Revisiting Sparse Learning Methods: A Comprehensive Comparison of Best Subset Selection and LASSO

Anonymous authors
Paper under double-blind review

Abstract

Understanding the comparative performance of L_0 and L_1 models is crucial for developing accurate and efficient machine learning systems, particularly in noisy, real-world settings. The current understanding in the literature is that L_1-penalized linear models perform better than L_0 models as noise increases. However, prior studies have largely relied on small and synthetic datasets and limited comparisons between differing optimizers, leaving practical implications for diverse applications underexplored. We fill these gaps in analysis by testing multiple different L_0 and L_1 based optimizers on a larger variety of real datasets, and demonstrate that performance differences between L_0 and L_1 models depend significantly on the choice of optimizer and dataset characteristics. In many cases, the difference in performance by changing the optimization algorithm, while leaving the regularization penalty constant, is larger than the differences in changing the penalty. Additionally, we demonstrate cases where an L_0-penalized model can be both sparser and more accurate than the L_1-penalized variants. Together, our results show that even convex L_1 models can vary significantly in performance according to optimizer implementation, and that L_0 penalized models are more viable for many smaller real-world and noisy situations than previously recognized.

1 Introduction

Methods for sparse regression and classification are useful for a multitude of reasons, especially when confronting problems with a large number of features. Induced sparsity can be important for reducing overfitting and improving model generalization on unseen data. The regularization reduces the variance of the model predictions, and this has been demonstrated to improve model generalization on real-world datasets. Moreover, sparsity can reduce required resources, and improve model interpretability. These are some of the reasons why methods for sparse linear and logistic regression are among the most commonly used tools in the toolbox for machine learning Hastie et al. (2015).

The LASSO Tibshirani (1996) is a widely used and highly successful regularization method for regression and classification problems and induces both coefficient sparsity as well as coefficient shrinkage. If we denote by $f(\boldsymbol{\theta})$ the loss function of a regression or classification problem, the LASSO in its primal form is given by

$$\min_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}) \tag{1}$$

$$\text{s.t.} \quad \|\boldsymbol{\theta}\|_1 < \kappa. \tag{2}$$

Being convex, it can equivalently be solved via its Lagrangian dual,

$$\min_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}) + \lambda_1 \|\boldsymbol{\theta}\|_1. \tag{3}$$

Solutions to the constrained optimization problem naturally induce sparsity – some subset of the coefficients $\boldsymbol{\theta}$ will be zero. In addition, the L_1 constraint (or penalty) on

$$\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{p} \|\theta_i\|_1 \tag{4}$$

induces shrinkage of the coefficient magnitudes.

The Best Subset Selection regularization scheme instead uses the L_0 pseudo-norm. Therefore, it induces sparsity but without any shrinkage of the coefficient magnitudes. In primal form,

$$\min_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}) \tag{5}$$

$$\text{s.t.} \quad \|\boldsymbol{\theta}\|_0 < k, \tag{6}$$

where

$$\|\boldsymbol{\theta}\|_0 = \sum_{i=1}^{p} 1(\theta_i \neq 0). \tag{7}$$

This presents a mixed-integer optimization problem. Exact solutions for regression problems via the leaps and bounds algorithm Furnival & Wilson (1974) were available in the `leaps` and `bestglm` packages, but could only solve problems with $p \sim 30$ features. Recent advances enabled mixed-integer optimizers such as Bertsimas et al. (2016) to tackle problem sizes roughly having a number of samples $n \sim 10^2$ and a number of features $p \sim 10^4$.

One can also construct feasible solutions via a Lagrangian dual,

$$\min_{\boldsymbol{\theta}} \quad f(\boldsymbol{\theta}) + \lambda_0 \|\boldsymbol{\theta}\|_0. \tag{8}$$

However, such feasible solutions are not guaranteed equivalence with those of the primal equation. There also exist first order methods Blumensath & Davies (2009); Bahmani et al. (2013) and second-order methods Yuan & Liu (2017); Zhou et al. (2021); Wang et al. (2021) for approximately solving the Best Subset Selection problem for regression and classification.

