THE 2-CODEWORD SCREENING TEST FOR LASSO PROBLEMS

Hao Wu

Hong Kong University of Science and Technology
Dept. of Computer Science and Engineering
Kowloon, Hong Kong

Peter J. Ramadge

Princeton University
Dept. of Electrical Engineering
Princeton, NJ

ABSTRACT

Solving a lasso problem is a practical approach for acquiring a sparse representation of a signal with respect to a given dictionary. Driven by the demand for sparse representations over large-scale data in machine learning and statistics, we explore lasso screening tests. These enhance solution efficiency via the elimination of codewords absent in the optimal solution prior to detailed computation. On basis of the concept of a region test and the recently introduced dome test, we propose *the 2-codeword test*, which uses two codewords together in a correlation screening test. In addition to the rejection rate as the performance measure, we introduce an innovative way to access the performance of a screening test, called the uncertainty measure, via a comparison with the optimal test.

***Index Terms*—** Optimization, Algorithms, Machine learning

1. INTRODUCTION

Given $\mathbf{x}, \mathbf{b}_i \in \mathbb{R}^p, i = 1, ..., m$, the lasso problem is defined as follows:

$$\min_{w_1,w_2,...,w_m} \frac{1}{2}\|\mathbf{x} - \sum_{i=1}^{m} w_i \mathbf{b}_i\|_2^2 + \lambda \sum_{i=1}^{m} |w_i|. \tag{1}$$

We assume $\|\mathbf{x}\|_2 = \|\mathbf{b}_i\|_2 = 1, i = 1, ..., m$. The lasso problem finds a representation of $\mathbf{x}$ over the set of m known codewords $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_m]$ with the sparsity of the representation controlled by the λ weighted ℓ_1 penalty. The optimizer $\tilde{\mathbf{w}}$ is used as a sparse feature vector for the corresponding $\mathbf{x}$ in subsequent processing tasks such as classification. Difficult problems in computer vision such as object recognition [1] seem amenable to this form of sparse representation.

The convex optimization (1) becomes more time consuming to solve as the size of the dictionary m grows large. This is quite common in applications such as face recognition [2]. To address this challenge, screening tests have been introduced to reduce the dictionary size prior to solving a lasso problem [3, 4, 5, 6, 7, 8, 9, 10]. For each target vector $\mathbf{x}$, the screening test efficiently identifies a set of $\mathbf{b}_i$ that receive zero weights $\tilde{w_i}$ in the optimal solution of (1). By removing these codewords, the size m of the dictionary $\mathbf{B}$ is reduced without affecting the optimal solution. Thus the performance of the algorithm for solving the lasso problem is improved.

To achieve such improvement, the test should be efficient in terms of time, e.g., linear in the dimension of data p and the size of the dictionary m, and it should not alter the optimal solution. In [3], the SAFE rule rejects $\mathbf{b}_i$ when $|\mathbf{x}^T\mathbf{b}_i| < \lambda + \lambda/\lambda_{\max} - 1$ where $\lambda_{\max} = \max_i |\mathbf{x}^T\mathbf{b}_i|$. In [5], three Sphere Tests (ST) are introduced. ST1 and ST2 place a threshold on $|\mathbf{x}^T\mathbf{b}_i|$ while ST3 investigates also using $|\mathbf{b}_*^T\mathbf{b}_i|$ where $\mathbf{b}_* = \arg\max_{\mathbf{b}_i} = |\mathbf{x}^T\mathbf{b}_i|$. These three tests attempt to bound the optimal solution $\tilde{\boldsymbol{\theta}}$ of the dual problem of (1) within a sphere (hence the term sphere test). In [6], the bounding region is reshaped into a dome by cutting the bounding sphere using a hyperplane defined by the codeword $\mathbf{b}_*$. The resulting test yields improved codeword rejection.

Following [10], in this paper we employ a more general notion of the dome test used in [6]. This is described in §3.1. Moreover, we use an additional codeword to construct a new sphere to bound the dual solution (§3.2). This procedure has the possibility of being applied recursively. This leads to a *2-codeword test* which has exhibits improved rejection power and computational time. This will be illustrated in §4. A new metric called the uncertainty measure is used to assess the performance of a screening test. This measures the potential room for improvement of a particular screening test.

