An Efficient Projection for $l_{1,\infty}$ Regularization

Ariadna Quattoni

Michael Collins

Xavier Carreras

Trevor Darrell

MIT CSAIL

Joint Sparsity

Goal :

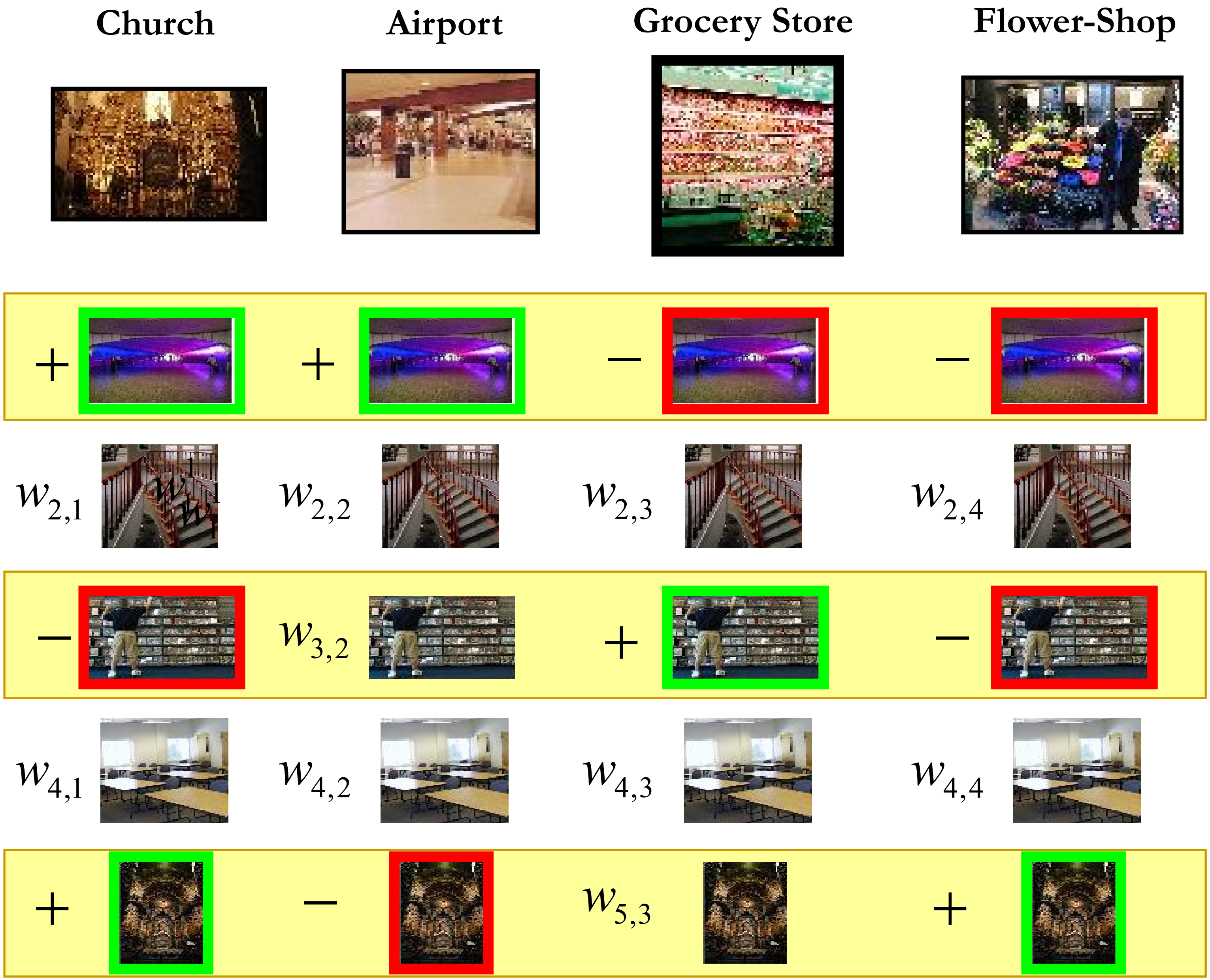
- Efficient training of jointly sparse models in high dimensional spaces.

Why? :

- Learn from fewer examples.
- Build more efficient classifiers.
- Interpretability.


