



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

Bruno Loureiro

Département d'Informatique
École Normale Supérieure & CNRS

brloureiro@gmail.com

Outline

1. Theory of machine learning? A statistical physics point of view
2. Two layer neural networks in the lazy regime
3. Two layer neural networks in the rich regime

Part I

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.



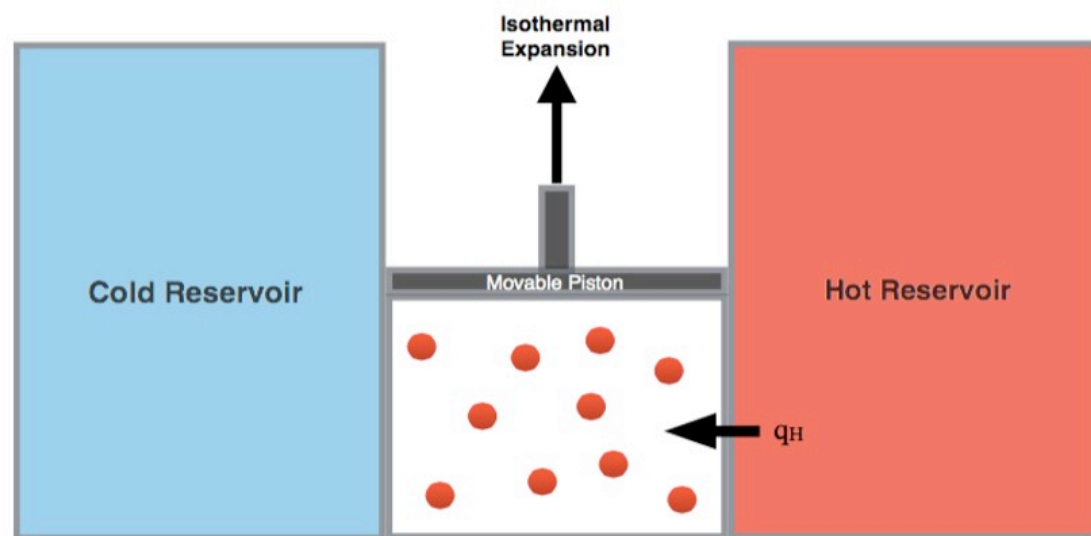
fridge

Theory of ~~machine learning~~?

Theory can mean different things.

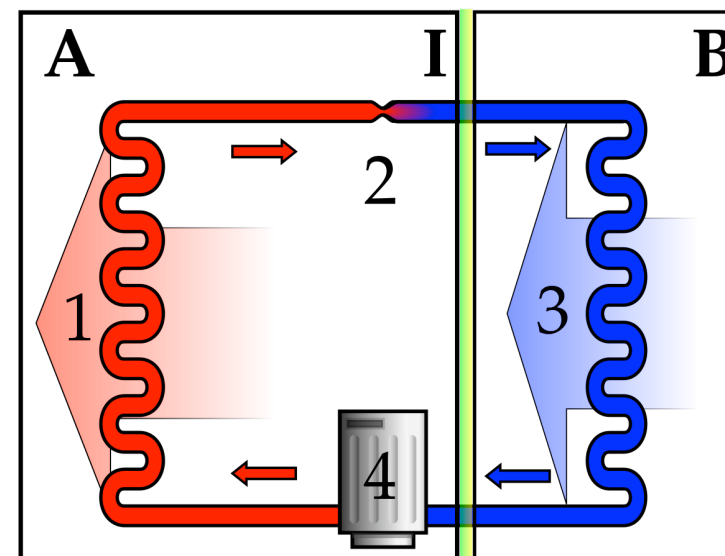
Physics

Fundamental laws that govern behaviour of the fridge



Engineering

How do I build a good fridge?



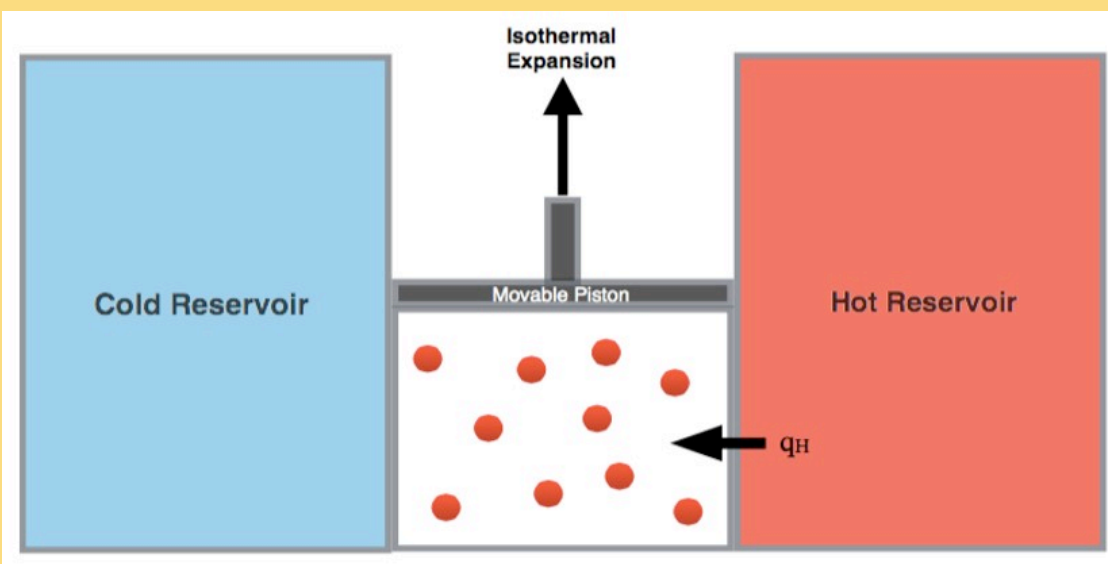
fridge

Theory of ~~machine learning~~?

Theory can mean different things.

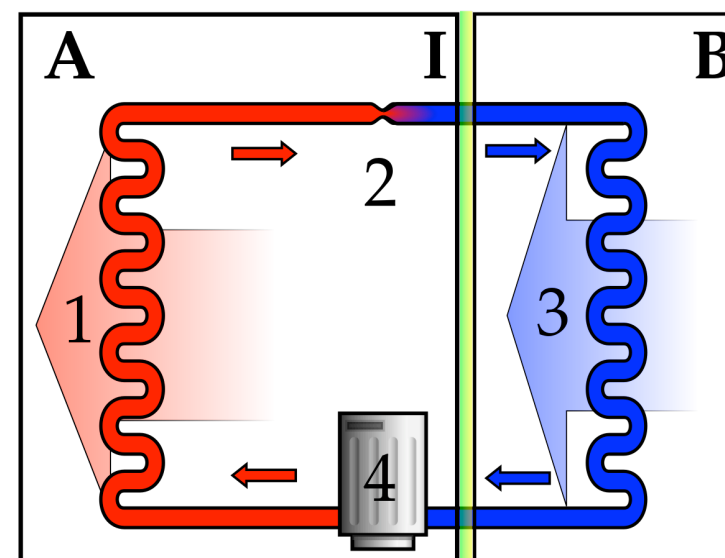
Physics

Fundamental laws that govern behaviour of the fridge



Engineering

How do I build a good fridge?

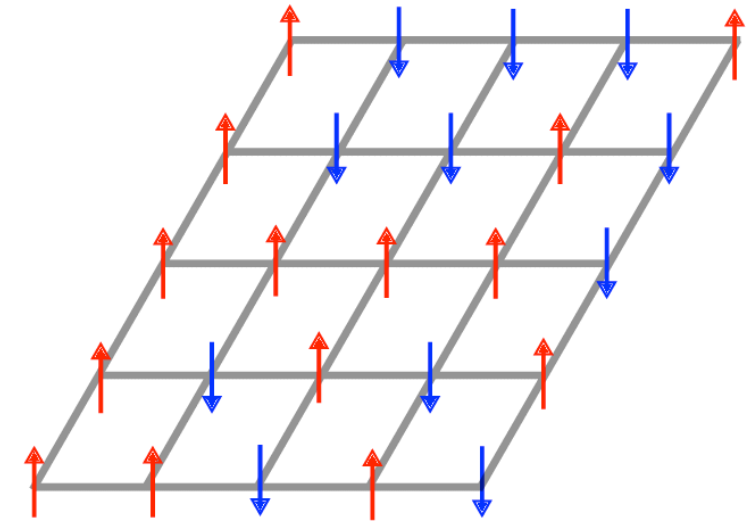


Theory of magnetism

a.k.a. the *Ising Model*

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

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

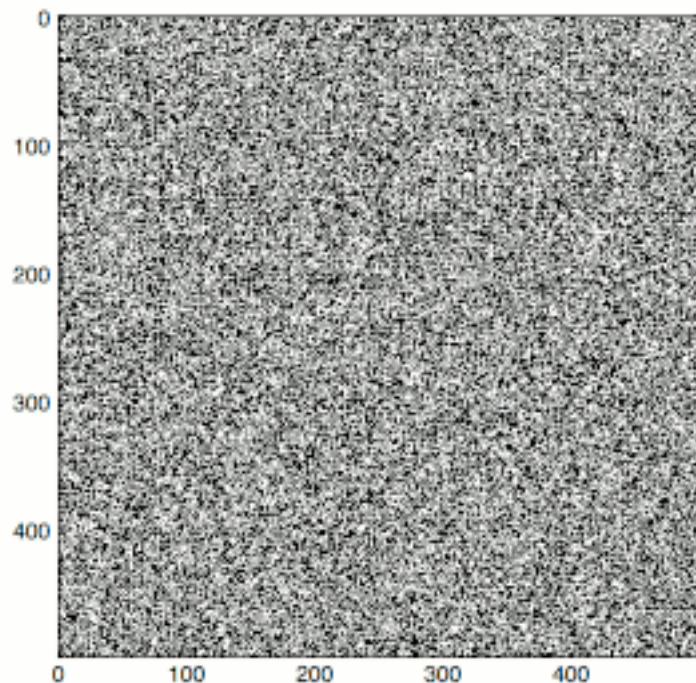
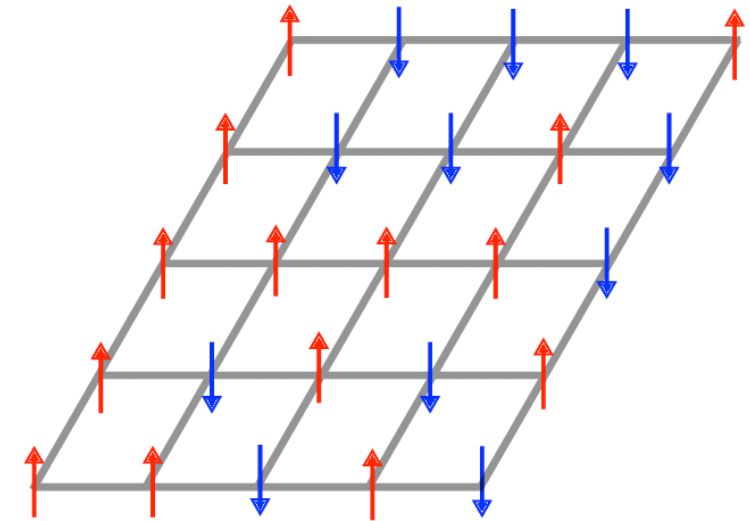


Theory of magnetism

a.k.a. the *Ising Model*

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

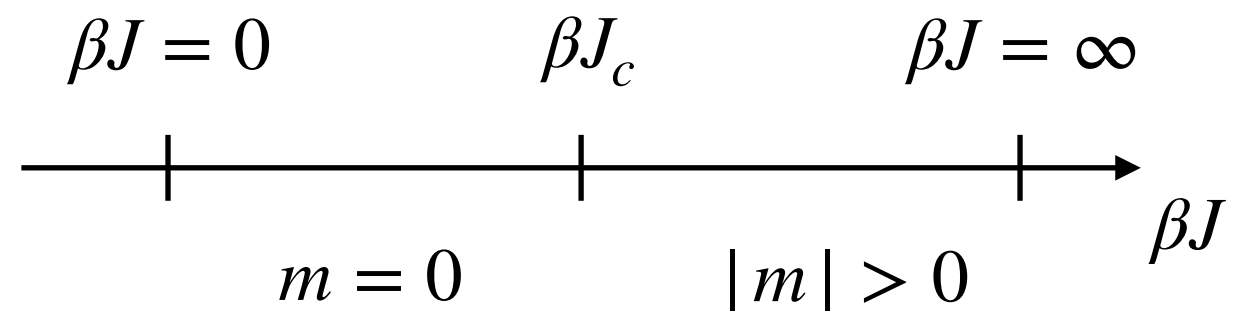
$$\mu_\beta(s) = \frac{1}{Z_{\beta,J,h}} e^{-\beta H_{J,h}(s)} \quad s \in \{-1, +1\}^N$$



$h = 0$

Order
parameter:

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



[Ising 1925; Onsager 1944]

Theory of machine learning?