2. PRELIMINARIES

Consider the Lagrangian dual problem of (1):

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda}\|_2^2 \\ s.t. \quad & |\boldsymbol{\theta}^T\mathbf{b}_i| \leq 1 \quad \forall i = 1, 2, ..., m. \end{aligned} \tag{2}$$

The solution of the primal form $\tilde{\mathbf{w}}$ and that of the dual $\tilde{\boldsymbol{\theta}}$ are related through:

$$\mathbf{x} = \sum_{i=1}^{m} \tilde{w}_i \mathbf{b}_i + \lambda\tilde{\boldsymbol{\theta}}, \ \tilde{\boldsymbol{\theta}}^T\mathbf{b}_i \in \begin{cases} \{\text{sign } \tilde{w_i}\} & \text{if} \quad \tilde{w_i} \neq 0, \\ [-1, 1] & \text{if} \quad \tilde{w_i} = 0. \end{cases} \tag{3}$$

A detailed derivation of these claims is available in the supplementary material of [5].

Observing the form of (3), one straightforward screening method called the *core rejection test* is given by:

$$\textbf{Reject } \mathbf{b}_i: \text{ if } |\tilde{\boldsymbol{\theta}}^T \mathbf{b}_i| < 1. \tag{4}$$

However, executing (4), requires knowing $\tilde{\boldsymbol{\theta}}$ in advance, which is not feasible since this is equivalent to solving the lasso problem. Despite this fact, the core rejection test serves as a benchmark for evaluating screening tests since it rejects every $\mathbf{b}_i$ with $\tilde{w}_i = 0$ (see (3)). We only consider screening tests which are *compliant* with the core rejection test. A test T_1 is compliant with T_2 if T_1 only rejects codewords that are rejected by T_2. Compliance with the core rejection test implies that the test only rejects $\mathbf{b}_i$ with $\tilde{w}_i \neq 0$. Tests mentioned in §1 are all compliant with the core rejection test.

Some geometric insight can also be deduced from (2). In the p-dimensional space containing $\mathbf{x}$ and $\mathbf{b}_i \in \mathbb{R}^p$, $i = 1, \ldots, m$, we seek a point $\tilde{\boldsymbol{\theta}}$ which is feasible for the dual problem and is closest to $\frac{\mathbf{x}}{\lambda}$. The feasible region of the dual problem (2) is determined by the set of linear constraints in (2) corresponding to hyperplanes perpendicular to each $\mathbf{b}_i$. The optimal dual solution $\tilde{\boldsymbol{\theta}}$ must lie in this region.

A screening test can be identified with a region $\mathcal{R}$ in this space and compliance can be interpreted in terms of an inclusive relationship between regions:

Lemma 1. *Let $\mathcal{R}_1$,$\mathcal{R}_2$ be two regions, $\mathcal{R}_1 \subset \mathcal{R}_2 \implies T_{\mathcal{R}_2}$ is compliant with $T_{\mathcal{R}_1}$.*

For the core rejection test, the region $\mathcal{R}$ is just the point $\tilde{\boldsymbol{\theta}}$. This confirms that the core rejection test $T_{\{\tilde{\boldsymbol{\theta}}\}}$ is optimal since the region can't be further reduced.

The dome test in [6] first bounds $\tilde{\boldsymbol{\theta}}$ within a sphere $(\mathbf{q}, r_0)$:

$$\mathcal{B} = \{\boldsymbol{\theta} : \|\mathbf{q} - \boldsymbol{\theta}\|_2 \leq r_0\} \tag{5}$$

where $\mathbf{q} = \mathbf{x}/\lambda$ and $r_0 = 1/\lambda - 1/\lambda_{\max}$. To see why $\mathcal{B}$ bounds $\tilde{\boldsymbol{\theta}}$, note that since $\forall i : |\mathbf{x}^T \mathbf{b}_i/\lambda_{\max}| \leq 1$, $\mathbf{x}/\lambda_{\max}$ is a feasible solution to the dual problem (2). Hence the distance from $\tilde{\boldsymbol{\theta}}$ to $\mathbf{x}/\lambda$ is at most the distance $\|\mathbf{x}/\lambda - \mathbf{x}/\lambda_{\max}\|_2 = r_0$. So $\tilde{\boldsymbol{\theta}}$ is contained in a ball centered at $\mathbf{x}/\lambda$ with radius $r_0 = 1/\lambda - 1/\lambda_{\max}$. Now augment this bound with the constraint $\boldsymbol{\theta}^T \mathbf{b}_* \leq 1$. This leads to the dome bound:

$$\mathcal{G} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \mathbf{x}/\lambda\|_2 \leq r_0, \mathbf{b}_*^T \boldsymbol{\theta} \leq 1\}. \tag{6}$$

