

Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches

Mark Schmidt¹, Glenn Fung², Rómer Rosales²

¹ Department of Computer Science University of British Columbia,

² IKM, Siemens Medical Solutions, USA

1 Further Experimental Results

We looked at a generalized version of the **Gauss-Seidel**, **Shooting**, **Grafting**, **Sub-Gradient**, **epsL1**, **Log-Barrier**, **EM**, **Log(norm(w))**, **SmoothL1**, **SQP**, **ProjectionL1**, and **Interior Point** methods and applied them to a number of datasets. Although the general-L1 framework make no assumptions about convexity, we have restricted our experiments to convex functions.

All methods were run until the same convergence criteria was met (*i.e.*, that the step length between iterates, change in function value between iterates, negative directional derivative, or optimality condition was less than 10^{-6}). We assessed the ability of the methods to optimize a loss function known only through a ‘black box’ function that returns the objective value and derivatives for a given parameter setting. Convergence was measured based on function evaluations; this is, the number of times the algorithm invoked the ‘black box’ (to make the comparisons fair, all of the implementations were designed and tuned with this in mind). The iterates were truncated to 250 such evaluations, and methods whose final loss was greater than 10^{-3} times the minimum found across the methods were assigned the maximum value of 250 evaluations to punish for low accuracy. This was only needed in a small minority of cases, since all methods typically either found a high accuracy solution, or reach the maximum number of iterations. We used a second-order (Hessian-based Newton) strategy across all methods examined (Quasi-Newton methods are left to the discussion).

Datasets used for Probit Regression and Smooth SVM classifiers (from the UCI repository¹):

1. Wisconsin Breast Cancer
2. Australian Heart
3. Pima Diabetes
4. Australian Credit

¹<http://www.ics.uci.edu/~mllearn/MLRepository.html>

5. Sonar
6. Ionosphere
7. German
8. Bright
9. Dim
10. Adult
11. Census
12. 2Norm

Datasets used for Multinomial Logistic Regression classifiers (from the UCI repository and the Statlog project²):

1. Iris
2. Glass
3. Wine
4. Vowel
5. Vehicle
6. LED
7. Satellite
8. Waveform21
9. DNA
10. Waveform40
11. Shuttle

²<http://www.liacc.up.pt/ML/old/statlog/>

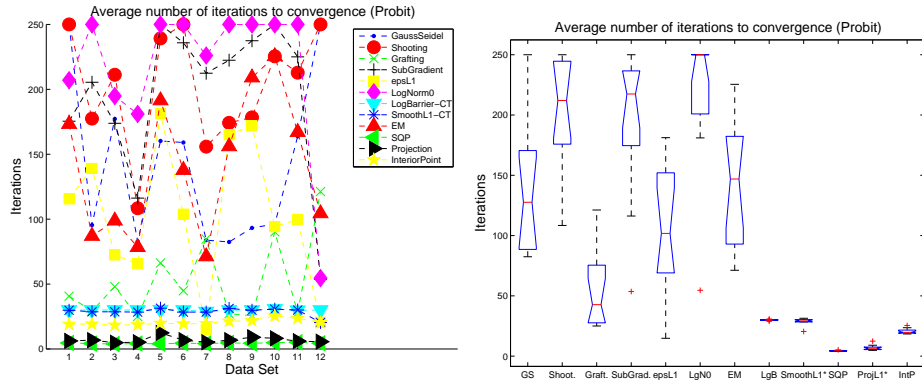


Figure 1: Distribution of function evaluations (averaged over λ) across 12 data sets to train a Probit Regression classifier with L1-regularization. Left: Detailed results for each dataset and method. Right: Summary results across all datasets (*=new methods).

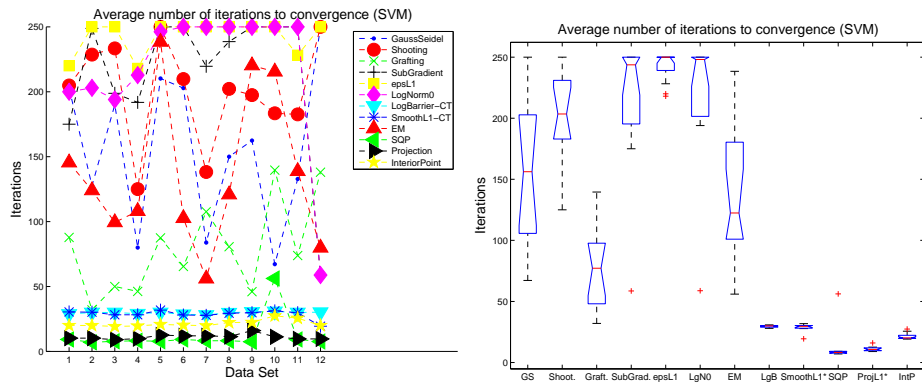


Figure 2: Distribution of function evaluations (averaged over λ) across 12 data sets to train a Smooth Support Vector Machine classifier with L1-regularization. Left: Detailed results for each dataset and method. Right: Summary results across all datasets (*=new methods).

