

# **기초미분과 최적화**

**윤정훈**

# 들어가며

- 대표적인 분류기 SVM에서는 여백이라는 개념을 정의하고 여백을 최대화하는 결정 직선을 찾는 것이 목적이다.
- 신경망에서는 모델을 추정하기 위하여, 목적함수 또는 손실함수를 최소로 하는 모수를 찾는 것이 목적이다.
- 베이시언 분류기에서는 분류 오류를 최소화 하는 분류기를 찾는 것이 목적이다.

# 최적화 문제

- ✓  $J(\theta)$  : 목적 함수(target function) 또는 비용 함수(cost function)이며, 우리가 최대화하거나 최소화하려는 함수이다.
- ✓  $\theta$  : 매개 변수

- 최대화 문제

$J(\theta)$ 를 최대로 하는  $\hat{\theta}$ 를 찾아라. 즉  $\hat{\theta} = \arg \max_{\theta} J(\theta)$

- 최소화 문제

$J(\theta)$ 를 최소로 하는  $\hat{\theta}$ 를 찾아라. 즉  $\hat{\theta} = \arg \min_{\theta} J(\theta)$

# 분석적 방법 vs 수치해석적 방법

- 분석적 방법

- 입력 : 목적 함수  $J(\theta)$
- 출력 :  $\hat{\theta}$  (최고 점 또는 최저 점)
- 알고리즘

①  $J(\theta)$ 를  $\theta$ 로 미분한다.

② 방정식  $\frac{\partial J(\theta)}{\partial \theta} = 0$ 을 만족하는  $\hat{\theta}$ 를 구한다.

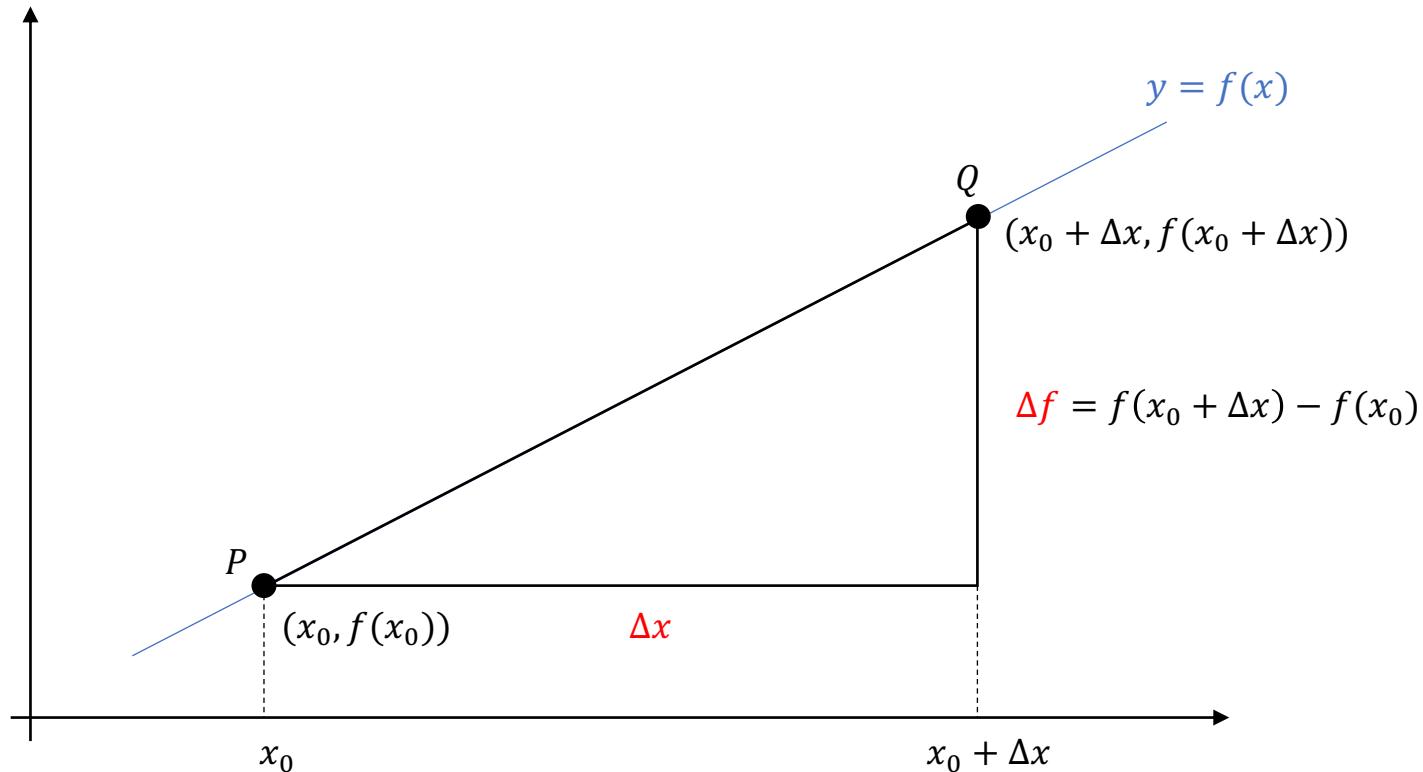
③  $\hat{\theta}$ 를 리턴한다.

# 분석적 방법 vs 수치해석적 방법

- 수치해석적 방법
  - 선형회귀모형, 로지스틱회귀모형에서 모수의 값을 추정하기 위하여, 정규방정식을 풀어야 한다. 정규방정식을 푸는 방법 중의 하나는 방정식의 초기해를 설정한 후, 초기해를 지속적으로 업데이트 하여, 우리가 구하고자 하는 해 즉 최대값 또는 최소값에 가장 근접하도록 하는 방법이다. 이러한 방법을 수치해석적으로 해를 구하는 방법이라고 한다.
  - 대표적인 이론이 경사하강법(Gradient Descent)이다. 이 이론을 정확하게 이해하기 위해서는 기본적인 미분, 편미분에 대한 개념을 정확하게 이해해야 한다.