Theory can mean different things.

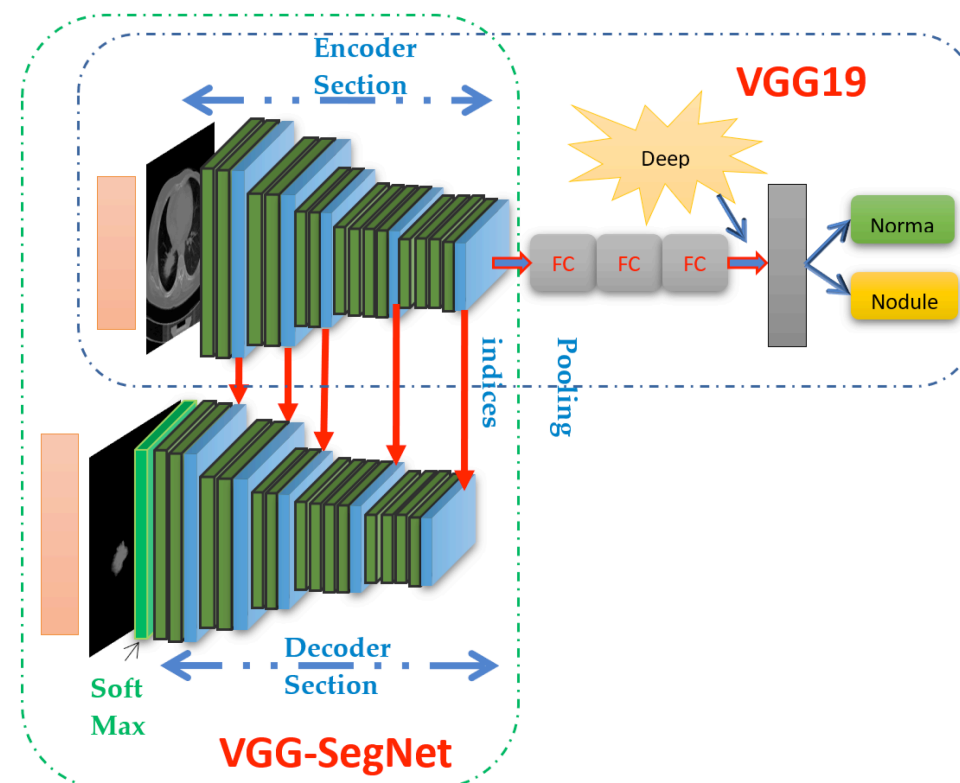
Theory

Fundamental principles
that govern learning



Engineering

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



Supervised Learning

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

Supervised Learning


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

Want: Learn $f: \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}

Supervised Learning

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


Want: Learn $f: \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}


$$f(x) = \begin{cases} y^\nu & \text{if } x \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

Supervised Learning

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

Want: Learn $f: \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}



$$f(x) = \begin{cases} y^\nu & \text{if } x \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation,
not learning!

Supervised Learning

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

Want: Learn $f: \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}


$$f(x) = \begin{cases} y^\nu & \text{if } x \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation,
not learning!



Introduce a “cost function” $\ell(y, f(x)) \geq 0$


minimise $\mathcal{R}(f) = \mathbb{E}_{(x,y) \sim \rho}[\ell(y, f(x))]$

Population
Risk

Supervised Learning

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

Want: Learn $f: \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}


$$f(x) = \begin{cases} y^\nu & \text{if } x \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation,
not learning!



Introduce a “cost function” $\ell(y, f(x)) \geq 0$

$$\text{minimise } \mathcal{R}(f) = \mathbb{E}_{(x,y) \sim \rho}[\ell(y, f(x))]$$

Population
Risk



- Problems:
- In practice, doesn't know ρ , only \mathcal{D}
 - How to minimise over $\{f: \mathbb{R}^d \rightarrow \mathbb{R}\}$?

Supervised Learning

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

Want: Learn $f: \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}

minimise $\mathcal{R}(f) = \mathbb{E}_{(x,y) \sim \rho}[\ell(y, f(x))]$

Population
Risk

minimise $\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{\nu \in [n]} [\ell(y^\nu, f(x^\nu))]$

Empirical
Risk



Problems:

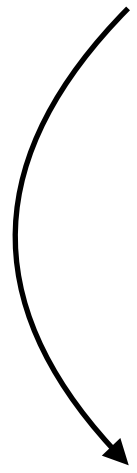
- In practice, doesn't know ρ , only \mathcal{D}
- How to minimise over $\{f: \mathbb{R}^d \rightarrow \mathbb{R}\}$?



Supervised Learning

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

Want: Learn $f_\Theta : \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}


$$\text{minimise } \mathcal{R}(\Theta) = \mathbb{E}_{(x,y) \sim \rho} [\ell(y, f_\Theta(x))]$$

Population
Risk

$$\text{minimise } \hat{\mathcal{R}}_n(\Theta) = \frac{1}{n} \sum_{\nu \in [n]} [\ell(y^\nu, f_\Theta(x^\nu))]$$

Empirical
Risk



Problems:

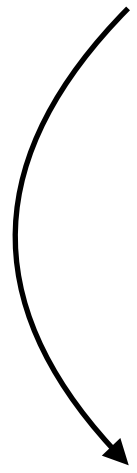
- In practice, doesn't know ρ , only \mathcal{D}
- How to minimise over $\{f : \mathbb{R}^d \rightarrow \mathbb{R}\}$?



Supervised Learning

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

Want: Learn $f_\Theta: \mathbb{R}^d \rightarrow \mathbb{R}$ from data \mathcal{D}



minimise $\mathcal{R}(\Theta) = \mathbb{E}_{(x,y) \sim \rho}[\ell(y, f_\Theta(x))]$

Population
Risk

minimise $\hat{\mathcal{R}}_n(\Theta) = \frac{1}{n} \sum_{\nu \in [n]} [\ell(y^\nu, f_\Theta(x^\nu))]$

Empirical
Risk



Problems:

- In practice, doesn't know ρ , only \mathcal{D}
- How to minimise over $\{f: \mathbb{R}^d \rightarrow \mathbb{R}\}$?



Stat. Learning Theory

Supervised binary classification $(x^\nu, y^\nu) \in \mathbb{R}^d \times \{-1, 1\}$, $\nu = 1, \dots, n$

Stat. Learning Theory

Supervised binary classification $(x^\nu, y^\nu) \in \mathbb{R}^d \times \{-1, 1\}$, $\nu = 1, \dots, n$

Theorem (Uniform convergence): with probability at least $1 - \delta$

$$\forall f_\Theta \in \mathcal{H} \quad \mathcal{R}(\Theta) - \hat{\mathcal{R}}_n(\Theta) \leq \text{Rad}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n}}$$

Where

$$\text{Rad}(\mathcal{H}) = \frac{1}{n} \mathbb{E} \left[\sup_{f_\Theta \in \mathcal{H}} \sum_{\nu \in [n]} y^\nu f_\Theta(x^\nu) \right]$$

Stat. Learning Theory

Supervised binary classification $(x^\nu, y^\nu) \in \mathbb{R}^d \times \{-1, 1\}$, $\nu = 1, \dots, n$

Theorem (Uniform convergence): with probability at least $1 - \delta$

$$\forall f_\Theta \in \mathcal{H} \quad \mathcal{R}(\Theta) - \hat{\mathcal{R}}_n(\Theta) \leq \text{Rad}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n}}$$

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

assignments. While we consider multiclass problems, it is straightforward to consider related binary classification problems for which the same experimental observations hold. Since our randomization tests suggest that many neural networks fit the training set with random labels perfectly, we expect that $\hat{\mathcal{R}}_n(\mathcal{H}) \approx 1$ for the corresponding model class \mathcal{H} . This is, of course, a trivial upper bound on the Rademacher complexity that does not lead to useful generalization bounds in realistic settings.

[Zhang, Bengio, Hardt, Recht, Vinyals 17']

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?

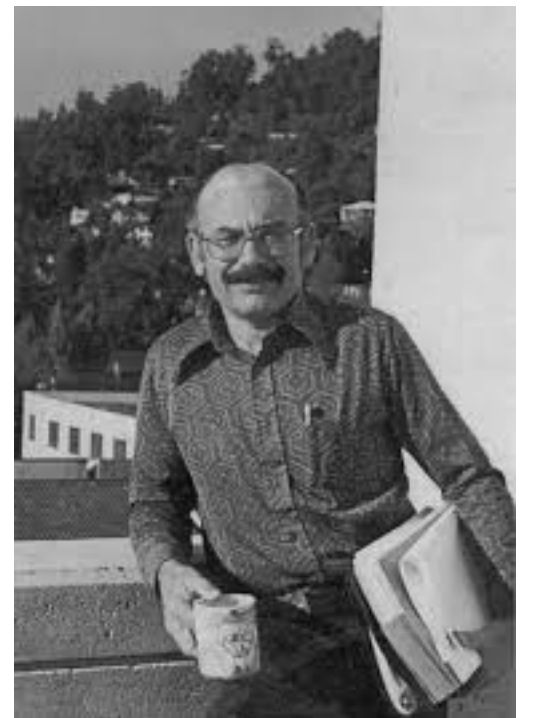
Many questions, few answers

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**



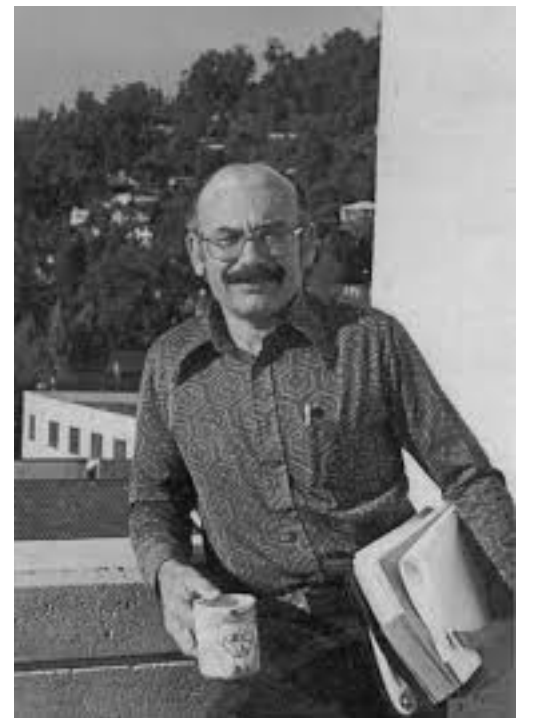
Many questions, few answers

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**



Bias-Variance decomposition

For $\ell(y, f_{\Theta}(x)) = (y - f_{\Theta}(x))^2$:

$$f_{\star}(x) = \operatorname{argmin}_f \mathcal{R}(f) = \mathbb{E}[y | x]$$

Bias-Variance decomposition

For $\ell(y, f_{\Theta}(x)) = (y - f_{\Theta}(x))^2$:

$$f_{\star}(x) = \operatorname{argmin}_f \mathcal{R}(f) = \mathbb{E}[y | x]$$

Hence, for $\hat{\Theta} = \hat{\Theta}(X, y)$ the excess risk is given by:

$$\mathcal{R}(\hat{\Theta}) - \mathcal{R}(f_{\star}) = \mathbb{E}[(f_{\star}(x) - f(x; \Theta))^2]$$

