An adaptive regularization method in Banach spaces

Gratton, S.; Jerad, S.; Toint, Ph L.

Published in:
Optimization Methods and Software

DOI:
10.1080/10556788.2023.2210253

Publication date:
2023

Document Version
Early version, also known as pre-print

An Adaptive Regularization Method in Banach Spaces

S. Gratton*, S. Jerad† and Ph. L. Toint‡

4 X 2021

Abstract

This paper considers optimization of nonconvex functionals in smooth infinite dimensional spaces. It is first proved that functionals in a class containing multivariate polynomials augmented with a sufficiently smooth regularization can be minimized by a simple linesearch-based algorithm. Sufficient smoothness depends on gradients satisfying a novel two-terms generalized Lipschitz condition. A first-order adaptive regularization method applicable to functionals with β-Hölder continuous derivatives is then proposed, that uses the linesearch approach to compute a suitable trial step. It is shown to find an ϵ-approximate first-order point in at most $\mathcal{O}(\epsilon^{-\frac{p+\beta}{p+\beta-1}})$ evaluations of the functional and its first p derivatives.

Keywords: nonlinear optimization, adaptive regularization, evaluation complexity, Hölder gradients, infinite-dimensional problems.

1 Introduction

The analysis of adaptive regularization (AR) algorithms for nonlinear (and potentially nonconvex) optimization has been a very active field in recent years (see [19, 24, 7, 8, 10, 4, 17, 5, 6, 23, 18, 3, 2, 13], to cite only a few). This sustained interest of the research community is motivated in part by the fact that these methods not only work well in practice, but also exhibit excellent worst-case evaluation complexity bounds: one can indeed prove that the number of function and derivatives evaluations which may be required to find an approximate critical point is small, at least compared to similar bounds for other standard methods such as linesearch-based Newton or trust-region algorithms [24, 8]. As it turns out, evaluation complexity results obtained for AR methods and nonconvex problems have been obtained, to the best of the authors' knowledge, in the context of $\mathbb{R}^n$. It is the purpose of this short note to show that this need not be the case, and that evaluation complexity bounds for computing approximate first-order critical point can be derived in infinite-dimensional Banach spaces.

The main motivation for this generalization is twofold. Our first aim is to cover a number of infinite-dimensional applications in optimal control and variational analysis, and show that adaptive regularization methods do make sense in that context. Indeed, our developement

*Université de Toulouse, INP, IRIT, Toulouse, France. Email: serge.gratton@enseeiht.fr. Work partially supported by 3IA Artificial and Natural Intelligence Toulouse Institute, French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004"

†ANITI, Université de Toulouse, INP, IRIT, Toulouse, France. Email: sadok.jerad@enseeiht.fr

‡NAXYS, University of Namur, Namur, Belgium. Email: philippe.toint@unamur.be

covers optimization problems in $L^p(\mathbb{R}^n)$, ℓ^p and Sobolev spaces $W^{m,p}(\mathbb{R}^n)$ [29] for $p \in (1, \infty)$ as well as in all Hilbert spaces.

Our second aim is to investigate the necessary methodological coherence when optimization algorithms are applied to large-scale discretized problems: it is then important to show that AR methods continue to make sense in the limit, as the discretization mesh converges to zero. This coherence, sometimes called "mesh independence", has long been considered as an important feature of numerical optimization methods [22, 1, 16, 20, 27]. For trust-region methods, this was studied in [26] in the Hilbert space context, and developed for Hilbert and Banach spaces in [15, Section 8.3]. Considering the question for AR algorithms therefore seems a natural development in this line of research. One might argue that most evaluation complexity results are "dimensionless", making this effort unnecessary. This argument however ignores an important point: problems in infinite dimension (and thus their discretizations) are often defined in spaces whose norms (and inner products when they exist) are not the standard Euclidean one. As a consequence, gradients must be measured in dual norms and thus approximate first-order points detected using these norms. This makes most existing complexity results applicable only though the use of norm-equivalence constants in large-scale finite dimensional approximations, whose value may significantly increase with dimension. The complexity estimates obtained using this approach can thus be severe overestimates for large-dimensional discretizations of infinite-dimensional variational problems. Considering the norm adapted to the problem may therefore provide substantially more robust evaluation complexity bounds, which is the point of view developed in this paper.

Our second objective however raises specific technical difficulties. While the outline of adaptive regularization methods is today quite well-known for finite dimensional spaces (see [6], for instance), its simple generalization to infinite dimensions is impossible. Indeed, the existence of a suitable step at a given iteration of the method in finite dimensions typically hinges on approaching a minimizer of a regularized model, which may no longer exist in infinite dimensions. Our analysis circumvents that problem by proposing a specialized optimization technique which guarantees an acceptable step for a class of function that, at variance with existing Lipschitz approaches, includes regularized polynomials.

Contributions. Having set the scene, we now make our contribution more precise.

- We first analyse the convergence of a first-order method for minimizing a regularized differentiable functions on a bounded set, where the first-order approximation error for the objective function's and the regularization's gradients satisfy a two-terms generalized Hölder condition. Significantly, this class includes regularized multivariate polynomials. To our knowledge, no such regularization has been considered before, even in finite dimensional spaces.

- Exploiting this result, we then propose an adaptive regularization algorithm whose step is found by minimizing a regularized polynomial and whose objective is to find first-order points of nonconvex functions having Hölder continuous p-th derivative (in the Fréchet sense). We analyze its evaluation complexity and show that the sharp complexity bound known [11] for the finite-dimensional case is recovered, in that the algorithm requires at most $\mathcal{O}\left(\epsilon^{-\frac{p+\beta}{p+\beta-1}}\right)$ evaluations of the function and its first p derivatives to compute such a point.

Outline. The paper is organized as follows. Section 2 considers the minimization of smooth regularized functionals in Banach spaces. Section 3 then introduces the class of Banach spaces of interest and details our general adaptive regularization algorithm for first-order minimization in these spaces, while Section 4 analyzes its evaluation complexity. We conclude the paper in Section 5 with a brief discussion of the new results and perspectives.

Notation Throughout the paper, $\|.\|_{\mathcal{V}}$ denotes the norm over the space $\mathcal{V}$. $\mathcal{B}(x,B)$ denotes the open ball centered at x of radius B. $\mathcal{L}(\mathcal{V}^{\otimes m};\mathbb{R})$ denotes the space of multilinear continuous functionals from $\mathcal{V}\times\mathcal{V}\cdots\times\mathcal{V}$ to $\mathbb{R}$ and $\mathcal{L}^m_{sym}(\mathcal{V}^{\otimes m};\mathbb{R})$ the subspace of $\mathcal{L}^m(\mathcal{V}^{\otimes m};\mathbb{R})$ that is m-linear symmetric. For a functional f defined from $\mathcal{V}$ to $\mathbb{R}$ that is p times Fréchet differentiable, $\nabla^k_x f(x)\in\mathcal{L}^k_{sym}(\mathcal{V}^{\otimes k};\mathbb{R})$ denotes the k-th derivative tensor for $k\in\{1,\ldots,p\}$. $\nabla^1_x f$ is an element of the dual space of $\mathcal{V}$ denoted $\mathcal{V}'$. The symbol $\langle\cdot,\cdot\rangle$ denotes the dual pairing between $\mathcal{V}$ and $\mathcal{V}'$, that is $\langle y,x\rangle \stackrel{\rm def}{=} y(x)$, for $y\in\mathcal{V}'$ and $x\in\mathcal{V}$. The norm in the dual space $\mathcal{V}'$ will be denoted as $\|.\|_{\mathcal{V}'}$. For $S\in\mathcal{L}^m_{sym}$, $S[v_1,v_2\ldots,v_m]\in\mathbb{R}$ denotes the result of applying S to $v_1,\ldots,v_m$. $S[v]^m$ is the result of applying S to m copies of v and $S[v]^l\in\mathcal{L}^{m-l}_{sym}(\mathcal{V}^{\otimes m-l};\mathbb{R})$ the result of applying it to l copies of v. We define the norm in $\mathcal{L}^m_{sym}(\mathcal{V}^{\otimes m};\mathbb{R})$ as

