

Subset Selection in Sparse Matrices*

Alberto Del Pia[†] Santanu S. Dey[‡] Robert Weismantel[§]

February 3, 2020

Abstract

In Subset Selection we search for the best linear predictor that involves a small subset of variables. From a computational complexity viewpoint, Subset Selection is NP-hard and few classes are known to be solvable in polynomial time. Using mainly tools from Discrete Geometry, we show that some sparsity conditions on the original data matrix allow us to solve the problem in polynomial time.

Key words: subset selection, linear least squares, polynomial-time algorithm, sparsity

1 Introduction

With the current developments in technology, data science and machine learning it is apparent that in the near future more and more variants of Mixed Integer Optimization models will emerge whose dimension is extremely large, if not huge. In view of the fact that mixed integer optimization is NP-hard, the aim is to identify parameters in the input data that lead to polynomial-time algorithms.

One natural parameter in this context is the sparsity of the input data. More precisely, the goal is to identify sparsity patterns in the data and exploit them to design algorithms that depend more on the complexity of the sparsity patterns and less on the ambient dimension of the problem. The hope is that in such a way, high dimensional problems, which are typically intractable simply due to their dimension, can be made algorithmically solvable by decomposing them into smaller ones. For example, consider a Linear Integer Optimization problem where one aims at optimizing a linear function over the integer points in a polyhedral region $\{x \in \mathbb{Z}^n \mid Ax \leq b\}$. Assume that the underlying constraint matrix A is the so-called 1-sum of matrices A^1, \dots, A^h , i.e.,

$$A = \begin{pmatrix} A^1 & & & 0 \\ & A^2 & & \\ & & \ddots & \\ 0 & & & A^h \end{pmatrix}.$$

***Funding:** A. Del Pia is partially funded by ONR grant N00014-19-1-2322. The work of R. Weismantel is supported by the Einstein Foundation Berlin. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

[†]Department of Industrial and Systems Engineering & Wisconsin Institute for Discovery, University of Wisconsin-Madison. E-mail: delpia@wisc.edu.

[‡]School of Industrial and Systems Engineering, Georgia Institute of Technology. E-mail: santanu.dey@isye.gatech.edu.

[§]Department of Mathematics, ETH Zürich. E-mail: robert.weismantel@ifor.math.ethz.ch.

A high dimensional Linear Integer Optimization problem of 1-sum structure can be easily solved by decomposing it into optimization problems associated with the smaller units A^1, \dots, A^h . In particular, if for every $i \in \{1, \dots, h\}$ the number of columns of A^i is a constant, then Lenstra’s algorithm [28] applied h times shows us how to solve the original high-dimensional optimization problem in polynomial time. We remark that in this paper we use the standard notion of polynomial-time algorithms in Discrete Optimization, and we refer the reader to the book [33] for a thorough introduction. A more involved sparsity structure in a matrix arises when there is an additional matrix $C = (c_1 | \dots | c_k)$ whose number of columns k is also assumed to be constant, but the matrix C itself is part of a larger matrix M such that it may interact with all the matrices A^i . More formally, the matrix M is of the form

$$M = (A|C) = \left(\begin{array}{ccc|ccc} A^1 & & & | & & | \\ & \ddots & & | & \dots & | \\ & & A^h & | & & | \\ & & & c_1 & \dots & c_k \end{array} \right). \quad (1)$$

The complexity for Linear Integer Optimization associated with a matrix of sparsity pattern (1) in which the number n_i of columns of each A^i is a constant and k is a constant is not known, unless one also requires that $\|M\|_\infty$ is a constant. In fact, extensive research in the field in the past decade on N-fold matrices [29, 22], 4-block decomposable matrices [21] and tree-fold matrices [7] finally led to the most general result about which matrices of the form (1) can be handled efficiently: Linear Integer Optimization can be solved in polynomial time if the so-called tree width of M or M^T is constant and the value $\|M\|_\infty$ is a constant, see [27]. What can one say about other variants of Mixed Integer Optimization associated with matrices of sparsity pattern (1) in which the number of columns of each A^i is a constant and k is a constant, but the parameter h is part of the input?

One important class of Mixed Integer Optimization problems arising in the context of statistics, data science and machine learning, is called Subset Selection and is formally introduced in Equation (2) below. In the underlying model all variables may attain real values. Among all variables, however, only a certain, a-priori given number of them are allowed to attain non-zero values. This latter combinatorial condition is the discrete flavor of the underlying model. The objective is a convex quadratic function in variables $x \in \mathbb{R}^d$ and the scalar variable μ . The remaining characters stand for rational data in the problem instance: M is an $m \times d$ matrix, b and c are m -vectors, and σ is a natural number. Finally, the support of a vector x is defined to be $\text{supp}(x) = \{i \mid x_i \neq 0\}$ and $\|\cdot\|$ denotes the Euclidean norm. Subset Selection is then defined as the optimization problem

$$\begin{aligned} \min \quad & \|Mx + c\mu - b\| \\ \text{s.t.} \quad & x \in \mathbb{R}^d, \mu \in \mathbb{R} \\ & |\text{supp}(x)| \leq \sigma. \end{aligned} \quad (2)$$

Problem (2) is a standard model in statistics, see the textbooks [31] and [20]. In this context, it is usually assumed that c is the vector of all ones. It is also often assumed that the columns of M and b are mean-centered (i.e., the sum of entries in the columns of M and in b is zero), in which case it can be shown that the optimal value of μ is zero. From an optimization point of view there is no complication if one does not assume mean-centering. Therefore, we retain the μ variable and the vector c .

From a computational complexity point of view, Subset Selection is – such as Integer Linear Optimization – NP-hard in general [37]. This complexity viewpoint then leads us naturally to the same question introduced in the context of Linear Integer Optimization: can we identify sparsity patterns in the input data for Subset Selection that allow us to design algorithms that depend

more on the complexity of the sparsity patterns and less on the ambient dimension of the problem? More precisely, given a matrix of sparsity pattern (1) where each of the blocks A^1, \dots, A^h and C have a constant number of columns, can Subset Selection be solved in polynomial time? This paper answers the question in the affirmative. While some other sparsity patterns have been explored in the context of compressive sensing [19], to the best of our knowledge the sparsity pattern in the form of (1) is unexplored. Formally, the following is the main result of this paper:

Theorem 1. *There is an algorithm that solves problem (2) in polynomial time, provided that M is of the form (1) and each of the blocks A^1, \dots, A^h and C have a constant number of columns.*

In particular, our algorithm solves $O(2^k h^{(k+1)^2(2\bar{\theta}+3)})$ linear least squares problems, where k is the number of columns of C , n_i is the number of columns of block A^i , $\theta := \max\{n_i : i = 1, \dots, h\}$, and $\bar{\theta} := (\theta - 1)\theta(\theta + 1)/2$. We remark that the presence of k and of n_1, \dots, n_h in the exponent is expected, since the problem is NP-hard even if only one among k, n_1, \dots, n_h is not a constant but a parameter which is part of the input. This is because the block with a non-constant number of columns can be used to solve a general Subset Selection problem.

The generality of the techniques used in our proofs yields, in fact, a result that is stronger than Theorem 1. Indeed, the 2-norm in the objective plays no special role in our derivation, and our proofs can be tweaked to deal with a more general p -norm, for any fixed $p \geq 1$.

This paper is organized as follows. After discussing related literature in Section 1.1, in Section 1.2 we present an overview of our algorithm and we introduce a reduction of problem (2) to separable form. In Section 2 we prove Theorem 1 under the additional assumption that each A^i block is a just a 1×1 -matrix, meaning that the matrix A is diagonal. Then, in Section 3 we give a complete proof of Theorem 1, and so we study the more general case where each A^i block can be any matrix with a constant number of columns. The polynomial-time algorithm presented in Section 3 clearly implies the polynomial solvability of the class of problems discussed in Section 2. We present our algorithms in this order, since the proof for the diagonal case features many key ideas that will also be used in the block diagonal case, but with significantly more technicalities.

1.1 Relevance of the topic and related work

In Discrete Optimization, few examples of sparsity patterns are known that allow us to solve the problem in a running time that grows moderately with the ambient dimension. The earliest occurrence of sparsity patterns studied in Discrete Optimization is sixty years old. It occurs in the famous maximum $s - t$ flow over time problem of Ford and Fulkerson [16]. See [26] and [35] for surveys about flow over time models.

Problem (2) and variations of it have been studied by several communities with quite different foci on theoretical and more computational aspects. Because of the breadth of the topics, a detailed description of the state-of-the-art for its many faces does not appear to be feasible. We list below several results with links to the computational complexity viewpoint on the subject, with comparison to our result, and point the reader to references.

