A Theory of Universal Learning

Raghu Arghal

Dept. of Electrical and Systems Engineering
University of Pennsylvania
rarghal@seas.upenn.edu

April 21, 2022

What is this presentation about?

A Theory of Universal Learning

Olivier Bousquet OBOUSQUET@GOOGLE.COM
Google, Brain Team
Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Toyota Technological Institute at Chicago
Shay Moran SMORAN@TECHNION.AC.IL
Technion
Ramon van Handel RVAN@MATH.PRINCETON.EDU
Princeton University
Amir Yehudayoff AMIR.YEHUDAYOFF@GMAIL.COM
Technion

Published in Symposium on Theory of Computing (STOC) '21.

Overview

Introduction and Motivation

The Basic Learning Problem

- Distribution P over labelled examples $(x, y) \in \mathcal{X} \times \{0, 1\}$
- Given n i.i.d. training samples
- Output classifier $\hat{h}_n : \mathcal{X} \to \{0, 1\}$

$$\min err(\hat{h}_n) = \min \mathbb{P}_{(x,y)\sim P}\{(x, y) : \hat{h}_n(x) \neq y\}$$

- Assume P realizable: concept class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$

$$\inf_{h \in \mathcal{H}} err(h) = 0$$

$$\inf_{\hat{h}_n} \sup_{P \in RE(\mathcal{H})} \mathbb{E}[err(\hat{h}_n)] \asymp \min\left(\frac{VC(\mathcal{H})}{n}, 1\right)$$

- $VC(\mathcal{H})$ denotes VC dimension i.e. the size of the largest set that can be shattered by $\mathcal{H}$

Classical Theory: The PAC Model Dichotomy

$$\inf_{\hat{h}_n} \sup_{P \in RE(\mathcal{H})} \mathbb{E}[err(\hat{h}_n)] \asymp \min\left(\frac{VC(\mathcal{H})}{n}, 1\right)$$

- $VC(\mathcal{H})$ denotes VC dimension i.e. the size of the largest set that can be shattered by $\mathcal{H}$

We have a dichotomy! Linear or nothing

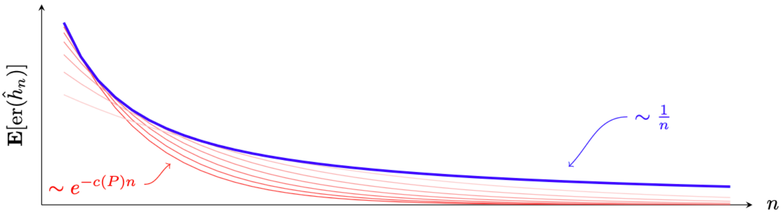


Figure: The PAC model only captures the pointwise supremum of the expected error i.e. the upper envelope of error decay

$$\inf_{\hat{h}_n} \sup_{P \in RE(\mathcal{H})} \mathbb{E}[err(\hat{h}_n)] \asymp \min\left(\frac{VC(\mathcal{H})}{n}, 1\right)$$

Minimax error convergence rate is not realistic and overly conservative!

Universal Learning Rate

A More Flexible Approach

Uniform learning rate (PAC):

$$\sup_{P \in RE(\mathcal{H})} \mathbb{E}[err(\hat{h}_n)]$$

A More Flexible Approach

Uniform learning rate (PAC):

$$\sup_{P \in RE(\mathcal{H})} \mathbb{E}[err(\hat{h}_n)]$$

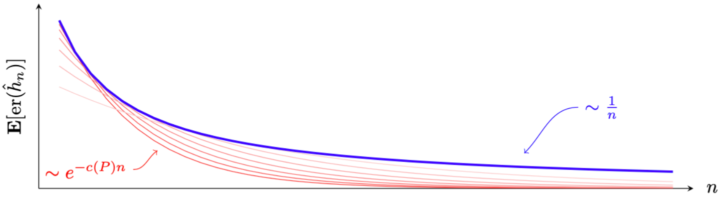
Universal learning rate:

$$\mathbb{E}[err(\hat{h}_n)] \; \forall P$$

Definition

$\mathcal{H}$ is learnable at rate R if $\exists \hat{h}_n$ s.t. $\forall P \in RE(\mathcal{H})$, $\exists c, C > 0$ s.t. $\mathbb{E}[err(\hat{h}_n)] \leq CR(cn)\ \forall n$