$$\|S\| \stackrel{\rm def}{=} \sup_{\|v_1\|_{\mathcal{V}}=\cdots=\|v_m\|_{\mathcal{V}}=1} |S[v_1,\ldots,v_m]|. \tag{1.1}$$

2 Gradient descent with a Hölder regularization

We start by considering the minimization, for x in the Banach space $\mathcal{V}$, of the regularized objective functional

$$\phi(x) \stackrel{\rm def}{=} \psi(x)+h(x), \tag{2.1}$$

where h is a general regularization term. This is motivated by the need to replace the problematic condition that the step of our yet to be defined regularization method is close to a minimizer by some more appropriate condition for infinite dimensional spaces, where ψ will play the role of the regularized model.

The space $\mathcal{V}$ and the functionals ϕ, ψ and h in (2.1) are assumed to satisfy the following properties.

AS.1

(i) There exists $\phi_{\min}\in\mathbb{R}$ such that, for all $x\in\mathcal{V}$, $\phi(x)\geq\phi_{\min}$. Moreover, the set $\mathcal{D}\stackrel{\rm def}{=}\{x\in\mathcal{V}\,,\,\phi(x)\leq\phi(0)\}$ is bounded in the sense that $\sup_{x\in\mathcal{D}}\|x\|_{\mathcal{V}}\leq\omega$ for some $\omega<\infty$.

(ii) ψ is a Fréchet differentiable function that satisfies the local two-terms Hölder condition

$$\forall\delta>0,\ \forall x\in\mathcal{B}(0,\delta),\ \forall y\in\mathcal{V},\ \|\nabla^1_x\psi(x)-\nabla^1_x\psi(y)\|_{\mathcal{V}'}\leq L_{1,\delta}\|x-y\|^{\beta_1}_{\mathcal{V}}+L_{2,\delta}\|x-y\|^{\beta_2}_{\mathcal{V}},$$

where $\beta_1>0$ and $\beta_2>0$, $L_{1,\delta}>0$ and $L_{2,\delta}>0$ are constants, the latter two depending on δ.

(iii) h is a convex Fréchet differentiable function whose gradient satisfies the local two-terms Hölder condition

$$\forall\delta>0,\ \forall x\in\mathcal{B}(0,\delta),\ \forall y\in\mathcal{V},\ \|\nabla^1_x h(x)-\nabla^1_x h(y)\|_{\mathcal{V}'}\leq L_{3,\delta}\|x-y\|^{\beta_3}_{\mathcal{V}}+L_{4,\delta}\|x-y\|^{\beta_4}_{\mathcal{V}},$$

where $\beta_3 > 0$ and $\beta_4 > 0$, $L_{3,\delta} > 0$ and $L_{4,\delta} > 0$ are constants, the latter two depending on δ. Moreover, $\min[\beta_3, \beta_4] \leq 1$

(iv) the space $\mathcal{V}$ is reflexive.

The conditions stated in **AS.1**(ii) and (iii) are verified by functionals with Hölder continuous gradient as proven in [13, Lemma 2.1] (β_1 is then equal to the Hölder exponent and $L_{1,\delta}$ equal to the Hölder constant). We use the more slightly more general conditions of **AS.1**(ii) in order to widen the class of allowed functionals and, in particular, to cover multivariate polynomials. Observe also that, should β_3 and β_4 both exceed one, then h must be affine and, since we do not exclude an affine ψ, **AS.1**(i) could then be violated. This potential contradiction justifies our assumption that $\min[\beta_3, \beta_4] \leq 1$.

The conditions stated in **AS.1**(ii) (for ψ) and (iii) (for h) are identical, and they obvioulsy combine to yield that

$$\forall \delta > 0,\ \forall x \in \mathcal{B}(0,\delta),\ \forall y \in \mathcal{V},\ \|\nabla_x^1 \phi(x) - \nabla_x^1 \phi(y)\|_{\mathcal{V}'} \leq L'_{1,\delta} \|x-y\|_{\mathcal{V}}^{\alpha_1} + L'_{2,\delta} \|x-y\|_{\mathcal{V}}^{\alpha_2}, \quad (2.2)$$

where $\alpha_1 = \min(\beta_1, \beta_2, \beta_3, \beta_4) \leq 1$, $\alpha_2 = \max(\beta_1, \beta_2, \beta_3, \beta_4)$ and $L'_{1,\delta} = L'_{2,\delta} = \sum_{i=1}^4 L_{i,\delta}$. We could clearly have assumed this condition on the gradient of ϕ directly, but we have preferred separate statements because **AS.1**(ii) and (iii) will be proved separately for the functionals of interest. We immediately verify that multivariate polynomials satisfy **AS.1**(ii). This result is crucial for our purposes, as it will allow us to compute a step in the ARp-BS algorithm defined below.

Lemma 2.1 Consider a multivariate polynomial functional $\psi : \mathcal{V} \to \mathbb{R}$ given, for $x \in \mathcal{V}$, by

$$\psi(x) = \psi_0 + \sum_{\ell=1}^{p} \frac{1}{l!} S_\ell[x]^\ell, \quad (2.3)$$

where $S_\ell \in \mathcal{L}^\ell_{sym}(\mathcal{V}^{\otimes \ell})$ for $\ell \in \{1, \ldots, p\}$. Then, ψ satisfies **AS.1**(ii).