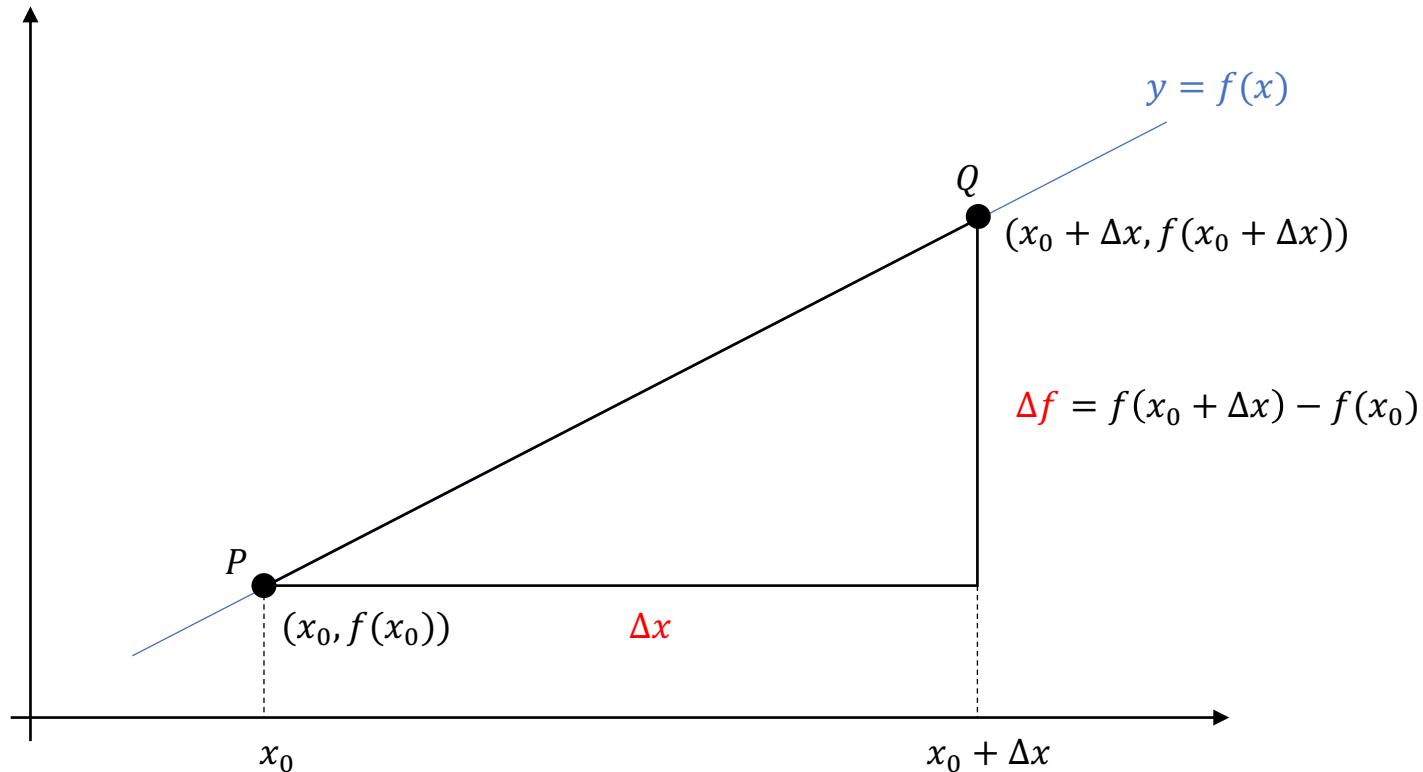
# 증분

- $x$ 의 증분( $\Delta x$ ) :  $x$ 의 증가량
- $y$ 의 증분( $\Delta y$ ) :  $y$ 의 증가량

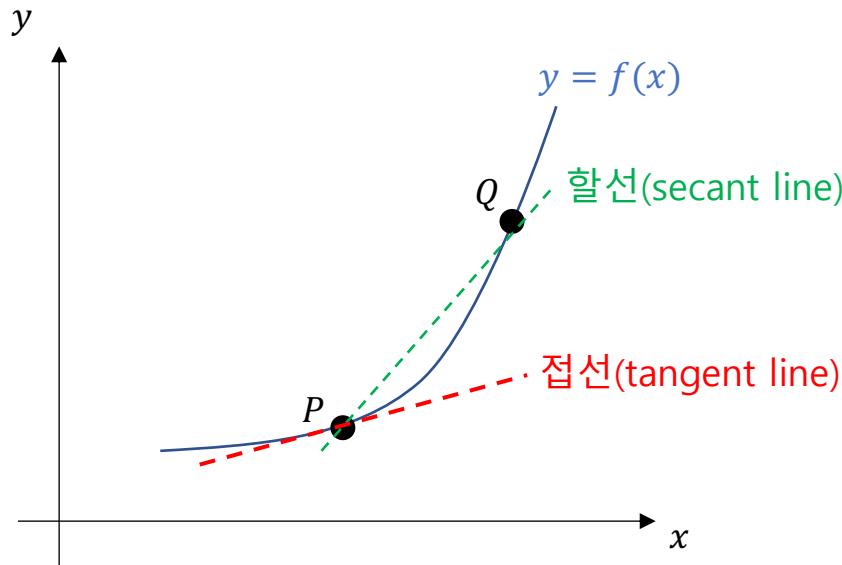


# 기울기(slope)

- 기울기 :  $\frac{y\text{의 증분}}{x\text{의 증분}} = \frac{\Delta y}{\Delta x} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$



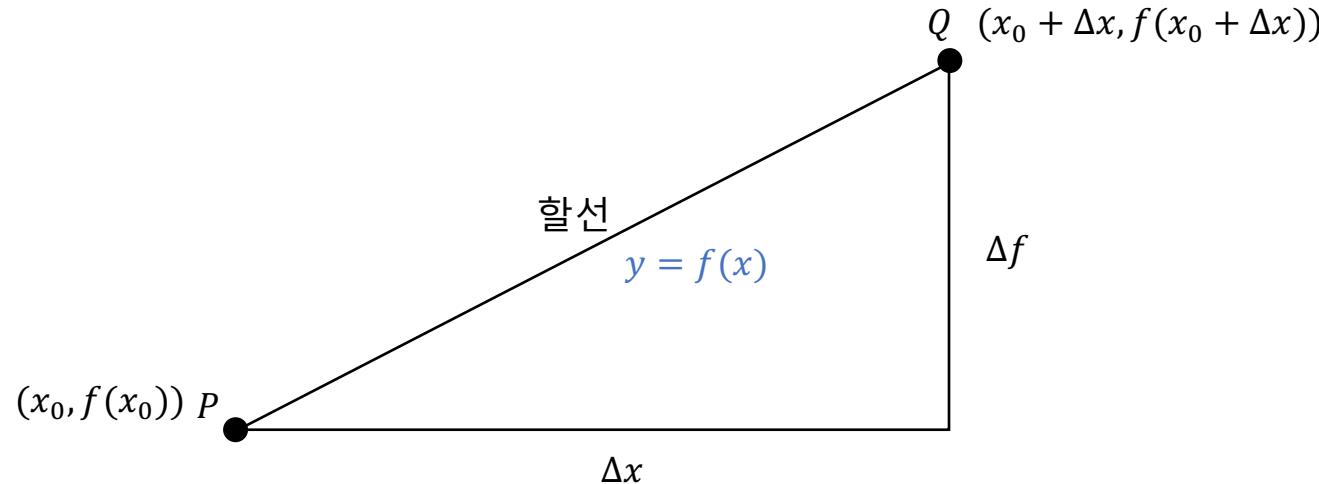
# 도함수(Derivatives)



도함수는 그래프  $f(x)$ 의 접선의 기울기이다. 그러면 접선은 무엇인가?

- 그래프의 한 지점(point)에서 만나는 선이 절대 아니다.
- 접선은 두 지점 사이의 할선이며, 두 지점 사이의 거리가 0으로 갈 때의 할선의 극값이다.

