Nonlinear Stepsize Control, Trust Regions and Regularizations for Unconstrained Optimization

by Ph. L. Toint[1]

Report 08/15 (revised) 10 February 2011

[1] Department of Mathematics,
FUNDP-University of Namur,
61, rue de Bruxelles, B-5000 Namur, Belgium.
Email: philippe.toint@fundp.ac.be

Nonlinear Stepsize Control, Trust Regions and Regularizations for Unconstrained Optimization

Ph. L. Toint

10 February 2011

Abstract

A class of algorithms for unconstrained optimization is introduced, which subsumes the classical trust-region algorithm and two of its newer variants, as well as the cubic and quadratic regularization methods. A unified theory of global convergence to first-order critical points is then described for this class.

Keywords: nonlinear optimization, unconstrained problems, global convergence.

1 Introduction

Unconstrained minimization and nonlinear least-squares problems are important instances of nonlinear programming, not only because of their own right, but also in view of the many other optimization problems which are reformulated as a (sequence of) problems of this type. As a consequence, algorithms which guarantee convergence to a local solution of nonconvex problems from arbitrary starting points are central and subject to intensive study. This "global" convergence property has traditionally been enforced by controlling the distance between two successive iterates of variants of Newton's method by either linesearch or trust-region techniques (see Nocedal and Wright, 1999, for a recent introduction to these techniques). These techniques are however strongly intertwined, and Shultz, Schnabel and Byrd (1985) and Toint (1988) independently observed that linesearch-based methods can often be reinterpreted as special cases of trust-region methods. Moreover, a common convergence theory can be derived that covers both classes (see Conn, Gould and Toint, 2000, Section 10.3, for a more recent exposition of this idea).

It is only recently that a third class of methods has emerged which also guarantees global convergence to local solutions for nonconvex problems. Elaborating on original ideas by Griewank (1981), Nesterov and Polyak (2006) and Weiser, Deuflhard and Erdmann (2007), Cartis, Gould and Toint (2009*a*) derived a general class of optimization methods where the distance between successive iterates is controlled by adaptive regularization. In this technique, an iteration-dependent cubic penalization of the steplength produces the desirable control. Remarkably, this class of algorithms enjoys all the good global convergence properties of trust-region methods, as well as an interesting worst-case complexity and, very crucially, a promising numerical efficiency. Unfortunately, this interesting development broke the unified setting where all efficient methods could be covered by a single convergence analysis. This situation deteriorated even further (from this specific point of view) when Nesterov (2007) proposed yet another (quadratic) regularization algorithm for nonlinear systems of equations. This method was then extended to general nonlinear least-squares problems by Bellavia, Cartis, Gould, Morini and Toint (2010), who also provided an again independent proof of global convergence for this algorithm. In parallel with this activity, new trust-region methods were also developed by Fan and Yuan (2001) for general unconstrained optimization and by Zhang and Wang (2003) and Fan (2006), Fan and Pan (2008*a*, 2008*b*) for the solution of nonlinear equations, where the rules of the trust-region updates are modified so that the radius converges to zero.

Because of this non-standard feature, each of these contributions again presents its own version of global convergence.

It is the purpose of this note to reconstruct a unifying framework in which global convergence can be proved for a class of methods, including standard and non-standard trust-region algorithms (and their linesearch avatars) as well as cubic and quadratic regularization schemes. This unification is based on the exploitation of technical similarities between the various existing proofs, but also has the advantage of providing a new potentially useful mechanism for nonlinear step control.

The paper is organized as follows. Section 2 introduces the new nonlinear stepsize control mechanism and the associated framework for unconstrained optimization algorithms, whose global convergence to first-order critical points is proved in Section 3. Some discussion of this result is finally provided in Section 4.

2 Nonlinear stepsize control

We first consider the general nonlinear and possibly nonconvex problem of unconstrained optimization, where one seeks to solve

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2.1}$$

for some objective function f from $\mathbb{R}^n$ into $\mathbb{R}$, which is assumed to be twice continuously differentiable and bounded below. The resolution methods considered here construct, from an arbitrary starting point $x_0 \in \mathbb{R}^n$, a sequence of iterates $\{x_k\}$ hopefully converging to a local solution of (2.1). The step s_k from the iterate x_k to x_{k+1} is constructed by the (possibly approximate) minimization of an iteration dependent model of the objective function $m_k(x_k + s)$ around x_k, subject to method specific step restrictions on $\|s\|$, where $\|\cdot\|$ is a norm on $\mathbb{R}^n$. This model is typically a suitable modification of the local quadratic

$$q_k(x_k + s) = f(x_k) + \langle s, g_k \rangle + \tfrac{1}{2}\langle s, H_k s \rangle \tag{2.2}$$

where g_k denotes the gradient $\nabla_x f(x_k)$ and where H_k is a symmetric matrix approximating the second-order behaviour of f in the neighbourhood of x_k. Depending on the choice of model and the specific stepsize restriction, various sufficient decrease conditions may then be imposed to facilitate convergence proofs. Typically, these conditions are derived from the minimization of the chosen model along the steepest descent direction.

- In *trust-region methods*, the objective function $f(x)$ is assumed to be twice continuously differentiable and the model chosen is exactly (2.2), that is

$$m_k(x_k + s) = q_k(x_k + s) \tag{2.3}$$

and the stepsize is restricted by the explicit constraint

$$\|s_k\| \le \Delta_k \tag{2.4}$$

for some adaptive "radius" $\Delta_k > 0$. In this case, it is well-known that sufficient decrease is guaranteed by the so-called Cauchy-point condition, stating that

$$m_k(x_k) - m_k(x_k^+) \ge \kappa_{\rm TR} \|g_k\| \min\left[\frac{\|g_k\|}{1 + \|H_k\|}, \Delta_k\right] \tag{2.5}$$

for some constant $\kappa_{\rm TR} \in (0,1)$ and where $x_k^+ = x_k + s_k$ (see Conn et al., 2000, Section 6.3, for a detailed derivation of this inequality originally due to Powell, 1970).

