Quartic Polynomial Sub-problem Solutions in Tensor Methods for Nonconvex Optimization

* Wenqi Zhu, † Coralia Cartis

Abstract

There has been growing interest in high-order tensor methods for nonconvex optimization in machine learning as these methods provide better/optimal worst-case evaluation complexity, stability to parameter tuning, and robustness to problem conditioning. The well-known pth-order adaptive regularization (ARp) method relies crucially on repeatedly minimising a nonconvex multivariate Taylor-based polynomial sub-problem. It remains an open question to find efficient techniques to minimise such a sub-problem for $p \geq 3$.

In this paper, we propose a second-order method (SQO) for the AR3 (ARp with $p = 3$) sub-problem. SQO approximates the special-structure quartic polynomial sub-problem from above and below by using second-order models that can be minimised efficiently and globally. We prove that SQO finds a local minimiser of a quartic polynomial, but in practice, due to its construction, it can find a much lower minimum than cubic regularization approaches. This encourages us to continue our quest for algorithmic techniques that find approximately global solutions for such polynomials.

1. Motivation and Problem Set-up

The evaluation complexity of finding an approximate local minimum for a nonconvex function has been of interest for several decades [23]. Recent works [1, 5, 6, 15] have shown that some optimization algorithms have better worst-case evaluation complexity when using higher-order derivative information of the objective function together with regularization techniques.

In this paper, we consider the unconstrained nonconvex optimization problem, $\min_{x \in \mathbb{R}^n} f(x)$ where $f : \mathbb{R}^n \to \mathbb{R}$ is p-times continuously differentiable and bounded below. Since optimizing a high-dimensional function f is often challenging, we will approach the problem using a derivative-based method. Namely, we find a polynomial function $m_p(x, s)$ that approximates $f(x + s)$ at $x = x_k$. We then update x_k iteratively by $s_k := \min_{s \in \mathbb{R}^n} m_p(x_k, s)$ and $x_{k+1} := x_k + s_k$ until we reach the approximate local minimum of f, i.e. $\|\nabla f(x_k)\| \leq \epsilon_1$ and $\lambda_{\min}\left(\nabla^2 f(x_k)\right) \geq -\epsilon_2$.

The model m_p is based on Taylor expansions. The pth-order Taylor expansion of $f(x_k + s)$ at x_k is $T_p(x_k, s) := f(x_k) + \sum_{j=1}^{p} \frac{1}{j!} \nabla^j f(x_k)[s]^j$, where $\nabla^j f(x_k) \in \mathbb{R}^{n^j}$ is a jth-order tensor and $\nabla^j f(x_k)[s]^j$ is the jth-order derivative of f at x_k along $s \in \mathbb{R}^n$. To ensure the method converges globally from an arbitrary starting point to first/second-order critical points, we add an adaptive $(p + 1)$-power regularization term to T_p. This gives us the pth-order regularized Taylor model,

$$m_p(x_k, s) = T_p(x_k, s) + \frac{\sigma_k}{p+1} \|s\|_2^{p+1} \tag{1.1}$$

where $\sigma_k > 0$ is adjusted adaptively to ensure progress towards optimality over the iterations. The case of $p = 1$ gives the steepest descent model and the case of $p = 2$ gives a Newton-like model; in this paper, we focus on the case of $p \geq 3$. This construction correspond to the well-known adaptive regularization algorithmic framework ARp [1, 5, 6]. Under Lipschitz continuity assumptions on

* wenqi.zhu@maths.ox.ac.uk, University of Oxford, UK
† cartis@maths.ox.ac.uk, University of Oxford, UK

$\nabla^p f(x)$, ARp methods require no more than $\mathcal{O}\left(\epsilon^{-\frac{p+1}{p-q+1}}\right)$ evaluations of f and its derivatives to compute a q-th order minimizer to an accuracy of ϵ, where $q \in \{1, 2\}$. More details can be found in Appendix A and in [1, 5, 7, Ch 4].

