0. はじめに

このシケプリでは、$\nabla f(\mathbf{x})^t$ のことを $f'(\mathbf{x})$ と表記してあります。これは、手持ちの数学の教科書でこのようになっていた（実際はベクトルが太字にさえなっていなかった）こともありますが、僕が妥協したことが一番の原因だったりします。ご了承下さい。また、1. は微分の復習なので読み飛ばしても問題ないと思います。

1. 微分の復習

1. では $f(\mathbf{x}+t\mathbf{d}) = f(\mathbf{x}) + tf'(\mathbf{x})\mathbf{d} + \frac{t^2}{2}\mathbf{d}^t\nabla^2\mathbf{f}(\mathbf{x})\mathbf{d} + O(t^3)$ の導出をメインとしています。これを導出するために、教科書の丸写しみたいなことをしてページ数を増やしてしまいました。ただ先生の授業とプリントが、シケプリなんていらないくらい丁寧で、序盤で唯一不親切な気がしたのが上の式の導出だったので … と言い訳。

定義 関数 $f : \mathbf{R} \to \mathbf{R}$ が点 x で微分可能とは次式を満たす $c \in \mathbf{R}$ が存在することである。

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = c \tag{1}$$

この時 c を f の x における**微分係数**または**導値**といい、$c = f'(a)$ と表す。

これを今更書く必要は無かったかもしれませんが、一応挙げておきました。次にこれを多変数関数に拡張します。

定義 関数 $f : \mathbf{R}^n \to \mathbf{R}$ が点 $\mathbf{x}$ で**微分可能**とは次式を満たす定ベクトル $\mathbf{c} = (c_1, ..., c_n) \in \mathbf{R}^n$ が存在することである。

$$\lim_{|h|\to 0} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \mathbf{ch}}{|h|} = 0 \tag{2}$$

この時 $\mathbf{c}$ を f の $\mathbf{x}$ における**微分係数**または**導値**といい、$\mathbf{c} = f'(\mathbf{x})$（授業でいうところの $\nabla f(\mathbf{x})$ です）と表す。

$\mathbf{c}$ だけ $\mathbf{x}$ や $\mathbf{h}$ と違って横ベクトルです。また $\mathbf{h}$ の方向によらず、その大きさが 0 に向かってくれさえすれば c がただ一つに決まることが、多変数関数の場合の微分のポイントです。

定理 f が $\mathbf{x}$ で微分可能である時、f は $\mathbf{x}$ で各座標について偏微分可能で、

$$f'(\mathbf{x}) = (\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n}) \tag{3}$$

が成り立つ。証明略。

数理科学　でやったと思いますが、$f'(\mathbf{x}) = (\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n})$ のように「定義」しないのは、仮に $(\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_n})$ が存在したとしても、(2) 式の意味で微分可能かどうかは不明だからです。このことから、微分可能のことをあえて**全微分可能**と表現することもあるようです。ただし、最適化手法で出てくるような関数はほとんど全微分可能なので、いきなりこの定理によって微分係数を与えてもさほど問題ないものと思います。

次の定義と定理はあまり最適化手法とあまり関係ありませんが、その次の連鎖律のために必要なので乗せておきました。

定義 関数$\mathbf{f}:\mathbf{R}^n\to\mathbf{R}^m(\mathbf{f}=(f_1,...,f_m)^t)$が点$\mathbf{x}$で微分可能とは次式を満たす行列$M\in\mathbf{R}^{m\times n}$が存在することである。

$$\lim_{|h|\to 0}\frac{\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-M\mathbf{h}}{|h|}=0 \tag{4}$$

この時Mを$\mathbf{f}$の$\mathbf{x}$における**微分係数**または**導値**といい、$M=\mathbf{f}'(\mathbf{x})$と表す。

定理 $\mathbf{f}$が$\mathbf{x}$で微分可能である時、$f_i(i=1,...,m)$は$\mathbf{x}$で各座標について偏微分可能で、

$$\mathbf{f}'(\mathbf{x})=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n}\end{bmatrix} \tag{5}$$

