Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Convex Matching Pursuit for Large-Scale Sparse Coding and Subset Selection

Mingkui Tan † and **Ivor W. Tsang** † and **Li Wang** ‡ and **Xinming Zhang** †
†Nanyang Technological University, Singapore
‡University of California, San Diego, USA
†{*tanm0097, IvorTsang, Zxinming*}*@ntu.edu.sg*, ‡*liw022@ucsd.edu*

Abstract

In this paper, a new convex matching pursuit scheme is proposed for tackling large-scale sparse coding and subset selection problems. In contrast with current matching pursuit algorithms such as subspace pursuit (SP), the proposed algorithm has a convex formulation and guarantees that the objective value can be monotonically decreased. Moreover, theoretical analysis and experimental results show that the proposed method achieves better scalability while maintaining similar or better decoding ability compared with state-of-the-art methods on large-scale problems.

Introduction

Sparse coding has been a fundamental task in many applications such as compressive sensing (CS) (Candès and Wakin 2008), image classification (Wright et al. 2009; Gao et al. 2010) and statistical signal processing (Davenport et al. 2010; Blumensath 2011). Given a design matrix $\mathbf{D} \in \mathbb{R}^{n\times m}$ and a noisy measurement $\mathbf{y} = \mathbf{D}\mathbf{x} + \boldsymbol{\xi} \in \mathbb{R}^n$ from a sparse signal $\mathbf{x}$, where $\boldsymbol{\xi} \in \mathbb{R}^n$ is an additive noise and $\|\mathbf{x}\|_0 \leq \kappa \ll m$, in CS, sparse coding recovers $\mathbf{x}$ via solving the following ℓ_0-norm constrained inverse problem:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2 : \;\; s.t. \;\; \|\mathbf{x}\|_0 \leq \kappa. \qquad (1)$$

Moreover, suppose that $\mathbf{D}$ is a dataset with m features and $\mathbf{y}$ is the output response vector, the problem in (1) becomes a feature selection task, which has been widely used in many data mining applications such as bio-informatics.

In most of the aforementioned applications, the number of measurement n is much smaller than m, which makes (1) an ill-conditioned problem. Furthermore, it is NP-hard due to the ℓ_0-norm constraint. In the past decade, many efficient algorithms have been developed to solve this problem under some restricted condition, which is usually expressed in terms of the restricted isometry property (RIP) (Candès and Tao 2005). A matrix $\mathbf{D}$ is said to satisfy the RIP of order κ if there is some $\sigma_\kappa \in [0, 1)$, for all $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq \kappa$, then $(1 - \sigma_\kappa)\|\mathbf{x}\|^2 \leq \|\mathbf{D}\mathbf{x}\|^2 \leq (1 + \sigma_\kappa)\|\mathbf{x}\|^2$.

Current sparse coding methods can be categorized into two groups, namely, matching pursuit (MP) (or greedy pursuit) algorithms which directly solve (1) and ℓ_1 relaxed algorithms which solve a ℓ_1 convex relaxation (Zibulevsky and Elad 2010; Yang et al. 2010). The basic scheme of MP algorithms is to iteratively identify the most possible supports of $\mathbf{x}$. Among these algorithms, orthogonal matching pursuit (OMP) algorithm is one of the most well-known matching pursuit algorithms (Tropp and Gilbert 2007). In OMP method, at each iteration, only one support is detected via $\arg\max_i |c_i|$ with $\mathbf{c} = \mathbf{D}'\mathbf{r}$ and is added to a support set $\mathcal{S}$, where $\mathbf{r} = \mathbf{y} - \mathbf{D}\mathbf{x}$ is the residue. Then an orthogonal projection is performed by solving $\min_{\mathbf{x}_\mathcal{S}} \|\mathbf{y} - \mathbf{D}_\mathcal{S}\mathbf{x}_\mathcal{S}\|^2$, where $\mathbf{D}_\mathcal{S}$ denotes the sub-design matrix with atoms selected by $\mathcal{S}$ and $\mathbf{x}_\mathcal{S}$ is the corresponding regressors. Finally, an update of $\mathbf{r}$ is given by $\mathbf{r} = \mathbf{y} - \mathbf{D}_\mathcal{S}\mathbf{x}_\mathcal{S}$. Recent studies reveal that, under the RIP condition, with $O(\kappa \log(m))$ measurements, OMP can uniformly recover the κ-sparse signals but may need more iterations (Zhang 2011). Regarding this drawback, many improved MP variants have been proposed. Typical methods include compressive sampling matching pursuit (CoSaMP) (Needell and Tropp 2009), subspace pursuit (SP) (Dai and Milenkovic 2009), accelerated iterative hard thresholding (AIHT) (Blumensath 2011), orthogonal matching pursuit with replacement (OMPR) (Jain, Tewari, and Dhillon 2011) and so on. The best recovery condition of the above algorithms so far has been shown in (Jain, Tewari, and Dhillon 2011; Giryes and Elad 2012). Typically, CoSaMP, SP, AIHT and OMPR can recover κ-sparse signal provided with that $\sigma_{4\kappa} < 0.35$, $\sigma_{3\kappa} < 0.35$, $\sigma_{3\kappa} < 1/\sqrt{32}$ and $\sigma_{2\kappa} < 0.499$, respectively.

Another type of sparse coding algorithms is based on the following ℓ_1 convex relaxation of (1):

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \; : \; \mathbf{y} = \mathbf{D}\mathbf{x} + \boldsymbol{\xi} \;\; \text{or} \;\; \min_{\mathbf{x}} \rho\|\mathbf{x}\|_1 + \|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2, \qquad (2)$$

where ρ is a trade-off parameter. Recent years have witnessed a fast development on ℓ_1 sparse coding methods (Zibulevsky and Elad 2010; Yang et al. 2010). Among these methods, alternating directions method (ADM) or augmented Lagrange multiplier (ALM) has shown fast convergence in solving the first formulation in (2); while fast iterative shrinkage-threshold algorithm (FISTA) with continuation shows the most promising performance on solving the latter one of (2). A more detailed review of ℓ_1 sparse coding methods can be found in (Zibulevsky and Elad 2010) and references therein. Recent studies discovered that if $\sigma_\kappa \leq 0.307$,

ℓ_1 methods can successfully decode all the κ-sparse signal (Cai, Wang, and Xu 2010). However, it has been shown that ℓ_1 methods are too expensive on large scale problems (Jain, Tewari, and Dhillon 2011).

In this paper, we propose a convex matching pursuit (CMP) scheme for sparse coding. The core contributions of CMP are listed in the following: (1) We introduce a variation norm such that CMP is formulated as a convex optimization problem and hence guarantees its convergence. (2) Via a matching pursuit strategy, CMP shows the best scalability when dealing with large-scale problems. (3) Last but not least, CMP can successfully recover any κ-sparse signals if the design matrix $\mathbf{D}$ satisfies RIP with $\sigma_\kappa \leq 0.307$.

