第7章　降下法

この章では，制約なし最小化問題の解法である降下法ついて紹介する．降下法は、目的関数の微分 (勾配) の情報を用いることにより、各反復において必ず目的関数の値を減らす (降下させる) 反復法である。この降下法で重要となるのは、以下によい探索方向とステップサイズを求めるかである。探索方向の違いによって、最急降下法、ニュートン法、準ニュートン法と沸けることができる。また、ステップサイズ求める手法として、アルミホのルールがある。

7.1　降下法の枠組み

この節では、降下法の枠組みについて説明をする。特にその探索方向とステップサイズについて言及する。さらにその枠組みとなるアルゴリズムの大域的収束性について調べる

ここでは、目的関数 $f : R^n \to R$ を制約なしで最小化する次の問題を扱う。

$$\min \quad f(x) \tag{7.1}$$

制約なし最小化問題の解法の1つである降下法は，目的関数の値を減らす (降下する) 点を $\{x^1, x^2, \ldots\}$ と生成していく反復法である．つまり，

$$f(x^0) > f(x^1) > \cdots$$

となる点列 $\{x^k\}$ を生成する．このような点列は、各反復において、 探索方向 d^k とステップサイズ t_k が求めた後、

$$x^{k+1} := x^k + t_k d^k$$

と与えられる．降下法では，探索方向 d^k が次の条件を満たしている必要がある．

[**降下方向**]　$\nabla f(x^k)^T d^k < 0$

このような条件を満たしてるベクトル d^k を**降下方向**とよぶ．図 7.1 より明らかなように，降下方向に進めば関数の値を減らすことができる．実際、微分の定義より、

$$f(x^{k+1}) = f(x^k + t_k d^k) = f(x^k) + t_k \nabla f(x^k)^T d^k + o(t_k)$$

となるので、

$$f(x^{k+1}) - f(x^k) = t_k \left(\nabla f(x^k)^T d^k + \frac{o(t_k)}{t_k} \right) \tag{7.2}$$

となり、ステップサイズ t_k が十分小さいとき右辺は負になる、つまり、$f(x^{k+1}) < f(x^k)$ が成り立つ。

しかしながら，ステップサイズが小さくないときは、逆に目的関数の値が増加する場合がある (図 7.1, d^2)．そのため，ステップサイズ t_k をうまく調節することによって，$f(x^{k+1})$ を $f(x^k)$ よ

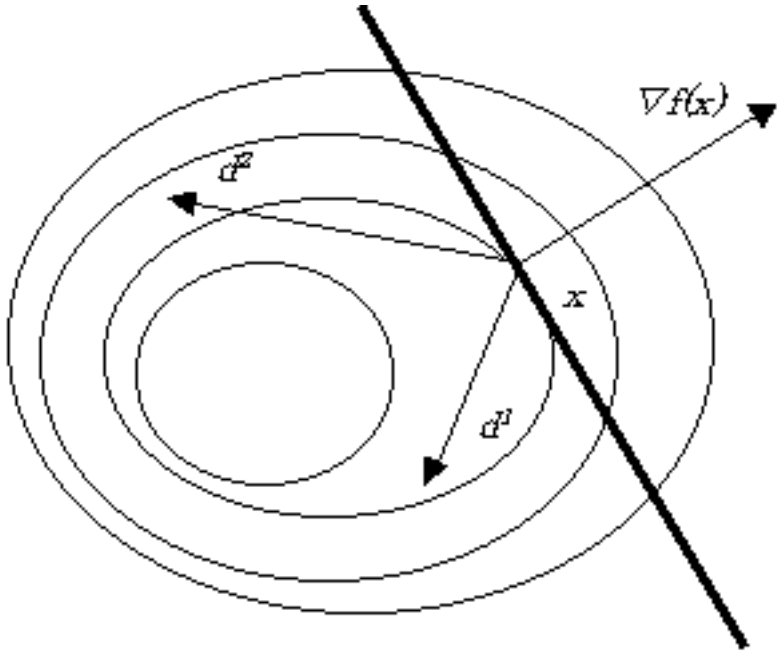


図 7.1: 降下方向

りも小さくする必要がある．この t_k の決め方はいくつか提案されている。最も有効に思われる手法は、目的関数を $\theta(t) := f(x^k + t_k d^k)$ として、次の 1 次元の最小化問題の解を t_k とすることである。

$$\begin{array}{ll} \min & \theta(t) \\ \text{subjectto} & l \le t \le u \end{array} \tag{7.3}$$