# 도함수(Derivatives)



- 도함수 :  $P \rightarrow Q$  일 때, 할선  $PQ$  기울기의 극한 값

$$- \lim_{\Delta x \rightarrow 0} \frac{\Delta f}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} = f'(x_0) = \frac{df(x_0)}{dx} = \frac{dy}{dx}$$

# 여러가지 함수의 도함수

- $y = \{f(x)\}^n \Rightarrow \frac{dy}{dx} = n \cdot f'(x) \cdot \{f(x)\}^{n-1}$
- $y = \log f(x) \Rightarrow \frac{dy}{dx} = \frac{f'(x)}{f(x)}$
- $y = e^{f(x)} \Rightarrow \frac{dy}{dx} = f'(x) \cdot e^{f(x)}$
- $y = f(x) \cdot g(x) \Rightarrow \frac{dy}{dx} = f'(x) \cdot g(x) + f(x) \cdot g'(x)$
- $y = \frac{f(x)}{g(x)} = \frac{f'(x) \cdot g(x) - f(x)g'(x)}{\{g(x)\}^2}$

# 도함수의 예제

- $y = \log_e x \Rightarrow y' = ?$

- $y = \log_e(x^3) \Rightarrow y' = ?$

- $y = e^{4x} \Rightarrow y' = ?$

- $y = xe^{3x} \Rightarrow y' = ?$

- $y = \frac{e^x}{1+e^x} \Rightarrow y' = ?$

# Chain Rule

- 합성함수  $y = f(g(t))$ 의 도함수  $\frac{dy}{dt}$ 
  - $\frac{dy}{dt} = \frac{dy}{dx} \frac{dx}{dt}$
- 합성함수  $y = (f \circ g)(x)$ 의 도함수  $\frac{d}{dx} f(g(x))$ 
  - $\frac{d}{dx} f(g(x)) = f'(g(x))g'(x)$
- 예를 들어,  $y = \sin x$  이고  $x = t^2$  일 때,  $\frac{dy}{dt}$ 를 구해보자.
  - $\frac{dx}{dt} = 2t$  ,  $\frac{dy}{dx} = \cos x$
  - $\frac{d}{dt}(\sin(t^2)) = \left(\frac{dy}{dx}\right) \left(\frac{dx}{dt}\right) = (\cos x)(2t) = 2t \cdot \cos(t^2)$

# 고차원의 도함수

- 고차원의 도함수는, 도함수의 도함수들을 의미한다.

- $f'(x) = \frac{df}{dx} = Df$

- $f''(x) = \frac{d^2f}{dx^2} = D^2f$

- $f'''(x) = \frac{d^3f}{dx^3} = D^3f$

- $f^{(n)}(x) = \frac{d^n f}{dx^n} = D^n f$

# 다변수함수의 미분

- 이전까지는 독립변수가 하나인 함수를 대상으로 도함수를 구하였다. 하지만 다양한 분야에서 사용되는 많은 이론은 여러 변수들에 의해 함수값이 결정되는 다변수함수의 형태를 띠고 있다.
- 예를 들면 경제학에서 사용하는 수요함수는 자신의 가격( $P_1$ )뿐만 아니라 타 재화의 가격( $P_2, \dots, P_n$ ) , 소득( $M$ ) , 기호( $T$ )에 의해서 결정된다. 따라서 수요함수는  $D = f(P_1, P_2, \dots, P_n, M, T)$ 로 나타낼 수 있다. 이 때 이들 독립변수가 변하면 수요의 변화는 어떻게 될까?

# 다변수함수의 미분의 종류

- 편도함수
- 전미분

# 1차편도함수

2변수함수  $z = f(x, y)$ 에서  $y$ 변수가 특정한 값에 고정되어 변하지 않는다고 가정하면( $y$ 변수를 상수로 가정하면) 2변수함수는 실질적으로 독립변수가 하나인 1변수함수가 된다. 이 함수를  $x$ 로 미분하면 도 함수가 구해지는데 이것을  $x$ 의 편도함수(partial derivative)라 한다. 마찬가지로  $x$ 변수가 특정한 값에 고정되어 변하지 않는다고 가정하면  $y$ 의 편도함수를 구할 수 있다.

# 1차편도함수

- 편도함수는 다음과 같이 정의된다.

함수  $f(x, y)$ 가 모든 점에서 미분 가능할 때  $x$ 와  $y$ 에 대한 각각의 편도함수는 다음과 같이 정의된다.

$$\frac{\partial f}{\partial x} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

$$\frac{\partial f}{\partial y} = \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}$$

# 1차편도함수

- 편도함수의 정의는 한 변수의 평균변화율을 나타내는 차원에서 실질적으로 동일한 의미를 갖는다.

편도함수를 나타내기 위해  $f_x$ ,  $\frac{\partial f}{\partial x}$ ,  $D_x f$  등의 기호로 표시한다.

- 어떤 특정한 점  $(a, b)$ 에서  $x$ 의 편미분계수는  $\frac{\partial f}{\partial x}|_{(a,b)}$  또는  $f_x(a, b)$ 로,  $y$ 의 편미분계수는  $\frac{\partial f}{\partial y}|_{(a,b)}$  또는  $f_y(a, b)$ 로 표시한다. 점  $(a, b)$ 에서 편미분계수  $f_x(a, b)$ 는  $y = b$ 로 고정되어 있는 상태에서  $x = a$ 에서  $x$ 단위 변화에 대한  $z$ 의 변화율을 나타낸다. 다변수함수의 경우에도 관심대상의 변수 외의 모든 변수는 상수 취급하므로 실질적으로 1변수함수가 된다. 이와 같이 편도함수를 구하는 것을 "편미분한다"라고 한다.

# 1차편도함수 예제1

- $f(x, y) = xy^2 + x^2y$ 의  $x$ 와  $y$ 의 편도함수를 구해보자.

$x$ 의 편도함수는  $y$ 를 상수로 하고  $x$ 에 대해서만 미분하므로  $f_x(x, y) = y^2 + 2xy$ 가 구해진다.

$y$ 의 편도함수는  $x$ 를 상수로 하고  $y$ 에 대해서만 미분하므로  $f_y(x, y) = 2xy + x^2$ 가 구해진다.

점  $(1,2)$ 에서,

$x$ 의 편미분계수를 구하면,  $f_x(1,2) = 2^2 + 2 \times 1 \times 2 = 8$  이고

$y$ 의 편미분계수를 구하면,  $f_y(1,2) = 2 \times 1 \times 2 + 1^2 = 5$  이다.

# 1차편도함수 예제2

- $f(x, y) = \ln(x^2 + 2xy - y^2)$ 에서  $f_x, f_y$ 를 구하고 점 (1,1)에서 각 변수의 편미분계수를 구하라.

$x^2 + 2xy - y^2 = u$ 라고 놓으면,  $f(x, y) = \ln u$ 가 된다. 연쇄법칙에 의하여

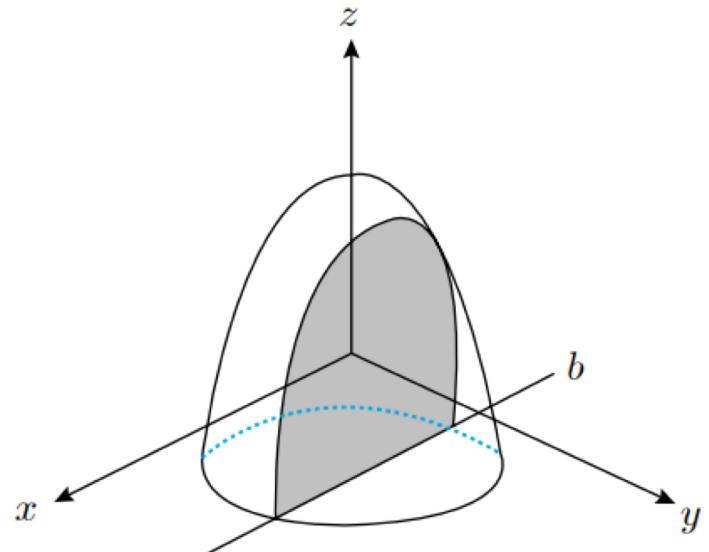
$$x\text{의 편도함수는 } f_x(x, y) = \frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial x} = \frac{1}{u} \cdot \frac{\partial}{\partial x} (x^2 + 2xy - y^2) = \frac{2x+2y}{x^2+2xy-y^2},$$

