Mixed integer programming (MIP) for machine learning

Stéphane Canu

asi.insa-rouen.fr/enseignants/~scanu

Joint work with

Ruobing Shen	Yuan LIU	Mehde Jammal	Ismaila Seck
Heidelberg (D)	INSA Rouen	Baalbeck (Lebanon)	INSA Rouen

and

G. Reinelt, P. Honeine, S. Ruan & G. Loosli

Journée « autour de l'optimisation » de l'axe DAC, Rouen

May 23, 2019

Road map

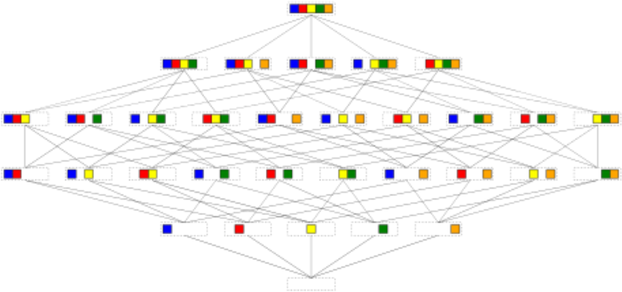
1. Examples of combinatorial problems in machine learning
   - L_0 norm

2. MIP for variable selection AND outlier detection
   - MIP for variable selection
   - L_0 proximal algorithm
   - Implementation

NP Hard =

Variable selection

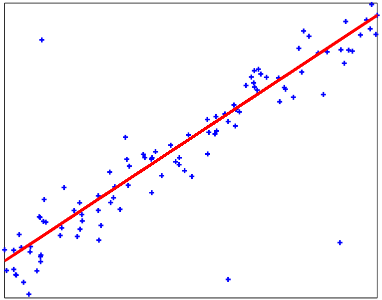
$$f(x_1, \ldots, x_j, \ldots, x_p) = \sum_{j=1}^{p=5} x_j w_j$$

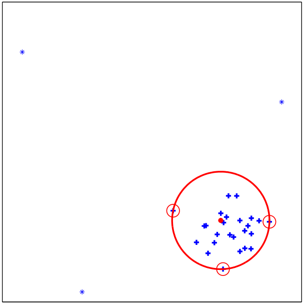


Fit the data and remove useless variables

Enumerate of all possible combinations and choose

Outlier detection

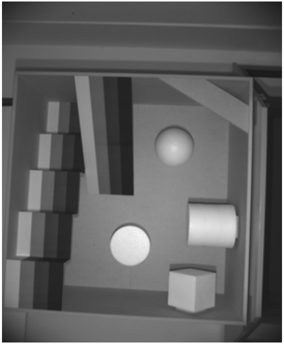


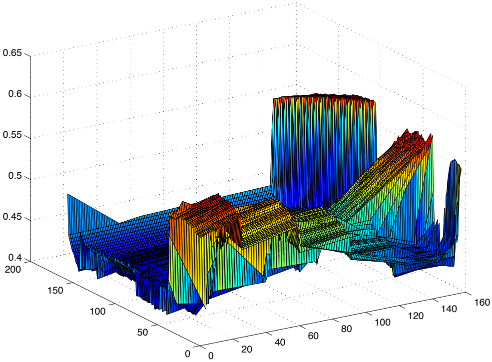


Fit the data and remove useless observations (outliers)

Enumerate of all possible point configurations and choose

Deepth estimation





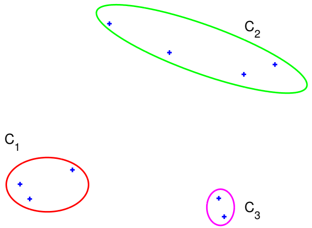
Fundamental hypothesis

piecewise linear model (counting the number of pieces)

Clustering

$$Z_{ij} = \begin{cases} 1 & \text{if observation } i \text{ belongs to cluster } j \\ 0 & \text{else} \end{cases}$$

Observation	Cluster 1	Cluster 2	Cluster 3
x_1	1	0	0
x_2	0	1	0
x_3	1	0	0
x_4	0	0	1
x_5	0	1	0
x_6	0	0	1
x_7	0	1	0
x_8	1	0	0
x_9	0	1	0



Minimize some energy within the clusters

Enumerate of all possible $(\{0,1\},\{0,1\},\{0,1\})^n$ configurations such that each point belongs to only one cluster.

This is a $k = 3$-partition problem.

