Subject A21 Predoc2 Date March 2nd 2022

Attendants

Part II

$$\theta^* \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x_i), y_i)$$

where $f_i(\theta) := l(f_\theta(x_i), y_i)$

$$f(\theta) := \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

$$\begin{cases} \theta_0 \in \mathbb{R}^d \\ \theta_{t+1} = \theta_t - \underbrace{\eta \nabla f_{i_t}(\theta_t)}_{\text{stochastic update}} \end{cases}$$

$\boxed{1.}$ In general, is θ^* unique?

It's unconstrained optimization problem

$\forall x \in \mathbb{R}^n$ に対し, $f(x^*) < f(x)$

x^* は global optimizer

$|x^* - x| < \varepsilon$ $\forall x \in \mathbb{R}^n$ に対し $f(x^*) < f(x)$

x^* は local optimizer

Subject

Date 2

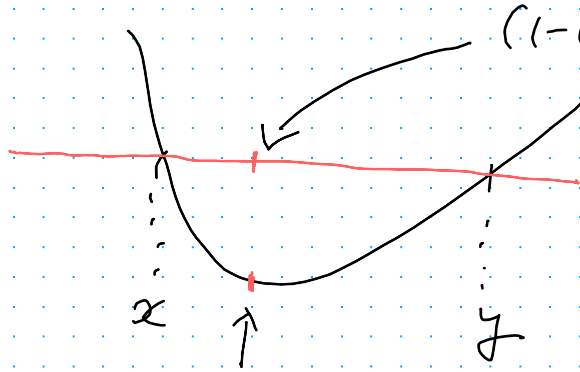
Attendants

Definition of Convex function

$f : \mathbb{R}^n \to (-\infty, \infty)$ $\quad \forall x, y \in \mathbb{R}^n \quad \alpha \in [0, 1]$

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

$(1-\alpha)f(x) + \alpha f(y)$



$f((1-\alpha)x + \alpha y)$

if

objective function is convex and

constraint function is also convex

convex problem

↓

local solution = global solution

Subject Date 3

Attendants

How to identify f is convex

$\Rightarrow$ calc Hessian

if $f(x) = x^2$

$$f''(x) = \frac{\partial}{\partial x} f'(x) = \frac{\partial}{\partial x}(2x) = 2 \geq 0$$

positive semidefinite

Iterative Algorithm

search direction

$$x^{k+1} = x^k + \alpha^k d^k$$

stepsize, a.k.a learning rate

How to find stepsize

1. line search (直線探索)

find minimum $\alpha \geq 0$ where min loss value when step to d^k direction

$$f(x^k + \alpha^k d^k) = \min_{\alpha} f(x_k + \alpha d_k)$$

but, in practice it is not used due to computational cost.

Subject Date

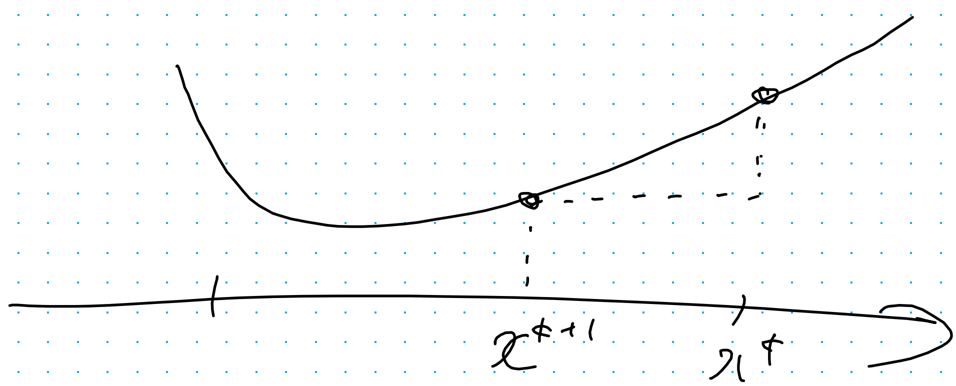
Attendants How to find good value of steps?

Armijo condition

Wolf condition

How to find d^k direction

$$f(x^{k+1}) - f(x^k) = f(x^k + \alpha^k d^k) - f(x^k)$$



$$\lim_{t \to +\infty} \frac{f(x^k + t d^k) - f(x^k)}{t} = \nabla f(x^k) d^k$$

$d^k = -\nabla f(x^k)$ gradient descent direction

Subject Date 5

Attendants

Assumption

convex - $\nabla^2 L(w)$ positive definite.

Lipschitz condition

$$|L(w^t) - L(w^{t+1})| \leq G \|w^t - w^{t+1}\|$$

$$\Rightarrow \|\nabla L(w^t)\|_2 < G$$

$$f(x_k + x) \simeq f(x_k) + \nabla f(x_k)^T x + \frac{1}{2} x^T H(x_k) x$$

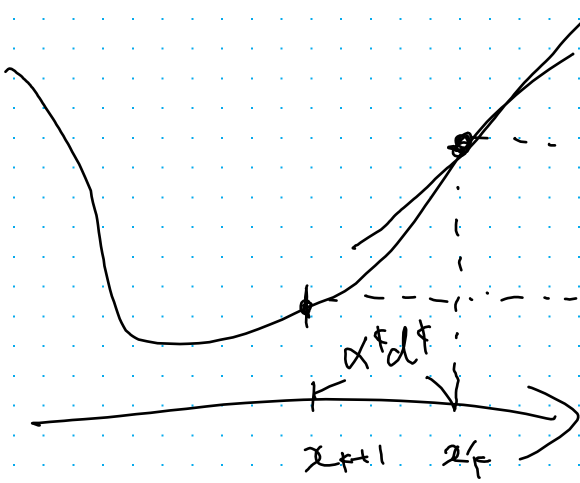
f の $\bar{x}$ における 2次のテイラー展開 $\bar{f}$

$$\bar{f}(x) = f(\bar{x}) + f'(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T f''(\bar{x})(x - \bar{x})$$

