The Curse of Dimensionality: Safe Screening Rules and Geometric Adaptation

Contents

1. Introduction

1.1 Motivation

The information age has challenged statisticians to develop novel methods where classical techniques are inadequate or outright fail. While the datasets of the past typically consisted of a few carefully chosen features (alternatively, covariates or regressors) with many observations relative to the number of features, modern datasets can have hundreds or thousands of features, even exceeding the number of data points. These so-called high-dimensional situations have been a key focus of statistics and machine learning in recent times.

The Lasso, introduced by Tibshirani [1], has been one of the most popular tools in such high-dimensional settings. It is an extension of linear regression, the key idea being to regularise the ordinary least-squares estimator with an ℓ_1 penalty. The Lasso has a powerful practical benefit that makes it particularly suited to tackle high-dimensional data: it promotes sparse solutions having many components equal to exactly zero. When the number of features is large, this allows for identification of a small set of features deemed to exhibit the strongest effects on the response – that is, the Lasso inherently performs feature selection. More precisely, it can be shown [2] that there always exists a Lasso solution with at most $\min\{n, p\}$ non-zero components, where n is the number of observations and p the number of features.

Many algorithms exist for solving the Lasso problem: ISTA, FISTA [3], and co-ordinate descent [4] to name a few. However, when the number of features is large – precisely the sort of situation in which the Lasso is a preferred tool – these algorithms can be slow. This has sparked research into techniques for accelerating these solvers in high dimensions.

The primary focus of this essay is on *safe screening rules* for the Lasso problem. These methods aim to speed up Lasso solvers by exploiting the sparsity of the solutions. The key idea is to identify features whose corresponding coefficients in the Lasso solutions are guaranteed to be zero. Such features can then be safely discarded from the problem, yielding a Lasso problem with a smaller number of features and hence reducing the computational burden of solving the problem. This identification is done before the solver is initiated, in a preprocessing step, and/or during the process of solving the Lasso; the hope is that the computational effort devoted to executing these safe screening rules is small compared to the gains from solving a smaller problem. The 'safety' of these rules is in contrast with rules which may wrongfully discard features (whose corresponding coefficients in the Lasso solutions are non-zero), such as the *strong rules* proposed by Tibshirani et al. [5].

As a secondary focus, we explore extensions of the safe screening idea to a general class of learning problems with sparsity-enforcing penalties, which encompasses the Lasso, Group

Lasso and ℓ_1-regularised logistic regression.

1.2 Outline of Essay

Chapter 2 reviews safe screening for the Lasso. We set the scene in Section 2.1, reviewing the Lasso problem and providing the necessary results for safe screening. Section 2.2 explains in detail the safe screening idea. In Section 2.3 we delve into concrete examples of safe screening tests. The tests of this section are static in nature: they are designed to be applied as a preprocessing step, eliminating features only once before the Lasso solver is initiated. In contrast, the dynamic screening strategies of Section 2.4 eliminate features in tandem with solving the Lasso problem. In Section 2.5 we explore a novel strain of screening strategies, termed gap-safe rules, which exhibit many advantages over other screening tests. Finally, Section 2.6 presents a series of numerical experiments on real and synthetic data, demonstrating and comparing the performance of various screening tests.

In Chapter 3 we discuss safe screening applied to other problems with sparsity-enforcing penalties. The primary focus is on extending the gap-safe screening tests to a general class of problems, whilst maintaining their attractive properties in the Lasso setting.

Throughout the essay, we often reference the ISTA, FISTA and co-ordinate descent algorithms, using these as representative examples of Lasso solvers. For completeness, we provide a brief review of these solvers in the Appendix. Certain proofs are also deferred to the Appendix.

1.3 Notation and Basic Definitions

j, k denote arbitrary elements of $\mathbb{N}$. The ℓ_1, ℓ_2 and $\ell_\infty$$-$norms on $\mathbb{R}^k$ are denoted by $||\cdot||_1, ||\cdot||_2$ and $||\cdot||_\infty$ respectively. The standard (Euclidean) inner product on $\mathbb{R}^k$ is denoted by $\langle\cdot,\cdot\rangle$. We write $\text{Corr}(x, y) = \langle x, y\rangle/||x||_2||y||_2$ for the correlation of $x, y \in \mathbb{R}^k$. We use the notation $[k]$ for the set $\{1, \ldots, k\}$. We define the *support* of $x \in \mathbb{R}^k$ to be the set of $i \in [k]$ for which $x_i \neq 0$. For $\mathcal{A} \subseteq [k]$, $x \in \mathbb{R}^k$ and $M \in \mathbb{R}^{j\times k}$ (a real $j \times k$ matrix), $x_\mathcal{A}$ denotes the restriction of x to the indices in $\mathcal{A}$ and $M_\mathcal{A}$ denotes the sub-matrix of M assembled from the columns with indices in $\mathcal{A}$. Given $c \in \mathbb{R}^k$, $r \geqslant 0$, we denote by $\mathcal{B}(c, r)$ the *closed* $\ell_2$$-$ball centred at c with radius r. For C a subspace of $\mathbb{R}^k$, we denote by $\mathcal{P}_C$ the projection $\mathbb{R}^k \to C$. Given a real number x, we define $\text{sign}(x)$ to be 1 if x is positive, -1 if x is negative, and 0 otherwise. We call a *halfspace* any subset of $\mathbb{R}^k$ of the form $\{x \in \mathbb{R}^k : w^T x \leqslant c\}$ and a *hyperplane* any subset of the form $\{x \in \mathbb{R}^k : w^T x = c\}$, for some $w \in \mathbb{R}^k$, $c \in \mathbb{R}$.

Given an optimisation problem $\min_{x\in\mathcal{C}} f(x)$ where $\mathcal{C} \subseteq \mathbb{R}^k$, we say $\hat{x}$ is a *solution* or *minimiser* of the problem if $\hat{x} \in \mathcal{C}$ and $f(\hat{x})$ equals $\inf_{x\in\mathcal{C}} f(x)$. If the problem is instead to maximise, we define $\hat{x}$ being a solution or maximiser analogously.

Throughout the essay, X denotes an $n \times p$ *feature* or *design matrix* with rows x_i corresponding to observations, $i \in [n]$, and columns X_i corresponding to features or explanatory variables, $i \in [p]$. y denotes a signal or response vector in $\mathbb{R}^n$.

2. Safe Screening for the Lasso

2.1 The Lasso Problem

In this section we review the Lasso problem, its dual and the relations between the solutions of the two problems. Detailed proofs and derivations are reserved for the Appendix.

The Lasso seeks to estimate a vector of parameters $\beta \in \mathbb{R}^p$ by a solution of the problem

$$\min_{\beta \in \mathbb{R}^p} P_\lambda(\beta) \quad \text{for} \quad P_\lambda(\beta) = \left(\frac{1}{2} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right) \tag{1}$$

where $\lambda > 0$ is a *regularisation parameter* controlling the extent of the ℓ_1 penalisation. We write $\hat{\beta}^{(\lambda)}$ for a solution or minimiser of (1) and call such $\hat{\beta}^{(\lambda)}$ a *Lasso solution* (for the particular value λ of the regularisation parameter). Note that there always exists such a minimiser and in fact there may exist many. That said, there is almost always a unique one – a thorough treatment of this matter is provided in [2]. In general, the larger λ is, the greater the ℓ_1 penalty and the sparser the Lasso solutions.

The following Theorem is key for safe screening.

Theorem 1 (Lasso Dual Problem and Optimality Conditions)**.** The dual problem

$$\max_{\theta \in \Delta_X} D_\lambda(\theta) \quad \text{for} \quad D_\lambda(\theta) = \frac{1}{2} ||y||_2^2 - \frac{\lambda^2}{2} \left|\left| \theta - \frac{y}{\lambda} \right|\right|_2^2,$$

where $\Delta_X = \{\theta \in \mathbb{R}^n : ||X^T\theta||_\infty \leqslant 1\}$ is the *dual feasible region*, has a unique solution $\hat{\theta}^{(\lambda)}$. Moreover, we have the following relations between the primal and dual objectives:

$$P_\lambda(\beta) \geqslant D_\lambda(\theta) \text{ for all } (\beta, \theta) \in \mathbb{R}^p \times \Delta_X \qquad \textit{(weak duality)} \tag{2}$$

$$\min_{\beta \in \mathbb{R}^p} P_\lambda(\beta) = P_\lambda(\hat{\beta}^{(\lambda)}) = D_\lambda(\hat{\theta}^{(\lambda)}) = \max_{\theta \in \Delta_X} D_\lambda(\theta) \qquad \textit{(strong duality)} \tag{3}$$

Further, primal and dual solutions are related in the following manner:

$$\begin{aligned} &\text{i)} \;\; y = X\hat{\beta}^{(\lambda)} + \lambda\hat{\theta}^{(\lambda)} \\ &\text{ii)} \;\; |X_i^T\hat{\theta}^{(\lambda)}| < 1 \implies \hat{\beta}_i^{(\lambda)} = 0 \\ &\text{iii)} \;\; \hat{\beta}_i^{(\lambda)} \neq 0 \implies X_i^T\hat{\theta}^{(\lambda)} = \operatorname{sign}(\hat{\beta}_i^{(\lambda)}) \end{aligned} \tag{4}$$

These *optimality conditions* are necessary and sufficient for $\hat{\beta}^{(\lambda)}$ to be a Lasso solution.

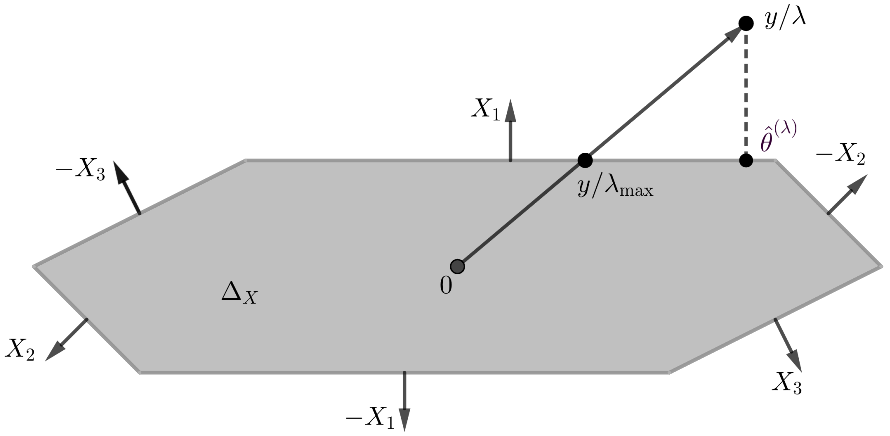

Figure 1: Illustration of the dual problem when $n = 2$, $p = 3$. Each feature corresponds to a pair of opposite faces of the polytope Δ_X, shaded in grey. $y/\lambda_{\max}$ lies on the boundary of Δ_X, where $\lambda_{\max} = ||X^T y||_\infty$. The dual solution $\hat{\theta}^{(\lambda)}$ is the projection of y/λ onto Δ_X. Here, $\lambda < \lambda_{\max}$.

Proof. Special case of Theorem 5 in Chapter 3. □

Remark 1. The dual feasible region Δ_X is an n-dimensional polytope defined by the intersection of the $2p$ *halfspaces* $\{\theta \in \mathbb{R}^n : v^T\theta \leqslant 1\}$, $v \in \{\pm X_i : i \in [p]\}$. The dual problem is equivalent to minimising $||\theta - y/\lambda||_2$ over Δ_X, hence the dual solution $\hat{\theta}^{(\lambda)}$ is the (Euclidean) projection of y/λ onto Δ_X. Figure 1 presents an illustration of the dual problem.

There is a key threshold of the regularisation parameter beyond which there is a unique, 'sparsest possible' Lasso solution: $\hat{\beta}^{(\lambda)} = 0$.

Proposition 1 (Critical Threshold: $\lambda_{\max}$)**.** $0 \in \mathbb{R}^p$ is a Lasso solution if and only if $\lambda \geqslant \lambda_{\max} := ||X^T y||_\infty$. Moreover, if this holds then $\hat{\beta}^{(\lambda)} = 0$ is the unique Lasso solution.

Proof. Special case of Theorem 6 in Chapter 3. □

$\lambda_{\max}$ is a known quantity – it is easily computed from X and y. In practice we compute $\lambda_{\max}$ and use it as an upper bound for our choice(s) of regularisation parameter λ since the trivial zero solution is of no use. Moreover, the case $\lambda_{\max} = 0$ is degenerate: it corresponds to all features being orthogonal to the response, hence the Lasso solution is 0 for all λ. In light of these facts:

Assumption. We henceforth assume $\lambda_{\max} > 0$ and $\lambda < \lambda_{\max}$.

2.2 Safe Screening Framework

The basic aim of safe screening tests is to identify, ahead of time, components of Lasso solutions that are zero. In this section we formally define such tests and discuss the best possible safe test, highlighting its impracticality. We then lay down a framework for constructing practical, albeit weaker, safe tests. Finally, we discuss two of the most common safe test architectures.

2.2.1 Basics

The following fact is simple but important enough to state as a theorem.

Theorem 2 (Discarding Features)**.** Suppose it is known that there exists a Lasso solution $\hat{\beta}^{(\lambda)}$ with support a subset of $\mathcal{A} \subseteq [p]$. Then any solution $\tilde{\beta}$ of the smaller Lasso problem

$$\min_{\beta \in \mathbb{R}^{\mathcal{A}}} \left(\frac{1}{2} ||y - X_{\mathcal{A}}\beta||_2^2 + \lambda ||\beta||_1 \right)$$

yields a solution $\hat{\beta}$ of the original Lasso problem (1) by defining $\hat{\beta}_{\mathcal{A}} = \tilde{\beta}$ and $\hat{\beta}_{[p]\backslash\mathcal{A}} = 0$. We say the features in $[p]\backslash\mathcal{A}$ are *discarded*, *screened out* or *rejected* by $\mathcal{A}$.

The smaller problem with the reduced feature matrix $X_{\mathcal{A}}$ is quicker to solve because the cost of solving the Lasso problem grows with the number of features; for ISTA, FISTA and co-ordinate descent solvers, this growth is linear in the number of features p. The idea of safe screening is to reduce computational costs in precisely this manner: discard as many features as possible and solve a smaller problem.

Definition 1 (Safe Screening Test)**.** Fixing $\lambda > 0$, a *safe screening test*, or more succinctly *safe rule* or *safe test*, is a function

$$T : [p] \to \{0, 1\}$$

such that there exists a Lasso solution $\hat{\beta}^{(\lambda)}$ with support a subset of $\mathcal{A}(T) = [p]\backslash\mathcal{S}(T)$, where

$$\mathcal{S}(T) = \{i \in [p] : T(i) = 1\}$$

is the set of *discarded*, *screened out* or *rejected* features.

Given a safe test T as above, the Lasso problem can be solved faster by taking $\mathcal{A} = \mathcal{A}(T)$ in Theorem 2, which amounts to discarding features with $T(i) = 1$.

Remark 2. We emphasise that a safe test T is related to a particular choice of λ – it may not be a safe test for other values of the regularisation parameter. The tests T we consider in this essay are defined in terms of λ, so that they are really a family of tests, one for each choice of the regularisation parameter. When discussing these tests, we do not specify a particular value of λ, with the understanding that λ is implicitly fixed and the statements we make apply for a range of λ, by default $\lambda \in (0, \lambda_{\max})$. Moreover, when we consider multiple such tests simultaneously, the regularisation parameter is the same for all of them.

2.2.2 Ultimate Safe Test

The optimality condition (4) ii) gives us our first safe test:

Corollary 1 (Ultimate Safe Test)**.** Fixing $\lambda > 0$, all Lasso solutions $\hat{\beta}^{(\lambda)}$ (by which we mean all solutions for this particular choice of λ) have support a subset of the *equicorrelation set*

$$\mathcal{E}^{(\lambda)} = \{i \in [p] : |X_i^T \hat{\theta}^{(\lambda)}| = 1\}.$$

Hence, the function

$$T(i) = \begin{cases} 1 & |X_i^T \hat{\theta}^{(\lambda)}| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

defines a safe test.

By this corollary, we may take $\mathcal{A} = \mathcal{E}^{(\lambda)}$ in Theorem 2. Note that $\mathcal{E}^{(\lambda)}$ is conservative in the sense that it only discards features that are guaranteed to be zero in *all* Lasso solutions $\hat{\beta}^{(\lambda)}$. It is in general the smallest such conservative set, as shown in [2]: fixing X and $\lambda > 0$, for Lebesgue−almost every $y \in \mathbb{R}^n$ there exists a Lasso solution with support exactly the equicorrelation set. However, it is possible for a Lasso solution to have support $\mathcal{A}$ which is a proper subset of $\mathcal{E}^{(\lambda)}$, and then $\mathcal{A}$ will discard more features than $\mathcal{E}^{(\lambda)}$; a trivial example of this is when the columns of X are identical. That said, in general the Lasso solution is unique and the equicorrelation set is precisely the support of the unique Lasso solution [2], so we do not dwell on identifying $\mathcal{A}$ smaller than $\mathcal{E}^{(\lambda)}$; all of our safe tests T will have $\mathcal{E}^{(\lambda)} \subseteq \mathcal{A}(T)$.

