Blended Matching Pursuit

Cyrille W. Combettes
Georgia Institute of Technology
Atlanta, GA, USA
cyrille@gatech.edu

Sebastian Pokutta
Zuse Institute Berlin and TU Berlin
Berlin, Germany
pokutta@zib.de

Abstract

Matching pursuit algorithms are an important class of algorithms in signal processing and machine learning. We present a *blended matching pursuit* algorithm, combining coordinate descent-like steps with stronger gradient descent steps, for minimizing a smooth convex function over a linear space spanned by a set of atoms. We derive sublinear to linear convergence rates according to the smoothness and sharpness orders of the function and demonstrate computational superiority of our approach. In particular, we derive linear rates for a large class of non-strongly convex functions, and we demonstrate in experiments that our algorithm enjoys very fast rates of convergence and wall-clock speed while maintaining a sparsity of iterates very comparable to that of the (much slower) orthogonal matching pursuit.

1 Introduction

Let $\mathcal{H}$ be a separable real Hilbert space, $\mathcal{D} \subset \mathcal{H}$ be a dictionary, and $f : \mathcal{H} \to \mathbb{R}$ be a smooth convex function. In this paper, we aim at solving the problem:

$$\text{Find a solution to } \min_{x \in \mathcal{H}} f(x) \text{ which is sparse relative to } \mathcal{D}. \tag{1}$$

Together with fast convergence, achieving high sparsity, i.e., keeping the iterates as linear combinations of a *small* number of atoms in the dictionary $\mathcal{D}$, is a primary objective and leads to better generalization, interpretability, and decision-making in machine learning. In signal processing, Problem (1) encompasses a wide range of applications, including compressed sensing, signal denoising, and information retrieval, and is often solved with the Matching Pursuit algorithm [Mallat and Zhang, 1993]. Our approach is inspired by the Blended Conditional Gradients algorithm [Braun et al., 2019] which solves the constrained setting of Problem (1), i.e., minimizing f over the convex hull $\text{conv}(\mathcal{D})$ of the dictionary, and is ultimately based on the Frank-Wolfe algorithm [Frank and Wolfe, 1956] a.k.a. Conditional Gradient algorithm [Levitin and Polyak, 1966]. It enhances the vanilla Frank-Wolfe algorithm by replacing the linear minimization oracle with a *weak-separation oracle* [Braun et al., 2017] and by blending the traditional Frank-Wolfe steps with *lazified* Frank-Wolfe steps and projected gradient steps, while still avoiding projections. Frank-Wolfe algorithms are particularly well-suited for problems with a desired sparsity in the solution (see, e.g., Jaggi [2013] and the references therein) however, from an optimization perspective, although they approximate the optimal descent direction $-\nabla f(x_t)$ via the linear minimization oracle $v_t^{\text{FW}} \leftarrow \arg\min_{\mathcal{D}} \langle \nabla f(x_t), v \rangle$, they move in the direction $v_t^{\text{FW}} - x_t$ in order to ensure feasibility, which provides less progress.

An analogy between Frank-Wolfe algorithms and the unconstrained Problem (1) was proposed by Locatello et al. [2017]. They unified the Frank-Wolfe and Matching Pursuit algorithms, and proposed a Generalized Matching Pursuit algorithm (GMP) and an Orthogonal Matching Pursuit algorithm (OMP) for solving Problem (1), which descend in the directions v_t^{FW}. Essentially, Locatello et al. [2017] established that GMP corresponds to the vanilla Frank-Wolfe algorithm and OMP corresponds to the Fully-Corrective Frank-Wolfe algorithm. GMP and OMP converge with similar rates in the various regimes, namely with a sublinear rate for smooth convex functions and with a linear rate for

smooth strongly convex functions, however they have different advantages: GMP converges (much) faster in wall-clock time while OMP offers (much) sparser iterates. The interest in these algorithms stems from the fact that they work in the general setting of smooth convex functions in Hilbert spaces and that their convergence analyses do not require incoherence or restricted isometry properties (RIP, Candès and Tao [2005]) of the dictionary, which are quite strong assumptions from an optimization standpoint. For an in-depth discussion of the advantages of GMP and OMP over other methods, e.g., in Tropp [2004], Gribonval and Vandergheynst [2006], Davenport and Wakin [2010], Shalev-Shwartz et al. [2010], Temlyakov [2013, 2014, 2015], Tibshirani [2015], Yao and Kwok [2016], and Nguyen and Petrova [2017], we refer the interested reader to Locatello et al. [2017]. In a follow-up work, Locatello et al. [2018] presented an Accelerated Matching Pursuit algorithm.

We aim at unifying the best of GMP (speed) and OMP (sparsity) into a single algorithm by blending them strategically. However, while the overall idea is reasonably natural, we face considerable challenges as many important features of Frank-Wolfe methods do not apply anymore in the Matching Pursuit setting and cannot be as easily overcome as in Locatello et al. [2017], requiring a different analysis. For example, Frank-Wolfe (duality) gaps are not readily available but they are crucial in monitoring the blending, and further key components, such as the weak-separation oracle, require modifications.

Contributions. We propose a *Blended Matching Pursuit* algorithm (BMP), a fast and sparse first-order method for solving Problem (1). Our method unifies the best of GMP (speed) and OMP (sparsity) into one algorithm, which is of fundamental interest for practitioners. We establish a continuous range of convergence rates between $\mathcal{O}(1/\epsilon^p)$ and $\mathcal{O}(\ln 1/\epsilon)$, where $\epsilon > 0$ is the desired accuracy and $p > 0$ depends on the properties of the function. In particular, we derive linear rates of convergence for a large class of smooth convex but non-strongly convex functions. Lastly, we demonstrate the computational superiority of BMP over state-of-the-art methods, with BMP converging the fastest in wall-clock time while maintaining its iterates at close-to-optimal sparsity, and this without requiring sparsity-inducing constraints.

Outline. We introduce notions and notation in Section 2. We present the Blended Matching Pursuit algorithm in Section 3 with the convergence analyses in Section 3.1. Computational experiments are provided in Section 4. Additional experiments and the proofs can be found in the Appendix.

2 Preliminaries

We work in a separable real Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ with induced norm $\| \cdot \|$. A set $\mathcal{D} \subset \mathcal{H}$ of normalized vectors is a *dictionary* if it is at most countable and $\mathrm{cl}(\mathrm{span}(\mathcal{D})) = \mathcal{H}$, and in this case its elements are referred to as *atoms*. For any set $\mathcal{S} \subseteq \mathcal{H}$, let $\mathcal{S}' := \mathcal{S} \cup -\mathcal{S}$ denote the *symmetrization of $\mathcal{S}$* and $D_{\mathcal{S}} := \sup_{u,v \in \mathcal{S}} \|u - v\|$ denote the *diameter of $\mathcal{S}$*. If $\mathcal{S}$ is closed and convex, let $\mathrm{proj}_{\mathcal{S}}$ denote the *orthogonal projection onto $\mathcal{S}$* and $\mathrm{dist}(\cdot, \mathcal{S}) := \| \mathrm{id} - \mathrm{proj}_{\mathcal{S}} \|$ denote the *distance to $\mathcal{S}$*. For Problem (1) to be feasible, we will assume f to be *coercive*, i.e., $\lim_{\|x\| \to +\infty} f(x) = +\infty$. Since f is convex, this is actually a mild assumption when $\arg\min_{\mathcal{H}} f \neq \varnothing$.

Let $f : \mathcal{H} \to \mathbb{R}$ be a Fréchet differentiable function. In the following, we use extended notions of smoothness and strong convexity by introducing *orders*, and we weaken and generalize the notion of strong convexity to that of *sharpness* (see, e.g., Roulet and d'Aspremont [2017] and Kerdreux et al. [2019] for recent work). We say that f is:

(i) *smooth of order $\ell > 1$* if there exists $L > 0$ such that for all $x, y \in \mathcal{H}$,
$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{\ell} \|y - x\|^{\ell},$$

(ii) *strongly convex of order $s > 1$* if there exists $S > 0$ such that for all $x, y \in \mathcal{H}$,
$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geqslant \frac{S}{s} \|y - x\|^{s},$$

(iii) *sharp of order $\theta \in]0, 1[$ on $\mathcal{K}$* if $\mathcal{K} \subset \mathcal{H}$ is a bounded set, $\varnothing \neq \arg\min_{\mathcal{H}} f \subset \mathrm{int}(\mathcal{K})$, and there exists $C > 0$ such that for all $x \in \mathcal{K}$,
$$\mathrm{dist}\left(x, \arg\min_{\mathcal{H}} f\right) \leqslant C \left(f(x) - \min_{\mathcal{H}} f\right)^{\theta}.$$

If needed, we may specify the constants by introducing f as *L-smooth*, *S-strongly convex*, or *C-sharp*. The following fact, whose result was already used in Nemirovskii and Nesterov [1985], provides a bound on the sharpness order of a smooth function. A proof is available in Appendix D.

Fact 2.1. *Let $f : \mathcal{H} \to \mathbb{R}$ be smooth of order $\ell > 1$, convex, and sharp of order $\theta \in]0, 1[$ on $\mathcal{K}$. Then $\theta \in]0, 1/\ell]$.*

2.1 On sharpness and strong convexity

Notice that if $f : \mathcal{H} \to \mathbb{R}$ is Fréchet differentiable and strongly convex of order $s > 1$, then $\operatorname{card}(\arg\min_{\mathcal{H}} f) = 1$. Let $\{x^*\} := \arg\min_{\mathcal{H}} f$. It follows directly from $\nabla f(x^*) = 0$ that for any bounded set $\mathcal{K} \subset \mathcal{H}$ such that $x^* \in \operatorname{int}(\mathcal{K})$, f is sharp of order $\theta = 1/s$ on $\mathcal{K}$. Thus, strong convexity implies sharpness. However, not every sharp function is strongly convex; moreover, the next example shows that not every sharp and convex function is strongly convex.

Example 2.2 (Distance to a convex set). *Let $\mathcal{C} \subset \mathcal{H}$ be a nonempty, closed, and bounded convex set, and $\mathcal{K} \subset \mathcal{H}$ be a bounded set such that $\mathcal{C} \subset \operatorname{int}(\mathcal{K})$. The function $f : x \in \mathcal{H} \mapsto \operatorname{dist}(x, \mathcal{C})^2 = \|x - \operatorname{proj}_{\mathcal{C}}(x)\|^2$ is convex, and it is sharp of order $\theta = 1/2$ on $\mathcal{K}$. Indeed, since $\arg\min_{\mathcal{H}} f = \mathcal{C}$ and $\min_{\mathcal{H}} f = 0$, we have for all $x \in \mathcal{K}$,*

$$\operatorname{dist}\left(x, \arg\min_{\mathcal{H}} f\right) = \|x - \operatorname{proj}_{\mathcal{C}}(x)\| = \left(f(x) - \min_{\mathcal{H}} f\right)^{1/2}.$$

Now, suppose $\mathcal{C}$ contains more than one element. Then, f has more than one minimizer. However, a function that is strongly convex of order $s > 1$ has no more than one minimizer. Therefore, f cannot be strongly convex of order s, for all $s > 1$. Notice that f is also a smooth function, of order $\ell = 2$.

Hence, sharpness is a more general notion of strong convexity. It is a local condition around the optimal solutions while strong convexity is a global condition. In fact, building on the Łojasiewicz inequality of Łojasiewicz [1963], [Bolte et al., 2007, Equation (15)] showed that sharpness always holds in finite dimensional spaces for reasonably well-behaved convex functions; see Lemma 2.3. Polynomial convex functions, the ℓ_p-norms, the Huber loss (see Appendix A.4), and the rectifier ReLU are simple examples of such functions.

Lemma 2.3. *Let $f : \mathbb{R}^n \to]-\infty, +\infty]$ be a lower semicontinuous, convex, and subanalytic function with $\{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\} \neq \varnothing$. Then for any bounded set $\mathcal{K} \subset \mathbb{R}^n$, there exists $\theta \in]0, 1[$ and $C > 0$ such that for all $x \in \mathcal{K}$,*

$$\operatorname{dist}\left(x, \arg\min_{\mathbb{R}^n} f\right) \leqslant C \left(f(x) - \min_{\mathbb{R}^n} f\right)^{\theta}.$$

Strong convexity is a standard requirement to prove linear convergence rates on smooth convex objectives but, regrettably, this considerably restricts the set of candidate functions. For our Blended Matching Pursuit algorithm, we will only require sharpness to establish linear convergence rates, thus including a larger class of functions.

2.2 Matching Pursuit algorithms

For $y \in \mathcal{H}$ and $f : x \in \mathcal{H} \mapsto \|y - x\|^2/2$, Problem (1) falls in the area of sparse recovery and is often solved with the Matching Pursuit algorithm [Mallat and Zhang, 1993]. The algorithm recovers a sparse representation of the signal y from the dictionary $\mathcal{D}$ by sequentially *pursuing* the best *matching* atom. At each iteration, it searches for the atom $v_t \in \mathcal{D}$ most correlated with the residual $y - x_t$, i.e., $v_t := \arg\max_{v \in \mathcal{D}} |\langle y - x_t, v \rangle|$, and adds it to the linear decomposition of the current iterate x_t to form the new iterate x_{t+1}, keeping track of the *active set* $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{v_t\}$. However, this does not prevent the algorithm from selecting atoms that have already been added in earlier iterations or that are redundant, hence affecting sparsity. The Orthogonal Matching Pursuit variant [Pati et al., 1993, Davis et al., 1994] overcomes this by computing the new iterate as the projection of the signal y onto $\mathcal{S}_t \cup \{v_t\}$; see Chen et al. [1989] and Tropp [2004] for analyses and Zhang [2009] for an extension to the stochastic case. Thus, $y - x_{t+1}$ becomes orthogonal to the active set.

In order to solve Problem (1) for any smooth convex objective, Locatello et al. [2017] proposed the Generalized Matching Pursuit (GMP) and Generalized Orthogonal Matching Pursuit (GOMP) algorithms (Algorithm 1); slightly abusing notation we will refer to the latter simply as Orthogonal

Matching Pursuit (OMP). The atom selection subroutine is implemented with a Frank-Wolfe linear minimization oracle $\arg\min_{v\in\mathcal{D}'}\langle\nabla f(x_t),v\rangle$ (Line 3). The solution v_t to this oracle is guaranteed to be a descent direction as it satisfies $\langle\nabla f(x_t),v_t\rangle\leqslant 0$ by symmetry of $\mathcal{D}'$, and $\langle\nabla f(x_t),v_t\rangle=0$ if and only if $x_t\in\arg\min_{\mathcal{H}}f$. Notice that for $y\in\mathcal{H}$ and $f:x\in\mathcal{H}\mapsto\|y-x\|^2/2$, the GMP and OMP variants of Algorithm 1 recover the original Matching Pursuit and Orthogonal Matching Pursuit algorithms respectively. In particular, up to a sign which does not affect the sequence of iterates, $\arg\max_{v\in\mathcal{D}}|\langle y-x_t,v\rangle|\Leftrightarrow\arg\min_{v\in\mathcal{D}'}\langle\nabla f(x_t),v\rangle$. In practice, the main difference in the case of general smooth convex functions is that the OMP variant (Line 6) is much more expensive, as a closed-form solution to this projection step is not available anymore. Hence, Line 6 is typically a sequence of projected gradient steps and OMP is significantly slower than GMP to converge.

