Learning in an Echo Chamber: Online Learning with Replay Adversary

Daniil Dmitriev*, Harald Eskelund Franck†, Carolin Heinzler†, Amartya Sanyal†

MACHINE LEARNING
UNIVERSITY OF COPENHAGEN

***University of Pennsylvania**
†University of Copenhagen

Registration and travel support
for this presentation was provided by SIGACT.

Echo Chamber

Everyone seems to agree with me, so I must be right

Sara Germain

Online Learning

For $t = 1, \ldots, T$:

Learner $\mathscr{A}$ outputs a hypothesis $\hat{h}_t \in 2^{\mathcal{X}}$,

Nature produces $x_t \in \mathcal{X}$ and reveals $(x_t, f^*(x_t))$ for some $f^* \in \mathscr{H}$,

Learner suffers loss $1\{f^*(x_t) \neq \hat{h}_t(x_t)\}$

Proper ($\hat{h}_t \in \mathscr{H}$) vs. Improper ($\hat{h}_t \in 2^{\mathcal{X}}$)

Realizable ($\exists f^* \in \mathscr{H}$)

Adaptive adversary (adversary choosing $x_t \in \mathcal{X}$ every round t) vs. Stochastic adversary ($\exists \mathscr{D} : x_t \overset{i.i.d.}{\sim} \mathscr{D}$)

Online Learning with Replay Adversary

For $t = 1, \ldots, T$:

Learner $\mathscr{A}$ outputs a hypothesis $\hat{h}_t \in 2^{\mathscr{X}}$,

Nature produces $x_t \in \mathscr{X}$

Replay adversary reveals (x_t, y_t) with $y_t = \begin{cases} f^*(x_t) \text{ for some } f^* \in \mathscr{H} \\ \hat{h}_i(x_t) \text{ for some } i < t \end{cases}$

Full loss **NOT** observed by learner

True label

Replay label

Learner suffers loss $1\{y_t \neq \hat{h}_t(x_t) \text{ and } y_t = f^*(x_t)\}$

Learning thresholds conservatively

$\mathscr{H}$ class of thresholds on finite interval [0,1], $|\mathscr{H}| = N$

The Learner

The Learner predicts with the smallest threshold (conservative and consistent).

$t = 1$

$\hat{h}_1 = 0$

$t = 1$

$(x_1, y_1) = (1/4, 1)$

$\hat{h}_1(1/4) = 0 \neq y_1$ ✘

A new positive point, which cannot be replay

→Update classifier

Learning thresholds conservatively

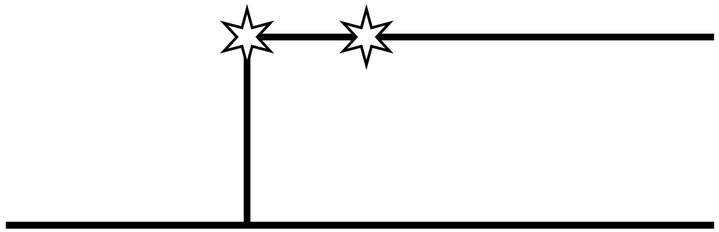
$\mathscr{H}$ class of thresholds on finite interval [0,1], $|\mathscr{H}| = N$

The Learner

The Learner predicts with the smallest threshold (conservative and consistent).

$t = 2$

$\hat{h}_2 = [1/4,1]$



$t = 2$

$(x_2, y_2) = (1/2,1)$

$\hat{h}_2(1/8) = 1 = y_2$ ✓

Correct classification, no update

Notation: a function $h : \mathscr{X} \to \{0,1\} \leftrightarrow \text{supp}(h) = h^{-1}(1)$

Learning thresholds conservatively

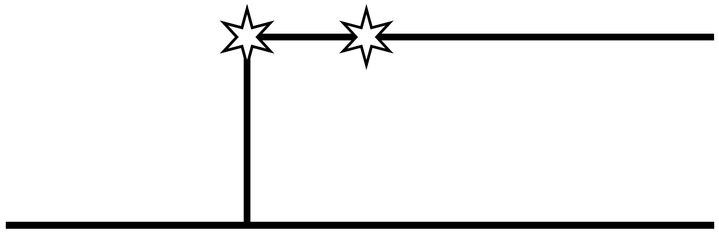
$\mathscr{H}$ class of thresholds on finite interval [0,1], $|\mathscr{H}| = N$

The Learner

The Learner predicts with the smallest threshold (conservative and consistent).

