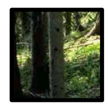
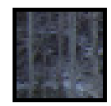
Multiple Instance Learning with Manifold Bags

Boris Babenko, Nakul Verma, Piotr Dollar, Serge Belongie

ICML 2011

Supervised Learning

- (example, label) pairs provided during training

( , +) ( , +) ( , -) ( , -)

Multiple Instance Learning (MIL)

- (**set of examples**, label) pairs provided
- MIL lingo: set of examples = **bag** of instances
- Learner does not see instance labels
- Bag labeled positive if at least one instance in bag is positive

[Dietterich et al. '97]

MIL Example: Face Detection

Instance: image patch

Instance Label: is face?

Bag: whole image

Bag Label: contains face?

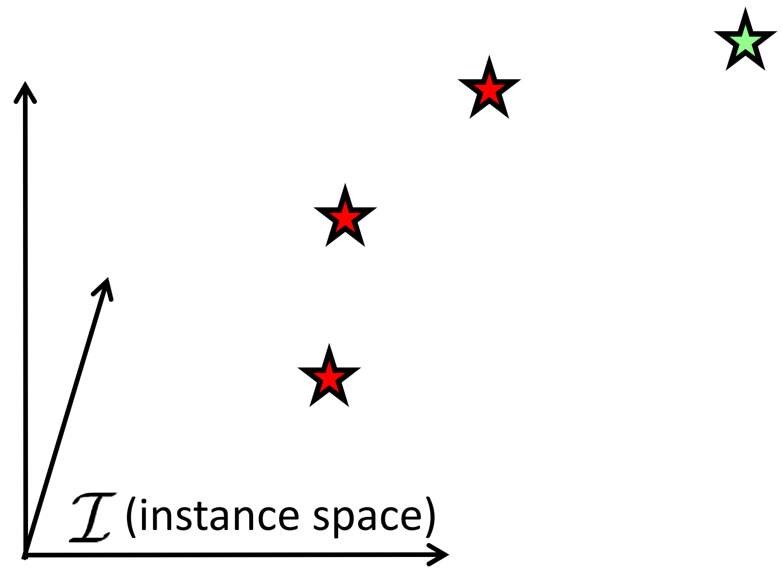
[Andrews et al. ’02, Viola et al. ’05, Dollar et al. 08, Galleguillos et al. 08]

PAC Analysis of MIL

- Bound bag **generalization** error in terms of **empirical** error
- Data model (bottom up)
  - Draw r instances and their labels from fixed distribution $\mathcal{D}_{\mathcal{I}}$
  - Create bag from instances, determine its label (max of instance labels)
  - Return bag & bag label to learner

Data Model

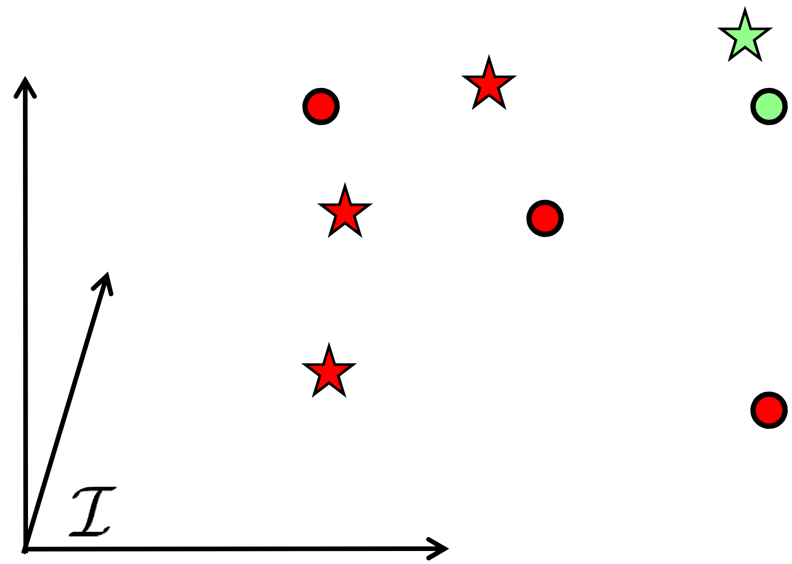
☆Bag 1: positive



Negative instance Positive instance

Data Model

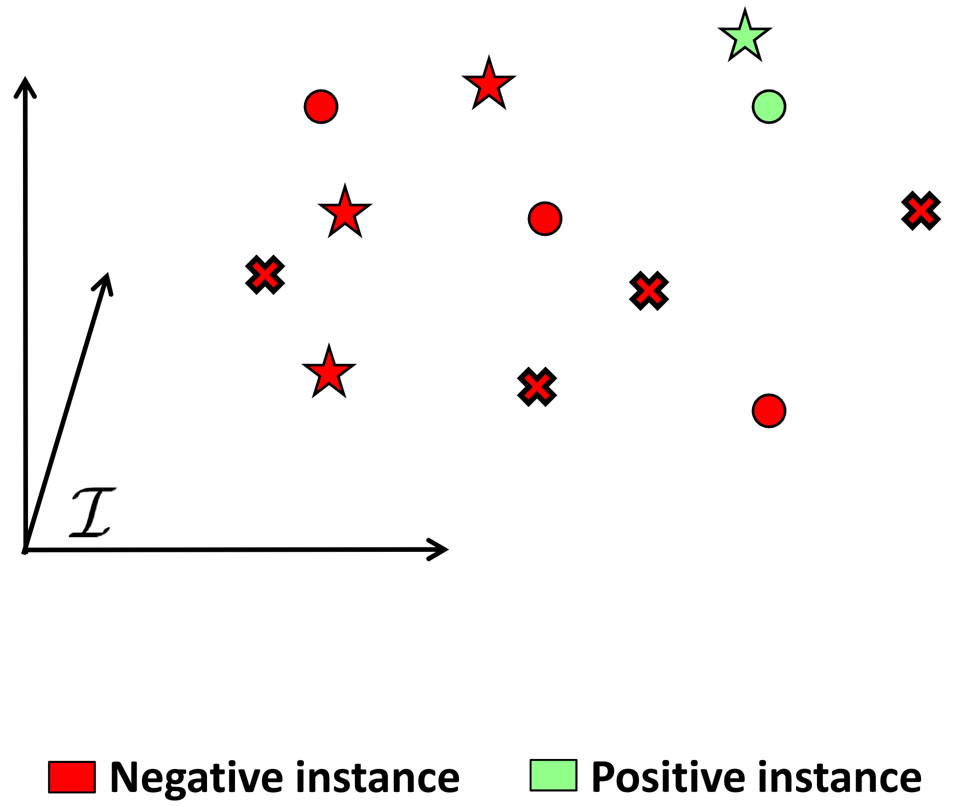
- Bag 2: positive



Negative instance Positive instance

Data Model

✖ **Bag 3: negative**



PAC Analysis of MIL

- Blum & Kalai (1998)
  - If: access to noise tolerant instance learner, instances drawn independently
  - Then: bag sample complexity **linear** in r
- Sabato & Tishby (2009)
  - If: can minimize empirical error on bags
  - Then: bag sample complexity **logarithmic** in r