Algorithm 1 Generalized/Orthogonal Matching Pursuit (GMP/OMP)

Input: Start atom $x_0\in\mathcal{D}$, number of iterations $T\in\mathbb{N}^*$.
Output: Iterates $x_1,\ldots,x_T\in\mathrm{span}(\mathcal{D})$.

1: $\mathcal{S}_0\leftarrow\{x_0\}$
2: **for** $t=0$ to $T-1$ **do**
3:     $v_t\leftarrow\arg\min_{v\in\mathcal{D}'}\langle\nabla f(x_t),v\rangle$
4:     $\mathcal{S}_{t+1}\leftarrow\mathcal{S}_t\cup\{v_t\}$
5:     *GMP variant*: $x_{t+1}\leftarrow\arg\min_{x_t+\mathbb{R}v_t}f$
6:     *OMP variant*: $x_{t+1}\leftarrow\arg\min_{\mathrm{span}(\mathcal{S}_{t+1})}f$
7: **end for**

2.3 Weak-separation oracle

We present in Oracle 2 the weak-separation oracle, a *modified* version of the one first introduced in Braun et al. [2017] and used in, e.g., Lan et al. [2017], Braun et al. [2019]. Note that the modification asks for an unconstrained improvement, whereas the original weak-separation oracle required an improvement relative to a reference point. As such, our variant here is even simpler than the original weak-separation oracle. The oracle is called in Line 11 by the Blended Matching Pursuit algorithm.

Oracle 2 Weak-separation $\mathrm{LPsep}_{\mathcal{D}}(c,\phi,\kappa)$

Input: Linear objective $c\in\mathcal{H}$, objective value $\phi\leqslant 0$, accuracy $\kappa\geqslant 1$.
Output: Either atom $v\in\mathcal{D}$ such that $\langle c,v\rangle\leqslant\phi/\kappa$ (positive call), or **false** ensuring $\langle c,z\rangle\geqslant\phi$ for all $z\in\mathrm{conv}(\mathcal{D})$ (negative call).

The weak-separation oracle determines whether there exists an atom $v\in\mathcal{D}$ such that $\langle c,v\rangle\leqslant\phi/\kappa$, and thereby relaxes the Frank-Wolfe linear minimization oracle. If not, then this implies that $\mathrm{conv}(\mathcal{D})$ can be *separated* from the ambient space by c and ϕ with the linear inequality $\langle c,z\rangle\geqslant\phi$ for all $z\in\mathrm{conv}(\mathcal{D})$. In practice, the oracle can be efficiently implemented using *caching*, i.e., first testing atoms that were already returned during previous calls as they may satisfy the condition here again. In this case, caching also preserves sparsity. If no active atom satisfies the condition, the oracle can be solved, e.g., by means of a call to a linear optimization oracle; see Braun et al. [2017] for an in-depth discussion. Lastly, we would like to briefly note that the parameter κ can be used to further promote positive calls over negative calls, by weakening the improvement requirement and therefore speeding up the oracle. Indeed, only negative calls need a full scan of the dictionary.

3 The Blended Matching Pursuit algorithm

We now present our Blended Matching Pursuit algorithm (BMP) in Algorithm 3. Note that although we blend steps, we maintain the explicit decomposition of the iterates $x_t=\sum_{j=1}^{n_t}\lambda_{t,i_j}a_{i_j}$ as linear combinations of the atoms.

Remark 3.1 (Algorithm design). *BMP actually does not require the atoms to have exactly the same norm and only needs the dictionary to be bounded, whether it be for ensuring the convergence rates or for computations; one could further take advantage of this to add weights to certain atoms. Line 6*

Algorithm 3 Blended Matching Pursuit (BMP)

Input: Start atom $x_0 \in \mathcal{D}$, parameters $\eta > 0$, $\kappa \geqslant 1$, and $\tau > 1$, number of iterations $T \in \mathbb{N}^*$.
Output: Iterates $x_1, \ldots, x_T \in \mathrm{span}(\mathcal{D})$.

1:   $\mathcal{S}_0 \leftarrow \{x_0\}$
2:   $\phi_0 \leftarrow \min\limits_{v \in \mathcal{D}'} \langle \nabla f(x_0), v \rangle / \tau$
3:   **for** $t = 0$ **to** $T - 1$ **do**
4:     $v_t^{\text{FW-}\mathcal{S}} \leftarrow \arg\min\limits_{v \in \mathcal{S}'_t} \langle \nabla f(x_t), v \rangle$
5:     **if** $\left\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \right\rangle \leqslant \phi_t / \eta$ **then**
6:       $\widetilde{\nabla} f(x_t) \leftarrow \mathrm{proj}_{\mathrm{span}(\mathcal{S}_t)}(\nabla f(x_t))$
7:       $x_{t+1} \leftarrow \arg\min\limits_{x_t + \mathbb{R}\widetilde{\nabla} f(x_t)} f$ {constrained step}
8:       $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$
9:       $\phi_{t+1} \leftarrow \phi_t$
10:    **else**
11:       $v_t \leftarrow \mathrm{LPsep}_{\mathcal{D}'}(\nabla f(x_t), \phi_t, \kappa)$
12:       **if** $v_t = $ **false then**
13:         $x_{t+1} \leftarrow x_t$ {dual step}
14:         $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$
15:         $\phi_{t+1} \leftarrow \phi_t / \tau$
16:       **else**
17:         $x_{t+1} \leftarrow \arg\min\limits_{x_t + \mathbb{R}v_t} f$ {full step}
18:         $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
19:         $\phi_{t+1} \leftarrow \phi_t$
20:       **end if**
21:    **end if**
22:    *Optional:* Correct $\mathcal{S}_{t+1}$
23: **end for**

---

is simply taking the component of $\nabla f(x_t)$ parallel to $\mathrm{span}(\mathcal{S}_t)$, which can be achieved by basic linear algebra and costs $\mathcal{O}(n \, \mathrm{card}(\mathcal{S}_t)^2)$ when $\mathcal{H} = \mathbb{R}^n$. The line searches Lines 7 and 17 can be replaced with explicit step sizes using the smoothness of f (see Fact B.2 in the Appendix). The purpose of (the optional) Line 22 is to reoptimize the active set $\mathcal{S}_{t+1}$, e.g., by reducing it to a subset that forms a basis for its linear span. One could also obtain further sparsity by removing atoms whose coefficient in the decomposition of the iterate is smaller than some threshold $\delta > 0$.

Blending. BMP aims at unifying the best of GMP and OMP. As seen in Section 2.2, an OMP iteration is a sequence of projected gradient (PG) steps. The idea is that the sequence of PG steps constituting an OMP iteration is actually overkill: there is a sweet spot where further optimizing over $\mathrm{span}(\mathcal{S}_t)$ is less effective than adding a new atom and taking a GMP step into a (possibly) new space. However, PG steps have the benefit of preserving sparsity, since no new atom is added. Furthermore, GMP steps require an expensive scan of the dictionary to output the descent direction $v_t^{\text{FW}} \leftarrow \arg\min_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle$. To remedy this, BMP blends *constrained steps* (PG steps, Line 7) with *full steps* (lazified GMP steps, Line 17) by promoting constrained steps as long as the progress in function value is *comparable* to that of a GMP step, else by taking a full step in an approximate direction v_t (with cheap computation via Oracle 2) such that the progress is *comparable* to that of a GMP step. Therefore, to monitor this blending of steps, we wish to compare $\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle$ and $\langle \nabla f(x_t), v_t \rangle$ to $\langle \nabla f(x_t), v_t^{\text{FW}} \rangle$, which quantities measure the progress in function value offered by a constrained step, a full step, and a GMP step respectively (see proofs in the Appendix).

Dual gap estimates. The aforementioned comparisons however cannot be made directly as the quantity $\langle \nabla f(x_t), v_t^{\text{FW}} \rangle$ is (deliberately) not computed; computing it requires an expensive complete scan of the dictionary. Instead, we use an estimation of this quantity, by introducing the *dual gap estimate* $|\phi_t|$. This designation comes from the fact that $-\langle \nabla f(x_t), v_t^{\text{FW}} \rangle$ is our equivalent of the *duality gap* from the constrained setting (see, e.g., Jaggi [2013]), and this will guide how we build our estimation. Indeed, since $\mathcal{D}'$ is symmetric and assuming $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{D}'))$, there exists (an unknown)

$\rho > 0$ such that $\{x_0, \ldots, x_T\} \cup \arg\min_{\mathcal{H}} f \subset \rho\operatorname{conv}(\mathcal{D}')$. Then for all $x^* \in \arg\min_{\mathcal{H}} f$,

$$\epsilon_t := f(x_t) - f(x^*) \leqslant \langle \nabla f(x_t), x_t - x^* \rangle$$

$$\leqslant \max_{u,v \in \rho\operatorname{conv}(\mathcal{D}')} \langle \nabla f(x_t), u - v \rangle = -2\rho\langle \nabla f(x_t), v_t^{\mathrm{FW}} \rangle, \qquad (2)$$

which is our desired inequality. We set $\phi_0 \leftarrow \langle \nabla f(x_0), v_0^{\mathrm{FW}} \rangle / \tau$ (Line 2) so $\epsilon_0 \leqslant 2\tau\rho|\phi_0|$ by (2). The criterion in Line 5 compares $\langle \nabla f(x_t), v_t^{\mathrm{FW}\text{-}\mathcal{S}} \rangle$ to ϕ_t. If this quantity is below the threshold ϕ_t, then a constrained step is not taken and the weak-separation oracle (Line 11, Oracle 2) is called to search for an atom v_t satisfying $\langle \nabla f(x_t), v_t \rangle \leqslant \phi_t$. If the oracle cannot find such an atom, then a full step is not taken and it returns a *negative call* with the certificate $\langle \nabla f(x_t), v_t^{\mathrm{FW}} \rangle > \phi_t$. In this case, BMP has detected an improved dual gap estimate and takes a *dual step* (Line 13): by (2), this implies that $\epsilon_t \leqslant 2\rho|\phi_t|$ so with $\phi_{t+1} \leftarrow \phi_t/\tau$ and $x_{t+1} \leftarrow x_t$, we recover $\epsilon_{t+1} \leqslant 2\tau\rho|\phi_{t+1}|$. Furthermore, observe that this update is a geometric rescaling which ensures that BMP requires only $N_{\mathrm{dual}} = \mathcal{O}(\ln 1/\epsilon)$ dual steps (see proofs). Thus, the total number of negative calls, i.e., the number of iterations requiring a complete scan of the dictionary, is only $\mathcal{O}(\ln 1/\epsilon)$. Therefore, for this and for the blending of steps, the dual gap estimates $|\phi_t|$ are the key to the speed-up realized by BMP.

Parameters. BMP involves three (hyper-)parameters $\eta > 0$, $\kappa \geqslant 1$, and $\tau > 1$ to be set before running the algorithm. The parameter η needs to be tuned carefully, as its value affects the criterion in Line 5 to promote either speed of convergence (e.g., $\eta \sim 0.1$, promoting full steps) or sparsity of the iterates (e.g., $\eta \sim 1000$, promoting constrained steps). In our experiments (see Section 4 and the Appendix), we found that setting $\eta \sim 5$ leads to close to both maximal speed of convergence and sparsity of the iterates, with the default choices $\kappa = \tau = 2$. In this setting, BMP converges (much) faster than GMP and has iterates with sparsity very comparable to that of OMP, and therefore it is possible to enjoy both properties of speed and sparsity simultaneously. Note that the value of κ also impacts the range of values of η to which BMP is sensitive, since the criterion (Line 5) tests $\min_{v \in \mathcal{S}'_t} \langle \nabla f(x_t), v \rangle \leqslant \phi_t/\eta$ while the weak-separation oracle asks for $v \in \mathcal{D}'$ such that $\langle \nabla f(x_t), v \rangle \leqslant \phi_t/\kappa$. In specific experiments, parameter tuning might further improve performance.

3.1 Convergence analyses

We start with the simpler case of smooth convex functions of order $\ell > 1$ (Theorem 3.2). Our main result is Theorem 3.3, which subsumes the case of strongly convex functions. To establish the convergence rates of GMP and OMP, Locatello et al. [2017] assume knowledge of an upper bound on $\sup\{\|x^*\|_{\mathcal{D}'}, \|x_0\|_{\mathcal{D}'}, \ldots, \|x_T\|_{\mathcal{D}'}\}$ where $\|\cdot\|_{\mathcal{D}'} : x \in \mathcal{H} \mapsto \inf\{\rho > 0 \mid x \in \rho\operatorname{conv}(\mathcal{D}')\}$ is the *atomic norm*. In Locatello et al. [2018], this is resolved by working with the atomic norm $\|\cdot\|_{\mathcal{D}'}$ instead of the Hilbert space induced norm $\|\cdot\|$ to, e.g., define smoothness and strong convexity of f and derive the proofs, but $\|\cdot\|_{\mathcal{D}'}$ itself can be difficult to derive in many applications. In contrast, we need neither the finiteness assumption nor to change the norm, however we assume f to be coercive to ensure feasibility of Problem (1), a reasonably mild assumption.

Theorem 3.2 (Smooth convex case). *Let $\mathcal{D} \subset \mathcal{H}$ be a dictionary such that $0 \in \operatorname{int}(\operatorname{conv}(\mathcal{D}'))$ and let $f : \mathcal{H} \to \mathbb{R}$ be smooth of order $\ell > 1$, convex, and coercive. Then the Blended Matching Pursuit algorithm (Algorithm 3) ensures that $f(x_t) - \min_{\mathcal{H}} f \leqslant \epsilon$ for all $t \geqslant T$ where $T = \mathcal{O}\big((L/\epsilon)^{1/(\ell-1)}\big)$.*

We now present our main result in its full generality. We provide the general convergence rates of BMP (Algorithm 3) in Theorem 3.3. Recall that sharpness is implied by strong convexity and that it is a very mild assumption in finite dimensional spaces as it is satisfied by all *well-behaved* convex functions (Lemma 2.3).

Theorem 3.3 (Smooth convex sharp case). *Let $\mathcal{D} \subset \mathcal{H}$ be a dictionary such that $0 \in \operatorname{int}(\operatorname{conv}(\mathcal{D}'))$ and let $f : \mathcal{H} \to \mathbb{R}$ be L-smooth of order $\ell > 1$, convex, coercive, and C-sharp of order $\theta \in \,]0, 1/\ell]$ on $\mathcal{K}$. Then the Blended Matching Pursuit algorithm (Algorithm 3) ensures that $f(x_t) - \min_{\mathcal{H}} f \leqslant \epsilon$ for all $t \geqslant T$ where*

$$T = \begin{cases} \mathcal{O}\left(C^{1/(1-\theta)} L^{1/(\ell-1)} \ln\left(\dfrac{C|\phi_0|}{\epsilon^{1-\theta}}\right)\right) & \text{if } \ell\theta = 1 \\[2.5ex] \mathcal{O}\left(\left(\dfrac{C^\ell L}{\epsilon^{1-\ell\theta}}\right)^{1/(\ell-1)}\right) & \text{if } \ell\theta < 1. \end{cases}$$

Moreover, $\operatorname{dist}(x_t, \arg\min_{\mathcal{H}} f) \to 0$ as $t \to +\infty$ at same rate.

