Subset selection with sparse matrices

Alberto Del Pia[1] Santanu S. Dey[2] Robert Weismantel[3]

[1]University of Wisconsin -Madison.

[2]Georgia Institute of Technology.

[3]ETH Zürich.

April 2018

1

Introduction

Subset selection for linear regression

$$\begin{aligned} \min_{x,\mu} \quad & ||Mx + \mathbf{1}\mu - b||^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \ \mu \in \mathbb{R} \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

- Rows of M are samples, columns of M corresponds to variables.
- A large number of variables can be observed, and we are interested in the value of a predictor variable b.

Subset selection for linear regression

$$\begin{aligned} \min_{x,\mu} \quad & ||Mx + \mathbf{1}\mu - b||^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \ \mu \in \mathbb{R} \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

- Rows of M are samples, columns of M corresponds to variables.
- A large number of variables can be observed, and we are interested in the value of a predictor variable b.
- Sparsity constraint: For interpretability, restrict to few variables. Also due to time or cost constraints, it is not feasible to sample all the variables every time a prediction is required.
- Goal: Select a small subset of variables to predict b in the future.

Subset selection for regression: Formulation and Approach

$$\begin{aligned} \min_{x,\mu} \quad & ||Mx + \mathbf{1}\mu - b||^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \ \mu \in \mathbb{R} \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

Different communities and different approaches:

- Enumeration: Best-subset regression.
- Greedy algorithms: Forward- and backward-stepwise selection, forward-stagewise regression, etc.
- Branch and bound: Leaps and bounds procedure.
- Convex optimization: Shrinkage methods: ridge regression, the lasso, least angle regression, etc.

Subset selection for regression: Complexity

$$
\begin{aligned}
\min \quad & \|Mx + \mathbf{1}\mu - b\|^2 \\
\text{s.t.} \quad & x \in \mathbb{R}^n, \ \mu \in \mathbb{R} \\
& |\operatorname{supp}(x)| \leq \sigma
\end{aligned}
$$

The problem is NP-hard [Welch, 1982]. Few polynomially solvable cases:

- When the covariance graph is a tree. [Das and Kempe, 2008]
- Approximate algorithm when the covariance has constant bandwidth. [Das, Kempe, 2008]
- Under some conditions, using l_1 norm to replace the cardinality constraint yields the exact solution with an overwhelming probability. [Donoho, 2006] [Candes, Romberg, Tao 2006]

2

Main Result

Our contribution

$$\begin{aligned} \min \quad & \|Mx + \mathbf{1}\mu - b\|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \ \mu \in \mathbb{R} \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

We wish to understand the problem from a computational complexity viewpoint.
Can be solved in polynomial time if:

(i) M is obtained from a diagonal matrix by adding a fixed number of extra columns.

$$\left(\begin{array}{ccc|ccc} d_1 & & & | & & | \\ & \ddots & & c^1 & \cdots & c^k \\ & & d_n & | & & | \end{array} \right)$$

Our contribution

$$\begin{aligned} \min \quad & \|Mx + \mathbf{1}\mu - b\|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n, \ \mu \in \mathbb{R} \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

We wish to understand the problem from a computational complexity viewpoint.
Can be solved in polynomial time if:

(ii) M is obtained by adding a fixed number of extra columns to a block diagonal matrix, where each block has a fixed number of variables.

$$\left(\begin{array}{ccc|ccc} A^1 & & & | & & | \\ & \ddots & & c_1 & \cdots & c_k \\ & & A^h & | & & | \end{array} \right)$$

First reduction

$$\begin{aligned} \min \quad & \|(A|c_1|\cdots|c_k)x + \mathbf{1}\mu - b\|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^{n+k},\ \mu \in \mathbb{R} \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

First reduction

$$\begin{aligned} \min \quad & \|(A|c_1|\cdots|c_k)x + \mathbf{1}\mu - b\|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^{n+k},\ \mu \in \mathbb{R} \\ & |\operatorname{supp}(x)| \le \sigma \end{aligned}$$

Lemma

We just need to show how to solve in polynomial time the problem

$$\begin{aligned} \min \quad & \left\| Ax - \left(b - \sum_{\ell=0}^{\tilde{k}} c_\ell \lambda_\ell \right) \right\|^2 \\ \textit{s.t.} \quad & x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}^{\tilde{k}} \\ & |\operatorname{supp}(x)| \le \tilde{\sigma} \end{aligned}$$

3

Diagonal Case

The diagonal case

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

The diagonal case

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

First, we consider a simpler problem by fixing the variables λ:

$$\begin{aligned} \min \quad & ||Dx - b||^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

The diagonal case

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

First, we consider a simpler problem by fixing the variables λ:

$$\begin{aligned} \min \quad & \left\| \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix} x - b \right\|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

The diagonal case

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

First, we consider a simpler problem by fixing the variables λ:

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} (d_j x_j - b_j)^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \le \sigma \end{aligned}$$

A simpler diagonal problem

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} (d_j x_j - b_j)^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

