OASIS:
Online Active Semi-Supervised Learning




Andrew Goldberg, Xiaojin Zhu*, Alex Furger, Jun-Ming Xu

University of Wisconsin-Madison

AAAI 2011

The Problem We Consider

1. At time t the world picks x_t, y_t, shows x_t
2. We predict y'_t
3. With small probability, world reveals y_t
4. If y_t not revealed we may query it
5. We update our model even if y_t unknown

Is this …

- semi-supervised learning?
  - Yes, but sequential input, active query
- online learning?
  - Yes, but learns on unlabeled items
- active learning?
  - Yes, but learns on un-queried items

OASIS = Online Active Semi-Supervised Learning

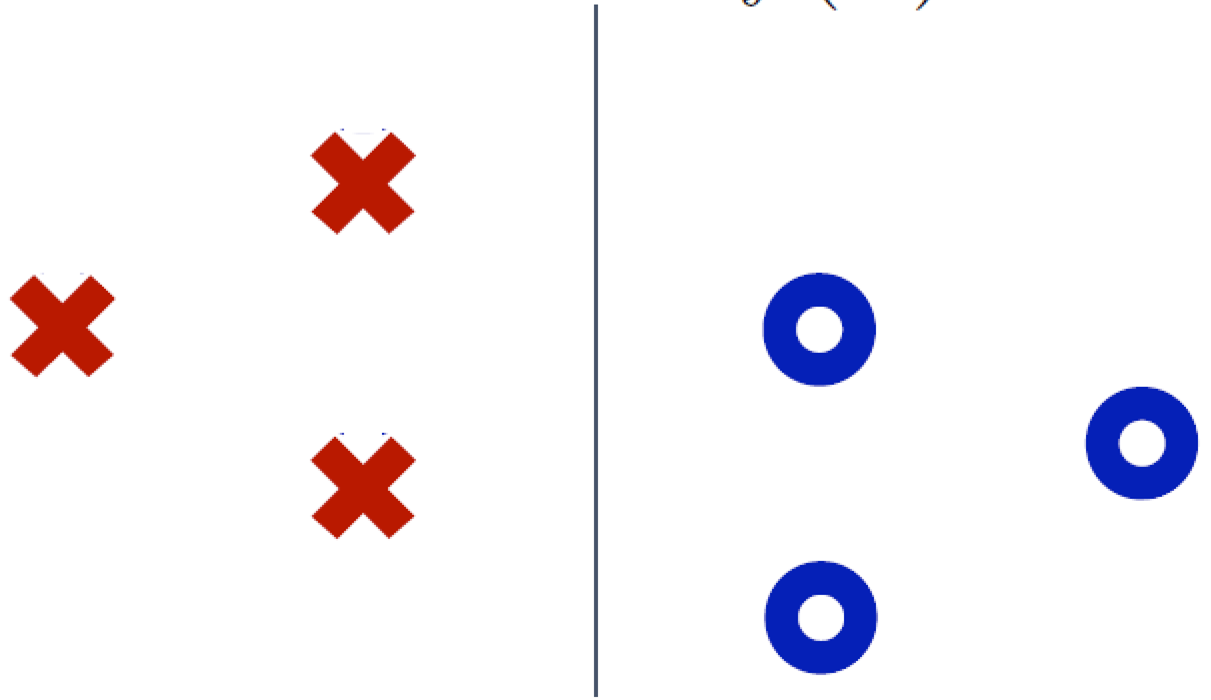
Main idea: Be Bayesian!

- Track all gaps with the posterior.
  - semi-supervised learning
  - online learning
  - active learning

  all naturally follow.

