LASS: a simple assignment model with Laplacian smoothing

Miguel Á. Carreira-Perpiñán and **Weiran Wang**

Electrical Engineering and Computer Science

University of California, Merced

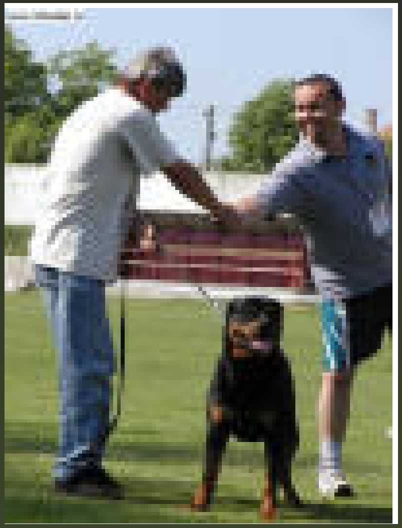
`http://eecs.ucmerced.edu`

Motivation

- We consider items and categories. Examples:
  - images and object categories (e.g. Flickr pictures)
  - documents and keywords (e.g. papers in a conference)
- Incomplete information because many categories

  partial tags/annotations for images; partial keywords for papers
- Categories have structure

  hierarchical; various intersection, inclusion and exclusion relations



Tags: dog, grass, man, sky

- Sometimes practical to tag an item as not associated with a certain category, particularly if this helps to make it distinctive

  "this paper is not about regression", "this patient does not have fever"
- In this type of applications, it is impractical for an item to be fully labeled over all categories, but it is natural for it to be associated or disassociated with a few categories.

Motivation: two sources of information

- This can be coded with item-category similarity values that are positive or negative, respectively, with the magnitude indicating the degree of association, and zero meaning indifference or ignorance.
  - Partial supervised information, specific for each item irrespectively of other items: the wisdom of the expert.
- Another practical source of information: similarity of a given item to other items, at least its nearest neighbors. We expect similar items to have similar assignment vectors, and this can be captured with an item-item similarity matrix and its graph Laplacian.
  - Partial unsupervised source of information, about an item in the context of other items: the wisdom of the crowd.

Problem statement

- We want to learn soft assignments of N items to K categories given two sources of information:
  - An item-category similarity matrix, which encourages items to be assigned to categories they are similar to (and to not be assigned to categories they are dissimilar to).
  - An item-item similarity matrix, which encourages similar items to have similar assignments. It propagates assignment information through the graph.

  Both matrices are sparse.
- The assignment $z_{nk} \in [0,1]$ indicates the degree of association of item n with category k. It may also be interpreted as a probability.
  So LASS transforms an incomplete matrix of item-category similarities into a complete matrix of item-category assignments.
- Given this information, we want a formulation as simple as possible.
- This is a new type of semisupervised learning
  different from (multiclass) classification.

LASS: quadratic program formulation

$$\min_{\mathbf{Z}} \quad \lambda \operatorname{tr}\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right) - \operatorname{tr}\left(\mathbf{G}^T\mathbf{Z}\right) = \frac{\lambda}{2}\sum_{n,m=1}^{N} w_{nm}\left\|\mathbf{z}_n - \mathbf{z}_m\right\|^2 - \sum_{n,k=1}^{N,K} g_{nk}z_{nk}$$

$$\text{s.t.} \quad \mathbf{Z}\mathbf{1}_K = \mathbf{1}_N, \ \mathbf{Z} \geq \mathbf{0}$$

- Sparse item-category similarity matrix $\mathbf{G}_{N\times K}$ contains similarity values that are positive or negative, with the magnitude indicating the degree of similarity, or zero meaning indifference or ignorance.
- Sparse item-item similarity matrix $\mathbf{W}_{N\times N}$ contains nonnegative similarity values of a given item to other items, with graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$ where $\mathbf{D} = \operatorname{diag}\left(\sum_{n=1}^{N} w_{nm}\right)$.
- One user parameter: $\lambda \geq 0$ trades off crowd vs expert information.
- We want to learn the assignment matrix $\mathbf{Z}_{N\times K}$ where z_{nk} is the assignment of item n to category k.

  $z_{nk} \in [0,1]$ and $\sum_{k=1}^{K} z_{nk} = 1$ for each $n = 1,\dots,N$.
- This is a convex quadratic program over NK variables.

LASS: particular cases

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \lambda \operatorname{tr}\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right) - \operatorname{tr}\left(\mathbf{G}^T\mathbf{Z}\right) \\ \text{s.t.} \quad & \mathbf{Z}\mathbf{1}_K = \mathbf{1}_N,\ \mathbf{Z} \geq \mathbf{0} \end{aligned}$$

- If $\lambda = 0$, then the problem becomes a linear program and separates over each item with solution is $z_{nk} = \delta(k - k_{\mathsf{max}}(n))$ where $k_{\mathsf{max}}(n) = \arg\max\{g_{nk},\ k = 1, \dots, K\}$, i.e., each item is assigned to its most similar category.
- If $\lambda = \infty$ or $\mathbf{G} = \mathbf{0}$, then the LASS problem is a QP with an infinite number of solutions of the form $\mathbf{z}_n = \mathbf{z}_m$ for each $n, m = 1, \dots, N$, i.e., all items have the same assignments.

