A Unified Optimization View on Generalized Matching Pursuit and Frank-Wolfe

Francesco Locatello
ETH Zürich

Rajiv Khanna*
UT Austin

Michael Tschannen*
ETH Zürich

Martin Jaggi
EPFL

Abstract

Two of the most fundamental prototypes of greedy optimization are the matching pursuit and Frank-Wolfe algorithms. In this paper, we take a unified view on both classes of methods, leading to the first explicit convergence rates of matching pursuit methods in an optimization sense, for general sets of atoms. We derive sublinear $(1/t)$ convergence for both classes on general smooth objectives, and linear convergence on strongly convex objectives, as well as a clear correspondence of algorithm variants. Our presented algorithms and rates are affine invariant, and do not need any incoherence or sparsity assumptions.

1 Introduction

During the past decade, greedy algorithms have attracted significant attention and led to many success stories in machine learning and signal processing (e.g., compressed sensing), and optimization in general. The most prominent representatives are matching pursuit (MP) algorithms on one hand (Mallat & Zhang, 1993), such as, e.g., orthogonal matching pursuit (OMP) (Chen et al., 1989; Tropp, 2004), and on the other hand Frank-Wolfe (FW)-type algorithms (Frank & Wolfe, 1956). Both operate in the setting of minimizing an objective over (combinations of) a given set of atoms, or dictionary elements.

The two classes of methods have very strong similarities, in the sense that they in each iteration rely on the very same subroutine, namely selecting the atom of largest inner product with the negative gradient, i.e., what we call the *linear minimization oracle* (LMO). Yet, the main difference is that MP methods optimize over the linear span of the atoms, while FW methods optimize over their convex hull.

Despite the vast literature on MP-type methods which typically gives recovery guarantees for sparse signals, surprisingly little is known about MP algorithms in terms of optimization, i.e., how many iterations are needed to reach a defined target accuracy. In particular, we are not aware of any general-purpose explicit convergence rates, which hold for an arbitrary given set of atoms (here "explicit" means that the result must not depend on iteration-dependent quantities). Indeed, in the context of sparse recovery, convergence rates typically come as a byproduct of the recovery guarantees and hence depend on very strong assumptions (from an optimization perspective), such as incoherence or restricted isometry properties of the atom set (Tropp, 2004; Davenport & Wakin, 2010). Motivated by this line of work, (Gribonval & Vandergheynst, 2006; Temlyakov, 2013, 2014; Nguyen & Petrova, 2014) specifically target convergence rates but still rely on incoherence properties. On the other hand, FW methods are well understood from an optimization perspective, with strong explicit convergence results available for a large class of input problems, see, e.g., (Jaggi, 2013; Lacoste-Julien & Jaggi, 2015) for a recent account.

In this paper, we provide a unified view on MP and FW algorithms from an optimization perspective. Our joint understanding of both classes of algorithms has several benefits:

- We provide a clear presentation of MP methods with their FW analogues in a unified context, for the task of general convex optimization over any set of atoms from a Hilbert space. Our view also includes weight-corrective variants of MP and FW which we are able to set in direct correspondence.
- Our derived convergence rates (sub-linear for the case of smooth objective, and linear/geometric for the case of smooth and strongly convex objective) are the first explicit optimization rates for MP methods, for general atom sets, to the best of our knowledge. We set the new rates and their complexity constants in context with existing FW rates. Our linear convergence rate of MP is expressed in terms of a new quantity called the *minimal intrinsic directional width* of the atoms.
- We allow for approximate subroutines in all proposed MP and FW variants, that is the use of an approximate linear oracle (LMO). The level of approximation quality is reflected in all convergence rates.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

- Additionally, we give affine invariant extensions of the MP and FW algorithm variants, as well as convergence rates in terms of affine invariant quantities. That is, the algorithms and rates will be invariant under affine transformations and re-parameterizations of the optimization domain (a property which was known for Newton's method and FW methods, but is novel in the MP context).

Motivation. The setting of optimization over linear or convex combinations of atoms has served as a very useful template in many applications, since the choice of the atom set conveniently allows to encode structure desired for the use case. Apart from many applications based on sparse vectors, the use of rank-1 atoms gives rise to structured matrix and tensor factorizations, see, e.g., (Wang et al., 2014; Yang et al., 2015; Yao & Kwok, 2016; Guo et al., 2017). For example, minimizing the Bregman Divergence over a set of structured rank-1 matrices yields an exponential family structured PCA (Gunasekar et al., 2014). Other applications include multilinear multitask learning (Romera-Paredes et al., 2013), matrix completion and image denoising (Tibshirani, 2015).

Complexity Constants and Coherence. While our sublinear convergence rates for MP and FW only depend on bounded norm of the iterates and on the diameter of the atom set, the linear rates also depend on our notion of minimal intrinsic directional width. In contrast to the notion of (cumulative) coherence commonly used in the context of MP and OMP (Gribonval & Vandergheynst, 2006), our width complexity notion is more robust, e.g., w.r.t. addition of new atoms, and leading to provably better bounds than coherence. Furthermore, our linear rates are significantly easier to interpret than the linear rates obtained for FW algorithm variants in (Lacoste-Julien & Jaggi, 2015) which rely on a complex geometric quantity called pyramidal width. Finally, we elucidate the relationship between FW algorithms and our proposed generalized MP variants, by showing that the iterates of FW converge to those of MP as $O(1/\alpha)$, if the atom set of FW is scaled by a growing factor α.

We note that a few recent works (Shalev-Shwartz et al., 2010; Temlyakov, 2013, 2014, 2015; Nguyen & Petrova, 2014; Yao & Kwok, 2016) proposed similar algorithms extending MPs to general smooth objective functions, although with less general convergence rates and without studying the algorithms in the larger context of MP and FW. The relation to these works is discussed in detail in Section 8.

Notation. Let $[d]$ be the set $\{1, 2, \dots, d\}$. Given a non-empty subset $\mathcal{A}$ of some vector space, let $\operatorname{conv}(\mathcal{A})$ be the convex hull of the set $\mathcal{A}$, and let $\operatorname{lin}(\mathcal{A})$ denote the linear span of the elements in $\mathcal{A}$. Given a closed set $\mathcal{A}$ we call its diameter $\operatorname{diam}(\mathcal{A}) = \max_{\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{A}} \|\mathbf{z}_1 - \mathbf{z}_2\|$ and its radius $\operatorname{radius}(\mathcal{A}) = \max_{\mathbf{z} \in \mathcal{A}} \|\mathbf{z}\|$. Note that for convex hulls of finite atom sets $\mathcal{A}$ we have $\operatorname{diam}(\operatorname{conv}(\mathcal{A})) = \operatorname{diam}(\mathcal{A})$, i.e., the diameter is attained at two vertices (Ziegler, 1995). $\|\mathbf{x}\|_{\mathcal{A}} := \inf\{c > 0 \colon \mathbf{x} \in c \cdot \operatorname{conv}(\mathcal{A})\}$ is the atomic norm of $\mathbf{x}$ over a set $\mathcal{A}$ (also known as the gauge function of $\operatorname{conv}(\mathcal{A})$). We call a subset $\mathcal{A}$ of a Hilbert space symmetric if $\mathcal{A} = -\mathcal{A}$. We write $\operatorname{clip}_{[0,1]}(s) := \max\{0, \min\{1, s\}\}$.

