Predicting the Labelling of a Graph via Minimum p-Seminorm Interpolation

Sparsity in Machine Learning and Statistics Workshop @ Cumberland Lodge, April 1-3, 2009

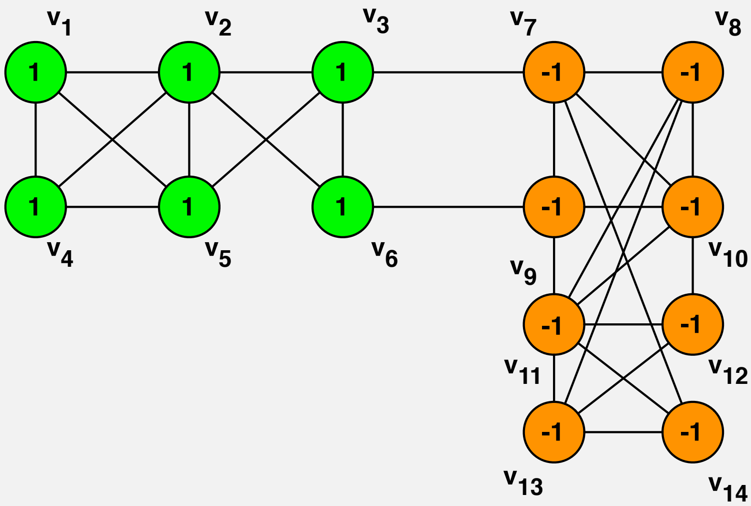
Mark Herbster (with Guy Lever)

University College London
Department of Computer Science
Centre for Computational Statistics and Machine Learning

Outline

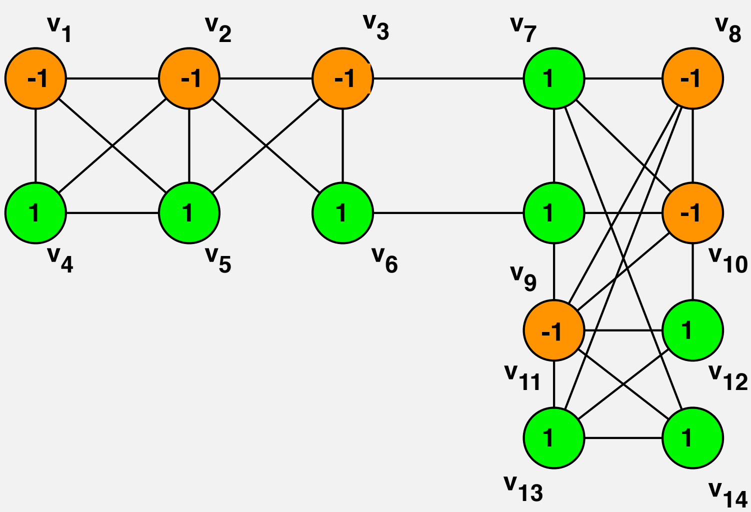
1. Online graph label prediction
2. Minimum p-seminorm interpolation
3. Mistake bound for interpolation
4. Applying interpolation to graph label prediction
5. p-Resistive networks interpretation
6. Mistake bound for graph label prediction

A prediction game



Nature asks	Learner predicts	Nature tells	Cum. Mistakes
v_1	1	1	0
v_5	1	1	0
v_2	1	1	0
v_{14}	1	-1	1
v_{10}	-1	-1	1
v_6	1	1	1
v_{11}	-1	-1	1
v_{13}	-1	-1	1
v_9	-1	-1	1
v_3	1	1	1
v_7	-1	-1	1
v_{12}	-1	-1	1
v_8	-1	-1	1
v_4	1	1	1

Let's try again



Nature asks	Learner predicts	Nature tells	Cum. Mistakes
v_1	1	-1	1
v_5	-1	1	2
v_2	1	-1	3
v_{14}	-1	1	4
v_{10}	1	-1	5
v_6	-1	1	6
v_{11}	1	-1	7
v_{13}	-1	1	8
v_9	-1	1	9
v_3	1	-1	10
v_7	-1	1	11
v_{12}	-1	1	12
v_8	1	-1	13
v_4	-1	1	14

