LASSO SCREENING WITH A SMALL REGULARIZATION PARAMETER

Yun Wang, Zhen James Xiang and Peter J. Ramadge

Department of Electrical Engineering, Princeton University

ABSTRACT

Screening for lasso problems is a means of quickly reducing the size of the dictionary needed to solve a given instance without impacting the optimality of the solution obtained. We investigate a sequential screening scheme using a selected sequence of regularization parameter values decreasing to the given target value. Using analytical and empirical means we give insight on how the values of this sequence should be chosen and show that well designed sequential screening yields significant improvement in dictionary reduction and computational efficiency for lightly regularized lasso problems.

***Index Terms*—** screening, sparse regression, regularized regression

1. INTRODUCTION

Screening of the lasso problem [1]:

$$\min_{w_i, i=1:p} \quad 1/2\|\mathbf{x} - \sum_{i=1}^{p} w_i \mathbf{b}_i\|_2^2 + \lambda \sum_{i=1}^{p} |w_i|, \quad (1)$$

is a means of quickly reducing the size of the dictionary $B = [\mathbf{b}_1, \ldots, \mathbf{b}_p]$ needed to solve a given instance $(\mathbf{x}, \lambda)$ without impacting the optimality of the solution obtained. This problem has recently attracted much attention [2–9]. This is partly motivated by the need to solve larger scale lasso problems efficiently using limited memory. Prior to solving the lasso problem, screening uses knowledge of $\mathbf{x}$, λ and B to quickly identify a set of codewords that receive zero weights ($\tilde{w}_i = 0$) in all solutions $\tilde{\mathbf{w}}$ of the current instance $(\mathbf{x}, \lambda)$. This makes it possible to solve the instance using a smaller dictionary. Screening thus enables the lasso to be solved faster and to be applied to larger scale problems.

Current screening methods are based on bounding the solution of the dual problem of (1) within a compact region $\mathcal{R}$. A tighter bound $\mathcal{R}$ enables the removal of more unneeded codewords from B. For example, $\mathcal{R}$ could be a sphere [2,4] or the intersection of a sphere and a half space (a dome) [3,5,9]. Let $\lambda_{\max} = \max_{\mathbf{b} \in \{\pm \mathbf{b}_i\}_{i=1}^{p}} \mathbf{x}^T \mathbf{b}$. Current tests perform well for $\lambda_{\max}/2 \leq \lambda < \lambda_{\max}$. However, when $\lambda/\lambda_{\max} < 0.3$, the tests fail to provide equivalent performance since known bounds $\mathcal{R}$ are not tight when λ is small. This situation occurs frequently since one often seeks lightly regularized solutions.

There are approaches which can help with this problem. For example, [2] used screening to help solve (1) for a sequence of instances $\{(\mathbf{x}, \lambda_k)\}_{k=1}^{N}$. The primary objective was to quickly obtain a dense sampling of the regularization path of the problem. At each step, the previously solved instance $(\mathbf{x}, \lambda_{k-1})$ was used to define a bound for the dual solution of the next instance $(\mathbf{x}, \lambda_k)$. This enables a tighter region bound for the current problem. But solving many instances along the regularization path is inefficient if we only require the solution of one instance $(\mathbf{x}, \lambda_t)$.

In [8] it is proposed to run K steps of the homotopy algorithm to find a solution at the K-th breakpoint on the regularization path of $\tilde{\mathbf{w}}(\lambda)$. This effectively solves a sequence of lasso problems (via homotopy) to obtain a solution $\tilde{\mathbf{w}}_K$ at $\lambda_K > \lambda_t$. This solution is then used to help screen the instance $(\mathbf{x}, \lambda_t)$. This has some advantages over a dense sampling of the regularization path. But it cannot directly control the values λ_j used nor how close λ_K is to λ_t. In the worst case, the regularization path is not well defined (a lasso problem need not have a unique solution) and even when it is well defined, it can have an exponential number of breakpoints ($O(3^p)$ where p is the number of codewords) [10].

We propose to adopt the best features of the above methods: a sequential approach that uses the previous solution to help screen the next instance, but to do so in an otherwise unconstrained form: Given $\mathbf{x}$ and λ_t, select N and a sequence $\{\lambda_k\}_{k=1}^{N}$ with $\lambda_N = \lambda_t$. Then efficiently solve the sequence of lasso problems $(\mathbf{x}, \lambda_k)$ to obtain the solution of the instance $(\mathbf{x}, \lambda_t)$. Note, we are only interested in the solution of one instance: $(\mathbf{x}, \lambda_t)$. The other instances are simply way points in the computation. The germ of this idea was proposed in [6].