ここで、l, u は $l < u$ となる定数である。$l = 0, u = 1$ とすれば、(7.2) より、この問題の解 t_k は、目的関数の値を減らすことができる。この問題は f が凸関数であるとき、前章で紹介した黄金分割法を用いることによって解くことができる。しかしながら、凸関数でない場合は目的関数を減らす t_k が見つからない場合がある。また、一般に黄金分割法は、解を求めるために時間がかかるため、ステップサイズを求める解法としては適切でない。

そこで、(7.3) を厳密に解くことを諦めて、目的関数を減らすステップサイズを高速に得る手法が提案されている。その中のひとつに**アルミホのルール**と呼ばれる方法があり，理論的にも実用的にも優れていることが知られている．

アルミホのルール: $\alpha, \beta \in (0,1)$ を選び，次式を満たす最小の非負の整数 l を求め，$t_k := \beta^l$ とする．

$$f(x^k + \beta^l d^k) - f(x^k) \le \alpha\beta^l \nabla f(x^k)^T d^k$$

実際にプログラミングするときは，$l = 0, 1, \ldots$ と順に上記の式に代入していき，上記の不等式が満たされたときに $t_k := \beta^l$ とする．アルミホのルールは、$p(t) = f(x^k + td^k) - f(x^k)$ とすると、

$$p(t) \le \alpha p'(0)t \tag{7.4}$$

となる t を求めることになる。ここで、$p(0) = 0$ であることに注意すると、t の関数 $p'(0)t$ は $p(t)$

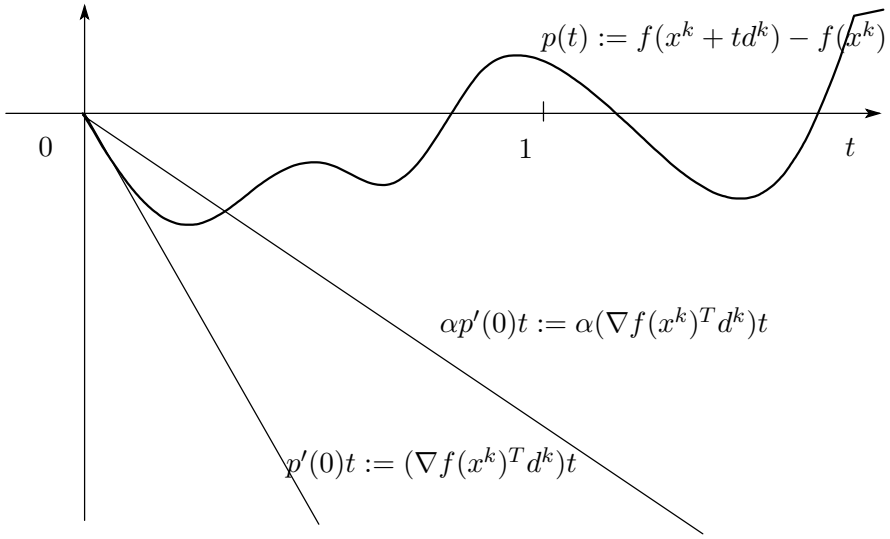


図 7.2: アルミホのルール

の 0 における接線になる (図 7.2)。そのため、t が小さいところでは、この接線と関数 p はほとんど同じになる。一方、その接線を $\alpha \in (0,1)$ 倍した関数 $\alpha p'(0)t$ は接線よりも大きくなる。このため、t が小さいときには、不等式 (7.4) が成り立つ。

このアルミホのルールを用いた降下法は以下のようになる.

降下法

ステップ 0 (初期設定): アルミホのルールのパラメータ $\alpha, \beta \in (0,1)$ を決める．適当に初期点 x^0 を決める．$k := 0$ とする.

ステップ 1 (終了判定): x^k が終了条件 $\|\nabla f(x^k)\| \leq \varepsilon$ を満たしているとき，x^k を解として終了.

ステップ 2 (探索方向の計算): 次式を満たす降下方向 d^k を定める.

$$\nabla f(x^k)^T d^k < 0$$

ステップ 3 (直線探索): アルミホのルールを用いて，ステップサイズ t_k を求める.

