Some applications in human behavior modeling (and open questions)

Jerry Zhu

University of Wisconsin-Madison

Simons Institute Workshop on Spectral Algorithms:
From Theory to Practice
Oct. 2014

Outline

Human Memory Search

Machine Teaching

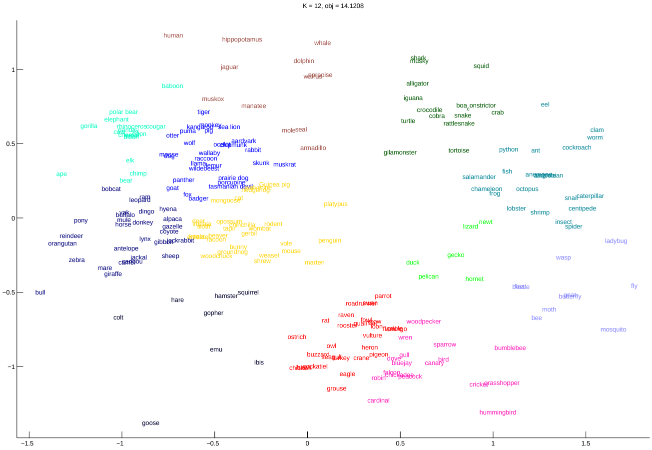
Verbal fluency

Say as many animals as you can without repeating in one minute.

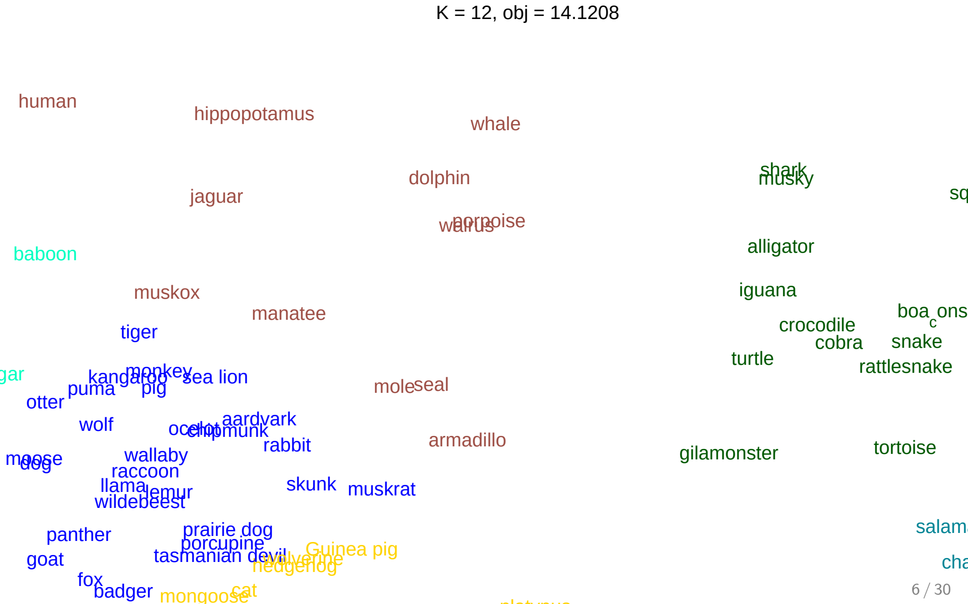
Semantic "runs"

1. cow, horse, chicken, pig, elephant, lion, tiger, porcupine, gopher, rat, mouse, duck, goose, horse, bird, pelican, alligator, crocodile, iguana, goose
2. elephant, tiger, dog, cow, horse, sheep, cat, lynx, elk, moose, antelope, deer, tiger, wolverine, bobcat, mink, rabbit, wolf, coyote, fox, cow, zebra
3. cat, dog, horse, chicken, duck, cow, pig, gorilla, giraffe, tiger, lion, ostrich, elephant, squirrel, gopher, rat, mouse, gerbil, hamster, duck, goose
4. cat, dog, sheep, goat, elephant, tiger, dog, deer, lynx, wolf, mountain goat, bear, giraffe, moose, elk, hyena, aardvark, platypus, lion, skunk, wolverine, raccoon
5. dog, cat, leopard, elephant, monkey, sea lion, tiger, leopard, bird, squirrel, deer, antelope, snake, beaver, robin, panda, vulture
6. deer, muskrat, bear, fish, raccoon, zebra, elephant, giraffe, cat, dog, mouse, rat, bird, snake, lizard, lamb, hippopotamus, elephant, skunk, lion, tiger

Memory search



Memory search



Censored random walk

[Abbott, Austerweil, Griffiths 2012]

- $\mathcal{V}$: n animal word types in English
- P: (dense) $n \times n$ transition matrix
- Censored random walk: observing only the first token of each type $x_1, x_2, \ldots, x_t, \ldots \Rightarrow a_1, \ldots, a_n$
- (star example)

The estimation problem

Given m censored random walks
$\mathcal{D} = \left\{ \left(a_1^{(1)}, ..., a_n^{(1)} \right), ..., \left(a_1^{(m)}, ..., a_n^{(m)} \right) \right\}$, estimate P.

