VI. Optimization

- Notation
  - 예를 들어 $L(\boldsymbol{x}) = \|\boldsymbol{b} - A\boldsymbol{x}\|^2$ 과 같은 loss function이 주어졌을 때, $L(\boldsymbol{x})$의 값을 최소로 만드는 입력값이 $\hat{\boldsymbol{x}}$일 때 $\operatorname{argmin} L(\boldsymbol{x}) = \hat{\boldsymbol{x}}$과 같이 표현함
- 단일 변수 함수의 테일러 급수 (Taylor series)
  - 변수 x 에 대한 임의의 함수 $f(x)$ 는 다항식으로 근사 가능함
    - $f(x) \approx c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots$
  - 다항식의 계수 $c_0, c_1, c_2, c_3, \ldots$는 어떻게 결정?
  - 위의 다항식으로 전개한 결과가 $f(x)$와 같다면,
    - $x = 0$ 을 등식에 대입했을 때, 양변이 같아야 함 ➔ $f(0) = c_0$
    - $x = 0$ 을 등식에 대입했을 때, 양변의 1차 미분이 같아야 함 ➔ $\left.\frac{df}{dx}\right|_{x=0} = c_1$
    - $x = 0$ 을 등식에 대입했을 때, 양변의 2차 미분이 같아야 함 ➔ $\left.\frac{d^2 f}{dx^2}\right|_{x=0} = c_2 \cdot 2$
    - $x = 0$ 을 등식에 대입했을 때, 양변의 3차 미분이 같아야 함 ➔ $\left.\frac{d^3 f}{dx^3}\right|_{x=0} = c_3 \cdot 3 \cdot 2$
    - 일반적으로 $c_n = \left.\frac{1}{n!}\frac{d^n f}{dx^n}\right|_{x=0}$
  - $f(x) \approx f(0) + \left.\frac{df}{dx}\right|_{x=0} x + \left.\frac{1}{2!}\frac{d^2 f}{dx^2}\right|_{x=0} x^2 + \left.\frac{1}{3!}\frac{d^3 f}{dx^3}\right|_{x=0} x^3 + \cdots$

VI. Optimization

- (계속) 단일 변수 함수의 테일러 급수 (Taylor series)
  - $f(x) \approx f(0) + \left.\frac{df}{dx}\right|_{x=0} x + \frac{1}{2!}\left.\frac{d^2 f}{dx^2}\right|_{x=0} x^2 + \frac{1}{3!}\left.\frac{d^3 f}{dx^3}\right|_{x=0} x^3 + \cdots$
  - 위의 결과는 $f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \cdots$의 형태를 가정하고 유도하였으므로, $x = 0$ 근처에서는 정확도가 높지만 0에서 먼 위치에 대해서는 정확도가 떨어짐
  - 만약 x가 0이 아닌 위치(x_0) 근처에서 정확도가 높은 다항식을 원한다면 아래와 같은 전개 가능
    - $f(x) \approx d_0 + d_1(x - x_0) + d_2(x - x_0)^2 + d_3(x - x_0)^3 + \cdots$
  - 위의 다항식으로 전개한 결과가 $f(x)$와 같다면,
    - $x = x_0$ 을 등식에 대입했을 때, 양변이 같아야 함 ➔ $f(x_0) = d_0$
    - $x = x_0$ 을 등식에 대입했을 때, 양변의 1차 미분이 같아야 함 ➔ $\left.\frac{df}{dx}\right|_{x=x_0} = d_1$
    - $x = x_0$을 등식에 대입했을 때, 양변의 2차 미분이 같아야 함 ➔ $\left.\frac{d^2 f}{dx^2}\right|_{x=x_0} = d_2 \cdot 2$
    - 일반적으로 $d_n = \frac{1}{n!}\left.\frac{d^n f}{dx^n}\right|_{x=x_0}$
  - $f(x) \approx f(x_0) + \left.\frac{df}{dx}\right|_{x=x_0}(x - x_0) + \frac{1}{2!}\left.\frac{d^2 f}{dx^2}\right|_{x=x_0}(x - x_0)^2 + \frac{1}{3!}\left.\frac{d^3 f}{dx^3}\right|_{x=x_0}(x - x_0)^3 + \cdots$
  - Example) $f(x) = 1 + 2x + 3x^2$를 $x = 1$ 근처에서의 다항식으로 나타내면?
    - $d_0 + d_1(x - 1) + d_2(x - 1)^2 = 6 + 8(x - 1) + \frac{1}{2!}6(x - 1)^2$
    - 만약 $f(x)$가 3차 이상의 식이었는데, 2차에서 멈췄다면 근사의 정확도가 떨어짐