Robustness of a NN

Let f be a neural network $\begin{array}{rccc} f: & [0,1]^p & \longrightarrow & \mathbb{R}^c \\ & x & \longmapsto & f(x) \end{array}$

Assume f is **piecewise linear** (e.g. $f(x) = V \; ReLU(Wx)$)

The neural network is ε–robust at x if $\varepsilon < \varepsilon'$ where

$$\varepsilon' = \begin{cases} \min\limits_{x' \in [0,1]^p} & \|x - x'\| \\ \text{with} & \text{argmax}_{i=1,\ldots,c} f_i(x') \neq y. \end{cases} \tag{1}$$

Think about x' as an attack

Optimize over possible configurations

Enumerate of all possible combinations in the piecewise linear model

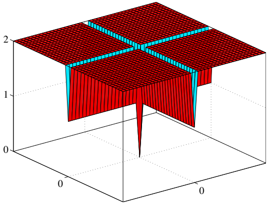
Tjeng, V., Xiao, K., and Tedrake, R. Evaluating robustness of neural networks with mixed integer programming. ICLR 2019

What's common?

Example (The counting function)

$$\begin{array}{rccl} c: & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \mathrm{w} & \longmapsto & c(\mathrm{w}) = \text{ the number of nonzero components } w_i \text{ of } \mathrm{w} \end{array}$$

It is often called the 0-norm denoted by $c(\mathrm{w}) = \|\mathrm{w}\|_0$.



Minimize a <u>nonconvex nonsmooth</u> target function or constraint

Nonconvex Nonsmooth problems in machine learning

Many lattice based problems

- variable selection, outlier detection, clustering,
- image processing, total variations,
- discrete artificial vision,
- sensor placement,
- distribution factorization,
- low rank factorization,
- NN robustness (piecewise linear optimization).

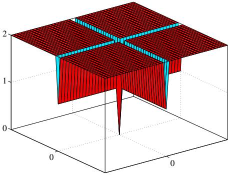
3 ways of solving combinatorial problems

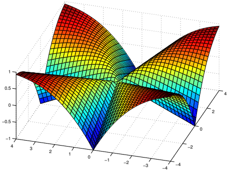
- local optimization (in general)
  - Continuous relaxations (L_1 penalty, DC...)
  - Combinatorial algorithms (greedy search, spanning tree...)
- global optimization
  - Mixed integer programming (difficult to scale to large problems)

Road map

1. Examples of combinatorial problems in machine learning
   - L_0 norm

2. MIP for variable selection AND outlier detection
   - MIP for variable selection
   - L_0 proximal algorithm
   - Implementation





Variable selection with binary variables

Definition (the least square variable selection problem)

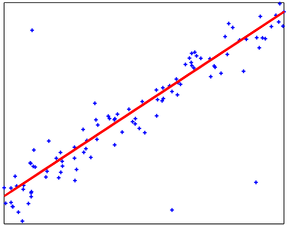
$$\begin{cases} \min\limits_{w \in \mathbb{R}^p} & \|X\mathrm{w} - \mathrm{y}\|^2 & \longleftarrow \text{fit the data} \\ \text{s.t.} & \|\mathrm{w}\|_0 \le k & \longleftarrow \text{with k variables} \end{cases}$$

- introduce p new binary variable $\mathrm{z} \in \{0,1\}^p$
- for useless variables: $z_j = 0 \Leftrightarrow w_j = 0$ $\quad\rightarrow\;$ a coupling mechanism
- $\|\mathrm{w}\|_0 = \sum_{j=1}^{p} z_j$

Definition (the LS variable selection problem with binary variables)

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p, \mathrm{z} \in \{0,1\}^p} & \|X\mathrm{w} - \mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{w}\|_0 = \sum_{i=1}^{p} z_i \le k \\ & z_j = 0 \Leftrightarrow w_j = 0, \qquad j = 1, p \end{cases}$$

Outlier detection with binary variables



Introducing outliers variables $\mathrm{o} \in \mathbb{R}^n$

$$\mathrm{y} = X\mathrm{w} + \varepsilon + \mathrm{o}, \qquad \mathrm{o}_i = \begin{cases} y_i - \mathrm{x}_i^t \mathrm{w} & \text{if } i \text{ outlier} \\ 0 & \text{else} \end{cases}$$

The least square (trimmed) regression problem with k outliers [GP02]

$$\begin{cases} \displaystyle\min_{\mathrm{w}\in\mathbb{R}^p, \mathrm{o}\in\mathbb{R}^n} & \frac{1}{2}\|X\mathrm{w} + \mathrm{o} - \mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{o}\|_0 \le k \end{cases}$$

Introduce binary variables

$$i = 1, n \qquad \begin{cases} t_i = 0 & (\mathrm{x}_i, y_i) \text{ is an outlier} & o_i \neq 0 \\ t_i = 1 & (\mathrm{x}_i, y_i) \text{ is NOT an outlier} & o_i = 0 \end{cases}$$

$$\|\mathrm{o}\|_0 = \sum_{i=1}^{n} (1 - t_i)$$

Bi robust regression

Variable selection AND outlier detection

$$\left\{ \begin{array}{cl} \min\limits_{\mathrm{w}\in\mathbb{R}^p} & \frac{1}{2}\,\|X\mathrm{w}-\mathrm{y}\|^2 \\ \text{s.t.} & \|w\|_0 \leq k_v \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{cl} \min\limits_{\mathrm{w}\in\mathbb{R}^p,\mathrm{o}\in\mathbb{R}^n} & \frac{1}{2}\,\|X\mathrm{w}+\mathrm{o}-\mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{o}\|_0 \leq k_o \end{array} \right.$$

