Fast Optimization Methods for L_1 Regularization: A Comparative Study and Two New Approaches

Mark Schmidt[1], Glenn Fung[2], Rómer Rosales[2]
[1]University of British Columbia, BC, Canada
[2]IKM CKS, Siemens Medical Solutions, PA, USA

SIEMENS

Overview

- Motivation
- Comparative Study
  - Subgradient strategies
  - Unconstrained approximations
  - Constrained formulations
- New Approaches
  - Differentiable convex approximation for L1 norm
  - Constrained optimization and two-metric projection
- Experimental Results
- Discussion

Motivation

- L_1 norm appears in various important machine learning problems
  - Finding optimal subset of features for a linear classifier is NP-hard (~L_0 norm).
    - # nonzero components of normal to hyper-plane classifier = # features it needs to employ.
  - Model selection in graphical models
    - MDL/BIC
    - AIC
  - L_1 norm is a reasonable convex approximation for the L_0 norm
  - Logarithmic sample complexity bounds [Ng04]

  (~number of data points relative to data dimensionality)

L_1 regularization. General problem

- We address optimization problems of the form:

$$\min_x f(x) \equiv L(x) + \lambda||x||_1$$

- $L(x)$: loss function (Logistic Regression, CRF,…)
- $\lambda||x||_1$: penalty on size of coefficients

- Properties of L_1-penalty:
  - Simultaneous *Regularization* and *Variable Selection* ☺
  - Logarithmic sample complexity with irrelevant variables ☺
  - Convex ☺
  - Non-differentiable ☹

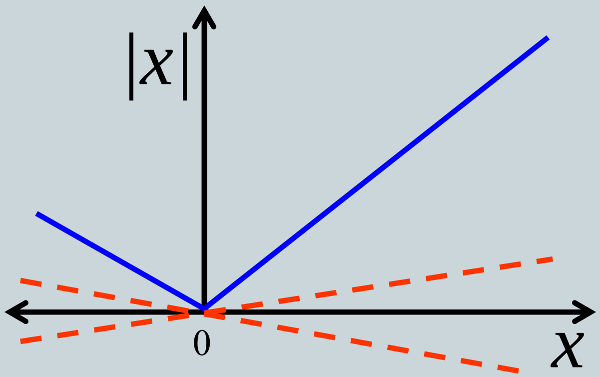
Overview of contributions

- Many approaches proposed to solve optimization problem for specific loss functions
- We consider the more general case where the loss function is continuous and twice-differentiable
- We give generalizations of some existing approaches, and outline 2 new approaches:
  - SmoothL1
  - ProjectionL1
- Our results indicate (consistently across datasets/loss functions)
  - Competitive with s-o-a (iterations for convergence)
  - Much more efficient (per iteration)

1 Subgradient strategies

Subgradients

- Let $f(x)$: $\mathcal{X} \rightarrow \mathbb{R}$ be a convex function
- Subderivative at point x: any real number c in $[a,b]$
  - a, b: one-sided derivatives at x
- Example $f(x)=|x|$



SIEMENS

Subgradient strategies for L1 regularization

- Gradient of $f(x)$: is not defined for $x = 0$
- First order optimality conditions at local minimizer $\bar{x}$

$$\begin{cases} \nabla_i L(\bar{x}) + \lambda \operatorname{sign}(\bar{x}_i) = 0, & |\bar{x}_i| > 0 \\ |\nabla_i L(\bar{x})| \le \lambda, & \bar{x}_i = 0 \end{cases}$$

- Define sub-gradient as follows:

$$\nabla_i f(x) = \begin{cases} \nabla_i L(x) + \lambda \operatorname{sign}(x_i), & |x_i| > 0 \\ \nabla_i L(x) + \lambda, & x_i = 0, \nabla_i L(x) < -\lambda \\ \nabla_i L(x) - \lambda, & x_i = 0, \nabla_i L(x) > \lambda \\ 0, & x_i = 0, -\lambda \le \nabla_i L(x) \le \lambda \end{cases}$$

Subgradient strategies

SIEMENS

- Solve iteratively: a few variables at a time
- Working set: variables that are free to change in iteration (optimization problem)
- Summary

Approach	{GaussSeidel} [Shevade-Keerthi03]	{Grafting} [Perkins03]	{Shooting} [Fu98]	Gen. {SubGrad}
Working set	y	y	n	y
Working (wk) set inclusion criteria	x_i w/largest sugradient (1)	x_i w/largest sugradient	n/a	x_i st satisf. optim. cond.
#var optim prob.	1	all in wk set	1 (cycle thru)	all in wk set
Step Method	1D line search	Newton	1D line search	Newton

Subgradient strategies

	Coordinate-wise	Joint
Incremental	Gauss-Seidel	Grafting
Full	Shooting	Gen SubGrad

SIEMENS

Subgradient Methods

- Main drawbacks of subgradient methods
  - Need special treatment for variables near zero
  - Does not guarantee to provide descent direction (some coordinates)
  - Optimize all variables jointly: slow convergence (does not guarantee descent direction)
  - Optimize coordinate-wise: inefficient
- Instability close to the singularity

2

Unconstrained approximations

Unconstrained approximations

- General idea: replace $f(x)$ with approximation $g(x)$ and solve <u>unconstrained</u> problem
  - $g(x)$ continuous and twice-differentiable
  - Solve (e.g., Newton iterations)
- {epsL1} [Lee et al.06] :

$$g(x) = L(x) + \lambda \sum_i \sqrt{x_i^2 + \epsilon}$$

- Log barrier functions:

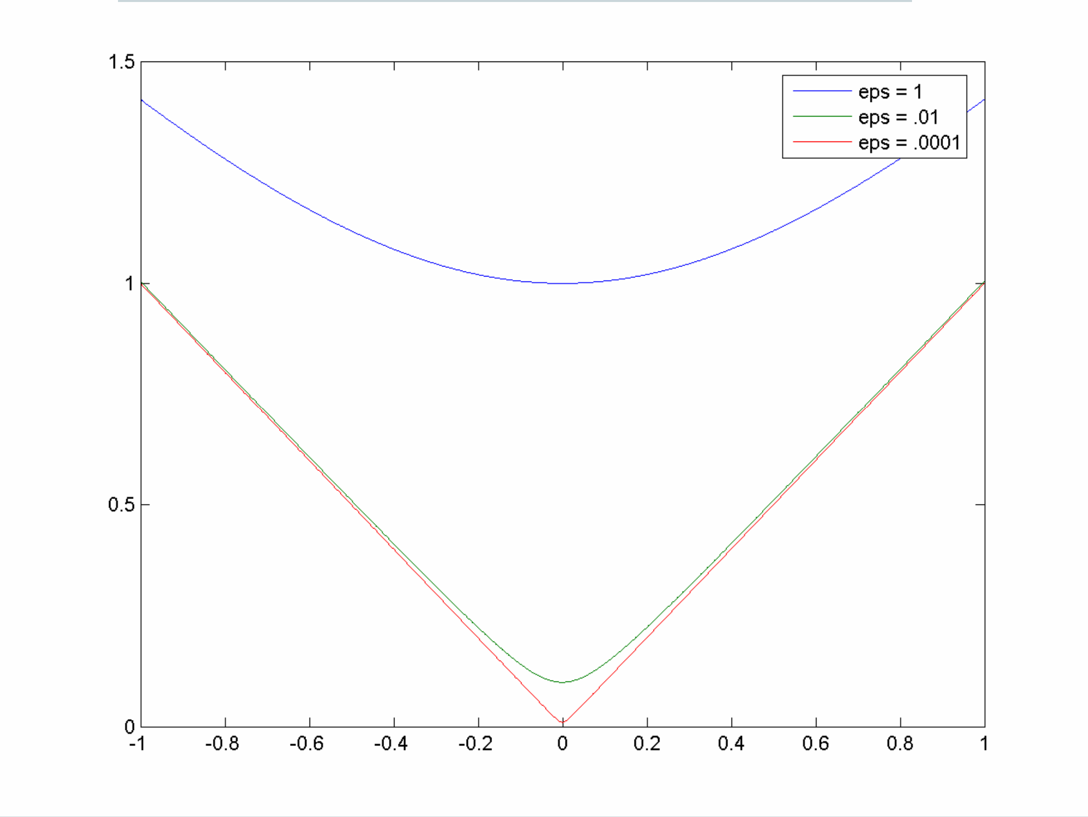
$$g(x) = L(x) + \lambda ||x||_1 - \mu \log c(x)$$

  - Example {LogNorm} $c(x) = ||x||_2^2$
- {SmoothL1*} …

Unconstrained approximations

- {epsL1}

$$g(x) = L(x) + \lambda \sum_i \sqrt{x_i^2 + \epsilon}$$



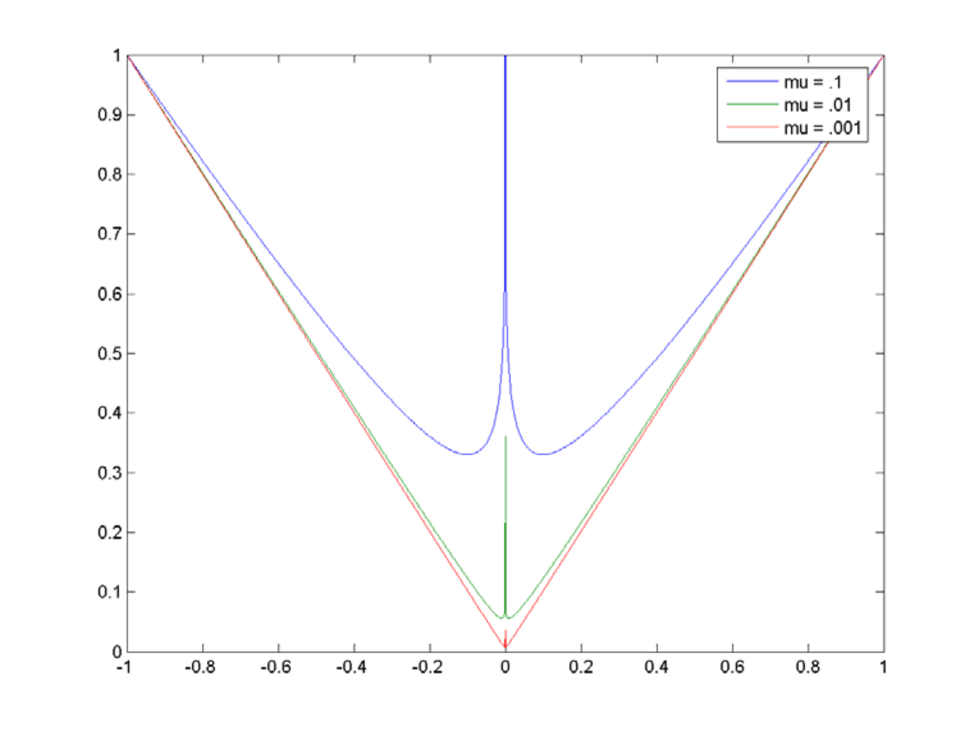
Unconstrained approximations

- Log barrier functions:

$$g(x) = L(x) + \lambda||x||_1 - \mu \log c(x)$$

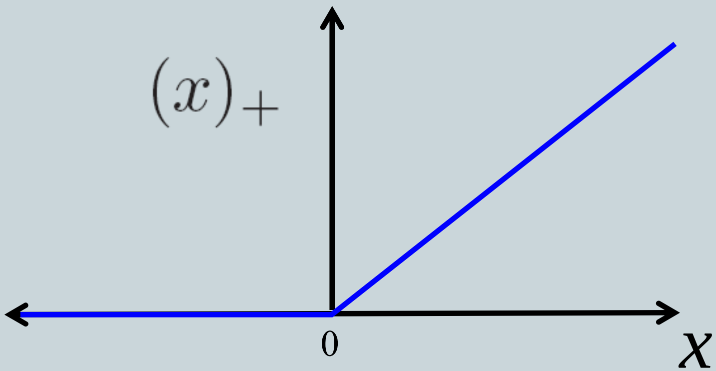
  - Example {LogNorm} $c(x) = ||x||_2^2$

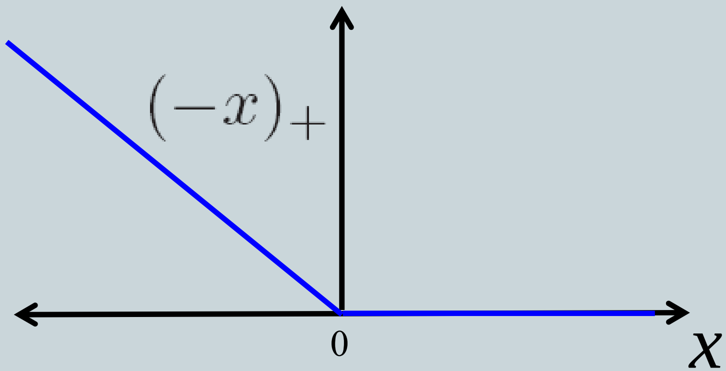
    (Smooth but infinite at 0)