With the help of Lemma 1, it is clear that this dome is compliant. We now proceed one step further: an additional codeword is used to construct the 2-codeword test based on the region defined by:

$$\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \mathbf{x}/\lambda\_2 \leq r_0, \mathbf{b}_1^T \boldsymbol{\theta} \leq 1, \mathbf{b}_2^T \boldsymbol{\theta} \leq 1\}. \tag{7}$$

Since the bounding region $\mathcal{C}$ is more constrained, this 2-codeword test has the potential to achieve a high rejection rate, thus reducing the time for solving the lasso problem.

3. METHODOLOGY

In this section, we first state the dome test in a more general form. Then we introduce an additional codeword and bound the region so formed by a much smaller sphere compared to that used in [5]. Strategies to select the codewords for screening are also discussed.

3.1. The general dome test

Given a sphere centered at $\mathbf{q}$ with radius r and a codeword $\mathbf{b}$ on the unit-sphere, a general dome is given by the region:

$$\mathcal{D} = \{\mathbf{z} : \mathbf{z}^T \mathbf{b} \leq 1, \|\mathbf{z} - \mathbf{q}\|_2 \leq r\}. \tag{8}$$

The corresponding dome test $T_{\mathcal{D}(\mathbf{q},r,\mathbf{b})}$ is specified by the following closed form rule.

Lemma 2. *$T_{\mathcal{D}(\mathbf{q},r,\mathbf{b})}$ rejects codeword $\mathbf{b}_i$ if*

$$V_l(\mathbf{b}^T \mathbf{b}_i) < \mathbf{q}^T \mathbf{b}_i < V_u(\mathbf{b}^T \mathbf{b}_i),$$

where $V_l(t)$ and $V_u(t)$ are defined as:

$$V_l(t) = \begin{cases} -1 + r(\psi_d t + \sqrt{(1 - \psi_d^2)(1 - t^2)}) & if\ t \leq \psi_d, \\ -(1 - r) & if\ t > \psi_d; \end{cases}$$

$$V_u(t) = \begin{cases} (1 - r) & if\ t < -\psi_d, \\ 1 + r(\psi_d t - \sqrt{(1 - \psi_d^2)(1 - t^2)}) & if\ t \geq -\psi_d; \end{cases}$$

with $\psi_d = (\mathbf{q}^T \mathbf{b} - 1)/r$.

Lemma 2 is a generalization of Theorem 1 in [6]. The latter is obtained by setting $\mathbf{q} = \mathbf{x}/\lambda$ and $\mathbf{b} = \mathbf{b}_* = \arg\max_{\mathbf{b}_i} = \mathbf{x}^T \mathbf{b}_i$. A proof of Lemma 2 is given in [10].

We observe that a better bounding sphere can be obtained by supplying a dual solution $\tilde{\boldsymbol{\theta}}_0$ of another lasso problem with sparsity parameter λ_0 (typically $\lambda_0 > \lambda$) for the same $\mathbf{x}$ and $\mathbf{B}$. The point $\tilde{\boldsymbol{\theta}}_0$ is a feasible point in the target lasso problem since $\forall i\ \boldsymbol{\theta}_0^T \mathbf{b}_i \leq 1$ holds. This can potentially reduce the sphere radius r_0 to $\|\mathbf{x}/\lambda_{\max} - \tilde{\boldsymbol{\theta}}_0\|$. On the other hand, solving an additional lasso problem may be time consuming and the gain may not be able to compensate for the loss in time. This trade-off will be discussed in §4.

