# A Brief Introduction to Learning Theory

Rongzhen Wang

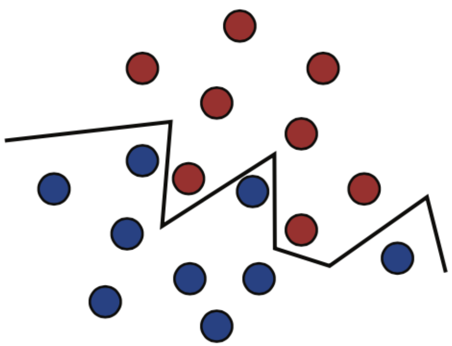
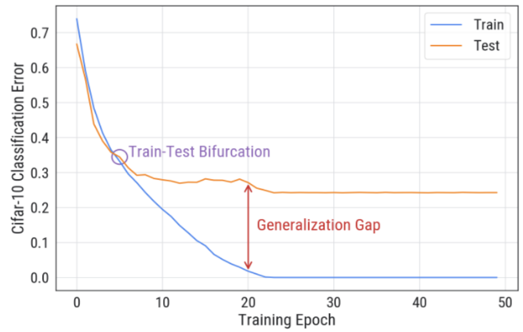
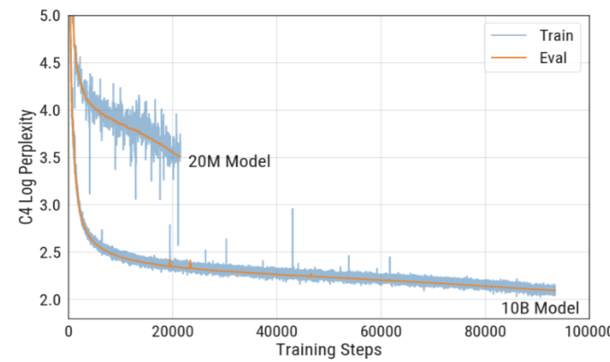
2024.11.4

# Theoretical Learning Guarantee

- Machine learning: to *find patterns in data* for *particular tasks*
- To design "good" machine learning algorithms, we need to answer an important question: *How to evaluate an algorithm?*

- Empirical benchmarks: test loss, user study…
  - Pros: generally applicable, seems promising
  - Cons: instance-by-instance, without guarantee
- What if we take an machine learning algorithm as a *mathematical model* so that we can *tune it arbitrarily* and its performance can be *theoretically guaranteed*?

# 3-stages in AI

| Traditional statistical machine learning | Deep learning | Large models |
| --- | --- | --- |
|  |  |  |
| • Small train error & sensitive test error | • Zero train error & non-increasing test error | • Aligned train error & test error |






[1]Belkin M, Hsu D, Ma S, et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. 2019.
[2] Xiao L. Rethinking Conventional Wisdom in Machine Learning: From Generalization to Scaling. 2024.

# 3-stages in AI

| Traditional statistical machine learning | Deep learning | Large models |
| --- | --- | --- |

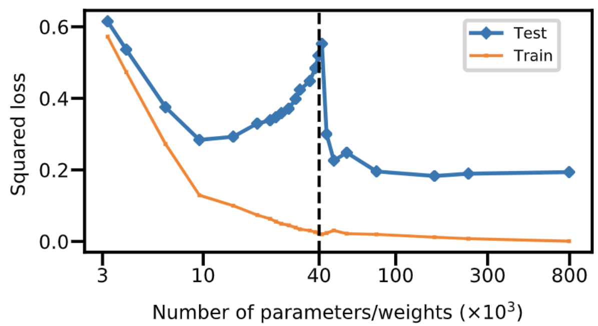
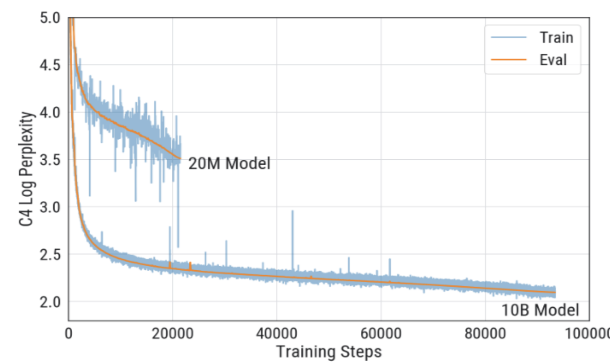



**Fig. 3.** Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of $H$ hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d + 1) \cdot H + (H + 1) \cdot K$. The interpolation threshold (black dashed line) is observed at $n \cdot K$.




- Small train error & sensitive test error
- Zero train error & non-increasing test error
- Aligned train error & test error


[1]Belkin M, Hsu D, Ma S, et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. 2019.
[2] Xiao L. Rethinking Conventional Wisdom in Machine Learning: From Generalization to Scaling. 2024.

# Outline

- Learning Theory Framework
  - PAC-learning ($\delta$-$\varepsilon$ correct) framework
  - Generalization, optimization, approximation
- Generalization: measure of model complexity
- Optimization: computational hardness
- Theoretical mysteries in deep learning
- Theoretical topics for large models

# Notations

- Consider *supervised learning* with *empirical risk minimization (ERM)*.
- Feature space $X$, label space $Y$, target distribution $D$, $S_m = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ is a simple random (i.i.d.) sample of size $m$ drawn from $D$
- Hypothesis $h: X \rightarrow Y$, hypothesis space $H = \{h: h_\theta, \theta \in \Theta\}$
- Loss function $L: Y \times Y \rightarrow \mathbb{R}$
- Population risk $R(h) = \mathbb{E}_{x,y\sim D}[L(h(x), y)]$, empirical risk $R_{S_m}(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i)$
- ERM $\hat{h}_{S_m,\mathrm{ERM}} = \min_{h\in H} R_{S_m}(h)$
- Optimization algorithm $A$ with output $\hat{h}_{S_m,A}$

