Optimization Convergence of Matching Pursuit Algorithms

Francesco Locatello
Max Planck / ETH Zurich
locatelf@ethz.ch

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

Rajiv Khanna
UT Austin
rajivak@utexas.edu

Martin Jaggi
EPFL
martin.jaggi@epfl.ch

***Abstract*—The matching pursuit algorithm and its variants are among the most commonly used methods for greedy optimization. In this paper, we present the first explicit convergence rates of matching pursuit methods in an optimization sense, for general sets of atoms. We present sublinear $O(1/t)$ convergence on general smooth objectives, and linear convergence on strongly convex objectives. Our algorithm variants and rates do not need any incoherence or sparsity assumptions. Direct applications of the presented algorithms are structured matrix and tensor factorization problems.**

I. Introduction

During the past decade, greedy algorithms have attracted significant attention and have led to many success stories for optimization in machine learning and signal processing (e.g. compressed sensing). Among the most prominent representatives of greedy algorithms are matching pursuit (MP) algorithms [2], including the orthogonal matching pursuit (OMP) [1], [3]. They are used to minimize an objective over the linear span of a given set of atoms, or dictionary elements. The setting of optimization over linear combinations of atoms has served as a very useful template in many applications, since the choice of the atom set conveniently allows to encode the desired structure for the application at hand. 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. [4], [5], [6].

II. Norm Corrective Matching Pursuit

Let $f\colon \mathcal{H} \to \mathbb{R}$ be convex and L-smooth function and let $\mathcal{A}$ be a bounded subset (atom set) of a Hilbert space $\mathcal{H}$. We consider problems of the form

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

To solve (1), we present the Norm-Corrective Generalized Matching Pursuit (GMP) in Algorithm 1 which is based on the quadratic upper bound $g_{\mathbf{x}_t}(\mathbf{x}) = f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$ and can be seen as an extension of MP and OMP to smooth functions f.

Algorithm 1 Norm-Corrective Generalized Matching Pursuit (GMP)

1: **init** $\mathbf{x}_0 \in \operatorname{lin}(\mathcal{A})$, and $\mathcal{S} := \{\mathbf{x}_0\}$
2: **for** $t = 0 \dots T$
3: Find $\mathbf{z}_t := (\text{Approx-})\operatorname{LMO}_{\mathcal{A}}(\nabla f(\mathbf{x}_t))$
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 $\mathbf{x}_{t+1} := \underset{\mathbf{z}\in\operatorname{lin}(\mathcal{S})}{\arg\min} \|\mathbf{z} - \mathbf{b}\|_2^2$
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). Note that the update step in line 6 of Algorithm 1 Variant 0 (line-search) has the closed-form solution $\gamma = -\frac{\langle \mathbf{x}_t - \mathbf{b}, \mathbf{z}_t\rangle}{\|\mathbf{z}_t\|^2}$. In each iteration, GMP queries a linear minimization oracle (LMO) which solves the optimization problem

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

for given $\mathbf{y} \in \mathcal{H}$ and $\mathcal{D} \subset \mathcal{H}$. As computing an exact solution (2), depending on $\mathcal{D}$, is often hard in practice, it is desirable to rely on an *approximate* LMO that returns an approximate minimizer $\tilde{\mathbf{z}}$ of (2) instead of the true solution $\mathbf{z}_t$ such that $\exists \delta \ s.t. \ \langle \mathbf{d}, \tilde{\mathbf{z}}\rangle \leq \delta\langle \mathbf{d}, \mathbf{z}\rangle$. We write $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\operatorname{lin}(\mathcal{A})} f(\mathbf{x})$ for an optimal solution of (1). Our rates crucially depend on a (possibly loose) upper bound on the atomic norm of the solution and iterates: Let $\rho > 0$ s.t.

$$\rho \geq \max\{\|\mathbf{x}^\star\|_{\mathcal{A}}, \|\mathbf{x}_0\|_{\mathcal{A}}, \dots, \|\mathbf{x}_T\|_{\mathcal{A}}\} \tag{3}$$

where $\|\mathbf{x}\|_{\mathcal{A}} := \inf\{c > 0 \colon \mathbf{x} \in c \cdot \operatorname{conv}(\mathcal{A})\}$ is the atomic norm of $\mathbf{x}$. If $\mathbf{x}^\star$ is not unique, we consider it to be of largest atomic norm.

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

Definition 2. *Given a bounded set $\mathcal{A}$, we define its* minimal intrinsic directional width *as*

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

The quantity $\operatorname{mDW}(\mathcal{A})$ is meaningful for both undercomplete and overcomplete atom sets, and plays a similar role as the coherence in coherence-based convergence analysis of MPs.

Theorem 3. *Let $\mathcal{A} \subset \mathcal{H}$ be a bounded set such that $\operatorname{mDW}(\mathcal{A}) > 0$, and let f be L-smooth and μ-strongly convex w.r.t. the given norm $\|.\|$ over $\rho\operatorname{conv}(\mathcal{A})$. Then, for $t \geq 0$, the suboptimality of the iterates of Algorithm 1 decays exponentially as*

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

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 of the employed approx. LMO.

We numerically investigate the tightness of the linear rate and illustrate the impact of the $\operatorname{mDW}(\mathcal{A})$ on the empirical rate of Algorithm 1 (Variant 0, with exact LMO) in Figure 1.

Additional Results. In a longer version of this paper, we additionally provide affine invariant variants of both the algorithms and their convergence analysis, as well as theoretical lower bounds, and a more detailed discussion of connections with Frank-Wolfe methods, and contrast our new complexity constants to existing coherence notions.

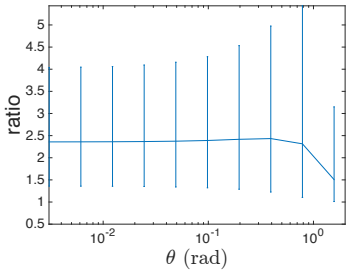

Fig. 1: Minimum, maximum and average ratio of the theoretical rate (Theorem 3) and empirical rate over 20 runs from random starting point in $\mathrm{conv}(\mathcal{A})$. We minimize the function $f(\mathbf{x}) = \|\mathbf{x}^\star - \mathbf{x}\|^2$ over the set $\mathcal{A} := \{\mathcal{A}_\theta \cup -\mathcal{A}_\theta\}_\top$ where $\mathcal{A}_\theta := \left\{ \binom{1}{0}, \binom{\cos\theta}{\sin\theta} \right\}$ with $\theta \in (0, \pi/2]$ and $\mathbf{x}^\star := (-1, 1)^\top$.

REFERENCES

[1] 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.

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

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

[4] 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.

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

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