LASS: existence and unicity of the solution

- Since the Hessian of the objective function is positive semidefinite, there can be multiple minima.
- Assume the graph Laplacian $\mathbf{L}$ corresponds to a connected graph and let $\mathbf{Z}^* \in \mathbb{R}^{N\times K}$ be a solution (minimizer) of the LASS problem. Then, any other solution has the form $\mathbf{Z}^* + \mathbf{1}_N\mathbf{p}^T$ where $\mathbf{p} \in \mathbb{R}^K$ satisfies the conditions:

$$\mathbf{p}^T\mathbf{1}_K = 0, \quad \mathbf{p}^T(\mathbf{G}^T\mathbf{1}_N) = 0, \quad \mathbf{Z}^* + \mathbf{1}_N\mathbf{p}^T \geq \mathbf{0}.$$

  Hence, the set of solutions is a convex polytope.

- Under the same assumptions, if $\max_k(\min_n(z^*_{nk})) = 0$ then the solution $\mathbf{Z}^*$ is unique.
- In practice, the solution is unique if the categories are sufficiently distinctive and λ is small enough.

Relationship with semisupervised learning (SSL)

- SSL minimizes $\mathrm{tr}\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right)$ given some of the $\mathbf{z}_n$ assignment vectors (= labeled items). Solution: sparse linear system.
- Similarities:
  - $\mathbf{L}$ plays the same role, i.e., to propagate label information in a smooth way according to the item-item graph.
  - Both rely on some given data to learn $\mathbf{Z}$: the similarity matrix $\mathbf{G}$ in LASS and the given labels $\mathbf{z}_n$ in SSL.
- Differences:
  - SSL is ill-suited for the partially labeled scenario because it is impractical to guess the full assignment for any item given its partial tags, while LASS optimizes over all the assignment values jointly.

    The semantics of the item-category similarities in LASS is that, where nonzero, they encourage the corresponding assignment towards relatively high or low values (for positive and negative similarities, respectively), and where zero, they reflect ignorance and are non-committing.

A simple and efficient algorithm

- Many possible algorithms for QP

  interior-point, gradient projection, active-set. . .

  We want to take advantage of the problem structure

  sparse $\mathbf{L}$, repeated for each category

- We apply the alternating direction method of multipliers (ADMM):

$$\boldsymbol{\nu} \leftarrow \frac{\rho}{K}(\mathbf{Y}-\mathbf{U})\mathbf{1}_K - \mathbf{h}$$

Lagrange multipliers for $\mathbf{Z}\mathbf{1} = \mathbf{1}$

$$\mathbf{Z} \leftarrow (2\lambda\mathbf{L} + \rho\mathbf{I})^{-1}(\rho(\mathbf{Y}-\mathbf{U}) + \mathbf{G} - \boldsymbol{\nu}\mathbf{1}_K^T)$$

Primal variables

$$\mathbf{Y} \leftarrow (\mathbf{Z}+\mathbf{U})_+$$

Auxiliary variables

$$\mathbf{U} \leftarrow \mathbf{U} + \mathbf{Z} - \mathbf{Y}$$

Lagrange multipliers for $\mathbf{Y} = \mathbf{Z}$

- Convergence to a global minimum is guaranteed for any $\rho > 0$.

  No other parameters to set (step sizes, etc.).

- Computational complexity: each iteration is $\mathcal{O}(NK)$

  Step over $\mathbf{Z}$: cache Cholesky factor, or use conjugate gradients.

Out-of-sample mapping

- Given a new, test item $\mathbf{x}$, along with its item-item and item-category similarities $\mathbf{w} = (w_n)$, $n = 1, \dots, N$ and $\mathbf{g} = (g_k)$, $k = 1, \dots, K$, respectively, we wish to find its assignment $\mathbf{z}(\mathbf{x})$ to each category.

  Or, its probability distribution over categories.

- We optimize LASS on a dataset consisting of the original training set augmented with $\mathbf{x}$, but keeping the original $\mathbf{Z}$ fixed to the values obtained during training. Hence, the only free parameter is the assignment vector $\mathbf{z}$ for the new point $\mathbf{x}$.

- Solution: $\mathbf{z}(\mathbf{x}) =$ Euclidean projection of $\bar{\mathbf{z}} + \gamma\mathbf{g} \in \mathbb{R}^K$ onto the probability simplex (which can be computed in $\mathcal{O}(K \log K)$):

$$\min_{\mathbf{z}} \left\| \mathbf{z} - (\bar{\mathbf{z}} + \gamma \mathbf{g}) \right\|^2 \quad \text{s.t.} \quad \mathbf{z}^T \mathbf{1}_K = 1,\ \mathbf{z} \geq \mathbf{0}$$

$$\gamma = \frac{1}{2\lambda \sum_{n=1}^N w_n} \qquad \bar{\mathbf{z}} = \sum_{n=1}^N \frac{w_n}{\sum_{n'=1}^N w_{n'}} \mathbf{z}_n$$

$\bar{\mathbf{z}}$ is the SSL out-of-sample mapping: weighted average of the training points' assignments.

