Learning Many Tasks with a Single Teacher

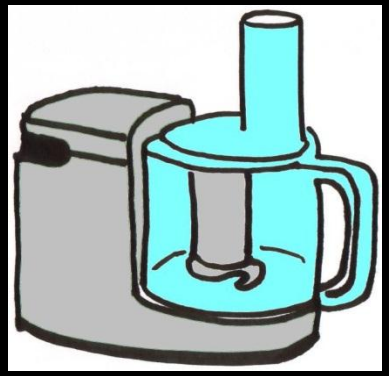
Intel Retreat
ML with 2020 sight

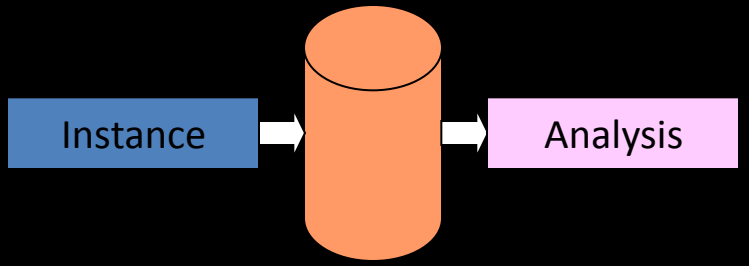
Haim Cohen
Koby Crammer

Haifa
May 2013

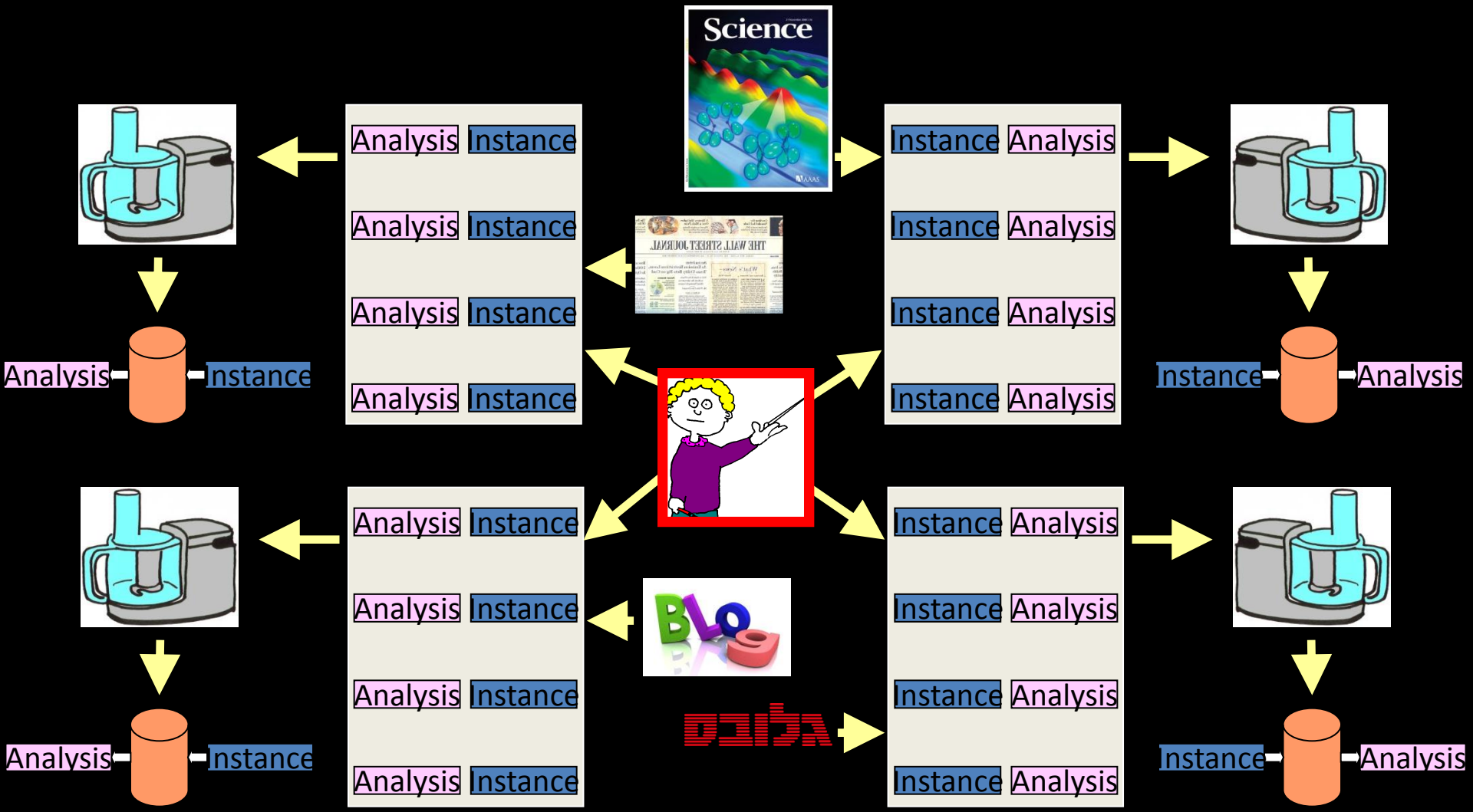
Supervised Learning

Instance	Analysis
Instance	Analysis
Instance	Analysis
Instance	Analysis





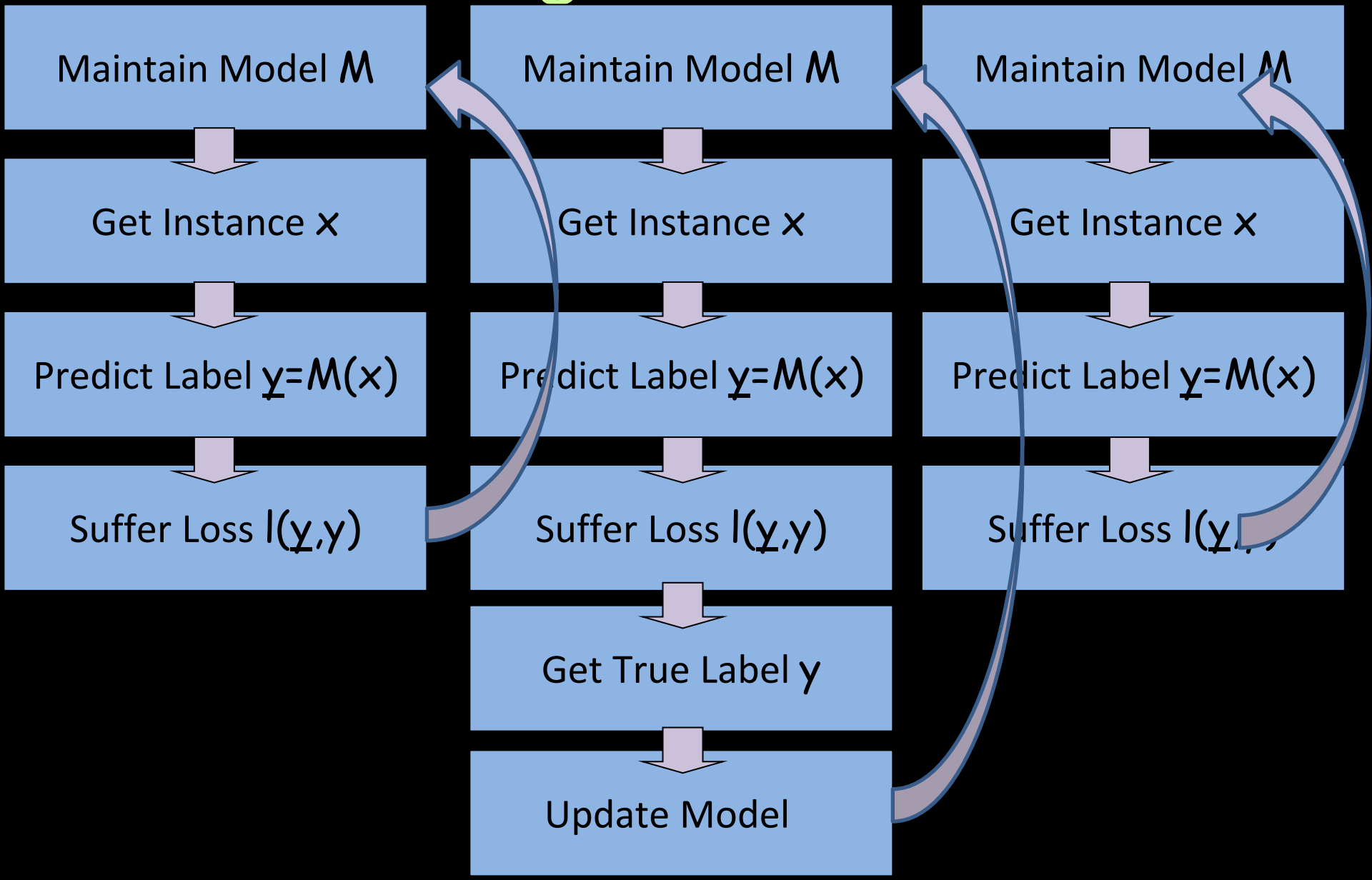
Multi-Task Supervised Learning



Outline

- Setting
- Algorithm
- Analysis
- Empirical Study
- Present and Future

Online Learning With Partial Feedback



How to choose which task?

- Random
- Confidence = abs(margin)
- Our: Probability is monotonically decreasing in the confidence of the model about its prediction

$$P(J = i) = \frac{\left(b + \left|\hat{p}_{i,t}\right| - \min_j \left|\hat{p}_{j,t}\right|\right)^{-1}}{D_t}$$

Algorithm

Parameters: $b \in \mathbb{R}, \quad b > 0.$

Initialization: $w_{i,0} = 0 \in \mathbb{R}^d, \quad (i = 1, \dots, K).$

for t $= 1, 2, \dots n$ **do**

1. Observe K instance vectors, $x_{i,t} \quad (i = 1, \dots, K)$ and calculate the margin $\hat{p}_{i,t} = w_{i,t-1}^T x_{i,t}$.
2. Predict K labels by $\hat{y}_{i,t} = \text{sign}(\hat{p}_{i,t})$.
3. Draw $J \in \{1, \dots, K\}$ with probability

$$P(J = i) = \frac{\left(b + |\hat{p}_{i,t}| - \min_j |\hat{p}_{j,t}|\right)^{-1}}{D_t}$$

   where $j, i \in \{1, \dots, K\}$ and D_t is the normalization factor. Set $Z_{J,t} = 1, \quad Z_{i,t} = 0, \quad \forall i \neq j.$
4. Query the true label, $y_{J,t} \in \{-1,1\}$.
5. Update the weight vector by the standard perceptron update rule

