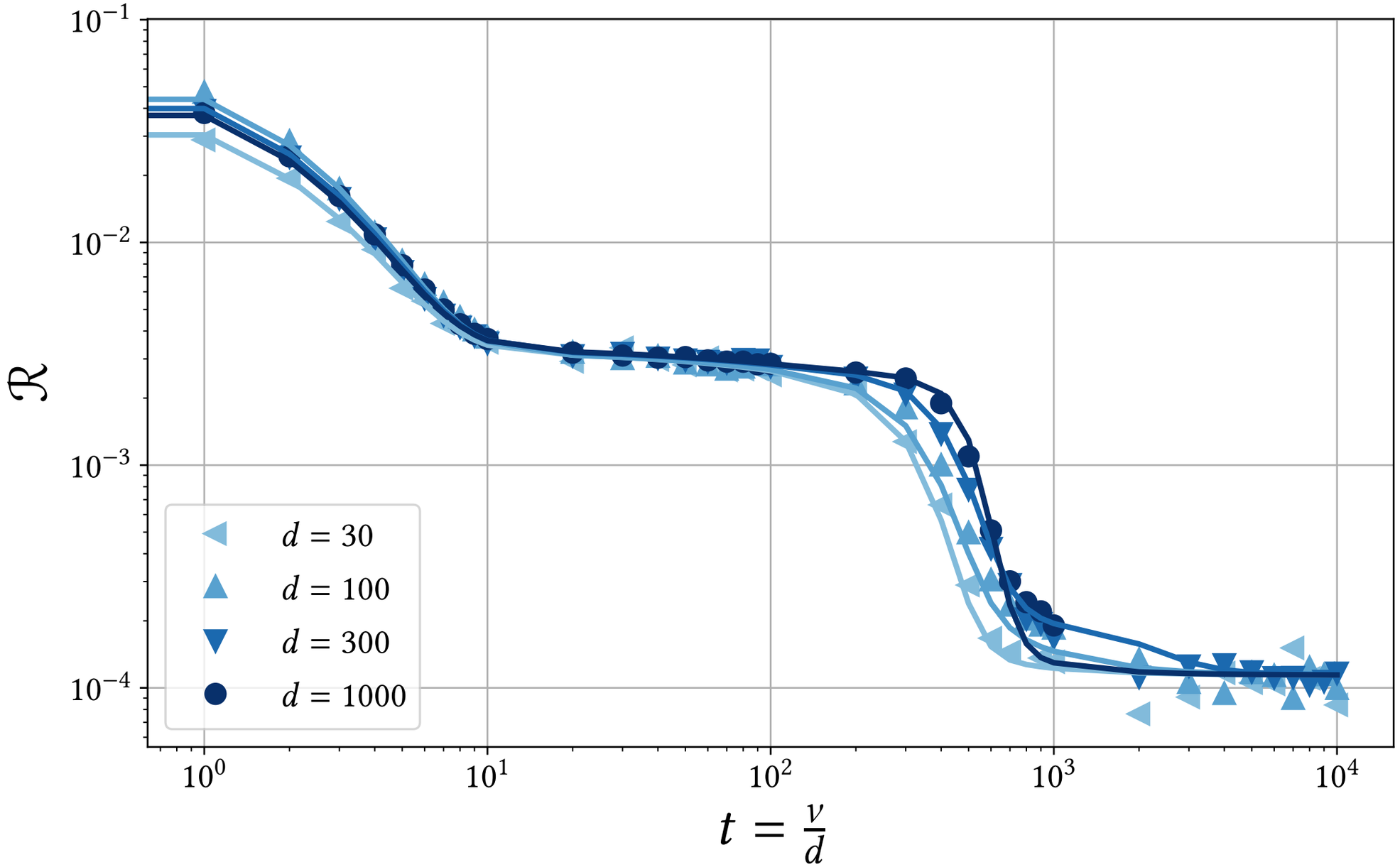
$$\dot{Q}(t) = \frac{\gamma}{p}\,\bar{\psi}_Q^{(1)}(M(t), Q(t)) + \frac{\gamma^2}{p^2}\,\bar{\psi}_Q^{(2)}(M(t), Q(t))$$

Note: number of samples seen at time $t = O(1)$ is $n \sim td$

Typical learning curve

$$k = 4$$
$$p = 8$$

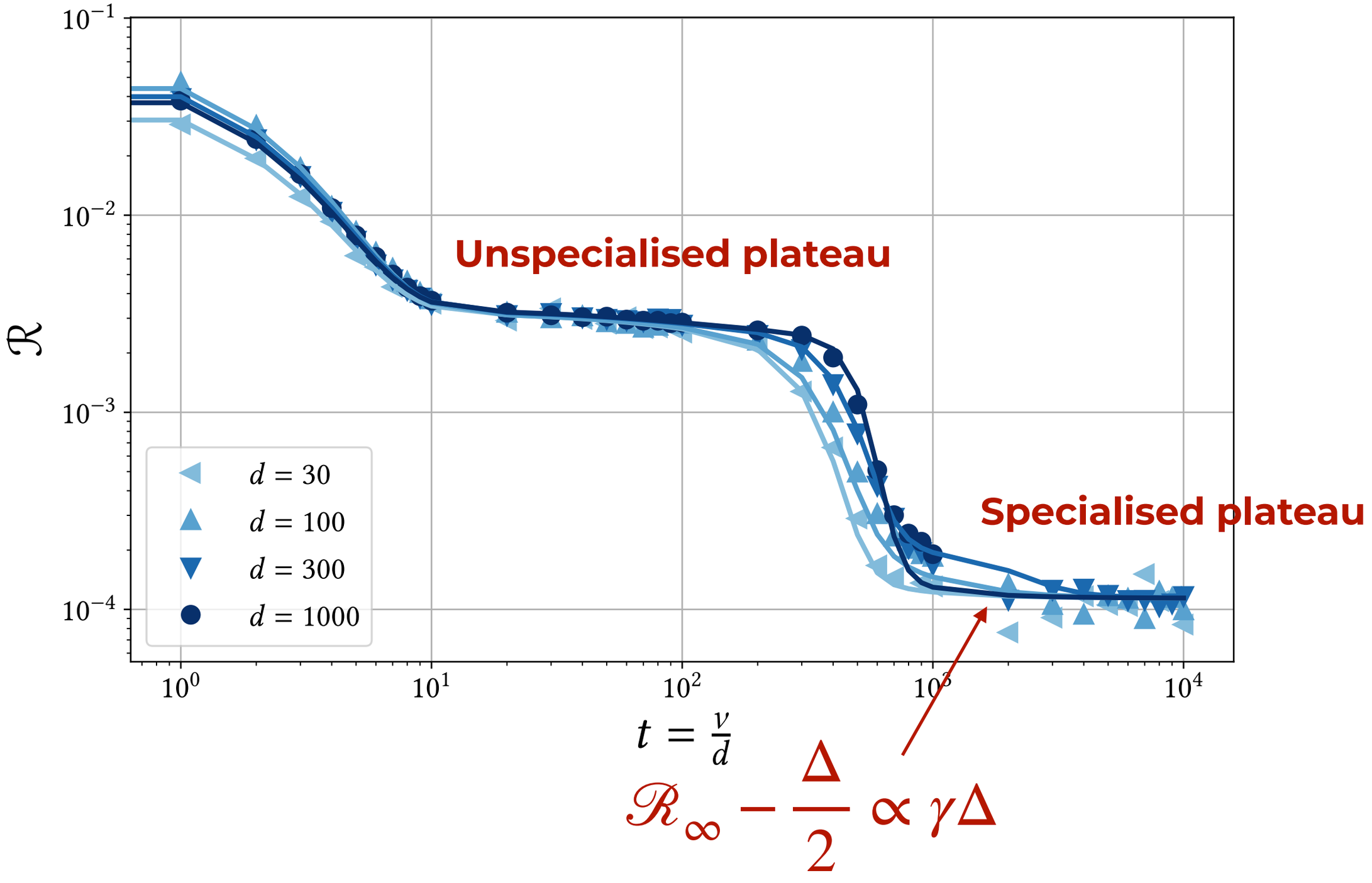
$$\mathbf{w}_i(t=0) \sim \mathcal{N}(0, \mathrm{I_d})$$
$$\sigma(x) = \mathrm{erf}\left(x/\sqrt{2}\right)$$