Corollary 1 provides us with the best possible safe rule: discard feature i if $i \notin \mathcal{E}^{(\lambda)}$. The problem is that we do not know $\hat{\theta}^{(\lambda)}$, so this test is impractical. The dual problem is a convex quadratic program, so in theory an approximation of $\hat{\theta}^{(\lambda)}$ could be obtained with standard techniques such as interior-point or active set methods. However, just as the primal problem is slow to solve in high dimensions, so too is the dual. Solving the dual, applying screening according to (5) and then solving the primal is not a viable option.

2.2.3 Safe Regions

Practical safe rules aim to relax (5) by constructing safe regions $\mathcal{R}$ containing $\hat{\theta}^{(\lambda)}$.

Definition 2 (Safe Region)**.** A subset $\mathcal{R} \subseteq \mathbb{R}^n$ is a *safe region* (for the particular value λ of the regularisation parameter) if $\hat{\theta}^{(\lambda)} \in \mathcal{R}$.

Safe regions naturally give rise to safe tests:

Proposition 2 (Safe Region Test)**.** Let $\mathcal{R}$ be a region containing $\hat{\theta}^{(\lambda)}$. Then

$$\sup_{\theta \in \mathcal{R}} |X_i^T \theta| < 1 \implies |X_i^T \hat{\theta}^{(\lambda)}| < 1 \implies \hat{\beta}_i^{(\lambda)} = 0 \text{ for all Lasso solutions } \hat{\beta}^{(\lambda)}.$$

Thus, the function

$$T_{\mathcal{R}}(i) = \begin{cases} 1 & \sup_{\theta \in \mathcal{R}} |X_i^T \theta| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

is a safe test.

Remark 3. We will often conflate a safe region $\mathcal{R}$ with its corresponding safe test $T_{\mathcal{R}}$, for example through a statement such as '$\mathcal{R}$ screens out feature $i \ldots$'.

Two extremes of safe regions are $\mathcal{R} = \{\hat{\theta}^{(\lambda)}\}$ and $\mathcal{R} = \mathbb{R}^n$. The former corresponds to the test (5), so is impractical, and the latter results in a trivial test which does not reject any features, so is useless. In constructing useful $\mathcal{R}$, in addition to practicability we look for the following two properties:

1) The *support function* $\mu_{\mathcal{R}}(v) = \sup_{\theta \in \mathcal{R}} v^T \theta$ should be quick to compute.

2) $\mathcal{R}$ should be as small as possible.

The first property ensures efficient execution of the safe rule – that is, efficient computation of $\sup_{\theta \in \mathcal{R}} |X_i^T \theta|$ and application of (6) – and requires that $\mathcal{R}$ is a simple geometric object. The ultimate goal of screening tests is to speed up Lasso solvers, so quick execution of these tests is of great importance. A test which discards a large number of features but takes longer to do so than solving the original Lasso problem is of no use.

The second property is desirable because it is in our favour to discard as many features as possible. If we have safe regions $\mathcal{R}_1 \subset \mathcal{R}_2$ then every feature discarded by $\mathcal{R}_2$ is also discarded by $\mathcal{R}_1$. With equal computational cost of executing the two tests, there would be no reason to choose $\mathcal{R}_2$ over $\mathcal{R}_1$.

Note that the support function of a region $\mathcal{R}$ is the same as the support function of the closure of its convex hull. For this reason, we assume all safe regions are closed and convex.

Remark 4. The Ultimate Safe Test (5) discards feature i if and only if $\hat{\theta}^{(\lambda)}$ is in the interior of the slab defined by $|X_i^T \theta| \leqslant 1$. The safe region relaxation (6) discards feature i if and only if the *entire region* $\mathcal{R}$ is contained in the interior of this slab. Figure 2 illustrates this.

All safe rules fit into the framework described in this section, differing only in their regions $\mathcal{R}$.

2.2.4 Typical Safe Region Architectures

An essential property of a safe region $\mathcal{R}$ is that its support function is quick to compute. As a result, almost all safe regions have one of two simple structures.

The simplest is a *sphere*:

Proposition 3 (Sphere Test Principle)**.** Suppose $\mathcal{R} = \mathcal{B}(c, r)$ is a safe region, where $c \in \mathbb{R}^n$, $r \geqslant 0$. The support function of $\mathcal{R}$ is

$$\mu_{\mathcal{R}}(v) = \sup_{\theta \in \mathcal{R}} v^T \theta = v^T c + r||v||_2.$$

Hence, $\mathcal{R}$ screens out feature i if and only if

$$|X_i^T c| + r||X_i||_2 < 1. \tag{7}$$

From (7) we see that to apply a sphere test it is required to compute $|X_i^T c|$ and $||X_i||_2$, taking $\mathcal{O}(n)$ operations. The total cost of applying the test on s features is $\mathcal{O}(ns)$, comparable to the cost of a single iteration of Lasso solvers like ISTA, FISTA and co-ordinate descent

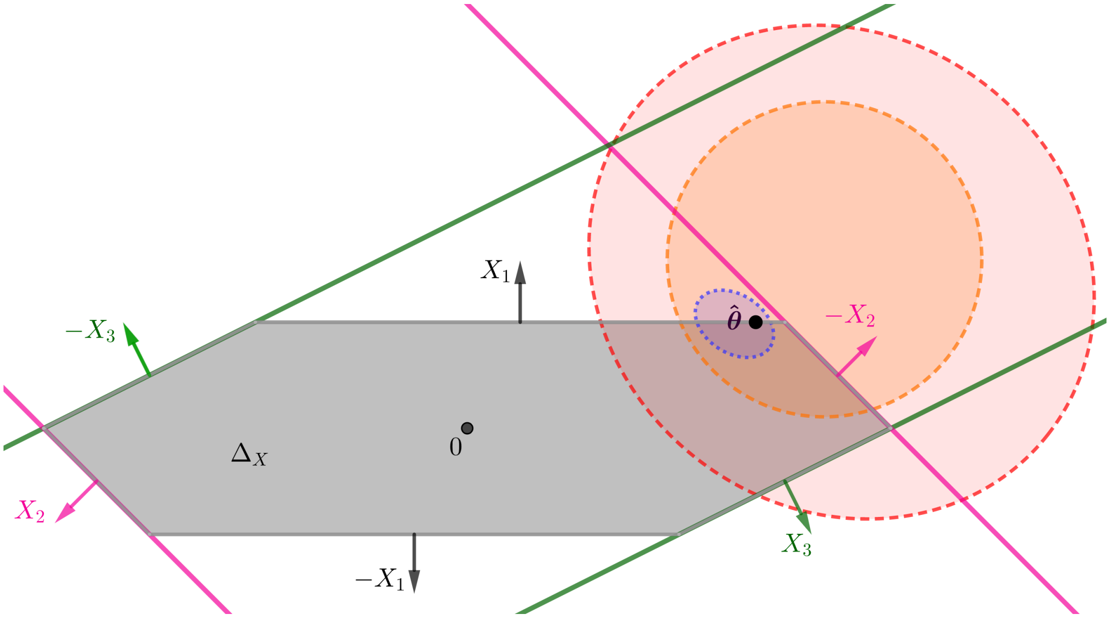

Figure 2: The dual solution $\hat{\theta}$ lies between the two pink and the two green hyperplanes (or rather, lines, in this $n = 2$ case). Given knowledge of $\hat{\theta}$, the features corresponding to these hyperplanes − features 2 and 3 − can be discarded.

Three nested safe regions are shown in blue, orange and red. These discard as follows: the blue region discards features 2 and 3, the orange region discards feature 3 but fails to discard feature 2, and the red region fails to discard any features.

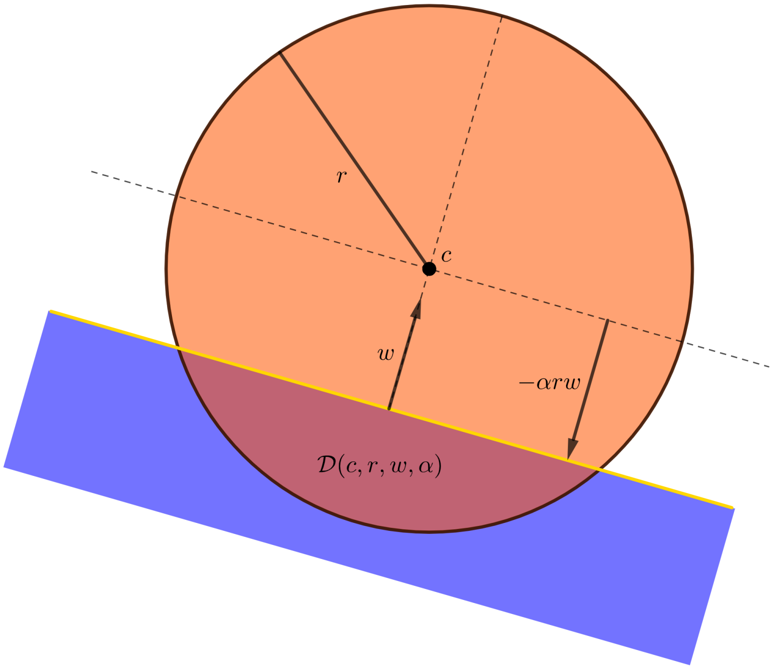

Figure 3: Illustration of the dome $\mathcal{D}(c, r, w, \alpha)$ (in dark red) for $0 < \alpha < 1$.

when there are s features. Moreover, the memory footprint of a sphere test is very small: it is required to store only c and r. Thus, sphere tests can be efficiently stored and executed.

A more complex type of region is a *dome*, the intersection of a halfspace and a sphere. We denote such a region by $\mathcal{D}(c, r, w, \alpha)$, where c is the centre of the sphere, $r \geqslant 0$ its radius, w is the unit normal vector of the halfspace pointing *out* of the halfspace, and α is such that $c - \alpha r w$ is the projection of c onto the hyperplane forming the boundary of the halfspace. For $\alpha > 1$ the dome is empty, for $\alpha = 0$ it is a hemisphere and for $\alpha \leqslant -1$ it is simply the sphere $\mathcal{B}(c, r)$. See Figure 3 for an illustration in the case $0 < \alpha < 1$. Note that, fixing c, r, and w, the volume of the dome is decreasing in α.

We allow for w to equal 0, in which case we assume the dome is simply the sphere $\mathcal{B}(c, r)$ and $\alpha = 0$.

Analogously to Proposition 3 we have:

Proposition 4 (Dome Test Principle, [12] Lemma 3)**.** Suppose $\mathcal{R} = \mathcal{D}(c, r, w, \alpha)$ is a safe region with $|\alpha| \leqslant 1$. The support function of $\mathcal{R}$ is

$$\mu_{\mathcal{R}}(v) = \sup_{\theta \in \mathcal{R}} v^T \theta = v^T c + \begin{cases} r||v||_2 & w^T v \leqslant -\alpha ||v||_2 \\ -\alpha r(w^T v) + r\sqrt{1-\alpha^2}\sqrt{||v||_2^2 - (w^T v)^2} & \text{otherwise} \end{cases}.$$

Hence, $\mathcal{R}$ screens out feature i if and only if

$$M_-(X_i) < c^T X_i < M_+(X_i) \tag{8}$$

where

$$M_\pm(v) = \pm \left(1 - \begin{cases} r||v||_2 & \pm w^T v \leqslant -\alpha ||v||_2 \\ \mp \alpha r(w^T v) + r\sqrt{1-\alpha^2}\sqrt{||v||_2^2 - (w^T v)^2} & \text{otherwise} \end{cases}\right).$$

The computational cost of applying a dome test via (8) is greater than that of a sphere test because of the need to compute an additional inner product, $w^T X_i$. That said, it is still $\mathcal{O}(n)$, and the cost of applying the test over all features is still approximately the same as one or two iterations of common iterative Lasso solvers. The memory footprint is also slightly greater than that of a sphere test but is still mild.

We point out that the computational cost of applying these tests is only half the story. Before they can be applied, they need to be constructed, i.e. their defining parameters need to be computed.

2.3 Static and Sequential Screening

The first safe screening tests were applied as a pre-processing step, screening out features only once, before the solver is initiated. We call this approach *static screening.*

A common scenario in practice, particularly when tuning the regularisation parameter, is to compute Lasso solutions over a grid of regularisation parameters $\lambda_{\max} \geqslant \lambda_1 > \cdots > \lambda_N > 0$. It is standard to use iterative solvers such as ISTA/FISTA/co-ordinate descent with a 'warm start strategy' whereby the solutions are computed in sequence, starting with

$\hat{\beta}^{(\lambda_1)}$ and using $\hat{\beta}^{(\lambda_i)}$ to initialise the computation of $\hat{\beta}^{(\lambda_{i+1})}$. In a similar vein, *sequential screening* strategies aim to leverage a previously computed Lasso solution to speed up the computation of the next.

The line between static and sequential strategies is blurred. Typically, the latter are also static, in the sense of being applied as a preprocessing step. The main distinction between the two is the assumption that a previous Lasso solution has been computed. In fact, static screening rules are usually special cases of sequential counterparts, where the known Lasso solution is taken to be the trivial solution $\hat{\beta}^{(\lambda_{\max})} = 0$.

Algorithm 1 presents the static screening strategy. A sequential strategy is structured in the same way, the only difference being that the test T is constructed with knowledge of a previous Lasso solution.

Algorithm 1: Static Screening

```
input $\lambda$, $X$, $y$, screening rule $T$
$\mathcal{A} \leftarrow$ set of $i$ for which $T(i) = 0$
$X \leftarrow X_{\mathcal{A}}$
$\hat{\beta}^{(\lambda)} \leftarrow$ solve Lasso with $\lambda$, $X$, $y$ the regularisation parameter, feature matrix and
    response respectively
```

In this section we present various static and sequential strategies in the literature. We state a few general facts on Euclidean projections and derive the tests from these. This obscures to some degree the original motivations behind these tests, though we believe this is outweighed by the benefits of a unifying framework: our presentation highlights that the tests can all be viewed as applications of basic results on projections.

Throughout, we assume knowledge (or approximate knowledge) of $\hat{\theta}^{(\lambda_0)}$ for some $\lambda_0 \in (0, \lambda_{\max}]$. This is appropriate for the sequential screening setting, since if we know $\hat{\beta}^{(\lambda_0)}$, then we also know $\hat{\theta}^{(\lambda_0)}$ through (4) i). Our goal is to locate $\hat{\theta}^{(\lambda)}$, the projection of y/λ onto Δ_X, to high precision. The tools at our disposal are $\hat{\theta}^{(\lambda_0)}$ and the properties of projections.

Proposition 5 (Existence of Projections)**.** Let $C \subseteq \mathbb{R}^n$ be a closed, non-empty set and $v \in \mathbb{R}^n$. There exists a point in C closest to v, i.e. a $\hat{c} \in C$ such that

$$||\hat{c} - v||_2 = \inf_{c \in C} ||c - v||_2. \tag{9}$$

We call $\hat{c}$ a (Euclidean) projection of v onto C. If, additionally, C is convex, then $\hat{c}$ is uniquely characterised by

$$\langle c - \hat{c}, v - \hat{c} \rangle \leqslant 0 \text{ for all } c \in C. \tag{10}$$

Noting Δ_X is closed and convex we deduce the following.

Corollary 2 (Fundamental Tools)**.** Let $v_0, v \in \mathbb{R}^n$ and $\hat{\theta}_0, \hat{\theta} \in \Delta_X$ their projections onto Δ_X. Then

$$||\hat{\theta} - v||_2 \leqslant ||\hat{\theta}_0 - v||_2, \tag{11}$$

$$\langle \hat{\theta} - \hat{\theta}_0, v_0 - \hat{\theta}_0 \rangle \leqslant 0, \tag{12}$$

$$\langle \hat{\theta}_0 - \hat{\theta}, v - \hat{\theta} \rangle \leqslant 0. \tag{13}$$

Adding (12) and (13) we obtain

$$||\hat{\theta} - \hat{\theta}_0||_2^2 \leqslant \langle \hat{\theta} - \hat{\theta}_0, v - v_0 \rangle. \tag{14}$$

Applying the Cauchy-Schwarz inequality to (14) we obtain

$$||\hat{\theta} - \hat{\theta}_0||_2 \leqslant ||v - v_0||_2. \tag{15}$$

2.3.1 Sequential Safe Regions

El Ghaoui et al. [8] coined the term 'safe' rule with their introduction of the original safe test, the SAFE sphere.

Proposition 6 (SAFE Sphere)**.** Recall that $\hat{\theta}^{(\lambda_0)}$ is known for some $\lambda_0 \in (0, \lambda_{\max}]$. The SAFE sphere

$$\mathcal{R}_{\text{SAFE}} = \mathcal{B}\left(\frac{y}{\lambda}, \left|\left|\frac{y}{\lambda} - \hat{\theta}^{(\lambda_0)}\right|\right|_2\right)$$

is a safe region, i.e. it contains $\hat{\theta}^{(\lambda)}$.

Proof. Apply (11) with $\hat{\theta}_0 = \hat{\theta}^{(\lambda_0)}$, $v = y/\lambda$, and $\hat{\theta} = \hat{\theta}^{(\lambda)}$. □

Given $\hat{\theta}^{(\lambda_0)}$, we apply the SAFE sphere by computing its centre c and radius r and checking the condition (7) for each $i \in [p]$. We then discard the features for which this condition holds. The remaining tests of this section are applied similarly: using (7) for sphere tests and (8) for dome tests.

