Solving the Elastic Net using SVMs

Alberto Torres Barrán

Universidad Autónoma de Madrid



October 31, 2014

Outline

1. Introduction

2. Equivalence between Lasso and ν-SVM

3. Experiments

4. Conclusions and further work

Lasso and Elastic Net problems

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the data matrix and $\mathbf{y} \in \mathbb{R}^n$ the response vector, where n is the number of patterns and d the dimension.

Constrained Lasso (C-Lasso)

$$\min_{\mathbf{w}} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \quad \text{s.t} \quad \|\mathbf{w}\|_1 \leq t. \tag{1}$$

Unconstrained Lasso (U-Lasso)

$$\min_{\mathbf{w}} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1. \tag{2}$$

Elastic Net

$$\min_{\mathbf{w}} \ \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda_2 \|\mathbf{w}\|_2^2 + \lambda_1 \|\mathbf{w}\|_1. \tag{3}$$

.

Equivalence between C-Lasso and U-Lasso

- Let $\mathbf{w}^0$ be the OLS estimator for the Linear Regression problem and $t^0 = \|\mathbf{w}^0\|_1$.
- Given $t < t^0$ and $d \leq n$, problem (1) has an unique solution $\mathbf{w}_t^*$, that also solves problem (2) for

$$\lambda = \frac{(\mathbf{X}\mathbf{w}_t^* - \mathbf{y})^t \mathbf{X}\mathbf{w}_t^*}{t}.$$

- Given $\lambda > 0$, problems (2) and (1) share the same solution $\mathbf{w}_\lambda^*$ for $t = \|\mathbf{w}_\lambda^*\|_1$.
- In general, both problems are equivalent [1], but you can only go from one problem to the other **after** it has been solved.

Elastic Net reduction to Lasso-type problem [2]

- Given a dataset $(\mathbf{X}, \mathbf{y})$ and (λ_1, λ_2), define an artificial dataset $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$ by

$$\widetilde{\mathbf{X}} = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_d \end{pmatrix}, \qquad \widetilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_d \end{pmatrix}.$$

- Note that $\widetilde{\mathbf{X}}$ is a $(n+d) \times d$ matrix and $\widetilde{\mathbf{y}}$ is a $(n+d)$ vector.
- Problem (3) can be written as a Lasso-type problem

$$\min_{\widetilde{\mathbf{w}}} \ \|\widetilde{\mathbf{X}}\widetilde{\mathbf{w}} - \widetilde{\mathbf{y}}\|_2^2 + \gamma\|\widetilde{\mathbf{w}}\|_1,$$

where $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ and $\widetilde{\mathbf{w}} = \sqrt{(1+\lambda_2)}\mathbf{w}$.

μ-NPP and ν-SVC problems

Given the sample $S = \{(\mathbf{x}_i, y_i) : y_i = \pm 1,\ i = 1, \ldots, N\}$ we can define the two following problems

μ–Nearest Point Problem (μ–NPP)

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2} \left\| \sum_i \alpha_i y_i \mathbf{x}_i \right\|^2 \quad \text{s.t.} \quad 0 \leq \boldsymbol{\alpha} \leq \mu,\ \sum_{y_i=1} \alpha_i = \sum_{y_i=-1} \alpha_i = 1$$

ν-Support Vector Classification (ν-SVC)

$$\min_{\mathbf{w}, \boldsymbol{\xi}, b, \rho} \quad \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{N}\sum_i \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^t \mathbf{x}_i + b) \geq \rho - \xi_i, \quad \xi_i \geq 0,\ i = 1, \ldots, N, \quad \rho \geq 0$$

Equivalence between μ-NPP and ν-SVC

μ–Margin Nearest Point Problem (μ-Margin NPP)

$$\min_{\mathbf{w},\boldsymbol{\xi},\gamma,\eta} \quad \frac{1}{2}\|\mathbf{w}\|^2 - \gamma + \eta + \mu \sum_i \xi_i$$

$$\text{s.t.} \quad \mathbf{w}^t\mathbf{x}_i \geq \gamma - \xi_i,\ \forall i \in \mathcal{I}^+, \quad \mathbf{w}^t\mathbf{x}_i \leq \eta + \xi_i,\ \forall i \in \mathcal{I}^-,$$

$$\xi_i \geq 0,\ i = 1, \ldots, N.$$

- μ-NPP is the dual of μ-Margin NPP.
- There is a one-to-one relationship between μ-Margin NPP and ν-SVC, given the transformations

$$\nu = \frac{2}{\mu N}, \quad \rho^* = \frac{\gamma^* - \eta^*}{2}, \quad b^* = -\frac{\gamma^* + \eta^*}{2}$$

and scaling the coefficients by $\frac{1}{\mu N}$ [3].

