Learning in Pessiland via Inductive Inference

Shuichi Hirahara
NII

Mikito Nanashima
TokyoTech

Online Complexity Seminar

Backgrounds

Our Results

Proof Techniques

Conclusion

Backgrounds

Our Results

Proof Techniques

Conclusion

Pessiland

NP is Hard on Avg

Too Great Heuristics for SAT ✗

No One-Way Functions

(Classic) Crypto ✗

Suppose Our World is Unfortunately Pessiland.

What Can We Do?

Average-Case Inverter for Poly-Time Functions

↓

Algorithms in Pessiland

Hardness

Learning in Pessiland

$\nexists$OWF $\Longrightarrow$ Learning

"BigBang"

1990 [Impagliazzo-Levin] No Better Ways to Generate Hard NP Instances than Picking Uniformly at Random

- High-Level Ideas of "Universal Extrapolation"
- No Formal Statements

Which kind of learning tasks can be done ?

93 [Blum-Frust-Kearns-Lipton]

Average-Case PAC Learning

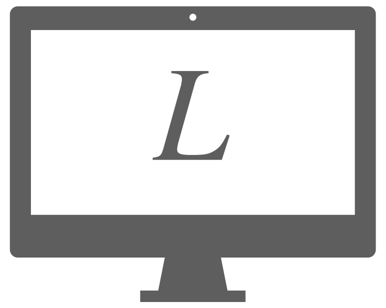
PAC Learning in Pessiland

[BFKL93] $\nexists$OWF$\Longrightarrow$ Avg PAC learning

a concept class $\mathscr{C} = \{\mathscr{C}_n\}_{n\in\mathbb{N}}$ $\mathscr{C}_n \subseteq \{f\colon \{0,1\}^n \to \{0,1\}\}$

an example distribution $\mathscr{E} = \{\mathscr{E}_n\}_{n\in\mathbb{N}}$ $\mathscr{E}_n$ is over $\{0,1\}^n$

a distribution over functions $\mathscr{F} = \{\mathscr{F}_n\}_{n\in\mathbb{N}}$ $\mathscr{F}_n$ is over $\mathscr{C}_n$



learner

in time poly(n, ϵ^{-1})

$\longleftarrow$ $(x^1, f(x^1)), \ldots, (x^m, f(x^m))$

$f \sim \mathscr{F}_n$

$x^1, \ldots, x^m \sim \mathscr{E}_n$

$\longrightarrow$ w.h.p. hypothesis h s.t. $\Pr_{x\sim\mathscr{E}}[h(x) \neq f(x)] \leq \epsilon$

[BFKL93] $\mathscr{C}$: efficiently evaluatable, $\mathscr{E}, \mathscr{F}$: samplable

Learning in Pessiland

$\nexists$OWF $\Longrightarrow$ Learning

"BigBang"

1990

[Impagliazzo-Levin] No Better Ways to Generate Hard NP Instances than Picking Uniformly at Random

- High-Level Ideas of "Universal Extrapolation"
- No Formal Statements

93

[Blum-Frust-Kearns-Lipton]

Average-Case PAC Learning

2006

[Naor-Rothblum]

Learning Adaptively Changing Distributions (ACDs)

Distributional Learning in Pessiland

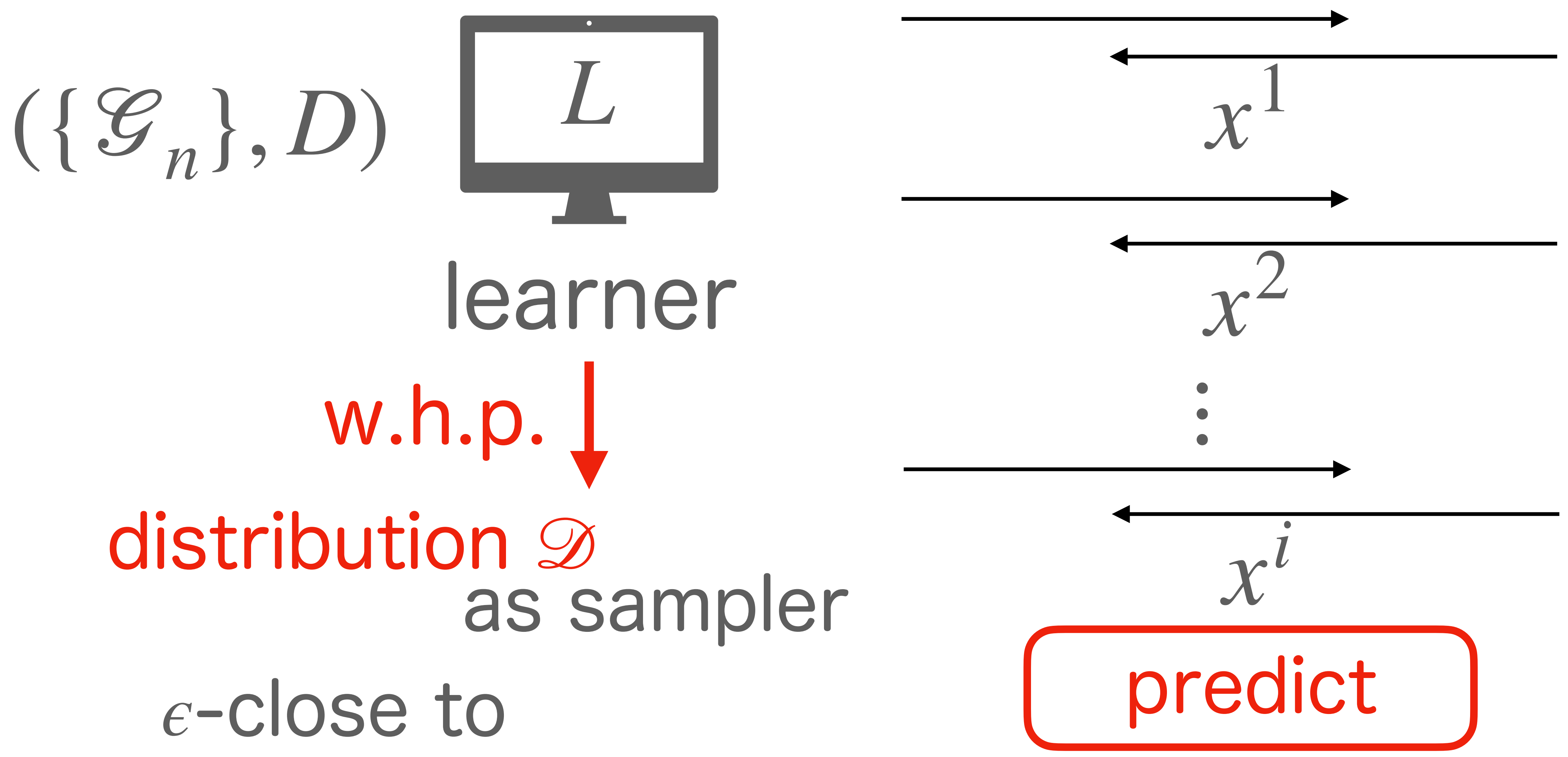
[NR06] $\nexists$OWF $\Longrightarrow$ Learning **known** ACDs

