Report: Finite Precision Quadratic Regularization Algorithm

D. Monnet, F. Rahbarnia

February 8, 2022

1 Problem statement and background

We consider the problem

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

with f a continuously differentiable function. We suppose that the computations are run in finite precision. As a consequence, it is not possible to evaluate f nor the gradient $\nabla_x f$ exactly, but only their finite precision couterparts denoted $\hat{f}$ and g. We further suppose that models for the errors ω_f and ω_g are provided as functions of x,

$$||f(x) - \hat{f}(x)|| \leq \omega_f(x), \quad ||\nabla_x f(x) - g(x)|| \leq \omega_g(x)||g(x)||. \tag{1}$$

The usual assuptions are made.

AS.1 There exsits κ_{low} such that,

$$\forall x \in \mathbb{R}^n,\ f(x) \geq \kappa_{low}. \tag{AS.1}$$

AS.2 f is continuously differentiable over $\mathbb{R}^n$.

AS.3 The gradient of f is Lipschitz continuous.

$$\exists L > 0,\ \forall x, y \in \mathbb{R}^n,\ ||\nabla_x f(x) - \nabla_x f(y)|| \leq L||x - y||. \tag{AS.3}$$

The regularized quadratic method is based on a order one model of the objective function. Let T be the Taylor expansion of f truncated at order 1,

$$T(x, s) = f(x) + \nabla_x f(x)^T s. \tag{2}$$

Since $\nabla_x f(x)$ is not available, it is not possible to use directly T. Instead, we can use the approximated Taylor serie defined with the inexact computed gradient g,

$$\bar{T}(x, s) = f(x) + g(x)^T s. \tag{3}$$

The regularized approximated Talor serie is used as the local model in the algorihtm, and defined as,

$$m(x_k, s) = \bar{T}(x_k, s) + \frac{1}{2}\sigma_k||s||^2, \tag{4}$$

with σ_k a parameters that control the step size at iteration k. Indeed, the step is chosen as the minimizer of the local model given by,

$$s_k = \underset{s \in \mathbb{R}^n}{\text{argmin}}\ m(x_k, s) = -\frac{g(x_k)}{\sigma_k}. \tag{5}$$

2 Inexact descent direction and pseudo-model

Add some material on finite precision computation somewhere, introduce δ, θ, γ and u. Cite [2].

Add prelimirary paragraph explaining that the usual proof of convergence relies on equalities/inequalities that do not hold when considering finite precision computations: $fl\left(\bar{T}(x_k, 0) - \bar{T}(x_k, s)\right) \neq -g^T s$. Try to introduce the idea of the pseudo model as a surrogate that helps prooving the convergence.

AS.5 Finite precision computations comply with IEEE 754 norm, and underflow and overflow do not occur during algorithm execution.

When computing s_k with finite precision arithmetic, it happens that the descent direction is not g but $\tilde{g}$ that includes rounding error. Denoting $\hat{s}_k$ the finite precision, one has,

$$\hat{s}_k = fl\left(-\frac{g_k}{\sigma_k}\right) = -\frac{\tilde{g}_k}{\sigma_k} \tag{6}$$

with $\tilde{g}_k = g_k.[(1+\delta_1), \ldots, (1+\delta_n)]^T$, where . denotes the element-wise multiplication operator. The difference between the norms of g_k and $\tilde{g}_k$ is bounded as,

$$||g_k - \tilde{g}_k|| \leq u||g_k||. \tag{7}$$

We define the *pseudo-Taylor serie* as,

$$\tilde{T}(x_k, s) = \hat{f}(x_k) + \tilde{g}^T s, \tag{8}$$

with. The pseudo-Taylor serie can be viewed as approximated Talor serie $\bar{T}$ but slightly modified, such that the rounding error makes sense for it. Indeed, $\tilde{T}(x_k, s)$ is expressed with $\tilde{g}_k$ which is the actual descent direction. As such, considering the pseudo-Taylor serie $\tilde{T}$ instead of the $\bar{T}$ makes it easier to prove the convergence of the algorithm when considering rounding errors due to finite precision computations (see Section 4).

