Complementarity Formulations for $L^0$-norm Optimization Problems

John E. Mitchell

Department of Mathematical Sciences
RPI, Troy, NY 12180 USA

SIAM Conference on Optimization
San Diego, May 2014
1. Introduction and Applications
2. L0-norm minimization
3. Conclusions
Outline

1. Introduction and Applications

2. L0-norm minimization

3. Conclusions
L0-norm minimization

Want to solve:

\[
\min_x \{ f(x) + \gamma \|x\|_0 : Ax \geq b, \ x \in \mathbb{R}^n \}
\]

where \( A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m, \) and \( \|x\|_0 = \text{card}\{x_i : x_i \neq 0\}. \)

The parameter \( \gamma \) reflects the relative importance of sparsity and \( f(x). \)

\( L_0 \)-norm minimization is of interest in feature selection, compressed sensing, misclassification minimization, and almost any problem where a sparse solution is desired.

Can be modeled as a **Mathematical Program with Complementarity Constraints or MPCC**.
Eg: Misclassification minimization

Given \((C, \gamma, \varepsilon) > 0\), and points \(x^i\) with observed values \(y_i\):

\[
\begin{align*}
\text{minimize} & \quad C \sum_{i=1}^{n} \max \left\{ |w^T x^i + b - y_i| - \varepsilon, 0 \right\} + \frac{1}{2} w^T w \\
\text{standard SVM objective} & \\
+ & \quad \gamma \sum_{i=1}^{n} \left\{ egin{array}{ll}
1 & \text{if } |w^T x^i + b - y_i| > \varepsilon \\
0 & \text{if } |w^T x^i + b - y_i| \leq \varepsilon
\end{array} \right\} \\
\text{number of misclassified points} & 
\end{align*}
\]

(Mangasarian)
Eg: Minimum portfolio revision

Given portfolio $x^0$ in standard unit simplex $\Delta_n \triangleq \{ x \in \mathbb{R}_+^n \mid 1^T x = 1 \}$, consider a portfolio revision problem as follows:

$$\min_{x \in \Delta_n} \frac{C}{2} x^T V x + c' \text{VaR}_\beta(x) + K \| x - x^0 \|_0$$

subject to $\mu^T x \geq R$ (expected portfolio return),

where $\text{VaR}_\beta(x)$ is the $\beta$-value-at-risk of the portfolio $x$ with $\beta \in (0, 1)$. 
Outline

1. Introduction and Applications
2. L0-norm minimization
3. Conclusions
**Complementarity formulation**

Want to solve:

$$\min_x \{ f(x) + \gamma \| x \|_0 : Ax \geq b, \ x \in \mathbb{R}^n \}$$

Equivalently:

$$\min_{x^\pm, \xi} f(x) + \sum_{i=1}^n (1 - \xi_i)$$

subject to

$$Ax \geq b$$

$$0 \leq \xi \leq 1$$

$$0 \leq \xi \perp x$$

Note that if $x_i = 0$ then we can choose $\xi_i = 1$, and if $x_i \neq 0$ then we must set $\xi_i = 0$.

Thus, the objective counts the number of nonzero components.

No need for any big-$M$ terms in this LPCC formulation.
Complementarity formulation

\[ \min_{x^\pm, \xi} \quad f(x) + \sum_{i=1}^{n} (1 - \xi_i) \]
subject to \[ Ax \geq b \]
\[ 0 \leq \xi \leq 1 \]
\[ 0 \leq \xi \perp x \]

This is a half-complementary formulation;
Complementarity formulation

$$\min_{x^\pm, \xi} f(x) + \sum_{i=1}^{n} (1 - \xi_i)$$
subject to
$$Ax \geq b$$
$$0 \leq \xi \leq 1$$
$$0 \leq \xi \perp x^+ + x^- \geq 0, \ 0 \leq x^+ \perp x^- \geq 0, \ x = x^+ - x^-$$

This is a half-complementary formulation; can also get a full-complementary formulation by splitting $x = x^+ - x^-$. 
KKT points

The Constant Rank Constraint Qualification holds. It also holds under additional assumptions with nonlinear constraints.

Let $x$ satisfy $Ax \geq b$ and assume $x$ minimizes $f(x)$ for some assignment of the complementarities, set $\xi$ to count the number of nonzero components:

This point is a local minimizer and a strongly stationary point of the MPCC formulation.
Relaxed NLP formulation

Let $\epsilon > 0$. Assume $f(x) \equiv 0$. Can get a relaxed formulation:

$$\min_{x^+, x^-} \sum_{i=1}^{n} (1 - \xi_i)$$

subject to $Ax^+ - Ax^- \geq b$

$0 \leq x^+, x^-$

$0 \leq \xi \leq 1$

$\xi_i (x_i^+ + x_i^-) \leq \epsilon \quad i = 1, \ldots, n$

CQ holds for this problem.