Typical learning curve

$$k = 4$$
$$p = 8$$

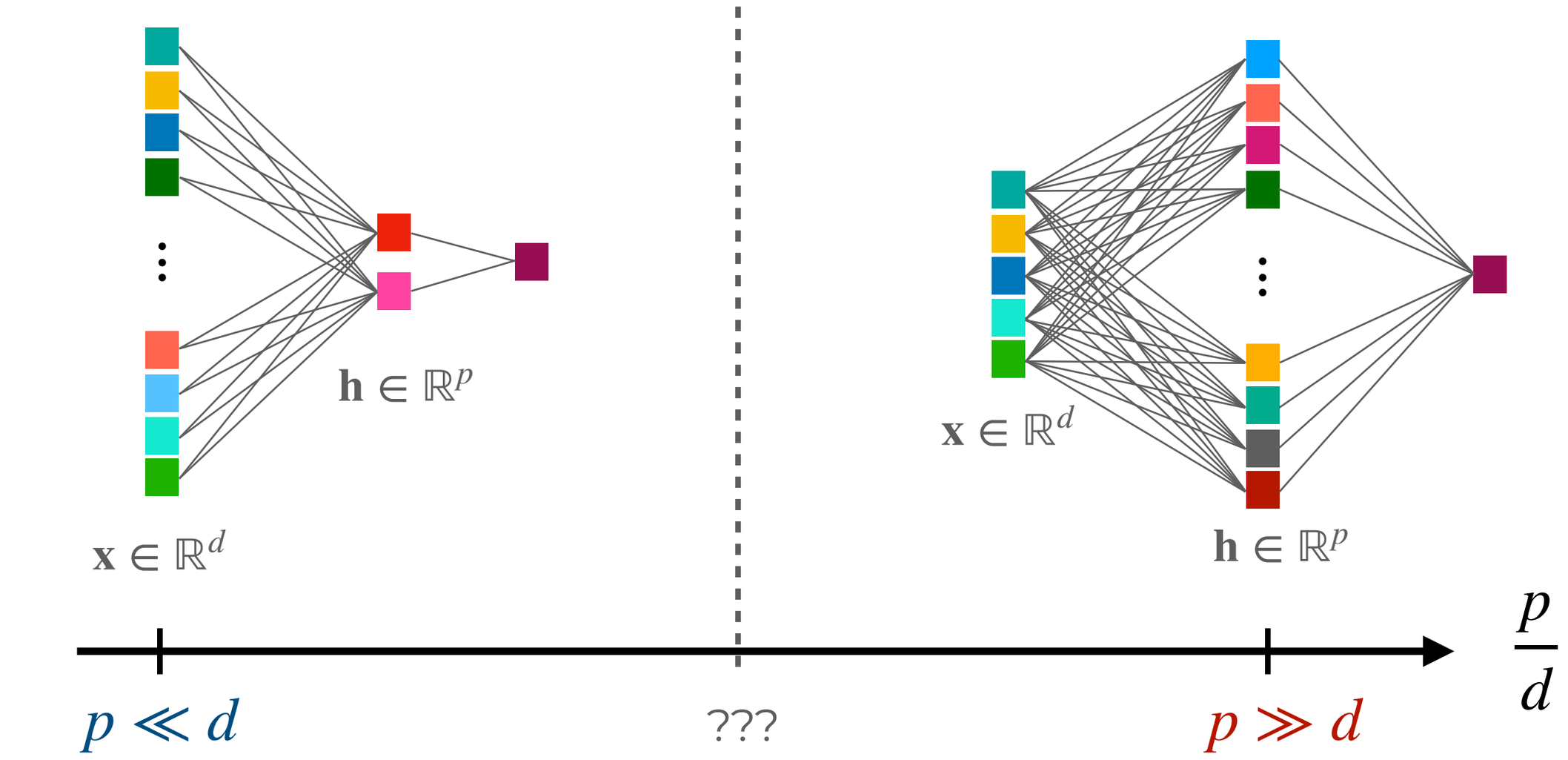
$$\mathbf{w}_i(t=0) \sim \mathcal{N}(0, \mathrm{I_d})$$
$$\sigma(x) = \mathrm{erf}\left(x/\sqrt{2}\right)$$



Bridging the two regimes

Narrow networks

Wide networks



(Saad & Solla)

(Mean-field limit)

Bridging the two regimes

Saad & Solla ‘95

$$d \to \infty$$

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

Our work

$$d \to \infty$$

$$k = O(1)$$

$$p \sim d^{\kappa} \qquad \kappa > 0$$

$$\gamma \sim d^{-\delta} \qquad \kappa + \delta > -\frac{1}{2}$$

Main theoretical result

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

Let $T \in \mathbb{R}_+$, $\delta t \geq c \max\left(\frac{\gamma}{dp}, \frac{\gamma^2}{dp^2}\right)$. Then $\forall 0 \leq \nu \leq \left\lfloor \frac{T}{\delta t} \right\rfloor$:

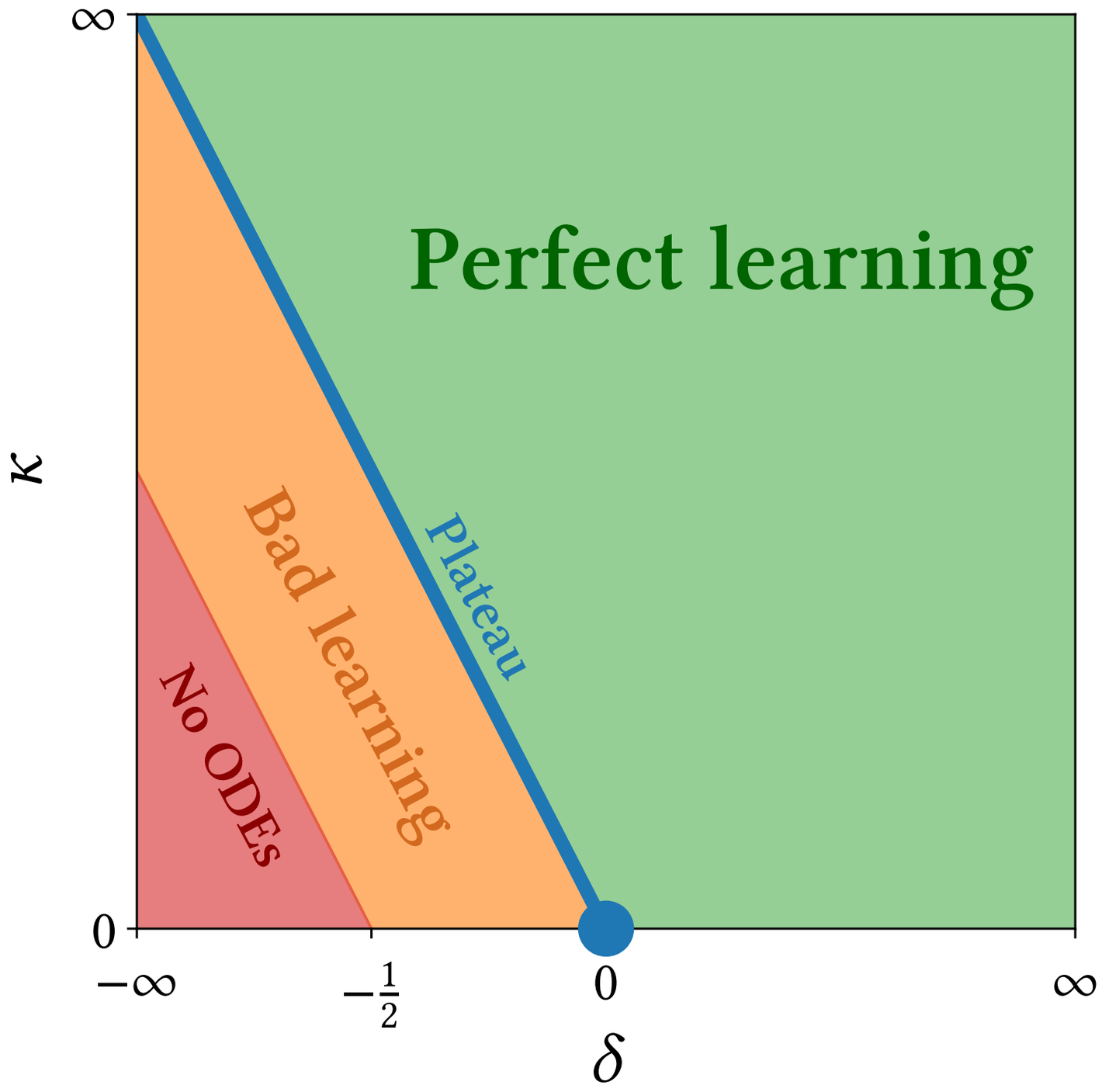
$$\mathbb{E}\,||\Omega^{\nu} - \bar{\Omega}(\nu\delta t)||_{\infty} \leq C(T)\log(p)\sqrt{\delta t}$$

