Multi-Label Prediction via Compressed Sensing

Daniel Hsu
UC San Diego
djhsu@cs.ucsd.edu

Sham M. Kakade
TTI-Chicago
sham@tti-c.org

John Langford
Yahoo! Research
jl@hunch.net

Tong Zhang
Rutgers University
tongz@rci.rutgers.edu

Abstract

We consider multi-label prediction problems with large output spaces under the assumption of *output sparsity* – that the target (label) vectors have small support. We develop a general theory for a variant of the popular error correcting output code scheme, using ideas from compressed sensing for exploiting this sparsity. The method can be regarded as a simple reduction from multi-label regression problems to binary regression problems. We show that the number of subproblems need only be logarithmic in the total number of possible labels, making this approach radically more efficient than others. We also state and prove robustness guarantees for this method in the form of regret transform bounds (in general), and also provide a more detailed analysis for the linear prediction setting.

1 Introduction

Suppose we have a large database of images, and we want to learn to predict who or what is in any given one. A standard approach to this task is to collect a sample of these images x along with corresponding labels $y = (y_1, \ldots, y_d) \in \{0,1\}^d$, where $y_i = 1$ if and only if person or object i is depicted in image x, and then feed the labeled sample to a multi-label learning algorithm. Here, d is the total number of entities depicted in the entire database. When d is very large (*e.g.* 10^3, 10^4), the simple one-against-all approach of learning a single predictor for each entity can become prohibitively expensive, both at training and testing time.

Our motivation for the present work comes from the observation that although the output (label) space may be very high dimensional, the actual labels are often sparse. That is, in each image, only a small number of entities may be present and there may only be a small amount of ambiguity in who or what they are. In this work, we consider how this sparsity in the output space, or *output sparsity*, eases the burden of large-scale multi-label learning.

Exploiting output sparsity. A subtle but critical point that distinguishes output sparsity from more common notions of sparsity (say, in feature or weight vectors) is that we are interested in the sparsity of $\mathbb{E}[y|x]$ rather than y. In general, $\mathbb{E}[y|x]$ may be sparse while the actual outcome y may not (*e.g.* if there is much unbiased noise); and, vice versa, y may be sparse with probability one but $\mathbb{E}[y|x]$ may have large support (*e.g.* if there is little distinction between several labels).

Conventional linear algebra suggests that we must predict d parameters in order to find the value of the d-dimensional vector $\mathbb{E}[y|x]$ for each x. A crucial observation – central to the area of compressed sensing [1] – is that methods exist to recover $\mathbb{E}[y|x]$ from just $O(k \log d)$ measurements when $\mathbb{E}[y|x]$ is k-sparse. This is the basis of our approach.

Our contributions. We show how to apply algorithms for compressed sensing to the output coding approach [2]. At a high level, the output coding approach creates a collection of subproblems of the form "Is the label in this subset or its complement?", solves these problems, and then uses their solution to predict the final label.

The role of compressed sensing in our application is distinct from its more conventional uses in data compression. Although we do employ a sensing matrix to compress training data, we ultimately are not interested in recovering data explicitly compressed this way. Rather, we *learn to predict compressed label vectors*, and then use sparse reconstruction algorithms to *recover uncompressed labels from these predictions*. Thus we are interested in reconstruction accuracy of predictions, averaged over the data distribution.

The main contributions of this work are:

1. A formal application of compressed sensing to prediction problems with output sparsity.

2. An efficient output coding method, in which the number of required predictions is only logarithmic in the number of labels d, making it applicable to very large-scale problems.

3. Robustness guarantees, in the form of regret transform bounds (in general) and a further detailed analysis for the linear prediction setting.

Prior work. The ubiquity of multi-label prediction problems in domains ranging from multiple object recognition in computer vision to automatic keyword tagging for content databases has spurred the development of numerous general methods for the task. Perhaps the most straightforward approach is the well-known one-against-all reduction [3], but this can be too expensive when the number of possible labels is large (especially if applied to the power set of the label space [4]). When structure can be imposed on the label space (*e.g.* class hierarchy), efficient learning and prediction methods are often possible [5, 6, 7, 8, 9]. Here, we focus on a different type of structure, namely output sparsity, which is not addressed in previous work. Moreover, our method is general enough to take advantage of structured notions of sparsity (*e.g.* group sparsity) when available [10]. Recently, heuristics have been proposed for discovering structure in large output spaces that empirically offer some degree of efficiency [11].

As previously mentioned, our work is most closely related to the class of output coding method for multi-class prediction, which was first introduced and shown to be useful experimentally in [2]. Relative to this work, we expand the scope of the approach to multi-label prediction and provide bounds on regret and error which guide the design of codes. The loss based decoding approach [12] suggests decoding so as to minimize loss. However, it does not provide significant guidance in the choice of encoding method, or the feedback between encoding and decoding which we analyze here.

The output coding approach is inconsistent when classifiers are used and the underlying problems being encoded are noisy. This is proved and analyzed in [13], where it is also shown that using a Hadamard code creates a robust consistent predictor when reduced to binary regression. Compared to this method, our approach achieves the same robustness guarantees up to a constant factor, but requires training and evaluating exponentially (in d) fewer predictors.

Our algorithms rely on several methods from compressed sensing, which we detail where used.

2 Preliminaries

Let $\mathcal{X}$ be an arbitrary input space and $\mathcal{Y} \subset \mathbb{R}^d$ be a d-dimensional output (label) space. We assume the data source is defined by a fixed but unknown distribution over $\mathcal{X} \times \mathcal{Y}$. Our goal is to learn a predictor $F : \mathcal{X} \to \mathcal{Y}$ with low expected ℓ_2^2-error $\mathbb{E}_x \|F(x) - \mathbb{E}[y|x]\|_2^2$ (the sum of mean-squared-errors over all labels) using a set of n training data $\{(x_i, y_i)\}_{i=1}^n$.

We focus on the regime in which the output space is very high-dimensional (d very large), but for any given $x \in \mathcal{X}$, the expected value $\mathbb{E}[y|x]$ of the corresponding label $y \in \mathcal{Y}$ has only a few non-zero entries. A vector is k-*sparse* if it has at most k non-zero entries.

3 Learning and Prediction

3.1 Learning to Predict Compressed Labels

Let $A : \mathbb{R}^d \to \mathbb{R}^m$ be a linear compression function, where $m \leq d$ (but hopefully $m \ll d$). We use A to compress (*i.e.* reduce the dimension of) the labels $\mathcal{Y}$, and learn a predictor $H : \mathcal{X} \to A(\mathcal{Y})$ of these compressed labels. Since A is linear, we simply represent $A \in \mathbb{R}^{m \times d}$ as a matrix.

Specifically, given a sample $\{(x_i, y_i)\}_{i=1}^n$, we form a compressed sample $\{(x_i, Ay_i)\}_{i=1}^n$ and then learn a predictor H of $\mathbb{E}[Ay|x]$ with the objective of minimizing the ℓ_2^2-error $\mathbb{E}_x \|H(x) - \mathbb{E}[Ay|x]\|_2^2$.