Proof. First observe that $\nabla_x^1 \psi(x) = \sum_{\ell=1}^p \frac{1}{(\ell-1)!} S_\ell[x]^{\ell-1}$. Suppose first that $p = 1$. Then $\|\nabla_x^1 \psi(x) - \nabla_x^1 \psi(y)\|_{\mathcal{V}'} = 0$ for all x, y and the condition of **AS.1**(ii) holds for arbitrary positive values of $L_{1,\delta}$, $L_{2,\delta}$, β_1 and β_2. Suppose therefore that $p > 1$. For

$x \in \mathcal{B}(0,\delta)$, $y \in \mathcal{V}$ and $u \in \mathcal{V}$, $\|u\|_{\mathcal{V}} = 1$ we then derive that

$$\begin{aligned}
\langle \nabla_x^1 \psi(y) - \nabla_x^1 \psi(x), u\rangle &= \sum_{\ell=1}^{p} \frac{1}{(\ell-1)!} \langle S_\ell [x+(y-x)]^{\ell-1} - S_\ell [x]^{\ell-1}, u\rangle, \\
&= \sum_{\ell=1}^{p} \frac{1}{(\ell-1)!} \left\langle \sum_{i=0}^{\ell-1} \binom{\ell}{i} S_\ell [x]^{\ell-1-i} [(y-x)]^i - S_\ell [x]^{\ell-1}, u \right\rangle, \\
&= \sum_{\ell=2}^{p} \frac{1}{(\ell-1)!} \left\langle \sum_{i=1}^{\ell-1} \binom{\ell}{i} S_\ell [x]^{\ell-1-i} [(y-x)]^i, u \right\rangle, \\
&\leq \sum_{\ell=2}^{p} \sum_{i=1}^{\ell-1} \frac{1}{(\ell-1)!} \binom{\ell}{i} \|S_\ell\| \|x\|_{\mathcal{V}}^{l-1-i} \|y-x\|_{\mathcal{V}}^i, \\
&\leq \sum_{\ell=2}^{p} \kappa_{\ell,\delta} \|y-x\|_{\mathcal{V}}^{\ell-1}, \qquad (2.4)
\end{aligned}$$

For $\|y-x\|_{\mathcal{V}} \leq 1$, an upper bound on the right hand side of (2.4) is given by

$$\langle \nabla_x^1 \psi(y) - \nabla_x^1 \psi(x), u\rangle \leq \sum_{\ell=2}^{p} \kappa_{\ell,\delta} \|y-x\|_{\mathcal{V}}, \qquad (2.5)$$

while, for $\|y-x\|_{\mathcal{V}} \geq 1$, it is given by

$$\langle \nabla_x^1 \psi(y) - \nabla_x^1 \psi(x), u\rangle \leq \sum_{\ell=2}^{p} \kappa_{\ell,\delta} \|y-x\|_{\mathcal{V}}^{p-1}, \qquad (2.6)$$

Combining (2.5), (2.6) and the fact that $\|u\|_{\mathcal{V}} = 1$ yields **AS.1**(ii) with $\beta_1 = 1$, $\beta_2 = p-1 \geq 1$ and $L_{1,\delta} = L_{2,\delta} = \sum_{\ell=2}^{p} \kappa_{\ell,\delta}$. □

We now analyze the following very simple first-order linesearch-based algorithm on the following page for the minimization of ϕ.

Note that the existence of the direction d_k in Step 1 is guaranteed by **AS.1**(iv) and James' theorem [21]. The reader has undoubtly recognized the Wolfe linesearch conditions in (2.7) and (2.8) (see [25]). Unfortunately, the general form of (2.2) prevents extending the standard convergence theory for such algorithms applied to functions with Lipschitz gradients [25, Theorem 3.2] to our case. However, a modest modification of the classical argument allows us to prove the following convergence result.

Theorem 2.2 Suppose that ψ, h and $\mathcal{V}$ verify **AS.1** and let $\{x_k\}_{k\geq 0}$ be the sequence generated by Algorithm 2.1. Then

$$\phi(x_{k+1}) < \phi(x_k) \quad \text{for all} \quad k \geq 0$$

and either the algorithm terminates in a finite number of iterations with an iterate x_k such that $\nabla_x^1 \phi(x_k) = 0$, or

$$\lim_{k\to\infty} \|\nabla_x^1 \phi(x_k)\|_{\mathcal{V}'} = 0.$$

Algorithm 2.1: A Simple First-Order Algorithm for Minimizing Regularized Functionals Satisfying (2.2)

Step 0: Initialization. The constants $0 < c_1 < c_2 < 1$ are given. Set $x_0 = 0$ and $k = 0$.

Step 1: Compute a search direction. Compute $\nabla_x^1\phi(x_k) \in \mathcal{V}'$. If $\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'} = 0$, terminate and return x_k. Otherwise, select a direction d_k such that $\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'} = -\langle\nabla_x^1\phi(x_k), d_k\rangle$ and $\|d_k\| = 1$.

Step 2: Linesearch. Compute t_k a stepsize satisfying

$$\phi(x_k + t_k d_k) \leq \phi(x_k) + t_k c_1 \langle\nabla_x^1\phi(x_k), d_k\rangle, \tag{2.7}$$

$$\langle\nabla_x^1\phi(x_k + t_k d_k), d_k\rangle \geq c_2\langle\nabla_x^1\phi(x_k), d_k\rangle. \tag{2.8}$$

Step 3: Define the next iterate. Set $x_{k+1} = x_k + t_k d_k$, increment k by one and return to Step 1.

Proof. Because of the first Wolfe condition (2.7), the values $\{\phi(x_k)\}$ produced by Algorithm 2.1 are strictly decreasing, proving the theorem's first statement. More guarantees that all x_k lie in the level set $\mathcal{D}$. Using now the second Wolfe condition (2.8), we obtain that

$$\langle\nabla_x^1\phi(x_{k+1}) - \nabla_x^1\phi(x_k), d_k\rangle \geq (c_2 - 1)\langle\nabla_x^1\phi(x_k), d_k\rangle = (1 - c_2)\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'},$$

which, together with the fact that both x_k and x_{k+1} belong to $\mathcal{D}$, (2.2) (with $\delta = \omega$) and $\|d_k\|_{\mathcal{V}} = 1$, ensures that

$$(1 - c_2)\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'} \leq \langle\nabla_x^1\phi(x_{k+1}) - \nabla_x^1\phi(x_k), d_k\rangle \leq L'_{1,\omega}t_k^{\alpha_1} + L'_{2,\omega}t_k^{\alpha_2},$$

with $\alpha_1 < \alpha_2$. If $t_k \leq 1$, we obtain from the last inequality that

$$t_k \geq \left(\frac{(1 - c_2)\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'}}{L'_{1,\omega} + L'_{2,\omega}}\right)^{\frac{1}{\alpha_1}}. \tag{2.9}$$

Conversely, if $t_k \geq 1$, then

$$t_k \geq \left(\frac{(1 - c_2)\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'}}{L'_{1,\omega} + L'_{2,\omega}}\right)^{\frac{1}{\alpha_2}}. \tag{2.10}$$

Therefore,

$$t_k \geq \mu \min\left[\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'}^{\frac{1}{\alpha_1}}, \|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'}^{\frac{1}{\alpha_2}}\right],$$

where

$$\mu = \min\left[\left(\frac{(1 - c_2)}{L'_{1,\omega} + L'_{2,\omega}}\right)^{\frac{1}{\alpha_2}}, \left(\frac{(1 - c_2)}{L'_{1,\omega} + L'_{2,\omega}}\right)^{\frac{1}{\alpha_1}}\right].$$

Combining this lower bound on t_k with the first Wolfe condition yields that

$$\phi(x_{k+1}) \leq \phi(x_k) - c_1\mu \min\left(\|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'}^{\frac{1}{\alpha_1}}, \|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'}^{\frac{1}{\alpha_2}} \right) \|\nabla_x^1\phi(x_k)\|_{\mathcal{V}'}, \tag{2.11}$$

