Taylor series

미적분학에서, **테일러 급수**(Taylor級數, 영어: Taylor series)는 도함수들의 한 점에서의 값으로 계산된 항의 무한 합으로 해석함수를 나타내는 방법이다.

매끄러운 함수 $f\colon \mathbb{R} \to \mathbb{R}$ 및 실수 $a \in \mathbb{R}$ (또는 정칙 함수 $f\colon \mathbb{C} \to \mathbb{C}$ 및 복소수 $a \in \mathbb{C}$)가 주어졌을 때, f의 **테일러 급수**는 다음과 같은 멱급수이다.

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{3!}f^{(3)}(a)(x-a)^3 + \cdots$$

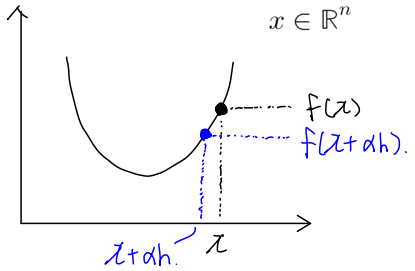
$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^{(n)}$$

다변수 테일러 급수도 마찬가지

$$optimization\ problem\ given\ f:\ \mathbb{R}^n \to \mathbb{R}$$

$$find\ x^* = argmin_{x \in \mathbb{R}^n} f(x)$$

Descent direction



alpha 는 step size, h= descent direction

Taylor series 에 의해

나머지항

$$f(x+\alpha h) = f(x) + \alpha \nabla f(x) h + O(\alpha^2), \alpha > 0$$

적당히 작은 alpha 에 대하여

$$f(x+\alpha h) \approx f(x) + \alpha \nabla f(x) h$$

$$if\ \ \nabla f(x) h < 0 \quad \longrightarrow \quad f(x+\alpha h) < f(x)$$

Def (1) vector h 가 $h^T \nabla f(x) < 0$ 를 만족하면 vector h 를 x 에서의 descent direction 이라 한다.

그렇다면 어떤 descent direction 이 가장 많이 감소하게 하는가?

임의의 unit vector h 에 대해 f(x) 의 방향 기울기 h를 directional derivative 생각

directional derivative 의 엄밀한 정의

단위 벡터 $v \in \mathbb{R}^n \ (\|v\| = 1)$, 점 $x_0 \in \mathbb{R}^n$ 에 대해 함수 $f : \mathbb{R}^n \to \mathbb{R}$ 의 x_0 에서의 v 방향으로의 방향미분(directional derivative) $D_v f(x)$ 는 다음과 같이 주어진 값이다.

$$D_v f(x) = \frac{d}{dt} f(x + tv) \mid_{t=0}$$

방향미분 정의에서 나오는 함수 $f(x+tv)$ 는 곡선 $c(t) : \mathbb{R} \to \mathbb{R}^n$, $c(t) = x + tv$ 와 함수 f 의 합성 함수 $h(t) = f(c(t))$ 이므로 연쇄법칙에 의하여 $\frac{dh}{dt} = D(f)c'(t) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}\right) \cdot v$ 가 된다.

식을 좀더 간단히 하기위해 다음 정의를 도입

기울기 벡터 정의

함수 $f : \mathbb{R}^n \to \mathbb{R}$ 이 미분가능하면 f 의 $(x_1, \cdots, x_n)$ 에서의 기울기 벡터(gradient)란 $\mathbb{R}^n$의 벡터

$$grad\ f = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right)$$

를 뜻한다.

$\nabla f\ or\ \nabla f(x_1, \cdots, x_n)$ 로 표기한다. 기울기 벡터는 $\mathbb{R}^n \to \mathbb{R}^n$ 으로 보내는 함수로써 vector field 가 된다.

기초정리

함수 $f : \mathbb{R}^n \to \mathbb{R}$ 이 미분가능하면 임의의 단위벡터 $v \in \mathbb{R}^n$, $(v = (v_1, v_2, \cdots, v_n))$ 에 대해 f 의 x 에서 v 방향으로의 방향미분은 다음과 같다.

$$D_v(f) = grad\ f(x) \cdot v = \nabla f(x) \cdot v$$
$$= \left(\frac{\partial f}{\partial x_1}\right) v_1 + \left(\frac{\partial f}{\partial x_2}\right) v_2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right) v_n$$

그렇다면 어떤 descent direction 이 가장 많이 감소하게 하는가?