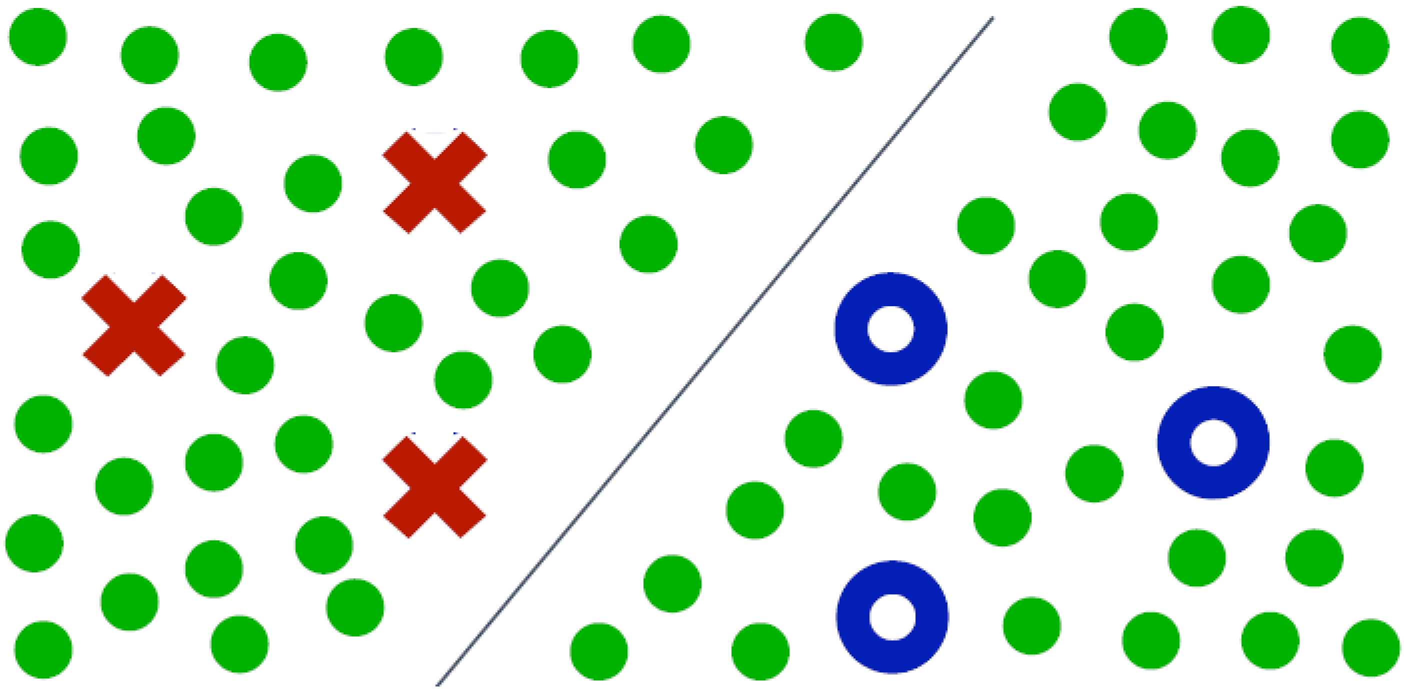
The Margin in Supervised Learning

- E.g. SVM linear classifier $f(\mathbf{x}) \doteq \mathbf{w}^\top \mathbf{x}$



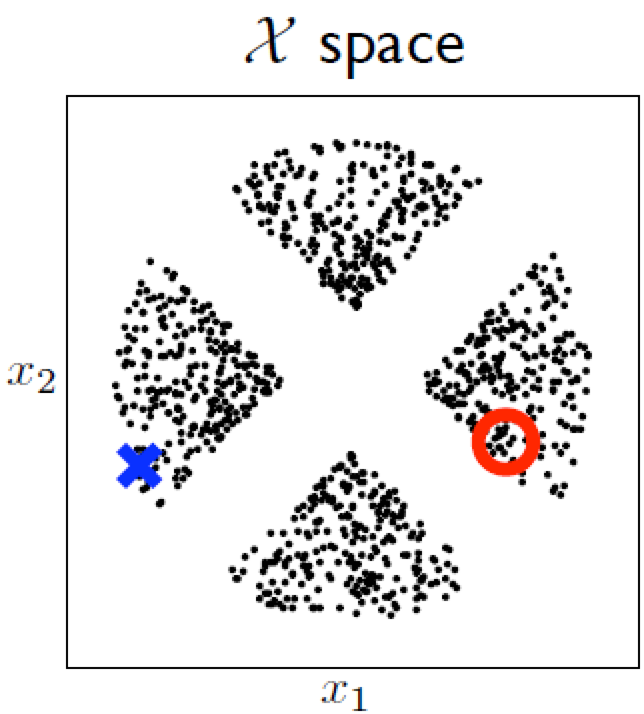
The Gap Assumption in SSL

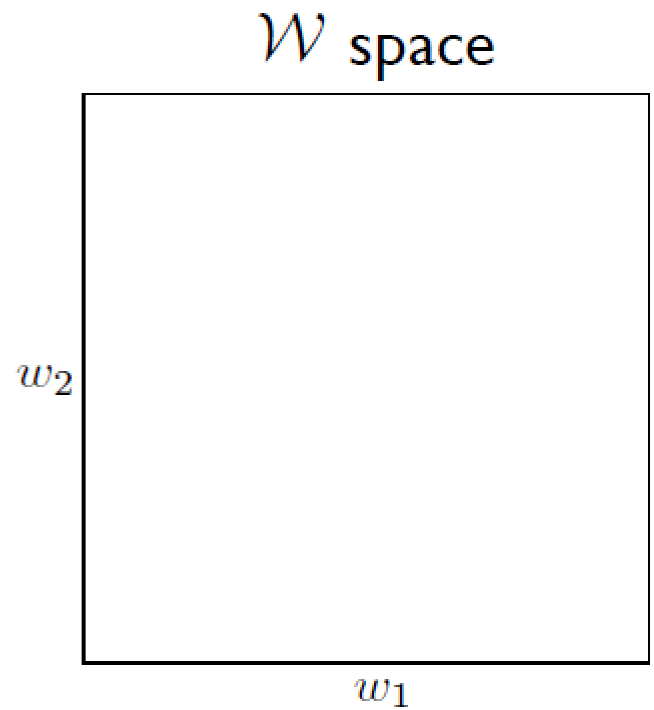
- S3VM: find the largest unlabeled margin



