TRADEOFFS IN IMPROVED SCREENING OF LASSO PROBLEMS

Yun Wang, Zhen James Xiang and Peter J. Ramadge

Department of Electrical Engineering, Princeton University

ABSTRACT

Recently, methods of screening the lasso problem have been developed that use the target vector $\mathbf{x}$ to quickly identify a subset of columns of the dictionary that will receive zero weight in the solution. Current classes of screening tests are based on bounding the dual lasso solution within a sphere or the intersection of a sphere and a half space. Stronger tests are possible but are more complex and incur a higher computational cost. To investigate this, we determine the optimal screening test when the dual lasso solution is bounded within the intersection of a sphere and two half spaces, and empirically investigate the trade-off that this test makes between screening power and computational efficiency. We also compare its performance both in terms of rejection power and efficiency to existing test classes. The new test always has better rejection, and for an interesting range of regularization parameters, offers better computational efficiency.

Index Terms— sparsity, screening, lasso problem

1. INTRODUCTION

We consider the lasso problem [1]:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \quad {}^1\!/_2\|\mathbf{x}-B\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. We call $\mathbf{x}$ the *target vector*, $B \in \mathbb{R}^{n\times p}$ the *dictionary* and its columns *codewords*. We will assume that $\mathbf{x}$ and the codewords $\mathbf{b}_i$ are normalized so that $\|\mathbf{x}\|_2 = 1$ and $\|\mathbf{b}_i\|_2 = 1$, $i = 1,\dots,p$.

Problem (1) seeks a sparse representation of $\mathbf{x} \in \mathbb{R}^n$ as a linear combination of a subset of the codewords in the dictionary $B \in \mathbb{R}^{n\times p}$. The ℓ_1 regularization encourages sparsity in a solution $\tilde{\mathbf{w}}$ (i.e., many components of $\tilde{\mathbf{w}}$ are 0). A solution $\tilde{\mathbf{w}}$ of (1) gives a nonlinear representation of $\mathbf{x}$ which can then be used in subsequent stages of data analysis. In a variety of applications ranging from signal processing to machine learning, this has proved to be a very effective means of data representation [2–10]. However, many of today's large-scale datasets pose a computational challenge for this method: the dictionary may be too big to fit into memory at once, it may take too long to solve (1) using allotted resources, and one may need to solve a large set of such problems. This poses a computation bottleneck on the method's applicability to large-scale problems.

To address this challenge, a screening test can use $\mathbf{x}$ to first determine a subset of codewords $\mathbf{b}_i$ with $\tilde{w}_i = 0$ [11–18]. Since these codewords will not be used to represent $\mathbf{x}$, this allows (1) to be solved using a smaller dictionary. This has two benefits: better computational efficiency and the ability to solve larger problems with fixed resources.

Screening has its origins in feature selection heuristics in which a feature (codeword) is selected/rejected based on an empirical measure of its relevance to $\mathbf{x}$. An important recent development has given statistical performance guarantees for such methods [19]. Others have extended such tests [20]. A strict form of screening insists that no codeword is incorrectly rejected. The first line of work in this direction [11, 12] applies this idea across a broad range of problems involving sparse regularization. Follow-on work has focused on screening for variations of the lasso problem [13–18]. These methods are based on bounding the solution of the dual problem of (1) within a compact region $\mathcal{R}$ and finding $\mu_{\mathcal{R}}(\mathbf{b}_i) = \max_{\boldsymbol{\theta}\in\mathcal{R}} \boldsymbol{\theta}^T\mathbf{b}_i$. For simple regions $\mathcal{R}$, $\mu_{\mathcal{R}}$ is readily computed and this yields a screening test for rejecting unneeded codewords. This has resulted in tests based on spherical bounds [11,13], the intersection of spheres and half spaces (domes) [12, 14, 18], elliptical bounds [16] and novel methods for selecting the parameters for these regions, e.g., [17]. As a result, we now have classes of lasso screening tests which can be quickly executed, require few codewords to be in memory at once, and can significantly reduce dictionary size and speedup the solution of lasso problems [14,15,17,18].

