Statistical physics view of a *“theory of machine learning”*

Bruno Loureiro
@ IdePHICS, EPFL

brloureiro@gmail.com

ACDL 2022, 26.08.2022

Statistical physics view of a

"theory of machine learning"

Theory of machine learning?

Theory can mean different things.

fridge
Theory of ~~machine learning~~?

Theory can mean different things.

Theory of ~~machine learning~~ fridge?

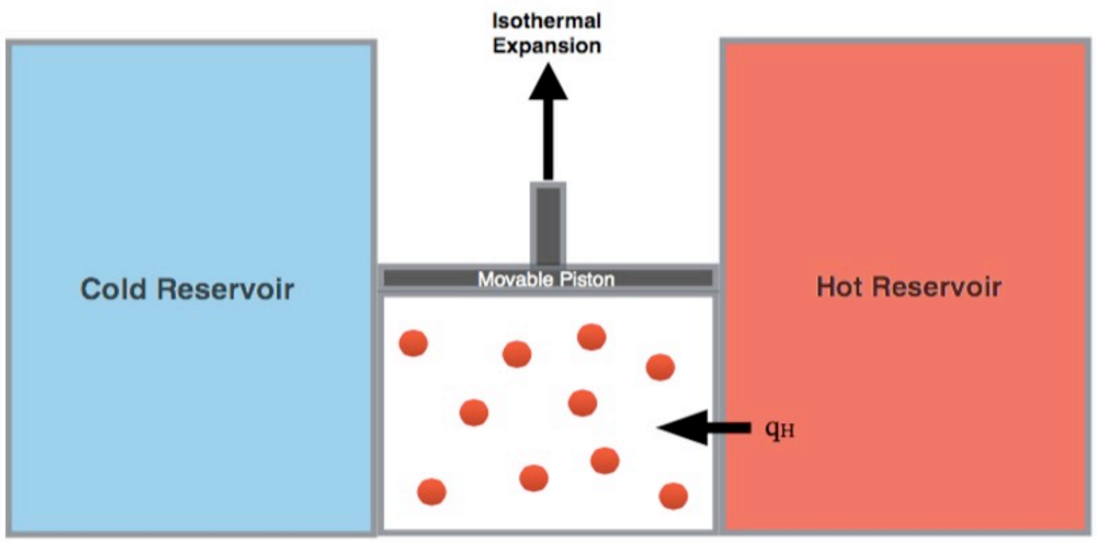
Theory can mean different things.

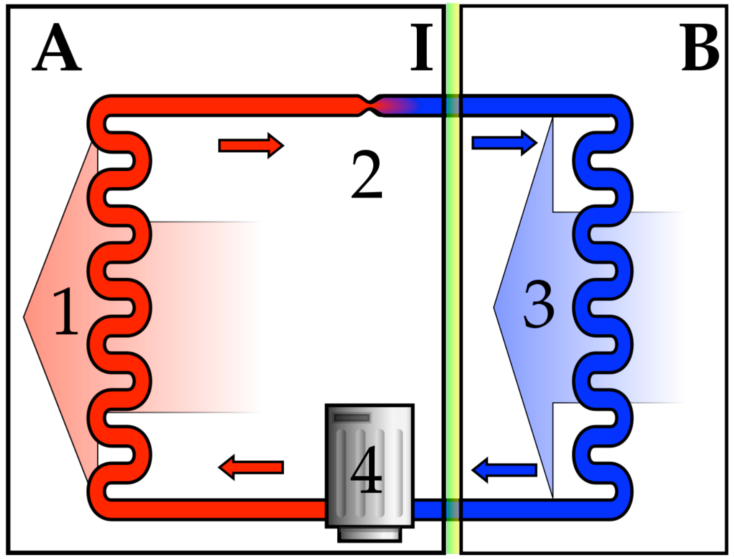
Physics

Fundamental laws that govern behaviour of the fridge

Engineering

How do I build a good fridge?



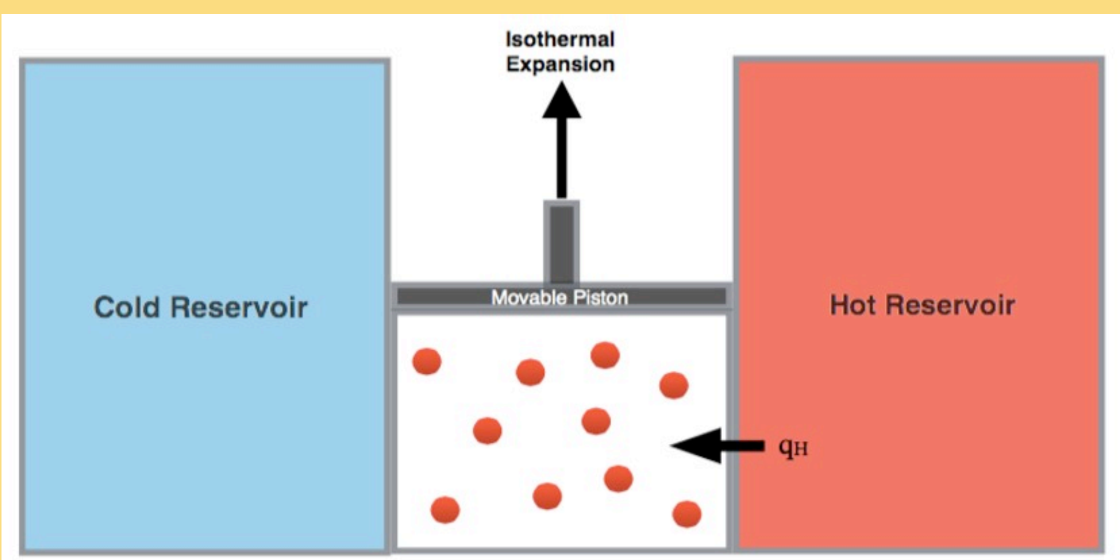


fridge
Theory of ~~machine learning~~?

Theory can mean different things.

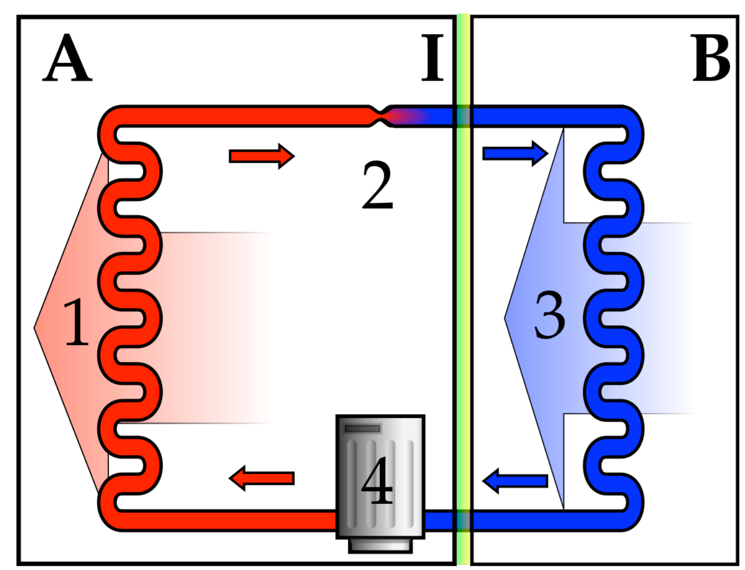
Physics

Fundamental laws that govern behaviour of the fridge



Engineering

How do I build a good fridge?

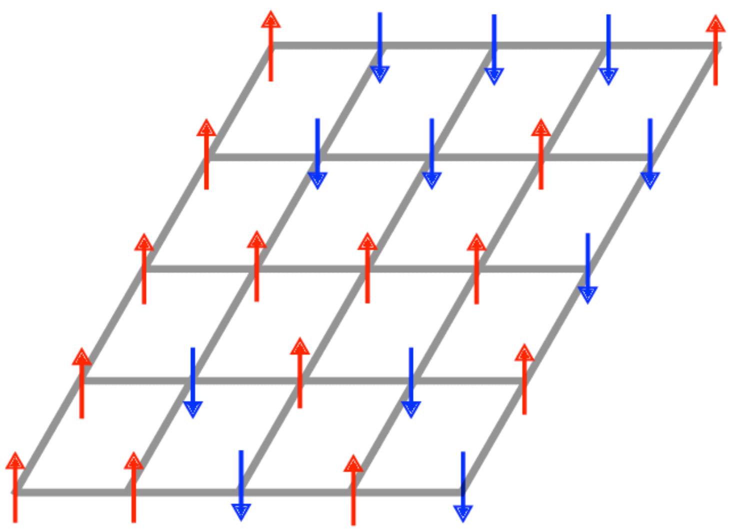


Theory of magnetism

a.k.a. the *Ising Model*

$$H_{J,h}(s) = -J \sum_{(ij)\in E} s_i s_j + h \sum_{i \in V} s_i$$

$$\mu_\beta(s) = \frac{1}{\mathcal{Z}_{\beta,J,h}} e^{-\beta H_{J,h}(s)} \quad s \in \{-1, +1\}^N$$

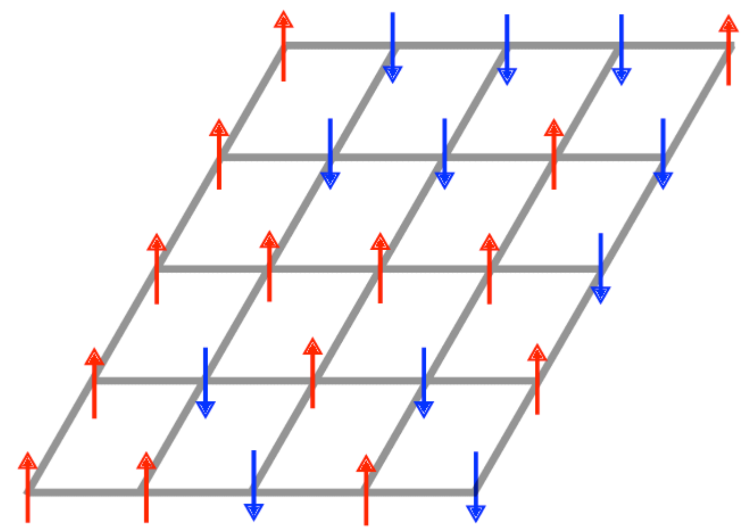


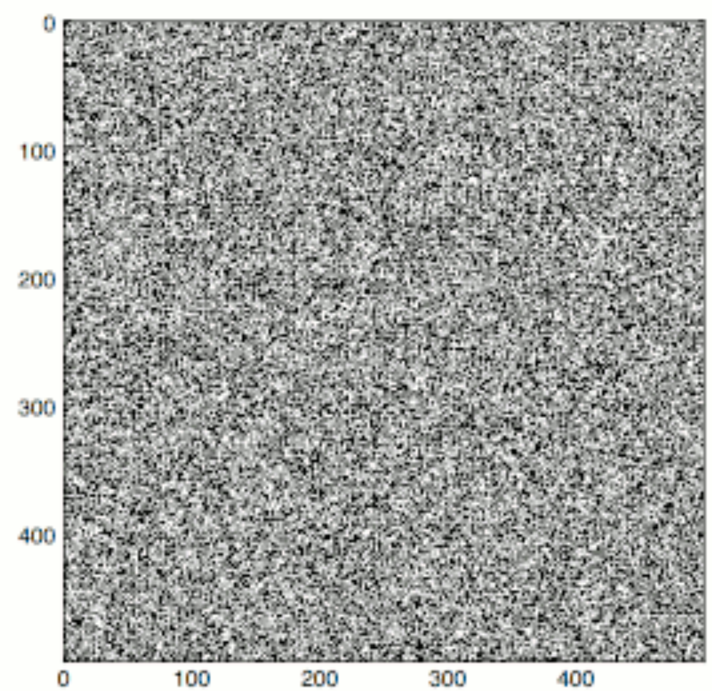
Theory of magnetism

a.k.a. the *Ising Model*

$$H_{J,h}(s) = -J \sum_{(ij)\in E} s_i s_j + h \sum_{i\in V} s_i$$

$$\mu_\beta(s) = \frac{1}{\mathcal{Z}_{\beta,J,h}} e^{-\beta H_{J,h}(s)} \quad s \in \{-1,+1\}^N$$





$h = 0$

Order parameter: $m = \frac{1}{|V|} \sum_{i\in V} s_i$

$\beta J = 0$ βJ_c $\beta J = \infty$

βJ

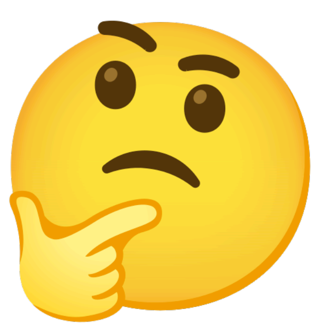
$m = 0$ $|m| > 0$

Theory of machine learning?

Theory can mean different things.

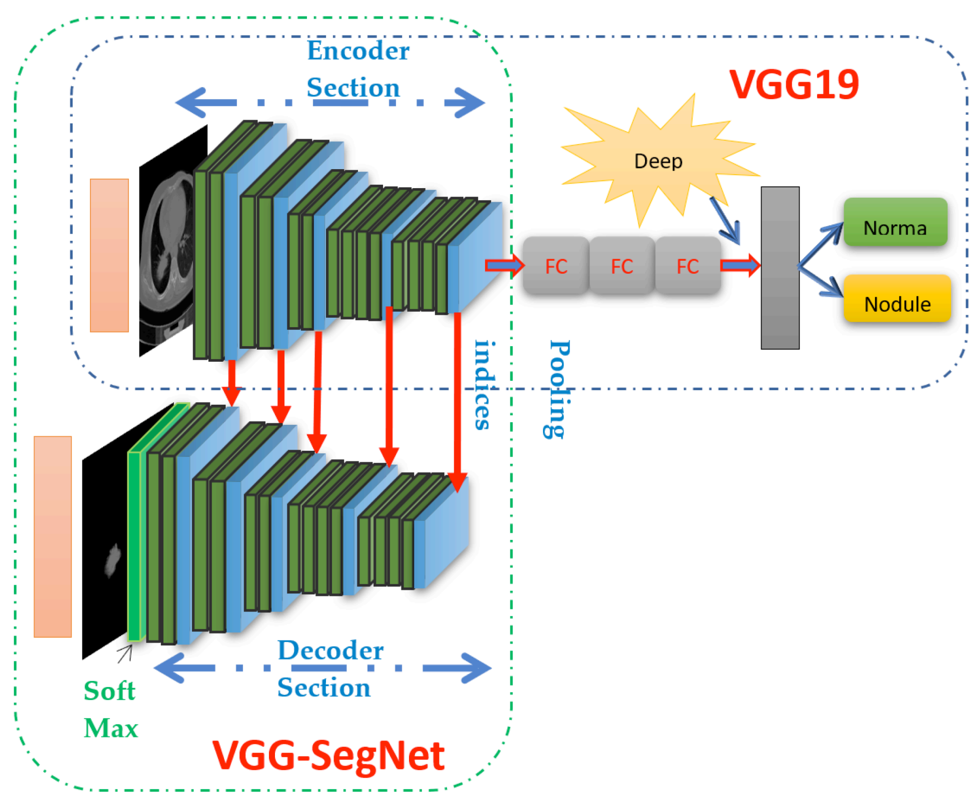
<u>Theory</u>

Fundamental principles that govern learning



<u>Engineering</u>

How do I build a state-of-the-art neural net?



