

# Question 5.1 on Online Agnostic Policy Learning: A Self-Contained Mathematical Write-Up

## Abstract

This note gives a self-contained write-up of Question 5.1 from Chapter 5 of Gene Li’s thesis *Agnostic Reinforcement Learning: Foundations and Algorithms*. The question asks whether bounded spanning capacity implies an *almost-learnability* guarantee for online agnostic policy learning, namely a sample complexity upper bound quasi-polynomial in the spanning capacity. The document develops the formal setup, defines spanning capacity and its relationship to coverability, records the basic examples from Sections 2.4.2–2.4.3, isolates a correct warm-up positive result for product policy classes, states the main positive and negative results already known for agnostic reinforcement learning, and explains precisely why these results leave Question 5.1 open. The presentation is intended to be mathematically rigorous and largely self-contained, while clearly distinguishing proved facts from conjectural interpretations.

## Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>   | <b>2</b>  |
| <b>2</b> | <b>Formal setup</b>   | <b>2</b>  |
| <b>3</b> | <b>Coverability and spanning capacity</b>   | <b>3</b>  |
| 3.1      | Basic properties . . . . .  | 3         |
| <b>4</b> | <b>The basic examples from Sections 2.4.2–2.4.3</b>   | <b>5</b>  |
| 4.1      | Contextual bandits, tabular classes, and finite classes . . . . .                             | 5         |
| 4.2      | Singletons, $\ell$ -tons, and active-policy classes . . . . .                                 | 5         |
| 4.3      | Product policy classes and parameter sharing . . . . .  | 7         |
| <b>5</b> | <b>Known results around the open problem</b>  | <b>8</b>  |
| 5.1      | Generative-model access: spanning capacity is the right parameter . . . . .                   | 8         |
| 5.2      | Online access beyond the product regime: polynomial dependence on $C(\Pi)$ is false . . . . . | 9         |
| 5.3      | Sunflowers: a general positive result beyond tabular or pure importance sampling . . . . .    | 9         |
| 5.4      | Neither ingredient alone is sufficient . . . . .  | 10        |
| <b>6</b> | <b>The open problem</b>   | <b>11</b> |
| 6.1      | What a positive answer would imply . . . . .  | 11        |
| 6.2      | Why the current theory stops here . . . . .   | 11        |
| 6.3      | The canonical heuristic . . . . .   | 12        |
| <b>7</b> | <b>A nearby companion open problem</b>  | <b>12</b> |
| <b>8</b> | <b>Concluding summary</b>   | <b>12</b> |

# 1 Introduction

A central theme in agnostic reinforcement learning is to understand how the complexity of a policy class  $\Pi$  governs the number of samples needed to compete with the best policy in  $\Pi$ . In the thesis *Agnostic Reinforcement Learning: Foundations and Algorithms* [2], a complexity measure called the *spanning capacity* is introduced and shown to characterize minimax sample complexity under strong simulator access. In the standard online interaction model, however, the picture is subtler. A theorem of Jia, Li, Rakhlin, Sekhari, and Srebro [1] rules out any sample-complexity upper bound polynomial in the spanning capacity alone, while a complementary theorem proves polynomial sample complexity once one assumes an additional structural condition called the *sunflower property*. The gap between these two results is exactly the open problem addressed here.

The purpose of this document is not to propose a new proof. Rather, it is to state the open problem cleanly, collect the exact definitions and examples needed to understand it, and place it against the sharpest known surrounding theorems. In particular, we record one additional warm-up theorem from the supplementary note accompanying this write-up: product policy classes already admit polynomial online sample complexity. This is useful because it shows that the real difficulty in Question 5.1 lies in genuinely non-product, parameter-shared classes.

## Status

To the best of my knowledge, the precise question stated below remains open in the public literature as of March 2026. Later related work on agnostic policy learning and stronger environment-access models continues to treat the online setting as unresolved in this regime [3, 2].

## 2 Formal setup

We work with finite-horizon episodic Markov decision processes. Throughout,  $[H] := \{1, \dots, H\}$ .

**Definition 2.1** (Finite-horizon MDP). An MDP is a tuple

$$M = (\mathcal{X}, \mathcal{A}, P, R, H, d_1),$$

where  $\mathcal{X} = \bigsqcup_{h=1}^H \mathcal{X}_h$  is a layered state space,  $\mathcal{A}$  is the action space,  $P$  is a transition kernel,  $R$  is a reward kernel,  $H$  is the horizon, and  $d_1$  is the initial-state distribution on  $\mathcal{X}_1$ . Starting from  $x_1 \sim d_1$ , the episode proceeds for  $H$  rounds:

$$a_h \in \mathcal{A}, \quad r_h \sim R(x_h, a_h), \quad x_{h+1} \sim P(\cdot | x_h, a_h).$$

We assume rewards are normalized so that  $\sum_{h=1}^H r_h \in [0, 1]$  almost surely.