Algorithms assuming properties of eigenvalues of $M^T M$ Gao and Li [18] present a polynomial time algorithm when the $d - l$ largest eigenvalues of $M^T M$ are identical, provided that l is a fixed number. Clearly, our result is significantly different since it is possible in our setting that the top eigenvalue of $M^T M$ is unique. Indeed, if M is a diagonal matrix with distinct entries, we have a unique maximum eigenvalue of $M^T M$. However, in this case one could argue that it is possible to scale variables and arrive at a resulting matrix where all the eigenvalues of $M^T M$ are equal. On the other hand, consider the matrix:

$$M = \left(I_n \mid \frac{1}{\sqrt{n}} \mathbf{1} \right),$$

where I_n is the n -times- n identity matrix and $\mathbf{1}$ is the column vector with each component equal to 1. Now each column has an equal two-norm. It is easily verified that the largest eigenvalue of $M^\top M$ is unique (in fact equal to 2 with the next eigenvalue being equal to 1), thus showing that our results are different from those in [18].

Algorithms assuming sparsity properties of the covariance matrix The covariance matrix is computed as $(M - \mathbf{1}a)^\top (M - \mathbf{1}a)$ where $\mathbf{1}$ is a column vector of 1s and a is a row vector with a_i being the average of the i^{th} column of M . Associated with the covariance matrix is the so-called covariance graph, whose nodes correspond to columns of M , i.e. the variables, and an edge between two nodes i and j reflects the fact that the corresponding variables have non-zero covariance, i.e. the entry (i, j) of $(M - \mathbf{1}a)^\top (M - \mathbf{1}a)$ is non-zero.

Das and Kempe [1] give exact algorithms when the covariance graph is sparse: either a tree, or an independent set with a fixed number of variables not participating in the independent set. Also, there exists a fully polynomial time approximation scheme for solving Subset Selection when the covariance matrix has constant bandwidth [1]. These results have been obtained by exploiting matrix perturbation techniques. It is easy to see that in our setting the covariance matrix can be very dense. For example, consider again the matrix:

$$M = \left(I_n \mid \frac{1}{\sqrt{n}} \mathbf{1} \right).$$

If we examine the MLE estimator of the covariance matrix (see [1]) of M (where rows are samples and columns are variables), then it is easy to verify that $\text{cov}(i, j) = -\frac{1}{n^2} \neq 0$ for $i, j \leq n$ and $i \neq j$.

Algorithms assuming restricted isometry property of M A series of papers [11, 3, 4, 5, 6] study this problem in the signal processing community under the name “sparse approximation” with applications to the sparse signal recovery from noisy observations. These papers prove various versions (of deterministic or randomized ‘noisy’ setting) of the following meta-theorem: “Suppose there exists a sparse (noisy) solution \hat{x} to $Mx = b$, that is $M\hat{x} = b + \text{‘noise’}$ and \hat{x} is σ -sparse. If the matrix M satisfies the “restricted isometry condition”, then there is a simple convex program whose optimal solution exactly recovers the solution \hat{x} .” These results serve as a theoretical foundation of why lasso [36] often works well experimentally.

To derive the convex program, examine a dual version of problem (2): given a scalar $\epsilon > 0$, a target vector b and a matrix M , the task is to solve

$$\min\{|\text{supp}(x)| : x \in \mathbb{R}^d, \|Mx - b\| \leq \epsilon\}. \quad (3)$$

or a simplified version of it where the support-objective is replaced by an l_1 -norm objective:

$$\min\{\|x\|_1 : x \in \mathbb{R}^d, \|Mx - b\| \leq \epsilon\}. \quad (4)$$

The latter problem can be attacked using convex optimization tools.

It is also easy to show that the M matrices we can provide a polynomial-time algorithm for, do not always satisfy the required restricted isometry condition. A key parameter is the S -restricted isometry constant [5] of M , denoted as δ_S , and defined as a scalar satisfying:

$$(1 - \delta_S) \|x\|^2 \leq \|M_T x\|^2 \leq (1 + \delta_S) \|x\|^2, \quad \forall x \in \mathbb{R}^{|T|},$$

where T is a subset of the index set of columns of M , $|T| \leq S$ and M_T is the matrix with columns corresponding to indices in T . One restricted isometry condition from [4]: $\delta_S + \theta_{S,S} + \theta_{S,2S} < 1$,

guarantees the exact recovery of optimal solution for $\sigma \leq S$ by solving a convex program, (where $\theta_{S,S}, \theta_{S,2S}$ are other non-negative parameters). Consider the matrix of the form

$$M = \left(D \mid G \right),$$

where D is a n -times- n diagonal matrix, $\mathbf{0}_{n-3,3}$ is the all-zero matrix with $n - 3$ rows and 3 columns, and G is the matrix:

$$G = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}.$$

Then it is easy to see that $\delta_3 \geq 1$ (and thus $\delta_3 + \theta_{3,3} + \theta_{3,6} \geq 1$), since $Gv = 0$ for v being the 3-dimensional all-ones vector. Thus, for such a M matrix the exact recovery results do not hold for $\sigma \geq 3$. In general, since in our result the extra k columns (other than the block diagonal matrix) that are allowed in the M matrix can be arbitrary, it is very easy to destroy any potential restricted isometry condition.

Other computational approaches Due to the vast applicability of Problem (2), many computational approaches are available to solve experimentally instances of (2) or variations of it. Among them are greedy algorithms (e.g., forward- and backward-stepwise selection [31, 14, 8, 9, 15] with provable approximation guarantees, forward-stagewise regression [13, 20]), branch and bound method [2, 24] (e.g., the leaps and bounds procedure [17]), and convex optimization (e.g., ridge regression [25], the lasso [36, 32]). See also [20, 10] for further references on this topic.

1.2 Overview of the algorithm and reduction to separable form

We briefly explain our overall strategy to prove Theorem 1. We first show that it suffices to consider problems of the form

$$\begin{aligned} \min \quad & \left\| Ax - \left(b - \sum_{\ell=1}^k c_\ell \lambda_\ell \right) \right\|^2 \\ \text{s. t.} \quad & x \in \mathbb{R}^n, \lambda \in \mathbb{R}^k \\ & |\text{supp}(x)| \leq \sigma, \end{aligned} \tag{5}$$

where $A \in \mathbb{R}^{m \times n}$ is block diagonal with blocks $A^i \in \mathbb{R}^{m_i \times n_i}$, for $i = 1, \dots, h$, and $b, c_1, \dots, c_k \in \mathbb{R}^m$, $\sigma \in \mathbb{N}$. In particular, $n = \sum_{i=1}^h n_i$. More specifically, to prove Theorem 1, it suffices to give a polynomial-time algorithm for problem (5) under the assumption that k, n_1, \dots, n_h are fixed numbers.

Next, we design an algorithm to solve the problem obtained from (5) by fixing the vector $\lambda \in \mathbb{R}^k$, namely

$$\text{Opt}(\sigma)_{|\lambda} := \min \left\{ \left\| Ax - \left(b - \sum_{\ell=1}^k c_\ell \lambda_\ell \right) \right\|^2 : x \in \mathbb{R}^n, |\text{supp}(x)| \leq \sigma \right\}.$$

This algorithm has the additional property that it only needs to evaluate a polynomial number of inequalities in order to identify an optimal support of the problem.

Our next task is to partition all possible vectors $\lambda \in \mathbb{R}^k$ in a number of regions such that in each region the algorithm yields the same optimal support. Normally, such a partition can be comprised of exponentially many regions, even for a polynomial-time algorithm. However, the special property of our algorithm for $\text{Opt}(\sigma)_{|\lambda}$ allows us to obtain a partition formed only by a polynomial number of regions.

We then construct the set \mathcal{X} of supports that contains, for each region, the corresponding optimal support. In particular, the set \mathcal{X} has the property that there is at least one optimal solution of (5) whose support is contained in some $\chi \in \mathcal{X}$. Finally, for each support $\chi \in \mathcal{X}$, we solve the linear least squares problem obtained from problem (5) by dropping the support constraint and by setting to zero all variables with indices not in χ . All these linear least squares problems can be solved in polynomial time (see, e.g., [34]), and the best solution obtained is optimal to (5).

The approach that we present relies on two key results. The first is an analysis of the proximity of optimal solutions with respect to two consecutive “support-conditions” $|\text{supp}(x)| \leq s$ and $|\text{supp}(x)| \leq s + 1$. This result is fairly technical and is presented in Section 3.1, as it is only needed in the general block diagonal case in order to obtain the algorithm for $\text{Opt}(\sigma)_{|\lambda}$. The second result is a tool from Discrete Geometry known as the *hyperplane arrangement theorem*. We now briefly describe this theorem, which is needed to upper bound the number of regions in the partition of the λ -space that we construct. A set H of n hyperplanes in a k -dimensional Euclidean space determine a partition of the space called the *arrangement of H* , or the *cell complex induced by H* . The hyperplane arrangement theorem states that this arrangement consists of $O(n^k)$ polyhedra and can be constructed with $O(n^k)$ operations. For more details, we refer the reader to the paper [12], and in particular to Theorem 3.3 therein.