3.2. The 2-codeword test

We now describe the construction of the 2-codeword test. Consider a bounding sphere centered at $\mathbf{q}$ with radius r and two codewords $\mathbf{b}_1$ and $\mathbf{b}_2$. Performing a test directly on the region defined by the sphere and two hyperplanes is complex (see[11]). An alternative is to first bound the dome formed by the sphere and codeword $\mathbf{b}_1$ within a new sphere. This is shown as the green sphere in Fig. 1. The parameters (center $\mathbf{q}_d$ and radius r) of this new sphere are

$$\mathbf{q}_d = \mathbf{q} - \psi_d \mathbf{b}_1, \quad r_d = r\sqrt{1 - \psi_d^2}, \tag{9}$$

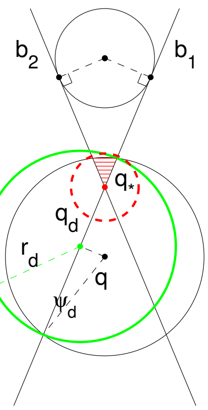

Fig. 1. The hyperplane of the codeword $\mathbf{b}_1$ cuts the sphere $\mathbf{q}$, producing a new sphere $(\mathbf{q}_d, r_d)$ with reduction factor ψ_d. Then the codeword $\mathbf{b}_2$ further cuts $\mathbf{q}_d$, acquiring the sphere $\mathbf{q}_*$ identified in Theorem. 1.

with $\psi_d = (\mathbf{q}^T\mathbf{b}_1 - 1)/r$ as defined in Lemma 2. Upon obtaining the new sphere $(\mathbf{q}_d, r_d)$, we perform a second dome test $T_{\mathcal{D}(\mathbf{q}_d,r_d,\mathbf{b}_2)}$ using the second codeword $\mathbf{b}_2$.

To select $\mathbf{b}_1$ and $\mathbf{b}_2$, we use a greedy strategy (see [10]) where $\mathbf{b}_1 = \arg\max_{\mathbf{b}} \mathbf{b}^T\mathbf{q}$ and $\mathbf{b}_2 = \arg\max_{\mathbf{b}} \mathbf{b}^T\mathbf{q}_d$. This has the advantage of simplicity and efficiency.

Finally, we construct a second new sphere $(\mathbf{q}_*, r_*)$ to bound the region defined by sphere $(\mathbf{q}, r)$ and the two codewords $\mathbf{b}_1, \mathbf{b}_2$ as described in (7). This region is illustrated in Fig.1 by the region shaded in red and the bounding sphere is illustrated by the dotted red circle.

We can now state the following result.

Theorem 1. *Define the sphere* $(\mathbf{q}_*, r_*)$ *as follows,*

$$\begin{aligned}\mathbf{q}_* &= \mathbf{q} - w_1\mathbf{b}_1 + w_2\mathbf{b}_2\\ r_* &= \sqrt{r^2 + 2(w_1+w_2) - |\mathbf{q}|^2 + |\mathbf{q}_*|^2}.\end{aligned} \tag{10}$$

w_1, w_2 *are constants depending on* $\mathbf{b}_1, \mathbf{b}_2$ *and* $\mathbf{q}$*:*

$$\begin{aligned}w_1 &= (\mathbf{b}_1^T\mathbf{q} - \mathbf{b}_1^T\mathbf{b}_2\mathbf{b}_2^T\mathbf{q} + \mathbf{b}_1^T\mathbf{b}_2 - 1)/(1-(\mathbf{b}_1^T\mathbf{b}_2)^2),\\ w_2 &= (\mathbf{b}_2^T\mathbf{q} - \mathbf{b}_1^T\mathbf{b}_2\mathbf{b}_1^T\mathbf{q} + \mathbf{b}_1^T\mathbf{b}_2 - 1)/(1-(\mathbf{b}_1^T\mathbf{b}_2)^2).\end{aligned}$$

Then this sphere bounds the region $\mathcal{C}$ *defined by* (7).

Proof. We are free to select $\mathbf{q}_*$ and r_* provided we bound the desired region. First consider the optimization problem:

$$\mathbf{q}_* = \arg\min_{\boldsymbol{\theta}} \|\mathbf{q} - \boldsymbol{\theta}\|_2, \boldsymbol{\theta} \in \mathcal{C}. \tag{11}$$

Letting $\mathbf{x}_0 = \mathbf{q}/\|\mathbf{q}\|_2$ and $\lambda_0 = 1/\|\mathbf{q}\|_2$, we can paraphrase (11) into the dual lasso problem:

$$\begin{aligned}\mathbf{q}_* = \arg\max_{\boldsymbol{\theta}} \quad & \frac{1}{2}\|\mathbf{x}_0\|_2^2 - \frac{\lambda_0^2}{2}\|\boldsymbol{\theta} - \frac{\mathbf{x}_0}{\lambda_0}\|_2^2,\\ \text{s.t.} \quad & |\boldsymbol{\theta}^T\mathbf{b}_i| \le 1, \quad i = 1,2\end{aligned} \tag{12}$$

