Deep Learning Theory

An Introduction & Recent Results

Khoat Than
Hanoi University of Science and Technology

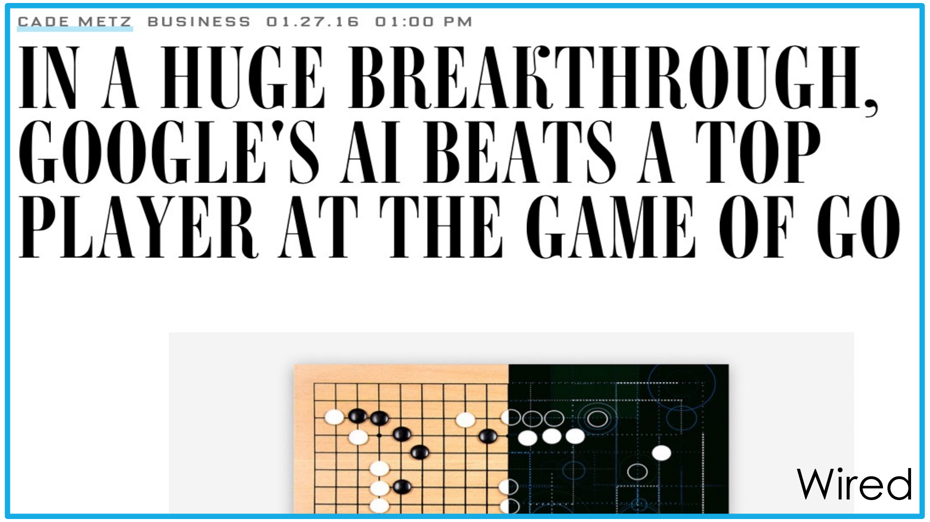
Summer school on Modern Machine Learning, SOICT, 9-2023

Contents

- Recent breakthroughs
- The open theoretical challenge
- Basic concepts in learning theory
- Some theories for deep neural networks

Some successes: AlphaGo (2016)

- AlphaGo of Google DeepMind: the world champion at Go (cờ vây), 3/2016
  - Go is a 2500-year-old game
  - Go is one of the most complex games
- AlphaGo learns from 30 millions human moves and plays itself to find new moves



CADE METZ BUSINESS 01.27.16 01:00 PM

IN A HUGE BREAKTHROUGH, GOOGLE'S AI BEATS A TOP PLAYER AT THE GAME OF GO

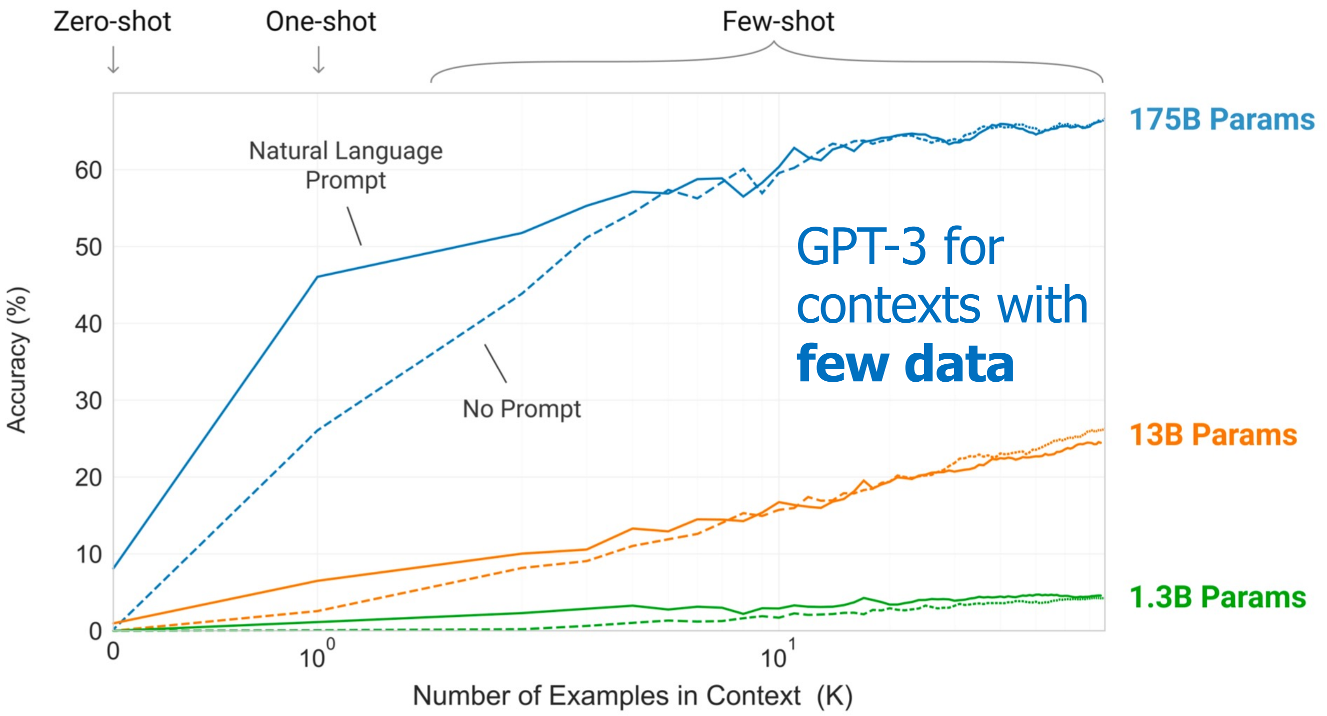
Wired

DeepMind

Some successes: GPT-3 (2020)

- Language generation (writing ability?)
  - A huge model was trained from a huge data set
  - This model, as *universal knowledge,* can be used for many problems

Con người không thể phân biệt bài viết 500 từ là do **máy** hay **người** viết



	Mean accura	95% Confidence Interval (low, hi)
Control	88%	84%–91%
GPT-3 175B	52%	48%–57%

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

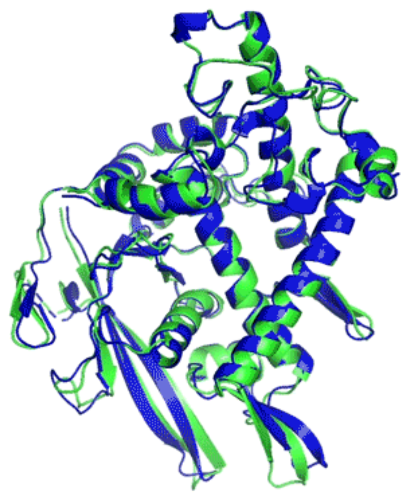
The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

OpenAI

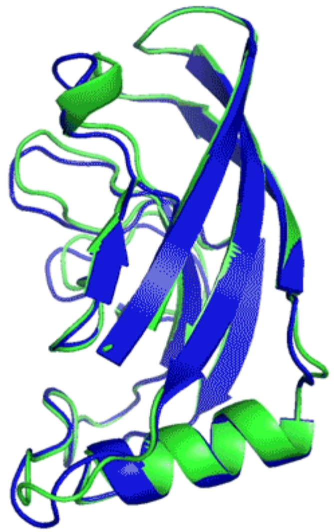
Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." NeurIPS (2020). **Best Paper Award**

Some successes: AlphaFold 2 (2021)

