ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747510

GENERALIZED MATCHING PURSUITS FOR THE SPARSE OPTIMIZATION OF SEPARABLE OBJECTIVES

Sebastian Ament *and* *Carla Gomes*

Cornell University
Department of Computer Science
Ithaca, NY 14850

ABSTRACT

Matching pursuit algorithms are a popular family of algorithms for compressed sensing and feature selection. Originally, Matching Pursuit (MP) was proposed as an algorithm for the least-squares objective, but has recently been generalized to arbitrary convex objectives. Here, we are concerned with the case of a general objective that is separable over observed data points, which encompasses most problems of practical interest: least-squares, logistic, and robust regression problems, and the class of generalized linear models. We propose efficient generalizations of Forward and Backward Stepwise Regression for this case, which take advantage of special structure in the Hessian matrix and are based on a locally quadratic approximation of the objective. Notably, the acquisition criterion of the generalized stepwise algorithms can be computed with the same complexity as the ones for the least-squares objective. We further propose a modification to the Newton step to avoid saddle points of non-convex objectives. Lastly, we demonstrate the generality and performance of the forward algorithm on least-squares, logistic, and robust regression problems, for which it compares favorably to generalized Orthogonal Matching Pursuit (OMP) on problems with moderate to large condition numbers.

Index Terms— Matching Pursuit, Optimization, Sparsity, Feature Selection, Compressed Sensing

1. INTRODUCTION

The optimization of objective functions under sparsity constraints is an important problem in signal processing, machine learning, and statistics with applications in medicine [1, 2], engineering [3], and materials science [4]. Given a set of *atoms* $\mathcal{D} \stackrel{\text{def}}{=} \{\boldsymbol{\varphi}_i\}$, referred to as a dictionary, and its matrix representation $\boldsymbol{\Phi} = [\boldsymbol{\varphi}_1 \dots \boldsymbol{\varphi}_m] \in \mathbb{C}^{n \times m}$, we define the sparse optimization problem of an objective $f : \mathbb{R}^n \to \mathbb{R}$ as

$$\min_{\mathbf{x}} f(\boldsymbol{\Phi}\mathbf{x}) \;\text{ subject to }\; \|\mathbf{x}\|_0 \leq k. \tag{1.1}$$

Matching pursuit algorithms are a family of greedy algorithms for this problem, were first proposed for the least-squares objective, $f(\boldsymbol{\Phi}\mathbf{x}) = \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2$ where $\mathbf{y}$ is a target vector, and iteratively add atoms to an *active set* $\mathcal{A}$. For example, Matching Pursuit (MP) and Orthogonal Matching Pursuit (OMP) add the atom that has the largest inner product with the residual $\mathbf{r} = \mathbf{y} - \boldsymbol{\Phi}\mathbf{x}$ to $\mathcal{A}$ during each iteration [5, 6]. For the least-squares objective, this acquisition criterion is equivalent to the component of the gradient with the largest magnitude:

$$\arg\max_{i \notin \mathcal{A}} |\langle \boldsymbol{\varphi}_i, \mathbf{r} \rangle| = \arg\max_{i \notin \mathcal{A}} |\partial_{x_i} f(\boldsymbol{\Phi}\mathbf{x})|, \tag{1.2}$$

which provides the basis for the existing generalization of matching pursuit algorithms to general objectives [7, 8, 9], and is also referred to as the *linear minimization oracle* [10].

Forward Stepwise Regression (FR) is a related algorithm which iteratively chooses the atom that, upon its addition to $\mathcal{A}$, minimizes the least-squares residual:

$$\arg\min_{i \notin \mathcal{A}} \|\mathbf{r}_{\mathcal{A} \cup i}\|_2^2 = \arg\min_{i \notin \mathcal{A}} |\langle \boldsymbol{\varphi}_i, \mathbf{r}_{\mathcal{A}} \rangle|^2 \;/\; \|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A}}}^2, \tag{1.3}$$