# Learning Theory Framework

## $\delta$-$\varepsilon$ correct with certain sample complexity

- Probability: $1-\delta$
- Approximately Correct: $\varepsilon$
- Sample complexity: $m_{\mathcal{A}}(\delta,\varepsilon)$

- *PAC-learning framework* [Valiant, 1984] :
  - We say an learning algorithm $\mathcal{A}$ (corresponding with $H, L, A$) is PAC-learnable with sample complexity $m_{\mathcal{A}}(\delta,\varepsilon)$, if there exists a function $m_{\mathcal{A}}(\delta,\varepsilon)$ such that for any $\delta,\varepsilon>0$, it holds that
  $$\forall\, m \geq m_{\mathcal{A}}(\delta,\varepsilon),\ \ \mathbb{P}\left(R\left(\hat{h}_{S_m,A}\right) \leq \varepsilon\right) \geq 1-\delta.$$

# Learning Theory Framework

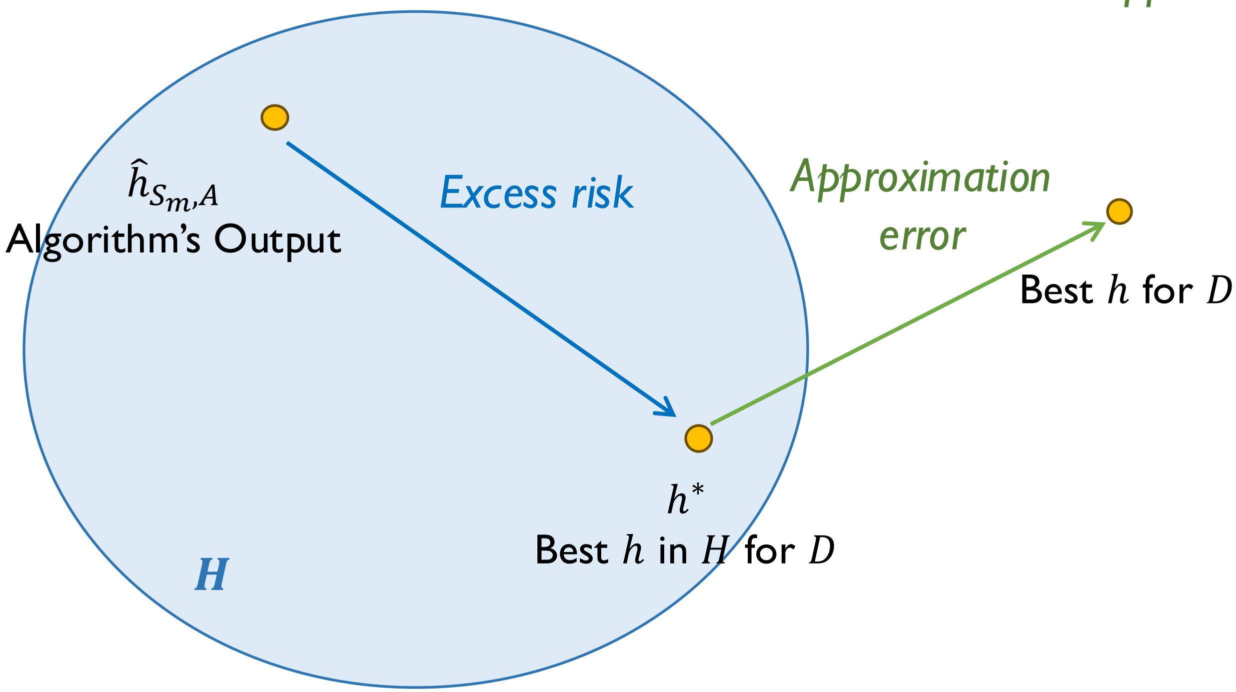
## $\delta$-$\varepsilon$ correct with certain sample complexity

- Probability: $1-\delta$
- Approximately Correct: $\varepsilon$
- Sample complexity: $m_{\mathcal{A}}(\delta,\varepsilon)$

- *PAC-learning framework* [Valiant, 1984] :
  - We say an learning algorithm $\mathcal{A}$ (corresponding with $H, L, A$) is PAC-learnable with sample complexity $m_{\mathcal{A}}(\delta,\varepsilon)$, if there exists a function $m_{\mathcal{A}}(\delta,\varepsilon)$ such that for any $\delta, \varepsilon > 0$, it holds that
$$\forall\, m \geq m_{\mathcal{A}}(\delta,\varepsilon),\ \mathbb{P}\left(R\left(\hat{h}_{S_m,A}\right) \leq \varepsilon\right) \geq 1-\delta.$$
Or for any $\delta > 0$, it holds with probability at least $1-\delta$ that
$$R\left(\hat{h}_{S_m,A}\right) \leq \varepsilon(\delta, \mathrm{m}).$$

