Some Matrix Aspects of Memoryless Nonlinear Programming Algorithms

Saman Babaie-Kafaki
SLOPE
Scientific Lab of Optimization and Engineering
Department of Mathematics
Faculty of Mathematics, Statistics, and Computer Science
Semnan University, Semnan, Iran
Email: sbk@seman.ac.ir
Website: sbk.profile.semnan.ac.ir

SOAL
Sharif Optimization and Applications Laboratory
Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran

Main Subjects

- **Nonlinear Programming: Models & Applications**
- **Some Concepts of Linear Algebra**
- **Quasi-Newton Methods**
- **Conjugate Gradient Methods**
- **Linear Regression: Some Modified Models**

J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer, New York, 2006.

D.S. Watkins. *Fundamentals of Matrix Computations.* John Wiley and Sons, New York, 2002.

Nonlinear Programming (NLP)

$$\min_{x\in\mathrm{R}^n} f(x) \qquad \text{subject to} \quad \begin{aligned} c_i(x) &= 0, \quad i \in \mathcal{E}, \\ c_i(x) &\geq 0, \quad i \in \mathcal{I}. \end{aligned}$$

$$\mathcal{E} = I = \varnothing \rightarrow \text{Unconstrained Optimization}$$

Penalty Method: $$Q(x;\mu) \overset{\text{def}}{=} f(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x) + \frac{\mu}{2}\sum_{i\in\mathcal{I}} \left([c_i(x)]^-\right)^2$$

Applications

Neural network training, Machine learning, Curve fitting, Compressed Sensing, Image restoration, Matrix approximations and so on.

Line Search Methods

$$x_{k+1} = x_k + s_k, \;\; s_k = \alpha_k d_k, \;\; k = 0, 1, \ldots$$

Wolfe Line Search:

$$\begin{aligned} f(x_k + \alpha_k d_k) - f(x_k) &\leq \delta \alpha_k \nabla f(x_k)^T d_k \\ \nabla f(x_k + \alpha_k d_k)^T d_k &\geq \sigma \nabla f(x_k)^T d_k \\ 0 < \delta < \sigma < 1 & \end{aligned}$$

Trust Region Methods

$$x_{k+1} = x_k + s_k, \;\; k = 0, 1, \ldots$$

$$\begin{aligned} \min_{s \in \mathbb{R}^n} \;\; & m_k(s) = g_k^T s + \frac{1}{2} s^T B_k s \\ s.t. \;\; & ||s|| \leq \Delta_k \end{aligned}$$

Notations: $g_k = \nabla f(x_k), \; y_k = g_{k+1} - g_k.$

Line Search-Based Algorithms

- **Steepest Descent (Gradient) Method:** $d_k = -g_k$
- **Newton Method:** $d_k = -\nabla^2 f(x_k)^{-1} g_k$
- **Quasi-Newton (QN) Method:** $d_k = -H_k g_k$
- **Conjugate Gradient (CG) Method:** $d_{k+1} = -g_{k+1} + \beta_k d_k$

Different Versions of the BFGS Updating Formula

Secant Condition: $H_{k+1}y_k = s_k$

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

$$\hat{H}_{k+1} = \theta_k H_k - \theta_k \frac{s_k y_k^T H_k + H_k y_k s_k^T}{s_k^T y_k} + \left(1 + \theta_k \frac{y_k^T H_k y_k}{s_k^T y_k}\right)\frac{s_k s_k^T}{s_k^T y_k}$$

$$\theta_k = \frac{s_k^T y_k}{y_k^T H_k y_k} \qquad\qquad \theta_k = \frac{s_k^T H_k^{-1} s_k}{s_k^T y_k}$$

$$\bar{H}_{k+1} = \theta_k I - \theta_k \frac{s_k y_k^T + y_k s_k^T}{s_k^T y_k} + \left(1 + \theta_k \frac{y_k^T y_k}{s_k^T y_k}\right)\frac{s_k s_k^T}{s_k^T y_k}$$

$$d_{k+1} = -\theta_k g_{k+1} + \theta_k \frac{y_k^T g_{k+1}}{s_k^T y_k} s_k + \theta_k \frac{s_k^T g_{k+1}}{s_k^T y_k} y_k - \left(1 + \theta_k \frac{||y_k||^2}{s_k^T y_k}\right)\frac{s_k^T g_{k+1}}{s_k^T y_k} s_k$$

Singular Value Decomposition

Theorem 4.1.1 *(SVD Theorem) Let $A \in \mathbb{R}^{n\times m}$ be a nonzero matrix with rank r. Then A can be expressed as a product*

$$A = U\Sigma V^T, \tag{4.1.2}$$

where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal, and $\Sigma \in \mathbb{R}^{n\times m}$ is a nonsquare "diagonal" matrix

