When is Inductive Inference Possible?

A new link between philosophy and learning theory

Zhou Lu

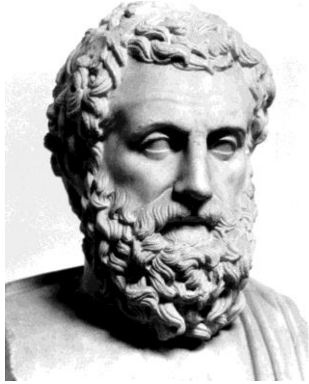
Outline

- **The problem of induction**
- Background
- Non-uniform online learning
- Characterizing inductive inference
- The agnostic setting
- Conclusion

Inductive vs deductive reasoning

Induction: Observations $\xrightarrow{?}$ Laws (axioms)

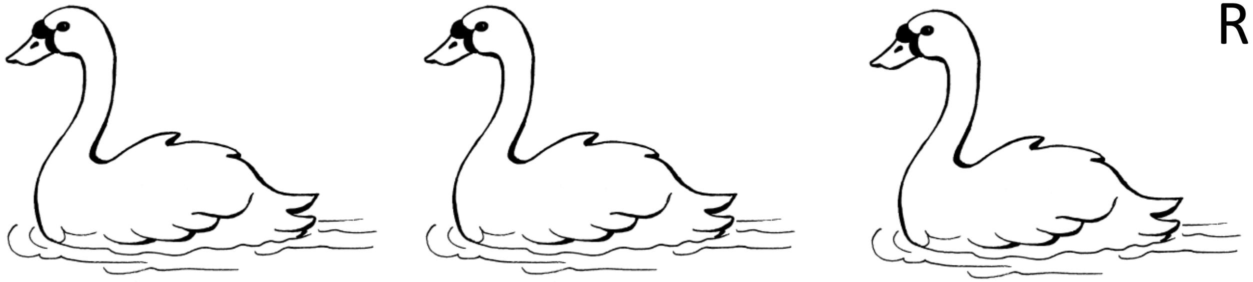
Deduction: Laws (axioms) $\xrightarrow{\text{logically}}$ Theorems



"Induction is the starting-point which knowledge even of the universal presupposes, while syllogism proceeds from the universals."

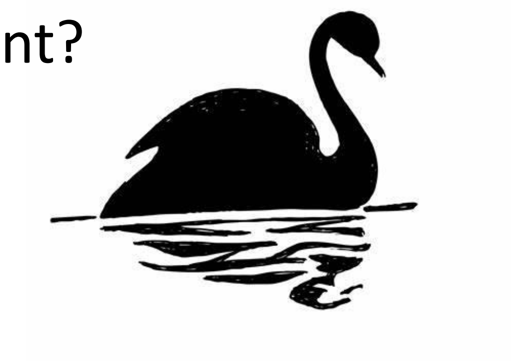
--- Nicomachean Ethics

The problem of induction



Rare event?

…

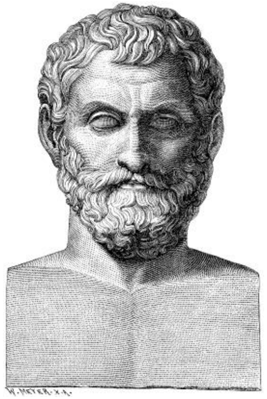


Without assumptions, there is no rigorous induction on unseen events.

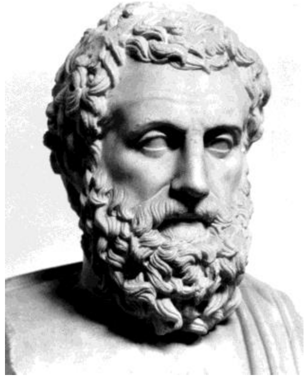
“There can be no demonstrative arguments to prove, that those instances, of which we have had no experience, resemble those, of which we have had experience.”