The SAFE sphere can be improved. The following sphere, which we call 'iSAFE' or 'improved SAFE', is thanks to Liu et al. [15]. It is a subset of the SAFE sphere, hence it discards every feature discarded by SAFE.

Proposition 7 (Improved SAFE Sphere)**.** The iSAFE sphere

$$\mathcal{R}_{\text{iSAFE}} = \mathcal{B}\left(\frac{1}{2}\left(\frac{y}{\lambda} + \hat{\theta}^{(\lambda_0)}\right), \frac{1}{2}\left|\left|\frac{y}{\lambda} - \hat{\theta}^{(\lambda_0)}\right|\right|_2\right)$$

contains $\hat{\theta}^{(\lambda)}$. Moreover, it is a subset of the SAFE sphere.

Proof. That $\hat{\theta}^{(\lambda)}$ is contained in the iSAFE sphere follows from (13) with $\hat{\theta}_0 = \hat{\theta}^{(\lambda_0)}$, $v = y/\lambda$, $\hat{\theta} = \hat{\theta}^{(\lambda)}$:

$$\begin{aligned} &\langle \hat{\theta}^{(\lambda_0)} - \hat{\theta}^{(\lambda)}, y/\lambda - \hat{\theta}^{(\lambda)} \rangle \leqslant 0 \\ \iff &\left|\left|\hat{\theta}^{(\lambda)} - \frac{1}{2}\left(\frac{y}{\lambda} + \hat{\theta}^{(\lambda_0)}\right)\right|\right|_2 \leqslant \frac{1}{2}\left|\left|\frac{y}{\lambda} - \hat{\theta}^{(\lambda_0)}\right|\right|_2. \end{aligned}$$

To see that the iSAFE sphere is contained in the SAFE sphere: letting c and r be the centre and radius respectively of the SAFE sphere, the iSAFE sphere has centre a distance $r/2$ from c and radius $r/2$. □

Remark 5. This improvement of SAFE comes from using the convexity of the dual feasible region. The inequality (11) used in constructing SAFE is derived from (9), which characterises the projection onto any closed set. But under the additional assumption of convexity, the stronger condition (10) characterises the projection. This stronger condition implies (13), from which the iSAFE is constructed.

Wang et al. [14] proposed the following family of dual polytope projection (DPP) safe spheres.

Proposition 8 (DPP Sphere)**.** The DPP sphere

$$\mathcal{R}_{\text{DPP}} = \mathcal{B}\left(\hat{\theta}^{(\lambda_0)}, \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| ||y||_2\right)$$

contains $\hat{\theta}^{(\lambda)}$.

Proof. Apply (15) with $\hat{\theta} = \hat{\theta}^{(\lambda)}$, $\hat{\theta}_0 = \hat{\theta}^{(\lambda_0)}$, $v = y/\lambda$ and $v_0 = y/\lambda_0$. □

Similarly to the SAFE sphere, the DPP sphere can be improved.

Proposition 9 (Improved DPP Sphere)**.** The iDPP sphere

$$\mathcal{R}_{\text{iDPP}} = \mathcal{B}\left(\hat{\theta}^{(\lambda_0)} + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) y, \frac{1}{2}\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| ||y||_2\right)$$

contains $\hat{\theta}^{(\lambda)}$. Moreover, it is a subset of the DPP sphere.

Proof. Rearrange (14) to obtain

$$\left|\left|\hat{\theta} - \left(\hat{\theta}_0 + \frac{1}{2}(v - v_0)\right)\right|\right|_2^2 \leqslant \frac{1}{4}\,||v - v_0||_2^2\,.$$

Choosing $\hat{\theta} = \hat{\theta}^{(\lambda)}$, $\hat{\theta}_0 = \hat{\theta}^{(\lambda_0)}$, $v = y/\lambda$ and $v_0 = y/\lambda_0$, it follows that $\hat{\theta}^{(\lambda)} \in \mathcal{R}_{\text{iDPP}}$.

To see that the iDPP sphere is contained in the DPP sphere: letting c and r be the centre and radius respectively of the DPP sphere, the iDPP sphere has centre a distance $r/2$ from c and radius $r/2$. □

Remark 6. Like the improvement to SAFE, this improvement relies on the convexity of the dual feasible region. The construction of the DPP sphere relied on (15), which is a general property of projections onto closed sets (subject to choosing the projections appropriately, if they are not unique). Meanwhile, the stronger condition (14), used in the construction of the iDPP sphere, is a property of projections onto closed and *convex* sets.

El Ghaoui et al. [9] proposed the following dome test for sequential screening. It refines the SAFE sphere by intersecting it with a hyperplane.

Proposition 10 (Dome Test)**.** The dome

$$\mathcal{R}_{\text{DT}} = \mathcal{D}\left(\frac{y}{\lambda}, \left|\left|\frac{y}{\lambda} - \hat{\theta}^{(\lambda_0)}\right|\right|_2, \frac{y/\lambda_0 - \hat{\theta}^{(\lambda_0)}}{||y/\lambda_0 - \hat{\theta}^{(\lambda_0)}||_2}, \text{Corr}\left(y/\lambda - \hat{\theta}^{(\lambda_0)}, y/\lambda_0 - \hat{\theta}^{(\lambda_0)}\right)\right)$$

contains $\hat{\theta}^{(\lambda)}$. Moreover, it is a subset of the SAFE sphere.

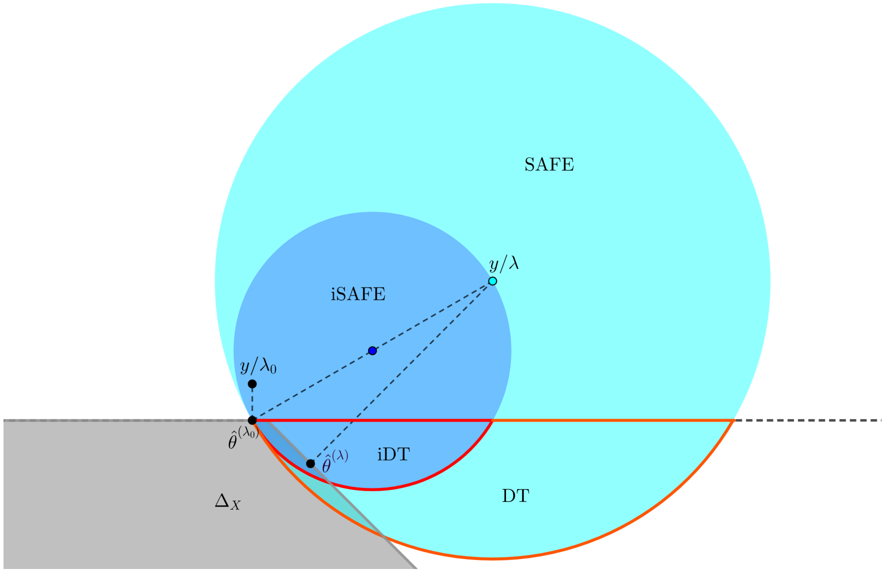

Figure 4: SAFE and iSAFE spheres, in light and dark blue respectively, and the DT and iDT domes, enclosed in orange and red. Here, $\lambda < \lambda_0 < \lambda_{\max}$.

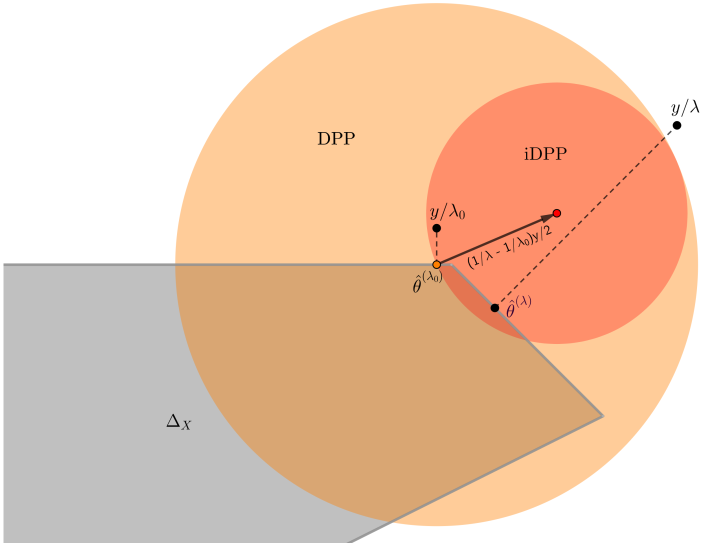
y/λ
DPP
iDPP
y/λ_0
$(1/\lambda - 1/\lambda_0)y/2$
$\hat{\theta}^{(\lambda_0)}$
$\hat{\theta}^{(\lambda)}$
Δ_X

Figure 5: DPP and iDPP spheres, in orange and red respectively, with $\lambda < \lambda_0 < \lambda_{\max}$.

Proof. This is the intersection of the SAFE sphere and the halfspace

$$\{\theta \in \mathbb{R}^n : \langle \theta - \hat{\theta}^{(\lambda_0)}, y/\lambda_0 - \hat{\theta}^{(\lambda_0)} \rangle \leqslant 0\}. \tag{16}$$

This halfspace contains $\hat{\theta}^{(\lambda)}$, by (12). □

Liu et al. [15] improved the above dome by using the iSAFE sphere in place of the SAFE sphere. The resulting dome is the best safe region we have so far in the sense of discarding the most features.

Proposition 11 (Improved Dome Test)**.** The dome

$$\mathcal{R}_{\text{iDT}} = \mathcal{D}\left(\frac{1}{2}\left(\frac{y}{\lambda} + \hat{\theta}^{(\lambda_0)}\right), \frac{1}{2}\left|\left|\frac{y}{\lambda} - \hat{\theta}^{(\lambda_0)}\right|\right|_2, \frac{y/\lambda_0 - \hat{\theta}^{(\lambda_0)}}{||y/\lambda_0 - \hat{\theta}^{(\lambda_0)}||_2}, \operatorname{Corr}\left(y/\lambda - \hat{\theta}^{(\lambda_0)}, y/\lambda_0 - \hat{\theta}^{(\lambda_0)}\right)\right)$$

contains $\hat{\theta}^{(\lambda)}$. Moreover, it is a subset of all of the regions we have presented so far.

Proof. $\mathcal{R}_{\text{iDT}}$ is the intersection of the iSAFE sphere and the halfspace (16), so is a safe region. Equivalently, $\mathcal{R}_{\text{iDT}}$ is the set of $\hat{\theta}$ satisfying (12) and (13), with $\hat{\theta}_0 = \hat{\theta}^{(\lambda_0)}$, $v = y/\lambda$, $v_0 = y/\lambda_0$. All of the regions we have presented so far are relaxations of (12) and (13) with these choices of $\hat{\theta}_0, v, v_0$. □

All of these regions perform better (are smaller and screen more features) when λ_0 is close to λ. For the spheres, this is clear by inspection of their radii. For the domes, not only do their spheres become small for λ_0 close to λ, but also $\alpha = \operatorname{Corr}(y/\lambda - \hat{\theta}^{(\lambda_0)}, y/\lambda_0 - \hat{\theta}^{(\lambda_0)})$ increases to one, which further contributes to their reduced volume.

In summary, the closer the known dual solution is to the target dual solution – or the closer the known Lasso solution is to the target Lasso solution – the more effective sequential screening is.

2.3.2 Static Safe Regions

We can regard $\lambda_{\max}$ as a default choice of λ_0, since we know the unique Lasso solution in this case: the zero vector. This choice gives rise to static screening rules: rules to be applied before initiating a Lasso solver and without knowledge of other (non-trivial) Lasso solutions.

Proposition 12 (Default Spheres)**.** The following spheres are safe regions:

$$\mathcal{R}_{\text{dSAFE}} = \mathcal{B}\left(\frac{y}{\lambda}, \left(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right)||y||_2\right) \qquad \text{(default SAFE sphere)}$$

$$\mathcal{R}_{\text{diSAFE}} = \mathcal{B}\left(\frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{\lambda_{\max}}\right)y, \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right)||y||_2\right) \qquad \text{(default improved SAFE sphere)}$$

$$\mathcal{R}_{\text{dDPP}} = \mathcal{B}\left(\frac{y}{\lambda_{\max}}, \left(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right)||y||_2\right) \qquad \text{(default DPP sphere)}$$

Moreover, these discard feature i if and only if

$$\frac{\lambda}{\lambda_{\max}} > \frac{||X_i||_2||y||_2 + |X_i^T y|}{||X_i||_2||y||_2 + \lambda_{\max}} \qquad \text{(dSAFE)} \qquad (17)$$

$$\frac{\lambda}{\lambda_{\max}} > \frac{||X_i||_2||y||_2 + |X_i^T y|}{||X_i||_2||y||_2 + 2\lambda_{\max} - |X_i^T y|} \qquad \text{(diSAFE)} \qquad (18)$$

$$\frac{\lambda}{\lambda_{\max}} > \frac{||X_i||_2||y||_2}{||X_i||_2||y||_2 + \lambda_{\max} - |X_i^T y|} \qquad \text{(dDPP)} \qquad (19)$$

Proof. That they are safe regions follows from Propositions 6, 7 and 8 with $\lambda_0 = \lambda_{\max}$. For the inequalities, apply Proposition 3 (Sphere Test Principle) with c, r the appropriate centres and radii. □

Remark 7. We did not provide a default improved DPP sphere because the iDPP and iSAFE spheres are identical when $\lambda_0 = \lambda_{\max}$. The diSAFE sphere screens the most features among these static rules, being contained in both the dSAFE and dDPP spheres.

The scenarios under which these spheres are most and least effective are apparent from (17), (18), and (19). A simple observation is that these tests screen most aggressively when λ is large, and in fact in the limit $\lambda \uparrow \lambda_{\max}$ the only features not screened out are those with $|X_i^T y| = \lambda_{\max}$. This is not surprising, since the larger λ is, the smaller the radii of the spheres. Keeping $||X_i||_2$ and $||y||_2$ fixed, the right-hand-sides of the inequalities are increasing in $|\text{Corr}(X_i, y)|$, indicating that features with low absolute correlation with the response are the easiest to discard.

In summary, these tests are most useful when particularly sparse solutions are sought (corresponding to large λ) or when a substantial number of the features have low correlation with the response. They are least useful when λ is small or when the features are highly correlated with the response.

In the original paper of El Ghaoui et al., the applications of interest involved seeking extremely sparse solutions, hence the dSAFE sphere proved very effective. Not only did it substantially accelerate Lasso solvers, up to an order of magnitude, it also extended the reach of the Lasso to extremely large problems whose feature matrices were too large to even be loaded into memory, by removing sufficiently many features for them to be loaded.

When λ_0 assumes the default value $\lambda_{\max}$, the dome tests of Propositions 10 and 11 are the intersections of spheres with trivial halfspaces, since $\hat{\theta}^{(\lambda_{\max})} = y/\lambda_{\max}$. Hence, they are simply the dSAFE and diSAFE spheres respectively. However, there is a natural candidate for a non-trivial halfspace, proposed by Xiang et al. [11]: the halfspace corresponding to the face of Δ_X on which $y/\lambda_{\max}$ lies.

Proposition 13 (Default Dome Tests)**.** Choosing $X^* \in \arg\max_{v \in \{\pm X_i : i \in [p]\}} v^T y$, we define the default dome

$$\begin{aligned}\mathcal{R}_{\text{dDT}} &= \{\theta : X^{*T}\theta \leqslant 1\} \cap \mathcal{R}_{\text{dSAFE}} \\ &= \mathcal{D}\left(\frac{y}{\lambda}, \left(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right)||y||_2, \frac{X^*}{||X^*||_2}, \text{Corr}(X^*, y)\right)\end{aligned}$$

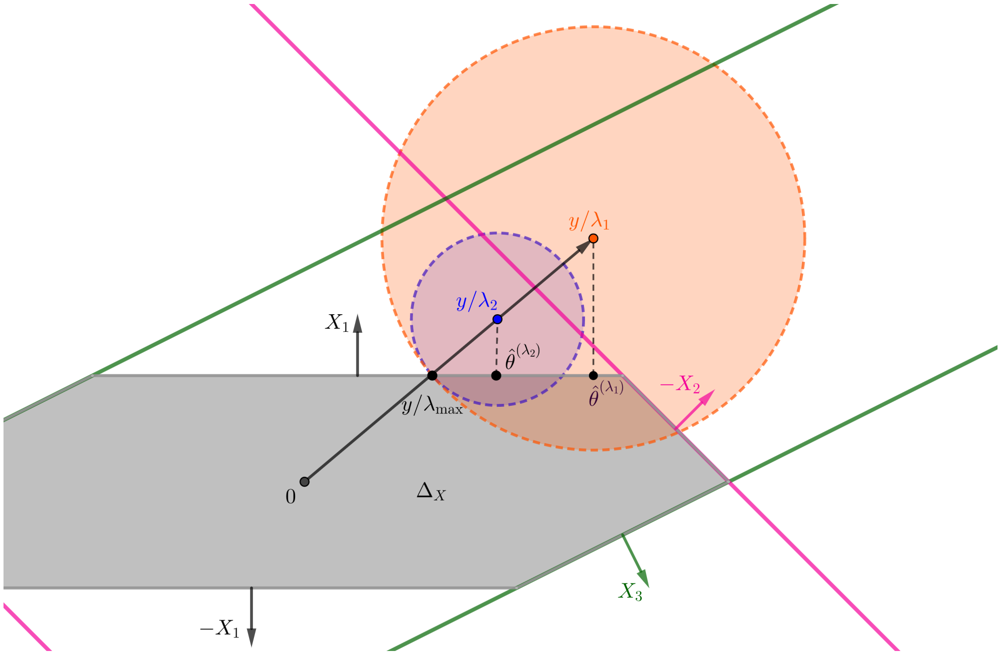

Figure 6: Illustration of the dSAFE sphere. Here, $\lambda_1 < \lambda_2 < \lambda_{\max}$. The orange and blue regions are the dSAFE spheres corresponding to $\lambda = \lambda_1$ and $\lambda = \lambda_2$ respectively.
For both values of λ, given $\hat{\theta}^{(\lambda)}$ we can discard features 2 and 3. The test is most effective for large λ: the blue sphere is smaller and discards feature 3 (though fails to discard feature 2), whereas the orange sphere fails to discard either feature.

and the default improved dome

$$\begin{aligned}\mathcal{R}_{\text{diDT}} &= \{\theta : {X^*}^T\theta \leqslant 1\} \cap \mathcal{R}_{\text{diSAFE}} \\ &= \mathcal{D}\left(\frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{\lambda_{\max}}\right)y,\ \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}}\right)||y||_2,\ \frac{X^*}{||X^*||_2},\ \text{Corr}(X^*, y)\right).\end{aligned}$$

These are safe regions, i.e. they contain $\hat{\theta}^{(\lambda)}$.

