차원의 저주를 넘어: 인공신경망의 고차원 근사 이론

고승찬

바야흐로 인공지능(artificial intelligence)의 시대다. 필자가 박사과정에서 공부하던 2016년 봄, 알파고가 이세돌 9단을 꺾었을 때 많은 사람이 느꼈던 충격을 기억한다. 그로부터 10년이 지난 지금, 인공지능은 더 이상 바둑판 위에서만 펼쳐지는 이야기가 아니다. 우리는 인공지능과 자연어로 대화하고, 인공지능이 그린 그림을 감상하며, 인공지능이 예측한 단백질 구조를 바탕으로 신약을 개발하는 시대에 살고 있다. 단백질 구조 예측에 기여한 연구자들이 노벨 화학상을, 인공신경망(artificial neural network) 연구의 선구자들이 노벨 물리학상을 수상한 사실은 현대 과학기술에서 인공지능이 차지하는 위상을 상징적으로 보여준다. 그러나 이처럼 눈부신 성공의 이면에는 다소 불편한 진실이 자리하고 있다. 딥러닝(deep learning)이 왜 이토록 잘 작동하는지 우리가 아직 충분히 이해하지 못하고 있다는 사실이다. 수억 개에서 수천억 개에 이르는 매개변수(parameter)를 지닌 인공신경망을 방대한 데이터로 훈련하면 놀라운 성능을 얻을 수 있다는 점은 무수한 실험을 통해 확인되었다. 하지만 이러한 성공을 떠받치는 수학적 원리는 여전히 안개 속에 머물러 있다.

과학기술의 역사에서는 경험적 성공이 새로운 이론의 정립을 촉발하고, 뒤이어 엄밀한 수학적 이해가 기술의 신뢰성과 응용 범위를 확장한 사례를 반복해서 찾아볼 수 있다. 증기기관은 충분한 이론적 토대가 마련되기 전에 발명되어 산업혁명을 이끌었다. 그러나 카르노(Nicolas Léonard Sadi Carnot)와 볼츠만(Ludwig Boltzmann)으로 이어지는 열역학 이론이 정립된 뒤에야 인류는 기관의 효율에 근본적인 한계가 있음을 이해하고, 그 한계 안에서 기관을 최적으로 설계하는 법을 익힐 수 있었다. 응용수학의 역사에서는 유한요소법(finite element method)이 유사한 사례를 제공한다. 미분방정식의 해를 계산하기 위한 이 기법의 원형은 1940년대 초 수학자 쿠랑(Richard Courant)이 변분 문제(variational problem)의 근사 해법으로 제시한 것이었으나, 당시에는 주목받지 못한 채 한동안 잊혔다. 이후 1950–60년대에 항공기와 구조물을 설계하던 공학자들이 이를 독자적으로 재발견하여 실용적인 계산 기법으로 발전시켰고, 유한요소법은 이론적 정당화가 마련되기 전부터 산업 현장에서 눈부신 성공을 거두었다. 그러나 유한요소법의 활용이 확산되면서 계산 결과의 신뢰성이 중요한 문제로 떠올랐다. 이에 1970년대를 전후하여 수학자들은 함수해석학(functional analysis)의 언어로 유한요소법의 수렴성(convergence)을 증명하고 오차를 정량적으로 추정하는 이론을 확립하였다. 이를 통해 방법의 적용 범위와 한계가 비로소 분명해졌고, 유한요소법은 오늘날 교량에서 반도체에 이르는 다양한 설계 분야에서 신뢰받는 표준 도구로 자리 잡았다.