A simpler diagonal problem

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} (d_j x_j - b_j)^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

- We can order indices $1, \ldots, n$ such that we have

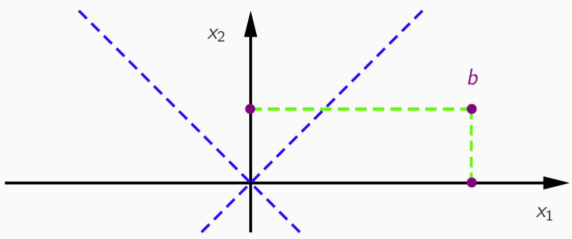
$$|b_{j_1}| \geq |b_{j_2}| \geq \cdots \geq |b_{j_n}| .$$

- The optimal support $\{j_1, j_2, \ldots, j_\sigma\}$ only depends on the ordering.
- An optimal solution is

$$x^*_{j_i} := \begin{cases} b_{j_i}/d_{j_i} & \text{for } i = 1, \ldots, \sigma \\ 0 & \text{otherwise} \end{cases}$$

A simpler diagonal problem

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} (d_j x_j - b_j)^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

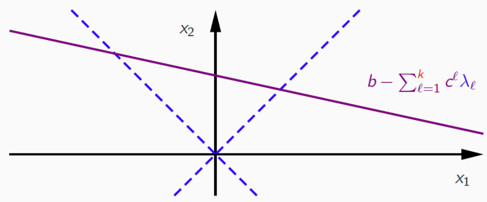


Back to the diagonal case

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

Back to the diagonal case

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



Back to the diagonal case

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

- Key Idea: We wish to partition all λ vectors based on the optimal support they yield.
- In each cell of the partition, we want to have an ordering of all the

$$\left| b_i - \sum_{\ell=1}^{k} c_i^{\ell} \lambda_{\ell} \right|$$

Back to the diagonal case

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

- For every two indices i, j, we subdivide all points $\lambda \in \mathbb{R}^k$ based on
$$\left| b_i - \sum_{\ell=1}^{k} c_i^{\ell} \lambda_{\ell} \right| \sim \left| b_j - \sum_{\ell=1}^{k} c_j^{\ell} \lambda_{\ell} \right|$$
- We just need two hyperplanes.
- In total we obtain $O(n^2)$ hyperplanes.
- By the *hyperplane arrangement theorem*, we obtain $O(n^{2k})$ polyhedra Q^t in $\mathbb{R}^k$.

Back to the diagonal case

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

Consider one polyhedron Q^t.

- We obtain an ordering of all the

$$\left| b_i - \sum_{\ell=1}^{k} c_i^{\ell} \lambda_{\ell} \right|$$

  that holds for all $\lambda \in Q^t$
- The σ indices highest in the ordering yield an optimal support for all λ in Q^t.

Back to the diagonal case

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

- Let $\mathcal{X}$ be the set containing all these $O(n^{2^k})$ supports.
- For each $\chi \in \mathcal{X}$, the original problem with the additional constraints
$$x_i = 0, \text{ for all } i \notin \chi,$$
can be solved in polynomial time.
- The original problem has an optimal solution with support contained in χ for some $\chi \in \mathcal{X}$.

4

The Block Diagonal Case

The block diagonal case

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

The block diagonal case

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

Same overall strategy:

1. Design an algorithm for the simpler problem obtained by fixing the variables λ.
2. Cover the space of the λ variables with a polynomial number of regions such that in each region the algorithm yields the same optimal support.

The block diagonal case

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

First, we consider a simpler problem by fixing the variables λ:

$$\begin{aligned} \min \quad & ||Ax - b||^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

The block diagonal case

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

First, we consider a simpler problem by fixing the variables λ:

$$\begin{aligned} \min \quad & \left\| \begin{pmatrix} A^1 & & \\ & \ddots & \\ & & A^h \end{pmatrix} \begin{pmatrix} x^1 \\ \vdots \\ x^h \end{pmatrix} - \begin{pmatrix} b^1 \\ \vdots \\ b^h \end{pmatrix} \right\|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

The block diagonal case

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

First, we consider a simpler problem by fixing the variables λ:

$$\begin{aligned} \min \quad & \sum_{i=1}^{h} \left|\left| A^i x^i - b^i \right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

4.1
Solving simpler block diagonal problem: "Generalizing" the sorting algorithm

A simpler block diagonal problem: failed attempt

$$\begin{aligned} \min \quad & \sum_{i=1}^{h} \left|\left|A^i x^i - b^i\right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

Dynamic programming

- $$\begin{aligned} OPT(i,j) = \min \quad & \left|\left|A^i x^i - b^i\right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq j \end{aligned}$$

- $$OPT(1,\dots,i,j) = min_{0 \leq u \leq j}\{OPT(1,\dots,i-1,u) + OPT(i,j-u)\}$$

A simpler block diagonal problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{h} \left|\left| A^i x^i - b^i \right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

A simpler block diagonal problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{h} \left|\left|A^i x^i - b^i\right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq \sigma \end{aligned}$$

- We have to decide how to distribute σ among the different blocks. We can do it recursively for support $1, 2, \ldots, \sigma$.