3.2 Predicting Sparse Labels

To obtain a predictor F of $\mathbb{E}[y|x]$, we compose the predictor H of $\mathbb{E}[Ay|x]$ (learned using the compressed sample) with a reconstruction algorithm $\mathsf{R} : \mathbb{R}^m \to \mathbb{R}^d$. The algorithm R maps predictions of compressed labels $h \in \mathbb{R}^m$ to predictions of labels $y \in \mathcal{Y}$ in the original output space. These algorithms typically aim to find a sparse vector y such that Ay closely approximates h.

Recent developments in the area of compressed sensing have produced a spate of reconstruction algorithms with strong performance guarantees when the compression function A satisfies certain properties. We abstract out the relevant aspects of these guarantees in the following definition.

Definition. An algorithm R is a *valid reconstruction algorithm for a family of compression functions* $(\mathcal{A}_k \subset \bigcup_{m \geq 1} \mathbb{R}^{m \times d} : k \in \mathbb{N})$ *and sparsity error* $\mathrm{sperr} : \mathbb{N} \times \mathbb{R}^d \to \mathbb{R}$, if there exists a function $f : \mathbb{N} \to \mathbb{N}$ and constants $C_1, C_2 \in \mathbb{R}$ such that: on input $k \in \mathbb{N}$, $A \in \mathcal{A}_k$ with m rows, and $h \in \mathbb{R}^m$, the algorithm $\mathsf{R}(k, A, h)$ returns an $f(k)$-sparse vector $\widehat{y}$ satisfying

$$\|\widehat{y} - y\|_2^2 \;\leq\; C_1 \cdot \|h - Ay\|_2^2 + C_2 \cdot \mathrm{sperr}(k, y)$$

for all $y \in \mathbb{R}^d$. The function f is the *output sparsity* of R and the constants C_1 and C_2 are the *regret factors*.

Informally, if the predicted compressed label $H(x)$ is close to $\mathbb{E}[Ay|x] = A\mathbb{E}[y|x]$, then the sparse vector $\widehat{y}$ returned by the reconstruction algorithm should be close to $\mathbb{E}[y|x]$; this latter distance $\|\widehat{y} - \mathbb{E}[y|x]\|_2^2$ should degrade gracefully in terms of the accuracy of $H(x)$ and the sparsity of $\mathbb{E}[y|x]$. Moreover, the algorithm should be agnostic about the sparsity of $\mathbb{E}[y|x]$ (and thus the sparsity error $\mathrm{sperr}(k, \mathbb{E}[y|x])$), as well as the "measurement noise" (the prediction error $\|H(x) - \mathbb{E}[Ay|x]\|_2$). This is a subtle condition and precludes certain reconstruction algorithm (*e.g.* Basis Pursuit [14]) that require the user to supply a bound on the measurement noise. However, the condition is needed in our application, as such bounds on the prediction error (for each x) are not generally known beforehand.

We make a few additional remarks on the definition.

1. The minimum number of rows of matrices $A \in \mathcal{A}_k$ may in general depend on k (as well as the ambient dimension d). In the next section, we show how to construct such A with close to the optimal number of rows.
2. The sparsity error $\mathrm{sperr}(k, y)$ should measure how poorly $y \in \mathbb{R}^d$ is approximated by a k-sparse vector.
3. A reasonable output sparsity $f(k)$ for sparsity level k should not be much more than k, *e.g.* $f(k) = O(k)$.

Concrete examples of valid reconstruction algorithms (along with the associated $\mathcal{A}_k$, sperr, etc.) are given in the next section.

4 Algorithms

Our prescribed recipe is summarized in Algorithms 1 and 2. We give some examples of compression functions and reconstruction algorithms in the following subsections.

Algorithm 1 Training algorithm

parameters sparsity level k, compression function $A \in \mathcal{A}_k$ with m rows, regression learning algorithm L

input training data $S \subset \mathcal{X} \times \mathbb{R}^d$

  for $i = 1, \ldots, m$ **do**

    $h_i \leftarrow \mathsf{L}(\{(x, (Ay)_i) : (x, y) \in S\})$

  end for

output regressors $H = [h_1, \ldots, h_m]$

Algorithm 2 Prediction algorithm

parameters sparsity level k, compression function $A \in \mathcal{A}_k$ with m rows, valid reconstruction algorithm R for $\mathcal{A}_k$

input regressors $H = [h_1, \ldots, h_m]$, test point $x \in \mathcal{X}$

output $\widehat{y} = \vec{\mathsf{R}}(k, A, [h_1(x), \ldots, h_m(x)])$

Figure 1: Training and prediction algorithms.

4.1 Compression Functions

Several valid reconstruction algorithms are known for compression matrices that satisfy a *restricted isometry property*.

Definition. A matrix $A \in \mathbb{R}^{m \times d}$ satisfies the (k, δ)-*restricted isometry property* (*(k, δ)-RIP*), $\delta \in (0, 1)$, if $(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2$ for all k-sparse $x \in \mathbb{R}^d$.

While some explicit constructions of (k, δ)-RIP matrices are known (*e.g.* [15]), the best guarantees are obtained when the matrix is chosen randomly from an appropriate distribution, such as one of the following [16, 17].

- All entries i.i.d. Gaussian $N(0, 1/m)$, with $m = O(k \log(d/k))$.
- All entries i.i.d. Bernoulli $B(1/2)$ over $\{\pm 1/\sqrt{m}\}$, with $m = O(k \log(d/k))$.
- m randomly chosen rows of the $d \times d$ Hadamard matrix over $\{\pm 1/\sqrt{m}\}$, with $m = O(k \log^5 d)$.

The hidden constants in the big-O notation depend inversely on δ and the probability of failure.

A striking feature of these constructions is the very mild dependence of m on the ambient dimension d. This translates to a significant savings in the number of learning problems one has to solve after employing our reduction.

Some reconstruction algorithms require a stronger guarantee of bounded *coherence* $\mu(A) \leq O(1/k)$, where $\mu(A)$ defined as

$$\mu(A) = \max_{1 \leq i < j \leq d} |(A^\top A)_{i,j}| / \sqrt{|(A^\top A)_{i,i}||(A^\top A)_{j,j}|}$$

It is easy to check that the Gaussian, Bernoulli, and Hadamard-based random matrices given above have coherence bounded by $O(\sqrt{(\log d)/m})$ with high probability. Thus, one can take $m = O(k^2 \log d)$ to guarantee $1/k$ coherence. This is a factor k worse than what was needed for (k, δ)-RIP, but the dependence on d is still small.

4.2 Reconstruction Algorithms

In this section, we give some examples of valid reconstruction algorithms. Each of these algorithm is valid with respect to the sparsity error given by

$$\mathrm{sperr}(k, y) \;=\; \|y - y_{(1:k)}\|_2^2 + \frac{1}{k}\|y - y_{(1:k)}\|_1^2$$

where $y_{(1:k)}$ is the best k-sparse approximation of y (*i.e.* the vector with just the k largest (in magnitude) coefficients of y).