The primal form of this problem is the lasso:

$$\min_{\mathbf{w}'} \quad \frac{1}{2}\|\mathbf{x}_0 - \mathbf{B}\mathbf{w}'\|_2^2 + \lambda_0\|\mathbf{w}'\|_1, \ \mathbf{B} = [\mathbf{b}_1 \ \mathbf{b}_2], \tag{13}$$

with the primal and dual solutions related by:

$$\mathbf{x}_0 = w_1'\mathbf{b}_1 + w_2'\mathbf{b}_2 + \lambda\boldsymbol{\theta}_0. \tag{14}$$

By changing the sign of $\mathbf{b}_i$ if necessary, we additionally assume $w_i' > 0$, $i = 1, 2$. Then we solve (13) by taking derivatives and setting these to zero. This yields

$$\begin{aligned}w_1' &= (\mathbf{b}_1^T\mathbf{x} - \mathbf{b}_1^T\mathbf{b}_2\mathbf{b}_2^T\mathbf{x} + \lambda_0\mathbf{b}_1^T\mathbf{b}_2 - \lambda_0)/(1-(\mathbf{b}_1^T\mathbf{b}_2)^2),\\ w_2' &= (\mathbf{b}_2^T\mathbf{x} - \mathbf{b}_1^T\mathbf{b}_2\mathbf{b}_1^T\mathbf{x} + \lambda_0\mathbf{b}_1^T\mathbf{b}_2 - \lambda_0)/(1-(\mathbf{b}_1^T\mathbf{b}_2)^2).\end{aligned}$$

Substituting the expressions for $\mathbf{x}_0$ and λ_0 and letting $w_1 = w_1'/\lambda_0, w_2 = w_2'/\lambda_0$, yields the first line of (10). Then we select r_* by considering the optimization problem:

$$\begin{aligned}r_*^2 &= \max_{\boldsymbol{\theta}} \ \|\boldsymbol{\theta} - \mathbf{q}_*\|^2,\\ \text{where } \mathbf{q}_* &= \mathbf{q} - w_1\mathbf{b}_1 - w_2\mathbf{b}_2,\\ &\boldsymbol{\theta}^T\mathbf{b}_1 \le 1, \boldsymbol{\theta}^T\mathbf{b}_2 \le 1, \|\boldsymbol{\theta} - \mathbf{q}\| \le r.\end{aligned} \tag{15}$$

By unwrapping the square, we have

$$\begin{aligned}&\|\boldsymbol{\theta} - \mathbf{q}_*\|^2 = \|\boldsymbol{\theta}\|^2 + \|\mathbf{q}_*\|^2 - 2\boldsymbol{\theta}^T\mathbf{q}_*\\ &= \|\boldsymbol{\theta}\|^2 - 2\boldsymbol{\theta}^T(\mathbf{q} - w_1\mathbf{b}_1 - w_2\mathbf{b}_2) + \|\mathbf{q}_*\|^2\\ &= \|\boldsymbol{\theta} - \mathbf{q}\|^2 + 2w_1\boldsymbol{\theta}^T\mathbf{b}_1 + 2w_2\boldsymbol{\theta}^T\mathbf{b}_2 - \|\mathbf{q}\|^2 + \|\mathbf{q}_*\|^2\\ &\le r^2 + 2(w_1+w_2) - \|\mathbf{q}\|^2 + \|\mathbf{q}_*\|^2.\end{aligned}$$

The last step is valid through enforcing all three constraints in (15). This yields the second line of (10) and completes the proof of Theorem 1. □

To summarize, we perform two general dome tests $T_{\mathcal{D}(\mathbf{q},r,\mathbf{b}_1)}$ and $T_{\mathcal{D}(\mathbf{q}_d,r_d,\mathbf{b}_2)}$, proceeding from sphere $(\mathbf{q}, r)$ to a new sphere $(\mathbf{q}_*, r_*)$. Hence, in principle, the test can be applied iteratively.