が成り立つ。証明略。

定理 (連鎖律) 二つの関数$g:\mathbf{R}^n\to\mathbf{R}^m$と$f:\mathbf{R}^m\to\mathbf{R}^p$が合成可能で両方とも微分可能であるとき、その合成関数の微分に関して次式が成り立つ。

$$(\mathbf{f}\circ\mathbf{g})'(\mathbf{x})=\mathbf{f}'(\mathbf{y})\mathbf{g}'(\mathbf{x}),(\mathbf{y}=\mathbf{g}(\mathbf{x}) \tag{6}$$

証明略

合成関数の微分が合成される関数の微分（この場合行列）の積で表されるということです。この定理は、名前こそあまり知られていないものの、非常に使用頻度が高いと思われます。

次に$f:\mathbf{R}^n\to\mathbf{R}$に関して、$f(\mathbf{x}+t\mathbf{d})$を$t$の関数と見た時（他を定ベクトルと考えた時）の微分を考えます。この関数は$g(t):\mathbf{R}\to\mathbf{R}$ですから、形は非常になじみのある関数です。実際$t$で微分してみると、

$$g'(t)=f'(\mathbf{x}+t\mathbf{d})\mathbf{d} \tag{7}$$

となります。ここで既に連鎖律が使われています。(6)での$\mathbf{g}=\mathbf{x}+t\mathbf{d}$とし、$t$を引数にベクトルを返す関数だと考えています。そしてこれをtで微分すると(5)式よりn行1列行列として$\mathbf{d}$を得ます。次に二階の導関数を求めるわけですが、$f'(\mathbf{x}+t\mathbf{d})$は横ベクトルであるため、うまく微分できません。そこで転置します。$g'(t)=\mathbf{d}^t\nabla f(\mathbf{x}+t\mathbf{d})$として、これを行列の積と考えて微分します。$\mathbf{d}^t$は変化しないのでその後ろだけ考えればよくて、

$$g''(t)=\mathbf{d}^t\nabla^2\mathbf{f}(\mathbf{x}+t\mathbf{d})\mathbf{d} \tag{8}$$

となります。$\nabla^2\mathbf{f}(\mathbf{x})$はヘシアンですが、これの定義式をtexで書くのは心が折れてしまうので、ノートの方を参照下さい。この式変形で解かるとおりヘシアンは、$\nabla f'(\mathbf{x})$という多変数ベクトル値関数の、(4)式の意味での微分係数として現れます。二階までの導関数を求めたことにより、fを2次まで展開することができます。その結果は

$$\begin{aligned}f(\mathbf{x}+t\mathbf{d})=g(t)&=g(0)+tg'(0)+\frac{t^2}{2}g''(0)+O(t^3)\\&=f(\mathbf{x})+tf'(\mathbf{x})\mathbf{d}+\frac{t^2}{2}\mathbf{d}^t\nabla^2\mathbf{f}(\mathbf{x})\mathbf{d}+O(t^3)\end{aligned} \tag{9}$$

となります。
$g'(t) = \mathbf{d}^t \nabla f(\mathbf{x} + t\mathbf{d})$ をよく見ると、一階微分を表す演算子が、$(d_1 \frac{\partial}{\partial x_1} +, ..., d_n \frac{\partial}{\partial x_n})$ となっていることがわかります。よって、

$$g^{(n)}(t) = (d_1 \frac{\partial}{\partial x_1} +, ..., d_n \frac{\partial}{\partial x_n})^n f(\mathbf{x} + t\mathbf{d}) \tag{10}$$

を得ます。これは n 回微分でも簡単に表せる利点を持ちますが、(9) の形式だと関数の極値問題を行列の正定値問題に帰着できるので主にこちらを使います。

2. 無制約最適化問題

一応ここからが最適化手法の本題です。だいたいノートの順番に則って書いたので、ノートを取ってない人には役に立つかもしれませんが、しっかりノートを取った人には無用かもしれません … 。このシケプリはノートを取ってない人用の試験対策向けという感じで作ってあります。そのくせ、ノート参照とか出てきますが、そこらへんは友達に借りてください。