Where $\bar{\Omega}(t)$ is the solution of an ODE:

$$\frac{\mathrm{d}\bar{\Omega}(\mathrm{t})}{\mathrm{dt}} = \psi(\bar{\Omega}(t))$$

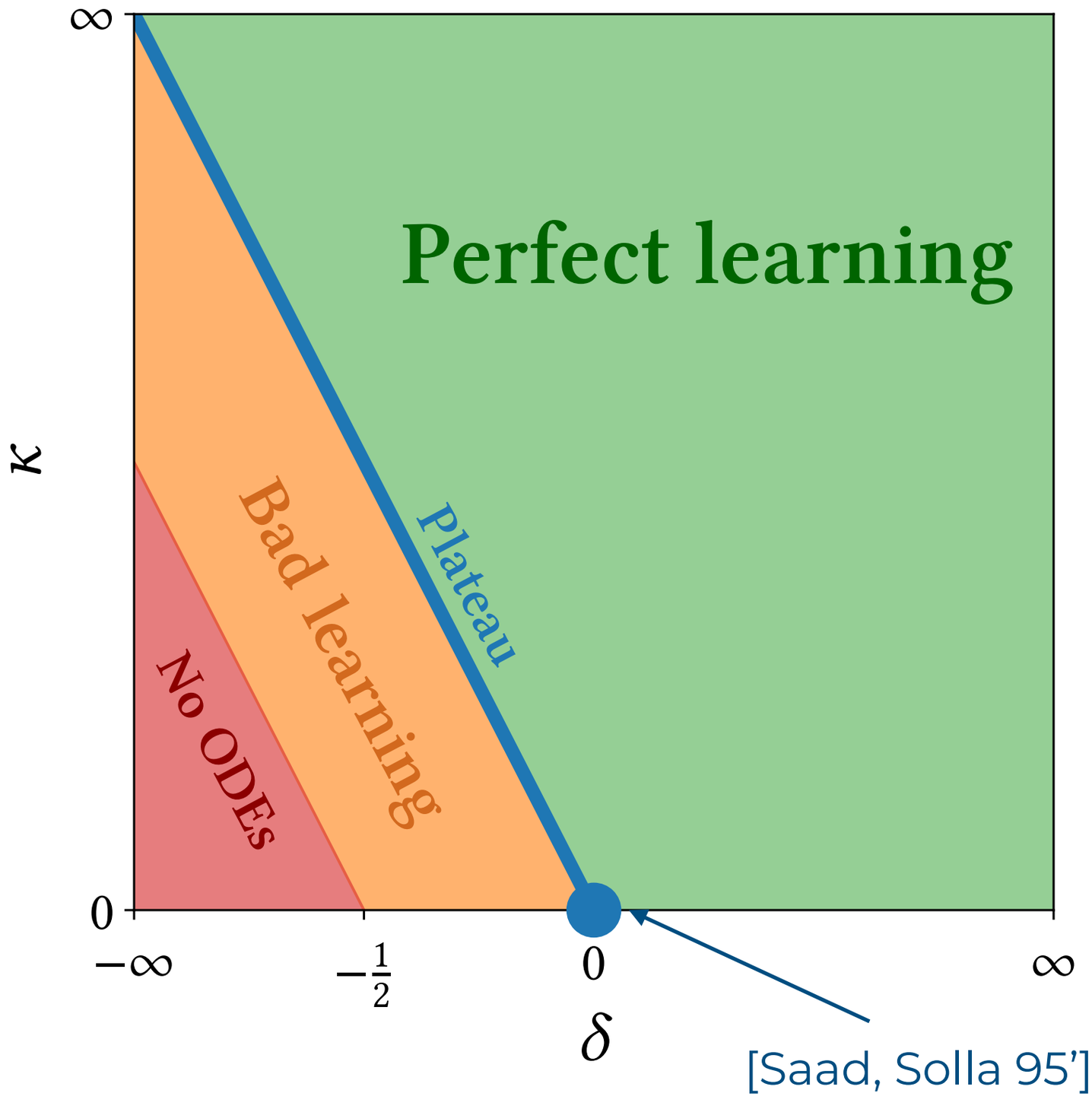
Phase diagram

