ORIGINAL PAPER

The randomized block coordinate descent method in the Hölder smooth setting

Leandro Farias Maia[1] · **David Huckleberry Gutman[1]**

Abstract

This work provides the first convergence analysis for the Randomized Block Coordinate Descent method for minimizing a function that is both Hölder smooth and block Hölder smooth. Our analysis applies to objective functions that are non-convex, convex, and strongly convex. For non-convex functions, we show that the expected gradient norm reduces at an $\mathcal{O}\left(k^{\frac{\gamma}{1+\gamma}}\right)$ rate, where k is the iteration count and γ is the Hölder exponent. For convex functions, we show that the expected suboptimality gap reduces at the rate $\mathcal{O}(k^{-\gamma})$. In the strongly convex setting, we show this rate for the expected suboptimality gap improves to $\mathcal{O}\left(k^{-\frac{2\gamma}{1-\gamma}}\right)$ when $\gamma > 1$ and to a linear rate when $\gamma = 1$. Notably, these new convergence rates coincide with those furnished in the existing literature for the Lipschitz smooth setting.

Keywords Randomized coordinate descent · Hölder · Convergence

1 Introduction

In this article, we provide non-asymptotic convergence rates for the Randomized Block Coordinate Descent (RBCD) method when applied to the problem

$$f^* := \min_{x \in \mathbb{R}^d} f(x), \tag{1}$$

where the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is Hölder smooth, a generalization of the standard (Lipschitz) smoothness, and block Hölder smooth. Formally, the

✉ Leandro Farias Maia
leandro.maia@tamu.edu

David Huckleberry Gutman
dhgutman@tamu.edu

[1] Department of Industrial and Systems Engineering, Texas A&M University, Emerging Technologies Building, 3131 TAMU, 101 Bizzell Street, College Station, TX 77843, USA

continuously differentiable function, f, is said to be *Hölder smooth* when its gradient, ∇f, is Hölder continuous, i.e. there exist $L > 0$ and $\gamma \in (0, 1]$ guaranteeing

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|^\gamma \quad \text{for all } x, y \in \mathbb{R}^d. \tag{2}$$

The popularity of block coordinate methods owes to their fitness for large-scale optimization problems emerging from applications in machine learning and statistics. Essentially, randomized block coordinate descent is a (random) block-wise adaptation of gradient descent. Instead of updating all coordinates simultaneously, the randomized block coordinate descent method updates a single, randomly selected coordinate block using only that block's partial gradient. The computational economy of these block gradient updates, relative to full gradient updates, are what make the randomized block coordinate descent method especially attractive for large-scale problems.

Given an initial point x^0, this cheap iterate update rule is somewhat more generally realized as

$$x^k = x^{k-1} - t_k \cdot P_{i_k} \nabla f(x^{k-1}), \quad k = 1, 2, \ldots, \tag{3}$$

where $t_k > 0$, i_k is selected randomly from $\{1, \ldots, m\}$, and $P_1, \ldots, P_m \in \mathbb{R}^{d \times d}$ are orthogonal projection matrices onto orthogonal subspaces that sum to $\mathbb{R}^d$. The "block coordinate" name originates from the archetypal choice for the orthogonal subspaces projected onto: spans of collections of coordinate vectors.

For coordinate descent methods, and indeed a preponderance of first-order methods, the intimate relationship between the selection of step-sizes and ∇f's regularity determines their convergence rates [2–7, 9, 12, 13]. Both Bredies [1] and Yashtini [13] study the interplay between step-size selection and convergence for gradient descent applied to (1) in the Hölder smooth regime. Bredies [1] established a $O(1/k^\gamma)$ convergence rate of the proximal gradient method, a generalization of the gradient descent method, for convex composite minimization. On the other hand, Yashtini [13] established that, given an appropriate step-size selection, gradient descent converges at a $O\left(1/k^{\frac{\gamma}{1+\gamma}}\right)$ for non-convex, Hölder smooth objective functions.

We are unaware of any studies of the randomized block coordinate descent method that assume Hölder smoothness or its block-wise adaptation, block Hölder smoothness. We say that the continuously differentiable function, f, is *block Hölder smooth* if for each $i = 1, \ldots, m$, there exists $L_i > 0$ such that

$$\|\nabla f(x + P_i u) - \nabla f(x)\| \leq L_i \|P_i u\|^\gamma \text{ for all } u \in \mathbb{R}^d. \tag{4}$$

The seminal articles [8, 11] studying the randomized block coordinate descent method all make the more restrictive assumption that the gradient is Lipschitz continuous. Recently, inspired by the work of both Bredies [1] and Yashtini [13], Gutman and Ho-Nguyen [3] produced a convergence analysis for the cyclic block coordinate descent method assuming Hölder and block Hölder smoothness in both the convex and non-convex settings. Thus, the goal of this paper is to extend this

analysis to the more popular randomized block coordinate descent method in the non-convex, convex, and even strongly convex settings.

We conclude this introduction with an outline of our article that includes a high-level description of each of our main contributions. This article is structured into four primary sections:

- *Section* 2 In this section, we introduce our RBCD step-size selection for Hölder smooth objective functions as well as the attendant notation. We also introduce two key lemmas (Lemmas 2.1 and 2.2) that support our analyses.
- *Section* 3 In this section, we present our convergence analysis for general, possibly non-convex objective functions satisfying Hölder and block Hölder smoothness conditions. For these objectives, our proposed step-size ensures RBCD shrinks the expected gradient norm at a $\mathcal{O}\left(1/k^{\frac{\gamma}{1+\gamma}}\right)$ rate (Theorem 3.1).
- *Section* 4 In this section, we present our convergence analysis under the further assumption that the objective function is convex. In this setting, RBCD with our step-size shrinks the expected suboptimality gap at a $O(1/k^\gamma)$ rate for non-strongly convex objective functions (Theorem 4.1). Notably, our rates for these objective functions coincide with those of [8] when $\gamma = 1$, or equivalently, when the objective is L-smooth.
- *Section* 5 In this section, we present our analysis under the further assumption that the objective function is strongly convex. This analysis depends upon the value of the Hölder exponent, γ. When $\gamma = 1$, we show RBCD converges at a linear rate (Theorem 5.1). When $\gamma \in (0, 1)$, we obtain a $\mathcal{O}\left(1/k^{\frac{2\gamma}{1-\gamma}}\right)$ rate of convergence (Theorem 5.1). Moreover, we show that our sublinear rates converge to our linear rates as $\gamma \to 1$ (Corollary 5.1). As for convex objectives, our rates for strongly convex objectives coincide with those of [8] when the objective is L-smooth.

2 Notation and step-size selection for RBCD under Hölder smoothness

This short section introduces the notation necessary for all of this article's developments, and the Hölder smoothness-based step-size selection for the RBCD method. It also exhibits two lemmas, Lemmas 2.1 and 2.2, that are used throughout the paper to aid the convergence analysis of the proposed method.