Not every $x$ is a local minimizer.

A sequence of global minimizers as $\epsilon \to 0$ will have a subsequence that converges to a global minimizer of the L0-norm problem.

Under certain conditions, the solution to the relaxed formulation is a solution to a reweighted version of the L1-norm minimization problem.
Penalty method

Can penalize violations of complementarity in many ways:

- Add \((x^+ + x^-)^T \xi\) to the objective function.
- Add \((x^+ + x^-)^T \xi + (x^+)^T x^-\) to the objective function. Corresponds to the AMPL keyword “complements” when using knitro.
- Add \(((x^+ + x^-)^T \xi)^2\) to the objective function.
- Add \(((x^+ + x^-)^T \xi + (x^+)^T x^-)^2\) to the objective function.
- Use the Fischer-Burmeister function, or some other NCP function.

Theoretical investigation is ongoing.

Surprisingly, the fourth formulation has worked well in our tests.
Penalty method

Can penalize violations of complementarity in many ways:

- Add \((x^+ + x^-)^T \xi\) to the objective function.
- Add \((x^+ + x^-)^T \xi + (x^+)^T x^-\) to the objective function.
  
  Corresponds to the AMPL keyword “complements” when using knitro.
- Add \(((x^+ + x^-)^T \xi)^2\) to the objective function.
- Add \(((x^+ + x^-)^T \xi + (x^+)^T x^-)^2\) to the objective function.
- Use the Fischer-Burmeister function, or some other NCP function.

Theoretical investigation is ongoing.

Surprisingly, the fourth formulation has worked well in our tests.
Test problems

Examine three classes of problems:

- Minimize L0-norm, without another objective \( f(x) \).
- Look at weighted combinations of \( f(x) \) and L0-norm.
- Look at signal recovery problems.

In each case, compare solutions obtained by a nonlinear programming package (KNITRO, SNOPT, CONOPT, MINOS) for our complementarity formulation with an L1-norm formulation.

Have no guarantee of global optimality.
Minimize number of non zeroes

\[
\min_{x \in \mathbb{R}^n} \|x\|_{0,1}
\]

s.t.
\[
Ax \geq b.
\]

Look at four formulations:

**MILP:** integer programming formulation with a big-M

**L1:** \(L1\)-approximation solved as an LP. Objective reweighted iteratively (Candes et al).

**fullcomp:** complementarity formulation solved as NLP

**halfcomp:** modified complementarity formulation, solved as NLP

Test problems

\(A\) and \(b\) generated randomly, each entry in \(U[-1, 1]\).
A is $30 \times 50$:
50 test problems, various solvers and start points
A is $300 \times 500$: performance profile of sparsity
Weighted combinations

$$\min_{x \in \mathbb{R}^n} \quad q(x) + \beta \|x\|_{0,1}$$
$$\text{s.t.} \quad Ax \geq b.$$ 

Solve for each norm for various different choices of $\beta$.

In our tests, $A$ is $30 \times 50$, $q(x)$ is strictly convex quadratic function. The L0-norm formulations are solved using KNITRO under AMPL.
Pareto frontier

![Graph showing Pareto frontier with various markers for L1, L0, MIQP, and L1-Debiased solutions.](image-url)
Signal recovery

\[
\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \beta \|x\|_{0,1}
\]

Look at different choices of regularizing parameter \(\beta\).

\(A\) is \(256 \times 1024\).
In original signal, 40 entries of \(x\) are nonzero.
Random noise is added to give \(b\).

L1-norm results are improved by debiasing:
fix zeroes in L1 solution at zero, then look for best least squares fit.
L1-debiased leads to many small nonzero components in our tests.
Signal recovery sparsity

![Signal recovery sparsity graph](image-url)

- L0-norm minimization
- Signal recovery

**Regularizing Parameter**

- Number of NZ in optimal solution

**Graph Details**

- L0
- L1
- L1-Debiased

**Axes**

- Number of NZ in optimal solution
- Regularizing Parameter

**Legend**

- y=40
Outline

1. Introduction and Applications
2. L0-norm minimization
3. Conclusions
Conclusions

A complementarity formulation can be very effective at finding sparse solutions to $L_0$-norm optimization problems, far more quickly than an MILP approach but more slowly than an $L_1$-norm LP approach.

In the first class, outperformed an $L_1$-norm formulation even with sophisticated iterative schemes for the $L_1$ approach.

For Pareto frontier, get better results than $L_1$ for small instance. So far, these results haven’t yet extended well to larger instances.

For signal recovery, able to recover signal effectively. Currently investigating different distributions and test instances.

At present, we only have a very partial understanding as to why the $L_0$-norm approach works so well. Theoretical investigation is ongoing!