- The *cubic regularization method* assumes that $f(x)$ has Lipschitz continuous gradients and chooses an indirect way to control the stepsize, in that the step (approximately) minimizes the model

$$m_k(x_k+s) = q_k(x_k+s) + \tfrac{1}{3}\sigma_k\|s\|^3 \tag{2.6}$$

on the whole of $\mathbb{R}^n$, where σ_k is an adaptive regularization parameter and $\|\cdot\|$ an ellipsoidal norm. For such steps, Cartis et al. (2009*a*) show that "sufficient decrease" is given by

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\rm CR}\|g_k\| \min\left[\frac{\|g_k\|}{1+\|H_k\|}, \sqrt{\frac{\|g_k\|}{\sigma_k}}\right] \tag{2.7}$$

for some constant $\kappa_{\rm CR} \in (0,1)$, while the stepsize resulting from the unconstrained minimization of (2.6) satisfies the bound

$$\|s_k\| \leq 3\max\left[\frac{\|H_k\|}{\sigma_k}, \sqrt{\frac{\|g_k\|}{\sigma_k}}\right]. \tag{2.8}$$

- The *quadratic regularization method* of Nesterov (2007) (as extended by Bellavia et al., 2010) only applies to nonlinear least-squares problems where

$$f(x) = \|F(x)\| \tag{2.9}$$

for some smooth function F from $\mathbb{R}^n$ to $\mathbb{R}^m$ with locally Lipschitz continuous Jacobian, and where $\|\cdot\|$ is the Euclidean norm. In this method, the model which is minimized to calculate the step is given by

$$m_k(x_k+s) = \|F(x_k) + J(x_k)s\| + \sigma_k\|s\|^2, \tag{2.10}$$

where $J(x)$ is the Jacobian of F at x and where σ_k is again an adaptive regularization parameter. Note that this model is non-differentiable, but Cartis, Gould and Toint (2009*b*) show that the problem of minimizing (2.10) on $\mathbb{R}^n$ can be reformulated as a smooth constrained problem. In this framework, the ensured "sufficient decrease" turns out to be given by

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\rm QR}\frac{\|J_k^TF_k\|}{\|F_k\|}\min\left[\frac{\|J_k^TF_k\|}{1+\|H_k\|}, \frac{\|J_k^TF_k\|}{\sigma_k\|F_k\|}\right] \tag{2.11}$$

for some constant $\kappa_{\rm QR} \in (0,1)$, where

$$H_k = J_k^TJ_k. \tag{2.12}$$

The restriction that (2.10) imposes on the stepsize is also indirect, as it can be proved that

$$\|s_k\| \leq 2\frac{\|J_k^TF_k\|}{\sigma_k\|F_k\|}. \tag{2.13}$$

The reader is referred to Bellavia et al. (2010) for further details on the derivation of (2.11) and (2.13).

- The unconstrained optimization *method by Fan and Yuan (2001)* considers problem (2.1) and is similar to the classical trust-region method in that its step is computed by minimizing the model (2.2) inside a trust region of radius Δ_k. However, this radius takes the form

$$\Delta_k = \mu_k\|g_k\| \tag{2.14}$$

for some parameter μ_k which is then updated in a manner similar to the classical trust-region radius update. In this algorithm, the sufficient decrease condition is given by

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\rm FY} \|g_k\| \min\left[\frac{\|g_k\|}{1+\|H_k\|}, \mu_k \|g_k\|\right] \tag{2.15}$$

for some constant $\kappa_{\rm FY} > 0$.

- As was the case for the quadratic regularization method, the *algorithm by Zhang and Wang (2003) and Fan (2006)* addresses the solution of nonlinear systems in the least-squares sense and considers (at variance with (2.9)), an objective function of the form

$$f(x) = \tfrac{1}{2}\|F(x)\|^2.$$

The step s_k is then computed by minimizing the Gauss-Newton model

$$m_k(x_k + s) = \tfrac{1}{2}\|F(x_k) + J(x_k)s\|^2, \tag{2.16}$$

within a trust-region of radius Δ_k, which is chosen as

$$\Delta_k = \nu^j \|F(x_k)\|^\gamma \tag{2.17}$$

for some $\nu \in (0,1)$ and $\gamma \in (\frac{1}{2}, 1)$, and where j is reset to zero on successful iterations (i.e., when a new iterate is accepted) and incremented by one otherwise[1]. This mechanism is therefore close to a "backtracking trust region", whose initial size is chosen as some power of $\|F(x_k)\|$. For this algorithm, sufficient reduction is described by the usual Cauchy condition for nonlinear least-squares problems, which then gives that

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\rm ZW} \|J_k^T F_k\| \min\left[\frac{\|J_k^T F_k\|}{1+\|H_k\|}, \nu^j \|F_k\|^\gamma\right] \tag{2.18}$$

for some constant $\kappa_{\rm ZW} \in (0,1)$ (where H_k is defined by (2.12)). This method was extended by Fan and Pan (2008*b*) to include the case $\gamma = 1$, and a variant was analyzed in Fan and Pan (2008*a*) where the model is now given, instead of (2.16), by

$$m_k(x_k + s) = \tfrac{1}{2}\|F(x_k) + J(x_k)s\|^2 + \theta\|F_k\|^2 \, \|s\|^2,$$

for some $\theta > 0$. The model reduction is then again given by (2.18) with H_k now being defined as $J_k^T J_k + \theta\|F_k\|^2$.