Our step-size selection is an adaptation of that used for the cyclic block setting from [3] to the immensely more popular randomized block setting. Thus, our notation is a synthesis of that article's notation and the notation of [8], one of the canonical works on randomized coordinate descent. We let $\tilde{L} := \{L_1, \dots, L_m\}$ denote the set of the block Hölder smoothness constants. For $\alpha \in \mathbb{R}$, we define the new constant, $S_\alpha(f)$, as

$$S_\alpha(f) = \sum_{i=1}^{m} L_i^\alpha.$$

When f is clear from context, we will simply write S_α. For the sake of concision, we set $\nabla_i f(x) := P_i \nabla f(x)$ for all $x \in \mathbb{R}^d$ and $1 \le i \le m$. We adopt the notation, $\nu := \frac{1+\gamma}{\gamma} > 1$, because the quantity $\frac{1+\gamma}{\gamma}$ frequently appears in our analysis.

Much of our analysis is framed in terms of $\tilde{L}$-weighted q-norms on $\mathbb{R}^d$, $\|\cdot\|_{\alpha,q}$. Given $\alpha \in \mathbb{R}$ and $q \ge 1$, we let

$$\|x\|_{\alpha,q} := \left[\sum_{i=1}^{m} L_i^\alpha \|P_i x\|^q \right]^{1/q}. \tag{5}$$

When $\alpha = 0$, $\|\cdot\|_{\alpha,q}$ reduces to the standard q-norm, which we write as $\|\cdot\|_q$. For simplicity, we let $\|\cdot\| := \|\cdot\|_2$. We bare three important notes about these weighted norms. First, $\|\cdot\|_{\alpha,q}$ generalizes the norm

$$x \mapsto \left[\sum_{i=1}^{m} L_i^\alpha \|P_i x\|^2 \right]^{1/2},$$

which plays a starring role throughout Nesterov's classical analysis of randomized coordinate descent methods from [8] in the block Lipschitz smooth setting. The flexibility provided by changing the exponents 2 and 1/2 to q and $1/q$, respectively, is critical to capturing our more general Hölderian convergence rates. Additionally, the parameter α permits us to simultaneously achieve RBCD's convergence rates for two different, common random block selection schemes:

(i) $\alpha = 0$ corresponds to selecting the blocks uniformly at random;
(ii) $\alpha = 1$ corresponds to selecting the i-th block with probability $L_i / \sum_{i=1}^{m} L_i$.

Finally, these weighted norms possess natural duality relationships and equivalences to the Euclidean norm, which we liberally use throughout our analysis and summarize in the below lemma.

Lemma 2.1 ((α, q)-Norm Duality and Equivalences) *Let $\alpha \in \mathbb{R}$, $p \in [1, \infty]$, and q be the Hölder conjugate of p, i.e. $q := \frac{p}{p-1}$. The following hold for $\|\cdot\|_{\alpha,p}$:*

1. *The Cauchy–Schwarz inequality*

$$|\langle x, y \rangle| \le \|x\|_{\alpha,q} \|y\|_{-\alpha\frac{p}{q},q} \tag{6}$$

   holds for all $x, y \in \mathbb{R}^d$. Equality is obtained if and only if $x = 0$ or

$$y = c \cdot \sum_{i=1}^{m} L_i^{\alpha} \|P_i x\|^{p-2} P_i x$$

for some $c \in \mathbb{R}$. Consequently, $\| \cdot \|_{-\alpha \frac{p}{q}, q}$ is the dual norm of $\| \cdot \|_{\alpha, q}$.

2. *If $p \geq 2$ and $\alpha, \beta \in \mathbb{R}$ then the norms $\| \cdot \|_{\alpha, p}$ and $\| \cdot \|_{\beta, 2}$ satisfy*

$$\left(\max_{1 \leq i \leq m} L_i^{\frac{\alpha}{p} - \frac{\beta}{2}} \right) \cdot \|x\|_{\beta, 2} \geq \|x\|_{\alpha, p} \geq \left(m^{\frac{1}{p} - \frac{1}{2}} \cdot \min_{1 \leq i \leq m} L_i^{\frac{\alpha}{p} - \frac{\beta}{2}} \right) \cdot \|x\|_{\beta, 2}$$

for all $x \in \mathbb{R}^d$.

We defer the proof of this lemma to the appendix ("Appendix 1") to maintain the focus of our exposition.

With all of the article's requisite notation in hand, we may introduce our main algorithm (Algorithm 1), and describe an associated descent lemma (Lemma 2.2). We note that our step-size, $-\|\nabla_i f(x^k)\|^{\nu-2}/L_i^{\nu-1}$, coincides with that proposed in [8] when $\gamma = 1$. Thus, we may view it as a generalization that accounts for the use of block Hölder smoothness in the place of standard block smoothness.

Algorithm 1 Randomized block coordinate descent method (RBCD)

Data: $x^0 \in \text{dom}(f)$, $\alpha \in [0, 1]$

for $k=0,1,2,\ldots$ **do**

> Choose
> $$i_k \sim (p_1, \ldots, p_m) := \left(\frac{L_1^{\alpha}}{\sum_{j=1}^{m} L_j^{\alpha}}, \ldots, \frac{L_m^{\alpha}}{\sum_{j=1}^{m} L_j^{\alpha}} \right);$$
>
> Update block i_k of x^k according to
>
> $$x^{k+1} := x^k - \frac{\|\nabla_i f(x^k)\|^{\nu-2}}{L_i^{\nu-1}} \cdot \nabla_i f(x^k). \tag{7}$$

end

A special case of the main descent lemma of [1], derived in [3], plays the same role in our analysis that it played for the cyclic block analysis in [3]. We directly quote this special case from [3] below.

Lemma 2.2 (Block Hölder Descent Lemma, [3], Lemma 1) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that satisfies the block Hölder smoothness condition. For any $i, 1 \leq i \leq m$,*

$$f(x + U_i y) \leq f(x) + \langle \nabla_i f(x), U_i y \rangle + \frac{L_i}{1 + \gamma} \|U_i y\|_2^{1+\gamma}. \tag{8}$$

Moreover, if x^+ is the minimizer of the right-hand side of (8), *i.e.*

$$x^+ = x - \frac{\|\nabla_i f(x)\|^{\nu-2}}{L_i^{\nu-1}} \cdot \nabla_i f(x),$$

then

$$f(x) - f(x^+) \geq \frac{1}{\nu L_i^{\nu-1}} \|\nabla_i f(x)\|^{\nu}.$$

3 Convergence analysis: general objectives

In this section, we layout our convergence rate analysis for non-convex objectives satisfying Hölder smoothness (2) and block Hölder smoothness (4). We will present the main convergence theorem (Theorem 3.1) after we elaborate our key Sufficient Decrease Lemma (Lemma 3.1). This lemma facilitates all of our convergence analyses.

Lemma 3.1 *(Sufficient Decrease) Let $\{x_n\}_{n=0}^{\infty}$ be the sequence generated by RBCD (Algorithm 1). If f satisfies our Hölder smoothness (2) and block Hölder smoothness (4) assumptions, then*

$$\frac{1}{\nu S_\alpha(f)} \|\nabla f(x_k)\|_{\alpha+1-\nu,\nu}^{\nu} \leq f(x_k) - \mathbb{E}\left[f(x^{k+1})\Big| x^k\right] \tag{9}$$

holds for all $k \geq 0$.