$$w_{J,t} = w_{J,t-1} + M_{J,t} y_{J,t} x_{J,t}$$

$$w_{i,t} = w_{i,t-1} \; \forall i \neq J.$$

end for

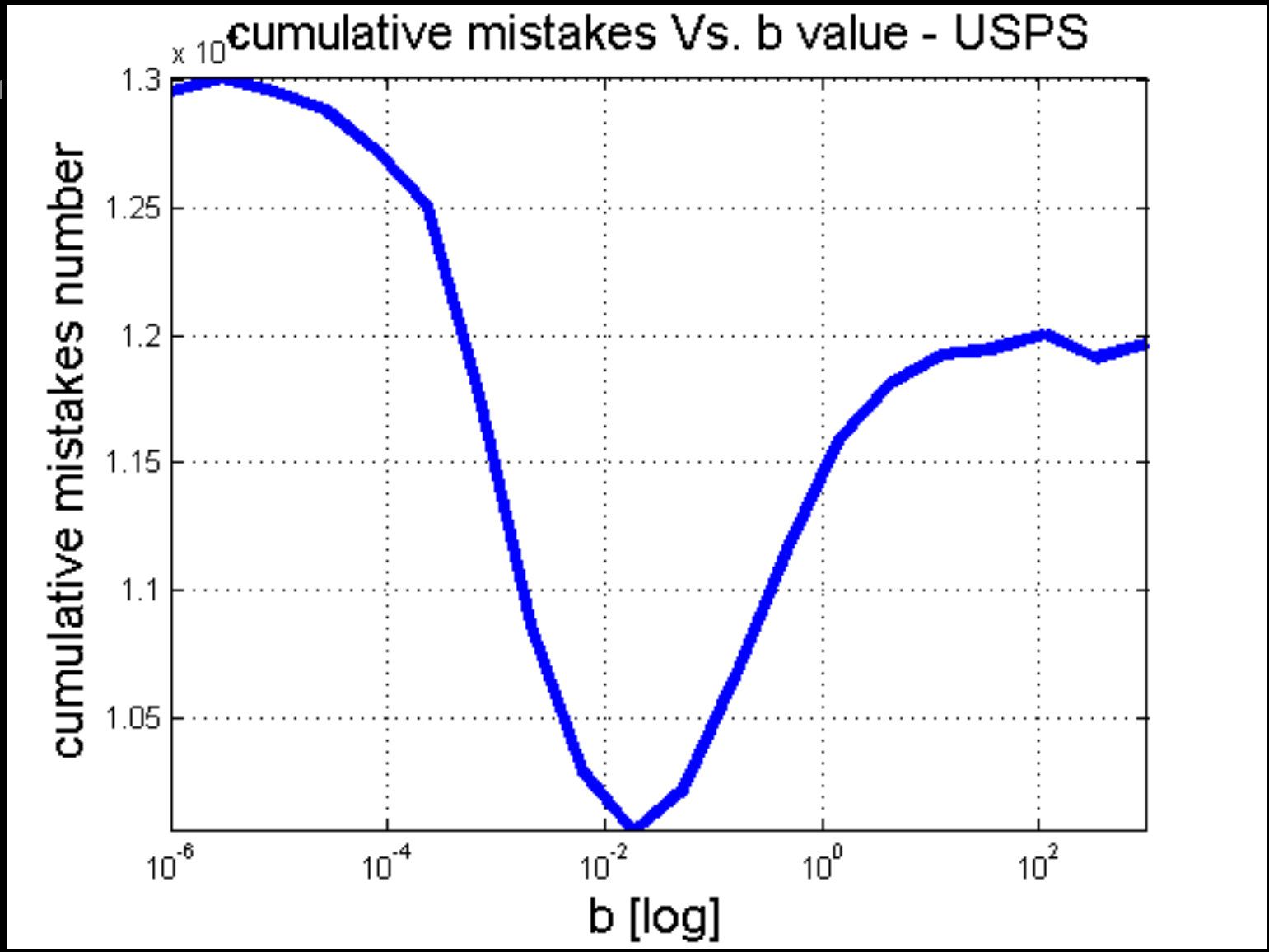
Experiments

- Data:
  - OCR
    - USPS, MNIST
  - Audio
    - Vocal Joistick
  - Text
    - Amazon reviews, News items, SPAM
- Bottom line:
  - Our algorithms improves over baselines

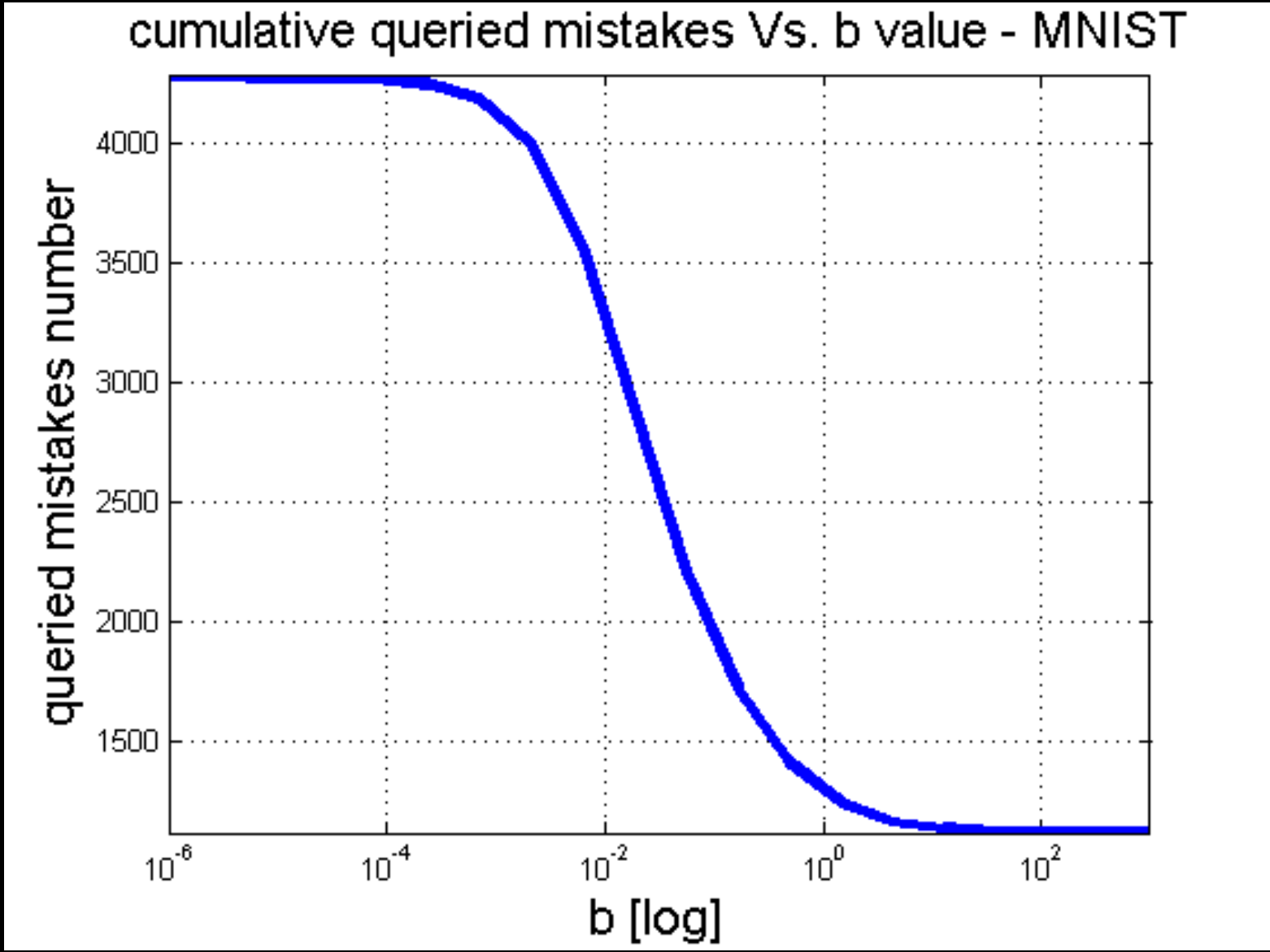
Experiments

- OCR, one-vs-one pairs (45)
- MNIST:
  - 60,000 training, 10,000 test
  - 28x28 pixels
  - 11,273 examples per task (avg 250 queries per task)
- USPS
  - 7,291 training, 2,007 test
  - 16x26 pixels
  - 1,098 per task (avg 24 queries per task)

