Sparse Recovery with Orthogonal Matching Pursuit under RIP

Tong Zhang
Statistics Department
Rutgers University, NJ
tzhang@stat.rutgers.edu

Abstract

This paper presents a new analysis for the orthogonal matching pursuit (OMP) algorithm. It is shown that if the restricted isometry property (RIP) is satisfied at sparsity level $O(\bar{k})$, then OMP can recover a $\bar{k}$-sparse signal in 2-norm. For compressed sensing applications, this result implies that in order to uniformly recover a $\bar{k}$-sparse signal in $\mathbb{R}^d$, only $O(\bar{k} \ln d)$ random projections are needed. This analysis improves earlier results on OMP that depend on stronger conditions such as mutual incoherence that can only be satisfied with $\Omega(\bar{k}^2 \ln d)$ random projections.

1 Introduction

Consider a signal $\bar{\mathbf{x}} \in \mathbb{R}^d$, and we observe its linear transformation plus noise as:

$$\mathbf{y} = A\bar{\mathbf{x}} + \text{noise},$$

where A is an $n \times d$ matrix. If we define an objective function

$$Q(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2, \tag{1}$$

then $\bar{\mathbf{x}}$ approximately minimizes $Q(\mathbf{x})$.

If $d > n$, then the solution of (1) is not unique. In order to recover $\bar{\mathbf{x}}$ based on optimizing (1), additional assumptions on $\bar{\mathbf{x}}$ is necessary. We are specifically interested in the case where $\bar{\mathbf{x}}$ is sparse. That is $\|\bar{\mathbf{x}}\|_0 \ll n$, where

$$\|x\|_0 = |\text{supp}(x)|, \quad \text{supp}(x) = \{j : x_j \neq 0\}.$$

It is known that under appropriate conditions, it is possible to recovery $\bar{\mathbf{x}}$ by solving (1) with sparsity constraint as follows:

$$\min_{\mathbf{x} \in \mathbb{R}^d} Q(\mathbf{x}) \qquad \text{subject to } \|\mathbf{x}\|_0 \leq k. \tag{2}$$

However, this optimization problem is generally NP-hard. Therefore one seeks computationally efficient algorithms that can approximately solve (2), with the goal of recovering sparse signal $\bar{\mathbf{x}}$. This paper considers the popular orthogonal matching pursuit algorithm (OMP), which has been widely used for this purpose. We are specifically interested in two issues: the performance of OMP in terms of optimizing $Q(\mathbf{x})$ and the performance of OMP in terms of recovering the sparse signal $\bar{\mathbf{x}}$.

2 Main Result

Our analysis considers a more general objective function $Q(\mathbf{x})$ that not necessarily takes the quadratic form in (1). However, we assume that $Q(\mathbf{x})$ is convex. For such a general convex objective function, we consider the fully (or totally) corrective greedy algorithm in Figure 1, which was analyzed in [6]. We will refine the analysis in this paper in order to show that the algorithm works under the RIP condition. This algorithm is a directly generalization of OMP which has been traditionally considered only for the quadratic objective function in (1). The algorithm has been known in the machine learning community as a version of boosting [9], and has also been proposed recently in the signal processing community [1]. In order to use notation consistent with the sparse recovery literature, in the current paper, we still refer to this more general algorithm as OMP even though it applies to objective functions other than (1).

Input: $Q(\mathbf{x})$ defined on $\mathbb{R}^d$, and initial feature set $F^{(0)} \subset \{1, \ldots, d\}$.
Output: $\mathbf{x}^{(k)}$
let $\mathbf{x}^{(0)} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x})$ subject to $\mathrm{supp}(\mathbf{x}) \subset F^{(0)}$
(default choice is $F^{(0)} = \emptyset$ with $\mathbf{x}^{(0)} = 0$)
for $k = 1, 2, \ldots$
  let $j = \arg\max_i |\nabla Q(\mathbf{x}^{(k-1)})_i|$
  let $F^{(k)} = \{j\} \cup F^{(k-1)}$
  let $\mathbf{x}^{(k)} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x})$ subject to $\mathrm{supp}(\mathbf{x}) \subset F^{(k)}$
end

Figure 1: Fully Corrective Greedy Boosting Algorithm (OMP)

The general problem of optimization under sparsity constraint is NP hard. In order to alleviate the difficulty, we consider approximate optimization under the restricted strong convexity assumption introduced below.

