Methodologies and Algorithms for Structured Mixed-Integer Nonlinear Optimization

Andrés Gómez

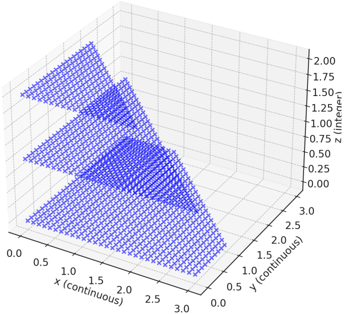
Department of Industrial & Systems Engineering
University of Southern California

MIP Workshop 2025

Agenda

Mixed-integer linear optimization (MILO)

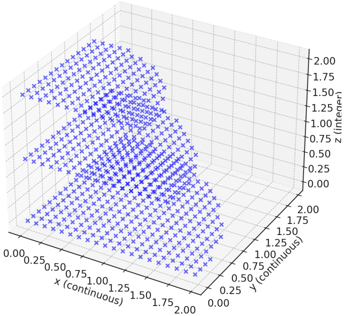
$$\begin{aligned} \min\ & \boldsymbol{c}^\top \boldsymbol{x} + \boldsymbol{d}^\top \boldsymbol{z} \\ \text{s.t. } & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{G}\boldsymbol{z} \le \boldsymbol{b} \\ & \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m \end{aligned}$$



- Usually solved using branch-and-cut
- NP-hard in theory, *often* solvable in practice

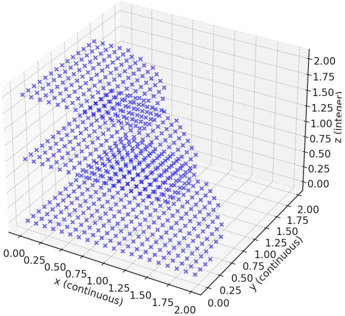
Mixed-integer nonlinear optimization (MINLO)

$$
\begin{aligned}
\min\ & f(\boldsymbol{x}, \boldsymbol{z}) \\
\text{s.t. } & g_i(\boldsymbol{x}, \boldsymbol{z}) \leq 0 \qquad\qquad i = 1, \dots, p \\
& \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m
\end{aligned}
$$



Mixed-integer nonlinear optimization (MINLO)

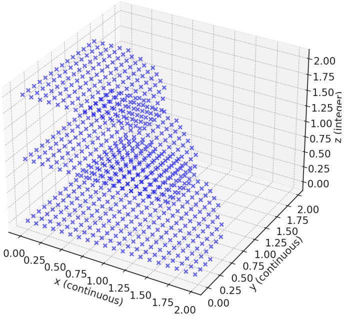
$$\begin{aligned} &\min f(\boldsymbol{x}, \boldsymbol{z}) \\ &\text{ s.t. } g_i(\boldsymbol{x}, \boldsymbol{z}) \leq 0 \qquad\qquad i = 1, \ldots, p \\ &\qquad \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m \end{aligned}$$



- Undecidable Includes *Hilbert's 10th problem*
    - Given polynomial g, does there exist $\boldsymbol{z} \in \mathbb{Z}^n$ satisfying $g(\boldsymbol{z}) = 0$

Mixed-integer nonlinear optimization (MINLO)

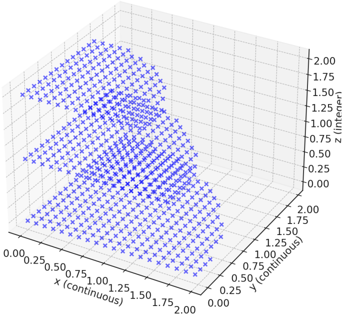
$$\begin{aligned} \min\ & f(\boldsymbol{x}, \boldsymbol{z}) \\ \text{s.t. } & g_i(\boldsymbol{x}, \boldsymbol{z}) \leq 0 \qquad\qquad i = 1, \ldots, p \\ & \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m \end{aligned}$$



- Undecidable Includes *Hilbert's 10th problem*
  - Given polynomial g, does there exist $\boldsymbol{z} \in \mathbb{Z}^n$ satisfying $g(\boldsymbol{z}) = 0$
- Unstructured Letting $g(z) = z - z^2$, $z \in \{0, 1\} \Leftrightarrow g(z) = 0$

Mixed-integer nonlinear optimization (MINLO)

$$\begin{aligned} \min\ & f(\boldsymbol{x}, \boldsymbol{z}) & \\ \text{s.t. } & g_i(\boldsymbol{x}, \boldsymbol{z}) \le 0 & i = 1, \dots, p \\ & \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m & \end{aligned}$$



- Undecidable Includes *Hilbert's 10th problem*
  - Given polynomial g, does there exist $\boldsymbol{z} \in \mathbb{Z}^n$ satisfying $g(\boldsymbol{z}) = 0$
- Unstructured Letting $g(z) = z - z^2$, $z \in \{0, 1\} \Leftrightarrow g(z) = 0$

$\rightarrow$ We assume continuous relaxation is "nice" (e.g., f and g_i are convex)

Mixed-integer nonlinear optimization (MINLO)

$$\begin{aligned} \min\ & f(\boldsymbol{x}, \boldsymbol{z}) & \\ \text{s.t. } & g_i(\boldsymbol{x}, \boldsymbol{z}) \le 0 & i = 1, \ldots, p \\ & \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m & \end{aligned}$$

Transformations

- Objective is linear: $\min y$ s.t. $f(\boldsymbol{x}, \boldsymbol{z}) \le y$
- Single constraint: $g(\boldsymbol{x}, \boldsymbol{z}) \le 0$ with $g(\boldsymbol{x}, \boldsymbol{z}) = \max_i g_i(\boldsymbol{x}, \boldsymbol{z})$
- Unconstrained: $F(\boldsymbol{x}, \boldsymbol{z}) = \begin{cases} f(\boldsymbol{x}, \boldsymbol{z}) & \text{if } g(\boldsymbol{x}, \boldsymbol{z}) \le 0 \\ \infty & \text{otherwise} \end{cases}$

Current "state-of-the-art" for MINLO

- Much less understood and mature than MILOs
- Concepts like number of variables/constraints are "uninformative"
- Most solvers and researchers are focused elsewhere
- Unlike MILOs, most of the heavy-lifting is left to the user

Agenda

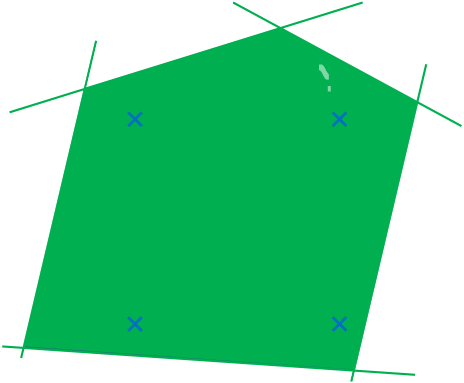
1. Introduction
2. Branch and bound
   - Branch and bound for MILO
   - Branch and bound for MINLO
3. Convexification

Agenda

Branch-and-cut for MILO

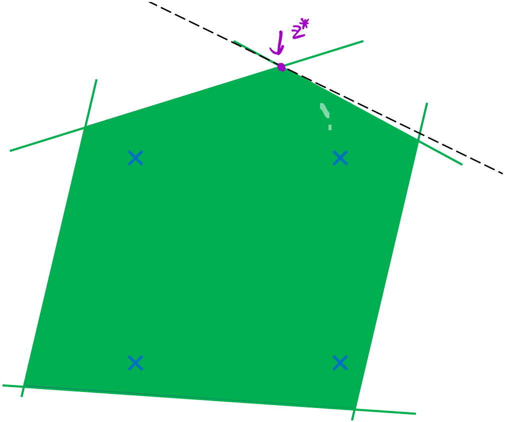
Discrete feasible region

Branch-and-cut for MILO



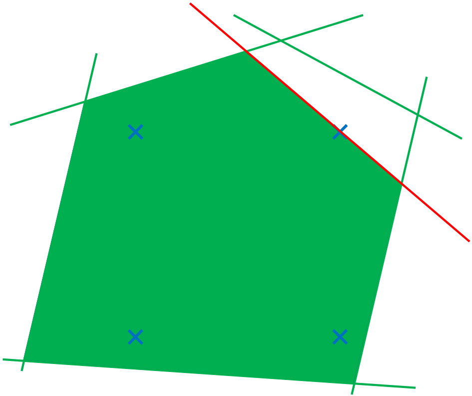
Linear programming relaxation

Branch-and-cut for MILO



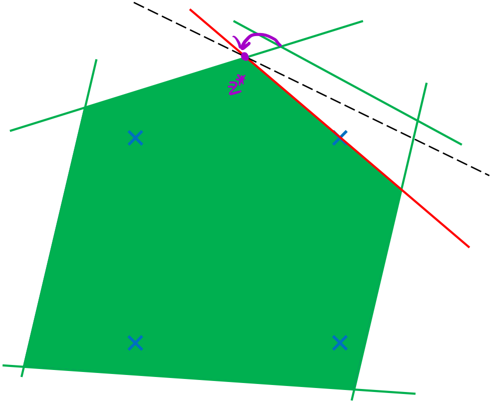
Solve (extreme point solution)

Branch-and-cut for MILO



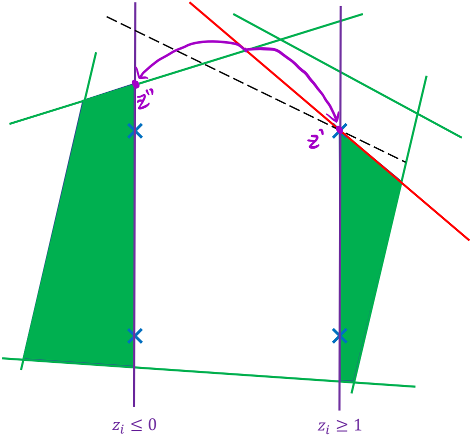
Improve relaxation (cutting plane)

Branch-and-cut for MILO



Solve (dual simplex)

Branch-and-cut for MILO



Branch and resolve (dual simplex with two independent subproblems)

Branch-and-cut for MILO

Algorithm To solve a mixed-integer linear program

- Start with a linear relaxation
- Dynamically refine using cutting planes
- Branch when needed
- Reoptimize using the simplex method
- When upper bound (best solution) = lower bound (relaxation), stop

- Other techniques
  - Heuristics, often based on rounding solutions from linear relaxations
  - Presolve, to improve the initial linear relaxation

Branch-and-cut for MILO

Algorithm To solve a mixed-integer linear program

- Start with a linear relaxation
- Dynamically refine using cutting planes
- Branch when needed
- Reoptimize using the simplex method
- When upper bound (best solution) = lower bound (relaxation), stop

- Other techniques
  - Heuristics, often based on rounding solutions from linear relaxations
  - Presolve, to improve the initial linear relaxation

Algorithms revolve around deriving and exploiting linear relaxations

Agenda

1. Introduction
2. Branch and bound
   - Branch and bound for MILO
   - Branch and bound for MINLO
3. Convexification

Branch-and-cut for MINLO

The same algorithm works!

