볼록함수를 목적함수로 둔 시계열 분수계 경사하강법의 최적화 불가능성

정보성°, 김도윤, 이동헌
고려대학교 수학과
{2018160026, doyoon_kim, holy}@korea.ac.kr

Impossibility of Optimizing Time-Fractional Gradient Descent With a Convex Function As the Objective Function

Bosung Jung, Doyoon Kim, Donghun Lee
Department of Mathematics, Korea University

요 약

현재의 기계학습 모델에서는 경사하강법에 다양한 목적함수와 그의 미분을 사용한다. 일부 연구는 목적함수와 옵티마이저에 분수계 미적분을 적용할 때의 성능 변화를 조사하고, 분수계 미적분을 경사하강법에 적용하였을 때 특정 조건 하에서 극솟값이 없음을 보이며 그 필요성을 주장하고 있다. 본 논문에서는 볼록함수에 시계열 분수계 경사하강법을 적용하여 알고리즘화하였고, 최솟값으로의 최적화가 불가능하단 것을 가장 간단한 볼록함수에 적용하여 수학적으로 증명하고 실험으로 검증하였다. 이러한 결과는 분수계 미적분을 사용하는 모델의 최적화 가능성을 보장하기 위해 주어진 목적함수에 대해서 추가적인 수학적 분석이 필요함을 시사한다.

1. 서 론

현재의 기계학습 모델들은 다양한 방식으로 목적함수를 최소화하도록 설계되어 있다. 미분가능한 목적함수를 $f(\cdot)$라고 하였을 때, 매개변수 x을 다음과 같이 반복 적용하는

$$x_{t+1} = x_t - \eta \frac{df}{dx}(x_t),\ \eta > 0 \tag{1}$$

경사하강법 알고리즘은, 매개변수 x를 t에 관한 함수로 여겨 이에 대한 도함수가 $\frac{df}{dx}(x(t))$와 같다는 다음 미분방정식 형태의 수식을 이산화한 것으로 볼 수 있다:

$$\frac{d}{dt}x(t) = -\frac{df}{dx}(x(t)). \tag{2}$$

경사하강법은 이 방정식의 해를 통해 $f(\cdot)$의 최솟값을 나타낼 수 있는 매개변수 x를 찾는 것을 목표로 둔다. 하지만 (2)는 목적함수가 최솟값뿐만이 아닌 극솟값에서도 수렴하기에 현재 학계에선 이 문제를 해결하기 위해 경사하강법에 Adam, RMSProp, Momentum등의 휴리스틱(heuristic)을 적용하여 사용하고 있다[1].

이러한 시도와 함께 등장한 것 중 하나가 분수계 미적분학의 적용이었다. 목적함수와 옵티마이저(Optimizer)에 분수계 미분을 적용하여 가변한 미분계수를 통해 기계학습 모델 내의 조절할 수 있는 하이퍼파라미터(hyperparameter)를 늘리는 효과가 있다[2, 3]. 또한, 수학적 분석을 통해 특정 조건을 만족한다면 경사하강법에 분수계 미분을 적용하였을 때 극솟값으로 수렴하지 않고 최솟값을 향해 수렴한다는 연구 결과들이 존재한다[4]. 하지만 분수계 미분이 적용된 경사하강법의 수렴성과 최적화 연구는 대부분 (2)의 우변을 분수계 미분으로 변경하여 이루어졌다. 좌변, 즉 시간변수의 미분에 분수계 미적분학을 적용하는 연구[5]는 많지 않다.

따라서, 본 논문에선 시간변수에 대한 분수계 미분을 적용한 경사하강법의 알고리즘화를 연구하고, 이를 바탕으로 일반적인 볼록함수에서의 시계열에 대한 분수계 경사하강법의 수렴 특성을 조사하려고 한다. 특히, 본 논문 2장에서는 분수계 미적분의 여러 정의 중 Caputo의 정의[6]를 활용한 시간변수의 분수계 미분 경사하강법을 제안하고, 3장에서는 볼록함수 $f(x) = x$를 목적함수로 채택하더라도 해당 알고리즘이 보편적인 종료 조건에서 최솟값으로 최적화되지 않는 것을 수식과 실험으로 검증하였다.

2. 분수계 미분과 경사하강법

2.1 분수계 미적 분의 두가지 정의 및 특징

분수계 미적분학은 통상적인 정수 n번의 미분과 적분을 실수 α번으로 확장한 개념으로서, 다양한 정의들이 있다[7]. 이 중 자주 사용되는 정의는 Riemann-Liouville의 정의와 Caputo의 정의이다. 이 두 정의 모두 주어진 함수 $g(\cdot)$가 구간 $[a, T]$, $T > a$에서 연속이고, $n-1 < \alpha < n, n \in \mathbb{N}$일 때를 가정한다.

