EURO J Comput Optim (2017) 5:501–527
DOI 10.1007/s13675-017-0081-7

ORIGINAL PAPER

On the construction of quadratic models for derivative-free trust-region algorithms

Adriano Verdério[1] · **Elizabeth W. Karas**[2] · **Lucas G. Pedroso**[2] · **Katya Scheinberg**[3]

Received: 30 March 2016 / Accepted: 12 January 2017 / Published online: 9 February 2017

Abstract We consider derivative-free trust-region algorithms based on sampling approaches for convex constrained problems and discuss two conditions on the quadratic models for ensuring their global convergence. The first condition requires the poisedness of the sample sets, as usual in this context, while the other one is related to the error between the model and the objective function at the sample points. Although the second condition trivially holds if the model is constructed by polynomial interpolation, since in this case the model coincides with the objective function at the sample set, we show that it also holds for models constructed by support vector regression. These two conditions imply that the error between the gradient of the trust-region model and the objective function is of the order of δ_k, where δ_k controls the diameter of the sample set. This allows proving the global convergence of a trust-region algorithm that uses two radii, δ_k and the trust-region radius. Preliminary numerical experiments are presented for minimizing functions with and without noise.

Keywords Derivative-free optimization · Trust region · Polynomial interpolation · Support vector regression

This work was partially supported by Capes-Brazil Grant PDSE 11348/12-7, NSF Grant DMS 13-19356, AFOSR Grant FA9550-11-1-0239, and CNPq-Brazil Grants 477611/2013-3 and 308957/2014-8.

✉ Adriano Verdério
verderio@utfpr.edu.br

[1] Department of Mathematics, Federal University of Technology - Paraná, 3165 Sete de Setembro Avenue, Curitiba, PR 80230-901, Brazil

[2] Department of Mathematics, Federal University of Paraná, Centro Politécnico, Jardim das Américas CP 19081, Curitiba, PR 81531-980, Brazil

[3] Department of Industrial and Systems Engineering, Lehigh University, Harold S. Mohler Laboratory 200 West Packer Avenue, Bethlehem, PA 18015-1582, USA

Mathematics Subject Classification 90C56 · 65K05 · 49M37

1 Introduction

We consider the nonlinear programming problem

$$\begin{array}{l} \text{minimize} \quad f(x) \\ \text{subject to } x \in \Omega, \end{array} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function and $\Omega \subset \mathbb{R}^n$ is a non-empty closed convex set. Although the objective function is smooth, we assume its derivatives are not available.

We are interested in derivative-free trust-region methods, which have been applied to unconstrained problems by Conn et al. (1997, 2009a, b, c), Ferreira et al. (2015), Powell (2006, 2012), and Scheinberg and Toint (2010), to box-constrained problems by Powell (2009) and Gratton et al. (2011), to linearly constrained problems by Gumma et al. (2014) and Powell (2015), to convex constrained problems by Conejo et al. (2013), to general constrained problems by Bueno et al. (2013), Conejo et al. (2015), Conn et al. (1998), and Ferreira et al. (2016), and to composite non-smooth optimization by Grapiglia et al. (2016) and Garmanjani et al. (2016).

Derivative-free trust-region algorithms make use of quadratic models that approximate the objective function relying only on zero-order information. In this work, we discuss conditions on the models to ensure global convergence of an algorithm. These conditions are relaxed versions of classical hypotheses (Conn et al. 2009c) and will allow us to move beyond the polynomial interpolation models that are largely used in derivative-free frameworks (Conn et al. 2008a, b, 2009c; Powell 2006). The support vector regression (Drucker et al. 1996; Smola and Schölkopf 2004), which is a well-known machine learning technique, is a motivating framework for constructing the models which may have a superior performance on noisy problems. We state that not all practical and theoretical aspects of support vector regression have yet been explored here, but we show that this technique can be useful and theoretically justified.

We consider the following assumptions on the objective function.

A 1 *The function f is continuously differentiable, and ∇f is Lipschitz continuous with constant $L_g > 0$ in a sufficiently large open bounded domain $\mathcal{X} \subset \mathbb{R}^n$.*

A 2 *The function f is bounded below in Ω.*

1.1 Structure of the paper

In Sect. 2, we propose a derivative-free trust-region algorithm which extends the ideas of Conejo et al. (2013) considering two radii, one that defines the trust region and another that controls the diameter of the sample set used to construct the trust-region models. The global convergence is proven under classical hypotheses on the problem and the assumption that the error between the gradient of the trust-region model and

the objective function is of the order of δ_k, where δ_k controls the diameter of the sample set. Section 3 shows that this assumption is implied by two conditions on the sample set, one on its geometry and another on the error between the trust-region model and the objective function at the points of this set. In Sect. 4, we prove that the first condition is fulfilled if the sample set is Λ-poised (Conn et al. 2009c) for some $\Lambda > 0$. Section 5 shows that models constructed by quadratic interpolation or support vector regression satisfy the second condition. Numerical experiments are presented in Sect. 6.

1.2 Notation

Throughout the paper, the symbol $\|\cdot\|$ denotes the Euclidean norm and $B(x^k, r) = \{x \in \mathbb{R}^n \mid \|x - x^k\| \leq r\}$.

2 Trust-region algorithm

In this section, we present a derivative-free trust-region algorithm based on Conejo et al. (2013) and we prove its global convergence under reasonable assumptions.

At each iteration $k \in \mathbb{N}$, we consider the current iterate $x^k \in \Omega$ and the quadratic model

$$m_k(x) = b_k + g_k^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top H_k (x - x^k), \tag{2}$$

where $b_k \in \mathbb{R}$, $g_k = \nabla m_k(x^k) \in \mathbb{R}^n$ and $H_k \in \mathbb{R}^{n\times n}$ is a symmetric matrix.

We define the stationarity measure at x^k for the problem of minimizing the model m_k over the set Ω by

$$\pi_k = \|P_\Omega(x^k - g_k) - x^k\|,$$

where P_Ω denotes the orthogonal projection onto Ω. Note that $x^* \in \Omega$ is stationary for the original problem (1) when

$$\|P_\Omega(x^* - \nabla f(x^*)) - x^*\| = 0.$$

Given the radius $\Delta_k > 0$, the trust-region computation solves approximately the subproblem

$$\begin{aligned} &\text{minimize} && m_k(x^k + d) \\ &\text{subject to} && x^k + d \in \Omega \\ &&& \|d\| \leq \Delta_k. \end{aligned} \tag{3}$$

We accept as an approximate solution of (3) any feasible direction d^k satisfying the efficiency condition

$$m_k(x^k) - m_k(x^k + d^k) \geq \theta_1 \pi_k \min\left\{\frac{\pi_k}{1 + \|H_k\|}, \Delta_k, 1\right\}, \tag{4}$$

where $\theta_1 > 0$ is a constant independent of k. Similar conditions are common in trust-region approaches and used by several authors in different contexts as by Conn et al. (2000, 2009c), Nocedal and Wright (2006), and Gratton et al. (2011).

The point $x^k + d^k$ is accepted as the new iterate when the ratio between the actual and the predicted reductions

$$\rho_k = \frac{f(x^k) - f(x^k + d^k)}{m_k(x^k) - m_k(x^k + d^k)} \tag{5}$$

is greater than or equal to a fixed constant $\eta > 0$. In this case, we define $x^{k+1} = x^k + d^k$ and repeat the process. Otherwise, the step d^k is rejected and the radius Δ_k is reduced.

The model is updated at the beginning of each iteration, and its quality is controlled by a second radius $\delta_k > 0$. We assume that the following assumption holds.

A3 *There exists a constant $\kappa_m > 0$ such that, for all $k \in \mathbb{N}$,*

$$\|\nabla f(x) - \nabla m_k(x)\| \leq \kappa_m \delta_k$$

for all $x \in \mathcal{X} \cap B(x^k, \delta_k)$.

Incorporating the radius δ_k is the main difference between our algorithm and the one proposed in Conejo et al. (2013), where it is proven that the radius Δ_k goes to zero. The motivation for such modification is the fact that, from a theoretical point of view, it is necessary that the term that controls the model quality converges to zero, but in practice it is desirable that the trust-region radius is as large as possible. The usage of two radii is not new. It was considered, for example, by Powell (2006, 2009), but without convergence analysis.

We present now the derivative-free trust-region algorithm.

Algorithm 1 *Derivative-free trust-region algorithm*

Data: $x^0 \in \Omega$, $\beta > 0$, $\delta_0 = \Delta_0 > 0$, $0 < \tau_1 < 1 \leq \tau_2$, $\eta_1 \in (0, 1)$, $0 \leq \eta < \eta_1 \leq \eta_2$.
Set $k = 0$.
REPEAT
  Construct the model m_k.
  IF $\delta_k > \beta\pi_k$, THEN
    $\delta_{k+1} = \tau_1\delta_k$, $d^k = 0$ *and* $x^{k+1} = x^k$.
    Choose $\Delta_{k+1} \in [\delta_{k+1}, \Delta_k]$.
  ELSE
    Find an approximate solution d^k *of* (3) *satisfying* (4).
    Compute ρ_k *by* (5).
    IF $\rho_k \geq \eta$, THEN
      $x^{k+1} = x^k + d^k$.
    ELSE
      $x^{k+1} = x^k$.
    IF $\rho_k < \eta_1$, THEN
      $\delta_{k+1} = \tau_1\delta_k$ *and* $\Delta_{k+1} = \tau_1\Delta_k$.
    ELSE
      IF $\rho_k > \eta_2$ *and* $\|d^k\| = \Delta_k$, THEN
        $\delta_{k+1} = \tau_2\delta_k$ *and* $\Delta_{k+1} = \tau_2\Delta_k$.
      ELSE
        $\delta_{k+1} = \delta_k$ *and* $\Delta_{k+1} = \Delta_k$.
  $k = k + 1$.