Branch-and-cut for MINLO

The same algorithm works!

...but how to solve the continuous relaxations?

Second order methods

- Hard to warm start (after branching or cuts)
- Memory intensive

$\rightarrow$ Adds up when $\#$ nodes $> 10^6$

First order methods

- May struggle in heavily constrained problems
- High-quality solutions difficult to obtain

$\rightarrow$ Numerical precision can be an issue

Branch-and-cut for MINLO

Numerical precision is a very real issue in MINLO

C:\WINDOWS\system32\cmd.exe - run.bat

```
      0     0        0.0029     3                    Cone: 235      847
      0     0        0.0029   161                       0.0029      874
*     0+    0                            0.1800         0.0029             98.38%
      0     0        0.0029   200        0.1800     MIRcuts: 1      959    98.37%
      0     0        0.0029   200        0.1800     MIRcuts: 1      987    98.37%
*     0+    0                            0.0035         0.0029             16.90%
      0     2        0.0029   200        0.0035         0.0029     1016    16.65%
Elapsed time = 1.66 sec. (3548.12 ticks, tree = 0.01 MB, solutions = 2)
      3     5        0.0029   198        0.0035         0.0029     1104    16.36%
      6     8        0.0029   195        0.0035         0.0029     1189    16.36%
      9    11        0.0029   193        0.0035         0.0029     1276    16.36%
Integer feasible solution rejected --- infeasible on original model
     10    12        0.0029   192        0.0035         0.0029     1305    16.36%
     13    15        0.0029   190        0.0035         0.0029     1400    16.36%
     16    18        0.0029   187        0.0035         0.0029     1490    16.36%
     19    21        0.0030   185        0.0035         0.0029     1583    16.36%
Integer feasible solution rejected --- infeasible on original model
     20    22        0.0029   184        0.0035         0.0029     1616    16.36%
     23    25        0.0029   181        0.0035         0.0029     1713    16.36%
Integer feasible solution rejected --- infeasible on original model
     32    34        0.0029   173        0.0035         0.0029     1981    16.36%
Elapsed time = 3.25 sec. (7110.85 ticks, tree = 0.01 MB, solutions = 2)
Integer feasible solution rejected --- infeasible on original model
     42    44        0.0029   165        0.0035         0.0029     2281    16.36%
Integer feasible solution rejected --- infeasible on original model
Integer feasible solution rejected --- infeasible on original model
     50    52        0.0029   157        0.0035         0.0029     2517    16.36%
Integer feasible solution rejected --- infeasible on original model
     60    62        0.0029   147        0.0035         0.0029     2819    16.36%
Integer feasible solution rejected --- infeasible on original model
```

Portfolio optimization

Given potential investments $\{1, \ldots, n\}$, find a small portfolio maximizing return and minimizing risk

- Decision variables $\boldsymbol{x} \in \mathbb{R}^n$, where $x_i = \%$ invested in security i
- Return $\boldsymbol{\mu} \in \mathbb{R}^n$, where $\mu_i =$ expected profit of investment i
  $\rightarrow$ Total return: $\boldsymbol{\mu}^\top \boldsymbol{x}$
- Risk $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, where $\Sigma_{ij} =$ covariance of returns from i and j
  $\rightarrow$ Variance of portfolio: $\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}$
- Size $\#$ of nonzero elements of $\boldsymbol{x}$ is small

Portfolio optimization

$$\begin{aligned} \max_{\boldsymbol{x},\boldsymbol{z}} \ & \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t. } & \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \leq \alpha \\ & \mathbf{1}^\top \boldsymbol{x} = 1 \\ & \mathbf{0} \leq \boldsymbol{x} \leq \boldsymbol{z} \\ & \mathbf{1}^\top \boldsymbol{z} \leq k \\ & \boldsymbol{x} \in \mathbb{R}^n, \ \boldsymbol{z} \in \{0,1\}^n \end{aligned}$$

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{z}} \ & \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \\ \text{s.t. } & \boldsymbol{\mu}^\top \boldsymbol{x} \geq \beta \\ & \mathbf{1}^\top \boldsymbol{x} = 1 \\ & \mathbf{0} \leq \boldsymbol{x} \leq \boldsymbol{z} \\ & \mathbf{1}^\top \boldsymbol{z} \leq k \\ & \boldsymbol{x} \in \mathbb{R}^n, \ \boldsymbol{z} \in \{0,1\}^n \end{aligned}$$

Which formulation is preferable?

Branch-and-cut for MIQO

$$\begin{aligned} \min\ & f(\boldsymbol{x}, \boldsymbol{z}) \\ \text{s.t. } & \boldsymbol{Ax} + \boldsymbol{Gz} \leq \boldsymbol{b} \\ & \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m \end{aligned}$$

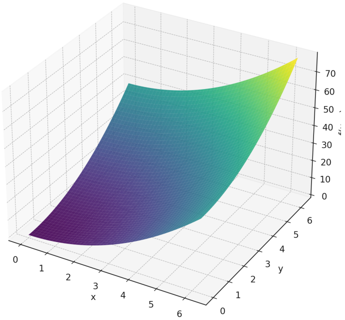
where f is quadratic

Continuous relaxations can be solved via the simplex method[1][2]

Keeping quadratic terms in the objective seems to help in MINLO

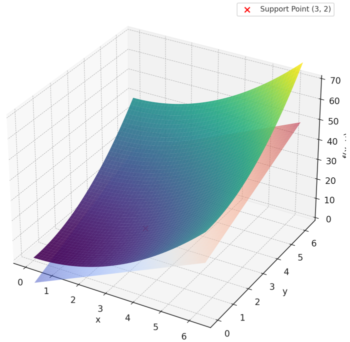
[1] Wolfe P (1959) The Simplex method for quadratic programming. *Econometrica*
[2] Van de Panne C and Whinston A (1964) Simplicial methods for quadratic programming. *Naval Research Logistics*

Linear outer approximations



Consider constraint $f(\boldsymbol{x}) \leq t$

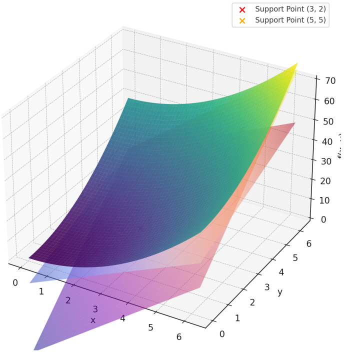
Linear outer approximations



Consider constraint $f(\boldsymbol{x}) \leq t$

Given $\bar{\boldsymbol{x}}$, can be relaxed as $f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq t$

Linear outer approximations



Consider constraint $f(\boldsymbol{x}) \leq t$

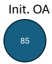
Given $\bar{\boldsymbol{x}}$, can be relaxed as $f(\bar{\boldsymbol{x}}) + \nabla f(\bar{\boldsymbol{x}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}) \leq t$

This process can be repeated for different support points

Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

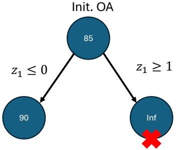
- Assume UB=100
- Construct an initial linear OA



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

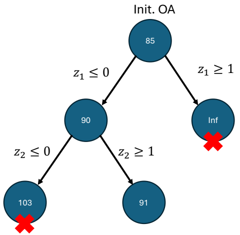
- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

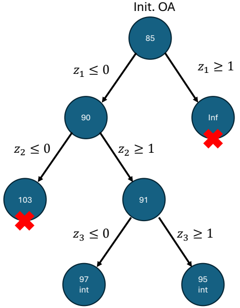
- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

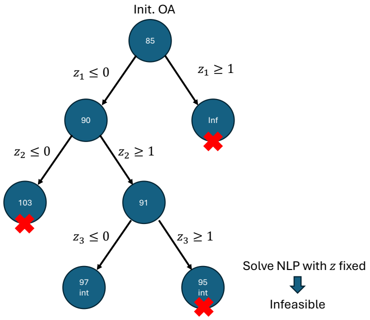
- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

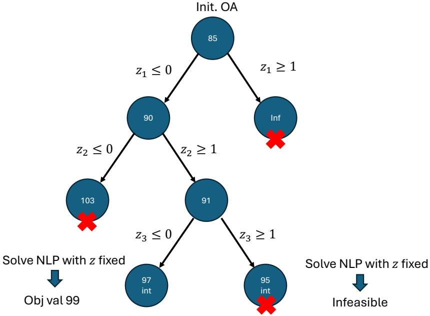
- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual
- Integer nodes might be infeasible



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

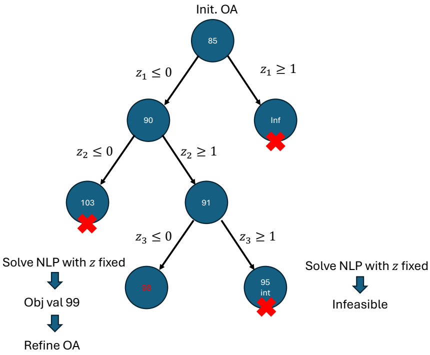
- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual
- Integer nodes might be infeasible
- Incumbents obj values need to be handled carefully



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

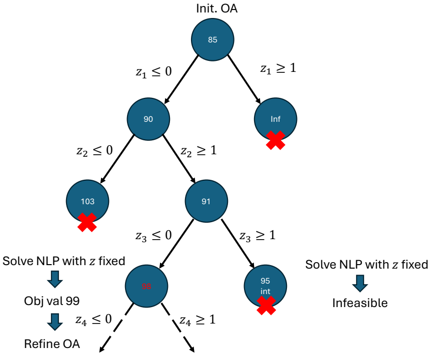
- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual
- Integer nodes might be infeasible
- Incumbents obj values need to be handled carefully



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

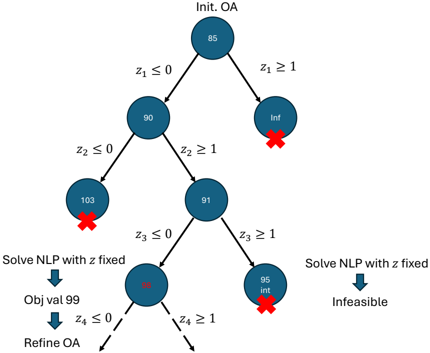
- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual
- Integer nodes might be infeasible
- Incumbents obj values need to be handled carefully
- No pruning at integer nodes



Linear outer approximations in branch-and-bound

How to integrate in branch-and-bound?