The Need for a Multi-Modal Posterior

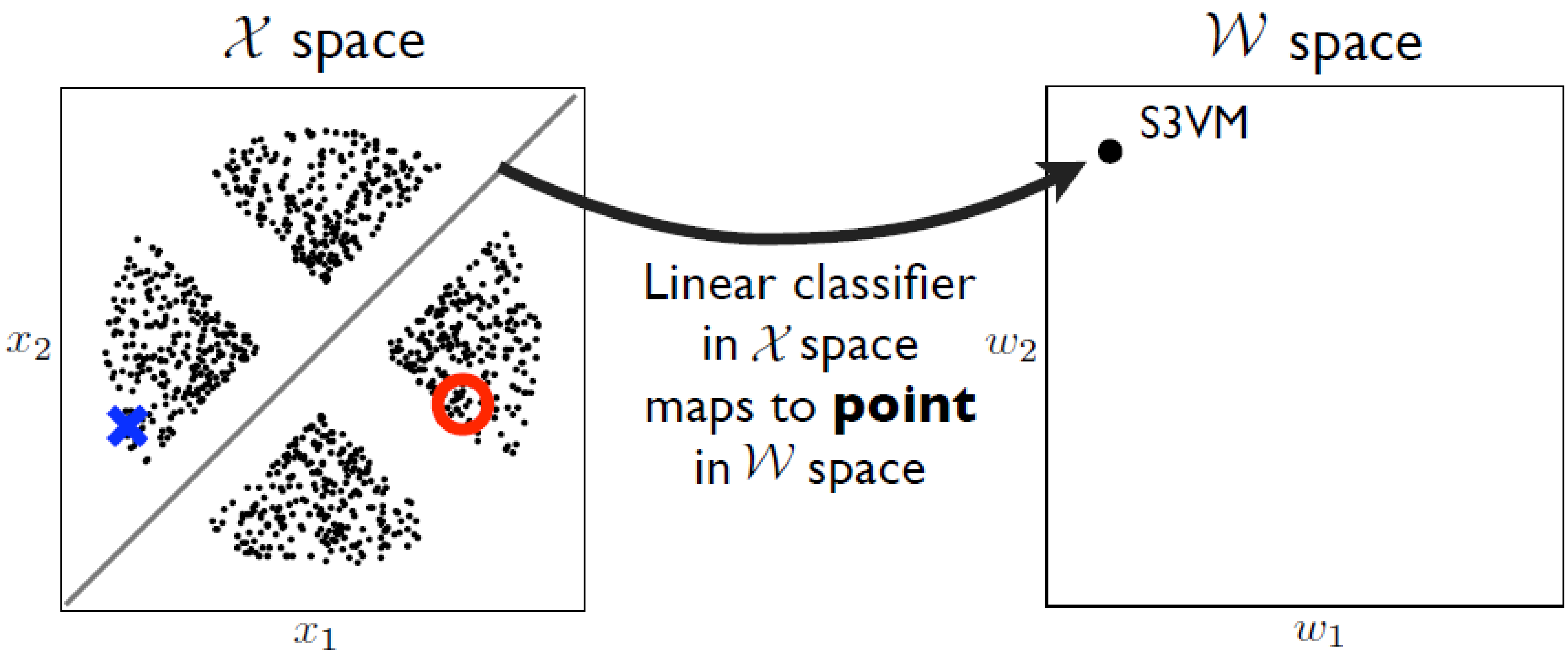
- There may be multiple candidate gaps





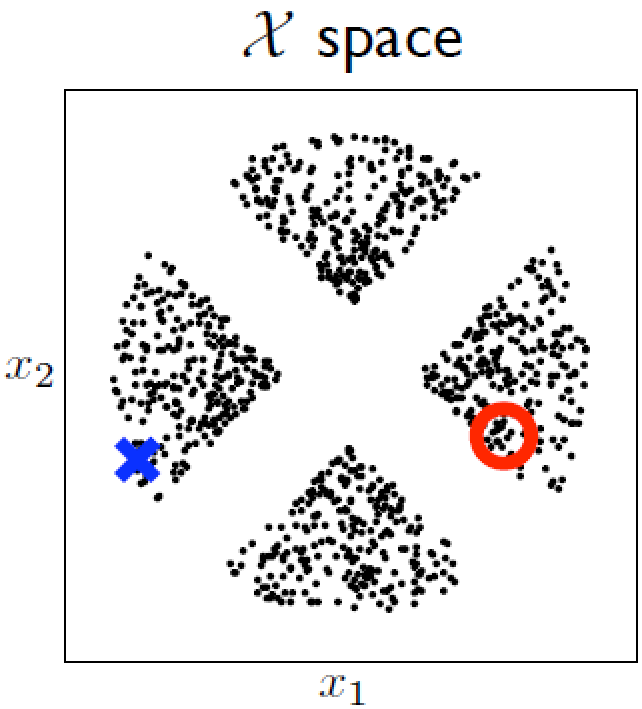
The Need for a Multi-Modal Posterior

- There may be multiple candidate gaps