필자는 오늘날의 딥러닝이 바로 이 역사의 어느 지점, 곧 열역학이 정립되기 이전의 증기기관과 수렴이론이 확립되기 이전의 유한요소법이 서 있던 자리에 놓여 있다고 생각한다. 딥러닝에 대한 수학적 이해는 단순한 지적 호기심의 대상이 아니다. 그것은 모델이 언제 성공하고 언제 실패하는지 이해하고, 주어진 문제를 해결하기 위해 어느 정도 규모의 모델과 데이터가 필요한지를 예측하며, 나아가 더욱 효율적이고 신뢰할 수 있는 인공지능을 설계하기 위한 필수적인 토대다. 이 글에서는 딥러닝을 이론적으로 이해하려는 여러 접근 가운데 인공신경망의 고차원 근사(high-dimensional approximation) 이론을 간략히 소개하고자 한다. 이를 위해 먼저 '차원의 저주(curse of dimensionality)'라 불리는 근본적인 장벽을 살펴보고, 이어서 이 장벽을 넘어설 하나의 이론적 열쇠를 제공하는 바론 공간(Barron space)의 개념을 설명한다.

신경망 근사로서의 딥러닝

딥러닝에서 말하는 '학습'이란 무엇을 의미할까. 수학자의 관점에서 보면 학습의 본질은 놀랄 만큼 고전적인 문제, 곧 '함수의 근사'로 요약된다. 손글씨 숫자를 인식하는 문제를 생각해 보자. 28×28 화소의 흑백 이미지는 각 화소의 밝기를 나열한 $28 \times 28 = 784$개의 숫자, 즉 784차원 공간의 한 점 $x \in \mathbb{R}^{784}$로 볼 수 있다. 그리고 '이 이미지는 어떤 숫자인가'라는 질문에 대한 이상적인 정답은, 각 이미지 x에 그에 해당하는 숫자를 대응시키는 어떤 함숫값 $f^*(x)$로 생각할 수 있다. 문제는 우리가 이 함수 f^*의 명시적인 공식을 전혀 모른다는 것이다. 우리가 가진 것은 오직 유한개의 예시, 즉 데이터 $(x_1, f^*(x_1)), \dots, (x_N, f^*(x_N))$뿐이다. 딥러닝의 학습이란, 이 데이터를 바탕으로 미지의 타깃 함수 f^*를 인공신경망이라는 도구로 최대한 가깝게

재구성하는 과정이다. 대화형 인공지능도 본질적으로 다르지 않다. 지금까지의 문맥이라는 고차원 입력을 받아 다음 단어의 확률분포를 출력하는, 지극히 복잡한 함수를 근사하고 있는 것이다.

그렇다면 근사의 도구인 인공신경망은 어떠한 함수인가. 가장 단순한 형태인, 은닉층(hidden layer)이 하나인 얕은 신경망(shallow neural network)은 다음과 같이 표현된다.

$$f_n(x) \;=\; \sum_{i=1}^{n} a_i\, \sigma(b_i \cdot x + c_i).$$

여기서 σ는 활성화 함수(activation function)라 불리는 고정된 비선형 함수다. 활성화 함수는 근사 함수 f_n에 비선형성을 부여하여 신경망이 더욱 다양한 형태의 함수를 표현할 수 있도록 한다. 각 항 $\sigma(b_i \cdot x + c_i)$는 하나의 '뉴런(neuron)'에 해당하는데, 기하학적으로는 입력 공간을 방향 b_i로 바라보며 세운 능선 모양의 함수(ridge function)다. 신경망은 이러한 단순한 능선 n개를 가중치 a_i로 중첩하여 복잡한 지형을 빚어내는 구조이며, 학습이란 데이터를 반영하여 매개변수 a_i, b_i, c_i를 조정함으로써 신경망 f_n을 타깃 함수 f^*에 가깝게 만드는 작업이다. 흔히 말하는 '층을 깊게 쌓는다'는 것은 이러한 구조를 여러 차례 합성하는 것에 해당한다.

