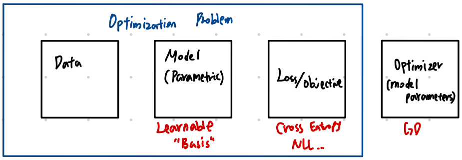
Gradient Descent



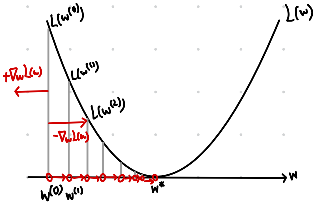
Gradient descent solution:

$$w^{(t+1)} = w^{(t)} - \varepsilon \nabla_w L(w^{(t)})$$

where:

$$\nabla_w L(w^{(t)}) = \Phi \Phi^T w^{(t)} - \Phi y$$

$$L(w) = \| \Phi^T w - y \|^2$$



How should we select the learning rate?

1. Let's dive deep into the innerworkings of gradient descent.

- Let $g^{(t)} = \nabla_w L(w^{(t)})$ and let H be the Hessian matrix s.t.,

$$[H^{(t)}]_{ij} = \frac{\partial^2 L}{\partial w_i \partial w_j}(w^{(t)})$$

ex)
$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- Recall the Taylor series of a function (second order)

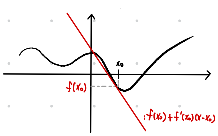
Taylor series of a function is an infinite sum of terms that are expressed in terms of the functions derivatives at a single point

☆ For most common functions, the function and the sum of its Taylor series are equal near this point.

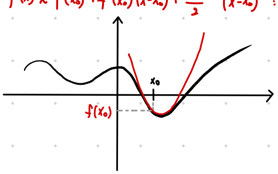
- Definition: $f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n.$

ex) $f: \mathbb{R} \to \mathbb{R}$

$f(x) \approx f(x_0) + f'(x_0)(x-x_0)$: First Order Taylor Series



$f(x) \approx f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2$: Second Order Taylor Series



Recap:

Examples of Hessian Matrices

① Let $f: \mathbb{R}^n \to \mathbb{R}$

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

② Let $f: \mathbb{R}^2 \to \mathbb{R}$

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial^2 y} \end{bmatrix}$$

let $f(x,y) = x^2y^3$

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(x^2y^3) = 2xy^3$$

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(x^2y^3) = 3x^2y^2$$

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial}{\partial x}(2xy^3) = 2y^3$$

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}(x^2y^3)\right) = \frac{\partial}{\partial x}(3x^2y^2) = 6xy^2$$

③ Let $g(x,y) = x^3 + 2y^2 + 3xy^2$

$$H_g = \begin{bmatrix} \frac{\partial^2 g}{\partial x^2} & \frac{\partial^2 g}{\partial x \partial y} \\ \frac{\partial^2 g}{\partial y \partial x} & \frac{\partial^2 g}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 6x & 6y \\ 6y & 4+6x \end{bmatrix}$$

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}(x^3+2y^2+3xy^2)\right) = \frac{\partial}{\partial x}(3x^2+3y^2) = 6x$$

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}(x^3+2y^2+3xy^2)\right) = \frac{\partial}{\partial x}(4y+6xy) = 6y$$

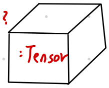
$$\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}(x^3+2y^2+3xy^2)\right) = \frac{\partial}{\partial y}(3x^2+3y^2) = 6y$$

$$\frac{\partial}{\partial y}\left(\frac{\partial}{\partial y}(x^3+2y^2+3xy^2)\right) = \frac{\partial}{\partial y}(4y+6xy) = 4+6x$$

When $f: \mathbb{R}^d \to \mathbb{R}$,

$$f(x) \approx f(x_0) + \nabla f(x_0) \cdot (x - x_0) + \frac{1}{2} \underbrace{(x-x_0)^T}_{1\times d} \underbrace{H(x_0)}_{d\times d} \underbrace{(x-x_0)}_{d\times 1}$$

☆ What if we wanted Taylor approximation of degree 3?



Taylor Series of $L(w)$ around $w^{(t)}$:

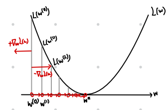
$$L(w) \approx L(w^{(t)}) + (w - w^{(t)})^T g^{(t)} + \frac{1}{2}(w - w^{(t)})^T H^{(t)} (w - w^{(t)})$$

① Substitute gradient descent update, $w^{(t+1)} = w^{(t)} - \varepsilon g^{(t)}$, into the Taylor series. : $L(w) \to L(w^{(t+1)})$

: Why?: To understand $L(w^{(t+1)}) - L(w^{(t)})$.