Riemann-Liouville의 분수계 미분의 정의는 다음과 같다:

$${}^{RL}_{a}D^{\alpha}_{x}g(x) = \frac{1}{\Gamma(n-\alpha)}\frac{d^n}{dx^n}\int_a^x (x-s)^{n-1-\alpha}g(s)ds. \tag{3}$$

이는 함수 $g(\cdot)$에 대해서 $n-\alpha$번만큼 적분한 다음, n번 만큼 미분하는 것이다. 이때, $\Gamma(\cdot)$는 감마함수이다.

반면 Caputo의 분수계 미분의 정의는 다음과 같다:

$${}^{C}_{a}D^{\alpha}_{x}g(x) = \frac{1}{\Gamma(n-\alpha)}\int_a^x (x-s)^{n-1-\alpha}\frac{d^n g}{ds^n}(s)ds. \tag{4}$$

이는 Riemann-Liouville과 다르게 n번의 미분 이후, 함수를 $n-\alpha$

번만큼 적분한 것을 볼 수 있다. 이런 차이로 인해 Caputo의 정의는 상수함수 $g(x) = K$를 미분했을 때 0이 되지만 Riemann-Liouville은

$$ {}_{a}^{RL}D_{x}^{\alpha}K = \frac{Kx^{-\alpha}}{\Gamma(n-\alpha)} $$

위 식과 같은 결과를 보이게 되어 상수함수를 한번 미분하면 0이 된다는 기존 미적분학과 다른 결과를 낸다. 이러한 차이로 인해 Caputo의 정의가 Riemann-Liouville보다 기존 미적분학의 성질을 잘 따른다고 평가를 받고, 이 때문에 현재 공학과 기계학습 모델에서는 Caputo의 정의가 많이 사용되고 있다[8].

반면, 분수계 적분은 두 가지 정의 모두 동일하게 다음과 같다:

$$ {}_{a}I_{x}^{\alpha}g(x) = \frac{1}{\Gamma(\alpha)}\int_{a}^{x}(x-s)^{\alpha-1}g(s)ds. $$

2.2 Caputo 분수계 미분을 이용한 경사하강법

본 장에선, Caputo의 분수계 미분의 정의를 사용하여 시간변수의 분수계 미분을 적용한 경사하강법 알고리즘을 설계한다. $\alpha \in (0,1)$일때 (2)의 좌변에 Caputo의 분수계 미분을 적용하면

$$ {}_{a}^{C}D_{t}^{\alpha}x(t) = -\frac{df}{dx}(x(t)),\ \alpha \in (0,1). \tag{2*} $$

이 식의 좌변을 분수계 미분의 정의(4)를 사용해 전개하여

$$ \frac{1}{\Gamma(1-\alpha)}\int_{a}^{t}(t-s)^{-\alpha}x'(s)ds = -\frac{df}{dx}(x(t)), $$

이후 해당 식에서 양변을 α만큼 t에 대해 적분하고, 좌변에 [9]의 결과를 적용하면

$$ {}_{a}I_{t}^{\alpha}\frac{1}{\Gamma(1-\alpha)}\int_{a}^{t}(t-s)^{-\alpha}x'(s)ds = -{}_{a}I_{t}^{\alpha}\frac{df}{dx}(x(t)), $$
$$ {}_{a}I_{t}^{\alpha}\,{}_{a}I_{t}^{1-\alpha}x'(t) = {}_{a}I_{t}^{1}x'(t) = -{}_{a}I_{t}^{\alpha}\frac{df}{dx}(x(t)), $$

처럼 좌변을 1계 적분으로 정리할 수 있다. 이를 계산하면,

$$ \int_{a}^{t}x'(s)ds = -\frac{1}{\Gamma(\alpha)}\int_{a}^{t}(t-s)^{\alpha-1}\frac{df}{dx}(x(s))ds, $$
$$ x(t)-x(a) = -\frac{1}{\Gamma(\alpha)}\int_{a}^{t}(t-s)^{\alpha-1}\frac{df}{dx}(x(s))ds. $$

이를 통해 (2*)를 경사하강법 알고리즘의 미분방정식 형태인 (2)와 같이 표현해 보면 다음과 같다:

$$ \frac{d}{dt}x(t) = -\frac{1}{\Gamma(\alpha)}\frac{d}{dt}\int_{a}^{t}(t-s)^{\alpha-1}\frac{df}{dx}(x(s))ds. \tag{5} $$

이렇게 구한 $\frac{d}{dt}x(t)$을 이산화하여 (1)처럼 반복 계산하면, 좌변에 분수계 미분이 적용된 경사하강법 알고리즘이 구성된다.