Bias-Variance decomposition

For $\ell(y, f_{\Theta}(x)) = (y - f_{\Theta}(x))^2$:

$$f_{\star}(x) = \operatorname{argmin}_f \mathcal{R}(f) = \mathbb{E}[y | x]$$

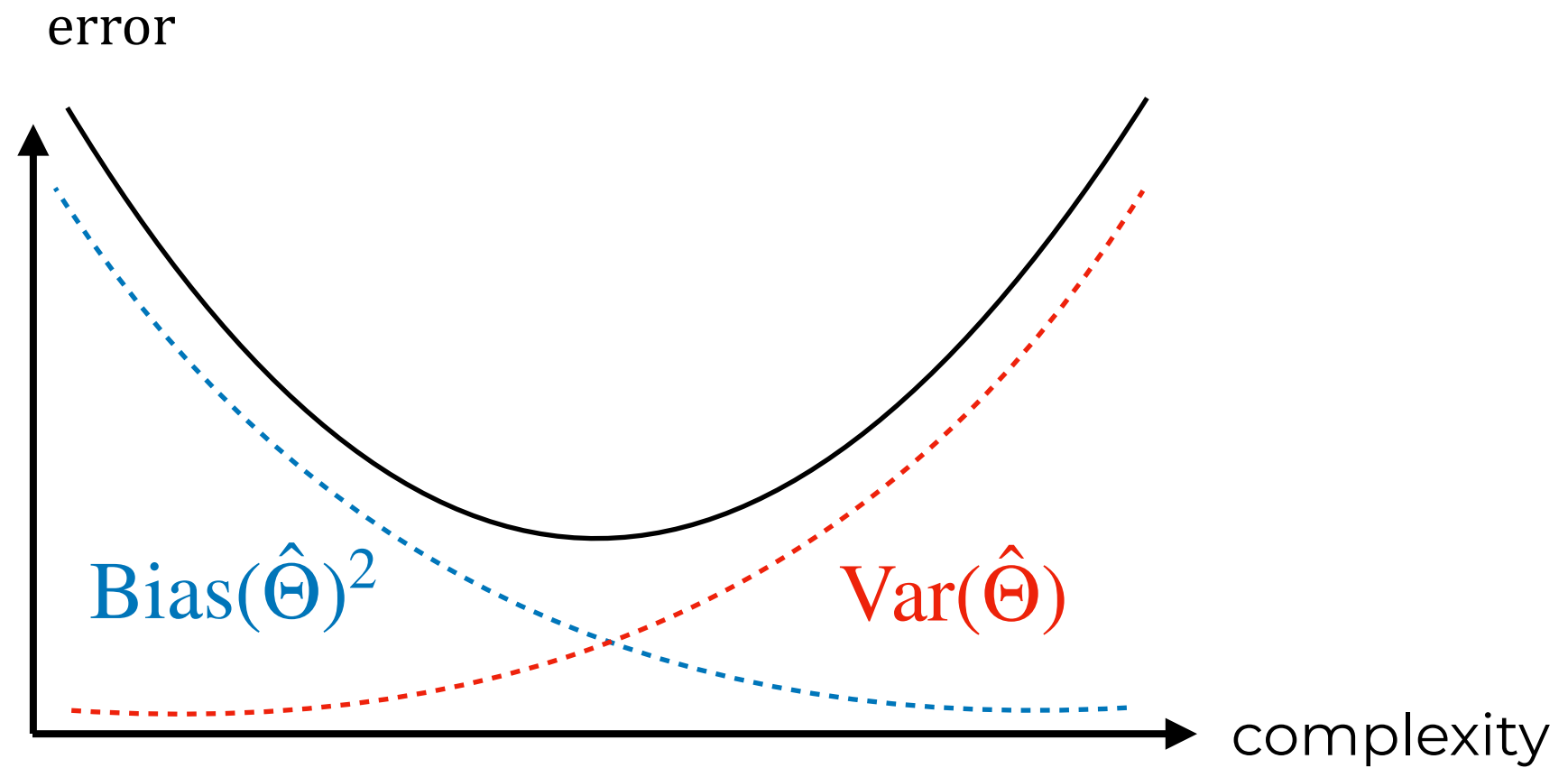
Hence, for $\hat{\Theta} = \hat{\Theta}(X, y)$ the excess risk is given by:

$$\begin{aligned} \mathcal{R}(\hat{\Theta}) - \mathcal{R}(f_{\star}) &= \mathbb{E}[(f_{\star}(x) - f(x; \hat{\Theta}))^2] \\ &= \mathbb{E}_X[\text{Bias}(\hat{\Theta})^2] + \mathbb{E}_X[\text{Var}(\hat{\Theta})] \end{aligned}$$

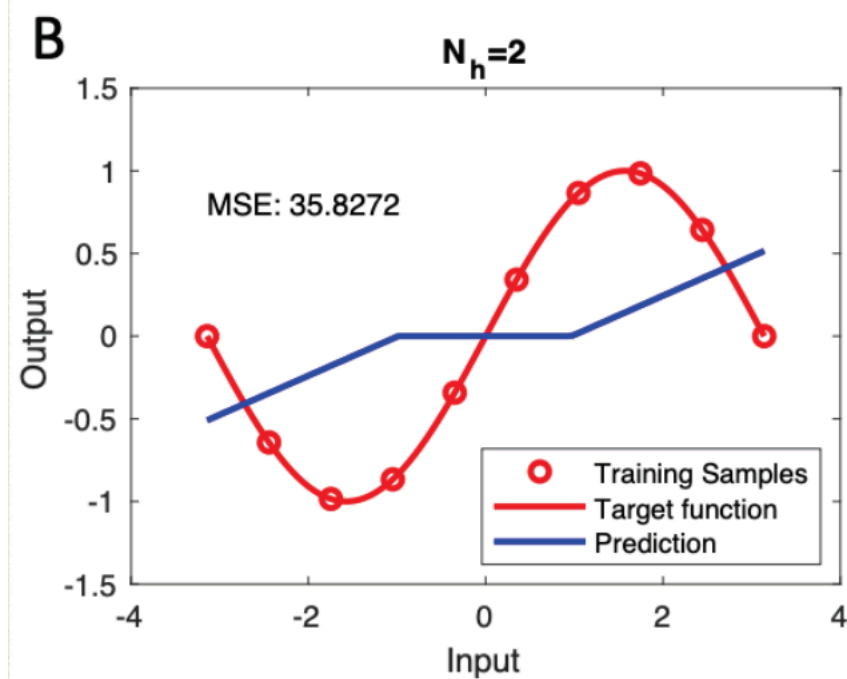
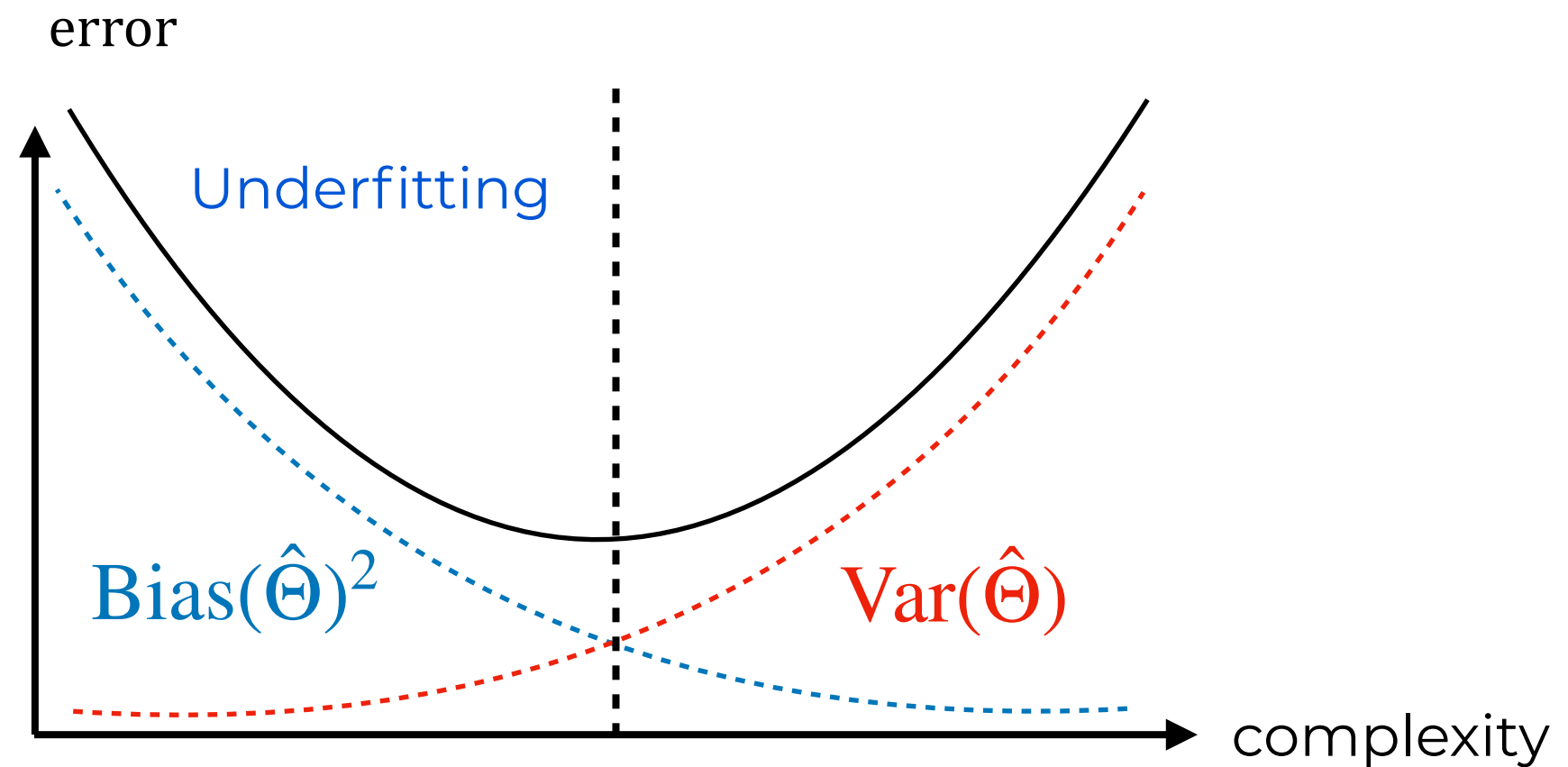
Where:

$$\begin{aligned} \text{Bias}(\hat{\Theta})^2 &= \mathbb{E}_x \left[\left(f_{\star}(x) - \mathbb{E}_y [f(x; \hat{\Theta})] \right)^2 \right] \\ \text{Var}(\hat{\Theta}) &= \mathbb{E}_{x,y} \left[\left(f(x; \hat{\Theta}) - \mathbb{E}_y [f(x; \hat{\Theta})] \right)^2 \right] \end{aligned}$$

