The Weighted Non-Negative Least-Squares Problem with Implicitly Characterized Points

Ali Fattahi
Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90095,
ali.fattahi.1@anderson.ucla.edu

Sriram Dasu
Marshall School of Business, University of Southern California, Los Angeles, Los Angeles, California 90089,
dasu@marshall.usc.edu

Reza Ahmadi
Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90095,
rahamadi@anderson.ucla.edu

The non-negative least-squares (NNLS) problem is defined as finding the Euclidean distance to a convex cone generated by a set of discrete points in $\mathbb{R}^n$. In this paper, we study NNLS when the discrete points are implicitly known and there are an exponentially large number of them (e.g., the set of integer feasible solutions of a mixed-integer program). This problem is motivated by a large auto manufacturer that produces mass customized products where the products are configured by choosing a subset of features. In addition, this problem can have applications in other manufacturing settings, machine learning, clustering, pattern recognition, and statistics. The NNLS can be formulated as a convex optimization problem, and there are efficient algorithms in the literature for solving it. However, NNLS with implicit points cannot be solved using the existing methods. Therefore, we first present a surrogate problem that enables us to find the optimal solution to this problem. Second, we design a Frank-Wolfe based solution approach for solving our surrogate problem. At each iteration $k$ of our proposed methodology, we generate a feasible solution (upper bound) and show that the upper bound converges to the optimal value at $O(1/k)$. Third, we develop a lower bound that enables us to estimate the optimality gap at each iteration and show that the lower bound also converges to the optimal value at $O(1/k)$. Finally, we test the effectiveness of our approach on a set of simulated and real industrial instances.

Keywords: Non-Negative Least-Squares, High-Dimensional Statistics, High-Dimensional Analytics, Frank-Wolfe.
1. Introduction

The non-negative least-squares (NNLS) problem is defined as finding the Euclidean distance between a target point and a convex cone generated by a set of discrete points in $\mathbb{R}^n$ (Lawson and Hanson 1995, Franc et al. 2005). We study NNLS in which the set of discrete points are not a priori given but are implicitly known (e.g., the set of integer feasible solutions of a mixed-integer program). In particular, we are interested in the instances where the number of discrete points is exponential in $n$—e.g., when the discrete points are the feasible solutions of a binary program, the number of solutions is $O(2^n)$.

Our problem is primarily motivated by an application in a large auto manufacturer, in which $n$ represents the number of options (e.g., engine E1, engine E2, transmission T1, and so on) that define the car configurations (end-products). In this setting, the configurations have to satisfy a set of feasibility constraints and hence the number of feasible configurations can be $O(2^n)$, which are only implicitly defined. The target point is an estimation of future demand of options over the planning horizon—e.g., the number of cars sold over the planning horizon with engine E1. The convex cone generated by the feasible configurations constitutes the set of feasible demand over the planning horizon (Fattahi et al. 2017a,b). The problem is to find a point in the convex cone that has the minimum weighted Euclidean distance to the target point (weights may reflect the importance or the price of options). In addition, this problem can generally be found in the manufacturing systems where products are configured based on a set of options available for the customers. Other applications include variants of classical clustering problems. In some clustering problems, the objective is to find the Euclidean distance between a point and the convex hull of a set of points in $\mathbb{R}^n$, called a cluster, whereas our problem is to find the Euclidean distance between a target point and the convex cone of a set of implicitly known points in $\mathbb{R}^n$.

Although the NNLS is a convex optimization problem and solved efficiently (see, for example, Boutsidis and Drineas (2009) and Potluru (2012)), NNLS with implicit points cannot be solved using the existing approaches. Therefore, we first develop a surrogate problem where by solving
it we find the optimal solution to our problem. Our surrogate problem enables us to design a
Frank-Wolfe based approach \cite{Frank and Wolfe 1956} for finding the optimal solution. We generate
a feasible solution (upper bound) and a lower bound at each iteration $k$ of our proposed approach.
We show that both the upper and lower bounds converge to the optimal value at $O(1/k)$. We
numerically investigate the effectiveness of our approach on a set of instances. Finally, we remark
that the computational complexity of our approach depends on the computational complexity of
the subproblem that we solve at each iteration. If the subproblem is polynomially solved, then the
overall computational complexity of our approach is polynomial in $n$.

We also note that a special case of this problem (non-weighted), where the discrete points are
the solutions of a satisfiability problem, is studied by Fattahi et al. \cite{Fattahi et al. 2017a}. The current paper
extends and generalizes their results by allowing the discrete points to be the integer solutions of a
mixed-integer program, providing a detailed convergence rate guarantee on the upper bound and
estimating the convergence rate of the lower bound.

The paper is organized as follows. Section 2 formally defines our problem. In section 3, we present
the surrogate problem and propose our algorithm for solving it. We establish the convergence rate
of the upper and lower bounds in section 4. Section 5 presents our numerical results. Section 6
concludes.

2. Problem Description

Consider a finite set $\mathcal{Y} = \{y^1, y^2, \ldots, y^m\} \subseteq \mathbb{R}^n$, a target point $\hat{z} \in \mathbb{R}^n$, and an $n \times n$ invertible matrix $W$, satisfying the following assumptions: (i) $y^T W^T W y'' \geq 0$, for all $y'$, $y'' \in \mathcal{Y}$,
(ii) there exists $y \in \mathcal{Y}$ such that $y \neq 0$, and (iii) $\hat{z} \neq 0$ (note that assumptions (ii) and (iii) are used to eliminate
the trivial cases). A complete list of notations is available in Appendix A.

In particular, we are interested in cases where $\mathcal{Y}$ is only implicitly known and $m$ is exponential
in $n$. Let $\mathcal{C}(\mathcal{Y})$ denote the convex cone generated by the set $\mathcal{Y}$. In fact, $\mathcal{C}(\mathcal{Y})$ contains all points $z$
such that $z = \sum_{j=1}^{m} \xi_j y^j$ for some non-negative coefficients $\xi_j$’s. We are interested in the problems
of the form:

$$\text{(P1): } \min h(z) = \|W(z - \hat{z})\|^2$$
s.t. \( z \in CC(\mathbb{Y}). \)

In fact, (P1) finds a point \( z^* \in CC(\mathbb{Y}) \) that has the closest weighted Euclidean distance to \( \hat{z} \). Fig. 1 provides an example in \( \mathbb{R}_+^3 \). For the ease of illustration, we assume \( \mathbb{Y} \) contains five points that are as shown in Fig. 1. Given vectors \( y^1, \ldots, y^5 \), we want to find the Euclidean distance between \( \hat{z} \) and the convex cone generated by \( y^1, \ldots, y^5 \). We assume, in this example, \( W \) is a 3×3 identity matrix.

![Figure 1](image-url)  
An illustrative example of (P1) where \( y^1 = (1, 1, 2) \), \( y^2 = (0, 2, 3) \), \( y^3 = (2, 1, 3) \), \( y^4 = (3, 0, 2) \), \( y^5 = (0, 0, 2) \), and \( \hat{z} = (1, 1, 0) \).

Problem (P1) has a strongly convex objective function. Appendix B presents two alternative formulations of (P1) and discusses their drawbacks. Problem (P1) is very difficult, and general purpose solvers cannot solve medium and large instances. In the remainder, we propose an effective approach for solving (P1).

3. Solution Methodology

We are particularly interested in special cases of (P1) where \( \mathbb{Y} \) is implicitly known, but we can solve \( \min_{y \in \mathbb{Y}} c^T y \) in a reasonable time\(^2\) for a given coefficient vector \( c \in \mathbb{R}^n \). For example, \( \mathbb{Y} \) can be the set of feasible solutions of a satisfiability problem, binary program, or mixed-integer program. Since we can solve \( \min_{y \in \mathbb{Y}} c^T y \) in a reasonable time, then we can sequentially construct \( \mathbb{Y} \). We note
that the optimal solution $z^*$ can be represented as a positive combination of at most $n$ points in $\mathcal{Y}$; hence, a cleverly designed algorithm may find the optimal solution to (P1) by solving $\min_{y \in \mathcal{Y}} c^T y$ at most $O(n)$ times.

We design our solution approach based on this observation. Our approach belongs to the class of methods that sequentially construct the feasible region and stop when a good solution is found. One such approach is the well-known Frank-Wolfe algorithm [Frank and Wolfe 1956, Demyanov and Rubinov 1970], also known as the conditional gradient method. In the literature, the Frank-Wolfe algorithm is used to minimize a convex objective function over a compact convex domain [Jaggi 2013, Lacoste-Julien and Jaggi 2015]. Reddi et al. (2016), Lafond et al. (2015), Hazan and Kale (2012) have recently extended the application of the Frank-Wolfe to minimizing a non-convex objective function and stochastic optimization.

In subsection 3.1, we present our surrogate problem. We propose our Frank-Wolfe based algorithm, discuss our lower bounding procedure, and provide an illustrative example in subsection 3.2. We analyze the subproblem that has to be solved at each iteration of the algorithm and present a modeling framework for it in subsection 3.3.

### 3.1. Surrogate Problem for Problem (P1)

We note that (P1) has a non-compact feasible region and the Frank-Wolfe type approaches cannot be directly applied to it; hence, as a critical step of our approach, we first present a surrogate problem (P2) that enables us to employ our solution method. We will show that one can determine the optimal solution of (P1) in polynomial time if the optimal solution of (P2) is known. Problem (P2) is given as follows:

$$
\text{(P2): } \min z \in \mathcal{H} f(z) = \|z - \hat{W}z\|_2^2
$$

s.t. $z \in \mathcal{C}(\bar{\mathcal{Y}})$, $\hat{z}^T W^T z = \hat{z}^T W^T \hat{z}$,

where $\bar{\mathcal{Y}} := \{W^1, \ldots, W^m, W^m\}$. Define $\mathcal{H} := \{z \in \mathbb{R}^n | \hat{z}^T W^T z = \hat{z}^T W^T \hat{z}\}$. In fact, $\mathcal{H}$ is the set of points in $\mathbb{R}^n$ that belong to the hyperplane $\hat{z}^T W^T z = \hat{z}^T W^T \hat{z}$. Moreover, let $\mathcal{F} \mathcal{H}$ denote the set of feasible solutions to (P2)—i.e., $\mathcal{F} \mathcal{H} := \{z \in \mathcal{H} | z \in \mathcal{C}(\bar{\mathcal{Y}})\}$. 

