Greedy Algorithms for Cone Constrained Optimization with Convergence Guarantees

Article · May 2017

4 authors, including:

Martin Jaggi
École Polytechnique Fédérale de Lausanne
109 PUBLICATIONS   **3,121** CITATIONS

Some of the authors of this publication are also working on these related projects:

Frank-Wolfe View project

Greedy Algorithms for Cone Constrained Optimization with Convergence Guarantees

Francesco Locatello
MPI for Intelligent Systems - ETH Zurich
locatelf@ethz.ch

Michael Tschannen
ETH Zurich
michaelt@nari.ee.ethz.ch

Gunnar Rätsch
ETH Zurich
raetsch@inf.ethz.ch

Martin Jaggi
EPFL
martin.jaggi@epfl.ch

Abstract

Greedy optimization methods such as Matching Pursuit (MP) and Frank-Wolfe (FW) algorithms regained popularity in recent years due to their simplicity, effectiveness and theoretical guarantees. MP and FW address optimization over the *linear span* and the *convex hull* of a set of atoms, respectively. In this paper, we consider the intermediate case of optimization over the *convex cone*, parametrized as the conic hull of a generic atom set, leading to the first principled definitions of non-negative MP algorithms for which we give explicit convergence rates and demonstrate excellent empirical performance. In particular, we derive sublinear ($\mathcal{O}(1/t)$) convergence on general smooth and convex objectives, and linear convergence ($\mathcal{O}(e^{-t})$) on strongly convex objectives, in both cases for general sets of atoms. Furthermore, we establish a clear correspondence of our algorithms to known algorithms from the MP and FW literature. Our novel algorithms and analyses target general atom sets and general objective functions, and hence are directly applicable to a large variety of learning settings.

1 Introduction

In recent years, greedy optimization algorithms have attracted significant interest in the domains of signal processing and machine learning thanks to their ability to process very large data sets. Arguably two of the most popular representatives are Frank-Wolfe (FW) [8, 14] and Matching Pursuit (MP) algorithms [22], in particular Orthogonal MP (OMP) [6, 35]. While the former targets minimization of a convex function over *bounded convex sets*, the latter apply to minimization over a *linear subspace*. In both cases, the domain is commonly parametrized by a set of atoms or dictionary elements, and in each iteration, both algorithms rely on querying a so-called *linear minimization oracle* (LMO) to find the direction of steepest descent in the set of atoms. The iterate is then updated as a *linear* or *convex combination*, respectively, of previous iterates and the newly obtained atom from the LMO. The particular choice of the atom set allows to encode structure such as sparsity and non-negativity (of the atoms) into the solution. This enables control of the trade-off between the amount of structure in the solution and approximation quality via the number of iterations, which was found useful in a large variety of use cases including structured matrix and tensor factorizations [36, 38, 39, 11].

In this paper, we target an important "intermediate case" between the two domain parameterizations given by the *linear span* and the *convex hull* of an atom set, namely the parameterization of the optimization domain as the *conic hull* of a possibly infinite atom set. In this case, the solution can be represented as a *non-negative* linear combination of the atoms, which is desirable in many applications, e.g., due to the physics underlying the problem at hand, or for the sake of interpretability.

Concrete examples include unmixing problems [7, 9, 1], model selection [21], and matrix and tensor factorizations [2, 16]. However, existing convergence analyses do not apply to the currently used greedy algorithms. In particular, all existing MP variants for the conic hull case [3, 25, 37] are not guaranteed to converge and may get stuck far away from the optimum (this can be observed in the experiments in Section 6). From a theoretical perspective, this intermediate case is of paramount interest in the context of MP and FW algorithms. Indeed, the atom set is not guaranteed to contain an atom aligned with a descent direction for all possible suboptimal iterates, as is the case when the optimization domain is the linear span or the convex hull of the atom set [26, 20]. Hence, while conic constraints have been widely studied in the context of a manifold of different applications, none of the existing greedy algorithms enjoys explicit convergence rates.

We propose and analyze new MP algorithms tailored for the minimization of smooth convex functions over the conic hull of an atom set. Specifically, our key contributions are:

- We propose the first (non-orthogonal) MP algorithm for optimization over conic hulls guaranteed to converge, and prove a corresponding *sublinear* convergence rate with *explicit constants*. Surprisingly, convergence is achieved without increasing computational complexity compared to ordinary MP.
- We propose new away-step, pairwise, and fully corrective MP variants, inspired by variants of FW [17] and generalized MP [20], respectively, that allow for different degrees of weight corrections for previously selected atoms. We derive corresponding sublinear and linear (for strongly convex objectives) convergence rates that solely depend on the geometry of the atom set.
- All our algorithms apply to general smooth convex functions. This is in contrast to all prior work on non-negative MP, which targets quadratic objectives [3, 25, 37]. Furthermore, if the conic hull of the atom set equals its linear span, we recover both algorithms and rates derived in [20] for generalized MP variants.
- We make no assumptions on the atom set which is simply a subset of a Hilbert space, in particular we do not assume the atom set to be finite.

Before presenting our algorithms (Section 3) along with the corresponding convergence guarantees (Section 4), we briefly review generalized MP variants. A detailed discussion of related work can be found in Section 5 followed by illustrative experiments on a least squares problem on synthetic data, and non-negative matrix factorization as well as non-negative garrote logistic regression as applications examples on real data (numerical evaluations of more applications and the dependency between constants in the rate and empirical convergence can be found in the supplementary material).

Notation. Given a non-empty subset $\mathcal{A}$ of some Hilbert space, let $\mathrm{conv}(\mathcal{A})$ be the convex hull of $\mathcal{A}$, and let $\mathrm{lin}(\mathcal{A})$ denote its linear span. Given a closed set $\mathcal{A}$, we call its diameter $\mathrm{diam}(\mathcal{A}) = \max_{\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{A}} \|\mathbf{z}_1 - \mathbf{z}_2\|$ and its radius $\mathrm{radius}(\mathcal{A}) = \max_{\mathbf{z} \in \mathcal{A}} \|\mathbf{z}\|$. $\|\mathbf{x}\|_{\mathcal{A}} := \inf\{c > 0 \colon \mathbf{x} \in c \cdot \mathrm{conv}(\mathcal{A})\}$ is the atomic norm of $\mathbf{x}$ over a set $\mathcal{A}$ (also known as the gauge function of $\mathrm{conv}(\mathcal{A})$). We call a subset $\mathcal{A}$ of a Hilbert space symmetric if it is closed under negation.

2 Review of Matching Pursuit Variants

Let $\mathcal{H}$ be a Hilbert space with associated inner product $\langle \mathbf{x}, \mathbf{y} \rangle$, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{H}$. The inner product induces the norm $\|\mathbf{x}\|^2 := \langle \mathbf{x}, \mathbf{x} \rangle$, $\forall \mathbf{x} \in \mathcal{H}$. Let $\mathcal{A} \subset \mathcal{H}$ be a compact set (the "set of atoms" or dictionary) and let $f \colon \mathcal{H} \to \mathbb{R}$ be convex and L-smooth (L-Lipschitz gradient in the finite dimensional case). If $\mathcal{H}$ is an infinite-dimensional Hilbert space, then f is assumed to be *Fréchet differentiable*. The generalized MP algorithm studied in [20], presented in Algorithm 1, solves the following optimization problem:

$$\min_{\mathbf{x} \in \mathrm{lin}(\mathcal{A})} f(\mathbf{x}). \tag{1}$$

In each iteration, MP queries a linear minimization oracle (LMO) solving the following linear problem:

$$\mathrm{LMO}_{\mathcal{A}}(\mathbf{y}) := \arg\min_{\mathbf{z} \in \mathcal{A}} \langle \mathbf{y}, \mathbf{z} \rangle \tag{2}$$

for a given query $\mathbf{y} \in \mathcal{H}$. The MP update step minimizes a quadratic upper bound $g_{\mathbf{x}_t}(\mathbf{x}) = f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}_t\|^2$ of f at $\mathbf{x}_t$, where L is an upper bound on the smoothness

constant of f with respect to a chosen norm $\|\cdot\|$. Optimizing this norm problem instead of f directly allows for substantial efficiency gains in the case of complicated f. For symmetric $\mathcal{A}$ and for $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$, $\mathbf{y} \in \mathcal{H}$, Algorithm 1 recovers MP (Variant 0) [22] and OMP (Variant 1) [6, 35], see [20] for details.

Algorithm 1 Norm-Corrective Generalized Matching Pursuit

1: **init** $\mathbf{x}_0 \in \text{lin}(\mathcal{A})$, and $\mathcal{S} := \{\mathbf{x}_0\}$
2: **for** $t = 0 \ldots T$
3:     Find $\mathbf{z}_t := (\text{Approx-})\text{LMO}_{\mathcal{A}}(\nabla f(\mathbf{x}_t))$
4:     $\mathcal{S} := \mathcal{S} \cup \{\mathbf{z}_t\}$
5:     Let $\mathbf{b} := \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$
6:     Variant 0:
        Update $\mathbf{x}_{t+1} := \underset{\substack{\mathbf{z} := \mathbf{x}_t + \gamma \mathbf{z}_t \\ \gamma \in \mathbb{R}}}{\arg\min} \|\mathbf{z} - \mathbf{b}\|^2$
7:     Variant 1:
        Update $\mathbf{x}_{t+1} := \underset{\mathbf{z} \in \text{lin}(\mathcal{S})}{\arg\min} \|\mathbf{z} - \mathbf{b}\|^2$
8:     *Optional:* Correction of some/all atoms $\mathbf{z}_{0\ldots t}$
9: **end for**

Approximate linear oracles. Solving the LMO defined in (2) exactly is often hard in practice, in particular when applied to matrix (or tensor) factorization problems, while approximate versions can be much more efficient. Algorithm 1 allows for an *approximate* LMO. For given quality parameter $\delta \in (0,1]$ and given direction $\mathbf{d} \in \mathcal{H}$, the approximate LMO for Algorithm 1 returns a vector $\tilde{\mathbf{z}} \in \mathcal{A}$ such that

$$\langle \mathbf{d}, \tilde{\mathbf{z}} \rangle \leq \delta \langle \mathbf{d}, \mathbf{z} \rangle, \qquad (3)$$

relative to $\mathbf{z} = \text{LMO}_{\mathcal{A}}(\mathbf{d})$ being an exact solution.

Discussion and limitations of MP. The analysis of the convergence of Algorithm 1 in [20] critically relies on the assumption that the origin is in the relative interior of $\text{conv}(\mathcal{A})$ with respect to its linear span. This assumption originates from the fact that the convergence of MP- and FW-type algorithms fundamentally depends on an *alignment assumption* of the search direction returned by the LMO (i.e., $\mathbf{z}_t$ in Algorithm 1) and the gradient of the objective at the current iteration (see *third premise* in [26]). Specifically, for Algorithm 1, the LMO is assumed to select a descent direction, i.e., $\langle \nabla f(\mathbf{x}_t), \mathbf{z}_t \rangle < 0$, so that the resulting weight (i.e., γ for Variant 0) is always positive. In this spirit, Algorithm 1 is a natural candidate to minimize f over the conic hull of $\mathcal{A}$. However, if the optimization domain is a cone, the alignment assumption does not hold as there may be non-stationary points $\mathbf{x}$ in the conic hull of $\mathcal{A}$ for which $\min_{\mathbf{z} \in \mathcal{A}} \langle \nabla f(\mathbf{x}), \mathbf{z} \rangle = 0$. Algorithm 1 is therefore not guaranteed to converge when applied to conic problems. The same issue arises for essentially all existing non-negative variants of MP, see, e.g., Alg. 2 in [25] and in Alg. 2 in [37]. We now present modifications corroborating this issue along with the resulting MP-type algorithms for conic problems and corresponding convergence guarantees.

3   Greedy Algorithms on Conic Hulls

The cone $\text{cone}(\mathcal{A} - \mathbf{y})$ tangent to the convex set $\text{conv}(\mathcal{A})$ at a point $\mathbf{y}$ is formed by the half-lines emanating from $\mathbf{y}$ and intersecting $\text{conv}(\mathcal{A})$ in at least one point distinct from $\mathbf{y}$. Without loss of generality we consider $\mathbf{0} \in \mathcal{A}$ and assume the set $\text{cone}(\mathcal{A})$ (i.e., $\mathbf{y} = \mathbf{0}$) to be closed. If $\mathcal{A}$ is finite the cone constraint can be written as $\text{cone}(\mathcal{A}) := \{\mathbf{x} : \mathbf{x} = \sum_{i=1}^{|\mathcal{A}|} \alpha_i \mathbf{a}_i \text{ s.t. } \mathbf{a}_i \in \mathcal{A}, \ \alpha_i \geq 0 \ \forall i\}$. We consider conic optimization problems of the form:

$$\min_{\mathbf{x} \in \text{cone}(\mathcal{A})} f(\mathbf{x}). \qquad (4)$$

Note that if the set $\mathcal{A}$ is symmetric or if the origin is in the relative interior of $\text{conv}(\mathcal{A})$ w.r.t. its linear span then $\text{cone}(\mathcal{A}) = \text{lin}(\mathcal{A})$. We will show later how our results recover known MP rates when the origin is in the relative interior of $\text{conv}(\mathcal{A})$.

As a first algorithm to solve problems of the form (4), we present the Non-Negative Generalized Matching Pursuit (NNMP) in Algorithm 2 which is an extension of MP to general f and non-negative weights.

Discussion: Algorithm 2 differs from Algorithm 1 (Variant 0) in line 4, adding the iteration-dependent atom $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}$ to the set of possible search directions[1]. We use the atomic norm for the

---

[1] This additional direction makes sense only if $\mathbf{x}_t \neq \mathbf{0}$. Therefore, we set $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} = 0$ if $\mathbf{x}_t = \mathbf{0}$, i.e., no direction is added.

Algorithm 2 Non-Negative Matching Pursuit

1: **init** $\mathbf{x}_0 = \mathbf{0} \in \mathcal{A}$
2: **for** $t = 0 \dots T$
3:     Find $\bar{\mathbf{z}}_t := (\text{Approx-})\text{LMO}_{\mathcal{A}}(\nabla f(\mathbf{x}_t))$
4:     $\mathbf{z}_t = \arg\min_{\mathbf{z} \in \left\{ \bar{\mathbf{z}}_t, \frac{-\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} \right\}} \langle \nabla f(\mathbf{x}_t), \mathbf{z} \rangle$
5:     $\gamma := \frac{\langle -\nabla f(\mathbf{x}_t), \mathbf{z}_t \rangle}{L \|\mathbf{z}_t\|^2}$
6:     Update $\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma \mathbf{z}_t$
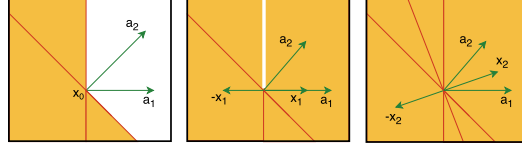7: **end for**

Figure 1: Two dimensional example for $T_{\mathcal{A}}(\mathbf{x}_t)$ where $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2\}$, for three different iterates $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$. The shaded area corresponds to $T_{\mathcal{A}}(\mathbf{x}_t)$ and the white area to $\text{lin}(\mathcal{A}) \setminus T_{\mathcal{A}}(\mathbf{x}_t)$.

normalization because it yields the best constant in the convergence rate. In practice, one can replace it with the Euclidean norm, which is often much less expensive to compute. This iteration-dependent additional search direction allows to reduce the weights of the atoms that were previously selected, thus admitting the algorithm to "move back" towards the origin while maintaining the cone constraint. This idea is informally explained here and formally studied in Section 4.1.