MIL Applications

- Recently MIL has become popular in applied areas (vision, audio, etc)
- Disconnect between theory and many of these applications

MIL Example: Face Detection (Images)

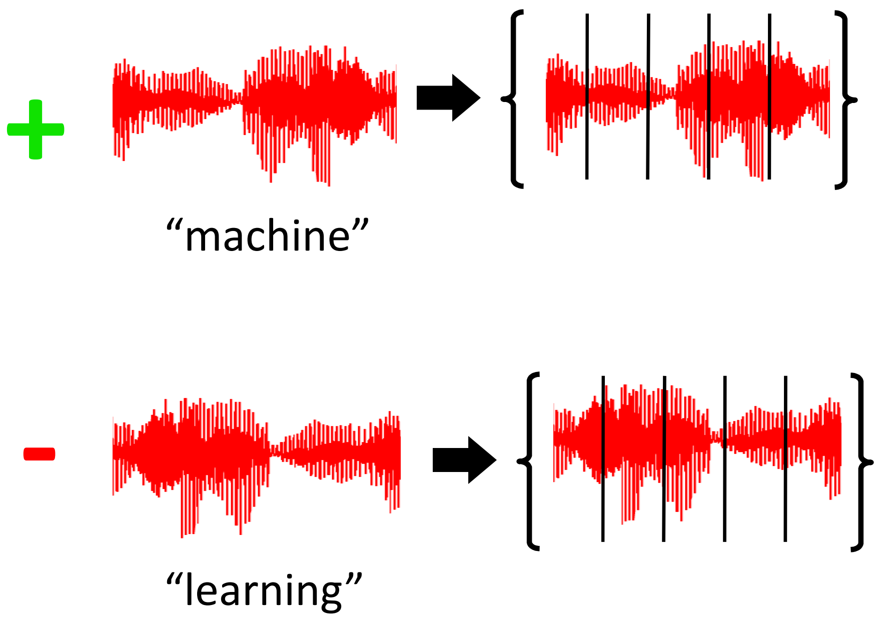
Bag: whole image

Instance: image patch

[Andrews et al. ’02, Viola et al. ’05, Dollar et al. 08, Galleguillos et al. 08]

MIL Example: Phoneme Detection (Audio)

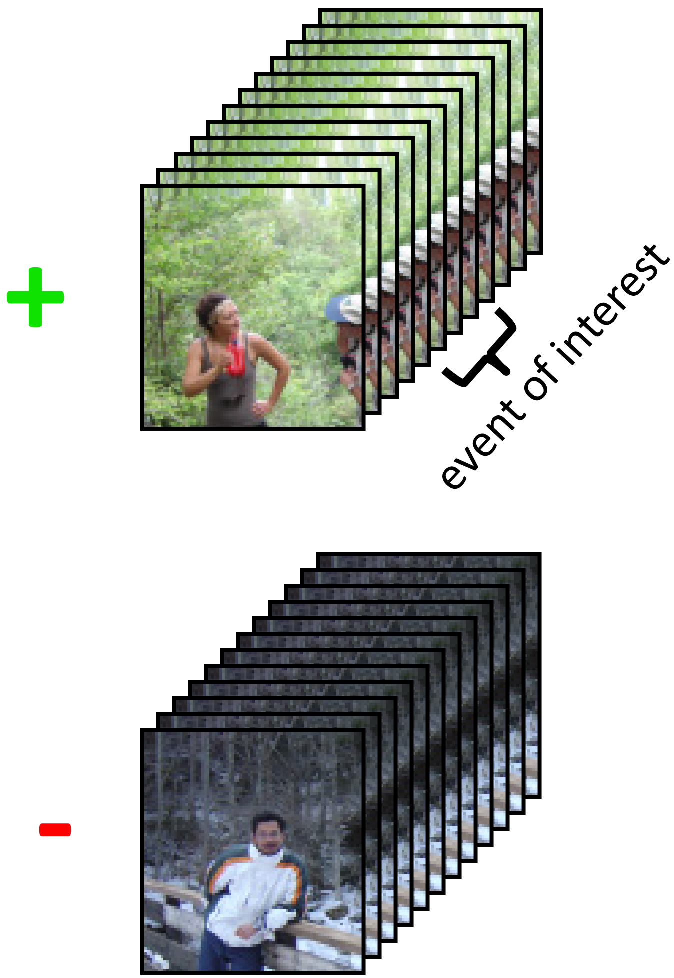
Detecting ‘sh’ phoneme



Bag: audio of word
Instance: audio clip

[Mandel et al. ‘08]

MIL Example: Event Detection (Video)