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & & \\ & \sigma_2 & & & & \\ & & \ddots & & & \\ & & & \sigma_r & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix} \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

Singular Value Decomposition

Theorem 4.1.3 *(Geometric SVD Theorem) Let $A \in \mathbb{R}^{n\times m}$ be a nonzero matrix with rank r. Then $\mathbb{R}^m$ has an orthonormal basis $v_1, \ldots, v_m$, $\mathbb{R}^n$ has an orthonormal basis $u_1, \ldots, u_n$, and there exist $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ such that*

$$Av_i = \begin{cases} \sigma_i u_i & i = 1, \ldots, r \\ 0 & i = r+1, \ldots, m \end{cases} \qquad A^T u_i = \begin{cases} \sigma_i v_i & i = 1, \ldots, r \\ 0 & i = r+1, \ldots, n. \end{cases} \tag{4.1.4}$$

Theorem 4.2.1 *Let $A \in \mathbb{R}^{n\times m}$ have singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$. Then $\|A\|_2 = \sigma_1$.*

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}|^2 \right)^{1/2} \longrightarrow \|A\|_F = (\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2)^{1/2}$$

$$|\det(A)| = \sigma_1 \times \sigma_2 \times \ldots \times \sigma_n$$

Maximum Magnification

Theorem 4.2.4 *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix with singular values $\sigma_1 \geq \ldots \geq \sigma_n > 0$. Then*

$$\kappa_2(A) = \frac{\sigma_1}{\sigma_n}.$$

Another expression for the condition number that was given in Chapter 2 is

$$\kappa_2(A) = \frac{\text{maxmag}(A)}{\text{minmag}(A)},$$

where

$$\text{maxmag}(A) = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2},$$

$$\text{minmag}(A) = \min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

Maximum Magnification

Example 2.2.15 Let us take another look at the ill-conditioned matrices

$$A = \begin{bmatrix} 1000 & 999 \\ 999 & 998 \end{bmatrix} \quad \text{and} \quad A^{-1} = \begin{bmatrix} -998 & 999 \\ 999 & -1000 \end{bmatrix}$$

from Example 2.2.8. Notice that

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1999 \\ 1997 \end{bmatrix}. \tag{2.2.16}$$

If we use the ∞-norm to measure lengths, the magnification factor $\| Ax \|_\infty / \| x \|_\infty$ is 1999, which equals $\| A \|_\infty$. Thus $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is a vector that is magnified maximally by A. Since the amount by which a vector is magnified depends only on its direction and not on its length, we say that $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is in a *direction of maximum magnification* by A.

Dai-Liao Conjugate Gradient Method

$$d_{k+1}^T(g_{k+1}-g_k)=\alpha_k d_{k+1}^T\nabla^2 f(x_k+\xi\alpha_k d_k)d_k$$

$$d_{k+1}^T y_k=-tg_{k+1}^T s_k \longrightarrow \beta_k^{DL}=\frac{g_{k+1}^T y_k}{d_k^T y_k}-t\frac{g_{k+1}^T s_k}{d_k^T y_k}$$

$$\beta_k^{DL+}=\max\left\{\frac{g_{k+1}^T y_k}{d_k^T y_k},0\right\}-t\frac{g_{k+1}^T s_k}{d_k^T y_k}$$

Hestenes-Stiefel Method: $\beta_k^{HS}=\dfrac{g_{k+1}^T y_k}{d_k^T y_k}$

CG_Descent Algorithm: $\beta_k^{HZ}=\dfrac{g_{k+1}^T y_k}{d_k^T y_k}-2\dfrac{||y_k||^2}{d_k^T y_k}\dfrac{g_{k+1}^T d_k}{d_k^T y_k}$

Dai-Kou Method: $\beta_k(\tau_k)=\dfrac{g_{k+1}^T y_k}{d_k^T y_k}-\left(\tau_k+\dfrac{||y_k||^2}{s_k^T y_k}-\dfrac{s_k^T y_k}{||s_k||^2}\right)\dfrac{g_{k+1}^T s_k}{d_k^T y_k}$