$$y\text{의 편도함수는 } f_y(x, y) = \frac{\partial z}{\partial y} = \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial y} = \frac{1}{u} \cdot \frac{\partial}{\partial y} (x^2 + 2xy - y^2) = \frac{2x-2y}{x^2+2xy-y^2} \text{ 가 성립한다.}$$

$$x\text{의 편미분계수 } f_x(1, 1) = \frac{4}{2} = 2 \text{ 이고,}$$

$$x\text{의 편미분계수 } f_y(1, 1) = \frac{0}{2} = 0 \text{ 이다.}$$

# 1차편도함수의 기하학적 의미



$z = f(x, y)$ 의 그래프

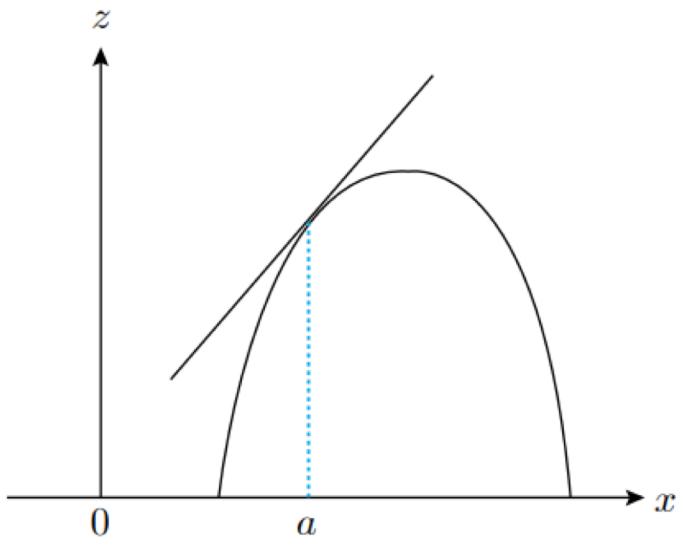
$y = b$ 는 기하학적으로

$xz$ 평면에 평행하고  $y = b$ 선을 지나는 평면이다.

이 평면으로  $z = f(x, y)$ 의 곡면을 절단하면

단면에 나타나는 함수는  $z = f(x, b)$ 이다.

# 1차편도함수의 기하학적 의미



$x$ 의 편미분계수 :  $f_x(x, y)$

$f_x(a, b)$ 는 함수

$z = f(x, b)$ 의  $x = a$ 에서 그은 접선의 기울기와 일치한다.

즉,  $y$ 방향으로 전혀 움직이지 않고

$x = a$ 에서  $x$ 축 방향으로 단위 변화할 때  $z$ 의 변화율을 나타낸다.

$y$ 의 편도함수  $f_y(a, b)$ 의 경우는,

$x = a$  평면으로  $z = f(x, y)$  공간을 자르고 나타난

함수  $z = f(a, y)$ 의  $y = b$ 에서 그은 접선의 기울기가 된다.

# 다변수함수의 편미분

- 다변수함수  $z = f(x_1, x_2, \dots, x_n)$ 에서  $x_i$ 를 제외한 모든 변수가 고정되었다고 하면( $x_i$ 를 제외한 모든 변수가 상수라 하면)  $x_i$ 의 편도함수는 다음과 같이 나타낸다.

$$\frac{\partial f}{\partial x_i} = f_i(x_1, x_2, \dots, x_n), \quad (i = 1, \dots, n)$$

ex)  $f(x, y, z) = xy + yz + xz$ 의 1차 편도함수를 구하라.

$x$ 의 1차편도함수는  $y, z$ 를 상수로 하고  $x$ 에 대해서만 미분하면

$$f_x = y + z$$

$$f_y = x + z$$

$$f_z = x + y$$

가 된다.

# 벡터의 편미분

- 다변수함수  $z = f(x_1, x_2 \dots, x_n)$ 라고 하고, 벡터  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ 이라고 하자. 즉,  $z = f(\mathbf{x})$ 이다. 이 때, 벡터  $\mathbf{x}$ 의 편미분  $\frac{\partial z}{\partial \mathbf{x}}$ 는 다음과 같이 나타낸다.

$$\frac{\partial z}{\partial \mathbf{x}} = \left( \frac{\partial z}{\partial x_1}, \frac{\partial z}{\partial x_2}, \dots, \frac{\partial z}{\partial x_n} \right)$$

ex)  $\mathbf{x} = (x_1, x_2, x_3)$ 에 대하여,  $f(\mathbf{x}) = f(x_1, x_2, x_3) = x_1x_2 + x_2x_3 + x_3x_1$ 의 편미분  $\frac{\partial z}{\partial \mathbf{x}}$ 을 구하라.

$$\frac{\partial f}{\partial x_1} = x_2 + x_3$$

$$\frac{\partial f}{\partial x_2} = x_1 + x_3$$

$$\frac{\partial f}{\partial x_3} = x_2 + x_1$$

이므로  $\frac{\partial z}{\partial \mathbf{x}} = (x_2 + x_3, x_1 + x_3, x_2 + x_1)$ 이다.

# 2차편도함수

- 함수  $z = f(x, y)$ 의 1차편도함수  $f_x(x, y), f_y(x, y)$ 는  $x, y$ 의 형태를 띤다. 따라서 1차편도함수가 미분 가능하면 편도함수 정의에 의해서 1차편도함수를 가지고 2차편도함수를 구할 수 있다.

$$\bullet \quad f_{xx} = \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial x} \right) = \frac{\partial^2 f}{\partial x^2}$$

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

$$\bullet \quad f_{yx} = \frac{\partial}{\partial x} \left( \frac{\partial f}{\partial y} \right) = \frac{\partial^2 f}{\partial y \partial x}$$

$$\bullet \quad f_{yy} = \frac{\partial}{\partial y} \left( \frac{\partial f}{\partial y} \right) = \frac{\partial^2 f}{\partial y^2}$$

# 2차편도함수 예제

- 함수  $f(x, y) = x^3 + 2xy^3$ 의 2차편도함수를 구하라.

$x, y$ 에 대한 편미분하면 1차편도함수는

$$- f_x = 3x^2 + 2y^3$$

$$- f_y = 6xy^2$$

2차편도함수는

$$- f_{xx} = 6x$$

$$- f_{xy} = 6y^2 = f_{yx}$$

$$- f_{yy} = 12xy$$

# 1차/2차 편도함수 예제

- 1차 편도함수를 구하여라.

- $f(x, y) = (x - 6y)(2x + 3y^2)$

- $g(t_1, t_2) = \frac{2t_2 + 4t_1}{t_2^2 - 3t_1}$

- $f(x, y, z) = (3x^2 + z^2)(z - y)$

- 2차 편도함수를 구하여라.