함수를 단순한 조각들의 합으로 근사한다는 발상 자체는 수학에서 매우 유서 깊은 것이다. 테일러 급수(Taylor series)는 함수를 다항식으로, 푸리에 급수(Fourier series)는 함수를 삼각함수의 합으로 근사한다. 인공신경망은 이 계보에 새로이 더해진 근사의 도구이며, 실제로 1989년 시벤코(George Cybenko)와 1991년 호르니크(Kurt Hornik)는 은닉층이 하나뿐인 신경망이라도 충분히 많은 뉴런을 사용하면 임의의 연속함수를 원하는 정확도까지 근사할 수 있음을 보였다. 이것이 이른바 보편 근사 정리(universal approximation theorem)다. 이 정리는 신경망의 표현력에 대한 이론적인 보증을 제공하였으나, 실용적인 관점에서는 만족스럽지 못하다. '충분히 많은 뉴런'이 과연 어느 정도로 많아야 하는지에 대해서는 아무런 정보를 제공하지 않기 때문이다. 원하는 정확도를 얻는 데 우주의 원자 수보다 많은 뉴런이 필요하다면, 근사가 원리적으로 가능하다는 사실은 실질적인 의미를 갖기 어렵다.

따라서 근사 이론의 본질적인 질문은 정량적인 형태로 제기되어야 한다. 뉴런을 n개 사용할 때 근사 오차는 n이 증가함에 따라 얼마나 빠르게 감소하는가? 근사 이론의 오랜 역사는 이 질문에 답하려면 타깃 함수 f^*에 대한 어떠한 가정이 반드시 필요하다는 사실을 가르쳐 준다. 아무런 규칙성 없이 제멋대로 요동하는 함수를 유한한 데이터와 유한한 수의 뉴런만으로 포착할 수는 없기 때문이다. 함수가 지닌 이러한 '좋은 성질', 이를테면 연속성, 미분 가능성, 도함수의 크기 등을 통틀어 함수의 정칙성(regularity)이라 부른다. 근사 이론의 기본적인 문법은 언제나 다음과 같은 형태를 취한다. '타깃 함수가 어떤 정칙성을 가지면, 근사 오차는 이만큼의 속도로 감소한다'. 그렇다면 인공신경망 근사를 설명하는 데 적합한 정칙성은 과연 무엇일까. 먼저 고전적인 후보들부터 살펴보자.

차원의 저주

해석학에서 함수의 정칙성을 측정하는 가장 표준적인 방법 가운데 하나는 함수가 어느 차수까지 미분 가능한지, 그리고 그 도함수들의 크기가 얼마나 잘 제어되는지를 살펴보는 것이다. 이를 수학적으로 정교하게 표현하는 대표적인 틀이 소볼레프 공간(Sobolev space)이다. 간략히 말하면, 소볼레프 공간은 일정 차수까지의 도함수가 적분 가능한 의미에서 유한한 크기를 갖는 함수들의 모임이다. 소볼레프 공간은 편미분방정식(partial differential equation)을 비롯한 현대 해석학의 표준 언어이므로, 함수 근사의 관점에서 신경망을 연구할 때도 자연스러운 출발점이 된다. 그런데 이 고전적인 잣대를 들이대는 순간, 신경망 근사 이론은 잘 알려진 벽에 부딪힌다. d차원 공간에서 정의된, s번 미분 가능한 정도의 정칙성을 갖는 함수를 뉴런 n개의 얕은 신경망으로 근사하면, 이론적으로 증명 가능한 최선의 오차는 대체로

$$(\text{오차}) \;\sim\; n^{-s/d}$$

의 규모로 나타난다는 것이 알려져 있다. 여기서 우리는 수렴속도(convergence rate)라 불리는 n의 지수에 주목할 필요가 있다. 분자에는 함수의 매끄러움 s가, 분모에는 입력의 차원 d가 있다. 차원 d가 커질수록 지수 s/d는 0에 가까워지고, 이 경우 오차의 감소는 매우 느려진다.

