Adaptive Region-Based Active Learning

Corinna Cortes
Google Research
New York, NY

Giulia DeSalvo
Google Research
New York, NY

Claudio Gentile
Google Research
New York, NY

Mehryar Mohri
Google Research & Courant
New York, NY

Ningshan Zhang
New York University
New York, NY

Introduction

■ **ARBAL** (Adaptive Region-Based Active Learning)

1. Adaptively partitions the input space into a finite number of regions
2. Subsequently seeks a distinct predictor for each region

- both phases actively requesting labels.

prove theoretical guarantees for both the generalization error and the label complexity

Algorithm

IWAL-query

The same hypothesis space H

Algorithm 1 ARBAL$(\mathcal{H}, \tau, \kappa, (\gamma_t)_{t\in[T]})$

$K \leftarrow 1, \mathcal{X}_1 \leftarrow \mathcal{X}, \mathcal{H}_1 \leftarrow \mathcal{H}$
for $t \in [T]$ **do**
  Observe x_t; set $k_t \leftarrow k$ such that $x_t \in \mathcal{X}_k$
  $p_t \leftarrow \max_{h,h' \in \mathcal{H}_{k_t}, y\in\mathcal{Y}} \ell(h(x_t), y_t) - \ell(h'(x_t), y_t)$
  $Q_t \leftarrow \textsc{Bernoulli}(p_t)$
  if $Q_t = 1$ **then**
    $y_t \leftarrow \textsc{Label}(x_t)$
  end if
  if $t \leq \tau$ and $K < \kappa$ **then**
    $\mathcal{X}_l, \mathcal{X}_r \leftarrow \textsc{Split}(\mathcal{X}_{k_t}, \gamma_t)$ `# split phase`
    if split **then**
      $K \leftarrow K+1, \mathcal{X}_{k_t} \leftarrow \mathcal{X}_l, \mathcal{X}_K \leftarrow \mathcal{X}_r$
      $\mathcal{H}_K \leftarrow \mathcal{H}, \mathcal{H}_{k_t} \leftarrow \mathcal{H}$
    end if
  else
    $\mathcal{H}_{k_t} \leftarrow \textsc{Update}(\mathcal{H}_{k_t})$ `# IWAL phase`
  end if
end for
return $\widehat{h}_T \leftarrow \sum_{k=1}^{K} 1_{x\in\mathcal{X}_k} \widehat{h}_{k,T}$

SPLIT phase

SPLIT splits a region if and only if the best-in-class error is likely to improve by a strictly positive amount

Algorithm 2 $\textsc{Split}(\mathcal{X}_k, \gamma)$

for $d \in [D]$ and $c \in \mathbb{R}$ **do**
  $(\mathcal{X}_l, \mathcal{X}_r) \leftarrow \textsc{RegSplit}(\mathcal{X}_k, d, c)$
  $\gamma_{d,c} \leftarrow \mathrm{p}_k \left[L_{k,t}(\widehat{h}_{k,t}) - L_{k,t}(\widehat{h}_{lr,t}) - \sqrt{\frac{2\sigma_T}{T_{k,t}}} \right]$
end for
$(d^*, c^*) \leftarrow \operatorname{argmax}_{d \in [D], c \in \mathbb{R}} \gamma_{d,c}$
if $\gamma_{d^*,c^*} \geq \gamma$ **then**
  $\mathcal{X}_l^* \leftarrow \{x \in \mathcal{X}_k : x[d^*] \leq c^*\}$ # split
  $\mathcal{X}_r^* \leftarrow \{x \in \mathcal{X}_k : x[d^*] > c^*\}$
  return $\mathcal{X}_l^*, \mathcal{X}_r^*$
else
  return $\emptyset$ # no split
end if

the splitting parameters (d, c)

$$L_{k,t}(h) = \frac{1}{T_{k,t}} \sum_{s \in [t], x_s \in \mathcal{X}_k} \frac{Q_s}{p_s} \ell(h(x_s), y_s)$$

$$\sigma_T = \kappa D \log \left[\frac{8T^3 |\mathcal{H}|^3 \kappa D}{\delta} \right]$$