Proof. We already know that the spheres are safe regions. The halfspace is also a safe region because the entirety of Δ_X is contained in it. Hence their intersections are safe regions. □

The dome $\mathcal{R}_{\text{diDT}}$ is our best static screening region, being a subset of all of the other regions. Both domes have small volume when $\text{Corr}(X^*, y)$ is large. Indeed, in this case they can prove to be very effective static screening rules.

2.3.3 Convergence to Zero Volume

The radii of the SAFE spheres can only be so small, minimised to a positive value when $\lambda_0 = \lambda$. As a result, even when these spheres are constructed with perfect knowledge of the dual solution, they are unable to identify it. In contrast, as $\lambda_0 \to \lambda$ the DPP spheres and dome tests converge to zero in volume. For λ_0 sufficiently close λ, they are equivalent to the Ultimate Safe Test.

The reason for the differing limiting behaviour is precisely that the SAFE spheres outright ignore that $\hat{\theta}^{(\lambda_0)}$ is the projection of y/λ_0 onto Δ_X, relying only on the knowledge that $\hat{\theta}^{(\lambda_0)}$ is an element of Δ_X. In particular, these spheres do not leverage the fact that $||y/\lambda - y/\lambda_0||_2$ can be used to bound $||\hat{\theta}^{(\lambda)} - \hat{\theta}^{(\lambda_0)}||_2$.

2.3.4 Computational Costs

We saw in Section 2.2.4 that both sphere and dome tests can be efficiently applied, but that the cost of constructing the tests is also important in analysing the computational effort demanded by them. We briefly address these for the tests presented.

In the static screening setting, where $\lambda_0 = \lambda_{\max}$, the cost of construction is dominated by the computation of $\lambda_{\max}$, taking $\mathcal{O}(np)$ operations.

In the sequential screening setting, after finishing the computation of $\hat{\beta}^{(\lambda_0)}$, it takes $\mathcal{O}(np)$ operations to compute $\hat{\theta}^{(\lambda_0)}$ via (4) i), or $\mathcal{O}(ns)$ if one leverages the sparsity of $\hat{\beta}^{(\lambda_0)}$, where s is the size of the support of $\hat{\beta}^{(\lambda_0)}$. Once $\hat{\theta}^{(\lambda_0)}$ is known, all of the tests take only a few additions or inner products over $\mathbb{R}^n$ to compute their defining parameters, such as the centres of their spheres. Thus, they can all be constructed in $\mathcal{O}(ns)$ time.

In summary, the total cost of construction plus execution of all of the tests presented is at most that of two or three iterations of the underlying solver. If a substantial number of features are screened out, the performance gains are immense. Even if a small number of features are screened out, the gain in reducing the problem size can outweigh the total cost of screening.

2.3.5 Problem of Inexact Knowledge

As pointed out by Fercoq et al. [19], we generally have only an approximation of $\hat{\theta}^{(\lambda_0)}$, unless $\lambda_0 = \lambda_{\max}$. This is certainly the case in the sequential screening setting when iterative solvers are used: we only obtain approximations of $\hat{\beta}^{(\lambda_i)}$, from which we obtain approximations of $\hat{\theta}^{(\lambda_i)}$ through (4) i). Neglecting this issue can produce unsafe rules which erroneously discard features.

Suppose we know $\hat{\theta}^{(\lambda_0)}$ to $\epsilon-$accuracy through an approximation $\tilde{\theta}$, i.e. $||\hat{\theta}^{(\lambda_0)} - \tilde{\theta}||_2 \leqslant \epsilon$. The sphere tests can be easily adjusted to reflect this by adding ϵ to their radii. Adjusting the dome tests is more difficult since one has to take account of all of the possible halfspaces that (16) could be, but it can also be done.

However, it is not so easy to obtain an estimate of $||\hat{\theta}^{(\lambda_0)} - \tilde{\theta}||_2$. Firstly, since we estimate $\hat{\theta}^{(\lambda_0)}$ using an estimate of $\hat{\beta}^{(\lambda_0)}$ and equation (4) i), we can only hope to estimate the accuracy of $\tilde{\theta}$ if we have an estimate of the accuracy of our supposed Lasso solution, thus requiring that our Lasso solver returns such estimates. Secondly, even if we have such an estimate, bounding $||\hat{\theta}^{(\lambda_0)} - \tilde{\theta}||_2$ may not be so straightforward: we require an estimate of the operator norm of X.

The weakness of the SAFE spheres discussed in Section 2.3.3 works to their benefit when it comes to inexact knowledge of $\hat{\theta}^{(\lambda_0)}$. Because these spheres use only the fact that $\hat{\theta}^{(\lambda_0)}$ is a dual feasible point and not that it is the projection of y/λ_0 onto $\hat{\theta}^{(\lambda_0)}$, we may replace $\hat{\theta}^{(\lambda_0)}$

in their definitions by any dual feasible point θ to obtain valid safe regions. Then, given an approximation $\tilde{\theta}$ of $\hat{\theta}^{(\lambda_0)}$, there is a simple and efficient way to obtain a dual feasible point θ from $\tilde{\theta}$, which will be close to $\hat{\theta}^{(\lambda_0)}$ if the original approximation was close. Though we may not be able to obtain concrete practical bounds on how close θ is to $\hat{\theta}^{(\lambda_0)}$, the SAFE spheres constructed from θ are, at the very least, guaranteed to be safe, so we can happily use them knowing that no features will be incorrectly discarded. These ideas were developed in the dynamic screening setting and are presented in Section 2.4 through the Dynamic SAFE spheres. We also provide a simple but novel extension to the DPP spheres and the dome tests, allowing them to be used in the absence of precise knowledge of how accurately $\hat{\theta}^{(\lambda_0)}$ is known.

2.4 Dynamic Screening

Bonnefoy et al. [17, 18] proposed a dynamic screening strategy where the screening tests are interlaced with an iterative solver and *evolve* along the solving process, rather than being applied only before its initiation, as in static screening. This strategy can take advantage of the fact that the optimisation algorithm converges to a Lasso solution by using improving approximations of the solution to construct safe regions of increasing precision.

Algorithm 2 describes the dynamic screening strategy. A static screening strategy would omit lines 8 and 9, where the latest iterate is used to (safely) screen the feature matrix.

Algorithm 2: Dynamic Screening

```
 1: $k \leftarrow 0$
 2: $\beta_0$ initial value
 3: $X \leftarrow$ screen X using a safe test
 4: $\beta_0 \leftarrow$ remove components corresponding to screened features
 5: repeat
 6:     $k \leftarrow k + 1$
 7:     $\beta_k \leftarrow$ update $\beta_{k-1}$
 8:     $X \leftarrow$ screen $X$ using a safe test constructed from $\beta_k$
 9:     $\beta_k \leftarrow$ remove components corresponding to screened features
10: until stopping criterion triggered
11: return $\beta_k$
```

Crucially, screening in this fashion maintains the convergence of the underlying Lasso solver.

Theorem 3 (Convergence Under Dynamic Screening)**.** Consider an iterative Lasso solver which, from any initial point β_0, produces iterates (β_k) converging to a Lasso solution. If dynamic (safe) screening is used in conjunction with the solver, as in Algorithm 2, the iterates still converge to a Lasso solution.

Proof. That the screening tests are safe guarantees that at any given iteration there exists a Lasso solution with support a subset of the set of remaining features. Since there are only so many features that can be discarded, no more features are screened out beyond some Kth iteration and the solver proceeds as normal, converging to a Lasso solution. □

The performance gain from dynamic screening comes from the continual reduction of the dimension of the feature matrix and of the iterates (β_k), in lines 8 and 9. This makes the update steps cheaper as the algorithm proceeds. For dynamic screening to be worthwhile, these gains have to outweigh the time spent constructing and applying the screening tests. Note that one should also track the set of removed features, which we have omitted in Algorithm 2.

We emphasise that the underlying Lasso solver in a dynamic screening strategy can be any convergent, iterative Lasso solver. That said, co-ordinate descent is particularly simple to adapt for the dynamic framework. Since the algorithm inherently acts co-ordinate by co-ordinate, the screening of X and the dimension reduction of β become straightforward to implement: one can simply maintain a list of active features, removing features from it when applying screening, and updating co-ordinates only from this list in line 7.

2.4.1 Key Ideas

The first key idea behind dynamic screening is to construct safe regions from arbitrary elements of $\mathbb{R}^n$. More precisely:

Definition 3 (Dynamic Safe Screening Test)**.** Fixing $\lambda > 0$, a *dynamic safe screening test*, or more succinctly *dynamic safe rule* or *dynamic safe test*, is a function $\mathcal{M}$ on $\mathbb{R}^n$ such that, given $z \in \mathbb{R}^n$, $T = \mathcal{M}(z)$ is a safe test (in the sense of Definition 1).

The second key idea is a scheme for generating a sequence converging to the dual solution. This allows for locating $\hat{\theta}^{(\lambda)}$ to greater and greater precision as the solver proceeds.

Corollary 3 (Convergence to Dual Solution)**.** Suppose (β_k) converges to the Lasso solution $\hat{\beta}^{(\lambda)}$. Then the sequence (R_k/λ), where $R_k = y - X\beta_k$ is the kth residual, converges to the dual solution $\hat{\theta}^{(\lambda)}$.

Proof. Follows from (4) ii). □

We combine these two ideas in the dynamic screening strategy of Algorithm 3, choosing $z = R_k/\lambda$ within each iteration. Over the course of the solving process, the dynamic safe test $\mathcal{M}$ is supplied with a sequence converging to $\hat{\theta}^{(\lambda)}$. A good choice of $\mathcal{M}$ should be able to take advantage of this to construct better and better safe tests.

In the next section we discuss various proposals for $\mathcal{M}$.

2.4.2 Dynamic Safe Regions

Our task is to construct mappings $\mathcal{M}$ from $\mathbb{R}^n$ to the set of safe tests.

Suppose $z \in \mathbb{R}^n$ is given. Rather than trying to construct a safe test directly from z, we first generate a dual feasible point $\theta \in \Delta_X$ from z. We then focus on constructing a safe region from θ. There is a natural and efficient way to obtain such a dual feasible point: simply rescale z so that it is dual feasible.

Algorithm 3: Dynamic Screening 2

```
input $\lambda$, $X$, $y$, $\mathcal{M}$, $\beta_0$
$k \leftarrow 0$
$R_0 \leftarrow y - X\beta_0$; $z \leftarrow R_0/\lambda$; $T \leftarrow \mathcal{M}(z)$
$X \leftarrow$ screen $X$ using $T$
$\beta_0 \leftarrow$ remove components corresponding to screened features
repeat
    $k \leftarrow k + 1$
    $\beta_k \leftarrow$ update $\beta_{k-1}$
    $R_k \leftarrow y - X\beta_k$; $z \leftarrow R_k/\lambda$; $T \leftarrow \mathcal{M}(z)$
    $X \leftarrow$ screen $X$ using $T$
    $\beta_k \leftarrow$ remove components corresponding to screened features
until stopping criterion triggered
return $\beta_k$
```

Lemma 1 (Dual Scaling)**.** Given $z \in \mathbb{R}^n$,

$$\mu^* := \operatorname*{arg\,min}_{\mu \in \mathbb{R}:\ \mu z \in \Delta_X} \left\| \mu z - \frac{y}{\lambda} \right\|_2 = \left[\frac{z^T y}{\lambda ||z||_2^2} \right]_{-||X^T z||_\infty^{-1}}^{||X^T z||_\infty^{-1}}$$

where

$$[a]_b^c := \min\big(\max(a, b), c\big).$$

We refer to $\theta = \mu^* z$ as the dual scaled version of z.

Remark 8. Whilst rescaling z to achieve dual feasibility, we also minimise the distance to y/λ. This serves as a proxy for minimising the distance to the (unknown) dual solution $\hat{\theta}^{(\lambda)}$.

Using this idea, Bonnefoy et al. proposed the first dynamic safe rules:

Proposition 14 (Dynamic SAFE/Dome [17, 18])**.** Let X^* be as in Proposition 13. Given $z \in \mathbb{R}^n$, let θ be its dual scaled version. The D-SAFE sphere

$$\mathcal{R}_{\text{D-SAFE}} = \mathcal{B}\left(\frac{y}{\lambda}, \left\|\frac{y}{\lambda} - \theta\right\|_2\right)$$

and the dome

$$\begin{aligned} \mathcal{R}_{\text{D-dDT}} &= \{\varphi : {X^*}^T \varphi \leqslant 1\} \cap \mathcal{R}_{\text{D-SAFE}} \\ &= \mathcal{D}\left(\frac{y}{\lambda}, \left\|\frac{y}{\lambda} - \theta\right\|_2, \frac{X^*}{||X^*||_2}, \frac{\langle X^*, y/\lambda\rangle - 1}{||X^*||_2 ||y/\lambda - \theta||_2}\right) \end{aligned}$$

are safe regions, i.e. they contain $\hat{\theta}^{(\lambda)}$. Hence, the functions $\mathcal{M}_{\text{D-SAFE}}$ and $\mathcal{M}_{\text{D-dDT}}$, defined by $z \mapsto T_{\mathcal{R}_{\text{D-SAFE}}}$ and $z \mapsto T_{\mathcal{R}_{\text{D-dDT}}}$ respectively, are dynamic safe tests.

Proof. In proving the SAFE sphere is a safe region we used only that $\hat{\theta}^{(\lambda_0)}$ is dual feasible, so replacing $\hat{\theta}^{(\lambda_0)}$ with θ still maintains that this is a safe region. The hyperplane $\{\varphi : {X^*}^T\varphi \leqslant 1\}$ contains Δ_X so in particular contains $\hat{\theta}^{(\lambda)}$. □

Remark 9. From here on we do not bother to specify the function $\mathcal{M}$ and speak of *dynamic safe regions* instead.

We can improve the regions of the preceding proposition by adapting the iSAFE sphere in the same way.

Proposition 15 (Improved Dynamic SAFE/Dome)**.** Let z, θ be as in Proposition 14. The D-iSAFE sphere

$$\mathcal{R}_{\text{D-iSAFE}} = \mathcal{B}\left(\frac{1}{2}\left(\frac{y}{\lambda} + \theta\right), \frac{1}{2}\left\|\frac{y}{\lambda} - \theta\right\|_2\right)$$

and the dome

$$\begin{aligned}\mathcal{R}_{\text{D-idDT}} &= \{\varphi : {X^*}^T\varphi \leqslant 1\} \cap \mathcal{R}_{\text{D-iSAFE}} \\ &= \mathcal{D}\left(\frac{1}{2}\left(\frac{y}{\lambda} + \theta\right), \frac{1}{2}\left\|\frac{y}{\lambda} - \theta\right\|_2, \frac{X^*}{\|X^*\|_2}, \frac{\langle X^*, y/\lambda + \theta\rangle - 2}{\|X^*\|_2\|y/\lambda - \theta\|_2}\right)\end{aligned}$$

are safe regions.

In addition to these regions, we propose novel adaptations of the full suite of tests in Section 2.3 for use in the dynamic setting. The SAFE spheres have been straightforward to generalise because they do not make use of the fact that $\hat{\theta}^{(\lambda_0)}$ is the projection of y/λ_0 onto Δ_X, only that it is dual feasible. This allows to construct SAFE spheres from any dual feasible point, and, using dual scaling, from any element of $\mathbb{R}^n$. For the other tests, given $z \in \mathbb{R}^n$ we need to be able to generate, in addition to the dual scaled θ, a point v whose projection onto Δ_X is θ. Then, while θ replaces $\hat{\theta}^{(\lambda_0)}$, v can replace y/λ_0.

