Picasso: A Sparse Learning Library for High Dimensional Data Analysis in R and Python

J. Ge, X. Li, H. Jiang, H. Liu, T. Zhang, M. Wang and T. Zhao*

Abstract

We describe a new library named picasso, which implements a unified framework of pathwise coordinate optimization for a variety of sparse learning problems (e.g., sparse linear regression, sparse logistic regression, sparse Poisson regression and sparse square root loss linear regression), combined with efficient active set selection strategies. Besides, the library allows users to choose different sparsity-inducing regularizers, including the convex ℓ_1, nonvoncex MCP and SCAD regularizers. The library is coded in C, has user-friendly R and Python wrappers, and can scale up to large problems efficiently with the memory optimized using sparse matrix output.

1 Introduction

The pathwise coordinate optimization is undoubtedly one the of the most popular solvers for a large variety of sparse learning problems. By leveraging the solution sparsity through a simple but elegant algorithmic structure, it significantly boosts the computational performance in practice (Friedman et al., 2007). Some recent progresses in (Zhao et al., 2017; Li et al., 2016) establish theoretical guarantees to further justify its computational and statistical superiority for both convex and nonvoncex sparse learning, which makes it even more attractive to practitioners.

We recently developed a new library named picasso, which implements a unified toolkit of pathwise coordinate optimization for solving a large class of convex and nonconvex regularized sparse learning problems. Efficient active set selection strategies are provided to guarantee superior statistical and computational preference. Specifically, we implement sparse linear regression, sparse logistic regression, sparse Poisson regression, scaled sparse linear regression, and undirected graphical model estimation (Tibshirani, 1996; Belloni et al., 2011; Sun and Zhang, 2012; Ravikumar et al., 2010; Liu and Wang, 2012; Sun and Zhang, 2013). The options of regularizers

*Xingguo Li is affiliated with Department of Electrical and Computer Engineering at University of Minnesota; Tong Zhang is affiliated with Tencent AI Lab; Jason Ge, Mengdi Wang, and Han Liu are affiliated with Department of Operations Research and Financial Engineering at Princeton University; Haoming Jiang and Tuo Zhao are affiliated with Schools of Industrial & Systems Engineering and Computational Science & Engineering at Georgia Institute of Technology, Atlanta, GA 30332; Emails: lixx1661@umn.edu, jiange@princeton.edu, tourzhao@gatech.edu; Xingguo Li and Jason Ge contributed equally; Tuo Zhao is the corresponding author.

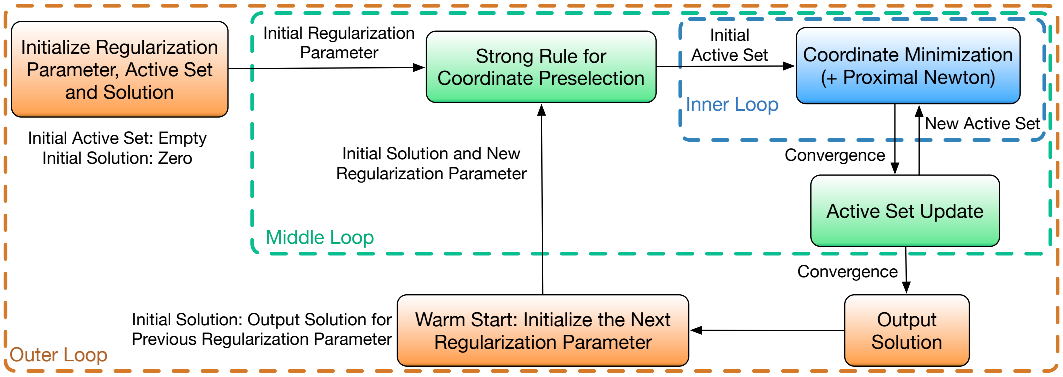

Initialize Regularization Parameter, Active Set and Solution
Initial Active Set: Empty
Initial Solution: Zero
Initial Regularization Parameter
Strong Rule for Coordinate Preselection
Initial Active Set
Coordinate Minimization (+ Proximal Newton)
Inner Loop
New Active Set
Convergence
Initial Solution and New Regularization Parameter
Active Set Update
Middle Loop
Convergence
Initial Solution: Output Solution for Previous Regularization Parameter
Warm Start: Initialize the Next Regularization Parameter
Output Solution
Outer Loop

