Improving Theoretical Line Search Techniques of Practical Numerical Optimization

Nofl Sh. Al-Shimari¹, Dr. Ahmed Sabah. Al-Jilawi²
¹College of Education for Pure S.C/University of Babylon
²College of Education for Pure S.C/University of Babylon

Emails: nawal.kazem@student.uobabylon.edu.iq
       aljelawy2000@yahoo.com

Abstract The paper aims to shed light on some important concepts in optimization so that the reader can have a broad knowledge and awareness of idealism, In addition to that we bring the reader closer to know the line of research and how to employ to solve various problems of optimization.

1. Introduction

Each new time Optimization teaching is being approached to hand issues that are a lot bigger and complex than before, as a result of the wide (and developing) Optimization of improvement in science. Improvement is a significant device in choice science and the investigation of actual frameworks. The way toward recognizing target factors and requirements for a given issue is known as modeling.

2. Mathematical Formulation

\[
\begin{align*}
\text{minimize} & \quad f^0(x) \\
\text{subjectto} & \quad f^i(x) \leq 0, i = \{1, ..., I\}, \\
& \quad g^i(x) = 0, i = \{1, ..., J\}, \\
& \quad x \in \Omega.
\end{align*}
\]

(2)

where \(x \in \mathbb{R}^n\) parameters and \(Q\) is objective function we requirement minimize or maximize. we want find \(x^*\) s.t \(Q(x^*) \leq Q(x)\) for all \(x \in X\). Always can say maximize \(Q(x) \equiv -\text{minimize} Q(x)\). In the figure above, the dark region represents the feasible set of linear(see [23, 29]) programming. And the regain boundaries illustrate the constraints. The general optimization problem in the form

\[
\begin{align*}
\text{minimize} & \quad f^0(x) \\
\text{subjectto} & \quad f^i(x) \leq 0, i = \{1, ..., I\}, \\
& \quad g^i(x) = 0, i = \{1, ..., J\}, \\
& \quad x \in \Omega.
\end{align*}
\]

Where I and J are sets of indicator for equality and inequality constraints[18]
(a) Some Problem from An Optimization

1. Finding the shortest way between the two cites.
2. Location problem Finding a suitable location between two cities to build an airport, so that the distance between the airport and the two cities is the minimal.[?]
3. Transportation Problem: Find the best way to satisfy the requirement of demand points using the capacities of supple point.

Background Mathematical (Basic Concept)

The Hessian matrix is defined as the (square $n \times n$) symmetric matrix[1 20, 25]

$$H = \nabla^2 f(x) = \begin{pmatrix}
\frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}
\end{pmatrix}$$

Example 1.1 $f(x, y) = 5x^2 + xy^3 - y^2$; $\mathcal{H} = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix} = \begin{bmatrix} 10 & 3y^2 \\ 3y^2 & 6xy - 2 \end{bmatrix}$ so $\begin{bmatrix} 10 \\ 3 \end{bmatrix} = \begin{bmatrix} 40 - 9 = 31 \geq 0 \text{ then } \mathcal{H} \succeq 0. $  

And Let $f: R^n \to R$ and $x \in R^n$ the gradient of $f$ at $x$ is define as the column vector

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

Example 1.2 Let $f(x, y) = 4 - x^2 - y^2$ than $\nabla f(x, y) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}) = (-2x, -2y)$

1. argmin : Let $S \subseteq \mathbb{R}^n$ be nonempty set if $f: S \to R$, the argument of the minimum is the set of element in $S$ that achieve the global minimum in $S$. argmin$f(x) = \{x \in S | f(y) \geq f(x), \forall y \in S\}$

For example if $f: R \to R$ is given by $f(x) = (x + 1)^2 + 3$ than argmin$f(x) = -1$ where $f(x) = (-1 + 1)^2 + 3 = 3$ is the minimum objective function .