Supervised learning

Data $\{\mathbf{x}^{\mu}, y^{\mu}\}_{\mu=1}^{n}$, sampled independently from $p_{x,y}$

Supervised learning

Data $\{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^n$, sampled independently from $p_{x,y}$

Typically, learn $\hat{y} = f_\theta(\mathbf{x})$ by minimising empirical risk

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{\mu=1}^{n} \ell \left(y^\mu, f_\theta(\mathbf{x}^\mu) \right) + r(\theta) \right]$$

Supervised learning

Data $\{\mathbf{x}^{\mu}, y^{\mu}\}_{\mu=1}^{n}$, sampled independently from $p_{x,y}$

Typically, learn $\hat{y} = f_{\theta}(\mathbf{x})$ by minimising empirical risk

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{\mu=1}^{n} \ell\left(y^{\mu}, f_{\theta}(\mathbf{x}^{\mu})\right) + r(\theta) \right]$$

Goal: Characterise error of predictor

$$\mathscr{R}(\hat{\theta}) = \mathbb{E}_{x,y}\left[\ell(y, f_{\hat{\theta}}(x))\right] \qquad \hat{\mathscr{R}}_n(\hat{\theta}) = \frac{1}{n} \sum_{\mu=1}^{n} \ell(y^{\mu}, f_{\hat{\theta}}(\mathbf{x}^{\mu}))$$

Supervised learning

Let $(x^\mu, y^\mu) \in \mathbb{R}^d \times \{-1,1\}$, $\mu = 1,\cdots,n$ denote the training data and , and let $\hat{y} = f_\theta(x)$ be a predictor belonging to some function class $\mathscr{H}$

Supervised learning

Let $(x^\mu, y^\mu) \in \mathbb{R}^d \times \{-1,1\}$, $\mu = 1,\cdots,n$ denote the training data and , and let $\hat{y} = f_\theta(x)$ be a predictor belonging to some function class $\mathscr{H}$

Theorem (informal):

$$\sup_{f_\theta \in \mathscr{H}} \mathscr{R}(f_\theta) - \hat{\mathscr{R}}_n(f_\theta) \leq \sqrt{\frac{d_{\mathrm{VC}}}{n}}$$

VC dimension: $d_{\mathrm{VC}} \propto$ number of parameters

Agnostic bounds: as few assumptions as possible on data x^μ and labels y^μ

Many questions, few answers

Despite the amazing progress on the engineering side, theory falls short.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

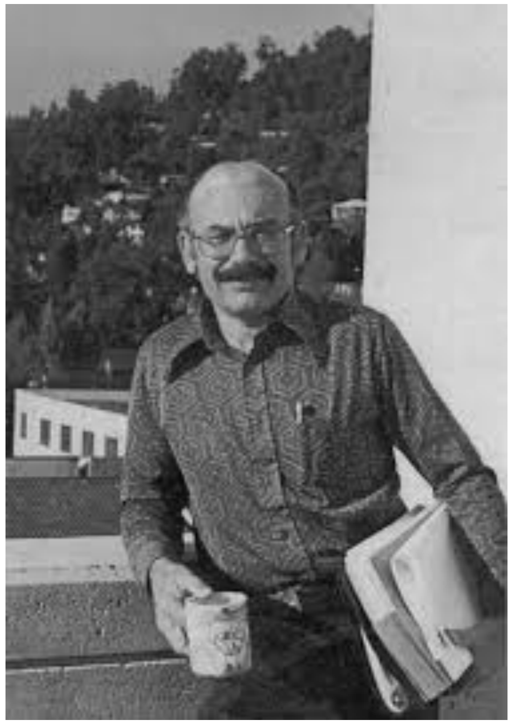
- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?

Many questions, few answers

Despite the amazing progress on the engineering side, theory falls short.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?



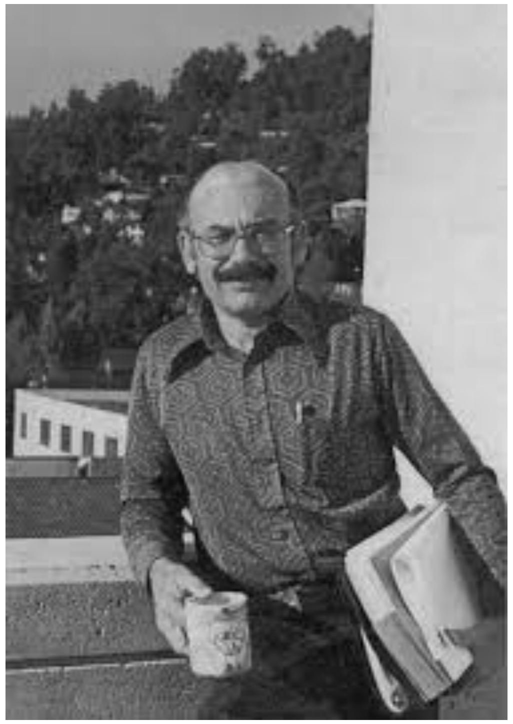
"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

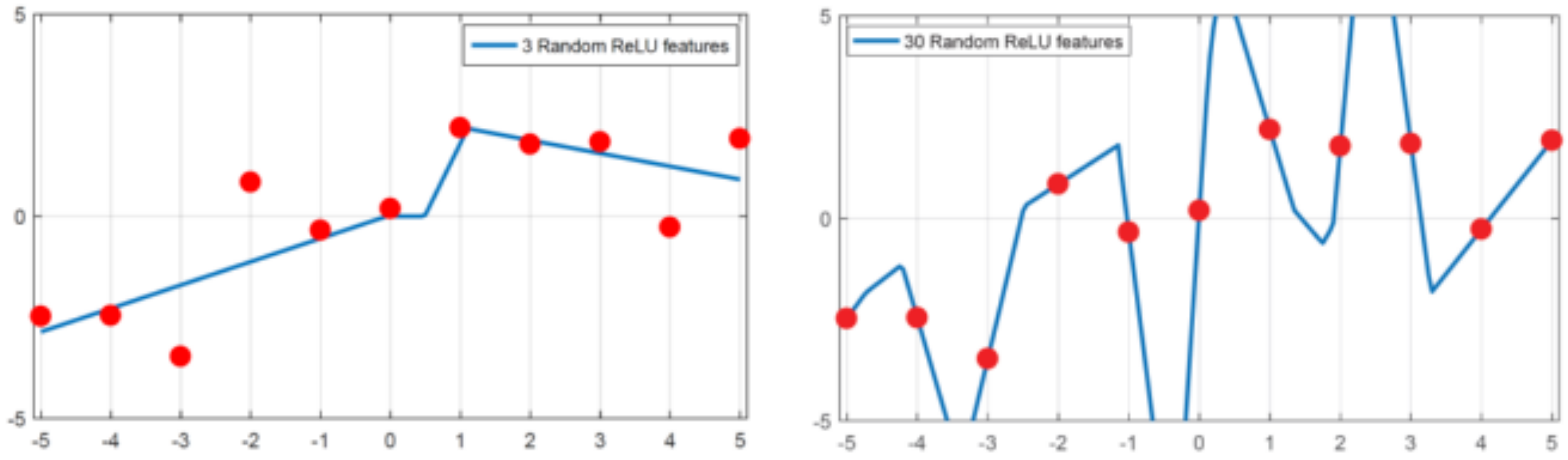
- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?



"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

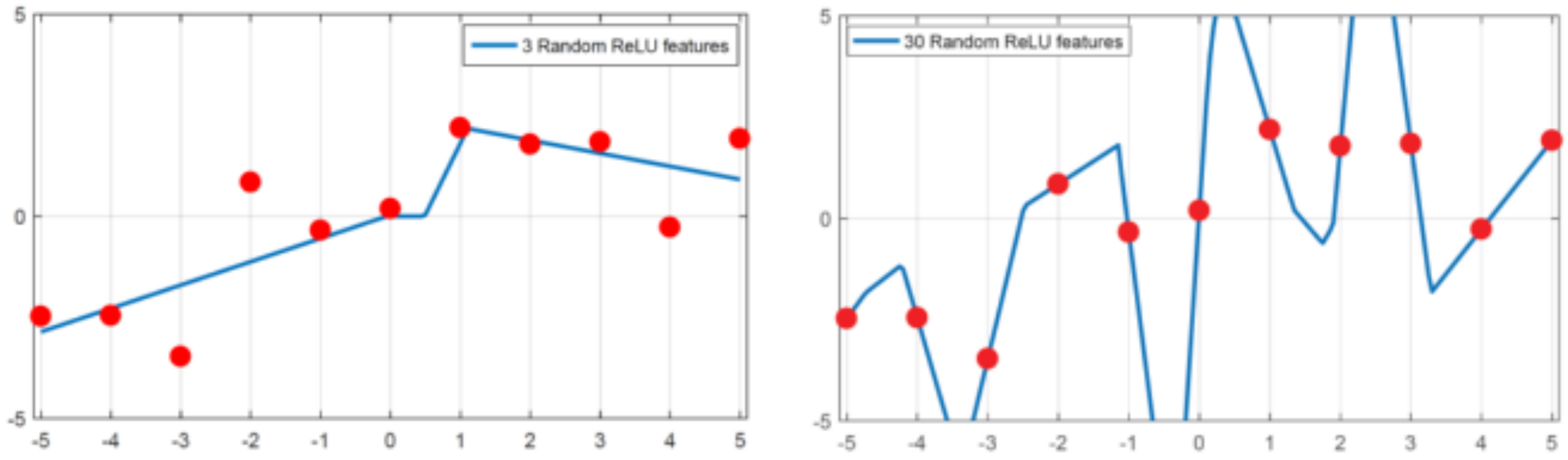
Overfitting

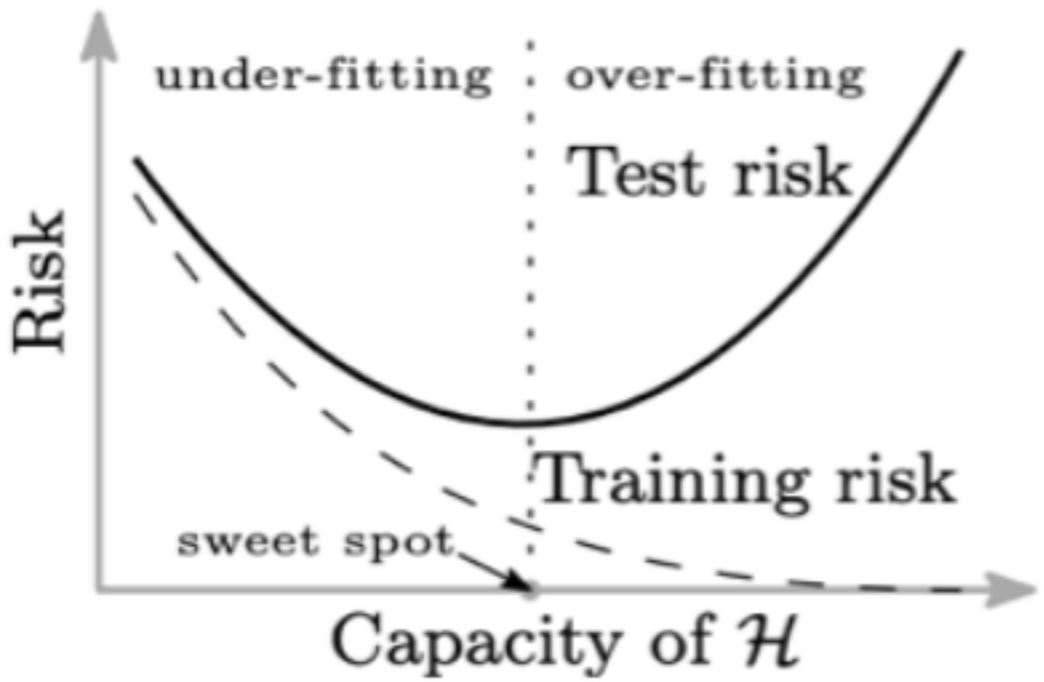
Figure from [Belkin 21']



Overfitting

Figure from [Belkin 21']





Overfitting

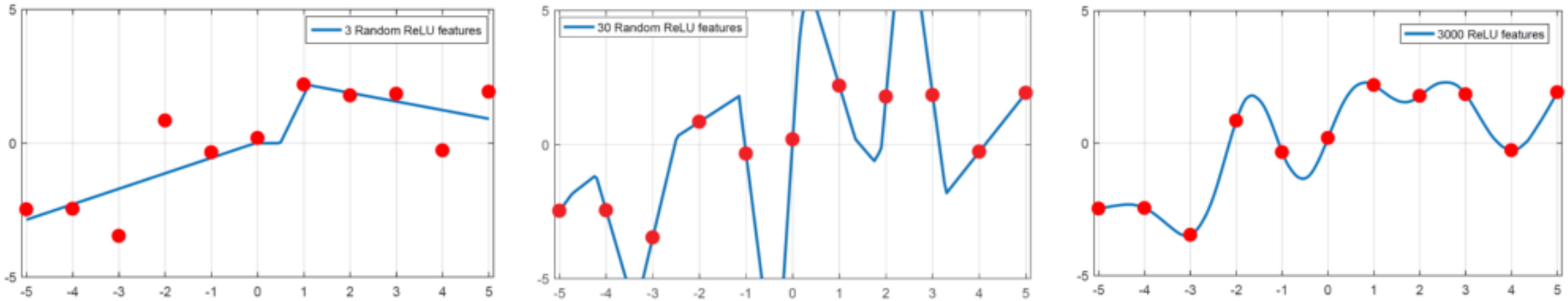
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

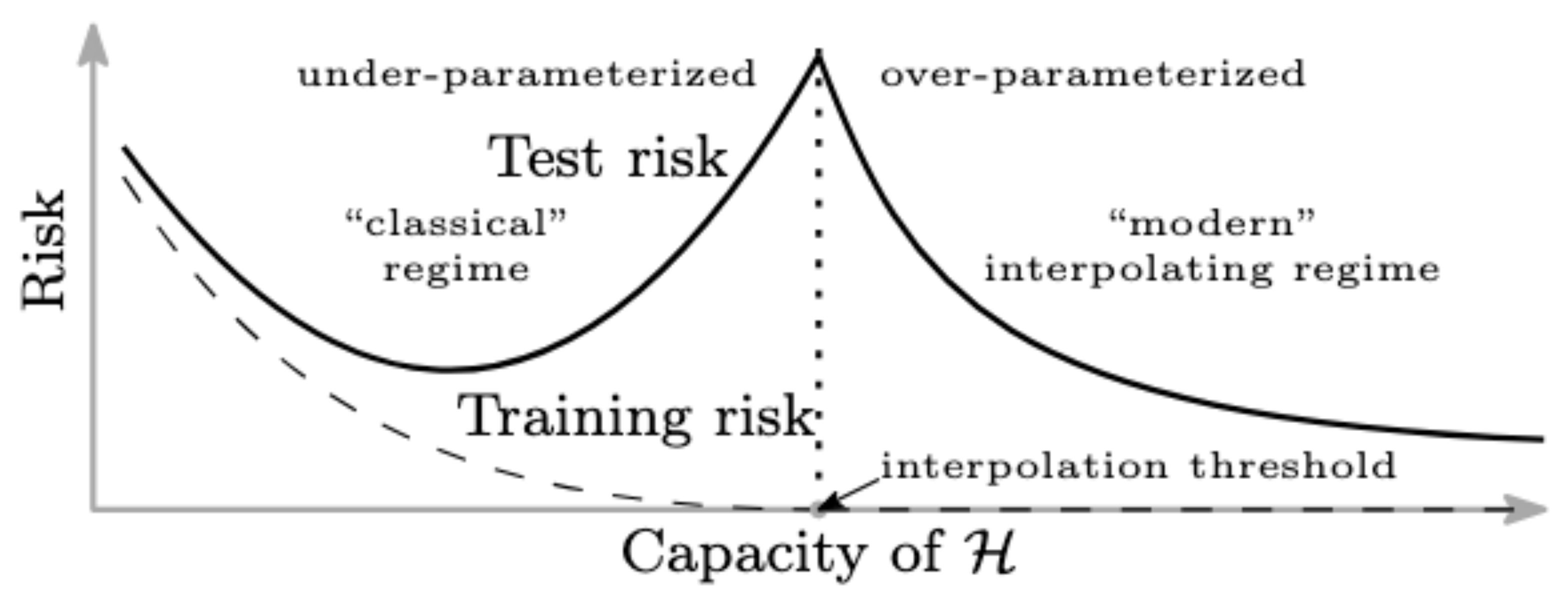
Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