These results show that to achieve the same criticality conditions, the evaluation complexity bound improves as we increase the order p. Consequently, an efficient high-order method (such as ARp) is crucial for a fast and accurate algorithm that optimizes high-dimensional nonconvex functions. Moreover, [2] proves that ARp has the optimal complexity for the pth-order methods. In the case of $p = 2$, we obtain the AR2 method (also known as *Adaptive Cubic regularization, ARC*[1]). AR2 requires at most $\mathcal{O}\left(\max\{\epsilon_1^{-3/2}, \epsilon_2^{-3}\}\right)$ evaluations of f and its derivatives to find the local minimum [3, 4]. Under a similar condition, many trust-region (TR) algorithms require at most $\mathcal{O}\left(\max\{\epsilon_1^{-2}, \epsilon_2^{-3}\}\right)$ evaluations. For instance, in the case of $p = 3$, the worst case evaluation complexity for the AR3 model is $\mathcal{O}\left(\max\{\epsilon_1^{-4/3}, \epsilon_2^{-2}\}\right)$. This means that the AR3 model achieves a better worst-case performance than methods that use only first/second derivatives. The superior theoretical complexity result of AR3 motivates us to find a algorithmic implement of it. Note that these complexities do not include the cost of minimising the sub-problem (1.1).

In practice, an efficient ARp algorithm relies critically on iteratively minimising (1.1). AR2 is the first non-trivial problem in the family that requires an efficient solver for the sub-problem. The sub-problem m_2 has been widely researched as part of the (adaptive) cubic regularization framework [3, 8, 10, 13, 19]. There are many scalable methods for solving the AR2 sub-problems [3, 9, 11, 12]. Specifically, [3] gives an iterative algorithm for finding the global minimizer of m_2 in both convex and nonconvex cases.

In the case of $p = 3$, m_3 is generally a nonconvex quartic regularized multivariate polynomial with respect to s. How to efficiently minimize m_3 remains an open question and will be the main topic of our paper. Following [19], Nesterov has proposed a series of second-order methods for minimising a convex m_3 model. Under the convexity assumption, the additional relative smoothness properties of the third-order tensor models allow the use of Bregman gradient methods to approximately minimise the m_3 model [15]. The first results of this type were based on the difference approximation of the third-order tensor term [17], while the other papers [14, 16] employ the framework of the high-order proximal-point operators. More recently, Nesterov gave a linearly convergent second-order method for minimizing convex quartic polynomials [18]. In this method, he estimated the third-order tensor term by a combination of the second and the fourth derivatives. In this paper, we extend this idea to nonconvex functions and nonconvex AR3 sub-problems.

Minimising the AR2 sub-problem: We first review a generalized form of the m_2 model and an algorithm to minimize it. In the rest of the paper, let x_k be fixed and we drop the k subscript. Consider the second-order model with rth-power of regularization,

$$m_2^r(s) = \tilde{f}_0 + \tilde{g}^T[s] + \frac{1}{2}\tilde{H}[s]^2 + \frac{1}{r}\sigma\|s\|_2^r, \tag{1.2}$$

where $\tilde{f}_0$, $\tilde{g}$ and $\tilde{H}$ denote functions and derivative values at any given point $\tilde{x}$. Clearly, when $r = 3$, (1.2) gives the ARC sub-problem. When $r = 4$, $m_2^4(s) = \tilde{f}_0 + \tilde{g}^T[s] + \frac{1}{2}\tilde{H}[s]^2 + \frac{1}{4}\sigma\|s\|_2^4$. Notice that m_3, the model that we are focused on, has the same regularization power as $m_2^4(s)$, but $m_2^4(s)$ does not have the third-order tensor term. The third derivative of a function is not an independent characteristic. It can be estimated by a combination of the second and the fourth derivatives.

1. In this paper, we also refer AR2 as ARC, ARC = AR2.

The idea of our proposed method, SQO, first bounds the third-order tensor term of $m_3(s)$ using a second-order term and quartic regularization term. Then, SQO minimizes $m_3(s)$ using a sequence of second-order models with quartic regularization (in the form of $m_2^4(s)$). Theorem 1 summarizes the necessary and sufficient optimality conditions for the global minimizer of $m_2^4(s)$.