A natural way to generate possible candidates for v is to start at θ and propagate along the normal vector to the face of Δ_X on which θ lies. We minimise the distance to y/λ to get the smallest possible bound on the distance between θ and $\hat{\theta}^{(\lambda)}$.

Lemma 2 (Dual Propagation)**.** Let $z \in \mathbb{R}^n$ and θ its dual scaled version. If $y^Tz \neq 0$, then, choosing $w \in \{X_i : i \in [p]\}$ such that $|w^T\theta| = 1$,

$$\operatorname*{arg\,min}_{v\in\{\theta\}+\text{span}(w)} \left\|v - \frac{y}{\lambda}\right\|_2 = \theta + \frac{w^T(y/\lambda - \theta)}{\|w\|_2^2}w.$$

Moreover, the projection of v onto Δ_X is θ. If instead $y^Tz = 0$ then $\theta = 0$ and we set $v = 0$. We refer to v as the dual propagated version of z.

Proof. The statement for the case $y^Tz = 0$ follows straightforwardly from Lemma 1. By assumption $\lambda < \lambda_{\text{max}}$, so y/λ lies outside of Δ_X. Then, assuming $y^Tz \neq 0$, θ lies on the boundary of Δ_X and we can indeed choose w as in the statement of the Lemma (though there may be multiple possible choices). Moreover, given any $v \in \{\theta\} + \text{span}(w)$, the projection of v onto $\mathcal{R} = \{\varphi : |w^T\varphi| \leqslant 1\}$ is θ. Since $\mathcal{R}$ contains Δ_X, θ is also the projection of v onto Δ_X. □

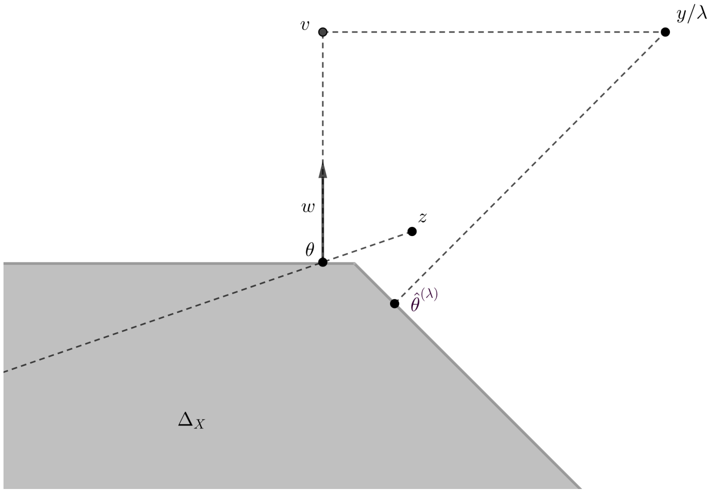

Figure 7: Dual scaling and dual propagation of an arbitrary $z \in \mathbb{R}^n$. The dual scaled point θ is the dual feasible point on the line $\operatorname{span}(z)$ closest to y/λ. w is the normal to the face of Δ_X on which θ lies. The dual propagated point v is the point on the line $\theta + \operatorname{span}(w)$ closest to y/λ.

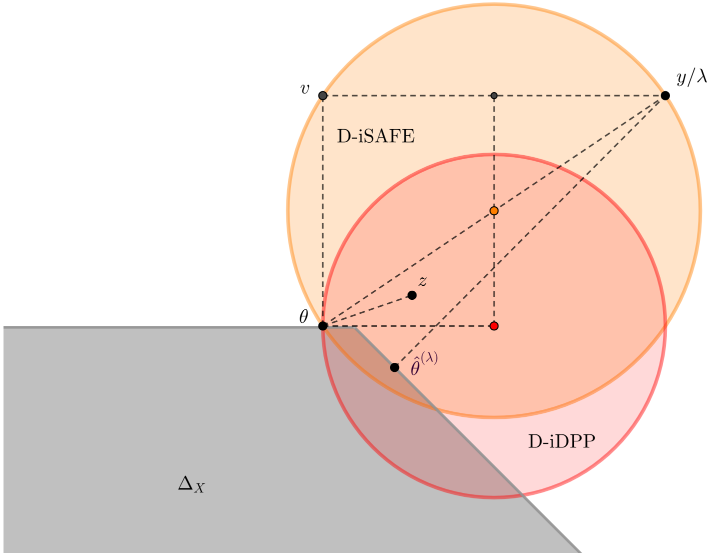

Figure 8: The dynamic improved SAFE and dynamic improved DPP spheres, in orange and red respectively, constructed from a point $z \in \mathbb{R}^n$ using dual scaling and dual propagation. θ and v are the dual scaled and dual propagated versions, respectively, of z.

Proposition 16 (Dynamic DPP Spheres)**.** Let $z \in \mathbb{R}^n$, θ its dual scaled version, and v its dual propagated version. The D-DPP sphere

$$\mathcal{R}_{\text{D-DPP}} = \mathcal{B}\left(\theta,\ \left\|\frac{y}{\lambda} - v\right\|_2\right)$$

and the D-iDPP sphere

$$\mathcal{R}_{\text{D-iDPP}} = \mathcal{B}\left(\theta + \frac{1}{2}\left(\frac{y}{\lambda} - v\right),\ \frac{1}{2}\left\|\frac{y}{\lambda} - v\right\|_2\right)$$

are safe regions.

Proposition 17 (Dynamic Dome Tests)**.** Let $z \in \mathbb{R}^n$, θ its dual scaled version, and v its dual propagated version. The domes

$$\mathcal{R}_{\text{D-DT}} = \mathcal{D}\left(\frac{y}{\lambda},\ \left\|\frac{y}{\lambda} - \theta\right\|_2,\ \frac{v-\theta}{||v-\theta||_2},\ \text{Corr}\left(y/\lambda - \theta,\ v - \theta\right)\right)$$

and

$$\mathcal{R}_{\text{D-iDT}} = \mathcal{D}\left(\frac{1}{2}\left(\frac{y}{\lambda} + \theta\right),\ \frac{1}{2}\left\|\frac{y}{\lambda} - \theta\right\|_2,\ \frac{v-\theta}{||v-\theta||_2},\ \text{Corr}\left(y/\lambda - \theta,\ v - \theta\right)\right)$$

are safe regions.

2.4.3 Computational Costs

The cost of dynamic screening is $\mathcal{O}(ns)$ per iteration when s features remain; the computational burden is chiefly in dual scaling the residuals R_k and in applying the latest screening test T. This cost is comparable to the cost of one iteration of the solver itself. The improvement of the tests from one iteration to the next is not enough to justify so large an additional cost in every single iteration. In practice, we recommend applying Algorithm 3 with screening not at every single iteration, but at every K iterations for some K.

2.4.4 Super-Efficient Dynamic Screening

Bonnefoy et al. considered dynamic screening specifically for first-order Lasso solvers – solvers which use the gradient of the data fidelity term, $X^T R_k$, such as ISTA/FISTA – in conjunction with the tests of Proposition 14. With this particular setup, dynamic screening can be implemented very efficiently.

Since $X^T R_k$ is naturally computed by a first-order solver, dual scaling of R_k takes only $\mathcal{O}(s)$ operations when s features remain. The other major cost of dynamic screening – the application of the tests – is also reduced: applying these tests requires computation of $X_i^T y$ and $||X_i||_2$, and possibly also $X_i^T X^*$ and $||X^*||_2$, which can all be computed once at the start of the algorithm and stored.

This dynamic strategy has just $\mathcal{O}(s+n)$ cost per iteration, which Bonnefoy et al. deemed cheap enough to incorporate screening into every iteration.

Both the use of a first-order algorithm and the specific choice of tests are key for this efficient implementation. The former is necessary for cheap dual scaling, and the latter for uniforming the computation involved in applying the tests.

The drawback is that these tests can quickly reach their peak performance beyond which they do not screen any more features and become useless computation. Moreover, their limiting regions (taking the iteration number to infinity) are not as small as those of other dynamic regions, hence they screen fewer features.

Consider, for example, the D-SAFE sphere. It can be shown that if

$$\frac{\lambda}{\lambda_{\max}} \leqslant \frac{||X_i||_2||y||_2 + |X_i^T y|}{||X_i||_2\lambda_{\max}||\hat{\theta}^{(\lambda)}||_2 + \lambda_{\max}} \tag{20}$$

then dynamic screening with D-SAFE is unable to reject feature i. Meanwhile, (17) characterises whether or not the default SAFE sphere discards feature i. The dSAFE sphere performs best when λ is large and close to $\lambda_{\max}$, but under this scenario the right-hand-sides of (20) and (17) are close, suggesting that the dynamic strategy does not screen many more features than the static test using dSAFE. Dynamic screening becomes more effective *relative* to static screening as $\lambda \downarrow 0$, though both strategies screen fewer and fewer features in this limit, screening no features at all for λ sufficiently small.

To summarise, though the two tests of Proposition 14 can be every efficiently implemented, they screen fewer features than other dynamic tests and they may not even benefit much from the dynamic framework.

2.4.5 Convergence to Zero Volume

It is desirable for the evolving safe regions to converge to zero in volume as the solver progresses. This would mean that the equicorrelation set would be identified in finite time. None of the dynamic regions we have presented have this property. Our smallest region is the dynamic improved dome $\mathcal{R}_{\text{D-iDT}}$, which grows to infinity in volume as $\lambda \downarrow 0$ if y is not perfectly correlated (or anti-correlated) with any of the features. However, if λ is such that $\hat{\theta}^{(\lambda)}$ lies in the interior of one of the faces of Δ_X, then the dynamic domes of Proposition 17 do converge to zero in volume.

2.4.6 Application to Sequential Screening

The dynamic regions proposed can be used to tackle the problem of inexact knowledge of $\hat{\theta}^{(\lambda_0)}$ in the sequential screening setting. Given an approximation $\tilde{\theta}$ of $\hat{\theta}^{(\lambda_0)}$, we simply choose $z = \tilde{\theta}$ in constructing the dynamic regions. These regions can then be used for sequential screening.

When $\tilde{\theta}$ is exactly $\hat{\theta}^{(\lambda_0)}$, the dynamic SAFE spheres are identical to their sequential counterparts. This is not the case for the regions of Propositions 16 and 17. This is because the dual propagated version v of $\hat{\theta}^{(\lambda_0)}$ is, in general, not equal to y/λ_0 and in fact may not even be such that $v - \hat{\theta}^{(\lambda_0)}$ is parallel to $y/\lambda_0 - \hat{\theta}^{(\lambda_0)}$; see Figure 10 for an illustration. Therefore, the dynamic regions can differ from their sequential counterparts even with perfect knowledge of $\hat{\theta}^{(\lambda_0)}$. This can result in them being larger *or* smaller than their sequential counterparts.

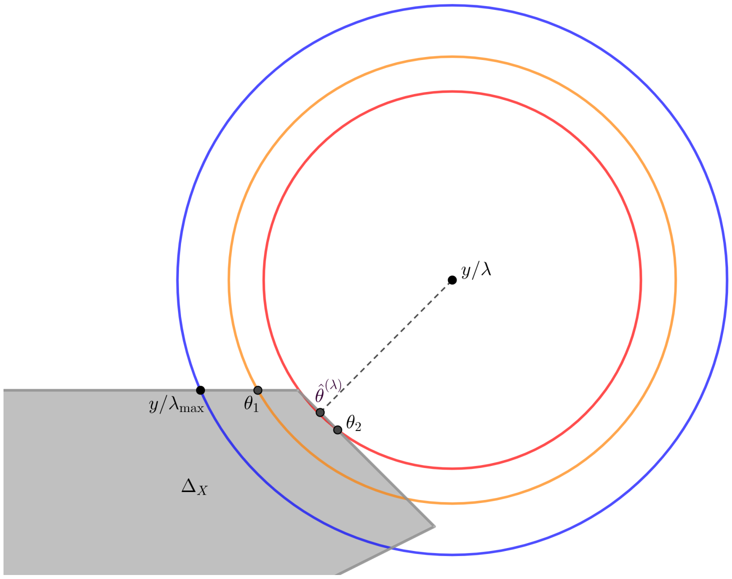

Figure 9: Dynamic SAFE spheres. The blue, orange and red spheres are the first (default), second and third safe tests over the course of a dynamic screening strategy. θ_i are the dual scaled versions of the ith residuals $R_i = y - X\beta_i$. The spheres are centred at y/λ; their radii decrease as the algorithm progresses.

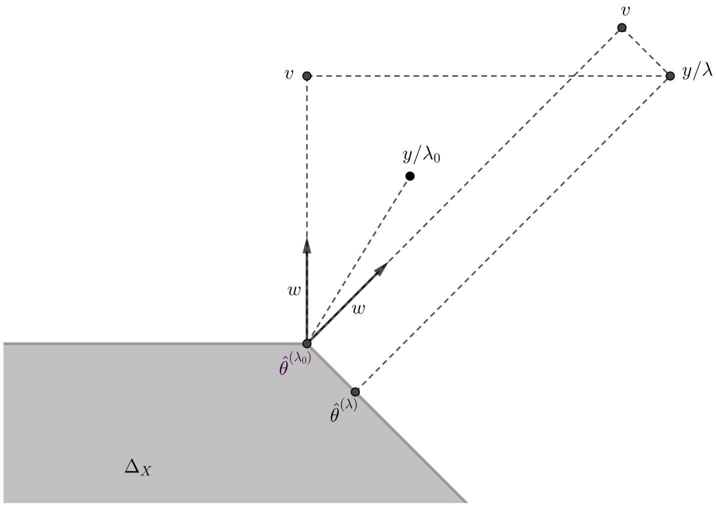

Figure 10: There are two possible dual propagated versions v of $\hat{\theta}^{(\lambda_0)}$ depending on the choice of face of Δ_X on which $\hat{\theta}^{(\lambda_0)}$ lies. Neither of them equal y/λ_0, so the dynamic DPP spheres of Proposition 16 constructed using $\hat{\theta}^{(\lambda_0)}$ do not equal their sequential counterparts. Additionally, neither choice of v has $v - \hat{\theta}^{(\lambda_0)}$ parallel to $y/\lambda_0 - \hat{\theta}^{(\lambda_0)}$, hence the dynamic dome tests of Proposition 17 are also not equal to their sequential counterparts − while their spheres are identical, their halfspaces differ.

2.5 Gap-Safe Screening

Fercoq et al. [19] proposed novel ‘gap-safe’ screening rules, so called because they leverage *duality gap* computations. These rules are inherently sequential and dynamic. In the sequential setting, they can deal with the problem of inexact knowledge with ease. In the dynamic setting, they can produce safe regions converging to zero in volume along the course of the optimisation. Hence, the equicorrelation set can be identified in a finite number of iterations; moreover, this is independent of the iterative Lasso solver used, the only requirement being that the solver is convergent.

The key idea of gap-safe rules is to *lower bound* the distance between $\hat{\theta}^{(\lambda)}$ and y/λ. None of the tests we have presented so far make use of such a lower bound, and in fact they tend to perform worst when λ is small and this distance is large; for example, the radii of the SAFE spheres becomes larger and larger as $\lambda \downarrow 0$, eventually becoming so large that no features are discarded.

Corollary 4 (Key Lower Bound)**.** Let $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$. Then

$$\left|\left|\frac{y}{\lambda} - \theta\right|\right|_2 \geqslant \widehat{R}_\lambda(\beta)$$

where

$$\widehat{R}_\lambda(\beta) := \frac{1}{\lambda}\sqrt{(||y||_2^2 - ||y - X\beta||_2^2 - 2\lambda||\beta||_1)_+}.$$

In particular, this holds when $\theta = \hat{\theta}^{(\lambda)}$.

Proof. Weak duality (2) states that

$$P_\lambda(\beta) = \frac{1}{2}||X\beta - y||_2^2 + \lambda||\beta||_1 \geqslant D_\lambda(\theta) = \frac{1}{2}||y||_2^2 - \frac{\lambda^2}{2}\left|\left|\frac{y}{\lambda} - \theta\right|\right|_2^2.$$

Rearranging this and noting the positive definiteness of the norm, we arrive at the claimed lower bound. □

Using this lower bound, we can locate $\hat{\theta}^{(\lambda)}$ within an annulus.

Corollary 5. Let $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$. Then

$$\widehat{R}_\lambda(\beta) \leqslant \left|\left|\frac{y}{\lambda} - \hat{\theta}^{(\lambda)}\right|\right|_2 \leqslant \check{R}_\lambda(\theta)$$

where

$$\check{R}_\lambda(\theta) := \left|\left|\frac{y}{\lambda} - \theta\right|\right|_2.$$

Proof. The first inequality follows from the previous Corollary. The second inequality follows from Proposition 14. □

We can now construct the gap-safe sphere.