We provide a brief intuition for forming our surrogate problem (P2). In fact, we aim to generate a new problem that has a compact feasible region. Therefore, we concentrate on the intersection of the cone $CC(\tilde{Y})$ and the hyperplane $\mathcal{H}$. Although the intersection is not compact in general, we will show in subsection 3.3 that there exists a compact subset of the intersection that contains the iterates of our algorithm.

The sets $\mathcal{H}$ and $\mathcal{F}H$ are illustrated in Fig. 2 for the example introduced in Fig. 1 (assuming that $W$ is a $3 \times 3$ identity matrix).

![Figure 2](image_url)

**Figure 2** Illustration of $\mathcal{H}$ (entire shaded region) and $\mathcal{F}H$ (darker region) (assumption: $W$ is an identity matrix).

Note that there is no $z^5$ since $y^5$ in Fig. 1 is orthogonal to $\hat{z}$.

The following theorem shows that by solving (P2), one can obtain an optimal solution for (P1) in polynomial time.

**Theorem 1.** (a) $z^* = 0$ solves (P1) if and only if (P2) is infeasible and (b) if (P2) has a feasible solution, then (P1) and (P2) have unique optimal solutions, which we denote by $z^*$ and $z^{**}$, respectively, and $z^* = \frac{\tilde{z}^T W^T W \hat{z}}{\hat{z}^T W T W \tilde{z}} W^{-1} z^{**}$.

**Proof.** (a, $\Rightarrow$) We use a contradiction to show if $z^* = 0$, then $\mathcal{F}H = \{\}$. Assume that $z^* = 0$ and $\mathcal{F}H \neq \{\}$. Then, there exists $\tilde{z} \in \mathcal{F}H$. Note that $\tilde{z} \neq 0$ (using the definition of $\mathcal{F}H$, we have $\tilde{z}^T W^T \tilde{z} = \tilde{z}^T W^T W \tilde{z} > 0$ because $W^T W$ is a positive definite matrix (because $W$ is invertible) and $\tilde{z} \neq 0$ (because of our assumption (iii) in section 2)).
Define \( \tilde{z} := \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T \hat{z}} W^{-1} \tilde{z} \neq 0 \) (note that \( \tilde{z} \) is well-defined because \( \tilde{z} \neq 0 \)). We show that \( \tilde{z} \) is feasible for (P1) and has a strictly better objective value than \( \mathbf{z}^* = 0 \). To prove the feasibility of \( \tilde{z} \), note the following. First, since \( \bar{z} \in CC(\hat{\mathbf{y}}) \), then \( W^{-1} \tilde{z} \in CC(\hat{\mathbf{y}}) \). Second, since \( \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T \hat{z}} > 0 \) and \( W^{-1} \tilde{z} \in CC(\hat{\mathbf{y}}) \), then \( \tilde{z} \in CC(\hat{\mathbf{y}}) \) (because of the definition of cone). Thus, \( \tilde{z} \) is feasible for (P1).

Next, note that:

\[
\begin{align*}
& -\frac{(\hat{z}^T W^T W \hat{z})^2}{\hat{z}^T \hat{z}} < 0 \\
\Rightarrow & \frac{(\hat{z}^T W^T W \hat{z})^2}{\hat{z}^T \hat{z}} - 2 \frac{(\hat{z}^T W^T \hat{z})^2}{\hat{z}^T \hat{z}} < 0 \\
\Rightarrow & \frac{(\hat{z}^T W^T W \hat{z})^2}{\hat{z}^T \hat{z}} - 2 \hat{z}^T W^T \hat{z} \hat{z}^T W \hat{z} \hat{z} - \hat{z}^T W^T W \hat{z} < 0 \\
\Rightarrow & \frac{(\hat{z}^T W^T W \hat{z})^2}{\hat{z}^T \hat{z}} - 2 \hat{z}^T W^T \hat{z} \hat{z}^T W \hat{z} \hat{z} + \hat{z}^T W^T W \hat{z} < \hat{z}^T W^T W \hat{z} \\
\Rightarrow & \left( \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T \hat{z}} - W \hat{z} \right)^T \left( \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T \hat{z}} - W \hat{z} \right) < (W \hat{z})^T (W \hat{z}) \\
\Rightarrow & h(\tilde{z}) < h(0),
\end{align*}
\]

and hence \( \tilde{z} \) has a strictly better objective value than \( \mathbf{z}^* = 0 \), which contradicts the optimality of \( \mathbf{z}^* = 0 \). Thus, the proof of (a, \( \Rightarrow \)) is complete.

(a, \( \Leftarrow \)) We first show (using a contradiction) that if \( \mathcal{FH} = \{\} \), then \( \hat{z}^T W^T W \hat{z} \leq 0 \), for all \( \tilde{z} \in CC(\hat{\mathbf{y}}) \). Assume that there exists \( \bar{z} \in CC(\hat{\mathbf{y}}) \) such that \( \hat{z}^T W^T W \hat{z} > 0 \). Define \( \bar{z} := \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T \hat{z}} W \tilde{z} \).

Because \( \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T \hat{z}} > 0 \) and \( W \tilde{z} \in CC(\hat{\mathbf{y}}) \), then \( \bar{z} \in CC(\hat{\mathbf{y}}) \). Moreover, \( \bar{z}^T W \bar{z} = \hat{z}^T W^T \left( \frac{\hat{z}^T W^T W \bar{z}}{\hat{z}^T \hat{z}} W \tilde{z} \right) = \hat{z}^T W^T W \tilde{z} \). Therefore, \( \bar{z} \in \mathcal{FH} \), meaning that \( \mathcal{FH} \neq \{\} \). This is a contradiction, and hence, we proved that: if \( \mathcal{FH} = \{\} \), then \( \hat{z}^T W^T W \hat{z} \leq 0 \), for all \( \tilde{z} \in CC(\hat{\mathbf{y}}) \).

Let \( \tilde{z} \in CC(\hat{\mathbf{y}}) \) be arbitrary. We show that the objective value of 0 is at least as good as the objective value of \( \tilde{z} \). We note that:

\[
h(\tilde{z}) = \|W(\tilde{z} - \bar{z})\|^2_2
= \|W(\tilde{z} - \hat{z}^T W^T W \bar{z} + \hat{z}^T W^T W \tilde{z} - \bar{z})\|^2_2
= \left\| (1 - \frac{\hat{z}^T W^T W \bar{z}}{\hat{z}^T \hat{z}^T W \bar{z}}) W \tilde{z} + \left( \frac{\hat{z}^T W^T W \tilde{z}}{\hat{z}^T \hat{z}^T W \tilde{z}} W \tilde{z} - W \tilde{z} \right) \right\|^2_2
= \left\| (1 - \frac{\hat{z}^T W^T W \tilde{z}}{\hat{z}^T \hat{z}^T W \tilde{z}}) W \tilde{z} \right\|^2_2 + \left\| \frac{\hat{z}^T W^T W \tilde{z}}{\hat{z}^T \hat{z}^T W \tilde{z}} W \tilde{z} - W \tilde{z} \right\|^2_2.
\]
meaning that the objective value of \( \hat{z} \) is greater than or equal to the objective value of 0. In the second line, we add and subtract \( \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T W^T W \hat{z}} \) to the expression inside the parentheses. To obtain the fourth line, we note that \( W \hat{z} \) and \( (\frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T W^T W \hat{z}})(W \hat{z} - W \hat{z}) \) are orthogonal—i.e., \( (W \hat{z})^T (\frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T W^T W \hat{z}})(W \hat{z} - W \hat{z}) = 0 \). To obtain the sixth line, we use the fact that \( \hat{z}^T W^T W \hat{z} \leq 0 \), which we proved earlier in this part. Since \( \hat{z} \) is an arbitrary feasible solution of (P1), then \( z^* = 0 \) solves (P1). Thus, the proof of (a, \( \iff \)) is complete.

(b) We first show (using a contradiction) that if \( \mathcal{F} \mathcal{H} \neq \{\} \), then \( z^T W^T W \hat{z} > 0 \). Assume that \( \mathcal{F} \mathcal{H} \neq \{\} \) and \( z^T W^T W \hat{z} \leq 0 \). Then, we have: \( h(z^*) = \|W(\hat{z} - z^*)\|^2 = \|W \hat{z}\|^2 + \|W z^*\|^2 - 2z^T W^T W \hat{z} \geq \|W \hat{z}\|^2 + \|W z^*\|^2 \geq \|W(\hat{z} - 0)\|^2 = h(0) \) (note that \( -2z^T W^T W \hat{z} \geq 0 \), and \( W z^* \neq 0 \) because \( W \) is an invertible matrix, and, in part (a), we showed that if \( \mathcal{F} \mathcal{H} \neq \{\} \), then \( z^* \neq 0 \)). This contradicts the optimality of \( z^* \) because 0 has a strictly better objective value (note that, using part (a), we must have \( z^* \neq 0 \) because we assume \( \mathcal{F} \mathcal{H} \neq \{\} \)). Hence, we proved that: if \( \mathcal{F} \mathcal{H} \neq \{\} \), then \( z^T W^T W \hat{z} > 0 \).

We note that in (P1) and (P2) the objective functions are strongly convex (due to invertibility of \( W \)) and the feasible regions are nonempty and convex sets; hence, (P1) and (P2) have unique optimal solutions, which we denote by \( z^* \) and \( z^{**} \), respectively.

Define \( \lambda := \frac{\hat{z}^T W^T W \hat{z}}{z^T W^T W \hat{z}} \). Note that \( \lambda W z^* \in \mathcal{F} \mathcal{H} \) for the following reasons. First, \( \lambda > 0 \) because \( \hat{z}^T W^T W \hat{z} > 0 \) and \( z^T W^T W \hat{z} > 0 \). Second, since \( W z^* \in \text{CC}(\tilde{Y}) \), then \( \lambda W z^* \in \text{CC}(\tilde{Y}) \). Finally, we note that:

\[
\hat{z}^T W^T (\lambda W z^*) = \frac{\hat{z}^T W^T W \hat{z}}{z^T W^T W \hat{z}} z^T W^T W z^* = \tilde{z}^T W^T W \hat{z},
\]

and hence, we proved that \( \lambda W z^* \in \mathcal{F} \mathcal{H} \).
We also note that \( \frac{1}{\lambda}W^{-1}z^* \in CC(\mathbb{Y}) \) for the following reasons. First, \( \frac{1}{\lambda} > 0 \). Second, \( W^{-1}z^* \in CC(\mathbb{Y}) \). Hence, in summary, we have shown that \( \frac{1}{\lambda}W^{-1}z^* \) and \( z^* \) are feasible for (P1), and \( z^* \) and \( \lambda Wz^* \) are feasible for (P2).

