Efficient Online Decision Tree Learning with Active Feature Acquisition

Arman Rahbar[1], Ziyu Ye[2], Yuxin Chen[2], Morteza Haghir Chehreghani[1]

[1]Chalmers University of Technology
[2]University of Chicago

IJCAI 2023

Agenda

- Introduction
- Proposed framework
- Results

Introduction

Decision trees:

- Interpretable

Online variants of decision trees:

- Medical diagnosis
- Intrusion detection

Classical online DT learners:

- Require all features of incoming data points
- Not fully online

Proposed framework

We take feature acquisition cost into account:

At each time step t

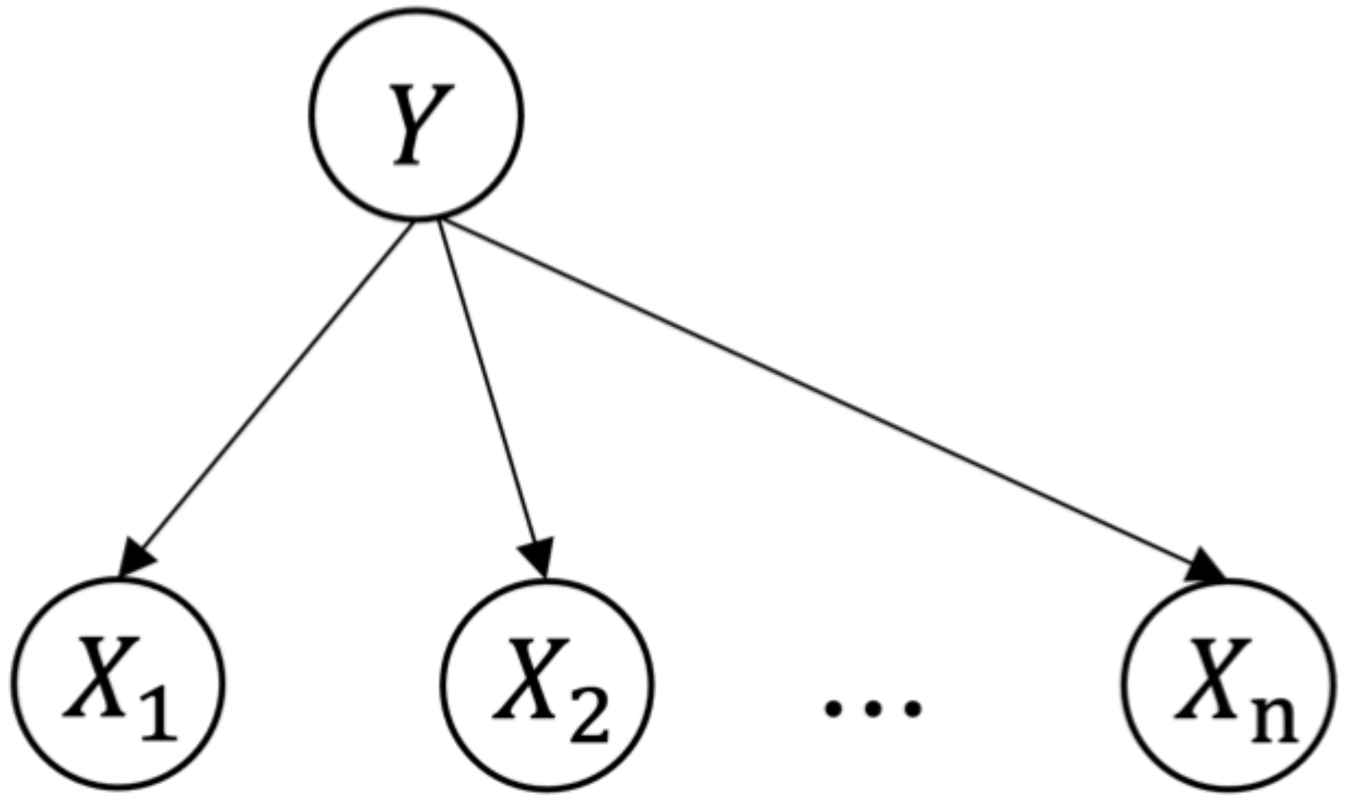
1. Receive a data point with unknown features
2. Make prediction (i.e., classify) with low feature acquisition cost
3. Receive correct prediction for learning

