

# Hölder Gradient Descent and Adaptive Regularization Methods in Banach Spaces for First-Order Points

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

4 IV 2021

## Abstract

This paper considers optimization of smooth nonconvex functionals in smooth infinite dimensional spaces. A Hölder gradient descent algorithm is first proposed for finding approximate first-order points of regularized polynomial functionals. This method is then applied to analyze the evaluation complexity of an adaptive regularization method which searches for approximate first-order points of functionals with  $\beta$ -Hölder continuous derivatives. It is shown that finding an  $\epsilon$ -approximate first-order point requires 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, 23, 7, 8, 10, 4, 17, 5, 6, 22, 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 [23, 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 motivation for this generalization is a matter of 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

---

\*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

an important feature of numerical optimization methods [21, 1, 16, 20, 25]. For trust-region methods, this was studied in [24] 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.

The outline of adaptive regularization methods is today quite well-known for finite dimensional spaces (see [6], for instance), but difficulties arise in the nonconvex infinite dimensional space case. The main problem is that the existence of a suitable step at a given iteration of the method typically hinges on approaching a minimizer of the 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.

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

- We first analyse the convergence of a method for minimizing polynomial functionals with a general differentiable convex regularization whose gradients satisfy a generalized Hölder condition. To our knowledge, no such regularization has been considered before, even in finite dimensional spaces.
- We then propose an adaptive regularization algorithm for finding first-order points of nonconvex functions having Hölder continuous  $p$ -th derivative (in the Fréchet sense) and analyze its evaluation complexity. We 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 regularized polynomials 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,  $\|\cdot\|_{\mathcal{V}}$  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 functions from  $\mathcal{V} \times \mathcal{V} \cdots \times \mathcal{V}$  to  $\mathbb{R}$  and  $\mathcal{L}_{sym}^m(\mathcal{V}^{\otimes m}; \mathbb{R})$  the subspace of  $\mathcal{L}^m(\mathcal{V}^{\otimes m}; \mathbb{R})$  that is  $m$ -linear symmetric. For a function  $f$  defined from  $\mathcal{V}$  to  $\mathbb{R}$  that is  $p$  times Fréchet differentiable,  $\nabla_x^k f(x) \in \mathcal{L}_{sym}^k(\mathcal{V}^{\otimes k}; \mathbb{R})$  denotes the  $k$ -th derivative tensor for  $k \in \{1, \dots, p\}$ .  $\nabla_x^1 f$  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  $\langle y, x \rangle \stackrel{\text{def}}{=} y(x)$ , for  $y \in \mathcal{V}'$  and  $x \in \mathcal{V}$ . The norm in the dual space  $\mathcal{V}'$  will be denoted as  $\|\cdot\|_{\mathcal{V}'}$ . For  $S \in \mathcal{L}_{sym}^m$ ,  $S[v_1, v_2, \dots, v_m] \in \mathbb{R}$  denotes the result of applying  $S$  to the vectors  $v_1, \dots, v_m$ .  $S[v]^m$  is the result of applying  $S$  to  $m$  copies of the vector  $v$  and  $S[v]^l \in \mathcal{L}_{sym}^{m-l}(\mathcal{V}^{\otimes m-l}; \mathbb{R})$  the result of applying  $l$  times the vector  $v$ . We define the norm in  $\mathcal{L}_{sym}^m(\mathcal{V}^{\otimes m}; \mathbb{R})$  as

$$\|S\| \stackrel{\text{def}}{=} \sup_{\|v_1\|_{\mathcal{V}}=\dots=\|v_m\|_{\mathcal{V}}=1} |S[v_1, \dots, v_m]|. \quad (1.1)$$

## 2 Gradient descent with a Hölder regularization

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

$$\phi(x) \stackrel{\text{def}}{=} \phi_0 + \sum_{\ell=1}^p \frac{1}{\ell!} S_\ell[x]^\ell + h(x), \quad (2.1)$$

where  $S_\ell \in \mathcal{L}_{sym}^\ell(\mathcal{V}^{\otimes \ell})$  for  $\ell \in \{1, \dots, p\}$  and  $h$  is a general regularization term. Note that the sum of the two first terms of the right-hand side have the form of a Taylor expansion (in the Fréchet sense). The functions  $\phi$  and  $h$  and the space  $\mathcal{V}$  are assumed to satisfy the following assumptions.