Notations and Preliminaries

In the sequel, we denote the transpose of vector/matrix by the superscript $'$, $\mathbf{0}$ as a vector with all entries equal to one, and $\mathrm{diag}(\mathbf{v})$ as a diagonal matrix with diagonal entries equal to $\mathbf{v}$. We also denote $\|\mathbf{v}\|_p$ and $\|\mathbf{v}\|$ as the ℓ_p-norm and ℓ_2-norm of a vector $\mathbf{v}$, respectively. Furthermore, we denote $\mathbf{v} \succeq \alpha$ if $v_i \geq \alpha, \forall i$ and $\mathbf{v} \preceq \alpha$ if $v_i \leq \alpha, \forall i$. We let $\mathbf{A} \odot \mathbf{B}$ represent the element-wise product of two matrices $\mathbf{A}$ and $\mathbf{B}$. Following (Rakotomamonjy et al. 2008), we also define $\frac{x_i}{0} = 0$ if $x_i = 0$ and ∞ otherwise. In the paper, we use the following minimax saddle-point theorem for deriving our proposed method.

Theorem 1. *(Sion 1958) Let $g(\mathbf{x}, \mathbf{z}) : \mathcal{X} \times \mathcal{Z} \rightarrow R$, where $\mathcal{X}$ and $\mathcal{Z}$ are compact convex subsets of linear topological spaces, $g(\cdot, \mathbf{z})$ is quasiconvex and lower semi-continuous for every $\mathbf{z} \in \mathcal{Z}$, and $g(\mathbf{x}, \cdot)$ is quasiconcave and upper semi-continuous for every $\mathbf{x} \in \mathcal{X}$. Then $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{z}\in\mathcal{Z}} g(\mathbf{x}, \mathbf{z}) = \max_{\mathbf{z}\in\mathcal{Z}} \min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}, \mathbf{z})$, in particular, each optimum is attainable.*

ℓ_1 Norm versus Variation Norm

Note that ℓ_1-norm of a vector $\mathbf{x}$ can be expressed as the following variation form (Jenatton, Audibert, and Bach 2009): $\|\mathbf{x}\|_1 = \sum_{i=1}^m |x_i| = \frac{1}{2} \min_{\boldsymbol{\eta}\succeq 0} \sum_{i=1}^m \frac{x_i^2}{\eta_i} + \eta_i$. One major issue for ℓ_1-norm regularization is that it is inefficient when solving very high dimensional and large-scale problems. To tackle this issue, inspired by the variation norm, we define a new variation norm parameterized by a scalar B. Let $\Lambda = \{\boldsymbol{\eta} | \sum \eta_i \leq B, 0 \leq \eta_i \leq 1\}$, we define the $\|\mathbf{x}\|_B$ norm as $\|\mathbf{x}\|_B = \sqrt{\min_{\boldsymbol{\eta}\in\Lambda} \sum_{i=1}^m \frac{x_i^2}{\eta_i}}$.

Proposition 1. *Given a vector $\mathbf{x} \in \mathbb{R}^m$ with $\|\mathbf{x}\|_0 = \hat{\kappa}$, where $\hat{\kappa} > 0$ defines the number of nonzero entries in $\mathbf{x}$. Consider the following minimization problem $\min_{\boldsymbol{\eta}\in\Lambda} \sum_{i=1}^m \frac{x_i^2}{\eta_i}$, we have: (1) Let $\boldsymbol{\eta}^*$ be the minimizer, then $\eta_i^* = 0$ if $|x_i| = 0$ and $\eta_1^* > \eta_2^*, ..., > \eta_{\hat{\kappa}}^* > 0$ if $|x_1| > |x_2|, ..., > |x_{\hat{\kappa}}| > 0$. (2) If $\hat{\kappa} \leq B$, then $\eta_i = 1$ for $|x_i| > 0$; otherwise, $\frac{|x_i|}{\eta_i} = \|\mathbf{x}\|_1 / B$ for all $|x_i| > 0$. (3) If $\hat{\kappa} \leq B$, then $\|\mathbf{x}\|_B = \|\mathbf{x}\|_2$; otherwise, $\|\mathbf{x}\|_B = \frac{\|\mathbf{x}\|_1}{\sqrt{B}}$.*

The proof can be found in Appendix A. Based on the definition of $\|\mathbf{x}\|_B$, suppose $\mathbf{y} = \mathbf{D}\hat{\mathbf{x}} + \boldsymbol{\xi}$, we consider to solve a regularized sparse inverse problem: $\min_{\hat{\mathbf{x}}} \|\hat{\mathbf{x}}\|_B^2 + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{D}\hat{\mathbf{x}}\|^2$, which is equivalent to the following problem:

$$\min_{\hat{\mathbf{x}}, \boldsymbol{\eta}\in\Lambda} \frac{1}{2}\sum_{i=1}^m \frac{\hat{x}_i^2}{\eta_i} + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{D}\hat{\mathbf{x}}\|^2, \tag{3}$$

where $\lambda > 0$ is a regularization parameter that trades off the model complexity and the fitness of the regressor. By defining $x_i = \frac{\hat{x}_i}{\eta_i}$, the above convex problem can be easily transformed as the following problem:

$$\min_{\boldsymbol{\eta}\in\Lambda} \min_{\mathbf{x}} \frac{1}{2}\sum_{i=1}^m \eta_i x_i^2 + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{D}\mathrm{diag}(\boldsymbol{\eta})\mathbf{x}\|^2. \tag{4}$$

Proposition 2. *Given a fixed $\boldsymbol{\eta} \in \Lambda$, the dual of the inner minimization problem of (4) regarding $\mathbf{x}$ can be given by*

$$\max_{\boldsymbol{\alpha}} \; \boldsymbol{\alpha}'\mathbf{y} - \frac{\boldsymbol{\alpha}'\mathbf{D}diag(\boldsymbol{\eta})\mathbf{D}'\boldsymbol{\alpha}}{2} - \frac{1}{2\lambda}\boldsymbol{\alpha}'\boldsymbol{\alpha}, \tag{5}$$

with regression error $\boldsymbol{\xi} = \frac{1}{\lambda}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^n$.

Define $f(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{1}{2}\boldsymbol{\alpha}'\mathbf{D}\mathrm{diag}(\boldsymbol{\eta})\mathbf{D}'\boldsymbol{\alpha} + \frac{1}{2\lambda}\boldsymbol{\alpha}'\boldsymbol{\alpha} - \boldsymbol{\alpha}'\mathbf{y}$, and $-f(\boldsymbol{\alpha}, \boldsymbol{\eta})$ is concave regarding $\boldsymbol{\alpha}$, so the globally optimal solution $\boldsymbol{\alpha}^*$ of (5) exists. Thus, we can find a bounded region $\mathcal{A} = [-\varpi, \varpi]^n$, where ϖ is a large number such that $\boldsymbol{\alpha}^* \in \mathcal{A}$. Now, both $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ in (5) are in compact domains, and according to Theorem 1, we have $\min_{\boldsymbol{\eta}\in\Lambda} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -f(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\eta}\in\Lambda} -f(\boldsymbol{\alpha}, \boldsymbol{\eta})$. By bringing in a new variable $\theta \in \mathbb{R}$, the latter problem can be further transformed to a convex QCQP problem (Tan, Wang, and Tsang 2010):

$$PD : \min_{\boldsymbol{\alpha}\in\mathcal{A}, \theta} \theta \; : \; f(\boldsymbol{\alpha}, \boldsymbol{\eta}) \leq \theta, \; \forall \, \boldsymbol{\eta} \in \Lambda. \tag{6}$$

Convex Matching Pursuit

Notice that there are infinite number of quadratic inequality constraints in (6), making it hard to solve. In this paper, we propose to solve it via an efficient central cutting plane algorithm (CCP) (Elzinga and Moore 1975; Kortanek and No 1993), which iteratively includes the most violated active constraint into the active constraint set and solves a series of reduced optimization problems. Its convergence behaviors have been thoroughly studied (Elzinga and Moore 1975; Kortanek and No 1993), and its efficiency and effectiveness have been verified on solving semi-infinite programs (SIP) (Kortanek and No 1993). As to our problem, the details of the CCP algorithm for solving (6) is shown in Algorithm 1. Here, θ^{k-1} represents the upper bound estimate of the original primal objective value, $\theta - \delta$ is the dual objective value of the reduced problem, and 2δ is the separation between them.

