Lecture 11
Planning under v* realizability (TensorPlan I.)

High accuracy planning?

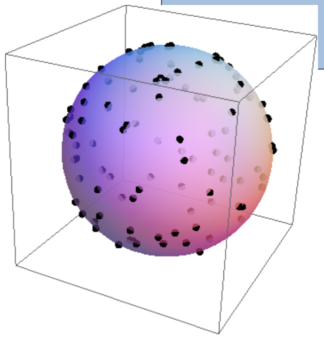
$\delta \ll \sqrt{d\varepsilon}$?

$\forall \pi : q^{\pi} \in_{\varepsilon} \mathcal{F}$

Yes

No

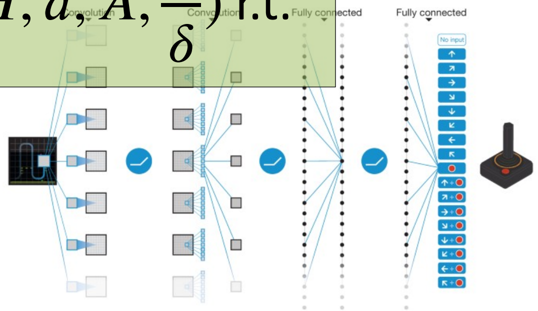
#queries=$2^{\Omega(d \wedge H)}$ ☠



API≈DQN +

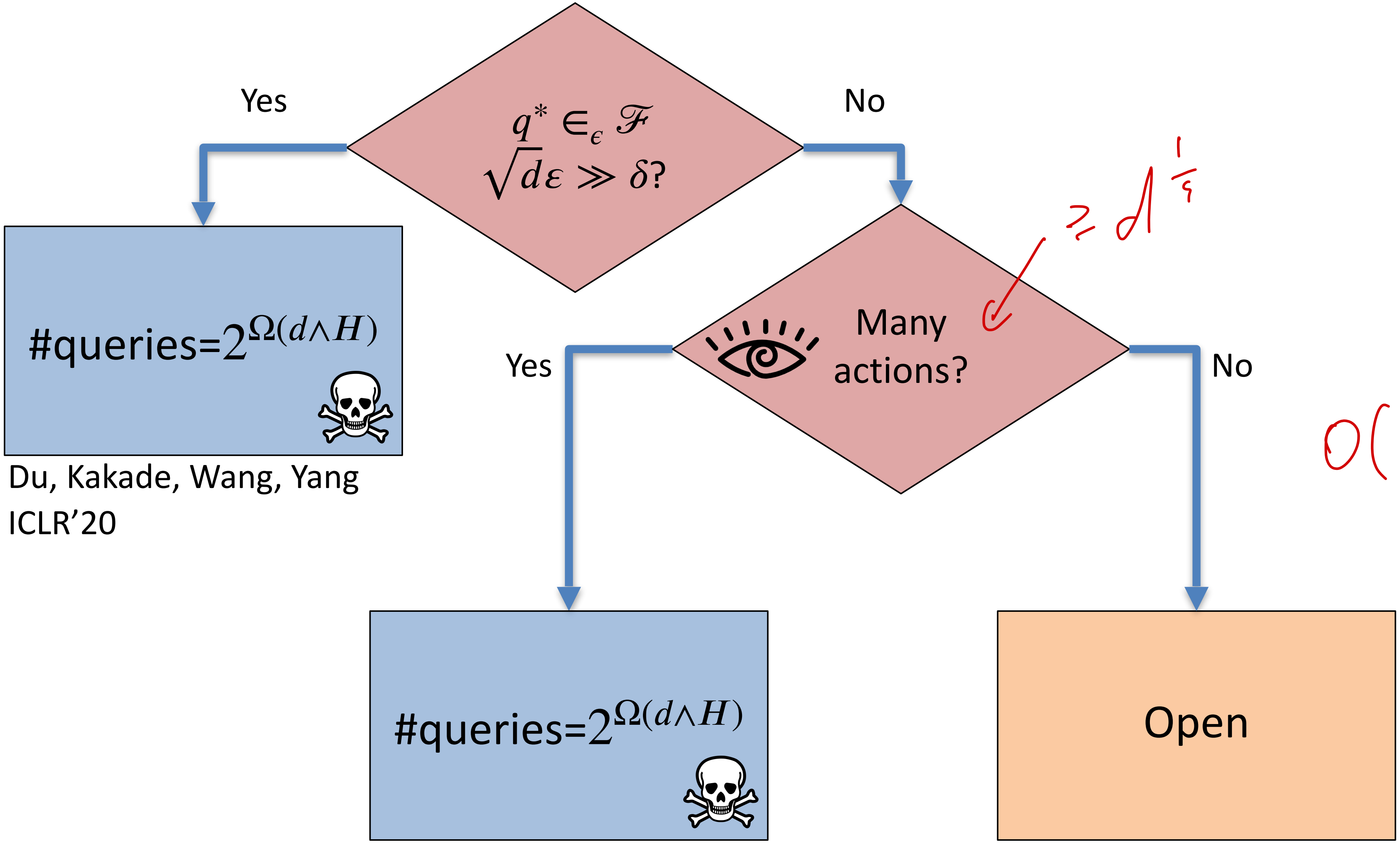
G-opt. design

$\mathrm{poly}(H, d, A, \frac{1}{\delta})$ r.t.



Du, Kakade, Wang, Yang ICLR’20

Lattimore, Sz, Weisz, ICML’20



$q^* \in_\epsilon \mathscr{F}$

$\sqrt{d}\varepsilon \gg \delta$?

Yes

No

#queries=$2^{\Omega(d \wedge H)}$ ☠