The main difficulty is that $\tilde{g}$ is unknown, and as a consequence $\tilde{T}$ is also unknown. This is an issue for relying the pseudo-Taylor serie since the algorithm requires the decrease to be computed. However, computing the decrease of the approximated Taylor serie $\bar{T}$ with finite precision arithmetic provide the decrease of the pseudo-Taylor serie with a relative error. This is what is stated by Lemma 1.

Lemma 1. *Let $\hat{\Delta T}_k$ be the finite precision computation of $\bar{T}(x_k, 0) - \bar{T}(x_k, \hat{s}_k)$. One has,*

$$\hat{\Delta T}_k = \left(\tilde{T}(x, 0) - \tilde{T}(x, \hat{s}_k)\right)(1 + \theta_{n+2}). \tag{9}$$

Proof. Not very rigorous, rethink how to properly write down finite precision computations analysis and name g_k elements.

$$\begin{aligned}
\hat{\Delta T}_k &= fl\left(-g_k \hat{s}_k\right) \\
&= -fl\left(g_k \frac{\tilde{g}}{\sigma_k}\right) \\
&= -fl\left(\sum_{i=1}^{n} \frac{\tilde{g}_{k,i}^2 (1+\delta)}{\sigma_k (1+\delta_i)}\right) \\
&= -\textstyle\sum_{i=1}^{n} \frac{\tilde{g}_{k,i}^2 (1+\delta)}{\sigma_k (1+\delta_i)} (1+\theta_n) \\
&= -\frac{\tilde{g}_{k,i}^2}{\sigma_k} (1+\theta_{n+2}) \\
&= -\tilde{g}\hat{s}_k (1+\theta_{n+2})
\end{aligned} \tag{10}$$

□

3 Algorithm

The algorithm details in this section implements strategies to deal with inexact evaluation of f, its gradient and the approximated Taylor serie. In order to make the equations easier to read, we define the constant $\alpha = \frac{1}{1-\gamma_{n+2}}$. We suppose that the machine precision is such that $1 - \gamma_{n+2} > 0$.

AS.4 There exists κ_g such that $\forall x$,

$$\frac{\omega_g(x) + u}{1-u} \leq \kappa_g \tag{AS.4}$$

Algorithm 1 Multi-precision trust region algorithm

Step 0: Initialization: Initial point x_0, initial value σ_0, final gradient accuracy ϵ, constant values η_0, η_1, η_2, γ_1, γ_2, γ_3 such that,

$$0 < \eta_1 \leq \eta_2 < 1 \quad 0 < \gamma_1 < 1 < \gamma_2 \leq \gamma_3, \quad \eta_0 \leq \frac{1}{2}\eta_1 \quad \eta_0 + \frac{\alpha\kappa_g}{2} \leq \frac{1}{2}(1-\eta_2) \tag{11}$$

Set $k = 0$, compute $f_0 = \hat{f}(x_0)$.

Step 1: Check for termination: If $k = 0$ or $x_k \neq x_{k-1}$, compute $g_k = g(x_k)$. Terminate if

$$||g_k|| \leq \frac{\epsilon}{1+\kappa_g} \tag{12}$$

With condition in [1], terminate if $\dfrac{\omega_g(x_k)+u}{1-u} > 1/\sigma_k$.

Step 2: Step calculation: Compute $\hat{s}_k = fl\,(g_k/\sigma_k)$.
Compute approximated taylor serie decrease $\hat{\Delta T} = fl\left(g_k^T \hat{s}_k\right)$.

Step 3: Evaluate the objective function: Compute $\hat{f}_k^+ = \hat{f}(x_k + s_k)$.
If $\omega_f(x_k) > \eta_0 \hat{\Delta T}$ or $\omega_f(x_k + \hat{s}_k) > \eta_0 \hat{\Delta T}$, return x_k.

Step 4: Acceptance of the trial point: Define the ratio

$$\rho_k = \frac{f_k - f_k^+}{\hat{\Delta T}} \tag{13}$$

If $\rho_k \geq \eta_1$, then $x_{k+1} = x_k + \hat{s}_k$, $f_{k+1} = f_k^+$. Otherwise set $x_{k+1} = x_k$, $f_{k+1} = f_k$.