In this manuscript, we revisit and challenge earlier findings by comparing several Best Subset Selection and LASSO optimizers on an extensive selection of datasets, with varying amounts of feature and label noise. We demonstrate that the choice of optimizer can be equally as important as the choice of regularization class under large levels of noise, and certain Best Subset Selection optimizers retain stable performance at moderate levels of noise.

2 Related Studies

Having coefficient shrinkage, the LASSO was believed to be superior to Best Subset Selection for datasets for data with lower signal-to-noise ratio (SNR) Hastie (2001). Some evidence was presented to this effect in previous studies for regression Hastie et al. (2020) and classification Dedieu et al. (2021). However, these studies primarily demonstrated results on simulated data and had very limited results on real datasets. Hastie et al. (2020) compared Best Subset Selection and LASSO for regression problems, and only used the mixed-integer optimization method provided by Bertsimas et al. (2016) for the Best Subset Selection. They concluded that LASSO gave better test accuracy in the low SNR regime, and worse accuracy in the high SNR regime, and the transition point in SNR depended on the problem dimensions: the number of training samples n and number of features p. This work only studied regression problems, rather than classification problems which are the focus of this manuscript. Moreover, this work performed comparisons exclusively on simulated/synthetic data, where the underlying data-generating process is known.

On the other hand, Dedieu et al. (2021) studied binary classification problems, which are also the focus of this manuscript, and compared their optimizer designed to solve the combined $L_0 + \alpha L_q$ penalty with LASSO. They found that combined $L_0 + \alpha L_2$ penalty could outperform LASSO, with the L_2 penalty inducing coefficient shrinkage and reducing variance. Unfortunately, their study did not share any results for the pure (L_0-only) best subset selector, which would have been highly relevant and informative. At a high level, their conclusions were largely similar to Hastie et al. (2020), however, these conclusions were based mostly on simulated datasets with very limited real datasets. The simulated data were generated as multivariate Gaussian features with various correlation strengths. For comparison on real datasets, they showed only three (Arcene, Dexter, and Dorothea) taken from the NIPS 2003 Feature Selection Challenge Guyon et al. (2004).

In contrast, in this study, we compare the performance of Best Subset Selection and LASSO methods for a wide set of binary classification problems. In addition, we compare a variety of optimizers within each regularization class. We compare four optimizers for the Best Subset Selection. In Iterative Hard Thresholding (IHT) Blumensath & Davies (2009): the weights are updated at each iteration by a projected gradient descent method,

$$\boldsymbol{\theta}_{t+1} = \Pi_k(\boldsymbol{\theta}_t - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t)), \tag{9}$$

where η is a learning rate, and the operator Π_k projects the weights onto the nearest point of the L_0 ball $|\boldsymbol{\theta}|_0 < k$. This projection is accomplished by sorting the weights $\boldsymbol{\theta}$ by their magnitude and keeping the k-largest while zeroing the rest:

$$\Pi_k(\boldsymbol{\theta}) = \boldsymbol{\theta'} \quad \text{where} \quad \theta'_i = \begin{cases} \theta_i & \text{if } |\theta_i| \geq |\theta_{[k]}| \\ 0 & \text{if } |\theta_i| < |\theta_{[k]}|, \end{cases} \tag{10}$$

where $\theta_{[k]}$ denotes the k-th largest element in the sorted list of $|\theta_i|$ values.

We also extend the Iterative Hard Thresholding method to include a proposal vector given by gradient descent with momentum (IHTM),

$$\boldsymbol{\theta}_{t+1} = \Pi_k(\boldsymbol{\theta}_t - \eta \boldsymbol{v}_t), \tag{11}$$

where

$$\boldsymbol{v}_t \equiv \beta \boldsymbol{v}_{t-1} + \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t)), \tag{12}$$

with momentum decay parameter $\beta = 0.9$.

We also compare the Best Subset Selection optimizers given by the coordinate descent plus local combinatorial search method described in Dedieu et al. (2021) and implemented in Hazimeh et al. (2023), which we refer to as L0Learn. We compare their optimizer with both the pure best subset L_0 selector as well as a mixed selector which has both L_0 and L_2 penalties. For LASSO, we compare two optimizers: LIBLINEAR Fan et al. (2008) and SAGA Defazio et al. (2014).