- c, C can depend on P $\rightarrow$ distribution-dependent learning rates



Example

Any finite class $\mathcal{H}$ is universally learnable at an exponential rate.

Example

Any finite class $\mathcal{H}$ is universally learnable at an exponential rate.

Proof.

Take $\epsilon = \min_{h \in \mathcal{H},\ err(h)>0} err(h)$ and let h^* be the target classifier.
For any $\hat{h}_n$ that fits all training data

$$P\{\hat{h}_n \neq h^*\} \leq |\mathcal{H}|(1-\epsilon)^n$$

Thus,

$$\mathbb{E}[err(\hat{h}_n)] \leq Ce^{-cn}$$

for some C, c depending on $|\mathcal{H}|, P$ □

Example

Any finite class $\mathcal{H}$ is universally learnable at an exponential rate.

Proof.

Take $\epsilon = \min_{h \in \mathcal{H},\ err(h)>0} err(h)$ and let h^* be the target classifier. For any $\hat{h}_n$ that fits all training data

$$P\{\hat{h}_n \neq h^*\} \leq |\mathcal{H}|(1-\epsilon)^n$$

Thus,

$$\mathbb{E}[err(\hat{h}_n)] \leq Ce^{-cn}$$

for some C, c depending on $|\mathcal{H}|, P$ $\square$

How much additional granularity does this provide?

A Fundamental ~~Di~~Trichotomy

Theorem

For every concept class $\mathcal{H}$ with $|\mathcal{H}| \geq 3$, exactly one of the following holds

1. *$\mathcal{H}$ is learnable with optimal rate e^{-n}.*
2. *$\mathcal{H}$ is learnable with optimal rate $\frac{1}{n}$.*
3. *$\mathcal{H}$ requires arbitrarily slow rates.*

Background

Definition

A Littlestone tree for $\mathcal{H}$ is a complete binary tree of depth $d \leq \infty$ such that each finite path emanating from the root is consistent with a concept $h \in \mathcal{H}$. We say that $\mathcal{H}$ has an infinite Littlestone tree if there is a Littlestone tree for $\mathcal{H}$ of depth $d = \infty$

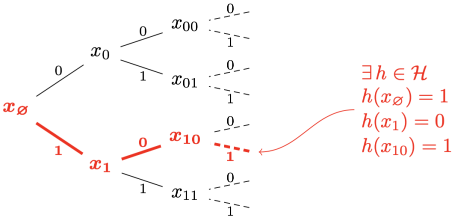


Figure: A Littlestone tree of depth 3

In online learning, finite Littlestone *dimension* yields algorithms that make finitely many errors on any adversarial sequence of points