The following theorem relates reconstruction quality to approximate sparse regression, giving a sufficient condition for any algorithm to be valid for RIP matrices.

Algorithm 3 Prediction algorithm with $\mathsf{R} = \text{OMP}$

parameters sparsity level k, compression function $A = [a_1|\dots|a_d] \in \mathcal{A}_k$ with m rows,
input regressors $H = [h_1, \dots, h_m]$, test point $x \in \mathcal{X}$
  $h \leftarrow [h_1(x), \dots, h_m(x)]^\top$ (predict compressed label vector)
  $\widehat{y} \leftarrow \vec{0}$, $J \leftarrow \emptyset$, $r \leftarrow h$
  for $i = 1, \dots, 2k$ **do**
    $j_* \leftarrow \arg\max_j |r^\top a_j| / \|a_j\|_2$ (column of A most correlated with residual r)
    $J \leftarrow J \cup \{j_*\}$ (add j_* to set of selected columns)
    $\widehat{y}_J \leftarrow (A_J)^\dagger h$, $\widehat{y}_{J^c} \leftarrow \vec{0}$ (least-squares restricted to columns in J)
    $r \leftarrow h - A\widehat{y}$ (update residual)
  end for
output $\widehat{y}$

Figure 2: Prediction algorithm specialized with Orthogonal Matching Pursuit.

Theorem 1. *Let $\mathcal{A}_k = \{(k + f(k), \delta)$-RIP matrices$\}$ for some function $f : \mathbb{N} \to \mathbb{N}$, and let $A \in \mathcal{A}_k$ have m rows. If for any $h \in \mathbb{R}^m$, a reconstruction algorithm R returns an $f(k)$-sparse solution $\widehat{y} = \mathsf{R}(k, A, h)$ satisfying*

$$\|A\widehat{y} - h\|_2^2 \le \inf_{y \in \mathbb{R}^d} C\|Ay_{(1:k)} - h\|_2^2,$$

then it is a valid reconstruction algorithm for $\mathcal{A}_k$ and sperr given above, with output sparsity f and regret factors $C_1 = 2(1 + \sqrt{C})^2/(1 - \delta)$ and $C_2 = 4(1 + (1 + \sqrt{C})/(1 - \delta))^2$.

Proofs are deferred to Appendix B.

Iterative and greedy algorithms. Orthogonal Matching Pursuit (OMP) [18], FoBa [19], and CoSaMP [20] are examples of iterative or greedy reconstruction algorithms. OMP is a greedy forward selection method that repeatedly selects a new column of A to use in fitting h (see Algorithm 3). FoBa is similar, except it also incorporates backward steps to un-select columns that are later discovered to be unnecessary. CoSaMP is also similar to OMP, but instead selects larger sets of columns in each iteration.

FoBa and CoSaMP are valid reconstruction algorithms for RIP matrices ($(8k, 0.1)$-RIP and $(4k, 0.1)$-RIP, respectively) and have linear output sparsity ($8k$ and $2k$). These guarantees are apparent from the cited references. For OMP, we give the following guarantee.

Theorem 2. *If $\mu(A) \le 0.1/k$, then after $f(k) = 2k$ steps of OMP, the algorithm returns $\widehat{y}$ satisfying*

$$\|A\widehat{y} - h\|_2^2 \le 23\|Ay_{(1:k)} - h\|_2^2 \quad \forall y \in \mathbb{R}^d.$$

This theorem, combined with Theorem 1, implies that OMP is valid for matrices A with $\mu(A) \le 0.1/k$ and has output sparsity $f(k) = 2k$.

ℓ_1 algorithms. Basis Pursuit (BP) [14] and its variants are based on finding the minimum ℓ_1-norm solution to a linear system. While the basic form of BP is ill-suited for our application (it requires the user to supply the amount of measurement error $\|Ay - h\|_2$), its more advanced path-following or multi-stage variants may be valid [21].

5 Analysis

5.1 General Robustness Guarantees

We now state our main regret transform bound, which follows immediately from the definition of a valid reconstruction algorithm and linearity of expectation.

Theorem 3 (Regret Transform)**.** *Let R be a valid reconstruction algorithm for $\{\mathcal{A}_k : k \in \mathbb{N}\}$ and $\mathrm{sperr} : \mathbb{N} \times \mathbb{R}^d \to \mathbb{R}$. Then there exists some constants C_1 and C_2 such that the following holds.*

Pick any $k \in \mathbb{N}$, $A \in \mathcal{A}_k$ with m rows, and $H : \mathcal{X} \to \mathbb{R}^m$. Let $F : \mathcal{X} \to \mathbb{R}^d$ be the composition of $\mathsf{R}(k, A, \cdot)$ *and H,* i.e. $F(x) = \mathsf{R}(k, A, H(x))$. *Then*

$$\mathbb{E}_x \|F(x) - \mathbb{E}[y|x]\|_2^2 \;\; \leq \;\; C_1 \cdot \mathbb{E}_x \|H(x) - \mathbb{E}[Ay|x]\|_2^2 \; + \; C_2 \cdot \mathrm{sperr}(k, \mathbb{E}[y|x]).$$

The simplicity of this theorem is a consequence of the careful composition of the learned predictors with the reconstruction algorithm meeting the formal specifications described above.

In order compare this regret bound with the bounds afforded by Sensitive Error Correcting Output Codes (SECOC) [13], we need to relate $\mathbb{E}_x \|H(x) - \mathbb{E}[Ay|x]\|_2^2$ to the average scaled mean-squared-error over all induced regression problems; the error is scaled by the maximum difference $L_i = \max_{y \in \mathcal{Y}} (Ay)_i - \min_y (Ay)_i$ between induced labels:

$$\bar{r} \;=\; \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_x \left(\frac{H(x)_i - \mathbb{E}[(Ay)_i|x]}{L_i} \right)^2 .$$

In k-sparse multi-label problems, we have $\mathcal{Y} = \{y \in \{0,1\}^d : \|y\|_0 \leq k\}$. In these terms, SECOC can be tuned to yield $\mathbb{E}_x \|F(x) - \mathbb{E}[y|x]\|_2^2 \;\leq\; 4k^2 \cdot \bar{r}$ for general k.

For now, ignore the sparsity error. For simplicity, let $A \in \mathbb{R}^{m \times d}$ with entries chosen i.i.d. from the Bernoulli $B(1/2)$ distribution over $\{\pm 1/\sqrt{m}\}$, where $m = O(k \log d)$. Then for any k-sparse y, we have $\|Ay\|_\infty \leq k/\sqrt{m}$, and thus $L_i \leq 2k/\sqrt{m}$ for each i. This gives the bound

$$C_1 \cdot \mathbb{E}_x \|H(x) - \mathbb{E}[Ay|x]\|_2^2 \;\leq\; 4C_1 \cdot k^2 \cdot \bar{r},$$

which is within a constant factor of the guarantee afforded by SECOC. Note that our reduction induces exponentially (in d) fewer subproblems than SECOC.

