制約なし最適化の解法②

山下信雄

講義内容

1. 具体的な直線探索法
    - 最急降下法
    - ニュートン法
    - 準ニュートン法

2. 信頼領域法

3. 確率的勾配法

直線探索法

ステップ１：　降下方向 d^k を求める.
ステップ２：　ステップ幅 t_k を定める.
ステップ３：　$x^{k+1} = x^k + t_k d^k$ として，ステップ１へ

適当な仮定のもとで，停留点($\nabla f(x^*) = 0$)
に 大域的収束する.

最急降下法

最急降下方向：

$$d^k = -\nabla f(x^k) = -\color{red}{I}\nabla f(x^k)$$

$$\nabla f(x^k)^\top d^k = -\nabla f(x^k)^\top \nabla f(x^k) = -\left\|\nabla f(x^k)\right\|^2$$

$$\left\|d^k\right\| = \left\|\nabla f(x^k)\right\|$$

$$\Downarrow$$

大域的収束の仮定を満たす

最急降下法の性質

良いところ:

- 各反復の計算が簡単
- 大域的収束する

悪いところ：

- 収束が遅い（高々一次収束）

ニュートン法

ニュートン方向：

$$d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$$

ニュートン方向は次の最小化問題(部分問題)の最適解

$$\begin{aligned} &\min\ m_k(d) \\ &\text{s.t.} \quad d \in R^n \end{aligned}$$

$$0 = \nabla m_k(d) = \nabla f(x^k) + \nabla^2 f(x^k) d$$

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

ここで $m_k(d) = f(x^k) + \nabla f(x^k)^\top d + \frac{1}{2} d^\top \nabla^2 f(x^k) d$

$f(x^k + d)$ の2次近似

ニュートン法の収束性

ニュートン方向は一般には降下方向にならない.

$\Rightarrow$　直線探索法では大域的収束しないことがある.

大域的収束と2次収束するための十分条件：

次の不等式をみたす定数 $c_1, c_2 > 0$ が存在する.

$$c_1 \|v\|^2 \le v^\top \nabla^2 f(x)^{-1} v \le c_2 \|v\|^2 \quad \forall v, x \in R^n$$

$$\nabla f(x^k)^\top d^k = -\nabla f(x^k)^\top \nabla^2 f(x^k)^{-1} \nabla f(x^k) \le -c_1 \|\nabla f(x^k)\|^2$$

$$\|d^k\| = \|\nabla^2 f(x^k)^{-1} \nabla f(x^k)\| \le c_2 \|\nabla f(x^k)\|$$

ニュートン法の2次収束性

ステップ幅は $t_k = 1$ とする.

$$\begin{aligned}\left\| x^{k+1} - x^* \right\| &= \left\| x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k) - x^* \right\| \\ &= \left\| \nabla^2 f(x^k)^{-1} \left(\nabla f(x^*) - \nabla f(x^k) - \nabla^2 f(x^k)(x^k - x^*) \right) \right\| \\ &\le c_2 \left\| \nabla f(x^*) - \nabla f(x^k) - \nabla^2 f(x^k)(x^k - x^*) \right\|\end{aligned}$$

f を2回連続的微分可能とすると

$$\nabla f(x^*) = \nabla f(x^k) + \nabla^2 f(x^k)(x^* - x^k) + O\left(\left\| x^* - x^k \right\|^2 \right)$$

このとき $\quad \left\| x^{k+1} - x^* \right\| \le O\left(\left\| x^k - x^* \right\|^2 \right)$

ニュートン法の性質のまとめ

よいところ

- 収束すれば，収束は速い(2次収束)

わるいところ

- 直線探索法では大域的収束しない
- 各反復で線形方程式(ニュートン方程式)を解かなければならない.

$$\nabla f(x^k) + \nabla^2 f(x^k)d = 0$$

一般に計算量は $O(n^3)$

準ニュートン法

準ニュートン方向：

$$d^k = -(B_k)^{-1}\nabla f(x^k) = -H_k\nabla f(x^k)$$

$H_k = B_k^{-1}$ が正定値行列のとき

$$\nabla f(x^k)^\top d^k = -\nabla f(x^k)^\top H_k \nabla f(x^k) < 0$$

$\Rightarrow$ 降下方向

近似ヘッセ行列の望ましい条件

近似ヘッセ行列 B_k とその逆行列 H_k

- H_k が正定値 ⇒ 大域的収束
- $H_k \approx \nabla^2 f(x^k)^{-1}$ ⇒ 速い収束

近似ヘッセ行列のセカント条件

$\nabla f(x_k)$のテーラー展開を考えると,

$$\nabla f(x_{k+1}) - \nabla f(x_k) \approx \nabla^2 f(x_{k+1})(x_{k+1} - x_k)$$

を得る.