Subject

Date 6

Attendants



$f(x_{k+1}) - f(x_k)$

$= f(x_k + \alpha^k d^k) - f(x_k)$

< 0

となるような d^k を求めたい

$$\lim_{\alpha_k \to 0} \frac{f(x_k + \alpha_k d_k) - f(x_k)}{\alpha_k d_k} = \nabla f(x_k)$$

$$\Leftrightarrow \lim_{\alpha_k \to 0} \frac{f(x_k + \alpha_k d_k) - f(x_k)}{\alpha_k} = \nabla f(x_k) d_k$$

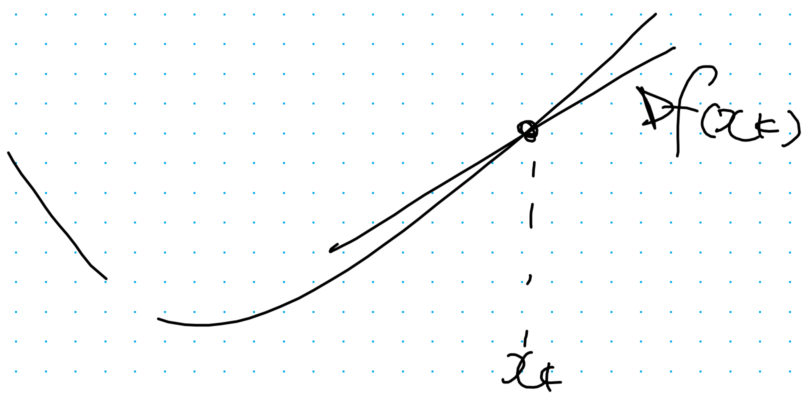
方向微係数

これが負となる d^k を降下方向

Subject Date 7

Attendants

$$\nabla f(x_k) d_k = |\nabla f(x_k)| |d_k| \cos\theta$$



$$d_k = \underset{d_k}{\mathrm{argmin}} \left\{ f(x_k + \alpha_k d_k) - f(x_k) \right\}$$

$$= \underset{d_k}{\mathrm{argmin}} \left\{ \lim_{\alpha \to 0} \frac{f(x_k + \alpha_k d_k) - f(x_k)}{\alpha_k} \right\}$$

$$= \underset{d_k}{\mathrm{argmin}} \underbrace{\left\{ \nabla f(x_k) d_k \right\}}$$

$$\min \left\{ \nabla f(x_k) d_k \right\} = \min \left\{ |\nabla f(x_k)| |d_k| \cos\theta \right\}$$

$\cos\theta = -1$ のとき

つまり $d_k = -\nabla f(x_k)$

Subject　　　　Date　8

Attendants 収束性

$$x^{k+1} = x^{k} - \alpha^{k} \nabla f(x^{k})$$

$\{x^k\}$ の点列が収束するには $\|\nabla f(x_k)\| \to 0$ $(k \to \infty)$

が必要十分条件

Zoutendijk 条件より

$$\sum_{k=0}^{\infty} \left(\frac{\nabla f(x_k)^T d_k}{\|d_k\|} \right) < \infty$$

$$\Longleftrightarrow \sum_{k=0}^{\infty} \|\nabla f(x_k)\| \cos\theta_k < \infty$$

ここで $\cos\theta_k = \left(\frac{-\nabla f(x_k) d_k}{\|\nabla f(x_k)\| \|d_k\|} \right)$

$d_k = -\nabla f(x_k)$ より

$$\cos\theta_k = 1$$

Subject　　　　　　　　　　　　　　　　　　Date 9

Attendants

つまり　　　　　　Zoutendijk 条件

$$\sum_{k=0}^{\infty} \|\nabla f(x_k)\| < \infty$$

(大域的収束定理)