Example: medical diagnosis:

At time t:

1. Patient $\mathbf{x}^t$ comes in
   - n medical tests: $x_1^t, x_2^t, \ldots, x_n^t$, results unknown initially and can be measured at a cost
2. Predict accurate treatment with a low cost

A New Problem Formulation



Data point: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$

R.V. for feature values: $X_i \in \mathcal{X} \triangleq \{0,1\}$

Random variable for labels: $Y_j \in \mathcal{Y} \triangleq \{y_1, y_2, \ldots, y_m\}$

Naive Bayes assumption

- Joint distribution: $\mathbb{P}[Y_j] \prod_{i=1}^{n} \mathbb{P}[X_i \mid Y_j]$

$\theta_{ij} \triangleq \mathbb{P}[X_i = 1 \mid Y_j]$, $\theta_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$

A New Problem Formulation

Random variable for *hypothesis* of data points:

full realization of features

$$H = [X_1, \ldots, X_n], h \in \mathcal{H} \triangleq \{0,1\}^n$$

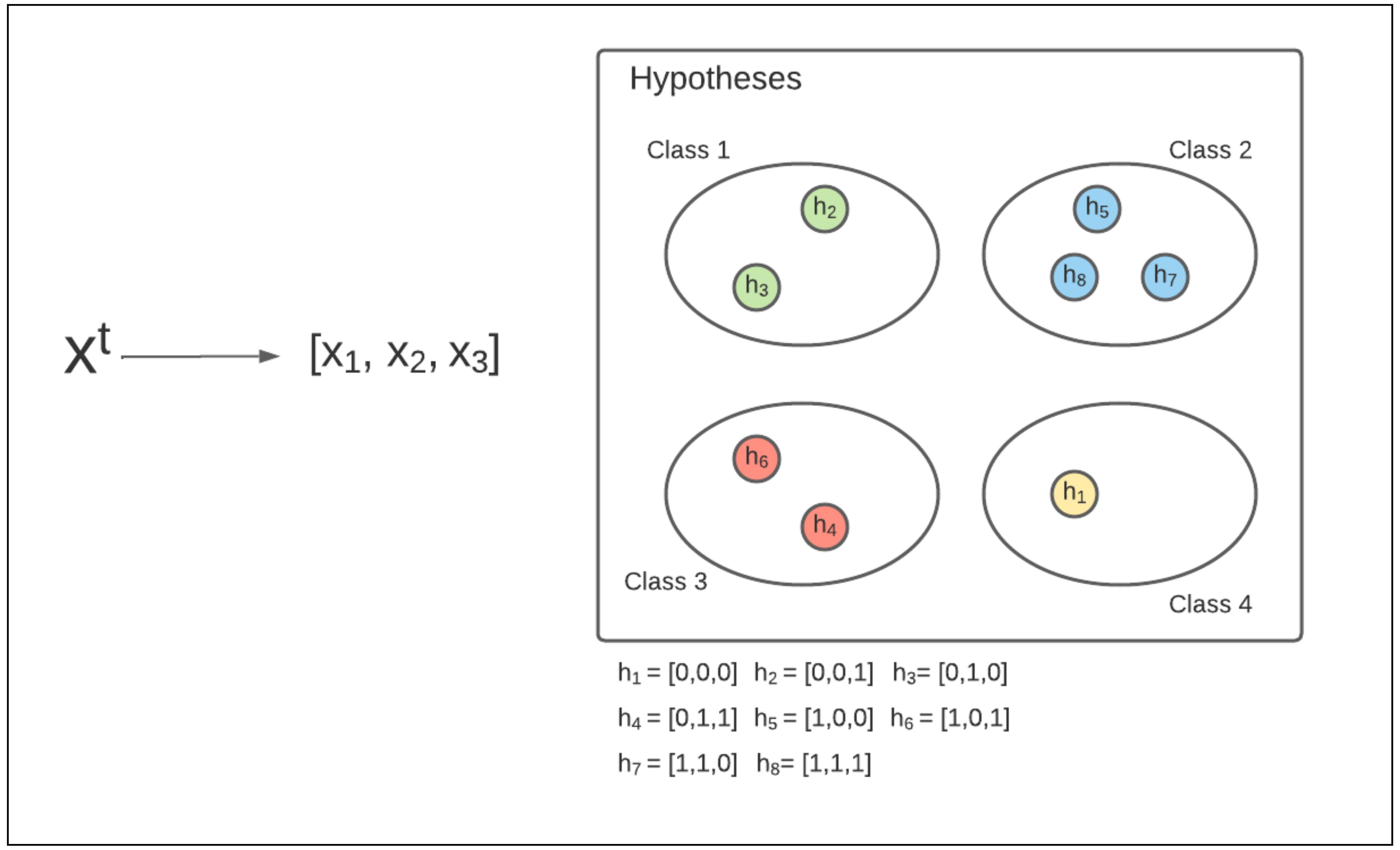
$\mathcal{H}$ partitioned into m disjoint