**Definition 2.2** (Policy class and value). A deterministic Markov policy is a function  $\pi : \mathcal{X} \rightarrow \mathcal{A}$ . A policy class is a set  $\Pi \subseteq \mathcal{A}^{\mathcal{X}}$  of deterministic policies. For  $\pi \in \Pi$ , its value in  $M$  is

$$V^M(\pi) := \mathbb{E} \left[ \sum_{h=1}^H r_h \right].$$

**Definition 2.3** (Agnostic PAC policy learning). Fix  $\varepsilon, \delta \in (0, 1)$ . An algorithm is  $(\varepsilon, \delta)$ -PAC for  $\Pi$  if for every MDP  $M$ , with probability at least  $1 - \delta$ , it returns  $\hat{\pi} \in \Pi$  satisfying

$$V^M(\hat{\pi}) \geq \sup_{\pi \in \Pi} V^M(\pi) - \varepsilon.$$

The *minimax online sample complexity*  $n_{\text{on}}(\Pi; \varepsilon, \delta)$  is the smallest number  $n$  such that some algorithm achieves this guarantee after at most  $n$  online episodes. Likewise,  $n_{\text{gen}}(\Pi; \varepsilon, \delta)$  denotes the corresponding minimax sample complexity under generative-model access.

**Remark 2.4.** The problem is *agnostic*:  $\Pi$  need not contain an optimal policy for the ambient MDP. The learner is required only to compete with the best policy in the reference class.

### 3 Coverability and spanning capacity

The structural quantity at the heart of Question 5.1 is the spanning capacity. It is best introduced alongside coverability.

**Definition 3.1** (State-action occupancy measure). For a policy  $\pi$  and an MDP  $M$ , let  $d_h^\pi(x, a)$  denote the probability that  $(x_h, a_h) = (x, a)$  at layer  $h$  when executing  $\pi$  in  $M$ .

**Definition 3.2** (Coverability). For a policy class  $\Pi$  and MDP  $M$ , define the coverability coefficient

$$C_{\text{cov}}(\Pi, M) := \inf_{\mu^1, \dots, \mu^H} \sup_{\pi \in \Pi, h \in [H]} \left\| \frac{d_h^\pi}{\mu_h} \right\|_\infty,$$

where the infimum is over distributions  $\mu_h$  on  $\mathcal{X}_h \times \mathcal{A}$ . Equivalently,

$$C_{\text{cov}}(\Pi, M) = \max_{h \in [H]} \sum_{(x, a) \in \mathcal{X}_h \times \mathcal{A}} \sup_{\pi \in \Pi} d_h^\pi(x, a).$$

The spanning capacity is the worst-case value of coverability over all MDPs with the same state and action spaces and horizon. Equivalently, it may be defined purely in terms of deterministic MDPs.

**Definition 3.3** (Reachability and spanning capacity). Fix a deterministic MDP  $M \in \mathcal{M}_{\text{det}}$ . For a layer  $h \in [H]$ , let

$$C_h^{\text{reach}}(\Pi, M) := |\{(x, a) \in \mathcal{X}_h \times \mathcal{A} : (x, a) \text{ is reachable at layer } h \text{ by some } \pi \in \Pi\}|.$$

Define

$$C_h(\Pi) := \sup_{M \in \mathcal{M}_{\text{det}}} C_h^{\text{reach}}(\Pi, M), \quad C(\Pi) := \max_{h \in [H]} C_h(\Pi).$$

The quantity  $C(\Pi)$  is the *spanning capacity* of  $\Pi$ .

**Remark 3.4.** Intuitively,  $C(\Pi)$  is the largest number of layer- $h$  state-action pairs that policies from  $\Pi$  can simultaneously expose inside a single deterministic MDP. It is therefore the natural “needle-in-a-haystack” parameter for agnostic policy learning.

#### 3.1 Basic properties

**Proposition 3.5** (Elementary upper bounds). *For every deterministic policy class  $\Pi \subseteq \mathcal{A}^{\mathcal{X}}$ ,*

$$C(\Pi) \leq \min\{|\mathcal{A}|^H, |\Pi|, |\mathcal{X}| |\mathcal{A}|\}.$$

*Proof.* Fix a deterministic MDP  $M$  and layer  $h \in [H]$ .