6

If f is not strongly convex then Locatello et al. [2017] only guarantee a sublinear convergence rate $\mathcal{O}(1/\epsilon)$ for GMP and OMP, while Theorem 3.3 can still guarantee higher convergence rates, up to linear convergence $\mathcal{O}(\ln 1/\epsilon)$ if $\ell\theta = 1$, using sharpness. Note that in the popular case of smooth strongly convex functions of orders $\ell = 2$ and $s = 2$, Theorem 3.3 guarantees a linear convergence rate as these functions are sharp of order $\theta = 1/2$ (with constant $C = \sqrt{2/S}$) and thus satisfy $\ell\theta = 1$. For completeness, we also study this special case in Appendix C, with a simpler proof. In conclusion, Theorem 3.3 extends linear convergence rates to a large class of non-strongly convex functions solving Problem (1).

Remark 3.4 (Optimality of the convergence rates). *Let $n \leqslant +\infty$ be the dimension of $\mathcal{H}$. Nemirovskii and Nesterov [1985] provided unimprovable rates when solving Problem (1) in different cases. These optimal rates are reported in Table 1, where we compare them to those of BMP proved in this paper (Theorems 3.2 and 3.3). The third column gives the lower bounds on complexity stated in Nemirovskii and Nesterov [1985, Equations (1.20), (1.21'), and (1.21)]. Note that our rates are dimension independent and hold globally across iterations. It remains an open question to determine whether the gap in the exponent can be closed by accelerating BMP.*

Table 1: Comparison of the rates of BMP vs. the lower bounds on complexity.

Properties of f	**BMP rate**	**Lower bound on complexity**
Smooth convex	$T(\epsilon) = \mathcal{O}\left(\dfrac{1}{\epsilon^{1/(\ell-1)}}\right)$	$T(\epsilon) = \Omega\left(\min\left\{n, \dfrac{1}{\epsilon^{1/(1.5\ell-1)}}\right\}\right)$
Smooth convex sharp with $\ell = 2, \theta = 1/2$	$T(\epsilon) = \mathcal{O}\left(\ln\left(\dfrac{1}{\epsilon}\right)\right)$	$T(\epsilon) = \Omega\left(\min\left\{n, \ln\left(\dfrac{1}{\epsilon}\right)\right\}\right)$
Smooth convex sharp with $\ell\theta < 1$	$T(\epsilon) = \mathcal{O}\left(\dfrac{1}{\epsilon^{(1-\ell\theta)/(\ell-1)}}\right)$	$T(\epsilon) = \Omega\left(\min\left\{n, \dfrac{1}{\epsilon^{(1-\ell\theta)/(1.5\ell-1)}}\right\}\right)$

4 Computational experiments

We implemented BMP in Python 3 along with GMP and OMP [Locatello et al., 2017], the Accelerated Matching Pursuit algorithm (accMP) [Locatello et al., 2018], and the Blended Conditional Gradients (BCG) [Braun et al., 2019] and Conditional Gradient with Enhancement and Truncation (CoGEnT) [Rao et al., 2015] algorithms for completeness. All algorithms share the same code framework to ensure fair comparison. No enhancement beyond basic coding was performed. We ran the experiments on a laptop under Linux Ubuntu 18.04 with Intel Core i7 3.5GHz CPU and 8GB RAM. The random data are drawn from Gaussian distributions. For GMP, OMP, BCG, and CoGEnT, we represented the dual gaps by $-\min_{v \in \mathcal{D}'}\langle \nabla f(x_t), v\rangle$, yielding a zig-zag plot dissimilar to the stair-like plot of the dual gap estimates $|\phi_t|$ of BMP. The Appendix contains additional experiments.

4.1 Comparison of BMP vs. GMP, OMP, BCG, and CoGEnT

Let $\mathcal{H}$ be the Euclidean space $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ and $\mathcal{D}$ be the set of signed canonical vectors $\{\pm e_1, \ldots, \pm e_n\}$. Suppose we want to learn the (sparse) source x^* from observed data $y \coloneqq Ax^* + w$, where $A \in \mathbb{R}^{m \times n}$ and where $w \sim \mathcal{N}(0, \sigma^2 I_m)$ is the noise in the observed y. The general and most intuitive formulation of the problem is $\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2$ s.t. $\|x\|_0 \leqslant \|x^*\|_0 =: s$ but the ℓ_0-pseudo norm constraint is nonconvex and makes the problem NP-hard and therefore intractable in many situations [Natarajan, 1995]. To remedy this, one can handle the sparsity constraint in various ways, either by completely removing it and relying on an algorithm inherently promoting sparsity (as done in BMP, GMP, and OMP), or through a convex relaxation of the constraint, often via the ℓ_1-norm, and then solving the new constrained convex problem $\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2$ s.t. $\|x\|_1 \leqslant \|x^*\|_1$ (as done in BCG and CoGEnT). We ran a comparison of these methods, where we favorably provided the constraint $\|x\|_1 \leqslant \|x^*\|_1$ for BCG and CoGEnT although x^* is unknown. We set $m = 500$, $n = 2000$, $s = 100$, and $\sigma = 0.05$. In BMP, we set $\kappa = \tau = 2$ and we chose $\eta = 5$; see Appendix A.1 for an in-depth sensitivity analysis of BMP with respect to η. We did not perform any additional correction of the active sets (Line 22). Note that [Rao et al., 2015, Table III] demonstrated the superiority of

CoGEnT over CoSaMP [Needell and Tropp, 2009], Subspace Pursuit [Dai and Milenkovic, 2009], and Gradient Descent with Sparsification [Garg and Khandekar, 2009] on an equivalent experiment and we therefore do not compare to those methods.
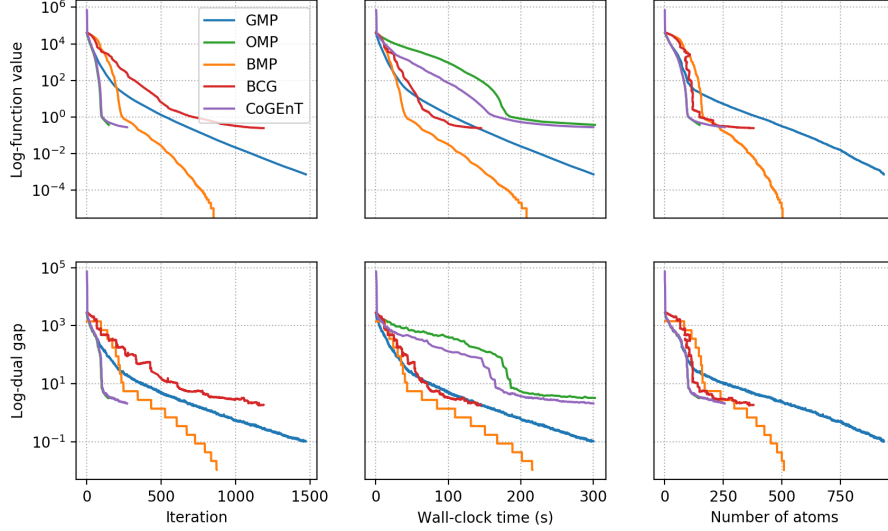

Figure 1: Comparison of BMP vs. GMP, OMP, BCG, and CoGEnT, with $\eta = 5$.

Figure 1 shows that BMP is the fastest algorithm in wall-clock time and has close-to-optimal sparsity. It is important to stress that, unlike BCG and CoGEnT, BMP achieves this while having no explicit sparsity-promoting constraint, regularization, nor information on x^*. Thus, when $\|x^*\|_1$ is not provided, which is the case in most applications, BCG and CoGEnT would require a hyper-parameter tuning of the sparsity-inducing constraint (or, equivalently, the Lagrangian penalty parameters), such as the radius of the ℓ_1-ball [Tibshirani, 1996], as used here, or the trace-norm-ball [Fazel et al., 2001]. OMP and CoGEnT converge faster per-iteration, as expected, given that they solve a reoptimization problem at each iteration, however this is very costly and the disadvantage becomes evident in wall-clock time performance. Note that another "obvious" choice for an algorithm would be projected gradient descent, however the provided sparsity is far from sufficient (see Appendix A.2).
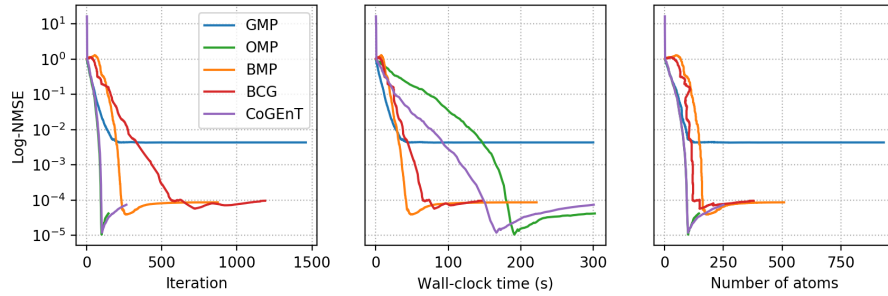

Figure 2: Comparison in NMSE of BMP vs. GMP, OMP, BCG, and CoGEnT, with $\eta = 5$.

In Figure 2, we compare the Normalized Mean Squared Error (NMSE) of the different methods. The NMSE at iterate x_t is defined as $\|x_t - x^*\|_2^2 / \|x^*\|_2^2$. The plots show a rebound occurring once the NMSE reaches $\sim 10^{-4}$, which is due to the algorithms overfitting to the noisy measurements y. A post-processing step can mitigate the rebound via early stopping or by removing atoms whose coefficient in the decomposition of the iterate are smaller than some threshold $\delta > 0$. We used early stopping on a validation set and present the test error $\|y_{\text{test}} - A_{\text{test}} x_T\|_2^2 / m_{\text{test}}$ on a test set in Table 2, where x_T is the solution iterate for each algorithm. For completeness, we also reported the results for the Gradient Hard Thresholding Pursuit (GraHTP) and Fast Gradient Hard Thresolding Pursuit (Fast GraHTP) algorithms [Yuan et al., 2018], for which we favorably set $k = \|x^*\|_0$. As expected, GMP

performs the worst on the test set because its NMSE does not achieve sufficient convergence (see Figure 2), highlighting the importance of a clean, i.e., sparse, decomposition into the dictionary $\mathcal{D}$.

Table 2: Test error achieved using early stopping on a validation set.

Algorithm	GMP	OMP	BMP	BCG	CoGEnT	GraHTP	Fast GraHTP
Test error	0.1917	0.0036	0.0037	0.0068	0.0043	0.0036	0.0037

The Appendix contains additional experiments on different objective functions: an arbitrarily chosen norm (Appendix A.3), the Huber loss (Appendix A.4), the distance to a convex set (Appendix A.5), and a logistic regression loss (Appendix A.6). The conclusions are identical.

4.2 Comparison of BMP vs. accMP

Locatello et al. [2018] recently provided an Accelerated Matching Pursuit algorithm (accMP) for solving Problem (1). We implemented the same code as theirs, using the exact same parametrization. The code framework matches the one we used for BMP. We ran BMP on their toy data example and compared the results against accMP (which they labeled *accelerated steepest* in their plot); notice that we recovered their (per-iteration) plot exactly. The experiment is to minimize $f : x \in \mathbb{R}^{100} \mapsto \|x - b\|_2^2/2$ over the linear span of $\mathcal{D}$, where $\mathcal{D}$ is dictionary of 200 randomly chosen atoms in $\mathbb{R}^{100}$ and $b \in \mathbb{R}^{100}$ is also randomly chosen. The parameters, kindly provided by the authors of Locatello et al. [2018], for accMP were $L = 1000$ and $\nu = 1$. We report the results in Figure 3.
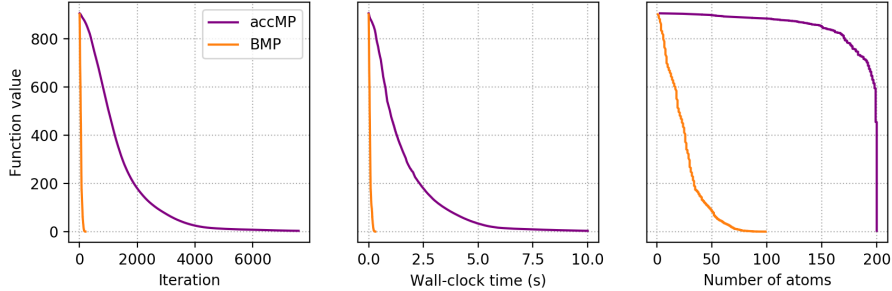

Figure 3: Comparison of BMP vs. accMP, with $\eta = 3$.

We see that BMP outperforms accMP in both speed of convergence and sparsity of the iterates. In fact, in terms of sparsity, accMP needs to use all available atoms to converge while BMP needs only half as much. Furthermore, accMP needs 75% of all available atoms to start converging significantly while BMP starts to converge instantaneously. We suspect that this is due to the following: accMP accelerates coordinate descent-like directions, which might be relatively bad approximations of the actual descent direction $-\nabla f(x_t)$, whereas BMP is working directly with (the projection of) $-\nabla f(x_t)$, achieving much more progress and offsetting the effect of acceleration.

5 Final remarks

We presented a Blended Matching Pursuit algorithm (BMP) which enjoys both properties of fast rate of convergence and sparsity of the iterates. More specifically, we derived linear convergence rates for a large class of non-strongly convex functions solving Problem (1), and we showed that our blending approach outperforms the state-of-the-art methods in speed of convergence while achieving close-to-optimal sparsity, and this without requiring sparsity-inducing constraints nor regularization. Although BMP already outperforms the Accelerated Matching Pursuit algorithm [Locatello et al., 2018] in our experiments, we believe it is also amenable to acceleration.

Acknowledgments

References

J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

G. Braun, S. Pokutta, and D. Zink. Lazifying conditional gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 566–575, 2017.

G. Braun, S. Pokutta, D. Tu, and S. Wright. Blended conditional gradients: the unconditioning of conditional gradients. In *Proceedings of the 36th International Conference on Machine Learning*, pages 735–743, 2019.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989.

L. Condat. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1):575–585, 2016.

W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

M. A. Davenport and M. B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Transactions on Information Theory*, 56(9):4395–4401, 2010.

G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions with matching pursuits. *Optical Engineering*, 33(7):2183–2191, 1994.

M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, pages 4734–4739, 2001.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

R. Garg and R. Khandekar. Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property. In *Proceedings of the 26th International Conference on Machine Learning*, pages 337–344, 2009.

R. Gribonval and P. Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *IEEE Transactions on Information Theory*, 52(1):255–261, 2006.

I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems 17*, pages 545–552. 2005.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, 2013.