無限級数が収束するために

$$\lim_{k \to \infty} \|\nabla f(x_k)\| = 0$$

勾配のノルムが0に収束していく

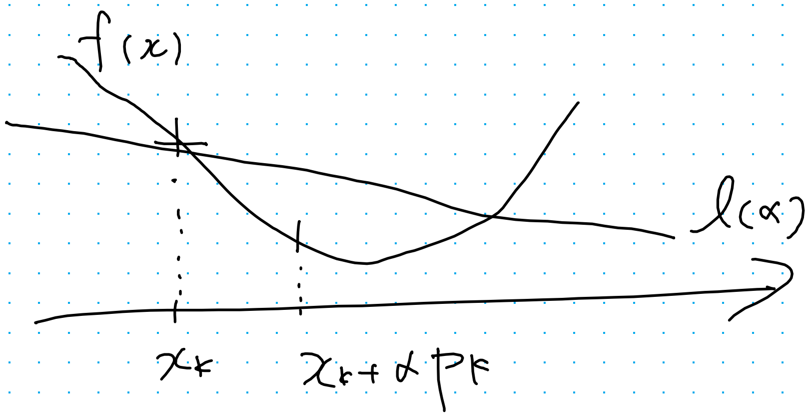
Subject　　　　　　　　　　　　Date 10

Attendants

Armijo 条件（不完全な直線探索条件）

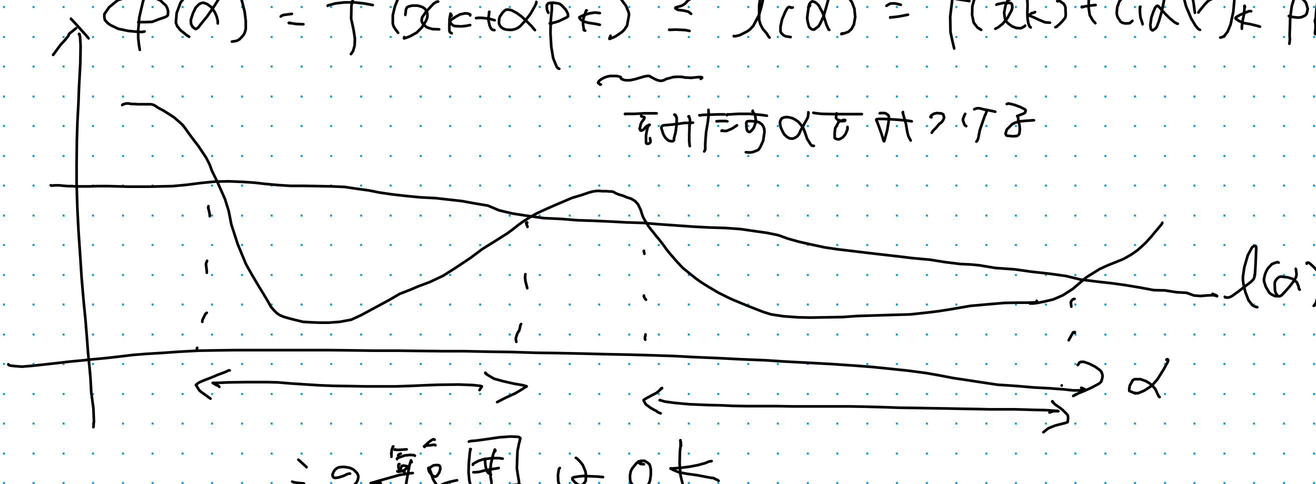
$$f(x_k+\alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k = l(\alpha)$$

方向性導関数



$$\phi(\alpha) = f(x_k+\alpha p_k) \leq l(\alpha) = f(x_k) + c_1 \alpha \nabla f_k^T p_k$$

を満たすαを見つける



$\phi(\alpha) \leq l(\alpha)$ なら十分な減少条件を満たす

Subject

Date / /

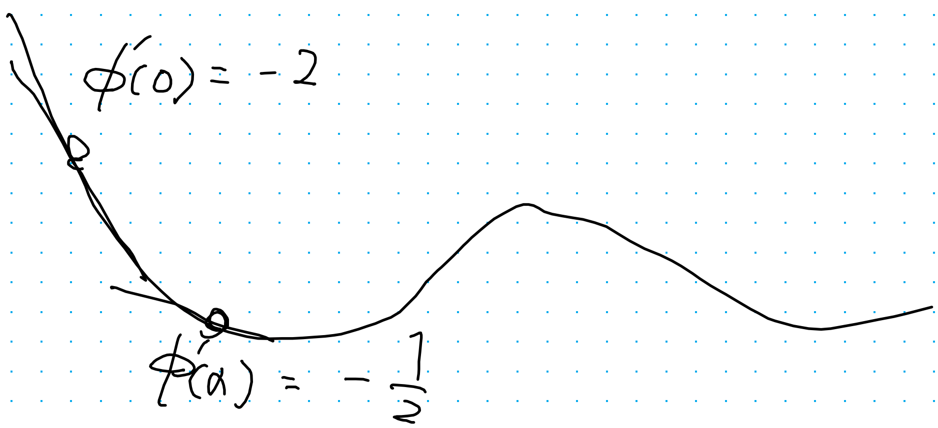
Attendants

小さすぎる α では意味がない

→ 曲率条件

α が $\underbrace{\nabla f(x_k+\alpha_k p_k)^T}_{= \phi'(\alpha_k)} p_k \geq c_2 \nabla f_k^T p_k$

$\phi'(\alpha_k)$ が $\phi'(0)$ より大きい



Wolfe 条件

十分な減少条件と曲率条件

$\phi(\alpha) < l(\alpha)$ をみたすα

$\phi'(\alpha) > l(\alpha)$ をみたすα

Zoutendijk 条件 (大域的収束定理)

p_kが降下方向, α_kがWolfe条件を満たす

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \le L \|x - \tilde{x}\|$$

が成り立つとすると

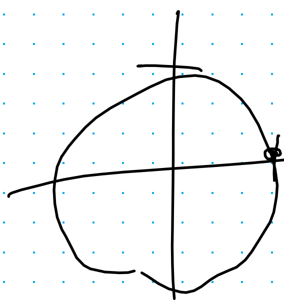
$$\sum_{k\ge 0} \cos^2\theta_k \|\nabla f_k\|^2 < \infty$$ が成立

where $\cos\theta_k = \dfrac{-\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}$

