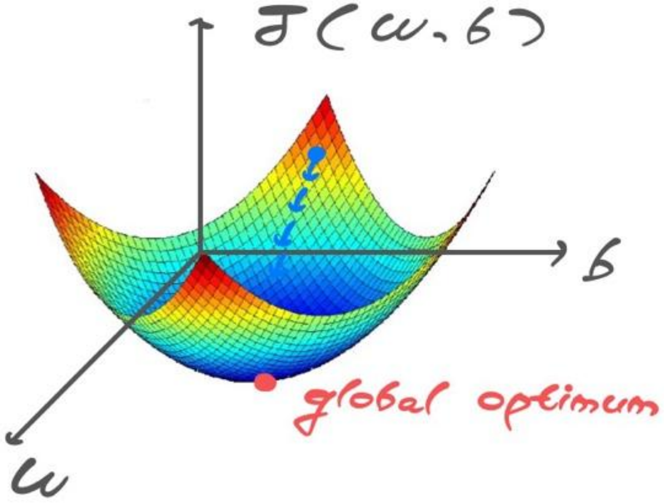
Gradient Descent

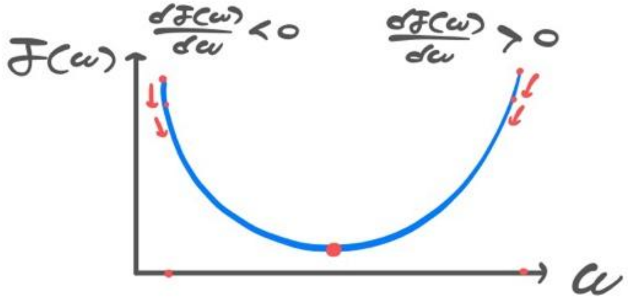
Convex Optimization

수학과 오서영

How to minimize cost function?

Gradient Descent
Want to find W, b that minimize J(W, b)





$$w \overset{\text{update}}{:=} w - \alpha \frac{\partial J(w, b)}{\partial w}$$

$$b := b - \alpha \frac{\partial J(w, b)}{\partial b}$$

α : learning rate

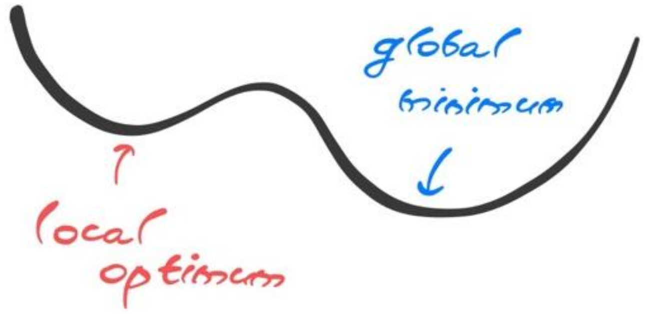
Introduction

Gradient Descent

- Simple and gives some kind of meaningful result for both convex and nonconvex optimization

-> It tries to improve the function value by moving in a direction related to the gradient (i.e first derivative)

1. For convex optimization, it gives the global optimum
2. For nonconvex optimization, it arrives at a local optimum



Convex function

Function f is convex

ie "$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$"

for $\forall x, y$ and $\alpha \in [0, 1]$

Introduction

Taylor expansion of a function all of whose derivatives exist at x

$$f(x+\eta) = f(x) + \eta f'(x) + \frac{\eta^2}{2} f''(x) + \frac{\eta^3}{3!} f'''(x) \cdots$$

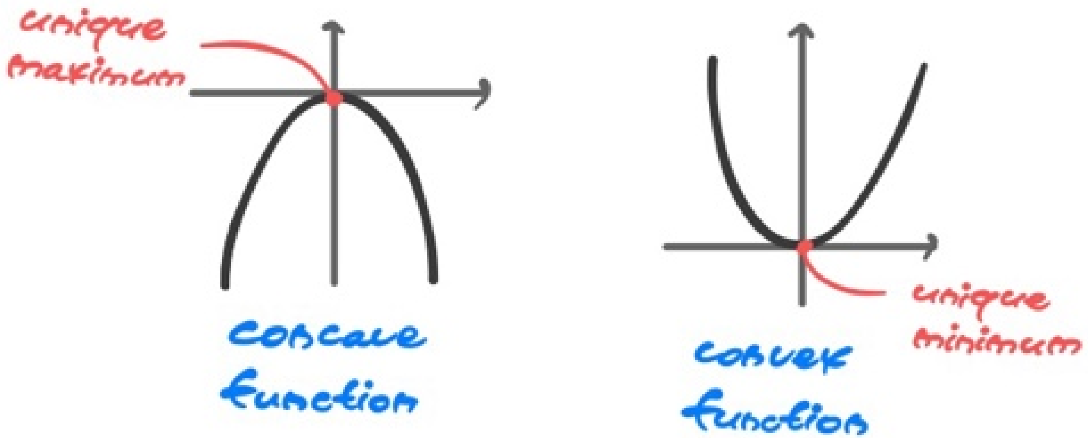
The function is convex

if $f''(x) \geq 0$ for $\forall x$

($\Leftrightarrow$ $f'(x)$ is increasing function of x)

Introduction

The minimum is attained when $f'(x) = 0$
since $f'(x)$ keeps increasing to the left and right of that
thus the global minimum is unique.