T. Kerdreux, A. d'Aspremont, and S. Pokutta. Restarting Frank-Wolfe. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1275–1283, 2019.

G. Lan, S. Pokutta, Y. Zhou, and D. Zink. Conditional accelerated lazy stochastic gradient descent. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1965–1974, 2017.

E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

F. Locatello, R. Khanna, M. Tschannen, and M. Jaggi. A unified optimization view on generalized matching pursuit and Frank-Wolfe. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 860–868, 2017.

F. Locatello, A. Raj, S. P. Karimireddy, G. Rätsch, B. Schölkopf, S. U. Stich, and M. Jaggi. On matching pursuit and coordinate descent. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3198–3207, 2018.

S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, 117, pages 87–89. Colloques Internationaux du CNRS, 1963.

S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

A. S. Nemirovskii and Y. E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

H. Nguyen and G. Petrova. Greedy strategies for convex optimization. *Calcolo*, 54(1):207–224, 2017.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

N. Rao, S. Shah, and S. Wright. Forward-backward greedy algorithms for atomic norm regularization. *IEEE Transactions on Signal Processing*, 63(21):5798–5811, 2015.

V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems 30*, pages 1119–1129. 2017.

S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20:2807–2832, 2010.

V. Temlyakov. Chebushev greedy algorithm in convex optimization. *arXiv preprint arXiv:1312.1244*, 2013.

V. Temlyakov. Greedy algorithms in convex optimization on Banach spaces. In *Proceedings of the 48th Asilomar Conference on Signals, Systems, and Computers*, pages 1331–1335, 2014.

V. Temlyakov. Greedy approximation in convex optimization. *Constructive Approximation*, 41(2):269–296, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

R. J. Tibshirani. A general framework for fast stagewise algorithms. *Journal of Machine Learning Research*, 16 (1):2543–2588, 2015.

J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Q. Yao and J. T. Kwok. Greedy learning of generalized low-rank model. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 2294–2300, 2016.

X.-T. Yuan, P. Li, and T. Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18 (166):1–43, 2018.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

A Additional computational experiments

We provide the sensitivity analysis to the experiment in Figure 1 in Appendix A.1, and the comparison to the projected gradient method in Appendix A.2. We then conduct additional experiments on a variety of objective functions: an arbitrarily chosen norm (Appendix A.3), the Huber loss (Appendix A.4), the distance to a convex set (Appendix A.5), and a logistic regression loss (Appendix A.6).

A.1 Sensitivity of BMP to the parameter η

Here we report the sensitivity analysis of BMP for the data in Section 4.1. We ran BMP (Algorithm 3) for values of η in $\{100, 10, 5, 2, 1\}$. We set $\kappa = 2$ and $\tau = 2$ and did not activate the correction of atoms (Line 22). We report the results in Figure 4.
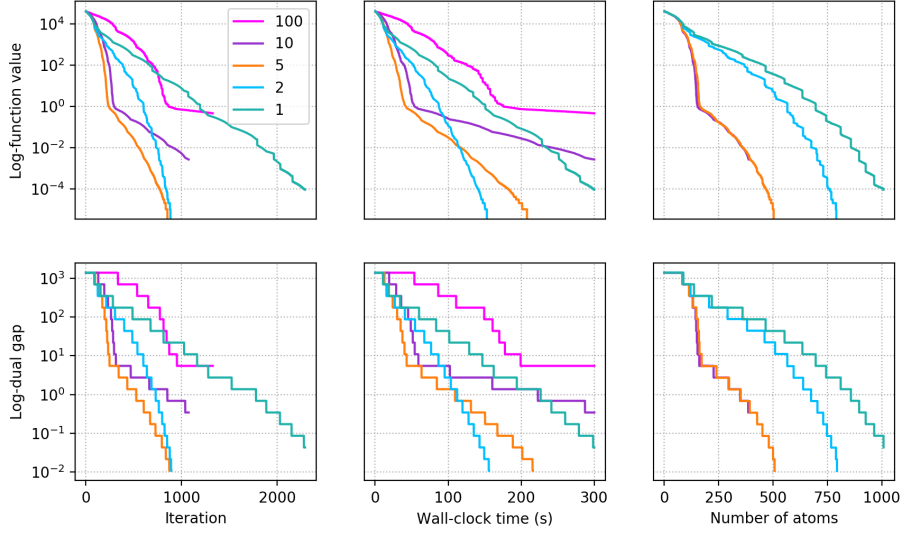

Figure 4: Sensitivity of BMP to the parameter η.

We see that $\eta = 5$ is at the sweet spot between speed of convergence and sparsity of the iterates. Higher values of η have similar levels of sparsity but they perform worse for speed of convergence. Lower values of η perform much worse in sparsity and are not better in speed; $\eta = 2$ offsets $\eta = 5$ after 100 seconds but the function value is already 10^{-2} at that point. Therefore, by setting $\eta = 5$ in this example we achieve both speed of convergence and sparsity of the iterates. Similar insights are obtained in the other experiments, so that $\eta \sim 5$ seems to be a good initial choice.

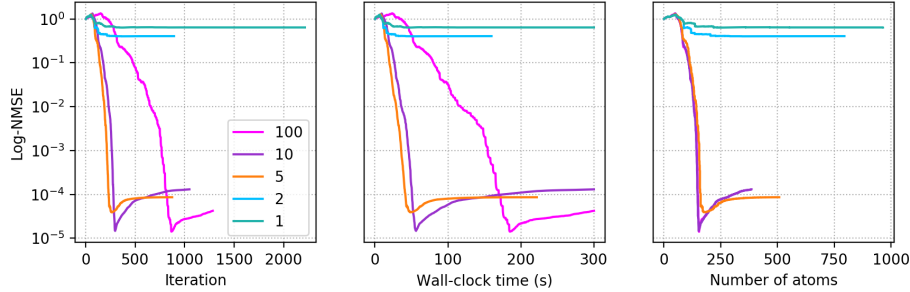For completeness, we present in Figure 5 the sensitivity of BMP to η in NMSE.

Figure 5: Sensitivity of BMP to the parameter η in NMSE.

A.2 Comparison with PGD

Projected gradient descent (PGD) is a natural candidate for the experiment in Section 4.1. However, it does not ensure sufficient sparsity of the iterates. We depict three configurations of PGD in Figure 6, each named "PGD:α" where PGD is ran with the constraint $\|x\|_1 \leqslant \alpha \|x^*\|_1$ and $\alpha \in \{1/2, 1, 2\}$. The implementation of PGD is in line with our general code framework and we used the method of Condat [2016] for projections onto the ℓ_1-ball. The number of atoms collected by the iterates in PGD are reported as the number of nonzero coordinates. Note that in BMP, GMP, and OMP we do not check if a selected atom v_t already satisfies $-v_t \in \mathcal{S}_t$ before adding it to $\mathcal{S}_t$, which is disadvantageous to these algorithms when evaluating their sparsity performance.
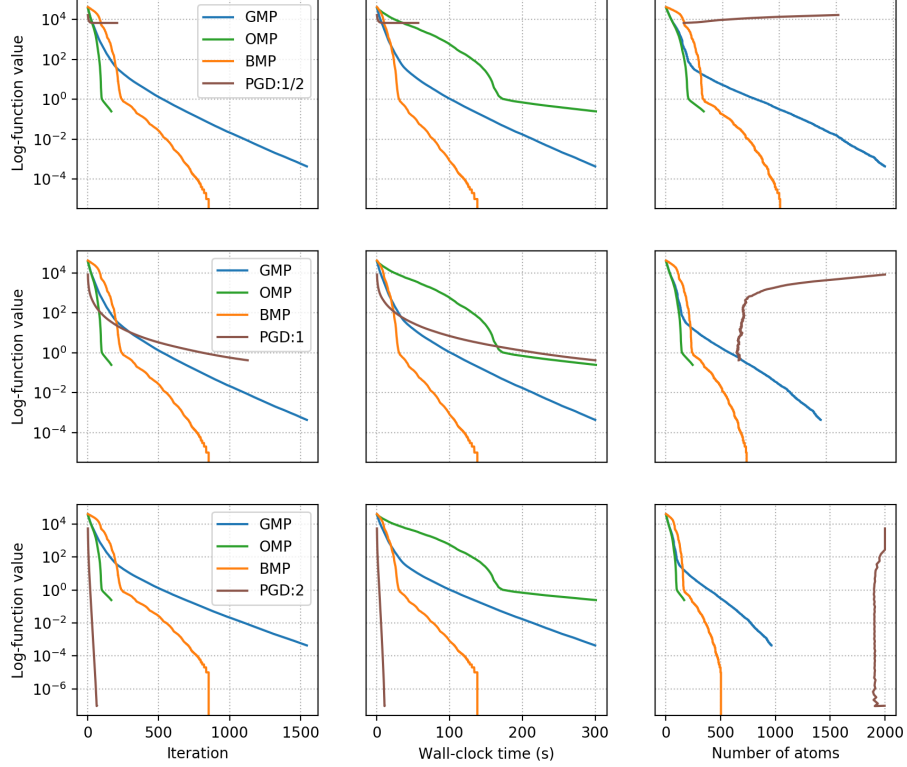

Figure 6: Comparison of PGD vs. the MP algorithms.

As expected, the constraint $\|x\|_1 \leqslant \|x^*\|_1$ provides the best results for PGD; the constraint $\|x\|_1 \leqslant 2\|x^*\|_1$ is too loose and basically produces no sparsity in the iterates (recall that the ambient space is $\mathbb{R}^{2000}$). In the configuration $\|x\|_1 \leqslant \|x^*\|_1$, PGD does not converge faster than OMP and produces significantly worse sparsity than OMP and BMP.

A.3 Regression with arbitrarily chosen norm

We set $m = 250$, $n = 1000$, $s = 50$, and $\sigma = 0.05$ and generated the data as in Section 4.1. We ran a comparison with the arbitrarily chosen $f : x \in \mathbb{R}^n \mapsto \|Ax - b\|_3^5$. We plot the results in Figure 7.
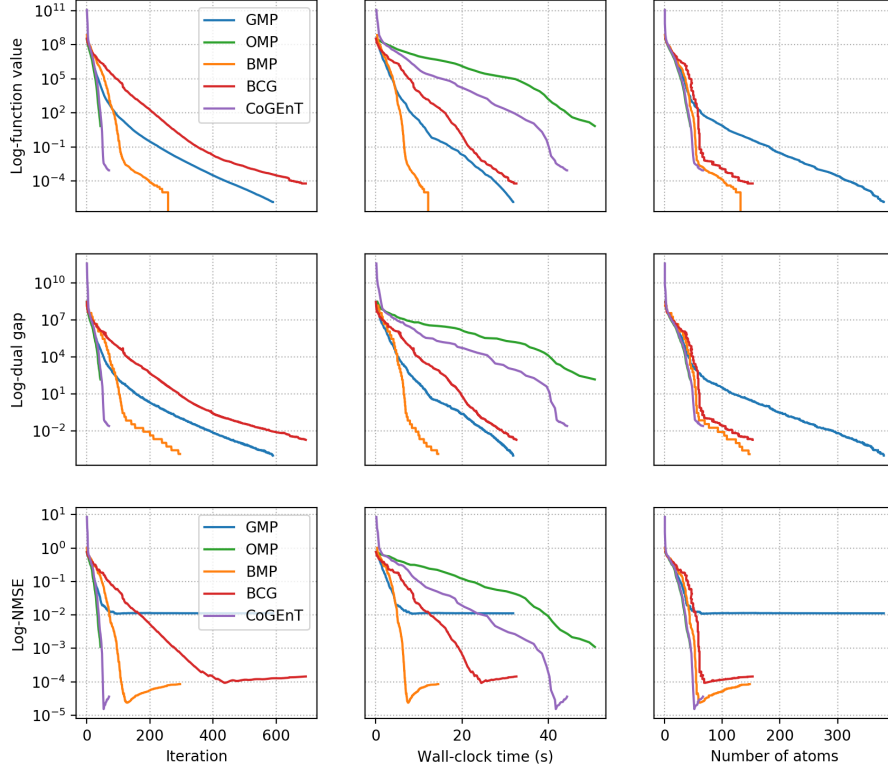

Figure 7: Comparison of BMP vs. GMP and OMP on $f : x \in \mathbb{R}^n \mapsto \|Ax - b\|_3^5$, with $\eta = 5$.

We see that BMP offers very close-to-optimal levels of sparsity while being faster than the other algorithms in wall-clock time. We provide a sensitivity analysis of BMP to the parameter η in Figure 8. We see that $\eta \sim 5$ is an appropriate choice combining the best of speed and sparsity.
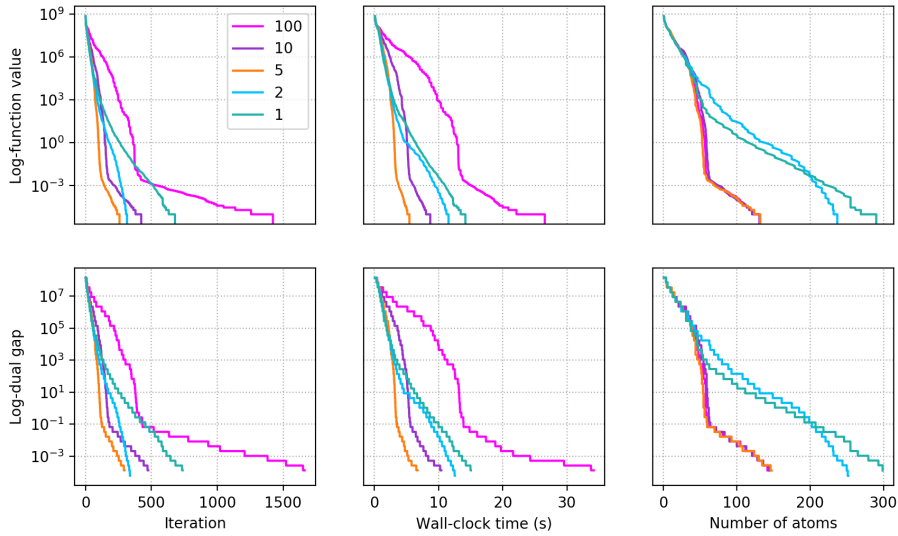

Figure 8: Sensitivity of BMP to the parameter η.

14

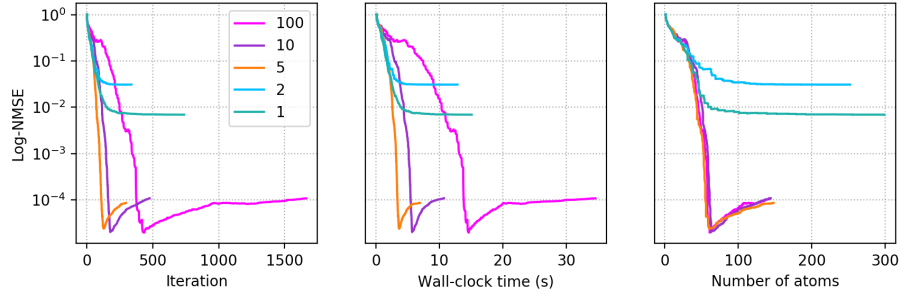For completeness, we present in Figure 9 the sensitivity of BMP to η in NMSE.