where $\mathbf{r}_{\mathcal{A}} = \mathbf{R}_{\mathcal{A}}\mathbf{y}$ is the residual, $\mathbf{R}_{\mathcal{A}} = \mathbf{I} - \boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\Phi}_{\mathcal{A}}^{+}$, and $\|\boldsymbol{\varphi}_i\|_{\mathbf{R}_{\mathcal{A}}}^2 = \boldsymbol{\varphi}_i^T \mathbf{R}_{\mathcal{A}} \boldsymbol{\varphi}_i$ is an energetic norm [11]. Because of the inner product with the residual on right side of the equality in (1.3), the forward algorithm is also known as Optimized Orthogonal Matching Pursuit (OOMP) [11] and Order-recursive Matching Pursuit (ORMP) [12]. Herein, we propose an efficient generalization of the stepwise algorithm, which employs a *quadratic minimization oracle* for the important special case of the problem (1.1) for which f takes the form

$$f(\boldsymbol{\Phi}\mathbf{x}) = \sum_i g_i\big([\boldsymbol{\Phi}\mathbf{x}]_i\big) \tag{1.4}$$

and $g_i : \mathbb{R} \to \mathbb{R}$, also referred to as ridge functions [13], are twice continuously differentiable. Unless otherwise stated, we also assume that g_i is convex, though in Section 3.5, we

This research was supported by NSF awards CCF-1522054 (Expeditions in computing) and AFOSR Multidisciplinary University Research Initiatives (MURI) Program FA9550-18-1-0136, ARO award W911NF-17-1-0187 for our compute cluster, US DOE Award No. DE-SC0020383, and an award from the Toyota Research Institute.

also show how to adapt to non-convexity. While the separability assumption is stronger than in prior work on generalized matching pursuits, it captures an important structure that occurs in most practical applications, including least-squares, logistic, and robust regression problems, and the entire class of generalized linear models [14].

Our main contributions are of algorithmic nature. We propose 1) a numerically stable algorithm for the computation of support-restricted Newton steps for (1.4), 2) generalizations of the forward and backward stepwise algorithms based on quadratic approximations to the objective, 3) efficient algorithms for the computation of the stepwise acquisition criteria, and 4) provide numerical experiments highlighting the advantages of the proposed algorithms.

2. RELATED WORK

There is a large body of related methods for the solution of the problem (1.1) under the least squares objective, ranging from relaxations of the counting norm [15] and greedy algorithms [16] to probabilistic methods [17]. [18] is one of the earliest works on generalizing matching pursuits to arbitrary objectives. [19] provides sparse recovery guarantees for generalized MP based on the restricted isometry property (RIP) of the dictionary. More recent work provides a unified optimization view on generalized matching pursuits and Frank-Wolfe algorithms [10], and jointly analyzed matching pursuits and coordinate descent, yielding sublinear convergence rate for smooth convex objectives [20]. On the algorithmic side, [21] proposes Blended Matching Pursuit (BMP), a particularly efficient first-order optimization algorithm for general convex objectives, which blends support-restricted gradient steps with coordinate-wise steps. The work of [22] relates restricted strong convexity of an objective to weak submodularity and uses this to provide approximation guarantees for the forward stepwise algorithm for the subset selection problem. [23] establishes an equivalence between a stepwise regression algorithm and a coordinate-wise optimization algorithm for Sparse Bayesian Learning and [24] proves conditions under which the backward stepwise algorithm solves the subset selection problem to optimality.

3. METHODS

First, we observe that the Hessian matrix of the objective (1.4) has a special structure which we can exploit to compute Newton steps for the coefficients in the active set $\mathcal{A}$ efficiently. For ease of notation, we define the vector functions $\mathbf{g} = \begin{bmatrix} g_1 & \dots & g_n \end{bmatrix}$, $\mathbf{g}' = \begin{bmatrix} g_1' & \dots & g_n' \end{bmatrix}$, and $\mathbf{g}''$ similarly. The gradient and Hessian of (1.4) with respect to $\mathbf{x}_\mathcal{A}$ are then

$$\begin{aligned} \nabla_\mathcal{A} f &= \mathbf{\Phi}_\mathcal{A}^T \mathbf{g}', \text{ and} \\ \mathbf{H}_\mathcal{A} f &= \mathbf{\Phi}_\mathcal{A}^T \mathbf{D}_{\mathbf{g}''} \mathbf{\Phi}_\mathcal{A}, \end{aligned} \tag{3.1}$$

where $\mathbf{D}_{\mathbf{g}''}$ is the diagonal matrix with $\mathbf{g}''$ on the diagonal.