Du, Kakade, Wang, Yang
ICLR’20

Many actions?

Yes

No

$\gtrsim d^{\frac{1}{4}}$

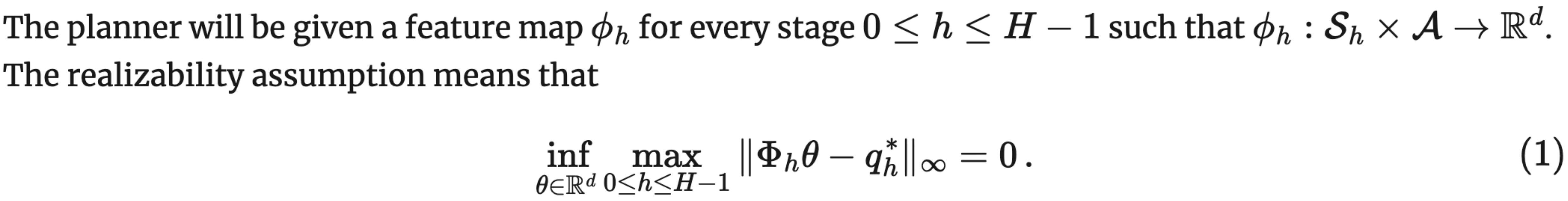
$O(1)$

#queries=$2^{\Omega(d \wedge H)}$ ☠

Weisz, Amortila, Sz
ALT’20

Open

Previous lecture: q* realizability

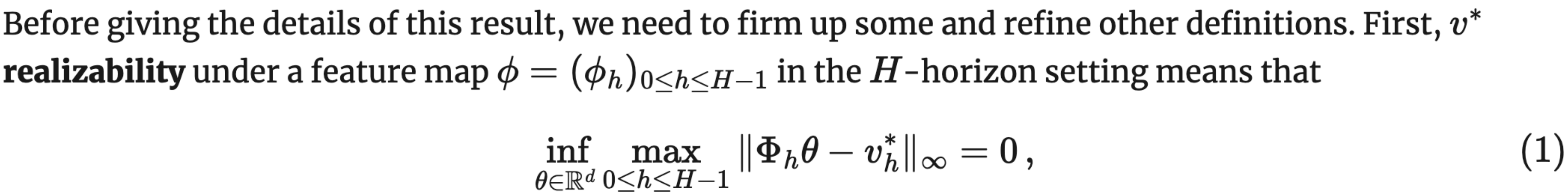


The planner will be given a feature map ϕ_h for every stage $0 \le h \le H-1$ such that $\phi_h : \mathcal{S}_h \times \mathcal{A} \to \mathbb{R}^d$. The realizability assumption means that

$$\inf_{\theta \in \mathbb{R}^d} \max_{0 \le h \le H-1} \|\Phi_h \theta - q_h^*\|_\infty = 0\,. \tag{1}$$