→ 探索方向p_kと最急降下方向$-\nabla f_k$間の角度

$\cos 90^\circ = 0$

$\cos^2\theta_k \|\nabla f_k\|^2 \to 0$



$\lim_{k\to\infty} \|\nabla f_k\| = 0$

$\cos\theta \neq 0 \iff \theta \neq 90^\circ$ でない場合 → 大域的収束が保証できる

$\boxed{\text{収束率}}$

$q \in (0,1)$, $\exists k'$

$$\frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|} \le q \quad \forall k > k' \text{ が成立}$$

$\Rightarrow$ 一次収束

$$\lim_{k\to\infty} \frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|} = 0$$

$\Rightarrow$ 超一次収束

$$\frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|^p} \le M$$

$M > 0$ と $\exists k'$ $\forall k \ge k'$ が成立 $\Rightarrow$ p次収束

最急降下法の収束率

$$f(x) = \frac{1}{2} x^T Q x - b^T x$$

$$x^* = Q^{-1} b \quad \because \; f'(x) = 0$$

$$\iff Qx - b = 0$$

$$\iff x = Q^{-1} b$$

最急降下法 $\quad x_{k+1} = x_k - \alpha_k \nabla f_k$

$$f(x_{k+1}) = f(x_k - \alpha_k \nabla f_k)$$

$$= \frac{1}{2} (x_k - \alpha_k \nabla f_k)^T Q (x_k - \alpha_k \nabla f_k) - b^T (x_k - \alpha_k \nabla f_k)$$

α_kについて解くと、 $\quad \alpha_k = \dfrac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}$

つまり、 $\quad x_{k+1} = x_k - \left(\dfrac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \right) \nabla f_k$

ここで重みつきノルム

$\|x\|_Q^2 = x^T Q x$ を導入

Subject Date 15

Attendants

$x^* = Q^{-1}b$ ナ')

$$\underbrace{\frac{1}{2}\|x - x^*\|_Q^2}_{} = \underbrace{f(x)}_{} - f(x^*)$$

$$\frac{1}{2}(x-x^*)^T Q (x-x^*) \qquad \frac{1}{2}x^T Q x - b^T x$$

$$\boxed{\begin{aligned} f(x^*) &= \frac{1}{2}Q^{-1}b\,\underline{\underline{Q}}\,Q^{-1}b - b^T Q^{-1}b \\ &= \frac{1}{2}Q^{-1}b^2 - Q^{-1}b^2 \\ &= -\frac{1}{2}Q^{-1}b^2 \end{aligned}}$$

$$= \frac{1}{2}x^T Q x - bx + \frac{1}{2}Q^{-1}b^2$$

$$= f(x) - f(x^*)$$

$$\frac{1}{2}(x - x^*)^2 Q$$

$$= \frac{1}{2}Q\left\{x^2 - 2xx^* + x^{*2}\right\}$$

$$= \frac{1}{2}x^T Q x - xQQ^{-1}b + \frac{1}{2}\underline{QQ}^{-1}bQ^{-1}b$$

Subject

Date 16

Attendants

$$\nabla f_k = Q(x_k - x^*)$$

$\Leftrightarrow$

$$\left(\nabla f_k (x_k - x^*) = \|x_k - x^*\|_Q^2 \right)$$

$$\boxed{\begin{aligned} f(x) &= \frac{1}{2} x^T Q x - b^T x \\ \nabla f(x) &= Qx - b \\ &= Q(x - Q^{-1} b) \\ &= Q(x - x^*) \end{aligned}}$$

$$\frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}$$

$$\|x_{k+1} - x^*\|_Q^2 = \|x_k - (\cdot\cdot)\nabla f_k - x^*\|_Q^2$$

$$= \|(x_k - x^*) - (\cdot\cdot)\nabla f_k\|^2$$

$$= (\cdot\cdot)^2 Q$$

$$= \left\{ (x_k - x^*)^2 - 2(\cdot\cdot)\nabla f_k (x_k - x^*) + ((\cdot\cdot)\nabla f_k)^2 \right\} Q$$

$$= \|x_k - x^*\|_Q^2 + \|(\cdot\cdot)\nabla f_k^2\|_Q^2 - 2(\cdot\cdot)\nabla f_k (x_k - x^*)$$

Subject

Date 18

Attendants

$$\|x_t - x^*\|_Q^2 \left\{ 1 + \frac{\left((\because)\nabla f_k\right)^2}{\nabla f_k (x - x^*)} - \frac{2(\because)\cancel{\nabla f_k (x-x^*)}}{\cancel{\nabla f_k (x-x^*)}} \right\}$$

$$= \|x_k - x^*\|_Q^2 \left\{ 1 + (\because)^2 \times \frac{\nabla f_k}{(x-x^*)} - 2(\because) \right\}$$

$$= \|\sim\|_Q^2 \left\{ 1 + (\because)\left\{ \frac{\nabla f_k}{(x-x^*)} - \frac{2(x-x^*)}{(x-x^*)} \right\} \right\}$$

$$= \|\sim\|_Q^2 \left\{ 1 + (\because) \frac{\nabla f_k - 2Q^{-1}\nabla f_k}{Q^{-1}\nabla f_k} \right\}$$

$$= \|x_t - x^*\|_Q^2 \left\{ 1 + \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \frac{\cancel{\nabla f_k}(1 - 2Q^{-1})}{Q^{-1}\cancel{\nabla f_k}} \right\}$$

$$= \|\sim\|_Q^2 \left\{ 1 + \frac{\nabla f_k^T \nabla f_k (Q-2)}{\nabla f_k^T Q \nabla f_k} \right\}$$