# Analysis of PAC-learning

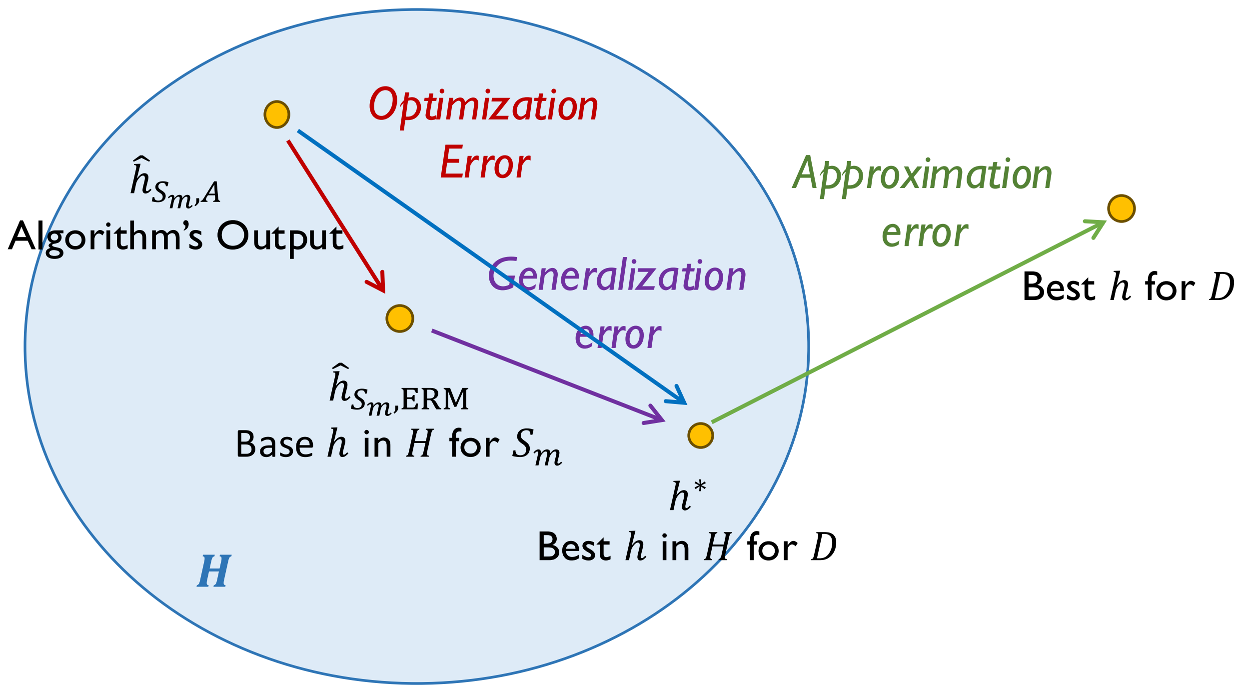
- Decomposition of population risk
  - Let $h^* \triangleq \arg\inf_{h \in H} R(h)$ , $R(\hat{h}_{S_m,A}) = \underbrace{R(\hat{h}_{S_m,A}) - R(h^*)}_{\text{Excess risk}} + \underbrace{R(h^*)}_{\text{Approximation error}}$



# Analysis of PAC-learning

- Decomposition of population risk
  - Let $h^* \triangleq \arg\inf_{h \in H} R(h)$, $R(\hat{h}_{S_m,A}) = \underbrace{R(\hat{h}_{S_m,A}) - R(h^*)}_{\text{Excess risk}} + \underbrace{R(h^*)}_{\text{Approximation error}}$
- Decomposition of excess risk
  - $$\begin{aligned} & R(\hat{h}_{S_m,A}) - R(h^*) \\ &= R(\hat{h}_{S_m,A}) - R_{S_m}(\hat{h}_{S_m,A}) + R_{S_m}(\hat{h}_{S_m,A}) - R_{S_m}(\hat{h}_{S_m,\mathrm{ERM}}) \\ &+ \underbrace{R_{S_m}(\hat{h}_{S_m,\mathrm{ERM}}) - R_{S_m}(h^*)}_{(\leq 0 \text{ by ERM})} + R_{S_m}(h^*) - R(h^*) \\ &\leq \underbrace{R(\hat{h}_{S_m,A}) - R_{S_m}(\hat{h}_{S_m,A}) + R_{S_m}(h^*) - R(h^*)}_{\text{Generalization error}} \\ &+ \underbrace{R_{S_m}(\hat{h}_{S_m,A}) - R_{S_m}(\hat{h}_{S_m,\mathrm{ERM}})}_{\text{Optimization error}} \end{aligned}$$

# Analysis of PAC-learning




**generalization + optimization + approximation**

# Generalization

- $R\left(\hat{h}_{S_m,A}\right) - R_{S_m}\left(\hat{h}_{S_m,A}\right) + R_{S_m}(h^*) - R(h^*)$
- $R_{S_m}(h^*) - R(h^*)$ can be bounded with concentration inequality.
  - Since $h^*$ is independent of $S_m$,
    $$\mathbb{E}\left[R_{S_m}(h^*)\right] = \mathbb{E}[R(h^*)].$$
  - By *Hoeffding's inequality*,
    $\mathbb{P}_{S_m \sim D^m}\left[\left|R_{S_m}(h^*) - R(h^*)\right| \leq \varepsilon\right] \geq 1 - 2\exp(-2m\varepsilon^2)$ .

# Generalization

- $R\left(\hat{h}_{S_m,A}\right) - R_{S_m}\left(\hat{h}_{S_m,A}\right) + R_{S_m}(h^*) - R(h^*)$
- $R_{S_m}(h^*) - R(h^*)$ can be bounded with concentration inequality.
  - Since $h^*$ is independent of $S_m$,
  $$\mathbb{E}\left[R_{S_m}(h^*)\right] = \mathbb{E}[R(h^*)].$$
  - By *Hoeffding's inequality*,
  $\mathbb{P}_{S_m \sim D^m}\left[\left|R_{S_m}(h^*) - R(h^*)\right| \leq \varepsilon\right] \geq 1 - 2\exp(-2m\varepsilon^2)$.
- Many techniques are developed to bound $R\left(\hat{h}_{S_m,A}\right) - R_{S_m}\left(\hat{h}_{S_m,A}\right)$.
  - *Uniform convergence* [Vapnik and Chervonenkis, 1968], which depends on the complexity of the hypothesis space $H$:
  $$\left|R\left(\hat{h}_{S_m,A}\right) - R_{S_m}\left(\hat{h}_{S_m,A}\right)\right| \leq \sup_{h \in H}\left|R(h) - R_{S_m}(h)\right|.$$
  - *Algorithmic stability* [Bousquet and Elisseeff, 2002], which characterizes the property of learning algorithm $\mathcal{A}$.

