

Graph-based and Bayesian approaches to Semi-supervised Learning

Zoubin Ghahramani

**Department of Engineering
University of Cambridge, UK**

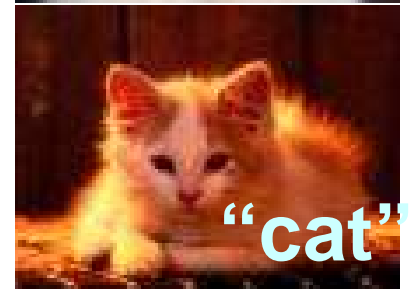
**Machine Learning Department
Carnegie Mellon University, USA**

`zoubin@eng.cam.ac.uk`

`http://learning.eng.cam.ac.uk/zoubin/`

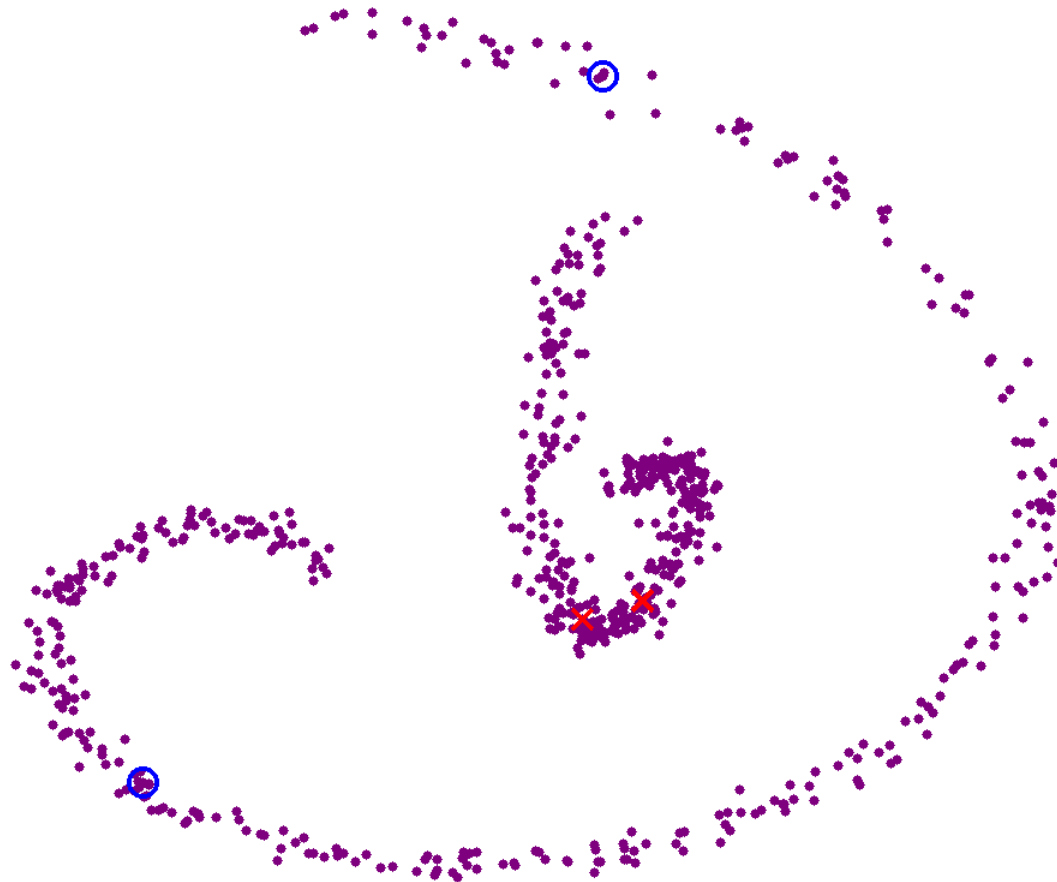
**IPAM Probabilistic Models of Cognition
Lectures July 2007**

Poverty of the Stimulus



Classification using Unlabelled Data

Assumption: there is information in the data distribution

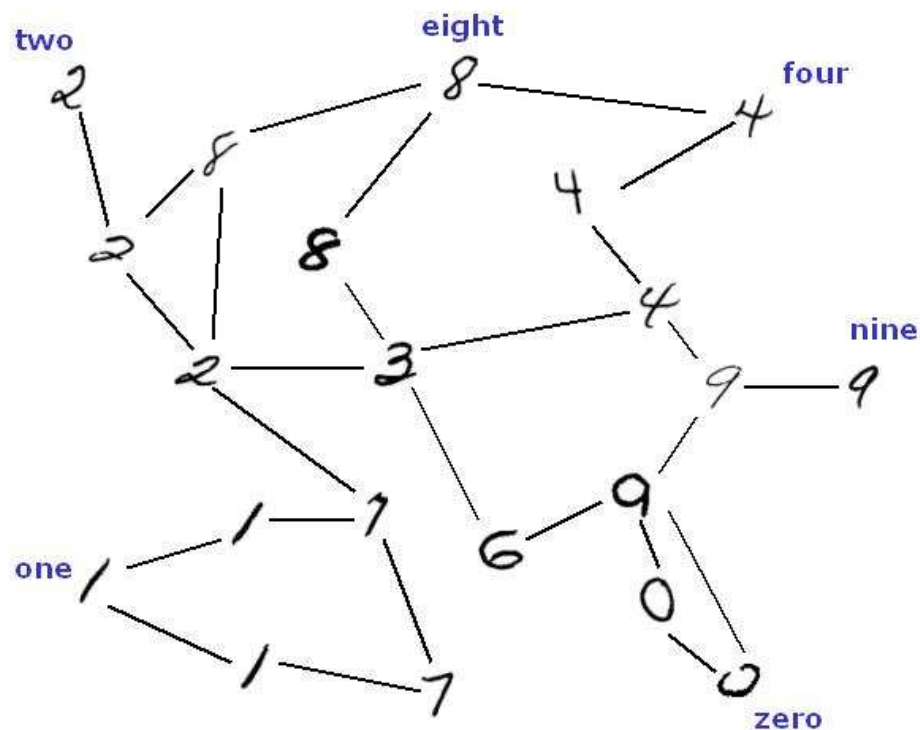


Outline

- Graph-based semi-supervised learning
- Some thoughts on fully Bayesian semi-supervised learning