Introduction

Gradient Descent

- Given a differentiable function of f(x) and an initial parameter of x1
- Iteratively moving the parameter to the lower value of f(x)
- By taking the direction of the negative gradient of f(x)

Why this works?

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + O(\|x-a\|^2)$$

Assume $a = x_1$ and $x = x_1 + h u$ unit direction vector for the partial deriv

$$f(x_1 + h u) = f(x_1) + h f'(x_1) u + h^2 O(1)$$

$$f(x_1 + h u) - f(x_1) \approx h f'(x_1) u$$

Introduction

$$u^* = \mathrm{argmin}_u \{ f(x_i + h u) - f(x_i) \}$$

$$= \mathrm{argmin}_u \, h f'(x_i) u = - \frac{f'(x_i)}{|f'(x_i)|}$$

$$\left(\because \; f(x_i + h u) \le f(x_i), \quad a \cdot b = |a||b| \cos\alpha \right)$$

$$\text{"} \; x_{t+1} \leftarrow x_t + h u^* = x_t - h \frac{f'(x_i)}{|f'(x_i)|}$$

Introduction

Example

$$f(x_1, x_2) = (1-x_1)^2 + 100(x_2 - x_1^2)^2$$

$$\frac{\partial}{\partial x_1} f(x_1, x_2) = -2(1-x_1) - 400x_1(x_2 - x_1^2)$$

$$\frac{\partial}{\partial x_2} f(x_1, x_2) = 200(x_2 - x_1^2)$$

global minimum = 0
at (1, 1)

initial point $\mathbf{x}^0 = (x_1^0, x_2^0) = (-1.3, 0.9)$

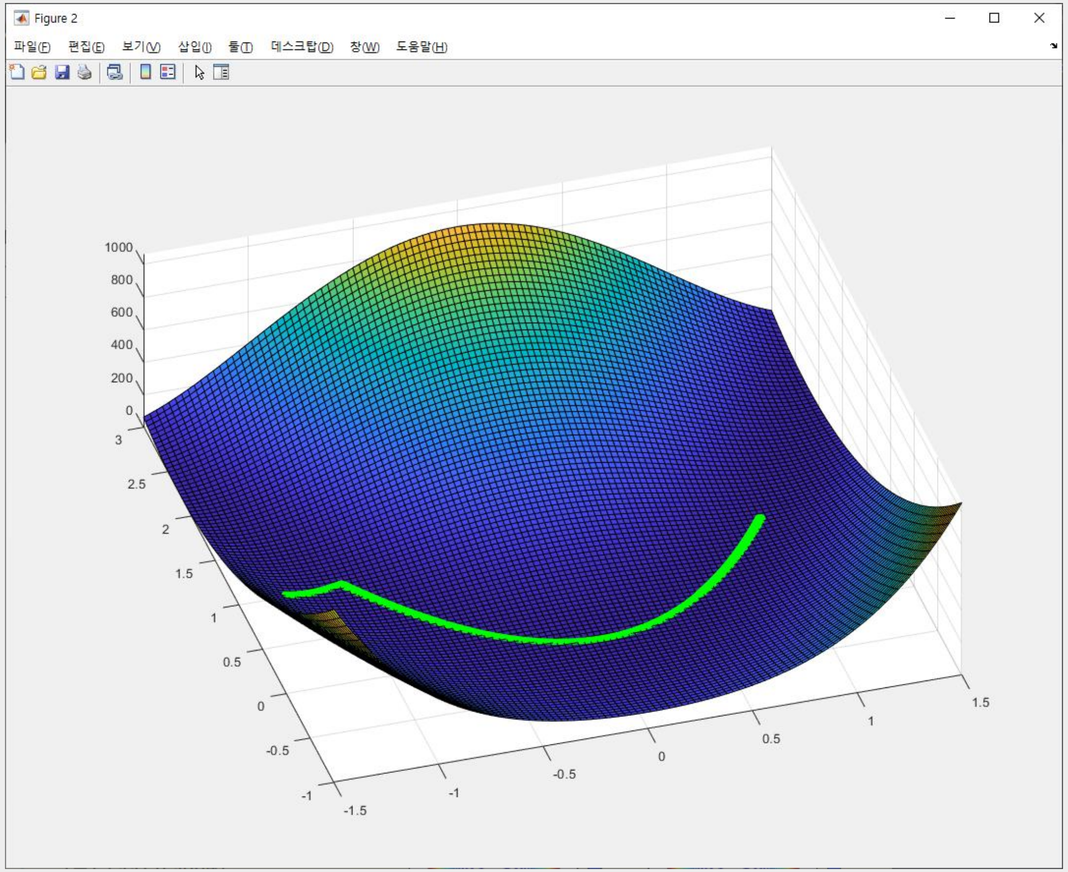
$$f'(\mathbf{x}^0) = \left(\frac{\partial}{\partial x_1} f(x_1, x_2), \frac{\partial}{\partial x_2} f(x_1, x_2)\right)$$