# Generalization via Uniform Convergence

## Finite hypothesis space

- Concentration inequality + union bound

- **Theorem 1 (Generalization bound — finite $H$)** *Let $H$ be a finite hypothesis set. Suppose $L \in [0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\forall h \in H, R(h) \leq R_{S_m}(h) + \sqrt{\frac{\log(|H|) + \log(2/\delta)}{2m}}.$$

# Generalization via Uniform Convergence

## Infinite hypothesis space

- The hypothesis sets typically used in machine learning are *infinite*.
- General idea: *reducing the infinite case to the analysis of finite sets of hypotheses* and then proceed as in the finite cases.

# Generalization via Uniform Convergence

## Infinite hypothesis space

- The hypothesis sets typically used in machine learning are *infinite*.
- General idea: *reducing the infinite case to the analysis of finite sets of hypotheses* and then proceed as in the finite cases.

- Via discretization:
  - Use finite set of hypotheses $\widetilde{H}$ to approximately cover infinite $H$
  - Derive bounds with $|\widetilde{H}|$ and additional error between $\widetilde{H}$ and $H$
- Via complexity:
  - Measure the variety of $H$ with its ability to fit a known-complexity space
  - Transfer the bounds as an complexity bound and further measure the complexity

# Generalization via Uniform Convergence

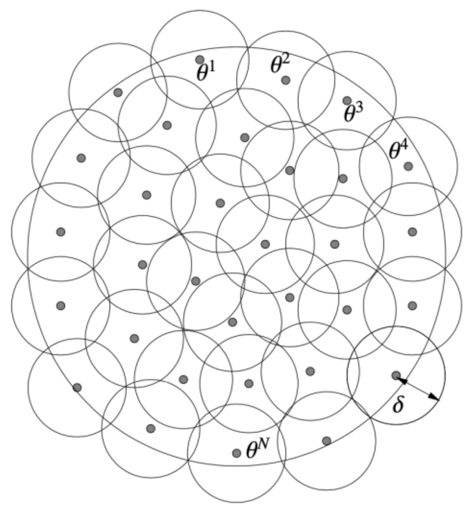
## Infinite hypothesis space via discretization

- Discretize parameter space with *covering number*:

**Definition 4.22.** $\mathcal{C}$ is an $\epsilon$-*cover* of $\mathcal{Q}$ with respect to metric $\rho$ if for all $v' \in \mathcal{Q}$, there exists $v \in \mathcal{C}$ such that $\rho(v, v') \leq \epsilon$.

**Definition 4.23.** The *covering number* is defined as the minimum size of an $\epsilon$-cover, or explicitly:

$$N(\epsilon, \mathcal{Q}, \rho) \triangleq (\text{min size of } \epsilon\text{-cover of } \mathcal{Q} \text{ w.r.t. metric } \rho). \tag{4.102}$$





[1] Tengyu Ma. Lecture Notes for Machine Learning Theory. 2022.

[2] Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. 2019.

# Generalization via Uniform Convergence

## Infinite hypothesis space via discretization

- Let $\mathcal{G} = \{g\colon (x, y) \mapsto L(h(x), y)\colon h \in \mathcal{H}\}$.
- **Theorem 2.1 (Generalization bound — infinite $H$, via function space discretization)** *Let $H$ be a finite hypothesis set. Suppose $L \in [0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\forall h \in H, R(h) \leq R_{S_m}(h) + \inf_{\epsilon > 0}\left[\epsilon + \sqrt{\frac{\log\big(3N(\epsilon, \mathcal{G}, L_1)\big) + \log(1/\delta)}{2m}}\right].$$

[1] Tong Zhang. Mathematical Analysis of Machine Learning Algorithms. 2022.

# Generalization via Uniform Convergence

## Infinite hypothesis space via discretization

- Let $\mathcal{G} = \{g\colon (x,y) \mapsto L(h(x),y)\colon h \in \mathcal{H}\}$.
- **Theorem 2.1 (Generalization bound — infinite $H$, via function space discretization)** *Let $H$ be a finite hypothesis set. Suppose $L \in [0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$, it holds that*

$$\forall h \in H, R(h) \le R_{S_m}(h) + \inf_{\epsilon>0}\left[\epsilon + \sqrt{\frac{\log\big(3N(\epsilon,\mathcal{G},L_1)\big) + \log(1/\delta)}{2m}}\right].$$

- Recalling in the finite case: $\forall h \in H, R(h) \le R_{S_m}(h) + \sqrt{\frac{\log(|H|)+\log\left(\frac{2}{\delta}\right)}{2m}}$.
  - $|H| \Rightarrow 3N(\epsilon,\mathcal{G},L_1)$

[1] Tong Zhang. Mathematical Analysis of Machine Learning Algorithms. 2022.

# Generalization via Uniform Convergence

## Infinite hypothesis space via discretization

- Covering number of concrete models
  - Lipschitz functions on bounded parameter space
    - If $L(h_\theta(x), y)$ is $\kappa$-Lipschitz w.r.t. $\theta$ and $\Theta = \{\theta \in \mathbb{R}^d\colon \|\theta\|_1 \le r\}$ , then $N(\epsilon, \mathcal{G}, L_1) \le (1 + 2\kappa r/\epsilon)^d$.
    - $R(h) \le R_{S_m}(h) + \tilde{O}\left(\sqrt{\frac{d}{m}}\right)$.
  - Deep neural networks
    - If $H = \left\{h\colon h_\theta(x) = {W_L}^T \sigma_{L-1}\left({W_{L-1}}^T \sigma_{L-2}(\cdots \sigma_1(W_1 x))\right), \|W_i\|_\infty \le B\right\}$, then $N(\epsilon, \mathcal{G}, L_1) \le \dfrac{\left(4(L+1)(B_x+1)(2B)^{L+2}(\Pi_{j=1}^L \rho_j)(\Pi_{j=0}^L d_j)\cdot \epsilon^{-1}\right)^S}{d_1! \times d_2! \times \cdots \times d_L!}$
    - $R(h) \le R_{S_m}(h) + \tilde{O}\left(\sqrt{\frac{LS}{m}}\right)$, where $S$ is the number of parameters.