VI. Optimization

- 다 변수 함수(multivariable function)의 테일러 급수 (Taylor series)
  - $f(x,y) \approx c_0 + c_x x + c_y y + c_{xx} x^2 + c_{xy} xy + c_{yy} y^2 + \cdots$
  - 위의 다항식으로 전개한 결과가 $f(x,y)$와 같다면,
    - $x = y = 0$을 등식에 대입했을 때, 양변이 같아야 함 ➔ $f(0,0) = c_0$
    - $x = y = 0$을 등식에 대입했을 때, 양변을 x에 대해 1차 편미분한 결과가 같아야 함 ➔ $\left.\frac{\partial f}{\partial x}\right|_{x=y=0} = c_x$
    - $x = y = 0$을 등식에 대입했을 때, 양변을 x에 대해 2차 편미분한 결과가 같아야 함 ➔ $\left.\frac{\partial^2 f}{\partial x^2}\right|_{x=y=0} = c_{xx} \cdot 2$
    - $x = y = 0$을 등식에 대입했을 때, 양변을 x에 대해 1차 편미분한 후 y에 대해 1차 편미분한 결과가 같아야 함 ➔ $\left.\frac{\partial^2 f}{\partial y \partial x}\right|_{x=y=0} = c_{xy}$
    - 일반적으로 $c_{x^m y^{(n-m)}} = \frac{1}{m!(n-m)!} \left.\frac{\partial^n f}{\partial x^m \partial y^{(n-m)}}\right|_{x=y=0} = \frac{1}{n!}\binom{n}{m} \left.\frac{\partial^n f}{\partial x^m \partial y^{(n-m)}}\right|_{x=y=0} = \frac{1}{n!}\binom{n}{m}\left(\frac{\partial}{\partial x}\right)^m \left(\frac{\partial}{\partial y}\right)^{n-m} f\Big|_{x=y=0}$
  - $f(x,y) \approx f(0,0) + \left.\frac{\partial f}{\partial x}\right|_{x=y=0} x + \left.\frac{\partial f}{\partial y}\right|_{x=y=0} y + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial x^2}\right|_{x=y=0} x^2 + \left.\frac{\partial^2 f}{\partial y \partial x}\right|_{x=y=0} xy + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial y^2}\right|_{x=y=0} y^2 + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!}\left(x\left(\frac{\partial}{\partial x}\right) + y\left(\frac{\partial}{\partial y}\right)\right)^n f\Big|_{x=y=0}$

VI. Optimization

- 다 변수 함수(multivariable function)의 테일러 급수 (Taylor series)
  - $f(x,y) \approx f(0,0) + \left.\frac{\partial f}{\partial x}\right|_{x=y=0} x + \left.\frac{\partial f}{\partial y}\right|_{x=y=0} y + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial x^2}\right|_{x=y=0} x^2 + \left.\frac{\partial^2 f}{\partial y \partial x}\right|_{x=y=0} xy + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial y^2}\right|_{x=y=0} y^2 + \cdots$
  - 단일 변수 함수와 마찬가지로 임의의 위치 (x_0, y_0)에서 위의 전개를 하면,
    - $f(x,y) \approx f(x_0,y_0) + \left.\frac{\partial f}{\partial x}\right|_{x_0,y_0} (x - x_0) + \left.\frac{\partial f}{\partial y}\right|_{x_0,y_0} (y - y_0) + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial x^2}\right|_{x_0,y_0} (x - x_0)^2 + \left.\frac{\partial^2 f}{\partial y \partial x}\right|_{x_0,y_0} (x - x_0)(y - y_0) + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial y^2}\right|_{x_0,y_0} (y - y_0)^2 + \cdots$
  - $x - x_0 \equiv \Delta x, y - y_0 \equiv \Delta y$로 표현하고, 이들 차이값을 성분으로 하는 벡터 $\Delta \boldsymbol{x} \equiv \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}$를 정의하면, 1차 다항식 항들은 아래와 같이 표현 가능
    - $\left.\frac{\partial f}{\partial x}\right|_{x_0,y_0} (x - x_0) + \left.\frac{\partial f}{\partial y}\right|_{x_0,y_0} (y - y_0) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}_{x_0,y_0} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = (\nabla f)^T \Delta \boldsymbol{x}$

