JOTA manuscript No.
(will be inserted by the editor)

Safe Feature Elimination for Non-Negativity Constrained Convex Optimization

James Folberth · Stephen Becker

Received: date / Accepted: date

Abstract Inspired by recent work on safe feature elimination for 1-norm regularized least-squares, we develop strategies to eliminate features from convex optimization problems with non-negativity constraints. Our strategy is safe in the sense that it will only remove features/coordinates from the problem when they are guaranteed to be zero at a solution. To perform feature elimination we use an accurate, but not optimal, primal-dual feasible pair, making our methods robust and able to be used on ill-conditioned problems. We supplement our feature elimination problem with a method to construct an accurate dual feasible point from an accurate primal feasible point; this allows us to use a first-order method to find an accurate primal feasible point, then use that point to construct an accurate dual feasible point and perform feature elimination. Under reasonable conditions, our feature elimination strategy will eventually eliminate all zero features from the problem. As an application of our methods we show how safe feature elimination can be used to robustly certify the uniqueness of non-negative least-squares problems. We give numerical examples on a well-conditioned synthetic non-negative least-squares problem and on a set of 40000 extremely ill-conditioned problems arising in a microscopy application.

Keywords feature elimination · dimension reduction · duality · NNLS

Mathematics Subject Classification (2000) 49N15 · 90C25 · 90C46

This is a post-peer-review, pre-copyedit version of an article published in Journal of Optimization Theory and Applications. The final authenticated version is available online at: `https://dx.doi.org/10.1007/s10957-019-01612-w`.

1 Introduction

There is an expanding body of work on safe feature elimination for 1-norm regularized optimization problems, particularly for 1-norm regularized least-squares (the lasso). Safe feature elimination removes features/columns of the dictionary/observation matrix when they are *guaranteed* not to be present in a solution. El Ghaoui et al.'s influential work in this direction [1] is based on using complementary slackness between primal and dual optimization problems to identify zero coordinates in a solution to the primal problem. Complementary slackness implies that if the dual optimal point satisfies an inequality constraint strictly, then the corresponding primal optimal coordinate must be equal to zero in any primal optimal point (we will make this statement precise shortly).

James Folberth, Corresponding author
University of Colorado
Boulder, Colorado
james.folberth@colorado.edu

Stephen Becker
University of Colorado at Boulder
Boulder, Colorado
stephen.becker@colorado.edu

Using duality to identify zero coordinates has been used before, of course; for instance duality has been used to eliminate features in linear programs [2]. What is novel is that safe feature elimination (SAFE) strategies are designed to avoid the use of the exact dual optimal point, which may be very expensive to compute in practice. SAFE strategies instead use an auxiliary dual feasible point to construct a compact set that is guaranteed to contain the dual optimal point. If all points in this compact set satisfy a dual inequality constraint strictly, then the exact dual optimal point also satisfies the inequality strictly and we can safely eliminate the corresponding primal coordinate. We refer the reader to [3] (and the extensions and generalizations in [4,5,6]) for a discussion of numerous such safe sets for the lasso and [7] for a survey on both safe and unsafe feature elimination strategies for lasso problems. Safe elimination strategies for support vector machines (SVM) have also been proposed in [8] and [9]. The work of [10] puts forth a more general theory of deriving SAFE rules for a wide class of convex problems that includes the lasso, SVM, and other problems as examples.

In this paper we develop a SAFE strategy for non-negativity constrained convex optimization problems which uses an accurate, but non-optimal, primal-dual feasible pair. This is similar to the SAFE strategy for the lasso proposed in [3], which is more robust than El Ghaoui et al.'s original in [1]. We show that under reasonable conditions, a sufficiently accurate primal-dual pair will eliminate all zero coordinates from the problem.

A recent technique in super-resolution fluorescence microscopy uses many tens of thousands of non-negative least-squares (NNLS) problems to form a super-resolved image. Motivated by these problems we focus our efforts on the case where only an accurate primal feasible point is known, as is usually the case when using first-order methods to solve the primal. To enable the use of SAFE we propose an efficient method to construct an accurate dual feasible point from a given primal feasible point. We also show that the construction depends continuously on the given primal feasible point, meaning that as the primal feasible point converges to an optimal point (e.g., as one iterates a first-order method) so too does the dual, and hence SAFE will eliminate all zero features.

We apply our SAFE strategy to the task of robustly certifying the uniqueness of solutions to NNLS problems. In a small synthetic numerical example we compare our method with an existing uniqueness sufficient condition that relies on a strong assumption on the structure of the data matrix. The strong assumption is that the columns are in general linear position (19), which can be checked only for very small matrices or if the matrix has a generating model of a certain form. In a real-data numerical example of a much larger size we use our SAFE strategy to certify the uniqueness of reconstructed images from a microscopy application. It is infeasible to check if the columns of the data matrix are in general linear position, so the existing uniqueness condition cannot be used. We instead find an approximate solution (a reconstructed image) to the NNLS problem using an efficient gradient method and use our SAFE strategy to confirm that the exact reconstructed image is unique.

Constructing the dual feasible point and performing feature elimination costs about as much as a primal gradient evaluation, which is to say that it is not expensive. Although we do not explore this direction in this work, the inexpensiveness of our SAFE strategy likely allows it to be used to decrease the cost of solving the primal problem with a first-order method, as has already been demonstrated for the lasso [7].

The rest of this paper is organized as follows. In Section 2 we state a general non-negativity constrained primal problem, develop a dual problem, and state KKT optimality conditions. Section 3 derives the general structure of a SAFE strategy using an accurate primal-dual feasible pair. We also give a simple, but effective, instantiation of this strategy. To enable SAFE to work with first-order methods, Section 4 derives and analyses a dual line search that allows us to construct an accurate dual feasible point from an accurate primal feasible point. Section 5 gives the proof that our SAFE strategy eventually eliminates all zero features. Sections 6 and 7 discuss robustly certifying solution uniqueness for NNLS problems.

2 Preliminaries

Let $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\pm\infty\}$ be the set of extended real values and $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geq 0\}$ be the non-negative orthant. We denote the standard inner product on $\mathbb{R}^n$ by both $\langle x, y\rangle$ and $x^T y$, and the induced norm by $\|x\| = \sqrt{x^T x}$. The convex conjugate of a function f is defined via

$$f^*(y) \coloneqq \sup_x \langle y, x\rangle - f(x).$$

We denote the domain of a function f by $\operatorname{dom} f$ and the ith coordinate of a vector x by x_i or $\{x\}_i$. Except for the convex conjugate, we use a superscript $*$ to denote the value of a quantity at an optimum, e.g., p^* for the optimal value of a primal optimization problem.

We consider a general optimization problem involving a convex objective f subject to a non-negativity constraint on the optimization variables:

$$\begin{array}{ll} \min_x & f(Ax) \\ \text{s.t.} & x \geq 0. \end{array} \tag{1}$$

For example, this generic problem structure captures non-negative least-squares (NNLS) with $f(z) = \frac{1}{2}\|z - b\|^2$. We assume the following throughout the paper:

- $f : \mathbb{R}^m \to \mathbb{R}$ is proper and convex,
- f has globally L-Lipschitz continuous gradient with $L > 0$ (f is L-smooth),
- $A \neq 0$ is a real $m \times n$ matrix,
- the primal optimal value is attained.

A sufficient condition for the primal optimal value to be attained is to apply Slater's constraint qualifications to the dual problem (2), and in many cases this simplifies. For example, if f is strongly convex, as is the case when $f(z) = \frac{1}{2}\|z - b\|^2$, then in particular f is supercoercive and hence f^* has full domain [11, Cor. 11.17, Prop. 14.15], thus a strictly feasible dual point is any ν satisfying $A^T\nu > 0$. Such a point is guaranteed in cases such as when the entries of A are positive, as occurs in the microscopy example of Subsection 7.2.

Note that one could generalize to f taking on values in the extended real line, i.e., $\operatorname{dom} f \subsetneq \mathbb{R}^n$, in which case we need to add the assumption that $\operatorname{int} \operatorname{dom} f \cap A\mathbb{R}^n_+$ is non-empty, since this implies Slater's constraint qualification holds for the primal, which is needed for strong duality and for Lemma 4.2.

We begin by deriving a dual problem to (1). The problem (1) fits naturally into the framework of Fenchel-Rockafellar duality (though one can use Lagrange duality to find the same dual problem; see [12]). See the text [13] for a self-contained introduction or [11] for a thorough treatment. We can write (1) as

$$p^* = \min_x f(Ax) + \iota_{\mathbb{R}^n_+}(x)$$

where $\iota_{\mathbb{R}^n_+}$ is the indicator function for the non-negative orthant, i.e. $\iota_{\mathbb{R}^n_+}(x)$ is 0 if $x \geq 0$ and $+\infty$ otherwise.