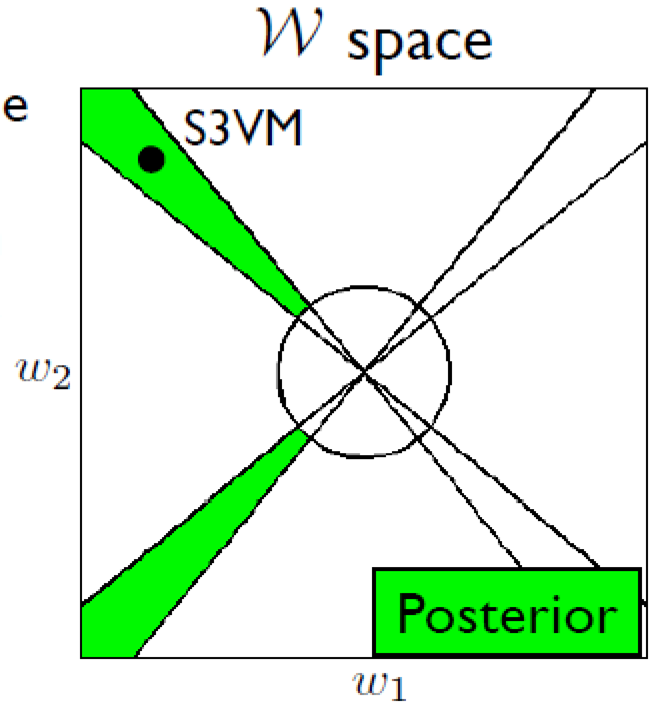


The Need for a Multi-Modal Posterior

- There may be multiple candidate gaps

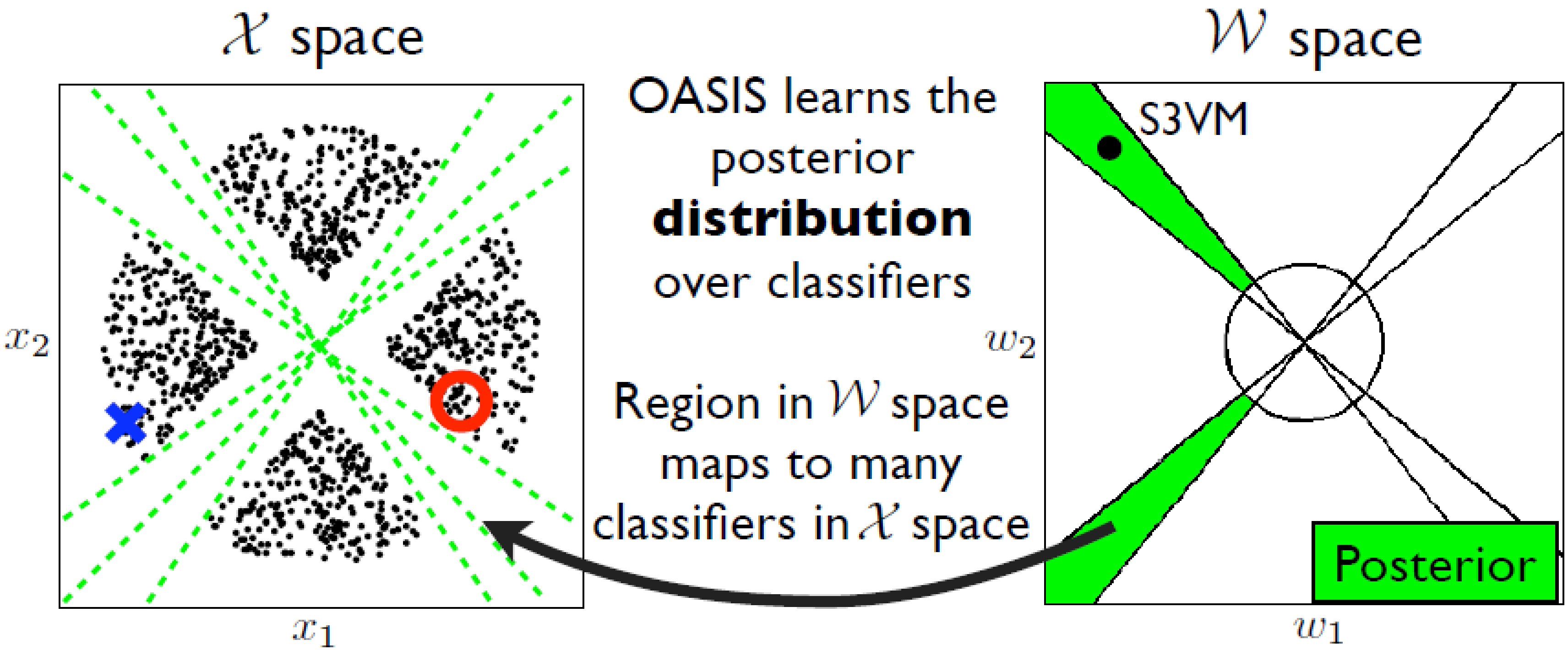


OASIS learns the posterior **distribution** over classifiers