To complete the proof of part (b), we show that \( z^* = \lambda Wz^* \) using a contradiction. Suppose to the contrary that \( z^* \neq \lambda Wz^* \) (or \( \frac{1}{\lambda}W^{-1}z^* \neq z^* \)). Since \( z^* \) is unique and \( \frac{1}{\lambda}W^{-1}z^* \) is feasible for (P1), then we have:

\[
\begin{align*}
  h(z^*) < h\left( \frac{1}{\lambda}W^{-1}z^* \right) \\
  \Rightarrow \left\| W\left( z^* - \frac{1}{\lambda}\hat{z} \right) - \left( 1 - \frac{1}{\lambda} \right) W\hat{z} \right\|_2^2 < \left\| W\left( \frac{1}{\lambda}W^{-1}z^* - \frac{1}{\lambda}\hat{z} \right) - \left( 1 - \frac{1}{\lambda} \right) W\hat{z} \right\|_2^2 \\
  \Rightarrow \left\| W\left( z^* - \frac{1}{\lambda}\hat{z} \right) \right\|_2^2 + \left\| \left( 1 - \frac{1}{\lambda} \right) W\hat{z} \right\|_2^2 < \left\| W\left( \frac{1}{\lambda}W^{-1}z^* - \frac{1}{\lambda}\hat{z} \right) \right\|_2^2 + \left\| \left( 1 - \frac{1}{\lambda} \right) W\hat{z} \right\|_2^2 \\
  \Rightarrow \left\| W\left( z^* - \frac{1}{\lambda}\hat{z} \right) \right\|_2^2 < \left\| W\left( \frac{1}{\lambda}W^{-1}z^* - \frac{1}{\lambda}\hat{z} \right) \right\|_2^2 \\
  \Rightarrow \left\| W(\lambda z^* - \hat{z}) \right\|_2^2 < \left\| W(W^{-1}z^* - \hat{z}) \right\|_2^2 \\
  \Rightarrow f(\lambda Wz^*) < f(z^*).
\end{align*}
\]

The third line follows from the fact that \( W(z^* - \frac{1}{\lambda}\hat{z}) \) and \( (1 - \frac{1}{\lambda})W\hat{z} \) are orthogonal, and \( W\left( \frac{1}{\lambda}W^{-1}z^* - \frac{1}{\lambda}\hat{z} \right) \) and \( (1 - \frac{1}{\lambda})W\hat{z} \) are orthogonal (by feasibility of \( z^* \) on the hyperplane in (P2)).

In the last line, note that \( \lambda Wz^* \) is feasible for (P2) and has a strictly better objective value than \( z^* \). This contradicts the optimality of \( z^* \) for (P2). Hence, \( z^* = \lambda Wz^* \) or \( z^* = \frac{1}{\lambda}W^{-1}z^* \).

So far, we have shown that \( z^* = \hat{\lambda}W^{-1}z^* \), where \( \hat{\lambda} := \frac{1}{\lambda} > 0 \). To complete the proof, we have to determine \( \hat{\lambda} \) (note that we must show \( \hat{\lambda} = \frac{\hat{z}^T W\hat{z}}{z^{**\top} z^{**}} \)). We can find \( \hat{\lambda} \) by solving the following optimization problem (because we already know that \( \hat{\lambda}W^{-1}z^* \) is the optimal solution of (P1)).

\[
\min_{\lambda > 0} h(\lambda W^{-1}z^*) = \left\| W(\lambda W^{-1}z^* - \hat{z}) \right\|_2^2.
\]

The objective function is convex in \( \lambda \). Using first order conditions, we obtain \( \hat{\lambda} = \frac{\hat{z}^T W\hat{z}}{z^{**\top} z^{**}} > 0 \). Therefore, we have:

\[
z^* = \hat{\lambda}W^{-1}z^* = \frac{\hat{z}^T W\hat{z}}{z^{**\top} z^{**}} W^{-1}z^* ,
\]

and hence the proof is complete. \( \Box \)
Algorithm 1 Solving problem (P2) (Assumption: $\mathcal{FH} \neq \emptyset$)

1: Let $y^o \in Y$ such that $\hat{z}^TW^Ty^o > 0$; Define $z^{(0)} := \frac{\hat{z}^TW^Ty^o}{\hat{z}^TWWy^o} \in \mathcal{FH}$; Define $L^{(0)} := 0$;
2: for $k = 0, 1, 2, \ldots$ do
3: Solve $z^{(k)} = \arg \min_{z \in \mathcal{CH}(\{z^{(0)}, \ldots, z^{(k)}\})} f(z)$;
4: if $f(z^{(k)}) \leq L^{(k)} + \varepsilon$ then
5: Return $z^{(k)}$;
6: end if
7: Solve $M(k)$ and denote the optimal solution by $z^{(k+1)}$;
8: Define $L^{(k+1)} := \max \left\{ L^{(k)}, \left( \max \left\{ 0, \frac{(z^{(k+1)} - \hat{z})^TW^{-1}z^{(k+1)} - \hat{z}}{\|z^{(k)} - \hat{z}\|^2} \right\} \right)^2 \right\}$;
9: if $f(z^{(k+1)}) \leq L^{(k+1)} + \varepsilon$ then
10: Return $z^{(k+1)}$;
11: end if
12: end for

Based on Theorem 1, we note the following. First, (P1) always has a finite optimal solution, which we denote by $z^*$. Second, (P2) can be infeasible or it has a bounded optimal solution, which we denote by $z^{**}$. Finally, if (P2) is infeasible, then $z^* = 0$ solves (P1); otherwise, the optimal solution of (P1) is found using equations $z^* = \frac{\hat{z}^TW^T\hat{z}z^*}{\hat{z}^TW^{-1}\hat{z}^*} W^{-1}z^{**}$ in polynomial time. Therefore, we need to solve problem (P2) to determine the optimal solution to (P1).

### 3.2. An Algorithm for Solving the Surrogate Problem

We present our method for solving (P2), which is also shown in Algorithm 1 (assuming $\mathcal{FH} \neq \emptyset$).

An initial feasible point is found that we denote by $z^{(0)} \in \mathcal{FH}$ (note that since $\mathcal{FH} \neq \emptyset$, then there exists $y^o \in Y$ such that $\hat{z}^TW^Ty^o > 0$). We denote by $L^{(k)}$ a lower bound on the optimal value of (P2) at iteration $k$. We will discuss determining $L^{(k)}$ later in this subsection. This lower bound is initially set to 0. At each iteration $k = 0, 1, 2, \ldots$, we first determine the best known point, denoted by $z^{(k)}$, by minimizing the objective function of (P2) over the convex hull of all feasible points known so far (Line (3) of Algorithm 1). If the difference between the objective value of $z^{(k)}$ and $L^{(k)}$ is less than or equal to an acceptable error $\varepsilon$, we stop (Line (4) of Algorithm 1). Otherwise,
we minimize the linear approximation of the objective function at \( z^{*\langle k \rangle} \) over the feasible region of (P2). In fact, we solve the following minimization problem:

\[
\mathcal{M}(k) : \min_{z \in \mathcal{F}H} z^T(z^{*\langle k \rangle} - W\hat{z}),
\]

and denote the optimal solution by \( z^{\langle k+1 \rangle} \) (see Line (7) of Algorithm 1).

An important feature of Algorithm 1 is the presentation of a lower bound at each iteration, which is computed as:

\[
\mathcal{L}^{\langle k+1 \rangle} := \max \left\{ \mathcal{L}^{\langle k \rangle}, \left( \max_0 \left\{ \frac{(z^{\langle k+1 \rangle} - W\hat{z})^T(z^{*\langle k \rangle} - W\hat{z})}{\|z^{*\langle k \rangle} - W\hat{z}\|^2} \right\} \right)^2 \right\},
\]

for all \( k \geq 0 \). Note that our lower bound is always well-defined because the denominator is non-zero (since if \( \|z^{*\langle k \rangle} - W\hat{z}\|^2 = 0 \) then \( f(z^{*\langle k \rangle}) = 0 \), and hence the algorithm must have stopped in Line (4)). The validity of this lower bound is proven in the following proposition.

**Proposition 1.** In Algorithm 1, \( f(z^{*\langle k \rangle}) \geq \mathcal{L}^{\langle k \rangle}, \) for all \( k \geq 0 \).