- Assume UB=100
- Construct an initial linear OA
- Branching, pruning by bound/infeasibility as usual
- Integer nodes might be infeasible
- Incumbents obj values need to be handled carefully
- No pruning at integer nodes



How to best construct linear outer approximations?

Constructing effective linear outer approximations

Assume support points $\{\bar{\boldsymbol{x}}^{\boldsymbol{j}}\}_{j=1}^{r}$ and approximate[3]

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} h_i(x_i)?$$

Direct Add r linear inequalities

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}^{\boldsymbol{j}}) + \nabla f(\bar{\boldsymbol{x}}^{\boldsymbol{j}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}^{\boldsymbol{j}}), \quad \forall j = 1, \ldots, r$$

[3] Tawarmalani M and Sahinidis N (2005) A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming*

Constructing effective linear outer approximations

Assume support points $\{\bar{\boldsymbol{x}}^{\boldsymbol{j}}\}_{j=1}^{r}$ and approximate[3]

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} h_i(x_i)?$$

Direct Add r linear inequalities

$$f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}}^{\boldsymbol{j}}) + \nabla f(\bar{\boldsymbol{x}}^{\boldsymbol{j}})^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}^{\boldsymbol{j}}), \;\; \forall j = 1, \ldots, r$$

Extended Add n variables and nr linear inequalities

$$f(\boldsymbol{x}) \geq \sum_{i=1}^{n} y_i$$

$$y_i \geq h_i(\bar{x}_i^j) + h_i'(\bar{x}_i^j)(x_i^j - \bar{x}_i^j), \;\; \forall i = 1, \ldots, n, \; j = 1, \ldots, r$$

[3] Tawarmalani M and Sahinidis N (2005) A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming*

Constructing effective linear outer approximations

Example Outer approximate function

$$f(\boldsymbol{x}) = |x_1| + |x_2| + |x_3| + |x_4|$$

Constructing effective linear outer approximations

Example Outer approximate function

$$f(\boldsymbol{x}) = |x_1| + |x_2| + |x_3| + |x_4|$$

Direct Add $2^n = 16$ linear inequalities

$$\begin{array}{lll} f(\boldsymbol{x}) \geq x_1 + x_2 + x_3 + x_4, & f(\boldsymbol{x}) \geq x_1 + x_2 + x_3 - x_4, & f(\boldsymbol{x}) \geq x_1 + x_2 - x_3 + x_4 \\ f(\boldsymbol{x}) \geq x_1 + x_2 - x_3 - x_4, & f(\boldsymbol{x}) \geq x_1 - x_2 + x_3 + x_4, & f(\boldsymbol{x}) \geq x_1 - x_2 + x_3 - x_4 \\ \vdots & & \end{array}$$

Constructing effective linear outer approximations

Example Outer approximate function

$$f(\boldsymbol{x}) = |x_1| + |x_2| + |x_3| + |x_4|$$

Direct Add $2^n = 16$ linear inequalities

$$\begin{aligned}
&f(\boldsymbol{x}) \geq x_1 + x_2 + x_3 + x_4, \quad && f(\boldsymbol{x}) \geq x_1 + x_2 + x_3 - x_4, \quad && f(\boldsymbol{x}) \geq x_1 + x_2 - x_3 + x_4 \\
&f(\boldsymbol{x}) \geq x_1 + x_2 - x_3 - x_4, \quad && f(\boldsymbol{x}) \geq x_1 - x_2 + x_3 + x_4, \quad && f(\boldsymbol{x}) \geq x_1 - x_2 + x_3 - x_4 \\
&\quad \vdots
\end{aligned}$$

Extended Add $n = 4$ variables and $2n = 8$ linear inequalities

$$f(\boldsymbol{x}) \geq \sum_{i=1}^{n} y_i$$

$$y_i \geq x_i, \; y_i \geq -x_i \quad i = 1, \ldots, 4$$

Constructing effective linear outer approximations

Proposition (Tawarmalani and Sahinidis 2005)

For separable functions, the extended formulation with support points $\{\bar{\boldsymbol{x}}^{\boldsymbol{j}}\}_{j=1}^{r}$ is equivalent to the direct linear outer approximation supported at every $\boldsymbol{x}$ such that for every $i \in [n]$ there exists $j \in [r]$ such that $x_i = \bar{x}_i^j$.

- Polynomial extended formulations $\Leftrightarrow$ Exponential direct OA
- Linear ineqs in extended space $\Leftrightarrow$ Nonlinear ineqs in original space

Outer approximations of nonlinear functions?

Quadratic functions

$$f(\boldsymbol{x}) = 5x_1^2 + 4x_2^2 + 9x_3^2 + 4x_1x_2 + 6x_1x_3 + 12x_2x_3$$

Outer approximations of nonlinear functions?

Quadratic functions

$$f(\boldsymbol{x}) = 5x_1^2 + 4x_2^2 + 9x_3^2 + 4x_1x_2 + 6x_1x_3 + 12x_2x_3$$
$$f(\boldsymbol{x}) = (x_1 + 2x_2 + 3x_3)^2 + 4x_1^2$$

Outer approximations of nonlinear functions?

Quadratic functions

$$f(\boldsymbol{x}) = 5x_1^2 + 4x_2^2 + 9x_3^2 + 4x_1x_2 + 6x_1x_3 + 12x_2x_3$$
$$f(\boldsymbol{x}) = (x_1 + 2x_2 + 3x_3)^2 + 4x_1^2$$
$$f(\boldsymbol{x}) = x_4^2 + 4x_1^2 \text{ with } x_4 = x_1 + 2x_2 + 3x_3$$

Outer approximations of nonlinear functions?

Quadratic functions

$$f(\boldsymbol{x}) = 5x_1^2 + 4x_2^2 + 9x_3^2 + 4x_1x_2 + 6x_1x_3 + 12x_2x_3$$
$$f(\boldsymbol{x}) = (x_1 + 2x_2 + 3x_3)^2 + 4x_1^2$$
$$f(\boldsymbol{x}) = x_4^2 + 4x_1^2 \text{ with } x_4 = x_1 + 2x_2 + 3x_3$$

Any convex quadratic function of rank k can be written as a separable function with k additional variables

$\rightarrow$ Cholesky decomposition, eigendecomposition...

Outer approximations of nonlinear functions?

Conic quadratic functions [4] (Handling the Lorentz cone)

$$x_0 \geq \sqrt{\sum_{i=1}^{n} x_i^2}$$

[4] Vielma JP et al (2017) Extended formulations in mixed-integer conic quadratic programming. *Mathematical Programming Computation*

Outer approximations of nonlinear functions?

Conic quadratic functions [4] (Handling the Lorentz cone)

$$x_0 \geq \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\Leftrightarrow \; x_0^2 \geq \sum_{i=1}^{n} x_i^2 \text{ and } x_0 \geq 0$$

[4] Vielma JP et al (2017) Extended formulations in mixed-integer conic quadratic programming. *Mathematical Programming Computation*

Outer approximations of nonlinear functions?

Conic quadratic functions [4] (Handling the Lorentz cone)

$$
\begin{aligned}
& x_0 \geq \sqrt{\sum_{i=1}^{n} x_i^2} \\
\Leftrightarrow\ & x_0^2 \geq \sum_{i=1}^{n} x_i^2 \text{ and } x_0 \geq 0 \\
\Leftrightarrow\ & x_0 \geq \sum_{i=1}^{n} x_i^2 / x_0 \text{ and } x_0 \geq 0 \\
\Leftrightarrow\ & x_0 \geq \sum_{i=1}^{n} y_i \text{ and } x_0 \geq 0,\ y_i \geq x_i^2 / x_0, \forall i \in [n]
\end{aligned}
$$

[4] Vielma JP et al (2017) Extended formulations in mixed-integer conic quadratic programming. *Mathematical Programming Computation*

Outer approximations of nonlinear functions?

Conic quadratic functions [4] (Handling the Lorentz cone)

$$
\begin{aligned}
& x_0 \geq \sqrt{\sum_{i=1}^{n} x_i^2} \\
\Leftrightarrow \; & x_0^2 \geq \sum_{i=1}^{n} x_i^2 \text{ and } x_0 \geq 0 \\
\Leftrightarrow \; & x_0 \geq \sum_{i=1}^{n} x_i^2 / x_0 \text{ and } x_0 \geq 0 \\
\Leftrightarrow \; & x_0 \geq \sum_{i=1}^{n} y_i \text{ and } x_0 \geq 0,\; y_i \geq x_i^2 / x_0, \forall i \in [n]
\end{aligned}
$$

20x speedup when first implemented

[4] Vielma JP et al (2017) Extended formulations in mixed-integer conic quadratic programming. *Mathematical Programming Computation*

Summary

- Lack of dual simplex hampers algorithms
- Several approach exist in the literature [5]
- Several popular approaches rely on linear outer approximations
- Effective implementations: integrated with branch-and-bound, calls to interior point method, addition of variables, reformulations...
- In practice, varying degrees of success

[5] Kronqvist J et al (2019) A review and comparison of solvers for convex MINLP. *Optimization and Engineering*

Agenda

1. Introduction
2. Branch and bound
3. Convexification
   - Convexification for MILO
   - Convexification for MINLO in sparse regression

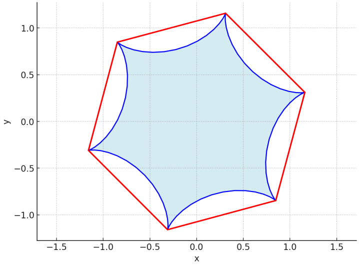
Agenda

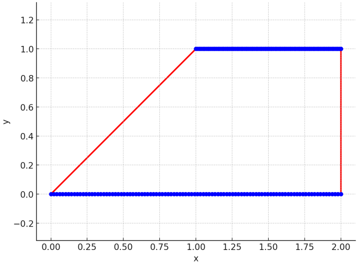
Convex hull

Definition

Given a set $X \subseteq \mathbb{R}^n$, the convex hull of X, denoted as $\mathrm{conv}(X)$, is

- The smallest convex set containing X
- The set of all convex combinations of points in X.





Convex optimization

Consider the optimization $$\min_{\boldsymbol{x} \in X} \boldsymbol{a}^\top \boldsymbol{x}$$

Proposition

If set X is convex, then any local minimum is a global minimum.

Intuition: Optimization over set X is "easy" under convexity

Convex optimization

Consider the optimization $$\min_{\boldsymbol{x}\in X} \boldsymbol{a}^\top \boldsymbol{x}$$

Proposition

If set X is convex, then any local minimum is a global minimum.