Intractable if at least d^1/4 actions

This lecture: v* realizability



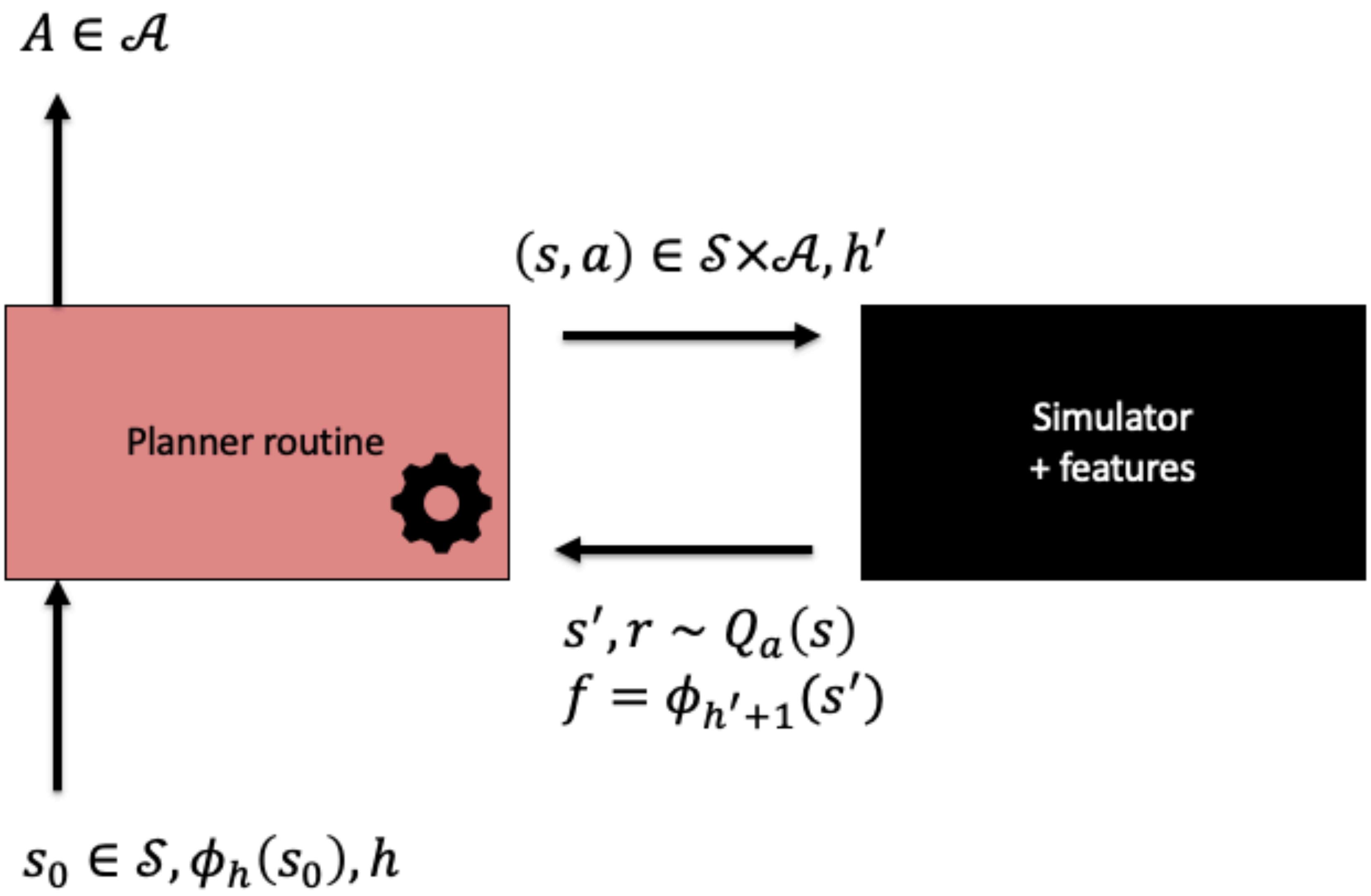
Before giving the details of this result, we need to firm up some and refine other definitions. First, v^* **realizability** under a feature map $\phi = (\phi_h)_{0 \le h \le H-1}$ in the H-horizon setting means that