$$p \sim d^{\kappa} \qquad \gamma \sim d^{-\delta}$$



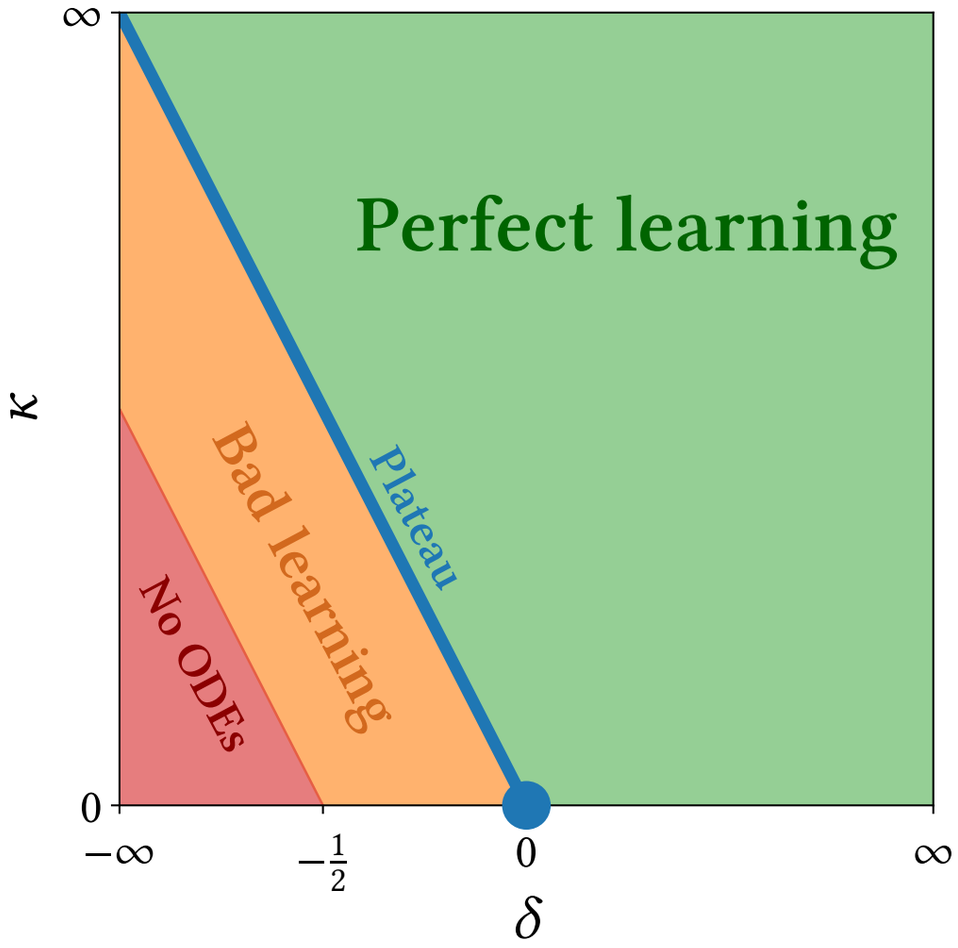
Phase diagram

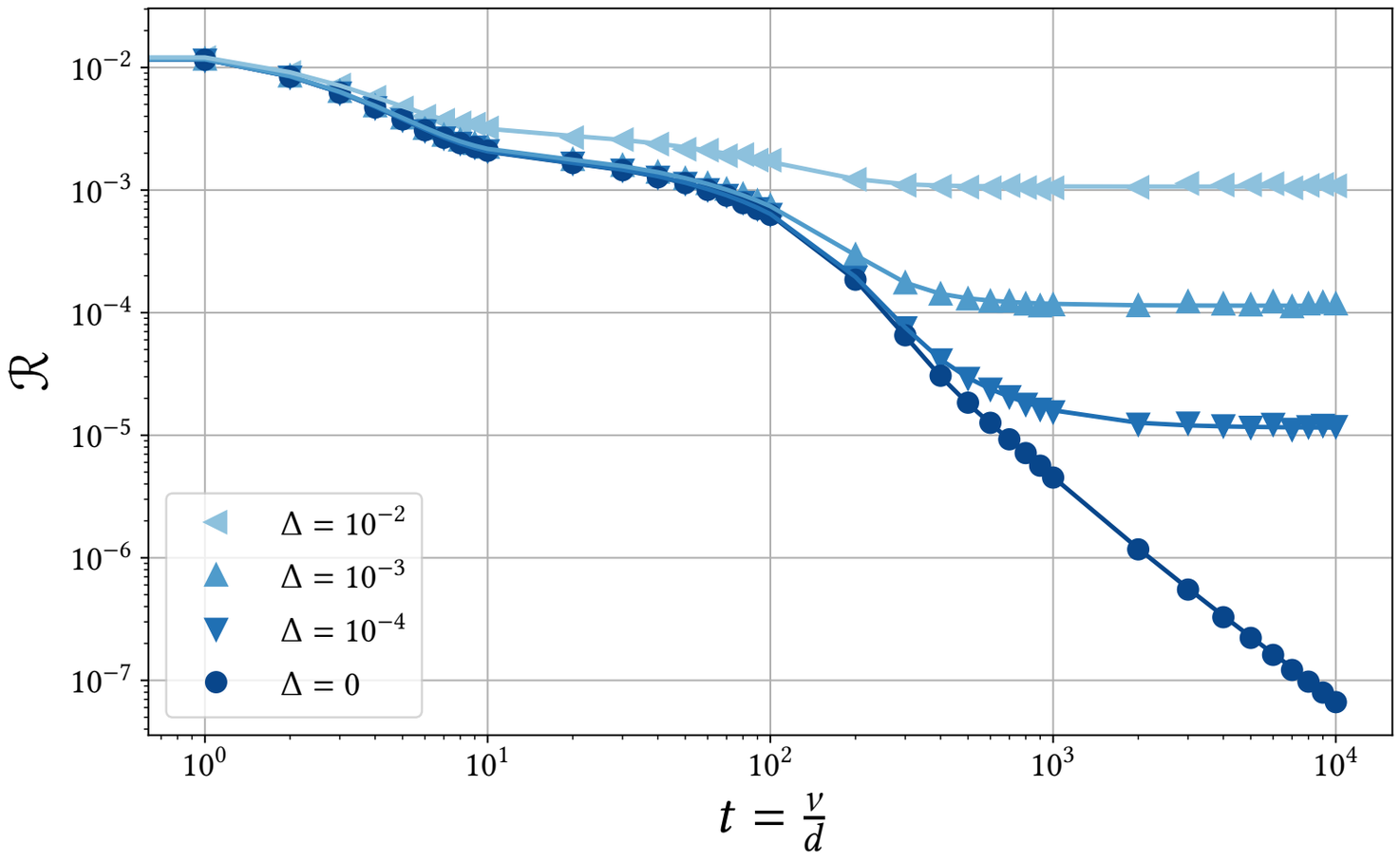
$$p \sim d^{\kappa} \qquad \gamma \sim d^{-\delta}$$