Parametrizing the l_1-ball as the convex hull

- We can represent the l_1-ball by the $2d$-dimensional simplex of probability vectors

$$\Delta_{2d} = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{2d}) \ : \ 0 \leq \boldsymbol{\alpha} \leq 1, \sum_{i=1}^{2d} \alpha_i = 1\}.$$

- The trick is to use two non-negative variables to represent each real variable [4]

$$\mathbf{w} = (\mathbf{I}_d \mid -\mathbf{I}_d)\boldsymbol{\alpha} \ \text{ for } \ \boldsymbol{\alpha} \in \Delta_{2d},$$

where $(A \mid B)$ denotes the horizontal concatenation of matrix A and B.

Lasso $\preceq$ SVM [4]

- First we scale the problem by $1/t$

$$\min_{\mathbf{w}} \ \|\mathbf{X}\widetilde{\mathbf{w}} - \widetilde{\mathbf{y}}\|_2^2 \quad \text{s.t} \quad \|\widetilde{\mathbf{w}}\|_1 \leq 1. \tag{4}$$

  where $\widetilde{\mathbf{w}} = \frac{1}{t}\mathbf{w}$ and $\widetilde{\mathbf{y}} = \frac{1}{t}\mathbf{y}$.

- Let $\widetilde{\mathbf{X}} = (\mathbf{X} \mid -\mathbf{X})$, the Lasso problem (4) can be rewritten as

$$\min_{\boldsymbol{\alpha}} \ \|\widetilde{\mathbf{X}}\boldsymbol{\alpha} - \widetilde{\mathbf{y}}\|_2^2 \quad \text{s.t} \quad 0 \leq \boldsymbol{\alpha} \leq 1, \sum_{i=1}^{2d} \alpha_i = 1. \tag{5}$$

- $\widetilde{\mathbf{X}}\boldsymbol{\alpha}$ lies in the convex hull spanned by the columns of $\widetilde{\mathbf{X}}$,

$$\mathcal{S} = \{\widetilde{\mathbf{X}}_j, \ 1 \leq j \leq 2d\}.$$

Lasso $\preceq$ SVM (cont.)

- Solving (5) is equivalent to solving the μ-NPP between the convex hull $\mathcal{S}$ and the singleton $\{\widetilde{\mathbf{y}}\}$ with $\mu = 1$.

- Any ν-SVM solver can be used to solve the NPP problem, for example LibSVM.

- Given $\mathbf{X}$, $\mathbf{y}$ and t, the algorithm is the following:

  1. Create extended dataset $\mathbf{A} = (\mathbf{X} \mid -\mathbf{X} \mid \frac{\mathbf{y}}{t})^t$ and artificial labels $\mathbf{b} = (\mathbf{1}_{2d} \mid -1)^t$.

  2. Solve the ν-SVM problem with data matrix $\mathbf{A}$, output vector $\mathbf{b}$ and $\nu = \frac{2}{2d+1}$, obtaining dual coefficients $\boldsymbol{\alpha}^* \in \mathbb{R}^{2d+1}$.

  3. Discard α_{2d+1} and recover original weights $\mathbf{w}^* = (\mathbf{I}_d \mid -\mathbf{I}_d)\boldsymbol{\alpha}^* \times t$.