We can directly write the dual problem as

$$\max_\nu -f^*(\nu) - (\iota_{\mathbb{R}^n_+})^*(-A^T\nu).$$

The conjugate of $\iota_{\mathbb{R}^n_+}$ is readily found to be $\iota_{\mathbb{R}^n_-}$, the indicator for the non-positive orthant. Writing this as a constraint on $A^T\nu$, we have the dual problem

$$\begin{array}{ll} d^* = \max_\nu & g(\nu) \\ \text{s.t.} & A^T\nu \geq 0, \end{array} \tag{2}$$

where we have defined the dual objective $g(\nu) \coloneqq -f^*(\nu)$. Since we assumed f to be proper, f^* is proper and convex. Further, by assuming f has L-Lipschitz continuous gradient, the "conjugate correspondence theorem" (Theorem 5.26 of [13]) implies that $g = -f^*$ is $1/L$-strongly concave. The strong concavity of g implies that the dual optimal point ν^* exists and is unique. It also provides us with the bound

$$\frac{1}{2L}\|\nu - \nu^*\|^2 \leq g(\nu^*) - g(\nu) \quad \forall \nu \in \operatorname{dom} g. \tag{3}$$

We will use this bound as a fundamental building block for our feature elimination procedure in Section 3.

Observe that Slater's condition holds for the primal problem (1). This implies that strong duality holds, so that the primal optimal value p^* and the dual optimal value d^* are equal. Note that Slater's condition holding for the primal problem also shows that the dual optimal value is attained (a ν^* exists that achieves $d^* = g(\nu^*)$) [12]. Furthermore, Slater's condition holds for the dual problem (2), which implies that the primal optimal value is attained and that the KKT conditions are necessary and sufficient for primal and dual optimal points.

The KKT conditions for the primal problem (1) can be written as

$$A^T \nabla f(Ax) - A^T \nu = 0 \tag{4}$$
$$x \geq 0 \tag{5}$$
$$A^T \nu \geq 0 \tag{6}$$
$$x_i \{A^T \nu\}_i = 0 \quad \forall i = 1, \ldots, n. \tag{7}$$

3 Safe Feature Elimination

Let x^* be a (primal) optimal point of (1) and ν^* the (dual) optimal point of (2). Let a_i be the ith column/feature of the matrix A. The complementary slackness condition (7) implies that if $\{A^T \nu^*\}_i > 0$, then $x_i^* = 0$. The key idea of safe feature elimination is that if we can certify that $\{A^T \nu^*\}_i = \langle a_i, \nu^* \rangle > 0$, then we can guarantee that $x_i^* = 0$. This allows us to eliminate the ith column of A, a_i, from the problem with a *guarantee* that it will not be present in a solution. What remains is to robustly determine for each column a_i if $\langle a_i, \nu^* \rangle > 0$ without knowledge of the exact solutions ν^* or x^*.

Observe that we do not require the precise value of $\langle a_i, \nu^* \rangle$; we merely need to certify that $\langle a_i, \nu^* \rangle$ is strictly positive to certify $x_i^* = 0$. This allows us to avoid the apparent need for the exact solution ν^*. Suppose we have a set of dual points N that is guaranteed to contain ν^*. We then find a lower bound for $\langle a_i, \nu^* \rangle$ by solving the "feature elimination subproblem"

$$\begin{array}{ll} \min_\nu & \langle a_i, \nu \rangle \\ \text{s.t.} & \nu \in N. \end{array} \tag{8}$$

That N contains ν^* makes (8) safe. The optimal value of (8) is guaranteed to be no larger than $\langle a_i, \nu^* \rangle$, so that if the optimal value is strictly positive we can certify that $\langle a_i, \nu^* \rangle > 0$. The feature elimination subproblem (8) tests for elimination of the single feature a_i, so to test for feature elimination on all of A we simply solve (8) for each column of A.

We now construct a simple, but effective, search set N without the use of any exact solutions. Let us assume that we have access to both a primal feasible point $\hat{x}$ and a dual feasible point $\hat{\nu}$, neither of which are assumed to be optimal. This gives the duality gap $\epsilon = f(A\hat{x}) - g(\hat{\nu})$. Since strong duality holds, the duality gap ϵ will shrink to zero as $\hat{x}$ and $\hat{\nu}$ become increasingly accurate. Using the strong-concavity bound (3), we have

$$\frac{1}{2L} \|\hat{\nu} - \nu^*\|^2 \leq g(\nu^*) - g(\hat{\nu}). \tag{9}$$

Since strong duality holds, $g(\nu^*) = f(Ax^*)$, from which we see

$$g(\nu^*) - g(\hat{\nu}) = f(Ax^*) - g(\hat{\nu}) \leq f(A\hat{x}) - g(\hat{\nu}) = \epsilon.$$

Combining these gives us a bound on the distance from $\hat{\nu}$ to ν^* in terms of the duality gap ϵ. We therefore define the search set N to be the set of all points satisfying this bound: $N \coloneqq \{\nu : \|\hat{\nu} - \nu\|^2 \leq 2L\epsilon\}$. As desired, N is guaranteed to contain ν^*, but is constructed using only the feasible points $\hat{x}$ and $\hat{\nu}$.

The associated feature elimination subproblem is

$$\begin{array}{ll} \min_\nu & \langle a_i, \nu \rangle \\ \text{s.t.} & \|\nu - \hat{\nu}\|^2 \leq 2L\epsilon. \end{array} \tag{10}$$

The problem (10) has a linear objective and the constraint set is a ball of radius $\sqrt{2L\epsilon}$ centered at $\hat{\nu}$. See Figure 1 for a diagram of the dual geometry for this problem. Taking the search set N to be a ball as we have done is very similar to the GAP SAFE sphere test of [3].

The optimal value is easily found in closed-form to be $\langle a_i, \hat{\nu} \rangle - \sqrt{2L\epsilon}\|a_i\|$. As $\hat{x}$ and $\hat{\nu}$ become more accurate, $\hat{\nu}$ approaches ν^* and ϵ shrinks to zero; as this occurs the optimal value of the subproblem approaches $\langle a_i, \nu^* \rangle$, giving more precise lower bounds and thus increasing the strength of the subproblem to eliminate features. It is therefore crucial to have accurate, feasible $\hat{x}$ and $\hat{\nu}$ in order to apply (10) effectively.

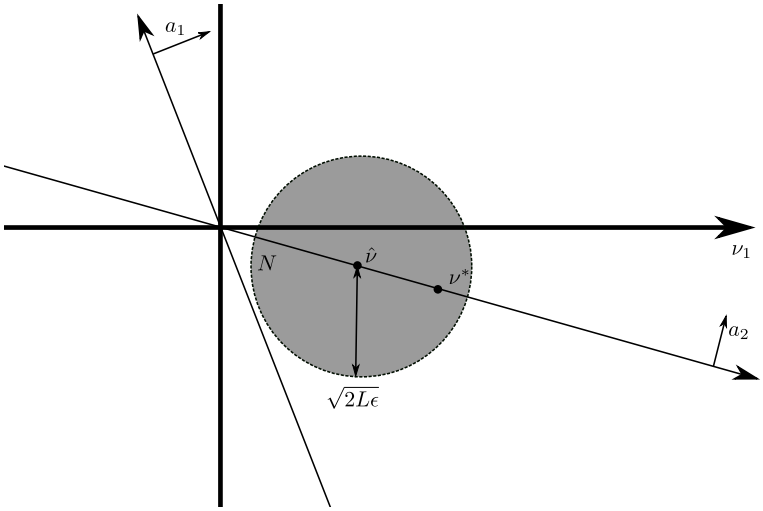

Fig. 1 Dual geometry of the feature elimination subproblem (10). The hyperplanes $\langle a_1, \nu \rangle = 0$ and $\langle a_2, \nu \rangle = 0$ are drawn, with the dual feasible set $\{\nu : A^T\nu \geq 0\}$ extending toward the upper right. The dual optimal point ν^* is guaranteed to be the search set N, which is a ball of radius $\sqrt{2L\epsilon}$ centered at $\hat{\nu}$. Since $\langle a_1, \nu \rangle > 0$ for all $\nu \in N$, the feature elimination subproblem (10) has strictly positive optimal value and so feature a_1 can be eliminated. The figure is drawn such that $\langle a_2, \nu^* \rangle = 0$, so a_2 cannot be eliminated.

4 Dual Line Search

In order to use feature elimination subproblem (10), we must have a primal feasible $\hat{x}$ and a dual feasible $\hat{\nu}$ that achieve a reasonably tight duality gap $\epsilon = f(\hat{x}) - g(\hat{\nu})$. When using a first-order method on the primal we have access to an accurate primal feasible point $\hat{x}$ simply by taking one of the iterates. But we typically do not have access to an accurate dual feasible point $\hat{\nu}$. Hence we derive an inexpensive method to find an accurate dual feasible $\hat{\nu}$ from an accurate primal feasible $\hat{x}$.