LS regression with variable selection AND outlier detection

Given k_v the number of variable required and k_o the number of outliers

$$\left\{ \begin{array}{cl} \min\limits_{w\in\mathbb{R}^p,\mathrm{o}\in\mathbb{R}^n} & \|Xw-y-\mathrm{o}\|^2 \\ \text{s.t.} & \|w\|_0 \leq k_v \\ & \|\mathrm{o}\|_0 \leq k_o. \end{array} \right. \tag{2}$$

Bi robust regression with binary variables

- Variables
- Outliers

$$\begin{cases} \min\limits_{w,\mathrm{o}} & \|Xw - y - \mathrm{o}\|^2 \\ \text{s.t.} & \|w\|_0 \le k_v \\ & \|\mathrm{o}\|_0 \le k_o. \end{cases}$$

p binary variables $z_j \in \{0,1\}$

$$\|\mathrm{w}\|_0 = \sum_{j=1}^{p} z_j \quad \text{and} \quad z_j = 0 \Leftrightarrow w_j = 0,$$

n binary variables $t_i \in \{0,1\}$

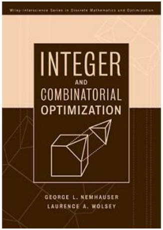
$$\|\mathrm{o}\|_0 = \sum_{i=1}^{n} (1-t_i) \text{ and } 1-t_i = 0 \Leftrightarrow o_i = 0,$$

Bi robust regression

$$\begin{cases} \min\limits_{w\in\mathbb{R}^p,\mathrm{o}\in\mathbb{R}^n, z\in\{0,1\}^p t\in\{0,1\}^n} & \|Xw - y - \mathrm{o}\|^2 \\ \text{s.t.} & \|w\|_0 = \sum z_j \le k_v \\ & z_j = 0 \Leftrightarrow w_j = 0, \qquad j = 1,p \\ & \|\mathrm{o}\|_0 = \sum t_i \le k_o \\ & 1 - t_i = 0 \Leftrightarrow o_i = 0, \qquad i = 1,n \end{cases} \tag{3}$$

So far...

- combinatorial problems can be formulated using binary variables
- we have mixed binary optimization problem
- How to solve them?
  - reformulations
  - towards stronger relaxations
  - nice initialization



Bi robust regression

Variable selection AND outlier detection

$$\left\{ \begin{array}{cl} \min\limits_{w \in \mathbb{R}^p} & \frac{1}{2} \|Xw - y\|^2 \\ \text{s.t.} & \|w\|_0 \le k_v \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{cl} \min\limits_{w \in \mathbb{R}^p, o \in \mathbb{R}^n} & \frac{1}{2} \|Xw + o - y\|^2 \\ \text{s.t.} & \|o\|_0 \le k_o \end{array} \right.$$

LS regression with variable selection AND outlier detection

Given k_v the number of variable required and k_o the number of outliers

$$\left\{ \begin{array}{cl} \min\limits_{w \in \mathbb{R}^p, o \in \mathbb{R}^n} & \|Xw - y - o\|^2 \\ \text{s.t.} & \|w\|_0 \le k_v \\ & \|o\|_0 \le k_o. \end{array} \right. \tag{4}$$

LS with fixed cardinality as a MIQP: the big M constraint

Assuming we know an upper bound M for w

$$\|\mathrm{w}\|_0 \leq k \qquad \Leftrightarrow \qquad \begin{cases} z_j \in \{0,1\}, \quad j = 1:p \\ \sum_{i=1}^{p} z_j \leq k \\ |w_j| \leq z_j M \end{cases}$$

For useless variables:

$z_j = 0 \quad \Rightarrow \quad w_j = 0$

LS with fixed cardinality as a MIQP [BKM15]

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p, \mathrm{z} \in \{0,1\}^p} & \frac{1}{2}\|X\mathrm{w} - \mathrm{y}\|_2^2 \\ \text{s.t.} & \sum_{j=1}^{p} z_j \leq k \\ \text{and} & |w_j| \leq z_j M \qquad j = 1, p \end{cases}$$

Variable selection AND outlier detection as a MILP

$$q \in \{1,2\} \qquad \begin{cases} \min\limits_{w\in\mathbb{R}^p, \mathrm{o}\in\mathbb{R}^n} & \|Xw - y - \mathrm{o}\|_q^q \\ \text{s.t.} & \|w\|_0 \le k_v \\ & \|\mathrm{o}\|_0 \le k_o. \end{cases}$$