Intuition: Optimization over set X is "easy" under convexity

Proposition

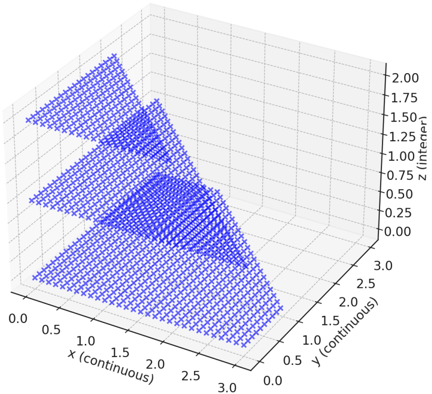
The optimization problem is equivalent to

$$\min_{\boldsymbol{x}\in conv(X)} \boldsymbol{a}^\top \boldsymbol{x},$$

i.e., there exist a solution that is optimal for both.

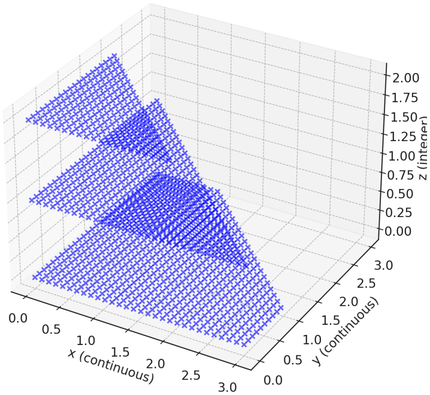
Intuition: Any optimization problem can be reduced to a convex problem

Convexification in mixed-integer linear optimization



What is the convex hull?

Convexification in mixed-integer linear optimization



$$x + y + z \leq 4$$
$$0 \leq x \leq 3$$
$$0 \leq y \leq 3$$
$$0 \leq z \leq 2, \; z \in \mathbb{Z}$$

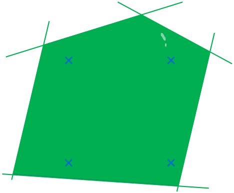
What is the convex hull?

Convexification in mixed-integer linear optimization

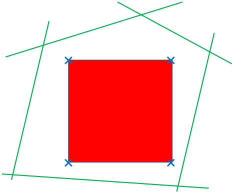
Proposition (Meyer 1974)

The convex hull of set $\{\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m : \boldsymbol{A}\boldsymbol{x} + \boldsymbol{G}\boldsymbol{z} \leq \boldsymbol{b}\}$ *is a polyhedron.*

Linear relaxation



Convex hull

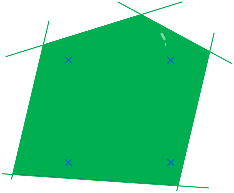


Convexification in mixed-integer linear optimization

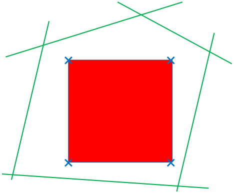
Proposition (Meyer 1974)

The convex hull of set $\{\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{z} \in \mathbb{Z}^m : \boldsymbol{Ax} + \boldsymbol{Gz} \leq \boldsymbol{b}\}$ *is a polyhedron.*

Linear relaxation

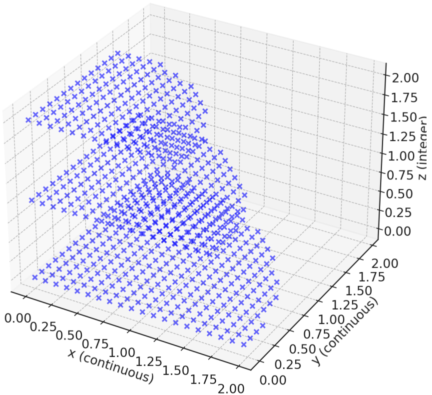


Convex hull



Instead of computing exact convex hulls, convexifications are dynamically added to branch-and-bound algorithms via cutting planes

Convexification in mixed-integer nonlinear optimization



$$x^2 + y^2 + z \leq 4$$
$$0 \leq x \leq 3$$
$$0 \leq y \leq 3$$
$$0 \leq z \leq 2, \ z \in \mathbb{Z}$$

What is the convex hull? How to implement in practice?

Agenda

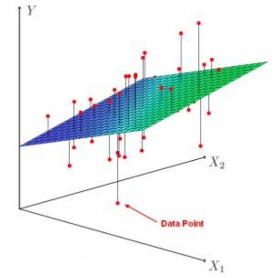
Least squares regression

Consider dataset[6] $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^m$ where $\boldsymbol{x_i} \in \mathbb{R}^n$

<u>Least squares with ridge regularization</u>

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^m (y_i - \boldsymbol{a_i}^\top \boldsymbol{x})^2 + \lambda \sum_{j=1}^n x_j^2$$

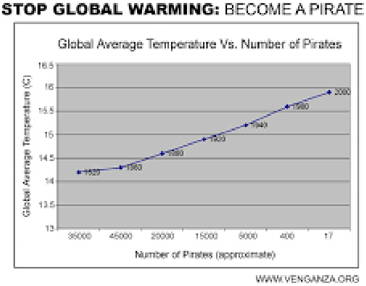
for some $\lambda \geq 0$



[6] Hoerl AE and Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*

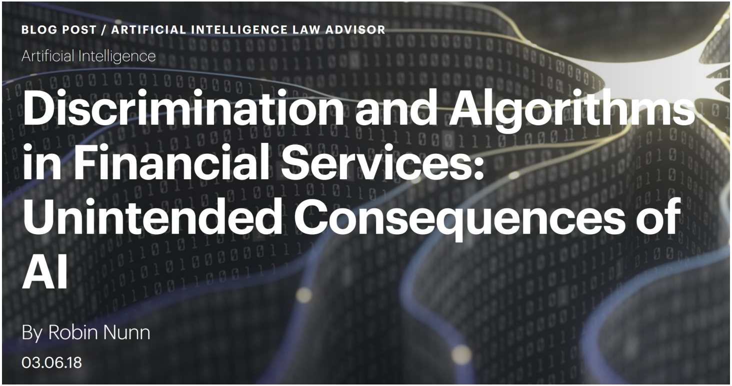
Shortcomings of ordinary least squares

Prone to overfitting



Can fail to make meaningful predictions out-of-sample

Shortcomings of ordinary least squares



In some cases, interpretability is far more important than accuracy

Linear regression

Least squares in action with the “Communities and crime” dataset

- Data with socio-economic data, law enforcement data and crime data (US census, LEMAS survey and FBI)
- $n = 100$ features, $m = 1993$ cities

Linear regression

Least squares in action with the "Communities and crime" dataset

- Data with socio-economic data, law enforcement data and crime data (US census, LEMAS survey and FBI)
- $n = 100$ features, $m = 1993$ cities

Solution metrics Optimal solution found in milliseconds, $R^2 = 0.84$

Linear regression

Least squares in action with the "Communities and crime" dataset

- Data with socio-economic data, law enforcement data and crime data (US census, LEMAS survey and FBI)
- $n = 100$ features, $m = 1993$ cities

Solution metrics Optimal solution found in milliseconds, $R^2 = 0.84$

Solution

HispPerCap	0.0266852	PctRecImmig8	0.0561141
HousVacant	0.200901	PctRecentImmig	-0.0213925
LandArea	0.0215684	PctSameCity85	0.0366286
LemasPctOfficDrugUn	0.0295357	PctSameHouse85	0.0108617
MalePctDivorce	0.109487	PctSameState85	0.00130485
MalePctNevMarr	0.0771952	PctSpeakEnglOnly	0.0196796
MedNumBR	0.0124658	PctTeen2Par	-0.0147065
MedOwnCostPctInc	-0.0132171	PctUnemployed	-0.0148311
MedOwnCostPctIncNoMtg	-0.0740048	PctUsePubTrans	-0.0304822
MedRent	0.0981299	PctVacMore6Mos	-0.0422866
MedRentPctHousInc	0.0525401	PctVacantBoarded	0.0537563
MedYrHousBuilt	0.0015325	PctWOFullPlumb	-0.00800257
NumIlleg	-0.0712478	PctWorkMom	-0.0756108
NumImmig	-0.105966	PctWorkMomYoungKids	-0.000117624
NumInShelters	0.144226	PctYoungKids2Par	-0.036353
NumStreet	0.186877	PersPerFam	0.0309863
NumUnderPov	0.0190623	PersPerOccupHous	0.0687344
OtherPerCap	0.0409795	PersPerOwnOccHous	-0.0114322
OwnOccHiQuart	0.0250454	PersPerRentOccHous	0.0191038
OwnOccLowQuart	-0.0585255	PopDens	0.0108191
OwnOccMedVal	0.00538469	RentHighQ	0.0262614
PctBSorMore	0.0219847	RentLowQ	-0.135663
PctBornSameState	0.00152456	RentMedian	0.0250774
PctEmplManu	-0.0393647	TotalPctDiv	0.032203
PctEmplProfServ	-0.00888359	agePct12t21	0.0242128
PctEmploy	0.0736816	agePct12t29	-0.0444774
PctFam2Par	-0.0370295	agePct16t24	-0.0332621
PctForeignBorn	0.0601135	agePct65up	0.0791789
PctHousLess3BR	0.0723027	blackPerCap	-0.0148187
PctHousNoPhone	0.0242783	householdsize	0.0170821
PctHousOccup	-0.0392774	indianPerCap	-0.0284771
PctHousOwnOcc	0.00969509	medFamInc	0.0181421
PctIlleg	0.196562	medIncome	0.0300747
PctImmigRec10	-0.00153328	numbUrban	-0.0920702
PctImmigRec5	-0.0117715	pctUrban	0.039124

Use parsimony

Occam's razor / Principle of parsimony (William of Ockham $\approx$ 1300)

Why did the tree fall?

- The wind
- Two meteorites crashed into earth. One hit the tree, the other hit the first meteorite, obliterating both and destroying the evidence

Use parsimony

Occam's razor / Principle of parsimony (William of Ockham $\approx$ 1300)

Why did the tree fall?

- The wind
- Two meteorites crashed into earth. One hit the tree, the other hit the first meteorite, obliterating both and destroying the evidence

Given two competing explanations, the simplest one is often right.