4.1 Finding an Accurate Dual Feasible $\hat{\nu}$ From an Accurate Primal Feasible $\hat{x}$

To leverage the accuracy of $\hat{x}$, we form $\nu' = \nabla f(A\hat{x})$, since if $\hat{x}$ were optimal, then $\nabla f(A\hat{x})$ would be the dual optimal point (see Lemma 4.2). But note that ν' is not guaranteed to be dual feasible since $\hat{x}$ is not necessarily optimal (i.e., $A^T\nu' \ngeq 0$ is possible). To fix this, perhaps the "best" approach is to solve the orthogonal projection problem

$$\begin{array}{ll} \min_{\hat{\nu}} & \frac{1}{2}\|\hat{\nu} - \nu'\|^2 \\ \text{s.t.} & A^T\hat{\nu} \geq 0, \end{array} \tag{11}$$

which finds the closest dual feasible point to ν'. This problem is closely related to the dual of NNLS, and the optimal $\hat{\nu}$ can be found by solving an appropriate NNLS primal problem (see (15) and Lemma 4.2). In the context of our microscopy NNLS example, this would mean we must solve an additional NNLS problem each time we would like to attempt feature elimination. With the number of NNLS problems already in the tens of thousands, this approach becomes rather unwieldy, and we can accept a less-than "best" $\hat{\nu}$ in exchange for improved speed. We only need $\hat{\nu}$ that does not spoil the accuracy provided by $\hat{x}$, thereby providing a small duality gap ϵ.

To that end let us assume we have access to a strictly dual feasible point ν^{strict}. We can use ν^{strict} to construct $\hat{\nu}$ nearby ν' that is also dual feasible using a simple line search: we find the closest dual feasible point to ν' along the line segment between ν' and ν^{strict} via

$$\begin{array}{ll} \min & t \\ \text{s.t.} & A^T((1-t)\nu' + t\nu^{\text{strict}}) \geq 0 \\ & 0 \leq t \leq 1. \end{array} \tag{12}$$

Once we have solved the line search for t^*, we form $\hat{\nu} = (1-t^*)\nu' + t^*\nu^{\text{strict}}$. We can view this line search as a not-necessarily-orthogonal projection onto the dual feasible set. We will give a few simple

methods to find a strictly dual feasible ν^{strict} in Subsection 4.1.1, and in Subsection 4.1.2 we will show that ν^{strict} being strictly dual feasible (instead of just dual feasible) is necessary for $\hat{\nu}$ from the line search to converge to ν^* as $\hat{x}$ converges to x^*.

See Figure 2 for a diagram of this line search in two dimensions. The boundary of the dual feasible set is given by two hyperplanes $\langle a_1, \nu\rangle = 0$ and $\langle a_2, \nu\rangle = 0$. We can see $\langle a_2, \nu'\rangle < 0$, so ν' is not dual feasible.

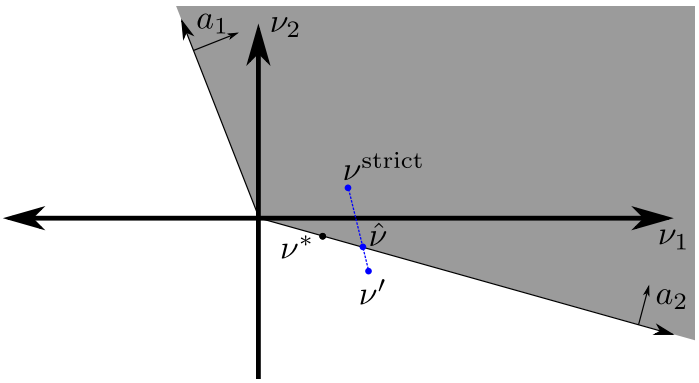

Fig. 2 Finding $\hat{\nu}$ from ν' and ν^{strict} via the dual line search (12).

The constraint $t \geq 0$ is used in the line search only so that $\hat{\nu} = \nu'$ in the case when ν' is already dual feasible. Additionally, the optimal value is never greater than 1, since the point ν^{strict} is assumed to be dual feasible. By precomputing $A^T\nu'$ and $A^T\nu^{\text{strict}}$, the optimal value of the line search and the resulting dual feasible point $\hat{\nu}$ can be found in closed-form, which is given in Subsection 4.1.2.

4.1.1 Finding a Strictly Dual Feasible ν^{strict}

For a given A, we can search for a strictly dual feasible point ν^{strict} via the optimization problem

$$\begin{array}{ll} \max_{\nu,t} & t \\ \text{s.t.} & A^T\nu \geq t \\ & \|\nu\| \leq 1 \end{array} \tag{13}$$

This problem maximizes a lower bound of $A^T\nu$ while the constraint $\|\nu\| \leq 1$ serves to keep ν bounded. If we take $\|\cdot\|$ to be the ℓ_1-norm, the program is an LP; if we take $\|\cdot\|$ to be the ℓ_2-norm, the program is a quadratically constrained quadratic program. If the optimal value $t^* > 0$, then (13) will have found a strictly dual feasible ν^{strict}. Conversely, if ν^{strict} is strictly dual feasible, then $\nu^{\text{strict}}/\|\nu^{\text{strict}}\| \leq 1$ and the optimal value $t^* \geq A^T\nu^{\text{strict}}/\|\nu^{\text{strict}}\| > 0$. Hence if there exists a strictly dual feasible point for a given A the program (13) will find one. The cost of this problem is not much of a concern (unless A is huge), as it only needs to be solved once to find ν^{strict}; once we have a strictly dual feasible ν^{strict}, we can use it for any primal problem of the form (1) with the same A.

There are other methods to find a suitable ν^{strict} without solving (13). If the sum of each row of A is positive, the point $\nu^{\text{strict}} = \mathbb{1}$ is strictly dual feasible. In particular, if A is elementwise positive (as is the case in our microscopy example in Subsection 7.2), $\nu^{\text{strict}} = \mathbb{1}$ is strictly dual feasible. We can also take $\nu^{\text{strict}} = \max\{0, \nu'\}$, where $\nu' = \nabla f(A\hat{x})$; if ν^{strict} has at least one positive entry, then it is strictly dual feasible. In our microscopy example A is elementwise positive and we find that using $\nu^{\text{strict}} = \max\{0, \nu'\}$ reliably produces strictly dual feasible points (and avoids the need for solving the program (13)).

4.1.2 The Dual Line Search is a Continuous Mapping

In Subsection 4.2 we will show how $\hat{\nu}$ from the dual line search converges to the dual optimal point ν^* as $\hat{x}$ converges to a primal optimal point. This will then be used to show that, under reasonable conditions, our dual line search and feature elimination strategy will eventually eliminate all zero

features from the problem. This means that if we perform sufficiently many iterations of a first-order method, we can eliminate all zero features from the problem. To enable that analysis, we find a closed-form solution to the line search and prove a lemma on the continuity of the mapping from $\nu' = \nabla f(A\hat{x})$ to $\hat{\nu}$ found via the line search.

We find the closed-form solution to the line search by identifying two cases:

1. If ν' is dual feasible, $t = 0$ is the minimum feasible value, which leads to $\hat{\nu} = \nu'$.
2. Otherwise, there is at least one index i such that $\langle a_i, \nu' \rangle = \{A^T\nu'\}_i < 0$. In this case, we must increase t until the all coordinates of $A^T((1-t)\nu' + t\nu^{\text{strict}})$ are non-negative.

We define the scalar-valued function

$$t(\lambda; \lambda_0) \coloneqq \begin{cases} 0 & \lambda \geq 0 \\ \dfrac{\lambda}{\lambda - \lambda_0} & \lambda < 0, \end{cases}$$

where λ is the scalar independent variable and λ_0 is a fixed parameter. We can write the dual feasible point returned from the line search as $\hat{\nu} = (1-t^*)\nu' + t^*\nu^{\text{strict}}$ where

$$t^* = \max_i \; t(a_i^T\nu'; a_i^T\nu^{\text{strict}}).$$

Lemma 4.1 *If ν^{strict} is strictly dual feasible (i.e., $A^T\nu^{strict} > 0$), then the dual line search* (12) *mapping ν' to $\hat{\nu}$ is continuous in ν'.*

Proof The strict dual feasibility assumption states that $a_i^T\nu^{\text{strict}} > 0$ for each i. The dual line search produces the point

$$\hat{\nu} = (1-t^*)\nu' + t^*\nu^{\text{strict}}.$$

To show continuity of the mapping $\nu' \mapsto \hat{\nu}$, it is sufficient to show t^* is continuous in ν'.