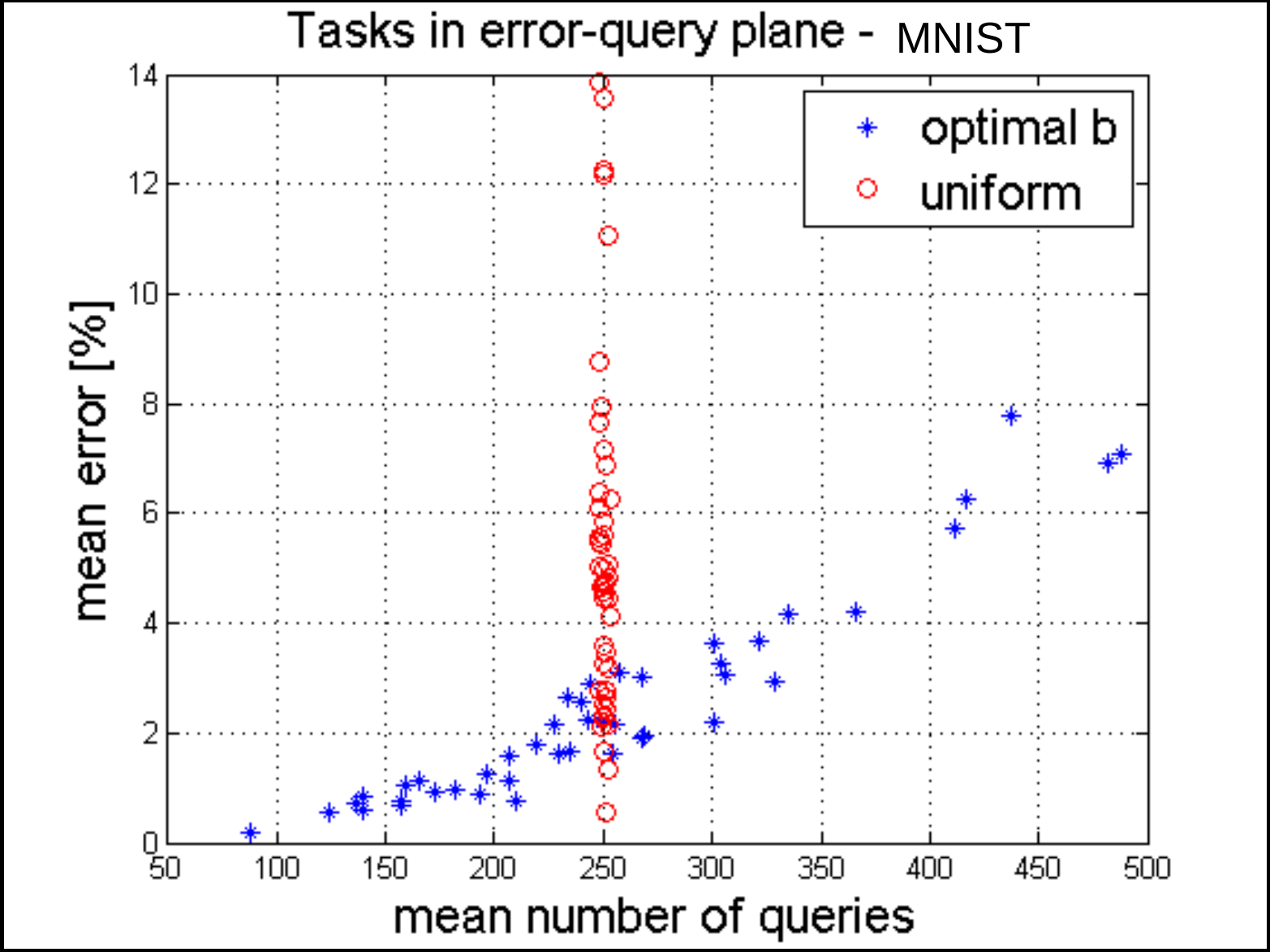
USPS: Cumulative Mistakes



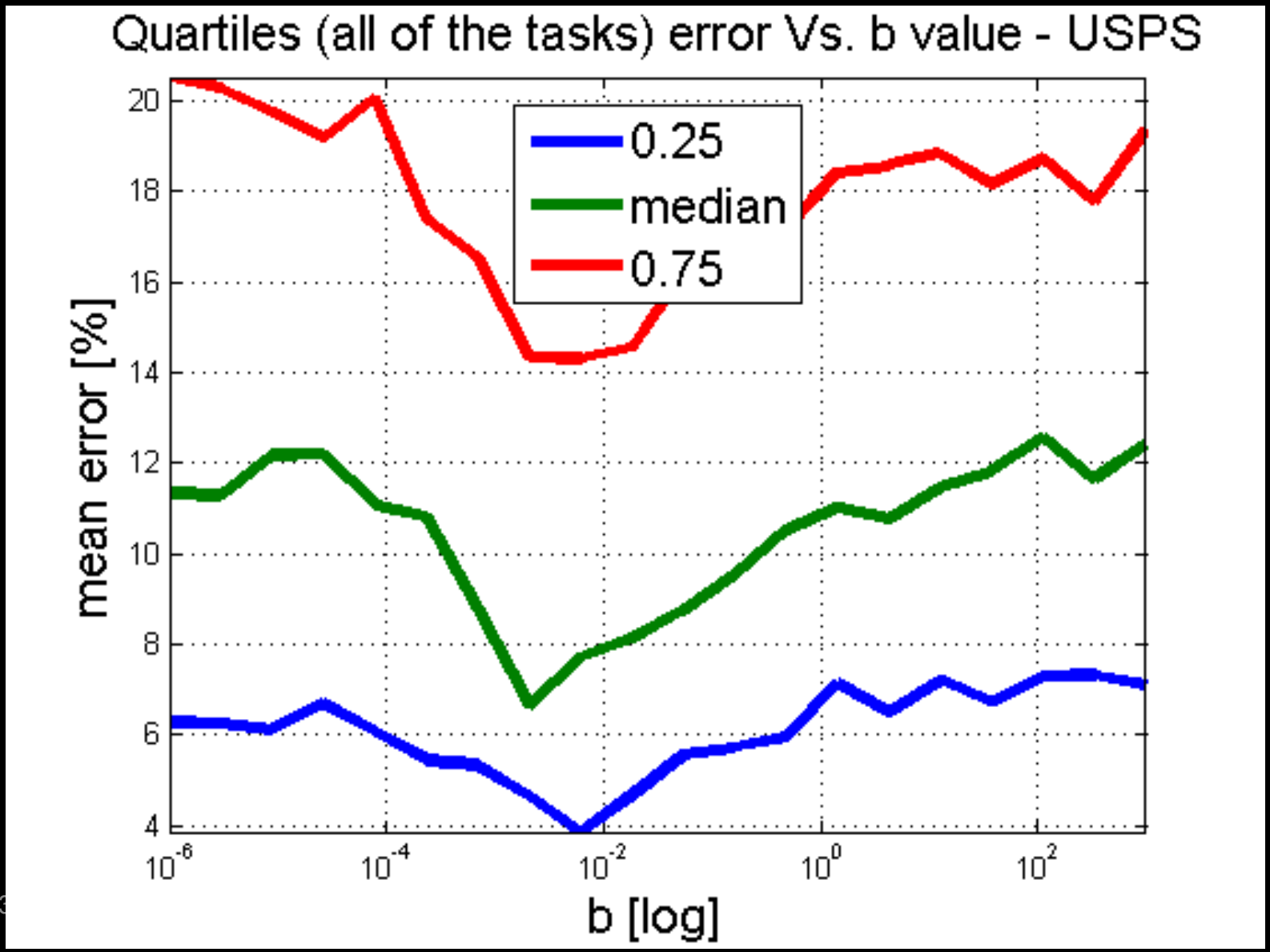
MNIST: Cumulative Queries Mistakes



MNIST


Tasks in error-query plane - MNIST
optimal b
uniform
mean error [%]
mean number of queries
0
2
4
6
8
10
12
14
50
100
150
200
250
300
350
400
450
500

