Part III Essay

May 2019

Introduction

This essay explores the topic of SAFE screening tests for the Lasso problem. Screening tests are a means of achieving dimensionality reduction, and can be used to greatly accelerate Lasso-solving algorithms. In particular, SAFE screening tests are able to achieve dimensionality reduction without the loss of any relevant information. These tests are based on exploiting geometric properties of the dual formulation of the Lasso problem. A variety of SAFE tests have been proposed in the literature, all of which seek to identify a subset of features guaranteed to have zero weight in the Lasso solution. In the following pages, we survey a number of increasingly sophisticated SAFE tests, and also provide an empirical comparison of their efficacy.

The Lasso

We begin by considering the standard linear regression setting: given n samples $\{(x_i, y_i)\}_{i=1}^n$, we aim to infer a linear function relating the response variable $y_i \in \mathbb{R}$ to the vector of regressors $x_i \in \mathbb{R}^p$. Specifically, we assume samples are generated according to the model:

$$Y = X\beta^0 + \epsilon$$

and seek to approximate the unknown vector of coefficients β^0. Here $Y \in \mathbb{R}^n$ denotes the vector of responses, $X \in \mathbb{R}^{n \times p}$ the feature matrix, and $\epsilon \in \mathbb{R}^n$ a vector of mean zero i.i.d. random errors. We will use the convention that x_i (lowercase) represents the ith row of X, and X_j (uppercase) the j-th column. When $n \geq p$, and X has full rank, we can solve for the unknown parameter vector using the Ordinary Least Squares (OLS) method:

$$\hat{\beta}^{OLS} := \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2$$

However, in the domain $p > n$, the least squares problem is ill-posed, and $\hat{\beta}^{OLS}$ can be chosen to fit the data exactly. This results in poor generalization error, and is known as over-fitting. A standard remedy for this problem is to add a regularization term to the loss function, which reduces estimator variance at the cost of introducing some small bias. A variety of different regularization strategies exist, which encode different kinds of prior knowledge.

In particular, in the $p > n$ setting, we often have prior reason to believe that the response can

be explained by a relatively small number of the covariates included in the feature matrix. In this situation, predictive accuracy can be greatly enhanced if the set of relevant features can be correctly identified. This is known as the problem of feature selection. In addition to improving predictive accuracy, feature selection can also improve model interpretability, which is a worthy goal in its own right.

When seeking to perform feature selection and regularization, a popular approach is to use the Lasso (Least Absolute Shrinkage and Selection Operator) [1]. Assuming that the response vector Y has been centered, and the columns of X centered and standardized, the Lasso estimator $\hat{\beta}_\lambda^L$ is defined as the minimizer over $\beta \in \mathbb{R}^p$ of the objective function:

$$\frac{1}{2}\left\|Y - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_1 \tag{1}$$

Here, λ is a free parameter, generally chosen by cross-validation, which governs the degree of regularization. In Figure 1A, we feature a comparison of the Least Squares and Lasso estimators on a synthetic data set. The defining feature of the Lasso is that it typically yields a sparse solution vector, making it a useful tool for feature selection. This property can be understood from two complementary perspectives. On one hand, the Lasso problem (eq 1) is equivalent to the problem:

$$\begin{aligned} &\min_{\beta\in\mathbb{R}^p} \left\|Y - X\beta\right\|_2^2 \\ &\text{subject to } \left\|\beta\right\|_1 \leq t \end{aligned} \tag{2}$$

where $t = \left\|\hat{\beta}_\lambda^L\right\|_1$. Geometrically, problem (2) is equivalent to solving for the lowest level set of $\left\|Y - X\beta\right\|_2^2$ that intersects the ℓ_1 ball $\{\beta \in \mathbb{R}^p : \left\|\beta\right\|_1 \leq t\}$ centered at the origin. If the point of intersection is at a corner along the jth axis, as illustrated in Figure 1B, the corresponding parameter β_j is equal to zero. This is frequently the case when p and λ are large, and thus the Lasso yields sparse solutions under these conditions.

The tendency of the Lasso to produce sparse solution vectors can also be understood from a Bayesian viewpoint. If we assume that that the coefficients β_j have independent Laplace prior distributions, $P(\beta_j) = \frac{1}{2\tau}\exp\left(-|\beta_j|/\tau\right)$, and that the random error is Gaussian with standard deviation σ, then the log posterior distribution of β is given by:

$$\log P(\beta|X,Y) = \log\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\left(\frac{1}{2\tau}\right)^p\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - x_i^T\beta)^2 - \frac{1}{\tau}\sum_{j=1}^{p}|\beta_j|$$

If we set $\tau = \frac{\sigma^2}{\lambda}$, it follows that the maximum a posteriori estimate is equivalent to the Lasso estimate. The fact that the Laplace distribution is sharply peaked at zero reflects the propensity of the Lasso to produce sparse solutions.

While the Lasso problem does not have an analytic solution, the KKT optimality conditions dictate that $\hat{\beta}_\lambda^L$ must satisfy:

$$X^T(Y - X\hat{\beta}_\lambda^L) = \lambda\hat{\nu} \tag{3}$$

where $\hat{\nu}$ is defined as:

$$\hat{\nu}_i \in \begin{cases} \text{sign}((\hat{\beta}_\lambda^L)_i) & \text{if } (\hat{\beta}_\lambda^L)_i \neq 0 \\ [-1,1] & \text{if } (\hat{\beta}_\lambda^L)_i = 0 \end{cases}$$

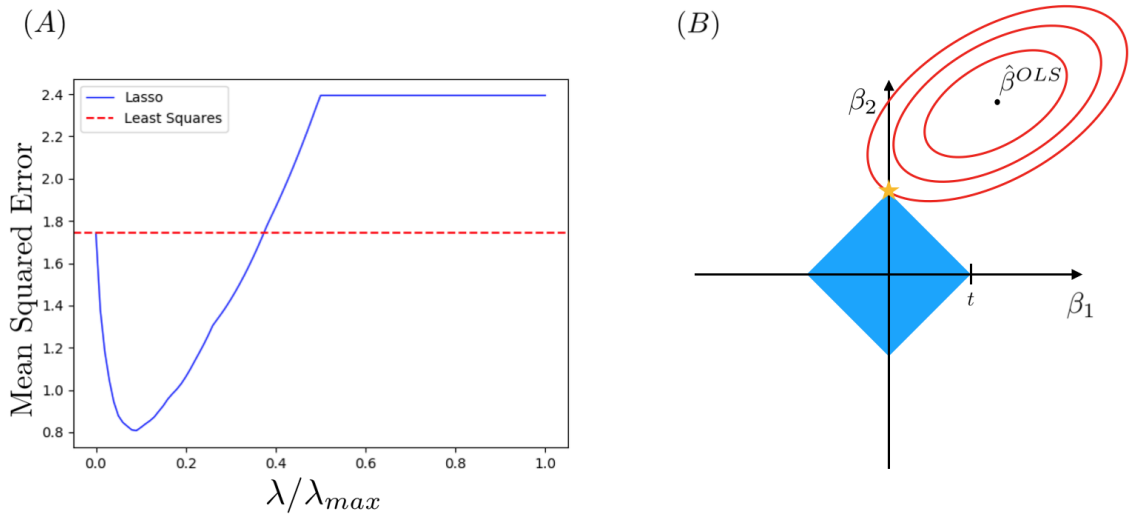

Figure 1: (A) Comparison of the mean squared error of the Least Squares (dotted red) and Lasso (blue) estimators on a synthetic data set with $n = 100$ and $p = 30$. The true parameter vector β^0 contained 15 non-zero entries. Features were drawn from independent $\mathcal{N}(0, 1)$ distributions, and the error terms were drawn from independent $\mathcal{N}(0, 0.2)$ distributions. The response vector was then computed as $Y = X\beta^0 + \epsilon$. The Lasso estimator was calculated across an equally spaced grid of 200 λ values ranging from 0 to λ_{max}. From the plot above, it is evident that the Lasso can outperform Least Squares if λ is appropriately tuned. (B) Geometric interpretation of the Lagrangian form of the Lasso problem (closely adapted from [2]). The red ellipses represent the contours of the function $\|Y - X\beta\|_2^2$, and the blue diamond represents the $\ell - 1$ ball $\{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq t\}$. The Lasso solution is given by the yellow star. In this example $(\hat{\beta}_\lambda^L)_2 = 0$.

These are necessary and sufficient conditions. Although Lasso solutions are not necessarily unique, it can be shown that if the entries of $X \in \mathbb{R}^{n \times p}$ are drawn from a continuous probability distributions, then the solution is unique with probability one [3]. Moreover, the fitted values $X\hat{\beta}_\lambda^L$ are always unique [2].

From these facts, it is straightforward to prove that if $\lambda \geq \lambda_{max} := \left\|X^T Y\right\|_\infty$, then all components of $\hat{\beta}_\lambda^L$ are necessarily equal to zero. By inspection, we see that $\hat{\beta}_\lambda^L = 0$ satisfies the KKT conditions, provided that $\lambda \geq \lambda_{max}$. To rule out the possibility of an additional solution, we note that the fitted value $X\hat{\beta}_\lambda^L = 0$ must be unique. Since the Lasso objective function must have the same value at all solutions, it follows that all solutions must have the same $\ell - 1$ norm. Therefore, the solution $\hat{\beta}_\lambda^L = 0$ is necessarily unique, completing the proof. As a consequence of this result, we only ever need to tune λ over the interval $(0, \lambda_{max})$.