When π_k is small, the iterate is probably close to a solution of the problem of minimizing the model m_k within the feasible set Ω. On the other hand, if δ_k is large, we cannot guarantee that the model represents the objective function accurately enough. Hence, when $\delta_k > \beta\pi_k$, the radius δ_k is reduced in the attempt of finding a more accurate model. Although we could always set $\beta = 1$, this parameter might be used to balance the magnitude of π_k and δ_k, according to the problem.

2.1 Convergence analysis

We start the convergence analysis proving that the Hessian of the model is bounded.

Lemma 1 *Suppose that Assumptions A1 and A3 hold. Then there exists* $\kappa_H > 1$ *such that, for all* $k \in \mathbb{N}$,

$$\|H_k\| \leq \kappa_H - 1.$$

Proof Consider an arbitrary direction $d \in \mathbb{R}^n$ with $\|d\| = \delta_k$. By the definition of the model (2), the triangle inequality and the hypotheses we have that

$$\begin{aligned}\|H_k d\| &= \|\nabla m_k(x^k + d) - \nabla m_k(x^k)\| \\ &\leq \|\nabla m_k(x^k + d) - \nabla f(x^k + d)\| + \|\nabla f(x^k + d) - \nabla f(x^k)\| + \|\nabla f(x^k) \\ &\quad -\nabla m_k(x^k)\| \\ &\leq (2\kappa_m + L_g)\delta_k.\end{aligned}$$

Thus,

$$\|H_k\| = \max_{\|d\|=\delta_k} \left\| H_k \frac{d}{\|d\|} \right\| = \frac{1}{\delta_k} \max_{\|d\|=\delta_k} \|H_k d\| \leq 2\kappa_m + L_g.$$

Defining $\kappa_H = 2\kappa_m + L_g + 1$, we complete the proof.

As a consequence of the previous lemma and Assumption A3, the Hypotheses H3 and H4 of Conejo et al. (2013) hold. Although the convergence analysis of Algorithm 1 is similar to the one presented in Conejo et al. (2013), the inclusion of the radius $\delta_k \leq \Delta_k$ demands some modifications on the proofs that justify their presentation here.

Consider the following sets of indices

$$\mathcal{S} = \{k \in \mathbb{N} \mid \rho_k \geq \eta\} \qquad \text{and} \qquad \overline{\mathcal{S}} = \{k \in \mathbb{N} \mid \rho_k \geq \eta_1\}.$$

The set $\mathcal{S}$ is the set of *successful* iterations and $\overline{\mathcal{S}} \subset \mathcal{S}$.

The following lemma states that if the trust-region radius is small enough, then the iteration is successful.

Lemma 2 (Conejo et al. 2013, Lemma 3.1) *Suppose that Assumptions A1 and A3 hold. Let L_g, κ_m, κ_H, and θ_1 be the constants given in A1, A3, Lemma 1, and (4), respectively. Consider the set*

$$\mathcal{K} = \left\{ k \in \mathbb{N} \mid \Delta_k \leq \min\left\{ \frac{\pi_k}{\kappa_H}, \frac{(1-\eta_1)\pi_k}{c}, \beta\pi_k, 1 \right\} \right\},$$

where $c = \dfrac{L_g + \kappa_m + \dfrac{\kappa_H}{2}}{\theta_1}$. *If $k \in \mathcal{K}$, then $k \in \overline{\mathcal{S}}$.*

From Assumption A3, we can see that the smaller δ_k, the better the model locally represents the objective function. Thus, it is reasonable to expect that δ_k goes to zero. This is proven in the following lemma.

Lemma 3 *Suppose that Assumptions A1–A3 hold. Then the sequence $\{\delta_k\}$ converges to zero.*

Proof If $\overline{\mathcal{S}}$ is finite, from the mechanism of the algorithm, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, $\delta_{k+1} = \tau_1 \delta_k$. Thus, the sequence $\{\delta_k\}$ goes to zero. Henceforth, consider that $\overline{\mathcal{S}}$ is infinite. For any $k \in \overline{\mathcal{S}}$, we have that ρ_k is computed and consequently, $\delta_k \leq \beta \pi_k$. Using this, the definition of ρ_k, (4), Lemma 1 and the fact that $\delta_k \leq \Delta_k$,

$$f(x^k) - f(x^{k+1}) \geq \eta_1 \theta_1 \pi_k \min\left\{\frac{\pi_k}{\kappa_H}, \Delta_k, 1\right\} \geq \eta \theta_1 \frac{\delta_k}{\beta} \min\left\{\frac{\delta_k}{\beta \kappa_H}, \delta_k, 1\right\}.$$

Once $\{f(x^k)\}$ is non-increasing by the mechanism of the algorithm and bounded below by Assumption A2, the left-hand side of the expression above goes to zero. Then,

$$\lim_{k \in \overline{\mathcal{S}}} \delta_k = 0. \tag{6}$$

Consider the set

$$\mathcal{U} = \left\{k \in \mathbb{N} \mid k \notin \overline{\mathcal{S}}\right\}.$$

If $\mathcal{U}$ is finite, by (6) we have that $\lim_{k \to \infty} \delta_k = 0$. Now suppose that $\mathcal{U}$ is infinite. For $k \in \mathcal{U}$, define ℓ_k as the last index in $\overline{\mathcal{S}}$ before k. By the algorithm, $\delta_k \leq \tau_2 \delta_{\ell_k}$, which implies that

$$\lim_{k \in \mathcal{U}} \delta_k \leq \tau_2 \lim_{k \in \mathcal{U}} \delta_{\ell_k} = \tau_2 \lim_{\ell_k \in \overline{\mathcal{S}}} \delta_{\ell_k}.$$

By (6), it follows that $\lim_{k \in \mathcal{U}} \delta_k = 0$, concluding the proof.

The next lemma shows that $\{\pi_k\}$ has a subsequence that converges to zero.

Lemma 4 *Suppose that Assumptions A1–A3 hold. Then* $\liminf_{k \to \infty} \pi_k = 0$.

Proof Suppose by contradiction that there exist $\varepsilon > 0$ and an integer $K > 0$ such that $\pi_k \geq \varepsilon$ for all $k \geq K$. Let η_1, β, κ_H, and c be the constants given in the algorithm and Lemmas 1 and 2. Take

$$\tilde{\Delta} = \min\left\{\frac{\varepsilon}{\kappa_H}, \frac{(1-\eta_1)\varepsilon}{c}, \beta\varepsilon, 1\right\}.$$

Consider $k \geq K$. If $\Delta_k \leq \tilde{\Delta}$, by Lemma 2, $k \in \overline{\mathcal{S}}$ and thus $\Delta_{k+1} \geq \Delta_k$. It follows that the radius Δ_k can only decrease if $\Delta_k > \tilde{\Delta}$. In this case, if $\delta_k > \beta \pi_k$, we have that

$$\Delta_{k+1} \geq \delta_{k+1} = \tau_1 \delta_k > \tau_1 \beta \pi_k \geq \tau_1 \beta \varepsilon \geq \tau_1 \tilde{\Delta}.$$

On the other hand, if $\delta_k \leq \beta \pi_k$, by the algorithm

$$\Delta_{k+1} \geq \tau_1 \Delta_k > \tau_1 \tilde{\Delta}.$$

In both cases, for all $k \geq K$,

$$\Delta_k \geq \min\left\{\tau_1 \tilde{\Delta}, \Delta_K\right\}. \tag{7}$$

Take $k \in \overline{\mathcal{S}}$ with $k \geq K$. By the definition of ρ_k given in (5), the condition (4), the contradiction hypothesis, and (7), it follows that

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\geq \eta_1 \theta_1 \pi_k \min\left\{\frac{\pi_k}{\kappa_H}, \Delta_k, 1\right\} \\ &\geq \eta_1 \theta_1 \varepsilon \min\left\{\frac{\varepsilon}{\kappa_H}, \min\left\{\tau_1 \tilde{\Delta}, \Delta_K\right\}, 1\right\}. \end{aligned}$$

By Assumption A2, the sequence $\{f(x^k)\}$ is bounded below, and since it is monotonically non-increasing, $f(x^k) - f(x^{k+1}) \to 0$. As the right-hand side of the inequality above is a positive constant, the set $\{k \in \overline{\mathcal{S}} \mid k \geq K\}$ is finite. Thus, for all k sufficiently large, $\delta_k > \beta \pi_k$ or $\rho_k < \eta_1$. However, by Lemma 3, $\delta_k \to 0$, and since $\pi_k \geq \varepsilon$ for all $k \geq K$, it follows that $\rho_k < \eta_1$ for all k sufficiently large, which implies that $\Delta_{k+1} = \tau_1 \Delta_k$. Consequently, $\Delta_k \to 0$, contradicting (7) and concluding the proof.

Assuming a sufficient decrease in the objective function, i.e., setting $\eta > 0$ in Algorithm 1, we prove that not only there exists a subsequence of $\{\pi_k\}$ converging to zero, but that the convergence prevails for the whole sequence.

