A comparison between line searches and trust regions for nonlinear optimization

Luís N. Vicente[1]

Departamento de Matemática
Universidade de Coimbra
3000 Coimbra, Portugal

Abstract

Line searches and trust regions are two techniques to globalize nonlinear optimization algorithms. We claim that the trust–region technique has built–in an appropriate regularization of ill–conditioned second–order approximation. The question we ask and then answer in this short paper supports this claim. We force the trust–region technique to act like a line search and we accomplish this by always choosing the step along the quasi–Newton direction. We obtain global convergence to a stationary point as long as the condition number of the second–order approximation is uniformly bounded, a condition that is required in line searches but not in trust regions.

Resumo

A pesquisa unidimensional e as regiões de confiança são técnicas de globalização de algoritmos para optimização não linear. A técnica de regiões de confiança incorpora também a regularização de aproximações de segunda ordem mal condicionadas. Neste artigo é discutida esta regularização numa situação em que a técnica de regiões de confiança é forcada a actuar como a pesquisa unidimensional, ao exigir–se que o passo seja sempre na direcção de quasi–Newton. Neste caso, a convergência global para um ponto estacionário é verdadeira desde que o número de condição da aproximação de segunda ordem seja limitado uniformemente, hipótese que tradicionalmente é assumida para a pesquisa unidimensional mas não para as regiões de confiança.

Keywords. line searches, trust regions, quasi–Newton.

1 Framework

Consider the unconstrained minimization problem

$$\text{minimize } f(x), \tag{1}$$

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is at least continuously differentiable, and $x \in \mathbb{R}^n$.

[1]This work was developed while the author was a graduate student at the Department of Computational and Applied Mathematics of Rice University. Support has been provided by INVOTAN (NATO scholarship), CCLA (Fulbright scholarship), FLAD, and NSF cooperative agreement CCR–9120008.

A quasi–Newton method for the solution of (1) generates a sequence of iterates $\{x_k\}$ and steps $\{s_k\}$ such that $x_{k+1} = x_k + s_k$. At x_k, a quadratic model of $f(x_k + s)$,

$$\Psi_k(s) = f(x_k) + g_k^T s + \frac{1}{2} s^T H_k s,$$

is formed, where $g_k = \nabla f(x_k)$ and H_k introduces curvature into the model. We assume that H_k is a symmetric positive definite matrix of order n. The quasi–Newton step s_k is given by $s_k = -H_k^{-1} g_k$ and hence is the unconstrained minimizer of $\Psi_k(s)$. Thus a quasi–Newton method consists of forming $x_{k+1} = x_k - H_k^{-1} g_k$, for $k = 0, 1, \ldots$, but it is well–known that such an algorithm is not globally convergent. If we want to start with any choice of x_0 and still guarantee convergence then we need a globalization strategy.

A line search strategy considers $-H_k^{-1} g_k$ to be a direction from which a step will be obtained. The step s_k is of the form $-\lambda_k H_k^{-1} g_k$, where the step length λ_k is chosen in an appropriate way.

The trust–region technique does not necessarily choose the quasi–Newton direction. Here a step is an approximate solution of the trust–region subproblem

$$\begin{aligned} \text{minimize} \quad & \Psi_k(s), \\ \text{subject to} \quad & \|s\| \leq \delta_k, \end{aligned} \tag{2}$$

where δ_k is the trust radius, and $\|\cdot\|$ denotes a norm in $\mathbb{R}^n$, assumed in this paper to be the ℓ_2 norm.

2 Line searches and trust regions

The global convergence result we are looking at is

$$\lim_{k \longrightarrow +\infty} \|g_k\| = 0. \tag{3}$$

Let us describe in detail the classical conditions under which both line searches and trust regions give us (3).