SPLIT phase

$$\sigma_T = \kappa D \log\left[\frac{8T^3|\mathcal{H}|^3\kappa D}{\delta}\right]$$

$$\boldsymbol{\sigma_T = 0.01/\sqrt{T_k}}$$

Lemma 1. *With probability at least $1-\delta/4$, for all binary trees with (at most) κ leaf nodes, the improvement in the minimal empirical error by splitting concentrates around the improvement in the best-in-class error:*

$$\left|[R_k(h_k^*) - R_k(h_{lr}^*)] - [L_{k,t}(\widehat{h}_{k,t}) - L_{k,t}(\widehat{h}_{lr,t})]\right| \le \sqrt{\frac{2\sigma_T}{T_{k,t}}}.$$

Corollary 2. *With probability at least $1-\delta/4$, for all splits made by* ARBAL, *the improvement in the best-in-class error is at least γ_t, where γ_t is the threshold at the time of split.*

$$\gamma_{d,c} \leftarrow \mathbf{p}_k\left[L_{k,t}(\widehat{h}_{k,t}) - L_{k,t}(\widehat{h}_{lr,t}) - \sqrt{\frac{2\sigma_T}{T_{k,t}}}\right]$$

Algorithm

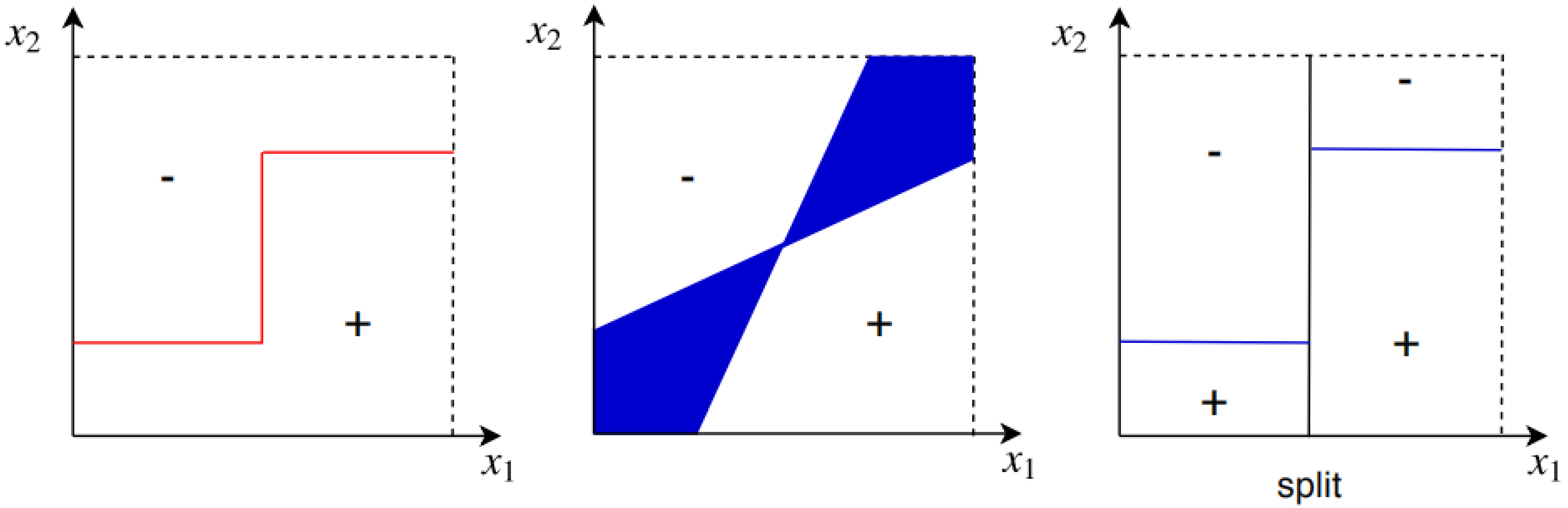
IWAL-query

The same hypothesis space H

Algorithm 1 $\textsc{ARBAL}(\mathcal{H}, \tau, \kappa, (\gamma_t)_{t\in[T]})$

$K \leftarrow 1, \mathcal{X}_1 \leftarrow \mathcal{X}, \mathcal{H}_1 \leftarrow \mathcal{H}$
for $t \in [T]$ **do**
  Observe x_t; set $k_t \leftarrow k$ such that $x_t \in \mathcal{X}_k$
  $p_t \leftarrow \max_{h,h' \in \mathcal{H}_{k_t}, y\in\mathcal{Y}} \ell(h(x_t), y_t) - \ell(h'(x_t), y_t)$
  $Q_t \leftarrow \textsc{Bernoulli}(p_t)$
  if $Q_t = 1$ **then**
    $y_t \leftarrow \textsc{Label}(x_t)$
  end if
  if $t \le \tau$ and $K < \kappa$ **then**
    $\mathcal{X}_l, \mathcal{X}_r \leftarrow \textsc{Split}(\mathcal{X}_{k_t}, \gamma_t)$ # split phase
    if split **then**
      $K \leftarrow K + 1, \mathcal{X}_{k_t} \leftarrow \mathcal{X}_l, \mathcal{X}_K \leftarrow \mathcal{X}_r$
      $\boxed{\mathcal{H}_K \leftarrow \mathcal{H}, \mathcal{H}_{k_t} \leftarrow \mathcal{H}}$
    end if
  else