What sets this formulation apart from the methods discussed above is that we are free to design the sequence $\{\lambda_k\}_{k=1}^{N}$. An equally spaced sequence, for example, is usually undesirable. A key idea from [3] will also be very important: a particular dome defined by the previous solution screens the next instance most effectively. We demonstrate both analytically and empirically why this is so and why uniform spacing of the λ_k is usually undesirable. We also show empirically that careful selection of the λ_k can yield sequential screening schemes that significantly enhance solving lightly regularized lasso problems.

2. SEQUENTIAL SCREENING

For simplicity, we assume all codewords $\mathbf{b}_i$ and the target vector $\mathbf{x}$ are normalized, i.e., $\|\mathbf{x}\|_2 = \|\mathbf{b}_i\|_2 = 1, i =$

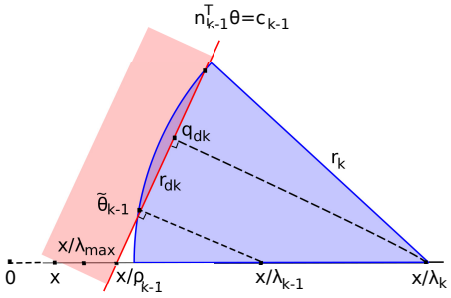

Fig. 1. An illustration of the dome (5) formed at step k.

$1, 2, \ldots, p$. The Lagrangian dual of (1) [2–5, 11–13]:

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

will be important in what follows. Let $\mathcal{F}$ denote the feasible set of (2). $\mathcal{F}$ is a closed polyhedron that depends only on the dictionary $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p]$.

Current screening tests bound the solution $\tilde{\boldsymbol{\theta}}$ of (2) within a region $\mathcal{R} \subset \mathbb{R}^n$ and thereby derive a corresponding screening test. One class of tests, [3, 5, 9], uses a dome region bound $D(\mathbf{q}, r; \mathbf{n}, c)$ formed by the intersection of a sphere $S(\mathbf{q}, r) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \mathbf{q}\|_2 \leq r\}$ and a half space $H(\mathbf{n}, c) = \{\boldsymbol{\theta} \colon \mathbf{n}^T\boldsymbol{\theta} \leq c\}$. This dome region has the following basic features. The dome center $\mathbf{q}_d$ is the point of intersection of the bounding hyperplane and the line passing through $\mathbf{q}$ in the direction $\mathbf{n}$. The signed distance from $\mathbf{q}$ to $\mathbf{q}_d$ is a fraction ψ_d of the sphere radius r. The dome radius r_d is the maximum distance one can move from $\mathbf{q}_d$ within the bounding hyperplane and dome. This is illustrated in Fig. 1 with $\mathbf{q} = \mathbf{x}/\lambda_k$.

We want $D(\mathbf{q}, r; \mathbf{n}, c)$ to contain the dual solution $\tilde{\boldsymbol{\theta}}$ of the current instance. For example, let $\boldsymbol{\theta}_F$ be a known point in $\mathcal{F}$. Then the sphere with center $\mathbf{q} = \mathbf{x}/\lambda$ and radius $r = \|\mathbf{x}/\lambda - \boldsymbol{\theta}_F\|_2$ must contain $\tilde{\boldsymbol{\theta}}$. Moreover, the dual solution $\tilde{\boldsymbol{\theta}}$ must satisfy the constraints of (2). Hence for any $\mathbf{b}_i$, $|\mathbf{b}_i^T\tilde{\boldsymbol{\theta}}| \leq 1$. In particular, let $\mathbf{b}_* \in \arg\max_{\mathbf{b} \in \{\pm\mathbf{b}_i\}_{i=1}^p} \mathbf{x}^T\mathbf{b}$ and $\lambda_{\max} = \mathbf{x}^T\mathbf{b}_*$. Then $\mathbf{b}_*^T\tilde{\boldsymbol{\theta}} \leq 1$. This yields the dome bound:

$$D^{''} = D(\mathbf{x}/\lambda, \|\mathbf{x}/\lambda - \boldsymbol{\theta}_F\|_2; \mathbf{b}_*, 1) \tag{3}$$

In particular, $\mathbf{x}/\lambda_{\max}$ is always a feasible point. Setting $\boldsymbol{\theta}_F = \mathbf{x}/\lambda_{\max}$ in (3) yields the dome region used in [5]. Determining this version of $D^{''}$ only requires knowledge of $(\mathbf{x}, \lambda)$ and some minor computation to find $\lambda_{\max}$ and $\mathbf{b}_*$.