From *"Language Models are Few-Shot Learners"*, Brown et al 2020

Overfitting

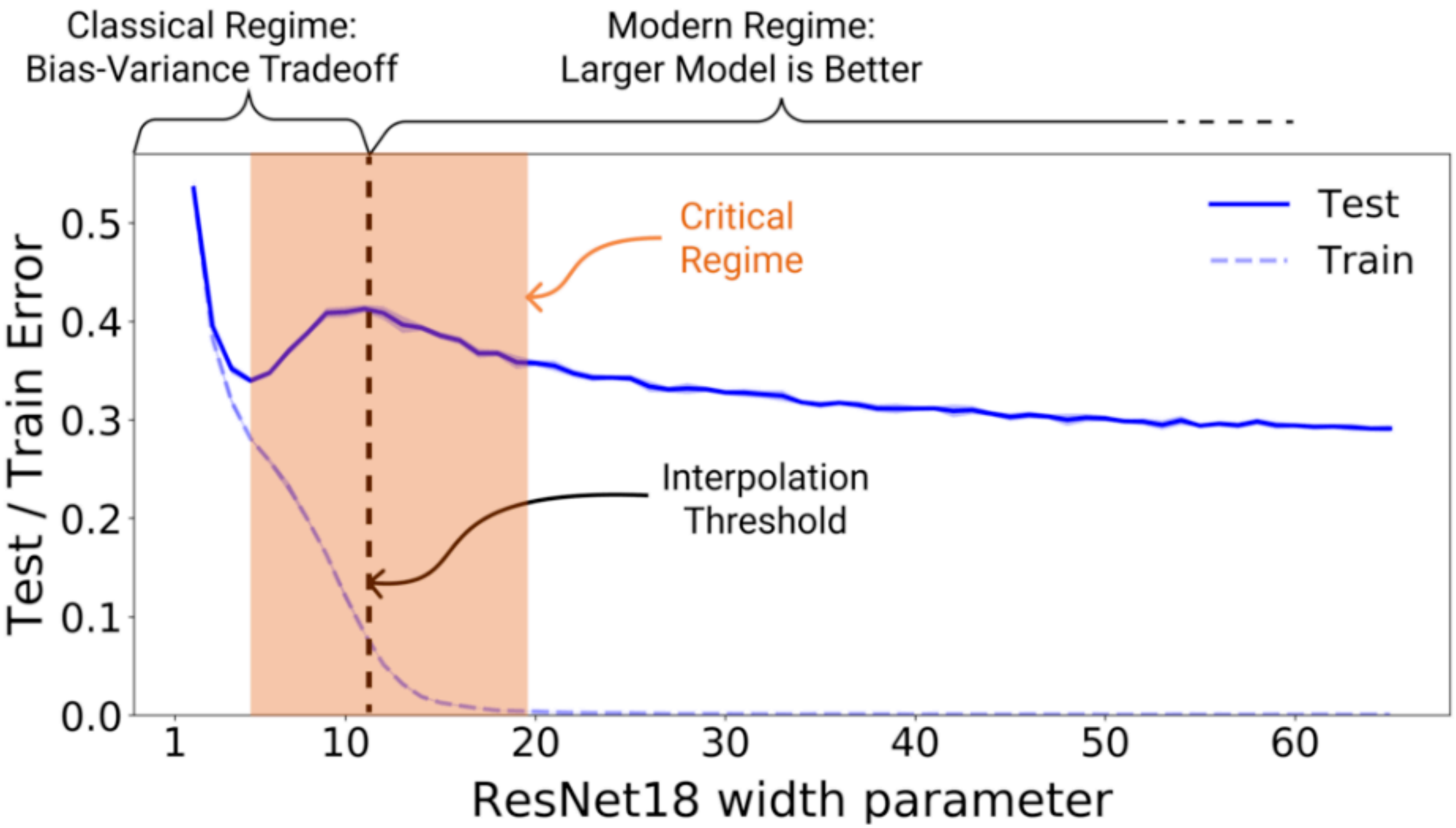
Figure from [Belkin 21']





Overfitting

Figure from [Nakkiran et al 19']

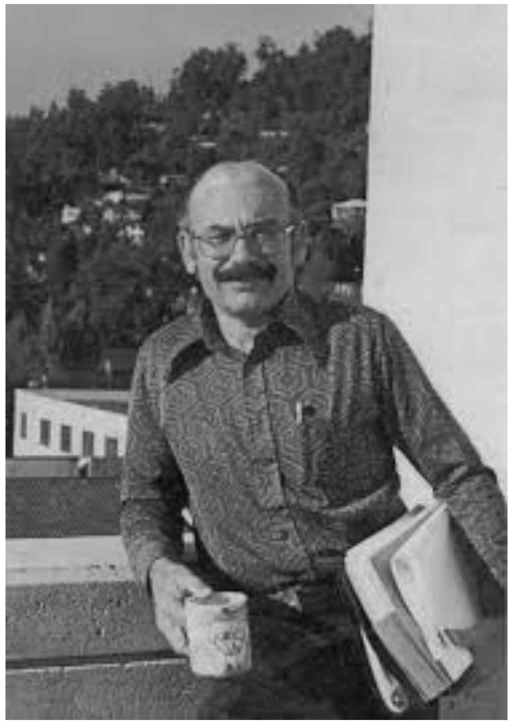


Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.

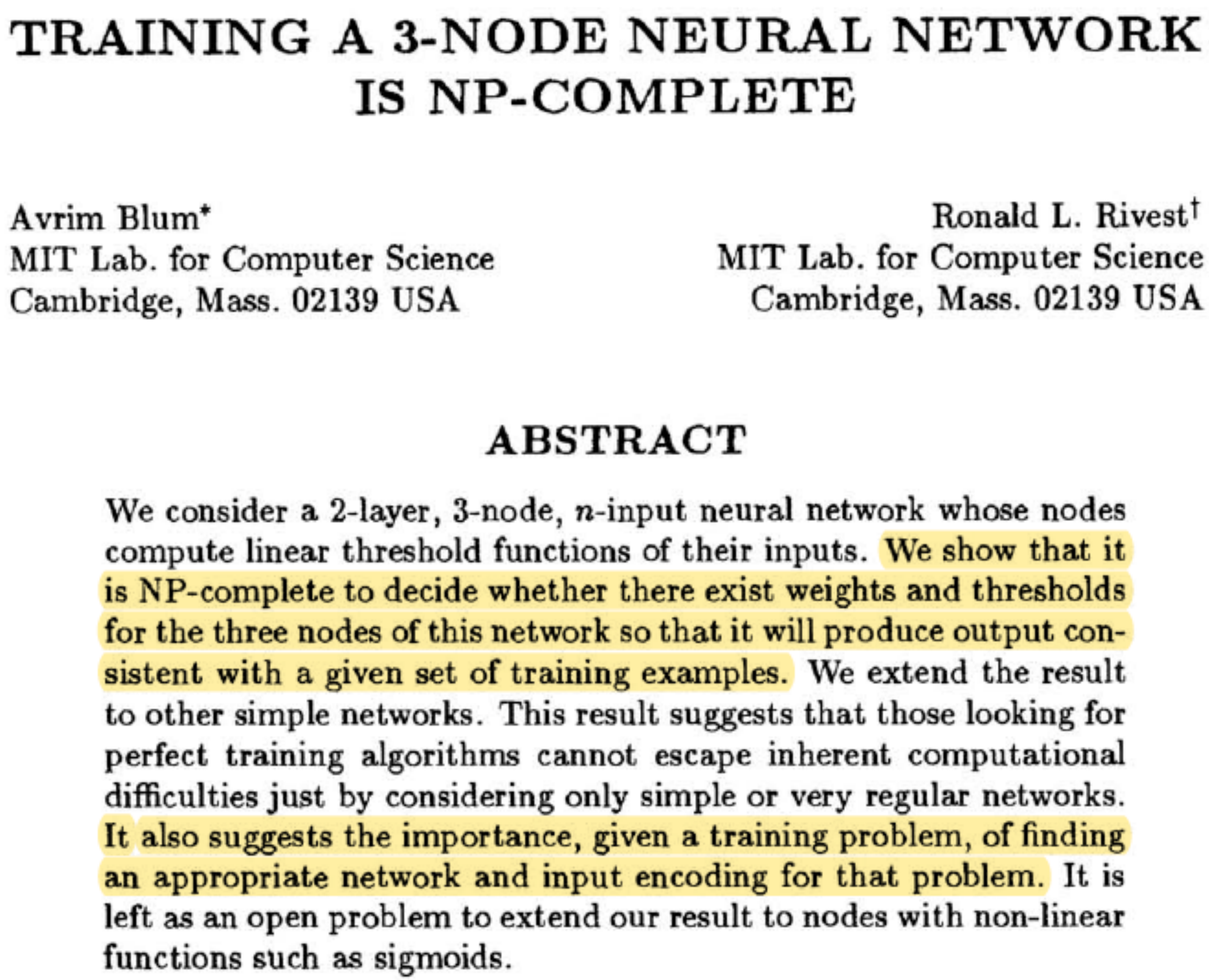
For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?



"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

Worst case can be hard



TRAINING A 3-NODE NEURAL NETWORK IS NP-COMPLETE

Avrim Blum*
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

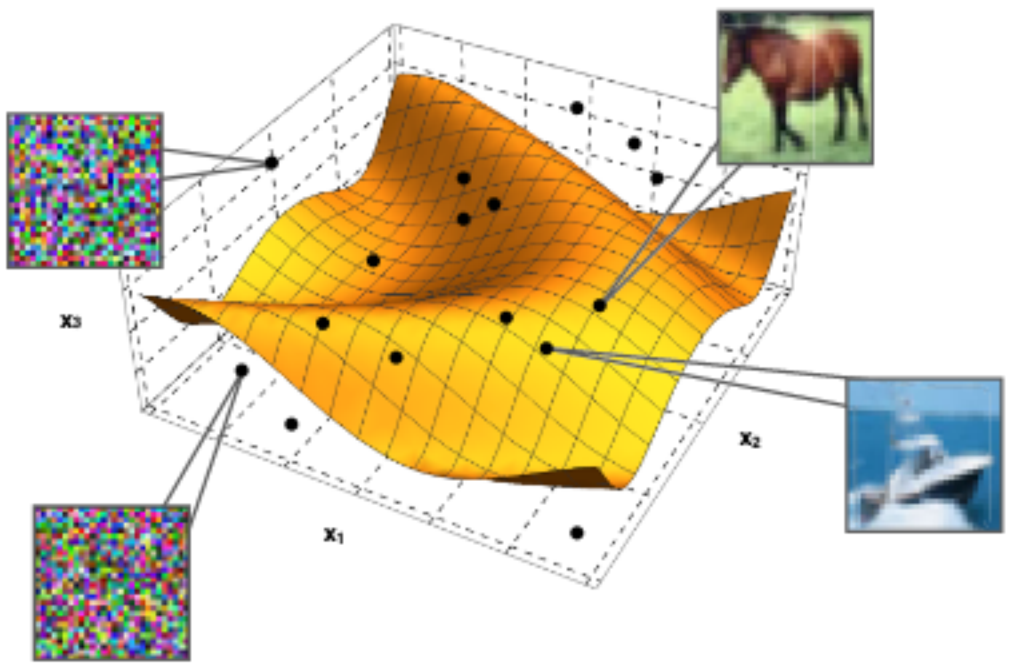
Ronald L. Rivest†
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

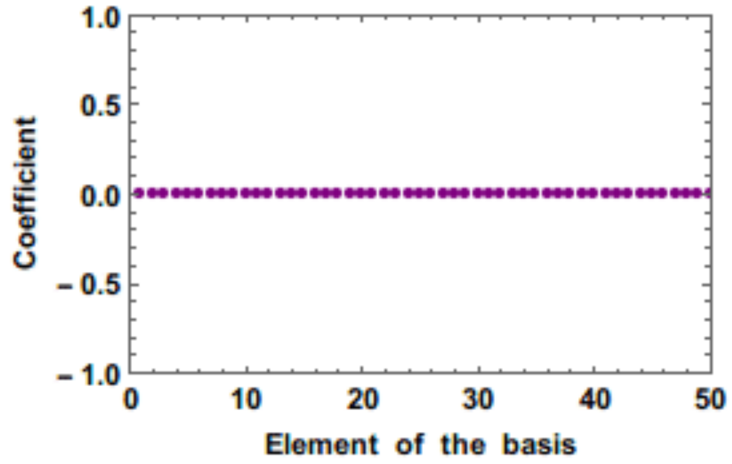
ABSTRACT

We consider a 2-layer, 3-node, n-input neural network whose nodes compute linear threshold functions of their inputs. We show that it is NP-complete to decide whether there exist weights and thresholds for the three nodes of this network so that it will produce output consistent with a given set of training examples. We extend the result to other simple networks. This result suggests that those looking for perfect training algorithms cannot escape inherent computational difficulties just by considering only simple or very regular networks. It also suggests the importance, given a training problem, of finding an appropriate network and input encoding for that problem. It is left as an open problem to extend our result to nodes with non-linear functions such as sigmoids.

Effective dimension?

How many features / samples needed to correctly learn?



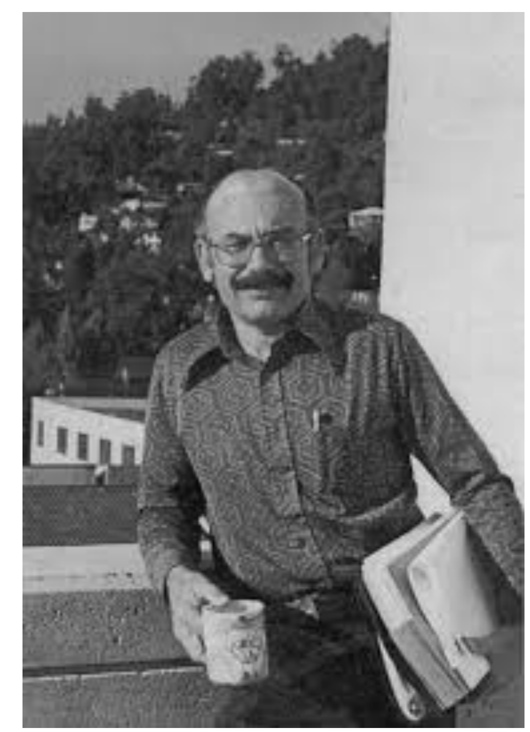


Many questions, few answers

Despite the amazing progress on the engineering side,
theory falls short.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?



"Reflections after refereeing papers for NIPS",
Leo Breiman, **1995**

Bad minima exist

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

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.

Breiman's suggestions

"Reflections after refereeing papers for NIPS", Leo Breiman, **1995**

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

Breiman's suggestions

"Reflections after refereeing papers for NIPS", Leo Breiman, **1995**

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

flexible maths

simple, solvable toy models

experiments

Smells of... physics.

Neural nets, before it was cool

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

Abstract. We calculate the number, $p = \alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction f_{min} of wrong bits. We find a critical line, $\alpha_c(K)$ with $\alpha_c(0) = 2$. The minimum fraction of wrong bits vanishes for $\alpha < \alpha_c(K)$ and increases from zero for $\alpha > \alpha_c(K)$. The calculations are done using a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is locally stable in a finite region of the K,α plane including the line, $\alpha_c(K)$ but there is a line above which the solution becomes unstable and replica symmetry must be broken.

Neural nets, before it was cool

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

hat an optimal
h the condition
er K. For each
d a critical line,
$\alpha < \alpha_c(K)$ and
e-point method
ymmetric. This
line, $\alpha_c(K)$ but
nmetry must be

The space of interactions in neural network models

E Gardner

Department of Physics, Edinburgh University, Mayfield Road, Edinburgh EH9 3JK, UK

Received 13 May 1987, in final form 27 July 1987

Abstract. The typical fraction of the space of interactions between each pair of N Ising spins which solve the problem of storing a given set of p random patterns as N-bit spin configurations is considered. The volume is calculated explicitly as a function of the storage ratio, $\alpha = p/N$, of the value $\kappa(>0)$ of the product of the spin and the magnetic field at each site and of the magnetisation, m. Here m may vary between 0 (no correlation) and 1 (completely correlated). The capacity increases with the correlation between patterns from $\alpha = 2$ for correlated patterns with $\kappa = 0$ and tends to infinity as m tends to 1. The calculations use a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is shown to be locally stable. A local iterative learning algorithm for updating the interactions is given which will converge to a solution of given κ provided such solutions exist.