    $\mathcal{H}_{k_t} \leftarrow \textsc{Update}(\mathcal{H}_{k_t})$ # IWAL phase
  end if
end for
return $\widehat{h}_T \leftarrow \sum_{k=1}^{K} 1_{x\in\mathcal{X}_k} \widehat{h}_{k,T}$

SPLIT phase



- ARBAL maintains throughout the split phase the original hypothesis space H
- The shrinkage of H only takes place during the IWAL phase

IWAL phase

■ **IWAL** (Importance Weighted Active Learning)[1]

Based on the largest possible disagreement among the current set of hypotheses on the current input: flips a coin $Q_t \in \{0, 1\}$ with bias $p_t = p(x_t)$

$$p_t = \max_{h, h' \in \mathcal{H}_t, y \in \mathcal{Y}} \ell(h(x_t), y) - \ell(h'(x_t), y)$$

$$\widehat{h}_T = \mathrm{argmin}_{h \in \mathcal{H}_T} \sum_{t=1}^{T} Q_t \ell(h(x_t), y_t)/p_t$$

$$\mathcal{H}_{k,t} = \left\{ h \in \mathcal{H}_{k,t-1} : L_{k,t}(h) \leq \min_{h \in \mathcal{H}_{k,t-1}} L_{k,t}(h) + \sqrt{\frac{8\sigma_T}{T_{k,t}}} \right\}$$

[1]Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. In *Proceedings of ICML*, pp. 49–56. ACM, 2009.

Theoretical analysis

$$\sigma_T = \kappa D \log\left[\frac{8T^3|\mathcal{H}|^3\kappa D}{\delta}\right]$$

- **with a fixed γ**

$$\boldsymbol{\sigma_T = 0.01/\sqrt{T_k}}$$

Theorem 3. *Assume that a run of* ARBAL *over T rounds has split the input space into K regions. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds:*

$$R(\widehat{h}_T) \leq R_U + \sqrt{\frac{32K\sigma_T}{T}} + \frac{16K\sigma_T}{T}$$

where $R_U = R^* - \gamma(K-1)$ is an upper bound on the best-in-class error obtained by ARBAL. *Moreover, with probability at least $1-\delta$, the expected number of labels requested, $\tau_T = \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathcal{D}_\mathcal{X}}\left[p_t | \mathcal{F}_{t-1}\right]$, satisfies*

$$\tau_T \leq \min\{2\theta r_0, 1\}\tau + 4\theta_{\max}(T-\tau)\left[R_U + 8\sqrt{\frac{K\sigma_T}{T-\tau}}\right] + \sqrt{32K\sigma_T}.$$