2.1 種々の言葉の定義

定義 関数 $f : \mathbf{R}^n \to \mathbf{R}$ に対して、$min f(\mathbf{x})$ を考えるような問題を**無制約最適化問題**という。
$f(\mathbf{x}) \geq f(\mathbf{x}^*), \forall \mathbf{x} \in \mathbf{R}$ であるとき、$\mathbf{x}^*$ を**大域的最適解**という。
$f(\mathbf{x}) \geq f(\mathbf{x}^*), \forall \mathbf{x} \in N(\mathbf{x}^*)$ であるとき、$\mathbf{x}^*$ を**局所最適解**という。ただし、$N(\mathbf{x}^*)$ とは、$\mathbf{x}^*$ の近傍である。
$min\{f(\mathbf{x}) | \mathbf{x} \in \mathbf{R}\}$ を f の**最適値**といい、f^* と表す。
$f'(\mathbf{x}) = \mathbf{0}$ のとき、$\mathbf{x}$ を**停留解**という。
極小でも極大でもない停留解を**鞍点**という。

定理というほどのものではありませんが、$\mathbf{x}^*$ が大域的最適解 $\Leftrightarrow f(\mathbf{x}^*) = f^*$ です。

例 $\mathbf{g} : \mathbf{R}^n \to \mathbf{R}^n$ に対して $\mathbf{g}(\mathbf{x}) = 0$ の解を求めることは、無制約最適化問題 $min \mathbf{g}(\mathbf{x})^t \mathbf{g}(\mathbf{x})$ に帰着される。

例 ペナルティ法　ノート参照。

定義 対称行列 A に対して、A の固有値が全て 0 より大きいとき、A を**正定値**といい、$A \succ 0$ と表す。$(\mathbf{x}^t A \mathbf{x} > 0, (\forall \mathbf{x} \neq \mathbf{0})$ と同値)
対称行列 A に対して、A の固有値が全て 0 以上のとき、A を**半正定値**といい、$A \succeq 0$ と表す。$(\mathbf{x}^t A \mathbf{x} \geq 0, (\forall \mathbf{x})$ と同値)
括弧内のことは証明が必要ですが省略します。また、ノートに正定値行列に関する公式があるので見ておいて下さい。(例えば、「A が正定値ならば逆行列が存在する」など)

定理 f が微分可能であるとき、$\mathbf{x} \in \mathbf{R}$ が（局所）最適解 $\Rightarrow f'(\mathbf{x}) = \mathbf{0}$
f が二回微分可能であるとき、$\mathbf{x} \in \mathbf{R}$ が（局所）最適解 $\Rightarrow f'(\mathbf{x}) = \mathbf{0}, \nabla^2 f(\mathbf{x})$ は半正定値
これらの定理は一般に逆は成り立ちません。

定理 f が二回微分可能で、$f'(\mathbf{x}) = \mathbf{0}, \nabla^2 f(\mathbf{x}) \succ 0 \Leftrightarrow \mathbf{x}$ は f の局所最適解。

こちらは、最適解の十分条件です。

定義 $\mathbf{x}_{k+1} = \mathbf{g}_k(\mathbf{x}_k)$ として点列 $\{\mathbf{x}_k\}$ を作るようなアルゴリズムを**反復法**と呼ぶ。
基礎的な反復法は、$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ の形をしています。

2.2 最急降下法

以下では $f(\mathbf{x}_k) = f_k, f'(\mathbf{x}_k) = f_k$ と置いています。アルゴリズムについては、プリント「最急降下法の概要」に詳しくでているのでそちらをご覧下さい。最急降下法の $\mathbf{d_k}$ は f の $\mathbf{x}_k$ におけるテイラー展開による 1 次モデル

$$f_k^L(\mathbf{x}_k + \mathbf{d}) = f_k + f_k'\mathbf{d} \tag{11}$$

((9) 式で $t = 1$ と置いて 2 次以下を無視する。）を最小にする向きになります。また、$\mathbf{x}_k$ が解に近づくにつれて $\mathbf{d}_k$ も勝手に小さくなってくれますが、だからといって α_k を定数と置いてしまうと収束しません。α_k の求め方は「正確な直線探索」「Armijo の基準」「Wolfe の基準」の 3 種類あります。「正確な直線探索」はプリントに出ています。後の 2 つは特に覚えなくても多分問題は無いと思います。一応ノートに書いてあるはずです。

2.3Newton 法

テイラー展開による f の 2 次モデルは、

$$f_k^Q(\mathbf{x}_k + \mathbf{d}) = f_k + f_k'\mathbf{d} + \frac{1}{2}\mathbf{d}^t\nabla^2\mathbf{f}_k\mathbf{d} \tag{12}$$