In all cases, the ratio of achieved vs. predicted reduction

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} \tag{2.19}$$

is computed and the (method dependent) adaptive parameter is modified to reduce the stepsize if this ratio is below some small $\eta_1 \in (0,1)$, or left unchanged or modified to allow an increase in the stepsize otherwise.

As is implied by the above description of these methods, it is obvious that they share much in structure. In particular, the sufficient decrease condition features in all five cases the minimum between two quantities, the first of which corresponding to the case where the minimization of q_k dominates the step definition and the second to the case where the stepsize is explicitly or implicitly limited. One may therefore expect something similar for all techniques which mix these two potentially conflicting objectives, and providing a single framework with the aim of unifying the convergence theory therefore seems a natural objective.

[1] The update described here is that proposed by Zhang and Wang (2003). Fan (2006) uses a very similar update, which is closer to that of Fan and Yuan (2001).

Our proposal is to do so by defining, in (2.22) below, a function Δ whose purpose is to provide an explicit bound on the step (therefore mimicking trust-region methods), but at the same time avoiding to identify the value of this bound automatically with the adaptive parameter used in the method (now in contrast with classical trust-region methods). More specifically, our function Δ will be assumed to be nonnegative and to depend on two variables. The first is a nonnegative adaptive parameter, which we will denote by δ, the second essentially reduces to some measure of first-order criticality computed at the current iterate. However, this description is not sufficient to cover the various forms of Cauchy conditions given by (2.5), (2.7), (2.11), (2.15) and (2.18). In order to cover all cases, we introduce four functions satisfying the following conditions.

A.1 There exists a continuous, bounded and nonnegative function $\omega(x)$ such that $\omega(x) = 0$ only if x is a first order critical point.

A.2 There exist three continuous nonnegative functions $\phi(x)$, $\psi(x)$ and $\chi(x)$, possibly undefined at roots of $\omega(x)$, such that, provided $\omega(x) > 0$, then $\min[\phi(x), \psi(x), \chi(x)]$ is zero at x only if x is a first order critical point.

A.3 The function $\chi(x)$ is bounded, in the sense that

$$\chi(x) \leq \kappa_\chi \quad \text{for all} \quad x. \tag{2.20}$$

Note that ϕ, ψ and χ need not be different. By convention, we use the notation

$$\phi_k = \phi(x_k), \quad \psi_k = \psi(x_k), \quad \chi_k = \chi(x_k) \quad \text{and} \quad \omega_k = \omega(x_k).$$

Using these functions, we may then state our sufficient-decrease and steplength conditions.

A.4: The step s_k produces a decrease in the model which is sufficient in the sense that

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\rm C} \psi_k \min\left[\frac{\phi_k}{1 + \|H_k\|}, \Delta(\delta_k, \chi_k)\right] \tag{2.21}$$

for some $\kappa_{\rm C} \in (0,1)$, where we define

$$\Delta(\delta, \chi) = \delta^\alpha \chi^\beta \tag{2.22}$$

for some powers $\alpha \in (0,1]$ and $\beta \in [0,1]$, and where H_k is the Hessian of the model at x_k.

A.5: The step s_k satisfies the bound

$$\|s_k\| \leq \kappa_{\rm s} \Delta(\delta_k, \chi_k) \quad \text{whenever} \quad \delta_k \leq \kappa_\delta \chi_k$$

for some $\kappa_{\rm s} \geq 1$ and $\kappa_\delta > 0$.

We complete our set of assumptions by requiring that the model approximates the objective function sufficiently well at the trial point.

A.6: We have that, for all $k \geq 0$,

$$f(x_k) = m_k(x_k) \quad \text{and} \quad f(x_k + s) - m_k(x_k + s) \leq \kappa_m \|s\|^2 \tag{2.23}$$

for some constant $\kappa_m > 0$.

Note that this assumption may require additional smoothness properties of the objective function to hold.

For future reference, we state a few immediate properties of the function $\Delta(\delta, \chi)$ defined by (2.22).

Theorem 2.1

1. *The function* $\Delta(\delta, \chi)$ *is concave for* $\delta, \chi \geq 0$.
2. *As a function of* δ*,* $\Delta(\delta, \chi)$ *is continuously differentiable and strictly increasing for* $\delta > 0$.
3. *For all* $\chi \geq 0$*,* $\Delta(0, \chi) = 0$.
4. $\Delta(\delta, \chi)$ *is a non-decreasing function of* χ *for* $\chi \geq 0$.

We are now in position to describe our algorithmic framework explicitly, as Algorithm 2.1 on this page. Note that the model choice at Step 1 of Algorithm 2.1 is relatively general in that it does not imply a particular choice of H_k in (2.2), nor does it impose the explicit use of regularization. As in a trust-region algorithm, we will say that iteration k is successful whenever $\rho_k \geq \eta_1$ and very successful whenever $\rho_k \geq \eta_2$.

Algorithm 2.1: Nonlinear Stepsize Control Algorithm

Step 0: Initialization. An initial point $x_0 \in \mathbb{R}^n$ and an initial stepsize parameter δ_0 are given, as well as constants $0 < \gamma_1 < \gamma_2 < 1$ and $0 < \eta_1 \leq \eta_2 < 1$. Set $k = 0$.