Equipped with these theoretical results, we now shift our focus to the subject of screening rules for the Lasso. To set the stage for this discussion, it will first be necessary to formulate the Lagrangian dual of the Lasso problem. Then, we will turn to the topic of SAFE rules, which are a class of screening rules inspired by the geometry of the dual problem.

Formulation of the Dual Problem

We can reformulate the Lasso as a constrained optimization problem by introducing an additional variable, Z:

$$\min_{\beta \in \mathbb{R}^p, Z \in \mathbb{R}^n} \frac{1}{2} \|Y - Z\|_2^2 + \lambda \|\beta\|_1$$
$$\text{subject to } Z - X\beta = 0$$

The Lagrangian function $L : \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ associated with this constrained optimization problem is:

$$L(\beta, Z, \mu) = \frac{1}{2} \|Y - Z\|_2^2 + \lambda \|\beta\|_1 + \mu^T (Z - X\beta)$$

In the above expression, μ_i is the dual variable associated with the equality constraint $(Z - X\beta)_i = 0$. The dual function, $G : \mathbb{R}^n \to \mathbb{R}$, is computed by minimizing the Lagrangian over the primal variables:

$$G(\mu) = \inf_{\beta \in \mathbb{R}^p, Z \in \mathbb{R}^n} L(\beta, Z, \mu)$$

Making the change of variables $\theta = \mu / \lambda$, we have:

$$G(\theta) = \inf_{Z \in \mathbb{R}^p} (\frac{1}{2} \|Y - Z\|_2^2 + \lambda \theta^T Z) + \inf_{\beta \in \mathbb{R}^p} (\lambda \|\beta\|_1 - \lambda \theta^T X \beta)$$

The left-hand term is minimized when $Z = Y - \lambda\theta$, and is thus equal to $\frac{1}{2} \|Y\|_2^2 - \frac{\lambda^2}{2} \left\|\theta - \frac{Y}{\lambda}\right\|_2^2$). The right-hand term can be rewritten as $-\lambda f^*(X^T\theta)$, where f^* is the convex conjugate of the absolute value function. Thus, the right-hand term is equal to:

$$\begin{cases} 0 & \text{if } \left\|X^T\theta\right\|_\infty \leq 1 \\ -\infty & \text{otherwise} \end{cases}$$

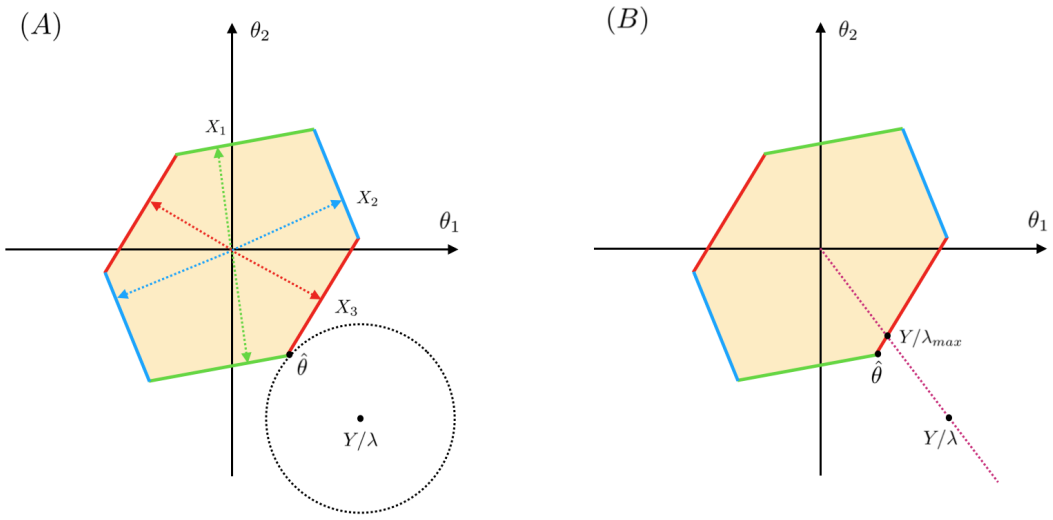

Figure 2: (A) An illustration of the geometry of the dual problem. In this example, there are three constraint vectors, X_1, X_2 and X_3, shown in green, blue and red. The feasible region, shaded in yellow, is bounded by planes normal to these vectors. The solution to the dual problem, labelled above as $\hat{\theta}$, is given by the projection of Y/λ onto the feasible region. Since $|X_1^T\hat{\theta}| = |X_3^T\hat{\theta}| = 1$, we see that the constraints X_1 and X_3 are active, while X_2 is inactive. Thus, X_2 can be eliminated from the feature matrix. (B) The point Y/λ_{max} lies on the boundary of the feasible set. When the regularization parameter is greater than λ_{max}, Y/λ lies within the feasible set, and therefore $\hat{\theta}(\lambda) = Y/\lambda$

Therefore, the dual problem is:

$$\begin{aligned} \max_{\theta \in \mathbb{R}^n} \frac{1}{2} \|Y\|_2^2 - \frac{\lambda^2}{2} \left\| \theta - \frac{Y}{\lambda} \right\|_2^2 \\ \text{subject to } \left\| X^T\theta \right\|_\infty \leq 1 \end{aligned} \tag{4}$$

The dual function is strongly convex and thus has a unique solution. Geometrically, the unique dual solution can be interpreted as the projection of of $\frac{Y}{\lambda}$ onto the polyhedral set $\Delta_X := \{\theta \in \mathbb{R}^n : \left\| X^T\theta \right\|_\infty \leq 1\}$. Furthermore, since the primal problem is convex, and satisfies Slater's condition, the primal and dual solutions are equal. The primal optimal point and dual optimal point, denoted $\hat{\beta}$ and $\hat{\theta}$ respectively, are related by:

$$Y = X\hat{\beta} + \lambda\hat{\theta} \tag{5}$$

Note that the superscript L and subscript λ have been dropped, as henceforth we are only concerned with the Lasso estimator. Recalling the KKT conditions for the Lasso problem (eq 3), we see that:

$$X^T\hat{\theta} = \hat{\nu} \tag{6}$$

where

$$\hat{\nu} \in \begin{cases} \text{sign}(\hat{\beta}_i) & \text{if } \hat{\beta}_i \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i = 0 \end{cases}$$

SAFE Screening Tests

Although in theory the Lasso estimator has desirable properties in the high-dimensional limit [2], in practice many of the standard optimization algorithms used to solve the Lasso problem are computationally infeasible for high-dimensional inputs [4]. Thus, there has been considerable interest in devising new methods to accelerate Lasso solvers [5], [6]. One approach (referred to as "screening") is to identify a subset of features guaranteed to have zero weight in the solution $\hat{\beta}^L_\lambda$, prior to solving the Lasso problem. Once identified, these features can be trimmed from the feature matrix, resulting in a smaller, more computationally tractable problem.

It is straightforward to prove that the removal of these features does not impact the optimality of the solution [7]. Let $\mathcal{I} = [1, ..p]$ denote the total set of features, and $\mathcal{S} \subset \mathcal{I}$ an arbitrary subset of these features. Additionally, given a vector $q \in \mathbb{R}^p$, let $q_{\downarrow\mathcal{S}}$ denote the vector in $\mathbb{R}^{|S|}$ obtained by sampling q at the indices in $\mathcal{S}$. In the same vein, for a feature matrix $X \in \mathbb{R}^{n\times p}$, let $X_{\downarrow\mathcal{S}}$ denote the matrix obtained by sampling the columns of X with indices in $\mathcal{S}$. Finally, for a vector $w \in \mathbb{R}^{|\mathcal{S}|}$, let $w^{\uparrow\mathcal{S}}$ denote the upsampling of w: the vector in $\mathbb{R}^p$ such that $(w^{\uparrow\mathcal{S}})_{\downarrow\mathcal{S}} = w$ and $(w^{\uparrow\mathcal{S}})_j = 0, \forall j \in \overline{\mathcal{S}}$. Now, if $\hat{\beta}$ is a solution to the original Lasso problem, and $\hat{\alpha}$ is a solution to the reduced Lasso problem on dictionary $X_{\downarrow\mathcal{S}}$, then by definition we have:

$$\begin{aligned}
&\frac{1}{2}\left\|Y - X\hat{\beta}\right\|_2^2 + \lambda\left\|\hat{\beta}\right\|_1 \\
\leq &\frac{1}{2}\left\|Y - X\hat{\alpha}^{\uparrow\mathcal{S}}\right\|_2^2 + \lambda\left\|\hat{\alpha}^{\uparrow\mathcal{S}}\right\|_1 \\
= &\frac{1}{2}\left\|Y - X_{\downarrow\mathcal{S}}\hat{\alpha}\right\|_2^2 + \lambda\left\|\hat{\alpha}\right\|_1 \\
\leq &\frac{1}{2}\left\|Y - X_{\downarrow\mathcal{S}}\hat{\beta}_{\downarrow\mathcal{S}}\right\|_2^2 + \lambda\left\|\hat{\beta}_{\downarrow\mathcal{S}}\right\|_1
\end{aligned} \tag{7}$$

Now, if $\hat{\beta}_j = 0, \forall j \in \overline{\mathcal{S}}$, then:

$$\frac{1}{2}\left\|Y - X\hat{\beta}\right\|_2^2 + \lambda\left\|\hat{\beta}\right\|_1 = \frac{1}{2}\left\|Y - X_{\downarrow\mathcal{S}}\hat{\beta}_{\downarrow\mathcal{S}}\right\|_2^2 + \lambda\left\|\hat{\beta}_{\downarrow\mathcal{S}}\right\|_1$$

Therefore, in this case, the chain of inequalities in (7) must in fact be a chain of equalities, and $\hat{\alpha}^{\uparrow\mathcal{S}}$ must be a solution to the original Lasso problem. In conclusion, if we can find a identify a subset of features $\overline{\mathcal{S}}$ guaranteed to have zero weight in the solution $\hat{\beta}$ (referred to as "inactive" features), it suffices to solve the Lasso problem on the reduced dictionary $X_{\downarrow\mathcal{S}}$, and then upsample the solution vector.