Theoretical analysis

- with a adaptive γ_t $\quad \rho$: $R_k(h_k^*) - R_k(h_{lr}^*) \geq \rho \quad \rho = 0.01$

Proposition 4. *Let* ARBAL *be run with* $\gamma_t = \rho\mathbb{P}(\mathcal{X}_{k_t})/2$. *Then, for any* $\delta > 0$, *with probability at least* $1 - \delta/2$, *the first split occurs before round* $\left\lceil 2\sigma_T(\frac{4}{\rho} + 1)^2 \right\rceil$.

$$\sigma_T = \kappa D \log\left[\frac{8T^3|\mathcal{H}|^3\kappa D}{\delta}\right]$$

$$\min\{\mathbb{P}(\mathcal{X}_l), \mathbb{P}(\mathcal{X}_r)\} \geq c\,\mathbb{P}(\mathcal{X}_k), \text{ with } 0 < c < 0.5$$

$$\sigma_T = 0.01/\sqrt{T_k}$$

Corollary 5. *Let* ARBAL *run with* $\gamma_t = \mathbb{P}(\mathcal{X}_{k_t})\rho/2$. *Then, with probability at least* $1 - \delta/2$, ARBAL *splits more than* $\min\left\{\log_{1/c}\left[\frac{\tau}{2\sigma_T(\frac{4}{\rho}+1)^2}\right], \kappa - 1\right\}$ *times by the end of the split phase.*

Experiment

logistic loss function $\boldsymbol{\sigma_T = 0.01/\sqrt{T_k}}$

The initial hypothesis set H consists of 3,000 randomly drawn hyperplanes with bounded norms.

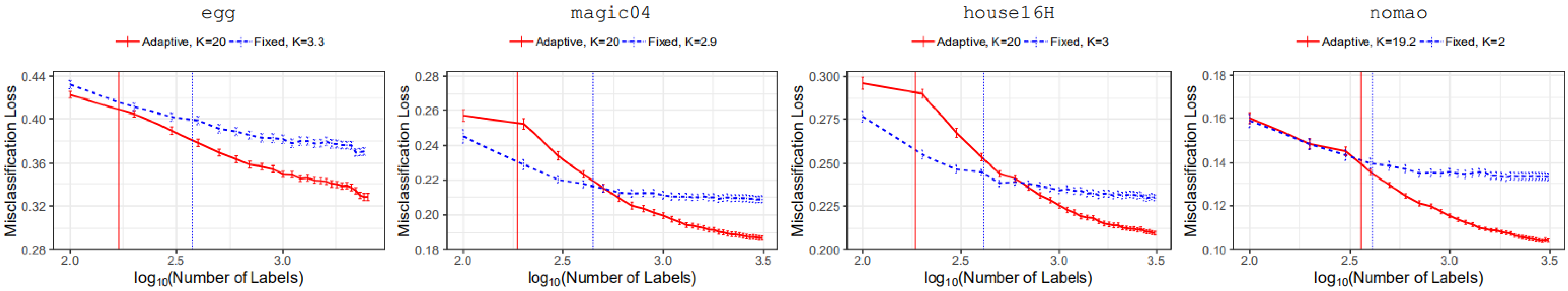


Figure 2: Misclassification loss of ARBAL with fixed and adaptive threshold γ on held out test data vs. number of labels requested ($\log_{10}$ scale), with $\kappa = 20$ and $\tau = 800$. The vertical lines indicate the end of the first (split) phase.

