Revisiting and Improving Semi-Supervised Learning : A Large Dimensional Approach

Xiaoyi Mai[1,2], Romain Couillet[1,2]

[1]LSS, CentraleSupélec, Université ParisSaclay, France.
[2]GIPSA-lab, Université Grenoble Alpes, France.

Abstract

- **Semi-Supervised Learning (SSL) on Graphs**: Learning a class-structured signal f from data graph and pre-labelled data.
- **Motivation**: Failure of classical algorithms in large dimensional regime.
- **Main Result**: Improved random-matrix inspired algorithm.

Preliminaries

- Data $x_1, \ldots, x_n \in \mathbb{R}^p$ in $\mathcal{C}_1$ or $\mathcal{C}_2$, seen as nodes in a graph.
- Data similarity matrix W. Usually,

$$W_{ij} = h(\|x_i - x_j\|^2) \geq 0$$

for some decreasing function h .

Smoothness Assumption:

- W_{ij} large implies tendency for x_i, x_j in the same class.
- Sought-after class-structured signal f smooth wrt W, i.e., with small smoothness penalty:

$$\mathcal{Q}(f) = \sum_{i,j} W_{ij}(f_i - f_j)^2 = f^T(D - W)f = f^T Lf.$$

Semi-Supervised Learning:

- $n_{[l]}$ labeled observations $\{(x_1, y_1), \ldots, (x_{n_{[l]}}, y_{n_{[l]}})\}$ with labels $y_i \in \{-1, 1\}$, and $n_{[u]}$ unlabeled samples $\{x_{n_{[l]}+1}, \ldots, x_n\}$.
- Objective: f with small $\mathcal{Q}(f)$ & in accordance with labeled data.

Curse of Dimensionality:

- Mixture model: $k \in \{1, 2\}$, $\mathbb{P}(x_i \in \mathcal{C}_k) = \rho_k$, $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(\mu_k, C_k)$.
- Large data asymptotics: $\frac{n_{[l]}}{p} \to c_{[l]} > 0$, $\frac{n_{[u]}}{p} \to c_{[u]} > 0$.
- Consequence of large p: distance concentration irrespective of class (at non-trivial regime of μ_k, C_k),

$$\frac{1}{p}\|x_i - x_j\|^2 = \tau + o_p(1), \quad \tau \equiv \frac{1}{p}\mathrm{tr}(\rho_1 C_1 + \rho_2 C_2)$$

Semi-Supervised Laplacian Regularization

- **Method:** find $f_{[u]}$ by minimizing $\mathcal{Q}(f)$ with $f_{[l]} = y_{[l]}$, e.g., by solving

$$\min_{f \in \mathbb{R}^n} f^\mathsf{T} Lf \quad \text{s.t.} \quad f_i = y_i, \quad 1 \leq i \leq n_{[l]}.$$

- **Solution:** $f_{[u]} = -L_{[uu]}^{-1} L_{[ul]} f_{[l]}$

where $L = I - D^{-1}W = \begin{bmatrix} L_{[ll]} & L_{[lu]} \\ L_{[ul]} & L_{[uu]} \end{bmatrix}$ with $D = \mathrm{diag}\{W1_n\}$.

- **Generalized Laplacian:** $L^{(\alpha)} = I - D^{-1-a}WD^a$.
- **Large Dimensional Behavior:** for $x_i \in \mathcal{C}_k$ unlabelled,

$$f_i = (c_{[l]}/c_0)(\rho_2 - \rho_1) + o_p(1)$$

Consequence: All f_i have the same sign if $\rho_2 \neq \rho_1$.

Amendment: Use balanced $f_{[l]} = \left(I_{n_{[l]}} - \frac{1}{n_{[l]}} 1_{n_{[l]}} 1_{n_{[l]}}^\mathsf{T}\right) y_{[l]}$.

$$\Downarrow$$

$$\sqrt{p} f_i = \eta(1 + a)(\mathrm{tr}C_2 - \mathrm{tr}C_1)/\sqrt{p} + o_p(1)$$

Consequence: All f_i have the same sign if $\mathrm{tr}C_1/\sqrt{p} \neq \mathrm{tr}C_2/\sqrt{p}$.

Amendment: Take $a = -1$.

$$\Downarrow$$

$$pf_i = g_i + o_p(1) \quad \text{where} \quad g_i \sim \mathcal{N}((-1)^k(1 - \rho_k)m, \sigma_k^2)$$

with σ_k^2/m^2 a decreasing function of c_l, but independent of $c_{[u]}$.