Graph-based Semi-supervised Learning

Labeled and Unlabeled Data as a Graph



- Idea: Construct a random field on graph
- Intuition: Similar examples have similar labels
- Information “propagates” from labeled examples
- Graph encodes prior intuition

Work with Xiaojin Zhu (U Wisconsin) and John Lafferty (CMU)

The Graph

- **nodes**: instances in $L \cup U$. Binary labels $\mathbf{y} \in \{0, 1\}^n$
- **edges**: **local similarity**. $n \times n$ symmetric weight matrix W assumed **given**.
- **energy**: $E(\mathbf{y}) = \frac{1}{2} \sum_{i,j} w_{ij} (y_i - y_j)^2$

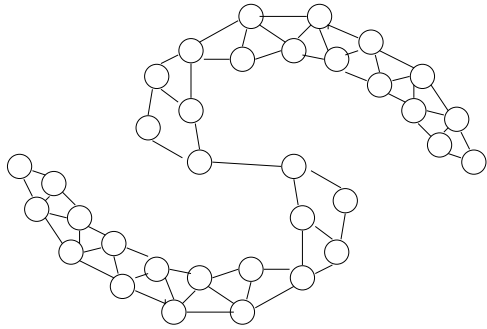


happy, low energy



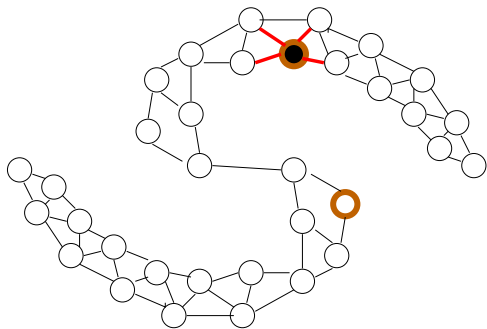
unhappy, high energy

Low energy \rightarrow Label Propagation

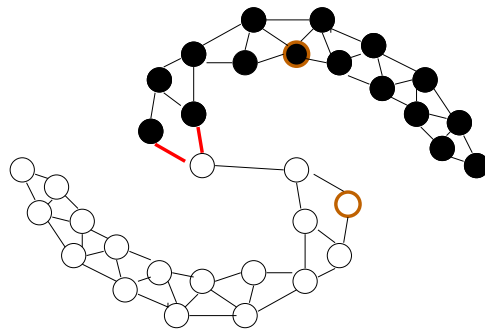


energy=0

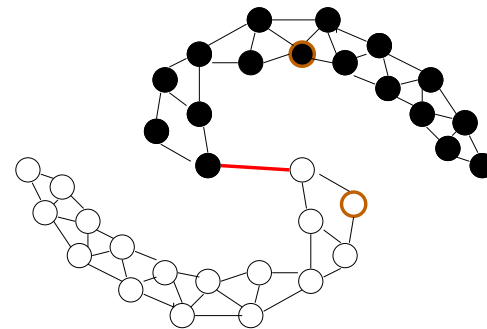
Conditioned on labeled data:



energy=4



energy=2



energy=1

Discrete Markov Random Fields

$$p(\mathbf{y}) \propto \exp(-E(\mathbf{y})) \mid_{\mathbf{y}_L=L}$$
$$y_i \in \{0, 1\}$$

[Zhu & Ghahramani 02]

We relaxed this to a
Gaussian random fields

Discrete Markov Random Fields, revisited

$$p(\mathbf{y}) \propto \exp(-E(\mathbf{y})) \mid_{\mathbf{y}_L=L}$$
$$y_i \in \{0, 1\}$$

Gaussian Random Fields

$$p(\mathbf{y}) \propto \exp(-E(\mathbf{y})) \mid_{\mathbf{y}_L=L}$$

$$y_i \in \mathbb{R}$$

The Laplacian

$$W = \begin{bmatrix} w_{11} & \dots & w_{1n} \\ & \dots & \\ w_{n1} & \dots & w_{nn} \end{bmatrix} \quad D = \begin{bmatrix} \sum w_{1\cdot} & & \mathbf{0} \\ & \dots & \\ \mathbf{0} & & \sum w_{n\cdot} \end{bmatrix}$$

The Laplacian $\Delta = D - W$

$$\Delta = \left[\begin{array}{c|c} \Delta_{LL} & \Delta_{LU} \\ \hline \Delta_{UL} & \Delta_{UU} \end{array} \right]$$

Gaussian Random Fields

$$\begin{aligned} p(\mathbf{y}) &\propto \exp(-E(\mathbf{y})) \mid_{\mathbf{y}_L=L} \\ &= \exp\left(-\frac{1}{2} \sum_{i,j} w_{ij} (y_i - y_j)^2\right) \mid_{\mathbf{y}_L=L} \\ &= \exp(-\mathbf{y}^\top \Delta \mathbf{y}) \mid_{\mathbf{y}_L=L} \end{aligned}$$