### 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  $\mathcal{D} \stackrel{\text{def}}{=} \{x \in \mathcal{V}, \phi(x) \leq \phi(0)\}$  is bounded in the sense that  $\sup_{x \in \mathcal{D}} \|x\|_{\mathcal{V}} \leq \omega$  for some  $\omega < \infty$ .
- (ii)  $h$  is a convex differentiable function whose gradient satisfies the local Hölder condition

$$\forall \delta > 0, \forall x \in \mathcal{B}(0, \delta), \forall y \in \mathcal{V}, \|\nabla_x^1 h(x) - \nabla_x^1 h(y)\|_{\mathcal{V}'} \leq \sum_{i=1}^k L_{i,\delta} \|x - y\|_{\mathcal{V}}^{\beta_i - 1},$$

where, for  $i \in \{1, \dots, k\}$ ,  $\beta_i > 1$  and  $L_{i,\delta}$  are a positive constants, the latter depending on  $\delta$ . Moreover,  $\beta_i \leq 2$  for at least one  $i \in \{1, \dots, k\}$ .

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

Observe that the condition stated in **AS.1**(ii) reduces to the standard  $\beta_1 - 1$ -Hölder continuity of the gradients of  $h$  whenever  $k = 1$ . Also note that, if all  $\beta_i$  were strictly larger than two,  $h$  would be affine.

We now use the property that  $S_\ell \in \mathcal{L}_{sym}^\ell(\mathcal{V}^{\otimes \ell}; \mathbb{R})$  to derive an upper bound of  $\phi(x + s)$  for all  $x \in \mathcal{D}, s \in \mathcal{V}$ . We then choose a specific  $s$  to obtain the next result.

**Lemma 2.1** There exists an integer  $m \geq p$  and constants  $\kappa_{i,\omega} > 0$  ( $i \in \{1, \dots, m\}$ ) such that, for all  $x \in \mathcal{D}$ , there exists a vector  $d$  in  $\mathcal{V}$ ,

$$\phi(x - td) \leq \phi(x) - \|\nabla_x^1 \phi(x)\|_{\mathcal{V}'} t + \sum_{i=1}^m \kappa_{i,\omega} t^{\gamma_i}, \quad (2.2)$$

where  $1 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_m$  and  $t \in \mathbb{R}$ .

**Proof.** Successively using the binomial expansion, the convexity of  $h$ , (1.1), the fact

that  $\|x\|_{\mathcal{V}} \leq \omega$  because  $x \in \mathcal{D}$  and **AS.1(ii)**, we derive that

$$\begin{aligned}
 \phi(x+s) &= \phi_0 + \sum_{\ell=1}^p \frac{1}{\ell!} S_{\ell}[x+s]^{\ell} + h(x+s), \\
 &= \phi(x) + \sum_{\ell=1}^p \frac{1}{(\ell-1)!} \langle S_{\ell}[x]^{\ell-1}, s \rangle + \sum_{\ell=2}^p \sum_{i=0}^{\ell-2} \frac{1}{\ell!} \binom{\ell}{i} S_{\ell}[x]^i [s]^{\ell-i} + h(x+s) - h(x), \\
 &\leq \phi(x) + \sum_{\ell=1}^p \frac{1}{(\ell-1)!} \langle S_{\ell}[x]^{\ell-1}, s \rangle + \sum_{\ell=2}^p \sum_{i=0}^{\ell-2} \frac{1}{\ell!} \binom{\ell}{i} \|S_{\ell}\| \|x\|_{\mathcal{V}}^i \|s\|_{\mathcal{V}}^{\ell-i} + \langle \nabla_x^1 h(x+s), s \rangle, \\
 &\leq \phi(x) + \sum_{\ell=1}^p \frac{1}{(\ell-1)!} \langle S_{\ell}[x]^{\ell-1}, s \rangle + \sum_{\ell=2}^p \sum_{i=0}^{\ell-2} \frac{1}{\ell!} \binom{\ell}{i} \|S_{\ell}\| \omega^i \|s\|_{\mathcal{V}}^{\ell-i} + \langle \nabla_x^1 h(x+s), s \rangle, \\
 &\leq \phi(x) + \langle \nabla_x^1 \phi(x), s \rangle + \sum_{\ell=2}^p \kappa_{\ell, \omega} \|s\|_{\mathcal{V}}^{\ell} + \langle \nabla_x^1 h(x+s) - \nabla_x^1 h(x), s \rangle, \\
 &\leq \phi(x) + \langle \nabla_x^1 \phi(x), s \rangle + \sum_{\ell=2}^p \kappa_{\ell, \omega} \|s\|_{\mathcal{V}}^{\ell} + \sum_{\ell=1}^k L_{\ell, \omega} \|s\|_{\mathcal{V}}^{\beta_{\ell}}.
 \end{aligned}$$

