# A THEORY OF UNIVERSAL LEARNING

Nov. 18, 2020

- Bousquet, Hanneke, Moran, van Handel, Yehudayoff (Nov 2020)

## ① Uniform learning ⇝ universal learning

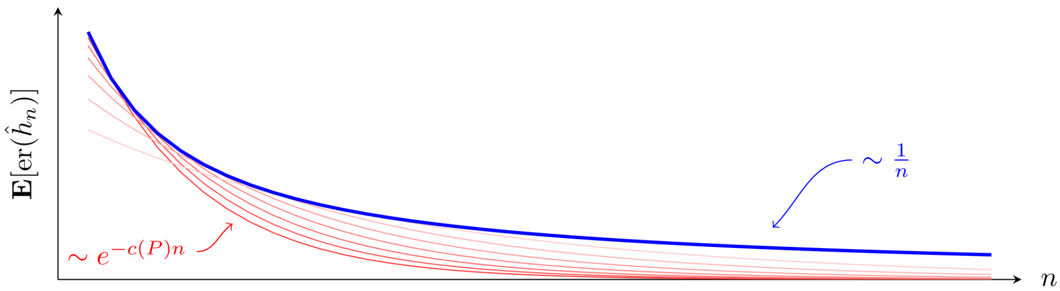
- standard PAC-learning (realizable), uniform convergence

$$\varepsilon_n = \inf_{\hat{h}} \sup_{P \in RE(\mathcal{H})} E[er_P(\hat{h}_n)]$$

- worst-case error rate characterized by VC dimension

$$\varepsilon_n \asymp \frac{VC(\mathcal{H})}{n} \quad \leftarrow \text{linear rate}$$

- observed learning curve often much better than linear




Even though the minimax risk decreases $\sim \frac{1}{n}$, for any fixed data distribution $P$, the error rate may decrease exponentially.

"One may argue that it is really the learning curve for $P$, rather than the PAC-error bound, that is observed in practice."

**QUESTION.** Is there a theory that can better explain this phenomenon?
→ move away from worst-case analysis to a data-dependent one?

- the term universal is used when a property (consistency, rate, etc.) holds for every realizable distribution $P$, but not uniformly.

Idea: define universal learnability so that constants in learning rates are allowed to depend on $P$ (but the rate may overall still be exponential, for example).

Result: there are only 3 possible rates

- exponential
- linear
- arbitrarily slow

## ② Preliminaries

- $\mathcal{X}$ — instance space / domain
- $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ — concept class
- $P$ — distribution over $\mathcal{X} \times \{0,1\}$
- $er_P(h) = P(h(X) \neq Y)$ — error rate
- $RE(\mathcal{H}) = \{ P : \inf_{h \in \mathcal{H}} er_P(h) = 0 \}$ — realizable distributions ← for our purposes, can just assume $\exists h \in \mathcal{H}$ s.t. $er_P(h) = 0$.
- $\hat{h}_n$ — classifier chosen by a learning algorithm given $n$ iid samples from $P$
  ← can be improper (ie. $\hat{h}_n \notin \mathcal{H}$ allowed)

**Definition** (learning rate). A concept class is universally learnable with rate $R : \mathbb{N} \to [0,1]$ if there exists a learning algorithm $\hat{h}_n$ s.t. for every $P \in RE(\mathcal{H})$, there exist constants $C, c > 0$ so that:

$$\mathbb{E}[er_P(\hat{h}_n)] \leq C R(cn) \qquad (\forall n \in \mathbb{N})$$

- we say that $\mathcal{H}$ is not learnable with rate faster than $R$ if $\forall \hat{h}_n$ learning algorithm, $\exists P \in RE(\mathcal{H})$ data distribution s.t. $\forall C, c > 0$

$$\mathbb{E}[er_P(\hat{h}_n)] \geq C R(cn) \quad \text{for infinitely many } n.$$

- $R$ is an optimal learning rate if $\mathcal{H}$ is learnable with rate $R$ but no faster
- $\mathcal{H}$ requires arbitrarily slow rates if for all $R(n) \downarrow 0$, $\mathcal{H}$ is not learnable faster than rate $R$.
  ↑ ie. depending on the choice of $P$, learning rate may be arbitrarily slow.