Theorem 1 *(Theorem 8.2.8 [7]) Let $r \geq 3$, any global minimizer of $m_2^r(s)$, s_*, satisfies $(\tilde{H} + \lambda_* I_n)s_* = -\tilde{g}$, where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, $\lambda_* \geq 0$, $\tilde{H} + \lambda_* I_n \succeq 0$, and $\lambda_* = \sigma\|s_*\|^{r-2}$. If $\tilde{H} + \lambda_* I_n$ is positive definite, then s_* is unique.*

Theorem 1 converts the problem of finding the global minimum of m_2^r into a problem of solving a nonlinear equation paired with a matrix system. Broadly speaking, we use the matrix system $(\tilde{H} + \lambda_* I_n)s_* = -\tilde{g}$ to express s^* as a function of λ_*, such that $s^* = s(\lambda_*)$. The nonlinear equation is re-written as $\lambda_* = \sigma\|s(\lambda_*)\|_2^{r-2}$. Due to non-linearity, λ_* does not have an explicit expression in general, but we can obtain an approximate solution via the Newton method for root finding. More details of the algorithm can be found Ch. 8 [7] where a literature survey of various approaches is also given.

2. Minimising the AR3 Sub-problem via Sequential Quadratic Optimization (SQO)

SQO is an iterative algorithm that solves the AR3 sub-problem by generating $\{s^{(i)}\}_{i\geq 0}$. We prove that $s^{(i+1)}$ converges to s_k which is a local minimizer of $m_3(x_k, s)$ for every major iteration k of AR3. Since SQO solves the sub-problem with k fixed, we drop the k subscript and write the third-order model with quartic regularization as $m_3(s) = f_0 + g^T[s] + \frac{1}{2}H[s]^2 + \frac{1}{6}T[s]^3 + \frac{1}{4}\sigma\|s\|_2^4$, where $f_0 = f(x_k)$, $g = \nabla f(x_k)$, $H = \nabla^2 f(x_k)$, and $T = \nabla^3 f(x_k)$. The 4th-order Taylor expansion of $m_3(s^{(i)} + s)$ at $s^{(i)}$ is

$$M(s^{(i)}, s) := m_3(s^{(i)}) + \nabla_s m_3(s^{(i)})^T s + \frac{1}{2}s^T \nabla_s^2 m_3(s^{(i)})s + \frac{1}{6}\nabla_s^3 m_3(s^{(i)})[s]^3 + \frac{\sigma}{4}\|s\|_2^4.$$

Since $m_3(s)$ is a 4th-degree multivariate polynomial, the Taylor expansion is exact, such that $M(s^{(i)}, s) = m_3(s + s^{(i)})$, and $\min_{s\in\mathbb{R}^n} m_3(s) = \min_{s\in\mathbb{R}^n} M(s^{(i)}, s)$.

The key idea for SQO is that we use a quadratic term with a quartic regularization term to approximate the third-order tensor term, such that

$$s^T H_-(s^{(i)})s + d_-\|s\|_2^4 \leq \frac{1}{6}\nabla_s^3 m_3(s^{(i)})[s]^3 \leq s^T H_+(s^{(i)})s + d_+\|s\|^4, \tag{2.1}$$

where $H_\pm \in \mathbb{R}^{n\times n}$, $d_\pm \in \mathbb{R}$ are specifically chosen and details are in Section 2.2. Let

$$M_\pm(s^{(i)}, s) := m_3(s^{(i)}) + \nabla_s m_3(s^{(i)})[s] + \left[\frac{1}{2}\nabla_s^2 m_3(s^{(i)}) + H_\pm(s^{(i)})\right][s]^2 + \left(\frac{\sigma}{4} + d_\pm\right)\|s\|_2^4,$$

and $d_\pm \geq 0$, then we can obtain a bound for m_3, such that $M_-(s^{(i)}, s) \leq m_3(s) = M(s^{(i)}, s) \leq M_+(s^{(i)}, s)$. Moreover, we can deduce that

$$\min_{s\in\mathbb{R}^n} M_-(s^{(i)}, s) \leq \min_{s\in\mathbb{R}^n} m_3(s) \leq \min_{s\in\mathbb{R}^n} M_+(s^{(i)}, s). \tag{2.2}$$

It is worth noting that $M_\pm$ is a model with quadratic order and quartic regularisation. $M_\pm$ has no third-order tensor term and is of the same form as m_2^4 model in (1.2). As explained in the previous section, we have an algorithm to find the global minimum for problems in the form of the m_2^4 model. SQO uses this algorithm and optimizes the regularized quartic polynomial model (AR3) by a sequence of second-order model with quartic regularization.