Rearranging the last equation, we obtain that

$$\phi(x+s) \leq \phi(x) + \langle \nabla_x^1 \phi(x), s \rangle + \sum_{i=1}^m \kappa_{i, \omega} \|s\|_{\mathcal{V}}^{\gamma_i},$$

where the exponents  $\gamma_i$  are in ascending order and strictly larger than one. We now use the reflexivity of  $\mathcal{V}$  to choose a  $d \in \mathcal{V}$  that verifies both  $\langle \nabla_x^1 \phi(x), d \rangle = \|\nabla_x^1 \phi(x)\|_{\mathcal{V}'}$  and  $\|d\|_{\mathcal{V}} = 1$ , we choose  $s = -td$  in the last inequality so that (2.2) follows.  $\square$

Looking at the steepest descent direction for minimizing (2.1), we are now lead to consider (2.2) and to characterize the minima of functions of the form

$$\Psi(t) \stackrel{\text{def}}{=} -\alpha t + \sum_{i=1}^m \kappa_i t^{\gamma_i}, \tag{2.3}$$

for  $t \in \mathbb{R}_+$ ,  $\alpha > 0$ ,  $\kappa_i > 0$  and  $1 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_m$ . This is the object of the next lemma.

**Lemma 2.2** A function of the form (2.3) admits a unique minimum  $t^*$  over  $\mathbb{R}_+$  and

$$\Psi(t^*) \leq -\min(\kappa_A \alpha^{\frac{\gamma_1}{\gamma_1-1}}, \kappa_B \alpha^{\frac{\gamma_m}{\gamma_m-1}}), \tag{2.4}$$

where  $\kappa_A$  and  $\kappa_B$  depend on  $\{\kappa_i\}_{i=1}^m$ .

**Proof.** Let us consider  $\Psi$  of the form (2.3). Clearly,  $\Psi$  is a strictly convex function as a sum of a linear function and a positive linear combination of powers strictly exceeding one. In addition,  $\Psi'(0) < 0$  and  $\Psi'(t) > 0$  for  $t \in \mathbb{R}_+$  sufficiently large. Thus, a unique

positive minimizer  $t^*$  exists such that  $\Psi'(t^*) = 0$ . Suppose first that  $t^* \geq 1$ . Our problem then reduces to the minimization of  $\Psi$  for  $t \geq 1$ . Define

$$t_1 \stackrel{\text{def}}{=} \left( \frac{\alpha}{\sum_{i=1}^m \kappa_i \gamma_i} \right)^{\frac{1}{\gamma_m - 1}} > 0. \quad (2.5)$$

Because  $t^* \geq 1$  and  $\psi'$  is a non decreasing function, we obtain that  $\psi'(1) \leq \psi'(t^*) = 0$  and thus that

$$-\alpha + \sum_{i=1}^m \kappa_i \gamma_i \leq 0,$$

which, together with the definition of  $t_1$  in (2.5), implies that  $(t_1)^{\gamma_m - 1} \geq 1$ , and the inequality  $\gamma_m > 1$  then ensures  $t_1 \geq 1$ . Using now the assumption that  $1 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_m$ , we deduce that, for  $t \geq 1$ ,