- Accurate prediction of Protein folding



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



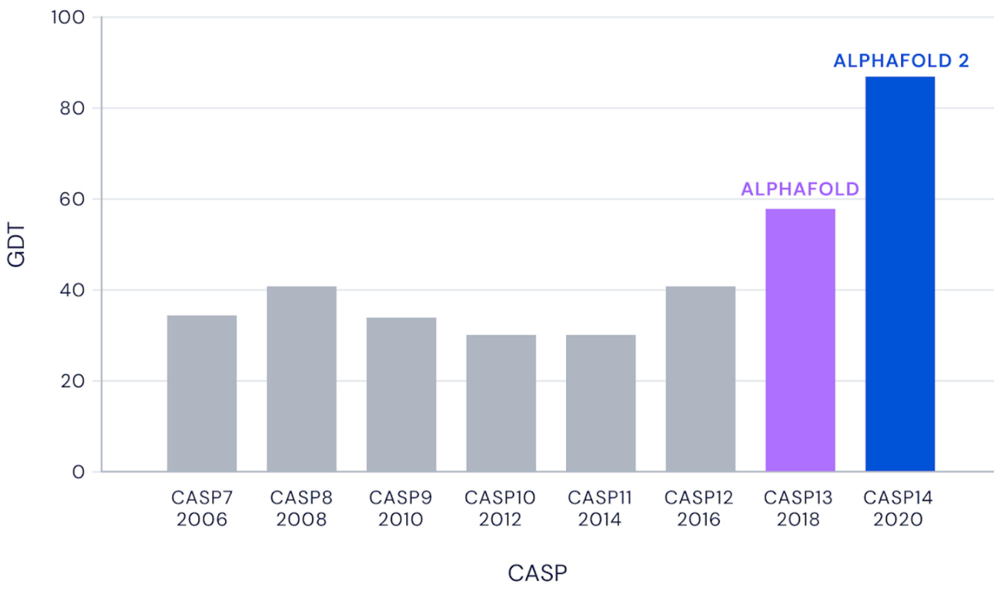
T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

DeepMind

This computational work represents a stunning advance on the protein-folding problem, **a 50-year-old grand challenge in biology.**

– Venki Ramakrishnan, Nobel Laureate



Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021).

Théâtre D'opéra Spatial

An AI-Generated Picture Won an Art Prize

@Jason Allen + Midjourney

https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

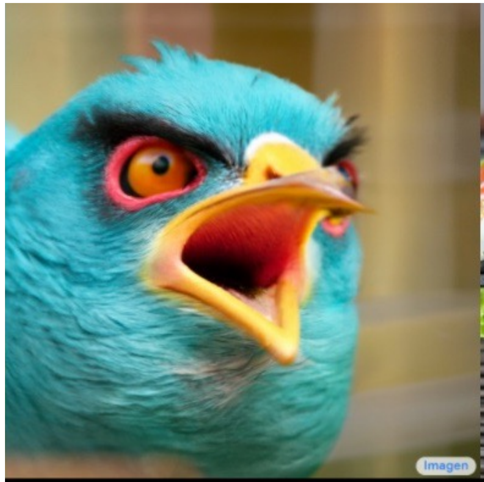
Some successes: Text-to-image (2022)

- Draw pictures by short descriptions

A bowl of soup

Imagen

Google



An extremely angry bird.

A cute corgi lives in a house made out of sushi.

Some successes: ChatGPT (2022)

- Human-level Chatting, Writing, QA,…

SO trời đẹp quá, làm cho tớ bài thơ đi

Tạm biệt mây mù, đón tới ánh sáng
Trời đẹp mới rực, tấm lòng tớ tăng
Sức mạnh của mùa xuân, giữa cánh đồng hoa
Vẻ đẹp tự nhiên, làm cho tớ mãi mãi mê hoặc.

Lấp lánh tràn đầy niềm vui, nụ cười tràn đầy niềm
Không cần bất kỳ điều gì, chỉ cần trời đẹp để tớ h

Why ChatGPT is about to change how you work, like it or not?
- Forbes, 2/2023

CNN BUSINESS Audio Live TV Log In

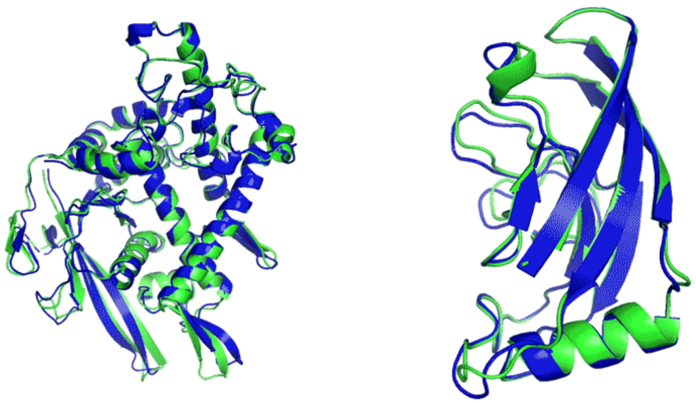
ChatGPT passes exams from law and business schools

By Samantha Murphy Kelly, CNN Business
Updated 1:35 PM EST, Thu January 26, 2023

OpenAI

Open question

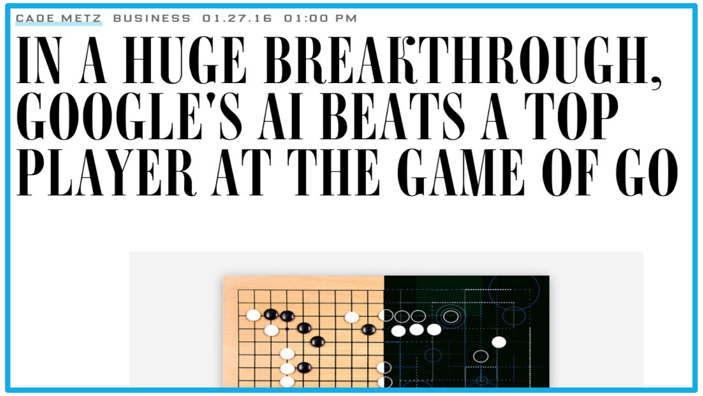
- **Why can deep neural networks perform well?**
  - Many breakthroughs in recognition, games, image synthesis, language generation, Protein folding prediction, …



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction



CADE METZ BUSINESS 01.27.16 01:00 PM

IN A HUGE BREAKTHROUGH, GOOGLE'S AI BEATS A TOP PLAYER AT THE GAME OF GO

Theoretical study

- **Approximation** (power of an architecture)
  - Pros: any continuous function can be approximated well by a deep neural network (NNs)
  - Cons: Unclear how to find a specific NN, based on a given training set
- **Optimization** (learning process)
  - Overparameterized NNs can have zero training error, but do not overfit
  - SGD can find global solutions to the training problems
  - Cons: good optimization does not imply good generalization ability
- **Generalization** (ability of trained NNs to perform on unseen data)
  - Existing standard theories cannot be used, due to vacuousness
  - Some theories work well for only NNs with one-hidden layer

Learning theory

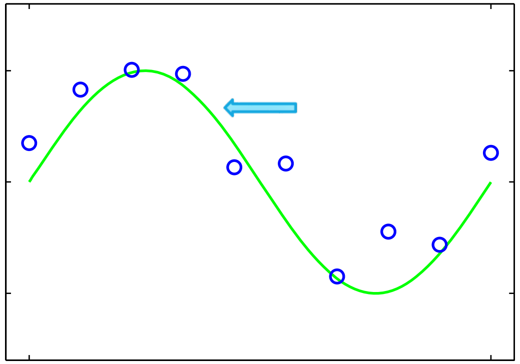
Basic concepts

The learning problem

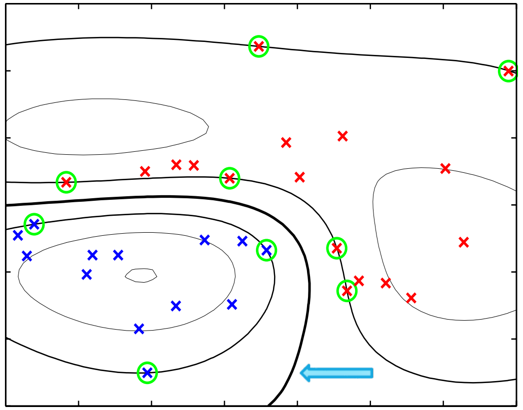
- There is an **unknown** (measurable) function

$$y^*: \mathcal{X} \rightarrow \mathcal{Y}$$

  - It maps each input $\boldsymbol{x} \in \mathcal{X}$ to a label (output) $y \in \mathcal{Y}$
  - Spaces: input space $\mathcal{X}$, output space $\mathcal{Y}$



- We can collect a dataset **D** = {($\mathbf{x}_1$, y_1), ($\mathbf{x}_2$, y_2), …, ($\mathbf{x}_M$, y_M)}
  - $y_i = y^*(\boldsymbol{x}_i)$ for any $i \in \{1, \dots, M\}$
  - Sometimes labels cannot be collected