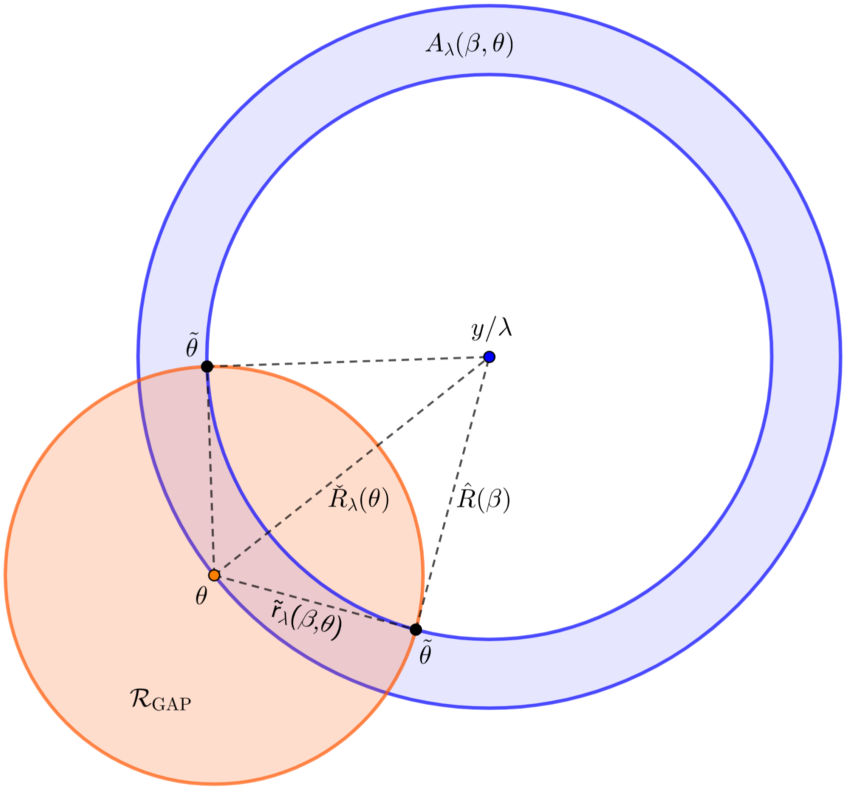
$A_\lambda(\beta,\theta)$
$\tilde{\theta}$
y/λ
$\check{R}_\lambda(\theta)$
$\hat{R}(\beta)$
θ
$\tilde{r}_\lambda(\beta,\theta)$
$\tilde{\theta}$
$\mathcal{R}_{\text{GAP}}$

Figure 11: Shown in orange and blue respectively are the gap-safe sphere and the annulus A_λ.

Theorem 4 (Gap-Safe Sphere)**.** Given any $(\beta,\theta)\in\mathbb{R}^p\times\Delta_X$, the gap-safe sphere

$$\mathcal{R}_{\text{GAP}}=\mathcal{B}(\theta,\tilde{r}_\lambda(\beta,\theta)),$$

where

$$\tilde{r}_\lambda(\beta,\theta)=\sqrt{\check{R}_\lambda(\theta)^2-\widehat{R}_\lambda(\beta)^2},$$

is a safe region, i.e. it contains $\hat{\theta}^{(\lambda)}$.

Proof. We know from Corollary 5 that $\hat{\theta}^{(\lambda)}$ belongs to the annulus

$$A_\lambda(\beta,\theta)=\left\{z:\widehat{R}_\lambda(\beta)\leqslant\left\|\frac{y}{\lambda}-z\right\|_2\leqslant\check{R}_\lambda(\theta)\right\}.$$

Since Δ_X and $\mathcal{B}(y/\lambda,\check{R}_\lambda(\theta))$ are convex, so too is their intersection. In fact, by Corollary 4,

$$\Delta_X\cap\mathcal{B}(y/\lambda,\check{R}_\lambda(\theta))=\Delta_X\cap A_\lambda(\beta,\theta),$$

hence the intersection of Δ_X with the annulus is convex. We know that $\hat{\theta}^{(\lambda)}$ is in this intersection and moreover that it must be the closest point in this intersection to y/λ.

We ask: how far can $\hat{\theta}^{(\lambda)}$ be from θ? Given the facts in the preceding paragraph, the points $\tilde{\theta}$ in Figure 11 are the points furthest from θ which could feasibly be $\hat{\theta}^{(\lambda)}$. The $\tilde{\theta}$ are on the boundary of the inner ball of the annulus, with the line from θ to $\tilde{\theta}$ tangent to this ball. Thus, $\tilde{\theta} - \theta$ and $y/\lambda - \tilde{\theta}$ are orthogonal, hence

$$\begin{aligned} \check{R}_\lambda(\theta)^2 &= ||\theta - \tilde{\theta}||_2^2 + \hat{R}_\lambda(\beta)^2 \\ ||\theta - \tilde{\theta}||_2^2 &= \check{R}_\lambda(\theta)^2 - \hat{R}_\lambda(\beta)^2 = \tilde{r}_\lambda(\beta,\theta)^2. \end{aligned}$$

Since $||\theta - \hat{\theta}^{(\lambda)}||_2 \leqslant ||\theta - \tilde{\theta}||_2$, it follows that $\hat{\theta}^{(\lambda)} \in \mathcal{B}(\theta, \tilde{r}_\lambda(\beta,\theta))$. □

Choosing $\beta = 0$ and $\theta = y/\lambda_{\max}$, as in the static screening setting, the gap-safe sphere is simply the dDPP sphere. Indeed, the gap-safe framework does not provide any innovations for static screening; sequential and dynamic screening is where it shines.

2.5.1 Dynamic Gap-Safe Screening

The volume of the gap-safe sphere can be bounded by the duality gap.

Definition 4 (Duality Gap). Given $(\beta,\theta) \in \mathbb{R}^p \times \Delta_X$, the *duality gap* is

$$G_\lambda(\beta,\theta) := P_\lambda(\beta) - D_\lambda(\theta).$$

Proposition 18 (Duality Gap Bound). For any $(\beta,\theta) \in \mathbb{R}^p \times \Delta_X$ the following holds

$$\tilde{r}_\lambda(\beta,\theta)^2 \leqslant r_\lambda(\beta,\theta)^2 := \frac{2}{\lambda^2} G_\lambda(\beta,\theta). \tag{21}$$

Proof. Use the fact that $\check{R}_\lambda(\theta)^2 = ||y/\lambda - \theta||_2^2$ and $\hat{R}_\lambda(\beta)^2 \geqslant (||y||_2^2 - ||y - X\beta||_2^2 - 2\lambda||\beta||_1)/\lambda^2$. □

The following result shows how to generate gap-safe spheres converging to zero in volume in the dynamic screening setting.

Corollary 6 (Gap-Safe Convergence). Let (β_k) be iterates generated from a convergent Lasso solver, possibly incorporated with safe screening tests in (some of) its iterations, as in Algorithm 2. Define θ_k to be the dual scaled versions of the residuals $R_k = y - X\beta_k$. Then the gap-safe spheres constructed from (β_k, θ_k) converge to zero in volume.

Proof. We have already established that incorporating dynamic safe screening into a convergent Lasso solver preserves the fact that its iterates converge to a Lasso solution. Then, $\theta_k \to \hat{\theta}^{(\lambda)}$, so $G_\lambda(\beta_k,\theta_k) \to P_\lambda(\hat{\beta}^{(\lambda)}) - D_\lambda(\hat{\theta}^{(\lambda)})$. By strong duality (3) this limit is zero. By Proposition 18, it follows that $\tilde{r}_\lambda(\beta_k,\theta_k) \to 0$. □

This result is very powerful. It implies that dynamic gap-safe screening identifies the equicorrelation set in finite time. That is, at some iteration K and beyond, the gap-safe sphere is equivalent to the Ultimate Safe Test. None of the dynamic regions presented in Section 2.4 were able to achieve this.

The computational cost of dynamic gap-safe screening is at best $\mathcal{O}(ns)$ per iteration, s the current number of active features. This is of the same order as one iteration of the underlying solver, so the screening should be applied every K iterations for some $K > 1$.

On top of the convergence in Corollary 6, the gap-safe framework offers another key practical benefit. The duality gap $G_\lambda(\beta_k, \theta_k)$ can be computed at negligible extra cost whilst constructing the gap-safe spheres. A sufficiently small duality gap makes for an appropriate criterion for termination of the Lasso solver: by weak duality, if $G_\lambda(\beta_k, \theta_k) \leqslant \epsilon$ then we have $0 \leqslant P_\lambda(\beta_k) - P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant \epsilon$. Thus, the gap-safe framework naturally ties in with using the duality gap as a stopping criterion.

2.5.2 Sequential Gap-Safe Screening

Suppose we are in the sequential screening setting: having just computed an approximation β_0 of $\hat{\beta}^{(\lambda_0)}$, we seek to screen features for $\hat{\beta}^{(\lambda)}$. We obtain an approximation $\tilde{\theta}$ of $\hat{\theta}^{(\lambda_0)}$ via $\tilde{\theta} = (y - X\beta_0)/\lambda_0$. Letting θ_0 be the dual scaled version of $\tilde{\theta}$, we can construct a gap-safe sphere from β_0 and θ_0 as in Theorem 4 and use this as a sequential screening rule; that is, we construct the sphere centred at θ_0 with radius $\tilde{r}_\lambda(\beta_0, \theta_0)$. The problem of inexact knowledge of $\hat{\theta}^{(\lambda_0)}$ is instantly solved because all we require to construct gap-safe spheres are $\beta \in \mathbb{R}^p$ and dual feasible θ.

The following result relates the radius of the gap-safe sphere constructed in this way to the previous solution's duality gap.

Proposition 19 (Sequential Gap-Safe, [19] Proposition 3)**.** Let $\lambda_0, \lambda > 0$. Let $(\beta_0, \theta_0) \in \mathbb{R}^p \times \Delta_X$. Then

$$\begin{aligned}\tilde{r}_\lambda(\beta_0, \theta_0)^2 &\leqslant r_\lambda(\beta_0, \theta_0)^2 \\ &= \left(\frac{\lambda_0}{\lambda}\right) r_{\lambda_0}(\beta_0, \theta_0)^2 + \left(1 - \frac{\lambda}{\lambda_0}\right) \left\|\frac{y - X\beta_0}{\lambda}\right\|_2^2 - \left(\frac{\lambda_0}{\lambda} - 1\right) ||\theta_0||_2^2.\end{aligned}$$

The key takeaway is that if the duality gap $G_{\lambda_0}(\beta_0, \theta_0)$ is small (as it will be if β_0 is a good approximation) and λ_0 is close to λ, the sequential gap-safe sphere is small.

The computational cost of applying the sequential gap-safe screening strategy is similar to that of the sequential tests of Section 2.3: about two or three iterations of the underlying solver, at most.

2.6 Experiments

We implement the safe screening strategies presented on two datasets: the standard Leukemia dataset and a synthetic dataset, which we call Random, generated using the `make_regression` function from the Python package Scikit-Learn. The former has $n = 72$ and $p = 7{,}128$ and the latter has $n = 50$, $p = 20{,}000$. All of the code used in our experiments is written in pure Python. Where relevant, the underlying Lasso solver used is the co-ordinate descent algorithm.

2.6.1 Static Screening

We compare the static screening methods of Section 2.3 by the percentage of features they screen out, as a function of λ. Figure 12 displays plots of these percentages for the two datasets.

As expected, diDT screens the most features and diSAFE screens more than dDPP and dSAFE. All of the tests screen more features as λ is increased, screening almost every feature as λ approaches $\lambda_{\max}$. Each test has a threshold of λ below which it is useless, screening no features; the better tests have smaller such thresholds.

All of the tests perform better for the Leukemia dataset than for the Random dataset. Remarkably, the dome tests reject nearly all of the features for the Leukemia data when $\lambda/\lambda_{\max} > 0.1$.

There are two notable differences in the plots. Firstly, the ranking of the tests differs slightly for the two datasets: the default dome test (dDT) performs better than the default iSAFE (diSAFE) for the Leukemia dataset but not for the Random dataset. Secondly, the performance boost of the domes, relative to their component spheres, appears to be larger for the Leukemia dataset. These observations can be explained as follows. The dome regions deliver the biggest improvement to their underlying spheres when $|\mathrm{Corr}(X^*, y)|$ is large, where X^* is normal to the face of Δ_X on which $y/\lambda_{\max}$ lies. For the Leukemia dataset, this quantity is very large: 0.97. For the Random dataset, it is only 0.57, hence the domes do not offer as large a performance boost and in particular the improvement delivered is not large enough for dDT to outperform diSAFE.

2.6.2 Dynamic Screening

In Figure 13 we compare various dynamic screening strategies applied to the Leukemia dataset. Specifically, we compare the number of features screened out over the course of the process of solving the Lasso. Each strategy screens features before the solver begins and then every 20 iterations, up to 500 iterations.

The plots show that the dynamic SAFE spheres and the dome tests do not benefit much from the dynamic screening framework: they do not show a great deal of improvement over the course of the algorithm. The DPP tests show more improvement, but still appear to plateau quickly for most values of λ. The gap-safe sphere clearly reaps the most benefit from the dynamic framework. Despite starting off by screening fewer iterations than the other strategies, it continuously improves, reaching a plateau only when it has screened out almost every feature.

2.6.3 Computational Gains

The end goal of safe screening is to solve the Lasso problem more quickly. We explore the performance boost delivered by safe screening through two experiments.

In Figure 14 we plot the speedup attained by various safe screening strategies. Each strategy is applied to solve the Lasso problem with the zero vector as the initial iterate, for a range of λ. The speedup is expressed by the time taken to attain a duality gap of less than 10^{-2}, as a fraction of the time taken using no screening strategy. The stopping criterion was checked every 10 iterations; in addition, the dynamic strategies conducted screening every 10 iterations.

The strategies perform best for large λ, delivering speedups of an order of magnitude or more as $\lambda \uparrow \lambda_{\max}$. At the opposite extreme, as $\lambda \downarrow 0$ the fraction of time saved is around 0; that said, for such λ the problem takes a long time to solve, so the speedup can still be

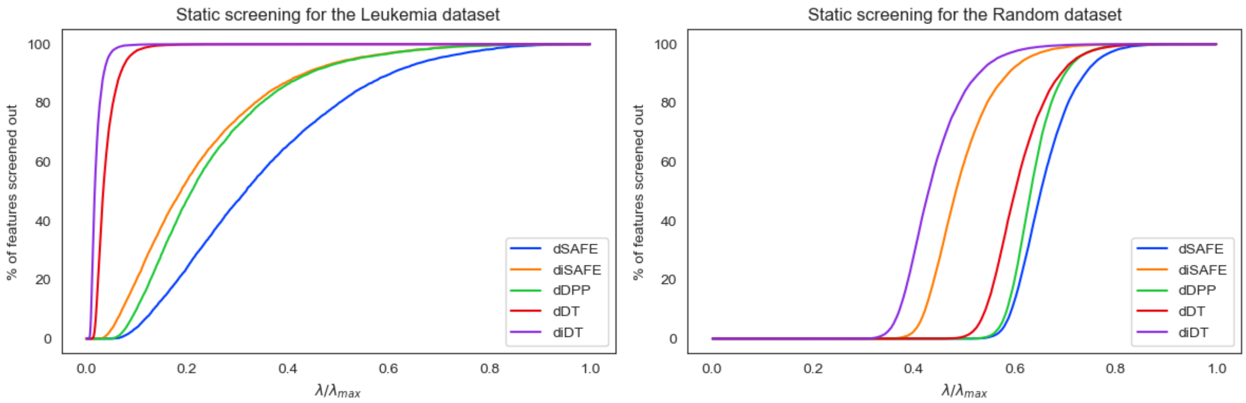

Figure 12: Comparison of static screening strategies by features rejected.

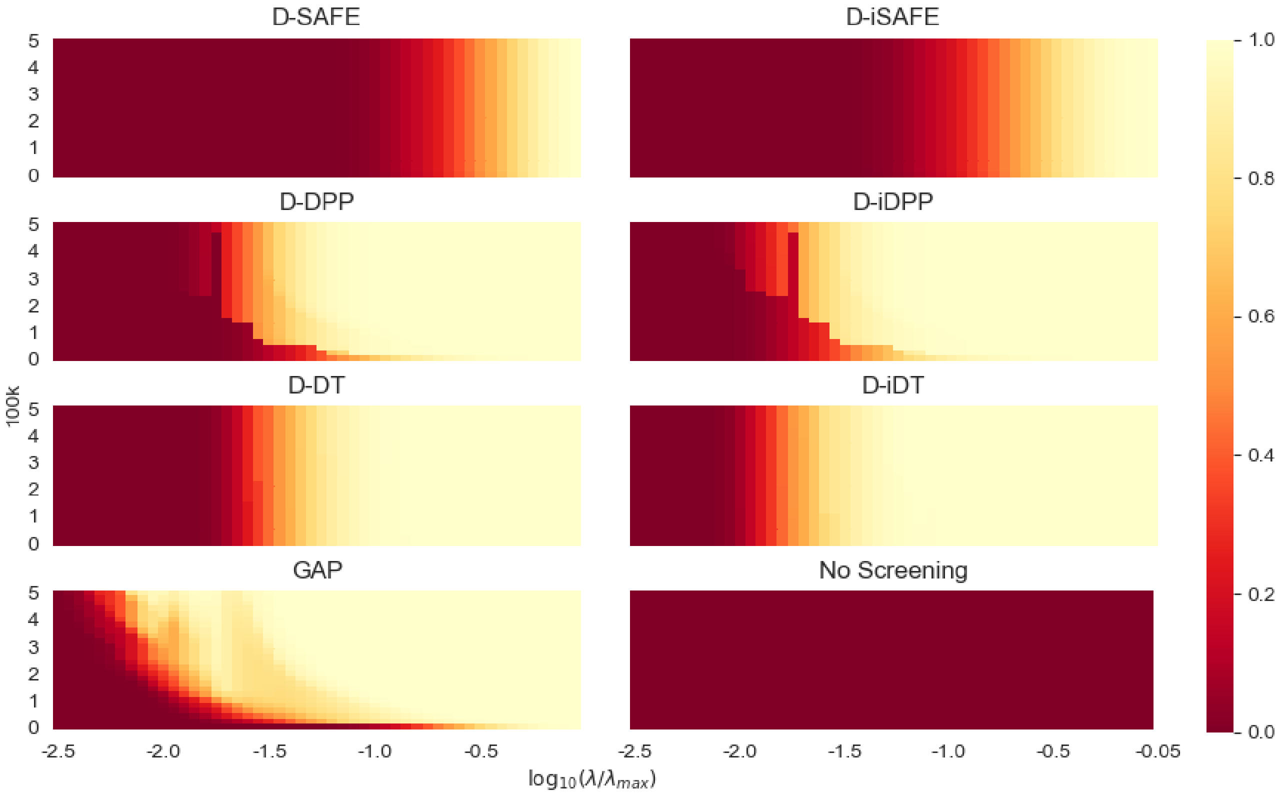

Figure 13: Comparison of dynamic screening strategies by features rejected, using the Leukemia dataset. The colours indicate the proportion of features that are rejected, as a function of the iteration k and the regularisation parameter λ. Better strategies have longer and thicker light yellow regions.

large in absolute terms. The gap-safe sphere is the top-performing strategy on the Random dataset for all but a small range of large λ values. In contrast, both static and dynamic iDT appear to outperform or at least closely match the gap-safe sphere on the Leukemia dataset. This can be explained by the fact that the Leukemia dataset is particularly adapted for the dome tests since it has $|\mathrm{Corr}(X^*, y)| = 0.97$.