Bias-variance trade-off

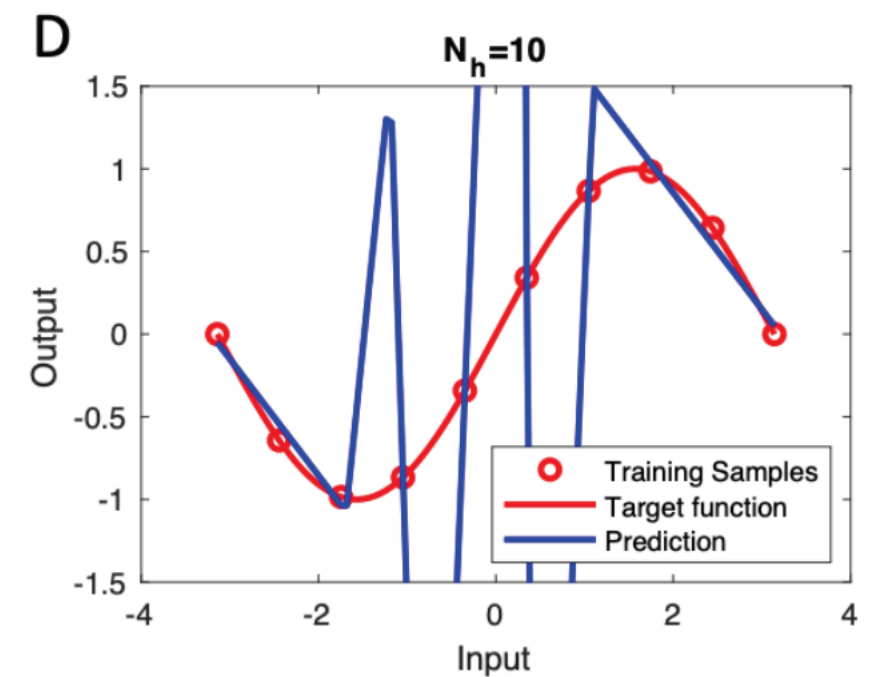
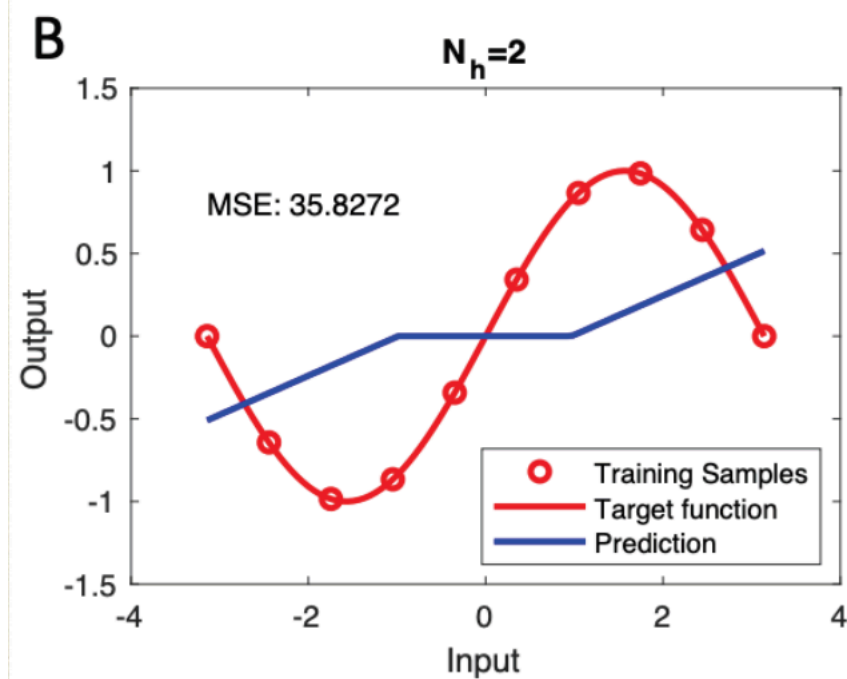
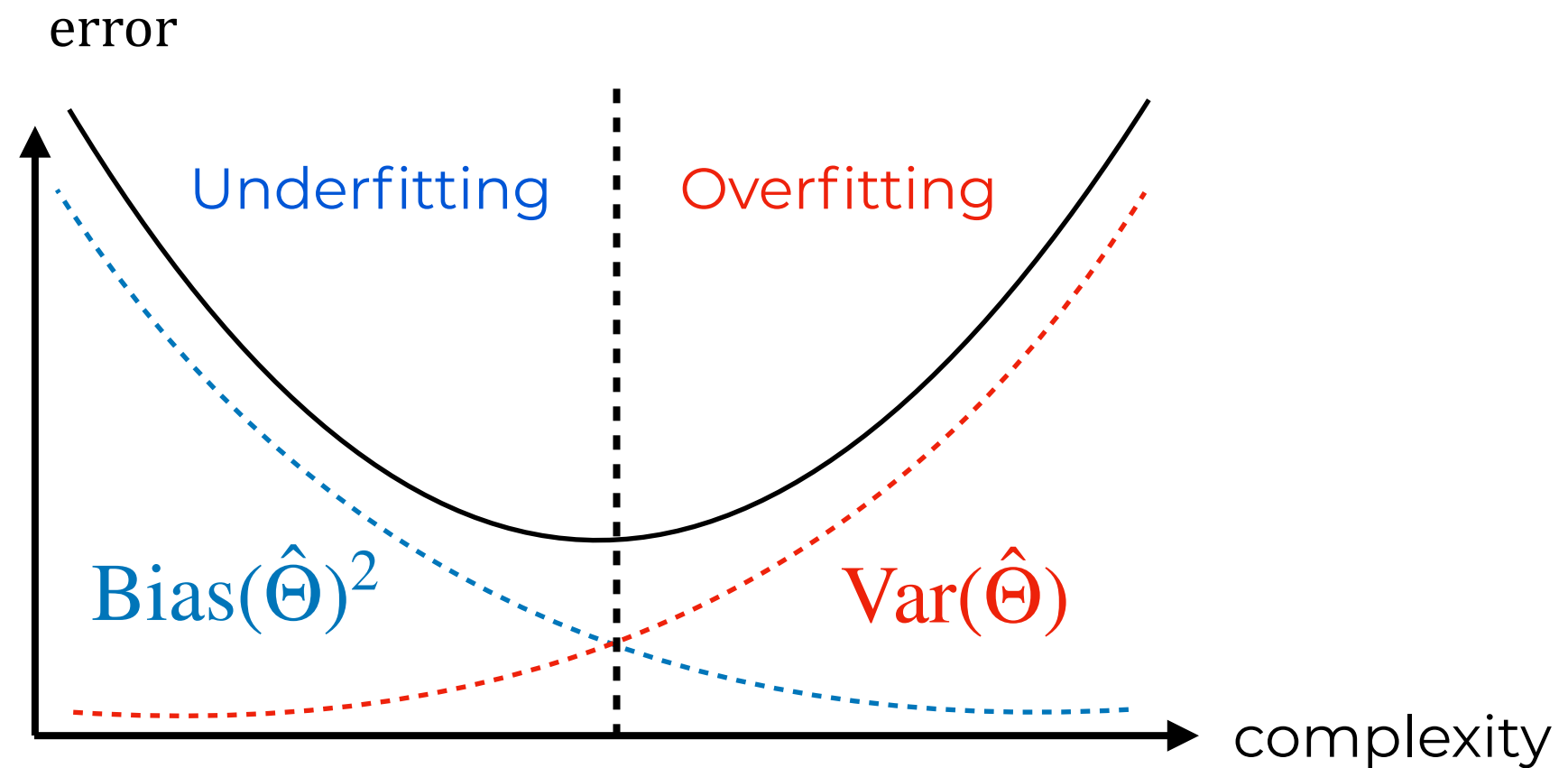


Bias-variance trade-off



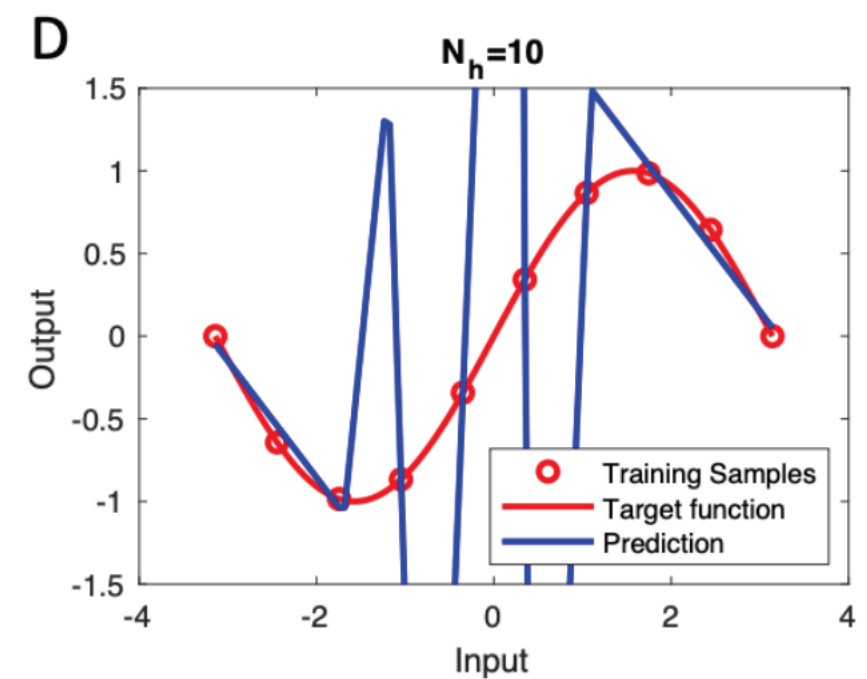
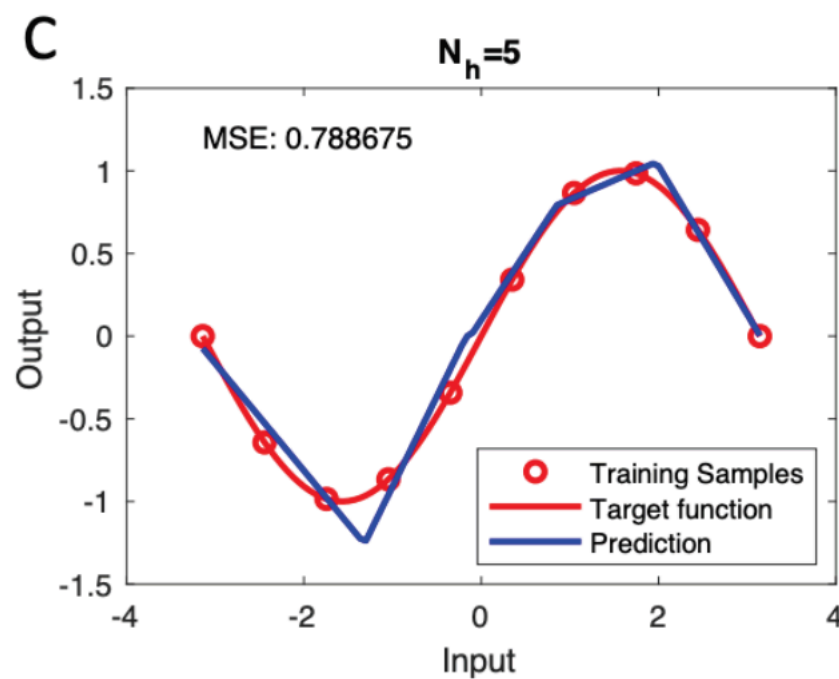
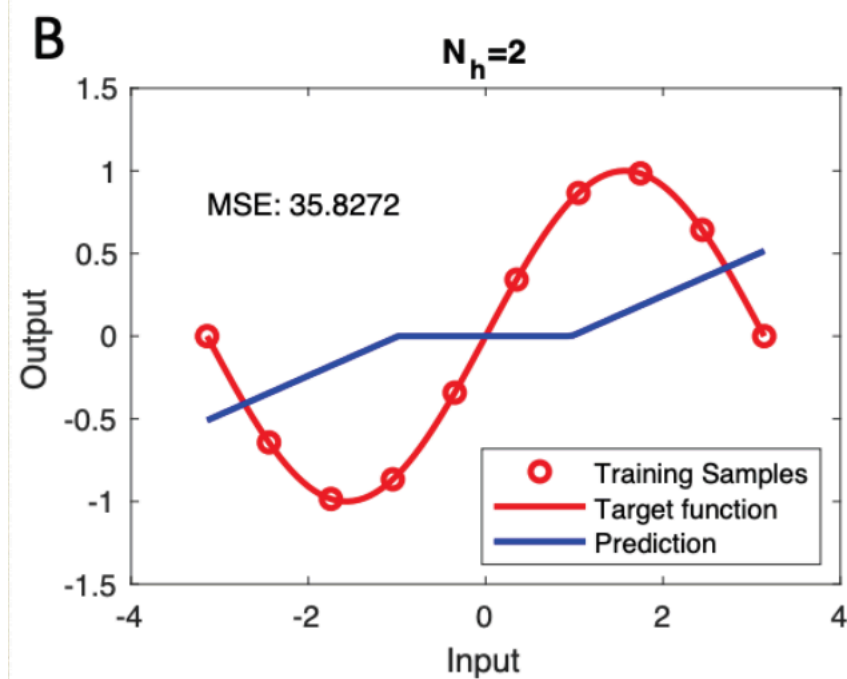
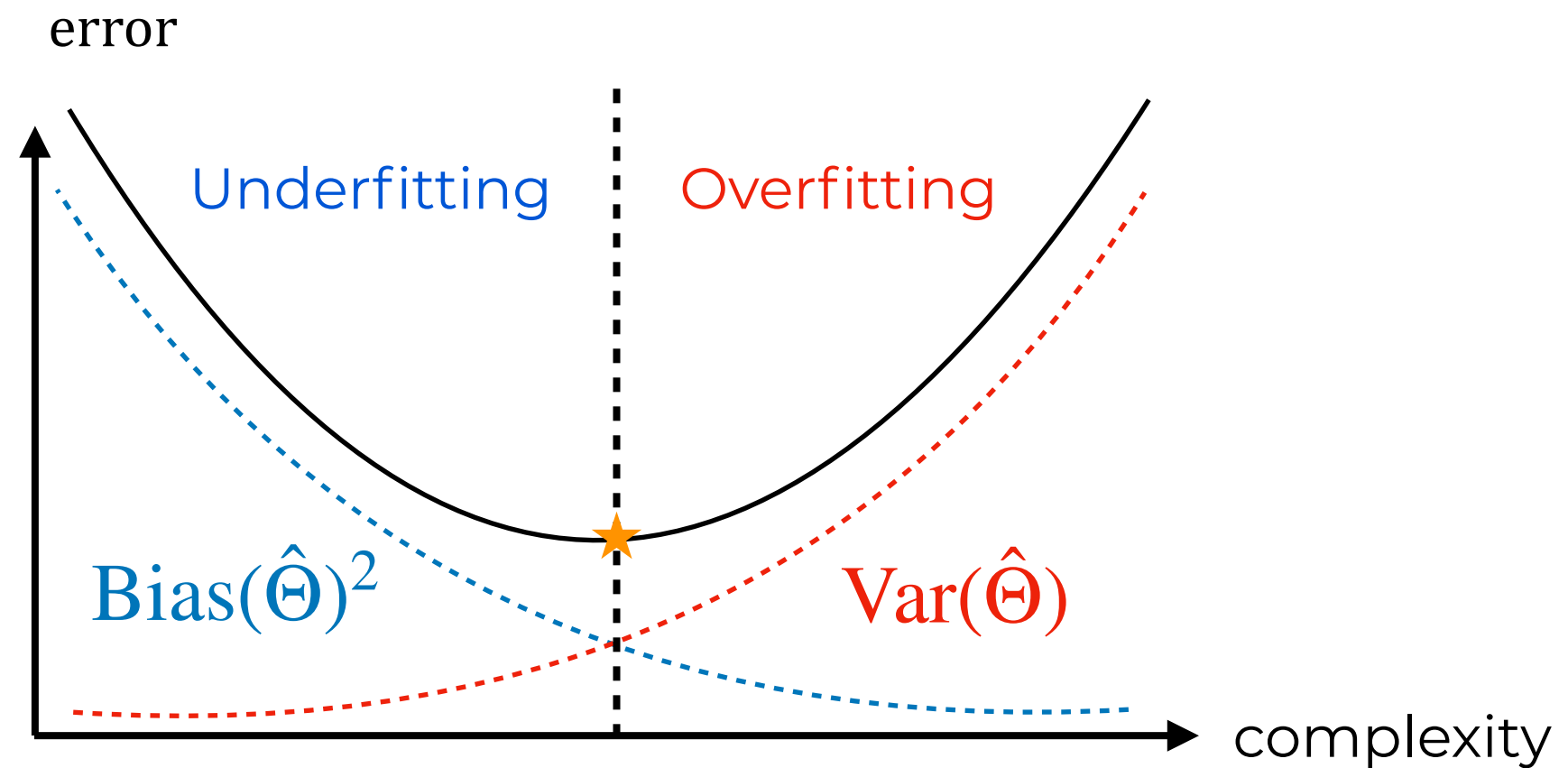
From [Advani, Saxe 17']