- $f(x_1, x_2, x_3) = 2x_1^3x_2 + 5x_3^4x_2$

- $f(x, y, z) = x - \sqrt{y^2 + z^2}$

- $f(x, y) = x \ln y$

# 전미분 들어가기

함수  $z = f(x, y)$ 의 1차편도함수  $f_x$ 는  $y$ 가 고정되어 있는 상태에서  $x$ 단위 변화에 대한  $z$ 의 변화율을 나타낸 것이다. 또한  $f_y$ 도  $x$ 가 고정되어 있는 상태에서  $y$ 단위 변화에 대한  $z$ 의 변화율을 나타낸 것이다. 두 독립변수  $x, y$ 가 동시에 변화할 때  $z$ 의 총변화량은 어떻게 될까?

# 전미분의 개념

두 사람이 조깅을 한다. A는 10분 동안 10km를 달리고, B는 30초 동안 0.75km를 달린다고 하자.

단위시간을 1분으로 하면 A는 1분 동안 1km( $v_A = \frac{dS_A}{dt_A}$ )를, B는 1분 동안 1.5km( $v_B = \frac{dS_B}{dt_B}$ ) 속도로 달린 것이다.

시간의 단위 변화( $x$ 의 단위변화)에 따른 각 개인의 위치 변화, 각 개인의 속도 개념으로 도함수를 이해한 것이다.

각 개인의 속도와 달린 시간을 알면 그 시간 동안 각 개인의 조깅거리를 구할 수 있다.

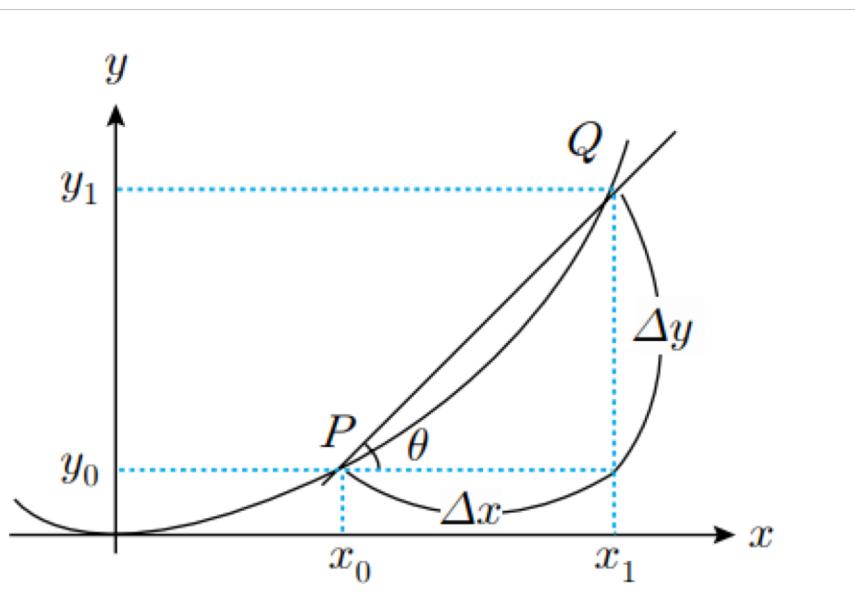
A가 2분 동안 달리고 B가 3분 동안 달렸다면 A가 달린 거리는  $1 \times 2 = 2\text{km}$ 이고 B가 달린 거리는  $1.5 \times 3 = 4.5\text{km}$ 가 된다.

이것을 일반화하면 일정 시간 동안( $dt_A$ ) A가 달린 거리 변화는  $dS_A = v_A dt_A$ 가 되고, 일정 시간 동안( $dt_B$ ) B가 달린 거리 변화는  $dS_B = v_B dt_B$ 가 된다.

이 때 A,B 두 사람이 달린 총거리 변화는  $ds = dS_A + dS_B = v_A dt_A + v_B dt_B$ 로 나타낼 수 있다.

# 전미분의 개념

$x$ 는 시간,  $y$ 는 거리로 생각해보자.



시간이  $\Delta x$ 만큼 변화할 때 위치는  $P$ 점에서  $Q$ 점으로

이동하였다. 이 때 속도는 사선  $PQ$ 의 기울기, 즉

$\frac{\Delta y}{\Delta x}$ 가 된다. 그러면 거리 변화  $\Delta y$ 는 속도  $\frac{\Delta y}{\Delta x}$ 에 시간의

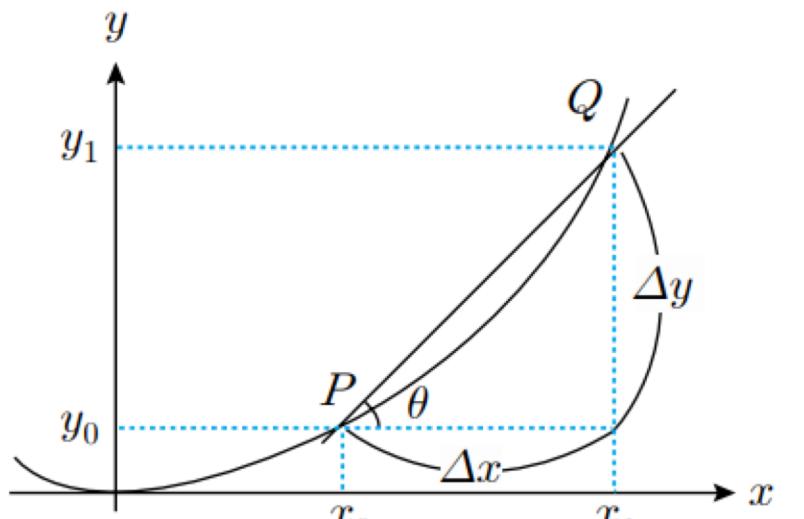
변화량  $\Delta x$ 를 곱한 것과 같다.  $\Delta y = \left(\frac{\Delta y}{\Delta x}\right) \Delta x$

이제  $\Delta x$ 가 0에 가까울 정도로 미세하게 변화한다면

$$dy = \left(\frac{dy}{dx}\right) dx = f'(x)dx \text{이다.}$$

여기서  $dx, dy$ 를  $x, y$ 의 미분이라 한다.

# 전미분의 개념



$$\left( \frac{dy}{dx} \right) = \frac{(dy)}{(dx)} = f'(x)$$

단위시간당 거리 변화율(속도)은 일정 시간 동안  
변화된 총거리를 시간으로 나눈 값과 같다.

즉, 한 묶음으로서의 도함수  $\left( \frac{dy}{dx} \right)$ 는 다시 2개의 미분  
 $dx, dy$ 의 비율로 나타낼 수 있다. 이처럼 도함수를  
이용하여  $x$ 의 무한소 변화에 대한  $y$ 의 변화량을 구하는  
것은 "전미분한다"라고 한다.

즉, 임의의 한 점  $x_0$  근방에서  $x$ 가 아주 미세하게 변했을 때  
 $y$ 의 변화량이  $dy = f'(x)dx$ 가 된다는 것을 의미한다.