Blue line: $\kappa + \delta = 0$

$$\delta t = 1/d$$



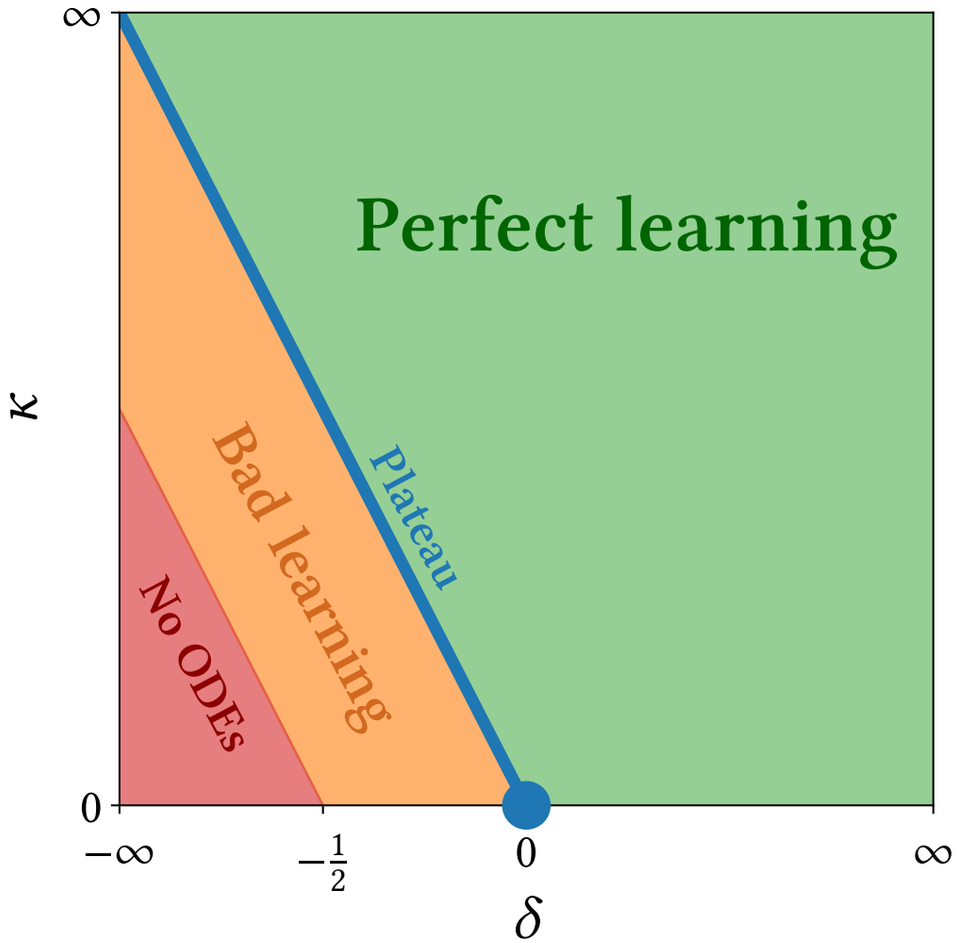


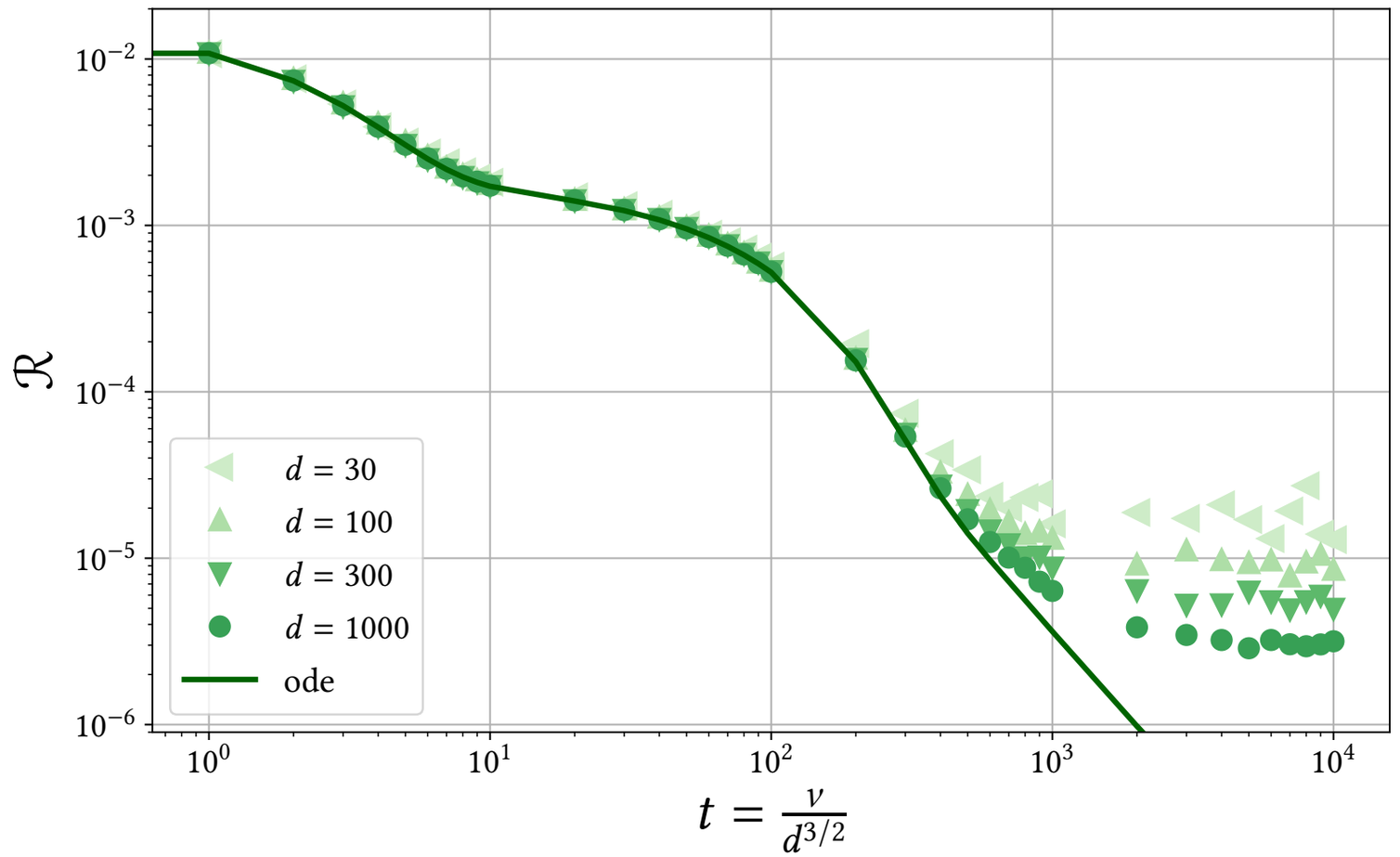
Extension of S&S regime to the whole blue line
(same phenomenology)