2 Matching Pursuit and Frank-Wolfe

We start by reviewing the MP (Mallat & Zhang, 1993), the OMP (Chen et al., 1989; Tropp, 2004), and the FW algorithm (Frank & Wolfe, 1956; Jaggi, 2013) in Hilbert spaces. The setting considered throughout this paper is the following. 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 non-empty bounded 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*.

In each iteration, both the MP/OMP and the FW algorithm query a so-called linear minimization oracle (LMO) which solves the optimization problem

$$\operatorname{LMO}_{\mathcal{D}}(\mathbf{y}) := \arg\min_{\mathbf{z} \in \mathcal{D}} \langle \mathbf{y}, \mathbf{z} \rangle \tag{1}$$

for given $\mathbf{y} \in \mathcal{H}$ and $\mathcal{D} \subset \mathcal{H}$. As computing an exact solution (1), depending on $\mathcal{D}$, is often hard in practice, it is desirable to rely on an *approximate* LMO that returns an approximate minimizer of (1). Different notions of approximate LMOs are discussed in more detail in Section 3.4.

MP and OMP, presented in Algorithm 1, aim at approximating a target point $\mathbf{y} \in \mathcal{H}$ as well as possible in the least-squares sense using no more than T atoms form a possibly countable or finite dictionary $\mathcal{A} \subset \mathcal{H}$.

Algorithm 1 (Orthogonal) Matching Pursuit

init $\mathbf{x}_0 \in \operatorname{lin}(\mathcal{A})$ $\mathcal{S} = \{\mathbf{x}_0\}$
for $t = 0 \dots T$
  Find $\mathbf{z}_t := (\text{Approx-})\operatorname{LMO}_{\mathcal{A} \cup -\mathcal{A}}(-\mathbf{y} + \mathbf{x}_t)$
  $\mathcal{S} = \mathcal{S} \cup \mathbf{z}_t$
  Update MP: $\mathbf{x}_{t+1} := \arg\min_{\mathbf{x} := \mathbf{x}_t + \gamma \mathbf{z}_t,\ \gamma \in \mathbb{R}} \|\mathbf{y} - \mathbf{x}\|^2$, or
  Update OMP: $\mathbf{x}_{t+1} := \arg\min_{\mathbf{x} \in \operatorname{lin}(\mathcal{S})} \|\mathbf{y} - \mathbf{x}\|^2$
end for

At each iteration, OMP adds a new atom to the active set $\mathcal{S}$ and computes the new iterate as the least-squares approximation of $\mathbf{y}$ in terms of the atoms in $\mathcal{S}$. As a result, the residual $\mathbf{r}_{t+1} := \mathbf{y} - \mathbf{x}_{t+1}$ is orthogonal to $\operatorname{lin}(\mathcal{S})$. This is in contrast to MP, which only minimizes the residual error $\|\mathbf{r}_{t+1}\|^2$ w.r.t. $\mathbf{z}_t$ so that $\mathbf{r}_{t+1}$ is orthogonal to $\mathbf{z}_t$, but not necessarily to all $\mathbf{z}_{t'}$, $t' \leq t - 1$. Note that MP

does not require to maintain the active set $\mathcal{S}$ as the update only relies on $\mathbf{z}_t$. Also note that in the signal processing literature MP and OMP are typically formulated using $\mathbf{z}_t := \arg\max_{\mathbf{z}\in\mathcal{A}} |\langle \mathbf{y} - \mathbf{x}_t, \mathbf{z}\rangle|$ in Line 3 of Algorithm 1 instead of $\mathbf{z}_t := \text{LMO}_{\mathcal{A}\cup-\mathcal{A}}(-\mathbf{y}+\mathbf{x}_t)$. The solution of this alternative LMO definition is equal to that of $\text{LMO}_{\mathcal{A}\cup-\mathcal{A}}$ up to the sign, so that the iterates $\mathbf{x}_t$ are identical for both definitions. Relying on $\text{LMO}_{\mathcal{A}\cup-\mathcal{A}}$ here allows to better illustrate the parallels between MP/OMP and FW.

We now turn to the FW algorithm (Frank & Wolfe, 1956; Jaggi, 2013), also referred to as conditional gradient in the literature. The FW algorithm, presented in Algorithm 2, targets the optimization problem

$$\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}), \tag{2}$$

where $\mathcal{D} \subset \mathcal{H}$ is convex and bounded. In many applications, $\mathcal{D}$ is the convex hull of a dictionary $\mathcal{A}$, i.e., $\mathcal{D} = \text{conv}(\mathcal{A})$, in which case $\text{LMO}_{\mathcal{D}}(\mathbf{y}) = \text{LMO}_{\mathcal{A}}(\mathbf{y})$.

Algorithm 2 Frank-Wolfe

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

At each iteration, the FW algorithm selects a new atom $\mathbf{z}_t$ from $\mathcal{D}$ by querying the LMO and computes the new iterate as a convex combination of $\mathbf{z}_t$ and the old iterate $\mathbf{x}_t$. As discussed in (Jaggi, 2013), the convex update can be performed either by line search (line 5 in Algorithm 2) or as a convex combination of all previously selected atoms $\mathbf{z}_{t'}$, $t' \leq t$.

The steps in Line 3 of MP (Algorithm 1) and Line 3 of FW (Algorithm 2) (finding the step direction) are identical up to symmetrization of $\mathcal{A}$. This is seen as follows. Recall that MP and OMP approximate $\mathbf{y}$ in the least-squares sense, i.e., they aim at minimizing $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$. For this choice of f we have $\nabla f(\mathbf{x}_t) = -\mathbf{y} + \boldsymbol{x}_t = -\mathbf{r}_t$, i.e., $\text{LMO}_{\mathcal{A}\cup-\mathcal{A}}(-\mathbf{r}_t) = \text{LMO}_{\mathcal{A}\cup-\mathcal{A}}(\nabla f(\mathbf{x}_t))$.

