Lass0: sparse non-convex regression by local search

William Herlands **Maria De-Arteaga** **Daniel Neill** **Artur Dubrawski**
Carnegie Mellon University
herlands@cmu.edu, mdeartea@andrew.cmu.edu, neill@cs.cmu.edu, awd@cs.cmu.edu

Abstract

We compute approximate solutions to L_0 regularized linear regression using L_1 regularization, also known as the Lasso, as an initialization step. Our algorithm, the Lass0 ("Lass-zero"), uses a computationally efficient stepwise search to determine a locally optimal L_0 solution given any L_1 regularization solution. We present theoretical results of consistency under orthogonality and appropriate handling of redundant features. Empirically, we use synthetic data to demonstrate that Lass0 solutions are closer to the true sparse support than L_1 regularization models. Additionally, in real-world data Lass0 finds more parsimonious solutions than L_1 regularization while maintaining similar predictive accuracy.

1 Introduction

Sparse approximate solutions to linear systems are desirable for providing interpretable results that succinctly identify important features. For $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$, L_0 regularization (Eq. 1[1]), called "best subset selection," is a natural way to achieve sparsity by directly penalizing non-zero elements of β. This intuition is fortified by theoretical justification. Foster and George [1] demonstrate that for general predictor matrix X, L_0 regularization achieves the asymptotic minimax rate of risk inflation. Unfortunately, it is well known that L_0 regularization is non-convex and NP hard [2].

$$\min_{\beta \in R^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \quad (1) \qquad\qquad \min_{\beta \in R^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \quad (2)$$

Despite the computational difficulty, the optimality of L_0 regularization has motivated approximation methods such as Zhang [3], who provide a Forward-Backward greedy algorithm with asymptotic guarantees. Additionally, integer programming has been used to find solutions for problems of bounded size [4, 5, 6].

Instead of L_0 regularization, it is common to use L_1 regularization (Eq. 2), known as the Lasso [7]. This convex relaxation of L_0 regularization achieves sparse solutions which are model consistent and unique under regularity conditions, which, among other things, limit the correlations between columns of X [8, 9]. Additionally, L_1 is a reasonable substitute for L_0 regularization because the L_1 norm is the best convex approximation to the L_0 norm [10]. However, on real-world data sets, L_1 regularization tends to select incorrect models since the L_1 norm shrinks all coefficients including those which are in the active set [11, 12]. This bias can be particularly troublesome in very sparse settings, where the predictive risk of L_1 can be arbitrarily worse than that of L_0 [13].

In order to take advantage of the computational tractability of L_1 regularization, and the optimality of L_0, we develop the Lass0 ("Lass-zero"), a method which uses an L_1 solution as an initialization step to find a locally optimal L_0 solution. At each computationally efficient step, the Lass0 improves upon the L_0 objective, often finding sparser solutions without sacrificing prediction accuracy.

[1] We consider the Lagrange form of subset selection. Since the problem is nonconvex this is weaker than the constrained form, meaning that all solutions of the Lagrange problem are solutions to a constrained problem.

Previous literature, such as SparseNet [12], also explored the relationship between L_1 and L_0 solutions. Yet unlike our approach, SparseNet reparameterizes the problem with MC+ loss and solves a generalized soft thresholding problem at each iteration requiring a large number of problems to solve to reach L_0. Alternatively, Lin et al. [14] use the L_0 objective as a criterion to select among different L_1 models from the LARS [15] solution set. However, they do not improve upon the L_1 results by optimizing L_0 directly, as in our work.

In the remainder of this paper, Section 2 details the Lass0 algorithm. Section 3 provides theoretical guarantees for convergence in the orthogonal case and elimination of redundant variables in the general case. Section 4 presents empirical results on synthetic and real world data. Finally, we conclude in Section 5 with directions for future work in the general context of non-convex optimization.

2 Lass0

We propose a new method for finding sparse approximate solutions to linear problems, which we call the Lass0. The full pseudocode of the Lass0 algorithm is presented in Algorithm 1, and we refer to the lines through this section. The method is initialized by a solution to L_1 regularization, β^{L_1}, given a particular λ. The Lass0 then uses an iterative algorithm to find a locally optimal solution that minimizes the objective function of the L_0 regularization (Eq. 3).

$$L_0(\beta, y, X, \lambda) = \|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \tag{3}$$

If $supp()$ indicates the support, the first step in the Lass0 is to compute $\beta = O\hat{L}S(supp(\beta^{L_1}), y, X)$, the ordinary least squares solution constrained such that every zero entry of β^{L_1} must remain zero. $O\hat{L}S()$ is formally defined as,

$$O\hat{L}S(F, y, X) = \min_{\beta} \|y - X\beta\|_2^2 \text{ s.t. } \beta_i = 0 \; \forall i \notin F \tag{4}$$

For each entry, β_i, of the resulting vector, we compute the effect of individually adding or removing it from $supp(\beta)$ in Lines 6 and 7. Note that by adding an entry to the support, we increase the penalty by λ, but potentially create a better estimate for y, resulting in a lower $\|y - X\beta\|_2^2$ loss term. Similarly, the opposite may be true when removing an entry from the support set.