First, each policy  $\pi \in \Pi$  reaches at most one state-action pair at layer  $h$ , so

$$C_h^{\text{reach}}(\Pi, M) \leq |\Pi|.$$

Second, the total number of state-action pairs available at layer  $h$  is at most  $|\mathcal{X}_h||\mathcal{A}| \leq |\mathcal{X}||\mathcal{A}|$ , hence

$$C_h^{\text{reach}}(\Pi, M) \leq |\mathcal{X}||\mathcal{A}|.$$

Third, every reachable state-action pair at layer  $h$  is determined by an action sequence of length  $h$ , and there are at most  $|\mathcal{A}|^h \leq |\mathcal{A}|^H$  such sequences in a deterministic MDP. Hence

$$C_h^{\text{reach}}(\Pi, M) \leq |\mathcal{A}|^H.$$

Taking the supremum over  $M \in \mathcal{M}_{\text{det}}$  and then the maximum over  $h \in [H]$  proves the claim.  $\square$

The next lemma, proved in the thesis and in [1], shows that spanning capacity is exactly the worst-case coverability over *all* MDPs, not merely deterministic ones.

**Lemma 3.6** (Worst-case coverability equals spanning capacity). *For every deterministic policy class  $\Pi$ ,*

$$\sup_{M \in \mathcal{M}_{\text{sto}}} C_{\text{cov}}(\Pi, M) = C(\Pi).$$

*Proof.* We sketch a self-contained proof. Fix an MDP  $M \in \mathcal{M}_{\text{sto}}$  and a layer  $h$ . Write

$$\Gamma_h(M) := \sum_{(x,a) \in \mathcal{X}_h \times \mathcal{A}} \sup_{\pi \in \Pi} d_h^\pi(x, a).$$

By the equivalent formula for coverability,

$$C_{\text{cov}}(\Pi, M) = \max_{h \in [H]} \Gamma_h(M).$$

We first show  $\Gamma_h(M) \leq C_h(\Pi)$ . Expand the occupancy of a policy at layer  $h$  by summing over all state-action prefixes:

$$\Gamma_h(M) = \sum_{x_1, a_1, \dots, x_h, a_h} \sup_{\pi \in \Pi} \mathbf{1}\{\pi \rightsquigarrow (x_1, a_1, \dots, x_h, a_h)\} \prod_{t=1}^{h-1} P(x_{t+1} \mid x_t, a_t),$$

where  $\pi \rightsquigarrow (x_1, a_1, \dots, x_h, a_h)$  means that the policy  $\pi$  is consistent with the state-action prefix. Bounding each transition probability by a supremum over the next state and repeatedly moving the suprema outside gives

$$\Gamma_h(M) \leq \sum_{a_1} \sup_{x_2} \sum_{a_2} \sup_{x_3} \cdots \sup_{x_h} \sum_{a_h} \mathbf{1}\{\exists \pi \in \Pi : \pi \rightsquigarrow (x_1, a_1, \dots, x_h, a_h)\}.$$

The right-hand side is exactly the cumulative reachability that can be realized at layer  $h$  inside a deterministic MDP obtained by choosing, at each prior state-action pair, a maximizing successor state. Hence

$$\Gamma_h(M) \leq C_h(\Pi).$$

Taking the maximum over  $h$  and then the supremum over  $M \in \mathcal{M}_{\text{sto}}$  yields

$$\sup_{M \in \mathcal{M}_{\text{sto}}} C_{\text{cov}}(\Pi, M) \leq C(\Pi).$$

For the reverse inequality, every deterministic MDP is a special case of a stochastic MDP, and for deterministic MDPs one has

$$\sup_{\pi \in \Pi} d_h^\pi(x, a) = \mathbf{1}\{(x, a) \text{ is reachable by some } \pi \in \Pi\}.$$

Therefore

$$C(\Pi) = \sup_{M \in \mathcal{M}_{\text{det}}} C_{\text{cov}}(\Pi, M) \leq \sup_{M \in \mathcal{M}_{\text{sto}}} C_{\text{cov}}(\Pi, M).$$

Combining the two inequalities completes the proof.  $\square$

## 4 The basic examples from Sections 2.4.2–2.4.3

We now record the examples emphasized in the thesis. Throughout this section the action space is  $\mathcal{A} = \{0, 1\}$ , and each layer  $h$  contains a large enough set of labeled states  $\mathcal{X}_h = \{x^{(i,h)} : i \in [K]\}$ , where  $K$  is chosen large enough so that all constructions below make sense.

### 4.1 Contextual bandits, tabular classes, and finite classes

**Example 4.1** (Contextual bandits). When  $H = 1$ , every deterministic MDP has a single initial state and at most  $|\mathcal{A}|$  reachable actions. Hence

$$C(\Pi) \leq |\mathcal{A}|$$

for every policy class  $\Pi$ .