2.1. Algorithms and Main Theorem

SQO (Algorithm 1) allows for both convex and nonconvex polynomial/iterations. The convergence result and complexity analysis are given in Theorem 2 and in its remark.

Algorithm 1 Sequential Quadratic Optimization (**SQO**)

Initialization: Choose $s^{(0)} = \mathbf{0}$ and compute $H_+(s^{(0)}) \in \mathbb{R}^n$ and $d_+ \in \mathbb{R}$.
ith iteration ($i \geq 0$): If $\|\nabla_s m(s^{(i)})\|_2 > \texttt{TOL}$ or $\|\nabla_s^2 m(s^{(i)})\|_2 < -\texttt{TOL}$,
$s^{(i+1)} = \text{argmin}_{s \in \mathbb{R}^n} M_+(s^{(i)}, s)$.

Figure 1 is a simple illustration of SQO for $n = 1$. Note that M_- is bounded below on each iteration[2]. SQO gives a strict descent in $m_3(s^{(i)})$ in each iteration, such that $m_3(s^{(0)}) > m_3(s^{(1)}) > \ldots$ In nonconvex iterations, SQO usually gives a particularly large descent in $m_3(s^{(i)})$. We give an explanation for these two properties of SQO in Theorem 2.

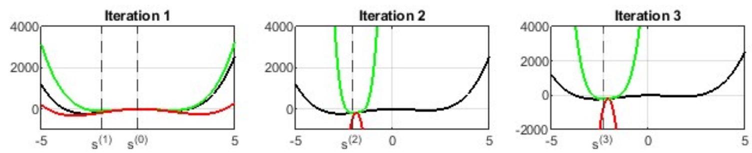


Figure 1: Minimizing $10s - 50s^2 + 5s^3 + 5s^4$. m_3, M_+, M_- plotted in black, green and red respectively.

Theorem 2 *Assume M_+ is constructed as described in Section 2.2 and $s^{(i)}$ is computed by Algorithm 1. Then, $m_3(s^{(i)})$ is strictly decreasing with i and converges to a local minimum of $m_3(s)$. In non-convex iterations, if we construct M_+ using hB-SQO and assume that $\|\nabla_s m(s^{(i)})\|_2 > \epsilon$, $\lambda_{\min}\left(\nabla_s^2 m(s^{(i)})\right) < -\epsilon$ and $\lambda_{\max}\left(\nabla_s^2 m(s^{(i)})\right) \neq 0$. Then, the iterations satisfy the complexity bound*

$$m_3(s^{(i)}) - m_* \leq \left(m_3(s^{(0)}) - m_*\right)\left(1 - (i+1)\hat{c}\epsilon^{4/3}\right)$$

where $m_* = \min_{s \in \mathbb{R}^n} m_3(s)$. $\hat{c}$ is a problem-dependent positive constant that depends on σ, $\lambda_{\min}\left(\nabla_s^2 m(s^{(i)})\right)$, $\lambda_{\max}\left(\nabla_s^2 m(s^{(i)})\right)$, and $\max_{u \in \mathbb{R}^n, \|u\|_2 = 1} \left|\nabla_s^3 m_3(s^{(i)})[u]^3\right|$.

The complexity result for convex iterations is given next and a sketch of proof is given in Appendix B.

Remark on Convex Iterations: [18] proves the result for convex quartic regularized polynomials. In the convex case, each iteration satisfies $m_3(s^{(i)}) - m_* \leq \left(m_3(s^{(0)}) - m_*\right)(1 - \alpha)^k$ where $\alpha > 0.193$.