$q = 1$

$$\begin{cases} \min\limits_{\mathrm{w}\in\mathbb{R}^p, \mathrm{o},\varepsilon^+,\varepsilon^-\in\mathbb{R}^n, \mathrm{z}\in\{0,1\}^p, \mathrm{t}\in\{0,1\}^n} & \sum\limits_{i=1}^n \varepsilon_i^+ + \varepsilon_i^- \\ \text{s.t.} & \varepsilon_i^+ - \varepsilon_i^- = \mathrm{x}_i^t\mathrm{w} + o_i - y_i \qquad i=1,n \\ & \sum\limits_{j=1}^p z_j \le k_v \\ & |w_j| \le z_j M_v \qquad j = 1,p \\ & \sum\limits_{i=1}^n (1-t_i) \le k_o \\ & |o_i| \le t_i M_o \qquad i=1,n \\ & 0 \le \varepsilon_i^+,\ 0 \le \varepsilon_i^- \qquad i = 1,n. \end{cases}$$

LSE with fixed cardinality as a MIQP with SOS constraints

Variable selection: $z_j = 0 \;\Rightarrow\; w_j = 0$ either $w_j = 0$ or $1 - z_j = 0$

Special ordered set (SOS) of type 1: at most one variable in the set can take a nonzero value,

$$w_j = 0 \text{ or } 1 - z_j = 0 \;\Leftrightarrow\; (w_j, 1 - z_j) : SOS$$

MIQP using special ordered set (SOS) of type 1

$$\begin{cases} \displaystyle\min_{\mathrm{w}\in\mathbb{R}^p,\varepsilon\in\mathbb{R}^n,z\in\{0,1\}^p} & \displaystyle\sum_{i=1}^{n} \frac{1}{2}\,(X_i^t\mathrm{w} - y_i)^2 & \longleftarrow \text{data loss} \\ \text{s.t.} & \displaystyle\sum_{j=1}^{p} z_j \le k & \longleftarrow \text{ at most } k \text{ non 0 variables} \\ & (w_j, 1 - z_j) : SOS & j = 1, p \end{cases}$$

Variable selection AND outlier detection as a MIQP

$$q \in \{1, 2\} \qquad \begin{cases} \min\limits_{w \in \mathbb{R}^p, \mathrm{o} \in \mathbb{R}^n} & \|Xw - y - \mathrm{o}\|_q^q \\ \text{s.t.} & \|w\|_0 \le k_v \\ & \|\mathrm{o}\|_0 \le k_o. \end{cases}$$

$q = 2$

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p, \mathrm{o} \in \mathbb{R}^n, \mathrm{z} \in \{0,1\}^p, \mathrm{t} \in \{0,1\}^n} & (y - Xw - \mathrm{o})^t (y - Xw - \mathrm{o}) \\ \text{s.t.} & \sum_{j=1}^{p} z_j = k_v \\ & \sum_{i=1}^{n} t_i \le k_o \\ & (w_j, 1 - z_j) : SOS & j = 1, p \\ & (\mathrm{o}_i, 1 - t_i) : SOS & i = 1, n. \end{cases}$$

Balls and Triks: the convex hull of the feasible set

$$Conv\left(\left\{\mathrm{w} \mid |w_j| \leq z_j M \text{ and } \sum_{j=1}^{p} z_j \leq k\right\}\right) = \left\{\mathrm{w} \mid \|w\|_\infty \leq M \ \text{ and } \ \|\mathrm{w}\|_1 \leq kM\right\}$$

MIQP: a more structured representation [BKM15]

$$\left\{\begin{array}{lll}
\displaystyle\min_{\mathrm{w}\in\mathbb{R}^p,\varepsilon\in\mathbb{R}^n,\mathrm{z}\in\{0,1\}^p} & \displaystyle\sum_{i=1}^{n} \tfrac{1}{2}\left(X_i^t\mathrm{w} - y_i\right)^2 & \longleftarrow \text{data loss} \\
\text{s.t.} & \displaystyle\sum_{j=1}^{p} z_j \leq k & \longleftarrow \ \text{at most } k \text{ non 0 variables} \\
 & (w_j, 1 - z_j) : SOS & \quad j = 1, p \\
 & |w_j| \leq M_\infty & \quad j = 1, p \\
 & \displaystyle\sum_{j=1}^{p} |w_j| \leq M_1 &
\end{array}\right.$$

[BKM15] claim: *Adding these bounds typically leads to improved performance of the MIO, especially in delivering lower bound certificates*