- We need to find y^* from **D**

- In practice, we often find a function *h* to **approximate** y^*

[Figure by C. Bishop]

Basic concepts

- Loss/cost function:

$$f\colon \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$$

  - $f(y,\hat{y})$: the cost/loss of prediction $\hat{y}$ about y
  - 0-1 loss: $f(y,\hat{y}) = \mathbf{1}_{y\neq\hat{y}}$
  - Square loss: $f(y,\hat{y}) = (y-\hat{y})^2$

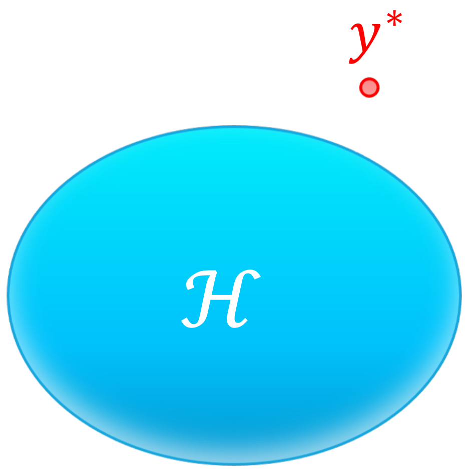
- Empirical loss: the loss of a function h on the training set **D**

$$F(\boldsymbol{D},h) = \frac{1}{M}\sum_{i=1}^{M} f(y_i, h(\boldsymbol{x}_i))$$

- Expected loss (risk): the loss of a function h over the whole space

$$F(P,h) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}[f(y,h(\boldsymbol{x}))]$$

  - *P* is the distribution where each $(\boldsymbol{x},y)$ is sampled



Learning goal

- Function space (*hypothesis space, model space*):
  a set $\mathcal{H}$ of functions, where a learner will select a good function $h \in \mathcal{H}$
  - Depends on input features: $h: \mathcal{X} \to \mathcal{Y}$
  - Represents our prior knowledge about a task
  - E.g. for a linear model:

$$\mathcal{H} = \{w_0 + w_1x_1 + \ldots + w_nx_n \,|\, \boldsymbol{w} = (w_0, w_1, \ldots, w_n) \in \mathbb{R}^{n+1}\}$$

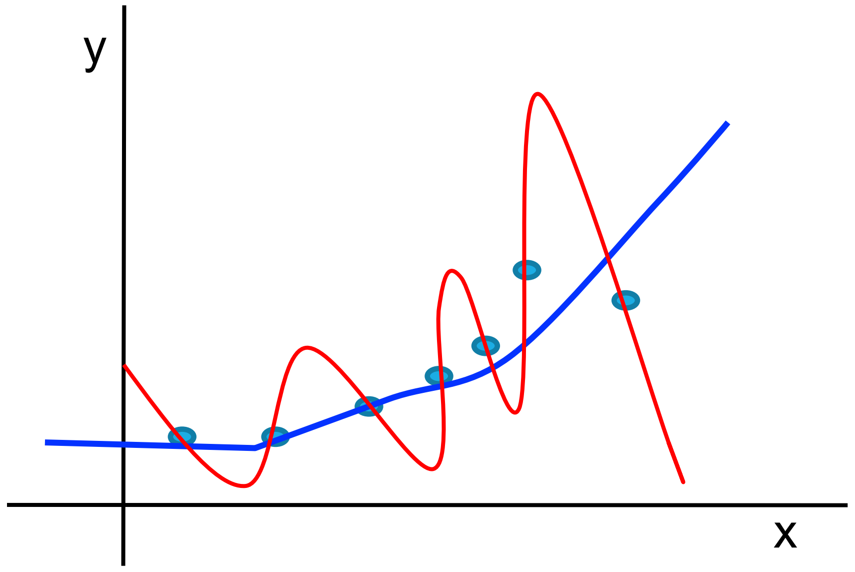
- Learner: a learning algorithm that can select one $h \in \mathcal{H}$
  - Input: a training set **D** and $\mathcal{H}$
- Learning goal: find one $h \in \mathcal{H}$ with *small expected loss*
  - h should generalize well on future data
- Ultimately, we want to find the best one in $\mathcal{H}$: $h^* = \arg\min_{h \in \mathcal{H}} F(P, h)$

Learning vs Fitting

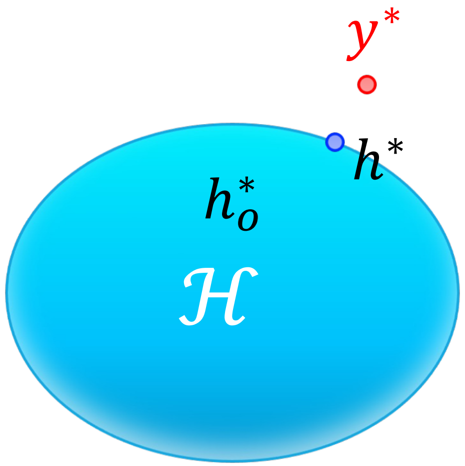
- Fitting:

$$h_o^* = \arg\min_{h \in \mathcal{H}} F(\boldsymbol{D}, h)$$

  - Minimize the training loss $F(\boldsymbol{D}, h)$
  - Focus on "Interpolation"

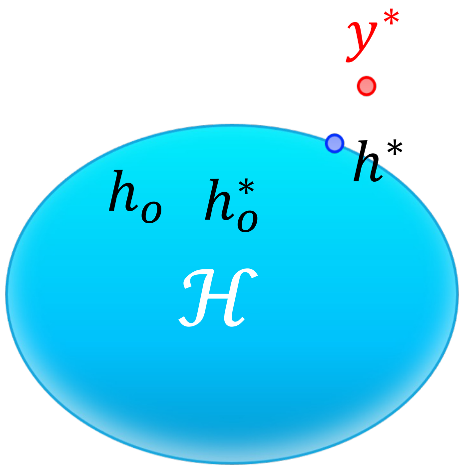


- Learning ≠ Fitting
  - Learning requires the learned model to generalize well on future data
  - Ability of "extrapolation"



Errors of a trained model

- After training, the learning algorithm will return $h_o \in \mathcal{H}$
- How well does it work with future data?
- Maybe: $h_o \neq h_o^*$ and $h_o \neq h^*$
  - $h_o^* = \arg\min_{h \in \mathcal{H}} F(\boldsymbol{D}, h)$ is the *minimizer of the empirical loss*
  - $h^* = \arg\min_{h \in \mathcal{H}} F(P, h)$ is the *best member* of family $\mathcal{H}$



- Note:

$$F(P, h_o) - F(P, y^*) = F(P, h_o) - F(P, h^*) + F(P, h^*) - F(P, y^*)$$

- Estimation error: $F(P, h_o) - F(P, h^*)$
  - How good is the learning algorithm?
- Approximation error: $F(P, h^*) - \cancel{F(P, y^*)}$
  - Capacity (representational power) of family $\mathcal{H}$

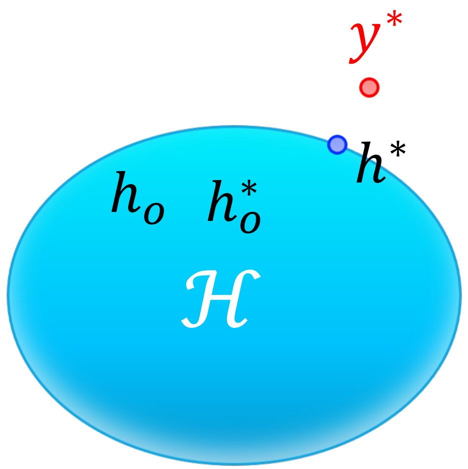
Bousquet et al. Introduction to statistical learning theory. In Machine Learning, LNAI, volume 3176. Springer, 2004.