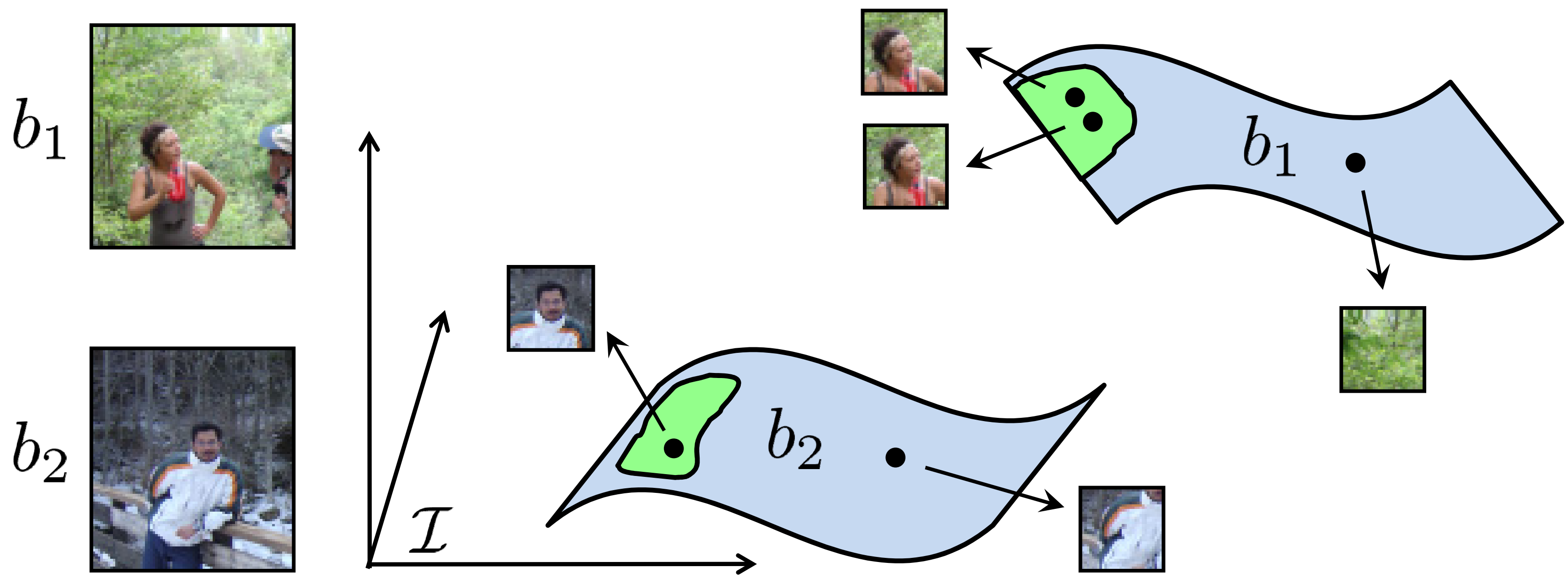
Bag: video
Instance: few frames

[Ali et al. '08, Buehler et al. '09, Stikic et al. '09]

Observations for these applications

- Top down process: draw entire bag from a **bag distribution**, then get instances
- Instances of a bag lie on a **manifold**

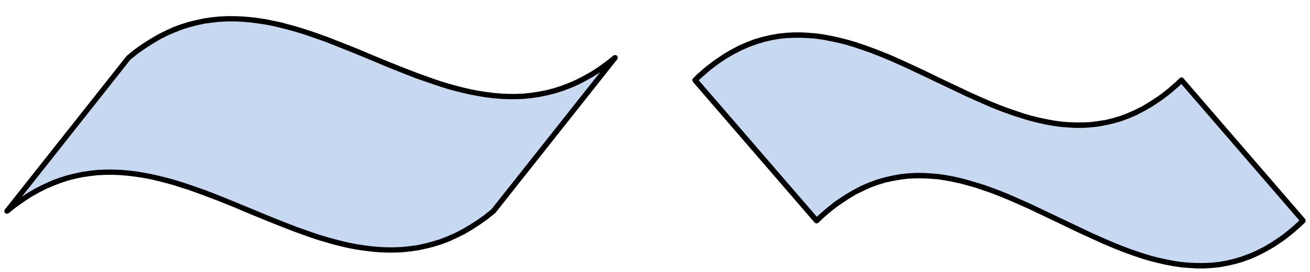
Manifold Bags



Negative region   Positive region

Manifold Bags

- For such problems:
  - Existing analysis not appropriate because number of instances is infinite
  - Expect sample complexity to scale with **manifold** parameters (curvature, dimension, volume, etc)



Manifold Bags: Formulation

- Manifold bag b drawn from **bag** distribution $\mathcal{D}_\mathcal{B}$
- Instance hypotheses:

$$h \in \mathcal{H}, \quad h : \mathcal{I} \to \{0, 1\}$$

- Corresponding bag hypotheses:

$$\bar{h} \in \overline{\mathcal{H}}, \quad \bar{h} : \mathcal{B} \to \{0, 1\}$$

$$\bar{h}(b) \stackrel{\text{def}}{=} \max_{x \in b} h(x)$$

Typical Route: VC Dimension

- Error Bound:

$$e \leq \hat{e} + O\left(\sqrt{\frac{VC(\overline{\mathcal{H}})}{m}}\right)$$

Typical Route: VC Dimension

- Error Bound:

$$e \leq \hat{e} + O\left(\sqrt{\frac{VC(\overline{\mathcal{H}})}{m}}\right)$$

generalization error

empirical error

# of training bags

Typical Route: VC Dimension

- Error Bound:

$$e \leq \hat{e} + O\left(\sqrt{\frac{VC(\overline{\mathcal{H}})}{m}}\right)$$

VC Dimension of <u>bag</u> hypothesis class

Relating $\overline{\mathcal{H}}$ to $\mathcal{H}$

- We do have a handle on $VC(\mathcal{H})$
- For **finite** sized bags, Sabato & Tishby:

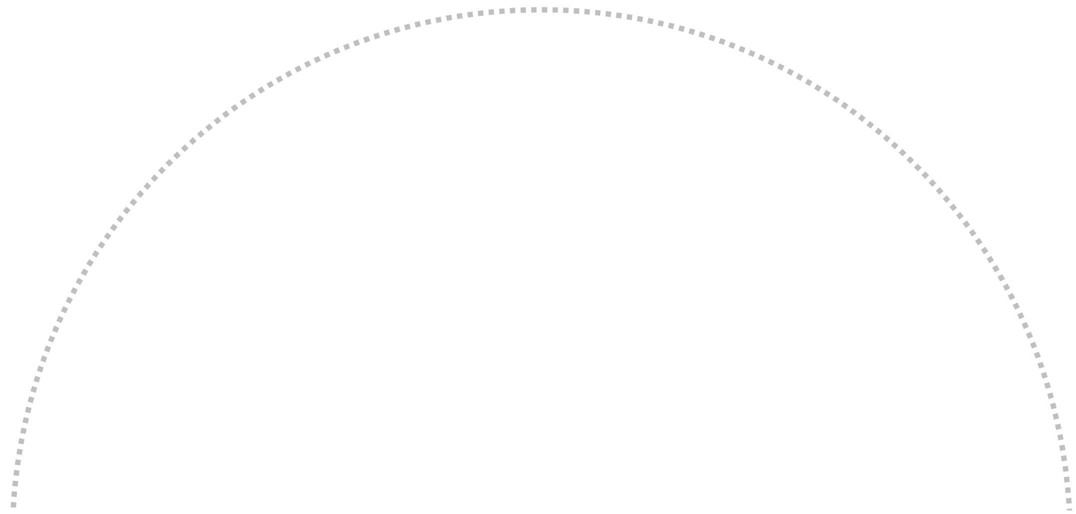
$$VC(\overline{\mathcal{H}}) \leq VC(\mathcal{H}) \log(r)$$

- Question: can we assume manifold bags are smooth and use a covering argument?