Figure 1: The pathwise coordinate optimization framework with 3 nested loops : (1) Warm start initialization; (2) Active set selection, and strong rule for coordinate preselection; (3) Active coordinate minimization.

include the ℓ_1, MCP, and SCAD regularizers (Fan and Li, 2001; Zhang, 2010). Unlike existing packages implementing heuristic optimization algorithms such as ncvreg or glmnet (Breheny, 2013; Friedman et al., 2010), our implemented algorithm picasso have strong theoretical guarantees that it attains a global linear convergence to a unique sparse local optimum with optimal statistical properties (e.g. minimax optimality and oracle properties). See more technical details in Zhao et al. (2017); Li et al. (2016).

2 Algorithm Design and Implementation

The algorithm implemented in picasso is mostly based on the generic pathwise coordinate optimization framework proposed by Zhao et al. (2017); Li et al. (2016), which integrates the warm start initialization, active set selection strategy, and strong rule for coordinate preselection into the classical coordinate optimization. The algorithm contains three structurally nested loops as shown in Figure 1:

(1) Outer loop: The warm start initialization, also referred to as the pathwise optimization scheme, is applied to minimize the objective function in a multistage manner using a sequence of decreasing regularization parameters, which yields a sequence of solutions from sparse to dense. At each stage, the algorithm uses the solution from the previous stage as initialization.

(2) Middle loop: The algorithm first divides all coordinates into active ones (active set) and inactive ones (inactive set) by a so-called strong rule based on coordinate gradient thresholding (Tibshirani et al., 2012). Then the algorithm calls an inner loop to optimize the objective, and update the active set based on efficient active set selection strategies. Such a routine is repeated until the active set no longer changes

(3) Inner loop: The algorithm conducts coordinate optimization (for sparse linear regression) or proximal Newton optimization combined with coordinate optimization (for sparse logistic regression, Possion regression, scaled sparse linear regression, sparse undirected graph estimation) only over active coordinates until convergence, with all inactive coordinates staying zero values. The active coordinates are updated efficiently using an efficient "naive update" rule that only operates on the non-zero coefficients. Better efficiency is achieved by the "covariance update" rule. See more details in (Friedman et al., 2010). The inner loop terminates when the successive descent is within a predefined numerical precision.

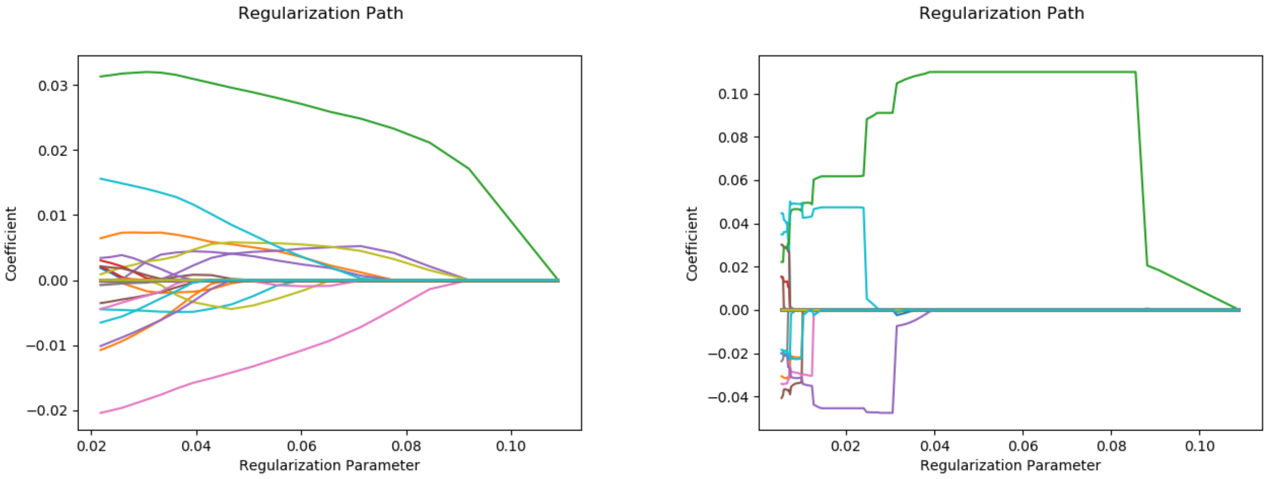
The warm start initialization, active set selection strategies, and strong rule for coordinate preselection significantly boost the computational performance, making pathwise coordinate optimization one of the most important computational frameworks for sparse learning. The library is implemented in `C` with the memory optimized using sparse matrix output, and called from `R` and `Python` by user-friendly interfaces. The numerical evaluations show that `picasso` is efficient and can scale to large problems.