Neural nets, before it was cool

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430
(Received 9 February 1990)

Learning from examples by a perceptron with binary synaptic parameters is studied. The examples are given by a reference (teacher) perceptron. It is shown that as the number of examples increases, the network undergoes a first-order transition, where it freezes into the state of the reference perceptron. When the transition point is approached from below, the generalization error reaches a minimal positive value, while above that point the error is constantly zero. The transition is found to occur at $\alpha_{GD} = 1.245$ examples per coupling.

configurations is considered. The volume is calculated explicitly as a function of the storage ratio, $\alpha = p/N$, of the value $\kappa(>0)$ of the product of the spin and the magnetic field at each site and of the magnetisation, m. Here m may vary between 0 (no correlation) and 1 (completely correlated). The capacity increases with the correlation between patterns from $\alpha = 2$ for correlated patterns with $\kappa = 0$ and tends to infinity as m tends to 1. The calculations use a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is shown to be locally stable. A local iterative learning algorithm for updating the interactions is given which will converge to a solution of given κ provided such solutions exist.

Neural nets, before it was cool

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430
(Received 9 February 1990)

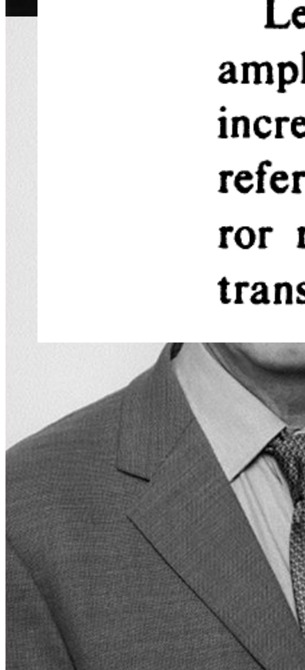
Learning fr
amples are giv
increases, the
reference perc
ror reaches a
transition is fc

Learning from Examples in Large Neural Networks

H. Sompolinsky[a] and N. Tishby
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

H. S. Seung
Department of Physics, Harvard University, Cambridge, Massachusetts 02138
(Received 29 May 1990)

A statistical mechanical theory of learning from examples in layered networks at finite temperature is studied. When the training error is a smooth function of continuously varying weights the generalization error falls off asymptotically as the inverse number of examples. By analytical and numerical studies of single-layer perceptrons we show that when the weights are discrete the generalization error can exhibit a discontinuous transition to perfect generalization. For intermediate sizes of the example set, the state of perfect generalization coexists with a metastable spin-glass state.



Neural nets, before it was cool

The statistical mechanics of learning a rule

Timothy L. H. Watkin* and Albrecht Rau†
Department of Physics, University of Oxford, Oxford OX1 3NP, United Kingdom

Michael Biehl
Physikalisches Institut, Julius-Maximilians-Univer…

A summary is presented of the statistical mechanical the… rapidly advancing area which is closely related to other in… cists. By emphasizing the relationship between neural net… such as spin glasses, the authors show how learning theo… new, exact analytical techniques.

dels

…dinburgh, EH9 3JZ, UK
…vette, France

Learning fr… amples are gi… increases, the reference perc… ror reaches a transition is f…

Learn…

A1

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

Information storage and retrieval in synchronous neural networks

José F. Fontanari and R. Köberle
Phys. Rev. A **36**, 2475 – Published 1 September 1987

a discontinuous transition of perfect generalization c…

…ork of the per-…ters which ren-…s of attraction) …s and study the size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

Neural nets, before it was cool

The statistical mechanics of learning a rule

Timothy L. H. Watkin* and Albrecht Rau†

Department of Physics, University of Oxford, Oxford OX1 3NP, United Kingdom

UK

Michael Biehl

Physikalisches Institut, Julius-Maximilians-Univer...

A summary is presented of the statistical mechanical the... rapidly advancing area which is closely related to other in... cists. By emphasizing the relationship between neural net... such as spin glasses, the authors show how learning theo... new, exact analytical techniques.

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

Learning from examples by

Learning from Examples in Large N...

H. Sompolinsky[(a)] and N. T...

AT&T Bell Laboratories, Murray Hill, N...

Information storage and retrieval in synchronous neural networks

José F. Fontanari and R. Köberle

Phys. Rev. A **36**, 2475 – Published 1 September 1987

A stat... studied. ... error falls off asymptotically as the inverse number of examples. ... single-layer perceptrons we show that when the weights are disc... a discontinuous transition to perfect generalization. For interme... of perfect generalization coexists with a metastable spin-glass sta...

...vork of the per- ...ters which ren- ...s of attraction) ...s and study the size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

And they were not alone...

Yann LeCun is with Levent Sagun and 3 others.
August 30

Stéphane Mallat's tutorial at the "Statistical Physics and Machine Learning back Together" summer school in Cargese, Corsica.

There is a long history of theoretical physicists (particularly condensed matter physicists) bringing ideas and mathematical methods to machine learning, neural networks, probabilistic inference, SAT problems, etc.

In fact, the wave of interest in neural networks in the 1980s and early 1990s was in part caused by the connection between spin glasses and recurrent nets popularized by John Hopfield. While this caused some physicists to morph into neuroscientists and machine learners, most of them left the field when interest in neural networks wanted in the late 1990s

With the prevalence of deep learning and all the theoretical questions that surround it, physicists are coming back!

Many young physicists (and mathaticians) are now working on trying to explain why deep learning works so well. This summer school is for them.

We need to find ways to connect this emerging community with the ML/AI community. It's not easy because (1) papers submitted by physicists to ML conferences rarely make it because of a lack of qualified reviewers; (2) conference papers don't count in a physicist's CV.

http://cargese.krzakala.org

ÉCOLE DE PHYSIQUE DES HOUCHES

Disordered Systems and Biological Organization

The key idea

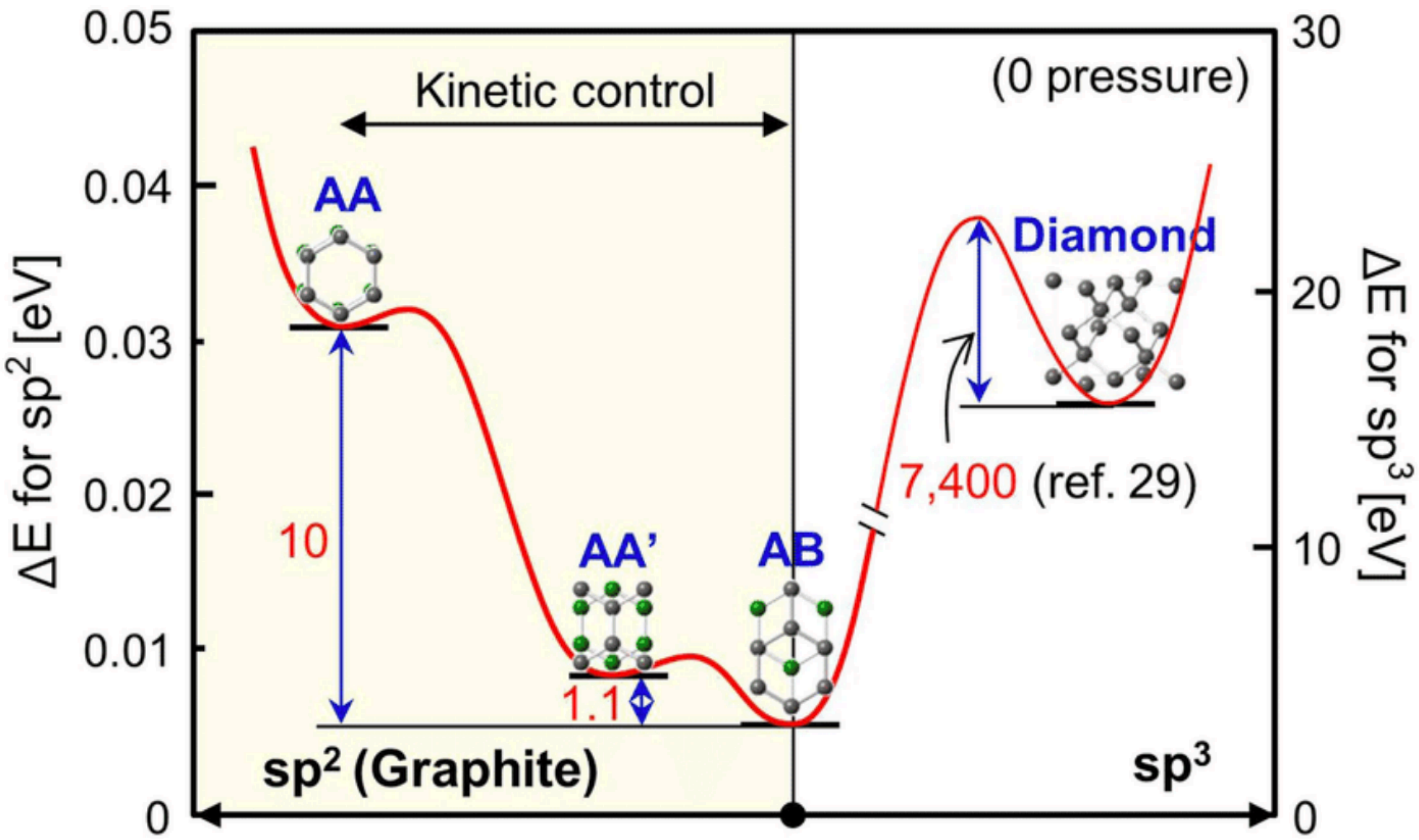
Idea: write this as Stat. Mech. problem

$$\mu_\beta(\theta) = \frac{\mathbf{1}}{\mathcal{Z}_\beta} \mathbf{e}^{-\beta \mathbf{H}(\theta)} \qquad H(\theta) = \frac{1}{n} \sum_{\mu=1}^{n} \ell\left(y^\mu, f_\theta(\mathbf{x}^\mu)\right) + r(\theta)$$

The key idea

Idea: write this as Stat. Mech. problem

$$\mu_\beta(\theta) = \frac{\mathbf{1}}{\mathcal{Z}_\beta} \mathbf{e}^{-\beta \mathbf{H}(\theta)} \qquad H(\theta) = \frac{1}{n} \sum_{\mu=1}^{n} \ell\left(y^\mu, f_\theta(\mathbf{x}^\mu)\right) + r(\theta)$$



Back to Breiman

"Reflections after refereeing papers for NIPS", Leo Breiman, **1995**

> For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:
>
> - Why don't heavily parameterized neural networks overfit the data?
> - What is the effective number of parameters?
> - Why doesn't backpropagation head for a poor local minima?

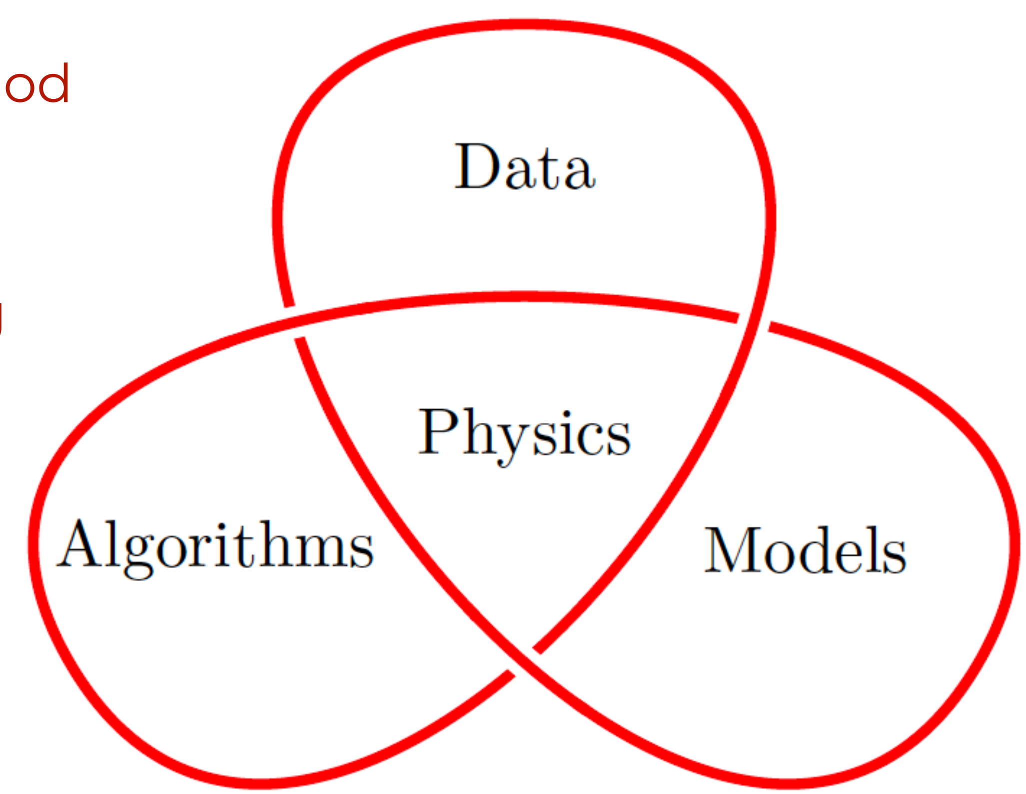
Teacher-student models

Cavity method

Gaussian universality

Approximate message passing

Replica method



Dynamical mean-field theory

Glassy landscapes (Kac-Rice)

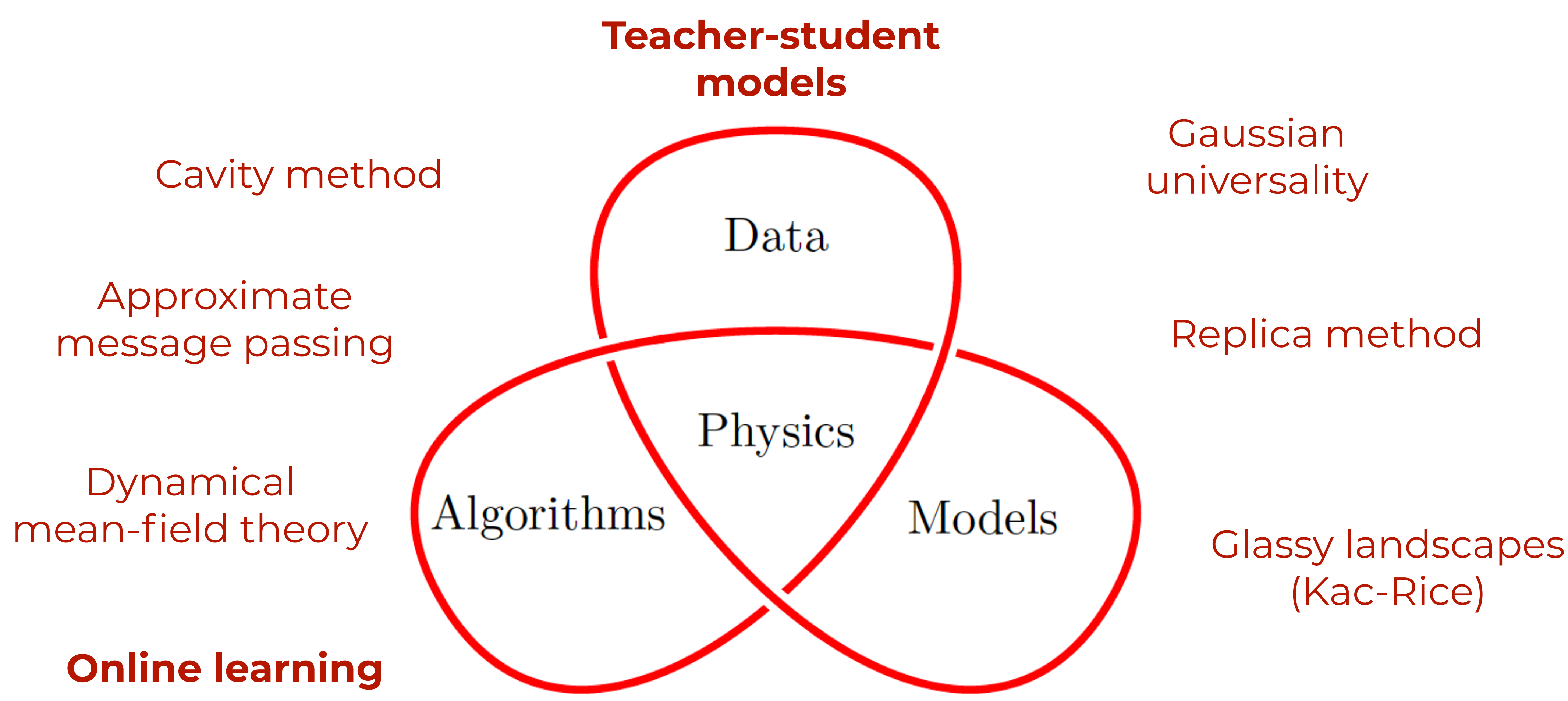
Online learning

Back to Breiman

"Reflections after refereeing papers for NIPS", Leo Breiman, **1995**

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?