$$\inf_{\theta \in \mathbb{R}^d} \max_{0 \le h \le H-1} \|\Phi_h \theta - v_h^*\|_\infty = 0\,, \tag{1}$$

also assume ||theta||_2<=B

Tractable sample complexity if O(1) actions

Online planning, local access



$A \in \mathcal{A}$

$(s, a) \in \mathcal{S} \times \mathcal{A}, h'$

Planner routine

Simulator
+ features

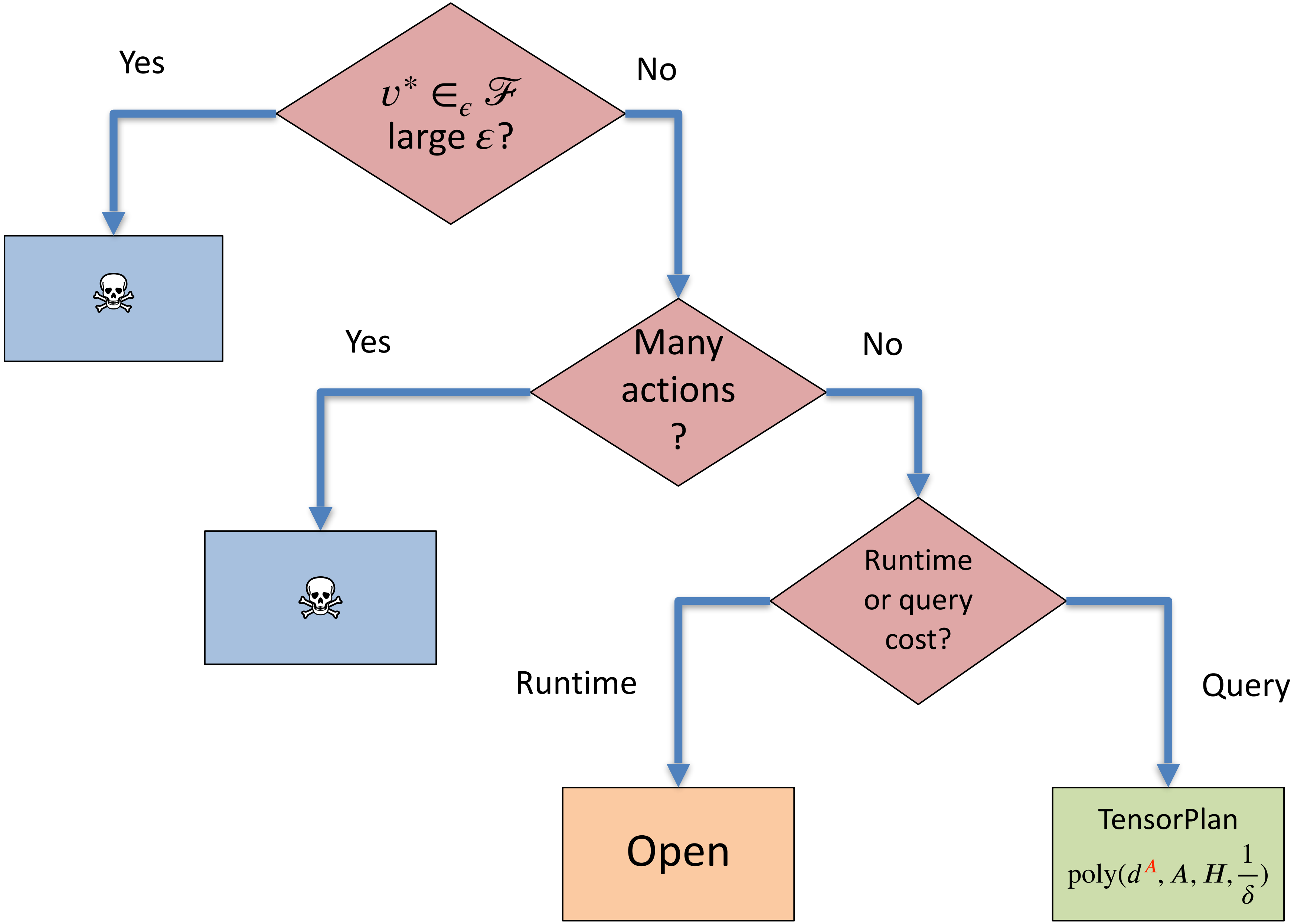
$s', r \sim Q_a(s)$
$f = \phi_{h'+1}(s')$