Bias-variance trade-off



From [Advani, Saxe 17']

Bias-variance trade-off



From [Advani, Saxe 17']

Bias-variance trade-off

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.

“Double descent” [Belkin '18]

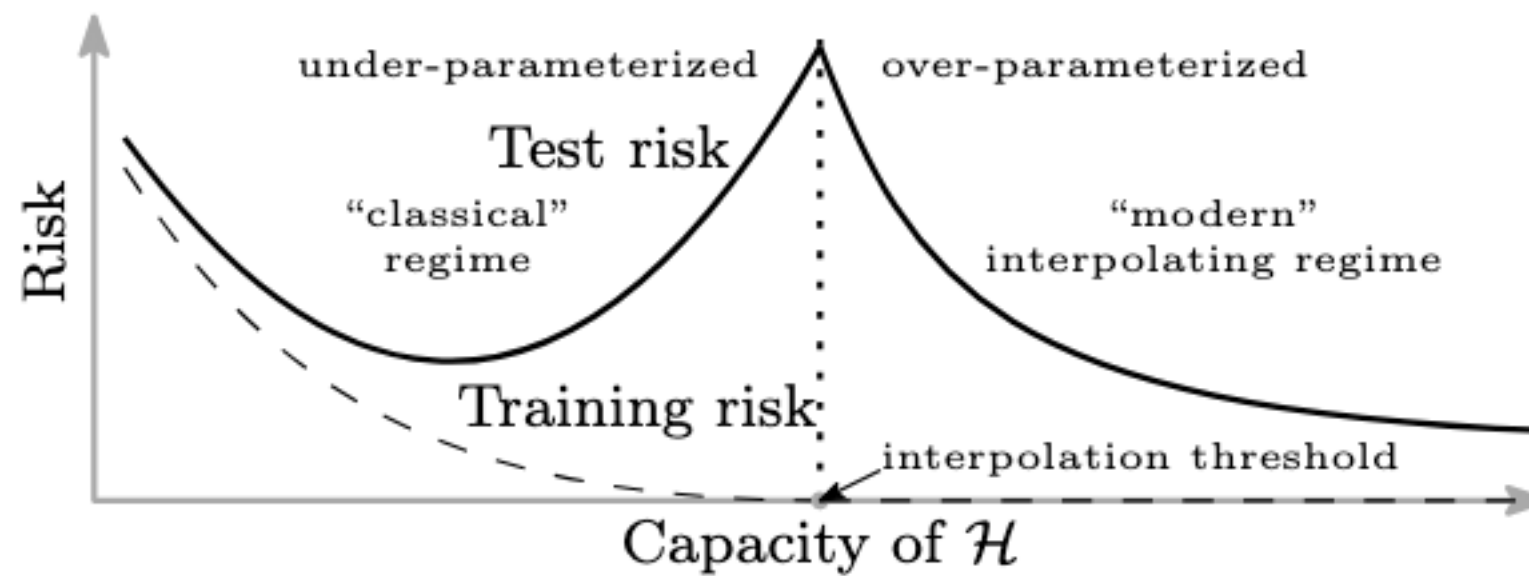
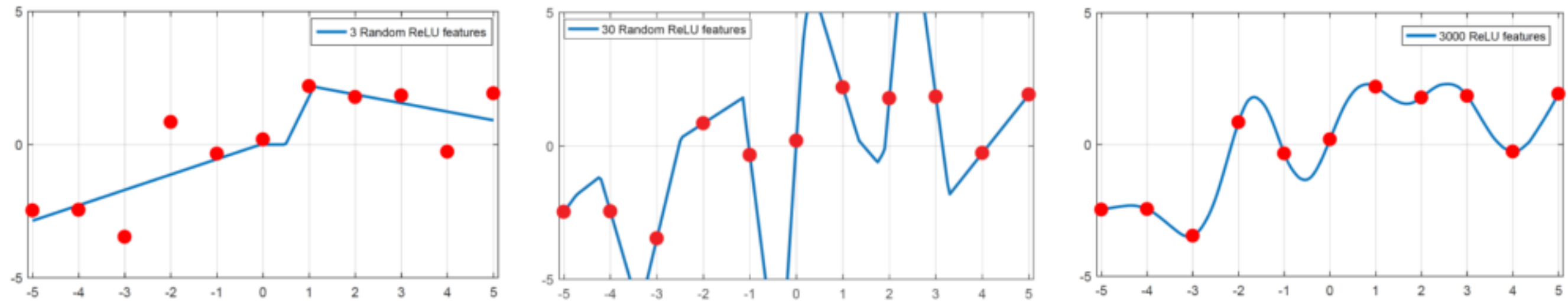
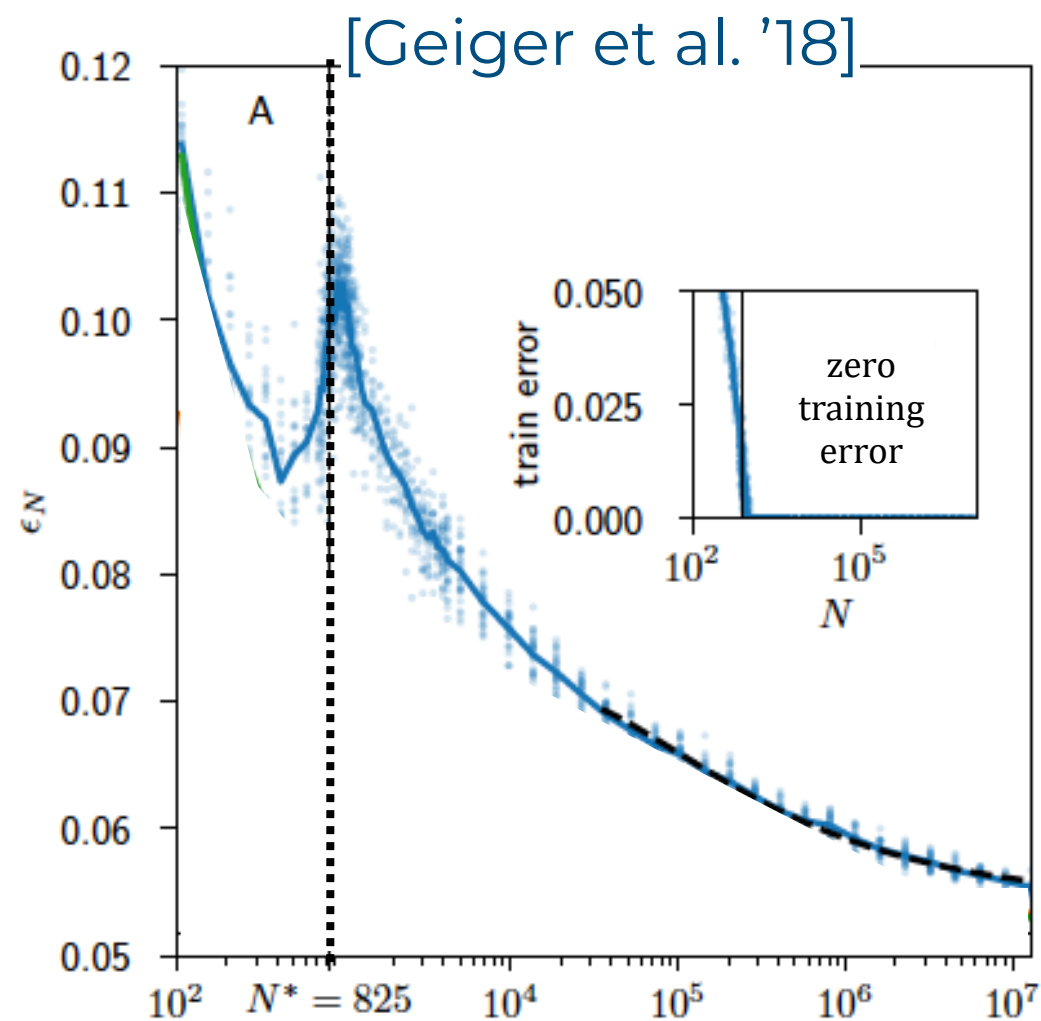