3 Greedy Algorithms in Hilbert Spaces

We present new greedy algorithms—inspired by MP, OMP, and FW—for the minimization of functions f over a convex and bounded set $\mathcal{D} \subset \mathcal{H}$, or over the linear span of a dictionary $\mathcal{A} \subset \mathcal{H}$. As MP, OMP, and FW, these algorithms alternate between querying the LMO defined in (1) and updating the current iterate $\mathbf{x}_t$. Common to all of our algorithms is that their update step minimizes an upper bound of f at $\mathbf{x}_t$, given as

$$g_{\mathbf{x}_t}(\mathbf{x}) := f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 \tag{3}$$

where L is an upper bound on the smoothness constant of f w.r.t. a chosen norm $\|.\|$. Optimizing this norm problem instead of the original f objective allows for substantial efficiency gains in the case of complicated f objective.

We note that our algorithms can be made *affine invariant*, i.e., invariant under affine transformations and reparameterizations of the domain, by simple modifications of the update steps. For simplicity of exposition, we present these algorithm versions, along with corresponding sublinear and linear convergence results later in Section 6.

3.1 Constrained Optimization

We consider constrained optimization problems of the form (2) with $\mathcal{D} := \text{conv}(\mathcal{A})$ for some dictionary $\mathcal{A} \subset \mathcal{H}$. Inspired by the fully-corrective Frank-Wolfe variant (see, e.g., (Holloway, 1974; Jaggi, 2013)) which, in each update step, re-optimizes the original objective over the convex hull of all previously selected atoms, $\text{conv}(\mathcal{S})$, we instead propose to minimize the simpler quadratic upper bound (3) over the atom selected at the current iteration (using line-search) or over $\text{conv}(\mathcal{S})$. We call this algorithm variant, presented in Algorithm 3, *norm-corrective Frank-Wolfe*.

Algorithm 3 Norm-Corrective Frank-Wolfe

1: **init** $\mathbf{x}_0 \in \text{conv}(\mathcal{A})$, and $\mathcal{S} := \{\mathbf{x}_0\}$
2: **for** $t = 0 \dots T$
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 $\mathbf{x}_{t+1} := \arg\min_{\substack{\mathbf{z}:=\mathbf{x}_t+\gamma(\mathbf{z}_t-\mathbf{x}_t)\\ \gamma\in[0,1]}} \|\mathbf{z} - \mathbf{b}\|_2^2$
Variant 1: Update $\mathbf{x}_{t+1} := \arg\min_{\mathbf{z}\in\text{conv}(\mathcal{S})} \|\mathbf{z} - \mathbf{b}\|_2^2$
7: *Optional:* Correction of some/all atoms $\mathbf{z}_{0\dots t}$
8: **end for**

The name "norm-corrective" is used to illustrate that the algorithm employs a simple squared norm surrogate function (or upper bound on f), which only depends on the smoothness constant L. This is in contrast to second-order optimization methods such as Newton's method, which rely on a non-uniform quadratic surrogate function at each iteration. Importantly, we do not need to know L (and the corresponding constant in the affine invariant algorithm versions in Section 6) exactly in any of the proposed algorithms; an upper bound is always sufficient to ensure convergence. Finding the closest point in norm can typically be performed much more efficiently than solving a general optimization problem, such as if we would minimize f over the same domain, which is what the "fully-corrective" algorithm variants require in each iteration. Approximately solving the

subproblem in Variant 1 can be done efficiently using projected gradient steps on the weights (as projection onto the simplex and L1 ball is efficient). Assuming a fixed quadratic subproblem as in Variant 1, the CoGEnT algorithm of (Rao et al., 2015) uses the same "enhancement" steps. The difference in the presentation here is that we address general f, so that the quadratic correction subproblem changes in every iteration in our case.

3.2 Optimization over the linear span of a dictionary

We now move on to optimization over linear span of a dictionary $\mathcal{A} \subset \mathcal{H}$, i.e., we consider problems of the form

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

To solve (4), we present the Norm-Corrective Generalized Matching Pursuit (GMP) in Algorithm 4 which is again based on the quadratic upper bound (3) and can be seen as an extension of MP and OMP to smooth functions f.

Algorithm 4 Norm-Corrective Generalized Matching Pursuit

1: **init** $\mathbf{x}_0 \in \text{lin}(\mathcal{A})$, and $\mathcal{S} := \{\mathbf{x}_0\}$
2: **for** $t = 0 \dots T$
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 $\mathbf{x}_{t+1} := \underset{\substack{\mathbf{z}:=\mathbf{x}_t+\gamma\mathbf{z}_t \\ \gamma\in\mathbb{R}}}{\arg\min} \|\mathbf{z}-\mathbf{b}\|_2^2$
Variant 1: Update $\mathbf{x}_{t+1} := \underset{\mathbf{z}\in\text{lin}(\mathcal{S})}{\arg\min} \|\mathbf{z}-\mathbf{b}\|_2^2$
7: *Optional:* Correction of some/all atoms $\mathbf{z}_{0\dots t}$
8: **end for**

Here, the updates in line 6 are again either over the most recently selected atom (Variant 0) or over all perviously selected atoms (Variant 1). However, the optimization is unconstrained as opposed to norm-corrective FW. Note that the update step in line 6 of Algorithm 4 Variant 0 (line-search) has the closed-form solution $\gamma = -\frac{\langle \mathbf{x}_t - \mathbf{b}, \mathbf{z}_t \rangle}{\|\mathbf{z}_t\|^2}$.

It is important to stress the fact that for Variant 1, at the end of iteration t, $\nabla f(\boldsymbol{x}_{t+1})$ is not always orthogonal to $\text{lin}(\mathcal{S})$ as it is the case for OMP (see the discussion in Section 2).

This difference is rooted in the fact that the OMP residual $\mathbf{r}_{t+1} = \mathbf{y} - \mathbf{x}_{t+1}$ (i.e., the gradient at iteration $t+1$, $\nabla f(\boldsymbol{x}_{t+1})$) can be obtained by projecting the $\mathbf{r}_t$ (i.e., the gradient at iteration t, $\nabla f(\boldsymbol{x}_t)$) onto the orthogonal complement of $\hat{\mathbf{z}}_t$, where $\hat{\mathbf{z}}_t$ is obtained by orthogonalizing $\mathbf{z}_t$ w.r.t. $\mathbf{z}_{t'}$, $t' \leq t-1$. In other words, the OMP update step maintains orthogonality of the gradient w.r.t. the atoms selected in all previous iterations, which is not the case for general smooth functions f due to varying curvature.

3.3 Discussion

The update step in line 6 in Algorithms 3 and 4 is very similar to a projected gradient descent step with a step-size of $1/L$ (i.e., $\mathbf{b} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$ is a gradient descent step with step size $1/L$ and the update step in line 6 is a projection of $\mathbf{b}$). However, the crucial difference to projected gradient descent is that the projection step is only partial, i.e., the projection is only onto $\text{conv}(\mathcal{S})$ and $\text{lin}(\mathcal{S})$ instead of the entire constraint set $\text{conv}(\mathcal{A})$ and $\text{lin}(\mathcal{A})$ for Algorithms 3 and 4, respectively.