We slightly improve the CCP algorithm (Elzinga and Moore 1975; Kortanek and No 1993) by solving (7), (8) and (9) exactly, leading to faster convergence. Notice that Algorithm 1 is also similar to the iterative procedure of MP algorithms in the sense that Problem (7) and (9) are coincided with the matching step in MP and Problem (8) is related to the projection step in MP. Hence, we name it as a convex matching pursuit (CMP). However, CMP apparently differs from MP in several aspects, such as different motivations and matching schemes. More importantly, as will be shown later, CMP can guarantee its convergence.

Algorithm 1 Convex matching pursuit for solving (6)

```
0: Initialize $\boldsymbol{\alpha}^0 = \mathbf{y}$, and find the most-violated constraint via
       $\boldsymbol{\eta}_0 = \arg\max_{\boldsymbol{\eta}} f(\boldsymbol{\alpha}^0, \boldsymbol{\eta}).$    (7)
1: Let $\theta^0 = f(\boldsymbol{\alpha}^0, \boldsymbol{\eta}_0)$, $\Psi_0 = \{\boldsymbol{\eta}_0\}$ and set $k = 1$.
2: Solve the following program:
       $\mathrm{SD}_k: \quad V_D = \max_{\boldsymbol{\alpha} \in \mathcal{A}, \theta, \delta} \delta$    (8)
       $s.t. \quad \theta + \delta \leq \theta^{k-1}, f(\boldsymbol{\alpha}, \boldsymbol{\eta}_k) \leq \theta - \delta, \forall \boldsymbol{\eta}_k \in \Psi_{k-1}.$
3: Let $(\boldsymbol{\alpha}^k, \theta^k, \delta^k)$ be the solution to $\mathrm{SD}_k$. If $\delta^k = 0$, stop.
4: Check the most violated constraint via
       $\boldsymbol{\eta}_k = \arg\max_{\boldsymbol{\eta}} f(\boldsymbol{\alpha}^k, \boldsymbol{\eta}).$    (9)
5: Let $\Psi_k = \Psi_{k-1} \cup \{\boldsymbol{\eta}_k\}$, if $f(\boldsymbol{\alpha}^k, \boldsymbol{\eta}_k) > \theta^k$, that is, $\boldsymbol{\alpha}^k$ is an
   infeasible solution to $PD$, add the constraint $f(\boldsymbol{\alpha}, \boldsymbol{\eta}_k) \leq \theta - \delta$
   to $\mathrm{SD}_k$. Let $k = k + 1$ and go to Step 2.
```

Matching versus Exploratory Matching

At each iteration in Algorithm 1, one needs to find the most violated $\boldsymbol{\eta}_k$ via solving (9). Let $\mathbf{z} = [z_1, \ldots, z_m]' = \mathbf{D}'\boldsymbol{\alpha}^k$, we have

$$\max_{\boldsymbol{\eta} \in \Lambda} \mathbf{z}' \mathrm{diag}(\boldsymbol{\eta}) \mathbf{z} = \max_{\boldsymbol{\eta} \in \Lambda} \sum_{i=1}^{m} \eta_i z_i^2. \tag{10}$$

It is easy to verify that the global solution to (10) can be obtained exactly by finding the B largest elements z_i^2, which is similar to the matching stage in MP algorithms (Dai and Milenkovic 2009). Although (10) can be solved exactly, the obtained solution $\boldsymbol{\eta}$ may not be optimal for the whole CCP algorithm due to the immediate solution $\boldsymbol{\alpha}^k$, which is obtained based on only a portion of possible supports of $\mathbf{x}$. To remedy this issue, we can explore a larger search space via including more possible supports and then do a projection with all included supports. Specifically, we can include k_e ($k_e > 1$) groups of size B elements via solving (10). And then we do projection via solving Problem SD with all selected atoms. Finally, we can rank those newly included $k_e B$ atoms according to the regressor $\mathbf{z}$ and choose those B atoms with the largest z_i^2 among the newly included $k_e B$ atoms as the most violated atoms. In summary, the (exploratory) matching scheme can be implemented as in Algorithm 2.

Notice that each $\boldsymbol{\eta}$ represents a group of possible supports of $\mathbf{x}$. With the exploratory search, one can possibly find better supports and hence the convergence speed as well as the decoding ability can be potentially improved. However, as the projection stage requires to solve (8), additional computational costs are needed. To differentiate the two matching schemes, we term the CMP with exploratory search as ECMP.

Algorithm 2 Matching and Exploratory Matching.

```
Given $\boldsymbol{\alpha} \in \mathbb{R}^n$, $B \in \mathbb{Z}_+$, two zero vectors $\mathbf{z} \in \mathbb{R}^m$ and $\boldsymbol{\eta} \in \mathbb{R}^m$,
design matrix $\mathbf{D}$, and the response vector $\mathbf{y} \in \mathbb{R}^n$.
1: Calculate $\mathbf{z} = \mathbf{D}'\boldsymbol{\alpha}$.
   *** Exploratory Search ***
   2: Find the $k_e B$ atoms with the largest $|z_j|$'s.
   3: Rank the $k_e B$ atoms in descending order and group them
      into $k_e$ groups.
   4: Do projection via solving Problem (SD) with the newly included $k_e$ groups.
   5: Obtain an updated solution $\mathbf{z}$.
6: Select the $B$ atoms with the largest $|z_i|$'s as the active atoms
   and set $\eta_i = 1$.
7: Return the obtained atom subset $\mathbf{D}_t = \mathbf{D}\mathrm{diag}(\boldsymbol{\eta}_t)$.
```

Solution to subproblem SD

Now, we are ready to solve the Problem (8). For convenience, here we omit the index k. Problem (8) is a nonsmooth QCQP problem defined by Ψ with $T = |\Psi|$, which can be formulated as a MKL problem (Rakotomamonjy et al. 2008), and can be solved via sub-gradient methods (Rakotomamonjy et al. 2008) or SQP methods. However, these methods are inefficient when dealing with large-scale problems with the high desired accuracy. In this paper, we propose to solve this problem efficiently in its primal via an Accelerated Proximal Gradient (APG) algorithm (Toh and Yun 2009; Ji and Ye 2009). At first, the following lemma is useful to derive the primal form of Problem SD.

