Cybenko's Universal Approximation Theorem and Why It Is Not Enough

MIMIC Seminar (May 4, 2026)

Speaker: Kyoungman Kim

I. Introduction

Definition 1 (Deep Neural Network)**.** A deep neural network (DNN) is a parameterized function obtained by composing affine maps with nonlinear activation functions.

Its objective is to approximate a target function $f : \mathbb{R}^n \to \mathbb{R}^m$. If Θ denotes the collection of parameters, this goal can be written concisely as $f \approx F_\Theta$.
Let σ be an activation function.
Thus, a fully connected feedforward neural network can be written as

$$F(x) = \sigma(W_N \sigma(\cdots \sigma(W_2 \sigma(W_1 x + b_1) + b_2) \cdots) + b_N)\,.$$

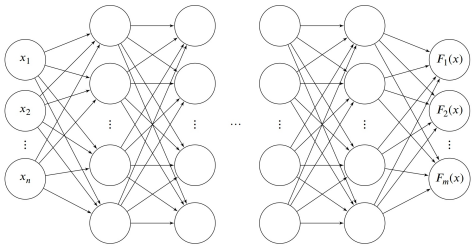


Figure 1: A visual representation of a feedforward neural network.

II. Approximation

Universal Approximation Theorem (Cybenko, 1989). Let σ be any continuous sigmoidal function. Then the set

$$S := \left\{ G(x) = \sum_{k=1}^{N} \alpha_k \sigma(y_k \cdot x + \theta_k) \;\middle|\; y_k \in \mathbb{R}^n, \quad \alpha_k \in \mathbb{R}, \quad \theta_k \in \mathbb{R}, \quad N \in \mathbb{N} \right\}$$

is uniformly dense in $C(I_n)$. In other words, for every $f \in C(I_n)$ and every $\varepsilon > 0$, there exists $G \in S$ such that $\|G - f\|_\infty < \varepsilon$.

Throughout this section, let μ denote a signed regular Borel measure on $I_n = [0, 1]^n$. We will rely on the following standard mathematical results.

Theorem 1 (Lebesgue's Dominated Convergence Theorem)**.** *Let X be a measure space, and let $\{f_n\}$ be a sequence of measurable functions on X such that $f(x) = \lim_{n\to\infty} f_n(x)$ exists for every $x \in X$. If there exists $g \in L^1(|\mu|)$ such that $|f_n(x)| \leq g(x)$ for all $n \in \mathbb{N}$ and $x \in X$, then $f \in L^1(|\mu|)$ and*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu.$$

Theorem 2 (Riesz Representation Theorem)**.** *If X is a locally compact Hausdorff space, then every bounded linear functional Φ on $C_0(X)$ is represented by a unique regular complex Borel measure μ, in the sense that $\Phi(f) = \int_X f \, d\mu$ for every $f \in C_0(X)$. Moreover, $\|\Phi\| = |\mu|(X)$.*

Here, $C_0(X)$ denotes the class of all continuous functions that vanish at infinity. In other words, for every $\varepsilon > 0$, there exists a compact set $K \subset X$ such that $|f(x)| < \varepsilon$ for all $x \notin K$.

Theorem 3 (Hahn–Banach Theorem: Geometric Form)**.** *Let V be a normed vector space, and let $A, B \subset V$ be two nonempty, closed, disjoint, convex subsets such that one of them is compact. Then there exist a nonzero continuous linear functional f, a constant $\alpha \in \mathbb{R}$, and $\varepsilon > 0$ such that*

$$f(x) \leq \alpha - \varepsilon \quad (\forall x \in A),$$
$$f(y) \geq \alpha + \varepsilon \quad (\forall y \in B).$$

Corollary 1. *Let V be a normed vector space over $\mathbb{R}$, and let $U \subset V$ be a proper closed linear subspace. Then there exists a continuous linear functional $f : V \to \mathbb{R}$ such that $f(x) = 0$ for all $x \in U$, while $f \neq 0$.*

Proof. Let $z \in V \setminus U$. Since $\{z\}$ is compact and U is closed and convex, the geometric Hahn–Banach theorem gives a continuous linear functional $f : V \to \mathbb{R}$ and a constant $\alpha > 0$ such that $f(x) < \alpha$ for all $x \in U$, while $f(z) > \alpha$.
Because f is linear and U is a subspace, we have $\lambda x_0 \in U$ for every $x_0 \in U$ and $\lambda \in \mathbb{R}$. Hence $f(\lambda x_0) = \lambda f(x_0) < \alpha$. If $\lambda > 0$, then $f(x_0) < \alpha/\lambda$, and if $\lambda < 0$, then $f(x_0) > \alpha/\lambda$. Taking the limits as $\lambda \to \infty$ and $\lambda \to -\infty$, respectively, forces $f(x_0) = 0$. Therefore, $f(x) = 0$ for all $x \in U$. Since $f(z) > \alpha > 0$, we also have $f \neq 0$. □

With these tools in hand, we are now ready to delve into Cybenko's theorem.