The total number of iterations T of Algorithms 3 and 4 controls the trade-off between approximation quality, i.e., how close $f(\mathbf{x}_T)$ is to the optimum $f(\mathbf{x}^\star)$, and the "structuredness" of the (approximate) solution $\mathbf{x}_T$. The structure is due to the fact that we only use T atoms from $\mathcal{A}$ and due to the structure of the atoms themselves (e.g., sparsity). A concrete example for an application of Algorithm 4 that requires such a structure is low-rank matrix factorization: Choosing for f a function measuring the approximation quality of a given matrix to a target matrix and rank-1 matrices with unit norm as atom set, T controls the rank of the solution matrix.

3.4 Approximate linear oracles and atom corrections

Recall that an exact LMO is often very costly, in particular when applied to matrix (or tensor) factorization problems, while approximate versions can be much more efficient. We now generalize all the presented Algorithms to allow for an *approximate* LMO. Different notions of such an LMO were already explored for the Frank-Wolfe framework in (Lacoste-Julien et al., 2013). Here, we focus on multiplicative errors and define two different approximate LMO s, one for Algorithm 3 and another one for Algorithm 4. We discuss their relationship in Section 7. Formally, for a given quality parameter $\delta_{\mathsf{FW}} \in (0, 1]$ and for a given direction $\mathbf{d} \in \mathcal{H}$, the approximate LMO for Algorithm 3 returns a vector $\tilde{\mathbf{z}} \in \mathcal{A}$ satisfying

$$\langle \mathbf{d}, \tilde{\mathbf{z}} - \mathbf{x}_t \rangle \leq \delta_{\mathsf{FW}} \min_{\mathbf{z}\in\mathcal{A}} \langle \mathbf{d}, \mathbf{z} - \mathbf{x}_t \rangle. \tag{5}$$

For given quality parameter $\delta_{\mathsf{MP}} \in (0, 1]$ and given direction $\mathbf{d} \in \mathcal{H}$, the approximate LMO for Algorithm 4 returns a vector $\tilde{\mathbf{z}} \in \mathcal{A}$ such that

$$\langle \mathbf{d}, \tilde{\mathbf{z}} \rangle \leq \delta_{\mathsf{MP}} \langle \mathbf{d}, \mathbf{z} \rangle, \tag{6}$$

where $\mathbf{z} = \text{LMO}_{\mathcal{A}}(\mathbf{d})$. We will often refer to the quality parameter simply as δ.

Further, as shown in line 7 of Algorithms 3 and 4, we also allow for correction of some/all atoms in the active set $\mathcal{S}$, see, e.g., (Laue, 2012; Guo et al., 2017), to obtain a better objective cost while maintaining the same (small) number of atoms.

4 Sublinear Convergence Rates

In this section we present sub-linear convergence guarantees for Algorithms 3 and 4. All proofs are deferred to the Appendix in the supplement.

Frank-Wolfe algorithm variants. We start with the convergence result for Algorithm 3, which targets optimization problems of the form (2). Let $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathrm{conv}(\mathcal{A})} f(\mathbf{x})$ be an optimal solution of (2).

Theorem 1. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set and let $f\colon \mathcal{H}\to\mathbb{R}$ be L-smooth w.r.t. a given norm $\|.\|$, over $\mathrm{conv}(\mathcal{A})$. Then, the Frank-Wolfe method (Algorithm 2), as well as Norm-Corrective Frank-Wolfe (Algorithm 3), converge for $t \geq 0$ as*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{2\left(\frac{1}{\delta}L\,\mathrm{diam}_{\|.\|}(\mathcal{A})^2 + \varepsilon_0\right)}{\delta t + 2}$$

where $\varepsilon_0 := f(\mathbf{x}_0) - f(\mathbf{x}^\star)$ is the initial error in objective, and $\delta \in (0,1]$ is the accuracy parameter of the employed approximate LMO *(Equation (5)).*

Matching pursuit algorithm variants. We now move on to Algorithm 4 which solves optimization problems over a linear span, as given in (4). We again write $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathrm{lin}(\mathcal{A})} f(\mathbf{x})$ for an optimal solution. Our rates will crucially depend on a (possibly loose) upper bound on the atomic norm of the solution and iterates: Let $\rho > 0$ s.t.

$$\rho \geq \max\left\{\|\mathbf{x}^\star\|_{\mathcal{A}}, \|\mathbf{x}_0\|_{\mathcal{A}} \ldots, \|\mathbf{x}_T\|_{\mathcal{A}}\right\}. \tag{7}$$

If the optimum is not unique, we consider $\mathbf{x}^\star$ to be one of largest atomic norm. We now present the convergence results for the Matching Pursuit algorithm variants.

Theorem 2. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded and symmetric set, and let $f\colon \mathcal{H}\to\mathbb{R}$ be L-smooth w.r.t. a given norm $\|.\|$, over $\rho\,\mathrm{conv}(\mathcal{A})$ with $\rho < \infty$. Then, Norm-Corrective Matching Pursuit (Algorithm 4), 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 $\varepsilon_0 := f(\mathbf{x}_0) - f(\mathbf{x}^\star)$ is the initial error in objective, and $\delta \in (0,1]$ is the relative accuracy of the employed approximate LMO *(6).*

The proof of Theorem 2 extends the FW convergence analysis from $\mathrm{conv}(\mathcal{A})$ to $\mathrm{lin}(\mathcal{A})$ by rescaling $\mathrm{conv}(\mathcal{A})$ so that it includes $\mathbf{x}^\star$ and $\mathbf{x}_t$ for all $t \leq T$, the reason for which the rate in Theorem 2 depends on the upper bound ρ on the atomic norm of $\mathbf{x}^\star$ and $\mathbf{x}_t$, $t \leq T$. The relationship between Norm-Corrective FW and Norm-Corrective GMP is systematically studied in Section 7.

Using well-known results from convex optimization, we can particularize Theorem 2 for $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2$ and obtain iterate-independent constants (i.e., constants independent of ρ) as follows.