Recall the alignment assumption of the search direction and the gradient of the objective at the current iterate discussed in Section 2 (see also [26]). Algorithm 2 obeys this assumption. The intuition behind this is the following. Whenever $\mathbf{x}_t$ is not a minimizer of (4) and $\min_{\mathbf{z} \in \mathcal{A}} \langle \nabla f(\mathbf{x}_t), \mathbf{z} \rangle = 0$, the vector $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}$ is aligned with $\nabla f(\mathbf{x}_t)$ (i.e., $\langle \nabla f(\mathbf{x}_t), -\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} \rangle < 0$), preventing the algorithm from stopping at a suboptimal iterate. To make this intuition more formal, let us define the set of feasible descent directions of Algorithm 2 at a point $\mathbf{x} \in \text{cone}(\mathcal{A})$ as:

$$T_{\mathcal{A}}(\mathbf{x}) := \left\{ \mathbf{d} \in \mathcal{H} \colon \exists \mathbf{z} \in \mathcal{A} \cup \left\{ -\frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathcal{A}}} \right\} \text{ s.t. } \langle \mathbf{d}, \mathbf{z} \rangle < 0 \right\}. \tag{5}$$

If at some iteration $t = 0, 1, \dots$ the gradient $\nabla f(\mathbf{x}_t)$ is not in $T_{\mathcal{A}}(\mathbf{x}_t)$ Algorithm 2 terminates as $\min_{\mathbf{z} \in \mathcal{A}} \langle \mathbf{d}, \mathbf{z} \rangle = 0$ and $\langle \mathbf{d}, -\mathbf{x}_t \rangle \geq 0$ (which yields $\mathbf{z}_t = 0$). Even though, in general, not every direction in $\mathcal{H}$ is a feasible descent direction, $\nabla f(\mathbf{x}_t) \notin T_{\mathcal{A}}$ only occurs if $\mathbf{x}_t$ is a constrained minimum of Equation 4:

Lemma 1. *If* $\tilde{\mathbf{x}} \in \text{cone}(\mathcal{A})$ *and* $\nabla f(\tilde{\mathbf{x}}) \notin T_{\mathcal{A}}$ *then* $\tilde{\mathbf{x}}$ *is a solution to* $\min_{\mathbf{x} \in \text{cone}(\mathcal{A})} f(\mathbf{x})$.

Initializing Algorithm 2 with $\mathbf{x}_0 = \mathbf{0}$ guarantees that the iterates $\mathbf{x}_t$ always remain inside $\text{cone}(\mathcal{A})$ even though this is not enforced explicitly (by convexity of f, see proof of Theorem 2 in Appendix D for details).

Limitations of Algorithm 2: Let us call *active* the atoms which have nonzero weights in the representation of $\mathbf{x}_t = \sum_{i=0}^{t-1} \alpha_i \mathbf{z}_i$ computed by Algorithm 2. Formally, the set of active atoms is defined as $\mathcal{S} := \{\mathbf{z}_i \colon \alpha_i > 0, i = 0, 1, \dots, t-1\}$. The main drawback of Algorithm 2 is that when the direction $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}$ is selected, the weight of *all* active atoms is reduced. This can lead to the algorithm alternately selecting $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}$ and an atom from $\mathcal{A}$, thereby slowing down convergence in a similar manner as the *zig-zagging* phenomenon well-known in the Frank-Wolfe framework [17]. In order to achieve faster convergence we introduce the corrective variants of Algorithm 2.

3.1 Corrective Variants

To achieve faster (linear) convergence (see Section 4.2) we introduce variants of Algorithm 2, termed Away-steps MP (AMP) and Pairwise MP (PWMP), presented in Algorithm 3. Here, inspired by the away-steps and pairwise variants of FW [8, 17], instead of reducing the weights of the active atoms uniformly as in Algorithm 2, the LMO is queried a second time on the active set $\mathcal{S}$ to identify the direction of steepest ascent in $\mathcal{S}$. This allows, at each iteration, to reduce the weight of a previously selected atom (AMP) or swap weight between atoms (PWMP). This selective "reduction" or "swap of weight" helps to avoid the zig-zagging phenomenon which prevent Algorithm 2 from converging linearly.

At each iteration, Algorithm 3 updates the weights of $\mathbf{z}_t$ and $\mathbf{v}_t$ as $\alpha_{\mathbf{z}_t} = \alpha_{\mathbf{z}_t} + \gamma$ and $\alpha_{\mathbf{v}_t} = \alpha_{\mathbf{v}_t} - \gamma$, respectively. To ensure that $\mathbf{x}_{t+1} \in \text{cone}(\mathcal{A})$, γ has to be clipped according to the weight which is currently on $\mathbf{v}_t$, i.e., $\gamma_{\max} = \alpha_{\mathbf{v}_t}$. If $\gamma = \gamma_{\max}$, we set $\alpha_{\mathbf{v}_t} = 0$ and remove $\mathbf{v}_t$ from $\mathcal{S}$ as the atom $\mathbf{v}_t$ is no longer active. If $\mathbf{d}_t \in \mathcal{A}$ (i.e., we take a regular MP step and not an away step), the line search is unconstrained (i.e., $\gamma_{\max} = \infty$).

4

For both algorithm variants, the second LMO query increases the computational complexity. Note that an exact search on $\mathcal{S}$ is feasible in practice as $|\mathcal{S}|$ has at most t elements at iteration t.

Taking an additional computational burden allows to update the weights of all active atoms in the spirit of OMP. This approach is implemented in the Fully Corrective MP (FCMP), Algorithm 4.

<table>
<tr><td>

Algorithm 3 Away-steps (AMP) and Pairwise (PWMP) Non-Negative Matching Pursuit

</td><td>

Algorithm 4 Fully Corrective Non-Negative Matching Pursuit (FCMP)

</td></tr>
<tr><td>

1: **init** $\mathbf{x}_0 = \mathbf{0} \in \mathcal{A}$, and $\mathcal{S} := \{\mathbf{x}_0\}$
2: **for** $t = 0 \ldots T$
3:     Find $\mathbf{z}_t := $ (Approx-)$\text{LMO}_{\mathcal{A}}(\nabla f(\mathbf{x}_t))$
4:     Find $\mathbf{v}_t := $ (Approx-)$\text{LMO}_{\mathcal{S}}(-\nabla f(\mathbf{x}_t))$
5:     $\mathcal{S} = \mathcal{S} \cup \mathbf{z}_t$
6:     *AMP:* $\mathbf{d}_t = \arg\min_{\mathbf{d} \in \{\mathbf{z}_t, -\mathbf{v}_t\}} \langle \nabla f(\mathbf{x}_t), \mathbf{d} \rangle$
7:     *PWMP:* $\mathbf{d}_t = \mathbf{z}_t - \mathbf{v}_t$
8:     $\gamma := \min\left\{ \frac{\langle -\nabla f(\mathbf{x}_t), \mathbf{d}_t \rangle}{L\|\mathbf{d}_t\|^2}, \gamma_{\max} \right\}$
       ($\gamma_{\max}$ see text)
9:     Update $\alpha_{\mathbf{z}_t}$, $\alpha_{\mathbf{v}_t}$ and $\mathcal{S}$ according to γ
       (γ see text)
10:    Update $\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma \mathbf{d}_t$
11: **end for**

</td><td>

1: **init** $\mathbf{x}_0 = \mathbf{0} \in \mathcal{A}, \mathcal{S} = \{\mathbf{x}_0\}$
2: **for** $t = 0 \ldots T$
3:     Find $\mathbf{z}_t := $ (Approx-)$\text{LMO}_{\mathcal{A}}(\nabla f(\mathbf{x}_t))$
4:     $\mathcal{S} := \mathcal{S} \cup \{\mathbf{z}_t\}$
5:     *Variant 0:*
       $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \text{cone}(\mathcal{S})} \|\mathbf{x} - (\mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t))\|^2$
6:     *Variant 1:*
       $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \text{cone}(\mathcal{S})} f(\mathbf{x})$
7:     Remove atoms with zero weights from $\mathcal{S}$
8: **end for**

</td></tr>
</table>

At each iteration, Algorithm 4 maintains the set of active atoms $\mathcal{S}$ by adding $\mathbf{z}_t$ and removing atoms with zero weights after the update. In Variant 0, the algorithm minimizes the quadratic upper bound $g_{\mathbf{x}_t}(\mathbf{x})$ on f at $\mathbf{x}_t$ (see Section 2) imitating a gradient descent step with projection onto a "varying" target, i.e., $\text{cone}(\mathcal{S})$. In Variant 1, the original objective f is minimized over $\text{cone}(\mathcal{S})$ at each iteration, which is in general more efficient than minimizing f over $\text{cone}(\mathcal{A})$ using a generic solver for cone constrained problems. For $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$, $\mathbf{y} \in \mathcal{H}$, Variant 1 recovers Algorithm 1 in [37] and the OMP variant in [3] which both only apply to this specific objective f.

3.2 Computational Complexity

We briefly discuss the computational complexity of the algorithms we introduced. For $\mathcal{H} = \mathbb{R}^d$, sums and inner products have cost $O(d)$. Let us assume that each call of the LMO has cost C on the set $\mathcal{A}$ and $O(td)$ on $\mathcal{S}$. The variants 0 and 1 of FCMP solve a cone problem at each iteration with cost h_0 and h_1, respectively. In general, h_0 can be much smaller than h_1. In Table 1 we report the cost per iteration for every algorithm along with the asymptotic convergence rates derived in Section 4.

algorithm	cost per iteration	convergence	$k(t)$
NNMP	$C + O(d)$	$O(1/t)$	-
PWMP	$C + O(d + td)$	$O\left(e^{-\beta k(t)}\right)$	$\frac{t}{3|\mathcal{A}|!+1}$
AMP	$C + O(d + td)$	$O\left(e^{-\frac{\beta}{2} k(t)}\right)$	$t/2$
FCMP v. 0	$C + O(d) + h_0$	$O\left(e^{-\beta k(t)}\right)$	$\frac{t}{3|\mathcal{A}|!+1}$
FCMP v. 1	$C + O(d) + h_1$	$O\left(e^{-\beta k(t)}\right)$	t

Table 1: Computational complexity versus convergence rate (see Section 4) for strongly convex objectives

4 Convergence Rates

In this section, we present convergence guarantees for Algorithms 2, 3, and 4. All proofs are deferred to the Appendix in the supplementary material. We write $\mathbf{x}^\star \in \arg\min_{\mathbf{x} \in \text{cone}(\mathcal{A})} f(\mathbf{x})$ for an optimal solution. Our rates will depend on the atomic norm of the solution and the iterates of the respective algorithm variant:

$$\rho = \max\left\{ \|\mathbf{x}^\star\|_{\mathcal{A}}, \|\mathbf{x}_0\|_{\mathcal{A}} \ldots, \|\mathbf{x}_T\|_{\mathcal{A}} \right\}. \tag{6}$$

If the optimum is not unique, we consider $\mathbf{x}^\star$ to be one of largest atomic norm. A more intuitive and looser notion is to simply upper-bound ρ by the diameter of the level set of the initial iterate $\mathbf{x}_0$ measured by the atomic norm. Then, boundedness follows since the presented method is a descent method (due to Lemma 1 and line search on the quadratic upper bound, each iteration strictly

decreases the objective and our method stops only at the optimum). This justifies the statement $f(\mathbf{x}_t) \leq f(\mathbf{x}_0)$. Hence, ρ must be bounded for any sequence of iterates produced by the algorithm, and the convergence rates presented in this section are valid as T goes to infinity. A similar notion to measure the convergence of MP was established in [20]. All of our algorithms and rates can be made *affine invariant*. We defer this discussion to Appendix B.

4.1 Sublinear Convergence

We now present the convergence results for the non-negative and Fully-Corrective Matching Pursuit algorithms. Sublinear convergence of Algorithm 3 is addressed in Theorem 3.

Theorem 2. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set with $\mathbf{0} \in \mathcal{A}$, $\rho := \max\{\|\mathbf{x}^\star\|_\mathcal{A}, \|\mathbf{x}_0\|_\mathcal{A}, \ldots, \|\mathbf{x}_T\|_\mathcal{A},\}$ and f be L-smooth over $\rho \operatorname{conv}(\mathcal{A} \cup -\mathcal{A})$. Then, Algorithms 2 and 4 converge for $t \geq 0$ as*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{4 \left(\frac{2}{\delta} L \rho^2 \operatorname{radius}(\mathcal{A})^2 + \varepsilon_0\right)}{\delta t + 4},$$

where $\delta \in (0, 1]$ is the relative accuracy parameter of the employed approximate LMO *(see Equation (3)).*

Relation to FW rates. By rescaling $\mathcal{A}$ by a large enough factor $\tau > 0$, FW with $\tau \mathcal{A}$ as atom set could in principle be used to solve (4). In fact, for large enough τ, only the constraints of (4) become active when minimizing f over $\operatorname{conv}(\tau \mathcal{A})$. The sublinear convergence rate obtained with this approach is up to constants identical to that in Theorem 2 for our MP variants, see [14]. However, as the correct scaling is unknown, one has to either take the risk of choosing τ too small and hence failing to recover an optimal solution of (4), or to rely on too large τ which can result in slow convergence. In contrast, knowledge of ρ is not required to run our MP variants.

Relation to MP rates. If $\mathcal{A}$ is symmetric, we have that $\operatorname{lin}(\mathcal{A}) = \operatorname{cone}(\mathcal{A})$ and it is easy to show that the additional direction $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|}$ in Algorithm 2 is never selected. Therefore, Algorithm 2 becomes equivalent to Variant 0 of Algorithm 1, while Variant 1 of Algorithm 1 is equivalent to Variant 0 of Algorithm 4. The rate specified in Theorem 2 hence generalizes the sublinear rate in [20, Theorem 2] for symmetric $\mathcal{A}$.

4.2 Linear Convergence

We start by recalling some of the geometric complexity quantities that were introduced in the context of FW and are adapted here to the optimization problem we aim to solve (minimization over $\operatorname{cone}(\mathcal{A})$ instead of $\operatorname{conv}(\mathcal{A})$).

Directional Width. The directional width of a set $\mathcal{A}$ w.r.t. a direction $\mathbf{r} \in \mathcal{H}$ is defined as:

$$dirW(\mathcal{A}, \mathbf{r}) := \max_{\mathbf{s}, \mathbf{v} \in \mathcal{A}} \left\langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{s} - \mathbf{v} \right\rangle \tag{7}$$

Pyramidal Directional Width [17]**.** The Pyramidal Directional Width of a set $\mathcal{A}$ with respect to a direction $\mathbf{r}$ and a reference point $\mathbf{x} \in \operatorname{conv}(\mathcal{A})$ is defined as:

$$PdirW(\mathcal{A}, \mathbf{r}, \mathbf{x}) := \min_{\mathcal{S} \in \mathcal{S}_\mathbf{x}} dirW(\mathcal{S} \cup \{\mathbf{s}(\mathcal{A}, \mathbf{r})\}, \mathbf{r}), \tag{8}$$

where $\mathcal{S}_\mathbf{x} := \{\mathcal{S} \mid \mathcal{S} \subset \mathcal{A}$ and $\mathbf{x}$ is a proper convex combination of all the elements in $\mathcal{S}\}$ and $\mathbf{s}(\mathcal{A}, \mathbf{r}) := \max_{\mathbf{s} \in \mathcal{A}} \langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{s} \rangle$.

Inspired by the notion of pyramidal width in [17], which is the minimal pyramidal directional width computed over the set of feasible directions, we now define the cone width of a set $\mathcal{A}$ where only the generating faces (g-faces) of $\operatorname{cone}(\mathcal{A})$ (instead of the faces of $\operatorname{conv}(\mathcal{A})$) are considered. Before doing so we introduce the notions of *face*, *generating face*, and *feasible direction*.