Additionally, we compare the empirical and practical performance of the methods on a wide variety of binary classification datasets from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html` at variable SNR. We perform this extensive experimentation on real datasets to give guidance that applies to real-world classification problems in which the ground truth data-generating process is not known. Additionally, these experiments are performed to mimic a typical machine learning workflow for real-world data: models are optimized on training data, hyper-parameters are optimized on held-out validation data, and expected performance in-distribution is estimated by performance on held-out testing data. Our experiments challenge the simple story that the relative strengths and weaknesses of Best Subset Selection and LASSO are mostly a function of the signal-to-noise ratio. Moreover, by comparing multiple optimizers for Best Subset Selection and LASSO, we also challenge the idea that the cost functions alone are predictive of performance as SNR is varied. Instead, we will show that differences between different optimizers can be as significant and relevant in determining performance.

3 Experimental Method and Results

In this section, we present comparisons between variants of the Iterative Hard Thresholding Blumensath & Davies (2009) and L0Learn Dedieu et al. (2021); Hazimeh et al. (2023) Best Subset Selection optimizers, the optimizer with mixed L_0 and L_2 penalty from L0Learn, and two LASSO optimizers: LIBLINEAR Fan et al. (2008) and SAGA Defazio et al. (2014). Their performance is compared on a wide variety of binary classification datasets from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html` as well as three datasets from the NIPS 2003 feature selection challenge, that are listed in Table 1.

The datasets are first balanced by label via undersampling. We add noise to the data by two different methods which are plotted separately:

1. we add Gaussian noise with standard deviation σ_X to the normalized features of each dataset, allowing us to explicitly vary the amount of noise present in the features not associated with the true label, or

Table 1: All the datasets included in comparisons between L_0 and L_1 optimizers studied.

a1a	arcene
australian	breast-cancer
cod-rna	colon-cancer
dexter	diabetes
dorothea	german.numer
gisette	heart
ijcnn1	ionosphere
leukemia	liver-disorders
madelon	phishing
sonar	splice
svmguide1	w1a

2. we flip the binary labels in the dataset with probability p_y.

We hold out separate validation and test data from the training data. Following a practical/real-world scenario, each model is optimized on the training data with a hyperparameter controlling sparsity. This is an integer k for Best Subset Selection, and real numbers λ_1 for the LASSO penalty or λ_2 for the L_2 penalty. The hyperparameters are optimized via fifty trials of optuna Akiba et al. (2019).

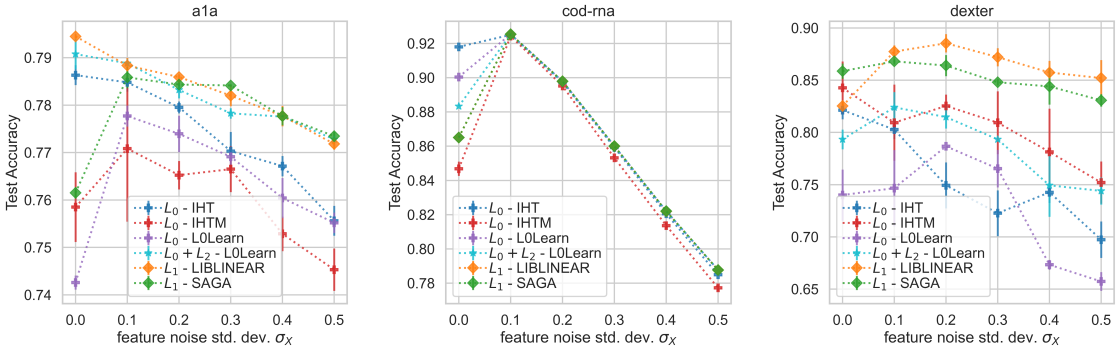
For Figs. 1a through 1i, the top 10% of trials according to validation set accuracy are selected, and their mean and standard deviation on the separate test set is shown. To facilitate more easy visual comparison, we point out that all L_0 methods have markers that are '+', while all L_1 methods have markers that are diamonds. We first discuss the optimizer behaviors on individual datasets with varying feature noise. Interestingly, only two out of twenty-one datasets, `a1a` (Fig. 1a) and `dexter` (Fig. 1c), show the behavior of Best Subset Selection compared to LASSO that we would expect based on prior studies as feature noise increases. Namely, for these datasets, there is clear degradation in the performance of *all* Best Subset Selection methods as noise is increased, with significantly lesser degradation in the performance of the LASSO methods.