--- A Treatise of Human Nature

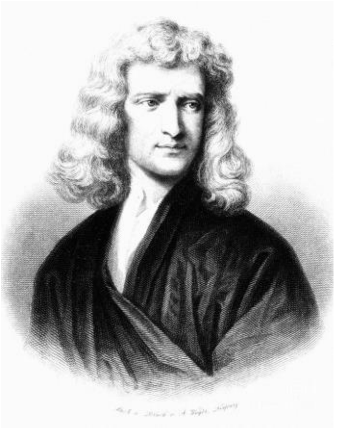
Example: evolution of physical models



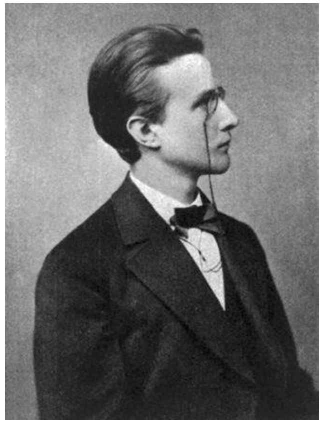
Single basic element



Four basic elements



Classical mechanics



Quantum mechanics

Legit until proven wrong!

Can we identify the true physical model? ✗

Let’s consider a weaker criterion.

Can we make few errors in pursuing it? ✓

With **assumptions** on the candidate set of models.

Outline

- The problem of induction
- **Background**
- Non-uniform online learning
- Characterizing inductive inference
- The agnostic setting
- Conclusion

Inductive Inference

- Given domain X, hypothesis class H containing binary functions.
- Nature selects the ground-truth h^*.
- For t in $1,2,\cdots$
    - Nature presents x_t to Learner.
    - Learner predicts y_t.
    - Nature reveals the true label $h^*(x_t)$.

A sufficient condition

It was proven in the 60s, that when the size of H is countable, we can guarantee only a finite number of errors is made.

Example: Solomonoff Induction

Hypothesis class: the set of all Turing machines. (countable size)

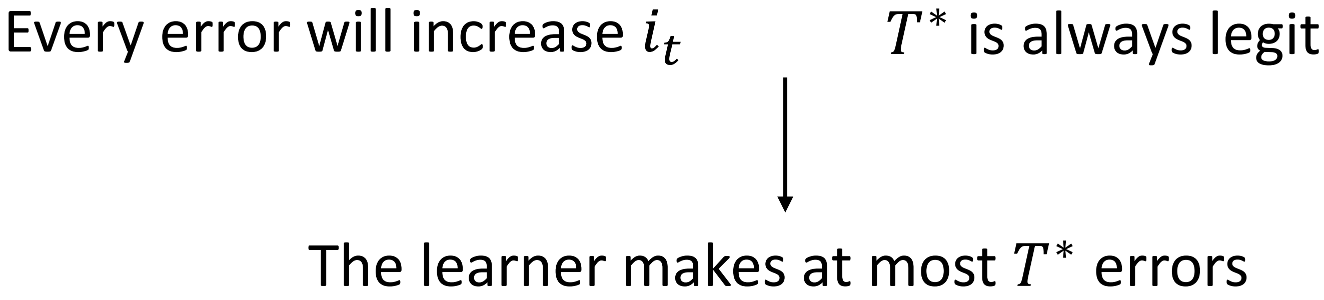
Assign larger prior weights to Turing machines with shorter description lengths.

Basic idea: Bayesian

At time t, the learner find the hypothesis h_{i_t}.

h_{i_t} has the **smallest** index i_t, among hypotheses consistent with history.

The learner then predicts with h_{i_t}.



Every error will increase i_t

T^* is always legit

The learner makes at most T^* errors

It’s not necessary!

Consider $H = \{f_c | f_c(x) = 1_{x=c}, c \in R\}$.

This class has an uncountable size.

However, we make at most one error, by keep predicting zero.

If we make an error, we immediately identify the ground-truth.

Question: what’s a sufficient and necessary condition?