$t = 3$

$\hat{h}_3 = [1/4,1]$



In general

Easy to see:

$\mathscr{M}_T = \Theta(\min\{N, T\})$

Learning thresholds NOT conservatively

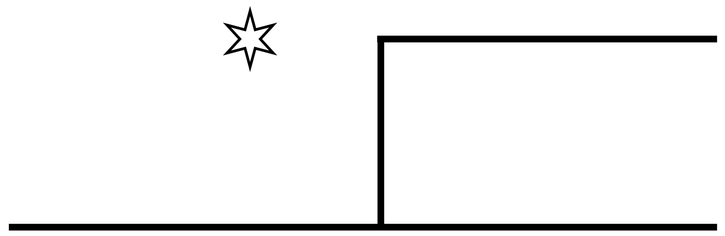
$\mathscr{H}$ class of thresholds on finite interval [0,1], $|\mathscr{H}| = N$

Other Learner

Predict with threshold left of smallest threshold (not conservative, but consistent).

$t = 1$

$\hat{h}_1 = [1/2,1]$



$t = 1$

$(x_1, y_1) = (1/4,1)$

$\hat{h}_1(1/4) = 0 \neq y_1$ ✘

A new positive point, which cannot be replay

→Update classifier

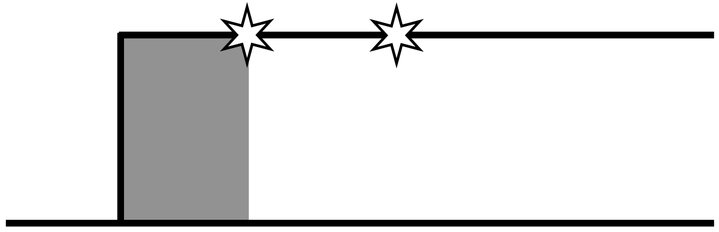
Learning thresholds NOT conservatively

$\mathscr{H}$ class of thresholds on finite interval [0,1], $|\mathscr{H}| = N$

Other Learner

Predict with threshold left of smallest threshold (not conservative, but consistent).

Trap Region
Label 0 and 1 could be true AND replay



$t = 2$

$\hat{h}_2 = [1/8,1]$

$t = 2$

$(x_2, y_2) = (1/2,1)$

$\hat{h}_2(1/2) = 1 = y_2$ ✓

Correct classification, no update

Learning thresholds NOT conservatively

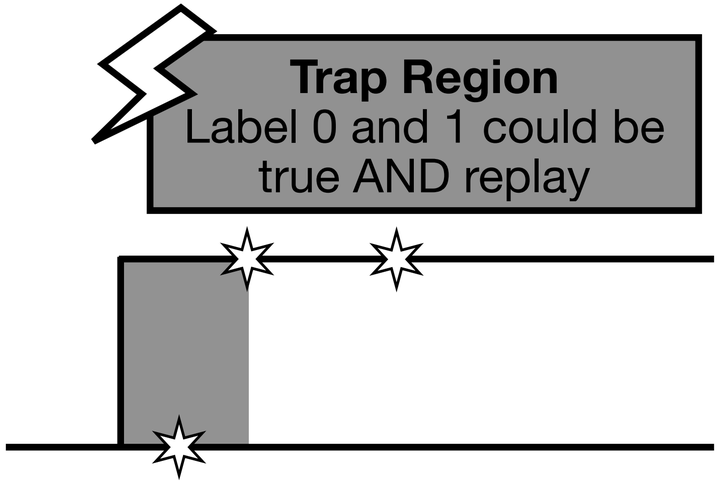
$\mathscr{H}$ class of thresholds on finite interval [0,1], $|\mathscr{H}| = N$

Other Learner

Predict with threshold left of smallest threshold (not conservative, but consistent).

$t = 2$

$\hat{h}_2 = [1/8,1]$



$t = 2$

$(x_2, y_2) = (3/16,0)$

$\hat{h}_2(3/16) = 1 \neq y_2$?