Error decomposition

- Estimation error

$$|F(P, h_o) - F(P, h^*)| \leq |F(\boldsymbol{D}, h_o) - F(\boldsymbol{D}, h_o^*)| + 2 \sup_{h \in \mathcal{H}} |F(P, h) - F(\boldsymbol{D}, h)|$$

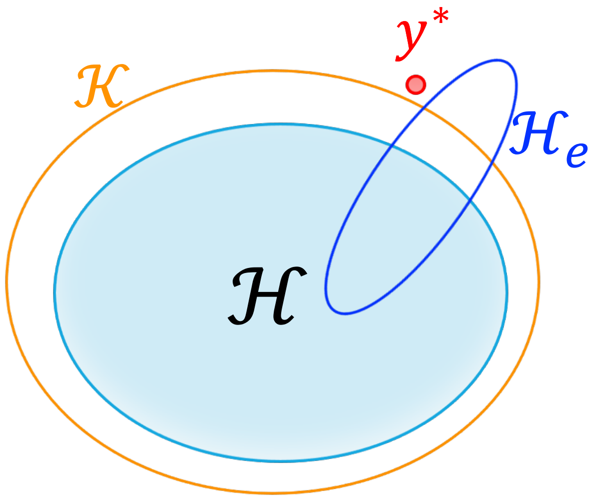
  - It can be decomposed into two types of error
- Optimization error: $F(\boldsymbol{D}, h_o) - F(\boldsymbol{D}, h_o^*)$
  - How close to optimality is h_o?
  - h_o may not be the global solution to the training problem
- Generalization error: $F(P, h_o) - F(\boldsymbol{D}, h_o)$
  - How far is the training loss from expected loss?
- In summary:



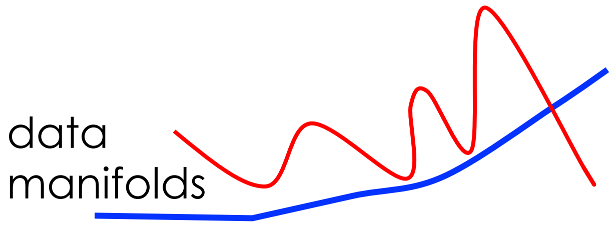
$$Error(h_o) := \text{Optimization error} + \text{Generalization error} + \text{Approximation error}$$

Errors by different factors

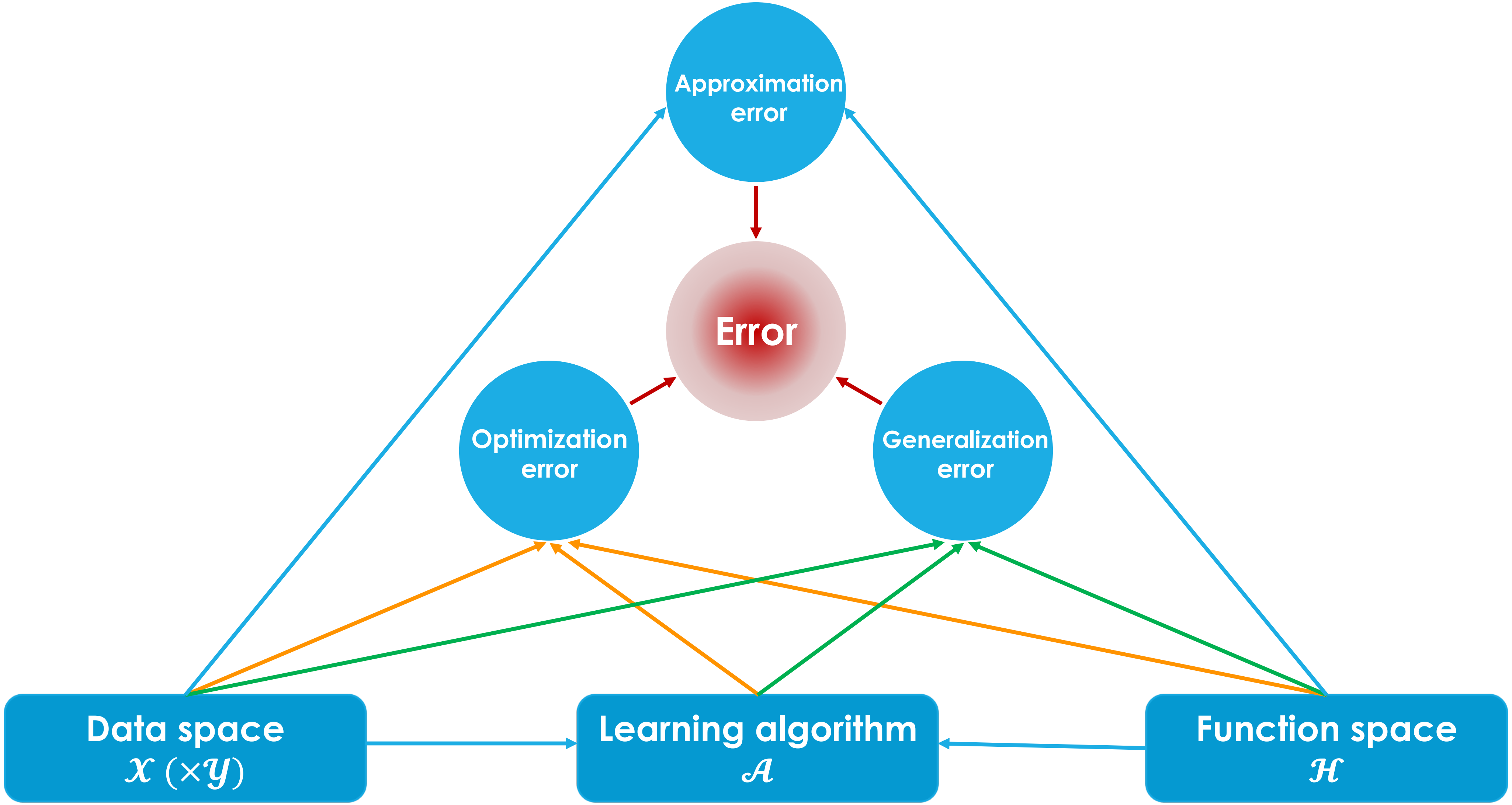
- **Function space $\mathcal{H}$**
  - A bigger space ($\mathcal{K}$), the (probably) smaller approximation error
  - More complex members, the (probably) smaller approximation error ➔ larger capacity
  - An *effective* space ($\mathcal{H}_e$) is enough ➔ not too big/complex



- **Training algorithm $\mathcal{A}$**
  - A better $\mathcal{A}$ implies smaller estimation error of the trained model
  - A **bad** $\mathcal{A}$ can provide small optimization error, but large generalization error ➔ overfitting
  - A good $\mathcal{A}$ can localize an *effective* subset $\mathcal{H}^* \subset \mathcal{H}$
- **Data**
  - Complexity of the data space
  - Representativeness of the training samples, …



A unified view



Bounding the error

- Study upper (and lower) bounds for the errors
- Approximation error:

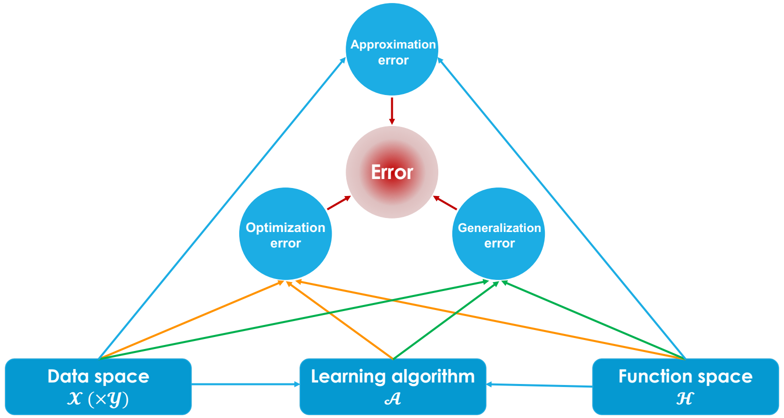
$$|F(P, y^*) - F(P, h^*)| \leq \epsilon_a$$

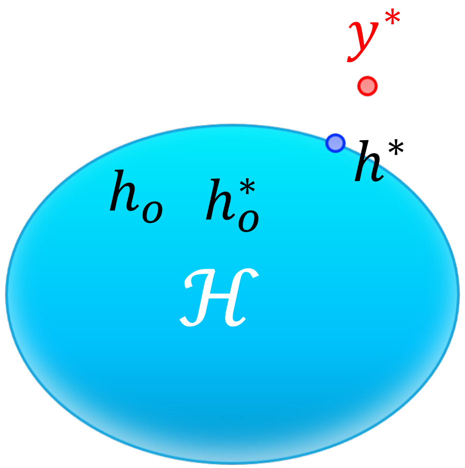
  - Capacity of family $\mathcal{H}$
  - The ability of $\mathcal{H}$ to approximate function y^*

- *Optimization error:*

$$|F(\boldsymbol{D}, h_o^*) - F(\boldsymbol{D}, h_o)| \leq \epsilon_o$$

  - Depending on the number of training iterations (epochs)
  - Capacity of learning algorithm $\mathcal{A}$





Bousquet et al. Introduction to statistical learning theory. In Machine Learning, LNAI, volume 3176. Springer, 2004.