Face of a convex set. Let us consider a set $\mathcal{K}$ with a $k-$dimensional affine hull along with a point $\mathbf{x} \in \mathcal{K}$. Then, $\mathcal{K}$ is a $k-$dimensional face of $\operatorname{conv}(\mathcal{A})$ if $\mathcal{K} = \operatorname{conv}(\mathcal{A}) \cap \{\mathbf{y} : \langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle = 0\}$ for some normal vector $\mathbf{r}$ and $\operatorname{conv}(A)$ is contained in the half-space determined by $\mathbf{r}$, i.e., $\langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \forall \mathbf{y} \in \operatorname{conv}(\mathcal{A})$. Intuitively, given a set $\operatorname{conv}(\mathcal{A})$ one can think of $\operatorname{conv}(\mathcal{A})$ being a $\dim(\operatorname{conv}(\mathcal{A}))-$dimensional face of itself, an edge on the border of the set a 1-dimensional face and a vertex a 0-dimensional face.

Face of a cone and g-faces. Similarly, a k–dimensional face of a cone is an open and unbounded set $\text{cone}(\mathcal{A}) \cap \{\mathbf{y} \colon \langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle = 0\}$ for some normal vector $\mathbf{r}$ and $\text{cone}(A)$ is contained in the half space determined by $\mathbf{r}$. We can define the generating faces of a cone as:

$$\text{g-faces}(\text{cone}(\mathcal{A})) := \{\mathcal{B} \cap \text{conv}(\mathcal{A}) \colon \mathcal{B} \in \text{faces}(\text{cone}(\mathcal{A}))\}.$$

Note that $\text{g-faces}(\text{cone}(\mathcal{A})) \subset \text{faces}(\text{conv}(\mathcal{A}))$ and $\text{conv}(\mathcal{A}) \in \text{g-faces}(\text{cone}(\mathcal{A}))$. Furthermore, for each $\mathcal{K} \in \text{g-faces}(\text{cone}(\mathcal{A}))$, $\text{cone}(\mathcal{K})$ is a k–dimensional face of $\text{cone}(\mathcal{A})$.

We now introduce the notion of **feasible directions**. A direction $\mathbf{d}$ is feasible from $\mathbf{x} \in \text{cone}(\mathcal{A})$ if it points inwards $\text{cone}(\mathcal{A})$, i.e., if $\exists \varepsilon > 0$ s.t. $\mathbf{x} + \varepsilon \mathbf{d} \in \text{cone}(\mathcal{A})$. Since a face of the cone is itself a cone, if a direction is feasible from $\mathbf{x} \in \text{cone}(\mathcal{K}) \setminus \mathbf{0}$, it is feasible from every positive rescaling of $\mathbf{x}$. We therefore can consider only the feasible directions on the generating faces (which are closed and bounded sets). Finally, we define the cone width of $\mathcal{A}$.

Cone Width.

$$\text{CWidth}(\mathcal{A}) := \min_{\substack{\mathcal{K} \in \text{g-faces}(\text{cone}(\mathcal{A})) \\ \mathbf{x} \in \mathcal{K} \\ \mathbf{r} \in \text{cone}(\mathcal{K} - \mathbf{x}) \setminus \{\mathbf{0}\}}} PdirW(\mathcal{K} \cap \mathcal{A}, \mathbf{r}, \mathbf{x}) \tag{9}$$

We are now ready to show the linear convergence of Algorithms 3 and 4.

Theorem 3. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set with $\mathbf{0} \in \mathcal{A}$ and let the objective function $f : \mathcal{H} \to \mathbb{R}$ be both L-smooth and μ-strongly convex over $\rho \, \text{conv}(\mathcal{A} \cup -\mathcal{A})$. Then, the suboptimality of the iterates of Algorithms 3 and 4 decreases geometrically at each step in which $\gamma < \alpha_{\mathbf{v}_t}$ (henceforth referred to as "good steps") as:*

$$\varepsilon_{t+1} \leq (1 - \beta) \, \varepsilon_t, \tag{10}$$

where $\beta := \delta^2 \frac{\mu \, \text{CWidth}(\mathcal{A})^2}{L \, \text{diam}(\mathcal{A})^2} \in (0, 1]$, $\varepsilon_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is the suboptimality at step t and $\delta \in (0, 1]$ is the relative accuracy parameter of the employed approximate LMO (3). For AMP (Algorithm 3), $\beta^{\text{AMP}} = \beta/2$. If $\mu = 0$ Algorithm 3 converges with rate $O(1/k(t))$ where $k(t)$ is the number of "good steps" up to iteration t.

Discussion. To obtain a linear convergence rate, one needs to upper-bound the number of "bad steps" $t - k(t)$ (i.e., steps with $\gamma \geq \alpha_{\mathbf{v}_t}$). We have that $k(t) = t$ for Variant 1 of FCMP (Algorithm 4), $k(t) \geq t/2$ for AMP (Algorithm 3) and $k(t) \geq t/(3|\mathcal{A}|! + 1)$ for PWMP (Algorithm 3) and Variant 0 of FCMP (Algorithm 4). This yields a global linear convergence rate of $\varepsilon_t \leq \varepsilon_0 \exp\left(-\beta k(t)\right)$. The bound for PWMP is very loose and only meaningful for finite sets $\mathcal{A}$. However, it can be observed in the experiments in the supplementary material (Appendix A) that only a very small fraction of iterations result in bad PWMP steps in practice. Further note that Variant 1 of FCMP (Algorithm 4) does not produce bad steps. Also note that the bounds on the number of good steps given above are the same as for the corresponding FW variants and are obtained using the same (purely combinatorial) arguments as in [17].

Relation to previous MP rates. The linear convergence of the generalized (not non-negative) MP variants studied in [20] crucially depends on the geometry of the set which is characterized by the Minimal Directional Width $\text{mDW}(\mathcal{A})$:

$$\text{mDW}(\mathcal{A}) := \min_{\substack{\mathbf{d} \in \text{lin}(\mathcal{A}) \\ \mathbf{d} \neq \mathbf{0}}} \max_{\mathbf{z} \in \mathcal{A}} \langle \frac{\mathbf{d}}{\|\mathbf{d}\|}, \mathbf{z} \rangle. \tag{11}$$

The following Lemma relates the Cone Width with the minimal directional width.

Lemma 4. *If the origin is in the relative interior of $\text{conv}(\mathcal{A})$ with respect to its linear span, then $\text{cone}(\mathcal{A}) = \text{lin}(\mathcal{A})$ and $\text{CWidth}(\mathcal{A}) = \text{mDW}(\mathcal{A})$.*

Now, if the set $\mathcal{A}$ is symmetric or, more generally, if $\text{cone}(\mathcal{A})$ spans the linear space $\text{lin}(\mathcal{A})$ (which implies that the origin is in the relative interior of $\text{conv}(\mathcal{A})$), there are no bad steps. Hence, by Lemma 4, the linear rate obtained in Theorem 3 for non-negative MP variants generalizes the one presented in [20, Theorem 7] for generalized MP variants.

Relation to FW rates.   Optimization over conic hulls with non-negative MP is more similar to FW than to MP itself in the following sense. For MP, every direction in $\mathrm{lin}(\mathcal{A})$ allows for unconstrained steps, from any iterate $\mathbf{x}_t$. In contrast, for our non-negative MPs, while some directions allow for unconstrained steps from some iterate $\mathbf{x}_t$, others are constrained, thereby leading to the dependence of the linear convergence rate on the cone width, a geometric constant which is very similar in spirit to the Pyramidal Width appearing in the linear convergence bound in [17] for FW. Furthermore, as for Algorithm 3, the linear rate of Away-steps and Pairwise FW holds only for good steps. We finally relate the cone width with the Pyramidal Width [17]. The Pyramidal Width is defined as

$$\mathrm{PWidth}(\mathcal{A}) := \min_{\substack{\mathcal{K} \in \mathrm{faces}(\mathrm{conv}(\mathcal{A})) \\ \mathbf{x} \in \mathcal{K} \\ \mathbf{r} \in \mathrm{cone}(\mathcal{K} - \mathbf{x}) \backslash \{\mathbf{0}\}}} PdirW(\mathcal{K} \cap \mathcal{A}, \mathbf{r}, \mathbf{x}).$$

We have $\mathrm{CWidth}(\mathcal{A}) \geq \mathrm{PWidth}(\mathcal{A})$ as the minimization in the definition (9) of $\mathrm{CWidth}(\mathcal{A})$ is only over the subset g-faces$(\mathrm{cone}(\mathcal{A}))$ of faces$(\mathrm{conv}(\mathcal{A}))$. As a consequence, the decrease per iteration characterized in Theorem 3 is larger than what one could obtain with FW on the rescaled convex set $\tau\mathcal{A}$ (see Section 4.1 for details about the rescaling). Furthermore, the decrease characterized in [17] scales as $1/\tau^2$ due to the dependence on $1/\mathrm{diam}(\mathrm{conv}(\mathcal{A}))^2$.

5 Related Work

The line of recent works by [30, 32, 33, 34, 24, 20] targets the generalization of MP from the least-squares objective to general smooth objectives and derives corresponding convergence rates (see [20] for a more in-depth discussion). However, only little prior work targets MP variants with non-negativity constraint [3, 25, 37]. In particular, the least-squares objective was addressed and no rigorous convergence analysis was carried out. [3, 37] proposed an algorithm equivalent to our Algorithm 4 for the least-squares case. More specifically, [37] then developed an acceleration heuristic, whereas [3] derived a coherence-based recovery guarantee for sparse linear combinations of atoms. Apart from MP-type algorithms, there is a large variety of non-negative least-squares algorithms, e.g., [19], in particular also for matrix and tensor spaces. The gold standard in factorization problems is projected gradient descent with alternating minimization, see [29, 2, 31, 15]. Other related works are [27], which is concerned with the feasibility problem on symmetric cones, and [12], which introduces a norm-regularized variant of problem (4) and solves it using FW on a rescaled convex set. To the best of our knowledge, in the context of MP-type algorithms, we are the first to combine general convex objectives with conic constraints and to derive corresponding convergence guarantees.

Boosting: In an earlier line of work, a flavor of the generalized MP became popular in the context of boosting, see [23]. The literature on boosting is vast, we refer to [28, 23, 5] for a general overview. Taking the optimization perspective given in [28], boosting is an iterative greedy algorithm minimizing a (strongly) convex objective over the linear span of a possibly infinite set called hypothesis class. The convergence analysis crucially relies on the assumption of the origin being in the relative interior of the hypothesis class, see Theorem 1 in [10]. Indeed, Algorithm 5.2 of [23] might not converge if the [26] alignment assumption is violated. Here, we managed to relax this assumption while preserving essentially the same asymptotic rates in [23, 10]. Our work is therefore also relevant in the context of (non-negative) boosting.

6 Illustrative Experiments

We illustrate the performance of the presented algorithms on three different exemplary tasks, showing that our algorithms are competitive with established baselines across a wide range of objective functions, domains, and data sets while not being specifically tailored to any of these tasks (see Section 3.2 for a discussion of the computational complexity of the algorithms). Additional experiments targeting KL divergence NMF, non-negative tensor factorization, and hyperspectral image unmixing can be found in the appendix.

Synthetic data.   We consider minimizing the least squares objective on the conic hull of 100 unit-norm vectors sampled at random in the first orthant of $\mathbb{R}^{50}$. We compare the convergence of Algorithms 2, 3, and 4 with the Fast Non-Negative MP (FNNOMP) of [37], and Variant 3 (line-search) of the FW algorithm in [20] on the atom set rescaled by $\tau = 10\|\mathbf{y}\|$ (see Section 4.1), observing linear convergence for our corrective variants.
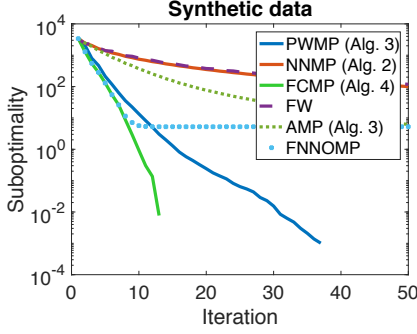
Figure 2: Synthetic data experiment.

Figure 2 shows the suboptimality ε_t, averaged over 20 realizations of $\mathcal{A}$ and $\mathbf{y}$, as a function of the iteration t. As expected, FCMP achieves fastest convergence followed by PWMP, AMP and NNMP. The FNNOMP gets stuck instead. Indeed, [37] only show that the algorithm terminates and not its convergence.

Non-negative matrix factorization. The second task consists of decomposing a given matrix into the product of two non-negative matrices as in Equation (1) of [13]. We consider the intersection of the positive semidefinite cone and the positive orthant. We parametrize the set $\mathcal{A}$ as the set of matrices obtained as an outer product of vectors from $\mathcal{A}_1 = \{\mathbf{z} \in \mathbb{R}^k : \mathbf{z}_i \geq 0 \, \forall \, i\}$ and $\mathcal{A}_2 = \{\mathbf{z} \in \mathbb{R}^d : \mathbf{z}_i \geq 0 \, \forall \, i\}$. The LMO is approximated using a truncated power method [40], and we perform atom correction with greedy coordinate descent see, e.g., [18, 11], to obtain a better objective value while maintaining the same (small) number of atoms. We consider three different datasets: The Reuters Corpus[2], the CBCL face dataset[3] and the KNIX dataset[4]. The subsample of the Reuters corpus we used is a term frequency matrix of 7,769 documents and 26,001 words. The CBCL face dataset is composed of 2,492 images of 361 pixels each, arranged into a matrix. The KNIX dataset contains 24 MRI slices of a knee, arranged in a matrix of size $262,144 \times 24$. Pixels are divided by their overall mean intensity. For interpretability reasons, there is interest to decompose MRI data into non-negative factorizations [57]. We compare PWMP and FCMP against the multiplicative (mult) and the alternating (als) algorithm of [2], and the greedy coordinate descent (GCD) of [13]. Since the Reuters corpus is much larger than the CBCL and the KNIX dataset we only used the GCD for which a fast implementation in C is available. We report the objective value for fixed values of the rank in Table 2, showing that FCMP outperform all the baselines across all the datasets. PWMP achieves smallest error on the Reuters corpus.

Non-negative garrote. We consider the non-negative garrote which is a common approach to model order selection [4]. We evaluate NNMP, PWMP, and FCMP in the experiment described in [21], where the non-negative garrote is used to perform model order selection for logistic regression (i.e., for a non-quadratic objective function). We evaluated training and test accuracy on 100 random splits of the sonar dataset from the UCI machine learning repository. In Table 3 we compare the median classification accuracy of our algorithms with that of the cyclic coordinate descent algorithm (NNG) from [21].

algorithm	Reuters $K = 10$	CBCL $K = 10$	CBCL $K = 50$	KNIX $K = 10$
mult	-	2.4241e3	1.1405e3	2.4471e03
als	-	2.73e3	3.84e3	2.7292e03
GCD	5.9799e5	2.2372e3	806	2.2372e03
PWMP	**5.9591e5**	2.2494e3	789.901	2.2494e03
FCMP	5.9762e5	**2.2364e3**	**786.15**	**2.2364e03**

Table 2: Objective value for least-squares non-negative matrix factorization with rank K.

	training accuracy	test accuracy
NNMP	0.8345 ± 0.0242	$\mathbf{0.7419} \pm 0.0389$
PWMP	$\mathbf{0.8379} \pm 0.0240$	$\mathbf{0.7419} \pm 0.0392$
FCMP	0.8345 ± 0.0238	$\mathbf{0.7419} \pm 0.0403$
NNG	0.8069 ± 0.0518	0.7258 ± 0.0602

Table 3: Logistic Regression with non-negative Garrote, median $\pm$ std. dev.

7 Conclusion

In this paper, we considered greedy algorithms for optimization over the convex cone, parametrized as the *conic hull* of a generic atom set. We presented a novel formulation of NNMP along with a comprehensive convergence analysis. Furthermore, we introduced corrective variants with linear convergence guarantees, and verified this convergence rate in numerical applications. We believe that the generality of our novel analysis will be useful to design new, fast algorithms with convergence guarantees, and to study convergence of existing heuristics, in particular in the context of non-negative matrix and tensor factorization.