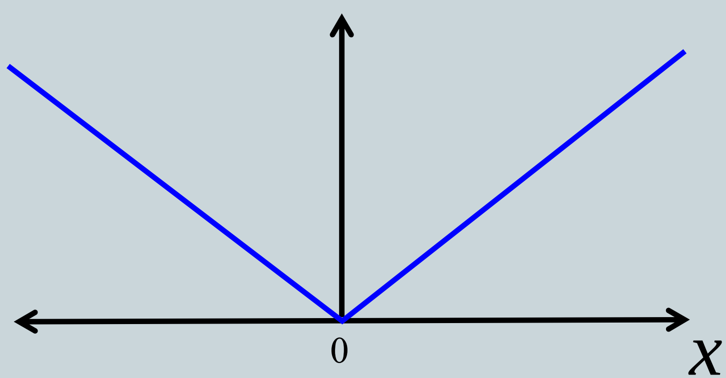
{SmoothL1*} approximation

- Define $(x)_+ = \max(x, 0)$



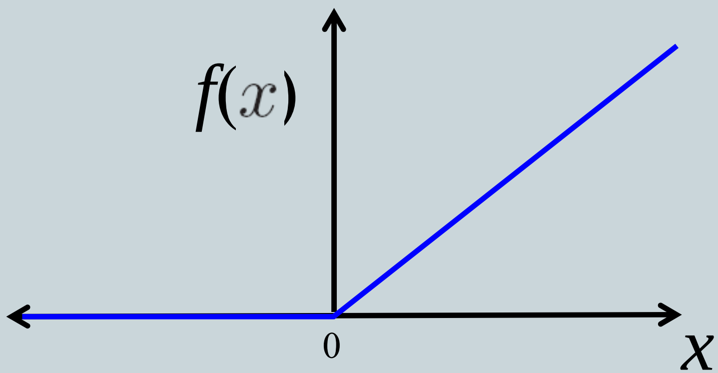


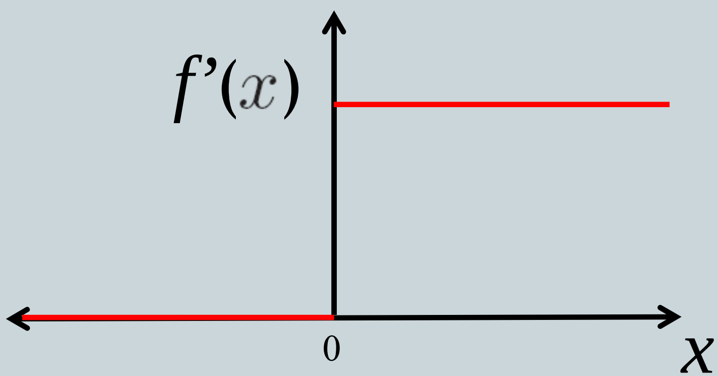
$$|x| = (x)_+ + (-x)_+$$



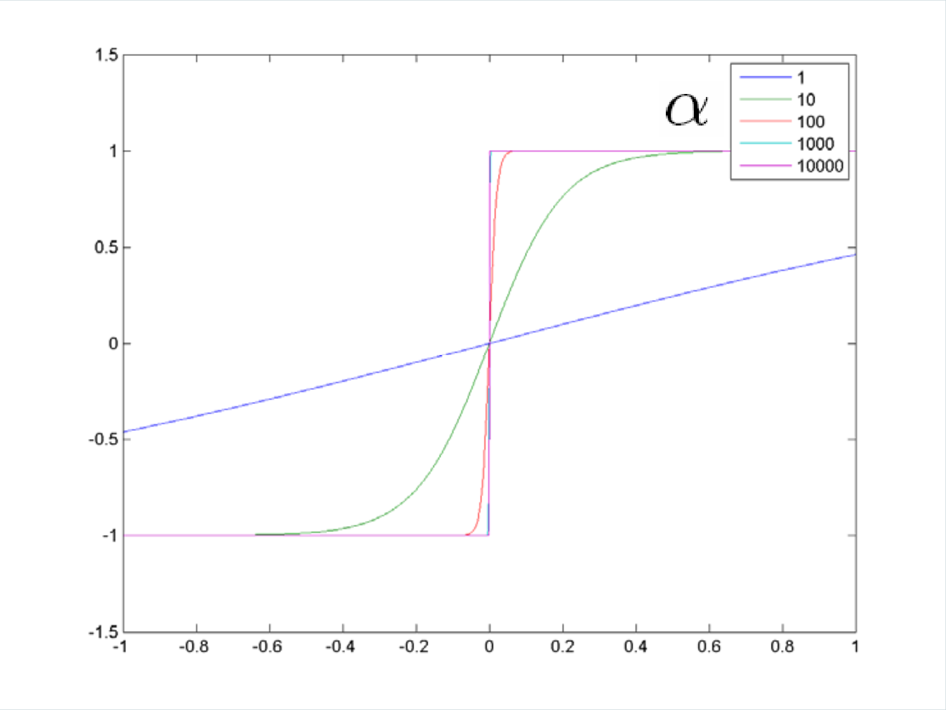
{SmoothL1*} approximation

- Define $f(x) = (x)_+ = \max(x, 0)$





Sigmoid function



x

{SmoothL1*} approximation

- Let $(x)_+ = \max(x, 0)$

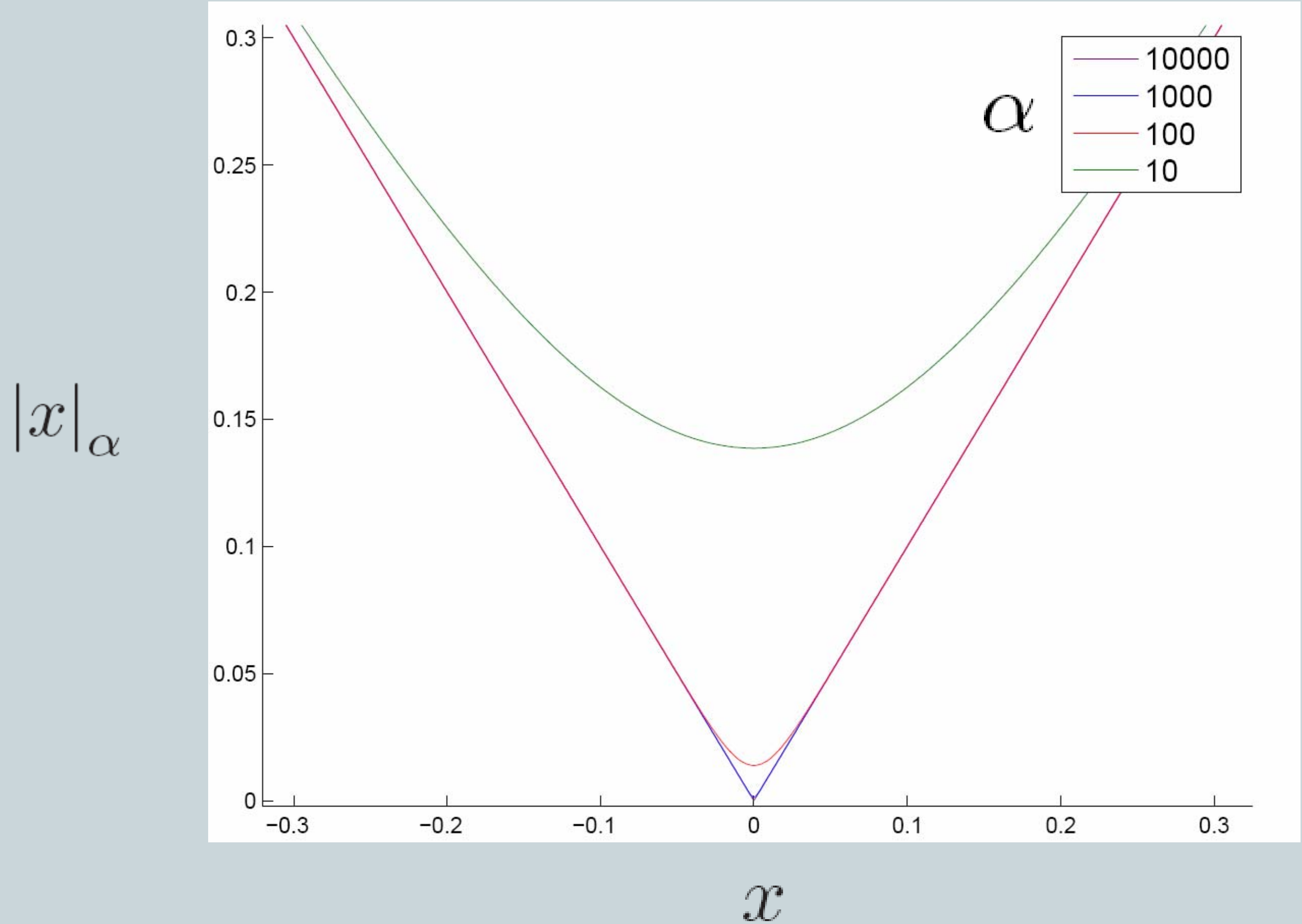
- Define $(x)_+ \approx p(x,\alpha) = x + \frac{1}{\alpha}\log(1+\exp(-\alpha x))$
  - ~Integral of sigmoid function [Chen-Mangasarian96]