((9) 式で $t = 1$ と置いて 3 次以下を無視する。）となります。ここで、求めなくてはいけないのは $\mathbf{d}$ なので、$\mathbf{d}$ を変数と見た勾配 ∇ を考えます。そしてこの微分のベクトルが $\mathbf{0}$ になったらそれを $\mathbf{d}_k$ とします。まず、f_k は $\mathbf{d}$ に関係ありません。$f_k'\mathbf{d}$ の $\mathbf{d}$ に関する勾配ベクトルは ∇f_k に、$\frac{1}{2}\mathbf{d}^t\nabla^2\mathbf{f}_k\mathbf{d}$ は $\nabla^2\mathbf{f}_k\mathbf{d}$ となります（計算過程略)。これにより、$\mathbf{d}$ による勾配が $\mathbf{0}$ になることの条件として Newton 方程式

$$\nabla^2\mathbf{f}_k\mathbf{d} = -\nabla f_k \tag{13}$$

を得ます。Newton 方程式は解けるとは限らないし、減少しているかもわからないところが難点です。

定義 $\mathbf{d}$ が点 $\mathbf{x}$ における f の**降下方向**とは、$f'(\mathbf{x})\mathbf{d} < 0$ が成り立つことである。

定理 $\nabla^2\mathbf{f}_k \succ 0$ のとき、Newton 方程式は一意的に解けて、$\mathbf{d}_k$ は降下方向。（証明はノート参照）

2.4 収束の早さ

定義 $\{\mathbf{x}_k\}$ が $\mathbf{x}^*$ に **1 次収束**するとは、

$$\lim_{k\to\infty} \frac{||\mathbf{x}_{k+1} - \mathbf{x}^*||}{||\mathbf{x}_k - \mathbf{x}^*||} = M, 0 \leq M < 1 \tag{14}$$

が成り立つことをいう。**超 1 次収束**するとは

$$\lim_{k\to\infty}\frac{||\mathbf{x}_{k+1}-\mathbf{x}^*||}{||\mathbf{x}_k-\mathbf{x}^*||}=0 \tag{15}$$

が成り立つことをいう。**2 次収束**するとは、1 次収束するうえでさらに

$$\lim_{k\to\infty}\frac{||\mathbf{x}_{k+1}-\mathbf{x}^*||}{||\mathbf{x}_k-\mathbf{x}^*||^2}=M\geq 0 \tag{16}$$

が成り立つことをいう。

定理 最急降下法は 1 次収束（プリント「最急降下法の収束速度について」参照)。Newton 法は 2 次収束（ノート参照）

2.5 準 Newton 法

準 Newton 法では 2 次モデルを

$$\tilde{f}_k^Q(\mathbf{x}_k+\mathbf{d})=f_k+f_k'\mathbf{d}+\frac{1}{2}\mathbf{d}^tB_k\mathbf{d} \tag{17}$$

とします。テイラー展開による 2 次モデルとは違うものなのでチルダがついています。ここで、B_k の条件の 1 つは正定値となることです。B_0 の条件はこれだけです。この条件が満たされれば、Newton 方程式によく似た式

$$B_k\mathbf{d}=-\nabla f_k \tag{18}$$

の解として、降下方向のベクトル $\mathbf{d}_k$ を得ることができます。次に $k+1$ での 2 次モデル

$$\tilde{f}_{k+1}^Q(\mathbf{x}_{k+1}+\mathbf{d})=f_{k+1}+f_{k+1}'\mathbf{d}+\frac{1}{2}\mathbf{d}^tB_{k+1}\mathbf{d} \tag{19}$$