**Example 4.2** (Tabular class). Let  $\Pi_{\text{tab}} := \mathcal{A}^{\mathcal{X}}$  be the full class of deterministic Markov policies on a finite state space. Then

$$C(\Pi_{\text{tab}}) \leq \min\{|\mathcal{A}|^H, |\mathcal{X}| |\mathcal{A}|\}$$

by Proposition 3.5, and this order is attained up to constants by the obvious tree and counting constructions.

**Example 4.3** (Finite classes). For every finite policy class,

$$C(\Pi) \leq |\Pi|.$$

This is the second bound in Proposition 3.5.

### 4.2 Singletons, $\ell$ -tons, and active-policy classes

**Definition 4.4** (Singletons). Define

$$\Pi_{\text{sing}} := \{\pi_{(i,h)} : i \in [K], h \in [H]\},$$

where  $\pi_{(i,h)}$  takes action 1 on the single state  $x^{(i,h)}$  and takes action 0 everywhere else.

**Definition 4.5** ( $\ell$ -tons). For an integer  $\ell \geq 1$ , define

$$\Pi_\ell := \{\pi_I : I \subseteq \mathcal{X}, |I| \leq \ell\}, \quad \pi_I(x) := \mathbf{1}\{x \in I\}.$$

Thus a policy in  $\Pi_\ell$  may take action 1 on at most  $\ell$  states in the entire layered state space.

**Definition 4.6** (One-active and all-active classes). For a bit-string  $b \in \{0, 1\}^H$ , define  $\pi_b$  as follows.

(a) The *one-active* class  $\Pi_{1\text{-act}}$  consists of all policies such that

$$\pi_b(x) = \begin{cases} b[h], & x = x^{(1,h)}, \\ 0, & \text{otherwise.} \end{cases}$$

(b) For each fixed  $j \in [K]$ , let  $\Pi_{j\text{-act}}$  be the analogous class with the distinguished state  $x^{(j,h)}$  in each layer. Define the *all-active* class

$$\Pi_{\text{act}} := \bigcup_{j=1}^K \Pi_{j\text{-act}}.$$

The thesis proves the following growth rates.

**Proposition 4.7** (Spanning-capacity estimates for the basic examples). *The following bounds hold.*

$$\begin{aligned} C(\Pi_{\text{sing}}) &= H + 1, \\ C(\Pi_\ell) &= \Theta(H^\ell) \quad (\ell \text{ fixed}), \\ C(\Pi_{1\text{-act}}) &= \Theta(H), \\ C(\Pi_{\text{act}}) &= \Theta(H^2). \end{aligned}$$

*Proof.* We prove the upper bounds that are used throughout the thesis; matching lower bounds follow from explicit deterministic layered constructions.

**Singletons.** Fix a deterministic MDP and a layer  $h$ . Consider the reference policy  $\pi_0$  that always plays 0. Any singleton policy differs from  $\pi_0$  at exactly one state and one layer. In a deterministic MDP, such a deviation can create at most one new branch at each earlier layer. Hence the number of reachable state-action pairs at layer  $h$  is at most  $h + 1$ . Conversely, one can realize  $h + 1$  reachable pairs at layer  $h$  by arranging  $h$  sequential opportunities to deviate from the all-zero path. Therefore  $C(\Pi_{\text{sing}}) = H + 1$ .

**$\ell$ -tons.** Let  $C_h(\Pi_\ell)$  denote the worst-case cumulative reachability at layer  $h$ . We prove by induction on  $h$  that

$$C_h(\Pi_\ell) \leq 2h^\ell.$$

The case  $h = 1$  is trivial. For the inductive step, fix a deterministic MDP. At the initial state, policies that take action 1 have only  $\ell - 1$  future opportunities to play action 1, whereas policies that take action 0 still belong to  $\Pi_\ell$ . Therefore

$$C_h(\Pi_\ell) \leq C_{h-1}(\Pi_{\ell-1}) + C_{h-1}(\Pi_\ell) \leq 2(h-1)^{\ell-1} + 2(h-1)^\ell \leq 2h^\ell.$$

Thus  $C(\Pi_\ell) \leq 2H^\ell$ . A matching lower bound of order  $H^\ell$  is realized by a layered construction in which the learner chooses the layers at which action 1 is used; this is the content of the thesis example [2, Section 2.4.3].

**One-active class.** For a deterministic MDP and a layer  $h$ , let  $\bar{X}_h$  be the set of states reachable by policies in  $\Pi_{1\text{-act}}$  at layer  $h$ . We claim  $|\bar{X}_h| \leq h$  by induction. The base case  $h = 1$  is obvious. If the claim holds at layer  $h$ , then from any reachable state at layer  $h$ , all policies take action 0 except possibly on the single distinguished state  $x^{(1,h)}$ , where one extra action is possible. Hence

$$|\bar{X}_{h+1}| \leq |\bar{X}_h| + 1 \leq h + 1.$$

Therefore  $C(\Pi_{1\text{-act}}) \leq 2H$ , and a matching linear lower bound is obtained by the obvious chain construction.