2.2. Upper and Lower bounds

We introduce two types of upper and lower bounds for $m_3(s)$ which use the information from gradient and from Hessian, respectively. Let L_g and L_T be the (local) Lipschitz constant for $m_3(s)$. The *Gradient Bound SQO (gB-SQO)* is

$$M_+(s^{(i)}, s) = m_3(s^{(i)}) + \nabla_s m_3(s^{(i)})^T s + s^T \left[\left(\frac{1}{2} - \frac{1}{3\tau}\right)\nabla_s^2 m_3(s^{(i)}) + \frac{L_g}{3\tau} I_n\right] s + \left(\frac{\sigma}{4} + \frac{\tau L_T}{12}\right)\|s\|_2^4$$

2. The lower bound of M_- is -2×10^7 which is outside the range of the plot for the 2nd and 3rd iteration.

where $\tau > 1$ is a constant. Let $\chi_1 := -\lambda_{\min}[\nabla_s^2 m_3(s^{(i)})]$. The *Hessian Bound SQO (hB-SQO) is*

$$M_+(s^{(i)}, s) = m_3(s^{(i)}) + \nabla_s m_3(s^{(i)})^T s + s^T \left[\frac{1}{2}\nabla_s^2 m_3(s^{(i)}) + a_+ \chi_1 I_n\right] s + \left(\frac{\sigma}{4} + b_+\right) \|s\|_2^4$$

where a_+ and b_+ are constants depending on χ_1, σ and $\lambda_{\max}[\nabla_s^3 m_3(s^{(i)})]$.

2.3. Preliminary Numerical Illustrations

Numerical Set-up: The test quartic regularised polynomials are nonconvex. $H \in \mathbb{R}^{n\times n}$ is a symmetric, normally distributed matrix. $T \in \mathbb{R}^{n\times n\times n}$ is a symmetric, uniformly distributed third-order tensor. The stopping criteria is $\|\nabla_s m(s^{(i)})\|_2 < \texttt{TOL}$ and $\lambda_{\min}\left(\nabla_s^2 m(s^{(i)})\right) > -\texttt{TOL}$ where $\texttt{TOL} = 10^{-5}$ unless otherwise stated.

SQO vs other solvers: (Table 1) SQO performs much better than MATLAB `fminsearch` (simplex search) and MATLAB `fminunc` (quasi-newton) in terms of functions and derivatives evaluation count and the size of the local minimum found. SQO finds a local minimum for $n = 2$ to $n = 300$ while `fminsearch` fails for $n > 5$ and `fminunc` produces inaccurate results for $n > 20$. SQO also locates a lower local minimum than ARC for large-size problems ($n > 50$). But it takes more function evaluations than ARC.

Stability and Accuracy of SQO: (Figure 2) SQO finds a highly accurate local minimum, giving a tolerance of 10^{-10} for problem size $n = 50$. SQO is stable for ill-conditioned quartic regularized problems (i.e. m_3 with ill-conditioned Hessian and random uniformly distributed tensor).

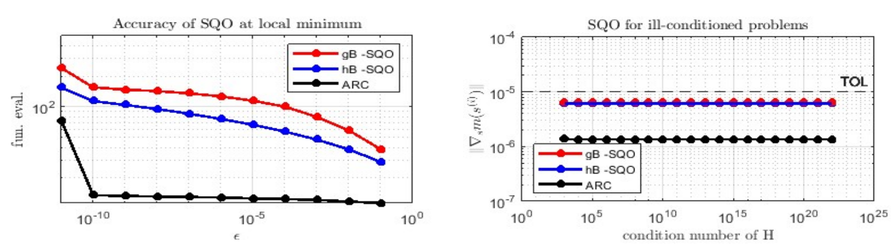
gB-SQO vs hB-SQO: The two solvers require similar function evaluations in finding a local minimum. gB-SQO usually produces a lower local minimum than hB-SQO (Table 1). Both methods guarantees function decrease in every iteration (Appendix C).

Table 1: Comparison of SQO and other solvers

	Minimum values attained						**Function evaluation counts**					
Dimension (n)	2	20	50	75	100	300	2	20	50	75	100	300
`fminsearch`	-3.9	-	-	-	-	-	115	-	-	-	-	-
`fminunc`	-3.9	-439	-	-	-	-	33	733*	-	-	-	-
ARC	-3.9	-439	$-1.11*10^5$	$-8.3*10^5$	$-6.9*10^6$	$-1.5*10^{10}$	8	13	13	14	15	19
gB-SQO	-3.9	-439	$-1.14*10^5$	$-1.0*10^6$	$-1.1*10^7$	$-1.8*10^{10}$	20	43	60	85	77	97
hB-SQO	-3.9	-439	$-1.13*10^5$	$-8.4*10^5$	$-8.1*10^6$	$-1.8*10^{10}$	20	44	57	68	76	79

$-$ represents max.iteration (> 5000) exceeded. Results avg. over 75 problems. Tolerance not achieved, $^*\|\nabla_s m_3(s^{(i)})\| = 2*10^{-4}$.