이것이 얼마나 심각한 문제인지 숫자로 체감해 보자. 미분 가능한($s = 1$) 함수를 오차 0.1 이내로 근사하고 싶다고 하자. 위 관계를 형식적으로 뒤집으면 필요한 뉴런의 수는 대략 $n \sim 10^d$이다. 차원이 $d = 2$라면 뉴런

백 개면 충분하다. 그러나 손글씨 이미지처럼 $d = 784$라면 10^{784}개의 뉴런이 필요하다는 터무니없는 계산이 나온다. 관측 가능한 우주의 원자 수가 대략 10^{80}개로 추정된다는 점과 비교해 보아도, 이는 물리적으로 감당할 수 없는 규모다. 이처럼 요구되는 자원이 차원에 대해 지수적으로(exponentially) 폭발하는 이 현상을, 응용수학자 벨만(Richard Bellman)의 표현을 빌려 차원의 저주라 부른다.

여기서 한 가지 의문이 생긴다. 이론은 일반적인 고차원 함수의 근사가 극도로 어렵다고 말하는데, 현실의 딥러닝은 수천 또는 수만 차원의 입력을 다루면서도 천문학적이지 않은 규모의 모델로 유용한 성능을 발휘한다. 물론 이는 이론과 현실의 모순이 아니다. 이론은 주어진 함수군에 속하는 모든 함수를 대상으로 한 최악의 경우에 관한 명제인 반면, 실제 문제에서 나타나는 타깃 함수는 그 가운데 매우 제한되고 구조적인 일부일 수 있기 때문이다. 따라서 우리는 자연스럽게 다음과 같은 결론에 이른다. 현실의 문제에서 등장하는 타깃 함수들은 '임의의 매끄러운 함수'가 아니라, 소볼레프 정칙성과는 결이 다른 어떤 특별한 구조를 지니고 있고, 신경망은 바로 그 구조를 활용하는 데 탁월한 도구라는 것이다. 그렇다면 그 '특별한 구조'를 수학적으로 어떻게 포착할 수 있을까? 이 질문에 대한 최초의, 그리고 가장 우아한 답 중 하나가 1993년 바론(Andrew Barron)의 연구에서 제시되었다.

바론 공간: 주파수 영역에서의 정칙성

바론의 아이디어를 이해하기 위해서는 먼저, 함수를 다양한 주파수(frequency)의 파동들로 분해하는 도구인 푸리에 해석(Fourier analysis)에 대한 직관이 필요하다. 주기함수 f는 적절한 조건 아래에서

$$f(x) \;=\; \sum_{n=-\infty}^{\infty} \widehat{f}(n)\, e^{inx}$$

형태의 무한합으로 전개된다. 여기서 e^{inx}는 오일러 공식(Euler's formula)에 의해 $\cos(nx)$와 $\sin(nx)$를 담고 있는, 정수 주파수 n으로 진동하는 파동이며, 푸리에 계수(Fourier coefficient)

$$\widehat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, e^{-inx}\, \mathrm{d}x$$

는 그 파동이 f에 포함된 세기(진폭)를 나타낸다. 주파수가 낮은 파동은 완만하게 굽이치는 성분을, 주파수가 높은 파동은 잘게 요동치는 성분을 담당하므로, 이 전개는 함수를 느린 변화에서 빠른 변화에 이르는 성분들로 층층이 분해하는 것이라 할 수 있다. 이러한 분해는 주기함수에 국한되지 않는다. $\mathbb{R}^d$ 위의 일반적인 함수에 대해서는 푸리에 변환(Fourier transform)

$$\widehat{f}(\omega) = \int_{\mathbb{R}^d} f(x)\, e^{-i\omega\cdot x}\, \mathrm{d}x$$

가 계수의 역할을 이어받으며, 함수는 푸리에 반전 공식(Fourier inversion formula)

$$f(x) \;=\; \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(\omega)\, e^{i\omega\cdot x}\, \mathrm{d}\omega$$