$$\Psi(t) \leq -\alpha t + \sum_{i=1}^m \kappa_i t^{\gamma_i},$$

and thus, since  $t_1 \geq 1$ ,

$$\begin{aligned} \Psi(t_1) &\leq -\alpha^{\frac{\gamma_m}{\gamma_m - 1}} \left( \sum_{i=1}^m \kappa_i \gamma_i \right)^{\frac{-1}{\gamma_m - 1}} + \alpha^{\frac{\gamma_m}{\gamma_m - 1}} \frac{\sum_{i=1}^m \kappa_i}{\left( \sum_{i=1}^m \kappa_i \gamma_i \right)^{\frac{\gamma_m}{\gamma_m - 1}}}, \\ &\leq -\alpha^{\frac{\gamma_m}{\gamma_m - 1}} \left( \sum_{i=1}^m \kappa_i \gamma_i \right)^{\frac{-1}{\gamma_m - 1}} \left( 1 - \frac{\sum_{i=1}^m \kappa_i}{\sum_{i=1}^m \kappa_i \gamma_i} \right), \\ &\stackrel{\text{def}}{=} -\alpha^{\frac{\gamma_m}{\gamma_m - 1}} \kappa_A. \end{aligned}$$

As  $\kappa_i > 0$  and  $\gamma_i > 1$  for all  $i$ , we obtain that  $\kappa_A > 0$  and hence  $\Psi(t^*) \leq \Psi(t_1) \leq -\kappa_A \alpha^{\frac{\gamma_m}{\gamma_m - 1}}$ , which corresponds to the first term in the minimum of (2.4).

Suppose now that  $t^* \leq 1$  and define

$$t_2 \stackrel{\text{def}}{=} \left( \frac{\alpha}{\sum_{i=1}^m \kappa_i \gamma_i} \right)^{\frac{1}{\gamma_1 - 1}} > 0. \quad (2.6)$$

As  $\psi'(t^*) = 0$ ,  $t^* \leq 1$  and  $1 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_m$ ,

$$-\alpha + \sum_{i=1}^m \kappa_i \gamma_i (t^*)^{\gamma_i - 1} = 0,$$

and thus

$$-\alpha + \sum_{i=1}^m \kappa_i \gamma_i (t^*)^{\gamma_1 - 1} \geq 0,$$

which, with the definition of  $t_2$  in (2.6) gives that

$$(t^*)^{\gamma_1-1} \geq t_2^{\gamma_1-1}.$$

The inequality  $\gamma_1 > 1$  then ensures that  $t_2 \leq t^* \leq 1$ . Using an argument similar to that used above but now for the case  $t^* \leq 1$ , we deduce that, for all  $t \leq 1$ ,

$$\Psi(t) \leq -\alpha t + \sum_{i=1}^m \kappa_i t^{\gamma_1},$$

and therefore, since  $t_2 \leq 1$ ,

$$\begin{aligned} \Psi(t_2) &\leq -\alpha^{\frac{\gamma_1}{\gamma_1-1}} \left( \sum_{i=1}^m \kappa_i \gamma_i \right)^{\frac{-1}{\gamma_1-1}} + \alpha^{\frac{\gamma_1}{\gamma_1-1}} \frac{\sum_{i=1}^m \kappa_i}{\left( \sum_{i=1}^m \kappa_i \gamma_i \right)^{\frac{\gamma_1}{\gamma_1-1}}}, \\ &\leq -\alpha^{\frac{\gamma_1}{\gamma_1-1}} \left( \sum_{i=1}^m \kappa_i \gamma_i \right)^{\frac{-1}{\gamma_1-1}} \left( 1 - \frac{\sum_{i=1}^m \kappa_i}{\sum_{i=1}^m \kappa_i \gamma_i} \right), \\ &\stackrel{\text{def}}{=} -\alpha^{\frac{\gamma_1}{\gamma_1-1}} \kappa_B. \end{aligned}$$

Rewriting the last line gives that  $\Psi(t^*) \leq \Psi(t_2) \leq -\kappa_B \alpha^{\frac{\gamma_1}{\gamma_1-1}}$ , which completes the proof.  $\square$

This result suggest the following algorithm for minimizing functions of the form (2.1).