The field is Gaussian: $\mathbf{y}_U \sim \mathcal{N}(f_U, \frac{1}{2}(\Delta_{UU})^{-1})$

The mean is $f_U = -(\Delta_{UU})^{-1} \Delta_{UL} \mathbf{y}_L$

The Mean f_U

The mean $f_U \equiv$ mode of Gaussian Random Field
 \equiv min energy state

- “soft labels”, unique
- harmonic

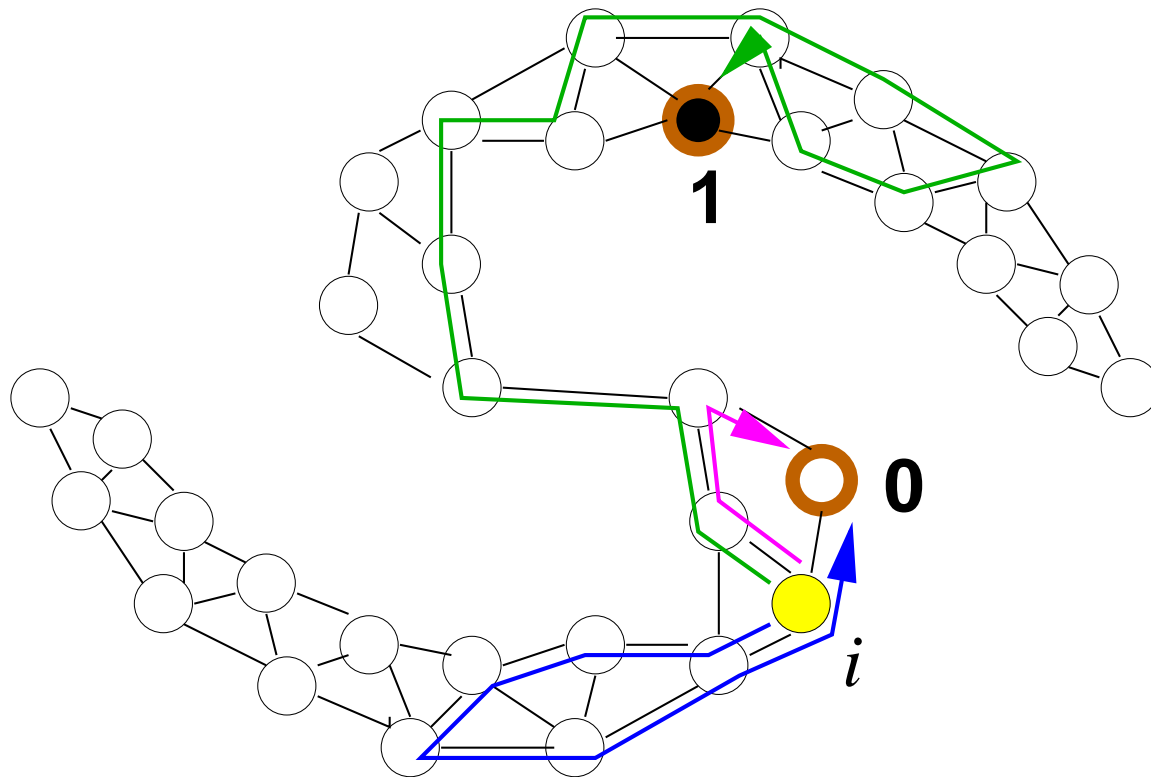
$$\Delta \mathbf{f} = 0 \quad \text{or} \quad f_i = \frac{\sum_{j \sim i} w_{ij} f_j}{\sum_{j \sim i} w_{ij}}, \quad i \in U$$
$$0 < f_i < 1$$

- Related to heat kernels etc. in spectral graph theory.