Observe that if $\lambda_0 > 0$ the function $t(\lambda; \lambda_0)$ is continuous for all λ. Since we take ν^{strict} strictly dual feasible, $a_i^T\nu^{\text{strict}} > 0$ for each i, meaning that $t(a_i^T\nu'; a_i^T\nu^{\text{strict}})$ depends continuously on ν' for each i. Since t^* is the pointwise maximum of continuous functions of ν', it is continuous in ν', completing the proof. □

The strict dual feasibility assumption in Lemma 4.1 is necessary for the continuity of the mapping. Let us look at the dual geometry when ν^{strict} is not strictly dual feasible. The non-strictly dual feasible point ν^{strict} is on the boundary of the dual feasible set, since it satisfies $a_i^T\nu^{\text{strict}} = 0$ for some a_i. When ν' is not dual feasible, the only dual feasible point on the line segment between ν' and ν^{strict} is ν^{strict}, so the dual line search returns $\hat{\nu} = \nu^{\text{strict}}$. Thus, when ν^{strict} is not strictly dual feasible, the dual line search will return one of two values: when ν' is not feasible, the line search returns ν^{strict}; when ν' is feasible, the line search returns ν'.

Figure 3 illustrates what would happen if ν' converges to ν^* while remaining dual infeasible. The line search is "stuck", returning $\hat{\nu} = \nu^{\text{strict}}$ for all ν'. So as $\nu' \to \nu^*$, $\hat{\nu} = \nu^{\text{strict}}$ is "stuck" and does not converge to ν^*. Picking ν^{strict} to be strictly dual feasible "unsticks" the dual line search, allowing the returned value $\hat{\nu}$ to converge to ν^* as $\nu' \to \nu^*$.

4.2 Convergence of Dual Sequence Given Primal Sequence

Using a first-order method to solve (1) typically provides a sequence x^k of primal feasible points that converges to a primal optimal point x^* (which may not be unique). For example, the projected gradient method produces such a sequence (see Theorem 10.24 of [13] for a proof of this). In Subsection 4.1, we discussed a dual line search that allows us to produce a dual feasible point $\hat{\nu}^k$ from $\nu'^k \coloneqq \nabla f(Ax^k)$ and a separate dual feasible point ν^{strict}.

In Subsection 3 we saw a simple feature elimination subproblem (10) that utilized an accurate, but not necessarily optimal, primal-dual pair. The strength of the subproblem depends on the size of the duality gap ϵ. In other words, as the duality gap shrinks, the lower bound produced by the subproblem increases, possibly eliminating the feature. In order for the duality gap ϵ to shrink to zero as x^k converges, we must also have that the dual line search produces $\hat{\nu}^k$ that converges to the dual optimal point ν^* (recall that strong duality holds for (1) and (2)). If ν^{strict} is strictly dual feasible, we will see that $\hat{\nu}^k \to \nu^*$ as $x^k \to x^*$, thus giving $\epsilon \to 0$ as desired.

First, we give a lemma that states $\nu^* = \nabla f(Ax^*)$ for any primal optimal x^*. This comes somewhat directly from the KKT conditions for the dual problem.

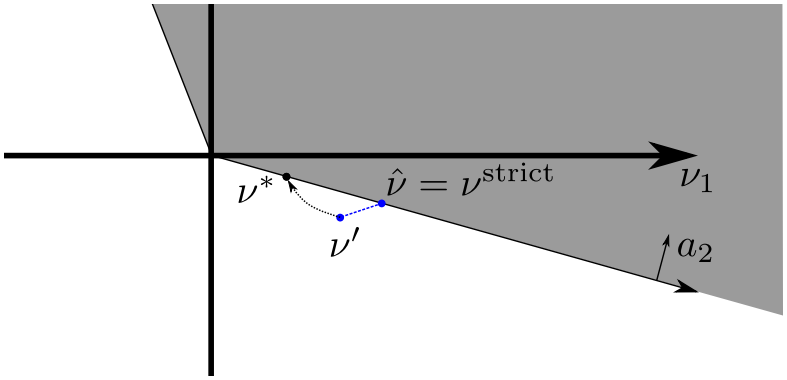

Fig. 3 Pathological case for the continuity of the dual line search when ν^{strict} is on the boundary of the dual feasible set (i.e., ν^{strict} is not strictly dual feasible).

Lemma 4.2 *Let x^* be a primal optimal point of* (1)*. Then $\nu = \nabla f(Ax^*)$ is the unique dual optimal point of* (2)*.*

Proof This is a known result coming from optimality conditions for the dual problem. For example, this is a consequence of Theorem 19.1 of [11]. □

Now we show that the dual line search produces a sequence of points that converges to the dual optimal point.

Theorem 4.1 *Let x^k be a sequence of primal feasible points that converge to a primal optimal point x^*, and let ν^{strict} be a strictly dual feasible point (i.e., $A^T\nu^{strict} > 0$). For each k, define the dual feasible point $\hat{\nu}^k$ by performing the dual line search* (12) *using $\nu'^k = \nabla f(Ax^k)$ and ν^{strict}. Then the sequence $\hat{\nu}^k$ of dual feasible points converges to the unique dual optimal point ν^*.*

Proof Note that the map $x^k \mapsto \nu'^k = \nabla f(Ax^k)$ is continuous by assumption. By continuity and Lemma 4.2, $\nu'^k \to \nu^*$ as $x^k \to x^*$. But ν'^k is not guaranteed to be dual feasible, hence our use of the dual line search. We seek to show that the dual feasible sequence $\hat{\nu}^k \to \nu^*$ as $x^k \to x^*$. By the triangle inequality,

$$\|\hat{\nu}^k - \nu^*\| \le \|\hat{\nu}^k - \nu'^k\| + \|\nu'^k - \nu^*\|.$$

We already have that $\|\nu'^k - \nu^*\| \to 0$, so to complete the proof it remains to show $\|\hat{\nu}^k - \nu'^k\| \to 0$.

Since $\hat{\nu}^k$ is computed using the dual line search using ν'^k and ν^{strict}, we have that

$$\|\hat{\nu}^k - \nu'^k\| = \|(1-t^k)\nu'^k + t^k\nu^{\text{strict}} - \nu'^k\| = t^k\|\nu'^k - \nu^{\text{strict}}\|,$$

where t^k is the optimal value of t from the dual line search for that particular ν'^k. Since $\nu'^k \to \nu^*$ and ν^* is unique, we know that $\|\nu'^k - \nu^{\text{strict}}\|$ is eventually bounded above by a constant. Therefore, we just need to show $t^k \to 0$ to imply that $\|\hat{\nu}^k - \nu'^k\| \to 0$.

Recalling the proof of Lemma 4.1, the function

$$\nu' \mapsto t(a_i^T\nu'; a_i^T\nu^{\text{strict}}) = \begin{cases} 0 & a_i^T\nu' \ge 0 \\ \dfrac{a_i^T\nu'}{a_i^T\nu' - a_i^T\nu^{\text{strict}}} & a_i^T\nu' < 0, \end{cases}$$

is continuous precisely when $a_i^T\nu^{\text{strict}} > 0$ (i.e., when ν^{strict} is strictly dual feasible). Since ν'^k converges to a dual feasible point, $a_i^T\nu'^k$ converges to a nonnegative value for each a_i. Therefore the limiting value of $t(a_i^T\nu'^k; a_i^T\nu^{\text{strict}})$ is 0 for each a_i, and thus $t^k = \max_i t(a_i^T\nu'^k; a_i^T\nu^{\text{strict}}) \to 0$. This then shows that $\|\hat{\nu}^k - \nu'^k\| \to 0$, which, using the triangle inequality above, implies that $\hat{\nu}^k \to \nu^*$ as desired. □

5 When Will the Screening Rule Eliminate all Zero Features?

Here we show that the feature elimination subproblem (10) will eliminate all zero features from the solution, under reasonable assumptions and with sufficiently small duality gap ϵ. Coupled with the use of a first-order method and the dual line search of Subsection 4.1, this means that our feature elimination strategy will eventually eliminate all features that can be eliminated.