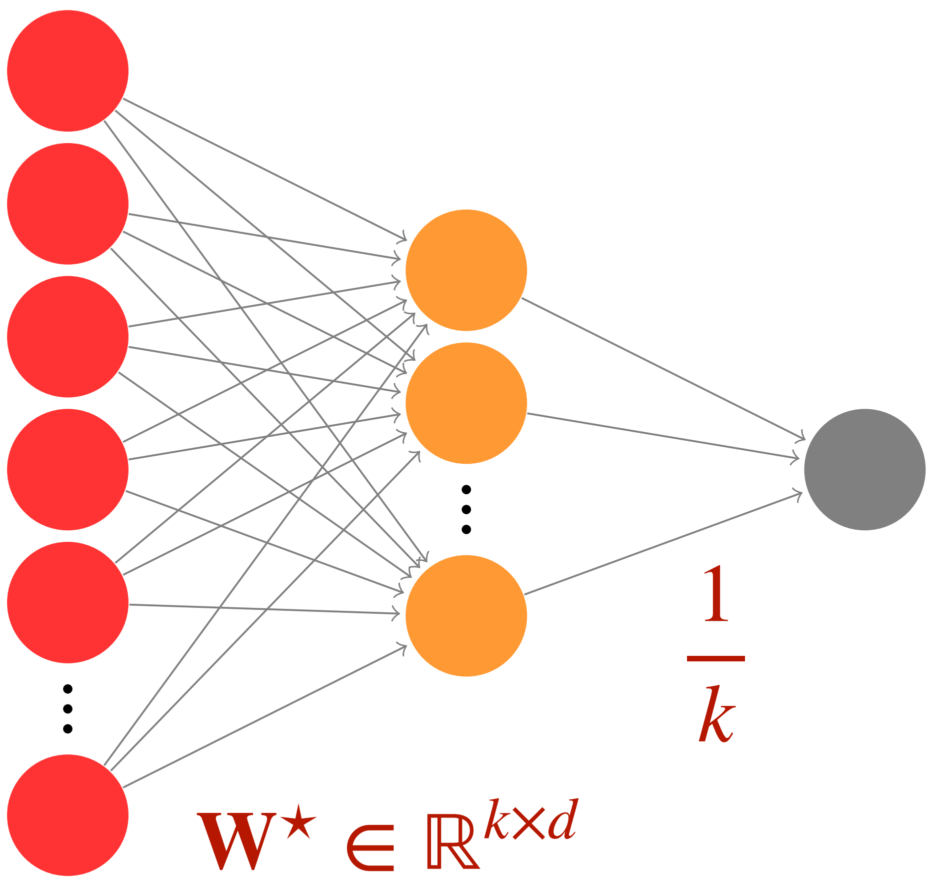


“Board” time

Teacher-student setting

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

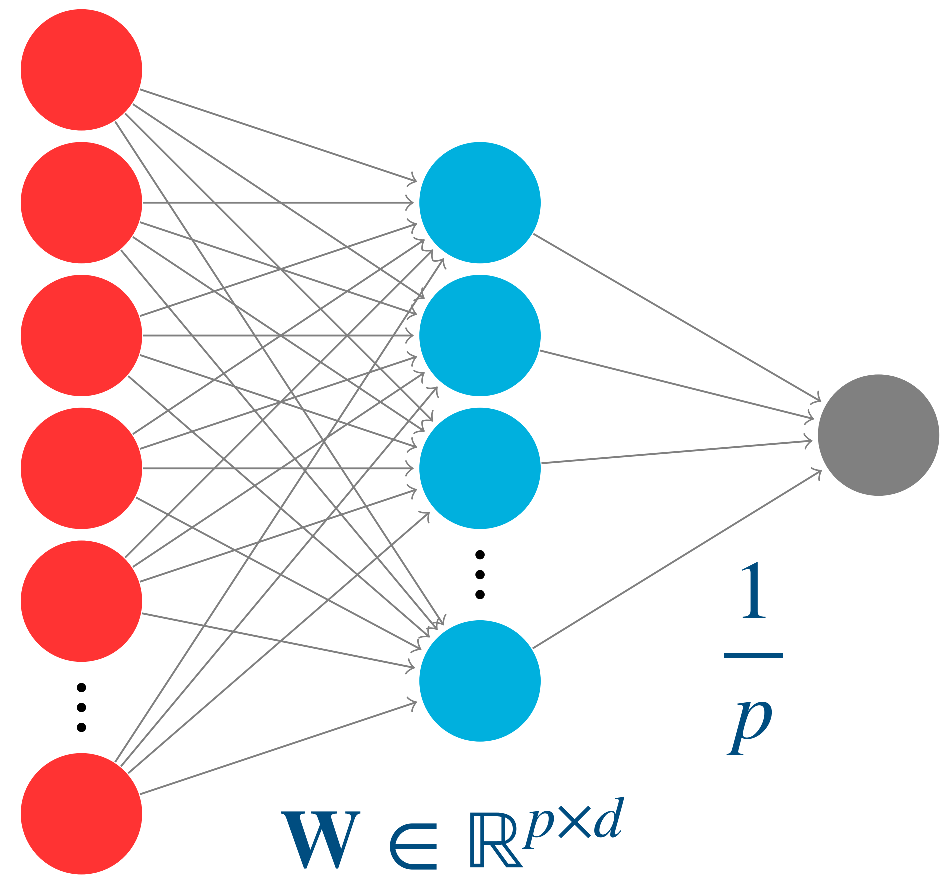
Teacher network



$$\frac{1}{k}$$

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

Student network



$$\frac{1}{p}$$

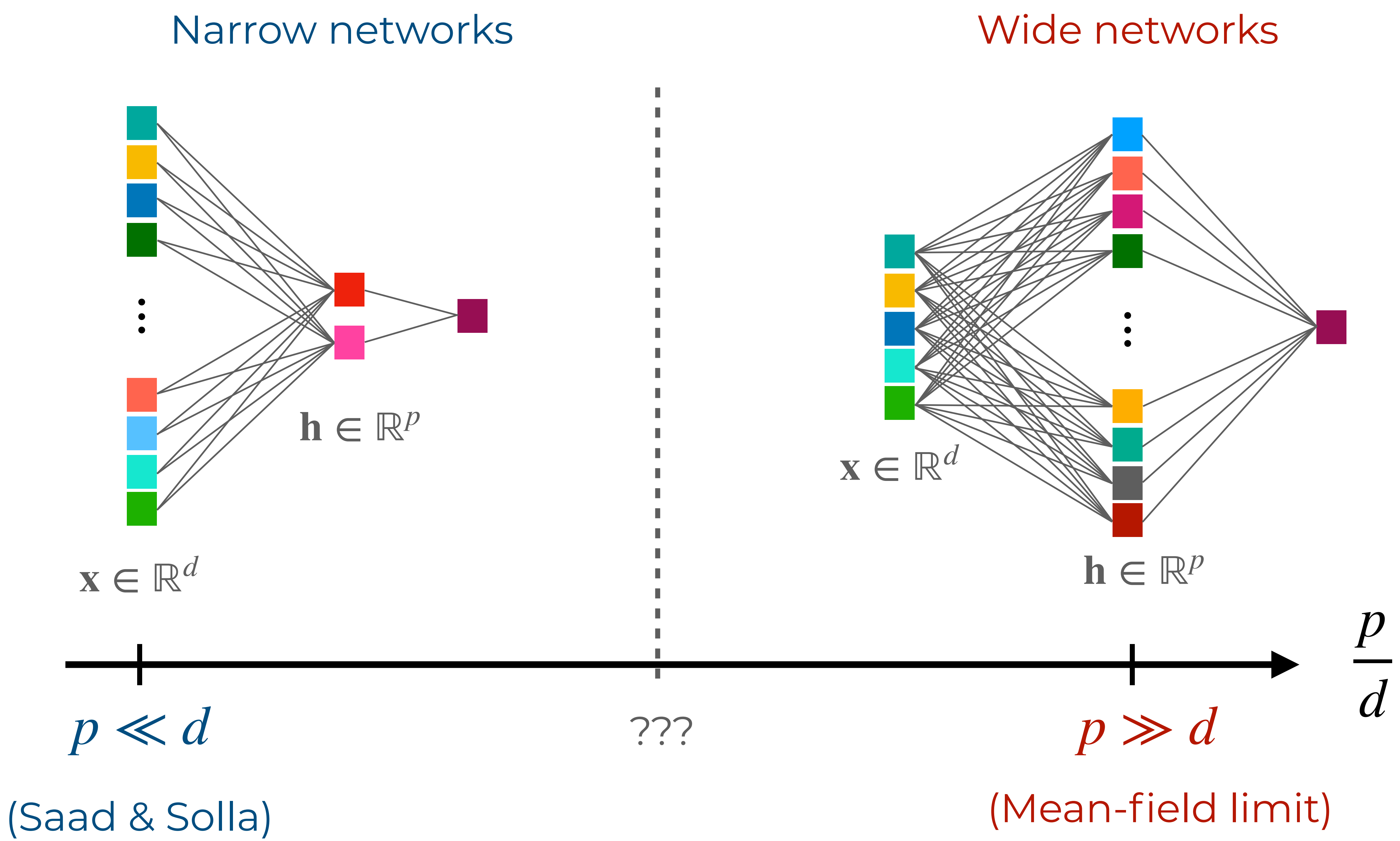
$$\mathbf{W} \in \mathbb{R}^{p \times d}$$

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

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

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

Bridging the two regimes



Infinite input limit

$$\mathbf{w}_i^{\nu+1} = \mathbf{w}_i^{\nu} - \gamma \nabla_{\mathbf{w}_i} \mathscr{R}$$

$$\downarrow$$

$$q_{jl}^{\nu+1} - q_{jl}^{\nu} = \frac{\gamma}{pd}(\mathscr{E}_j^{\nu}\lambda_l^{\nu} + \mathscr{E}_l^{\nu}\lambda_j^{\nu}) + \frac{\gamma^2}{p^2 d}\mathscr{E}_j^{\nu}\mathscr{E}_l^{\nu}$$

$$m_{jr}^{\nu+1} - m_{jr}^{\nu} = \frac{\gamma}{pd}\mathscr{E}_j^{\nu}\lambda_r^*$$

$$\mathscr{E}_j^{\nu} \equiv \sigma'(\lambda_j^{\nu})\left[\frac{1}{k}\sum_{r=1}^{k}\sigma(\lambda_r^*) - \frac{1}{p}\sum_{i=1}^{p}\sigma(\lambda_i^{\nu}) + \sqrt{\Delta}\zeta^{\nu}\right]$$

Main theoretical result

Theorem (Veiga, Stephan, **BL**, Krzakala, Zdeborová 22')

For $k = O(1), p \sim d^{\kappa}, \gamma \sim d^{-\delta}, \delta t = \max\left(\mathrm{d}^{-(1+\kappa+\delta)}, \mathrm{d}^{-(1+2(\delta+\kappa))}\right)$,

$$\Omega^{\nu+1} = \Omega^{\nu} + \delta t\ \psi(\Omega^{\nu}) \quad \xrightarrow{d \to \infty} \quad \frac{\mathrm{d}\bar{\Omega}(\mathrm{t})}{\mathrm{dt}} = \psi_{\kappa+\delta}(\bar{\Omega}(t))$$

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

Main theoretical result

For $k = O(1), p \sim d^{\kappa}, \gamma \sim d^{-\delta}, \delta t = \max\left(\mathrm{d}^{-(1+\kappa+\delta)}, \mathrm{d}^{-(1+2(\delta+\kappa))}\right)$,

$$\Omega^{\nu+1} = \Omega^{\nu} + \delta t \;\; \psi(\Omega^{\nu}) :$$

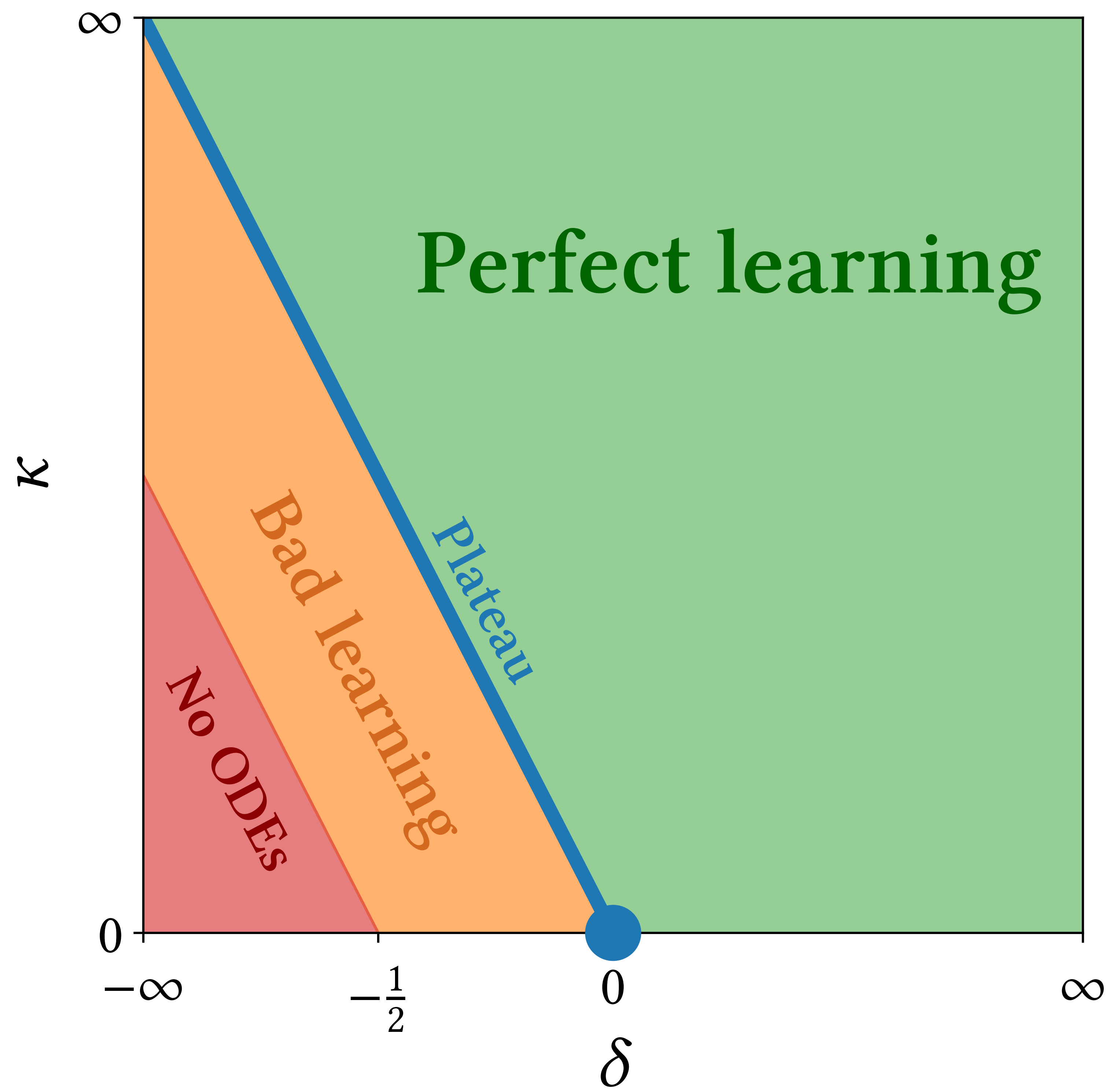
$$q_{jl}^{\nu+1} - q_{jl}^{\nu} = \frac{1}{d^{1+\kappa+\delta}} I_{\text{learning}}(\Omega^{\nu}) + \frac{1}{d^{1+2(\kappa+\delta)}} I_{\text{noise}}(\Omega^{\nu})$$

$$m_{jr}^{\nu+1} - m_{jr}^{\nu} = \frac{1}{d^{1+\kappa+\delta}} I^{*}_{\text{learning}}(\Omega^{\nu})$$

Interplay between learning and noise terms!