---

[2] http://www.nltk.org/book/ch02.html
[3] http://cbcl.mit.edu/software-datasets/FaceData2.html
[4] http://www.osirix-viewer.com/resources/dicom-image-library/

References

[1] Jonas Behr, André Kahles, Yi Zhong, Vipin T Sreedharan, Philipp Drewe, and Gunnar Rätsch. Mitie: Simultaneous rna-seq-based transcript identification and quantification in multiple samples. *Bioinformatics*, 29(20):2529–2538, 2013.

[2] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

[3] Alfred M Bruckstein, Michael Elad, and Michael Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54(11):4813–4820, 2008.

[4] P Bühlmann and B Yu. Boosting, model selection, lasso and nonnegative garrote. Technical Report 127, Seminar für Statistik ETH Zürich, 2005.

[5] Peter Bühlmann and Bin Yu. Boosting. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):69–74, 2010.

[6] Sheng Chen, Stephen A Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of control*, 50(5):1873–1896, 1989.

[7] Ernie Esser, Yifei Lou, and Jack Xin. A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM Journal on Imaging Sciences*, 6(4):2010–2046, 2013.

[8] M Frank and P Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 1956.

[9] Nicolas Gillis and Robert Luce. A fast gradient method for nonnegative sparse regression with self dictionary. *arXiv preprint arXiv:1610.01349*, 2016.

[10] Alexander Grubb and J Andrew Bagnell. Generalized boosting algorithms for convex optimization. *arXiv preprint arXiv:1105.2054*, 2011.

[11] Xiawei Guo, Quanming Yao, and James T Kwok. Efficient sparse low-rank tensor completion using the Frank-Wolfe algorithm. In *AAAI Conference on Artificial Intelligence*, 2017.

[12] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

[13] Cho-Jui Hsieh and Inderjit S Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1064–1072. ACM, 2011.

[14] Martin Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.

[15] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[16] Jingu Kim and Haesun Park. Fast nonnegative tensor factorization with an active-set-like method. In *High-Performance Scientific Computing*, pages 311–326. Springer, 2012.

[17] Simon Lacoste-Julien and Martin Jaggi. On the Global Linear Convergence of Frank-Wolfe Optimization Variants. In *NIPS 2015*, pages 496–504, 2015.

[18] Sören Laue. A Hybrid Algorithm for Convex Semidefinite Optimization. In *ICML*, 2012.

[19] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 15. SIAM, 1995.

[20] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and frank-wolfe. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[21] Enes Makalic and Daniel F Schmidt. Logistic regression with the nonnegative garrote. In *Australasian Joint Conference on Artificial Intelligence*, pages 82–91. Springer, 2011.

[22] Stéphane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[23] Ron Meir and Gunnar Rätsch. An introduction to boosting and leveraging. In *Advanced lectures on machine learning*, pages 118–183. Springer, 2003.

[24] Hao Nguyen and Guergana Petrova. Greedy strategies for convex optimization. *Calcolo*, pages 1–18, 2014.

[25] Robert Peharz, Michael Stark, and Franz Pernkopf. Sparse nonnegative matrix factorization using l0-constraints. In IEEE, editor, *Proceedings of MLSP*, pages 83 – 88, Aug 2010.

[26] Javier Pena and Daniel Rodriguez. Polytope conditioning and linear convergence of the frank-wolfe algorithm. *arXiv preprint arXiv:1512.06142*, 2015.

[27] Javier Pena and Negar Soheili. Solving conic systems via projection and rescaling. *Mathematical Programming*, pages 1–25, 2016.

[28] Gunnar Rätsch, Sebastian Mika, Manfred K Warmuth, et al. On the convergence of leveraging. In *NIPS*, pages 487–494, 2001.

[29] F Sha, LK Saul, and Daniel D Lee. Multiplicative updates for nonnegative quadratic programming in support vector machines. *Advances in Neural Information Processing Systems*, 15, 2002.

[30] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading Accuracy for Sparsity in Optimization Problems with Sparsity Constraints. *SIAM Journal on Optimization*, 20:2807–2832, 2010.

[31] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799. ACM, 2005.

[32] Vladimir Temlyakov. Chebushev Greedy Algorithm in convex optimization. *arXiv.org*, December 2013.

[33] Vladimir Temlyakov. Greedy algorithms in convex optimization on Banach spaces. In *48th Asilomar Conference on Signals, Systems and Computers*, pages 1331–1335. IEEE, 2014.

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

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

[36] Zheng Wang, Ming jun Lai, Zhaosong Lu, Wei Fan, Hasan Davulcu, and Jieping Ye. Rank-one matrix pursuit for matrix completion. In *ICML*, pages 91–99, 2014.

[37] Mehrdad Yaghoobi, Di Wu, and Mike E Davies. Fast non-negative orthogonal matching pursuit. *IEEE Signal Processing Letters*, 22(9):1229–1233, 2015.

[38] Yuning Yang, Siamak Mehrkanoon, and Johan A K Suykens. Higher order Matching Pursuit for Low Rank Tensor Learning. *arXiv.org*, March 2015.

[39] Quanming Yao and James T Kwok. Greedy learning of generalized low-rank models. In *IJCAI*, 2016.

[40] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, 14(1):899–925, April 2013.
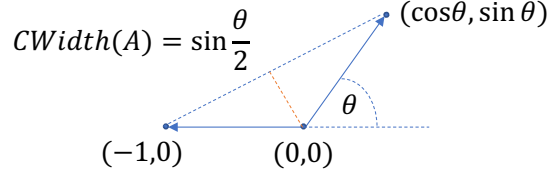
$$CWidth(A) = \sin\frac{\theta}{2}$$

$(\cos\theta, \sin\theta)$

θ

$(-1,0)$ $(0,0)$

Figure 3: $\mathrm{CWidth}(\mathcal{A})$ for the set $\mathcal{A} := \{\mathcal{A}_\theta \cup -\mathcal{A}_\theta\}$ where $\mathcal{A}_\theta := \left\{ \binom{0}{0}, \binom{-1}{0}, \binom{\cos\theta}{\sin\theta} \right\}$ with $\theta \in (0, \pi/2]$
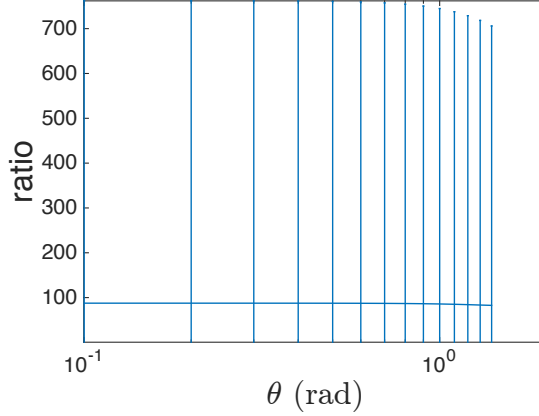
Figure 4: Ratio of theoretical and empirical rate for $\mathcal{A} := \{\mathcal{A}_\theta \cup -\mathcal{A}_\theta\}$ where $\mathcal{A}_\theta := \left\{ \binom{0}{0}, \binom{-1}{0}, \binom{\cos\theta}{\sin\theta} \right\}$ with $\theta \in (0, \pi/2)$ and 20 random target points $\binom{-\alpha_1}{\alpha_2}$ with $\alpha_i > 0$.

A   Additional experiments

A.1   An illustrative experiment: Tightness of Theorem 3

We now consider the setting depicted in Figure 3. We consider the set $\mathcal{A} := \{\mathcal{A}_\theta \cup -\mathcal{A}_\theta\}$ where $\mathcal{A}_\theta := \left\{ \binom{0}{0}, \binom{-1}{0}, \binom{\cos\theta}{\sin\theta} \right\}$ with $\theta \in (0, \pi/2)$. For this set $\mathrm{CWidth}(\mathcal{A})$ can be computed in closed form as $\mathrm{CWidth}(\mathcal{A}) = \sin(\theta/2)$. We then perform 20 runs of Algorithm 3 and report the ratio between the theoretical rate and the empirical one. The result is depicted in Figure 4. There, we considered an iteration starting from the origin minimizing the distance function to 20 random points $\binom{-\alpha_1}{\alpha_2}$ with $\alpha_i > 0$. The vertical bars shows minimal and maximal values.

A.2   Real Data

Hyperspectral image unmixing. One of the classical applications of non-negative least squares are unmixing problems [7] such as hyperspectral image unmixing. Scalable unmixing approaches such as SPA [42] first extract a self-dictionary from a target image. Each pixel is then projected on
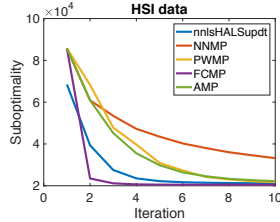

Figure 5: Hyperspectral Imaging. We report suboptimality in a non-negative least squares task on real data, respectively.
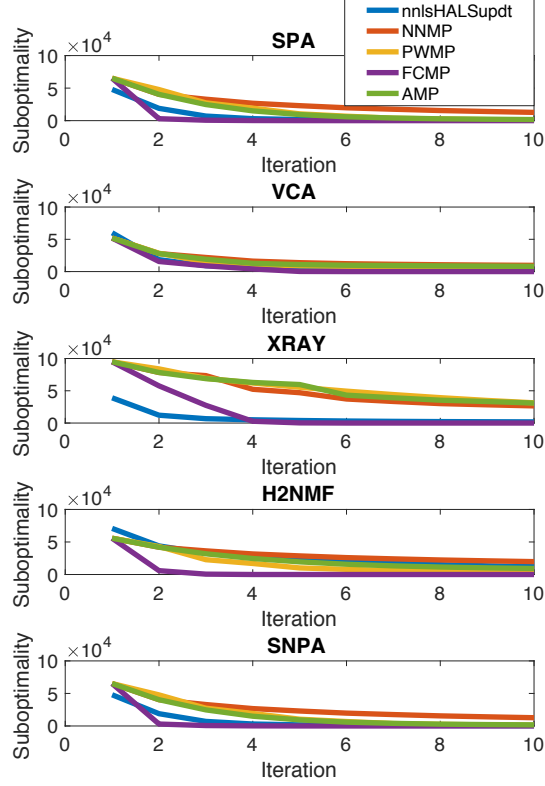
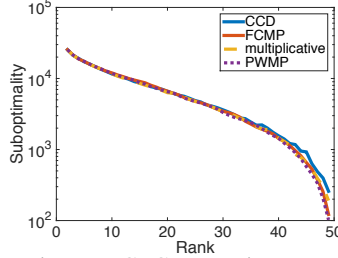Figure 6: SPA [42], VCA [65], XRAY [58], H2NMF [50], SNPA [48]

Figure 7: CBCL KL-Divergence

the conic hull of the dictionary to estimate the abundance of each material. A standard technique is the hierarchical alternating least squares of [49] (nnlsHALSupdt). In Figure 5 (right), we compare the suboptimality of different methods as a function of the iteration. The dictionary is extracted from the undersampled Urban HSI Dataset[5] using SPA. This dataset contains 5,929 pixels, each associated with 162 hyperspectral features. The number of dictionary elements is 6, motivated by the fact that 6 different physical materials are depicted in this HSI data [9]. Therefore, FCMP converges after 6 iterations. For PWMP only 1.5% of the iterations were bad steps on average for all dictionaries. Therefore, our corrective methods are proven to be competitive also on real data and the effect of the bad steps is negligible. We test other dictionaries for the Hyperspectral Imaging task. The result is depicted in Figure 6.

KL-divergence non-negative low-rank matrix factorization. The third task targets non-negative matrix factorization by minimization of the (non-least squares) KL-divergence-based objective function in Equation (3) in [13]. We again use the CBCL face dataset and we compare FCMP (Variant 1) and PWMP against the multiplicative algorithm from [62] (multiplicative) and the cyclic coordinate descent (CCD) from [13]. We use the same approximate LMO and parametrization of $\mathcal{A}$ as for the

---

[5]download at `http://bit.ly/fgnsr`

algorithm	relative error
multiplicative	0.2991
hals	0.2927
anls-as	0.2912
anls-bpp	0.2914
PWMP	0.2913
FCMP	**0.2909**

Table 4: Non-negative tensor factorization on the KNIX dataset with rank 20

least-squares non-negative matrix factorization, and we set the L to 0.1. The atom correction was implemented using the CCD algorithm. In this experiment the use of Variant 0 of FCMP is crucial as it allows for a much easier update step. The objective value as a function of the rank is depicted in Figure 7. We note that all the algorithms yield comparable objective value up to rank 35. For higher rank, FCMP and PWMP achieve a slightly smaller objective value.

Least Squares Non-negative Tensor Factorization. For this task we again use the KNIX dataset but now we arrange the scans to form a tensor of dimensionality $512 \times 512 \times 24$. We compare against the alternating nonnegativity-constrained least squares with block principal pivoting [16] (anlss-bpp) (which is also used in the FCMP and PWMP for the corrections), the active set method in [55] (anls-as), the hierarchical alternating least squares of [46] (hals) and the multiplicative updating algorithm (multiplicative) of [66]. The LMO for the tensor factorization is implemented with the tensor power method [41]. The result is depicted in Table 4.

In this section we showed that the merit of our algorithms is not limited to their theoretical properties. Indeed, our algorithms are competitive with several approaches and can be successfully used in a manifold of different tasks and datasets while not being tailored to any specific cost function.

B   Affine Invariant Algorithms and Rates

In this section, present affine invariant versions of all presented algorithms, along with sub-linear and linear convergence guarantees. An optimization method is called *affine invariant* if it is invariant under linear or affine transformations of the input problem: If one chooses any re-parameterization of the domain $\mathcal{Q}$ by a *surjective* linear or affine map $\mathbf{M} : \hat{\mathcal{Q}} \to \mathcal{Q}$, then the "old" and "new" optimization problems $\min_{\mathbf{x} \in \mathcal{Q}} f(\mathbf{x})$ and $\min_{\hat{\mathbf{x}} \in \hat{\mathcal{Q}}} \hat{f}(\hat{\mathbf{x}})$ for $\hat{f}(\hat{\mathbf{x}}) := f(\mathbf{M}\hat{\mathbf{x}})$ look the same to the algorithm. We still require the set $\mathcal{Q}$ to contain the origin. In the following, we assume that after the transformation the origin is still on the border of $\text{conv}(\mathcal{Q})$. If the origin is contained in the relative interior of $\text{conv}(\mathcal{Q})$ we recover the existing affine invariant rates of [20].

B.1   Affine invariant non-negative MP

To define an affine invariant upper bound on the objective function f, we use a variation of the affine invariant definition of the *curvature constant* from [14], adapted for MP in [20]:

$$C_{f,\mathcal{A}}^{\mathsf{MP}} := \sup_{\substack{\mathbf{s} \in \mathcal{A},\, \mathbf{x} \in \text{conv}(\mathcal{A}) \\ \gamma \in [0,1] \\ \mathbf{y} = \mathbf{x} + \gamma \mathbf{s}}} \frac{2}{\gamma^2} D(\mathbf{y}, \mathbf{x}), \tag{12}$$

where for cleaner exposition, we have used the shorthand notation $D(\mathbf{y}, \mathbf{x})$ to denote the difference of $f(\mathbf{y})$ and its linear approximation at $\mathbf{x}$

$$D(\mathbf{y}, \mathbf{x}) := f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{y} - \mathbf{x}, \nabla f(\mathbf{x}) \rangle.$$

Bounded curvature $C_{f,\mathcal{A}}$ closely corresponds to smoothness of the objective f. More precisely, if ∇f is L-Lipschitz continuous on $\text{conv}(\mathcal{A})$ with respect to some arbitrary chosen norm $\|.\|$, i.e. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$, where $\|.\|_*$ is the dual norm of $\|.\|$, then

$$C_{f,\mathcal{A}} \leq L \, \text{radius}_{\|.\|}(\mathcal{A})^2, \tag{13}$$

where $\text{radius}_{\|.\|}(.)$ denotes the $\|.\|$-radius, see Lemma 15 in [20]. The curvature constant $C_{f,\mathcal{A}}$ is affine invariant as it does not depend on any norm. It combines the complexity of the domain $\text{conv}(\mathcal{A})$

and the curvature of the objective function f into a single quantity. Throughout this section, we assume availability of a finite constant $\rho > 0$ upper-bounding the atomic norms $\|.\|_{\mathcal{A}}$ of the optimum $\mathbf{x}^\star$, as well as the iterate sequence $(\mathbf{x}_t)_{t=0}^T$ until the current iteration, as defined in (6). We now present the affine invariant version of the non-negative MP algorithm (Algorithm 2) in Algorithm 5. The algorithm uses the curvature constant $C_{f,\rho(\mathcal{A}\cup-\mathcal{A})}^{\mathsf{MP}}$ over the re-scaled set $\rho \operatorname{conv}(\mathcal{A}\cup-\mathcal{A})$, rather than $\operatorname{conv}(\mathcal{A}\cup-\mathcal{A})$.