Step 5: Regularization parameter update:

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{min}, \gamma_1\sigma_k), \sigma_k] & \text{if } \rho_k \geq \eta_2 \\ [\sigma_k, \gamma_2\sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2) \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \text{if } \rho_k \in < \eta_1 \end{cases} \tag{14}$$

$k = k+1$, go to Step 1.

4 Proof of convergence

Lemma 2.

$$|f(x_k+s) - \tilde{T}(x_k,s)| \le \frac{1}{2}L||s||^2 + (\omega_g(x_k)+u)||g_k||\,||s||. \tag{15}$$

Proof. First, recall that [1]

$$|f(x_k+s) - T(x_k,s)| \le \frac{1}{2}L||s||^2. \tag{16}$$

With Equation (16) and (7), it follows that,

$$\begin{aligned}
|f(x_k+s) - \tilde{T}(x_k,s)| &\le |f(x_k+s) - T(x_k,s)| + |T(x_k,s) - \tilde{T}(x_k,s)| \\
&\le \frac{1}{2}L||s||^2 + ||\nabla_x f(x_k+s) - \tilde{g}_k||\,||s|| \\
&\le \frac{1}{2}L||s||^2 + (\omega_g(x_k+s)+u)||g_k||\,||s||
\end{aligned} \tag{17}$$

□

Lemma 3. *Correspond to condition* $\omega_g(x_k) < 1/\sigma_k$*.*

$$\sigma_k \ge \frac{\frac{1}{2}L+1}{\alpha(1-\eta_2-2\eta_0)} \implies \rho_k \ge \eta_2 \tag{18}$$

Proof.

$$\begin{aligned}
|\rho - 1| &\le \frac{|f(x_k) - \hat{f}(x_k)|}{\Delta \hat{T}_k} + \frac{|f(x_k+s_k) - \hat{f}(x_k+s_k)|}{\Delta \hat{T}_k} + |\frac{f(x_k+s_k) - f(x_k)}{\Delta \hat{T}_k} - 1| \\
&\le 2\eta_0 + \alpha \frac{f(x_k+\hat{s}_k) - \tilde{T}(x_k,\hat{s}_k)}{-\tilde{g}_k^T \hat{s}_k} \\
&= 2\eta_0 + \alpha \frac{\frac{1}{2}L||\hat{s}_k||^2 + (\omega_g(x_k)+u)||g_k||\,||\hat{s}_k||}{||\tilde{g}_k||^2/\sigma_k} \\
&\le 2\eta_0 + \alpha\left[\frac{1}{2}\frac{L}{\sigma_k} + \frac{\omega_g(x_k)+u}{1-u}\right]
\end{aligned} \tag{19}$$

Since Step 1 enforces $\frac{\omega_g(x_k)+u}{1-u} \le 1/\sigma_k$, it follows that

$$|\rho - 1| \le 2\eta_0 + \alpha \frac{\frac{1}{2}L+1}{\sigma_k}. \tag{20}$$

Therefore, having $\sigma_k \ge \alpha \frac{\frac{1}{2}L+1}{1-\eta_2-2\eta_0}$ implies that $|\rho - 1| \le 1-\eta_2$. □

Lemma 4. *Correspond to new strategy involving* κ_g

$$\frac{1}{\sigma_k} \le \left[1 - \eta_2 - \eta_0 - \frac{\alpha\kappa_g}{2}\right]\frac{1}{L\alpha} \implies \rho_k \ge \eta_2, \tag{21}$$

with $\alpha = \dfrac{1}{1-\gamma_{n+2}}$.

Proof. As for the proof of Lemma 3, one has,

$$\begin{aligned} |\rho - 1| &\leq 2\eta_0 + \alpha \left[\tfrac{1}{2} \frac{L}{\sigma_k} + \frac{\omega_g(x_k) + u}{1 - u} \right] \\ &= \end{aligned} \tag{22}$$

With Assumption (AS.4), one obtains,

$$|\rho - 1| \leq 2\eta_0 + \alpha\kappa_g + \frac{\alpha L}{2} \frac{1}{\sigma_k}. \tag{23}$$