Definition 3. *The effective inradius of a convex set $\mathcal{A}$, denoted by $\mathrm{inr}(\mathcal{A})$, is the radius of the largest d-dimensional Euclidean ball which can be inscribed in $\mathcal{A}$, where d is the dimension of the subspace spanned by $\mathrm{lin}(\mathcal{A})$.*

Corollary 4. *Let $\mathcal{A} \in \mathbb{R}^d$ be a finite symmetric set of atoms, or the convex hull of a finite set of atoms, and let $\tilde{\rho} \geq \max\{\|\mathbf{x}^\star\|, \|\mathbf{x}_0\|, \ldots, \|\mathbf{x}_T\|\}$, $\tilde{\rho} < \infty$. Then, under the conditions of Theorem 2, Algorithm 4 converges both with $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{2\tilde{\rho}^2\,\mathrm{diam}_{\|.\|}(\mathcal{A})^2}{\delta^2\,\mathrm{inr}(\mathrm{conv}(\mathcal{A}))^2(t+2)}$. If further $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2$, then $\tilde{\rho}$ can be replaced by $\|\mathbf{y}\|$.*

The effective inradius $\mathrm{inr}(\mathrm{conv}(\mathcal{A}))$ generally depends on the ambient space dimension d. For example, the effective inradius of the L1-ball scales as $O(\sqrt{d})$. Hence, if $\mathcal{A}$ is the L1-ball, Corollary 4 tells us that we need to take T at least on the order of d to obtain an $O(1)$ error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$.

5 Linear Convergence Rates

It is possible to obtain faster convergence rates for some classes of objective functions, still over arbitrary dictionaries. In this section, we present linear convergence rates for our generalized matching pursuit, Algorithm 4. While linear rates have recently been demonstrated for Frank-Wolfe algorithm variants for strongly convex objectives by (Lacoste-Julien & Jaggi, 2015), we are not aware of any existing *explicit* linear convergence rates for matching pursuit algorithms (see Section 8 for a discussion).

We begin our analysis by proposing a new geometric complexity measure of the atom set which we call the *minimal intrinsic directional width*. It builds upon the classic geometric width as follows:

Definition 5. *The* directional width of a set $\mathcal{A}$ *as a function of a given non-zero vector* $\mathbf{d}$ *is defined as*

$$W_{\mathcal{A}}(\mathbf{d}) := \max_{\mathbf{z}\in\mathcal{A}} \langle \tfrac{\mathbf{d}}{\|\mathbf{d}\|}, \mathbf{z}\rangle.$$

In general, the directional width can be zero depending on the choice of $\mathbf{d}$. Building upon the the concept of directional width, we are ready to define our main complexity constant, which will be crucial to our linear convergence guarantees.

Definition 6. *Given a bounded set $\mathcal{A}$, we define its* minimal intrinsic directional width *as*

$$\mathrm{mDW}(\mathcal{A}) := \min_{\substack{\mathbf{d}\in\mathrm{lin}(\mathcal{A})\\ \mathbf{d}\neq\mathbf{0}}} W_{\mathcal{A}}(\mathbf{d})\,.$$

A crucial aspect of the preceding definition is that only directions in $\mathrm{lin}(\mathcal{A})$ are allowed, hence the name intrinsic. If the minimum was not over $\boldsymbol{d} \in \mathrm{lin}(\mathcal{A})$, the width would be zero whenever $\mathcal{A}$ does not span the ambient space.

Properties. Note that $\mathrm{mDW}(\mathcal{A}) > 0$ implies that the origin is in the relative interior of $\mathrm{conv}(\mathcal{A})$ and hence the atomic is well defined $\forall \mathbf{x} \in \mathrm{lin}(\mathcal{A})$ (which ensures that $\rho < \infty$). Furthermore, note how for a fixed sequence of iterates and $\mathbf{x}^\star$ the value of ρ is a monotone decreasing function of the $\mathrm{mDW}(\mathcal{A})$. Moreover, any symmetric set satisfies the property $\mathrm{mDW}(\mathcal{A}) > 0$. For example, the L1

ball in $\mathbb{R}^d$ has $\mathrm{mDW}(\mathcal{A}) = \frac{1}{\sqrt{d}}$. The quantity $\mathrm{mDW}(\mathcal{A})$ is meaningful for both undercomplete and overcomplete, possibly continuous, atom sets, and plays a similar role as the coherence in coherence-based convergence analysis of MPs (this is discussed in more detail at the end of this section).

We now present our main linear convergence result for optimization over the linear span of atoms as defined in (4). As we will only consider strongly convex objective functions f, the optimum $\mathbf{x}^\star$ is unique here, as opposed to the general context of our sub-linear rates.

Theorem 7. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set such that $\mathrm{mDW}(\mathcal{A}) > 0$, and let the objective function $f \colon \mathcal{H} \to \mathbb{R}$ be both globally L-smooth and globally μ-strongly convex w.r.t. a given norm $\|.\|$ over $\rho \operatorname{conv}(\mathcal{A})$. Then, for $t \geq 0$, the suboptimality of the iterates of Algorithm 4 decays exponentially as*

$$\varepsilon_{t+1} \leq \left(1 - \delta^2 \frac{\mu \,\mathrm{mDW}(\mathcal{A})^2}{L \operatorname{radius}_{\|.\|}(\mathcal{A})^2}\right) \varepsilon_t,$$

where $\varepsilon_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is the suboptimality at step t, and $\delta \in (0, 1]$ is the relative accuracy parameter of the employed approximate LMO *(6).*

Even though $\mathrm{mDW}(\mathcal{A})$ can take on values larger than 1 (depending on $\mathcal{A}$) the rate in Theorem 7 is always valid as $\mathrm{mDW}(\mathcal{A}) \,/\, \operatorname{radius}_{\|.\|}(\mathcal{A}) < 1$ for any non-empty $\mathcal{A}$.

We present an additional illustrative experiment measuring the practical dependence of the convergence upon the defined $\mathrm{mDW}(\mathcal{A})$ quantity in Appendix A.

Lower Bounds. We continue by presenting a lower bound on the decay of the suboptimality of the iterates for GMP. This lower bound depends on the width $W_{\mathcal{A}}$, which shows that this quantity plays a fundamental role for the convergence of GMP. We first consider the general strongly convex and smooth functions and the particularize the result for the least-squares function $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$, which allows to compute the update in closed-form. Furthermore, we consider only the case of the exact oracle ($\delta = 1$ in Equation (6)).

Theorem 8. *Assume that $\mathbf{x}^\star \in \operatorname{lin}(\mathcal{A})$ and let $\mathbf{z}_t$ be the atom selected at iteration t by the* LMO*. Then, under the assumptions of Theorem 7, the suboptimality of the iterates of Algorithm 4 Variant 0 with an exact* LMO *($\delta = 1$) does not decay faster than*

$$\varepsilon_{t+1} \geq \left(1 - \frac{W_{\mathcal{A}}(-\nabla f(\mathbf{x}_t))^2}{\|\mathbf{z}_t\|^2} \frac{2L - \mu}{\mu}\right) \varepsilon_t$$

Note that the lower bound on the exponential decay given in Theorem 8 depends on the iteration t. We now particularize the result for the least-squares function.