Definition 2 (Discriminatory Function)**.** We say that a function $f : \mathbb{R} \to \mathbb{R}$ is discriminatory if

$$\int_{I_n} f(y \cdot x + \theta) \, d\mu(x) = 0 \qquad \forall y \in \mathbb{R}^n, \ \forall \theta \in \mathbb{R}$$

implies $\mu = 0$.

Definition 3 (Sigmoidal Function)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is called sigmoidal if $\lim_{t\to\infty} f(t) = 1$ and $\lim_{t\to-\infty} f(t) = 0$.

Lemma 1. *Every bounded, Borel measurable, sigmoidal function is discriminatory.*

Proof.

Step 1. Applying the Lebesgue Dominated Convergence Theorem (LDCT). Let f be a bounded, Borel measurable, sigmoidal function. Suppose that

$$\int_{I_n} f(y \cdot x + \theta)\, d\mu(x) = 0 \qquad \forall y \in \mathbb{R}^n,\ \forall \theta \in \mathbb{R}.$$

For fixed $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}$, and $\beta \in \mathbb{R}$, define

$$\begin{aligned}\gamma(x) &:= \lim_{\lambda \to \infty} f\big(\lambda(y \cdot x + \theta) + \beta\big) \\ &= \begin{cases} 1, & y \cdot x + \theta > 0, \\ f(\beta), & y \cdot x + \theta = 0, \\ 0, & y \cdot x + \theta < 0. \end{cases}\end{aligned}$$

Then $f\big(\lambda(y \cdot x + \theta) + \beta\big) \to \gamma(x)$ pointwise. Since f is bounded and μ is finite, the dominated convergence theorem yields

$$\int_{I_n} \gamma(x)\, d\mu(x) = \lim_{\lambda \to \infty} \int_{I_n} f\big(\lambda(y \cdot x + \theta) + \beta\big)\, d\mu(x) = 0.$$

It follows that

$$\begin{aligned}0 &= \int_{I_n} \gamma(x)\, d\mu(x) = \int_{H^+_{y,\theta}} 1\, d\mu(x) + \int_{\Pi_{y,\theta}} f(\beta)\, d\mu(x) + \int_{H^-_{y,\theta}} 0\, d\mu(x) \\ &= \mu(H^+_{y,\theta}) + f(\beta)\mu(\Pi_{y,\theta}),\end{aligned}$$

where $H^+_{y,\theta} := \{x \in I_n : y \cdot x + \theta > 0\}$, $\Pi_{y,\theta} := \{x \in I_n : y \cdot x + \theta = 0\}$, and $H^-_{y,\theta} := \{x \in I_n : y \cdot x + \theta < 0\}$.
Since this identity holds for every β, taking $\beta \to \infty$ gives $\mu(H^+_{y,\theta}) + \mu(\Pi_{y,\theta}) = 0$, while taking $\beta \to -\infty$ gives $\mu(H^+_{y,\theta}) = 0$. Hence

$$\mu(H^+_{y,\theta}) = 0, \qquad \mu(\Pi_{y,\theta}) = 0.$$

If we were only considering positive measures, the proof would already be complete. For signed measures, we need one more argument.

Step 2. Showing that the one-dimensional pushforward measure vanishes. Fix $y \in \mathbb{R}^n$, and define

$$T_y : I_n \to \mathbb{R}, \qquad T_y(x) = y \cdot x.$$

Let

$$\nu_y := (T_y)_{\#}\mu$$

be the pushforward signed measure on $\mathbb{R}$. Thus, for every Borel set $B \subset \mathbb{R}$,

$$\nu_y(B) = \mu(T_y^{-1}(B)) = \mu\big(\{x \in I_n : y \cdot x \in B\}\big).$$

From the previous step, we know that for every $a \in \mathbb{R}$,

$$\mu\big(\{x \in I_n : y \cdot x > a\}\big) = 0, \qquad \mu\big(\{x \in I_n : y \cdot x = a\}\big) = 0.$$

Hence

$$\nu_y((a, \infty)) = 0, \qquad \nu_y([a, \infty)) = 0 \qquad \forall a \in \mathbb{R}.$$

Therefore ν_y vanishes on all half-lines. Since the half-lines generate the Borel σ-algebra on $\mathbb{R}$, a standard monotone class argument implies that $\nu_y = 0$. (See Appendix)

Step 3. Vanishing Fourier transform and conclusion.
Consequently, for every $h \in B_b(\mathbb{R})$,

$$\int_{I_n} h(y \cdot x)\, d\mu(x) = \int_{\mathbb{R}} h(t)\, d\nu_y(t) = 0.$$

Taking $h(t) = \cos t$ and $h(t) = \sin t$, we obtain

$$\int_{I_n} e^{i\, y \cdot x}\, d\mu(x) = \int_{I_n} \big(\cos(y \cdot x) + i \sin(y \cdot x)\big)\, d\mu(x) = 0.$$

Since $y \in \mathbb{R}^n$ was arbitrary, the Fourier transform of μ is identically zero.