Figure 2: Stability and accuracy analysis



Results avg. over 25 problems. At $\texttt{TOL} = 10^{-11}$, one problem exceed max. iter for all three solvers.

3. Extensions and Conclusion

In this paper, we proposed a second-order method (SQO) that minimizes a nonconvex quartic regularized polynomial (i.e. the AR3 sub-problem) to high accuracy and high numerical stability. Our preliminary numerical results show that SQO can locate a much lower minimum than cubic regularisation approaches. There is a vast literature in multivariate polynomial optimization that explores global descent algorithms [20–22]. In future work, we will explore the properties of the tensor term at a local minimum and extend SQO to provably find the approximate global minimizer for AR3.

Acknowledgments: This work was supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

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] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021.

[3] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[4] Coralia Cartis, Nicholas IM Gould, and Ph L Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012.

[5] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. *SIAM Journal on Optimization*, 30(1):513–541, 2020.

[6] Coralia Cartis, Nick IM Gould, and Ph L Toint. A concise second-order complexity analysis for unconstrained optimization using high-order regularized models. *Optimization Methods and Software*, 35(2):243–256, 2020.

[7] Coralia Cartis, Nicholas Ian Mark Gould, and Ph L Toint. Evaluation complexity of algorithms for nonconvex optimization. *MOS-SIAM Series on Optimization.*, 2022.

[8] Jean-Pierre Dussault. Arcq: a new adaptive regularization by cubics. *Optimization Methods and Software*, 33(2):322–335, 2018.

[9] Yong Hsia, Ruey-Lin Sheu, and Ya-xiang Yuan. Theory and application of p-regularized subproblems for p¿ 2. *Optimization Methods and Software*, 32(5):1059–1077, 2017.

[10] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR, 2017.

[11] Stefano Lucidi, Laura Palagi, and Massimo Roma. On some properties of quadratic programs with a convex quadratic constraint. *SIAM Journal on Optimization*, 8(1):105–122, 1998.

[12] José Mario Martínez. Local minimizers of quadratic functions on euclidean balls and spheres. *SIAM Journal on Optimization*, 4 (1):159–176, 1994.

[13] José Mario Martínez and Marcos Raydan. Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization. *Journal of Global Optimization*, 68(2):367–385, 2017.

[14] Yurii Nesterov. Inexact accelerated high-order proximal-point methods. Technical report, 2020.

[15] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186(1):157–183, 2021.

[16] Yurii Nesterov. Inexact high-order proximal-point methods with auxiliary search procedure. *SIAM Journal on Optimization*, 31 (4):2807–2828, 2021.

[17] Yurii Nesterov. Superfast second-order methods for unconstrained convex optimization. *Journal of Optimization Theory and Applications*, 191(1):1–30, 2021.

[18] Yurii Nesterov. Quartic regularity. *arXiv preprint arXiv:2201.04852*, 2022.

[19] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[20] Liqun Qi. Extrema of a real polynomial. *Journal of Global Optimization*, 30(4):405–433, 2004.

[21] Liqun Qi and Kok Lay Teo. Multivariate polynomial minimization and its application in signal processing. *Journal of Global Optimization*, 26(4):419–433, 2003.

[22] Liqun Qi, Zhong Wan, and Yu-Fei Yang. Global minimization of normal quartic polynomials based on global descent directions. *SIAM Journal on Optimization*, 15(1):275–302, 2004.

[23] Stephen A Vavasis. Black-box complexity of local minimization. *SIAM Journal on Optimization*, 3(1):60–80, 1993.

Appendix A. Complexity Theory and ARp Algorithm

Assumption: $f : \mathbb{R}^n \to \mathbb{R}$ is p-times continuously differentiable and bounded below f_{low}. The pth-derivative of f is globally Lipschitz continuous. That is, there exists a constant $L_p \geq 0$, such that $\|\nabla^p f(x) - \nabla^p f(y)\| \leq (p-1)! L_p \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$.