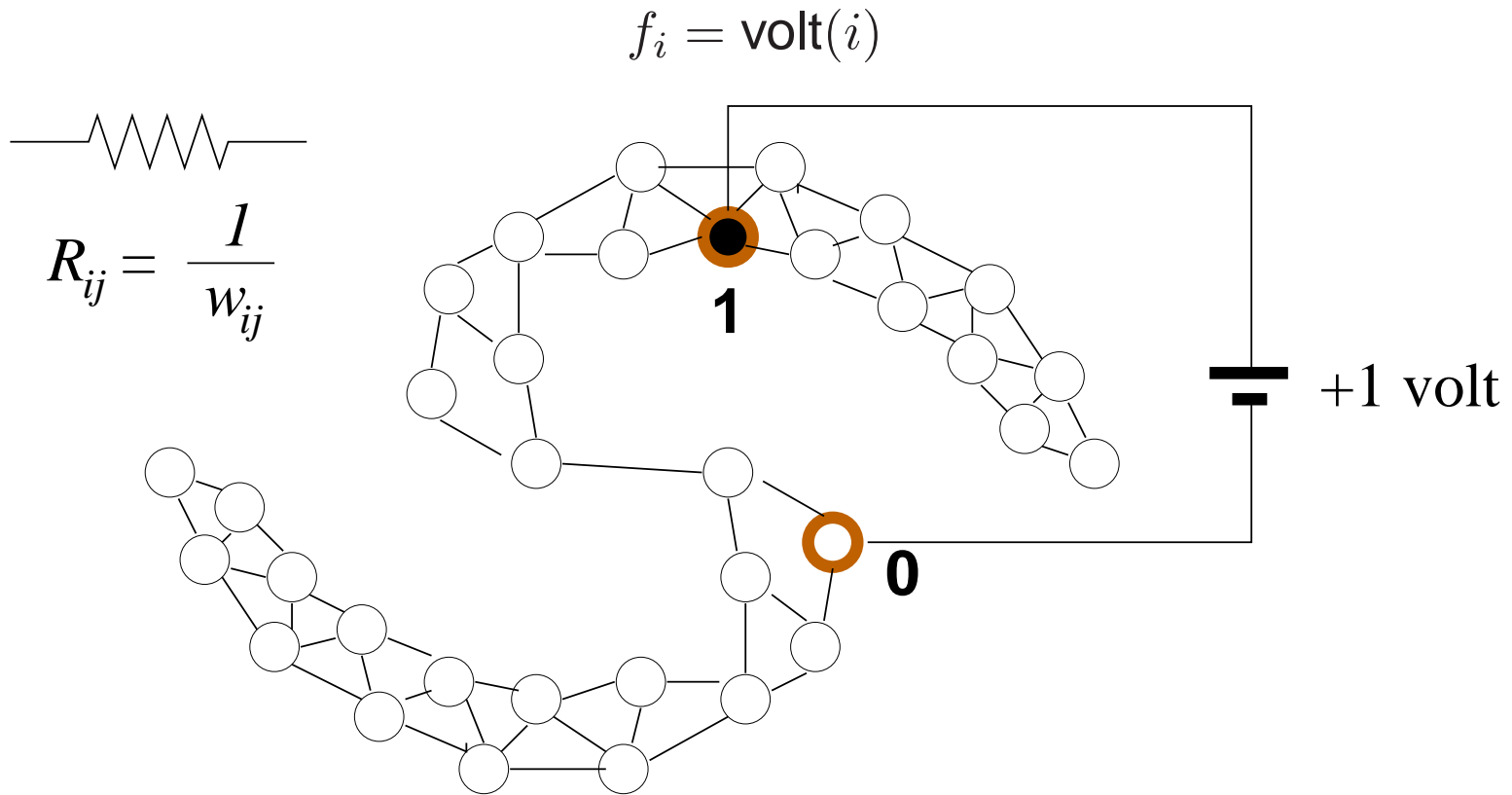
f_U Interpretation: Random Walks

$$P(j|i) = \frac{w_{ij}}{\sum_k w_{ik}}$$

$$f_i = P(\text{reach label 1} | \text{from } i)$$



f_U Interpretation: Electric Networks



Active Semi-Supervised Learning

[Zhu, Lafferty, Ghahramani, 2003]

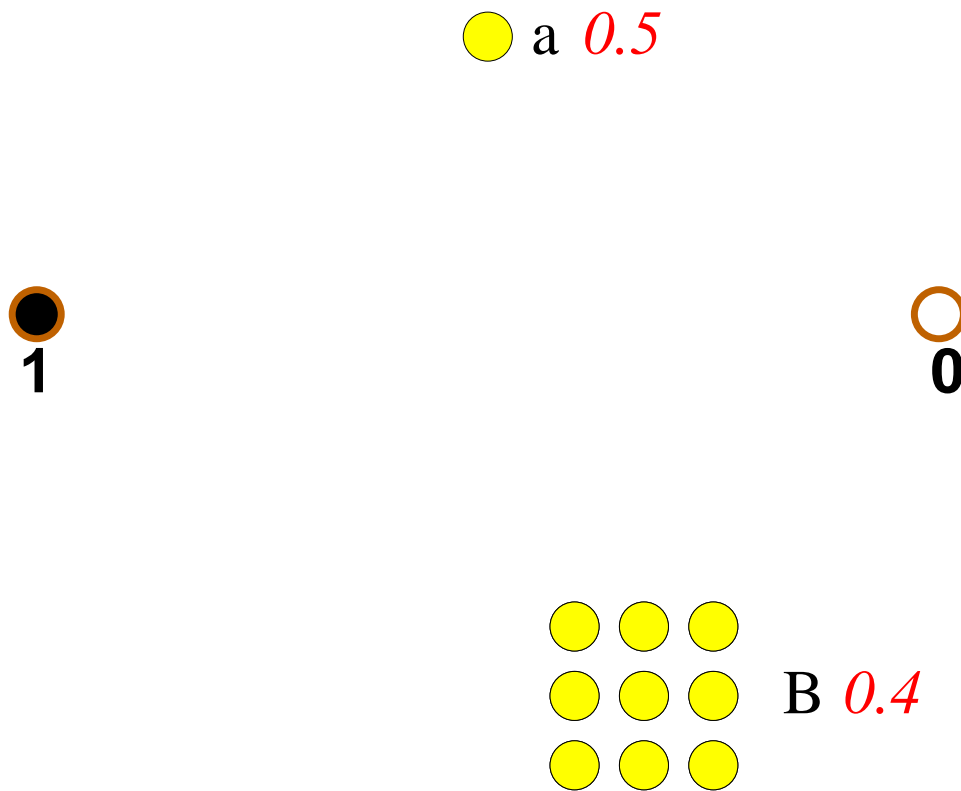
Semi-supervised learning uses U to help classification.

Active learning (pool based) selects queries in U to ask for labels.

Put it together, we have a better query selection criterion than naively selecting the point with maximum label ambiguity.

Active Learning

Select a query to **minimize the estimated generalization error**, not by maximum ambiguity.