Lemma

Given an optimal solution with $\sigma = s$, there exists an optimal solution for $\sigma = s + 1$, where the difference in support of the two solutions is no more than $O(\theta^3)$ supports, where θ is the maximum number of variables in a block.

- The optimal support only depends on an ordering of $O(h^{\theta^3})$ values of the form

$$d(I, J) := \sum_{i \in I} OPT(i, j_{i1}) - OPT(i, j_{i2})$$

where $|I| \leq \theta^3$ and $\sum_{i \in I} j_{i1} - j_{i2} = 1$.

4.2
Back to the block diagonal case

Back to the block diagonal case

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

Back to the block diagonal case

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

- We wish to partition all λ vectors based on the optimal support they yield.
- In each cell of the partition, we want to have an ordering of all the

$$d(I, J) := \sum_{i \in I} OPT(i, j_{i1}) - OPT(i, j_{i2})$$

$$\begin{aligned} OPT(i, j) = \min \quad & \left\| A^i x^i - b^i \right\|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \le j \end{aligned}$$

Partitioning the space

$$d(I, J) := \sum_{i \in I} OPT(i, j_{i1}) - OPT(i, j_{i2})$$

$$\begin{aligned} OPT(i, j) = \min \quad & \left|\left| A^i x^i - b^i \right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq j \end{aligned}$$

Issue 1: There is no such polyhedral decomposition of the λ space.
Why? There are quadratic functions involved.

Solution: Linearization

- Instead of the λ space, we define a new space $\mathcal{S}$ where we can linearize any quadratic monomial.
- The dimension of $\mathcal{S}$ is $O(k^2)$.

Partitioning the space

$$d(I,J) := \sum_{i \in I} OPT(i, j_{i1}) - OPT(i, j_{i2})$$

$$\begin{aligned} OPT(i,j) = \min \quad & \left|\left|A^i x^i - b^i\right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq j \end{aligned}$$

Issue 2: The objective value of each $OPT(i,j)$ depends on its optimal support.

Solution: Two-level of partition

- In level 1, we partition $\mathcal{S}$ in polyhedra $P^t \subseteq \mathcal{S}$, such that in each of them every $OPT(i,j)$ has a fixed optimal support.

Partitioning the space

$$d(I, J) := \sum_{i \in I} OPT(i, j_{i1}) - OPT(i, j_{i2})$$

$$\begin{aligned} OPT(i, j) = \min \quad & \left|\left| A^i x^i - b^i \right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq j \end{aligned}$$

Level 1 partition:

- For each fixed support, the optimal value of every $OPT(i, j)$ is a quadratic function.
- By comparing all pairs of quadratic functions, we obtain $O(h2^{2^\theta}) = O(h)$ hyperplanes in $\mathcal{S}$.
- By the *hyperplane arrangement theorem*, we obtain $O(h^{k^2})$ polyhedra P^t in $\mathcal{S}$.
- The best quadratic function among those corresponding to i, j yields the optimal support of $OPT(i, j)$.

Partitioning the space

$$d(I, J) := \sum_{i \in I} OPT(i, j_{i1}) - OPT(i, j_{i2})$$

$$\begin{aligned} OPT(i, j) = \min \quad & \left|\left|A^i x^i - b^i\right|\right|^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \\ & |\operatorname{supp}(x)| \leq j \end{aligned}$$

Level 2 partition:

- In each P^t every expression is a linear function in $\mathcal{S}$.
- For every two such expression, we subdivide all points in $\mathcal{S}$ based on which is largest.
- In total we obtain $O(h^{2\theta^3})$ hyperplanes.
- By the *hyperplane arrangement theorem*, we obtain $O(h^{2\theta^3 k^2})$ polyhedra $Q^{t,u} \subseteq P^t$.

Wrapping up the block diagonal case

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

Consider one polyhedron $Q^{t,u}$.

- We obtain an ordering of all the $d(I, J)$ that holds for all $\lambda \in Q^{t,u}$
- The algorithm for the simpler block diagonal problem yields the same optimal support for all λ in $Q^{t,u}$.

Wrapping up the block diagonal case

$$f \min \quad \left\| Ax - \left(b - \sum_{\ell=1}^{k} c_\ell \lambda_\ell \right) \right\|^2$$
$$\text{s.t.} \quad x \in \mathbb{R}^n, \ \lambda \in \mathbb{R}^k$$
$$|\operatorname{supp}(x)| \le \sigma$$

- Let $\mathcal{X}$ be the set containing all these $O(h^{2\theta^3 k^2})$ supports.
- For each $\chi \in \mathcal{X}$, the original problem with the additional constraints
$$x_i = 0, \ \text{for all } i \notin \chi,$$
can be solved in polynomial time.
- The original problem has an optimal solution with support contained in χ for some $\chi \in \mathcal{X}$.

Thank You