If a line search is used one has to ask the step $s_k = -\lambda_k H_k^{-1} g_k$ to satisfy the Armijo–Goldstein–Wolfe conditions:

$$f(x_k + s_k) \leq f(x_k) + \alpha_1 g_k^T s_k, \tag{4}$$

$$\nabla f(x_k + s_k)^T s_k \geq \alpha_2 g_k^T s_k, \tag{5}$$

where α_1 and α_2 are constants fixed for all k and satisfying $0 < \alpha_1 < \alpha_2 < 1$. After an s_k, or equivalently a λ_k, has been found that satisfies these conditions, a new iterate x_{k+1} is formed by setting $x_{k+1} = x_k + s_k = x_k - \lambda_k H_k^{-1} g_k$. A key ingredient to obtain global convergence to a stationary point is to keep the angle $\theta_k \in [0, \frac{\pi}{2}]$ between g_k and $-H_k^{-1} g_k$ uniformly bounded away from $\pi/2$. Let $\mathbf{cn}(H_k) = \|H_k\| \|H_k^{-1}\| \geq 1$ be the condition number of the matrix H_k. If $\mathbf{cn}(H_k)$ is uniformly bounded, i.e., if there exists a $\nu > 0$ such that

$$\mathbf{cn}(H_k) \leq \nu$$

for every k, then we have

$$cos(\theta_k) = \frac{g_k^T H_k^{-1} g_k}{\|g_k\| \, \|H_k^{-1} g_k\|} \geq \frac{1}{\nu}. \tag{6}$$

The inequality (6) is proved using

$$\frac{g_k^T H_k^{-1} g_k}{\|g_k\|\,\|H_k^{-1} g_k\|} \geq \frac{\lambda_{min}(H_k^{-1})\|g_k\|^2}{\|H_k^{-1}\|\,\|g_k\|^2} = \frac{1}{\lambda_{max}(H_k)\|H_k^{-1}\|} = \frac{1}{\|H_k\|\,\|H_k^{-1}\|},$$

where $\lambda_{min}(H_k^{-1})$ and $\lambda_{max}(H_k)$ denote the smallest and largest eigenvalues of H_k^{-1} and H_k, respectively. The lower bound (6) on $cos(\theta_k)$ is crucial to establish the following result.

Theorem 2.1 *Let f be bounded below and ∇f be uniformly continuous. If s_k satisfies (4)–(5) and the condition number $\mathbf{cn}(H_k)$ of H_k is uniformly bounded, then $\{x_k\}$ satisfies* (3).

Some of the ground work that led to this result was provided by Armijo [1] and Goldstein [7]. It was established by Wolfe [23], [24] and Zoutendijk [25], under the assumption that the gradient is Lipschitz continuous. However this condition can be relaxed and one can see that uniform continuity is enough (see Fletcher [5], Theorem 2.5.1). Some practical line–search algorithms are described by Moré and Thuente [10]. For more references see also the books [3], [11], and [13] and the review papers [4] and [12].

Now let us describe how the trust–region technique works. A step s_k has to decrease the quadratic model $\Psi_k(s)$ from $s = 0$ to $s = s_k$. The way s_k is computed determines the magnitude of the predicted decrease $\Psi_k(0) - \Psi_k(s_k)$ and influences the type of global convergence of the trust–region algorithm. One can ask s_k to satisfy two classical conditions, either fraction of Cauchy decrease (simple decrease) or fraction of optimal decrease.

The first condition forces the predicted decrease to be at least as large as a fraction of the decrease given for $\Psi_k(s)$ by the Cauchy step c_k. This step is defined as the solution of the one–dimensional problem minimize $\Psi_k(s)$ subject to $\|s\| \leq \delta_k,\ s \in span\{-g_k\}$, and it is given by

$$c_k = \begin{cases} -\frac{\|g_k\|^2}{g_k^T H_k g_k} g_k & \text{if } \frac{\|g_k\|^3}{g_k^T H_k g_k} \leq \delta_k, \\ \\ -\frac{\delta_k}{\|g_k\|} g_k & \text{otherwise.} \end{cases} \tag{7}$$

The step s_k is said to satisfy a fraction of Cauchy decrease for the trust–region subproblem (2) if

$$\Psi_k(0) - \Psi_k(s_k) \geq \beta_1 \Big(\Psi_k(0) - \Psi_k(c_k) \Big), \tag{8}$$

where $\beta_1 \in (0, 1]$ is fixed across all iterations. Two widely used algorithms to compute steps that satisfy (8) are the dogleg algorithm ([2], [14], and [17]) and the conjugate–gradient algorithm ([20] and [22]).