To prove the second theorem statement, we first note that the definition of the algorithm ensures the identity $\nabla_x^1\phi(x_k) = 0$ whenever termination occurs after a finite number of iterations. Assume therefore that the algorithm generates an infinite sequence of iterates and that

$$\|\nabla_x^1\phi(x_{k_i})\|_{\mathcal{V}'} \geq \epsilon, \tag{2.12}$$

for some $\epsilon > 0$ and some subsequence $\{k_i\}_{i=1}^{\infty}$. Summing over all iterations k_i and using **AS.1**(i), we obtain that

$$\begin{aligned} +\infty > \phi(0) - \phi_{\min} &\geq \sum_{i=1}^{\infty} c_1\mu \min\left(\|\nabla_x^1\phi(x_{k_i})\|_{\mathcal{V}'}^{\frac{\alpha_1+1}{\alpha_1}}, \|\nabla_x^1\phi(x_{k_i})\|_{\mathcal{V}'}^{\frac{\alpha_2+1}{\alpha_2}} \right), \\ &\geq c_1\mu \sum_{i=1}^{\infty} \min[\epsilon^{\frac{\alpha_1+1}{\alpha_1}}, \epsilon^{\frac{\alpha_2+1}{\alpha_2}}], \end{aligned} \tag{2.13}$$

which is a contradiction since the right-hand side diverges to $+\infty$. Hence (2.12) cannot hold and the second conclusion of the theorem holds. □

Thus a vanilla linesearch gradient-descent algorithm with the standard Wolfe conditions applied to infinite-dimensional functionals verifying **AS.1** yields asymptotic first-order stationarity. This is significant for our purpose of developing an adaptive regularization algorithm using a model defined by a regularized polynomial. Note that the iteration complexity of this algorithm in terms of $\epsilon \in (0, 1]$ can easily be derived from (2.13) since

$$\phi(0) - \phi_{\min} \geq c_1\mu \sum_{i=1}^{N_\epsilon} \min[\epsilon^{\frac{\alpha_1+1}{\alpha_1}}, \epsilon^{\frac{\alpha_2+1}{\alpha_2}}] \geq N_\epsilon c_1\mu\epsilon^{\frac{\alpha_1+1}{\alpha_1}}. \tag{2.14}$$

where N_ϵ denotes the total number of iterations to achieve the algorithm's termination condition. The iteration complexity for the linesearch Algorithm 2.1 is therefore $\mathcal{O}\left(\epsilon^{-\frac{1+\alpha_1}{\alpha_1}}\right)$ as a function of the requested accuracy ϵ of the gradient's norm. Note that $\alpha_1 \leq 1$ and hence this bound cannot be better than $\mathcal{O}\left(\epsilon^{-2}\right)$. This is reminiscent of Theorem 3.2 in [12], where the evaluation complexity of an adaptive regularization method in $\mathbb{R}^n$ is analyzed for functions with Hölder continuous gradients and a more specific regularization $h(x) = \|x\|_2^r$.

3 An adaptive regularization algorithm in Banach spaces

We now consider developing an adaptive regularization method for finding first-order points for the problem

$$\min_{x\in\mathcal{V}} f(x), \tag{3.1}$$

and make our assumptions on the problem more precise.

AS.2 f is p times continuously Fréchet differentiable with $p \geq 1$.
AS.3 There exists a constant $f_{\rm low}$ such that $f(x) \geq f_{\rm low}$ for all $x \in \mathcal{V}$.
AS.4 The p-th derivative tensor $\nabla_x^p f(x) \in \mathcal{L}(\mathcal{V}^p; \mathbb{R})$ is globally Hölder continuous, that is, there exist constants $L > 0$ and $\beta \in (0, 1]$ such that

$$\|\nabla_x^p f(x) - \nabla_x^p f(y)\| \leq L\|x - y\|_{\mathcal{V}}^{\beta}, \text{ for all } x, y \in \mathcal{V}. \tag{3.2}$$

For brevity, **AS.2** and **AS.4** will be denoted by $f \in \mathcal{C}^{p,\beta}(\mathcal{V}; \mathbb{R})$.
Let $T_{f,p}(x, s)$ be the Taylor series of the functional $f(x + s)$ truncated at order p.

$$T_{f,p}(x, s) \stackrel{\rm def}{=} f(x) + \sum_{l=1}^{p} \frac{1}{l!} \nabla_x^l f(x)[s]^l. \tag{3.3}$$

The gradient $\nabla_x^1 f(x)$ belongs to the dual space $\mathcal{V}'$ and will be denoted by $g(x)$. Thus, for a requested accuracy $\epsilon \in (0, 1]$, we are interested in finding an ϵ-approximate first-order critical point, that is a point x_ϵ such that $\|g(x_\epsilon)\|_{\mathcal{V}'} \leq \epsilon$.

3.1 Smooth Banach spaces

In a generic Banach space, we can only ensure "a decrease principle" as stated in [14, Theorem 5.22]. To obtain more conclusive results, we need to introduce additional assumptions. We choose to work with the class of *uniformly q smooth Banach spaces*. For the sake of completeness, we briefly recall the context. Given a Banach space $\mathcal{V}$, we first define its module of smoothness, for $t \geq 0$, by

$$\rho_{\mathcal{V}}(t) \stackrel{\rm def}{=} \sup_{\|x\|_{\mathcal{V}}=1, \|y\|_{\mathcal{V}}=t} \left\{ \frac{\|x + y\|_{\mathcal{V}} + \|x - y\|_{\mathcal{V}}}{2} - 1 \right\}, \tag{3.4}$$

and immediately deduce from the triangular inequality that $\rho_{\mathcal{V}}(t) \leq t$. We now say that $\mathcal{V}$ is a uniformly smooth Banach space if and only if $\lim_{t\to 0} \frac{\rho_{\mathcal{V}}(t)}{t} = 0$. Going one step further, we say that a Banach space $\mathcal{V}$ is *uniformly q smooth* for some $q \in (1, 2]$ if and only if

$$\exists \kappa_{\mathcal{V}} > 0, \ \rho_{\mathcal{V}}(t) \leq \kappa_{\mathcal{V}} t^q. \tag{3.5}$$

It is easy to see that, if $\mathcal{V}$ is uniformly q smooth, it is also uniformly q' smooth for all $1 < q' < q$. Indeed, one can easily show[(1)] that $\rho_{\mathcal{V}}(t) \leq \max(1, \kappa_{\mathcal{V}}) t^{q'}$ from definition (3.4) and inequality (3.5).

We motivate our choice of this particular class of Banach spaces by giving a few examples. $L^p(\mathbb{R}^n)$, $1 < p < \infty$, are uniformly smooth Banach spaces. In particular, $L^p(\mathbb{R}^n)$ is uniformly 2 smooth for $p \geq 2$ and uniformly p smooth for $1 < p \leq 2$. The same results apply for ℓ^p and the Sobolev spaces $W^{m,p}(\mathbb{R}^n)$. Moreover, all Hilbert spaces are 2 smooth Banach. See [29] for more details, in particular for the fact that q cannot be larger than 2.