In the remainder of this section we show the reduction to form (5). Problem (5) has a key benefit over the original Subset Selection problem (2) with matrix M of the form (1): the support constraint is applied only to the variables associated with the columns of A . A key consequence is that, once we fix the variables that are not subject to support constraints, then the objective function is decomposable in a sum of quadratic functions, and each one depends only on the components of x corresponding to a single block of A . This property will be crucial in the design of our algorithm.

Lemma 1. *To prove Theorem 1, it suffices to give a polynomial-time algorithm for problem (5), provided that k, n_1, \dots, n_h are fixed numbers.*

Proof. Consider problem (2) where the matrix M is of the form (1). Recall that k is the number of columns of C , and that n_i is the number of columns of block A^i , for $i = 1, \dots, h$. Let $n := \sum_{i=1}^h n_i$ be the number of columns of A . The structure of the matrix M implies that the objective function of problem (2) can be written in the form

$$\left\| Ax - \left(b - c\mu - \sum_{\ell=1}^k c_\ell x_{n+\ell} \right) \right\|,$$

For each subset \mathcal{L} of $\{1, \dots, k\}$, we consider the subproblem obtained from (2) by setting $x_{n+\ell} = 0$ for all $\ell \in \{1, \dots, k\} \setminus \mathcal{L}$, and by restricting the cardinality constraint to the n -dimensional vector (x_1, \dots, x_n) . Formally,

$$\begin{aligned} \min \quad & \left\| Ax - \left(b - c\mu - \sum_{\ell \in \mathcal{L}} c_\ell x_{n+\ell} \right) \right\| \\ \text{s. t.} \quad & x_1, \dots, x_n \in \mathbb{R}, \quad x_{n+\ell} \in \mathbb{R}, \forall \ell \in \mathcal{L}, \quad \mu \in \mathbb{R} \\ & |\text{supp}(x_1, \dots, x_n)| \leq \sigma - |\mathcal{L}|. \end{aligned} \tag{6}$$

We claim that to solve problem (2) we just need to solve the 2^k distinct subproblems of the form (6) and return the best among the 2^k optimal solutions found. To see this, let $(\bar{x}, \bar{\mu})$ be feasible for (2) and consider the subproblem (6) corresponding to the set $\mathcal{L} := \{\ell \in \{1, \dots, k\} : \bar{x}_{n+\ell} \neq 0\}$. Then the restriction of $(\bar{x}, \bar{\mu})$ obtained by dropping the zero components $\bar{x}_{n+\ell}$, for $\ell \in \{1, \dots, k\} \setminus \mathcal{L}$, is feasible for this subproblem and has the same objective value. Now let $(\hat{x}, \hat{\mu})$ be feasible for a subproblem (6) corresponding to a subset \mathcal{L} of $\{1, \dots, k\}$. Consider now the extension of $(\hat{x}, \hat{\mu})$ obtained by adding components $\hat{x}_{n+\ell} = 0$ for all $\ell \in \{1, \dots, k\} \setminus \mathcal{L}$. This vector is feasible for (2) and has the same objective value. This concludes the proof of the equivalence of (2) with all the 2^k subproblems (6).

Notice that, the assumption in Theorem 1 that each of the blocks A^1, \dots, A^h and C have a constant number of columns means that in each subproblem (6), n_1, \dots, n_h and $|\mathcal{L}|$ are fixed numbers.

Consider now a subproblem (6). Let $\mathcal{L}' := \mathcal{L} \cup \{0\}$, and $k' := |\mathcal{L}'| = |\mathcal{L}| + 1$. We introduce a k' -dimensional vector λ as

$$\lambda_\ell := \begin{cases} \mu & \text{if } \ell = 0, \\ x_{n+\ell} & \text{if } \ell \in \mathcal{L}, \end{cases}$$

and we define k' vectors $c'_\ell \in \mathbb{R}^m$, for $\ell \in \mathcal{L}'$, as

$$c'_\ell := \begin{cases} c & \text{if } \ell = 0, \\ c_\ell & \text{if } \ell \in \mathcal{L}. \end{cases}$$

Finally, we define $\sigma' := \sigma - |\mathcal{L}|$. It follows that a subproblem (6) can be written in the form (5), where the parameters k, σ, c_ℓ in (5) correspond to the parameters k', σ', c'_ℓ just defined. \square

To avoid confusion, we remark that the parameters k, σ, c^ℓ in (5) do not need to match the corresponding parameters k, σ, c^ℓ in model (2).

2 The diagonal case

In this section we prove a weaker version of Theorem 1 obtained by further assuming that the block A of the matrix M given by (1) is diagonal, i.e., each A^i block is just a 1×1 -matrix. Then $A \in \mathbb{R}^{n \times n}$, and we denote by a_{11}, \dots, a_{nn} the diagonal entries of A .

Theorem 2. *There is an algorithm that solves problem (2) in polynomial time, provided that M is of the form (1), A is diagonal, and k is a fixed number.*

In particular, our algorithm solves $O(2^k n^{2(k+1)})$ linear least squares problems. The presence of k in the exponent is expected, since the problem is NP-hard if k is not a constant but a parameter which is part of the input.

For brevity, in this section, every time we refer to problem (2), problem (5), and problem $\text{Opt}(\sigma)_{|\lambda}$, we assume that the matrices M and A satisfy the assumptions of Theorem 2.

Instead of considering directly problem (5), we first consider the simpler problem $\text{Opt}(\sigma)_{|\lambda}$. For notational simplicity, we define the vector

$$b' := b - \sum_{\ell=1}^k c_\ell \lambda_\ell.$$

Lemma 2. *Problem $\text{Opt}(\sigma)_{|\lambda}$ can be solved in polynomial time. In particular, an optimal solution is given by*

$$x_j^* := \begin{cases} b'_j/a_{jj} & \text{if } j \in K \\ 0 & \text{if } j \notin K, \end{cases}$$

where K is a subset of $H := \{i \in \{1, \dots, n\} : a_{ii} \neq 0\}$ of cardinality $\max\{\sigma, |H|\}$ with the property that for each $i \in K$ and each $j \in H \setminus K$, we have $|b'_i| \geq |b'_j|$.

Proof. The objective function of Problem $\text{Opt}(\sigma)_{|\lambda}$ is $\sum_{j=1}^n (a_{jj}x_j - b'_j)^2$. The separability of the objective function implies that the objective value of a vector x will be at least $\sum_{j \notin H} (b'_j)^2 + \sum_{j \in H \setminus \text{supp}(x)} (b'_j)^2$. The support constraint and the definition of the index set K then imply that the optimum value of $\text{Opt}(\sigma)_{|\lambda}$ is at least $\sum_{j \notin H} (b'_j)^2 + \sum_{j \in H \setminus K} (b'_j)^2$. The latter is the objective value of the vector x^* , thus it is an optimal solution to $\text{Opt}(\sigma)_{|\lambda}$. \square

We are now ready to prove Theorem 2.

Proof of Theorem 2. Lemma 1 implies that, in order to prove Theorem 2, it suffices to give a polynomial-time algorithm for problem (5), where A is diagonal, provided that k is a fixed number. In the remainder of the proof we show how to solve problem (5).

Claim 1. *We can construct in polynomial time an index set T with $|T| = O(n^{2k})$, polyhedra $Q^t \subseteq \mathbb{R}^k$, for $t \in T$, that cover \mathbb{R}^k , and sets $\chi^t \subseteq \{1, \dots, n\}$ of cardinality σ , for each $t \in T$, with the following property: For every $t \in T$, and for every λ such that $\lambda \in Q^t$, the problem $\text{Opt}(\sigma)_{|\lambda}$ has an optimal solution with support contained in χ^t .*

Proof of claim. Lemma 2 implies that in order to understand the optimal support of problem $\text{Opt}(\sigma)_{|\lambda}$ for a fixed vector λ , it is sufficient to compare all quantities $|b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell|$, for $i \in \{1, \dots, n\}$ with $a_{ii} \neq 0$. So let i and j be two distinct indices in $\{1, \dots, n\}$ with $a_{ii}, a_{jj} \neq 0$. We wish to subdivide all points $\lambda \in \mathbb{R}^k$ based on which of the two quantities $|b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell|$ and $|b_j - \sum_{\ell=1}^k (c_\ell)_j \lambda_\ell|$ is larger. In order to do so, consider the inequality

$$\left| b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell \right| \geq \left| b_j - \sum_{\ell=1}^k (c_\ell)_j \lambda_\ell \right|.$$

It is simple to check that the set of points in \mathbb{R}^k that satisfy the above inequality can be written as the union of polyhedra using linear inequalities corresponding to the four hyperplanes in \mathbb{R}^k defined by equations

$$\begin{aligned} b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell &= 0, & b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell &= b_j - \sum_{\ell=1}^k (c_\ell)_j \lambda_\ell, \\ b_j - \sum_{\ell=1}^k (c_\ell)_j \lambda_\ell &= 0, & b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell &= - \left(b_j - \sum_{\ell=1}^k (c_\ell)_j \lambda_\ell \right). \end{aligned}$$

By considering these four hyperplanes for all possible distinct pairs of indices in $\{1, \dots, n\}$, we obtain at most $4(n^2 - n) = O(n^2)$ hyperplanes in \mathbb{R}^k . These hyperplanes subdivide \mathbb{R}^k into a number of full-dimensional polyhedra. By the hyperplane arrangement theorem, this subdivision consists of polyhedra Q^t , for $t \in T$, where T is an index set with $|T| = O((n^2)^k) = O(n^{2k})$. Since k is fixed, $|T|$ is polynomial in n and the subdivision can be obtained in polynomial time.