Church
Airport
Grocery Store
Flower-Shop
$w_{1,1}$ $w_{1,2}$ $w_{1,3}$ $w_{1,4}$
$w_{2,1}$ $w_{2,2}$ $w_{2,3}$ $w_{2,4}$
$w_{3,1}$ $w_{3,2}$ $w_{3,3}$ $w_{3,4}$
$w_{4,1}$ $w_{4,2}$ $w_{4,3}$ $w_{4,4}$
$w_{5,1}$ $w_{5,2}$ $w_{5,3}$ $w_{5,4}$


Church
Airport
Grocery Store
Flower-Shop
$w_{1,1}$
$w_{1,2}$
$-$
$w_{1,4}$
$w_{2,1}$
$w_{2,2}$
$w_{2,3}$
$w_{2,4}$
$w_{3,1}$
$w_{3,2}$
$+$
$w_{3,4}$
$w_{4,1}$
$w_{4,2}$
$w_{4,3}$
$w_{4,4}$
$w_{5,1}$
$w_{5,2}$
$w_{5,3}$
$w_{5,4}$


Church
Airport
Grocery Store
Flower-Shop
+ + − −
$w_{2,1}$ $w_{2,2}$ $w_{2,3}$ $w_{2,4}$
− $w_{3,2}$ + −
$w_{4,1}$ $w_{4,2}$ $w_{4,3}$ $w_{4,4}$
+ − $w_{5,3}$ +

$l_{1,\infty}$ Regularization

- How do we promote joint (i.e. row) sparsity ?

$$W = \begin{pmatrix} W_{1,1} & W_{1,2} & \cdots & W_{1,m} \\ W_{2,1} & W_{2,2} & \cdots & W_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ W_{d,1} & W_{d,2} & \cdots & W_{d,m} \end{pmatrix}$$

Coefficients for feature 2

Coefficients for task 2

$$\boxed{||W||_{1,\infty} = \sum_{i=1}^{d} \max_{k}(|W_{i,k}|)}$$

The l_∞ norm on each row promotes non-sparsity on each row. → Share parameters

An l_1 norm on the maximum absolute values of the coefficients across tasks promotes sparsity. → Use few features

Contributions

- An efficient projected gradient method for $\mathbf{l}_{1,\infty}$ regularization

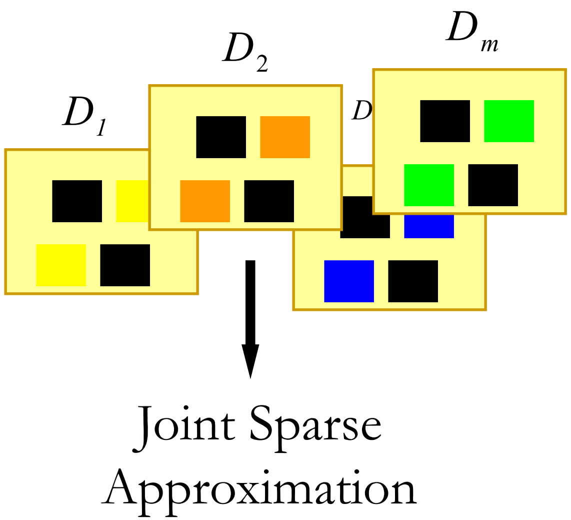
- Our projection works on O(n log n) time, same cost as $\mathbf{l}_1$ projection

- Experiments in Multitask image classification problems

- We can discover jointly sparse solutions

- $\mathbf{l}_{1,\infty}$ regularization leads to better performance than $\mathbf{l}_2$ and $\mathbf{l}_1$ regularization

Multitask Application



Collection of Tasks

$$\mathbf{D} = \{D_1, D_2, \ldots, D_m\}$$

$$\mathrm{D}_k = \{(x_1^k, y_1^k), \ldots, (x_{n_k}^k, y_{n_k}^k)\}$$

$$\mathbf{x} \in \mathbb{R}^d \quad y \in \{+1, -1\}$$

$$\arg\min_W \sum_{i=1}^{m} \frac{1}{|D_k|} \sum_{(x,y)\in D_k} L(f_k(x), y) + Q \sum_{i=1}^{d} \max_k(|W_{i,k}|)$$

$l_{1,\infty}$ Regularization: Constrained Convex Optimization Formulation

$$\boxed{\arg\min_{W} \sum_{i=1}^{m} \frac{1}{|D_k\|} \sum_{(x,y)\in D_k} L(f_k(x), y)}$$

A convex function

$$\boxed{s.t. \quad \sum_{i=1}^{d} \max_{k}(|W_{i,k}|) \leq C}$$

Convex constraints

- We use a Projected SubGradient method.

Main advantages: simple, scalable, guaranteed convergence rates.

- Projected SubGradient methods have been recently proposed:
  - l_2 regularization, i.e. SVM [Shalev-Shwartz et al. 2007]
  - l_1 regularization [Duchi et al. 2008]

Euclidean Projection into the $l_{1\text{-}\infty}$ ball

$$
\begin{aligned}
\mathbf{P}_{\mathbf{1},\infty}: \quad & \min_{B,\boldsymbol{\mu}} && \frac{1}{2}\sum_{i,j}(B_{i,j}-A_{i,j})^2 \\
& \text{s.t.} && \forall i,j \;\; B_{i,j} \le \mu_i \\
& && \sum_i \mu_i = C \\
& && \forall i,j \;\; B_{i,j} \ge 0 \\
& && \forall i \;\; \mu_i \ge 0
\end{aligned}
$$