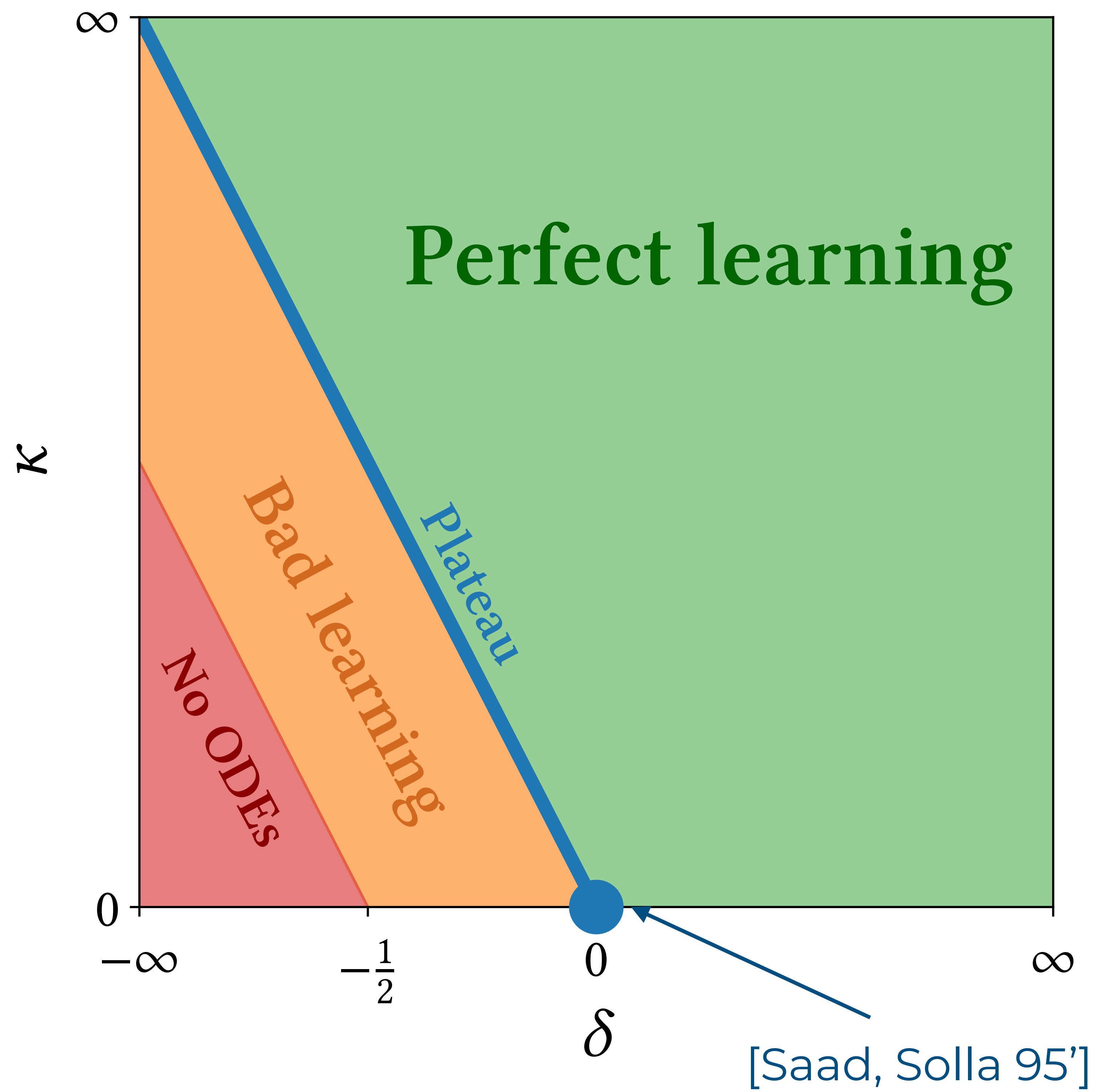
Phase diagram

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



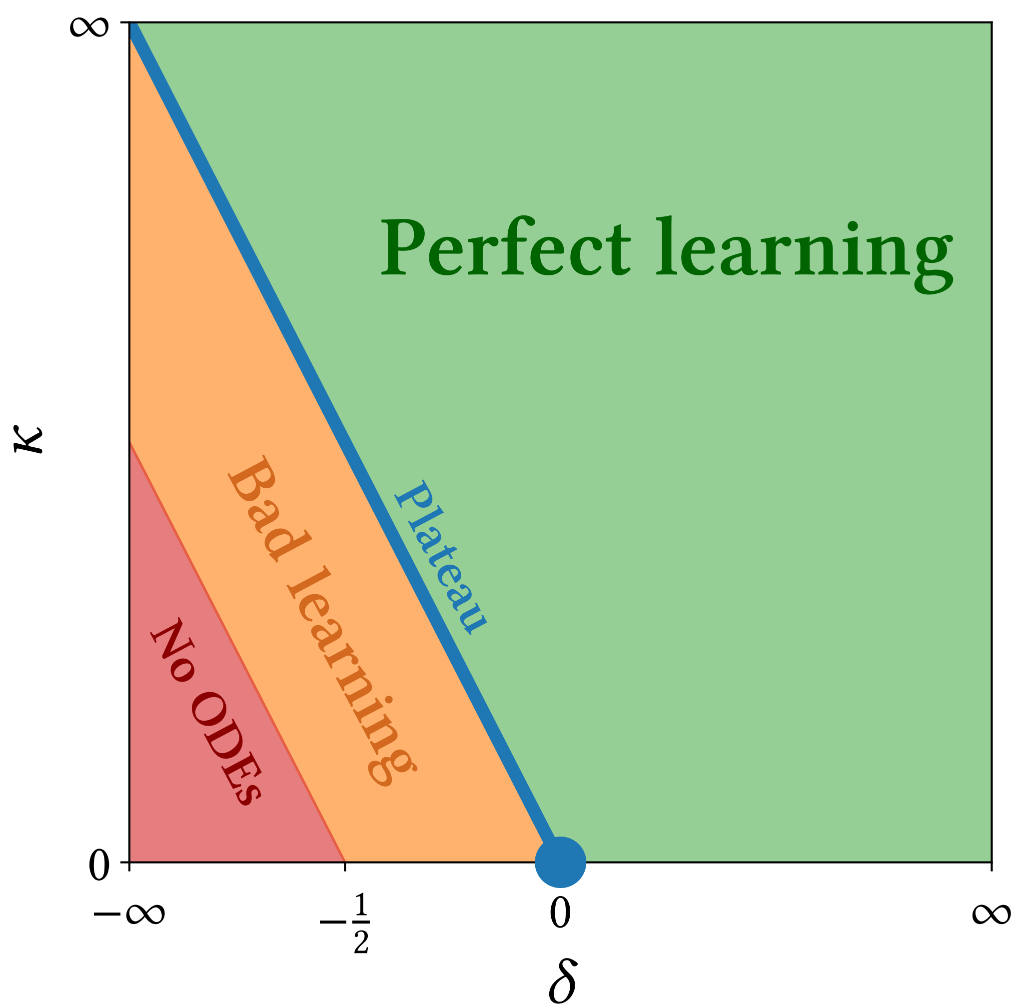
Phase diagram

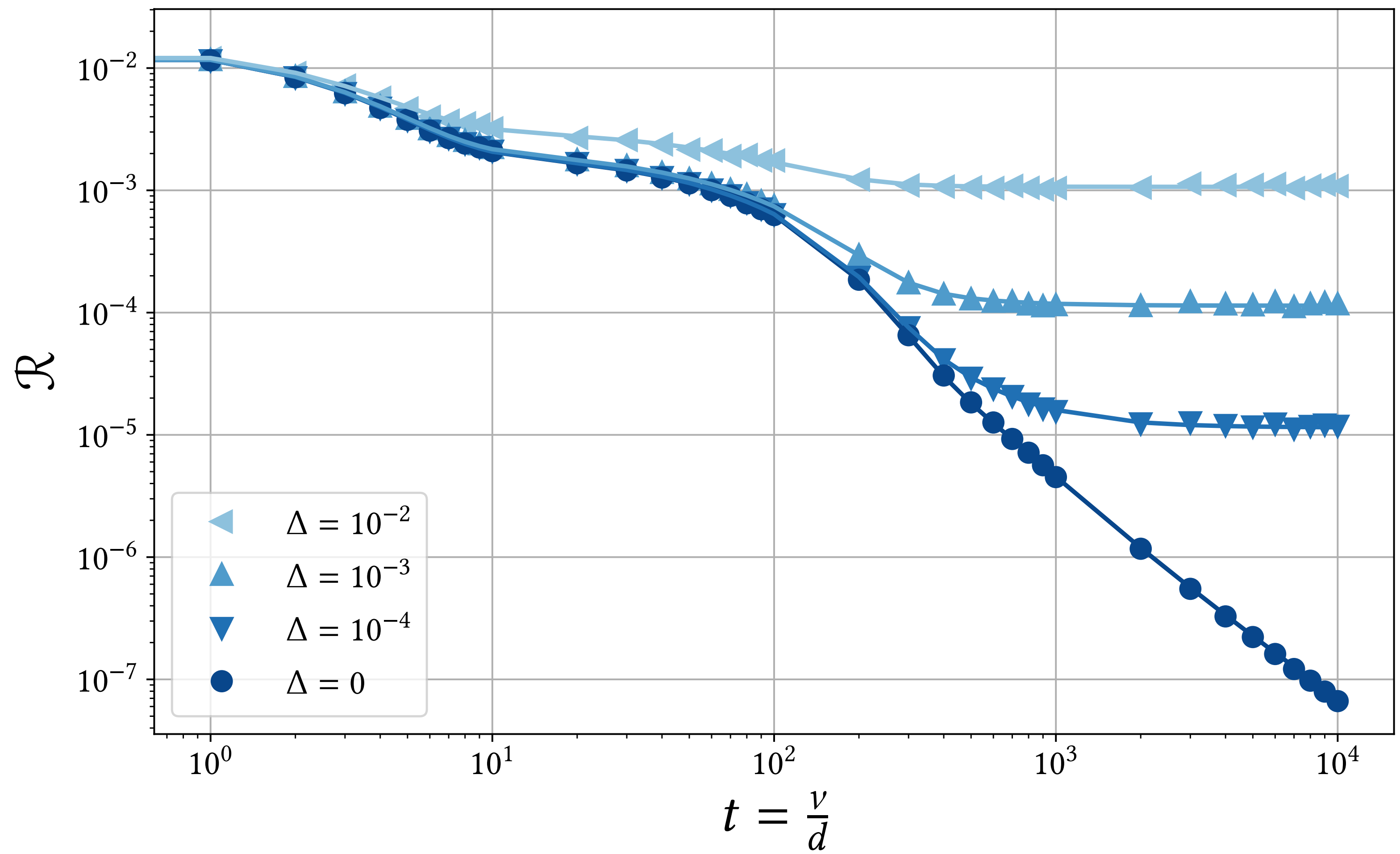
$p \sim d^{\kappa}$ $\gamma \sim d^{-\delta}$