Theorem (informal) Inductive inference is possible, iff H is a countable union of online learnable classes.

We obtain this result by a new link to online learning.

Online learning (Littlestone, 1988)

- Given domain X, and hypothesis class H.
- For t in $1,2,\cdots$
  - Nature presents x_t to Learner.
  - Learner predicts y_t.
  - Nature selects h_t^* consistent with history: $\forall i \leq t, h_t^*(x_i) = h_i^*(x_i)$.
  - Nature reveals the true label $h_t^*(x_t)$.

A class H is online learnable, if there exists a **uniform** constant m and an algorithm, such that for any Nature’s choice, the algorithm makes at most m errors.

We denote the min-max number of errors as the Littlestone dimension.

It’s similar to VC dimension, but only requires shattering every branch of a tree.

Example: rational thresholds

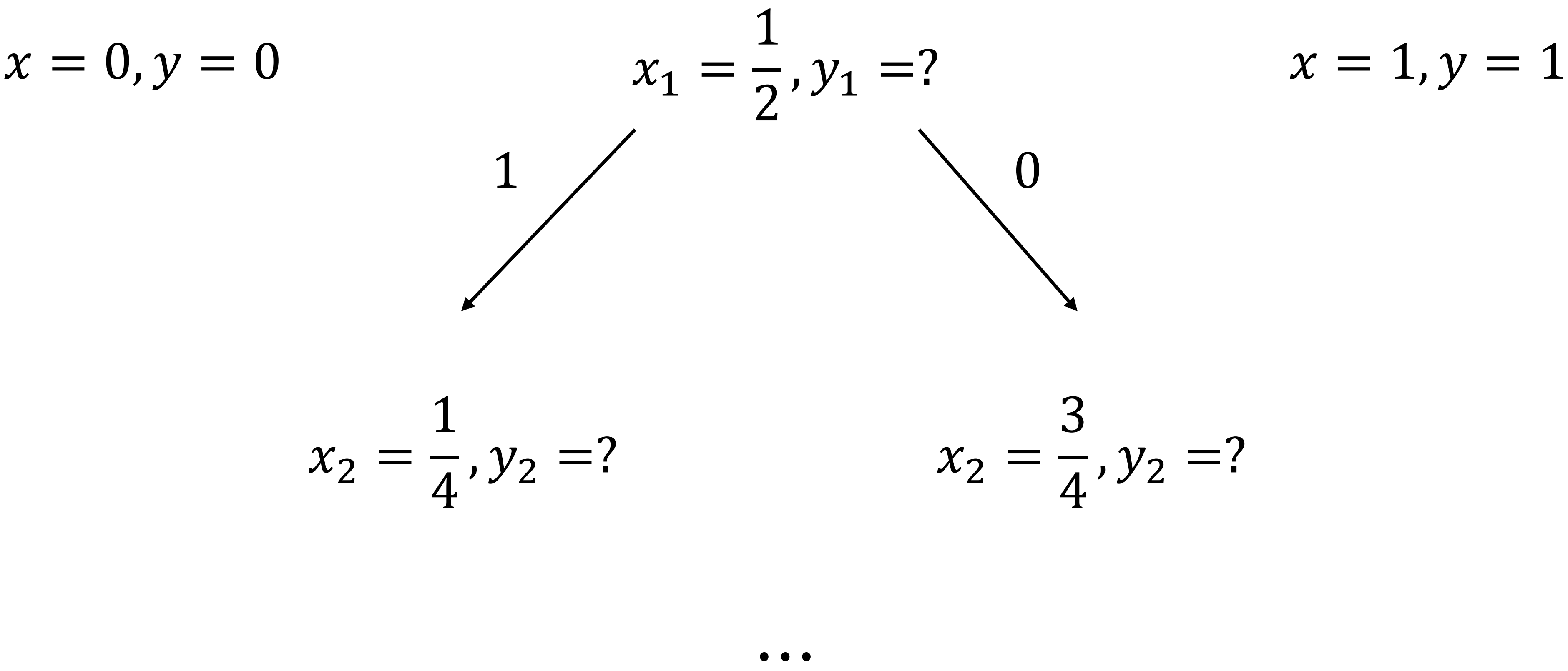
Consider $X = R$, and H being the set of all rational threshold functions:

$$H = \{h : h(x) = 1_{x \geq c}, c \in Q\}$$

It has a countable size, thus learnable in inductive inference.