Subject

Date 19

Attendants

最急降下法の収束率

$$\|x_{k+1} - x^*\|_Q^2 \le \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 \|x_k - x^*\|_Q^2$$

誤差ノルムが以上を満たす

$\Rightarrow$ f_k は f^* に一次収束する

条件数 $c(Q) = \dfrac{\lambda_n}{\lambda_1}$ が大きくなると

収束性が悪化する

$$r \in \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1\right) \quad k \to \infty \text{ に対し}$$

$$f(x_{k+1}) - f(x^*) = r^2 |f(x_k) - f(x^*)|$$

e.g. $\lambda_n - \lambda_1$

$0.5 - 0.1 = 0.4$

$0.5 + 0.1 = 0.6$ $\quad \frac{2}{3} = 0.66.$

e.g. $\lambda_n - \lambda_1$

$0.5 - 0.3 = 0.2$

$0.5 + 0.3 = 0.8$ $\quad \frac{1}{4} = 0.25$

収束rateが良い

Subject　　Date 20

Attendants Newton Method.

$$P_k = -\underbrace{\nabla^2 f_k^{-1}} \nabla f_k$$

$H = \nabla^2 f_k$ が正定値とは限らない

→ trust region, modify H

2次収束をするが大域収束する保証なし

$$f(x_k + d) \simeq f(x_k) + \nabla f(x_k) d + \frac{1}{2} \nabla^2 f(x_k) d^2$$

テイラー展開　aでの$f(x)$の展開

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

マクローリン展開　0での$f(x)$の展開

x_kの関数

$f(x)$をx_kで展開

$$f(x) = f(x_k) + \nabla f(x_k)(x - x_k) + \frac{1}{2} \nabla^2 f(x_k)(x - x_k)^2$$

$x \leftarrow x_k + d$を代入　$f(x_k+d) = f(x_k) + \nabla f(x_k) d + \cdots$

Subject

Date 21

Attendants

$$f(x_k+d) = f(x_k) + \nabla f(x_k)\, d + \underbrace{\tfrac{1}{2}\nabla^2 f(x_k)\, d^2}_{2次近似} \cdots$$

$\nabla f(x^*) = 0$ (一次の最適性条件より)

$$d = \underset{d}{\arg\min}\ f(x_k) + \nabla f(x_k)\, d + \frac{1}{2}\nabla^2 f(x_k)\, d^2 = J$$

一次の最適性条件

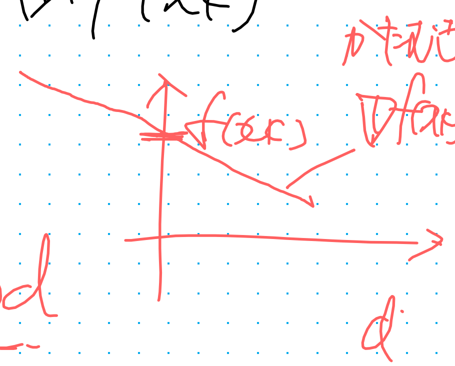
$$\frac{\partial J}{\partial d} = \nabla f(x_k) + \nabla^2 f(x_k)\, d = 0$$

のとき dはJを最小にする

$$\iff d = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

GDは一次近似

$$d = \underset{d}{\arg\min}\ f(x_k) + \nabla f(x_k)\, d$$



$\nabla f(x_k) > 0$ $\mathrm{sign}(d) = -1$,

Subject

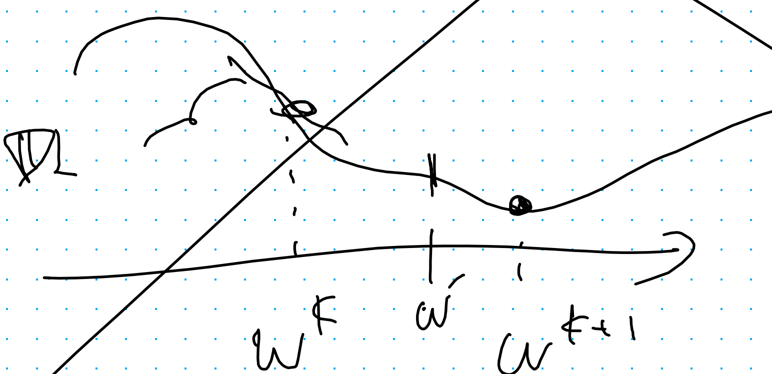
Date 22

Attendants

$w^* = \underset{w}{\mathrm{argmin}}\, L(w)$

重みベクトルの更新式

$$w^{(k+1)} = \underset{u}{\mathrm{argmin}}\, L\left\{ u - \left(w^{(k)} - \alpha^{k} \nabla L(w^{(k)}) \right) \right\}$$