3 Examples of R User Interface

We illustrate the user interface by analyzing the eye disease data set in `picasso`.

```
> # Load the data set
> library(picasso); data(eyedata)
> # Lasso
> out1 = picasso(x,y,method="l1",type.gaussian="naive",nlambda=20,
+                lambda.min.ratio=0.2)
> # MCP regularization
> out2 = picasso(x,y,method="mcp", gamma = 1.25, prec=1e-4)
> # Plot solution paths
> plot(out1); plot(out2)
```

The program automatically generates a sequence of regularization parameters and estimate the corresponding solution paths based on the ℓ_1 and MCP regularizers respectively. For the ℓ_1 regularizer, the number of regularization parameters as 20, and the minimum regularization parameter as `0.2*lambda.max`. Here `lambda.max` is the smallest regularization parameter yielding an all zero solution (automatically calculated by the library). For the MCP regularizer, we set the concavity parameter as $\gamma = 1.25$, and the pre-defined accuracy as 10^{-4}. Here `nlambda` and `lambda.min.ratio` are omitted, and therefore set by the default values (`nlambda=100` and `lambda.min.ratio=0.01`). We further plot two solution paths in Figure 4.



(a) ℓ_1

(b) MCP

Figure 2: The solution paths of the ℓ_1-regularized and MCP-regularized sparse linear regression.

4 Examples of Python User Interface

We illustrate the usage of Python interface under the same setting as the above section [1].

```
> # Load the library and the data set
> from pycasso import core
> import numpy as np
> data = np.load("eyedata.npy").item()
> x = data["data"]
> y = data["label"]
> #Lasso
> s1 = core.Solver(x,y,penalty="l1",type_gaussian="naive",nlambda=20,\
>      lambda_min_ratio=0.2)
> # MCP regularization
> s2 = core.Solver(x,y,penalty="mcp", gamma = 1.25, prec=1e-4)
> # Plot solution paths
> s1.plot(); s2.plot()
```

The results are stored in out1 and out2. We further plot two solution paths and see the same result in Figure 4.

[1] The library is named ”Pycasso” for Python to avoid the conflict with https://pypi.python.org/pypi/Picasso/0.0.2.

5 Numerical Simulation

To demonstrate the superior efficiency of our package, we compare `picasso` with a popular `R` package `ncvreg` for nonconvex regularized sparse regression, and with the most popular `R` package `glmnet` for convex regularized sparse regression. All experiments are evaluated on an Intel Xeon CPU E5-2667 v4 3.20GHz and under R version 3.4.3. Timings of the CPU execution are recored in seconds and averaged over 10 replications on a sequence of 100 regularization parameters. All algorithms are compared on the same regularization path and the convergence threshold are adjusted so that similar objective function gaps are achieved.