# 1변수함수의 전미분 예제

- 함수  $y = x^2 + \ln x$ 를 전미분하라.

$$f'(x) = 2x + \frac{1}{x} \text{이므로}$$

$$dy = f'(x)dx = \left(2x + \frac{1}{x}\right)dx \text{가 된다.}$$

# 2변수함수의 전미분

$dS = dS_A + dS_B = v_A dt_A + v_B dt_B$  식에서 총거리 변화는  $A$ 의 속도( $v_A$ )와  $A$ 가 달린 시간( $dt_A$ )의 곱과  $B$ 의 속도( $v_B$ )와  $B$ 가 달린 시간( $dt_B$ )의 곱을 합해서 구해진다.

여기서 총거리 변화는 바로 2변수함수에서의 전미분을 의미한다.

그런데  $A$ 의 속도  $v_A$ 는  $B$ 가 간 거리와 관계없이  $A$ 의 단위시간당 거리로,  $B$ 의 속도  $v_B$ 도  $A$ 가 간 거리와 관계없이  $B$ 의 단위시간당 거리로 구해지므로 각 개인의 속도는 편미분의 의미를 갖는다.

즉 일반적으로, 2변수함수  $z = f(x, y)$ 를 전미분하면  $dz = \left(\frac{\partial z}{\partial x}\right) dx + \left(\frac{\partial z}{\partial y}\right) dy = f_x dx + f_y dy$ 이다.

즉  $z$ 의 총변화량은  $x$ 변수에 의한  $z$ 의 총변화량과  $y$ 변수에 의한  $z$ 의 총변화량의 합으로 표시된다.

# 다변수함수의 전미분

- 전미분 개념을 다변수함수  $y = f(x_1, x_2, \dots, x_n)$ 으로 확장하면 다음과 같다.

- $dy = \left(\frac{\partial y}{\partial x_1}\right)dx_1 + \left(\frac{\partial y}{\partial x_2}\right)dx_2 + \cdots + \left(\frac{\partial y}{\partial x_n}\right)dx_n$

- $dy = f_1dx_1 + f_2dx_2 + \cdots + f_ndx_n$

각 변수의 편미분에 변수의 변화량을 곱한 값의 합으로 표시된다.

ex)  $z = x^2y^2 - 3x^2 + 2$ 를 전미분하라.

1차편도함수는 각각

$$f_x = 2xy^2 - 6x ,$$

$$f_y = 2x^2y \text{ 이므로}$$

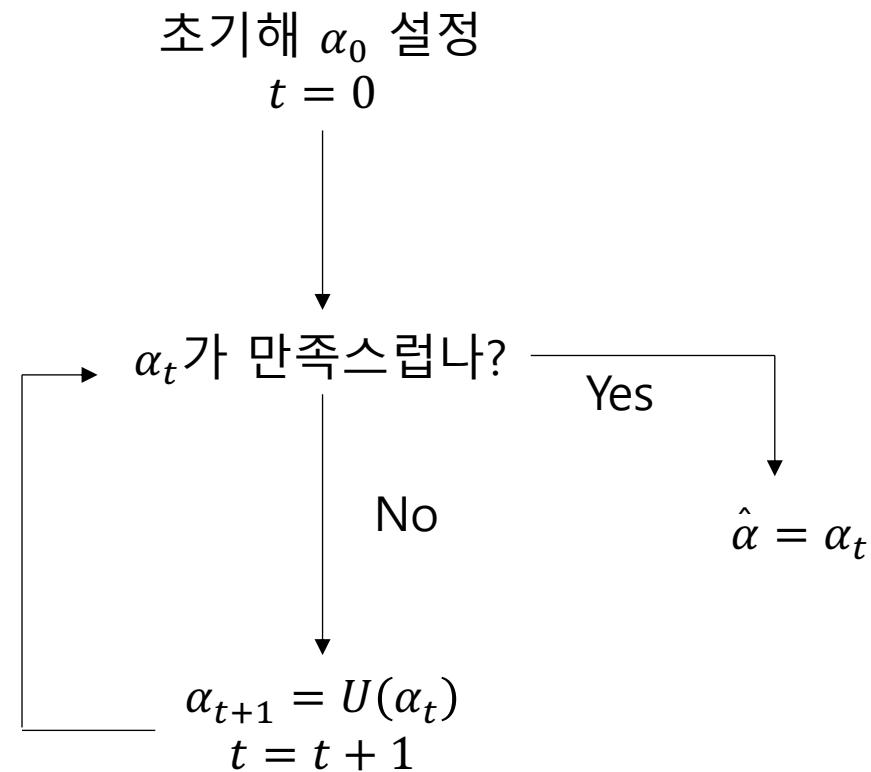
전미분을 하면,

$$dz = f_xdx + f_ydy = (2xy^2 - 6x)dx + (2x^2y)dy \text{ 가 된다.}$$

# 경사하강법 들어가기

- machine learning에서는 매개 변수(parameter, 선형회귀에서는  $\theta_0, \theta_1$ )가 수십~수백 차원의 벡터인 경우가 대부분이다. 또한 목적 함수(선형회귀에서는  $\sum \epsilon_i^2$ )가 모든 구간에서 미분 가능하다는 보장이 항상 있는 것도 아니다.
- 따라서 한 번의 수식 전개로 해를 구할 수 없는 상황이 적지 않게 있다.
- 이런 경우에는 초기 해에서 시작하여 해를 반복적으로 개선해 나가는 수치적 방법을 사용한다. (미분이 사용됨)

# 경사하강법의 개념



# 경사하강법의 정의

- **Gradient Descent**

현재 위치에서 경사가 가장 급하게 하강하는 방향을 찾고, 그 방향으로 약간 이동하여 새로운 위치를 잡는다.  
이러한 과정을 반복함으로써 가장 낮은 지점(즉 최저 점)을 찾아 간다.

- **Gradient Ascent**

- 현재 위치에서 경사가 가장 급하게 상승하는 방향을 찾고, 그 방향으로 약간 이동하여 새로운 위치를 잡는다.  
이러한 과정을 반복함으로써 가장 높은 지점(즉 최대 점)을 찾아 간다.

# 경사하강법 알고리즘

$J$  = 목적함수

$$\alpha_{t+1} = \alpha_t - \rho \frac{\partial J}{\partial \alpha} \Big|_{\alpha_t}$$

$\frac{\partial J}{\partial \alpha} \Big|_{\alpha_t}$  :  $\alpha_t$ 에서의 도함수  $\frac{\partial J}{\partial \alpha}$ 의 값

$\alpha_t$ 에서의 미분값은 음수이다.

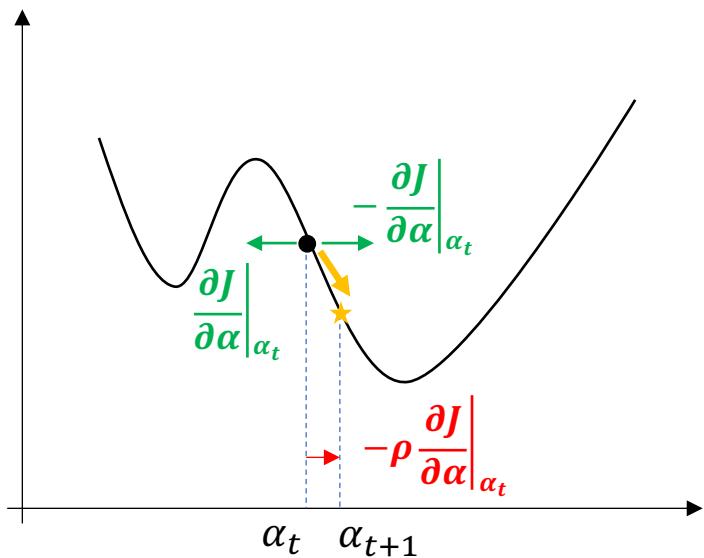
그래서  $\frac{\partial J}{\partial \alpha} \Big|_{\alpha_t}$  를 더하게 되면

왼쪽으로 이동하게 된다.