Online Learning Model

- Aim: learn a function $\boldsymbol{u} : V \to \{-1, +1\}$ corresponding to a labeling of a graph $\mathcal{G} = (V, E)$ and $V = \{1, \ldots, n\}$.
- Learning proceeds in trials

for $t = 1, \ldots, \ell$ **do**

1. Nature selects $v_t \in V$
2. Learner predicts $\hat{y}_t \in \{-1, +1\}$
3. Nature selects $y_t \in \{-1, +1\}$
4. If $\hat{y}_t \neq y_t$ then mistakes $=$ mistakes $+ 1$

- Learner's goal: *minimize* mistakes
- Bound: mistakes $\leq$ *f(complexity*$(\mathcal{G}, \boldsymbol{u})$)

Minimum p-Seminorm Interpolation

$(\boldsymbol{\Psi}, p)$-seminorm

Given a linear map $\boldsymbol{\Psi} : \mathbb{R}^n \to \mathbb{R}^m$ define

$$\|\boldsymbol{u}\|_{\boldsymbol{\Psi},p} := \|\boldsymbol{\Psi}\boldsymbol{u}\|_p$$

Minimum $(\boldsymbol{\Psi}, p)$-seminorm Interpolation

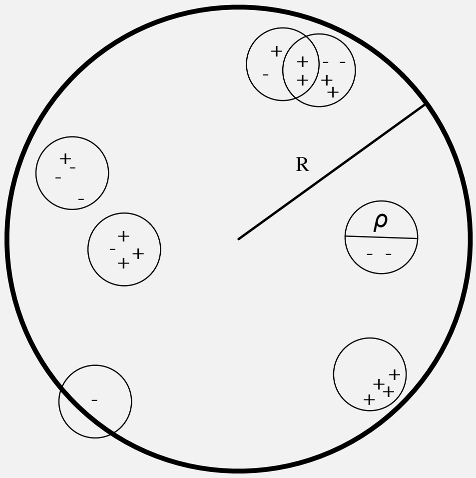
Input: $\{(i_t, y_t)\}_{t=1}^{\ell} \subseteq \{1, 2, \ldots, n\} \times \{-1, 1\}$.
for $t = 1, \ldots, \ell$ **do**
  Receive query: $i_t \in \{1, \ldots, n\}$
  Predict: $\hat{y}_{i_t} = \text{sign}(\boldsymbol{e}_{i_t}^{\top} \mathbf{w}_t)$
  Receive label: y_{i_t}
  if $\hat{y}_{i_t} = y_{i_t}$ **then** mistakes $=$ mistakes $+ 1$
  $\mathbf{w}_{t+1} = \arg\min_{\boldsymbol{u} \in \mathbb{R}^n} \{\|\boldsymbol{u}\|_{\boldsymbol{\Psi},p} : u_{i_1} = y_{i_1}, \ldots, u_{i_t} = y_{i_t}\}$

Duals and Covers

Definitions

Dual seminorm: $\|\boldsymbol{f}\|^*_{\Psi,p} := \sup\{|\boldsymbol{f}^\top \boldsymbol{u}| : \boldsymbol{u} \in \mathbb{R}^n, \|\boldsymbol{u}\|_{\Psi,p} \leq 1\}$

Metric: $d_{\Psi,p}(i,j) := \|\boldsymbol{e}_i - \boldsymbol{e}_j\|^*_{\Psi,p}$



Input space X of radius R with cover number $\mathcal{N}(X, \rho, d_{\Psi,p}) = 7$.