VI. Optimization

- 다 변수 함수(multivariable function)의 테일러 급수 (Taylor series)
  - $f(x,y) \approx f(x_0,y_0) + \left.\frac{\partial f}{\partial x}\right|_{x_0,y_0}(x-x_0) + \left.\frac{\partial f}{\partial y}\right|_{x_0,y_0}(y-y_0) + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial x^2}\right|_{x_0,y_0}(x-x_0)^2 + \left.\frac{\partial^2 f}{\partial y \partial x}\right|_{x_0,y_0}(x-x_0)(y-y_0) + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial y^2}\right|_{x_0,y_0}(y-y_0)^2 + \cdots$
  - 위 전개의 2차 다항식 항들은 아래와 같이 표현 가능
    - $\frac{1}{2!}\left.\frac{\partial^2 f}{\partial x^2}\right|_{x_0,y_0}\Delta x^2 + \left.\frac{\partial^2 f}{\partial y \partial x}\right|_{x_0,y_0}\Delta x \Delta y + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial y^2}\right|_{x_0,y_0}\Delta y^2 = \frac{1}{2}[\Delta x \quad \Delta y]\begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}_{x_0,y_0}\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \frac{1}{2}(\Delta \boldsymbol{x})^T H(\Delta \boldsymbol{x})$
  - H는 Hessian 행렬이라고 부르고, 주어진 스칼라 함수의 2차 편미분들을 성분으로 가지는 대칭 행렬임
  - 위의 식을 일반화하여 정리해 보면,
    - $f(\boldsymbol{x}_0 + \Delta \boldsymbol{x}) \approx f(\boldsymbol{x}_0) + {(\nabla f)_{\boldsymbol{x}_0}}^T \Delta \boldsymbol{x} + \frac{1}{2}(\Delta \boldsymbol{x})^T H_{\boldsymbol{x}_0}(\Delta \boldsymbol{x}) + \cdots$

VI. Optimization

- 벡터 함수
  - 벡터의 각각의 성분들이 (독립적인) 함수들로 이루어져 있을 때 이를 벡터 함수라 부름
    - $\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix}$, Ex) $\boldsymbol{f}(x, y, z) = \begin{bmatrix} 2x + 3y \\ 1 \\ y \cdot z \\ 0 \end{bmatrix}$
  - 위의 예와 같이 벡터 함수의 성분의 개수($m = 4$)와 각각의 성분 함수들이 가지는 변수의 개수 (n =3)는 독립적임
- 벡터 함수의 Jacobian 행렬
  - 벡터 함수의 각 성분들의 gradient를 구했을 때, 이들을 행벡터로 가지는 행렬을 주어진 벡터 함수의 Jacobian 행렬이라고 부름
  - $\boldsymbol{J} = \begin{bmatrix} (\nabla f_1)^T \\ (\nabla f_2)^T \\ \vdots \\ (\nabla f_m)^T \end{bmatrix} = \begin{bmatrix} \frac{\partial \boldsymbol{f}}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$, Ex) $\boldsymbol{J} = \begin{bmatrix} 2 & 3 & 0 \\ 0 & 0 & 0 \\ 0 & z & y \\ 0 & 0 & 0 \end{bmatrix}$
  - Hessian 행렬 ($H_{ij} = \partial^2 f / \partial x_i \partial x_j$)은 스칼라 함수의 gradient 인 $\nabla f = \boldsymbol{F}$ 벡터 함수의 Jacobian 행렬임