Known classes of bounding regions (spheres, domes) are parameterized, e.g.,spherical bounds by a center $\mathbf{q}$ and radius r. This leads to an associated parameter selection problem: For a given region class (e.g., spherical) select the region parameters (e.g., $\mathbf{q}$ and r) to best bound the dual solution of (1) subject to available information and a computation budget. Better means of selecting parameters yield stronger and more time efficient tests. Some new ideas directed to this purpose have recently been proposed in [15–18].

Here we investigate an alternative question: is it worthwhile to develop more complex lasso screening tests that are structurally distinct from the above known classes. Clearly, there are classes of tests based on a spherical bound and k half spaces for $k \geq 0$; $k = 0$ yields the sphere tests and $k = 1$ yields the dome tests. As k increases one obtains more powerful tests but the tests are also more complex and time consuming to execute. To investigate this question we examine the case $k = 2$. This allows us to determine where things stand in the trade-off between test power and computational efficiency, particularly in comparison to the existing classes of tests $k = 0, 1$. If $k = 2$ yields significant payback despite its additional complexity, it will be a worthwhile addition the stable of lasso screening tests.

Each of the test classes $k = 0, 1, 2, \dots,$ has an associated parameter selection problem. Efficient methods for investing computation to best accomplish this are very important. In this regard, we expect recent results on this problem for $k = 0, 1$ [17, 18] will also be applicable when $k = 2$.

2. PRELIMINARIES

We consider the Lagrangian dual of (1), [11–14, 21–23]:

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

The solutions $\tilde{\mathbf{w}} = (\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_p)^T$ and $\tilde{\boldsymbol{\theta}}$ of the primal and dual problems are related through:

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

Letting $\mathfrak{B} = \{\pm\mathbf{b}_i\}_{i=1}^p$, allows the constraints in (2) to be written as $\forall\mathbf{b} \in \mathfrak{B} : \boldsymbol{\theta}^T\mathbf{b} \leq 1$. Let $\lambda_{\max} = \max_i \mathbf{x}^T\mathbf{b}_i$, and $\mathbf{b}_* \in \arg\max_{\mathbf{b}\in\mathfrak{B}} \mathbf{x}^T\mathbf{b}$. So $\lambda_{\max} = \mathbf{x}^T\mathbf{b}_*$. Note that $\boldsymbol{\theta} = \mathbf{x}/\lambda_{\max}$ is always a feasible point of (2). The dual problem (2) seeks the closest feasible point to $\mathbf{x}/\lambda$. If $\lambda > \lambda_{\max}$, then $\mathbf{x}/\lambda$ itself is feasible, making it the optimal solution. In this case, by (2), $\tilde{w}_i = 0, i = 1, \dots, p$. Hence we assume $0 \leq \lambda \leq \lambda_{\max}$. By (3), the dual problem provides a sufficient (but impractical) condition for excluding $\mathbf{b}_i$:

$$\tilde{\boldsymbol{\theta}}^T\mathbf{b}_i < 1 \Rightarrow \tilde{w}_i = 0. \tag{4}$$

A practical test is obtained as follows. If we know that $\tilde{\boldsymbol{\theta}} \in \mathcal{R}$ with $\mathcal{R}$ compact, then

$$\mu_{\mathcal{R}}(\mathbf{b}_i) = \max_{\boldsymbol{\theta}\in\mathcal{R}} \boldsymbol{\theta}^T\mathbf{b}_i < 1 \Rightarrow \tilde{w}_i = 0. \tag{5}$$

Examples of such regions $\mathcal{R}$ are given in §3. Selecting $\mathcal{R}$ involves a trade-off between rejection power and computation efficiency. The tighter the bound, the more codewords it could reject but the more complex the execution becomes.

3. TWO HYPERPLANE TEST

We now introduce a new lasso screening test: the Two Hyperplane Test (THT). This test corresponds to bounding $\tilde{\boldsymbol{\theta}}$ in the intersection of a sphere $S(\mathbf{q}, r) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \mathbf{q}\|_2 \leq r\}$ and two closed half spaces $H_i = \{\boldsymbol{\theta} : \mathbf{n}_i^T\boldsymbol{\theta} \leq c_i\}$, $i = 1, 2$, where $\mathbf{n}_i$ is the the unit normal to H_i and $c_i \geq 0$, $i = 1, 2$. This region is denoted by $\mathcal{D}(\mathbf{q}, r; \mathbf{n}_1, c_1; \mathbf{n}_2, c_2)$.