Bounds on Generalization Error

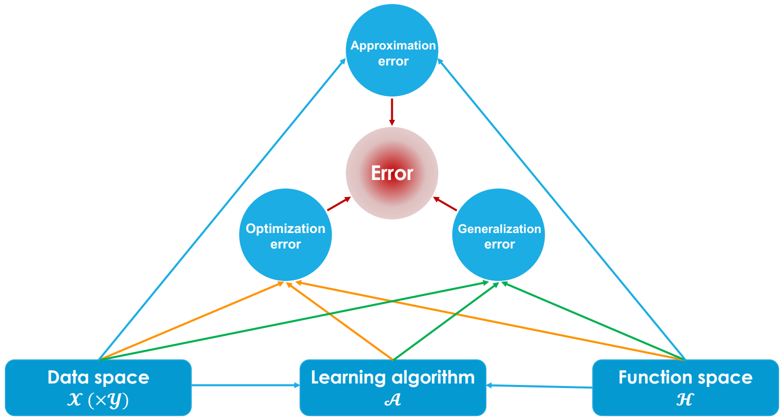
$$|F(P, h_o) - F(\boldsymbol{D}, h_o)| \leq \epsilon_g$$

- Generalizability of a learned function h_o

- Uniform bounds:

$$\sup_{h \in \mathcal{H}} |F(P, h) - F(\boldsymbol{D}, h)| \leq \epsilon_g$$

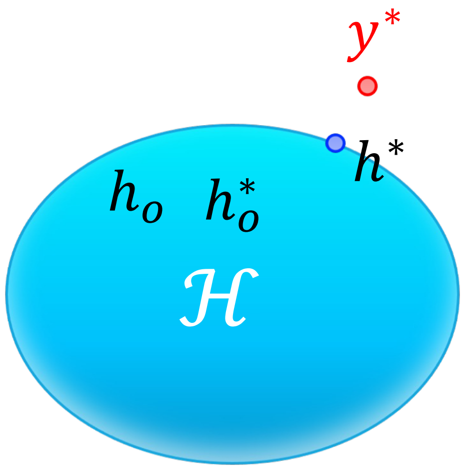
  - Generalizability of the worst member
  - May not be a good way to explain a learned function h_o



- PAC-Bayes bounds:

$$|\mathbb{E}_{h \in \mathcal{H}}[F(P, h) - F(\boldsymbol{D}, h)]| \leq \epsilon_g$$

  - Study the error on average over $\mathcal{H}$
  - See the goodness on average over the model family
  - May not explain a learned function h_o



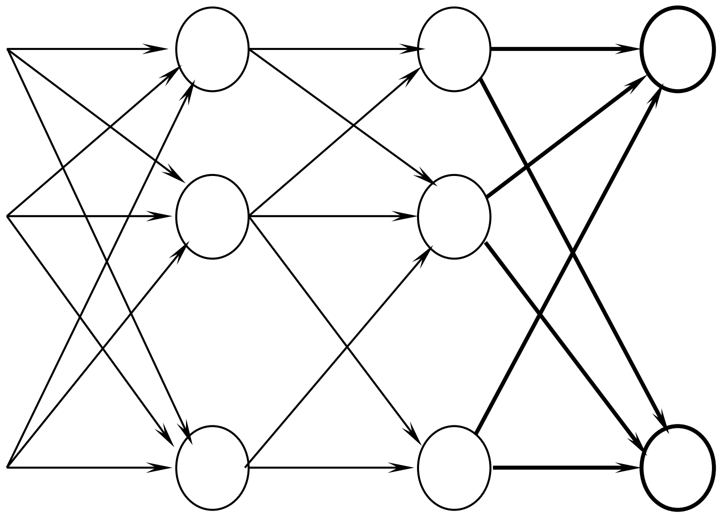
Nagarajan & Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*. 2019.

Theoretical results for deep neural networks

A short summary

Neural network

- Artificial neural networks (ANN):
  - Biologically inspired by human brain
  - A rich family to represent complex functions
- An ANN:
  - Consists of many neurons, organized in a layer-wise manner
  - Each *neuron* computes a simple function
  - A neuron can have few *connections* to other neurons
- Each configuration about #neurons, #layers, #connections, … ➔ an architecture
- Shallow vs. Deep NNs:
  - One hidden layer >< many hidden layers



Mathematical description

$$h(\boldsymbol{x}, \boldsymbol{W}) = g_K(\boldsymbol{W}_K h_{K-1}), \qquad \text{where} \;\; h_i = g_i(\boldsymbol{W}_i h_{i-1}), \qquad h_0 = \boldsymbol{x}$$

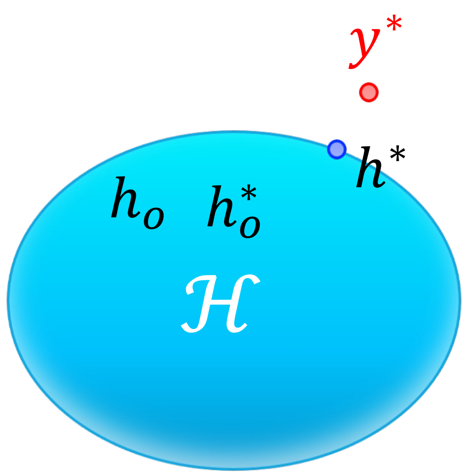
(feedforward network)

- An NN with K layers
- $\boldsymbol{W}_i$ is the weight matrix at layer i
- h_i is the output of layer i
- g_i is the activation function at layer i

- A NN maps an input $\boldsymbol{x}$ to an output $y = h(\boldsymbol{x}, \boldsymbol{W})$
- Function space:

$$\mathcal{H} = \{h(\boldsymbol{x}, \boldsymbol{W}) \mid \boldsymbol{W}_1, \ldots \boldsymbol{W}_K \text{ are real matrices}\}$$

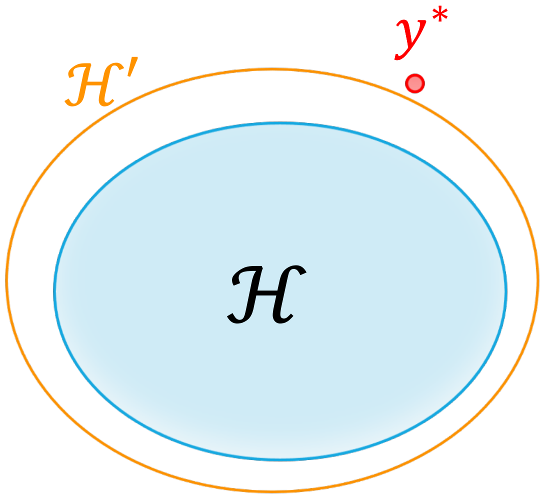
- *Training:* often find weights **W**, by minimizing a loss $F(\boldsymbol{D}, h)$



$Error(h_o) \approx$ Optimization error +Generalization error +Approximation error