Use parsimony

Best subset selection

- Let k be the target complexity of the model. Among all $\binom{n}{k}$ subsets of k features, find the one that best fits the model

[7] Furnival G and Wilson R (1974) Regressions by leaps and bounds. *Technometrics*

Use parsimony

Best subset selection

- Let k be the target complexity of the model. Among all $\binom{n}{k}$ subsets of k features, find the one that best fits the model

- Solve

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^{m}\left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda\sum_{j=1}^{n} x_j^2$$

$$\text{s.t. } \sum_{j=1}^{n} \mathbb{1}_{\{x_j\neq 0\}} \leq k$$

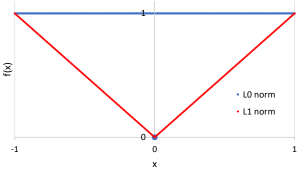
- Implemented[7] in R packages for $n < 30$

[7] Furnival G and Wilson R (1974) Regressions by leaps and bounds. *Technometrics*

Relaxations

Lasso/ elastic net (Tibshirani 1996, Zou and Hastie 2005)

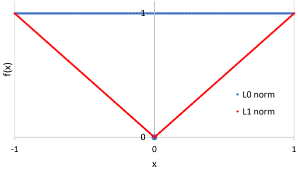
- The best convex underestimator of the "ℓ_0-norm" function $f(x) = \mathbb{1}_{\{x \neq 0\}}$ on $-1 \leq x \leq 1$ is the ℓ_1-norm $|x|$



Relaxations

Lasso/ elastic net (Tibshirani 1996, Zou and Hastie 2005)

- The best convex underestimator of the "ℓ_0-norm" function $f(x) = \mathbb{1}_{\{x \neq 0\}}$ on $-1 \leq x \leq 1$ is the ℓ_1-norm $|x|$



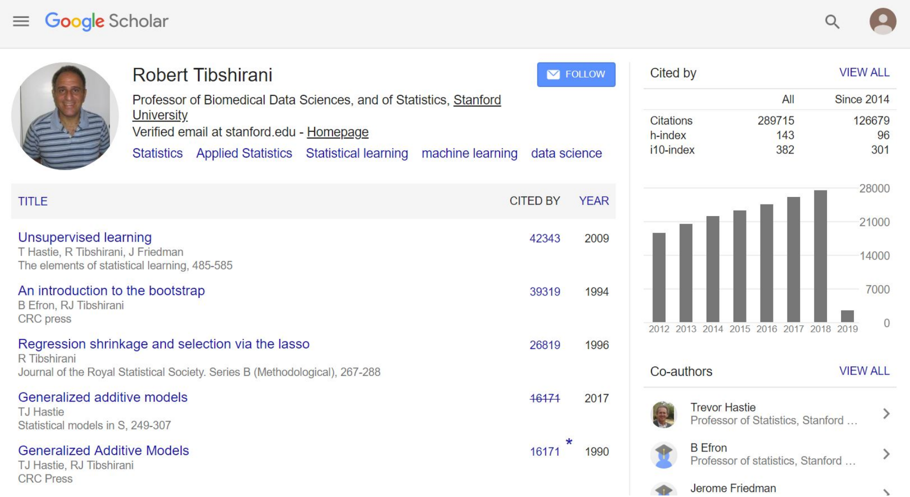
- Solve

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \sum_{i=1}^{m} \left(y_i - \boldsymbol{a}_i^\top \boldsymbol{x} \right)^2 + \lambda \sum_{j=1}^{n} x_j^2$$

$$\text{s.t. } \sum_{j=1}^{n} |x_j| \leq \kappa$$

where κ is a parameter (to be tuned) controlling sparsity vs accuracy

Relaxations



Relaxations

Google Scholar

Albert Einstein

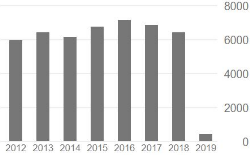
FOLLOW

Institute of Advanced Studies, Princeton
No verified email
Physics

TITLE	CITED BY	YEAR
Can quantum-mechanical description of physical reality be considered complete? A Einstein, B Podolsky, N Rosen Physical review 47 (10), 777	17493	1935
Uber einen die Erzeugung und Verwandlung des Lichtes betreffenden heurischen Gesichtpunkt A Einstein Ann. Phys. 17, 132-148	11362 *	1905
On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat A Einstein Annalen der Physik 17, 549-560	9690 *	1905
Zur Elektrodynamik bewegter Körper A Einstein	5504 *	
Sitzungsber. K A Einstein Preuss. Akad. Wiss., Phys. Math. Kl 3, 18	4950 *	1925
Graviton Mass and Inertia Mass A Einstein Ann Physik 35, 898	4814 *	1911

Cited by VIEW ALL

	All	Since 2014
Citations	124162	33880
h-index	113	64
i10-index	370	206



Relaxations

Lasso in action with the "Communities and crime" dataset ($n = 100$)

Solution metrics Optimal solution found in milliseconds

Solutions

Large κ ($R^2 = 0.81$)

HousVacant	0.237656
LemasPctOfficDrugUn	0.000206104
MalePctDivorce	0.0260223
NumStreet	0.14165
PctHousNoPhone	0.0266273
PctIlleg	0.311418
PctPersDenseHous	0.197454
PctVacantBoarded	0.0405917
PopDens	0.0193849
pctWPubAsst	0.0445904
racepctblack	0.186306

Small κ ($R^2 = 0.25$)

LandArea	0.121248
NumIlleg	0.685344
NumImmig	0.183812
PctPersDenseHous	0.000258902

Mixed-integer optimization

Mixed-integer optimization [8][9]

- Best subset selection can be formulated as a MIO
- Letting binary variable $z_j = 1$ iff feature j is included, solve

$$\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\{0,1\}^n} \quad & \sum_{i=1}^{m}\left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda \sum_{j=1}^{n} x_j^2 \\
\text{s.t.} \quad & \sum_{j=1}^{n} z_j \le k \\
& -Mz_j \le x_j \le Mz_j \qquad \forall j = 1, \dots, n
\end{aligned}$$

[8] Bertsimas D et al (2016) Best subset selection via a modern optimization lens. *The Annals of Statistics*

[9] Cozad A et al (2014) Learning surrogate models for simulation-based optimization. *AIChE.*

Mixed-integer optimization

MIO in action with the "Communities and crime" dataset ($n = 100$)

Solution ($R^2 = 0.81$)

HousVacant	0.250896
MalePctDivorce	0.135992
PctIlleg	0.524062
PctPersDenseHous	0.175159

Mixed-integer optimization

MIO in action with the "Communities and crime" dataset ($n = 100$)

Solution ($R^2 = 0.81$)

HousVacant	0.250896
MalePctDivorce	0.135992
PctIlleg	0.524062
PctPersDenseHous	0.175159

Solution metrics 20 hours to optimality, millions of nodes (Gurobi, 2022)

Mixed-integer optimization

Is best subset selection really worth it?[10]

- Best subset is slow
- Lasso is better in some situations, and can be improved otherwise

Of course![11]?

- Solution times are appropriate in many cases
- Lasso is better in very low SNR regimes, and best subset can be adapted

Both methods have merits![12]?

---

[10] Hastie T, Tibshirani R, Tibshirani R (2020) Best subset, forward stepwise or Lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*

[11] Mazumder R, Radchenko P, Dedieu A (2023) Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *Operations Research*

[12] Chen Y, Taeb A, Bühlmann P (2020) A look at robustness and stability of ℓ_1- versus ℓ_0-regularization: Discussion of papers by Bertsimas et al. and Hastie et al. *Statistical Science*

Improving the formulation

How good is the convex relaxation? $\boldsymbol{z} \in \{0,1\}^n \to \mathbf{0} \leq \boldsymbol{z} \leq \mathbf{1}$

$$\min_{\boldsymbol{x} \in \mathbb{R}^n, \mathbf{0} \leq \boldsymbol{z} \leq \mathbf{1}} \sum_{i=1}^{m} \left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x} \right)^2 + \lambda \sum_{j=1}^{n} x_j^2$$

$$\text{s.t.} \sum_{j=1}^{n} z_j \leq k$$

$$-Mz_j \leq x_j \leq Mz_j \qquad \forall j = 1, \ldots, n$$

Improving the formulation

How good is the convex relaxation? $\boldsymbol{z} \in \{0,1\}^n \to \mathbf{0} \leq \boldsymbol{z} \leq \mathbf{1}$

$$\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n, \mathbf{0}\leq\boldsymbol{z}\leq\mathbf{1}} \quad & \sum_{i=1}^{m} \left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda \sum_{j=1}^{n} x_j^2 \\
\text{s.t.} \quad & \sum_{j=1}^{n} z_j \leq k \\
& -Mz_j \leq x_j \leq Mz_j \qquad \forall j = 1,\dots,n
\end{aligned}$$

In an optimal solution, $z_j^* = |x_j|/M$

$\implies$ The continuous relaxation is in fact lasso!

Improving the formulation

How good is the convex relaxation? $\boldsymbol{z} \in \{0,1\}^n \to \mathbf{0} \le \boldsymbol{z} \le \mathbf{1}$

$$\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n, \mathbf{0}\le\boldsymbol{z}\le\mathbf{1}} \quad & \sum_{i=1}^{m}\left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda \sum_{j=1}^{n} x_j^2 \\
\text{s.t.} \quad & \sum_{j=1}^{n} z_j \le k \\
& -Mz_j \le x_j \le Mz_j \qquad \forall j = 1,\dots,n
\end{aligned}$$

In an optimal solution, $z_j^* = |x_j|/M$

$\Longrightarrow$ The continuous relaxation is in fact lasso!

Since lasso is not a good approximation, this formulation is slow...

Improving the formulation

Improve the convex relaxation Need to exploit nonlinearities

$$
\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\{0,1\}^n,\boldsymbol{t}\in\mathbb{R}^n_+} \; & \sum_{i=1}^{m}\left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda \sum_{j=1}^{n} t_j \\
\text{s.t. } & x_j^2 \le t_j \qquad \forall j = 1,\ldots,n \\
& \sum_{j=1}^{n} z_j \le k \\
& -Mz_j \le x_j \le Mz_j \qquad \forall j = 1,\ldots,n
\end{aligned}
$$

Improving the formulation

Improve the convex relaxation Need to exploit nonlinearities

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\{0,1\}^n, \boldsymbol{t}\in\mathbb{R}^n_+} \sum_{i=1}^{m}\left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda \sum_{j=1}^{n} t_j$$

$$\text{s.t. } x_j^2 \le t_j \qquad \forall j = 1, \ldots, n$$

$$\sum_{j=1}^{n} z_j \le k$$

$$-Mz_j \le x_j \le Mz_j \qquad \forall j = 1, \ldots, n$$

What is the convex hull of

$$S = \left\{x \in \mathbb{R}, z \in \{0,1\}, t \in \mathbb{R} : x^2 \le t, x(1-z) = 0\right\}?$$

Improving the formulation

$$S = \underbrace{\left\{(x, z, t) \in \mathbb{R}^3 : 0 \le t, x = z = 0\right\}}_{S_1} \cup \underbrace{\left\{(x, z, t) \in \mathbb{R}^3 : x^2 \le t, z = 1\right\}}_{S_2}?$$