For the `cod-rna` (Fig. 1b) and `ijcnn1` (Fig. 1d)the performance of all methods degrades systematically at large levels of added noise. However, there is no statistically significant increase in the performance gap between the L_1 and L_0 methods as noise increases when one includes all possible L_0 optimizers (L_0 - L0Learn performs as well as both Lasso methods at the highest level of noise). Perhaps more interestingly, certain datasets such as `leukemia` (Fig. 1e) and `madelon` (Fig. 1f) datasets exhibit performance that is *contrary* to prior studies themes, with the performance of certain L_0 - methods initially increasing as noise is increased, sometimes closing the performance gap or even overtaking the performance of some L_1 methods.

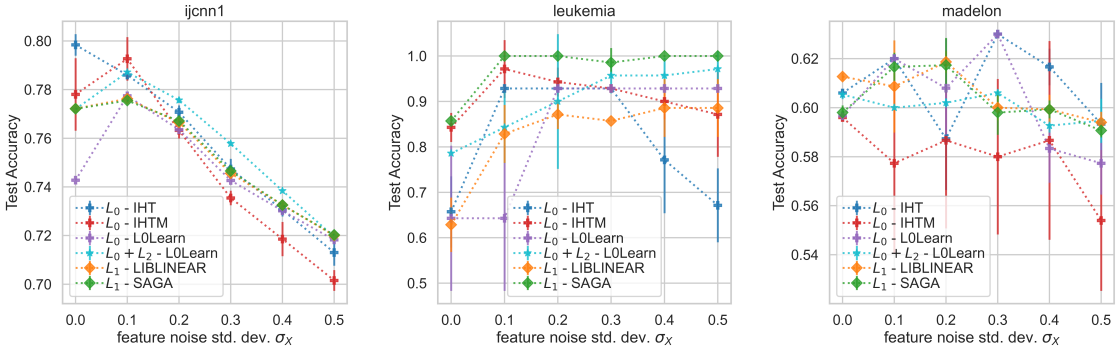
Now, we discuss the behaviors on individual datasets as the label noise is increased. Certain datasets exhibit what we would likely suspect, that as label noise is increased, test performance systematically decreases. This includes the `a1a` (Fig. 1g) and `gisette` (Fig. 1h) datasets among others. For certain datasets such as `arcene` (Fig. 1i), the divergence between the performance of the two LASSO optimizers grows significantly, again highlighting that the specific optimizer can be as important a choice as the regularization penalty (LASSO vs best subset) in test performance.

Perhaps the easiest conclusion to draw from these comparisons is the absence of a simple pattern or 'story' between Best Subset Selection and LASSO as it pertains to the effect of coefficient shrinkage; on the contrary, the performance differences among the different Best Subset Selection optimization methods IHT, IHTM, and L0Learn, or between LIBLINEAR and SAGA LASSO optimizers, are often as large or larger than the differences between L_0 and L_1 methods. And, similarly for the differences between the LASSO optimizers.