3. 최적화 불가능성에 대한 수학적 검증

이번 장은 (5)에 목적함수 $f(x) = x$를 적용해 해당 함수에서 경사하강법이 최적화되지 않음을 수학적으로 증명하고자 한다. 해당 목적함수는 실수 전체에서 정의된 함수이고 a가 바뀐다고 해도 동일한 결과를 보일 수 있으니, 계산의 편의상 $a = 0$로 설정하겠다. 이를 전부 반영하여 (5)을 전개하면:

$$ \begin{aligned} \frac{d}{dt}x(t) &= -\frac{1}{\Gamma(\alpha)}\frac{d}{dt}\int_{a}^{t}(t-s)^{\alpha-1}\frac{df}{dx}(x(s))ds \\ &= -\frac{1}{\Gamma(\alpha)}\frac{d}{dt}\int_{a}^{t}(t-s)^{\alpha-1}ds \\ &= -\frac{1}{\Gamma(\alpha)}\frac{d}{dt}\frac{1}{\alpha}t^{\alpha}. \end{aligned} $$

이를 통해 (2*)을 해당 목적함수에서 (2)와 같이 표현하면 다음과 같이 표현된다.

$$ \frac{d}{dt}x(t) = -\frac{1}{\Gamma(\alpha)}t^{\alpha-1}. \tag{6} $$

이때, a의 값이 변경된다면 t대신에 $t-a$가 들어가게 되고 이는 (6)의 결과를 t에 대해 평행 이동한 것과 같다.

보편적인 경사하강법은 주어진 $\epsilon > 0$에 대하여,

$$ \left|\frac{d}{dt}x(t)\right| \le \epsilon \tag{7} $$

를 만족하면, '매개변수의 변화가 없다,' 즉, '목적함수의 극솟값에 도달했다'고 판단하여 위 식을 만족하는 x 에서 훈련을 멈추게 된다. 목적함수 $f(x) = x$의 경우, 경사하강법이 함수의 최솟값으로 최적화되려면 $x \to -\infty$로 발산해야하기 때문에 (7)의 조건을 만족하지 않아야 한다. 하지만 $f(x) = x$의 경우 종료 조건을 만족하는 분수계 경사하강법의 t값은,

$$ \left|-\frac{1}{\Gamma(\alpha)}t^{\alpha-1}\right| \le \epsilon $$

이므로, 정리하여서

$$ t \ge \left(\frac{1}{\Gamma(\alpha)\epsilon}\right)^{1/(1-\alpha)} $$

을 만족하는 t에서 멈추게 된다.

그림 1은 범위 내 α에서 $x'(t)$를 시각화한 그림들이다. 그래프의 점선은 $\epsilon = 10^{-2}$을 나타내고, 그림의 빨간 점 이후로 (7)을 만족하여 경사하강법이 멈추게 된다. 그림 2는 실제 $\alpha = 0.3$에서 (2*)를 이산화하였을 때 매개변수 x가 어떻게 이동하는지 시각화한 것이다. 각 빨간 점들은 이전 x에서 30번의 경사하강을 거친 다음의 x이다. 모든 점에서 $f(\cdot)$의 미분계수가 1로 같음에도 불구하고 빨간 점들의 간격이 점점 좁아짐을 볼 수 있다. 하지만 $f(x) = x$는 최솟값이 $-\infty$로 발산하고 극솟값 또한 존재하지 않