Theorem 5.1 *Assume that strict complementary slackness holds (i.e., $x_i^* = 0$ iff $\langle a_i, \nu^* \rangle > 0$). Define $\mathcal{I} \coloneqq \{i : \langle a_i, \nu^* \rangle > 0\}$ to be the index set of zero features. Let the pair $\hat{x}, \hat{\nu}$ produce a duality gap estimate ϵ such that*

$$\sqrt{\epsilon} < \frac{1}{\sqrt{2L}} \min_{i \in \mathcal{I}} \frac{\langle a_i, \hat{\nu} \rangle}{\|a_i\|}.$$

Then the feature elimination problem (10) *will eliminate all zero features.*

Proof Since we have assumed strict complementary slackness, a zero feature $x_i^* = 0$ always corresponds to $\langle a_i, \nu^* \rangle > 0$. Since ν^* is unique, the set $\mathcal{I}$ uniquely determines the indexes of zero features. If we did not assume strict complementary slackness, then a zero feature $x_i^* = 0$ may be associated with $\langle a_i, \nu^* \rangle = 0$, which cannot be eliminated by the subproblem (10). So we see that strict complementary slackness implies that each zero feature corresponds to a strictly satisfied dual inequality. We must show that the subproblem (10) produces a strictly positive lower bound for $\langle a_i, \nu^* \rangle$ for every $i \in \mathcal{I}$.

For the subproblem to produce a strictly positive lower bound for the ith feature, the search set N must be contained strictly in the interior of the halfspace $\langle a_i, \nu \rangle \geq 0$. The search set N is a (closed) ball of radius $\sqrt{2L\epsilon}$ centered at $\hat{\nu}$. Recall from Section 3 that the optimal value of (10) is $\langle a_i, \hat{\nu} \rangle - \sqrt{2L\epsilon}\|a_i\|$ for the ith feature. Up to a scaling factor, this is the minimum distance between the search set N and the hyperplane $\langle a_i, \nu \rangle = 0$. Therefore the search set N is strictly separated from the hyperplane precisely when the optimal value is strictly positive:

$$\langle a_i, \hat{\nu} \rangle - \sqrt{2L\epsilon}\|a_i\| > 0 \iff \sqrt{\epsilon} < \frac{1}{\sqrt{2L}} \frac{\langle a_i, \hat{\nu} \rangle}{\|a_i\|}.$$

By assumption on the size of ϵ, this condition is satisfied for each $i \in \mathcal{I}$, so feature elimination will eliminate all zero features. □

Theorem 5.1 shows that under strict complementarity and if the duality gap ϵ is sufficiently small, then feature elimination will eliminate all zero features from the problem. We used knowledge of the the exact dual optimal point only to assist in quantifying how small ϵ must be in order to imply that feature elimination will work. But even without knowledge of the dual optimal point, we still know that if ϵ is sufficiently small, then feature elimination will have worked. Indeed, by combining Theorems 4.1 and 5.1, we have the following:

Corollary 5.1 *Assume that strict complementary slackness holds (i.e., $x_i^* = 0$ iff $\langle a_i, \nu^* \rangle > 0$). Let x^k be a sequence of primal feasible points that converges to x^* (e.g., from a first-order method) and let $\hat{\nu}^k$ be the sequence of dual feasible points produced as in Theorem 4.1. Then the duality gap $\epsilon = f(x^k) - g(\hat{\nu}^k) \to 0$ as $k \to \infty$ and Theorem 5.1 will eventually apply. This means that the feature elimination subproblem* (10) *will eventually eliminate all zero features.*

This tells us that if we do enough iterations of a first-order method, perform the dual line search, and then do feature elimination, we will eliminate all possible features. But we don't know how many iterations are sufficient (without knowledge of the dual optimal point, that is). Furthermore, if strict complementarity does not hold, we can only eliminate zero features that correspond to $\langle a_i, \nu^* \rangle > 0$; a zero feature that corresponds to $\langle a_i, \nu^* \rangle = 0$ cannot be eliminated. These issues notwithstanding, we can still use feature elimination very effectively in practice, including certifying that underdetermined NNLS problems have unique solutions.

6 Certifying NNLS Solution Uniqueness

Here we consider applying safe feature elimination to the problem of certifying the uniqueness of the primal solution. We consider the case of NNLS, where $f(z) = \frac{1}{2}\|z - b\|^2$, which reduces the primal problem (1) to

$$\begin{array}{ll} \min_x & \frac{1}{2}\|Ax - b\|^2 \\ \text{s.t.} & x \geq 0. \end{array} \tag{14}$$

We have assumed throughout that the $m \times n$ matrix A is full-rank. But we have not yet assumed anything about the shape of A, which may be "overdetermined" ($m \geq n$) or "underdetermined" ($m < n$). In the overdetermined case, the primal objective is $\sigma_{\min}(A^T A)$-strongly convex, where

$\sigma_{\min}(A^TA)$ is the minimum singular value of A^TA. Since A is overdetermined, the Hessian A^TA of the NNLS problem is non-singular, so $\sigma_{\min}(A^TA) > 0$. The NNLS primal problem (14) therefore has a unique optimal point [13]. In the underdetermined case the Hessian A^TA is singular and the primal objective is neither strongly nor strictly convex, so there is no such uniqueness guarantee.

To attempt to certify the uniqueness of solutions to underdetermined problems we use our feature elimination strategy to reduce the problem to an overdetermined, full-rank NNLS problem. Suppose we eliminate r features/columns of A. This allows us to form the reduced matrix A_{red} with those r columns removed and with the guarantee that the removed columns are not used by a solution of the original problem. If $r \geq n - m$, so that the reduced matrix A_{red} is overdetermined, and if A_{red} is full-rank, then the reduced NNLS problem has a strongly convex objective and has a unique solution. Since our feature elimination strategy is safe, the solution to the original NNLS problem is guaranteed to be the same as the solution to the reduced problem (with appropriate zero padding), meaning that the original NNLS problem has a unique solution. Thus we have a procedure to robustly certify the uniqueness of NNLS problems via our safe feature elimination strategy, which requires an accurate, but not optimal, primal-dual pair.

Note that to certify uniqueness we need not eliminate all zero features, as was the goal of Theorem 5.1. In a sense, certifying uniqueness is an easier problem than eliminating all zero features; indeed, to certify uniqueness we need only eliminate sufficiently many ($r \geq n - m$) features. We therefore do not require (full) strict complementary slackness, as was assumed in Theorem 5.1. There may be some zero features $x_i^* = 0$ paired with $\langle a_i, \nu^* \rangle = 0$ but that does not concern us as long as there are sufficiently many $x_i^* = 0$ such that $\langle a_i, \nu^* \rangle > 0$, enabling us to certify the uniqueness of x^*.

The above uniqueness certification procedure was described just for the least-squares objective (i.e., for NNLS problems), but it can be generalized to the problem $\min_{x \geq 0} f(Ax)$ where f is strictly convex and satisfies the assumptions of SAFE (i.e., proper, L-smooth). When A is underdetermined the objective $f(Ax)$ is no longer strictly convex. But if we use SAFE to eliminate sufficiently many features such that the reduced matrix is overdetermined and full-rank, the reduced objective is strictly convex and therefore a minimizer is unique. One can also modify this uniqueness certification technique to work with ℓ_1-regularized problems (like lasso), for instance using the GAP SAFE rules of [3], which are similar to (10).

6.1 A Small NNLS Example

Let us illustrate our procedure with a small example. For NNLS, the dual problem (2) reduces to

$$\begin{array}{ll} \max_\nu & g(\nu) = -\frac{1}{2}\|\nu + b\|^2 + \frac{1}{2}\|b\|^2 \\ \text{s.t.} & A^T\nu \geq 0. \end{array} \tag{15}$$

Suppose we have a primal feasible point $\hat{x}$ and dual feasible point $\hat{\nu}$. This allows us to compute the duality gap $\epsilon = f(A\hat{x}) - g(\hat{\nu})$. The basic feature elimination subproblem (10) reduces to

$$\begin{array}{ll} \min_\nu & \langle a_i, \nu \rangle \\ \text{s.t.} & \|\nu - \hat{\nu}\|^2 \leq 2\epsilon, \end{array} \tag{16}$$

where we have used that the dual objective g is 1-strongly convex (since f has 1-Lipschitz continuous gradient).

Consider the following matrix with randomly chosen entries

$$A = \begin{bmatrix} 1 & 6 & -1 & 8 & 0 \\ -2 & 7 & 1 & 8 & 2 \\ 3 & 1 & 4 & 1 & -5 \end{bmatrix},$$

and right-hand side (RHS) $b = \begin{bmatrix} -1 & 2 & 1 \end{bmatrix}^T$. We note that there is nothing special about these entries; the entries are the first few digits of the golden ratio, the base of the natural logarithm, and pi, with some negative signs added. Projected gradient descent (PGD) for NNLS produces the iteration

$$x^+ \leftarrow x - tA^T(Ax - b) \tag{17}$$

where we pick step size $t = 1/\|A\|^2$. Starting with $x = 0$ and iterating 250 times yields the primal feasible point $\hat{x} \doteq \begin{bmatrix}0 & 0 & 0.9282 & 0 & 0.5409\end{bmatrix}^T$.[1]

We now find a dual feasible point $\hat{\nu}$ with the dual line search of Section 4. First we find a strictly dual feasible point ν^{strict} via the program (13). Because it makes the numbers more presentable on paper, we opt to rescale the solution ν^{strict} so that $\|\nu^{\text{strict}}\|_1 = 1$, which gives $\nu^{\text{strict}} \doteq \begin{bmatrix}0.56 & 0.34 & 0.1\end{bmatrix}^T$. Then we perform the dual line search (12) with ν^{strict} and $\nu' = A\hat{x} - b$, giving us the dual feasible point $\hat{\nu} \doteq \begin{bmatrix}0.1387 & 0.0552 & 0.0209\end{bmatrix}^T$. Together $\hat{x}$ and $\hat{\nu}$ produce the duality gap $\epsilon \doteq 0.0069$.