그러면 목적함수의 값이 증가하는  
방향으로 이동하게 된다.

따라서  $\frac{\partial J}{\partial \alpha} \Big|_{\alpha_t}$  를 빼준다.

그리고 적당한  $\rho$ (스텝크기, 학습률)를 곱해주어서  
조금만 이동하게 한다.



# 경사하강법 알고리즘

$J$  = 목적함수

## Gradient Descent

$\frac{\partial J}{\partial \alpha} \Big|_{\alpha_t}$  :  $\alpha_t$ 에서의 도함수  $\frac{\partial J}{\partial \alpha}$ 의 값

$$\alpha_{t+1} = \alpha_t - \rho \frac{\partial J}{\partial \alpha} \Big|_{\alpha_t}$$

## Gradient Ascent

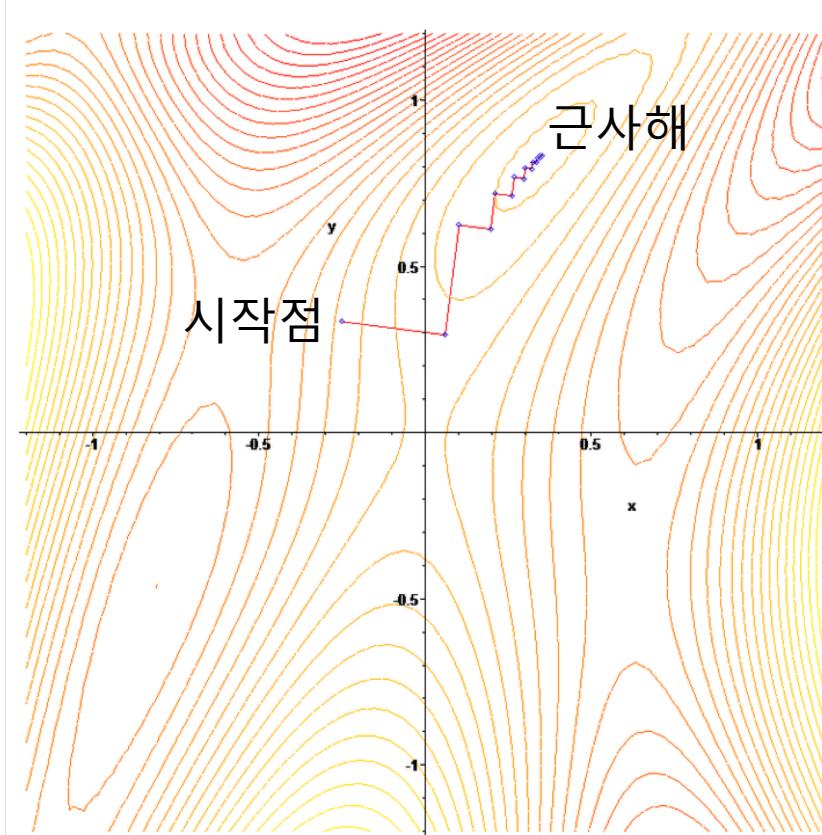
$$\alpha_{t+1} = \alpha_t + \rho \frac{\partial J}{\partial \alpha} \Big|_{\alpha_t}$$

Gradient Descent, Gradient Ascent는 전형적인 Greedy algorithm이다.

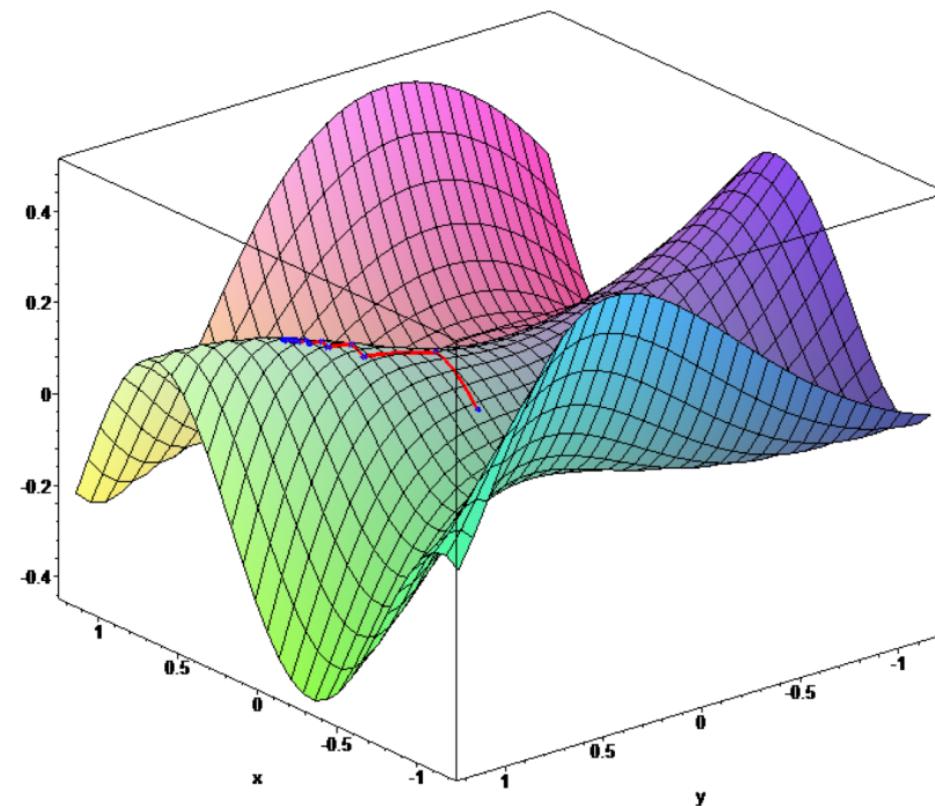
과거 또는 미래를 고려하지 않고 현재 상황에서 가장 유리한 다음 위치를 찾아 Local optimal point로 끝날 가능성을 가진 알고리즘이다.