The second condition is more stringent and relates the predicted decrease to the decrease given on $\Psi_k(s)$ by the optimal solution s_k^* of the trust–region subproblem (2). The step s_k is said to satisfy a fraction of optimal decrease for the trust–region subproblem (2) if

$$\Psi_k(0) - \Psi_k(s_k) \geq \beta_2 \Big(\Psi_k(0) - \Psi_k(s_k^*) \Big), \tag{9}$$

where $\beta_2 \in (0, 1]$ is fixed across all iterations. Algorithms to compute s_k that satisfy the fraction of optimal decrease (9) have been proposed in [9] and [19]. It is a simple matter to see that (9) implies (8).

The predicted decrease $pred(s_k)$ given by s_k is defined as $\Psi_k(0) - \Psi_k(s_k)$. The actual decrease $ared(s_k)$ is given by $f(x_k) - f(x_k + s_k)$. The trust–region strategy relates the acceptance of s_k and the update of the trust radius with the ratio $r_k = \frac{ared(s_k)}{pred(s_k)}$ in the following way:

If $r_k < \eta$ then s_k is rejected, $x_{k+1} = x_k$, and $\delta_{k+1} = \gamma \|s_k\|$.

If $r_k \geq \eta$ then s_k is accepted, $x_{k+1} = x_k + s_k$, and $\delta_{k+1} \geq \delta_k$.

Here γ and η are uniformly fixed and such that $0 < \gamma, \eta < 1$. Of course the rules to update the trust radius can be much more involved, but the above suffices to prove convergence results and to understand the trust–region mechanism.

Theorem 2.2

Let f be bounded below and ∇f be uniformly continuous. If s_k satisfies (8) and $\|H_k\|$ is uniformly bounded, then $\{x_k\}$ satisfies (3).

If in addition, f is twice continuously differentiable and s_k satisfies (9), then $\{x_k\}$ has a limit point x_* such that $\nabla^2 f(x_*)$ is positive semi–definite.

The global convergence to a stationary point was established by Powell [15] and Thomas [21]. The global convergence to a point where the Hessian is positive semi–definite was established by Sorensen [18]. Related results can be found in references [6], [8], [9], and [17]. The assumption on $\|H_k\|$ can be weakened. Powell [16] proved a convergence result in the case where there is a bound on the second–order approximation H_k that depends linearly on the iteration counter k.

3 The scaled quasi–Newton step

A major difference between the results that describe global convergence to a stationary point is that a uniform bound on H_k^{-1} is required for line searches but not for trust regions. Of course we are not making a fair comparison because the form of the step for trust regions was left unspecified whereas for line searches the step was taken along the quasi–Newton direction. In order to compare these global convergence results, let us take away the flexibility that the trust–region technique has to pick a direction and force it to move along the quasi–Newton direction. In other words the step s_k is now given by $-\xi_k H_k^{-1} g_k$, where

$$\xi_k = \begin{cases} \frac{\delta_k}{\|H_k^{-1} g_k\|} & \text{if} \quad \|H_k^{-1} g_k\| > \delta_k, \\ \\ 1 & \text{otherwise.} \end{cases} \tag{10}$$

We call this step a scaled quasi–Newton step and denote it by s_k^{N}.

If we want to establish global convergence to a stationary point, we need to make sure that the scaled quasi–Newton step satisfies the fraction of Cauchy decrease condition (8). The natural question to ask is: under what conditions does the scaled quasi–Newton step satisfy (8) ? We can go even further and ask: what do we need to assume to guarantee that such a step also satisfies the fraction of optimal decrease condition (9) ?