Out-of-sample mapping (cont.)

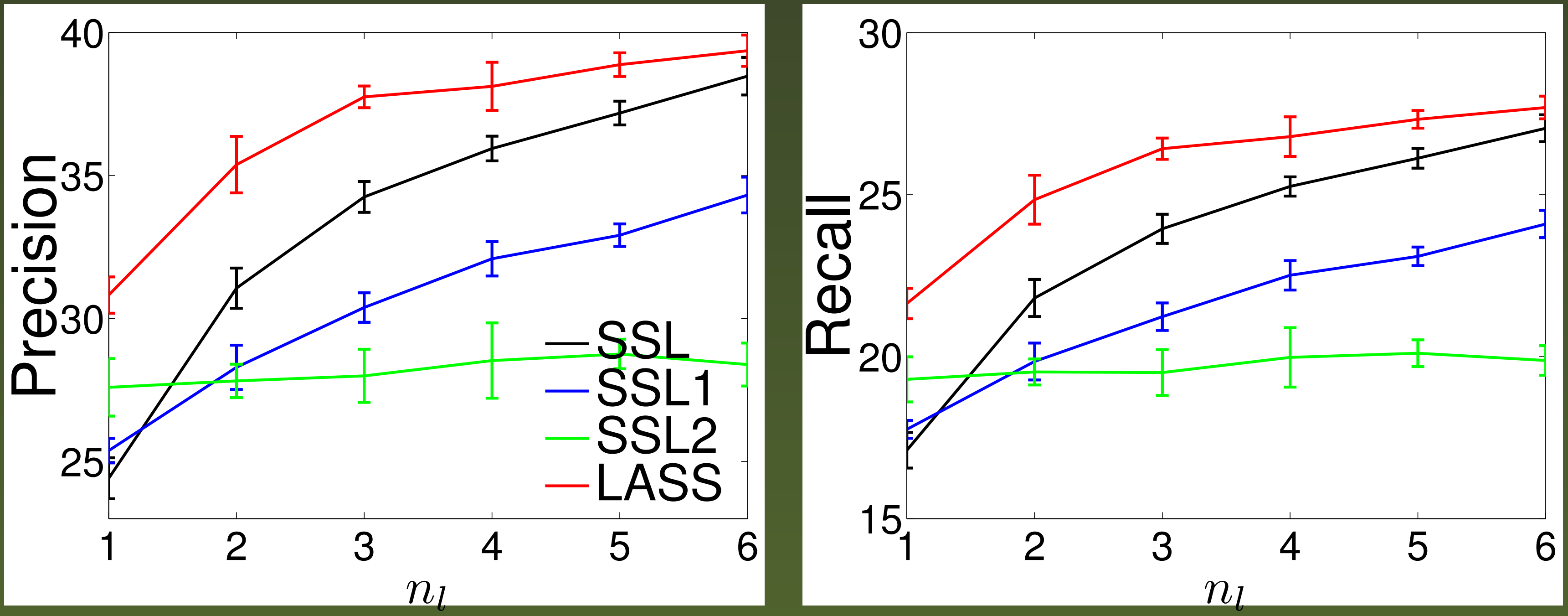
$$\min_{\mathbf{z}} \|\mathbf{z} - (\bar{\mathbf{z}} + \gamma \mathbf{g})\|^2 \quad \text{s.t.} \quad \mathbf{z}^T \mathbf{1}_K = 1,\ \mathbf{z} \geq \mathbf{0}$$

- So $\bar{\mathbf{z}} + \gamma \mathbf{g}$ is itself an average between the SSL (wisdom of the crowd) and the item-category similarities (wisdom of the expert).
- Tradeoff between the crowd ($\mathbf{w}$) and expert ($\mathbf{g}$) wisdoms:
  - $\lambda = 0$ or $\mathbf{w} = \mathbf{0}$: $\mathbf{x}$ is assigned to its most similar category.
  - $\lambda = \infty$ or $\mathbf{g} = \mathbf{0}$: $\mathbf{x}$'s assignment is the average of its neighbors'.

  . . . which could be used at test time to explore what-if scenarios.
- $\mathbf{z}(\mathbf{x})$ is different from the simple average of $\bar{\mathbf{z}}$ and $\mathbf{g}$ and may produce exact 0s or 1s for some entries (unlike with SSL).
- $\mathbf{z}(\mathbf{x})$ is a nonparametric, piecewise linear mapping.

Experimental results: ESP Game

- Dataset: $6\,100$ images with a total of 267 non-empty categories, with 7.2 categories per image on average.
- $\mathbf{G}$: for $4\,600$ images, we give positive similarity $(+1)$ for a random subset of size n_l from the categories it is tagged with, and give negative similarities (-1) randomly for 5 out of the 20 most frequent categories it is not tagged with, the other $1\,500$ images are completely unlabeled and used for testing.
- $\mathbf{W}$: 10-nearest-neighbor graph based on various image features.