Proposition 8.50 (Folland)

Let $\mu_1, \mu_2, \ldots, \mu \in M(\mathbb{R}^n)$. If $\|\mu_k\| \leq C < \infty$ for all k, and $\widehat{\mu_k}(\xi) \to \widehat{\mu}(\xi)$ pointwise for every $\xi \in \mathbb{R}^n$, then $\mu_k \to \mu$ vaguely, i.e.

$$\int_{\mathbb{R}^n} f\, d\mu_k \to \int_{\mathbb{R}^n} f\, d\mu \qquad \forall f \in C_0(\mathbb{R}^n).$$

In our application, extend μ to all of $\mathbb{R}^n$ by setting it to be zero outside I_n. Since $\widehat{\mu}(y) = 0$ for all $y \in \mathbb{R}^n$, we have $\widehat{\mu} = \widehat{0}$. Applying Proposition 8.50 to the constant sequence $\mu_k = \mu$ with limiting measure 0, we obtain

$$\int_{\mathbb{R}^n} g\, d\mu = \int_{\mathbb{R}^n} g\, d0 = 0 \qquad \forall g \in C_0(\mathbb{R}^n).$$

By the uniqueness part of the Riesz representation theorem, $\mu = 0$. Therefore, f is discriminatory. □

Theorem 4 (Universal Approximation Theorem, Cybenko 1989)**.** *Let σ be any continuous sigmoidal function. Then the set*

$$S := \left\{ G(x) = \sum_{k=1}^{N} \alpha_k \sigma(y_k \cdot x + \theta_k) \;\middle|\; y_k \in \mathbb{R}^n, \quad \alpha_k \in \mathbb{R}, \quad \theta_k \in \mathbb{R}, \quad N \in \mathbb{N} \right\}$$

is uniformly dense in $C(I_n)$. In other words, for every $f \in C(I_n)$ and every $\varepsilon > 0$, there exists $G \in S$ such that $\|G - f\|_\infty < \varepsilon$.

Proof. Suppose, for contradiction, that $\overline{S} \neq C(I_n)$. By Corollary 1, there exists a continuous linear functional $F : C(I_n) \to \mathbb{R}$ such that $F(g) = 0$ for all $g \in S$, while $F \neq 0$. By the Riesz representation theorem, there exists a Borel measure μ such that

$$F(g) = \int_{I_n} g(x)\, d\mu(x) \qquad \forall g \in C(I_n).$$

However, since $\sigma(y \cdot x + \theta) \in S$, it follows that

$$\int_{I_n} \sigma(y \cdot x + \theta)\, d\mu(x) = 0 \qquad \forall y \in \mathbb{R}^n,\ \forall \theta \in \mathbb{R}.$$

Because σ is discriminatory, we must have $\mu = 0$. Therefore $F(g) = 0$ for all $g \in C(I_n)$, contradicting $F \neq 0$. Hence S is uniformly dense in $C(I_n)$. □

Remark.

1. The proof of the universal approximation theorem is *not constructive*. It guarantees the existence of suitable parameters, but it does not provide a method for finding them. In practice, finding such parameters requires solving a highly non-convex optimization problem.

2. Cybenko's theorem applies specifically to scalar-valued functions with sigmoidal activations, and therefore does not directly cover modern activations such as ReLU. Universal approximation theorems for ReLU networks, including vector-valued settings, have been established in more recent work.

Recent Results

	Classical UAT	Modern UAT
Width	Unbounded	Bounded
Depth	Bounded	Unbounded

This structural distinction is one reason modern approaches emphasize *deep* learning. Moreover, recent work has established remarkable theoretical bounds on the minimum width required for universal approximation.

Reference	Function Class	Activation	Bounds
Hanin and Sellke (2017)	$C([0,1]^{d_x}, \mathbb{R}^{d_y})$	ReLU	$d_x + 1 \le w_{\min} \le d_x + d_y$
Park et al. (2021)	$L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max\{d_x + 1, d_y\}$
Kim et al. (2024)	$L^p([0,1]^{d_x}, \mathbb{R}^{d_y})$	ReLU	$w_{\min} = \max\{d_x, d_y, 2\}$

In the following sections, we examine several gaps between traditional theory and practice in modern deep learning.