Step 1: Step computation. Choose a model $m_k(x_k + s)$ satisfying A.6 and find a step s_k which sufficiently reduces the model in the sense of A.4 and for which $\|s_k\|$ satisfies A.5.

Step 2: Step acceptance. Compute $f(x_k + s_k)$ and the ratio ρ_k given by (2.19). Set $x_{k+1} = x_k + s_k$ if $\rho_k \geq \eta_1$, and $x_{k+1} = x_k$ otherwise.

Step 3: Stepsize parameter update. Choose

$$\delta_{k+1} \in \left\{ \begin{array}{lll} [\gamma_1 \delta_k, \gamma_2 \delta_k] & \text{if} & \rho_k < \eta_1, \\ [\gamma_2 \delta_k, \delta_k] & \text{if} & \rho_k \in [\eta_1, \eta_2), \\ [\delta_k, +\infty] & \text{if} & \rho_k \geq \eta_2. \end{array} \right. \tag{2.24}$$

Increment k by one and go to Step 1.

We now verify that the algorithm class we have defined covers the cases of interest which we have mentioned before,

- Consider the trust-region method first. In this case, it is easy to verify that the choices

  $$\omega(x) = 1, \quad \psi(x) = \phi(x) = \chi(x) = \|\nabla_x f(x)\|,$$
  $$\delta = \Delta, \quad \alpha = 1, \quad \beta = 0.$$

  are adequate, and Taylor's theorem implies that

  $$f(x_k + s_k) - q_k(x_k + s_k) \leq \kappa_{\rm H} \|s_k\|^2, \tag{2.25}$$

  for $\kappa_{\rm H}$, an upper bound on the norms of H_k and $\nabla_{xx} f(x)$ (see Theorem 6.4.1, p. 133, in Conn et al., 2000). A.6 thus immediately follows with $\kappa_m = \kappa_{\rm H}$. Moreover, (2.5) then implies A.4 with $\kappa_{\rm C} = \kappa_{\rm TR}$, and A.5 with $\kappa_{\rm s} = 1$ directly follows from (2.4).

- The case of the cubic regularization algorithm can be similarly handled, with the choices
$$\omega(x) = 1, \quad \psi(x) = \phi(x) = \chi(x) = \|\nabla_x f(x)\|,$$
$$\delta = \frac{1}{\sigma}, \quad \alpha = \tfrac{1}{2}, \quad \beta = \tfrac{1}{2}.$$
Note that the condition that $\delta_k \leq \chi_k/\kappa_{\rm H}^2$, where $\kappa_{\rm H}$ is again an upper bound on $\|H_k\|$ and $\|\nabla_{xx} f(x)\|$, implies that
$$\sqrt{\delta_k}\kappa_{\rm H} \leq \sqrt{\chi_k}$$
and hence
$$\frac{\|H_k\|}{\sigma_k} \leq \delta_k \kappa_{\rm H} \leq \sqrt{\delta_k \chi_k} = \Delta(\delta_k, \chi_k) = \sqrt{\frac{\|g_k\|}{\sigma_k}},$$
which in turn ensures that (2.8) implies A.5 with $\kappa_\delta = \frac{1}{\kappa_{\rm H}^2}$ and $\kappa_{\rm s} = 3$. A.4 immediately follows from (2.7) with $\kappa_{\rm C} = \kappa_{\rm CR}$. The bound (2.25) and (2.6) also imply that
$$f(x_k + s_k) - m_k(x_k + s_k) \leq \kappa_{\rm H}\|s_k\|^2 - \tfrac{1}{3}\sigma_k\|s_k\|^3 \leq \kappa_{\rm H}\|s_k\|^2,$$
and hence A.6 again follows with $\kappa_m = \kappa_{\rm H}$.

- Consider now the quadratic regularization method. In this case, we may choose
$$\omega(x) = \|F(x)\|, \quad \psi(x) = \chi(x) = \frac{\|J(x)^T F(x)\|}{\|F(x)\|}, \quad \phi(x) = \|J(x)^T F(x)\|$$
$$\delta = \frac{1}{\sigma}, \quad \alpha = 1, \quad \beta = 1.$$
These identifications give that
$$\Delta(\delta_k, \chi_k) = \frac{\|J_k^T F_k\|}{\sigma_k \|F_k\|}.$$
The condition (2.11) gives A.4 with $\kappa_{\rm C} = \kappa_{\rm QR}$, while (2.13) gives A.5 with $\kappa_{\rm S} = 2$. A.6 (with $\kappa_m = L/2$) also follows from the mean-value theorem and the Lipschitz continuity (with constant L) assumption on the Jacobian (see Lemma 3.5 in Bellavia et al., 2010). Note that, in this case, the gradient of the objective function and the model do not coincide, but are merely collinear.

- If we now turn to the method by Fan and Yuan (2001), we see that the choices
$$\omega(x) = 1, \quad \psi(x) = \phi(x) = \chi(x) = \|\nabla_x f(x)\|,$$
$$\delta = \mu \quad \alpha = 1, \quad \beta = 1$$
are adequate to cover this case and give (2.14) back. As for the trust-region method, A.4 with $\kappa_{\rm C} = \kappa_{\rm FY}$ is ensured by (2.15), and A.6 (with $\kappa_{\rm m} = \kappa_{\rm H}$) follows from Taylor's theorem. The step bound in A.5 (with $\kappa_{\rm s} = 1$) is guaranteed by the explicit trust-region constraint.