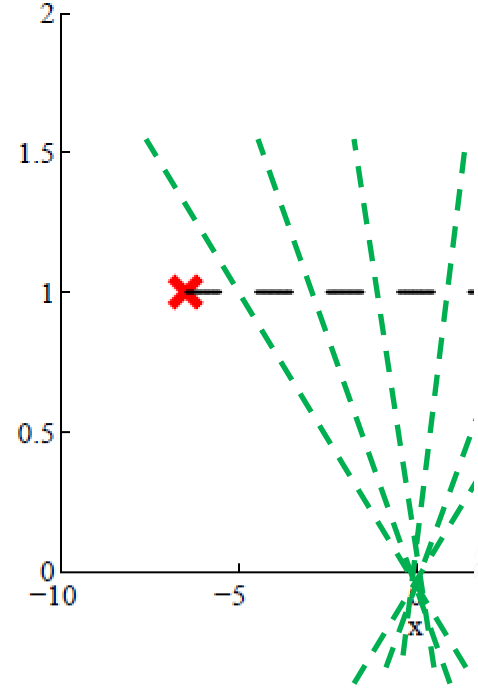
The Need for a Multi-Modal Posterior

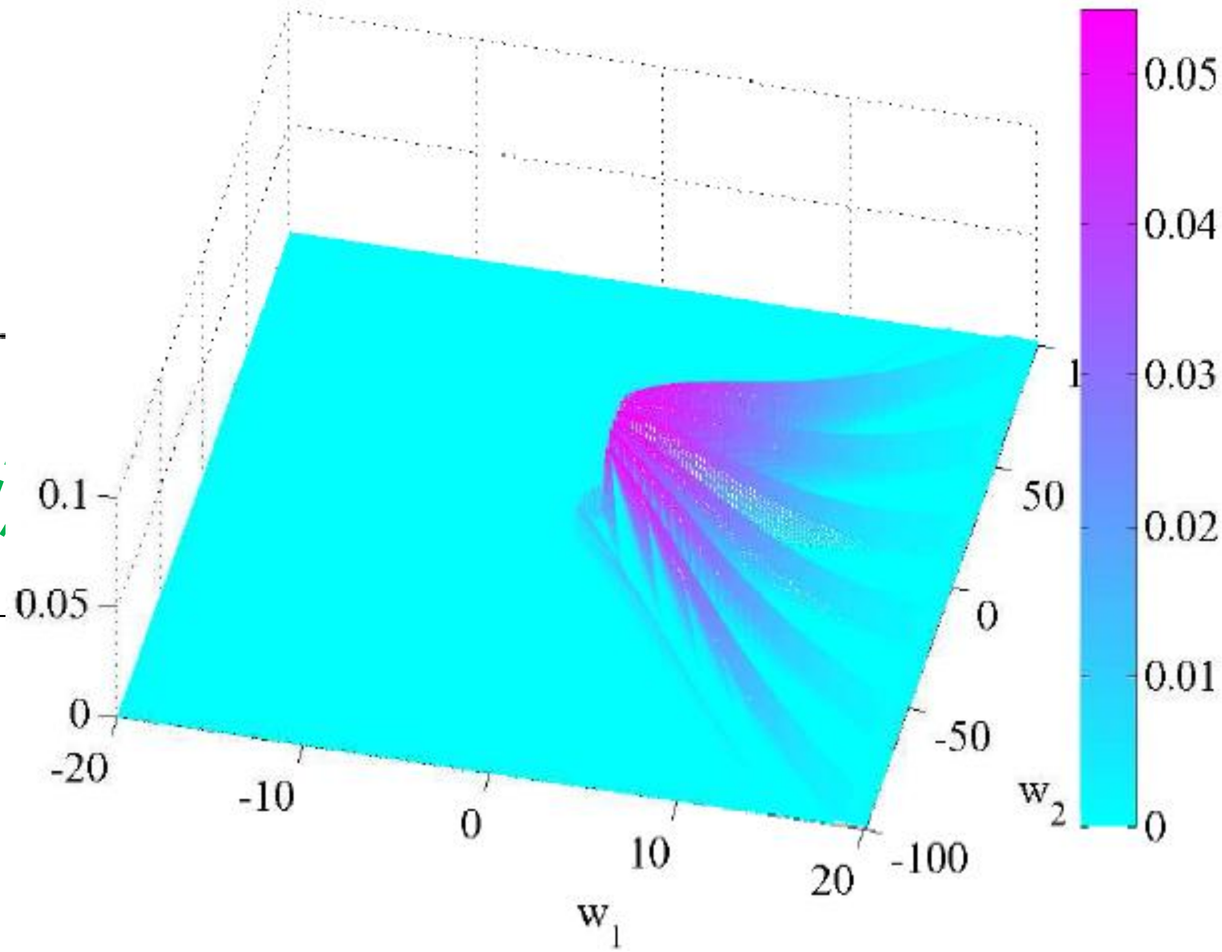
- There may be multiple candidate gaps



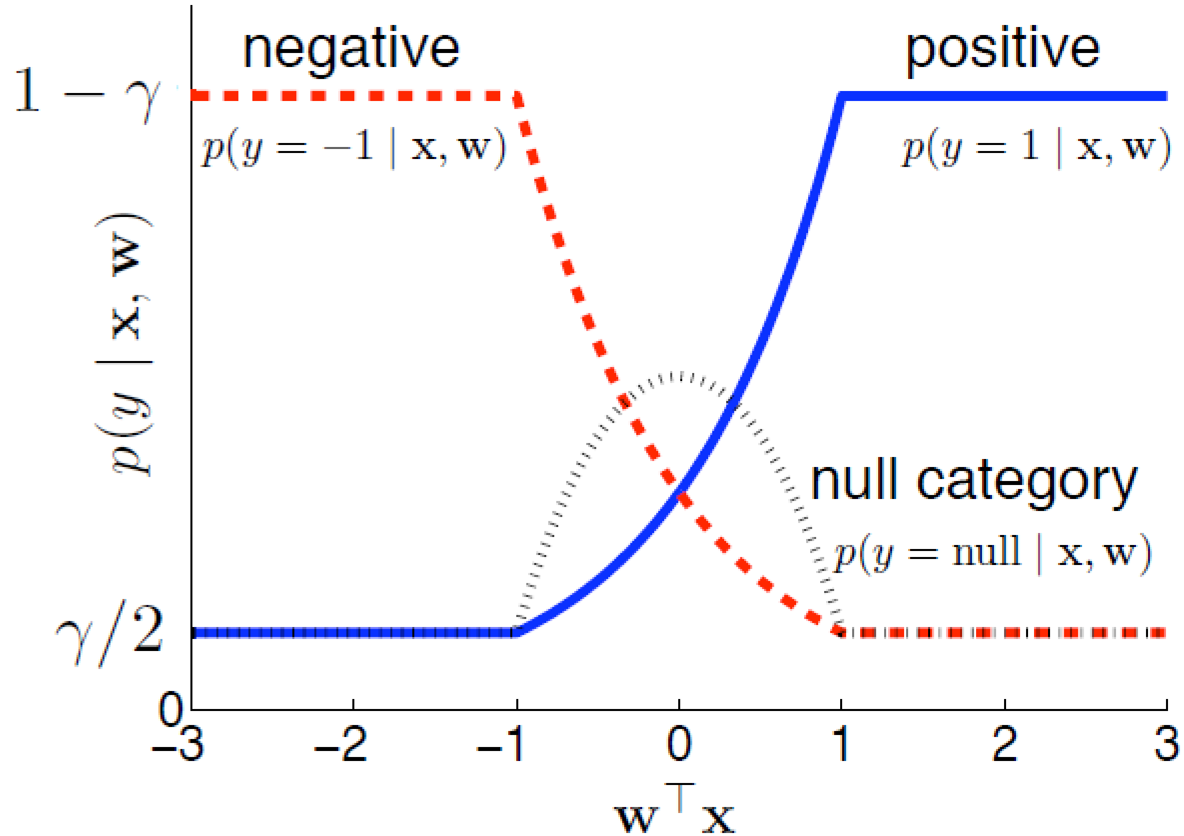
Another Example of Multi-modal Posterior

