고전 역학의 라그랑지안을 이용한 미분 기하학적 global minimum 탐색 알고리즘

김준식[0,1,2], 오장민[2], 김종찬[1], 장병탁[2]
[1]서울대학교 물리학부
[2] 서울대학교 컴퓨터 공학과 바이오 지능 연구실
{shick, jckim}@phya.snu.ac.kr, {jmoh, btzhang}@bi.snu.ac.kr

A Novel Global Minimum Search Algorithm based on the Geodesic of Classical Dynamics Lagrangian

Joon Shik Kim, Jangmin O, Jong Chan Kim, Byoung-Tak Zhang
School of Physics, Seoul National University
Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University

요 약

뉴럴네트워크에서 학습은 에러를 줄이는 방법으로 구현 된다. 이 때 parameter 공간에서 Risk function 은 multi-minima potential 로 표현 될 수 있으며 우리의 목적은 global minimum weight 좌표를 얻는 것이다. 이전의 연구로는 Attouch et al. 의 damped oscillator 방정식을 이용한 방법이 있고, Qian 의 critically damped oscillator 를 통한 steepest descent 의 momentum 과 learning parameter 유도가 있다. 우리는 이 두 연구를 참고로 manifold 상에서 최단 경로인 geodesic 을 Newton 역학의 Lagrangian 에 적용함으로써 adaptive steepest descent 학습법을 얻었다. 우리는 이 새로운 방법을 Rosenbrock 과 Griewank 포텐셜들에 적용하여 그 성능을 알아 본다.

1. 서 론

어떤 양을 여러 경로로 적분해 본 후 그 경로 적분을 최소로 하는 경로를 따라 빛이나 물체가 움직인다는 것은 물리학에서 잘 알려진 사실이다. 일 예로서 Newton 역학은 Lagrange 역학으로 새롭게 formulation 될 수 있으며 이때 예로서 있으며 이 때의 새로운 양식은 Hamilton' s principle [1] 이라는 일종의 변분법으로 표현된다. Hamilton' s principle 이란 물체의 운동에너지에서 위치 에너지를 뺀 양을 Lagrangian 으로 정의한 뒤 이를 시간으로 적분한 Action 을 최소화 시키는 경로가 바로 Newton 역학에서 구한 물체의 경로와 같다는 것이다.

우리는 이러한 개념을 뉴럴 네트워크 학습에 적용하고자 한다. 결국 affine parameter 라는 시간 좌표에서 크기가 상수 1인 새로운 Lagrangian 을 정의하고 이에 변분법(calculus of variation)을 적용하여 [2] 2차의 adaptively damped oscillator 방정식을 얻었다.

이와 비슷한 기존의 연구로는 Qian [3] 의 critical damped oscillator를 이용하여 momentum 과 learning rate 를 유도하는 연구가 있다. 또한 Attouch et al. [4] 의 damped oscillator를 이용하여 global minimum 탐색을 하는 연구가 있다.

우리는 변분법으로 새롭게 얻은 2차의 adaptively damped oscillator 방정식을 Qian [3] 의 방법으로 discretize 시켰다. 그리고 이렇게 얻은 1차의 adaptive steepest descent 규칙을 Attouch et al. [4] 의 example 들인 Rosenbrock 과 Griewank potential 함수들에 대해 global minimum 탐색을 수행하여 보았다.

우리는 다양한 시작점에서 학습을 수행하였으며 대부분의 경우 한번에 global minimum을 찾았다.

이 새로운 adaptive steepest descent 규칙은 global minimum 탐색과 연관된 뉴럴 네트워크, 경제학, 그리고 게임 이론에 적용될 수 있을 것으로 기대된다.

2. 2차의 adaptively damped oscillator 방정식의 유도

먼저 고전 역학의 Lagrangian L_C 는 다음과 같이 정의된다.

$$L_c = \frac{m}{2}\sum_i \left(\frac{dw_i}{dt^2}\right)^2 - V \qquad (1)$$

여기서 m은 물체의 질량, w_i는 좌표, 그리고 V 는 potential 함수이다.

우리는 affine parameter σ 상에서 크기가 상수 1인 새로운 Lagrangian L_N을 다음과 같이 정의 한다.

$$ds^2 = L_N d\sigma^2 = L_C dt^2, \qquad (2)$$

여기서 ds 는 manifold 상에서의 미세거리이다.

식 (1) 과 (2)로부터 L_N을 구한후 아래의 Euler-Lagrangian 방정식에 대입한다.

$$\frac{dL_N}{dw_i} - \frac{d}{d\sigma}\frac{dL_N}{d\dot{w}_i} = 0, \qquad (3)$$

위에서 w_0는 시간 t, 그리고 $w_j (j \neq 0)$ 은 weight space coordinate 이다. 또한 w_i 위의 dot(.) 은 affine parameter σ 에 대한 미분을 의미한다.