**All-active class.** At the first layer, at most one distinguished chain  $\Pi_{j\text{-act}}$  can branch by taking action 1; all policies that take action 0 merge back to a common successor. Consequently,

$$C_h(\Pi_{\text{act}}) \leq C_{h-1}(\Pi_{\text{act}}) + \max_j C_{h-1}(\Pi_{j\text{-act}}).$$

Using the one-active estimate  $\max_j C_{h-1}(\Pi_{j\text{-act}}) \lesssim h$ , telescoping gives

$$C_h(\Pi_{\text{act}}) \lesssim \sum_{t=1}^{h-1} t \lesssim h^2.$$

Thus  $C(\Pi_{\text{act}}) \lesssim H^2$ , and a matching quadratic lower bound is again obtained by an explicit layered construction.  $\square$

**Remark 4.8.** The examples above are the motivating evidence for Question 5.1. Each has bounded spanning capacity, and each is online learnable with polynomial sample complexity by the sunflower theorem discussed later. The difficulty is to decide whether *every* bounded-spanning-capacity class is at least quasi-polynomially learnable.

### 4.3 Product policy classes and parameter sharing

The thesis emphasizes an important distinction between *product* policy classes and policy classes with *parameter sharing across layers*. This distinction is directly relevant to Question 6.1.

**Definition 4.9** (Product policy class). Assume throughout this subsection that  $\mathcal{A} = \{0, 1\}$ . A policy class  $\Pi \subseteq \mathcal{A}^{\mathcal{X}}$  is called a *product policy class* if there exist classes

$$\Pi_h \subseteq \{0, 1\}^{\mathcal{X}_h}, \quad h \in [H],$$

such that

$$\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_H.$$

Equivalently, a policy in  $\Pi$  is obtained by choosing independently one layer-wise decision rule from each  $\Pi_h$ .

In such classes there is no cross-layer parameter sharing: choices made at one layer place no combinatorial restrictions on the choices available at later layers. By contrast, many of the thesis's motivating examples—notably singleton-type and array-of-combination-lock classes—are *not* product classes.

The thesis records the following useful recurrence for the spanning capacity of product classes.

**Proposition 4.10** (Inductive characterization for product classes [2, Proposition 2.3]). *Let  $\Pi = \Pi_1 \times \cdots \times \Pi_H$  be a binary-action product policy class. For each layer  $h$ , define*

$$M_h := |\{(x, a) \in \mathcal{X}_h \times \mathcal{A} : \exists \pi_h \in \Pi_h \text{ with } \pi_h(x) = a\}|$$

and

$$N_h := |\{x \in \mathcal{X}_h : \exists \pi_h, \pi'_h \in \Pi_h \text{ with } \pi_h(x) \neq \pi'_h(x)\}|.$$

If  $C_0(\Pi) := 1$ , then for every  $h \in [H]$ ,

$$C_h(\Pi) = \min\{M_h, 2C_{h-1}(\Pi), C_{h-1}(\Pi) + N_h\}.$$

This recurrence makes transparent why product classes are substantially easier to analyze: all of the layer- $h$  combinatorics are encoded by the one-layer disagreement quantity  $N_h$ , with no additional blow-up coming from cross-layer coupling.

**Theorem 4.11** (Warm-up positive result for product classes). *Let  $\Pi = \Pi_1 \times \cdots \times \Pi_H$  be a finite binary-action product policy class. Then there is an online agnostic PAC algorithm whose sample complexity is polynomial in*

$$C(\Pi), \quad H, \quad \varepsilon^{-1}, \quad \log |\Pi|, \quad \log(1/\delta).$$

*In particular, product policy classes satisfy Question 6.1 in a much stronger form: for this subclass one already has a polynomial, rather than merely quasi-polynomial, dependence on the spanning capacity.*

*Comment.* A complete proof is given in the supplementary note accompanying this write-up. The key point for the present discussion is conceptual: the absence of parameter sharing across layers allows one to combine the recurrence in Proposition 4.10 with a layer-wise online exploration-and-estimation argument, yielding a polynomial-in- $C(\Pi)$  bound. Since the purpose of this note is to frame the genuinely open regime rather than to reprove that warm-up theorem in full detail, we record only the result and its significance here.  $\square$

**Remark 4.12.** Theorem 4.11 sharpens the interpretation of Question 6.1. The unresolved difficulty does not already occur for classes without cross-layer coupling. Rather, the real challenge is to understand *non-product* classes, where parameter sharing can keep the spanning capacity small while still creating intricate sequential dependencies that are invisible in a one-layer analysis.