$$s_k = x_{k+1} - x_k,\quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

とする．近似ヘッセ行列が

セカント条件

$$\boxed{y_k = B_{k+1} s_k \ \text{ or } \ H_{k+1} y_k = s_k}$$

を満たせば，ヘッセ行列と似た性質をもつ.

⇒　速い収束が期待できる.

Broyden–Fletcher–Goldfarb–Shanno (BFGS)更新

BFGS更新：

$$B_{k+1} = B_k - \frac{B_k s_k (B_k s_k)^\top}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{s_k^\top y_k}$$

$$H_{k+1} = H_k - \frac{H_k y_k s_k^\top + s_k (H_k y_k)^\top}{s_k^\top y_k} + \left(1 + \frac{y_k^\top H_k y_k}{s_k^\top y_k}\right) \frac{s_k s_k^\top}{s_k^\top y_k}$$

- セカント条件を満たす．

$$B_{k+1} s_k = B_k s_k - \frac{B_k s_k (B_k s_k)^\top s_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top s_k}{s_k^\top y_k} = y_k$$

- H_kが正定値行列で，$s_k^\top y_k > 0$ であれば正定値行列となる．

準ニュートン法の性質

適当な条件のもとで

・大域的収束する.

・超一次収束する.

・ H_{k+1} の更新と，ベクトル $\nabla f(x^k)$ と H_{k+1} の掛け算は $O(n^2)$

$$H_{k+1} = H_k - \frac{H_k y_k s_k^\top + s_k (H_k y_k)^\top}{s_k^\top y_k} + \left(1 + \frac{y_k^\top H_k y_k}{s_k^\top y_k}\right) \frac{s_k s_k^\top}{s_k^\top y_k}$$

注意：行列と行列の掛け算は $O(n^3)$

準ニュートン法の問題点

BFGS更新：
$$H_{k+1} = H_k - \frac{H_k y_k s_k^\top + s_k (H_k y_k)^\top}{s_k^\top y_k} + \left(1 + \frac{y_k^\top H_k y_k}{s_k^\top y_k} \right) \frac{s_k s_k^\top}{s_k^\top y_k}$$

$s_k = (1, 1, \ldots, 1)^\top$ のとき，

$$\boldsymbol{s}_k \boldsymbol{s}_k^\top = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

ヘッセ行列が疎であっても，H_{k+1} は密な行列となる．

$\Rightarrow$ 大規模な問題には使えない！

補足：スパース性(疎性)について

スパース性：ベクトルや行列の成分がほとんど 0 になる性質

＊ 0 の成分は，足し算や掛け算で計算する必要がない．

例： $f(x) = \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 + x_n^2$

ヘッセ行列は3重対角行列． $\nabla^2 f(x) = \begin{pmatrix} 2 & -2 & 0 & \cdots & 0 \\ -2 & 4 & -2 & \ddots & \vdots \\ 0 & -2 & 4 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -2 \\ 0 & \cdots & 0 & -2 & 4 \end{pmatrix}$

ニュートン方程式は $O(n)$で計算できる．

準ニュートン法＋直線探索

[良いところ]
- 大域的収束かつ超1次収束
- 実装が簡単

[悪いところ]
- H_k は非ゼロ要素がほとんどないため,
  大規模な問題(数万変数以上)には適用できない.

