# Exploiting Cluster Structure to Predict the Labeling of a Graph

ALT 2008

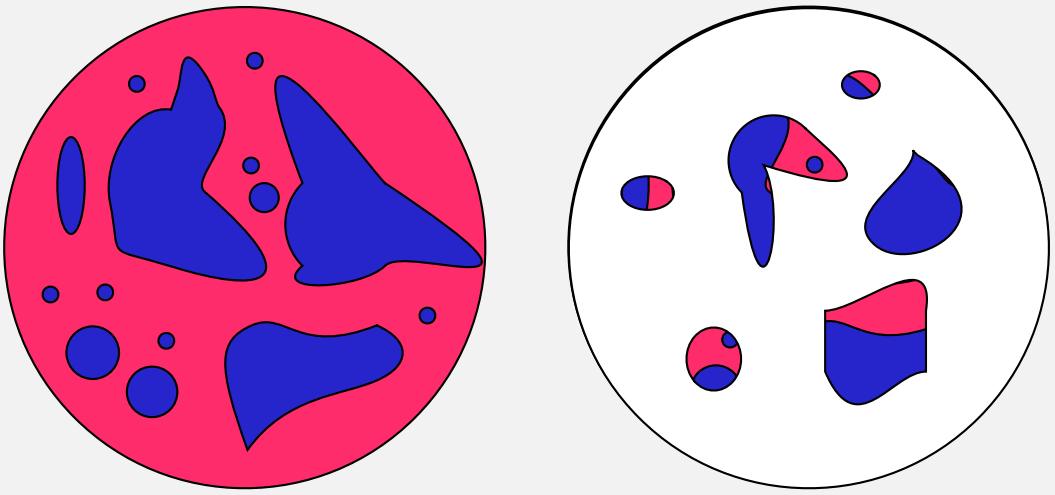
Mark Herbster

University College London
Department of Computer Science

16 October, 2008

## Overview

1. Give perceptron-like algorithm for graph label prediction
2. Improve on Perceptron bound when cluster-structure

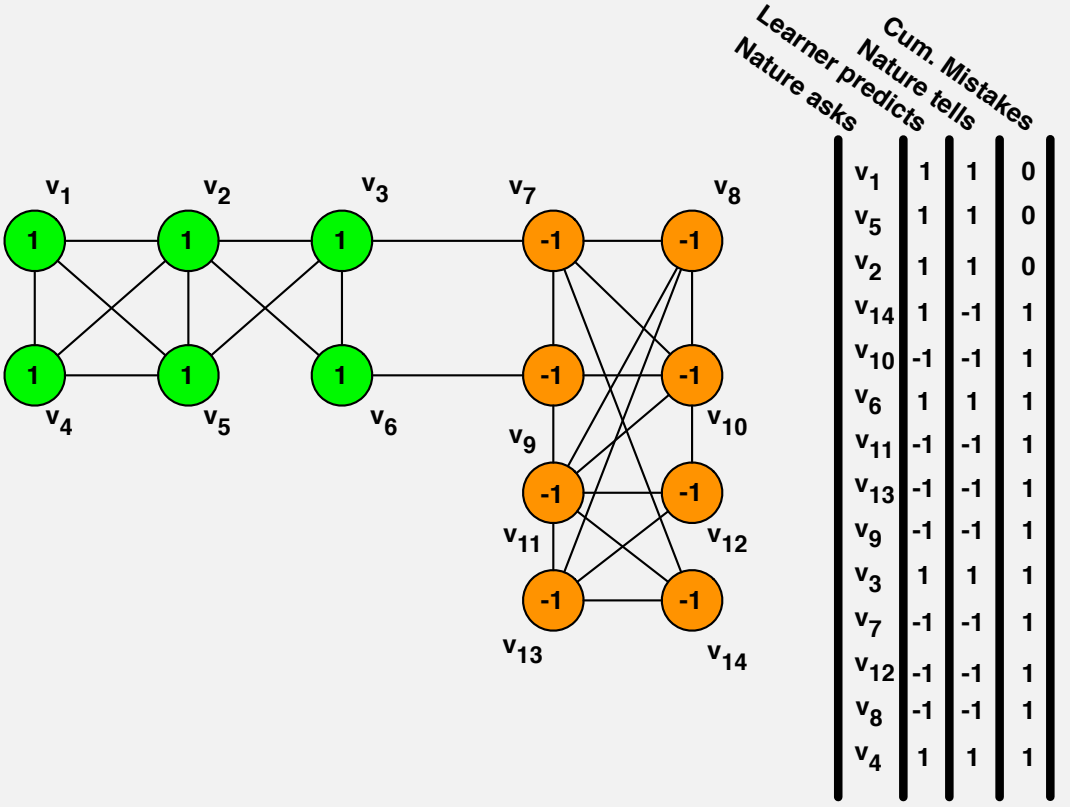


*"Default" Assumption* – *"Cluster" Assumption*

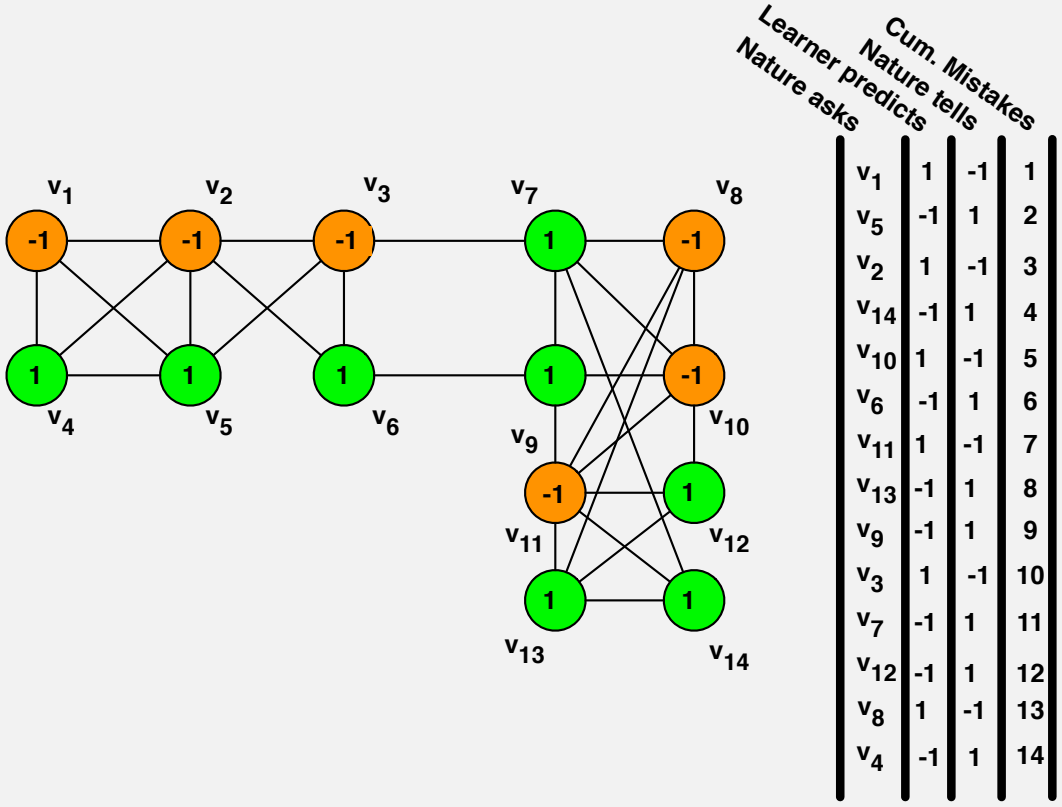
# Outline

- Review: Online graph label prediction
- Review: Predicting labeling of a graph with a perceptron
- Problem: Multiple clusters (Perceptron fails)
- Solution: Pounce algorithm
- Bound: Via cover of input space
- Application: Heat kernel
- Moral: Value of unlabeled data

# A prediction game



# Let's try again




# Online Learning Model

- Aim: learn a function $\boldsymbol{u} : V \to \{-1, +1\}$ corresponding to a labeling of a graph $G = (V, E)$ 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*($\boldsymbol{u}$))