Proof Expanding the expectation-defining sum, and applying the block descent lemma (Lemma 2.2), we compute

$$
\begin{aligned}
f(x_k) - \mathbb{E}\left[f(x^{k+1})\Big| x^k\right] &= \mathbb{E}\left[f(x_k) - f(x^{k+1})\Big| x^k\right] \\
&= \sum_{i=1}^{m} \left(\frac{L_i^\alpha}{\sum_{j=1}^{m} L_j^\alpha}\right) \cdot \left[f(x_k) - f\left(x^k - \frac{\|\nabla_i f(x^k)\|^{\nu-2}}{L_i^{\nu-1}} \cdot \nabla_i f(x^k)\right)\right] \\
&\overset{\text{Lemma 2.2}}{\geq} \frac{1}{\nu S_\alpha} \sum_{i=1}^{m} L_i^{\alpha+1-\nu} \|\nabla_i f(x^k)\|^{\nu} = \frac{1}{\nu S_\alpha} \|\nabla f(x_k)\|_{\alpha+1-\nu,\nu}^{\nu}.
\end{aligned}
$$

Rearranging the inequality and taking total expectations completes the proof. $\qquad\square$

Next, we present the centerpiece of this section, our main convergence theorem for non-convex objective functions.

Theorem 3.1 (RBCD Convergence: General Objective Functions) *Let* $\{x_n\}_{n=0}^\infty$ *be the sequence generated by RBCD* (*Algorithm* 1). *If f satisfies our Hölder smoothness* (2) *and block Hölder smoothness* (4) *assumptions, then*

$$\min_{0 \le j \le k} \mathbb{E}\left[\|\nabla f(x^j)\|_{1+\alpha-\nu,\nu}\right] \le \left(\nu S_\alpha(f)\right)^{\frac{1}{\nu}} \cdot \left(\frac{f(x^0) - f^*}{k+1}\right)^{\frac{1}{\nu}} = \mathcal{O}\left(k^{-\frac{1}{\nu}}\right)$$

holds for all $k \ge 0$. *Consequently, we have the convergence rate measured in the norm* $\|\cdot\|_{\beta,2}$,

$$\min_{0 \le j \le k} \mathbb{E}\left[\|\nabla f(x^j)\|_{\beta,2}\right] \le \left(\frac{\max\limits_{1 \le i \le m} L_i^{\frac{\beta}{2} - \frac{1+\alpha-\nu}{\nu}}}{m^{\frac{\nu-2}{2\nu}}}\right) \cdot \left(\nu S_\alpha(f)\right)^{\frac{1}{\nu}} \cdot \left(\frac{f(x^0) - f^*}{k+1}\right)^{\frac{1}{\nu}} = \mathcal{O}\left(k^{-\frac{1}{\nu}}\right),$$

holds for all $k \ge 0$.

Proof For each $k \ge 0$, observe that

$$\min_{0 \le j \le k} \mathbb{E}\left[\|\nabla f(x^j)\|_{\alpha+1-\nu,\nu}\right]^\nu \le \frac{1}{(k+1)} \cdot \sum_{j=0}^k \mathbb{E}\left[\|\nabla f(x^j)\|_{\alpha+1-\nu,\nu}\right]^\nu$$

$$\le \frac{1}{(k+1)} \cdot \sum_{j=0}^k \mathbb{E}\left[\|\nabla f(x^j)\|_{\alpha+1-\nu,\nu}^\nu\right] \qquad (10)$$

$$\le \nu S_\alpha \cdot \frac{1}{(k+1)} \cdot \sum_{j=0}^k \left(\mathbb{E}\left[f(x^j)\right] - \mathbb{E}\left[f(x^{j+1})\right]\right)$$

$$= \nu S_\alpha \cdot \frac{f(x^0) - \mathbb{E}[f(x^{k+1})]}{k+1} \le \nu S_\alpha \cdot \frac{f(x^0) - f^*}{k+1}, \qquad (11)$$

where we apply Jensen's inequality to the expectation operator for the convex function $x \mapsto x^\nu$ in (10), and Lemma 3.1 in (10). Taking ν-th roots of both sides of the resultant inequality above, produces our first result.

The result in terms of the $\|\cdot\|_{\beta,2}$ follows immediately from Lemma 2.1. □

4 Convergence analysis: convex objectives

In this section, we forward our convergence analysis of RBCD (Theorem 4.1) for convex objective functions that are both Hölder (2) and block Hölder (4) smooth. First, we present a Technical Recurrence Lemma (Lemma 4.1) that helps produce our convergence rates in this section, and a subset of the convergence rates for strongly convex objective functions in the sequel. Next, we exhibit a techical

lemma (Lemma 4.2) that permits us to express our rates in terms of the diameter of the initial sublevel set. Finally, the section concludes with our main convergence theorem (Theorem 4.1) and a comparison of these rates to those furnished for smooth and convex functions in [8].

As promised, we begin this section with a Technical Recurrence Lemma that supports the derivation of our convergence rates.

Lemma 4.1 (Technical Recurrence, [10, Chapter 2, Lemma 6]) *If* $\{A_k\}_{k\geq 0}$ *is a nonnegative sequence of real numbers satisfying the recurrence*

$$A_{k+1} \leq A_k - \theta A_k^r$$

for some $\theta \geq 0$ *and* $r > 1$, *then*

$$A_k \leq \frac{A_0}{(1 + (r-1)\theta A_0^{r-1}k)^{\frac{1}{r-1}}}.$$

The following lemma permits us to express our convergence rates here and in the sequel section in terms of the initial sublevel set's diameter.

Lemma 4.2 *Under the Block Hölder Smoothness assumption* (4) *and coercivity of f, f satisfies*

$$f(x) - f^* \leq \left(\frac{\nu S_\alpha(f)}{2}\right)^{\frac{1}{\nu-1}} \cdot \left(\frac{\nu-1}{\nu}\right) \cdot R(x)^{\frac{\nu}{\nu-1}}_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}}$$

for all $x \in \mathbb{R}^d$, *where* $R_{\beta,q}(x) := \max\left\{\|y - x^*\|_{\beta,q} : f(x^*) = f^*, f(y) \leq f(x)\right\} < \infty$.

We defer the proof of this lemma to the appendix ("Appendix 1").

Finally, equipped with these tools, we present and prove the theorem that establishes RBCD's convergence rate for convex functions. Afterward, we explain its relationship to its analogue for smooth and convex functions in [8].

Theorem 4.1 (RBCD Convergence: Convex Objective Functions) *Let* $\{x_n\}_{n=1}^\infty$ *be the sequence generated by RBCD (Algorithm* 1). *If f is a convex and coercive function that satisfies our Hölder smoothness* (2) *and block Hölder smoothness* (4) *assumptions, then*

$$\mathbb{E}[f(x^k)] - f^* \leq \frac{\left(\nu S_\alpha(f) R_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}}(x^0)^\nu\right)^{\frac{1}{\nu-1}}(\nu-1)}{\left[2\nu^{\nu-1} + (\nu-1)^\nu k\right]^{\frac{1}{\nu-1}}} = \mathcal{O}\left(k^{-\frac{1}{\nu-1}}\right),$$

where $R_{\beta,q}(x^0) := \max_y\left\{\|y - x^*\|_{\beta,q} : f(x^*) = f^*, f(y) \leq f(x^0)\right\} < \infty$.