Now we consider the sparsity error. In the extreme case $m = d$, $\mathbb{E}[y|x]$ is allowed to be fully dense ($k = d$) and $\mathrm{sperr}(k, \mathbb{E}[y|x]) = 0$. When $m = O(k \log d) < d$, we potentially incur an extra penalty in $\mathrm{sperr}(k, \mathbb{E}[y|x])$, which relates how far $\mathbb{E}[y|x]$ is from being k-sparse. For example, suppose $\mathbb{E}[y|x]$ has small ℓ_p norm for $0 \leq p < 2$. Then even if $\mathbb{E}[y|x]$ has full support, the penalty will decrease polynomially in $k \approx m/\log d$.

5.2 Linear Prediction

A danger of using generic reductions is that one might create a problem instance that is even harder to solve than the original problem. This is an oft cited issue with using output codes for multi-class problems. In the case of linear prediction, however, the danger is mitigated, as we now show. Suppose, for instance, there is a perfect linear predictor of $\mathbb{E}[y|x]$, *i.e.* $\mathbb{E}[y|x] = B^\top x$ for some $B \in \mathbb{R}^{p \times d}$ (here $\mathcal{X} = \mathbb{R}^p$). Then it is easy to see that $H = BA^\top$ is a perfect linear predictor of $\mathbb{E}[Ay|x]$:

$$H^\top x \;=\; AB^\top x \;=\; A\mathbb{E}[y|x] \;=\; \mathbb{E}[Ay|x].$$

The following theorem generalizes this observation to imperfect linear predictors for certain well-behaved A.

Theorem 4. *Suppose $\mathcal{X} \subset \mathbb{R}^p$. Let $B \in \mathbb{R}^{p \times d}$ be a linear function with*

$$\mathbb{E}_x \left\| B^\top x - \mathbb{E}[y|x] \right\|_2^2 \;=\; \epsilon.$$

Let $A \in \mathbb{R}^{m \times d}$ have entries drawn i.i.d. from $N(0, 1/m)$, and let $H = BA^\top$. Then with high probability (over the choice of A),

$$\mathbb{E}_x \|H^\top x - A\mathbb{E}[y|x]\|_2^2 \;\leq\; \left(1 + O(1/\sqrt{m})\right) \epsilon.$$

Remark 5. *Similar guarantees can be proven for the Bernoulli-based matrices. Note that d does not appear in the bound, which is in contrast to the expected spectral norm of A: roughly $1 + O(\sqrt{d/m})$.*

Theorem 4 implies that the errors of *any* linear predictor are not magnified much by the compression function. So a good linear predictor for the original problem implies an almost-as-good linear predictor for the induced problem. Using this theorem together with known results about linear prediction [22], it is straightforward to derive sample complexity bounds for achieving a given error relative to that of the best linear predictor in some class. The bound will depend polynomially in k but only logarithmically in d. This is cosmetically similar to learning bounds for feature-efficient algorithms (*e.g.* [23, 22]) which are concerned with sparsity in the weight vector, rather than in the output.

6 Experimental Validation

We conducted an empirical assessment of our proposed reduction on two labeled data sets with large label spaces. These experiments demonstrate the feasibility of our method – a sanity check that the reduction does in fact preserve learnability – and compare different compression and reconstruction options.

6.1 Data

Image data.[1] The first data set was collected by the ESP Game [24], an online game in which players ultimately provide word tags for a diverse set of web images.

The set contains nearly 68000 images, with about 22000 unique labels. We retained just the 1000 most frequent labels: the least frequent of these occurs 39 times in the data, and the most frequent occurs about 12000 times. Each image contains about four labels on average. We used half of the data for training and half for testing.

We represented each image as a bag-of-features vector in a manner similar to [25]. Specifically, we identified 1024 representative SURF features points [26] from 10×10 gray-scale patches chosen randomly from the training images; this partitions the space of image patches (represented with SURF features) into Voronoi cells. We then built a histogram for each image, counting the number of patches that fall in each cell.

Text data.[2] The second data set was collected by Tsoumakas et al. [11] from `del.icio.us`, a social bookmarking service in which users assign descriptive textual tags to web pages.

The set contains about 16000 labeled web page and 983 unique labels. The least frequent label occurs 21 times and the most frequent occurs almost 6500 times. Each web page is assigned 19 labels on average. Again, we used half the data for training and half for testing.

Each web page is represented as a boolean bag-of-words vector, with the vocabulary chosen using a combination of frequency thresholding and χ^2 feature ranking. See [11] for details.

Each binary label vector (in both data sets) indicates the labels of the corresponding data point.

6.2 Output Sparsity

We first performed a bit of exploratory data analysis to get a sense of how sparse the target in our data is. We computed the least-squares linear regressor $\widehat{B} \in \mathbb{R}^{p \times d}$ on the training data (without any output coding) and predicted the label probabilities $\widehat{p}(x) = \widehat{B}^\top x$ on the test data (clipping values to the range $[0, 1]$). Using $\widehat{p}(x)$ as a surrogate for the actual target $\mathbb{E}[y|x]$, we examined the relative ℓ_2^2 error of $\widehat{p}$ and its best k-sparse approximation $\epsilon(k, \widehat{p}(x)) = \sum_{i=k+1}^{d} \widehat{p}_{(i)}(x)^2 / \|\widehat{p}(x)\|_2^2$, where $\widehat{p}_{(1)}(x) \geq \ldots \geq \widehat{p}_{(d)}(x)$.

Examining $\mathbb{E}_x \epsilon(k, \widehat{p}(x))$ as a function of k, we saw that in both the image and text data, the fall-off with k is eventually super-polynomial, but we are interested in the behavior for small k where it appears polynomial k^{-r} for some r. Around $k = 10$, we estimated an exponent of 0.50 for the image data and 0.55 for the text data. This is somewhat below the standard of what is considered sparse (*e.g.* vectors with small ℓ_1-norm show k^{-1} decay). Thus, we expect the reconstruction algorithms will have to contend with the sparsity error of the target.

6.3 Procedure

We used least-squares linear regression as our base learning algorithm, with no regularization on the image data and with ℓ_2-regularization with the text data ($\lambda = 0.01$) for numerical stability. We did not attempt any parameter tuning.

[1] `http://hunch.net/~learning/ESP-ImageSet.tar.gz`
[2] `http://mlkd.csd.auth.gr/multilabel.html`

The compression functions we used were generated by selecting m random rows of the 1024×1024 Hadamard matrix, for $m \in \{100, 200, 300, 400\}$. We also experimented with Gaussian matrices; these yielded similar but uniformly worse results.

We tested the greedy and iterative reconstruction algorithms described earlier (OMP, FoBa, and CoSaMP) as well as a path-following version of Lasso based on LARS [21]. Each algorithm was used to recover a k-sparse label vector $\widehat{y}^k$ from the predicted compressed label $H(x)$, for $k = 1, \ldots, 10$. We measured the ℓ_2^2 distance $\|\widehat{y}^k - y\|_2^2$ of the prediction to the true test label y. In addition, we measured the precision of the predicted support at various values of k using the 10-sparse label prediction. That is, we ordered the coefficients of each 10-sparse label prediction $\widehat{y}^{10}$ by magnitude, and measured the precision of predicting the first k coordinates $|\operatorname{supp}(\widehat{y}^{10}_{(1:k)}) \cap \operatorname{supp}(y)|/k$. Actually, for $k \geq 6$, we used $\widehat{y}^{2k}$ instead of $\widehat{y}^{10}$.