However, it has an infinite Littlestone dimension, thus not online learnable.

Nature can enforce an error in each round in online learning.



$x = 0, y = 0$

$x_1 = \frac{1}{2}, y_1 = ?$

$x = 1, y = 1$

1

0

$x_2 = \frac{1}{4}, y_2 = ?$

$x_2 = \frac{3}{4}, y_2 = ?$

$\cdots$

Each round, Nature can choose the opposite label.

Outline

- The problem of induction
- Background
- **Non-uniform online learning**
- Characterizing inductive inference
- The agnostic setting
- Conclusion

Non-uniform online learning

It’s a variant of classic online learning. Two changes:

- It requires Nature to fix h^* in advance.
- It allows **non-uniform** error bounds depending on h^*.

It is inductive inference without the size constraint on H.

It’s a generalization of non-uniform PAC learning to the online setting.

Inductive Inference = Non-uniform Online Learning

- Given domain X, hypothesis class H containing binary functions.
- Nature selects the ground-truth h^*.
- For t in $1,2,\cdots$
  - Nature presents x_t to Learner.
  - Learner predicts y_t.
  - Nature reveals the true label $h^*(x_t)$.

Non-uniform online learnability

Definition A class H is non-uniform online learnable, if there exists an algorithm A,

$$\exists m: H \to N, \forall h \in H, \forall x \in X^{\infty}, err\_A(x, h) \leq m(h).$$

If H has a countable size, it’s reduced to classic inductive inference.

Notice here the error bound can vary with the ground-truth h^*.

Non-uniform stochastic online learning

In non-uniform online learning, Nature can adaptively choose x_t.

An easier yet natural setting: Nature can only choose a distribution, each x_t is iid.

Definition We say a class H is non-uniform stochastic online learnable, if there exists an algorithm A, such that

$$\exists m: H \to N, \forall h \in H, \forall \mu, P_{x \sim \mu^\infty}\left(err_A(x,h) \le m(h)\right) = 1.$$

Outline

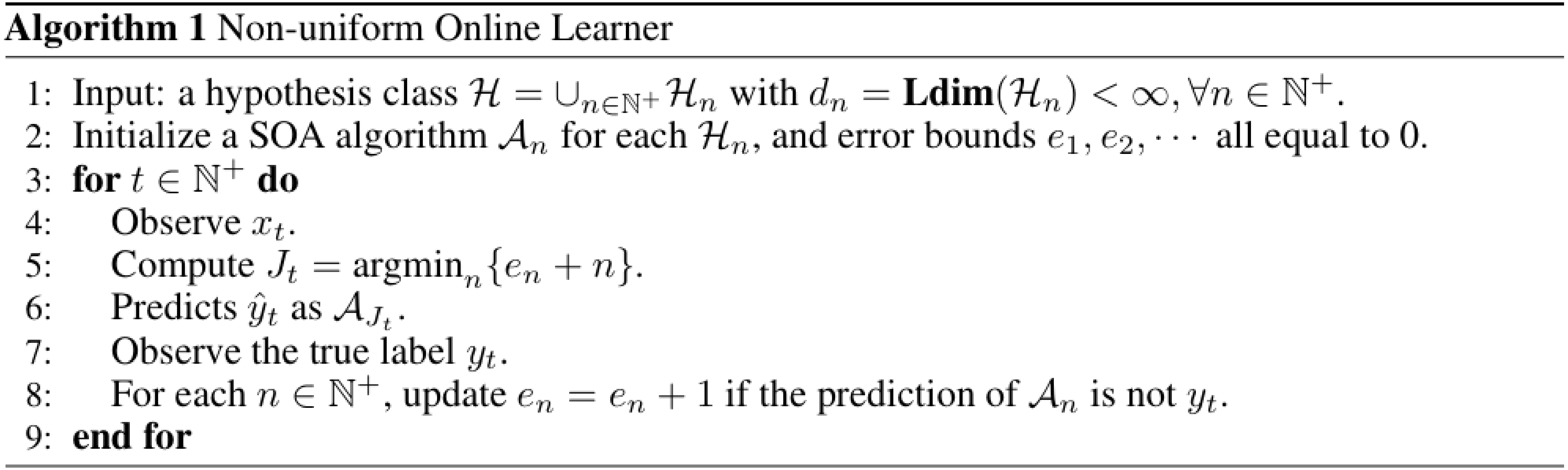
- The problem of induction
- Background
- Non-uniform online learning
- **Characterizing inductive inference**
- The agnostic setting
- Conclusion