Lemma 1. *In Algorithm 1, let $\{\nu^k\}$ and $\{\hat{\mu}_t^k\}$ be the optimal dual variables of Problem SD in kth iteration, then we have $\nu^k = \frac{1}{2}$, $\sum_t \hat{\mu}_t^k = \frac{1}{2}$ and $\theta^k + \delta^k = \theta^{k-1}$. Furthermore, let $\boldsymbol{\mu} \in \Pi = \{\boldsymbol{\mu} | \boldsymbol{\mu} \succeq 0, \sum_{t=1}^{T} \mu_t = 1\}$, then Problem SD is equivalent to the following problem:*

$$\min_{\boldsymbol{\mu} \in \Pi} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \frac{1}{2}\theta^{k-1} - \frac{1}{2} \sum_{\boldsymbol{\eta}_t \in \Psi} \mu_t f(\boldsymbol{\alpha}, \boldsymbol{\eta}_t). \tag{11}$$

Proof. The Lagrangian function of Problem SD is $\mathcal{L}(\boldsymbol{\alpha}, \theta, \delta, \nu, \hat{\mu}) = \delta - \nu(\theta - \theta^{k-1} + \delta) - \sum_t \hat{\mu}_t (f(\boldsymbol{\alpha}, \boldsymbol{\eta}_t) - \theta + \delta)$. When setting the gradient of $\mathcal{L}(\cdot)$ with respect to θ and δ to zeros, we get the optimal dual variables $\{\nu^k\}$ and $\{\hat{\mu}_t^k\}$ satisfy $\sum_t \hat{\mu}_t^k + \nu^k = 1$ and $\sum_t \hat{\mu}_t^k - \nu^k = 0$. Alternatively, we have $\sum_t \hat{\mu}_t^k = \frac{1}{2}$ and $\nu^k = \frac{1}{2} > 0$. From the KKT condition, $\theta^k + \delta^k = \theta^{k-1}$ holds. Since the objective function is concave in $\boldsymbol{\alpha}$ and convex in $\hat{\mu}_t$, and both $\boldsymbol{\alpha}$ and $\hat{\mu}_t$ are in convex compact domains, we can exchange the order of max and min operators using Theorem 1. Finally, letting $\mu_t = 2\hat{\mu}_t$ completes the proof. □

With Lemma 1, let $\mathbf{x}$ be a supervector concatenating all $\mathbf{x}_k$'s and $\boldsymbol{\xi} = \mathbf{y} - \mathbf{D} \sum_{k=1}^{T} \mathrm{diag}(\boldsymbol{\eta}_k) \mathbf{x}^k$, and we have the following results regarding the primal form of problem (SD).

Theorem 2. *The primal form of $\min_{\boldsymbol{\mu} \in \Pi} \max_{\boldsymbol{\alpha} \in \mathcal{A}} - \sum_{\boldsymbol{\eta}_t \in \Psi} \mu_t f(\boldsymbol{\alpha}, \boldsymbol{\eta}_t)$ can be expressed as:*

$$F(\mathbf{x}) = \frac{1}{2} \Big(\sum_{k=1}^{T} \|\mathbf{x}_k\| \Big)^2 + p(\mathbf{x}), \tag{12}$$

where $p(\mathbf{x}) = \frac{\lambda}{2} \|\boldsymbol{\xi}\|^2$. Once its optimality is achieved, the dual variables $\boldsymbol{\alpha}$ in (11) can be recovered by $\boldsymbol{\alpha} = \lambda \boldsymbol{\xi}$.

Proof. By applying the conic duality theory, the proof parallels the results in (Bach, Lanckriet, and Jordan 2004). □

The primal problem (12) is non-smooth and non-separable. In this paper, we propose to solve it via the APG algorithm. Note that, with the primal form, we only need to solve a reduced problem defined by Ψ_k, which can gain much better efficiency. On the other hand, from Theorem 2, with the complementary slackness, we can easily recover the dual variable $\boldsymbol{\alpha}$ via the training errors $\boldsymbol{\xi}$, which will be used to find new possible supports via matching.

Let $\Omega(\mathbf{x}) = (\sum_{k=1}^{T} \|\mathbf{x}_k\|)^2/2$ in the APG algorithm, we adopt the following quadratic approximation of $F(\mathbf{x})$ at the point $\mathbf{v}$: $Q_\tau(\mathbf{x}, \mathbf{v}) = \Omega(\mathbf{x}) + p(\mathbf{v}) + \nabla p(\mathbf{v})'(\mathbf{x} - \mathbf{v}) + \frac{\tau}{2}\|\mathbf{x} - \mathbf{v}\|^2 = \Omega(\mathbf{x}) + p(\mathbf{v}) - \frac{1}{2\tau}\|\nabla p(\mathbf{v})\|^2 + \frac{\tau}{2}\|\mathbf{x} - \mathbf{g}\|^2$, where $\tau > 0$ and $\mathbf{g} = \mathbf{v} - \frac{1}{\tau}\nabla p(\mathbf{v})$. And then we need to solve the following Moreau Projection problem(Martins et al. 2010):

$$\min_{\mathbf{x}} \frac{\tau}{2}\|\mathbf{x} - \mathbf{g}\|^2 + \Omega(\mathbf{x}),$$

which has a unique optimal solution. Specifically, let $S_\tau(\mathbf{v})$ be the optimal solution, it can be easily calculated via Algorithm 2 in (Martins et al. 2010). Finally, we can solve (12) via an APG method in Algorithm 3, which is adapted from (Ji and Ye 2009; Toh and Yun 2009). When Algorithm 3 terminates, $\{\mathbf{x}^k\}$ converges to the optimal solution to (12).

Algorithm 3 APG for solving (12).

```
Initialization: set $\mathbf{x}^0 = 0$, $\eta \in (0, 1)$, $t^0 = t^1 = 1$, $k = 0$ and
$\tau^0 = L_0$ is the initial guess of the Lipshitz constant $L$.
1: Set $\mathbf{v}^k = \mathbf{x}^k + \frac{t^{k-1}-1}{t^k}(\mathbf{x}^k - \mathbf{x}^{k-1})$.
2: Set $\tau = \eta\tau_{k-1}$.
   For $j = 0, 1, ...,$
      Set $\mathbf{g} = \mathbf{v}^k - \frac{1}{\tau}\nabla p(\mathbf{v}^k)$, compute $S_\tau(\mathbf{g})$.
      If $F(S_\tau(\mathbf{g})) \leq Q(S_\tau(\mathbf{g}), \mathbf{v}^k)$,
         set $\tau_k = \tau$, stop;
      Else
         $\tau = \min\{\eta^{-1}\tau, L_0\}$.
      End
   End
3: Set $\mathbf{x}^{k+1} = S_{\tau_k}(\mathbf{g})$.
4: Compute $t^{k+1} = \frac{1+\sqrt{(1+4(t^k)^2)}}{2}$.
5: Quit if stopping condition achieves. Otherwise, go to step 1.
```

Regarding the above APG algorithm, the following convergence rate is guaranteed.