- Basic observation: $|x| = (x)_+ + (-x)_+$

- Combining these:

$$\begin{aligned} |x| &= (x)_+ + (-x)_+ \approx p(x,\alpha) + p(-x,\alpha) \\ &= \frac{1}{\alpha}\left[\log(1+\exp(-\alpha x)) + \log(1+\exp(\alpha x))\right] \\ &\stackrel{\text{def}}{=} |x|_\alpha \end{aligned}$$

{SmoothL1*} approximation



{SmoothL1*} approximation

- Implementation details
  - Newton steps
  - Continuation strategy: increase $alpha$ between steps
    - (in practice $alpha = 1.5alpha$)
  - General line search methods can be used (advantage over log barrier)
  - No doubling of number of variables (as in most constrained formulations)

SIEMENS

Unconstrained approximations

- {EM}-based approach [Figuereido03]

$$x_i|\tau_i \sim N(0, \tau_i)$$

$$p(\tau_i|\sqrt{\lambda}) = \frac{\sqrt{\lambda}}{2}\exp(\frac{-\tau_i\sqrt{\lambda}}{2})$$

  - Integrating over τ_i yields the Laplacian prior over x_i
  - E-Step: compute posterior for τ_i
  - M-Step: update x, loss function is an expectation over τ of the L_2 norm (derived from conditional Gaussian prior)

3

Constrained approaches

SIEMENS

Constrained approaches

- Redefine as a constrained optimization problem

$$\min_x L(x) \quad s.t. ||x||_1 \leq t$$

  - Special case, L=logistic regression, can be solved by enforcing constraint in IRLS iterations using LARS [Efron et al.03] (Least Angle Regression).
  - Not possible in general for other L.
- Generalize by redefining IRLS-LARS as {SQP}
  - Solve a quadratic approximation of L, subject to linear constrains
  - Superlinear convergence

Constrained Formulations

- {SQP} formulation (generalizes IRLS-LARS)
- Let $x^+ = max(0, x) \;\; x^- = -min(0, x)$

$$\min_{x^+, x^-} L(x^+ - x^-) + \lambda \sum_i [x_i^+ + x_i^-] \quad s.t. \forall_i x_i^+ \geq 0, x_i^- \geq 0$$

- SQP formulation for calculating descent direction

$$\min_d \nabla(L(x^+ - x^-)^T + \lambda 1)^T d + \frac{1}{2} d^T \nabla^2 L(x^+ - x^-) d$$

$$s.t. \forall_i x_i^+ + d_i^+ \geq 0, x_i^- + d_i^- \geq 0$$

{ProjectionL1*}

SIEMENS

- Define problem as:

$$\min_{x^+,x^-} L(x^+ - x^-) + \lambda \sum_i [x_i^+ + x_i^-] \quad s.t. \underline{\forall_i x_i^+ \geq 0, x_i^- \geq 0}$$

- Observation: non-negative bound constraints.
  - Can be easily handled by Gradient Projection Method
  - Projection into constrain set

$$x^* := [x^* - t\nabla f(x^+ - x^-)]^+$$

  - Simple projection into non-negative orthant

{ProjectionL1*}

- Two-metric projection [Gafni-Bertsekas84]

$$x^* := [x^* - t\nabla^2 f(x^*)^{-1} \nabla f(x^*)]^+$$

  - Newton-like scaling
  - Superlinear convergence (like SQP)
  - Lower iteration cost than SQP
  - Not guarantee descent for arbitrary Hessian but guaranteed if optimize for x not in active set

SIEMENS

Active set

- Active set of constraints for non-negative bounds.

$$\{i | x_i^* = 0, \nabla L(x^+ - x^-) + \lambda > 0\}$$

- At each step, optimize wrt variables whose bound constraint is non-active

SIEMENS

Summary

Optimization Method	Approx Objective	Sub-Gradient	Explicit Constraints
Gauss-Seidel [16]	N	Y	N
Shooting [15]	N	Y	N
Grafting [6]	N	Y	N
Sub-Gradient	N	Y	N
epsL1 [11]	Y	N	N
Log(norm(x))	Y	N	N
EM [4]	Y*	Y***	N
Log-Barrier [14]	Y*	N	Y
SmoothL1 [ThisPaper]	Y*	N	N
SQP [11]	N	N	Y
ProjectionL1 [ThisPaper]	Y	Y***	Y
Interior Point [5]	Y**	N	Y

* Improve approximations between iterations

** Constrained objective improved over iterations

*** Correct gradient but only for ws

Experimental Results

- Methods compared
  - Gauss-Seidel
  - Shooting
  - Grafting
  - Sub-Gradient
  - EpsL1
  - Log Barrier
  - EM
  - Log-Norm
  - SmoothL1*
  - SQP
  - ProjectionL1*
  - Interior Point

SIEMENS

Experimental Results

- Stopping criteria
  - Step Length between iterations < 10^{-6}
  - Change in function value between iteration < 10^{-6}
  - Negative directional derivative < 10^{-6}
- Methods only know $f(x)$ through black box. For given x, $f(x)$ and derivatives
- Convergence measured based on number of black box invocations
- All methods typically found the optimal solution or reached max evaluations allowed (max=250)