ステップ 4 (更新): $x^{k+1} := x^k + t_k d^k$ とする．$k := k+1$ として，ステップ 1 へ.

Remark 1 ステップ *0* におけるアルミホのルールのパラメータは，$\alpha = 0.0001$, $\beta = 0.5$ ぐらいにすることが多い．α を小さくすることによって，アルミホのルールをみたす整数 l を速く見つけることができる.

Remark 2 ステップ *1* における終了条件のパラメータ ε は，$\varepsilon = n \times 10^{-6}$ にすることが多い．これは，次元 n が大きいとき，誤差が大きくなることを許容する意味がある．また，10^{-6} はコンピュータの性能や扱う関数の性質によっては変更する必要がある.

この降下法によって生成される点列 $\{x^k\}$ に対して，次の収束に関する定理が示されている.

定理 1 降下方向の列 $\{d^k\}$ に対して，次式を満たす正の定数 γ_1, p_1 が存在するとする．

$$-\nabla f(x^k)^T d^k \geq \gamma_1 \|\nabla f(x^k)\|^{p_1} \tag{7.5}$$

このとき，アルミホのルールによってステップサイズは有限回で求めることができる．さらに，次の条件をみたす正の定数 γ_2, p_2 が存在するとする．

$$\|d^k\| \leq \gamma_2 \|\nabla f(x^k)\|^{p_2} \tag{7.6}$$

このとき，生成された点列の任意の集積点 x^* は f の停留点である．つまり，

$$\nabla f(x^*) = 0$$

である．

ここで、この定理において注意しなければならないことをいくつかあげておこう。まず、条件 (7.6) であるが、これは、探索方向 d^k が良い降下方向を向いていることを要求している。この条件を満たさない降下方向 d^k は、目的関数の勾配 $-\nabla f(x^k)$ とほとんど直交している。そのため、いくら降下方向だからといって、その方向に進んでも、目的関数の改善はほとんど得られない。次に注意しなければならないのは，必ずしも生成される点がすべて収束するとは言っていない点である。また、集積点が存在しても、その集積点は停留点 (最適性の 1 次の必要条件) である。もし関数 f が凸関数であれば、最適性の十分条件より、その点は大域的最適解となるが、そうでないときは”局所的最小解”のようなものが求まっているに過ぎない。

それでは、この定理の証明を与えよう。この証明においては、まず、点列が必ず生成される、つまり降下法が良定義であることを示す。その後、その点列が集積点を持つとき、その点列の中から収束する部分列を取りだし、その部分列が収束した先が停留点になることを示す。

証明. まず, 降下法が良定義であること、x^k が f の停留点でない場合, アルミホのルールを満たす β^l が有限回で求められることを示す. (x^k が f の停留点である場合、そこでアルゴリズムは終了する。) ここで, 有限回で求められないとすると, すべての正数 l に対して,