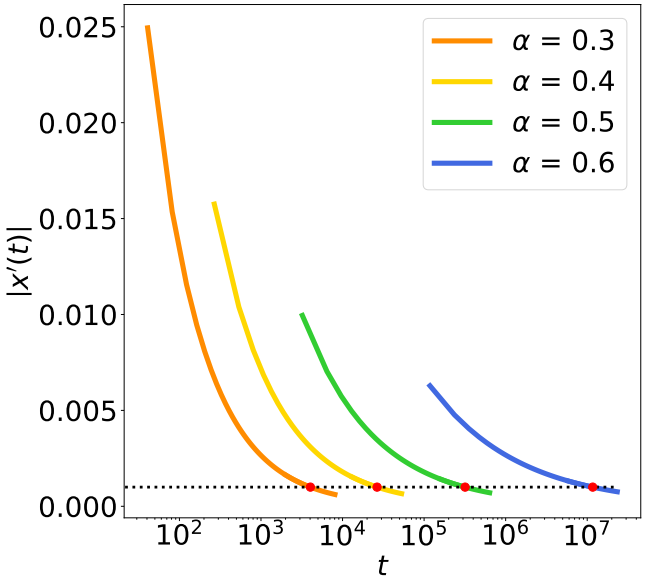

그림 1: 다양한 α에서의 $x'(t)$

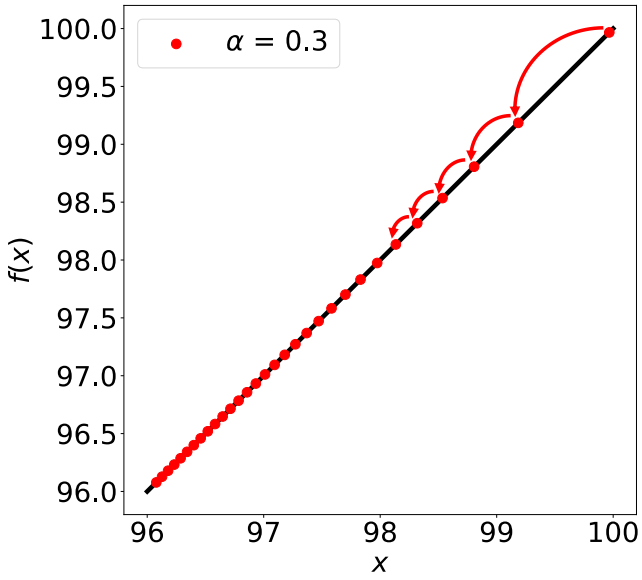

그림 2: $\alpha = 0.3$에서의 경사하강이 느려지는 예시

기에 경사하강법이 멈춘 곳은 목적함수의 최솟값이 될 수 없다.

비록, $x(t)$는 수학적으로 다음과 같이 표현되어,

$$x(t) = x(0) - \frac{1}{\alpha\Gamma(\alpha)}t^{\alpha}$$

$t \to \infty$에 따라 발산하지만, 이산화를 통한 알고리즘에선 그림 1에서 볼 수 있듯이 주어진 종료 조건에서 t가 멈추게 되므로 최적화될 수 없다. 이를 통해 시계열에 분수계 미적분을 적용한 경사하강법에선 단순한 볼록함수와 보편적인 종료조건에서 최적화될 수 없음을 보였다.

4. 결 론

본 논문은 시계열에 대한 분수계 경사하강법을 적용하였을 때, 이를 이산화하여 알고리즘화하는 방법을 제시하였고 이를 간단한 볼록함수 $f(x) = x$를 통해 수식화하였다. 이와 함께 $\alpha \in (0, 1)$계 미분에선 해당 경사하강법이 보편적인 종료조건에서 최솟값으로 최적화되지 않는다는 것을 시각화하였다.

이러한 경사하강법의 최적화 불가능성은 매개변수가 다양해질수록, 목적함수가 복잡해질수록 더 커질 것이다. 이를 통해 기존의 경사하강법에 분수계 미적분을 도입한 모델들을 사용하면서 최적화가 보장되는 조건을 찾는 수학적인 분석이 필요할 것임을 시사한다.

참고 문헌

[1] S. Sun, Z. Cao, H. Zhu, and J. Zhao, "A survey of optimization methods from a machine learning perspective," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3668–3681, 2020.

[2] O. Herrera-Alcántara, "Fractional derivative gradient-based optimizers for neural networks and human activity recognition," *Applied Sciences*, vol. 12, no. 18, 2022.

[3] D. P. Hapsari, I. Utoyo, and S. W. Purnami, "Fractional gradient descent optimizer for linear classifier support vector machine," in *2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE)*, pp. 1–5, 2020.

[4] J. Wang, Y. Wen, Y. Gou, Z. Ye, and H. Chen, "Fractional-order gradient descent learning of bp neural networks with caputo derivative," *Neural Networks*, vol. 89, pp. 19–30, 2017.

[5] J. Xie and S. Li, "Training neural networks by time-fractional gradient descent," *Axioms*, vol. 11, no. 10, 2022.

[6] M. Caputo, "Linear Models of Dissipation whose Q is almost Frequency Independent—II," *Geophysical Journal International*, vol. 13, pp. 529–539, 11 1967.

[7] M. Candan and M. Çubukçu, "Implementation of caputo type fractional derivative chain rule on back propagation algorithm," *Applied Soft Computing*, vol. 155, p. 111475, 2024.

[8] C. Bao, Y.-F. PU, and Y. Zhang, "Fractional-order deep backpropagation neural network," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–10, 07 2018.

[9] R. Almeida, "A caputo fractional derivative of a function with respect to another function," *Communications in Nonlinear Science and Numerical Simulation*, vol. 44, p. 460–481, Mar. 2017.