Balls and Triks: the convex hull of the feasible set

$$\mathcal{S} = \Big\{ \mathrm{w}, \mathrm{o} \mid \sum_{j=1}^{p} z_j \leq k_v,\ |w_j| \leq z_j M_v,\ \sum_{i=1}^{n} (1 - t_i) \leq k_o,\ |\tau_i| \leq t_i M_o \Big\},$$

$$Conv(\mathcal{S}) = \Big\{ \mathrm{w}, \mathrm{o} \mid ||\mathrm{w}||_\infty \leq M_v\ ||\mathrm{o}||_\infty \leq M_o\ ||\mathrm{w}||_1 \leq k_v M_v,\ ||\mathrm{o}||_1 \leq k_o M_o \Big\}$$

$$\begin{cases} \displaystyle\min_{\mathrm{w},\mathrm{o},\mathrm{z}\in\{0,1\}^p,\mathrm{t}\in\{0,1\}^n} & \frac{1}{q}\|X\mathrm{w} + \mathrm{o} - \mathrm{y}\|_q^q & & \\ \text{s.t.} & \displaystyle\sum_{j=1}^{p} z_j \leq k_v, & (w_j, 1 - z_j) : SOS & j = 1, p \\ & \displaystyle\sum_{i=1}^{n} (1 - t_i) \leq k_o, & (\tau_i, 1 - t_i) : SOS & i = 1, n \\ & \|\mathrm{w}\|_1 \leq k_v M_v,\ \ \|\mathrm{w}\|_\infty \leq M_v & & \\ & \|\mathrm{o}\|_1 \leq k_o M_o,\ \ \|\mathrm{o}\|_\infty \leq M_o, & & \end{cases} \tag{5}$$

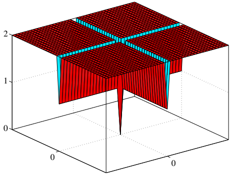
with problem-dependent constants M_v and M_o.

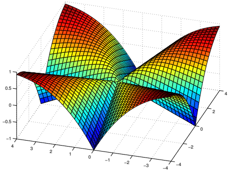
So far...

- birobust regression as a MIP
  - for variable selection AND outlier detection in regression
  - and in quantile regression, SVM, logistic regression
  - reformulation (practical matter)
- efficient software for moderate size problem
- for large size: use first order algorithms

Road map

1. Examples of combinatorial problems in machine learning
   - L_0 norm
2. MIP for variable selection AND outlier detection
   - MIP for variable selection
   - L_0 proximal algorithm
   - Implementation





Variable selection: a specific case with a closed-form solution

Definition (the least square variable selection problem with $X = Id$)

given $k < p$

$$\begin{cases} \min\limits_{\mathrm{u} \in \mathbb{R}^p} & \|\mathrm{u} - \mathrm{w}\|^2 & \longleftarrow \text{fit the data} \\ \text{s.t.} & \|\mathrm{u}\|_0 \leq k & \longleftarrow \text{with k variables} \end{cases}$$

sort $|\mathrm{w}|$: $|w_{(1)}| \geq |w_{(2)}| \geq \dots |w_{(j)}| \geq \dots |w_{(p)}|$

Closed-form solution: the hard thresholding operator

$$u_i^\star = H_k(\mathrm{w}) = \begin{cases} w_j & \text{if } j \in \{(1), \dots, (k)\} \\ 0 & \text{else} \end{cases}$$

Proximity operator

Definition (Proximity operator [Mor62])

The Proximity operator of a function h is:

$$\begin{array}{rccl} \mathbf{prox}_h : & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \mathrm{w} & \longmapsto & \mathbf{prox}_h(\mathrm{w}) = \underset{\mathrm{u}\in\mathbb{R}^p}{\arg\min}\ h(\mathrm{u}) + \frac{1}{2}\|\mathrm{u}-\mathrm{w}\|^2 \end{array}$$

Example

$$\begin{array}{lll} h(\mathrm{w}) = 0 & \mathbf{prox}_h(\mathrm{w}) = \mathrm{w} & \\ h(\mathrm{w}) = \rho\,\mathbf{pen}_\lambda(\mathrm{w}) & \mathbf{prox}_h(\mathrm{w}) = \mathbf{shr}_{\rho\lambda}(\mathrm{w}) & \text{shrinkage} \\ h(\mathrm{w}) = \mathbb{I}_C(\mathrm{w}) & \mathbf{prox}_h(\mathrm{w}) = \underset{\mathrm{u}\in C}{\arg\min}\ \frac{1}{2}\|\mathrm{u}-\mathrm{w}\|^2 & \text{projection} \end{array}$$

The proximity operator as a projection

$$\mathbf{prox}_{\mathbb{I}_{\{\|\mathrm{w}\|_0\le k\}}}(\mathrm{w}) = \underset{\|\mathrm{u}\|_0\le k}{\arg\min}\ \frac{1}{2}\|\mathrm{u}-\mathrm{w}\|^2 = H_k(\mathrm{w}) = \begin{cases} w_i & \text{if } i \in \{(1),\dots,(k)\} \\ 0 & \text{else} \end{cases}$$

The projected gradient (L_0 projection or proximal)

for solving $\begin{cases} \min_{w \in \mathbb{R}^p} & \frac{1}{2}\|X\mathrm{w} - \mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{w}\|_0 \leq k_v \end{cases}$

Algorithm 1 L_0 gradient projection algorithm [BD09]

Data: $X, \mathrm{y}, \mathrm{w}$ initialization
Result: w
while *not converged* **do**
- $g \leftarrow \nabla g(\mathrm{w}) = X^\top(X\mathrm{w} - y)$, — the gradient
- $\rho \leftarrow$choose a stepsize
- $d \leftarrow \mathrm{w} - \rho g$, — forward (explicit)
- $\mathrm{w} \leftarrow H_k(d)$, — the projection–proximal step, backward (implicit)

end

if $\varepsilon \leq \rho \leq \frac{1}{\|X^\top X\|}$, it converges towards a local minimum [ABS13] since its objective function satisfies the Kurdyka-Lojasiewicz inequality.