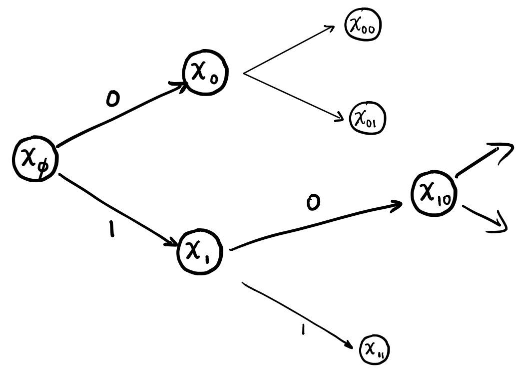
**Theorem.** For every concept class $|\mathcal{H}| \geq 3$, exactly one of the following holds:

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

# Characterization of rates

**Definition** (Littlestone tree). A Littlestone tree for $\mathcal{H}$ is a complete binary tree

- depth $d \leq \infty$




- each node is a point $x \in \mathcal{X}$
  - indexed by the path to reach it from the root

- let $y \in \{0,1\}^d$ define a path through the tree

  $x_\phi \to x_{y_1} \to x_{y_1 y_2} \to \cdots \to x_{y_{\leq d-2}} \to x_{y_{d-1}}$

  then there exists $h \in \mathcal{H}$ s.t. $h(x_{y_{\leq k}}) = y_{k+1}$ $\quad 0 \leq k < d$

  i.e. if we can reach a node $x$ through a path $(y_1, y_2, \ldots, y_k)$


  then $\exists h^{(0)}$ and $h^{(1)}$ s.t. $h^{(i)}(x_{y_{\leq i}}) = y_{i+1}$ and $h^{(0)}(x_{y_{\leq k}}) = 0$

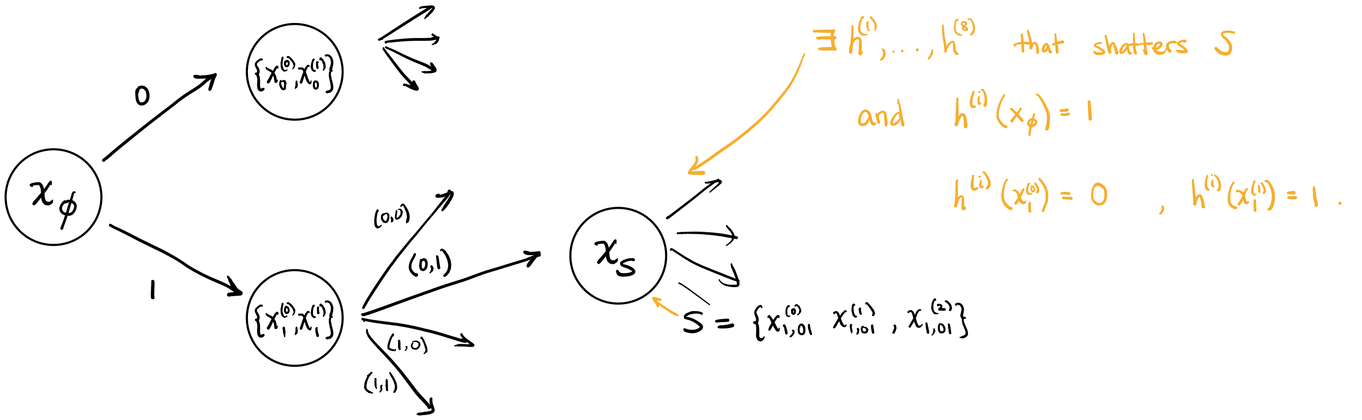
If $d = \infty$, then we say that $\mathcal{H}$ has an infinite Littlestone tree.

If $\mathcal{H}$ has a Littlestone tree with depth $d$ but not $d+1$, we say it has Littlestone dimension $d$.

**REMARK**. $\mathcal{H}$ has an infinite Littlestone tree $\implies$ $\mathcal{H}$ has infinite Littlestone dimension.
However, the converse does not hold (infinite Littlestone dimension merely implies the depth can be arbitrarily large, but not that $d = \infty$ exists).

In the Littlestone tree, we have that for any path to a node $x$, there exist $h^{(0)}$ and $h^{(1)}$ that are consistent on the path and $h^{(0)}(x) = 0$ while $h^{(1)}(x) = 1$.

- Now define a related object where nodes at the $k^{th}$ layer is $S \in \mathcal{X}^k$
  - for every path to a node, there exists $h^{(1)}, \dots, h^{(2^k)}$ consistent on the path and shatters $S$.