Theorem 3. *(Ji and Ye 2009) Let $\{\mathbf{x}^{k+1}\}$, $\{\mathbf{v}^k\}$ and $\{t^k\}$ be the sequences generated by APG and L be the Lipschitz constant of $p(\mathbf{x})$, for any $k \geq 1$, we have*

$$F(\mathbf{x}^k) - F(\mathbf{x}^*) \leq \frac{2L||\mathbf{x}^0 - \mathbf{x}^*||^2}{\eta(k+1)^2}.$$

Convergence and Performance Guarantees

Since (8) and (7) can be solved exactly, we have the following results regarding convergence property of Algorithm 1.

Lemma 2. *(Lemma 3.1 in (Kortanek and No 1993)) The generated sequence $\{\delta^k\}$ ($\delta^k \geq 0$) in Algorithm 1 will converge to 0.*

Theorem 4. *There exists $\hat{k}$, such that $\boldsymbol{\alpha}^{\hat{k}-1}$ is feasible for problem PD and Algorithm 1 stops in the $\hat{k}$th iteration with $\boldsymbol{\alpha}^{\hat{k}}$ to be the optimal solution of Problem PD.*

The proof parallels the proof of Theorem 3.1 in (Kortanek and No 1993), which is omitted due to the page limit.

Theorem 5. *The function difference of successive iterations in Algorithm 1 will converge to 0.*

Proof. From Lemma 1, we have $\delta^k = \theta^{k-1} - \theta^k$. And the function difference between two iterations is δ^k. From Lemma 2, δ^k will monotonically decrease (*i.e.* $\delta^k/\delta^{k-1} \leq 1$) and converge to 0, which completes the proof. □

Now we discuss the RIP condition under which `CMP` can successfully recover the κ-sparse signals.

Lemma 3. *Problem $P_0 : \min_{\mathbf{x}} \|\mathbf{x}\|_1 + \lambda_1\|\mathbf{y} - \mathbf{Ax}\|^{\mathbf{2}}$ and $P_1 : \frac{1}{2}\min_{\mathbf{x}} \|\mathbf{x}\|_1^2 + \lambda_2\|\mathbf{y} - \mathbf{Ax}\|^{\mathbf{2}}$ have the same optimal solution $\mathbf{x}^*$ by adjusting the parameters λ_1 and λ_2.*

Proof. Let $l_1(\mathbf{x}) = \lambda_1\|\mathbf{y} - \mathbf{Ax}\|^{\mathbf{2}}$ and $l_2(\mathbf{x}) = \lambda_2\|\mathbf{y} - \mathbf{Ax}\|^{\mathbf{2}}$. Their solutions are unique. The KKT condition for Problem (P_0) is $\nabla_{x_i} l_1(x_i) + 1 = 0$ if $x_i > 0$; $\nabla_{x_i} l_1(x_i) - 1 = 0$ if $x_i < 0$; $|\nabla_{x_i} l_1(x_i)| \leq 1$ if $x_i = 0$. For any $\lambda_1 > 0$, once we obtain the optimal solution $\mathbf{x}^*(\mathbf{x}^* \neq 0)$ for P_0, we have $\|\mathbf{x}^*\|_1|\nabla_{x_i^*} l_1(x_i^*)| + \|\mathbf{x}^*\|_1 = 0$ if $x_i^* > 0$; $\|\mathbf{x}^*\|_1|\nabla_{x_i^*} l_1(x_i^*)| - \|\mathbf{x}^*\|_1 = 0$ if $x_i^* < 0$; $\|\mathbf{x}^*\|_1|\nabla_{x_i^*} l_1(x_i^*)| \leq \|\mathbf{x}^*\|_1$ if $x_i^* = 0$, which is the KKT condition for Problem (P_1) by setting $\lambda_2 = \|\mathbf{x}^*\|_1\lambda_1$. □

Theorem 6. `CMP` *algorithms can recover the κ-sparse signal under RIP condition with $\sigma_\kappa = 0.307$ given that $B < \kappa$.*

Proof. With Lemma 3 and the equivalence of the two problems in (2), we can immediately complete the proof by adapting the proof from (Cai, Wang, and Xu 2010) together with the equivalence of Problem (3), (6), (P_0), and (P_1). □

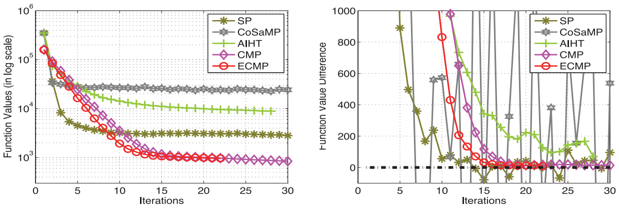
Experiments

Experimental Settings

In this section, we will compare the performance of `CMP` and `ECMP` with other baseline methods. The aforementioned methods, namely, `OMP`, `AIHT`, `SP`, `CoSaMP`, `FISTA` and `ADM` are adopted as baseline methods, which have shown good overall performance and their Matlab implementations are also available[1]. Although `OMPR` shows relatively better RIP condition, it is sensitive to its learning rate η according to our study. In addition, its performance is close to `AIHT` (Jain, Tewari, and Dhillon 2011) and hence we did not include it for comparison. Finally, as $(||\mathbf{w}||_1)^2$ is closely related to $||\mathbf{x}||_B^2$ and can be directly solved via the `FISTA` algorithm, we also implement it and denote it as `FISTA`-l_{12}.

All the methods are re-implemented in C++, where several implementation issues have been carefully considered for fair comparison. For `MP` algorithms, the efficient conjugate gradient descent (`CGD`) algorithm together with warm

[1] http://sites.google.com/site/igorcarron2/cscodes; http://www.eecs. berkeley.edu/ yang/software/l1benchmark/index.html.



(a) Function value (b) Function value difference

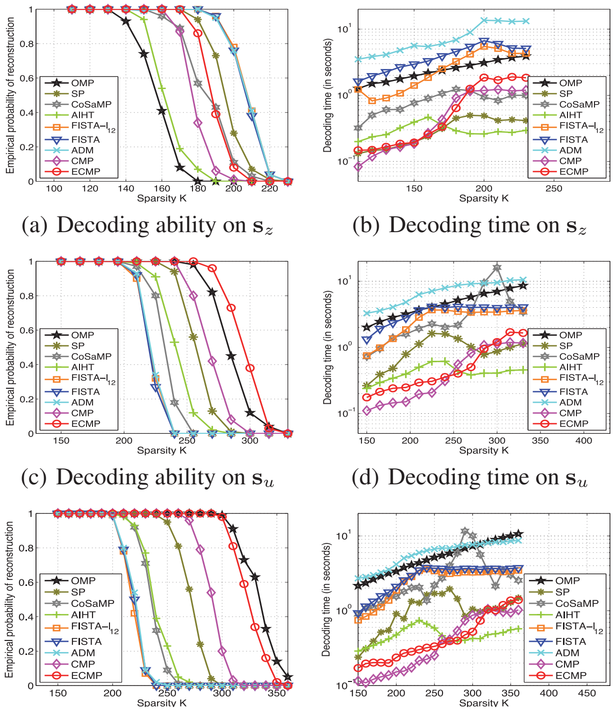
Figure 1: Convergence behavior of various MP algorithms.

start is adopted to solve the projection problem (Blumensath and Davies 2008b). For FISTA and ADM, the continuation and debiasing (implemented by CGD) are adopted to improve their convergence speed and decoding performance, respectively (Figueiredo, Nowak, and Wright 2007; Yang et al. 2010). All the experiments are conducted on Intel(R) Core(TM) i7 CPU with 64-bit operating system.