Therefore, whith $1/\sigma_k \leq [1 - \eta_2 - 2\eta_0 - \alpha\kappa_g] \dfrac{1}{L\alpha}$, the inequality rewrites

$$|\rho - 1| \quad \leq \tfrac{1}{2}(1 - \eta_2) + \eta_0 + \frac{\alpha\kappa_g}{2} \tag{24}$$

and since the parameters are chosen such that $\eta_0 + \dfrac{\alpha\kappa_g}{2} \leq \tfrac{1}{2}(1 - \eta_2)$, it follows that $|\rho - 1| \leq 1 - \eta_2$. □

Lemma 5. *For all $k > 0$,*

$$1/\sigma_k \geq 1/\sigma_{max} = \gamma_3 \left[1 - \eta_2 - \eta_0 - \frac{\alpha\kappa_g}{2} \right] \frac{1}{L\alpha} \tag{25}$$

Proof. Straightforward from Lemma 4 □

Let $S_k = \{0 \leq j \leq k \,|\, \rho_j \geq \eta_1\}$ be the set of successful iterations and $U_k = \{0 \leq j \leq k \,|\, \rho_j < \eta_1\}$ be the set of unsuccessful iterations.

Lemma 6. *For all $k > 0$,*

$$|S_k| \left(1 + \frac{|\log(\gamma_1)|}{\log(\gamma_2)} \right) + \frac{1}{\log(\gamma_2)} \log \left(\frac{\sigma_{max}}{\sigma_0} \right), \tag{26}$$

Proof. From the update formula of σ_k (14), one has for each $k > 0$,

$$\forall j \in S_k,\ \gamma_1\sigma_j \leq \max\left[\gamma_1\sigma_j, \sigma_{min}\right] \text{ and } \forall i \in U_k,\ \gamma_2\sigma_i \leq \gamma_{i+1}.$$

It follows that,

$$\sigma_0 \gamma_1^{|S_k|} \gamma_2^{|U_k|} \leq \sigma_k. \tag{27}$$

With Lemma 5 one obtains,

$$|S_k| \log \gamma_1 + |U_k| \log \gamma_2 \leq \log \left(\frac{\sigma_{max}}{\sigma_0} \right). \tag{28}$$

Since $\gamma_2 > 1$, it follows that,

$$|U_k| \leq -|S_k| \frac{\log(\gamma_1)}{\log(\gamma_2)} + \frac{1}{\log(\gamma_2)} \log \left(\frac{\sigma_{max}}{\sigma_0} \right). \tag{29}$$

Since $k = |S_k| + |U_k|$, the statement of Lemma 6 holds true. □

Theorem 1. *Algorithm 1 needs at most*

$$\epsilon^{-2}\kappa_s(f(x_0)-\kappa_{low}), \quad \kappa_s = \alpha\sigma_{max}(1+\kappa_g)^2(\eta_1-2\eta_0)$$

successful iteration and at most

$$\epsilon^{-2}\kappa_s(f(x_0)-\kappa_{low})\left(1+\frac{|\log(\gamma_1)|}{\log(\gamma_2)}\right)+\frac{1}{\log(\gamma_2)}\log\left(\frac{\sigma_{max}}{\sigma_0}\right)$$

iteration to provide an iterate x_k such that $\nabla_x f(x_k) \leq \epsilon$.

Proof.

$$\begin{aligned} f(x_0)-\kappa_{low} \quad &\geq (\eta_1-2\eta_0)\sum_{j\in S_k}\hat{\Delta T}_j \\ &\geq (\eta_1-2\eta_0)(1-\gamma_{n+2})\sum_{j\in S_k}\tilde{T}(x_j,0)-\tilde{T}(x_j,\hat{s}_j) \\ &\geq (\eta_1-2\eta_0)(1-\gamma_{n+2})\sum_{j\in S_k}||\tilde{g}_j||^2/\sigma_j \\ &\geq (\eta_1-2\eta_0)(1-\gamma_{n+2})|S_k|\frac{\epsilon^2}{(1+\kappa_g)^2\sigma_{max}} \end{aligned} \tag{30}$$

This proves the first statement of Theorem 1. Using the above inequality with Lemma 6 proves the second statement. □

References

[1] Ernesto G Birgin, JL Gardenghi, José Mario Martínez, Sandra Augusta Santos, and Ph L Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1):359–368, 2017.

[2] Nicholas J Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.