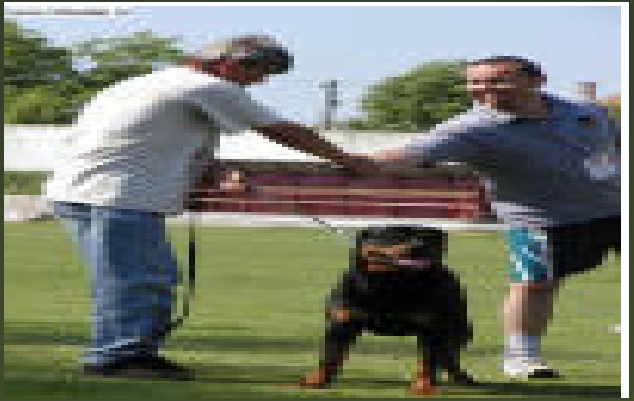


Experimental results: ESP Game (cont.)

Sample predictions on test images: ground truth, predicted (top z_k), mistakes.

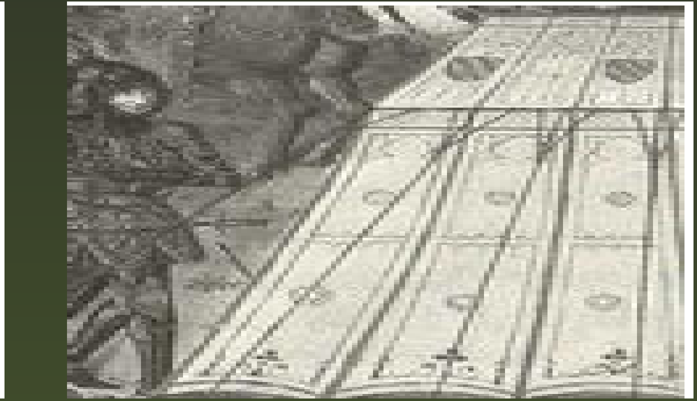
dog grass green man sky white

grass man sky green white tree . . .



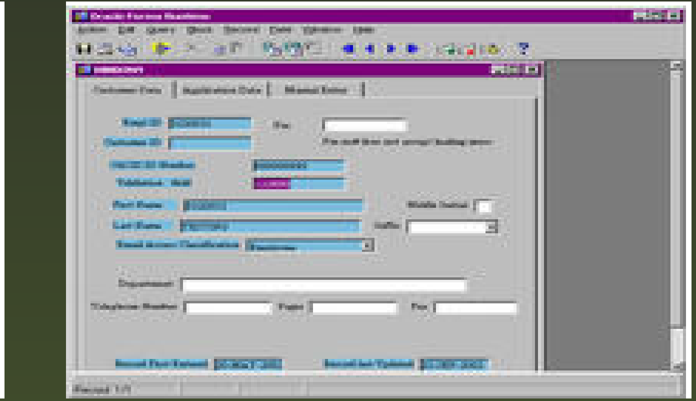
black drawing man old soldier tent white

black old drawing white tent sketch man . . .



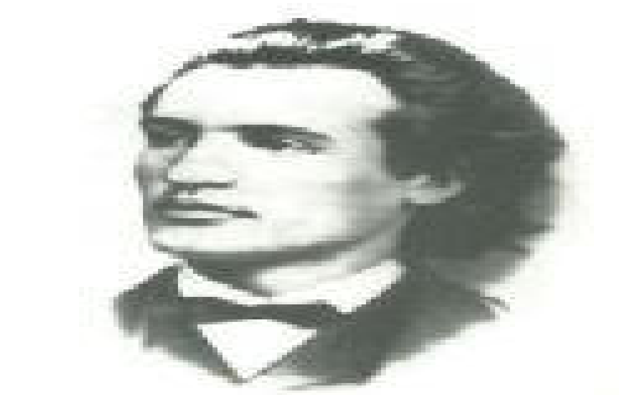
blue computer gray purple screen window

computer screen gray window blue white . . .



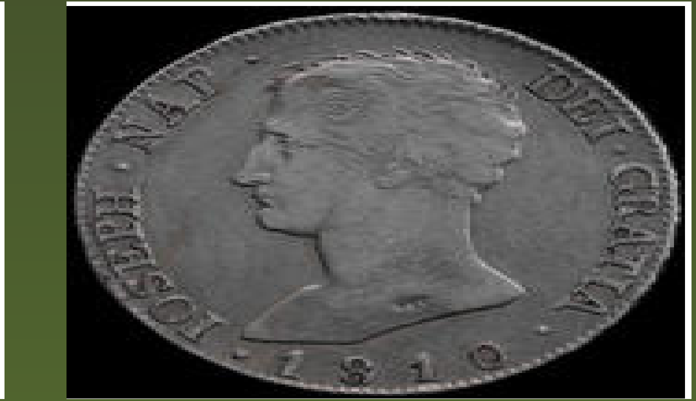
black drawing hair man nose old white

black white drawing man hair circle tie . . .