Figure 9: Sensitivity of BMP to the parameter η in NMSE.

A.4 Huber loss

The Huber loss [Huber, 1964] is a smooth combination of the squared and absolute losses. The absolute loss is robust to outliers in the dataset, however its gradient is piecewise constant and not defined at the origin. This leads to instability of the solutions. The Huber loss overcomes this by behaving like the squared loss around the origin:

$$h_\delta : t \in \mathbb{R} \mapsto \begin{cases} t^2/2 & \text{if } |t| \leqslant \delta \\ \delta(|t| - \delta/2) & \text{else} \end{cases}$$

where $\delta > 0$ defines this region around the origin. We set $m = 250$, $n = 1000$, $s = 50$, and $\sigma = 0.05$ and generated the data as in Section 4.1. In this experiment we aim at minimizing the smooth convex function

$$f : x \in \mathbb{R}^n \mapsto \sum_{i=1}^{m} h_{10}(a_i^\top x - y_i)$$

where $a_1^\top, \ldots, a_m^\top \in \mathbb{R}^{1 \times n}$ are the rows of A. We plot the results in Figure 10.

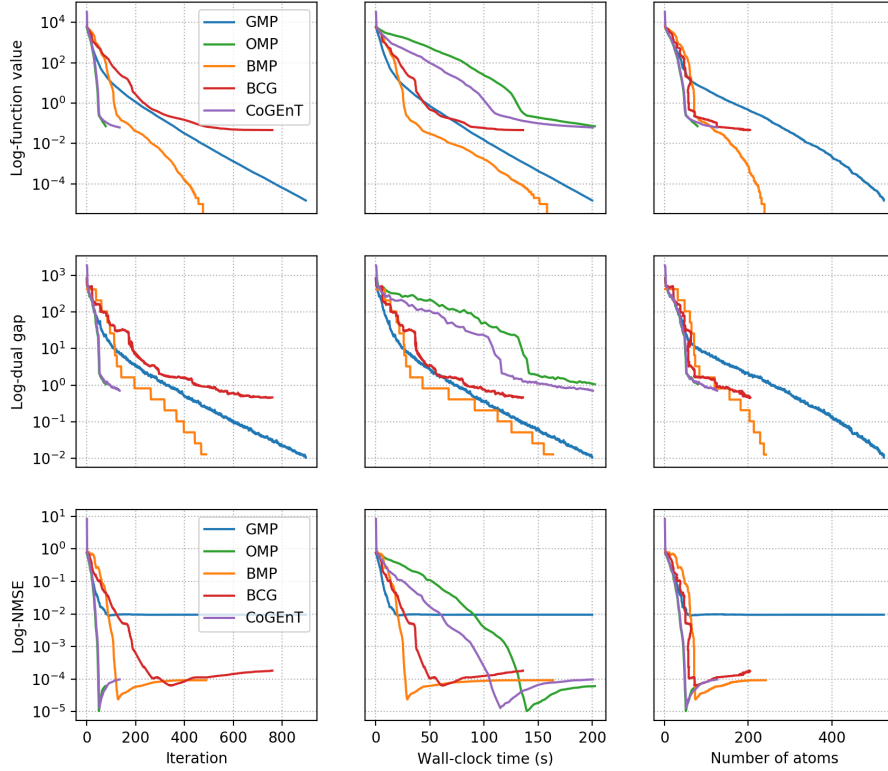Figure 10: Comparison of BMP vs. GMP and OMP on $f : x \in \mathbb{R}^n \mapsto \sum_{i=1}^{m} h_{10}(a_i^\top x - y_i)$, with $\eta = 5$.

Again, BMP has very close-to-optimal levels of sparsity while being the fastest algorithm to converge. We provide a sensitivity analysis of BMP to the parameter η in Figure 11. We see that $\eta \sim 5$ is an appropriate choice combining the best of speed and sparsity.
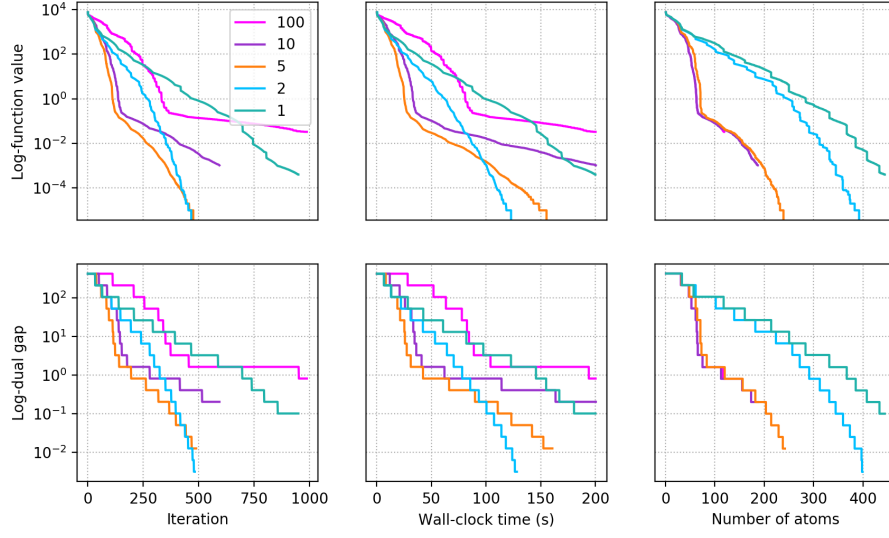
Figure 11: Sensitivity of BMP to the parameter η.

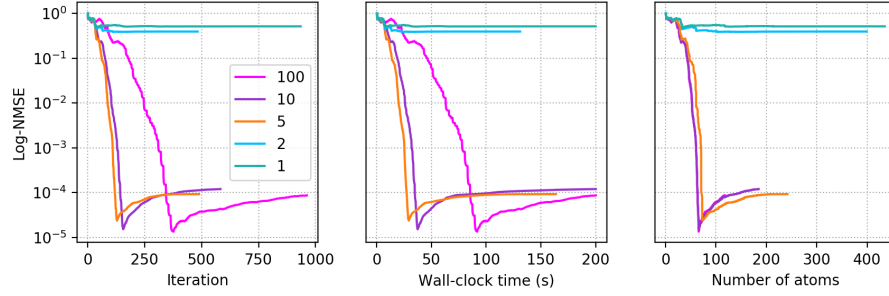For completeness, we present in Figure 12 the sensitivity of BMP to η in NMSE.

Figure 12: Sensitivity of BMP to the parameter η in NMSE.

A.5 Distance to a convex set

Here we compared BMP vs. GMP and OMP on an arbitrarily chosen problem. We used

$$f : x \in \mathbb{R}^{500} \mapsto d_{\overline{\mathcal{B}}(0,1)}(Ax - b)^2 = \|(Ax - b) - \mathrm{proj}_{\overline{\mathcal{B}}(0,1)}(Ax - b)\|_2^2$$

and $\mathcal{D}$ a dictionary of 750 atoms randomly chosen in $\mathbb{R}^{500}$, where $A \in \mathbb{R}^{500 \times 500}$ and $b \in \mathbb{R}^{500}$ are also randomly chosen. We did not reduce f to a closed-form expression simplifying computations. This is not a setting where BCG or CoGEnT can be applied. We depict two configurations of BMP: one with emphasis on sparsity of the iterates and one with emphasis on speed of convergence. The parameters η_{sparse} and η_{fast} were both optimized. We plot the results in Figure 13.
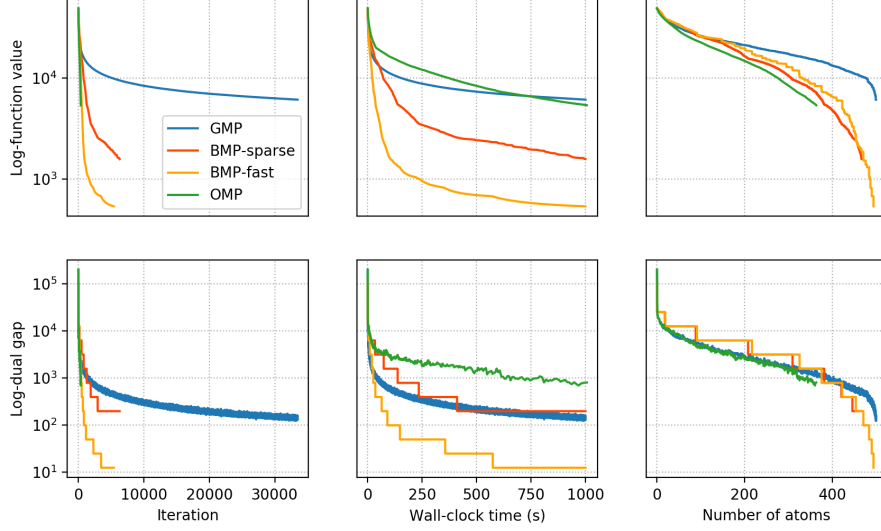
Figure 13: Comparison of BMP vs. GMP and OMP on $f : x \in \mathbb{R}^{500} \mapsto d_{\overline{\mathcal{B}}(0,1)}(Ax - b)^2$, with $\eta_{\mathrm{sparse}} = 10$ and $\eta_{\mathrm{fast}} = 2$.

We provide a sensitivity analysis of BMP to the parameter η in Figure 14. The scaling is not exactly the same as in Figure 13 due to the randomness in the generation of the data. We see that $\eta \sim 2$ is an appropriate choice combining the best of speed and sparsity.

Figure 14: Sensitivity of BMP to the parameter η.

A.6 Logistic regression

Here we compared BMP vs. GMP and OMP on the Gisette dataset [Guyon et al., 2005] available at https://archive.ics.uci.edu/ml/datasets/Gisette. We did not have access to the true parameters x^* so we could not produce the NMSE plots. The objective function is the logistic loss

$$f : x \in \mathbb{R}^n \mapsto \frac{1}{m} \sum_{i=1}^{m} \ln(1 + \exp(-y_i a_i^\top x))$$

with the labels $y_i \in \{-1, 1\}$, and the dictionary is the set of signed canonical vectors $\mathcal{D} = \{\pm e_1, \ldots, e_n\}$. We have $n = 5000$ and in order to reduce the running time, we chose $m = 1000$ and we slightly enhanced the code framework by replacing $\min_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle$ with $\min_{i \in [\![1,n]\!]} -|[\nabla f(x_t)]_i|$. Note that this lessens the speed-up provided by the weak-separation oracle in BMP. We represented the dual gaps of BMP by $\max_{i \in [\![1,n]\!]} |[\nabla f(x_t)]_i|$, like for GMP and OMP, and thus yielding a similar zig-zag plot.
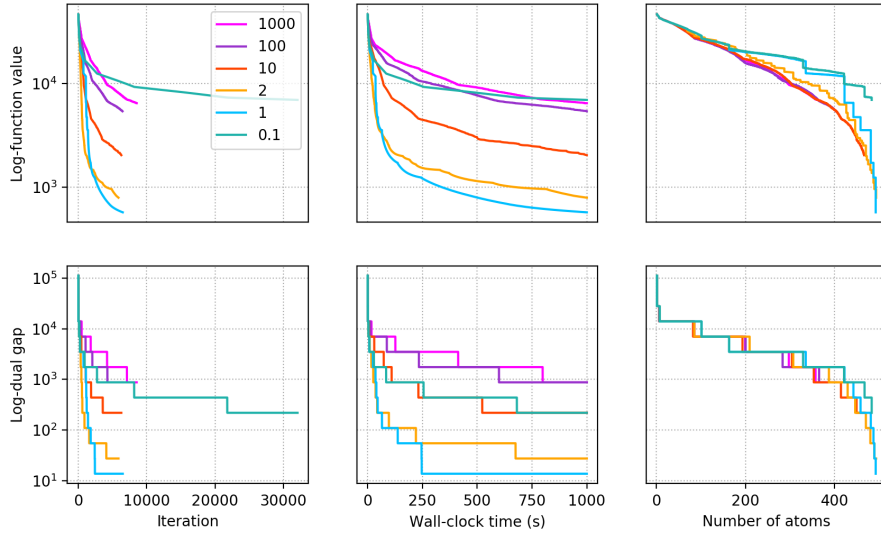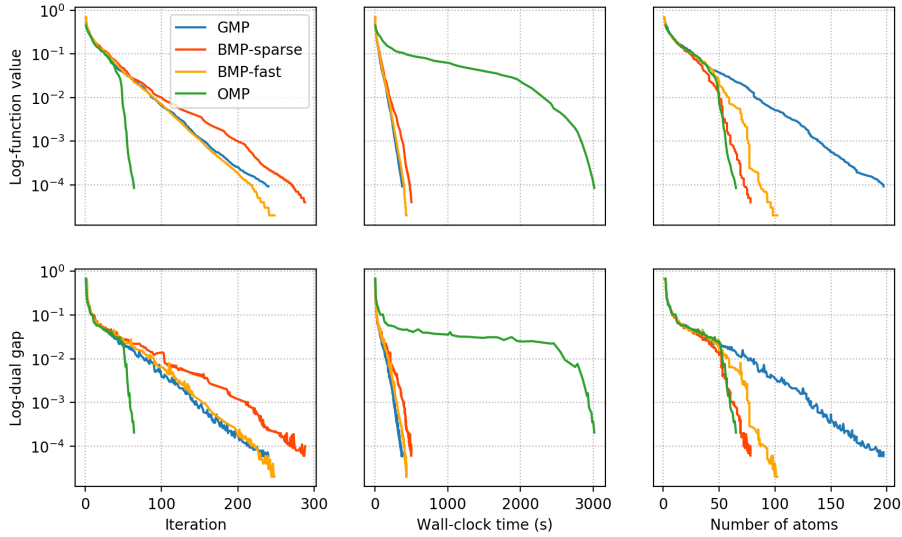

Figure 15: Comparison of BMP vs. GMP and OMP on the Gisette dataset with $\eta_{\text{sparse}} = 3$ and $\eta_{\text{fast}} = 2$.

In this situation, Figure 15 shows that BMP converges as fast as GMP while producing iterates with much higher sparsity, equivalent to that of OMP. Hence, BMP hits the sweet spot of speed of convergence and sparsity of the iterates.