ACD $(\{\mathscr{G}_n\}, D)$

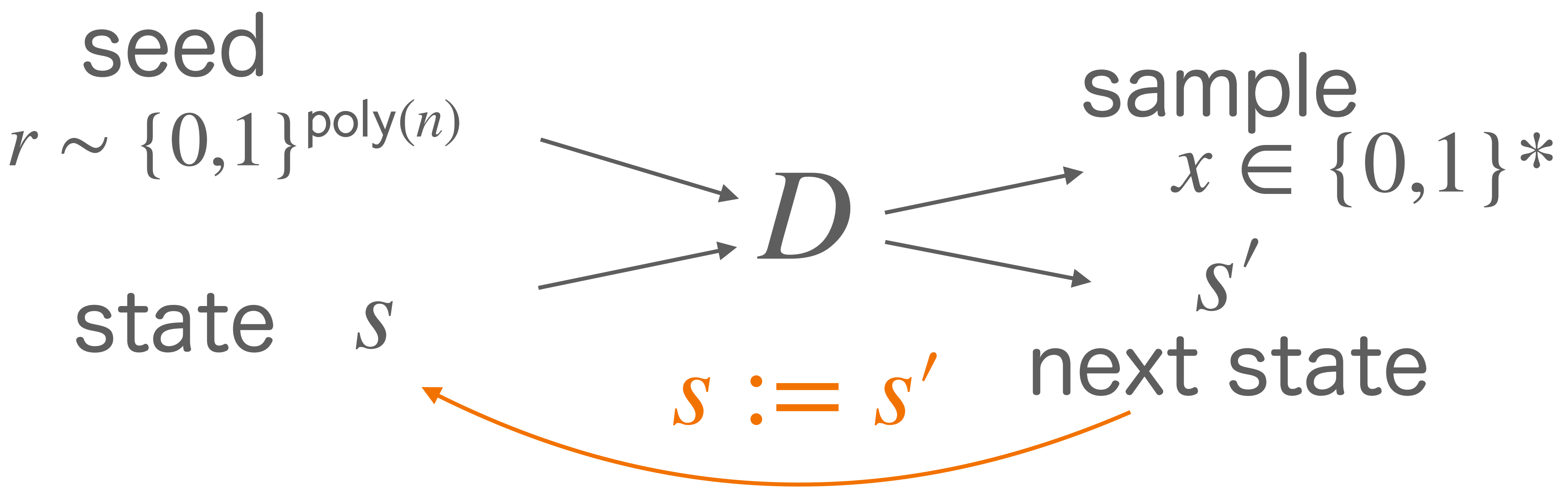
$\mathscr{G}$: samplable D : poly-time sampler

internal state s $= s_0$

Initialization: $s_0 \sim \mathscr{G}_n$
initial state **target**



ϵ-close to
conditional dist of x^{i+1}
given $x^1, \cdots, x^i$ & initial state s_0



[NR06] L uses the knowledge of $\{\mathscr{G}_n\}$ and D

Why is the knowledge of distributions important?