USPS


Quartiles (all of the tasks) error Vs. b value - USPS
0.25
median
0.75
mean error [%]
20
18
16
14
12
10
8
6
4
10^{-6}
10^{-4}
10^{-2}
10^{0}
10^{2}
b [log]

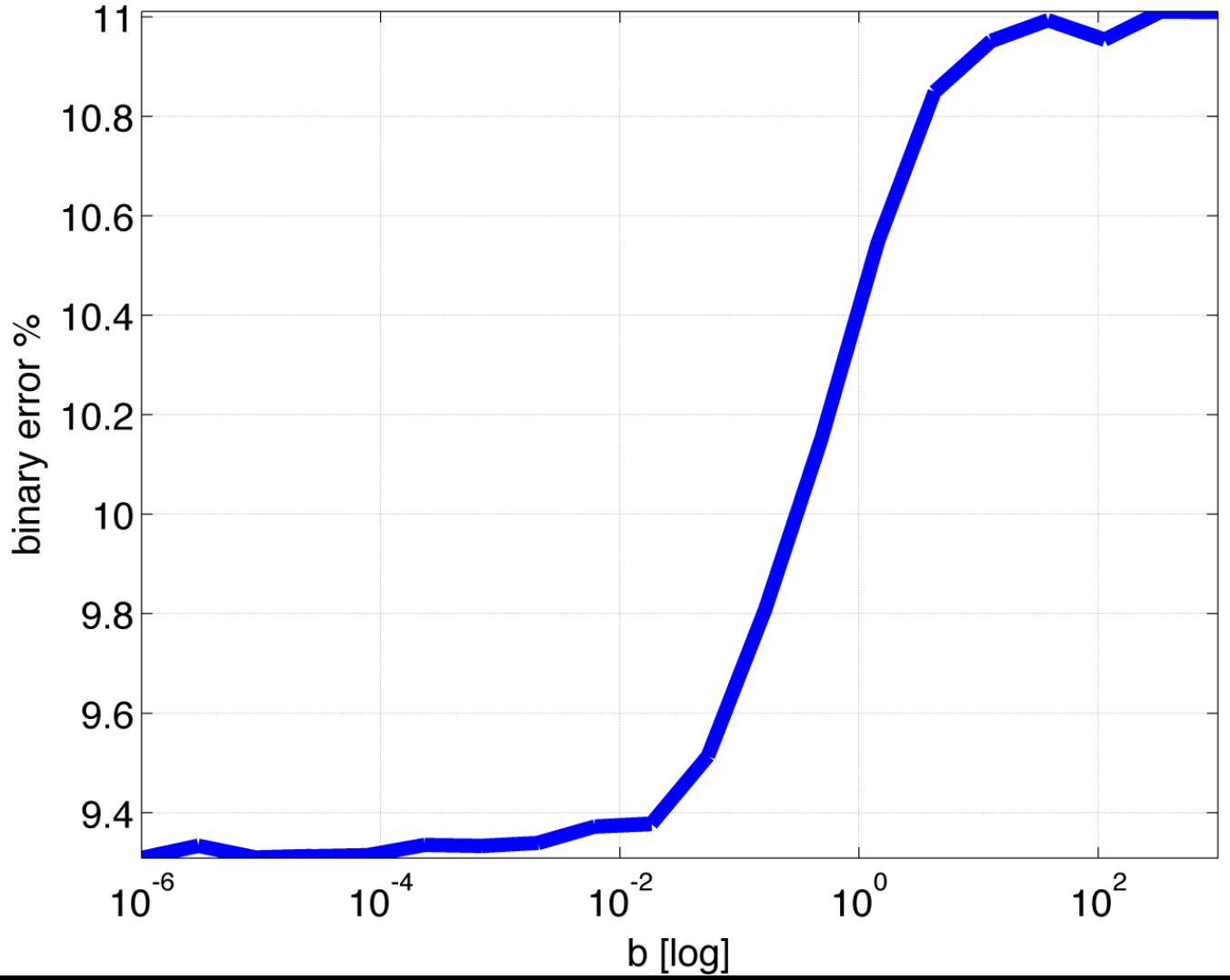
Experiments

- Vocal-Joystick

- Predict one of four/eight vowel (directions)

- V4: 279,484 training, 116,193 test
- V8: 572,911 training 236,680
- ~130,000 training per pair, ~4,640 queries per task
- 13 MFCC -> 27 features