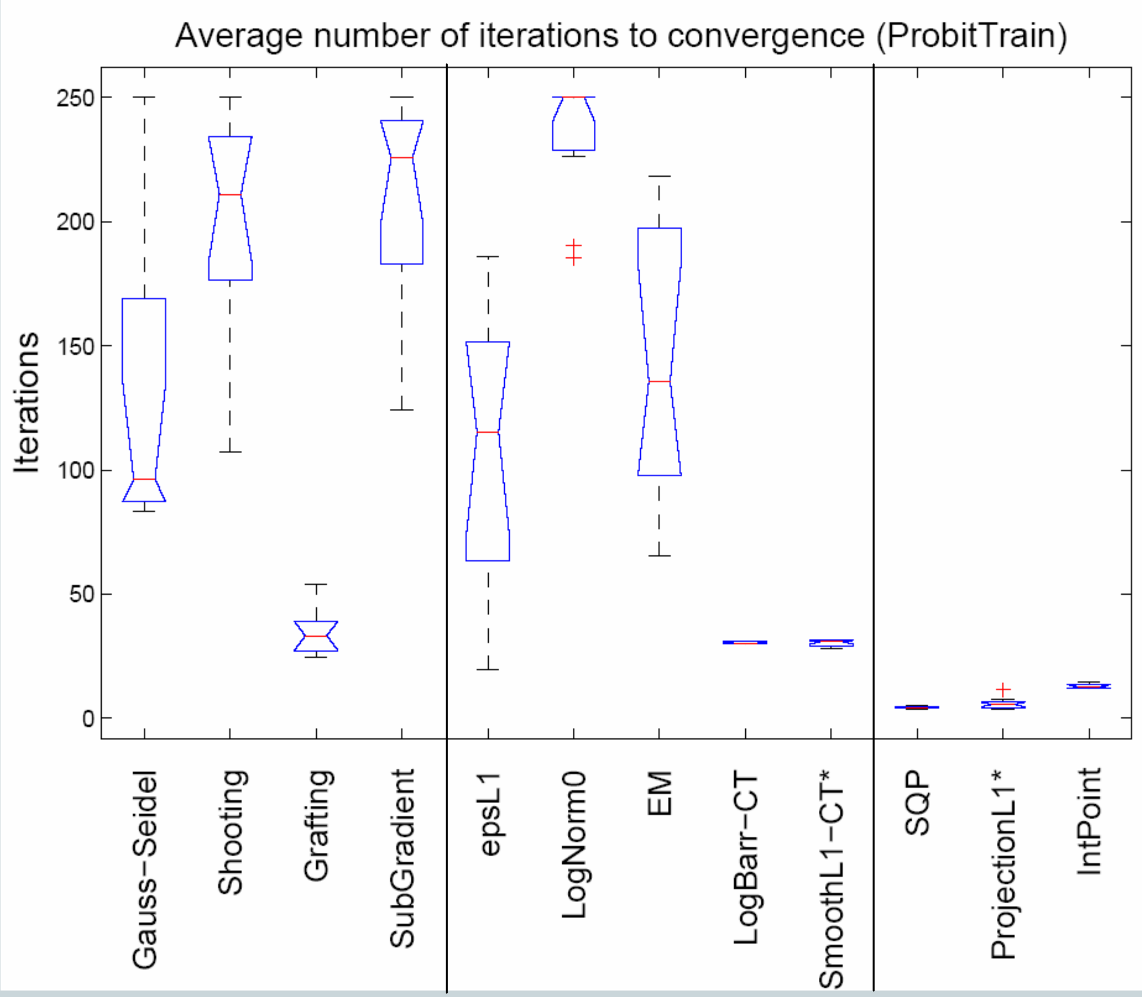
Experimental evaluation

- Binary Classification
  - Probit Regression $L(x) = \log(\phi(\frac{y_i x^T z_i}{\sqrt(2)}))$
  - Smooth SVM $L(x) = (1 - y_i x^T z_i)^+$

- Initialized with $x=0$ (or x=0.01)
- Define λ_{max} s.t. optimal solution: $x=0$
- Test for λ_{max} * $[.1, .3, .5, .7, .9]$
- 12 datasets UCI repository

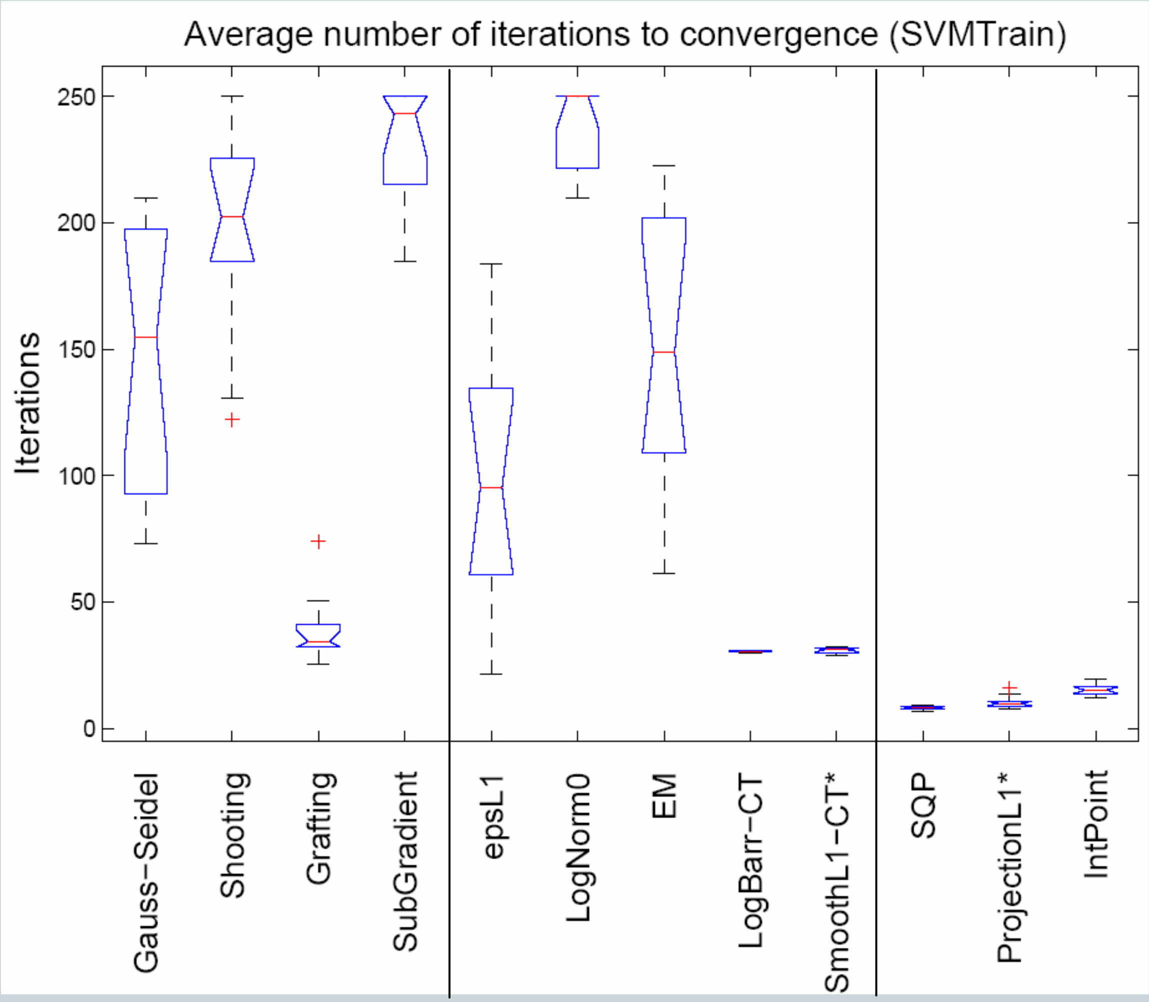
Experimental results (Probit)

