Les Cahiers du GERAD ISSN: 0711–2440

Corrigendum: A proximal quasi-Newton trust-region method for nonsmooth regularized optimization

Supplementary material

A. Aravkin, R. Baraldi, G. Leconte, D. Orban

G–2021–12

March 2021
Revised: March 2024

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2021
– Bibliothèque et Archives Canada, 2021

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2021
– Library and Archives Canada, 2021

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Corrigendum: A proximal quasi-Newton trust-region method for nonsmooth regularized optimization

Supplementary material

Aleksandr Aravkin [a]

Robert Baraldi [b]

Geoffroy Leconte [c]

Dominique Orban [c]

[a] *Department of Applied Mathematics, University of Washington, Seattle WA., USA*

[b] *Optimization and Uncertainty Quantification, Sandia National Laboratories, Albuquerque, NM, USA*

[c] *GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec), Canada*

saravkin@uw.edu
rjbaral@sandia.gov
geoffroy.leconte@polymtl.ca
dominique.orban@gerad.ca

March 2021
Revised: March 2024
Les Cahiers du GERAD
G–2021–12

Abstract : The purpose of the present note is to bring clarifications to certain concepts and surrounding notation of [1]. All results therein continue to hold. The clarifications ensure complete agreement between the paper and our implementations.

Acknowledgements: The research of A. Aravkin is partially supported by the Washington Research Foundation. The work of R. Baraldi was sponsored by the Department of Energy Office of Science, Office of Advanced Scientific Computing Research's John von Neumann Fellowship. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE–NA0003525. The research of D. Orban is supported by an NSERC Discovery grant.

In (3.2), we should have defined $\varphi(s; x) := \varphi_{cp}(s; x) + \frac{1}{2}\nu^{-1}\|s\|^2$ where φ_{cp} is the first order Taylor approximation

$$\varphi_{cp}(s; x) = f(x) + \nabla f(x)^T s, \tag{1}$$

and "cp" stands for "Cauchy point". In (3.4), $s_{k,1}$ could have been defined equivalently as the solution of

$$s_{k,1} \in \underset{s}{\operatorname{argmin}} \; \varphi_{cp}(s; x_k) + \tfrac{1}{2}\nu^{-1}\|s\|^2 + \psi(s; x_k) + \chi(s; \Delta_k). \tag{2}$$

The first step of the proximal-gradient algorithm applied to minimize $\varphi_{cp}(s; x_k) + \psi(s; x_k) + \chi(s; \Delta_k)$ with step size ν is then precisely $s_{k,1}$.

In addition to the definition of ξ in (3.6), we also define

$$\xi_{cp}(\Delta; x_k, \nu) := f(x_k) + h(x_k) - \varphi_{cp}(s_{k,1}; x_k) - \psi(s_{k,1}; x_k). \tag{3}$$

Note that for $\varphi = \varphi_{cp}$,

$$\begin{aligned}\xi(\Delta_k; x_k, \nu_k) &= (f + h)(x_k) - (\varphi_{cp}(s_{k,1}; x_k) + \tfrac{1}{2}\nu_k^{-1}\|s_{k,1}\|^2 + \psi(s_{k,1}; x_k)) \\ &\leqslant (f + h)(x_k) - (\varphi_{cp}(s_{k,1}; x_k) + \psi(s_{k,1}; x_k)) \\ &= \xi_{cp}(\Delta_k; x_k, \nu_k).\end{aligned}$$

Thus $\xi_{cp}(\Delta_k; x_k, \nu_k) = 0 \implies \xi(\Delta_k; x_k, \nu_k) = 0$, and, by Proposition 3.3, x_k is first-order stationary. Using the definition of $s_{k,1}$, we have

$$\varphi_{cp}(s_{k,1}; x_k) + \tfrac{1}{2}\nu_k^{-1}\|s_{k,1}\|^2 + \psi(s_{k,1}; x_k) \leqslant \varphi_{cp}(0; x_k) + \psi(0; x_k).$$