VI. Optimization

- Gradient 벡터의 의미
  - 다 변수 함수 $f(x,y)$ 에 대해, $f(x_0+\Delta x, y_0+\Delta y)$를 Taylor급수로 1차항까지만 근사시키면 아래와 같이 표현

$$f(x_0+\Delta x, y_0+\Delta y) \approx f(x_0,y_0) + \left(\frac{\partial f}{\partial x}\right)\Delta x + \left(\frac{\partial f}{\partial y}\right)\Delta y = f(x_0,y_0) + (\nabla f)^T \Delta \boldsymbol{x}$$

  - 따라서 $(\nabla f)^T \Delta \boldsymbol{x}$는 현재 위치 (x_0,y_0) 에서 $(\Delta x, \Delta y)$의 방향으로 움직였을 때 함수 값의 변화를 나타냄
  - 따라서 $\nabla f = 0$ 이라는 조건은 단순히 x축 방향 또는 y축 방향으로 움직였을 때 기울기의 변화가 0이라는 의미가 아니라 xy평면내 임의의 방향으로 움직였을 때 변화가 0이라는 의미임
  - 거꾸로 $\nabla f \neq 0$ 임에도 불구하고 $(\nabla f)^T \Delta \boldsymbol{x} = 0$ 이라는 조건의 의미?
    - $(\Delta x, \Delta y)$의 방향이 등고선(또는 등치선 contour)을 따라 움직인다는 것을 의미
    - 거꾸로 gradient 벡터는 등고선과 직교하는 방향을 가리키고 있음을 의미
  - 움직인 거리 $\sqrt{\Delta x^2 + \Delta y^2}$ 가 같을 때 (즉 벡터 $(\Delta x, \Delta y)$의 크기가 같을 때), 함수 값의 변화 ($(\nabla f)^T \Delta \boldsymbol{x}$)가 최대가 되려면, gradient 벡터 ∇f와 위치 변화 벡터 $(\Delta x, \Delta y)$의 방향이 평행해야 함

  ➔ steepest gradient

VI.1 Minimum Problems

- 함수의 최소값 판별
  - 임의의 함수가 주어졌을 때, 다음과 같이 $\boldsymbol{x}_0 = (x_0, y_0)$ 를 중심으로 $\Delta\boldsymbol{x} = (\Delta x, \Delta y)$의 2차항까지 Taylor급수로 근사 가능
    - $f(\boldsymbol{x}_0 + \Delta\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + (\nabla f)_{\boldsymbol{x}_0}{}^T \Delta\boldsymbol{x} + \frac{1}{2}(\Delta\boldsymbol{x})^T H_{\boldsymbol{x}_0}(\Delta\boldsymbol{x})$
  - 만약 $\boldsymbol{x}_0 = (x_0, y_0)$ 에서 gradient가 0이 되면 모든 방향으로의 기울기가 0이므로 최소값일 가능성이 존재함
    - $f(\boldsymbol{x}_0 + \Delta\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \frac{1}{2}(\Delta\boldsymbol{x})^T H_{\boldsymbol{x}_0}(\Delta\boldsymbol{x})$
  - 이 함수가 $\boldsymbol{x}_0 = (x_0, y_0)$ 에서 최소값을 가지려면, $\frac{1}{2}(\Delta\boldsymbol{x})^T H_{\boldsymbol{x}_0}(\Delta\boldsymbol{x})$가 $\Delta\boldsymbol{x} = \boldsymbol{0}$ 이외의 지점에서는 항상 0보다 커야 함
  - 주어진 함수가 Taylor급수로 근사 가능한 함수인 경우에는 Hessian 행렬인 $H_{\boldsymbol{x}_0}$은 대칭 행렬이 되어야 함
  - $\Delta\boldsymbol{x} \neq \boldsymbol{0}$ 인 경우 $\frac{1}{2}(\Delta\boldsymbol{x})^T H_{\boldsymbol{x}_0}(\Delta\boldsymbol{x}) > 0$를 만족하기 위해서는 Hessian 행렬의 모든 eigenvalue들이 0보다 커야 함