Algorithm 5 Affine Invariant Non-Negative Matching Pursuit

Same as Algorithm 2 except:

5:      $\gamma := \frac{\langle -\nabla f(\mathbf{x}_t), \rho^2 \mathbf{z}_t \rangle}{C_{f,\rho(\mathcal{A}\cup-\mathcal{A})}^{\mathsf{MP}}}$

A sub-linear convergence guarantee for Algorithm 5 is presented in the following theorem.

Theorem 5. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set with $\mathbf{0} \in \mathcal{A}$, $\rho := \max\{\|\mathbf{x}^\star\|_{\mathcal{A}}, \|\mathbf{x}_0\|_{\mathcal{A}}, \ldots, \|\mathbf{x}_T\|_{\mathcal{A}}\} < \infty$. Assume f has smoothness constant $C_{f,\rho(\mathcal{A}\cup-\mathcal{A})}^{\mathsf{MP}}$. Then, Algorithm 5 converges for $t \geq 0$ as*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{4\left(\frac{2}{\delta} C_{f,\rho(\mathcal{A}\cup-\mathcal{A})}^{\mathsf{MP}} + \varepsilon_0\right)}{\delta t + 4},$$

where $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO (3).

Exact knowledge of $C_{f,\rho(\mathcal{A}\cup-\mathcal{A})}^{\mathsf{MP}}$ is not required; the same theorem also holds if any upper bound on $C_{f,\rho(\mathcal{A}\cup-\mathcal{A})}^{\mathsf{MP}}$ is used in the algorithm and resulting rate instead.

B.2   Affine invariant corrective MP

An affine invariant version of AMP and PWMP, Algorithm 3, is presented in Algorithm 6. Note that Variant 1 of the fully corrective non-negative MP in Algorithm 4 is already affine invariant as it does not rely on any norm. Note that sublinear convergence is guaranteed with the rate indicated by Theorem 5 since each step of the affine invariant FCMP yields at least as much improvement as the affine invariant NNMP, Algorithm 5.

Since the update step in Algorithm 5 and the resulting upper bound on the progress in objective, based on the curvature constant (13), we used in the proof of Theorem 5 are not enough to ensure linear convergence, we use a different notion of curvature based on [59].

$$C_{f,\mathcal{A}}^{\mathsf{A}} = \sup_{\substack{\mathbf{s}\in\mathcal{A},\mathbf{x}\in\operatorname{conv}(\mathcal{A}) \\ \mathbf{v}\in\mathcal{S} \\ \gamma\in[0,1] \\ \mathbf{y}=\mathbf{x}+\gamma(\mathbf{s}-\mathbf{v})}} \frac{2}{\gamma^2} D(\mathbf{y},\mathbf{x}).$$

The following positive step size quantity relates the dual certificate value of the descent direction

---

Algorithm 6 Affine invariant AMP and PWMP

---

same as Algorithm 3 except for:

5:      $\gamma := \frac{\langle -\nabla f(\mathbf{x}_t), \rho^2 \mathbf{d}_t \rangle}{C_{f,\rho(\mathcal{A}\cup-\mathcal{A})}^{\mathsf{A}}}$

$\mathbf{x}^\star - \mathbf{x}$ with the MP selected atom,

$$\gamma(\mathbf{x}, \mathbf{x}^\star) := \frac{\langle -\nabla f(\mathbf{x}), \mathbf{x}^\star - \mathbf{x} \rangle}{\langle -\nabla f(\mathbf{x}), \mathbf{s}(\mathbf{x}) - \mathbf{v}(\mathbf{x}) \rangle}, \tag{14}$$

for $\mathbf{s}(\mathbf{x}) := \arg\min_{\mathbf{s}\in\mathcal{A}} \langle \nabla f(\mathbf{x}), \mathbf{s} \rangle$ and $\mathbf{v}(\mathbf{x}) := \min_{\mathcal{S}\in\mathcal{S}_{\mathbf{x}}} \arg\max_{\mathbf{s}\in\mathcal{S}} \langle \nabla f(\mathbf{x}), \mathbf{s} \rangle$ where $\mathcal{S} \in \mathcal{S}_{\mathbf{x}}$ is the active set. We now define the affine invariant surrogate of strong convexity.

$$\mu_{f,\rho\mathcal{A}}^A := \inf_{\mathbf{x}\in\operatorname{conv}(\rho\mathcal{A})} \inf_{\substack{\mathbf{x}^\star\in\operatorname{conv}(\rho\mathcal{A}) \\ \langle\nabla f(\mathbf{x}), \mathbf{x}^\star - \mathbf{x}\rangle < 0}} \frac{2}{\gamma(\mathbf{x},\mathbf{x}^\star)} D(\mathbf{x}^\star, \mathbf{x}).$$

Theorem 6. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set containing the origin and let the objective function $f: \mathcal{H} \to \mathbb{R}$ have smoothness constant $C^{\mathrm{A}}_{f,\rho(\mathcal{A}-\mathcal{A})}$ and strong convexity constant $\mu^{A}_{f,\rho\mathcal{A}}$*

Then, the suboptimality of the iterates of Algorithm 3 and 4 decreases geometrically at each step in which $\gamma < \alpha_{\mathbf{v}_t}$ (henceforth referred to as "good steps") as:

$$\varepsilon_{t+1} \leq (1 - \beta)\,\varepsilon_t, \tag{15}$$

where $\beta := \delta^2 \frac{\mu^{A}_{f,\rho\mathcal{A}}}{C^{\mathrm{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}} \in (0,1]$, $\varepsilon_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is the suboptimality at step t and $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO (Equation (3)). For AMP (Algorithm 3), $\beta^{\mathrm{AMP}} = \beta/2$. If $\mu^{A}_{f,\rho\mathcal{A}} = 0$ Algorithm 3 converges with rate $O(1/k(t))$ where $k(t)$ is the number of "good steps" up to iteration t.

C   Proof of Lemma 1

If $\tilde{\mathbf{x}} \in \mathrm{cone}(\mathcal{A})$ and $\nabla f(\tilde{\mathbf{x}}) \notin T_{\mathcal{A}}$ then $\tilde{\mathbf{x}}$ is a solution to $\min_{\mathbf{x} \in \mathrm{cone}(\mathcal{A})} f(\mathbf{x})$.

Proof. We will prove this lemma by contradiction assuming that $\mathbf{x}^\star \neq \tilde{\mathbf{x}}$ and $\nabla f(\tilde{\mathbf{x}}) \notin T_{\mathcal{A}}$. Now, by convexity of f we have that:

$$f(\mathbf{x}^\star) \geq f(\tilde{\mathbf{x}}) + \langle \nabla f(\tilde{\mathbf{x}}), \mathbf{x}^\star - \tilde{\mathbf{x}} \rangle$$

Since $\mathbf{x}^\star \neq \tilde{\mathbf{x}}$ we have also that $f(\mathbf{x}^\star) < f(\tilde{\mathbf{x}})$. Therefore:

$$0 < f(\tilde{\mathbf{x}}) - f(\mathbf{x}^\star) \leq \langle -\nabla f(\tilde{\mathbf{x}}), \mathbf{x}^\star - \tilde{\mathbf{x}} \rangle$$

which we rewrite as $\langle \nabla f(\tilde{\mathbf{x}}), \mathbf{x}^\star \rangle + \langle \nabla f(\tilde{\mathbf{x}}), -\tilde{\mathbf{x}} \rangle < 0$. Now we note that by the assumption that $\nabla f(\tilde{\mathbf{x}}) \notin T_{\mathcal{A}}$ we have that both these inner products are non negative which is absurd. To draw this conclusion note that $\mathbf{x}^\star \in \mathrm{cone}(\mathcal{A})$ we have that $\mathbf{x}^\star = \sum_i \alpha_i \mathbf{z}_i$ where $\mathbf{z}_i \in \mathcal{A}$ and $\alpha_i \geq 0 \,\forall\, i$. $\square$

D   Sublinear Rates

Theorem' 2. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set with $\mathbf{0} \in \mathcal{A}$, $\rho := \max\{\|\mathbf{x}^\star\|_{\mathcal{A}}, \|\mathbf{x}_0\|_{\mathcal{A}}, \ldots, \|\mathbf{x}_T\|_{\mathcal{A}},\}$ and f be L-smooth over $\rho\,\mathrm{conv}(\mathcal{A} \cup -\mathcal{A})$. Then, Algorithms 2 and 4 with $\mathbf{x}_0 = \mathbf{0}$ converges for $t \geq 0$ as*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{4\left(\frac{2}{\delta}L\rho^2\,\mathrm{radius}(\mathcal{A})^2 + \varepsilon_0\right)}{\delta t + 4},$$

where $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO (3).

Proof. We separately prove the convergence for the two algorithms.

non-negative MP: Recall that $\tilde{\mathbf{z}}_t$ is the atom returned by the inexact LMO after the comparison with $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}$ at the current iteration t. We distinguish the two cases in which $\tilde{\mathbf{z}}_t \neq -\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}$ (**case A**) and $\tilde{\mathbf{z}}_t = -\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}$ (**case B**). Let us call $\bar{\mathcal{A}} := \mathcal{A} \cup \left\{-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}\right\}$. Note that $\mathrm{radius}(\bar{\mathcal{A}}) = \mathrm{radius}(\mathcal{A})$.

Recall that in the Algorithm the step size γ is computed at each iteration via line search minimizing the quadratic upper bound on f and no further clipping is made. The reason being that f is convex, therefore, for $t > 0$ we have $f(\mathbf{x}_t) \leq f(\mathbf{0})$. Hence the minimum of f over the line between $\mathbf{x}_t$ and the origin must lie between these two points making clipping unnecessary.

We start by upper bounding f on $\rho \operatorname{conv}(\bar{\mathcal{A}})$ using smoothness as follows:

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) \; \leq \; & \min_{\gamma \in \mathbb{R}_{\geq 0}} g_{\mathbf{x}_t}(\mathbf{x}_t + \gamma \tilde{\mathbf{z}}_t) \\
= \; & \min_{\gamma \in [0,1]} g_{\mathbf{x}_t}(\mathbf{x}_t + \gamma \rho \tilde{\mathbf{z}}_t) \\
\leq \; & \min_{\gamma \in [0,1]} f(\mathbf{x}_t) + \gamma \langle \nabla f(\mathbf{x}_t), \rho \tilde{\mathbf{z}}_t \rangle \\
& + \frac{L}{2} \gamma^2 \rho^2 \|\tilde{\mathbf{z}}_t\|^2 \\
\leq \; & \min_{\gamma \in [0,1]} f(\mathbf{x}_t) + \gamma \langle \nabla f(\mathbf{x}_t), \rho \tilde{\mathbf{z}}_t \rangle \\
& + \frac{L}{2} \gamma^2 \rho^2 \operatorname{radius}(\mathcal{A})^2
\end{aligned}
$$

(16)

We now treat separately the linear term for **case A** and **case B**.

case A: We start from the definition of inexact LMO (Equation (3)). We then have:

$$
\langle \nabla f(\mathbf{x}_t), \tilde{\mathbf{z}}_t \rangle \leq \delta \langle \nabla f(\mathbf{x}_t), \mathbf{z}_t \rangle
$$

where $\mathbf{z}_t$ is the true minimizer of the linear problem (LMO). In other words, it holds that $\langle \nabla f(\mathbf{x}_t), \mathbf{z}_t \rangle \leq \langle \nabla f(\mathbf{x}_t), \mathbf{z} \rangle \ \forall \ \mathbf{z} \in \operatorname{conv}(\bar{\mathcal{A}})$ due to the $\arg\min$ in line 4 of Algorithm 2. Therefore, since $\mathbf{x}^\star \in \rho \operatorname{conv}(\mathcal{A})$ it holds that:

$$
\langle \nabla f(\mathbf{x}_t), -\rho \mathbf{z}_t \rangle \geq \langle \nabla f(\mathbf{x}_t), -\mathbf{x}^\star \rangle.
$$

Using the same argument, since $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} \in \bar{\mathcal{A}}$ and $\rho \geq \|\mathbf{x}_t\|_{\mathcal{A}}$, we have that $-\mathbf{x}_t \in \rho \operatorname{conv}(\bar{\mathcal{A}})$. Therefore:

$$
\langle \nabla f(\mathbf{x}_t), -\rho \mathbf{z}_t \rangle \geq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t \rangle.
$$

We can now bound the linear term of in (16) as:

$$
\begin{aligned}
\langle \nabla f(\mathbf{x}_t), -2\frac{\rho}{\delta}\tilde{\mathbf{z}}_t \rangle &= \langle \nabla f(\mathbf{x}_t), -\frac{\rho}{\delta}\tilde{\mathbf{z}}_t \rangle + \langle \nabla f(\mathbf{x}_t), -\frac{\rho}{\delta}\tilde{\mathbf{z}}_t \rangle \\
&\geq \langle \nabla f(\mathbf{x}_t), -\rho \mathbf{z}_t \rangle + \langle \nabla f(\mathbf{x}_t), -\rho \mathbf{z}_t \rangle \\
&\geq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle \\
&\geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) =: \varepsilon_t
\end{aligned}
$$

where in the inequalities we used the the inexact oracle definition (see Section 2), the fact that both $-\mathbf{x}_t$ and $\mathbf{x}^\star \in \rho \operatorname{conv}(\bar{\mathcal{A}})$ and convexity respectively.

case B: Using line 4 of Algorithm 2 along with the inexact oracle definition we obtain:

$$
\langle \nabla f(\mathbf{x}_t), -\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} \rangle \leq \delta \min_{\mathbf{z} \in \mathcal{A}} \langle \nabla f(\mathbf{x}_t), \mathbf{z} \rangle.
$$

Therefore, since $\mathbf{x}^\star \in \rho \operatorname{conv}(\mathcal{A})$ we can write:

$$
\begin{aligned}
\langle \nabla f(\mathbf{x}_t), -\frac{\rho}{\delta}\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} \rangle &\leq \min_{\mathbf{z} \in \mathcal{A}} \langle \nabla f(\mathbf{x}_t), \rho \mathbf{z} \rangle \\
&\leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}^\star \rangle
\end{aligned}
$$

We also have $\langle \nabla f(\mathbf{x}_t), -\mathbf{x}_t \rangle \leq 0$ and $\frac{\rho}{\delta \|\mathbf{x}_t\|_{\mathcal{A}}} > 1$, which yields:

$$
\langle \nabla f(\mathbf{x}_t), -\mathbf{x}_t \rangle \geq \langle \nabla f(\mathbf{x}_t), -\frac{\rho}{\delta}\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} \rangle
$$

Putting these inequalites together we obtain:

$$
\begin{aligned}
\langle \nabla f(\mathbf{x}_t), \frac{2}{\delta}\rho\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}} \rangle &\geq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t \rangle + \max_{\mathbf{z} \in \mathcal{A}} \langle \nabla f(\mathbf{x}_t), -\rho \mathbf{z} \rangle \\
&\geq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t \rangle - \langle \nabla f(\mathbf{x}_t), \mathbf{x}^\star \rangle \\
&\geq \varepsilon_t
\end{aligned}
$$

combining A and B   By combining case A and case B we obtain:

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) + \min_{\gamma \in [0,1]} \left\{ -\frac{\delta}{2}\gamma \varepsilon_t + \frac{\gamma^2}{2}L\rho^2 \operatorname{radius}(\mathcal{A})^2 \right\}$$

Now, subtracting $f(\mathbf{x}^\star)$ from both sides and setting $C := L\rho^2 \operatorname{radius}(\mathcal{A})^2$, we get

$$\begin{aligned}
\varepsilon_{t+1} \quad &\le \varepsilon_t + \min_{\gamma \in [0,1]} \left\{ -\frac{\delta}{2}\gamma \varepsilon_t + \frac{\gamma^2}{2}C \right\} \\
&\le \varepsilon_t - \frac{2}{\delta' t + 2}\delta' \varepsilon_t + \frac{1}{2}\left(\frac{2}{\delta' t + 2}\right)^2 C,
\end{aligned}$$

where we set $\delta' := \delta/2$ and used $\gamma = \frac{2}{\delta' t + 2} \in [0,1]$ to obtain the second inequality. Finally, we show by induction

$$\varepsilon_t \le \frac{4\left(\frac{2}{\delta}C + \varepsilon_0\right)}{t+4} = 2\frac{\left(\frac{1}{\delta'}C + \varepsilon_0\right)}{\delta' t + 2}$$

for $t \ge 0$.