This procedure yields a new solution vector $\beta^{(i)}$ for each i. The $\beta^{(i)}$ which minimizes the L_0 objective function is selected as β' in Line 8. Then, in Line 9, we accept β' only if it is strictly better than the solution we began with, β. The iterative algorithm terminates whenever there is no improvement.

Algorithm 1 Lass0

1: Input: L_1 solution, β^{L_1}
2: $F = supp(\beta^{L_1})$
3: $\beta = O\hat{L}S(F)$
4: **while** True **do**
5:     $F = supp(\beta_i)$
6:     **For all** $i \in F$ **do:** $\beta^{(i)} = O\hat{L}S(F \setminus \{i\}, y, X)$
7:     **For all** $i \notin F$ **do:** $\beta^{(i)} = O\hat{L}S(F \cup \{i\}, y, X)$
8:     $\beta' = \arg\min_i L_0(\beta^{(i)}, y, X, \lambda)$
9:     **if** $L_0(\beta', y, X, \lambda) < L_0(\beta, y, X, \lambda)$ **then**
10:         $\beta = \beta'$
11:     **else**Break
12:     **end if**
13: **end while**

This procedure is equivalent to greedy coordinate minimization where we warm-start the optimization procedure with the L_1 regularization solution. Additionally, we note that any L_p regularization with norm $p < 1$ is non-convex. While the present work focuses on L_0 regularization, the Lass0 can be applied to approximate solutions to any other non-convex L_p regularization with minimal changes.

3 Theoretical properties

Theorem 1. *Assuming that X is orthogonal, the Lass0 solution is the L_0 regularization solution.*

Proof. Recall that Lass0 is initialized with the L_1 regularization solution. With an orthogonal set of covariates, it is well known that the solution to L_1 regularization, β^{L_1}, is soft-thresholding of the components of $X^T y$ at level λ (Eq. 5). Additionally, it is well known that in this case the solution to L_0 regularization, β^{L_0}, is hard-thresholding of the components of $X^t y$ at level $\sqrt{2\lambda}$ (Eq. 6).

$$\beta_j^{L_1} = \begin{cases} X_j^T y - \lambda & \text{if} \quad X_j^T y > \lambda \\ 0 & \text{if} \quad |X_j^T y| < \lambda \\ X_j^T y + \lambda & \text{if} \quad X_j^T y < -\lambda \end{cases} \quad (5) \qquad \beta_j^{L_0} = \begin{cases} X_j^T y & \text{if} \quad X_j^T y > \sqrt{2\lambda} \\ 0 & \text{if} \quad |X_j^T y| < \sqrt{2\lambda} \\ X_j^T y & \text{if} \quad X_j^T y < -\sqrt{2\lambda} \end{cases} \quad (6)$$

We will prove that the Lass0 solution, $\beta^{Lass0} = \beta^{L_0}$. Since the solutions to L_1 and L_0 regularization depend on λ, the proof is divided in three cases to cover all possible values of λ, and we use the same regularization parameter λ for both algorithms.

i. **Case $\boldsymbol{\lambda = 2}$:** Since $\sqrt{2\lambda} = \lambda$, therefore $supp(\beta^{L_0}) = supp(\beta^{L_1})$. Note that in the orthogonal case, the least squares solution is $(X^T X)^{-1} X^t y = X^t y$. In the first step of the Lass0 algorithm we find $\hat{OLS}(supp(\beta^{L_1}), y, X)$ which corresponds to setting $\beta_k = (X^T y)_k \; y \; \forall k \in supp(\beta^{L_1})$, and $\beta_k = 0$ otherwise. Therefore, in the first step the algorithm will reach the hard-thresholding at level $\sqrt{2\lambda}$ and terminate.

ii. **Case $\boldsymbol{\lambda > 2}$:** Since $\sqrt{2\lambda} < \lambda$, then $supp(\beta^{L_0}) \supseteq supp(\beta^{L_1})$. Let $\beta = \hat{OLS}(supp(\beta^{L_1}), y, X)$ and let $\beta^{new} = \hat{OLS}(supp(\beta) \setminus \{k\}, y, X)$. The Lass0 will only choose β^{new} and remove element k from $supp(\beta^{L_1})$ if,

$$\frac{1}{2}\|y - X\beta^{new}\|_2^2 - \frac{1}{2}\|y - X\beta\|_2^2 < \lambda \quad (7)$$