Critical Cases

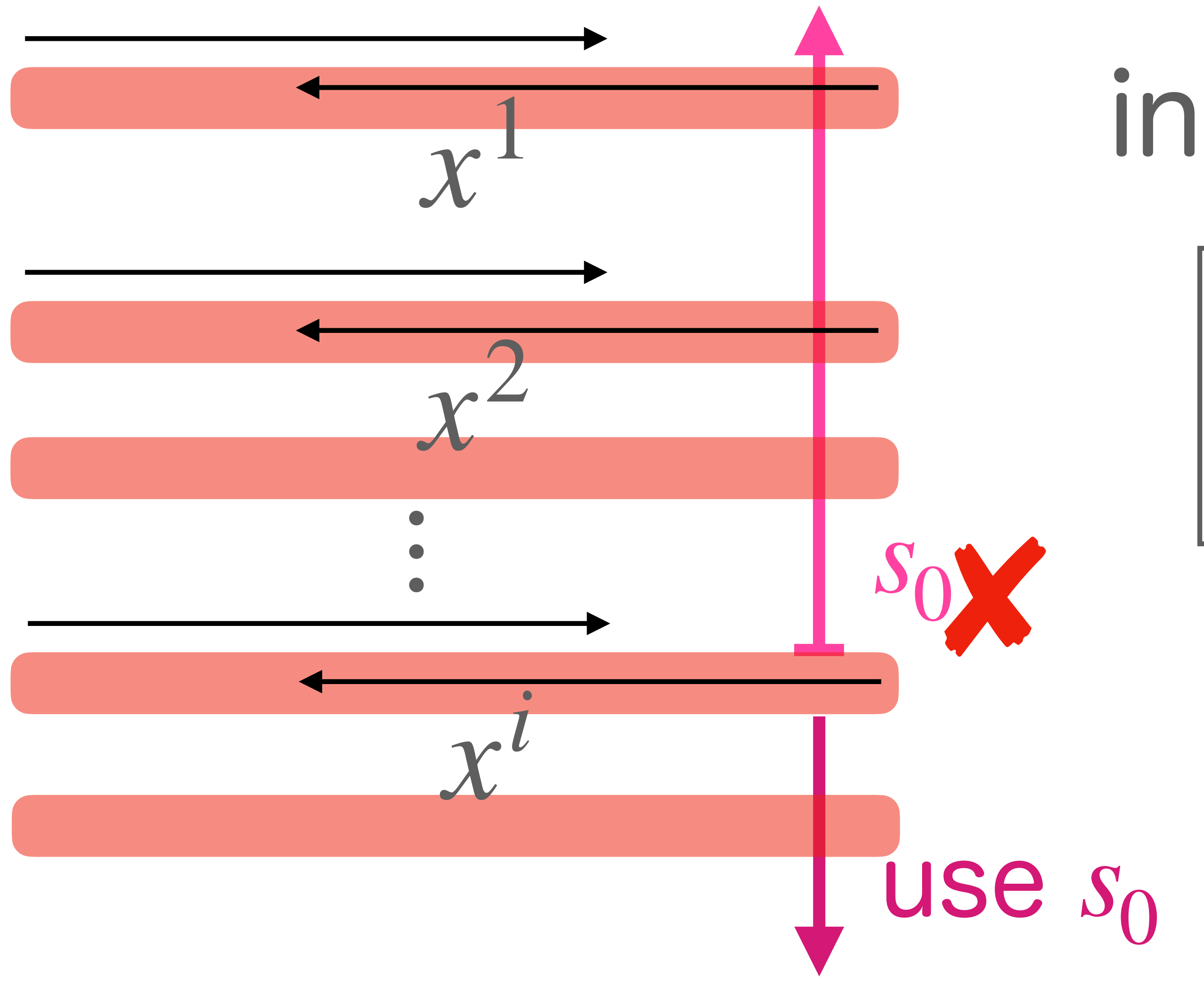
$(\{\mathscr{G}_n\}, D)$ ✗

learner L

w.h.p.

distribution $\mathscr{D}$ as sampler

ϵ-close to conditional dist of x^{i+1} given $x^1, \cdots, x^i$ & initial state s_0



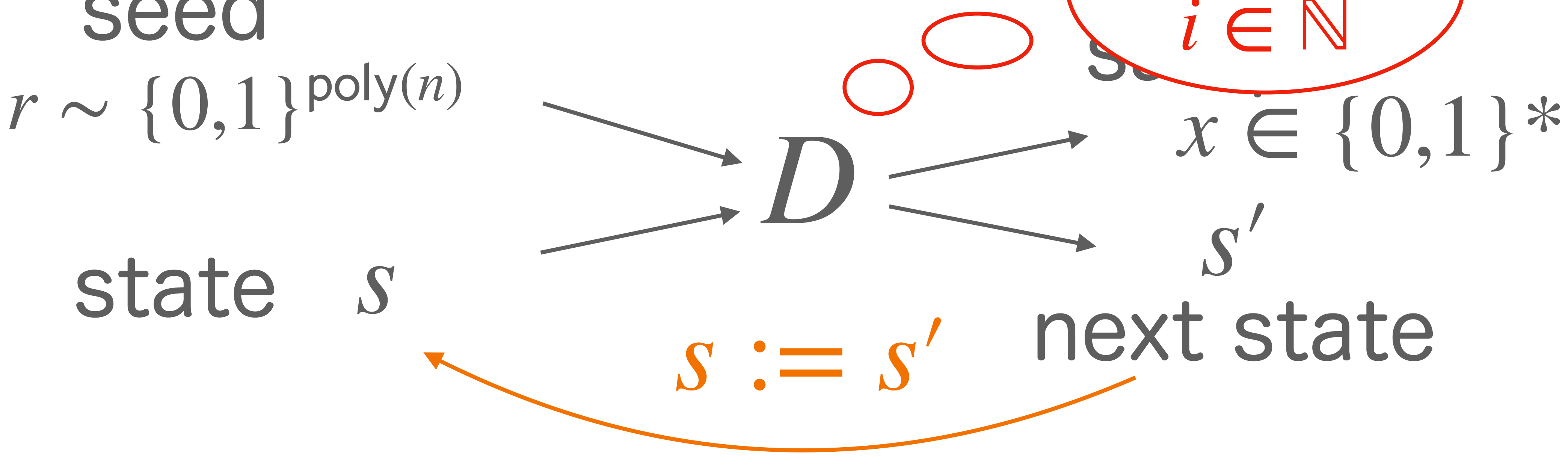
ACD $(\{\mathscr{G}_n\}, D)$

$\mathscr{G}$: samplable D : poly-time sampler

internal state s $= s_0$

Initialization: $s_0 \sim \mathscr{G}_n$
initial state **target**

specific $i \in \mathbb{N}$



[NR06] L uses the knowledge of $\{\mathscr{G}_n\}$ and D

to test whether the next sample is predictable

Improved Learning in Pessiland ?

[BFKL] examples and target function are separately selected

→ Q. Average-Case PAC Learning on Joint Distribution?

$\mathscr{D} \sim \mathscr{G}$ $(x, b) \sim \mathscr{D}$

↑
description of joint distribution

	Separated (BFKL)	Joint
	proper learning	→ DistNP-hard [Pitt-Valiant]

Q. Agnostic Learning ? $\Pr_{(x,b)}[h(x) \neq b] \leq \min_{f \in \mathscr{C}} \Pr_{(x,b)}[f(x) \neq b] + \epsilon$

[NR] learning known ACDs

→ Q. Learning unknown ACDs ?

Backgrounds

Our Results

Proof Techniques

Conclusion

Our Contribution

Resolve Q. Joint Distribution? Q. Learning unknown ACDs ?
Q. Agnostic Learning ?

1990

Impagliazzo-Levin, No Better Ways to Generate Hard NP Instances than Picking Uniformly at Random

- High-Level Ideas of “Universal Extrapolation” revisit
- No Formal Statements

Unified & Simplified & Improved

93

Blum-Frust-Kearns-Lipton

Average-Case PAC Learning

2006

Naor-Rothblum

Learning Adaptively Changing Distributions (ACDs)

In this talk...

