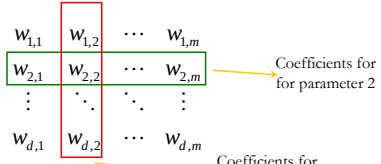
Problem:
Efficient training of jointly sparse models in high dimensional spaces.

Approach:
The $L_{1,\infty}$ norm has been proposed for sparse joint regularization.

Contributions:
We derive an efficient projected gradient method for L1-INF regularization. Our projection works on O(n log n) time, same cost as L_1 projection.
We test our algorithm in a multi-task image annotation problem and show that our algorithm can discover jointly sparse solutions and leads to better performance than L_2 and L_1 regularization.

Joint Regularization Penalty

- How do we penalize solutions that use too many features?

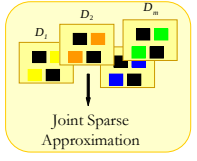


$$L_{1,\infty}(W) = \sum_{i=1}^{d} \max_k(|W_{ik}|)$$

The L_∞ norm on each row promotes non-sparsity on each row. → Share features

An L_1 norm on the maximum absolute values of the coefficients across tasks promotes sparsity. → Use few features

Application: Multitask Learning



Collection of Tasks

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

$D_k = \{(x_1^k, y_1^k), ..., (x_{n_k}^k, y_{n_k}^k)\}$

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

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

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

Ariadna Quattoni, Xavier Carreras, Michael Collins, Trevor Darrell
MIT Computer Science and Artificial Intelligence Laboratory

Constrained Convex Optimization Formulation

$$\arg\min_w \sum_{k=1}^{m} \frac{1}{|D_k|} \sum_{(x,y)\in D_k} l(f_k(x), y)$$ A convex function

$$s.t. \sum_{i=1}^{d} \max_k(|W_{ik}|) \le C$$ Convex constraints

We use a Projected SubGradient method.
Advantages: simple, scalable, guaranteed convergence rates.

These methods have been proposed for:
L_2 regularization [Shalev-Shwartz et al. 2007]
L_1 regularization [Duchi et al. 2008]

Euclidean Projection into the $L_{1,\infty}$ ball

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

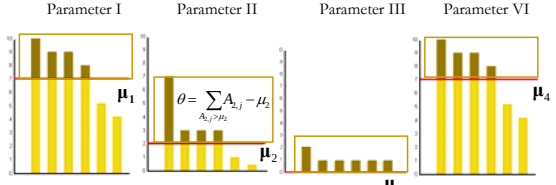
Characterization of the solution:

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

$$A_{i,j} \ge \mu_i \Rightarrow B_{i,j} = \mu_i$$
$$A_{i,j} \le \mu_i \Rightarrow B_{i,j} = A_{i,j}$$
$$\mu_i = 0 \Rightarrow B_{i,j} = 0$$

At the optimal solution of $P_{1,\infty}$ there exists a constant $\theta \ge 0$ such that for every i either:

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



Mapping to a simpler problem

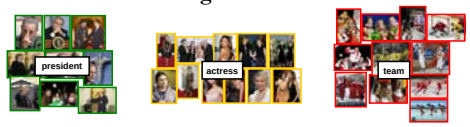
$$\mathbf{M}_{1,\infty}: \quad \text{find} \ \ \boldsymbol{\mu}, \theta$$
$$\text{s.t.} \ \ \sum_i \mu_i = C$$
$$\sum_{j:A_{i,j}\ge\mu_i} (A_{i,j} - \mu_i) = \theta, \ \forall i \text{ s.t. } \mu_i > 0$$
$$\sum_j A_{i,j} \le \theta, \ \forall i \ \text{ s.t. } \ \mu_i = 0$$
$$\forall i \ \ \mu_i \ge 0 \ ; \ \ \theta \ge 0$$

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 $M_{1,\infty}$.

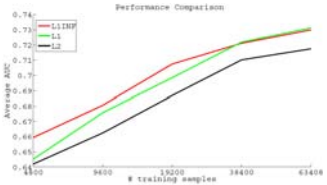
The total cost of the algorithm is dominated by a sort of the entries of **A**

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

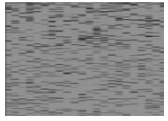
Dataset: Image Annotation

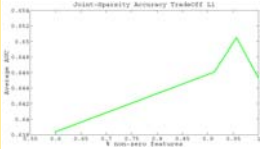


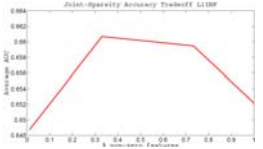
40 top content words
Image representation: Vocabulary Tree (Nister 2006)
11000 dimensions



Differences are statistically significant







Dataset: Indoor Scene Recognition

67 indoor scenes.
Image representation: Similarities to a set of unlabeled images.
2000 dimensions