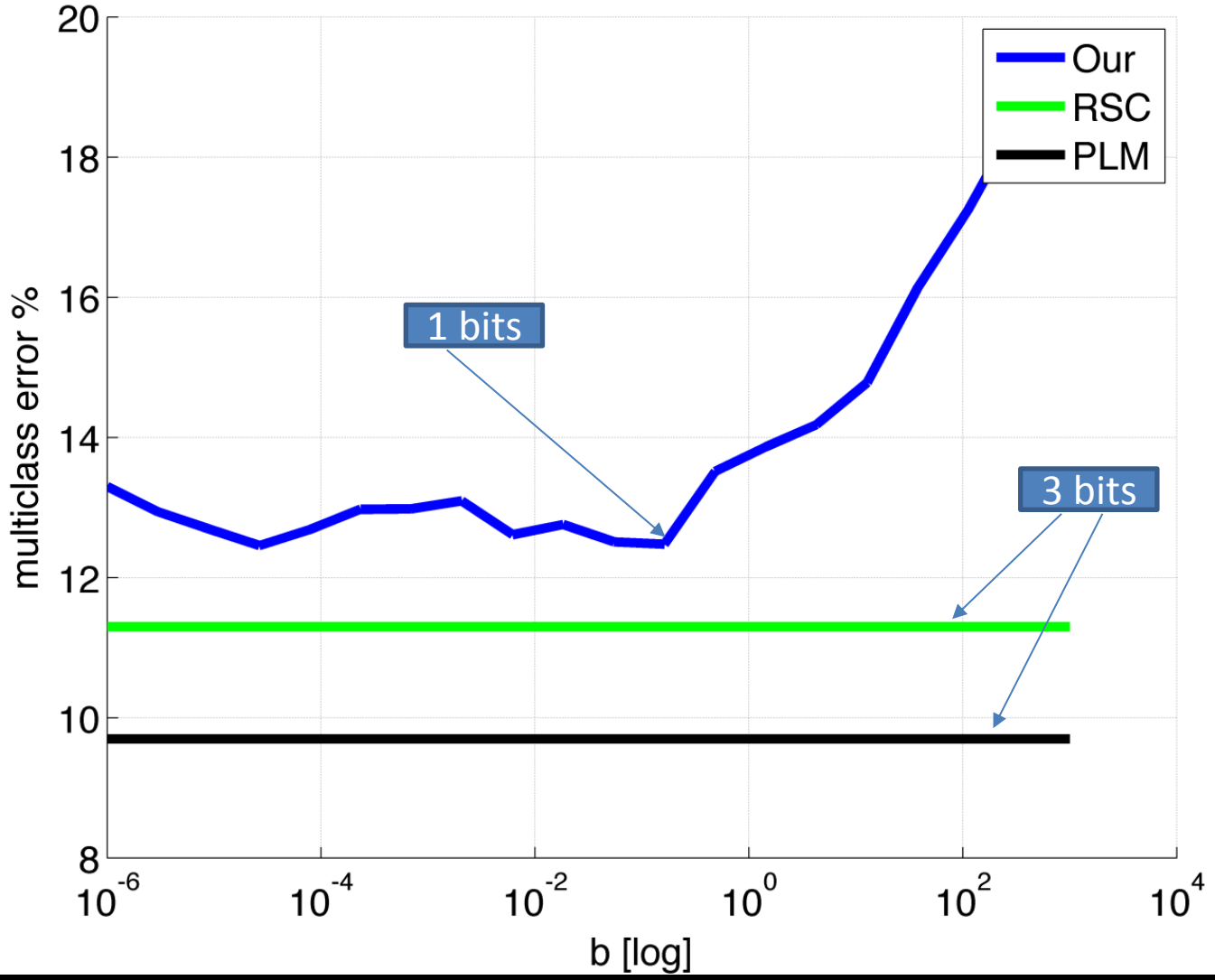
VJ


11
10.8
10.6
10.4
10.2
10
9.8
9.6
9.4
binary error %
10^-6
10^-4
10^-2
10^0
10^2
b [log]

Three-way feedback

- Tasks: all-pairs, goal multi-class prediction
- Given input and a pair of labels, feedback is:
  - not-relevant if true label is not included in pair
  - Identity of label form pair, if it is
- Problem:
  - If not relevant, current algorithm will ignore input, may repeat querying same pair
- Fix:
  - Update even if not in pair, lower rate

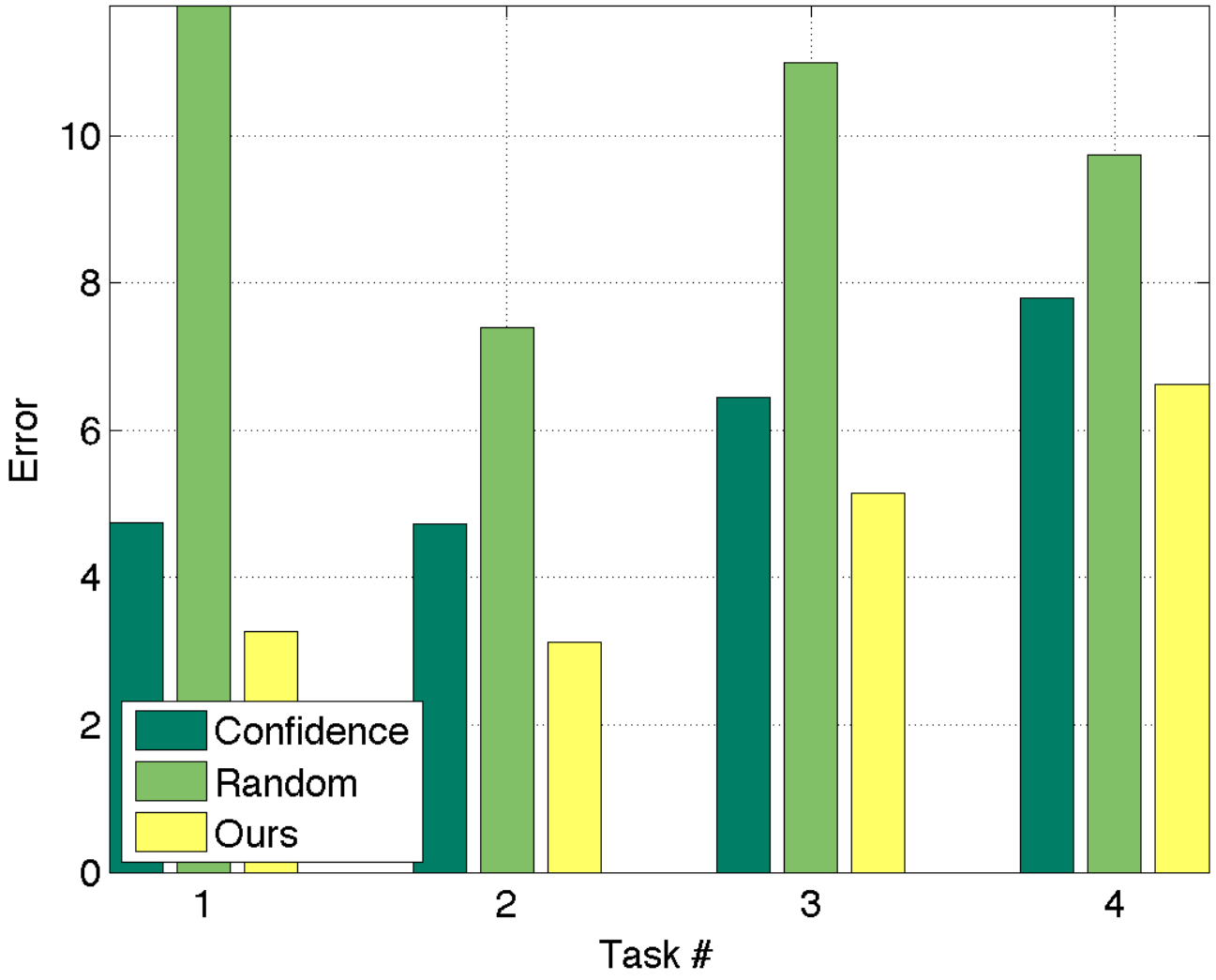
VJ


Our
RSC
PLM
1 bits
3 bits
multiclass error %
b [log]
20
18
16
14
12
10
8
10^{-6}
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}