Definition 2.1 (Restricted Strong Convexity Constants) *Given any $s \geq 0$, define restricted strong convexity constants $\rho_-(s)$ and $\rho_+(s)$ as follows: for all $\|\mathbf{x} - \mathbf{x}'\|_0 \leq s$, we require*

$$\rho_-(s)\|\mathbf{x} - \mathbf{x}'\|_2^2 \leq Q(\mathbf{x}') - Q(\mathbf{x}) - \nabla Q(\mathbf{x})^\top(\mathbf{x}' - \mathbf{x}) \leq \rho_+(s)\|\mathbf{x} - \mathbf{x}'\|_2^2.$$

If the objective function takes the quadratic form (1), then the above definition becomes sparse eigenvalues of $A^\top A$, which is used in defining the restricted isometry property (RIP) [2].

In order to recover the target $\bar{\mathbf{x}}$, we have to assume that $\bar{\mathbf{x}}$ is sparse and approximately optimizes $Q(\mathbf{x})$. If a target $\bar{\mathbf{x}}$ is a global optimal solution, then $\nabla Q(\bar{\mathbf{x}}) = 0$. However, this paper deals with sparse approximate optimal solution, where $\nabla Q(\bar{\mathbf{x}}) \approx 0$. In particular, we introduce the following definition, which is convenient to apply.

Definition 2.2 (Restricted Gradient Optimal Constatnt) *Given $\bar{\mathbf{x}} \in \mathbb{R}^d$ and $s > 0$, we define the restricted gradient optimal constant $\delta_s(\bar{\mathbf{x}})$ as:*

$$|\nabla Q(\bar{\mathbf{x}})^\top \mathbf{u}| \leq \delta_s(\bar{\mathbf{x}})\|\mathbf{u}\|_2$$

for all $\mathbf{u} \in \mathbb{R}^d$ such that $\|\mathbf{u}\|_0 \leq s$.

Proposition 2.1 *We have* $\delta_s(\bar{\mathbf{x}}) \leq \sqrt{s}\|\nabla Q(\bar{\mathbf{x}})\|_\infty$ *and* $\delta_s(\bar{\mathbf{x}}) \leq \|\nabla Q(\bar{\mathbf{x}})\|_2$. *Moreover, if* $Q(\bar{\mathbf{x}}) \leq \inf_{\|\mathbf{x}\|_0 \leq \|\bar{\mathbf{x}}\|_0 + s} Q(\mathbf{x}) + \epsilon$, *then*

$$\delta_s(\bar{\mathbf{x}}) \leq 2\sqrt{\rho_+(s)\epsilon}.$$

Proof The first two inequalities are straight-forward. For the third inequality, we note that for $\|\mathbf{u}\|_0 \leq s$:

$$\begin{aligned}\inf_{\|\mathbf{x}\|_0 \leq \|\bar{\mathbf{x}}\|_0 + s} Q(\mathbf{x}) \leq & \inf_\eta Q(\bar{\mathbf{x}} + \eta\mathbf{u}) \\ \leq & \inf_\eta [Q(\bar{\mathbf{x}}) + \eta \nabla Q(\bar{\mathbf{x}})^\top \mathbf{u} + \rho_+(s)\eta^2 \|\mathbf{u}\|_2^2] \\ = & Q(\bar{\mathbf{x}}) - |\nabla Q(\bar{\mathbf{x}})^\top \mathbf{u}|^2/(4\rho_+(s)\|\mathbf{u}\|_2^2).\end{aligned}$$

The result follows by rearranging the above inequality. ■

The following theorem is the main result of this paper, which shows that OMP can approximately recover a sparse signal $\bar{\mathbf{x}}$ in 2-norm if a certain condition on the strong convexity constants hold.

Theorem 2.1 *Consider the OMP algorithm. Let* $\bar{\mathbf{x}} \in \mathbb{R}^d$ *and* $\bar{F} = \mathrm{supp}(\bar{\mathbf{x}})$. *If there exists* s *such that*

$$s \geq |\bar{F} \cup F^{(0)}| + 4|\bar{F} - F^{(0)}|(\rho_+(1)/\rho_-(s)) \ln(20\rho_+(|\bar{F} - F^{(0)}|)/\rho_-(s)),$$

then when $k = s - |\bar{F} \cup F^{(0)}|$, *we have*

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + 2.5\delta_s(\bar{\mathbf{x}})^2/\rho_-(s)$$

and

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2 \leq \sqrt{6}\delta_s(\bar{\mathbf{x}})/\rho_-(s).$$

Proof The detailed proof relies on a number of technical lemmas that are left to the appendix.