# Perceptron Bound (Novikoff)

## Theorem [Novikoff]:

Given a sequence $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^{\ell} \subseteq \mathcal{H} \times \{-1, 1\}$ then the mistakes of the perceptron are bounded by

$$M \leq \|\boldsymbol{u}\|^2 R$$

with $R = \max_t(\|\boldsymbol{x}_t\|^2)$ for all $\boldsymbol{u} \in \mathcal{H}$ such that

$$\langle \boldsymbol{u}, \boldsymbol{x}_t \rangle y_t \geq 1$$

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

# Signed Laplacian

## Matrix-Edge Decomposition

If $\mathbf{G}$ is a symmetric matrix then

$$\mathbf{G} = \sum_{(i,j)\in E^+} a_{ij}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top + \sum_{(i,j)\in E^-} b_{ij}(\boldsymbol{e}_i + \boldsymbol{e}_j)(\boldsymbol{e}_i + \boldsymbol{e}_j)^\top + \sum_{i=1}^{n} c_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$$

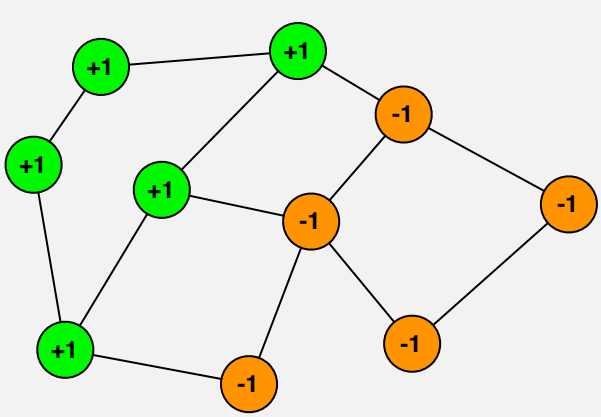
where $(\boldsymbol{e}_i)_j = \delta_{ij}$ with $a_{ij}, b_{ij} > 0$ and $c_i \in \mathbb{R}$ and thus

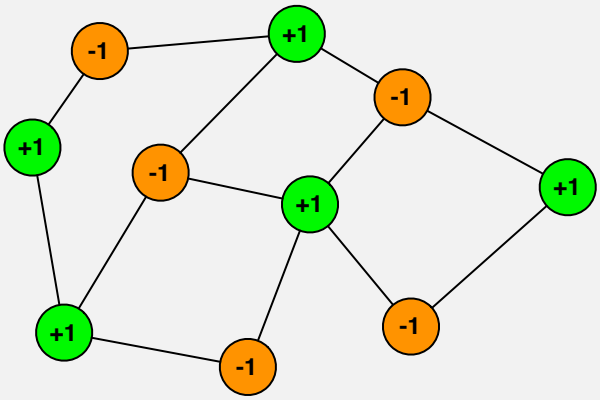
$$\|\boldsymbol{u}\|_{\mathbf{G}}^2 = \boldsymbol{u}^\top \mathbf{G} \boldsymbol{u} = \sum_{(i,j)\in E^+} a_{ij}(u_i - u_j)^2 + \sum_{(i,j)\in E^-} b_{ij}(u_i + u_j)^\top + \sum_{i=1}^{n} c_i u_i^2$$

## Definitions

- $\mathbf{G}$ is a graph Laplacian if $E^- = \emptyset$ and $c_1 = \cdots = c_n = 0$
- $\mathbf{G}$ is a signed graph Laplacian if $c_1 \geq 0, \ldots, c_n \geq 0$ (p.s.d.)

