近接勾配法

(3.5) 式

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2 \tag{3.5}$$

の最小化問題に対するアプローチとして，(3.10) 式，(3.11) 式

$$\gamma := \beta + \nu X^T(y - X\beta) \tag{3.10}$$

$$\beta = \left(1 - \frac{\nu\lambda}{\|\gamma\|_2}\right)_+ \gamma \tag{3.11}$$

を繰り返すことを前節で考えたが，一般には上式が (3.5) 式の最適解に収束することは保証されない。以下では $\nu > 0$ の値を上手く設定することによって (3.10),(3.11) 式のループによって正しい解に収束することを示す.

まず $\lambda = 0$ の場合であれば (3.11) 式において

$$\beta = \gamma$$

となるため，(3.10) 式と (3.11) 式のループは

$$\beta_{t+1} \leftarrow \beta_t + \nu X^T(y - X\beta_t)\gamma$$

とかけることに注意したい．ここで, $g(\beta) = \|y - X\beta\|_2^2$ と定めれば $\nabla g(\beta) = -X^T(y - X\beta)$ とかけることより，上式は

$$\beta_{t+1} \leftarrow \beta_t - \nu\nabla g(\beta_t)$$

と書くことにする．この方法は g の減少の大きい方向に β を更新していくという意味で，勾配法と呼ばれている.

以下では，(3.5) 式の第 2 項を $h(\beta) := \lambda\|\beta\|_2$ とおいて，$\lambda > 0$ での勾配法を

$$\beta_{t+1} \leftarrow \beta_t - \nu\{\nabla g(\beta_t) + \partial h(\beta)\} \tag{1}$$

とした場合の収束性について考えたい．ここで，関数 h に対して写像 $\mathrm{prox}_h : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$ を

$$\mathrm{prox}_h(z) := \mathrm{argmin}_{\theta \in \mathbb{R}^p} \left\{\frac{1}{2}\|z - \theta\|_2^2 + h(\theta)\right\}$$

とすることで，(1) 式は

$$\beta_{t+1} \leftarrow \mathrm{prox}_{\nu h}(\beta_t - \nu\nabla g(\beta_t))$$

とかけること，すなわち (1) 式で更新される β は

$$\varphi(\theta)=\frac{1}{2}\|\beta_t-\nu\nabla g(\beta_t)-\theta\|_2^2+\nu h(\beta)$$

を最小にする θ と一致することを以下で確認する．

最初に $\varphi(\theta)$ を最小にする θ' について考える；まず $\beta_t-\nu\nabla g(\beta_t)=0$ であるとき，$\varphi(\theta)=\frac{1}{2}\|\theta\|_2^2+\nu h(\theta)\geq 0$ であり，等号成立は $\theta=0$ のときに限ることより，求める θ' は $\theta'=0$ であることがわかる．

次に $\beta_t-\nu\nabla g(\beta_t)\neq 0$ であるとき，$\theta\neq 0$ における $\dfrac{\partial\varphi(\theta)}{\partial\theta}$ を求めると

$$\frac{\partial\varphi(\theta)}{\partial\theta}=-(\beta_t-\nu\nabla g(\beta_t)-\theta)+\frac{\nu\lambda\theta}{\|\theta\|_2}$$

となるので，$\varphi(\theta)$ を最小にする θ' は

$$\beta_t-\nu\nabla g(\beta_t)-\theta'=\frac{\nu\lambda\theta'}{\|\theta'\|_2} \tag{2}$$

を満たしている．ここで $\tilde{\beta}_t:=\beta_t-\nu\nabla g(\beta_t)$ とおくと，上式はある実数 $\alpha\neq 0$ が存在して θ' が

$$\theta'=\alpha\tilde{\beta}_t$$

とかけることを意味している．したがってこれを (2) 式に代入することで

$$\begin{aligned}\tilde{\beta}_t-\alpha\tilde{\beta}_t&=\frac{\nu\lambda\alpha}{|\alpha|\|\tilde{\beta}_t\|_2}\tilde{\beta}_t\\ \alpha\tilde{\beta}_t&=\left(1-\frac{\nu\lambda\alpha}{|\alpha|\|\tilde{\beta}_t\|_2}\right)\tilde{\beta}_t\end{aligned}$$

が成立することがわかり，$\tilde{\beta}_t\neq 0$ であることから上式の係数を比較して

$$\begin{aligned}\alpha&=1-\frac{\nu\lambda\alpha}{|\alpha|\|\tilde{\beta}_t\|_2}\\ &=1-\frac{\nu\lambda}{\|\tilde{\beta}_t\|}\cdot\mathrm{sign}(\alpha)\end{aligned}$$