4. EXPERIMENTS

We look into the percentage of codewords being rejected and the computational time for performing tests and solving the reduced lasso problem, on three data sets: RAND (Randomized generated data set with size m=10,000 dictionary, 50 instances for testing, dimensionality p=28), MNIST500 (m=5,000, p=28*28, obtained from the first 500 images from the MNIST data set [12], with 50 testing images sampled from the test set) and YALEBXF (Yale B Extended Frontal, m=2,376,p=192*168, selected from 2,414 pictures of 38 objects published in [13], by sampling one picture for each object to be the test set). Instead of controlling λ for sparsity, we plot the rejection rate graph against $\lambda/\lambda_{\max}$ being the x-axis, where $\lambda_{\max}$ reflects the correlation between the

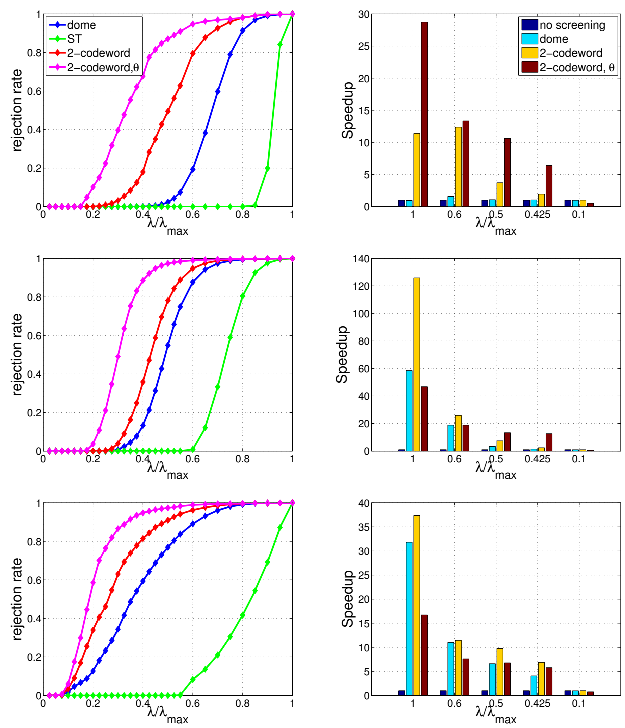

Fig. 2. Left column: rejection power varying with desired sparsity; Right column: speedup (times) compared to no screening; Rows: data set RAND, MNIST and YALEBXF

dictionary and the testing codeword. We compare the 2-codeword test with the dome test and the ST1/SAFE test, where we recursively perform two iterations.

We also look at the performance when supplied with a $\tilde{\boldsymbol{\theta}}_0$ obtained from the lasso with sparsity $\lambda/\lambda_{max}+0.2$ (1 is used if exceeding 1). As revealed in Fig. 2, in data set MNIST and YALEBXF, 2-codeword test outperforms by about 20% at the same sparsity level comparing to the dome test. With $\tilde{\boldsymbol{\theta}}_0$ supplied, the rejection rate is even higher. Involving $\tilde{\boldsymbol{\theta}}_0$ does not always improve the overall performance. By looking at $\lambda/\lambda_{\max}=0.425$ (the 4^{th} bar cluster in Speedup column), the 2-codeword test rejects about 30% and 50% in the first two data sets where $\tilde{\boldsymbol{\theta}}_0$ claims advantage in time, much lower than 85% in YALEBXF where the gain in rejection rate does not make up the loss in time to acquire $\tilde{\boldsymbol{\theta}}_0$. When $\lambda/\lambda_{\max}$ is small, the speedup tends to 1 due to the low rejection rate, where acquiring $\tilde{\boldsymbol{\theta}}_0$ appears to be much more costly, resulting a speedup less than 1. In all our experiments, we use the featuresign algorithm from [14] to solve the lasso problems.