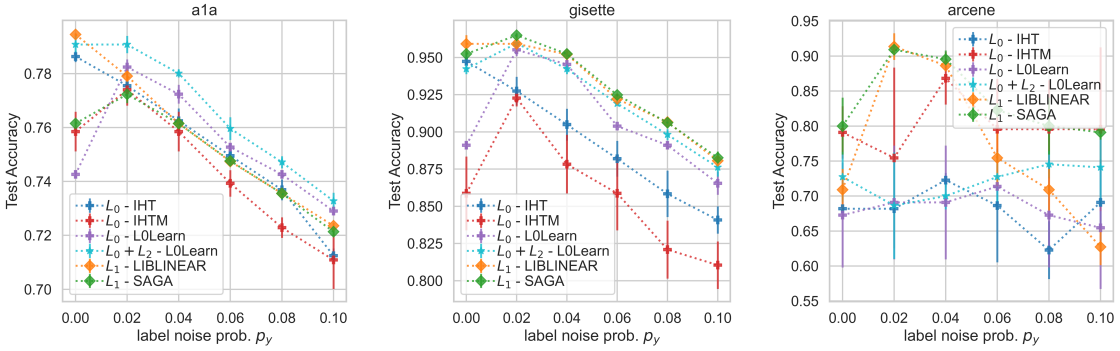
In Fig. 2a we show plots demonstrating the performance of each optimizer over all datasets studied as feature noise is varied. At the very largest levels of added feature noise ($\sigma_X = 0.5$), the previous understanding seems to hold well. That is, both LIBLINEAR and SAGA L_1 optimizers have test performance which is



(a) Plot a (b) Plot b (c) Plot c



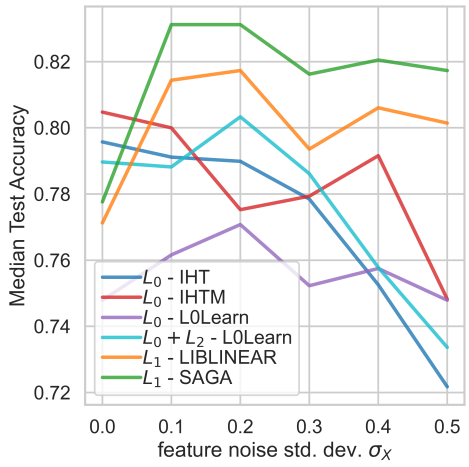
(d) Plot d (e) Plot e (f) Plot f



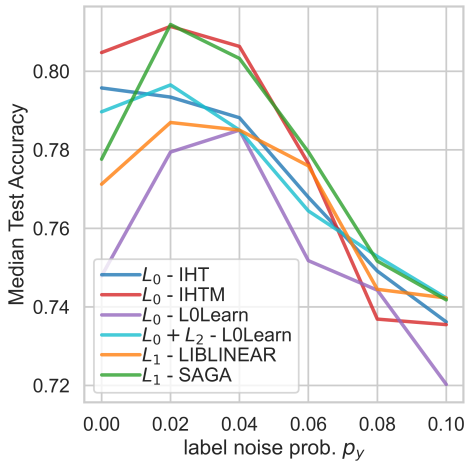
(g) Plot g (h) Plot h (i) Plot i

systematically slightly higher than all of the L_0 optimizers, including L0Learn's mixed $L_0 + L_2$ penalty. However, that the $L_0 + L_2$ penalty (with coefficient shrinkage) does not systematically improve performance at the largest levels of noise w.r.t. pure L_0 methods is surprising, and contrary to the L0Learn GitHub page, which 'strongly recommends' using the mixed penalty, justified by the same concerns regarding signal-to-noise ratio and overfitting without shrinkage Hazimeh et al. (2023). However, we see that the situation is more nuanced. Both the LIBLINEAR and SAGA L_1 methods actually showed increases in performance at small to intermediate feature noise levels. Evidently, small amounts of feature noise act as regularizations for some L_1 methods and actually increase generalization. Additionally, the L0Learn optimizer showed stable performance across noise levels without a systematic decrease.

In Fig. 2b we show plots demonstrating the performance of each optimizer over all datasets studied as label noise is varied. In this case, we see that all optimizers, either L_0 or L_1, have similar trends, where degradation is only significant at the largest amount of label noise (10%). Importantly, we see that the SAGA L_1 method



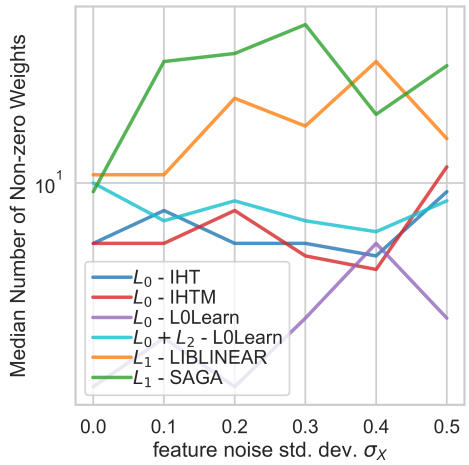
(a) Test accuracy as a function of added feature noise.