Lemma 5 *Suppose that Assumptions A1–A3 hold and $\eta > 0$. Then*

$$\lim_{k \to \infty} \pi_k = 0.$$

Proof Suppose by contradiction that for some $\varepsilon > 0$, the set

$$\mathbb{N}' = \{k \in \mathbb{N} \mid \pi_k \geq \varepsilon\}$$

is infinite. Given $k \in \mathbb{N}'$, consider ℓ_k the first index such that $\ell_k > k$ and $\pi_{\ell_k} \leq \varepsilon/2$. The existence of ℓ_k is ensured by Lemma 4. So, $\pi_k - \pi_{\ell_k} \geq \dfrac{\varepsilon}{2}$. Using the definition of π_k, the triangle inequality, and the linearity and contraction properties of the projection operator, we have

$$\begin{aligned} \frac{\varepsilon}{2} &\leq \|P_\Omega(x^k - g_k) - x^k\| - \|P_\Omega(x^{\ell_k} - g_{\ell_k}) - x^{\ell_k}\| \\ &\leq \|P_\Omega(x^k - g_k) - x^k - P_\Omega(x^{\ell_k} - g_{\ell_k}) + x^{\ell_k}\| \\ &\leq \|P_\Omega(x^k - x^{\ell_k} + g_{\ell_k} - g_k)\| + \|x^k - x^{\ell_k}\| \\ &\leq 2\|x^k - x^{\ell_k}\| + \|g_k - g_{\ell_k}\|. \end{aligned}$$

Using again the triangle inequality and Assumptions A1 and A3, for $k \in \mathbb{N}'$,

$$\begin{aligned}\frac{\varepsilon}{2} &\leq 2\|x^k - x^{\ell_k}\| + \|g_k - \nabla f(x^k)\| + \|\nabla f(x^k) - \nabla f(x^{\ell_k})\| + \|\nabla f(x^{\ell_k}) - g_{\ell_k}\| \\ &\leq (2 + L_g)\|x^k - x^{\ell_k}\| + \kappa_m(\delta_k + \delta_{\ell_k}). \end{aligned} \tag{8}$$

On the other hand, by Lemma 3, $\delta_k \to 0$. Thus, there exists $k_0 \in \mathbb{N}$ such that for $k \geq k_0$, $\delta_k < \frac{\varepsilon}{8\kappa_m}$, where κ_m is the constant given in Assumption A3. Consequently, for $k \in \mathbb{N}'$, $k \geq k_0$,

$$\kappa_m(\delta_k + \delta_{\ell_k}) \leq \frac{\varepsilon}{4}. \tag{9}$$

Applying this in (8), it follows that

$$\|x^k - x^{\ell_k}\| \geq \frac{\varepsilon}{4(2 + L_g)}. \tag{10}$$

For $k \in \mathbb{N}'$, $k \geq k_0$, consider the set

$$C_k = \{j \in \mathcal{S} | k \leq j < \ell_k\},$$

which is non-empty. In fact, if $C_k = \emptyset$, then $x^k = x^{\ell_k}$, which means that $x^i \notin \mathcal{S}$ for $k \leq i < \ell_k$. In this case, by Assumption A3,

$$\begin{aligned}\frac{\varepsilon}{2} &\leq 2\|x^k - x^{\ell_k}\| + \|g_k - g_{\ell_k}\| \\ &= \|g_k - g_{\ell_k}\| \\ &\leq \|g_k - \nabla f(x^k)\| + \|\nabla f(x^{\ell_k}) - g_{\ell_k}\| \\ &\leq \kappa_m(\delta_k + \delta_{\ell_k}),\end{aligned}$$

which contradicts (9). So, by the definition of C_k, (4) and Lemma 1,

$$f(x^k) - f(x^{\ell_k}) = \sum_{j \in C_k} \left(f(x^j) - f(x^{j+1})\right) \geq \sum_{j \in C_k} \eta\theta_1\pi_j \min\left\{\frac{\pi_j}{\kappa_H}, \Delta_j, 1\right\}.$$

From the definition of ℓ_k, we have that $\pi_j > \varepsilon/2$ for all $j \in C_k$. Thus,

$$f(x^k) - f(x^{\ell_k}) \geq \sum_{j \in C_k} \eta\theta_1\frac{\varepsilon}{2} \min\left\{\frac{\varepsilon}{2\kappa_H}, \Delta_j, 1\right\} \geq \eta\theta_1\frac{\varepsilon}{2} \min\left\{\frac{\varepsilon}{2\kappa_H}, \sum_{j \in C_k}\Delta_j, 1\right\}. \tag{11}$$

But, from (10),

$$\frac{\varepsilon}{4(2 + L_g)} \leq \|x^k - x^{\ell_k}\| \leq \sum_{j \in C_k} \|x^j - x^{j+1}\| \leq \sum_{j \in C_k} \Delta_j.$$

So, as $\eta > 0$, it follows that the right-hand side of (11) is a positive constant. On the other hand, by Assumption A2, the sequence $\{f(x^k)\}$ is bounded below, and by the algorithm, it is monotonically non-increasing. Consequently, $f(x^k) - f(x^{\ell_k}) \to 0$, which is a contradiction, completing the proof.

Now we present the main result of this section.

Theorem 1 *Suppose that Assumptions A1–A3 hold. Then*

(i) If $\eta = 0$, $\liminf_{k\to\infty} \|P_\Omega(x^k - \nabla f(x^k)) - x^k\| = 0$.
(ii) If $\eta > 0$, $\lim_{k\to\infty} \|P_\Omega(x^k - \nabla f(x^k)) - x^k\| = 0$.

Proof By the triangle inequality, the projection operator properties, and Assumption A3, we have that

$$\begin{aligned}
&\|P_\Omega(x^k - \nabla f(x^k)) - x^k\| \\
&\quad = \|P_\Omega(x^k - \nabla f(x^k)) - P_\Omega(x^k - g_k) + P_\Omega(x^k - g_k) - x^k\| \\
&\quad \le \|P_\Omega(x^k - \nabla f(x^k)) - P_\Omega(x^k - g_k)\| + \|P_\Omega(x^k - g_k) - x^k\| \\
&\quad \le \|\nabla f(x^k) - g_k\| + \|P_\Omega(x^k - g_k) - x^k\| \\
&\quad \le \kappa_m \delta_k + \pi_k.
\end{aligned}$$

Using Lemmas 3, 4, and 5, we conclude the proof.

By Theorem 1, we conclude that if $\eta > 0$ and Algorithm 1 generates a sequence $\{x^k\}$ with an accumulation point x^*, then x^* is a stationary point of first order (Conn et al. 2000). One way to ensure the existence of an accumulation point is to assume that the level set $\left\{x \in \mathbb{R}^n \mid f(x) \le f(x^0)\right\}$ is bounded.

3 Ensuring Assumption A3

In this section, we discuss hypotheses in the construction of the models to ensure Assumption A3.

Let $\mathcal{P}_n^a$ be the space of polynomials of degree less than or equal to a in $\mathbb{R}^n$, whose dimension is

$$\dim \mathcal{P}_n^a = \frac{(n+a)!}{n!a!}.$$

In particular, $\dim \mathcal{P}_n^1 = n + 1$ and $\dim \mathcal{P}_n^2 = (n+1)(n+2)/2$.

Consider the set $X = \{x^0, x^1, \ldots, x^p\} \subset B(x^0, \delta)$, where $p = \dim \mathcal{P}_n^a - 1$. Assume that the function value in each point of X is known. Our task is to build a linear or quadratic model $m \in \mathcal{P}_n^a$ that approximates the function f in $B(x^0, \delta)$. The quality of the model depends on the geometry of the set X as we will discuss ahead.

In the linear case, $p = n$ and we define the matrix

$$L_\ell = \begin{pmatrix} (x^1 - x^0)^\top \\ \vdots \\ (x^n - x^0)^\top \end{pmatrix} = \begin{pmatrix} x_1^1 - x_1^0 & \cdots & x_n^1 - x_n^0 \\ \vdots & \ddots & \vdots \\ x_1^n - x_1^0 & \cdots & x_n^n - x_n^0 \end{pmatrix}, \tag{12}$$

and the scaled matrix

$$\widehat{L}_\ell = \frac{1}{\delta} L_\ell. \tag{13}$$

In the quadratic case, $p = q$ with $q = (n^2 + 3n)/2$ and we define the matrix

$$L_q = \begin{pmatrix} \left(\bar{\varphi}(x^1 - x^0)\right)^\top \\ \vdots \\ \left(\bar{\varphi}(x^q - x^0)\right)^\top \end{pmatrix}, \tag{14}$$

where

$$\bar{\varphi}(x) = \left(x_1, x_2, \ldots, x_n, \frac{1}{2}x_1^2, x_1x_2, x_1x_3, \ldots, x_1x_n, \frac{1}{2}x_2^2, \ldots, x_{n-1}x_n, \frac{1}{2}x_n^2\right)^\top$$

is the vector whose elements are the polynomials of the natural basis, as defined in Section 3.1 of Conn et al. (2009c). We also consider the scaled matrix

$$\widehat{L}_q = L_q \begin{pmatrix} D_\ell^{-1} & 0 \\ 0 & D_q^{-1} \end{pmatrix}, \tag{15}$$

where $D_\ell = \delta I_{n\times n}$ and $D_q = \delta^2 I_{(q-n)\times(q-n)}$.

In the previous sections, we state some assumptions to prove convergence of a trust-region method. To achieve Assumption A3, assume that the following assumptions hold.

A 4 *If the model to be constructed is linear, then the matrix L_ℓ is non-singular and there exists a constant $\kappa_\ell > 0$ such that $\|\widehat{L}_\ell^{-1}\| \leq \kappa_\ell$. If the model is quadratic, then the matrix L_q is non-singular and there exists a constant $\kappa_q > 0$ such that $\|\widehat{L}_q^{-1}\| \leq \kappa_q$.*

A 5 *There is a constant $\kappa \geq 0$ such that for all $x^i \in X \subset B(x^0, \delta)$,*

$$|m(x^i) - f(x^i)| \leq \kappa\delta^2.$$

Assumption A5 is weaker than asking that the model interpolates the objective function in the sample set, as required in theoretical analysis from Conn et al. (2009c). Our task is to prove a similar result assuming this weaker hypothesis. Before proving it, we state an auxiliary lemma proven in Dennis Jr and Schnabel (1996).