[1] Tong Zhang. Mathematical Analysis of Machine Learning Algorithms. 2022.

# Generalization via Uniform Convergence

## Infinite hypothesis space via complexity

- Motivation:
  - The generalization bounds based on discretization are mostly dependents on the dimension of parameters.
  - High dimension does not imply high variety.
- Can we characterize the variety of $H$?

# Generalization via Uniform Convergence

## Infinite hypothesis space via complexity

- Motivation:
  - The generalization bounds based on discretization are mostly dependents on the dimension of parameters.
  - High dimension does not imply high variety.
- Can we characterize the variety of $H$?
- Empirical Rademacher complexity [Bartlett and Mendelson, 2002]: ability to mimic or express randomness

**Definition 3.1 (Empirical Rademacher complexity)** *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[a, b]$ and $S = (z_1, \ldots, z_m)$ a fixed sample of size $m$ with elements in $\mathcal{Z}$. Then, the* empirical Rademacher complexity *of $\mathcal{G}$ with respect to the sample $S$ is defined as:*

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right], \qquad (3.1)$$

*where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\top$, with $\sigma_i$s independent uniform random variables taking values in $\{-1, +1\}$.*[3] *The random variables $\sigma_i$ are called* Rademacher variables.

# Generalization via Uniform Convergence

## Infinite hypothesis space via complexity

- **Theorem 2.2 (Generalization bound — infinite $H$, via empirical Rademacher complexity)**

**Theorem 3.3** *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample $S$ of size $m$, each of the following holds for all $g \in \mathcal{G}$:*

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

- Recalling in the finite case: $\forall h \in H, R(h) \leq R_{S_m}(h) + \sqrt{\frac{\log(|H|)+\log(2/\delta)}{2m}}$.
  - $\sqrt{\frac{\log(|H|)}{2m}} \Rightarrow \widehat{\mathfrak{R}}_S(\mathcal{G})$

[3] Mohri M. Foundations of machine learning. 2018.

# Generalization via Uniform Convergence

## Infinite hypothesis space via complexity

- Rademacher complexity of deep neural networks
  - Covering number upper bounds Rademacher complexity

**Lemma A.5.** *Let $\mathcal{F}$ be a real-valued function class taking values in $[0,1]$, and assume that $\mathbf{0} \in \mathcal{F}$. Then*

$$\mathfrak{R}(\mathcal{F}_{|S}) \leq \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n}\int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}_{|S}, \varepsilon, \|\cdot\|_2)}d\varepsilon.\right)$$

  - Covering number bound of deep neural networks
  - Rademacher complexity bounds for deep neural networks

**Lemma A.8.** *Let fixed nonlinearities $(\sigma_1, \ldots, \sigma_L)$ and reference matrices $(M_1, \ldots, M_L)$ be given where $\sigma_i$ is $\rho_i$-Lipschitz and $\sigma_i(0) = 0$. Further let margin $\gamma > 0$, data bound $B$, spectral norm bounds $(s_i)_{i=1}^L$, and $l_1$ norm bounds $(b_i)_{i=1}^L$ be given. Then with probability at least $1-\delta$ over an iid draw of $n$ examples $((x_i, y_i))_{i=1}^n$ with $\sqrt{\sum_i \|x_i\|_2^2} \leq B$, every network $F_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^k$ whose weight matrices $\mathcal{A} = (A_1, \ldots, A_L)$ obey $\|A_i\|_\sigma \leq s_i$ and $\|A_i^\top - M_i^\top\|_{2,1} \leq b_i$ satisfies*

$$\Pr\left[\arg\max_j F_{\mathcal{A}}(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_\gamma(f) + \frac{8}{n} + \frac{72B\ln(2W)\ln(n)}{\gamma n}\left(\prod_{i=1}^L s_i\rho_i\right)\left(\sum_{i=1}^L \frac{b_i^{2/3}}{s_i^{2/3}}\right)^{3/2} + 3\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

- $R(h) \leq R_{S_m}(h) + \tilde{O}\left(\sqrt{\frac{LBb}{m}}\right)$, where $B/b$ are data/weight matric normalizations.

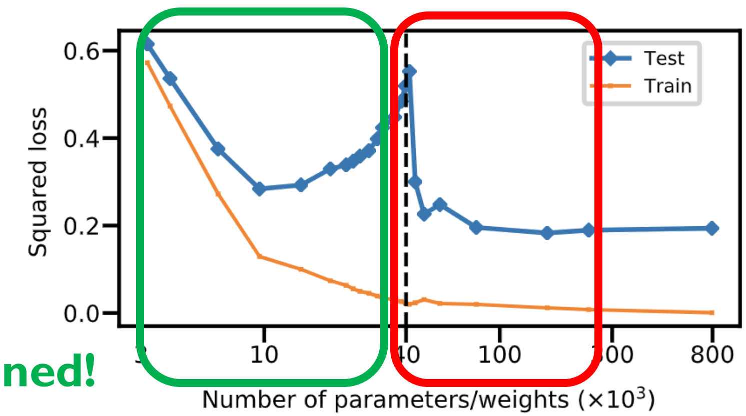
[1] Bartlett P L, Foster D J, Telgarsky M J. Spectrally-normalized margin bounds for neural networks. 2017.

# Theoretical Mysteries in Deep Learning

## Generalization

Traditional statistical machine learning

Deep learning




Have been well explained!

**Fig. 3.** Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of $H$ hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d+1) \cdot H + (H+1) \cdot K$. The interpolation threshold (black dashed line) is observed at $n \cdot K$.

- Small train error & sensitive test error
- Zero train error & non-increasing test error


[1]Belkin M, Hsu D, Ma S, et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. 2019.