The challenge is therefore to devise a screening test that can identify which features are inactive. A variety of fast screening tests have been proposed in the literature, but many of these are heuristic in nature, and are not guaranteed to preserve the set of active features [5]. However, a recently developed sub-class of screening tests, known as SAFE (SAfe Feature Elimination) tests, do enjoy this theoretical guarantee [8]. The key insight in the development of SAFE tests comes from considering the geometry of the dual problem. As a direct consequence of equation (6), we know that:

$$|X_k^T\hat{\theta}| < 1 \rightarrow \hat{\beta}_k = 0 \tag{8}$$

Thus, if we knew the dual solution $\hat{\theta}$, it would be straightforward to identify the set of inactive features. However, solving the dual problem is a computationally intensive task. Hence, the approach of SAFE tests is instead to bound the dual solution within a compact "safe region", $\mathcal{R} \in \mathbb{R}^n$, and then to solve for the maximum of $|X_k^T\theta|$ over all points in the safe region. If we can verify that $\max_{\theta\in\mathcal{R}} |X_k^T\theta| < 1$, then it follows that $|X_k^T\hat{\theta}| < 1$, and therefore $\hat{\beta}_k = 0$. A major benefit of this approach is that each feature can be tested independently, and thus the procedure is easy to parallelize [8].

A variety of SAFE tests have been proposed in the literature, which differ based on the specifications of the safe region. Ideally, this region should be as constrained as possible, since narrowing the region can only increase the number of features eliminated. This is clear, because if $\mathcal{R}_1 \subseteq \mathcal{R}_2$, it follows that $\max_{\theta\in\mathcal{R}_1} |X_k^T\theta| \leq \max_{\theta\in\mathcal{R}_2} |X_k^T\theta|$, and so any feature rejected by an $\mathcal{R}_2$ SAFE test is also rejected by an $\mathcal{R}_1$ SAFE test. Furthermore, it is desirable that the support function $\sigma_\mathcal{R} : \mathbb{R}^n \to \mathbb{R}$, defined as $\sigma_\mathcal{R}(X_k) = \max_{\theta\in\mathcal{R}} X_k^T\theta$, is simple to evaluate over the region. It will be convenient to formulate SAFE tests as functions, $\mathcal{T}_\mathcal{R}$, mapping a given feature vector to a binary output:

$$\mathcal{T}_\mathcal{R}(X_k) = \begin{cases} 1 & \text{if } \max(\sigma_\mathcal{R}(X_k), \sigma_\mathcal{R}(-X_k)) < 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Under this convention, features mapped to '1' can be safely eliminated. We will now review several of the different types of SAFE screening tests that have been proposed in the literature, and compare the efficacy of these tests.

Static SAFE Tests

The concept of a SAFE test was originally introduced by El Ghaoui et al. in the 2010 paper "Safe Feature Elimination in Sparse Supervised Learning". The test proposed in this paper, which was later refined in [9] [10], is meant to be applied before solving the Lasso problem. Since screening occurs only once, it has become known in the literature as a "static" SAFE test (in contrast to "dynamic" tests, which are meant to be interlaced throughout the Lasso-solving algorithm). Two main classes of static SAFE tests have been proposed: sphere tests, and dome tests [7]. These differ based on the constraints employed to bound the dual optimal point.

Sphere Tests

In a sphere test, the dual optimal point $\hat{\theta}$ is bounded within a closed ball $\mathcal{B}(c, r) = \{\theta : \|\theta - c\|_2 \leq r\}$ with center c and radius r. Sphere tests are simple to implement, because the support function for a sphere can be evaluated quickly. For all $\theta \in \mathcal{B}(c, r)$, it holds that:

$$\begin{aligned} X_k^T\theta &= X_k^T c + X_k^T(\theta - c) \\ &\leq X_k^T c + \left\|X_k^T\right\|_2 \|\theta - c\|_2 \qquad \text{(Cauchy Schwarz)} \\ &\leq X_k^T c + r \|X_k\|_2 \end{aligned}$$

where equality holds when $\theta - c$ is aligned with X_k. Therefore, the support function is given by $\sigma_{\mathcal{B}(c,r)}(X_k) = X_k^T c + r \|X_k\|_2$, and so the condition for eliminating feature X_k is:

$$\mathcal{T}_{\mathcal{B}(c,r)}(X_k) = \begin{cases} 1 & \text{if } |X_k^T c| + r \|X_k\|_2 < 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Now, we address the problem of selecting the parameters c and r. Ideally, we would like to choose a set of parameters that give a tight bound on the dual optimal point, and are cheap to compute. Unfortunately, these goals are often in conflict. As a starting point, we will review the "Basic SAFE Lasso test" proposed in [8], which serves as a building block for many other tests. More complex sphere tests will be discussed in the sections on sequential and dynamic safe rules.

The Basic SAFE Lasso test takes advantage of the fact that the dual optimal point $\hat{\theta}$ is the closest point in the feasible region to Y/λ. Thus, for any dual feasible point $\theta_F \in \Delta_X$, it follows that:

$$\left\| \hat{\theta} - Y/\lambda \right\|_2 \leq \|\theta_F - Y/\lambda\|_2 \tag{11}$$

This implies that $\hat{\theta}$ must be contained within the sphere centered at $c = Y/\lambda$ with radius $r = \|\theta_F - Y/\lambda\|_2$. In the Basic SAFE Lasso test, the particular θ_F used in this bound is Y/λ_{max}, which is dual feasible since $\left\|X^T Y/\lambda_{max}\right\|_\infty = \frac{1}{\lambda_{max}} \left\|X^T Y\right\|_\infty = 1$. Therefore, the Basic SAFE Lasso test is given by:

$$\mathcal{T}_{\mathcal{B}(c,r)}(X_k) = \begin{cases} 1 & \text{if } \frac{1}{\lambda}|X_k^T Y| + (\frac{1}{\lambda} - \frac{1}{\lambda_{max}}) \|Y\|_2 \|X_k\|_2 < 1 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

which follows from substituting $c = Y/\lambda$ and $r = (\frac{1}{\lambda} - \frac{1}{\lambda_{max}})$ into equation 10. If Y and the columns of X are standardized, this test becomes:

$$\mathcal{T}_{\mathcal{B}(c,r)}(X_k) = \begin{cases} 1 & \text{if } |\cos \alpha_k| < \lambda + \frac{\lambda}{\lambda_{max}} - 1 \\ 0 & \text{otherwise} \end{cases}$$

where α_k is the angle between X_k and Y. Thus, in this standardized case, we see that the Basic SAFE Lasso test reduces to a correlation test. A drawback of this test is that it is ineffective for small values of λ/λ_{max}. If the variables are standardized, we see that no features at all are eliminated in the regime:

$$\frac{\lambda}{\lambda_{max}} \leq \min_{j \in [p]} \left(\frac{1 + |\cos \alpha_j|}{1 + \lambda_{max}} \right) \tag{13}$$

Dome Tests

Dome tests bound the dual optimal point $\hat{\theta}$ within the intersection of a closed ball $\mathcal{B}(c, r)$ and a closed half-space $\mathcal{H}(n, p) := \{\theta : n^T \theta \leq p\}$, as pictured in Figure 3A [7]. If we take the convention that n is a unit vector, and $p \geq 0$, then every dome can be uniquely characterized by the four parameters c, r, n and p. In order to simplify the the derivations in this section, it will be convenient to define the additional feature ψ_d, which denotes the signed distance, in the direction of -n, between the sphere center c and the bounding hyper-plane, as a fraction of the radius of the sphere. To ensure that the dome is non-degenerate, we will require that $-1 < \psi_d < 1$. By basic Euclidean

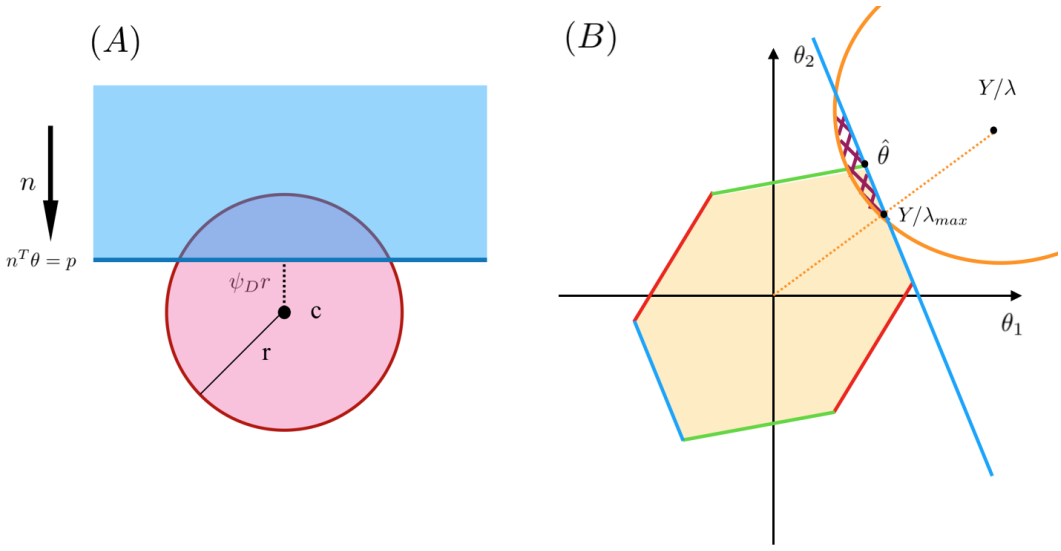

Figure 3: (A) A dome $\mathcal{D}(c, r, n, p)$ formed by the intersection of the closed ball $\mathcal{B}(c, r) = \{\theta : \|\theta - c\|_2 \leq r\}$ and the half-space $\mathcal{H}(n, p) = \{\theta : n^T\theta \leq p\}$. (B) An illustration of the SAFE region (marked by purple gridlines) employed in the default dome test.

geometry, it follows that $\psi_d = \frac{n^T c - p}{r}$. As in the discussion on sphere tests, we will begin by solving for the support function of an arbitrary dome $\mathcal{D}(c, r; n, p) := \{\theta : n^T\theta \leq p, \|\theta - c\|_2 \leq r\}$. Next, we will turn to the question of parameter selection, and review a sample dome test from the literature.