Lemma 6 (Dennis Jr and Schnabel 1996, Lemma 4.1.12) *Assume that A1 holds. Then for all $x \in \mathcal{X}$ and $d \in \mathbb{R}^n$ such that $x + d \in \mathcal{X}$,*

$$|f(x + d) - f(x) - \nabla f(x)^\top d| \leq \frac{1}{2} L_g \|d\|^2.$$

Theorem 2 *Suppose that Assumptions A1, A2, A4, and A5 hold. Then there exist constants* $\kappa_H, \kappa_g, \kappa_f \geq 0$ *such that, for all* $x \in \mathcal{X} \cap B(x^0, \delta)$,

$$\|\nabla^2 m(x)\| \leq \kappa_H, \tag{16}$$

$$\|\nabla f(x) - \nabla m(x)\| \leq \kappa_g \delta \tag{17}$$

and

$$|f(x) - m(x)| \leq \kappa_f \delta^2. \tag{18}$$

Proof There are two cases to be considered.

Linear case If the model is linear, it can be written as $m(x) = w^\top x + b$, with $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. In this case, we can set $\kappa_H = 0$.

By the triangle inequality, the fact that $\nabla m(x) = w$, Lemma 6 and Assumption A5, for all $i = 1, \ldots, n$, and $x \in \mathcal{X} \cap B(x^0, \delta)$,

$$\begin{aligned} \| \left(\nabla f(x^0) - \nabla m(x)\right)^\top (x^i - x^0)\| &\leq |\nabla f(x^0)^\top (x^i - x^0) + f(x^0) - f(x^i)| + \\ &\quad + |f(x^i) - w^\top x^i - b| + |w^\top x^0 + b - f(x^0)| \\ &\leq \frac{1}{2} L_g \delta^2 + 2\kappa \delta^2. \end{aligned}$$

By the definition of matrix L_ℓ in (12), we have

$$\left\| L_\ell \left(\nabla f(x^0) - \nabla m(x)\right)\right\| \leq \sqrt{n} \left\| L_\ell \left(\nabla f(x^0) - \nabla m(x)\right)\right\|_\infty \leq \sqrt{n} \left(\frac{1}{2} L_g + 2\kappa\right) \delta^2.$$

Then, by (13) and Assumption A4,

$$\left\| \nabla f(x^0) - \nabla m(x)\right\| \leq \|L_\ell^{-1}\| \left\| L_\ell \left(\nabla f(x^0) - \nabla m(x)\right)\right\| \leq \kappa_\ell \sqrt{n} \left(\frac{1}{2} L_g + 2\kappa\right) \delta. \tag{19}$$

From the triangle inequality and Assumption A1, it follows that, for all $x \in \mathcal{X} \cap B(x^0, \delta)$,

$$\begin{aligned} \|\nabla f(x) - \nabla m(x)\| &\leq \|\nabla f(x) - \nabla f(x^0)\| + \|\nabla f(x^0) - \nabla m(x)\| \\ &\leq \left(L_g + \kappa_\ell \sqrt{n} \left(\frac{1}{2} L_g + 2\kappa\right)\right) \delta. \end{aligned}$$

Defining $\kappa_g = L_g + \kappa_\ell \sqrt{n} \left(\frac{1}{2} L_g + 2\kappa\right)$, we obtain (17).

By the Taylor expansion of m around x^0, the triangle inequality, Lemma 6, Assumption A5, the Cauchy–Schwarz inequality and (19), we can see that, for all $x \in \mathcal{X} \cap B(x^0, \delta)$,

$$
\begin{aligned}
|f(x) - m(x)| &\leq |f(x) - f(x^0) - \nabla f(x^0)^\top (x - x^0)| + |f(x^0) - m(x^0)| \\
&\quad |\left(\nabla f(x^0) - \nabla m(x^0)\right)^\top (x - x^0)| \\
&\leq \left(\frac{1}{2} L_g + \kappa + \left(\frac{1}{2} L_g + 2\kappa\right) \sqrt{n} \kappa_\ell\right) \delta^2.
\end{aligned}
$$

Defining $\kappa_f = \frac{1}{2} L_g + \kappa + \left(\frac{1}{2} L_g + 2\kappa\right) \sqrt{n} \kappa_\ell$, we obtain (18).

Quadratic case If the model is quadratic, for all $i = 0, 1, \ldots, q$ and for all $x \in \mathcal{X} \cap B(x^0, \delta)$,

$$
m(x^i) = m(x) + \nabla m(x)^\top (x^i - x) + \frac{1}{2} (x^i - x)^\top H (x^i - x) \tag{20}
$$

with $H \in \mathbb{R}^{n \times n}$ symmetric.

By the mean value theorem, for all $i = 0, 1, \ldots, q$, there exists $y^i \in [x^0, x^i] \subset B(x^0, \delta)$ such that

$$
f(x^i) = f(x^0) + \nabla f(y^i)^\top (x^i - x^0).
$$

Therefore, by (20),

$$
\begin{aligned}
m(x^i) - f(x^i) = m(x) + \nabla m(x)^\top (x^i - x) + \frac{1}{2} (x^i - x)^\top H (x^i - x) - f(x^0) \\
- \nabla f(y^i)^\top (x^i - x^0),
\end{aligned} \tag{21}
$$

for all $i = 0, 1, \ldots, q$.

Taking $i = 0$ in (21) and subtracting from (21), for $i = 1, ..., q$, we have that

$$
\begin{aligned}
&m(x^i) - f(x^i) + f(x^0) - m(x^0) + \nabla f(y^i)^\top (x^i - x^0) \\
&\quad = \nabla m(x)^\top (x^i - x^0) + \frac{1}{2} (x^i - x)^\top H (x^i - x) - \frac{1}{2} (x^0 - x)^\top H (x^0 - x) \\
&\quad = \nabla m(x)^\top (x^i - x^0) + \frac{1}{2} (x^i - x^0)^\top H (x^i - x^0) - (x^i - x^0)^\top H (x - x^0).
\end{aligned}
$$

Subtracting $\nabla f(x)^\top (x^i - x^0)$ on both sides, we get

$$
\begin{aligned}
&m(x^i) - f(x^i) + f(x^0) - m(x^0) + (\nabla f(y^i) - \nabla f(x))^\top (x^i - x^0) \\
&\quad = \left(\nabla m(x) - \nabla f(x) - H(x - x^0) + \frac{1}{2} H (x^i - x^0)\right)^\top (x^i - x^0).
\end{aligned}
$$

By the triangle inequality, the Cauchy–Schwarz inequality, and Assumptions A1 and A5, we have that, for all $x \in \mathcal{X} \cap B(x^0, \delta)$, and for $i = 1, \ldots, q$,

$$\left| \left(\nabla m(x) - \nabla f(x) - H(x - x^0) + \frac{1}{2} H(x^i - x^0) \right)^\top (x^i - x^0) \right| \leq 2(\kappa + L_g)\delta^2,$$

since $\|y^i - x\| \leq 2\delta$. Defining

$$r^g(x) = \nabla m(x) - \nabla f(x) - H(x - x^0) \tag{22}$$

and $r^H(x)$ as a vector in $\mathbb{R}^{q-n}$ that stores the elements H_{kk}, $k = 1, \ldots, n$ and $H_{k\ell}$, $1 \leq \ell < k \leq n$, from the definition of L_g in (14), it follows that

$$\left\| L_q \begin{pmatrix} r^g(x) \\ r^H(x) \end{pmatrix} \right\| \leq \sqrt{q} \left\| L_q \begin{pmatrix} r^g(x) \\ r^H(x) \end{pmatrix} \right\|_\infty \leq 2\sqrt{q}(\kappa + L_g)\delta^2.$$

Using (15) and Assumption A4,

$$\begin{aligned} \left\| \begin{pmatrix} D_\ell r^g(x) \\ D_q r^H(x) \end{pmatrix} \right\| &\leq \left\| \widehat{L}_q^{-1} \right\| \left\| \widehat{L}_q \begin{pmatrix} D_\ell r^g(x) \\ D_q r^H(x) \end{pmatrix} \right\| = \left\| \widehat{L}_q^{-1} \right\| \left\| L_q \begin{pmatrix} r^g(x) \\ r^H(x) \end{pmatrix} \right\| \\ &\leq 2\kappa_q \sqrt{q}(\kappa + L_g)\delta^2. \end{aligned}$$

Consequently,

$$\|r^g(x)\| \leq \|D_\ell^{-1}\| \|D_\ell r^g(x)\| \leq 2\kappa_q \sqrt{q}(\kappa + L_g)\delta, \tag{23}$$

and

$$\|H\| \leq \|H\|_F \leq \sqrt{2}\|r^H(x)\| \leq \sqrt{2}\|D_q^{-1}\| \|D_q r^H(x)\| \leq 2\kappa_q \sqrt{2q}(\kappa + L_g). \tag{24}$$

Defining $\kappa_H = 2\kappa_q \sqrt{2q}(\kappa + L_g)$, we obtain (16).

By (22), (23), and (24),

$$\|\nabla m(x) - \nabla f(x)\| \leq \|r^g(x)\| + \|H\| \, \|x - x^0\| \leq 2\kappa_q \sqrt{q}(1 + \sqrt{2})(\kappa + L_g)\delta.$$

Defining $\kappa_g = 2\kappa_q \sqrt{q}(1 + \sqrt{2})(\kappa + L_g)$, we obtain (17).