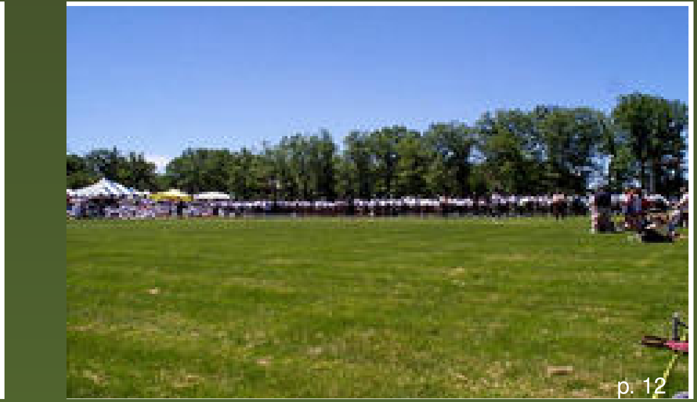
black coin man money old round silver white

black old round coin money man woman gray . . .



field grass green people sky tree

grass sky green man tree tent . . .



Conclusions

- LASS is a simple quadratic programming model for learning nonparametrically assignments that combines two complementary sources of information: the crowd wisdom and the expert wisdom.
- LASS transforms an incomplete matrix of item-category similarities into a complete matrix of item-category assignments.
- It is particularly attractive when fully labeling an item is impractical, or when categories have a complex structure and items can genuinely belong to multiple categories to different extents.
- It provides a different way to incorporate supervision to that of traditional SSL, which is ill-suited for the partially tagged setting.
- It seems able to predict reasonable assignments from very little information and minimal assumptions.

Matlab code: `http://eecs.ucmerced.edu`.

Relationship with semisupervised learning (SSL)

- Assume some items are fully labeled. Call $\mathbf{Z}_u$ of $N_u \times K$ and $\mathbf{Z}_l$ of $N_l \times K$ the matrices of labels for the unlabeled and labeled items, respectively, where $N = N_l + N_u$, and $\mathbf{Z}^T = (\mathbf{Z}_u^T\ \mathbf{Z}_l^T)$.
- SSL minimizes $\operatorname{tr}\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right)$ over $\mathbf{Z}_u$, with fixed $\mathbf{Z}_l$:

$$\begin{aligned}\min_{\mathbf{Z}_u} \operatorname{tr}\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right) &= \min_{\mathbf{Z}_u} \operatorname{tr}\left(\begin{pmatrix}\mathbf{Z}_u\\ \mathbf{Z}_l\end{pmatrix}^T \begin{pmatrix}\mathbf{L}_u & \mathbf{L}_{ul}\\ \mathbf{L}_{ul}^T & \mathbf{L}_l\end{pmatrix} \begin{pmatrix}\mathbf{Z}_u\\ \mathbf{Z}_l\end{pmatrix}\right)\\ &= \min_{\mathbf{Z}_u} \operatorname{tr}\left(\mathbf{Z}_u^T\mathbf{L}_u\mathbf{Z}_u + 2\mathbf{Z}_l^T\mathbf{L}_{ul}^T\mathbf{Z}_u\right) + \text{constant}\\ &\Rightarrow \mathbf{Z}_u = -\mathbf{L}_u^{-1}\mathbf{L}_{ul}\mathbf{Z}_l = \mathbf{L}_u^{-1}\mathbf{W}_{ul}\mathbf{Z}_l.\end{aligned}$$

- The SSL solution involves a sparse linear system of $N_u \times N_u$, and is a weighted average of the assignments of labeled items.

Relationship with semisupervised learning (SSL) (cont.)

- Similarities:
  - $\mathbf{L}$ plays the same role, i.e., to propagate label information in a smooth way according to the item-item graph.
  - Both rely on some given data to learn $\mathbf{Z}$: the similarity matrix $\mathbf{G}$ in LASS and the given labels $\mathbf{Z}_l$ in SSL.
- Differences:
  - SSL is ill-suited for the partially labeled scenario because it is impractical to guess the full assignment for any item given its partial tags, while LASS optimizes over all the assignment values jointly.

    The semantics of the item-category similarities in LASS is that, where nonzero, they encourage the corresponding assignment towards relatively high or low values (for positive and negative similarities, respectively), and where zero, they reflect ignorance and are non-committing.
  - In SSL, the assignment to each category can be solved independently; in LASS, the assignment to all categories are coupled due to the simplex constraints.

A simple and efficient algorithm

- Many possible algorithms for QP

  interior-point, gradient projection, active-set. . .

  We want to take advantage of the problem structure

  sparse $\mathbf{L}$, repeated for each category

- We apply the alternating direction method of multipliers (ADMM):
  - First transform the problem into

$$\min_{\mathbf{Z}\mathbf{1}_K=\mathbf{1}_N,\mathbf{Y}} \quad \lambda \operatorname{tr}\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right) - \operatorname{tr}\left(\mathbf{G}^T\mathbf{Z}\right) + \mathbf{1}_{\geq 0}(\mathbf{Y})$$
$$\text{s.t.} \quad \mathbf{Y} = \mathbf{Z}$$

where $\mathbf{1}_{\geq 0}(\mathbf{Y}) = \begin{cases} 0 & \text{if } \mathbf{Y} \geq \mathbf{0} \\ \infty & \text{otherwise} \end{cases}$ is the indicator function of the nonnegative orthant.

