Dual Greedy Algorithm for Conic Optimization Problem

S. P. Sidorov, S. V. Mironov, M. G. Pleshakov^{*}

Saratov State University, Russian Federation http://www.sgu.ru

Abstract. In the paper we propose an algorithm for finding approximate sparse solutions of convex optimization problem with conic constraints and examine convergence properties of the algorithm with application to the index tracking problem and unconstrained l_1 -penalized regression.

Keywords: greedy algorithm; constrained convex optimization; conic optimization problem; index tracking problem.

1 Introduction

Greedy algorithms have been intensively studied since the 80s of the last century, and their main credit consists in obtaining constructive methods for finding the best *m*-term approximations. The main contribution to the development of greedy algorithms was made by J. Friedman [1], S. Mallat [2], J. Zhang [3], P. Huber [4], L. Jones Jones, A. Barron [6], R. DeVore, V.N. Temlyakov [7], S.V. Konyagin [8] and others.

Let X be a Banach space with norm $\|\cdot\|_X$. A set of elements \mathcal{D} from the space X is called a dictionary if each element $g \in \mathcal{D}$

- has norm bounded by one, $||g|| \leq 1$;
- the closure of span \mathcal{D} is X, i.e. $\overline{\operatorname{span}\mathcal{D}} = X$.

A dictionary \mathcal{D} is called symmetric if $-g \in \mathcal{D}$ for every $g \in \mathcal{D}$. In this paper we assume that the dictionary \mathcal{D} is symmetric.

Let E be a convex function defined on X. The problem of convex optimization is to find an approximate solution to the problem

$$E(x) \to \min_{x \in X}.$$
 (1)

In many applications it is necessary to find solutions of the problem (1), that are sparse with respect to \mathcal{D} , i.e. we are interested in solving the following problem:

$$E(x) \to \inf_{x \in \Sigma_m(\mathcal{D})},$$
 (2)

* This work was supported in the Russian Fund for Basic Research under Grant 14-01-00140. Copyright © by the paper's authors. Copying permitted for private and academic purposes.

In: A. Kononov et al. (eds.): DOOR 2016, Vladivostok, Russia, published at http://ceur-ws.org

where $\Sigma_m(\mathcal{D})$ is the set of all *m*-term polynomials with respect to \mathcal{D} :

$$\Sigma_m(\mathcal{D}) = \left\{ x \in X : x = \sum_{i=1}^m c_i g_i, g_i \in \mathcal{D} \right\}.$$
 (3)

The paper of V.N.Temlyakov [9] examines greedy algorithms for finding an approximate solution to the problem (2). It is shown that greedy algorithms with respect to the dictionary \mathcal{D} solve the problem (1) as well.

The paper [9] proposes the weak greedy Chebyshev algorithm and proves some estimates of the convergence rate of the algorithm based on the geometric properties of the function E. The development of ideas presented in [9] can be found in [10], [11].

Let Y be the *m*-dimensional Euclidean space, $Y = \mathbb{R}^m$. Let E be a convex function defined on X. Many convex constrained optimization problems can be expressed in the conic form:

$$E(x) \to \min_{A(x)+b \in K},\tag{4}$$

where $A: X \to Y$ is a linear operator, $b \in Y$ and K is a closed cone in Y.

Recall that K is said to be a cone in Y if $a_1x_1 + a_2x_2 \in K$ for any $x_1, x_2 \in K$ and $a_1, a_2 \in [0, \infty)$.

At first sight the constraint $A(x) + b \in K$ seems to be too specialized, but as it is pointed out in [12], any convex subset of $Y = \mathbb{R}^m$ may be represented in the conic form. In particular, as it was shown in [12] both Dantzig selector problem [13] and LASSO regression problem [14] can be reformulated in the form (4).

We are interested in finding an approximate solution to the problem (4). Furthermore, in many applications it is necessary to find solutions of (4) that are sparse with respect to the dictionary \mathcal{D} :

$$E(x) \to \inf_{x \in \Sigma_m(\mathcal{D}), \ A(x) + b \in K}.$$
(5)

In the paper we propose the dual weak greedy Chebyshev algorithm for finding approximate sparse solutions of convex optimization problem with conic constraint and examine convergence properties of the algorithm.

2 Dual weak greedy Chebyshev algorithm

It may turn out that the solution to the problem (4) can be inefficient from the computational point of view since projection onto the set $\{x : A(x) + b \in K\}$ may be expensive, as well as the finding of a single feasible point. For example, projection onto the set $\{x : \|y - Ax\|_2 \le \epsilon\}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, is very computational expensive, while projection onto the dual cone is trivial.