$s_0 \in \mathcal{S}, \phi_h(s_0), h$

Theorem (query-efficient planning under v^*-realizability): For any integers $A, H > 0$ and reals $B, \delta > 0$, there exists an online planner $\mathcal{P}$ with the following properties:

1. The planner $\mathcal{P}$ is δ-sound for the H-horizon planning problem and the class of MDP-feature-map pairs (M, ϕ) such that v^* is B-realizable under ϕ and M has at most A actions and its rewards are bounded in $[0, 1]$;
2. The number of queries used by the planner in each of its call is at most

$$\operatorname{poly}\left(\left(\frac{dH}{\delta}\right)^A, B\right)$$



$v^* \in_\epsilon \mathcal{F}$

large ε?

Yes

☠

No

Many actions ?

Yes

☠

No

Runtime or query cost?

Runtime

Open

Query

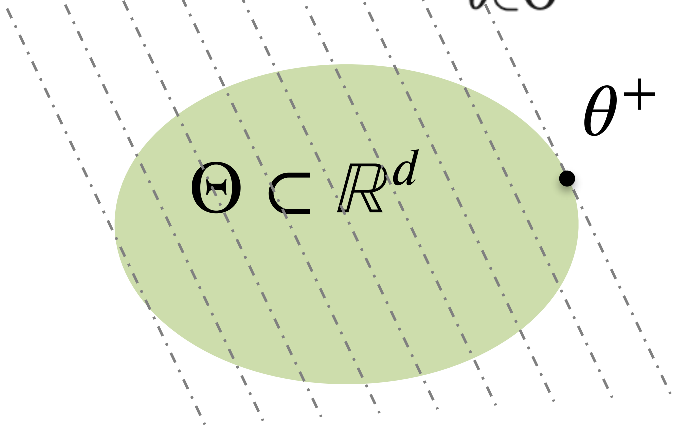
TensorPlan

poly$(d^{A}, A, H, \frac{1}{\delta})$

Weisz, Amortila, Janzer,
Abbasi-Yadkori, Jiang, Sz'20

TensorPlan: Generate + test/rollouts

$$\theta^+ = \arg\max_{\theta\in\Theta} \phi_0(s_0)^\top\theta\,.$$



$\theta \mapsto \pi_\theta$:

$\pi_\theta(s) = a$ for the action a such that

(*) $v_h(s;\theta) = r_a(s) + P_a(s)^\top v_{h+1}(\,\cdot\,;\theta)$

No max!!