**Algorithm 2.1: A First-Order Gradient Algorithm for Minimizing Regularized Polynomials**

**Step 0: Initialization.** Set  $x_0 = 0$  and  $k = 0$ .

**Step 1: Compute a search direction.** Compute  $\nabla_x^1 \phi(x_k) \in \mathcal{V}'$  and 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  $\|d_k\| = 1$ . If  $\|\nabla_x^1 \phi(x_k)\|_{\mathcal{V}'} = 0$ , stop and return the sequence  $(x_0, x_1, \dots, x_k)$ .

**Step 2: Stepsize definition.** Compute  $t_k$  a global minimizer of  $\phi(x_k - t d_k)$ .

**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.

Note that the selection of  $d_k$  in Step 1 is possible because  $\mathcal{V}$  is reflexive, and that the minimization in Step 2 is possible because it occurs in a one-dimensional space.

We now prove the following convergence result.

**Theorem 2.3** Suppose that  $\phi$ ,  $h$  and  $\mathcal{V}$  verify **AS.1** and let  $\{x_k\}_{k \geq 0}$  be the sequence generated by Algorithm 2.1. Then

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

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

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

**Proof.** Recall that  $\mathcal{D} = \{x \in \mathcal{V} \mid \phi(x) \leq \phi(x_0) = \phi(0)\}$  and that inequality (2.2) is valid if  $x \in \mathcal{D}$ . Since the left hand-side of the inequality (2.2) for  $x = x_0$  verifies the conditions of Lemma 2.2, and denote by  $t_0^*$  the minimizer of Lemma 2.2. We may apply this lemma and deduce that,

$$\phi(x_1) \leq \phi(x_0 - t_0^* d_0) \leq \phi(0) - \min(\kappa_A \|\nabla_x^1 \phi(0)\|_{\mathcal{V}'}^{\frac{\gamma_1}{\gamma_1-1}}, \kappa_B \|\nabla_x^1 \phi(0)\|_{\mathcal{V}'}^{\frac{\gamma_m}{\gamma_m-1}}),$$

where now  $\kappa_A$  and  $\kappa_B$  are strictly positive and depend on  $\omega$  (the radius of  $\mathcal{D}$ ) and the Lipschitz constant  $L_{i,\omega}$ , themselves depending on  $\omega$ . As  $\|\nabla_x^1 \phi(x_0)\|_{\mathcal{V}'} > 0$ ,  $\phi(x_1) < \phi(x_0) = \phi(0)$  and therefore  $x_1 \in \mathcal{D}$ .

Suppose now that  $x_{k-1} \in \mathcal{D}$  and that  $\|\nabla_x^1 \phi(x_{k-1})\|_{\mathcal{V}'} > 0$ . We may again apply Lemma 2.2 to the left handside of inequality (2.2) with  $x$  chosen as  $x_{k-1}$  and by denoting  $t_{k-1}^*$  the minimizer of the left hand-side, we deduce that

$$\phi(x_k) \leq \phi(x_{k-1} - t_{k-1}^* d_{k-1}) \leq \phi(x_{k-1}) - \min(\kappa_A \|\nabla_x^1 \phi(x_{k-1})\|_{\mathcal{V}'}^{\frac{\gamma_1}{\gamma_1-1}}, \kappa_B \|\nabla_x^1 \phi(x_{k-1})\|_{\mathcal{V}'}^{\frac{\gamma_m}{\gamma_m-1}}),$$

thus  $x_k$  and the complete sequence  $\{x_k\}_{k \geq 0}$  belong to  $\mathcal{D}$  and the first conclusion of the theorem holds. To prove the second part, 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

$$\|\nabla_x^1 \phi(x_{k_i})\|_{\mathcal{V}'} \geq \epsilon, \tag{2.7}$$

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

$$\begin{aligned} +\infty > \phi(0) - \phi_{\min} &\geq \sum_i \min(\kappa_A \|\nabla_x^1 \phi(x_{k_i})\|_{\mathcal{V}'}^{\frac{\gamma_1}{\gamma_1-1}}, \kappa_B \|\nabla_x^1 \phi(x_{k_i})\|_{\mathcal{V}'}^{\frac{\gamma_m}{\gamma_m-1}}), \\ &\geq \sum_i \min[\kappa_A \epsilon^{\frac{\gamma_1}{\gamma_1-1}}, \kappa_B \epsilon^{\frac{\gamma_m}{\gamma_m-1}}], \end{aligned}$$

which is a contradiction since the right-hand side diverges to  $+\infty$ . Hence (2.7) cannot hold and the second conclusion of the theorem is valid.  $\square$

Thus a vanilla gradient-descent algorithm applied to a  $p$ -th degree polynomial augmented by a convex regularization term with Hölder gradient will yield asymptotic first-order stationarity.