Each observed step is an absorbing random walk

(with Kwang-Sung Jun)

- $P(a_{k+1} \mid a_1, \ldots, a_k)$ may contain infinite latent steps
- Instead, model this observed step as an absorbing random walk with absorbing states $\mathcal{V} \setminus \{a_1, \ldots, a_k\}$
- $P \Rightarrow \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$

Maximum Likelihood

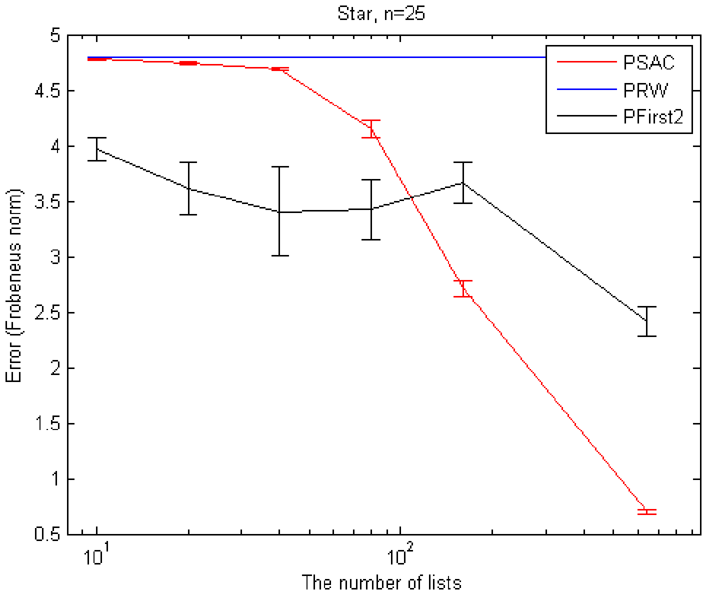
- Fundamental matrix $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$, N_{ij} is the expected number of visits to j before absorption when starting at i
- $p(a_{k+1} \mid a_1, \ldots, a_k) = \sum_{i=1}^{k} N_{ki} R_{i1}$
- log likelihood $\sum_{i=1}^{m} \sum_{k=1}^{n} \log p(a_{k+1}^{(i)} \mid a_1^{(i)}, ..., a_k^{(i)})$
- Nonconvex, gradient method

Other estimators

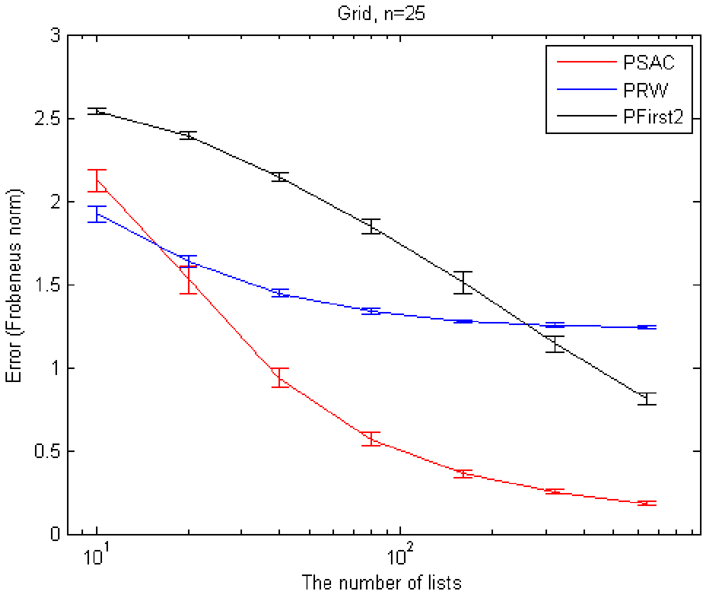
- PRW: pretend $a_1, \ldots, a_n$ not censored
- PFirst2: Use only a_1, a_2 in each walk (consistent)