decision regions / labels / classes

Query q has cost $c(q)$

prediction $\hat{y}$ has loss $l(\hat{y}, y)$

Goal: low $c(q)$ and low $l(\hat{y}, y)$



Offline Prediction

$$\theta_{ij} \sim Beta(\alpha_{ij}, \beta_{ij})$$

$$\theta_{ij} \triangleq \mathbb{P}[X_i = 1 \mid Y_j]$$

Assume:

θ_{ij} for all ij and $P(Y)$ are known

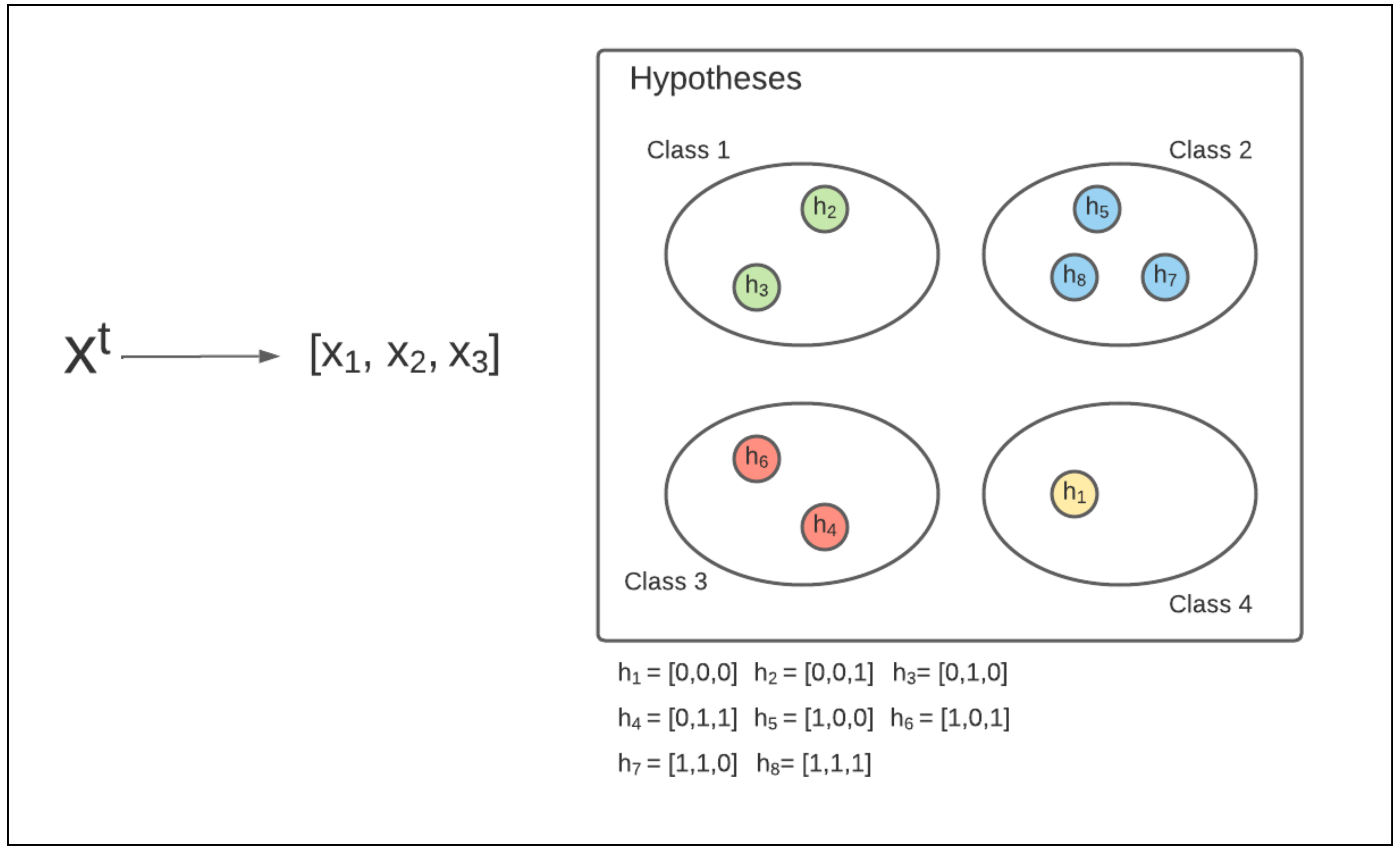
- We know the distribution of H

Decision regions are known

Task: find decision region with low cost

Instance of:

Decision Region Determination (DRD)



Decision Region Determination

Policy: mapping from current observations to features

Goal of DRD:

$$\pi^* = \arg\min_{\pi} \text{cost}(\pi)$$
$$\text{s.t. } \pi \text{ finds correct decision region}$$

NP-hard

EC^2 [1] finds near-optimal policy and gives a sequence of features to query

Online Prediction

We only have some prior knowledge (distributions):

- Use posterior sampling

Online Prediction

Online Prediction

1

Online Prediction

1. **Sample the environment and receive data point**

Online Prediction

1 **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

Online Prediction

1. **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2.

Online Prediction

1 **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2 **Sample decision regions and use EC^2 on the sampled environment**

Online Prediction

1. **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2. **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}_{\mathcal{F}}^t$

Online Prediction

1 **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2 **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3

Online Prediction

1 **Sample the environment and receive data point**

$\boldsymbol{\theta}^t \sim \text{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$

2 **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3 **Make prediction and observe true label**

Online Prediction

1 **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2 **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3 **Make prediction and observe true label**

$$y^t_j$$

Online Prediction

1 **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2 **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3 **Make prediction and observe true label**

$$y^t_j$$

4

Online Prediction

1. **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2. **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}_{\mathcal{F}}^t$

3. **Make prediction and observe true label**

$$y_j^t$$

4. **Update the knowledge**

Online Prediction

1 **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2 **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3 **Make prediction and observe true label**

$$y^t_j$$

4 **Update the knowledge**

for each $(i, x_i) \in \boldsymbol{x}^t_{\mathcal{F}}$ **do**
  if $x_i = 1$ **then** $\alpha^t_{ij} \leftarrow \alpha^{t-1}_{ij} + 1$
  else $\beta^t_{ij} \leftarrow \beta^{t-1}_{ij} + 1$