$d = x_k - x^*$

$$f(x) = f(\alpha) + f'(\alpha)(x-\alpha) + \cdots$$

$\alpha \leftarrow x_k$ で展開

$$f(x) = f(x_k) + \nabla f(x_k)(x - x_k)$$

$x \leftarrow x_k + t d_k$ とすると

$$f(x_k + t d_k) = f(x_k) + \nabla f(x_k)\, t d_k$$

Subject Date 23

Attendants Iterative Method

$$\frac{f(x_k + t d_k) - f(x_k)}{t} = \nabla f(x_k) d_k < 0$$

$t \to 0$ となるような探索方向 d_k を

$f(x)$ の降下方向と呼ぶ

<line - search>

local & global convergence

q-linear convergence $\frac{\|x_k - x^*\|}{\|x_{k+1} - x^*\|} < c$

q-quadratic convergence

q-cubic convergence $\frac{\|x_k - x^*\|}{\|x_{k+1} - x^*\|^3} < c$

q-superlinear convergence

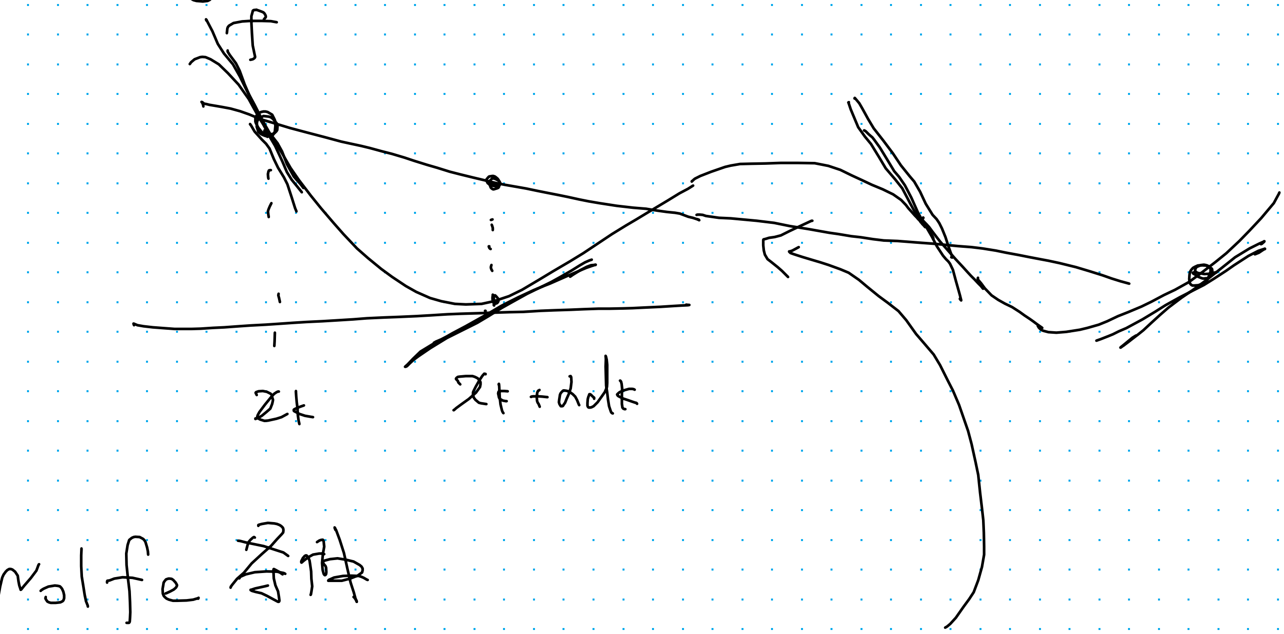
$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$

Subject

Date 24

Attendants αの探索

Armijo 条件



Wolfe 条件

$$f(x_k+\alpha d_k) \leq f(x_k) + \xi_1 \alpha \nabla f(x_k)^T d_k$$

に加えて

$$\xi_2 \nabla f(x_k)^T d_k \leq \nabla f(x_k+\alpha d_k)^T d_k$$

Subject

Date 25

Attendants Zoutendijk 条件

Assumption

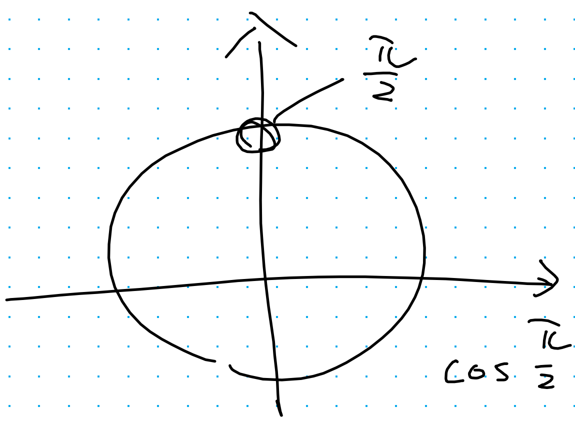
$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$$

search direction

$$d_k \quad \text{s.t.} \quad \nabla f(x_k)^T d_k < 0$$

$$\|$$

$$\|\nabla f(x_k)\| \|d_k\| \cos\theta < 0$$



$\cos\frac{\pi}{2} = 0$ これはダメ.

$$d_k = \underset{d_k}{\arg\min} \ \nabla f(x_k)^T d_k$$

$$= -\nabla f(x_k)$$

use wolfe condition

$$\sum_{k=0}^{\infty} \left(\frac{\nabla f(x_k)^T d_k}{\|d_k\|} \right)^2 < \infty.$$