4 Global convergence for the scaled quasi–Newton step

We prove in this section that the answer to the questions formulated above is the existence of a uniform bound on the condition number of H_k.

Theorem 4.1 *If the condition number* $\mathbf{cn}(H_k)$ *of* H_k *is uniformly bounded, then the scaled quasi–Newton step* $s_k^{\mathsf{N}} = -\xi_k H_k^{-1} g_k$ *satisfies the fraction of Cauchy decrease condition (8).*

Proof. If $\xi_k = 1$, s_k^{N} is the optimal solution of the trust–region subproblem (2) and there is nothing else to prove. So, suppose that $\|H_k^{-1} g_k\| > \delta_k$. It follows from this and $\xi_k < 1$ that

$$\begin{aligned}
\Psi_k(0) - \Psi_k(s_k^{\mathsf{N}}) &= \frac{1}{2}\xi_k(2-\xi_k) g_k^T H_k^{-1} g_k \\
&> \frac{1}{2}\xi_k g_k^T H_k^{-1} g_k \\
&= \frac{1}{2}\delta_k \|g_k\| \frac{g_k^T H_k^{-1} g_k}{\|g_k\| \|H_k^{-1} g_k\|}. \qquad (11)
\end{aligned}$$

According to the definition of c_k given by (7), we either have $\frac{\|g_k\|^3}{g_k^T H_k g_k} \leq \delta_k$ in which case

$$\begin{aligned}
\Psi_k(0) - \Psi_k(c_k) &= \tfrac{\|g_k\|^4}{g_k^T H_k g_k} - \tfrac{1}{2}\tfrac{\|g_k\|^4}{(g_k^T H_k g_k)^2} g_k^T H_k g_k \\
&\leq \tfrac{1}{2}\delta_k \|g_k\|,
\end{aligned}$$

or $\frac{\|g_k\|^3}{g_k^T H_k g_k} > \delta_k$ which in turn gives

$$\begin{aligned}
\Psi_k(0) - \Psi_k(c_k) &= \delta_k \|g_k\| - \tfrac{1}{2}\tfrac{\delta_k^2}{\|g_k\|^2} g_k^T H_k g_k \\
&\leq \tfrac{1}{2}\delta_k \|g_k\|.
\end{aligned}$$

From this, (6), and (11), we get

$$\Psi_k(0) - \Psi_k(s_k^{\mathsf{N}}) \geq \frac{1}{2\nu}\Big(\Psi_k(0) - \Psi_k(c_k)\Big).$$

Thus s_k^{N} satisfies (8) with $\beta_1 = \frac{1}{2\nu}$. □

The following example is taken from [2] and indicates that without the uniform bound on the condition number, the scaled quasi–Newton step might not satisfy the fraction of Cauchy decrease condition.

Example 4.1 *Let us drop the subscripts* k *and consider* $H = diag(1, \epsilon^2, \epsilon^4)$ *and* $g = (\epsilon^2, \epsilon^2, \epsilon^3)^T$*, where* ϵ *is positive and small. With these choices we have*

$$H^{-1}g = \left(\epsilon^2, 1, \frac{1}{\epsilon}\right)^T, \quad \|H^{-1}g\| = \mathcal{O}\left(\frac{1}{\epsilon}\right), \quad g^T H^{-1} g = \mathcal{O}(\epsilon^2), \quad \textit{and} \quad \frac{\|g\|^3}{g^T H g} = \mathcal{O}(\epsilon^2).$$

Note that

$$\frac{g^T H^{-1} g}{\|g\| \|H^{-1} g\|} = \mathcal{O}(\epsilon).$$

If δ *is chosen very small, say* $\delta = \mathcal{O}(\epsilon^3)$ *then by (7) and (10),* $c = -\frac{\delta}{\|g\|} g$ *and* $\xi = \frac{\delta}{\|H^{-1}g\|}$*. As a result,* $\Psi(0) - \Psi(s^{\mathsf{N}}) = \mathcal{O}(\epsilon^6)$ *and* $\Psi(0) - \Psi(c) = \mathcal{O}(\epsilon^5)$*, which shows that as* ϵ *gets smaller and smaller the fraction of Cauchy decrease condition becomes more and more difficult to satisfy.*