The first inequality of the theorem is a direct consequence of Lemma A.5. The second inequality is a consequence of the first inequality and Lemma A.2:

$$\rho_-(s)\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2^2 \leq 2\left[Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}})\right] + \delta_s(\bar{\mathbf{x}})^2/\rho_-(s) \leq 6\delta_s(\bar{\mathbf{x}})^2/\rho_-(s).$$

This implies the second inequality. ■

The following result gives a simpler interpretation of the above theorem, where it is easy to check that the condition of the theorem is satisfied. For the quadratic objective (1), the condition $\rho_+(\|\bar{\mathbf{x}}\|_0) \leq 2\rho_-(31\|\bar{\mathbf{x}}\|_0)$ is referred to as RIP in [2].

Corollary 2.1 *Consider the OMP algorithm with* $F^{(0)} = \emptyset$. *Let* $\bar{\mathbf{x}} \in \mathbb{R}^d$. *If the RIP condition* $\rho_+(\|\bar{\mathbf{x}}\|_0) \leq 2\rho_-(31\|\bar{\mathbf{x}}\|_0)$ *holds, then when* $k = 30\|\bar{\mathbf{x}}\|_0$, *we have*

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + 2.5\delta_s(\bar{\mathbf{x}})^2/\rho_-(s)$$

and

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2 \leq \sqrt{6}\delta_s(\bar{\mathbf{x}})/\rho_-(s),$$

where $s = 31\|\bar{\mathbf{x}}\|_0$.

For quadratic objective, a simple instantiation of $\delta_s(\bar{\mathbf{x}})$ using Proposition 2.1 leads to the following sparse recovery result that is relatively simple to interpret.

Corollary 2.2 *If $Q(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2$. Consider the OMP algorithm with $F^{(0)} = \emptyset$. Let $\bar{\mathbf{x}} \in \mathbb{R}^d$. If the RIP condition $\rho_+(\|\bar{\mathbf{x}}\|_0) \leq 2\rho_-(31\|\bar{\mathbf{x}}\|_0)$ holds, then when $k = 30\|\bar{\mathbf{x}}\|_0$, we have*

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2 \leq 2\sqrt{6}\rho_+(s)^{1/2}\|A\bar{\mathbf{x}} - \mathbf{y}\|_2/\rho_-(s),$$

where $s = 31\|\bar{\mathbf{x}}\|_0$.

3 Discussion

In this paper we proved a new result for the orthogonal matching pursuit algorithm. It is shown that if the restricted isometry property (RIP) is satisfied at sparsity level $O(\bar{k})$, then OMP can recover a $\bar{k}$-sparse signal in 2-norm. For compressed sensing applications, this result implies that in order to uniformly recover a $\bar{k}$-sparse signal in $\mathbb{R}^d$, only $n = O(\bar{k}\ln d)$ random projections are needed [2].

Our result is stronger than previous results for OMP that relied on different conditions. For example, [7] considered the problem of recovering the support set of a sparse signal under a stronger condition (also see [11]). A similar analysis was employed in [8], where it was shown that for any fixed sparse signal $\bar{\mathbf{x}}$ with $\bar{k} = \|\bar{\mathbf{x}}\|_0$, OMP can recover the signal with large probability using $O(\bar{k}\ln d)$ measurements. However, this result is not uniform with respect to all $\bar{k}$-sparse signals $\bar{\mathbf{x}}$ (that is, for any set of random projections, there exist $\bar{k}$-sparsity signals that fail the analysis). In comparison, the RIP condition holds uniformly by definition, and hence our result applies uniformly to all $\bar{k}$-sparse signals. Although some previous results apply uniformly to all $\bar{k}$-sparse signals, such as those in [3], they depend on the stronger mutual incoherence condition. In particular a result similar to Corollary 2.2 but under the mutual incoherence condition can be found in [4]. Unfortunately the mutual incoherence condition can only be satisfied with $\Omega(\bar{k}^2\ln d)$ random projections.