Subject　　　　Date 2-6

Attendants

$$\sum_{k=0}^{\infty} \left(\|\nabla f(x_k)\| \cos\varphi\right)^2 < \infty$$

Zoutendijk (ズーテンダイク条件)

$$\lim_{k\to\infty} \frac{\nabla f(x_k)\, d_k}{\|d_k\|} = 0$$ ← 無限級数の収束条件より

$$\Longleftrightarrow \lim_{k\to\infty} \|\nabla f(x_k)\| \cos\varphi = 0$$

すべてのφに対し $\cos\varphi > \delta$を満たすδが存在するような降下方向を生成できるなら

$$\lim_{k\to\infty} \|\nabla f(x_k)\| \to 0$$

生成される点列の大域的収束性を表している

Subject

Date 27

Attendants

$$\cos\varphi = \frac{-\nabla f(x_k)\, d_k}{\|\nabla f(x_k)\|\ \|d_k\|}$$

より $\cos\varphi = \dfrac{(-\nabla f(x_k))^2}{\|\nabla f(x_k)\|^2} = 1$

最急降下法の収束率

$$\|x_{k+1} - x^*\|_A \leq \left|\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right| \|x_k - x^*\|_A$$

$0 < \lambda_1 \leq \lambda_n$ はそれぞれ行列 A の最小固有値と最大固有値

$$\|u\|_A = \sqrt{u^T A u}$$

$$\frac{|\lambda_1 - \lambda_n|}{|\lambda_1 + \lambda_n|} = \left|\frac{\frac{\lambda_1}{\lambda_n} - 1}{\frac{\lambda_1}{\lambda_n} + 1}\right| = 1 \text{ となって収束おそい}$$

$\frac{\lambda_n}{\lambda_1}$

条件数が大きいとき

Subject

Date 28

Attendants

strong convex → H is positive definite

$$eig(\nabla^2 f(x)) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

$$= Q \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} Q^T$$

where Q is orthogonal

$$z = Q^T w$$

Subject　　　　　　　　　　　　　　　　Date 29

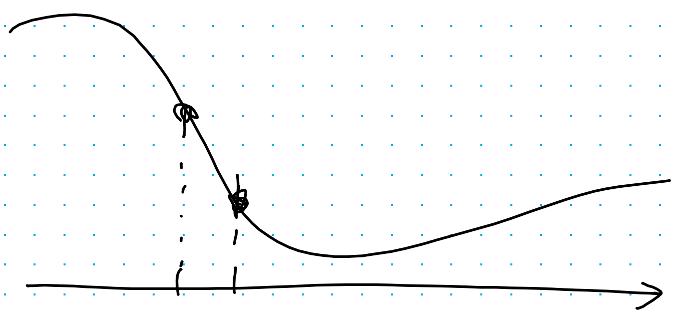
Attendants

方向微分 (directional derivative)

u方向のfの傾きを表す

$\frac{d}{d\alpha} f(x+\alpha u)$ 　$\alpha=0$で　$u^T \nabla_x f(x)$

chain rule　$\frac{\partial f}{\partial \alpha} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \alpha}$



$= \nabla_x f(x+\alpha u) \times \frac{\partial}{\partial \alpha}(x+\alpha u)$

$= u^T \times \nabla_x f(x+\alpha u)$

この傾きは $\lim_{\alpha \to 0} \frac{f(x+\alpha u) - f(x)}{\alpha u}$　$\alpha \to 0$で

$\downarrow$

$\underline{u^T \nabla_x f(x)}$

fを最小化する方向を考える

方向微分は$f(x+\alpha u)$のαに関する微分を$\alpha=0$で評価

$\downarrow$

$\frac{\partial f}{\partial \alpha} = u^T \nabla_x f(x)$

$\min_{u,\, u^T u = 1} u^T \nabla_x f(x) = \min_{u} \|u\|_2 \|\nabla_x f(x)\|_2 \cos\theta$

Subject

Date 30

Attendants

$\|u\|_2 = 1$ より

$$\min_u \underbrace{\|\nabla_x f(x)\|_2}_{u\text{に依存しない}} \cos\theta$$

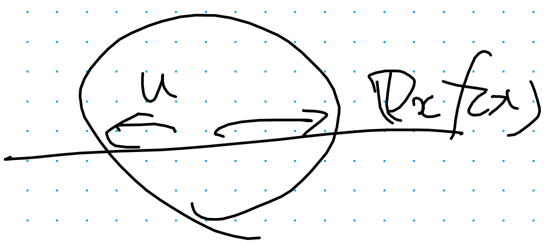
$$= \min_u \cos\theta$$

これは u と $\nabla_x f(x)$ が 反対向き

$\cos\theta = -1$

$\theta = 180^\circ$

で最小になる



これを最急降下法という

$$x_{t+1} = x_t - \varepsilon \nabla_x f(x)$$

Jacobian

Matrix $f : \mathbb{R}^m \to \mathbb{R}^n$

$J \in \mathbb{R}^{n \times m}$ $\quad J_{ij} = \frac{\partial}{\partial x_j} f(x)_i$