2. $epi(f)$ let $S \subseteq \mathbb{R}^n$ be an nonempty convex set the epigraph of a function $f: S \to R$ denote by $epi(f) \subseteq R^{n+1}$

$$epi(f) = \{(x, \lambda) | f(x) \leq \lambda, x \in S, \lambda \in R\}$$

3. cone : A set $C \subseteq \mathbb{R}^n$ is a cone when with every $x \in C$ the whole ray $\{\lambda x | \lambda \geq 0\}$ also belongs to the set $C$. this means $\lambda x \in C, \forall x \in C$ and $\lambda \geq 0$

Cone is may or not may be convex also Cone my or my not contains the origin for example
• \( \{x \in \mathbb{R}^n | x_1 x_2 = 0 \} \) cone that contains origin and non convex
• \( \{x \in \mathbb{R}^n | x_1 x_2 \neq 0 \} \) cone that not contains origin and non convex

4. **Global minimum and local minimum** [15]: A point \( x^* \) is called a local minimize if \( f(x^*) \leq f(x) \) forall \( x \in \mathcal{B}(x^*, \delta) \) where \( \delta > 0 \). A point \( x^* \) is called a strict local minimum if \( f(x^*) < f(x) \) forall \( x \in \mathcal{B}(x^*, \delta) \) where \( \delta > 0 \) with \( x \neq x^* \).

   A point \( x^* \) is called a global minimize if \( f(x^*) \leq f(x) \) for all \( x \in \mathbb{R} \), A point \( x^* \) is called a strict global minimum if\([1, 3]\) \( f(x^*) < f(x) \) forall \( x \in \mathbb{R}^n \), with \( x \neq x^* \)

5. **KKT condition**: Assume \( x^* \) maximizes the maximize \( f(x) \)
subject to \( g_i(x) \leq 0 (\mu_i) \forall i \in I \)
• primal constraints \( g_i(x) \leq 0 \) \( i \in \{1, \ldots, I \} \) and \( h_i(x) = 0 \)
• complementary slackness \( \mu_i g_i(x^*) = 0 \)
• Dual constraints \( \mu_i \geq 0 \)
• \( \forall \) L. w. r. t. \( x \) will be equal to zero.
\[
\nabla f(x^*) - \sum_{i=1}^{I} \lambda_i \nabla h_i(x^*) - \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) = 0
\]

3. Unconstrained Optimization

(b) **First Order Necessary Condition**
   if \( x^* \) is a local minimizer of \( f(x) \) and \( f(x) \) is contentiously differentiablein an \( (B(x^*, \epsilon), f \in \mathcal{C}^1 \) with \( \epsilon > 0 \) then
\[
\nabla f(x^*) = 0
\]
That is \( f(x) \) is stationary at \( x^* \)

(c) **Second Order Necessary Conditions**
   if \( x^* \) is a local minimizer of \( f(x) \) and \( \nabla^2 f(x) \) is contentiously differentiable in an \( ((B(x^*, \epsilon) \) with \( \epsilon > 0, f \in \mathcal{C}^2 \) then
\[
\nabla f(x^*) = 0
\]
\[
d \nabla^2 f(x^*) d^T \geq 0
\]

(d) **Second Order Sufficient Condition**
   suppose \( f(x) \) and \( \nabla^2 f(x) \) is contentiously differentiable in an open neighborhood of \( x^* \) if the following to condition are satisfy then \( x^* \) is a local minimum of \( f(x) \).
\[
\nabla f(x) = 0
\]
\[
d \nabla^2 f(x^*) d^T > 0
\]
this means \( \nabla^2 f(x) > 0 \) where \( d = x + ad \) is belong to feasible direction.

**Some Important Theory Of Line Search Methods**

4. **Line Search Methods**
   How to find \( x^{k+1} \) such that \( f(x^{k+1}) < f(x^k) \) ? Every cycle of a linesearch technique processes a search direction \( d^k \) and then afterward chooses how far to move along that direction. the iteration[10] is given by
\[
x^{k+1} = x^k + d^k
\]
1. **How to find the \( d^k \)?**

[42] consider the first order approximation to \( f(x) \) about \( x^k \),
\[
f(x) = f(x^k) + g^T d = f(x^k) + g^T d \] \[11, 30\]

2. **Descent Direction Set**

\[
DS = \{ d \in \mathbb{R}^n : g^T d < 0 \}
\]

Let \( g^k \neq 0 \) and \( d^k = -A^k g^k \) where \( A^k \) is a symmetric matrix

3. **How to determine \( \alpha^k \)?**

(a) Exact line search ??