を考えます。これを変数を $\mathbf{d}$ と考えて勾配を取ると、Newton 方程式を導出したときと同じ計算により、$\nabla f_{k+1}+B_{k+1}\mathbf{d}$ を得ます。これが、$\mathbf{d}=\mathbf{0}, -\mathbf{x}_{k+1}+\mathbf{x}_k$ のときそれぞれ、$\nabla f_{k+1}, \nabla f_k$ に一致してくれれば 2 次モデルは精度の高いものといえます（$\nabla\tilde{f}_{k+1}^Q$ と ∇f が 2 点 $\mathbf{x}_{k+1}, \mathbf{x}_k$ で等しくなるため)。前者は明らかに成り立ちますが、後者が成り立つことは

$$B_{k+1}(\mathbf{x}_{k+1}-\mathbf{x}_k)=\nabla f_{k+1}-\nabla f_k \tag{20}$$

と同値です。上の式であたえられる条件ををセカント条件といいます。B_1 以下では正定値であることの他に、この条件がつきます。これらの条件を満たすように $\{\mathbf{B}_k\}$ を作るやりかたはいくつか知られていて、BFGS 公式、DFP 公式などと呼ばれています。公式はプリント「準 Newton 法の概要」に載っています。さらに、(18) 式を解くのは手間がかかるので、$\{\mathbf{B}_k\}$ を計算するのではなく、その逆行列 $\{\mathbf{H}_k\}$ を計算しようという公式を H 公式と呼びます。多分こちらの方が実用的です。

定理 準 Newton 法は超 1 次収束する。（プリント「準 Newton 法の超 1 次収束性」参照）

2.6 信頼領域法

信頼領域法の 2 次モデルは

$$f_k^Q(\mathbf{x}_k + \mathbf{d}) = f_k + f_k'\mathbf{d} + \frac{1}{2}\mathbf{d}^t B_k \mathbf{d} \tag{21}$$

となります。（さっきはテイラー展開と違うとの理由でチルダをつけましたが今回は省略です。）そして更新公式は、$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$ となります。$\mathbf{d}_k$ は制約条件付き最適化問題

$$min f_k^Q(\mathbf{x}_k + \mathbf{d}) s.t. ||\mathbf{d}|| \leq \Delta k \tag{22}$$

の最適解として与えられます。（条件式が閉空間を作っているので、必ず最適解を持ちます。）$\mathbf{d}_*$ が部分問題の最適解であることの必要条件がノートに載っているので、そちらを参照してみて下さい。

$\{B_k\}$の求め方は手持ちのノートに書いていなかったのですが、おそらく準 Newton 法と同じように求めるのだと思います（このようにすれば、B_k は正定値です）。そして、まず領域の内部に解があると仮定して、$-B_k^{-1}\nabla f_k$ を求め（B_k が正定値なら無制約下での最適解はこれ 1 つ）、それがちゃんと条件の領域の内部にあったらそれを $\mathbf{d}_k$ とし、領域から出てしまったら部分問題の解は領域上にあるので、ラグランジュ未定乗数法で領域上の解を求めるのだと思います。

また信頼領域法では、求めた $\mathbf{d}_k$ が妥当かどうかチェックするようです。もし妥当じゃないと判定されたら、Δk を小さくしてやり直すようです。これについてはノートを参照して下さい。

2.7 凸関数

定義 関数 $f : \mathbf{R}^n \to \mathbf{R}$ で次の条件をみたすものを**凸関数**という。

$$\forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, \forall \lambda \in [0,1] \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \tag{23}$$

さらに次の条件をみたすものを**狭義凸関数**という。

$$\forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, \forall \lambda \in (0,1) \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) > f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \tag{24}$$

定理 f が微分可能であるとき、f が凸であることの必要十分条件は

$$\forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, f(\mathbf{y}) \geq f(\mathbf{x}) + f'(\mathbf{x}) + (\mathbf{y} - \mathbf{x}) \tag{25}$$

f が微分可能な凸関数のとき、$\nabla f(\mathbf{x}) = \mathbf{0} \Rightarrow \mathbf{x}$ は f の大域的最適解
f が 2 回連続微分可能のとき、f が凸関数 $\Leftrightarrow \nabla^2 f\mathbf{x} \succeq 0 (\forall \mathbf{x})$

例　2 次関数　$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^t Q \mathbf{x} + \mathbf{p}^t \mathbf{x} + r$ において f が凸であることと Q が半正定値であることは同値

ところで、上の定理において凸を狭義凸に、$\geq, \succeq$ を $>, \succ$ に変えたものは少し定理が必要なはずで、2 回連続微分可能で f が凸関数でも、$\nabla^2 f\mathbf{x} \succ 0 (\forall \mathbf{x})$ は言えないはずです。それ以外は問題ないと思います。