From here on, we assume that

AS.5 $\mathcal{V}$ is a uniformly q smooth space.

Uniformly smooth Banach spaces are also reflexive (See [29, Proposition 1.e.3, p.61]), so that **AS.1**(iv) automatically holds. Let us now define the set

$$J_p(x) \stackrel{\rm def}{=} \left\{ v^* \in \mathcal{V}', \ \langle v^*, x\rangle = \|x\|_{\mathcal{V}}^p, \ \|v^*\|_{\mathcal{V}'} = \|x\|_{\mathcal{V}}^{p-1} \right\}. \tag{3.6}$$

[(1)] If $t \in [0, 1]$ this follows from (3.5) and $q' < q$. If $t > 1$, $\rho_{\mathcal{V}}(t) \leq t \leq t^{q'}$.

It is known [28] that $J_p(x)$ is the subdifferential of the functional $\frac{1}{p}\|\cdot\|_{\mathcal{V}}^p$, $p \geq 1$ at x. We may now introduce another characterization of uniform smoothness.

Theorem 3.1 Let

$\mathcal{F} \stackrel{\rm def}{=} \{\psi : \mathbb{R} \to \mathbb{R} \mid \psi(0) = 0, \psi \text{ is convex, non decreasing and } \exists \kappa_{\mathcal{F}} > 0 \mid \psi(t) \leq \kappa_{\mathcal{F}} \rho_{\mathcal{V}}(t)\}$.

Then, for any $1 < p < \infty$, the following statements are equivalent.

(i) $\mathcal{V}$ is a uniformly smooth Banach space.

(ii) J_p is single valued and there exists $\varphi_p(t) = \frac{\psi_p(t)}{t}$ where $\psi_p \in \mathcal{F}$ and such that

$$\|J_p(x) - J_p(y)\|_{\mathcal{V}'} \leq \max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}})^{p-1} \varphi_p\left(\frac{\|x-y\|_{\mathcal{V}}}{\max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}})}\right). \tag{3.7}$$

Proof. [29, Theorem 2]. □

As we will be only working with $\|.\|_{\mathcal{V}}^p$ for $p > 1$ in the rest of the paper, we define $J_p(x)$ as the unique value in the set (3.6). As the subdifferential of $\|.\|_{\mathcal{V}}^p$ reduces to a singleton for $p > 1$ and $\|.\|_{\mathcal{V}}^p$ is a convex function, $\|.\|_{\mathcal{V}}^p$ is Fréchet differentiable for $p > 1$ since it verifies [14, Condition 4.16]. The reader is referred to [28] or [29] for more extensive coverage of characterizations of the norm in uniformly smooth Banach spaces.

For all $\ell > 1$, we now prove an upper bound of the norm of $\|J_\ell(x) - J_\ell(y)\|_{\mathcal{V}'}$ in terms of $\|x-y\|_{\mathcal{V}}$ in a uniform q smooth Banach space. Let us first remind the useful inequality $(x+y)^r \leq \max(1, 2^{r-1})(x^r + y^r)$ for all $x, y \geq 0$ and all $r \geq 0$, before stating the next crucial lemma.

Lemma 3.2 Suppose that $\mathcal{V}$ is a uniformly q smooth Banach space and that $x \in \mathcal{B}(0, \omega)$. Then for all $\ell > 1$, there exist constants $\kappa_\omega, \kappa_\ell > 0$ such that

$$\|J_\ell(x) - J_\ell(y)\|_{\mathcal{V}'} \leq \kappa_\omega \|x-y\|_{\mathcal{V}}^{\min[q,\ell]-1} + \kappa_\ell \|x-y\|_{\mathcal{V}}^{\ell-1}, \tag{3.8}$$

where κ_ω and κ_ℓ depend only on ω, ℓ, $\kappa_{\mathcal{F}}$ and $\kappa_{\mathcal{V}}$.

Proof. As $\ell > 1$, if $q > \ell$, we can use our remark above and decrease the q smooth order until $q' = \min[q, \ell] \leq \ell$. We now develop the upper bound (ii) of Theorem 3.1 and use the definition of the set $\mathcal{F}$ to derive that

$$\begin{aligned}\|J_l(x) - J_l(y)\|_{\mathcal{V}'} &\leq \max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}})^{\ell-1} \kappa_{\mathcal{F}} \kappa_{\mathcal{V}} \left(\frac{\|x-y\|_{\mathcal{V}}}{\max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}})}\right)^{q'-1}, \\ &\leq \max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}})^{l-q'} \kappa_{\mathcal{F}} \kappa_{\mathcal{V}} \|x-y\|_{\mathcal{V}}^{q'-1}.\end{aligned}$$

Using now the inequalities $\max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}}) \leq \|x\|_{\mathcal{V}} + \|x-y\|_{\mathcal{V}}$ and $\ell \geq q'$, we obtain that

$$\begin{aligned}
\|J_l(x) - J_l(y)\|_{\mathcal{V}'} &\leq \kappa_{\mathcal{F}}\kappa_{\mathcal{V}}(\|x\|_{\mathcal{V}} + \|x-y\|_{\mathcal{V}})^{\ell-q'}\|x-y\|_{\mathcal{V}}^{q'-1}, \\
&\leq \kappa_{\mathcal{F}}\kappa_{\mathcal{V}} \max(1, 2^{\ell-q'-1})(\|x\|_{\mathcal{V}}^{\ell-q'} + \|x-y\|_{\mathcal{V}}^{\ell-q'})\|x-y\|_{\mathcal{V}}^{q'-1}, \\
&\leq \kappa_{\mathcal{F}}\kappa_{\mathcal{V}} \max(1, 2^{\ell-q'-1})\, \omega^{\ell-q'}\|x-y\|_{\mathcal{V}}^{q'-1} \\
&\qquad + \kappa_{\mathcal{F}}\kappa_{\mathcal{V}} \max(1, 2^{\ell-q'-1})\|x-y\|_{\mathcal{V}}^{\ell-1}, \\
&\leq \kappa_{\omega}\|x-y\|_{\mathcal{V}}^{q'-1} + \kappa_{\ell}\|x-y\|_{\mathcal{V}}^{\ell-1}.
\end{aligned}$$

□

It results from this theorem that the primal representation of the gradient of a regularization term of the form $\|s\|_{\mathcal{V}}^{\alpha}$ does satisfy the condition of **AS.1**(iii). This will be crucial as it will allow applying Algorithm 2.1 to a model consisting of a multivariate polynomial (satisfying **1S.1**(ii)) augmented by such a regularization term.