# 경사하강법 예제1

- Wiki의 예제

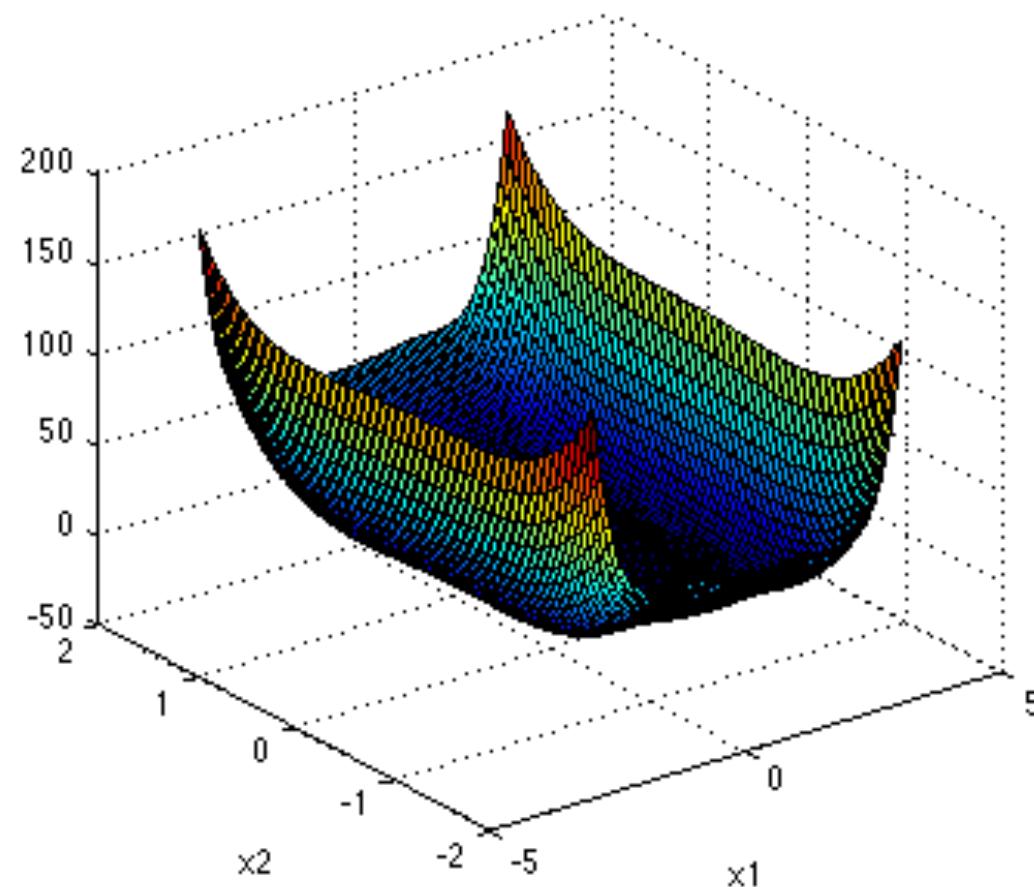


$$F(x, y) = \sin\left(\frac{1}{2}x^2 - \frac{1}{4}y^2 + 3\right) \cos(2x + 1 - e^y).$$



# 경사하강법 예제2

- 낙타 등 함수(six-hump camelback function)



# 경사하강법 예제2

- 낙타 등 함수(six-hump camelback function)

✓  $J(\Theta) = \left(4 - 2.1\theta_1^2 + \frac{\theta_1^4}{3}\right)\theta_1^2 + \theta_1\theta_2 + (-4 + 4\theta_2^2)\theta_2^2$

✓ 초기값을  $\Theta_0 = (-0.5, 0.5)^T$ 로 하고 학습률을  $\rho = 0.01$ 로 하자.

✓  $J'(\Theta) = \frac{\partial J}{\partial \Theta} = \left(\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}\right)^T = (2\theta_1^5 - 8.4\theta_1^3 + 8\theta_1 + \theta_2, 16\theta_2^3 - 8\theta_2 + \theta_1)^T$

- $\Theta_1$ 을 구해보자

①  $\frac{\partial J}{\partial \Theta}|_{\Theta_0} = (-2.5125, -2.5)^T$

②  $\Theta_1 = \Theta_0 - 0.01 \times \frac{\partial J}{\partial \Theta}|_{\Theta_0} = (-0.5, 0.5)^T - 0.01 \times (-2.5125, -2.5)^T = (-0.4748, 0.525)^T$

# 경사하강법 예제2

- 낙타 등 함수(six-hump camelback function)

✓  $J(\Theta) = \left(4 - 2.1\theta_1^2 + \frac{\theta_1^4}{3}\right)\theta_1^2 + \theta_1\theta_2 + (-4 + 4\theta_2^2)\theta_2^2$

✓ 초기값을  $\Theta_0 = (-0.5, 0.5)^T$ 로 하고 학습률을  $\rho = 0.01$ 로 하자.

✓  $J'(\Theta) = \frac{\partial J}{\partial \Theta} = \left(\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}\right)^T = (2\theta_1^5 - 8.4\theta_1^3 + 8\theta_1 + \theta_2, 16\theta_2^3 - 8\theta_2 + \theta_1)^T$

- $\Theta_2$ 을 구해보자!

①  $\frac{\partial J}{\partial \Theta}|_{\Theta_1} = (-2.4228, -2.3596)^T$

②  $\Theta_2 = \Theta_1 - 0.01 \times \frac{\partial J}{\partial \Theta}|_{\Theta_1} = (-0.4748, 0.525)^T - 0.01 \times (-2.4228, -2.3596)^T = (-0.4506, 0.5486)^T$

## 경사하강법 예제2

$$\Theta_0 = (-0.5, 0.5)^T$$

$$\Theta_1 = (-0.4748, 0.525)^T$$

$$\Theta_2 = (-0.4506, 0.5486)^T$$

위의 값을 대입하여  $J(\Theta)$ 를 계산하면 아래와 같다.

$$J(\Theta_0) = -0.12604$$

$$J(\Theta_1) = -0.24906$$

$$J(\Theta_2) = -0.36036$$

# 학습률의 영향

- Gradient에서 알 수 있는 것은 함수값이 가장 빠르게 증가하는 방향이다. 그 방향으로 얼만큼을 가야하는지는 알려주지 않는다. 얼만큼 가야하는지를 의미하는 학습률(learning rate  $\rho$ )는 Gradient를 사용하는 모델을 학습시킬 때 있어 가장 중요한 hyperparameter이다.
- 학습률을 너무 크게 하면 :  
    최저 점을 중심으로 좌우를 왔다갔다하는 진자 현상이 발생한다.
- 학습률을 너무 작게 하면 :  
    수렴 속도가 느려진다.

# 학습 데이터 수에 따른 경사하강법 종류

- Mini-batch gradient descent - MGD
  - 우리가 구하고자 하는 모델의 파라미터를 한 번 업데이트하려고 학습데이터 전체를 계산에 사용하는 것은 낭비가 될 수 있다. 학습데이터의 전체가 아닌 배치(batches)만 이용해서 gradient를 계산하는 것이다.
  - 예를들어 120만개 중에 256개짜리 배치만을 이용하여 gradient를 구하고 파라미터를 업데이트한다.
  - 학습데이터가 서로 상관관계가 있기 때문에 전체 데이터를 보지 않고 배치만 이용하여도 이 방법이 효과적임

# 학습 데이터 수에 따른 경사하강법 종류

- Stochastic gradient descent - SGD
  - 온라인 그라디언트 하강이라고도 한다.
  - Mini-batch gradient descent의 배치 크기가 데이터 한 개 일때 이다. 즉 모델의 파라미터를 계산할 때, 데이터 하나에 대하여 모수를 업데이트한다.
  - 모델의 파라미터를 계산할 때, 행렬 및 벡터의 연산이기 때문에, 한 예제에서 100번 계산하는 것보다, 100개의 예제에서 1번 계산하는게 더 빠르다.
  - 엄밀하게는 데이터 한 개에 대하여, 계산한 후 파라미터를 업데이트 하는 것이 SGD이나, 많은 사람들이 MGD를 의미하면서 SGD라고 부르기도 한다.