A tight characterization

Theorem (main result) H is non-uniform online learnable, if and only if it's a countable union of Littlestone classes.

The proof is simple and standard. It leverages the structural risk minimization technique from Vapnik and Chervonenkis 74.

The learning algorithm



Algorithm 1 Non-uniform Online Learner

```
Input: a hypothesis class $\mathcal{H} = \cup_{n \in \mathbb{N}^+} \mathcal{H}_n$ with $d_n = \mathbf{Ldim}(\mathcal{H}_n) < \infty, \forall n \in \mathbb{N}^+$.
Initialize a SOA algorithm $\mathcal{A}_n$ for each $\mathcal{H}_n$, and error bounds $e_1, e_2, \cdots$ all equal to 0.
for $t \in \mathbb{N}^+$ do
    Observe $x_t$.
    Compute $J_t = \text{argmin}_n \{e_n + n\}$.
    Predicts $\hat{y}_t$ as $\mathcal{A}_{J_t}$.
    Observe the true label $y_t$.
    For each $n \in \mathbb{N}^+$, update $e_n = e_n + 1$ if the prediction of $\mathcal{A}_n$ is not $y_t$.
end for
```

Index: prior

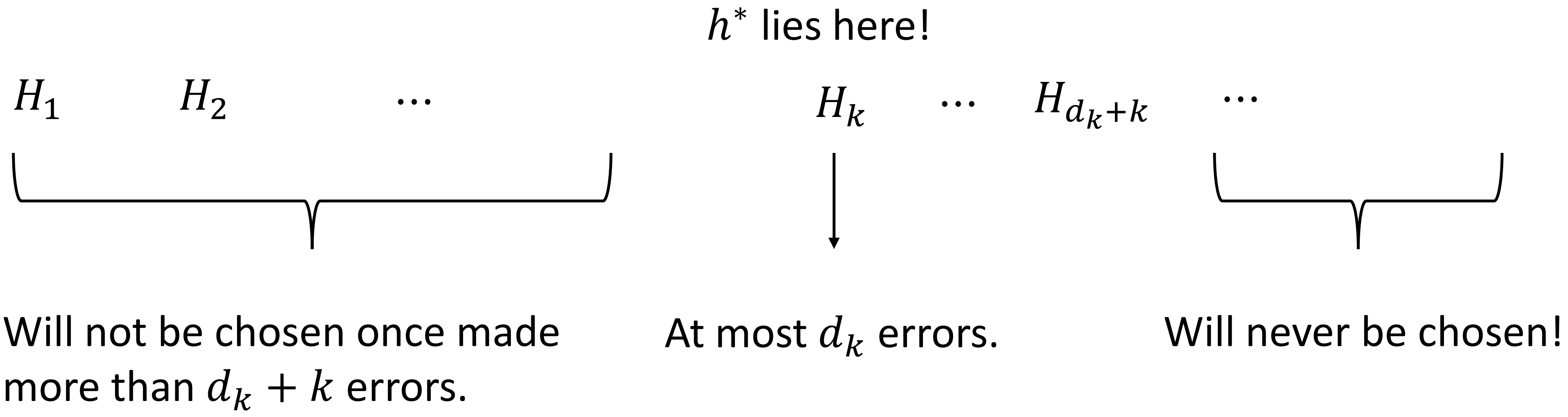
Error bound: posterior

J_t is the index minimizing error bound plus index.