1. $\nexists$ Infinitely Often OWF

"Adversaries can invert functions for all sufficiently large parameters"

$\nexists$ (standard) OWF $\longrightarrow$ infinitely many size n

accuracy, confidence $\leq 1/\text{poly}(n)$

fixed as poly-time functions in n

2. No details for the choices of parameters

I do not discuss confidence parameters so much

minimize $\mathrm{K}^{\text{your\_time}}(\text{the paper} \mid \text{this talk})$

Learning ACDs in Pessiland

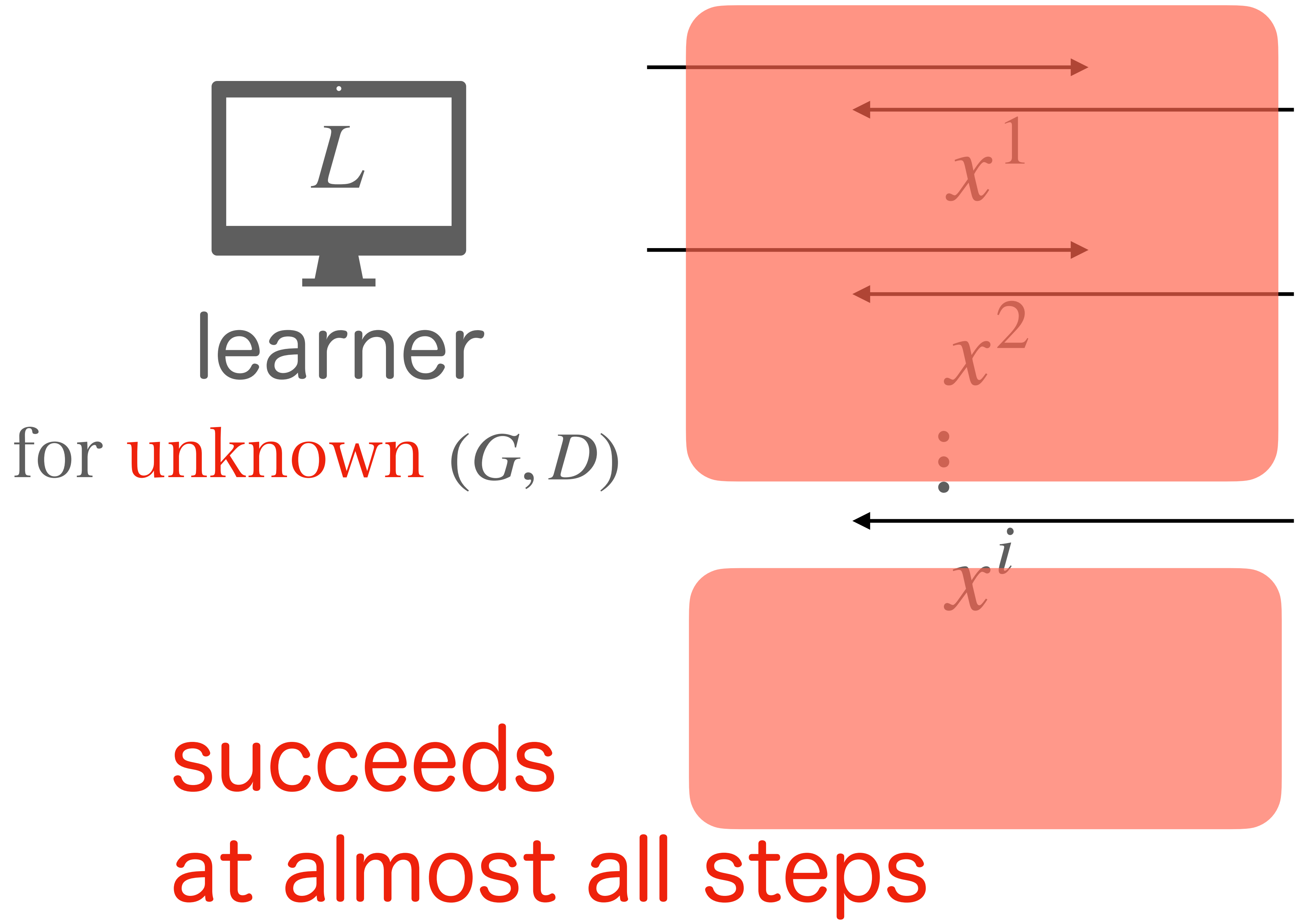
Q. Learning unknown ACDs ?

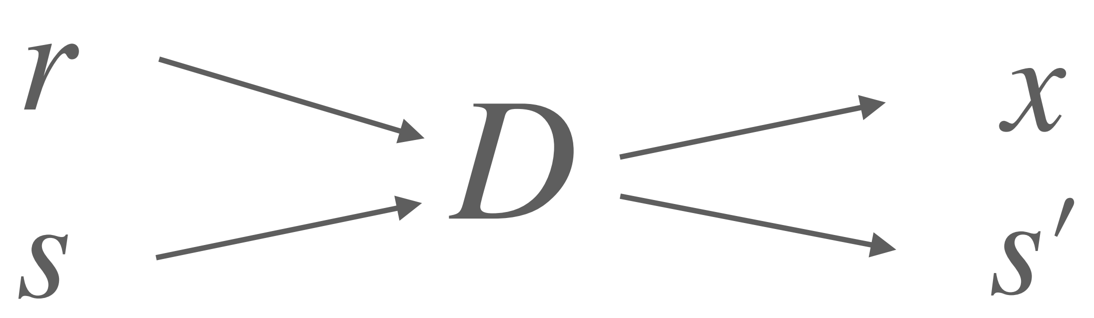
Thm. 1

∄OWF iff ∃ poly-time learner for all (unknown) ACDs $(\mathscr{G}, D)$
with sample complexity $O(s\epsilon^{-2})$ for s-bit initial states and accuracy ϵ

improved from $O(s\epsilon^{-4})$ [NR06]

How can we avoid the critical cases?





choose a prediction stage
uniformly at random

Agnostic Learning in Pessiland

Q. Joint Distribution?　　Q. Agnostic Learning ?

Thm. 2

$\nexists$OWF iff $\exists$ poly-time agnostic learner for $\mathscr{F} = \{f\colon \{0,1\}^n \to \{0,1\}^{\mathrm{poly}(n)}\}$ (with 0-1 loss)

(= learning by information theoretically optimal hypothesis)

on avg under a joint dist $\mathscr{D}$ on samples, where $\mathscr{D} \sim \mathscr{G}$, $\mathscr{G}$ is samplable

with sample complexity $O(s\epsilon^{-2})$ (for $s = |\mathscr{D}|$, accuracy ϵ)

optimal in general

[Previous]		[Ours]
PAC	▶	Agnostic
Separated Distributions		Joint Distributions
Binary Labels		Multi Labels

Improper learning
(General Hypothesis)

Improved Learning in Pessiland

$\nexists$OWF $\iff$ Worst-Case Learning

exp-time in Computational Depth of Secrets

U: universal TM

$$\mathrm{K}(x) = \min\{p \in \mathbb{N} : \exists \Pi \in \{0,1\}^p \text{ s.t. } U(\Pi) = x\}$$