Finally, note that, for all $x \in \mathcal{X} \cap B(x^0, \delta)$,

$$m(x) = m(x^0) + \nabla m(x^0)^\top (x - x^0) + \frac{1}{2}(x - x^0)^\top H(x - x^0).$$

Thus, by the triangle inequality, the Cauchy–Schwarz inequality, Lemma 6, and Assumption A5,

$$\begin{aligned}|f(x)-m(x)| &\le |f(x)-f(x^0)-\nabla f(x^0)^\top (x-x^0)| + |f(x^0)-m(x^0)| \\ &\quad + |\left(\nabla f(x^0)-\nabla m(x^0)\right)^\top (x-x^0)| + \frac{1}{2}|(x-x^0)H(x-x^0)|. \\ &\le \left(\frac{1}{2}L_g + \kappa + \kappa_q\sqrt{q}(\kappa+L_g)(2+3\sqrt{2})\right)\delta^2.\end{aligned}$$

Defining $\kappa_f = \frac{1}{2}L_g + \kappa + \kappa_q\sqrt{q}(\kappa+L_g)(2+3\sqrt{2})$, we obtain (18), which concludes the proof.

Theorem 2 states that if the function f satisfies A1–A2, Assumptions A4 and A5 imply A3. Thus, we show that the quadratic model provides sufficient first-order accuracy. It is worth mentioning that it is possible to prove second-order error bounds for the quadratic models, similarly to Theorem 3.16 of Conn et al. (2009c), if A4 and A5 are made appropriately stronger. However, we do not focus on the second-order error bounds here, because we do not need them for our first-order convergence results. In the next sections, we will discuss how A4 and A5 can be ensured.

4 Sample set geometry

The main task of this section is to establish conditions on the sample set to ensure Assumption A4. First we define a measure of poisedness for a set of points.

Definition 1 (Conn et al. 2009c, *Definition 3.6*) Let $\phi = \{\phi_0(x), \phi_1(x), \ldots, \phi_p(x)\}$ be a basis in $\mathcal{P}_n^a$ and consider $\Lambda > 0$. A set $X = \{x^0, \ldots, x^p\} \subset \mathcal{X}$ is Λ-poised in $B(x^0, \delta)$ with respect to the base ϕ if for each $x \in B(x^0, \delta)$ there exists $\lambda(x) \in \mathbb{R}^{p+1}$ such that

$$\sum_{i=0}^{p} \lambda_i(x)\phi_j(x^i) = \phi_j(x) \quad \text{for all} \quad j = 0, \ldots, p, \qquad \text{with} \qquad \|\lambda(x)\|_\infty \le \Lambda.$$

Note that these conditions can be written as

$$M(\phi, X)^\top \lambda(x) = \phi(x) \qquad \text{with} \qquad \|\lambda(x)\|_\infty \le \Lambda, \tag{25}$$

where

$$M(\phi, X) = \begin{pmatrix} \phi_0(x^0) & \phi_1(x^0) & \phi_2(x^0) & \cdots & \phi_p(x^0) \\ \phi_0(x^1) & \phi_1(x^1) & \phi_2(x^1) & \cdots & \phi_p(x^1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_0(x^p) & \phi_1(x^p) & \phi_2(x^p) & \cdots & \phi_p(x^p) \end{pmatrix} \quad \text{and} \quad \phi(x) = \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_p(x) \end{pmatrix}.$$

The definition of Λ-poisedness is independent of the choice of the basis. In fact, given $[I]_A^B$ the change in basis matrix from ϕ^A to ϕ^B, the solution $\lambda(x)$ of (25) does not depend of the basis ϕ, once

$$M(\phi^A, X)^\top \lambda(x) = \phi^A(x) \Rightarrow [I]_A^B M(\phi^A, X)^\top \lambda(x) = [I]_A^B \phi^A(x)$$

which implies that $M(\phi^B, X)^\top \lambda(x) = \phi^B(x)$.

The poisedness constant Λ does not depend on the scaling of the sample set and is invariant with respect to shifts of coordinates (Conn et al. 2009c, Lemmas 3.8 and 3.9).

Now we state an auxiliary lemma.

Lemma 7 *(Conn et al. 2009c, Lemma 3.13) Let $v \in \mathbb{R}^n$ be a unitary right singular vector corresponding to the largest singular value of a non-singular matrix $A \in \mathbb{R}^{n\times n}$. Then, for any vector $z \in \mathbb{R}^n$,*

$$|v^\top z| \|A\| \leq \|Az\|.$$

The following result ensures that Assumption A4 holds when the model to be constructed is linear.

Lemma 8 *Let $X = \{x^0, x^1, \ldots, x^n\}$ be a Λ-poised set in $B(x^0, \delta)$ with respect to the basis ϕ in $\mathcal{P}_n^1$ and $\widehat{L}_\ell$ be the matrix defined in (13). Then*

$$\|\widehat{L}_\ell^{-1}\| \leq \Lambda\sqrt{n}.$$

Proof Consider v a unitary right singular vector of $\widehat{L}_\ell^{-1}$ associated with the largest singular value σ_1. There exists a unitary vector $u \in \mathbb{R}^n$ such that $\widehat{L}_\ell^{-1} v = \sigma_1 u$. Consequently,

$$\|\widehat{L}_\ell^{-1} v\| = \sigma_1 \|u\| = \sigma_1 = \|\widehat{L}_\ell^{-1}\|. \tag{26}$$

Since the Λ-poisedness does not depend on the scaling of the sample set and is invariant with respect to a shift of coordinates, it follows that

$$\widehat{X} = \left\{0, \frac{x^1 - x^0}{\delta}, \ldots, \frac{x^n - x^0}{\delta}\right\}$$

is Λ-poised in $B(0, 1)$. So there exists $\lambda(v) \in \mathbb{R}^n$ with $\|\lambda(v)\|_\infty \leq \Lambda$ such that $\widehat{L}_\ell \lambda(v) = v$. Thus $\lambda(v) = \widehat{L}_\ell^{-1} v$. Using this, (26) and the fact that X is a Λ-poised set, we have

$$\|\widehat{L}_\ell^{-1}\| = \|\lambda(v)\| \leq \sqrt{n}\|\lambda(v)\|_\infty \leq \Lambda\sqrt{n},$$

which concludes the proof.

To prove a similar result for the quadratic case, we first state an auxiliary lemma.

Lemma 9 (Conn et al. 2009c, Lemma 6.7) *Let $v^\top \bar{\phi}(x)$ be a quadratic polynomial, where $\|v\|_\infty = 1$ and $\bar{\phi}$ is the natural basis in $\mathcal{P}_n^2$. Then*

$$\max_{x \in B(0,1)} |v^\top \bar{\phi}(x)| \geq \frac{1}{4}.$$

Given $\bar{v} \in \mathbb{R}^{q+1}$ with $\|\bar{v}\| = 1$, there exist $\beta \in (0, \sqrt{q+1})$ and $v \in \mathbb{R}^{q+1}$ such that $\|v\|_\infty = 1$ and $v = \beta \bar{v}$. Then, by Lemma 9,

$$\max_{x \in B(0,1)} |\bar{v}^\top \bar{\phi}(x)| = \max_{x \in B(0,1)} \frac{1}{\beta} |v^\top \bar{\phi}(x)| \geq \frac{1}{\sqrt{q+1}} \max_{x \in B(0,1)} |v^\top \bar{\phi}(x)| \geq \frac{1}{4\sqrt{q+1}}. \tag{27}$$

The following result ensures that the Assumption A4 holds when the model to be constructed is quadratic.

Lemma 10 *Let $X = \{x^0, x^1, \ldots, x^q\}$ be a Λ-poised set in $B(x^0, \delta)$ with respect to the basis ϕ in $\mathcal{P}_n^2$ and $\widehat{L}_q$ be the matrix defined in (15). Then,*

$$\|\widehat{L}_q^{-1}\| \leq 4\Lambda\sqrt{(q+1)^3}.$$

Proof Consider $\widehat{X} = \left\{0, \frac{x^1 - x^0}{\delta}, \ldots, \frac{x^q - x^0}{\delta}\right\}$ which is Λ-poised in $B(0,1)$ with respect to $\bar{\phi}$ the natural basis in $\mathcal{P}_n^2$. Consider also $\widehat{M} = \widehat{M}(\bar{\phi}, \widehat{X})$ the matrix of the linear system in (25) and L_q defined in (15). Then

$$\widehat{M} = \begin{pmatrix} 1 & 0 \\ e & \widehat{L}_q \end{pmatrix} \quad \text{and} \quad \widehat{M}^{-1} = \begin{pmatrix} 1 & 0 \\ -\widehat{L}_q^{-1} e & \widehat{L}_q^{-1} \end{pmatrix}.$$

By the equivalence between the Frobenius and the Euclidean norms,

$$\|\widehat{L}_q^{-1}\| \leq \|\widehat{L}_q^{-1}\|_F \leq \|\widehat{M}^{-1}\|_F \leq \sqrt{q+1}\|\widehat{M}^{-1}\|. \tag{28}$$

For each $x \in B(0,1)$, the solution $\lambda(x) \in \mathbb{R}^{q+1}$ of (25) satisfies

$$\Lambda \geq \|\lambda(x)\|_\infty \geq \frac{1}{\sqrt{q+1}}\|\lambda(x)\| = \frac{1}{\sqrt{q+1}}\|\widehat{M}^{-\top}\bar{\phi}(x)\|.$$

Let $\bar{v}$ be a unitary right singular vector associated with the largest singular value of $\widehat{M}^{-\top}$. Consider $x \in \mathbb{R}^n$ the maximizer of $|\bar{v}^\top \bar{\phi}(x)|$ in $B(0,1)$. By the inequality above, Lemma 7 and (27),

$$\sqrt{q+1}\,\Lambda \geq \|\widehat{M}^{-\top}\bar{\phi}(x)\| \geq |\bar{v}^\top \bar{\phi}(x)|\|\widehat{M}^{-\top}\| \geq \frac{1}{4\sqrt{q+1}}\|\widehat{M}^{-1}\|.$$

By (28),

$$\|\widehat{L}_q^{-1}\| \leq \sqrt{q+1}\|\widehat{M}^{-1}\| \leq 4\Lambda\sqrt{(q+1)^3},$$

which concludes the proof.

