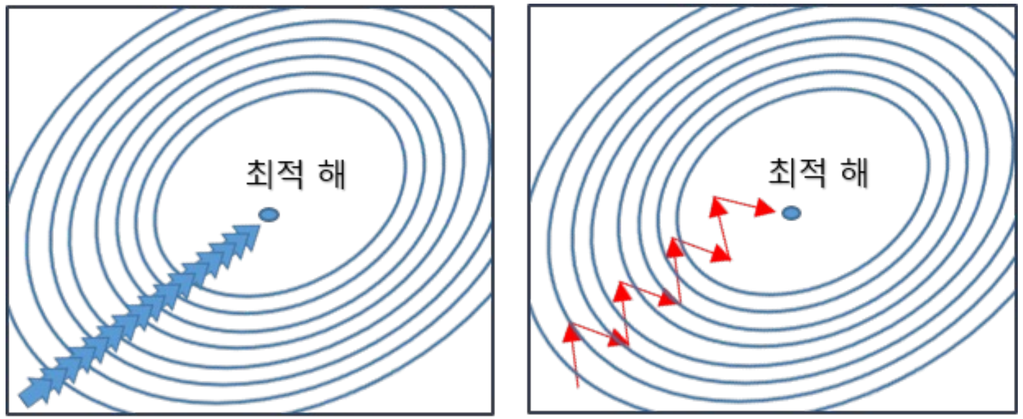
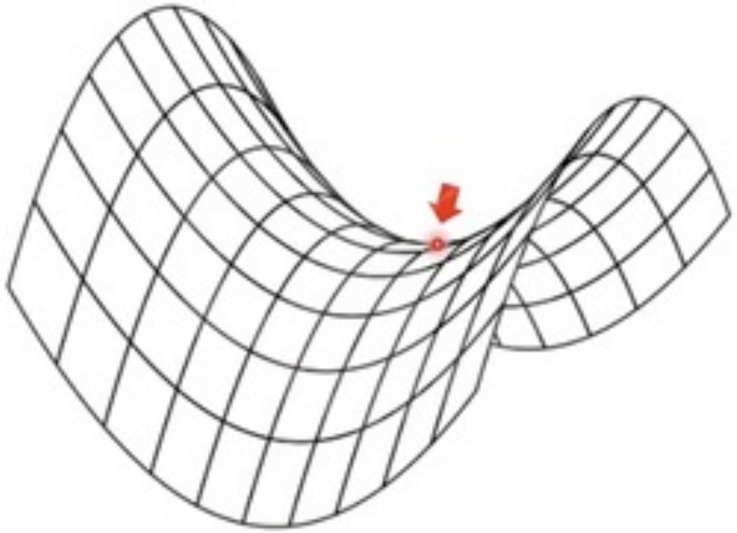
04. 최적화

딥러닝 스터디 기초 1반
최민아

확률적 경사 하강법 (SGD)