provided that $\theta^* \in \Theta$,

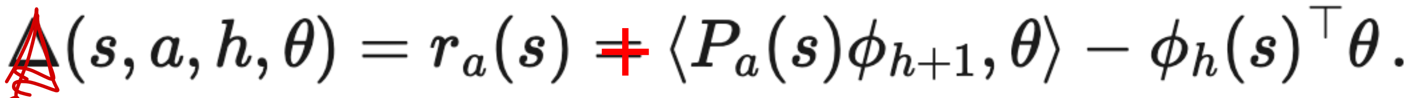
$$v_0(s_0;\theta^+) \ge v_0^*(s_0)\,,$$

Given θ^+, roll out with π_{θ^+} to check whether

1. $v_1\big(s_0;\theta\big)$ is achieved by π_{θ^+}
2. (*) holds

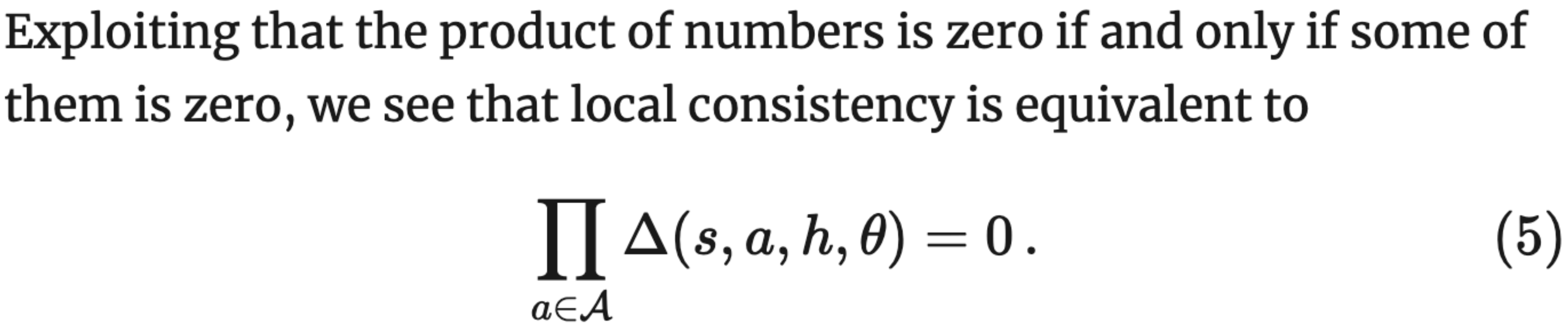
$$v_h(s;\theta) := \theta^\top\phi_h(s)$$

Simplifying assumption: we can measure this exactly



$$\Delta(s,a,h,\theta) = r_a(s) + \langle P_a(s)\phi_{h+1}, \theta\rangle - \phi_h(s)^\top\theta\,.$$

expected next
feature vector
given (s,a)



Exploiting that the product of numbers is zero if and only if some of them is zero, we see that local consistency is equivalent to

$$\prod_{a\in\mathcal{A}} \Delta(s,a,h,\theta) = 0\,. \tag{5}$$

$$\Delta(s,a,h,\theta) = \langle \overline{r_a(s)\,(P_a(s)\phi_{h+1} - \phi_h(s))}, \overline{1\,\theta}\rangle$$

Now, recall that the tensor product ⊗ of vectors satisfies the following property:

$$\prod_a \langle x_a, y_a\rangle = \langle \otimes_a x_a, \otimes_a y_a\rangle\,,$$

$$\underbrace{(d+1) \times (d+1) \times (\quad)}_{A}$$

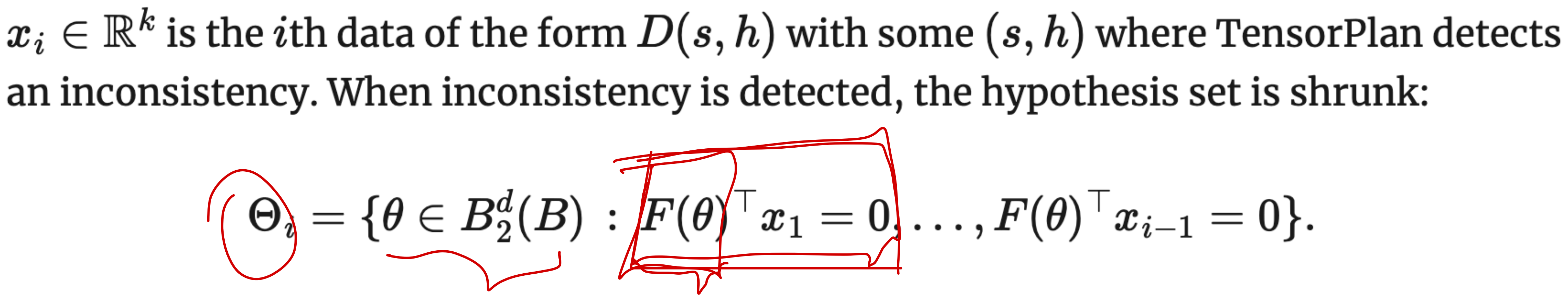
we see that (5), and thus local consistency, is equivalent to

k-dim vector

$$\langle \underbrace{\otimes_a \overline{r_a(s)\,(P_a(s)\phi_{h+1} - \phi_h(s))}}_{D(s,h)}, \underbrace{\otimes_a \overline{1\,\theta}}_{F(\theta)}\rangle = 0\,.$$