When the sample set is Λ-poised, Lemmas 8 and 10 ensure Assumption A4 in the linear and quadratic case, respectively.

5 Model building

In this section, we discuss techniques to construct the models in derivative-free context which ensure Assumption A5. The standard one is polynomial interpolation (Conn et al. 2008a, b, 2009b, c; Powell 2002, 2006, 2008, 2009, 2012), which trivially satisfies the hypothesis. Alternatively, the models can be constructed by support vector regression (Drucker et al. 1996), as we will see in details later in this section.

5.1 Polynomial interpolation

Let ϕ be a basis of the space $\mathcal{P}_n^a$ of polynomials with degree less than or equal to a in $\mathbb{R}^n$, which dimension is $s+1$. As any polynomial $m \in \mathcal{P}_n^a$ can be written as a linear combination of the elements of ϕ, we have that

$$m(x) = \sum_{j=0}^{s} \mu_j \phi_j(x) = \mu^\top \phi(x),$$

where $\mu = (\mu_0, \mu_1, \ldots, \mu_s)^\top$ is a vector in $\mathbb{R}^{s+1}$ and $\phi(x) = (\phi_0(x), \ldots, \phi_s(x))^\top$ is the vector with the elements of ϕ.

We say that m interpolates the function f at $\bar{x}$ when $m(\bar{x}) = f(\bar{x})$. If $m \in \mathcal{P}_n^a$ interpolates f in the set $X = \{x^0, x^1, \ldots, x^p\} \subset \mathbb{R}^n$, then the coefficients $\mu_0, \ldots, \mu_s$ can be found by the interpolation conditions

$$m(x^i) = \sum_{j=0}^{s} \mu_j \phi_j(x^i) = f(x^i), \quad i = 0, \ldots, p,$$

which can be written as the linear system

$$M(\phi, X)\mu_\phi = f(X), \tag{29}$$

where

$$M(\phi, X) = \begin{pmatrix} \phi_0(x^0) & \phi_1(x^0) & \cdots & \phi_s(x^0) \\ \phi_0(x^1) & \phi_1(x^1) & \cdots & \phi_s(x^1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x^p) & \phi_1(x^p) & \cdots & \phi_s(x^p) \end{pmatrix},$$

$$\mu_\phi = \begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_s \end{pmatrix} \quad \text{and} \quad f(X) = \begin{pmatrix} f(x^0) \\ f(x^1) \\ \vdots \\ f(x^p) \end{pmatrix}.$$

The set X is said *poised* for polynomial interpolation when the linear system (29) has a unique solution. This occurs if, and only if, $p = s$ and the matrix $M(\phi, X)$ is non-singular. Since all bases in a finite-dimensional vector space are equivalent, the poisedness is independent of the choice of the basis ϕ. Some authors call a poised set for polynomial interpolation as unisolvent set (Quarteroni et al. 2007).

It is easy to see that a set X that is Λ-poised by Definition 1 is also poised for polynomial interpolation and that interpolation models satisfy Assumption A5. Therefore, polynomial interpolation in Λ-poised sets can be used to construct the models in the derivative-free trust-region algorithm presented in Sect. 2.

5.2 Support vector regression

We present an alternative way to build models for derivative-free trust-region methods. The idea is to construct an approximation of the function f, from a sample set $X = \{x^0, \ldots, x^p\}$, by using the technique of support vector machines developed for pattern classification by Vapnik (1998) and extended to regression by Drucker et al. (1996).

5.2.1 Linear support vector regression

In this section, we discuss how to determine $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ for constructing a linear model $m(x) = w^\top x + b$ that approximates the function f within $\varepsilon > 0$ accuracy, if possible. Ideally, we would like to solve the following convex quadratic optimization problem

$$\begin{aligned} \underset{(w,b)}{\text{minimize}} \ & \frac{1}{2} w^\top w \\ \text{subject to} \ & f(x^i) - w^\top x^i - b \leq \varepsilon, \ i = 0, \ldots, p \\ & w^\top x^i + b - f(x^i) \leq \varepsilon, \ i = 0, \ldots, p. \end{aligned}$$

However, since this problem does not always have a feasible solution, we allow some error and penalize it in the objective function. Thus, as presented in Section 9.1 by Schölkopf and Smola (2002), we consider the problem

$$\begin{aligned} \underset{(w,b,\xi,\xi')}{\text{minimize}} \ & \frac{1}{2} w^\top w + C \sum_{i=0}^{p} \left(\xi_i + \xi_i'\right) \\ \text{subject to} \ & f(x^i) - w^\top x^i - b \leq \varepsilon + \xi_i, \ i = 0, \ldots, p \\ & w^\top x^i + b - f(x^i) \leq \varepsilon + \xi_i', \ i = 0, \ldots, p \\ & \xi_i, \xi_i' \geq 0, \qquad\qquad i = 0, \ldots, p, \end{aligned} \tag{30}$$

where $\xi, \xi' \in \mathbb{R}^p$ are the error vectors and $C > 0$ is the penalty parameter. If ξ and ξ' are null, then m is considered an ε-approximation of f in X.

The dual problem of (30) is given by

$$\begin{array}{ll} \underset{z}{\text{minimize}} & \frac{1}{2} z^\top Q z + v^\top z \\ \text{subject to} & Az = 0 \\ & 0 \leq z \leq C, \end{array} \tag{31}$$

where

$$Q = \begin{pmatrix} PP^\top & -PP^\top \\ -PP^\top & PP^\top \end{pmatrix} \quad \text{with} \quad P = \begin{pmatrix} (x^0)^\top \\ (x^1)^\top \\ \vdots \\ (x^p)^\top \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} -f(P) + \varepsilon e \\ f(P) + \varepsilon e \end{pmatrix} \tag{32}$$

and $A = \left(-e^\top, e^\top\right)$ with $e = (1, 1, \ldots, 1)^\top$.

As suggested in Schölkopf and Smola (2002), from the dual solution $z^* = (\alpha, \alpha')$, with $\alpha, \alpha' \in \mathbb{R}^p$, we compute w and b that define the model as

$$w = P^\top(\alpha - \alpha')$$

and

$$b = f(x^i) - w^\top x^i - \varepsilon \qquad \text{or} \qquad b = f(x^j) - w^\top x^j + \varepsilon$$

for any $i, j \in \{1, \ldots, p\}$ such that $0 < \alpha_i, \alpha'_j < C$.

5.2.2 Quadratic support vector regression

For building a quadratic model q, we lift the data to a space of higher dimension, called feature space, and construct a linear model in it. Defining

$$\varphi(x) = (x_1^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \ldots, x_2^2, \ldots, \sqrt{2}x_{n-1}x_n, x_n^2, x_1, x_2, \ldots, x_n)^\top,$$

the quadratic model can be written as $q(x) = w^\top \varphi(x) + b$, which is linear in the feature space. Following the approach of the last section, we consider the dual problem (31), in which the matrix P is now given by

$$P = \begin{pmatrix} (\varphi(x^0))^\top \\ (\varphi(x^1))^\top \\ \vdots \\ (\varphi(x^p))^\top \end{pmatrix}.$$

The coefficient $\sqrt{2}$ in some terms of the definition of φ provides a simple expression for the entries of $PP^\top$, which are obtained by $\varphi(x^i)^\top\varphi(x^j) = (x^i)^\top x^j + ((x^i)^\top x^j)^2$, implying that the mapping φ does not have to be explicitly computed.

For using support vector regression models in the derivative-free trust-region algorithm presented in Sect. 2, we need to ensure that Assumptions A4 and A5 are fulfilled. If the sample set in each iteration is Λ-poised, we have by Lemmas 8 and 10 that Assumption A4 holds in the linear and quadratic cases, respectively. On the other hand, given $c_1, c_2 > 0$, the support vector regression model m can be constructed solving (30) with $\varepsilon = c_1\delta^2$ and C so that $\max_{1\le i\le p}\{\xi_i, \xi_i'\} \le c_2\delta^2$, which implies that the error $|f(x^i) - m(x^i)|$ is not greater than $(c_1 + c_2)\delta^2$ for all $x^i \in X$ and A5 holds. The existence of such C is guaranteed by the fact that the interpolation model is a feasible solution of (30) with null error in the sample set. Thus, we can start with an arbitrary C and increase it if sufficient accuracy is not reached. So, if in each iteration the sample set is Λ-poised and C is sufficiently large, then, by Theorem 2, Assumption A3 holds.

6 Numerical experiments

In this section, we describe numerical experiments to illustrate the practical performance of the algorithm. The tests were performed in a notebook ACER Intel Core i7-5500U, CPU 2.40GHz, with 8GB RAM, 64-bit, using `Matlab` 2015a, v. 8.5.

Algorithm 1 was run with $\Delta_0 = \delta_0 = 1$, $\beta = 1$, $\tau_1 = 0.6$, $\tau_2 = 1.5$, $\eta = 0.1$, $\eta_1 = 0.3$, $\eta_2 = 0.6$, and $\Delta_{k+1} = \Delta_k$ whenever a new trust-region radius should be taken in the interval $[\delta_k, \Delta_k]$. The subproblems (3) were solved by the `Matlab` routine `trust` with default parameters. We declare success when the sample set radius became too small: $\|\delta_k\| \le 10^{-8}$, as suggested in Conejo et al. (2013).