Q^t := dist. of $U(w)$ executed in t steps for $w \sim \{0,1\}^t$

$$\mathrm{q}^t(x) := -\log \Pr[x \sim \mathrm{Q}^t]$$

$\mathrm{q}^t(x) \approx \mathrm{pK}^{t'}(x)$ introduced in [GKLO22]

$\mathrm{pK}^{\mathrm{poly}(t)}(x) \lesssim \mathrm{q}^t(x)$ Optimal coding [LOZ22]

$\mathrm{pK}^t(x) \gtrsim \mathrm{q}^{\mathrm{poly}(t)}(x)$ Domination

$$\mathrm{cd}^t(x) := \mathrm{q}^t(x) - \mathrm{K}(x)$$

$$\approx \mathrm{pK}^{t'}(x) - \mathrm{pK}^{\infty}(x)$$

for any samplable distribution $\mathscr{D} = \{\mathscr{D}_n\}$

$\mathrm{cd}^{\mathsf{poly}}(x) = O(\log n)$ w.h.p. $x \sim \mathscr{D}_n$

Slow Growth Law

$x \longrightarrow$ M:PPT $\longrightarrow x'$

$\mathrm{cd}^{O(t+time_M)}(x') \lesssim \mathrm{cd}^t(x)$ w.h.p.

Improved Learning in Pessiland

$\nexists$OWF $\iff$ Worst-Case Learning

exp-time in Computational Depth of Secrets

Thm. 3

The following are equivalent:

1. $\nexists$ OWF
2. learning unknown ACDs in time $\mathrm{poly}(n, t, \epsilon^{-1}, 2^{\mathrm{cd}^t(s_0)})$, where s_0 is initial state (worst case on initial states)
3. agnostic learning on unknown joint dists $\mathscr{D}$ over samples in time $\mathrm{poly}(n, t, \epsilon^{-1}, 2^{\mathrm{cd}^t(|\mathscr{D}|)})$ (worst case on joint distributions over samples)

Note Thm1 & 2 are implied by Thm3

Our Approach

Step I State “Universal Extrapolation” formally

Step II Translate “Universal Extrapolation” into Learning

Our Approach

Step I State "Universal Extrapolation" formally

Step II Translate "Universal Extrapolation" into Learning

Formulating Universal Extrapolation

Our Proposal Universal Extrapolation = Extrapolation under Q^t

(Time-Bounded Universal Distribution)

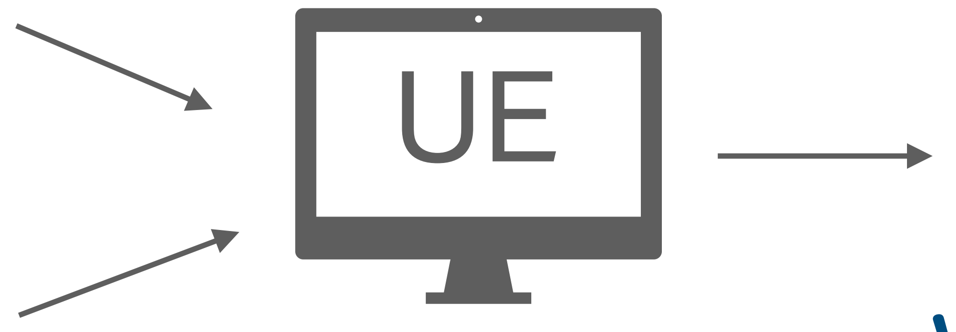
Notation distribution $\mathscr{D}$ prefix $x \in \{0,1\}^*$ $k \in \mathbb{N}$

$\mathsf{Next}_k(x; \mathscr{D})$ = distribution of k bits following x w.r.t. $\mathscr{D}$

prefix $x \in \{0,1\}^*$

paramaters

$k \in \mathbb{N}$ $t \in \mathbb{N}$ $\epsilon \in (0,1)$



$y \in \{0,1\}^{\leq k} \approx \mathsf{Next}_k(x; \mathrm{Q}^t)$

within statistical distance ϵ

Lemma (informal) $\nexists$ (io)OWF $\implies$ $\exists$UE that works in worst case on x

in time $\mathrm{poly}(|x|, k, t, \epsilon^{-1}, 2^{\mathrm{cd}^t(x)})$

(UE is given 2^α (in unary) and works for every x with $\mathrm{cd}^t(x) < \alpha$)

Lemma (informal) $\nexists$ (io)OWF $\Longrightarrow$ $\exists$UE that works in worst case on x in time $\mathrm{poly}(|x|, k, t, \epsilon^{-1}, 2^{\mathrm{cd}^t(x)})$

Distributional Inverting

Inverting

$f(x)$ $(x \sim \{0,1\}^*)$ $\longrightarrow$ Inv $\longrightarrow$ some x' $f(x') = f(x)$

Distributional Inverting

$f(x)$ $(x \sim \{0,1\}^*)$ $\longrightarrow$ $DInv$ $\longrightarrow$ some x' $f(x') = f(x)$

simulate unif sampling from $\{x' : f(x') = f(x)\}$

Thm [IL89]