↳ We want this to be negative.

: We want $L(w^{(2)}) \le L(w^{(1)})$



$g^{(t)} = \nabla_w L(w^{(t)})$

$$L(w^{(t+1)}) \approx L(w^{(t)}) + (w^{(t+1)} - w^{(t)})^T g^{(t)} + \frac{1}{2}(w^{(t+1)} - w^{(t)})^T H^{(t)} (w^{(t+1)} - w^{(t)})$$

Since $w^{(t+1)} = w^{(t)} - \varepsilon \cdot g^{(t)}$, then $w^{(t+1)} - w^{(t)} = -\varepsilon \cdot g^{(t)}$

Then, $L(w^{(t+1)}) \approx L(w^{(t)}) - \varepsilon (g^{(t)})^T g^{(t)} - \frac{1}{2}\varepsilon (g^{(t)})^T H^{(t)} \varepsilon g^{(t)}$

$\approx L(w^{(t)}) - \varepsilon (g^{(t)})^T g^{(t)} + \frac{1}{2}\varepsilon^2 (g^{(t)})^T H^{(t)} g^{(t)}$

And $L(w^{(t+1)}) = L(w^{(t)} - \varepsilon g^{(t)})$

· We want $L(w^{(t+1)}) \le L(w^{(t)})$ for GD to converge

Since $L(w^{(t+1)}) \approx L(w^{(t)}) - \varepsilon (g^{(t)})^T g^{(t)} + \frac{1}{2}\varepsilon^2 (g^{(t)})^T H^{(t)} g^{(t)}$

$L(w^{(t+1)}) - L(w^{(t)}) \approx -\varepsilon (g^{(t)})^T g^{(t)} + \frac{1}{2}\varepsilon^2 (g^{(t)})^T H^{(t)} g^{(t)}$

Hence, since we want $L(w^{(t+1)}) \le L(w^{(t)})$, which is equivalent to $L(w^{(t+1)}) - L(w^{(t)}) \le 0$,

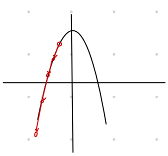
we need to find $-\varepsilon g^T g + \frac{\varepsilon^2}{2} g^T H g \le 0$

① we know $g^T g \ge 0$ (always positive), hence $-\varepsilon g^T g$ is always negative

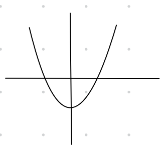
② Then, what if $\frac{\varepsilon^2}{2} g^T H g$ is negative?

: then $-\varepsilon g^T g + \frac{\varepsilon^2}{2} g^T H g \le 0$ is satisfied

. Hence, if $g^T H g \le 0$ (as long as $\varepsilon > 0$), then regardless of ε, $L(w^{(t+1)}) \le L(w^{(t)})$



$g^T H g \le 0$: Concave



$g^T H g \ge 0$: Convex

③ But, what if $g^T H g > 0$?

: Since we want $-\varepsilon g^T g + \frac{\varepsilon^2}{2} g^T H g \le 0$

$$= \frac{\varepsilon^2 g^T H g}{2} \le \varepsilon g^T g$$

$$= \frac{\varepsilon g^T H g}{2} \le g^T g$$

$$= \varepsilon \le \frac{2 g^T g}{g^T H g}$$

Hence, when $\varepsilon \le \frac{2g^Tg}{g^THg}$, $L(w^{(t+1)}) \le L(w^{(t)})$

In Summary:

- When $(g^{(t)})^T H^{(t)} g^{(t)} \le 0$, we have no issue (i.e., if $L(w)$ is concave)
- But, when $(g^{(t)})^T H^{(t)} g^{(t)} > 0$, to satisfy $L(w^{(t+1)}) < L(w^{(t)})$, we need to pick ε carefully.

Convergence Rate : How fast is it going to converge?

- Assume that H has a spectral radius smaller than L (Lipschitz constant):

s.t. $v^T H(w) v \le L \|v\|^2$

Lipschitz Constant

In other words, $\lambda_{max}(H(w)) \le L \ \forall w.$

Spectral radius of $H(w)$

Then, $\frac{2\cancel{g^Tg}}{L \cdot \cancel{g^Tg}} \le \frac{2g^Tg}{g^THg}$, which implies $\varepsilon < \frac{2}{L} \le \frac{2g^Tg}{g^THg}$

Hence, if we set $\varepsilon < \frac{2}{L}$, we will guarantee convergence

I think it will be related to the concept of Smoothness. But not sure yet.