III. Optimization

To find suitable parameters, we minimize the empirical loss function

$$L(x) := \frac{1}{n}\sum_{k=1}^{n} f_k(x),$$

A. Non-convergence of the ADAM Optimizer

Algorithm: ADAM (Adaptive Moment Estimation)

The ADAM update rule is defined as follows:

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\theta J(\theta_t), & v_t = \beta_2 v_{t-1} + (1-\beta_2)\big(\nabla_\theta J(\theta_t)\big)^2, \\ \widehat{m}_t = \dfrac{m_t}{1-\beta_1^t}, & \widehat{v}_t = \dfrac{v_t}{1-\beta_2^t}, \\ \theta_{t+1} = \theta_t - \dfrac{\eta}{\sqrt{\widehat{v}_t} + \varepsilon}\widehat{m}_t. \end{cases}$$

Although Adam often finds high-quality minima in practice, it does not have a general convergence guarantee, even in certain online convex optimization settings.

Theorem 5 (Reddi et al., 2018). *For any constants $\beta_1, \beta_2 \in [0, 1)$ such that $\beta_1 < \sqrt{\beta_2}$, there exists an online convex optimization problem for which ADAM has nonzero average regret; that is,*

$$\frac{R_T}{T} \not\to 0 \qquad \text{as } T \to \infty.$$

Proof sketch. Consider $\mathcal{X} = [-1, 1]$ and the sequence of functions

$$f_t(x) = \begin{cases} Cx, & t \equiv 1 \pmod{C}, \\ -x, & \text{otherwise}, \end{cases} \qquad C \in \mathbb{N}.$$

One can choose C such that

$$\begin{cases} (1-\beta_1)\beta_1^{C-1} C \le 1 - \beta_1^{C-1}, \\ \beta_2^{(C-2)/2} C^2 \le 1, \\ \dfrac{3(1-\beta_1)}{2\sqrt{1-\beta_2}} \left(1 + \dfrac{\gamma(1-\gamma^{C-1})}{1-\gamma}\right) + \dfrac{\beta_1(C/2-1)}{1-\beta_1} < \dfrac{C}{3}, \\ \gamma = \dfrac{\beta_1}{\sqrt{\beta_2}} < 1. \end{cases}$$

Under these conditions, the average regret satisfies $\lim_{T\to\infty} \frac{R(T)}{T} = \frac{2}{C} \neq 0$. □

B. The Edge of Stability

For full-batch gradient descent, define the sharpness at time t by $s_t := \lambda_{\max}(\nabla^2 L(\theta_t))$.

Theorem 6 (Descent Lemma). *If $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$ and $s_t \le \ell$, then*

$$L(\theta_{t+1}) \le L(\theta_t) - \frac{\eta(2 - \eta\ell)}{2} \|\nabla L(\theta_t)\|^2.$$

Thus, classical optimization theory guarantees monotone decrease when $\ell < 2/\eta$. In particular, if $s_t < 2/\eta$, training is stable and if $s_t > 2/\eta$ training is unstable
However, Cohen et al. observed that neural network training often approaches the *edge of stability*, where the sharpness hovers near the critical threshold $2/\eta$. In this regime,

$$s_t \approx \frac{2}{\eta} \qquad \text{or even} \qquad s_t \gtrsim \frac{2}{\eta}.$$

Although the classical monotonicity condition may fail, the loss can still decrease overall. This behavior is not explained by the descent lemma.

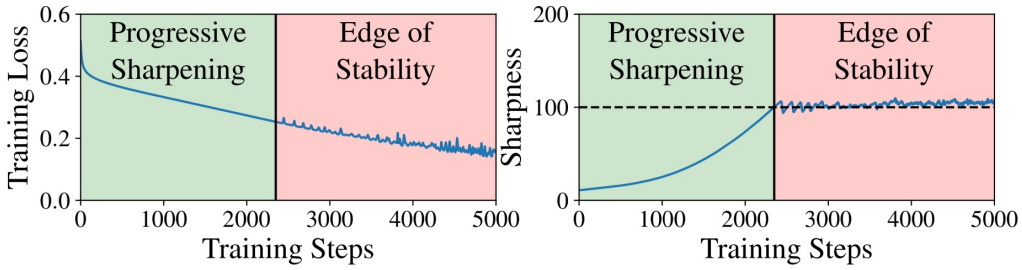


Figure 1: **Progressive Sharpening and Edge of Stability:** We train an MLP on CIFAR10 with learning rate $\eta = 2/100$. It reaches instability after around 2200 training steps after which the sharpness hovers at $2/\eta = 100$, which is denoted by the horizontal dashed line.

Figure 2: Progressive sharpening and the edge of stability.

IV. Generalization

Once training is complete, the model's true performance must be evaluated on unseen data.