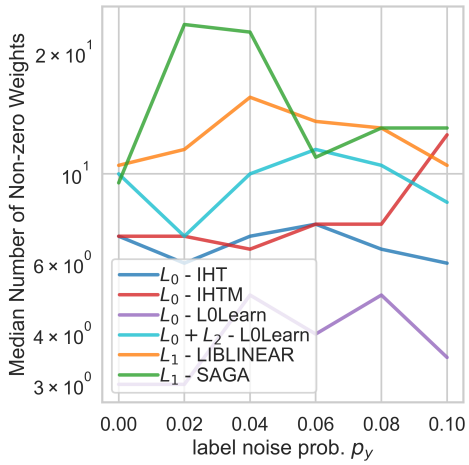
(b) Test accuracy as a function of added label noise.

Figure 2: The test accuracy performance (the median taken over datasets) for each optimizer as a function of added feature noise (a) and label noise (b).

has a moderate advantage over all pure L_0 optimizers at this noise level, but the LIBLINEAR solver does not.



(a) Model sparsity as a function of added feature noise.

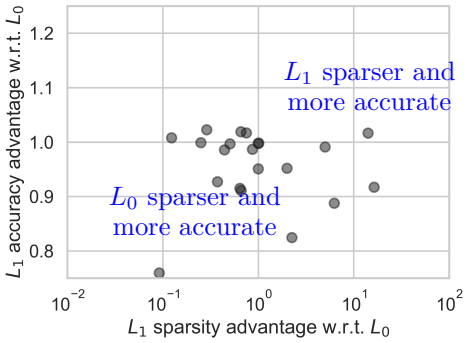


(b) Optimizer performance as a function of added label noise.

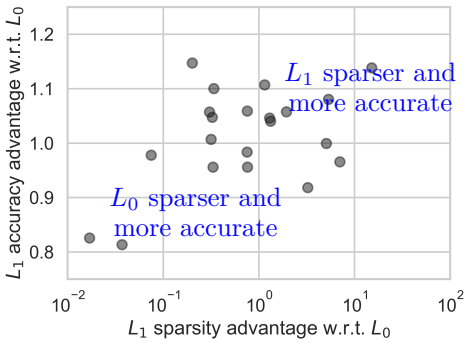
Figure 3: Model sparsity (measured by the number of nonzero weights, median over datasets) and optimizer performance as a function of added feature noise (a) and label noise (b).

In Fig. 3a we show plots demonstrating the sparsity, measured by the number of nonzero weights, of each optimizer over all datasets studied as feature noise is varied. Overall, we see that L_0 methods tend to produce sparser (fewer nonzero weights) models than both L_1 methods. The sparsest models are produced by the L0Learn L_0 optimizer, which is significantly sparser than even the other L_0 optimizers. But even the iterative hard thresholding method produces sparser models than either L_1 solver.

In Fig. 3b we show plots demonstrating the sparsity, measured by the number of nonzero weights, of each optimizer over all datasets studied as label noise is varied. Overall, the situation is similar to the previous one. L_0 methods are sparser on average, with L0Learn's optimizer producing the sparsest models. However, there is considerable variation in the sparsity of different L_0 models, with L0Learn's $L_0 + L_2$ optimizer producing models with comparable sparsity to the LIBLINEAR L_1 optimizer.



(a) Datasets without added noise.



(b) Datasets with the largest added feature noise.

Figure 4: Comparison of the best-performing L_0 and L_1 optimizers in terms of sparsity and test performance. (a) Results for datasets without added noise. (b) Results for datasets with the largest amount of added feature noise.