$\nexists$ (io)OWF $\Longrightarrow$ for every poly-time function $f = \{f_n\}$

$\exists DInv$: PPT s.t. $\forall n, \epsilon^{-1}, \delta^{-1} \in \mathbb{N}$

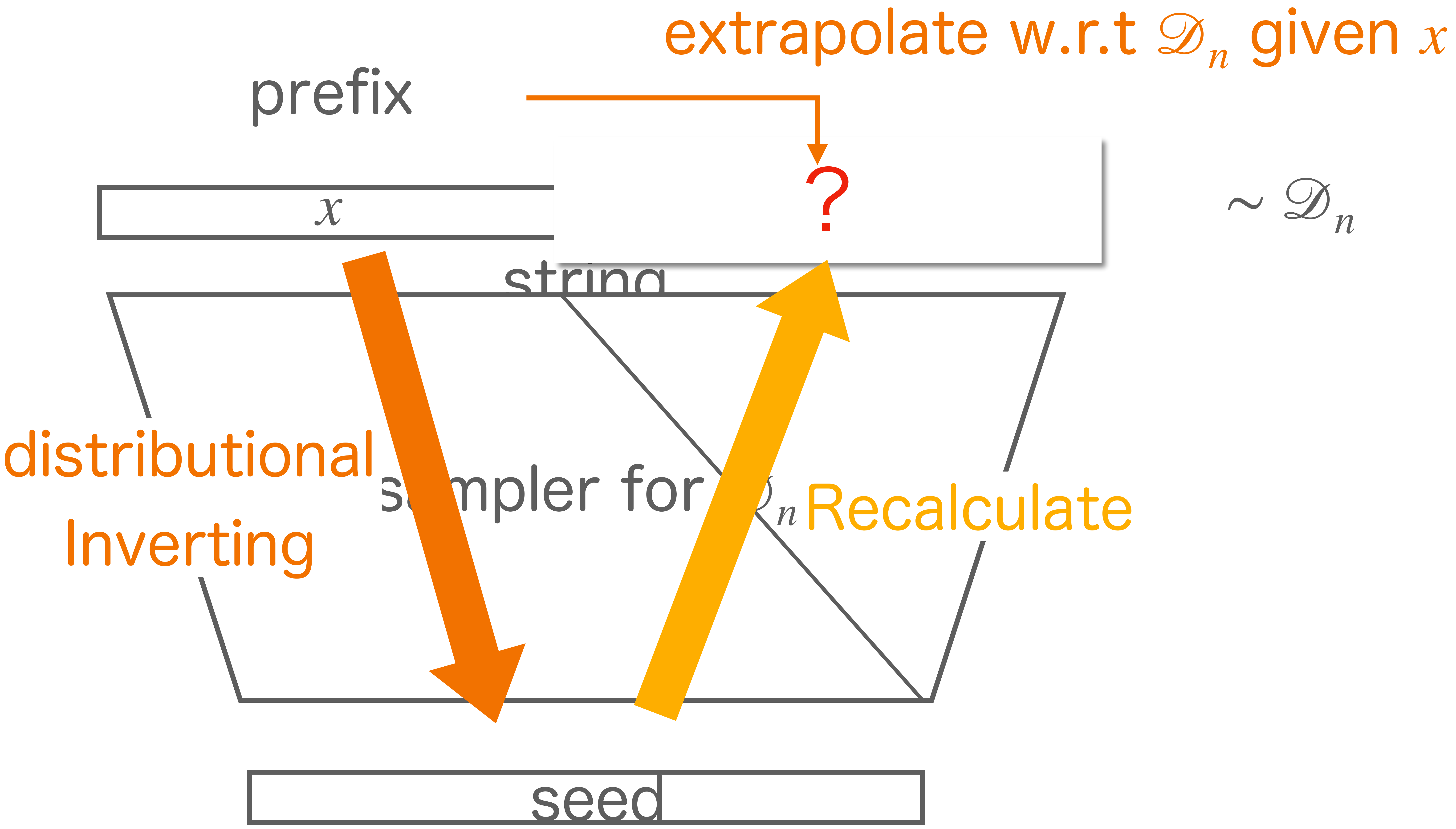
$$\Delta_{TV}\left(DInv(f_n(x); 1^n, 1^{\epsilon^{-1}}, 1^{\delta^{-1}}), \text{Unif over } f_n^{-1}(f_n(x))\right) \leq \epsilon$$

w.p. $\geq 1 - \delta$ over $x \sim \{0,1\}^{\mathrm{poly}(n)}$

Lemma (informal) $\nexists$ (io)OWF $\Longrightarrow$ $\exists$UE that works in worst case on x in time $\mathrm{poly}(|x|, k, t, \epsilon^{-1}, 2^{\mathrm{cd}^t(x)})$

Distributional Inverting → Distribution-Specific Extrapolation [Ost91, OW93, NR06, . . .]

$\mathscr{D} = \{\mathscr{D}_n\}$ samplable distribution



Lemma (informal) $\nexists$ (io)OWF $\Longrightarrow$ $\exists$UE that works in worst case on x in time $\mathrm{poly}(|x|, k, t, \epsilon^{-1}, 2^{\mathrm{cd}^t(x)})$

Distribution-Specific Extrapolation for Q^t

$\exists\mathrm{UE}'$: PPT s.t. $\forall t, \ell, k, \epsilon^{-1}, \delta^{-1} \in \mathbb{N}$

$$\Delta_{TV}\left(\mathrm{UE}'(x), \mathrm{Next}_k(x; \mathrm{Q}^t)\right) \le \epsilon \quad \text{w.p. } \ge 1-\delta \text{ over } x \sim \mathrm{Q}^t_{\le \ell}$$

the first ℓ bits of Q^t

Based on [AF09] $E := E_{t,\ell,k,\epsilon,\delta} = \left\{ x : \Delta_{TV}\left(\mathrm{UE}'(x), \mathrm{Next}_k(x; \mathrm{Q}^t)\right) > \epsilon \right\}$ error set

$\delta := 2^{-\alpha}$ Goal: $x \in E \Longrightarrow \mathrm{cd}^t(x) \ge \alpha - O(\log t\ell k \epsilon^{-1} \alpha)$

$$2^{-\alpha} \ge \sum_{x \in E} \Pr[x \sim \mathrm{Q}^t_{\le \ell}] \ge \sum_{x \in E} \Pr[x \sim \mathrm{Q}^t] \qquad \sum_{x \in E} 2^{\alpha} \Pr[x \sim \mathrm{Q}^t] \le 1$$

computable given UE$'$ & parameters

inefficiently computable distribution $\{\mathscr{E}_{t,\ell,k,\epsilon,\delta}\}$ $\forall x \in E$ $\Pr[x \sim \mathscr{E}] = 2^{\alpha} \Pr[x \sim \mathrm{Q}^t]$

$\forall x \in E$ $\mathrm{K}(x) \le -\log \Pr[x \sim \mathscr{E}] + O(\log t\ell k \epsilon \alpha)$ optimal coding

$$= -\alpha + \underbrace{(-\log \Pr[x \sim \mathrm{Q}^t])}_{\mathrm{q}^t(x)} + O(\log t\ell k \epsilon \alpha)$$

Step 1: Summary

Our Proposal Universal Extrapolation = Extrapolation under Q^t

(Time-Bounded Universal Distribution)

prefix $x \in \{0,1\}^*$

paramaters

$k \in \mathbb{N}$ $t \in \mathbb{N}$ $\epsilon \in (0,1)$

→ UE → $y \in \{0,1\}^k$ $\approx$ $\mathrm{Next}_k(x; \mathrm{Q}^t)$

with statistical distance within ϵ

Lemma (informal) $\nexists$ (io)OWF $\Longrightarrow$ $\exists$UE that works in worst case on x

in time $\mathrm{poly}(|x|, k, t, \epsilon^{-1}, 2^{\mathrm{cd}^t(x)})$

(UE is given 2^α (in unary) and works for every x with $\mathrm{cd}^t(x) < \alpha$)