Yet such inequality will never hold, since $\beta_k = (X^T y)_k$ and it would imply

$$\beta_k (X^T y)_k - \frac{1}{2}(\beta_k)^2 < \lambda \quad \Rightarrow \quad \frac{1}{2}(X^T y)_k^2 < \lambda \quad \Rightarrow \quad \frac{1}{2}\lambda^2 < \lambda \quad (8)$$

Which contradicts $\lambda > 2$. Thus Lass0 will never remove an element from $supp(\beta^{L_1})$.

Similarly, if we let $\beta^{new} = \hat{OLS}(supp(\beta) \cup \{k\}, y, X)$ Lass0 will only choose β^{new} and add element k to $supp(\beta^{L_1})$ if,

$$\frac{1}{2}\|y - X\beta^{new}\|_2^2 - \frac{1}{2}\|y - X\beta\|_2^2 < -\lambda \quad \Rightarrow \quad -\frac{1}{2}(X^T y)_k^2 < -\lambda \quad (9)$$

Thus Lass0 will add element k to $supp(\beta^{L_1})$ if and only if $\sqrt{2\lambda} < (X^T y)_k$. Therefore, $supp(\beta^{Lass0}) = supp(\beta^{L_0})$. Furthermore, since β^{Lass0} is optimized by OLS, $\beta^{Lass0} = \beta^{L_0}$.

iii. **Case $\boldsymbol{\lambda < 2}$:** Since $\sqrt{2\lambda} > \lambda$, then $supp(\beta^{L_0}) \subseteq supp(\beta^{L_1})$. The result that $\beta^{Lass0} = \beta^{L_0}$ follows from an analogous argument to the above, omitted for the sake of brevity. ■

For sparse solutions, it is important to know how a given algorithm will behave when faced with strongly correlated features. For example, the elastic net [16] assigns identical coefficients to identical variables. In contrast, L_1 regularization picks one of the strongly correlated features. The latter behavior is desirable in situations where including both variables in the support would be considered redundant. We now prove that when two variables are strongly correlated, Lass0 behaves similarly to L_1 regularization: it only selects one among a group of strongly correlated features.

Theorem 2. *Assume that $\mathbf{x_i} = k\mathbf{x_j}$, then either $\beta_i^{Lass0} = 0$ or $\beta_j^{Lass0} = 0$ (or both).*

Proof. Let β be the solution at any step of Lass0. We will prove that if both indices $\{i, j\} \in supp(\beta)$, meaning $\beta_i \neq 0$ and $\beta_j \neq 0$, at least one of them will become zero in the solution.

Without loss of generality, let β^{new} be the least squares solution that preserves all the constraints of β and also enforces $\beta_i^{new} = 0$. Let $\beta_j^{new} = k\beta_i + \beta_j$, then $||y - X\beta^{new}||_2^2 = ||y - X\beta||_2^2$, implying $L_0(\beta_{new}, y, X, \lambda) < L_0(\beta, y, X, \lambda)$. ■

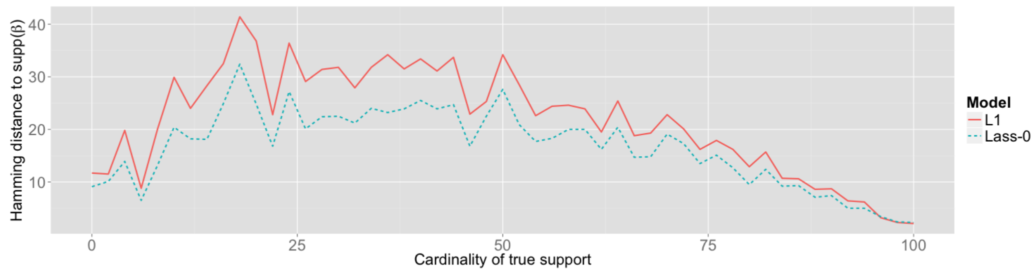

Figure 1: Average Hamming distance between $supp(\beta^{Lass\mathtt{0}})$ or $supp(\beta^{L_1})$ and the true $supp(\beta)$ over 10 CV tests. The Lass0 consistently chooses models closer to the true support.