Active Learning

generalization error

$$\text{err} = \sum_{i \in U} \sum_{y_i=0,1} (\text{sgn}(f_i) \neq y_i) P_{\text{true}}(y_i)$$

approximation

$$P_{\text{true}}(y_i = 1) \leftarrow f_i$$

estimated generalization error

$$\hat{\text{err}} = \sum_{i \in U} \min(f_i, 1 - f_i)$$

Active Learning

estimated generalization error **after querying** x_k **and receiving label** y_k

$$\hat{\text{err}}^{+(x_k, y_k)} = \sum_{i \in U} \min \left(f_i^{+(x_k, y_k)}, 1 - f_i^{+(x_k, y_k)} \right)$$

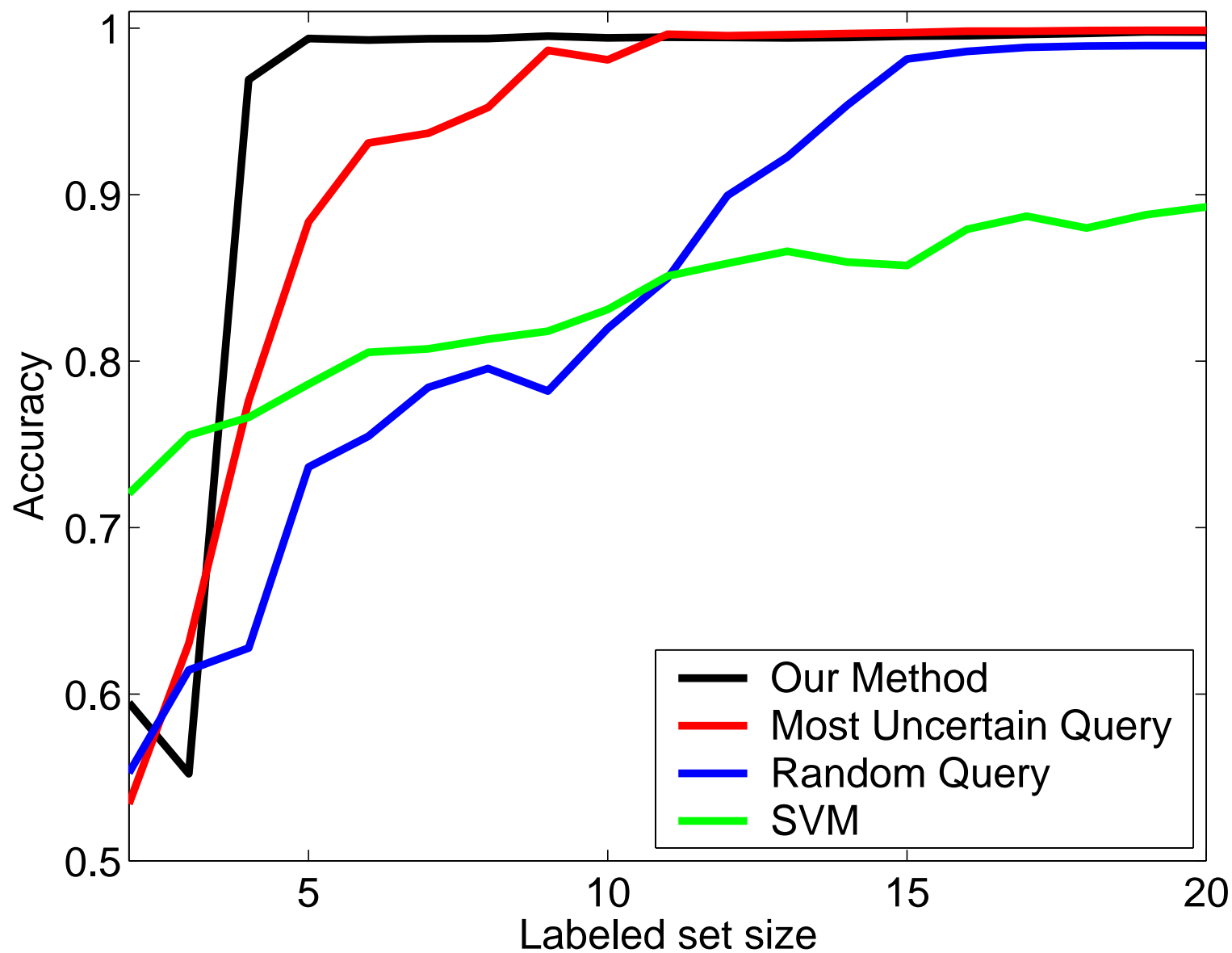
‘re-train’ is fast for the harmonic function

$$f_U^{+(x_k, y_k)} = f_U + (y_k - f_k) \frac{(\Delta_{UU})_{\cdot k}^{-1}}{(\Delta_{UU})_{kk}^{-1}}$$