$$f(x^k + \beta^l d^k) - f(x^k) > \alpha \beta^l \nabla f(x^k)^T d^k$$

が成り立つことになる. このとき両辺を β^l で割り, $l \to \infty$ とすると, 微分の定義より、

$$\nabla f(x^k)^T d^k \geq \alpha \nabla f(x^k)^T d^k$$

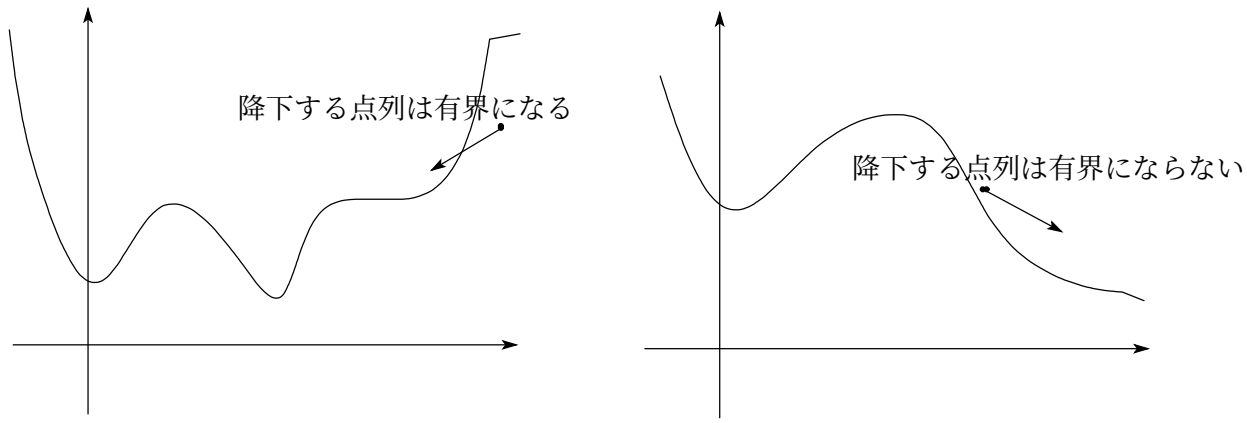
となる。 $0 < \alpha < 1$ であることから,

$$\nabla f(x^k)^T d^k \geq 0$$

となる．定理の仮定 (7.5) より，$\nabla f(x^k) = 0$ となり，x^k が f の停留点でないことに矛盾する．よって, アルミホのルールをみたすステップサイズは有限回で求めることができる．

次に降下法によって生成される任意の集積点 x^* が，停留点となることを示す．x^* は集積点であるので、x^* に収束する部分列 $\{x^k\}_K$ が存在する．ここで，K は非負整数の集合の部分集合である．このとき，点列 $\{\nabla f(x^k)\}_K$ は有界であり，$\nabla f(x^*)$ に収束する．

一方，アルミホのルールより $\{f(x^k)\}$ は単調に減少するので，$\{f(x^k)\}$ はある値に収束するか，$-\infty$ に発散する．部分列 $\{f(x^k)\}_K$ が $f(x^*)$ に収束するため，$\{f(x^k)\}$ も $f(x^*)$ に収束する．よっ



強圧性がある関数　　強圧性がない関数

図 7.3: 強圧性

て，アルミホのルールより，$\{\beta^{l_k}\nabla f(x^k)^T d^k\}$ も 0 に収束する．このとき、(i)$l_k \to \infty$ であるか、(ii)$\nabla f(x^k)^T d^k \to 0$ のどちらかが成り立つ。

まず、(i)$l_k \to \infty$ の場合を考える．(7.6) より，$\{d^k\}_K$ は有界となるので，一般性を失わずに，$\{d^k\}_K$ は d^* に収束するとする．アルミホのルールより

$$f(x^k + \beta^{l_k-1} d(x^k)) - f(x^k) > -\alpha\beta^{l_k-1}\nabla f(x^k)^T d^k$$

が成り立つので，両辺を β^{l_k} で割って, $k \in K \to \infty$ とすると

$$\nabla f(x^*)^T d^* > 0$$

を得る．これは，仮定 (7.5) に矛盾する．よって、(i) の場合はない。(ii)$\{\nabla f(x^k)^T d^k\}$ が 0 に収束する場合は、(7.6) より $\nabla f(x^*) = 0$ を得る． □

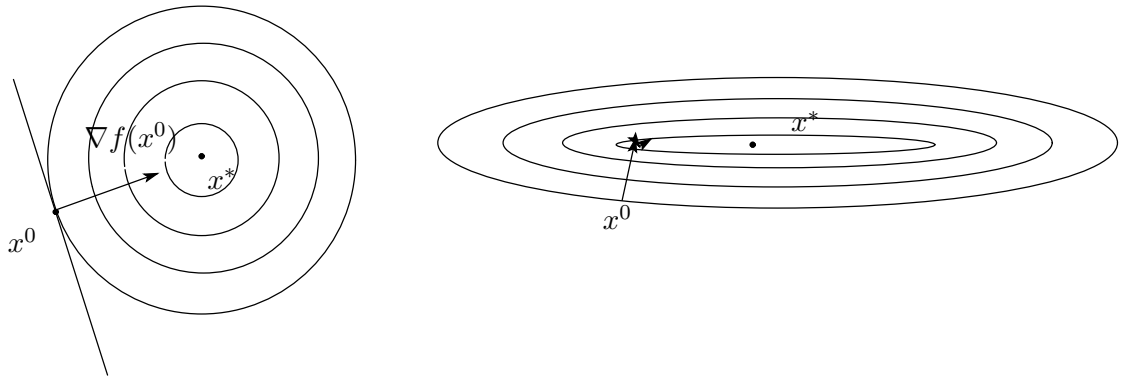
証明の前に述べたように、生成された点列に集積点が存在したときに始めて、停留点が求まることが保証されている。このように集積点が求める十分条件として、

$$\lim_{\|x\|\to\infty} f(x) = +\infty$$

がある (7.3)。この性質を強圧性という。さらに、停留点が大域的最小解となる十分条件は f が凸関数であることである。このことをまとめると以下の系を得る。

系 1 降下法に生成される点列が定理 *1* の仮定を満たしているとする。このとき、f が強圧的であれば、集積点が存在し、その集積点は停留点である。さらに f が凸関数であれば、大域的最小解である。