3.2 The ARp-BS algorithm

Adaptive regularization methods are iterative schemes which compute a step form an iterate x_k by building, for $f \in \mathcal{C}^{p,\beta}(\mathcal{V}; \mathbb{R})$, a regularized model $m_k(s)$ of $f(x_k + s)$ of the form

$$m_k(s) \stackrel{\rm def}{=} T_{f,p}(x_k, s) + \frac{\sigma_k}{(p+\beta)!}\|s\|_{\mathcal{V}}^{p+\beta}, \; p \geq 1. \tag{3.9}$$

As in [11] but at variance with [12], we will assume here that β, the degree of Hölder continuity of the p-th derivative tensor of f, is known. The p-th order Taylor series is "regularized" by adding the term $\frac{\sigma_k}{(p+\beta)!}\|s\|_{\mathcal{V}}^{p+\beta}$, where σ_k is known as the "regularization parameter". This term guarantees that the functionnal $m_k(s)$ is bounded below and thus makes the procedure of finding a step s_k by (approximately) minimizing $m_k(s)$ well-defined. In our uniform q smooth setting, $m_k(s)$ is Fréchet differentiable but this is unfortunately insufficient to derive results on the Lipschitz continuity of its gradient, which makes the use of more standard gradient-descent methods impossible.
Our proposed algorithm is similar in spirit to ARC [8] and proceeds as follows. At a given iterate x_k, a step s_k is first computed by approximately minimizing (3.9). Once the step is computed, the value of the objective functional at the trial point $x_k + s_k$ is then evaluated. If the decrease in f from x_k to $x_k + s_k$ is comparable to that predicted by the p-th order Taylor series, the trial point is accepted as the new iterate and the regularization parameter is (possibly) reduced. If this is not the case, the trial point is rejected and the regularization parameter is increased. The resulting algorithm is formally stated as the ARp-BS algorithm on the next page.

The ARp-BS algorithm follows the main lines of existing ARp methods [8, 6]. However, as we have already mentioned, the existence of a minimizer of $m_k(s)$ may not be guaranteed in infinite dimensions and hence a point $s^\star$ such that $\nabla_s^1 m_k(s^\star) = 0$ may not exist. As a consequence, standard proofs that a step satisfying both (3.12) and (3.13) exists no longer apply. We thus need to check that a suitable step can still de found in our context. This is achieved using Algorithm 2.1.

Algorithm 3.1: p-th order adaptive regularization in a uniform q smooth Banach Space (ARp-BS)

Step 0: Initialization: An initial point $x_0 \in \mathcal{V}$, a regularization parameter σ_0 and a requested final gradient accuracy $\epsilon \in (0,1]$ are given. The constants η_1, η_2, γ_1, γ_2, γ_3, $\chi \in (0,1)$, and $\sigma_{\min}$ are also given such that

$$\sigma_{\min} \in (0, \sigma_0], 0 < \eta_1 \leq \eta_2 < 1 \quad \text{and} \quad 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \tag{3.10}$$

Compute $f(x_0)$ and set $k = 0$.

Step 1: Check for termination: Evaluate $g_k = \nabla_x^1 f(x_k)$. Terminate with $x_\epsilon = x_k$ if

$$\|g(x_k)\|_{\mathcal{V}'} \leq \epsilon. \tag{3.11}$$

Step 2: Step calculation: Evaluate $f(x_k)$ and $\{\nabla_x^i f(x_k)\}_{i=2}^p$. Compute a step s_k which sufficiently reduces the model m_k defined in (3.9) in the sense that

$$m_k(s_k) < m_k(0), \tag{3.12}$$

and

$$\|\nabla_s^1 m_k(s_k)\|_{\mathcal{V}'} \leq \max\left[\chi\epsilon, \theta\|s_k\|_{\mathcal{V}}^{p+\beta-1}\right]. \tag{3.13}$$

Step 3: Acceptance of the trial point. Compute $f(x_k + s_k)$ and define

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_{f,p}(x_k, 0) - T_{f,p}(x_k, s_k)}. \tag{3.14}$$

If $\rho_k \geq \eta_1$, then define $x_{k+1} = x_k + s_k$; otherwise define $x_{k+1} = x_k$.

Step 4: Regularization parameter update. Set

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1\sigma_k), \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2\sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \text{if } \rho_k < \eta_1. \end{cases} \tag{3.15}$$

Increment k by one and go to Step 1.

Theorem 3.3 Suppose that **AS.2**, **AS.4** and **AS.5** hold. Suppose also that $\|g(x_k)\|_{\mathcal{V}'} > 0$. Then a step satisfying both (3.12) and (3.13) always exists.

Proof. First note that **AS.2** and **AS.4** imply that $p + \beta > 1$. In order to apply Algorithm 2.1 to the problem of minimizing (3.9), we just need to prove that $m_k(s)$ satisfies **AS.1** of Section 2. We have that

$$m_k(s) \geq m_k(0) - \sum_{i=1}^{p} \|\nabla_x^i f(x)\| \|s\|_{\mathcal{V}}^i + \frac{\sigma_k}{(p+\beta)!}\|s\|_{\mathcal{V}}^{p+\beta} \to \infty \text{ as } \|s\|_{\mathcal{V}} \to \infty,$$

and thus m_k is a coercive functional verifying **AS.1**(i). Lemma 2.1 ensures that the Taylor series term $T_{f,p}(x_k, s)$ satisfies **AS.1**(ii). Lemma 3.2 (applied with $\delta = \omega$, $\ell = p + \beta - 1$, $L_{3,\delta} = \kappa_\ell$, $\beta_3 = \min[q, \ell] - 1 \in (0, 1]$, $L_{4,\delta} = \kappa_\omega$ and $\beta_4 = \ell + \beta - 1 > 0$) then ensures that $\|.\|_{\mathcal{V}}^{p+\beta}$ satisfies **AS.1**(iii). We already noted that, being uniformly smooth, $\mathcal{V}$ must be reflexive, which ensures that **AS.1**(iv) holds. All the requirements of **AS.1** in Section 2 are therefore met and, since $\nabla_s^1 m_k(0) = g(x_k)$, Theorem 2.2 applies to the functional $m_k(s)$. As a consequence, a suitable step s_k such that $m_k(s_k) < m_k(0)$ and $\|\nabla_s^1 m_k(s_k)\|_{\mathcal{V}'} \leq \chi\epsilon$ exists. □

Observe that equation (2.14) and the fact that $\alpha_1 = \min[q, p + \beta] - 1$ and $\alpha_2 = p + \beta - 1$ (all the other powers ranging from 2 to p), imply that, for our iterative gradient descent Algorithm 2.1,

$$\lim_{i \to \infty} \min \left[\kappa_A \|\nabla_s^1 m(s_i)\|_{\mathcal{V}'}^{\frac{\min[q,p+\beta]}{\min[q,p+\beta]-1}}, \kappa_B \|\nabla_s^1 m(s_i)\|_{\mathcal{V}'}^{\frac{p+\beta}{p+\beta-1}} \right] = 0.$$

As a consequence, the first term in the minimum indicates that the smoother the space, the faster the convergence for $p \geq 2$.
Following well-established practice, we now define

$$\mathcal{S} \stackrel{\text{def}}{=} \{k \geq 0 \mid x_{k+1} = x_k + s_k\} = \{k \geq 0 \mid \rho_k \geq \eta_1\},$$

the set of indexes of "successful iterations", and

$$\mathcal{S}_k \stackrel{\text{def}}{=} \mathcal{S} \cap \{1, \ldots, k\},$$

the set of indexes of successful iterations up to iteration k. We also recall a well-known result bounding the total number of iterations in terms of the number of successful ones.