Characterization of the solution

Let $\boldsymbol{\mu}$ *be the optimal maximums of problem* $\mathrm{P}_{1,\infty}$.
The optimal matrix B *of* $\mathrm{P}_{1,\infty}$ *satisfies that:*

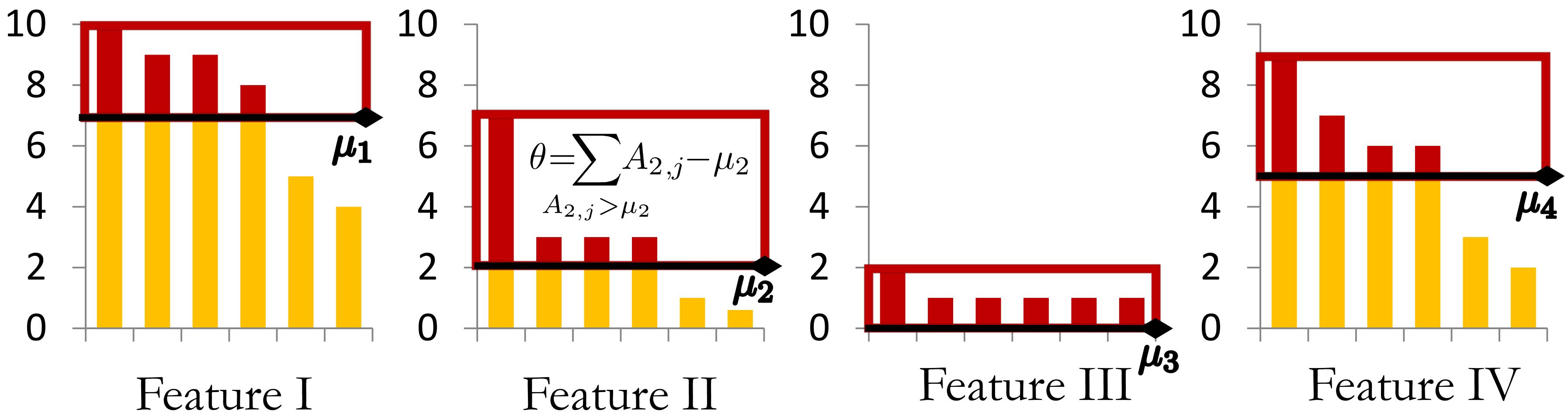
$$\begin{aligned} A_{i,j} \geq \mu_i \quad &\Rightarrow \quad B_{i,j} = \mu_i \\ A_{i,j} \leq \mu_i \quad &\Rightarrow \quad B_{i,j} = A_{i,j} \\ \mu_i = 0 \quad &\Rightarrow \quad B_{i,j} = 0 \end{aligned}$$

Characterization of the solution

At the optimal solution of $\mathrm{P}_{1,\infty}$ *there exists a constant* $\theta \geq 0$ *such that for every* i *either:*