The screening test corresponding to dome region bounds is known [9] and specific instances, differing in the selection of the dome parameters, have been explored in [3, 5, 9]. However, when (1) is lightly regularized the screening test for known dome region bounds is not very effective.

To address this we consider a sequence of instances $\{(\mathbf{x}, \lambda_k)\}_{k=1}^N$, with $\lambda_1 > \cdots > \lambda_N = \lambda_t$. The idea is that we sequentially screen and solve each of these instances to eventually obtain the solution of the desired instance $(\mathbf{x}, \lambda_t)$. Let $\tilde{\boldsymbol{\theta}}_k$ denote solution to the dual problem for $(\mathbf{x}, \lambda_k)$. Then $\tilde{\boldsymbol{\theta}}_{k-1}$ is available for screening the next instance $(\mathbf{x}, \lambda_k)$. Using the dome (3) with $\boldsymbol{\theta}_F = \tilde{\boldsymbol{\theta}}_{k-1}$ yields the dome bound:

$$D_k^{'} = D(\mathbf{x}/\lambda_k, \|\mathbf{x}/\lambda_k - \tilde{\boldsymbol{\theta}}_{k-1}\|_2; \mathbf{b}_*, 1). \tag{4}$$

By Pythagoras' theorem the dome radius r_{dk} of $D_k^{'}$ satisfies $r_{dk}^2 = r_k^2 - r_k^2\psi_{dk}^2$, with $r_k = \|\mathbf{x}/\lambda_k - \tilde{\boldsymbol{\theta}}_{k-1}\|_2$ (Fig. 1). By definition we have $\mathbf{x}/\lambda_k - \mathbf{b}_* r_k\psi_{dk} = \mathbf{q}_d k$. Hence $r_k\psi_{dk} = \mathbf{b}_*^T\mathbf{x}/\lambda_k - 1 = \lambda_{\max}/\lambda_k - 1$. Combining and simplifying these expressions yields:

$$r_{dk} = \frac{1 - \lambda_{\max}^2}{\lambda_k^2} + \frac{2}{\lambda_k}(\lambda_{\max} - \mathbf{x}^T\tilde{\boldsymbol{\theta}}_{k-1}) + \|\tilde{\boldsymbol{\theta}}_{k-1}\|_2^2 - 1.$$

So if $\tilde{\boldsymbol{\theta}}(\lambda)$ is bounded as λ goes to 0 (a reasonable assumption), then the radius of the bounding sphere grows like $1/\lambda_k$. Moreover, using the fixed half space $H(\mathbf{b}_*, 1)$ allows the dome radius r_{dk} to grow like $1/\lambda_k^2$. This will impede the effectiveness of the corresponding screening test. The same conclusion holds for any bounding sphere with radius $O(1/\lambda_k)$ and a fixed half space.

The limitations of $D_k^{'}$ emphasize the need to extract as much information as possible from the solution of the previous instance. Following [3], we will use the previous solution to form the hyperplane passing through $\tilde{\boldsymbol{\theta}}_{k-1}$ with its normal aligned with $\mathbf{x}/\lambda_{k-1} - \tilde{\boldsymbol{\theta}}_{k-1}$. Using standard convex analysis one can show that this hyperplane separates $\mathcal{F}$ from $\mathbf{x}/\lambda_{k-1}$. Combining this hyperplane with the sphere adapted to $\tilde{\boldsymbol{\theta}}_{k-1}$ used above, yields the dome bound:

$$D_k = D(\mathbf{x}/\lambda_k, \|\mathbf{x}/\lambda_k - \tilde{\boldsymbol{\theta}}_{k-1}\|_2; \mathbf{n}_{k-1}, c_{k-1}) \tag{5}$$

where $\mathbf{n}_{k-1} = (\mathbf{x}/\lambda_{k-1} - \tilde{\boldsymbol{\theta}}_{k-1})/\|\mathbf{x}/\lambda_{k-1} - \tilde{\boldsymbol{\theta}}_{k-1}\|_2$ and $c_{k-1} = \mathbf{n}_{k-1}^T\tilde{\boldsymbol{\theta}}_{k-1}$. For this bounding region both the sphere radius and the half space are adapted to the solution of the previous instance. As λ_k decreases the sphere radius still increases roughly like $1/\lambda_k$. But this time, as $\mathbf{x}/\lambda_k$ recedes the hyperplane shifts and rotates to cut a smaller dome.