The support function $\sigma_{\mathcal{D}(c,r;n,p)}(X_k)$ can be expressed as a constrained optimization problem in standard form:

$$\begin{aligned} \min_{\theta \in \mathbb{R}^n} \quad & -X_k^T\theta \\ \text{subject to} \quad & (\theta - c)^T(\theta - c) - r^2 \leq 0 \\ & n^T\theta - p \leq 0 \end{aligned} \tag{14}$$

Since affine functions are convex, and the intersection of convex sets is also a convex, this is a convex optimization problem. Moreover, since $-1 < \psi_d < 1$, Slater's condition holds, and hence the duality gap is zero. Therefore, we can choose to solve the dual problem instead. This is given by:

$$\max_{\mu,\gamma \geq 0} \min_{\theta \in \mathbb{R}^n} -X_k^T\theta + \mu\left((\theta - c)^T(\theta - c) - r^2\right) + \gamma(n^T\theta - p) \tag{15}$$

Setting the derivative with respect to θ equal to zero yields:

$$\theta = c + \frac{1}{2\mu}(X_k - \gamma n) \tag{16}$$

This is the link between the primal and dual problems. After substituting this expression for θ back into (15), the problem becomes:

$$\max_{\mu,\gamma\geq 0} -\frac{1}{4\mu}\left\|X_k\right\|_2^2 - X_k^T c - \frac{\gamma^2}{4\mu} - \mu r^2 + \frac{\gamma}{2\mu}X_k^T n + \gamma r\psi_d$$

Holding μ constant, we see that the dual function is quadratic in γ, and is maximized when:

$$\gamma = \max(0, 2\mu r\psi_d + X_k^T n) \tag{17}$$

Furthermore, as a function of μ, the dual function is maximized when:

$$\mu = \frac{1}{2r}\left\|X_k - \gamma n\right\|_2 \tag{18}$$

By substituting equation (17) into (18), working through a case check, and then substituting the result for μ back into equation (17) it can be verified that:

$$\mu = \begin{cases} \frac{1}{2r}\sqrt{\frac{\|X_k\|_2^2 - (X_k^T n)^2}{1-\psi_d^2}} & \text{if } X_k^T n \geq -\psi_d \left\|X_k\right\|_2 \\ \frac{1}{2r}\left\|X_k\right\|_2 & \text{if } X_k^T n < -\psi_d \left\|X_k\right\|_2 \end{cases}$$

$$\gamma = \begin{cases} X_k^T n + \psi_d\sqrt{\frac{\|X_k\|_2^2 - (X_k^T n)^2}{1-\psi_d^2}} & \text{if } X_k^T n \geq -\psi_d \left\|X_k\right\|_2 \\ 0 & \text{if } X_k^T n < -\psi_d \left\|X_k\right\|_2 \end{cases}$$

Substituting these values into (16) gives the value of θ which minimizes the primal problem. After simplifying, we see that the support function is equal to:

$$\sigma_{\mathcal{D}(c,r;n,p)}(X_k) = X_k^T c + M_{\mathcal{D}}(X_k, n)$$

where $M_{\mathcal{D}}(X_k, n)$ is given by:

$$\begin{cases} -\psi_d r X_k^T n + r\sqrt{\left\|X_k\right\|_2^2 - (X_k^T n)^2}\sqrt{1-\psi_d^2} & \text{if } X_k^T n \geq -\psi_d \left\|X_k\right\|_2 \\ r\left\|X_k\right\|_2 & \text{if } X_k^T n < -\psi_d \left\|X_k\right\|_2 \end{cases}$$

When $X_k^T n = -\psi_d \left\|X_k\right\|_2$, the two piecewise terms are equal, and so this function is continuous. If $X_k^T n < -\psi_d \left\|X_k\right\|_2$, we see that the support function is equivalent to that of the ball $\mathcal{B}(c, r)$. However, when $X_k^T n \geq -\psi_d \left\|X_k\right\|_2$, it can be easily verified that $\sigma_{\mathcal{D}(c,r;n,p)}(X_k) \leq \sigma_{\mathcal{B}(c,r)}(X_k)$, and thus dome tests have greater rejection power. The SAFE test for a non-degenerate dome $\mathcal{D}(c, r; n, p)$ is:

$$\mathcal{T}_{\mathcal{D}(c,r;n,p)}(X_k) = \begin{cases} 1 & \text{if } M_{\mathcal{D}}(-X_k, n) - 1 < X_k^T c < 1 - M_{\mathcal{D}}(X_k, n) \\ 0 & \text{otherwise} \end{cases}$$

Having arrived at the general form of a dome SAFE test, we now turn to address the question of parameter selection. As selection of the parameters c and r was covered in the sphere test section, we focus on the choice of n and p. A natural starting point is to consider the set of hyper-planes which bound the dual feasible region. As $\hat{\theta}$ is by definition dual feasible, it must satisfy $\left\|X^T\hat{\theta}\right\|_\infty \leq 1$.

Therefore, defining $\mathcal{F} = \{\pm X_i\}_{i=1}^p$, we see that $\forall f \in \mathcal{F}$, the half-space $\mathcal{H}(f/\|f\|_2, 1/\|f\|_2)$ contains $\hat{\theta}$. Out of this set, we can choose the half-space which gives the smallest bounding dome (or equivalently, the largest value of $\psi_{\mathcal{D}}$). This is the half-space $\mathcal{H}(f_s/\|f_s)\|_2, 1/\|f_s\|_2)$, such that

$$f_s = \arg\max_{f \in \mathcal{F}} \frac{f^T c - 1}{\|f\|_2}$$

If the columns of X are standardized, we see that f_s is just the element of $\mathcal{F}$ most correlated with c. Thus, if we use the spherical bound provided by the Basic SAFE Lasso test, then the corresponding dome (pictured in Figure 3B) is:

$$\mathcal{D}(Y/\lambda, 1/\lambda - 1/\lambda_{max}, f_s, 1) \tag{19}$$

where $f_s = \arg\max_{f \in \mathcal{F}} f^T Y$. Solving for the quantity f_s requires no additional work, because it falls out of the calculation for λ_{max}. This is referred to as the "Default Dome Test" [7]; sequential screening will enable more complex dome tests.

Sequential SAFE Tests

It is often the case that we seek to solve the Lasso problem over a grid $\{\lambda_k\}_{k=1}^G$ of λ values (in the process of cross-validation, for example) [10]. Sequential SAFE tests leverage the fact that consecutive values λ_k, λ_{k+1} in the grid are likely to have similar dual solutions $\hat{\theta}_k, \hat{\theta}_{k+1}$. Thus, knowledge of the Lasso solution $\hat{\beta}_k$ (with corresponding dual solution $\hat{\theta}_k = \frac{1}{\lambda_k}(Y - X\hat{\beta}_k)$) can enable us to construct a tighter SAFE region around $\hat{\theta}_{k+1}$.

Specifically, if we know that $\hat{\theta}_k$ is the dual solution when the regularization parameter is λ_k, then for all $\theta \in \Delta_X$ it holds that:

$$(Y/\lambda_k - \hat{\theta}_k)^T \theta \leq (Y/\lambda_k - \hat{\theta}_k)^T \hat{\theta}_k \tag{20}$$

Proof: Since Δ_X is convex, for any arbitrary $\theta \in \Delta_X$ and $\alpha \in (0, 1)$ we have $\hat{\theta}_k + \alpha(\theta - \hat{\theta}_k) \in \Delta_X$. Furthermore, by definition of the dual problem we know that $\hat{\theta}_k = \arg\min_{\theta \in \Delta_X} \|Y/\lambda_k - \theta\|_2^2$. This implies:

$$\left\|Y/\lambda_k - \hat{\theta}_k\right\|_2^2 \leq \left\|Y/\lambda_k - \hat{\theta}_k - \alpha(\theta - \hat{\theta}_k)\right\|_2^2$$

$$2\alpha(Y/\lambda_k - \hat{\theta}_k)^T \theta - \alpha^2 \left\|\theta - \hat{\theta}_k\right\|_2^2 \leq 2\alpha(Y/\lambda_k - \hat{\theta}_k)^T \hat{\theta}_k$$

Dividing by 2α and then taking $\alpha \to 0$ gives the desired result.