A. The Insufficiency of Classical Statistical Learning Theory

Let $\mathcal{Z}$ be a data space, and let D be a distribution on $\mathcal{Z}$. Given an i.i.d. sample $S = \{z_1, \ldots, z_m\} \sim D^m$, let $\mathcal{F}$ be a class of hypothesis functions $f : \mathcal{Z} \to \mathbb{R}$.

Definition 4 (Empirical Rademacher Complexity of $\mathcal{F}$)**.** The empirical Rademacher complexity of $\mathcal{F}$ with respect to a sample S is defined as

$$\widehat{R}_m(\mathcal{F}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right],$$

where the σ_i are independent random variables drawn uniformly from $\{-1, 1\}$.

Definition 5 (Rademacher Complexity of $\mathcal{F}$)**.** The Rademacher complexity of $\mathcal{F}$ is the expectation of the empirical Rademacher complexity over the data distribution:

$$R_m(\mathcal{F}) = \mathbb{E}_D \left[\widehat{R}_m(\mathcal{F}) \right].$$

Intuitively, Rademacher complexity measures the ability of a function class $\mathcal{F}$ to correlate with random signs on a fixed sample. In other words, it quantifies the model class's ability to fit pure noise.

Theorem 7. *Fix a distribution $D|_{\mathcal{Z}}$ and a confidence parameter $\gamma \in (0, 1)$. If*

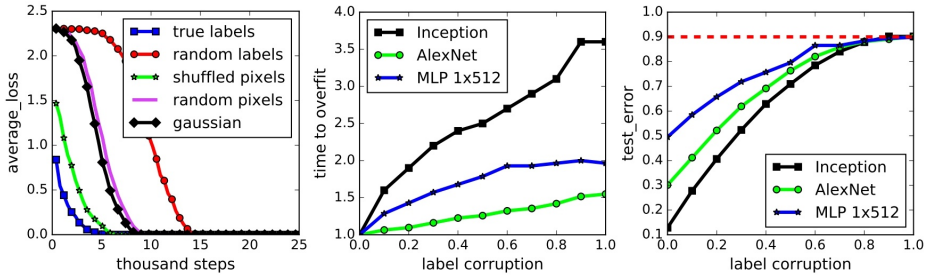
$$\mathcal{F} \subset \{f : \mathcal{Z} \to [a, a + 1]\}$$

and $S = \{z_1, \ldots, z_m\}$ is drawn i.i.d. from $D|_{\mathcal{Z}}$, then with probability at least $1 - \gamma$ over the draw of S, the following bounds hold for every $f \in \mathcal{F}$:

$$\mathbb{E}_D[f(z)] \leq \widehat{\mathbb{E}}_S[f(z)] + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\gamma)}{m}}, \tag{1}$$

$$\mathbb{E}_D[f(z)] \leq \widehat{\mathbb{E}}_S[f(z)] + 2\widehat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(1/\gamma)}{m}}. \tag{2}$$

However, the seminal paper *Understanding Deep Learning Requires Rethinking Generalization* (Zhang et al., ICLR 2017) empirically showed that such bounds can become essentially vacuous for highly overparameterized deep models.



(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 3: Empirical results illustrating the memorization capacity of deep networks.

Surprisingly, modern overparameterized models can memorize completely random labels and easily achieve 100% training accuracy. Consequently, the generalization bound may reduce to

$$\text{Generalization Error} \lesssim \underbrace{\text{Training Error}}_{\approx 0} + \underbrace{\text{Capacity to Fit Randomness}}_{\text{maximized}}.$$

Thus, the traditional upper bound becomes loose and uninformative. Moreover, achieving zero training error does not necessarily imply either good generalization or catastrophic overfitting in the modern regime. Also, explicit regularization techniques, such as ℓ_2 weight decay and dropout, are not always essential for deep networks to generalize well.

B. The Double Descent Phenomenon

Conventional statistical wisdom, often summarized by Occam's razor, is insufficient to explain the behavior of modern deep neural networks.

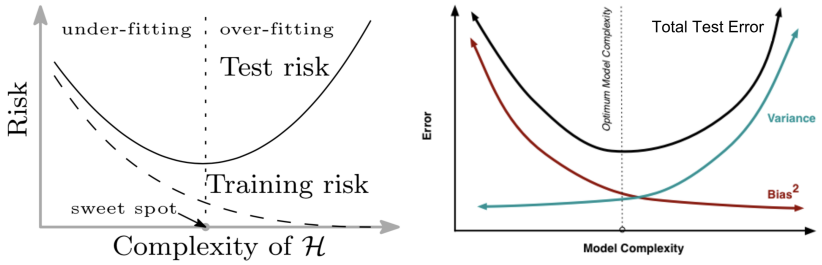


Figure 4: The classical bias–variance tradeoff and the traditional "sweet spot."

The Classical Perspective

In traditional machine learning, one expects a U-shaped risk curve with a clear transition from underfitting to overfitting.