SIEMENS



Distribution of function evaluations (averaged over λ) across data sets

Experimental results (~SVM)

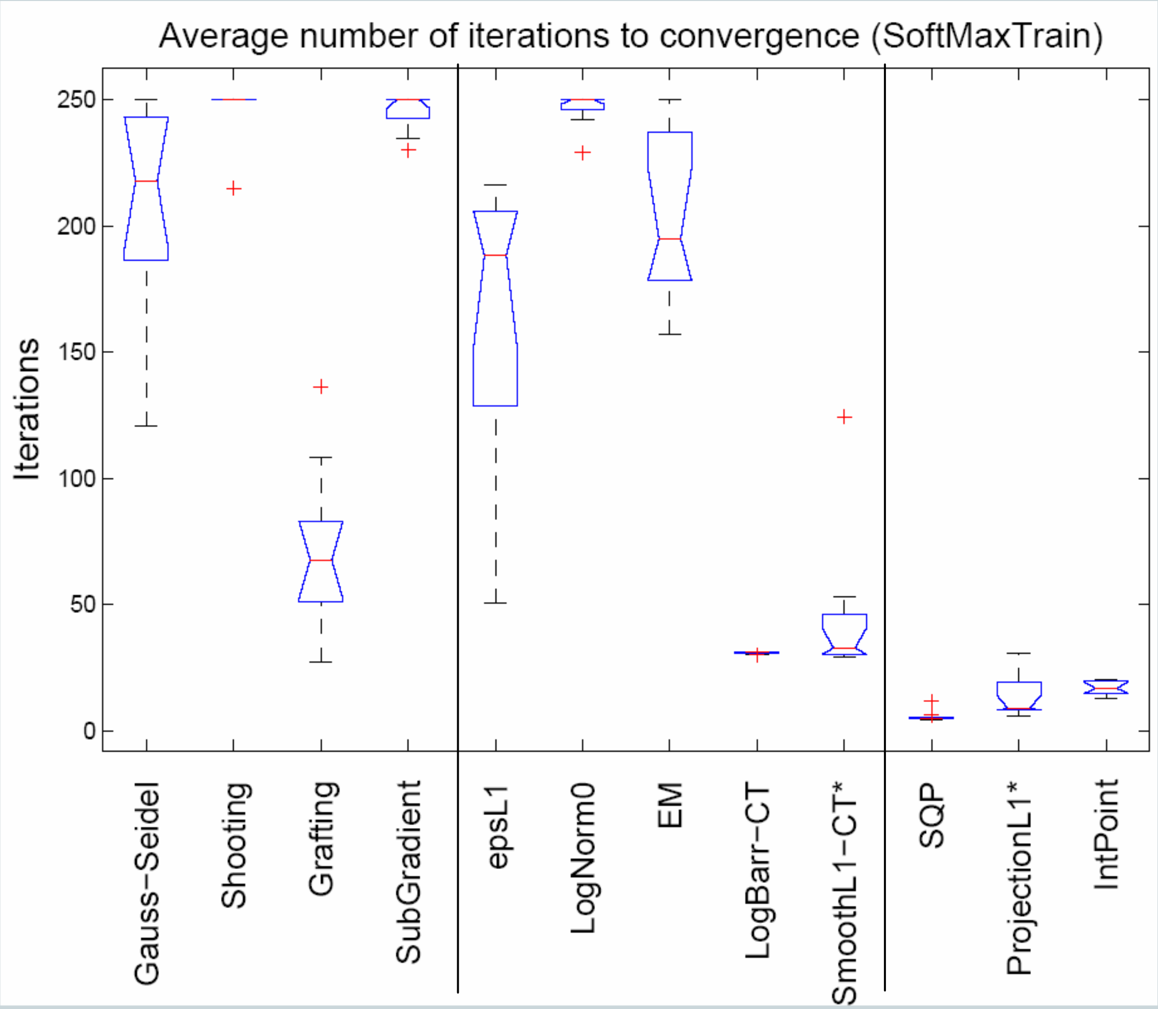


Distribution of function evaluations (averaged over λ) across data sets

Experimental evaluation

- Multinomial Classification
  - Multinomial Log Regression (Soft Max)
  - 11 Datasets UCI Repository + StatLog Project