위의 식 (3) 에서 얻은 미분 방정식을 연립하여 풀면 아래와 같은 2차의 adaptively damped oscillator 방정식을 얻는다.

$$\frac{d^2 w_i}{dt^2} = \frac{d}{dt}(\ln V)\frac{dw_i}{dt} - \frac{1}{m}\frac{dV}{dw_i} \qquad (4)$$

3. Adaptively damped oscillator 의 수렴성

Section 2 에서 유도한 식 (4) 를 얻는 과정에서 아래와 같은 affine parameter 에서의 dynamics 방정식을 거친다.

$$\frac{d^2 w_i}{d\sigma^2} = -\eta \frac{d}{dw_i}(-\frac{1}{V}), \qquad (5)$$

여기서 η 는 learning rate로 $1/m$ 에 해당한다.

이제 potential V가 global minimum 에서 V=0 의 값을 가진다고 가정하자. 그러면 식 (5) 에서 learning mass 는 affine parameter σ 시간에서 -∞ 의 singular 한 potential 값을 가지는 global minimum point로 무한대의 속력을 가지고 충돌한다.

이를 ordinary time t 에서 바라다보면 식 (4) 에서 알 수 있듯이 adaptive 한 damping 을 받아 global minimum point로 수렴하게 된다.

이로서 우리는 식 (4) 의 2차 adaptively damped oscillator 방정식의 수렴성을 affine parameter 에서의 singular crashing behavior 로부터 알 수 있다.

4. 1차의 adaptive steepest descent 규칙 유도

식 (4) 의 2차 미분방정식을 Qian[3] 의 논문처럼 다음과 같이 1차 update 식으로 바꿀 수 있다.

$$\frac{w_i(t+\Delta t) + w_i(t-\Delta t) - 2w_i(t)}{\Delta t^2} = \frac{\ln V(t) - \ln V(t-\Delta t)}{\Delta t}\frac{w_i(t+\Delta t) - w_i(t)}{\Delta t} - \frac{1}{m}\frac{dV}{dw_i}, \qquad (6)$$

$$w_i(t+\Delta t) - w_i(t) = \frac{1}{1-(\ln V(t) - \ln V(t-\Delta t))}(w(t) - w_i(t-\Delta t)) - \frac{(\Delta t)^2}{m(1-(\ln V(t) - \ln V(t-\Delta t))}\frac{dV}{dw_i} \qquad (7)$$

5. Adaptive steepest descent 규칙의 global minimum 탐색 문제에의 적용

우선 우리는 식 (7) 에서 얻은 1차의 update 식을 아래와 같이 주어지는 Rogenbrock potential 에 적용하여 보았다

$$V(x,y) = 100(y-x^2)^2 + (1-x)^2, \qquad (8)$$

결과는 그림 1과 같으며 global minimum point (1,1) 로 잘 수렴함을 알 수 있다.

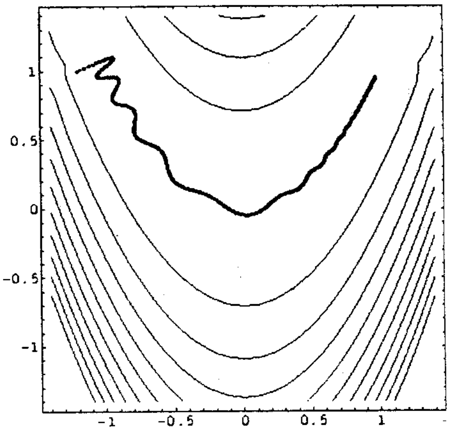

그림 1. Rosenbrock potential 에서의 global minimum 탐색. m=10000, Δt=1, 그리고 global minimum point 는 (1,1) 이다. 시작점은 (-1.2,1) 이다.

둘째로, 다음과 같이 주어지는 Griewank potential 에 adaptive steepest descent 규칙을 적용해 보았다.

$$V(x,y) = (2x^2 + y^2 - xy)/50 - \cos(x)\cos(y/\sqrt{2}) + 1 \qquad (9)$$

그림 2와 그림 3은 각각 처음 두 초기값이 (10,-10), (7,7) 일 때의 탐색양상이다. 그림 3에서는 learning mass 가 두개의 local minima 를 빠져 나오는 것을 볼 수 있다. Attouch et al. 의 경우 여러 번의 다시 시작함으로 local minima 들을 돌아 다니는 결과를 보였으며 우리의 결과는 한번에 global minimum 을 찾는다는 데서 그 장점을 보인다.