Online Prediction

1. **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \text{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2. **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3. **Make prediction and observe true label**

$$y^t_j$$

4. **Update the knowledge**

for each $(i, x_i) \in \boldsymbol{x}^t_{\mathcal{F}}$ **do**
  if $x_i = 1$ **then** $\alpha^t_{ij} \leftarrow \alpha^{t-1}_{ij} + 1$
  else $\beta^t_{ij} \leftarrow \beta^{t-1}_{ij} + 1$

Online Prediction

1. **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2. **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3. **Make prediction and observe true label**

$$y^t_j$$

4. **Update the knowledge**

for each $(i, x_i) \in \boldsymbol{x}^t_{\mathcal{F}}$ **do**
  if $x_i = 1$ **then** $\alpha^t_{ij} \leftarrow \alpha^{t-1}_{ij} + 1$
  else $\beta^t_{ij} \leftarrow \beta^{t-1}_{ij} + 1$

Online Prediction

1 **Sample the environment and receive data point**

$$\boldsymbol{\theta}^t \sim \mathrm{Beta}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1})$$

2 **Sample decision regions and use EC^2 on the sampled environment**

Observe $\boldsymbol{x}^t_{\mathcal{F}}$

3 **Make prediction and observe true label**

$$y^t_j$$

4 **Update the knowledge**

$$\begin{aligned}
&\textbf{for } \text{each } (i, x_i) \in \boldsymbol{x}^t_{\mathcal{F}} \textbf{ do}\\
&\quad \textbf{if } x_i = 1 \textbf{ then } \alpha^t_{ij} \leftarrow \alpha^{t-1}_{ij} + 1\\
&\quad \textbf{else } \beta^t_{ij} \leftarrow \beta^{t-1}_{ij} + 1
\end{aligned}$$

Extensions:

- Real-valued features
- Concept drift

Results

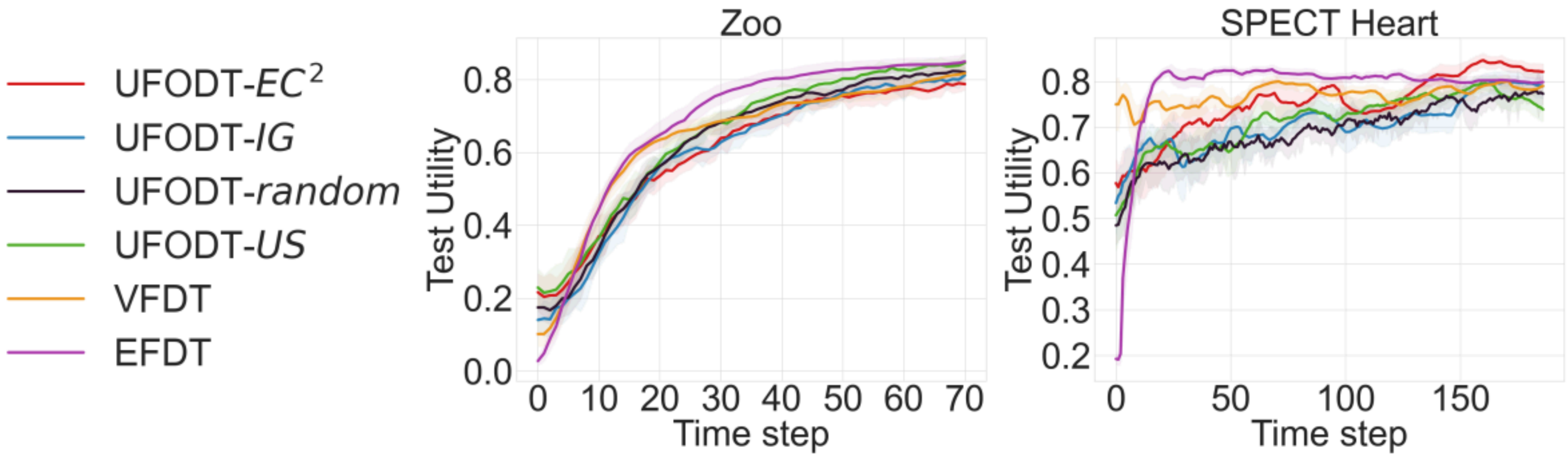
Theoretical analysis:

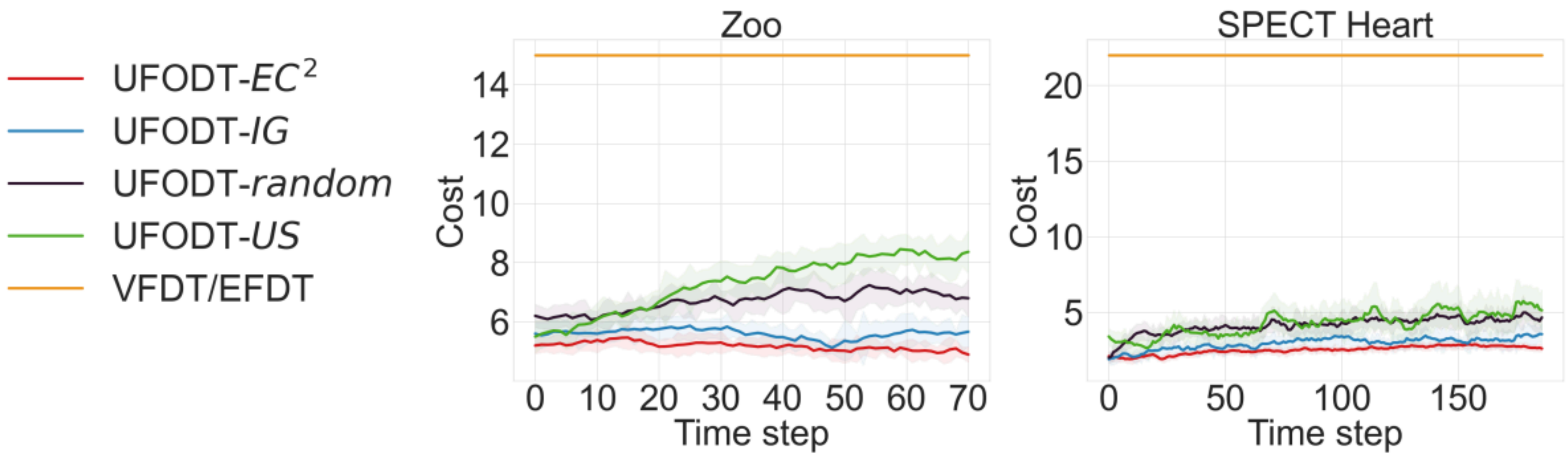
$$\Delta^t \triangleq \mathbb{U}(\pi^*_{\boldsymbol{\theta}^\star}) - \mathbb{U}(\pi^{\mathrm{EC}^2}_{\boldsymbol{\theta}^t})$$

$$\textbf{\textit{Regret}}(T) = \sum_{t=1}^{T} \Delta^t$$

Theorem: $\mathbb{E}[\mathrm{Regret}(T)] = O(LS\sqrt{nLT\log(SnLT)})$

Results


UFODT-EC²
UFODT-IG
UFODT-random
UFODT-US
VFDT
EFDT
Zoo
SPECT Heart
Test Utility
Time step
0.0
0.2
0.4
0.6
0.8
0.3
0.5
0.7
0 10 20 30 40 50 60 70
0 50 100 150


UFODT-EC²
UFODT-IG
UFODT-random
UFODT-US
VFDT/EFDT
Zoo
SPECT Heart
Cost
Time step
6
8
10
12
14
5
15
20
0 10 20 30 40 50 60 70
0 50 100 150

References

[1] Daniel Golovin, Andreas Krause, and Debajyoti Ray. 2010. Near-optimal Bayesian active learning with noisy observations. In Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1 (NIPS'10). Curran Associates Inc., Red Hook, NY, USA, 766–774.