Figure from [Belkin 21']

“Double descent”

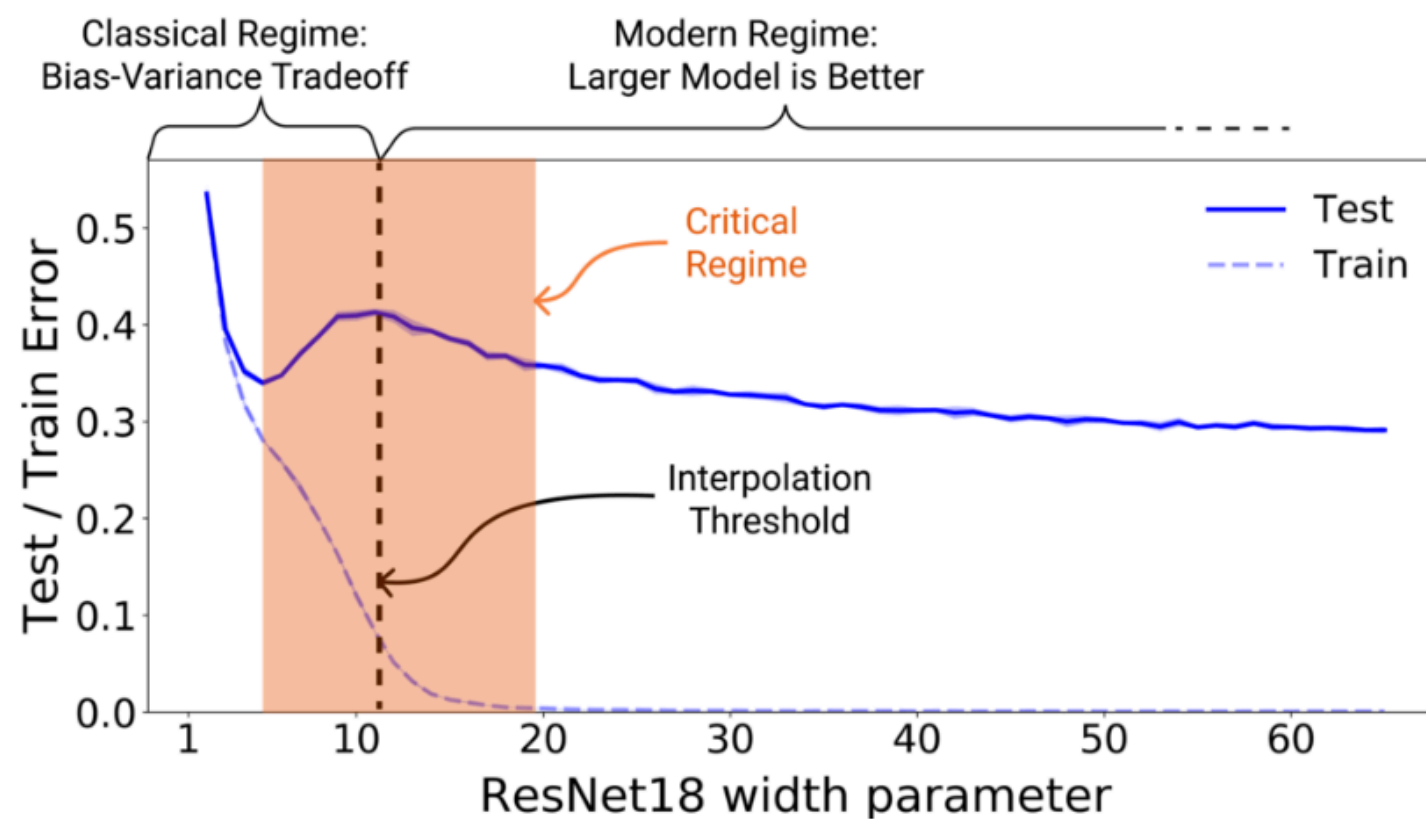
[Belkin '18]



Number of parameters

Parity-MNIST, 5 layers,
fully-connected, no
regularisation

[Nakkiran et al. '19]



CIFAR10, no regularisation

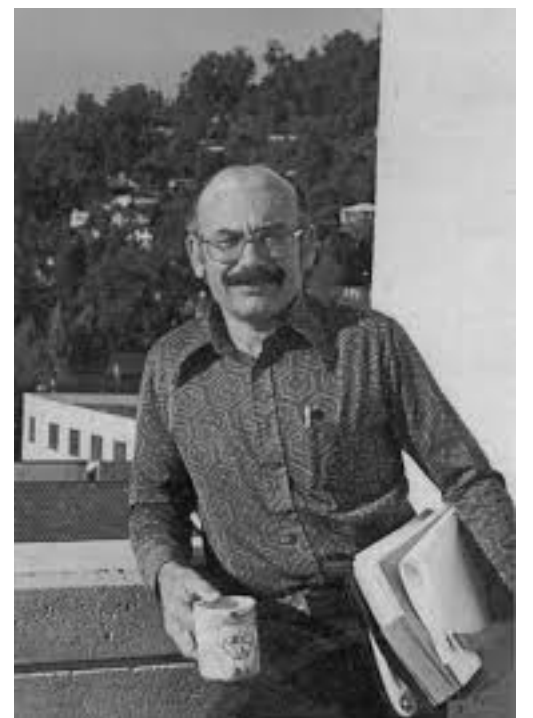
Many questions, few answers

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

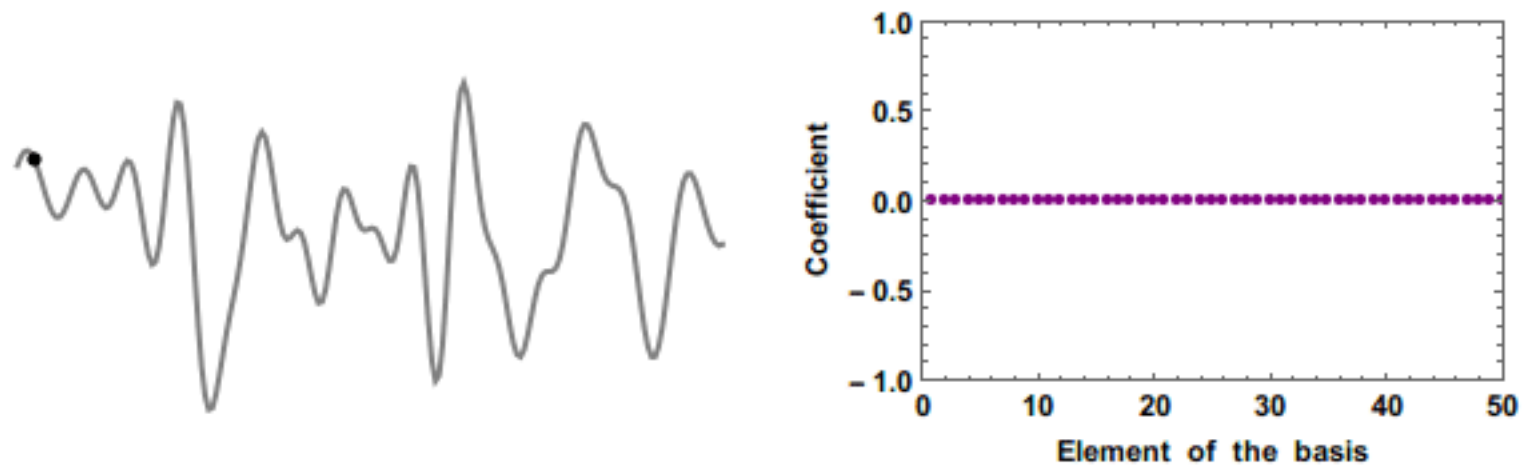
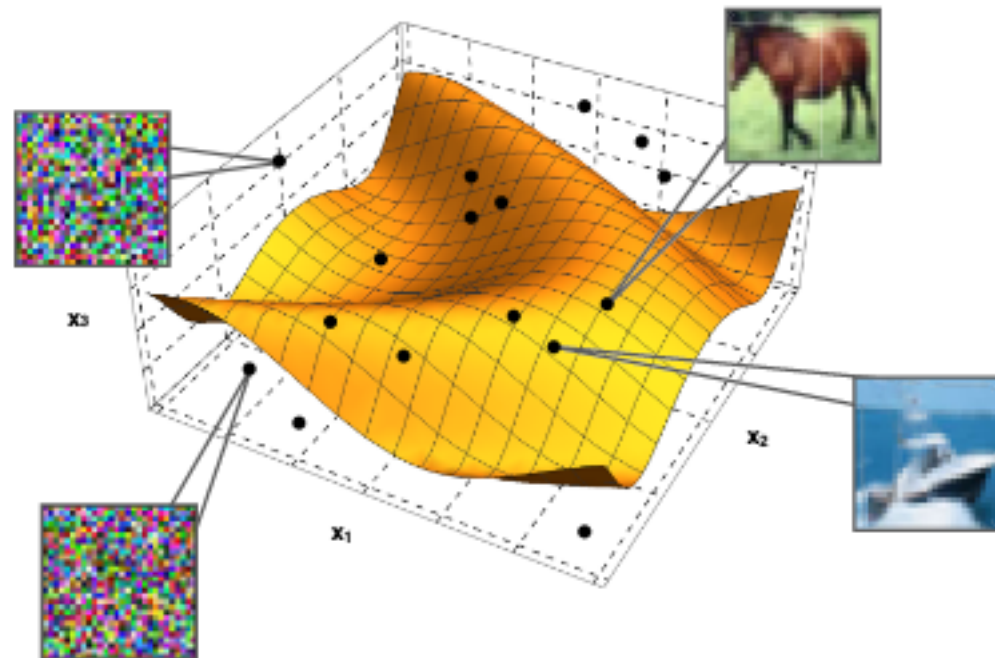
Ronald L. Rivest†
MIT Lab. for Computer Science
Cambridge, Mass. 02139 USA

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?



