Conditional Gradient with Enhancement and Truncation for Atomic-Norm Regularization

Nikhil Rao[†] **Parikshit Shah**[⋆] **Stephen Wright**[#]
[†]Department of Electrical and Computer Engineering
[#] Department of Computer Sciences
University of Wisconsin - Madison
[⋆] Wisconsin Institutes for Discovery

Abstract

In many applications in signal and image processing, communications, system identification, and elsewhere, one aims to recover a signal that has a simple representation in a given basis or frame. Key devices for obtaining such representations are objects called atoms, and functions called atomic norms. These concepts unify the idea of simple representations across several known applications, and motivate extensions to new problem classes of interest. In important special cases, fast and efficient algorithms are available to solve the reconstruction problems, but an approach that works well for general atomic-norm paradigm has not been forthcoming to date. In this paper, we combine a greedy selection scheme based on the conditional gradient approach with backward steps that reduce the size of the basis. Our scheme achieves the same convergence rate as the forward greedy scheme alone, provided that backward steps are taken only when they do not increase the objective too much.

1 INTRODUCTION

The problem of selecting simple models from data in a tractable way (via convex optimization, for example) is widespread in applications in communications, machine learning, image processing, genetics, and other fields. The notion of "simplicity" varies across applications. In signal and image processing, we often wish the vector of coefficients for the selected basis to be sparse. In matrix-completion problems arising in recommendation systems, we seek *low-rank* matrices.

A common conceptual framework for the notion of simplicity of representations has been proposed in [1]. Here the object $\boldsymbol{x}$[1] to be recovered is assumed to be a conic combination of a modest number of atoms $\boldsymbol{a}$, which form the basic building blocks of signals of interest and which are drawn from an atomic set $\mathcal{A}$. We seek a subset $\mathcal{A}_t$ and scalar coefficients $c_{\boldsymbol{a}}$ for $\boldsymbol{a} \in \mathcal{A}_t$, such that

$$\boldsymbol{x} = A_t c := \sum_{\boldsymbol{a} \in \mathcal{A}_t} c_{\boldsymbol{a}} \boldsymbol{a}, \quad \text{with } c_{\boldsymbol{a}} \geq 0 \text{ for all } \boldsymbol{a} \in \mathcal{A}_t. \tag{1}$$

(Here A_t denotes a linear operator from $\mathbb{R}^{|\mathcal{A}_t|}$ to the space occupied by $\boldsymbol{x}$.) We write $\boldsymbol{x} \in \text{co}(\mathcal{A}_t, \tau)$ for some given $\tau \geq 0$, if there is a representation of the form (1) such that $\sum_{\boldsymbol{a} \in \mathcal{A}_t} c_{\boldsymbol{a}} \leq \tau$.

Given a vector $\boldsymbol{x} \in \mathbb{R}^p$ and an atomic set, the *atomic norm* is defined in [1] as follows:

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \inf \left\{ \sum_{\boldsymbol{a} \in \mathcal{A}} c_a : \boldsymbol{x} = \sum_{\boldsymbol{a} \in \mathcal{A}} c_a \boldsymbol{a}, \quad c_a \geq 0 \ \ \forall \boldsymbol{a} \in \mathcal{A} \right\}. \tag{2}$$

[1]We use boldface letters $\boldsymbol{x}, \boldsymbol{y}$ etc. to denote variables in the problem.

By carefully choosing the atomic set, one can model signals that arise in a wide variety of applications [1]. We consider in this paper problems of the form

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 \text{ subject to } \|\boldsymbol{x}\|_{\mathcal{A}} \leq \tau. \tag{3}$$

Efficient algorithms have been devised for special cases of the problem (3) such as ℓ_1-constrained least squares [2–4] and nuclear-norm-constrained least squares [5–7]. For the general form (3), [8] proposed a greedy method based on the conditional gradient algorithm, often known as "Frank-Wolfe" [9]. Conditional gradient (CG) has enjoyed renwewed popularity in big data applications, due to its simplicity and global convergence properties [8, 10–12]. This approach typically adds one atom to the basis at each iteration. A drawback is that they tend to take too many iterations, thus have too many atoms in the basis at termination. Our algorithm modifies this basic procedure by allowing atoms to be purged from the current basis, and allowing the current basis to be modified and possibly reduced in size. We refer to these modifications collectively as "backward steps," and call our algorithm Enhanced Conditional Gradient (ECG). As our experiments demonstrate, ECG tends to have sparser and better solutions than CG, without sacrificing theoretical performance guarantees.

By modifying the analysis of conditional gradient [8] for backward steps, we show sublinear convergence at a $1/T$ rate for our approach. When a strict Slater-type condition holds, a *linear* rate of convergence can be proved by adapting the arguments in [12]. We can show that the algorithm is fairly robust, and even when it is only possible to select approximate optimal atoms, the algorithm displays good practical performance.