Data	NRMSE L_1	NRMSE Lass0	p	$\lvert supp(\beta^{L_1})\rvert$	$\lvert supp(\beta^{Lass\mathtt{0}})\rvert$
Pyrimidines	101.4 ± 47.5	103.1 ± 42.3	28	16.1 ± 5.5	7.6 ± 5.1
Ailerons	42.2 ± 1.8	42.5 ± 1.9	41	24 ± 3.3	6.9 ± 1.1
Triazines	98.9 ± 14.5	97.5 ± 19.7	61	17.9 ± 9.2	7.3 ± 6.6
Airplane stocks	36.1 ± 5.2	36.4 ± 5.6	10	8.5 ± 0.5	7.8 ± 0.9
Pole Telecomm	73.3 ± 2.1	73.4 ± 2.1	49	22.7 ± 1.1	24.5 ± 0.9
Bank domains	69.9 ± 3	70.7 ± 3.1	33	9.2 ± 2.7	5.2 ± 8.2
Pumadyn domains	89 ± 2.2	88.9 ± 2.4	33	5.2 ± 8.2	1 ± 0
Breast Cancer	93.5 ± 15.3	96.7 ± 19.7	33	16 ± 8.7	18.8 ± 5.7
Mice	103.5 ± 5	105 ± 6.6	100	17 ± 6.6	6.3 ± 4.3

Table 1: Mean and standard deviation from Lass0 and L_1 regularization on real data for 10 CV runs

4 Experimental results

We generate synthetic data from a linear model $y = X\beta + \epsilon$, where each sample is generated $X_j \sim N(\mu, \Sigma)$ using Σ with high correlation. The coefficients are generated as $\beta \sim$ Uniform$(-1, 1)$, with sparsity enforced by setting some β_i to zero. We compare $supp(\beta^{Lass\mathtt{0}})$ and $supp(\beta^{L_1})$ against the true underlying support, $supp(\beta)$. We use 10-fold cross validation (CV) testing and report the average Hamming distance between the estimated and true supports. Figure 1 shows Hamming distances over different levels of sparsity in the true support. The Lass0 consistently yields models which are closer to the true support than the optimally chosen L_1 model.

We evaluate the Lass0 on nine real-world data sets sourced from the publicly available repositories [17, 18]. Table 1 shows the mean and standard deviation for the normalized root mean squared error (NRMSE) and cardinality of the support for the estimated β. For all data sets, both regularization methods produce very similar NRMSE values. However, in most cases the Lass0 reduced the size of the active set, often by 50% or more. Combined with the above results showing that the Lass0 yields models closer to the true sparse synthetic model, we see that the Lass0 tends to produce sparser, more fidelitous models than L_1 regularization.

5 Future work

We intend to build upon Theorem 1 to support our empirical observations. Additionally, we expect that this paper's general approach can be applied to other non-convex optimization problems. While convex relaxations may yield interesting problems in their own right, they are often good approximations to non-convex solutions. Using convex results to initialize an efficient search for a locally optimal non-convex solution can combine the strengths of convex and non-convex formulations.

Acknowledgments

We thank Ryan Tibshirani for his advice and fruitful suggestions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1252522 and the National Science Foundation award No. IIS-0953330.

References

[1] Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

[2] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[3] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems*, pages 1921–1928, 2009.

[4] Hiroshi Konno and Rei Yamamoto. Choosing the best set of variables in regression analysis using integer programming. *Journal of Global Optimization*, 44(2):273–282, 2009.

[5] Ryuhei Miyashiro and Yuichi Takano. Subset selection by mallows cp: A mixed integer programming approach. *Expert Systems with Applications*, 42(1):325–331, 2015.

[6] Cristian Gatu and Erricos John Kontoghiorghes. Branch-and-bound algorithms for computing the best-subset regression models. *Journal of Computational and Graphical Statistics*, 2012.

[7] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[8] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[9] Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[10] Carlos Ramirez, Vladik Kreinovich, and Miguel Argaez. Why l1 is a good approximation to l0: A geometric explanation. *Journal of Uncertain Systems*, 7, 2013.

[11] Jerome H Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3):722–738, 2012.

[12] Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495), 2011.

[13] Dongyu Lin, Emily Pitler, Dean P Foster, and Lyle H Ungar. In defense of l0. In *Workshop on Feature Selection,(ICML 2008)*, 2008.

[14] Dongyu Lin, Dean P Foster, and Lyle H Ungar. A risk ratio comparison of l0 and l1 penalized regressions. *University of Pennsylvania, Tech. Rep*, 2010.

[15] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[16] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[17] William Valdar, Leah C Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O Cookson, Martin S Taylor, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38(8):879–887, 2006.

[18] M. Lichman. UCI machine learning repository, 2013.