B Prerequisites for proofs

Fact B.1. *Let $f : \mathcal{H} \to \mathbb{R}$ be a coercive function and $(x_t)_{t \in \mathbb{N}}$ be a sequence of iterates in $\mathcal{H}$ such that $f(x_{t+1}) \leqslant f(x_t)$ for all $t \in \mathbb{N}$. Then $(x_t)_{t \in \mathbb{N}}$ is bounded.*

Proof. By assumption, $f(x_t) \leqslant f(x_0)$ for all $t \in \mathbb{N}$, so $\limsup_{t \to +\infty} f(x_t) \leqslant f(x_0) < +\infty$. Suppose $(x_t)_{t \in \mathbb{N}}$ is unbounded. Then there exists $\varphi : \mathbb{N} \to \mathbb{N}$ strictly increasing such that the subsequence $(x_{\varphi(t)})_{t \in \mathbb{N}}$ satisfies $\lim_{t \to +\infty} \|x_{\varphi(t)}\| = +\infty$. By coercivity, this implies that $\lim_{t \to +\infty} f(x_{\varphi(t)}) = +\infty$, and therefore $\limsup_{t \to +\infty} f(x_t) \geqslant +\infty$. This is absurd. $\square$

Fact B.2. *Let $f : \mathcal{H} \to \mathbb{R}$ be differentiable, $M > 0$, $\mu > 1$, and $x, v \in \mathcal{H}$ such that $\langle \nabla f(x), v \rangle \leqslant 0$. Define*

$$g : \gamma \in \mathbb{R}_+ \mapsto f(x) + \gamma \langle \nabla f(x), v \rangle + \frac{M}{\mu} \gamma^\mu \|v\|^\mu.$$

Then

$$\min_{\mathbb{R}_+} g = f(x) - \frac{\langle -\nabla f(x), v\rangle^{\overline{\mu}}}{\overline{\mu} M^{\overline{\mu}-1} \|v\|^{\overline{\mu}}}$$

where $\overline{\mu} := \mu/(\mu - 1) > 1$.

Proof. Let $\overline{\mu} = \mu/(\mu - 1)$. We have $\overline{\mu} - 1 = 1/(\mu - 1)$, and g is differentiable with

$$\forall \gamma \in \mathbb{R}_+, \ g'(\gamma) \geqslant 0 \Leftrightarrow \langle \nabla f(x), v\rangle + M\gamma^{\mu-1}\|v\|^{\mu} \geqslant 0$$

$$\Leftrightarrow \gamma \geqslant \left(\frac{\langle -\nabla f(x), v\rangle}{M\|v\|^{\mu}}\right)^{1/(\mu-1)} = \frac{\langle -\nabla f(x), v\rangle^{\overline{\mu}-1}}{M^{\overline{\mu}-1}\|v\|^{\overline{\mu}}}.$$

Therefore, using $\mu(\overline{\mu} - 1) = \overline{\mu}$ and $1 - 1/\mu = 1/\overline{\mu}$,

$$\min_{\mathbb{R}_+} g = f(x) + \frac{\langle -\nabla f(x), v\rangle^{\overline{\mu}-1}}{M^{\overline{\mu}-1}\|v\|^{\overline{\mu}}}\nabla f(x)v + \frac{M}{\mu}\left(\frac{\langle -\nabla f(x), v\rangle^{\overline{\mu}-1}}{M^{\overline{\mu}-1}\|v\|^{\overline{\mu}}}\right)^{\mu}\|v\|^{\mu}$$

$$= f(x) - \frac{\langle -\nabla f(x), v\rangle^{\overline{\mu}}}{M^{\overline{\mu}-1}\|v\|^{\overline{\mu}}} + \frac{1}{\mu}\frac{\langle -\nabla f(x), v\rangle^{\mu(\overline{\mu}-1)}}{M^{\mu(\overline{\mu}-1)-1}\|v\|^{\mu(\overline{\mu}-1)}}$$

$$= f(x) - \left(1 - \frac{1}{\mu}\right)\frac{\langle -\nabla f(x), v\rangle^{\overline{\mu}}}{M^{\overline{\mu}-1}\|v\|^{\overline{\mu}}}$$

$$= f(x) - \frac{\langle -\nabla f(x), v\rangle^{\overline{\mu}}}{\overline{\mu} M^{\overline{\mu}-1}\|v\|^{\overline{\mu}}}.$$

$\square$

Corollary B.3. *Let $f : \mathcal{H} \to \mathbb{R}$ be S-strongly convex of order $s = 2$ with $\{x^*\} := \arg\min_{\mathcal{H}} f$. Then for all $x \in \mathcal{H}$,*

$$f(x^*) \geqslant f(x) - \frac{\langle \nabla f(x), x - x^*\rangle^2}{2S\|x - x^*\|^2}.$$

Proof. Let $x \in \mathcal{H}$. By strong convexity, for all $\gamma \in \mathbb{R}_+$,

$$f(x + \gamma(x^* - x)) \geqslant f(x) + \gamma\langle \nabla f(x), x^* - x\rangle + \frac{S}{2}\gamma^2\|x^* - x\|^2. \tag{3}$$

Let $v := x^* - x$, then $\langle \nabla f(x), v\rangle \leqslant 0$ by convexity. Applying Fact B.2 to the right-hand side of (3), since $\overline{s} = s/(s - 1) = 2$,

$$f(x + \gamma(x^* - x)) \geqslant f(x) - \frac{\langle \nabla f(x), x - x^*\rangle^2}{2S\|x - x^*\|^2}$$

so, with $\gamma = 1$,

$$f(x^*) \geqslant f(x) - \frac{\langle \nabla f(x), x - x^*\rangle^2}{2S\|x - x^*\|^2}.$$

$\square$

C   The smooth strongly convex case

Theorem C.1 (Smooth strongly convex case). *Let $\mathcal{D} \subset \mathcal{H}$ be a dictionary such that $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{D}'))$ and let $f : \mathcal{H} \to \mathbb{R}$ be L-smooth of order $\ell = 2$ and S-strongly convex of order $s = 2$. Then the Blended Matching Pursuit algorithm (Algorithm 3) ensures that $f(x_t) - \min_{\mathcal{H}} f \leqslant \epsilon$ for all $t \geqslant T$ where*

$$T = \mathcal{O}\left(\frac{L}{S}\ln\frac{|\phi_0|}{\epsilon}\right).$$

Moreover, $\|x_t - x^*\| \to 0$ as $t \to +\infty$ at same rate.

Proof. Let $\epsilon > 0$ and $T = N_{\text{dual}} + N_{\text{full}} + N_{\text{constrained}}$ where N_{dual}, N_{full}, and $N_{\text{constrained}}$ are the number of dual steps (Line 13), full steps (Line 17), and constrained steps (Line 7) taken in total respectively, and $\epsilon_t := f(x_t) - \min_{\mathcal{H}} f$. Similarly to Braun et al. [2017], we introduce *epoch starts* at iteration $t = 0$ or any iteration immediately following a dual step. Our goal is to bound the number of epochs and the number of iterations within each epoch. Notice that $0 \leqslant \epsilon_{t+1} \leqslant \epsilon_t$ and $\phi_t \leqslant \phi_{t+1} \leqslant 0$ for $t \in \mathbb{N}$.

Denote $\{x^*\} := \arg\min_{\mathcal{H}} f$. Let $t \in \mathbb{N}$ be an iteration of the algorithm and $v_t^{\text{FW}} \in \arg\min_{v \in \mathcal{D}'}\langle \nabla f(x_t), v \rangle = \arg\min_{z \in \text{conv}(\mathcal{D}')}\langle \nabla f(x_t), z \rangle$. We can assume that $f(x_t) > f(x^*)$ otherwise the iterates have already converged. By convexity, $\langle \nabla f(x_t), x^* - x_t \rangle < 0$. Since $0 \in \text{int}(\text{conv}(\mathcal{D}'))$, there exists $r > 0$ such that $\mathcal{B}(0, r) \subseteq \text{conv}(\mathcal{D}')$. Therefore, $\frac{r(x^* - x_t)}{2\|x^* - x_t\|} \in \text{conv}(\mathcal{D}')$ so

$$\min_{z \in \text{conv}(\mathcal{D}')}\langle \nabla f(x_t), z \rangle = \langle \nabla f(x_t), v_t^{\text{FW}} \rangle \leqslant \left\langle \nabla f(x_t), \frac{r(x^* - x_t)}{2\|x^* - x_t\|} \right\rangle < 0$$

and it follows that

$$\frac{\langle \nabla f(x_t), x_t - x^* \rangle}{\|x_t - x^*\|} \leqslant -\frac{2}{r}\langle \nabla f(x_t), v_t^{\text{FW}} \rangle. \tag{4}$$

By Corollary B.3,

$$f(x^*) \geqslant f(x_t) - \frac{\langle \nabla f(x_t), x_t - x^* \rangle^2}{2S\|x_t - x^*\|^2}.$$

Combining with (4) we obtain

$$\epsilon_t = f(x_t) - f(x^*) \leqslant \frac{2}{r^2 S}\langle \nabla f(x_t), v_t^{\text{FW}} \rangle^2. \tag{5}$$

Let t be a dual step (Line 13). Then the weak-separation oracle call (Line 11) yields $\langle \nabla f(x_t), v_t^{\text{FW}} \rangle \geqslant \phi_t$. By (5) and Line 15,

$$\epsilon_t \leqslant \frac{2}{r^2 S}\phi_t^2 \tag{6}$$

$$= \frac{2}{r^2 S}\left(\frac{\phi_0}{\tau^{n_{\text{dual}}}}\right)^2 \tag{7}$$

where n_{dual} is the number of dual steps taken before t. Therefore, by (7) and since $\tau > 1$, N_{dual} is finite with

$$N_{\text{dual}} \leqslant \left\lceil \frac{1}{2}\log_\tau\left(\frac{2\phi_0^2}{r^2 S \epsilon}\right)\right\rceil. \tag{8}$$

If a full step is taken (Line 17), then the weak-separation oracle (Line 11) returns $v_t \in \mathcal{D}'$ such that $\langle \nabla f(x_t), v_t \rangle \leqslant \phi_t/\kappa$. By smoothness,

$$f(x_{t+1}) = \min_{\gamma \in \mathbb{R}} f(x_t + \gamma v_t)$$

$$\leqslant \min_{\gamma \in \mathbb{R}} f(x_t) + \gamma\langle \nabla f(x_t), v_t \rangle + \frac{L}{2}\gamma^2\|v_t\|^2$$

$$= f(x_t) - \frac{\langle \nabla f(x_t), v_t \rangle^2}{2L\|v_t\|^2}$$

$$\leqslant f(x_t) - \frac{(\phi_t/\kappa)^2}{2L(D_{\mathcal{D}'}/2)^2}$$

where we used $\|v_t\| \leqslant D_{\mathcal{D}'}/2$ (by symmetry). Therefore, the primal progress is at least

$$f(x_t) - f(x_{t+1}) \geqslant \frac{2\phi_t^2}{\kappa^2 L D_{\mathcal{D}'}^2}. \tag{9}$$

21

Last, if a constrained step is taken (Line 7), then by smoothness,

$$f(x_{t+1}) = \min_{\gamma \in \mathbb{R}} f\big(x_t + \gamma \widetilde{\nabla} f(x_t)\big)$$

$$\leqslant \min_{\gamma \in \mathbb{R}} f(x_t) + \gamma \langle \nabla f(x_t), \widetilde{\nabla} f(x_t) \rangle + \frac{L}{2}\gamma^2 \big\|\widetilde{\nabla} f(x_t)\big\|^2$$

$$= f(x_t) - \frac{\langle \nabla f(x_t), \widetilde{\nabla} f(x_t) \rangle^2}{2L\big\|\widetilde{\nabla} f(x_t)\big\|^2}$$

$$= f(x_t) - \frac{\big\|\widetilde{\nabla} f(x_t)\big\|^2}{2L}$$

$$\leqslant f(x_t) - \frac{\langle \widetilde{\nabla} f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle^2}{2L\big\|v_t^{\text{FW-}\mathcal{S}}\big\|^2}$$

$$= f(x_t) - \frac{\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle^2}{2L\big\|v_t^{\text{FW-}\mathcal{S}}\big\|^2}$$

$$\leqslant f(x_t) - \frac{(\phi_t/\eta)^2}{2L(D_{\mathcal{D}'}/2)^2}$$

where the last three lines respectively come from the Cauchy-Schwarz inequality, $v_t^{\text{FW-}\mathcal{S}} \in \mathrm{span}(\mathcal{S}_t)$, $\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle \leqslant \phi_t/\eta$ (Line 5), and $\big\|v_t^{\text{FW-}\mathcal{S}}\big\| \leqslant D_{\mathcal{D}'}/2$ (by symmetry). Therefore, the primal progress is at least

$$f(x_t) - f(x_{t+1}) \geqslant \frac{2\phi_t^2}{\eta^2 L D_{\mathcal{D}'}^2}. \tag{10}$$

whose lower bound only differs by a constant factor $(\kappa/\eta)^2$ from that of a full step (9).

Now, we have

$$T = N_{\text{dual}} + N_{\text{full}} + N_{\text{constrained}}$$

$$= N_{\text{dual}} + \sum_{\substack{t=0 \\ t \text{ epoch start}}}^{T-1} \left(N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \right) \tag{11}$$

where $N_{\text{full}}^{(t)}$ and $N_{\text{constrained}}^{(t)}$ are the number of full steps and constrained steps taken during epoch t respectively. Let $t > 0$ be an epoch start. Thus, $t - 1$ is a dual step. By (6), since $x_t = x_{t-1}$ and $\phi_t = \phi_{t-1}/\tau$,

$$\epsilon_t \leqslant \frac{2\tau^2 \phi_t^2}{r^2 S}. \tag{12}$$

This also holds for $t = 0$ by (5) and Line 2 (and actually for all $t \in \mathbb{N}$). By (9) and (10), since $\phi_s = \phi_t$ for all nondual steps s in the epoch starting at t,

$$\epsilon_t \geqslant \sum_{s \in \text{epoch}(t)} \big(f(x_s) - f(x_{s+1})\big)$$

$$\geqslant \left(N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \right) \frac{2\phi_t^2}{\max\{\kappa^2, \eta^2\} L D_{\mathcal{D}'}^2}. \tag{13}$$

Combining (12) and (13),

$$N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \leqslant \frac{\tau^2 \max\{\kappa^2, \eta^2\} L D_{\mathcal{D}'}^2}{r^2 S}. \tag{14}$$

Therefore, by (11) and (14),

$$T \leqslant N_{\text{dual}} + (N_{\text{dual}} + 1) \frac{\tau^2 \max\{\kappa^2, \eta^2\} L D_{\mathcal{D}'}^2}{r^2 S}.$$

By (8), we conclude that the algorithm converges with

$$T = \mathcal{O}\left(\frac{L}{S} \ln \frac{|\phi_0|}{\epsilon}\right).$$

Finally, by strong convexity and since $\nabla f(x^*) = 0$,

$$\frac{S}{2}\|x_t - x^*\|^2 \leqslant f(x_t) - f(x^*) - \langle \nabla f(x^*), x_t - x^* \rangle = \epsilon_t$$

for all $t \in \mathbb{N}$. Thus, $\|x_t - x^*\| \to 0$ as $t \to +\infty$. $\qquad\square$

D   Main proofs

Fact 2.1. *Let $f : \mathcal{H} \to \mathbb{R}$ be smooth of order $\ell > 1$, convex, and sharp of order $\theta \in\,]0, 1[$ on $\mathcal{K}$. Then $\theta \in\,]0, 1/\ell]$.*

Proof. Let $x \in \mathcal{K} \backslash \arg\min_{\mathcal{H}} f$ and $x^* := \mathrm{proj}_{\arg\min_{\mathcal{H}} f}(x)$. By sharpness, smoothness, and $\nabla f(x^*) = 0$,

$$\mathrm{dist}\left(x, \arg\min_{\mathcal{H}} f\right) = \|x - x^*\| \leqslant C(f(x) - f(x^*))^\theta \leqslant C\left(\frac{L}{\ell}\right)^\theta \|x - x^*\|^{\ell\theta}.$$

Therefore,

$$\frac{1}{C}\left(\frac{\ell}{L}\right)^\theta \leqslant \|x - x^*\|^{\ell\theta - 1}.$$

As the left-hand side is constant and x can be arbitrarily close to x^*, we conclude that $\ell\theta \leqslant 1$. $\quad\square$

Theorem 3.2 (Smooth convex case). *Let $\mathcal{D} \subset \mathcal{H}$ be a dictionary such that $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{D}'))$ and let $f : \mathcal{H} \to \mathbb{R}$ be smooth of order $\ell > 1$, convex, and coercive. Then the Blended Matching Pursuit algorithm (Algorithm 3) ensures that $f(x_t) - \min_{\mathcal{H}} f \leqslant \epsilon$ for all $t \geqslant T$ where $T = \mathcal{O}\big((L/\epsilon)^{1/(\ell-1)}\big)$.*

Proof. Let $\epsilon > 0$ and $T = N_{\mathrm{dual}} + N_{\mathrm{full}} + N_{\mathrm{constrained}} \in \mathbb{N} \cup \{+\infty\}$ where N_{dual}, N_{full}, and $N_{\mathrm{constrained}}$ are the number of dual steps (Line 13), full steps (Line 17), and constrained steps (Line 7) taken in total respectively. The objective f is continuous and coercive so $\arg\min_{\mathcal{H}} f \neq \varnothing$. Let $\epsilon_t := f(x_t) - \min_{\mathcal{H}} f$ for $t \in \mathbb{N}$. Similarly to Braun et al. [2017], we introduce *epoch starts* at iteration $t = 0$ or any iteration immediately following a dual step. Our goal is to bound the number of epochs and the number of iterations within each epoch. Notice that $0 \leqslant \epsilon_{t+1} \leqslant \epsilon_t$ and $\phi_t \leqslant \phi_{t+1} \leqslant 0$ for $t \in \mathbb{N}$.