Our next experiment concerns a common scenario in practice. We applied various strategies to compute Lasso solutions over a grid of regularisation parameters $\lambda_1 > ... > \lambda_N$. We chose the default grid used by the glmnet and Scikit-Learn packages: $\lambda_i = \lambda_{\max} 10^{-3(i-1)/(N-1)}$, $N = 100$. We first solved for $\hat{\beta}^{(\lambda_1)}$ and then initialised the computation for $\hat{\beta}^{(\lambda_{i+1})}$ using the output of the previous computation, $\hat{\beta}^{(\lambda_i)}$. A duality gap of less than ϵ was used as the stopping criterion for each Lasso problem, which was checked every 10 iterations. We used the Leukemia dataset and varied ϵ over the set $\{10^{-3}, 10^{-2}, 10^{-1}\}$. Figure 15 displays the results.

The gap-safe sphere is clearly the top performer for all values of ϵ, solving the Lasso problem up to 5x faster than the other strategies. Notably, the relative performance of the gap-safe sphere increases as ϵ decreases, indicating that it is particularly useful when accurate solutions are sought.

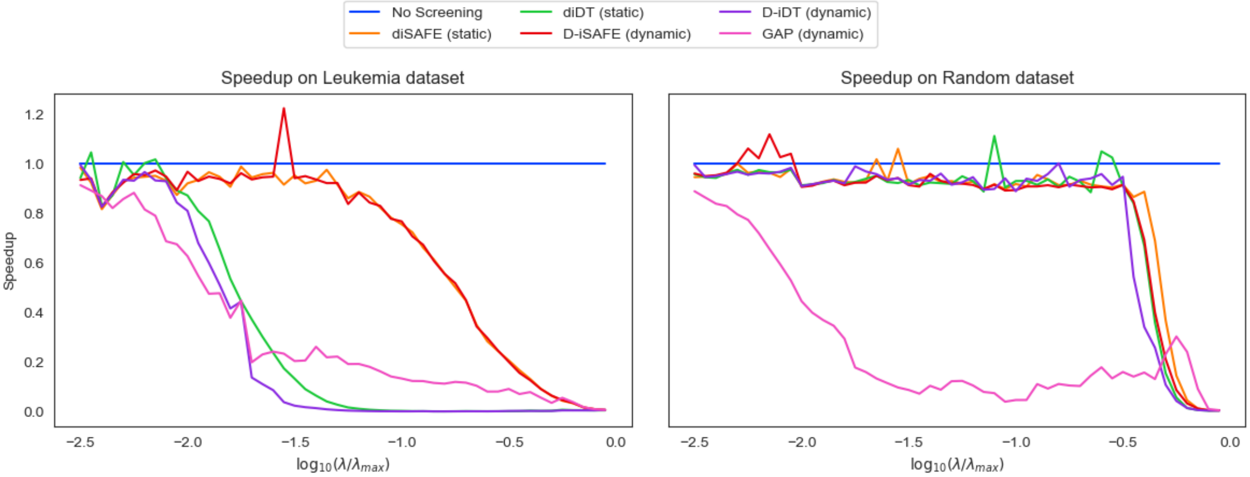

Figure 14: Speedup obtained by various safe screening strategies, expressed as the ratio of the time taken to solve the Lasso to the time taken with no screening. Lower values are better.

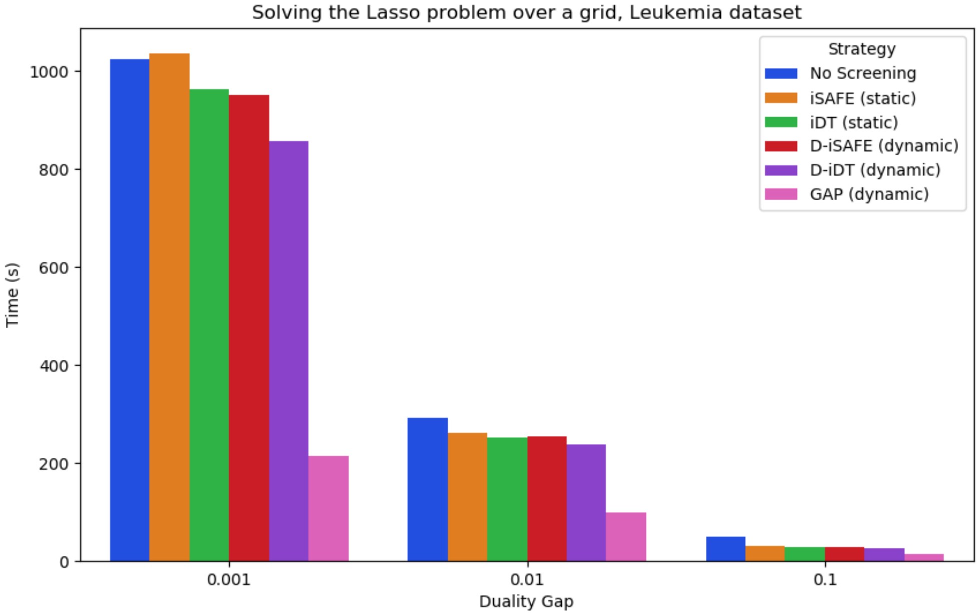

Figure 15: Comparison of various safe screening strategies by the time taken to solve the Lasso problem over a grid of regularisation parameters, using the Leukemia dataset.
Note that the static iSAFE and iDT are simply the D-iSAFE and D-iDT at iteration 0. The dynamic strategies continue to evolve whereas the static strategies screen only at iteration 0.

3. Safe Screening for Other Problems

Whilst the Lasso has been our primary focus, safe screening can be extended to other learning problems with sparsity-enforcing penalties, such as the Group Lasso and ℓ_1-regularised logistic regression. The principle is the same: identify inactive features (or groups of features) in advance and remove them, leaving a smaller problem to solve.

In this Chapter, we generalise gap-safe screening to a broad range of penalised learning problems, following closely the work of Ndiaye et al. [20]. Safe screening rules for problems other than the Lasso have been proposed earlier in the literature, for example in [8, 9, 16], though these approaches have been highly problem-specific. In contrast, gap-safe screening fluidly generalises to a large class of problems, maintaining its attractive properties in the Lasso setting.

3.1 Results from Convex Analysis

We briefly review some results from convex analysis that will be used in this Chapter.

Throughout, let $f : \mathbb{R}^k \to (-\infty, \infty]$ such that $f(x) < \infty$ for some $x \in \mathbb{R}^k$. The *dual function* or *Fenchel conjugate* of f is the function $f^* : \mathbb{R}^k \to (-\infty, +\infty]$ defined by $f^*(y) = \sup_{x \in \mathbb{R}^k} \left(\langle x, y \rangle - f(x) \right)$; this is convex, even if f is not. If f is convex, we say f is *strongly convex* with parameter μ (or simply μ-strongly convex) if $f(x) - \mu||x||_2^2/2$ is convex, and f is *smooth* with parameter ν (or simply $\nu-$smooth) if f is differentiable with ν-Lipschitz continuous gradient. We say $g \in \mathbb{R}^k$ is a *subgradient* of f at x if $f(x+h) \geqslant f(x) + \langle g, h \rangle$ for all $h \in \mathbb{R}^k$ and denote by $\partial f(x)$ the *subdifferential* of f at x, which we define as the set of all subgradients of f at x. For a norm Ω on $\mathbb{R}^k$, we define its *dual norm* Ω^D via $\Omega^D(y) = \max_{\Omega(x) \leqslant 1} \langle x, y \rangle$; this is also a norm on $\mathbb{R}^k$. We point out that the ℓ_1 and $\ell_\infty-$norms are the duals of one another and that the ℓ_2-norm is its own dual. Moreover, Ω and its dual satisfy a generalised Cauchy-Schwarz inequality: $|\langle x, y \rangle| \leqslant \Omega(x)\Omega^D(y)$ for $x, y \in \mathbb{R}^k$.

Proposition 20 (Subgradient Optimality Condition)**.** f is minimised at $x \in \mathbb{R}^k$ if and only if $0 \in \partial f(x)$.

Proposition 21 (Fenchel Inequality)**.** f and its dual f^* satisfy the *Fenchel inequality*

$$x^T y \leqslant f(x) + f^*(y)$$

for $x, y \in \mathbb{R}^k$. Moreover, equality holds if and only if $y \in \partial f(x)$.

Proposition 22 ([7] Theorem 1)**.** Assume f is convex and $(1/\gamma)-$smooth. Then f^* is $\gamma-$strongly convex.

Proposition 23. Assume $f : \mathbb{R} \to \mathbb{R}$ is convex. Then f is differentiable at $x \in \mathbb{R}$ with derivative g if and only if $\partial f(x) = \{g\}$.

3.2 A General Learning Problem

In this section we formulate a general learning problem which encompasses the Lasso problem as a special case.

We seek to estimate a vector of parameters $\beta \in \mathbb{R}^p$ by a solution of the following problem:

$$\min_{\beta \in \mathbb{R}^p} P_\lambda(\beta) \text{ for } P_\lambda(\beta) = \left(\sum_{i=1}^{n} f_i(x_i^T \beta) + \lambda \Omega(\beta_{\mathcal{P}}) \right) \tag{22}$$

where:

- $\lambda > 0$ is the *regularisation parameter* controlling the extent of the penalisation through Ω.
- $f_i : \mathbb{R} \to \mathbb{R}$ are convex and $(1/\gamma)-$smooth, $\gamma > 0$.
- $\mathcal{P} \subseteq [p]$ is a non-empty set of *penalised* parameters, with $\mathcal{U} = [p] \backslash \mathcal{P}$ a possibly non-empty set of *unpenalised* parameters. Further, Ω is a group-decomposable norm on $\mathbb{R}^{\mathcal{P}}$:

  $$\Omega(\beta_{\mathcal{P}}) = \sum_{g \in \mathcal{G}} \Omega_g(\beta_g)$$

  where $\mathcal{G}$ is a collection of disjoint, non-empty subsets of $[p]$, each subset $g \in \mathcal{G}$ corresponding to a group of parameters, and Ω_g are norms on $\mathbb{R}^g$.
- A solution of (22) exists for all $\lambda > 0$.

We write $\hat{\beta}^{(\lambda)}$ for a solution of (22).

Remark 10. Typically the f_i arise as transformed versions of a common function f; for example, for the Lasso we have $f_i(\xi) = f(\xi - y_i)$ where $f(\xi) = \frac{1}{2}\xi^2$ and y_i is the ith response. The grouping of parameters via $\mathcal{G}$ allows for models which penalise parameters in a group-wise fashion, such as the Group Lasso. We have allowed for the existence of a group $\mathcal{U}$ of unpenalised parameters, which the problem in [20] does not permit. This further generalisation is useful because, while unpenalised parameters such as intercept terms can be safely removed in the linear model, this is not the case for other models.

The following Theorem is a generalisation of Theorem 1.

Theorem 5 (Dual Problem and Optimality Conditions)**.** The dual problem

$$\max_{\theta \in \Delta_X} D_\lambda(\theta) \text{ for } D_\lambda(\theta) = -\sum_{i=1}^{n} f_i^*(-\lambda \theta_i),$$

where

$$\Delta_X = \{\theta \in \mathbb{R}^n : \Omega^D(X_{\mathcal{P}}^T \theta) \leqslant 1 \text{ and } X_{\mathcal{U}}^T \theta = 0\}$$

is the *dual feasible region*, has a unique solution/maximiser $\hat{\theta}^{(\lambda)}$. Moreover, we have the following relations between the primal and dual objectives:

$$P_\lambda(\beta) \geqslant D_\lambda(\theta) \text{ for all } (\beta, \theta) \in \mathbb{R}^p \times \Delta_X \qquad \textit{(weak duality)} \tag{23}$$

$$\min_{\beta \in \mathbb{R}^p} P_\lambda(\beta) = P_\lambda(\hat{\beta}^{(\lambda)}) = D_\lambda(\hat{\theta}^{(\lambda)}) = \max_{\theta \in \Delta_X} D_\lambda(\theta) \qquad \textit{(strong duality)} \tag{24}$$

Further, primal and dual solutions are related in the following manner:

$$\begin{aligned} &\text{i)} \ \ \lambda\hat{\theta}^{(\lambda)} + J(X\hat{\beta}^{(\lambda)}) = 0 \text{ where } J(z) := (f_1'(z), \ldots, f_n'(z)) \\ &\text{ii)} \ \ \Omega_g^D(X_g^T\hat{\theta}^{(\lambda)}) < 1 \implies \hat{\beta}_g^{(\lambda)} = 0 \\ &\text{iii)} \ \ \hat{\beta}_g^{(\lambda)} \neq 0 \implies \hat{\beta}_g^{(\lambda)T}(X_g^T\hat{\theta}^{(\lambda)}) = \Omega_g(\hat{\beta}_g^{(\lambda)}) \end{aligned} \tag{25}$$

These *optimality conditions* are necessary and sufficient for $\hat{\beta}^{(\lambda)}$ to be a solution.

Proof. Provided in the Appendix. □

Remark 11. In the Lasso case, the dual problem had a convenient interpretation: it was the minimisation of the Euclidean distance to y/λ over Δ_X. This allowed for safe regions to be constructed using properties of projections onto closed and convex sets. In this general setting, we do not have such an interpretation of the dual problem and so the screening tests of Section 2.3 do not generalise easily.

Remark 12. The most restrictive assumption of the form of problem (22) (aside from convexity) is that of smoothness of the f_i. Moreover, an implicit assumption of ours is that J and the f_i^* are known analytically and can be evaluated. Many common convex learning problems fit into this framework, penalised logistic regression being an example. We refer the reader to [20] for a detailed discussion of popular learning problems as specialisations of (22).

Next, we generalise Proposition 1.

Theorem 6 (Critical Threshold: $\lambda_{\max}$)**.** If $\mathcal{U}$ is non-empty, let $\hat{\beta}_U$ be a solution of the problem (22) with all penalised parameters forced to 0, so that the minimisation is over $\beta_{\mathcal{U}} \in \mathbb{R}^{\mathcal{U}}$. Define

$$\lambda_{\max} = \begin{cases} \Omega^D\left(X_{\mathcal{P}}^T\, J(X_{\mathcal{U}}\hat{\beta}_{\mathcal{U}})\right) & \mathcal{U} \text{ non-empty} \\ \Omega^D\left(X_{\mathcal{P}}^T\, J(0)\right) & \mathcal{U} \text{ empty} \end{cases}.$$

There exists a solution $\hat{\beta}^{(\lambda)}$ of (22) with $\hat{\beta}_{\mathcal{P}}^{(\lambda)} = 0$ if and only if $\lambda \geqslant \lambda_{\max}$. Moreover, if $\lambda > \lambda_{\max}$ then all solutions $\hat{\beta}^{(\lambda)}$ have $\hat{\beta}_{\mathcal{P}}^{(\lambda)} = 0$, and if f_i are *strictly* convex then this also holds for $\lambda = \lambda_{\max}$.

Proof. Provided in the Appendix. □

This Theorem is useful in the same way as Proposition 1: knowledge of $\lambda_{\max}$ aids in choosing λ. Computing $\lambda_{\max}$ is more involved when $\mathcal{U}$ is non-empty, since $\hat{\beta}_{\mathcal{U}}$ is required.

That said, finding $\hat{\beta}_{\mathcal{U}}$ involves minimising a convex, differentiable objective, and moreover $\mathcal{U}$ is typically small (for example, containing a single intercept term), so the computation is not too cumbersome.