[courtesy of Kwang-Sung Jun]



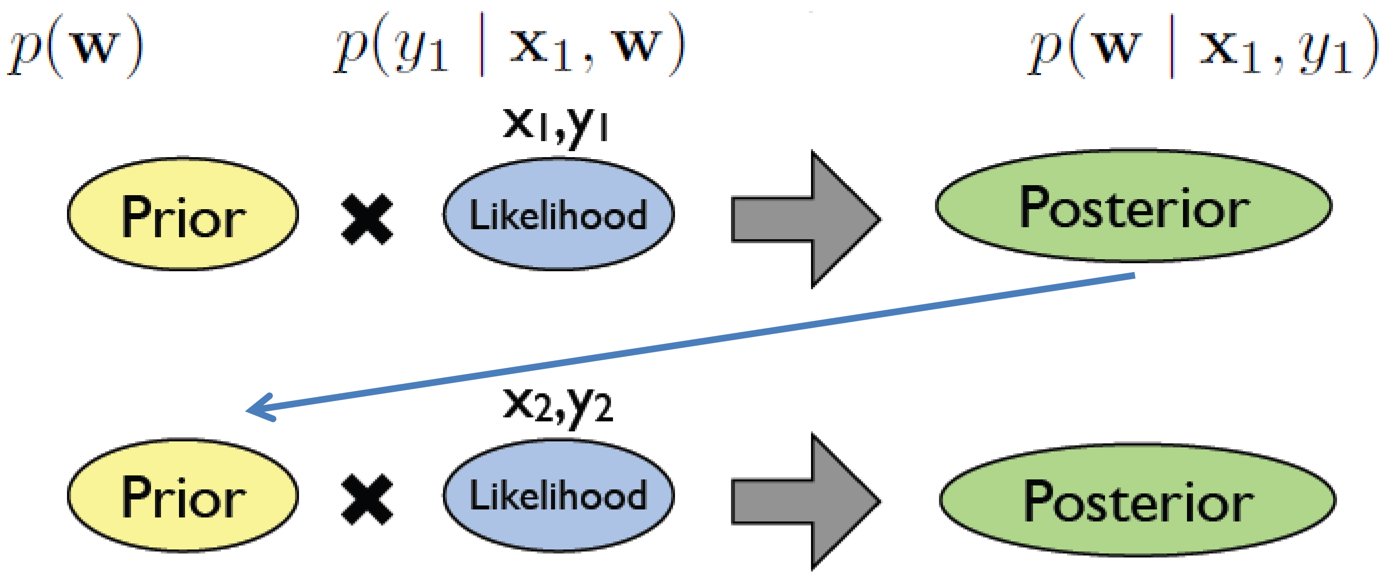


Life is Easy Being Bayesian: Likelihood



- The "null-category" likelihood pushes *w* away from unlabeled points. ➔ semi-supervised learning
- Inspired by [Lawrence & Jordan NIPS'04]

Life is Easy Being Bayesian: Update



- Sequential Bayesian update ➔ online learning
  - assume iid, not adversarial
  - Cauchy prior

Life is Easy Being Bayesian: Predict

- Predict label

$$\hat{y}_t = f(\mathbf{x_t}) = \underset{y \in \{-1,1\}}{\operatorname{argmax}}\, p(y \mid \mathbf{x_t}, D_{t-1})$$

- Integrate out w

$$p(y \mid \mathbf{x}_t, D_{t-1}) \quad = \quad \int p(y \mid \mathbf{x}_t, \mathbf{w}')p(\mathbf{w}' \mid D_{t-1})d\mathbf{w}'$$

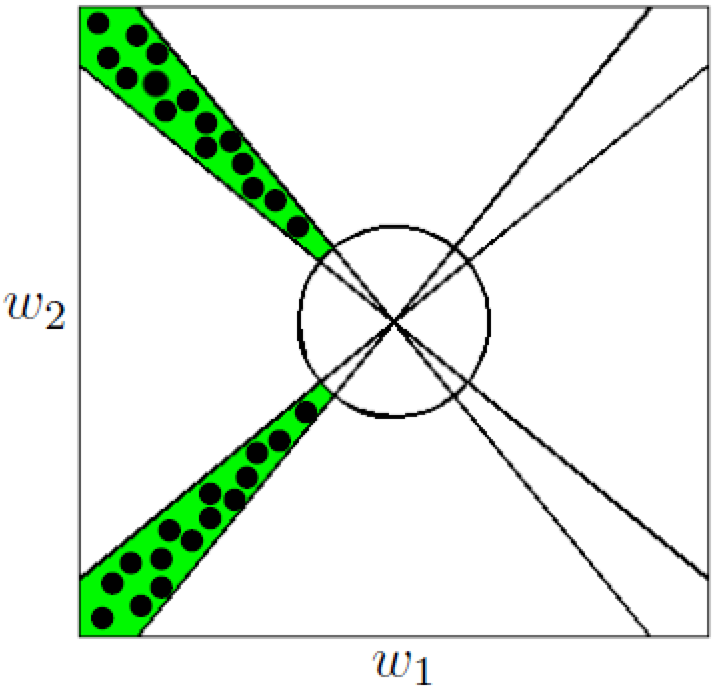
- If the posterior strongly disagree on x_t, ask for its label ➔ active learning