When $t = 0$ we get $\varepsilon_0 \le \left(\frac{1}{\delta'}C + \varepsilon_0\right)$. Therefore, the base case holds. We now prove the induction step assuming $\varepsilon_t \le \frac{2\left(\frac{1}{\delta'}C + \varepsilon_0\right)}{\delta' t + 2}$ as :

$$\begin{aligned}
\varepsilon_{t+1} &\le \left(1 - \frac{2\delta'}{\delta' t + 2}\right)\varepsilon_t + \frac{1}{2}C\left(\frac{2}{\delta' t + 2}\right)^2 \\
&\le \left(1 - \frac{2\delta'}{\delta' t + 2}\right)\frac{2\left(\frac{1}{\delta'}C + \varepsilon_0\right)}{\delta' t + 2} \\
&\quad + \frac{1}{2}\left(\frac{2}{\delta' t + 2}\right)^2 C + \frac{2}{(\delta' t + 2)^2}\delta' \varepsilon_0 \\
&= \frac{2\left(\frac{1}{\delta'}C + \varepsilon_0\right)}{\delta' t + 2}\left(1 - \frac{2\delta'}{\delta' t + 2} + \frac{\delta'}{\delta' t + 2}\right) \\
&\le \frac{2\left(\frac{1}{\delta'}C + \varepsilon_0\right)}{\delta'(t+1) + 2}.
\end{aligned}$$

Remembering that we set $C = L\rho^2 \operatorname{radius}(\mathcal{A})^2$ concludes the proof.

Fully Corrective non-negative MP: The proof is trivial considering that:

$$f(\mathbf{x}_{t+1}) = \min_{\mathbf{x} \in \operatorname{cone}(\mathcal{S} \cup \mathbf{s}(\mathcal{A},\mathbf{r}))} f(\mathbf{x}) \tag{17}$$

$$\le \min_{\mathbf{x} \in \operatorname{cone}(\mathcal{S} \cup \mathbf{s}(\mathcal{A},\mathbf{r}))} g_{\mathbf{x}_t}(\mathbf{x}) \tag{18}$$

$$\le \min_{\gamma \in \mathbb{R}_{\ge 0}} g_{\mathbf{x}_t}(\mathbf{x}_t + \gamma \tilde{\mathbf{z}}_t) \tag{19}$$

where $\tilde{\mathbf{z}}_t \in A \cup \left\{\frac{-\mathbf{x}_t}{\|\mathbf{x}_t\|_{\mathcal{A}}}\right\}$ as the search space in Equation (18) strictly contain the one in Equation (19). Equation (19) is also the beginning of the proof of the sublinear rate for NNMP which then concludes the proof. □

E   Linear Rate

Theorem' 3. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set containing the origin and let the objective function $f : \mathcal{H} \to \mathbb{R}$ be both L-smooth and μ-strongly convex over $\rho \operatorname{conv}(\mathcal{A} \cup -\mathcal{A})$.*

Then, the suboptimality of the iterates of Algorithm 3 decreases geometrically at each step in which $\gamma < \alpha_{\mathbf{v}_t}$ (henceforth referred to as "good steps") as:

$$\varepsilon_{t+1} \le (1 - \beta)\,\varepsilon_t, \tag{20}$$

where $\beta := \delta^2 \frac{\mu \operatorname{CWidth}(\mathcal{A})^2}{L \operatorname{diam}(\mathcal{A})^2} \in (0,1]$, $\varepsilon_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is the suboptimality at step t and $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO (Equation (3)). For AMP (Algorithm 3), $\beta^{\mathrm{AMP}} = \beta/2$. If $\mu = 0$ Algorithm 3 converges with rate $O(1/k(t))$ where $k(t)$ is the number of "good steps" up to iteration t.

Proof. Let us consider the case of PWMP.

Consider the atoms $\tilde{\mathbf{z}}_t \in \mathcal{A}$ and $\tilde{\mathbf{v}}_t \in \mathcal{S}$ selected by the LMO at iteration t. Due to the smoothness property of f it holds that:

$$f(\mathbf{x}_{t+1}) \leq \min_{\gamma \in \mathbb{R}} f(\mathbf{x}_t) + \gamma \langle \nabla f(\mathbf{x}_t), \tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t \rangle$$

$$+ \frac{L}{2}\gamma^2 \|\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t\|^2.$$

for a good step (i.e. $\gamma < \alpha_{\mathbf{v}_t}$). Note that this also holds for variant 0 of Algorithm 4.

We minimize the upper bound with respect to γ setting $\gamma = -\frac{1}{L}\langle \nabla f(\mathbf{x}_t), \frac{\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t}{\|\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t\|^2}\rangle$. Subtracting $f(\mathbf{x}^\star)$ from both sides and replacing the optimal γ yields:

$$\varepsilon_{t+1} \leq \varepsilon_t - \frac{1}{2L}\left\langle \nabla f(\mathbf{x}_t), \frac{\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t}{\|\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t\|}\right\rangle^2 \tag{21}$$

Now writing the definition of strong convexity, we have the following inequality holding for all $\gamma \in \mathbb{R}$:

$$f(\mathbf{x}_t + \gamma(\mathbf{x}^\star - \mathbf{x}_t)) \geq f(\mathbf{x}_t) + \gamma\langle \nabla f(\mathbf{x}_t), \mathbf{x}^\star - \mathbf{x}_t\rangle +$$

$$\gamma^2 \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{x}_t\|^2$$

We now fix $\gamma = 1$ in the LHS and minimize with respect to γ in the RHS:

$$\varepsilon_t \leq \frac{1}{2\mu}\left\langle \nabla f(\mathbf{x}_t), \frac{\mathbf{x}^\star - \mathbf{x}_t}{\|\mathbf{x}^\star - \mathbf{x}_t\|}\right\rangle^2$$

Combining this with (21) yields:

$$\varepsilon_t - \varepsilon_{t+1} \geq \frac{\mu}{L}\frac{\left\langle \nabla f(\mathbf{x}_t), \frac{\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t}{\|\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t\|}\right\rangle^2}{\left\langle \nabla f(\mathbf{x}_t), \frac{\mathbf{x}^\star - \mathbf{x}_t}{\|\mathbf{x}^\star - \mathbf{x}_t\|}\right\rangle^2}\varepsilon_t \tag{22}$$

We now use Theorem 8 to conclude the proof. For Away-steps MP the proof is trivially extended since $2\min_{\mathbf{z}\in\mathcal{A}-\mathcal{S}}\langle \nabla f(\mathbf{x}_t), \mathbf{z}\rangle \leq \min_{\mathbf{z}\in\mathcal{A},\mathbf{v}\in\mathcal{S}}\langle \nabla f(\mathbf{x}_t), \mathbf{z} - \mathbf{v}\rangle$. Therefore, we obtain the same smoothness upper bound of the PWMP. The rest of the proof proceed as for PWMP with the additional $\frac{1}{2}$ factor.

Sublinear Convergence for $\mu = 0$ If $\mu = 0$ we have for PWMP:

$$f(\mathbf{x}_{t+1}) \leq \min_{\gamma \leq \alpha_{\mathbf{v}_t}} f(\mathbf{x}_t) + \gamma\langle \nabla f(\mathbf{x}_t), \tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t\rangle \tag{23}$$

$$+ \frac{L}{2}\gamma^2\|\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t\|^2. \tag{24}$$

which can be rewritten for a good step (i.e. no clipping is necessary) as:

$$\varepsilon_{t+1} \leq \varepsilon_t + \min_{\gamma \in [0,1]}\left\{-\frac{\delta}{2}\gamma\varepsilon_t + \frac{\gamma^2}{2}L\rho^2 \operatorname{diam}(\mathcal{A})^2\right\}$$

using the same arguments of the proof of Theorem 2. Unfortunately, $\alpha_{\mathbf{v}_t}$ limits the improvement. On the other hand, we can repeat the induction of Theorem 2 for only the good steps. Therefore:

$$\varepsilon_{t+1} \leq \varepsilon_t - \frac{2}{\delta't+2}\delta'\varepsilon_t + \frac{1}{2}\left(\frac{2}{\delta't+2}\right)^2 C,$$

where we set $\delta' := \delta/2$, $C = L\rho^2 \operatorname{diam}(\mathcal{A})^2$ and used $\gamma = \frac{2}{\delta't+2} \in [0,1]$ (since it is a good step this produce a valid upper bound). Finally, we show by induction

$$\varepsilon_t \leq \frac{4\left(\frac{2}{\delta}C + \varepsilon_0\right)}{t+4} = 2\frac{\left(\frac{1}{\delta'}C + \varepsilon_0\right)}{\delta'k(t)+2}$$

where $k(t) \geq 0$ is the number of good steps at iteration t.

When $k(t) = 0$ we get $\varepsilon_0 \leq \left(\frac{1}{\delta'}C + \varepsilon_0\right)$. Therefore, the base case holds. We now prove the induction step assuming $\varepsilon_t \leq \frac{2\left(\frac{1}{\delta'}C + \varepsilon_0\right)}{\delta'k(t)+2}$ as :

$$
\begin{aligned}
\varepsilon_{t+1} &\leq \left(1 - \frac{2\delta'}{\delta'k(t)+2}\right)\varepsilon_t + \frac{1}{2}C\left(\frac{2}{\delta'k(t)+2}\right)^2 \\
&\leq \left(1 - \frac{2\delta'}{\delta'k(t)+2}\right)\frac{2\left(\frac{1}{\delta'}C+\varepsilon_0\right)}{\delta'k(t)+2} \\
&\quad + \frac{1}{2}\left(\frac{2}{\delta'k(t)+2}\right)^2 C + \frac{2}{(\delta'k(t)+2)^2}\delta'\varepsilon_0 \\
&= \frac{2\left(\frac{1}{\delta'}C+\varepsilon_0\right)}{\delta'k(t)+2}\left(1 - \frac{2\delta'}{\delta'k(t)+2} + \frac{\delta'}{\delta'k(t)+2}\right) \\
&\leq \frac{2\left(\frac{1}{\delta'}C+\varepsilon_0\right)}{\delta'(k(t)+1)+2}.
\end{aligned}
$$

For AFW the procedure is the same but the linear term of Equation 23 is divided by two. We proceed as before with the only difference that we call $\delta' = \delta/4$. $\qquad\square$

E.1 Proof sketch for linear rate convergence of FCMP

Theorem' 3. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set containing the origin and let the objective function $f \colon \mathcal{H} \to \mathbb{R}$ be both L-smooth and μ-strongly convex over $\rho\operatorname{conv}(\mathcal{A} \cup -\mathcal{A})$.*

Then, the suboptimality of the iterates of Algorithm 4 decreases geometrically at each step in which $\gamma < \alpha_{\mathbf{v}_t}$ (henceforth referred to as "good steps") as:

$$
\varepsilon_{t+1} \leq (1 - \beta)\,\varepsilon_t, \tag{25}
$$

where $\beta := \delta^2 \frac{\mu \operatorname{CWidth}(\mathcal{A})^2}{L \operatorname{diam}(\mathcal{A})^2} \in (0,1]$, $\varepsilon_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is the suboptimality at step t and $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO (Equation (3)).

Proof. The proof is trivial noticing that:

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) &= \min_{\mathbf{x}\in\operatorname{cone}(\mathcal{S}\cup\mathbf{s}(\mathcal{A},\mathbf{r}))} f(\mathbf{x}) \\
&\leq \min_{\mathbf{x}\in\operatorname{cone}(\mathcal{S}\cup\mathbf{s}(\mathcal{A},\mathbf{r}))} g_{\mathbf{x}_t}(\mathbf{x}) \\
&\leq \min_{\gamma\leq\alpha_{\mathbf{v}_t}} g_{\mathbf{x}_t}(\mathbf{x}_t + \gamma(\mathbf{z}_t - \mathbf{v}_t))
\end{aligned}
$$

which is the beginning of the proof of Theorem 3. Note that there are no bad steps for variant 1. Since we minimize f at each iteration, $\mathbf{v}_t$ is always zero and each step is unconstrained (i.e., no bad steps). $\qquad\square$

F Pyramidal Width

Let us first recall some definitions from [17].

Directional Width

$$
dirW(\mathcal{A}, \mathbf{r}) := \max_{\mathbf{s},\mathbf{v}\in\mathcal{A}} \left\langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{s} - \mathbf{v} \right\rangle \tag{26}
$$

Pyramidal Directional Width

$$
PdirW(\mathcal{A}, \mathbf{r}, \mathbf{x}) := \min_{\mathcal{S}\in\mathcal{S}_{\mathbf{x}}} dirW(\mathcal{S} \cup \{\mathbf{s}(\mathcal{A},\mathbf{r})\}, \mathbf{r}) \tag{27}
$$

Where $\mathcal{S}_{\mathbf{x}} := \{\mathcal{S} \mid \mathcal{S} \subset \mathcal{A}$ such that $\mathbf{x}$ is a proper convex combination of all the elements in $\mathcal{S}\}$ and $\mathbf{s}(\mathcal{A}, \mathbf{r}) := \max_{\mathbf{s}\in\mathcal{A}}\langle\frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{s}\rangle$.

Pyramidal Width

$$PWidth(\mathcal{A}) := \min_{\substack{\mathcal{K}\in\text{faces}(\text{conv}(\mathcal{A})) \\ \mathbf{x}\in\mathcal{K} \\ \mathbf{r}\in\text{cone}(\mathcal{K}-\mathbf{x})\setminus\{\mathbf{0}\}}} PdirW(\mathcal{K}\cap\mathcal{A},\mathbf{r},\mathbf{x})$$

Inspired by the notion of pyramidal width we now define the cone width of a set $\mathcal{A}$.

Cone Width

$$\text{CWidth}(\mathcal{A}) := \min_{\substack{\mathcal{K}\in\text{g-faces}(\text{cone}(\mathcal{A})) \\ \mathbf{x}\in\mathcal{K} \\ \mathbf{r}\in\text{cone}(\mathcal{K}-\mathbf{x})\setminus\{\mathbf{0}\}}} PdirW(\mathcal{K}\cap\mathcal{A},\mathbf{r},\mathbf{x})$$

The linear rate analysis is dominated by the fact that, similarly as in FW, many step directions are constrained (the ones pointing outside of the cone). So these arguments are in line with [17] and the techniques are adapted here. Lemma 7 is a minor modification of [[17], Lemma 5], see also their Figure 3. If the gradient is not feasible, the vector with maximum inner product must lie on a facet. Furthermore, it has the same inner product with the gradient and with its orthogonal projection on that facet. While first proof of Lemma 7 follows [17], we also give a different proof which does not use the KKT conditions.