講義内容

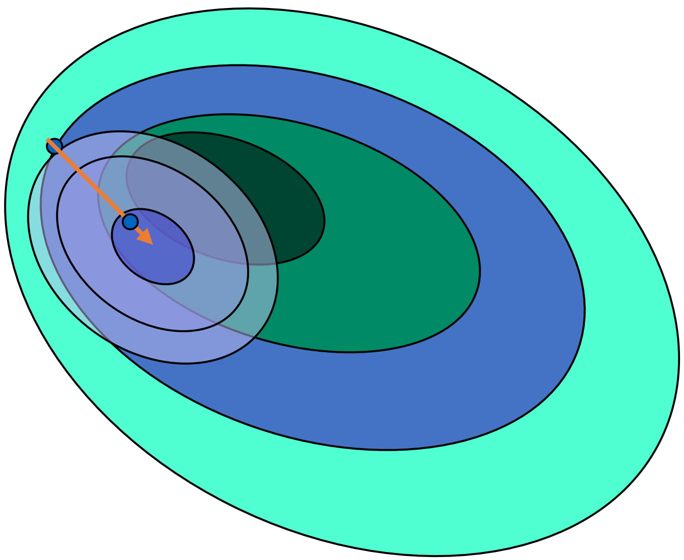
1. 具体的な直線探索法
    - 最急降下法
    - ニュートン法
    - 準ニュートン法

2. 信頼領域法

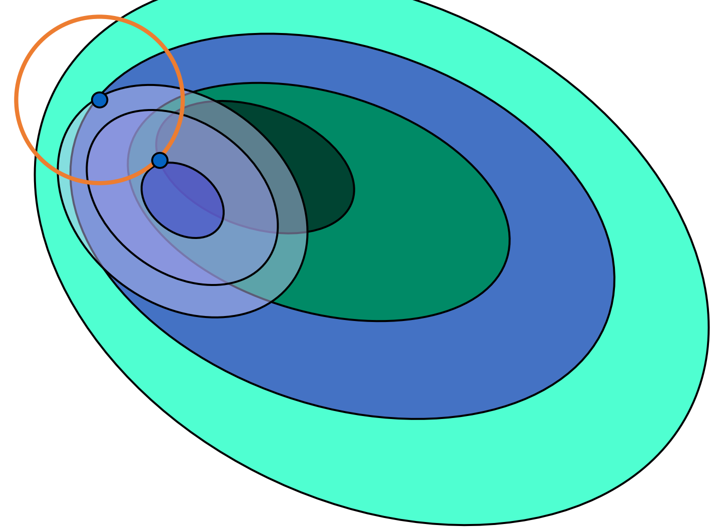
3. 確率的勾配法

大域的収束性を持たせる技術

- 直線探索法と信頼領域法 -



直線探索法



信頼領域法

信頼領域法

モデル関数のコンパクト集合上での最小解を探索報告とする．

$$\begin{array}{ll} \min & m_k(d) \\ \text{s.t.} & \|d\| \le \Delta_k \end{array}$$

- この部分問題には最適解 d^k が存在する．
- Δ_k を信頼半径という．
- $B(x_k; \Delta) = \{x \mid \|x - x_k\| \le \Delta\}$ を信頼領域という．

モデル関数は 2 次近似関数なので，信頼半径が小さいときは，

$$f(x^{\mathrm{tri}}) < f(x_k)$$

が成り立つ．ただし， $x^{\mathrm{tri}} = x^k + d^k$

信頼半径の更新

$$m_k(0) = f(x^k)$$

$$m_k(d^k) = f(x^k) + \nabla f(x^k)^\top d^k + \frac{1}{2}(d^k)^\top \nabla^2 f(x^k) d^k$$

$$= f(x^k) + \nabla f(x^k)^\top (x^{\mathrm{tri}} - x^k) + \frac{1}{2}(x^{\mathrm{tri}} - x^k)^\top \nabla^2 f(x^k)(x^{\mathrm{tri}} - x^k)$$

$$\approx f(x^{\mathrm{tri}})$$

モデルの信頼性を測る指標

$$\rho_k = \frac{\text{目的関数の減少量}}{\text{モデル関数の減少量}} = \frac{f(x_k) - f(x^{\mathrm{tri}})}{m_k(0) - m_k(d^k)}$$

$$\alpha_1 < 1 < \alpha_2$$

$$0 < \eta_1 < \eta_2 < 1$$

【信頼半径の更新】

$$\Delta_{k+1} = \begin{cases} \alpha_1 \left\| x^{\mathrm{tri}} - x_k \right\| & \text{if } \rho_k < \eta_1 \\ \Delta_k & \text{if } \eta_1 \le \rho_k < \eta_2 \\ \max\{\alpha_2 \left\| x^{\mathrm{tri}} - x_k \right\|, \Delta_k\} & \text{if } \eta_2 \le \rho_k \end{cases}$$

ニュートン法＋信頼領域法

[問題点]

- ヘッセ行列の計算

  ⇒　自動微分
- 非凸の部分問題の求解

  ⇒　高速の近似解法の開発

  （線形方程式を1回解く程度の計算時間）

大規模な問題では,

準ニュートン法＋直線探索

よりも優れている.