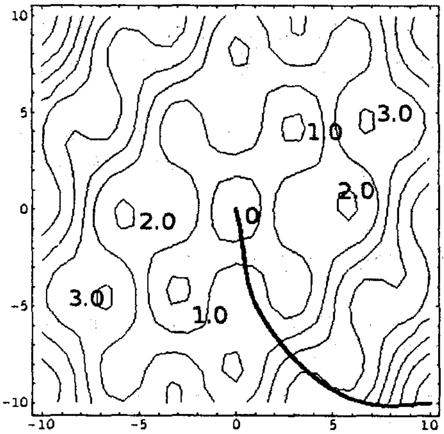

그림 2. Griewank potential 에서의 global minimum 탐색. m=1000, Δt=1, 그리고 global minimum point 는 (0,0) 이다. 두 초기 좌표는 (10,-10) 이다.

지금까지의 두 가지 potential 함수들은 Attouch et al. 의 논문에 나오는 예제들이며 우리의 학습 규칙이 한번에 global minimum 을 찾는 좋은 수행 결과를 보인다. 참고로 Attouch et al. 의 결과는 여러 번의 새로운 shooting 을 통하여 결국 global minimum 을 찾는다.

여기서 우리가 주의할 점은 global minimum 의 함수값이 0 임을 잊지 않아야 한다는 것이다. 만일 함수 값이 0 이 아니라면 Section 3 의 수렴성이 보장되지 않는다.

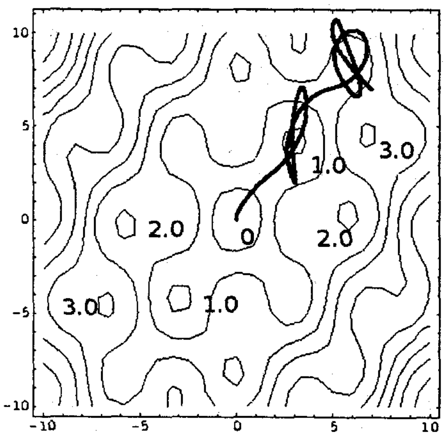

그림 3. Griewank potential 에서의 global minimum 탐색. m=1000, Δt=1, 그리고 global minimum point 는 (0,0) 이다. 두 초기 좌표는 (7,7) 이다

6. 결론

우리는 앞에서 새로운 adaptive steepest descent 규칙을 Newton 역학의 geodesic으로부터 유도해 내었다. 이를 Rosenbrock 과 Griewank potential 들에 적용하였을 때 우수한 성능을 보임을 확인할 수 있었다.

이 연구의 의미는 변분법(calculus of variation) 이라는 최소적분경로를 구하는 원리를 사용하여 최소 함수값을 가지는 parameter 좌표를 구하는 데서 찾을 수 있다. 즉 학습이라는 인지적 기능을 물리 법칙과 수학 원리를 사용하여 구현해 내었다는데 이 연구의 중요성이 있다.

감사의 글

박성찬, 계범석, 이지우 그리고 노영균 님들의 친절한 논의에 감사 드린다. 이 연구는 교육 인적 자원부의 BK21 program 과 과학기술부의 NRL program 의 지원을 받았다.

참고 문헌

1. Marion, J.B. and Thornton, S.T., *Classical dynamics of particles and systems*, Saunders College Pub., FortWorth, 1995.
2. Martin, J.L., *General Relativity: A First Course for Physicists*, Prentice Hall Europe, Hertfordshire, 1995.
3. Qian, N., On the momentum term in gradient descent learning algorithms, *Neural Networks*, 12, 145-151 (1999).
4. Attouch, H., Goudou, X. and Redont, P., The heavy ball with friction method I. The continuous dynamical system: Global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamics system, *communications in contemporary mathematics*, 2, 1-34 (2000).
5. Rumelhart, D.E., Hinton, G.E. and Williams, R.J.: Learning representations by back-propagating errors, *Nature*, 323, 533-536 (1986)
6. Cabot, A., Inertia gradient-like dynamics system controlled by a stabilizing term, *Journal of optimization theory and applications*, 120, 275-303 (2004).
7. Edelman, A., Arias, T.A. and Smith, S.T., The geometry of algorithms with orthogonality constraints, *Siam J. Matrix Anal. Appl*, 20, 303-353 (1998).
8. Nishimori, Y. and Akaho, S., Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold, *Neurocomputing*, 67, 106-135 (2005).
9. Amari, S.,Natural gradient works efficiently in learning, *Neural Computation*, 10, 251-276 (1998).
10. Caiani, L., Casetti, L., Clementi, C. and Pettini, M., Geometry of Dynamics, Lyapunov exponents, and phase transition, *Physical Review Letters*, 79, 4361-4364 (1997).
11. Torii, M. and Hagan, M.T., Stability of steepest descent with momentum for quadratic functions, *IEEE Transactions on Neural Networks*, 13, 752-756 (2002).