If we instead found $\hat{\nu}$ via the orthogonal projection subproblem (11), we would find the improved duality gap $\epsilon \doteq 0.0013$. But recall that the orthogonal projection subproblem is closely related to the NNLS dual problem and is computationally expensive to solve. Avoiding this expense is precisely the motivation for the dual line search, and we see for this example that the dual line search is not terribly worse than the orthogonal projection.

We are now ready to solve the feature elimination subproblem (16) once for each of the five columns of A. Using the closed-form solution given in Section 3, we find the following lower bounds on $A^T\nu^*$: $\begin{bmatrix}-0.34 & 0.17 & -0.49 & 0.26 & -0.61\end{bmatrix}^T$. The lower bounds for $\langle a_2, \nu^*\rangle$ and $\langle a_4, \nu^*\rangle$ are strictly positive, so we can eliminate them from the problem; the lower bounds for the remaining columns are non-positive, so the test is inconclusive. The reduced matrix is

$$A_{\text{red}} = \begin{bmatrix} 1 & -1 & 0 \\ -2 & 1 & 2 \\ 3 & 4 & -5 \end{bmatrix},$$

which is overdetermined and full-rank. We can therefore certify that the original NNLS problem has a unique solution. In fact, only 206 iterations of PGD are required to certify uniqueness, though this is unknown *a priori*.

With the solution certified to be unique, we can bound the distance from $\hat{x}$ to the unique solution x^* via the strong convexity of the reduced primal problem:

$$\|\hat{x} - x^*\|^2 \leq \frac{2}{\sigma_{\min}(A_{\text{red}})^2}\left(f(A\hat{x}) - f(Ax^*)\right) \leq \frac{2}{\sigma_{\min}(A_{\text{red}})^2}\epsilon \doteq 0.066. \tag{18}$$

6.2 An Alternative Method to Certify Uniqueness

Slawski and Hein, as part of their analysis of NNLS problems in [14], prove a lemma on the uniqueness of NNLS solutions. Their result is very similar to existing results for ℓ_1-regularized least-squares and related problems [15,16]. The lemma relies on a strong assumption on the columns of A, but provides a simple condition to certify the uniqueness of a solution. We discuss this condition first, state their lemma, and finally discuss how to use their lemma to certify uniqueness in practice.

For an index set $\mathcal{J} \subseteq \{1, \ldots, n\}$, we denote by $A_{\mathcal{J}}$ the submatrix of A formed by taking column j for $j \in \mathcal{J}$. The columns of the matrix $A \in \mathbb{R}^{m\times n}$ are said to be in general linear position (GLP) in $\mathbb{R}^m$ if the following condition holds:

$$\forall \mathcal{J} \subseteq \{1, \ldots, n\},\ |\mathcal{J}| = \min\{m, n\},\ \forall x \in \mathbb{R}^{|\mathcal{J}|}, A_{\mathcal{J}}x = 0 \implies x = 0. \tag{19}$$

In other words, every subset of $\min\{m, n\}$ columns is linearly-independent. For brevity, we will say "A is in GLP" to mean "the columns of A are in GLP". It is easy to see that A in GLP implies that A is full-rank, but the converse is not true: GLP is strictly stronger than full-rank. A being in GLP is also related to the spark of A, where $\text{spark}(A)$ is defined in [17] to be minimum number of columns that form a linearly dependent set. If $m < n$, then A is in GLP iff $\text{spark}(A) = m + 1$.

Unlike computing the rank of a matrix, computing the spark of A and determining if A is in GLP may be prohibitively difficult in the worst case. The straightforward computation to determine if A is in GLP requires computing combinatorially many determinants. Indeed, determining if A is in GLP (equivalently, if $\text{spark}(A) = m + 1$) is coNP-complete [18,19]; computing $\text{spark}(A)$ is NP-hard in general [19]. So numerically verifying that A is in GLP is likely intractable except for very small A.

But these are worst-case results, when we know nothing about the matrix A; there are matrices that are known to be in GLP or have known spark. For example, if the entries of $A \in \mathbb{R}^{m\times n}$ are

[1] We use $\doteq$ to denote equality up the number of digits shown.

drawn i.i.d. from an absolutely continuous distribution, then A is in GLP with probability one [15]. Though it is complex, another example is $A = \begin{bmatrix} I_n & F_n \end{bmatrix}$ where I_n is the $n \times n$ identity matrix and F_n is the $n \times n$ discrete Fourier transform matrix. When n is a perfect square, the spark is known to be exactly $2\sqrt{n}$, and hence it is not in GLP for $n > 1$ [17].

There are also lower bounds for $\operatorname{spark}(A)$ [17,20]. One such bound is $\operatorname{spark}(A) > 1/\mu(A)$, where

$$\mu(A) = \max_{i \neq j} \frac{|\langle a_i, a_j \rangle|}{\|a_i\| \|a_j\|}$$

is called the coherence parameter of A. For the $A \in \mathbb{R}^{1681 \times 2822}$ from our microscopy example in Subsection 7.2, we have $\mu(A) \approx 0.99$ which gives the uninformative bound $\operatorname{spark}(A) \geq 2$.

Assuming we know that A is in GLP (e.g., if A is drawn with entries from a continuous distribution like the standard normal distribution), the following lemma from [14] gives a simple condition implying the uniqueness of the NNLS solution.

Lemma 6.1 (Lemma 5 from [14]) *Let the columns of $A \in \mathbb{R}^{m \times n}$, $m < n$, be in GLP. If the NNLS optimal value is strictly positive,*

$$p^* = \min_{x \geq 0} \frac{1}{2} \|Ax - b\|^2 > 0,$$

then the NNLS problem has a unique solution. Furthermore there are at most $m - 1$ non-zero values in the solution.

For underdetermined NNLS problems with A in GLP, we can certify uniqueness simply by certifying $p^* > 0$. Assuming we know that A is in GLP, this is simple to check and certify in practice, including when using a first-order method that produces only primal points. We can produce a dual feasible point $\hat{\nu}$ from a primal feasible point $\hat{x}$ using the dual line search from Section 4. If we have that $g(\hat{\nu}) > 0$, then $p^* > 0$ by weak duality and the solution is certified to be unique. But of course if A is not known to be in GLP we cannot invoke Lemma 6.1.

The small example problem in the previous subsection has A in GLP, which can be checked directly since it is so small. It takes 286 iterations of PGD to certify that $p^* > 0$, which is slightly more than the 206 iterations needed for SAFE to certify uniqueness.

To certify uniqueness using safe feature elimination, there must be at least $n - m$ zero features with strict complementarity. If this condition does not hold, safe feature elimination will never certify uniqueness. If A is in GLP and $p^* > 0$, then one can certify uniqueness using Lemma 6.1. But notice from Lemma 6.1 that under such conditions, there are at least $n - m + 1$ zero features in the solution; so safe feature elimination will also certify uniqueness, provided the solution exhibits enough strict complementarity. Even if feature elimination fails to certify uniqueness, it still provides certificates that features are not present in the solution. This is a positive result, whereas Lemma 6.1 provides no additional benefit when it fails to certify uniqueness.

7 Certifying NNLS Solution Uniqueness - Examples

7.1 Synthetic Data Examples

Let us now see how safe feature elimination performs on a larger synthetic example. We construct a random NNLS problem by drawing a random 50×100 matrix A and 50×1 RHS b each with entries drawn i.i.d. from the standard normal distribution $\mathcal{N}(0, 1)$. Such an NNLS problem may not necessarily have a unique solution. To check if it does we find a high-accuracy solution using MATLAB's `lsqnonneg`, which implements an active set method from [21]. Using the numerically optimal solution and noting that A is in GLP with probability one, we check if $p^* > 0$ to certify that the solution is unique. If it not certified to be unique, we draw another random NNLS problem until we have a problem with a unique solution.