Inconsistency wrt unlabeled data

Solution: Centering Regularization

Failure of Laplacian Regularization & Distance Concentration:

$$s_{[u]} = -L_{[ul]}^{(a)} f_{[l]} = \left(D^{-1-a}WD^a\right)_{[ul]} f_{[l]}$$

$$f_{[u]} = L_{[uu]}^{(a)-1} s_{[u]} \simeq \left(I_{n_{[n]}} + \frac{1}{n_{[l]}} 1_{n_{[u]}} 1_{n_{[u]}}^\mathsf{T}\right) s_{[u]} \simeq s_{[u]} + \frac{1}{n_{[l]}}(1_{n_{[u]}}^\mathsf{T} s_{[u]}) 1_{n_{[u]}}$$

Ineffective learning from unlabelled data subgraph $L_{[uu]}^{(a)}$

Regularization with Centered Similarity Matrix:

$$\hat{W} = PWP \quad \text{with} \quad P \equiv I_n - \frac{1}{n} 1_n 1_n^\mathsf{T}$$

- Method: find $f_{[u]}$ with balanced $f_{[l]} = \left(I_{n_{[l]}} - \frac{1}{n_{[l]}} 1_{n_{[l]}} 1_{n_{[l]}}^\mathsf{T}\right) y_{[l]}$ by minimizing smoothness penalty on $\hat{W}$, i.e.,

$$\min_{f_{[u]} \in \mathbb{R}^{n_{[u]}}} \sum_{i,j} \hat{W}_{ij}(f_i - f_j)^2 = -f^\mathsf{T} \hat{W} f$$

$$\text{s.t.} \quad \|f_{[u]}\|^2 = n_{[u]} e^2$$

- Solution: $\hat{f}_{[u]} = \left(\alpha I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1} \hat{W}_{[ul]} f_{[l]}$ with $\alpha > \|\hat{W}_{[uu]}\|$

Advantages:

- $\hat{W}$ orthogonal to 1_n.
- Preserved difference between inter- and intra-class similarities.
- Balanced degrees as $\hat{d}_i = \sum_{j=1}^n \hat{W}_{ij} = 0$, for all i.

High Dimensional Performance: for $x_i \in \mathcal{C}_k$ unlabelled,

$$\hat{f}_i = \hat{g}_i + o_p(1) \quad \text{where} \quad \hat{g}_i \sim \mathcal{N}((-1)^k(1 - \rho_k)\hat{m}, \hat{\sigma}_k^2)$$

with $\hat{\sigma}_k^2/\hat{m}^2$ a decreasing function of both c_l and $c_{[u]}$ and

$$\lim_{\alpha \to +\infty} \hat{\sigma}_k^2/\hat{m}^2 = \sigma_k^2/m^2.$$

Consistent SSL for high dimensional data

Experimentation

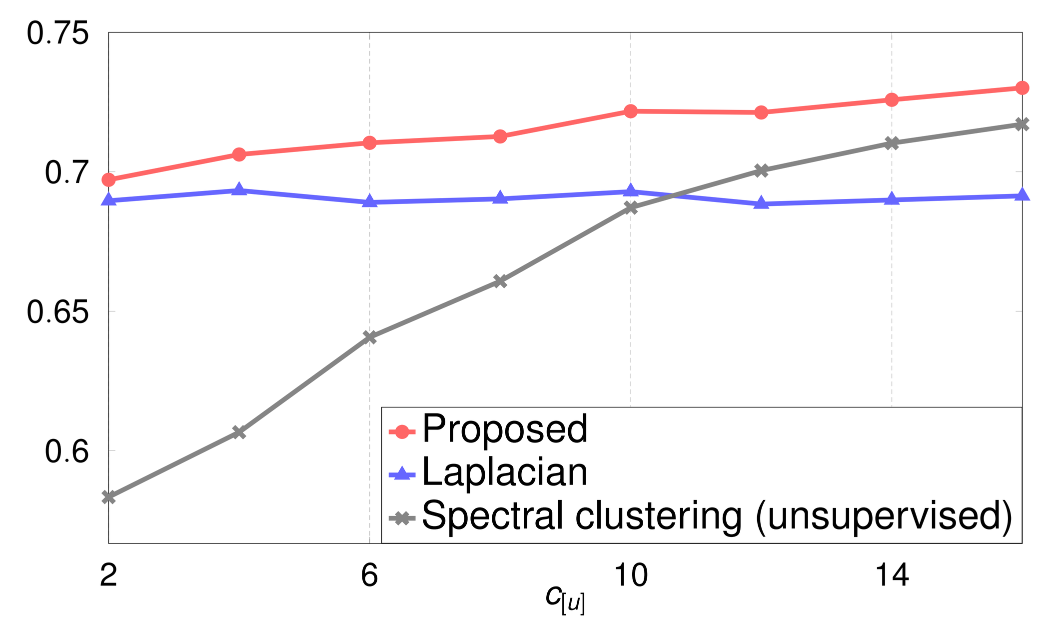

Figure: Accuracy as a function of $c_{[u]}$ for Gaussian data with $p = 80$, $h(t) = e^{-t}$. Averaged over $50000/n_{[u]}$ iterations.