## 5 Known results around the open problem

### 5.1 Generative-model access: spanning capacity is the right parameter

The starting point is the simulator-access theory of spanning capacity.

**Theorem 5.1** (Upper bound under generative access [2, Theorem 4.1]). *For every deterministic policy class  $\Pi$ ,*

$$n_{\text{gen}}(\Pi; \varepsilon, \delta) \leq O\left(\frac{H C(\Pi)}{\varepsilon^2} \log \frac{|\Pi|}{\delta}\right).$$

**Theorem 5.2** (Lower bound under generative access [2, Theorem 4.2]). *For every deterministic policy class  $\Pi$ ,*

$$n_{\text{gen}}(\Pi; \varepsilon, \delta) \geq \Omega\left(\frac{C(\Pi)}{\varepsilon^2} \log \frac{1}{\delta}\right).$$

Together, Theorems 5.1 and 5.2 show that spanning capacity characterizes minimax agnostic PAC learnability in the generative-model setting, up to an  $H \log |\Pi|$  factor.

A closely related negative theorem, important for context, shows that one cannot generally replace the worst-case parameter  $C(\Pi)$  by the instance-dependent coverability coefficient  $C_{\text{cov}}(\Pi, M)$ , even under stronger access models [2, 3]. Thus the online problem is *not* simply about finding a finer MDP-dependent complexity measure that automatically interpolates between easy and hard instances.

## 5.2 Online access beyond the product regime: polynomial dependence on $C(\Pi)$ is false

The main negative theorem for online interaction is the following.

**Theorem 5.3** (Online lower bound [2, Theorem 5.1]; [1, Section 5]). *There exist universal constants  $h_0 \in \mathbb{N}$  and  $c \in (0, 1)$  such that the following holds. Fix any  $H \geq h_0$ ,  $\ell \in \{2, \dots, H\}$ , and*

$$\varepsilon \in \left(2^{-cH}, \frac{1}{100H}\right) \quad \text{such that} \quad \varepsilon^{-\ell} \leq 2^H.$$

*Then there exists a policy class  $\Pi^{(\ell)}$  of size  $1/(6\varepsilon^\ell)$  with*

$$C(\Pi^{(\ell)}) \leq O(H^{4\ell+2}),$$

*and a family of binary-action horizon- $H$  MDPs such that any  $(\varepsilon/16, 1/8)$ -PAC online algorithm must collect at least*

$$\min \left\{ \frac{1}{120\varepsilon^\ell}, 2^{H/3-3} \right\}$$

*episodes in expectation on some MDP in the family.*

**Corollary 5.4.** *There is no general online agnostic PAC upper bound of the form*

$$n_{\text{on}}(\Pi; \varepsilon, \delta) \leq \text{poly}(C(\Pi), H, \varepsilon^{-1}, \log(1/\delta), \log |\Pi|)$$

*with polynomial dependence on  $C(\Pi)$  alone.*

*Proof.* Fix  $\ell$  as a constant. Then Theorem 5.3 gives policy classes with  $C(\Pi^{(\ell)}) = \text{poly}(H)$  but sample complexity at least  $\Omega(\varepsilon^{-\ell})$ , which is superpolynomial in  $H$  for the permitted parameter regime. Hence no bound polynomial in  $C(\Pi)$  can hold uniformly.  $\square$

**Remark 5.5.** Crucially, Theorem 5.3 refutes only *polynomial* dependence on  $C(\Pi)$ . It does *not* rule out quasi-polynomial dependence such as  $\varepsilon^{-O(\log C(\Pi))}$ . This is why Question 5.1 remains genuinely open.

## 5.3 Sunflowers: a general positive result beyond tabular or pure importance sampling

The known positive theorem for online RL adds another structural condition.

**Definition 5.6** (Petal [2, Definition 5.2]). Let  $\bar{\Pi}$  be a collection of policies and let  $\bar{X} \subseteq \mathcal{X}$ . A policy  $\pi$  is called an  $\bar{X}$ -*petal* on  $\bar{\Pi}$  if the following holds: whenever a partial trajectory  $\tau = (x_h, a_h, \dots, x_{h'}, a_{h'})$  is consistent with  $\pi$ , either  $\tau$  is also consistent with some policy in  $\bar{\Pi}$ , or else  $\tau$  passes through a state in  $\bar{X}$  after its starting point.

**Definition 5.7** (Sunflower property [2, Definition 5.3]). A policy class  $\Pi$  is a  $(K, D)$ -*sunflower* if there exists a set  $\Pi_{\text{core}}$  of Markov policies with  $|\Pi_{\text{core}}| \leq K$  such that for every  $\pi \in \Pi$ , there exists a set  $X_\pi \subseteq \mathcal{X}$  with  $|X_\pi| \leq D$  for which  $\pi$  is an  $X_\pi$ -petal on  $\Pi_{\text{core}}$ .