Green region: $\kappa + \delta > 0$

$$\delta t = 1/d^{1+\kappa+\delta}$$

$$\kappa = 0 \qquad \delta = 1/2$$



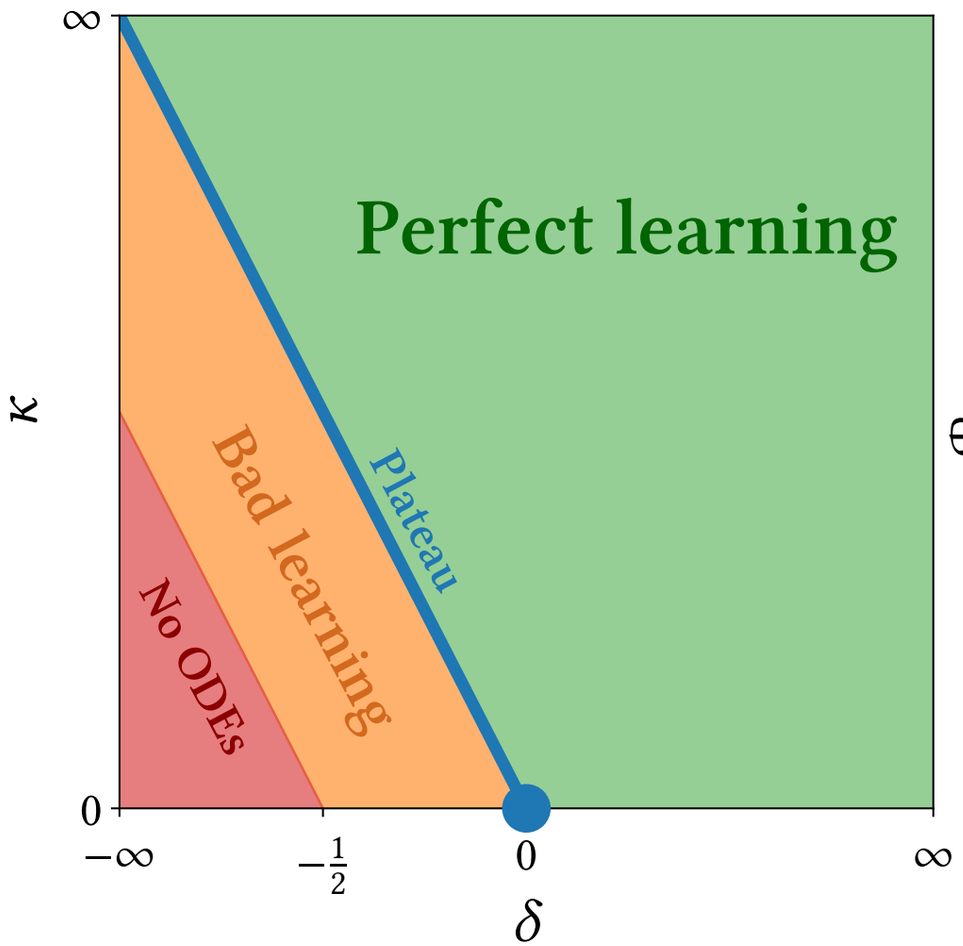


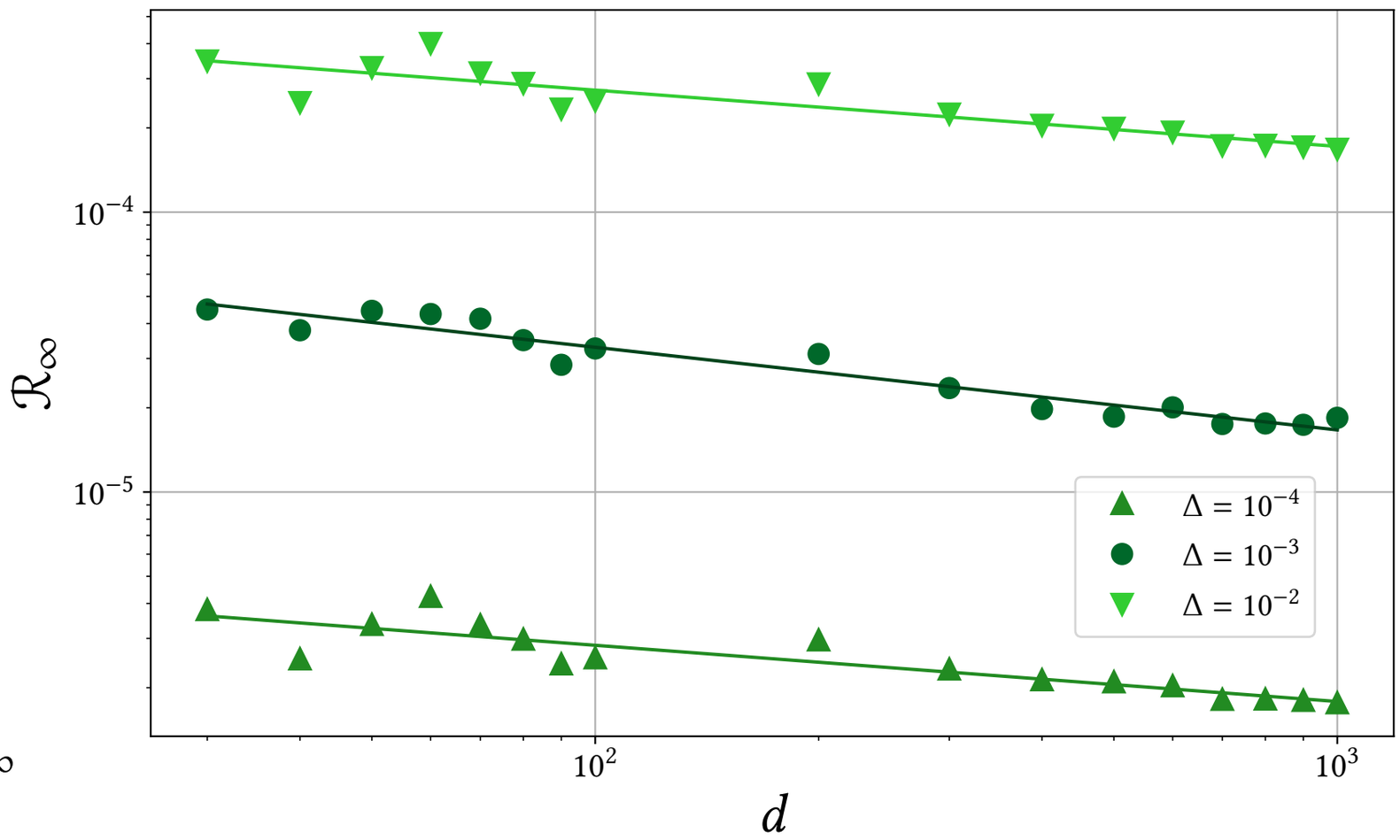
Perfect learning is achieved!