Lemma 3.4 Suppose that the ARp-BS algorithm is used and that $\sigma_k \leq \sigma_{\max}$ for some $\sigma_{\max} > 0$. Then

$$k \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0}\right). \tag{3.16}$$

Proof. See [6, Theorem 2.4]. □

4 Evaluation complexity for the ARp-BS algorithm

Before discussing our analysis of evaluation complexity, we first restate some classical lemmas of ARp algorithms, starting with Hölder error bounds.

Lemma 4.1 Suppose that$f \in \mathcal{C}^{p,\beta}(\mathcal{V};\mathbb{R})$ holds and that $k \in \mathcal{S}$. Then

$$|f(x_{k+1}) - T_{f,p}(x_k, s_k)| \leq \frac{L}{(p+\beta)!}\|s_k\|_{\mathcal{V}}^{p+\beta}, \tag{4.1}$$

and

$$\|g_{k+1} - \nabla_s^1 T_{f,p}(x_k, s_k)\|_{\mathcal{V}'} \leq \frac{L}{(p-1+\beta)!}\|s_k\|_{\mathcal{V}}^{p-1+\beta}. \tag{4.2}$$

Proof. This is a direct extension of [13, Lemma 2.1] since the proof in this reference only involves **AS.2**, **AS.4** and unidimensional integrals. □

From now on, the analysis follows that presented in [6] quite closely.

Lemma 4.2

$$\Delta T_{f,p}(x_k, s_k) \stackrel{\rm def}{=} T_{f,p}(x_k, 0) - T_{f,p}(x_k, s_k) \geq \frac{\sigma_k}{(p+\beta)!}\|s_k\|_{\mathcal{V}}^{p+\beta}. \tag{4.3}$$

Proof. Direct from (3.12) and (3.9). □

Lemma 4.3 Suppose that $f \in \mathcal{C}^{p,\beta}(\mathcal{V};\mathbb{R})$. Then, for all $k \geq 0$,

$$\sigma_k \leq \sigma_{\max} \stackrel{\rm def}{=} \gamma_3 \max\left[\sigma_0, \frac{L}{(1-\eta_2)}\right]. \tag{4.4}$$

Proof. See [6, Lemma 2.2]. Using (3.14), (4.1), and (4.3), we obtain that

$$|\rho_k - 1| \leq \frac{(p+\beta)!|f(x_k+s_k) - T_{f,p}(x_k, s_k)|}{\sigma_k\|s_k\|_{\mathcal{V}}^{p+\beta}} \leq \frac{L}{\sigma_k}.$$

Thus, if $\sigma_k \geq L/(1-\eta_2)$, then $\rho_k \geq \eta_2$ ensures that iteration k is successful and (3.15) implies that $\sigma_{k+1} \leq \sigma_k$. The mechanism of the algorithm then guarantees that (4.4) holds. □

The next lemma remains in the spirit of [6, Lemma 2.3], but now takes the condition (3.13) into account.

Lemma 4.4 Suppose that $f \in \mathcal{C}^{p+\beta}(\mathcal{V};\mathbb{R})$ holds and that $k \in \mathcal{S}$ before termination. Then

$$\|s_k\|_{\mathcal{V}}^{p-1+\beta} \geq \epsilon \min\left[\frac{(1-\chi)(p+\beta-1)!}{L+\sigma_{\max}}, \frac{(p+\beta-1)!}{L+\sigma_{\max}+\theta(p+\beta-1)!}\right]. \tag{4.5}$$

Proof. Successively using the fact that termination does not occur at iteration k and condition (3.13), we deduce that

$$\begin{aligned}
\epsilon &< \|g(x_{k+1})\|_{\mathcal{V}'}, \\
&\leq \|g(x_{k+1}) - \nabla_s^1 T_{f,p}(x_k,s_k)\|_{\mathcal{V}'} + \|\nabla_s^1 m_k(s_k)\|_{\mathcal{V}'} + \frac{\sigma_k}{(p+\beta-1)!}\|J_{p+\beta}(s_k)\|_{\mathcal{V}'}, \\
&\leq \frac{L}{(p-\beta+1)!}\|s_k\|_{\mathcal{V}}^{p-1+\beta} + \max\left[\chi\epsilon, \theta\|s_k\|_{\mathcal{V}}^{p-\beta+1}\right] + \frac{\sigma_k}{(p+\beta-1)!}\|s_k\|_{\mathcal{V}}^{p+\beta-1}.
\end{aligned}$$

By treating each case in the maximum separately, we obtain that either

$$(1-\chi)\epsilon \leq \left(\frac{L}{(p+\beta-1)!} + \frac{\sigma_k}{(p+\beta-1)!}\right)\|s_k\|_{\mathcal{V}}^{p-1+\beta},$$

or

$$\epsilon \leq \left(\frac{L}{(p+\beta-1)!} + \frac{\sigma_k}{(p+\beta-1)!} + \theta\right)\|s_k\|_{\mathcal{V}}^{p-1+\beta}.$$

Combining the two last inequalities gives that

$$\|s_k\|_{\mathcal{V}}^{p-1+\beta} \geq \min\left[\frac{(1-\chi)\epsilon(p+\beta-1)!}{L+\sigma_{\max}}, \frac{(p+\beta-1)!\epsilon}{L+\sigma_{\max}+\theta(p+\beta-1)!}\right].$$

This in turn directly implies (4.5). □

We may now resort to the standard "telescoping sum" argument to obtain the desired evaluation complexity result.

Theorem 4.5 Suppose that **AS.2**–**AS.5** hold. Then the ARp-BS algorithm requires at most

$$\kappa_{\text{ARpBS}}\frac{f(x_0)-f_{\text{low}}}{\epsilon^{\frac{p+\beta}{p+\beta-1}}},$$

successful iterations and evaluations of $\{\nabla_x^i f\}_{i=1,2,\ldots,p}$ and at most

$$\kappa_{\text{ARpBS}}\frac{f(x_0)-f_{\text{low}}}{\epsilon^{\frac{p+\beta}{p+\beta-1}}}\left(1+\frac{|\log\gamma_1|}{\log\gamma_2}\right)+\frac{1}{\log\gamma_2}\log\left(\frac{\sigma_{\max}}{\sigma_0}\right),$$

evaluations of f to produce a vector $x_\epsilon \in \mathcal{V}$ such that $\|g(x_\epsilon)\|_{\mathcal{V}'} \leq \epsilon$, where

$$\kappa_{\text{ARpBS}} = \frac{(p+\beta-1)!}{\eta_1\sigma_{\min}}\min\left[\frac{(1-\chi)(p+\beta-1)!}{L+\sigma_{\max}}, \frac{(p+\beta-1)!}{L+\sigma_{\max}+(p+\beta-1)!\theta}\right]^{\frac{p+\beta}{p+\beta-1}}.$$

Proof. Let k be the index of an iteration before termination. Then, using **AS.3**, the definition of successful iterations, (4.3) and (4.5), and the fact that computing an appropriate step is of constant order of complexity, we obtain that

$$f(x_0) - f_{\rm low} \geq \sum_{i=0, i\in\mathcal{S}}^{k} f(x_i) - f(x_{i+1}) \geq \eta_1 \sum_{i\in\mathcal{S}_k} \Delta T_{f,2}(x_i, s_i) \geq \frac{|\mathcal{S}_k|}{\kappa_{\rm ARpBS}} \epsilon^{\frac{p+\beta}{p+\beta-1}}.$$

Thus

$$|\mathcal{S}_k| \leq \kappa_{\rm ARpBS} \frac{f(x_0) - f_{\rm low}}{\epsilon^{\frac{p+\beta}{p+\beta-1}}},$$

for any k before termination. The first conclusion follows since the derivatives are only evaluated once per successful iteration. Applying now Lemma 3.4 gives the second conclusion.