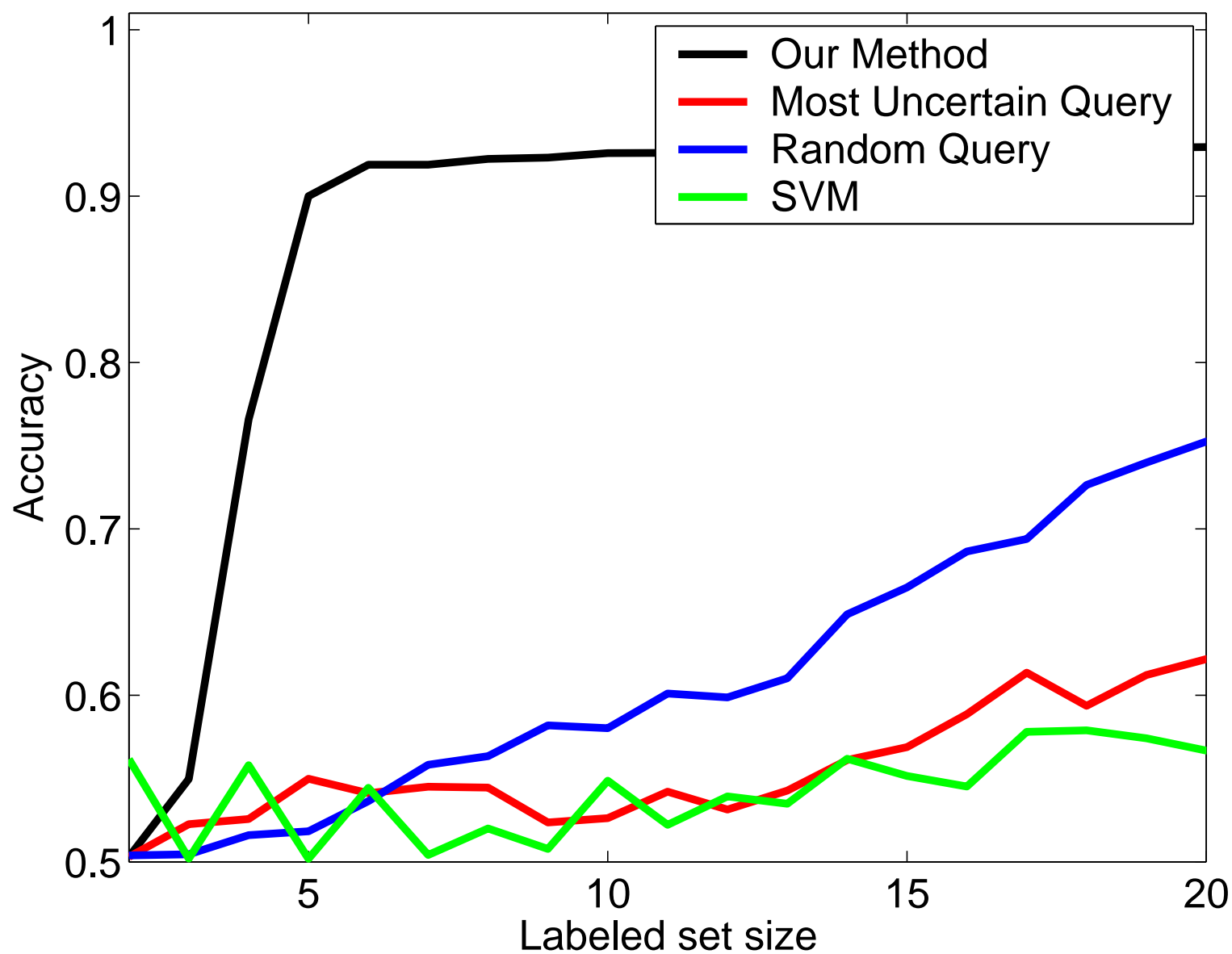
select query k^* s.t.

$$k^* = \arg \min_k (1 - f_k) \hat{\text{err}}^{+(x_k, 0)} + f_k \hat{\text{err}}^{+(x_k, 1)}$$

OCR Digits “1” vs. “2” ($|L \cup U| = 2200$)



20 Newsgroups PC vs. MAC ($|L \cup U| = 1943$)



Part II: Some thoughts on Bayesian semi-supervised learning

Moving forward...

- We have good methods for transduction.
- But we don't seem to have a single unified Bayesian framework for inductive SSL.
- How would we view this problem from a fully Bayesian framework?

Bayesian Semi-Supervised Learning

x inputs, y labels:

$$p(x, y) = p(x)p(y|x) = p(y)p(x|y)$$

Usually we assume some model with parameters:

- Discriminative:

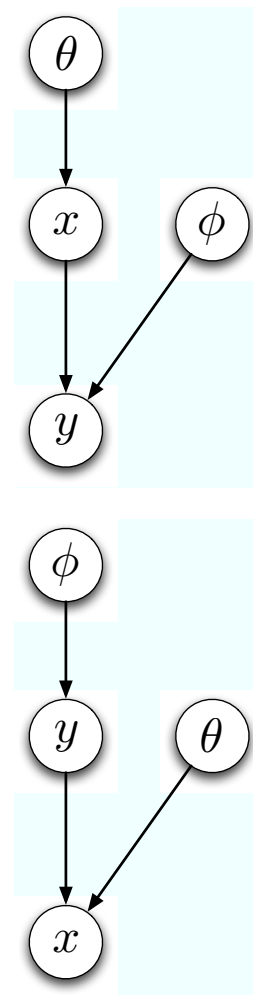
$$p(x, y|\theta, \phi) = p(x|\theta)p(y|x, \phi)$$

SSL possible if θ is somehow related to ϕ , works well when $p(y|x, \phi)$ is very flexible (e.g. non-parametric, kernel-based).

- Generative:

$$p(x, y|\theta, \phi) = p(y|\phi)p(x|y, \theta)$$

SSL possible but these methods are not currently widely used.



Bayesian Semi-Supervised Learning

Generative:

$$p(x, y | \theta, \phi) = p(y | \phi) p(x | y, \theta)$$

Limitations of the Generative approach:

- Often we don't *want* to model the full x .
(Solution: maybe we can model some features of x ?)
- Our models of $p(x | y, \theta)$ are usually too inflexible.
(Solution: use non-parametric methods?)

Some examples:

- Kemp et al (2003) Semi-supervised learning with trees.
- Radford Neal's entry using Dirichlet Diffusion trees into the NIPS feature selection competition.

From a Bayesian perspective, semi-supervised learning is just another missing data problem!