We now fix one polyhedron Q^t , for some $t \in T$. By checking, for each hyperplane that we have constructed above, in which of the two half-spaces lies Q^t , we obtain a total order on all the expressions $|b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell|$, for $i \in \{1, \dots, n\}$ with $a_{ii} \neq 0$. The obtained total order is global, in the sense that, for each fixed λ with $\lambda \in Q^t$, it induces a consistent total order on the values obtained by fixing λ in the expressions $|b_i - \sum_{\ell=1}^k (c_\ell)_i \lambda_\ell|$. This total order induces an ordering $i_1^t, i_2^t, \dots, i_n^t$ of the indices $i \in \{1, \dots, n\}$ with $a_{ii} \neq 0$ such that, for every $\lambda \in Q^t$, we have

$$\left| b_{i_1^t} - \sum_{\ell=1}^k (c_\ell)_{i_1^t} \lambda_\ell \right| \geq \left| b_{i_2^t} - \sum_{\ell=1}^k (c_\ell)_{i_2^t} \lambda_\ell \right| \geq \dots \geq \left| b_{i_n^t} - \sum_{\ell=1}^k (c_\ell)_{i_n^t} \lambda_\ell \right|.$$

By Lemma 2, for each λ such that $\lambda \in Q^t$, the problem $\text{Opt}(\sigma)_{|\lambda}$ has an optimal solution with support contained in $\chi^t := \{i_1^t, i_2^t, \dots, i_\sigma^t\}$. \diamond

Let \mathcal{X} be the set containing all index sets χ^t obtained in Claim 1, namely

$$\mathcal{X} := \{\chi^t : t \in T\}.$$

Claim 2. *There exists an optimal solution (x^*, λ^*) of problem (5) such that*

$$\text{supp}(x^*) \subseteq \chi \text{ for some } \chi \in \mathcal{X}.$$

Proof of claim. Let (x^*, λ^*) be an optimal solution of problem (5). Then x^* is an optimal solution of the restricted problem $\text{Opt}(\sigma)_{|\lambda^*}$. Let Q^t , for $t \in T$, be a polyhedron such that $\lambda^* \in Q^t$, and let $\chi^t \in \mathcal{X}$ be the corresponding index set. By Claim 1, the problem $\text{Opt}(\sigma)_{|\lambda^*}$ has an optimal solution \tilde{x} with support contained in χ^t . This implies that the solution (\tilde{x}, λ^*) is also optimal for problem (5). \diamond

For each $\chi \in \mathcal{X}$, each problem (5), with the additional constraints $x_i = 0$, for all $i \notin \chi$, is a linear least squares problem, since the cardinality constraint can be dropped, and it can then be solved in polynomial time. The best solution among the obtained ones is an optimal solution of (5). To find an optimal solution to (5), we have solved $O(n^{2k})$ linear least squares problems.

Since in the proof of Lemma 1 we have $k' \leq k + 1$, to find an optimal solution to (2) we need to solve $O(2^k n^{2(k+1)})$ linear least squares problems. This concludes the proof of Theorem 2. \square

Note that the algorithm for $\text{Opt}(\sigma)_{|\lambda}$ given in Lemma 2 plays a key role in the proof of Theorem 2. In fact, it allowed us to subdivide the space of the λ variables in a polynomial number of regions such that in each region the algorithm given in Lemma 2 yields the same optimal support. This part of the proof will be much more involved for Theorem 1. In fact, in the proof of our main result, in order to derive a similar subdivision, we need to study the proximity of optimal solutions with respect to two consecutive “support-conditions” $|\text{supp}(x)| \leq s$ and $|\text{supp}(x)| \leq s + 1$.

3 The block diagonal case

In this section we prove Theorem 1. Every time we refer to problem (2), problem (5), and problem $\text{Opt}(\sigma)_{|\lambda}$, we assume that the matrices M and A satisfy the assumptions of Theorem 1. In particular, the matrix M is as in (1), and the matrix $A \in \mathbb{R}^{m \times n}$ is block diagonal with blocks $A^i \in \mathbb{R}^{m_i \times n_i}$, for $i = 1, \dots, h$.

Our overall strategy is similar to the one we used to prove Theorem 2 and is described in Section 1.2. Our first task is the design of a polynomial-time algorithm to solve the simpler problem $\text{Opt}(\sigma)_{|\lambda}$.

Note that there are many possible polynomial-time algorithms that one can devise for problem $\text{Opt}(\sigma)_{|\lambda}$, and a particularly elegant one can be obtained with a dynamic programming approach similar to the classic dynamic programming recursion for knapsack [30]. For each $i = 1, \dots, h$, let $x^i \in \mathbb{R}^{n_i}$, $b^i \in \mathbb{R}^{m_i}$, and $c_\ell^i \in \mathbb{R}^{m_i} \forall \ell = 1, \dots, k$ such that

$$x = \begin{pmatrix} x^1 \\ \vdots \\ x^h \end{pmatrix}, \quad b = \begin{pmatrix} b^1 \\ \vdots \\ b^h \end{pmatrix}, \quad c_\ell = \begin{pmatrix} c_\ell^1 \\ \vdots \\ c_\ell^h \end{pmatrix}.$$

Denote by $\text{Opt}(1, \dots, i; j)$ the subproblem of $\text{Opt}(\sigma)_{|\lambda}$ on blocks A^1, \dots, A^i , with $i \in \{1, \dots, h\}$ and support j , with $j \in \{0, \dots, \sigma\}$, namely,

$$\begin{aligned} \min \quad & \left\| \begin{pmatrix} A^1 & & \\ & \ddots & \\ & & A^i \end{pmatrix} x - \left(\begin{pmatrix} b^1 \\ \vdots \\ b^i \end{pmatrix} - \sum_{\ell=1}^k \begin{pmatrix} c_\ell^1 \\ \vdots \\ c_\ell^i \end{pmatrix} \lambda_\ell \right) \right\|^2 \\ \text{s. t.} \quad & x \in \mathbb{R}^{n_1 + \dots + n_i} \\ & |\text{supp}(x)| \leq j. \end{aligned}$$

By exploiting the separability of the objective function, it can be checked that the recursion

$$\text{Opt}(1, \dots, i; j) = \min_{t=0, \dots, j} \{ \text{Opt}(1, \dots, i-1; j-t) + \text{Opt}(i; t) \}, \quad \forall j \in \{0, \dots, \sigma\},$$

for $i = 1, \dots, h$, yields an optimal solution for problem $\text{Opt}(\sigma)_{|\lambda}$.

We view this dynamic programming recursion as a ‘‘horizontal approach’’ where we fix the support but then proceed a recursion over the blocks. We did not see how this algorithm can be used in the proof of Theorem 1 to cover the space of the λ variables with a sub-exponential number of regions such that in each region the algorithm yields the same optimal support. In the next section, we present a combinatorial algorithm that not only can solve problem $\text{Opt}(\sigma)_{|\lambda}$, but that also allows us to obtain a polynomial covering of the space of the λ variables. This algorithm is based on a sort of ‘‘vertical approach’’ where we incorporate all blocks simultaneously and then develop a recursion to enlarge the support condition.

3.1 A proximity theorem

For every $i \in \{1, \dots, h\}$, and every $j \in \{0, \dots, n_i\}$, let $\text{Val}(i; j)$ be a real number. For $\sigma \in \{0, \dots, \sum_{i=1}^h n_i\}$, we define

$$\text{Val}(\sigma) := \min \left\{ \sum_{i=1}^h \text{Val}(i; j_i) : \sum_{i=1}^h j_i = \sigma, j_i \in \{0, \dots, n_i\} \right\}. \quad (7)$$

Note that problem $\text{Opt}(\sigma)_{|\lambda}$ can be polynomially transformed to a problem $\text{Val}(\sigma)$. This can be seen by exploiting the block diagonal structure of the matrix A , and by defining $\text{Val}(i; j_i)$ to be the optimal value of the problem restricted to block i and support j_i . The details of this reduction are given at the beginning of the proof of Theorem 1.

