Second-order Taylor expansion

박주찬

DICE Lab
School of Computer Science and Engineering
KOREATECH
green261535@gmail.com

Introduction

First-order optimization에서는 loss 그래프에서 한점에 대한 선형 직선을 **Fig. 1.** 와 같이 그릴 수 있다.

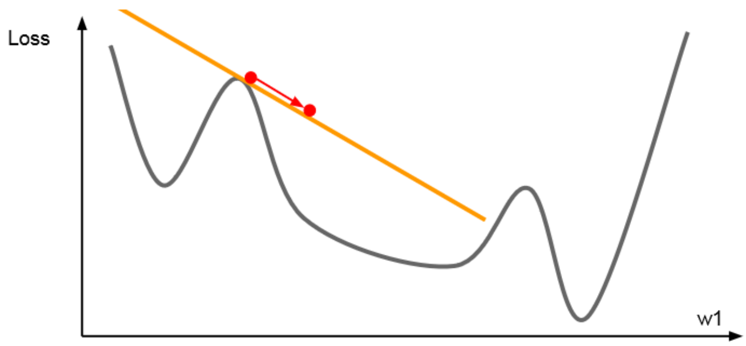

Fig. 1.

Second-order optimization에서는 loss 그래프에서 한점에 대한 2차 그래프를 **Fig. 2.** 와 같이 그린다. 이것을 토대로 loss가 낮아지는 지점으로 weight를 업데이트 한다. 보통 weight을 업데이트 할 때에는 1차원 선형 그래프를 가지고 미분값을 구해 업데이트를 하지만 2차 그래프를 그렸을 때는 어떻게 gradient update 식이 나오는지 알아보려고 한다.

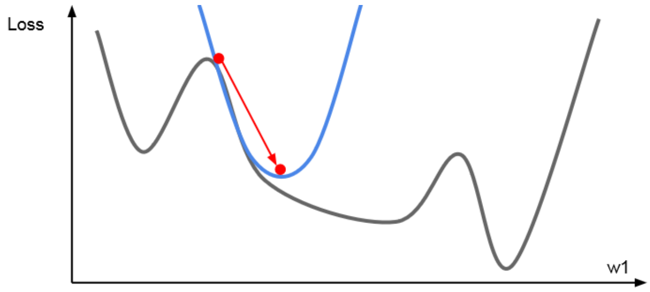

Fig. 2.

2차 그래프의 식을 Taylor expansion에 의하면 (1)번 식과 같고,

$$J(\theta) \approx J(\theta_0) + (\theta - \theta_0)^T \times \nabla_\theta J(\theta_0) + \frac{1}{2} \times (\theta - \theta_0)^T \times H \times (\theta - \theta_0) \tag{1}$$

gradient descent를 하기 위한 θ 업데이트 식은 (2)번 식과 같다.

$$\theta^* = \theta_0 - H^{-1} \times \nabla_\theta J(\theta_0) \tag{2}$$

J함수는 loss function이고, θ는 J함수의 매개변수 즉 파라미터이다. H는 Hessian matrix를 의미한다.
먼저 (1)번식이 왜 이렇게 정의할 수 있는지 증명해보고, weight를 업데이트 하는 식(2)에 대한 증명을 살펴보겠다.

(1)번에 대한 증명

$J(\theta)$함수가 있다고 하자.
$J(\theta)$함수에서 한 점 $(a, J(a))$가 있을 때, 이 점에서의 기울기는 $J(a)'$이며 직선의 방정식은 $J(\theta) - J(a) = J(a)' \times (\theta - a)$이다.

$$J(\theta) = J(a) + J(a)' \times (\theta - a) \tag{3}$$

직선의 방정식은 1차항으로 만들었다면 2차원 방정식은 2차항을 추가함으로써 수식을 만들 수 있다. 2차항에 대한 수식을 만들면 다음과 같다.

$$J(\theta) \approx J(a) + J(a)' \times (\theta - a) + \frac{1}{2} \times J(a)'' \times (\theta - a)^2 \tag{4}$$

2차원이 아닌 고차원 n차원이라고 하면 일반화 식은 다음과 같다.

$$J(\theta) \approx \sum_{i=0}^{n} \frac{J'(a)}{i!} \times (\theta - a)^i \tag{5}$$

위 식들은 θ와 a가 one variable(scalar)였다.
이제 θ와 a가 one variable(scalar)가 아닌 multi variable(vector)에 대해서 식이 어떻게 바뀌는지 증명해 보겠다.
$\theta = \begin{bmatrix} x \\ y \end{bmatrix}, \theta_0 = \begin{bmatrix} a \\ b \end{bmatrix}$ 라고 하자. (6)

그러면(4)번 식을 다음과 같이 풀어서 표현 할 수 있다.

$$\begin{aligned} J(x,y) \approx\ & J(a,b) + \nabla J_x(a,b) \times (x-a) + \nabla J_y(a,b) \times (y-b) \\ & + \frac{1}{2} \times \{\nabla J_{xx}(a,b) \times (x-a)^2 \\ & + 2 \times \nabla J_{xy}(a,b) \times (x-a) \times (y-b) \\ & + \nabla J_{yy}(a,b) \times (y-b)^2 \end{aligned} \tag{7}$$