となることがわかる．ここで，仮に $\alpha<0$ であるとすると

$$0>\alpha=1+\frac{\nu\lambda}{\|\tilde{\beta}_t\|}>0$$

となってしまい，矛盾が生じる．したがって $\alpha>0$ であることがわかるので

$$\alpha=1-\frac{\nu\lambda}{\|\tilde{\beta}_t\|}$$

となることがわかる．これを $\theta' = \alpha\tilde{\beta}_t$ に代入することで求める θ' は

$$\begin{aligned}
\theta' &= \left(1 - \frac{\nu\lambda}{\|\tilde{\beta}_t\|}\right)\tilde{\beta}_t \\
&= \tilde{\beta}_t - \frac{\nu\lambda}{\|\tilde{\beta}_t\|}\tilde{\beta}_t \\
&= \beta_t - \nu\nabla g(\beta_t) - \frac{\nu\lambda\tilde{\beta}_t}{\|\tilde{\beta}_t\|} \\
&= \beta_t - \nu\nabla g(\beta_t) - \nu\frac{\partial h}{\partial\theta}(\tilde{\beta}_t) \\
\therefore \theta' &= \beta_t - \nu\nabla g(\beta_t) - \nu\frac{\partial h}{\partial\theta}(\tilde{\beta}_t)
\end{aligned}$$

となることがわかる．
以上より $\varphi(\theta)$ を最小にする θ' は

$$\theta' = \begin{cases} \beta_t - \nu\left\{\nabla(\beta_t) - \dfrac{\partial h}{\partial\theta}(\tilde{\beta}_t)\right\} & (\beta_t - \nu\nabla(\beta_t) \neq 0 \text{ のとき}) \\ 0 & (\beta_t - \nu\nabla(\beta_t) = 0 \text{ のとき}) \end{cases}$$

と書くことができるが，関数 h の劣微分が 0 を含むことから改めて

$$\theta' \in \beta_t - \nu\left\{\nabla(\beta_t) - \partial h(\tilde{\beta}_t)\right\}$$

とかくことができ，これは (1) 式と一致することがわかる．

上記より，新たに提案された勾配法は

$$\beta_{t+1} \leftarrow \mathrm{prox}_{\nu h}(\beta_t - \nu\nabla g(\beta_t))$$

とかけることが確認できた．ここで $\mathrm{prox}_{\nu h}(\beta_t - \nu\nabla g(\beta_t))$ が実際には

$$\mathrm{prox}_{\nu h}(\beta_t - \nu\nabla g(\beta_t)) = \left(1 - \frac{\nu\lambda}{\|\tilde{\beta}_t\|}\right)_+ \tilde{\beta}_t$$

とかけていたことから更新式は

$$\begin{aligned}
&\gamma \leftarrow \beta - \nu\nabla g(\beta_t) = \beta + \nu X^T(y - X\beta) \\
&\beta \leftarrow \left(1 - \frac{\nu\lambda}{\|\gamma\|_2}\right)_+ \gamma
\end{aligned}$$

とかけることがわかる．これは (3.10),(3.11) の更新式に他ならない．

最後に ν を上手くとることができれば，上記の更新式によって $f(\beta)$ を最小にする β_* を求められることを示す.
まず X^TX の最大固有値を L とすれば，任意の $x, y, z \in \mathbb{R}^p$ に対して

$$(x-y)^T\nabla^2 g(z)(x-y) \le L\|x-y\|_2^2$$

が成立することを示す；X^TX が対称行列であることから，ある直交行列 P が存在して

$$P^T(X^TX)P = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}$$

が成立する．さらに L の定め方より，任意の $x = (x_1, \cdots, x_p)^T \in \mathbb{R}^p$ に対して I を p 次単位行列として

$$\begin{aligned} x^TP^TLIPx &= L \cdot x^TP^TPx \\ &= L\sum_{i=1}^{p} x_i^2 \ge \sum_{i=1}^{p} \lambda_i x_i^2 \\ &= x^T \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} x \\ &= x^TP^T(X^TX)Px \end{aligned}$$

が任意の $x \in \mathbb{R}^p$ に対して成立していることより，$x \to P^T(x-y)$ とおけば

$$\begin{aligned} (x-y)^T\nabla^2 g(z)(x-y) &= \left\{P^T(x-y)\right\}^T P^T(X^TX)P\left\{P^T(x-y)\right\} \\ &\le \left\{P^T(x-y)\right\}^T P^TLIP\left\{P^T(x-y)\right\}^T \\ &= L\|x-y\|_2^2 \\ \therefore (x-y)^T\nabla^2 g(z)(x-y) &\le L\|x-y\|_2^2 \end{aligned}$$

となり，求めたい不等式が導ける.