Definition 1. Let s and q be nonnegative integers. Given a feasible solution $j^s = (j_1^s, \dots, j_h^s)$ for $\text{Val}(s)$, we say that a feasible solution $j^{s+1} = (j_1^{s+1}, \dots, j_h^{s+1})$ for $\text{Val}(s+1)$ is q -close to j^s if

$$\sum_{i \in I^-} (j_i^s - j_i^{s+1}) = q,$$

where $I^- := \{i \in \{1, \dots, h\} : j_i^{s+1} < j_i^s\}$.

Clearly, if j^{s+1} is q -close to j^s , we also have that

$$\sum_{i \in I^+} (j_i^{s+1} - j_i^s) = q + 1,$$

where $I^+ := \{i \in \{1, \dots, h\} : j_i^{s+1} > j_i^s\}$. This yields to a l_1 -proximity bound of $\|j^{s+1} - j^s\|_1 \leq 2q + 1$.

A *weak composition* of an integer q into p parts is a sequence of p non-negative integers that sum up to q . Two sequences that differ in the order of their terms define different weak compositions. It is well-known that the number of weak compositions of a number q into p parts is $\binom{q+p-1}{p-1} = \binom{q+p-1}{q}$. For more details on weak compositions see, for example, [23].

Our next result establishes that optimal solutions for $\text{Val}(s)$ and $\text{Val}(s+1)$ are close to each other.

Lemma 3 (Proximity of optimal solutions). *Given an optimal solution j^s for $\text{Val}(s)$, there exists an optimal solution j^{s+1} for $\text{Val}(s+1)$ that is q -close to j^s , for some $q \in \{0, \dots, (\theta-1)\theta(\theta+1)/2\}$, where $\theta := \max\{n_i : i = 1, \dots, h\}$.*

Proof. Let j^{s+1} be an optimal solution for $\text{Val}(s+1)$ such that

$$\sum_{i=1}^h |j_i^{s+1} - j_i^s| \text{ is minimal.} \quad (8)$$

Let $I^+ := \{i \in \{1, \dots, h\} : j_i^{s+1} > j_i^s\}$ and $I^- := \{i \in \{1, \dots, h\} : j_i^{s+1} < j_i^s\}$. Note that

$$\sum_{i \in I^+} (j_i^{s+1} - j_i^s) = \sum_{i \in I^-} (j_i^s - j_i^{s+1}) + 1.$$

For $p = 1, \dots, \theta$, let

$$\begin{aligned} x_p &:= |\{i \in I^+ : j_i^{s+1} - j_i^s = p\}| \\ y_p &:= |\{i \in I^- : j_i^s - j_i^{s+1} = p\}|, \end{aligned}$$

This yields to the two equations

$$\sum_{i \in I^+} (j_i^{s+1} - j_i^s) = \sum_{p=1}^{\theta} p x_p \quad \text{and} \quad \sum_{i \in I^-} (j_i^s - j_i^{s+1}) = \sum_{p=1}^{\theta} p y_p.$$

If $\sum_{p=1}^{\theta} p y_p \leq (\theta-1)\theta(\theta+1)/2$, then the statement is verified. Otherwise, $\sum_{p=1}^{\theta} p y_p > (\theta-1)\theta(\theta+1)/2$. This implies that there exists $v \in \{1, \dots, \theta\}$ with $y_v \geq \theta$. To see this, note that if $y_p \leq \theta-1$, for all $p = 1, \dots, \theta$, then we obtain

$$\sum_{p=1}^{\theta} p y_p \leq (\theta-1) \sum_{p=1}^{\theta} p = (\theta-1)\theta(\theta+1)/2.$$

From the fact that $\sum_{p=1}^{\theta} px_p = \sum_{p=1}^{\theta} py_p + 1 > (\theta - 1)\theta(\theta + 1)/2$, it also follows that there exists $u \in \{1, \dots, \theta\}$ with $x_u \geq \theta$.

In particular we have that

$$x_u \geq \theta \geq v \text{ and } y_v \geq u.$$

Thus there exists a subset \tilde{I}^+ of I^+ such that $|\tilde{I}^+| = v$, and $j_i^{s+1} - j_i^s = u$ for all $i \in \tilde{I}^+$. Symmetrically, there exists a subset \tilde{I}^- of I^- such that $|\tilde{I}^-| = u$, and $j_i^s - j_i^{s+1} = v$ for all $i \in \tilde{I}^-$.

Let \tilde{j}^{s+1} be obtained from j^{s+1} as follows

$$\tilde{j}_i^{s+1} = \begin{cases} j_i^{s+1} - u = j_i^s & \text{if } i \in \tilde{I}^+ \\ j_i^{s+1} + v = j_i^s & \text{if } i \in \tilde{I}^- \\ j_i^{s+1} & \text{otherwise.} \end{cases}$$

Since $\sum_{i=1}^h \tilde{j}_i^{s+1} = \sum_{i=1}^h j_i^{s+1} - uv + uv = s + 1$, we have that \tilde{j}^{s+1} is a feasible solution for $\text{Val}(s + 1)$. Moreover, we have that

$$\sum_{i=1}^h |\tilde{j}_i^{s+1} - j_i^s| = \sum_{i=1}^h |j_i^{s+1} - j_i^s| - 2uv < \sum_{i=1}^h |j_i^{s+1} - j_i^s|.$$

In the remainder of the proof, we show that \tilde{j}^{s+1} is an optimal solution for $\text{Val}(s + 1)$. This will conclude the proof, since it contradicts the choice of j^{s+1} in Eq. (8).

Let \tilde{j}^s be obtained from j^s as follows

$$\tilde{j}_i^s = \begin{cases} j_i^s + u = j_i^{s+1} & \text{if } i \in \tilde{I}^+ \\ j_i^s - v = j_i^{s+1} & \text{if } i \in \tilde{I}^- \\ j_i^s & \text{otherwise.} \end{cases}$$

Since $\sum_{i=1}^h \tilde{j}_i^s = \sum_{i=1}^h j_i^s + uv - uv = s$, we have that \tilde{j}^s is a feasible solution for $\text{Val}(s)$. Since j^s is an optimal solution for $\text{Val}(s)$ we obtain

$$\sum_{i=1}^h \text{Val}(i; \tilde{j}_i^s) - \sum_{i=1}^h \text{Val}(i; j_i^s) = \sum_{i \in \tilde{I}^+ \cup \tilde{I}^-} (\text{Val}(i; j_i^{s+1}) - \text{Val}(i; j_i^s)) \geq 0.$$

Consider now the objective value of the feasible solution \tilde{j}^{s+1} for $\text{Val}(s + 1)$.

$$\begin{aligned} \sum_{i=1}^h \text{Val}(i; \tilde{j}_i^{s+1}) &= \sum_{i=1}^h \text{Val}(i; j_i^{s+1}) + \sum_{i \in \tilde{I}^+ \cup \tilde{I}^-} (\text{Val}(i; j_i^s) - \text{Val}(i; j_i^{s+1})) \\ &\leq \sum_{i=1}^h \text{Val}(i; j_i^{s+1}). \end{aligned}$$

This shows that the solution \tilde{j}^{s+1} is optimal for $\text{Val}(s + 1)$. □

Assume now that we know an optimal solution j^s for $\text{Val}(s)$ and we wish to obtain an optimal solution for $\text{Val}(s + 1)$. By Lemma 3, we just need to consider the feasible solutions for $\text{Val}(s + 1)$

that are q -close to j^s , for some $q \leq (\theta - 1)\theta(\theta + 1)/2 =: \bar{\theta}$. We denote the family of these solutions by $\text{Aug}(j^s)$, formally

$$\text{Aug}(j^s) := \left\{ j^{s+1} : j_i^{s+1} \in \{0, \dots, n_i\}, i \in \{1, \dots, h\}, \sum_{i=1}^h j_i^{s+1} = s + 1, \right. \\ \left. j^{s+1} \text{ is } q\text{-close to } j^s \text{ for some } q \leq \bar{\theta} \right\}. \quad (9)$$

For each solution $j^{s+1} \in \text{Aug}(j^s)$, the corresponding objective function value is obtained from $\text{Val}(s)$ by adding the difference,

$$d(j^s, j^{s+1}) := \sum_{i \in I^+ \cup I^-} (\text{Val}(i; j_i^{s+1}) - \text{Val}(i; j_i^s)). \quad (10)$$

We denote by $\mathcal{D}(j^s)$ the family of the values $d(j^s, j^{s+1})$, for each solution $j^{s+1} \in \text{Aug}(j^s)$,

$$\mathcal{D}(j^s) := \{d(j^s, j^{s+1}) : j^{s+1} \in \text{Aug}(j^s)\}.$$

From our discussions it follows that, in order to select the optimal solution for $\text{Val}(s + 1)$, we only need to know which value in $\mathcal{D}(j^s)$ is the smallest.