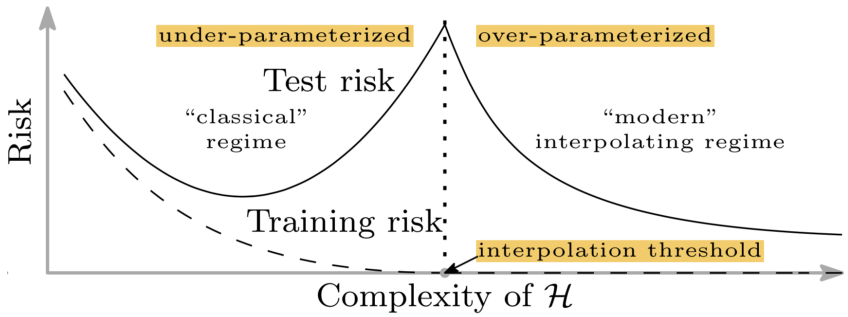


Figure 5: The modern double descent risk curve.

The Modern Regime

1. Parameter count is an inadequate measurement for a model's true complexity.
2. As model capacity increases beyond the interpolation threshold, where training error reaches zero, the test error can decrease again, defying the classical overfitting picture.

					Double-Descent		
Dataset	Architecture	Opt.	Aug.	% Noise	Model	Epoch	Figure(s)
CIFAR 10	CNN	SGD	✓	0	✗	✗	5, 27
			✓	10	✓	✓	5, 27, 6
			✓	20	✓	✓	5, 27
				0	✗	✗	5, 25
				10	✓	✓	5
				20	✓	✓	5
		SGD + w.d.	✓	20	✓	✓	21
		Adam		0	✓	–	25
	ResNet	Adam	✓	0	✗	✗	4, 10
			✓	5	✓	–	4
			✓	10	✓	✓	4, 10
			✓	15	✓	✓	4, 2
			✓	20	✓	✓	4, 9, 10
		Various	✓	20	–	✓	16, 17, 18
(subsampled)	CNN	SGD	✓	10	✓	–	11a
		SGD	✓	20	✓	–	11a, 12
(adversarial)	ResNet	SGD		0	Robust err.	–	26
CIFAR 100	ResNet	Adam	✓	0	✓	✗	4, 19, 10
			✓	10	✓	✓	4, 10
			✓	20	✓	✓	4, 10
	CNN	SGD		0	✓	✗	20
IWSLT '14 de-en	Transformer	Adam		0	✓	✗	8, 24
(subsampled)	Transformer	Adam		0	✓	✗	11b, 23
WMT '14 en-fr	Transformer	Adam		0	✓	✗	8, 24

Figure 6: Summary table of experiment results

V. Implicit Regularization

Almost every nuance of the optimizer can induce a different implicit bias.

- GD prefers parameters with low norm, e.g. ℓ_2 norm, ℓ_1 norm, etc.

  [Gunasekar et al.'17, Li–Liang'18, Jacot et al.'18, Du et al.'18, Li et al.'18, Du–Lee'18, Woodworth et al.'20, Li et al.'20, HaoChen et al.'20, ...]

- GD converges to max-margin predictors.

  [Zhang–Yu'05, Telgarsky'13, Soudry et al.'17, Ji–Telgarsky'18, Gunasekar et al.'18, Nacson et al.'19, Lyu et al.'19, ...]

- GD converges to flat or stable minima.

  [Jastrzebski et al.'19&'20, Cohen et al.'20, Arora et al.'21, Lyu et al.'22, Damian et al.'22, Zhu et al.'22, Li et al.'22, Chen et al.'22, Ma et al.'22, Agarwala et al.'22, Ahn et al.'22, Wu et al.'23, ...]

- SGD with a large learning rate or a small batch size prefers flatter minima.

  [Kleinberg–Li–Yuan'18, Wu–Ma'18, Wei et al.'19, Blanc et al.'20, Damian–Lee–Ma'21, Li–Wang–Arora'21, Ma–Ying'21, Smith et al.'21, Li–Wei–Ma'21, Wu–Wang–Su'22, Wu–Su'23, ...]

VI. Conclusion

Many empirical phenomena in modern deep learning remain only partially explained by current theory. Cybenko's theorem guarantees that neural networks have substantial expressive power, but it does not explain *why* gradient-based training can find useful solutions, nor *how* these solutions generalize so well to unseen data.
Thus, a realistic theory of deep learning must go beyond expressivity:

$$\text{Approximation} \quad + \quad \text{Optimization} \quad + \quad \text{Generalization} \quad + \quad \text{Implicit Bias}$$

This is why Cybenko's universal approximation theorem is a starting point, not a complete explanation of modern deep learning.