2 ALGORITHM

Algorithm 1 CoGEnT: Conditional Gradient with Enhancement and Truncation

1: **Input:** Characterization of $\mathcal{A}$, bound τ, acceptance threshold $0 < \eta < 1$;
2: **Initialize** $\boldsymbol{x}_0 = \tau\boldsymbol{a}_0$, $\boldsymbol{a}_0 \in \mathcal{A}$, $t \leftarrow 0$, $\mathcal{A}_t \leftarrow \{\boldsymbol{a}_0\}$;
3: **repeat**
4: $\quad\boldsymbol{a}_{t+1} \leftarrow \arg\min_{\boldsymbol{a}\in\mathcal{A}}\langle\nabla f(\boldsymbol{x}_t), \boldsymbol{a}\rangle$; {FORWARD STEP}
5: $\quad\tilde{\mathcal{A}}_{t+1} \leftarrow \mathcal{A}_t \cup \{\boldsymbol{a}_{t+1}\}$;
6: $\quad\gamma_{t+1} \leftarrow \arg\min_{\gamma\in[0,1]} f(\boldsymbol{x}_t + \gamma(\tau\boldsymbol{a}_t - \boldsymbol{x}_t))$;
7: $\quad$Choose $\tilde{c}^{t+1}$ to be any nonnegative coefficient vector such that $f(\tilde{A}_{t+1}\tilde{c}^{t+1}) \leq f(\boldsymbol{x}_t + \gamma_{t+1}(\tau\boldsymbol{a}_t - \boldsymbol{x}_t))$, and set $\tilde{\boldsymbol{x}}_{t+1} = \tilde{A}_{t+1}\tilde{c}^{t+1}$;
8: $\quad$Define acceptability threshold $F_{t+1} := \eta f(\boldsymbol{x}_t) + (1-\eta)f(\tilde{\boldsymbol{x}}_{t+1})$;
9: $\quad$Find possibly reduced basis $\mathcal{A}_{t+1}$ such that $|\mathcal{A}_{t+1}| \leq |\tilde{\mathcal{A}}_{t+1}|$, coefficients $c^{t+1} \geq 0$ with $\|c^{t+1}\|_1 \leq \tau$, and modified iterate $\boldsymbol{x}_{t+1} = A_{t+1}c^{t+1}$ satisfying $f(\boldsymbol{x}_{t+1}) \leq F_{t+1}$; {BACKWARD STEP}
10: **until convergence**

Algorithm 1 describes our approach. It tracks closely the Forward-Backward approch of [13], the main difference being that the "backward step" may be more elaborate than the simple removal of one basis element, as considered there. Furthermore, the "forward step" also can be seen as a more enhanced version of the standard conditional gradient method. Each forward step selects a new atom greedily and uses it to improve the objective. We choose the new coefficients c^{t+1} and iterate $\boldsymbol{x}_{t+1}$ to do as least as well as an optimal step from the current iterate $\boldsymbol{x}_t$ toward the new (scaled) atom $\tau\boldsymbol{a}_{t+1}$. One choice that clearly satisfies this assumption would be to optimize over the new expanded basis, as follows:

$$c^{t+1} := \arg\min_{c} f(A_{t+1}c) \text{ subject to } c \geq 0,\ \|c\|_1 \leq \tau. \tag{4}$$

We can solve this subproblem by means of a gradient projection procedure. (Projection onto the scaled simplex is an efficient operation, requiring $O(n_t \log n_t)$ operations, where $n_t = |\mathcal{A}_t|$.) If we start from the modified coefficients obtained in the calculation of γ_{t+1}, we can stop this procedure after any number of iterations and still satisfy the condition in step 7.

The backward step (step 9) aims to reduce the basis without sacrificing too much of the improvement in f gained from the latest forward step. This step can be skipped, which is equivalent to setting $\mathcal{A}_{t+1} \leftarrow \tilde{\mathcal{A}}_{t+1}$

and $c^{t+1} \leftarrow \tilde{c}^{t+1}$. Possibly the simplest nontrivial implementation of this step is to remove a single element of the basis $\tilde{\mathcal{A}}_t$, chosen so as to have the least effect on the objective (as in [13]). When f has the least-squares form defined in (3), we have

$$f(\tilde{\boldsymbol{x}}_{t+1} - c_{\boldsymbol{a}}\boldsymbol{a}) = f(\tilde{\boldsymbol{x}}_{t+1}) - c_{\boldsymbol{a}}\langle \nabla f(\tilde{\boldsymbol{x}}_{t+1}), \boldsymbol{a}\rangle + \frac{1}{2}c_{\boldsymbol{a}}^2 \|\Phi \boldsymbol{a}\|_2^2. \tag{5}$$

The quantities $\|\Phi \boldsymbol{a}\|_2^2$ can be computed efficiently and stored as soon as each atom $\boldsymbol{a}$ enters the current basis $\mathcal{A}_t$, and the quantity $\nabla f(\tilde{\boldsymbol{x}}_{t+1})$ may be needed for the next forward step, so the cost of evaluating these quantities is not excessive. Removal of an atom can be followed by reoptimization over the coefficients for the reduced basis (by again using gradient projection over the simplex) to improve the objective value. By extending the single-atom-removal procedure, we can remove more than one atom in a single backward step.