It is interesting to compare the new OMP result in this paper to that of Lasso, which is also known to work under RIP. However, a more refined comparison illustrates differences between the known theoretical results for these two methods. For OMP, the result in Theorem 2.1 can be applied as long as the condition

$$s/|\bar{F} \cup F^{(0)}| \geq 4|\bar{F} - F^{(0)}|(\rho_+(1)/\rho_-(s))\ln(20\rho_+(|\bar{F} - F^{(0)}|)/\rho_-(s))$$

is satisfied. With $F^{(0)} = \emptyset$, this roughly requires $(\rho_+(1)/\rho_-(s))\ln(\rho_+(\bar{k})/\rho_-(s))$ to grow sub-linearly as a function of s in order to apply the theory. In comparison, the known condition for Lasso (e.g., this has been made explicit in [10, 12]) requires $\rho_+(s)/\rho_-(s)$ to grow sub-linearly as a function of s. To compare the two conditions, we note that the condition for OMP is weaker in terms of of the upper convexity constant as there is no explicit dependency on $\rho_+(s)$; however, the dependency on $\rho_-(s)$ is stronger in OMP than Lasso due to the logarithmic term. Although it is unclear how tight these conditions are, the comparison nevertheless indicates that even though both algorithms work under RIP, there are still finer differences in their theoretical analysis: Lasso is slightly more favorable in terms of its dependency on the lower strong convexity constant, while the OMP is more favorable in terms of its dependency on the upper strong convexity constant. We further conjecture that the extra logarithmic dependency $\ln(\rho_+(\bar{k})/\rho_-(s))$ in OMP is necessary. In practice, it is known that some times Lasso performs better while other times OMP performs better. Therefore some

discrepancy in their theoretical analysis is expected. The theory in this paper significantly narrows the previous theoretical gap between these two sparse recovery methods by positively answering the open question of whether OMP can recovery sparse signals under RIP. Therefore our result allows practitioners to apply OMP with more confidence than previously expected.

References

[1] T. Blumensath and M. E. Davies. Gradient pursuit for non-linear sparse signal modelling. In *European Signal Processing Conference (EUSIPCO)*, 2008.

[2] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.

[3] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info. Theory*, 52(1):6–18, 2006.

[4] D. Hsu, S. M. Kakade, J. Langford, and T. Zhang. Multi-label prediction via compressed sensing. In *NIPS' 09*, 2009.

[5] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. Technical report, Rutgers University, January 2009. A short version appears in ICML'09. Available from http://arxiv.org/abs/0903.3002.

[6] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity. Technical report, TTI, May 2009.

[7] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Info. Theory*, 50(10):2231–2242, 2004.

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

[9] M. Warmuth, J. Liao, and G. Ratsch. Totally corrective boosting algorithms that maximize the margin. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

[10] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimension al linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[11] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

[12] T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.

A Technical Lemmas

We need a number of technical lemmas. Lemma A.3 and Lemma A.4, key to the proof, are based on earlier work of the author with collaborators. The first three lemmas use the following notations.

Let $F, \bar{F}$ be two subsets of $\{1, \ldots, d\}$. Let $\mathrm{supp}(\bar{\mathbf{x}}) \subset \bar{F}$, and

$$\mathbf{x} = \arg \min_{\mathbf{z}:\mathrm{supp}(\mathbf{z}) \subset F} Q(\mathbf{z}).$$

Lemma A.1 *We have*

$$Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) \leq 1.5\rho_+(s)\|\bar{\mathbf{x}}_{\bar{F}-F}\|_2^2 + 0.5\delta_s(\bar{\mathbf{x}})^2/\rho_+(s)$$

for all $s \geq |\bar{F} - F|$.

Proof Let $\mathbf{x}' = \bar{\mathbf{x}}_{\bar{F}\cap F}$, then by definition of $\mathbf{x}$, we know that $Q(\mathbf{x}) \leq Q(\mathbf{x}')$. Therefore