Corollary 9. *Let $\mathcal{A} := \{\pm \mathbf{e}_i\} \subset \mathbb{R}^d$ be the vertices of the L1 ball. Suppose we are minimizing $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$ over the linear span of $\mathcal{A}$ with $\mathbf{y} \in \mathbb{R}^d$. Let $\mathbf{x}_0$ be the starting point of the Matching Pursuit Algorithm and assume that $\forall i \in [d]\ (\mathbf{x}_0)_i \neq y_i$. Then*

$$\varepsilon_{t+1} \geq \left(1 - \frac{1}{d - t}\right) \varepsilon_t \,.$$

This result is discussed in more detail in Appendix B.8.

Relationship between $\mathrm{mDW}(\mathcal{A})$ and cumulative coherence. It is interesting to compare the rate in Theorem 2 with the coherence-based rates from the literature, such as (Gribonval & Vandergheynst, 2006). In order to relate the two notions of cumulative coherence and directional width, we need some additional assumptions. We only consider the least-squares function in $\mathbb{R}^d$ and assume that its minimizer over $\mathbb{R}^d$ lies in the span of the atom set $\mathcal{A}$. Further, we require symmetry so that the definition of LMO given in Equation (1) is equivalent (up to the sign) to the one used for MP in (Gribonval & Vandergheynst, 2006).

Theorem 10. *Let $\mathcal{A} \subset \mathbb{R}^d$ be a symmetric set of $2n$ atoms with $\|\mathbf{s}\|_2 = 1$ for all $\mathbf{s} \in \mathcal{A}$. Let $\mathcal{B}$ be a set such that $\mathcal{A} = \mathcal{B} \cup -\mathcal{B}$ with $\mathcal{B} \cap -\mathcal{B} = \emptyset$ and $|\mathcal{B}| = n$. Then, the cumulative coherence of the set $\mathcal{B}$, defined as $\mu(\mathcal{B}, m) := \max_{\mathcal{I} \subset \mathcal{B}, |\mathcal{I}| = m} \max_{\mathbf{s}_i \in \mathcal{B} \setminus \mathcal{I}} \sum_{\mathbf{s}_j \in \mathcal{I}} |\langle \mathbf{s}_i, \mathbf{s}_j \rangle|$, $m < n$, is lower-bounded as $\mu(\mathcal{B}, n-1) \geq 1 - n \cdot \mathrm{mDW}(\mathcal{A})^2$.*

In essence, Theorem 10 shows that if the directional width is close to zero, the cumulative coherence is close to 1 with a factor that depends on n. Note that by increasing the number of atoms, both the cumulative coherence and $\mathrm{mDW}(\mathcal{A})$ grow. Recall that when the cumulative coherence is 1, according to the rate for MP in (Gribonval & Vandergheynst, 2006) there is no linear convergence. Furthermore, our rate is more robust than the one in (Gribonval & Vandergheynst, 2006) in the following sense. An adversary could add an atom to the dictionary, making the coherence 1. In contrast, adding an atom cannot make $\mathrm{mDW}(\mathcal{A}) = 0$. In addition, if the atom is added so that $\mathrm{mDW}(\mathcal{A})$ is arbitrarily small, the cumulative coherence is arbitrarily close to 1 by Theorem 10. Finally, the linear rate for MP presented in (Gribonval & Vandergheynst, 2006) assumes that the optimum can be represented exactly using m atoms. Therefore, the rate depends on $\mu(\mathcal{B}, m-1)$ while $\mathrm{mDW}(\mathcal{A})$ can be compared only to the cumulative coherence of the whole set (i.e., $\mu(\mathcal{B}, n-1)$) since it is an intrinsic property of the atom set.

6 Affine Invariant Algorithms and Rates

We now present affine invariant versions of Algorithms 3 and 4, along with sub-linear and linear convergence guarantees. An optimization method is called *affine invariant* if it is invariant under 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.

6.1 Affine Invariant Frank-Wolfe

To define an affine invariant upper bound on the objective function f, we use the affine invariant definition of the *curvature constant* from (Jaggi, 2013)

$$C_{f,\mathcal{A}} := \sup_{\substack{\mathbf{s}\in\mathcal{A},\,\mathbf{x}\in\operatorname{conv}(\mathcal{A}) \\ \gamma\in[0,1] \\ \mathbf{y}=\mathbf{x}+\gamma(\mathbf{s}-\mathbf{x})}} \frac{2}{\gamma^2} D(\mathbf{y},\mathbf{x}), \tag{8}$$

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 $\mathbf{x}$, i.e.,

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

Bounded curvature $C_{f,\mathcal{A}}$ closely corresponds to smoothness of the objective f. More precisely, if ∇f is L-Lipschitz continuous on $\operatorname{conv}(\mathcal{A})$ 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

$$C_{f,\mathcal{A}} \leq L \operatorname{diam}_{\|.\|}(\mathcal{A})^2 \,, \tag{9}$$

where $\operatorname{diam}_{\|.\|}(.)$ denotes the $\|.\|$-diameter, see (Jaggi, 2013, Lemma 7). The curvature constant $C_{f,\mathcal{A}}$ is affine invariant, does not depend on any norm. It combines the complexity of the domain $\operatorname{conv}(\mathcal{A})$ and the curvature of the objective function f into a single quantity.

We are now ready to present the affine invariant version of the Norm-Corrective Frank-Wolfe algorithm (Algorithm 3).

Algorithm 5 Affine Invariant Frank-Wolfe

Same as Algorithm 3, except,
5. $\gamma := \operatorname{clip}_{[0,1]}\left[\langle -\nabla f(\mathbf{x}_t), \mathbf{z}_t - \mathbf{x}_t\rangle / C_{f,\mathcal{A}}\right]$
6. Update $\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma(\mathbf{z}_t - \mathbf{x}_t)$

The following theorem characterizes the sub-linear convergence rate of Algorithm 5.