We used correlation decoding (CD) as a baseline method, as it is a standard decoding method for ECOC approaches. CD predicts using the top k coordinates in $A^\top H(x)$, ordered by magnitude. For mean-squared-error comparisons, we used the least-squares approximation of $H(x)$ using these k columns of A. Note that CD is not a valid reconstruction algorithm when $m < d$.

6.4 Results

As expected, the performance of the reduction, using any reconstruction algorithm, improves as the number of induced subproblems m is increased (see figures in Appendix A) When m is small and $A \notin \mathcal{A}_K$, the reconstruction algorithm cannot reliably choose $k \geq K$ coordinates, so its performance may degrade after this point by over-fitting. But when the compression function A is in $\mathcal{A}_K$ for a sufficiently large K, then the squared-error decreases as the output sparsity k increases up to K. Note the fact that precision-at-k decreases as k increases is expected, as fewer data will have at least k correct labels.

All of the reconstruction algorithms at least match or out-performed the baseline on the mean-squared-error criterion, except when $m = 100$. When A has few rows, (1) $A \in \mathcal{A}_K$ only for very small K, and (2) many of its columns will have significant correlation. In this case, when choosing $k > K$ columns, it is better to choose correlated columns to avoid over-fitting. Both OMP and FoBa explicitly avoid this and thus do not fare well; but CoSaMP, Lasso, and CD do allow selecting correlated columns and thus perform better in this regime.

The results for precision-at-k are similar to that of mean-squared-error, except that choosing correlated columns does not necessarily help in the small m regime. This is because the extra correlated columns need not correspond to accurate label coordinates.

In summary, the experiments demonstrate the feasibility and robustness of our reduction method for two natural multi-label prediction tasks. They show that predictions of relatively few compressed labels are sufficient to recover an accurate sparse label vector, and as our theory suggests, the robustness of the reconstruction algorithms is a key factor in their success.

Acknowledgments

We thank Andy Cotter for help processing the image features for the ESP Game data. This work was completed while the first author was an intern at TTI-C in 2008.

References

[1] David Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, 2006.

[2] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[3] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, (5):101–141, 2004.

[4] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[5] A. Clare and R.D. King. Knowledge discovery in multi-label phenotype data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, 2001.

[6] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2003.

[7] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Incremental algorithms for hierarchical classification. *Journal of Machine Learning Research*, 7:31–54, 2006.

[8] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[9] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, (7):1601–1626, 2006.

[10] J. Huang, T. Zhang, and D. Metaxax. Learning with structured sparsity. In *ICML*, 2009.

[11] G. Tsoumakas, I. Katakis, and I. Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, 2008.

[12] Erin Allwein, Robert Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[13] J. Langford and A. Beygelzimer. Sensitive error correcting output codes. In *Proc. Conference on Learning Theory*, 2005.

[14] Emmanuel Candès, Justin Romberg, and Terrence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–122, 2006.

[15] R. DeVore. Deterministic constructions of compressed sensing matrices. *J. of Complexity*, 23:918–925, 2007.

[16] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

[17] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *Proc. Conference on Information Sciences and Systems*, 2006.

[18] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[19] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Proc. Neural Information Processing Systems*, 2008.

[20] D. Needell and J.A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 2007.

[21] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[22] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proc. Neural Information Processing Systems*, 2008.

[23] Andrew Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *ICML*, 2004.

[24] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proc. ACM Conference on Human Factors in Computing Systems*, 2004.

[25] Marcin Marszałek, Cordelia Schmid, Hedi Harzallah, and Joost van de Weijer. Learning object representations for visual object class recognition. In *Visual Recognition Challange Workshop, in conjunction with ICCV*, 2007.

[26] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

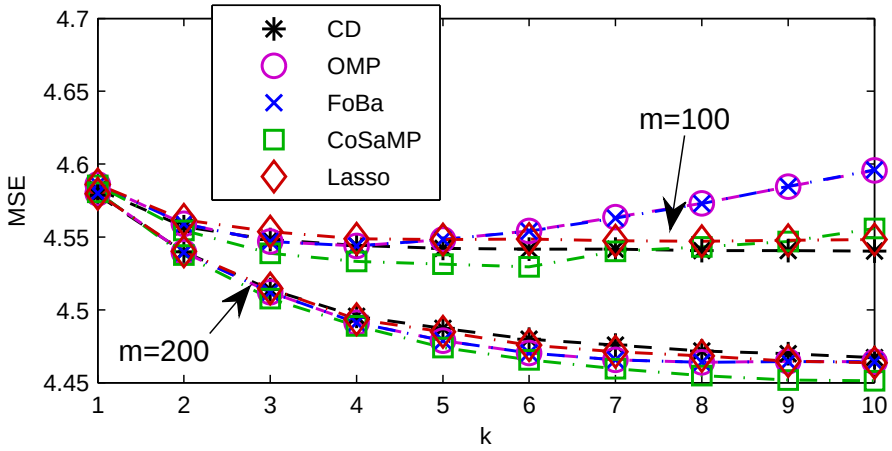
[27] David Donoho, Michael Elad, and Vladimir Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info. Theory*, 52(1):6–18, 2006.

[28] Sanjoy Dasgupta. *Learning Probability Distributions*. PhD thesis, University of California, 2000.

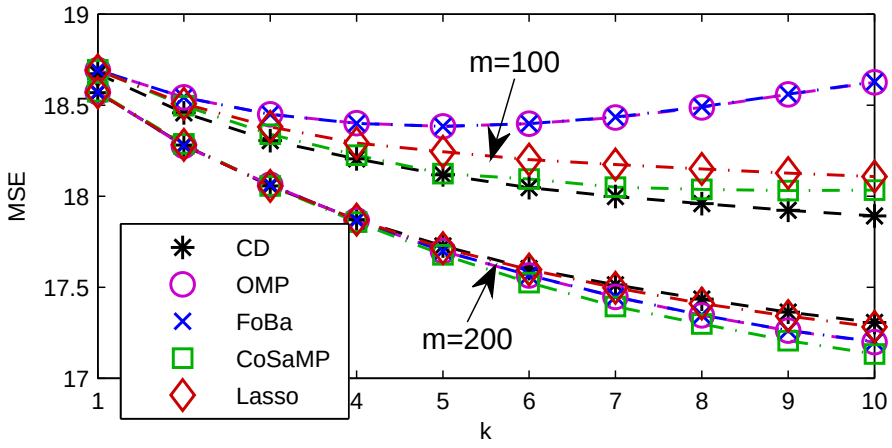
A Figures from Experimental Results

In each plot, the top set of lines corresponds to $m = 100$, and the bottom set to $m = 200$. At $m = 300, 400$, the performance is nearly the same as one-against-all, *i.e.* $m = 1024$, so we omit these plots.