Summary

- Semi-supervised learning with harmonic functions
- Active semi-supervised learning using harmonic functions by minimizing expected generalization error
- Many open questions...

For a good recent survey see:

Xiaojin Zhu (2005) Semi-supervised Learning Literature Survey.

Appendix

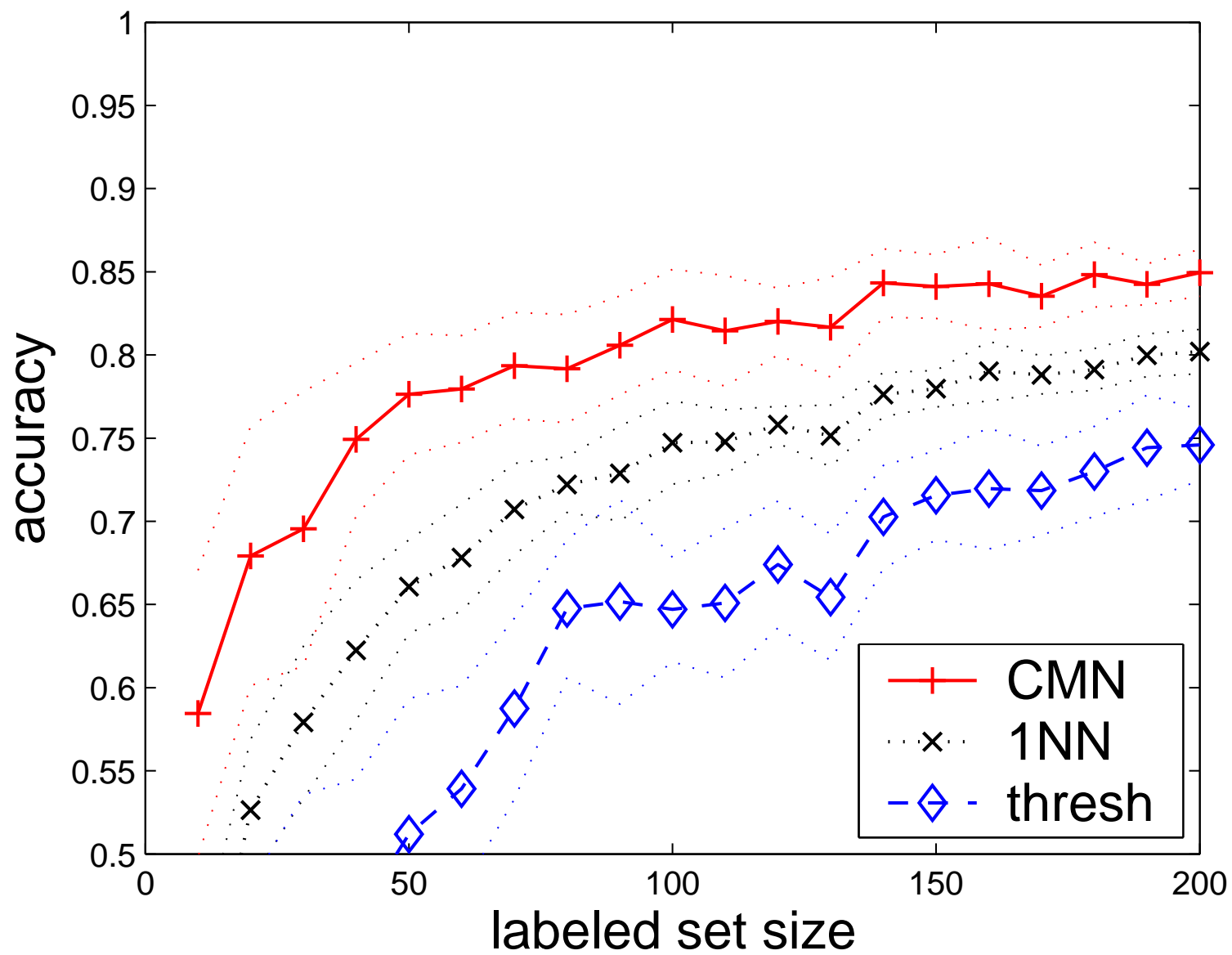
Classification

- naive: threshold f_U at 0.5. Classification often unbalanced.
- incorporating Class Priors ([heuristic](#))