**Proof.** First, note that \( \mathcal{L}^{\langle 0 \rangle} = 0 \), which is valid; hence, we focus on proving the validity of \( \mathcal{L}^{\langle k \rangle} \) for \( k \geq 1 \). Note that we can represent the lower bound as:

\[
\mathcal{L}^{\langle k+1 \rangle} := \max \left\{ 0, \max_{k' = 0, 1, \ldots, k} \left( \max_0 \left\{ \frac{(z^{\langle k'+1 \rangle} - W\hat{z})^T(z^{*\langle k' \rangle} - W\hat{z})}{\|z^{*\langle k' \rangle} - W\hat{z}\|^2} \right\} \right)^2 \right\},
\]

for all \( k \geq 0 \). Thus, it suffices to prove:

\[
f(z^{*\langle k' \rangle}) \geq \left( \max_0 \left\{ \frac{(z^{\langle k'+1 \rangle} - W\hat{z})^T(z^{*\langle k' \rangle} - W\hat{z})}{\|z^{*\langle k' \rangle} - W\hat{z}\|^2} \right\} \right)^2,
\]

for all \( k' \geq 0 \). We use a contradiction. Suppose that there exist \( k' \) and \( \bar{z} \in \mathcal{F}H \) such that:

\[
f(\bar{z}) < \left( \max_0 \left\{ \frac{(z^{\langle k'+1 \rangle} - W\hat{z})^T(z^{*\langle k' \rangle} - W\hat{z})}{\|z^{*\langle k' \rangle} - W\hat{z}\|^2} \right\} \right)^2.
\]

Note that if \( ((z^{\langle k'+1 \rangle} - W\hat{z})^T(z^{*\langle k' \rangle} - W\hat{z})) / (\|z^{*\langle k' \rangle} - W\hat{z}\|^2) \leq 0 \), then we have \( f(\bar{z}) < 0 \) which is a contradiction; hence, assume that \( ((z^{\langle k'+1 \rangle} - W\hat{z})^T(z^{*\langle k' \rangle} - W\hat{z})) / (\|z^{*\langle k' \rangle} - W\hat{z}\|^2) > 0 \). Thus, we have:

\[
f(\bar{z}) < \left( \frac{(z^{\langle k'+1 \rangle} - W\hat{z})^T(z^{*\langle k' \rangle} - W\hat{z})}{\|z^{*\langle k' \rangle} - W\hat{z}\|^2} \right)^2.
\]
\[\Rightarrow \|\tilde{z} - W\hat{z}\|_2 < \frac{(z^{(k') + 1} - W\hat{z})^T(z^* - W\hat{z})}{\|z^* - W\hat{z}\|_2}\]
\[\Rightarrow \|\tilde{z} - W\hat{z}\|_2 \|z^{(k')} - W\hat{z}\|_2 < (z^{(k') + 1} - W\hat{z})^T(z^* - W\hat{z})\]
\[\Rightarrow (\tilde{z} - W\hat{z})^T(z^* - W\hat{z}) \leq \|\tilde{z} - W\hat{z}\|_2 \|z^{(k')} - W\hat{z}\|_2 < (z^{(k') + 1} - W\hat{z})^T(z^* - W\hat{z})\]
\[\Rightarrow z^T(z^* - W\hat{z}) < z^{(k') + 1}T(z^* - W\hat{z}).\]

The fourth line follows from the Cauchy-Schwartz inequality. Note that the last line contradicts the optimality of \(z^{(k') + 1}\) for problem \(M(k')\). Hence, the proof is complete. \(\square\)

We note that \(L^{(k)}\) is nondecreasing in \(k\) due to the way it is defined. We stop Algorithm 1 if the difference between the current objective value and the lower bound is less than or equal to an acceptable error \(\varepsilon\) (see Lines (9)-(11) of Algorithm 1). Our experiment shows that \(L^{(k)}\) is significantly effective in terminating the execution of Algorithm 1.

We graphically show the steps of applying Algorithm 1 to the example presented in Figs. 1 and 2. Recall that this example assumes \(W\) is a 3 \times 3 identity matrix. For the ease of illustration, we assume \(\varepsilon = 0\). Let \(z^{(0)} \in \mathcal{FH}\) be as shown in Fig. 3(a), and \(L^{(0)} = 0\). In iteration 0, we find \(z^{(0)} = z^{(0)}\). We then minimize \(z^T(z^{(0)} - \tilde{z})\) over the feasible region \(\mathcal{FH}\) and find \(z^{(1)}\). Using the obtained \(z^{(1)}\), we compute \(L^{(1)}\), which is shown in Fig. 3(a). The termination criteria is not satisfied because \(L^{(1)} < f(z^{(0)}) = \|z^{(0)} - \tilde{z}\|^2_2\). Thus, we continue to iteration \(k = 1\) (see Fig. 3(b)). We first find \(z^{(1)}\) as the closest point to \(\tilde{z}\) in the convex hull of \(\{z^{(0)}, z^{(1)}\}\). Then, solving \(M(1)\), we find \(z^{(1)}\) (or \(z^{(0)}\)). We then compute \(L^{(2)}\) and note that the termination criteria is met because \(L^{(2)} = f(z^{(1)}) = \|z^{(1)} - \tilde{z}\|^2_2\). Thus, we stop and \(z^{* (1)}\) is the optimal solution of (P2)—i.e., \(z^{**} = z^{* (1)}\).

We remark that if Algorithm 1 is not equipped with \(L^{(k)}\), it iterates forever and cannot guarantee the optimality of \(z^{* (1)}\).

### 3.3. Analysis of the Subproblem \(M(k)\)

The applicability of Algorithm 1 depends on finding a bounded optimal solution \(z^{(k + 1)}\) by solving \(M(k)\). This is guaranteed if, for example, the feasible region of \(M(k)\) is compact. We remark that \(\mathcal{FH}\) is a closed and convex polyhedron, but not necessarily compact because it can be unbounded.
\[ z^{(0)} = z^{(0)} \]

(a) Iteration 0: \( z^{(1)} \) minimizes \( M(0) \);
\( L^{(1)} \) is found and \( L^{(1)} < f(z^{(0)}) \).

(b) Iteration 1: \( z^{(1)} \) (or \( z^{(0)} \)) minimizes \( M(1) \);
\( L^{(2)} \) is found and \( L^{(2)} = f(z^{(1)}) \); Stop!

Figure 3  Graphical illustration of Algorithm 1 on the example presented in Figs. 1 and 2.

(see, for example, Fig. 2). We must ensure there exists a nonempty and compact polyhedron that contains the optimal solutions of \( M(k) \). This result is not obvious as illustrated in Fig. 4. Suppose \( \mathcal{F} \mathcal{H} \), \( W\hat{z} \), and \( z^{(0)} \) are as shown in Fig. 4 and \( \mathcal{F} \mathcal{H} \) is unbounded with extreme directions \( d^1 \) and \( d^2 \). The objective value of \( M(0) \) improves in the direction of \(- (z^{(0)} - W\hat{z}) \). Hence, the objective value approaches to \(- \infty \) in the direction of \( d^2 \), and there exists no optimal solution. If this case happens, we cannot generate \( z^{(1)} \) and hence cannot proceed. A situation similar to Fig. 4 never happens as we show in the following theorem.

**Theorem 2.** If \( \mathcal{F} \mathcal{H} \neq \{ \} \), then, for all \( k \geq 0 \), \( M(k) \) has a bounded optimal solution, which is an extreme point of \( \mathcal{F} \mathcal{H} \) and in the form of \( \lambda W y \), where \( 0 < \lambda < +\infty \) and \( y \in \mathbb{Y} \).

**Proof.** Note that if for all \( W y \in \tilde{\mathbb{Y}} \), there exists \( 0 < \lambda < +\infty \) such that \( \tilde{z}^T W^T (\lambda W y) = \tilde{z}^T W^T W \hat{z} \), then \( \mathcal{F} \mathcal{H} \) is a bounded set (also because \( \mathbb{Y} \) has a finite number of elements). This case happens if \( \tilde{z}^T W^T W y > 0 \), for all \( y \in \mathbb{Y} \). On the other hand, if there exists \( y \in \mathbb{Y} \) such that \( \tilde{z}^T W^T W y \leq 0 \), then \( \mathcal{F} \mathcal{H} \) can be unbounded.\(^5\)

We first show that \( M(k) \) has a bounded optimal value. Note that for all \( z \in \mathcal{F} \mathcal{H} \) we have:
\[ z^T W \hat{z} = \hat{z}^T W^T W \hat{z} = \Pi, \] where \( \Pi \) is a constant. Then, for all \( z \in \mathcal{F} \mathcal{H} \), the objective value of \( M(k) \) is
Figure 4 An example to highlight the importance of Theorem 2.

equal to $z^T z^{*(k)} - II$. On the other hand, since both $z$ and $z^{*(k)}$ are some nonnegative combinations of $Wy$'s, by Assumption (i), the objective is lower bounded by $-II$.

We next show that $FH$ has at least one extreme point and does not contain lines ($FH$ contains a line if there exist $z', z'' \in FH$, $z' \neq z''$, such that $\theta z' + (1 - \theta)z'' \in FH$, for all $\theta \in \mathbb{R}$). First, $FH$ is a nonempty polyhedron because of the assumption of the theorem. Second, note that $CC(\tilde{Y})$ does not contain lines because, for all $z', z'' \in CC(\tilde{Y})$, we have:

$$z'^T z'' = \left( \sum_{j_1=1}^{m} \xi'_{j_1} Wy_{j_1} \right) T \left( \sum_{j_2=1}^{m} \xi''_{j_2} Wy_{j_2} \right) = \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} \xi'_{j_1} \xi''_{j_2} y_{j_1} y_{j_2} W^T W y_{j_2} \geq 0,$$

due to our assumption (i) in section 2 hence, $FH$ does not contain lines because $FH \subseteq CC(\tilde{Y})$.

In addition, since $FH$ is closed and nonempty, then $FH$ has at least one extreme point (due to Proposition 2.1.2 of Bertsekas (2009)).

Thus, since $FH$ is a closed subset of $\mathbb{R}^n$ with at least one extreme point and since the optimal value of $M(k)$ is attained over $FH$, then an extreme point of $FH$ must be optimal (due to Proposition 2.4.1 of Bertsekas (2009)). This extreme point is in the form of $\lambda Wy$, where $0 < \lambda < +\infty$ and $y \in Y$ (there might be other alternative optimal solutions). Let $CFH$ denote the convex hull of all $\lambda Wy$'s, where $0 < \lambda < +\infty$ and $y \in Y$. Thus, we proved that the optimal solution of $M(k)$ belongs to $CFH$. Moreover, $CFH$ is a nonempty and compact polyhedron. □

The following important results are noted based on Theorem 2. First, $M(k)$ has a bounded optimal solution, which is an extreme point of $FH$, and hence we can limit our search region to the
extreme points of $\mathcal{F}_H$ in the form of $\lambda W y$ where $0 < \lambda < +\infty$ and $y \in \mathbb{Y}$. Second, solving $M(k)$ with the limited search region guarantees that an obtained optimal solution is always bounded. Thus, we reformulate $M(k)$ based on these results as follows.

$$
M(k) : \min (\lambda W y)^T (z^{*(k)} - W \hat{z})
$$

s.t. $y \in \mathbb{Y}, \lambda \geq 0$

$$
\hat{z}^T W^T (\lambda W y) = \hat{z}^T W^T \hat{z}.
$$

Note that $\lambda \neq 0$ because $\hat{z}^T W^T W \hat{z} > 0$. By $M(k)$, we always refer to this new formulation in the rest of the paper. Moreover, we denote by $z^{(k+1)}$ the optimal value of $\lambda W y$.