In Fig. 4a we plot the relative (dis)advantages of LASSO methods and Best Subset Selection methods from a typical model selection perspective. This is calculated as follows. For each dataset (without any added noise), we take the top-performing L_0 method by validation accuracy and the top-performing L_1 method by validation accuracy. The test accuracy and number of nonzero weights for each are saved. Then we compute ratios: the L_1 sparsity advantage w.r.t. L_0 is given $\text{NNZ}_{L_0}/\text{NNZ}_{L_1}$ (less nonzero weights is more sparse and more desirable), and the L_1 accuracy advantage w.r.t. L_0 is given Test Acc._{L_1}/Test Acc._{L_0}. We can read this plot as follows: In the left half, the best L_1 method produces a *less sparse* model than the best L_0 method, while in the right half, it produces a *more sparse* model than L_0. In the top half, the best L_1 method is *more accurate* than the best L_0 method, while in the bottom half, it is *less accurate*. We see that for all datasets but one when L_1 is more accurate (top half), it is also less sparse or equivalently sparse. Therefore, there is sometimes a tradeoff between accuracy and sparsity. On the other hand, when L_0 is more accurate, it can be either more or less sparse roughly equally often. In such cases that the best L_0 method is more accurate and more sparse, it is a Pareto-optimal choice over the best L_1 method.

Similarly, in Fig. 4b we plot the relative (dis)advantages of LASSO methods and Best Subset Selection methods from a typical model selection perspective, but now on the datasets with the largest amount of added feature noise $\sigma_X = 0.5$. In this case, we see that LASSO methods, when they are more accurate (top half) can also be more sparse (right half). However, we see that there are still several datasets for which the best subset selector has both the sparsity and accuracy advantage even at these large levels of noise (lower signal-to-noise).

4 Conclusion

This manuscript provides a thorough evaluation of sparse learning techniques, challenging common assumptions about LASSO and Best Subset Selection across real-world datasets. Our experiments show that the strengths and weaknesses of these methods vary significantly, especially under different noise levels.

A key insight is that optimizer-specific behaviors can heavily influence performance, both in terms of test accuracy and model sparsity, sometimes more than the choice between L_1 and L_0 regularization. This

underscores the importance of considering the interactions of both the regularization and the optimization strategy with noise in the data in practice.

Our results indicate that the traditional view linking sparse learning performance primarily to signal-to-noise ratio and coefficient shrinkage is overly simplistic. The choice of optimizer plays a critical role, with varied performance under the same noise conditions. Notably, iterative hard thresholding L_0 methods demonstrate small but systematic performance gains at low to moderate noise levels. Also, the SAGA L_1 method has a moderate advantage over the best L_0 optimizer (IHT) at large levels of label noise, but the LIBLINEAR L_1 solver does not. For all datasets but one, when the best L_1 optimizer is more accurate, it is also less or equivalently sparse. This study offers actionable insights through a thorough comparison across multiple datasets and optimizers.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *The Journal of Machine Learning Research*, 14(1):807–841, 2013.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. 2016.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Antoine Dedieu, Hussein Hazimeh, and Rahul Mazumder. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. *Journal of Machine Learning Research*, 22(135):1–47, 2021.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *the Journal of machine Learning research*, 9:1871–1874, 2008.

George M. Furnival and Robert W. Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974. ISSN 00401706. URL `http://www.jstor.org/stable/1267601`.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. *Advances in neural information processing systems*, 17, 2004.

Trevor Hastie. Tibshirani r. friedman j.: The elements of statistical learning. *Friedman J.: The elements of statistical learning*, 2001.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8, 2015.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

Hussein Hazimeh, Rahul Mazumder, and Tim Nonet. L0learn: A scalable package for sparse learning using l0 regularization. *Journal of Machine Learning Research*, 24(205):1–8, 2023.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Rui Wang, Naihua Xiu, and Shenglong Zhou. An extended newton-type algorithm for 2-regularized sparse logistic regression and its efficiency for classifying large-scale datasets. *Journal of Computational and Applied Mathematics*, 397:113656, 2021.

Xiao-Tong Yuan and Qingshan Liu. Newton-type greedy selection methods for ℓ_0-constrained minimization. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2437–2450, 2017.

Shenglong Zhou, Naihua Xiu, and Hou-Duo Qi. Global and quadratic convergence of newton hard-thresholding pursuit. *Journal of Machine Learning Research*, 22(12):1–45, 2021.