As $\hat{\theta}_{k+1}$ is necessarily dual feasible, inequality (20) tells us it is bounded by the half space $\mathcal{H}(n, p)$, where:

$$n = \frac{Y/\lambda_k - \hat{\theta}_k}{\left\|Y/\lambda_k - \hat{\theta}_k\right\|_2} \qquad p = n^T \hat{\theta}_k$$

This choice of half-space can be combined with any valid sphere test to give a dome test. One natural choice is to take $c = Y/\lambda$ and $r = \left\|Y/\lambda - \hat{\theta}_k\right\|_2$. We will refer to this particular test as "Sequential Screening Test 1".

Another sequential screening strategy that has been pursued involves constructing a SAFE sphere around the point $\hat{\theta}_k$ [11]. We will call this "Sequential Screening Test 2". The key insight in this approach is that projection onto a convex set is non-expansive. Therefore:

$$\begin{aligned} \left\| \hat{\theta}_{k+1} - \hat{\theta}_k \right\|_2 &\leq \|Y/\lambda_{k+1} - Y/\lambda_k\|_2 \\ &= |1/\lambda_{k+1} - 1/\lambda_k| \, \|Y\|_2 \end{aligned} \tag{21}$$

Proof: By equation (20) and dual feasibility of $\hat{\theta}_{k+1}$, we know that:

$$(Y/\lambda_k - \hat{\theta}_k)^T(\hat{\theta}_{k+1} - \hat{\theta}_k) \leq 0$$

By symmetry, we also have that:

$$(Y/\lambda_{k+1} - \hat{\theta}_{k+1})^T(\hat{\theta}_k - \hat{\theta}_{k+1}) \leq 0$$

Adding these two inequalities and rearranging yields:

$$\left\| \hat{\theta}_{k+1} - \hat{\theta}_k \right\|_2^2 \leq (\hat{\theta}_{k+1} - \hat{\theta}_k)^T(Y/\lambda_{k+1} - Y/\lambda_k)$$

Applying the Cauchy-Schwarz inequality and cancelling gives (21).

This result tells us that the sphere centered at $\hat{\theta}_k$ with radius $|1/\lambda_{k+1} - 1/\lambda_k| \, \|Y\|_2$ necessarily contains $\hat{\theta}_{k+1}$. If λ_k and λ_{k+1} are close in value, as is often the case when we are evaluating over a grid, this SAFE sphere has a very small radius. We can also take the intersection of this sphere with any of the half space constraints previously discussed to achieve even greater screening. One point to be wary of is that these calculations assume knowledge of the *exact* dual solution for a given value of λ. However, it is often the case when solving the Lasso problem over a grid that we only know the value of the dual solution to within some pre-defined error threshold. To ensure that no active features are wrongly eliminated, it is important to build this threshold into the screening criterion.

Dynamic SAFE Tests

Whereas Sequential SAFE tests leverage the computation done for a previous value of λ, Dynamic SAFE tests leverage the computations done *during* the Lasso-solving algorithm to create more refined SAFE tests [12]. Dynamic screening can be applied to any standard first-order algorithm which produces a sequence of iterates $\{\beta^{(j)}\}_{j \geq 0}$ converging to the optimal $\hat{\beta}$. The critical insight is that these iterates can be used to improve our estimation of the dual optimal point, and thus can be used to construct narrower SAFE regions. At every iteration of the Lasso-solving algorithm, we can therefore conduct a new and improved "dynamic" screening test leveraging this narrower bound.

Just as with static SAFE tests, a variety of dynamic SAFE tests have been proposed which employ different constraints to bound the dual optimal point. The most straightforward form of dynamic screening, first proposed in [12], involves iteratively constructing tighter SAFE spheres around the dual optimal point. As discussed in the section on sphere tests, knowledge of any dual feasible

point θ_F enables us to bound $\hat{\theta}$ within a sphere of radius $\|\theta_F - Y/\lambda\|_2$ centered at Y/λ. The closer that θ_F is to $\hat{\theta}$, the tighter this bound will be. Now, the convergence of $\{\beta^{(j)}\}_{j\geq 0}$ towards $\hat{\beta}$ also implies convergence of $\{\theta^{(j)}\}_{j\geq 0}$ towards $\hat{\theta}$, where $\theta^{(j)} = \frac{1}{\lambda}(Y - X\beta^{(j)})$ is given by the primal-dual link. This suggests that we might use $\{\theta^{(j)}\}_{j\geq 0}$ to construct a tighter spherical bound around $\hat{\theta}$. However, in order for the sphere test to apply, we must scale $\theta^{(j)}$ to ensure it is within the dual feasible region. Among all scaled versions of $\theta^{(j)}$ that are dual feasible, the one closest to Y/λ is $\mu_j\theta^{(j)}$, where:

$$\mu_j = \min\left(\max\left(\frac{Y^T\theta^{(j)}}{\lambda\left\|\theta^{(j)}\right\|_2^2}, \frac{-1}{\left\|X^T\theta^{(j)}\right\|_\infty}\right), \frac{1}{\left\|X^T\theta^{(j)}\right\|_\infty}\right) \tag{22}$$

Proof: The solution follows from solving the constrained optimization problem:

$$\begin{aligned} \min_{\mu\in\mathbb{R}} \quad & \left\|\mu\theta^{(j)} - Y/\lambda\right\|_2^2 \\ \text{subject to} \quad & \left\|X^T\mu\theta^{(j)}\right\|_\infty \leq 1 \end{aligned}$$

The objective function is convex quadratic in μ, and by a taking derivative we see that it is minimized when $\mu = \frac{Y^T\theta^{(j)}}{\lambda\left\|\theta^{(j)}\right\|_2^2}$. However, if this value is outside the feasible interval, $[\frac{-1}{\|X^T\theta_j\|_\infty}, \frac{1}{\|X^T\theta_j\|_\infty}]$, then we see that the function is maximized by taking the endpoint value $\mu = \text{sign}(Y^T\theta^{(j)})\frac{1}{\left\|X^T\theta^{(j)}\right\|_\infty}$. This result can be written compactly as (22).

Thus, if the jth iterate of the Lasso solving algorithm yields the vector $\beta^{(j)}$, the corresponding dynamic screening sphere test is given by:

$$\mathcal{T}_{\mathcal{B}(c,r)}(X_k) = \begin{cases} 1 & \text{if } \frac{1}{\lambda}|X_k^TY| + \frac{1}{\lambda}\left\|\mu_j(Y - X\beta^{(j)}) - Y\right\|_2 \|X_k\|_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

Computationally speaking, this test adds minimal overhead when interlaced with any optimization algorithm that already involves computing the gradient $-X^T(Y - X\beta^{(j)})$. However, similar to the static sphere test previously discussed, the dynamic sphere test is only effective when λ/λ_{max} is above a certain threshold. Specifically, no features are eliminated when:

$$1 \leq \min_{k\in[p]} \frac{1}{\lambda}|X_k^TY| + \frac{1}{\lambda}\left\|\mu_j(Y - X\beta^{(j)}) - Y\right\|_2 \|X_k\|_2 \tag{23}$$

This condition can made independent of the value of $\beta^{(j)}$ by noting:

$$\left\|\hat{\theta} - Y/\lambda\right\|_2 \leq \frac{1}{\lambda}\left\|\mu_j(Y - X\beta^{(j)}) - Y\right\|_2$$

This inequality follows from the fact that $\hat{\theta}$ is the closest dual feasible point to $\frac{Y}{\lambda}$, and $\frac{\mu_j}{\lambda}(Y - X\beta^{(j)})$ is dual feasible. Substituting this bound into equation (23), applying the triangle inequality, and rearranging terms, we see that the dynamic sphere test is useless once:

$$\frac{\lambda}{\lambda_{max}} \leq \min_{k\in[p]}\left(\frac{\|X_k\|_2\|Y\|_2 + |X_k^TY|}{\lambda_{max}(\left\|\hat{\theta}\right\|_2\|X_k\|_2 + 1)}\right)$$

However, by employing additional constraints, it is possible to devise a dynamic screening test guaranteed to eliminate all inactive features in finite time, for all values of λ. This brings us to our final class of SAFE tests: Gap SAFE tests.

Gap SAFE Tests

Gap SAFE tests employ both an upper and lower bound on the distance between $\hat{\theta}$ and Y/λ, leading to SAFE regions which converge to zero in diameter [13]. The key improvement comes from using the weak duality theorem as a constraint on the location of $\hat{\theta}$. Combined with the bound provided by the Basic Sphere Test, this yields the following result:

Consider any $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$. Denote $R_1(\beta) := \frac{1}{\lambda}(\|Y\|_2^2 - \|Y - X\beta\|_2^2 - 2\lambda \|\beta\|_1)_+^{1/2}$, $R_2(\theta) := \|\theta - Y/\lambda\|_2$, and $\widetilde{r}(\beta, \theta) := \sqrt{R_2(\theta)^2 - R_1(\beta)^2}$. Then it necessarily follows that:

$$\hat{\theta}^{(\lambda)} \in \mathcal{B}(\theta, \widetilde{r}(\beta, \theta)) \tag{24}$$

Proof: By weak duality, we know that the dual objective function evaluated at any dual feasible point provides a lower bound on the primal objective function. Therefore, for any $\theta \in \Delta_X$ and $\beta \in \mathbb{R}^p$, it holds that:

$$\frac{1}{2}\|Y\|_2^2 - \frac{\lambda^2}{2}\left\|\theta - \frac{Y}{\lambda}\right\|_2^2 \leq \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Rearranging, we see that:

$$\frac{\sqrt{(\|Y\|_2^2 - \|Y - X\beta\|_2^2 - 2\lambda\|\beta\|_1)_+}}{\lambda} \leq \left\|\theta - \frac{Y}{\lambda}\right\|_2 \tag{25}$$