Proof The bulk of this proof centers on an application of the Technical Recurrence Lemma (Lemma 4.1). In the context of that lemma, we let $A_i = \mathbb{E}[f(x^i)] - f^*$ for

each $i \geq 0$. By definition, $A_i \geq 0$ for each $i \geq 0$. To simplify notation, we let $R := R_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}}(x^0)$.

With this notation, we may restate the sufficient decrease inequality (9) of Lemma 3.1 as

$$\mathbb{E}\left[\|\nabla f(x_k)\|_{\alpha+1-\nu,\nu}^{\nu}\right] \leq \nu S_\alpha \cdot \left(A_k - A_{k+1}\right),$$

or, equivalently,

$$A_{k+1} \leq A_k - \frac{1}{\nu S_\alpha}\mathbb{E}\left[\|\nabla f(x_k)\|_{\alpha+1-\nu,\nu}^{\nu}\right]. \tag{12}$$

Thus, to apply the Technical Recurrence Lemma (Lemma 4.1) we need only bound the expectation on the right below by A_k^ν. By the Cauchy–Schwarz inequality (Lemma 2.1) for $\|\cdot\|_{1+\alpha-\nu,\nu}$ and its dual $\|\cdot\|_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}}$, we achieve for any optimum x^*, that

$$f(x^k) - f^* \leq \langle x^k - x^*, \nabla f(x^k)\rangle$$
$$\leq \|x^k - x^*\|_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}} \|\nabla f(x^k)\|_{1+\alpha-\nu,\nu} \leq R\|\nabla f(x^k)\|_{1+\alpha-\nu,\nu}.$$

Raising each side of the above inequality to the power ν, taking expectations, and applying Jensen's inequality to the convex function $x \mapsto x^{\nu/2}$, we conclude

$$A_k^\nu = \left(\mathbb{E}[f(x^k)] - f^*\right)^\nu \leq \mathbb{E}\left[\left(f(x^k) - f^*\right)^\nu\right]$$
$$\leq R^\nu \mathbb{E}\left[\|\nabla f(x^k)\|_{1+\alpha-\nu,\nu}^\nu\right].$$

Stringing together our work above, equation (12) yields the recurrence

$$A_{k+1} \leq A_k - \frac{1}{\nu S_\alpha R^\nu}A_k^\nu$$

for each $k \geq 0$. We are now permitted to apply the Technical Recurrence Lemma (Lemma 4.1) with $r = \nu$ and $\theta = \frac{1}{\nu S_\alpha R^\nu}$ to produce

$$\mathbb{E}[f(x^k)] - f^* \leq \frac{f(x^0) - f^*}{\left[1 + (\nu-1)\cdot\theta\cdot\left(f(x^0)-f^*\right)^{\nu-1}\cdot k\right]^{\frac{1}{\nu-1}}}.$$

We dedicate the remainder of this proof to simplifying this convergence bound. By factoring $f(x^0) - f^*$ out of both the numerator and denominator, we may equivalently write the right-hand side of this bound as

$$\frac{1}{\left[\frac{1}{(f(x^0)-f^*)^{\nu-1}} + (\nu-1)\cdot\theta\cdot k\right]^{\frac{1}{\nu-1}}}.$$

By considering $x = x^0$ in Lemma 4.2, raising both sides to the power $\nu - 1$ and applying the norm equivalence inequality from Lemma 2.1, we further see

$$\left[f(x^0) - f(x^*) \right]^{\nu-1} \leq \left(\frac{\nu S_\alpha}{2} \right) \cdot \left(\frac{\nu - 1}{\nu} \right)^{\nu-1} \cdot R^\nu$$

so the right-hand side of our bound simplifies to

$$\left[\frac{1}{\left(\frac{\nu S_\alpha}{2} \right) \cdot \left(\frac{\nu-1}{\nu} \right)^{\nu-1} \cdot R^\nu} + (\nu - 1) \cdot \left(\frac{1}{\nu S_\alpha R^\nu} \right) \cdot k \right]^{-\frac{1}{\nu-1}} = \frac{\left(\nu S_\alpha R^\nu \right)^{\frac{1}{\nu-1}} (\nu - 1)}{\left(2\nu^{\nu-1} + (\nu - 1)\nu k \right)^{\frac{1}{\nu-1}}},$$

which concludes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Notably, our rate matches that provided by Nesterov [8] in the standard block smooth setting, i.e. when $\nu = 2$ we recover the convergence rate,

$$\mathbb{E}[A_{k+1}] \leq \frac{2}{k+4} S_\alpha(f) R^2(x_1) = \mathcal{O}(k^{-1}),$$

from [8].

5 Convergence analysis: strongly convex objectives

In this final section, we conclude the paper with a convergence analysis of RBCD (Algorithm 1) for strongly convex objective functions that are both Hölder (2) and block Hölder (4) smooth. We say that $f : \mathbb{R}^d \to \mathbb{R}$ is σ-strongly convex with respect to the norm $\| \cdot \|_{1-\alpha,2}$, where $\sigma > 0$, if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\sigma \|x - y\|_{1-\alpha,2}^2 \qquad (13)$$

for all $x, y \in \mathbb{R}^d$. The section begins with our main theorem (Theorem 5.1), which provides rates in both the L-smooth and Hölder smooth settings. Next, we compare these rates with those in the previous section and [8]. Finally, we show that the smooth setting's linear rate is achieved in the limit as $\nu \to 2$, or equivalently, $\gamma \to 1$ (Corollary 5.1).

Without further ado, we present our main convergence theorem for strongly convex objective functions.

Theorem 5.1 (RBCD Convergence: Strongly Convex Objective Functions) *Let $\{x_n\}_{n=1}^\infty$ be the sequence generated by RBCD (Algorithm 1). Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is σ-strongly convex and satisfies both the Hölder and block Hölder smoothness assumptions (2) and (4). The following hold*:

1. (*Linear Rate—Smooth Setting*) If $v = 2$, *i.e. f is smooth, then*

$$
\begin{aligned}
\mathbb{E}[f(x_k)] - f^* &\leq \left(1 - \frac{\sigma}{S_\alpha(f)}\right)^k \cdot \frac{S_\alpha(f)^{\frac{1}{v-1}}(v-1)R(x^0)^{\frac{v}{v-1}}}{v^{\frac{v-2}{v}}2^{\frac{1}{v-1}}} \\
&= \mathcal{O}\left(\exp\left(-\frac{\sigma}{S_\alpha(f)}k\right)\right).
\end{aligned}
\tag{14}
$$