Datasets and experiments

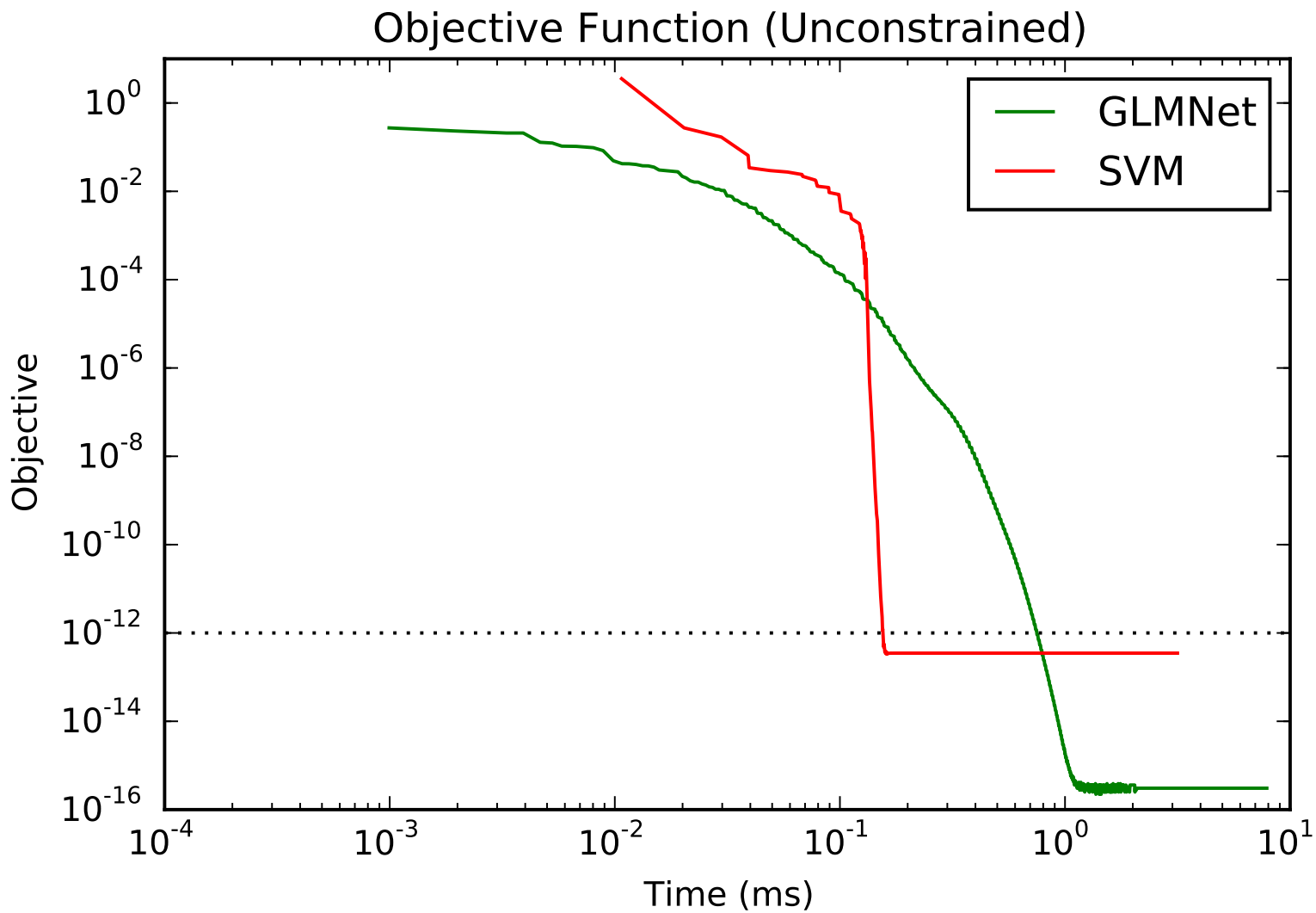
Dataset	Observations	Features
prostate	67	9
housing	378	13
cpusmall	6143	13
year	46215	91

- 4 datasets, 100 experiments
- 3 different λ values: "optimal" one selected by cross-validation λ^*, $2\lambda^*$ and $\lambda^*/2$.
- Solvers:
  - LIBSVM with support for dense data, SMO algorithm implementation in C [5].
  - Scikit-learn Cython implementation of GLMNet, coordinate descent algorithm [6].