Since this bound applies for all $\theta \in \Delta_X$, it must apply for $\hat{\theta}$. Thus, we see that $\hat{\theta}$ is located at least a distance of $R_1(\beta)$ away from the point Y/λ. Now, we also know from equation (11) that $\hat{\theta}$ is necessarily contained within the ball of radius $R_2(\theta)$ centered at Y/λ. Therefore, combining these two bounds, we see that $\hat{\theta}$ must be located within the annulus $A(Y/\lambda, R_1(\beta), R_2(\theta)) := \{z \in \mathbb{R}^n : R_1(\beta) \leq \|z - Y/\lambda\|_2 \leq R_2(\theta)$. Now, we know that the dual feasible region Δ_X is convex, and since the intersection of convex sets is convex, it follows that $\Delta_X \cap \mathcal{B}(Y/\lambda, R_2(\theta))$ is also convex. Furthermore, since (25) tells us that every $\theta \in \Delta_X$ is at least a distance of $R_1(\beta)$ from Y/λ, it follows that $\Delta_X \cap \mathcal{B}(Y/\lambda, R_2(\theta)) = \Delta_X \cap A(Y/\lambda, R_1(\beta), R_2(\theta))$. Therefore, since $\Delta_X \cap \mathcal{B}(Y/\lambda, R_2(\theta))$ is convex, we see that $\Delta_X \cap A(Y/\lambda, R_1(\beta), R_2(\theta))$ is also convex.

As θ and $\hat{\theta}$ are both contained within the set $\Delta_X \cap A(Y/\lambda, R_1(\beta), R_2(\theta))$, and this set is convex, it follows that every point on the line $[\theta, \hat{\theta}]$ is also contained within this set (where $[\theta, \hat{\theta}]$ denotes the line given by $\alpha\theta + (1-\alpha)\hat{\theta}$ for $\alpha \in [0, 1]$). Moreover, since every point on this line is dual feasible, and $\hat{\theta}$ is the closest dual feasible point to Y/λ, we have that $\hat{\theta}$ is the closest point on the line $[\theta, \hat{\theta}]$ to Y/λ.

Now, suppose for sake of contradiction that $\hat{\theta}$ is located a distance further than $\sqrt{R_2(\theta)^2 - R_1(\beta)^2}$ from θ. Since $\left\|Y/\lambda - \hat{\theta}\right\|_2 \geq R_1$ (as $\hat{\theta}$ is contained within the annulus), this would imply that

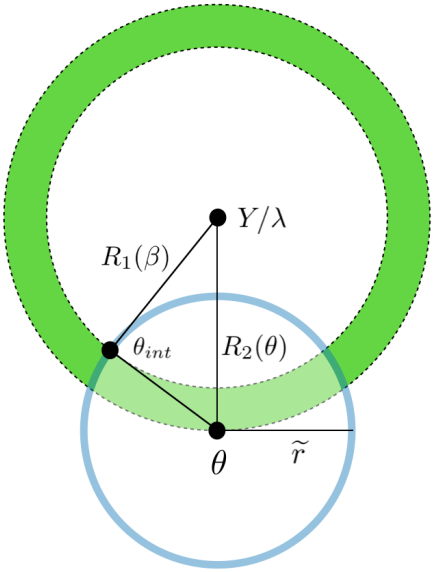

Figure 4: The Gap SAFE sphere is outlined in light blue. This sphere is centered at θ, and has radius $\widetilde{r}$. The annulus $A(Y/\lambda, R_1(\beta), R_2(\theta))$ is shown in green. The dual optimal point must be located within the intersection of the blue sphere and green annulus.

$\left\|\theta - \hat{\theta}\right\|_2^2 + \left\|Y/\lambda - \hat{\theta}\right\|_2^2 > R_1^2$. Equivalently, this means that the angle formed by the line segments $[\hat{\theta}, \theta]$ and $[\hat{\theta}, Y/\lambda]$ is necessarily acute. As such, the scalar projection of $[\hat{\theta}, Y/\lambda]$ onto $[\hat{\theta}, \theta]$ must be positive. This implies that there exists a point on the line $[\hat{\theta}, \theta]$ which is closer to Y/λ than $\hat{\theta}$ is. However, since we proved that all points on the line $[\hat{\theta}, \theta]$ are dual-feasible, this means that $\hat{\theta}$ is not the closest dual feasible point to Y/λ, a contradiction. Thus, we conclude that $\hat{\theta}$ must necessarily be located within a distance of $\widetilde{r}(\beta, \theta) = \sqrt{R_2(\theta)^2 - R_1(\beta)^2}$ from θ. This completes the proof.

Now, having established that $\hat{\theta}$ can be bounded within a sphere of radius $\widetilde{r}(\beta, \theta)$, we turn to examine the question of how this radius changes throughout the course of the optimization algorithm. Expanding the definition of $\widetilde{r}(\beta, \theta)$, we see:

$$\widetilde{r}(\beta, \theta)^2 = \|\theta - Y/\lambda\|_2^2 + \frac{1}{\lambda^2}(\|Y\|_2^2 - \|Y - X\beta\|_2^2 - 2\lambda \|\beta\|_1)_+ \leq \frac{2}{\lambda^2} G(\beta, \theta)$$

where $G(\beta, \theta) = \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 - \frac{1}{2}\|Y\|_2^2 + \frac{\lambda^2}{2}\|\theta - Y/\lambda\|_2^2$ is the duality gap. Therefore, since that strong duality holds, we see that the radius of the Gap SAFE region goes to zero as β, θ approach $\hat{\beta}, \hat{\theta}$. Accordingly, the Gap SAFE test described above is termed a converging SAFE test.

Numerical Results

Coordinate Descent

Having finished our review of the SAFE screening literature, we now assess the practical utility of screening through numerical experiments. In all experiments, we solve the Lasso problem using cyclic coordinate descent. Minimizing the Lasso objective function over the coordinate β_k yields the update:

$$\beta_k \leftarrow ST_\lambda\left(\beta_k + X_k^T(Y - X\beta)\right)$$

where $ST_\lambda(x)$ is the soft-thresholding operator, defined as:

$$ST_\lambda(x) := \text{sign}(x)\max(|x| - \lambda, 0)$$

This update follows from the subgradient optimality condition. Although cyclic coordinate descent is not guaranteed to converge to the global minimum of all convex non-differentiable functions, convergence is assured in the case of the Lasso problem [14].

A pseudo-code implementation of cyclic coordinate descent is presented in Algorithm 1. The actual code for all experiments was written in Python. One key detail to note is that strategic updating of the residual (line 7) allows us to perform the coordinate update step (line 6) in just $\mathcal{O}(n)$ operations. Therefore, one pass of cyclic coordinate descent takes a total of $\mathcal{O}(np)$ operations. The algorithm terminates when the duality gap falls below a pre-defined threshold ϵ. Evaluation of the duality gap involves first constructing a dual feasible point θ. This is done by taking $\theta = \frac{\mu}{\lambda}(Y - X\beta)$, where μ is defined as in equation (22). For computational efficiency, the duality gap is only evaluated once every 10 iterations.

Algorithm 1: Coordinate Descent

Input : $X, Y, \beta^{(0)}, \lambda, \epsilon$

`/*` $\beta^{(0)}$ `is an initial estimate of the Lasso solution; can use` $0 \in \mathbb{R}^p$ `as default */`

1 $\beta \leftarrow \beta^{(0)}$
2 $R \leftarrow Y - X\beta$
3 **for** j **in** `range` $(max\ iterations)$ **:**
4 **for** k **in** `range` (p) **:**
5 $\beta_k^{prev} \leftarrow \beta_k$
6 $\beta_k \leftarrow ST_\lambda(\beta_k + X_k^T R)$
7 $R \leftarrow R + X_k^T(\beta_k^{prev} - \beta_k)$
8 **if** $j \pmod{10} = 0$ **:**
9 **if** `Duality Gap(`X, Y, β, λ`)` $\leq \epsilon$ **:**
10 **break**
11 **return** β

Implementation of Screening Tests

We compare the performance of seven different screening tests in terms of rejection power and computational efficiency. The first two tests we consider are the Basic Sphere Test and the Default Dome Test, given by equations (12) and (19) respectively. These are both straightforward to implement: features are screened in a pre-processing step, and then the reduced feature matrix is passed as input into the coordinate descent algorithm. For all tests discussed, we set $\beta^{(0)}$ equal to the Lasso solution at the previous λ when solving the Lasso problem over a grid (referred to as a "warm start").

The next screening tests we consider are Sequential Screening Tests 1 and 2 (see "Sequential SAFE Tests" section for details). Similar to the static screening tests, these sequential tests are only applied as a pre-processing step, prior to running the coordinate descent algorithm. Furthermore, they are only relevant when solving the Lasso problem over a grid.

The fifth and sixth tests we examine are the dynamic versions of the Basic Sphere Test and Default Dome Test. These are interwoven with the coordinate decent algorithm, and screening is conducted once for every ten passes of coordinate descent. By storing recycled inner products in memory, the screening step is completed in $\mathcal{O}(n+p)$ time.