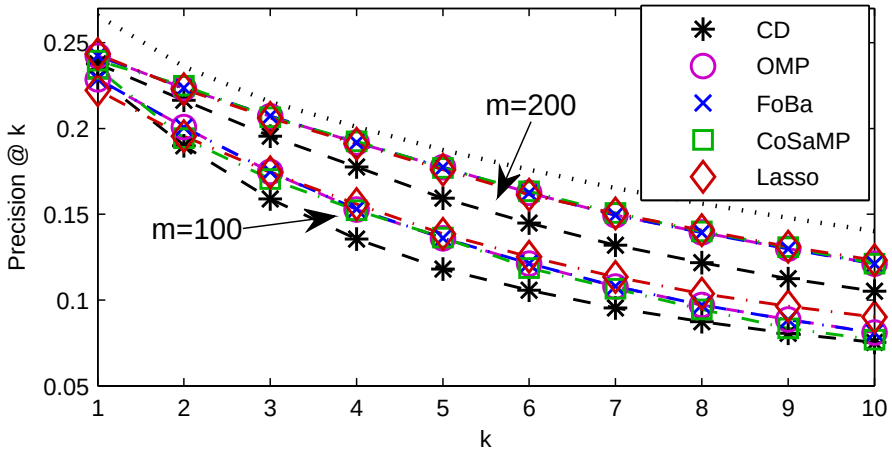
Mean-squared-error versus output sparsity k, $m \in \{100, 200\}$, image data.



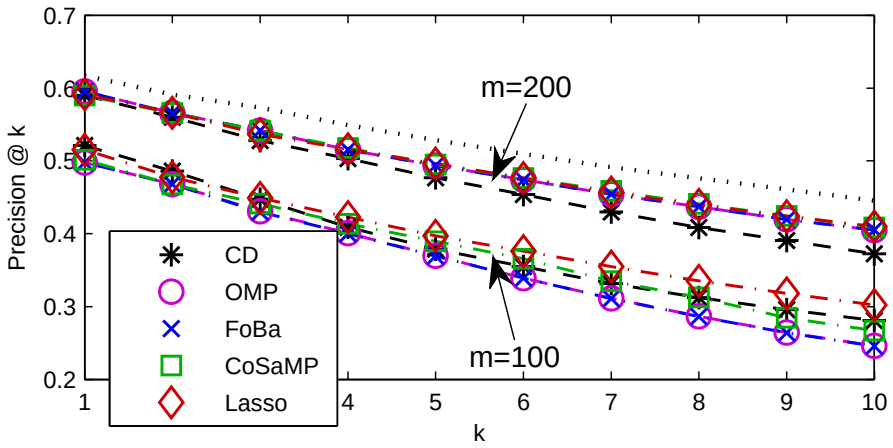
Mean-squared-error versus output sparsity k, $m \in \{100, 200\}$, text data.



Mean precision-at-k versus output sparsity k, $m \in \{100, 200\}$, image data.



Mean precision-at-k versus output sparsity k, $m \in \{100, 200\}$, text data.



B Proofs

B.1 Proof of Theorem 1

Let $\ell = k + f(k)$, $y \in \mathbb{R}^d$, and assume without loss of generality that $|y_1| \geq \ldots \geq |y_d|$. We need to show that

$$\|\widehat{y} - y\|_2^2 \;\leq\; C_1 \cdot \|Ay - h\|_2^2 \;+\; C_2 \cdot (\|\Delta\|_2^2 + k^{-1}\|\Delta\|_1^2)$$

where $\Delta = y - y_{(1:k)}$. Using the triangle inequality, the (ℓ, δ)-RIP of $A \in \mathcal{A}_k$, and the hypothesis that $\|A\widehat{y} - h\|_2^2 \leq C\|Ay_{(1:k)} - h\|_2^2$, we have

$$\begin{aligned}
\|\widehat{y} - y\|_2 &\leq \|\widehat{y} - y_{(1:k)}\|_2 + \|\Delta\|_2 \\
&\leq (1-\delta)^{-1/2}\|A\widehat{y} - Ay_{(1:k)}\|_2 + \|\Delta\|_2 \\
&\leq (1-\delta)^{-1/2}\left(\|A\widehat{y} - h\|_2 + \|h - Ay_{(1:k)}\|_2\right) + \|\Delta\|_2 \\
&\leq (1-\delta)^{-1/2}(1+\sqrt{C})\|Ay_{(1:k)} - h\|_2 + \|\Delta\|_2 \\
&\leq (1-\delta)^{-1/2}(1+\sqrt{C})\left(\|Ay - h\|_2 + \|A\Delta\|_2\right) + \|\Delta\|_2. \qquad (1)
\end{aligned}$$

We need to relate $\|A\Delta\|_2$ to $\|\Delta\|_2$ and $\|\Delta\|_1$. Write $\Delta = \sum_{i\geq 0} y_{J_i}$, where $J_i = \{k+i\ell+1, \ldots, k+(i+1)\ell\}$ and $y_J \in \mathbb{R}^d$ is the vector whose jth component is y_j if $j \in J$ and is 0 otherwise. Note that each y_{J_i} is ℓ-sparse, $\|y_{J_{i+1}}\|_1 \leq \|y_{J_i}\|_1$, and $\|y_{J_{i+1}}\|_\infty \leq \ell^{-1}\|y_{J_i}\|_1$. By Hölder's inequality,

$$\|y_{J_{i+1}}\|_2 \;\leq\; (\|y_{J_{i+1}}\|_\infty \|y_{J_{i+1}}\|_1)^{1/2} \;\leq\; (\ell^{-1}\|y_{J_i}\|_1^2)^{1/2} \;=\; \ell^{-1/2}\|y_{J_i}\|_1,$$

and so

$$\sum_{i\geq 0}\|y_{J_i}\|_2 \;\leq\; \|y_{J_0}\|_2 + \sum_{i\geq 0}\|y_{J_{i+1}}\|_2 \;\leq\; \|y_{J_0}\|_2 + \ell^{-1/2}\sum_{i\geq 0}\|y_{J_i}\|_1 \;\leq\; \|\Delta\|_2 + \ell^{-1/2}\|\Delta\|_1.$$

By the triangle inequality and the (ℓ, δ)-RIP of A, we have

$$\|A\Delta\|_2 \;\leq\; \sum_{i\geq 0}\|Ay_{J_i}\|_2 \;\leq\; \sum_{i\geq 0}(1+\delta)^{1/2}\|y_{J_i}\|_2 \;\leq\; (1+\delta)^{1/2}(\|\Delta\|_2 + \ell^{-1/2}\|\Delta\|_1).$$

Combining this final inequality with (1) gives

$$\|\widehat{y} - y\|_2 \;\leq\; C_0 \cdot \|Ay - h\|_2 \;+\; (1 + C_0(1+\delta)^{1/2}) \cdot (\|\Delta\|_2 + \ell^{-1/2}\|\Delta\|_1)$$

where $C_0 = (1-\delta)^{-1/2}(1+\sqrt{C})$. Now squaring both sides and simplifying using the fact $(x+y)^2 \leq 2x^2 + 2y^2$ concludes the proof.