Green region: $\kappa + \delta > 0$

$$\delta t = 1/d^{1+\kappa+\delta}$$

$$\kappa = 0 \qquad \delta = 1/2$$



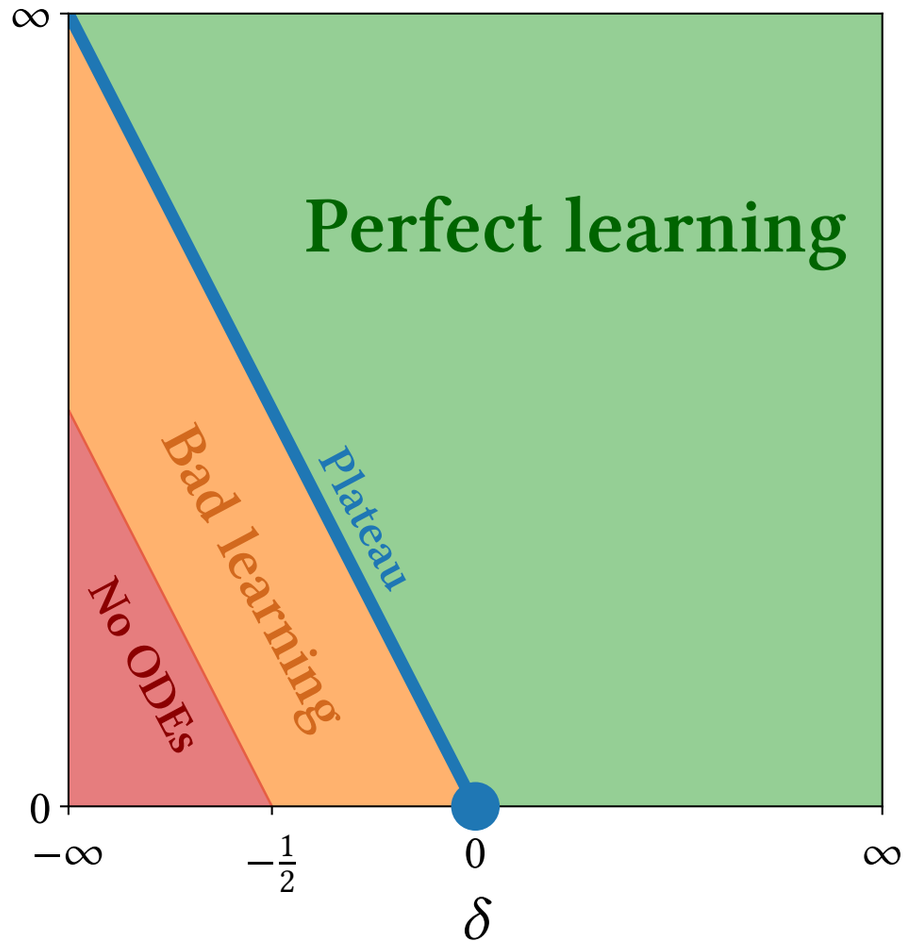


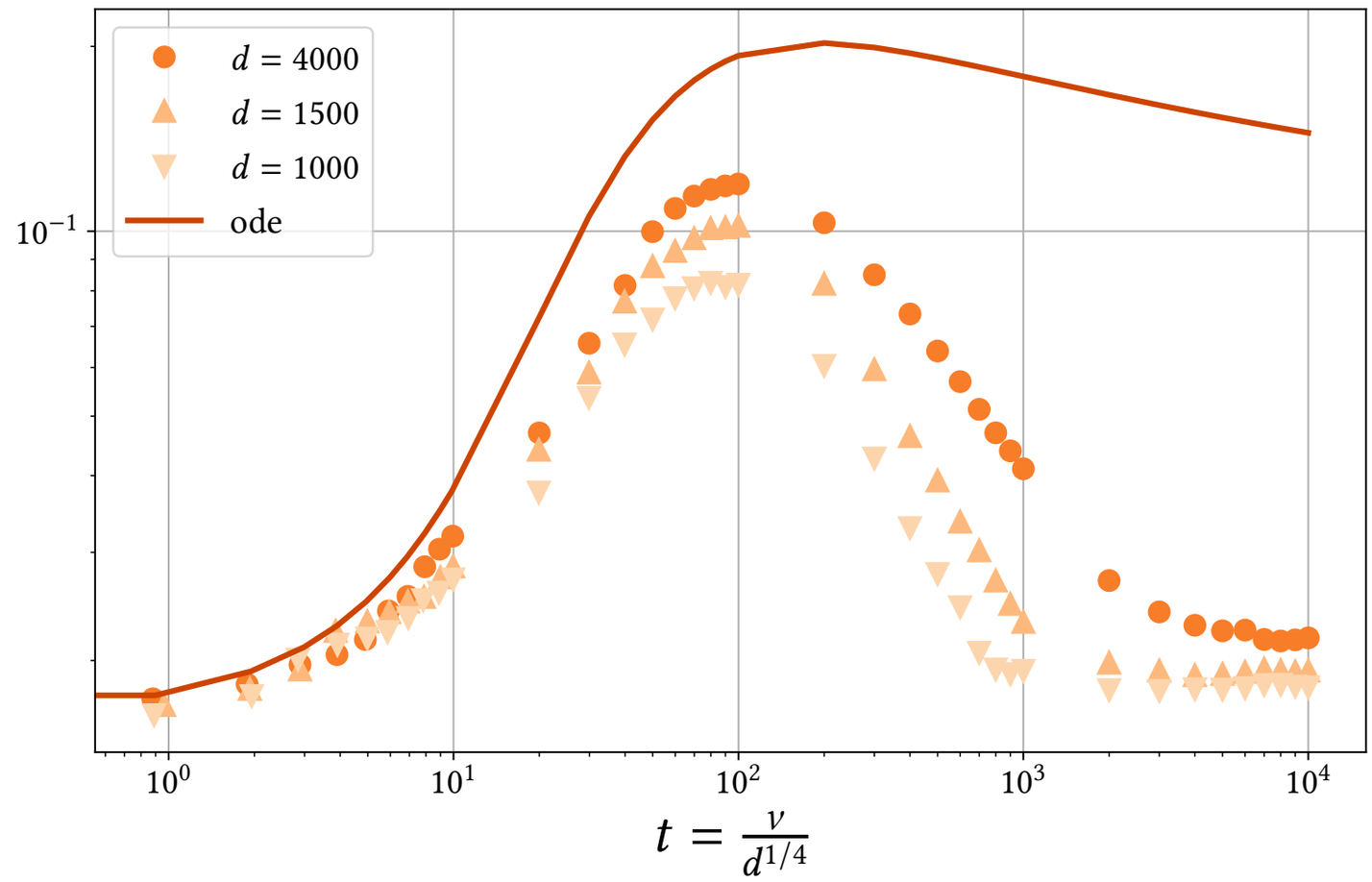
$$\mathscr{R}_\infty - \frac{\Delta}{2} \sim d^{-\delta}$$

Orange region: $0 > \kappa + \delta > -1/2$

$$\delta t = 1/d^{1+2(\kappa+\delta)}$$

$$\kappa = 0 \qquad \delta = -3/8$$



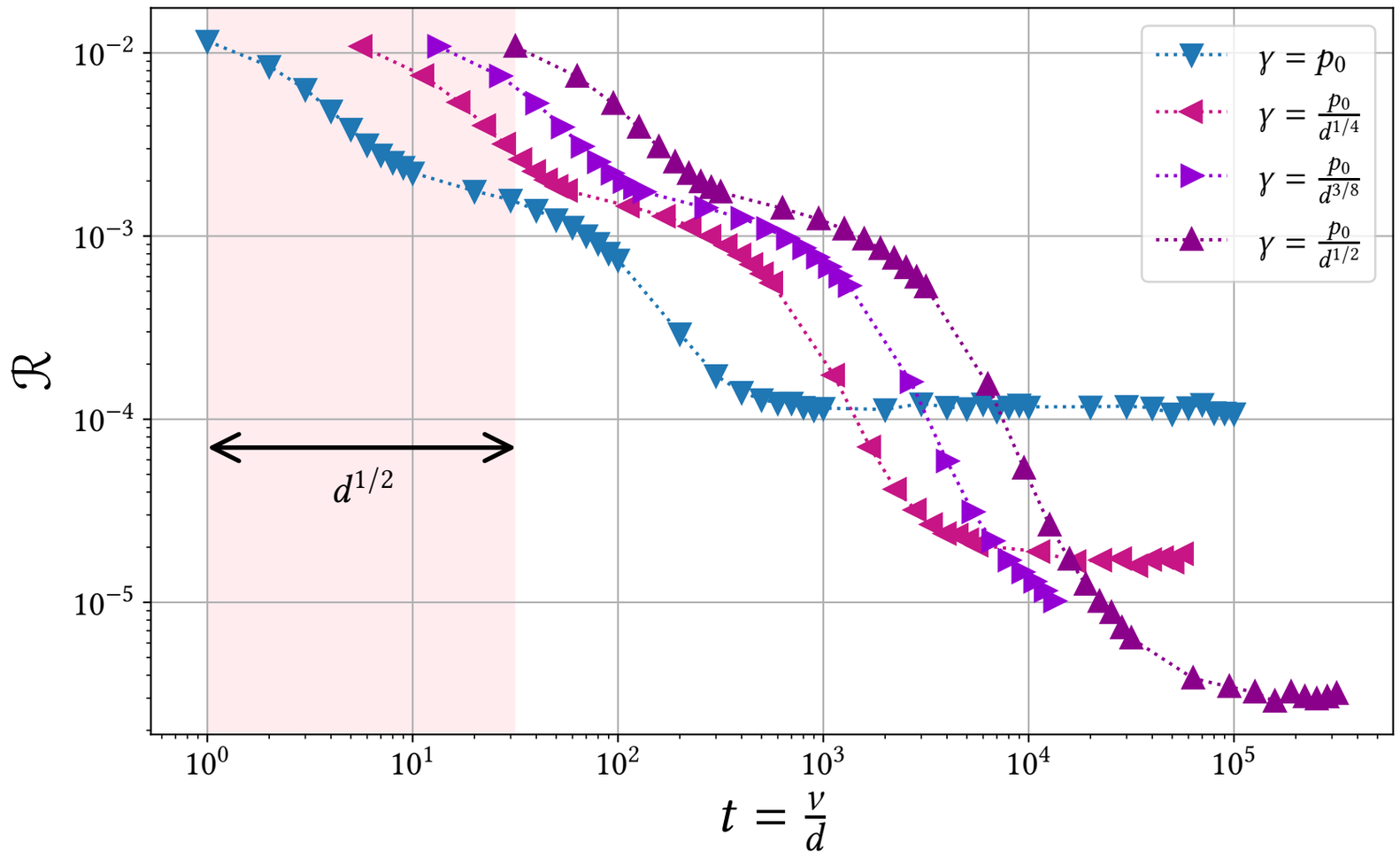


γ growing with d (weird!)