Theorem 4.2 *If the condition number* $\mathbf{cn}(H_k)$ *of* H_k *is uniformly bounded, then the scaled quasi–Newton step* $s_k^{\mathsf{N}} = -\xi_k H_k^{-1} g_k$ *satisfies the fraction of optimal decrease condition (9).*

Proof. Again if $\xi_k = 1$, s_k^{N} is the optimal solution of the trust–region subproblem (2) and there is nothing else to prove. Let us assume that $\|H_k^{-1} g_k\| > \delta_k$. Since $\|s_k^*\| \le \delta_k < \|H_k^{-1} g_k\| \le \|H_k^{-1}\| \|g_k\|$, we have

$$\begin{aligned} \Psi_k(0) - \Psi_k(s_k^*) &= -g_k^T s_k^* - \tfrac{1}{2}(s_k^*)^T H_k (s_k^*) \\ &\le \|s_k^*\| \|g_k\| + \tfrac{1}{2}\|s_k^*\|^2 \|H_k\| \\ &\le \delta_k \|g_k\| + \tfrac{\nu}{2}\delta_k \|g_k\| \\ &= \left(1 + \tfrac{\nu}{2}\right) \delta_k \|g_k\|. \end{aligned}$$

From this, (6), and (11), we get

$$\begin{aligned} \Psi_k(0) - \Psi_k(s_k^{\mathsf{N}}) &\ge \tfrac{1}{2}\delta_k \|g_k\| \tfrac{g_k^T H_k^{-1} g_k}{\|g_k\| \|H_k^{-1} g_k\|} \\ &\ge \tfrac{1}{2}\tfrac{1}{\nu}\tfrac{1}{1+\frac{\nu}{2}} \Big(\Psi_k(0) - \Psi_k(s_k^*)\Big) \\ &\ge \tfrac{1}{2\nu+\nu^2} \Big(\Psi_k(0) - \Psi_k(s_k^*)\Big), \end{aligned}$$

and we see that the scaled quasi–Newton step satisfies (9) with $\beta_2 = \frac{1}{2\nu+\nu^2}$. □

Theorem 2 in [2] shows that if a step satisfies the fraction of Cauchy decrease (8) and there exists a uniform bound on the condition number of H_k, then such a step also satisfies the fraction of optimal decrease condition (9). Thus we could prove Theorem 4.2 by appealing to this earlier result, in conjunction with Theorem 4.1.

5 Final remarks

There are other interesting relationships between line searches and trust regions. For instance, the criteria to accept a step are very similar. Suppose that a line search only requires the Armijo–Goldstein–Wolfe condition (4) to accept a step s_k. This condition can be rewritten as

$$\frac{f(x_k) - f(x_k + s_k)}{-g_k^T s_k} \ge \alpha_1, \tag{12}$$

and it becomes evident how similar this is to the condition

$$\frac{f(x_k) - f(x_k + s_k)}{-g_k^T s_k - s_k^T H_k s_k} \ge \eta,$$

used in the trust–region technique. One can see that trust regions use curvature to accept or reject a step but line searches do not. However many practical implementations of line searches include second–order information in the sufficient decrease condition (4), or (12).

One final comment about the regularization issue is in order. It is also possible to regularize a line search by adding to H_k a positive multiple μI of the identity matrix. Of course one must choose μ and this becomes a performance issue that does not arise in trust regions. The solution s_k^* of the trust–region subproblem (2) satisfies the first–order necessary optimality conditions

$$\begin{aligned} &\Big(H_k + \mu I\Big) s_k^* = -g_k, \\ &\mu(\delta_k - \|s_k^*\|) = 0, \\ &\mu \geq 0,\ \|s_k^*\| \leq \delta_k. \end{aligned}$$

Here the parameter μ is implicitly defined by the size of the trust–region radius δ_k.