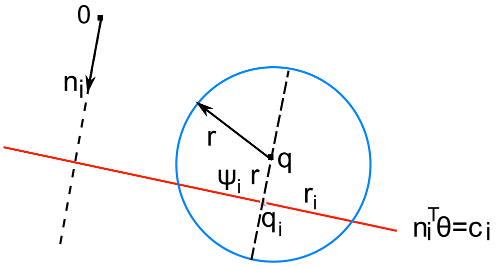

Fig. 1. Illustration of the region parameters, $i = 1, 2$.

Each hyperplane $P_i = \{\boldsymbol{\theta} : \mathbf{n}_i^T\boldsymbol{\theta} = c_i\}$ intersects the sphere forming a dome with base center $\mathbf{q}_i$ and base radius r_i. See Fig. 1. Denote the ratio between the signed distance from $\mathbf{q}$ to $\mathbf{q}_i$ and the sphere radius r by ψ_i. Then:

$$\psi_i = (\mathbf{n}_i^T\mathbf{q} - c_i)/r, \ \mathbf{q}_i = \mathbf{q} - \psi_i r\mathbf{n}_i, \ r_i = r\sqrt{1 - \psi_i^2}. \tag{6}$$

To ensure each intersection $H_i \cap S(\mathbf{q}, r)$ is nonempty and proper, we need $-1 \leq \psi_i \leq 1$. and to ensure the two half spaces intersect within the sphere, we need $\psi_1 + \psi_2 \leq \mathbf{n}_1^T\mathbf{n}_2$. These conditions ensure $\mathcal{D}(\mathbf{q}, r; \mathbf{n}_1, c_1; \mathbf{n}_2, c_2)$ is a nonempty, proper subset of the sphere and of each half space.

To find $\mu_{\mathcal{D}}(\mathbf{b}) = \max_{\boldsymbol{\theta}\in\mathcal{D}} \boldsymbol{\theta}^T\mathbf{b}$ we need to solve:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & (-\boldsymbol{\theta}^T\mathbf{b}) \\ \text{s.t.} \quad & (\boldsymbol{\theta} - \mathbf{q})^T(\boldsymbol{\theta} - \mathbf{q}) - r^2 \leq 0 \\ & \mathbf{n}_1^T\boldsymbol{\theta} - c_1 \leq 0 \\ & \mathbf{n}_2^T\boldsymbol{\theta} - c_2 \leq 0 \end{aligned} \tag{7}$$

This can be solved in closed form [18] and once $\mu_{\mathcal{D}}$ is known the THT is readily determined.

Theorem 1. *The Two Hyperplane Test (THT) for the region* $\mathcal{D}(\mathbf{q}, r; \mathbf{n}_1, c_1; \mathbf{n}_2, c_2)$ *is given by:*

$$V_l(\mathbf{n}_1^T\mathbf{b}_i, \mathbf{n}_2^T\mathbf{b}_i) < \mathbf{q}^T\mathbf{b}_i < V_u(\mathbf{n}_1^T\mathbf{b}_i, \mathbf{n}_2^T\mathbf{b}_i) \Rightarrow \tilde{w}_i = 0 \tag{8}$$

where $V_u(t_1, t_2) =$

$$\begin{cases} 1 - r, & \textit{if} \quad (a); \\ 1 + rt_2\psi_2 - r\sqrt{1 - t_2^2}\sqrt{1 - \psi_2^2}, & \textit{if} \quad (b), \\ 1 + rt_1\psi_1 - r\sqrt{1 - t_1^2}\sqrt{1 - \psi_1^2}, & \textit{if} \quad (c), \\ 1 + \frac{r}{1-\tau^2}\left[(\psi_1 - \tau\psi_2)t_1 + (\psi_2 - \tau\psi_1)t_2\right] & \\ \quad - \frac{r}{1-\tau^2} f(\psi_1, \psi_2) f(t_1, t_2), & \textit{otherwise,} \end{cases}$$