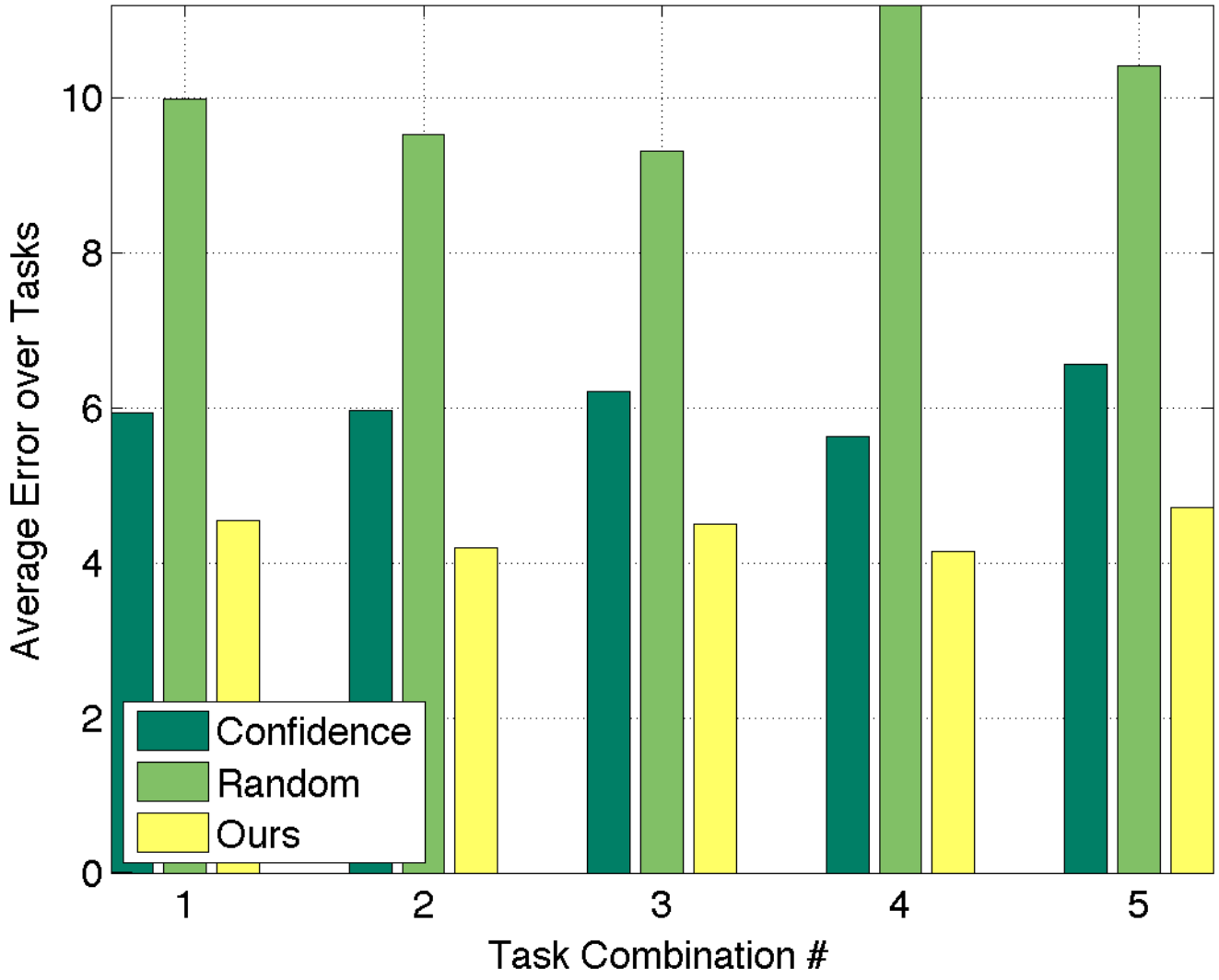
Text Categorization

- Data:
  - Sentiment in Amazon Reviews
  - Topic in Reuters news items
  - Spam (or not) in emails
- About 4,000 documents per task
- 5 Combinations of 4/8 tasks