$$\mu_i > 0 \quad \text{and} \quad \sum_j (A_{i,j} - B_{i,j}) = \theta$$

$$\mu_i = 0 \quad \text{and} \quad \sum_j A_{i,j} \leq \theta$$



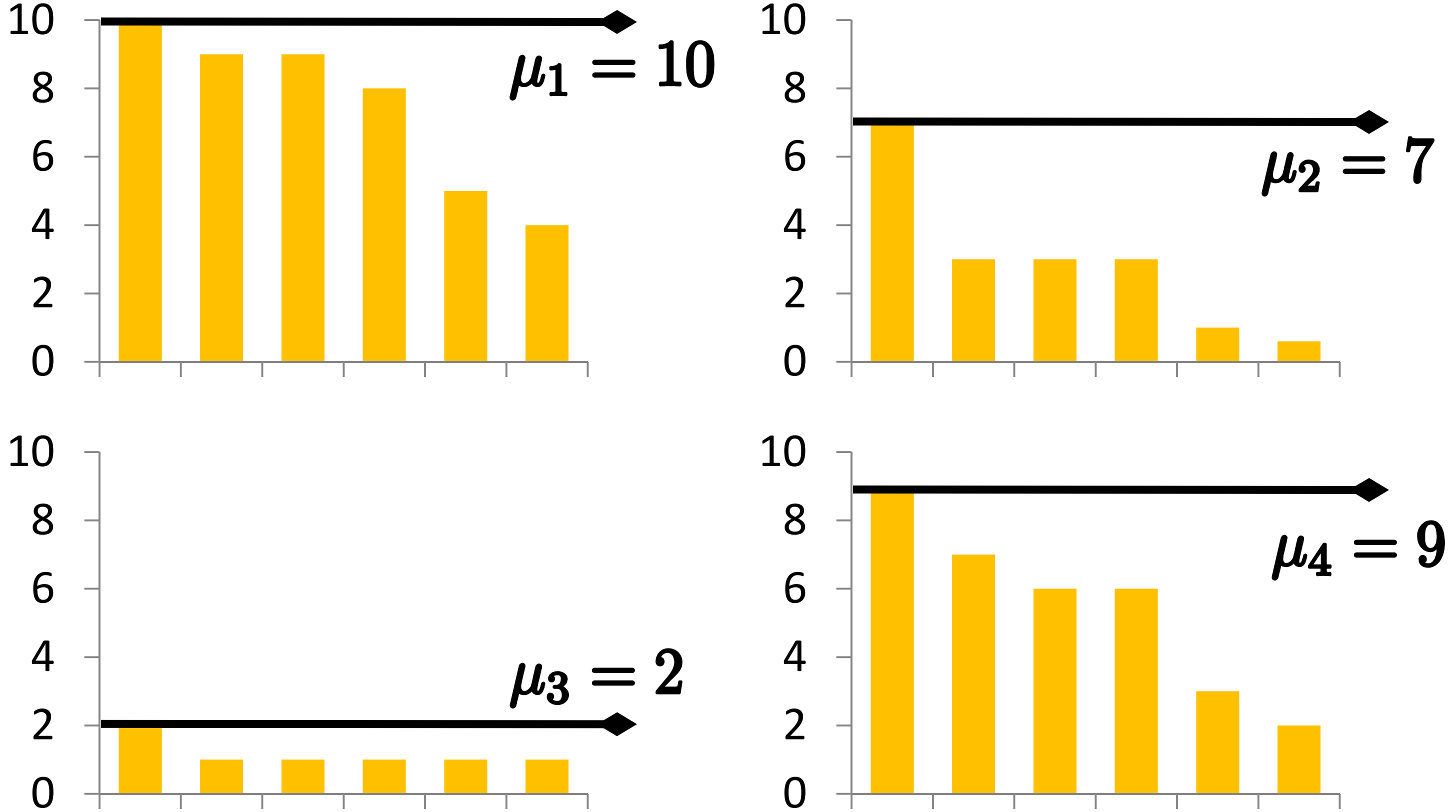
Mapping to a simpler problem

- We can map the projection problem to the following problem which finds the optimal maximums **μ**:

$$
\begin{aligned}
\mathbf{M_{1,\infty}}: \quad & \text{find} \quad \boldsymbol{\mu}\ ,\ \theta \\
& \text{s.t.} \quad \sum_i \mu_i = C \\
& \sum_{j:A_{i,j} \geq \mu_i} (A_{i,j} - \mu_i) = \theta\ ,\ \forall i \text{ s.t. } \mu_i > 0 \\
& \sum_j A_{i,j} \leq \theta\ ,\ \forall i \text{ s.t. } \mu_i = 0 \\
& \forall i \ \ \mu_i \geq 0 \ ; \ \ \theta \geq 0
\end{aligned}
$$

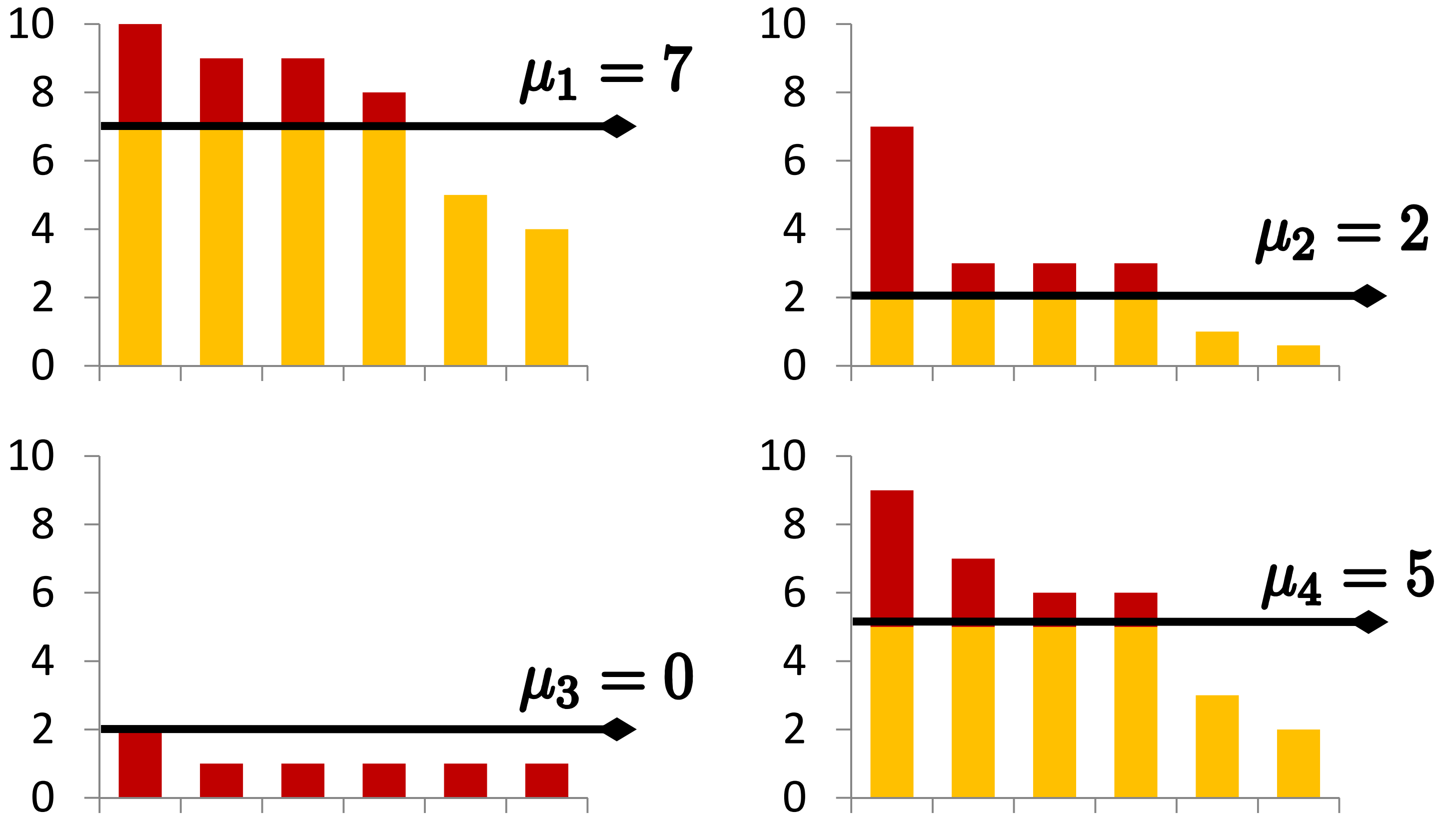
For any matrix A and a constant C such that $C < ||A||_{1,\infty}$, there is a unique solution $\boldsymbol{\mu}^*, \theta^*$ to the problem $\mathrm{M}_{1,\infty}$.