定理 1 の仮定を満たすように降下方向を選んでやれば，問題の停留点 (多くの場合局所最小解) を求めることができることがわかる．次の節で紹介する最急降下法，ニュートン法，準ニュートン法は，この定理の仮定をみたす降下方向を与える手法である．ここでは，降下法の中でも，特に重要な**最急降下法**，**ニュートン法**，**準ニュートン法**の説明をおこなう．



条件数が 1 に近い　　条件数が 0 に近い

図 7.4: 最急降下法と条件数

7.2　降下法の例

7.2.1　最急降下法

最急降下法は，降下方向として，目的関数の勾配の反対方向、

$$[\textbf{最急降下方向}] \quad d^k := -\nabla f(x^k)$$

を用いる手法である．この方向を最急降下方向と呼ぶ。最急降下方向と呼ばれるのは、この方向が次の最小化問題の解になっているからである。(方向が一致するだけで大きさは同じにならない。)

$$\begin{array}{ll} \min & \nabla f(x^k)^T d \\ \text{subjectto} & \|d\| = 1 \end{array}$$

最急降下方向は、

$$\nabla f(x^k)^T d^k = -\|\nabla f(x^k)\|^2$$

となるので，定理 1 の仮定をみたす．そのため、最急降下方向を用いた降下法は大域的収束する。しかしながら、最急降下法の収束はそれほど速くない。ここで， 収束する停留点 x^* のヘシアンは、最適性の 2 次の必要条件より半正定値行列となる。この最大固有値の最小固有値を $\kappa_{\max}$, $\kappa_{\min}$ としたとき、$\kappa := \kappa_{\min}/\kappa_{\max}$ をヘシアンの条件数と呼ぶ。最急降下法は、十分大きい k に対して、

$$\|x^{k+1} - x^*\| \leq \tau \|x^k - x^*\|$$

となる $\tau \in (0,1)$ が存在する、つまり、1 次収束することが示されている．そして、この定数 τ は $\tau \approx 1 - \kappa$ となる。(条件数は必ず 1 以下になることに注意しよう。) そのため、条件数が 1 に近いときは速く収束するが、小さいときはほとんど収束しない (図 7.4)。最急降下法の性質をまとめると以下のようになる。

良い点 1: 目的関数の勾配のみ必要で、ヘシアンは必要ない。

良い点 2: 大域的収束性がある。

悪い点 1: 一般に収束が遅い。

ここで、$n \times n$ の対称かつ正則な行列 B を用いて、$y = Bx$ という変数変換を考えよう。B が正則であるため、$x = B^{-1}y$ である。この変数変換を問題 (7.1) に施すと、

$$\min \quad f(B^{-1}y)$$

という y に対する問題なる。この問題の目的関数は $\hat{f}(y) = f(B^{-1}y)$ である。この $\hat{f}$ に対して最急降下法を実行することを考える。解 $y^* = Bx^*$ におけるヘシアンは、行列 B およびヘシアンが対称行列であることに注意すると、

$$\nabla \hat{f}(y^*) = \nabla f(B^{-1}y^*)B^{-1} = \nabla f(x^*)B^{-1}$$

となる。そのため、この関数における最急降下法の速さは、行列 $\nabla f(x^*)B^{-1}$ の条件数に依存する。もし、$B = \nabla f(x^*)$ であれば、$\nabla f(x^*)B^{-1}$ は単位行列となるため、条件数は 1 となり、非常に速い収束が期待できる。もちろん、$\nabla f(x^*)$ をあらかじめ知ることは難しいので、このような変数変換を行なって高速化を試みるときは、何らかの手法で $\nabla f(x^*)$ を推測する必要がある。