Strong finite size effects: $\mathbb{E}\,||\Omega^{\nu} - \bar{\Omega}(\nu\delta t)||_{\infty} \sim \dfrac{\log d}{d^{\frac{1}{2}+\delta+\kappa}}$

Fundamental trade-off

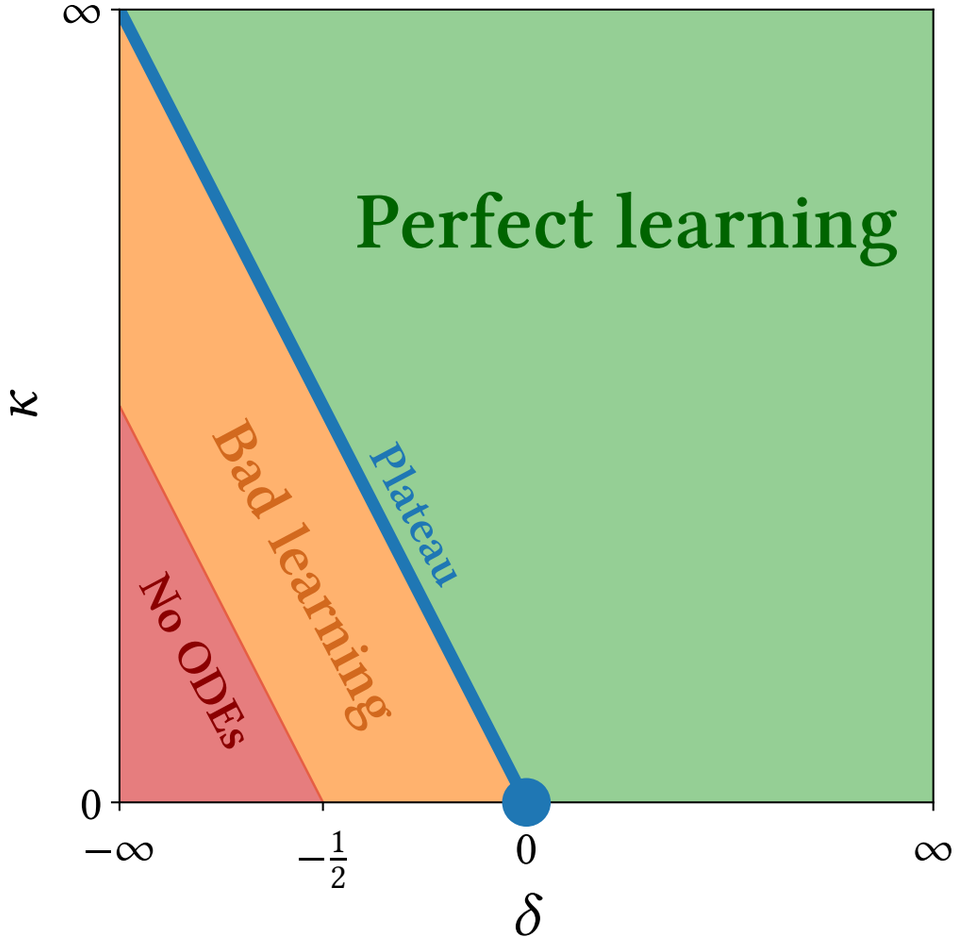
$$\kappa = 0 \qquad d = 1000 \qquad \Delta = 10^{-3}$$

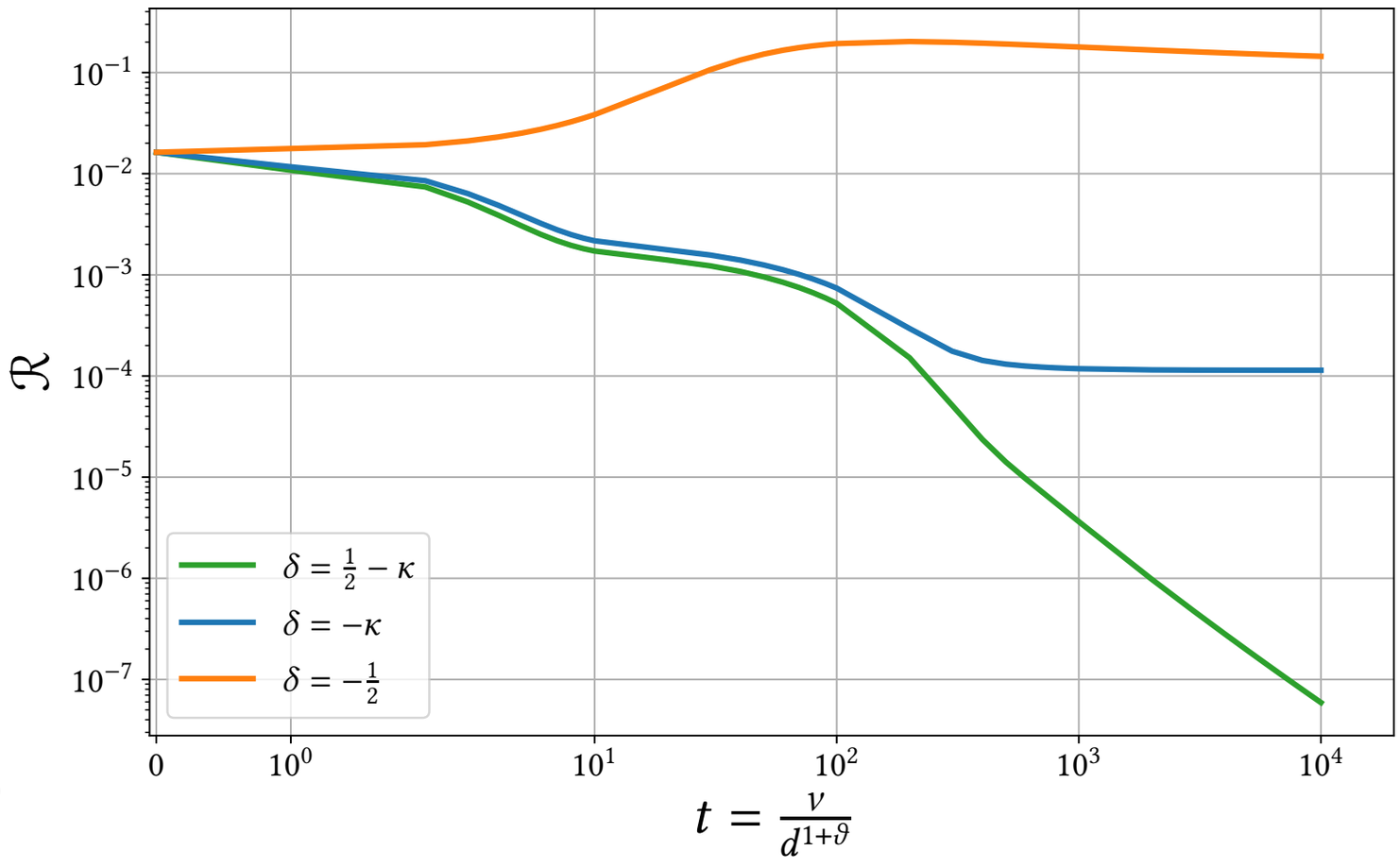


$$n \sim d^{1+\delta}$$

Lowering γ by a factor $d^{-\delta}$ requires d^{δ} more samples

Summary





$\theta = 0$ $\theta = \kappa + \delta$ $\theta = 2(\kappa + \delta)$

Conclusion

- Exact and deterministic theory for one-pass SGD for 2-layer NN in the high-dimensional limit

Conclusion

- Exact and deterministic theory for one-pass SGD for 2-layer NN in the high-dimensional limit
- Wide hidden-layer helps achieving perfect learning

Conclusion

- Exact and deterministic theory for one-pass SGD for 2-layer NN in the high-dimensional limit
- Wide hidden-layer helps achieving perfect learning
- Full phase diagram describing cross-over between different regimes

Thank you!