Sufficient Descent Condition: $d_k^T g_k\le -c||g_k||^2,\ k=0,1,\ldots$

Search Direction Matrix of the Dai-Liao Method

$$Q_{k+1} = I - \frac{s_k y_k^T}{s_k^T y_k} + t\frac{s_k s_k^T}{s_k^T y_k} \longrightarrow d_{k+1} = -Q_{k+1} g_{k+1}, \ \ k = 0, 1, \ldots$$

$$\{u_k^i\}_{i=1}^{n-2} \xrightarrow{\text{Mutually Orthonormal}} Q_{k+1} u_k^i = Q_{k+1}^T u_k^i = u_k^i, \ \ i = 1, 2, \ldots, n-2$$

$$s_k^T u_k^i = y_k^T u_k^i = 0, \ \ i = 1, 2, \ldots, n-2$$

$$||Q_{k+1}||_F^2 = \mathrm{tr}(Q_{k+1}^T Q_{k+1}) \longrightarrow \sigma_k^{-2} + \sigma_k^{+2} = t^2 \frac{||s_k||^4}{(s_k^T y_k)^2} + \frac{||s_k||^2 ||y_k||^2}{(s_k^T y_k)^2}$$

$$\det(Q_{k+1}) = t\frac{||s_k||^2}{s_k^T y_k} \longrightarrow \sigma_k^- \sigma_k^+ = t\frac{||s_k||^2}{s_k^T y_k}$$

$$\sigma_k^{\pm} = \frac{1}{2}\frac{\sqrt{(t||s_k||^2 + s_k^T y_k)^2 + ||s_k||^2||y_k||^2 - (s_k^T y_k)^2}}{s_k^T y_k} \pm \frac{1}{2}\frac{\sqrt{(t||s_k||^2 - s_k^T y_k)^2 + ||s_k||^2||y_k||^2 - (s_k^T y_k)^2}}{s_k^T y_k}$$

Well-Conditioning: A Singular Value Analysis

$$\kappa(Q_{k+1}) = \frac{\sigma_k^+}{\sigma_k^-} \leq \frac{\left(1 + t\frac{||s_k||^2}{s_k^T y_k} + \frac{||s_k||||y_k||}{s_k^T y_k}\right)^2}{t\frac{||s_k||^2}{s_k^T y_k}} \longrightarrow t_{k_1}^* = \frac{s_k^T y_k}{||s_k||^2} + \frac{||y_k||}{||s_k||}$$

$$\kappa(Q_{k+1}) \leq \frac{\sigma_k^{+2} + \sigma_k^{-2}}{\sigma_k^- \sigma_k^+} = t\frac{||s_k||^2}{s_k^T y_k} + \frac{1}{t}\frac{||y_k||^2}{s_k^T y_k} \longrightarrow t_{k_2}^* = \frac{||y_k||}{||s_k||}$$

Descent Property: An Eeigenvalue Analysis

$$d_{k+1}^T g_{k+1} = -g_{k+1}^T Q_{k+1}^T g_{k+1} = -g_{k+1}^T \frac{Q_{k+1}^T + Q_{k+1}}{2} g_{k+1}$$

$$\lambda_k^{\pm} = \frac{1 + t\frac{||s_k||^2}{s_k^T y_k} \pm \sqrt{\left(t\frac{||s_k||^2}{s_k^T y_k} - 1\right)^2 + \frac{||s_k||^2 ||y_k||^2}{(s_k^T y_k)^2} - 1}}{2}$$

$$t > \frac{1}{4}\left(\frac{||y_k||^2}{s_k^T y_k} - \frac{s_k^T y_k}{||s_k||^2}\right) \longrightarrow \lambda_k^- > 0$$

$$\downarrow$$

$$t_k^{p,q} = p\frac{||y_k||^2}{s_k^T y_k} - q\frac{s_k^T y_k}{||s_k||^2}, p > \frac{1}{4} \text{ and } q < \frac{1}{4}$$

Maximum Magnification by the DL Search Direction Matrix

Based on the above discussions, when g_{k+1} approximately lies in the direction of the maximum magnification by Q_{k+1}, we may have

$$d_{k+1} = -Q_{k+1}g_{k+1} \longrightarrow ||g_{k+1}|| \ll ||d_{k+1}||.$$

In this situation, because $||d_{k+1}||$ may be extremely large in contrast to $||g_{k+1}||$, computational errors may appear with a great probability. Especially, when $||g_{k+2}|| \approx ||g_{k+1}||$ and the line search is approximately exact (as probable near the solution), we have

$$d_{k+2} \approx -g_{k+2} + \beta_{k+1}^{HS} d_{k+1}, \ \beta_{k+1}^{HS} = \frac{g_{k+2}^T y_{k+1}}{d_{k+1}^T y_{k+1}},$$

and thus, for not so small values of β_{k+1}^{HS}, the vector g_{k+2} may be swamped by $\beta_{k+1}^{HS} d_{k+1}$. So, a part of information may be lost. In addition, the directions d_{k+2} and d_{k+1} get approximately parallel and so, the method fails to generate a new search direction.