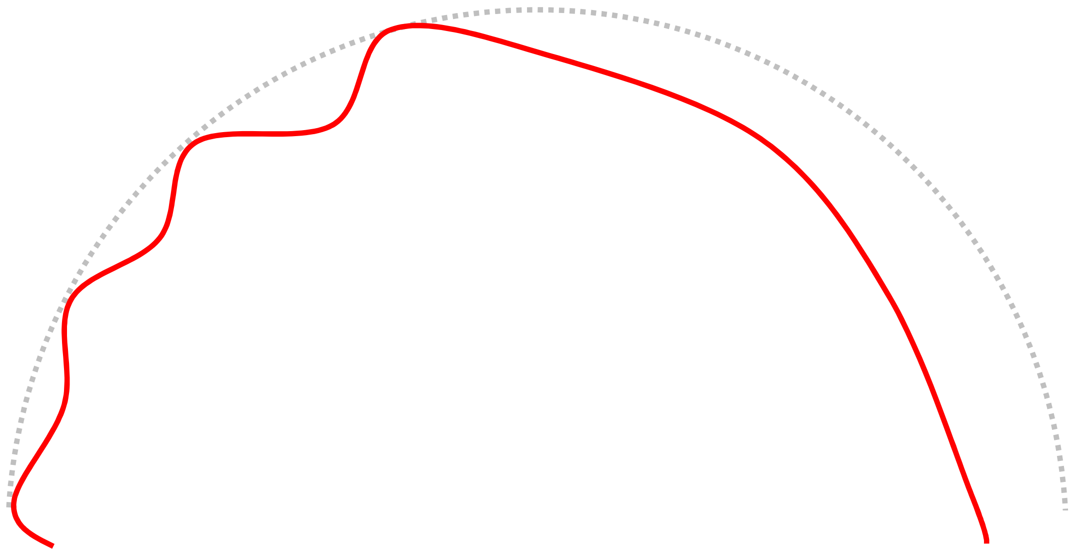
VC of bag hypotheses is unbounded!

- Let $\mathcal{H}$ be half spaces (hyperplanes)
- For arbitrarily smooth bags can always construct any number of bags s.t. **all possible** labelings achieved by $\overline{\mathcal{H}}$
- Thus, $VC(\overline{\mathcal{H}})$ unbounded!

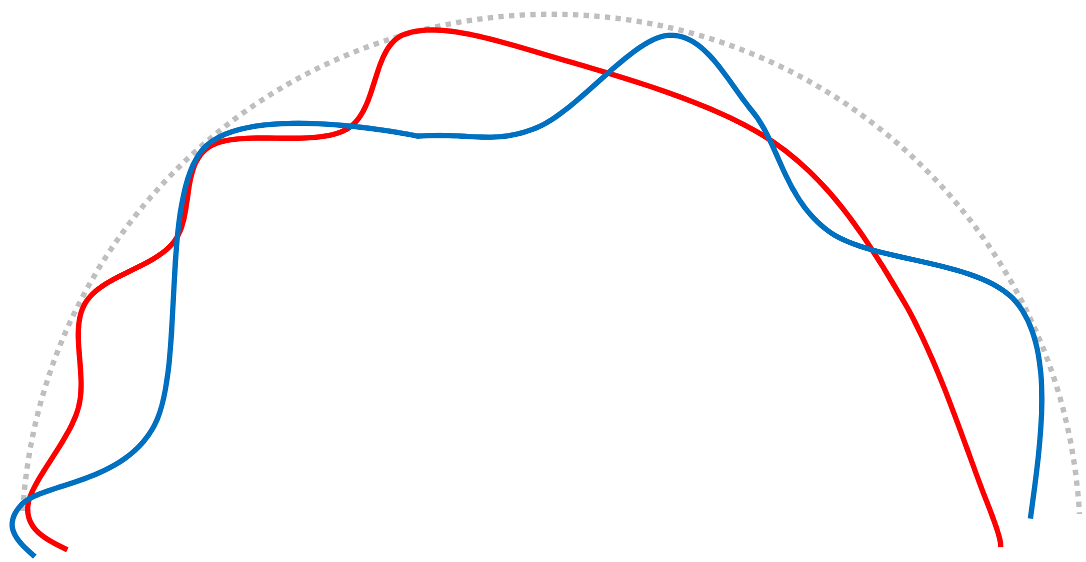
Example (3 bags)



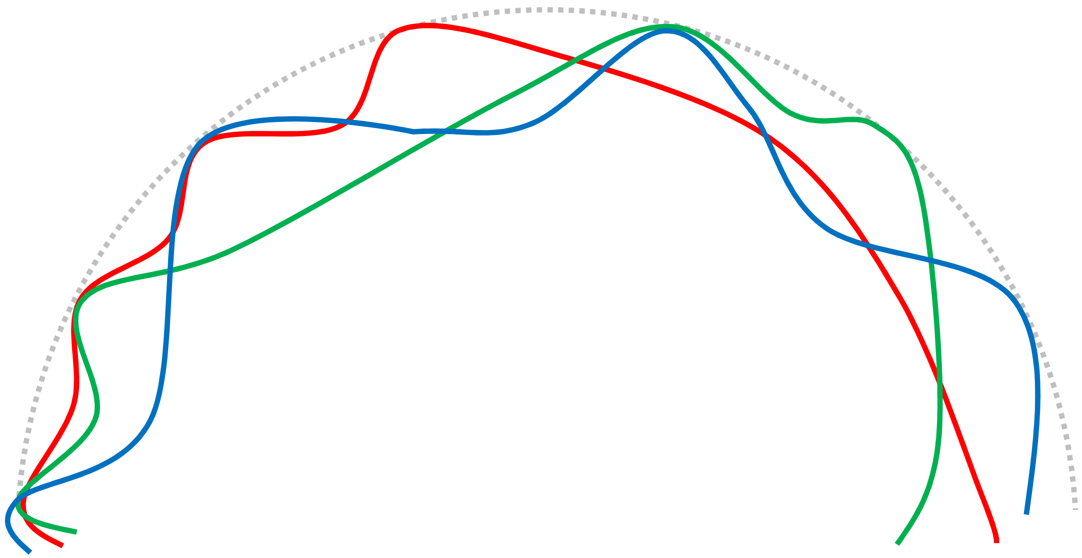
Example (3 bags)



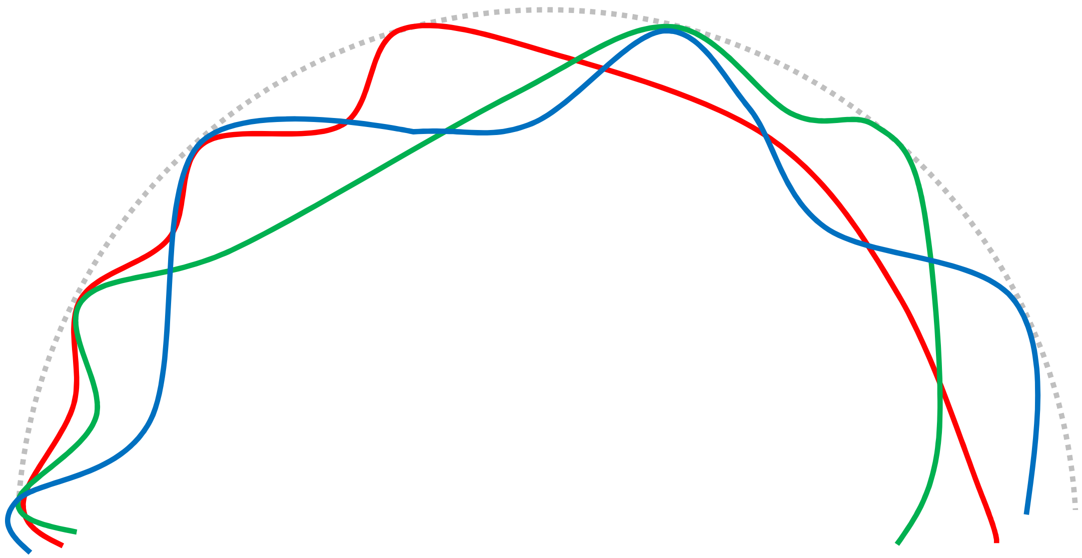
Example (3 bags)



