구간별 정밀도 제어를 활용한 합성 Minimax 다항식의 ReLU 함수 근사

최윤제, 이은상
세종대학교

chldbswp123@gmail.com, eslee3209@sejong.ac.kr

ReLU Approximation via Composite Minimax Polynomials with Interval-Wise Precision Control

요 약

본 논문에서는 minimax 근사 다항식에서 구간별로 정밀도를 다르게 두는 근사 방식과 근사 다항식의 합성을 이용하여 부호 함수를 근사하며 더 나아가 인공신경망에서 활성화 함수로 많이 사용되는 ReLU 함수를 최적으로 근사하는 방법을 제안한다. 우리는 이 근사 방식이 이전 연구[1]와 비교했을 때, 15 차, 15 차 다항식의 합성 근사 환경에서 ReLU 함수에 대한 평균절대오차를 약 27% 정도 감소시키는 것을 확인하였다.

Ⅰ. 서론

최근 인공지능 서비스가 급속히 확산되면서, 서버 측에 민감한 데이터를 위탁하는 과정에서 프라이버시 침해에 대한 우려가 커지고 있다. 이에 따라 Privacy-Preserving Machine Learning(PPML)에 대한 관심이 높아지고 있으며, 그 구현 방법 중 하나로 동형암호가 주목받고 있다. 동형암호는 암호화된 상태에서 대수 연산을 지원하는 특수 암호화 기법으로, 데이터를 복호화하지 않고 연산을 수행함으로써 개인정보 유출 위험을 원천적으로 차단할 수 있다는 장점이 있다. 그러나 최대/최소 함수나 비교함수처럼 비다항식 함수에 대해서는 계산 효율이 크게 저하된다는 한계가 존재한다.

본 연구에서는 비다항식 함수인 부호 함수를 다항식으로 근사하기 위해, 합성 다항식 구조와 구간별 정밀도 차등을 적용한 minimax 근사기법을 결합한 새로운 방법을 제안한다. 특히, 구간별로 서로 다른 오차를 부여해 최소제곱 근사 방식을 사용한 선행연구[2]의 아이디어를 참고하되, 본 연구는 이를 minimax 근사에 적용하면서 구간 가중화와 합성 구조를 결합하였다. 그 결과, 전 구간에 동일한 정밀도를 부여하던 기존 방식[1]과 달리, 구간별 중요도에 따라 근사 정밀도를 조절함으로써 보다 효율적이고 정확한 근사를 달성한다. 나아가, 이렇게 얻은 부호 함수 근사 다항식을 활용하여 인공신경망에서 널리 사용되는 $ReLU$ 활성화 함수를 동형암호 친화적으로 근사하는 것을 최종 목표로 한다.

Ⅱ. 본론

기존 연구[1]는 $[-1,1]$ 구간에서 부호함수를 minimax 근사다항식으로 근사하고 2~3 번 합성하여 정밀도를 높이는 방식으로 효과적인 근사를 달성했다. 부호함수에 대한 근사다항식을 $p(x)$ 라고 하면 $x(1+p(x))/2$ 는 $\mathrm{ReLU(x)}$ 를 잘 근사하는 함수가 되며, 이는 주어진 정밀도 파라미터 α에 대해 입력 $[-1,1]$에서 $2^{-\alpha}$ 의 최대 오차를 가지고, 전 범위 $[-1,1]$ 에서 동일한 최대오차를 가지게 된다.

본 연구에선 입력 분포를 반영해서 분포가 많은 범위는 에러에 더 민감한 범위라 정하고 분포가 적은 범위는 덜 민감한 범위라고 정한다. 그리고 민감한 범위와 덜 민감함 범위에 오차를 달리 두는 최적화된 ReLU 근사방법을 제안한다. 즉, 근사다항식의 입력분포에 따라 범위별로 다르게 근사정확도를 둔다. 그리고 입력 분포를 표준편차가 1/6 인 정규분포를 따른다고 설정하였고, minimax 근사방법을 사용하여 15 차, 15 차 합성 다항식 근사로 상황을 정하고 연구를 진행하였다.

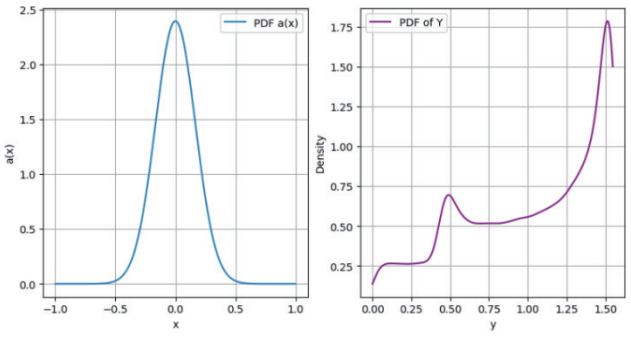

그림 1. 왼쪽 그림은 표준편차가 1/6 인 정규분포, 오른쪽 그림은 표준편차가 1/6 인 정규분포를 입력으로 하여, 부호 함수에 대한 15 차 minimax 근사 다항식을 통과시킨 결과로 얻어진 출력 분포를 나타냄

그림 1 의 오른쪽 그림에서 확인되듯, 표준편차가 1/6 인 정규분포를 부호 함수의 15 차 minimax 근사다항식에 통과시킨 결과, 출력 분포는 균등하지 않게 나타난다. 이에 본 연구는 이 비균등성을 반영하여 두 번째 15 차 근사다항식에서 구간별 민감도를 달리 설정하였다. 구체적으로 분포가 민감한 구간에서는 minimax 근사 오차를 더 작게 설정해 정밀도를 높이고, 덜 민감한 구간에서는 비교적 크게 설정해 거친 근사를 허용한다. 예시로 $[0.45245,1]$ 구간에서는 $[1-1.5E, 1+1.5E]$, $[1,1.54758]$ 구간에서는 $[1-0.5E, 1+0.5E]$ 안에

들도록 제약한다. 여기서 E 는 해당 구간의 기준 오차(최대 편차 상한)를 의미한다.