### 3 An adaptive regularization algorithm in Banach spaces

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

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

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  $f_{\text{low}}$  such that  $f(x) \geq f_{\text{low}}$  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

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

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  $p$ .

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

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  $\epsilon$ -approximate first-order critical point, that is a point  $x_{\epsilon}$  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

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

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  $\lim_{t \rightarrow 0} \frac{\rho_{\mathcal{V}}(t)}{t} = 0$ . 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

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

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<sup>(1)</sup> that  $\rho_{\mathcal{V}}(t) \leq \max(1, \kappa_{\mathcal{V}}) t^{q'}$  from definition (3.4) and inequality (3.5).

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

We motivate our choice of this particular class of Banach spaces by giving a few examples.  $L^p(\mathbb{R})$ ,  $1 < p < \infty$ , are uniformly smooth Banach spaces. In particular,  $L^p(\mathbb{R})$  is uniformly 2 smooth for  $p \geq 2$  and uniformly  $p$  smooth for  $1 < p \leq 2$ . The same results apply for  $\ell^p$  and the Sobolev spaces  $W_m^p(\mathbb{R})$  [27]. Moreover, all Hilbert spaces are 2 smooth Banach.

**Lemma 3.1** Let  $\mathcal{H}$  be a Hilbert space. Then  $\mathcal{H}$  is a 2 smooth Banach space with

$$\rho_{\mathcal{H}}(t) \leq \frac{t^2}{2}. \quad (3.6)$$

**Proof.** Because of the definition of  $\rho_{\mathcal{H}}$  in (3.4), we have that

$$\begin{aligned} \rho_{\mathcal{H}}(t) &= \sup \left\{ \frac{\|x+y\|_{\mathcal{V}} + \|x-y\|_{\mathcal{V}}}{2} - 1, \|x\|_{\mathcal{V}} = 1, \|y\|_{\mathcal{V}} = t \right\}, \\ &= \sup \left\{ \frac{\sqrt{1+t^2+2\langle y, x \rangle} + \sqrt{1+t^2-2\langle y, x \rangle}}{2} - 1, \|x\|_{\mathcal{V}} = 1, \|y\|_{\mathcal{V}} = t \right\}. \end{aligned}$$

Thus, when maximizing over  $\langle y, x \rangle \in [-t, t]$ ,

$$\rho_{\mathcal{H}}(t) = \sqrt{1+t^2} - 1 = \frac{t^2}{\sqrt{1+t^2} + 1} \leq \frac{t^2}{2}.$$

□

One might wonder if it is possible for the  $q$  smooth order to be strictly superior to 2 in (3.5). We now show that this is impossible. Indeed, for any Banach space  $\mathcal{V}$ , we have that,  $\rho_{\mathcal{V}}(t) \geq \rho_{\mathcal{H}}(t) = \frac{t^2}{\sqrt{1+t^2}+1}$  [27]. Suppose now  $\rho_{\mathcal{V}}(t) \leq ct^m$  with  $m > 2$ . Using the last two inequalities, we obtain that:  $ct^{m-2} \geq \frac{1}{\sqrt{1+t^2}+1}$  for all  $t$  strictly positive. But this inequality is impossible for small enough  $t$  and hence our supposition about  $m$  is false and  $m \in (1, 2]$ .