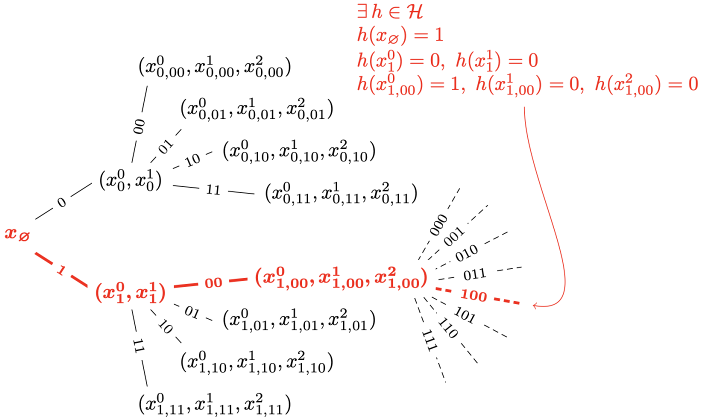


Figure: A VCL tree of depth 3

Gale-Stewart Games

- Fix sets $\mathcal{X}_t, \mathcal{Y}_t, t \geq 1$
- In each round player A (P_A) selects an element $x_t \in \mathcal{X}_t$, and then player B (P_B) selects $y_t \in \mathcal{Y}_t$
- define the winning set of P_B as $W \subseteq \prod_{t \geq 1}(\mathcal{X}_t \times \mathcal{Y}_t)$
- If $(x_1, y_1, x_2, \dots) \in W$, P_B wins; else, P_A wins
- W is called finitely decidable if for every sequence in W there is some finite n such that $(x_1, y_1, \dots, x_n, y_n, x'_{n+1}, y'_{n+1}, \dots) \in W$ for all x', y' Such a finitely decidable infinite game is called a <u>Gale-Stewart game</u>

Theorem

In a Gale-Stewart game, one of the players has a winning strategy.

Can be shown via topological argument: A's winning sequence is a closed set

Main Result

A Fundamental ~~Di~~Trichotomy

Theorem

For every concept class $\mathcal{H}$ with $|\mathcal{H}| \geq 3$, exactly one of the following holds

1. *$\mathcal{H}$ is learnable with optimal rate e^{-n}.*
2. *$\mathcal{H}$ is learnable with optimal rate $\frac{1}{n}$.*
3. *$\mathcal{H}$ requires arbitrarily slow rates.*

Theorem

For every concept class $\mathcal{H}$ with $|\mathcal{H}| \geq 3$ the following hold:

1. *If $\mathcal{H}$ does not have an infinite Littlestone tree, then $\mathcal{H}$ is learnable with optimal rate e^{-n}.*
2. *If $\mathcal{H}$ has an infinite Littlestone tree but does not have an infinite VCL tree, then $\mathcal{H}$ is learnable with optimal rate $\frac{1}{n}$.*
3. *If $\mathcal{H}$ has an infinite VCL tree, then $\mathcal{H}$ requires arbitrarily slow rates.*

Claims

1. <u>Exponential Rates</u>
   Any $\mathcal{H}$ is learnable at an exponential rate iff it has no infinite Littlestone tree. Otherwise it is learnable no faster than linear.

2. <u>Linear Rates</u>
   Any $\mathcal{H}$ is learnable at rate $\frac{1}{n}$ iff it has no infinite VCL tree. Otherwise $\mathcal{H}$ requires arbitrarily slow rates.

Claims

1. Exponential Rates
   Any $\mathcal{H}$ is learnable at an exponential rate iff it has no infinite Littlestone tree. Otherwise it is learnable no faster than linear.

2. Linear Rates
   Any $\mathcal{H}$ is learnable at rate $\frac{1}{n}$ iff it has no infinite VCL tree. Otherwise $\mathcal{H}$ requires arbitrarily slow rates.

Proof Outline

1. Construct a Gale-Stewart game
2. Translate the game into an online learning result
3. Use data-splitting and voting to obtain the rate bound

Exponential Rates Proof Outline

Consider the following game

- Player A proposes a point $x_1 \in \mathcal{X}$
- B proposes a label for that point $y_1 \in \{0, 1\}$
- Repeat ad infinitum

- B wins if at some point, there are no classifiers in $\mathcal{H}$ that can fit the entire sequence

Exponential Rates Proof Outline

Consider the following game

- Player A proposes a point $x_1 \in \mathcal{X}$
- B proposes a label for that point $y_1 \in \{0, 1\}$
- Repeat ad infinitum

- B wins if at some point, there are no classifiers in $\mathcal{H}$ that can fit the entire sequence

Recall

Theorem

In a Gale-Stewart game, one of the players has a winning strategy.