2. (*Sublinear Rate—Hölder Smooth Setting*) If $v > 2$, *i.e. f is Hölder smooth but not smooth, then*

$$
\mathbb{E}[f(x^k)] - f^* \leq \frac{C_0}{\left(C_1 + C_2 k\right)^{\frac{2}{v-2}}} = \mathcal{O}\left(k^{-\frac{2}{v-2}}\right),
$$

where

$$
C_0 = (2vS_\alpha(f))^{\frac{2}{v-2}} m^{\frac{1}{v}}(v-1)R(x^0)^{\frac{v}{v-1}}, \quad C_1 = 2^{\frac{v-2}{2(v-1)}} m^{\frac{v-2}{2v}} S_\alpha(f)^{\frac{v-1}{v}} v^{\frac{v^2-2v+4}{2v}}
$$

$$
C_2 = R(x^0)^{\frac{v(v-2)}{2(v-1)}}(v-1)^{\frac{v-2}{2}}(v-2)(2\sigma)^{\frac{v}{2}} \min_{1\leq i\leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}}
$$

Proof Let $R := R_{(1+\alpha-v)(1-v),\frac{v}{v-1}}(x^0)$ to simplify notation. Both parts of the theorem speedily follow from the recurrence

$$
A_{k+1} \leq A_k - A_k^{\frac{v}{2}} \cdot \left(\frac{(2\sigma)^{\frac{v}{2}} \min_{1\leq i\leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}}}{vS_\alpha m^{\frac{1}{2}-\frac{1}{v}}}\right),
\tag{15}
$$

where $A_i = \mathbb{E}[f(x^i)] - f^*$ for each $i \geq 0$. After establishing (15), we will separately show how each of the Theorem's two parts result from it.

As in the proof of convergence for non-strongly convex functions, the sufficient decrease inequality (9) of Lemma 3.1 implies (12), which we recall is

$$
A_{k+1} \leq A_k - \frac{1}{vS_\alpha}\mathbb{E}\left[\|\nabla f(x_k)\|_{\alpha+1-v,v}^v\right].
$$

Glancing at (15) and this latest inequality, it becomes immediately clear that we ought to bound $\mathbb{E}\left[\|\nabla f(x_k)\|_{\alpha+1-v,v}^v\right]$ below by $A_k^{\frac{v}{2}} = \mathbb{E}\left[f(x_k) - f^*\right]^{v/2}$, appropriately scaled. To this end, strong convexity now makes it's main appearance. Using the standard argument of fixing $x \in \mathbb{R}^d$ in σ-strong convexity's defining inequality (13) and minimizing it over $y \in \mathbb{R}^d$, we achieve the Polyak–Łojasiewicz (PL) inequality

$$
\frac{1}{2\sigma}\left(\|\nabla f(x)\|_{1-\alpha,2}^*\right)^2 \geq f(x) - f^*.
$$

Setting $x = x_k$, raising both sides to the power $v/2$, and then taking expectations, we see

$$\frac{1}{(2\sigma)^{\frac{v}{2}}} \mathbb{E}\left[\left(\|\nabla f(x_k)\|_{1-\alpha,2}^*\right)^v\right] \geq \mathbb{E}\left[(f(x_k) - f^*)^{v/2}\right],$$

which, by Jensen's inequality applied to the convex function $x \mapsto x^{v/2}$, produces

$$\frac{1}{(2\sigma)^{\frac{v}{2}}} \mathbb{E}\left[\left(\|\nabla f(x_k)\|_{1-\alpha,2}^*\right)^v\right] \geq \left(\mathbb{E}\left[f(x_k) - f^*\right]\right)^{v/2} = A_k^{\frac{v}{2}} \qquad (16)$$

The main bound (15) is secured by twice applying the (α, q)-Norm Duality Equivalence Lemma (Lemma 2.1) to connect the recurrence inequality (12) and the PL-derived bound (16),

$$A_k - A_{k+1} \overset{(12)}{\geq} \frac{1}{vS_\alpha} \mathbb{E}\left[\|\nabla f(x_k)\|_{\alpha+1-v,v}^v\right]$$

$$\overset{\text{Lemma 2.1}}{\geq} \frac{1}{vS_\alpha} \cdot \left(\frac{\min\limits_{1\leq i \leq m} L_i^{\frac{\alpha+1-v}{v} - \frac{\alpha-1}{2}}}{m^{\frac{1}{2}-\frac{1}{v}}}\right) \cdot \mathbb{E}\left[\|\nabla f(x_k)\|_{\alpha-1,2}^v\right]$$

$$\overset{\text{Lemma 2.1}}{=} \frac{1}{vS_\alpha} \cdot \left(\frac{\min\limits_{1\leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}}}{m^{\frac{1}{2}-\frac{1}{v}}}\right) \cdot \mathbb{E}\left[\left(\|\nabla f(x_k)\|_{1-\alpha,2}^*\right)^v\right]$$

$$\overset{(16)}{\geq} \frac{1}{vS_\alpha} \cdot \left(\frac{\min\limits_{1\leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}}}{m^{\frac{1}{2}-\frac{1}{v}}}\right) \cdot \left[(2\sigma)^{\frac{v}{2}} A_k^{\frac{v}{2}}\right].$$

Now, we are prepared to prove the theorem's two constituent parts.

1. If $v = 2$, then the main recurrence inequality (15) becomes

$$A_{k+1} \leq A_k - A_k \cdot \left(\frac{\sigma}{S_\alpha}\right) = A_k\left(1 - \frac{\sigma}{S_\alpha}\right),$$

which by backward induction is equivalent to our desired bound,

$$\mathbb{E}[f(x_k)] - f^* = A_k \leq \left(1 - \frac{\sigma}{S_\alpha}\right)^k \cdot A_0 = \left(1 - \frac{\sigma}{S_\alpha}\right)^k \cdot [f(x_0) - f^*]$$

$$\leq \left(1 - \frac{\sigma}{S_\alpha}\right)^k \cdot \frac{S_\alpha^{\frac{1}{v-1}} (v-1) R^{\frac{v}{v-1}}}{v^{\frac{v-2}{v}} 2^{\frac{1}{v-1}}}, \qquad (17)$$

where we have applied Lemma 4.2 in the first line.

2. The $v > 2$ result requires a verification that is as straightforward as, but more tedious than, that of 1. Applying the Technical Recurrence Lemma (Lemma 4.1) with $r = v/2$, and $\theta = \frac{(2\sigma)^{\frac{v}{2}}}{v S_\alpha m^{\frac{v-2}{2v}}} \cdot \min_{1 \leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}}$ we see

$$\mathbb{E}[f(x^k)] - f^* \leq \frac{f(x^0) - f^*}{\left[1 + \left(\frac{v-2}{2}\right) \cdot \frac{(2\sigma)^{\frac{v}{2}}}{v S_\alpha m^{\frac{v-2}{2v}}} \cdot \min_{1 \leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}} \cdot \left(f(x^0) - f^*\right)^{\frac{v-2}{2}} \cdot k\right]^{\frac{2}{v-2}}}. \tag{18}$$