$f(x, y) = \sqrt{1 - \tau^2 + 2\tau xy - x^2 - y^2}$, $\tau = \mathbf{n}_1^T\mathbf{n}_2$, *conditions* (a), (b) *and* (c) *given by:*

$$\begin{aligned} &(a) \quad t_1 < -\psi_1, t_2 < -\psi_2, \\ &(b) \quad t_2 \geq -\psi_2, \frac{t_1 - \tau t_2}{\sqrt{1 - t_2^2}} < \frac{-\psi_1 + \tau\psi_2}{\sqrt{1 - \psi_2^2}}, \\ &(c) \quad t_1 \geq -\psi_1, \frac{t_2 - \tau t_1}{\sqrt{1 - t_1^2}} < \frac{-\psi_2 + \tau\psi_1}{\sqrt{1 - \psi_1^2}}, \end{aligned}$$

and $V_l(t_1, t_2) = -V_u(-t_1, -t_2)$.

As a class of tests, THT is guaranteed to be the most powerful among all classes that bound $\tilde{\boldsymbol{\theta}}$ within the intersection of a sphere and up to two half spaces, i.e., a THT test can eliminate at least as many codewords as any test based on the same or fewer structural elements. Theorem 1 also shows that THT uses only the $3p$ correlations $\{\mathbf{q}^T\mathbf{b}_i, \mathbf{n}_1^T\mathbf{b}_i, \mathbf{n}_2^T\mathbf{b}_i\}_{i=1}^p$. Since each correlation can be computed in $O(n)$ time, THT has linear-time complexity $O(pn)$.

To obtain specific instances of THT we must consider the parameter selection problem. The inequality constraints in (2) show that the feasible set $\mathcal{F}$ of the dual problem is a nonempty closed polyhedron which depends only on the codewords in the pool. The point $\boldsymbol{\theta}_F = \mathbf{x}/\lambda_{\max}$ is always in $\mathcal{F}$ and can be taken as a default feasible dual solution. The distance $\|\tilde{\boldsymbol{\theta}} - \mathbf{x}/\lambda\|_2$ can be no greater than $\|\boldsymbol{\theta}_F - \mathbf{x}/\lambda\|_2$. This provides a default spherical bound for $\tilde{\boldsymbol{\theta}}$, with parameters $\mathbf{q} = \mathbf{x}/\lambda$, $r = \|\boldsymbol{\theta}_F - \mathbf{x}/\lambda\|_2$. Recent results in [17, 18] indicate how to refine this initial spherical bound. In addition, if many problem instances $(\mathbf{x}_j, \lambda_j)$ are solved using a common dictionary B, the dual solution of a previously solved instance could be better feasible dual point than the default indicated above. Next we select two half space bounds. The inequality constraints in (2) are natural half space bounds on $\tilde{\boldsymbol{\theta}}$. We can select two such half spaces with the objective of minimizing the area of intersection with the sphere. The intersection of the first half space $\{\boldsymbol{\theta} : \mathbf{n}_1\boldsymbol{\theta} \leq c_1\}$ with the sphere $S(\mathbf{q}, r) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \mathbf{q}\|_2 \leq r\}$ forms a dome. Using (6), the area of the dome is minimized when r_1 is minimized, and hence when $\mathbf{n}_1^T\mathbf{q}$ is maximized. The second hyperplane, intersects this dome to form the final bounding region. Again, to minimize the intersection, $\mathbf{n}_2$ is chosen from $\mathfrak{B}/\mathbf{n}_1$ to be most correlated with the center of the first dome $\mathbf{q}_1$. We refer to this form of THT as a Codeword based THT (C-THT):

$$\mathbf{n}_1 = \arg\max\nolimits_{\mathbf{b}\in\mathfrak{B}} \mathbf{b}^T\mathbf{q}, \qquad c_1 = 1; \tag{9}$$

$$\mathbf{n}_2 = \arg\max\nolimits_{\mathbf{b}\in\mathfrak{B}\setminus\{\mathbf{n}_1\}} \mathbf{b}^T\mathbf{q}_1 \quad c_2 = 1. \tag{10}$$