먼저 (7)번 식 중에 $\nabla J_x(a,b) \times (x-a) + \nabla J_y(a,b) \times (y-b)$ 를 간단히 하면 다음과 같다.

$$\begin{aligned} [x-a \quad y-b] \times \begin{bmatrix} \nabla J_x(a,b) \\ \nabla J_y(a,b) \end{bmatrix} &= [x-a \quad y-b] \times \begin{bmatrix} \nabla J_x(\theta_0) \\ \nabla J_y(\theta_0) \end{bmatrix} \\ &= (\theta - \theta_0)^T \times \nabla J_\theta(\theta_0) \end{aligned} \tag{8}$$

(8)번 식 중에서 $\frac{1}{2} \times \{\nabla J_{xx}(a,b) \times (x-a)^2 + 2 \times \nabla J_{xy}(a,b) \times (x-a) \times (y-b) + \nabla J_{yy}(a,b) \times (y-b)^2$ 를 간단히 하면 다음과 같다.

$$\begin{aligned} & \frac{1}{2} \times \{\nabla J_{xx}(a,b) \times (x-a)^2 + 2 \times \nabla J_{xy}(a,b) \times (x-a) \times (y-b) \\ & \qquad + \nabla J_{yy}(a,b) \times (y-b)^2 \\ & = \frac{1}{2} \times [x-a \quad y-b] \times \begin{bmatrix} \nabla J_{xx}(a,b) & \nabla J_{xy}(a,b) \\ \nabla J_{yx}(a,b) & \nabla J_{yy}(a,b) \end{bmatrix} \times \begin{bmatrix} x-a \\ y-b \end{bmatrix} \\ & = \frac{1}{2} \times (\theta - \theta_0)^T \times H \times (\theta - \theta_0) \end{aligned} \tag{9}$$

$$\text{Hessian matrix} = \ \mathrm{H} \ = \begin{bmatrix} \nabla \mathrm{J}_{xx}(\mathrm{a,b}) & \nabla \mathrm{J}_{xy}(\mathrm{a,b}) \\ \nabla \mathrm{J}_{yx}(\mathrm{a,b}) & \nabla \mathrm{J}_{yy}(\mathrm{a,b}) \end{bmatrix}$$
$$= \begin{bmatrix} \nabla \mathrm{J}_{xx}(\theta_0) & \nabla \mathrm{J}_{xy}(\theta_0) \\ \nabla \mathrm{J}_{yx}(\theta_0) & \nabla \mathrm{J}_{yy}(\theta_0) \end{bmatrix}$$

(7)번 식에 (8), (9)을 이용하면 다음과 같은 (1)번 식이 된다.

$$\mathrm{J(x,y)} \approx \mathrm{J(a,b)} + (\theta - \theta_0)^T \times \nabla \mathrm{J}_\theta(\theta_0) + \frac{1}{2} \times (\theta - \theta_0)^T \times H \times (\theta - \theta_0) \tag{10}$$
$$\mathrm{J}(\theta) \approx \mathrm{J}(\theta_0) + (\theta - \theta_0)^T \times \nabla \mathrm{J}_\theta(\theta_0) + \frac{1}{2} \times (\theta - \theta_0)^T \times H \times (\theta - \theta_0) \tag{11}$$

(2)번에 대한 증명

(1)번에 대한 증명을 위에서 했으므로 (1)번을 사용해서 (2)를 증명하겠다.
(1)번 식에서 계산하기 쉽게 $\theta - \theta_0 = Z$라고 정의하자.
Θ와 θ_0는 각각 2x1 벡터로 (6)에서 정의했으므로 Z도 2x1 벡터가 된다.
$Z = \begin{bmatrix} \mathrm{z}_1 \\ \mathrm{z}_2 \end{bmatrix}$라고 하자.

(1) 번 식에서 $\theta - \theta_0 = Z$로 두면 다음과 같다.

$$\mathrm{J}(\theta_0) + (\theta - \theta_0)^T \times \nabla \mathrm{J}_\theta(\theta_0) + \frac{1}{2} \times (\theta - \theta_0)^T \times H \times (\theta - \theta_0)$$
$$= \mathrm{J}(\theta_0) + Z^T \times \nabla \mathrm{J}_\theta(\theta_0) + \frac{1}{2} \times Z^T \times H \times Z$$

우리는 여기서 Z의 변화량에 대한 $\mathrm{J}(\theta)$의 변화량이 0이 되는 지점의 Z값을 구할 것이다. 한마디로 $\frac{\partial \mathrm{J}(\theta)}{\partial \mathrm{Z}}$가 0이 되는 Z를 구하면 된다.

$$\frac{\partial \mathrm{J}(\theta)}{\partial \mathrm{Z}} = \frac{\partial}{\partial \mathrm{Z}} \left\{ \mathrm{J}(\theta_0) + Z^T \times \nabla \mathrm{J}_\theta(\theta_0) + \frac{1}{2} \times Z^T \times H \times Z \right\} = 0$$