Approximation error: classical

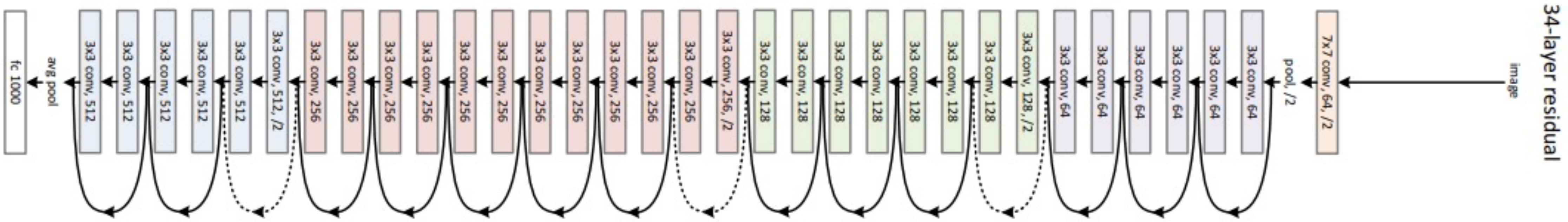
$$\|y^* - h\| \leq \epsilon_a$$

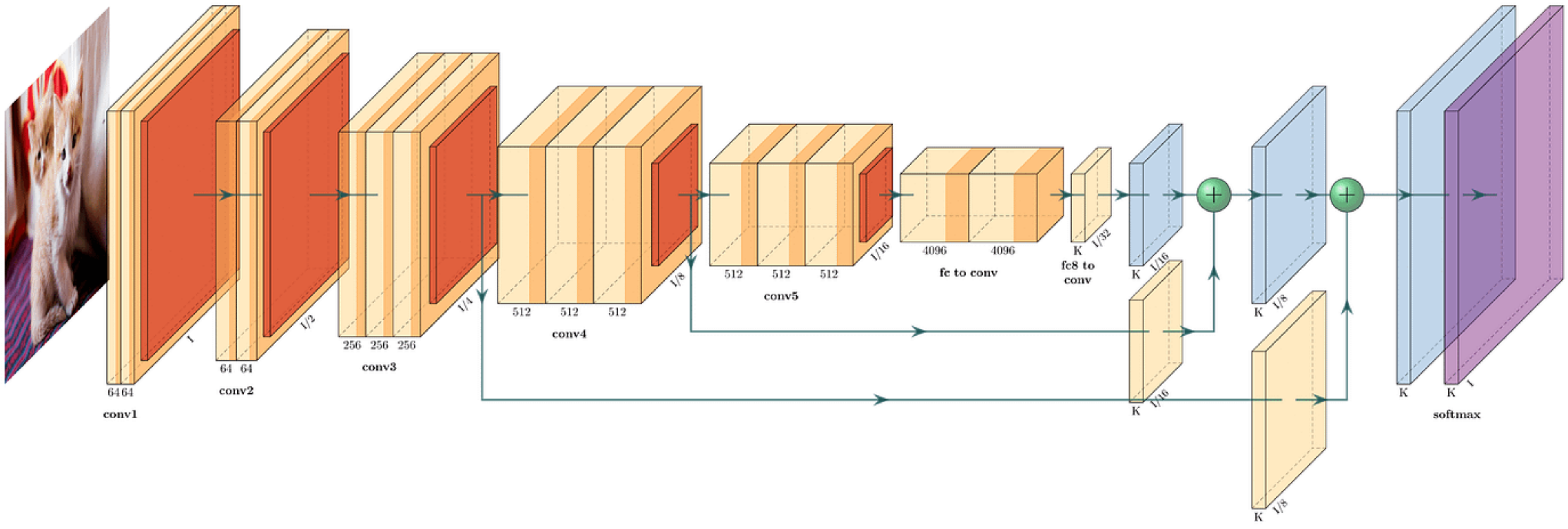


- Increase capacity ➔ approximate better
  - Larger family $\mathcal{H}'$
  - More complex NNs ➔ stronger representational power
  - E.g., wider or deeper NNs
- Any binary function can be learnt (approximately well) by a feedforward network using one hidden layer, when the **width goes to infinity**
- Any bounded continuous function can be learnt (approximately) by a *feedforward network* using one hidden layer [Cybenko, 1989; Hornik, 1991]

Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*.
Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257.

Approximation error: modern





Approximation error: modern

- Any continuous function can be approximated arbitrarily well by Convolutional neural network, when the depth is large [Zhou, 2020]
- Any Lebesgue-integrable function can be approximated arbitrarily well by a ResNet wit
- **Deep** NNs a ating Lipschitz functions [Po
  - Shallow NNs cannot
  - To approximate a Lipschitz function (mapping $[0,1]^n$ to $\mathbb{R}$) with error $O(N^{-\sqrt{L}})$, width $\max\{n, 5N+13\}$ and depth $64nL+3$ are sufficient

Universal approximators

Lin, H., & Jegelka, S. (2018). ResNet with one-neuron hidden layers is a universal approximator. *NeurIPS*.
Lu, J., Shen, Z., Yang, H., & Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*.
Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., & Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*.
Zhou, D. X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*.

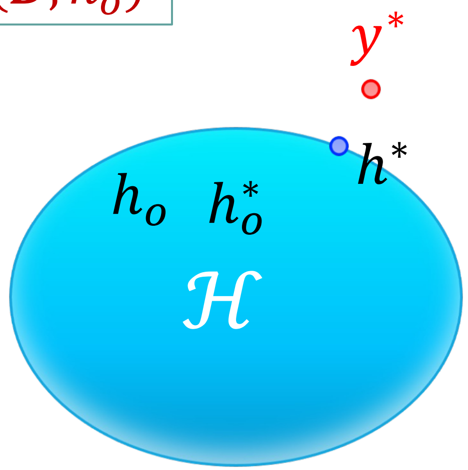
Approximation: existence $\nrightarrow$ method

Unclear
how to find such DNNs,
based on a training set

Optimization error

- Training is often by minimizing a loss $F(\boldsymbol{D}, h)$
- The training loss is *highly non-convex*
- Theory:
  - Exponentially large number of iterations may be needed
  - Intractable in the worst case [Nesterov, 2018]
- Practice:
  - Often have zero training error ➔ global solution h_o^*?
  - Easily perfectly fit random labelling of data [Zhang et al. 2021] (training seems to be easy!)
- Contradiction? What's missing?

$\boxed{F(\boldsymbol{D}, h_o) - F(\boldsymbol{D}, h_o^*)}$



Nesterov, Y. (2018). *Lectures on convex optimization*. Springer.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*.

Optimization: theoretically easy

- Gradient descent (GD) achieves zero training loss in polynomial time for a deep over-parameterized ResNet [Du et al. 2019]
  - Over-parameterization: #parameters ≫ training size
- GD can find a global optimum when the width of the last hidden layer of an MLP exceeds the number of training samples [Nguyen, 2021]
- Stochastic gradient descent (SGD) can find global minima on the training objective of DNNs in polynomial time [Allen-Zhu et al. 2019]
  - Architecture: MLP, CNN, ResNet

Du, S., Lee, J., Li, H., Wang, L., & Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*.
Nguyen, Q. (2021). On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning*.
Allen-Zhu, Z., Li, Y., & Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*.