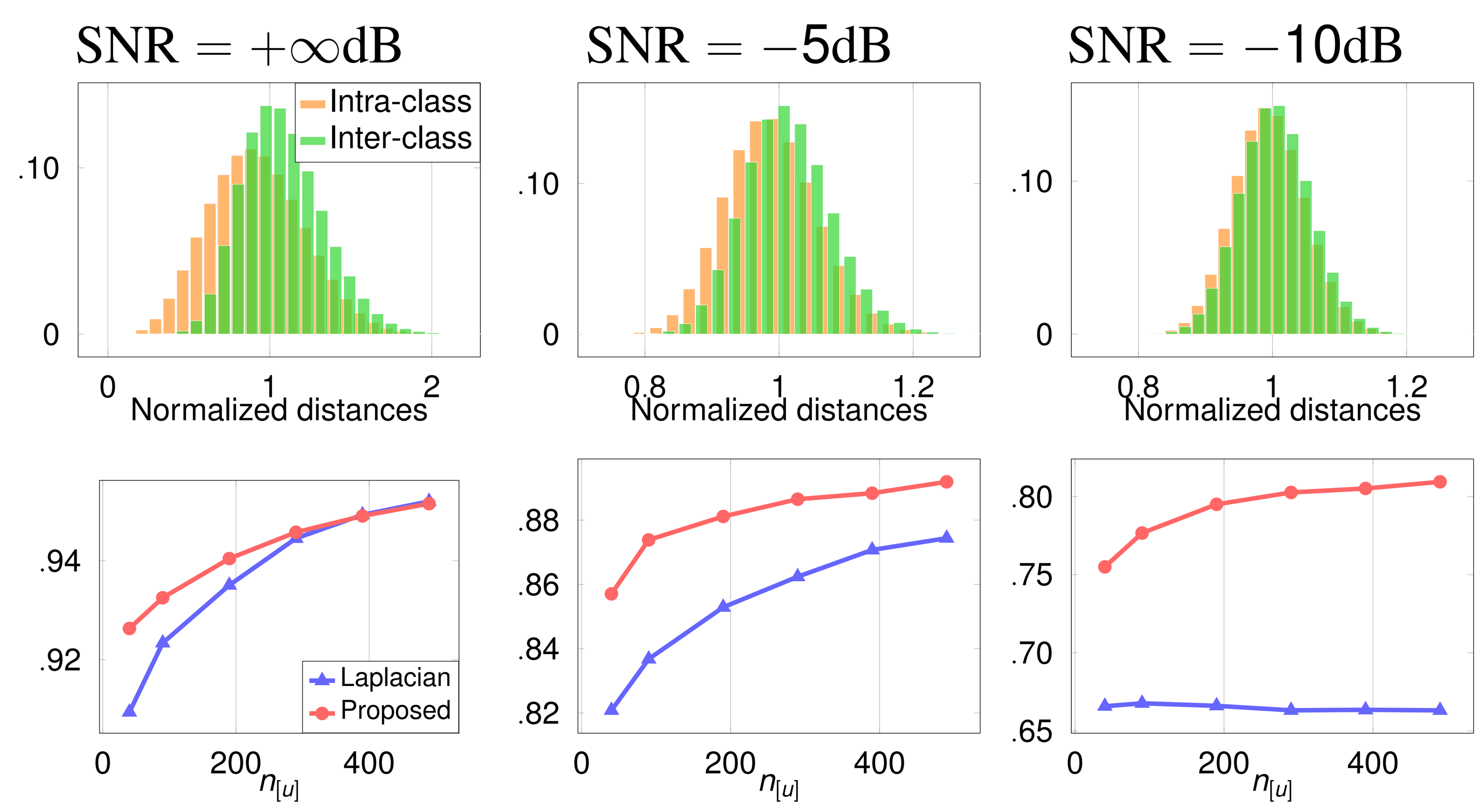

Figure: Top: distribution of normalized pairwise distances for noisy MNIST data (8,9). Bottom: average accuracy as a function of $n_{[u]}$ with $n_{[l]} = 10$, computed over 1000 random realizations.

Email: {xiaoyi.mai,romain.couillet}@l2s.centralesupelec.fr