We next define the set \mathcal{D} as the union of all sets $\mathcal{D}(j^s)$, for any feasible solution j^s of problem $\text{Val}(s)$, for any $s \in \{0, \dots, \sum_{i=1}^h n_i\}$. Formally, the set \mathcal{D} is defined as

$$\mathcal{D} := \left\{ d(j^s, j^{s+1}) : j^s \text{ is a feasible solution for } \text{Val}(s), \text{ for some } s \in \{0, \dots, \sum_{i=1}^h n_i\}, \right. \\ \left. j^{s+1} \in \text{Aug}(j^s) \right\}. \quad (11)$$

The next result implies that the set \mathcal{D} contains a number of values that is polynomial in h , provided that θ is fixed. This fact could be surprising on the first glance since the number of feasible solutions j^s for $\text{Val}(s)$ is of exponential order θ^h .

Lemma 4. *The set \mathcal{D} contains $O((\bar{\theta} + h)^{\bar{\theta}+1}\theta^{2\bar{\theta}+1})$ values.*

Proof. We count the number of all possible values $d(j^s, j^{s+1})$ in \mathcal{D} . Fix $q \in \{0, \dots, \bar{\theta}\}$, and let us consider all the values $d(j^s, j^{s+1})$ corresponding to a feasible solution j^s for $\text{Val}(s)$ and a feasible solution j^{s+1} for $\text{Val}(s + 1)$ such that j^{s+1} is q -close to j^s . First, we construct the possible positive differences $d_i := j_i^{s+1} - j_i^s$, for $i \in I^+$. Since the total sum of differences d_i , for $i \in I^+$, equals $q + 1$, and the number of weak compositions of $q + 1$ into h parts is $\binom{q+h}{q+1}$, we conclude that there are $O((q + h)^{q+1})$ ways of choosing the set I^+ and constructing differences $j_i^{s+1} - j_i^s$, for $i \in I^+$. Similarly, there are $O((q + h)^q)$ ways of choosing the set I^- and constructing differences $j_i^s - j_i^{s+1}$, for $i \in I^-$. For each $i \in I^+ \cup I^-$, there are at most θ possible indices j_i^s , yielding a total number of $\theta^{|I^+ \cup I^-|} \leq \theta^{2q+1}$ possible set of indices j_i^s , for $i \in I^+ \cup I^-$. In total, we obtain $O((q + h)^{q+1}\theta^{2q+1})$ possible values.

From the fact that $q \in \{0, \dots, \bar{\theta}\}$, it follows that the cardinality of the set \mathcal{D} is bounded by $O((\bar{\theta} + h)^{\bar{\theta}+1}\theta^{2\bar{\theta}+1})$. \square

Proposition 1. *Given a total order on all the values in \mathcal{D} , we can construct an optimal solution for $\text{Val}(\sigma)$, for any $\sigma \in \{0, \dots, \sum_{i=1}^h n_i\}$, in time $O(\sigma(\bar{\theta} + h)^{\bar{\theta}+1}\theta^{2\bar{\theta}+1})$.*

Proof. An optimal solution for $\text{Val}(0)$ is $j^0 = (0, \dots, 0)$. Let $s \in \{0, \dots, \sum_{i=1}^h n_i\}$ and assume that we have an optimal solution j^s for $\text{Val}(s)$. We show how we can construct an optimal solution j^{s+1} for $\text{Val}(s+1)$. Consider all the values in $\mathcal{D}(j^s)$. Since $\mathcal{D}(j^s) \subseteq \mathcal{D}$, we can inquire a total order of $\mathcal{D}(j^s)$ from the given total order of \mathcal{D} . Thus $\mathcal{D}(j^s)$ has a minimum element. Since a minimum element can be found in linear time with respect to the cardinality of the set, we obtain a running time $O(s(\bar{\theta} + h)^{\bar{\theta}+1}\theta^{2\bar{\theta}+1})$ as a consequence of Lemma 4. In view of Lemma 3, the solution in $\text{Aug}(j^s)$ corresponding to the minimum element in $\mathcal{D}(j^s)$ is an optimal solution j^{s+1} for $\text{Val}(s+1)$. This argument applied in an inductive manner leads to an optimal solution for $\text{Val}(\sigma)$ for any $\sigma \in \{0, \dots, \sum_{i=1}^h n_i\}$. \square

3.2 Proof of main result

We are now ready to present the formal proof of our main theorem.

Proof of Theorem 1. From Lemma 1, in order to prove Theorem 1, we only need to show that there is a polynomial-time algorithm that solves problem (5), where $A \in \mathbb{R}^{m \times n}$ is block diagonal with blocks $A^i \in \mathbb{R}^{m_i \times n_i}$, for $i = 1, \dots, h$, provided that k, n_1, \dots, n_h are fixed numbers. In the remainder of the proof we describe our algorithm to solve problem (5).

First, we consider problem $\text{Opt}(\sigma)_{|\lambda}$ and we rewrite it by exploiting the separability of the objective function. For each $i = 1, \dots, h$, let $x^i \in \mathbb{R}^{n_i}$, $b^i \in \mathbb{R}^{m_i}$, and $c_\ell^i \in \mathbb{R}^{m_i} \forall \ell = 1, \dots, k$ such that

$$x = \begin{pmatrix} x^1 \\ \vdots \\ x^h \end{pmatrix}, \quad b = \begin{pmatrix} b^1 \\ \vdots \\ b^h \end{pmatrix}, \quad c_\ell = \begin{pmatrix} c_\ell^1 \\ \vdots \\ c_\ell^h \end{pmatrix}.$$

Define the subproblem of $\text{Opt}(\sigma)_{|\lambda}$ on block A^i , with $i \in \{1, \dots, h\}$, and support j , for $j \in \{0, \dots, n_i\}$. Formally,

$$\text{Opt}(i; j)_{|\lambda} := \min \left\{ \left\| A^i x^i - \left(b^i - \sum_{\ell=1}^k c_\ell^i \lambda_\ell \right) \right\|^2 : x^i \in \mathbb{R}^{n_i}, |\text{supp}(x^i)| \leq j \right\}.$$

We can finally rewrite $\text{Opt}(\sigma)_{|\lambda}$ in the form

$$\text{Opt}(\sigma)_{|\lambda} = \min \left\{ \sum_{i=1}^h \text{Opt}(i; j_i)_{|\lambda} : \sum_{i=1}^h j_i = \sigma, j_i \in \{0, \dots, n_i\} \right\}. \quad (12)$$

Note that in this new form, the decision variables are the integers j_i , for $i = 1, \dots, h$, and the variables x do not appear explicitly. We observe that we have reduced ourselves to the same setting described in Section 3.1. In fact, for each fixed λ , each $\text{Opt}(i; j_i)_{|\lambda}$ can be calculated by solving a fixed number of linear least squares problems. Thus, problem $\text{Opt}(\sigma)_{|\lambda}$ is now a problem of the form $\text{Val}(\sigma)$, as defined in (7) and can be solved efficiently as a consequence of Proposition 1. Hence, problem $\text{Opt}(\sigma)_{|\lambda}$ can be solved for each fixed λ . However, in order to solve our original problem, we have to solve $\text{Opt}(\sigma)_{|\lambda}$ for every $\lambda \in \mathbb{R}^k$. In order to do this, we now think of $\text{Opt}(\sigma)_{|\lambda}$ and $\text{Opt}(i; j)_{|\lambda}$ as functions that associate to each $\lambda \in \mathbb{R}^k$ a real number.

Next, we define a space \mathcal{S} that is an extended version of the space \mathbb{R}^k of variables λ_ℓ , for $\ell = 1, \dots, k$. The space \mathcal{S} contains all the variables λ_ℓ , for $\ell = 1, \dots, k$, and it also contains one variable for each product of two variables $\lambda_{\ell_1} \lambda_{\ell_2}$, with $\ell_1, \ell_2 \in \{1, \dots, k\}$. The dimension of the space \mathcal{S} is therefore $(k^2 + k)/2 \leq k^2$. Note that, for each $\lambda \in \mathbb{R}^k$, there exists a unique corresponding point in \mathcal{S} , that we denote by $\text{ext}(\lambda)$, obtained by computing all the products $\lambda_{\ell_1} \lambda_{\ell_2}$, for $\ell_1, \ell_2 \in \{1, \dots, k\}$.