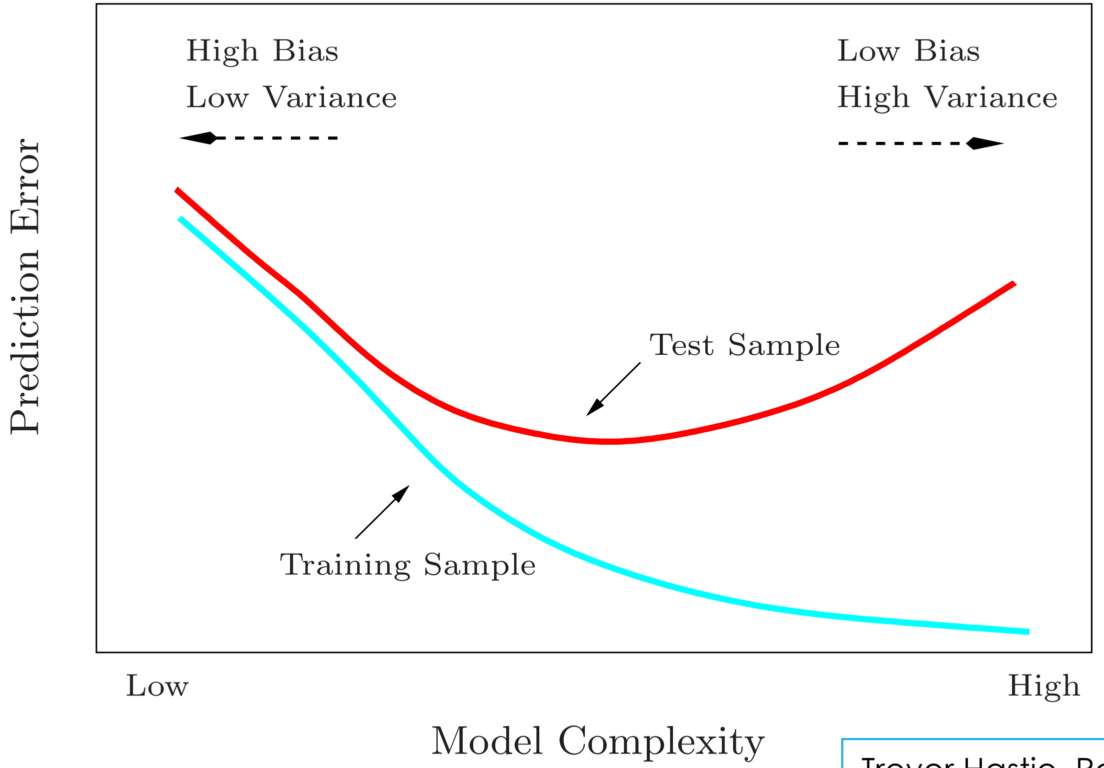
Optimization: reminder

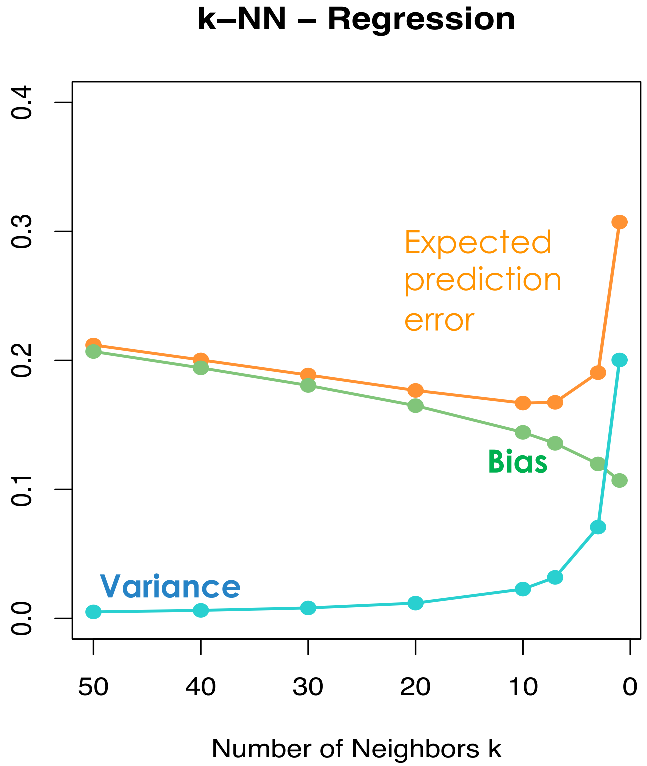
However

global optimality
of the training problem
does not imply
good predictive ability

Bias-Variance tradeoff: classical view

- The more complex the model is, the more data points it can capture, and the lower the bias can be
  - However, higher complexity will make the model "move" more to capture the data points, and hence its variance will be larger.





Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

Bias-Variance: modern behavior

- Modern phenomenon:

  *Very rich models such as DNNs are trained to **exactly fit** the data, but often obtain **high accura**[obscured]**y** on test data* [Belkin et al., 2019]

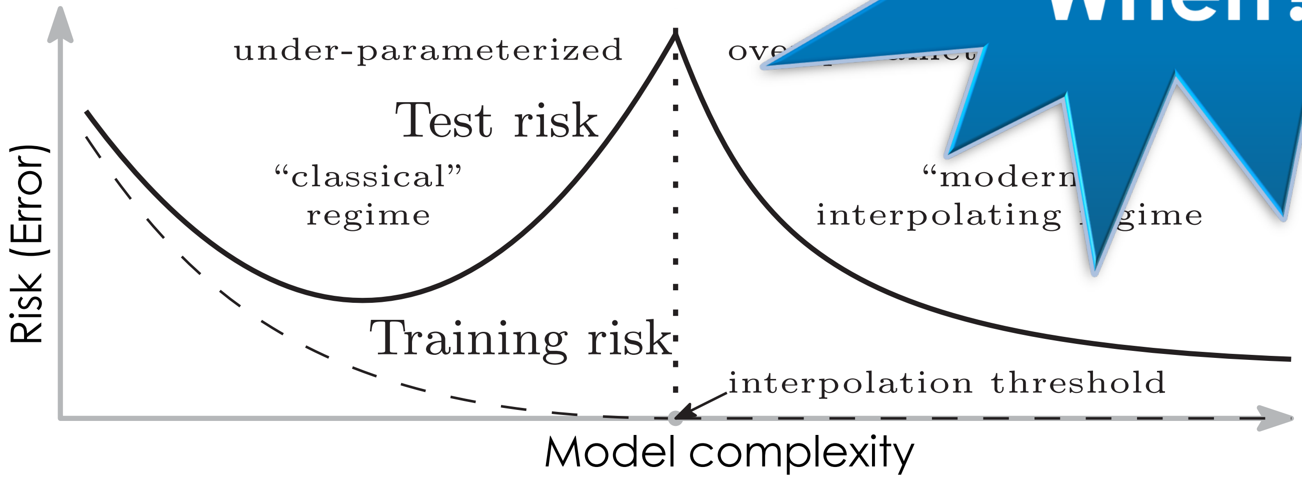
  - $Bias \cong 0$
  - GPT-4, ResNets, StyleGAN, D[obscured]

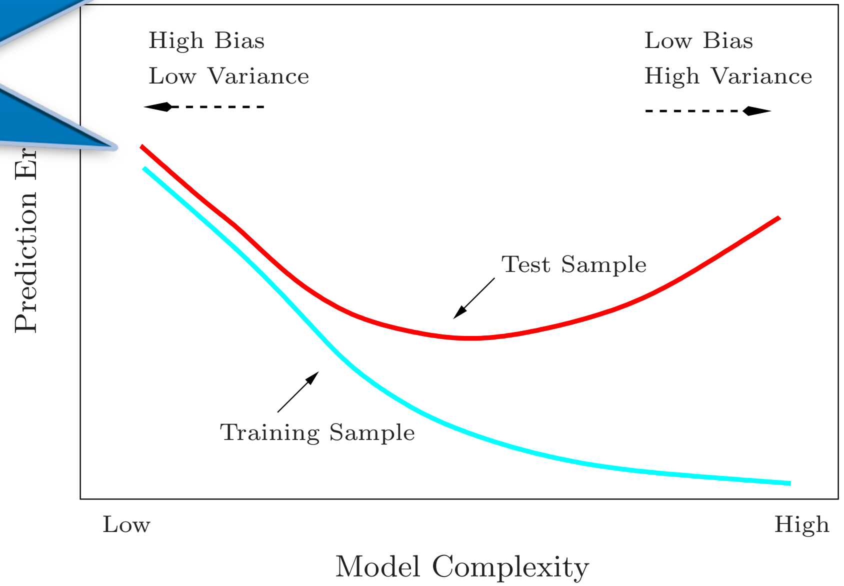
- Classical view:

  more complex model

  - Lower bias, higher variance

Why?
When?





Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849-15854.

Generalization ability: long-standing open

- **Main goal:** small expected loss $F(P, h_o)$
  - Practice: training loss $F(\boldsymbol{D}, h_o) \cong 0$ for overparameterized NNs
- Why can a trained DNN generalize well?
  (Generalization: ability to well perform on unseen data)
- We want to assure, for $\delta > 0$,

$$\Pr(|F(P, h_o) - F(\boldsymbol{D}, h_o)| \leq \epsilon) \geq 1 - \delta$$

  - Generalization gap should be small with a high probability over the random choice of **D**
  - How fast does $F(\boldsymbol{D}, h_o)$ converge to $F(P, h_o)$? (as the training size m increases)

$Error(h_o) :=$
Approximation error
+Optimization error
+Generalization error

A long-standing challenge in DL theory

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT press.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*.