에 의해 모든 주파수의 파동 $e^{i\omega\cdot x}$를 각각 가중치 $\widehat{f}(\omega)$만큼씩 겹쳐 놓은 중첩으로 표현된다. 이산적인 주파수에 대한 합이 연속적인 주파수에 대한 적분으로 바뀌었을 뿐, '함수는 파동들의 합'이라는 구조는 동일한 것이다. 이 관점에서 함수의 매끄러움은 스펙트럼(spectrum)의 감쇠(decay)라는 언어로 번역된다. 위의 전개를 미분하면 각 파동에 주파수가 곱해지므로, 도함수의 크기가 적절하게 유지되기 위해서는 고주파 성분이 충분히 작아야 하기 때문이다. 요컨대 함수가 매끄럽다는 것은 고주파 성분이 미미하다는 것, 즉 $\widehat{f}(\omega)$가 $|\omega|$가 커짐에 따라 빠르게 감쇠한다는 것과 대응된다.

이와 같은 관점에서 바론은 다음과 같은 양에 주목하였다. 주어진 함수 f에 대하여, 그는

$$\|f\|_{\mathcal{B}} \;=\; \int_{\mathbb{R}^d} |\omega|\, |\widehat{f}(\omega)|\, \mathrm{d}\omega$$

로 정의되는 값을 고려하였고, 이 값이 유한한 함수를 오늘날 우리는 바론 함수, 그 모임을 바론 공간이라 부른다. 식의 의미를 풀어 보면, 주파수 ω 성분의 세기 $|\widehat{f}(\omega)|$에 주파수의 크기 $|\omega|$를 곱하여 모든 주파수에

걸쳐 적분한 것으로, 고주파일수록 큰 값의 $|\omega|$를 곱하여 합산하는 셈이다. 따라서 이 값이 유한하기 위해서는 고주파 성분이 충분히 빠르게 감쇠하여야 한다. 요컨대 $\|f\|_{\mathcal{B}}$ 역시 함수의 매끄러움을 측정하는 하나의 잣대인데, 미분 가능한 횟수를 세는 방식이 아니라 전체 주파수 스펙트럼에 총량의 제한을 두는 방식이라는 점에서 고전적 정칙성과 구별된다.

이 정의가 신경망과 잘 어울리는 이유는 무엇일까? 핵심은 반전 공식의 파동 $e^{i\omega\cdot x}$와 뉴런 $\sigma(b\cdot x+c)$가 같은 기하학적 구조를 지닌다는 데 있다. 둘 다 한 방향(ω 혹은 b)으로만 변하고 그에 수직인 방향으로는 일정한, 이른바 능선 함수다. 따라서 함수가 파동들의 중첩으로 표현된다면, 각 파동을 뉴런으로 대체하여 뉴런들의 중첩으로도 표현할 수 있으리라 기대할 수 있다. 여기서 바론의 결정적인 아이디어가 등장한다. 반전 공식의 적분을 신경망의 형태로 표현한 뒤, $\|f\|_{\mathcal{B}}$으로 나누어 정규화하면, $|\widehat{f}(\omega)|$에 비례하는 확률분포를 주파수 공간 위에 정의할 수 있고, 이때 f는 그 분포에 대한 파동들의 기댓값으로 표현된다. 즉 고주파일수록 뽑힐 확률이 낮은 추첨의 형태가 된다. 함수가 이처럼 기댓값으로 주어진다면, 이를 유한한 n개의 항으로 근사하는 문제는 평균을 표본으로 재현하는 문제로 환원되는데, 이는 잘 알려진 확률론의 영역이다. 요컨대 무작위 추출(random sampling)을 통해 적분값을 근사하는 몬테카를로(Monte Carlo) 방법은 차원에 관계없이 $1/\sqrt{n}$의 수렴 속도를 유지한다. 바론은 이 논증을 함수 근사에 적용하였다. 위의 분포로부터 주파수 $\omega_1,\dots,\omega_n$을 추출하고 각 파동을 뉴런으로 대체한 뒤 평균을 취하면, 표준적인 분산 계산에 의해

$$\left\| f - \frac{1}{n}\sum_{i=1}^{n} a_i\,\sigma(b_i\cdot x + c_i) \right\| \;\lesssim\; \frac{\|f\|_{\mathcal{B}}}{\sqrt{n}}$$