This intermediate form of our convergence rate will facilitate the proof of our later convergence rate interpolation result (Corollary 5.1) so we have labeled it. For now though, we focus on processing this expression of the rate into its final form. The first step is to simply re-arrange this to

$$\mathbb{E}[f(x^k)] - f^* \leq \frac{(2v S_\alpha)^{\frac{2}{v-2}} m^{\frac{1}{v}} (f(x^0) - f^*)}{\left[2v S_\alpha m^{\frac{v-2}{2v}} + (v-2)(2\sigma)^{\frac{v}{2}} \min_{1 \leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}} \cdot \left(f(x^0) - f^*\right)^{\frac{v-2}{2}} k\right]^{\frac{2}{v-2}}}.$$

By factoring $f(x^0) - f^*$ out of both the numerator and denominator, and by applying Lemma 4.2 to the previous expression, its right-hand side simplifies to

$$= \frac{(2v S_\alpha)^{\frac{2}{v-2}} m^{\frac{1}{v}}}{\left[\frac{v S_\alpha m^{\frac{v-2}{2v}}}{(f(x^0) - f^*)^{\frac{v-2}{2}}} + (v-2)(2\sigma)^{\frac{v}{2}} \min_{1 \leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}} k\right]^{\frac{2}{v-2}}}$$

$$\overset{\text{Lemma 4.2}}{\leq} \frac{(2v S_\alpha)^{\frac{2}{v-2}} m^{\frac{1}{v}}}{\left[\frac{v S_\alpha m^{\frac{v-2}{2v}}}{\left(\frac{S_\alpha^{\frac{v-2}{2(v-1)}} (v-1)^{\frac{v-2}{2}} R^{\frac{v(v-2)}{2(v-1)}}}{v^{\frac{(v-2)^2}{2v}} 2^{\frac{v-2}{2(v-1)}}}\right)} + (v-2)(2\sigma)^{\frac{v}{2}} \min_{1 \leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}} k\right]^{\frac{2}{v-2}}}$$

$$= \frac{(2v S_\alpha)^{\frac{2}{v-2}} m^{\frac{1}{v}}}{\left[\frac{2^{\frac{v-2}{2(v-1)}} m^{\frac{v-2}{2v}} S_\alpha^{\frac{v-1}{v}} v^{\frac{v^2-2v+4}{2v}}}{(v-1)^{\frac{v-2}{2}} R^{\frac{v(v-2)}{2(v-1)}}} + (v-2)(2\sigma)^{\frac{v}{2}} \min_{1 \leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}} k\right]^{\frac{2}{v-2}}}$$

$$= \frac{(2v S_\alpha)^{\frac{2}{v-2}} m^{\frac{1}{v}} (v-1) R^{\frac{v}{v-1}}}{\left[2^{\frac{v-2}{2(v-1)}} m^{\frac{v-2}{2v}} S_\alpha^{\frac{v-1}{v}} v^{\frac{v^2-2v+4}{2v}} + R^{\frac{v(v-2)}{2(v-1)}} (v-1)^{\frac{v-2}{2}} (v-2)(2\sigma)^{\frac{v}{2}} \min_{1 \leq i \leq m} L_i^{\frac{(\alpha+1)(2-v)}{2v}} k\right]^{\frac{2}{v-2}}}$$

□

We now make two crucial comparisons for the rates above. First, it is noteworthy that when $v = 2$ in the strongly-convex regime, we recover the same linear rate as that in [8]. Second, our strongly convex sublinear rate in the $v > 2$ setting,

$\mathcal{O}(k^{-\frac{2}{\nu-2}})$, is indeed faster than the $\mathcal{O}(k^{-\frac{1}{\nu-1}})$ rate occurring in the merely convex case.

To conclude this article, we demonstrate that, in the strongly convex case, when $\nu \to 2$ the intermediate form (18) of the strongly convex sublinear rate of Theorem 5.1 converges to its $\nu = 2$ linear rate of convergence.

Corollary 5.1 *(Interpolation of Linear and Sublinear Rate) In the strongly convex setting of Theorem 5.1, if $\nu \to 2$, then the sublinear rate converges to a linear rate of convergence. More formally, for the convergence bound*

$$\mathbb{E}[f(x^k)] - f^* \leq \frac{f(x^0) - f^*}{\left[1 + \left(\frac{\nu-2}{2}\right) \cdot \frac{(2\sigma)^{\frac{\nu}{2}}}{\nu S_\alpha m^{\frac{\nu-2}{2\nu}}} \cdot \min_{1\leq i\leq m} L_i^{-\frac{(\alpha+1)(2-\nu)}{2\nu}} \cdot \left(f(x^0) - f^*\right)^{\frac{\nu-2}{2}} \cdot k\right]^{\frac{2}{\nu-2}}},$$

for $k \geq 0$, we observe the limiting result

$$\lim_{\nu \to 2^+} \frac{f(x^0) - f^*}{\left[1 + \left(\frac{\nu-2}{2}\right) \cdot \frac{(2\sigma)^{\frac{\nu}{2}}}{\nu S_\alpha m^{\frac{\nu-2}{2\nu}}} \cdot \min_{1\leq i\leq m} L_i^{-\frac{(\alpha+1)(2-\nu)}{2\nu}} \cdot \left(f(x^0) - f^*\right)^{\frac{\nu-2}{2}} \cdot k\right]^{\frac{2}{\nu-2}}}$$

$$\leq \left(1 - \frac{\sigma}{2S_\alpha}\right)^k \cdot \left(f(x^0) - f^*\right)$$

Proof The form of the sublinear convergence bound here was established by the immediately preceeding theorem in equation (18). The use of said rate in the proof of this corollary was foreshadowed there.

To prove our main limit result, suppose for a moment that

$$\lim_{x \to 0^+} \left[1 + g(x) \cdot x\right]^{-\frac{1}{x}} = e^{-g(0)} \tag{19}$$

holds for any continuously differentiable $g : [0, \infty) \to [0, \infty)$ such that $g(0) > 0$. Restating our sublinear convergence rate, we see that

$$\frac{\mathbb{E}[f(x^k)] - f^*}{f(x^0) - f^*} \leq \left[1 + g\left(\frac{\nu-2}{2}\right) \cdot \left(\frac{\nu-2}{2}\right)\right]^{\frac{2}{2-\nu}},$$

where

$$g(x) = \frac{(2\sigma)^{x+1}}{(2x+2)S_\alpha m^{\frac{x}{2x+2}}} \cdot \min_{1\leq i\leq m} L_i^{-\frac{(\alpha+1)x}{2x+2}} \cdot A_0^x \cdot k.$$

Observe that g is continuous differentiable and $g(0) = \sigma/S_\alpha > 0$. Thus, it follows that

$$\frac{\mathbb{E}[f(x^k)] - f^*}{f(x^0) - f^*} \leq \lim_{\nu \to 2^+} \left[1 + g\left(\frac{\nu - 2}{2}\right) \cdot \left(\frac{\nu - 2}{2}\right) \right]^{\frac{2}{2-\nu}}$$

$$\overset{(19)}{=} e^{-g(0)} = \left(e^{-\frac{\sigma}{S_\alpha}} \right)^k$$

$$\leq \left(\frac{1}{1 + \frac{\sigma}{S_\alpha}} \right)^k \leq \left(1 - \frac{\sigma}{2S_\alpha} \right)^k,$$

where we applied the standard inequalities $e^x \geq 1 + x$ and $(1 + x)^{-1} < (1 - x/2)$ for all $x > 0$ in the last two lines. Thus, we only need to prove (19) to finish the proof.