Example (3 bags)

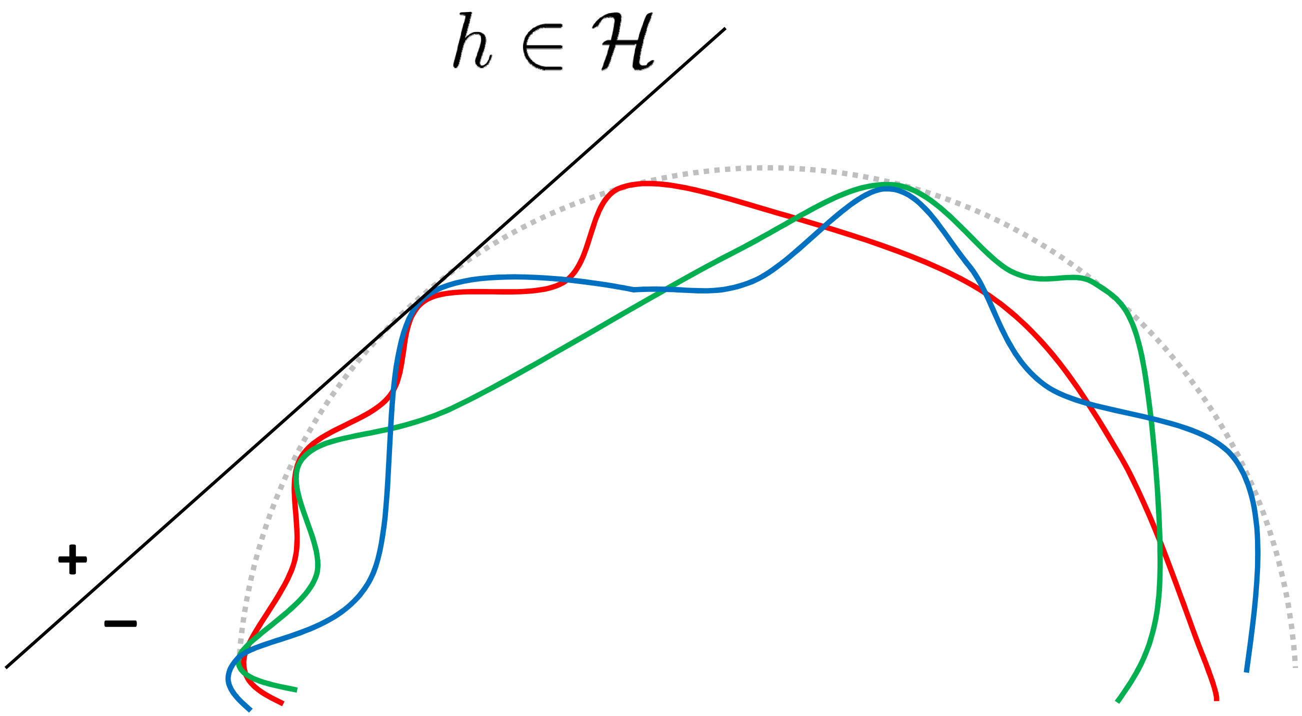


Example (3 bags)



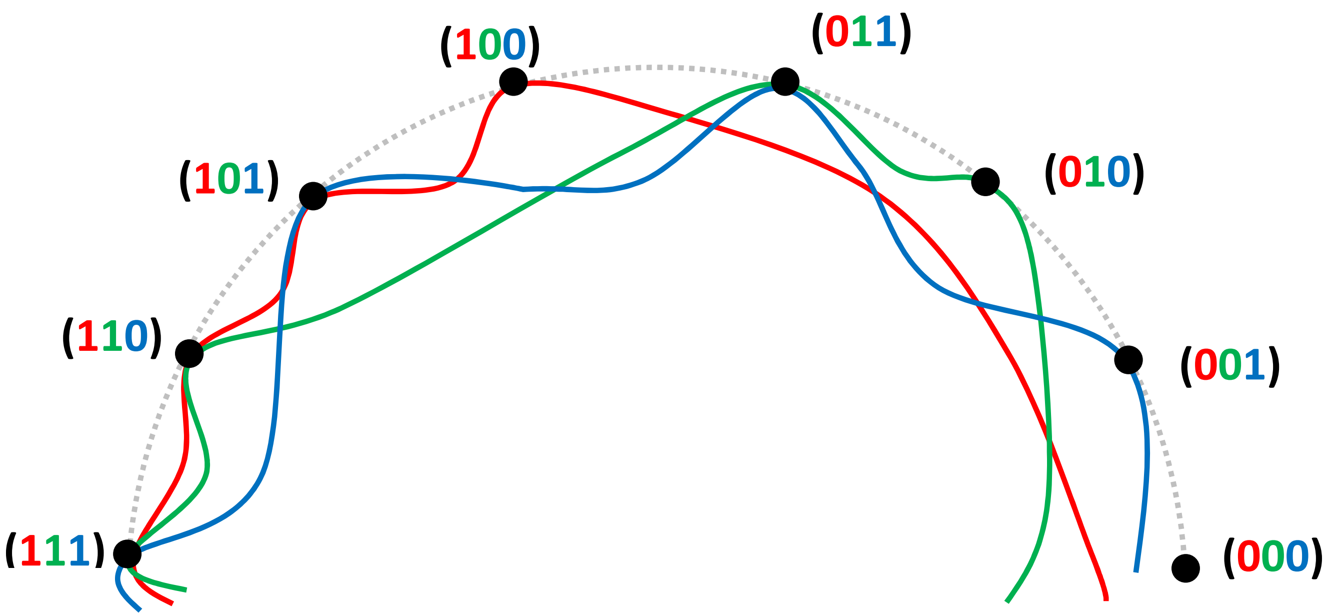
Want labeling **(101)**

Example (3 bags)



Achieves labeling (**101**)

Example (3 bags)