- If A has a winning strategy, we can use it to construct an infinite Littlestone tree
- Thus if there is no infinite Littlestone tree, then B has a winning strategy
- There is some finite m such that any candidate classifier can be contradicted with m points
- Express that strategy as $g_{S_m} : \{x_i, y_i\}_{i=1}^{m} \times \mathcal{X} \to \{0, 1\}$

Exponential Rates Proof Outline

Online learning setting: Observe X_i, predict $\hat{Y}_i$, observe Y_i, ...

Exponential Rates Proof Outline

Online learning setting: Observe X_i, predict $\hat{Y}_i$, observe Y_i, ... Let's

use B's winning strategy to make an online learner:

1. Initialize $m = 0, S_m = \{\}, \hat{f}_m(x) = 1 - g_{S_m}(x)$
2. For each $i = 1, 2, \ldots$
   2.1 Predict $\hat{f}_m(X_i)$
   2.2 If prediction is incorrect
      - Increment m
      - Append new pair (X_i, Y_i) to S_m
      - $\hat{f}_m(x) = 1 - g_{S_m}(x)$

Exponential Rates Proof Outline

Online learning setting: Observe X_i, predict $\hat{Y}_i$, observe Y_i, ... Let's

use B's winning strategy to make an online learner:

1. Initialize $m = 0, S_m = \{\}, \hat{f}_m(x) = 1 - g_{S_m}(x)$
2. For each $i = 1, 2, \ldots$
   - 2.1 Predict $\hat{f}_m(X_i)$
   - 2.2 If prediction is incorrect
     - Increment m
     - Append new pair (X_i, Y_i) to S_m
     - $\hat{f}_m(x) = 1 - g_{S_m}(x)$

Because B wins after finite time, this algorithm will make finitely many mistakes

Exponential Rates Proof Outline

- The online learner yields a consistent algo in the original setting
- By splitting data into batches, training multiple classifiers, and voting we can achieve exponential rate via Hoeffding's

Exponential Rates Proof Outline

- The online learner yields a consistent algo in the original setting
- By splitting data into batches, training multiple classifiers, and voting we can achieve exponential rate via Hoeffding's

<u>Any $\mathcal{H}$ is learnable at an exponential rate iff it has no infinite Littlestone tree.</u>

Example

Consider the class of threshold functions $\mathcal{H} := \{\mathbf{1}_{x \geq t}, t \in \mathbb{N}\}$. This class is learnable at an exponential rate.

Exponential Rates

Example

Consider the class of threshold functions $\mathcal{H} := \{\mathbf{1}_{x \geq t}, t \in \mathbb{N}\}$. This class is learnable at an exponential rate.

Proof.

Once the corresponding Littlestone tree branches right, it can only branch left finitely many times $\square$

Example

Consider the class of threshold functions $\mathcal{H} := \{\mathbf{1}_{x \geq t}, t \in \mathbb{N}\}$. This class is learnable at an exponential rate.

Proof.

Once the corresponding Littlestone tree branches right, it can only branch left finitely many times □

Note that this example has VC dimension 1, but VC only provides a linear learning rate

Penn
University *of* Pennsylvania

Example

Consider the class of disjoint unions of finite sets. Define $\mathcal{X} = \cup_k \mathcal{X}_k$ where $|\mathcal{X}_k| = k$. Let $\mathcal{H} = \cup_k \mathcal{H}_k$ where $\mathcal{H}_k = \{\mathbf{1}_S : S \subseteq \mathcal{X}_k\}$. This class is learnable at an exponential rate.

Example

Consider the class of disjoint unions of finite sets. Define $\mathcal{X} = \cup_k \mathcal{X}_k$ where $|\mathcal{X}_k| = k$. Let $\mathcal{H} = \cup_k \mathcal{H}_k$ where $\mathcal{H}_k = \{\mathbf{1}_S : S \subseteq \mathcal{X}_k\}$. This class is learnable at an exponential rate.