<u>Definition</u> (VC-Littlestone tree). A VCL($\mathcal{H}$)-tree of depth $d \le \infty$ is a collection:

$$\{x_u \in \mathcal{X}^{k+1} : 0 \le k < d,\ u \in \{0,1\}^1 \times \{0,1\}^2 \times \cdots \times \{0,1\}^k\}$$

such that for every $n < d$ and $y \in \{0,1\}^1 \times \{0,1\}^2 \times \cdots \times \{0,1\}^{n+1}$, there exists a concept $h \in \mathcal{H}$ s.t. $h(x^i_{y_{\le k}}) = y^i_{k+1}$ for all $0 \le i \le k$, $0 \le k \le n$.

We say that $\mathcal{H}$ has an infinite VCL tree if there exists one with $d = \infty$.

<u>THEOREM</u>. For every concept class $\mathcal{H}$ with $|\mathcal{H}| \ge 3$, the following holds:

- if $\mathcal{H}$ does not have an infinite Littlestone tree, then it has optimal learning rate $e^{-n}$.
- if $\mathcal{H}$ has an infinite Littlestone tree but no infinite VCL-tree, then it has optimal learning rate $\frac{1}{n}$.
- if $\mathcal{H}$ has an infinite VCL-tree, then it requires arbitrarily slow learning rate.

## EXAMPLES

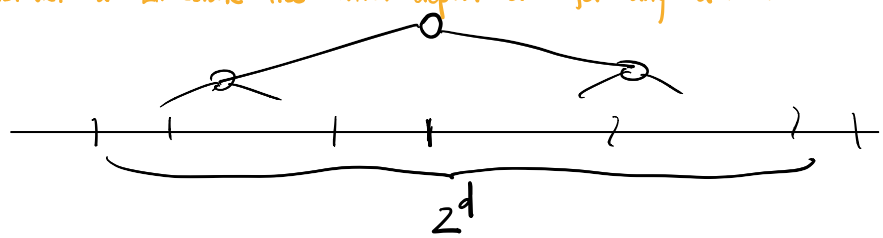
① **Threshold functions on $\mathbb{N}$.** Let $\mathcal{X} = \mathbb{N}$ and $\mathcal{H} = \{ \mathbb{1}_{\geq t} : t \in \mathbb{N} \}$.

- $VC(\mathcal{H}) = 1$, has no infinite Littlestone tree, Littlestone dimension is infinite

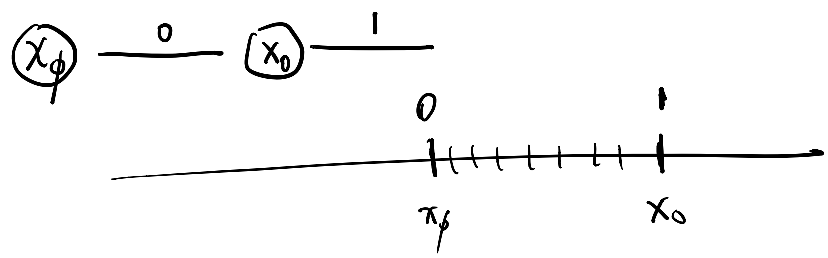
  has no infinite Littlestone tree $\Downarrow$ learnable with exponential rate

  Littlestone dimension is infinite $\Downarrow$ no online learning algorithm with uniformly bounded mistakes

Exercise. Construct a Littlestone tree with depth $d$ for any $d < \infty$.




Exercise. Prove that any Littlestone tree for $\mathcal{H}$ is finite.




② **Threshold functions on $\mathbb{R}$.** Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{ \mathbb{1}_{\geq t} : t \in \mathbb{R} \}$.

- $VC(\mathcal{H}) = 1$, has an infinite Littlestone tree

  $VC(\mathcal{H}) = 1$ $\Downarrow$ no infinite VCL-tree

  has an infinite Littlestone tree $\Downarrow$ learnable with linear rate (optimal)

③ **Nonlinear manifold.** Let $\mathcal{X}$ be a Polish space and $g : \mathcal{X} \to \mathbb{R}^d$ $(d < \infty)$ be measurable ($g$ is a coordinate function).

- Let $\mathcal{H} = \{ \mathbb{1}_{Ag = 0} : A \in \mathbb{R}^{k \times d} \}$

Then $\mathcal{H}$ has finite Littlestone dimension.