$(x, z, t) \in \text{conv}(S)$ if and only $\exists (x_i, z_i, t_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$\begin{aligned}
&x = \lambda_1 x_1 + \lambda_2 x_2,\ z = \lambda_1 z_1 + \lambda_2 z_2,\ t = \lambda_1 t_1 + \lambda_2 t_2 \\
&\lambda_1 + \lambda_2 = 1,\ \lambda_1 \ge 0,\ \lambda_2 \ge 0 \\
&(x_1, z_1, t_1) \in S_1 \Leftrightarrow x_1 = z_1 = 0, t_1 \ge 0 \\
&(x_2, z_2, t_2) \in S_2 \Leftrightarrow x_2^2 \le t_2, z_2 = 1
\end{aligned}$$

Improving the formulation

$$S = \underbrace{\left\{(x,z,t)\in\mathbb{R}^3 : 0\leq t, x=z=0\right\}}_{S_1} \cup \underbrace{\left\{(x,z,t)\in\mathbb{R}^3 : x^2\leq t, z=1\right\}}_{S_2}?$$

$(x,z,t)\in \text{conv}(S)$ if and only $\exists (x_i,z_i,t_i)\in S_i$ and $\lambda_i\in\mathbb{R}^i$ such that

$$\begin{aligned}
&x=\lambda_1 x_1+\lambda_2 x_2,\ z=\lambda_1 z_1+\lambda_2 z_2,\ t=\lambda_1 t_1+\lambda_2 t_2\\
&\lambda_1+\lambda_2=1,\ \lambda_1\geq 0,\ \lambda_2\geq 0\\
&(x_1,z_1,t_1)\in S_1 \Leftrightarrow x_1=z_1=0, t_1\geq 0\\
&(x_2,z_2,t_2)\in S_2 \Leftrightarrow x_2^2\leq t_2, z_2=1
\end{aligned}$$

Change of variables: $\tilde{x}_i=x_i\lambda_i,\ \tilde{z}_i=z_i\lambda_i,\ \tilde{t}_i=t_i\lambda_i$

Improving the formulation

$$S = \underbrace{\left\{(x,z,t) \in \mathbb{R}^3 : 0 \le t, x = z = 0\right\}}_{S_1} \cup \underbrace{\left\{(x,z,t) \in \mathbb{R}^3 : x^2 \le t, z = 1\right\}}_{S_2}?$$

$(x,z,t) \in \text{conv}(S)$ if and only $\exists (x_i, z_i, t_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$\begin{aligned}
& x = \tilde{x}_1 + \tilde{x}_2,\ z = \tilde{z}_1 + \tilde{z}_2,\ t = \tilde{t}_1 + \tilde{t}_2 \\
& \lambda_1 + \lambda_2 = 1,\ \lambda_1 \ge 0,\ \lambda_2 \ge 0 \\
& (x_1, z_1, t_1) \in S_1 \Leftrightarrow \tilde{x}_1 = \tilde{z}_1 = 0, \tilde{t}_1 \ge 0 \\
& (x_2, z_2, t_2) \in S_2 \Leftrightarrow (\tilde{x}_2/\lambda_2)^2 \le \tilde{t}_2/\lambda_2, \tilde{z}_2/\lambda_2 = 1
\end{aligned}$$

Change of variables: $\tilde{x}_i = x_i\lambda_i,\ \tilde{z}_i = z_i\lambda_i,\ \tilde{t}_i = t_i\lambda_i$

Improving the formulation

$$S = \underbrace{\left\{(x,z,t)\in\mathbb{R}^3 : 0\le t, x=z=0\right\}}_{S_1} \cup \underbrace{\left\{(x,z,t)\in\mathbb{R}^3 : x^2\le t, z=1\right\}}_{S_2}?$$

$(x,z,t)\in \text{conv}(S)$ if and only $\exists (x_i,z_i,t_i)\in S_i$ and $\lambda_i\in\mathbb{R}^i$ such that

$$\begin{aligned}
&x=\tilde{x}_1+\tilde{x}_2,\ z=\tilde{z}_1+\tilde{z}_2,\ t=\tilde{t}_1+\tilde{t}_2\\
&\lambda_1+\lambda_2=1,\ \lambda_1\ge 0,\ \lambda_2\ge 0\\
&(\tilde{x}_1,\tilde{z}_1,\tilde{t}_1)\in\lambda_1 S_1 \Leftrightarrow \tilde{x}_1=\tilde{z}_1=0, \tilde{t}_1\ge 0\\
&(\tilde{x}_2,\tilde{z}_2,\tilde{t}_2)\in\lambda_2 S_2 \Leftrightarrow \tilde{x}_2^2/\lambda_2\le\tilde{t}_2, \tilde{z}_2=\lambda_2
\end{aligned}$$

Change of variables: $\tilde{x}_i=x_i\lambda_i,\ \tilde{z}_i=z_i\lambda_i,\ \tilde{t}_i=t_i\lambda_i$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$
\begin{aligned}
& x = \tilde{x}_1 + \tilde{x}_2,\ z = \tilde{z}_1 + \tilde{z}_2, t = \tilde{t}_1 + \tilde{t}_2 \\
& \lambda_1 + \lambda_2 = 1,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0 \\
& \tilde{x}_1 = \tilde{z}_1 = 0, \tilde{t}_1 \geq 0 \\
& \tilde{x}_2^2 / \lambda_2 \leq \tilde{t}_2,\ \tilde{z}_2 = \lambda_2
\end{aligned}
$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$
\begin{aligned}
& x = \tilde{x}_1 + \tilde{x}_2,\ z = \tilde{z}_1 + \tilde{z}_2, t = \tilde{t}_1 + \tilde{t}_2 \\
& \lambda_1 + \lambda_2 = 1,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0 \\
& \tilde{x}_1 = \tilde{z}_1 = 0, \tilde{t}_1 \geq 0 \\
& \tilde{x}_2^2/\lambda_2 \leq \tilde{t}_2,\ \tilde{z}_2 = \lambda_2
\end{aligned}
$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$\begin{aligned}
& x = \tilde{x}_2,\ z = \tilde{z}_2, t = \tilde{t}_1 + \tilde{t}_2 \\
& \lambda_1 + \lambda_2 = 1,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0 \\
& \tilde{t}_1 \geq 0 \\
& \tilde{x}_2^2/\lambda_2 \leq \tilde{t}_2,\ \tilde{z}_2 = \lambda_2
\end{aligned}$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$
\begin{aligned}
& x = \tilde{x}_2,\ z = \tilde{z}_2, t = \tilde{t}_1 + \tilde{t}_2 \\
& \lambda_1 + \lambda_2 = 1,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0 \\
& \tilde{t}_1 \geq 0 \\
& \tilde{x}_2^2/\lambda_2 \leq \tilde{t}_2,\ \tilde{z}_2 = \lambda_2
\end{aligned}
$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$\begin{aligned}
&x = \tilde{x}_2,\ z = \lambda_2, t = \tilde{t}_1 + \tilde{t}_2\\
&\lambda_1 + \lambda_2 = 1,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0\\
&\tilde{t}_1 \geq 0\\
&\tilde{x}_2^2/\lambda_2 \leq \tilde{t}_2
\end{aligned}$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$x = \tilde{x}_2,\ z = \lambda_2, t = \tilde{t}_1 + \tilde{t}_2$$
$$\lambda_1 + \lambda_2 = 1,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0$$
$$\tilde{t}_1 \geq 0$$
$$\tilde{x}_2^2/\lambda_2 \leq \tilde{t}_2$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$
\begin{aligned}
& t = \tilde{t}_1 + \tilde{t}_2 \\
& \lambda_1 + z_2 = 1,\ \lambda_1 \geq 0,\ z \geq 0 \\
& \tilde{t}_1 \geq 0 \\
& x^2/z \leq \tilde{t}_2
\end{aligned}
$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$
\begin{aligned}
& t = \tilde{t}_1 + \tilde{t}_2 \\
& \lambda_1 + z_2 = 1, \ \lambda_1 \geq 0, \ z \geq 0 \\
& \tilde{t}_1 \geq 0 \\
& x^2/z \leq \tilde{t}_2
\end{aligned}
$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$
\begin{aligned}
& t = \tilde{t}_1 + \tilde{t}_2 \\
& z \leq 1,\ z \geq 0 \\
& \tilde{t}_1 \geq 0 \\
& x^2/z \leq \tilde{t}_2
\end{aligned}
$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$
\begin{aligned}
& t = \tilde{t}_1 + \tilde{t}_2 \\
& z \leq 1,\ z \geq 0 \\
& \tilde{t}_1 \geq 0 \\
& x^2/z \leq \tilde{t}_2
\end{aligned}
$$

Improving the formulation

$(x, z, t) \in \text{conv}(S)$ if and only[13] $\exists(\tilde{x}_i, \tilde{z}_i, \tilde{t}_i) \in S_i$ and $\lambda_i \in \mathbb{R}^i$ such that

$$t \geq x^2/z,\ 0 \leq z \leq 1$$

Proposition (Frangioni and Gentile 2006)

The convex hull of set

$$S = \left\{x \in \mathbb{R}, z \in \{0,1\}, t \in \mathbb{R} : x^2 \leq t, x(1-z) = 0\right\}$$

is

$$conv(S) = \left\{(x, z, t) \in \mathbb{R}^3 : x^2 \leq tz,\ 0 \leq z \leq 1\right\}$$

[13] Frangioni A and Gentile C (2006) Perspective cuts for a class of convex 0-1 mixed-integer programs. *Mathematical Programming*

Improving the formulation

Improve the convex relaxation [14][15]

$$\begin{aligned}\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\{0,1\}^n,\boldsymbol{t}\in\mathbb{R}^n_+} \quad & \sum_{i=1}^m \left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda \sum_{j=1}^n t_j \\ \text{s.t.} \quad & x_j^2 \le t_j \qquad \forall j = 1,\dots,n \\ & \sum_{j=1}^n z_j \le k \\ & -Mz_j \le x_j \le Mz_j \qquad \forall j = 1,\dots,n\end{aligned}$$

[14] Dong H et al (2018) Regularization vs relaxation: A convexification perspective of statistical variable selection. *Optimization Online*

[15] Xie W and Deng X (2020) Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization*

Improving the formulation

Improve the convex relaxation [14][15]

$$\begin{aligned}
\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\{0,1\}^n,\boldsymbol{t}\in\mathbb{R}^n_+} \quad & \sum_{i=1}^{m}\left(y_i - \boldsymbol{a_i}^\top \boldsymbol{x}\right)^2 + \lambda \sum_{j=1}^{n} t_j \\
\text{s.t.} \quad & x_j^2 \le t_j z_j \qquad \forall j = 1,\ldots,n \\
& \sum_{j=1}^{n} z_j \le k \\
& -Mz_j \le x_j \le Mz_j \qquad \forall j = 1,\ldots,n
\end{aligned}$$

The perspective reformulation! (Disjunctive programming)