Proof.

Similar to the previous slide, once you hit a positive point and the Littlestone tree branches right, you can only branch right finitely many more times. $\square$

This class has <u>unbounded VC dimension</u> but is still learnable at an exponential rate!

Linear Rates Proof Outline

Consider the following game

- Player A proposes a point $x_1 \in \mathcal{X}$
- B proposes a label for that point $y_1 \in \{0, 1\}$
- A proposes two points
- B proposes two labels
- ...
- B wins if at some point, there are no classifiers in $\mathcal{H}$ that can fit the entire sequence

Linear Rates Proof Outline

Consider the following game

- Player A proposes a point $x_1 \in \mathcal{X}$
- B proposes a label for that point $y_1 \in \{0, 1\}$
- A proposes two points
- B proposes two labels
- ...
- B wins if at some point, there are no classifiers in $\mathcal{H}$ that can fit the entire sequence

Recall

- If A has a winning strategy, then there is an infinite VCL tree
- If no infinite VCL tree, then B has a winning strategy

Use B's winning strategy to make an online learner

1. Initialize $m = 0, S_m = \{\}$, B's winning strategy is $g_{S_m}(x_1, \ldots, x_{m+1})$
2. For each $i = 1, 2, \ldots$
   2.1 If there exists $m + 1$ points that match B's prediction
      - Increment m
      - Append $\{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_{m+1}}, Y_{i_{m+1}})\}$ to S_m

Linear Rates Proof Outline

Use B's winning strategy to make an online learner

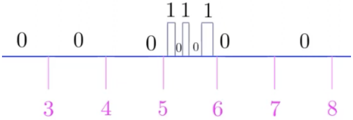
1. Initialize $m = 0, S_m = \{\}$, B's winning strategy is $g_{S_m}(x_1, \ldots, x_{m+1})$
2. For each $i = 1, 2, \ldots$
   2.1 If there exists $m + 1$ points that match B's prediction
      - Increment m
      - Append $\{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_{m+1}}, Y_{i_{m+1}})\}$ to S_m

From realizability assumption

- We know this terminates at some point
- For some m, every $m + 1$ points have a pattern that cannot be fit by the class
- Analogous to VC dim m
- Again apply data-splitting and voting to obtain our $\frac{1}{n}$ rate

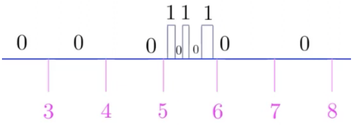
Example

Consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{x \to h(x)\mathbb{I}[x \in (i-1, i] : i \in \mathbb{N}, h \in \mathcal{H}_i\}$ where $\mathcal{H}_i$ have finite VC dimension (e.g. unions of intervals). This class is learnable at a linear rate.



Example

Consider $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{x \to h(x)\mathbb{I}[x \in (i-1, i] : i \in \mathbb{N}, h \in \mathcal{H}_i\}$ where $\mathcal{H}_i$ have finite VC dimension (e.g. unions of intervals). This class is learnable at a linear rate.



Proof.

Once the VCL tree branches once (i.e. you encounter a positive example), you are left with a class of finite VC dimension. This bounds the size of possible shattered sets and, hence, the depth of the VCL tree. □

Conclusion

Summary and Next Steps

Summary

- Reframed fundamental learning theory questions (universal/uniform rates)
- Uncovered a fundamental trichotomy of error convergence rates
- Fully characterized the classes that fit into each of three rates
- Showed intimate connections between convergence rates, combinatorial structures, and online learning

Summary

- Reframed fundamental learning theory questions (universal/uniform rates)
- Uncovered a fundamental trichotomy of error convergence rates
- Fully characterized the classes that fit into each of three rates
- Showed intimate connections between convergence rates, combinatorial structures, and online learning

Next Steps

- Extension to agnostic setting
- Extension to noisy setting
- Understanding or bounding distribution-dependent constants

Thank You!