Alternatively, if we have solved instance $(\mathbf{x}_0, \lambda_0)$ yielding primal and dual solutions $\tilde{\mathbf{w}}_0$ and $\tilde{\boldsymbol{\theta}}_0$ (see (3)), then $\tilde{\boldsymbol{\theta}}_0$ must satisfy the inequalities defining the unique projection of $\mathbf{x}_0/\lambda_0$ onto the closed convex set $\mathcal{F}$ [24]: for each $\boldsymbol{\theta} \in \mathcal{F}$,

$$(\mathbf{x}_0/\lambda_0 - \tilde{\boldsymbol{\theta}}_0)^T(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_0) \leq 0 \tag{11}$$

Using some algebra and (3), (11) can be written as:

$$(B\mathbf{w}_0)^T\boldsymbol{\theta} \leq (B\mathbf{w}_0)^T\tilde{\boldsymbol{\theta}}_0. \tag{12}$$

Since $0 \in \mathcal{F}$, the right hand side is nonnegative. Hence (12) bounds $\mathcal{F}$ (thus $\tilde{\boldsymbol{\theta}}$) in the half space $\mathbf{n}_1^T\boldsymbol{\theta} \leq c_1$ with

$$\mathbf{n}_1 = B\mathbf{w}_0/\|B\mathbf{w}_0\|_2, \qquad c_1 = \mathbf{n}_1^T\tilde{\boldsymbol{\theta}}_0. \tag{13}$$

One can then select $\mathbf{n}_2$ and c_2 using the method in (10).

4. EXPERIMENTS

We now empirically examine the performance of the THT, compare it to existing tests, and examine the trade-off between improved screening and computational efficiency that it imposes. We do so using two metrics: rejection power and speed-up. Rejection power is the percentage of codewords rejected by the test. Speed-up is the ratio between the time to solve the original lasso problem and the sum of the time to do screening plus solve the reduced lasso problem. Speed-up measures how much faster the lasso problem can be solved with the help of screening. Note that it takes into account the time spent on screening *and* the cost of solving the screened problem. Both of the above metrics are important since one can be traded-off against the other.

We use the following real and synthetic data sets. **RAND**: 10,000 28-dimensional vectors randomly generated using the MATLAB *rand* function); **MNIST500**: 5000 images of size $n = 28 \times 28 = 784$, consisting of random samples of 500 images of each digit in the MNIST data set; **YALEBXF**: 2,414 frontal face images of size $n = 192 \times 168 = 32,256$ of the 38 subjects in the extended Yale B face recognition data set.

For each data set, we screen and solve 64 lasso problems, each with a distinct, randomly selected target $\mathbf{x}$, and use the remaining vectors as codewords. We report the average performance with standard error over these instances.

We experimentally compare the following screening tests: dome test (DT, with default $\boldsymbol{\theta}_F$ and with $\boldsymbol{\theta}_F = \tilde{\boldsymbol{\theta}}$) [14]; Codeword based THT (C-THT, with default $\boldsymbol{\theta}_F$ and with $\boldsymbol{\theta}_F = \tilde{\boldsymbol{\theta}}$). For DT and C-THT, using $\boldsymbol{\theta}_F = \tilde{\boldsymbol{\theta}}$ is obviously impractical; it provides a performance upper bound for a sphere centered at $\mathbf{x}/\lambda$.

The results of using THT to screen lasso problems on these datasets are shown in Fig. 2. As expected, C-THT exhibits superior rejection: it can reject $10\% - 30\%$ more codewords than the equivalent DT. But this comes at a price. In speedup, C-THT outperforms DT when $\lambda/\lambda_{\max}$ is small, but the reverse holds when $\lambda/\lambda_{\max}$ is large. The simpler test (DT) is more computationally efficient for larger values of $\lambda/\lambda_{\max}$. To make the comparison fair, both DT and C-THT utilize ONLY the codewords to construct the hyperplanes in their respective bounding regions. Of course, the performance of the tests can always be improved by using better parameter selection. We only include results using the FeatureSign [25] lasso solver but experiments indicate consistent speedup results across the FeatureSign, Grafting [26] and FISTA [27] lasso solvers.