また，(3.13) 式について

$$\begin{aligned} \mathrm{prox}_{\nu h}(\beta_t - \nu\nabla g(\beta_t)) &= \mathrm{argmin}_{\theta\in\mathbb{R}^p}\left\{\frac{1}{2}\|\beta_t - \nu\nabla g(\beta_t) - \theta\|_2^2 + \nu h(\theta)\right\} \\ &= \mathrm{argmin}_{\theta\in\mathbb{R}^p}\left\{\frac{1}{2\nu}\|\beta_t - \nu\nabla g(\beta_t) - \theta\|_2^2 + h(\theta) + g(\beta_t)\right\} \\ &= \mathrm{argmin}_{\theta\in\mathbb{R}^p}\left\{\frac{1}{2\nu}\|\beta_t - \theta\|_2^2 + (\beta_t - \theta)^T\nabla g(\beta_t) + h(\theta) + g(\beta_t)\right\} \end{aligned}$$

が成立することより，$\nu = \dfrac{1}{L}$ とした (3.13) 式を

$$\begin{aligned} Q(x,y) &:= g(y) + (x-y)^T \nabla g(y) + \frac{L}{2}\|x-y\|_2^2 + h(x) \\ p(y) &:= \mathrm{argmin}_{x \in \mathbb{R}^p} Q(x,y) \end{aligned}$$

を用いて

$$\beta_{t+1} \leftarrow p(\beta_t) \tag{3}$$

と書きなおしておく．[*1]

以上で定めた定数 L，写像 Q, p に対し，次の命題 7 が成立する

命題 7

(3) 式によって生成された列 $\{\beta_t\}$ は，f の最小化問題に対する最適解を β_* として次式を満たす

$$f(\beta_k) - f(\beta_*) \le \frac{L\|\beta_0 - \beta_*\|_2^2}{2k}$$

証明には以下の補題 1 を用いる．

補題 1

任意の $x, y \in \mathbb{R}^p$ に対して

$$\frac{2}{L} f(x) - f(p(y)) \ge \|p(y) - y\|_2^2 + 2(y-x)^T(p(y) - y)$$

が成立する．

Proof. 自然数 k を任意にとって固定する．補題 1 で $x = \beta_*, y = \beta_t$ とおき，さらに $p(\beta) = \beta_{t+1}$ であることを用いれば

$$\begin{aligned} \frac{2}{L}\{f(\beta_*) - f(\beta_{t+1})\} &\ge \|\beta_{t+1} - \beta_t\|_2^2 + 2(\beta_t - \beta_*)^T(\beta_{t+1} - \beta_t) \\ &= \langle \beta_{t+1} - \beta_t, \beta_{t+1} - \beta_t + 2(\beta_t - \beta_*) \rangle \\ &= \langle (\beta_{t+1} - \beta_*) - (\beta_t - \beta_*), (\beta_{t+1} - \beta_*) + (\beta_t - \beta_*) \rangle \\ &= \|\beta_{t+1} - \beta_*\|_2^2 - \|\beta_t - \beta_*\|_2^2 \end{aligned}$$

が任意の自然数 $t \le k$ に対して成立することがわかる．したがって上の不等式を $t = 0, \cdots, k-1$ について足し合わせれば

$$\frac{2}{L}\left\{ k f(\beta_*) - \sum_{t=0}^{k-1} f(\beta_{t+1}) \right\} \ge \|\beta_* - \beta_k\|_2^2 - \|\beta_* - \beta_0\|_2^2 \tag{4}$$

[*1] この方法は ISTA(Iterative Shrinkage-Thresholding Algorithm) と呼ばれる

が成立することがわかる．さらに $x=y=\beta_t$ として再び補題 1 を適用すると

$$\frac{2}{L}\left\{f(\beta_t)-f(\beta_{t+1})\right\} \geq \|\beta_{t+1}-\beta_t\|_2^2$$

となることより，上式の両辺を t 倍し，さらに $t=0,\cdots,k-1$ について加えることで

$$\begin{aligned}\frac{2}{L}\sum_{t=0}^{k-1} t\left\{f(\beta_t)-f(\beta_{t+1})\right\} &= \frac{2}{L}\sum_{t=0}^{k-1}\left\{tf(\beta_t)-(t+1)f(\beta_{t+1})+f(\beta_{t+1})\right\} \qquad (5)\\ &= \frac{2}{L}\left\{-kf(\beta_k)+\sum_{t=0}^{k-1} f(\beta_{t+1})\right\} \geq \sum_{t=0}^{k-1} t\|\beta_t-\beta_{t+1}\|_2^2 \qquad (6)\end{aligned}$$