Although it is outside the scope of the present work, it is interesting to note that the uniqueness of the solution to random NNLS problems of this form appears to depend sharply on the shape of A. If $m > n/2$, the solution appears to be unique with high probability for large m, n; if $m < n/2$, the solution appears to be non-unique with high probability for large m, n. "Phase transitions" of a

similar form are analyzed in [22] and it seems quite possible to extend their results to random NNLS problems of the form used here.

In preparation for the dual line search, we find a strictly dual feasible point ν^{strict} by solving the program (13) once and precomputing $A^T\nu^{\text{strict}}$. We run 7500 iterations of projected gradient descent starting with $x = 0$, with each iteration giving a primal feasible point $\hat{x}$. At each iteration we use the dual line search (12) to construct a dual feasible point $\hat{\nu}$. Performing the dual line search requires computing $A^T\hat{\nu}$ plus $\mathcal{O}(n)$ work, which is on the order of a single gradient evaluation. Using $\hat{x}$ and $\hat{\nu}$, we find the duality gap and use the feature elimination subproblem (16) to eliminate features.

Figure 4 shows the result of using feature elimination to certify the uniqueness of the random NNLS problem. After about 1500 iterations, the duality gap is small enough that feature elimination has started to eliminate features. Just after 3000 iterations, sufficiently many features are eliminated to certify uniqueness (A in GLP implies that the reduced matrix A_{red} is full-rank, but we can also verify this numerically). In accordance with Corollary 5.1 we find that SAFE eventually eliminates all zero features.

For comparison, we also use SAFE with the orthogonal projection (11) instead of the dual line search. Solving the orthogonal projection subproblem at each step of projected gradient descent is tractable for this small problem, but is impractical for larger problems. We see that the dual line search, which scales well to large problems, performs only a bit worse than the orthogonal projection.

Figure 4 also shows that Lemma 6.1 certifies uniqueness for quite a large duality gap. For the problem used for Figure 4, ν^{strict} is sufficient to certify $p^* > 0$, which certifies uniqueness before even the first iteration. While impressive, this is not "for free" since we still solve the program (13) to find ν^{strict}. For other instances more PGD iterations are required, but it is typical for these problems that $p^* > 0$ is certified before SAFE has eliminated sufficiently many features. After SAFE has certified uniqueness, the reduced primal problem is strongly convex, which allows us to bound the distance from the primal iterate $\hat{x}$ to the true solution x^* *à la* (18). Though not always possible (unlike a bound on the duality gap, which we can always find), this provides quite strong information relating the iterate $\hat{x}$ to the optimal point x^*.

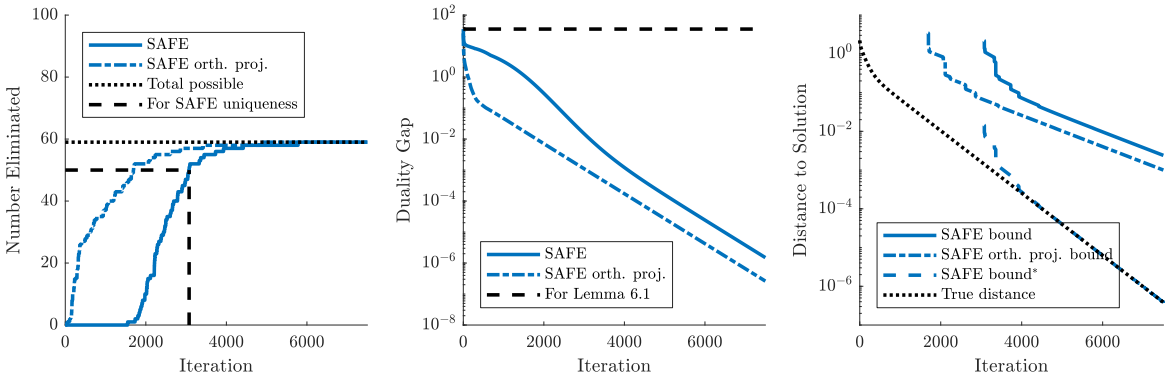

Fig. 4 Certifying uniqueness for a synthetic 50×100 NNLS problem. The dashed line in the left figure shows the minimum number of eliminated features to certify uniqueness; the dotted line shows the maximum number of features that can be eliminated; the dash-dot line shows SAFE using the orthogonal projection (11) instead of the dual line search. In the middle figure, the dashed line shows the duality gap when $p^* > 0$ is certified and Lemma 6.1 can be invoked. The right figure shows the bound (18) on the distance from $\hat{x}$ to the optimal point x^*. The line labeled SAFE bound* uses $f(A\hat{x}) - f(Ax^*)$ in place of the duality gap $\epsilon = f(A\hat{x}) - g(\hat{\nu})$ in the bound (18) (i.e., uses only the first inequality of (18)). Though this requires knowledge of x^*, this shows that the slower convergence of the bound (18) (which we can compute without knowledge of x^* or ν^*) is due to the suboptimality of $\hat{\nu}$.

7.2 Microscopy Uniqueness Example

We now consider a challenging set of NNLS problems arising from a new technique in super-resolution fluorescence microscopy. In this instance, the image formation process involves solving 40000 NNLS problems each using the same matrix A but different RHS b. The solutions of the NNLS problems

are then assembled into the final image. We refer the reader to [23] for details on the microscope and NNLS problem setup.

It is natural to ask if the final, super-resolved image is uniquely determined given the data. If each of the 40000 NNLS problems has a unique solution, then the final image is unique. We answer that question in the affirmative by using feature elimination to certify the uniqueness of each NNLS problem. In fact these microscopy problems motivated our development of feature elimination, in particular developing them to work with just a primal feasible point $\hat{x}$ coming from a first-order method. Note that we do not know *a priori*, and cannot verify numerically, if A is in GLP, so we cannot use Lemma 6.1.

The matrix A is 1681×2822 and has 2-norm condition number $\kappa_2(A) = 2.4 \times 10^{20}$ (computed using `dgesvj` compiled to use quadruple precision [24]). Even though these NNLS problems are extremely ill-conditioned, first-order methods are well-suited to solve them since each NNLS problem uses the same A. This structure allows us to combine gradient computations for many RHS into matrix-matrix products with A and A^T, instead of repeated matrix-vector products with A and A^T. High-performance matrix-matrix product implementations take advantage of modern hierarchical memory computers to achieve higher performance than repeated matrix-vector products [25, 26]. The result is an order of magnitude speedup in the gradient evaluation time (throughput, specifically). For further improved speed, we implement an optimal/accelerated first-order method from [27] instead of using PGD; we will refer to this method as AT. We include many of the implementation tricks from TFOCS [28], including an adaptive step size selection method. Our implementation uses a GPU for the matrix-matrix products and array operations in the iteration, leading to further improved runtime.

Like the example in Subsection 7.1, we will iterate AT for some number of iterations, then stop and perform feature elimination. We perform the dual line search using $\nu^{\text{strict}} = \max\{0, \nu'\}$, as mentioned in Subsection 4.1.1. Using the closed-form solution to the dual line search requires computing $A^T\hat{\nu}$ and $A^T\nu^{\text{strict}}$, which is on the order of the cost of a gradient evaluation. That is to say the dual line search is not terribly expensive, though we do not generally want to do it after each iteration of AT.

Table 1 shows the results of using the strong concavity subproblem (10) (we also show results for subproblem (20), which we discuss shortly). We show the number of iterations of AT, the total number of solutions certified to be unique, and the number of features eliminated across all NNLS problems. There are 40000 NNLS problems, each with 2822 features, giving approximately 113 million features total. As the accuracy of the primal feasible points $\hat{x}$ increases, the duality gap closes and more features are eliminated from the problem. But even at 500K iterations, not all problems are certified to have a unique solution.

	Solutions Certified Unique		Features Eliminated	
Iterations	using (10)	using (20)	using (10)	using (20)
10,000	7237	9510	20.0%	24.8%
50,000	18,339	23,174	46.2%	56.9%
100,000	23,512	28,826	57.9%	68.3%
500,000	34,462	38,094	81.6%	87.8%
500,000 + `lsqnonneg`	–	**40,000**	–	91.0%

Table 1 Number of problems (40000 total) certified to have unique solutions using feature elimination with the strong concavity subproblem (10) and with the strong concavity plus partial dual feasibility subproblem (20). Adding partial dual feasibility constraints in (20) can eliminate sufficiently many features to certify solution uniqueness at fewer iterations than using strong concavity alone in (10).

We know from Corollary 5.1 that we could simply perform more iterations of AT to shrink the duality gap. But instead let us construct a stronger feature elimination subproblem, allowing us to expend a little more work in solving the new subproblem to avoid computing more iterations of AT. We do this by introducing another constraint on (10) to shrink the search set N while still ensuring that $\nu^* \in N$. One of many ways to do this is by adding the dual feasibility constraint: $A^T\nu \geq 0$. This has the possibility to shrink N, thereby increasing the feature elimination lower bound, while guaranteeing that $\nu^* \in N$. Thus this leads to a stronger but still safe subproblem.