Sample in Trap Region

→ No matter what the learner predicts $\exists f^*$ s.t. the learner errs

The Learner: Closure-Based Algorithm

Definition

Define the $\mathscr{H}$**-closure** as $\operatorname{clos}_{\mathscr{H}}(Y) := \bigcap_{h \in \mathscr{H} : Y \subseteq h} h$ for any $Y \subseteq \mathscr{X}$

$\mathscr{H}$ is **intersection-closed*** if $\forall S \subseteq \mathscr{H} : \bigcap_{h \in S} h \in \mathscr{H}$.

* (over arbitrary intersections)

The Learner

The Learner predicts $\hat{h}_t = \operatorname{clos}_{\mathscr{H}}(\{x_i \mid y_i = 1, i \in \{1, \ldots t-1\}\})$

Measure of Complexity

Definition

Define Threshold dimension $\mathrm{ThD}(\mathscr{H})$ largest k, such that
$\exists x_1, \ldots, x_k \in \mathscr{X}$ and $h_0, h_1, \ldots, h_k \in \mathscr{H}$ with $h_i(x_j) = 1\{j \leq i\}$

$$\mathrm{ExThD}(\mathscr{H}) := \min_{f \subseteq \mathscr{X}} \mathrm{ThD}(\overline{\mathscr{H}^f})$$

Complexity measure for Learning with Replay Adversary

Where $\overline{\mathscr{H}} := \{\bigcap_{h \in S} h \mid S \subseteq \mathscr{H}\}$

and $\mathscr{H}^f := \left\{ h^f \mid h \in \mathscr{H} \right\}$ and $h^f := \left\{ x \in \mathscr{X} \mid h(x) \neq f(x) \right\}$

For $\mathscr{H}$ thresholds on finite intervals:
$\mathrm{ExThD}(\mathscr{H}) = C \cdot \mathrm{ThD}(\mathscr{H}) = C \cdot N$

For $\mathscr{H}$ intersection-closed:
$\mathrm{ExThD}(\mathscr{H}) = C \cdot \mathrm{ThD}(\mathscr{H})$

For $\mathscr{H}$ general:
$\mathrm{ExThD}(\mathscr{H})$

Results I

Number of mistakes: $\mathscr{M}_T(\mathscr{A}) = \sum_{t=1}^{T} 1\{y_t \neq \hat{h}_t(x_t) \text{ and } y_t = f^*(x_t)\}$

$\text{ExThD}(\mathscr{H}) = O(N)$

Hypothesis Class	**Adaptive Adv.** $\mathcal{M}_T$	**Stochastic Adv.** $\mathbb{E}[\mathcal{M}_T]$
Thresholds on $[N]$	$\Theta\big(\min\{N, T\}\big)$	$\Theta\big(\min\{N, \log T\}\big)$

Results I

Number of mistakes: $\mathscr{M}_T(\mathscr{A}) = \sum_{t=1}^{T} 1\{y_t \neq \hat{h}_t(x_t) \text{ and } y_t = f^*(x_t)\}$

$\text{ExThD}(\mathscr{H}) = O(N)$

Hypothesis Class	**Adaptive Adv.** $\mathcal{M}_T$	**Stochastic Adv.** $\mathbb{E}[\mathcal{M}_T]$
Thresholds on $[N]$	$\Theta\left(\min\{N, T\}\right)$	$\Theta\left(\min\{N, \log T\}\right)$
Intersection-Closed $\mathcal{H}$	$\Theta\left(\text{ThD}\left(\mathcal{H}\right)\right)$	$O\left(\min\{\text{ThD}\left(\mathcal{H}\right), d_{\text{VC}}(\mathcal{H}) \log T\}\right)$ $\Omega\left(\min\{\text{ThD}\left(\mathcal{H}\right), \log T\}\right)$

$\text{ExThD}(\mathscr{H}) = O(\text{ThD}(\mathscr{H}))$

Same upper bound would depend on $d_{\text{VC}}(\overline{\mathscr{H}^f})$

Gap for dependence on T

Results II

A finite $\mathscr{H}$ is **properly learnable** in the replay setting,

$\Leftrightarrow$

there exists a f-representation of the class $\mathscr{H}$ such that $\mathscr{H}^f := \{h^f \mid h \in \mathscr{H}\}$ is **intersection-closed**.

Open Questions/Extensions

- Open Question: Close gap for Stochastic Adversary
- Open Question: Existence of f-representation for $\mathscr{H}^f$ intersection-closed?
- Extension: Limit the power of the Replay Adversary (e.g. Replay only after k times)

Everyone seems to agree with me, so I must be right

Sara Germain