**Theorem 5.8** (POPLER theorem [2, Theorem 5.2]; [1, Section 6]). *Suppose  $\Pi$  has spanning capacity  $C(\Pi)$  and is a  $(K, D)$ -sunflower. Then there is an online algorithm (POPLER) such that for every MDP  $M$ , with probability at least  $1 - \delta$ , the algorithm returns an  $\varepsilon$ -optimal policy after*

$$\tilde{O} \left( \left( \frac{1}{\varepsilon^2} + \frac{HD^6 C(\Pi)}{\varepsilon^4} \right) K^2 \log \frac{|\Pi|}{\delta} \right)$$

*online episodes.*

**Remark 5.9.** This theorem interpolates between two classical extremes:

- pure importance sampling, corresponding morally to small  $K$  and  $D = 0$ ;
- tabular exploration, corresponding morally to  $K = 0$  and large  $D$ .

The proof relies on *policy-specific Markov reward processes* and on the ability to estimate many policies simultaneously by combining importance sampling with state-identification subroutines.

**Proposition 5.10** (The basic examples are sunflowers [2, Section 5.3.2]). *Each of the policy classes from Proposition 4.7 is a  $(K, D)$ -sunflower with  $K, D = \text{poly}(H)$ . More concretely:*

$$\begin{aligned} \Pi_{\text{sing}} & \text{ is an } (H + 1, 1)\text{-sunflower,} \\ \Pi_{\ell} & \text{ is an } (H + 1, \ell)\text{-sunflower,} \\ \Pi_{1\text{-act}} & \text{ is an } (H + 1, H)\text{-sunflower,} \\ \Pi_{\text{act}} & \text{ is an } (H + 1, H)\text{-sunflower.} \end{aligned}$$

*Proof sketch.* For each class, the core consists of the all-zero policy together with the  $H$  policies that play action 1 throughout a single layer. The corresponding petal set  $X_{\pi}$  is the set of states on which  $\pi$  may differ from the core. Any partial trajectory consistent with  $\pi$  either avoids those special states and is therefore also consistent with some core policy, or else it passes through  $X_{\pi}$ . This is exactly the petal condition.  $\square$

The sunflower theorem shows that all the motivating examples with small spanning capacity are indeed online learnable with polynomial sample complexity. This leaves open the possibility that bounded spanning capacity may still imply a weaker, quasi-polynomial guarantee in full generality.

## 5.4 Neither ingredient alone is sufficient

It is useful to record two converse facts.

**Proposition 5.11** (Spanning capacity alone is insufficient). *This is Corollary 5.4, a consequence of Theorem 5.3.*

**Proposition 5.12** (Sunflower structure alone is insufficient). *There exists a policy class with  $(K, D) = (1, H)$ -sunflower structure but exponential spanning capacity  $C(\Pi) = 2^H$ . Consequently, bounded  $(K, D)$  alone does not imply sample-efficient agnostic PAC learning.*

*Proof.* Consider a complete binary tree of depth  $H$  as a deterministic MDP, and define one policy  $\pi_{\tau}$  for each root-to-leaf trajectory  $\tau = (x_1, a_1, \dots, x_H, a_H)$  by

$$\pi_{\tau}(x) = \begin{cases} a_h, & x = x_h \text{ lies on } \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Then all  $2^H$  state-action pairs at depth  $H$  are reachable, so  $C(\Pi) = 2^H$ . On the other hand, if  $\pi_0$  denotes the always-zero policy, then every  $\pi_{\tau}$  is an  $X_{\tau}$ -petal on  $\{\pi_0\}$ , where  $X_{\tau}$  is the set of states lying on  $\tau$ . Thus the class is a  $(1, H)$ -sunflower.  $\square$

## 6 The open problem

We can now state the question cleanly.

**Question 6.1** (Question 5.1; almost learnability from spanning capacity). Does every policy class with bounded spanning capacity satisfy a quasi-polynomial online agnostic PAC bound? More precisely, is it true that

$$n_{\text{on}}(\Pi; \varepsilon, \delta) \lesssim \varepsilon^{-O(\log C(\Pi))} \text{poly}\left(H, \log |\Pi|, \log \frac{1}{\delta}\right)$$

for every finite deterministic policy class  $\Pi$ ?

The phrasing “almost learnable” refers to the fact that the conjectured upper bound is much weaker than a polynomial bound in  $C(\Pi)$ , but still far stronger than the trivial bound obtained by naive enumeration over  $\Pi$ .