Pf. Fix a Littlestone tree and consider $x_\phi, x_1, x_{11}, x_{111}, \dots$ (relabel as $x^0, x^1, x^2, \dots$).

Let $V_j = \{ A \in \mathbb{R}^{k \times d} : Ag(x^i) = 0 \text{ for } i = 0, \dots, j \}$ (ie. kernel of $g(x^{\leq j})$).

- Clearly, $V_{j+1} \subseteq V_j$. But if $V_{j+1} = V_j$, then

$$h(x^i) = 1 \quad \forall i \leq j \quad \Rightarrow \quad h(x^{j+1}) = 1 .$$

$\rightsquigarrow$ contradiction since $\nexists h(x^{j+1}) = 0$. $\square$

# ③ Adversarial Setting & Gale-Stewart Games

Online learning problem. Consider a game between an adversary and a learner:

For $t = 1, 2, 3, \ldots$

- adversary chooses point $x_t \in \mathcal{X}$
- learner guesses label $\hat{y}_t \in \{0,1\}$
- adversary reveals label $y_t \in \{0,1\}$

Win conditions: learner wins if after some $T < \infty$, no longer makes mistakes
adversary wins if it forces the learner to make infinitely mistakes

Constraint: at any point in time $t$, $\{(x_1, y_1), \ldots, (x_t, y_t)\}$ must be realizable by some $h \in \mathcal{H}$.

Remark. In the usual online learning setting, we want to bound the total number of mistakes (uniformly). Here, we just want to know it is finite.

THEOREM. For any concept class $|\mathcal{H}|$, we have the dichotomy:

- If $\mathcal{H}$ does not have an infinite Littlestone tree, then there exists a strategy so that the learner makes at most finitely many mistakes against any adversary
- If $\mathcal{H}$ has an infinite Littlestone tree, then there is an adversary that forces the learner to make a mistake every round.

In the following, we'll move back and forth between the online learning setting and a game:

| Online Learning Setting | | Game Setting |
|---|---|---|
| - for $t = 1, 2, 3, \ldots$ | | - for $t = 1, 2, 3, \ldots$ |
| • choose $x_t$ | $\longleftrightarrow$ | • choose $\xi_t$ |
| • guess $\hat{y}_t$ | | • choose $\eta_t$ |
| • update $y_t$ | $\longleftrightarrow$ (to choose $\eta_t$) | |

## Gale-Stewart Games

Let $x_1, \dots, x_t \in \mathcal{X}$ and $y_1, \dots, y_t \in \{0,1\}$. Let $\mathcal{H}_{x_1 y_1 \dots x_t y_t}$ be the version space consistent with $(x_1,y_1), \dots, (x_t,y_t)$:

$$\mathcal{H}_{x_1,y_1,\dots,x_t,y_t} := \{h \in \mathcal{H} : h(x_i) = y_i \text{ for } 1 \le i \le t\}.$$

Then, the adversary attempting to force the learner make a mistake each round will try to keep

$$\mathcal{H}_{x_1, 1-\hat{y}_1, \dots, x_t, 1-\hat{y}_t} \neq \phi$$

as long as possible.

↳ If $\mathcal{H}_{x_1, 1-\hat{y}_1, \dots, x_t, 1-\hat{y}_t} = \phi$, then the only consistent choice for $\hat{y}_t$ is $y_t$.

**The Game:** two players, $P_A$ and $P_L$

- in each round $t$:
  - $P_A$ chooses $\xi_t \in \mathcal{X}$ and shows it to $P_L$
  - $P_L$ chooses $\eta_t \in \{0,1\}$ ← read: $\eta_t = 1 - \hat{y}_t$

**Win condition.** $P_L$ wins if $\mathcal{H}_{\xi_1, \eta_1, \dots, \xi_\tau, \eta_\tau} = \phi$ for some finite $\tau < \infty$.

$P_A$ wins if the game continues indefinitely.

**Remark.** The winning sequences for $P_L$ are finitely decidable.

$\Rightarrow$ this infinite game is a Gale-Stewart game $\Rightarrow$ either $P_A$ or $P_L$ has a winning strategy.

**Lemma.** Player $P_A$ has a winning strategy if and only if $\mathcal{H}$ has an infinite Littlestone tree.

*Proof.* ($\Rightarrow$) if $P_A$ has a winning strategy, then can construct an infinite Littlestone tree.