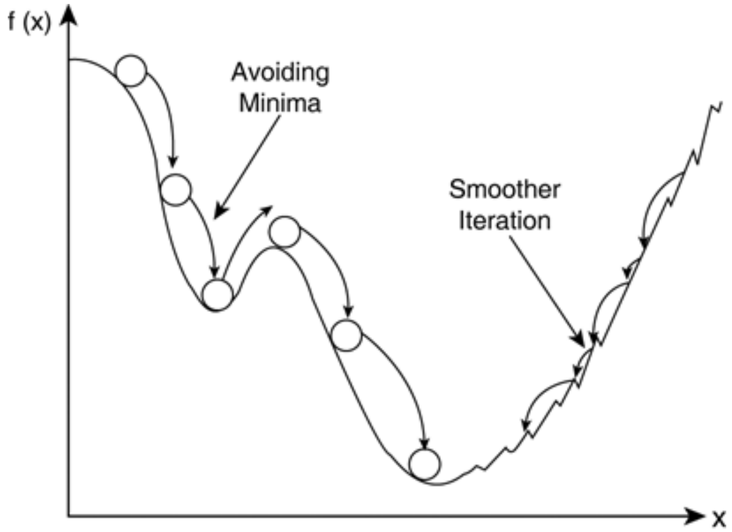


경사 하강법 확률적 경사 하강법



- 고정 학습률
- 협곡과 안장점
- 진동

SGD 모멘텀



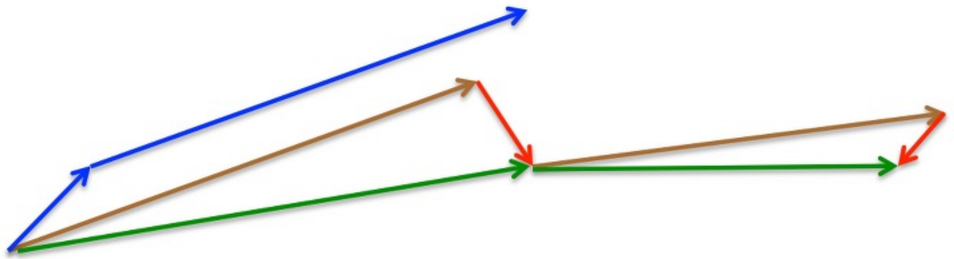
현재 속도 현재 그레이디언트

$$v_{t+1} = \rho v_t + \nabla f(x_t)$$

$$x_{t+1} = x_t - \alpha v_{t+1}$$

- 가장 가파른 곳으로 내려가는 과정에서 지금까지 진행하는 속도에 관성을 주는 방식
- 관성으로 임계점 탈출과 빠른 학습
- 현재의 속도 벡터 + 그레이디언트 벡터 가속도로 인한 오버슈팅

네스테로프 모멘텀



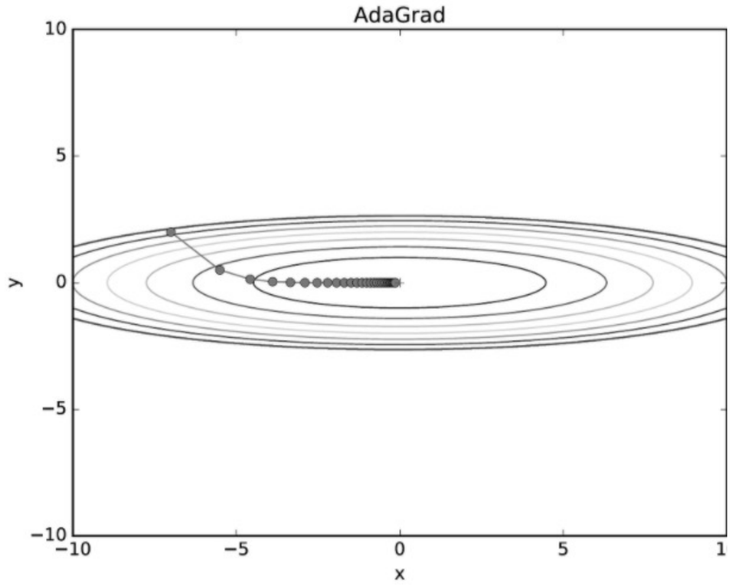
$$v_{t+1} = \rho v_t - \alpha \nabla f(x_t + \rho v_t)$$
$$x_{t+1} = x_t + v_{t+1}$$

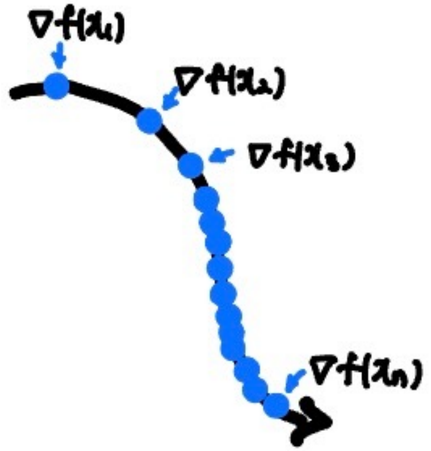
Change of variables $\tilde{x}_t = x_t + \rho v_t$ and rearrange:

$$v_{t+1} = \rho v_t - \alpha \nabla f(\tilde{x}_t)$$
$$\tilde{x}_{t+1} = \tilde{x}_t - \rho v_t + (1+\rho) v_{t+1}$$
$$= \tilde{x}_t + v_{t+1} + \rho(v_{t+1} - v_t)$$

- 관성을 이용해 현재 속도로 한 걸음 미리 간 지점에서 내리막길로 내려가는 방식
- 속도 벡터 + 현재 속도로 미리 가 본 위치의 그레이디언트 벡터
- 오버슈팅 억제

AdaGrad





기울기 벡터 = $(\nabla f(x_1), \nabla f(x_2), \cdots, \nabla f(x_n))$

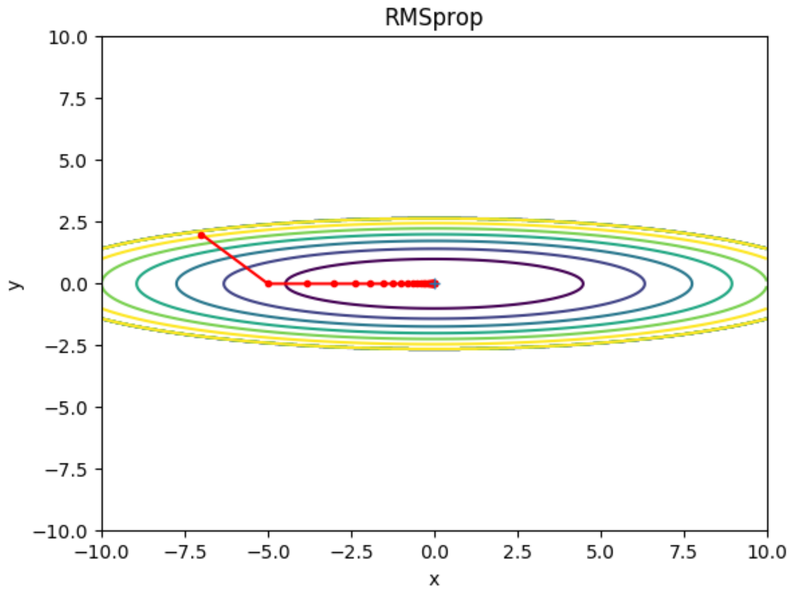
곡면의 변화량 = $r_{t+1} = \nabla f(x_1)^2 + \nabla f(x_2)^2 + \cdots + \nabla f(x_t)^2$