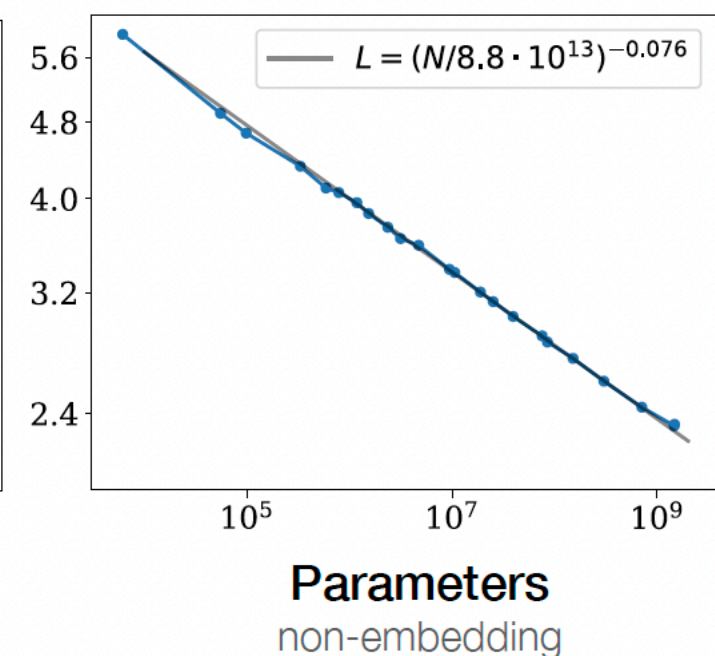
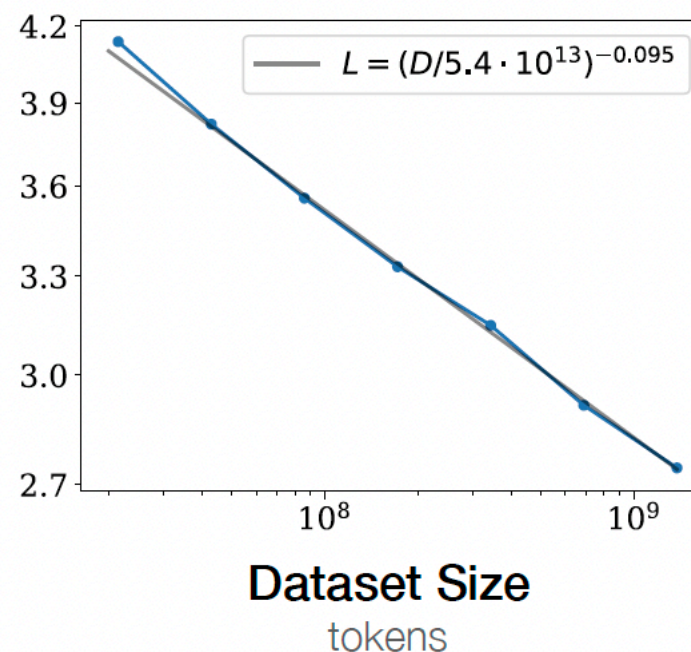
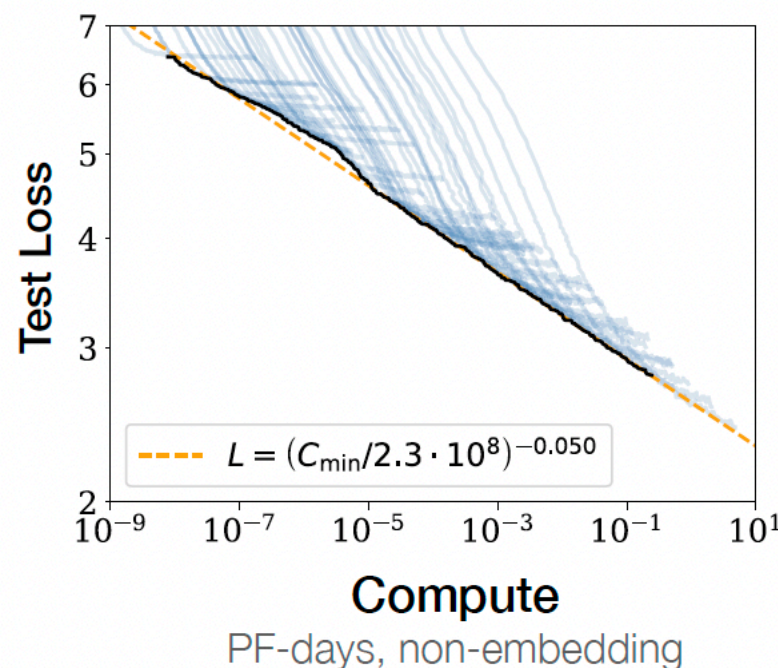
Neural scaling laws

[Kaplan et al. '20]

Scaling Laws for Neural Language Models

Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.



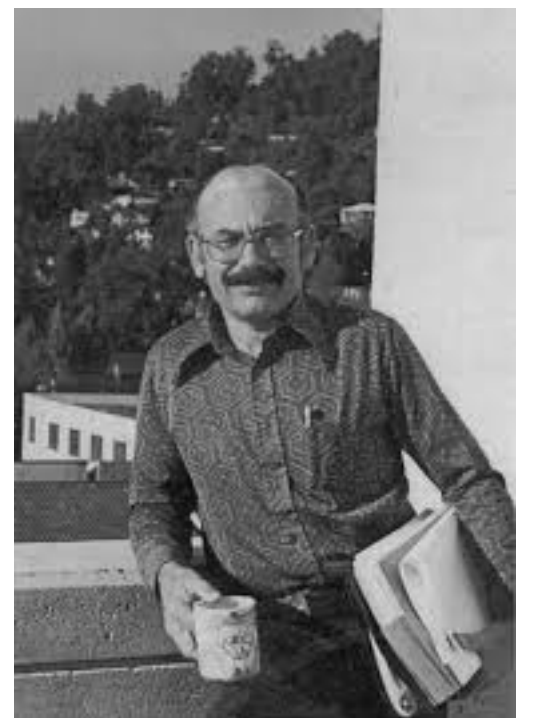
Many questions, few answers

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.

c.f. [Hopfield 1982; Amit, Gutfreund, Sompolinsky 1985]

The capacity problem



Optimal storage properties of neural network models

E Gardner[†] and B Derrida[‡]

[†] Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK

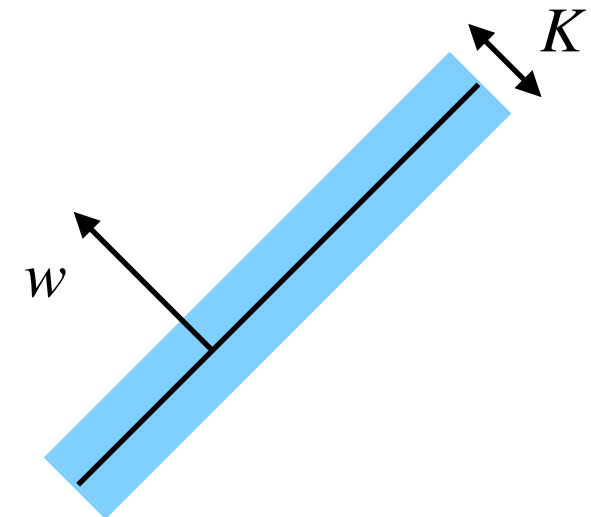
[‡] Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

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.

Given $(x^\nu, y^\nu)_{\nu \in [n]}$,
wants:

$$y^\nu(w^\top x^\nu) \geq K$$



The capacity problem



Optimal storage properties of neural network models

E Gardner[†] and B Derrida[‡]

[†] Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK

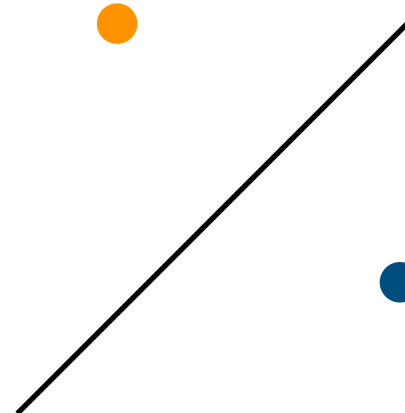
[‡] Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

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.

$$d = 2$$

$$n = 2$$



The capacity problem



Optimal storage properties of neural network models

E Gardner[†] and B Derrida[‡]

[†] Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK

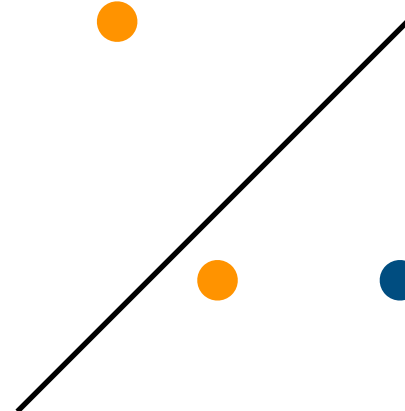
[‡] Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

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.

$$d = 2$$

$$n = 3$$



The capacity problem



Optimal storage properties of neural network models

E Gardner[†] and B Derrida[‡]

[†] Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK

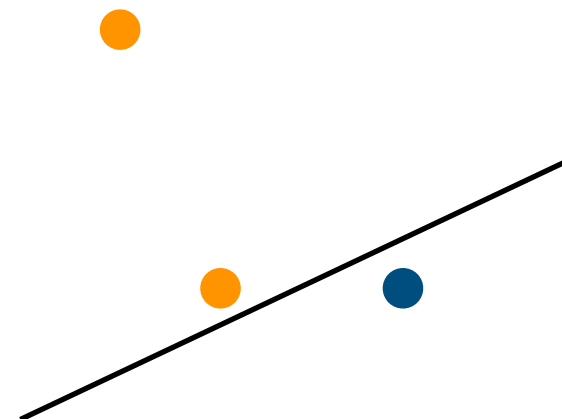
[‡] Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

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.

$$d = 2$$

$$n = 3$$



The capacity problem



Optimal storage properties of neural network models

E Gardner[†] and B Derrida[‡]

[†] Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK

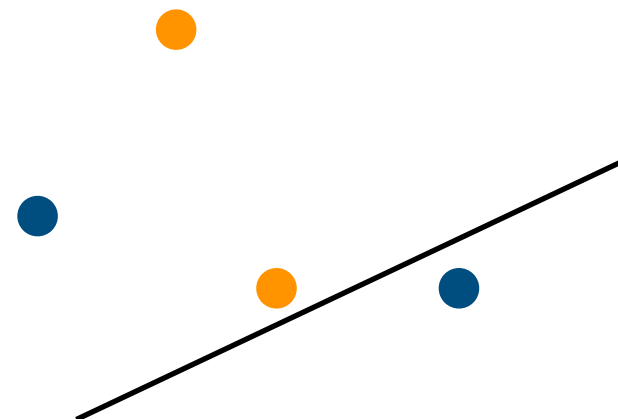
[‡] Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

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.

$$d = 2$$

$$n = 4$$



The capacity problem



Optimal storage properties of neural network models

E Gardner[†] and B Derrida[‡]

[†] Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK

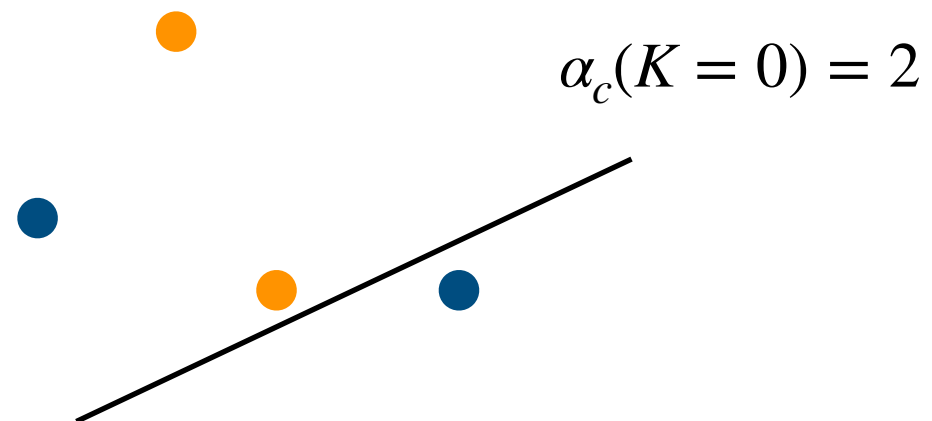
[‡] Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

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.

$$d = 2$$

$$n = 4$$



The capacity problem



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.

$$\hat{\mathcal{R}}_n(\Theta) = \frac{1}{2} \sum_{\mu=1}^n \mathbb{I} [y^\mu \neq \text{sign}(w^\top x^\mu - \kappa)] \quad x^\mu \sim \mathcal{N}(0_d, 1/dI_d)$$

$$\mu_\beta(\Theta) = \frac{e^{-\beta \hat{\mathcal{R}}_n(\Theta)}}{Z_\beta} \quad y^\mu \sim \text{Rad}(1/2)$$

$$w \in \mathbb{S}^{d-1}, \{-1, +1\}^d$$

c.f. CSP, sphere packing, etc.