References

[1] Maria-Florina Balcan. CS 8803: Machine Learning Theory. *Lecture notes*, 2011.

[2] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences (PNAS)*, 116(32):15849–15854, 2019.

[4] Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[6] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2023.

[7] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. 2nd edition, Wiley, 1999.

[8] Leonardo Ferreira Guilhoto. An overview of artificial neural networks for mathematicians. *University of Chicago Mathematics REU*, 2018.

[9] Namjun Kim, Chanho Min, and Sejun Park. Minimum width for universal approximation using ReLU networks on compact domain. In *International Conference on Learning Representations (ICLR)*, 2024.

[10] Tengyu Ma and Alex Damian. Recent advances in the generalization theory of neural networks. Tutorial at the *International Conference on Machine Learning (ICML)*, 2023.

[11] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018.

[12] Walter Rudin. *Real and Complex Analysis*. 3rd edition, McGraw–Hill, 1987.

[13] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Appendix

A. Monotone Class Argument

We record the measure-theoretic steps used in the proof of Lemma 1.

Definition 6 (Monotone Class)**.** Let X be a set. A collection $\mathcal{M}$ of subsets of X is called a monotone class if it satisfies the following two properties:

1. If $E_1 \subset E_2 \subset \cdots$ and $E_k \in \mathcal{M}$ for every $k \in \mathbb{N}$, then
$$\bigcup_{k=1}^{\infty} E_k \in \mathcal{M}.$$
2. If $E_1 \supset E_2 \supset \cdots$ and $E_k \in \mathcal{M}$ for every $k \in \mathbb{N}$, then
$$\bigcap_{k=1}^{\infty} E_k \in \mathcal{M}.$$

Theorem 8 (Monotone Class Theorem)**.** *Let $\mathcal{A}$ be an algebra of subsets of X. Then the smallest monotone class containing $\mathcal{A}$ is equal to the σ-algebra generated by $\mathcal{A}$. In other words,*

$$m(\mathcal{A}) = \sigma(\mathcal{A}),$$

where $m(\mathcal{A})$ denotes the smallest monotone class containing $\mathcal{A}$.

Proof. Since every σ-algebra is a monotone class, and since $\sigma(\mathcal{A})$ contains $\mathcal{A}$, we immediately have

$$m(\mathcal{A}) \subset \sigma(\mathcal{A}).$$

Thus it remains to show that $m(\mathcal{A})$ is itself a σ-algebra.
First, we show that $m(\mathcal{A})$ is closed under finite intersections. Fix $A \in \mathcal{A}$, and define

$$\mathcal{M}_A := \{E \in m(\mathcal{A}) : E \cap A \in m(\mathcal{A})\}.$$

We claim that $\mathcal{M}_A$ is a monotone class. Indeed, if $E_1 \subset E_2 \subset \cdots$ and $E_k \in \mathcal{M}_A$, then

$$\left(\bigcup_{k=1}^{\infty} E_k\right) \cap A = \bigcup_{k=1}^{\infty} (E_k \cap A).$$

Since $E_k \cap A \in m(\mathcal{A})$ for every k, and since $m(\mathcal{A})$ is a monotone class, we get

$$\left(\bigcup_{k=1}^{\infty} E_k\right) \cap A \in m(\mathcal{A}).$$

Hence $\bigcup_{k=1}^{\infty} E_k \in \mathcal{M}_A$. The decreasing case is similar. Therefore $\mathcal{M}_A$ is a monotone class.
Moreover, since $\mathcal{A}$ is an algebra, for every $B \in \mathcal{A}$,

$$B \cap A \in \mathcal{A} \subset m(\mathcal{A}).$$

Thus $\mathcal{A} \subset \mathcal{M}_A$. Since $m(\mathcal{A})$ is the smallest monotone class containing $\mathcal{A}$, it follows that

$$m(\mathcal{A}) \subset \mathcal{M}_A.$$

Therefore, for every $E \in m(\mathcal{A})$ and every $A \in \mathcal{A}$,

$$E \cap A \in m(\mathcal{A}).$$

Now fix $E \in m(\mathcal{A})$, and define

$$\mathcal{N}_E := \{F \in m(\mathcal{A}) : E \cap F \in m(\mathcal{A})\}.$$

By the same argument as above, $\mathcal{N}_E$ is a monotone class. Also, by the previous paragraph, every $A \in \mathcal{A}$ belongs to $\mathcal{N}_E$. Hence

$$\mathcal{A} \subset \mathcal{N}_E.$$

Since $m(\mathcal{A})$ is the smallest monotone class containing $\mathcal{A}$, we obtain

$$m(\mathcal{A}) \subset \mathcal{N}_E.$$

Thus, for all $E, F \in m(\mathcal{A})$,

$$E \cap F \in m(\mathcal{A}).$$

So $m(\mathcal{A})$ is closed under finite intersections.
Next, we show that $m(\mathcal{A})$ is closed under complements. Define