We note that the effectiveness of Algorithm 1 depends on how the subproblem $M(k)$ is modeled and solved at each iteration. We present a general framework for formulating $M(k)$ when $\mathbb{Y}$ is the set of integer feasible solutions of mixed-integer programs. Before we present our framework, we remark that if $\mathbb{Y}$ is given explicitly, then the solution of $M(k)$ is obtained by computing:

$$
\min_{y \in \mathbb{Y}}: \left\{ \left( \frac{\hat{z}^T W^T W \hat{z}}{\hat{z}^T W^T W y} \right) W y \right\}^T (z^{*(k)} - W \hat{z}),
$$

which can be performed in a time that is polynomial in $m$.

We next present our modeling framework. We assume that $y \in \mathbb{Y}$ is in the form of $y = (y_1, \ldots, y_n)$ such that $1 - 2^U \leq y_i \leq 2^U$, for all $i = 1, \ldots, n$, where $(1 - 2^U)$ and $2^U$ are known lower and upper bounds on the values of $y_i$’s. For example, if $U = 3$, then it is known that all integer variables are between $-7$ and $8$. We note that $U$ can be selected sufficiently big so that the lower and upper bounds are valid for all $y_i$’s, but for achieving the most effective model, we must set $U$ to the smallest possible integer value. We formulate $M(k)$ as the following mixed-integer linear programming problem:

$$
M(k) : \min (W x)^T (z^{*(k)} - W \hat{z})
$$

s.t. $\hat{z}^T W^T W x = \hat{z}^T W^T W \hat{z}$

$$
y_i = (1 - 2^U) + \sum_{u=0}^{U} 2^u v_{iu}, \ \forall i
$$
\[ x_i = (1 - 2^U)\lambda + \sum_{u=0}^{U} 2^u s_{iu}, \quad \forall i \]

\[ 0 \leq s_{iu} \leq M v_{iu}, \quad \forall i, u \]

\[ M(v_{iu} - 1) \leq s_{iu} - \lambda \leq M(1 - v_{iu}), \quad \forall i, u \]

\[ y \in \mathbb{Y}, \quad v_{iu} \in \{0, 1\}, \quad \forall i, u, \quad \lambda \geq 0 \]

where \( x_i, y_i, v_{iu}, s_{iu}, \) and \( \lambda \) are decision variables, \( i = 1, \ldots, n, \ u = 0, \ldots, U, \) and \( M \) is a sufficiently big number (an upper bound on the value of \( \lambda \)). For each \( y_i, \) we define \( U + 1 \) binary variables \( v_{iu}. \) The values of \( v_{iu} \)'s are determined through Eq. (3). For each \( v_{iu}, \) we define a continuous variable \( s_{iu} \) such that \( s_{iu} = \lambda v_{iu}. \) Eqs. (5) and (6) ensure that \( s_{iu} = \lambda v_{iu}, \) for all \( i, u. \) Finally, we define continuous variables \( x_i \) such that \( x_i = \lambda y_i, \) which is guaranteed by Eqs. (1), (2), and (4).

Our formulation of \( M(k) \) requires \( n \) integer variables, \( n(U + 2) + 1 \) continuous variables, and \( n(U + 1) \) binary variables. Note that this is a general formulation and, for example, by setting \( U = 0, \) we achieve a model for cases where \( \mathbb{Y} \) is the set of feasible solutions to satisfiability problems and binary programs.

We finally remark that the computational complexity of Algorithm 1 is the product of the number of iterations that it takes to find a solution with a pre-specified optimality gap and the computational effort in each iteration. If the computational requirement of solving \( M(k) \) is polynomially bounded, then the overall computational complexity of Algorithm 1 is polynomial in \( n. \) On the other hand, if \( M(k) \) is NP-hard, then the overall computational complexity is exponential in \( n. \) We also emphasize that the computational complexity of \( M(k) \) is application-dependent.

### 4. Convergence Rate Guarantee

At each iteration \( k \) of Algorithm 1, we obtain a lower bound \( (L^{(k+1)}) \) and an upper bound \( (f(z^{(k)})) \) on the optimal value \( (f(z^*)) \). In this section, we show that both \( f(z^{(k)}) - f(z^*) \) and \( f(z^*) - L^{(k+1)} \) are upper bounded by \( O(1/k) \)—i.e., both upper bound and lower bound converge to the optimal value at \( O(1/k). \) This also shows that \( f(z^{(k)}) - L^{(k+1)} \) converges to 0 at the same rate, which guarantees that the algorithm will stop after a finite number of iterations for a given \( \varepsilon > 0. \)
4.1. Upper Bound Convergence

It is generally known that in the Frank-Wolfe methods, after \( k \geq 1 \) iterations, \( f(z^{*}(k)) - f(z^{**}) \) is upper bounded by \( \frac{2C_{f}}{k+2} \) where \( C_{f} \) is the curvature constant of the objective function over the domain (see, for example, Frank and Wolfe (1956), and Jaggi (2013)). As one of our contributions, we further specify a detailed upper bound on \( C_{f} \). The following theorem states the convergence rate of our upper bound procedure.

**Theorem 3 (Upper Bound Convergence).** In Algorithm 1, for \( k \geq 1 \), we have:

\[
    f(z^{*}(k)) - f(z^{**}) \leq \frac{8}{k+2} \left( \hat{z}^{T}W^{T}W\hat{z} \right)^{2} \max_{\hat{z}^{T}W^{T}Wy > 0, y \in Y} \frac{y^{T}W^{T}Wy}{(\hat{z}^{T}W^{T}Wy)^{2}}.
\]

In addition, if \( W \) is an \( n \times n \) identity matrix, \( Y \subseteq \{0,1\}^{n} \), and \( 0 \leq \hat{z}_{i} \leq 1 \), for all \( i \), then we have:

\[
    f(z^{*}(k)) - f(z^{**}) \leq \frac{8}{k+2} \left( \frac{\hat{z}^{T}\hat{z}}{\hat{z}_{\text{min},nz}} \right)^{2} (1 + \Delta),
\]

where \( \hat{z}_{\text{min},nz} \) is the smallest nonzero entry of \( \hat{z} \) and \( \Delta \) is the number of zero entries of \( \hat{z} \).

Note that the upper bounds in Theorem 3 can be arbitrarily large since, for example in the second result, one can construct an instance such that \( \hat{z}_{\text{min},nz} \) is arbitrarily small. However, given that \( \hat{z}_{\text{min},nz} > 0 \), the upper bound always exists and is finite. The existence of a finite value is all we need to establish the convergence rate of our algorithm.

In Theorem 2 we showed that if \( \mathcal{F}H \neq \{} \), then the optimal solution of \( M(k) \) belongs to \( C\mathcal{F}H \), which is a nonempty and compact polyhedron. Note that in this section, we always assume \( \mathcal{F}H \neq \{} \) (which was also assumed in the presentation of Algorithm 1). Our analysis of the convergence rate depends on the existence of a finite curvature constant, which is defined as:

\[
    C_{f} := \sup_{z^{1},z^{2} \in C\mathcal{F}H, 0 \leq \gamma \leq 1} \frac{2}{\gamma^{2}} \left( f(z^{1} + \gamma(z^{2} - z^{1})) - f(z^{1}) - \gamma(z^{2} - z^{1})\nabla f(z^{1}) \right).
\]

It is well-known that the curvature constant for \( \frac{1}{2}\|z\|^{2}_{2} \) is the squared Euclidean diameter of the domain (see, for example, Jaggi (2013)). We provide a detailed upper bound on \( C_{f} \) using the structure of our problem.
LEMMA 1. In (P2), we have: 
\[ C_f \leq 4 \left( z^T W^T \hat{z}_\lambda \right)^2 \max_{\hat{z}_\lambda W^T y > 0, \ y \in \mathbb{Y}} \frac{y^T W^T y}{\left( z^T W^T \hat{z}_\lambda \right)^2} \]

PROOF. Using Theorem 2, \( \mathcal{CFH} \) is the convex hull of all \( \lambda W y \)'s such that \( y \in \mathbb{Y} \), \( \hat{z}^T W^T y > 0 \), and \( \lambda = \frac{\hat{z}^T W^T \hat{z}_\lambda}{\hat{z}^T W^T y} \). Thus, noting that \( C_f \) is equal to \( 2 \) times the squared Euclidean diameter of the domain \( \mathcal{CFH} \) (this can be shown by substituting \( f \) and \( \nabla f \) in the definition of \( C_f \)), we have:

\[
C_f = 2 \max_{z^1, z^2 \in \mathcal{CFH}} \| z^1 - z^2 \|^2_2
\]

\[
= 2 \max_{\hat{z}^T W^T (\lambda W y) = \hat{z}^T W^T \hat{z}_\lambda, \ i = 1, 2, \ \lambda \in \mathbb{R}^+} \| \lambda_1 W y - \lambda_2 W y \|^2_2
\]

\[
\leq 4 \max_{\hat{z}^T W^T (\lambda W y) = \hat{z}^T W^T \hat{z}_\lambda, \ i = 1, 2, \ \lambda \in \mathbb{R}^+} \lambda^2 y^T W^T y
\]

In the second line, we use the fact that \( z^1 \) and \( z^2 \) must be extreme points of \( \mathcal{CFH} \); hence, they should be in the form of \( z^i = \lambda_i W y^i \) where \( 0 < \lambda_i < +\infty \), \( y^i \in \mathbb{Y} \), and \( \lambda_i W y^i \) must belong to the hyperplane \( \hat{z}^T W^T z = \hat{z}^T W^T \hat{z}_\lambda \). In the third line, we eliminate \(-2\lambda_1 \lambda_2 y^1 W^T W y^2\) because it is always non-positive (see our assumption (i) in section 2). As a result, the problem decomposes into two identical problems (fourth line). In the fifth line, we substitute \( \lambda \) with \( \frac{\hat{z}^T W^T \hat{z}_\lambda}{\hat{z}^T W^T \hat{z}} \), but we have to ensure that the denominator is positive. Hence, the proof is complete. □

In Lemma 1 the upper bound is finite because for all \( y \in \mathbb{Y} \), the numerator (\( y^T W^T W y \)) is finite and the denominator ((\( z^T W^T W y \))^2) is nonzero. This upper bound can be conveniently simplified for some special cases of our problem. We discuss an important special case where \( W \) is an \( n \times n \) identity matrix, \( \mathbb{Y} \subseteq \{0, 1\}^n \), and \( 0 \leq \hat{z}_i \leq 1 \), for all \( i \).