Let $x^* \in \arg\min_{\mathcal{H}} f$. The function f is coercive and $f(x_{t+1}) \leqslant f(x_t)$ for $t \in \mathbb{N}$, so by Fact B.1 the sequence of iterates is bounded. Define $\rho := \sup_{t \in \mathbb{N}} \|x_t - x^*\| < +\infty$. Note that ρ is independent of T. Let $t \in \mathbb{N}$ be an iteration of the algorithm, and $v_t^{\mathrm{FW}} \in \arg\min_{v \in \mathcal{D}'}\langle \nabla f(x_t), v \rangle = \arg\min_{z \in \mathrm{conv}(\mathcal{D}')}\langle \nabla f(x_t), z \rangle$. We can assume that $f(x_t) > f(x^*)$ otherwise the iterates have already converged. By convexity, $\langle \nabla f(x_t), x^* - x_t \rangle < 0$. Since $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{D}'))$, there exists $r > 0$ such that $\mathcal{B}(0, r) \subseteq \mathrm{conv}(\mathcal{D}')$. Thus, $\frac{r(x^* - x_t)}{2\|x^* - x_t\|} \in \mathrm{conv}(\mathcal{D}')$ so

$$\min_{z \in \mathrm{conv}(\mathcal{D}')}\langle \nabla f(x_t), z \rangle = \left\langle \nabla f(x_t), v_t^{\mathrm{FW}} \right\rangle \leqslant \left\langle \nabla f(x_t), \frac{r(x^* - x_t)}{2\|x^* - x_t\|} \right\rangle < 0$$

i.e.,

$$\langle \nabla f(x_t), x_t - x^* \rangle \leqslant \frac{2\|x_t - x^*\|}{r}\left\langle -\nabla f(x_t), v_t^{\mathrm{FW}} \right\rangle$$

$$\leqslant \frac{2\rho}{r}\left\langle -\nabla f(x_t), v_t^{\mathrm{FW}} \right\rangle. \tag{15}$$

By convexity,

$$f(x_t) - f(x^*) \leqslant \langle \nabla f(x_t), x_t - x^* \rangle$$

23

so with (15),

$$\epsilon_t \leqslant \frac{2\rho}{r} \left\langle -\nabla f(x_t), v_t^{\text{FW}} \right\rangle. \tag{16}$$

Let t be a dual step (Line 13). Then the weak-separation oracle call (Line 11) yields $\left\langle \nabla f(x_t), v_t^{\text{FW}} \right\rangle \geqslant \phi_t$. By (16) and Line 15,

$$\epsilon_t \leqslant \frac{2\rho}{r} |\phi_t| \tag{17}$$

$$= \frac{2\rho}{r} \frac{|\phi_0|}{\tau^{n_{\text{dual}}}} \tag{18}$$

where n_{dual} is the number of dual steps taken before t. Therefore, by (18) and since $\tau > 1$,

$$N_{\text{dual}} \leqslant \left\lceil \log_\tau \left(\frac{2\rho|\phi_0|}{r\epsilon} \right) \right\rceil. \tag{19}$$

If a full step is taken (Line 17), then the weak-separation oracle (Line 11) returns $v_t \in \mathcal{D}'$ such that $\langle \nabla f(x_t), v_t \rangle \leqslant \phi_t/\kappa$. By smoothness and using Fact B.2 with $\bar{\ell} := \ell/(\ell-1) > 1$,

$$
\begin{aligned}
f(x_{t+1}) &\leqslant \min_{\gamma \in \mathbb{R}_+} f(x_t + \gamma v_t) \\
&\leqslant \min_{\gamma \in \mathbb{R}_+} f(x_t) + \gamma \langle \nabla f(x_t), v_t \rangle + \frac{L}{\ell}\gamma^\ell \|v_t\|^\ell \\
&= f(x_t) - \frac{\langle -\nabla f(x_t), v_t \rangle^{\bar{\ell}}}{\bar{\ell} L^{\bar{\ell}-1} \|v_t\|^{\bar{\ell}}} \\
&\leqslant f(x_t) - \frac{|\phi_t/\kappa|^{\bar{\ell}}}{\bar{\ell} L^{\bar{\ell}-1} (D_{\mathcal{D}'}/2)^{\bar{\ell}}}
\end{aligned}
$$

where we used $\|v_t\| \leqslant D_{\mathcal{D}'}/2$ (by symmetry). Therefore, the primal progress is at least

$$f(x_t) - f(x_{t+1}) \geqslant \frac{2^{\bar{\ell}}|\phi_t|^{\bar{\ell}}}{\bar{\ell}\kappa^{\bar{\ell}} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}}. \tag{20}$$

Last, if a constrained step is taken (Line 7), then by smoothness and using Fact B.2 with $-\widetilde{\nabla} f(x_t)$,

$$
\begin{aligned}
f(x_{t+1}) &\leqslant \min_{\gamma \in \mathbb{R}_+} f\left(x_t - \gamma \widetilde{\nabla} f(x_t)\right) \\
&\leqslant \min_{\gamma \in \mathbb{R}_+} f(x_t) - \gamma \langle \nabla f(x_t), \widetilde{\nabla} f(x_t) \rangle + \frac{L}{\ell}\gamma^\ell \|\widetilde{\nabla} f(x_t)\|^\ell \\
&= f(x_t) - \frac{\langle \nabla f(x_t), \widetilde{\nabla} f(x_t) \rangle^{\bar{\ell}}}{\bar{\ell} L^{\bar{\ell}-1} \|\widetilde{\nabla} f(x_t)\|^{\bar{\ell}}} \\
&= f(x_t) - \frac{\|\widetilde{\nabla} f(x_t)\|^{\bar{\ell}}}{\bar{\ell} L^{\bar{\ell}-1}} \\
&\leqslant f(x_t) - \frac{|\langle \widetilde{\nabla} f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle|^{\bar{\ell}}}{\bar{\ell} L^{\bar{\ell}-1} \|v_t^{\text{FW-}\mathcal{S}}\|^{\bar{\ell}}} \\
&\leqslant f(x_t) - \frac{|\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle|^{\bar{\ell}}}{\bar{\ell} L^{\bar{\ell}-1} \|v_t^{\text{FW-}\mathcal{S}}\|^{\bar{\ell}}} \\
&\leqslant f(x_t) - \frac{|\phi_t/\eta|^{\bar{\ell}}}{\bar{\ell} L^{\bar{\ell}-1} (D_{\mathcal{D}'}/2)^{\bar{\ell}}}
\end{aligned}
$$

where the last three lines respectively come from the Cauchy-Schwarz inequality, $v_t^{\text{FW-}\mathcal{S}} \in \text{span}(\mathcal{S}_t)$, $\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle \leqslant \phi_t/\eta$ (Line 5), and $\left\| v_t^{\text{FW-}\mathcal{S}} \right\| \leqslant D_{\mathcal{D}'}/2$ (by symmetry). Therefore, the primal progress is at least

$$f(x_t) - f(x_{t+1}) \geqslant \frac{2^{\bar{\ell}}|\phi_t|^{\bar{\ell}}}{\bar{\ell}\eta^{\bar{\ell}} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}} \tag{21}$$

whose lower bound only differs by a constant factor $(\kappa/\eta)^{\bar{\ell}}$ from that of a full step (20).

Now, we have

$$\begin{aligned} T &= N_{\text{dual}} + N_{\text{full}} + N_{\text{constrained}} \\ &= N_{\text{dual}} + \sum_{\substack{t=0 \\ t \text{ epoch start}}}^{T-1} \left(N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \right) \end{aligned} \tag{22}$$

where $N_{\text{full}}^{(t)}$ and $N_{\text{constrained}}^{(t)}$ are the number of full steps and constrained steps taken during epoch t respectively. Let $t > 0$ be an epoch start. Thus, $t-1$ is a dual step. By (17), since $x_t = x_{t-1}$ and $\phi_t = \phi_{t-1}/\tau$,

$$\epsilon_t \leqslant \frac{2\rho\tau}{r}|\phi_t|. \tag{23}$$

This also holds for $t = 0$ by (16) and Line 2 (and actually for all $t \in [\![0, T]\!]$). By (20) and (21), since $\phi_s = \phi_t$ for all nondual steps s in the epoch starting at t,

$$\begin{aligned} \epsilon_t &\geqslant \sum_{s \in \text{epoch}(t)} \left(f(x_s) - f(x_{s+1}) \right) \\ &\geqslant \left(N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \right) \frac{2^{\bar{\ell}}|\phi_t|^{\bar{\ell}}}{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}} \end{aligned} \tag{24}$$

Combining (23) and (24),

$$N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \leqslant \frac{2\rho\tau}{r} \frac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}}|\phi_t|^{1-\bar{\ell}}. \tag{25}$$

Therefore, by (22), (25), and $\bar{\ell} > 1$,

$$\begin{aligned} T &\leqslant N_{\text{dual}} + \frac{2\rho\tau}{r} \frac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}} \sum_{t=0}^{N_{\text{dual}}} \left(\frac{|\phi_0|}{\tau^t} \right)^{1-\bar{\ell}} \\ &= N_{\text{dual}} + \frac{2\rho\tau}{r} \frac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}}|\phi_0|^{1-\bar{\ell}} \frac{\tau^{(\bar{\ell}-1)(N_{\text{dual}}+1)} - 1}{\tau^{\bar{\ell}-1} - 1}. \end{aligned}$$

By (19),

$$\begin{aligned} T &\leqslant \log_\tau\left(\frac{2\rho|\phi_0|}{r\epsilon} \right) + 1 + \frac{2\rho\tau}{r} \frac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}} \frac{|\phi_0|^{1-\bar{\ell}}}{\tau^{\bar{\ell}-1} - 1} \left(\tau^{(\bar{\ell}-1)\left(\log_\tau\left(\frac{2\rho|\phi_0|}{r\epsilon}\right)+2\right)} - 1 \right) \\ &= \log_\tau\left(\frac{2\rho|\phi_0|}{r\epsilon} \right) + 1 + \frac{2\rho\tau}{r} \frac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\} L^{\bar{\ell}-1} D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}} \frac{|\phi_0|^{1-\bar{\ell}}}{\tau^{\bar{\ell}-1} - 1} \left(\tau^{2(\bar{\ell}-1)} \left(\frac{2\rho|\phi_0|}{r\epsilon} \right)^{\bar{\ell}-1} - 1 \right). \end{aligned}$$

We conclude that the algorithm converges with

$$T = \mathcal{O}\left(\left(\frac{L}{\epsilon} \right)^{1/(\ell-1)} \right).$$

$\square$

25

Theorem 3.3 (Smooth convex sharp case). *Let $\mathcal{D} \subset \mathcal{H}$ be a dictionary such that $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{D}'))$ and let $f : \mathcal{H} \to \mathbb{R}$ be L-smooth of order $\ell > 1$, convex, coercive, and C-sharp of order $\theta \in \,]0, 1/\ell]$ on $\mathcal{K}$. Then the Blended Matching Pursuit algorithm (Algorithm 3) ensures that $f(x_t) - \min_{\mathcal{H}} f \leqslant \epsilon$ for all $t \geqslant T$ where*

$$
T = \begin{cases}
\mathcal{O}\left(C^{1/(1-\theta)} L^{1/(\ell-1)} \ln\left(\dfrac{C|\phi_0|}{\epsilon^{1-\theta}} \right) \right) & \text{if } \ell\theta = 1 \\[3ex]
\mathcal{O}\left(\left(\dfrac{C^\ell L}{\epsilon^{1-\ell\theta}} \right)^{1/(\ell-1)} \right) & \text{if } \ell\theta < 1.
\end{cases}
$$

Moreover, $\mathrm{dist}(x_t, \arg\min_{\mathcal{H}} f) \to 0$ as $t \to +\infty$ at same rate.

Proof. Let $\epsilon > 0$. By Theorem 3.2, there exists $T \in \mathbb{N}$ such that $f(x_T) - \min_{\mathcal{H}} f \leqslant \epsilon$. Let $\epsilon_t := f(x_t) - \min_{\mathcal{H}} f$ for $t \in [\![0, T]\!]$ and $T = N_{\mathrm{dual}} + N_{\mathrm{full}} + N_{\mathrm{constrained}}$ where N_{dual}, N_{full}, and $N_{\mathrm{constrained}}$ are the number of dual steps (Line 13), full steps (Line 17), and constrained steps (Line 7) taken in total respectively. Similarly to Braun et al. [2017], we introduce *epoch starts* at iteration $t = 0$ or any iteration immediately following a dual step. Our goal is to bound the number of epochs and the number of iterations within each epoch. Notice that $0 \leqslant \epsilon_{t+1} \leqslant \epsilon_t$ and $\phi_t \leqslant \phi_{t+1} \leqslant 0$ for $t \in [\![0, T]\!]$.