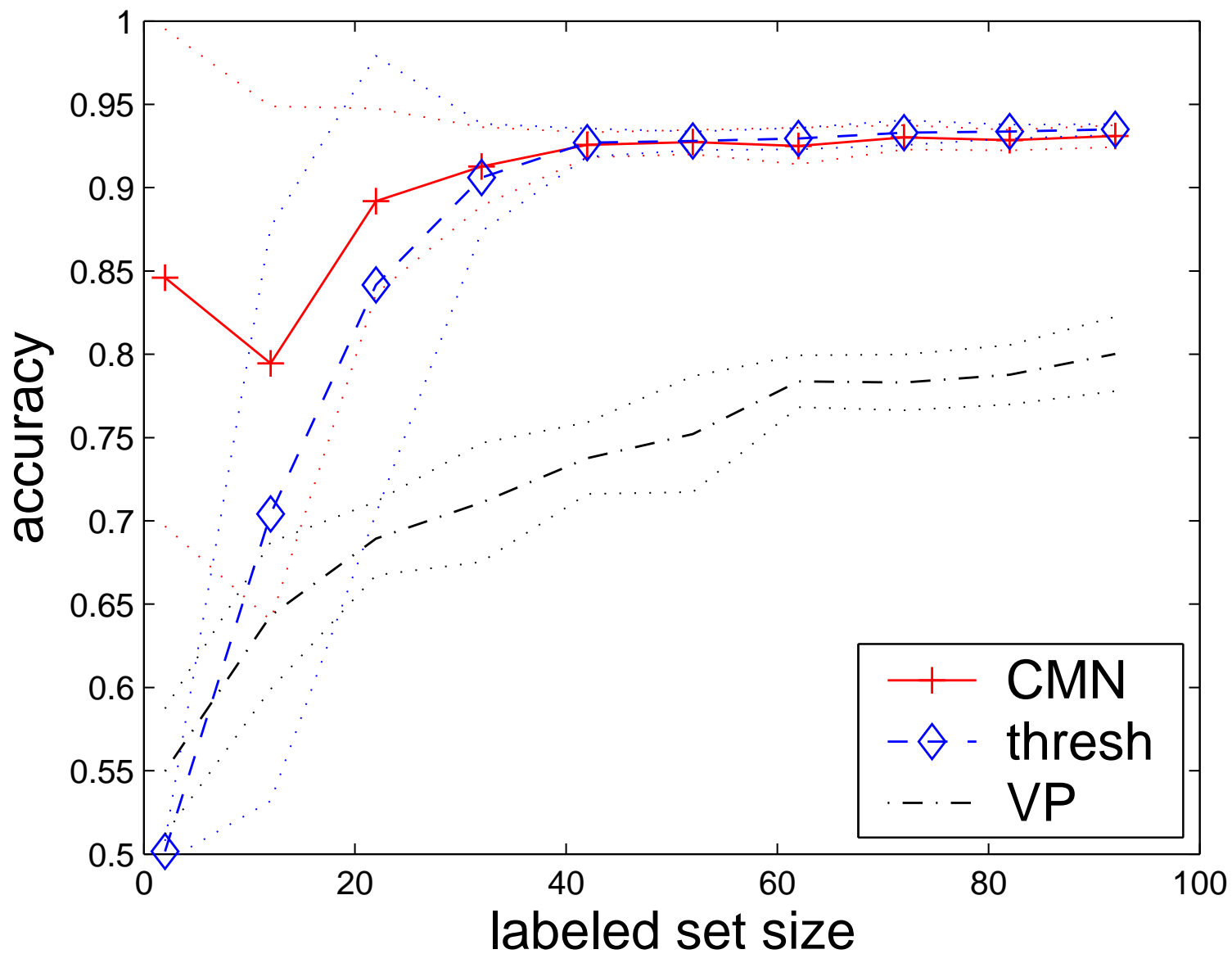
e.g. prior: 90% class 1

$$\begin{array}{ll} \text{minimize} & E(\mathbf{y}) = \mathbf{y}^\top \Delta \mathbf{y} \\ \text{subject to} & y_L = L \\ & \text{and } \frac{\sum f_U}{|U|} = 0.9 \end{array}$$

OCR Ten Digits ($|L \cup U| = 4000$)



20-Newsgroups (PC vs. MAC, $|L \cup U| = 1943$)



Threads?

Hyperparameter Learning

Learn the graph weights (or hyperparameters):

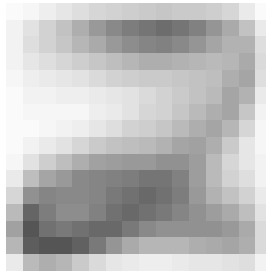
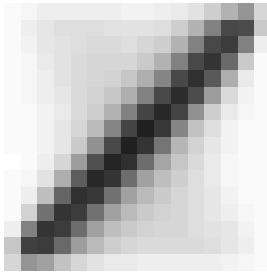
- $w_{ij} = \exp \left(-\sum_{d=1}^m \frac{(x_{id} - x_{jd})^2}{\sigma_d^2} \right)$, length scales;
- k NN unweighted graph, k ;
- ϵ NN unweighted graph, ϵ , etc.;

Hyperparameter Learning

- Minimize entropy on U (maximize label confidence);
- Evidence maximization with Gaussian process classifiers [tech report CMU-CS-03-175].

Hyperparameter Learning

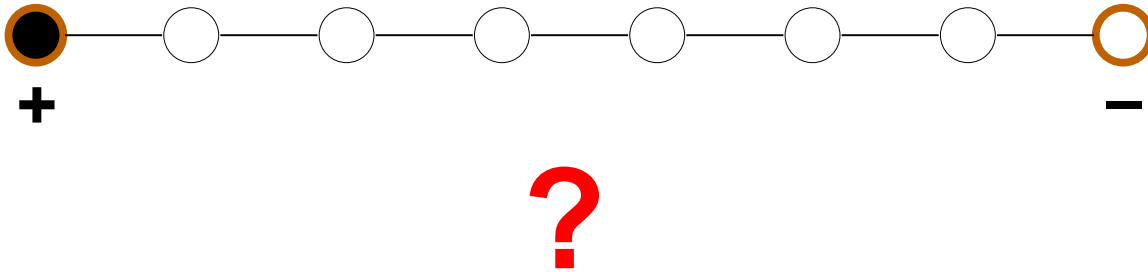
OCR Digits “1” vs. “2”, $|L| = 92$, $|U| = 2108$.



	H (bits)	GF acc
start	0.6931	$94.70 \pm 1.19 \%$
end	0.6542	$98.02 \pm 0.39 \%$

Graph Mincut

- graph mincut \equiv min energy \equiv discrete MRF mode
- Multiple minima, may be unbalanced



- Hard to compute when multi-class.

[Blum & Chawla 01]