- Bounds depend on the cover of X in metric $d_{\Psi,p}$
- Both the number of balls $\mathcal{N}(X, \rho, d_{\Psi,p})$ and their diameter ρ
- Note $\mathcal{N}(X, \rho, d_{\Psi,p}) \in \{n, n-1, \ldots, 1\}$ and $\rho \in [0, 2R]$

Minimum p-Seminorm Interpolation Bound

Theorem

The mistakes M of the interpolation are bounded by

$$M \leq \mathcal{N}(X, \rho, d_{\Psi,p}) + \frac{\rho^2 \|\boldsymbol{u}\|^2_{\Psi,p}}{p-1} ,$$

for all $0 < \rho$, and for all $\boldsymbol{u} \in \mathbb{R}^n$ such that $p \in (1, 2]$ and

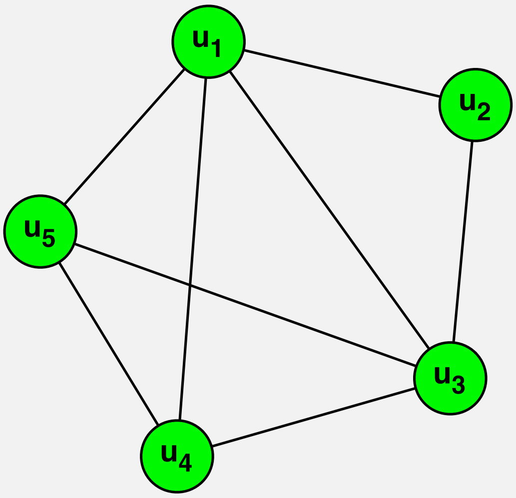
$$\boldsymbol{u}(i_t) = y_t \quad \text{(consistent labelings)}$$

for all $t = 1, \ldots, \ell$.

Edge Map $\Psi_{\mathcal{G}} : |V| \to |E|$

Example

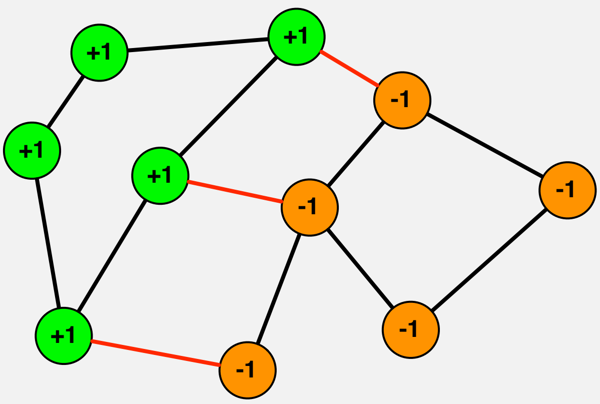
$$\boldsymbol{u} = (u_1, u_2, u_3, u_4, u_5)$$



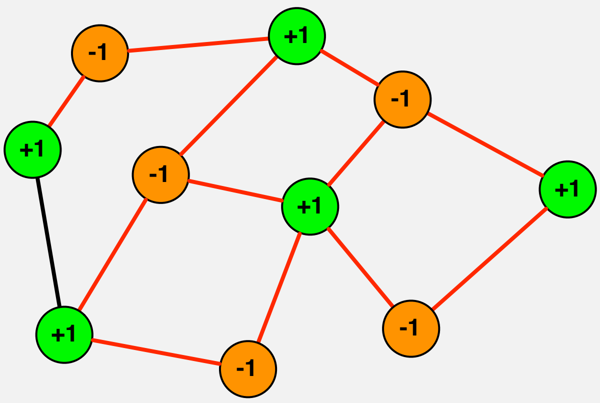
$$\Psi_{\mathcal{G}}\boldsymbol{u} = (u_1 - u_2, u_1 - u_3, u_1 - u_4, u_1 - u_5, u_2 - u_3, u_3 - u_4, u_4 - u_5)$$

† Extends to weighted graphs

Graph p-seminorm illustrated



$$\|\boldsymbol{u}\|^p_{\Psi_{\mathcal{G}},p} = 3 \times 2^p$$



$$\|\boldsymbol{u}\|^p_{\Psi_{\mathcal{G}},p} = 12 \times 2^p$$

p-Resistive Networks

- Define energy dissipation (power) of a state as

$$P(\boldsymbol{u}) := \|\boldsymbol{u}\|^p_{\Psi_{\mathcal{G}},p} = \sum_{(i,j)\in E(\mathbf{G})} |u_i - u_j|^p$$