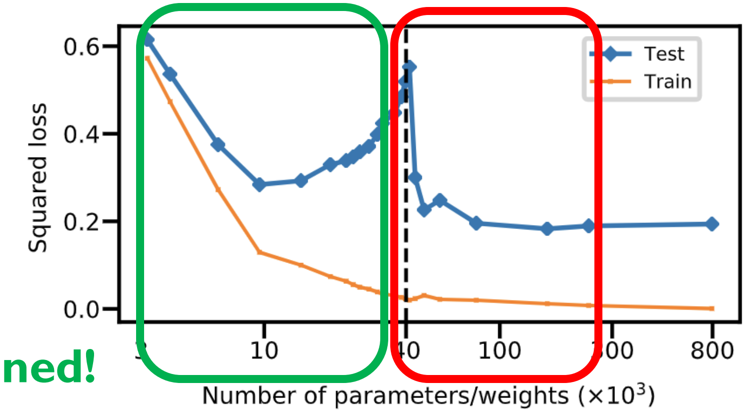
[2] Xiao L. Rethinking Conventional Wisdom in Machine Learning: From Generalization to Scaling. 2024.

# Theoretical Mysteries in Deep Learning

## Generalization

Traditional statistical machine learning

Deep learning




**Have been well explained!**

**Topic for generalization in deep learning era:**
How to explain this descent generalization error?

**Fig. 3.** Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of $H$ hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d+1) \cdot H + (H+1) \cdot K$. The interpolation threshold (black dashed line) is observed at $n \cdot K$.

- Small train error & sensitive test error
- Zero train error & non-increasing test error


[1]Belkin M, Hsu D, Ma S, et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. 2019.

[2] Xiao L. Rethinking Conventional Wisdom in Machine Learning: From Generalization to Scaling. 2024.

# Optimization

- Finding a zero-loss (or approximately zero-loss) solution for a 2-layer neural network is NP-complete
  - NP-complete: If a problem is proven to be NP-complete, it means that it is difficult to find its solution in polynomial time under existing algorithms.
  - Consequence: Running an algorithm in polynomial time, the optimization error $R_{S_m}(\hat{h}_{S_m,A}) - R_{S_m}(\hat{h}_{S_m,\mathrm{ERM}})$ is not guaranteed to be small.

[1] Blum A, Rivest R. Training a 3-node neural network is NP-complete. 1988.

[2] Bartlett P, Ben-David S. Hardness results for neural network approximation problems. 1999.

# Theoretical Mysteries in Deep Learning

## Optimization

Traditional statistical machine learning

Deep learning

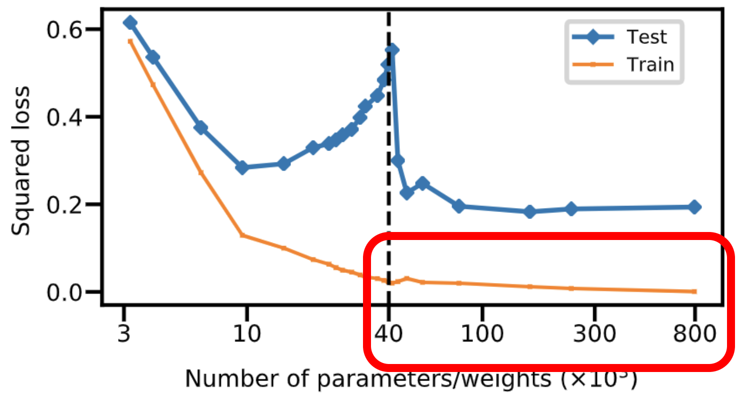
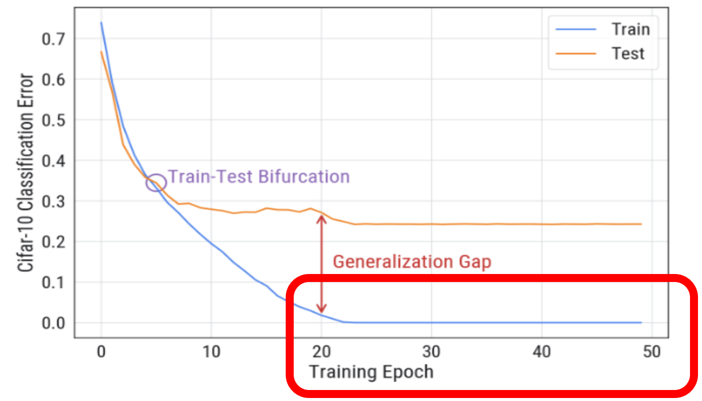



**Fig. 3.** Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of $H$ hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d+1) \cdot H + (H+1) \cdot K$. The interpolation threshold (black dashed line) is observed at $n \cdot K$.




- Small train error & sensitive test error
- Zero train error & non-increasing test error

**Topic for optimization in deep learning era:**

How to explain this zero training risk?


[1]Belkin M, Hsu D, Ma S, et al. Reconciling modern machine-learning practice and the classical bias–variance trade-off. 2019.

[2] Xiao L. Rethinking Conventional Wisdom in Machine Learning: From Generalization to Scaling. 2024.

# Theoretical Mysteries in Deep Learning
## Generalization

- Benign Overfitting

### Benign Overfitting

- Deep networks can achieve zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for noisy problems ($R^* \gg 0$).
- No tradeoff between fit to training data and complexity!
- Deep networks seem to operate in the overfitting regime ($\hat{R}(f) \ll R^*$) but still predict accurately.
- A new statistical phenomenon: *benign overfitting*.

[1] LIDS@80: Session 3 Keynote — Peter Bartlett (University of California, Berkeley). https://www.youtube.com/watch?v=RQz4JEw9ag4.