# Examples




$\|\boldsymbol{u}\|^2 = 3 \times 4$ positive unit-edges $\|\boldsymbol{u}\|^2 = 12 \times 4$

$\|\boldsymbol{u}\|^2 = 10 \times 4$ negative unit-edges $\|\boldsymbol{u}\|^2 = 1 \times 4$

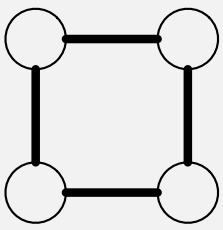
# Resistance distance

The *resistance distance* [KR93] between vertex $\boldsymbol{v}_p$ and $\boldsymbol{v}_q$,

$$\|\boldsymbol{v}_p - \boldsymbol{v}_q\|_{\mathbf{G}}^2 = (\boldsymbol{e}_p - \boldsymbol{e}_q)^\top \mathbf{G}^+ (\boldsymbol{e}_p - \boldsymbol{e}_q)$$

is the *effective resistance* between $\boldsymbol{v}_p$ and $\boldsymbol{v}_q$. The graph is the circuit and edge $(i, j)$ has resistance $a_{ij}^{-1}$.

- Resistance Diameter : $R_{\mathbf{G}} := \max_{p,q \in V} \|\boldsymbol{v}_p - \boldsymbol{v}_q\|_{\mathbf{G}}^2$
- Geodesic distance upper bounds the resistance desistance



*Resistance Diameter:* $R_{\mathbf{G}} = 1$

# Predicting the Labeling of a Graph with the Perceptron

### Theorem[HP06]:

The mistakes of perceptron are bounded by

$$M \leq 2\|\boldsymbol{u}\|^2 R_{\mathbf{G}} + 2$$

for all consistent labelings $\boldsymbol{u} \in [-1, 1]^n$.

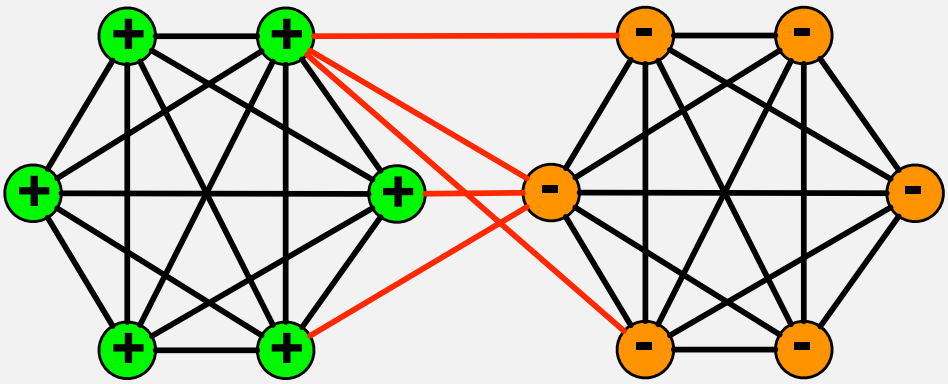
### Proof.

If $\mathbf{G}$ is a Laplacian use kernel $\mathbf{G}^+ + 11^\top R_{\mathbf{G}}$.

### Observations

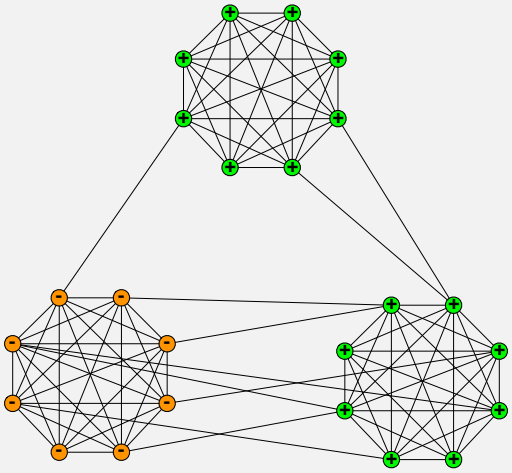
- Optimal $\boldsymbol{u}^*$ are the voltages which minimizes energy
- Optimal $\|\boldsymbol{u}^*\|^2$ is bounded by each label-separating cut.

# Bound: 2 Prototype Clusters




- Two $m$-cliques with $\ell$ edges ($\ell < m$) between cliques
- Norm: $\|\boldsymbol{u}\|^2 = 4\ell$
- Resistance diameter: $R_{\mathbf{G}} \leq 5/\ell$
- Perceptron: $M \leq 42$ (independent of $\ell$ and $m$)
- Does this generalize to multiple clusters? No!