Proof sketch

The If direction:

- A uniform error bound is achievable iff H has finite Littlestone dimension.
- A_n learns class H_n with at most d_n errors.
- Suppose h^* lies in H_k, then $J_t \leq e_k + k \leq d_k + k$ for any t.
- Notice that A_n will never be chosen once $e_n + n > d_k + k$.
- As a result, only $A_1, A_2, \cdots, A_{d_k+k}$ can be invoked.
- Then A makes at most $(d_k + k)^2$ errors.



h^* lies here!

H_1 H_2 $\cdots$ H_k $\cdots$ H_{d_k+k} $\cdots$

Will not be chosen once made more than $d_k + k$ errors.

At most d_k errors.

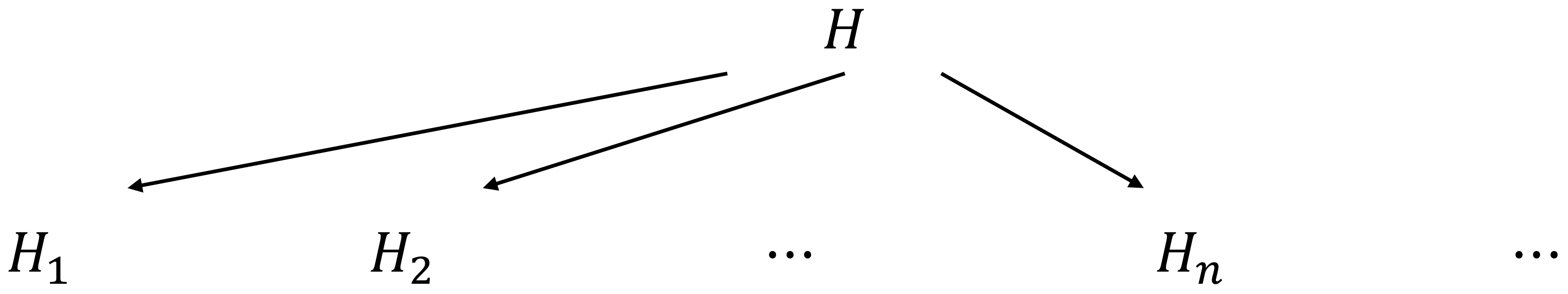
Will never be chosen!

Two cases: no error, good! If error, error bound increases.

e_k is bounded by d_k, error bound can't increase too many times.

The Only If direction:

- For any h, denote $d(h)$ as the finite error bound admitted.
- We denote $H_n = \{h: d(h) = n\}$.
- We have that $H = \cup_n H_n$.
- We only need to prove each H_n is learnable with a uniform error bound.
- The same algorithm for H has an error bound n on H_n.



H

H_1 H_2 $\cdots$ H_n $\cdots$

All h with error bound 1.

All h with error bound 2.

All h with error bound n.

Every H_n is online learnable.

H is the union of all H_n.

Theorem: H is non-uniform stochastic online learnable, if and only if it's a countable union of Littlestone classes.

Adaptive x_t: strongest Nature.

Distribution μ: weakest Nature.

The characterization for the two settings are the same.

It implies the same characterization is also tight for any setting in between.

Outline

- The problem of induction
- Background
- Non-uniform online learning
- Characterizing inductive inference
- **The agnostic setting**
- Conclusion

So far, we considered the realizable setting. Not very realistic.