# Theoretical Mysteries in Deep Learning

## Generalization I

- Implicit regularization

**Regularization in the overfitting regime ($c \ll R^*$)**

$$\begin{aligned} \min \quad & \Omega(f) \\ \text{s.t.} \quad & \hat{R}(f) \leq c. \end{aligned}$$

**Implicit Regularization**

- Stochastic gradient descent finds deep networks satisfying the (overfitting) constraint, and these deep networks predict accurately.
- What is the regularizer $\Omega$?
- The boundaries between the key issues of *optimization, estimation, and approximation* are blurred.

[1] LIDS@80: Session 3 Keynote — Peter Bartlett (University of California, Berkeley).
https://www.youtube.com/watch?v=RQz4JEw9ag4.

# Theoretical Mysteries in Deep Learning

## Generalization I

- Implicit regularization of linear cases

**Progress in Implicit Regularization**

- Linear. $f : x \mapsto \langle \theta, x \rangle$: $\Omega(f) = \|\theta - \theta_0\|$.
- Polynomial. $\theta_i$ replaced by $\theta_i^\alpha$: $\Omega(f)$ like a Huber norm.

  (Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, 2017)
- Logistic regression (Soudry, Hoffer, Srebro, 2017)
- Linear convolutional: $\Omega(f)$ penalizes norm of Fourier transform.

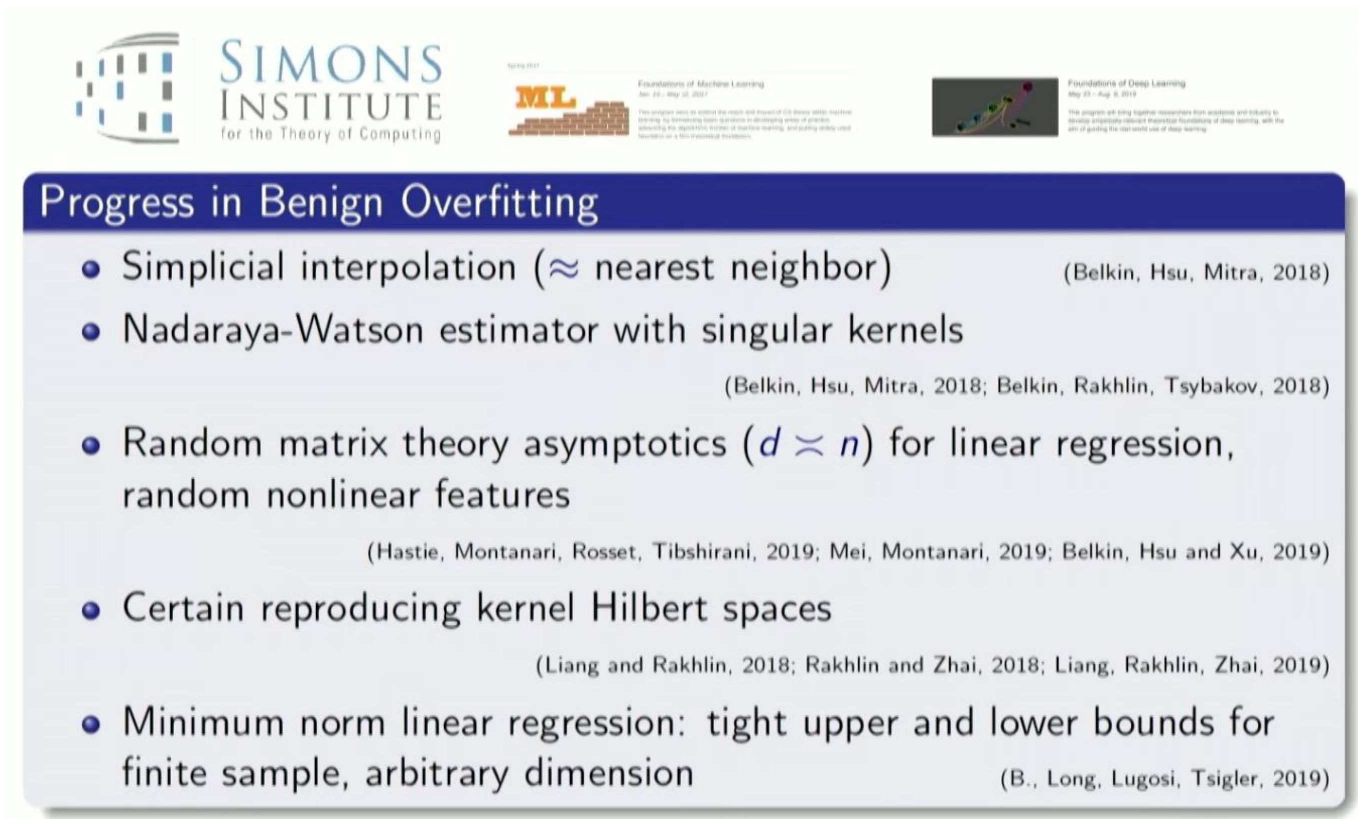
  (Gunasekar, Lee, Soudry, Srebro, 2018)

[1] LIDS@80: Session 3 Keynote — Peter Bartlett (University of California, Berkeley). https://www.youtube.com/watch?v=RQz4JEw9ag4.

# Theoretical Mysteries in Deep Learning

## Generalization 2

- Generalization performance of benign overfitting models




[1] LIDS@80: Session 3 Keynote — Peter Bartlett (University of California, Berkeley). https://www.youtube.com/watch?v=RQz4JEw9ag4.

# Theoretical Mysteries in Deep Learning

## Generalization 2



(B., Long, Lugosi, Tsigler, 2019)

### Characterizing benign overfitting in linear regression

For $\ell(f) = (f(x) - y)^2$, $\Omega(x \mapsto \langle x, \theta \rangle) = \|\theta\|_2$, and $\begin{pmatrix} x \\ y \end{pmatrix} = \Phi z$ where $\Phi$ is a bounded linear operator and $z$ has subgaussian, independent entries,

$$c_1 \left( \frac{d^*}{n} + \frac{n}{R_{d^*}} + \phi\left(\frac{1}{n}\right) \right) \leq \mathbb{E}R(\hat{f}) - R^* \leq c_2 \left( \frac{d^*}{n} + \frac{n}{R_{d^*}} + \frac{1}{\sqrt{n}} \right),$$

where $d^* = \min\{d : r_d \geq c_3 n\}$, $r_d$ and $R_d$ are effective ranks of the covariance of $x$ in the subspace orthogonal to the $d$ highest variance directions, and $\phi$ is increasing.