7.2.2 ニュートン法

テーラー展開を用いて関数 f の近似を行うとき，1 次までで近似した関数

$$\hat{f}(x) := f(x^k) + \nabla f(x^k)^T(x - x^k)$$

よりも 2 次の情報を用いて近似した関数

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

の方がよりよい近似となっている．そのため 2 次関数 $\tilde{f}$ の最小点 $\tilde{x}$ は，元の関数 f の最小点のよい近似解になっていると考えられる．このよう点 $\tilde{x}$ を与える探索方向は，1 次の必要条件より，

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

となる．この探索方向を用いる手法がニュートン法である．

[ニュートン方向] $\quad d^k := -\nabla^2 f(x^k)^{-1}\nabla f(x^k)$

なお、最急降下法の最後ところで述べた変数変換 $y = Bx$ を行なう最急降下法に対して、各反復において $B = \nabla^2 f(x^k)$ としたものが、ニュートン法となっている。

ここで，$\nabla^2 f(x)^{-1}$ が一様に正定値である，つまり，すべての $x \in R^n$ に対して

$$v^T\nabla^2 f(x)^{-1}v \geq \mu\|v\|^2 \ \ \forall v \in R^n$$

となる正の定数 μ が存在するとしよう．このとき，すべての k に対して，

$$\begin{aligned} \nabla f(x^k)^T d^k &= -\nabla f(x^k)^T\nabla^2 f(x^k)^{-1}\nabla f(x^k) \\ &\leq -\mu\|\nabla f(x^k)\|^2 \end{aligned}$$

となり，定理 1 の仮定を満たしていることがわかる．しかしながら，$\nabla^2 f(x)^{-1}$ が一様に正定値であるという仮定はかなり厳しい条件である．そのため，問題によっては、定理 1 の仮定を満たされず、大域的収束性が保証されない場合がある。また、さらに悪いことに、ニュートン方向

が降下方向にならない場合もある。例えば、決定変数が 1 次元の目的関数 $f(x) = x^4 - 2x^2$ の最小化を考えてみよう。このとき、$\nabla f(x) = 4x^3 - 4x$ であり、$\nabla^2 f(x) = 12x - 4$ である。今、$x^0 = 1/4$ とすると、$\nabla f(1/4) = -15/16$, $\nabla^2 f(1/4) = -1$ である。このため、ニュートン方向は $d = -\nabla^2 f(1/4)^{-1} \nabla f(1/4) = -15/16$ となり、$d^T \nabla f(1/4) = (15/16)^2 > 0$ となって、降下方向にすらならない。

一方ニュートン法の収束率は，次のように解析することができる．まず，この最小化問題の解を x^* としよう．さらに，$G(x) \equiv \nabla f(x)$ と定義したとき，G は微分可能かつ G' は局所的リプシッツ連続としよう．このとき，

$$\|G'(x)(x - x^*) - G(x) + G(x^*)\| = O(\|x - x^*\|^2) \tag{7.7}$$

が成り立つ．一方，$G'(x)$ が x^* において正定値行列であれば，

$$\|G'(x)\| \leq C \ \ \forall x \in N \tag{7.8}$$

を満たす正の定数 C が存在する．これらの性質 (7.7), (7.8) のもとで，ニュートン法の反復 $x^{k+1} = x^k - G'(x^k)^{-1} G(x^k)$ は，x^k が解 x^* に十分近いとき

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|x^k - x^* - G'(x^k)^{-1}(G(x^k) - G(x^*))\| \\ &\leq \|G'(x^k)^{-1}\| \|G'(x^k)(x^k - x^*) - G(x^k) + G(x^*)\| \\ &= O(\|x^k - x^*\|^2) \end{aligned}$$

を満たす．よって，ニュートン法が 2 次収束することがわかる．

ニュートン法において，ニュートン方向を求めるためには、線形方程式を解く必要がある。一般に、線形方程式の解を求めるためには、$O(n^3)$ かかることが知られている．これは，問題の規模が 10 倍になれば，その計算コストが 1000 倍になることを意味しており，大規模な問題において有効ではない．