Proof Use Distributional Inverter for Q^t

Q. How can we obtain learners (e.g., agnostic learners)?

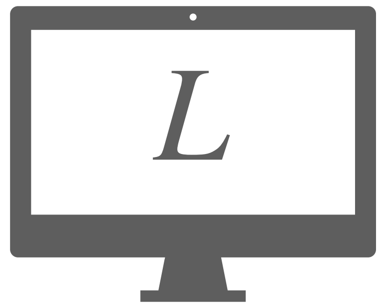
Step2 UE → Learning

offline stream

efficiently & randomly

$x^1 \ y^1 \ x^2 \ y^2 \ \cdots \ x^m \ y^m$ ← secret info. (sampler) $\Pi \in \{0,1\}^*$

possibly correlated



$x^1 \ y^1 \ x^2 \ y^2 \ \cdots \ x^{i-1} \ y^{i-1} \ x^i$ y^i **?**

predict

$x^1, y^1, x^2, y^2 \ldots \in \{0,1\}^*$

x : advice information

y : target

EX. learning ACDs

$\Pi \approx$ initial state x : empty y : sample

simulate the distribution of y^i

agnostic learning (with 0-1 loss)

$\Pi \approx$ joint distribution over samples

x : example y : label

answer the best possible y^i

Cheating Learner

offline stream

efficiently & randomly

$x^1 \ y^1 \ x^2 \ y^2 \ \cdots \ x^m \ y^m$ ← secret info. (sampler) $\Pi \in \{0,1\}^*$

possibly correlated

$x^1 \ y^1 \ x^2 \ y^2 \ \cdots \ x^{i-1} \ y^{i-1} \ x^i$ y^i **?**

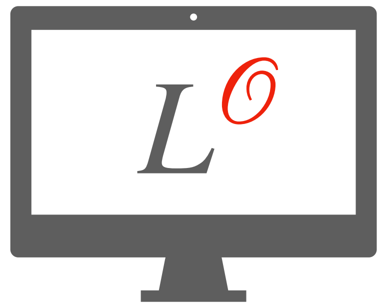
$x^1, y^1, x^2, y^2 \ldots \in \{0,1\}^*$

x : advice information

y : target

condition

$h^{\mathcal{O}}$



$y^i \ y^i \ y^i$ ← $\mathcal{O}$

w.r.t. Π

EX. learning ACDs

$\Pi \approx$ initial state x : empty y : sample

simulate the distribution of y^i

✅ by outputting $y^i \sim \mathcal{O}$

agnostic learning (with 0-1 loss)

$\Pi \approx$ joint distribution over samples

x : example y : label

answer the best possible y^i

✅ by collecting $y^i, y^{i'}, \ldots \sim \mathcal{O}$

UE $\rightarrow$ Learning

offline stream

efficiently & randomly

secret info. (sampler) $\Pi \in \{0,1\}^*$

$x^1 \; y^1 \; x^2 \; y^2 \; \cdots \; x^m \; y^m$

possibly correlated

$x^1 \; y^1 \; x^2 \; y^2 \; \cdots \; x^{i-1} \; y^{i-1} \; x^i$ $\quad y^i$?

$x^1, y^1, x^2, y^2 \ldots \in \{0,1\}^*$

x : advice information

y : target

condition

h L

$y^i \; y^i \; y^i$ $\leftarrow$ $\mathcal{O}$

w.r.t. Π

$\approx$

UE

w.h.p. when $i \sim [m]$

for statistical distance within ϵ

$m = O(|\Pi|\epsilon^{-2})$

Q. Why ?

learning ACDs ✓ agnostic learning (with 0-1 loss) ✓

sample complexity: linear in $|\Pi|$ time complexity: exp in $|\Pi|$

Solomonoff's Inductive Inference [Sol64, LV19]

When extrapolating symbols, attach a higher probability for a more precise hypothesis, particularly, with an exponential rate on the description size

$q^{poly}(\cdot\,|\,\cdot)$ or $pK^{poly}(\cdot\,|\,\cdot)$

$(\because 2^{-q^{poly}} = Q^{poly})$

by conditional coding [HILNO23]

Domination + Chain Rule for KL divergence

efficiently & randomly

$$x^1 \quad y^1 \quad x^2 \quad y^2 \quad \cdots \quad x^m \quad y^m \longleftarrow \boxed{\Pi \in \{0,1\}^*}$$

When $t \gg \text{time}(\Pi)$

$$\Pr\left[x^1y^1\ldots x^my^m \sim Q^t\right] \geq 2^{-O(|\Pi|)}\Pr\left[x^1y^1\ldots x^my^m \sim \Pi\right] \qquad \textbf{(Domination)}$$

$$\log \frac{\Pr\left[x^1y^1\ldots x^my^m \sim \Pi\right]}{\Pr\left[x^1y^1\ldots x^my^m \sim Q^t\right]} \leq O(|\Pi|)$$

Taking the expectation over $x^1, y^1, \ldots, x^m, y^m \sim \Pi$

$$\mathrm{KL}\left(\Pi \,\|\, Q^t\right) \leq O(|\Pi|)$$

Solomonoff's Inductive Inference [Sol64, LV19]

$\mathrm{KL}\left(\Pi \,\middle\|\, Q^t\right) \le O(|\Pi|)$

$\mathcal{O}$ 

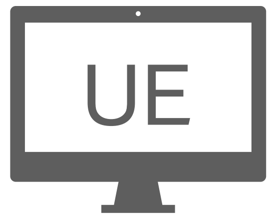
true distribution

$x^1\ y^1\ x^2\ y^2\ \cdots\ y^{i-1}\ x^i\ \ y^i\ \cdots\ x^m\ y^m \longleftarrow \boxed{\Pi \in \{0,1\}^*}$

$X^1\ Y^1 \quad X^m\ Y^m$

$\tilde{X}^1\ \tilde{Y}^1 \quad \tilde{X}^m\ \tilde{Y}^m$

$x^1\ y^1\ x^2\ y^2\ \cdots\ y^{i-1}\ x^i\ \ y^i\ \cdots\ x^m\ y^m \longleftarrow Q^t$



ignore the statistical error

$$O(|\Pi|) \ge \mathrm{KL}\left(X^1Y^1\cdots X^mY^m \,\middle\|\, \tilde{X}^1\tilde{Y}^1\cdots\tilde{X}^m\tilde{Y}^m\right)$$