이 성립함을 증명할 수 있고, 이 부등식이 바론 정리의 핵심이다. 우변에서 뉴런 수 n의 지수는 $-1/2$로 고정되어 있으며, 차원 d는 어디에도 등장하지 않는다. 소볼레프 정칙성 아래에서의 $n^{-s/d}$와 극명하게 대비되는, 차원의 저주로부터 자유로운 근사율이다. 상수 $\|f\|_{\mathcal{B}}$가 감당 가능한 크기라는 전제 아래, 오차 0.1을 얻는 데에는 차원이 784이든 백만이든 뉴런 백 개 규모면 충분한 것이다.

이 결과는 어떻게 해석되어야 할까? 저주가 소멸한 것은 아니다. 저주는 바론이 제시한 양이 유한하고 적당한 크기라는 가정 속으로 자리를 옮겼을 뿐이다. 고차원에서 임의의 함수가 바론 함수가 되는 것은 아니며, 바론 함수라 하더라도 그 값이 차원에 따라 지수적으로 커진다면 정리는 실질적인 의미를 갖지 못한다. 그럼에도 이것은 중요한 관점의 전환을 시사한다. 함수 근사의 난이도는 입력 공간의 차원만으로 결정되지 않으며, 타깃 함수가 지닌 스펙트럼 구조에 크게 좌우된다는 것이다. 실제로 바론이 보였듯 가우시안(Gaussian) 함수를 비롯한 여러 함수족은 차원에 비해 온건한 크기의 바론 정칙성을 만족한다. 아울러 이 논의는 '왜 딥러닝에서 신경망을 사용하는가'라는 물음에도 일정 부분 답을 제시한다. 바론 공간의 함수들을 근사할 때, 타깃 함수와 무관하게 미리 고정해 둔 기저의 선형결합을 사용하는 방식으로는 차원의 저주를 피할 수 없음이 알려져 있다. 신경망의 힘은 뉴런의 방향 b_i를 타깃 함수에 맞추어 적응적으로 선택할 수 있다는 데서 비롯되는 것이다.

로그-바론 공간: 심층 신경망의 근사

글을 마무리하기에 앞서, 이와 같은 관점에서 진행된 필자의 최근 연구를 간략하게 소개하려 한다. 앞서 소개한 바론의 이론은 가장 단순한 신경망 구조, 곧 두 개 층으로 이루어진 신경망의 근사에 관한 것이다. 그러나 실제로 쓰이는 신경망은 여러 층을 쌓아 올린 심층 신경망(deep neural network)이므로, 바론의 이론을 이러한 구조로 확장할 필요가 있다. 실제로 최근 랴오(Yulei Liao)와 밍(Pingbing Ming) 등의 후속 연구를 통해 바론의 이론은 깊은 신경망으로 확장되었다. 그런데 이러한 결과를 포함한 종래의 모든 연구는 한 가지 특징을 공유한다. 근사를 진행하는 방식이 모두 신경망의 깊이(depth)는 고정한 채 폭(width)을 넓히는 데 있다는 점이다. 은닉층 하나에 뉴런을 더 많이 두거나, 층수는 정해 둔 채 각 층의 폭을 넓히는 식이다. 그러나 현대 딥러닝의 성공은 오히려 폭보다 깊이에서 비롯되며, 경험적으로도 층을 깊게 쌓는 편이 넓게 펼치는 편보다 더 우수한 성능을 내는 경우가 보고되고 있다. 그렇다면 폭을 고정한 채 깊이를 늘려 가는 깊은 신경망에 대해서도 차원에 무관한 근사 이론을 정립할 수 있을까? 이것이 필자와 공동 연구자들이 최근 연구에서 답하고자 한 물음이다.