Theorem 3 *(Theorem 3.6 [6]) Let $p \geq 2$, $f \in C^{p,1}(\mathbb{R}^n)$. Then, AR$p$ requires at most*

$$\left\lceil \kappa_{1,2} \cdot (f(x_0) - f_{\text{low}}) \cdot \max \left[\epsilon_1^{-\frac{p+1}{p}}, \epsilon_2^{-\frac{p+1}{p-1}} \right] + \kappa_{1,2} \right\rceil \tag{A.1}$$

function and derivative evaluations to reach the approximate local minimum with first and second-order criticality as $\|\nabla f(x_k)\|_2 \leq \epsilon_1$, $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_2$.

Algorithm 2 Adaptive pth-order Regularization Model: ARp

Compute s_k such that $m_p(s_k) < f(x_k)$, $\|\nabla_s m_p(s_k)\|_2 \leq \theta_1 \|s_k\|_2^p$ and $\lambda_{\min}(\nabla^2 m_p(s_k)) \geq -\theta_2 \|s_k\|_2^{p-1}$.
Compute $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{f(x_k) - T_p(x_k, s_k)}$.
Update x_k**:** $x_{k+1} := x_k + s_k$ if $\rho_k > \eta = 0.1$, or $x_{k+1} := x_k$ otherwise.
Update σ_k**:** $\sigma_{k+1} = \frac{\sigma_k}{\gamma_1} = 2\sigma_k$ when $\rho_k < \eta$; else $\sigma_{k+1} = \max\{\gamma_2 \sigma_k, \sigma_{\min}\} = \max\{\frac{1}{2}\sigma_k, \sigma_{\min}\}$.

Appendix B. Sketch of Proof for Theorem 2

Sketch of proof: The key step is that we use M_- to safeguard the function value monotonically decreases. Using M_-, we form an *adjustable nonconvex bound* for m_3,

$$M_{-,\alpha}(s^{(i)}) := m_3(s^{(i)}) + \nabla_s m_3(s^{(i)})^T s + \alpha H_-(s^{(i)})[s]^2 + \alpha^{-3} \left(\frac{\sigma}{4} - d_- \right) \|s\|_2^4,$$

where $0 < \alpha < 1$. For a fixed $s^{(i)}$, $M_{-,\alpha}$ increases as $\alpha \to 0$. We prove that there exists a $\hat{\alpha} > 0$ such that $\min_{s \in \mathbb{R}^n} M_-(s^{(i)}) < \min_{s \in \mathbb{R}^n} m_3(s) < \min_{s \in \mathbb{R}^n} M_+(s^{(i)}) < \min_{s \in \mathbb{R}^n} M_{-,\hat{\alpha}}(s^{(i)})$. Technical analysis on $\hat{\alpha}$, $H_\pm$ and $d_\pm$ gives the complexity bound. $\square$

Appendix C. Contraction Factor and Rate of Convergence for SQO

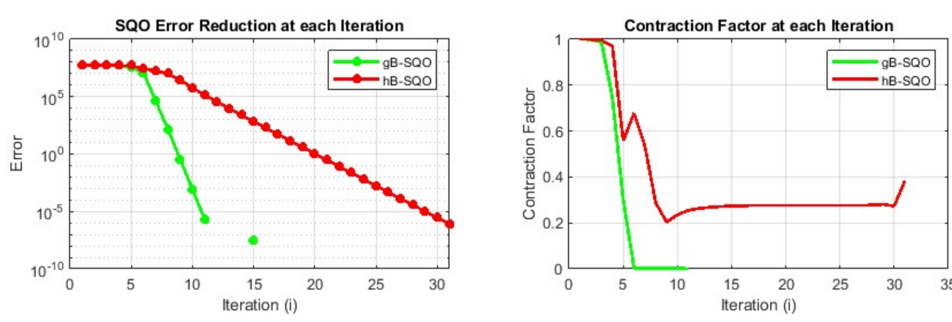


Figure 3: The error is calculated by $e_i := m_3(s^{(i)}) - m_*$. The contraction factor is e_{i+1}/e_i. Problem size is $n = 100$.