본 연구에서는 표준편차가 1/6인 정규분포를 $\phi(x)$라 하고, 근사 다항식을 $p(x)$, 대상 함수는 $f(x)$, 오차는 $E(x) = p(x) - f(x)$ 로 두었을 때, 다음의 평균절대오차(MAE)를 최소화하는 것을 목표로 한다.

$$\mathrm{MAE} = \int_{-1}^{1} \phi(x)|E(x)|dx$$

ReLU 근사는 부호함수 $sgn(x)$ 근사로 환원될 수 있다. 이에 따라 ε>0을 작은 수로 두고, 해석을 두 영역으로 나누어 진행한다.

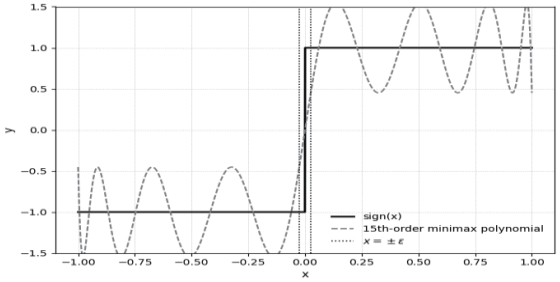

그림 2. 부호함수의 15 차 minimax 근사 다항식의 그래프

1. $|x| \leq \varepsilon$ (부호 함수의 0 근방)

그림 2에서 보이듯이, $sgn(x)$ 의 불연속성 때문에 이 구간에서 근사 오차를 균일하게 작게 만드는 것은 본질적으로 어렵다. 다만 ε 을 충분히 작게 선택하면, 입력 x의 절대값 자체가 작아져 이 구간이 전체 MAE에 영향을 거의 끼치지 못하도록 만들 수 있다.

2. $|x| > \varepsilon$ (불연속점 바깥)

여기서는 ReLU(x) 의 MAE 를 최소화하는 대신 $sgn(x)$ 근사로 문제를 바꿔 다룬다. $Sgn(x)$ 가 홀함수 이므로 $p(x)$도 홀다항식으로 두면, 범위를 $x \geq 0$ 에 한정할 수 있다. 이때 오차항은 결국

$$E(x) = \frac{1+sgn(x)}{2} - \frac{1+p(x)}{2}$$

$$E(x) = \frac{1}{2}|1 - p(x)|$$

로 귀결되고, 따라서 $|x| > \varepsilon$ 영역의 목표

$$MAE \propto \int_{\varepsilon}^{1} \phi\left(x\right)|1 - p(x)|\,dx$$

를 최소화하는 것으로 정리된다..

가중함수	MAE
-	7.68468153978e-05
$1.0 + 0..686 * sgn(-x + 1.0)$	5.94451455143e-05
$1.0 + 0.85 * \frac{0.65 - x}{\sqrt{(0.65 - x)^2} + 0.1}$	**5.61095352826e-05**

표 1. $\boldsymbol{ReLU}$함수의 합성 근사다항식의 MAE

마지막으로, 표 1 은 두 번째 15 차 합성 근사다항식에 대해 구간별 가중함수를 도입하여 minimax 근사 오차의 중요도를 차등화 했을 때의 $ReLU(x)$ 에 대한 MAE 를 요약한 결과이다. 이를 통해 불연속점 인근($|x| \leq \varepsilon$)을 적절히 제어하면서, $|x| > \varepsilon$ 구간의 정밀도를 향상시킬 수 있음을 확인하였다.

Ⅲ. 결론

본 연구는 minimax 합성 근사 방법을 통해 $ReLU$함수를 15 차, 15 차 합성 다항식으로 근사하되, 두 번째 근사 단계에서 구간별 가중함수를 도입하여 민감 구간의 minimax 오차를 더 엄격히 제어하고 비민감 구간에서는 비교적 느슨하게 허용하는 방법을 제안했다.. 그 결과, 표준편차 1/6 의 입력 분포(가우시안)를 가정한 환경에서 $ReLU$의 평균절대오차(MAE)를 유의미하게 감소시켰으며, 동형암호 기반 연산에서의 활성화 근사 품질을 향상시킬 수 있음을 보였다. 다만 실무 수준의 성능을 위해서는 여전히 추가적인 MAE 절감이 요구되며, 이에 대한 후속 연구의 여지가 크다.

참 고 문 헌

[1] E. Lee, J. -W. Lee, J. -S. No and Y. -S. Kim, "Minimax Approximation of Sign Function by Composite Polynomial for Homomorphic Comparison," in *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 6, pp. 3711-3727, 1 Nov.-Dec. 2022

[2] J. Lee, E. Lee, Y.-S. Kim, Y. Lee, J.-W. Lee, Y. Kim, and J.-S. No, "Optimized layerwise approximation for efficient private inference on fully homomorphic encryption," 2023, arXiv:2310.10349.