Proximal alternating linearized minimization (PALM)

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p, \mathrm{o} \in \mathbb{R}^n} & \frac{1}{2}\|X\mathrm{w} + \mathrm{o} - \mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{w}\|_0 \le k_v \\ & \|\mathrm{o}\|_0 \le k_o \end{cases}$$

given o

$$\begin{cases} \min\limits_{w \in \mathbb{R}^p} & \frac{1}{2}\|X\mathrm{w} + \mathrm{o} - \mathrm{y}\|^2 \\ \text{s.t.} & \|\mathrm{w}\|_0 \le k_v \end{cases}$$

given w

$$\begin{cases} \min\limits_{\mathrm{o} \in \mathbb{R}^n} & \frac{1}{2}\|\mathrm{o} - (\mathrm{y} - X\mathrm{w})\|^2 \\ \text{s.t.} & \|\mathrm{o}\|_0 \le k_o \end{cases}$$

Algorithm 2 Proximal alternating linearized minimization (PALM) [BST14]

Data: X, y initialization $\mathrm{w}, \mathrm{o} = 0$

Result: w, o

while *not converged* **do**

$d \leftarrow \mathrm{w} - \rho_v X^\top (X\mathrm{w} + \mathrm{o} - \mathrm{y})$, variable selection

$\mathrm{w} \leftarrow H_{k_v}(d)$,

$\delta \leftarrow \mathrm{o} - \rho_o (X\mathrm{w} + \mathrm{o} - y)$, eliminating outliers

$\mathrm{o} \leftarrow H_{k_o}(\delta)$,

end

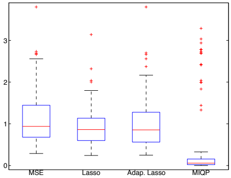
Prox summary

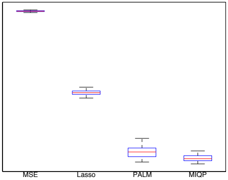
- PALM is fast and scalable
- convergence profs towards a local minimum
- improvement:
  - accelerations: FISTA and others
  - Newton proximal
  - more improvement with randomization

Road map

1. Examples of combinatorial problems in machine learning
   - L_0 norm

2. MIP for variable selection AND outlier detection
   - MIP for variable selection
   - L_0 proximal algorithm
   - Implementation





Combine the best of the two worlds

Combine the best of the two worlds [BKM15]

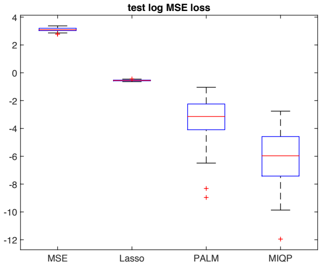
1. $(\mathrm{w}, \mathrm{o}) \leftarrow$ PALM alternating proximal gradient method
2. use w and o as a warm start for MIP (with Cplex)
   - $(\mathrm{w}, \mathrm{o}) \leftarrow$ Polish coefficients on the active set
   - initialize the constants M_v, M_o

Experimental setup

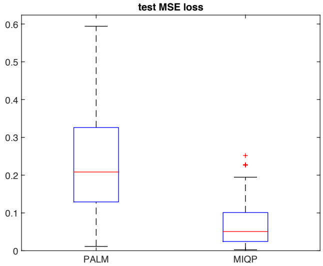
- My mac
- Matlab
- Cplex 12.6.1 (`cplexmiqp`)
- time out = 5 min

Variable selection AND outlier detection on a toy dataset

- $\mathrm{y} = X\mathrm{w} + \mathrm{o} + \varepsilon$
- $n = 300$ observations with $p = 25$ variables
- linear model with ε a centered Gaussian noise with SNR ≈ 1
- $k_o = 50$ outliers and $k_v = 5$ non zeros variables
- 100 repetitions



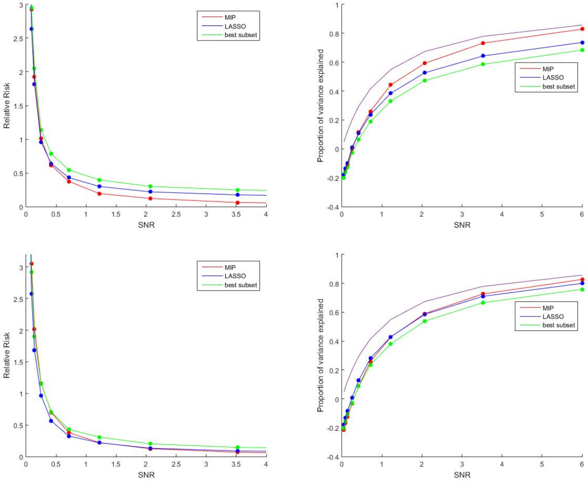
(log) performances



zoom

Best Subset, Forward Stepwise, or Lasso . . . or DR MIP

n=500, p=100, $k_v = 5$ and $k_o = 1\%$ of outliers



- $w^\star = (1, \ldots, 1, 0, \ldots, 0)^\top$
- $w^\star = (-2, 0, 0, 0.8, 3.22, 0, 0, 1.8, 0, -0.95, 0, \ldots, 0)^\top$

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692.