So our proposed sequential screening test is structured as follows. Given the target instance $(\mathbf{x}, \lambda_t)$, we screen and solve a designed sequence of instances $\{(\mathbf{x}, \lambda_k)\}_{k=1}^N$, with $\lambda_N = \lambda_t$ and $\lambda_1 < \lambda_{\max}$ (normally λ_1 close to $\lambda_{\max}$). For the first instance $(\mathbf{x}, \lambda_1)$, we screen and solve the lasso problem using the region $D^{''}$. This is effective when λ_1 is large. At each subsequent λ_k, $k = 2 : N$, the dual solution of $(\mathbf{x}, \lambda_{k-1})$ is used to construct region D_k given in (5). We expect this bounding region to be reasonably tight. Hence a high percentage of the dictionary to be discarded at each λ_k including λ_t. Since we expect to discard a high fraction of codewords at each step, we also expect the total computation time for both screening and solving the entire sequence will be competitive with, or even less than, the time to screen and

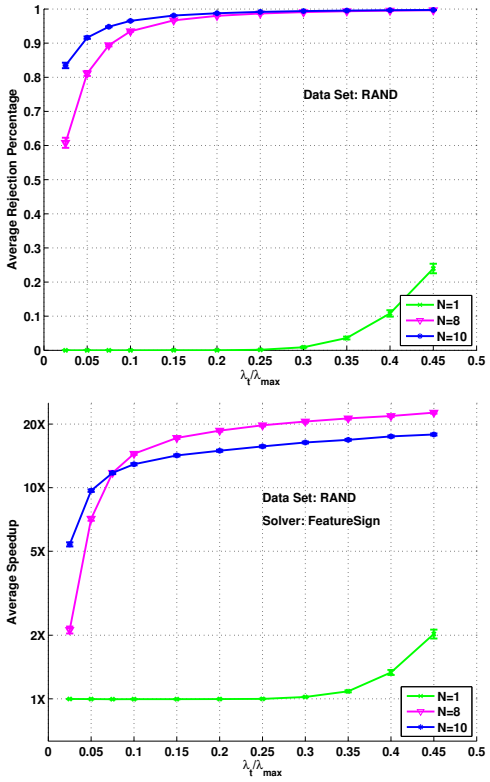

Fig. 2. Comparison of sequential screening ($N = 8, 10$, geometric spacing) and one-shot screening ($N = 1$) on **RAND**. Top: average rejection percentage. Bottom: average speedup

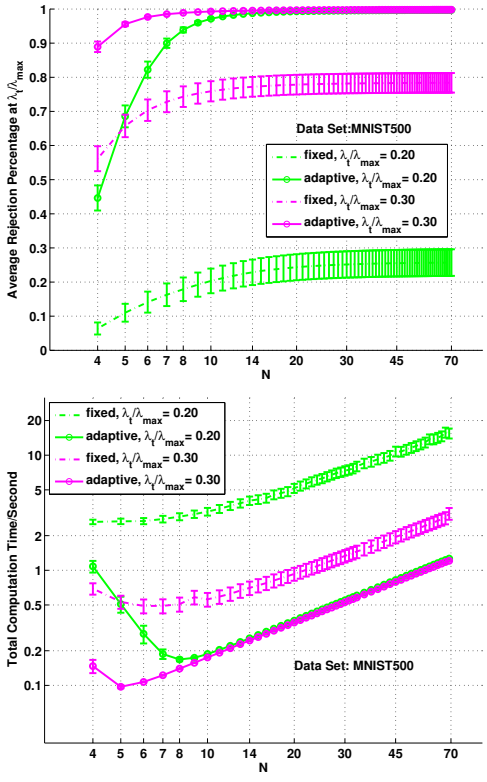

Fig. 3. Comparison of dome (4) and dome (5) used for sequential screening with geometric spacing on **MNIST500**. Top: average rejection percentage. Bottom: total computation time

solve the one instance at λ_t. We offer empirical verification of this fact in §3. One problem remains to be considered: the design of the sequence $\{\lambda_k\}_{k=1}^N$.