が成立することがわかる．以上で得られた不等式 (4),(6) の両辺を加えることで

$$\begin{aligned}\frac{2k}{L}\{f(\beta_*)-f(\beta_k)\} &\geq \|\beta_*-\beta_k\|_2^2-\|\beta_*-\beta_0\|_2^2+\sum_{t=0}^{k-1} t\|\beta_t-\beta_{t+1}\|_2^2\\ &\geq -\|\beta_*-\beta_0\|_2^2\\ \therefore f(\beta_k)-f(\beta_*) &\leq \frac{L\|\beta_0-\beta_*\|_2^2}{2k}\end{aligned}$$

が得られる．これが求めたかった不等式である． □

ISTA の修正アルゴリズムとして，次を考える；実数列 $\{\alpha_t\}$ を

$$\alpha_1=1, \quad \alpha_{t+1}:=\frac{1+\sqrt{1+4\alpha_t^2}}{2}$$

によって定め，さらに初期値 β_0 に対して $\gamma_1=\beta_0$ として更新式を

$$\begin{aligned}\beta_t &\leftarrow p(\gamma_t)\\ \gamma_{t+1} &\leftarrow \beta_t+\frac{\alpha_t-1}{\alpha_{t+1}}(\beta_t-\beta_{t-1})\end{aligned}$$

によって定める．この手法は FISTA[*2] と呼ばれ，得られる列 $\{\beta_t\}$ は次の命題 8 を満たす

命題 8

FISTA によって生成された数列 $\{\beta_t\}$ は，f の最小化問題に対する最適解を β_* として次式を満たす

$$f(\beta_k)-f(\beta_*) \leq \frac{4L\|\beta_0-\beta_*\|_2^2}{(k+1)^2}$$

証明には以下の補題 2,3,4 を用いる

[*2] Fast Iterative Shrinkage-Thresholding Algorithm

補題 2

FISTA で考える列 $\{\alpha_k\}$ と得られる列 $\{\beta_k\}$，及び最適解 β_* に対して次式が成立する

$$\frac{2}{L}\left\{\alpha_k^2(f(\beta_k)-f(\beta_*))-\alpha_{k+1}^2(f(\beta_{k+1})-f(\beta_*))\right\}\geq t_{k+1}-t_k$$

ただし $t_k := \|\alpha_k\beta_k-(\alpha_k-1)\beta_{k-1}-\beta_*\|_2^2$ とする．

補題 3

ある実数 c が存在して，2 つの非負値実数列 $\{a_k, b_k\}$ が任意の自然数 $k\geq 1$ について

$$a_k-a_{k+1}\geq b_{k+1}-b_k \text{ かつ } a_1+b_1\leq c$$

を満たすならば，任意の自然数 k に対して $a_k\leq c$ が成立する．

補題 4

FISTA で考える列 $\{\alpha_k\}$ は任意の自然数 $k\geq 1$ に対して

$$\alpha_k\geq\frac{k+1}{2}$$

が成立する．

Proof. 補題 2 より

$$\frac{2}{L}\alpha_k^2(f(\beta_k)-f(\beta_*))-\frac{2}{L}\alpha_{k+1}^2(f(\beta_{k+1})-f(\beta_*))\geq t_{k+1}-t_k$$

が成立する．ここで，$s_k := \frac{2}{L}\alpha_k^2(f(\beta_k)-f(\beta_*))$ とすることで上式は

$$s_k-s_{k+1}\geq t_{k+1}-t_k \tag{7}$$

と書き直せる．さらに $\{s_k\},\{t_k\}$ は共に非負値実数列で (7) 式及び

$$s_1+t_1\leq 2\|\beta_0-\beta_*\|_2^2$$

を満たすので，$c=2\|\beta_0-\beta_*\|_2^2$ として補題 3 を適用すれば

$$s_k\leq c\Leftrightarrow\frac{2}{L}\alpha_k^2(f(\beta_k)-f(\beta_*))\leq 2\|\beta_0-\beta_*\|_2^2$$

が得られる．上式に補題 4 を適用すれば，任意の自然数 k に対して

$$\begin{aligned}f(\beta_k)-f(\beta_*)&\leq L\|\beta_0-\beta_*\|_2^2\cdot\frac{1}{\alpha_k^2}\\&\leq L\|\beta_0-\beta_*\|_2^2\cdot\frac{4}{(k+1)^2}=\frac{4L\|\beta_0-\beta_*\|_2^2}{(k+1)^2}\end{aligned}$$

が成立することがわかる．これが求めたかった不等式である． □