LEMMA 2. In (P2), if \( W \) is an \( n \times n \) identity matrix, \( \mathbb{Y} \subseteq \{0, 1\}^n \), and \( 0 \leq \hat{z}_i \leq 1 \), for all \( i \), then we have: 
\[ C_f \leq 4 \left( \frac{\hat{z}_i}{\hat{z}_{\min, nz}} \right)^2 (1 + \Delta) \] where \( \hat{z}_{\min, nz} \) is the smallest nonzero entry of \( \hat{z} \) and \( \Delta \) is the number of zero entries of \( \hat{z} \).
Proof. We first show the following result: Let $0 \leq a_1 \leq a_2 \leq \cdots \leq 1$. Then, for all $\ell \geq 1$, such that $a_\ell > 0$, we have:

$$\frac{\ell}{(a_1 + \cdots + a_\ell)^2} \geq \frac{\ell + 1}{(a_1 + \cdots + a_{\ell+1})^2}.$$  

To prove this result, we note that $1 \leq \frac{\ell + 1}{\ell}$, for all $\ell \geq 1$, and then:

$$\frac{(a_1 + \cdots + a_\ell)^2}{\ell} \leq \frac{(\ell + 1)(a_1 + \cdots + a_\ell)^2}{\ell^2(\ell + 1)} = \frac{(\ell + 1)(a_1 + \cdots + a_\ell)^2}{\ell^2(\ell + 1)} + \frac{(2\ell)(a_1 + \cdots + a_\ell)^2}{\ell^2(\ell + 1)} = \frac{(a_1 + \cdots + a_\ell)^2}{(a_1 + \cdots + a_\ell + a_{\ell+1})^2} + \frac{2(a_1 + \cdots + a_\ell)(a_1 + \cdots + a_\ell)}{(\ell + 1)}.$$  

noting that $\frac{a_1 + \cdots + a_\ell}{\ell} \leq a_{\ell+1}$.  

On the other hand, using Lemma 1 and noting that $Y \subseteq \{0, 1\}^n$, we have:

$$C_f \leq 4(\hat{z}^T \hat{z})^2 \max_{\hat{z}^T y > 0, \ y \in \{0, 1\}^n} \frac{y^T y}{(\hat{z}^T y)^2} \leq 4(\hat{z}_{\min, nz}^T \hat{z})^2 (1 + \Delta).$$

Note that $y^T y$ is the number of ones in $y$, and the denominator $\hat{z}^T y$ must be positive. Assume that $\hat{z}_i$’s are sorted in a non-decreasing order. Using the result that we showed earlier in this proof, the fraction is maximized if we select all zero $\hat{z}_i$’s and exactly one smallest nonzero ($\hat{z}_{\min, nz}$). Thus, the proof is complete. □

4.2. Lower Bound Convergence

Establishing that $L^{(k+1)}$ converges to $f(z^{**})$ is important since it guarantees that the algorithm will stop after a finite number of iterations for a given $\varepsilon > 0$. The following theorem states the convergence rate of our lower bounding procedure.
Theorem 4 (Lower Bound Convergence). In Algorithm 1, there exists $K^*$ such that, for $k \geq K^*$, we have:

$$f(z^{**}) - L^{(k+1)} \leq \frac{8\beta}{k+2} \left(z^T W^T \hat{z} \right)^2 \max_{\hat{z}^T W^T y > 0, \ y \in Y} \frac{y^T W^T y}{(\hat{z}^T W^T y)^2},$$

where $\beta = 3.375$. In addition, if $W$ is an $n \times n$ identity matrix, $Y \subseteq \{0, 1\}^n$, and $0 \leq \hat{z}_i \leq 1$, for all $i$, then, for $k \geq K^*$, we have:

$$f(z^{**}) - L^{(k+1)} \leq \frac{8\beta}{k+2} \left(z_{\min, nz}^T \hat{z} \right)^2 (1 + \Delta),$$

where $z_{\min, nz}$ is the smallest nonzero entry of $\hat{z}$ and $\Delta$ is the number of zero entries of $\hat{z}$.

We first present three Lemmas that show the steps of the proof. We will then show how Theorem 4 is proven using these three lemmas.

Lemma 3. $(z - W\hat{z})^T (z^{**} - W\hat{z}) \geq \|z^{**} - W\hat{z}\|_2^2$, for all $z \in CFH$.

Proof. First, note that if $z^{**} = W\hat{z}$, then both sides of the inequality become 0; hence, we assume $z^{**} \neq W\hat{z}$. It is enough to show $z^T (z^{**} - W\hat{z}) \geq z^{**T} (z^{**} - W\hat{z})$, for all $z \in CFH$. We use a contradiction. Assume there exists $\bar{z} \in CFH$ such that $\bar{z}^T (z^{**} - W\hat{z}) < z^{**T} (z^{**} - W\hat{z})$. Obviously, $\bar{z} \neq z^{**}$. Consider the following problem:

$$\min_{0 \leq \alpha \leq 1} \|\alpha \bar{z} + (1 - \alpha)z^{**} - W\hat{z}\|_2^2,$$

which has a strongly convex objective function in $\alpha$; hence, this problem has a unique optimal solution. Using first order condition, the unique optimal value of $\alpha$ is equal to $\frac{(\bar{z} - z^{**})^T (z^{**} - W\hat{z})}{(\bar{z} - z^{**})^T (\bar{z} - z^{**})} > 0$.

Thus, there exists $\bar{z} \in CFH$ such that $\bar{z} \neq z^{**}$ and $\bar{z}$ has a strictly better objective value than $z^{**}$. This contradicts the optimality of $z^{**}$. Hence, the proof is complete. □

Lemma 4. If $z^{**} \neq W\hat{z}$, then there exists $K$ such that $(z - W\hat{z})^T (z^{*(k)} - W\hat{z}) > 0$, for all $z \in CFH$ and $k \geq K$. 

Proof. We first show that: for \( \epsilon > 0 \), there exists \( K \) such that if \( k \geq K \), then \( \|z^{*(k)} - z^{**}\|_2 < \epsilon \).

Let \( \epsilon > 0 \) be given and define \( K := \max\left\{ 1, \left\lfloor \frac{2C_f}{\epsilon^2} - 2 \right\rfloor + 1 \right\} \). Note that because of the definition of \( K \), we have \( \frac{2C_f}{K + 2} < \epsilon^2 \). Then, for \( k \geq K \) we have:

\[
\begin{align*}
\|z^{*(k)} - z^{**}\|_2^2 &= \|(z^{*(k)} - W\hat{z}) - (z^{**} - W\hat{z})\|_2^2 \\
&= \|z^{*(k)} - W\hat{z}\|_2^2 + \|z^{**} - W\hat{z}\|_2^2 - 2(z^{*(k)} - W\hat{z})^T(z^{**} - W\hat{z}) \\
&\leq \|z^{*(k)} - W\hat{z}\|_2^2 - \|z^{**} - W\hat{z}\|_2^2 \\
&\leq f(z^{*(k)}) - f(z^{**}) \leq \frac{2C_f}{k + 2} \leq \frac{2C_f}{K + 2} < \epsilon^2,
\end{align*}
\]

and hence \( \|z^{*(k)} - z^{**}\|_2 < \epsilon \), for \( k \geq K \). Note that we use the result of Lemma 3 to obtain the third line, i.e., \( (z^{*(k)} - W\hat{z})^T(z^{**} - W\hat{z}) \geq \|z^{**} - W\hat{z}\|_2^2 \). In the last line, we use the upper bound convergence result that was discussed in subsection 4.1, i.e., \( f(z^{*(k)}) - f(z^{**}) \leq \frac{2C_f}{k + 2} \), for \( k \geq 1 \).

Define \( Q := \max_{z \in CFH} \|z - W\hat{z}\|_2 \) and note that \( 0 < Q < \infty \). Moreover, let \( \epsilon := \frac{1}{Q}\|z^{**} - W\hat{z}\|_2^2 > 0 \). Therefore, there exists \( K \) such that:

\[
\|z^{*(k)} - z^{**}\|_2 < \frac{1}{Q}\|z^{**} - W\hat{z}\|_2^2, \ \forall k \geq K.
\]

Combining this result with Lemma 3, we obtain:

\[
\begin{align*}
(z - W\hat{z})^T(z^{**} - W\hat{z}) &> Q\|z^{*(k)} - z^{**}\|_2 \\
\Rightarrow (z - W\hat{z})^T(z^{**} - z^{*(k)} + z^{*(k)} - W\hat{z}) &> Q\|z^{*(k)} - z^{**}\|_2 \\
\Rightarrow (z - W\hat{z})^T(z^{*(k)} - W\hat{z}) &> (W\hat{z} - z)^T(z^{**} - z^{*(k)}) + Q\|z^{*(k)} - z^{**}\|_2 \\
&\geq -\|W\hat{z} - z\|_2\|z^{**} - z^{*(k)}\|_2 + Q\|z^{*(k)} - z^{**}\|_2 \\
\Rightarrow (z - W\hat{z})^T(z^{*(k)} - W\hat{z}) &> (Q - \|W\hat{z} - z\|_2\|z^{**} - z^{*(k)}\|_2 \geq 0 \\
\Rightarrow (z - W\hat{z})^T(z^{*(k)} - W\hat{z}) &> 0,
\end{align*}
\]

for all \( z \in CFH \) and \( k \geq K \). Note that the fourth line follows because \( (W\hat{z} - z)^T(z^{**} - z^{*(k)}) \geq -\|W\hat{z} - z\|_2\|z^{**} - z^{*(k)}\|_2 \), and the fifth line follows because \( Q \geq \|W\hat{z} - z\|_2 \), for all \( z \in CFH \) (due to the definition of \( Q \)). Hence, the proof is complete. □
Lemma 5. If $z^{**} \neq W\hat{z}$, then:

$$f(z^{**}) - LB^{(k+1)} = g(z^{(k)}) - (f(z^{(k)}) - f(z^{**})) - \frac{1}{4} \frac{g^2(z^{(k)})}{f(z^{(k)})},$$

for all $k \geq 0$, where $LB^{(k+1)} := \left(\frac{(z^{(k+1)} - W\hat{z})^T(z^{*}(k) - W\hat{z})}{\|z^{*}(k) - W\hat{z}\|_2}\right)^2$, and $g(z) := \max_{z \in CFH} (z - \tilde{z})^T \nabla f(z)$.

