**옵티마이저(optimizer)**

The Optimizer base class provides methods to compute gradients for a loss and apply gradients to variables.

A collection of subclasses implement classic optimization algorithms such as GradientDescent and Adagrad.

옵티마이저란 모델에 대하여 학습을 진행함에 있어서 최적화된 결과를 얻기 위하여 수행되는 객체입니다.

즉, 최적화된 w와 b의 값을 구하기 위하여 사용하는 알고리즘으로 이해하면 됩니다.

경사 하강법으로 부족해서 많은 옵티마이저를 만들어서 사용하고 있습니다.

|  |
| --- |
| 구현할 내용 |
| https://keras.io/optimizers/  http://www.denizyuret.com/2015/03/alec-radfords-animations-for.html |

### SGD(Stochastic gradient descent) 옵티마이저

확률적 경사 하강법(SGD)은 특정 지점에서 하강을 수행할 때 가장 크게 기울어진 방향(경사도가 가장 급격한)으로 일정한

거리만큼 이동하려는 optimizer입니다.

학습률이 크면 많이 이동하고, 탐색 경로가 zigzag(비효율적으로 이동)로 이동합니다.

SGD의 단점은 비등방성(anisotropy), 즉 한 방향으로 쭉 나아 가려는 성질을 가지고 있습니다.

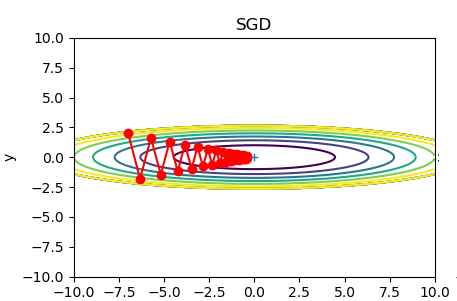
SGD의 수식은 다음과 같이 작성 가능합니다.

w는 갱신될 가중치 변수를, α는 학습률을 의미합니다.

는 w에 대한 손실 함수(loss function)의 기울기를 의미합니다.

학습률은 0.1 혹은 0.001 등의 초기 값을 미리 설정하여 사용합니다.

공식에서 보다시피 기울어진 방향으로 일정한 거리만큼 이동하겠다는 단순한 방법입니다.



### Momemtum 옵티마이저

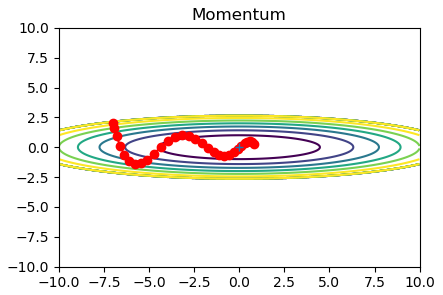
모멘텀(Momemtum)이라는 단어는 물리학에서 운동량이라고 하며, 관성/탄력/가속도라는 의미로 이해해도 좋습니다.

탄력을 받아서 기울기 방향으로 힘을 받아서 물체가 가속되는 특징이 있습니다.

관성을 유지하면 쭉 아래로 이동하며, 공이 그릇의 곡면을 따라서 굴러 가듯이 움직입니다.

학습률 이외에 **별도의 모멘텀 항을 이용**하여 "처음에는 크게 그리고 점점 작게"라는 개념을 포함할 수 있습니다.

SGD에 비하여 zigzag가 조금 덜합니다.



모멘텀의 수식은 다음과 같이 작성 가능합니다.

μ는 공기 저항이나 마찰 계수 정도로 이해하면 되고, v는 속도(velocity)입니다.

μ는 0.9 등의 값으로 설정합니다.

기울기 방향으로 가속도의 힘을 받아서 물체가 가속이 된다는 물리 법칙을 적용합니다.뭋

### AdaGrad 옵티마이저

일반적으로 신경망에서는 학습률의 값이 너무 적으면 학습 시간이 너무 길어지고, 너무 크게 되면 발산하여 올바른 학습을 수행할 수 없는 경우가 많습니다.

"국소적 최적해"에 빠지지 않고, 효율적인 정답을 구하기 위하여 학습률을 '처음에는 크게 진행하다가 그리고, 점차 작게'하는 방법이 바람직합니다.

AdaGrad 옵티마이저는 **학습을 진행해 나가면서 학습률을 조금씩 줄여 나가면서 진행하는 방식**입니다.

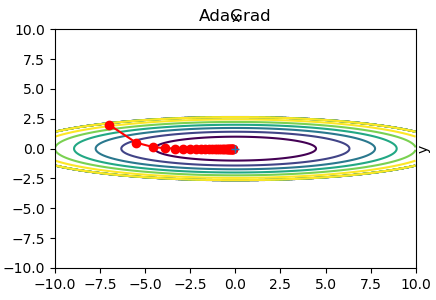
과거의 기울기를 제곱하여 더해 나가는 방식입니다.

진행할수록 갱신 강도(strength)는 약해집니다.

무한 학습을 하게 되면 어느 순간 갱신되는 량이 0이 되는 문제(경사 소실 문제)가 있습니다.

수식에서 h는 기존 w에 대한 손실 함수의 기울기를 제곱하여 더해주는 역할을 수행합니다.

⊙ 기호는 요소별 곱셈을 의미합니다.



### RMSProp 옵티마이저

AdaGrad는 무한 학습시 갱신되는 량이 0이 되는 문제가 있었습니다.

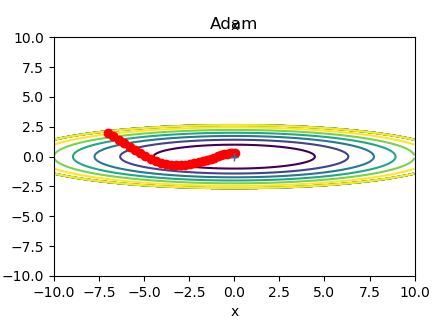
이러한 문제에 대하여 보폭에 대한 민감도를 조금 보완한 방식이 RMSProp(Root Mean Square Propagation)입니다.

먼 과거의 기울기는 고려하지 않고, 최근의 새로운 기울기 정보만을 반영하는 방식입니다.

이것을 지수 이동 평균(Exponential Moving Average) 방식이라고 합니다.

### Adam 옵티마이저

Adam 방식은 Momentum 방식과 RMSProp 방식의 장점들을 모아서 보완한 방식입니다.(2015년 제안)



### 어떤 옵티마이저를 사용할 것인가?

어떠한 옵티마이저를 사용할 것인가 ?

애석하게도 모든 문제에서 항상 뛰어난 기법이란 것은 없다.

학습률, 함수의 수식 등 여러 개의 요소가 복합적입니다.