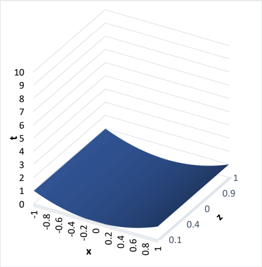
[14] Dong H et al (2018) Regularization vs relaxation: A convexification perspective of statistical variable selection. *Optimization Online*

[15] Xie W and Deng X (2020) Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization*

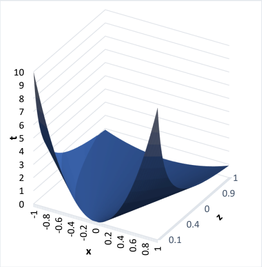
Improving the formulation

Constraint $tz \geq x^2$ with $t, z \geq 0$ is convex and SOCP representable

It represents a <u>substantial</u> improvement in the relaxation quality



Graph of $t = x^2$ (big-M)

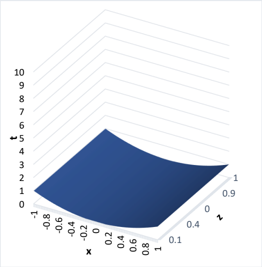


Graph of $t = x^2/z$

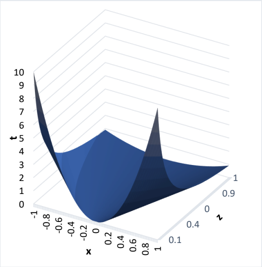
Improving the formulation

Constraint $tz \geq x^2$ with $t, z \geq 0$ is convex and SOCP representable

It represents a <u>substantial</u> improvement in the relaxation quality



Graph of $t = x^2$ (big-M)



Graph of $t = x^2/z$

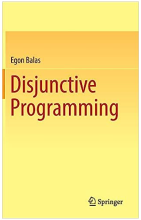
Solution times in "Communities and crime": 2s ($15,000\times$ speedup)

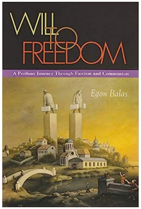
Disjunctive programming

Disjunctive programming was invented by Egon Balas in the 80s

`https://www.wsj.com/articles/`

`egon-balas-jailed-and-tortured-in-romania-found-salvation-in-math-11553869800`





Generalizes to nonlinear optimization[16]

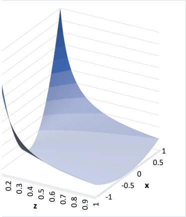
[16] Ceria S and Soares J (1999) Convex programming for disjunctive convex optimization. *Mathematical Programming*

Disjunctive Programming

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, consider

$$f^\pi(\boldsymbol{x}, \lambda) = \begin{cases} \lambda f(\boldsymbol{x}/\lambda) & \text{if } \lambda > 0 \\ \lim_{\lambda \to 0^+} \lambda f(\boldsymbol{x}/\lambda) & \text{if } \lambda = 0 \\ +\infty & \textit{otherwise}. \end{cases}$$

- f^π is convex and homogeneous
- If $f(\boldsymbol{x}) = a_0 + \boldsymbol{a}^\top \boldsymbol{x}$, then $f^\pi(\boldsymbol{x}, \lambda) = a_0\lambda + \boldsymbol{a}^\top \boldsymbol{x}$
- If $f(x) = x^2$, then $f^\pi(x, z) = x^2/z$ with $0/0 = 0$ and $x^2/0 = +\infty$ if $x \neq 0$

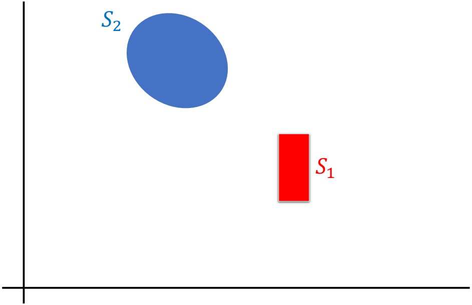


Disjunctive programming

For $i \in \{1, \ldots, k\}$, let $S_i = \{x \in \mathbb{R}^n : g_{ij}(x) \leq 0,\ j = 1 \ldots, m\}$ convex.
Then $x \in \text{cl conv}\left(\bigcup_{i=1}^k S_i\right)$ iff $\exists x^i \in \mathbb{R}^n$ and $\lambda^i \in \mathbb{R}_+$ such that

$$x^i \in S_i^\pi(\lambda) = \left\{x \in \mathbb{R}^n : g_{ij}^\pi(x, \lambda^i) \leq 0,\ j = 1 \ldots, m\right\}$$

$$x = \sum_{i=1}^k x^i, \text{ and } 1 = \sum_{i=1}^k \lambda^i.$$

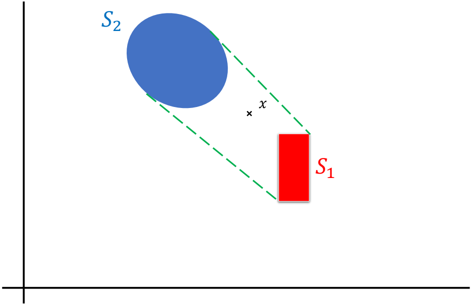


Disjunctive programming

For $i \in \{1, \ldots, k\}$, let $S_i = \{x \in \mathbb{R}^n : g_{ij}(x) \leq 0,\ j = 1 \ldots, m\}$ convex.
Then $x \in \text{cl conv}\left(\bigcup_{i=1}^{k} S_i\right)$ iff $\exists x^i \in \mathbb{R}^n$ and $\lambda^i \in \mathbb{R}_+$ such that

$$x^i \in S_i^\pi(\lambda) = \left\{x \in \mathbb{R}^n : g_{ij}^\pi(x, \lambda^i) \leq 0,\ j = 1 \ldots, m\right\}$$

$$x = \sum_{i=1}^{k} x^i, \text{ and } 1 = \sum_{i=1}^{k} \lambda^i.$$

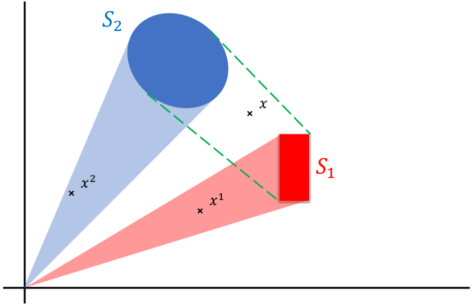


Disjunctive programming

For $i \in \{1, \ldots, k\}$, let $S_i = \{x \in \mathbb{R}^n : g_{ij}(x) \leq 0,\ j = 1 \ldots, m\}$ convex.
Then $x \in \text{cl conv}\left(\bigcup_{i=1}^k S_i\right)$ iff $\exists x^i \in \mathbb{R}^n$ and $\lambda^i \in \mathbb{R}_+$ such that

$$x^i \in S_i^\pi(\lambda) = \left\{x \in \mathbb{R}^n : g_{ij}^\pi(x, \lambda^i) \leq 0,\ j = 1 \ldots, m\right\}$$

$$x = \sum_{i=1}^k x^i, \text{ and } 1 = \sum_{i=1}^k \lambda^i.$$

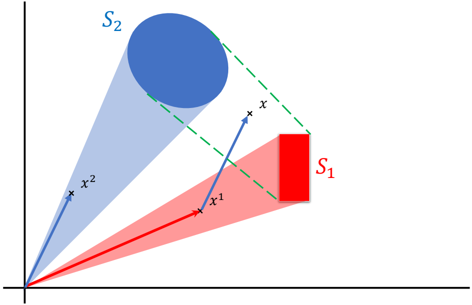


Disjunctive programming

For $i \in \{1, \ldots, k\}$, let $S_i = \{x \in \mathbb{R}^n : g_{ij}(x) \leq 0,\ j = 1 \ldots, m\}$ convex.
Then $x \in \text{cl conv}\left(\bigcup_{i=1}^k S_i\right)$ iff $\exists x^i \in \mathbb{R}^n$ and $\lambda^i \in \mathbb{R}_+$ such that

$$x^i \in S_i^\pi(\lambda) = \left\{x \in \mathbb{R}^n : g_{ij}^\pi(x, \lambda^i) \leq 0,\ j = 1 \ldots, m\right\}$$

$$x = \sum_{i=1}^k x^i, \text{ and } 1 = \sum_{i=1}^k \lambda^i.$$



Disjunctive programming

Implications of disjunctive programming Given any disjunctive set, we can create an equivalent convex (conic-representable) representation by creating k copies x^i of variables x, and mk constraints $g_{ij}^{\pi}(x^i, \lambda^i) \leq 0$. In other words, formulation increases by a factor of k.

Is it useful?

Disjunctive programming

Implications of disjunctive programming Given any disjunctive set, we can create an equivalent convex (conic-representable) representation by creating k copies x^i of variables x, and mk constraints $g_{ij}^{\pi}(x^i, \lambda^i) \leq 0$. In other words, formulation increases by a factor of k.

Is it useful? If used carefully

- Number of additional variables can grow exponentially
- Fourier–Motzkin elimination can be difficult in closed form
- Cuts from disjunctive programming may be hard to implement

Implementation of disjunctive programming

$$S = \bigcup_{i=1}^{\ell} \{\boldsymbol{x} \in \mathbb{R}^n : f_i(\boldsymbol{x}) \leq 0\}$$

Implementation 1 Add variables $\{(\boldsymbol{x}^i, \lambda_i) \in \mathbb{R}^{n+1}\}_{i=1}^{\ell}$ and formulate as

$$\boldsymbol{x} = \sum_{i=1}^{\ell} \boldsymbol{x^i},\ 1 = \sum_{i=1}^{\ell} \lambda_i,\ \boldsymbol{\lambda} \geq \boldsymbol{0}$$
$$f_i^{\pi}(\boldsymbol{x^i}, \lambda_i) \leq 0 \quad \forall \in \{1, \ldots, \ell\}$$

- Can be effective when ℓ is small (e.g., $\ell = 2$)
- But number of variables and constraints may be prohibitive...

Implementation of disjunctive programming

$$S = \bigcup_{i=1}^{\ell} \{\boldsymbol{x} \in \mathbb{R}^n : f_i(\boldsymbol{x}) \leq 0\}$$

Implementation 2 Add linear cuts using representation

$$\begin{aligned} 0 \geq \max_{\boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}, \boldsymbol{\gamma}\in\mathbb{R}_+^{\ell}} & \boldsymbol{\alpha}^\top \boldsymbol{x} + \beta \\ & + \min_{\{(\boldsymbol{x^i}, \lambda_i)\}} \left\{ -\sum_{i=1}^{\ell} \boldsymbol{\alpha}^\top \boldsymbol{x^i} - \sum_{i=1}^{\ell} \beta\lambda_i + \sum_{i=1}^{\ell} \gamma_i f_i^\pi(\boldsymbol{x^i}, \lambda_i) \right\} \end{aligned}$$

- Given fixed $\boldsymbol{x}$, solve separation (max) and add linear cut
- Requires computing Fenchel conjugates (min)
- But linear cuts can be ineffective...