The dual problem of the problem (4) is

$$F(\lambda) \to \max_{\lambda \in K^*},$$
 (6)

where $F(\lambda)$ is the Lagrange dual function:

$$F(\lambda) := \inf_{x} L(x, \lambda) = \inf_{x} (E(x) - \langle \lambda, A(x) + b \rangle),$$

278 S. P. Sidorov, S. V. Mironov, M. G. Pleshakov

and K^* is the dual cone,

$$K^* := \{ \lambda \in Y : \langle \lambda, y \rangle \ge 0 \ \forall y \in K \}.$$

Let $A^* : \mathbb{R}^m \to \mathbb{R}^n$ denote the adjoint of the linear operator A. Let E^* be the convex conjugate of E, i.e.

$$E^*(z) = \sup_x (\langle z, x \rangle - E(x)).$$

Then the dual problem (6) can be rewritten in the following form:

$$-E^*(A^*(\lambda)) - \langle b, \lambda \rangle \to \max_{\lambda \in K^*}.$$

Recall that the following inequality holds for any optimal primal/dual pair x and λ :

$$E(x) - F(\lambda) = E(x) + E^*(A^*(\lambda)) + \langle b, \lambda \rangle \ge \langle x, A^*(\lambda) \rangle + \langle b, \lambda \rangle = \langle A(x) + b, \lambda \rangle \ge 0.$$

If the optimal solutions of the primal problem x and the dual problem λ are strictly feasible then $E(x) = F(\lambda)$ and there exist (not necessary unique) points x^*, λ^* , such that $E(x^*) = F(\lambda^*) = L(x^*, \lambda^*)$ and satisfying optimality conditions

$$A(x^*) + b \in K, \ \lambda^* \in K^*, \ \langle A(x^*) + b, \lambda^* \rangle = 0, \ A^*(\lambda^*) \in \partial E(x^*), \tag{7}$$

where ∂E is the subgradient of E.

Let $\tau := \{t_m\}_{m=1}^{\infty}, t_m \in [0,1]$, be a weakness sequence. Let λ^* satisfy (7).

We suppose that function E is Fréchet differentiable. We note that it follows from convexity of E that for any x,y

$$E(y) \ge E(x) + \langle E'(x), y - x \rangle,$$

where E'(x) denotes Fréchet differential of E at x. Denote

$$Q(x) := E(x) - \langle A(x) + b, \lambda^* \rangle.$$

Dual weak greedy Chebyshev algorithm. Let $G_0 := 0$. For each $m \ge 1$ we define G_m by induction as follows.

1. (Gradient greedy step) Find the element $\phi_m \in \mathcal{D}$ satisfying

$$\langle -Q'(G_{m-1}), \phi_m \rangle \ge t_m \sup_{g \in \mathcal{D}} \langle -Q'(G_{m-1}), g \rangle.$$

2. (Chebyshev-type step) Find real numbers c_i^* , i = 1, ..., m, such that

$$Q\left(\sum_{i=1}^{m} c_i^* \phi_i\right) = \inf_{c_i} Q\left(\sum_{i=1}^{m} c_i \phi_i\right).$$

3. Let $G_m = \sum_{i=1}^m c_i^* \phi_i$.

The gradient greedy step maximizes a certain functional determined by gradient information from the previous steps of the algorithm. The Chebyshev-type step finds the best linear combination of m approximants $\{\phi_i\}_{i=1}^m$.

Let $\Omega := \{x \in X : E(x) \le E(0)\}$ and suppose that Ω is bounded. The modulus of smoothness of function E on the bounded set Ω is defined as follows:

$$\rho(E,u) = \frac{1}{2} \sup_{x \in \Omega, \|y\|=1} |E(x+uy) + E(x-uy) - 2E(x)|.$$
(8)

Let $A_1(\mathcal{D})$ denote the closure (in X) of the convex hull of \mathcal{D} .

Using results and ideas of [9] we can prove the following proposition.

Theorem 1. Let E be a uniformly smooth convex function with modulus of smoothness $\rho(E, u) \leq \gamma u^q$, $1 < q \leq 2$. Let $V := \{x \in X : A(x) + b \in K\}$. Let $\epsilon > 0$ and element $f^{\epsilon} \in \mathcal{D}$ be such that

$$Q(f^{\epsilon}) \leq \inf_{x \in \mathcal{D}} Q(x) + \epsilon, \quad f^{\epsilon}/C(\epsilon) \in A_1(\mathcal{D}),$$

for some real number $C(\epsilon) \geq 1$. Let p = q/(q-1). Then

$$E(G_m) - \inf_{x \in \mathcal{D}_m(\mathcal{D}) \cap V} E(x) \leq \\ \leq \max\left(2\epsilon, C(q, \gamma)A(\epsilon) \left(C(E, A, b, K, q, \gamma) + \sum_{k=1}^m t_k^p\right)^{1-q}\right), \quad (9)$$

where $C(q, \gamma)$, $C(E, A, b, K, q, \gamma)$ are positive constants not depending on k.