But the resulting feature elimination subproblem is too difficult to solve for our purposes. We relax the subproblem by enforcing dual feasibility for only a single column at a time with $a_j^T\nu \geq 0$.

Since we can pick any $j \in \{1, \ldots, n\}$, we solve the subproblem for each column and take the largest lower bound:

$$\begin{aligned} \max_{1 \le j \le n} \min_{\nu} \ & \langle a_i, \nu \rangle \\ \text{s.t.} \quad & \|\nu - \hat{\nu}\|^2 \le 2\epsilon \\ & \langle a_j, \nu \rangle \ge 0. \end{aligned} \tag{20}$$

Each inner problem is a "dome subproblem", since the feasible set is the intersection of a ball and a halfspace. A closed-form solution to the dome subproblem exists (see [1] for instance), allowing us to compute the optimal value cheaply and accurately. The dome subproblem optimal value uses the inner products $\langle a_j, \hat{\nu} \rangle$ and $\langle a_i, a_j \rangle$. With all required inner products computed, evaluation of the optimal value requires $\mathcal{O}(1)$ work.

Observe that we can compute the all the required inner products for all dome subproblems as $A^T\hat{\nu}$ and A^TA. Since A is fixed we precompute A^TA and discard this one-time cost. This brings the cost of computing the optimal value of (20) to about the cost of a gradient evaluation plus $\mathcal{O}(n)$ work for n evaluations of the dome subproblem optimal value.

Table 1 shows the analogous results when using the strong concavity and partial dual feasibility subproblem (20). We see a marked improvement in the number of solutions certified to be unique, though we still fall a bit short of certifying uniqueness for all 40000 problems. This appears to be due to a few particularly slow-to-converge problems where the accuracy of $\hat{x}$ is still quite low. We fix this by computing a high-accuracy solution for the remaining 1996 problems using MATLAB's `lsqnonneg` (note that we do not use `lsqnonneg` for the 5538 remaining problems when using subproblem (10)). This results in a sufficiently accurate $\hat{x}$ and we certify the remaining problems as having unique solutions. The final image is constructed by assembling the individual NNLS solutions, so by certifying that all NNLS solutions are unique we also guarantee that the final image is uniquely determined from the data.

8 Conclusions

We have developed a safe feature elimination strategy for non-negativity constrained convex optimization problems which uses an accurate, but non-optimal, primal-dual feasible pair. We show that under reasonable conditions, a sufficiently accurate primal-dual pair will eliminate all zero coordinates from the problem. To enable our methods to work with optimization algorithms that produce only primal points we also developed a dual line search to construct an accurate dual feasible point from an accurate primal feasible point. This allows us to use a first-order method to solve the primal, use the dual line search to cheaply construct a dual feasible point, and then use SAFE to eliminate features. We demonstrate the use of SAFE to robustly certify the uniqueness of a non-negative least-squares solution in a small synthetic data example and also for a large-scale, extremely ill-conditioned problem set arising from a microscopy application. Once an NNLS solution has been certified unique, safe feature elimination also provides a bound on the distance to the unique optimal point. Possible future directions of this work include strengthening the feature elimination subproblems and dual line search, and extending the uniqueness certification technique to 1-norm regularized problems like lasso. Relaxing the requirement for global strong concavity of the dual objective g (which came from the assumed global L-smoothness of f) via the characterization in [29] may also be fruitful. Another promising line of work is incorporating the feature elimination into active-set methods [30,31] which typically rely on estimating active and inactive features.

Acknowledgements Stephen Becker acknowledges the donation of a Tesla K40c GPU from NVIDIA.

References

1. Ghaoui, L.E., Viallon, V., Rabbani, T.: Safe feature elimination in sparse supervised learning. Pacific Journal of Optimization **8**(4) (2012)
2. Thompson, G.L., Tonge, F.M., Zionts, S.: Techniques for removing nonbinding constraints and extraneous variables from linear programming problems. Management Science **12**(7), 588–608 (1966)
3. Fercoq, O., Gramfort, A., Salmon, J.: Mind the duality gap: safer rules for the lasso. In: Proceedings of the 32nd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 37, pp. 333–342. PMLR, Lille, France (2015)

4. Ndiaye, E., Fercoq, O., Gramfort, A., Salmon, J.: GAP safe screening rules for sparse multi-task and multi-class models. In: Advances in Neural Information Processing Systems, pp. 811–819 (2015)
5. Ndiaye, E., Fercoq, O., Gramfort, A., Salmon, J.: GAP safe screening rules for sparse-group lasso. In: Advances in Neural Information Processing Systems, pp. 388–396 (2016)
6. Ndiaye, E., Fercoq, O., Gramfort, A., Salmon, J.: Gap safe screening rules for sparsity enforcing penalties. The Journal of Machine Learning Research **18**(1), 4671–4703 (2017)
7. Xiang, Z.J., Wang, Y., Ramadge, P.J.: Screening tests for lasso problems. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(5), 1008–1027 (2017)
8. Ogawa, K., Suzuki, Y., Takeuchi, I.: Safe screening of non-support vectors in pathwise SVM computation. In: Proceedings of the 30th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 28, pp. 1382–1390. PMLR, Atlanta, Georgia, USA (2013)
9. Zimmert, J., de Witt, C.S., Kerg, G., Kloft, M.: Safe screening for support vector machines. In: NIPS Workshop on Optimization in Machine Learning (OPT) (2015)
10. Raj, A., Olbrich, J., Gärtner, B., Schölkopf, B., Jaggi, M.: Screening rules for convex problems. arXiv preprint arXiv:1609.07478 (2016)
11. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces. Springer (2017)
12. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge university press, Cambridge, United Kingdom (2004)
13. Beck, A.: First-Order Methods in Optimization, vol. 25. SIAM (2017)
14. Slawski, M., Hein, M., et al.: Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. Electronic Journal of Statistics **7**, 3004–3056 (2013)
15. Tibshirani, R.J.: The lasso problem and uniqueness. Electronic Journal of Statistics **7**, 1456–1490 (2013)
16. Zhang, H., Yin, W., Cheng, L.: Necessary and sufficient conditions of solution uniqueness in 1-norm minimization. Journal of Optimization Theory and Applications **164**(1), 109–122 (2015)
17. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proceedings of the National Academy of Sciences **100**(5), 2197–2202 (2003)
18. Alexeev, B., Cahill, J., Mixon, D.G.: Full spark frames. Journal of Fourier Analysis and Applications **18**(6), 1167–1194 (2012)
19. Tillmann, A.M., Pfetsch, M.E.: The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. IEEE Transactions on Information Theory **60**(2), 1248–1259 (2014)
20. Tropp, J.A.: Greed is good: Algorithmic results for sparse approximation. IEEE Transactions on Information Theory **50**(10), 2231–2242 (2004)
21. Lawson, C.L., Hanson, R.J.: Solving least squares problems, vol. 15. SIAM (1995)
22. Amelunxen, D., Lotz, M., McCoy, M.B., Tropp, J.A.: Living on the edge: Phase transitions in convex programs with random data. Information and Inference: A Journal of the IMA **3**(3), 224–294 (2014)
23. Yu, J.Y., Becker, S.R., Folberth, J., Wallin, B.F., Chen, S., Cogswell, C.J.: Achieving superresolution with illumination-enhanced sparsity. Optics Express **26**(8), 9850–9865 (2018)
24. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK Users' Guide, third edn. Society for Industrial and Applied Mathematics, Philadelphia, PA (1999)
25. Golub, G.H., Van Loan, C.F.: Matrix computations, vol. 3. JHU Press, Baltimore, MD, USA (1998)
26. Goto, K., Geijn, R.A.: Anatomy of high-performance matrix multiplication. ACM Transactions on Mathematical Software (TOMS) **34**(3), 12 (2008)
27. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM Journal on Optimization **16**(3), 697–725 (2006)
28. Becker, S.R., Candès, E.J., Grant, M.C.: Templates for convex cone problems with applications to sparse signal recovery. Mathematical Programming Computation **3**(3), 165 (2011)
29. Goebel, R., Rockafellar, R.T.: Local strong convexity and local Lipschitz continuity of the gradient of convex functions. Journal of Convex Analysis **15**(2), 263 (2008)
30. Ferreau, H.J., Kirches, C., Potschka, A., Bock, H.G., Diehl, M.: qpOASES: A parametric active-set algorithm for quadratic programming. Mathematical Programming Computation **6**(4), 327–363 (2014)
31. van den Berg, E.: A hybrid quasi-Newton projected-gradient method with application to lasso and basis-pursuit denoising. Mathematical Programming Computation (2019)