Two variants of Algorithm 1 were considered. In the first one, the models were built by polynomial interpolation, as suggested in Sect. 5.1, while in the second one the models were constructed by support vector regression, as discussed in Sect. 5.2.

Independently of the chosen technique to build the models, the sample sets consisted of $(n+1)(n+2)/2$ points. The first set was constructed by taking steps of size δ_0 from x^0 in the positive and negative coordinate directions, resulting in $2n + 1$ points. The remaining points were obtained by Algorithm 6.2 of Conn et al. (2009c). The sample set was updated by replacing at most a point per iteration. A new iterate replaced the most distant sample point from it. In the iteration in which the trial point was rejected, it replaced the farthest sample point only if was closer to the current iterate. Otherwise, the sample set was kept unchanged for the next iteration. No safeguard was implemented to ensure the sample set well-poisedness.

When interpolation models were considered, they were obtained by solving the linear system (29). On the other hand, models constructed using support vector regression were obtained by solving problem (31), with $C = 10^8$, by the `Matlab` routine `quadprog` with default parameters. Several values for the parameter C were tested, such as $C = 10^j$ with $j = 1, \ldots, 10$ and also dynamical choices. Our tests showed that taking a fixed large C is more efficient than starting with a small value and increasing it when necessary. We adopted $C = 10^8$, which provided the best performance.

The following stopping criteria, with the respective exit flags, were considered as follows:

(1) The algorithm succeeded: $\|\delta_k\| \leq 10^{-8}$.
(−1) The number of iterations exceeded 5000.
(−2) The predicted reduction became too small: $m_k(x_k) - m_k(x_k + d_k) \leq 10^{-32}$.
(−3) The norm of the gradient of the model at the current point became too small: $\|g_k\| \leq 10^{-32}$.

The set of test problems consisted of all 35 problems from the Moré, Garbow, and Hillstrom collection (Moré et al. 1981), with default starting point x^0. The numerical experiments are discussed in two sections. First, we consider the functions as presented in the collection, and then the same set of problems adding noise.

6.1 Functions without noise

In this section, we discuss the numerical experiments for the problems from the Moré, Garbow, and Hillstrom collection. We set $\varepsilon = 10^{-8}$ in (32).

Table 1 presents the results. The first three columns contain information about the problems that consist of minimizing the sum of the square of m functions in $\mathbb{R}^n$. The label P indicates the number of the problem in the collection. The next columns display the number of function evaluations $\#f$, the minimum function value f^* and the exit flag for each instance of the algorithm. Finally, the column f_{MGH} displays the solution available in the collection.

As suggested in Birgin and Gentil (2012), given $\varepsilon_f > 0$ we considered that an instance of the algorithm solved a problem if it found a point $\overline{x}$ such that

$$\frac{f(\overline{x}) - f_{\min}}{\max\{1, |f_{\min}|\}} \leq \varepsilon_f, \tag{33}$$

where $f_{\min}$ is the smallest function value found among the instances that are being compared.

Figure 1 presents the performance profiles (Dolan and Moré 2009) related to the number of objective function evaluations, as usual in derivative-free optimization. In the figure on the left, we considered $\varepsilon_f = 10^{-4}$ in (33). Algorithm 1 with polynomial interpolation models solved all the 35 problems, while the algorithm using support vector regression models solved 22 problems, which represents 62.9% of the total amount. The algorithm using polynomial interpolation models was the most efficient for 74.3%, versus 31.4% achieved by the algorithm with support vector regression models. On the other hand, in the right figure, it was used $\varepsilon_f = 10^{-1}$. Algorithm 1 with polynomial interpolation models solved all problems being the most efficient in 57.1% of the problems. When support vector regression models were used, the algorithm solved 29 problems, being the most efficient in 48.6% of them.

Table 1 Numerical results for functions without noise

P	n	m	Interpolation			SVR			MGH
			$\#f$	f^*	Exit	$\#f$	f^*	Exit	f_{MGH}
1	2	2	102	3.8083e−17	1	134	8.3169e−06	1	0.0000e+00
2	2	2	50	4.8984e+01	1	64	4.8984e+01	1	4.8984e+01
3	2	2	476	3.0168e−19	1	50	1.3520e−01	1	0.0000e+00
4	2	3	171	6.6283e−11	1	41	9.7570e+11	−3	0.0000e+00
5	2	3	38	4.1242e−16	1	54	3.0937e−08	1	0.0000e+00
6	2	10	58	1.2436e+02	1	457	1.2436e+02	1	1.2436e+02
7	3	3	136	3.2112e−17	1	166	1.0816e−06	1	0.0000e+00
8	3	15	77	8.2149e−03	1	89	8.2149e−03	1	8.2149e−03
9	3	15	32	1.1279e−08	1	45	2.8035e−08	1	1.1279e−08
10	3	16	4437	1.1006e+02	1	88	7.0016e+06	1	8.7946e+01
11	3	20	389	3.8284e−06	1	77	3.6793e−02	1	0.0000e+00
12	3	20	172	1.3360e−17	1	106	2.4101e−02	1	0.0000e+00
13	4	4	307	9.3195e−14	1	194	2.5403e−05	1	0.0000e+00
14	4	6	514	1.9777e−15	1	501	1.7657e−03	1	0.0000e+00
15	4	11	149	3.0751e−04	1	140	4.0657e−04	1	3.0751e−04
16	4	20	133	8.5822e+04	1	440	8.5825e+04	1	8.5822e+04
17	5	33	437	7.0942e−05	1	21	8.7903e−01	−3	5.4649e−05
18	6	13	519	1.6961e−07	1	591	2.3280e−01	1	5.6556e−03
19	11	65	115	2.0934e+00	1	115	2.0934e+00	1	4.0138e−02
20	6	31	308	2.2877e−03	1	346	9.1860e−03	1	2.2877e−03
21	8	8	1263	5.3845e−14	1	889	7.6726e−03	1	0.0000e+00
22	8	8	861	1.2894e−09	1	426	3.3634e−04	1	0.0000e+00
23	10	11	3672	7.0877e−05	1	646	7.6432e−05	1	7.0877e−05
24	10	20	1492	2.9405e−04	1	394	2.9872e−04	1	2.9366e−04
25	10	12	937	1.3251e−13	1	539	2.1971e+00	1	0.0000e+00
26	10	10	814	2.7951e−05	1	366	5.1668e−05	1	0.0000e+00
27	10	10	73	4.6154e−24	1	95	4.2309e−10	1	0.0000e+00
28	10	10	232	3.9582e−16	1	110	7.8487e−04	1	0.0000e+00
29	10	10	165	1.0177e−13	1	146	1.8783e−07	1	0.0000e+00
30	6	6	132	9.5501e−16	1	127	1.2753e−06	1	0.0000e+00
31	5	5	153	9.0113e−16	1	135	9.8880e−07	1	0.0000e+00
32	6	6	81	2.2388e−25	1	57	2.0578e−10	1	0.0000e+00
33	6	6	30	1.1538e+00	1	46	1.1538e+00	1	1.1538e+00
34	6	6	30	2.6667e+00	1	46	2.6667e+00	1	2.6667e+00
35	9	9	92	2.8883e−02	1	92	2.8883e−02	1	0.0000e+00

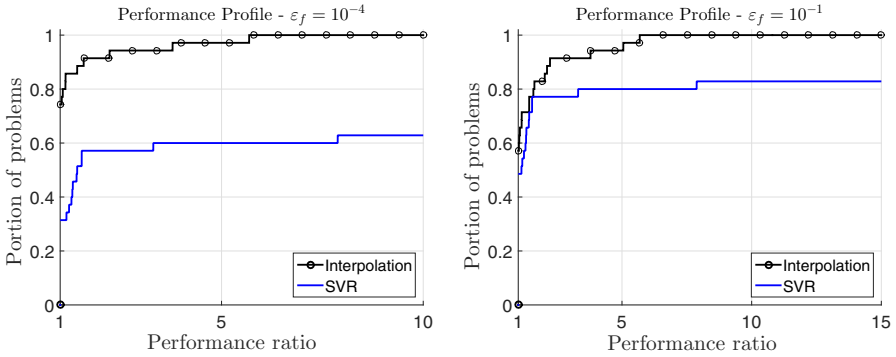

Fig. 1 Performance profiles for functions without noise

6.2 Noisy functions

In this section, we discuss numerical results for the Moré, Garbow, and Hillstrom collection adding a uniform random noise between -10^{-3} and 10^{-3} to the functions. Noisy optimization problems are common in applications, for example when the objective function is associated with a simulation. In this situation, the objective function values are often inaccurate and polynomial interpolation may be a bad choice for building the models. On the other hand, support vectors regression is likely to produce models that do not try to capture the noise accurately. We compare the two alternatives for constructing the models. For support vectors regression models, we set $\varepsilon = 10^{-3}$, the same order of the noise.

Table 2 presents the results, and Fig. 2 shows the performance profiles for noisy functions. In the figure on the left, we considered $\varepsilon_f = 10^{-4}$ in (33). Algorithm 1 with support vector regression models solved 25 problems (71.4% of the total amount), while the algorithm using polynomial interpolation models solved 17 problems (48.6%). The algorithm using support vector regression models was the most efficient for 65.7%, versus 42.9% achieved by the algorithm with polynomial interpolation models. On the other hand, in the right figure, it was used $\varepsilon_f = 10^{-1}$. When support vector regression models were used, the algorithm solved 32 problems, being the most efficient in 65.7% of them. With polynomial interpolation models solved 27 problems, being the most efficient in 42.9% of the problems.