講義内容

1. 具体的な直線探索法
   - 最急降下法
   - ニュートン法
   - 準ニュートン法

2. 信頼領域法

3. 確率的勾配法

大規模な最適化問題

$$
\begin{aligned}
&\min \frac{1}{M}\sum_{i=1}^{M} f_i(x) \\
&\text{s.t.} \quad x \in R^n
\end{aligned}
$$

大規模となるもの

- 変数の数 *n*
- 目的関数を構成する関数の数 *M*

応用例：　データ解析

$$\min \frac{1}{M}\sum_{i=1}^{M}\theta(x,a^i)$$

- a^i はデータ.　θ は損失関数
- $\frac{1}{M}\sum_{i=1}^{M}\theta(x,a^i)$ は経験的損失とよばれる.

想定する問題

- Mが大きい，大規模な問題
  → 関数値や勾配の計算が大変

- 似たような関数 f_i (データ a^i)が多い

例： ディープラーニング，SVM, L1正則化問題, etc

確率勾配法

以下では，目的関数を $f(x)=\dfrac{1}{M}\sum_{i=1}^{M} f_i(x)$ とする．

最急降下法： $x^{k+1}=x^k-t_k\nabla f(x^k)$

$\nabla f_i(x^k)$の計算量が $O(n)$ のとき，

1回の反復の計算量は $O(nM)$

確率的勾配法： $i_k\in\{1,2,\ldots,M\}$をランダムに選ぶ

$$x^{k+1}=x^k-t_k\nabla f_{i_k}(x^k)$$

1回の反復の計算量は $O(n)$

いろいろな呼び方

- Stochastic gradient descent method
  i_k をランダムにとってくるとき
  ニューラルネットワークの学習では誤差逆伝播法という

- Incremental gradient method
  i_k を $\{1, 2, \cdots, M\}$ から順番に取ってくるとき

- Online gradient descent method
  逐次的に　暫定解 x^k を``実行(利用)”するとき

確率的勾配 $g^k = \nabla f_{i_k}(x^k)$ の特徴

- i_k を確率 $p_i = \frac{1}{M}$ で取ってきたとき，探索方向の期待値は目的関数の最急降下方向と一致する．

$$E_i[g^k] = E_i[\nabla f_i(x^k)] = \sum_{i=1}^{M} p_i \nabla f_i(x^k) = \frac{1}{M}\sum_{i=1}^{M} \nabla f_i(x^k) = \nabla f(x^k)$$

- 分散 $E_i\left[\left(g^k - E\left[\nabla f_i(x^k)\right]\right)^2\right]$ は 0 とならない．

→ 分散が小さいとき，最急降下法に近づく

ステップ幅 t_k を固定したとき

例： $\min \frac{1}{2}(x-1)^2 + \frac{1}{2}(x+1)^2$

$$f_1(x) = \frac{1}{2}(x-1)^2,\ f_2(x) = \frac{1}{2}(x+1)^2$$

最適解は $x^* = 0$

$x^0 = -0.5, t_k = 0.5$ とした最急降下法

$$x^1 = x^0 - t_k f'(x^0) = -0.5 - 0.5 \times \{(-1.5) + 0.5\} = 0$$

$x^0 = -0.5, t_k = 0.5$ とした確率的勾配降下法

$$i_k = 1: \quad x^1 = x^0 - t_k f_1'(x^0) = -0.5 - 0.5 \times \{-1.5\} = 0.25$$

$$i_k = 2: \quad x^1 = x^0 - t_k f_2'(x^0) = -0.5 - 0.5 \times \{0.5\} = -0.75$$

ステップ幅を固定すると，一般に収束しない

ステップ幅の取り方と収束

Diminishing Rule:

$$t_k \to 0,\ \sum_{k=1}^{\infty} t_k = +\infty$$

Diminishing Ruleを用いた確率勾配降下法は大域的収束する.

- f_i が凸であれば最適解に収束．そうでないときは停留点に収束.