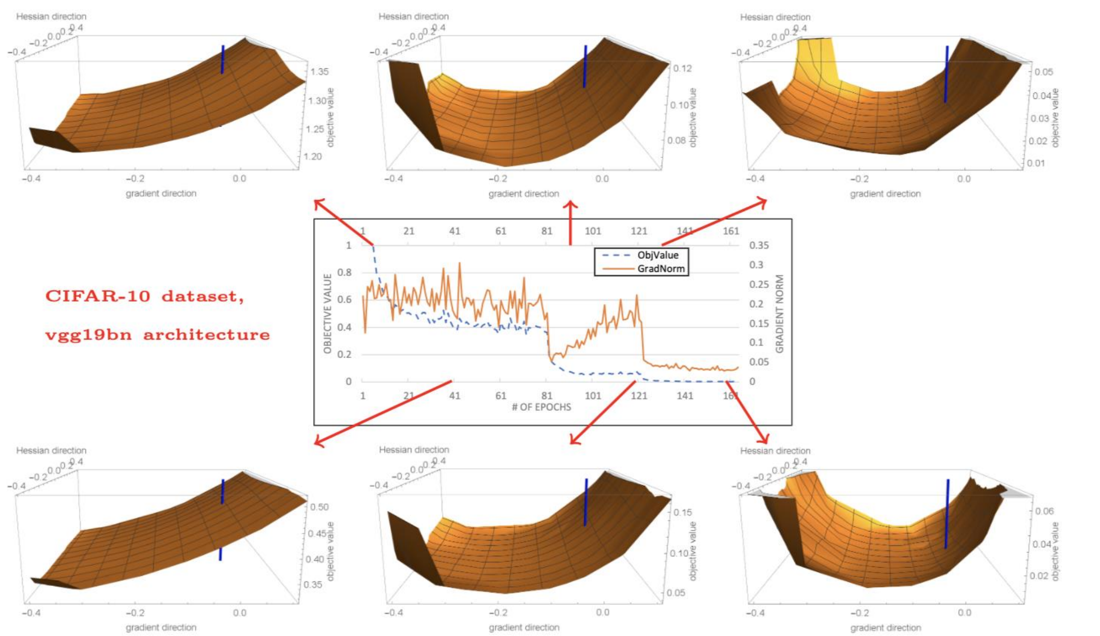
That is, benign overfitting occurs iff there is a subspace where the covariance has small magnitude, high dimension, and low eccentricity.

[1] LIDS@80: Session 3 Keynote — Peter Bartlett (University of California, Berkeley). https://www.youtube.com/watch?v=RQz4JEw9ag4.

# Theoretical Mysteries in Deep Learning

## Optimization

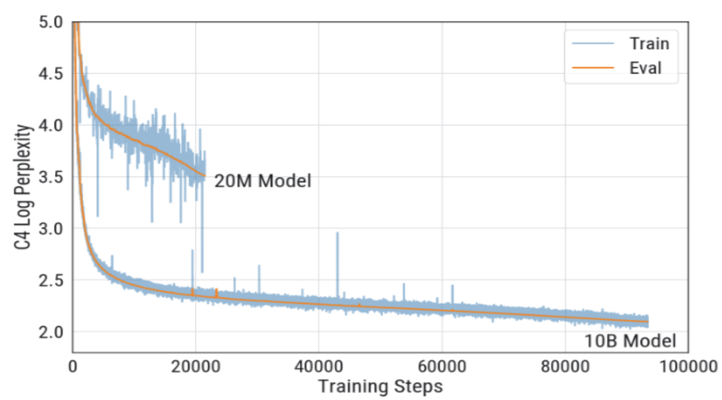
- Good loss landscape for over-parameterized neural network:
  - Near the GD/SGD training trajectory, there is no local minima and the objective is semi-smooth.
  - Consequence: SGD can find global minima on the training objective of DNNs in polynomial time.




[1] Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization, 2019.

# Theoretical Topics for Large Models

- Non-zero training loss, zero test loss, but great **ability**
- Generalization:
  - Evaluate precise output on specific tasks with new benchmarks, e.g., ICL
- Optimization:
  - Properties and improvement of large/infinite models, e.g., muP
- Approximation:
  - Expressiveness of certain architectures/models, e.g., Transformers.



# My works

- Stability of gradient-based bilevel algorithms, NeurIPS 2024

$$\epsilon_{\text{gen}} \leq \epsilon_{\text{stab}} \leq L\epsilon_{\text{arg}}$$

$$\sum_{t=1}^{T} \prod_{s=t+1}^{T} \left(1 + \alpha_s(1 - 1/m)\gamma\right) \frac{2\alpha_t L'}{m} \leq \epsilon_{\text{arg}} \leq \sum_{t=1}^{T} \prod_{s=t+1}^{T} \left(1 + \alpha_s(1 - 1/m)\gamma\right) \frac{2\alpha_t L}{m},$$

- Density estimation guarantee for conditional generative models

$$\mathbb{E}_Y[d_{TV}(P_{X|Y;\hat{\theta}}, P_{X|Y;\theta^*})] \leq \frac{1}{2}\sqrt{(\epsilon + 4)\epsilon + 2(\epsilon + 2)\left(\epsilon + \frac{2}{n}\log\frac{N_{\text{UB}}(\epsilon, \mathcal{P}(\Theta))}{\delta}\right)}.$$

# Thanks