  - Then apply the augmented Lagrangian method and solve it with alternating optimization.

$$\mathcal{L}(\mathbf{Y},\mathbf{Z},\mathbf{U},\boldsymbol{\nu}) = \lambda \operatorname{tr}\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right) - \operatorname{tr}\left(\mathbf{G}^T\mathbf{Z}\right) + \mathbf{1}_{\geq 0}(\mathbf{Y}) + \operatorname{tr}\left(\boldsymbol{\nu}^T(\mathbf{Z}\mathbf{1}-\mathbf{1})\right) + \operatorname{tr}\left(\mathbf{U}^T(\mathbf{Y}-\mathbf{Z})\right) + \frac{\rho}{2}\left\|\mathbf{Y}-\mathbf{Z}\right\|^2$$

A simple and efficient algorithm (cont.)

- Choose penalty parameter $\rho > 0$ and set

$$\mathbf{h} = -\tfrac{1}{K}\mathbf{G}\mathbf{1}_K + \tfrac{\rho}{K}\mathbf{1}_N, \qquad \mathbf{R}\mathbf{R}^T = 2\lambda\mathbf{L} + \rho\mathbf{I} \; (\text{Cholesky decomposition})$$

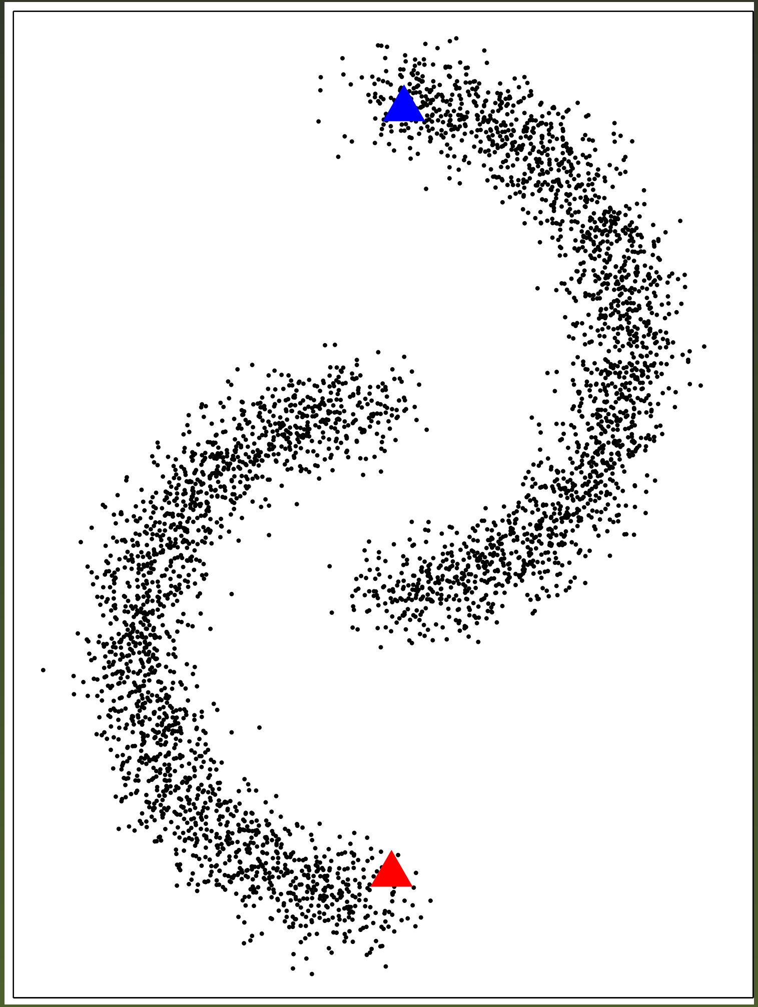
- We then iterate, in order, until convergence:

$$\begin{aligned}
\boldsymbol{\nu} &\leftarrow \tfrac{\rho}{K}(\mathbf{Y} - \mathbf{U})\mathbf{1}_K - \mathbf{h} && \text{Lagrange multipliers for } \mathbf{Z}\mathbf{1} = \mathbf{1} \\
\mathbf{Z} &\leftarrow (2\lambda\mathbf{L} + \rho\mathbf{I})^{-1}(\rho(\mathbf{Y} - \mathbf{U}) + \mathbf{G} - \boldsymbol{\nu}\mathbf{1}_K^T) && \text{Primal variables} \\
\mathbf{Y} &\leftarrow (\mathbf{Z} + \mathbf{U})_+ && \text{Auxiliary variables} \\
\mathbf{U} &\leftarrow \mathbf{U} + \mathbf{Z} - \mathbf{Y} && \text{Lagrange multipliers for } \mathbf{Y} = \mathbf{Z}
\end{aligned}$$