Lemma 7. *Let $\mathbf{x}$ be a reference point inside a polytope $\mathcal{K}\in\text{g-faces}(\text{cone}(\mathcal{A}))$ and $\mathbf{r}\in\text{lin}(\mathcal{K})$ is not a feasible direction from $\mathbf{x}$. Then, a feasible direction in $\mathcal{K}$ minimizing the angle with $\mathbf{r}$ lies on a facet $\mathcal{K}'$ of $\mathcal{K}$ that includes $\mathbf{x}$:*

$$\max_{\mathbf{e}\in\text{cone}(\mathcal{K}-\mathbf{x})}\left\langle\mathbf{r},\frac{\mathbf{e}}{\|\mathbf{e}\|}\right\rangle = \max_{\mathbf{e}\in\text{cone}(\mathcal{K}'-\mathbf{x})}\left\langle\mathbf{r},\frac{\mathbf{e}}{\|\mathbf{e}\|}\right\rangle$$

$$= \max_{\mathbf{e}\in\text{cone}(\mathcal{K}'-\mathbf{x})}\left\langle\mathbf{r}',\frac{\mathbf{e}}{\|\mathbf{e}\|}\right\rangle$$

where $\mathbf{r}'$ is the orthogonal projection of $\mathbf{r}$ onto $\text{lin}(\mathcal{K}')$

Proof. Let us center the problem in $\mathbf{x}$. We rewrite the optimization problem as:

$$\max_{\mathbf{e}\in\text{cone}(\mathcal{K}),\|\mathbf{e}\|=1}\langle\mathbf{r},\mathbf{e}\rangle$$

and suppose by contradiction that $\mathbf{e}$ is in the relative interior of the cone. By the KKT necessary conditions we have that $\mathbf{e}^\star$ is collinear with $\mathbf{r}$. Therefore $\mathbf{e}^\star = \pm\mathbf{r}$. Now we know that $\mathbf{r}$ is not feasible, therefore the solution is $\mathbf{e}^\star = -\mathbf{r}$. By Cauchy-Schwarz we know that this solution is minimizing the inner product which is absurd. Therefore, $\mathbf{e}^\star$ must lie on a face of the cone. The last equality is trivial considering that $\mathbf{r}'$ is the orthogonal projection of $\mathbf{r}$ onto $\text{lin}(\mathcal{K}')$.

Alternative proof. This proof extends the traditional proof technique of [17] to infinitely many constraints. We also reported the FW inspired proof for the readers that are more familiar with the FW analysis. Using proposition 2.11 of [45] (we also use their notation) the first order optimality condition minimizing a function J in a general Hilbert space given a closed set $\mathcal{K}$ is that the directional derivative computed at the optimum $\bar{u}$ satisfy $J'(\bar{u})v \geq 0 \ \forall v \in \mathcal{T}(\mathcal{K}-\bar{u})$. Let us now assume that $\bar{u}$ is in the relative interior of $\mathcal{K}$. Then $\mathcal{T}(\mathcal{K}-\bar{u}) = \mathcal{H}$. Furthermore, $J'(\bar{u})v = \langle\mathbf{r},v\rangle$ which is clearly not greater or equal than zero for any element of $\mathcal{H}$. $\qquad\square$

Theorem 8 is the key argument to conclude the proof of Theorem 3 from Equation (22): we have to bound the ratio of those inner products with the cone width.

Theorem 8. *Let $\mathbf{r} = -\nabla f(\mathbf{x}_t)$, $\mathbf{x}\in\text{cone}(\mathcal{A})$, $\mathcal{S}$ be the active set and $\mathbf{z}$ and $\mathbf{v}$ obtained as in Algorithm 3. Then, using the notation from Lemma 7:*

$$\frac{\langle\mathbf{r},\mathbf{d}\rangle}{\langle\mathbf{r},\hat{\mathbf{e}}\rangle} \geq \text{CWidth}(\mathcal{A}) \tag{28}$$

where $\mathbf{d} := \mathbf{z} - \mathbf{v}$, $\hat{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|}$ and $\mathbf{e} = \mathbf{x}^\star - \mathbf{x}_t$.

Proof. As we already discussed we can consider $\mathbf{x} \in \text{conv}(\mathcal{A})$ instead of $\mathbf{x}_t \in \text{cone}(\mathcal{A})$ since both the cone and the set of feasible direction are invariant to a rescaling of $\mathbf{x}$ by a strictly positive constant. Let us center all the vectors in $\mathbf{x}$, then $\hat{\mathbf{e}}$ is just a vector with norm 1 in some face. As $\mathbf{x}$ is not optimal, by convexity we have that $\langle \mathbf{r}, \hat{\mathbf{e}} \rangle > 0$. By Cauchy-Schwartz we know that $\langle \mathbf{r}, \hat{\mathbf{e}} \rangle \leq \|\mathbf{r}\|$ since $\langle \mathbf{r}, \hat{\mathbf{e}} \rangle > 0$ and $\|\hat{\mathbf{e}}\| = 1$. By definition of $\mathbf{d}$ we have:

$$
\begin{aligned}
\langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{d} \rangle &= \max_{\mathbf{z} \in \mathcal{A}, \mathbf{v} \in \mathcal{S}} \langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{z} - \mathbf{v} \rangle \\
&\geq \min_{\mathcal{S} \subset \mathcal{S}_{\mathbf{x}}} \max_{\mathbf{z} \in \mathcal{A}, \mathbf{v} \in \mathcal{S}} \langle \frac{\mathbf{r}}{\|\mathbf{r}\|}, \mathbf{z} - \mathbf{v} \rangle \\
&= PdirW(\mathcal{A}, \mathbf{r}, \mathbf{x}).
\end{aligned}
$$

Now, if $\mathbf{r}$ is a feasible direction from $\mathbf{x}$ Equation (28) is proved (note that $PdirW(\mathcal{A}, \mathbf{r}, \mathbf{x}) \geq \text{CWidth}(\mathcal{A})$ as $\text{conv}(\mathcal{A}) \in \text{g-faces}(\text{cone}(\mathcal{A}))$ and $\text{conv}(\mathcal{A}) \cap \mathcal{A} = \mathcal{A}$). If $\mathbf{r}$ is not a feasible direction it means that $\mathbf{x}$ is on a face of $\text{cone}(\mathcal{A})$ and $\mathbf{r}$ points to the exterior of $\text{cone}(\mathcal{A})$ from $\mathbf{x}$. We then project $\mathbf{r}$ on the faces of $\text{cone}(\mathcal{A})$ containing $\mathbf{x}$ until it is a feasible direction. We start by lower bounding the ratio of the two inner products replacing $\hat{\mathbf{e}}$ with a vector of norm 1 in the cone that has maximum inner product with $\mathbf{r}$ (with abuse of notation we still call it $\hat{\mathbf{e}}$). We then write:

$$
\frac{\langle \mathbf{r}, \mathbf{d} \rangle}{\langle \mathbf{d}, \hat{\mathbf{e}} \rangle} \geq \left(\max_{\mathbf{z} \in \mathcal{A}, \mathbf{v} \in \mathcal{S}} \langle \mathbf{r}, \mathbf{z} - \mathbf{v} \rangle \right) \cdot \left(\max_{\mathbf{e} \in \text{cone}(\mathcal{A} - \mathbf{x})} \langle \mathbf{r}, \frac{\mathbf{e}}{\|\mathbf{e}\|} \rangle \right)^{-1}
$$

Let us assume that $\mathbf{r}$ is not feasible but without loss of generality is in $\text{lin}(\mathcal{A})$ since orthogonal components to $\text{lin}(\mathcal{A})$ does not influence the inner product with elements in $\text{lin}(\mathcal{A})$. Using Lemma 7 we know that:

$$
\begin{aligned}
\max_{\mathbf{e} \in \text{cone}(\mathcal{K} - \mathbf{x})} \langle \mathbf{r}, \frac{\mathbf{e}}{\|\mathbf{e}\|} \rangle &= \max_{\mathbf{e} \in \text{cone}(\mathcal{K}' - \mathbf{x})} \langle \mathbf{r}, \frac{\mathbf{e}}{\|\mathbf{e}\|} \rangle \\
&= \max_{\mathbf{e} \in \text{cone}(\mathcal{K}' - \mathbf{x})} \langle \mathbf{r}', \frac{\mathbf{e}}{\|\mathbf{e}\|} \rangle
\end{aligned}
$$

Let us now consider the reduced cone $\text{cone}(\mathcal{K}')$ as $\mathbf{r} \in \text{lin}(\mathcal{K}')$. For the numerator we obtain:

$$
\max_{\mathbf{z} \in \mathcal{A}, \mathbf{v} \in \mathcal{S}} \langle \mathbf{r}, \mathbf{z} - \mathbf{v} \rangle \overset{\mathcal{K}' \subset \mathcal{A}}{\geq} \max_{\mathbf{z} \in \mathcal{K}'} \langle \mathbf{r}, \mathbf{z} \rangle + \max_{\mathbf{v} \in \mathcal{S}} \langle -\mathbf{r}, \mathbf{v} \rangle
$$

Putting numerator and denominator together we obtain:

$$
\frac{\langle \mathbf{r}, \mathbf{d} \rangle}{\langle \mathbf{d}, \hat{\mathbf{e}} \rangle} \geq \left(\max_{\substack{\mathbf{z} \in \mathcal{K}' \\ \mathbf{v} \in \mathcal{S}}} \langle \mathbf{r}', \mathbf{z} - \mathbf{v} \rangle \right) \cdot \left(\max_{\mathbf{e} \in \text{cone}(\mathcal{K}' - \mathbf{x})} \langle \mathbf{r}', \frac{\mathbf{e}}{\|\mathbf{e}\|} \rangle \right)^{-1}
$$

Note that $\mathcal{S} \subset \mathcal{K}'$. Indeed, $\mathbf{x}$ is a proper convex combination of the elements of $\mathcal{S}$ and $\mathbf{x} \in \mathcal{K}' \subset \text{conv}(\mathcal{A})$. Now if $\mathbf{r}'$ is a feasible direction in $\text{cone}(\mathcal{K}' - \mathbf{x})$ we obtain the cone width since $\text{cone}(\mathcal{K}')$ is a face of $\text{cone}(\mathcal{A})$. If not we reiterate the procedure projecting onto a lower dimensional face $\mathcal{K}''$. Eventually, we will obtain a feasible direction. Since $\langle \mathbf{r}, \hat{\mathbf{e}} \rangle \neq 0$ we will obtain $\mathbf{r}_{final} \neq \mathbf{0}$. $\qquad\square$

Lemma 4. *If the origin is in the relative interior of* $\text{conv}(\mathcal{A})$ *with respect to its linear span, then* $\text{cone}(\mathcal{A}) = \text{lin}(\mathcal{A})$ *and* $\text{CWidth}(\mathcal{A}) = \text{mDW}(\mathcal{A})$.

Proof. Let us first rewrite the definition of cone width:

$$
\text{CWidth}(\mathcal{A}) := \min_{\substack{\mathcal{K} \in \text{g-faces}(\text{cone}(\mathcal{A})) \\ \mathbf{x} \in \mathcal{K} \\ \mathbf{r} \in \text{cone}(\mathcal{K} - \mathbf{x}) \setminus \{\mathbf{0}\}}} PdirW(\mathcal{K} \cap \mathcal{A}, \mathbf{r}, \mathbf{x}).
$$

The minimum is over all the feasible directions of the gradient from every point in the domain. It is not restrictive to consider $\mathbf{r}$ parallel to $\text{lin}(\mathcal{A})$ (because the orthogonal component has no influence). Therefore, from every point $\mathbf{x} \in \text{lin}(\mathcal{A})$ every $\mathbf{r} \in \text{lin}(\mathcal{A})$ is a feasible direction. The geometric constant then becomes:

$$
\text{CWidth}(\mathcal{A}) = \min_{\substack{\mathcal{K} \in \text{g-faces}(\text{cone}(\mathcal{A})) \\ \mathbf{x} \in \mathcal{K} \\ \mathbf{r} \in \text{lin}(\mathcal{A}) \setminus \{\mathbf{0}\}}} PdirW(\mathcal{K} \cap \mathcal{A}, \mathbf{r}, \mathbf{x})
$$

Let us now assume by contradiction that for any $\mathcal{K} \in$ g-faces we have:
$$\mathbf{0} \notin \arg\min_{\mathbf{x}\in\mathcal{K}} \min_{\mathbf{r}\in\mathrm{lin}(\mathcal{A})\setminus\{\mathbf{0}\}} PdirW(\mathcal{K}\cap\mathcal{A},\mathbf{r},\mathbf{x}) \tag{29}$$
Therefore, $\exists\mathbf{v}\in\mathcal{S}$ such that $\mathbf{v}\neq\mathbf{0}$ for any of the $\mathbf{x}$ minimizing (29). By definition, we have $\mathbf{0}\in\mathcal{S}$, which yields $\max_{\mathbf{v}\in\mathcal{S}}\langle\mathbf{r},-\mathbf{v}\rangle\geq 0$ for every $\mathbf{r}$. Therefore, $\langle\mathbf{r},\mathbf{z}-\mathbf{v}\rangle\geq\langle\mathbf{r},\mathbf{z}\rangle$ which is absurd because we assumed zero was in the set of minimizers of (29). So $\mathbf{0}$ minimize the cone directional width which yields $\mathcal{S}_{\mathbf{x}}=\{\mathbf{0}\}$ and $\mathbf{v}=\mathbf{0}$. In conclusion we have:
$$\mathrm{CWidth}(\mathcal{A}) = \min_{\mathbf{d}\in\mathrm{lin}(\mathcal{A})} \max_{\mathbf{z}\in\mathcal{A}}\langle\frac{\mathbf{d}}{\|\mathbf{d}\|},\mathbf{z}\rangle = \mathrm{mDW}(\mathcal{A})$$
$\square$

G Affine Invariant Sublinear Rate

Theorem' 5. *Let $\mathcal{A}\subset\mathcal{H}$ be a bounded set with $\mathbf{0}\in\mathcal{A}$, $\rho := \max\left\{\|\mathbf{x}^\star\|_\mathcal{A},\|\mathbf{x}_0\|_\mathcal{A},\ldots,\|\mathbf{x}_T\|_\mathcal{A}\right\} < \infty$. Assume f has smoothness constant $C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}$. Then, Algorithm 5 converges for $t\geq 0$ as*
$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{4\left(\frac{2}{\delta}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} + \varepsilon_0\right)}{\delta t + 4},$$
where $\delta\in(0,1]$ is the relative accuracy parameter of the employed approximate LMO *(3).*

Proof. Recall that $\tilde{\mathbf{z}}_t$ is the atom returned by the inexact LMO after the comparison with $-\frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_\mathcal{A}}$ at the current iteration t.

We start by upper bounding f on $\rho\,\mathrm{conv}(\bar{\mathcal{A}})$ using smoothness as follows:
$$f(\mathbf{x}_{t+1}) \leq \min_{\gamma\in[0,1]} f(\mathbf{x}_t) + \gamma\langle\nabla f(\mathbf{x}_t),\rho\tilde{\mathbf{z}}_t\rangle + \frac{\gamma^2}{2}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}$$
We now proceed bounding the linear term as done in the proof of Theorem 2 for **case A** and **case B** obtaining:
$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \min_{\gamma\in[0,1]}\left\{-\frac{\delta}{2}\gamma\varepsilon_t + \frac{\gamma^2}{2}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}\right\}$$
Now, subtracting $f(\mathbf{x}^\star)$ from both sides we get
$$\begin{aligned}\varepsilon_{t+1} &\leq \varepsilon_t + \min_{\gamma\in[0,1]}\left\{-\frac{\delta}{2}\gamma\varepsilon_t + \frac{\gamma^2}{2}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}\right\}\\ &\leq \varepsilon_t - \frac{2}{\delta't+2}\delta'\varepsilon_t + \frac{1}{2}\left(\frac{2}{\delta't+2}\right)^2 C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})},\end{aligned}$$
where we set $\delta' := \delta/2$ and used $\gamma = \frac{2}{\delta't+2}\in[0,1]$ to obtain the second inequality. Finally, we show by induction
$$\varepsilon_t \leq \frac{4\left(\frac{2}{\delta}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} + \varepsilon_0\right)}{t+4} = 2\frac{\left(\frac{1}{\delta'}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} + \varepsilon_0\right)}{\delta't+2}$$
for $t\geq 0$.