- Finally, the method by Zhang and Wang (2003) and Fan (2006) is covered by the choices
$$\omega(x) = 1, \quad \psi(x) = \phi(x) = \|J(x)^T F(x)\|, \quad \chi(x) = \|F(x)\|^\gamma,$$
$$\delta = \nu^j \quad \alpha = 1, \beta = 1.$$
Condition (2.18) gives A.4 with $\kappa_{\rm C} = \kappa_{\rm ZW}$ and A.5 (with $\kappa_{\rm s} = 1$) follows from (2.17). In this case, we use the flexibility left in (2.24) to reset δ to 1 at every successful iteration and we choose $\gamma_1 = \gamma_2 = \alpha$. A.6 (with $\kappa_{\rm m} = \kappa_{\rm H}$) directly follows from Taylor's theorem. The case where $\gamma = 1$ is not special in our context.

We note that the sequence $\{\omega_k\}$ is non-increasing in all four cases.

3 Global convergence to first-order critical points

As a consequence of A.1 and A.2, proving global convergence to first-order critical points can be reduced to proving that ω_k or one of ψ_k, ϕ_k and χ_k approaches zero asymptotically. It is the objective of this section to prove this result.

Since our general class of algorithms involves an explicit constraint on the stepsize, it is not surprising that the necessary convergence theory borrows its basic organisation and a number of the technicalities to that available for trust-region methods (we refer the reader to Chapter 6 of Conn et al. (2000) for a detailed exposition of this topic).

We start by the obvious observation that, if

$$\liminf_{k\to\infty} \omega_k = 0, \tag{3.1}$$

then we obtain our desired convergence result from A.1. In particular, there must be a first-order critical limit point of the sequence of iterates $\{x_k\}$ if this sequence is bounded. Moreover, we may replace the limit inferior by a true limit (and all limit points, if any, must be first-order critical) if the sequence $\{\omega_k\}$ is non-increasing. We therefore focus, in the remaining of our analysis, on the case where

$$\omega_k \geq \epsilon_\omega \quad (k \geq 0), \tag{3.2}$$

for some $\epsilon_\omega \in (0,1]$. In this case, A.2 ensures that $\psi(x)$, $\phi(x)$ and $\chi(x)$ are all proper criticality measures.

As in the trust-region case, global convergence depends on the additional assumption that the Hessians of the model and objective function are uniformly bounded.

A.7: There exists a constant $\kappa_{\rm H} \geq 1$ such that, for all k, $1 + \|H_k\| \leq \kappa_{\rm H}$.

Note that we assume, without loss of generality, that

$$\kappa_m \leq \kappa_{\rm H}. \tag{3.3}$$

Also note that the method by Fan and Pan (2008*a*) is then covered by our analysis as well, because the Hessian of the model is, in this case, only modified by a diagonal matrix whose norm is bounded by $\theta\|F_k\|^2 \leq \theta\|F_0\|^2$, which does not affect A.7.

We now prove that iteration k must be very successful if the stepsize parameter and the step bound $\Delta(\delta_k, \chi_k)$ are small enough compared to a criticality measure at x_k.

Lemma 3.1 *If*

$$\Delta(\delta_k, \chi_k) \leq \frac{\kappa_{\rm C}(1-\eta_2)}{\kappa_{\rm H}\kappa_{\rm s}^2} \min[\phi_k, \psi_k] \quad \textit{and} \quad \delta_k \leq \kappa_\delta \chi_k, \tag{3.4}$$

then iteration k is very successful and $\delta_{k+1} \geq \delta_k$.

Proof. On one hand, we know from the second part of A.6 that

$$f(x_k+s_k) - m_k(x_k+s_k) \leq \kappa_m \|s_k\|^2 \leq \kappa_{\rm H}\kappa_{\rm s}^2 \Delta(\delta_k, \chi_k)^2 \tag{3.5}$$

where we used A.5 and (3.3) to derive the last inequality. On the other hand, the first part of (3.4) and the bounds $1-\eta_2 \in (0,1)$, A.7, $\kappa_{\rm s} \geq 1$ and $\kappa_{\rm C} \in (0,1)$ imply that

$$\Delta(\delta_k, \chi_k) \leq \frac{\phi_k}{1+\|H_k\|}$$

and hence A.4 gives that

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\rm C}\psi_k \Delta(\delta_k, \chi_k). \tag{3.6}$$

Combining this inequality with (3.5), we obtain, using the first part of A.6 and the first part of (3.4), that

$$1 - \rho_k = \frac{f(x_k + s_k) - m_k(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)} \le \frac{\kappa_{\rm H}\kappa_{\rm s}^2 \Delta(\delta_k, \chi_k)}{\kappa_{\rm C}\psi_k} \le 1 - \eta_2.$$

As a consequence, $\rho_k \ge \eta_2$, the iteration is very successful and $\delta_{k+1} \ge \delta_k$. □

We next show that the stepsize bound cannot shrink to zero unless the current iterate is first-order critical.