Subject

Date 31

Attendants

$$\mathbb{D}^2 f(x) = \frac{\partial^2}{\partial x_i \partial x_j}$$: curvature

関数の入力が多次元の場合、2階微分が多数存在する

↳ これをまとめたのを Hessian Matrix と呼ぶ

$$H(f)(x)_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$$

Jacobian of gradient = Hessian

実対称行列 → 固有値の集合と固有ベクトルの直交基底に分解できる

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

新しい点 x は $x_0 - \varepsilon g$ とすると

$$f(x_0 - \varepsilon g) \approx f(x_0) - g^T \varepsilon g + \frac{1}{2}(\varepsilon g)^2 H$$

$$\approx f(x_0) - \varepsilon g^T g + \frac{1}{2} \varepsilon^2 g^T H g$$

Subject　　　　　　　　　　　　　　Date 32

Attendants

$g^T H g > 0$ の場合，最適なstep幅

$$\varepsilon^* = \frac{g^T g}{g^T H g}$$

最悪の場合は g が H の最大固有値 λ_{max} に対応する
固有ベクトルと同じ方向の場合

$$\Leftrightarrow \frac{1}{\lambda_{max}} \text{ がstep幅}$$

Hessian
条件数が悪い場合，GDはあまり機能しない

∵ある方向には急激に g が大きくなるけど
　〃　には　ゆっくり g が変わる

→ 勾配が長く負である方向を優先的に探索すべきと判断できない

ニュートン法

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

$$f(x_0 + d) \approx f(x_0) + f'(x_0) d + \frac{1}{2} f''(x_0) d^2$$

$$\frac{\partial f}{\partial d} = f'(x_0) + f''(x_0) d = 0$$ のとき $f(x_0 + d)$ が最小

Subject

Date 33

Attendants

$$f'(x_0) + f''(x_0) d = 0$$

$$\Longleftrightarrow \quad d = -f''(x_0)^{-1} f'(x_0)$$

$$= -[\nabla^2 f(x_0) \nabla f(x_0)$$

Lipschitz Continuous

$$\forall x, \forall y, \quad |f(x) - f(y)| \leq L \|x - y\|_2$$

Karush-Kuhn-Tucker 法は

constrained optimization に対する

一般的な解を与える

generalized Lagrangian

Lagrange Function

$$L(x, \lambda, \alpha) = f(x) + \sum_i \lambda_i g^{(i)}(x) + \sum_j \alpha_j h^{(j)}(x)$$

$$\forall i \quad g^{(i)}(x) = 0, \quad \forall j \quad h^{(j)}(x) \leq 0$$

Subject　　　　　　　　　　　　　　　　　　　　　　　　　　　　Date　34

Attendants　線形二乗法

最小

$$f(x) = \frac{1}{2}\|Ax-b\|_2^2$$

$$\nabla_x f(x) = A^T(Ax-b) = A^TAx - A^Tb$$

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

表面的には、この分解によってモデル $p(x)$ の教師学習を

n個の supervised task に分解して解く。

$$p(y \mid x) = \frac{p(x,y)}{\sum_{y'} p(x,y')}$$

Linear Regression　$\hat{y} = w^T x$

Mean Square Error

$$MSE_{test} = \frac{1}{m}\sum_i (\hat{y}^{(test)} - y^{(test)})_i^2$$

$$\nabla_w MSE_{train} = 0$$

m

Subject

Date 35

Attendants

$$y = w^T x$$

$$\nabla_w MSE_{train} = 0$$

$$\Rightarrow \nabla_w \frac{1}{m} \| \hat{y} - y \|_2^2 = 0$$

$$\Leftrightarrow \nabla_w (w^T x - y)^T (w^T x - y) = 0$$

$$\Leftrightarrow x(w^T x - y) + (w^T x - y)^T x$$

$$\Leftrightarrow 2w^T x^2 - 2yx = 0$$

$$\Leftrightarrow w = yx(x^{-2})$$

$$= (X^T X)^{-1} X^T Y$$

Subject

Attendants Memo @ Zoom w/ Ioannis

Date

$$l(\hat{y}, y) = (\hat{y} - y)^2$$

$$f_i(\theta^*) = 0, \quad \nabla f_i(\theta^*) = 0 \quad \leftarrow \text{assumption 1}$$

$$f_i(\theta) = l(f_\theta(x_i), y_i) = (\hat{y} - y)^2$$

$$\nabla_\theta f_i(\theta) = 2(\nabla_\theta \hat{y} - \nabla_\theta y)$$

finite sum

emirical Risk

$\leftarrow \underset{\theta}{\text{argmin}}$

output

set

$$y = x^3 - x$$
$$= x(x^2 - 1)$$

$f(\theta)$

Subject Date

Attendants Memo @ Zoom w/ Ioannis

$$\frac{\partial}{\partial z}(y_i - z)^2 = 2(y_i - z)$$

$$z = f_{\theta^*}(x_i) = y_i$$

$$\to 0$$

---

interpolation $\hat{y} = y$

$\overbrace{\text{make it}}$

If grad $\neq 0$ $\to$

no constraints at problem

then we could

$\Rightarrow$ smaller value of Loss

Subject Date

Attendants Memo @ Zoom w/ Ioannis

interporation

¿

possible with NN

at least 2 layers

if we made it.
wide enough

$\nabla_\theta l(y_i, f_\theta(x_i))$

Subject Date

Attendants Memo @ Zoom w/ Ioannis

Gradient is always exists.

↓

w/o Assum 4.

grad $\neq 0$

⇓

loss given small

contrary loss $\neq 0$.