이를 위해 우리는 바론 공간보다 정칙성 가정을 완화한 로그-바론 공간(log-Barron space)을 도입하였다. 바론이 제시한 양이 고주파 성분에 $|\omega|$에 비례하는 가중치를 부여하였다면, 로그-바론 공간을 정의하는 다음

양은 그 가중치를 로그 규모로 낮춘다.

$$\|f\|_{\mathcal{B}^{\log}} \;=\; \int_{\mathbb{R}^d} \log_2\bigl(2+|\omega|\bigr)\,|\widehat{f}(\omega)|\,\mathrm{d}\omega.$$

즉, 가중함수가 $|\omega|$에서 $\log|\omega|$로 완만해진 만큼, 이 값이 유한하기 위해 요구되는 고주파 감쇠 역시 훨씬 느슨해진다. 실제로 바론 공간은 로그-바론 공간에 포함되며, 후자는 바론 조건을 만족하지 않는 함수, 곧 스펙트럼이 한층 느리게 감쇠하는 함수까지 폭넓게 아우른다. 요컨대 로그-바론 공간은 바론 공간보다 매끄러움을 덜 요구하는, 더 넓은 함수들의 모임이다.

이 완화된 공간 위에서 필자와 공동 연구자들은 다음을 증명하였다. 로그-바론 함수는 깊이를 m에 비례하여 늘려 가는 심층 신경망 F_m으로 근사되며, 그 오차는

$$\|f - F_m\|_{L^2(\Omega)} \;\lesssim\; \frac{\|f\|_{\mathcal{B}^{\log}}}{\sqrt{m}}$$

을 만족한다. 바론의 정리에서 폭 n이 담당하던 역할을 이제 깊이 m이 이어받되, 수렴 속도 $m^{-1/2}$의 지수에는 여전히 차원 d가 등장하지 않는다. 증명의 골격은 바론의 그것과 같은 확률적 논증, 곧 함수를 뉴런들의 기댓값으로 표현하여 표본 평균으로 근사하는 방식이다. 다만 표본으로 뽑힌 각 성분을 좁고 깊은 하위 신경망으로 구현한 뒤 이들을 하나의 깊은 신경망으로 병합하는 구성이 핵심을 이룬다. 이는 층수가 늘어나는 좁고 깊은 신경망에 대하여 차원에 무관한 수렴률을 정량적으로 확립한 최초의 결과다. 이로써 깊이를 늘리는 것은 폭을 넓히는 것과 마찬가지로, 나아가 더 넓은 함수 부류에 대하여 차원의 저주를 넘어서는 유효한 전략임이 이론적으로 뒷받침되었으며, 이는 현대의 딥러닝 아키텍처가 왜 깊이를 선호하는지에 대한 하나의 수학적 설명을 제공한다.

남은 과제와 전망

이제까지의 논의를 되짚어 보자. 딥러닝의 학습은 본질적으로 고차원 함수의 근사 문제이며, 고전적 정칙성 이론에서는 차원의 저주라는 근본적 한계에 직면한다. 바론은 함수를 주파수의 관점에서 재해석함으로써, 입력 차원에 직접 의존하지 않는 신경망 근사를 가능하게 하는 함수의 구조적 성질을 포착하였다. 이러한 이론은 현대 인공지능 모델의 작동 원리를 이해하는 데에도 중요한 출발점을 제공한다. 특히 2017년에 제안된 트랜스포머(transformer)는 오늘날 거대 언어 모델(large language model)의 핵심 구조이지만, 앞서 다룬 얕은 신경망과는 수학적으로 상당히 다른 대상이다. 트랜스포머는 고정 차원의 벡터를 입력받는 함수가 아니라, 가변 길이 수열을 다른 수열로 대응시키며, 핵심 연산인 어텐션(attention)에는 토큰(token) 사이의 곱셈적 상호작용이 내재한다. 문맥은 수십만 개의 토큰으로 이루어질 수 있고 각 토큰은 고차원 벡터로 표현되므로, 이를 단순한 벡터 함수로 환원할 경우 함수의 입력 차원은 앞서 다룬 신경망의 경우보다 훨씬 더 커진다. 그럼에도 실제 거대 언어 모델은 지수적으로 증가하지 않는 계산 자원으로 언어 데이터의 복잡한 규칙성을 상당한 수준까지 학습한다. 이는 바론의 이론이 시사한 바와 마찬가지로, 현실의 학습 대상이 임의의 고차원 함수가 아니라 특정한 구조적 성질을 지니며, 트랜스포머가 그 구조를 효과적으로 활용하고 있음을 의미한다. 다만 그 성질이 무엇이며, 그것이 트랜스포머의 구조 및 근사 능력과 어떻게 연결되는지를 엄밀하게 규명하는 이론은 아직 초기 단계에 머물러 있다.