Acknowledgments

The author would like to thank John Dennis, Mahmoud El–Alem, Matthias Heinkenschloss, Richard Tapia, and Virginia Torczon for many interesting discussions on the topic of this paper.

References

[1] L. Armijo. Minimization of functions having Lipschitz–continuous first partial derivatives. *Pacific J. Math.*, 16:1–3, 1966.

[2] R. H. Byrd, R. B. Schnabel, and G. A. Shultz. Approximate solution of the trust region problem by minimization over two–dimensional subspaces. *Math. Programming*, 40:247–263, 1988.

[3] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice–Hall, Englewood Cliffs, New Jersey, 1983.

[4] J. E. Dennis and R. B. Schnabel. A view of unconstrained optimization. In G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd, editors, *Handbooks in Operations Research and Management Science.* North Holland, Amsterdam, 1988. (Vol. 1, Optimization).

[5] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, Chichester, second edition, 1987.

[6] D. M. Gay. Computing optimal locally constrained steps. *SIAM J. Sci. Statist. Comput.*, 2:186–197, 1981.

[7] A. A. Goldstein. On steepest descent. *SIAM J. Control Optim.*, 3:147–151, 1965.

[8] J. J. Moré. Recent developments in algorithms and software for trust regions methods. In A. Bachem, M. Grotschel, and B. Korte, editors, *Mathematical programming. The state of art*, pages 258–287. Springer Verlag, New York, 1983.

[9] J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM J. Sci. Statist. Comput.*, 4:553–572, 1983.

[10] J. J. Moré and D. Thuente. Line search algorithms with guaranteed sufficient decrease. *ACM Trans. Math. Software*, 20:286–307, 1994.

[11] S. G. Nash and A. Sofer. *Linear and Nonlinear Programming.* McGraw–Hill, New York, 1996.

[12] J. Nocedal. Theory of algorithms for unconstrained optimization. *Acta Numerica*, pages 199–242, 1992.

[13] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[14] M. J. D. Powell. A new algorithm for unconstrained optimization. In J. B. Rosen, O. L. Mangasarian, and K. Ritter, editors, *Nonlinear Programming.* Academic Press, New York, 1970.

[15] M. J. D. Powell. Convergence properties of a class of minimization algorithms. In O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, editors, *Nonlinear Programming 2*, pages 1–27. Academic Press, New York, 1975.

[16] M. J. D. Powell. On the global convergence of trust region algorithms for unconstrained minimization. *Math. Programming*, 29:297–303, 1984.

[17] G. A. Shultz, R. B. Schnabel, and R. H. Byrd. A family of trust–region–based algorithms for unconstrained minimization with strong global convergence properties. *SIAM J. Numer. Anal.*, 22:47–67, 1985.

[18] D. C. Sorensen. Newton's method with a model trust region modification. *SIAM J. Numer. Anal.*, 19:409–426, 1982.

[19] D. C. Sorensen. Minimization of a large scale quadratic function subject to an ellipsoidal constraint. Technical Report TR94–27, Department of Computational and Applied Mathematics, Rice University, 1994.

[20] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Numer. Anal.*, 20:626–637, 1983.

[21] S. W. Thomas. *Sequential Estimation Techniques for Quasi-Newton Algorithms.* PhD thesis, Cornell University, Ithaca, New York, 1975.

[22] Ph. L. Toint. Towards an efficient sparsity exploiting Newton method for minimization. In I. S. Duff, editor, *Sparse Matrices and Their Uses*, pages 57–87. Academic Press, New York, 1981.

[23] P. Wolfe. Convergent conditions for ascent methods. *SIAM Rev.*, 11:226–235, 1969.

[24] P. Wolfe. Convergent conditions for ascent methods. II: Some corrections. *SIAM Rev.*, 13:185–188, 1971.

[25] G. Zoutendijk. Nonlinear Programming, Computational Methods. In J. Abadie, editor, *Integer and Nonlinear Programming*, pages 37–86. North–Holland, Amsterdam, 1970.