Undesirable Effects Related to the Maximum Magnification Direction

$$\cos\theta_{k+1} = -\frac{g_{k+1}^T d_{k+1}}{||g_{k+1}||\ ||d_{k+1}||} \xrightarrow{\text{exact line search}} \cos\theta_{k+1} = \frac{||g_{k+1}||}{||d_{k+1}||}$$

$$\{||d_k||/||g_k||\}_{k\geq 0} \text{ is bounded} \longrightarrow \lim_{k\to\infty} g_k = 0$$

$$g_{k+1}^T v_1 = 0 \longrightarrow t_k^{\pm} = \frac{\left(||y_k||^2 - a_k^2||s_k||^2 - \frac{(s_k^T y_k)^2}{||s_k||^2}\right)}{2a_k||s_k||^2} \pm \frac{\sqrt{\left(||y_k||^2 - a_k^2||s_k||^2 - \frac{(s_k^T y_k)^2}{||s_k||^2}\right)^2 + 4a_k^2||s_k||^2||y_k||^2}}{2a_k||s_k||^2}$$

$$a_k = \frac{s_k^T y_k}{||s_k||^2} - \frac{g_{k+1}^T y_k}{g_{k+1}^T s_k}$$

Linear Regression

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n\times p} \xrightarrow{\beta=?} y = X\beta + \varepsilon$$

Design Matrix

Errors

$$\mathbf{y} \approx \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\text{OLSE} \rightarrow \hat{\beta} = \arg\min_{\beta} \left(y - X\beta\right)^T \left(y - X\beta\right) = S^{-1} X^T y$$

$$p < n \rightarrow S = X^T X \ \textit{is Positive Definite.}$$

Outliers: Least Trimmed Squares Estimation

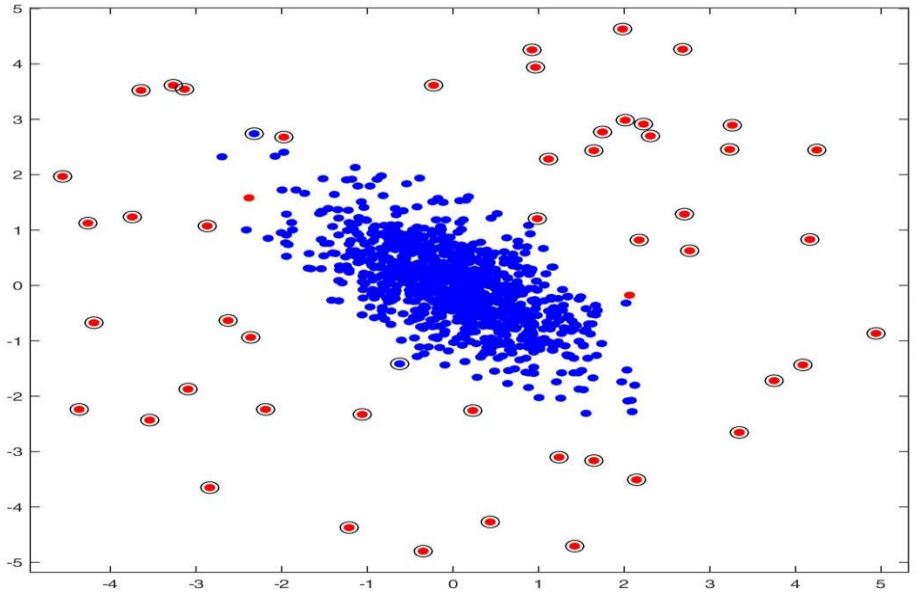
$$\min_{\beta,z} \quad f(\beta,z) = (y - X\beta)^T Z (y - X\beta)$$

$$s.t. \quad z^T e = h \qquad \rightarrow \hat{\beta}^{LTS} = S(z)^{-1} X^T Z y$$

$$z_i \in \{0,1\} \qquad i = 1,2,\ldots,n$$

$$Z = \text{diag}\underbrace{(z_1, z_2, \cdots, z_n)}_{z}, \; e = \text{ones}(n,1), \; S(z) = X^T Z X$$