$$\begin{aligned} Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) \leq & Q(\mathbf{x}') - Q(\bar{\mathbf{x}}) \\ = & Q(\mathbf{x}') - Q(\bar{\mathbf{x}}) - \nabla Q(\bar{\mathbf{x}})^\top(\mathbf{x}' - \bar{\mathbf{x}}) + \nabla Q(\bar{\mathbf{x}})^\top(\mathbf{x}' - \bar{\mathbf{x}}) \\ \leq & \rho_+(s)\|\bar{\mathbf{x}}_{\bar{F}-F}\|_2^2 + \delta_s(\bar{\mathbf{x}})\|\bar{\mathbf{x}}_{\bar{F}-F}\|_2 \\ \leq & \rho_+(s)\|\bar{\mathbf{x}}_{\bar{F}-F}\|_2^2 + 0.5\delta_s(\bar{\mathbf{x}})^2/\rho_+(s) + 0.5\rho_+(s)\|\bar{\mathbf{x}}_{\bar{F}-F}\|_2^2, \end{aligned}$$

which implies the lemma. ■

Lemma A.2 *We have:*

$$\rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2 \leq 2\left[Q(\mathbf{x}) - Q(\bar{\mathbf{x}})\right] + \delta_s(\bar{\mathbf{x}})^2/\rho_-(s)$$

for all $s \geq |F \cup \bar{F}|$.

Proof From

$$\begin{aligned} Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) = & Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) - \nabla Q(\bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}}) + \nabla Q(\bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}}) \\ \geq & \rho_-(s)\|\bar{\mathbf{x}} - \mathbf{x}\|_2^2 - \delta_s(\bar{\mathbf{x}})\|\bar{\mathbf{x}} - \mathbf{x}\|_2 \\ \geq & 0.5\rho_-(s)\|\bar{\mathbf{x}} - \mathbf{x}\|_2^2 - 0.5\delta_s(\bar{\mathbf{x}})^2/\rho_-(s), \end{aligned}$$

we obtain the desired inequality. ■

The next lemma shows that each greedy search makes reasonable progress. This proof is essential identically to a similar result in [6] but with refined notations which is used in the current paper. We thus include the proof for completeness. It allows the readers to verify more easily that the proof in [6] remains unchanged with our new definitions.

Lemma A.3 *Let* $\mathbf{e}_i \in \mathbb{R}^d$ *be the vector of zeros except for the* i*-th component being one. If* $\bar{F} - F \neq \emptyset$*, then for all* $s \geq |F \cup \bar{F}|$*:*

$$\min_\alpha Q(\mathbf{x} + \alpha\mathbf{e}_j) \;\leq\; Q(\mathbf{x}) - \frac{\rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|^2}{\rho_+(1)\left(\sum_{i\in\bar{F}-F}|\bar{\mathbf{x}}_i|\right)^2}(Q(\mathbf{x}) - Q(\bar{\mathbf{x}})),$$

where $j = \arg\max_i |\nabla Q(\mathbf{x})_i|$.

Proof For all $j \in \bar{F} - F$ and $\eta > 0$, we define

$$Q_j(\eta) = Q(\mathbf{x}) + \eta \operatorname{sgn}(\bar{\mathbf{x}}_j) \nabla Q(\mathbf{x})_j + \eta^2 \rho_+(1).$$

It follows from the definition of $\rho_+(1)$ that

$$Q(\mathbf{x} + \eta \operatorname{sgn}(\bar{\mathbf{x}}_j) \mathbf{e}_j) \ \leq \ Q_j(\eta).$$

Since the choice of $j = \arg\max_i |\nabla Q(\mathbf{x})_i|$ achieves the minimum of $\min_i \min_\eta Q_i(\eta)$, the lemma is a direct consequence of the following stronger statement:

$$\min_i Q_i(\eta) \ \leq \ Q(\mathbf{x}) - \frac{\left(Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) + \rho_-(s)\|\mathbf{x} - \mathbf{x}_i\|^2\right)^2}{4\rho_+(1) \left(\sum_{i \in \bar{F} - F} |\bar{\mathbf{x}}_i|\right)^2}, \tag{3}$$

with an appropriate choice of η; this is because

$$\left(Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) + \rho_-(s)\|\mathbf{x} - \mathbf{x}_i\|^2\right)^2 \geq 4\rho_-(s)(Q(\mathbf{x}) - Q(\bar{\mathbf{x}}))\|\mathbf{x} - \mathbf{x}_i\|^2.$$

Therefore, we now turn to prove that (3) holds. Denote $u = \sum_{i \in \bar{F} - F} |\bar{\mathbf{x}}_i|$, we obtain that

$$\begin{aligned} u \min_i Q_i(\eta) \leq & \sum_{i \in \bar{F} - F} |\bar{\mathbf{x}}_i| Q_i(\eta) \\ \leq & \ u\, Q(\mathbf{x}) + \eta \sum_{i \in \bar{F} - F} \bar{\mathbf{x}}_i \nabla Q(\mathbf{x})_i + u\, \rho_+(1)\eta^2. \end{aligned} \tag{4}$$