Proof. Note that since $z^{**} \neq W\hat{z}$, then $z^{(k)} \neq W\hat{z}$, for all $k \geq 0$. Thus, $f(z^{(k)}) > 0$ and $LB^{(k+1)}$ is well-defined, for all $k \geq 0$. We have:

$$LB^{(k+1)} = \left(\frac{(z^{(k+1)} - W\hat{z})^T(z^{*}(k) - W\hat{z})}{\|z^{*}(k) - W\hat{z}\|_2}\right)^2$$

$$= \left(\frac{(z^{(k+1)} - z^{*}(k)) + (z^{*}(k) - W\hat{z})^T(z^{*}(k) - W\hat{z})}{\|z^{*}(k) - W\hat{z}\|_2}\right)^2$$

$$= \left(\frac{(z^{*}(k) - W\hat{z})^T(z^{*}(k) - W\hat{z})}{\|z^{*}(k) - W\hat{z}\|_2} - \left(\frac{z^{*}(k) - z^{(k+1)}}{\|z^{*}(k) - W\hat{z}\|_2}\right)^2\right)^2$$

$$= \left(\sqrt{f(z^{(k)}) - \frac{1}{2}} \sqrt{f(z^{(k)})}\right)^2$$

$$= -g(z^{*}(k)) + f(z^{*}(k)) + \frac{1}{4} \frac{g^2(z^{*}(k))}{f(z^{*}(k))}.$$

To obtain the fourth line, we note that $g(z^{*}(k)) = \max_{z \in CFH} (z^{*}(k) - \tilde{z})^T \nabla f(z^{*}(k)) = 2(z^{*}(k) - z^{(k+1)})^T(z^{*}(k) - W\hat{z})$. Finally, the result follows by multiplying the obtained equation by -1 and adding $f(z^{**})$ to both sides. □

Completing the Proof of Theorem 4 Using Lemma 5 if $z^{**} \neq W\hat{z}$, then there exists $K$ such that $(z^{(k+1)} - W\hat{z})^T(z^{*}(k) - W\hat{z}) > 0$, for all $k \geq K$. Similar to our discussion in the proof of Proposition 1, we have:

$$f(z^{**}) - L^{(k+1)} = f(z^{**}) - \max \left\{ 0, \max_{k'=0,1,\ldots,k} \left( \max \left\{ 0, \frac{(z^{(k'+1)} - W\hat{z})^T(z^{*}(k') - W\hat{z})}{\|z^{*}(k') - W\hat{z}\|_2} \right) \right) \right\}$$

$$\leq f(z^{**}) - \max_{k'=K,K+1,\ldots,k} \left( \max \left\{ 0, \frac{(z^{(k'+1)} - W\hat{z})^T(z^{*}(k') - W\hat{z})}{\|z^{*}(k') - W\hat{z}\|_2} \right) \right)^2$$

$$= f(z^{**}) - \max_{k'=K,K+1,\ldots,k} \left( \frac{(z^{(k'+1)} - W\hat{z})^T(z^{*}(k') - W\hat{z})}{\|z^{*}(k') - W\hat{z}\|_2} \right)^2$$

$$= \min_{k'=K,K+1,\ldots,k} \{ f(z^{**}) - LB^{(k'+1)} \}$$

$$\leq \min_{k'=K,K+1,\ldots,k} g(z^{*}(k')).$$
for all \( k \geq K \). The last line follows because of Lemma 5 and noting that \( f(z^{**}) - \mathcal{L}(k+1) \leq g(z^{*(k)}) \), for all \( k \geq 0 \) (because \( f(z^{*(k)}) - f(z^{**}) \geq 0 \), \( g^2(z^{*(k)}) \geq 0 \), and \( f(z^{*(k)}) > 0 \), for all \( k \geq 0 \)). Jaggi (2013) shows that if \( k \geq 2 \), then:

\[
\min_{k' \in \{\lceil \frac{2}{3} (k+2) \rceil - 2, \ldots, k \}} g(z^{*(k')}) \leq 2C_f \beta k + 2,
\]

where \( \beta = 3.375 \). Let \( K^* \) be such that: (i) \( K^* \geq K \), and (ii) if \( k \geq K^* \) then \( \lceil \frac{2}{3} (k+2) \rceil - 2 \geq K \).

Thus, for \( k \geq K^* \), we have:

\[
f(z^{**}) - \mathcal{L}(k+1) \leq \min_{k' = K, K+1, \ldots, k} g(z^{*(k')}) \\
\leq \min_{k' = \{\frac{2}{3} (k+2)\} - 2, \ldots, k} g(z^{*(k')}) \\
\leq \frac{2C_f \beta}{k + 2}.
\]

We complete the proof by applying the results that we showed in Lemmas 1 and 2. Moreover, note that our above proof is for the case \( z^{**} \neq W\hat{z} \); however, it can be easily verified that if \( z^{**} = W\hat{z} \), then \( f(z^{**}) = 0 \) and hence the upper bounds of Theorem 4 hold. □

We finally note that one can obtain the following lower bound on \( \mathcal{L}(k+1) \) using Theorem 4 (and noting that \( \mathcal{L}(k) \geq 0 \), for all \( k \geq 0 \)). In Algorithm 1, there exists \( K^* \) such that, for \( k \geq K^* \), we have:

\[
\mathcal{L}(k+1) \geq \max \left\{ 0, f(z^{**}) - \frac{8\beta}{k + 2} \left( \hat{z}^T W^T W \hat{z} \right)^2 \max_{\hat{y}^T W^T y > 0, y \in \mathcal{V}} \frac{y^T W^T x (\hat{z}^T W^T y)^2}{z_{\min,nz}} \right\}.
\]

In addition, if \( W \) is an \( n \times n \) identity matrix, \( \mathcal{V} \subseteq \{0,1\}^n \), and \( 0 \leq \hat{z}_i \leq 1 \), for all \( i \), then, for \( k \geq K^* \), we have:

\[
\mathcal{L}(k+1) \geq \max \left\{ 0, f(z^{**}) - \frac{8\beta}{k + 2} \left( \hat{z}_{\min,nz} \right)^2 (1 + \Delta) \right\},
\]

where \( \hat{z}_{\min,nz} \) is the smallest nonzero entry of \( \hat{z} \) and \( \Delta \) is the number of zero entries of \( \hat{z} \).

5. Numerical Results

In this section, we first present the result of testing our algorithm on a limited set of randomly generated instances to generally show the quality of the upper and lower bounds and the actual and theoretical convergence behavior.6 In these instances, the implicit points are characterized by mixed-integer linear programs (MILPs). Second, we look at cases where the discrete points are clustered and show how this impacts the performance of our algorithm. Finally, we apply our algorithm to a set of industrial instances provided to us by a global auto manufacturer.
5.1. When the Implicit Points are Characterized by MILPs

In our instances, we assume that $\mathcal{Y}$ is the set of integer feasible solutions of $Ay + B\chi \leq b$ where $y \in \mathbb{Z}_+^n$, $\chi \in \mathbb{R}_+^q$, $A$ and $B$ are $n \times p$ and $q \times p$ matrices, respectively, and $b \in \mathbb{R}^n$. Each entry of $A$ and $B$ is 0 with probability $\pi$ and generated using a uniform distribution between -10 and 10, with probability $1 - \pi$. Each entry of $b$ is generated uniformly between 0 and 100. Each entry of $\hat{z}$ is generated uniformly between 0 and 1. Matrix $W$ is assumed to be diagonal where each diagonal entry is generated uniformly between 0 and 10. Table 1 shows the parameter setting and results for 9 sets of problems solved. For each set, we solve 10 instances and report the average statistics.

Fig. 5 shows the average convergence of the lower and upper bounds for the problem sets 2, 5, and 8 as well as the theoretical lower and upper bounds. Note that the lower and upper bounds converge in a reasonable number of iterations. Furthermore, we note that the actual convergence rates are much faster than the theoretical ones. Moreover, although the theoretical lower bounds are zero in the first 100 iterations, the actual lower bounds (dashed red curves) rapidly converge to the optimal value. Table 1 summarizes the results of the 9 problem sets.

5.2. When the Implicit Points Form Clusters

In this experiment, we set the number of implicit points, $|\mathcal{Y}|$, to 1000, and let $n$, the dimension of $\mathbb{R}^n$, vary between 100 and 1000 with the increments of 100. For the ease of experimentation, we let $W$ be an identity matrix. We consider 1, 2, and 5 clusters for each problem size. Let $NC$ denote
Table 1  Summary statistic (The algorithm is implemented in IBM ILOG CPLEX Optimization Studio 12.6.1 on a PC with Processor Intel(R) Core(TM) i5-2520M CPU 2.50GHz, 4.00 GB of RAM, and 64-bit Operating System.)