Best Subset, Forward Stepwise, or Lasso . . . or DR MIP

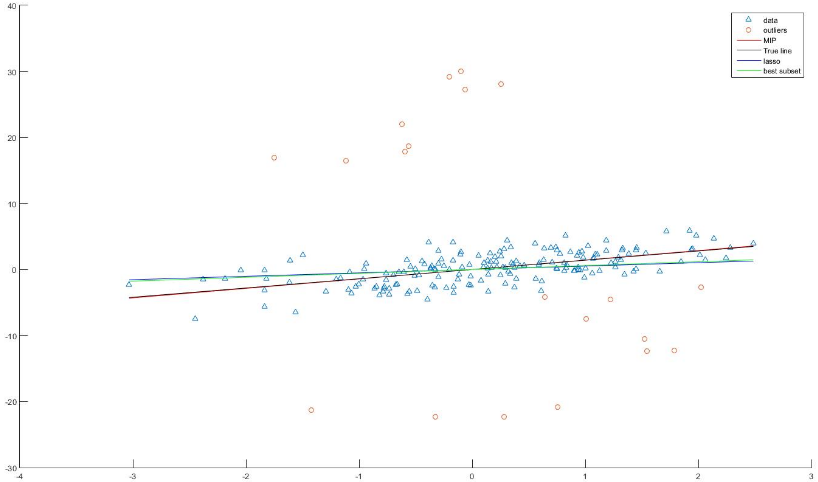
Dataset	# of instances	# of attributes	Origin
Boston housing	489	3	Github
Body fat	252	15	lib.stat.cmu.edu/
Forest fires	512	12	UCI
Facebook metrics	500	19	UCI
Real estate evaluation	414	7	UCI
Concrete slump test	103	10	UCI
Auto mpg	398	8	UCI
Diabetes	442	10	stat.ncsu.edu
Concrete compressive strength	1030	9	UCI

	Best subset	Lasso	MIP	PALM	k_o %	k_v^B	k_v^L	k_v^M
Boston housing	0.288 (0.08)	0.294 (0.09)	0.285 (0.09)	**0.284 (0.09)**	5	3	3	3
Body fat	0.006 (0.01)	0.006 (0.01)	0.006 (0.01)	**0.006 (0.01)**	12.5	1	3	1
Forest fires	**0.992 (1.22)**	0.998 (1.24)	1.012 (1.26)	0.994 (1.23)	45	2	1	3
Facebook metrics	**0 (ϵ)**	9.e-5 (9.e-5)	1.e-4 (2.e-4)	4.e-4 (3.e-4)	2.5	3	3	4
Real estate	0.434 (0.19)	**0.430 (0.19)**	0.446 (0.17)	0.439 (0.17)	15	5	6	6
Concrete slump	0.129 (0.02)	**0.122 (0.02)**	0.158 (0.05)	0.169 (0.04)	22.5	7	7	5
Auto mpg	0.186 (0.03)	**0.186 (0.03)**	0.201 (0.03)	0.200 (0.04)	7.5	6	6	5
Diabetes	0.514 (0.08)	**0.504 (0.08)**	0.551 (0.07)	0.534 (0.08)	40	7	7	6
Concrete copres.	0.392 (0.06)	**0.392 (0.05)**	0.495 (0.12)	0.467 (0.12)	22.5	8	7	8

with 5% outliers

	Best subset	Lasso	MIP	PALM	k_o %	k_v^B	k_v^L	k_v^M
Boston housing	0.312 (0.06)	0.302 (0.03)	0.301 (0.06)	**0.290 (0.06)**	32.5	3	3	3
Body fat	0.016 (0.01)	0.031 (0.03)	**0.005 (0.01)**	0.006 (0.01)	20	1	3	3
Forest fires	1.306 (1.37)	**1.005 (1.25)**	1.186 (1.27)	1.754 (1.24)	27.5	5	3	5
Facebook metrics	0.629 (0.70)	0.532 (0.52)	**0.139 (0.16)**	0.399 (0.81)	27.5	4	5	2
Real estate	0.475 (0.16)	0.462 (0.17)	**0.445 (0.16)**	0.445 (0.16)	15	4	6	5
Concrete slump	0.244 (0.05)	0.266 (0.09)	**0.145 (0.05)**	0.145 (0.07)	17.5	5	7	6
Auto mpg	0.202 (0.04)	0.218 (0.04)	**0.196 (0.04)**	0.195 (0.04)	17.5	4	5	5
Diabetes	0.535 (0.08)	**0.524 (0.09)**	0.555 (0.07)	0.553 (0.09)	25	6	8	7
Concrete copres.	0.404 (0.05)	**0.403 (0.06)**	0.412 (0.05)	0.411 (0.05)	12.5	7	7	8