Just as (4) ii) was key for safe screening for the Lasso, (25) ii) is the basis of safe screening in this more general setting. The idea is to construct a safe region $\mathcal{R}$ containing $\hat{\theta}^{(\lambda)}$ and to use (25) ii) to screen group $g \in \mathcal{G}$.

3.3 Generalised Gap-Safe Screening

It turns out that the smoothness assumption on f_i is all that is needed in constructing the gap-safe sphere; this is indeed satisfied for the Lasso, with $\gamma = 1$.

Theorem 7 (Generalised Gap-Safe Sphere)**.** Given any $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$, the gap-safe sphere

$$\mathcal{R}_{\text{GAP}} = \mathcal{B}\left(\theta, \sqrt{2G_\lambda(\beta,\theta)/\gamma\lambda^2}\right)$$

contains $\hat{\theta}^{(\lambda)}$, where

$$G_\lambda(\beta,\theta) = P_\lambda(\beta) - D_\lambda(\theta)$$

is the duality gap.

Proof. Define $d := -D_\lambda$. For convenience, we define d to be ∞ outside of Δ_X. By Proposition 22, f_i^* are $\gamma-$strongly convex, so d is $\gamma\lambda^2-$strongly convex. Since $\hat{\theta}^{(\lambda)}$ minimises d, by Proposition 20 we must have $0 \in \partial d(\hat{\theta}^{(\lambda)})$. By [7] Lemma 2, for $\theta \in \mathbb{R}^n$

$$d(\theta) \geqslant d(\hat{\theta}^{(\lambda)}) + \frac{\gamma\lambda^2}{2}||\theta - \hat{\theta}^{(\lambda)}||_2^2,$$

hence for $\theta \in \Delta_X$

$$D_\lambda(\hat{\theta}^{(\lambda)}) - D_\lambda(\theta) \geqslant \frac{\gamma\lambda^2}{2}||\theta - \hat{\theta}^{(\lambda)}||_2^2.$$

By weak duality (23), $P_\lambda(\beta) \geqslant D_\lambda(\hat{\theta}^{(\lambda)})$. The claim follows. □

Remark 13. The proof of this theorem presented in [20] assumes D_λ is differentiable at $\hat{\theta}^{(\lambda)}$, but this need not be true. This error is minor; our proof corrects it through the fact $0 \in \partial d(\hat{\theta}^{(\lambda)})$.

The final ingredient of gap-safe screening is a method for generating dual feasible θ from arbitrary $\beta \in \mathbb{R}^p$, which will allow the gap-safe sphere to be used in the static, sequential or dynamic screening settings in the same way as for the Lasso. In the dynamic screening setting, we should like the gap-safe spheres to converge to zero in volume if the underlying solver is convergent; for this, we will require that if $\beta_k \to \hat{\beta}^{(\lambda)}$ then $\theta_k \to \hat{\theta}^{(\lambda)}$.

We first require a scheme for constructing dual feasible θ from arbitrary $z \in \mathbb{R}^n$, analogous to dual scaling.

Lemma 3 (Generalised Dual Scaling)**.** Given $z \in \mathbb{R}^n$, define

$$\tilde{z} = \begin{cases} \mathcal{P}_{\mathrm{Ker}(X_{\mathcal{U}}^T)}(z) & \mathcal{U} \text{ non-empty} \\ z & \mathcal{U} \text{ empty} \end{cases}$$

Next, defining $\theta = \mu \tilde{z}$ where

$$\mu = \begin{cases} 1 & \Omega^D(X_{\mathcal{P}}^T \tilde{z}) \leqslant 1 \\ 1/\Omega^D(X_{\mathcal{P}}^T \tilde{z}) & \text{otherwise} \end{cases},$$

θ is dual feasible, i.e. $\theta \in \Delta_X$. We call it the dual scaled version of z.

Next, we require a candidate z given any $\beta \in \mathbb{R}^p$. For this, just as we used (4) i) for the Lasso, we use (25) i), choosing $z = -J(X\beta)/\lambda$.

The following result shows that the dynamic gap-safe spheres converge to zero volume just as in the Lasso setting.

Corollary 7 (Generalised Gap-Safe Convergence)**.** Let (β_k) be iterates generated from a convergent, iterative solver for the problem (22), possibly incorporated with safe screening tests in (some of) its iterations. Define θ_k to be the dual scaled versions of $-J(X\beta_k)/\lambda$. Then the gap-safe spheres constructed from (β_k, θ_k) converge to zero in volume.

Proof. Similarly to the Lasso case, incorporating dynamic safe screening into a convergent solver preserves the fact that its iterates converge to a Lasso solution. Then, $-J(X\beta_k)/\lambda \to \hat{\theta}^{(\lambda)}$ by (25) i), hence also $\theta_k \to \hat{\theta}^{(\lambda)}$. It follows that $G_\lambda(\beta_k, \theta_k) \to P_\lambda(\hat{\beta}^{(\lambda)}) - D_\lambda(\hat{\theta}^{(\lambda)})$. By strong duality (24) this limit is zero. □

4. Discussion

We have presented various safe screening strategies for accelerating algorithms solving the Lasso problem. Many more safe screening strategies exist in the literature – our focus has been to motivate a few of the most pivotal as applications of basic facts on projections, thereby highlighting some of the high-level ideas behind their construction. Moreover, though all of our safe tests relied on spherical or dome-shaped safe regions, one can concoct smaller safe regions of greater complexity – for example, by intersecting a sphere with not one but two hyperplanes – at the expense of more difficult execution of the tests; Xiang et al. [12] provide examples.

The gap-safe safe tests are unique among safe screening strategies for their use of duality gap computations. Thanks to these, the gap-safe dynamic screening strategy is able to identify the equicorrelation set in finite time. To the best of our knowledge, this is the only screening strategy to exhibit this property.

Our numerical experiments demonstrate the practical effectiveness of safe screening. The gap-safe framework stands out as a particularly versatile and performant screening strategy. It can sometimes be outperformed by other strategies, for example in the setting where the features are highly correlated with the response and the regularisation parameter is large; one could potentially squeeze out more performance by blending the gap-safe strategy with other strategies, but we do not think the gains will be substantial on most data sets.

In Chapter 3, we demonstrated that the gap-safe spheres can be generalised to a wide class of convex learning problems subject to a certain smoothness assumption. Importantly, the attractive properties of the gap-safe framework in the Lasso setting are preserved. This opens the door for speedups on popular problems such as ℓ_1-regularised logistic regression and the Group Lasso.

Though we did not explore it in this essay, another idea worthy of mention is that of safe screening of *observations*, rather than features, in the context of support vector machines. Analogously to the Lasso solution depending only on a small number of features, the solution of the SVM problem depends only on a small portion of the observations. Screening rules aimed at leveraging this fact were first proposed by Ogawa et al. [13].

References

[1] R. Tibshirani. Regression shrinkage and selection via the lasso. In *Journal of the Royal Statistical Society: Series B*, 1996.

[2] R. J. Tibshirani. The lasso problem and uniqueness. In *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[3] A. Beck, M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. In *SIAM Journal on Imaging Sciences*, 2(1), 183-202, 2009.

[4] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani. Pathwise coordinate optimization. In *Annals of Applied Statistics*, 1(2):302-332, 2007.

[5] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. In *Journal of the Royal Statistical Society: Series B*, 74(2):245-266, 2012.

[6] S. Boyd, L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004. ISBN 0521833787.

[7] Xingyu Zhou. On the Fenchel duality between strong convexity and Lipschitz continuous gradient. *arXiv:1803.06573v1 [math.OC]*, 2018.

[8] L. El Ghaoui, V. Viallon, T. Rabbani. Safe feature elimination in sparse supervised learning. EECS Department, University of California, Berkeley, Tech. Rep., 2010.

[9] L. El Ghaoui, V. Viallon, T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv:1009.4219v2 [cs.LG]*, 2011.

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

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

[12] Z.J. Xiang, Y. Wang, P. J. Ramadge. Screening tests for lasso problems. *arXiv: 1405.4897v1 [cs.LG]*, 2014.

[13] K. Ogawa, Y. Suzuki, I. Takeuchi. Safe Screening of Non-Support Vectors in Pathwise SVM Computation. In *International Conference on Machine Learning*, 2013.

[14] J. Wang, P. Wonka, J. Ye. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*, 2013.

[15] J. Liu, Z. Zhao, J. Wang, J. Ye. Safe screening with variational inequalities and its application to lasso. In *International Conference on Machine Learning*, 2014.

[16] J. Wang, J. Zhou, J. Liu, P. Wonka, J. Ye. A safe screening rule for sparse logistic regression. In *International Conference on Machine Learning*, 2014.

[17] A. Bonnefoy, V. Emiya, L. Ralaivola, R. Gribonval. A dynamic screening principle for the lasso. In *European Signal Processing Conference*, 2014.

[18] A. Bonnefoy, V. Emiya, L. Ralaivola, R. Gribonval. Dynamic screening: accelerating first-order algorithms for the lasso and group-lasso. In *IEEE Transactions on Signal Processing*, 2015.

[19] O. Fercoq, A. Gramfort, J. Salmon. Mind the duality gap: safer rules for the lasso. In *International Conference on Machine Learning*, 2015.

[20] E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon. Gap safe screening rules for sparsity enforcing penalties. In *Journal of Machine Learning Research*, 2017.

Appendix

Lasso Solvers

We provide a brief review of three popular Lasso solvers: ISTA, FISTA and co-ordinate descent.

Each of these solvers is iterative. They start with an initial value or guess β_0 of a Lasso solution, a common choice of β_0 being the p-dimensional zero vector. Given the iterate β_{k-1}, we form the next iterate β_k according to some update rule. The hope is that the sequence of iterates (β_k) converge to a solution of the Lasso problem. We do this until some *stopping criterion* is triggered. Algorithm 4 presents the strategy.

Algorithm 4: Solving Strategy

```
input $\beta_0$
$k \leftarrow 0$
repeat
    $k \leftarrow k + 1$
    $\beta_k \leftarrow$ update $\beta_{k-1}$
until stopping criterion triggered
```

The three algorithms differ only in their update rules and can be applied to problems other than the Lasso problem (1). It can be shown that they are all *convergent*: given any initial value β_0, the sequence of iterates (β_k) converges to a Lasso solution.

We do not go into detail on the update rules. We point out two facts:

1) ISTA and FISTA compute $X^T R_k$ during their update steps, where $R_k = y - X\beta_k$ is the kth residual. The co-ordinate descent update does not require this quantity.

2) The update step of co-ordinate descent is a for loop over co-ordinates. That is, the update step involves updating each co-ordinate one-by-one, using the latest values of all of the other co-ordinates.

Proofs

Proof of Theorem 5

We rewrite the problem (22) as the constrained problem

$$\min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \left(\sum_{i=1}^n f_i(z_i) + \lambda\Omega(\beta_{\mathcal{P}})\right) \text{ subject to } z = X\beta. \tag{26}$$

By forming the Lagrangian function

$$L_\lambda(\beta, z, \theta) = \left(\sum_{i=1}^n f_i(z_i) + \lambda\Omega(\beta_{\mathcal{P}})\right) + \lambda\theta^T(z - X\beta)$$

we obtain the dual function

$$\begin{aligned} g_\lambda(\theta) &= \inf_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} L_\lambda(\beta, z, \theta) \\ &= \begin{cases} -\sum_{i=1}^n f_i^*(-\lambda\theta_i) & \theta \in \Delta_X \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

The dual problem is to maximise this.

By definition of g_λ, it follows that $L_\lambda(\beta, z, \theta) \geqslant g_\lambda(\theta)$ for all $(\beta, z, \theta) \in \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n$. Weak duality follows:

$$P_\lambda(\beta) = L_\lambda(\beta, X\beta, \theta) \geqslant g_\lambda(\theta) \text{ for all } (\beta, \theta) \in \mathbb{R}^p \times \mathbb{R}^n.$$

By the assumption that a solution of (22) exists, $\min_{\beta\in\mathbb{R}^p} P_\lambda(\beta)$ is finite. Further, Slater's condition is trivially satisfied by the constrained problem (26), hence strong duality holds and there exists a dual solution ([6] Chapter 5).

By Proposition 22, f_i^* are $\gamma-$strongly convex hence strictly convex. Noting that Δ_X is closed and convex, the dual problem is the maximisation of a strictly concave function over a closed, convex set. Hence, any maximiser is unique. Thus, there exists a unique dual solution $\hat{\theta}^{(\lambda)}$.

It is straightforward to show that

$$L(\beta, X\beta, \theta) = \inf_{\beta'\in\mathbb{R}^p, z'\in\mathbb{R}^n} L(\beta', z', \theta) \tag{27}$$

is sufficient for β, θ to be primal and dual solutions, respectively. By existence of solutions to (22) and strong duality, it is in fact also necessary.

The optimality conditions (25) follow straightforwardly from (27). For (25) i), it is also necessary to invoke Propositions 21 and 23. □

Proof of Theorem 6

Throughout, we assume $\mathcal{U}$ is non-empty. The case $\mathcal{U}$ empty follows by setting $X_{\mathcal{U}}$ to have all entries zero.

It follows from convexity of f_i and the fact that Ω is a norm that, for $\lambda_1 > \lambda_2$, we have

$$\Omega(\hat{\beta}_{\mathcal{P}}^{(\lambda_1)}) \leqslant \Omega(\hat{\beta}_{\mathcal{P}}^{(\lambda_2)}). \tag{28}$$

Hence, if there exists $\hat{\beta}^{(\lambda^\star)}$ with $\hat{\beta}_{\mathcal{P}}^{(\lambda^\star)} = 0$, then $\hat{\beta}_{\mathcal{P}}^{(\lambda)} = 0$ for all $\hat{\beta}^{(\lambda)}$ for all $\lambda > \lambda^\star$.

Decomposing arbitrary $\beta \in \mathbb{R}^p$ as $\beta = (\beta_{\mathcal{P}}, \beta_{\mathcal{U}})$, we have, using the subgradient optimality condition, that

$$(0, v) \in \arg\min_{\beta} P_\lambda(\beta) \iff v = \hat{\beta}_{\mathcal{U}} \text{ and } \Omega^D(X_{\mathcal{P}}^T\, J(X_{\mathcal{U}}\hat{\beta}_{\mathcal{U}})) \leqslant \lambda \tag{29}$$

for some $\hat{\beta}_{\mathcal{U}}$ defined as in the statement of the theorem. Thus, for $\lambda \geqslant \lambda_{\max}$, there exists $\hat{\beta}^{(\lambda)}$ with $\hat{\beta}_{\mathcal{P}}^{(\lambda)} = 0$, where $\lambda_{\max}$ is defined as in the statement of the theorem.

Note that $\hat{\beta}_{\mathcal{U}}$ may not be unique; we have worked with a fixed choice. However, $\lambda_{\max}$ is independent of the particular choice of $\hat{\beta}_{\mathcal{U}}$. To see this, suppose otherwise: there exist $\hat{\beta}_{\mathcal{U}}^{(1)}$, $\hat{\beta}_{\mathcal{U}}^{(2)}$ giving rise to $\lambda_{\max}^{(1)}$, $\lambda_{\max}^{(2)}$ with $\lambda_{\max}^{(1)} < \lambda_{\max}^{(2)}$. Then, for $\lambda > \lambda_{\max}^{(1)}$, all solutions $\hat{\beta}^{(\lambda)}$ have $\hat{\beta}_{\mathcal{P}}^{(\lambda)} = 0$ and the vector $(0, \hat{\beta}_{\mathcal{U}}^{(2)})$ attains the minimum of P_λ. By the subgradient optimality condition,

$$0 \in \arg\min_{\beta_{\mathcal{P}}} P_\lambda((\beta_{\mathcal{P}}, \hat{\beta}_{\mathcal{U}}^{(2)})) \iff \Omega^D(X_{\mathcal{P}}^T\, J(X_{\mathcal{U}}\hat{\beta}_{\mathcal{U}}^{(2)})) \leqslant \lambda. \tag{30}$$

Choosing $\lambda \in (\lambda_{\max}^{(1)}, \lambda_{\max}^{(2)})$, using (30) we deduce that there exists $\hat{\beta}^{(\lambda)} = (v, \hat{\beta}_{\mathcal{U}}^{(2)})$ with $v \neq 0$. This contradicts the fact that for $\lambda > \lambda_{\max}^{(1)}$ all solutions $\hat{\beta}^{(\lambda)}$ have $\hat{\beta}_{\mathcal{P}}^{(\lambda)} = 0$.

For the final statement, assume f_i are strictly convex. Then f_i' are invertible, hence so too is J. By uniqueness of $\hat{\theta}^{(\lambda)}$ and (25) i), the fitted values $X\hat{\beta}^{(\lambda)}$ are the same for all solutions $\hat{\beta}^{(\lambda)}$. It follows that $\Omega(\hat{\beta}_{\mathcal{P}}^{(\lambda)})$ is the same for all solutions. Since there exists $\hat{\beta}^{(\lambda_{\max})}$ with $\hat{\beta}_{\mathcal{P}}^{(\lambda_{\max})} = 0$ and Ω is a norm, it follows that all solutions $\hat{\beta}^{(\lambda_{\max})}$ have $\hat{\beta}_{\mathcal{P}}^{(\lambda_{\max})} = 0$. □