The proof is a simply straightforward computation:

$$\lim_{x \to 0^+} \left[1 + g(x) \cdot x \right]^{-\frac{1}{x}} = \lim_{x \to 0^+} \exp\left(-\frac{\ln\left[1 + g(x) \cdot x \right]}{x} \right)$$

$$= \exp\left(-\lim_{x \to 0^+} \frac{\ln\left[1 + g(x) \cdot x \right]}{x} \right)$$

$$= \exp\left(-\lim_{x \to 0^+} \frac{g(x) + x \cdot g'(x)}{1 + x \cdot g(x)} \right)$$

$$= \exp(-g(0)),$$

where continuity of $x \mapsto e^x$ is used in the second line, l'Hôpital's rule is used in the third line, and the definition of g is used in the final line. $\qquad\square$

Appendix 1: Proof of Lemma 2.1

In this section of the technical appendix, we prove Lemma 2.1.

1. We begin by choosing $x, y \in \mathbb{R}^d$. The inequality is trivial if $x = 0$ so we assume $x \neq 0$. By the standard and p-norm versions of the Cauchy–Schwarz inequality, we compute

$$\langle x, y \rangle = \sum_{i=1}^m \langle P_i x, P_i y \rangle \leq \sum_{i=1}^m \|P_i x\| \|P_i y\| = \sum_{i=1}^m \left(L_i^{\alpha/p} \|P_i x\| \right) \left(\frac{\|P_i y\|}{L_i^{\alpha/p}} \right)$$

$$= \left\langle \left(L_1^{\alpha/p} \|P_1 x\|, \dots, L_1^{\alpha/p} \|P_1 x\| \right), \left(\frac{\|P_1 y\|}{L_1^{\alpha/p}}, \dots, \frac{\|P_m y\|}{L_m^{\alpha/p}} \right) \right\rangle$$

$$\leq \left\| \left(L_1^{\alpha/p} \|P_1 x\|, \dots, L_m^{\alpha/p} \|P_m x\| \right) \right\|_p \left\| \left(\frac{\|P_1 y\|}{L_1^{\alpha/p}}, \dots, \frac{\|P_m y\|}{L_m^{\alpha/p}} \right) \right\|_q = \|x\|_{\alpha,p} \|y\|_{-\alpha\frac{p}{q},q}$$

By the standard Cauchy–Schwarz inequality, the first inequality is obtained with equality if and only if one of $P_i x$ and $P_i y$ is a scalar multiple of the other

for each $i = 1, \dots, m$. By the Cauchy–Schwarz inequality for p-norms, and our assumption that $x \neq 0$, the second inequality obtains equality if and only if there is some $c \in \mathbb{R}$ such that

$$c^q \cdot \left(L_1^\alpha \|P_1 x\|^p, \dots, L_m^\alpha \|P_m x\|^p \right) = \left(\frac{\|P_1 y\|^q}{L_1^{\alpha q/p}}, \dots, \frac{\|P_m y\|^q}{L_m^{\alpha q/p}} \right).$$

Assuming both inequalities hold then, we conclude there are $c_1, \dots, c_m, c \in \mathbb{R}$ such that $P_i y = c_i \cdot P_i x$ and $c^q \cdot L_i^\alpha \|P_i x\|^p = \frac{\|P_i y\|^q}{L_i^{\alpha q/p}}$ for $i = 1, \dots, m$. Fixing $1 \leq i \leq m$, and combining the equalities, we see that $c^q \cdot L_i^\alpha \|P_i x\|^p = c_i^q \cdot \frac{\|P_i x\|^q}{L_i^{\alpha q/p}}$, so

$$c_i = c \cdot L_i^{\alpha\left(1 - \frac{q}{p}\right)} \|P_i x\|^{\frac{p}{q} - 1} = c \cdot L_i^\alpha \|P_i x\|^{p-2}$$

where we use the definition of q as p's Hölder conjugate to produce the second equality. This completes the proof of 1.

2. Given $x \in \mathbb{R}^d$, consider the vector $(\|P_j x\|_j)_{j=1}^m$. For any $p \geq 2$, the norm equivalence inequality yields

$$\|x\|_{\alpha, p} = \left(\sum_{j=1}^m L_j^\alpha \|P_j x\|_2^p \right)^{\frac{1}{p}} = \left(\sum_{j=1}^m \|L_j^{\alpha/p} \cdot P_j x\|_2^p \right)^{\frac{1}{p}} \leq \left(\sum_{j=1}^m \|L_j^{\alpha/p} \cdot P_j x\|_2^2 \right)^{\frac{1}{2}}$$

But,

$$\left(\sum_{j=1}^m \|L_j^{\alpha/p} \cdot P_j x\|_2^2 \right)^{\frac{1}{2}} = \left(\sum_{j=1}^m L_j^{2\alpha/p} \|P_j x\|_2^2 \right)^{\frac{1}{2}} \leq \|x\|_{\beta, 2} \cdot \max_{1 \leq j \leq m} L_m^{\frac{\alpha}{p} - \frac{\beta}{2}}$$

as, for any i, with $1 \leq i \leq m$, $L_i^{2\alpha/p} \leq L^\beta \left(\max_{1 \leq j \leq m} L_m^{\frac{\alpha}{p} - \frac{\beta}{2}} \right)^2$ and we complete the first part of the inequality. The second part is quite similar. By the equivalence norm inequality,

$$\|x\|_{\beta, 2} = \left(\sum_{j=1}^m L_j^\beta \|P_j x\|_2^2 \right)^{\frac{1}{2}} = \left(\sum_{j=1}^m \|L_j^{\beta/2} \cdot P_j x\|_2^2 \right)^{\frac{1}{2}} \leq m^{\frac{1}{2} - \frac{1}{p}} \cdot \left(\sum_{j=1}^m \|L_j^{\beta/2} \cdot P_j x\|_2^p \right)^{\frac{1}{p}}$$

But

$$\left(\sum_{j=1}^m \|L_j^{\beta/2} \cdot P_j x\|_2^p \right)^{\frac{1}{p}} = \left(\sum_{j=1}^m L_j^{\beta p/2} \|P_j x\|_2^p \right)^{\frac{1}{p}} \leq \|x\|_{\alpha, p} \cdot \max_{1 \leq j \leq m} L_j^{\frac{\beta}{2} - \frac{\alpha}{p}} = \frac{\|x\|_{\alpha, p}}{\min\limits_{1 \leq j \leq m} L_j^{\frac{\alpha}{p} - \frac{\beta}{2}}}$$

and we are done.