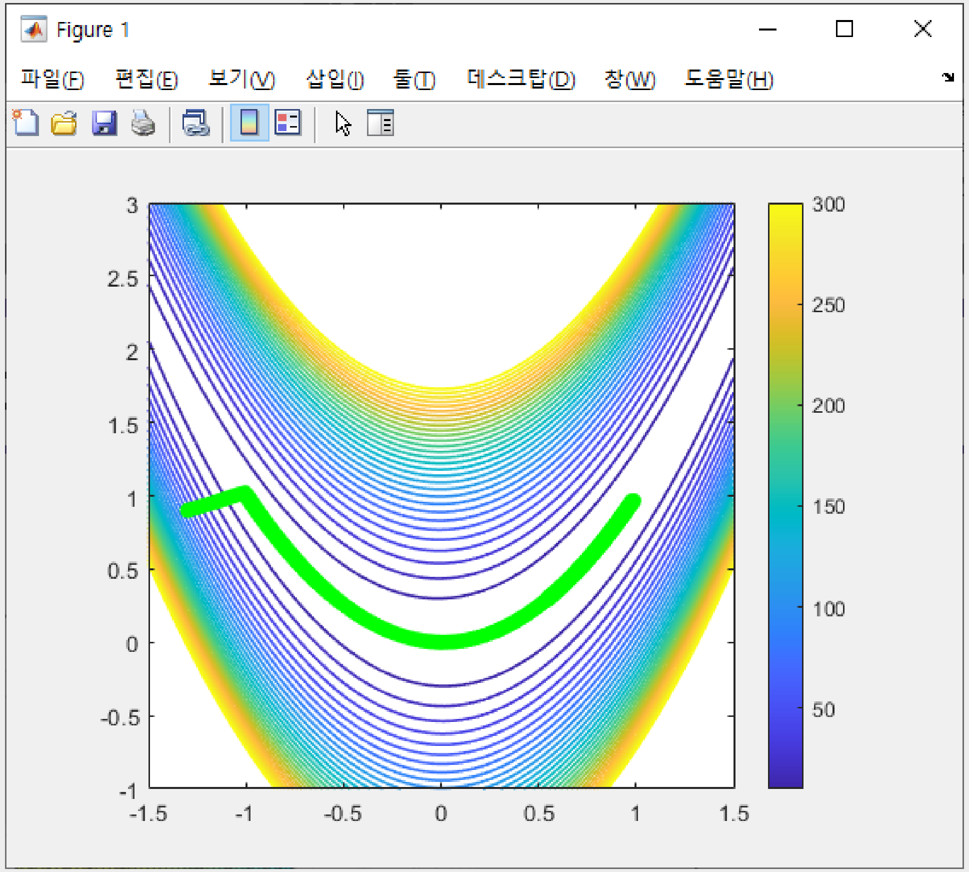
$$= (-415.4, -158)$$

$$\mathbf{x}^1 \leftarrow \mathbf{x}^0 - \sigma \frac{f'(\mathbf{x}^0)}{|f'(\mathbf{x}^0)|}$$

repeat the update until converge

Introduction





1. Gradient descent for convex functions : univariate case

We start at some x_0 and

at step i update x_i to $x_{i+1} = x_i + \eta f'(x)$ for some $\eta < 0$

i.e move in direction where f decreases.

$f(x_{i+1}) = f(x_i) + \eta f'(x_i) + \frac{\eta^2}{2} f''(x_i)$ ignore terms that involve η^3 or higher

Newton's method

Best value of η most reduction in one step

" $\eta = -\frac{f'(x)}{2f''(x)}$ " which gives $f(x_{i+1}) = f(x_i) - \frac{(f'(x_i))^2}{2f''(x_i)}$

$\longrightarrow$ the algorithm makes progress so long as $f''(x_i) > 0$

2. Convex multivariate functions

If higher derivatives also exist, the multivariate Taylor expansion for an n-variate function f is

$$\text{“} f(x+y) = f(x) + \nabla f(x) \cdot y + y^T \underbrace{\nabla^2 f(x)}_{\text{Hessian}} y + \cdots \text{”}$$

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

- f is convex

  iff the Hessian is positive semidefinite

  (i.e $y^T \nabla^2 f y \geq 0$ for $\forall y$)

3. Gradient Descent for Constrained Optimization

- Constrained optimization consists of solving the following where K is a convex set and f is convex function.

  " $\min f(x)$ s.t $x \in K$ "

Lagrange Multiplier

new optimization problem

$$\min f(x) \text{ s.t } g(x) = c$$

$$\exists \alpha \in \mathbb{R} \text{ s.t } \min f(x) + \alpha (g(x) - c)$$

References

[1] Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.