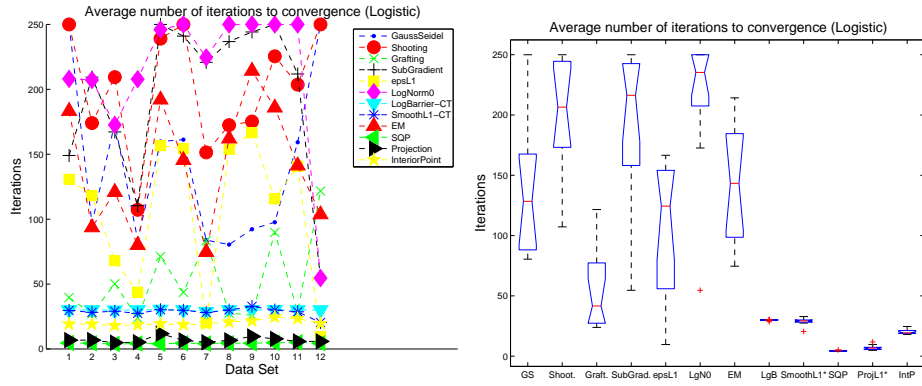


Figure 3: Distribution of function evaluations (averaged over λ) across 12 data sets to train a Logistic Regression classifier with L1-regularization. Left: Detailed results for each dataset and method. Right: Summary results across all datasets (*=new methods).

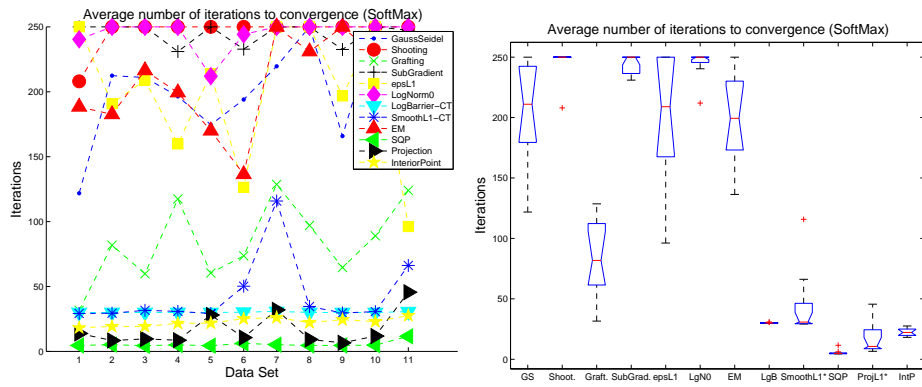


Figure 4: Distribution of function evaluations (averaged over λ) across 11 data sets to train a Multinomial Logistic Regression classifier with L1-regularization. Left: Detailed results for each dataset and method. Right: Summary results across all datasets (*=new methods).

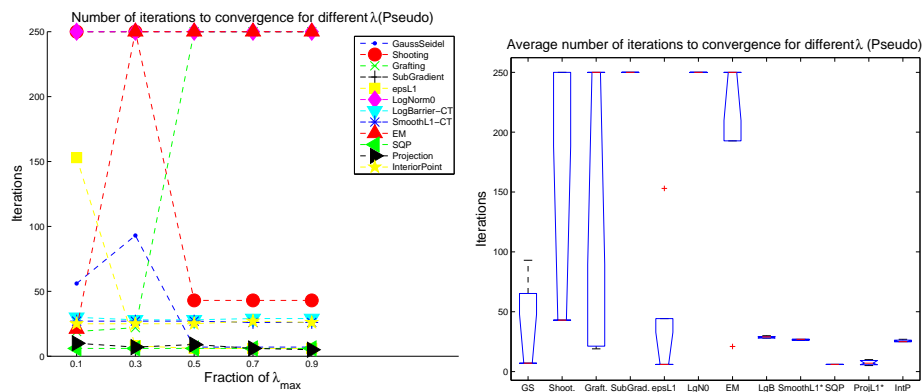


Figure 5: Left: Distribution of function evaluations on the image patch classification data set to train an L1-regularized 2D Conditional Random Field evaluated for various λ values. Left: Detailed Results for each λ and method. Right: Summary results across all λ values (*=new methods)