In some applications (for example, matrix completion) we cannot attain the final goal of a compact representation of the solution by adding and removing atoms from the basis. Atoms added at the start contain spurious components, which are cancelled out by atoms added at later iterations. We can thus consider a more general backward step that allows wholesale reorganization of the basis $\tilde{\mathcal{A}}_{t+1}$ to obtain a new, smaller basis $\mathcal{A}_{t+1}$ with fewer elements. In the case of atoms that are rank-one matrices, we could take the iterate $\tilde{\boldsymbol{x}}_{t+1}$ generated by the latest forward step, form a singular value decomposition, and form the new basis $\mathcal{A}_{t+1}$ and corresponding $\boldsymbol{x}_{t+1}$ from the rank-one matrices that correspond to the largest singular values. This approach would be competitive with the popular singular value thresholding (SVT) approach of [14]. One iteration of SVT requires calculation of the leading part of the singular value decomposition, which is about the same cost as our proposed backward step.

3 EXPERIMENTS

We report results on latent group lasso applications where we purge atoms in our backward steps using the quadratic expansion method (5), and matrix completion where we use the SVT-based backward step mentioned above.

3.1 Latent Group Lasso

Latent group Lasso [15] recovers signals whose support can be expressed as a union of groups. That the penalty can be expressed as an atomic norm is shown in [15, 16]. In [17], the authors use the concept to group parent-child pairs in DWT coefficients, and perform image recovery. CG and ECG approach can be viewed as a "greedy" analogue of the latent group lasso approach. CG and ECG do not require replication of variables (as was done in [17]), and hence avoids inflating othe problem dimension.Solving the greedy step (4) in this case amounts to the following operation

$$\hat{G} = \arg\max_{G \in \mathcal{G}} \| - [\nabla(f(\boldsymbol{x}_t))]_G \|, \quad [\boldsymbol{a}_{t+1}]_{\hat{G}} = -[\nabla f(\boldsymbol{x}_t))]_{\hat{G}} / \|[\nabla f(\boldsymbol{x}_t))]_{\hat{G}}\|, \quad [\boldsymbol{a}_{t+1}]_i = 0 \text{ for } i \notin \hat{G}.$$

We consider some standard one-dimensional signals [18], and aim to recover the parent child DWT coefficients modeled into groups. In each case, we considered a length 1024 signal, and obtained 300 Gaussian measurements corrupted with AWGN $\sigma = 0.01$. Each signal was scaled to lie between 0 and 1, and we restricted ourselves to 200 iterations of the algorithm. MSE results for the signals are shown in Table 1. Note that in all cases, the CG method selects 200 atoms, the same as the number of iterations for which we run our method, while ECG selects far fewer atoms to represent the signal (see final column) while producing closer fits to the ground truth.

To compare time taken and problem sizes, we considered M group sparse signals with $\lfloor \frac{M}{10} \rfloor$ groups active, and each group of size 50. The last 30 indices of each group overlap with the first 30 of the next group. We then took $n = \lceil \frac{p}{2} \rceil$ noisy ($\sigma = 0.1$) Gaussian measurements, with p being the ambient dimension. Table 2 compares the Latent Group Lasso (LGL) using replication with our method in terms of time taken. Note that certain methods to solve the replicated problem (for e.g: [2]) admit an efficient method to compute variables without explicitly replicating them, precluding the need to store a matrix of size $\mathbb{R}^{n \times \tilde{p}}$, $\tilde{p}$ being the replicated dimension. However, they do entail storing a sparse matrix of size $\mathbb{R}^{p \times \tilde{p}}$. In the (limited) simulations we performed, we noticed that the change in runtime was not significant when we used one procedure over the other. All times reported are in seconds.