以上のことより、ニュートン法の特徴は以下のようにまとめられる。

良い点 1: 2 次収束する。

悪い点 1: 大域的収束が保証されない。

悪い点 2: 各反復に線形方程式を解く必要がある。

7.2.3　準ニュートン法

準ニュートン法は，局所的に高速な収束と，大域的収束をあわせ持った手法である．また、各反復において線形方程式を解く必要がなく、反復点の計算が高速である。

まず，正定値対称行列 B_k が与えれているとき，次の方向は降下方向になることに注意しよう．

$$d = -B_k^{-1} \nabla f(x^k) \tag{7.9}$$

実際、$\nabla f(x^k) \neq 0$ であれば、

$$d^T \nabla f(x^k) = -\nabla f(x^k)^T B_k^{-1} \nabla f(x^k) < 0$$

となる。この方向 (7.9) は，B_k が $\nabla^2 f(x^k)$ に近似できていたら、ニュートン方向とほとんど同じになるため、高速な収束が期待できる．今、$\nabla f(x)$ のテーラー展開を考えると，

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

となるので，B_k もこの条件をみたすようにすることを考える．つまり、

$$\begin{aligned} s^k &= x^{k+1} - x^k \\ y^k &= \nabla f(x^{k+1}) - \nabla f(x^k) \end{aligned}$$

としたとき，

$$y^k = B_k s^k$$

となるようにする．この条件をセカント条件と呼ぶ。もちろん，この条件だけでは，B_k を 1 つに特定することはできない．そのため，これまでに数々の B_k の更新規則が提案されている．次の更新規則は **BFGS (Broyden-Fletcher-Goldfarb-Shanno) 公式**と呼ばれ，$O(n^2)$ の計算量で計算できる．

BFGS 公式：

$$B_{k+1} = B_k - \frac{B_k s_k (B_k s_k)^T}{(s^k)^T B_k s_k} + \frac{y^k (y^k)^T}{(s^k)^T y^k}$$

ここで，方向 d^k を求めるためには，B_k^{-1} を計算する必要があるが，

$$H_k = B_k^{-1}$$

としたとき，Sherman-Morrison の公式[1] を使うことによって，

$$H_{k+1} = H_k - \frac{H_k y^k (s^k)^T + s^k (H_k y^k)^T}{(s^k)^T y^k} + \left(1 + \frac{(y^k)^T H_k y^k}{(s^k)^T y^k}\right) \frac{s^k (s^k)^T}{(s^k) y^k} \tag{7.10}$$

とすることができる．そのため，B_0^{-1} を求めておけば，各反復で B_k^{-1} を求める必要がない．つまり $O(n^2)$ で降下方向 d^k を求めることができる．

準ニュートン法では、B_0 が $\nabla^2 f(x^*)$ に十分近く、$\nabla^2 f(x^*)$ が正則であるときに超 1 次収束することが示されている。しかし、あらかじめ $\nabla^2 f(x^*)$ を推定することは難しいので、$B_0 = I$ とすることが多い。そのため、必ずしも超一次収束するとは限らない．そこで，B_k をなるべく $\nabla^2 f(x^k)$ 近づけるために、例えば、一定回数ごとに $B_k = \nabla^2 f(x^k)$ とするなどの、工夫が必要となる．

今までは準ニュートン法の良い点ばかりを説明してきた。準ニュートン法の欠点として、計算容量、つまりメモリーがたくさん必要になるということがあげられる。次元 n が大きい大規模な問題では、目的関数のヘシアンが疎の行列になることが多い[2] . 行列が疎な場合、0 のところを記憶しないようにするデータ構造を使うことによって、計算容量を小さいままにすることができる。しかしながら、準ニュートン法で用いる行列 B_k は、例えヘシアンが疎な行列であっても、疎な行列にならない。そのため、n^2 のデータ容量が必ず必要となる。これは、問題の決定変数の数が多いときには、その問題を解くことができるコンピュータが限られてくることを意味している。

[1] A を $n \times n$ の正則行列，u, v を n 次元ベクトルとする．このとき，$1 + v^T A^{-1} u \neq 0$ であれば，以下の式が成り立つ．

$$(A + uv^T) = A^{-1} - \frac{1}{1 + v^T A^{-1} u} A^{-1} u v^T A^{-1}.$$

[2] 疎の行列とは, 要素に 0 が多い行列である。

以上のことより、準ニュートン法の特徴をまとめると以下のようになる。

良い点 1：　各反復が早い。

良い点 2： 大域的収束性がある。

良い点 3： B_0 を適切に選べば超 1 次収束する。

悪い点 1： B_0 が悪ければ、収束が遅い。

悪い点 2：　大規模な問題において、すべて要素がつまった B_k を用意しなければならない。(計算容量が大きくなる)