Claim 3. We can construct in polynomial time an index set T with $|T| = O(h^{k^2})$, polyhedra $P^t \subseteq \mathcal{S}$, for $t \in T$, that cover \mathcal{S} , and sets $v^t(i; j) \subseteq \{1, \dots, n_i\}$ of cardinality j , for each $i \in \{1, \dots, h\}$, $j \in \{0, \dots, n_i\}$, and $t \in T$, with the following property: For every $i \in \{1, \dots, h\}$, $j \in \{0, \dots, n_i\}$, $t \in T$, and for every λ such that $\text{ext}(\lambda) \in P^t$, the problem $\text{Opt}(i; j)_{|\lambda}$ has an optimal solution with support contained in $v^t(i; j)$.

Proof of claim. Let $i \in \{1, \dots, h\}$. First, we show that for every index set $v \subseteq \{1, \dots, n_i\}$, the best solution for problem $\text{Opt}(i; j)_{|\lambda}$ with support v has an objective value that is a quadratic function in λ . To see this, let L be the linear subspace of \mathbb{R}^{m_i} defined by $L := \{A^i x^i : \text{supp}(x^i) \subseteq v\}$. Consider the affine linear function $p : \mathbb{R}^k \rightarrow \mathbb{R}^{m_i}$ defined by $p^i(\lambda) := b^i - \sum_{\ell=1}^k c_\ell^i \lambda_\ell$. Then the best solution for problem $\text{Opt}(i; j)_{|\lambda}$ with support v has objective value

$$\text{dist}^2(L, p^i(\lambda)) = \|\text{proj}_L(p^i(\lambda)) - p^i(\lambda)\|^2. \quad (13)$$

The projection $\text{proj}_L(p^i(\lambda))$ can be written as a linear function in λ . In order to see this, as a consequence of the Gram-Schmidt orthogonalization, we can assume that the columns of A^i are pairwise orthogonal. The projection of $p^i(\lambda)$ onto L is simply the sum of scalar products of $p^i(\lambda)$ with the columns of the matrix A^i , which is a linear function. Therefore the expression on the right hand side of (13) is a quadratic function in λ .

Let $j \in \{0, \dots, n_i\}$, and let v^1, v^2 be two different index sets contained in $\{1, \dots, n_i\}$ of cardinality j . We wish to obtain a hyperplane that subdivides all points $\text{ext}(\lambda) \in \mathcal{S}$ based on which of the two supports v^1 and v^2 yields a better solution for the problem $\text{Opt}(i; j)_{|\lambda}$. To this end consider the equation whose left-hand side and right-hand side are two expressions of the type (13) corresponding to the two index sets v^1 and v^2 , namely

$$\|\text{proj}_{L^1}(p^i(\lambda)) - p^i(\lambda)\|^2 = \|\text{proj}_{L^2}(p^i(\lambda)) - p^i(\lambda)\|^2,$$

where $L^\beta := \{A^i x^i : \text{supp}(x^i) \subseteq v^\beta\}$ for $\beta \in \{1, 2\}$. Our argument implies that this is a quadratic equation. Thus, by linearizing all the quadratic terms, we obtain a hyperplane in the space \mathcal{S} . As desired, this hyperplane subdivides all points $\text{ext}(\lambda) \in \mathcal{S}$ based on which of the two supports yields a better solution for the problem $\text{Opt}(i; j)_{|\lambda}$.

By considering the hyperplanes of this form corresponding to all possible distinct pairs of index sets in $\{1, \dots, n_i\}$ of cardinality j , we obtain fewer than $\binom{n_i}{j}^2$ hyperplanes in \mathcal{S} . By considering these hyperplanes for all possible $j \in \{0, \dots, n_i\}$, we obtain at most 2^{2n_i} hyperplanes in \mathcal{S} , which is a fixed number since n_i is fixed. Then, by considering all $i \in \{1, \dots, h\}$, we obtain at most $\sum_{i=1}^h 2^{2n_i} \leq h \max\{2^{2n_i} : i = 1, \dots, h\} = O(h)$ hyperplanes in \mathcal{S} . These hyperplanes subdivide \mathcal{S} into a number of full-dimensional polyhedra. By the hyperplane arrangement theorem [12], this subdivision consists of at most $O(h^{\dim(\mathcal{S})}) = O(h^{k^2})$ polyhedra P^t , for $t \in T$. Since k is fixed, $|T|$ is polynomial in h and the subdivision can be obtained in polynomial time.

Let $i \in \{1, \dots, h\}$, $j \in \{0, \dots, n_i\}$, and $t \in T$. By checking, for each hyperplane corresponding to two index sets contained in $\{1, \dots, n_i\}$ of cardinality j , in which of the two half-spaces lies P^t , we obtain a total order on all the expressions of the type (13) corresponding to index sets contained in $\{1, \dots, n_i\}$ of cardinality j . The obtained total order is global, i.e., for each fixed λ with $\text{ext}(\lambda) \in P^t$, it induces a consistent total order on the values obtained by fixing λ in the expressions (13). Consider a minimum element of the obtained total order, and the corresponding index set $v^t(i; j) \subseteq \{1, \dots, n_i\}$ of cardinality j . This index set has then the property that problem $\text{Opt}(i; j)_{|\lambda}$ has an optimal solution with support contained in $v^t(i; j)$ for all λ such that $\text{ext}(\lambda) \in P^t$. \diamond

As in Section 3.1, let $\theta := \max\{n_i : i = 1, \dots, h\}$, and $\bar{\theta} := (\theta - 1)\theta(\theta + 1)/2$.

Claim 4. Let $t \in T$. We can construct in polynomial time an index set U^t with $|U^t| = O(h^{2k^2(\bar{\theta}+1)})$, polyhedra $Q^{t,u} \subseteq P^t$, for $u \in U^t$, that cover P^t , and index sets $\chi^{t,u} \subseteq \{1, \dots, n\}$ of cardinality σ , for each $u \in U^t$, with the following property: For every $t \in T$, $u \in U^t$, and for every λ such that $\text{ext}(\lambda) \in Q^{t,u}$, the problem $\text{Opt}(\sigma)_{|\lambda}$ has an optimal solution with support contained in $\chi^{t,u}$.

Proof of claim. To prove this claim, we first construct all polyhedra $Q^{t,u}$, for $u \in U^t$. Then we show how to construct the index sets with the desired property.

Let $s \in \{0, \dots, n\}$. Let $j^s = (j_1^s, \dots, j_h^s)$ such that $j_i^s \in \{0, \dots, n_i\}$ for every $i \in \{1, \dots, h\}$, and $\sum_{i=1}^h j_i^s = s$. Define $\text{Aug}(j^s)$ as in (9) and let $j^{s+1} \in \text{Aug}(j^s)$. Define the sets $I^- := \{i \in \{1, \dots, h\} : j_i^{s+1} < j_i^s\}$ and $I^+ := \{i \in \{1, \dots, h\} : j_i^{s+1} > j_i^s\}$. For λ such that $\text{ext}(\lambda) \in P^t$, we define the expression

$$d(j^s, j^{s+1})_{|\lambda} := \sum_{i \in I^+ \cup I^-} (\text{Opt}(i; j_i^{s+1})_{|\lambda} - \text{Opt}(i; j_i^s)_{|\lambda}).$$

For λ such that $\text{ext}(\lambda) \in P^t$, consider the set $\mathcal{D}_{|\lambda}$ defined by

$$\mathcal{D}_{|\lambda} := \left\{ d(j^s, j^{s+1})_{|\lambda} : \text{there exists } s \in \{0, \dots, \sum_{i=1}^h n_i\} \text{ such that } \right. \\ \left. j^s \text{ is a feasible solution for } \text{Opt}(s)_{|\lambda} \text{ and } j^{s+1} \in \text{Aug}(j^s) \right\}.$$

Note that for each fixed λ , each $d(j^s, j^{s+1})_{|\lambda}$ is a value of the form $d(j^s, j^{s+1})$, as defined in (10), and the set $\mathcal{D}_{|\lambda}$ reduces to the set \mathcal{D} , as defined in (11).