Lemma 3.2 *Assume now that (3.2) holds and that, for some $\epsilon > 0$,*

$$\min[\psi_k, \phi_k, \chi_k] \ge \epsilon \quad \textit{for all} \quad k \ge 0. \tag{3.7}$$

Then there exists a constant $\Delta_{\min}(\epsilon) > 0$ such that

$$\Delta(\delta_k, \chi_k) \ge \Delta_{\min}(\epsilon) \quad \textit{for all} \quad k \ge 0. \tag{3.8}$$

Proof. Suppose that iteration $k > 0$ is such that

$$\Delta(\delta_k, \chi_k) < \left[1 + \kappa_\Delta\left(\frac{1}{\gamma_1} + \frac{\kappa_\chi}{\epsilon}\right)\right]^{-1} \frac{\kappa_{\rm C}(1-\eta_2)\epsilon}{\kappa_{\rm H}\kappa_{\rm s}^2} \stackrel{\rm def}{=} \underline{\Delta}(\epsilon) \tag{3.9}$$

where $\kappa_\Delta = \max[\alpha, \beta$, and

$$\delta_k \le \gamma_1 \min\left[\kappa_\delta \epsilon, \left[\frac{\underline{\Delta}(\epsilon)}{\kappa_\chi^\beta}\right]^{1/\alpha}\right] \stackrel{\rm def}{=} \underline{\delta}(\epsilon), \tag{3.10}$$

Assume now, for the purpose of deriving a contradiction, that iteration $k-1$ is not very successful. Then (2.24) implies that

$$\gamma_1 \delta_{k-1} \le \delta_k \le \delta_{k-1}. \tag{3.11}$$

This then gives that

$$\delta_{k-1} \le \kappa_\delta \epsilon \le \kappa_\delta \chi_{k-1}, \tag{3.12}$$

where we used (3.11), (3.10) and (3.7) successively. We also verify that

$$\begin{array}{rcl}
\Delta(\delta_{k-1}, \chi_{k-1}) & \le & \Delta(\delta_k, \chi_k) + (\delta_{k-1} - \delta_k)\dfrac{\partial\Delta}{\partial\delta}(\delta_k, \chi_k) + (\chi_{k-1} - \chi_k)\dfrac{\partial\Delta}{\partial\chi}(\delta_k, \chi_k) \\
& \le & \Delta(\delta_k, \chi_k) + \dfrac{1-\gamma_1}{\gamma_1}\,\delta_k \dfrac{\partial\Delta}{\partial\delta}(\delta_k, \chi_k) + \dfrac{1 - (\epsilon/\kappa_\chi)}{(\epsilon/\kappa_\chi)}\,\chi_k \dfrac{\partial\Delta}{\partial\chi}(\delta_k, \chi_k) \\
& \le & \left[1 + \kappa_\Delta\left(\dfrac{1}{\gamma_1} + \dfrac{\kappa_\chi}{\epsilon}\right)\right]\Delta(\delta_k, \chi_k) \\
& < & \dfrac{\kappa_{\rm C}\epsilon\,(1-\eta_2)}{\kappa_{\rm H}\kappa_{\rm s}^2} \\
& \le & \dfrac{\kappa_{\rm C}(1-\eta_2)}{\kappa_{\rm H}\kappa_{\rm s}^2}\min[\psi_{k-1}, \phi_{k-1}]
\end{array}$$

where we used the concave nature of Δ to deduce the first inequality, (3.11) and the inequality

$$\frac{\chi_k}{\chi_{k-1}} \ge \frac{\epsilon}{\kappa_\chi}$$

(itself resulting from A.3 and (3.7)) to deduce the second, (2.22) and (3.9) to obtain the third and (3.10) to obtain the fourth. The last equality finally results from (3.7). Hence, Lemma 3.1 and (3.12) ensure that iteration $k-1$ is very successful. But this contradicts our assumption and thus iteration $k-1$ must be very successful, implying that $\delta_k \geq \delta_{k-1}$ as soon as (3.9) and (3.10) hold. Therefore, the first iteration k such that (3.10) occurs must be such that (3.9) fails at iteration $k-1$. But, using (2.22), this implies that

$$\delta_{k-1} \geq \left[\frac{\underline{\underline{\Delta}}(\epsilon)}{\chi_{k-1}^{\beta}}\right]^{1/\alpha}$$

The mechanism of the algorithm then implies that

$$\delta_k \geq \gamma_1 \left[\frac{\underline{\underline{\Delta}}(\epsilon)}{\chi_{k-1}^{\beta}}\right]^{1/\alpha} \geq \gamma_1 \left[\frac{\underline{\underline{\Delta}}(\epsilon)}{\kappa_{\chi}^{\beta}}\right]^{1/\alpha}.$$

This clearly violates (3.10), which is also a contradiction. As a consequence, an iteration satisfying (3.10) cannot exist, and we therefore obtain that, for all $k > 0$,

$$\delta_k \geq \underline{\delta}(\epsilon).$$

The definition (2.22) and (3.7) then imply that, for $k > 0$,

$$\Delta(\delta_k, \chi_k) \geq \Delta(\underline{\delta}(\epsilon), \epsilon) > 0,$$

yielding in turn that, for all $k \geq 0$,

$$\Delta(\delta_k, \chi_k) \geq \min\left[\Delta(\underline{\delta}(\epsilon), \epsilon), \Delta(\delta_0, \epsilon)\right] \stackrel{\rm def}{=} \Delta_{\min}(\epsilon). \tag{3.13}$$

□

This proof is the most significant deviation from the trust-region convergence analysis, because the relation between the parameter δ and the radius $\Delta(\delta, \chi)$ is now indirect and potentially nonlinear. From here on, the global first-order convergence analysis (for the case where (3.2) holds) follows the trust-region theory closely, with the obvious substitution of $\Delta(\delta_k, \chi_k)$ for Δ_k and where $\|g_k\|$ is replaced by $\min[\psi_k, \phi_k, \chi_k]$, which is a proper criticality measure when (3.2) holds (as ensured by A.2). We outline this analysis below.