물론 성과가 전혀 없는 것은 아니다. 트랜스포머의 보편 근사 정리는 이미 확립되었으며, 어텐션이 특정 함수 부류를 얼마나 효율적으로 표현할 수 있는지에 관한 정량적 결과도 축적되고 있다. 그러나 바론 공간이 얕은 신경망에 대해 수행한 역할, 곧 트랜스포머가 차원의 저주를 완화하며 근사할 수 있는 자연스러운 함수공간을 규정하고 명시적인 근사율을 제시하는 이론은 아직 구체적으로 확립되지 않았다. 이러한 이론의 의의는 학문적 성취에 그치지 않는다. 오늘날 거대 언어 모델의 발전을 이끄는 경험적 스케일링 법칙(empirical scaling laws)은 막대한 계산 비용과 전력 소비를 수반한다. 어떤 구조의 함수가 어느 정도의 자원으로 근사될 수 있는지를 밝히는 정량적 이론은 그 법칙에 수학적 근거를 제공하고, 궁극적으로는 동일한 성능을 더 적은 자원으로 달성하는 아키텍처의 설계로 이어질 수 있다. 증기기관의 발전을 열역학이 뒷받침하고, 유한요소법의 발전을 수치해석학이 이끌었던 것과 마찬가지다.

이러한 논의에서 흥미로운 점은, 인공지능이라는 가장 현대적인 기술의 중심에서 푸리에 해석, 확률론, 함수해석학과 같은 고전적인 수학이 핵심적인 역할을 담당한다는 사실이다. 한 세기 전 조화해석학자들이 설계해 놓은 도구가 오늘날 인공지능을 이해하는 언어가 되리라고는 누구도 예상하지 못하였을 것이다. 그런

의미에서 인공지능을 연구하는 수학자에게 가장 필요한 덕목은 자신의 전공 도구에 머무르지 않고 필요한 이론을 폭넓게 배워 연결하려는 개방적인 태도가 아닐까 한다.

인공지능의 실험적 발전에 비해 수학적 이해는 여전히 뒤처져 있다. 그러나 증기기관과 열역학, 유한요소법과 그 수렴 이론이 보여 주듯, 기술은 이론이 경험을 따라잡을 때 비로소 성숙한 학문적 체계를 갖춘다. 차원의 저주를 넘어서는 함수공간에 관한 연구는 그러한 진전의 한 단면이며, 앞으로도 다양한 수학의 이론이 새로운 방식으로 결합되어 인공지능의 원리를 밝히고 더 풍성한 결실로 이어지기를 기대한다.

References

[1] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information Theory, 39, 930–945 (1993).

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

[3] K. Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural Networks, 4, 251–257 (1991).

[4] C. Song, S. Ko, Y. Hong, *Sobolev approximation of deep ReLU networks in log-Barron space*, preprint, arXiv:2601.01295 (2026).

[5] A. Vaswani et al., *Attention is all you need*, Advances in Neural Information Processing Systems, 30 (2017).