Star graph

x-axis: m, y-axis: $\|\widehat{P} - P\|_F^2$

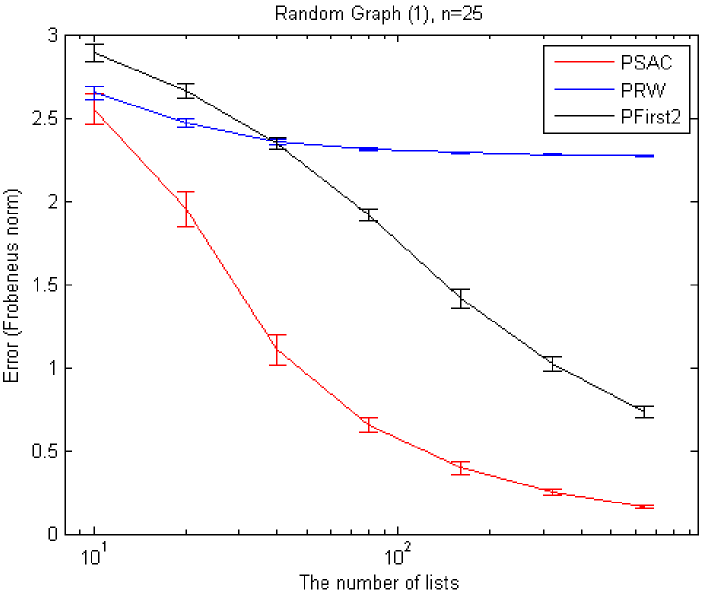


2D Grid

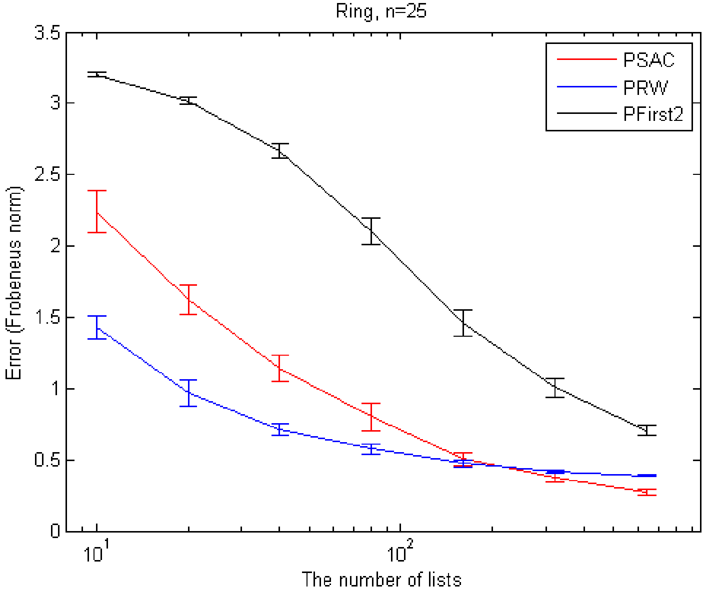


Grid, n=25
PSAC
PRW
PFirst2
Error (Frobeneus norm)
The number of lists
3
2.5
2
1.5
1
0.5
0
10^1
10^2

Erdös-Rényi with $p = \log(n)/n$



Ring graph



Ring, n=25
PSAC
PRW
PFirst2
Error (Frobeneus norm)
The number of lists
3.5
3
2.5
2
1.5
1
0.5
0
10^1
10^2

Questions

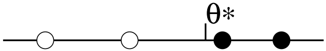
- Consistency?
- Rate?

Outline

Human Memory Search

Machine Teaching

Learning a noiseless 1D threshold classifier

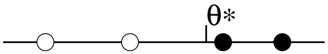


$$x \sim \text{uniform}[0,1]$$

$$y = \begin{cases} -1, & x < \theta^* \\ 1, & x \geq \theta^* \end{cases}$$

$$\Theta = \{\theta : 0 \leq \theta \leq 1\}$$

Learning a noiseless 1D threshold classifier



$$x \sim \text{uniform}[0,1]$$

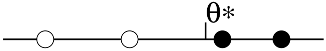
$$y = \begin{cases} -1, & x < \theta^* \\ 1, & x \geq \theta^* \end{cases}$$

$$\Theta = \{\theta : 0 \leq \theta \leq 1\}$$

Passive learning:

1. given training data $D = (x_1, y_1) \ldots (x_n, y_n) \stackrel{iid}{\sim} p(x, y)$
2. finds $\hat{\theta}$ consistent with D