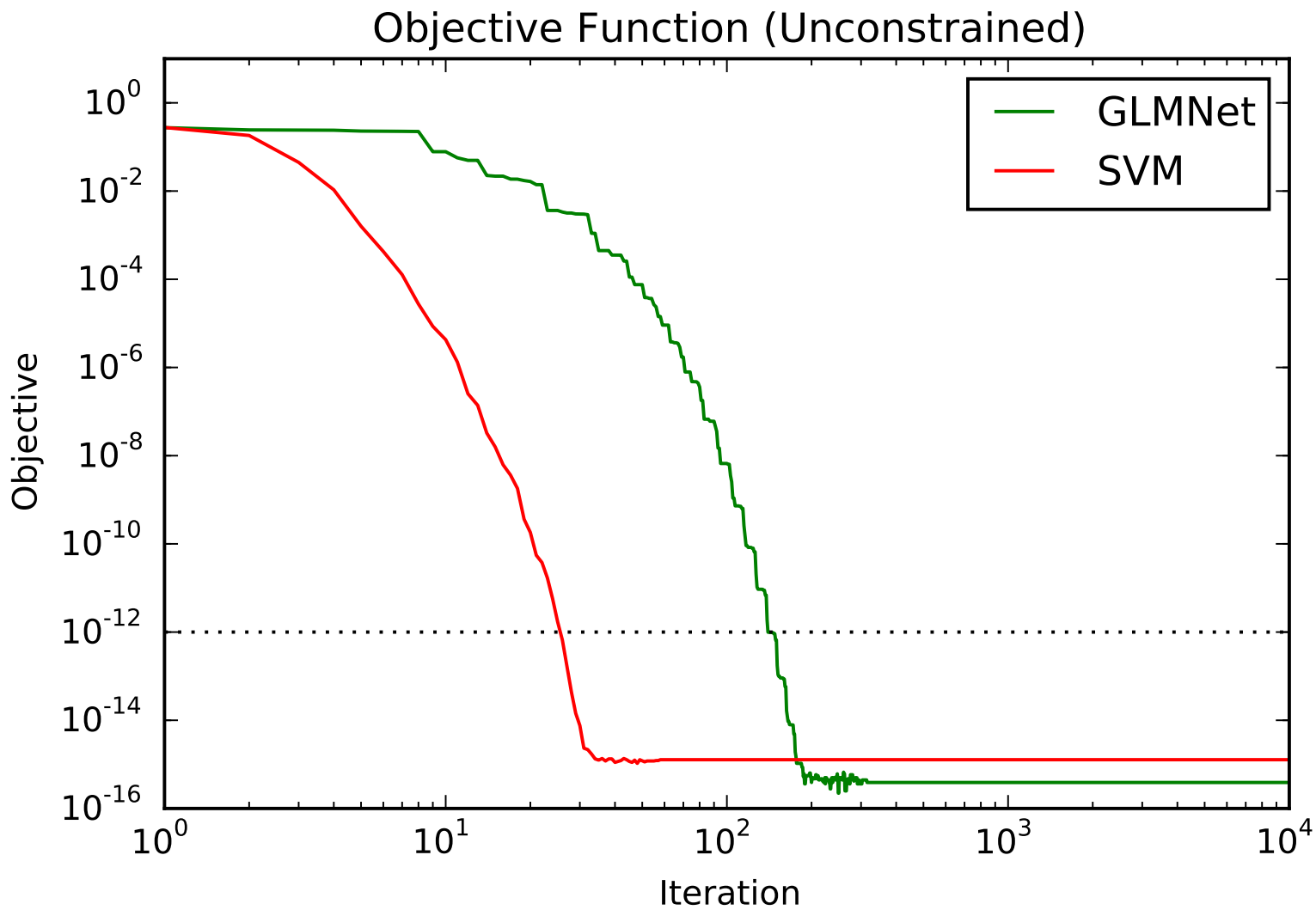
Results

		SVM		GLMNet	
Dataset	λ	Iter.	Time (ms)	Iter.	Time (ms)
prostate	0.001018	**86**	**0.035**	181	0.053
	0.002035	**75**	**0.033**	174	0.057
	0.004071	**52**	**0.024**	173	0.075
housing	0.004093	**144**	**0.166**	1010	0.832
	0.008186	**114**	**0.156**	933	0.749
	0.016371	**89**	**0.151**	666	0.520
cpusmall	0.016309	**42**	**1.319**	207	1.583
	0.032618	**36**	**1.168**	192	1.364
	0.065237	**27**	1.019	141	**0.880**
year	0.000227	**650**	1160.140	3606	**359.020**
	0.000453	**512**	1356.712	3528	**337.945**
	0.000907	**353**	1084.539	6123	**599.532**