We first compare the timing performance and the optimization performance for sparse linear regression objective function. We choose the (n,d) pairs as $(500,5000)$ and $(1000,10000)$ respectively, where n is the number of observation in the response vector $y \in \mathbb{R}^n$ and d is the dimension of the parameter vector $\theta \in \mathbb{R}^d$. For the design matrix $X \in \mathbb{R}^{n\times d}$, we generate each row independently from a d-dimensional normal distribution $N(\mathbf{0},\Sigma)$. For the well-conditioned case, we choose $\Sigma_{ij} = 0.25$ for $i \neq j$ and $\Sigma_{ii} = 1$. For the ill-conditioned case, we choose $\Sigma_{ij} = 0.75$ for $i \neq j$ and $\Sigma_{ii} = 1$. Then we have $y = X\theta + \varepsilon$, where θ has all 0 entries except that the first 20 entries are drawn from $[0,1]$ uniform distribution. We also compare the timing performance for sparse logistic regression. The generations of X and θ follow from the settings for sparse linear regression. The response vector y has independent Bernoulli$\left(\frac{\exp(X_{i*}^\top\theta)}{1+\exp(X_{i*}^\top\theta)}\right)$ entries.

From the summary in Table 1 and Table 2, we see that while achieving almost identical optimization objective function, `picasso` is as fast as `glmnet` and `ncvreg` for L1 regularization, and **out-performs `ncvreg` for SCAD/MCP regularization (especially for sparse logistic regression)**. Moreover, we remark that `picasso` performs stably for various settings and tuning parameters compared with `ncvreg`. Especially when the tuning parameters are relatively small (corresponding to denser estimators), `ncvreg` may converge very slow or fail to converge, which we did not show here. To avoid such a scenario, we choose the sequence of tuning parameters under the criteria that `ncvreg` attains the optimal performance in terms of the parameter estimation, while the estimators are not too dense so that it fails to converge.

We also compare square-root lasso solver `scalreg` and `flare`. The simulation is the same as the sparse linear regression setup but here we're minimizing the L1 penalized square root mean squared error (square-root lasso problem). From Table 3 we see that `picasso` **significantly outperforms the alternatives for square-root lasso problems**. `picasso` is the best square root lasso solver we have seen so far in R.

6 Conclusion

The `picasso` library demonstrates significantly improved computational and statistical performance over existing libraries such as `ncvreg` for nonconvex regularized sparse learning. Besides, `picasso` also shows improvement over the popular libraries such as `glmnet` for sparse linear re-

Table 1: Average timing performance (in seconds) and optimal objective function values with standard errors in the parentheses on sparse linear regression.

Sparse Linear Regression (Well-Conditioned)					
Method	Package	$n = 500, d = 5000$		$n = 1000, d = 10000$	
		Time	Obj. Value	Time	Obj.Value
ℓ_1 norm	`picasso`	0.176(0.068)	21.072	0.466(0.003)	24.030
	`glmnet`	0.190(0.082)	21.112	0.438(0.003)	24.013
	`ncvreg`	0.220(0.079)	21.113	0.548(0.006)	24.012
MCP	`picasso`	0.290(0.088)	17.676	0.470(0.020)	22.067
	`ncvreg`	0.342(0.015)	17.620	0.594(0.014)	22.066
SCAD	`picasso`	0.252(0.045)	20.641	0.650(0.008)	23.773
	`ncvreg`	0.302(0.071)	20.610	0.746(0.014)	23.809
Sparse Linear Regression (Ill-Conditioned)					
ℓ_1 norm	`picasso`	0.128(0.021)	29.655	0.492(0.011)	30.256
	`glmnet`	0.232(0.024)	29.658	0.900(0.017)	30.259
	`ncvreg`	0.188(0.031)	29.654	0.692(0.011)	30.257
MCP	`picasso`	0.064(0.008)	29.915	0.262(0.003)	30.348
	`ncvreg`	0.080(0.007)	30.609	0.272(0.040)	30.349
SCAD	`picasso`	0.124(0.015)	29.655	0.508(0.006)	30.256
	`ncvreg`	0.188(0.009)	29.654	0.680(0.034)	30.257

gression under ill-conditioned settings. Moreover, the algorithm implemented in `picasso`, which guarantees a global linear convergence to a unique sparse local optimum with optimal statistical properties. Overall, the `picasso` library has the potential to serve as a powerful toolbox for high dimensional sparse learning. We will continue to maintain and support this library.