All possible labelings

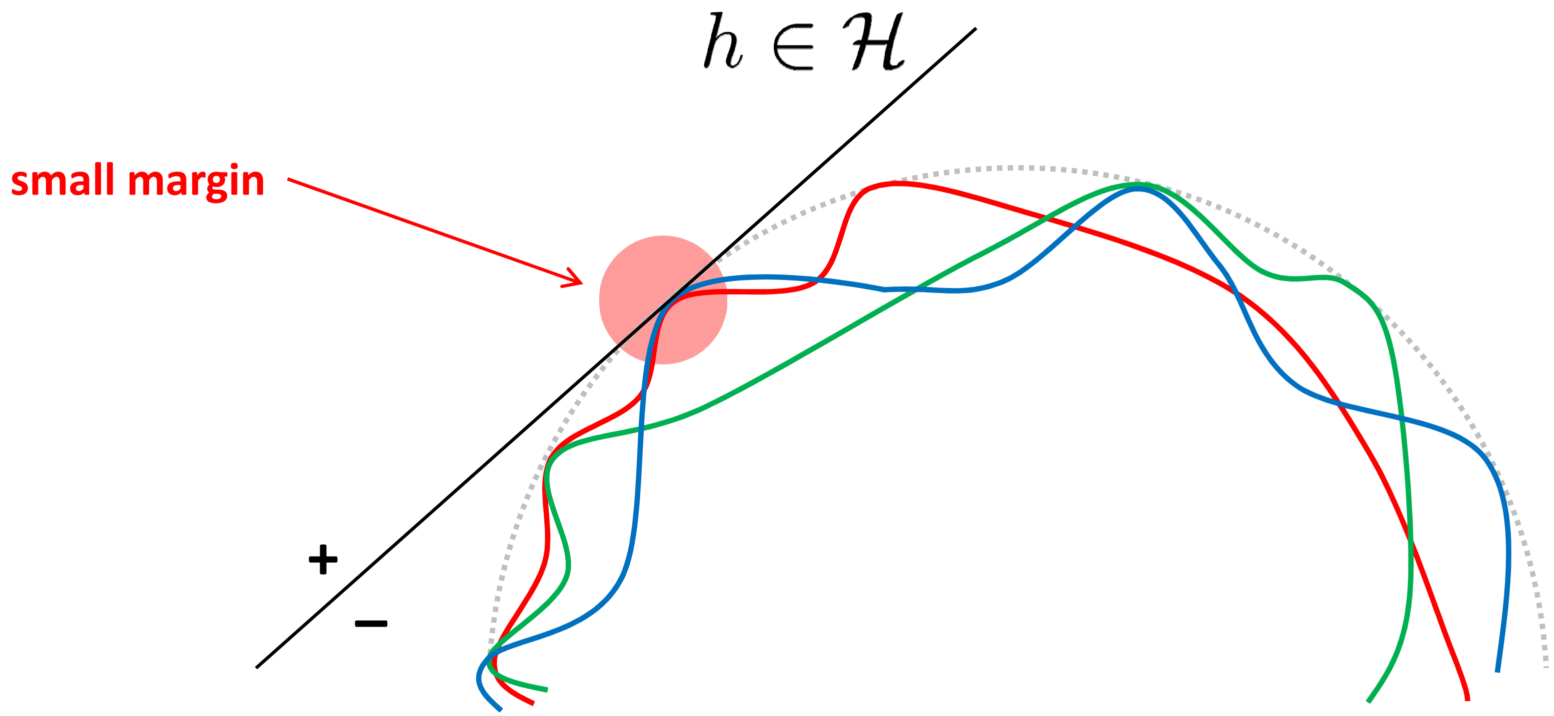
Issue

- Bag hypothesis class too powerful
  - For positive bag, need to only classify 1 instance as positive
  - Infinitely many instances -> too much flexibility for bag hypothesis
- Would like to ensure a non-negligible portion of positive bags is labeled positive

Solution

- Switch to real-valued hypothesis class
  - $h_r \in \mathcal{H}_r : \mathcal{I} \to [0, 1]$
  - corresponding bag hypothesis also real-valued
  - binary label via thresholding
  - true labels still binary
- Require that h_r is (lipschitz) **smooth**
- Incorporate a notion of **margin**

Example (3 bags)



Fat-shattering Dimension

- $\mathrm{F}_\gamma(\overline{\mathcal{H}_r})$ = "Fat-shattering" dimension of real-valued hypothesis class [Anthony & Bartlett '99]
  - Analogous to VC dimension
- Relates **generalization** error to **empirical** error at margin γ
  - i.e. not only does binary label have to be correct, margin has be to $\geq \gamma$

Fat-shattering of Manifold Bags

- Error Bound:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{\mathrm{F}_{\gamma/8}(\overline{\mathcal{H}}_r)}{m}}\right)$$

Fat-shattering of Manifold Bags

- Error Bound:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{\mathrm{F}_{\gamma/8}(\overline{\mathcal{H}}_r)}{m}}\right)$$

generalization error

# of training bags

empirical error at margin γ

Fat-shattering of Manifold Bags

- Error Bound:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{\mathrm{F}_{\gamma/8}(\overline{\mathcal{H}}_r)}{m}}\right)$$

fat shattering of <u>bag</u> hypothesis class

Fat-shattering of Manifold Bags

- Bound $\mathrm{F}_\gamma(\overline{\mathcal{H}_r})$ in terms of $\mathrm{F}_\gamma(\mathcal{H}_r)$
  - Use covering arguments – approximate manifold with finite number of points
  - Analogous to Sabato & Tishby's analysis of finite size bags

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

generalization error

complexity term

empirical error at margin γ

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

fat shattering of <u>instance</u> hypothesis class

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

number of training bags

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

manifold dimension

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{V m}{\gamma^2 \kappa^n}\right)}\right)$$

manifold volume

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

term depends (inversely) on smoothness of manifolds & smoothness of instance hypothesis class

Error Bound

- With high probability:

$$e \leq \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}(\mathcal{H})}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

- Obvious strategy for learner:
  - Minimize empirical error & maximize margin
  - This is what most MIL algorithms already do

Learning from Queried Instances

- Previous result assumes learner has access **entire** manifold bag
- In practice learner will only access small number of instances (ρ)

- Not enough instances -> might not draw a pos. instance from pos. bag

Learning from Queried Instances

- Bound

$$e \le \hat{e}_\gamma + O\left(\sqrt{\frac{n^2 \mathrm{F}_{\gamma/16}}{m} \log^2\left(\frac{Vm}{\gamma^2 \kappa^n}\right)}\right)$$

holds with failure probability increased by δ **if**

$$\rho \ge \Omega\left(\left(V/\kappa^n\right)\left(n + \ln\left(\frac{mV}{\kappa^n \delta}\right)\right)\right)$$