Screening is easy when λ is large but hard, without a tight region bound, when λ is small. Hence uniform spacing of the λ_k is not the best use resources. It is better to space the larger values of λ_k further apart and the smaller values much closer together. One can get some insight on this by considering Fig. 1. The segment of a circle (in blue) represents the bounding sphere and the solid line (in red) and associated shading (also in red) indicates the bounding half space, both at step k. The dome (5) is the intersection of these regions. The point $\mathbf{x}/\rho_k$ is the point of intersection of the line through 0 and $\mathbf{x}$ and the hyperplane $\mathbf{n}_k^T\boldsymbol{\theta} = c_k$. The similarity of the two right angle triangles yields a useful formula for the radius of the dome:

$$r_{dk} = (\frac{1}{\lambda_k} - \frac{1}{\lambda_{k-1}})(\frac{1}{\lambda_{k-1}} - \frac{1}{\rho_{k-1}})^{-1} a_{k-1}. \qquad (6)$$

Here $a_{k-1} = \|\tilde{\boldsymbol{\theta}}_{k-1} - \mathbf{x}/\rho_{k-1}\|_2$. Let the λ_k be uniformly spaced by Δ. Then by (6), at $k = N$ we have:

$$r_{dN}^u = \frac{\Delta}{\lambda_t}\frac{1}{(1 - \lambda_{N-1}/\rho_{N-1})} a_{N-1} \geq \frac{\Delta}{\lambda_t} a_{N-1}.$$

So provided a_{N-1} is bounded away from 0, the dome radius is unbounded as λ_t gets smaller. To ensure r_{dN}^u remain bounded as λ_t approaches 0 requires $N = O(1/\lambda_t)$. On the other hand, consider geometric spacing: $\lambda_k = \alpha\lambda_{k-1}$ with $0 < \alpha < 1$. Using (6) this yields:

$$r_{dN}^g = \frac{1-\alpha}{\alpha}\frac{\rho_{N-1}}{(\rho_{N-1} - \lambda_{N-1})} a_{N-1}$$

Assume a_{N-1} and ρ_{N-1} are bounded and ρ_{N-1} is bounded away from 0. Then to ensure r_{dN}^g is bounded we need α bounded away from 0 and this requires $N = O(\log(1/\lambda_t))$. This indicates a logarithmic difference in N between uniform and geometric spacing. We empirically demonstrate this in the next section.

3. EXPERIMENTS

We ran experiments on the datasets: **RAND**: 10,000 28-dimensional vectors randomly generated using the Matlab *rand()* function; and **MNIST500**: 5000 images of size $n = 28 \times 28 = 784$, obtained from the first 500 images of each digit in the MNIST data set. For the basic screening test to be used at each step in sequential screening we selected the Two Hyperplane Test (THT) and its codeword

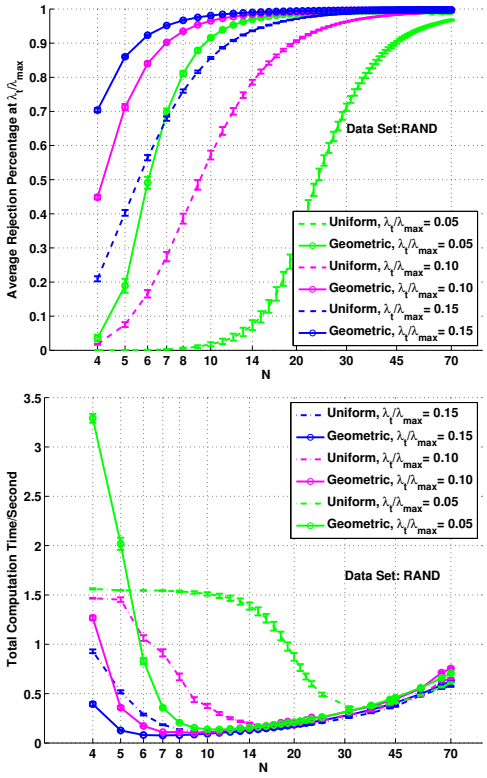

Fig. 4. Geometric vs uniform spacing on **RAND**. Top: average rejection percentage at λ_t. Bottom: total computation time