□

Theorem 4.5 extends the result of [6] in the case $\beta = 1$ and some results of [13] to uniform q smooth Banach spaces. We recall that L^p, ℓ^p and $W^{m,p}$ are uniform q smooth spaces for $1 < p < \infty$, and hence that Lemma 3.2 and Theorem 4.5 apply in these spaces. We may also consider the finite dimensional case where $\mathbb{R}^n$ is equipped with the norm $\|x\|_r = (\sum_{i=1}^n |x_i|^r)^{\frac{1}{r}}$. We know that, for all $1 < r < \infty$, this is a uniform $\min(r, 2)$ smooth space, and therefore Theorem 3.5 again applies. We could of course have obtained convergence of the adaptive regularization algorithm in this case using results for the Euclidean norm and introducing norm-equivalence constants in our proofs and final result, but this is avoided by the approach presented here. This could be significant when the dimension is large and the norm-equivalence constants grow.

Finally note that the evaluation complexity of Algorithm 2.1 discussed at the end of Section 2 is interesting but irrelevant for the evaluation complexity of the ARp-BS algorithm, because the former only evaluates the model m_k without requiring any evaluations of f or its derivatives beyond those already performed in ARp-BS.

5 Discussion

We have proposed a generalized Hölder condition and a gradient-descent algorithm for minimizing polynomial functionals with a general convex regularization term in Banach spaces, and have applied this result to show the existence of a suitable step in an adaptive regularization method for unconstrained minimization in q smooth Banach spaces. We have also analyzed the evaluation complexity of this latter algorithm and have shown that, under standard assumptions, it will find an ϵ-approximate first-order critical point in at most $\mathcal{O}\left(\epsilon^{-\frac{p+\beta}{p+\beta-1}}\right)$ evaluations of the functional and its first p derivatives, which is identical to the bound known for minimization in (finite-dimensional) Euclidean spaces. Since these bounds are known to be sharp [11], so is ours.

It would be interesting to consider convergence to second-order points, but the infinite dimensional framework causes more difficulties. Indeed, considering second-order derivatives as in [13] is impossible since we do not know if a power of the norm is twice differentiable.

As an example, consider $L^r([0,1])$ for $p > 1$, where

$$\nabla_f^1 \left(\frac{\|f\|^p_{L^r([0,1])}}{p} \right) = \|f\|^{p-r}_{L^r([0,1])} f|f|^{r-2}.$$

The right-hand side of the last equation involves an absolute value which is only differentiable for specific values of r. It is interesting to study the case of $r = 2$ with the objective of extending our analysis to the second order. Another line of future work is to extend these results to metrizable spaces (using the Bregman divergence or the Wasserstein distance) and to the complexity of second order adaptive regularization in an infinite-dimensional Hilbert space.

References

[1] E. L. Allgower, K. Böhmer, F. A. Potra, and W. C. Rheinboldt. A mesh-independence principle for operator equations and their discretizations. *SIAM Journal on Numerical Analysis*, 23(1):160–169, 1986.

[2] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint. Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2881–2915, 2019.

[3] E. Bergou, Y. Diouane, and S. Gratton. On the use of the energy norm in trust-region and adaptive cubic regularization subproblems. Optimization Online, April 2017.

[4] T. Bianconcini, G. Liuzzi, B. Morini, and M. Sciandrone. On the use of iterative methods in cubic regularization for unconstrained optimization. *Computational Optimization and Applications*, 60(1):35–57, 2015.

[5] T. Bianconcini and M. Sciandrone. A cubic regularization algorithm for unconstrained optimization using line search and nonmonotone techniques. *Optimization Methods and Software*, 31(5):1008–1035, 2016.

[6] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming, Series A*, 163(1):359–368, 2017.

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

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Mathematical Programming, Series A*, 130(2):295–319, 2011.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. *Optimization Methods and Software*, 27(2):197–219, 2012.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM Journal on Optimization*, 22(1):66–86, 2012.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. In B. Sirakov, P. de Souza, and M. Viana, editors, *Invited Lectures, Proceedings of the 2018 International Conference of Mathematicians (ICM 2018), vol. 4, Rio de Janeiro*, pages 3729–3768. World Scientific Publishing Co Pte Ltd, 2018.

[12] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Universal regularization methods – varying the power, the smoothness and the accuracy. *SIAM Journal on Optimization*, 29(1):595—-615, 2019.

[13] C. Cartis, N. I. M. Gould, and Ph. 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.

[14] F. H. Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*. Number 264 in Graduate Texts in Mathematics. Springer Verlag, Heidelberg, Berlin, New York, 2013.

[15] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

[16] P. Deuflhard and F. A. Potra. Asymptotic mesh independence for Newton–Galerkin methods via a refined Mysovskii theorem. *SIAM Journal on Numerical Analysis*, 29(5):1395–1412, 1992.

[17] G. N. Grapiglia, J. Yuan, and Y. Yuan. On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization. *Mathematical Programming, Series A*, 152:491–520, 2015.

[18] S. Gratton, C. W. Royer, and L. N. Vicente. A decoupled first/second-order steps technique for nonconvex nonlinear unconstrained optimization with improved complexity bounds. Technical Report TR 17-21, Department of Mathematics, University of Coimbra, Coimbra, Portugal, 2017.

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

[20] M. Heinkenschloss. Mesh independence for nonlinear least squares problems with norm constraints. *SIAM Journal on Optimization*, 3(1):81–117, 1993.

[21] Robert C. James. Reflexivity and the sup of linear functionals. *Israel Journal of Mathematics*, 13(3-4):289–300, September 1972.

[22] C. T. Kelley and E. W. Sachs. Quasi-Newton methods and unconstrained optimal control problems. *SIAM Journal on Control and Optimization*, 25(6):1503–1517, 1987.

[23] J. M. Martínez. On high-order model regularization for constrained optimization. *SIAM Journal on Optimization*, 27:2447–2458, 2017.

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

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

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

[27] M. Ulbrich and S. Ulbrich. Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional problems with pointwise bounds. *SIAM Journal on Control and Optimization*, 38(6):1938–1984, 2000.

[28] H.-K. Xu. Inequalities in Banach spaces with applications. *Nonlinear Analysis: Theory, Methods & Applications*, 16(12):1127–1138, 1991.

[29] Z.-B. Xu and G. F. Roach. Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces. *Journal of Mathematical Analysis and Applications*, 157(1):189–210, 1991.