When $t = 0$ we get $\varepsilon_0 \leq \left(\frac{1}{\delta'}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} + \varepsilon_0\right)$. Therefore, the base case holds. We now prove the induction step assuming $\varepsilon_t \leq \frac{2\left(\frac{1}{\delta'}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}+\varepsilon_0\right)}{\delta't+2}$ as :
$$\begin{aligned}\varepsilon_{t+1} &\leq \left(1-\frac{2\delta'}{\delta't+2}\right)\varepsilon_t + \frac{1}{2}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}\left(\frac{2}{\delta't+2}\right)^2\\ &\leq \left(1-\frac{2\delta'}{\delta't+2}\right)\frac{2\left(\frac{1}{\delta'}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}+\varepsilon_0\right)}{\delta't+2}\\ &\quad + \frac{1}{2}\left(\frac{2}{\delta't+2}\right)^2 C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} + \frac{2}{(\delta't+2)^2}\delta'\varepsilon_0\\ &= \frac{2\left(\frac{1}{\delta'}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}+\varepsilon_0\right)}{\delta't+2}\left(1-\frac{2\delta'}{\delta't+2}+\frac{\delta'}{\delta't+2}\right)\\ &\leq \frac{2\left(\frac{1}{\delta'}C^{\mathsf{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}+\varepsilon_0\right)}{\delta'(t+1)+2}.\end{aligned}$$

$\square$

We next explore the relationship of $C^{\text{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}$ and the smoothness parameter. Recall that f is *L-smooth* with respect to a given norm $\|.\|$ over a set $\mathcal{Q}$ if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{Q}, \tag{30}$$

where $\|.\|_*$ is the dual norm of $\|.\|$.

Lemma 9. *Assume f is L-smooth with respect to a given norm $\|.\|$, over the set $\text{conv}(\mathcal{A})$. Then,*

$$C^{\text{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} \leq L\,\rho^2\,\text{radius}_{\|.\|}(\mathcal{A})^2 \tag{31}$$

Proof. By the definition of smoothness of f with respect to $\|.\|$,

$$D(\mathbf{y}, \mathbf{x}) \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Hence, from the definition of $C^{\text{MP}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}$,

$$
\begin{aligned}
C^{\text{MP}}_{f,\mathcal{A}} &\leq \sup_{\substack{\mathbf{s}\in\rho\mathcal{A},\mathbf{x}\in\text{conv}(\rho\mathcal{A}) \\ \gamma\in[0,1] \\ \mathbf{y}=\mathbf{x}+\gamma\mathbf{s}}} \frac{2}{\gamma^2}\frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\
&= L\rho^2 \sup_{\mathbf{s}\in\mathcal{A}} \|\mathbf{s}\|^2 \\
&= L\,\rho^2\,\text{radius}_{\|.\|}(\mathcal{A})^2 .
\end{aligned}
$$

$\square$

H   Affine Invariant Linear Rate

Theorem' 6. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set containing the origin and let the objective function $f\colon \mathcal{H} \to \mathbb{R}$ have smoothness constant $C^{\text{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}$ and strong convexity constant $\mu^{A}_{f,\rho\mathcal{A}}$*

Then, the suboptimality of the iterates of Algorithm 3 and 4 decreases geometrically at each step in which $\gamma < \alpha_{\mathbf{v}_t}$ (henceforth referred to as "good steps") as:

$$\varepsilon_{t+1} \leq (1-\beta)\,\varepsilon_t, \tag{32}$$

where $\beta := \delta^2 \frac{\mu^{A}_{f,\rho\mathcal{A}}}{C^{\text{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}} \in (0,1]$, $\varepsilon_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is the suboptimality at step t and $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO (Equation (3)). For AMP (Algorithm 3), $\beta^{\text{AMP}} = \beta/2$. If $\mu^{A}_{f,\rho\mathcal{A}} = 0$ Algorithm 3 converges with rate $O(1/k(t))$ where $k(t)$ is the number of "good steps" up to iteration t.

Proof. Let us first consider the PWMP update. Using the definition of $C^{\text{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}$ we upper-bound f on $\rho\,\text{conv}(\mathcal{A})$ as follows

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq \min_{\gamma\in[0,1]} f(\mathbf{x}_t) + \gamma\langle\nabla f(\mathbf{x}_t), \rho\tilde{\mathbf{z}}_t - \rho\tilde{\mathbf{v}}_t\rangle \\
&\quad + \frac{\gamma^2}{2}C^{\text{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} \\
&= \min_{\gamma\in\mathbb{R}} f(\mathbf{x}_t) + \gamma\langle\nabla f(\mathbf{x}_t), \rho\tilde{\mathbf{z}}_t - \rho\tilde{\mathbf{v}}_t\rangle \\
&\quad + \frac{\gamma^2}{2}C^{\text{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})} \\
&= f(\mathbf{x}_t) - \frac{\rho^2}{2C^{\text{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}}\langle\nabla f(\mathbf{x}_t), \tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t\rangle^2 .
\end{aligned}
$$

This upper bound holds for Algorithm 6 every time $\rho\gamma < \alpha_{\mathbf{v}}$ as $\rho\gamma$ minimizing the RHS of the first equality coincides with the update of Algorithm 5 Line 5. The first equality holds as $C^{\text{A}}_{f,\rho(\mathcal{A}\cup-\mathcal{A})}$

is defined on $\rho \operatorname{conv}(\mathcal{A})$ and $\rho \operatorname{conv}(\mathcal{A})$ contains all iterates by definition, so that the unconstrained minimum lies in $[0, 1]$ assuming $\rho \gamma < \alpha_{\mathbf{v}}$.

Using $\varepsilon_t = f(\mathbf{x}^\star) - f(\mathbf{x}_t)$, we can lower bound the error decay as

$$\varepsilon_t - \varepsilon_{t+1} \geq \frac{\rho^2}{2 C^{\mathrm{A}}_{f, \rho(\mathcal{A} \cup -\mathcal{A})}} \left\langle \nabla f(\mathbf{x}_t), \tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t \right\rangle^2 . \tag{33}$$

Starting from the definition of $\mu^A_{f, \rho \mathcal{A}}$ we get,

$$
\begin{aligned}
\frac{\gamma(\mathbf{x}_t, \mathbf{x}^\star)^2}{2} \mu^A_{f, \rho \mathcal{A}} &\leq \quad f(\mathbf{x}^\star) - f(\mathbf{x}_t) - \langle \nabla f(\mathbf{x}_t), \mathbf{x}^\star - \mathbf{x}_t \rangle \\
&= \quad -\varepsilon_t \\
&\quad + \quad \gamma(\mathbf{x}_t, \mathbf{x}^\star) \langle -\nabla f(\mathbf{x}_t), \mathbf{s}(\mathbf{x}_t) - \mathbf{v}(\mathbf{x}) \rangle,
\end{aligned}
$$

which gives

$$
\begin{aligned}
\varepsilon_t &\leq \quad -\frac{\gamma(\mathbf{x}_t, \mathbf{x}^\star)^2}{2} \mu^A_{f, \rho \mathcal{A}} \\
&\quad + \quad \gamma(\mathbf{x}_t, \mathbf{x}^\star) \langle -\nabla f(\mathbf{x}_t), \mathbf{s}(\mathbf{x}_t) - \mathbf{v}(\mathbf{x}) \rangle \tag{34} \\
&\leq \quad \frac{\langle -\nabla f(\mathbf{x}_t), \mathbf{s}(\mathbf{x}_t) - \mathbf{v}(\mathbf{x}) \rangle^2}{2 \mu^A_{f, \rho \mathcal{A}}} \\
&= \quad \frac{\langle -\nabla f(\mathbf{x}_t), \rho(\tilde{\mathbf{z}}_t - \tilde{\mathbf{v}}_t) \rangle^2}{2 \delta^2 \mu^A_{f, \rho \mathcal{A}}} \tag{35}
\end{aligned}
$$

where the last inequality is by the quality of the approximate LMO as used in the algorithm, as defined in (3).

Combining equations (33) and (35), we have

$$\varepsilon_t - \varepsilon_{t+1} \geq \delta^2 \frac{\mu^A_{f, \rho \mathcal{A}}}{C^{\mathrm{A}}_{f, \rho(\mathcal{A} \cup -\mathcal{A})}} \varepsilon_t,$$

which proves the claimed result. The proof for AMP and FCMP follows directly using the same argument used in the proof of Theorem 3. The upper bound used in the FCMP is the affine invariant notion of smoothness. The proof steps for the sublinear convergence is the same as the one of Theorem 3 replacing C with $C^{\mathrm{A}}_{f, \rho(\mathcal{A} \cup -\mathcal{A})}$. $\qquad \square$

Lemma 10. *If f is μ strongly convex over the domain $\operatorname{conv}(\rho \mathcal{A})$ with respect to some arbitrary cholsen norm $\| \cdot \|$, then*

$$\mu^A_{f, \rho \mathcal{A}} \geq \mu \operatorname{CWidth}(\mathcal{A})^2$$

Proof. From the strong convexity:

$$
\begin{aligned}
\mu^A_{f, \rho \mathcal{A}} &= \inf_{\mathbf{x} \in \operatorname{conv}(\rho \mathcal{A})} \inf_{\substack{\mathbf{x}^\star \in \operatorname{conv}(\rho \mathcal{A}) \\ \langle \nabla f(\mathbf{x}), \mathbf{x}^\star - \mathbf{x} \rangle < 0}} \frac{2}{\gamma(\mathbf{x}, \mathbf{x}^\star)} D(\mathbf{x}^\star, \mathbf{x}) \\
&\geq \inf_{\substack{\mathbf{x}, \mathbf{x}^\star \in \operatorname{conv}(\rho \mathcal{A}), \\ \langle -\nabla f(\mathbf{x}), \mathbf{x}^\star - \mathbf{x} \rangle > 0}} \mu \left(\frac{\langle -\nabla f(\mathbf{x}), \mathbf{s}(\mathbf{x}) - \mathbf{v}(\mathbf{x}) \rangle}{\langle -\nabla f(\mathbf{x}), \frac{\mathbf{x}^\star - \mathbf{x}}{\|\mathbf{x}^\star - \mathbf{x}\|_{\mathcal{A}}} \rangle} \right)^2 \\
&\geq \mu \operatorname{CWidth}(\mathcal{A})^2
\end{aligned}
$$

where in the last inequality we used Theorem 8.

The proof for away-steps uses the same argument we used in the norm based rate. $\qquad \square$

Lemma 11. *Assume f is L-smooth with respect to a given norm $\|.\|$, over the set $\operatorname{conv}(\mathcal{A})$. Then,*

$$C^{\mathrm{A}}_{f, \rho(\mathcal{A} \cup -\mathcal{A})} \leq L \rho^2 \operatorname{diam}_{\|.\|}(\mathcal{A})^2 \tag{36}$$

Proof. By the definition of smoothness of f with respect to $\|.\|$,

$$D(\mathbf{y}, \mathbf{x}) \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

Hence, from the definition of $C^{\mathsf{MP}}_{f,\rho(\mathcal{A} \cup -\mathcal{A})}$,

$$
\begin{aligned}
C^{\mathsf{MP}}_{f,\mathcal{A}} \quad &\leq \quad \sup_{\substack{\mathbf{s}\in\rho\mathcal{A},\mathbf{x}\in\mathrm{conv}(\rho\mathcal{A}) \\ \mathbf{v}\in\mathcal{S} \\ \gamma\in[0,1] \\ \mathbf{y}=\mathbf{x}+\gamma(\mathbf{s}-\mathbf{v})}} \frac{2}{\gamma^2}\frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 \\
&= \quad L\rho^2 \sup_{\substack{\mathbf{x}\in\mathrm{conv}(\rho\mathcal{A}) \\ \mathbf{s}\in\mathcal{A} \\ \mathbf{v}\in\mathcal{S}}} \|\mathbf{s}-\mathbf{v}\|^2 \\
&= \quad L\,\rho^2\,\mathrm{diam}_{\|.\|}(\mathcal{A})^2\,.
\end{aligned}
$$

$\square$

Supplementary References

[41] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[42] Mário César Ugulino Araújo, Teresa Cristina Bezerra Saldanha, Roberto Kawakami Harrop Galvao, Takashi Yoneyama, Henrique Caldas Chame, and Valeria Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

[43] Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

[44] P Bühlmann and B Yu. Boosting, model selection, lasso and nonnegative garrote. Technical Report 127, Seminar für Statistik ETH Zürich, 2005.

[45] Martin Burger. Infinite-dimensional optimization and optimal design. 2003.

[46] Andrzej Cichocki and PHAN Anh-Huy. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[47] Ernie Esser, Yifei Lou, and Jack Xin. A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM Journal on Imaging Sciences*, 6(4):2010–2046, 2013.

[48] Nicolas Gillis. Successive nonnegative projection algorithm for robust nonnegative blind source separation. *SIAM Journal on Imaging Sciences*, 7(2):1420–1450, 2014.

[49] Nicolas Gillis and François Glineur. Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4):1085–1105, 2012.

[50] Nicolas Gillis, Da Kuang, and Haesun Park. Hierarchical clustering of hyperspectral images using rank-two nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):2066–2078, 2015.

[51] Nicolas Gillis and Robert Luce. A fast gradient method for nonnegative sparse regression with self dictionary. *arXiv preprint arXiv:1610.01349*, 2016.

[52] Xiawei Guo, Quanming Yao, and James T Kwok. Efficient sparse low-rank tensor completion using the Frank-Wolfe algorithm. In *AAAI Conference on Artificial Intelligence*, 2017.

[53] Cho-Jui Hsieh and Inderjit S Dhillon. Fast coordinate descent methods with variable selection for nonnegative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1064–1072. ACM, 2011.

[54] Martin Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.

[55] Hyunsoo Kim, Haesun Park, and Lars Elden. Non-negative tensor factorization based on alternating large-scale non-negativity-constrained least squares. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 1147–1151. IEEE, 2007.

[56] Jingu Kim and Haesun Park. Fast nonnegative tensor factorization with an active-set-like method. In *High-Performance Scientific Computing*, pages 311–326. Springer, 2012.

[57] Ivica Kopriva and Andrzej Cichocki. Nonlinear band expansion and 3d nonnegative tensor factorization for blind decomposition of magnetic resonance image of the brain. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 490–497. Springer, 2010.

[58] Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *ICML (1)*, pages 231–239, 2013.

[59] Simon Lacoste-Julien and Martin Jaggi. An Affine Invariant Linear Convergence Analysis for Frank-Wolfe Algorithms. In *NIPS 2013 Workshop on Greedy Algorithms, Frank-Wolfe and Friends*, December 2013.

[60] Simon Lacoste-Julien and Martin Jaggi. On the Global Linear Convergence of Frank-Wolfe Optimization Variants. In *NIPS 2015*, pages 496–504, 2015.

[61] Sören Laue. A Hybrid Algorithm for Convex Semidefinite Optimization. In *ICML*, 2012.

[62] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[63] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and frank-wolfe. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[64] Enes Makalic and Daniel F Schmidt. Logistic regression with the nonnegative garrote. In *Australasian Joint Conference on Artificial Intelligence*, pages 82–91. Springer, 2011.

[65] José MP Nascimento and José MB Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4):898–910, 2005.

[66] Max Welling and Markus Weber. Positive tensor factorization. *Pattern Recognition Letters*, 22(12):1255–1261, 2001.

[67] Mehrdad Yaghoobi, Di Wu, and Mike E Davies. Fast non-negative orthogonal matching pursuit. *IEEE Signal Processing Letters*, 22(9):1229–1233, 2015.

[68] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, 14(1):899–925, April 2013.