4.1. Using THT in sequential screening

For many data sets, when $\lambda/\lambda_{\max}$ is small (< 0.3), existing screening tests (including C-THT) reject few codewords and yield modest speedup. Prior work has suggested that a sequential screening scheme can significantly improve per-

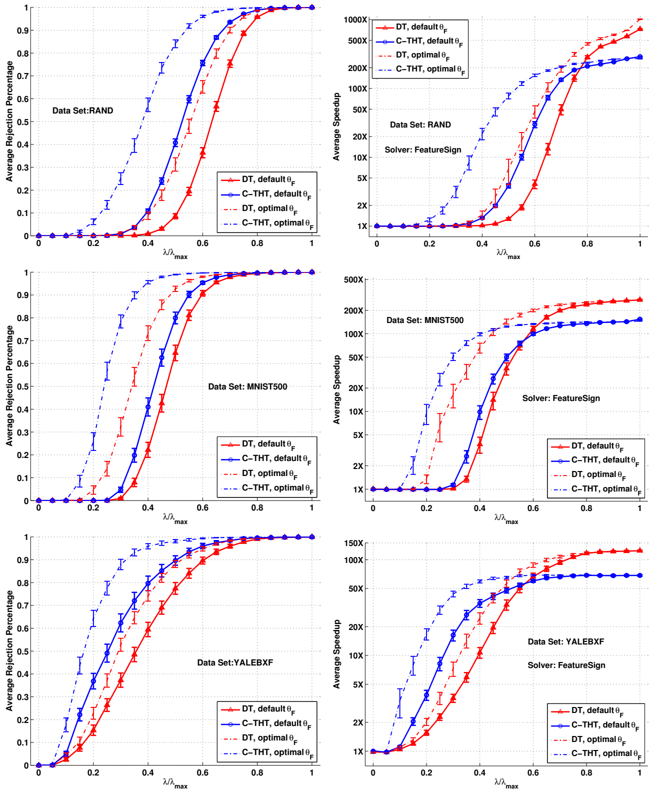

Fig. 2. Results for three data sets. Left: average rejection percentage; Right: average speedup using the FeatureSign solver.

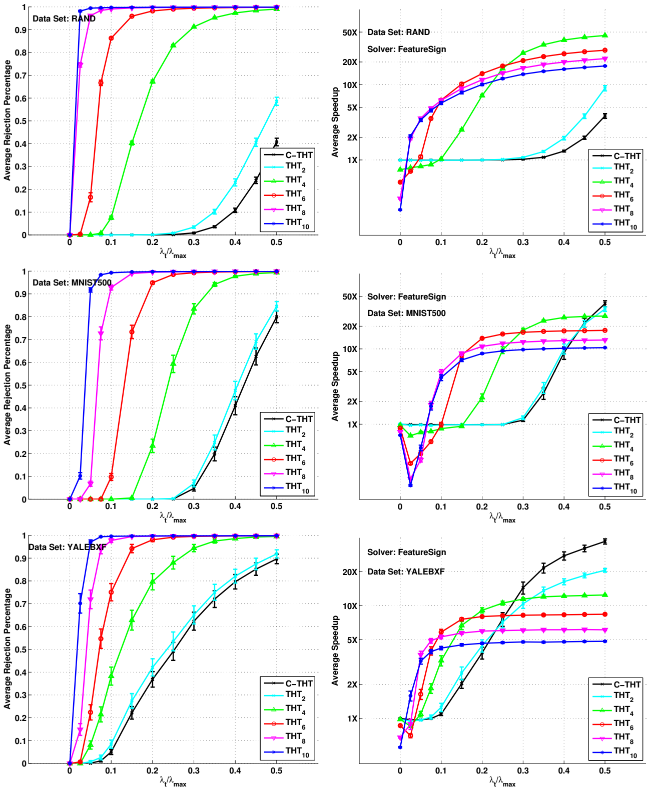

Fig. 3. THT used in sequential screening. Left: average rejection percentage; Right: average speedup using the FeatureSign solver.