Experimental results (SoftMax)



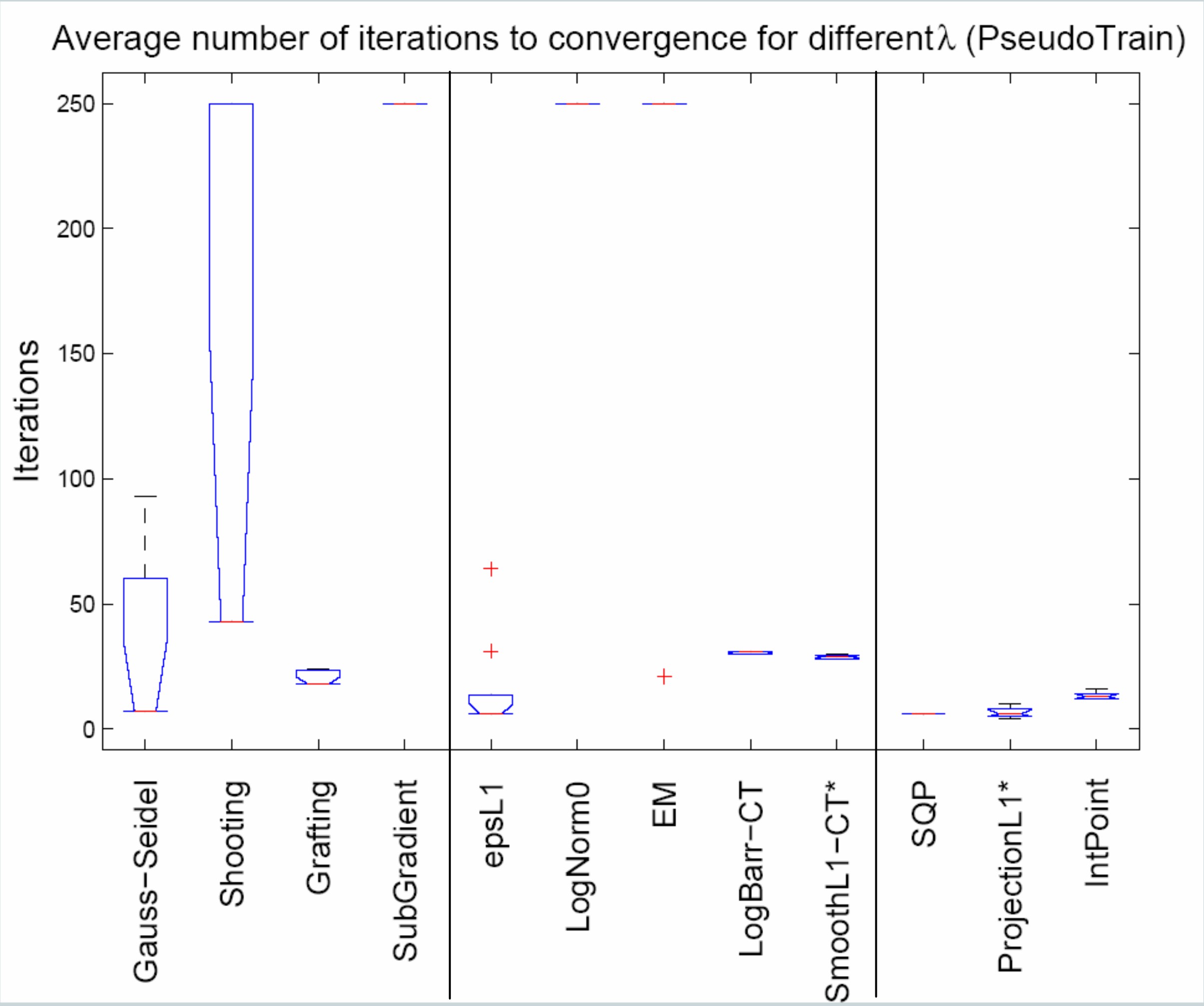
Experimental evaluation

- Structured Classification
  - CRF (2D). Pseudo-likelihood

$$l(x, v) = \log(1 + \exp(y_i x^T z_i + \sum_{j \in nei(i)} y_i y_j v^T z_{ij}))$$

  - Image Patch classification problem [Kumar-Hebert03]

Experimental results (CRF)



SIEMENS

Summary

Optimization Method	Approx Objective	Sub-Gradient	Explicit Constraints	Convergence Ranking	Iteration Speed Ranking
Gauss-Seidel [16]	N	Y	N	6	1
Shooting [15]	N	Y	N	8	1
Grafting [6]	N	Y	N	4	2
Sub-Gradient	N	Y	N	9	2
epsL1 [11]	Y	N	N	5	2
Log(norm(x))	Y	N	N	10	2
EM [4]	Y*	Y***	N	7	2
Log-Barrier [14]	Y*	N	Y	3	3
SmoothL1 [ThisPaper]	Y*	N	N	3	2
SQP [11]	N	N	Y	1	4
ProjectionL1 [ThisPaper]	Y	Y***	Y	1	3
Interior Point [5]	Y**	N	Y	2	3

* Improve approximations between iterations

** Constrained objective improved over iterations

*** Correct gradient but only for WS

SIEMENS

Summary

- Number of iterations
  - Best: {SQP}
  - Constrained approaches better than unconstrained approximations
- Iteration time
  - Best constrained approach {ProjectionL1*}
  - Constrained approaches highest iteration cost
  - {SmoothL1*} best unconstrained
- Overall run-time
  - Best {ProjectionL1*}
  - {SmoothL1*} may be best suitable for many variables
  - {SQP} best in problems with very expensive function evaluations (due to low #iterations)