임의의 unit vector h 에 대해 f(x) 의 방향 기울기 h를 directional derivative 생각

$f(x+\alpha h)$ 가 가장 작은 값을 가지게 하려면 $f(x+\alpha h) \approx f(x) + \alpha \nabla f(x) h$

$$if \quad \nabla f(x) h < 0 \quad \longrightarrow \quad f(x+\alpha h) < f(x)$$

$$\min_{h, h^T h = 1} h^T \nabla f(x) = \min_{h, h^T h = 1} \|h\| \, \|\nabla f(x)\| \, cos\theta$$

$$= \min_{h, h^T h = 1} \|\nabla f(x)\| \, cos\theta$$

h와 관련 없는 항 제거 $$= \min_{h, h^T h = 1} cos\theta \qquad \theta = 180^\circ \rightarrow -1$$

즉 h vector 와 f(x) 사이의 각도가 180 도면 된다.

$$x + \alpha h = x - \alpha \nabla f(x)$$

$\alpha = \epsilon, \; h = g$ 라 하자.

Hessian 은 최적의 learning rate를 결정함

W0점에서의 테일러시리즈로 함수를 근사할 때

$$L(w_1) = L(w_0 - \epsilon g) = L(w_0) + g^T(w_0 - \epsilon g - w_0) + \frac{1}{2}(w_0 - \epsilon g - w_0)^T H(w_0 - \epsilon g - w_0) + O(\epsilon^2)$$
$$= L(w_0) - \epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g + O(\epsilon^2)$$

$$w, g \in \mathbb{R}^n \ , \ \epsilon \in \mathbb{R} \qquad \epsilon > 0$$
$$g = \nabla L(w_0) \ , \ H = hessian\ matrix\ of\ L(w_0)$$

만약 목적함수 L이 테일러 급수 2차항까지 잘 근사된다면,

$$L(w_1) = L(w_0 - \epsilon g) \approx L(w_0) - \epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g$$

$L(w_1) < L(w_0)$ 여야 하므로 $\quad -\epsilon g^T g + \frac{1}{2}\epsilon^2 g^T H g < 0$

만약 H가 negative definite 이면 step size가 클수록 L(w1)이 감소를 의미
반대로 H 가 positive definite 이면 모든 term 이 양수이므로 둘의 차이를 최대화

$$\max_{\epsilon}(L(w_0) - L(w_1)) = \max_{\epsilon}\left(\epsilon g^T g - \frac{1}{2}\epsilon^2 g^T H g\right)$$

ϵ에 대한 함수가 quadratic form이므로

$$\frac{\partial(L(w_0) - L(w_1))}{\partial \epsilon} = 0 = g^T g - \epsilon g^T H g \qquad \epsilon^* = \frac{g^T g}{g^T H g}$$

$$H = Q \wedge Q^T, \ \wedge = \begin{pmatrix} \lambda_{max} & \cdots & \\ & & \lambda_{min} \end{pmatrix}$$

최악의 경우 람다 max 에 대응하는 eigen vector 일 때

$$\epsilon^* = \frac{1}{\lambda_{max}}$$

다른 방향의 경우도 hessian의 eigen vector들의 선형결합으로 이루어진 방향임.

Newton’s method

목적함수가 이차함수로 잘 근사된다고 가정했을 때, h 가 충분히 직을 때

$$L(w_1) = L(w_0 + h) \approx L(w_0) + \nabla L(w_0)^T (w_0 + h - w_0) + \frac{1}{2}(w_0 + h - w_0)^T H(w_0 + h - w_0)$$

$$L(w_1) = L(w_0 + h) \approx L(w_0) + \nabla L(w_0)^T h + \frac{1}{2} h^T H h$$

L(w1)을 미분했을 때 기울기가 0인 곳을 한 방에 찾아가 보려고

$$L(h) = L(w_0) + \nabla L(w_0)^T h + \frac{1}{2} h^T H h$$

$$\frac{\partial L(h)}{\partial h} = 0 = \nabla L(w_0) + Hh \qquad\qquad Hh = -\nabla L(w_0)$$

이때 H의 inverse matrix가 존재하면 $h^* = -H^{-1} \nabla L(w_0)$