VI.1 Minimum Problems

- Example: $f(x,y) = x^4 + x^2 + xy + y^2$의 함수가 주어졌을 때, 원점에서 최소값을 가지는지 확인하시오.
  - 원점에서의 gradient는 0
  - 2차항까지 Taylor급수로 구하면, $f(x,y) \approx x^2 + xy + y^2$
  - Hessian 행렬: $\begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}_{x_0=y_0=0} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$
  - Eigenvalue: $\lambda = 1, 3$
  - 따라서 $f(x,y)$는 원점에서 최소값을 가짐

VI.1 Minimum Problems

- Newton's method
  - 임의의 함수 f가 주어졌을 때, 특정한 위치 $\boldsymbol{x}_k = (x_k, y_k)$ 를 중심으로 $\Delta\boldsymbol{x} = (\Delta x, \Delta y)$의 2차항까지 Taylor급수로 근사했다고 가정
$$f(\boldsymbol{x}_k + \Delta\boldsymbol{x}) \approx f(\boldsymbol{x}_k) + (\nabla f)_{\boldsymbol{x}_k}{}^T \Delta\boldsymbol{x} + \frac{1}{2}(\Delta\boldsymbol{x})^T H_{\boldsymbol{x}_k}(\Delta\boldsymbol{x})$$
  - $\boldsymbol{x}_k = (x_k, y_k)$는 함수 f를 최소값을 만드는 위치가 아니므로 $(\nabla f)_{\boldsymbol{x}_k}$는 0이 아님
  - 상수값 $f(\boldsymbol{x}_k)$를 제외한 나머지를 $L(\boldsymbol{x})$로 표기
$$L(\Delta\boldsymbol{x}) \equiv \frac{1}{2}(\Delta\boldsymbol{x})^T H_{\boldsymbol{x}_k}(\Delta\boldsymbol{x}) + (\nabla f)_{\boldsymbol{x}_k}{}^T \Delta\boldsymbol{x}$$
  - $L(\Delta\boldsymbol{x})$이 위와 같은 형태로 주어졌을 때, $L(\Delta\boldsymbol{x})$가 최소가 되면 $L(\Delta\boldsymbol{x})$의 gradient 는 0이 된다는 점을 이용하면
  - $\nabla\big(L(\Delta\boldsymbol{x})\big) \equiv H_{\boldsymbol{x}_k}(\Delta\boldsymbol{x}) + (\nabla f)_{\boldsymbol{x}_k} = \boldsymbol{0}$ ➔ $\Delta\boldsymbol{x} = -H_{\boldsymbol{x}_k}^{-1}(\nabla f)_{\boldsymbol{x}_k}$
  - 따라서 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta\boldsymbol{x} = \boldsymbol{x}_k - H_{\boldsymbol{x}_k}^{-1}(\nabla f)_{\boldsymbol{x}_k}$

VI.1 Minimum Problems

- (계속) Newton's method
  - $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta\boldsymbol{x} = \boldsymbol{x}_k - H_{\boldsymbol{x}_k}^{-1}(\nabla f)_{\boldsymbol{x}_k}$
  - 장점: 수렴 속도가 빠름
  - 단점: Hessian 행렬까지 계산이 필요하므로 계산에 많은 시간이 소요되고, activation 함수의 경우 2차 미분의 계산이 어려운 경우도 많음
  - Example) $y = x^2 + 2x$
    - $x = 0$ 근처에서 Taylor전개
    - 1차 미분만으로는 수렴에 한계를 가짐
    - 2차 미분은 한번에 최소점을 찾을 수 있음.
    - 하지만 원래 함수가 2차 이상의 경우에는 여러 번 반복 필요

VI.1 Minimum Problems

- Gradient descent
  - $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta\boldsymbol{x} = \boldsymbol{x}_k - s_k(\nabla f)_{\boldsymbol{x}_k}$
  - 장점: Newton방법에 비해 1차 미분만으로 해결 가능
  - 단점: Newton방법에 비해 수렴 속도가 느림
  - Example) $F(x, y) = x^2 + 4y^2$
    - Newton 방법과 $\nabla F(x, y)$의 방향과 크기 비교
  - s_k는 learning rate 또는 step size로 불리고, 일반적으로 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - s_k(\nabla f)_{\boldsymbol{x}_k}$ 직선을 따라 움직일 때 F를 최소로 하는 값을 선택
  - Zigzag 형태로 수렴이 느림