We follow the testing strategy described in (Dai and Milenkovic 2009) to conduct the experiments. At first, three different sizes of Gaussian random matrices $\mathbf{D} \in \mathbb{R}^{n \times m}$, namely, $2^{10} \times 2^{13}$, $2^{12} \times 2^{15}$ and $2^{13} \times 2^{17}$, are generated as the design matrices. Then, three types of κ-sparse signals, Zero-one signal (denoted by $\mathbf{s}_z$ and each nonzero entry is either 1 or -1), Uniform signal (denoted by $\mathbf{s}_u$ and each nonzero entry is sampled from a uniform distribution $\mathcal{U}(-1,1)$) and Gaussian signal (denoted by $\mathbf{s}_g$ and each nonzero entry is sampled from a Gaussian distribution $\mathcal{N}(0,1)$) are generated to produce the measurement $\mathbf{y} = \mathbf{D}\mathbf{x} + \boldsymbol{\xi}$ with Gaussian noise sampled from $\mathcal{N}(0, 0.05^2)$. Via varying the sparsity κ, we can obtain different signals. The decoding ability and decoding time of each algorithm regarding κ are reported, respectively. The decoding ability is measured by the empirical probability of successful reconstruction from $\mathbf{y}$ among M independent experiments (Dai and Milenkovic 2009). For FISTA, we set the trade-off parameter ρ in (3) as $\rho = 0.005||\mathbf{D}'\mathbf{y}||_\infty$, which is suggested in many ℓ_1 packages. For ECMP and CMP, we set the regularization parameter $\lambda = 1000.0/||\mathbf{y}||$. For MP algorithms, a guess of the sparsity κ is required. In real problems, we have no idea of the ground-truth sparsity κ. In our simulation, we guess it by $\hat{\kappa} = \kappa + 0.2\kappa$. For CMP and ECMP, we guess $B = \hat{\kappa}/6$ for zero-one signal and $B = \hat{\kappa}/8$ for the other two types of signals. In addition, we set $\kappa_e = 5$ for ECMP. We use the default settings for other parameters for baseline methods as in their Matlab implementations.

Experimental Results

In the first experiments, we show the convergence behaviors of SP, CoSaMP, AIHT, CMP and ECMP on $\mathbf{D} \in 2^{10} \times 2^{13}$ with Gaussian signal of sparsity $\kappa = 360$ as in Figure 1. From Figure 1(a), we can observe that the objective values of AIHT, CMP and ECMP monotonically decrease, which verifies $\theta^k \leq \theta^{k-1}$ in Algorithm 1. However, this property does not hold for SP and CoSaMP under this setting. Figure 1(b) shows the function value difference for comparing methods, we can observe that CMP and ECMP manifest a stable and monotonic decrement for their function value difference (*i.e.* $\delta^k/\delta^{k-1} \leq 1$) where $\delta^k = \theta^{k-1} - \theta^k$; while the function value difference of SP, CoSaMP, AIHT are very fluctuated. In short, CMP and ECMP demonstrate better convergence results than others. For AIHT, the non-increasing function values can be theoretically guaranteed (Blumensath and Davies 2008a). However, because the optimization problem in AIHT is non-convex and AIHT is only guaranteed to find local solutions, which will limit its decoding ability (Blumensath and Davies 2008a).



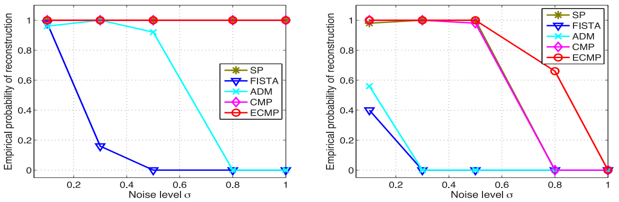
(a) Decoding ability on $\mathbf{s}_z$ (b) Decoding time on $\mathbf{s}_z$

(c) Decoding ability on $\mathbf{s}_u$ (d) Decoding time on $\mathbf{s}_u$

(e) Decoding ability on $\mathbf{s}_g$ (f) Decoding time on $\mathbf{s}_g$

Figure 2: Decoding results on $\mathbf{D} \in \mathbb{R}^{2^{10} \times 2^{13}}$.

To thoroughly compare various methods, we tested them on different sizes of design matrices mentioned above. On $\mathbf{D} \in \mathbb{R}^{2^{10} \times 2^{13}}$, a relatively small problem, we compared all mentioned algorithms. For each κ, we run $M = 200$ independent trials. The empirical probability of successful reconstruction and decoding time for the three kinds of signals, $\mathbf{s}_z$, $\mathbf{s}_u$ and $\mathbf{s}_g$, are presented in Figure 2. From Figure 2, we can observe following facts. Firstly, on Zero-one signal, ℓ_1-norm based methods, including FISTA-l_{12}, show the best decoding ability; while SP and ECMP show better decoding performance over other MP methods. Secondly, on Uniform and Gaussian signals, ECMP and OMP show relatively better decoding ability than other methods; while ℓ_1-norm based methods achieve much worse results. It is possible that ℓ_1-norm based methods assume the Zero-one signal in their objective (see (2)). Thirdly, on all three cases, ECMP shows improved decoding performance compared with CMP, which verify the validity of exploratory search in matching. However, on all cases, CMP shows the fastest decoding speed when κ is not too large; while ECMP can be faster than CMP when CMP fails to decode the signals. Moreover, from Figure 2, AIHT also shows good convergence. However, it shows relatively worse decoding ability than other methods, indi-



(a) Decoding ability on $\mathbf{s}_z$ (b) Decoding ability on $\mathbf{s}_u$

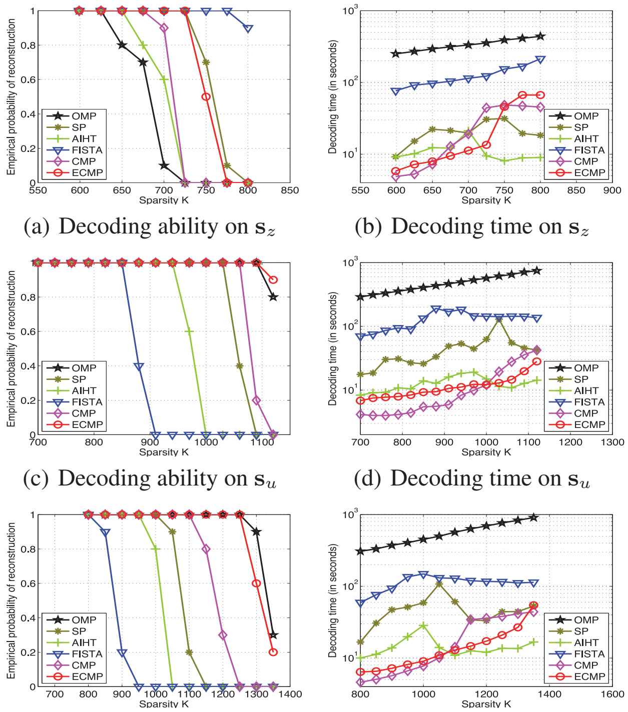
Figure 3: Decoding results with different scale of noises on $\mathbf{D} \in \mathbb{R}^{2^{10} \times 2^{13}}$.

cating that it is easy to converge to the local minima. Finally, all the ℓ_1-norm based methods show close decoding abilities on all signals. Considering that ADM is ρ parameter free, we can conclude that the decoding ability of FISTA cannot be further improved too much via tuning ρ.