Signal	MSE CoGEnT	MSE CG	#Atoms Selected
Piece Polynomial	$\mathbf{1.38 \times 10^{-4}}$	2.767×10^{-4}	44
Blocks	$\mathbf{2.126 \times 10^{-4}}$	7.593×10^{-4}	52
HeaviSine	**0.0004**	0.0005	64
Piecewise Regular	**0.0028**	0.0083	62

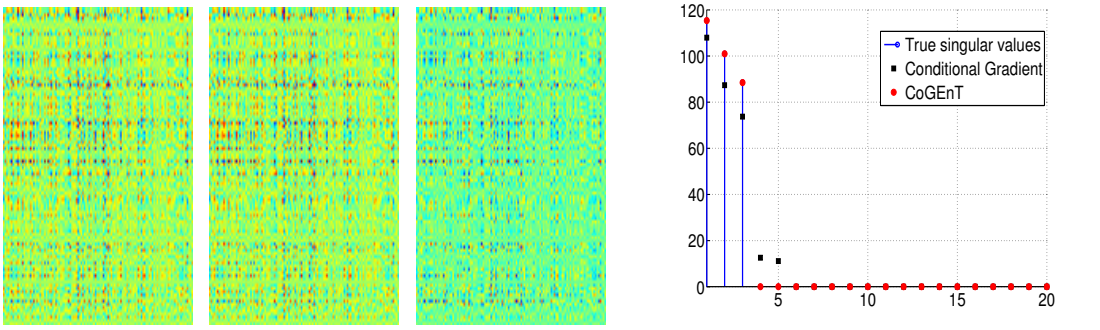
Table 1: Recovery of some 1d test signals in the presence of AWGN ($\sigma = 0.01$). After 200 iterations, ECG recovers more accurate and sparser solutions.

M	True Dimension	Replicated Dimension	time CoGEnT	time LGL
100	2030	5000	14.9	22.2
1000	20030	50000	210.9	461.6
1200	24030	60000	358.64	778.2
1500	30030	75000	574.9	1376.6
2000	40030	100000	852.02	2977

Table 2: Recovery times compared to Latent Group Lasso with Replication

3.2 Matrix Completion

We generated a 100×120 random matrix with rank $r = 3$, and observed only 30% of its entries at random. We choose the parameter τ to be the one that gave best results. A debiasing step is applied at the end, where the coefficients are chosen by solving a least-squares problem over the final basis with nonnegative coefficients but the bound involving τ removed. We see in Figure 1 that CoGEnT recovers the original matrix well; the three singular values are almost exact. CG gives a solution with five nonzero singular values.



(a) Original Matrix of rank 3 (b) Matrix recovered from CoGEnT (c) Matrix recovered from conditional gradient (d) Singular values of recovered matrices and ground truth

Figure 1: Matrix completion using CoGEnT and CG. Note that CoGEnT recovers the true matrix almost exactly and correctly identifies the rank.

4 Conclusions

We have described a method for atomic-norm-constrained minimization that enhances the conditional gradient method by allowing periodic reduction and refinement of the basis. Effectiveness of this approach in producing more compact and more accurate solutions on two problem classes has been demonstrated. We have also tested the approach on off-grid compressed sensing, standard l-1 recovery, group learning over graphs, and several other applications. Our experience with these applications will be reported in future publications.

References

[1] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. Willsky, "The convex geometry of linear inverse problems," *preprint arXiv:1012.0621v1*, 2010.

[2] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *Transactions on Signal Processing*, vol. 57, pp. 2479–2493, 2009.

[3] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, pp. 49–67, 2006.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, pp. 267–288, 1996.

[5] L. Jacob, G. Obozinski, and J. P. Vert, "Group lasso with overlap and graph lasso," *Proceedings of the 26th International Conference on machine Learning*, 2009.

[6] R. Jenatton, J. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, 2011.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the royal statistical society. Series B*, vol. 68, pp. 49–67, 2006.

[8] A. Tewari, P. Ravikumar, and I. Dhillon, "Greedy algorithms for structurally constrained high dimensional problems," *Advances in Neural Information Processing Systems*, vol. 24, pp. 882–890, 2011.

[9] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 2006.

[10] M. Jaggi, "Revisiting {Frank-Wolfe}: Projection-free sparse convex optimization," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 427–435.

[11] J. C. Dunn, "Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals," *SIAM Journal on Control and Optimization*, vol. 17, no. 2, pp. 187–211, 1979.

[12] A. Beck and M. Teboulle, "A conditional gradient method with linear rate of convergence for solving convex linear systems," *Mathematical Methods of Operations Research*, vol. 59, no. 2, pp. 235–247, 2004.

[13] T. Zhang, "Adaptive forward-backward greedy algorithm for learning sparse representations," *IEEE Transactions on Information Theory*, vol. 57, pp. 4689–4708, 2011.

[14] J.-F. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, 2010.

[15] G. Obozinski, L. Jacob, and J. Vert, "Group lasso with overlaps: The latent group lasso approach," *Preprint arXiv:1110.0413v1 [stat.ML]*, Oct 2011.

[16] N. S. Rao, B. Recht, and R. D. Nowak, "Universal measurement bounds for structured sparse signal recovery," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 942–950.

[17] N. Rao, R. Nowak, S. Wright, and N. Kingsbury, "Convex approaches to model wavelet sparsity patterns," *IEEE Conference on Image Processing*, pp. 1917–1920, 2011.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.