A natural generalization is the agnostic setting. No restriction on Nature.

Learner’s goal is to perform as well as the best hypothesis in some class.

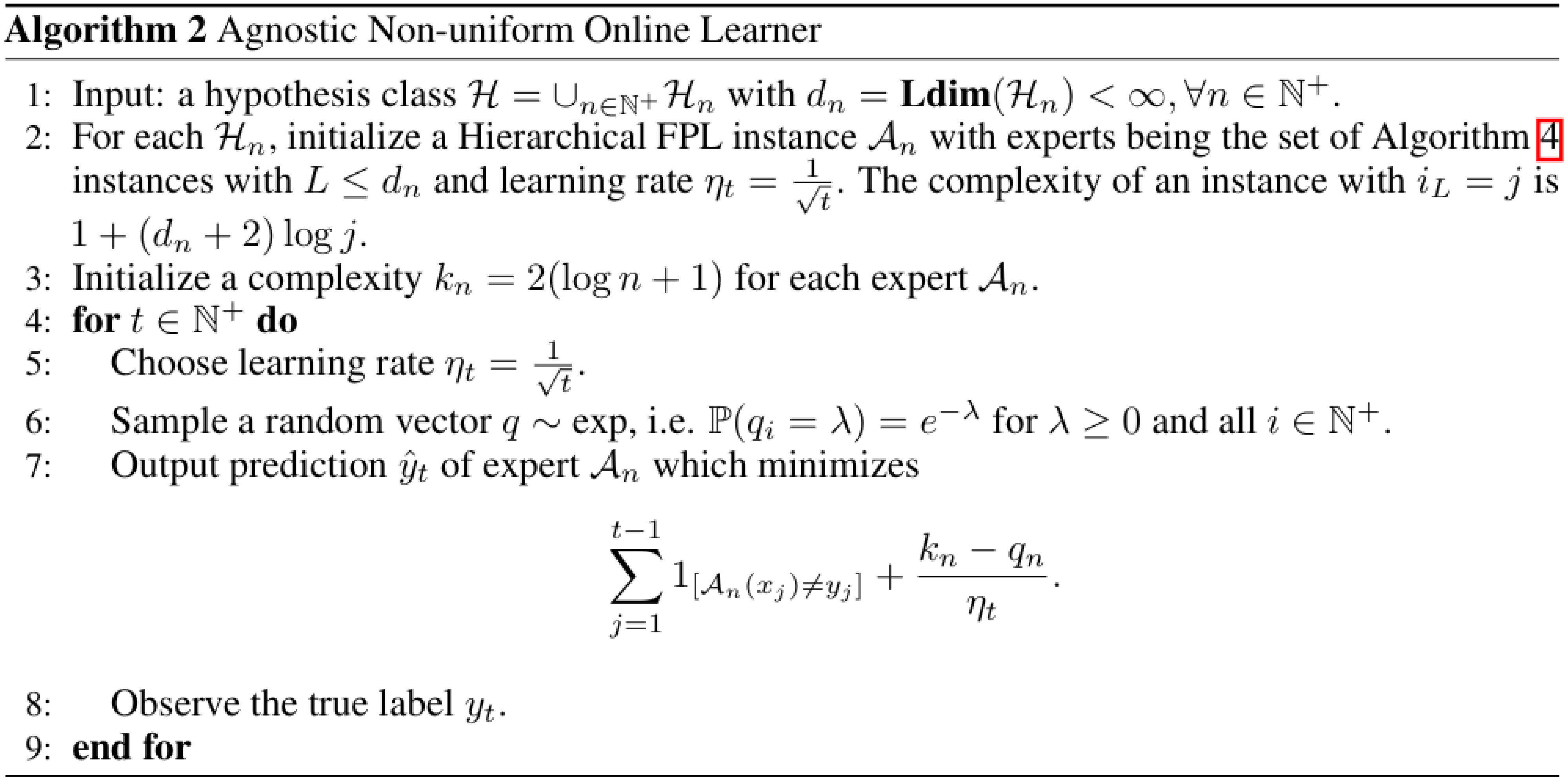
Agnostic online learnability

Definition We say a class H is agnostic non-uniform online learnable with rate r, if there exists an algorithm A, such that