Algorithm 1: Numerically Stable Newton's Method

Data: $\mathbf{\Phi}_\mathcal{A} \in \mathbb{C}^{n \times k}$, convex $g_i \in C^2$, initial $\mathbf{x}_\mathcal{A}$
Result: Minimizer $\mathbf{x}_\mathcal{A}$

```
while not converged do
    z ← Φ_A x_A
    g̃' ← D_{g''(z)}^{-1/2} g'(z)
    Φ̃_A ← D_{g''(z)}^{1/2} Φ_A
    Q̃, R̃ ← qr(Φ̃_A)
    d ← R̃^{-1} Q̃^T g̃'
    x_A ← x_A − stepsize(x, d) d
end
```

3.1. Stable Computation of Restricted Newton Steps

In case of collinearilties in $\mathbf{\Phi}$, the Hessian might be ill-conditioned. To derive a numerically stable direction method for the computation of the Newton step, note that it can be rewritten as

$$\begin{aligned} (\mathbf{H}_\mathcal{A} f)^{-1} \nabla_\mathcal{A} f &= (\mathbf{\Phi}_\mathcal{A}^T \mathbf{D}_{\mathbf{g}''} \mathbf{\Phi}_\mathcal{A})^{-1} \mathbf{\Phi}_\mathcal{A}^T \mathbf{g}' \\ &= (\tilde{\mathbf{\Phi}}_\mathcal{A}^T \tilde{\mathbf{\Phi}}_\mathcal{A})^{-1} \tilde{\mathbf{\Phi}}_\mathcal{A}^T \tilde{\mathbf{g}}' \\ &= \tilde{\mathbf{\Phi}}_\mathcal{A}^+ \tilde{\mathbf{g}}', \end{aligned} \tag{3.2}$$

where $\tilde{\mathbf{\Phi}}_\mathcal{A} = \mathbf{D}_{\mathbf{g}''}^{1/2} \mathbf{\Phi}_\mathcal{A}$ and $\tilde{\mathbf{g}}' = \mathbf{D}_{\mathbf{g}''}^{-1/2} \mathbf{g}'$. This is the solution to a canonical least-squares problem with the overdetermined system $\tilde{\mathbf{\Phi}}$ and can be solved for in a numerically stable manner with a QR factorization of $\tilde{\mathbf{\Phi}}$. Due to the separable objective and the resulting diagonal structure of $\mathbf{D}_{\mathbf{g}''}$, $\tilde{\mathbf{\Phi}}$ can be formed in $\mathcal{O}(nk)$ operations and factorized in $\mathcal{O}(nk^2)$. Algorithm 1 shows code for an efficient and stable implementation of Newton iterations based on this observation. We employ it for the support optimization step of both OMP and FR.

3.2. Generalized Forward Regression

Stepwise regression algorithms constitute a popular class of feature selection algorithms. Instead of adding an atom to $\mathcal{A}$ based on the largest magnitude of the gradient, stepwise algorithms update the atom that leads to the largest improvement or smallest deterioration in the objective. While efficient methods exist for the least squares objective [11, 12, 24], an application of this strategy to general objectives would require the solution of $\mathcal{O}(m)$ separate regression problems during each iteration of the algorithm, which quickly becomes expensive.

Here, we generalize the stepwise algorithms by calculating the stepwise acquisition criterion using a *quadratic approximation* to the objective function. Given efficient calculations with the Hessian matrix are possible in the case of a separable objective, the generalized stepwise algorithms have the same complexity as their canonical counterparts. In par-

ticular, the quadratic approximation of f at the point $\mathbf{x}_t$ is

$$\begin{aligned} f(\mathbf{\Phi x}) - f(\mathbf{\Phi x}_t) \approx q(\boldsymbol{\delta}) &\stackrel{\text{def}}{=} \nabla[f]^T \boldsymbol{\delta} + \boldsymbol{\delta}^T \mathbf{H}[f] \boldsymbol{\delta}/2 \\ &= \mathbf{g}_t'^T \mathbf{\Phi} \boldsymbol{\delta} + \boldsymbol{\delta}^T (\mathbf{\Phi}^T \mathbf{D}_{\mathbf{g}_t''} \mathbf{\Phi}) \boldsymbol{\delta}/2, \end{aligned} \tag{3.3}$$