($\Leftarrow$) if $\mathcal{H}$ has an infinite Littlestone tree, can force $P_L$ to make a mistake each round. □

Recall above theorem:

**THEOREM.** For any concept class $\mathbb{H}$, we have the dichotomy:

① If $\mathcal{H}$ does not have an infinite Littlestone tree, then there exists a strategy so that the learner makes at most finitely many mistakes against any adversary.

② If $\mathcal{H}$ has an infinite Littlestone tree, then there is an adversary that forces the learner to make a mistake every round.

Proof. We already saw ②.

① If $\mathcal{H}$ has no infinite Littlestone tree, then $P_L$ has a winning strategy.

- if we knew a priori that the adversary would force a mistake each round

  $\Rightarrow$ the learner should choose $\hat{y}_t = 1 - \eta_t$.

  ↳ winning strategy $\Rightarrow$ can only make finitely many mistakes

- we don't know that adversary will force mistake each round

  ↳ update the game only if in the online learning setting, the learner made a mistake. Only finitely many updates possible.

ALGORITHM (adversarial)

- initialize $\tau \leftarrow 0$
- for $t = 1, 2, \dots$
  - receive $x_t$
  - guess $\hat{y}_t \leftarrow 1 - \eta_{\tau+1}(\xi_1, \dots, \xi_\tau, x_t)$
  - if $y_t \neq \hat{y}_t$:
    - $\xi_{\tau+1} \leftarrow x_t$
    - $\tau \leftarrow \tau + 1$

let this correspond to the winning strategy for $P_L$.

□

## A note on measurability.

The algorithm given in the theorem lets us pass from adversarial settings to probabilistic ones:

- randomly choose data $(X_1, Y_1), \dots, (X_t, Y_t) \sim P$ to simulate the game

  ↳ we obtain some classifier $\hat{h}_t = H(X_1, Y_1, \dots, X_t, Y_t)$ via the Gale-Stewart strategy

$$H : \Omega^t \longrightarrow \mathcal{H}$$

Question: do we know that $H$ is measurable?

↳ if it is not, we can't even make sense of the statement

$$\Pr[\operatorname{er}(\hat{h}_t) < \varepsilon].$$

see Appendix in paper for example of this happening.

- it turns out that if $\mathcal{H}$ has any reasonable parametrization, then we're ok.

## ④ Exponential Rates

Assumptions: $\mathcal{X}$ is Polish, $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ satisfies certain measurability conditions, $|\mathcal{H}| > 2$.
Learner is presented $(X_1, Y_1), (X_2, Y_2), \dots \sim P$ iid.

**Theorem.** If $\mathcal{H}$ does not have an infinite Littlestone tree, then $\mathcal{H}$ is learnable with optimal rate $e^{-n}$.

Intuition:

- $\mathcal{H}$ does not have an infinite Littlestone tree
  ↳ let $P \in RE(\mathcal{H})$ ⟶ $(X_1, Y_1), \dots \sim P$ input into adversarial algorithm
  will eventually stop making mistakes a.s.
- Let $T$ be the random variable corresponding to the time of the last mistake
  ↳ some distribution on $\mathbb{N}$, since $T < \infty$ a.s.

$\Rightarrow \exists t^*$ s.t. $P(T \leq t^*) = \Omega(1)$.
so, we might expect a learning rate $\sim \exp(-n/t^*)$ is possible.

**Upper bound:** constructing an algorithm

ⓐ Adversarial algorithm is consistent in probabalistic setting

- let $\hat{y}_{t+1}$ be the classifier $\hat{y}_{t+1}(x) := \hat{y}_{t+1}(X_1, Y_1, \dots, X_t, Y_t, x)$ obtained from the adversarial algorithm with input $(X_1, Y_1), \dots, (X_n, Y_t) \overset{iid}{\sim} P$

**Lemma.** If $\mathcal{H}$ has no infinite Littlestone tree, then $\Pr[er(\hat{y}_t) > 0] \to 0$ as $t \to \infty$.

*Proof.* Intuition: online learning algorithm will eventually make no more mistakes ⤳ LLN $\Rightarrow$ $er(\hat{h}_t) \to 0$ a.s.