- A network $\mathcal{C}$ is a graph with potential constraints $(u_{i_1}, y_1), \ldots, (u_{i_\ell}, y_\ell)$.
- The power of a network is defined by

$$\mathbf{w}(\mathcal{C}) = \min_{\boldsymbol{u}\in\mathbb{R}^n} \{P(\boldsymbol{u}) : u_{i_1} = y_1, \ldots, u_{i_\ell} = y_\ell\}$$

- Analogues of usual “electric network” theory now follow ...

1. Kirchoff’s laws
2. Ohm’s law
3. conservation of energy principle,
4. the “rules” of resistors in parallel and in series,
5. Rayleigh’s monotonicity principle

p-Resistance

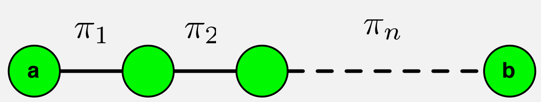
Definition

The (effective) p-resistance between any two vertices a and b is

$$r_{\mathcal{G},p}(a,b) = (\|\boldsymbol{e}_a - \boldsymbol{e}_b\|^*_{\Psi_{\mathcal{G}},p})^p.$$

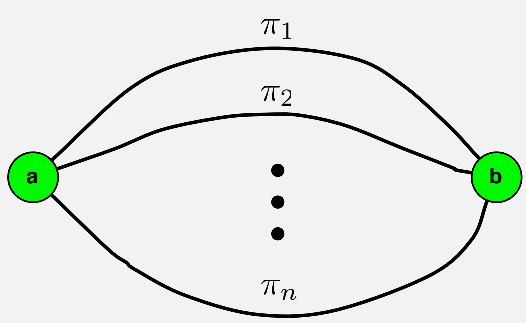
Resistors in series

$$r_{\mathcal{G},p}(a,b) = \left(\sum_{i=1}^n \pi_i^{\frac{1}{p-1}}\right)^{p-1}$$



Resistors in parallel

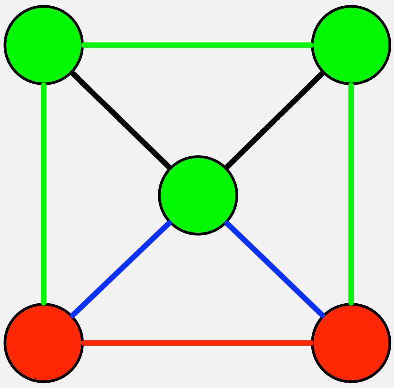
$$r_{\mathcal{G},p}(a,b) = \left(\sum_{i=1}^n \frac{1}{\pi_i}\right)^{-1}$$



Wide Diameter

Definition

A graph has wide diameter (k, Δ_k) if there exist k edge disjoint paths of length no more than Δ_k between all pairs of vertices.



Connectivity: $k = 1$, Wide Diameter: $\Delta_1 = 2$

Connectivity: $k = 2$, Wide Diameter: $\Delta_2 = 2$

Connectivity: $k = 3$, Wide Diameter: $\Delta_3 = 3$

- The wide diameter is used as a basis for upper bounding p-resistance with the resistors in parallel and series laws.