where $\boldsymbol{\delta} = \mathbf{x} - \mathbf{x}_t$, $\mathbf{g}_t' = \mathbf{g}(\mathbf{\Phi x}_t)$, and $\mathbf{g}_t'' = \mathbf{g}''(\mathbf{\Phi x}_t)$. Based on equation (3.3), the minimum of $q(\boldsymbol{\delta})$ is achieved for $\boldsymbol{\delta} = -\mathbf{H}^{-1}\mathbf{g}_t'$ and the corresponding predicted decrease in function value is equal to $-\mathbf{g}_t'^T \mathbf{H}[f]^{-1} \mathbf{g}_t'/2$. Restricting the support of $\boldsymbol{\delta}$ to $\mathcal{A} \cup i$, where $\mathcal{A}$ is the support of $\mathbf{x}_t$, and $i \notin \mathcal{A}$, we can use the formula for a block matrix inverse to derive the predicted decrease in value upon adding a single atom and setting all parameters of the augmented support to their optimal values under the quadratic approximation q. It can be shown that, based on equation (3.3), adding the i^{th} atom and setting the values of the augmented support to their optimal value under $q(\boldsymbol{\delta})$, leads to a predicted change in function value

$$\min_{\boldsymbol{\delta} \in \text{span}(\mathbf{\Phi}_{\mathcal{A} \cup i})} q(\boldsymbol{\delta}) = -\nabla_i^2 \;/\; \sigma_i, \tag{3.4}$$

where $\sigma_i = \mathbf{H}_{\mathcal{A}\cup i} \backslash \mathbf{H}_{\mathcal{A}}$ is the Schur complement of $\mathbf{H}_{\mathcal{A}}$ in $\mathbf{H}_{\mathcal{A}\cup i}$. Due to the particular structure of $\mathbf{H}$, $\mathbf{H}_{\mathcal{A}\cup i} \backslash \mathbf{H}_{\mathcal{A}}$ can be computed stably and efficiently using a QR factorization of a modified dictionary. Indeed, with similar reasoning as for the Newton step, we can rewrite the Schur complement as

$$\mathbf{H}_{\mathcal{A}\cup i} \backslash \mathbf{H}_{\mathcal{A}} = \|\tilde{\boldsymbol{\varphi}}_i\|_2^2 - \|\tilde{\mathbf{Q}}_{\mathcal{A}}^T \tilde{\boldsymbol{\varphi}}_i\|_2^2, \tag{3.5}$$

where $\tilde{\mathbf{Q}}_{\mathcal{A}}, \tilde{\mathbf{R}}_{\mathcal{A}} = \tilde{\mathbf{\Phi}}_{\mathcal{A}}$ and $\tilde{\boldsymbol{\varphi}}_i = \mathbf{D}_{\mathbf{g}''}^{1/2} \boldsymbol{\varphi}_i$. We can take advantage of this structure to efficiently compute the acquisition criterion for the generalized forward regression algorithm in $\mathcal{O}(nmk)$ operations, the same complexity as the efficient algorithm for the least-squares objective. This improves on the $\mathcal{O}(mnk^2)$ operations that are even required for naïvely computing $\mathcal{O}(m)$ separate least-squares regressors.

3.3. Generalized Matching Pursuits

Algorithm 2 shows the structure of a generalized forward greedy algorithm, and provides a unified view on MP, OMP, and FR. The initials in the beginning of Lines 5 to 8 indicate which algorithms execute the respective line, and shows the subtle differences between the three variants. For example, the matching pursuit heuristic is shown in Line 4, while the proposed generalized forward regression heuristic is alternatively shown in Line 5. In our implementation, the optimization of the atoms in the active set in Line 6 is carried out by the Newton algorithm sketched in Algorithm 1. We compare the three algorithms in experiments in Section 4.

3.4. Generalized Backward Regression

Notably, we can also generalize the backward stepwise elimination criterion, based on the quadratic approximation to the objective. Solving for the predicted change in the objective value upon removing an atom and setting the remaining to their optimal values under q yields:

$$\min_{\boldsymbol{\delta} \in \text{span}(\mathbf{\Phi}_{\mathcal{A}\setminus i})} q(\boldsymbol{\delta}) = |x_i|^2 / \gamma_i, \tag{3.6}$$

where $\boldsymbol{\gamma} = \text{diag}(\mathbf{H}_{\mathcal{A}}^{-1}) = \text{diag}([\tilde{\mathbf{R}}_{\mathcal{A}}^T \tilde{\mathbf{R}}_{\mathcal{A}}]^{-1})$. Minimizing (3.6) over i yields the index of the atom that is to be eliminated. As pointed out by [24] for the least-squares objective, this is similar to magnitude pruning, and differs only in the γ_i factor, which takes into account the local curvature of the objective. While the use of backward steps in conjunction with forward steps has proven effective in achieving a very high degree of sparsity [23, 25], we defer an experimental study of the generalized backward criterion to future work and instead focus on tackling non-convex objectives.