Proof. We have $\rho(Q, u) = \rho(E, u)$. It follows from Theorem 4.2 of [9] that

$$Q(G_m) - \inf_{x \in D} Q(x) \le \max\left(2\epsilon, C(q, \gamma)A(\epsilon)\left(C(Q, q, \gamma) + \sum_{k=1}^m t_k^p\right)^{1-q}\right).$$
(10)

Then (9) follows from (10) and

$$Q(G_m) - \inf_{x \in \Omega} Q(x) = E(G_m) - \langle A(G_m) + b, \lambda^* \rangle - \inf_{x \in \Omega} (E(x) - \langle A(x) + b, \lambda^* \rangle) \ge$$

$$\ge E(G_m) - \langle A(G_m) + b, \lambda^* \rangle - \inf_{x \in \Omega} (E(x) - \langle A(G_m) + b, \lambda^* \rangle) =$$

$$= E(G_m) - \inf_{x \in \Omega} E(x) \ge E(G_m) - \inf_{x \in \Sigma_m(\mathcal{D}) \cap V} E(x).$$

3 Applications

3.1 Index tracking problem

Greedy algorithms showed an excellent performance in solution of practical problems of machine learning and optimization. This section will show the use of such techniques in

280 S. P. Sidorov, S. V. Mironov, M. G. Pleshakov

solving the index tracking problem. Index tracking is a passive financial strategy that tries to replicate the performance of a given index or benchmark. The aim of investor is to find the weights of assets in her/his portfolio that minimize the tracking error, i.e. difference between the performance of the index and the portfolio.

For any q > 0 and $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, let $||x||_q := (\sum_{i=1}^n |x_i|^q)^{1/q}$ and $||x||_0 = \lim_{q \to 0+} ||x||_q =$ (the number of non-zero elements of x). If $q \ge 1$ then $||x||_q$ denotes l_q -norm of $x \in \mathbb{R}^n$. Let n be the number of investable assets. Denote r_{ti} the return of asset i at time $t, 1 \le i \le n, 1 \le t \le k, R = (r_{ti})$ is the $k \times n$ matrix. A portfolio is defined to be a vector of weights, $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. We do not allow the portfolio changes over time and do not take into account transaction costs. We will assume that

- 1. one unit of capital is available, i.e. $x^T \mathbf{1}_n = 1$, where $\mathbf{1}_n$ denotes the vector from \mathbb{R}^n in which every component is equal to 1;
- 2. short selling is allowed, i.e. weights x_i can be negative.

Let I_t be the index return at time $t, 1 \leq t \leq k$, and $I = (I_1, \ldots, I_k)^T \in \mathbb{R}^k$. In the traditional index tracking optimization, the objective is to find a portfolio which has minimal tracking error variance, the sum of squared deviations between portfolio returns and market index returns (see e.g. [15]):

$$x^* = \arg\min\frac{1}{k} \|I - Rx\|_2^2 \quad s.t. \quad x^T \mathbf{1}_n = 1.$$
(11)

It should be noted that the standard Markovitz model is a special case of index tracking portfolio model (11) (see, for example [16], [17]). Since the problem (11) is the problem of convex optimization, it can be easily solved by Lagrange method.

In this section we examine algorithm for solving the problem (11) with the cardinality constraint:

$$x^* = \arg\min\frac{1}{k} \|I - Rx\|_2^2 \quad s.t. \quad x^T \mathbf{1}_n = 1, \ \|x\|_0 \le K,$$
(12)

where K is the limit on the number of assets in the portfolio with non-zero weights. It is supposed that K is substantially smaller than $n, K \ll n$.

Classical methods usually require convexity and relatively well-behaved objective functions and they are usually based on gradients for descent direction. Therefore, standard optimization techniques can be no longer used if we add constraints on the number of assets in portfolio. In such problems, classical optimization methods do not work efficiently and many researchers have to resort to heuristic optimization [18].

For solution to the problem (12), in this paper we propose to use greedy algorithms. The choice of the greedy algorithm for our analysis is based on the fact that greedy algorithms showed an excellent performance in the papers [19], [16] for practical problem solution, and therefore we may assume that they look promising for solving the cardinality constrained index tracking problem. On the other hand, greedy algorithms do not necessarily yield an optimal solution.