The above inequality, combined with the facts that $\varphi(0; x_k) = f(x_k)$ and $\psi(0; x_k) = h(x_k)$, leads to

$$\xi_{cp}(\Delta_k; x_k, \nu_k) = f(x_k) + h(x_k) - (\varphi_{cp}(s_{k,1}; x_k) + \psi(s_{k,1}; x_k)) \geqslant \tfrac{1}{2}\nu_k^{-1}\|s_{k,1}\|^2. \tag{4}$$

In Step Assumption 3.1, (3.8b) should read

$$m_k(0; x_k) - m_k(s_k; x_k) \geqslant \kappa_{mdc}\xi_{cp}(\Delta_k; x_k, \nu_k). \tag{5}$$

The justification of (3.8b) (below Model Assumption 3.2) is no longer relevant with the above change. However, the following proposition shows that (5) in Step Assumption 3.1 can still be satisfied if

$$\varphi(s; x_k, B_k) = f(x_k) + \nabla f(x_k)^T s + \tfrac{1}{2}s^T B_k s = \varphi_{cp}(s; x_k) + \tfrac{1}{2}s^T B_k s, \tag{6}$$

where $\{B_k\}$ is bounded. In addition, we require the existence of a maximal trust-region radius $\Delta_{max} \geqslant \Delta_0$. This can be enforced in Algorithm 3.1 at the end of Step 11 with the following assignment:

$$\Delta_{k+1} \leftarrow \min(\Delta_{k+1}, \Delta_{max}). \tag{7}$$

Proposition 1. *Let Algorithm* 3.1 *be modified according to* (7)*, and* $\varphi(s; x_k, B_k)$ *defined in* (6) *be used in Step* 5*. Let Model Assumption* 3.2 *be satisfied, and the sequence of Hessian approximations* $\{B_k\}$ *generated by Algorithm* 3.1 *be bounded. Then, there exists* $\kappa_{mdc} \in (0, 1)$ *such that* (5) *holds for all* k*.*

Proof. By definition of s_k, computed in Step 8 of Algorithm 3.1 using (6) in Step 5,

$$\begin{aligned}\varphi(s_k; x_k, B_k) + \psi(s_k; x_k) = m(s_k; x_k, B_k) \leqslant m(s_{k,1}; x_k, B_k) &= \varphi_{cp}(s_{k,1}; x_k) + \tfrac{1}{2}s_{k,1}^T B_k s_{k,1} + \psi(s_{k,1}; x_k) \\ &\leqslant \varphi_{cp}(s_{k,1}; x_k) + \tfrac{1}{2}\|B_k\|\|s_{k,1}\|^2 + \psi(s_{k,1}; x_k).\end{aligned}$$

We multiply the above inequality by -1 and add $m(0; x_k, B_k) = m(0; x_k, \nu_k) = f(x_k) + h(x_k)$ to both sides to obtain

$$m(0; x_k, B_k) - m(s_k; x_k, B_k) \geqslant \xi_{cp}(\Delta_k; x_k, \nu_k) - \tfrac{1}{2}\|B_k\|\|s_{k,1}\|^2.$$

To satisfy (5), it is sufficient to show that there exists $\kappa_{mdc} \in (0, 1)$ such that

$$\xi_{cp}(\Delta_k; x_k, \nu_k) - \tfrac{1}{2}\|B_k\|\|s_{k,1}\|^2 \geqslant \kappa_{mdc}\xi_{cp}(\Delta_k; x_k, \nu_k),$$

i.e.,

$$(1 - \kappa_{mdc})\xi_{cp}(\Delta_k; x_k, \nu_k) \geqslant \tfrac{1}{2}\|B_k\|\|s_{k,1}\|^2.$$

Using (4), it is also sufficient to show that there exists $\kappa_{mdc} \in (0, 1)$ such that

$$(1 - \kappa_{mdc})\nu_k^{-1} \geqslant \|B_k\|. \tag{8}$$

If $\|B_k\| = 0$, (8) is satisfied for any $\kappa_{mdc} \in (0, 1)$. Otherwise, because (6) indicates that the Lipschitz constant $L(x_k)$ of $\nabla\varphi(\cdot; x_k, B_k)$ in Model Assumption 3.2 is $L(x_k) = \|B_k\|$, the definition of ν_k in Step 4 of Algorithm 3.1 yields

$$\|B_k\|\nu_k = \frac{\|B_k\|}{\|B_k\| + \alpha^{-1}\Delta_k^{-1}} \leqslant \frac{1}{1 + \alpha^{-1}\Delta_k^{-1}\|B_k\|^{-1}} \leqslant \frac{1}{\alpha^{-1}\Delta_{max}^{-1}(\sup_{k\in\mathbb{N}}\|B_k\|)^{-1} + 1} \in (0, 1), \tag{9}$$

because $\{B_k\}$ is bounded by assumption. We deduce from (9) that (8), and therefore (5) hold. □