Given a descent direction \( d^k \) determine \( \alpha^k \) by solving the optimization problem .
\[
\alpha^k = \arg\min_{\alpha \geq 0} \phi(\alpha) = f(x^k + \alpha d^k)
\]

At each iteration we need to solve the one dimensional optimization problem to find \( \alpha^k \). In most of case we have not really interesting to solve this problem our mine problem to minimize \( f(x) \) . this means we find some \( \alpha^k \) s.t
\[
f(x^{k+1}) < f(x^k)
\]

(b) **Inexact linesearch**
- Small decrease in function values relative to the steep length
- In this case the choice of \( \alpha^k \) is crucial
  i. **Armijo’s condition** \( \phi_1(\alpha) = f(x^k) + c_1 g^T d^k c_1 \in (0, 1) \) choose \( \alpha^k \) such that
  \[
f(x^k + \alpha d^k) \leq \phi_1(\alpha^k)
\]
  ii. **Goldenstein’s condition** \( \phi_2(\alpha) = f(x^k) + c_2 g^T d^k, c_2 \in (c_1, 1) \)
  iii. **Armijo-Goldstein conditions** choose \( \alpha \) such that
  \[
  \phi_2(\alpha^k) \leq f(x^k + \alpha d^k) \leq \phi_1(\alpha^k)
  \]

iv. **Wolfe’s condition** \( \phi' \geq c_2 \phi'(0), c_2 \in (c_1, 1) \)

**Backtracking Line Search**[4]

Pick the correct development size at each of gradient update. The method is summed up as:

At each iteration, start with step size \( t = 1 \)
\[
f(x - t\nabla f(x)) > f(x) - \alpha \| \nabla f(x) \|_2^2
\]
shrivle \( t = \beta t \), where \( 0 < \beta < 1, 0 < \alpha \leq 1/2 \) are some fixed parameters

Perform gradient update of \( x : x^+ = x - t\nabla f(x) \)

This keeps the step from getting too small ... but does not prevent too large steps keeps to decrease in \( f \)

**Theorem 1.1** suposst any iteration, where \( p^k \) is a descent direction and \( \alpha^k \) satisfies the “Wolfe conditions”: consider \( f \) is bounded below in \( \mathbb{R}^n \) and that \( f \in C^1 \) in an open set \( O \) consist the level set \( \{ x : f \leq f(x_0) \} \), where \( x_0 \) is the entail point of the iteration. Assume also that the gradient \( \nabla f \) is **Lip** continuous on \( O \), that is, \( \exists a constant B > 0 \) such that
\[
\| \nabla f(x) - \nabla f(\tilde{x}) \| \leq B \| x - \tilde{x} \|, \text{forall } x, \tilde{x} \in O
\]

Then
\[
\sum_{k \geq 0} \cos^2 \theta^k \| \nabla f^k \|^2 < \infty
\]

**PROOF.**

\[
(\nabla f^{k+1} - \nabla f^k)^T p^k \geq (c_2 - 1) \| \nabla f^k p^k \|
\]

while the Lipschitz condition [3]
\[
(\nabla f^{k+1} - \nabla f^k)^T p^k \leq \lambda^k L \| p^k \|^2
\]

By joining these two relations, we have \( \alpha^k \geq \frac{c^2 - 1}{L} \frac{\| p^k \|^2}{\| p^k \|^2} \)

By replace this inequality into the 1st “Wolfe condition”
we can write this relation as

\[ f^{k+1} \leq f^k - c_1 \left( \frac{\cos^2 \theta^k \| \nabla f^k \|^2}{\| p^k \|^2} \right) \]

where \( c = c_1(1 - c_2)/L \). By adding this articulation over\([1, 13, 15]\) all indices \( \leq k \) we obtain

\[ f^{k+1} \leq f^0 - c \sum_{j=0}^{k} \cos^2 \theta_j \| \nabla f_j \|^2 \]

since \( f \) is bounded below, we have that\([4]\) \( f^0 - f^{k+1} < B \) wher B constant, for all \( k \).