Lemma 3.3 *Suppose (3.2) holds and that there are only finitely many successful iterations. Then $x_k = x_*$ for all k sufficiently large and x_* is first-order critical.*

Proof. The mechanism of the algorithm implies that $x_k = x_{k_0+1} \stackrel{\rm def}{=} x_*$ for all $k \geq k_0$, where k_0 is the index of the last successful iteration. Moreover, since all iterations beyond k_0 are unsuccessful, we have that the sequence $\{\delta_k\}$ converges to zero. A.6 then implies that $\{\Delta_k\}$ also converges to zero, which, by Lemma 3.2, is impossible unless $\{\min[\psi_k, \phi_k, \chi_k]\}$ converges to zero as well. But $\psi_k = \psi_{k_0+1} = \psi(x_*)$ for all $k \geq k_0$ and similarly $\phi_k = \phi(x_*)$ and $\chi_k = \chi(x_*)$. A.2 then ensures that x_* is first-order critical. □

Theorem 3.4 *We have that*

$$\liminf_{k\to\infty} \omega_k = 0 \quad or \quad \liminf_{k\to\infty} \min[\psi_k, \phi_k, \chi_k] = 0, \tag{3.14}$$

and at least one limit point of the sequence $\{x_k\}$ (if any exists) is first-order critical.

Proof. If (3.1) holds, then the desired conclusion follows from A.1. Otherwise, that is if (3.2) holds, then we distinguish two cases. If there are only finitely many successful iterations, then Lemma 3.3 ensures the desired conclusion. Suppose therefore that there are infinitely many successful iterations and that (3.7) holds. Then, A.4 and A.7 give that, for each successful iteration,

$$f(x_k) - f(x_{k+1}) \geq \eta_1(m_k(x_k) - m_k(x_k + s_k)) \geq \kappa_{\rm C}\eta_1\epsilon \min\left[\frac{\epsilon}{\kappa_{\rm H}}, \Delta(\delta_k, \chi_k)\right].$$

Using now Lemma 3.2, we deduce that, for every such iteration,

$$f(x_k) - f(x_{k+1}) \geq \kappa_{\rm C} \eta_1 \epsilon \min\left[\frac{\epsilon}{\kappa_{\rm H}}, \Delta_{\min}(\epsilon)\right] > 0,$$

and the objective function then decreases at least by a positive constant. Since the number of successful iterations is infinite, one concludes that $\{f(x_k)\}$ must tend to $-\infty$, which contradicts our assumption that $f(x)$ is bounded below. Hence (3.7) cannot hold and we deduce that the second limit in (3.14) holds. The conclusion then follows from A.2 since it guarantees that $\min[\psi(x), \phi(x), \chi(x)]$ is a criticality measure under (3.2). □

Theorem 3.5 *Suppose that the sequence $\{\omega_k\}$ is non-increasing. Then we have that*

$$\lim_{k\to\infty} \omega_k = 0 \quad \textit{or} \quad \lim_{k\to\infty} \min[\psi_k, \phi_k, \chi_k] = 0. \tag{3.15}$$

and all limit points of the sequence $\{x_k\}$ (if any) are first-order critical.

Proof. The desired conclusion immediately follows from monotonicity and A.1 if (3.1) holds. Otherwise, an obvious extension of Theorems 6.4.6, p. 136-138, in Conn et al. (2000), where the continuity of the criticality measure $\min[\psi(x), \phi(x), \chi(x)]$ here replaces that of the objective function's gradient, allows us to derive that the second limit of (3.15) holds. Again, A.1 and A.2 allow us to conclude that every limit point (if any) is first-order critical. □

We conclude our discussion with the comment that our theory does not require the same definition of $\Delta(\delta, \chi)$ to be used at every iteration. One could vary this definition as the algorithm proceeds, as long as A.5–A.8 remain satisfied.

4 Discussion

We have shown how a single framework can be used to prove global convergence to first-order critical points for several methods for unconstrained optimization and nonlinear least-squares, for which the analysis was so far distinct. This is achieved by defining a variant of the trust-region radius which depends possibly nonlinearly on some stepsize parameter which is updated from iteration to iteration.

While we have focussed in this paper on algorithms for unconstrained optimization, the same ideas can be applied more widely. For instance, projection-based trust-region algorithms for optimization of a (possibly nonconvex) objective function $f(x)$ over a convex set $\mathcal{C}$ (see Conn, Gould and Toint, 1988, Burke, Moré and Toraldo, 1990, Conn, Gould, Sartenaer and Toint, 1993, or Chapter 12 in Conn et al., 2000) also involve a trust-region mechanism and a Cauchy condition of the form

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\rm CTR} \tau_k \min\left[\frac{\tau_k}{1 + \|H_k\|}, \Delta_k, 1\right] \tag{4.1}$$

where $\kappa_{\rm CTR} \in (0,1)$ and where the continuous criticality measure $\tau(x)$ is defined, for some norm $\|\cdot\|_n$, by

$$\tau(x) = |\min_{x+d\in\mathcal{C}, \|d\|_n \leq 1} \langle \nabla_x f(x), d\rangle|.$$

Algorithms of this type therefore fit in our framework with the choices

$$\omega(x) = \chi(x) = 1, \quad \psi(x) = \phi(x) = \min[\tau(x), 1], \quad \delta = \Delta,$$

$$\alpha = 1, \quad \beta = 0, \quad \kappa_\Delta = 1, \quad \kappa_\delta = +\infty,$$

thereby demonstrating that our approach is not limited to unconstrained problems.