Experiment

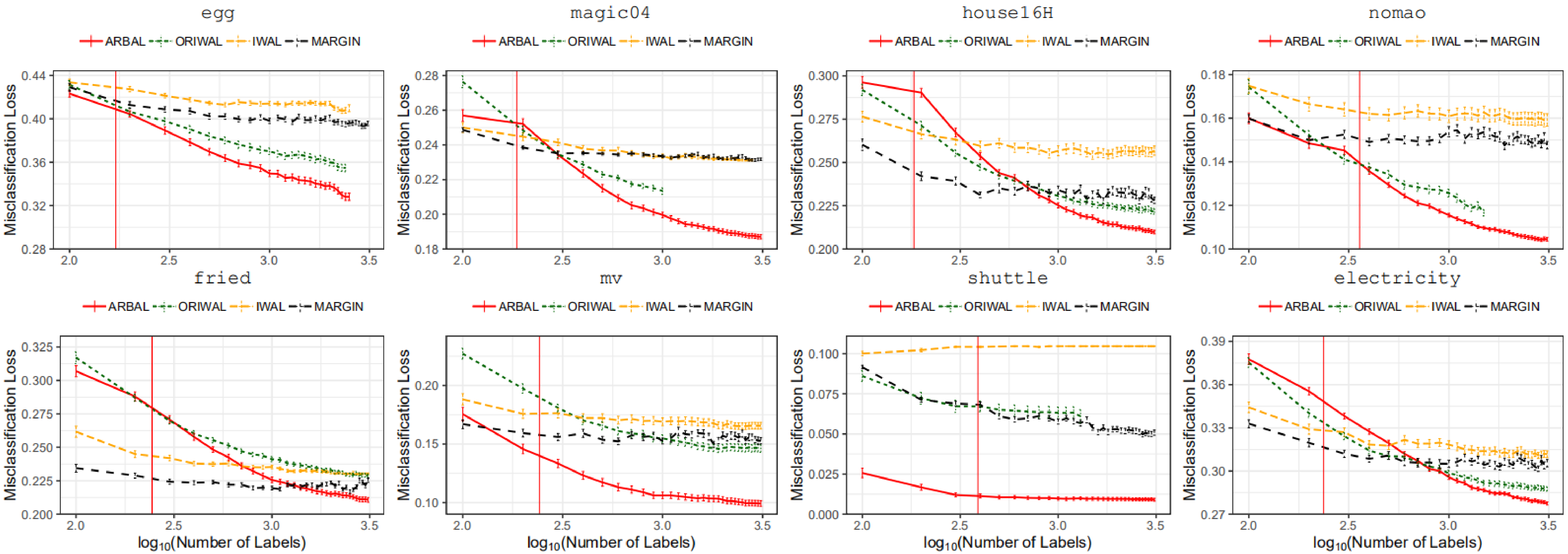


Figure 3: Misclassification loss of ARBAL(with adaptive γ_t), ORIWAL, IWAL, and MARGIN on hold out test data vs. number of labels requested ($\log_{10}$ scale), with $\kappa = 20$ and $\tau = 800$. The ARBAL curves are repetitions from Figure 2.

Theoretical analysis

Define the distance $\rho(f,g)$ between two hypotheses $f,g \in \mathcal{H}$ as $\rho(f,g) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left|\ell(f(x),y) - \ell(g(x),y)\right|$

The generalized disagreement coefficient $\theta(\mathcal{D},\mathcal{H})$ of a class of functions $\mathcal{H}$ with respect to distribution $\mathcal{D}$ is defined as the minimum value of θ, such that for all $r > 0$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_X}\left[\sup_{h\in\mathcal{H},\rho(h,h^*)\le r, y\in\mathcal{Y}} \left|\ell(h(x),y)-\ell(h^*(x),y)\right|\right] \le \theta r$$

disagreement coefficient $\theta_k = \theta(\mathcal{D}_k,\mathcal{H})$

$r_0 = \max_{h\in\mathcal{H}} \rho(h,h^*)$. Let $\mathcal{F}_t$ denotes the σ-algebra generated by $(x_1,y_1,Q_1),\ldots,(x_t,y_t,Q_t)$.