Hence, We have

\[ \sum_{k=0}^{\infty} \cos^2 \theta^k \| \nabla f \|^2 < \infty \]

5. Convergence rate of steepest descent

[1, 10, 11] The objective function is quadratic and the line searches are exact. Let us suppose that

\[ f(x) = \frac{1}{2} x^T O x - b^T x \]

where \( O = O^T \) and \( O \succeq 0 \). The \( \lambda \) is given by \( \nabla f(x) = O x - b \) and the minimizer \( x^* \) is the only solution of the linear system \( O x = b \). It is compute the \( \alpha^k \) that minimizes \( f(x^k - \alpha \nabla f^k) \). By differentiating the function

\[ f(x^k - \alpha \nabla f^k) = \frac{1}{2} \left( (x^k - \alpha \nabla f^k)^T O(x^k - \alpha \nabla f^k) - b^T(x^k - \alpha \nabla f^k) \right) \]

w.r.t \( \alpha \), and check the derivative to zero, we obtain

\[ \alpha_k = \frac{\nabla f^T x^k}{\nabla f^T \nabla f^k} \]

If we use this exact minimizer \( \alpha^k \),

\[ x^{k+1} = x^k - \frac{\nabla f^T x^k}{\nabla f^T \nabla f^k} \nabla f^k \]

so this standard estimates the contrast between the current objective value and the optimal value. \( \nabla f^k = O(x_k - x^*) \), we can derive the equality

\[ \| x^{k+1} - x^* \|^2 = \left( 1 - \frac{\nabla f^T x^k}{\nabla f^T \nabla f^k} \right) \| x^k - x^* \|^2 \]

Theorem 1.2 Suppose that \( f: \mathbb{R}^n \to \mathbb{R} \), where \( f \in C^2 \) and that the iterates generated by the "steepest-descent method" \( CD \) with \( \alpha_k = \text{arg} \min \{ f(x + \alpha \nabla f(x)) | \alpha \in \mathbb{R} \} \) searches converge to a point \( x^* \) atwhich the Hessian matrix\( \nabla^2 f \) \( \succeq \varepsilon \), let \( s \) be any scalar satisfying

\[ s \in \left( \frac{\lambda^1}{\lambda^2 + \cdots + \lambda^i}, 1 \right) \]

such that \( \lambda^1 \leq \lambda^2 \leq \cdots \leq \lambda^n \) are the eigenvalues of \( \nabla^2 f(x^*) \). Then \( \forall x^k \) adequately huge, we have

\[ f(x^{k+1}) - f(x^*) \leq r^2 [f(x^k) - f(x^*)] \]

We cannot agree the rate of convergence to get better if an inexact line search is used. Therefore, Theorem prove that the "steepest descent method" can have an unsatisfactorily slow rate of convergence\([14]\).
Example 1.3 if $\kappa(Q) = 900, f(x_0) = 1, \text{ and } f(x^*) = 0$,
The theorem proposes that the capacity worth will even now be about 0.09 after one 1000 iterations of the “DC method” with exact line search.

Theorem 1.3 Assume that $f: \mathbb{R}^n \to \mathbb{R}$: And let $f \in C^2$. Consider the iteration $x^{k+1} = x^k + \alpha^k p^k$, where $p^k$ is a descent direction and $\alpha^k$ satisfies the “Wolfe conditions” with $c_1 \leq 1/2$. If $\{x^k\}$ converges to a point $x^*$ s.t $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$, and if the search direction satisfies

$$\lim_{k \to \infty} \frac{\|p^k + \nabla^2 f(x^*) p^k\|}{\|p^k\|} = 0$$

then (i) the $\alpha^k = 1$ is allowable for all $k \geq k_0$ and (ii) if $\alpha^k = 1 \forall k > k_0(x^k) \to x^* p^k$. It is easy to see that if $c_1 > 1/2$, then the line search would shut out the minimizer of a quadratic, and unit step lengths may not be reasonable.

If $p^k$ is a "quasi-Newton search direction" then

$$p^k = -B^T g^k$$

$$\lim_{k \to \infty} \frac{\|p^k\|}{\|p^k\|} = 0$$

We have the excellent result that a excellent [10] liner convergence rate can be accomplished regardless of whether the sequence of "quasi-Newton" matrices $B^k$ does not converge to $\nabla^2 f(x^*)$; it gets the job done that the $B^k$ become progressively precise approximations to $\nabla^2 f(x^*)$ along the search directions $p^k$. Above condition is both NS for the convergence of "quasi-Newton methods".

Theorem 1.4 Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is $(f \in C^2)$. Consider the iteration $x^{k+1} = x^k + p^k \alpha^k = 1$. Let us assume also that $\{x^k\}$ converges to a point $x^*$ such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$. Then $\{x^k\}$ converges excellent if and only if (1.3) holds.