Life is Hard Being Bayesian!

$$p(y \mid \mathbf{x}_t, D_{t-1}) \;=\; \int p(y \mid \mathbf{x}_t, \mathbf{w}')p(\mathbf{w}' \mid D_{t-1})d\mathbf{w}'$$

intractable

- Particle filtering



Posterior approximated by *m* weighted particles:

$$p(\mathbf{w} \mid D_{t-1}) \approx \sum_{i=1}^{m} \beta_i \delta(\mathbf{w} - \mathbf{w}^{(i)})$$

Prediction using particles:

$$p(y \mid \mathbf{x}_t, D_{t-1}) \;=\; \int p(y \mid \mathbf{x}_t, \mathbf{w}')p(\mathbf{w}' \mid D_{t-1})d\mathbf{w}'$$

$$\approx \sum_{i=1}^{m} \beta_i p(y \mid \mathbf{x}_t, \mathbf{w}^{(i)})$$

Particle Filtering Details

- Update weight β_i by a multiplicative factor:

  $p(y = y_t | \mathbf{x}_t, \mathbf{w}_{t-1}^{(i)})$ if y_t is revealed or queried

  $p(y \in \{-1, 1\} | \mathbf{x}_t, \mathbf{w}_{t-1}^{(i)})$ if unlabeled

- Occasional resample-move to rejuvenize particles
  - A single step of Metropolis-Hastings sampling

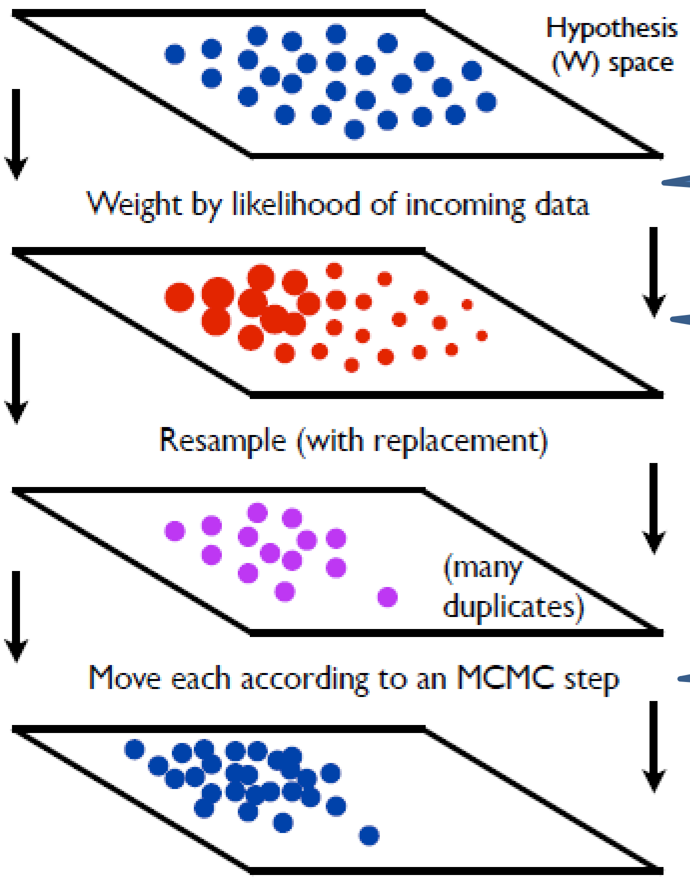
Active Learning using Particles

- Each incoming unlabeled point has a score:

$$\operatorname{score}(\mathbf{x}) = \left| \sum_{i=1}^{m} \beta_i \operatorname*{argmax}_{y \in \{-1,1\}} p(y \mid \mathbf{x}, \mathbf{w}^{(i)}) \right|$$

- Query for label if *score(x)* < s_0

The Complete Algorithm



If unlabeled and score(x)$<s_0$, query its label

The null category likelihood for gap assumption

Approximate Metropolis-Hastings with a small buffer

Experiments: List of Algorithms

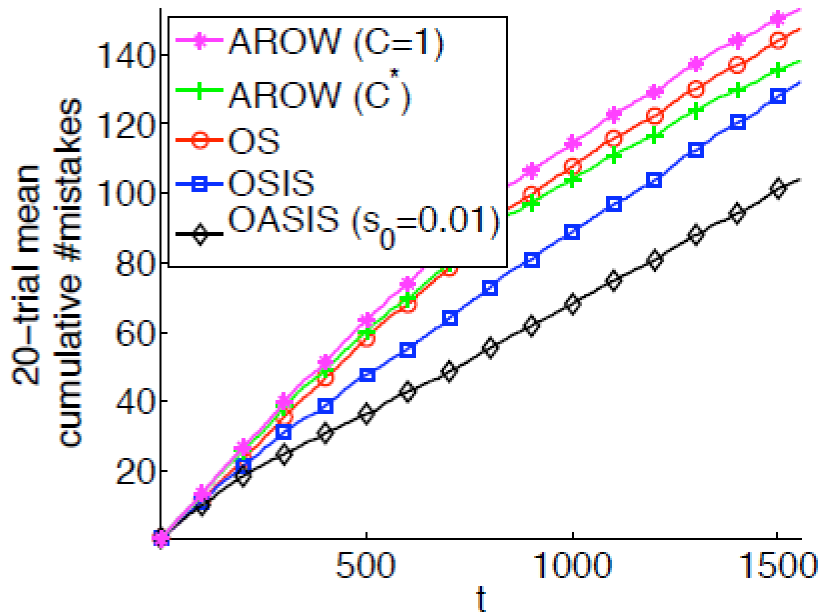
	Online	Active	SSL
OASIS	X	X	X
OSIS	X		X
OS	X		
AROW (C=1)	X		
AROW (C*) (test-set-tuned)	X		