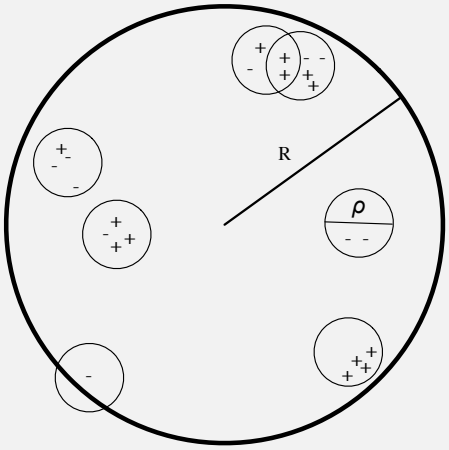
# Problem: 3 Clusters



*3 clusters (one in isolation)*

- Three $m$-cliques
- Two $m$-cliques with $\ell$ edges ($cm < \ell < m$) between cliques
- An "isolated" clique with $\Theta(1)$ outgoing edges.
- Norm: $\|\boldsymbol{u}\|^2 = \Theta(\ell)$, Resistance diameter: $R_{\mathbf{G}} = \Theta(1)$
- Problem: perceptron: $M \leq \Theta(\ell)$ (dependent on $\ell$)

# Pounce Bound Motivation




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

- Bounds to be dependent on structure of input space $X$.
- Novikoff is only dependent on $X$ through radius $R$.
- Expectation is that a typical ambient input space is only sparsely populated (cf manifold/cluster hypotheses).
- Pounce will depend on the cover of $X$.
- In particular the number of balls $\mathcal{N}(X, \rho)$ of diameter $\rho$.

## Pounce Algorithm

**Notation:** $\mathcal{V}_{\mathbf{G}} := \{\boldsymbol{v}_i := \boldsymbol{e}_i^\top \mathbf{G}^+ : i \in \{1, 2, \ldots, n\}\}$
**Input:** $\{(\boldsymbol{v}_{i_t}, y_t)\}_{t=1}^{\ell} \subseteq \mathcal{V}_{\mathbf{G}} \times \{-1, 1\}$.
**Initialization:** $\boldsymbol{w}_2 = \mathbf{0}$; $\mathcal{M} = \{1\}$.
**for** $t = 2, \ldots, \ell$ **do**
  **Receive:** $i_t \in \{1, \ldots, n\}$
    $\eta_t = \arg\min_{j \in \mathcal{M}} \|\boldsymbol{v}_{i_t} - \boldsymbol{v}_{i_j}\|$
  **Predict:** $\hat{y}_t = \text{sign}(y_{\eta_t} + \boldsymbol{w}_t(i_t) - \boldsymbol{w}_t(i_{\eta_t}))$
  **Receive:** $y_t$
    **if** $\hat{y}_t = y_t$ **then**
      $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$
    **else**
      $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \frac{y_t - y_{\eta_t} - (\boldsymbol{w}_t(i_t) - \boldsymbol{w}_t(i_{\eta_t}))}{\|\boldsymbol{v}_{i_t} - \boldsymbol{v}_{i_{\eta_t}}\|^2}(\boldsymbol{v}_{i_t} - \boldsymbol{v}_{i_{\eta_t}})$
      $\mathcal{M} = \mathcal{M} \cup \{t\}$
**end**

## Pounce Bound

### Theorem

The mistakes $M$ of POUNCE are bounded by

$$M \leq \mathcal{N}(X, \rho) + \|\boldsymbol{u}\|^2 \rho + 1 \ ,$$

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

$$\boldsymbol{u}(i_t) y_t \geq 1$$

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

- **Definition:** $\mathcal{N}(X, \rho)$ is the minimum number of balls of squared diameter $\rho$ that cover $X$.
- **Three Clusters:** $|\mathcal{M}| \leq 20$ with ($\rho = \frac{2}{m-1}$, $\mathcal{N}(X, \rho) = 3$, and $\|\boldsymbol{u}^*\|^2 < 8\ell$)

# Application: Heat Kernel

The Heat Kernel is used to build Laplacian from data.