To test the robustness of various methods over disturbances, in Figure 3, we showed the decoding ability of SP, FISTA, ADM, CMP and ECMP under different levels of noise. Here only the Zero-one signals (of sparsity $k = 140$) and the Uniform signals (of sparsity $k = 220$) with Gaussian noise $\mathcal{N}(0, \sigma^2)$ are studied. And the sparsity is set such that all methods can exactly recover the signals under the noise level $\sigma = 0.05$ from Figure 2. In addition, we vary $\sigma \in \{0.1, 0.3, 0.5, 0.8, 1.0\}$ to generate different levels of noise. From Figure 3(a) on Zero-one signals, we can observe that matching pursuit algorithms have better stability than the ℓ_1-regularized methods over the increasing noises. From Figure 3(b) on Uniform signals, we can see that ECMP shows better stability than its counterparts with larger level of noise.

In the next two experiments, we focus on the scalability of various methods. Because of the extensive computational cost for experiments, we only include OMP, SP, AIHT, FISTA, CMP and ECMP for comparisons on the design matrix $\mathbf{D} \in \mathbb{R}^{2^{12} \times 2^{15}}$ with 10 trials for each parameter κ. The decoding ability and time of various methods are shown in Figure 4. The experimental results have similar observations to those obtained on $\mathbf{D} \in \mathbb{R}^{2^{10} \times 2^{13}}$. Again, CMP and ECMP obtain the best decoding performance over all methods except for the Zero-one signals. Particularly, CMP and ECMP show much better decoding efficiency than other methods.

In the final experiment, we do the simulation on the largest problem $\mathbf{D} \in \mathbb{R}^{2^{13} \times 2^{17}}$. For computational issues, we only compare SP, AIHT, CMP, and ECMP on Gaussian signals with 10 independent trials. The experimental results are shown in Figure 5. From Figure 5, we can observe that both CMP and ECMP show much better decoding ability on Gaussian signals. And from Figure 5(b), CMP and ECMP can gain better decoding efficiency on large-scale problems.

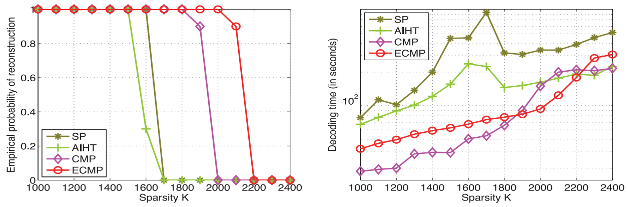


(a) Decoding ability on $\mathbf{s}_z$ (b) Decoding time on $\mathbf{s}_z$

(c) Decoding ability on $\mathbf{s}_u$ (d) Decoding time on $\mathbf{s}_u$

(e) Decoding ability on $\mathbf{s}_g$ (f) Decoding time on $\mathbf{s}_g$

Figure 4: Decoding results on $\mathbf{D} \in \mathbb{R}^{2^{10} \times 2^{15}}$.



(a) Decoding ability on $\mathbf{s}_g$ (b) Decoding time on $\mathbf{s}_g$

Figure 5: Decoding results on $\mathbf{D} \in \mathbb{R}^{2^{13} \times 2^{17}}$

Conclusion

A Convex Matching Pursuit scheme is presented to handle large-scale sparse coding and subset selection problems. Unlike current MP algorithms, our proposed scheme solves a parameterized variation norm regularized problem that attains the convexity property. Hence it can surely converge. Different from the ℓ_1 norm regularized methods, CMP (ECMP) can gain much better efficiency on dealing with large-scale problems and better decoding ability on Gaussian signals. Extensive experiments demonstrated state-of-the-art performance of the proposed methods over baseline methods in terms of both decoding ability and decoding time on large scale problems.

Appendix A: Proof of Proposition 1

Proof. (1): We prove it by contradiction. Firstly, suppose $\boldsymbol{\eta}^*$ is a minimizer and there exists $l \in \{1 \dots m\}$, such that $x_l = 0$ but $\eta_l^* > 0$. Let $0 < \epsilon < \eta_l^*$, and choose one $j \in \{1 \dots m\}$, such that $|x_j| > 0$. Define new solution $\hat{\boldsymbol{\eta}}$ in the following way: $\hat{\boldsymbol{\eta}}_j = \eta_j^* + \eta_l^* - \epsilon$, and $\hat{\boldsymbol{\eta}}_l = \epsilon$, $\hat{\boldsymbol{\eta}}_k = \eta_k^*$ for $k \in \{1 \dots m\} \backslash \{j, l\}$. Then it is easy to check $\sum_{i=1}^m \hat{\eta}_i = \sum_{i=1}^m \eta_i^* \leq B$, i.e. $\hat{\eta}$ is also a feasible point, but $\sum_{i=1}^m \frac{x_i^2}{\hat{\eta}_i} < \sum_{i=1}^m \frac{x_i^2}{\eta_i^*}$, which contradict $\boldsymbol{\eta}^*$ is the minimizer. Secondly, if $|x_i| > 0$ and $\eta_i^* = 0$, by the definition, $\frac{x_i^2}{0} = \infty$. As we expect to get the finite minimum, so if $|x_i| > 0$, we have $\eta_i^* > 0$. Suppose there exist two pairs $\{x_i, \eta_i^*\}$ and $\{x_j, \eta_j^*\}$ such that $|x_i| > |x_j|$ but $\eta_i^* < \eta_j^*$, define a new $\hat{\boldsymbol{\eta}}$ in the following way: $\hat{\boldsymbol{\eta}}_j = \eta_i^*$, and $\hat{\boldsymbol{\eta}}_i = \eta_j^*$, $\hat{\boldsymbol{\eta}}_k = \eta_k^*$ for $k \in \{1 \dots m\} \backslash \{i, j\}$, then we have $\sum_{i=1}^m \frac{x_i^2}{\hat{\eta}_i} - \sum_{i=1}^m \frac{x_i^2}{\eta_i^*} = \left(\frac{x_i^2}{\hat{\eta}_j} + \frac{x_j^2}{\hat{\eta}_i}\right) - \left(\frac{x_i^2}{\eta_i^*} + \frac{x_j^2}{\eta_j^*}\right) = \frac{(\eta_i^* - \eta_j^*)(x_i^2 - x_j^2)}{\eta_i^* \eta_j^*} < 0$, which contradict the fact that $\boldsymbol{\eta}^*$ is a minimizer. Inductively, we get (1).