$$\mathcal{C} := \{E \in m(\mathcal{A}) : E^c \in m(\mathcal{A})\}.$$

We claim that $\mathcal{C}$ is a monotone class. If $E_1 \subset E_2 \subset \cdots$ and $E_k \in \mathcal{C}$, then

$$\left(\bigcup_{k=1}^{\infty} E_k\right)^c = \bigcap_{k=1}^{\infty} E_k^c.$$

Since $E_1^c \supset E_2^c \supset \cdots$ and $E_k^c \in m(\mathcal{A})$, the monotone class property gives

$$\bigcap_{k=1}^{\infty} E_k^c \in m(\mathcal{A}).$$

Therefore $\bigcup_{k=1}^{\infty} E_k \in \mathcal{C}$. The decreasing case is similar. Hence $\mathcal{C}$ is a monotone class.
Since $\mathcal{A}$ is an algebra, it is closed under complements. Therefore

$$\mathcal{A} \subset \mathcal{C}.$$

Again, by the minimality of $m(\mathcal{A})$,

$$m(\mathcal{A}) \subset \mathcal{C}.$$

Thus $m(\mathcal{A})$ is closed under complements.
Finally, since $m(\mathcal{A})$ is closed under complements and finite intersections, it is also closed under finite unions. If $E_1, E_2, \ldots \in m(\mathcal{A})$, then

$$\bigcup_{j=1}^{k} E_j \in m(\mathcal{A}) \qquad \forall k \in \mathbb{N}.$$

Since

$$\bigcup_{j=1}^{k} E_j \uparrow \bigcup_{j=1}^{\infty} E_j,$$

the monotone class property implies that

$$\bigcup_{j=1}^{\infty} E_j \in m(\mathcal{A}).$$

Therefore $m(\mathcal{A})$ is a σ-algebra.
Since $m(\mathcal{A})$ is a σ-algebra containing $\mathcal{A}$, we have

$$\sigma(\mathcal{A}) \subset m(\mathcal{A}).$$

Together with $m(\mathcal{A}) \subset \sigma(\mathcal{A})$, this gives

$$m(\mathcal{A}) = \sigma(\mathcal{A}).$$

□

Proposition 1 (A Monotone Class Argument)**.** *Let ν be a finite signed Borel measure on $\mathbb{R}$. Suppose that*

$$\nu((a,\infty)) = 0 \qquad \forall a \in \mathbb{R}.$$

Then $\nu = 0$.

Proof. First, since $(-m,\infty) \uparrow \mathbb{R}$ as $m \to \infty$, the continuity from below of finite signed measures gives

$$\nu(\mathbb{R}) = \lim_{m\to\infty} \nu((-m,\infty)) = 0.$$

For $a < b$, we have

$$(a,b] = (a,\infty) \setminus (b,\infty).$$

Hence

$$\nu((a,b]) = \nu((a,\infty)) - \nu((b,\infty)) = 0.$$

Also,

$$\nu((-\infty,b]) = \nu(\mathbb{R}) - \nu((b,\infty)) = 0.$$

Let $\mathcal{A}$ be the algebra generated by the half-lines (a,∞). Equivalently, $\mathcal{A}$ consists of finite disjoint unions of intervals of the forms

$$(a,b], \qquad (a,\infty), \qquad (-\infty,b],$$

together with $\emptyset$ and $\mathbb{R}$. By the computations above,

$$\nu(A) = 0 \qquad \forall A \in \mathcal{A}.$$

Define

$$\mathcal{M} := \{B \in \mathcal{B}(\mathbb{R}) : \nu(B) = 0\}.$$

We claim that $\mathcal{M}$ is a monotone class. Indeed, if $B_1 \subset B_2 \subset \cdots$ and $B_k \in \mathcal{M}$, then by continuity from below,

$$\nu\left(\bigcup_{k=1}^{\infty} B_k\right) = \lim_{k\to\infty} \nu(B_k) = 0.$$

Similarly, if $B_1 \supset B_2 \supset \cdots$ and $B_k \in \mathcal{M}$, then by continuity from above,

$$\nu\left(\bigcap_{k=1}^{\infty} B_k\right) = \lim_{k\to\infty} \nu(B_k) = 0.$$

Thus $\mathcal{M}$ is a monotone class containing $\mathcal{A}$.
By the monotone class theorem,

$$\sigma(\mathcal{A}) \subset \mathcal{M}.$$

Since the half-lines generate the Borel σ-algebra on $\mathbb{R}$, we have

$$\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R}).$$

Therefore,

$$\nu(B) = 0 \qquad \forall B \in \mathcal{B}(\mathbb{R}).$$

Hence $\nu = 0$. □