- Claim: if $P \in RE(\mathcal{H})$, then for all $t$, $X_1, Y_1, \dots, X_t, Y_t \sim P$ is a valid input a.s. to the online learning problem

  obvious if $P \in RE(\mathcal{H}) \Rightarrow \exists h \in \mathcal{H}$ s.t. $h(X) = Y$ a.s. for $(X,Y) \sim P$

  - only slightly more involved if we take $P \in RE(\mathcal{H}) \Rightarrow \inf_{h \in \mathcal{H}} P(h(X) \neq Y) = 0$.

- Because the learner has a winning strategy
  $\Rightarrow$ the time of the last mistake $T < \infty$ a.s.

- If $t > T$, then by the Law of Large Numbers

$$\Pr(er(\hat{y}_t) = 0) = \Pr\left( \lim_{S \to \infty} \frac{1}{S} \sum_{s=t+1}^{t+S} \mathbb{1}_{\{\hat{y}_t(X_s) \neq Y_s\}} = 0 \right)$$

$P(T \leq t) \uparrow 1$

$$\geq \Pr\left( \lim_{S \to \infty} \frac{1}{S} \sum_{s=t+1}^{t+S} \mathbb{1}_{\{\hat{y}_t(X_s) \neq Y_s\}} = 0, \ T \leq t \right) = P(T \leq t). \quad \square$$

## (b) Finite Sample Bounds

**Idea.** If we knew $t^*$ s.t. $P(er(\hat{y}_{t^*}) > 0) = O(1)$

↳ then construct $\hat{h}_{nt^*}$ by taking a majority vote over $n$ independent copies of $\hat{y}_{t^*}$

Hoeffding's $\Rightarrow$ exponential rate in $n$.

**Problem.** We don't know $t^*$, since it depends on $P$. But, we can estimate using data!

**Notation.** For a fixed $P \in RE(H)$, define $t^*$ so that $P(er(\hat{y}_{t^*}) > 0) \leq \frac{1}{8}$

previous lemma says this exists

Let $T_{good} = \{ t \in [t^*] : P(er(\hat{y}_t) > 0) \leq \frac{3}{8} \}$

**Goal.** Show that we can find an estimate $\hat{t}$ that falls into $T_{good}$ w.h.p.

↳ take majority vote over $n$ copies of $\hat{y}_{\hat{t}}$ instead of $\hat{y}_{t^*}$

**Lemma.** There exists a universally measurable $\hat{t}_n = \hat{t}_n(X_1, Y_1, \ldots, X_n, Y_n)$ whose definition doesn't depend on $P$ s.t. there exists $C, c > 0$ (dependent on $P, t^*$) such that:

$$P(\hat{t}_n \in T_{good}) \geq 1 - Ce^{-cn}.$$

*Proof sketch.*

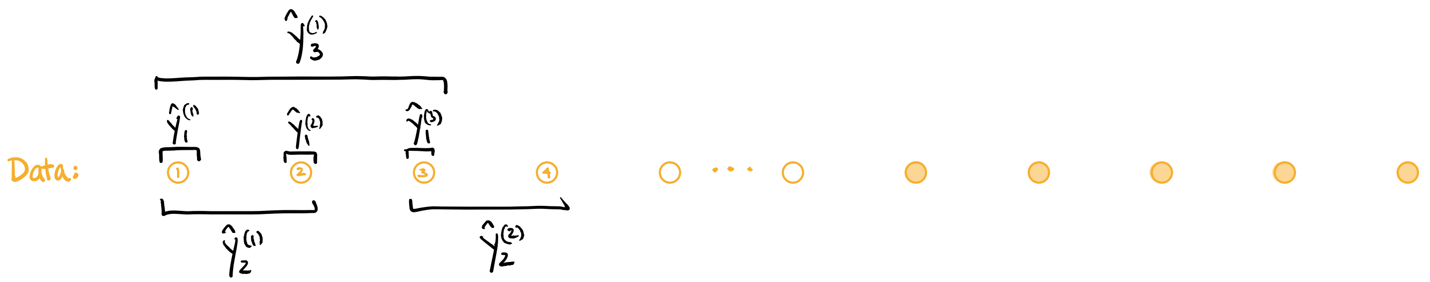
Algorithm to compute $\hat{t}_n$

- split data $(X_1, Y_1), \ldots, (X_n, Y_n)$ into two halves
  - $(X_1, Y_1), \ldots, (X_{\lfloor \frac{n}{2} \rfloor}, Y_{\lfloor \frac{n}{2} \rfloor})$ to estimate $\hat{y}_k$'s
  - $(X_{\lfloor \frac{n}{2} \rfloor + 1}, Y_{\lfloor \frac{n}{2} \rfloor + 1}), \ldots, (X_n, Y_n)$ as hold out data
- for $t = 1, \ldots, \lfloor \frac{n}{2} \rfloor$
  - compute as many independent copies of $\hat{y}_t^{(i)}$ as possible ($\lfloor n/2t \rfloor$ copies)
  - use hold out data to determine whether $\hat{y}_t^{(i)}$ makes any mistakes
  - let $\hat{e}_t$ be the fraction of $\hat{y}_t^{(1)}, \ldots, \hat{y}_t^{(\lfloor n/2t \rfloor)}$ that makes a mistake