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

$$\delta t = 1/d$$



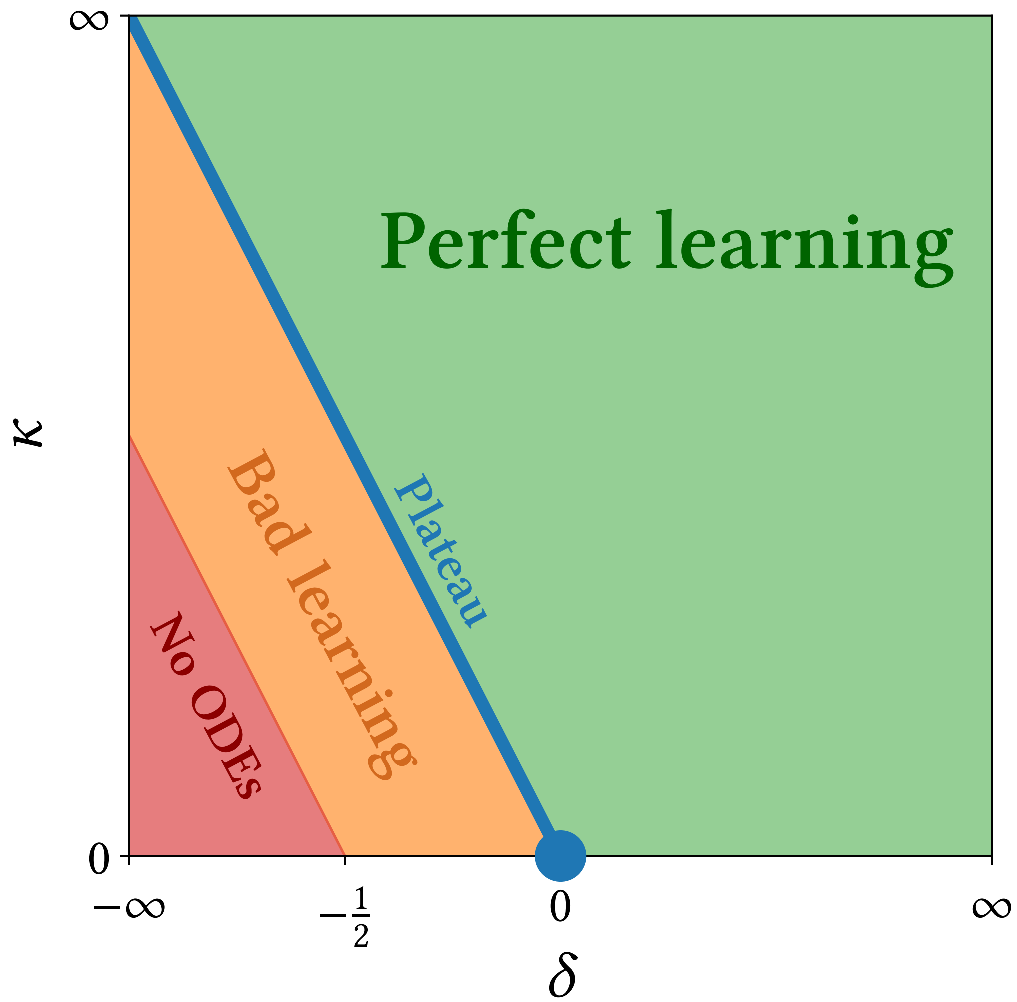


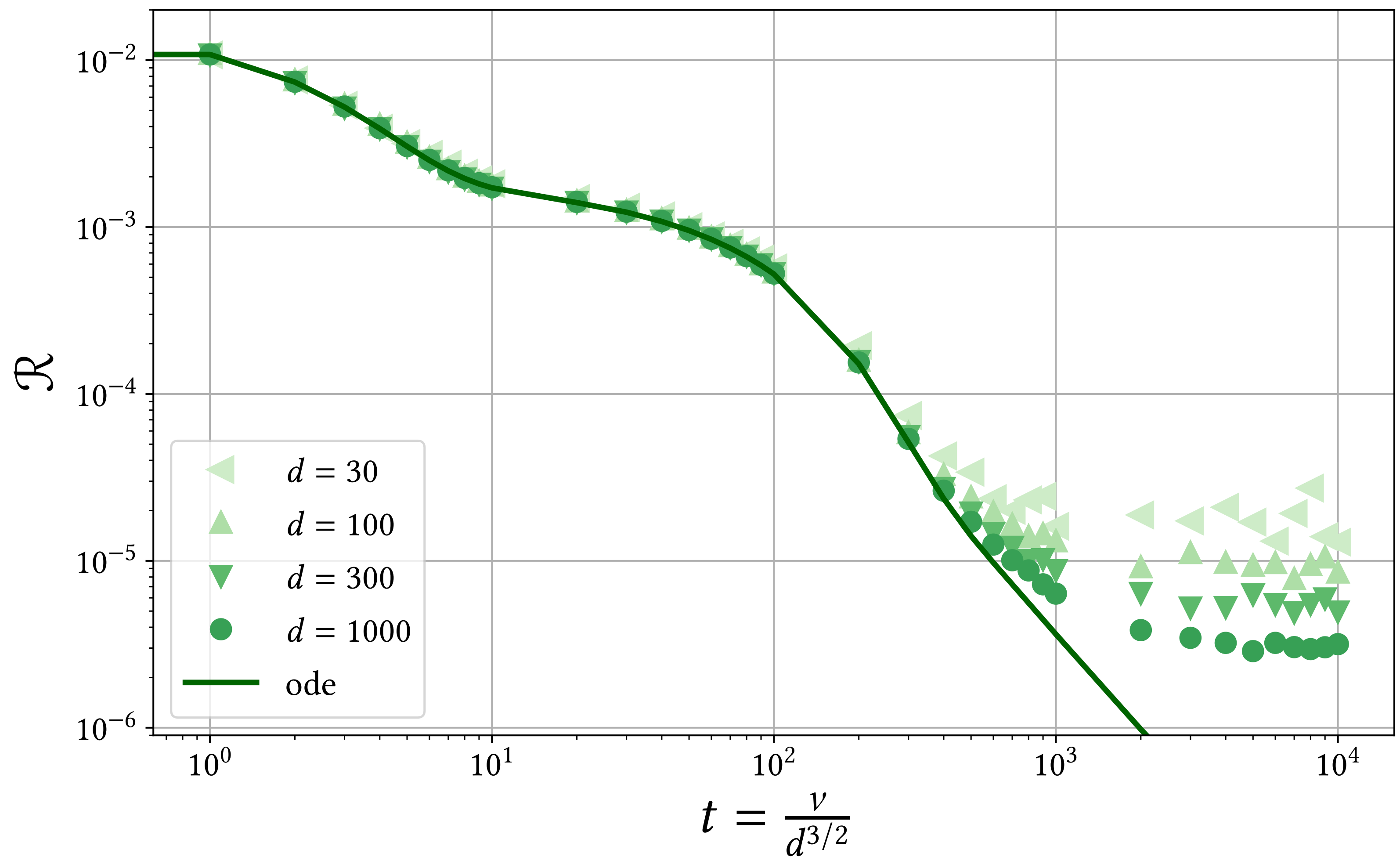
Extension of S&S regime to the whole blue line
(same phenomenology)

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

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

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



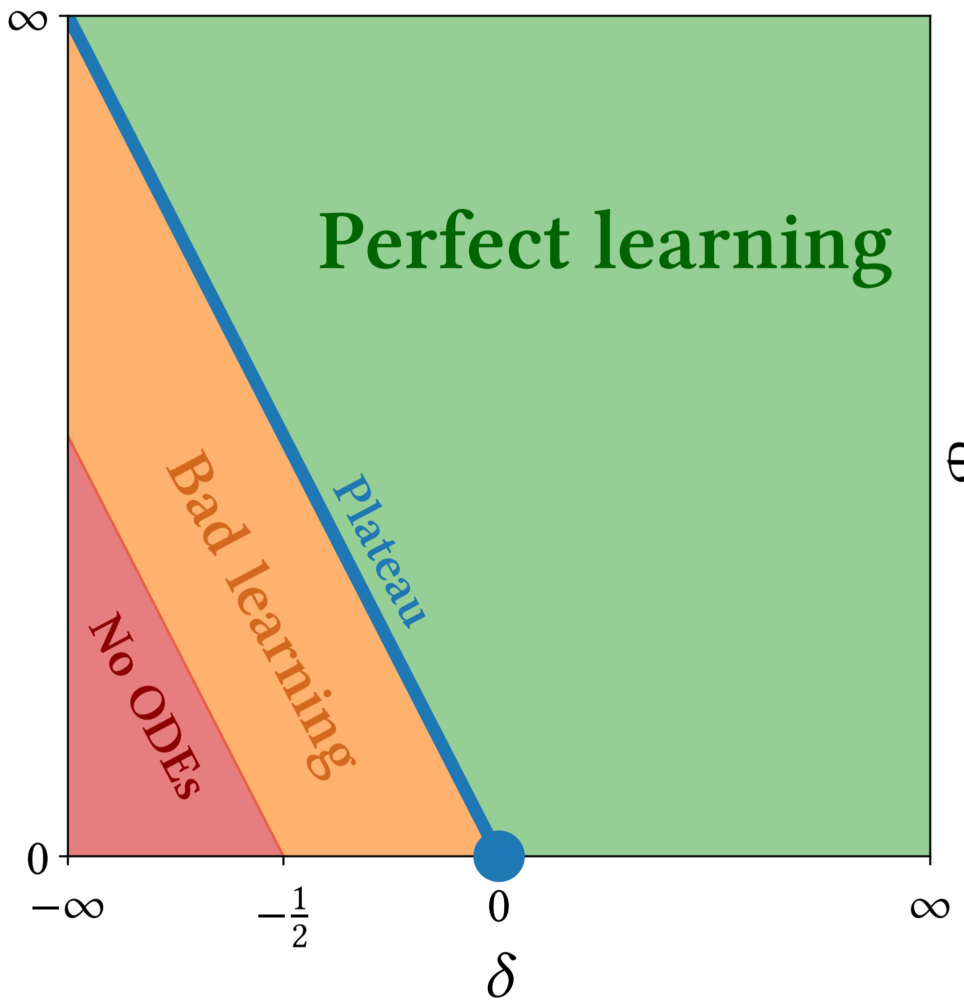


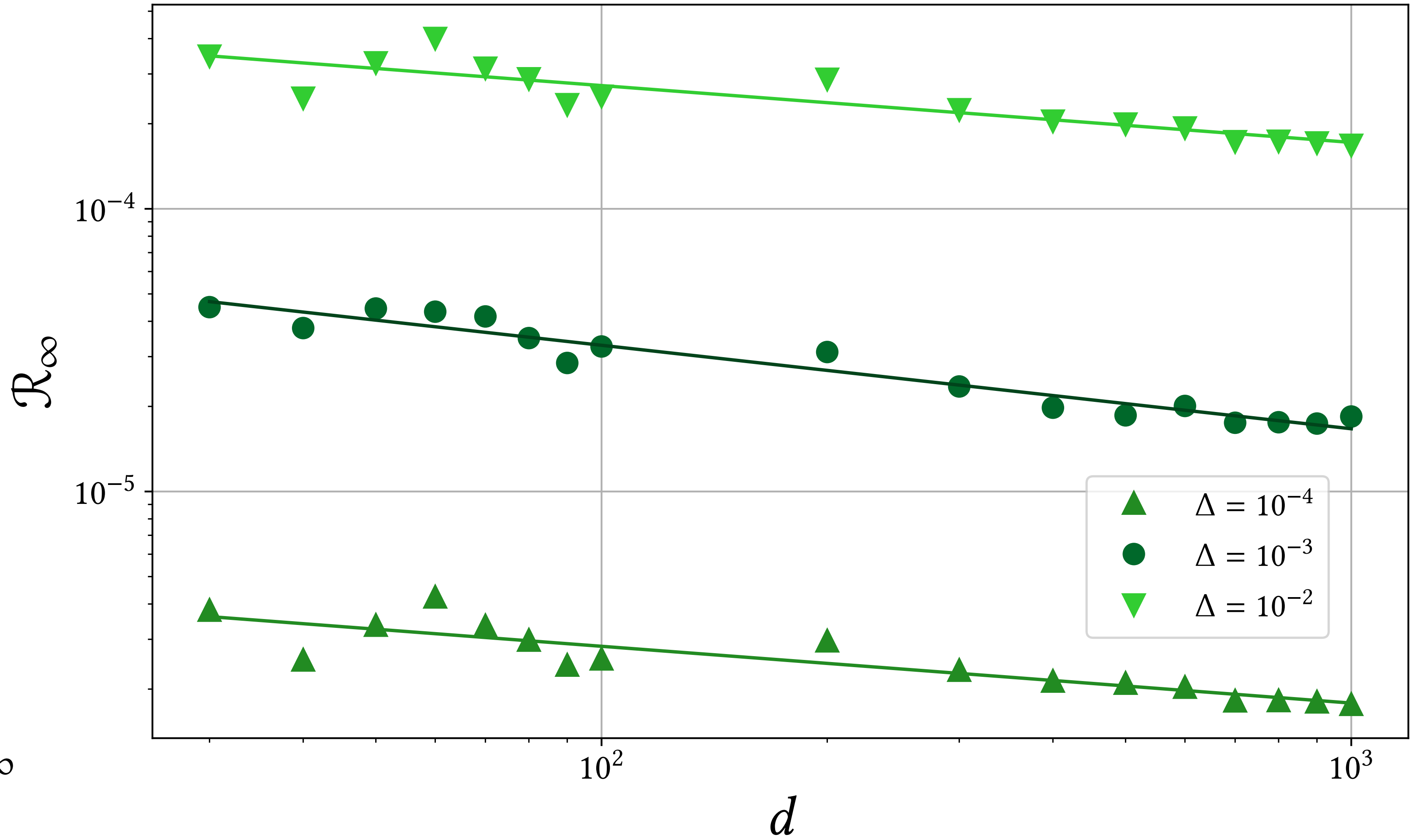
Perfect learning is achieved for any finite hidden layer width!