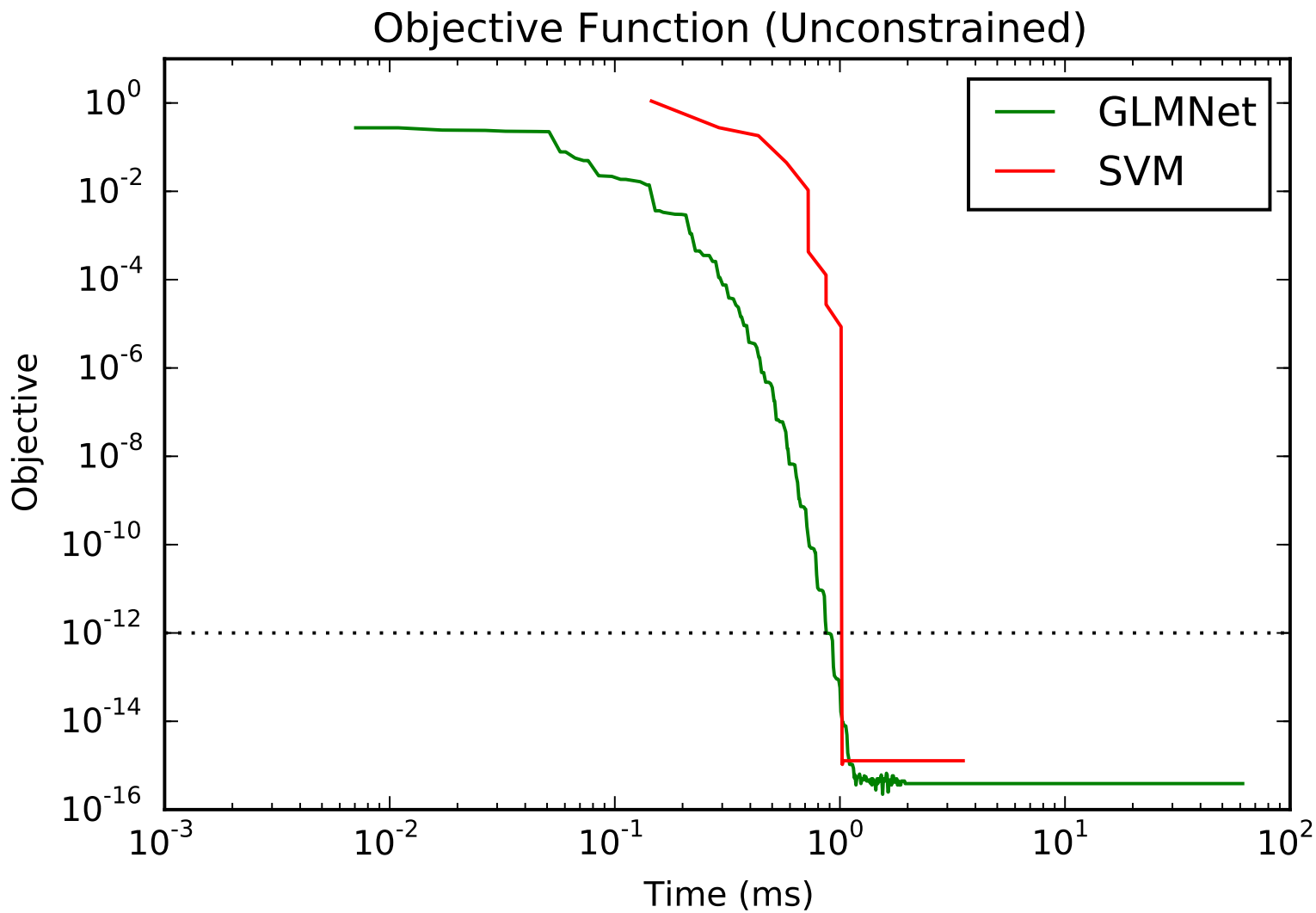
housing (time)



cpusmall (iterations)



cpusmall (time)



Conclusions and further work

- Reducing Lasso to ν-SVM opens the possibility of using fast SVM solvers for constrained Lasso.
- Some theoretical properties can also be translated between the two settings [4].
- Straight application of LibSVM is competitive with GLMNet.
- There is room for improvement modifying LibSVM for the Lasso reduction (idea: precompute the kernel).
- Alternative for large dimensional problems ($2d > n$): train the primal SVM problem instead of the dual.

References I

[1] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach.
On the lasso and its dual.
Journal of Computational and Graphical Statistics, 9:319–337, 1999.

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

[3] Jorge López and José R. Dorronsoro.
Linear Convergence Rate for the MDM Algorithm for the Nearest Point Problem.
Pattern Recognition, 2014.
(to appear).

References II

[4] Martin Jaggi.
An equivalence between the lasso and support vector machines.
CoRR, abs/1303.1152, 2013.

[5] Chih-Chung Chang and Chih-Jen Lin.
LIBSVM: A library for support vector machines.
ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.
Software available at
`http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
Regularization paths for generalized linear models via coordinate descent.
Journal of Statistical Software, 33(1):1–22, 2010.

Thank you for your attention.

Questions?

alberto.torres@uam.es