The capacity problem



Optimal storage properties of neural network models

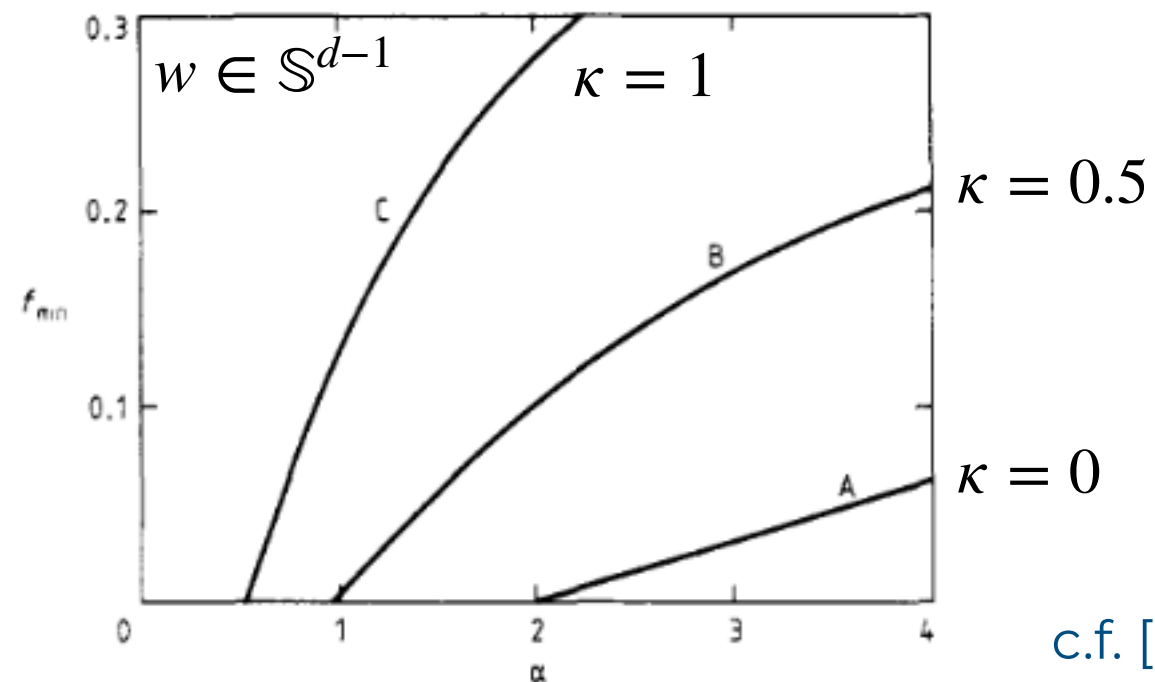
E Gardner[†] and B Derrida[‡]

[†] 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



c.f. [Cover 1967]

The capacity problem



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

Rademacher complexity and spin glasses:

A link between the replica and statistical theories of learning

$$\mathcal{H} = \{f(x; \Theta) = \text{sign}(w^\top x - \kappa) : w \in \mathbb{R}^d, \kappa \geq 0\}$$

$$\text{Rad}(\mathcal{H}) = \frac{1}{2} \left(e_{\text{g.s.}}(\alpha) - 1 \right) = C(\kappa) \sqrt{\frac{d}{n}}$$

[Abbara, Aubin, Krzakala, Zdeborová 2020;
Haussler, Kearns, Oppen, Schapire 1991]

The capacity problem



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

there is a critical line, $\alpha < \alpha_c(K)$ and the 2-point method is symmetric. This line, $\alpha_c(K)$ but symmetry 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.

The capacity problem



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

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.

The capacity problem



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

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 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)

Learning from examples is given. As the number of examples increases, the reference error reaches a transition is first order.

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.

The capacity problem



Optimal storage properties of neural network models

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

Edinburgh, EH9 3JZ, UK
Orsay, France

Michael Biehl
Physikalisches Institut, Julius-Maximilians-Universität, Am Hubland, D-8700 Würzburg, Germany

A summary is presented of the statistical mechanical theory of learning a rule with a neural network, a rapidly advancing area which is closely related to other inverse problems frequently encountered by physicists. By emphasizing the relationship between neural networks and strongly interacting physical systems, such as spin glasses, the authors show how learning theory has provided a workshop in which to develop new, exact analytical techniques.

perceptron-like Neural

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

Learn

A1

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

work of the per-
eters which ren-
s of attraction)
s and study the

a discontinuous transition
of perfect generalization c

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 Cargèse, 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 waned 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 mathematicians) 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>



Disordered Systems and Biological Organization

13	M. MEZARD	
	On the statistical physics of spin glasses.	119
16	J.J. HOPFIELD, D.W. TANK	
	Collective computation with continuous variables.	155
20	M.A. VIRASORO	
	Ultrametricity, Hopfield model and all that.	197
18	G. WEISBUCH, D. d'HUMIERES	
	Determining the dynamic landscape of Hopfield networks.	187
23	L. PERSONNAZ, I. GUYON, G. DREYFUS	
	Neural network design for efficient information retrieval.	227
24	Y. LE CUN	
	Learning process in an asymmetric threshold network.	233
30	D. GEMAN, S. GEMAN	
	Bayesian image analysis.	301

The key idea

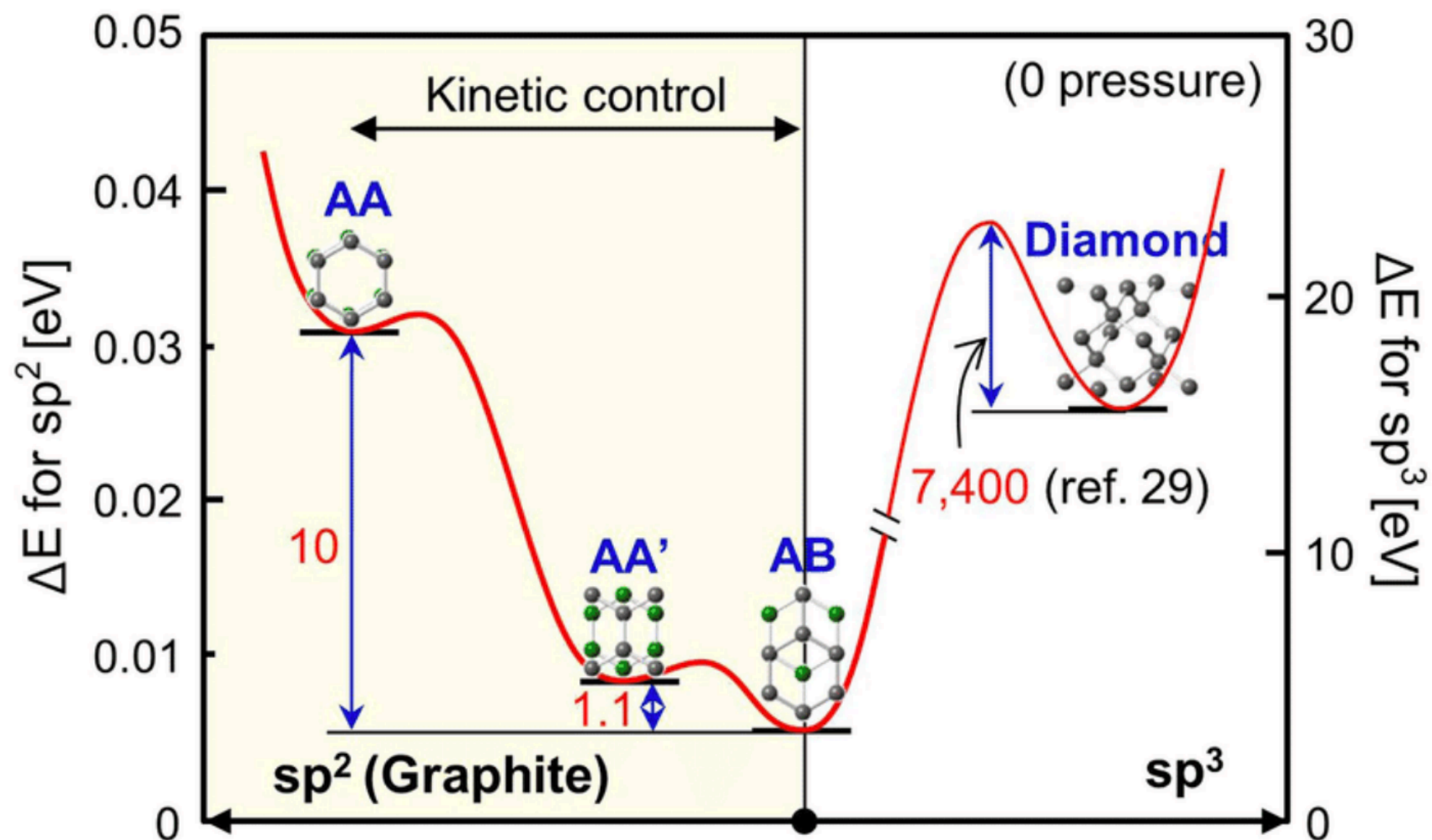
Idea: write this as Stat. Mech. problem

$$\mu_{\beta}(\Theta) = \frac{1}{Z_{\beta}} e^{-\beta H(\Theta)} \quad H(\Theta) = \frac{1}{n} \sum_{\nu \in [n]} \ell(y^{\nu}, f(x^{\nu}; \Theta)) + r(\Theta)$$

The key idea

Idea: write this as Stat. Mech. problem

$$\mu_{\beta}(\Theta) = \frac{1}{Z_{\beta}} e^{-\beta H(\Theta)} \quad H(\Theta) = \frac{1}{n} \sum_{\nu \in [n]} \ell(y^{\nu}, f(x^{\nu}; \Theta)) + r(\Theta)$$

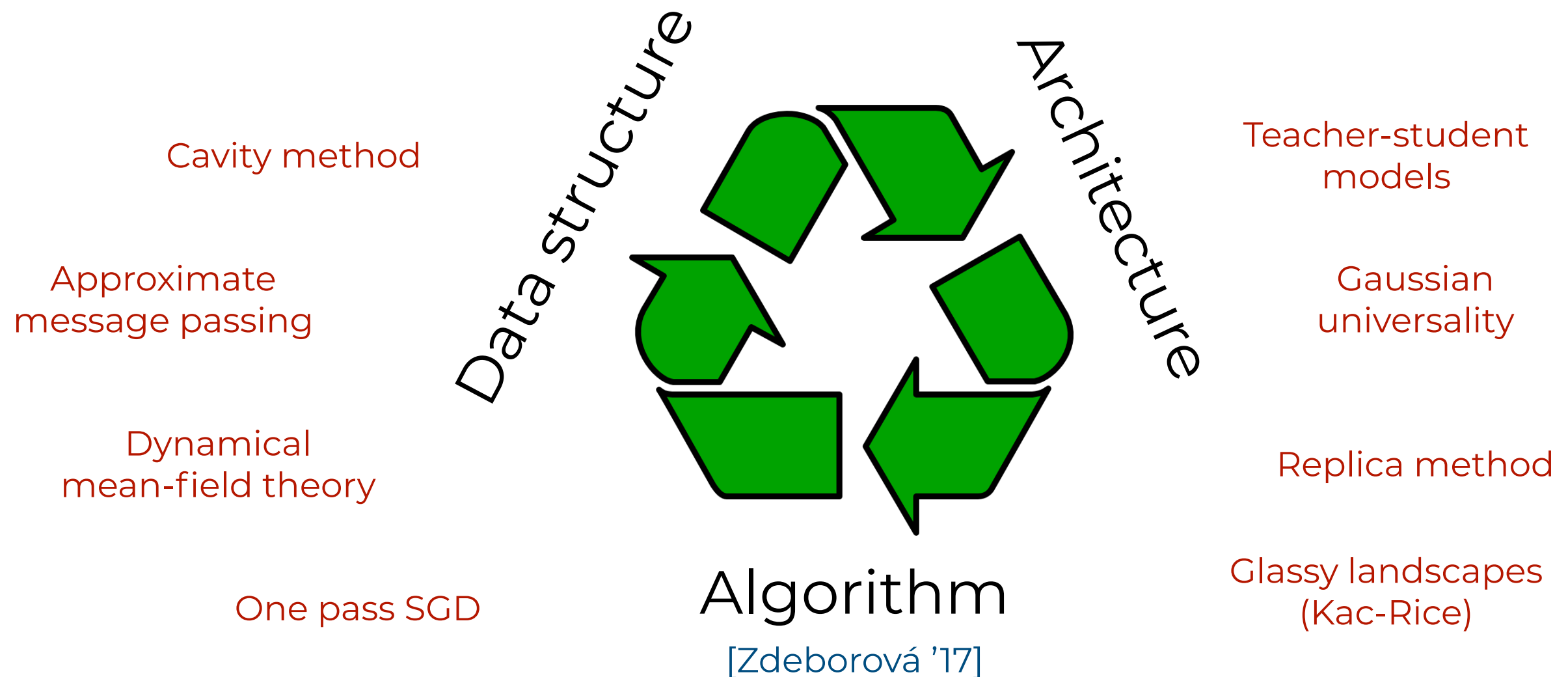


Back to Breiman

“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?



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?