$\mathbb{R}^{(d+1)^A}$

$k := (d+1)^A$



$x_i \in \mathbb{R}^k$ is the ith data of the form $D(s, h)$ with some (s, h) where TensorPlan detects an inconsistency. When inconsistency is detected, the hypothesis set is shrunk:

$$\Theta_i = \{\underbrace{\theta \in B_2^d(B)} : \boxed{F(\theta)^\top x_1 = 0}, \ldots, F(\theta)^\top x_{i-1} = 0\}.$$

k -dim

Why does TensorPlan stop?

dim(Theta_i) = dim(Theta_{i-1}) - 1

dim(Theta_1) = k := (d+1)^A

Soundness: when TensorPlan stops

Take a trajectory $S_0^{(i)}, A_0^{(i)}, \ldots, S_{H-1}^{(i)}, A_{H-1}^{(i)}, S_H^{(i)}$ generated during the ith rollout of m rollouts. Since there is no inconsistency along it, for any $0 \le t \le H-1$ we have

$$r_{A_t^{(i)}}(S_t^{(i)}) = v_t(S_t^{(i)};\theta^+) - \langle P_{A_t^{(i)}}(S_t^{(i)}), v_{t+1}(\cdot;\theta^+)\rangle\,. \qquad (6)$$

Hence, with probability $1-\zeta$,

$$v_0^{\pi_{\theta^+}}(s_0) \ge \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{\overset{H}{\cancel{t}}-1} r_{A_t^{(i)}}(S_t^{(i)}) - H\sqrt{\frac{\log(1/\zeta)}{2m}}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{\overset{H}{\cancel{t}}-1} v_t(S_t^{(i)};\theta^+) - \langle\langle P_{A_t^{(i)}}(S_t^{(i)}), v_{t+1}(\cdot;\theta^+)\rangle\rangle - H\sqrt{\frac{\log(2/\zeta)}{2m}}$$

$$\ge \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{\overset{H}{\cancel{t}}-1} v_t(S_t^{(i)};\theta^+) - v_{t+1}(\overset{\downarrow}{S_{t+1}^{(i)}};\theta^+) - (H+2B)\sqrt{\frac{\log(2/\zeta)}{2m}}$$

$$= v_0(s_0;\theta^+) - v_{H+1}(\cdot\,,\cdot) - (H+2B)\sqrt{\frac{\log(2/\zeta)}{2m}}\,,$$

Soundness: since no correct hypothesis ever eliminated, when TensorPlan returns,

$$v_0^{\pi_{\theta^+}}(s_0) \geq v_0(s_0;\theta^+) - \delta \geq \max_{\theta\in\Theta^\circ} v_0^{\pi_\theta}(s_0) - \delta\,.$$

$\rightarrow \left[\dim\left(\{\mathcal{F}(\theta) : \theta \in \Theta^{\theta}\}\right) \right.$

Next lecture:

$\Delta(s, a, h, \theta)$ cannot be measured exactly

Argument goes from $\boxed{\text{exact orthogonality/dim-reduction}}$ of hypothesis space…

…to its noisy version: eluder dimension!

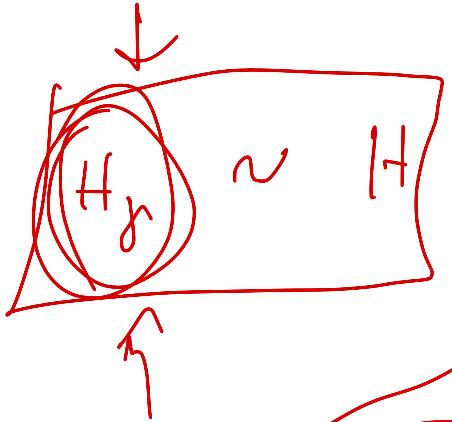
Open questions:

- TensorPlan with q*-realizability, stochastic transitions?
  (problem: max inside expectation)
- Computational efficiency?
- ~~TensorPlan with O(d) actions?~~ Δ

Questions from Slack:

Jiamin He [2:07 AM]
Last time Gellert said the result can be translated to the infinite horizon setting. Can this also be applied to the result for TensorPlan? If yes, why not use the infinite horizon setting and get rid of the subscript h? If no, what are the main difficulties?
+8

$$\sum q^*(s,\pi(s)) - v^*(s) \begin{cases} \geq \exp(d) \\ \exp(\min(d,H)) \end{cases}$$



H_γ
$\sim$
H