    $\hat{e}_t$ estimates $e_t$ = fraction of $\hat{y}_t^{(i)}$ that has $er(\hat{y}_t) > 0$.
- return $\hat{t}_n := \inf \{ t \leq \lfloor \frac{n}{2} \rfloor : \hat{e}_t < \frac{1}{4} \}$

smallest $t$ s.t. less than 1/4 of trials make no mistake

## Illustration of algorithm.




↳ for each $t = 1, \dots, \lfloor \frac{n}{2} \rfloor$, use estimates $\hat{y}_t^{(i)}$ to estimate fraction of $\hat{y}_t$ s.t. $er(\hat{y}_t) > 0$.

## Analysis sketch.

① With high probability (i.e. exponential in $n$), $\hat{t}_n \leq t^*$

⇝ we will need to estimate $e_t$ no more than $t^*$ times (and $t^* = O(1)$).

② With high probability, if $t$ is too small so that a large fraction of $\hat{y}_t$ have a positive error rate, then a large fraction of $\hat{y}_t^{(i)}$ will have error rates bounded away from 0, $er(\hat{y}_t^{(i)}) > \varepsilon > 0$.

⇝ we will be able to detect these $\hat{y}_t^{(i)}$ with $\varepsilon$-error using hold out data whp.

In short, Hoeffding's + union bounds. (Details: Lemma 4.4). □

**Corollary.** $\mathcal{H}$ has at most exponential learning rate.

*Proof.* Let $\hat{h}_n$ be the majority vote of the classifiers $\hat{y}^{(i)}_{t_n}$.

- WTS: majority of classifiers never make mistakes
  - Previous lemma $\Rightarrow$ $P(er(\hat{y}^{(i)}_{t_n}) > 0) \leq \frac{3}{8}$ w.h.p. (exponential in $n$)
  - Hoeffding's $\Rightarrow$ majority won't err w.h.p. (exponential in $n/t^*$)

    gap
  - Union bound

$$\Rightarrow \quad P(er(\hat{h}_n) > 0) \leq Ce^{-cn}.$$

$$E[er(\hat{h}_n)] \leq P(er(\hat{h}_n) > 0)$$ implies the result. □

**Lower bound:**

**Lemma** (lower bound). For any learning algorithm $\hat{h}_n$, there exists a realizable distribution $P$ such that $E[er(\hat{h}_n)] \geq 2^{-n-2}$ for infinitely many $n$.

Intuition: there exists $x, x' \in \mathcal{X}$ and $h_1, h_2 \in \mathcal{H}$ s.t. $\underbrace{h_1(x) = h_2(x)}_{\text{agree on } x}$ but $\underbrace{h_1(x') \neq h_2(x')}_{\text{disagree on } x'}$.

- consider $P$ where ½ of mass on $x$ and $x'$ each
  - ↳ there is an exponentially small probability that we haven't drawn $x'$
  - $\Rightarrow$ expected error may be lower bounded by a related exponential.

apply probabilistic method

□

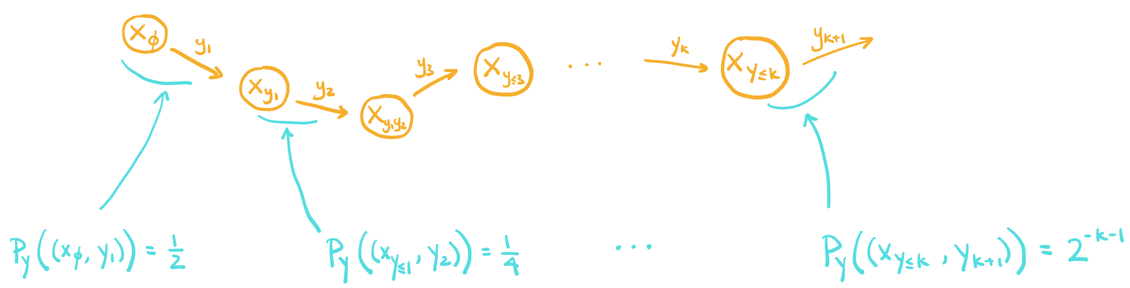
**Theorem.** If $\mathcal{H}$ has infinite Littlestone tree, then for any learning algorithm $\hat{h}_n$, there exists $P \in RE(\mathcal{H})$ s.t. $E[er(\hat{h}_n)] \geq \frac{1}{32n}$ for infinitely many $n$.