<table>
<thead>
<tr>
<th>Problem Set</th>
<th>n</th>
<th>p</th>
<th>q</th>
<th>π</th>
<th>Upper Bounds</th>
<th>Avg. number of iterations</th>
<th>Avg. time per iteration (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>200</td>
<td>0</td>
<td>0.9</td>
<td>1</td>
<td>43.20</td>
<td>0.20</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>200</td>
<td>0</td>
<td>0.9</td>
<td>3</td>
<td>49.11</td>
<td>4.78</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>200</td>
<td>0</td>
<td>0.9</td>
<td>8</td>
<td>34.50</td>
<td>36.45</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>100</td>
<td>0</td>
<td>0.8</td>
<td>1</td>
<td>61.00</td>
<td>0.37</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>100</td>
<td>0</td>
<td>0.8</td>
<td>3</td>
<td>44.20</td>
<td>40.13</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>100</td>
<td>0</td>
<td>0.8</td>
<td>8</td>
<td>26.20</td>
<td>76.91</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>200</td>
<td>10</td>
<td>0.9</td>
<td>1</td>
<td>46.70</td>
<td>0.24</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>200</td>
<td>10</td>
<td>0.9</td>
<td>3</td>
<td>48.30</td>
<td>0.63</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>200</td>
<td>10</td>
<td>0.9</td>
<td>8</td>
<td>42.20</td>
<td>11.04</td>
</tr>
</tbody>
</table>

the number of clusters and $CF$ denote cluster coefficient—a measure of the closeness of the $y$’s in a cluster. If $CF = 0$, then all $y$’s are randomly distributed in $\mathbb{R}^n_+$. In this case, we create each $y$ by uniformly generating $n$ nonnegative numbers. If $CF > 0$, then we assume that the data points from $(\frac{(\phi-1)|Y|}{NC}+1)$ to $(\frac{|Y|}{NC})$ are in the $\phi$’th cluster and the $(\frac{(\phi-1)|Y|}{NC}+1)$’th data point is the center of the $\phi$’th cluster, for all $\phi = 1, \ldots, NC$. The center of cluster $\phi$, denoted by $y^c_\phi$, is created randomly. Then, a new point is generated in cluster $\phi$ using the formula $(CFy^c_\phi + \tilde{y})$ where $\tilde{y}$ is generated randomly in $\mathbb{R}^n_+$. Obviously, if $CF = 0$, then there is no clustering, and as $CF$ increases, the angle between the points and the center of the cluster becomes smaller. Note that the lengths of $y$’s are not important as we are interested in the rays generated by $y$’s.

Fig. 6 summarizes the result of applying our approach for 1, 2, and 5 clusters, where the horizontal and vertical axis show the number of dimensions and the number of iterations. Each point indicates the average of the results of 50 randomly generated instances. Fig. 6 shows that, as expected, the number of iterations (almost linearly) increases in $n$. As $CF$ increases, the number of iterations decreases. Finally, the number of iterations increases in the number of clusters.

5.3. Industrial Instances

We next test our methodology on 2 industrial instances. In each of these instances, $n$ is the number of options (features, as we discuss in the introduction) and the constraint set for defining $Y$ is similar to satisfiability constraints. These two instances have 200 and 395 options, and 500 and
2,208 constraints, respectively. Fig. 7 shows our results. Note that the theoretical lower bounds become positive after about 100 iterations, and the actual lower bounds are very effective.

In summary, our computational experiment verifies the performance of our proposed methodology for solving problem (P1) and the effectiveness of the lower and upper bounding procedures.

6. Conclusions

In this paper, we present a new methodology for solving NNLS when the discrete points are implicitly characterized. This class of problems are found in manufacturing systems, clustering, machine learning, and statistics. We first define a surrogate problem that enables us to apply our approach, which is similar to a variant of the Frank-Wolfe algorithm. At each iteration, we find a lower bound and an upper bound and show that their difference converges to 0 at a rate of $O(1/k)$. 
We perform an experiment on a set of numerical examples to illustrate the effectiveness of our approach when the discrete points are the solutions of MILPs and when they form clusters. We finally show the results of applying our method to two industrial instances.

Endnotes

1. We acknowledge that, in general, Assumption (i) is restrictive and difficult to check and guarantee. In the application that motivated our paper, all discrete points are in the same orthant and $W$ is diagonal, which satisfies Assumption (i). Furthermore, in many similar practical applications, this assumption is simply satisfied. Theoretically, one needs to solve $\min_{y',y''\in \mathbb{Y}} y'^TW^TWy''$ in order to check Assumption (i). We note that this problem has a nonlinear objective function because it contains components that are the product of two discrete variables, e.g., $y'_i y''_i$. Since $y$’s are discrete, then $y'_i y''_i$ can be linearized, which increases the number of variables polynomially. The resulting problem may still be significantly difficult to solve.

2. “Reasonable time” does not imply polynomial time; rather, it means that the problem is solved fast.

3. Note that we use $\min_{y\in \mathbb{Y}} \mathbf{c}^T y$ to motivate our approach since this problem has basically similar objective function and constraints to the subproblem that we will introduce in subsection 3.2.

4. One could delete Lines 4-6 to have a more compact representation; however, this may result in an unnecessary computation in Lines 7-8.

5. Note that if there exists $\tilde{y} \in \mathbb{Y}$ such that $\tilde{z}^TW^TW\tilde{y} \leq 0$, then there exists no $\lambda > 0$ satisfying $\lambda W\tilde{y} \in \mathcal{F\mathcal{H}}$.

6. Note that the existing approaches for solving NNLS cannot be applied for the NNLS with implicit points since enumerating all feasible points is not practical. Therefore, our method cannot be compared to the existing methods.

Acknowledgments

We would like to thank the review team for helping us greatly improve our paper.
References


Potluru VK (2012) Frugal Coordinate Descent for Large-Scale NNLS. *AAAI*.


Appendix A: Notations

**Abbreviations:**

NNLS Non-negative least squares problem
Our formulation of NNLS with implicit points
Our surrogate formulation of NNLS with implicit points

Notations:

\( \mathbb{Y} = \{ \mathbf{y}^1, \ldots, \mathbf{y}^m \} \) — The set of discrete points in \( \mathbb{R}^n \).
\( m \) — The number of points in \( \mathbb{Y} \).
\( n \) — Dimension of \( \mathbb{R}^n \).
\( \hat{z} \) — The target point.
\( \mathcal{CC}(\mathbb{Y}) \) — The convex cone of the set \( \mathbb{Y} \).
\( \mathcal{CH}(.) \) — The convex hull of a given set.
\( \xi_1, \ldots, \xi_m \) — The non-negative coefficients of \( \mathbf{y} \)'s.
\( W \) — An \( n \times n \) invertible matrix.
\( h(z) = \|W(z - \hat{z})\|_2^2 \) — The objective function of (P1).
\( f(z) = \|z - W\hat{z}\|_2^2 \) — The objective function of (P2).
\( z^*, z^{**} \) — The optimal solutions of (P1) and (P2), respectively.
\( \lambda, x_i, v_{iu}, s_{iu} \) — The variables used in the formulation of \( \mathcal{M}(k) \).
\( 1 - 2U, 2U \) — The known lower and upper bounds on the values of \( y_i \)'s.
\( \mathcal{CFH} \) — The compact subset of \( \mathcal{FH} \) that includes the optimal solution of \( \mathcal{M}(k) \).
\( C_f \) — The curvature constant of \( f \) over the domain \( \mathcal{CFH} \).
\( \Delta \) — The number of zero entries of \( \hat{z} \).

Appendix B: Alternative Formulations of NNLS with Implicit Points

A possible approach is to formulate the feasible region of (P1) as a set of mixed-discrete nonlinear programming constraints, as we discuss below. The optimal solution \( z^* \) may be in the interior or boundary of \( \mathcal{CC}(\mathbb{Y}) \); hence, \( z^* \) can be represented as a non-negative combination of at most \( n \) points in \( \mathbb{Y} \). Thus, there exist \( \xi_1, \ldots, \xi_n \in \mathbb{R}_+ \) and \( \mathbf{y}^1, \ldots, \mathbf{y}^n \in \mathbb{Y} \) such that \( z^* = \sum_{i=1}^{n} \xi_i \mathbf{y}^i \). Therefore, (P1) can be formulated as:

\[
\begin{align*}
\min & \quad h(z) = \|W(z - \hat{z})\|_2^2 \\
\text{s.t.} & \quad z = \sum_{i=1}^{n} \xi_i \mathbf{y}^i, \quad \xi_1, \ldots, \xi_n \in \mathbb{R}_+, \quad \mathbf{y}^1, \ldots, \mathbf{y}^n \in \mathbb{Y},
\end{align*}
\]

which is a significantly difficult problem because of the nonlinearity in the objective function and constraints, and the existence of \( O(n) \) continuous and \( O(n^2) \) discrete variables (note that nonlinearity is due to the fact that both \( \xi \) and \( \mathbf{y} \)'s are decision variables). An special case of our
problem is when $Y$ is the set of feasible solutions to a binary linear program. In this special case, the feasible region of (P1) can be defined through linear constraints by using $O(n^2)$ continuous and $O(n^2)$ binary variables. This problem is still very difficult because if, for example, $n = 100$, then (P1) is defined using 10,000 binary variables.

Another possible approach is to formulate (P1) as follows:

$$(\text{P1'}): \min \tilde{h}(z, \lambda) = \|W(\lambda z - \hat{z})\|_2^2$$

s.t. $z \in \mathcal{CH}(Y)$, $\lambda \geq 0$,

where $\mathcal{CH}(Y)$ is the convex hull of $Y$. This alternative formulation is not advantageous because the objective function is not necessarily convex. This is shown in the following lemma.

**Lemma 6.** $\tilde{h}(z, \lambda) = \|W(\lambda z - \hat{z})\|_2^2$ is not necessarily convex over the domain $z \in \mathcal{CH}(Y)$, $\lambda \geq 0$.

**Proof.** Consider an instance of (P1’) with $n = 1$, $Y = \{0, 1\}$, $\hat{z} = 0.5$, and $W = 1$. For this instance, (P1’) can be written as $\min_{0 \leq z \leq 1, \lambda \geq 0} (\lambda z - 0.5)^2$. Fig. 8 shows the graph of $(\lambda z - 0.5)^2$ over the domain $0 \leq z \leq 1$, $0 \leq \lambda \leq 1$. Note that the objective function is not convex (consider, for example, the diagonal from (0,0) to (1,1)). □

![Figure 8](image-url)

**Figure 8** The graph of $(\lambda z - 0.5)^2$ over the domain $0 \leq z \leq 1$, $0 \leq \lambda \leq 1$

**Author Biographies**

**Ali Fattahi** is a PhD Candidate in Decisions, Operations & Technology Management at UCLA Anderson School of Management. His research interests are in the areas of parts and product portfolio in mass customization, electricity peak load demand management, and high-dimensional statistics.
Sriram Dasu is an associate professor of Data Sciences and Operations at USC Marshall School of Business. His research interest are healthcare operations, global health, service operations with focus on customer psychology, and supply chain management.

Reza Ahmadi is a professor of Decisions and Operations Management at UCLA Anderson School of Management. His recent research interests are in the areas of parts and product design, supply chain with gray markets, and networks and micro-finance.