Algorithm 2: Generalized Forward Algorithm

Data: Matrix $\mathbf{\Phi} \in \mathbb{C}^{n \times m}$, desired sparsity k
Result: Support set $\mathcal{A}$, optimized coefficients $\mathbf{x}_{\mathcal{A}}$

1 $\mathcal{A} \leftarrow \emptyset$
2 **while** $|\mathcal{A}| < k$ **do**
3 $\quad \nabla \leftarrow \nabla[f](\mathbf{\Phi}_{\mathcal{A}} \mathbf{x}_{\mathcal{A}})$
4 $\quad$ *MP / OMP*: $i^* \leftarrow \arg\max_{i \notin \mathcal{A}} |\nabla_i|$
5 $\quad$ *FR*: $i^* \leftarrow \arg\max_{i \notin \mathcal{A}} |\nabla_i|^2 \;/\; \sigma_i$ (3.4)
6 $\quad$ *FR / OMP*: $\mathbf{x}_{\mathcal{A}} \leftarrow \arg\min_{\mathbf{z}} f(\mathbf{\Phi}_{\mathcal{A}\cup i^*} \mathbf{z})$
7 $\quad$ *MP*: $\mathbf{x}_{\mathcal{A}} \leftarrow \arg\min_z f(\mathbf{\Phi}_{\mathcal{A}} \mathbf{x}_{\mathcal{A}} + \boldsymbol{\varphi}_{i^*} z)$
8 $\quad \mathcal{A} \leftarrow \mathcal{A} \cup i^*$
9 **end**

3.5. Non-Convex Separable Objectives

Certain non-convex objectives are of great practical interest, like the Cauchy likelihood and more generally the Student-t distribution for robust regression. In this case, we propose a modification of the Newton step that is similar to the saddle-point free Newton method of [26], which takes the absolute value of the Hessian matrix to ensure escape of saddle points and convergence to a local minimizer. Using the same justification, we can simply take the absolute value of the diagonal matrix $\mathbf{D}_{\mathbf{g}''}$ in the expression (3.1) for the Hessian of a separable objective to calculate an optimization direction for non-convex objectives:

$$(\mathbf{\Phi}_{\mathcal{A}} \mathbf{D}_{|\mathbf{g}''|} \mathbf{\Phi}_{\mathcal{A}})^{-1} \mathbf{\Phi}_{\mathcal{A}}^T \mathbf{g}'. \tag{3.7}$$

This has the advantage of avoiding an Eigen-decomposition of the Hessian during each iteration, as would be required by the general saddle-point-free Newton method. For a detailed justification of this procedure as a trust-region method, see [26]. Note that even on a non-convex objective, the modified Newton iterations will eventually evade saddle points and