Interestingly, the unified framework can be used to design new unconstrained optimization methods. For instance, one can define a nonlinear trust-region algorithm where the model (2.3) is minimized under the constraint (2.4) (which ensures (2.5)), but where the radius is now determined by the choice

$$\psi(x) = \phi(x) = \chi(x) = \|\nabla_x f(x)\|, \quad \omega(x) = 1, \quad \alpha = \tfrac{1}{2}, \quad \beta = \tfrac{1}{2}, \quad \kappa_\Delta = \frac{1}{2}, \quad \text{and} \quad \kappa_\delta = +\infty.$$

Such a method can be viewed as a hybrid between the standard trust-region algorithm and the cubic overestimation/regularization method, but this is only one example. The numerical exploration of the algorithmic possibilities within the new framework is potentially interesting, but is beyond the scope of this note.

We also observe that, although a convergence theory has been obtained for a large class of algorithms, it does not, at this stage, provide a useful handle for analyzing the worst-case function-evaluation complexity of the algorithms in the class. Indeed, relations (3.9) and (3.13) coupled with A.4 suggest that the model decrease at the Cauchy point might be as small as a multiple of ϵ^3, which is in turn likely to yield a worst-case complexity of $O(\epsilon^{-3})$ iterations to achieve first-order criticality within ϵ. This is worse than the known bounds for the trust-region and cubic-regularization methods, which is known[2] to be of $O(\epsilon^{-2})$ (see Cartis, Gould and Toint, 2010*a*, and Cartis, Gould and Toint, 2010*b*). Whether our analysis can be refined to yield less pessimistic complexity estimates is unclear, and it may be that this is the price to pay for increased generality.

The convergence theory for standard trust-region algorithms admits itself a large number of useful generalizations. The author expects that many of them can be adapted to the more general context discussed here. In particular, one may think of extensions in the spirit of the retrospective trust-region algorithm by Bastin, Malmedy, Mouffe, Toint and Tomanos (2010), or to multilevel frameworks presented by Gratton, Sartenaer and Toint (2008). It is also of interest to see whether the same type of idea can be applied more widely, for instance to the theory of convergence to second-order critical points, or to algorithms for equality- or inequality-constrained problems beyond projection-based methods.

Acknowledgements

The author is indebted to B. Morini, D. Tomanos and M. Porcelli for their suggestions on the manuscript, and to two anonymous referees whose comments on a first draft of this paper led to the present (tardy) revision.

References

F. Bastin, V. Malmedy, M. Mouffe, Ph. L. Toint, and D. Tomanos. A retrospective trust-region method for unconstrained optimization. *Mathematical Programming, Series A*, **123**(2), 395–418, 2010.

S. Bellavia, C. Cartis, N. I. M. Gould, B. Morini, and Ph. L. Toint. Convergence of a regularized euclidean residual algorithm for nonlinear least-squares. *SIAM Journal on Numerical Analysis*, **48**(1), 1–29, 2010.

J. V. Burke, J. J. Moré, and G. Toraldo. Convergence properties of trust region methods for linear and convex constraints. *Mathematical Programming, Series A*, **47**(3), 305–336, 1990.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results.

[2] When considering Cauchy decrease only

Mathematical Programming, Series A, 2009*a*. DOI: 10.1007/s10107-009-0286-5, 51 pages.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Trust-region and other regularisation of linear least-squares problems. *BIT*, **49**(1), 21–53, 2009*b*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Mathematical Programming, Series A*, 2010*a*. DOI: 10.1007/s10107-009-0337-y.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. *SIAM Journal on Optimization*, **20**(6), 2833–2852, 2010*b*.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. Global convergence of a class of trust region algorithms for optimization with simple bounds. *SIAM Journal on Numerical Analysis*, **25**(182), 433–460, 1988. See also same journal 26:764–767, 1989.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 01 *in* 'MPS-SIAM Series on Optimization'. SIAM, Philadelphia, USA, 2000.

A. R. Conn, N. I. M. Gould, A. Sartenaer, and Ph. L. Toint. Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints. *SIAM Journal on Optimization*, **3**(1), 164–221, 1993.

J. Fan. Convergence rate of the trust region method for nonlinear equations under local error bound condition. *Computational Optimization and Applications*, **2**(34), 215–227, 2006.

J. Fan and J. Pan. An improved trust region algorithm for nonlinear equations. *Computational Optimization and Applications*, **48**(1), 59–70, 2008*a*.

J. Fan and J. Pan. A modified trust region algorithm for nonlinear equations with new updating rule of trust region radius. *International Journal of Computer Mathematics*, **(submitted)**, 2008*b*.

J. Fan and Y. Yuan. A new trust region algorithm with trust region radius converging to zero. *in* D. Li, ed., 'Proceedings of the 5th International Conference on Optimization: Techniques and Applications (ICOTA 2001, Hong Kong)', pp. 786–794, 2001.

S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, **19**(1), 414–444, 2008.

A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom, 1981.

Yu. Nesterov. Modified Gauss-Newton scheme with worst-case guarantees for global performance. *Optimization Methods and Software*, **22**(3), 469–483, 2007.

Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming, Series A*, **108**(1), 177–205, 2006.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Series in Operations Research. Springer Verlag, Heidelberg, Berlin, New York, 1999.

M. J. D. Powell. A new algorithm for unconstrained optimization. *in* J. B. Rosen, O. L. Mangasarian and K. Ritter, eds, 'Nonlinear Programming', pp. 31–65, London, 1970. Academic Press.

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 Journal on Numerical Analysis*, **22**(1), 47–67, 1985.

Ph. L. Toint. Global convergence of a class of trust region methods for nonconvex minimization in Hilbert space. *IMA Journal of Numerical Analysis*, **8**(2), 231–252, 1988.

M. Weiser, P. Deuflhard, and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, **22**(3), 413–431, 2007.

J. Zhang and Y. Wang. A new trust region method for nonlinear equations. *Mathematical Methods of Operations Research*, **58**, 283–298, 2003.