From here on, we assume that

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

Uniformly smooth Banach spaces are also reflexive (See [27, Proposition 1.e.3, p61]), so that **AS.1**(iii) automatically holds. Let us now define the set

$$J_p(x) \stackrel{\text{def}}{=} \left\{ v^* \in \mathcal{V}^*, \langle v^*, x \rangle = \|x\|_{\mathcal{V}}^p, \|v^*\|_{\mathcal{V}'} = \|x\|_{\mathcal{V}}^{p-1} \right\}. \quad (3.7)$$

It is known [26] 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 3.2** Let

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

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

$$\|J_p(x) - J_p(y)\|_{\mathcal{V}'} \leq \max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}})^{p-1} \varphi_p \left( \frac{\|x - y\|_{\mathcal{V}}}{\max(\|x\|_{\mathcal{V}}, \|y\|_{\mathcal{V}})} \right). \quad (3.8)$$

**Proof.** [27, Theorem 2]. □

As we will be only working with  $\|\cdot\|_{\mathcal{V}}^p$  for  $p > 1$  in the rest of the paper, we define  $J_p(x)$  as the unique value in the set (3.7). As the subdifferential of  $\|\cdot\|_{\mathcal{V}}^p$  reduces to a singleton for  $p > 1$  and  $\|\cdot\|_{\mathcal{V}}^p$  is a convex function,  $\|\cdot\|_{\mathcal{V}}^p$  is Fréchet differentiable for  $p > 1$  since it verifies [14, Condition 4.16]. The reader is referred to [26] or [27] 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 3.3** 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

$$\|J_{\ell}(x) - J_{\ell}(y)\|_{\mathcal{V}'} \leq \kappa_{\omega} \|x - y\|_{\mathcal{V}}^{\min[q, \ell]-1} + \kappa_{\ell} \|x - y\|_{\mathcal{V}}^{\ell-1}, \quad (3.9)$$

where  $\kappa_{\omega}$  and  $\kappa_{\ell}$  depend only on  $\omega, \ell, \kappa_{\mathcal{F}}$  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 3.2 and use the definition of the set  $\mathcal{F}$  to derive that

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

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

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

□

### 3.2 The AR $p$ -BS algorithm

Adaptive regularization methods are iterative schemes which compute a step from 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

$$m_k(s) \stackrel{\text{def}}{=} T_{f,p}(x_k, s) + \frac{\sigma_k}{(p + \beta)!} \|s\|_{\mathcal{V}}^{p + \beta}, \quad p \geq 1. \quad (3.10)$$

As in [11] but at variance with [12], we will assume here that  $\beta$ , 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  $\sigma_k$  is known as the “regularization parameter”. This term guarantees that the functional  $m_k(s)$  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 (3.10). 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 AR $p$ -BS algorithm on the next page.

While the AR $p$ -BS algorithm follows the main lines of existing AR $p$  methods [8, 6]. Because we are in an infinite dimensional space, the existence of a minimizer of  $m_k(s)$  may not be guaranteed and hence a point  $s^*$  such that  $\nabla_s^1 m_k(s^*) = 0$  may not exist. As a consequence, standard proofs that a step satisfying both (3.13) and (3.14) exists no longer apply. We thus need to check that this is still the case in our context. This is achieved using Algorithm 2.1.

**Theorem 3.4** Suppose that **AS.2**, **AS.4** and **AS.5** hold. Suppose also that  $\|g(x_k)\|_{\mathcal{V}'} > 0$ . Then a step satisfying both (3.13) and (3.14) always exists.

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

**Step 0: Initialization:** An initial point  $x_0 \in \mathcal{V}$ , a regularization parameter  $\sigma_0$  and a requested final gradient accuracy  $\epsilon \in (0, 1]$  are given. The constants  $\eta_1, \eta_2, \gamma_1, \gamma_2, \gamma_3, \chi \in (0, 1)$ , and  $\sigma_{\min}$  are also given such that

$$\sigma_{\min} \in (0, \sigma_0], 0 < \eta_1 \leq \eta_2 < 1 \quad \text{and} \quad 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \quad (3.11)$$

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

**Step 1: Check for termination:** Terminate with  $x_\epsilon = x_k$  if

$$\|g(x_k)\|_{\mathcal{V}'} \leq \epsilon. \quad (3.12)$$

**Step 2: Step calculation:** Compute a step  $s_k$  which sufficiently reduces the model  $m_k$  in the sense that

$$m_k(s_k) < m_k(0), \quad (3.13)$$

and

$$\|\nabla_s^1 m_k(s_k)\|_{\mathcal{V}'} \leq \max \left[ \chi \epsilon, \theta \|s_k\|_{\mathcal{V}}^{p+\beta-1} \right]. \quad (3.14)$$

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

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_{f,p}(x_k, 0) - T_{f,p}(x_k, s_k)}. \quad (3.15)$$

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

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1 \sigma_k), \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2 \sigma_k, \gamma_3 \sigma_k] & \text{if } \rho_k < \eta_1. \end{cases} \quad (3.16)$$

Increment  $k$  by one and go to Step 1.