7 Conclusion

In this work, we discuss the global convergence of a derivative-free trust-region algorithm for minimizing a function on a closed convex set. The algorithm has a very simple structure and considers two radii: the trust-region radius and another one that controls the sample set. The global convergence is proven assuming that the gradient of the model is a good approximation for the gradient of the objective function at the current point. We proved that this property holds for models constructed by polynomial interpolation and by support vector regression. Preliminary numerical results are

Table 2 Numerical results for noisy functions

P	n	m	Interpolation			SVR		
			$\#f$	f^*	Exit	$\#f$	f^*	Exit
1	9	9	104	2.3837e−02	1	101	1.2482e−01	1
2	2	2	58	4.8983e+01	1	59	4.8983e+01	1
3	2	2	49	1.3284e−01	1	41	1.3445e−01	1
4	2	3	155	−4.3311e−04	1	40	9.7615e+11	−3
5	2	3	55	−6.1080e−04	1	43	3.8450e−03	1
6	2	10	67	1.2436e+02	1	143	1.2441e+02	1
7	3	3	123	−6.3114e−04	1	104	−2.1018e−04	1
8	3	15	73	1.1097e−02	1	56	1.0903e−02	1
9	3	15	45	−9.5246e−04	1	31	−9.8771e−04	1
10	3	16	4598	8.7977e+01	1	79	6.9835e+06	1
11	3	20	48	2.8578e−01	1	33	2.8562e−01	1
12	3	20	131	4.1528e−03	1	94	2.3220e−02	1
13	4	4	95	1.0322e−01	1	95	1.6449e−02	1
14	4	6	141	6.9339e+00	1	408	3.6963e−01	1
15	4	11	50	4.4872e−03	1	40	4.1383e−03	1
16	4	20	132	8.5822e+04	1	250	8.5826e+04	1
17	5	33	60	8.7454e−01	1	21	8.7920e−01	−3
18	6	13	70	2.6871e−01	1	77	2.6769e−01	1
19	11	65	115	2.0930e+00	1	115	2.0930e+00	1
20	6	31	109	1.3171e+00	1	147	7.3389e−02	1
21	8	8	178	3.4323e+01	1	514	4.0085e−01	1
22	8	8	124	7.5034e+01	1	249	3.9143e−02	1
23	10	11	684	1.2086e+01	1	465	1.4970e−02	1
24	10	20	113	8.8654e+00	1	219	4.1323e−02	1
25	10	12	580	9.8506e+00	1	407	3.4543e+00	1
26	10	10	103	6.6826e−03	1	106	6.2733e−03	1
27	10	10	96	−4.5497e−04	1	104	−4.0186e−04	1
28	10	10	100	−2.0086e−04	1	100	−2.0086e−04	1
29	10	10	102	3.6947e−03	1	105	7.5452e−03	1
30	6	6	74	3.0424e+00	1	128	2.7303e−03	1
31	5	5	178	1.1909e−01	1	124	−5.2808e−05	1
32	6	6	88	2.1977e−01	1	72	6.1060e−04	1
33	6	6	61	1.1530e+00	1	63	1.1530e+00	1
34	6	6	61	2.6657e+00	1	65	2.6658e+00	1
35	9	9	92	2.8117e−02	1	92	2.8117e−02	1

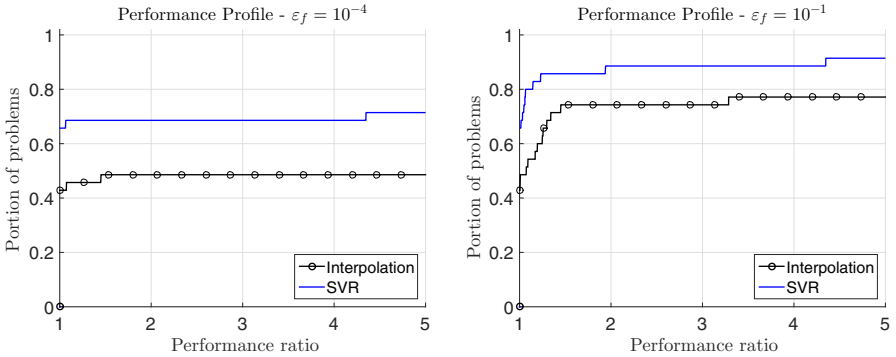

Fig. 2 Performance profiles for noisy functions

presented comparing the performance of the algorithm with the two techniques to construct the models. As expected, the tests showed that the algorithm with polynomial interpolation models is more robust and efficient for solving smooth problems. On the other hand, the algorithm with support vectors regression models is more robust and efficient for noisy problems, since they do not try to capture the noise accurately.

References

Birgin EG, Gentil JM (2012) Evaluating bound-constrained minimization software. Comput Optim Appl 53(2):347–373

Bueno LF, Friedlander A, Martínez JM, Sobral FNC (2013) Inexact restoration method for derivative-free optimization with smooth constraints. SIAM J Optim 23(2):1189–1213

Conejo PD, Karas EW, Pedroso LG (2015) A trust-region derivative-free algorithm for constrained optimization. Optim Method Softw 30(6):1126–1145

Conejo PD, Karas EW, Pedroso LG, Ribeiro AA, Sachine M (2013) Global convergence of trust-region algorithms for convex constrained minimization without derivatives. Appl Math Comput 220:324–330

Conn AR, Gould NIM, Toint PL (2000) Trust-region methods. MPS-SIAM Series on optimization. Society for industrial and applied mathematics, Philadelphia

Conn AR, Gould NIM, Vicente LN (2009a) Global convergence of general derivative-free trust-region algorithms to first and second order critical points. SIAM J Numer Anal 20:387–415

Conn AR, Scheinberg K, Toint PL (1997) On the convergence of derivative-free methods for unconstrained optimization: tributes to M.J.D. Powell. In: Buhmann MD, Iserles A (eds) Approximation theory and optimization. Cambridge University Press, Cambridge, pp 83–108

Conn AR, Scheinberg K, Toint PL (1998) A derivative free optimization algorithm in practice. In: Proceedings of the 7th AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization, St Louis, CO, USA

Conn AR, Scheinberg K, Vicente LN (2008a) Geometry of interpolation sets in derivative free optimization. Math Program 111:141–172

Conn AR, Scheinberg K, Vicente LN (2008b) Geometry of sample sets in derivative-free optimization: polynomial regression and undetermined interpolation. IMA J Numer Anal 28(4):721–748

Conn AR, Scheinberg K, Vicente LN (2009b) Global convergence of general derivative-free trust-region algorithms to first and second order critical points. SIAM J Optim 20(1):387–415

Conn AR, Scheinberg K, Vicente LN (2009c) Introduction to derivative-free optimization. MPS-SIAM series on optimization. Society for industrial and applied mathematics, Philadelphia

Dennis Jr, JE, Schnabel RB (1996) Numerical methods for unconstrained optimization and nonlinear equations (classics in applied mathematics, 16). Society for industrial and applied mathematics

Dolan ED, Moré JJ (2009) Benchmarking optimization software with performance profiles. Math Program 91(2):201–213
Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik V (1996) Support vector regression machines. In: Advances in neural information processing systems 9, NIPS, Denver, CO, USA, 2–5 December , pp 155–161
Ferreira PS, Karas EW, Sachine M (2015) A globally convergent trust-region algorithm for unconstrained derivative-free optimization. Comput Appl Math 34(3):1075–1103
Ferreira PS, Karas EW, Sachine M, Sobral FNC (2016) Global convergence of a derivative-free inexact restoration filter algorithm for nonlinear programming. Optimization, to appear
Garmanjani R, Júdice D, Vicente LN (2016) Trust-region methods without using derivatives: worst case complexity and the non-smooth case. SIAM J Optimiz 26(4):1987–2011
Grapiglia GN, Yuan J, Yuan Y (2016) A derivative-free trust-region algorithm for composite nonsmooth optimization. Comput Appl Math 35(2):475–499
Gratton S, Toint PL, Tröltzsch A (2011) An active set trust-region method for derivative-free nonlinear bound-constrained optimization. Optim Method Softw 26:873–894
Gumma EAE, Hashim MHA, Montaz Ali M (2014) A derivative-free algorithm for linearly constrained optimization problems. Comput Optim Appl 57(3):599–621
Moré JJ, Garbow BS, Hillstrom KE (1981) Testing unconstrained optimization software. ACM Trans Math Softw 7(1):17–41
Nocedal J, Wright SJ (2006) Numerical optimization. Springer series in operations research, 2nd edn. Springer, Berlin
Powell MJD (2002) UOBYQA: unconstrained optimization by quadratic approximation. Math Program 92:555–582
Powell MJD (2006) The NEWUOA software for unconstrained optimization without derivatives. In: Di Pillo G, Roma M (eds) Large-scale nonlinear optimization. Springer, New York, pp 255–297
Powell MJD (2008) Developments of NEWUOA for minimization without derivatives. IMA J Numer Anal 28:649–664
Powell MJD (2009) The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP NA2009/06, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
Powell MJD (2012) On the convergence of trust region algorithms for unconstrained minimization without derivatives. Comput Optim Appl 53(2):527–555
Powell MJD (2015) On fast trust region methods for quadratic models with linear constraints. Math Program Comput 7(3):237–267 To appear
Quarteroni A, Sacco R, Saleri F (2007) Numerical mathematics., Texts in applied mathematicsSpringer, Paris
Scheinberg K, Toint PL (2010) Self-correcting geometry in model-based algorithms for derivative-free unconstrained optimization. SIAM J Optim 20(6):3512–3532
Schölkopf B, Smola AJ (2002) Learning with kernels: support vector machines, regularization, optimization and beyond. The MIT Press, Cambridge
Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222
Vapnik V (1998) Statistical learning theory. Wiley, Hoboken