Graph prediction

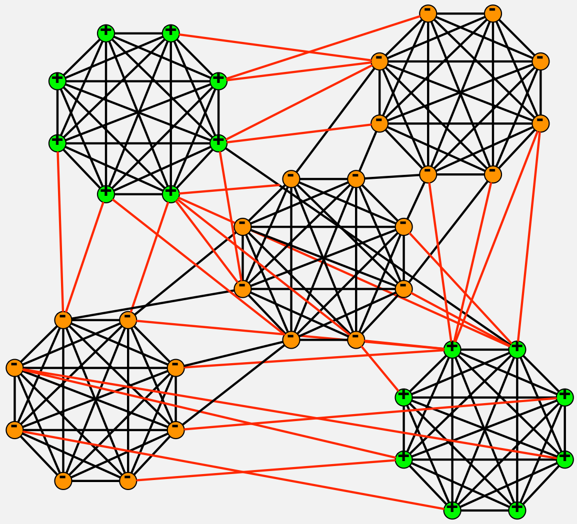
Theorem

The number of mistakes, M, incurred by minimum p-seminorm interpolation predicting the labeling of $\mathcal{G} = (V, E)$ with $p := \frac{c}{c-1}$ is bounded by

$$M \leq \begin{cases} N + \frac{4e^2 \Phi^2_{\mathbf{G}}(\boldsymbol{u})[\log(\Delta_k) + \log(\frac{k}{\Phi_{\mathbf{G}}(\boldsymbol{u})}) - 1]}{k^2} & \frac{k\Delta_k}{\Phi_{\mathbf{G}}(\boldsymbol{u})} > e^2, \quad p \in (1, 2) \\ N + \frac{4\Phi_{\mathbf{G}}(\boldsymbol{u})\Delta_k}{k} & \frac{k\Delta_k}{\Phi_{\mathbf{G}}(\boldsymbol{u})} \leq e^2, \quad p = 2 \end{cases}$$

where $c = \min(\log[\frac{k\Delta_k}{\Phi_{\mathbf{G}}(\boldsymbol{u})}], 2)$ and $V_1 \cup \ldots \cup V_N = V_{\mathcal{G}}$ is any vertex set partition with induced subgraphs $G_1, \ldots, G_N$ of maximum wide diameter $\Delta_k := \max\{\Delta_k(G_i) : i = 1, \ldots, N\}$, $\Phi_{\mathbf{G}}(\boldsymbol{u})$ is the "cutsize" of $\boldsymbol{u} \in \{-1, 1\}^n$ any consistent labelling.

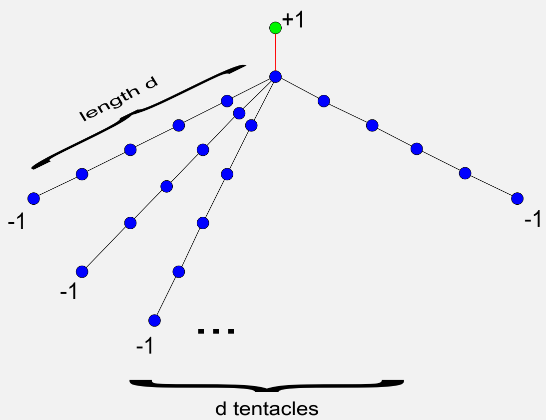
Example: Prototypical Clusters



$c \times m$-cliques with ℓ cut edges

- Cover: $N = c$, Cutsize: $\Phi_{\mathbf{G}}(\boldsymbol{u}) = \ell$
- Connectivity: $k = m - 1$, Wide Diameter: $\Delta_{m-1} = 2$
- When $p = 2$ then $M \leq c + \frac{8\ell}{m-1}$
- Example: two clusters $\ell < m \rightarrow M \leq 10$
- Furthermore $8\ell - 1 < m \rightarrow M \leq 2$ (optimal)

Example: Octopus



A d-tentacle octopus graph with length d tentacles ($n = d^2 + 1$)

- Cover: $N = 1$, Cutsize: $\Phi_{\mathbf{G}}(\boldsymbol{u}) = 1$
- Connectivity: $k = 1$, Diameter: $\Delta_1 = 2d$
- When $p = 1 + \frac{1}{\log(2d)-1}$ then $\mathcal{M}_A \leq 1 + 4e^2(\log(2d) - 1)$
- When $p = 2$ then $\sqrt{n} \leq \mathcal{M}_A \leq O(\sqrt{n})$

Thanks