Take-home Message

- Increasing m reduces **complexity term**
- Increasing ρ reduces **failure probability**
  - Seems to contradict previous results (smaller bag size r is better)
  - Important difference between r and ρ !
  - If ρ is too small we may only get negative instances from a positive bag
- Increasing m requires extra labels, increasing ρ does not

Iterative Querying Heuristic (IQH)

- Problem: want many instances/bag, but have computational limits
- Heuristic solution:
  - Grab small number of instances/bag, run standard MIL algorithm
  - Query more instances from each bag, only keep the ones that get high score from current classifier
- At each iteration, train with small # of instances

Experiments

- Synthetic Data (will skip in interest of time)
- Real Data
  - INRIA Heads (images)
  - TIMIT Phonemes (audio)

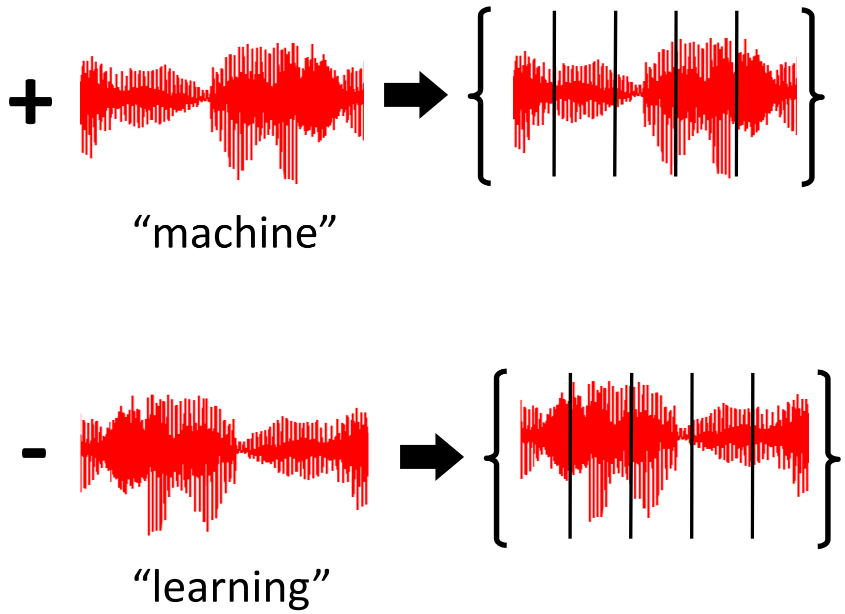
INRIA Heads


pad=16


pad=32

[Dalal et al. '05]

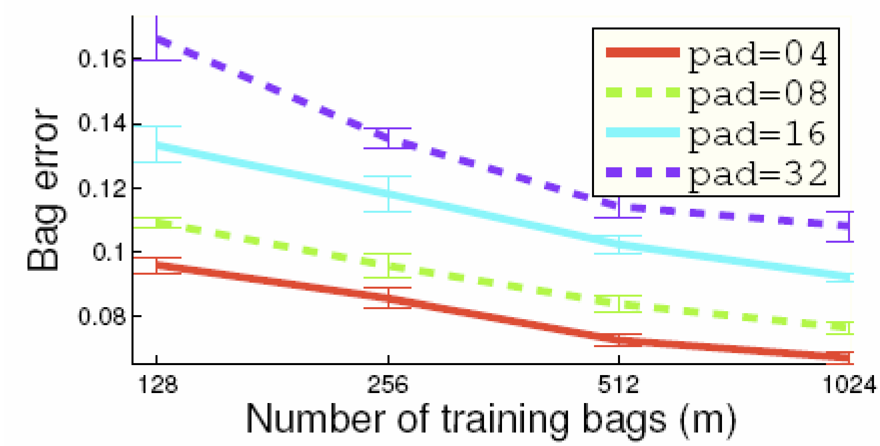
TIMIT Phonemes



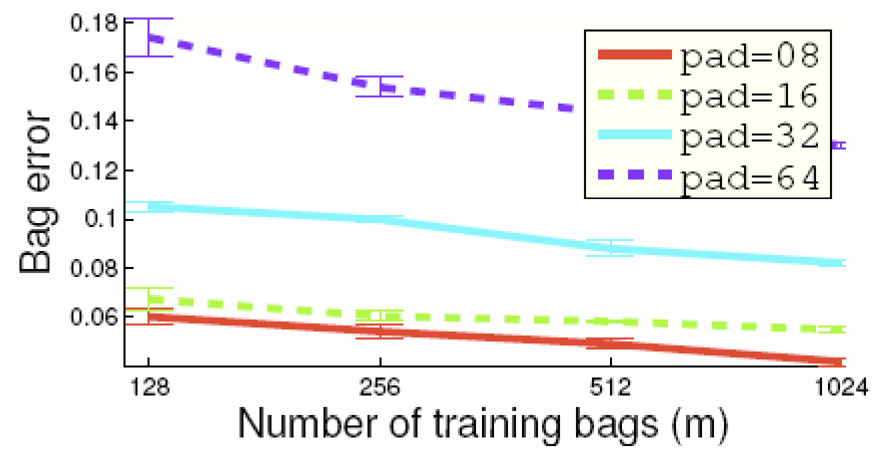
[Garofolo et al., '93]

Padding (volume)