Rewriting the Taylor expansion we have:

$$L(w^{(t+1)}) = L(w^{(t)}) - \varepsilon (g^{(t)})^T g^{(t)} + \frac{1}{2}\varepsilon^2 (g^{(t)})^T H^{(t)} g^{(t)} \le L(w^{(t)}) - \varepsilon \cdot \|g^{(t)}\|^2 + \frac{L}{2}\varepsilon^2 \|g^{(t)}\|^2 \quad \text{since} \quad g^THg \le L\|g\|^2$$

$$= L(w^{(t)}) + \left(\frac{L}{2}\varepsilon^2 - \varepsilon\right)\|g^{(t)}\|^2$$

Since $\varepsilon < \frac{2}{L}$, let's assume $\varepsilon = \frac{1}{L}$

Then,

$$L(w^{(t+1)}) \le L(w^{(t)}) + \left(\frac{L}{2}\cdot\frac{1}{L^2} - \frac{1}{L}\right)\|g^{(t)}\|^2 = L(w^{(t)}) - \frac{\|g^{(t)}\|^2}{2L}$$

So, we have

$$L(w^{(t+1)}) \le L(w^{(t)}) - \frac{\|g^{(t)}\|^2}{2L}$$

Lipschitz

$$\Rightarrow \frac{\|g^{(t)}\|^2}{2L} \le L(w^{(t)}) - L(w^{(t+1)})$$

$$\Rightarrow \|g^{(t)}\|^2 \le 2L\left(L(w^{(t)}) - L(w^{(t+1)})\right)$$

Let's now take sum over t steps of gradient descent

$$\sum_{k=0}^{t-1} \|g^{(k)}\|^2 \le 2L \cdot \sum_{k=0}^{t-1} \left(L(w^{(k)}) - L(w^{(k+1)})\right)$$

Lower bound left-hand-side by:

$$t g_m \le \sum_{k=1}^{t} \|g^{(k)}\|^2 \quad \text{where} \quad g_m = \min_{j \in \{1,\dots,t\}} \|g^{(j)}\|^2$$

The right-hand-side is equal to

$$\sum_{k=0}^{t-1} \left(L(w^{(k)}) - L(w^{(k+1)}) \right) = L(w^{(0)}) - L(w^{(1)}) + L(w^{(1)}) - L(w^{(2)}) + \dots + L(w^{(t-1)}) - L(w^{(t)})$$

$$= L(w^{(0)}) - L(w^{(t)}) \le L(w^{(0)}) - L(w^{*}) \quad \text{because} \quad L(w^{*}) \le L(w^{(t)})$$

Therefore, we can write:

$$t \cdot g_m \le \sum_{k=0}^{t-1} \|g^{(k)}\|^2 \le 2L \sum_{k=0}^{t-1} \left(L(w^{(k)}) - L(w^{(k+1)}) \right) \le 2L \left(L(w^{(0)}) - L(w^{(*)}) \right)$$

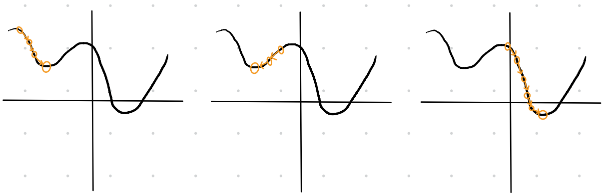
$$\Rightarrow g_m \le \frac{2L\left(L(w^{(0)}) - L(w^{(*)})\right)}{t}$$

Hence, $g_m = \nabla_w L$ goes to zero at a rate $O(\frac{1}{t})$

- How many iterations do we need to satisfy $\|\nabla_w L\|^2 \le \delta$? : $O(\frac{1}{\delta})$

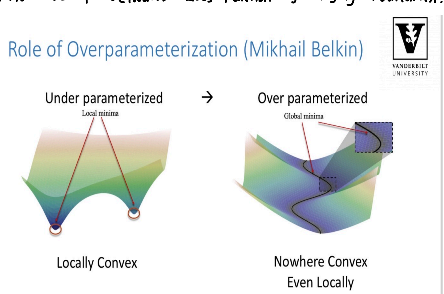
Gradient Descent in Non-convex Problems.

- Sensitive to initialization and could get stuck in local optima



Local Optima　　Local Optima　　Global Optima

And Neural Network's Loss Function is Highly Nonconvex. Why does the optimization work?



Infinitely many Solution