Let $t \in [\![0, T]\!]$ be an iteration of the algorithm, $v_t^{\mathrm{FW}} \in \arg\min_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle = \arg\min_{z \in \mathrm{conv}(\mathcal{D}')} \langle \nabla f(x_t), z \rangle$, and $x_t^* := \mathrm{proj}_{\arg\min_{\mathcal{H}} f}(x_t)$. We can assume that $f(x_t) > f(x_t^*)$ otherwise the iterates have already converged. By convexity, $\langle \nabla f(x_t), x_t^* - x_t \rangle < 0$. Since $0 \in \mathrm{int}(\mathrm{conv}(\mathcal{D}'))$, there exists $r > 0$ such that $\mathcal{B}(0, r) \subseteq \mathrm{conv}(\mathcal{D}')$. Therefore, $\frac{r(x_t^* - x_t)}{2\|x_t^* - x_t\|} \in \mathrm{conv}(\mathcal{D}')$ so

$$
\min_{z \in \mathrm{conv}(\mathcal{D}')} \langle \nabla f(x_t), z \rangle = \langle \nabla f(x_t), v_t^{\mathrm{FW}} \rangle \leqslant \left\langle \nabla f(x_t), \frac{r(x_t^* - x_t)}{2\|x_t^* - x_t\|} \right\rangle < 0
$$

i.e.,

$$
\frac{r\langle \nabla f(x_t), x_t - x_t^* \rangle}{-2\langle \nabla f(x_t), v_t^{\mathrm{FW}} \rangle} \leqslant \|x_t - x_t^*\|. \tag{26}
$$

The sharpness of f implies that $\arg\min_{\mathcal{H}} f \subset \mathrm{int}(\mathcal{K})$. Let $r_t^* \in \,]0, \|x_t - x_t^*\|[$ such that $\mathcal{B}(x_t^*, r_t^*) \subseteq \mathcal{K}$, and let $\rho := \min\{r_0^*/\|x_0 - x_0^*\|, \ldots, r_T^*/\|x_T - x_T^*\|\} \in \,]0, 1[$. Then, $x_t^* + \rho(x_t - x_t^*) \in \mathcal{B}(x_t^*, r_t^*) \subseteq \mathcal{K}$. By convexity, $x_t^* = \mathrm{proj}_{\arg\min_{\mathcal{H}} f}(x_t^* + \rho(x_t - x_t^*))$: indeed, $\langle x_t^* - x^*, x_t - x_t^* \rangle \geqslant 0$ for all $x^* \in \arg\min_{\mathcal{H}} f$ by the Hilbert projection theorem, thus

$$
\|(x_t^* + \rho(x_t - x_t^*)) - x^*\|^2 = \|x_t^* - x^*\|^2 + \rho^2\|x_t - x_t^*\|^2 + 2\rho\langle x_t^* - x^*, x_t - x_t^* \rangle
$$
$$
\geqslant \rho^2\|x_t - x_t^*\|^2 \tag{27}
$$

where (27) is an equality if and only if $x^* = x_t^*$. Hence, using sharpness,

$$
\begin{aligned}
\rho\|x_t - x_t^*\| &= \|(x_t^* + \rho(x_t - x_t^*)) - x_t^*\| \\
&\leqslant C\big(f(x_t^* + \rho(x_t - x_t^*)) - f(x_t^*)\big)^\theta \\
&\leqslant C\big(f(x_t^*) + \rho(f(x_t) - f(x_t^*)) - f(x_t^*)\big)^\theta \\
&= C\rho^\theta(f(x_t) - f(x_t^*))^\theta \\
&\leqslant C\rho^\theta\langle \nabla f(x_t), x_t - x_t^* \rangle^\theta
\end{aligned} \tag{28}
$$

where the second and last inequalities come from convexity. Combining with (26), we get

$$
\frac{r\langle \nabla f(x_t), x_t - x_t^* \rangle}{-2\langle \nabla f(x_t), v_t^{\mathrm{FW}} \rangle} \leqslant \frac{C}{\rho^{1-\theta}} \langle \nabla f(x_t), x_t - x_t^* \rangle^\theta
$$

so, by convexity, we obtain the primal bound

$$
f(x_t) - f(x_t^*) \leqslant \langle \nabla f(x_t), x_t - x_t^* \rangle \leqslant \frac{1}{\rho}\left(-\frac{2C}{r} \langle \nabla f(x_t), v_t^{\mathrm{FW}} \rangle \right)^{1/(1-\theta)}
$$

i.e.,

$$\epsilon_t \leqslant \frac{1}{\rho}\left(-\frac{2C}{r}\left\langle\nabla f(x_t), v_t^{\mathrm{FW}}\right\rangle\right)^{1/(1-\theta)}. \tag{29}$$

Let t be a dual step (Line 13). Then the weak-separation oracle call (Line 11) yields $\left\langle\nabla f(x_t), v_t^{\mathrm{FW}}\right\rangle \geqslant \phi_t$. By (29) and Line 15,

$$\epsilon_t \leqslant \frac{1}{\rho}\left(\frac{2C}{r}|\phi_t|\right)^{1/(1-\theta)} \tag{30}$$

$$= \frac{1}{\rho}\left(\frac{2C}{r}\frac{|\phi_0|}{\tau^{n_{\mathrm{dual}}}}\right)^{1/(1-\theta)} \tag{31}$$

where n_{dual} is the number of dual steps taken before t. Therefore, by (31) and since $\tau > 1$ and $\theta \in {]}0,1{[}$,

$$N_{\mathrm{dual}} \leqslant \left\lceil\log_\tau\left(\frac{2C|\phi_0|}{r\rho^{1-\theta}\epsilon^{1-\theta}}\right)\right\rceil. \tag{32}$$

If a full step is taken (Line 17), then the weak-separation oracle (Line 11) returns $v_t \in \mathcal{D}'$ such that $\langle\nabla f(x_t), v_t\rangle \leqslant \phi_t/\kappa$. By smoothness and using Fact B.2 and $\bar{\ell} := \ell/(\ell-1) > 1$,

$$f(x_{t+1}) \leqslant \min_{\gamma\in\mathbb{R}_+} f(x_t + \gamma v_t)$$

$$\leqslant \min_{\gamma\in\mathbb{R}_+} f(x_t) + \gamma\langle\nabla f(x_t), v_t\rangle + \frac{L}{\ell}\gamma^\ell\|v_t\|^\ell$$

$$= f(x_t) - \frac{\langle-\nabla f(x_t), v_t\rangle^{\bar{\ell}}}{\bar{\ell}L^{\bar{\ell}-1}\|v_t\|^{\bar{\ell}}}$$

$$\leqslant f(x_t) - \frac{|\phi_t/\kappa|^{\bar{\ell}}}{\bar{\ell}L^{\bar{\ell}-1}(D_{\mathcal{D}'}/2)^{\bar{\ell}}}$$

where we used $\|v_t\| \leqslant D_{\mathcal{D}'}/2$ (by symmetry). Therefore, the primal progress is at least

$$f(x_t) - f(x_{t+1}) \geqslant \frac{2^{\bar{\ell}}|\phi_t|^{\bar{\ell}}}{\bar{\ell}\kappa^{\bar{\ell}}L^{\bar{\ell}-1}D_{\mathcal{D}'}^{\bar{\ell}}}. \tag{33}$$

Last, if a constrained step is taken (Line 7), then by smoothness and using Fact B.2,

$$f(x_{t+1}) \leqslant \min_{\gamma\in\mathbb{R}_+} f\left(x_t - \gamma\widetilde{\nabla}f(x_t)\right)$$

$$\leqslant \min_{\gamma\in\mathbb{R}_+} f(x_t) - \gamma\langle\nabla f(x_t), \widetilde{\nabla}f(x_t)\rangle + \frac{L}{\ell}\gamma^\ell\|\widetilde{\nabla}f(x_t)\|^\ell$$

$$= f(x_t) - \frac{\langle\nabla f(x_t), \widetilde{\nabla}f(x_t)\rangle^{\bar{\ell}}}{\bar{\ell}L^{\bar{\ell}-1}\|\widetilde{\nabla}f(x_t)\|^{\bar{\ell}}}$$

$$= f(x_t) - \frac{\|\widetilde{\nabla}f(x_t)\|^{\bar{\ell}}}{\bar{\ell}L^{\bar{\ell}-1}}$$

$$\leqslant f(x_t) - \frac{|\langle\widetilde{\nabla}f(x_t), v_t^{\mathrm{FW\text{-}}\mathcal{S}}\rangle|^{\bar{\ell}}}{\bar{\ell}L^{\bar{\ell}-1}\|v_t^{\mathrm{FW\text{-}}\mathcal{S}}\|^{\bar{\ell}}}$$

$$= f(x_t) - \frac{|\langle\nabla f(x_t), v_t^{\mathrm{FW\text{-}}\mathcal{S}}\rangle|^{\bar{\ell}}}{\bar{\ell}L^{\bar{\ell}-1}\|v_t^{\mathrm{FW\text{-}}\mathcal{S}}\|^{\bar{\ell}}}$$

$$\leqslant f(x_t) - \frac{|\phi_t/\eta|^{\bar{\ell}}}{\bar{\ell}L^{\bar{\ell}-1}(D_{\mathcal{D}'}/2)^{\bar{\ell}}}$$

where the last three lines respectively come from the Cauchy-Schwarz inequality, $v_t^{\text{FW-}\mathcal{S}} \in \text{span}(\mathcal{S}_t)$, $\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle \leqslant \phi_t/\eta$ (Line 5), and $\|v_t^{\text{FW-}\mathcal{S}}\| \leqslant D_{\mathcal{D}'}/2$ (by symmetry). Therefore, the primal progress is at least

$$f(x_t) - f(x_{t+1}) \geqslant \frac{2^{\bar{\ell}}|\phi_t|^{\bar{\ell}}}{\bar{\ell}\eta^{\bar{\ell}}L^{\bar{\ell}-1}D_{\mathcal{D}'}^{\bar{\ell}}}. \tag{34}$$

whose lower bound only differs by a constant factor $(\kappa/\eta)^{\bar{\ell}}$ from that of a full step (33).

Now, we have

$$
\begin{aligned}
T &= N_{\text{dual}} + N_{\text{full}} + N_{\text{constrained}} \\
&= N_{\text{dual}} + \sum_{\substack{t=0 \\ t \text{ epoch start}}}^{T-1} \left(N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \right)
\end{aligned}
\tag{35}
$$

where $N_{\text{full}}^{(t)}$ and $N_{\text{constrained}}^{(t)}$ are the number of full steps and constrained steps taken during epoch t respectively. Let $t > 0$ be an epoch start. Thus, $t-1$ is a dual step. By (30), since $x_t = x_{t-1}$ and $\phi_t = \phi_{t-1}/\tau$,

$$\epsilon_t \leqslant \frac{1}{\rho}\left(\frac{2\tau C}{r}|\phi_t| \right)^{1/1-\theta}. \tag{36}$$

This also holds for $t = 0$ by (29) and Line 2 (and actually for all $t \in [\![0, T]\!]$). By (33) and (34), since $\phi_s = \phi_t$ for all nondual steps s in the epoch starting at t,

$$
\begin{aligned}
\epsilon_t &\geqslant \sum_{s \in \text{epoch}(t)} \left(f(x_s) - f(x_{s+1}) \right) \\
&\geqslant \left(N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \right) \frac{2^{\bar{\ell}}|\phi_t|^{\bar{\ell}}}{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\}L^{\bar{\ell}-1}D_{\mathcal{D}'}^{\bar{\ell}}}.
\end{aligned}
\tag{37}
$$

Combining (36) and (37),

$$N_{\text{full}}^{(t)} + N_{\text{constrained}}^{(t)} \leqslant \frac{1}{\rho}\left(\frac{2\tau C}{r} \right)^{1/(1-\theta)} \frac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\}L^{\bar{\ell}-1}D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}}|\phi_t|^{1/(1-\theta)-\bar{\ell}}. \tag{38}$$

Therefore, by (35) and (38),

$$T \leqslant N_{\text{dual}} + \frac{1}{\rho}\left(\frac{2\tau C}{r} \right)^{1/(1-\theta)} \frac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\}L^{\bar{\ell}-1}D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}} \sum_{t=0}^{N_{\text{dual}}} \left(\frac{|\phi_0|}{\tau^t} \right)^{1/(1-\theta)-\bar{\ell}} \tag{39}$$

$$= \begin{cases} N_{\text{dual}} + \dfrac{1}{\rho}\left(\dfrac{2\tau C}{r} \right)^{1/(1-\theta)} \dfrac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\}L^{\bar{\ell}-1}D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}} (N_{\text{dual}} + 1) & \text{if } \ell\theta = 1 \\[3ex] N_{\text{dual}} + \dfrac{1}{\rho}\left(\dfrac{2\tau C}{r} \right)^{1/(1-\theta)} \dfrac{\bar{\ell}\max\{\kappa^{\bar{\ell}}, \eta^{\bar{\ell}}\}L^{\bar{\ell}-1}D_{\mathcal{D}'}^{\bar{\ell}}}{2^{\bar{\ell}}}|\phi_0|^{1/(1-\theta)-\bar{\ell}} \dfrac{\left(\tau^{\bar{\ell}-1/(1-\theta)} \right)^{N_{\text{dual}}+1} - 1}{\tau^{\bar{\ell}-1/(1-\theta)} - 1} & \text{if } \ell\theta < 1 \end{cases}$$

where, if $\alpha := \frac{1-\ell\theta}{(\ell-1)(1-\theta)} = \bar{\ell} - \frac{1}{1-\theta}$, by (32) we have

$$
\begin{aligned}
\left(\tau^{\bar{\ell}-1/(1-\theta)} \right)^{N_{\text{dual}}+1} &= (\tau^{\alpha})^{N_{\text{dual}}+1} \\
&= \exp\left(\alpha \ln(\tau)(N_{\text{dual}}+1) \right) \\
&\leqslant \exp\left(\alpha \ln(\tau)\left(\log_\tau\left(\frac{2C|\phi_0|}{r\rho^{1-\theta}\epsilon^{1-\theta}} \right) + 2 \right) \right) \\
&= \exp\left(\alpha \ln\left(\frac{2C|\phi_0|}{r\rho^{1-\theta}\epsilon^{1-\theta}} \right) + 2\alpha \ln(\tau) \right) \\
&= \tau^{2\alpha}\left(\frac{2C|\phi_0|}{r\rho^{1-\theta}\epsilon^{1-\theta}} \right)^{\alpha}.
\end{aligned}
$$

By (32), we conclude that

$$
T = \begin{cases} \mathcal{O}\left(C^{1/(1-\theta)}L^{1/(\ell-1)}\ln\left(\dfrac{C|\phi_0|}{\epsilon^{1-\theta}}\right)\right) & \text{if } \ell\theta = 1 \\ \mathcal{O}\left(\left(\dfrac{C^\ell L}{\epsilon^{1-\ell\theta}}\right)^{1/(\ell-1)}\right) & \text{if } \ell\theta < 1. \end{cases}
$$

Finally, by (28),

$$
\|x_t - x_t^*\| \leqslant \frac{C}{\rho^{1-\theta}}\epsilon_t^\theta
$$

for all $t \in \mathbb{N}$. Thus, $\|x_t - x_t^*\| \to 0$ as $t \to +\infty$. $\qquad\square$