(2): The argument holds trivially when $||\mathbf{x}||_0 = \hat{\kappa} \leq B$. When $||\mathbf{x}||_0 = \hat{\kappa} > B$, WLOG, we assume $|x_i| > 0$ for the first $\hat{\kappa}$ elements. From (1), we have $\eta_i > 0$ and $\sum_{i=1}^{\hat{\kappa}} \eta_i \leq B$. Note that $\sum_{i=1}^{\hat{\kappa}} \frac{x_i^2}{\eta_i}$ is convex regarding $\boldsymbol{\eta}$. The KKT condition of the problem is: $-x_i^2/\eta_i^2 + \gamma - \zeta_i + \nu_i = 0, \zeta_i \eta_i = 0, \nu_i(1 - \eta_i) = 0, \gamma(B - \sum_{i=1}^{\hat{\kappa}} \eta_i) = 0, \gamma \geq 0, \zeta_i \geq 0, \nu_i \geq 0, i \in \{1 \dots \hat{\kappa}\}$, where γ, ζ_i and ν_i are the dual variables for the constraints $\sum_{i=1}^{\hat{\kappa}} \eta_i \leq B$, $\eta_i > 0$ and $1 - \eta_i \geq 0$ respectively. As $\eta_i > 0$, we have $\zeta_i = 0$ for $i \in \{1 \dots \hat{\kappa}\}$ from the KKT condition. By the first equality in KKT condition, we have $\eta_i = |x_i|/\sqrt{\gamma + \nu_i}$. As $\sum_{i=1}^{\hat{\kappa}} \eta_i \leq B < \hat{\kappa}$, which implies there exist some $\eta_i < 1$ and $\nu_i = 0$. For $i \in \{1 \dots \hat{\kappa}\}$, $|x_i| > 0$ and $\eta_i > 0$, $\eta_i = |x_i|/\sqrt{\gamma + \nu_i}$ implies $\gamma \neq 0$, and by complementary condition, we have $\sum_{i=1}^{\hat{\kappa}} \eta_i = B$. By substituting η_i back to the objective function, we have $\sum_{i=1}^{\hat{\kappa}} |x_i| \sqrt{\gamma + \nu_i}$. To get the minimum, we need all $\nu_i = 0$. Then for $i \in \{1 \dots \hat{\kappa}\}$, $\eta_i = \frac{|x_i|}{\sqrt{\gamma}}$, and $\sum_{i=1}^{\hat{\kappa}} \frac{|x_i|}{\sqrt{\gamma}} = B$, by simple calculation, we get $\sqrt{\gamma} = ||\mathbf{x}||_1/B$ and $\eta_i = B|x_i|/||\mathbf{x}||_1$.

(3): With the results of (2), if $\hat{\kappa} \leq B$, we have $\sum_{i=1}^m \frac{x_i^2}{\eta_i} = \sum_{i=1}^{\hat{\kappa}} x_i^2$, so $\|\mathbf{x}\|_B = \|\mathbf{x}\|_2$. And if $\hat{\kappa} > B$, we have $\sum \frac{x_i^2}{\eta_i} = \sum \frac{|x_i|}{\eta_i} |x_i| = \frac{||\mathbf{x}||_1}{B} \sum |x_i| = \frac{(||\mathbf{x}||_1)^2}{B}$. Hence we have $\|\mathbf{x}\|_B = \sqrt{\sum \frac{x_i^2}{\eta_i}} = \frac{||\mathbf{x}||_1}{\sqrt{B}}$. □

References

Bach, F. R.; Lanckriet, G. R. G.; and Jordan, M. I. 2004. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*.

Blumensath, T., and Davies, M. E. 2008a. Iterative thresholding for sparse approximations. *J. Fourier. Anal. Appl.* 14(5):629–654.

Blumensath, T., and Davies, M. 2008b. Gradient pursuits. *IEEE Trans. Signal Proces.* 56(6):2370–2382.

Blumensath, T. 2011. Accelerated iterative hard threshoding. *Signal Proces.* 92:752–756.

Cai, T. T.; Wang, L.; and Xu, G. 2010. New bounds for restricted isometry constants. *IEEE Trans. Info. Theory* 56(9):4388–4394.

Candès, E. J., and Tao, T. 2005. Decoding by linear programming. *IEEE Trans. Info. Theory* 51(12):4203–4215.

Candès, E. J., and Wakin, M. 2008. An introduction to compressive sampling. *IEEE Signal Proces. Mag.* 25(2):21–30.

Dai, W., and Milenkovic, O. 2009. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Info. Theory* 55(5):2230–2249.

Davenport, M.; Boufounos, P.; Wakin, M.; and Baraniuk, R. 2010. Signal processing with compressive measurements. *IEEE J. Sel. Topics Signal Proces.* 4(2):445–460.

Elzinga, J., and Moore, T. G. 1975. A central cutting plane algorithm for the convex programming problem. *Math. Programming* 8:134–145.

Figueiredo, M. A. T.; Nowak, R. D.; and Wright, S. J. 2007. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Topics Signal Proces.*

Gao, S.; Tsang, I. W.; Chia, L.-T.; and Zhao, P. 2010. Local features are not lonely - Laplacian sparse coding for image classification. In *CVPR*.

Giryes, R., and Elad, M. 2012. RIP-based near-oracle performance guarantees for subspace-pursuit, cosamp, and iterative hard-thresholding. *IEEE Trans. Signal Proces.* 60(3):1465 – 1468.

Jain, P.; Tewari, A.; and Dhillon, I. S. 2011. Orthogonal matching pursuit with replacement. In *NIPS*.

Jenatton, R.; Audibert, J.-Y.; and Bach, F. 2009. Structured variable selection with sparsity-inducing norms. Technical report.

Ji, S., and Ye, J. 2009. An accelerated gradient method for trace norm minimization. In *ICML*.

Kortanek, K. O., and No, H. 1993. A central cutting plane algorithm for convex semi-infinite programming problems. *SIAM J. Optimization* 3(4):901–918.

Martins, A. F.; Figueiredo, M. A. T.; Aguiar, P. M. Q.; Smith, N. A.; and Xing, E. P. 2010. Online multiple kernel learning for structured prediction. Technical report.

Needell, D., and Tropp, J. 2009. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis* 26(3):301–321.

Rakotomamonjy, A.; Bach, F.; Grandvalet, Y.; and Canu, S. 2008. SimpleMKL. *JMLR* 9:2491–2521.

Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics* 8(1):171–176.

Tan, M.; Wang, L.; and Tsang, I. 2010. Learning sparse SVM for feature selection on very high dimensional datasets. In *ICML*.

Toh, K.-C., and Yun, S. 2009. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Technical report.

Tropp, J., and Gilbert, A. 2007. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Info. Theory* 53(12):4655–4666.

Wright, J.; Yang, A.; Ganesh, A.; Sastry, S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(2):210–227.

Yang, A.; Ganesh, A.; Ma, Y.; and Sastry, S. 2010. Fast L1-minimization algorithms and an application in robust face recognition: A review. In *ICIP*.

Zhang, T. 2011. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Info. Theory* 57:6215–6221.

Zibulevsky, M., and Elad, M. 2010. L1-L2 optimization in signal and image processing. *IEEE Signal Proces. Mag.* 27(3):76–88.