**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 (3.10), we just need to prove that  $m_k(s)$  satisfies **AS.1** of Section 2. We have that

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

and thus  $m_k$  is a coercive functional verifying **AS.1(i)**. Lemma 3.3 (applied with  $k = 2$ ,  $\delta = \omega$ ,  $\ell = p + \beta$ ,  $L_{1,\delta} = \kappa_\ell$ ,  $\beta_1 = \min[q, \ell] \in (1, 2]$ ,  $L_{2,\delta} = \kappa_\omega$  and  $\beta_2 = \ell > 1$ ) then ensures that  $\|\cdot\|_{\mathcal{V}}^{p+\beta}$  satisfies **AS.1(ii)**. We already noted that, being uniformly smooth,  $\mathcal{V}$  must be reflexive, which ensures that **AS.1(iii)** 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 2.3 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.  $\square$

Observe that equation (2.2) and the fact that  $\gamma_1 = \min[q, p + \beta]$  and  $\gamma_m = p + \beta$  (all the other powers ranging from 2 to  $p$ ), imply that, for our iterative gradient descent,

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

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

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

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 3.5** Suppose that the AR $_p$ -BS algorithm is used and that  $\sigma_k \leq \sigma_{\max}$  for some  $\sigma_{\max} > 0$ . Then

$$k \leq |\mathcal{S}_k| \left( 1 + \frac{|\log \gamma_1|}{\log \gamma_2} \right) + \frac{1}{\log \gamma_2} \log \left( \frac{\sigma_{\max}}{\sigma_0} \right). \quad (3.17)$$

**Proof.** See [6, Theorem 2.4].  $\square$

## 4 Evaluation complexity for the AR $_p$ -BS algorithm

Before discussing our analysis of evaluation complexity, we first restate some classical lemmas of AR $_p$  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

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

and

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

**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.  $\square$

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

**Lemma 4.2**

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

**Proof.** Direct from (3.13) and (3.10).  $\square$

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

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

**Proof.** See [6, Lemma 2.2]. Using (3.15), (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 (3.16) implies that  $\sigma_{k+1} \leq \sigma_k$ . The mechanism of the algorithm then guarantees that (4.4) holds.  $\square$

The next lemma remains in the spirit of [6, Lemma 2.3], but now takes the condition (3.14) 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

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

**Proof.** Successively using the fact that termination does not occur at iteration  $k$  and condition (3.14), 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).  $\square$

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 AR $_p$ -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  $\{\nabla_x^i f\}_{i=1,2,\dots,p}$  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

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

**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

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

Thus

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

for any  $k$  before termination. The first conclusion follows since the derivatives are only evaluated once per successful iteration. Applying now Lemma 3.5 gives the second conclusion.  $\square$

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  $L^p$ ,  $\ell^p$  and  $W_m^p$  are uniform  $q$  smooth spaces for  $1 < p < \infty$ , and hence that Lemma 3.3 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.

## 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  $\epsilon$ -approximate first-order critical point in at most  $\mathcal{O}(\epsilon^{-\frac{p+\beta}{p+\beta-1}})$  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 [9] 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

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

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 Bergman 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. Deuffhard 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] 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.
- [22] J. M. Martínez. On high-order model regularization for constrained optimization. *SIAM Journal on Optimization*, 27:2447–2458, 2017.
- [23] Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming, Series A*, 108(1):177–205, 2006.
- [24] 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.
- [25] 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.
- [26] H.-K. Xu. Inequalities in Banach spaces with applications. *Nonlinear Analysis: Theory, Methods & Applications*, 16(12):1127–1138, 1991.
- [27] 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.