Efficient Algorithm



$$||A||_{1,\infty} = 28 \qquad \theta = 0$$

Efficient Algorithm



$$||A||_{1,\infty} = 14 \qquad \theta = 8$$

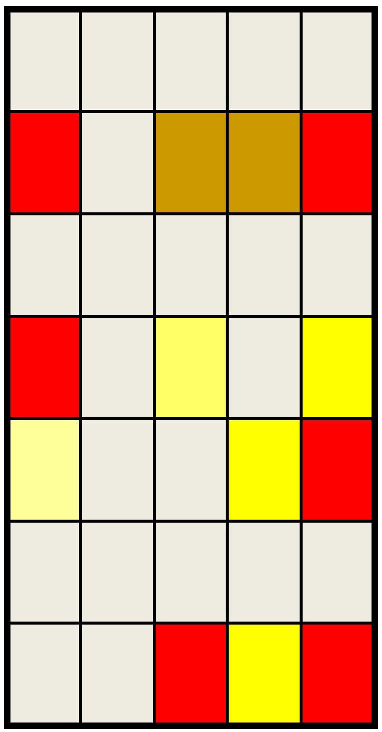
Complexity

- The total cost of the algorithm is dominated by sorting the entries of $\mathbf{A}$.

- The total cost is in the order of: $O(dm\log(dm))$

Synthetic Experiments

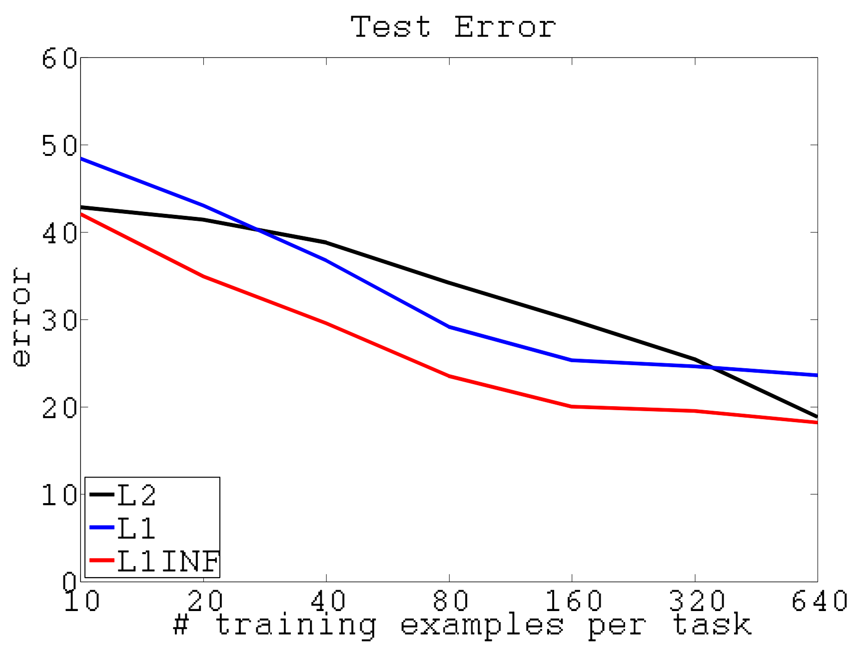
- Generate a jointly sparse parameter matrix $\mathbf{W}$:

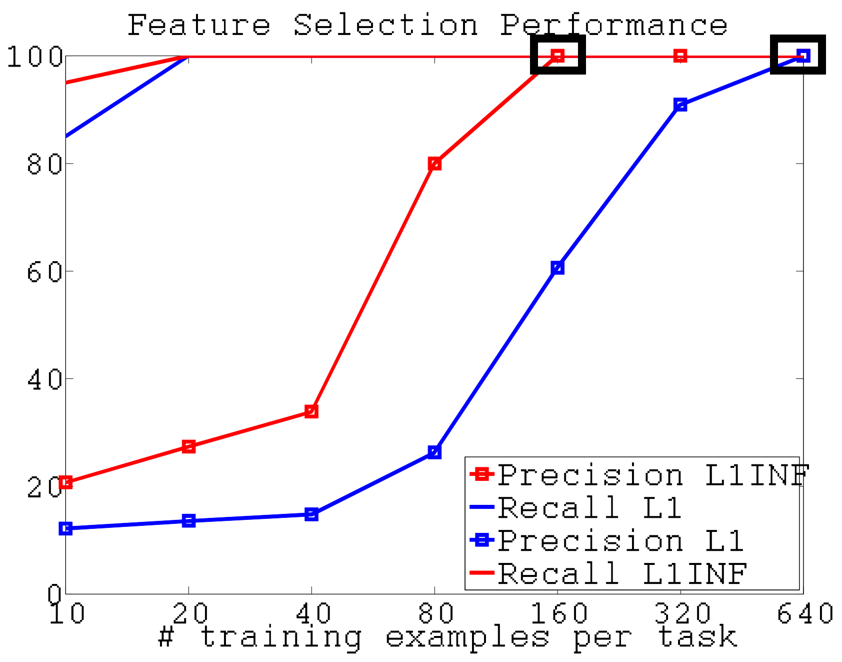


- For every task we generate pairs: (x_i^k, y_i^k)
  where: $y_i^k = \text{sign}(w_k^t x_i^k)$

- We compared three different types of regularization :
  - $l_{1,\infty}$ projection
  - l_1 projection
  - l_2 projection