Table 2: Average timing performance (in seconds) and optimal objective function values with standard errors in the parentheseson for solving square root lasso problem.

Sparse Logistic Regression (Well-Conditioned)					
Method	Package	$n = 500, d = 5000$		$n = 1000, d = 10000$	
		Time	Obj. Value	Time	Obj.Value
ℓ_1 norm	`picasso`	0.138(0.093)	0.346	0.324(0.006)	0.363
	`glmnet`	0.186(0.088)	0.346	0.600(0.011)	0.363
	`ncvreg`	0.168(0.051)	0.346	0.276(0.006)	0.363
MCP	`picasso`	0.098(0.093)	0.242	0.102(0.003)	0.215
	`ncvreg`	0.112(0.085)	0.292	0.126(0.003)	0.244
SCAD	`picasso`	0.100(0.079)	0.248	0.098(0.003)	0.221
	`ncvreg`	0.114(0.076)	0.314	0.162(0.003)	0.271
Sparse Logistic Regression (Ill-Conditioned)					
ℓ_1 norm	`picasso`	0.086(0.003)	0.325	0.438(0.037)	0.335
	`glmnet`	0.208(0.013)	0.325	1.236(0.153)	0.335
	`ncvreg`	0.156(0.031)	0.325	1.104(0.096)	0.335
MCP	`picasso`	0.026(0.003)	0.175	0.100(0.021)	0.170
	`ncvreg`	0.052(0.012)	0.222	0.264(0.006)	0.228
SCAD	`picasso`	0.028(0.009)	0.183	0.106(0.003)	0.181
	`ncvreg`	0.104(0.007)	0.253	1.028(0.019)	0.275

Table 3: Average timing performance (in seconds) and optimal objective function values with standard errors in the parentheses on square root lasso problem.

Square Root Lasso Problem (Well-Conditioned)				
Package	$n = 500, d = 5000$		$n = 1000, d = 10000$	
	Time	Obj. Value	Time	Obj.Value
`picasso`	0.368(0.045)	2.677	0.360(0.000)	4.454
`flare`	1.512(0.040)	3.336	5.324(0.062)	5.188
`scalreg`	1.680(0.034)	2.867	40.202(0.608)	4.492
Square Root Lasso Problem (Ill-Conditioned)				
`picasso`	0.040(0.002)	5.388	0.146(0.003)	5.495
`flare`	13.092(0.113)	5.979	297.356(2.772)	5.959
`scalreg`	3.354(0.427)	5.395	49.120(10.986)	5.507

References

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.

Breheny, P. (2013). ncvreg: Regularization paths for scad-and mcp-penalized regression models. *R package version* 2–6.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1** 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33** 1–13.

Li, X., Ge, J., Jiang, H., Wang, M., Hong, M. and Zhao, T. (2016). Boosting pathwise coordinate optimization in high dimensions: Sequential screening and proximal sub-sampled newton algorithm. Tech. rep., Georgia Tech.

Liu, H. and Wang, L. (2012). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. Tech. rep., Massachusett Institute of Technology.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D. et al. (2010). High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics* **38** 1287–1319.

Sun, T. and Zhang, C. (2012). Scaled sparse linear regression. *Biometrika* To appear.

Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research* **14** 3385–3418.

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

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. and Tibshirani, R. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 245–266.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.

Zhao, T., Liu, H. and Zhang, T. (2017). Pathwise coordinate optimization for nonconvex sparse learning: Algorithm and theory. *Annals of Statistics* .