B.2 Proof of Theorem 2

We first begin with two simple lemmas.

Lemma 6. *Suppose OMP is run for k iterations starting with $y^{(0)} = \vec{0}$, and produces intermediate solutions $y^{(1)}, y^{(2)}, \ldots, y^{(k)}$. Then there exists some $0 \le i < k$ such that if j_i is the column selected in step i, then $(a_{j_i}^\top (h - Ay^{(i)}))^2 \le \|h\|_2^2/k$.*

Proof. Let $r^{(i)} = h - Ay^{(i)}$. Suppose column j_i is added to J in step i. Let $\widetilde{y}^{(i+1)} = y^{(i)} + \alpha_i e_{j_i}$, where $\alpha_i = a_{j_i}^\top r^{(i)}$ and e_{j_i} is the j_ith elementary vector. Then

$$\begin{aligned}\|r^{(i)}\|_2^2 - \|r^{(i+1)}\|_2^2 \ge \|r^{(i)}\|_2^2 - \|h - A\widetilde{y}^{(i+1)}\|_2^2 \;&=\; \|r^{(i)}\|_2^2 - \|h - A(y^{(i)} + \alpha_i e_{j_i})\|_2^2 \\ &= \|r^{(i)}\|_2^2 - \|r^{(i)} - \alpha_i a_{j_i}\|_2^2 \;=\; 2\alpha_i a_{j_i}^\top r^{(i)} - \alpha_i^2 \|a_{j_i}\|_2^2 \;=\; (a_{j_i}^\top r^{(i)})^2.\end{aligned}$$

Moreover, $\sum_{i=0}^{k-1} \|r^{(i)}\|_2^2 - \|r^{(i+1)}\|_2^2 = \|r^{(0)}\|_2^2 - \|r^{(k)}\|_2^2 \le \|h\|_2^2$, so there is some $i \in \{0, 1, \ldots, k-1\}$ such that $(a_{j_i}^\top r^{(i)})^2 \le \|r^{(i)}\|_2^2 - \|r^{(i+1)}\|_2^2 \le \|h\|_2^2/k$. ☐

Lemma 7. *If $y \in \mathbb{R}^d$ is k-sparse and $\mu(A) \le \delta/(k-1)$, then $\|Ay\|_2^2 \ge (1-\delta)\|y\|_2^2$.*

This result also appears in Appendix A1 of [27]. We reproduce the proof here.

Proof. Expanding $\|Ay\|_2^2$, we have

$$\|Ay\|_2^2 \;=\; \sum_{i=1}^{k} \|a_i\|^2 y_i^2 + \sum_{i \ne j} y_i y_j (a_i^\top a_j) \;\ge\; \|y\|_2^2 - \left| \sum_{i \ne j} y_i y_j (a_i^\top a_j) \right|,$$

so we need to show this latter summation is at most $\delta\|y\|_2^2$. Indeed,

$$\begin{aligned}\left| \sum_{i \ne j} y_i y_j (a_i^\top a_j) \right| &\le \sum_{i \ne j} |y_i y_j| |a_i^\top a_j| && \text{(triangle inequality)} \\ &\le \mu(A) \sum_{i \ne j} |y_i y_j| && \text{(definition of coherence)} \\ &= \mu(A) \left(\sum_{i=1}^{k} \sum_{j=1}^{k} |y_i||y_j| - \sum_{i=1}^{k} y_i^2 \right) \\ &= \mu(A) \left(\left(\sum_{i=1}^{k} |y_i| \right)^2 - \|y\|_2^2 \right) \\ &\le \mu(A)(k\|y\|_2^2 - \|y\|_2^2) && \text{(Cauchy-Schwarz)} \\ &= \mu(A)(k-1)\|y\|_2^2 \\ &\le \delta\|y\|_2^2 && \text{(assumption on } \mu(A))\end{aligned}$$

which concludes the proof. ☐

We are now ready to prove Theorem 2. Without loss of generality, we assume that the columns of $A = [a_1|\ldots|a_d]$ are normalized (so $\|a_j\|_2 = 1$) and that the support of y is (some subset of) $\{1, \ldots, k\}$ (so y is k-sparse).

In addition to the vector $\widehat{y}$ returned by OMP and the vector y we want to compare to, we consider two other solution vectors:

- y': a $(2k-1)$-sparse solution obtained by running up to $k-1$ iterations of OMP starting from y. Lemma 6 implies that there exists such a vector y' with the following property: if j^* is the column OMP would select when the current solution is y', then

$$(a_{j^*}^\top (h - Ay'))^2 \quad\le\quad \|h - Ay\|_2^2/k. \tag{2}$$

Since y' is obtained by starting with y, it can only have smaller squared-error than y. Without loss of generality, let the support of y' be (some subset of) $\{1, \ldots, 2k\}$.

- $\widehat{y}'$: the actual solution produced by OMP (starting from $\vec{0}$) just before OMP chooses a column $j \notin \text{supp}(y')$. Note that if OMP never chooses a column $j \notin \text{supp}(y')$ within $2k$ steps, then $\|A\widehat{y} - h\|_2^2 \leq \|A\widehat{y}' - h\|_2^2 \leq \|Ay - h\|_2^2$ and the theorem is proven. Therefore we assume that this event does occurs and so $\widehat{y}'$ is defined. Since $\widehat{y}'$ precedes the final solution $\widehat{y}$ returned by OMP, it can only have larger squared-error than $\widehat{y}$.

We will bound $\|h - A\widehat{y}\|_2$ as follows:

$$\begin{aligned}
\|h - A\widehat{y}\|_2 &\leq \|h - A\widehat{y}'\|_2 && \text{(since } \widehat{y}' \text{ precedes } \widehat{y}\text{)} \\
&\leq \|h - Ay'\|_2 + \|A(\widehat{y}' - y')\|_2 && \text{(triangle inequality)} \\
&\leq \|h - Ay\|_2 + \|A(\widehat{y}' - y')\|_2. && \text{(since } y \text{ precedes } y'\text{)}
\end{aligned}$$

We thus need to bound $\|A(\widehat{y}' - y')\|_2$ in terms of $\|h - Ay\|_2$.