Generalization: VC dimension

- Vapnik–Chervonenkis (VC) dimension:
  - Measure of the capacity (complexity, expressive power, richness) of a set of functions
  - The cardinality of the largest set of points that the learning algorithm can shatter
  - A higher VC dim ➔ richer model family $\mathcal{H}$
- Example: in n-dimensional space
  - Linear models: $VC(\mathcal{H}) = n + 1$
  - ReLU networks with W weights: $VC(\mathcal{H}) = \Omega(W \log W)$

Bartlett, P. L., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*.

- Classical bound: for any $\delta > 0$, with probability at least $1 - \delta$

$$F(P, h) - F(\boldsymbol{D}, h) \leq \sqrt{\frac{2}{m} VC(\mathcal{H}) \log \frac{2e.m}{VC(\mathcal{H})} + \frac{1}{m} \log \frac{2}{\delta}}$$

- Vacuous/meaningless for modern DNNs, due to $W \gg m$ (training size)

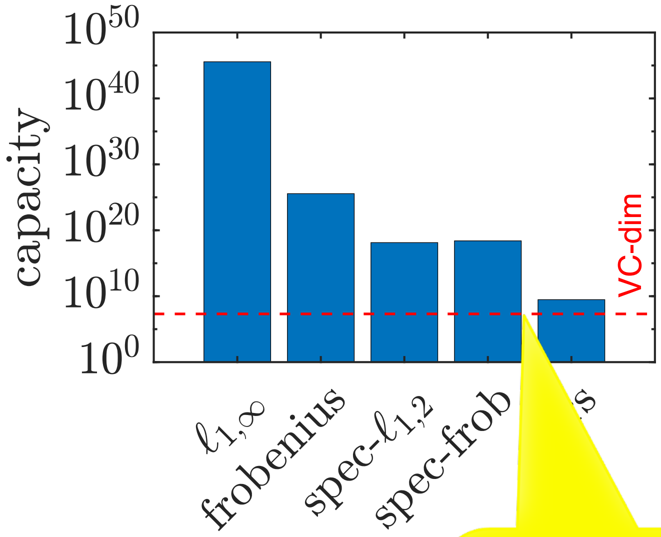
Generalization: Weight norm

- DNN: $h(\boldsymbol{x}, \boldsymbol{W}) = g_K(\boldsymbol{W}_K h_{K-1})$
- Bartlett: **#params** is not important
  - Size of weights may be more important
- Neyshabur et al.; Golowich et al.:

$$F(P,h) - F(\boldsymbol{D},h) \leq O(\|\boldsymbol{W}_1\|_F \cdots \|\boldsymbol{W}_K\|_F)/\sqrt{m}$$

- Bartlett et al.:

$$F(P,h) - F(\boldsymbol{D},h) \leq O(\|\boldsymbol{W}_1\|_2 \cdots \|\boldsymbol{W}_K\|_2)/\sqrt{m}$$



Uninformative for modern DNNs

Arora, S., Ge, R., Neyshabur, B., & Zhang, Y. (2018). Stronger generalization bounds for deep nets via a compression approach. In *ICML*.
Bartlett, P. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*.
Bartlett, P. L., Foster, D. J., & Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. *Neural Information Processing Systems*.
Golowich, N., Rakhlin, A., & Shamir, O. (2020). Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*.
Neyshabur, B., Bhojanapalli, S., & Srebro, N. (2018). A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks. In *ICLR*.

Generalization: PAC-Bayes

- Consider $\mathbb{E}_{h\sim\rho}[F(P,h) - F(\boldsymbol{D},h)]$
  - Generalization error on average over $\mathcal{H}$
  - ρ is the posterior distribution of h
- McAllester: with probability at least $1-\delta$

$$\mathbb{E}_{h\sim\rho}[F(P,h) - F(\boldsymbol{D},h)] \leq \sqrt{\frac{KL(\rho||\mu) + \log(m/\delta)}{2m-1}}$$

  - μ is the prior distribution of h
  - KL is the Kullback-Leibler divergence

- The "distance" between posterior ρ and prior μ:
  - Plays important role
  - Depends on the *bias* of a learning algorithm
- Unclear how fast can ρ approach μ?
- Do not directly consider the complexity of family $\mathcal{H}$

Meaningful bounds appeared

McAllester, D. A. (2003). PAC-Bayesian stochastic model selection. *Machine Learning*, *51*(1), 5-21.

Generalization: non-vacuous bounds

- We can optimize the PAC-Bayes bound
  - Find the posterior ρ^* that minimizes $KL(\rho||\mu)$
- Dziugaite & Roy: non-vacuous bounds
  - MLP with 3 layers, SGD a gorithm, M IST d aset
- Zhou et al.: compressibility
  - Use SOTA compression alg
    bound for ImageNet, L
- Lotfi et al., 2022:
  - Propose compression alg. to find nonvacuous bounds for LeNet-5, ResNet-18, MobileViT

Stochastic
DNNs

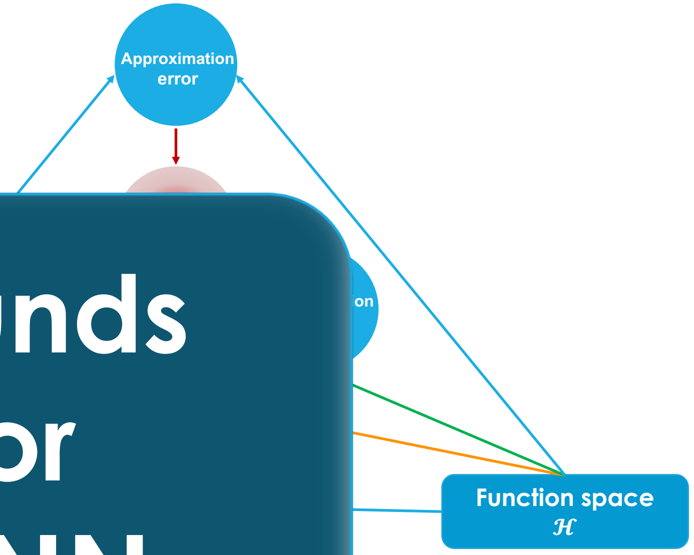
Dataset	Data-independent priors	
	Err. Bound (%)	SOTA (%)
MNIST	**11.6**	21.7 [59]
+ SVHN Transfer	**9.0**	16.1†
FashionMNIST	**32.8**	46.5†
+ CIFAR-10 Transfer	**28.2**	30.1†
CIFAR-10	**58.2**	89.9†
+ ImageNet Transfer	**35.1**	54.2†
CIFAR-100	**94.6**	100†
+ ImageNet Transfer	**81.3**	98.1†
ImageNet	**93.5**	96.5 [73]

Biggs & Guedj, 2022:
- Non-vacuous bounds for a (special) **deterministic networks**
- MNIST and Fashion-MNIST datasets

Dziugaite, G., & Roy, D. (2017). Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *UAI*.
Zhou, W., Veitch, V., Austern, M., Adams, R., & Orbanz, P. (2019). Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach. In *ICLR*.
Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., & Wilson, A. G. (2022). PAC-bayes compression bounds so tight that they can explain generalization. *In NeurIPS*.
Biggs, F., & Guedj, B. (2022). Non-vacuous generalisation bounds for shallow neural networks. In *ICML*.

Generalization: long-standing open

- Some other approaches:
  - Neural tangent kernel, Mean field
  - Algorithr



Current meaningful bounds however are mostly for stochastic or shallow NNs

Unclear about Big pretrained models, Deep NNs in practice

Unclear about Why many tricks in DL improve performance

Take-home messages

- Deep neural networks are universal approximators
- Theoretically clear about:
  - Approximation ability
  - Optimization (learning process)
- Long-standing open challenge about Generalization ability

Thank you

Q&A