$\frac{\partial \mathrm{J}(\theta)}{\partial \mathrm{z}_1}$ 를 구하면 다음과 같다.

$$\frac{\partial}{\partial z_1}\left\{ J(\theta_0) + Z^T \times \nabla J_\theta(\theta_0) + \frac{1}{2} \times Z^T \times H \times Z \right\}$$
$$= \frac{\partial}{\partial z_1}\left\{ J(\theta_0) + \begin{bmatrix} z_1 & z_2 \end{bmatrix} \times \begin{bmatrix} \nabla J_x(\theta_0) \\ \nabla J_y(\theta_0) \end{bmatrix} + \frac{1}{2} \times \begin{bmatrix} z_1 & z_2 \end{bmatrix} \times \begin{bmatrix} \nabla J_{xx}(\theta_0) & \nabla J_{xy}(\theta_0) \\ \nabla J_{yx}(\theta_0) & \nabla J_{yy}(\theta_0) \end{bmatrix} \times \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right\} \quad (11)$$
$$= \nabla J_x(\theta_0) + z_1 \times \nabla J_{xx}(\theta_0) + \frac{1}{2} \times \nabla J_{xy}(\theta_0) \times z_2 + \frac{1}{2} \times \nabla J_{yx}(\theta_0) \times z_2$$
$$= \nabla J_x(\theta_0) + z_1 \times \nabla J_{xx}(\theta_0) + \nabla J_{xy}(\theta_0) \times z_2$$

$\frac{\partial J(\theta)}{\partial z_2}$ 를 구하면 다음과 같다.

$$\frac{\partial}{\partial z_2}\left\{ J(\theta_0) + Z^T \times \nabla J_\theta(\theta_0) + \frac{1}{2} \times Z^T \times H \times Z \right\}$$
$$= \frac{\partial}{\partial z_2}\left\{ J(\theta_0) + \begin{bmatrix} z_1 & z_2 \end{bmatrix} \times \begin{bmatrix} \nabla J_x(\theta_0) \\ \nabla J_y(\theta_0) \end{bmatrix} + \frac{1}{2} \times \begin{bmatrix} z_1 & z_2 \end{bmatrix} \times \begin{bmatrix} \nabla J_{xx}(\theta_0) & \nabla J_{xy}(\theta_0) \\ \nabla J_{yx}(\theta_0) & \nabla J_{yy}(\theta_0) \end{bmatrix} \times \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right\} \quad (12)$$
$$= \nabla J_x(\theta_0) + z_2 \times \nabla J_{yy}(\theta_0) + \frac{1}{2} \times \nabla J_{xy}(\theta_0) \times z_1 + \frac{1}{2} \times \nabla J_{yx}(\theta_0) \times z_1$$
$$= \nabla J_y(\theta_0) + z_2 \times \nabla J_{yy}(\theta_0) + \nabla J_{yx}(\theta_0) \times z_1$$

$\frac{\partial J(\theta)}{\partial Z}$를 구하면 다음과 같다.

$$\frac{\partial J(\theta)}{\partial Z} = \begin{bmatrix} \frac{\partial J(\theta)}{\partial z_1} \\ \frac{\partial J(\theta)}{\partial z_2} \end{bmatrix}$$
$$= \begin{bmatrix} \nabla J_x(\theta_0) + z_1 \times \nabla J_{xx}(\theta_0) + \nabla J_{xy}(\theta_0) \times z_2 \\ \nabla J_y(\theta_0) + z_2 \times \nabla J_{yy}(\theta_0) + \nabla J_{yx}(\theta_0) \times z_1 \end{bmatrix} \quad (12)$$
$$= \begin{bmatrix} \nabla J_x(\theta_0) \\ \nabla J_y(\theta_0) \end{bmatrix} + \begin{bmatrix} \nabla J_{xx}(\theta_0) & \nabla J_{xy}(\theta_0) \\ \nabla J_{yx}(\theta_0) & \nabla J_{yy}(\theta_0) \end{bmatrix} \times \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$
$$= \nabla J_\theta(\theta_0) + H \times Z$$

$\nabla J_\theta(\theta_0) + H \times Z = 0$를 Z에 대한 식으로 간단히 나타내면 아래와 같다.

$$Z = -H^{-1} \times \nabla J_\theta(\theta_0)$$

그러므로 $\theta^* = \theta_0 - Z = \theta_0 - H^{-1} \times \nabla_\theta J(\theta_0)$ 이 된다.

Conclusion

θ 가 multi variable일 때, Second-order optimization loss 그래프의 식은 Taylor expansion에 의해 $J(\theta) \approx J(\theta_0) + (\theta - \theta_0)^T \times \nabla_\theta J(\theta_0) + \frac{1}{2} \times (\theta - \theta_0)^T \times H \times (\theta - \theta_0)$가 된다.
gradient descent를 하기 위한 θ 업데이트 식은 $\theta^* = \theta_0 - H^{-1} \times \nabla_\theta J(\theta_0)$이다.

References

1. http://cs231n.stanford.edu/ lecture 8
2. https://ratsgo.github.io/