Synthetic Experiments

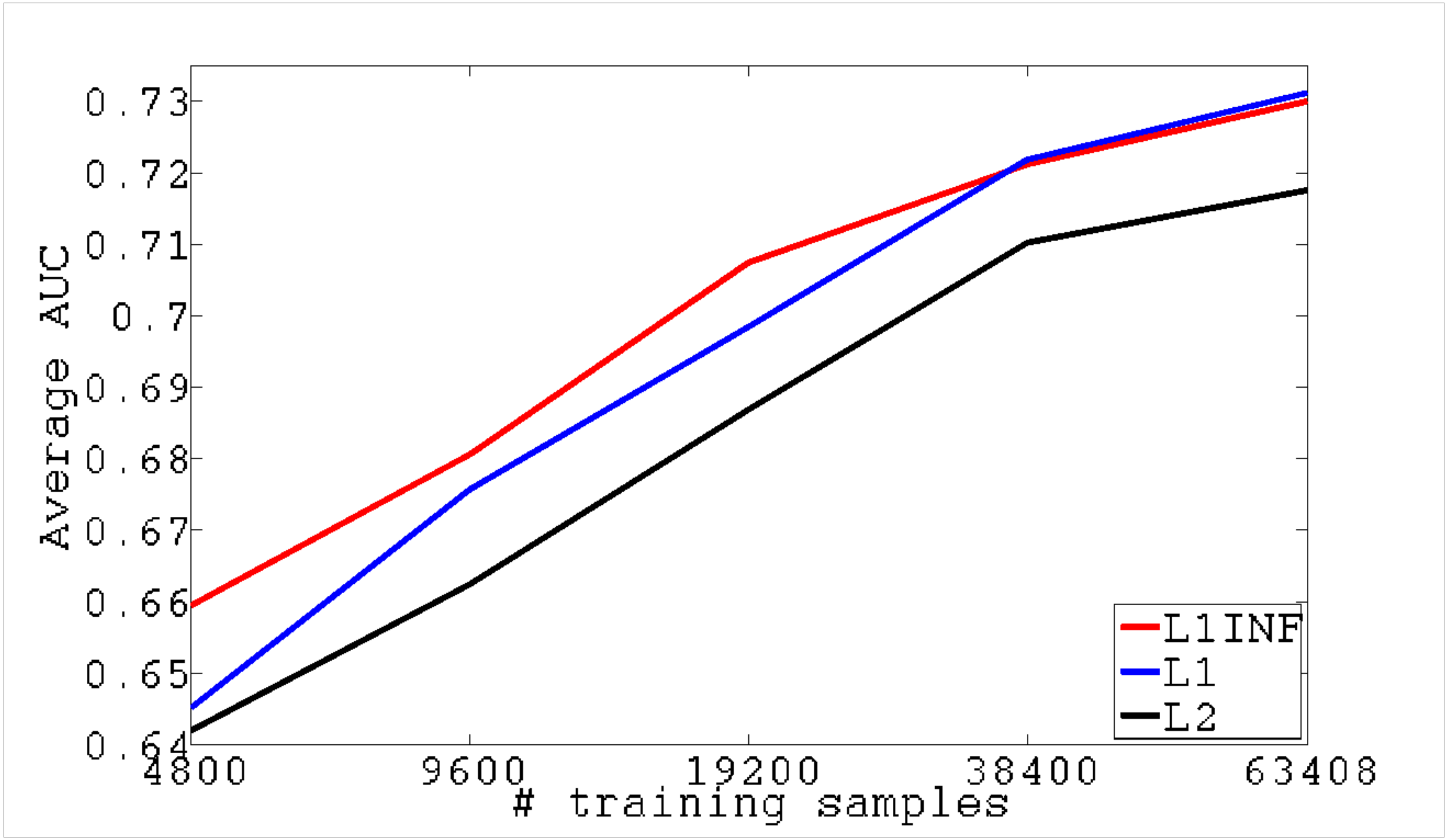




Dataset: Image Annotation

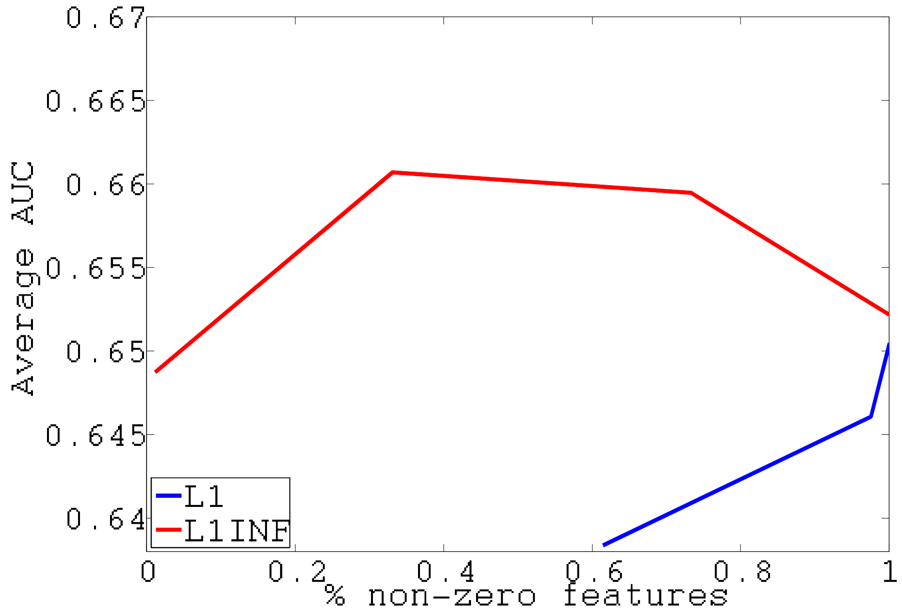
- 40 top content words
- Raw image representation: Vocabulary Tree
(Nister and Stewenius 2006)
- 11000 dimensions

Results

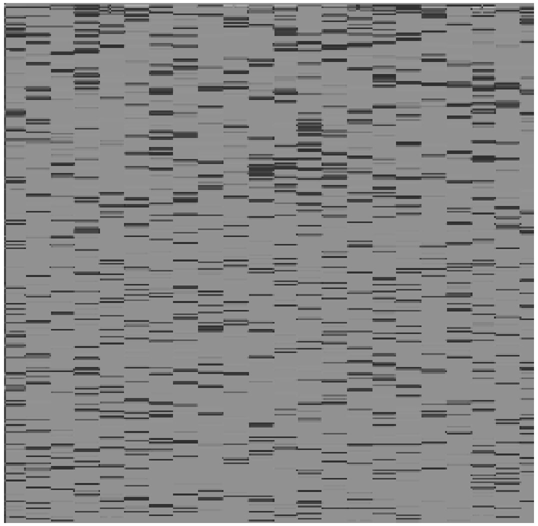


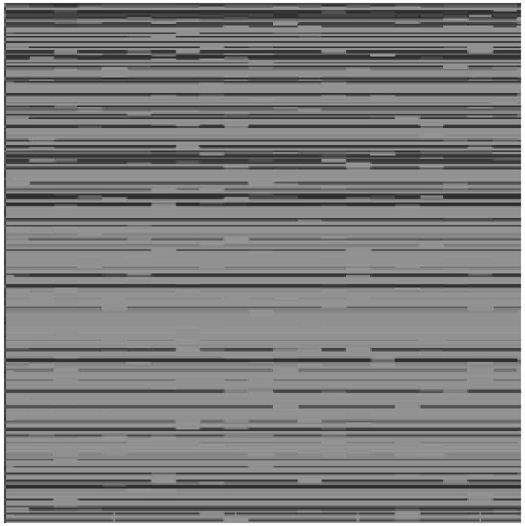
Most of the differences are statistically significant