OSIS=Online Semi-Supervised Learning
OS = Online Supervised learning
AROW = Adaptive Regularization of Weight Vectors
[Crammer et al. NIPS 09]

Experiments: Procedure

- 20 trials of T iterations
- Start with 2 labeled points
- To control the total number of labels:
  - First run OASIS, record the number of queries a
  - Run other algorithms with $2+a$ labeled points
- Same exact $x_1 \dots x_T$ sequence across algorithms

Results on Letter



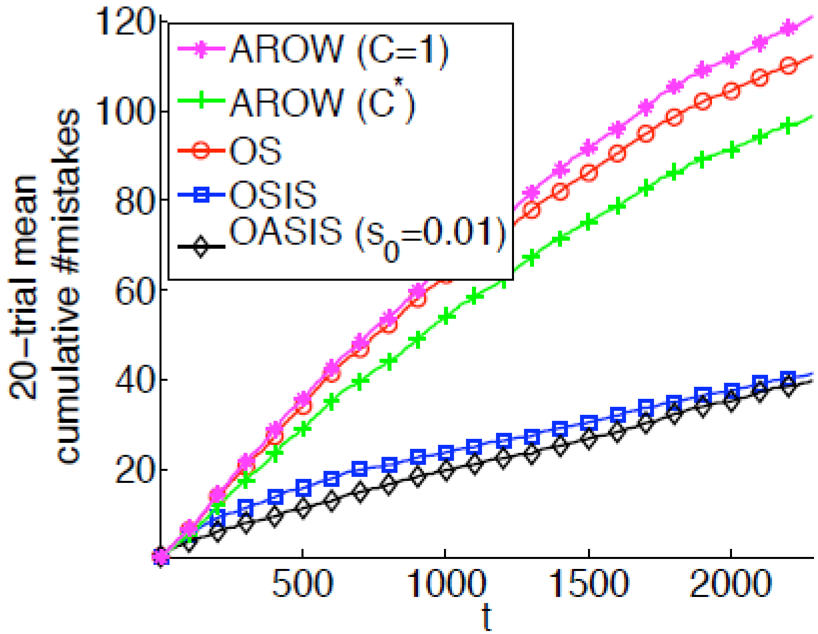
(a) letter A vs B ($d = 16$)

$T = 1555, l = 2$

$a = 5.10(1.92)$

OASIS ≫ OS, AROW, OSIS ≈ OS, AROW; active learning is key

Results on Pendigits



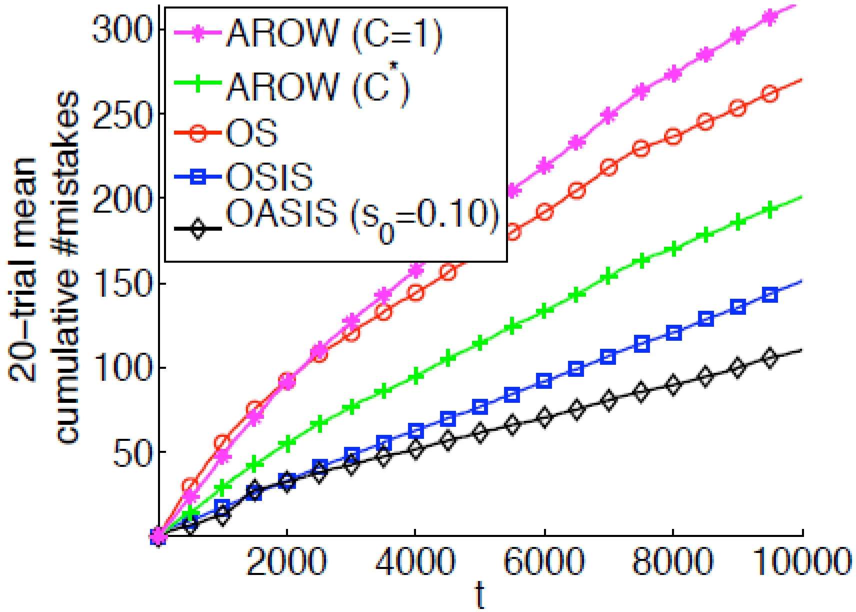
(b) pendigits 0 vs 1 ($d = 16$)

$T = 2286, l = 2$

$a = 2.60(1.14)$

OASIS, OSIS ≫ OS, AROW; semi-supervised learning is key

Results on MNIST



(c) MNIST 0 vs 1 ($d = 10$)

$T = 10000, l = 2$

$a = 10.30(5.01)$

OASIS » OSIS » OS, AROW;
SSL + active learning are both key

Summary

- Online + active + semi-supervised learning
- Full Bayesian on gap assumption
- Particle filtering
- Future work:
  - Theory
  - Adversarial setting

Acknowledgments: NSF IIS-0916038, AFOSR FA9550-09-1-0313, and NSF IIS-0953219. We thank Rob Nowak for helpful discussions.