The constraint $x^T 1_n = 1$ can be rewritten in the conic form $A(x) + b \in K$ with

$$A = \begin{pmatrix} 1 & 1 & \dots & 1 \\ -1 & -1 & \dots & -1 \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^2, \ b = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \ K = \mathbb{R}^2_+ := \{ x \in \mathbb{R}^2 : \ x \ge 0 \}.$$
(13)

The following proposition follows from Theorem 1 and finds the rate of convergence of the dual greedy algorithm for the index tracking problem.

Corollary 1. Let $E(x) := \frac{1}{k} ||I - Rx||_2^2$ and $V := \{x \in \mathbb{R}^n : x^T \mathbf{1}_n = 1\}$. Let the dictionary \mathcal{D} be such that $\mathcal{D} = \{\pm e_j\}_{j=1}^n$, where $e_j \in \mathbb{R}^n$ with $e_{ji} = 1$ if i = j and $e_{ji} = 0$ otherwise. Then

$$E(G_m) - \inf_{x \in \Sigma_m(\mathcal{D}) \cap V} E(x) = O(m^{-1}),$$

where G_m is the element obtained in the step m of the dual greedy algorithm with A, b, K defined in (13).

Proof. We have $K^* = K = \mathbb{R}^2_+$. It easy to verify that $\rho(E, u) \leq \gamma u^2$ with $\gamma = \frac{1}{2k} \sup_{\|y\|=1} \|Ry\|$:

$$\begin{split} \rho(E,u) &= \\ &= \frac{1}{2k} \sup_{x \in \Omega, \|y\|=1} \left(\|I - Rx - uRy\|_2^2 + \|I - Rx + uRy\|_2^2 - 2\|I - Rx\|_2^2 \right) \leq \\ &\leq \frac{1}{2k} \sup_{x \in \Omega, \|y\|=1} (2u^2 \|Ry\|) = \frac{u^2}{2k} \sup_{\|y\|=1} \|Ry\| = \gamma u^2. \end{split}$$

Primal greedy algorithms for the index tracking problems were empirically examined in the papers [16] (in l_2 -norm) and [20] (in l_1 -norm). The analysis of heuristic algorithms for portfolio optimization problems with the cardinality constraint can be found in the work [18].

3.2 Unconstrained l₁-penalized regression

Let us consider the problem of recovering an unknown vector $x_0 \in \mathbb{R}^n$ from the data $y \in \mathbb{R}^k$ and the model

$$y = Bx_0 + z,\tag{14}$$

where B is a known $k \times n$ matrix, z is a noise. In many practical problems there are fewer observations/measurements than unknowns, i.e. $k \ll n$. Recent works have shown that accurate estimation is often possible under reasonable sparsity constraints on x_0 . One practically and theoretically effective estimator is unconstrained l_1 -penalized regression.

Unconstrained l_1 -penalized regression can be written as follows:

$$||y - Bx||_2^2 + \lambda ||x||_1 \to \min.$$
 (15)

where λ is a positive real parameter.

Homotopy method for solving the problem (15) was proposed in papers [21], [22]. The method is also known as Least Angle Regression or LARS [23].

282 S. P. Sidorov, S. V. Mironov, M. G. Pleshakov

Let us to rewrite the problem (15) in the following way:

$$\|y - Bx\|_2 \to \min, \ s.t. \ \|x\|_1 \le \epsilon, \tag{16}$$

where ϵ is a positive scalar. The real ϵ should be adjusted so that the true x_0 is feasible, at least with high probability, when the noise term z is stochastic.

The equivalent conic formulation of the problem (16) is

$$E(x) \to \|y - Bx\|_2, \ A(x) \to (Bx, 0), \ b \to (-y, \epsilon), \ K \to \mathcal{L}_1^k, \tag{17}$$

where $\mathcal{L}_{1}^{k} := \{(y, t) \in \mathbb{R}^{k+1} : \|y\|_{1} \le t\}.$

The following corollary of Theorem 1 shows that the rate of convergence of the dual greedy algorithm for the unconstrained l_1 -penalized regression (16) is m^{-1} .

Corollary 2. Let $E(x) := ||y - Bx||_2$ and $V := \{x \in \mathbb{R}^n : ||x||_1 \le \epsilon\}$. Let the dictionary \mathcal{D} consists of all columns of matrix B multiplied by ± 1 . Then

$$E(G_m) - \inf_{x \in \mathcal{D}_m(\mathcal{D}) \cap V} E(x) = O(m^{-1}),$$

where G_m is the element obtained in the step m of the dual greedy algorithm with A, b, K defined in (17).