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

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

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



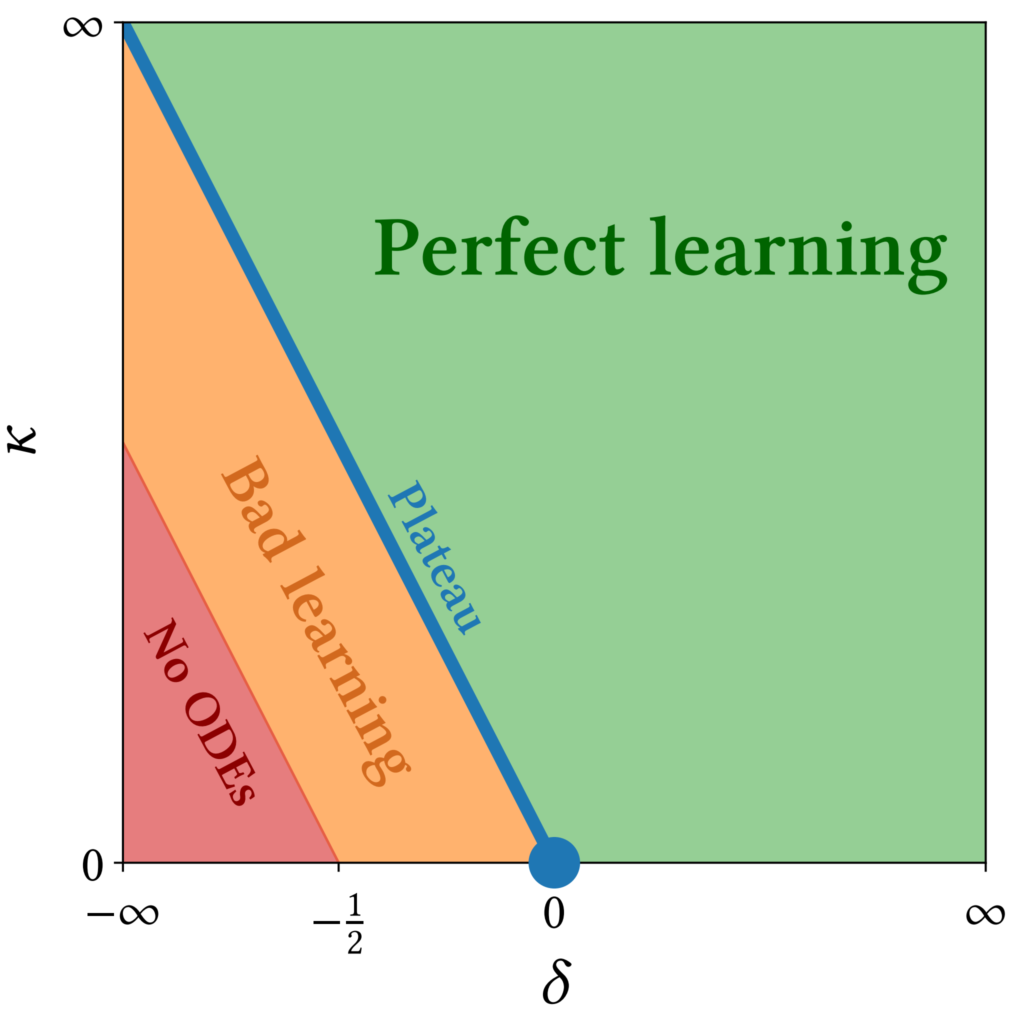


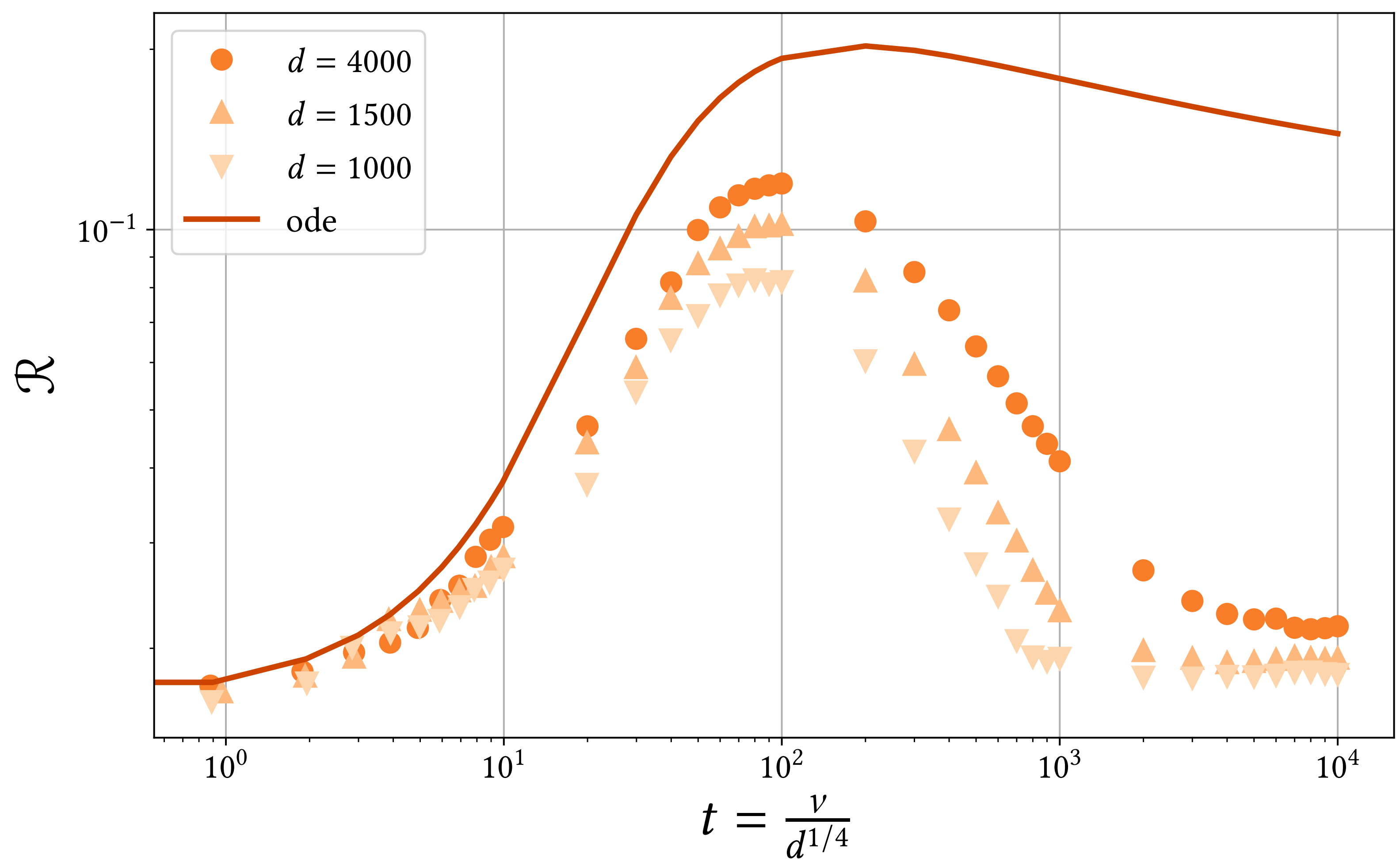
$$\mathcal{R}_\infty \sim d^{-\delta}$$

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

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

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



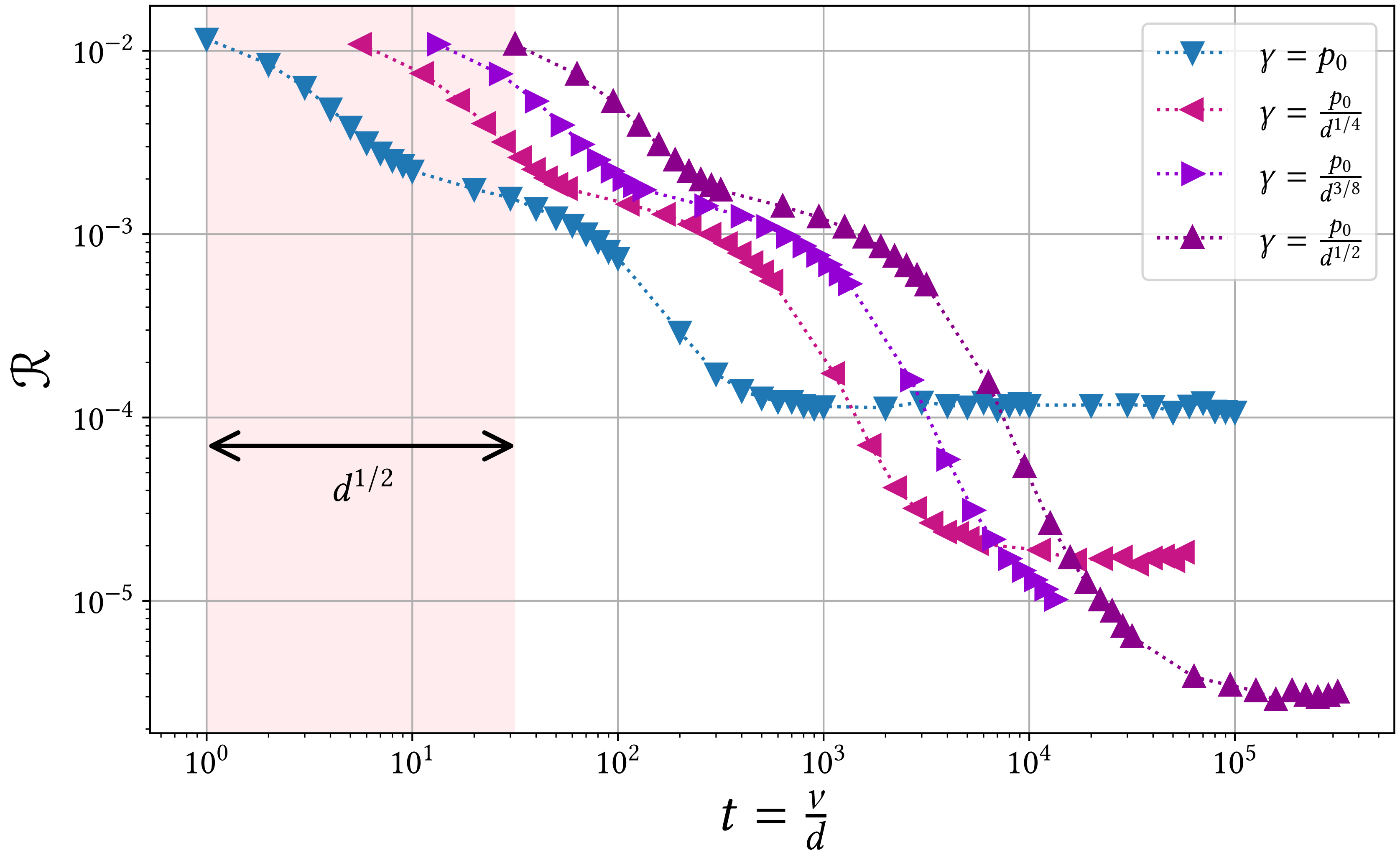


γ growing with d (weird!)

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

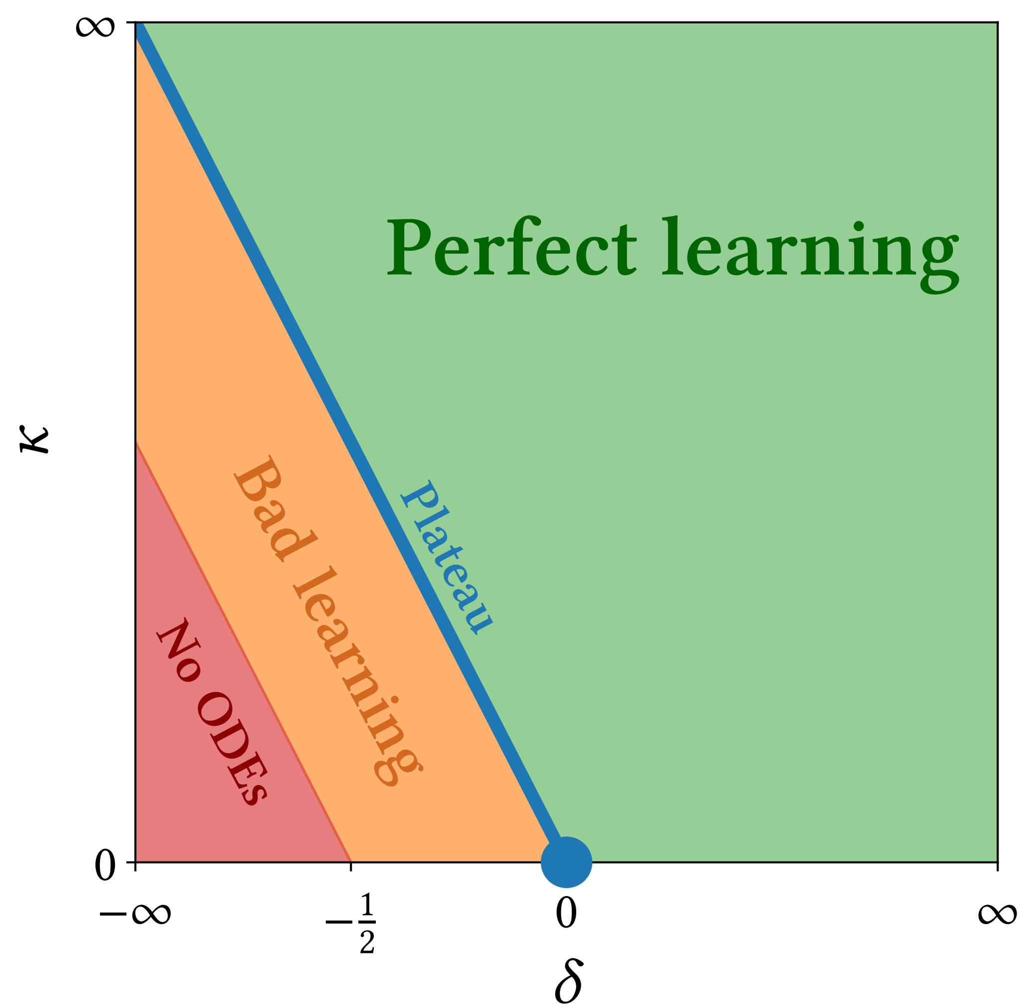
$$\kappa = 0 \qquad d = 1000 \qquad \Delta = 10^{-3}$$

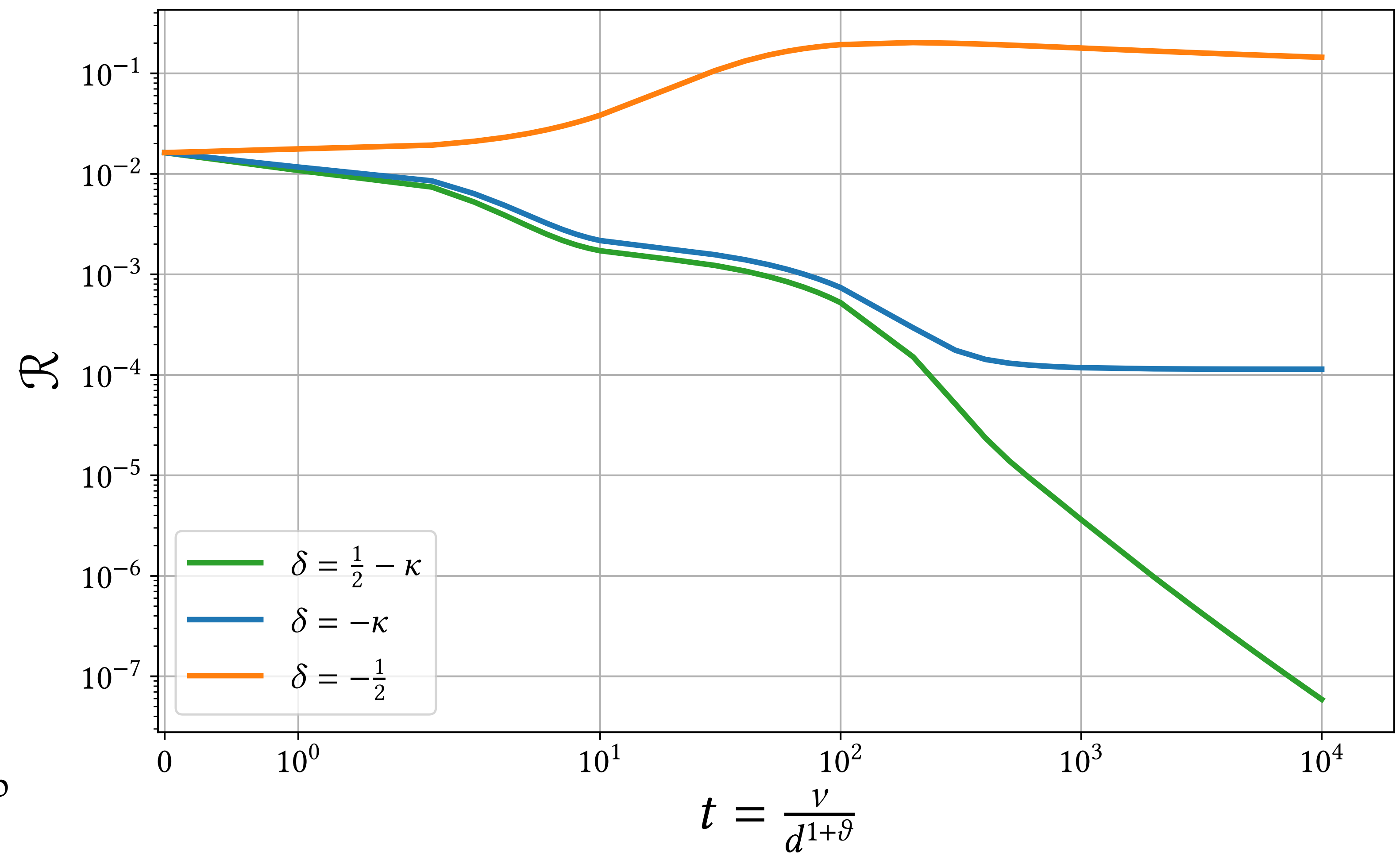


$$n \sim d^{1+\delta}$$

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

Summary



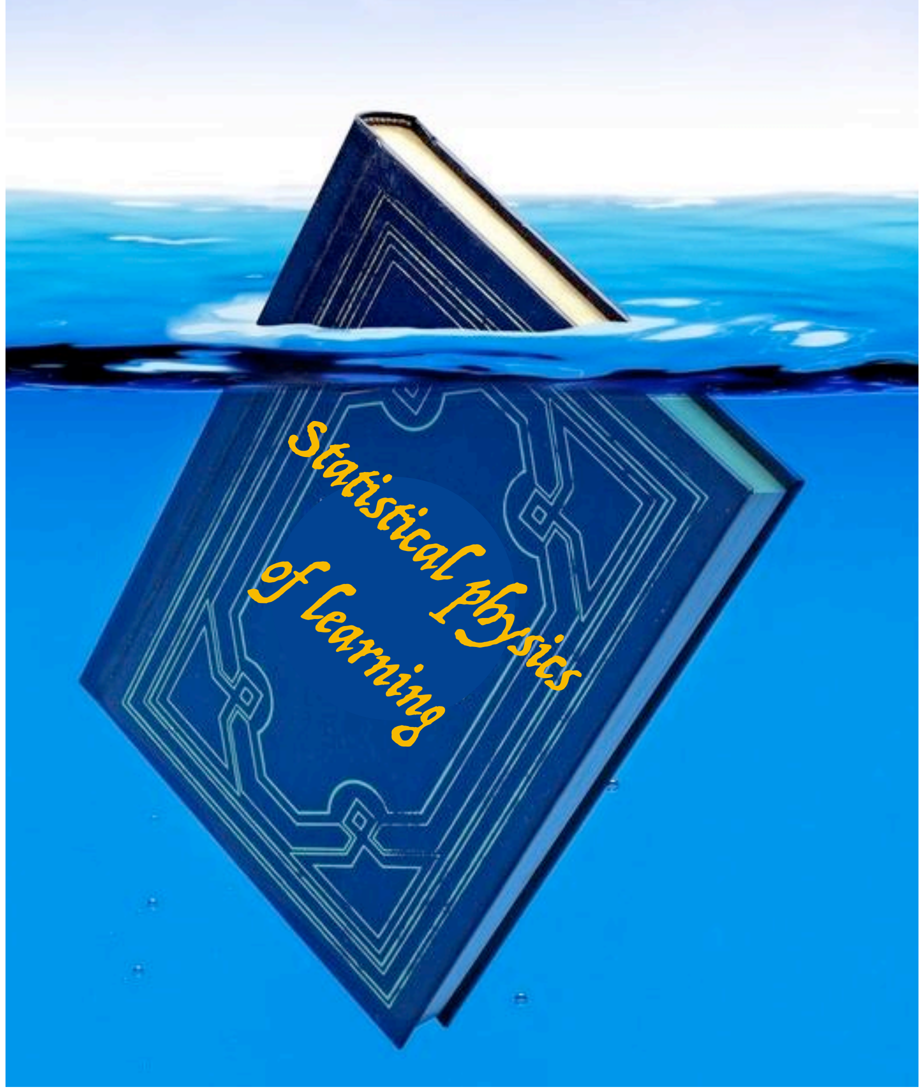


$\theta = 0$ $\quad$ $\theta = \kappa + \delta$ $\quad$ $\theta = 2(\kappa + \delta)$

Sum-up

- What do we mean by "theory"
- Why statistical physics has anything to do with that?
- A concrete example: phase diagram for one-pass SGD dynamics for 2-layer neural networks

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



brloureiro@gmail.com