Learning a noiseless 1D threshold classifier



$$x \sim \text{uniform}[0, 1]$$

$$y = \begin{cases} -1, & x < \theta^* \\ 1, & x \geq \theta^* \end{cases}$$

$$\Theta = \{\theta : 0 \leq \theta \leq 1\}$$

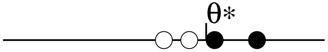
Passive learning:

1. given training data $D = (x_1, y_1) \ldots (x_n, y_n) \overset{iid}{\sim} p(x, y)$
2. finds $\hat{\theta}$ consistent with D

Risk $|\hat{\theta} - \theta^*| = O(n^{-1})$

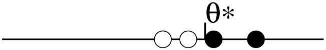
Active learning (sequential experimental design)

Binary search



Active learning (sequential experimental design)

Binary search



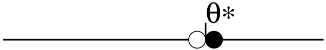
Risk $|\hat{\theta} - \theta^*| = O(2^{-n})$

Machine teaching

What is the minimum training set a helpful teacher can construct?

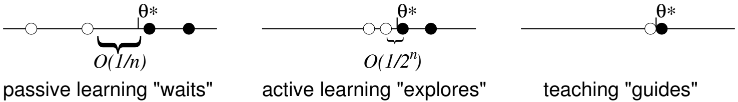
Machine teaching

What is the minimum training set a helpful teacher can construct?



Risk $|\hat{\theta} - \theta^*| = \epsilon, \forall \epsilon > 0$

Comparing the three



The teacher knows θ^* and the learning algorithm.

Example 2: Teaching a Gaussian distribution

Given a training set $x_1 \dots x_n \in \mathbb{R}^d$, let the learning algorithm be

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top$$

How to teach $N(\mu^*, \Sigma^*)$ to the learner quickly?

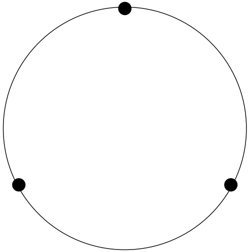
Example 2: Teaching a Gaussian distribution

Given a training set $x_1 \dots x_n \in \mathbb{R}^d$, let the learning algorithm be

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

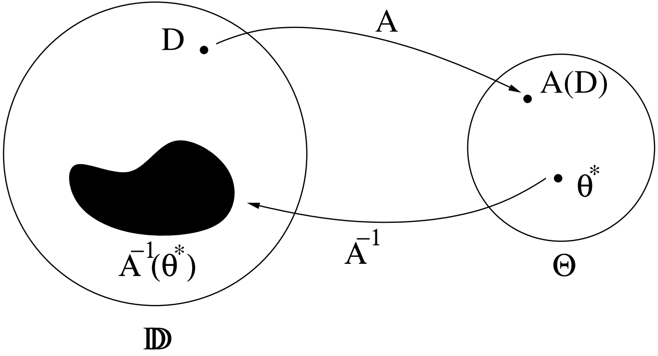
$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top$$

How to teach $N(\mu^*, \Sigma^*)$ to the learner quickly?



non-iid, $n = d + 1$

An Optimization View of Machine Teaching



$$\min_{D \in \mathbb{D}} \quad |D|$$
$$\text{s.t.} \quad A(D) = \theta^*$$

The objective is the teaching dimension [Goldman, Kearns 1995] of θ^* with respect to A, Θ.

Example 3: Works on Humans

Human categorization on 1D stimuli [Patil, Z, Kopeć, Love 2014]

human training set	human test accuracy
machine teaching	72.5%
iid	69.8%

(statistically significant)

New task: teaching humans how to label a graph

Given:

- a graph $G = (V, E)$
- target labels $y^* : V \mapsto \{-1, 1\}$
- a label-completion cognitive model A (graph diffusion algorithm) such as:
  - mincut
  - harmonic function [Z, Gharahmani, Lafferty 2003]
  - local global consistency [Zhou et al. 2004]
  - $\ldots$

Find the smallest seed set:

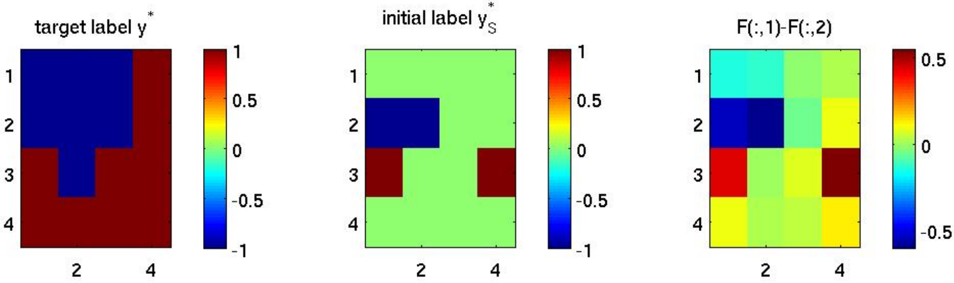
$$\begin{aligned} \min_{S \subseteq V} \quad & |S| \\ \text{s.t.} \quad & A(y^*(S)) = y^* \end{aligned}$$

(inverse problem of semi-supervised learning)

Example: $A =$ local-global consistency [Zhou et al. 2004]

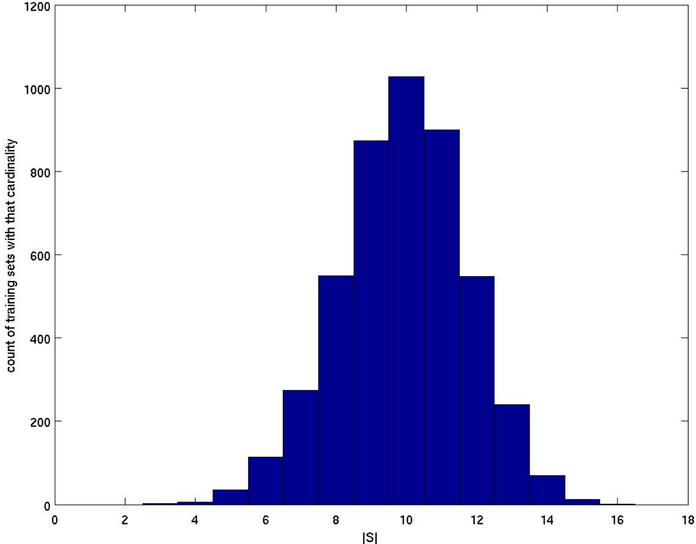
$$F = (1-\alpha)(I - \alpha D^{-1/2} W D^{-1/2})^{-1} y^*(S)$$

$$y = \text{sgn}\left(F \begin{pmatrix} 1 \\ -1 \end{pmatrix}\right)$$

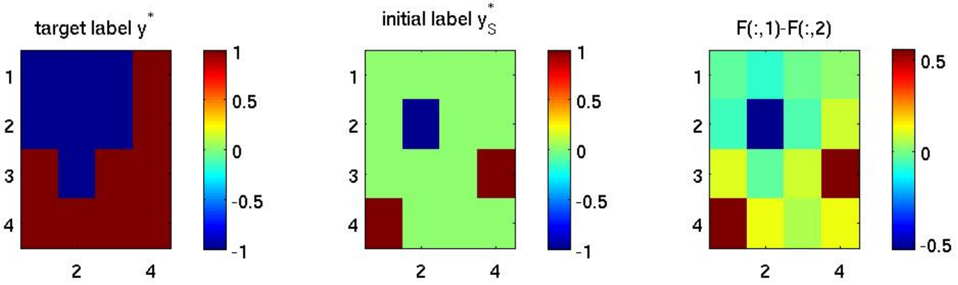


$|A^{-1}(y^*)|$ is large

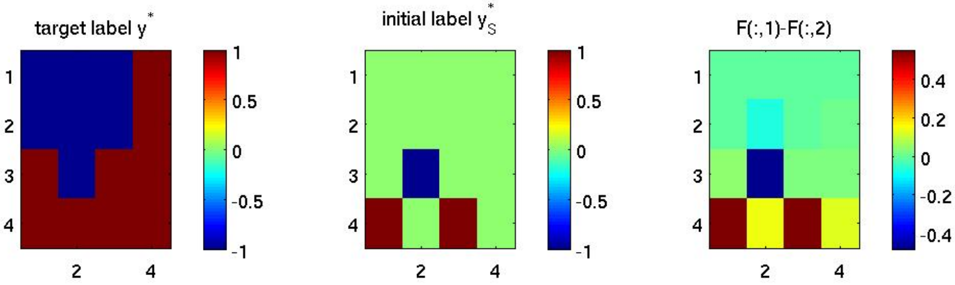
Turns out 4649 out of $2^{16} = 65536$ training sets S lead to the target label completion



Optimal seed set 1: $|S| = 3$



Optimal seed set 2: $|S| = 3$



Questions

- How does $|S|$ relate to (spectral) properties of G?
- How to solve for S?