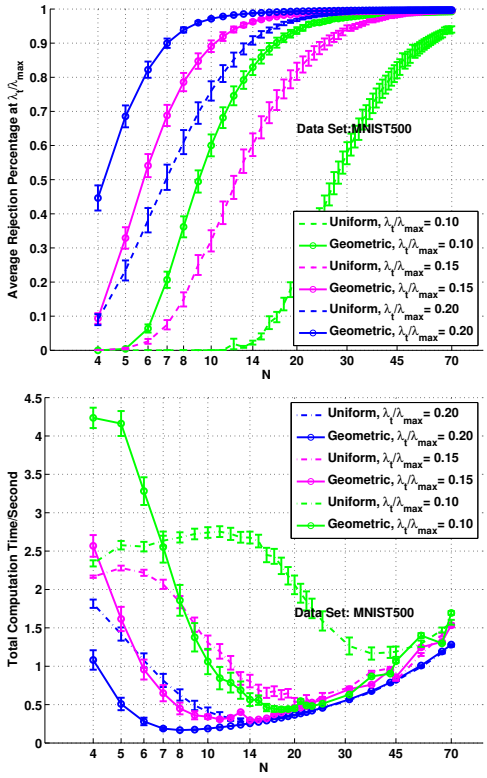

Fig. 5. Geometric vs uniform on **MNIST500**. Top: average rejection percentage at λ_t. Bottom: total computation time

based derivative (C-THT) [9]. For each data set, we construct 64 lasso problems, each with a distinct, randomly selected target $\mathbf{x}$, and use the remaining vectors as codewords. Results are reported with standard errors over these instances. Performance is evaluated by the percentage of codewords discarded (rejected) and the speedup factor. The reported rejection percentage is the percentage of codewords rejected by the test at λ_t. The speedup factor is the ratio of the time to solve the lasso problem without any screening to the time to perform the screening *and* solve the lasso problem. For a fair comparison of time efficiency, for a one-shot test ($N = 1$), the denominator of the speedup factor is the time to screen plus solve at λ_t. For sequential screening, the denominator is the total computation time, i.e., the total time to screen and solve the entire sequence $\{(\mathbf{x}, \lambda_k)\}_{k=1}^N$. We report results using the FeatureSign lasso solver [14], but our experiments indicate consistent results using several lasso solvers.

As Fig. 2 shows, sequential screening significantly outperforms one-shot screening in both metrics. Fig. 3 compares the performance of the adaptive hyperplane (5) and the fixed hyperplane (4). The fixed hyperplane dome fares poorly against the adaptive hyperplane, in agreement with our analysis. We then use the dome (5) to compare the performance of geometric and uniform spacing of the sequence $\{\lambda_k\}_{k=1}^N$. In each case, the sequence starts from $\lambda_1 = 0.95\lambda_{\max}$ and ends at $\lambda_N = \lambda_t$. We consider three λ_t's each plotted in a different color in Fig. 4 and Fig. 5. For each λ_t the geometric spacing is plotted in a solid line with circle marker and the uniform spacing is plotted in a dotted line. Geometric spacing has consistently better rejection in less computation time (when N is appropriately chosen) than uniform spacing. To achieve the same rejection percentage and/or total computation time, uniform spacing requires a much large value of N than geometric spacing. This is agreement with our theoretical analysis. As λ_t approaches 0, both spacing curves shift to the right, implying a smaller λ_t needs more points for the same performance. However, the margin between geometric and uniform spacing is much larger as λ_t approaches 0.

4. CONCLUSION

To solve a lightly regularized lasso problem (small λ_t), we have proposed screening and solving a designed sequence of instances $(\mathbf{x}, \lambda_k)$ for a decreasing sequence $\{\lambda_k\}_{k=1}^N$ with $\lambda_N = \lambda_t$. We examined both analytically and empirically the impact of the regularization sequence on the performance of the method. We have shown that when λ_t is small, a well designed sequential screening algorithm outperforms both uniform sampling along the regularization path and one-shot screening tests in both rejection power and computation time.

5. 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] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

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

[10] J. Mairal and B. Yu, "Complexity analysis of the lasso regularization path," in *Proceedings of the 29th Int. Conf. on Machine Learning (ICML 2012), Edinburgh, Scotland*, 2012.

[11] 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.

[12] 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.

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

[14] 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.