Finally, we consider the Gap SAFE Test given by equation (24). We experiment with three different implementations of this test: a standard version, an active set version, and a working set version. The standard version is similar in structure to the dynamic sphere and dome tests; screening is conducted once every ten iterations on the full set of potentially active feature vectors. The active set implementation (only relevant when solving over a grid) improves solver initialization by first solving the Lasso problem over the active set of the previous λ value. Since neighboring λ values in a grid are likely to have similar active sets, and generally only a small fraction of features are active, this strategy can enable us to hone in on the solution more quickly. Lastly, we consider a working set implementation of the Gap SAFE Test, modelled on "Algorithm A5G" as described in [15]. In the working set implementation, the coordinate descent update is performed only on a subset of features considered most likely to be active. Features are prioritized based on correlation with the residual, and this prioritization is reassessed dynamically as the residual changes. We take the size of the working set to be at most twice the size of the current active set.

We apply these screening tests to three different data sets. The first is the Wisconsin breast cancer data set, which contains $n = 569$ observations each with $p = 30$ features. Features describe the characteristics of cell nuclei present in cancerous breast tissue. The entries of the response vector are binary valued, signifying whether the cancer is benign or malignant. The second data set we use is synthetically generated, and has dimensions $n = 20$, $p = 1000$. The true parameter vector β used to generate the data contained 10 non-zero entries. Features were drawn from independent $\mathcal{N}(0, 1)$ distributions, and error terms were drawn from independent $\mathcal{N}(0, 0.2)$ distributions. The response vector was then computed as $Y = X\beta + \epsilon$. Our final data set is the standard Leukemia data set, which has dimensions $n = 72, p = 7129$. The feature matrix contains gene expression measurements conducted on 72 Leukemia patients. The response vector is binary valued, and signifies the type of Leukemia (AML vs. ALL).

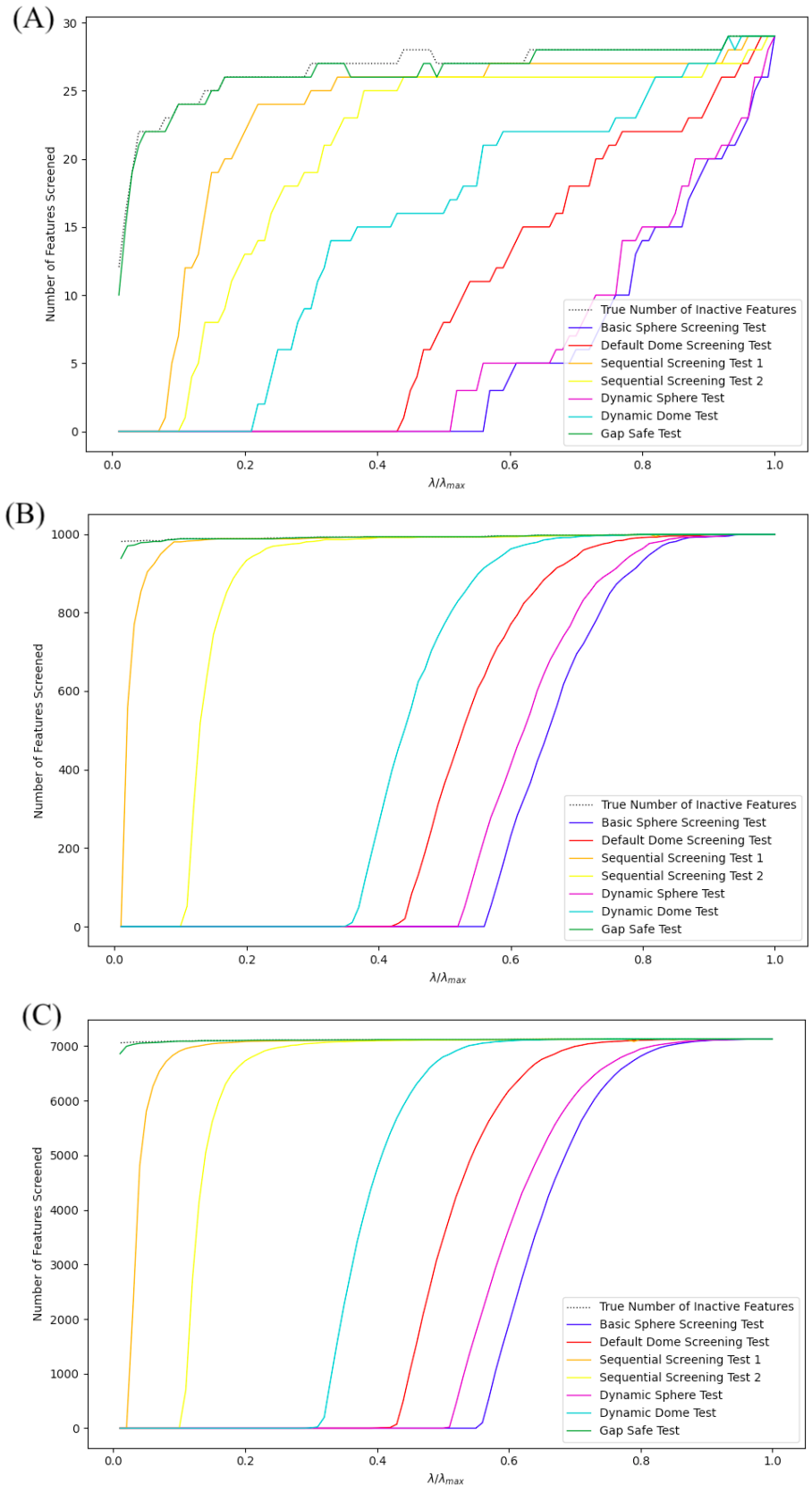

Figure 5: Number of features screened as a function of λ for the (A) breast cancer data set, (B) synthetic data set and (C) leukemia dataset.

Rejection Power

In Figure (5), we present the number of features screened as a function of λ for each of the three data sets. The duality gap threshold used is $\epsilon = 10^{-6}$. In each instance, the Gap SAFE test is the most effective, eliminating nearly all possible inactive features. When the duality gap threshold was lowered to $\epsilon = 10^{-8}$, the Gap SAFE test was observed to eliminate all possible inactive features, as predicted by theory. The least effective test on each data set is the Basic Sphere test, followed closely by the dynamic version of the Basic Sphere test. As the λ_{max} values of the three data sets are $0.7936, 0.7825$ and 0.7939 respectively, the bound given in (13) predicts that no features will be eliminated in the regimes $\frac{\lambda}{\lambda_{max}} < 0.5575$, $\frac{\lambda}{\lambda_{max}} < 0.5610$, and $\frac{\lambda}{\lambda_{max}} < 0.5574$. In practice, we see from Figure (5) that these bounds are tight. Moreover, the dynamic version of the Basic Sphere Test appears to offer little improvement over the static version. In comparison, the Default Dome test, and its corresponding dynamic version, offer improved screening power over a wider range of λ values. However, these tests are also ineffective for values of $\frac{\lambda}{\lambda_{max}}$ below 0.4. Sequential screening tests 1 and 2 are useful over an even wider range of λ values, but are not able to match the performance Gap SAFE test for very small values of λ. This is a crucial shortcoming, because solving the Lasso problem is very computationally intensive in this regime, and thus screening can make a large difference.

Time of Computation

In Figures (6) and (7), we feature a comparison of the time of computation for each of the different screening tests on the Leukemia data set. As the ultimate aim of screening is to accelerate Lasso solving algorithms, it is important to verify that its computational cost does not negate its benefits. Figure (6) deals with the situation of solving the Lasso problem over a grid of 100 equally spaced values of $\frac{\lambda}{\lambda_{max}}$ between 0 and 1. From sub-figure (6A), it is apparent that the time of computation is significantly greater for small values of $\frac{\lambda}{\lambda_{max}}$. In sub-figure (6B), we see more clearly that for large values of $\frac{\lambda}{\lambda_{max}}$, the Basic Sphere Test, Default Dome Test, Dynamic Sphere Test, and Dynamic Dome Test do provide some speed-up over baseline coordinate descent algorithm with no screening. However, it is already computationally inexpensive to solve the Lasso problem in this domain, so the gain is not substantial. Sequential Tests 1 and 2 are more successful over a broader range of $\frac{\lambda}{\lambda_{max}}$ values, but provide minimal benefit as $\frac{\lambda}{\lambda_{max}}$ approaches 0. When considering performance over the entire grid of $\frac{\lambda}{\lambda_{max}}$ values, it is clear that the Gap SAFE Tests are by far the most effective. As illustrated in sub-figure (6C), the standard Gap SAFE Test, Active Set Gap SAFE Test, and Working Set Gap SAFE Test outperformed baseline coordinate descent by factors of 6.0, 30.1, and 24.5, respectively. This is in large part due to their effectiveness at small values of $\frac{\lambda}{\lambda_{max}}$.

In Figure (7), we consider the "one-shot" setting, where the Lasso problem is solved at just a single value of λ. This value was chosen to be $\frac{\lambda}{\lambda_{max}} = 0.032397$, which is the optimal value as selected through cross-validation. Whereas previously the value of $\beta^{(0)}$ was initialized to be the Lasso solution at the preceding value of λ, now it is initialized at zero. We do not include Sequential Tests 1 and 2, or the Active Set Gap SAFE Test in this comparison, as these tests are only meaningful when solving the Lasso problem over a grid. From Figure (7), it is apparent that the Working Set Gap SAFE Test offers considerable improvement over the standard Gap SAFE Test. Moreover, the advantage afforded by both Gap SAFE Tests increases as the duality gap decreases, when compared to baseline coordinate descent.