Results



0.67
0.665
0.66
0.655
0.65
0.645
0.64
Average AUC
L1
L1INF
0
0.2
0.4
0.6
0.8
1
% non-zero features

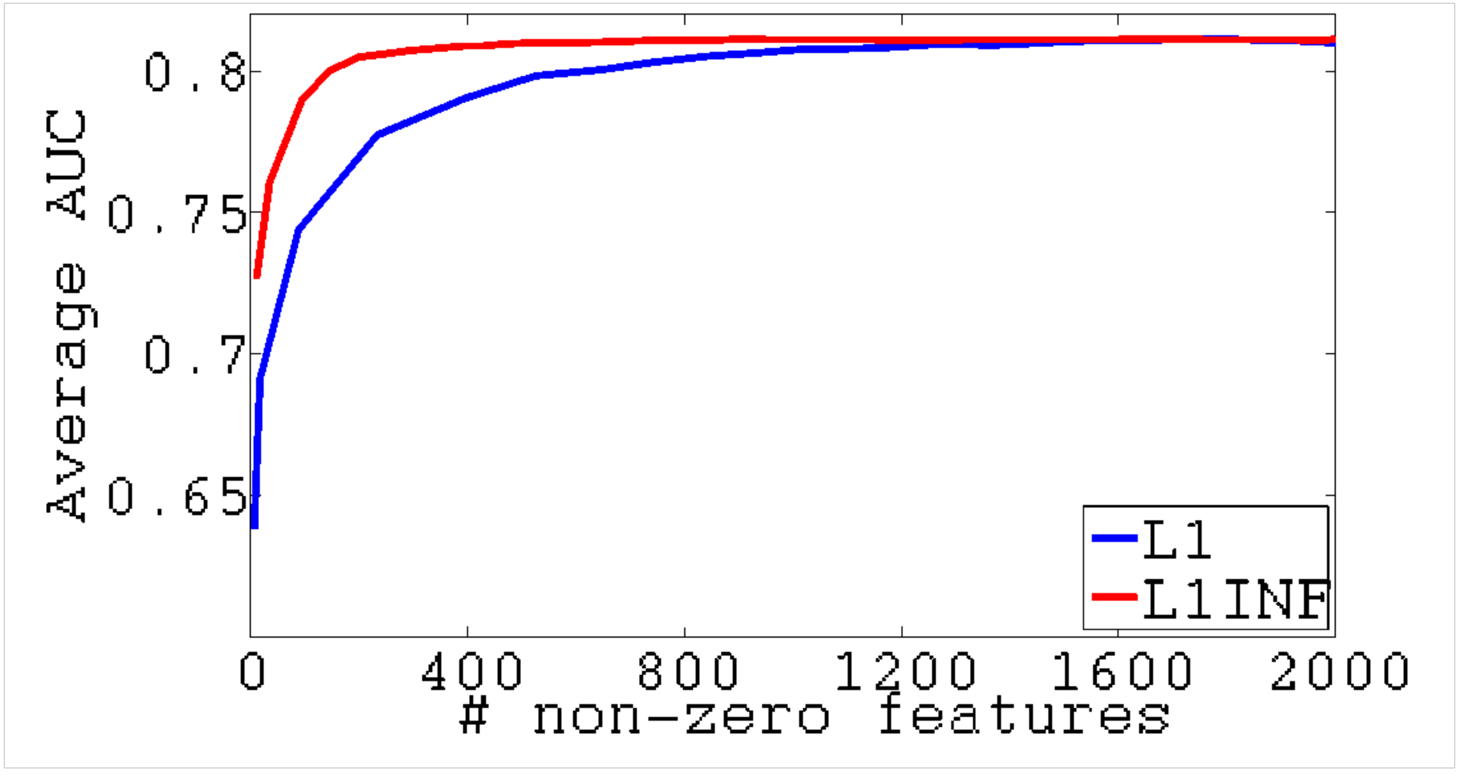




Dataset: Indoor Scene Recognition

- 67 indoor scenes.
- Raw image representation: similarities to a set of unlabeled images.
- 2000 dimensions.

Results



Average AUC
0.8
0.75
0.7
0.65
0
400
800
1200
1600
2000
non-zero features
L1
L1INF

Conclusions

- We proposed an efficient global optimization algorithm for $l_{1,\infty}$ regularization.

- A simple an efficient tool to implement an $l_{1,\infty}$ penalty, similar to standard l_1 and l_2 penalties.

- We presented experiments on image classification tasks and shown that our method can recover jointly sparse solutions.