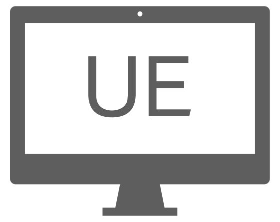
$$= \sum_{i=1}^{m} \mathrm{KL}\left((Y^i \mid X^1Y^1\cdots Y^{i-1}X^i) \,\middle\|\, (\tilde{Y}^i \mid \tilde{X}^1\tilde{Y}^1\cdots\tilde{Y}^{i-1}\tilde{X}^i)\right)$$

(Chain Rule)

$$+ \sum_{i=1}^{m} \mathrm{KL}\left((X^i \mid X^1Y^1\cdots X^{i-1}Y^{i-1}) \,\middle\|\, (\tilde{X}^i \mid \tilde{X}^1\tilde{Y}^1\cdots\tilde{X}^{i-1}\tilde{Y}^{i-1})\right)$$

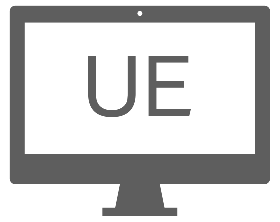
$$\ge \sum_{i=1}^{m} \mathrm{KL}\left((Y^i \mid X^1Y^1\cdots Y^{i-1}X^i) \,\middle\|\, (\tilde{Y}^i \mid \tilde{X}^1\tilde{Y}^1\cdots\tilde{Y}^{i-1}\tilde{X}^i)\right)$$

 i

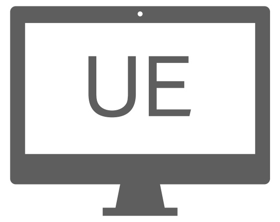


i

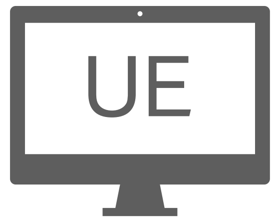
Solomonoff's Inductive Inference [Sol64, LV19]

$O(|\Pi|) \geq \sum_{i=1}^{m} \mathrm{KL}($  $_i \,\|\,$ 

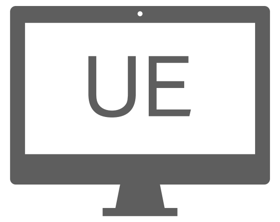
$_i\ )$

$\dfrac{O(|\Pi|)}{m} \geq \mathrm{E}_{i\sim[m]}[\mathrm{KL}($  $_i \,\|\,$ 

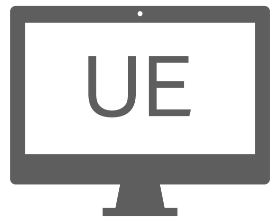
$_i\ )]$

$m \gg \dfrac{|\Pi|}{\epsilon^2} \implies \mathrm{E}_{i\sim[m]}[\mathrm{KL}($  $_i \,\|\,$ 

$_i\ )] \ll \epsilon^2$

(Markov) $\mathrm{KL}($  $_i \,\|\,$ 

$_i\ ) \ll \epsilon^2$ w.h.p. over $i \sim [m]$

(Pinsker) $\Delta_{TV}($  $_i\ ,$ 

$_i\ ) \leq \epsilon$ w.h.p. over $i \sim [m]$

Remarks

$x^1\ y^1\ x^2\ y^2\ \cdots\ x^{i-1}\ y^{i-1}\ x^i$ y^i **?**

condition

h ✗ L ✗ $y^i\ y^i\ y^i$ ← $\mathcal{O}$ ✗

$\approx$ UE

when $i \sim [m]$ for stat dist within ϵ

$m = O(|\Pi|\epsilon^{-2})$

Tasks poly-time cheating learners can do

= Tasks poly-time learners can do with UE

Not sample optimal in agnostic learning

Why? accuracy → query complexity of cheating learner

for optimal sample complexity

Extend universal prediction [MF98] to computational cases

(statistical cases)

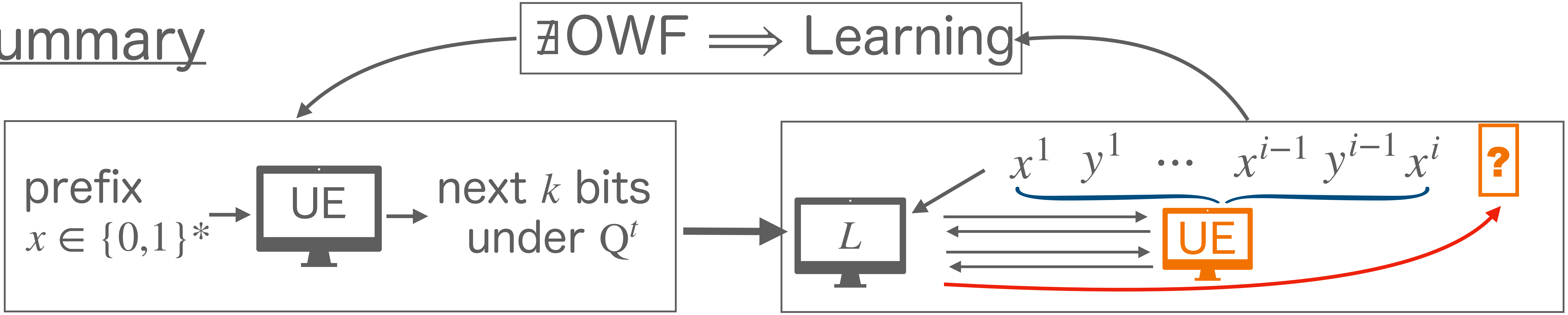
Backgrounds

Our Results

Proof Techniques

Conclusion

Summary



Thm.

The following are equivalent:

1. ∄ OWF
2. learning unknown ACDs in time $\mathrm{poly}(t, \epsilon^{-1}, 2^{\mathrm{cd}^t(s_0)})$, where s_0 is initial state
3. agnostic learning on unknown joint dists $\mathscr{D}$ over samples in time $\mathrm{poly}(t, \epsilon^{-1}, 2^{\mathrm{cd}^t(|\mathscr{D}|)})$

Q. weakest assumption for learning in time $\mathrm{poly}(\mathrm{cd})$? AIOWF or HSG or $\cdots$?