Since we assume that $\mathbf{x}$ is optimal over F, we get that $\nabla Q(\mathbf{x})_i = 0$ for all $i \in F$. Additionally, $\mathbf{x}_i = 0$ for $i \notin F$ and $\bar{\mathbf{x}}_i = 0$ for $i \notin \bar{F}$. Therefore,

$$\begin{aligned} \sum_{i \in \bar{F} - F} \bar{\mathbf{x}}_i \nabla Q(\mathbf{x})_i &= \sum_{i \in \bar{F} - F} (\bar{\mathbf{x}}_i - \mathbf{x}_i) \nabla Q(\mathbf{x})_i \\ &= \sum_{i \in \bar{F} \cup F} (\bar{\mathbf{x}}_i - \mathbf{x}_i) \nabla Q(\mathbf{x})_i \\ &= \nabla Q(\mathbf{x})^\top (\bar{\mathbf{x}} - \mathbf{x}) \quad . \end{aligned}$$

Combining the above with the definition of $\rho_-(s)$, we obtain that

$$\sum_{i \in \bar{F} - F} \bar{\mathbf{x}}_i \nabla Q(\mathbf{x})_i \leq Q(\bar{\mathbf{x}}) - Q(\mathbf{x}) - \rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2 \ .$$

Combining the above with (4) we get

$$u \min_i Q_i(\eta) \leq \ u\, Q(\mathbf{x}) + \eta \left[Q(\bar{\mathbf{x}}) - Q(\mathbf{x}) - \rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2\right] + u\, \rho_+(1)\eta^2.$$

Setting $\eta = [Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) + \rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2]/(2u\rho_+(1))$ and rearranging the terms, we conclude our proof of (3). ■

The direct consequence of the previous lemma is the following result, which is critical in our analysis. The idea of using a nesting approximating sequence has appeared in [5], but the current version is improved. The change is necessary for the purpose of this paper.

Lemma A.4 *Consider the OMP algorithm. Consider* $\bar{F}_1, \bar{F}_2, \ldots, \bar{F}_L \subset \bar{F} \cup F^{(0)}$, *and* $\bar{F}_0 = \bar{F} \cap F^{(0)}$. *Assume that* $\min_{\mathbf{x}:\mathrm{supp}(\mathbf{x})\subset \bar{F}_j} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_j$, $q_0 \geq q_1 \geq \cdots \geq q_L \geq 0$, *and let* $\mu \geq \sup_{j=1,\ldots,L-1}(q_{j-1}/q_j)$. *If* $s \geq |F^{(k)} \cup \bar{F}|$ *and*

$$k = \sum_{j=1}^{L} \left\lceil |\bar{F}_j - F^{(0)}|(\rho_+(1)/\rho_-(s)) \ln(2\mu) \right\rceil,$$

then

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + q_L + \mu^{-1} q_{L-1}.$$

Proof Note that for any $\mathrm{supp}(\mathbf{x}) \subset F$ and $\mathrm{supp}(\bar{\mathbf{x}}) \subset \bar{F}$, we have

$$\frac{\rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|^2}{\rho_+(1)\left(\sum_{i\in \bar{F}-F} |\bar{\mathbf{x}}_i|\right)^2} \geq \frac{\rho_-(s)}{\rho_+(1)|\bar{F} - F|}.$$

Therefore Lemma A.3 implies that at any k such that $s \geq |F^{(k)} \cup \bar{F}|$ and $\ell = 0, \ldots, L$, we have

$$Q(\mathbf{x}^{(k+1)}) \;\leq\; Q(\mathbf{x}^{(k)}) - \frac{\rho_-(s)}{\rho_+(1)|\bar{F}_\ell - F^{(k)}|} \max\left(0, Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell\right),$$

where we simply replace the target vector $\bar{\mathbf{x}}$ in Lemma A.3 by the optimal solution over $\bar{F}_\ell$, and replace $\mathbf{x}$ by $\mathbf{x}^{(k)}$. It implies that

$$\begin{aligned}\max(0, Q(\mathbf{x}^{(k+1)}) - Q(\bar{\mathbf{x}}) - q_\ell) \;&\leq\; \left[1 - \frac{\rho_-(s)}{\rho_+(1)|\bar{F}_\ell - F^{(k)}|}\right] \max\left(0, Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell\right) \\ &\leq \exp\left[-\frac{\rho_-(s)}{\rho_+(1)|\bar{F}_\ell - F^{(k)}|}\right] \max\left(0, Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell\right).\end{aligned}$$