Miss leading test error: 1d illustration

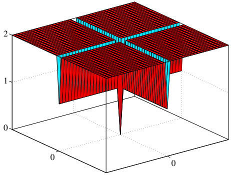


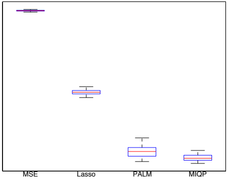
testerror with outliers (Lasso) $<$ testerror with outliers (MIP)

Road map (done)

1. Examples of combinatorial problems in machine learning

2. MIP for variable selection AND outlier detection





Conclusion

- Machine learning with MIP

pros	cons
it works	it does not scale
global optimum	only linear or quadratic
flexible	show some instability
that is what we want to do	it's not what we want to do

- Future work
  - efficient generic solver
  - efficient implementation: parallelization, randomisation, GPU
  - efficient hyper parameter calibration

[ABS13] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[BBF$^+$16] Pietro Belotti, Pierre Bonami, Matteo Fischetti, Andrea Lodi, Michele Monaci, Amaya Nogales-Gómez, and Domenico Salvagnin. On handling indicator constraints in mixed integer programming. *Computational Optimization and Applications*, 65(3):545–566, 2016.

[BD09] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[BKM15] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *arXiv preprint arXiv:1507.03133*, 2015.

[BST14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[GP02] A Giloni and M Padberg. Least trimmed squares regression, least median squares regression, and mathematical programming. *Mathematical and Computer Modelling*, 35(9):1043–1060, 2002.

[GW89] Martin Grötschel and Yoshiko Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming*, 45(1-3):59–96, 1989.

[LCHR19] Yuan Liu, Stephane Canu, Paul Honeine, and Su Ruan. Mixed integer programming for sparse coding: Application to image denoising. *IEEE Transactions on Computational Imaging*, 2019.

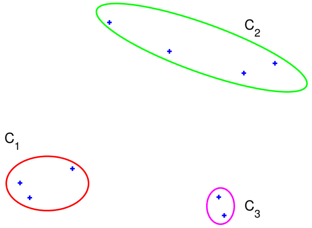
[Mor62] Jean-Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899, 1962.

[SRC17] Ruobing Shen, Gerhard Reinelt, and Stéphane Canu. A first derivative potts model for segmentation and denoising using ILP. In *Operations Research Proceedings 2017, Selected Papers of the Annual International Conference of the German Operations Research Society (GOR), Freie Universiät Berlin, Germany, September 6-8, 2017.*, pages 53–59, 2017.

[SSST06] Burcu Sağlam, F Sibel Salman, Serpil Sayın, and Metin Türkay. A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *European Journal of Operational Research*, 173(3):866–879, 2006.

Clustering as a MIP: the binary variables

$$w_{ij} = \begin{cases} 1 & \text{if } \mathrm{x}_i, \mathrm{x}_j \text{ in the same cluster} \\ 0 & \text{else.} \end{cases}$$



$$z_{ik} = \begin{cases} 1 & \text{if } \mathrm{x}_i \ \text{ in cluster } k \\ 0 & \text{else.} \end{cases}$$

$$W \in \{0,1\}^{n^2} \qquad\qquad Z \in \{0,1\}^{n \times q}$$

W and Z are connected since $W = ZZ^\top$.

Clustering as a MIP

Grötschell-Wakabayashi formulation [GW89]

$$\begin{cases} \displaystyle\min_{W\in\{0,1\}^{n^2}} & \displaystyle\sum_i^n \sum_j^n w_{ij}\|\mathrm{x}_i - \mathrm{x}_j\|^2 \\ \text{with} & w_{ij} + w_{jk} - w_{ik} \leq 1 \qquad i = 1, n, j = 1, n, k = 1, n. \end{cases}$$

[SSST06].

$$\begin{cases} \displaystyle\min_{R_k, Z\in\{0,1\}^{n\times q}} & \max(R_1, \dots, R_k, \dots, R_q) \\ \text{with} & (z_{ik} + z_{jk} - 1)\|\mathrm{x}_i - \mathrm{x}_j\|^2 \leq R_k, \quad i, j = 1, \dots, n;\ k = 1, \dots, \\ \text{and} & \displaystyle\sum_{k=1}^q z_{ik} = 1 \qquad i = 1, \dots, n \\ & \displaystyle\sum_{i=1}^n z_{ik} \geq 1 \qquad k = 1, \dots, q. \end{cases}$$

[LCHR19] [SRC17]