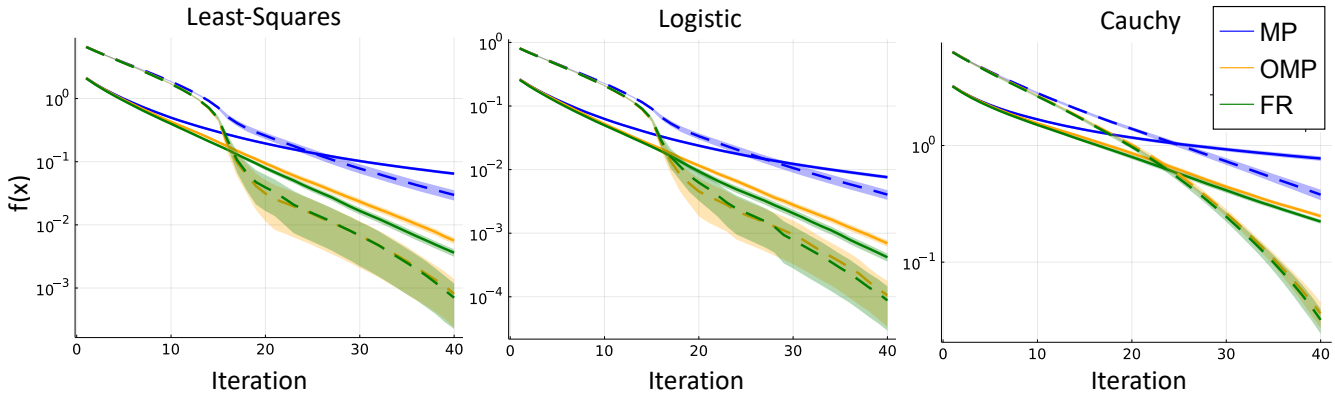

Fig. 3.1. Empirical mean and standard error of the function value of MP (blue), OMP (orange), and FR (green) during the optimization of the least-squares regression (left), logistic regression (center), and robust regression with the Cauchy likelihood (right). The dashed lines correspond to results using dictionaries with a condition number $\kappa(\mathbf{\Phi}) = 1$, while the solid lines correspond to $\kappa(\mathbf{\Phi}) = 100$. The shaded areas around the lines indicate *twice* the standard error of the corresponding means.

land in local minima in which the Hessian is positive (semi)-definite, allowing for the application of the generalized stepwise acquisition formulae (3.4) and (3.6) above.

4. EXPERIMENTS

The following experiments aim to highlight the generality of the algorithm for sparse optimization problems ranging from least-squares, logistic, to robust regression. For all of our experiments, we created dictionaries $\mathbf{\Phi}$ of size 64×128 with condition number $\kappa(\mathbf{\Phi})$ as follows. We define $\mathbf{S}_\kappa$ as a diagonal matrix with uniformly spaced values between $1/\kappa$ and 1 on the diagonal, and let $\mathbf{\Phi}_\kappa = \mathbf{U}\mathbf{S}_\kappa\mathbf{V}^*$, where $\mathbf{U}, \mathbf{V}$ are two orthonormal matrices computed by an SVD of a random matrix. The ground truth coefficients $\mathbf{x}$ were created with sparsity $k = 16$ and each non-zero element was Rademacher distributed. We let $\mathbf{b} = \mathbf{\Phi}\mathbf{x} + \boldsymbol{\epsilon}$ where the perturbation vectors $\boldsymbol{\epsilon}$ are uniformly distributed on the 10^{-2}-hypersphere. To highlight the generality of the algorithms, we report results on three different objectives:

1. Least-squares regression, where $g_i(z) = (z - b_i)^2$.
2. Logistic regression, where $g_i(z) = \log(1 + \exp(z)) - zy_i$, where $y_i = (1 + \exp(-b_i))^{-1}$.
3. A robust regression, where $g_i(z) = (1 + (z - b_i)^2)^{-1}$.

For the robust regression problem, we corrupted the target vector with outliers by adding ± 1 to three entries in addition to the Gaussian noise $\boldsymbol{\epsilon}$ that is added for all problems. Note that the Cauchy likelihood (3.) is non-convex, testing the method proposed in Section 3.5. We ran each experiment with 128 independently instantiated dictionaries and recorded the mean and standard error of the function values during the execution of Matching Pursuit (MP), Orthogonal MP (OMP), and Forward Regression (FR). See Figure 3.1 for the results.

On the instances with a higher condition number $\kappa(\mathbf{\Phi}) = 100$ (solid lines), FR improves on OMP for all three objectives *on average*, though this advantage is not guaranteed for every problem instantiation individually. For $\kappa(\mathbf{\Phi}) = 1$ (dashed lines) no difference larger than the magnitude of statistical fluctuations is present, highlighting that FR primarily has an edge for systems with collinearities. This is in line with the theoretical work of [22], which proves stronger approximation guarantees for FR than OMP for subset selection.

Notably, solving 128 separate one-dimensional logistic regression problems takes 125 ms using a Newton solver. The computation of the forward regression acquisition index via (3.4) takes 9.79 μs, and combined with the Newton optimization of the support, 967 μs. The timings were recorded on a 2021 MacBook Pro with an M1 Pro and 32 GB of RAM. Since the time for the index calculation is negligible compared to the numerical optimization, the speedup of (3.4) compared to canonical greedy algorithm is approximately m-fold in this scenario.