Theorem 11. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set and let $f\colon \mathcal{H} \to \mathbb{R}$ be a convex function with curvature $C_{f,\mathcal{A}}$ over $\mathcal{A}$ as defined in (8). Then, the Affine Invariant Frank-Wolfe algorithm (Algorithm 5) converges for $t \geq 0$ as*

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{2\left(\frac{1}{\delta}C_{f,\mathcal{A}} + \varepsilon_0\right)}{\delta t + 2}$$

where $\varepsilon_0 := f(\mathbf{x}_0) - f(\mathbf{x}^\star)$ is the initial error in objective, and $\delta \in (0,1]$ is the accuracy parameter of the employed approximate LMO *(Equation (5)).*

6.2 Affine Invariant Generalized Matching Pursuit

To design an affine invariant MP algorithm we will rely on the following slight variation of $C_{f,\mathcal{A}}$ (defined in (8)) using $\mathbf{y} = \mathbf{x} + \gamma\mathbf{s}$ instead of $\mathbf{y} = \mathbf{x} + \gamma(\mathbf{s}-\mathbf{x})$, i.e.,

$$C^{\mathsf{MP}}_{f,\mathcal{A}} := \sup_{\substack{\mathbf{s}\in\mathcal{A},\,\mathbf{x}\in\operatorname{conv}(\mathcal{A}) \\ \gamma\in[0,1] \\ \mathbf{y}=\mathbf{x}+\gamma\mathbf{s}}} \frac{2}{\gamma^2} D(\mathbf{y},\mathbf{x}). \tag{10}$$

Throughout this section, we again assume availability of a finite constant $\rho > 0$ as an upper bound of the atomic norms $\|.\|_{\mathcal{A}}$ of the optimum $\mathbf{x}^\star$, as well as the iterate sequence $(\mathbf{x}_t)_{t=0}^T$ up to the current iteration, as defined in (7). We now present the affine invariant version of the Norm-Corrective GMP algorithm (Algorithm 4, Variant 0) in Algorithm 6. The algorithm uses the bounded curvature $C^{\mathsf{MP}}_{f,\rho\mathcal{A}}$ over the rescaled set $\rho\operatorname{conv}(\mathcal{A})$, rather than $\operatorname{conv}(\mathcal{A})$.

Algorithm 6 Affine Invariant Generalized Matching Pursuit

Same as Algorithm 4 except,
5. Variant 1: $\gamma := \langle -\nabla f(\mathbf{x}_t), \rho^2\mathbf{z}_t\rangle / C^{\mathsf{MP}}_{f,\rho\mathcal{A}}$
6. Update $\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma\mathbf{z}_t$
5. Variant 2: $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\operatorname{lin}(\mathcal{S})} f(\mathbf{x})$

A sub-linear convergence guarantee for Algorithm 6 is presented in the following theorem.

Theorem 12. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded and symmetric set such that $\rho < \infty$. Then, Algorithm 6 converges for $t \geq 0$ as*

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

where $\delta \in (0,1]$ is the relative accuracy parameter of the employed approximate LMO *(6).*

Exact knowledge of $C^{\mathsf{MP}}_{f,\rho\mathcal{A}}$ is not required: The same theorem also holds if any upper bound on $C^{\mathsf{MP}}_{f,\rho\mathcal{A}}$ is used in the algorithm and resulting rate instead. Note further that the convergence guarantee in Theorem 12 is linear invariant only as the assumption of $\mathcal{A}$ being symmetric precludes affine maps involving translations.

We proceed by establishing a linear convergence guarantee for Algorithm 6. For lower-bounding the error at iteration t, we need to define an affine invariant analog of strong convexity over the requisite domain. The following positive step size quantity relates the dual certificate value of the descent direction $\mathbf{x}^\star - \mathbf{x}$ with the MP selected atom,

$$\gamma(\mathbf{x},\mathbf{x}^\star) := \frac{\langle -\nabla f(\mathbf{x}), \mathbf{x}^\star - \mathbf{x}\rangle}{\langle -\nabla f(\mathbf{x}), \mathbf{s}(\mathbf{x})\rangle}\,, \tag{11}$$

for $\mathbf{s}(\mathbf{x}) := \arg\min_{\mathbf{s}\in\mathcal{A}} \langle \nabla f(\mathbf{x}), \mathbf{s}\rangle$.

A quantity similar to (11) but using a different direction $\mathbf{s}(\mathbf{x})$ was also used by (Lacoste-Julien & Jaggi, 2015) to study linear convergence of FW variants. We now define the complexity measure $\mu^{\mathsf{MP}}_{f,\mathcal{A}}$, which serves as an affine invariant notion of strong convexity of the objective f, over the domain $\operatorname{conv}(\mathcal{A})$.

$$\mu^{\mathsf{MP}}_{f,\mathcal{A}} := \inf_{\mathbf{x}\in\operatorname{conv}(\mathcal{A})} \inf_{\substack{\mathbf{x}^\star\in\operatorname{conv}(\mathcal{A}) \\ \langle\nabla f(\mathbf{x}),\mathbf{x}^\star-\mathbf{x}\rangle<0}} \frac{2}{\gamma(\mathbf{x},\mathbf{x}^\star)^2} D(\mathbf{x}^\star,\mathbf{x}). \tag{12}$$

In the following, our results will depend on $\mu^{\mathsf{MP}}_{f,\rho\mathcal{A}}$, which is this quantity $\mu^{\mathsf{MP}}_{f,\mathcal{A}}$ taken over the scaled set $\rho\mathcal{A}$ instead of $\mathcal{A}$.

This is analogous to the smoothness parameter $C^{\mathsf{MP}}_{f,\mathcal{A}}$ as we have seen in the previous results. Theorem 13 characterizes the linear convergence of Algorithm 6.

Theorem 13. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set. Then, Algorithm 6 converges linearly as*

$$\varepsilon_{t+1} \leq \left(1 - \delta^2 \frac{\mu^{\mathsf{MP}}_{f,\rho\mathcal{A}}}{C^{\mathsf{MP}}_{f,\rho\mathcal{A}}}\right)\varepsilon_t$$

where $\varepsilon_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is the suboptimality at step t, and $\delta \in (0, 1]$ is the relative accuracy parameter of the employed approximate LMO *(6).*

Discussion: Note that the new affine invariant convergence rates in Theorems 11, 12, and 13 do imply the rates presented earlier for their norm-based algorithm counterparts in Theorems 1, 2, and 7, respectively, for any choice of norm. This follows simply establishing the relationships between $C_{f,\mathcal{A}}$ and L (see (9)) and accordingly for the strong convexity notion $\mu^{\mathsf{MP}}_{f,\mathcal{A}}$ compared to μ. For the latter, it is not hard to show that if $\mathrm{mDW}(\mathcal{A}) > 0$, $\mu^{\mathsf{MP}}_{f,\mathcal{A}} \geq \mu\,\mathrm{mDW}(\mathcal{A})^2$, see Lemma 16 in the appendix. The affine invariant convergence guarantees are therefore more general than the norm-based ones.

7 On the Relationship Between Matching Pursuit and Frank-Wolfe

The sub-linear convergence rates for MP and FW are related by the constant ρ that essentially simulates a "blown up" set in which the analysis of FW can be applied. In this section, we explore this relationship.

Let $\alpha\mathcal{A} := \{\alpha\mathbf{z} \mid \mathbf{z} \in \mathcal{A}\}$, and assume $\alpha \geq \frac{\rho}{\delta}$. We will consider Norm-Corrective FW (Algorithm 3) on the set $\alpha\mathcal{A}$ and analyze its behavior when α grows to infinity, relating the iterates of Algorithm 3 with the ones of Algorithm 4.