Let $d(j^s, j^{s+1})_{|\lambda}$ and $d(k^s, k^{s+1})_{|\lambda}$ be two distinct expressions in $\mathcal{D}_{|\lambda}$. We wish to subdivide all points $\text{ext}(\lambda) \in \mathcal{S}$ based on which of the two expressions is larger. In order to do so, consider the equation

$$d(j^s, j^{s+1})_{|\lambda} = d(k^s, k^{s+1})_{|\lambda}. \quad (14)$$

We show that (14) is a quadratic equation in λ . Consider a single $\text{Opt}(i; j)_{|\lambda}$ that appears in the expression defining $d(j^s, j^{s+1})_{|\lambda}$, and let $v^t(i; j)$ be the corresponding index set from Claim 3. Let L be the linear subspace of \mathbb{R}^{m_i} defined by $L := \{A^i x^i : \text{supp}(x^i) \subseteq v^t(i; j)\}$, and let $p^i(\lambda) := b^i - \sum_{\ell=1}^k c_\ell^i \lambda_\ell$. From Claim 3, for all λ such that $\text{ext}(\lambda) \in P^t$, we have that $\text{Opt}(i; j)_{|\lambda}$ can be written as the quadratic function in λ of the form (13), namely

$$\text{Opt}(i; j)_{|\lambda} = \|\text{proj}_L(p^i(\lambda)) - p^i(\lambda)\|^2.$$

The expression $d(j^s, j^{s+1})_{|\lambda}$ is a linear combination of expressions $\text{Opt}(i; j)_{|\lambda}$. Hence, it can also be written as a quadratic function in λ . The same argument shows that also the expression $d(k^s, k^{s+1})_{|\lambda}$ can be written as a quadratic function in λ . Hence (14) is a quadratic equation in λ . By linearizing all the quadratic terms, we obtain a hyperplane in the space \mathcal{S} .

As a consequence of Lemma 4, the set $\mathcal{D}_{|\lambda}$ contains $O(h^{\bar{\theta}+1})$ expressions. Thus, by considering the corresponding hyperplane for all possible distinct pairs of expressions in $\mathcal{D}_{|\lambda}$, we obtain a total number of $O(h^{2(\bar{\theta}+1)})$ hyperplanes. These hyperplanes subdivide \mathcal{S} into a number of full-dimensional polyhedra. The hyperplane arrangement theorem [12] implies that this subdivision consists of at most $O((h^{2(\bar{\theta}+1)})^{\dim(\mathcal{S})}) = O(h^{2k^2(\bar{\theta}+1)})$ polyhedra R^u , for $u \in U^t$. Since k and $\bar{\theta}$ are fixed, $|U^t|$ is polynomial in h and the subdivision can be obtained in polynomial time. Define $Q^{t,u} := P^t \cap R^u$, for every $u \in U^t$.

We now fix one polyhedron $Q^{t,u}$, for some $u \in U^t$. By checking, for each hyperplane that we have constructed above, in which of the two half-spaces lies $Q^{t,u}$, we obtain a total order on all the expressions in \mathcal{D}_λ . The obtained total order induces a consistent total order on the values obtained by fixing λ in the expressions in \mathcal{D}_λ , for each fixed λ with $\text{ext}(\lambda) \in Q^{t,u}$. Since problem $\text{Opt}(\sigma)_{|\lambda}$ can be written in the form (12), Proposition 1 implies that we can obtain an optimal support $\chi^{t,u} \subseteq \{1, \dots, \sum_{i=1}^h n_i\}$ for problem $\text{Opt}(\sigma)_{|\lambda}$, for each fixed λ with $\text{ext}(\lambda) \in Q^{t,u}$. Note that, since the total order is independent on λ , also the obtained support is independent on λ . Therefore the claim follows. \diamond

Let \mathcal{X} be the set containing all index sets $\chi^{t,u}$ obtained in Claim 4, namely

$$\mathcal{X} := \{\chi^{t,u} : t \in T, u \in U^t\}.$$

Claim 5. *There exists an optimal solution (x^*, λ^*) of problem (5) such that*

$$\text{supp}(x^*) \subseteq \chi \text{ for some } \chi \in \mathcal{X}.$$

Proof of claim. Let (x^*, λ^*) be an optimal solution of problem (5). Then x^* is an optimal solution of the restricted problem $\text{Opt}(\sigma)_{|\lambda^*}$. Let $Q^{t,u}$, for $t \in T, u \in U^t$, be a polyhedron such that $\text{ext}(\lambda^*) \in Q^{t,u}$, and let $\chi^{t,u} \in \mathcal{X}$ be the corresponding index set. From Claim 4, the problem $\text{Opt}(\sigma)_{|\lambda^*}$ has an optimal solution \tilde{x} with support contained in $\chi^{t,u}$. This implies that the solution (\tilde{x}, λ^*) is also optimal for problem (5). \diamond

For each $\chi \in \mathcal{X}$, each problem (5), with the additional constraints $x_i = 0$, for all $i \notin \chi$, is a linear least squares problem, since the cardinality constraint can be dropped, and it can then be solved in polynomial time. The best solution among the obtained ones is an optimal solution of (5). To find an optimal solution to (5), we have solved $O(h^{k^2}(h^{2(\bar{\theta}+1)})^{k^2}) = O(h^{k^2(2\bar{\theta}+3)})$ linear least squares problems.

Since in the proof of Lemma 1 we have $k' \leq k + 1$, to find an optimal solution to (2) we need to solve $O(2^k h^{(k+1)^2(2\bar{\theta}+3)})$ linear least squares problems. This concludes the proof of Theorem 1. \square

References

- [1] Das A. and D. Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the 40th ACM Symposium on Theory of Computing (STOC)*, pages 45–54, 2008.
- [2] EML Beale, MG Kendall, and DW Mann. The discarding of variables in multivariate analysis. *Biometrika*, 54(3-4):357–366, 1967.
- [3] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. In *IEEE Transactions on Information Theory*, volume 52, pages 489–509, 2006.
- [4] Emmanuel Candes and Terence Tao. Decoding by linear programming. *arXiv preprint math/0502327*, 2005.
- [5] Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.
- [6] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

- [7] L. Chen and D. Marx. Covering a tree with rooted subtrees – parameterized and approximation algorithms. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2801–2820, 2018.
- [8] Christophe Couvreur and Yoram Bresler. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM Journal on Matrix Analysis and Applications*, 21(3):797–808, 2000.
- [9] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.
- [10] Gamarnik David and Zadik Ilias. High dimensional regression with binary coefficients. estimating squared error and a phase transtition. In *Conference on Learning Theory*, pages 948–953, 2017.
- [11] D.L. Donoho. Compressed sensing. In *IEEE Transactions on Information Theory*, volume 52, pages 1289–1306, 2006.
- [12] H. Edelsbrunner, J. O’Rourke, and R. Seidel. Constructing arrangements of lines and hyperplanes with applications. *SIAM Journal on Computing*, 15(2):341–363, 1986.
- [13] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- [14] MA Efroymsen. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.
- [15] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, Sahand Negahban, et al. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568, 2018.
- [16] Lester Randolph Ford and Delbert Ray Fulkerson. Constructing maximal dynamic flows from static flows. *Operations Research*, (6):419–433, 1953.
- [17] George M Furnival and Robert W Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.
- [18] J. Gao and D. Li. A polynomial case of the cardinality-constrained quadratic optimization problem. *Journal of Global Optimization*, 56:1441–1455, 2013.
- [19] A. Gilbert and P. Indyk. Sparse recovery using sparse matrices. *Proceedings of the IEEE*, 98(6):937–947, June 2010.
- [20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2009.
- [21] Raymond Hemmecke, Matthias Köppe, and Robert Weismantel. A polynomial-time algorithm for optimizing over N-fold 4-block decomposable integer programs. In Friedrich Eisenbrand and F.Bruce Shepherd, editors, *Integer Programming and Combinatorial Optimization (IPCO)*, volume 6080, 2010.
- [22] Raymond Hemmecke, Shmuel Onn, and Lyubov Romanchuk. N-fold integer programming in cubic time. *Mathematical Programming*, 137(1-2):325–341, 2013.

- [23] S. Heubach and T. Mansour. *Combinatorics of Compositions and Words*. Discrete Mathematics and Its Applications. CRC Press, 2009.
- [24] RR Hocking and RN Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967.
- [25] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- [26] Bruce Edward Hoppe. *Efficient dynamic network flow algorithms*. PhD thesis, Cornell University, 1995.
- [27] M. Koutecký, A. Levin, and S. Onn. A parametrized strongly polynomial algorithm for block structured integer programs. In *45 th International Colloquium on Automata, Languages, and Programming*, pages 85:1–85:14. ICALP, 2018.
- [28] Hendrik W. Lenstra. Integer programming with a fixed number of variables. *Math. Oper. Res.*, 8:538–548, 1983.
- [29] Jesús A. De Loera, Raymond Hemmecke, Shmuel Onn, and Robert Weismantel. N-fold integer programming. *Discrete Optimization*, 5(2):231–241, 2000.
- [30] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Inc., New York, NY, USA, 1990.
- [31] Alan Miller. *Subset selection in regression*. Chapman and Hall/CRC, 2002.
- [32] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013.
- [33] A. Schrijver. *Theory of Linear and Integer Programming*. Wiley, Chichester, 1986.
- [34] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning*. Cambridge University Press, 2014.
- [35] Martin Skutella. An introduction to network flows over time. In William Cook, László Lovász, and Jens Vygen, editors, *Research Trends in Combinatorial Optimization*. Springer Berlin Heidelberg, 2009.
- [36] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [37] W.J. Welch. Algorithm complexity: three NP-hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15:17–25, 1982.