$$\exists m: H \to N, \forall x \in X^{\infty}, \forall y^{*} \in \{0,1\}^{\infty}, \forall h \in H, \forall T \in N,$$

$$E\left[\sum_{t=1}^{T} 1_{y_t \neq y_t^{*}} - \sum_{t=1}^{T} 1_{h(x_t) \neq y_t^{*}}\right] \leq m(h)r(T)$$

The learning algorithm



Algorithm 2 Agnostic Non-uniform Online Learner

```
Input: a hypothesis class $\mathcal{H} = \cup_{n \in \mathbb{N}^+} \mathcal{H}_n$ with $d_n = \mathbf{Ldim}(\mathcal{H}_n) < \infty, \forall n \in \mathbb{N}^+$.
For each $\mathcal{H}_n$, initialize a Hierarchical FPL instance $\mathcal{A}_n$ with experts being the set of Algorithm $\boxed{4}$ instances with $L \leq d_n$ and learning rate $\eta_t = \frac{1}{\sqrt{t}}$. The complexity of an instance with $i_L = j$ is $1 + (d_n + 2) \log j$.
Initialize a complexity $k_n = 2(\log n + 1)$ for each expert $\mathcal{A}_n$.
for $t \in \mathbb{N}^+$ do
    Choose learning rate $\eta_t = \frac{1}{\sqrt{t}}$.
    Sample a random vector $q \sim$ exp, i.e. $\mathbb{P}(q_i = \lambda) = e^{-\lambda}$ for $\lambda \geq 0$ and all $i \in \mathbb{N}^+$.
    Output prediction $\hat{y}_t$ of expert $\mathcal{A}_n$ which minimizes
        $\sum_{j=1}^{t-1} 1_{[\mathcal{A}_n(x_j) \neq y_j]} + \frac{k_n - q_n}{\eta_t}.$
    Observe the true label $y_t$.
end for
```

Theorem When $H = \cup_n H_n$ is a countable union of Littlestone classes, each with Littlestone dimension d_n, then it's agnostic non-uniform learnable at rate $r(T) = \tilde{O}(\sqrt{T})$ and $m(h) = \log n + \sqrt{d_n}$, for any $h \in H_n$.

$r(T) = O(\sqrt{T})$ and $m(h) = \sqrt{d_n}$ is optimal for learning a single H_n.

Algorithm: two hierarchies of FPL.

A trichotomy

For any H, if it can't be written as a countable union of Littlestone classes, it's not agnostic non-uniform online learnable at rate $T^{1-\epsilon}$, for any $\epsilon > 0$.

When $|H| > 1, r(T) = \Omega(\sqrt{T})$.

Theorem There are only three possible rates of agnostic non-uniform online learning

- H is learnable at rate 0 iff $|H| = 1$.
- H is learnable at rate $\widetilde{\Theta}(\sqrt{T})$ iff H is a countable union of Littlestone classes.
- H has an arbitrarily slow rate iff H isn't a countable union of Littlestone classes.

Outline

- The problem of induction
- Background
- Non-uniform online learning
- Characterizing inductive inference
- The agnostic setting
- **Conclusion**

Conclusion

We introduced a new framework, non-uniform online learning, as a general form of inductive inference.

We show that inductive inference is possible, if and only if the hypothesis class is a countable union of online learnable classes.

This characterization is proven tight across many settings.

Future directions

A weaker criterion is consistency, allowing error bound m to depend on x in addition.

We obtain a necessary condition and leave a tight characterization as future work.

Our learning algorithms are uncomputable, which is unavoidable in general.

Can we build computable learner to achieve approximation results?

Thank you!