4.1. The uncertainty measure

We also investigate how close we are to the optimal screening test, through the introduction of the *uncertainty measure*. Recall the core rejection test in (4), where the optimal dual solution is required. Here we loosen the criterion a bit and suppose we have restricted the $\tilde{\boldsymbol{\theta}}$ within a certain uncertainty range. We perform a family of reduced core rejection tests based on the rules $|\tilde{\boldsymbol{\theta}}^T\mathbf{b}_i| < 1-\epsilon|\tilde{\boldsymbol{\theta}}|$, where ϵ denotes the uncertainty ratio over the norm of $\tilde{\boldsymbol{\theta}}$, or the *uncertainty measure*. The uncertainty measure of a screening test at certain sparsity level indicates its ability in bounding $\tilde{\boldsymbol{\theta}}$. In Fig. 3, an intersecting point of 2-codeword curve and $\epsilon=15\%$ curve indicates that the 2-codeword test has uncertainty measure of 15% at $\lambda/\lambda_{max}=0.65$ on the MNIST data set. When $\tilde{\boldsymbol{\theta}}_0$ is supplied, the ability to locate $\tilde{\boldsymbol{\theta}}$ is improved since the starting sphere is much tighter. Notice that the uncertainty measure is not directly correlated with the rejection rate, making it more significant. On YALEBXF, the rejection rate 85% at $\lambda/\lambda_{max}=0.45$ has uncertainty measure less than 10%, while its counterpart on MNIST having uncertainty measure larger than 25%. Intuitively, it suggests that the 2-codeword test has more room to improve on MNIST than YALEBXF, which agrees with the fact that when supplied with $\tilde{\boldsymbol{\theta}}_0$, the 2-codeword test achieves larger improvement on MNIST compared to YALEBXF. By examining the behavior of the uncertainty measure, we can determine whether to supply a $\tilde{\boldsymbol{\theta}}_0$, or other further optimization such as another recursive step.

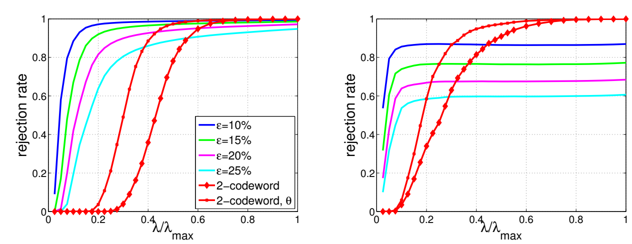

Fig. 3. Uncertainty Measure: Error range of approximating $\tilde{\boldsymbol{\theta}}$ on MNIST (left) and YALEBXF (right).

5. CONCLUSION

We have introduced the new 2-codeword screening test. By restricting the bounding region $\mathcal{R}$ into a much smaller sphere compared to the dome in [6], the 2-codeword test achieves significant improvement on the rejection rate, while the time complexity remains roughly the same as the dome test. The performance in terms of rejection rate is improved about 20% above the dome test at the same sparsity level. The 2-codeword test is faster (screen+solve) than previous tests, suggesting that the time loss in the extra tests is reclaimed by solving the lasso with a much smaller size dictionary, completing the task of a screening test in an efficient way. We also introduce the uncertainty measure to assess the performance of a screening test, and to decide whether further optimization is feasible.

6. REFERENCES

[1] T. Zhang K. Yu and Y. Gong, "Nonlinear learning using local coordinate coding," *Advances in Neural Information Processing Systems*, vol. 3, 2009.

[2] Arvind Ganesh Zihan Zhou Hossein Mobahi Andrew Wagner, John Wright and Yi Ma, "Towards a practical face recognition system: Robust alignment and illumination by sparse representation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2 2012, vol. 34.

[3] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," Tech. Rep. UCB/EECS-2010-126, EECS Department, University of California, Berkeley, Sep 2010.

[4] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Arxiv preprint arXiv:1011.2234*, 2010.

[5] Zhen J. Xiang, Hao Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, Eds., pp. 900–908. 2011.

[6] Zhen J. Xiang and P. J. Ramadge, "Fast lasso screening tests based on correlations," in *International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[7] Z. J. Xiang, "Combining structural knowledge with sparsity in machine learning and signal processing," *Ph.D. Thesis, Department of Electrical Engineering*, Aug. 2012.

[8] L. Dai and K. Pelckmans, "An ellipsoidal based, two-stage screening test for bpdn," in *20th European Signal Processing Conference*, Aug 2012.

[9] J. Jie Wang, B. Lin, P. Gong, P. Wonka, and J. Ye, "Lasso screening rules via dual polytope projection," *arXiv:1211.3966v1 [cs.LG]*, Nov. 2012.

[10] Zhen J. Xiang, Yun Wang, and P. J. Ramadge, "Screening tests for lasso problems," Tech. Rep., Princeton University, 2012.

[11] Y. Wang, Zhen J. Xiang, and P. J. Ramadge, "Tradeoffs in improved screening of lasso problems," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, 2013.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[13] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[14] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., pp. 801–808. MIT Press, Cambridge, MA, 2007.