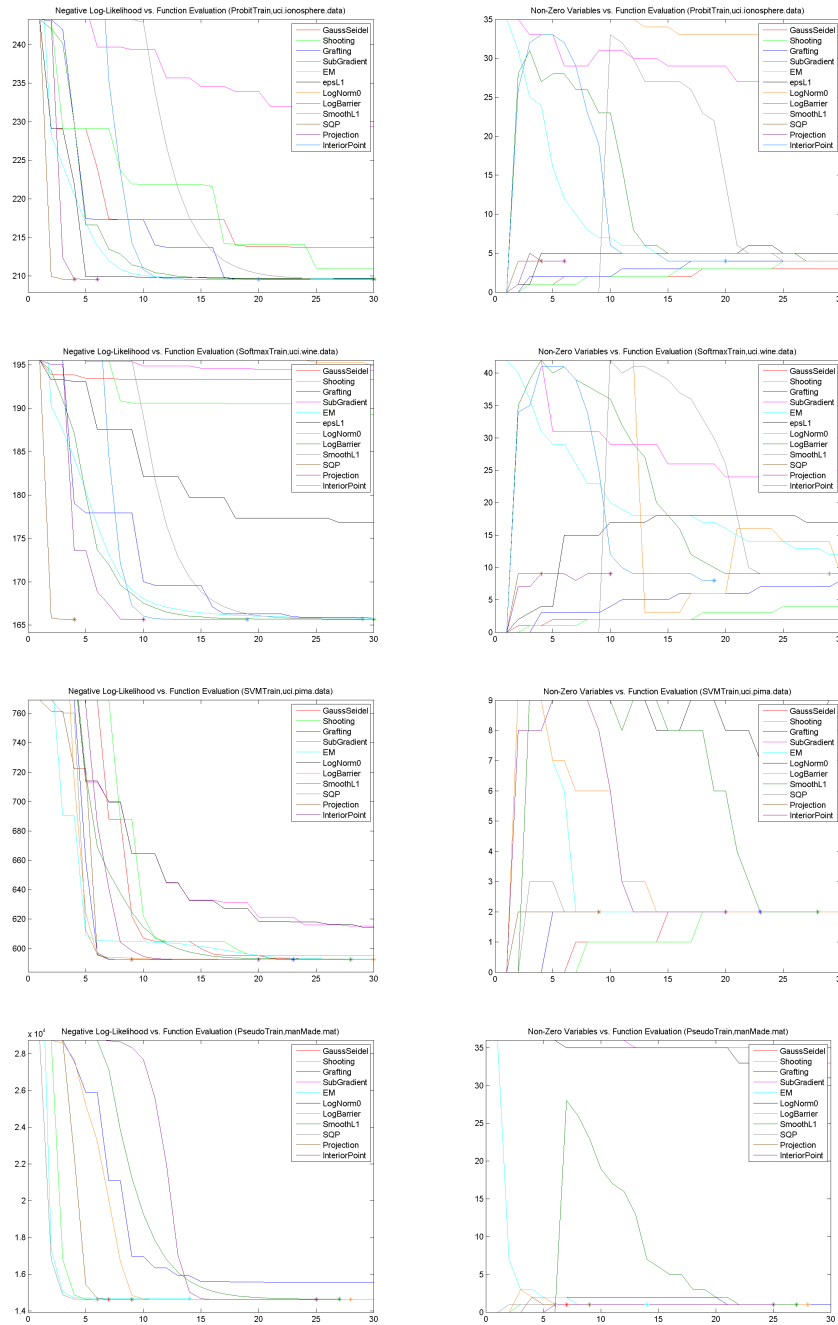


Figure 6: Loss Value (left) and Number of Non-Zero Coefficients (right) versus Function Evaluation for Training an L1-penalized Probit Regression classifier on the Ionosphere data, L1-penalized Softmax Regression on the Wine data, L1-penalized Support Vector Machine classifier on the Dim data, and L1-penalized 2D Conditional Random Field on the ManMade data. λ was always set to the mid-point of the regularization path. Stars indicate the termination point of the algorithm.