Intuition: if we are given an infinite Littlestone tree, we can construct a distribution over $RE(\mathcal{H})$ so that $\underset{P}{E}\left[\limsup_{n\to\infty} n\, er_P(\hat{h}_n)\right] = \Omega(1)$. This implies that there exists some $P \in RE(\mathcal{H})$ s.t. $er_P(\hat{h}_n) = \Omega(\frac{1}{n})$ for infinitely many $n$.

another application of probabilistic method

Proof. Given an infinite Littlestone tree, we can choose a random infinite path by a sequence $y = (y_1, y_2, \dots) \in \{0,1\}^{\mathbb{N}}$ where $y_i \overset{iid}{\sim} Ber(\frac{1}{2})$.

- Define $P_y$ on $\mathcal{X} \times \{0,1\}$ by selecting $(x_{y_{\leq k}}, y_{k+1})$ with probability $2^{-k-1}$:




i.e. the depth of our sample in the tree follows a Geometric($\frac{1}{2}$) distribution.

- Notice that if we draw $n$ iid samples from $P_y$, the max depth is $\lg(n)$ w.p. $\frac{1}{2}$

Pf. Let $(x_{y_{\leq t_1}}, y_{t_1+1}), \dots, (x_{y_{\leq t_n}}, y_{t_n+1})$ be the draws. Then:

$$P_y(\max\{t_1, \dots, t_n\} < k) = (1 - 2^{-k})^n$$

$\Rightarrow$ when $\underline{k_n} = \lceil 1 + \lg(n) \rceil$ then $(1 - 2^{-k_n})^n \geq \frac{1}{2}$.

mass decreases linearly

- The probability mass of $(x_{y_{\leq k_n}}, y_{k_n+1})$ is $2^{-k_n-2} \geq \frac{1}{8n}$.

- $\hat{h}_n \perp\!\!\!\perp y_{k_n+1} \mid \max\{t_1, \dots, t_n\} < k_n = \lg(n)$ ← $\hat{h}_n$ will err half of the time

the classifier will be conditionally independent with the label for $x_{y_{\leq k_n}}$ provided that all training samples came from depth at most $k_n$.

$$\Rightarrow P(\hat{h}_n(X) \neq Y \mid \max\{t_1, \dots, t_n\} < k_n) \geq \frac{1}{2} \cdot \frac{1}{8n} = \frac{1}{16n}.$$

$\frac{1}{2}$: Pr (mistake on $X \mid X = x_{y_{\leq k_n}}$)  $\frac{1}{8n}$: $Pr(X = x_{y_{\leq k_n}})$

$$\Rightarrow n \cdot P(\hat{h}_n(X) \neq Y \text{ and } \max\{t_1, \dots, t_n\} < k_n) \geq \frac{1}{32}$$

$$\Rightarrow n \cdot \mathop{E}_{Y}\Big[\underbrace{P(\hat{h}_n(X) \neq Y \mid y)}_{er_{P_y}(\hat{h}_n)}\Big] \geq \frac{1}{32}$$

$$\overset{\text{Fatou's}}{\Rightarrow} \quad \mathop{E}_{Y}\Big[\limsup_{n\to\infty} er_{P_Y}(\hat{h}_n)\Big] \geq \limsup_{n\to\infty} n \cdot \mathop{E}_{Y}\big[er_{P_Y}(\hat{h}_n)\big] \geq \frac{1}{32}.$$

$\square$

<u>Summary of Exponential Rates</u>. The following are equivalent:

① $\mathcal{H}$ is learnable at exponential rates but not faster.

② $\mathcal{H}$ has no infinite Littlestone tree.

③ There exists a learning algorithm that is eventually correct.

④ There exists a learning algorithm that is eventually correct with exponential rates.