Let $\widehat{r} = h - A\widehat{y}'$ and $r = h - Ay'$. Then

$$\begin{aligned}
\|A(\widehat{y}' - y')\|_2^2 &= (A\widehat{y}' - Ay')^\top A(\widehat{y}' - y') \\
&= (h - Ay')^\top A(\widehat{y}' - y') - (h - A\widehat{y}')^\top A(\widehat{y}' - y') \\
&\leq \|h - Ay'\|_2 \|A(\widehat{y}' - y')\|_2 + |(h - A\widehat{y}')^\top A(\widehat{y}' - y')| \quad \text{(Cauchy-Schwarz)} \\
&= \|r\|_2 \|A(\widehat{y}' - y')\|_2 + |\widehat{r}^\top A(\widehat{y}' - y')|.
\end{aligned}$$

Using the fact $x \leq b\sqrt{x} + c \Rightarrow x \leq (4/3)(b^2 + c)$ (which in turn follows from the quadratic formula and the fact $2xy \leq x^2 + y^2$), the above inequality implies

$$\frac{3}{4}\|A(\widehat{y}' - y')\|_2^2 \quad \leq \quad \|r\|_2^2 + |\widehat{r}^\top A(\widehat{y}' - y')|. \tag{3}$$

We now work on bounding the second term on the righthand side. Let $j > 2k$ be the column chosen by OMP when the current solution is $\widehat{y}'$. Then we have

$$|a_j^\top \widehat{r}| \quad \geq \quad |a_\ell^\top \widehat{r}| \quad \forall \ell \leq 2k. \tag{4}$$

Also, since $\widehat{y}' - y'$ has support $\{1, \ldots, 2k\}$, we have that

$$A(\widehat{y}' - y') \quad = \quad A_{\{1:2k\}}(\widehat{y}' - y') \tag{5}$$

where $A_{\{1:2k\}}$ is the same as A except with zeros in all but the first $2k$ columns. Then,

$$\begin{aligned}
|\widehat{r}^\top A(\widehat{y}' - y')| &= |\widehat{r}^\top A_{\{1:2k\}}(\widehat{y}' - y')| && \text{(Equation (5))} \\
&\leq \|\widehat{r}^\top A_{\{1:2k\}}\|_\infty \|\widehat{y}' - y'\|_1 && \text{(Hölder's inequality)} \\
&\leq |a_j^\top \widehat{r}| \|\widehat{y}' - y'\|_1 && \text{(Inequality (4))} \\
&\leq \left(|a_j^\top r| + |a_j^\top A(\widehat{y}' - y')|\right) \|\widehat{y}' - y'\|_1 && \text{(triangle inequality)} \\
&\leq \left(|a_j^\top r| + \|a_j^\top A_{\{1:2k\}}\|_\infty \|\widehat{y}' - y'\|_1\right) \|\widehat{y}' - y'\|_1 && \text{(Equation (5) and Hölder)} \\
&\leq |a_j^\top r| \|\widehat{y}' - y'\|_1 + \mu(A)\|\widehat{y}' - y'\|_1^2 && \text{(definition of coherence)} \\
&\leq \frac{5k}{2}(a_j^\top r)^2 + \frac{1}{10k}\|\widehat{y}' - y'\|_1^2 + \mu(A)\|\widehat{y}' - y'\|_1^2 && \text{(since } xy \leq (x^2 + y^2)/2\text{)} \\
&\leq \frac{5k}{2}(a_j^\top r)^2 + \frac{1}{5k}\|\widehat{y}' - y'\|_1^2 && \text{(since } \mu(A) \leq 0.1/k\text{)} \\
&\leq \frac{5k}{2}(a_j^\top r)^2 + \frac{1}{5k}(2k\|\widehat{y}' - y'\|_2^2) && \text{(Cauchy-Schwarz)} \\
&\leq \frac{5k}{2}(a_j^\top r)^2 + \frac{1}{2}\|A(\widehat{y}' - y')\|_2^2. && \text{(Lemma 7)}
\end{aligned}$$

Continuing from Inequality (3), we have

$$\|A(\widehat{y}' - y')\|_2^2 \quad \leq \quad 4\|r\|_2^2 + 10k(a_j^\top r)^2.$$

Since $(a_j^\top r)^2 \leq (a_{j^*}^\top r)^2$, where $j^* \leq 2k$ is the column that OMP would select when the current solution is y', and since $(a_{j^*}^\top r)^2 \leq \|h - Ay\|_2^2/k$ (by Inequality (2)), we have that

$$\begin{aligned} \|A(\widehat{y}' - y')\|_2^2 &\leq 4\|r\|_2^2 + 10\|h - Ay\|_2^2 \\ &\leq 14\|h - Ay\|_2^2. \end{aligned}$$

Therefore,

$$\|h - A\widehat{y}'\|_2 \leq (1 + \sqrt{14})\|h - Ay\|_2.$$

Squaring both sides gives the conclusion.

B.3 Proof of Theorem 4

We use the following Chernoff bound for sums of χ^2 random variables, a proof of which can be found in the Appendix A of [28].

Lemma 8. *Fix any $\lambda_1 \geq \ldots \geq \lambda_D > 0$, and let $X_1, \ldots, X_D$ be i.i.d. χ^2 random variables with one degree of freedom. Then $\Pr[\sum_{i=1}^D \lambda_i X_i > (1+\gamma)\sum_{i=1}^D \lambda_i] \leq \exp(-(D\gamma^2/24) \cdot (\lambda/\lambda_1))$ for any $0 < \gamma < 1$, where $\lambda = (\lambda_1 + \ldots + \lambda_D)/D$.*

Write $A = (1/\sqrt{m})[\theta_1 | \cdots | \theta_m]^\top$, where each θ_i is an independent d-dimensional Gaussian random vector $N(0, I_d)$. Define $v_x = B^\top x - \mathbb{E}[y|x]$ so $\epsilon = \mathbb{E}_x\|v_x\|_2^2$, and assume without loss of generality that v_x has full d-dimensional support. Using this definition and linearity of expectation, we have

$$\mathbb{E}_x\|Av_x\|_2^2 = \frac{1}{m}\mathbb{E}_x \sum_{i=1}^m (\theta_i^\top v_x)^2 = \frac{1}{m}\sum_{i=1}^m \theta_i^\top (\mathbb{E}_x v_x v_x^\top)\theta_i.$$

Our goal is to show that this quantity is $(1 + O(1/\sqrt{m}))\epsilon$ with high probability. Since $N(0, I_d)$ is rotationally invariant and $\mathbb{E}_x v_x v_x^\top$ is symmetric and positive definite, we may assume $\mathbb{E}_x v_x v_x^\top$ is diagonal and has eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d > 0$. Then, the above expression simplifies to

$$\frac{1}{m}\sum_{i=1}^m \theta_i^\top (\mathbb{E}_x v_x v_x^\top)\theta_i = \frac{1}{m}\sum_{i=1}^m \sum_{j=1}^d \lambda_j \theta_{ij}^2.$$

Each θ_{ij}^2 is a χ^2 random variable with one degree of freedom, so $\mathbb{E}\theta_{ij}^2 = 1$. Thus, the expected value of the above quantity is $\sum_{j=1}^d \operatorname{trace}(\mathbb{E}_x v_x v_x^\top) = \mathbb{E}_x \operatorname{trace}(v_x v_x^\top) = \mathbb{E}_x\|v_x\|_2^2$. Now applying Lemma 8, with $D = md$ variables and $\lambda = (\lambda_1 + \ldots + \lambda_d)/d$, we have $\Pr[(1/m)\sum_{i,j} \lambda_j \theta_{ij}^2 > (1+t)\epsilon] \leq \exp(-(mdt^2/24)(\lambda/\lambda_1)) \leq \exp(-mt^2/24)$ (using the fact $\lambda_1 \leq d\lambda$). This bound is δ when $t = \sqrt{(24/m)\ln(1/\delta)}$.