$$r_{t+1} = r_t + \nabla f(x_t)^2$$

$$x_{t+1} = x_t - \frac{\alpha}{\sqrt{r_{t+1} + \varepsilon}} \odot \nabla f(x_t)$$

- 곡면의 변화에 따라 적응적으로 학습률을 정하는 방식
- 전체 경로의 변화량 측정, 적응적 학습률
- 적응적 학습률로 인한 조기 중단

RMSProp



그레이디언트 제곱의 지수가중이동평균

$$r_{t+1} = \beta r_t + (1-\beta)\nabla f(x_t)^2$$

$$x_{t+1} = x_t - \frac{\alpha}{\sqrt{r_{t+1}+\epsilon}} \odot \nabla f(x_t)$$

적응적 학습률

$$r_t = \beta r_{t-1} + (1-\beta)\nabla f(x_{t-1})^2$$

$$r_{t+1} = \beta\left(\beta r_{t-1} + (1-\beta)\nabla f(x_{t-1})^2\right) + (1-\beta)\nabla f(x_t)^2$$

$$= \beta^2 r_{t-1} + \beta(1-\beta)\nabla f(x_{t-1})^2 + (1-\beta)\nabla f(x_t)^2$$

$$= \beta^2 r_{t-1} + (1-\beta)\left(\nabla f(x_t)^2 + \beta\nabla f(x_{t-1})^2\right)$$

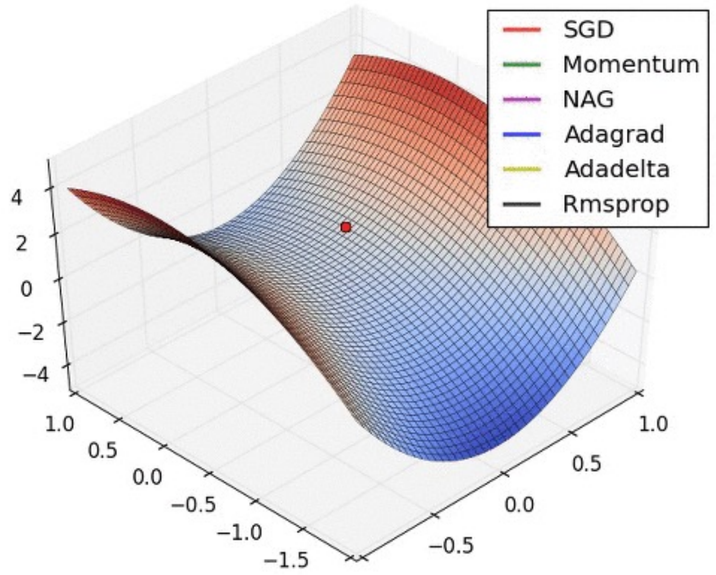
$$\vdots$$

$$= \beta^t r_1 + (1-\beta)\left(\nabla f(x_t)^2 + \beta\nabla f(x_{t-1})^2 + \cdots + \beta^{t-1}\nabla f(x_1)^2\right)$$

최근경로 반영↑ / 오래된 경로 반영↓

- 곡면 변화량에 따라 학습률을 적용적으로 결정하는 방식
- 최근 경로의 변화량 측정, 곡면 변화량이 계속 증가하는 현상 방지
- 지수가중이동평균

Adam



$$V_{t+1} = \beta_1 V_t + (1-\beta_1)\nabla f(x_t)$$ (1차 관성)

$$r_{t+1} = \beta_2 r_t + (1-\beta_2)\nabla f(x_t)^2$$ 2차 관성

$$x_{t+1} = x_t - \frac{\alpha}{\sqrt{r_{t+1}+\varepsilon}} \odot V_{t+1}$$

V_{t+1}: 관성을 강한 속도

최근 경로의 곡면의 변화량

$$V_{t+1} = \frac{V_{t+1}}{(1-\beta_1^t)}$$

$$r_{t+1} = \frac{r_{t+1}}{(1-\beta_2^t)}$$

- SGD 모멘텀 + RMSProp
- 진행하던 속도에 관성을 주고 동시에 최근 경로의 곡면 변화량에 따라 적응적 학습률
- 초기 경로 편향 문제