Therefore for any $k' \leq k$ and $\ell = 1, \ldots, L$, we have

$$Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell \;\leq\; \exp\left[-\frac{\rho_-(s)(k-k')}{\rho_+(1)|\bar{F}_\ell - F^{(k')}|}\right] \max\left(0, Q(\mathbf{x}^{(k')}) - Q(\bar{\mathbf{x}}) - q_\ell\right). \tag{5}$$

We are now ready to prove the lemma by induction on L. If $L = 1$, we can set $k' = 0$ in (5). Since

$$Q(\mathbf{x}^{(0)}) - Q(\bar{\mathbf{x}}) - q_1 \leq q_0,$$

we obtain that when

$$k = \left\lceil |\bar{F}_1 - F^{(0)}|(\rho_+(1)/\rho_-(s)) \ln(2\mu) \right\rceil,$$

we have

$$Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_1 \;\leq\; \exp\left[-\frac{\rho_-(s)k}{\rho_+(1)|\bar{F}_1 - F^{(0)}|}\right] q_0 \leq (2\mu)^{-1} q_0.$$

Therefore the lemma holds. Now assume that the lemma holds at $L = m - 1$ for some $m > 1$. That is, with

$$k' = \sum_{j=1}^{m-1} \left\lceil |\bar{F}_j - F^{(0)}|(\rho_+(1)/\rho_-(s)) \ln(2\mu) \right\rceil,$$

we have

$$Q(\mathbf{x}^{(k')}) \leq Q(\bar{\mathbf{x}}) + q_{m-1} + \mu^{-1} q_{m-2}.$$

This implies that when $L = m$:

$$Q(\mathbf{x}^{(k')}) - Q(\bar{\mathbf{x}}) - q_L \leq q_{L-1} + \mu^{-1} q_{L-2} - q_L \leq 2q_{L-1}.$$

We thus obtain from (5) that

$$Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_L \;\leq\; \exp\left[-\frac{\rho_-(s)(k-k')}{\rho_+(1)|\bar{F}_L - F^{(0)}|}\right](2q_{L-1}) \leq (2\mu)^{-1}(2q_{L-1}).$$

This finishes the induction. ■

The following lemma is a slightly stronger version of the theorem, which we can prove more easily by induction.

Lemma A.5 *Consider the OMP algorithm. If there exist k and s such that $|\bar{F} \cup F^{(k)}| \leq s$ and*

$$k = \left\lceil 4|\bar{F} - F^{(0)}|(\rho_+(1)/\rho_-(s)) \ln(20\rho_+(|\bar{F} - F^{(0)}|)/\rho_-(s)) \right\rceil,$$

then

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + 2.5\delta_s(\bar{\mathbf{x}})^2/\rho_-(s). \tag{6}$$

Proof We prove this result by induction on $|\bar{F} - F^{(0)}|$. If $|\bar{F} - F^{(0)}| = 0$, then the bound in (6) holds trivially because $Q(\mathbf{x}^{(k)}) \leq Q(\mathbf{x}^{(0)}) \leq Q(\bar{\mathbf{x}})$.

Assume that the claim holds with $|\bar{F} - F^{(0)}| \leq m-1$ for some $m > 0$. Now we consider the case of $|\bar{F} - F^{(0)}| = m$. Without loss of generality, we assume for notational convenience that $\bar{F} - F^{(0)} = \{1, \ldots, m\}$, and $|\bar{\mathbf{x}}_j|$ in $\bar{F} - F^{(0)}$ is arranged in descending order so that $|\bar{\mathbf{x}}_1| \geq |\bar{\mathbf{x}}_2| \geq \cdots \geq |\bar{\mathbf{x}}_m|$. Let L be the smallest positive integer such that for all $1 \leq \ell < L$, we have

$$\sum_{i=2^{\ell-1}}^{m} \bar{\mathbf{x}}_i^2 < \mu \sum_{i=2^{\ell}}^{m} \bar{\mathbf{x}}_i^2,$$

but

$$\sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2 \geq \mu \sum_{i=2^{L}}^{m} \bar{\mathbf{x}}_i^2, \tag{7}$$

where $\mu = 10\rho_+(m)/\rho_-(s)$. We have $L \leq \lfloor \log_2 m \rfloor + 1$ because the second inequality is automatically satisfied when $L = \lfloor \log_2 m \rfloor + 1$ (the right hand side is zero in this case).