In the proof of Theorem 3.4, (4) guarantees that the inequality

$$\xi_{cp}(\Delta_k; x_k, \nu_k) \geqslant \tfrac{1}{2}\nu_k^{-1}\|s_{k,1}\|^2 \geqslant \tfrac{1}{2}(L(x_k) + \alpha^{-1}\Delta_k^{-1})\|s_{k,1}\|^2 \geqslant \tfrac{1}{2}\alpha^{-1}\Delta_k^{-1}\|s_{k,1}\|^2, \tag{10}$$

continues to hold, where $L(x_k)$ is the Lipschitz constant of the gradient of $\varphi(\cdot; x_k)$. Thus, the theorem remains valid.

In the simple case where $h = 0$ and $s_{k,1} = -\nu_k\nabla f(x_k)$ is inside the trust-region,

$$\xi_{cp}(\Delta_k; x_k, \nu_k) = -\nabla f(x_k)^T s_{k,1} = \nu_k\|\nabla f(x_k)\|^2, \tag{11}$$

which suggests that the appropriate criticality measure (analogous to the gradient norm in the smooth case) should be $\nu_k^{-1/2}\xi_{cp}(\Delta_k; x_k; \nu_k)^{1/2}$ instead of $\nu_k^{-1}\xi_{cp}(\Delta_k; x_k, \nu_k)^{1/2}$. The first-order optimality condition in (3.13) then becomes

$$\nu_k^{-1/2}\xi_{cp}(\Delta_{min}; x_k, \nu_k)^{1/2} \leqslant \epsilon \quad (0 < \epsilon < 1). \tag{12}$$

This changes the bound in (3.15) to

$$|S(\epsilon)| \leqslant \frac{(f + h)(x_0) - (f + h)_{low}}{\eta_1\kappa_{mdc}\nu_{min}\epsilon^2} = O(\epsilon^{-2}), \tag{13}$$

i.e., ν_{min}^2 becomes ν_{min}, but does not alter the order of complexity.

Similarly, the limits considered in Corollary 3.9 and Theorem 3.11 should be $\liminf_{k\to\infty} \nu_k^{-1/2}\xi_{cp}(\Delta_k; x_k, \nu_k)^{1/2}$ and $\lim_{k\to\infty} \nu_k^{-1/2}\xi_{cp}(\Delta_k; x_k, \nu_k)^{1/2}$, respectively.

In addition to the definition of $\xi(\sigma; x)$ in (6.4), we should have also defined

$$\xi_{cp}(\sigma; x) := f(x) + h(x) - (\varphi(s; x) + \psi(s; x)), \tag{14}$$

where

$$s \in \arg\min_s m(s; x, \sigma). \tag{15}$$

However, (6.6) is still valid.

After (6.10), by analogy with Algorithm 3.1, the stationarity measure should be $\sigma_k^{1/2}\xi_{cp}(\sigma_{max}; x_k)^{1/2}$.

In Section 7, all our numerical experiments use $\xi_{cp}(\Delta_k; x_k, \nu_k)^{1/2}$ and $\xi(\sigma_k; x_k)^{1/2}$ as proxies for the stationarity measures. Even though the factors $\nu_k^{-1/2}$ and $\sigma_k^{1/2}$, respectively, are not included in the proxies, the measures remain valid as long as $\nu_{min} > 0$ and $\sigma_{max} < \infty$, which is the case under our assumptions.

References

[1] A. Y. Aravkin, R. Baraldi, and D. Orban. A proximal quasi-Newton trust-region method for nonsmooth regularized optimization. SIAM J. Optim., 32(2):900–929, 2022.