5. OUTLOOK

We note that the methods put forth herein are also applicable to non-separable objective, but would incur a $\mathcal{O}(n^3)$ cost for a Cholesky factorization of the Hessian of the input of f, which is required to form the modified matrix $\tilde{\mathbf{\Phi}}$ in this case. Adding support for regularization terms to the methods proposed herein could improve extrapolation performance for predictive problems and conditioning of the optimization problem. We hope to inspire further research on the theoretical properties of the novel quadratic minimization oracle that we proposed in generalizing Stepwise Regression. Due to the general applicability of the methods, we believe they could be of use to practitioners looking for general feature selection and sparse optimization algorithms.

6. REFERENCES

[1] J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen, "Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 940–947.

[2] Z. Zhang, T. Jung, S. Makeig, and B. D. Rao, "Compressed sensing for energy-efficient wireless telemonitoring of noninvasive fetal ecg via block sparse bayesian learning," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 2, pp. 300–309, 2013.

[3] Wai Lam Chan, Kriti Charan, Dharmpal Takhar, Kevin F Kelly, Richard G Baraniuk, and Daniel M Mittleman, "A single-pixel terahertz imaging system based on compressed sensing," *Applied Physics Letters*, vol. 93, no. 12, pp. 121105, 2008.

[4] Luca M Ghiringhelli, Jan Vybiral, Emre Ahmetcik, Runhai Ouyang, Sergey V Levchenko, Claudia Draxl, and Matthias Scheffler, "Learning physical descriptors for materials science by compressed sensing," *New Journal of Physics*, vol. 19, no. 2, pp. 023017, 2017.

[5] Stéphane G Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[6] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar conference on signals, systems and computers*. IEEE, 1993, pp. 40–44.

[7] Rahul Garg and Rohit Khandekar, "Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property," in *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, ICML '09, p. 337–344, Association for Computing Machinery.

[8] Xiao-Tong Yuan, Ping Li, and Tong Zhang, "Gradient hard thresholding pursuit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6027–6069, 2017.

[9] Amir Beck and Yonina C Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," *SIAM Journal on Optimization*, vol. 23, no. 3, pp. 1480–1509, 2013.

[10] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi, "A unified optimization view on generalized matching pursuit and frank-wolfe," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 860–868.

[11] Laura Rebollo-Neira and David Lowe, "Optimized orthogonal matching pursuit approach," *IEEE signal processing Letters*, vol. 9, no. 4, pp. 137–140, 2002.

[12] S. F. Cotter, R. Adler, R. D. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 146, no. 5, pp. 235–244, 1999.

[13] Hemant Tyagi and Volkan Cevher, "Learning non-parametric basis independent models from point queries via low-rank methods," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 389–412, 2014.

[14] Peter McCullagh and John A Nelder, *Generalized linear models*, Routledge, 2019.

[15] Emmanuel J Candès, Justin Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[16] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[17] Michael E Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[18] Rémi Gribonval and Pierre Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 255–261, 2005.

[19] Tong Zhang, "Sparse recovery with orthogonal matching pursuit under RIP," *IEEE transactions on information theory*, vol. 57, no. 9, pp. 6215–6221, 2011.

[20] Francesco Locatello, Anant Raj, Sai Praneeth Karimireddy, Gunnar Raetsch, Bernhard Schölkopf, Sebastian Stich, and Martin Jaggi, "On matching pursuit and coordinate descent," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds., Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 3198–3207, PMLR.

[21] Cyrille Combettes and Sebastian Pokutta, "Blended matching pursuit," in *Advances in Neural Information Processing Systems*, 2019, pp. 2042–2052.

[22] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, Sahand Negahban, et al., "Restricted strong convexity implies weak submodularity," *The Annals of Statistics*, vol. 46, no. 6B, pp. 3539–3568, 2018.

[23] Sebastian Ament and Carla Gomes, "Sparse bayesian learning via stepwise regression," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 264–274.

[24] Sebastian Ament and Carla Gomes, "On the optimality of backward regression: Sparse recovery and subset selection," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5599–5603.

[25] Tong Zhang, "Adaptive forward-backward greedy algorithm for sparse learning with linear models," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., pp. 1921–1928. Curran Associates, Inc., 2009.

[26] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," *arXiv preprint arXiv:1406.2572*, 2014.