We can now define

$$\bar{F}_\ell = (\bar{F} \cap F^{(0)}) \cup \{i : 1 \leq i \leq \min(m, 2^\ell - 1)\} \quad (\ell = 0, 1, 2, \ldots, L).$$

Lemma A.1 implies that for $\ell = 0, 1, \ldots, L$:

$$\min_{\mathbf{x} \subset \bar{F}_\ell} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_\ell, \quad q_\ell = 1.5\rho_+(m) \sum_{i=2^\ell}^{m} \bar{\mathbf{x}}_i^2 + 0.5\delta_s(\bar{\mathbf{x}})^2/\rho_+(m).$$

Moreover $q_{\ell-1} \leq \mu q_\ell$ when $\ell = 1, \ldots, L-1$. We can thus apply Lemma A.4 to conclude that when

$$k = \sum_{j=1}^{L} \left\lceil (2^j - 1)(\rho_+(1)/\rho_-(s)) \ln(2\mu) \right\rceil \leq 2^{L+1}(\rho_+(1)/\rho_-(s)) \ln(2\mu) - 1, \tag{8}$$

we have

$$\begin{aligned} Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) \leq & 1.5\rho_+(m) \sum_{i=2^L}^{m} \bar{\mathbf{x}}_i^2 + 1.5\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2 + 0.5(1+\mu^{-1})\delta_s(\bar{\mathbf{x}})^2/\rho_+(m) \\ \leq & 3\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2 + 0.5(1+\mu^{-1})\delta_s(\bar{\mathbf{x}})^2/\rho_+(m), \end{aligned}$$

where (7) is used to derive the second inequality.

Now, if $2\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2 \leq (1+\mu^{-1})\delta_s(\bar{\mathbf{x}})^2/\rho_-(s)$, then the above inequality implies that (6) holds automatically, which finishes the induction. Therefore in the following, we only consider the case this is not true. That is,

$$2\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2 > (1+\mu^{-1})\delta_s(\bar{\mathbf{x}})^2/\rho_-(s).$$

Now Lemma A.2 implies that

$$\begin{aligned} \rho_-(s)\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2^2 \leq & 2(Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}})) + \delta_s(\bar{\mathbf{x}})^2/\rho_-(s) \\ \leq & 6\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2 + (2+\mu^{-1})\delta_s(\bar{\mathbf{x}})^2/\rho_-(s) \\ < & 10\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2 = \rho_-(s) \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2. \end{aligned}$$

It implies that

$$\sum_{i=m-|\bar{F}-F^{(k)}|+1}^{m} \bar{\mathbf{x}}_i^2 \leq \sum_{i \in \bar{F}-F^{(k)}} \bar{\mathbf{x}}_i^2 \leq \|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2^2 < \sum_{i=2^{L-1}}^{m} \bar{\mathbf{x}}_i^2.$$

Therefore $m - |\bar{F} - F^{(k)}| + 1 > 2^{L-1}$. That is, $|\bar{F} - F^{(k)}| \leq m - 2^{L-1}$. It follows from the induction hypothesis that after another

$$\lceil 4(m - 2^{L-1})(\rho_+(1)/\rho_-(s)) \ln(2\mu) \rceil$$

OMP iterations, (6) holds. Therefore by combining this estimate with (8), we know that the total number of OMP iterations for (6) to hold (starting with $F^{(0)}$) is no more than

$$\begin{aligned} & \lceil 4(m - 2^{L-1})(\rho_+(1)/\rho_-(s)) \ln(2\mu) \rceil + 2^{L+1}(\rho_+(1)/\rho_-(s)) \ln(2\mu) - 1 \\ \leq & \lceil 4m(\rho_+(1)/\rho_-(s)) \ln(2\mu) \rceil. \end{aligned}$$

This finishes the induction step for the case $|\bar{F} - F^{(0)}| = m$. ■