formance in this situation [11, 12, 15, 17, 20]. This can be done using THT as follows. Denote the value of interest of the regularization parameter by λ_t. To solve the target lasso problem $(\mathbf{x}, \lambda_t)$, we bring in a sequence $\{\lambda_k\}_{k=1}^N$ that decreases to the given target value $\lambda_N = \lambda_t$. Then we screen (using THT) and solve each of the instances $(\mathbf{x}, \lambda_k)$ to obtain the solution at $\lambda_N = \lambda_t$. At λ_k, the known solution ($\tilde{\mathbf{w}}_{k-1}$, $\tilde{\boldsymbol{\theta}}_{k-1}$) to the previous instance $(\mathbf{x}, \lambda_{k-1})$ is used to provide the parameters of a THT test for $(\mathbf{x}, \lambda_k)$. The parameters of the spherical bound are selected by existing methods (we used $\|\tilde{\boldsymbol{\theta}}_k - \mathbf{x}/\lambda_k\|_2 \leq \|\tilde{\boldsymbol{\theta}}_{k-1} - \mathbf{x}/\lambda_k\|_2$). The first hyperplane constraint is obtained via (13) and the second is drawn from the codewords using (10). Compared with C-THT, this method further boosts performance significantly, as shown in Fig. 3. In the figure, the subscript on THT denotes N the number of λ value's that are used to solve the target problem. As can be seen from Fig. 3, for $\lambda_t/\lambda_{\max}$ as low as 0.1, where C-THT barely rejected any codewords, THT$_{10}$ rejects more than 90% of the codewords while concurrently provide around 5X speedup. This is a considerable improvement in both codeword rejection and speedup despite solving a sequence of 10 lasso problems to obtain the desired solution. More details on this approach, including the design of the sequence $\{\lambda_k\}$ are given in the companion paper [18].

5. CONCLUSION

We have investigated the trade-off between rejection power and computation efficiency of more complex screening tests by developing and examining a specific new class: the Two Hyperplane Test. The experimental results show that when used as a standalone screening test, THT can indeed reject significantly more codewords than current classes of tests. Moreover, when the regularization parameter is small, THT can do so while also providing significant speed-up. Since this is a range of λ that occurs frequently in applications, this makes THT quite interesting. These are positive and encouraging results. At the other extreme, when $\lambda/\lambda_{\max}$ is moderate to large, the experimental results show that one is probably better off using a simpler class of screening tests such as the dome tests. We also examined the use of THT as a building block in sequential screening schemes. Our experimental results in this application indicated that THT can significantly boost screening performance for very small values of λ.

One might ask why not use 3 hyperplanes or even 4? This will clearly yield better codeword rejection. However, it will do so at the expense of greater complexity and greater execution time. The results for THT used in a sequential screening scheme suggest that such additional complexity may not be warranted at this time.

6. REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[2] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *Image Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 53–69, 2008.

[3] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 2272–2279.

[4] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[5] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Towards a practical face recognition system: robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

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

[7] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Acoustics Speech and Signal Processing, 2010 IEEE International Conference on*, 2010, pp. 4370–4373.

[8] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] K. Chang, J. Jang, and C. S. Iliopoulos, "Music genre classification via compressive sampling," in *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, 2010, pp. 387–392.

[10] S. Prasad, P. Melville, A. Banerjee, and V. Sindhwani, "Emerging topic detection using dictionary learning," in *ACM Conference on Information and Knowledge Management*, 2011.

[11] 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.

[12] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems," *arXiv:1009.4219v2 [cs.LG]*, 2011.

[13] Z. J. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Advances in Neural Information Processing Systems*, 2011.

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

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

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

[17] 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.

[18] Z. J. Xiang, Y. Wang, and P. J. Ramadge, "Screening Tests for Lasso Problems," Tech. Rep., Princeton University, Dec. 2012.

[19] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[20] 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.

[21] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319–337, June 2000.

[22] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large scale ℓ_1-regularized least squares," *IEEE Selected Topics in Signal Processing*, vol. 1, pp. 606–617, 2007.

[23] R. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

[24] Jean-Baptiste Hiriart-Urruty and Claude Lemarechal, *Fundamentals of Convex Analysis*, Springer, 2001.

[25] H. Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*, 2007, vol. 19, p. 801.

[26] Simon Perkins and James Theiler, "Online featur selection using grafting," in *International Conference on Machine Learning*, 2003, pp. 592–599.

[27] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.