## Discrepancy Function

- $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$
- $d(x, y) = d(y, x)$
- $(d(x, y) = 0) \iff (x = y)$

## Heat Kernel Laplacian

Given $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, discrepancy $d$, and $a > 0$ then

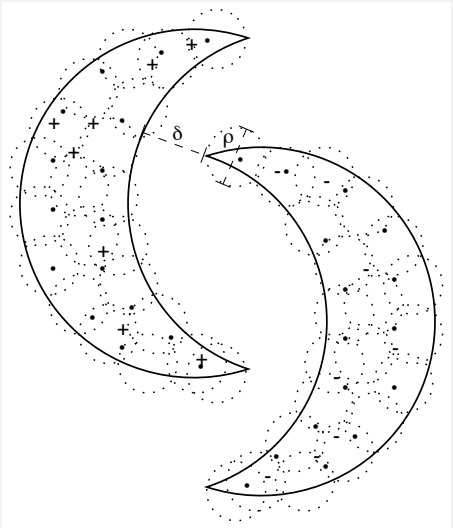
$$G^a_{ij} := \begin{cases} -e^{-ad(x_i,x_j)} & i \neq j \\ \sum_{k \neq i}^{n} e^{-ad(x_i,x_k)} & i = j \end{cases}$$

$a$ is a scale parameter

# (Component) Separating Cover

## Definitions

- $\mathcal{N}^\circ(\mathcal{X}, \mathcal{Y}, d)$ is the *separating cover number* the minimal number of balls of diameter $\rho$ covering $\mathcal{X}$ with $\mathcal{Y}(x) \neq \mathcal{Y}(x') \to d(x, x') > \rho$
- $x, x' \in X$ are *$\rho$-path-connected* if $d(x, x') \leq \rho$ or $\exists x''$ such that $d(x, x'') \leq \rho$ and $x'', x'$ are $\rho$-path-connected.
- $\mathcal{C}^\circ(X, \mathcal{Y}, d)$ is the *component-separating cover number* ... $\rho$-path-connected sets ... $\mathcal{Y}(x) \neq \mathcal{Y}(x') \to d(x, x') > \rho$.




$\mathcal{N}^\circ(\mathcal{X}, \mathcal{Y}, d) = 42$
$\mathcal{C}^\circ(X, \mathcal{Y}, d) = 2$

# Heat Kernel Bound

**Theorem**

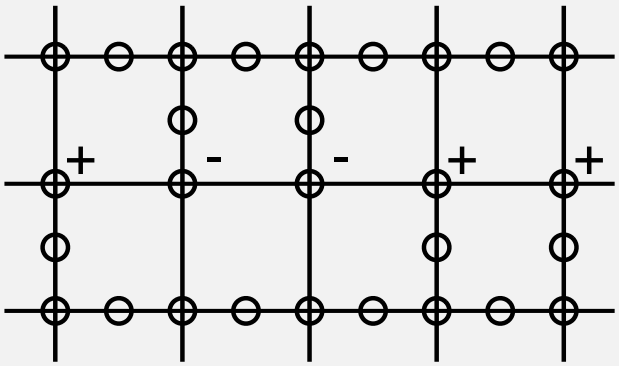
There exists an $a' > 0$ such that for all $a > a'$ the mistakes $M$ of the POUNCE algorithm with Laplacian $\mathbf{G}^a$ are bounded by

$$M \leq \mathcal{C}^\circ(X, \mathcal{Y}, d) + 1 \leq \mathcal{N}^\circ(\mathcal{X}, \mathcal{Y}, d) + 1 \quad (*)$$

Observe

- As $X$ increases potentially $\mathcal{C}^\circ(X, \mathcal{Y}, d) \ll \mathcal{N}^\circ(\mathcal{X}, \mathcal{Y}, d)$

# The Value of Unlabeled Data




*Two Path connected components*

- Center row of $m$ labels is the task.
- Center row is labeled randomly.
- Expected mistakes of any algorithm not using unlabeled data is $M = \frac{m}{2}$.
- Aligned unlabeled data $\longrightarrow$ POUNCE's bound is $M \leq 3$.
- Example implies a limitation in the type bounds provable for non-transductive algorithms.

## Thanks

Thank You!