Outliers →



Collinearity: Ridge Regression

$$For\ a\ positive\ definite\ matrix\ A \rightarrow \kappa(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

$$\min_{\beta} \quad f(\beta) = (y - X\beta)^T (y - X\beta)$$

$$s.t. \qquad \beta^T \beta \le \phi^2$$

$$\downarrow$$

$$\min_{\beta} f(\beta) = (y - X\beta)^T (y - X\beta) + k\left(\beta^T \beta - \phi^2\right)$$

$$\downarrow$$

$$S(k) = X^T X + kI_p = S + kI_p \rightarrow \hat{\beta}(k) = S(k)^{-1} X^T y$$

Ridge Least Trimmed Squares Estimation

$$\min_{\beta,z} \quad f(\beta,z) = (y-X\beta)^T Z(y-X\beta) + k\left(\beta^T\beta - \phi^2\right)$$
$$s.t. \quad z^T e = h$$
$$z_i \in \{0,1\} \qquad i = 1,2,\ldots,n$$

$$\downarrow$$

$$S(k,z) = X^T ZX + kI_p \rightarrow \hat{\beta}^{RLTS}(k,z) = S(k,z)^{-1} X^T y$$

Well-Conditioned Least Trimmed Squares Estimation

$$\min_{\beta,z} \quad f(\beta,z) = (y-X\beta)^T Z(y-X\beta) + M\kappa\underbrace{\left(X^T ZX\right)}_{S(z)}$$
$$s.t. \quad z^T e = h$$
$$z_i \in \{0,1\} \qquad i = 1,2,\ldots,n$$

Byrd-Nocedal Measure Function

$$\psi(A) = \mathrm{tr}(A) - \ln(\det(A)) \geq \ln\left(\kappa(A)\right)$$

$$\begin{aligned} \min_{\beta,z} \quad & f(\beta,z) = (y - X\beta)^T Z (y - X\beta) + M\psi\left(S(z)\right) \\ s.t. \quad & z^T e = h \\ & z_i \in \{0,1\} \qquad i = 1,2,\ldots,n \end{aligned}$$

Piazza-Politi Upper Bound for Condition Number

$$\kappa(A) \leq \varphi(A) = \frac{2}{|\det(A)|}\left(\frac{||A||_F}{\sqrt{n}}\right)^n$$

$$\begin{aligned} \min_{\beta,z} \quad & f(\beta,z) = (y - X\beta)^T Z (y - X\beta) + M\varphi\left(S(z)\right) \\ s.t. \quad & z^T e = h \\ & z_i \in \{0,1\} \qquad i = 1,2,\ldots,n \end{aligned}$$

High-Dimensional Regression

$$p > n \rightarrow S = X^T X \text{ is Positive Semidefinite.}$$

$$\downarrow$$

$$\mathrm{S}(z) \approx \theta I + uu^T$$

Theorem 4.2.15 *Let* $A \in \mathbb{R}^{n\times m}$ *with* $\operatorname{rank}(A) = r > 0$. *Let* $A = U\Sigma V^T$ *be the SVD of* A, *with singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. *For* $k = 1, \ldots, r-1$, *define* $A_k = U\Sigma_k V^T$, *where* $\Sigma_k \in \mathbb{R}^{n\times m}$ *is the diagonal matrix* $\operatorname{diag}\{\sigma_1, \ldots, \sigma_k, 0 \ldots, 0\}$. *Then* $\operatorname{rank}(A_k) = k$, *and*

$$\sigma_{k+1} = \| A - A_k \|_2 = \min \{ \| A - B \|_2 \mid \operatorname{rank}(B) \leq k \}.$$

That is, of all matrices of rank k *or less,* A_k *is closest to* A.

References

- Y.H. Dai and L.Z. Liao. New conjugacy conditions and related nonlinear conjugate gradient methods. *Appl. Math. Optim.*, 43(1):87–101, 2001.
- W.W. Hager and H. Zhang. Algorithm 851: CG_Descent, a conjugate gradient method with guaranteed descent. *ACM Trans. Math. Software*, 32(1):113–137, 2006.
- Y.H. Dai and C.X. Kou. A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search. *SIAM J. Optim.*, 23(1):296–320, 2013.
- S. Babaie–Kafaki and R. Ghanbari. The Dai–Liao nonlinear conjugate gradient method with optimal parameter choices. *European J. Oper. Res.*, 234(3):625–630, 2014.
- Z. Aminifard and S. Babaie–Kafaki. An optimal parameter choice for the Dai–Liao family of conjugate gradient methods by avoiding a direction of the maximum magnification by the search direction matrix. *4OR*, 17:317–330, 2019.
- M. Roozbeh, S. Babaie–Kafaki, and A. Naeimi Sadigh. A heuristic approach to combat multicollinearity in least trimmed squares regression analysis. *Appl. Math. Model.*, 57:105–120, 2018.

Any Question? You can contact me by sbk@semnan.ac.ir**, please.**