Implementation of disjunctive programming

$$S = \bigcup_{i=1}^{\ell} \{\boldsymbol{x} \in \mathbb{R}^n : f_i(\boldsymbol{x}) \leq 0\}$$

Implementation 3 Add nonlinear cuts (e.g., Fourier-Motzkin with duality)

- May achieve a good compromise between convex hull and linear cuts
- But adding nonlinear cuts in branch-and-bound is not easy
  - Not supported in OA branch-and-bound solvers
  - Could require column generation to implement effectively
  - May be of different classes than original function

Low-rank functions

What is the convex hull of

$$S = \left\{ \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \{0,1\}^n, t \in \mathbb{R} : \left(\boldsymbol{a}^\top \boldsymbol{x} \right)^2 \leq t,\ \boldsymbol{x} \circ (\mathbf{1} - \boldsymbol{z}) \right\}$$

where "$\circ$" is the entrywise product, i.e., $\boldsymbol{x} \circ (\mathbf{1} - \boldsymbol{z}) \Leftrightarrow x_i(1 - z_i) = 0$

Low-rank functions

What is the convex hull of

$$S = \left\{ \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \{0,1\}^n, t \in \mathbb{R} : \left(\boldsymbol{a}^\top \boldsymbol{x}\right)^2 \leq t,\ \boldsymbol{x} \circ (\mathbf{1} - \boldsymbol{z}) \right\}$$

where "$\circ$" is the entrywise product, i.e., $\boldsymbol{x} \circ (\mathbf{1} - \boldsymbol{z}) \Leftrightarrow x_i(1 - z_i) = 0$

Disjunctive programming

$$S = \bigcup_{\bar{\boldsymbol{z}} \in \{0,1\}^n} \left\{ (\boldsymbol{x}, \boldsymbol{z}, t) \in \mathbb{R}^{2n+1} : (\boldsymbol{a}^\top \boldsymbol{x})^2 \leq t,\ \boldsymbol{x} \circ (\mathbf{1} - \bar{\boldsymbol{z}}) = \mathbf{0} \right\}$$

$\Longrightarrow$ Exponential number of variables/constraints

Low-rank functions

Consider optimization over set S,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\{0,1\}^n} \boldsymbol{c}^\top\boldsymbol{z}+\boldsymbol{d}^\top\boldsymbol{x}+\left(\boldsymbol{a}^\top\boldsymbol{x}\right)^2 \text{ s.t. } \boldsymbol{x}\circ(\mathbf{1}-\boldsymbol{z})$$

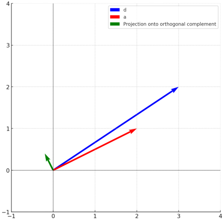
Low-rank functions

Consider optimization over set S,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{z}\in\{0,1\}^n} \boldsymbol{c}^\top\boldsymbol{z}+\boldsymbol{d}^\top\boldsymbol{x}+\left(\boldsymbol{a}^\top\boldsymbol{x}\right)^2 \text{ s.t. } \boldsymbol{x}\circ(\mathbf{1}-\boldsymbol{z})$$

$\boldsymbol{d}\neq\mu\boldsymbol{a}$ for some $\mu\in\mathbb{R} \implies \exists\boldsymbol{h}\in\mathbb{R}^n$ such that $\boldsymbol{h}^\top\boldsymbol{a}=0$ and $\boldsymbol{h}^\top\boldsymbol{d}<0$

$\implies$ Unbounded, letting $\boldsymbol{z}=\mathbf{1}$ and $\boldsymbol{x}=\gamma\boldsymbol{h}$ with $\gamma\to\infty$



Low-rank functions

Consider optimization over set S,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\{0,1\}^n} \boldsymbol{c}^\top \boldsymbol{z} + \boldsymbol{d}^\top \boldsymbol{x} + \left(\boldsymbol{a}^\top \boldsymbol{x}\right)^2 \text{ s.t. } \boldsymbol{x} \circ (\mathbf{1} - \boldsymbol{z})$$

If $\boldsymbol{d} = \mu \boldsymbol{a}$, then optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\{0,1\}^n, y\in\mathbb{R}} \boldsymbol{c}^\top \boldsymbol{z} + \mu y + y^2 \text{ s.t. } y = \boldsymbol{a}^\top \boldsymbol{x},\ \boldsymbol{x} \circ (\mathbf{1} - \boldsymbol{z})$$

has an optimal solution with at most one non-zero x_i

$\implies$ All extreme points of conv(S) have at most one non-zero x_i

Low-rank functions

Given any convex function[17] $f : \mathbb{R}^k \to \mathbb{R}$ and $\boldsymbol{A} \in \mathbb{R}^{k \times n}$, define

$$S = \{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \{0,1\}^n, t \in \mathbb{R} : t \geq f(\boldsymbol{Ax}),\ \boldsymbol{x} \circ (\boldsymbol{1} - \boldsymbol{z}) = \boldsymbol{0}\}$$

Proposition (Han and Gómez 2024)

$$cl\ conv(S) = cl\ conv\left(\bigcup_{\mathcal{I} \subseteq [n] : |\mathcal{I}| \leq k} V(\mathcal{I}) \cup R \right)$$

where

$$V(\mathcal{I}) = \{\{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \{0,1\}^n, t \in \mathbb{R} : t \geq f(\boldsymbol{Ax}),\ x_i = 0\ \forall i \notin \mathcal{I},\ z_i = 1\ \forall i \in \mathcal{I}\}\}$$
$$R = \{\{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \{0,1\}^n, t \in \mathbb{R} : t \geq 0,\ \boldsymbol{Ax} = \boldsymbol{0},\ \boldsymbol{z} = \boldsymbol{1}\}\}$$

[17] Han S and Gómez A (2024) Compact extended formulations for low rank functions with indicators. *Mathematics of Operations Research*

Low-rank functions

Computing

$$\mathrm{cl\ conv}\left(\bigcup_{\mathcal{I}\subseteq[n]:|\mathcal{I}|\leq k} V(\mathcal{I})\cup R\right)\ \text{vs.}\ \mathrm{cl\ conv}\left(\bigcup_{\mathcal{I}\subseteq[n]} V(\mathcal{I})\right)$$

involves $\mathcal{O}(n^k)$ vs 2^n disjunctions

Low-rank functions

Computing

$$\text{cl conv}\left(\bigcup_{\mathcal{I}\subseteq[n]:|\mathcal{I}|\leq k} V(\mathcal{I}) \cup R\right) \text{ vs. } \text{cl conv}\left(\bigcup_{\mathcal{I}\subseteq[n]} V(\mathcal{I})\right)$$

involves $\mathcal{O}(n^k)$ vs 2^n disjunctions

For special case of $k = 1$,

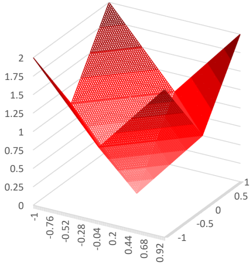
$$S = \left\{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \{0,1\}^n, t \in \mathbb{R} : (\boldsymbol{a}^\top \boldsymbol{x})^2 \leq t, \boldsymbol{x} \circ (\boldsymbol{1} - \boldsymbol{z}) = \boldsymbol{0}\right\},$$

we find after Fourier-Motzkin elimination that

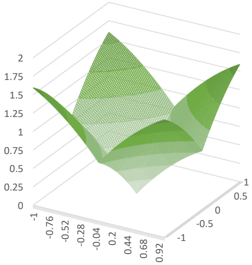
$$\text{cl conv}(S) = \left\{(\boldsymbol{x}, \boldsymbol{z}, t) \in \mathbb{R}^{2n+1} : (\boldsymbol{a}^\top \boldsymbol{x})^2 / \min\{1, \boldsymbol{1}^\top \boldsymbol{z}\} \leq t,\ \boldsymbol{0} \leq \boldsymbol{z} \leq \boldsymbol{1}\right\}$$

Rank-one convexification in sparse regression

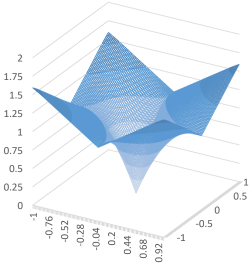
Can be interpreted as strong regularization[18]



Lasso as regularization



Perspective as regularization



Rank-one as regularization

- In tall instances ($n << m$), solution from relaxation is integral in practice
- But relaxation is more sophisticated (SOCP$\rightarrow$ SDP)

[18] Atamtürk A and Gómez A (2025) Rank-one convexifications for sparse regression. *Journal of Machine Learning Research.*

Full implementation

Tailored branch-and-bound algorithm based on perspective relaxation[19]

- Project out unnecessary variables
- Coordinate descent to solve relaxations
- Active sets
- Dual bounds
- Strong branching

[19] Hazimeh H et al (2022) Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Mathematical Programming*

Full implementation

Tailored branch-and-bound algorithm based on perspective relaxation[19]

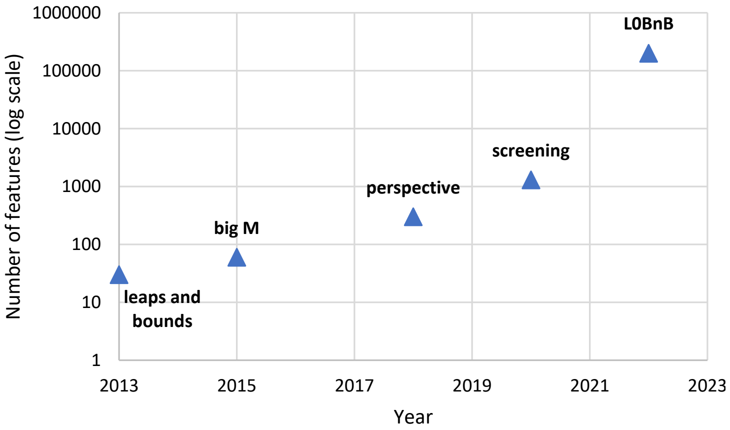
- Project out unnecessary variables
- Coordinate descent to solve relaxations
- Active sets
- Dual bounds
- Strong branching

p	L0BnB	GRB	MSK	B
10^3	**0.7**	70	92	(4%)
10^4	**3**	(15%)	1697	–
10^5	**34**	–	–	–
10^6	**1112**	–	–	–

Time comparison (in seconds) with Gurobi, Mosek and Baron

[19] Hazimeh H et al (2022) Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Mathematical Programming*

The journey so far...



Dimension of problems that can be comfortably solved

Conclusion

- Convexification is harder than in MILO
- Some methods extend (less intuitive)
  - Disjunctive programming
  - RLTs
  - Lifting
- Implementation is non-trivial
- ... but it can work