INRIA Heads

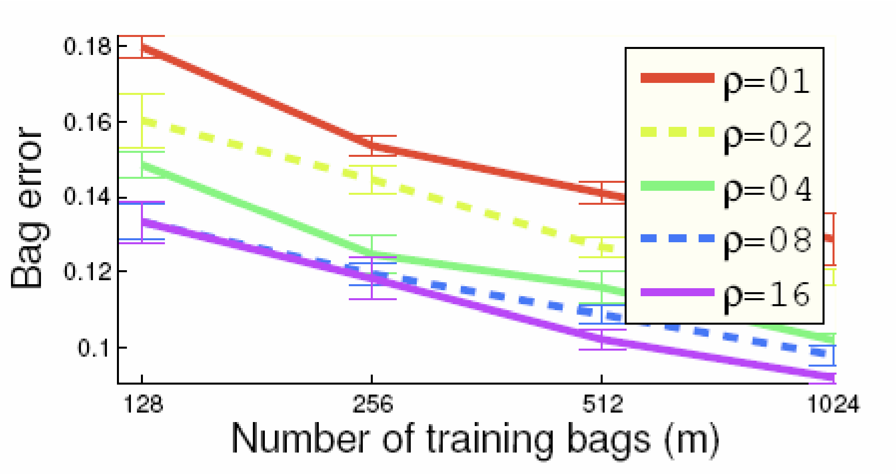


TIMIT Phonemes

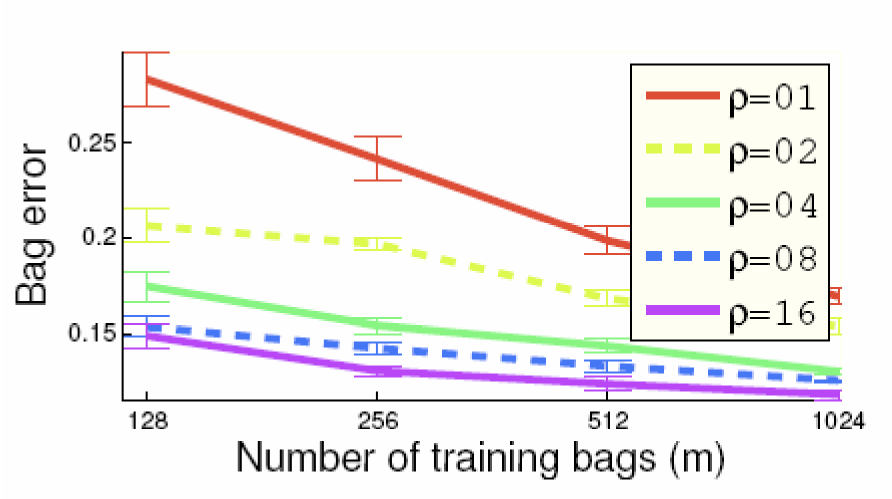


Number of Instances (ρ)

INRIA Heads

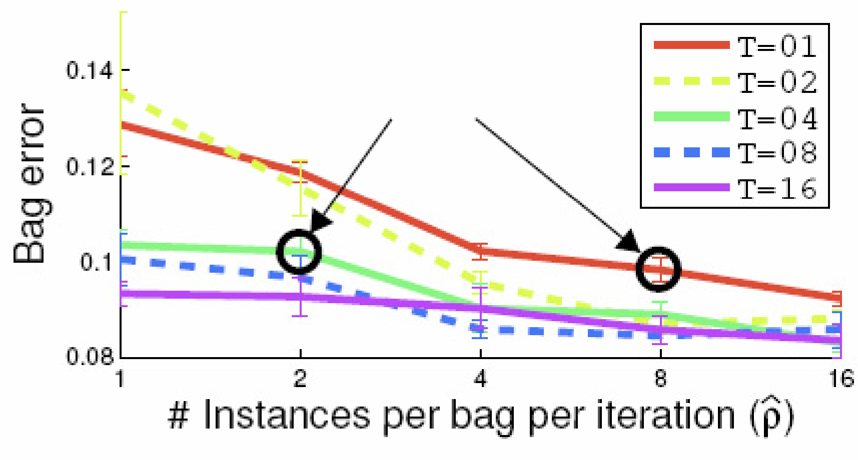


TIMIT Phonemes

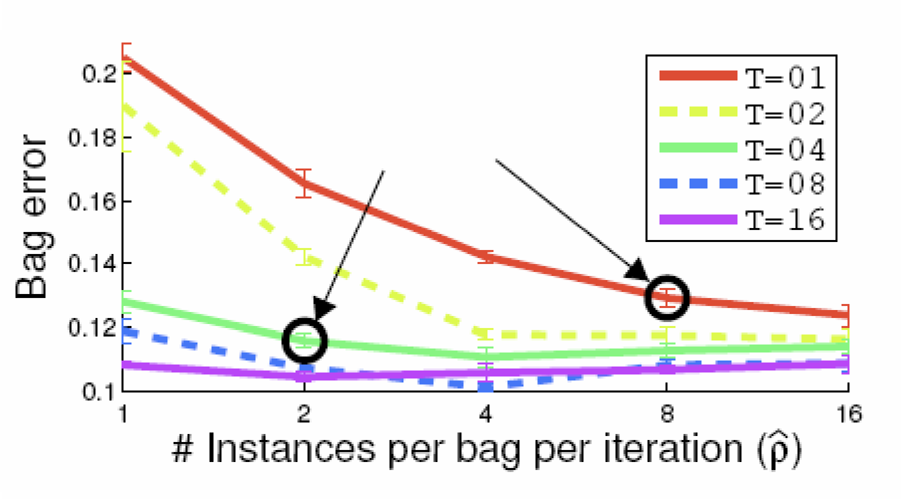


Number of Iterations (heuristic)

INRIA Heads



TIMIT Phonemes



Conclusion

- For many MIL problems, bags modeled better as **manifolds**
- PAC Bounds depend on **manifold properties**
- Need **many instances** per manifold bag
- **Iterative** approach works well in practice, while keeping comp. requirements low
- Further algorithmic development taking advantage of manifold would be interesting

Thanks

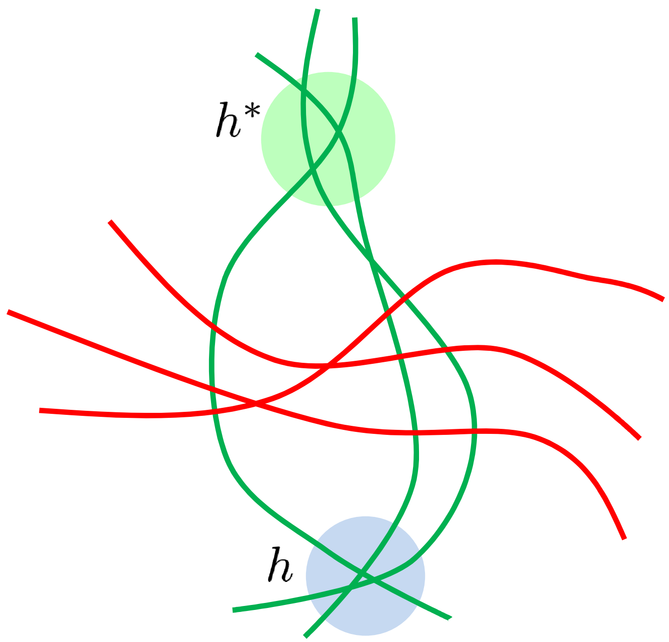
- Happy to take questions!

Why not learn directly over bags?

- Some MIL approaches do this
  - Wang & Zucker '00, Gartner et al. '02
- In practice, instance classifier is desirable
- Consider image application (face detection)
  - Face can be anywhere in image
  - Need features that are extremely robust

Why not instance error?

- Consider this example:



- In practice instance error tends to be low (if bag error is low)

Doesn’t VC have lower bound?

- Subtle issue with FAT bounds
  - If the distribution is terrible, $\hat{e}_\gamma$ will be high
- Consider SVMs with RBF kernel
  - VC dimension of linear separator is n+1
  - FAT dimension only depends on margin (Bartlett & Shawe-Taylor, 02)

Aren’t there finite number of image patches?

- We are **modeling** the data as a manifold
- In practice, everything gets discretized
- Actual number of instances (e.g. image patches with any scale/orientation) may be huge – existing bounds still not appropriate