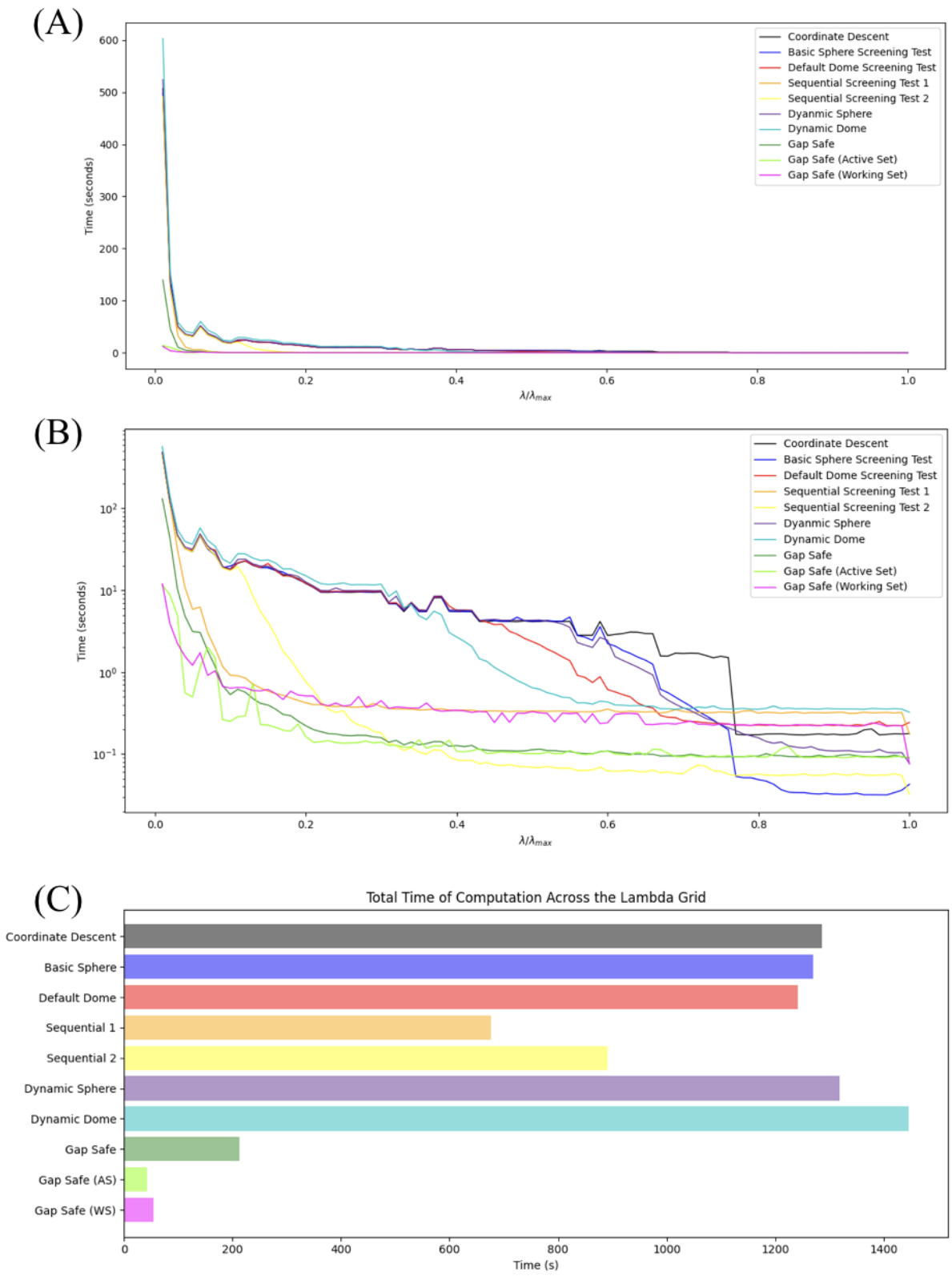

Figure 6: (A) Time of computation as a function of λ for the leukemia data set. (B) Graph from (A) presented on a logarithmic scale. (C) Total time of computation integrated across the grid of λ values used in (A).

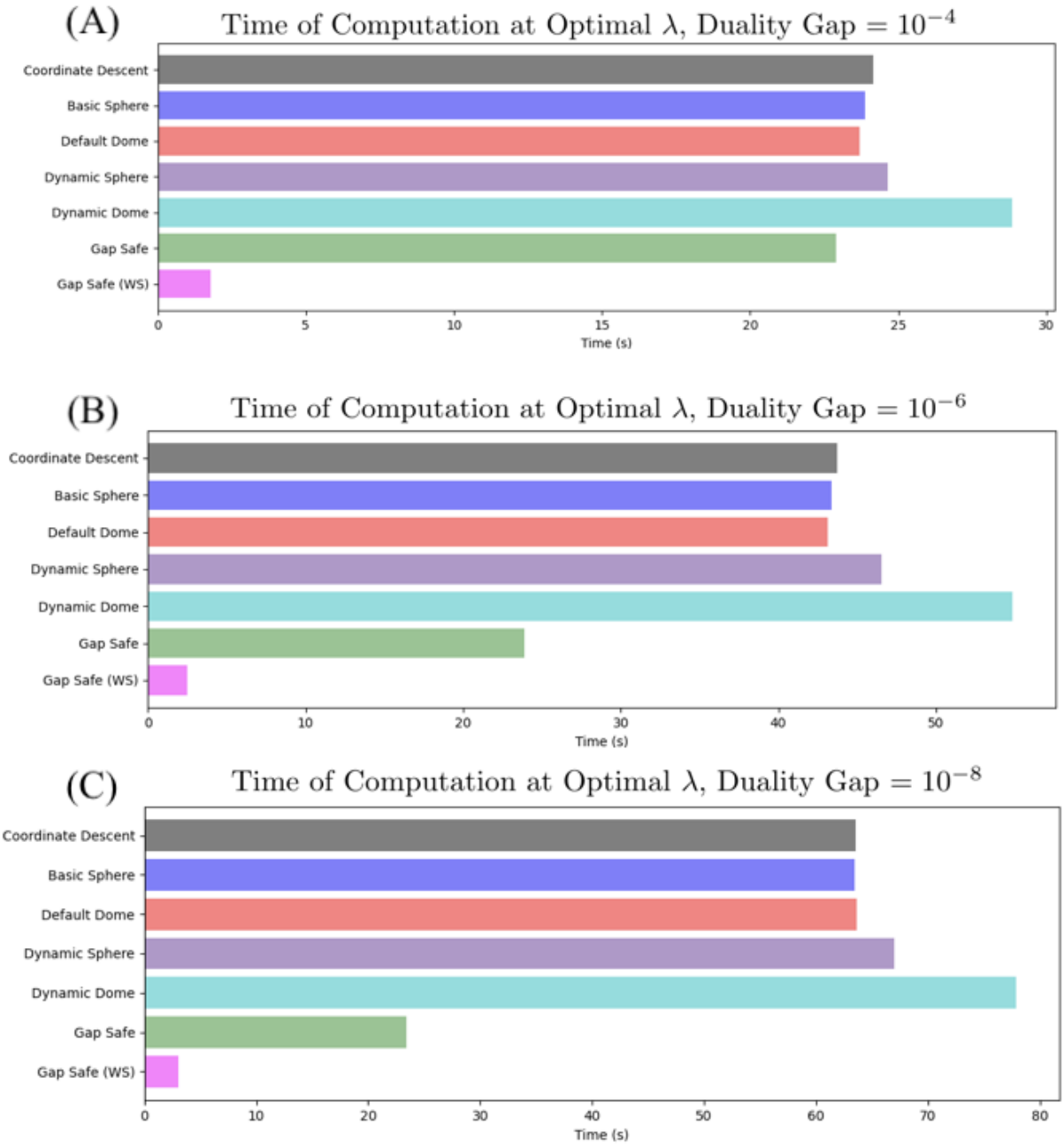

Figure 7: Time to solve the Lasso problem at $\frac{\lambda}{\lambda_{max}} = 0.032397$, which is the optimal value of the regularization parameter, as chosen through cross-validation. The duality gap is set to 10^{-4} in (A), 10^{-6} in (B), and 10^{-8} in (C).

Conclusion

In summary, we have examined a number of SAFE screening tests for the Lasso problem, which employ different constraints to bound the dual optimal point. We have tried to emphasize the relationships between these tests, and have grouped them into a framework accordingly. Numerical experiments demonstrate that the Gap SAFE test offers the greatest screening power, and reduction in computing time. Moreover, the Gap SAFE test can be further improved through the use of active set and working set strategies.

References

[1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[2] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[3] Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

[4] Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky. An interior-point method for large-scale _1-regularized least squares. *IEEE journal of selected topics in signal processing*, 1(4):606–617, 2007.

[5] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

[6] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[7] Zhen James Xiang, Yun Wang, and Peter J Ramadge. Screening tests for lasso problems. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):1008–1027, 2016.

[8] L El Ghaoui, V Viallon, and T Rabbani. Safe feature elimination in sparse supervised learning. Technical report, UC/EECS-2010-126, EECS Dept., University of California at Berkeley, 2010.

[9] Zhen James Xiang and Peter J Ramadge. Fast lasso screening tests based on correlations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2137–2140. IEEE, 2012.

[10] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

[11] Jie Wang, Jiayu Zhou, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. In *Advances in neural information processing systems*, pages 1070–1078, 2013.

[12] Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, and Remi Gribonval. Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. *IEEE Transactions on Signal Processing*, 63(19):5121–5132, 2015.

[13] Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Mind the duality gap: safer rules for the lasso. *arXiv preprint arXiv:1505.03410*, 2015.

[14] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494, 01 2001.

[15] Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. From safe screening rules to working sets for faster lasso-type solvers, 2017.