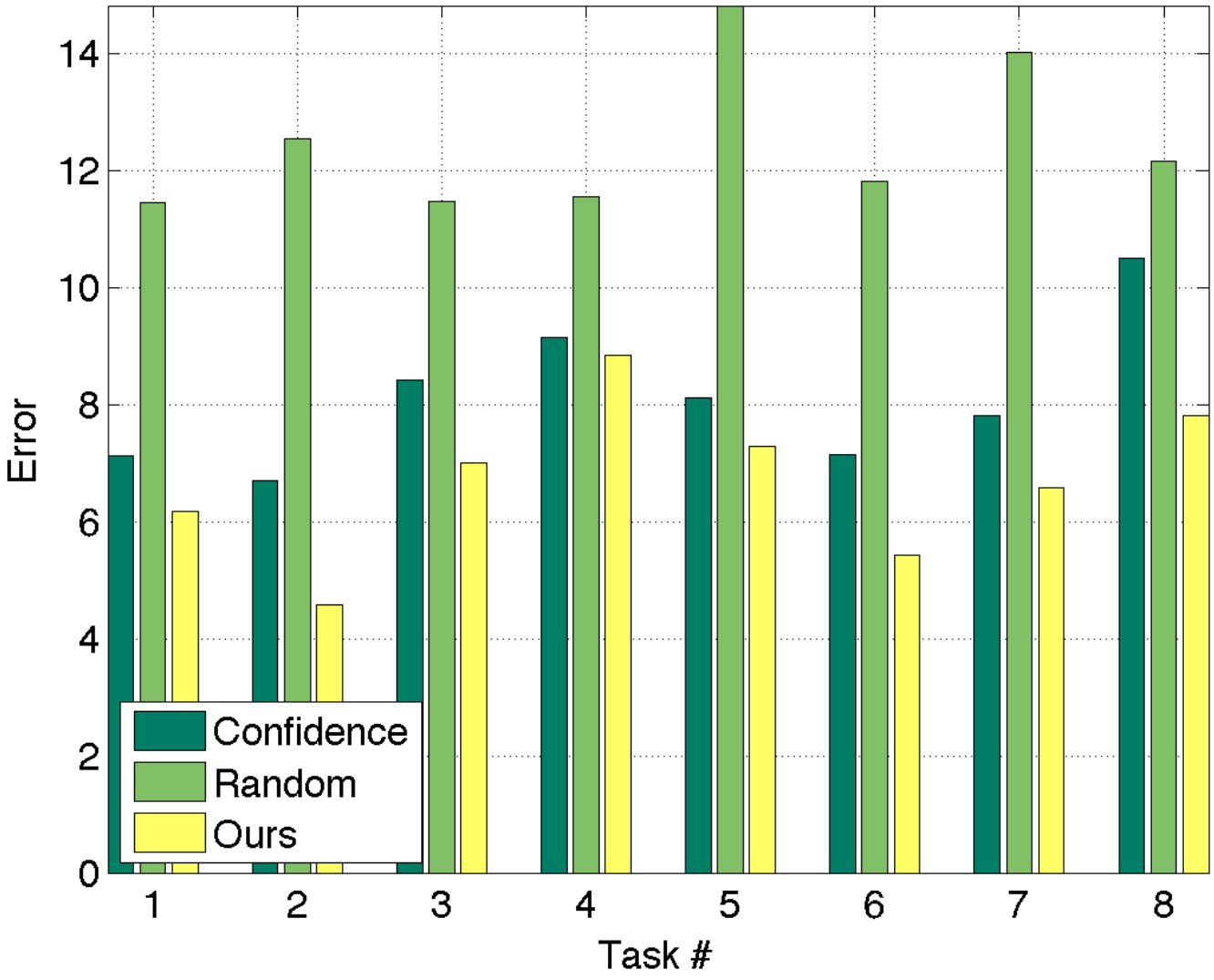
Combinations of 4 Tasks



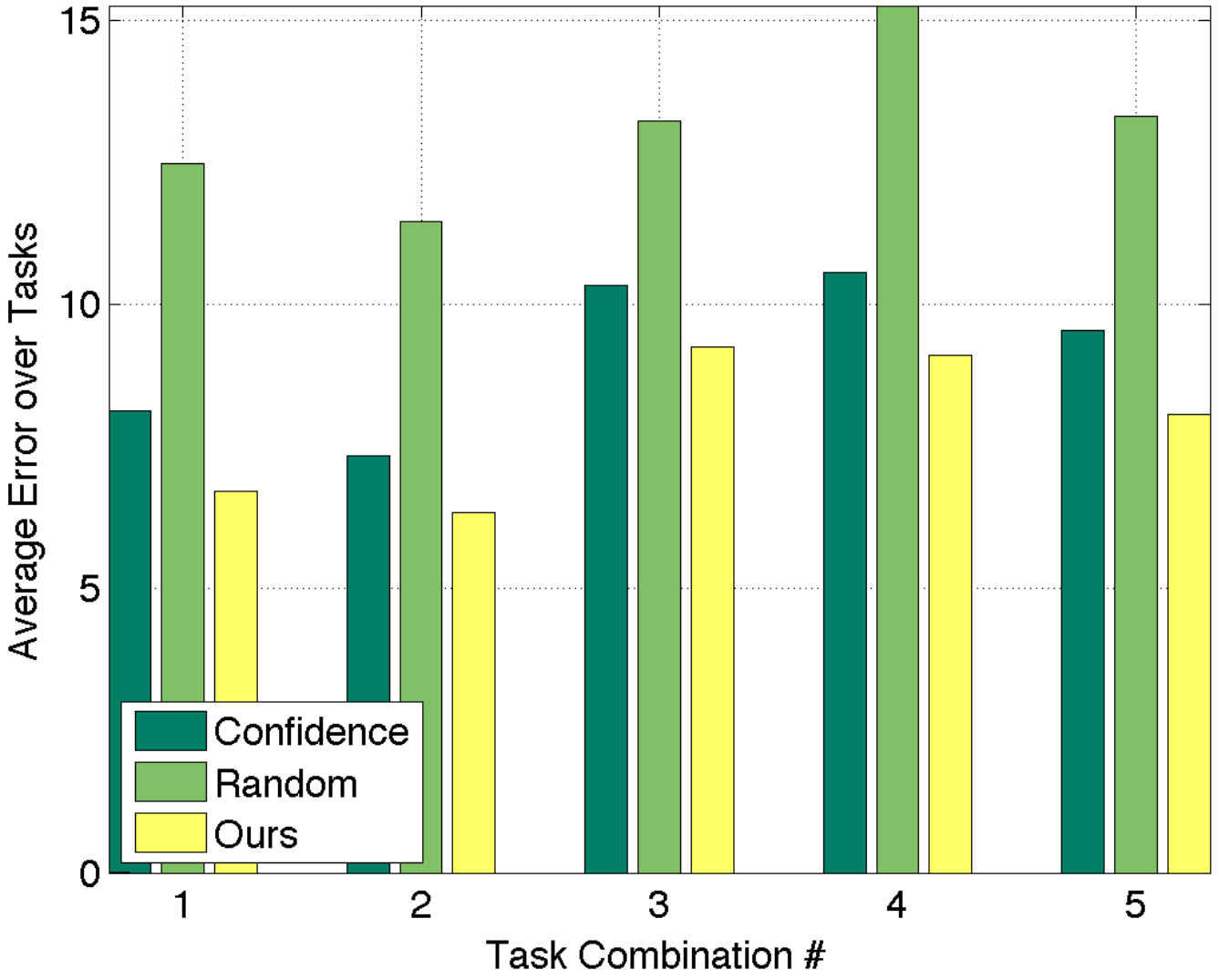
Combinations of 4 Tasks



Combinations of 8 Tasks



Combinations of 8 Tasks



Analysis

Theorem: If a multitask selective sampling perceptron algorithm run on K tasks with K parallel example pairs sequences $(x_{i,1}, y_{i,1}), \ldots, (x_{i,n}, y_{i,n}) \in \mathbb{R}^d \times \{-1,1\}$, $i = 1, \ldots, K$ with input parameter $b > 0$, then for all $\gamma > 0$, all $u_i \in \mathbb{R}^d$ and all $n \geq 1$,

$$\mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{n} M_{i,t}\right] \leq \frac{K}{\gamma}\left(1 + \frac{X^2}{2b}\right)\bar{L}_{\gamma,n} + \frac{K(2b + X^2)^2 U}{8\gamma^2 b}$$

Where $X = \max_{i,t} |x_{i,t}|$, $\bar{L}_{\gamma,n} = \mathbb{E}[\sum_{i=1}^{K}\sum_{t=1}^{n} Z_{i,t} M_{i,t} \ell_{\gamma,i,t}(u_i)]$ and $U = \sum_{i=1}^{K} \|u_i\|^2$.

Analysis

Corollary: There exists a value of the input parameter b, such that the expected number of mistakes is bounded

$$\mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{n} M_{i,t}\right] \leq K\bar{L}_{1,n} + 2K\sqrt{X^2 U \bar{L}_{1,n}}$$

Where $X = \max_{i,t} |x_{i,t}|$,$\bar{L}_{\gamma,n} = \mathbb{E}[\sum_{i=1}^{K}\sum_{t=1}^{n} Z_{i,t} M_{i,t} \ell_{\gamma,i,t}(u_i)]$ and $U = \sum_{i=1}^{K} \|u_i\|^2$.

Multi-Task: Previous Setting

- Full training data
- Share datasets
- Explore relation/similarity between tasks
  - Dekel, Long, Singer 2006
  - Argyriou, Evgeniou, Pontil, 2008
  - Agarwal, Raklin, Bartlettt, 2008
  - Saha, Rai, Daume III, Venkatasubramanian 2011
  - Kakade, Shalev-Shwartz, Tweri 2012
  - …

Multi-Task: Our Setting

- Unlabeled training data
- Shared annotator
- No assumptions about relation/similarity
- Assume:
  - Some tasks are harder than others
  - Some inputs are harder than other
  - Not simultaneously

Relative Partial-Feedback advantages

- Simple:
  - Focused binary feedback request
- Intuitive:
  - Can be interpreted as an intuitive question
- Easy:
  - Feedback from non-experts

Larger Picture

- Design, analyze and experiment with interactive algorithms that minimize their required feedback
  - Amount of inputs to be labeled
  - Amount of feedback per input
  - Type of feedback (direct / indirect)
  - Simple, Intuitive, Easy, Informative
  - “Seamless Supervision”