Appendix 2: Proof of Lemma 4.2

Suppose we are able to prove for all $x, y \in \mathbb{R}^d$ that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \cdot \left(\frac{\nu - 1}{\nu} \right) \cdot \|u\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}} \quad (20)$$

Then x^*, the first-order condition $\nabla f(x^*)$ naturally holds so (20) implies

$$\begin{aligned}
f(x) - f(x^*) &\leq \langle \nabla f(x^*), x - x^* \rangle + \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \cdot \left(\frac{\nu - 1}{\nu} \right) \cdot \|x - x^*\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}} \\
&= \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \cdot \left(\frac{\nu - 1}{\nu} \right) \cdot \|x - x^*\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}} \\
&\leq \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \cdot \left(\frac{\nu - 1}{\nu} \right) \cdot R(x)_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}}
\end{aligned}$$

and this completes the proof. Thus, it suffices to prove (20). Suppose that

$$\frac{2}{\nu S_\alpha} \|\nabla f(x) - \nabla f(y)\|_{1+\alpha-\nu, \nu}^{\nu-1} \leq \|x - y\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}} \quad (21)$$

holds for all $x, y \in \mathbb{R}^d$. Then, given $u \in \mathbb{R}^d$, the integral formulation of the mean value theorem states

$$f(x + u) - f(x) = \int_0^1 \langle \nabla f(x + tu), u \rangle dt,$$

Thus, the Cauchy–Schwartz inequality and our previous inequality imply

$$\begin{aligned}
f(x + u) - f(x) - \langle \nabla f(x), u \rangle &= \int_0^1 \langle \nabla f(x + tu) - \nabla f(x), u \rangle dt \\
&\leq \int_0^1 \|\nabla f(x + tu) - \nabla f(x)\|_{1+\alpha-\nu, \nu} \|u\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}} dt \\
&\leq \int_0^1 \left(\frac{\nu S_\alpha}{2} \|tu\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}} \right)^{\frac{1}{\nu-1}} \|u\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}} dt \\
&= \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \|u\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}} \int_0^1 t^{\frac{1}{\nu-1}} dt \\
&= \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \cdot \left(\frac{\nu - 1}{\nu} \right) \cdot \|u\|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}},
\end{aligned}$$

so taking $u = y - x$ in (21) completes the proof. Consequently, we now need only prove the Hölderian co-coercivity condition (21), which we do presently.

Given $y \in \mathbb{R}^d$, the function $x \mapsto \phi(x) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle$ is readily seen to be Hölder block smooth because f. Moreover, ϕ has the same block Hölder smoothness constants and $\nabla \phi(x) = \nabla f(x) - \nabla f(y)$. Thus, by the Block Hölder Descent Lemma (Lemma 2.2),

$$f(x) - f(y) - \langle \nabla f(y), y - x \rangle = \phi(y) - \min_{z \in \mathbb{R}^d} \phi(z) \geq \max_{1 \leq i \leq m} \left[\frac{1}{\nu L_i^{\nu-1}} \| \nabla_i f(x) - \nabla_i f(y) \|^\nu \right]$$

$$\geq \frac{1}{\nu S_\alpha} \sum_{i=1}^m \frac{L_i^\alpha}{L_i^{\nu-1}} \| \nabla_i f(x) - \nabla_i f(y) \|^\nu = \frac{1}{\nu S_\alpha} \| \nabla f(x) - \nabla f(y) \|_{1+\alpha-\nu,\nu}^\nu,$$

which we may restate as

$$f(x) \geq f(y) + \langle \nabla f(y), y - x \rangle + \frac{1}{\nu S_\alpha} \| \nabla f(x) - \nabla f(y) \|_{1+\alpha-\nu,\nu}^\nu.$$

Adding this inequality to its analogue with the roles of x and y reversed, we see produce

$$\frac{2}{\nu S_\alpha} \| \nabla f(x) - \nabla f(y) \|_{1+\alpha-\nu,\nu}^\nu \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

By Cauchy–Schwarz, we then see that

$$\frac{2}{\nu S_\alpha} \| \nabla f(x) - \nabla f(y) \|_{1+\alpha-\nu,\nu}^\nu \leq \| \nabla f(x) - \nabla f(y) \|_{1+\alpha-\nu,\nu} \| x - y \|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}},$$

or equivalently

$$\frac{2}{\nu S_\alpha} \| \nabla f(x) - \nabla f(y) \|_{1+\alpha-\nu,\nu}^{\nu-1} \leq \| x - y \|_{(1+\alpha-\nu)(1-\nu), \frac{\nu}{\nu-1}}.$$

Given $u \in \mathbb{R}^d$, the integral formulation of the mean value theorem states

$$f(x + u) - f(x) = \int_0^1 \langle \nabla f(x + tu), u \rangle dt,$$

so the Cauchy–Schwarz inequality and the previous inequality imply

$$f(x + u) - f(x) - \langle \nabla f(x), u \rangle = \int_0^1 \langle \nabla f(x + tu) - \nabla f(x), u \rangle dt$$

$$\leq \int_0^1 \|\nabla f(x + tu) - \nabla f(x)\|_{1+\alpha-\nu,\nu} \|u\|_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}} dt$$

$$\leq \int_0^1 \left(\frac{\nu S_\alpha}{2} \|tu\|_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}} \right)^{\frac{1}{\nu-1}} \|u\|_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}} dt$$

$$= \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \|u\|_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}} \int_0^1 t^{\frac{1}{\nu-1}} dt$$

$$= \left(\frac{\nu S_\alpha}{2} \right)^{\frac{1}{\nu-1}} \cdot \left(\frac{\nu - 1}{\nu} \right) \cdot \|u\|_{(1+\alpha-\nu)(1-\nu),\frac{\nu}{\nu-1}}^{\frac{\nu}{\nu-1}}.$$

Taking $u = y - x$ completes the proof.

Data availability No data were used or generated in the writing of this paper.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bredies, K.: A forward–backward splitting algorithm for the minimization of non-smooth convex functionals in Banach space. Inverse Problems **25**(1), 015005 (2008)
2. Grimmer, B.: General Hölder smooth convergence rates follow from specialized rates assuming growth bounds. J. Optim. Theory Appl. **197**(1), 51–70 (2023)
3. Gutman, D.H., Ho-Nguyen, N.: Cyclic coordinate descent in the Hölder smooth setting. Oper. Res. Lett. **50**(5), 458–462 (2022)
4. Lee, M., Shekhar, S., Javidi, T.: Multi-scale zero-order optimization of smooth functions in an RKHS. In: 2022 IEEE International Symposium on Information Theory (ISIT), pp. 288–293. IEEE (2022)
5. Liu, Y., Wang, Y., Singh, A.: Smooth bandit optimization: generalization to Hölder space. In: International Conference on Artificial Intelligence and Statistics, pp. 2206–2214. PMLR (2021)
6. Nemirovskii, A., Nesterov, Y.: Optimal methods of smooth convex minimization. USSR Comput. Math. Math. Phys. **25**, 21–30 (1986)
7. Nemirovskii, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience Series in Discrete Mathematics, Wiley (1983)
8. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
9. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. **152**(1), 381–404 (2015)
10. Polyak, B.: Introduction to Optimization. Optimization Software (1987)
11. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**(1), 1–38 (2014)
12. Shibaev, I., Dvurechensky, P., Gasnikov, A.: Zeroth-order methods for noisy Hölder-gradient functions. Optim. Lett. **16**(7), 2123–2143 (2022)

13. Yashtini, M.: On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients. Optim. Lett. **10**, 1361–1370 (2016)