Theorem 14. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set and let $f\colon \mathcal{H} \to \mathbb{R}$ be a L-smooth convex function. Let $\alpha > 0$ and let us fix $t > 0$ with iterate $\mathbf{x}_t$. There exists a polynomial function $\theta(f, \mathbf{x}_t, \alpha)$ such that if $\theta(f, \mathbf{x}_t, \alpha) \leq 1$ the new iterate $\mathbf{x}^{\mathsf{FW}_\alpha}_{t+1}$ of Frank-Wolfe (Algorithm 2) using the set $\alpha\mathcal{A}$ converges to the new iterate $\mathbf{x}^{\mathsf{MP}}_{t+1}$ of Matching Pursuit (Algorithm 4) applied on the linear span of the set $\mathcal{A}$ with rate:*

$$\left\|\mathbf{x}^{\mathsf{FW}_\alpha}_{t+1} - \mathbf{x}^{\mathsf{MP}}_{t+1}\right\| \in O\left(\frac{1}{\alpha}\right).$$

In particular, when α grows to infinity, the condition $\theta(f, \mathbf{x}_t, \alpha) \leq 1$ always holds (for all steps t). Otherwise, the difference of the iterates satisfies $\left\|\mathbf{x}^{\mathsf{FW}_\alpha}_{t+1} - \mathbf{x}^{\mathsf{MP}}_{t+1}\right\| \in O(\alpha)$.

Our analysis shows that, in some sense, FW can be suitable to solve the optimization problem (4). Indeed, if we knew the atomic norm of the iterates and the optimum in advance (which is usually not the case in practice), we could just consider a large enough convex set and run FW (Algorithm 2) on $\alpha\mathcal{A}$ with $\alpha = \rho$ (ρ as defined in Section 4) for an exact oracle (this can be seen in the proof of Theorem 2).

8 Relation to Prior Generalizations of MP

Shalev-Shwartz et al. (2010), Temlyakov (2013, 2014, 2015), and Nguyen & Petrova (2014) propose and study algorithms similar to Algorithm 4—although using the objective function directly in the update step instead of a quadratic upper bound—for the optimization of smooth functions on Banach spaces. Nguyen & Petrova (2014) consider orthonormal bases as dictionaries only. The sub-linear rates derived in (Temlyakov, 2013, 2014, 2015; Nguyen & Petrova, 2014) are similar to ours, whereas the linear rates in (Temlyakov, 2013, 2014) critically rely on incoherence and approximate sparsity (of the optimal solution) assumptions. Most importantly, these linear rates only hold for a finite number of iterations that is related to the sparsity level of the solution. Note that the linear rates for (least-squares) MP and OMP in (Gribonval & Vandergheynst, 2006) hold under similar incoherence and sparsity assumptions. The linear rates for a fully-corrective GMP variant in (Shalev-Shwartz et al., 2010) holds under a (sparsity-based) restricted strong convexity assumption.

Much more general rates are known for the class of random pursuit algorithms — which are derivative-free and use random directions instead of an LMO — as shown by (Stich et al., 2013). These rates only apply to the unconstrained setting $\mathcal{A} = \mathbb{R}^d$ (so do not cover the general Hilbert-space case) and do scale with the dimension as $\Theta(d)$, whereas our rates are dimension independent (but need an LMO).

In the statistics community, very related methods are studied under the names of, e.g., forward selection and stage-wise algorithms, see (Tibshirani, 2015) for a recent overview. The stage-wise framework considers the evolution of the solution—the regularization path—as the scaling of the constraint set grows (or the corresponding regularizer weakens). Our results can help to also equip such algorithms with explicit convergence rate, at any fixed regularizer value.

To the best of our knowledge, the only prior work on greedy optimization that also relies on a quadratic upper bound of the (smooth) objective function in the update step is (Yao & Kwok, 2016). However, (Yao & Kwok, 2016) specifically targets matrix completion, considers the set of unit norm rank-one matrices as dictionary only, and obtains problem-specific (i.e., matrix-specific) and iterate-dependent (implicit) sub-linear and linear rates. Hence, the setting considered here, i.e., functions on Hilbert spaces and general dictionaries, and the linear rate depending only on geometric properties of the dictionary enjoy much higher generality.

Finally, recovery guarantees for sparse solutions of convex optimization problems using generalized MPs were proposed, e.g., in (Blumensath & Davies, 2008; Zhang, 2011).

Acknowledgments: The authors thank Zaid Harchaoui and Gunnar Rätsch for fruitful discussions. FL is supported by the Max Plank-ETH Center for Learning Systems.

References

Blumensath, Thomas and Davies, Mike E. Gradient pursuit for non-linear sparse signal modelling. In *European Signal Processing Conference*, pp. 1–5. IEEE, 2008.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A., and Willsky, Alan S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

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

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

DeVore, Ronald A. and Temlyakov, Vladimir N. Some remarks on greedy algorithms. *Adv. Comput. Math.*, 5(1):173–187, 1996.

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

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

Gunasekar, Suriya, Ravikumar, Pradeep, and Ghosh, Joydeep. Exponential family matrix completion under structural constraint. In *International Conference on Machine Learning (ICML)*, jun 2014.

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

Holloway, Charles A. An extension of the frank and Wolfe method of feasible directions. *Mathematical Programming*, 6(1):14–27, 1974.

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

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

Lacoste-Julien, Simon, Jaggi, Martin, Schmidt, Mark, and Pletscher, Patrick. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. In *ICML 2013 - Proceedings of the 30th International Conference on Machine Learning*, 2013.

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

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

Nguyen, Hao and Petrova, Guergana. Greedy strategies for convex optimization. *Calcolo*, pp. 1–18, 2014.

Rao, Nikhil, Shah, Parikshit, and Wright, Stephen J. Forward - Backward Greedy Algorithms for Atomic Norm Regularization. *IEEE Transactions on Signal Processing*, 63(21):5798–5811, 2015.

Romera-Paredes, Bernardino, Aung, Hane, Bianchi-Berthouze, Nadia, and Pontil, Massimiliano. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1444–1452, 2013.

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

Soltanolkotabi, Mahdi and Candès, Emmanuel J. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

Stich, S U, Müller, C L, and Gärtner, Bernd. Optimization of Convex Functions with Random Pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, April 2013.

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

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

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

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

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

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

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

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

Zhang, Tong. Sparse Recovery With Orthogonal Matching Pursuit Under RIP. *IEEE Transactions on Information Theory*, 57(9): 6215–6221, 2011.

Ziegler, Günter M. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer Verlag, 1995.