### 6.1 What a positive answer would imply

A positive answer to Question 6.1 would place the minimax online sample complexity between the following lower and upper scales:

$$\Omega\left(\frac{C(\Pi)}{\varepsilon^2} \log \frac{1}{\delta}\right) \lesssim n_{\text{on}}(\Pi; \varepsilon, \delta) \lesssim \varepsilon^{-O(\log C(\Pi))} \text{poly}\left(H, \log |\Pi|, \log \frac{1}{\delta}\right).$$

The left-hand inequality is inherited from the generative-model lower bound, since online interaction is only harder than simulator access. The right-hand inequality is exactly the conjectural quasi-polynomial guarantee.

Such a theorem would say that, although online interaction may be exponentially harder than generative-model access, the sole policy-class parameter  $C(\Pi)$  still controls learnability up to a quasi-polynomial loss.

### 6.2 Why the current theory stops here

The known results leave a precise gap:

- (i) Theorems 5.1 and 5.2 show that  $C(\Pi)$  is exactly the right complexity measure under generative-model access.
- (ii) Theorem 4.11 shows that in the restricted but important subclass of product policies, one already has a polynomial online guarantee.
- (iii) Theorem 5.3 shows that  $C(\Pi)$  does *not* yield polynomial online sample complexity in general.
- (iv) Theorem 5.8 shows that  $C(\Pi)$ , together with the sunflower property, *does* yield polynomial online sample complexity.

What is not known is whether the sunflower theorem can be replaced, in full generality, by a weaker statement depending only on  $C(\Pi)$ , at the cost of allowing a quasi-polynomial dependence on  $1/\varepsilon$ .

### 6.3 The canonical heuristic

The thesis proposes the following heuristic route [2, Section 5.4]: perhaps the POPLER analysis can be generalized so that the core and petal sets are not fixed in advance by the policy class alone, but instead are discovered in an MDP-dependent manner during learning. This would amount to finding a *data-driven sunflower decomposition* for the actually reachable portion of the environment.

At present, however, this is only a heuristic. No theorem of this form is known, and no matching stronger lower bound is known either.

## 7 A nearby companion open problem

For completeness, we also record the companion question from the same section of the thesis.

**Question 7.1** (Question 5.2; necessity of sunflower-type structure). Is some form of the sunflower property necessary for online agnostic PAC learning? More concretely, must every online-learnable policy class admit a decomposition whose complexity is measured by a small core size and small petal size?

This question is logically distinct from Question 6.1, but they are closely related. A positive answer to Question 5.2 would suggest that sunflower-like structure is a fundamental combinatorial shadow of online learnability. A negative answer would point toward some different structural mechanism underlying efficient online learning.

## 8 Concluding summary

The current mathematical picture can be summarized as follows.

- (1) Spanning capacity  $C(\Pi)$  is exactly the right minimax complexity measure for agnostic policy learning under generative-model access.
- (2) In the restricted subclass of product policy classes, bounded spanning capacity already implies polynomial online sample complexity.
- (3) In the full online interaction model, bounded spanning capacity alone does not imply polynomial sample complexity.
- (4) Adding the sunflower property restores polynomial sample complexity and covers all of the motivating examples with small spanning capacity.
- (5) The central unresolved issue is therefore whether bounded spanning capacity still implies a weaker quasi-polynomial guarantee for genuinely non-product classes, of the form  $\varepsilon^{-O(\log C(\Pi))}$ .

Thus Question 5.1 isolates a natural and very sharp boundary in the current theory of agnostic reinforcement learning. It asks whether the spanning capacity remains, in a weaker but still meaningful sense, the governing combinatorial parameter for online policy learning.

## References

- [1] Z. Jia, G. Li, A. Rakhlin, A. Sekhari, and N. Srebro. When is agnostic reinforcement learning statistically tractable? In *Advances in Neural Information Processing Systems*, 2023. arXiv:2310.06113.
- [2] G. Li. *Agnostic Reinforcement Learning: Foundations and Algorithms*. PhD thesis, Toyota Technological Institute at Chicago, 2025. arXiv:2506.01884.
- [3] A. Krishnamurthy, G. Li, and A. Sekhari. The role of environment access in agnostic reinforcement learning. arXiv:2504.05405, 2025.
- [4] Supplementary note on Question 5.1 (provided by the user, 2026). Unpublished note containing a warm-up polynomial upper bound for product policy classes.
- [5] M. Kearns, Y. Mansour, and A. Ng. Approximate planning in large POMDPs via reusable trajectories. In *Advances in Neural Information Processing Systems*, 1999.
- [6] T. Xie, D. J. Foster, Y. Bai, N. Jiang, and S. M. Kakade. The role of coverage in online reinforcement learning. arXiv:2210.04157, 2022.