**Proof.** We show that([17, 8, 49])

$$p_k - p^*_k = Q(\|p^k\|)$$

where $p^*_k = -\nabla^2 f_k^{-1} \nabla f_k$ is the Newton step.

$$p^k - p^*_k = \nabla^2 f_k^{-1}(\nabla^2 f_k p^k + \nabla f_k)$$

$$= \nabla^2 f_k^{-1}(\nabla^2 f_k - B_k) p^k$$

$$= Q(\|\nabla^2 f_k - B_k\| p^k)$$

$$= Q(\|p^k\|)$$

where we have utilized the way that $\|\nabla^2 f_k^{-1}\|$ is bounded above for $x^k$ sufficiently close to $x^*$ since the limiting Hessian $\nabla^2 f(x^*) \succeq 0$

$$\|x^k + p^k - x^*\| \leq \|x^k + p^*_k - x^*\| + \|p^k - p^*_k\| = Q(\|x_k - x^*\|^2) + Q(\|p^k\|)$$

A simple doctrinaire of this inequality expose that[5, 7, 13] $\|p^k\| = Q(\|x^k - x^*\|)$, so we get

$$\|x^k + p^k - x^*\| \leq Q(\|x^k - x^*\|)$$

6. Conclusion

In this paper, We presented a simple picture of optimization and we touched upon some of the linesearch. In addition to that, we presented some important and popular theories about the line of research. As well as how to choose the descent direction $d^k$ and how to choose the right step $\alpha^k > 0$. 

---

[10] Linear convergence rate

[17, 8, 49] Newton's method
References

[1] A. Fiacco and G. McCormick, Sequential unconstrained minimization techniques for nonlinear programming. A primal dual method. ManagementSci, 10 (1968).

[2] Angelo Miele, EE Cragg, RR Iyer, and AV Levy. Use of the augmented penalty function in mathematical programming problems, part 1. Journal of Optimization Theory and Applications, 8(2):115–130, 1971.

[3] Andrzej Ruszczynski. Nonlinear optimization. Princeton university press, 2011.

[4] Christian Kanzow, Daniel Steck, and Daniel Wachsmuth. An augmented lagrangian method for optimization problems in banach spaces. SIAM Journal on Control and Optimization, 56(1):272–291, 2018.

[5] Dimitri P Bertsekas and Athena Scientific. Convex optimization algorithms. Athena Scientific Belmont, 2015.

[6] Magnus R Hestenes. Multiplier and gradient methods. Journal of optimization theory and applications, 4(5):303–320, 1969.

[7] D. P. Bertsekas, Nonlinear programming, Athena scienti_c Belmont, 1999.

[8] Dimitri P Bertsekas. Constrained optimization and Lagrange multiplier methods. Academic press, 2014.

[9] R. Courant et al., Variational methods for the solution of problems of equilibrium and vibrations.

[10] Judah Ben Rosen, Olvi L Mangasarian, and Klaus Ritter. Nonlinear Programming: Proceedings of a Symposium Conducted by the Mathematics Research Center, the University of Wisconsin, Madison, May 4-6, 1970. Number 25. Elsevier, 2014.

[11] JB Hiriart-Urruty and C Lemaréchal. Convex analysis and minimization algorithms. no. 305-306 in grund. der math. wiss, 1993.

[12] Kaisa Joki. Nonsmooth optimization: Bundle methods. 2013.

[13] M. M. El-Alem, A global convergence theory for a class of trust region algorithms for constrained optimization, PhD thesis, Rice University, 1988.

[14] Michael James Lowe. Nonlinear programming: augmented Lagrangian techniques for constrained minimization. PhD thesis, Montana State University-Bozeman, College of Engineering, 1974.

[15] Marko Mäkelä. Survey of bundle methods for nonsmooth optimization. Optimization methods and software, 17(1):1–29, 2002.

[16] M. Hestenes, Multiplier and gradient methods, in” computing methods in optimization problems-2”, edited by la zadeh, lwneustadt, and avbalakrishnan, 1969.

[17] P Dimitri Bertsekas. Nonlinear programming. 1999. Athena, Scientific, Belmont CA.

[18] Stephan Dempe and Jonathan F Bard. Bundle trust-region algorithm for bilinear bilevel programming. Journal of Optimization Theory and Applications, 110(2):265–288, 2001.

[19] Yu Du and Andrzej Ruszczyński. Rate of convergence of the bundle method. Journal of Optimization Theory and Applications, 173(3):908–922, 2017.