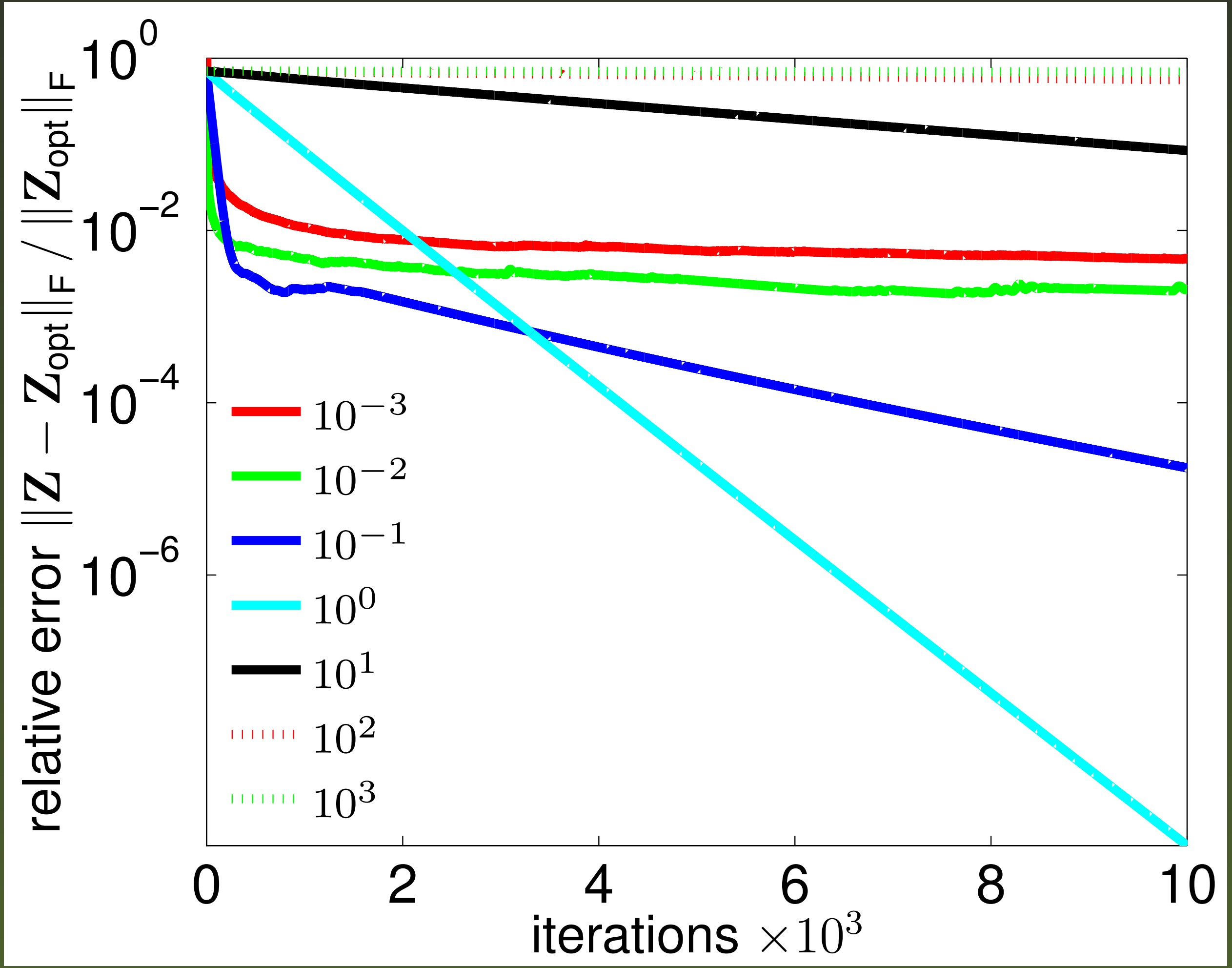
- Convergence to a global minimum is guaranteed for any $\rho > 0$. No other parameters to set (step sizes, etc.).
- Computational complexity: each iteration is $\mathcal{O}(NK)$ (plus `chol(L)`).
- Initialize $\mathbf{Y} = \mathbf{U} = \mathbf{0}$, stop when $\|\mathbf{Z}^{(\tau+\Delta)} - \mathbf{Z}^{(\tau)}\|_1 <$ `tol`.

Choice of penalty parameter ρ

We estimate the optimal ρ as $\rho^* = 2\lambda\sqrt{\sigma_{\min}(\mathbf{L})\sigma_{\max}(\mathbf{L})}$.
Convergence speed of ADMM for different ρ:



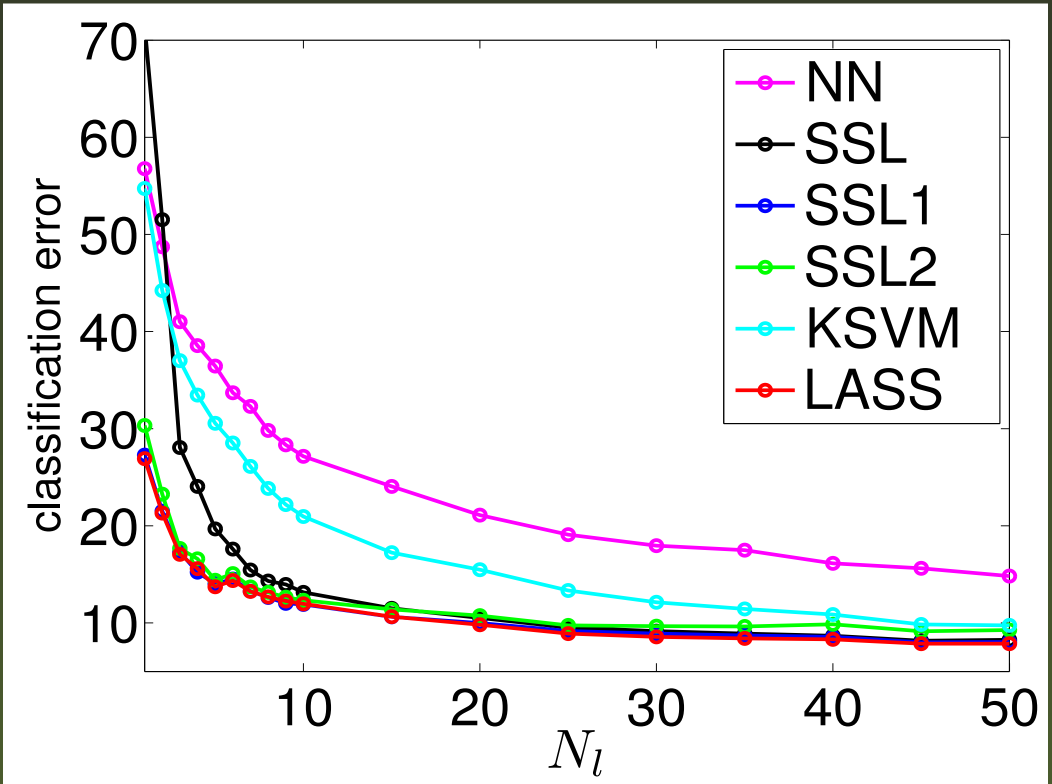
Dataset



Error (in log scale) vs iterations for different ρ/ρ^*

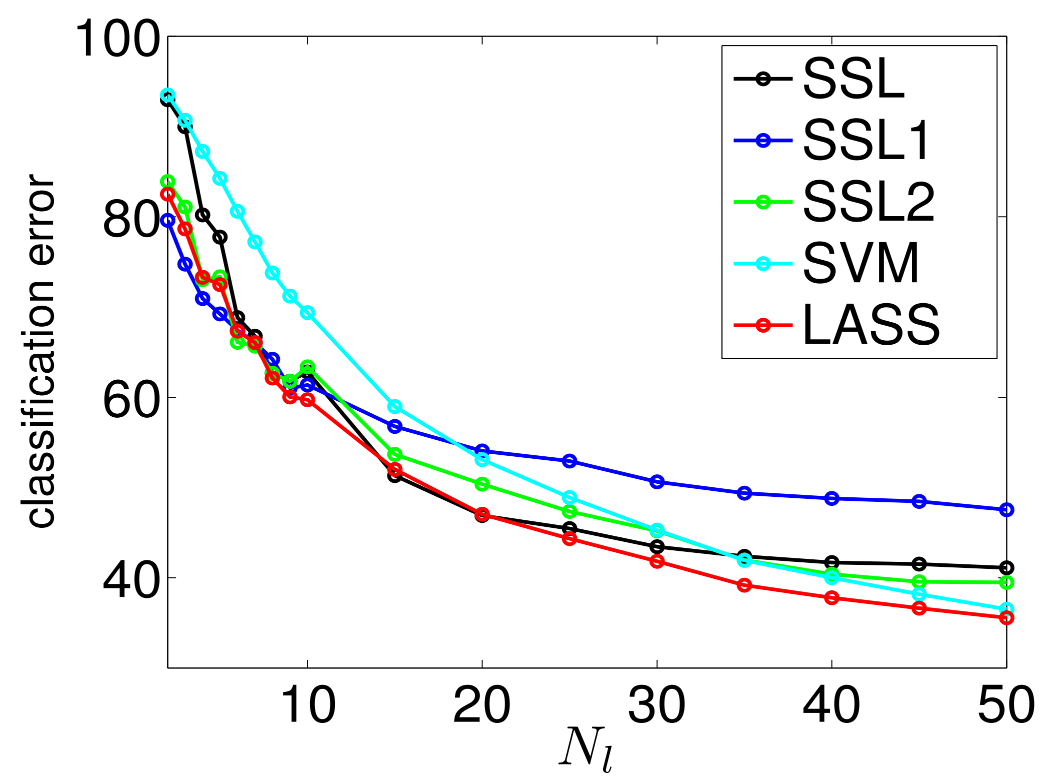
Experimental results: MNIST

- Dataset: $10\,000$ MNIST digit images $(0–9)$.
- $\mathbf{G}$: randomly select N_l images from each class and assign the correct $(+1)$ label.
- $\mathbf{W}$: 10-nearest-neighbor graph.



Experimental results: 20 Newsgroup

- Dataset: $11\,269$ documents, 27 topics (each document belongs to 1–3 topics).
- $\mathbf{G}$: randomly select N_l images from each document to give one $+1$ label and five -1 labels.
- $\mathbf{W}$: 10-nearest-neighbor graph based on TFIDF features.


classification error
N_l
SSL
SSL1
SSL2
SVM
LASS