Proof. We have
$$\rho(E, u) \leq \gamma u^2$$
 with $\gamma = \frac{1}{2k} \sup_{\|y\|=1} \|By\|$.

4 Conclusion

In this paper we have used greedy-type algorithms to solve (approximately) the problem

$$E(x) \to \inf_{x \in \Sigma_m(\mathcal{D}), A(x) + b \in K},$$

where E is a convex function defined on a Banach space $X, A : X \to Y$ is a linear operator, $b \in Y$ and K is a closed cone in Y. Since the solution to the problem (4) can be inefficient from the computational point of view, we propose the dual greedy algorithm for the conic optimization problem. Theorem 1 finds estimates on the rate of convergence for the dual greedy algorithm. We should mention that the approach used in this paper is based on the ideas and results developed in the paper [9]. Based on Theorem 1 we proved that the rate of convergence of dual greedy algorithms for index tracking problem (12) and the unconstrained l_1 -penalized regression (16) is m^{-1} .

References

- Friedman J. Greedy Function Approximation: A Gradient Boosting Machine The Annals of Statistics 2001, Vol. 29, No. 5, 1189-1232
- Davis G., Mallat S., Avellaneda M. Adaptive greedy approximation Constr. Approx. 13 (1997), pp. 57-98

- Zhang Z., Shwartz S., Wagner L., Miller W. A greedy algorithm for aligning DNA sequences J. Comput. Biol. 2000 Feb-Apr; (1-2): 203-14
- 4. Huber P. J. Projection pursuit Ann. Statist., 13(1985), pp. 435-525
- Jones L. On a conjecture of Huber concerning the convergence of projection pursuit regression Ann. Statist. 15 (1987), 880-882
- Barron A. R., Cohen A., Dahmen W., DeVore R. A. Approximation and Learning by Greedy Algorithms The Annals of Statistics 2008, Vol. 36, No. 1, 64-94
- DeVore R. A., Temlyakov V.N. Some remarks on greedy algorithms Advances in Computational Mathematics 5(1996) 173-187
- Konyagin S. V., Temlyakov V. N. A remark on greedy approximation in Banach spaces East J. Approx. 5 (1999), no. 3, 365-379
- V. N. Temlyakov, Greedy approximation in convex optimization, Constructive Approximation 41 (2) (2015) 269-296.
- H. Nguyen, G. Petrova, Greedy strategies for convex optimization, Calcolo 41 (2) (2016) 1-18.
- V. N. Temlyakov, Dictionary descent in optimization, Analysis Mathematica 42 (1) (2016) 69-89.
- S. R. Becker, E. J. Candès and M. C. Grant, Templates for convex cone problems with applications to sparse signal recovery, *Mathematical Programming Computation*, 3 (3) (2011) 165-218.
- 13. E. Candes and T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n. Annals of Statistics **35** (2007) 2313-2351.
- Tibshirani R. Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society (Series B) 58 (1996) 267-288.
- Roll R. A Mean/Variance Analysis of Tracking Error. The Journal of Portfolio Management, 18(4):13-22, 1992.
- Takeda A., Niranjan M., ya Gotoh J., Kawahara Y. Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. *Computational Management Science*, 10(1):21-49, 2013.
- Brodie J., Daubechies I., De Mol Ch., Giannone D., Loris I. Sparse and stable Markowitz portfolios. Proc. of the National Academy of Sciences of the USA, 106(30):12267-12272, 2009.
- Beasley J. E., Meade N., Chang T.-J. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3):621-643, 2003.
- Das A., Kempe D. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th Int. Conf. on Machine Learning (ICML-11)*, pages 1057–1064, New York, NY, USA, June 2011. ACM.
- Sidorov S. P., Faizliev A. R., Khomchenko A. A. Algorithms for l₁-Norm Minimization of Index Tracking Error and Their Performance. Int. J. of Mathematics in Operational Research, 2016, to appear
- Osborne M. R., Presnell B., and Turlach B. A. A New Approach to Variable Selection in Least Squares Problems. IMA Journal of Numerical Analysis, 20(3):389-403, 2000.
- Osborne M. R., Presnell B., and Turlach B. A. On the LASSO and Its Dual. Journal of Computational and Graphical Statistics, 9(2):319-337, 2000.
- Efron B., HastieT., Johnstone I., Tibshirani R. Least Angle Regression. Annals of Statistics, 32(2):407-499, 2004.